SPAD 7.0

Data management . Analyse des Données . Data Mining

GUIDE
DU
DATA MINER

Statistiques descriptives - Analyses factorielles - Typologies
Modèle linéaire – Analyses discriminantes –
Scoring – Arbres de décision

Coheris Spad
63 bd de Ménilmontant - 75011 PARIS
Tel : 01.42.81.17.47 – Fax : 01.42.81.16.96
spad@spad.eu – www.spad.eu
Siret : 478 173 644 00020 - APE : 722A
Numéro de déclaration formation : 11-75-41226-75

Introduction à SPAD

Guide du Data Miner

ISBN : 2-906711-X

Pour tous renseignements complémentaires sur le logiciel SPAD, les formations et Etudes/Conseils, consultez le site www.spad.eu ou écrivez-nous :

Thème	E-mail
Logiciels SPAD	info@spad.eu
Support Technique SPAD	support@spad.eu
Formation	formations@spad.eu
Etudes-Consulting	consulting@spad.eu
Librairie	publications@spad.eu

Pour tous renseignements complémentaires sur l'offre du groupe COHERIS (CRM, BI, Data Mining, Data Quality Management, Merchandising Sfa), consultez le site www.coheris.com ou contactez :

Coheris Spad
63 bd de Ménilmontant - 75011 PARIS
Tel : 01.42.81.17.47 – Fax : 01.42.81.16.96
spad@spad.eu – www.spad.eu
Siret : 478 173 644 00020 - APE : 722A
Numéro de déclaration formation : 11-75-41226-75

Avant-propos

Dans ce manuel, le lecteur découvrira les principales méthodes d'analyse de données et de data mining du logiciel SPAD au travers d'exemples illustrés et commentés.
Il apprendra à paramétrer et interpréter chaque méthode ainsi qu'à manipuler les principaux éditeurs graphiques associés.

Dans le manuel « Guide de l'utilisateur », le lecteur sera guidé pour faire ses premiers pas avec le logiciel SPAD. Il apprendra à manipuler l'interface et découvrira l'ensemble des possibilités offertes pour la gestion et la préparation des données.

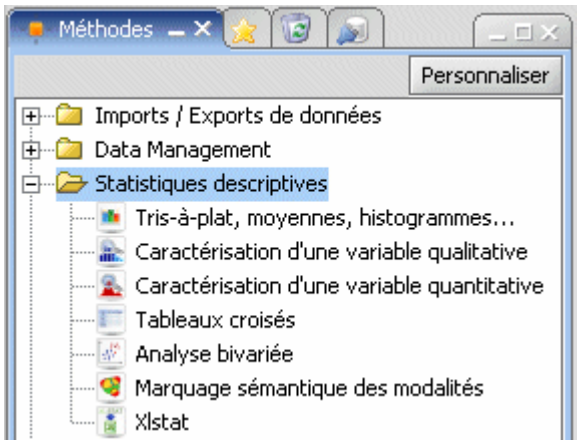
L'ambition de ce manuel est de donner un aperçu assez global du logiciel. Elle n'est pas de faire du lecteur un expert car la description de toutes les possibilités du logiciel engendrerait un volume propre à décourager les meilleures volontés.

Cependant, l'aide en ligne, disponible à tout endroit du logiciel, apportera les réponses aux questions de l'utilisateur au moment où elles se posent.

L'équipe de SPAD qui a participé collectivement à ce travail, remercie celles et ceux qui ont apporté le plus grand soin à la relecture de ce document. Mais naturellement, nous assumons la responsabilité des erreurs et imperfections que le lecteur attentif ne manquera pas de découvrir.

Table des matières

STATISTIQUES DESCRIPTIVES AVEC SPAD



STATS : Tris à plat / Histogramme

DEMOD : Caractérisation automatique d'une variable qualitative

DESCO : Caractérisation automatique d'une variable quantitative

TABLE : Tableaux croisés

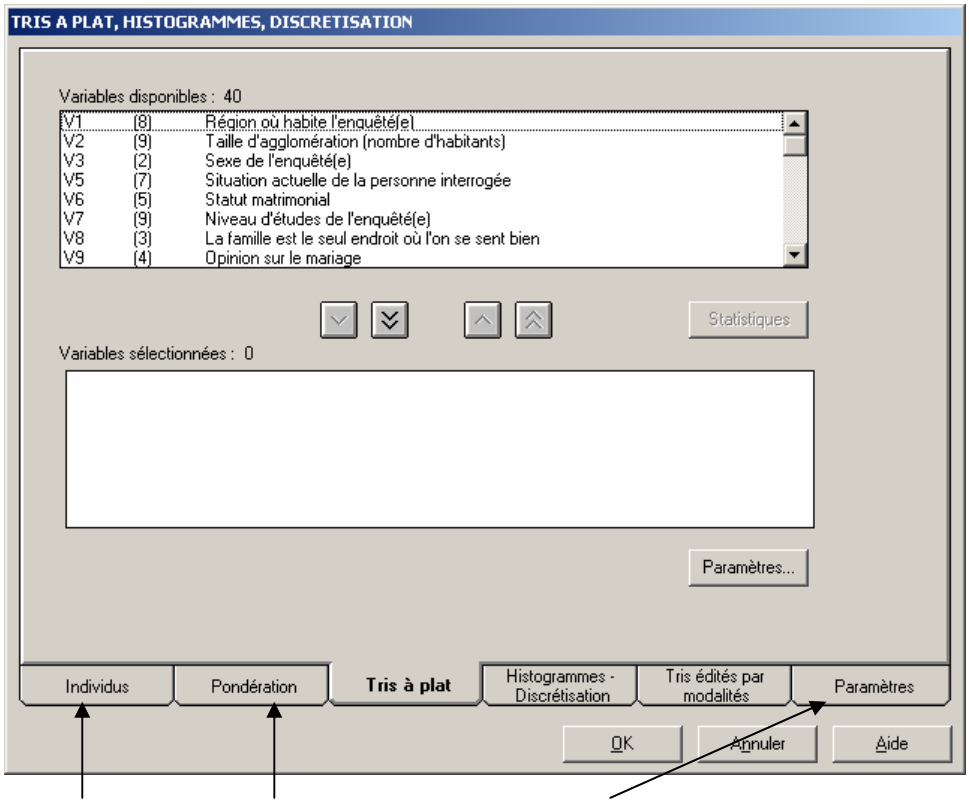
BIVAR : Analyse bivariée

STATS - Tris à plat / Histogrammes

Cette procédure fournit l'ensemble des statistiques élémentaires sur les variables nominales ou continues.

La base ENQUETE.SBA, fichier d'enquête d'opinion, va être utilisée comme exemple.

La fenêtre des paramètres qui s'ouvre est structurée en **onglets** qui regroupent logiquement les différents paramètres de la méthode.



Les onglets « **Individus** », « **Pondération** » et « **Paramètres** » apparaissent dans presque toutes les méthodes de SPAD.

Individus : Sélection des individus par filtre logique, échantillonnage ou sur liste
Pondération : Choix ou calcul des poids de redressement
Paramètres : Options et paramètres de calcul et d'édition

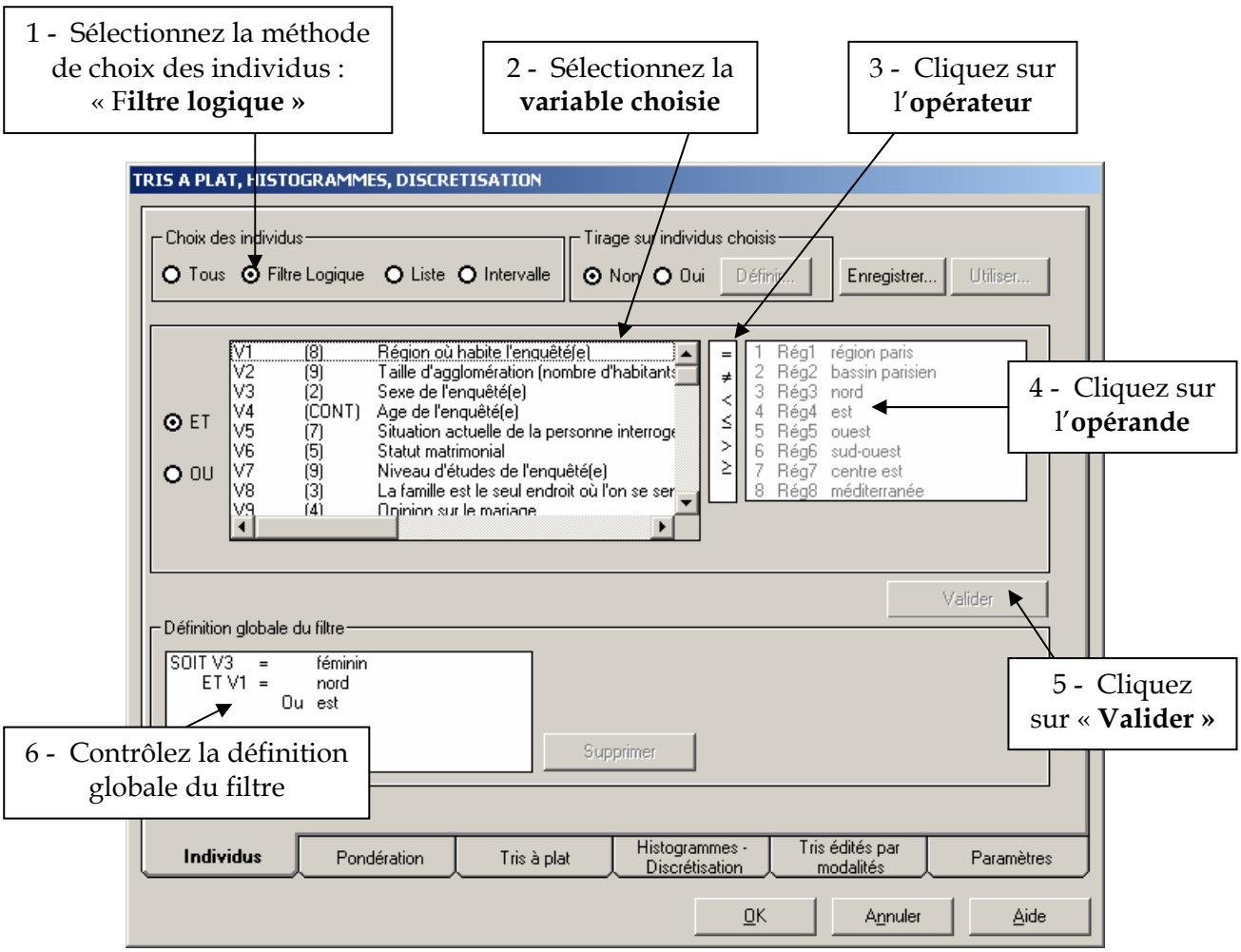
Les fiches **Individus** et **Pondération** sont identiques quelque soit la méthode.
Nous allons les présenter une fois pour toute.

L'ONGLET « **INDIVIDUS** »

La fiche « individus » permet de sélectionner les individus avec une des méthodes suivantes :

- ✓ tous les individus disponibles
- ✓ un ou plusieurs filtres logiques
- ✓ une liste nominative d'individus
- ✓ une sélection dans un ou plusieurs intervalles
- ✓ un tirage aléatoire

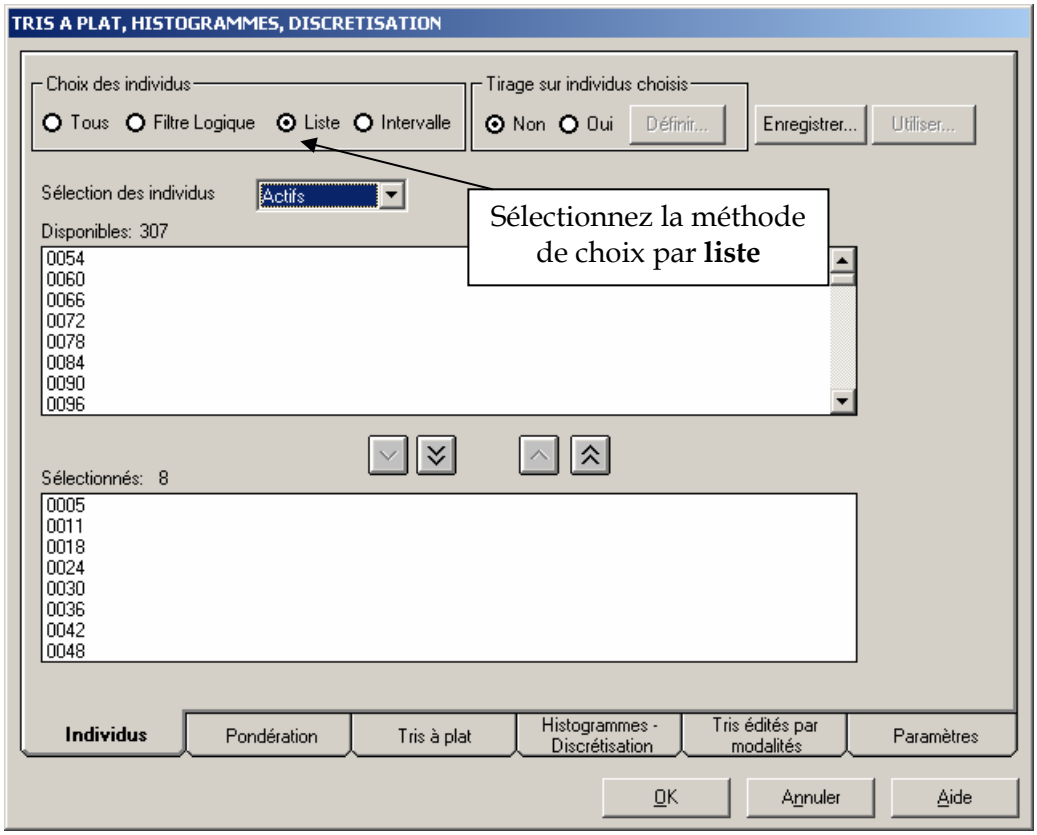
Appliquer un « **FILTRE LOGIQUE** »



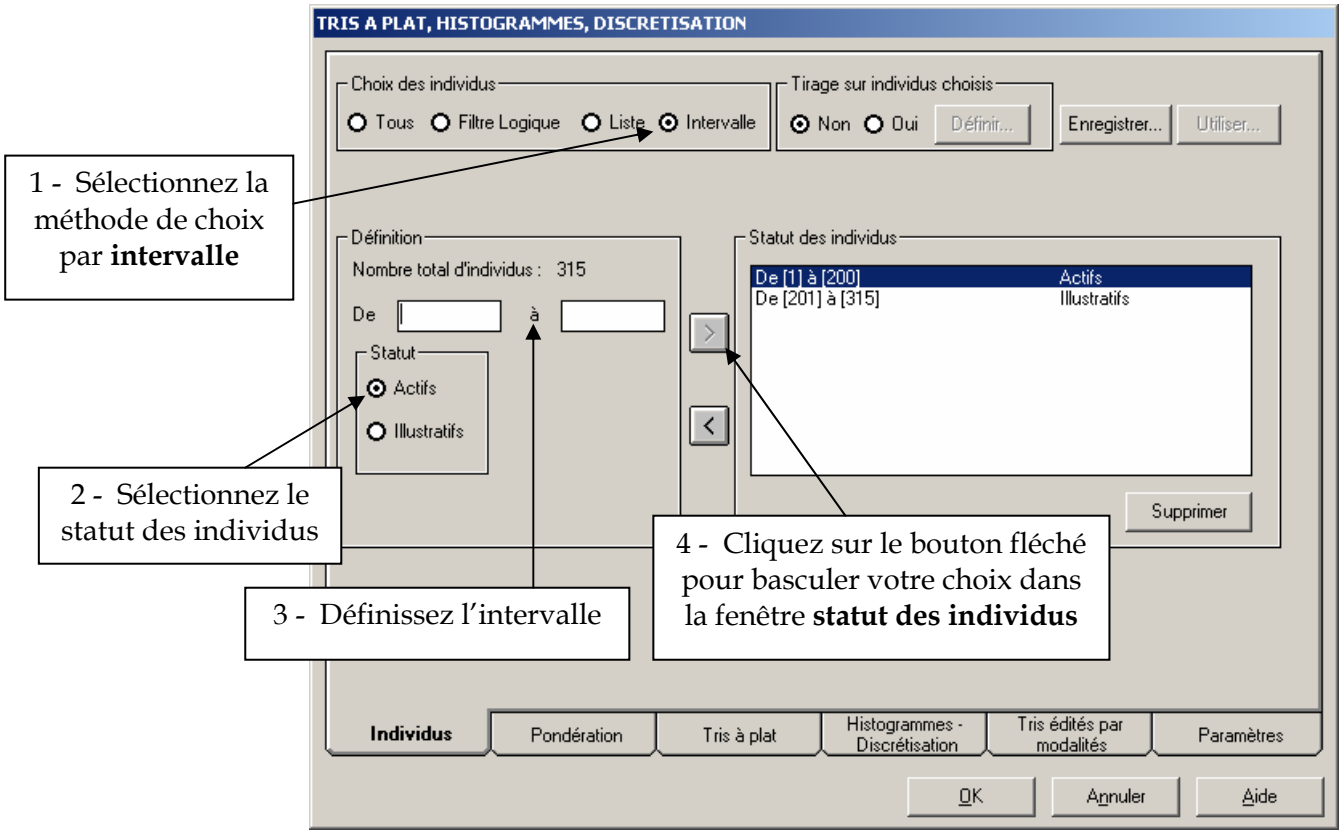
En cas d'erreur, vous pouvez supprimer une expression de filtre en sélectionnant l'expression à écarter et en cliquant sur « **Supprimer** ».

Les individus satisfaisant au filtre sont considérés comme actifs, les autres, comme illustratifs.

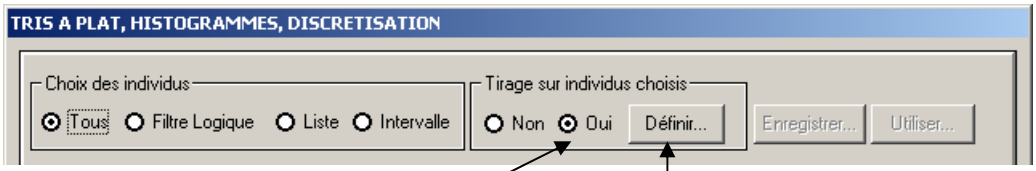
Sélectionner les individus dans une « **LISTE** »



Sélectionner les individus par « **INTERVALLE** »



Effectuer un « TIRAGE ALEATOIRE »

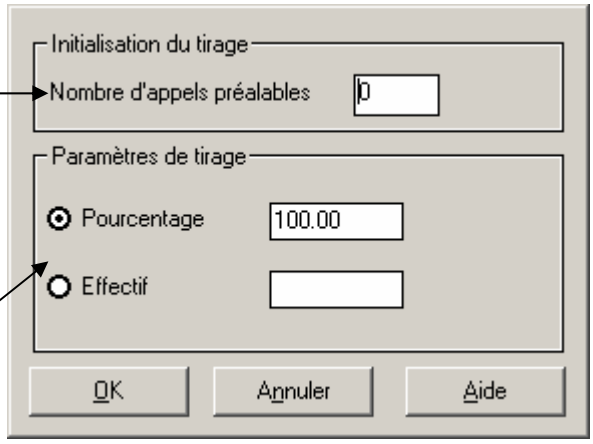


1 - Cliquez sur « O**ui** » pour effectuer un tirage aléatoire

2 - Cliquez sur **définir** pour paramétrer le tirage aléatoire

Indiquez éventuellement le nombre d'appels préalable au tirage au sort. Lors d'une autre exécution de la sélection, vous ne changerez la valeur que si vous souhaitez obtenir des tirages différents.

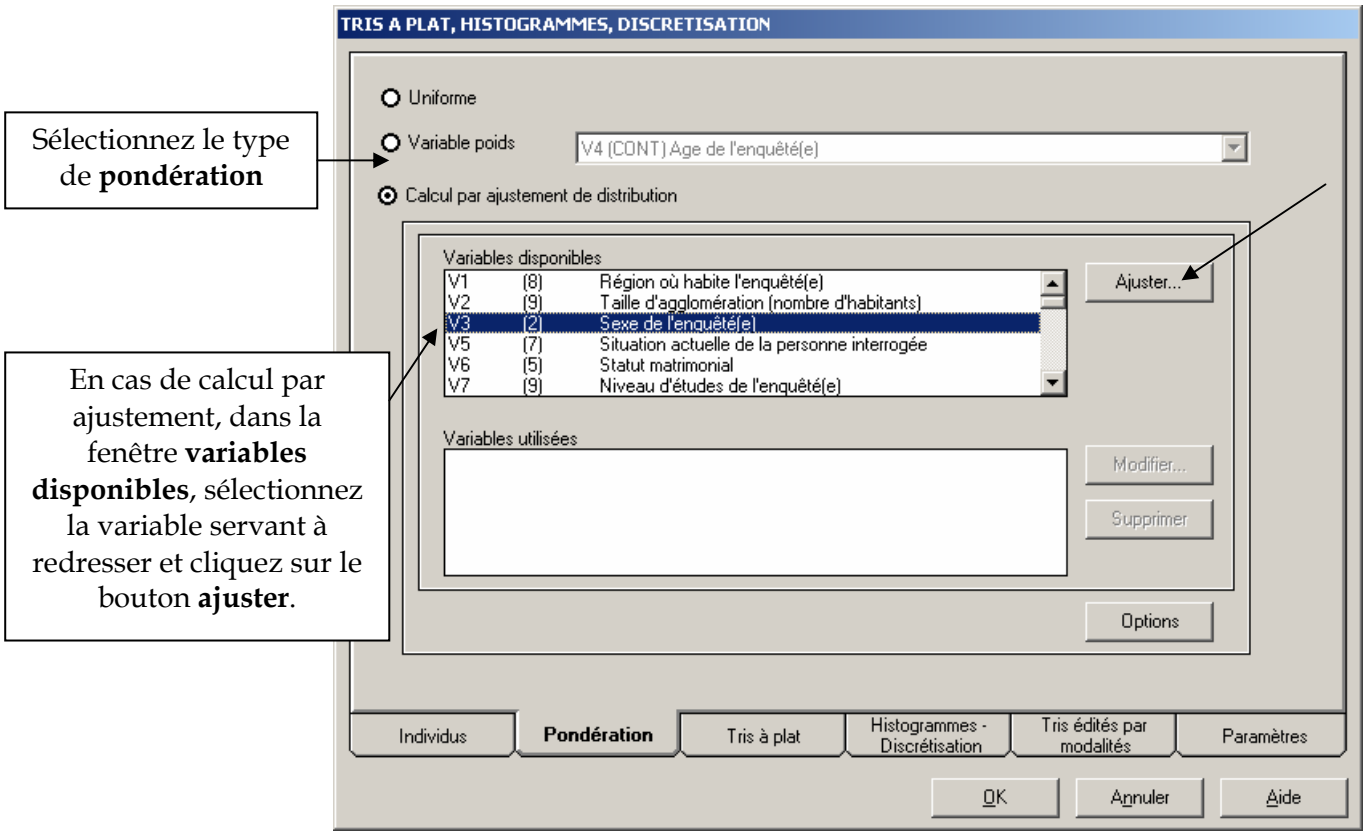
Saisissez le **pourcentage** du tirage au sort ou l'**effectif** de l'échantillon après tirage

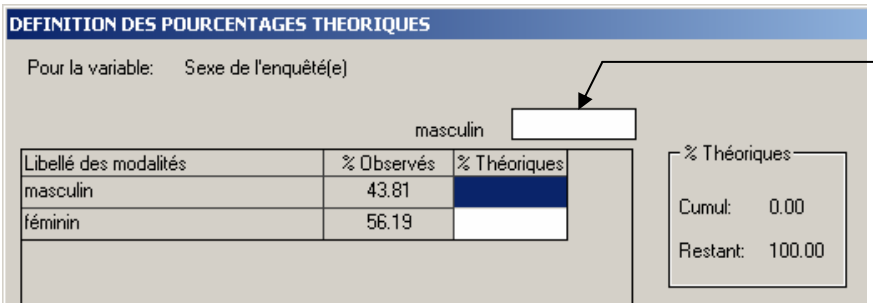


L'ONGLET « **PONDERATION** »

L'onglet pondération ouvre une fiche pour redresser la répartition des individus dans l'échantillon :

- ✓ selon une variable poids existante dans le fichier
- ✓ en fonction d'un ou plusieurs pourcentages théoriques (calcul par ajustement)





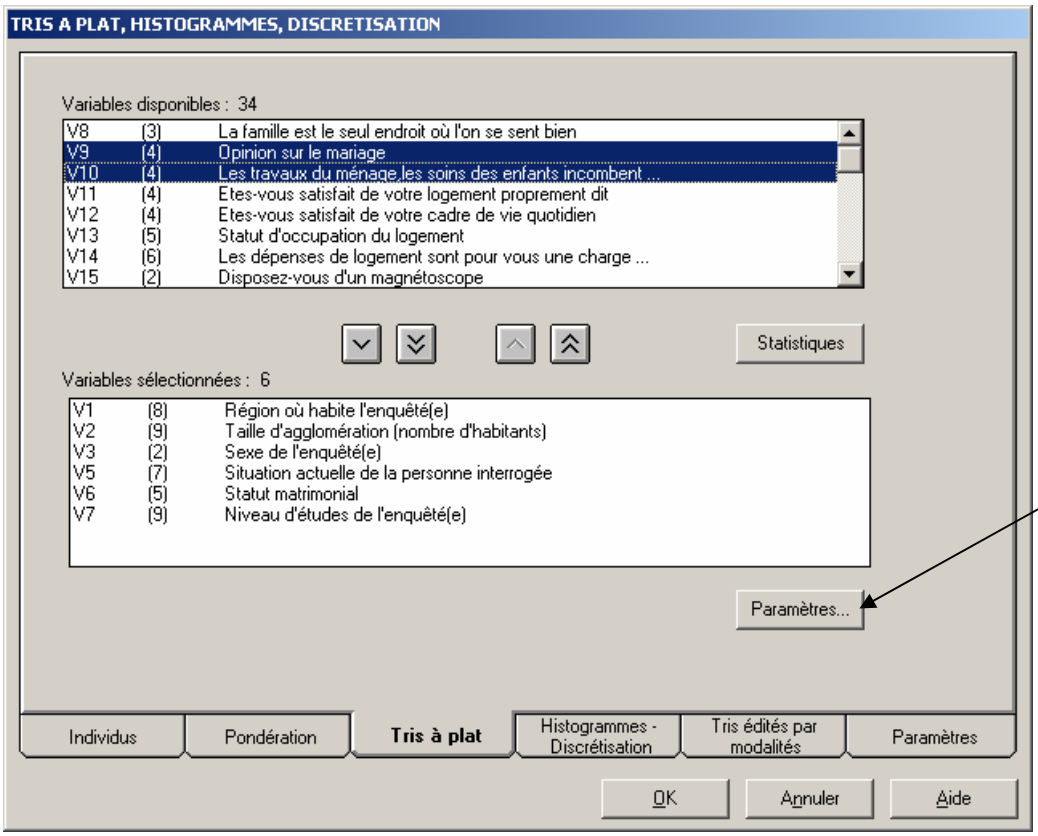
Vous pouvez renouveler cette opération pour une autre variable. Ainsi vous obtiendrez un redressement en fonction de plusieurs variables.

Attention : Le calcul de la pondération dans l'onglet pondération d'une méthode est temporaire (la variable poids n'est pas conservée). Cette fonctionnalité permet de faire rapidement des essais et de mesurer l'influence de la pondération sur les résultats de la méthode. Une fois obtenue une variable de pondération satisfaisante, il est préférable de créer définitivement la variable de pondération avec le menu « **Outils** » – « **Redressement** » du menu principal.

Nous allons maintenant nous intéresser aux onglets spécifiques de la méthode STATS.

L'ONGLET « **TRIS A PLAT** »

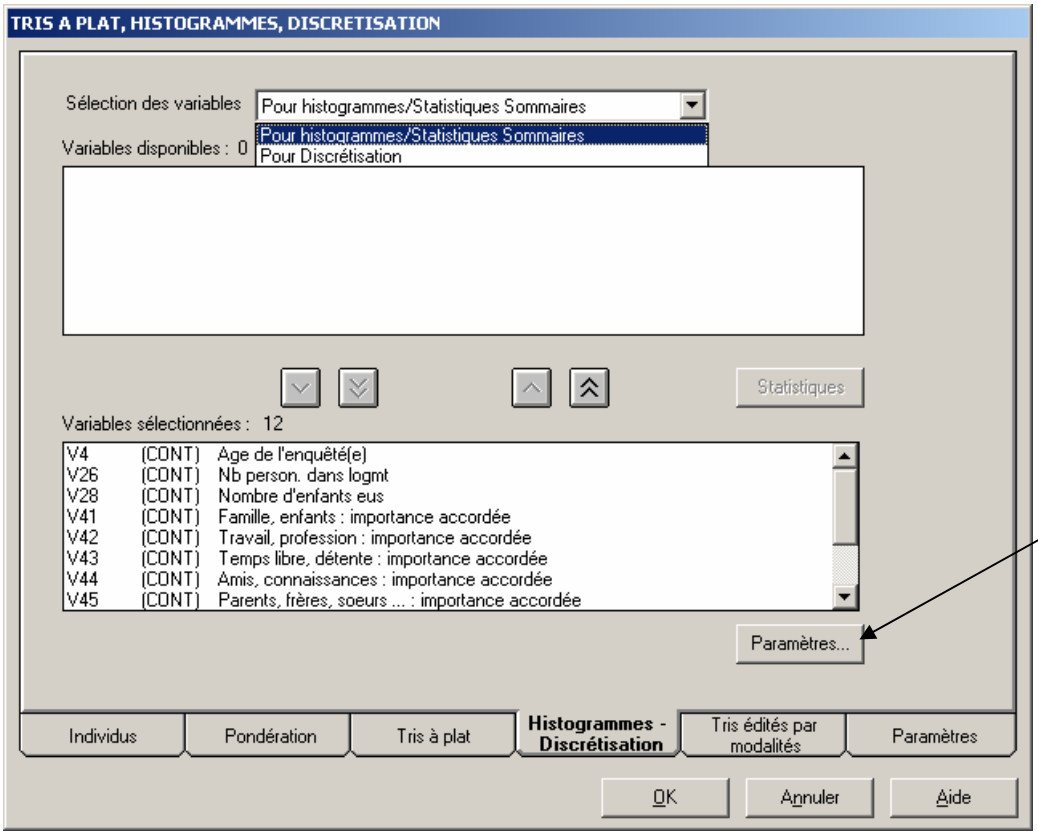
On va sélectionner les variables nominales pour lesquelles on souhaite obtenir le tri à plat.



Le bouton « **Paramètres** » permet d'obtenir l'édition ou non des modalités de poids nuls et gère l'édition des données manquantes : elles peuvent être recodées pour créer une modalité appelée « réponse manquante », sinon elles ne sont pas éditées.

L'ONGLET « HISTOGRAMMES - DISCRETISATION »

Cet onglet sert à sélectionner les variables continues pour obtenir l'édition des statistiques principales et les histogrammes.



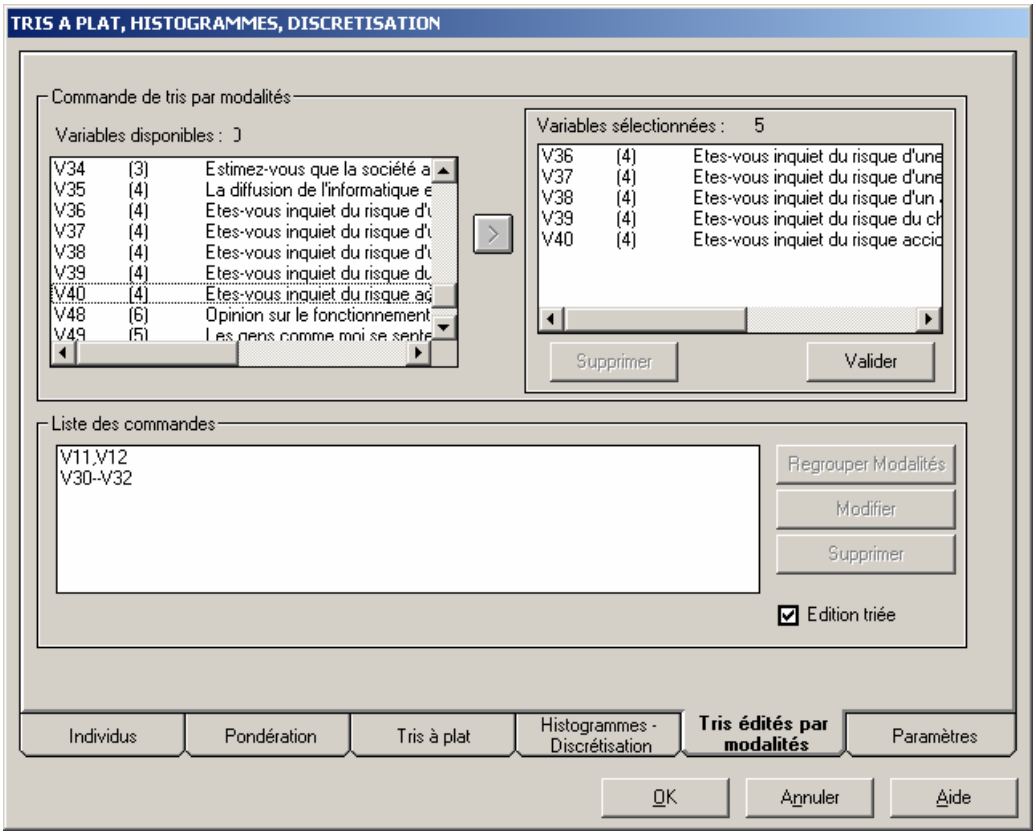
Le bouton « **Paramètres** » gère les éditions et contrôle la forme des histogrammes : nombre de classes, borne minimale et maximale, unité des barres des histogrammes.
Ces paramètres peuvent être choisis localement pour chaque variable continue ou globalement pour l'ensemble des histogrammes.

On peut également sélectionner les variables continues devant faire l'objet d'une « **discrétisation** ». Chaque valeur distincte de la variable est alors listée avec sa fréquence.
L'intérêt de la « discrétisation » est de connaître toutes les valeurs distinctes de la variable.
La « discrétisation » peut aussi être une étape intermédiaire vers une mise en classes précise d'une variable continue.

On ne peut pas demander à la fois l'histogramme et la discrétisation d'une même variable continue.

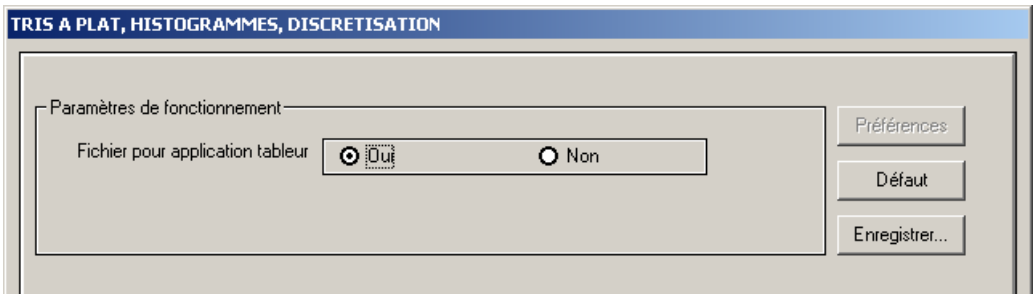
L'ONGLET « **TRIS EDITES PAR MODALITES** »

Cette commande permet de traiter des « batteries » de variables similaires : ces variables doivent avoir le même nombre de modalités.



L'ONGLET « **PARAMETRES** »

Cet onglet permet d'exporter les résultats vers Excel pour en faire une édition soignée et y ajouter des graphiques personnalisés.



Lorsque vous avez passé en revue tous les onglets, y compris les sous fiches optionnelles, validez votre paramétrage en cliquant sur « **OK** ».

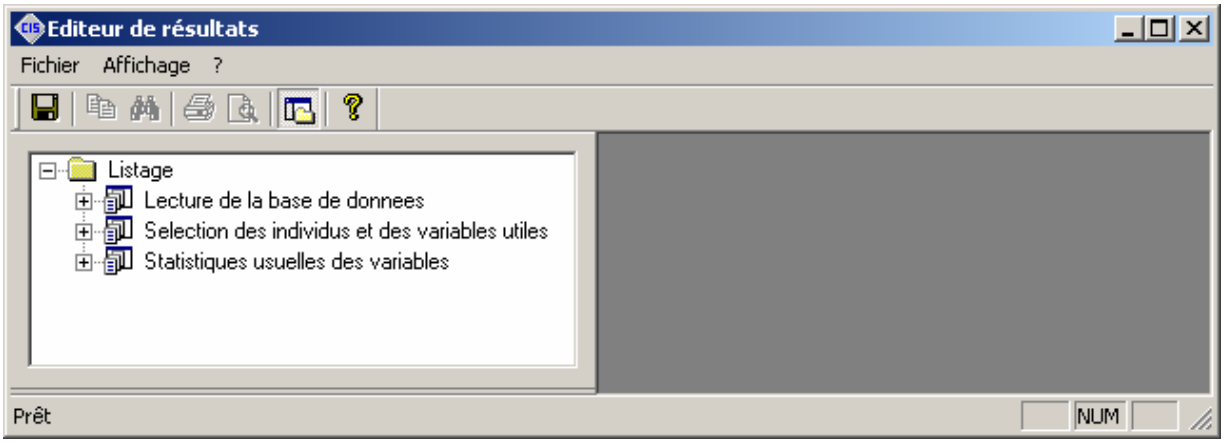
Accédez au menu contextuel de la méthode et cliquez sur Exécuter.

RESULTATS DE L'EXECUTION DE LA METHODE

Suite à l'exécution de la méthode, les résultats sont accessibles par clic droit sur la méthode ou à partir de la vue « Exécutions ». Ces icônes permettent d'accéder aux **résultats numériques** et **graphiques** de la méthode STATS exécutée.

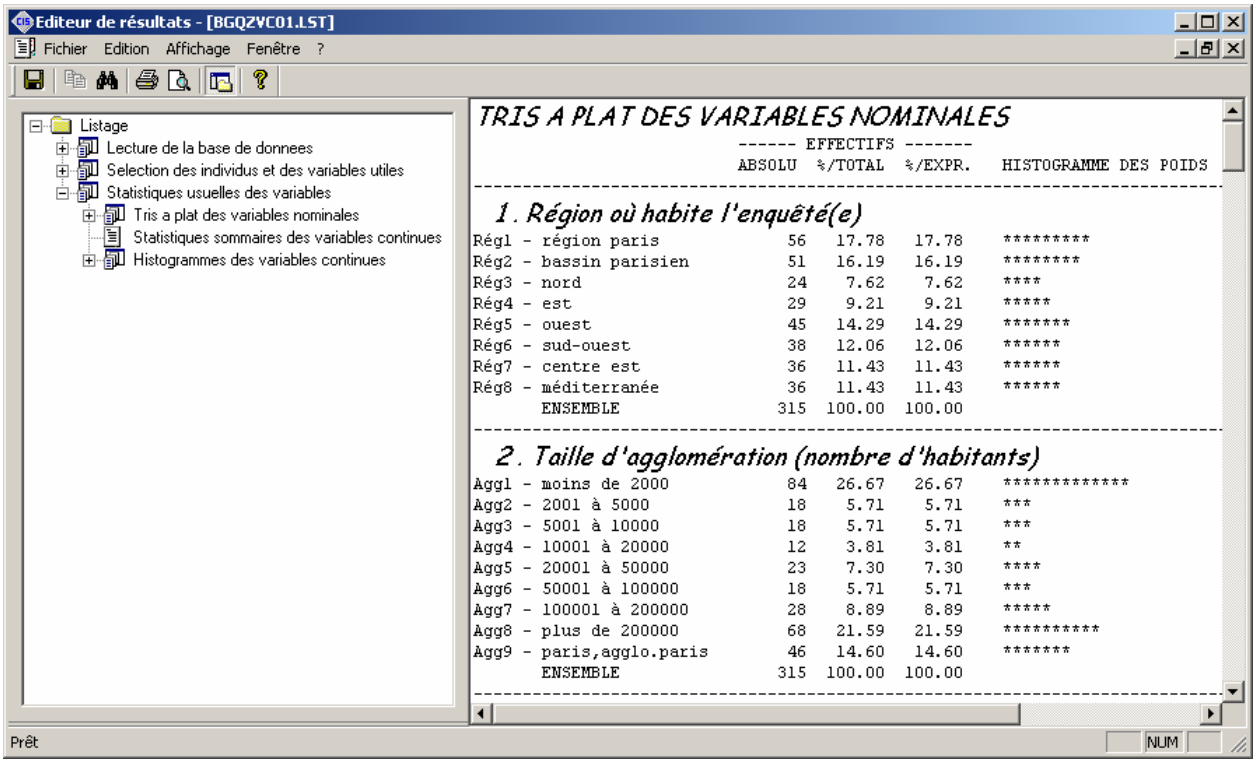
L'EDITEUR DE RESULTATS

L'éditeur de résultats se présente sous la forme d'une nouvelle fenêtre.



Les informations du listage sont structurées sous forme d'une arborescence.
En cliquant sur [+], vous ouvrez une arborescence et en cliquant sur [-] vous la refermez.

Par un double-clique sur un titre, vous affichez les résultats correspondants dans une nouvelle fenêtre.



L'option « **Mise en forme** » du menu « **Fichier** » facilite la lecture des résultats à l'écran.

LES RESULTATS DE STATS

TRIS A PLAT DES VARIABLES NOMINALES

	ABSOLU	%/TOTAL	%/EXPR.	HISTOGRAMME DES POIDS
	------ EFFECTIFS -------			
1 . Région où habite l'enquêté(e)				
Rég1 - région paris	56	17.78	17.78	*********
Rég2 - bassin parisien	51	16.19	16.19	********
Rég3 - nord	24	7.62	7.62	****
Rég4 - est	29	9.21	9.21	*****
Rég5 - ouest	45	14.29	14.29	*******
Rég6 - sud-ouest	38	12.06	12.06	******
Rég7 - centre est	36	11.43	11.43	******
Rég8 - méditerranée	36	11.43	11.43	******
ENSEMBLE	315	100.00	100.00	
2 . Taille d'agglomération (nombre d'habitants)				
Agg1 - moins de 2000	84	26.67	26.67	*************
Agg2 - 2001 à 5000	18	5.71	5.71	***
Agg3 - 5001 à 10000	18	5.71	5.71	***
Agg4 - 10001 à 20000	12	3.81	3.81	**
Agg5 - 20001 à 50000	23	7.30	7.30	****
Agg6 - 50001 à 100000	18	5.71	5.71	***
Agg7 - 100001 à 200000	28	8.89	8.89	*****
Agg8 - plus de 200000	68	21.59	21.59	**********
Agg9 - paris,agglo.paris	46	14.60	14.60	*******
ENSEMBLE	315	100.00	100.00	
3 . Sexe de l'enquêté(e)				
Sex1 - masculin	138	43.81	43.81	**********
Sex2 - féminin	177	56.19	56.19	**********
ENSEMBLE	315	100.00	100.00	

TRIS A PLAT DES VARIABLES DISCRETISEES

	ABSOLU	%/TOTAL	%/EXPR.	% CUM.	HISTOGRAMME DES POIDS
	--------- EFFECTIFS ----------				
41 . Famille, enfants : importance accordée					
1.000	5	1.59	1.59	1.59	*
2.000	2	0.63	0.63	2.22	*
3.000	2	0.63	0.63	2.86	*
4.000	8	2.54	2.54	5.40	**
5.000	11	3.49	3.49	8.89	**
6.000	16	5.08	5.08	13.97	***
7.000	271	86.03	86.03	100.00	**********
ENSEMBLE	315	100.00	100.00		
42 . Travail, profession : importance accordée					
1.000	14	4.44	4.44	4.44	**
2.000	5	1.59	1.59	6.03	*
3.000	6	1.90	1.90	7.94	*
4.000	13	4.13	4.13	12.06	**
5.000	48	15.24	15.24	27.30	*******
6.000	61	19.37	19.37	46.67	*********
7.000	168	53.33	53.33	100.00	**********
ENSEMBLE	315	100.00	100.00		

STATISTIQUES SOMMAIRES DES VARIABLES CONTINUES

EFFECTIF TOTAL : 315
POIDS TOTAL : 315.00

NUM . IDEN - LIBELLE	EFFECTIF	POIDS	MOYENNE	ECART-TYPE	MINIMUM	MAXIMUM	MIN.2	MAX.2
4 . Age - Age de l'enquêté(e)	315	315.00	43.756	16.581	18.000	86.000	19.000	83.000
26 . Nbpr - Nb person. dans logm	315	315.00	3.063	1.408	1.000	8.000	2.000	7.000
28 . Nbef - Nombre d'enfants eus	315	315.00	1.860	1.671	0.000	9.000	1.000	8.000
50 . PrFm - Prestat° familiales	283	283.00	533.795	926.899	0.000	5100.000	15.000	4980.000
51 . Salr - Salaire mens. de l'e	267	267.00	4408.547	4575.339	0.000	40000.000	300.000	24000.000

TRIS A PLAT DE VARIABLES GROUPEES

	EFFECTIFS ABSOLU	EFFECTIFS %/TOTAL	EFFECTIFS %/EXPR.
DISTRIBUTION DE LA REPONSE : beaucoup			
POUR LES VARIABLES			
Etes-vous inquiet du risque d'une maladie grave	146.00	46.35	46.35
Etes-vous inquiet du risque du chômage	125.00	39.68	39.68
Etes-vous inquiet du risque d'un accident de la route	115.00	36.51	36.51
Etes-vous inquiet du risque d'une agression dans la rue	92.00	29.21	29.21
Etes-vous inquiet du risque accident centrale nucléaire	89.00	28.25	28.25
DISTRIBUTION DE LA REPONSE : assez			
POUR LES VARIABLES			
Etes-vous inquiet du risque d'un accident de la route	98.00	31.11	31.11
Etes-vous inquiet du risque d'une maladie grave	76.00	24.13	24.13
Etes-vous inquiet du risque du chômage	71.00	22.54	22.54
Etes-vous inquiet du risque accident centrale nucléaire	53.00	16.83	16.83
Etes-vous inquiet du risque d'une agression dans la rue	46.00	14.60	14.60
DISTRIBUTION DE LA REPONSE : un peu			
POUR LES VARIABLES			
Etes-vous inquiet du risque accident centrale nucléaire	94.00	29.84	29.84
Etes-vous inquiet du risque d'une agression dans la rue	94.00	29.84	29.84
Etes-vous inquiet du risque d'un accident de la route	64.00	20.32	20.32
Etes-vous inquiet du risque d'une maladie grave	54.00	17.14	17.14
Etes-vous inquiet du risque du chômage	51.00	16.19	16.19
DISTRIBUTION DE LA REPONSE : pas du tout			
POUR LES VARIABLES			
Etes-vous inquiet du risque d'une agression dans la rue	83.00	26.35	26.35
Etes-vous inquiet du risque accident centrale nucléaire	79.00	25.08	25.08
Etes-vous inquiet du risque du chômage	68.00	21.59	21.59
Etes-vous inquiet du risque d'une maladie grave	39.00	12.38	12.38
Etes-vous inquiet du risque d'un accident de la route	38.00	12.06	12.06

Caractérisation d'une variable qualitative

L'intérêt de cette procédure est de caractériser une variable qualitative (ou nominale) particulière en explorant automatiquement l'ensemble des liaisons qu'elle entretient avec toutes les autres variables du fichier, quelque soit leur type.

Le tableau suivant résume les possibilités de caractérisation statistique proposées par la procédure DEMOD :

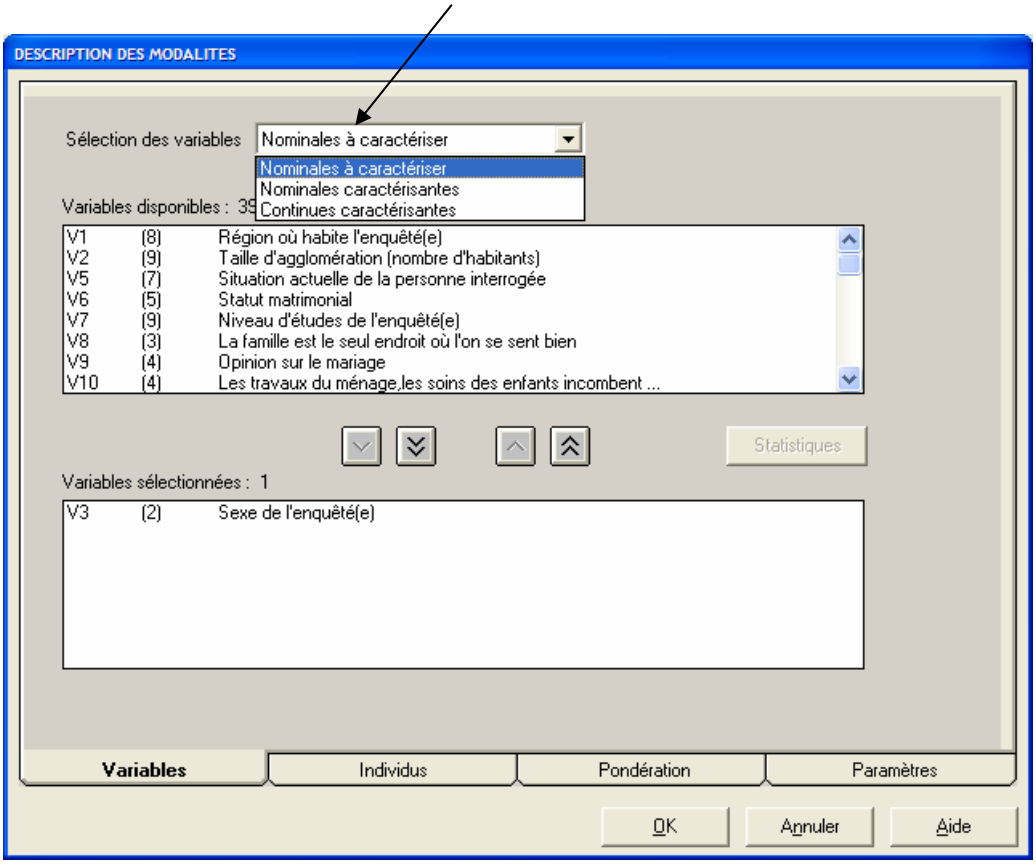
Eléments à caractériser	Eléments caractérisants
• Des groupes d'individus (définis par les modalités de la variable nominale à caractériser) *En d'autres termes, on décrit chaque modalité de la variable à caractériser par l'ensemble des éléments caractérisants.*	• Les modalités • Les variables nominales • Les variables continues
• La variable nominale à caractériser *On recherche parmi tous les éléments caractérisants ceux dont la liaison avec la variable nominale à caractériser est la plus significative.*	• Les modalités • Les variables nominales • Les variables continues

Un groupe d'individus est défini par une modalité de la variable à caractériser. Il y a donc autant de groupes d'individus que de modalités dans la variable à caractériser. On parlera aussi de *classe* pour faire la distinction entre cette modalité à caractériser et les modalités des variables caractérisantes.

Double cliquez sur l'icône de la méthode DEMOD pour accéder au paramétrage.

L'ONGLET « **VARIABLES** »

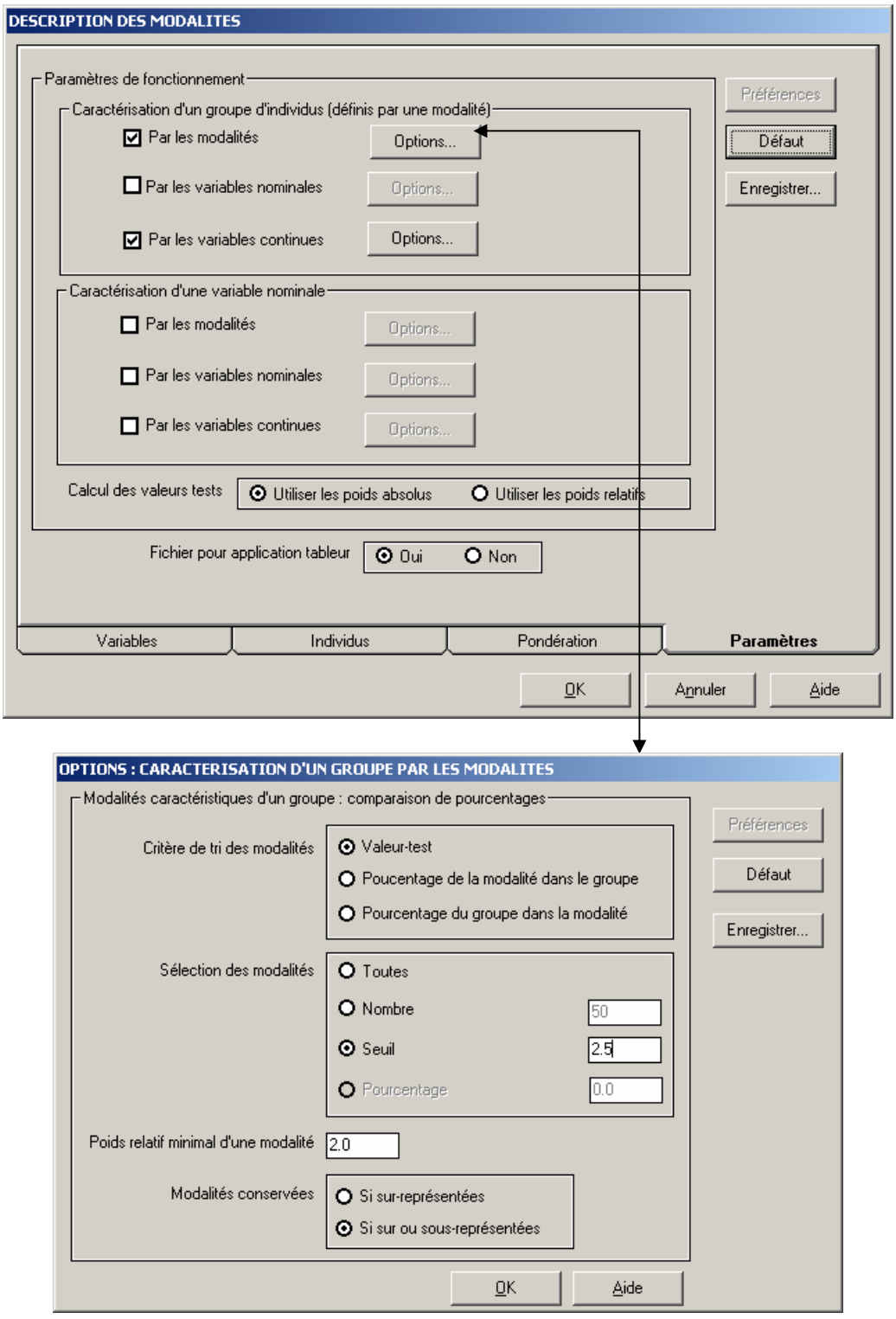
Le menu déroulant « **Sélection des variables** » permet de sélectionner la ou les variables nominales à caractériser ainsi que les variables nominales ou continues caractérisantes.



Dans cet exemple, la variable à caractériser est la variable V3 « Sexe de l'enquêté(e) ». Toutes les autres variables du fichier, nominales et continues, sont sélectionnées comme caractérisantes.

L'ONGLET « PARAMETRES »

Cet onglet spécifie les paramètres de fonctionnement et d'édition de la méthode DEMOD.



Lors que vous avez fini vos paramétrages, validez par « **OK** » et exécutez la méthode **DEMOD** en cliquant sur le bouton droit de la souris et en sélectionnant la commande « **Exécuter méthode** ».

LES RESULTATS DE **DEMOD**

Demod-5 : Caractérisation d'un groupe d'individus par les modalités

Caractérisation par les modalités des classes de la variable
Sexe de l'enquêté(e)
Classe: masculin (Effectif: 138 - Pourcentage: 43.81)

Libellés des variables	Modalités caractéristiques	% de la modalité dans la classe	% de la modalité dans	% de la classe dans la modalité	Valeur-Test	Proba	Poids
Exercez-vous en ce moment une activité professionnelle	oui, plein temps	63.77	45.40	61.54	5.71	0.000	143
Situation actuelle de la personne interrogée	actif	68.84	54.60	55.23	4.40	0.000	172
Avez-vous été au chômage ces douze derniers mois ?	non	63.77	49.52	56.41	4.37	0.000	156
Avez-vous des conflits travail - vie personnelle	non	42.03	30.79	59.79	3.69	0.000	97
Avez-vous souffert récemment d'un état dépressif	non	92.75	84.13	48.30	3.66	0.000	265
Avez-vous souffert récemment de nervosité	non	62.32	50.79	53.75	3.51	0.000	160
Niveau d'études de l'enquêté(e)	cep et cap	26.09	17.78	64.29	3.25	0.001	56
Avez-vous souffert récemment de maux de tête	non	73.19	63.49	50.50	3.06	0.001	200
Regardez-vous la télévision ?	assez souvent	31.88	24.13	57.89	2.70	0.003	76
Les gens comme moi se sentent souvent seuls	pas du tout d'accord	56.52	47.94	51.66	2.58	0.005	151
Etes-vous inquiet du risque accident centrale nucléaire	assez	22.46	16.83	58.49	2.20	0.014	53
Etes-vous inquiet du risque d'une maladie grave	assez	30.43	24.13	55.26	2.17	0.015	76
Avez-vous eu des enfants	non	27.54	21.90	55.07	1.99	0.023	69
Exercez-vous en ce moment une activité professionnelle	oui, temps partiel	5.07	9.21	24.14	-2.08	0.019	29
Avez-vous eu des enfants	oui	71.01	77.14	40.33	-2.15	0.016	243
Exercez-vous en ce moment une activité professionnelle	non	27.54	35.24	34.23	-2.42	0.008	111
Etes-vous satisfait de votre état de santé	pas du tout	0.00	3.17	0.00	-2.77	0.003	10
Avez-vous souffert récemment de maux de tête	oui	26.81	36.51	32.17	-3.06	0.001	115
Exercez-vous en ce moment une activité professionnelle	n'a jamais travaillé	3.62	10.16	15.63	-3.35	0.000	32
Avez-vous souffert récemment de nervosité	oui	37.68	49.21	33.55	-3.51	0.000	155
Avez-vous souffert récemment d'un état dépressif	oui	7.25	15.87	20.00	-3.66	0.000	50
Statut matrimonial	veuf(ve)	0.00	6.03	0.00	-4.24	0.000	19
Avez-vous des conflits travail - vie personnelle	*Reponse manquante*	31.16	45.08	30.28	-4.30	0.000	142
Avez-vous été au chômage ces douze derniers mois ?	*Reponse manquante*	31.16	45.08	30.28	-4.30	0.000	142
Situation actuelle de la personne interrogée	ménagère s.prof.	0.00	16.51	0.00	-7.87	0.000	52

% de la modalité dans la classe (MOD/CLA) :

Effectif de la modalité dans la classe divisé par l'effectif de la classe

% de la modalité dans l'échantillon (GLOBAL) :

Effectif de la modalité dans la population globale divisé par l'effectif de l'ensemble de la population

% de la classe dans la modalité (CLA/MOD) :

Effectif de la modalité dans la classe divisé par l'effectif de la modalité dans la population globale

Valeur-test :

Lorsque la valeur-test est positive, cela signifie que la modalité est sur-représentée dans la classe. La modalité est sous-représentée si la valeur-test est négative.

Probabilité :

La probabilité évalue l'importance de l'écart entre le pourcentage de la modalité caractérisante dans la classe (ie modalité à caractériser) et le pourcentage de cette modalité dans la population globale. Plus la probabilité est faible, plus l'écart est jugé significatif et plus la valeur-test associée à cette probabilité est forte (la valeur-test est le fractile de la loi normale correspondant à la même probabilité).

Poids :

Poids des individus dans la modalité caractérisante

Demod-13 : Caractérisation d'un groupe d'individus par les variables continues

Caractérisation par les variables continues des modalités de la variable
Sexe de l'enquêté(e)

masculin (Poids = 138.00 Effectif = 138)

Variables caractéristiques	Moyennes dans la modalité	Moyenne générale	Ecart-type dans la	Ecart-type général	Valeur-Test	Probabilité
Salaire mens. de l'enquêté	6533.190	4408.550	5486.120	4575.340	6.69	0.000
Famille, enfants : importance accordée	6.493	6.651	1.320	1.062	-2.33	0.010

féminin (Poids = 177.00 Effectif = 177)

Variables caractéristiques	Moyennes dans la modalité	Moyenne générale	Ecart-type dans la	Ecart-type général	Valeur-Test	Probabilité
Famille, enfants : importance accordée	6.774	6.651	0.785	1.062	2.33	0.010
Salaire mens. de l'enquêté	2751.330	4408.550	2742.020	4575.340	-6.69	0.000

Moyenne dans la modalité :

Moyenne pondérée du groupe pour la variable continue

Moyenne générale :

Moyenne pondérée dans l'échantillon de la variable continue

Interprétation :

On note que les variables « Salaire men. de l'enquêté » est la variable continue la plus caractéristique du groupe d'individus « masculin ».

Cette classe se caractérise par des personnes ayant un salaire plus important que la moyenne. Ainsi, les individus de cette classe ont 6533 F en salaire moyen contre 4408 dans l'échantillon.

Caractérisation d'une variable quantitative

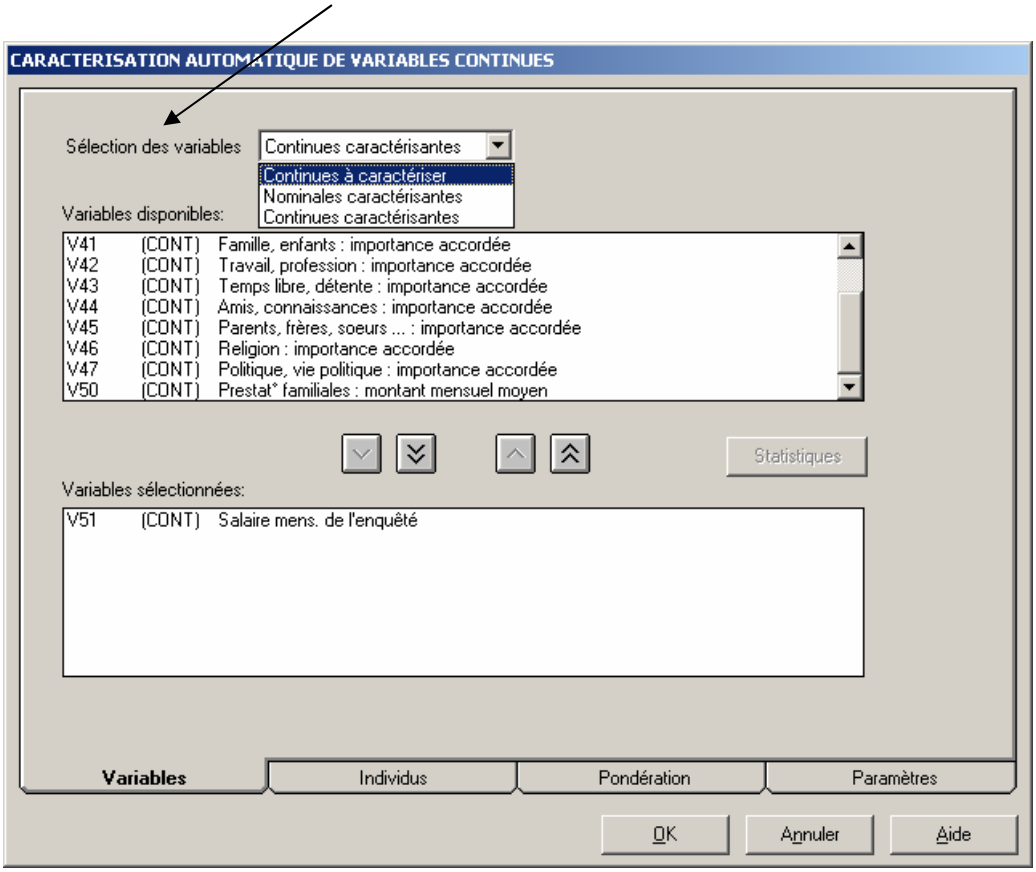
La procédure DESCO permet d'obtenir la caractérisation d'une ou plusieurs variables quantitatives (ou continues) en explorant l'ensemble des liaisons qu'elle entretient avec toutes les autres variables du fichier quelque soit leur type.
Pour les réponses manquantes relatives à une variable continue, les individus sont éliminés de l'analyse.

L'ONGLET « VARIABLES »

Une variable continue est caractérisable par les autres variables continues et par les variables nominales dites caractérisantes.

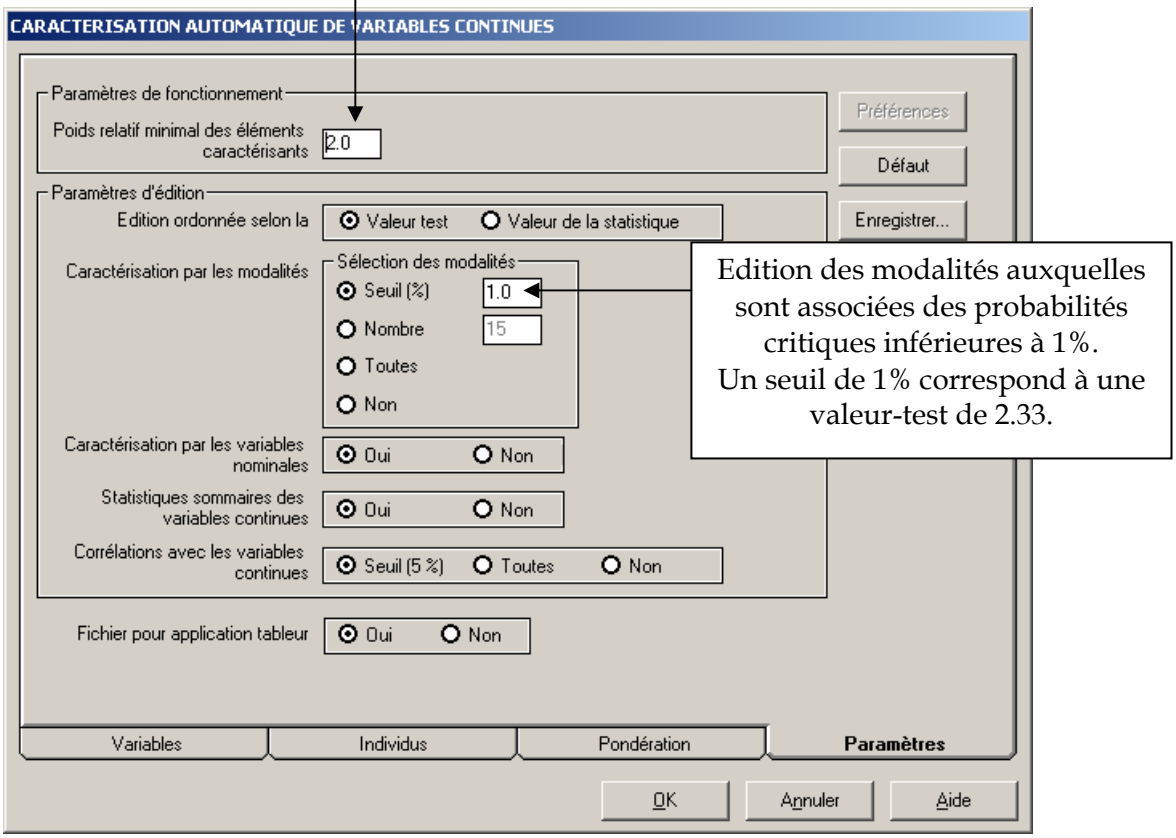
L'objet de cet onglet est de sélectionner la ou les variables à caractériser et les variables caractérisantes.

Le menu déroulant « **Sélection des variables** » donne accès au choix de variables continues à caractériser et au choix des variables nominales et continues caractérisantes.



L'ONGLET « **PARAMETRES** »

Le paramètre « **Poids relatif des éléments caractérisants** » permet de ne pas éditer les modalités et les variables continues dont le poids (en %) est inférieur à un seuil (2% par défaut).



LES RESULTATS DE DESCO

DESCRIPTION DE : Salaire mens. de l'enquêté
CARACTERISATION PAR LES MODALITES
DE LA VARIABLE : Salaire mens. de l'enquêté

SUR 267.0 INDIVIDUS ACTIFS MOYENNE = 4408.547
ECART-TYPE = 4575.339

V.TEST	PROB.	MOYENNE	E-TYPE	MODALITES	LIBELLE DE LA VARIABLE	POIDS
8.16	0.000	7060.53	4921.82	oui, plein temps	Exercez-vous en ce moment une activité professionnelle	114.00
7.58	0.000	6496.32	4736.16	actif	Situation actuelle de la personne interrogée	136.00
7.28	0.000	6617.07	4883.30	non	Avez-vous été au chômage ces douze derniers mois ?	123.00
6.69	0.000	6533.19	5486.12	masculin	Sexe de l'enquêté(e)	117.00
4.60	0.000	6452.63	5414.05	non	Avez-vous des conflits travail - vie personnelle	76.00
4.25	0.000	6698.25	6784.83	assez souvent	Regardez-vous la télévision ?	57.00
3.73	0.000	6331.15	3880.83	oui	Avez-vous des conflits travail - vie personnelle	61.00
3.47	0.000	6797.37	6049.03	études sup. g.écoles	Niveau d'études de l'enquêté(e)	38.00
3.35	0.000	4860.06	4834.30	non	Avez-vous souffert récemment d'un état dépressif	217.00
3.18	0.001	5291.85	5418.67	non	Avez-vous souffert récemment de nervosité	135.00
3.10	0.001	6950.00	5579.71	oui	Disposez-vous d'un piano	28.00
2.89	0.002	6529.41	5935.61	oui	Disposez-vous d'une résidence secondaire	34.00
2.88	0.002	6330.00	7536.22	oui	Disposez-vous d'un magnétoscope	40.00
2.65	0.004	5937.26	6786.27	région paris	Région où habite l'enquêté(e)	51.00
2.43	0.008	5179.34	5246.40	très	L'enquêté(e) a-t-il été intéressé par l'enquête	117.00
-2.47	0.007	3319.48	2735.76	beaucoup	Etes-vous inquiet du risque accident centrale nucléaire	77.00
-2.54	0.006	1971.43	1864.75	chômeur	Situation actuelle de la personne interrogée	21.00
-2.57	0.005	760.00	1356.61	étudiant	Situation actuelle de la personne interrogée	10.00
-2.66	0.004	2606.41	3255.77	beaucoup moins bien	Opinion sur l'évolution du niveau de vie personnel	39.00
-2.86	0.002	3726.34	3277.03	tous les jours	Regardez-vous la télévision ?	155.00
-2.88	0.002	4069.97	3721.48	non	Disposez-vous d'un magnétoscope	227.00
-2.89	0.002	4099.07	4253.85	non	Disposez-vous d'une résidence secondaire	233.00
-3.10	0.001	4110.81	4346.66	non	Disposez-vous d'un piano	239.00
-3.18	0.001	3505.18	3271.07	oui	Avez-vous souffert récemment de nervosité	132.00
-3.35	0.000	2449.00	2373.53	oui	Avez-vous souffert récemment d'un état dépressif	50.00
-3.49	0.000	2263.04	2043.80	aucun diplôme	Niveau d'études de l'enquêté(e)	46.00
-4.36	0.000	832.14	1563.89	n'a jamais travaillé	Exercez-vous en ce moment une activité professionnelle	28.00
-4.85	0.000	2691.10	3397.40	non	Exercez-vous en ce moment une activité professionnelle	103.00
-6.54	0.000	488.54	1396.02	ménagère s.prof.	Situation actuelle de la personne interrogée	48.00
-6.69	0.000	2751.33	2742.02	féminin	Sexe de l'enquêté(e)	150.00
-7.28	0.000	2311.41	3196.29	*Reponse manquante*	Avez-vous été au chômage ces douze derniers mois ?	130.00
-7.28	0.000	2311.41	3196.29	*Reponse manquante*	Avez-vous des conflits travail - vie personnelle	130.00
		4408.55	4575.34	ENSEMBLE		267.00

CARACTERISATION PAR LES VARIABLES NOMINALES
DE LA VARIABLE : Salaire mens. de l'enquêté

V.TEST	PROBA.	NUM . LIBELLE DE LA VARIABLE	DEG.LIB.DEN	FISHER
8.56	0.000	5 . Situation actuelle de la personne interrogée	261	21.44
8.48	0.000	18 . Exercez-vous en ce moment une activité professionnelle	263	31.95
7.50	0.000	20 . Avez-vous été au chômage ces douze derniers mois ?	264	35.01
7.28	0.000	19 . Avez-vous des conflits travail - vie personnelle	264	32.89
6.98	0.000	3 . Sexe de l'enquêté(e)	265	53.58
3.48	0.000	7 . Niveau d'études de l'enquêté(e)	258	3.87
3.47	0.000	33 . Regardez-vous la télévision ?	263	6.57
3.38	0.001	24 . Avez-vous souffert récemment d'un état dépressif	265	11.69
3.21	0.001	23 . Avez-vous souffert récemment de nervosité	265	10.50
3.12	0.002	16 . Disposez-vous d'un piano	265	9.94
2.90	0.004	17 . Disposez-vous d'une résidence secondaire	265	8.58
2.89	0.004	15 . Disposez-vous d'un magnétoscope	265	8.50
2.04	0.021	52 . L'enquêté(e) a-t-il été intéressé par l'enquête	264	3.92
1.92	0.054	21 . Avez-vous souffert récemment de maux de tête	265	3.74
1.77	0.039	30 . Opinion sur l'évolution du niveau de vie personnel	261	2.38
1.56	0.059	25 . Etes-vous satisfait de votre état de santé	263	2.51
1.33	0.092	40 . Etes-vous inquiet du risque accident centrale nucléaire	263	2.16
1.31	0.189	29 . Vous imposez-vous régulièrement des restrictions	265	1.73
1.24	0.107	8 . La famille est le seul endroit où l'on se sent bien	264	2.24
1.12	0.132	1 . Région où habite l'enquêté(e)	259	1.61
1.07	0.143	39 . Etes-vous inquiet du risque du chômage	263	1.82

1.03	0.151	35 . La diffusion de l'informatique est une chose ...	263	1.78
1.02	0.154	34 . Estimez-vous que la société a besoin de se transformer	264	1.86
0.92	0.179	49 . Les gens comme moi se sentent souvent seuls	263	1.64
0.89	0.186	31 . Opinion sur l'évolution du niveau de vie des français	260	1.48
0.86	0.194	36 . Etes-vous inquiet du risque d'une maladie grave	263	1.58
0.79	0.428	22 . Avez-vous souffert récemment de mal au dos	265	0.63
0.78	0.217	11 . Etes-vous satisfait de votre logement proprement dit	263	1.49
0.65	0.257	37 . Etes-vous inquiet du risque d'une agression dans la rue	263	1.35
0.45	0.327	13 . Statut d'occupation du logement	262	1.16
0.22	0.412	27 . Avez-vous eu des enfants	264	0.88
0.13	0.446	38 . Etes-vous inquiet du risque d'un accident de la route	263	0.89
0.10	0.459	6 . Statut matrimonial	262	0.91
0.08	0.469	9 . Opinion sur le mariage	263	0.85
-0.15	0.561	32 . Opinion sur les conditions de vie à venir	261	0.79
-0.21	0.585	12 . Etes-vous satisfait de votre cadre de vie quotidien	263	0.65
-0.23	0.591	14 . Les dépenses de logement sont pour vous une charge ...	260	0.77
-0.53	0.702	10 . Les travaux du ménage,les soins des enfants incombent .	263	0.47
-0.59	0.724	2 . Taille d'agglomération (nombre d'habitants)	258	0.66
-0.64	0.740	48 . Opinion sur le fonctionnement de la justice en 1986	261	0.55

STATISTIQUES SOMMAIRES DES VARIABLES CONTINUES

EFFECTIF TOTAL 315
POIDS TOTAL 315.00

NUM . IDEN - LIBELLE	EFFECTIF	POIDS	MOYENNE	ECART-TYPE	MINIMUM	MAXIMUM
4 . Age - Age de l'enquêté(e)	267	267.00	43.61	16.88	18.00	83.00
26 . Nbpr - Nb person. dans logm	267	267.00	3.04	1.43	1.00	8.00
28 . Nbef - Nombre d'enfants eus	267	267.00	1.85	1.69	0.00	9.00
41 . Fami - Famille, enfants : i	267	267.00	6.65	1.07	1.00	7.00
42 . Trav - Travail, profession	267	267.00	5.90	1.57	1.00	7.00
43 . Lois - Temps libre, détente	267	267.00	5.30	1.43	0.00	7.00
44 . Amis - Amis, connaissances	267	267.00	5.18	1.41	1.00	7.00
45 . Part - Parents, frères, soe	267	267.00	5.63	1.44	1.00	7.00
46 . Reli - Religion : importanc	267	267.00	3.15	1.96	1.00	7.00
47 . Poli - Politique, vie polit	267	267.00	3.15	1.79	1.00	7.00
50 . PrFm - Prestat° familiales	244	244.00	583.10	966.04	0.00	5100.00
51 . Salr - Salaire mens. de l'e	267	267.00	4408.55	4575.34	0.00	40000.00

CORRELATION AVEC LES VARIABLES CONTINUES
DE LA VARIABLE : Salaire mens. de l'enquêté

V.TEST	PROBA.	CORR.	NUM . LIBELLE DE LA VARIABLE	POIDS
-2.53	0.006	-0.162	50 . Prestat° familiales : montant mensuel moyen	244.000

TABLE - Tableaux croisés

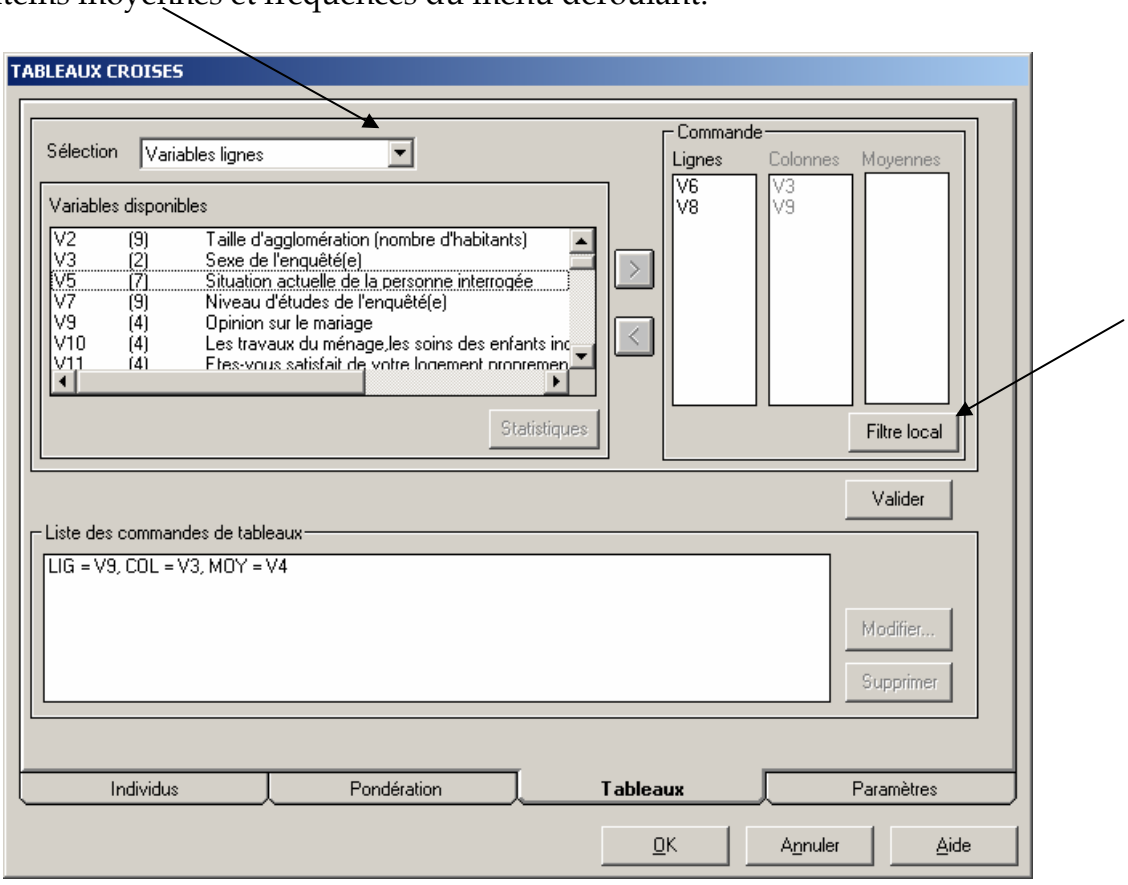
La procédure TABLE est conçue pour le calcul et l'édition massive de tableaux croisés.
On obtient à partir de cette procédure des tableaux de contingence, des tableaux de moyennes ou encore des tableaux de fréquences.

L'ONGLET « TABLEAUX »

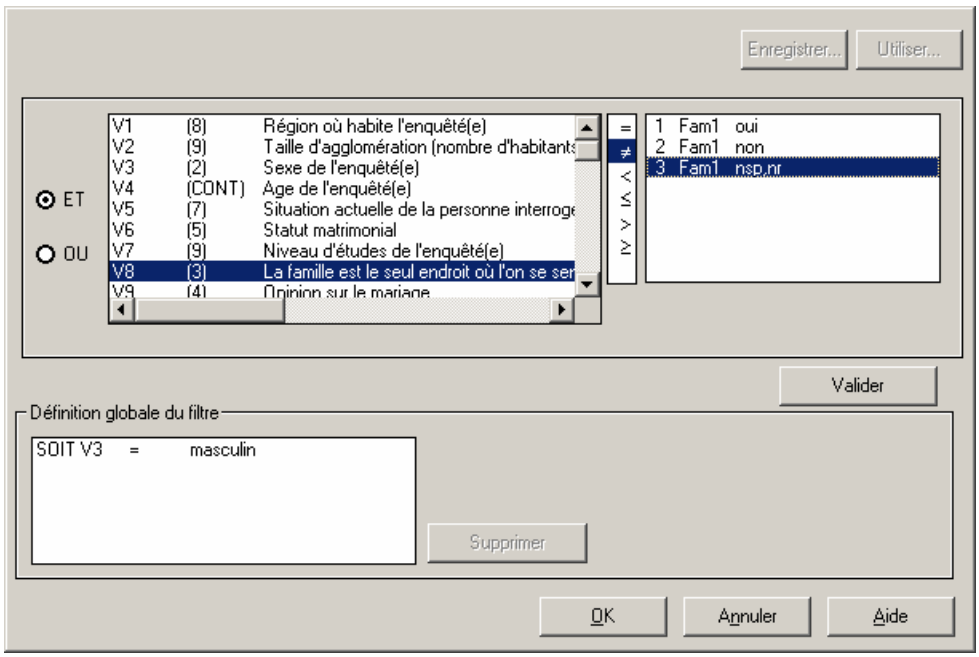
L'objet de cet onglet est de choisir l'ensemble des tableaux croisés que l'on souhaite éditer.

Les cases de chaque tableau contiendront soit les effectifs des individus si ceux-ci ont des poids uniformes, soit les poids des individus si ceux-ci sont pondérés.

Le menu « **Sélection** » offre également la possibilité de choisir des variables continues avec les items moyennes et fréquences du menu déroulant.

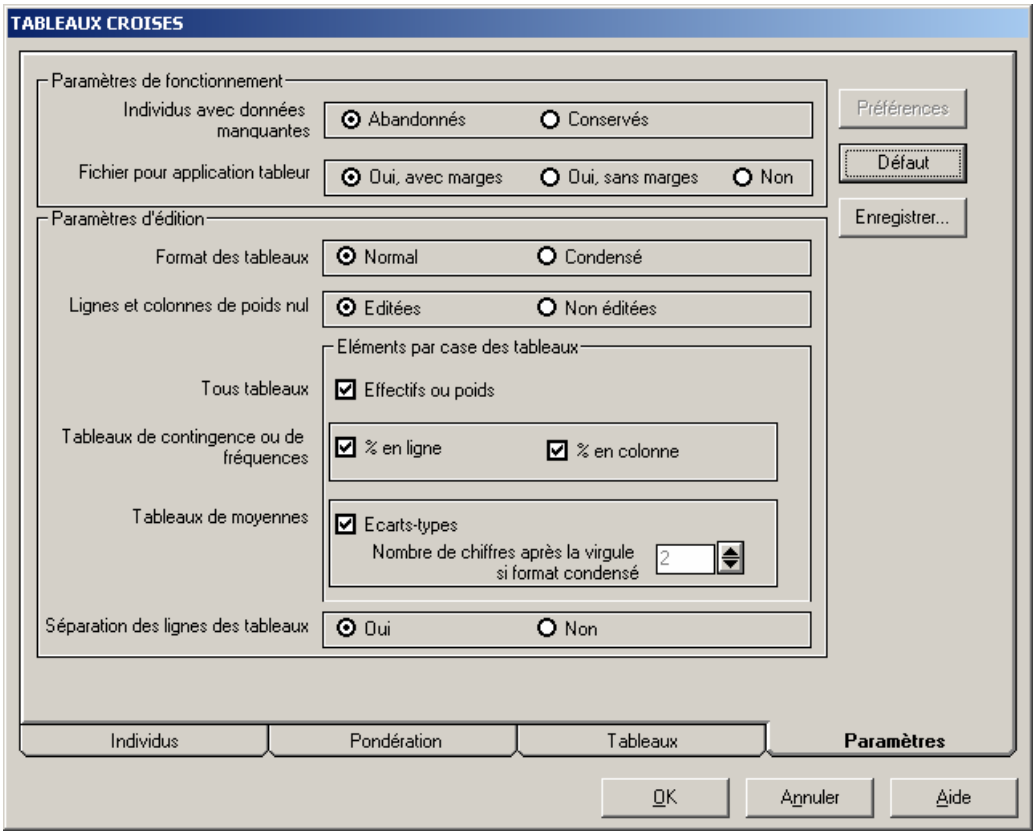


Si une variable est présente dans la liste des variables « **moyennes** », chaque case du tableau contiendra la moyenne pondérée de cette variable pour les individus de la case.
Si une variable est choisie dans la liste des variables « **fréquences** », chaque case du tableau contiendra la somme pondérée de la fréquence de cette variable pour les individus de la case.
En cliquant sur le bouton « **Filtre local** », il est possible de choisir un filtre sur les individus pour la commande de tableaux en cours.



Ce filtre permet de ne calculer les tableaux que pour des individus satisfaisant à ce filtre.

L'ONGLET « **PARAMETRES** »



LES RESULTATS DE TABLE

EDITION DES TABLEAUX

TABLEAU 1 EN LIGNE : Statut matrimonial *POIDS TOTAL : 315.*

EN COLONNE : Sexe de l'enquêté(e)

POIDS / % COLONNE / % LIGNE	masculin	féminin	ENSEMBLE
célibataire	23 16.67 54.76	19 10.73 45.24	42 13.33 100.00
marié(e)	105 76.09 47.09	118 66.67 52.91	223 70.79 100.00
concubinage	7 5.07 43.75	9 5.08 56.25	16 5.08 100.00
séparé(e) divorcé	3 2.17 20.00	12 6.78 80.00	15 4.76 100.00
veuf(ve)	0 0.00 0.00	19 10.73 100.00	19 6.03 100.00
ENSEMBLE	138 100.00 43.81	177 100.00 56.19	315 100.00 100.00

KHI2 = 21.29 / 4 DEGRES DE LIBERTE / 0 EFFECTIFS THEORIQUES INFERIEURS A 5
PROBA (KHI2 > 21.29) = 0.000 / V.TEST = 3.45

TABLEAU 2 EN LIGNE : Opinion sur le mariage *POIDS TOTAL : 315.*

EN COLONNE : Sexe de l'enquêté(e)

MOYENNES DE : Age de l'enquêté(e)

POIDS / MOYENNE / ECART-TYPE	masculin	féminin	ENSEMBLE
indissoluble	41 45.829 17.234	40 48.325 17.084	81 47.062 17.206
dissout si pb. grave	39 43.000 14.739	69 46.362 18.260	108 45.148 17.148
dissout si accord	50 41.300 15.442	64 38.484 14.330	114 39.719 14.893
ne sait pas	8 50.250 15.618	4 41.250 8.842	12 47.250 14.377
ENSEMBLE	138 43.645 16.007	177 43.842 17.015	315 43.756 16.581

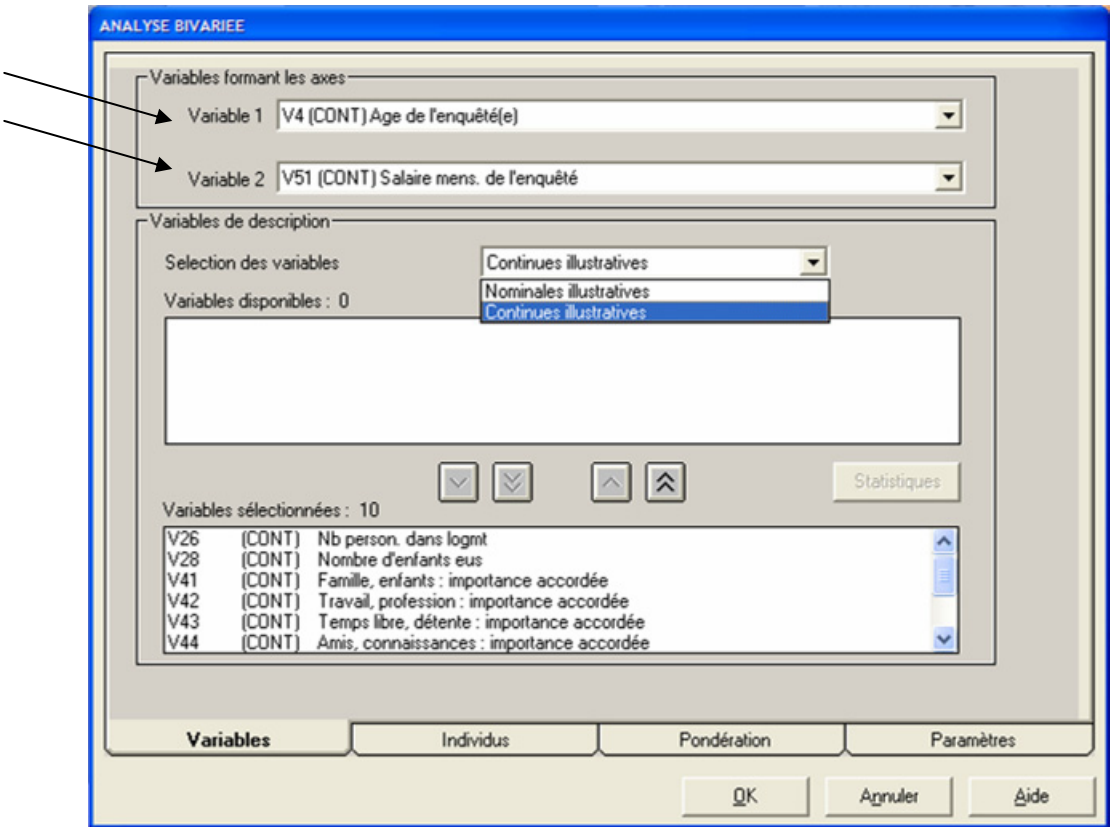
BIVAR - Analyse bivariée

L'analyse bivariée est une procédure essentiellement graphique permettant de visualiser les liaisons que deux variables continues prises ensemble entretiennent avec les autres variables.

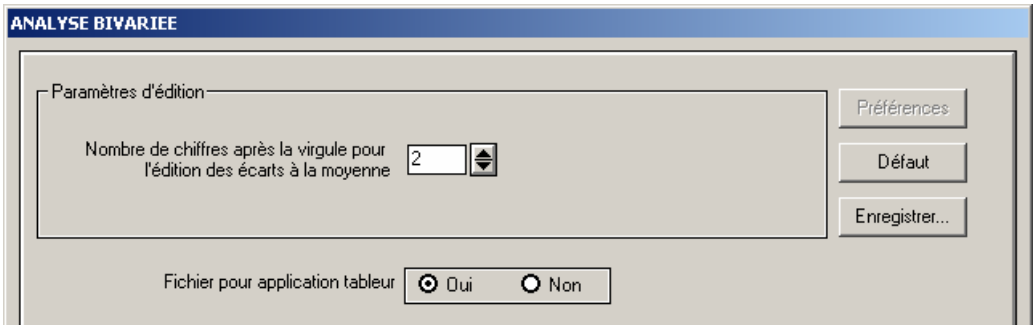
L'ONGLET « VARIABLES »

Cet onglet permet de sélectionner les deux variables dans l'encadrement « **Variables formant les axes** ». Il ne peut y avoir que deux variables continues actives.

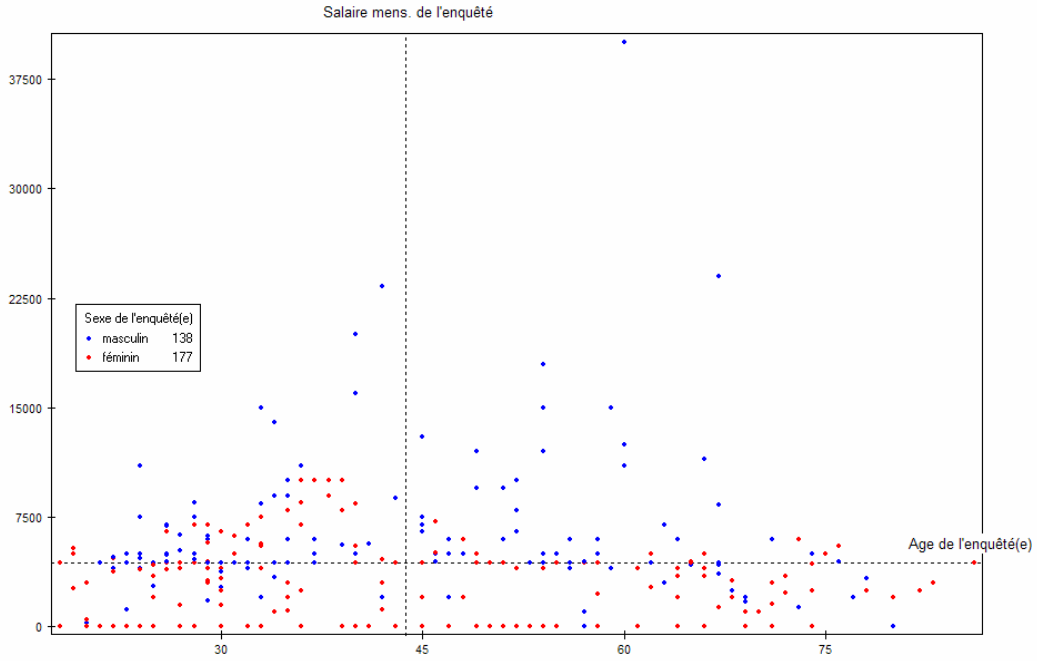
L'échantillon peut être aussi décrit par des variables nominales et continues. Ces variables sont sélectionnées à l'aide du menu déroulant dans l'encadrement « **Variables de description** ».

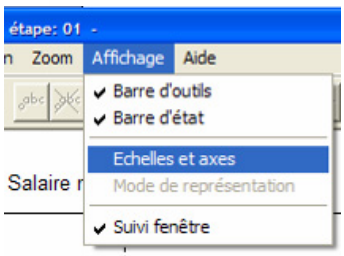


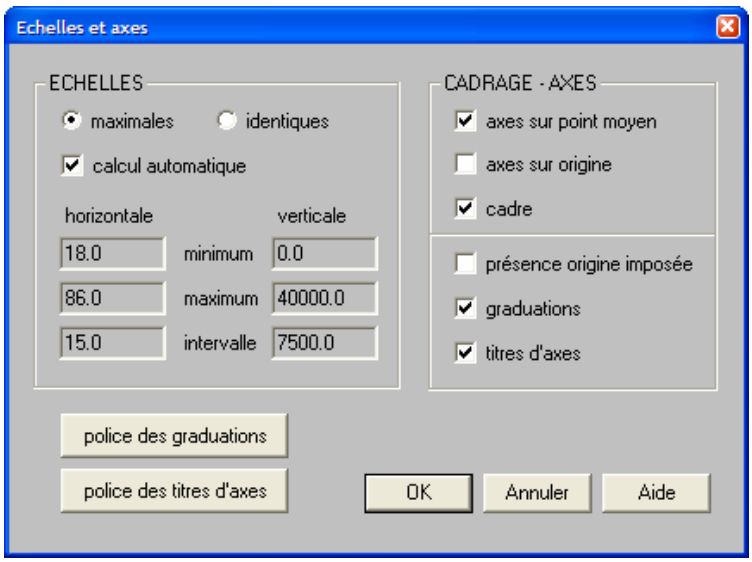
L'ONGLET « PARAMETRES »



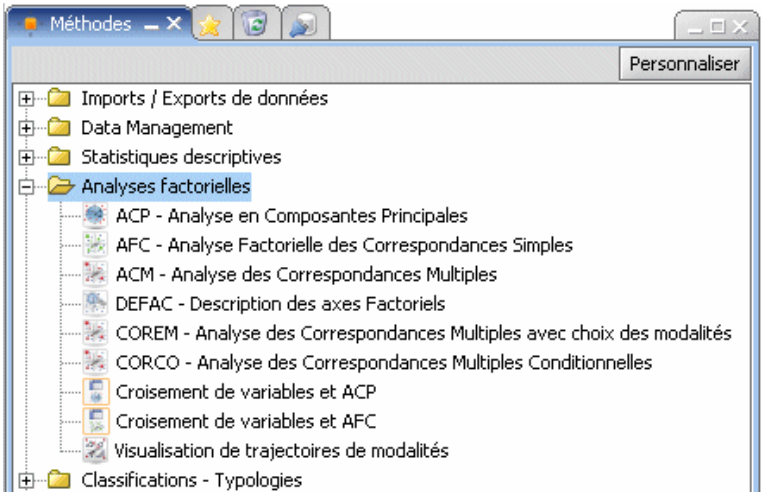
L'outil graphique de la méthode BIVAR est l'éditeur de graphiques factoriels.
L'outil est présenté dans la partie « Les analyses factorielles » - « Editeur Graphiques Factoriels ».







LES ANALYSES FACTORIELLES AVEC SPAD



Analyse en Composantes Principales(ACP)

Analyse Factorielle des Correspondances simples (AFC)

Analyse des Correspondances Multiples(ACM)

Description des axes factoriels

Les méthodes factorielles établissent des représentations synthétiques de vastes tableaux de données, en général sous forme de représentations graphiques. Ces méthodes ont pour objet de réduire les dimensions des tableaux de données de façon à représenter les associations entre individus et entre variables dans des espaces de faibles dimensions.

Les méthodes diffèrent selon la nature des variables analysées : il peut s'agir de variables continues, de variables nominales ou de catégories dans le cas des tableaux de contingences. Les lignes peuvent être des individus ou des catégories.

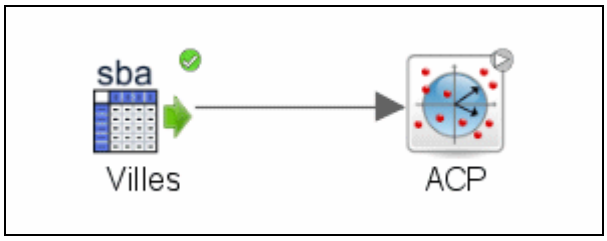
VOCABULAIRE

Variables actives	Variables qui participent à la construction des axes de la représentation factorielle
Variables illustratives	Toutes les variables qui n'ont pas participé à la construction des axes mais permettent d'illustrer les représentations factorielles
Contribution	Mesure la participation d'un élément (modalité, variable, fréquence ou individu) à la construction d'un axe factoriel
Cosinus	Mesure la qualité de la représentation d'un élément (modalité, variable, fréquence ou individu) sur un axe factoriel

ACP - Analyse en Composantes Principales

Cette procédure effectue l'analyse en composantes principales d'un ensemble d'individus caractérisés par des variables continues. L'analyse peut être normée (analyse de la matrice de corrélations entre les variables) ou non normée (analyse de la matrice des covariances). Elle permet l'introduction de variables continues et nominales en éléments illustratifs.

- Importer la base Sba Villes.sba de la connexion Bases Sba.
- Glissez-déposez la méthode Analyse en composantes principales sur la base importée.



Le problème est de comparer un certain nombre de grandes villes selon le niveau moyen des salaires dans une douzaine de professions afin de vérifier la cohérence de la description par rapport à nos connaissances économiques globales.

On s'intéresse à 51 villes.
Les données recueillies ne font pas seulement référence aux salaires mais elles constituent un ensemble plus vaste de 40 variables concernant aussi les prix et quelques autres indicateurs essentiellement économiques.
Les villes sont réparties dans 10 régions du monde (variable 2) et les observations sont connues à deux dates (1991 et 1994 : variable 1) bien que l'on s'intéressera uniquement à l'année 1994.

KIDEN	Année	Region du monde	Instituteur	Chauffeur d'autobus	Mécanicien autos	Manoeuvre du bâtiment
AbuDhabi94	Edition 1994	PROCHE ORIENT	19 500	11 400	9 200	3 500
Amsterdam94	Edition 1994	CENTRAL EUROPE	23 800	24 900	14 300	13 000
Athenes94	Edition 1994	SUD EUROPE	10 100	11 300	6 000	9 700
Bangkok94	Edition 1994	SUD ASIE ET AUSTRALI	4 100	3 400	2 600	1 700
Bogota94	Edition 1994	SUD AMERIQUE	4 100	4 100	6 500	1 700
Bombay94	Edition 1994	SUD ASIE ET AUSTRALI	1 600	1 700	1 300	800
Bruxelles94	Edition 1994	CENTRAL EUROPE	16 000	14 900	12 200	13 200
Budapest94	Edition 1994	EUROPE ORIENTAL	2 100	3 000	2 200	1 900
BuenosAires94	Edition 1994	SUD AMERIQUE	4 500	4 500	8 200	4 500
Caracas94	Edition 1994	SUD AMERIQUE	2 500	900	1 300	900
Chicago94	Edition 1994	NORD AMERIQUE	36 900	31 600	29 500	28 000
Copenhague94	Edition 1994	NORD EUROPE	20 500	17 100	23 600	19 000
Dublin94	Edition 1994	NORD EUROPE	19 800	13 600	10 300	10 100
Dusseldorf94	Edition 1994	CENTRAL EUROPE	32 900	24 000	16 600	15 800
Frankfurt94	Edition 1994	CENTRAL EUROPE	34 600	20 400	17 900	15 300
Geneve94	Edition 1994	CENTRAL EUROPE	49 200	37 200	29 800	21 300
Helsinki94	Edition 1994	NORD EUROPE	17 100	13 700	12 600	12 800
Hongkong94	Edition 1994	EST ASIE	16 500	12 200	1 400	1 300
Houston94	Edition 1994	NORD AMERIQUE	23 800	26 900	27 700	19 500
Jakarta94	Edition 1994	SUD ASIE ET AUSTRALI	1 200	2 200	4 200	1 800
Johannesburg94	Edition 1994	AFRIQUE	9 400	5 800	10 200	3 300
Lagos94	Edition 1994	AFRIQUE	600	900	1 400	700
Lisboa94	Edition 1994	SUD EUROPE	11 300	6 300	7 300	6 400
London94	Edition 1994	NORD EUROPE	19 900	12 700	14 500	13 400
LosAngeles94	Edition 1994	NORD AMERIQUE	27 900	27 600	25 000	25 700
Luxembourg94	Edition 1994	CENTRAL EUROPE	50 800	41 600	16 400	15 500
Madrid94	Edition 1994	SUD EUROPE	16 600	14 600	12 900	7 900
Manama94	Edition 1994	PROCHE ORIENT	14 800	7 600	5 300	2 600
Manila94	Edition 1994	SUD ASIE ET AUSTRALI	3 300	1 800	1 800	1 300
Mexico94	Edition 1994	NORD AMERIQUE	5 000	6 800	8 000	1 600
Milan94	Edition 1994	SUD EUROPE	13 000	16 700	13 100	9 800
Montreal94	Edition 1994	NORD AMERIQUE	21 200	21 500	21 100	20 400
Nairobi94	Edition 1994	AFRIQUE	600	400	700	200
NewYork94	Edition 1994	NORD AMERIQUE	29 000	26 200	24 500	25 700
Nicosia94	Edition 1994	SUD EUROPE	16 200	8 700	8 100	7 400
Oslo94	Edition 1994	NORD EUROPE	19 100	16 800	16 700	16 800
Panama94	Edition 1994	SUD AMERIQUE	5 300	8 200	5 300	4 200
Paris94	Edition 1994	CENTRAL EUROPE	16 200	19 400	13 100	9 300
Prague94	Edition 1994	EUROPE ORIENTAL	1 800	2 800	2 900	1 700
RiodeJaneiro94	Edition 1994	SUD AMERIQUE	2 000	4 700	3 000	1 400
SaoPaulo94	Edition 1994	SUD AMERIQUE	3 600	4 700	5 400	1 900
Seoul94	Edition 1994	EST ASIE	13 200	12 800	12 300	9 700
Sidney94	Edition 1994	SUD ASIE ET AUSTRALI	18 800	11 600	11 900	12 900
Singapore94	Edition 1994	EST ASIE	9 800	8 600	8 400	4 600
Stockholm94	Edition 1994	NORD EUROPE	17 400	15 300	14 800	18 200
Taipei94	Edition 1994	EST ASIE	23 700	19 100	23 700	13 800
Tel-Aviv94	Edition 1994	PROCHE ORIENT	9 100	10 300	11 500	6 800
Tokyo94	Edition 1994	EST ASIE	35 700	32 400	26 500	25 900
Toronto94	Edition 1994	NORD AMERIQUE	25 500	19 700	20 100	20 300
Vienna94	Edition 1994	CENTRAL EUROPE	18 500	19 200	17 800	16 200
Zurich94	Edition 1994	CENTRAL EUROPE	56 800	46 100	30 500	26 100

LES VARIABLES

1 .	Edition (1991 / 1994)	(2 modalités)
2 .	Région du monde	(10 modalités)
3 .	I_prix sans loyer	(CONTINUE)
4 .	I_prix avec loyer	(CONTINUE)
5 .	I_salaires bruts	(CONTINUE)
6 .	I_salaires nets	(CONTINUE)
7 .	Heures travail annuelles	(CONTINUE)
8 .	Vacances annuelles payées	(CONTINUE)
9 .	Pouvoir d'achat brut	(CONTINUE)
10 .	Pouvoir d'achat net	(CONTINUE)
11 .	Kg pain = temps de travail	(CONTINUE)
12 .	Hamb = temps de travail	(CONTINUE)
13 .	Denrées alimentaires	(CONTINUE)
14 .	Panier complet	(CONTINUE)
15 .	Dames vêtements	(CONTINUE)
16 .	Hommes vêtements	(CONTINUE)
17 .	4 pièces appart meublé	(CONTINUE)
18 .	3 pièces appart non meublé	(CONTINUE)
19 .	Loyer normale	(CONTINUE)
20 .	Appareils ménagers	(CONTINUE)
21 .	Bus tram ou métro	(CONTINUE)
22 .	Taxi	(CONTINUE)
23 .	Voitures	(CONTINUE)
24 .	Restaurant	(CONTINUE)
25 .	Nuit d'hôtel	(CONTINUE)
26 .	Services diverses	(CONTINUE)
27 .	Impôts et cotisations sociales en % salaire brut	(CONTINUE)
28 .	Salaire horaire net	(CONTINUE)
29 .	**Instituteur**	**(CONTINUE)**
30 .	**Chauffeur d'autobus**	**(CONTINUE)**
31 .	**Mécanicien autos**	**(CONTINUE)**
32 .	**Manœuvre du bâtiment**	**(CONTINUE)**
33 .	**Tourneur**	**(CONTINUE)**
34 .	**Cuisinier chef**	**(CONTINUE)**
35 .	**Chef de service**	**(CONTINUE)**
36 .	**Ingénieur**	**(CONTINUE)**
37 .	**Caissier de banque**	**(CONTINUE)**
38 .	**Secrétaire direction**	**(CONTINUE)**
39 .	**Vendeuse**	**(CONTINUE)**
40 .	**Ouvrière du textile**	**(CONTINUE)**
41 .	Revenu annuel net moyen	(CONTINUE)

LES PARAMETRES DE LA METHODE **ACP**

L'ONGLET « **VARIABLES** »

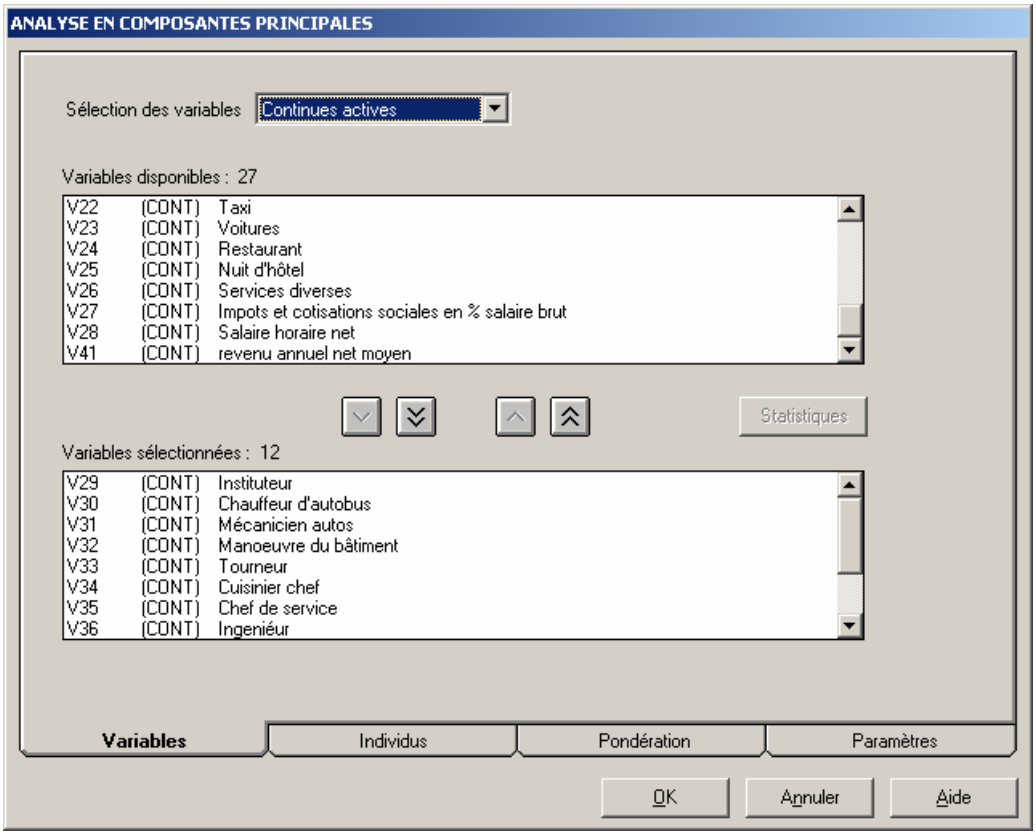
Pour comparer les villes entre elles, il est certes possible de prendre en compte toutes les variables disponibles. Cela conduira à comparer les villes en tenant compte simultanément du niveau des prix, des salaires, des impôts, … Les différences observées entre les villes sont alors difficiles à interpréter car elles peuvent avoir des causes multiples et de nature très différentes.

Il est plus « sage » de sélectionner un groupe de variables, ce groupe étant homogène par rapport à un thème défini et cohérent avec l'objectif propre de l'étude.

Les variables choisies, appelées variables actives, constituent donc les seuls éléments utilisés pour comparer les villes entre elles. Cela signifie pas que le reste de l'information soit abandonné : il servira ensuite à illustrer ou suggérer des explications pour les similitudes et différences observées entre les villes.

Dans notre exemple, nous décidons de prendre comme variables actives l'ensemble des revenus nets perçus dans les 12 professions retenues.

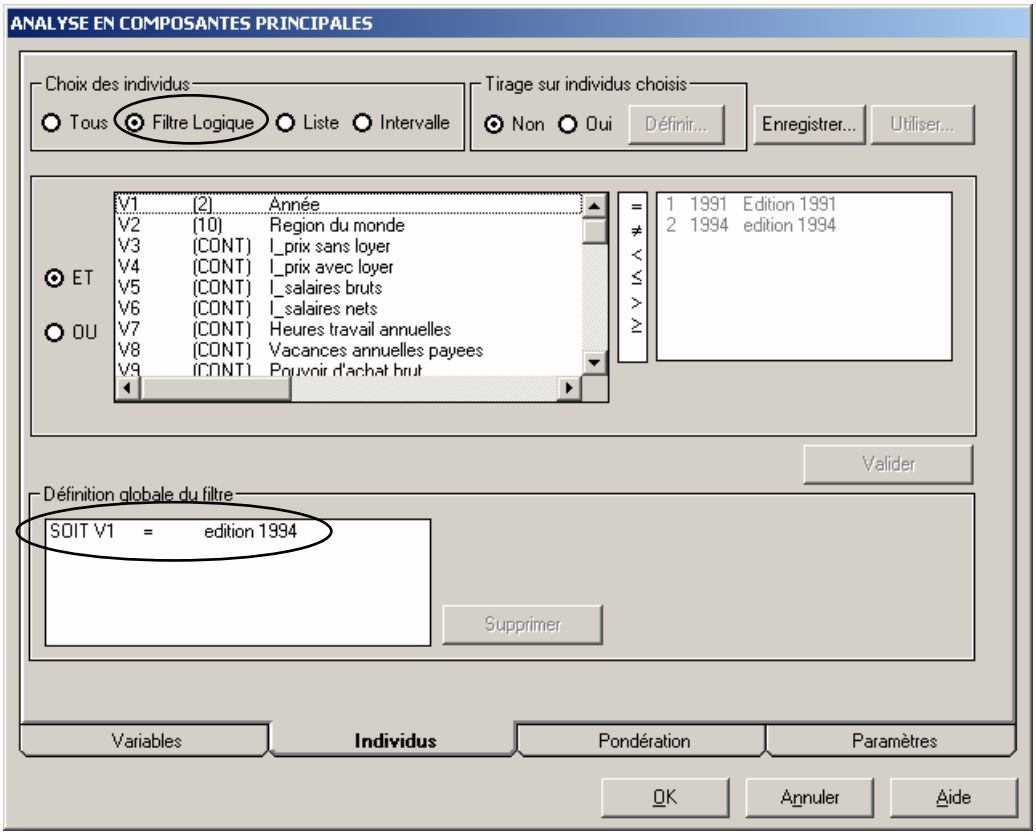
Deux villes seront proches si les rémunérations sont analogues dans l'ensemble de ces 12 professions, indépendamment de ce qui peut les différencier par ailleurs (taille, densité…).



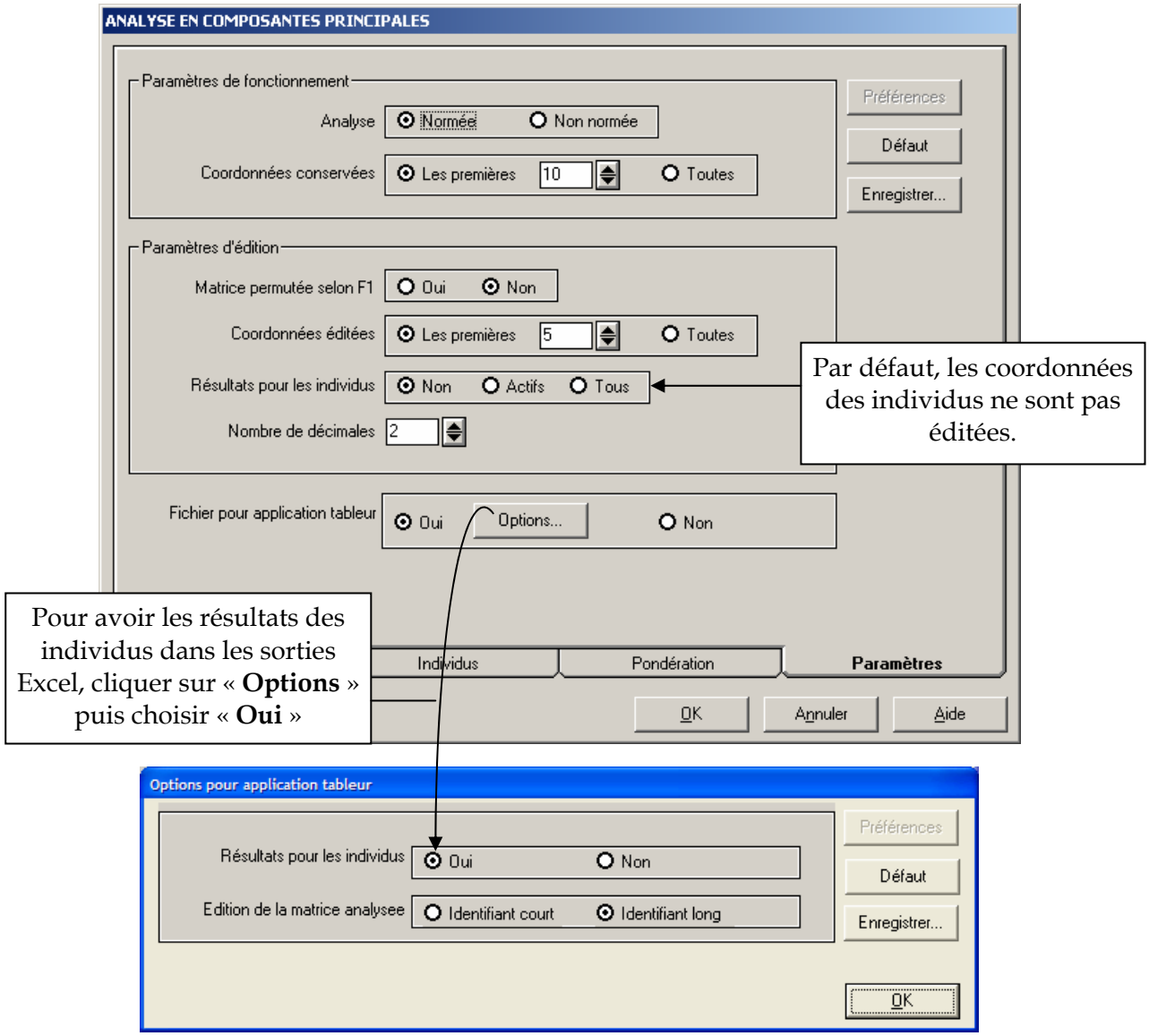
L'ONGLET « **INDIVIDUS** »

On s'intéresse uniquement à l'année 1994.
Pour cela, on définit un filtre logique :

« **Année = Edition 1994** »



L'ONGLET « **PARAMETRES** »



ACP NORMEE ET ACP NON NORMEE

Dans le cas d'une ACP normée (centrage, réduction), toutes les variables sont situées à une même distance de l'origine et participent de manière égale à l'inertie totale du nuage.
Dans le cas d'une ACP non normée (centrage), la distance de la variable à l'origine est égale à la variance de la variable.

Dans la pratique, il est souvent justifié de donner aux variables une importance égale en utilisant l'ACP normée. C'est en particulier indispensable quand les variables actives sont mesurées avec des unités différentes.

Dans notre exemple, toutes les variables actives sont mesurées avec la même unité. Nous pouvons penser réaliser une ACP non normée. Cependant, réaliser cette analyse aura tendance à donner plus d'importance aux professions à salaires élevés, car ce sont celles qui ont les plus grands écarts-types.

LES RESULTATS DE LA METHODE ACP

ANALYSE EN COMPOSANTES PRINCIPALES

STATISTIQUES SOMMAIRES DES VARIABLES CONTINUES

EFFECTIF TOTAL : 51 POIDS TOTAL : 51.00

NUM . IDEN - LIBELLE	EFFECTIF	POIDS	MOYENNE	ECART-TYPE	MINIMUM	MAXIMUM
29 . INST - Instituteur	51	51.00	16801.96	13243.42	600.00	56800.00
30 . CHAU - Chauffeur d'autobus	51	51.00	14311.76	10819.58	400.00	46100.00
31 . MECA - Mécanicien autos	51	51.00	12384.31	8520.83	700.00	30500.00
32 . MANO - Manoeuvre du bâtimen	51	51.00	10343.14	8239.82	200.00	28000.00
33 . OUTI - Tourneur	51	51.00	15145.10	10244.30	800.00	38700.00
34 . CUIS - Cuisinier chef	51	51.00	15615.68	8768.42	500.00	33900.00
35 . CHEF - Chef de service	51	51.00	30933.34	21250.57	1500.00	95000.00
36 . INGE - Ingénieur	51	51.00	24664.71	14019.08	1600.00	59700.00
37 . CAIS - Caissier de banque	51	51.00	18749.02	13413.83	1200.00	58800.00
38 . SECR - Secrétaire direction	51	51.00	13311.76	7569.80	1400.00	31500.00
39 . VEND - Vendeuse	51	51.00	9658.82	6064.53	400.00	24700.00
40 . OUVR - Ouvrière du textile	51	51.00	9247.06	6429.78	300.00	23800.00

MATRICE DES CORRELATIONS

	INST	CHAU	MECA	MANO	OUTI	CUIS	CHEF	INGE	CAIS	SECR	VEND	OUVR
INST	1.00											
CHAU	0.96	1.00										
MECA	0.84	0.89	1.00									
MANO	0.83	0.88	0.95	1.00								
OUTI	0.91	0.94	0.93	0.93	1.00							
CUIS	0.75	0.76	0.80	0.72	0.76	1.00						
CHEF	0.78	0.74	0.64	0.59	0.69	0.82	1.00					
INGE	0.81	0.82	0.74	0.70	0.80	0.82	0.87	1.00				
CAIS	0.82	0.80	0.70	0.64	0.72	0.79	0.89	0.85	1.00			
SECR	0.92	0.93	0.88	0.86	0.92	0.80	0.80	0.87	0.87	1.00		
VEND	0.88	0.89	0.89	0.86	0.88	0.85	0.79	0.85	0.85	0.94	1.00	
OUVR	0.88	0.92	0.89	0.92	0.94	0.71	0.65	0.81	0.73	0.93	0.89	1.00
	INST	CHAU	MECA	MANO	OUTI	CUIS	CHEF	INGE	CAIS	SECR	VEND	OUVR

Le coefficient de corrélation linéaire indique la force de la liaison **linéaire** entre deux variables continues. Ce coefficient prend des valeurs comprises entre –1 et 1.

MATRICE DES VALEURS-TESTS

	INST	CHAU	MECA	MANO	OUTI	CUIS	CHEF	INGE	CAIS	SECR	VEND	OUVR
INST	99.99											
CHAU	13.87	99.99										
MECA	8.82	10.07	99.99									
MANO	8.51	9.69	12.74	99.99								
OUTI	10.82	12.31	11.91	11.90	99.99							
CUIS	6.89	7.09	7.85	6.50	7.03	99.99						
CHEF	7.47	6.79	5.44	4.88	6.10	8.34	99.99					
INGE	8.11	8.28	6.81	6.22	7.76	8.37	9.64	99.99				
CAIS	8.33	7.86	6.19	5.46	6.51	7.64	10.27	9.10	99.99			
SECR	11.12	11.73	9.79	9.18	11.18	7.86	7.93	9.63	9.62	99.99		
VEND	9.72	10.20	10.20	9.14	9.96	9.08	7.74	9.02	8.94	12.15	99.99	
OUVR	9.89	11.58	10.19	11.29	12.22	6.35	5.58	8.11	6.65	11.83	10.32	99.99
	INST	CHAU	MECA	MANO	OUTI	CUIS	CHEF	INGE	CAIS	SECR	VEND	OUVR

Cette matrice est directement liée à la précédente. SPAD a retranscrit le test sous-jacent de nullité de la corrélation, en terme de valeur-test. Plus la valeur-test sera élevée, plus la liaison linéaire sera forte. On peut également affirmer qu'une valeur-test inférieure à 2 (en valeur absolue) indique qu'il n'y a pas de liaison linéaire entre les variables.

VALEURS PROPRES

```
APERCU DE LA PRECISION DES CALCULS : TRACE AVANT DIAGONALISATION ..  12.0000
                                     SOMME DES VALEURS PROPRES ....  12.0000
```

HISTOGRAMME DES 12 PREMIERES VALEURS PROPRES

NUMERO	VALEUR PROPRE	POURCENTAGE	POURCENTAGE CUMULE	
1	10.1390	84.49	84.49	**********
2	0.8612	7.18	91.67	*******
3	0.3248	2.71	94.37	***
4	0.1715	1.43	95.80	**
5	0.1484	1.24	97.04	**
6	0.0973	0.81	97.85	*
7	0.0682	0.57	98.42	*
8	0.0525	0.44	98.86	*
9	0.0505	0.42	99.28	*
10	0.0332	0.28	99.55	*
11	0.0309	0.26	99.81	*
12	0.0226	0.19	100.00	*

RECHERCHE DE PALIERS (DIFFERENCES TROISIEMES)

PALIER ENTRE	VALEUR DU PALIER	
1 -- 2	-8358.41	**********
2 -- 3	-252.72	**
3 -- 4	-158.36	*
9 -- 10	-29.14	*
5 -- 6	-8.64	*

RECHERCHE DE PALIERS ENTRE (DIFFERENCES SECONDES)

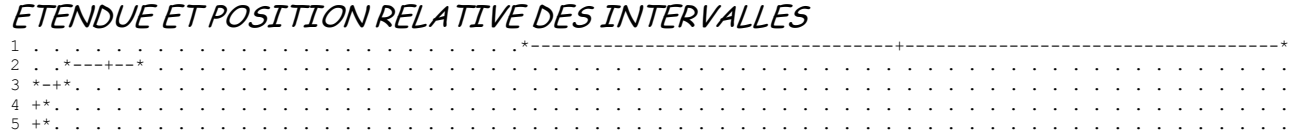
PALIER ENTRE	VALEUR DU PALIER	
1 -- 2	8741.43	**********
2 -- 3	383.02	***
3 -- 4	130.30	*
5 -- 6	22.02	*
7 -- 8	13.76	*
6 -- 7	13.38	*

INTERVALLES LAPLACIENS D'ANDERSON
INTERVALLES AU SEUIL 0.95

SPAD édite les bornes supérieures et inférieures des intervalles de confiance approchées au seuil 95% des valeurs propres correspondant à chaque axe factoriel. L'ampleur de l'intervalle donne une indication sur la stabilité de la valeur propre vis-à-vis des fluctuations dues à l'échantillonnage.

NUMERO	BORNE INFERIEURE	VALEUR PROPRE	BORNE SUPERIEURE
1	6.1645	10.1390	14.1135
2	0.5236	0.8612	1.1988
3	0.1975	0.3248	0.4521
4	0.1042	0.1715	0.2387
5	0.0902	0.1484	0.2066

ETENDUE ET POSITION RELATIVE DES INTERVALLES



COORDONNEES DES VARIABLES SUR LES AXES 1 A 5
VARIABLES ACTIVES

VARIABLES	COORDONNEES					CORRELATIONS VARIABLE-FACTEUR					ANCIENS AXES UNITAIRES				
IDEN - LIBELLE COURT	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
INST - Instituteur	0.94	0.04	0.21	-0.16	-0.13	0.94	0.04	0.21	-0.16	-0.13	0.30	0.05	0.37	-0.39	-0.34
CHAU - Chauffeur d'autobus	0.96	0.13	0.15	-0.08	-0.09	0.96	0.13	0.15	-0.08	-0.09	0.30	0.14	0.27	-0.19	-0.25
MECA - Mécanicien autos	0.92	0.27	-0.19	-0.07	0.03	0.92	0.27	-0.19	-0.07	0.03	0.29	0.29	-0.34	-0.16	0.07
MANO - Manoeuvre du bâtime	0.90	0.37	-0.11	-0.01	0.02	0.90	0.37	-0.11	-0.01	0.02	0.28	0.40	-0.19	-0.01	0.04
OUTI - Tourneur	0.95	0.24	0.02	0.01	-0.11	0.95	0.24	0.02	0.01	-0.11	0.30	0.26	0.04	0.03	-0.28
CUIS - Cuisinier chef	0.87	-0.24	-0.40	-0.05	-0.06	0.87	-0.24	-0.40	-0.05	-0.06	0.27	-0.25	-0.71	-0.12	-0.16
CHEF - Chef de service	0.84	-0.49	0.01	-0.04	-0.12	0.84	-0.49	0.01	-0.04	-0.12	0.26	-0.53	0.02	-0.11	-0.30
INGE - Ingeniéur	0.90	-0.27	0.03	0.30	-0.09	0.90	-0.27	0.03	0.30	-0.09	0.28	-0.30	0.05	0.72	-0.22
CAIS - Caissier de banque	0.88	-0.38	0.13	-0.11	0.20	0.88	-0.38	0.13	-0.11	0.20	0.28	-0.41	0.23	-0.26	0.53
SECR - Secrétaire directio	0.97	0.00	0.10	0.03	0.11	0.97	0.00	0.10	0.03	0.11	0.31	0.00	0.17	0.08	0.29
VEND - Vendeuse	0.96	-0.01	-0.08	0.00	0.17	0.96	-0.01	-0.08	0.00	0.17	0.30	-0.02	-0.14	0.00	0.44
OUVR - Ouvrière du textile	0.94	0.25	0.10	0.17	0.06	0.94	0.25	0.10	0.17	0.06	0.29	0.27	0.18	0.40	0.16

Comme il s'agit d'une analyse normée, les corrélations coïncident avec les coordonnées.

Les anciens axes unitaires sont les coefficients de la liaison linéaire entre les variables et les axes. On peut ainsi dire :

$$Axe1 = 0.30\left(\frac{Instit - Moy(Instit)}{Ecart\,type(Instit)}\right) + 0.30\left(\frac{Chauf - Moy(Chauf)}{Ecart\,type(Chauf)}\right) + 0.29...$$

Le fait le plus marquant de cette analyse est le facteur taille, très dominant, porté par la première composante. Ce facteur reflète pratiquement exclusivement la disparité des villes quand au niveau moyen des salaires. Les autres facteurs sont en quelque sorte « écrasés » par la force de ce phénomène dans le tableau des données.
Dans ce cas, il peut être intéressant de reprendre l'analyse en cherchant à éliminer des données cette connaissance que nous avons sur les salaires des villes.
On y parvient, par exemple, en divisant les salaires de chaque profession par le salaire moyen de la ville.

Remarque :
SPAD n'édite pas les contributions et les Cos² des variables continues actives, cependant, on peut aisément les obtenir à partir des résultats ci-dessus :

$$Cos^2(j,\alpha) = Coordonnée^2(j,\alpha) \quad en\ ACP\ normée$$

et

$$Contribution(j,\alpha) = Ancien\ axe\ unitaire^2(j,\alpha)$$

COORDONNEES, CONTRIBUTIONS ET COSINUS CARRES DES INDIVIDUS INDIVIDUS ACTIFS (AXES 1 A 5)

INDIVIDUS			COORDONNEES					CONTRIBUTIONS					COSINUS CARRES				
IDENTIFICATEUR	P.REL	DISTO	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
AbuDhabi94	1.96	27.16	2.17	-4.61	-0.75	0.42	-0.18	0.9	48.4	3.4	2.1	0.4	0.17	0.78	0.02	0.01	0.00
Amsterdam94	1.96	3.22	1.58	0.36	0.58	0.22	-0.12	0.5	0.3	2.0	0.6	0.2	0.77	0.04	0.10	0.02	0.00
Athenes94	1.96	4.24	-1.91	0.51	0.08	-0.17	0.04	0.7	0.6	0.0	0.3	0.0	0.86	0.06	0.00	0.01	0.00
Bangkok94	1.96	9.87	-2.97	-0.82	0.22	0.10	-0.25	1.7	1.5	0.3	0.1	0.8	0.89	0.07	0.00	0.00	0.01
Bogota94	1.96	7.14	-2.47	-0.66	-0.14	0.31	-0.26	1.2	1.0	0.1	1.1	0.9	0.86	0.06	0.00	0.01	0.01
Bombay94	1.96	21.11	-4.56	0.31	0.26	-0.28	-0.02	4.0	0.2	0.4	0.9	0.0	0.99	0.00	0.00	0.00	0.00
Bruxelles94	1.96	0.74	0.61	0.17	-0.25	0.12	0.26	0.1	0.1	0.4	0.2	0.9	0.50	0.04	0.08	0.02	0.09
Budapest94	1.96	17.74	-4.19	0.24	0.16	-0.28	-0.01	3.4	0.1	0.2	0.9	0.0	0.99	0.00	0.00	0.00	0.00
BuenosAires94	1.96	5.39	-0.89	-1.23	0.31	0.20	1.59	0.2	3.4	0.6	0.4	33.5	0.15	0.28	0.02	0.01	0.47
Caracas94	1.96	18.07	-4.24	0.06	0.01	-0.04	-0.07	3.5	0.0	0.0	0.0	0.1	1.00	0.00	0.00	0.00	0.00
Chicago94	1.96	23.64	4.42	1.25	-0.41	-0.61	-1.08	3.8	3.5	1.0	4.2	15.4	0.82	0.07	0.01	0.02	0.05
Copenhague94	1.96	7.43	2.37	0.83	-0.74	0.42	0.49	1.1	1.6	3.3	2.0	3.1	0.76	0.09	0.07	0.02	0.03
Dublin94	1.96	0.79	-0.27	0.19	0.63	0.22	0.10	0.0	0.1	2.4	0.6	0.1	0.10	0.05	0.50	0.06	0.01
Dusseldorf94	1.96	8.32	2.72	-0.24	0.64	-0.08	-0.21	1.4	0.1	2.5	0.1	0.6	0.89	0.01	0.05	0.00	0.01
Frankfurt94	1.96	10.12	3.05	-0.63	0.25	0.01	-0.15	1.8	0.9	0.4	0.0	0.3	0.92	0.04	0.01	0.00	0.00
Geneve94	1.96	42.20	6.36	0.30	0.75	-0.76	0.40	7.8	0.2	3.4	6.6	2.1	0.96	0.00	0.01	0.01	0.00
Helsinki94	1.96	0.49	0.03	0.51	0.03	0.05	0.04	0.0	0.6	0.0	0.0	0.0	0.00	0.53	0.00	0.00	0.00
Hongkong94	1.96	3.61	-1.03	-0.54	0.57	0.73	-0.49	0.2	0.7	2.0	6.1	3.2	0.30	0.08	0.09	0.15	0.07
Houston94	1.96	15.21	3.45	0.78	-1.37	0.23	-0.04	2.3	1.4	11.3	0.6	0.0	0.78	0.04	0.12	0.00	0.00
Jakarta94	1.96	16.92	-4.08	0.20	0.11	-0.21	-0.14	3.2	0.1	0.1	0.5	0.3	0.98	0.00	0.00	0.00	0.00
Johannesburg94	1.96	4.88	-2.08	0.02	0.12	0.20	-0.12	0.8	0.0	0.1	0.4	0.2	0.88	0.00	0.00	0.01	0.00
Lagos94	1.96	23.54	-4.81	0.43	0.34	-0.25	0.04	4.5	0.4	0.7	0.7	0.0	0.98	0.01	0.00	0.00	0.00
Lisboa94	1.96	5.00	-2.17	0.28	0.14	-0.09	0.21	0.9	0.2	0.1	0.1	0.6	0.94	0.02	0.00	0.00	0.01
London94	1.96	0.76	-0.02	0.63	0.19	0.21	-0.18	0.0	0.9	0.2	0.5	0.4	0.00	0.52	0.05	0.06	0.04
LosAngeles94	1.96	18.89	3.64	1.80	0.34	1.09	-0.33	2.6	7.4	0.7	13.7	1.4	0.70	0.17	0.01	0.06	0.01
Luxembourg94	1.96	32.79	5.24	-0.69	1.91	0.09	-0.78	5.3	1.1	22.1	0.1	8.0	0.84	0.01	0.11	0.00	0.02
Madrid94	1.96	0.89	-0.06	0.00	-0.32	-0.09	0.27	0.0	0.0	0.6	0.1	1.0	0.00	0.00	0.11	0.01	0.08
Manama94	1.96	7.17	-0.82	-2.05	-0.70	-0.81	-0.62	0.1	9.6	2.9	7.5	5.1	0.09	0.59	0.07	0.09	0.05
Manila94	1.96	16.51	-4.05	0.03	0.10	-0.19	-0.12	3.2	0.0	0.1	0.4	0.2	0.99	0.00	0.00	0.00	0.00
Mexico94	1.96	8.63	-2.83	0.07	-0.12	0.32	-0.25	1.5	0.0	0.1	1.2	0.8	0.93	0.00	0.00	0.01	0.01
Milan94	1.96	0.69	0.02	0.34	-0.22	0.14	0.37	0.0	0.3	0.3	0.2	1.9	0.00	0.17	0.07	0.03	0.20
Montreal94	1.96	5.68	2.17	0.77	-0.40	0.09	0.20	0.9	1.3	1.0	0.1	0.5	0.83	0.10	0.03	0.00	0.01
Nairobi94	1.96	23.45	-4.82	0.26	0.20	-0.25	-0.03	4.5	0.2	0.2	0.7	0.0	0.99	0.00	0.00	0.00	0.00
NewYork94	1.96	23.01	4.60	0.30	-0.99	0.35	-0.35	4.1	0.2	6.0	1.4	1.7	0.92	0.00	0.04	0.01	0.01
Nicosia94	1.96	3.56	-1.78	0.27	0.27	-0.26	-0.10	0.6	0.2	0.4	0.8	0.1	0.89	0.02	0.02	0.02	0.00
Oslo94	1.96	3.98	1.66	0.73	-0.02	0.37	0.38	0.5	1.2	0.0	1.6	1.9	0.69	0.13	0.00	0.03	0.04
Panama94	1.96	5.97	-2.22	-0.62	-0.02	-0.20	0.33	1.0	0.9	0.0	0.4	1.4	0.83	0.06	0.00	0.01	0.02
Paris94	1.96	5.31	1.65	-1.41	-0.07	0.53	0.05	0.5	4.5	0.0	3.3	0.0	0.51	0.37	0.00	0.05	0.00
Prague94	1.96	18.69	-4.29	0.31	0.13	-0.36	-0.01	3.6	0.2	0.1	1.5	0.0	0.98	0.01	0.00	0.01	0.00
RiodeJaneiro94	1.96	12.22	-3.40	-0.28	0.41	-0.02	0.05	2.2	0.2	1.0	0.0	0.0	0.95	0.01	0.01	0.00	0.00
SaoPaulo94	1.96	10.33	-3.18	0.01	0.13	-0.04	-0.11	2.0	0.0	0.1	0.0	0.2	0.98	0.00	0.00	0.00	0.00
Seoul94	1.96	0.69	-0.61	0.04	-0.01	-0.21	0.32	0.1	0.0	0.0	0.5	1.4	0.54	0.00	0.00	0.07	0.15
Singapore94	1.96	2.64	-1.16	0.16	0.32	0.69	-0.18	0.3	0.1	0.6	5.5	0.4	0.51	0.01	0.04	0.18	0.01
Stockholm94	1.96	2.16	0.67	0.98	-0.13	0.19	0.51	0.1	2.2	0.1	0.4	3.5	0.21	0.45	0.01	0.02	0.12
Sidney94	1.96	0.62	0.03	0.21	-0.32	-0.20	0.33	0.0	0.1	0.6	0.5	1.4	0.00	0.07	0.17	0.07	0.17
Taipei94	1.96	6.07	1.64	0.27	-1.36	-1.02	-0.10	0.5	0.2	11.2	11.9	0.1	0.45	0.01	0.30	0.17	0.00
Tel-Aviv94	1.96	3.35	-1.41	0.00	-0.98	-0.15	-0.25	0.4	0.0	5.8	0.3	0.8	0.59	0.00	0.29	0.01	0.02
Tokyo94	1.96	46.73	6.72	-0.72	-0.13	-0.05	0.47	8.7	1.2	0.1	0.0	2.9	0.97	0.01	0.00	0.00	0.00
Toronto94	1.96	4.86	1.77	1.06	-0.57	-0.16	-0.20	0.6	2.5	1.9	0.3	0.5	0.64	0.23	0.07	0.01	0.01
Vienna94	1.96	4.07	1.86	0.11	-0.36	0.61	-0.02	0.7	0.0	0.8	4.3	0.0	0.85	0.00	0.03	0.09	0.00
Zurich94	1.96	65.45	7.90	-0.29	1.18	-1.12	0.33	12.1	0.2	8.4	14.3	1.4	0.95	0.00	0.02	0.02	0.00

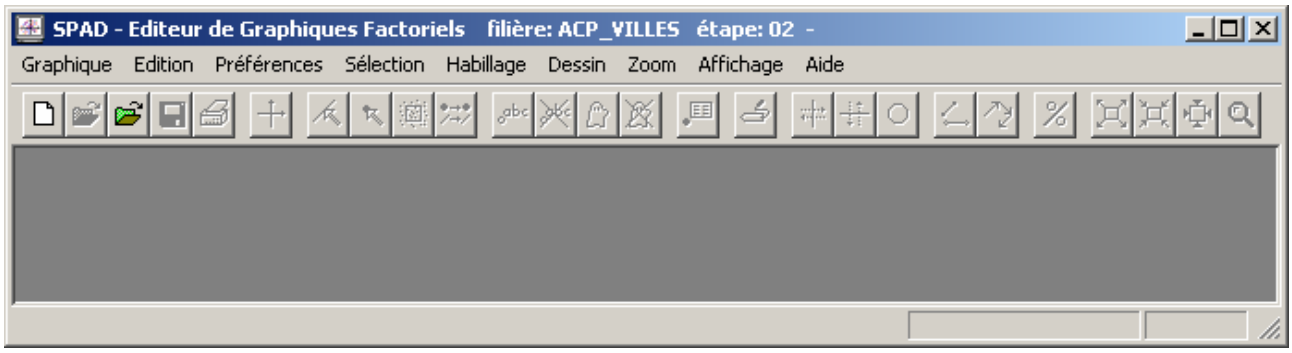
DISTO : distance à l'origine = carré de la distance de la ville au centre de gravité de toutes les villes. Cela permet de trouver facilement quelles sont les villes les plus « moyennes » (les plus proches du centre de gravité) et celles qui sont les plus « originales » (celles qui sont à la plus grande distance du centre de gravité). La distance au centre est en quelque sorte un critère « d'originalité » de l'élément.

$$Contribution(i,\alpha) = \frac{p_i \psi_{i\alpha}^2}{\lambda_\alpha} \quad et \quad Cos^2(i,\alpha) = \frac{\psi_{i\alpha}^2}{d^2(i,G)}$$

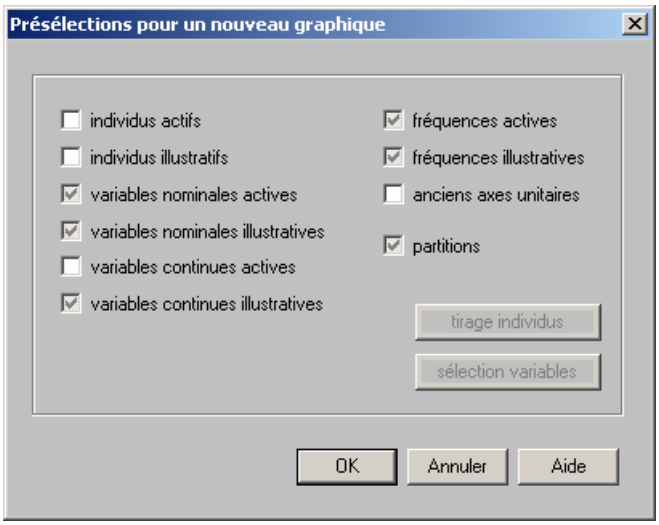
$$\sum_{i=1}^{n} Contribution(i,\alpha) = 100 \quad et \quad \sum_{\alpha=1}^{p} Cos^2(i,\alpha) = 1$$

L'EDITEUR DE GRAPHIQUES FACTORIELS

Pour accéder à **l'Editeur de plan factoriel**, double cliquez sur l'icône  .



Pour créer un graphique factoriel, sélectionnez « **Graphique** » - « **Nouveau** » ce qui affiche une sous-fenêtre de « **Présélection pour un nouveau graphique** ».

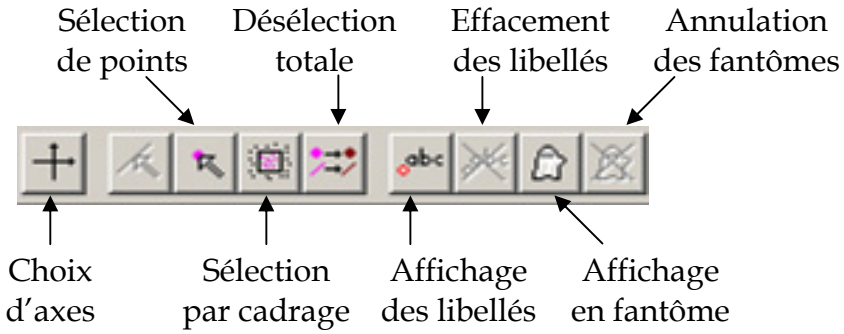


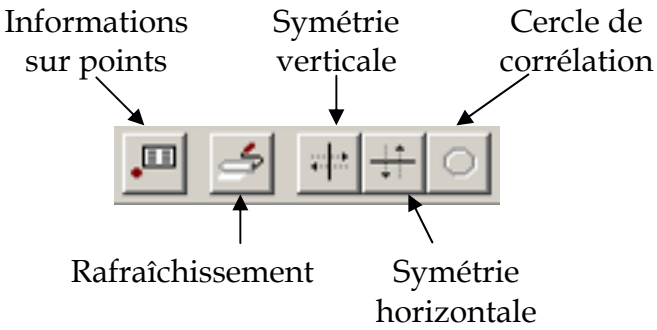
La présélection initiale pour un nouveau graphique est importante.

Il faut présélectionner les catégories de variables et d'individus dont vous pensez avoir besoin simultanément pour dessiner le plan factoriel.

Si vous ne présélectionnez, par exemple, que les variables, il vous est impossible d'introduire les individus dans ce graphique par la suite.

LA BARRE D'OUTILS DE L'EDITEUR DE PLAN FACTORIEL





LA SAUVEGARDE DES GRAPHIQUES

La **sauvegarde INTERNE** est liée à la méthode ; c'est-à-dire qu'en cas de ré-exécution de la méthode ou de suppression (par l'utilisateur) des résultats, ces sauvegardes sont détruites. L'intérêt de la sauvegarde au format interne est que toutes les fonctionnalités d'habillage de l'éditeur de plan factoriel restent disponibles.

La **sauvegarde au format ARCHIVE** est une sauvegarde indépendante de la méthode, certains habillage ne sont alors plus possibles, en particulier sur les habillages des individus.

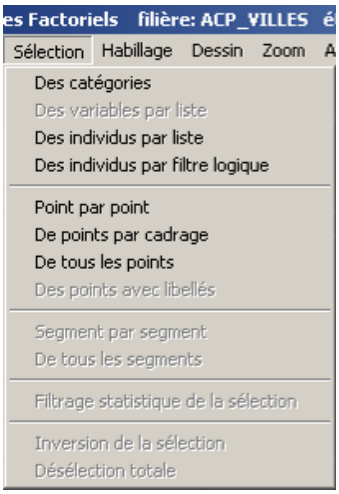
Conclusion : En cours d'élaboration de graphiques sur une analyse terminée, il est préférable de faire des sauvegardes internes de façon à bénéficier de toutes les possibilités d'habillage à chaque rappel d'un graphique. On utilisera les sauvegardes archives si l'on considère que les graphiques sont terminées.

LA REGLE FONDAMENTALE D'HABILLAGE D'UN GRAPHIQUE

L'enchaînement clé pour tout habillage d'un plan factoriel sera toujours :

Sélection / Habillage (action) / Désélection

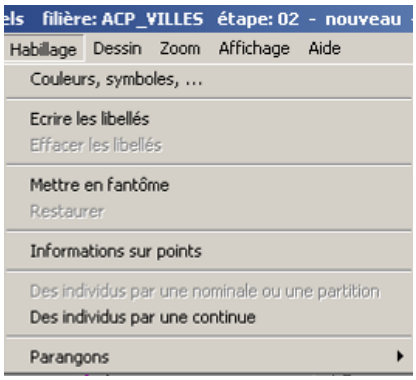
On commence par **sélectionner** un point, un groupe de points ou tous les points d'un même type, puis on **habille** la sélection faite, enfin on **désélectionne** pour voir le résultat de l'habillage.



Le menu « **Sélection** » permet d'effectuer des sélections de points, soit par groupe de points, soit point par point.

La sélection peut s'effectuer soit à partir du menu « Sélection » soit à partir des boutons de la barre d'outils.

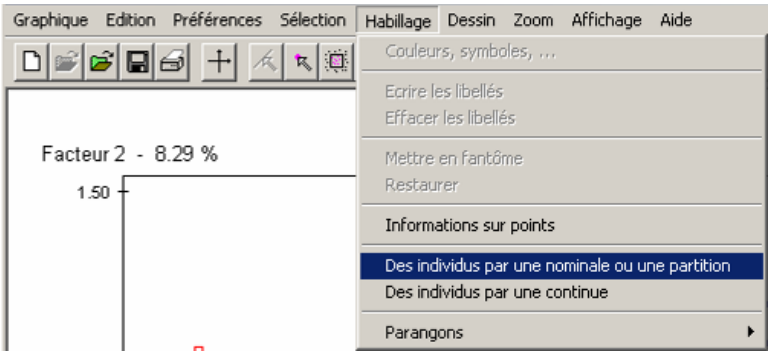
Le menu « **Habillage** » permet d'habiller les points sélectionnés au préalable. On peut utiliser de la même façon le menu « Habillage » ou les boutons de la barre d'outils.

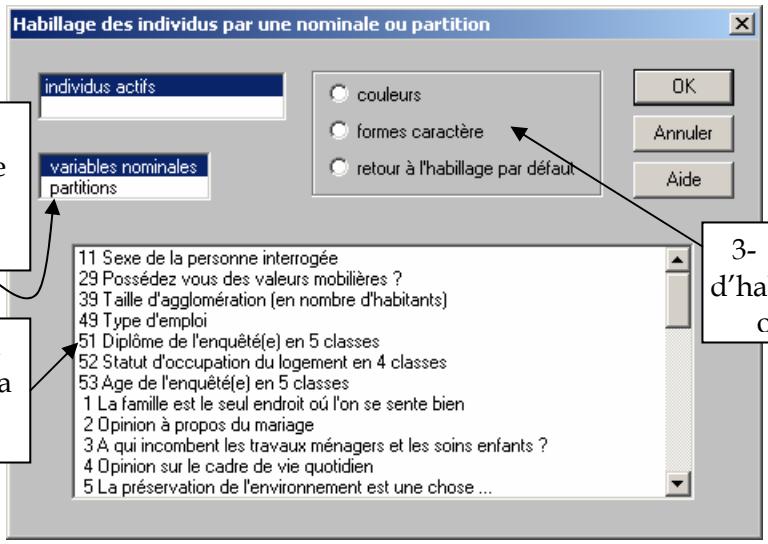


On **désélectionne** enfin pour voir l'effet de l'habillage en cliquant sur le bouton  de la barre d'outils ou par la commande « **Désélection totale** » du menu « **Sélection** » ou « **Ctrl + D** ».

L'éditeur de plan factoriel est totalement interactif.
Lors de certaines manipulations comme notamment le déplacement des libellés, le graphique peut présenter des chevauchements ou des doubles affichages, il faut dans ce cas **rafraîchir** le graphique en cliquant sur le bouton  de la barre d'outils ou par la commande « **Rafraîchir** » du menu « Dessin ».

HABILLAGE EN FONCTION DES VALEURS D'UNE VARIABLE NOMINALE OU D'UNE PARTITION

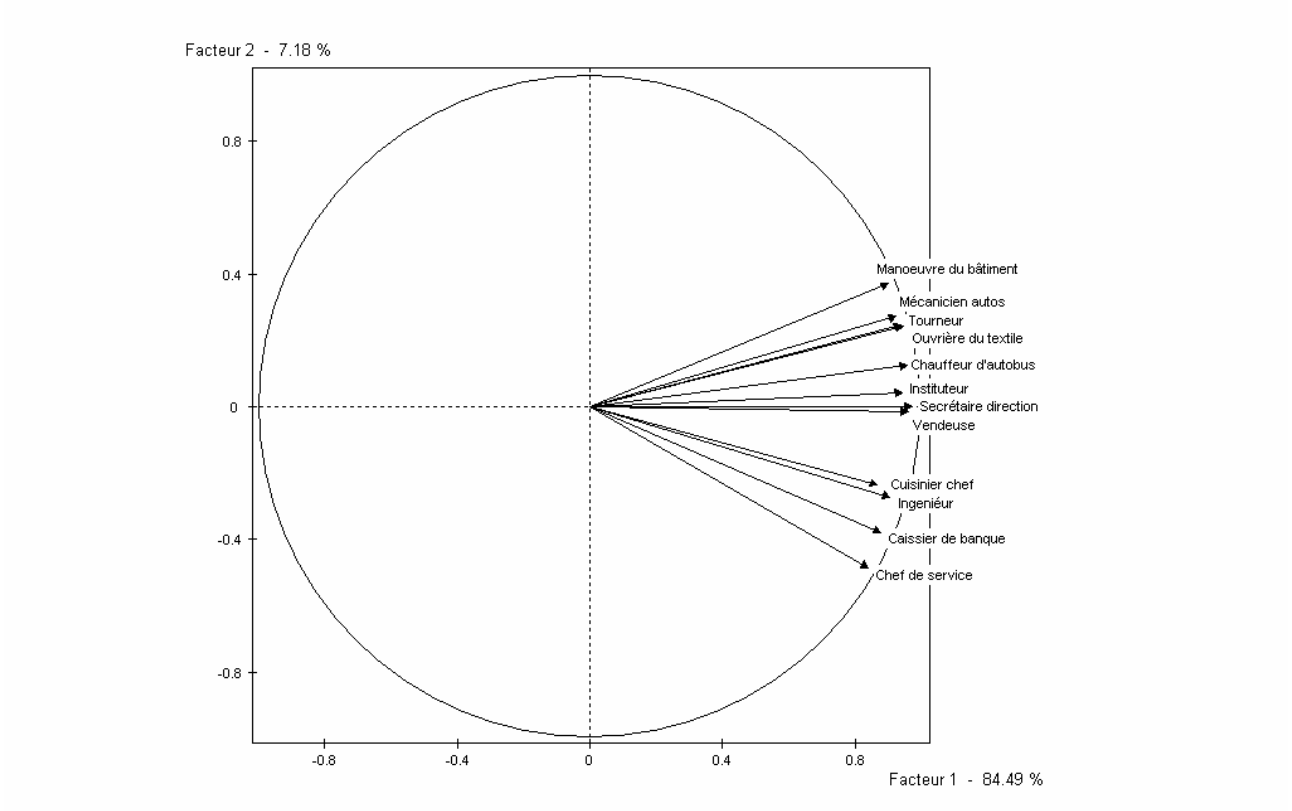


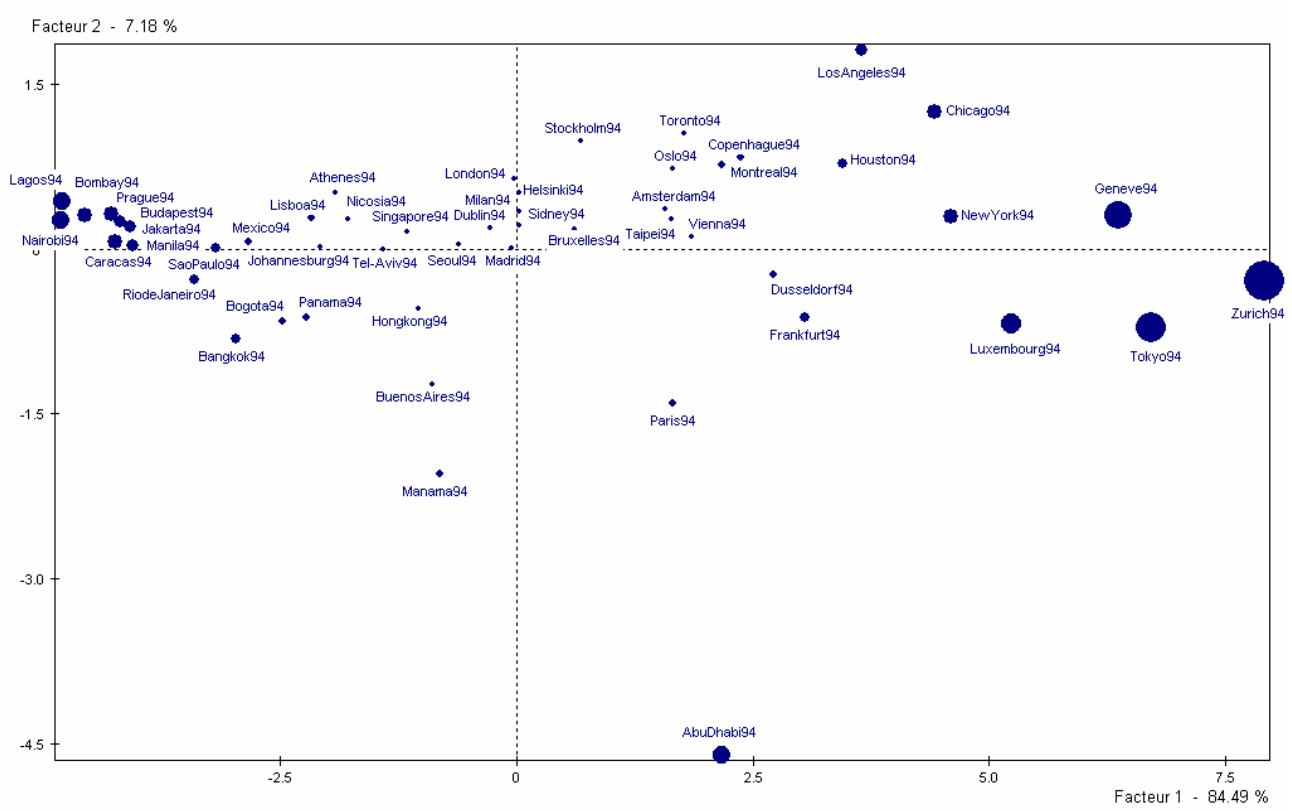


1- Sélectionner le type de la variable (nominale ou partition)

2- Sélectionner la variable qui servira pour l'habillage

3- Choisir le type d'habillage (couleurs ou caractères)


Facteur 2 - 7.18 %
0.8
0.4
0
-0.4
-0.8
-0.8
-0.4
0
0.4
0.8
Facteur 1 - 84.49 %
Manoeuvre du bâtiment
Mécanicien autos
Tourneur
Ouvrière du textile
Chauffeur d'autobus
Instituteur
Secrétaire direction
Vendeuse
Cuisinier chef
Ingeniéur
Caissier de banque
Chef de service


Facteur 2 - 7.18 %
1.5
-1.5
-3.0
-4.5
-2.5
0
2.5
5.0
7.5
Facteur 1 - 84.49 %
LosAngeles94
Chicago94
Stockholm94
Toronto94
Copenhague94
Oslo94
Houston94
Montreal94
Geneve94
Lagos94
Bombay94
Athenes94
London94
Helsinki94
Amsterdam94
Prague94
Lisboa94
Nicosia94
Milan94
Sidney94
NewYork94
Budapest94
Singapore94
Dublin94
Taipei94
Vienna94
Jakarta94
Mexico94
Nairobi94
Manila94
Bruxelles94
Caracas94
SaoPaulo94
Johannesburg94
Tel-Aviv94
Seoul94
Madrid94
Dusseldorf94
Zurich94
RiodeJaneiro94
Bogota94
Panama94
Hongkong94
Frankfurt94
Luxembourg94
Tokyo94
Bangkok94
BuenosAires94
Paris94
Manama94
AbuDhabi94

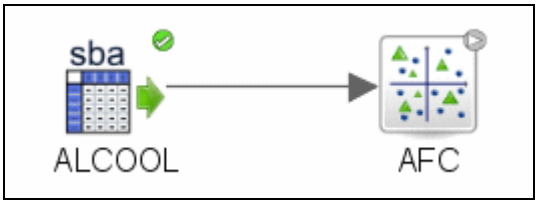
AFC - Analyse des Correspondances

Cette procédure effectue l'analyse des correspondances (binaires) sur un tableau de contingence ou d'un tableau de nombres non négatifs.

Nous effectuerons l'analyse du tableau suivant auquel on a ajouté les marges.
On cherche à étudier la perception de différentes boissons alcoolisées.

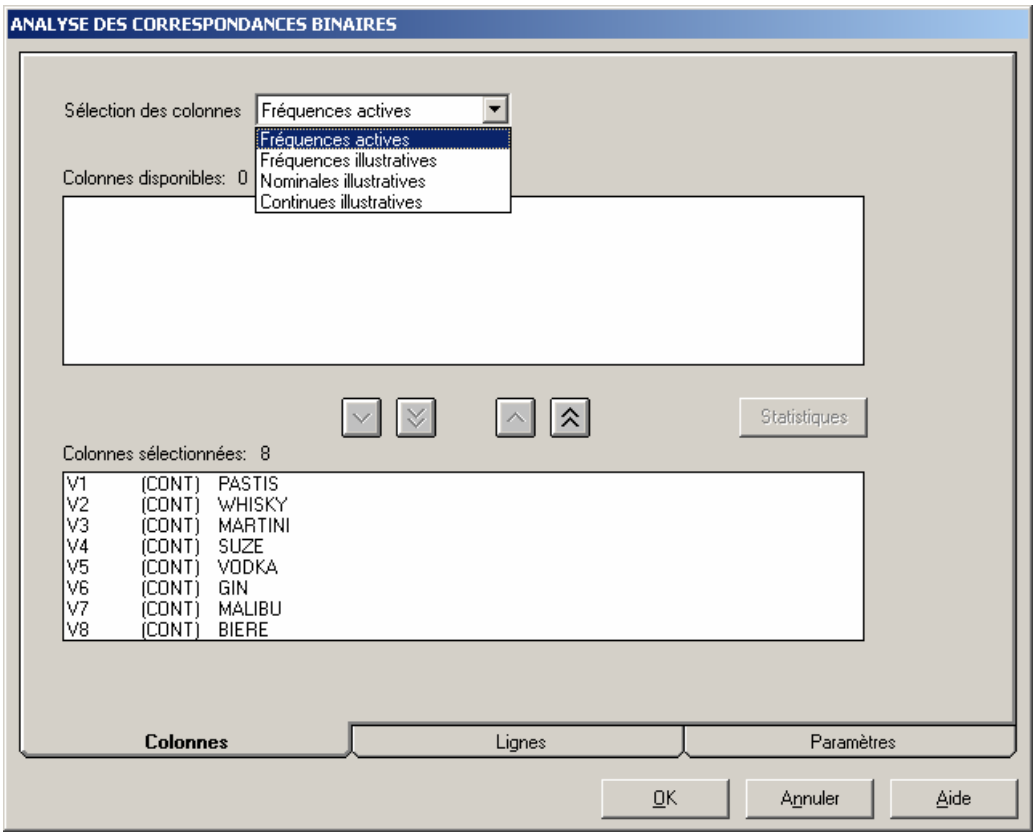
	PASTIS	WHISKY	MARTINI	SUZE	VODKA	GIN	MALIBU	BIERE	TOTAL
Aime le goût	49	50	42	18	25	23	25	59	291
Avec des amis	83	83	76	60	69	68	69	74	582
Pour se détendre	61	61	51	32	38	39	39	72	393
Qui revient cher	60	88	42	41	75	70	61	19	456
Rafraîchissante, désaltérante	78	22	18	19	17	19	14	80	267
Peu élégante, peu distinguée	26	11	13	17	13	11	13	29	133
Produit sympathique	64	64	56	34	45	42	46	68	419
Bien avant les repas	88	79	85	64	45	46	37	41	485
Bien dans la journée	24	21	12	10	13	12	13	85	190
Bien dans la soirée	7	61	12	11	53	50	48	54	296
Toute l'année	83	87	85	79	83	82	80	90	669
Appréciée des jeunes	45	77	36	16	65	69	76	89	473
Volontiers avec invités	88	92	87	60	70	67	67	81	612
Vieillotte, dépassée	12	4	13	38	5	6	8	7	93
Aussi bien hommes que femmes	50	62	69	43	49	51	61	60	445
Très proche	38	41	27	11	16	18	17	49	217
Par habitude	36	30	24	16	19	19	17	40	201
Fait snob, m'as-tu vu ?	3	35	9	8	28	25	21	4	133
On peut mélanger	43	87	29	32	82	80	43	40	436
La nuit/Bar/Disco	12	91	27	16	84	81	72	67	450
TOTAL	950	1 146	813	625	894	878	827	1 108	7 241

- Importer la base Sba Alcool.sba de la connexion Bases Sba.
- Glissez-déposez la méthode Analyse Factorielle des Correspondances simples.



LES PARAMETRES DE LA METHODE **AFC**

L'ONGLET « **COLONNES** »

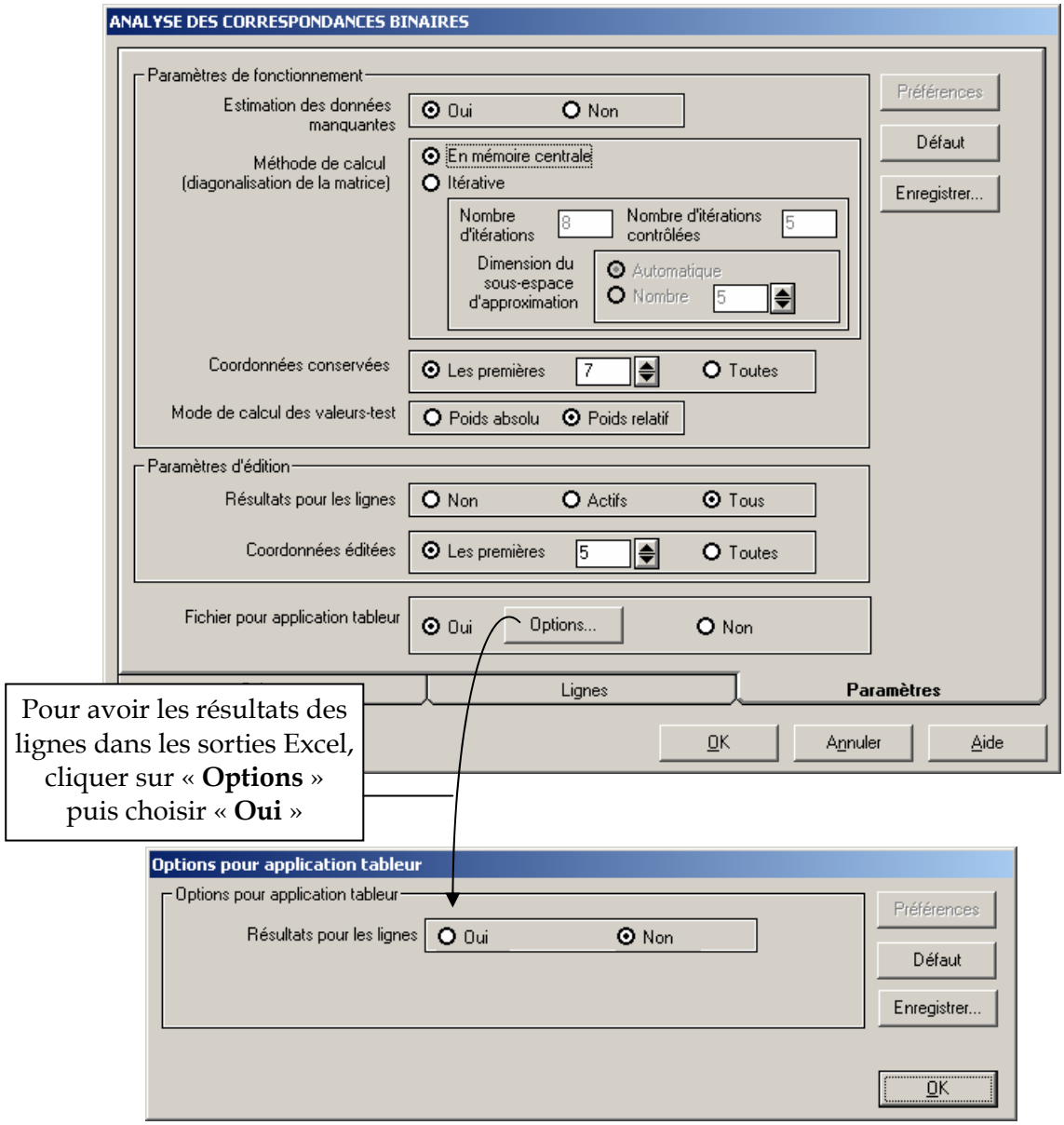


Fréquences actives : toutes les fréquences (colonnes)

L'ONGLET « **LIGNES** »

Onglet identique aux onglets « **Individus** » des méthodes de description statistique.

L'ONGLET « **PARAMETRES** »



Pour avoir les résultats des lignes dans les sorties Excel, cliquer sur « **Options** » puis choisir « **Oui** »

LES RESULTATS DE LA METHODE AFC

ANALYSE DES CORRESPONDANCES

VALEURS PROPRES

```
APERCU DE LA PRECISION DES CALCULS : TRACE AVANT DIAGONALISATION ..   0.1345
                                     SOMME DES VALEURS PROPRES ....   0.1345
```

HISTOGRAMME DES 7 PREMIERES VALEURS PROPRES

NUM	VALEUR PROPRE	POURCENT	POURCENT CUMULE	
1	0.0664	49.37	49.37	**********
2	0.0449	33.34	82.72	*******
3	0.0124	9.24	91.96	**
4	0.0069	5.14	97.09	*
5	0.0029	2.18	99.27	
6	0.0008	0.63	99.90	
7	0.0001	0.10	100.00	

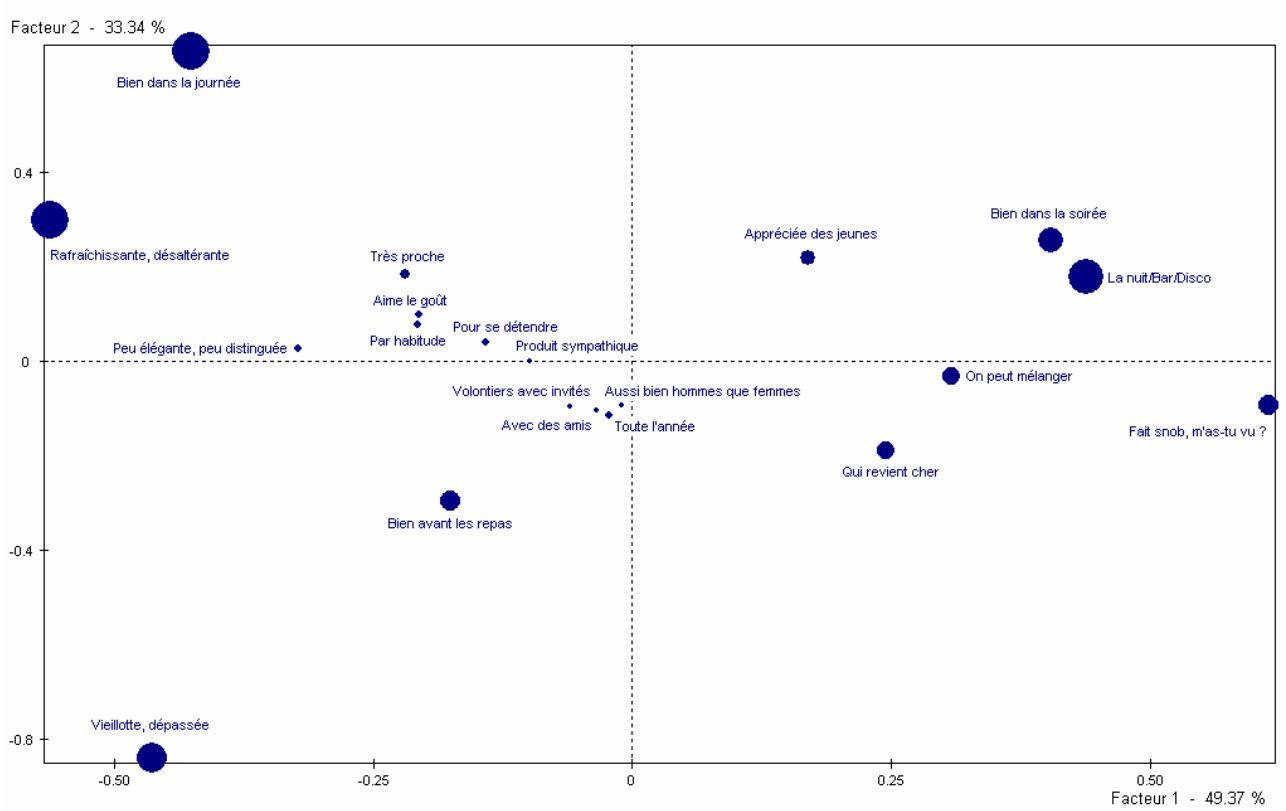
COORDONNEES, CONTRIBUTIONS DES FREQUENCES SUR LES AXES 1 A 5 FREQUENCES ACTIVES

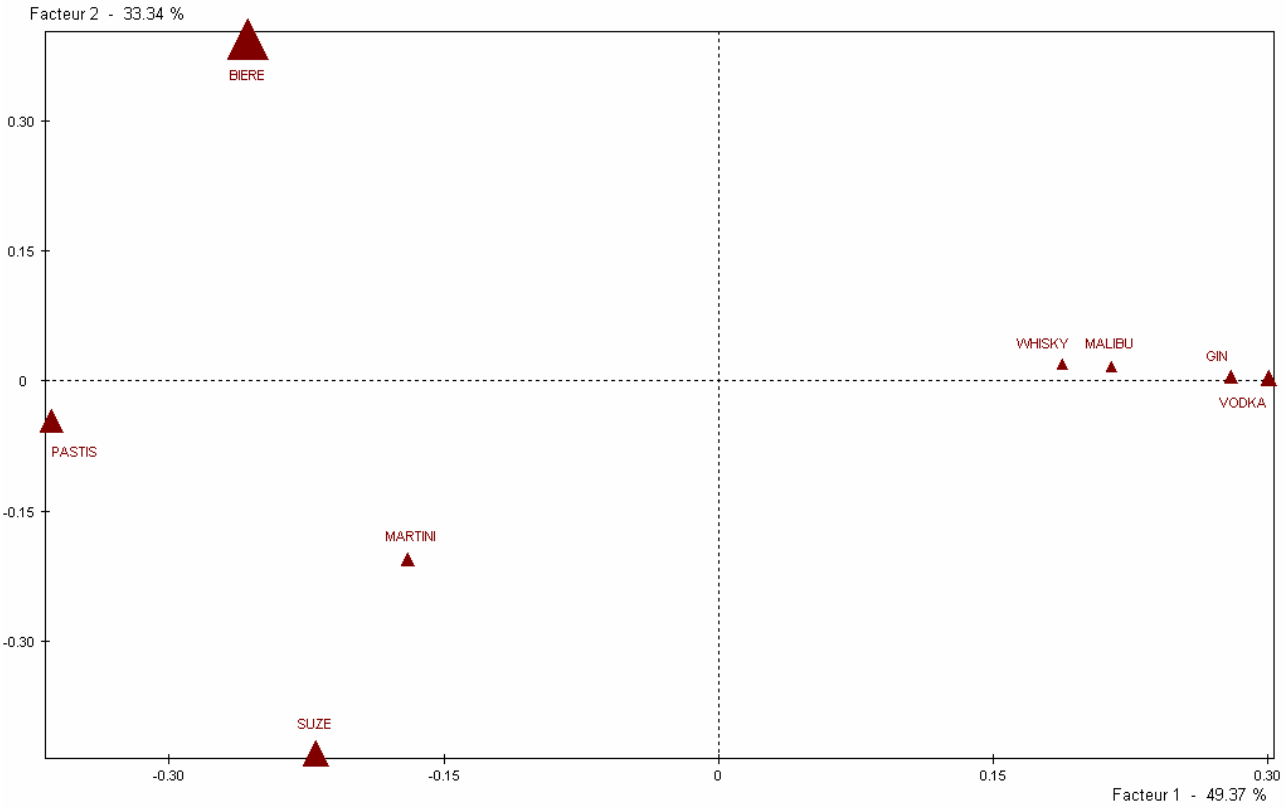
FREQUENCES			COORDONNEES					CONTRIBUTIONS					COSINUS CARRES				
IDEN - LIB COURT	P.REL	DISTO	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
PAST - PASTIS	13.12	0.17	-0.36	-0.05	0.16	0.11	-0.04	26.3	0.6	26.5	23.5	8.3	0.76	0.01	0.14	0.07	0.01
WHIS - WHISKY	15.83	0.05	0.19	0.02	0.09	-0.02	0.09	8.4	0.1	9.7	0.6	39.5	0.67	0.01	0.15	0.00	0.14
MART - MARTINI	11.23	0.11	-0.17	-0.21	0.09	-0.17	0.00	4.9	10.5	7.2	49.7	0.0	0.26	0.38	0.07	0.28	0.00
SUZE - SUZE	8.63	0.30	-0.22	-0.43	-0.24	0.05	0.04	6.3	35.6	40.7	3.2	3.9	0.16	0.62	0.20	0.01	0.00
VODK - VODKA	12.35	0.10	0.30	0.00	-0.01	0.06	0.00	16.8	0.0	0.0	7.2	0.0	0.94	0.00	0.00	0.04	0.00
GIN - GIN	12.13	0.08	0.28	0.00	-0.01	0.06	-0.01	14.3	0.0	0.1	5.9	0.7	0.94	0.00	0.00	0.04	0.00
MALI - MALIBU	11.42	0.07	0.21	0.02	-0.06	-0.07	-0.11	7.9	0.1	3.0	8.7	45.9	0.67	0.00	0.05	0.08	0.17
BIER - BIERE	15.30	0.23	-0.26	0.39	-0.10	-0.02	0.02	15.2	53.1	12.7	1.1	1.7	0.28	0.67	0.04	0.00	0.00

P.REL : poids relatif de la fréquence active. Le poids relatif se calcule de la façon suivante : (n_q * 100) / n avec n_q l'effectif de la fréquence active et ne l'effectif total. Par exemple, pour la fréquence active « Pastis », P.REL = (950 * 100) / 7241 = 13.12 .

COORDONNEES, CONTRIBUTIONS ET COSINUS CARRES DES INDIVIDUS AXES 1 A 5

INDIVIDUS			COORDONNEES					CONTRIBUTIONS					COSINUS CARRES				
IDENTIFICATEUR	P.REL	DISTO	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
Aime le goût	4.02	0.08	-0.21	0.10	0.12	-0.08	0.06	2.6	0.9	4.4	3.9	5.2	0.55	0.13	0.18	0.09	0.05
Avec des amis	8.04	0.01	-0.04	-0.10	0.00	-0.01	-0.04	0.1	1.9	0.0	0.2	3.9	0.09	0.79	0.00	0.01	0.10
Pour se détendre	5.43	0.03	-0.14	0.04	0.04	-0.04	0.02	1.6	0.2	0.7	1.1	0.7	0.79	0.07	0.06	0.06	0.02
Qui revient cher	6.30	0.12	0.25	-0.19	0.09	0.11	-0.03	5.7	5.0	4.2	10.1	1.8	0.51	0.30	0.07	0.09	0.01
Rafraîchissante, désalté	3.69	0.48	-0.56	0.30	0.07	0.25	-0.07	17.5	7.3	1.5	33.0	6.9	0.66	0.19	0.01	0.13	0.01
Peu élégante, peu distin	1.84	0.14	-0.32	0.03	-0.12	0.11	-0.08	2.9	0.0	2.0	3.0	3.9	0.76	0.01	0.10	0.08	0.05
Produit sympathique	5.79	0.01	-0.10	0.00	0.05	-0.04	-0.01	0.8	0.0	1.2	1.6	0.2	0.67	0.00	0.18	0.13	0.01
Bien avant les repas	6.70	0.14	-0.18	-0.30	0.11	-0.03	0.06	3.1	13.0	6.7	0.8	8.5	0.23	0.64	0.09	0.01	0.03
Bien dans la journée	2.62	0.69	-0.43	0.66	-0.25	-0.04	0.11	7.2	25.1	13.0	0.5	10.6	0.26	0.63	0.09	0.00	0.02
Bien dans la soirée	4.09	0.25	0.40	0.26	-0.12	-0.01	0.03	10.0	6.0	5.1	0.0	0.9	0.66	0.27	0.06	0.00	0.00
Toute l'année	9.24	0.02	-0.02	-0.11	-0.08	-0.01	-0.03	0.1	2.7	4.2	0.3	3.7	0.02	0.60	0.26	0.01	0.05
Appréciée des jeunes	6.53	0.09	0.17	0.22	-0.02	-0.03	-0.09	2.8	7.0	0.2	0.7	17.5	0.33	0.55	0.01	0.01	0.09
Volontiers avec invités	8.45	0.02	-0.06	-0.10	0.03	-0.04	-0.01	0.5	1.7	0.7	2.2	0.2	0.23	0.57	0.07	0.11	0.00
Vieillotte, dépassée	1.28	1.41	-0.46	-0.84	-0.68	0.11	0.08	4.1	20.2	47.5	2.3	2.9	0.15	0.50	0.33	0.01	0.00
Aussi bien hommes que fe	6.15	0.03	-0.01	-0.09	-0.02	-0.14	-0.06	0.0	1.2	0.3	16.3	6.6	0.00	0.28	0.02	0.59	0.10
Très proche	3.00	0.11	-0.22	0.19	0.13	-0.05	0.10	2.2	2.3	4.1	1.0	9.6	0.42	0.30	0.15	0.02	0.08
Par habitude	2.78	0.05	-0.21	0.08	0.06	0.02	0.03	1.8	0.4	0.8	0.2	0.6	0.80	0.11	0.06	0.01	0.01
Fait snob, m'as-tu vu ?	1.84	0.40	0.61	-0.09	0.03	0.02	0.09	10.4	0.4	0.1	0.1	4.6	0.95	0.02	0.00	0.00	0.02
On peut mélanger	6.02	0.13	0.31	-0.03	0.03	0.16	0.07	8.6	0.1	0.5	22.3	11.4	0.72	0.01	0.01	0.19	0.04
La nuit/Bar/Disco	6.21	0.23	0.44	0.18	-0.07	-0.02	0.01	17.9	4.4	2.7	0.3	0.1	0.84	0.14	0.02	0.00	0.00


Facteur 2 - 33.34 %
Bien dans la journée
Rafraîchissante, désaltérante
Très proche
Aime le goût
Par habitude
Pour se détendre
Produit sympathique
Peu élégante, peu distinguée
Appréciée des jeunes
Bien dans la soirée
La nuit/Bar/Disco
On peut mélanger
Volontiers avec invités
Aussi bien hommes que femmes
Avec des amis
Toute l'année
Fait snob, m'as-tu vu ?
Qui revient cher
Bien avant les repas
Vieillotte, dépassée
0.4
0
-0.4
-0.8
-0.50
-0.25
0
0.25
0.50
Facteur 1 - 49.37 %

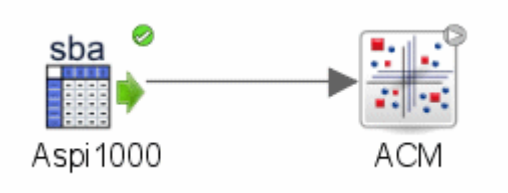

Facteur 2 - 33.34 %
BIERE
WHISKY
MALIBU
GIN
VODKA
PASTIS
MARTINI
SUZE
0.30
0.15
0
-0.15
-0.30
-0.30
-0.15
0
0.15
0.30
Facteur 1 - 49.37 %

ACM - Analyse des Correspondances Multiples

Cette procédure effectue l'analyse des correspondances multiples d'un ensemble d'individus caractérisés par des variables nominales. Si des variables continues (illustratives) sont présentes, on calcule les corrélations entre ces variables et les axes de l'analyse.

Nous effectuerons l'analyse du fichier ASPI1000.SBA.

- Importer la base ASPI1000.sba de la connexion « Bases Sba ».
- Glissez-déposez la méthode Analyse des Correspondances Multiples.

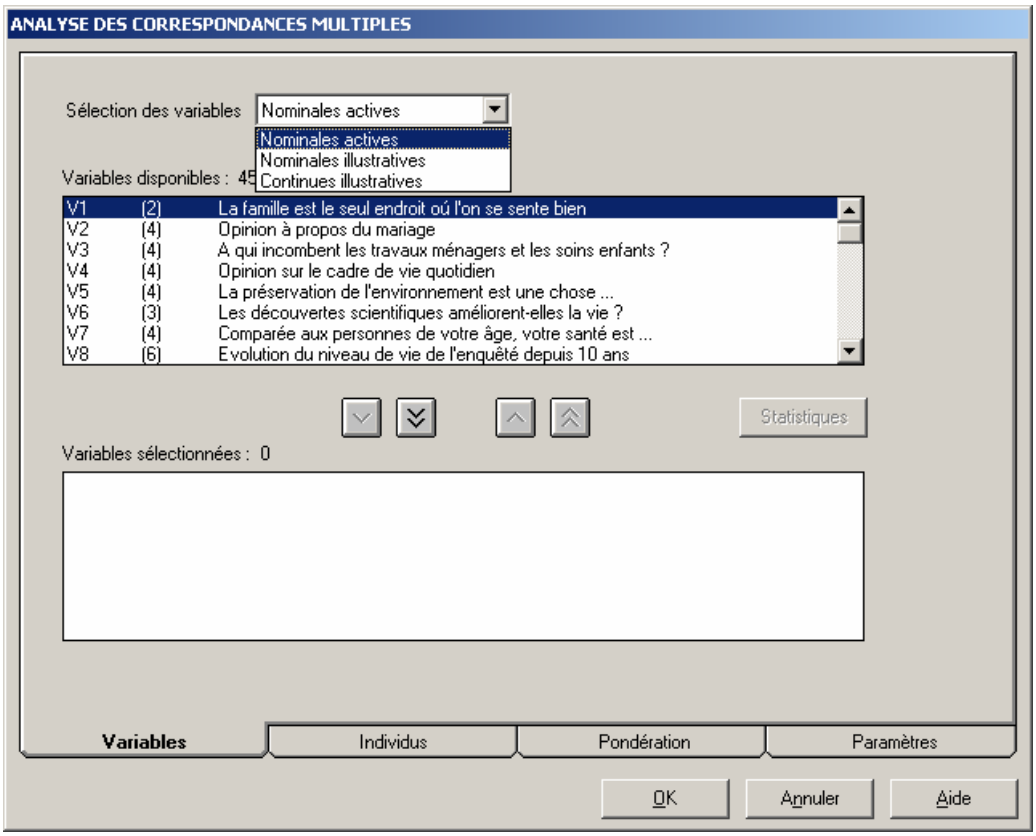


DESCRIPTION DES VARIABLES DE LA BASE ASPI1000.SBA

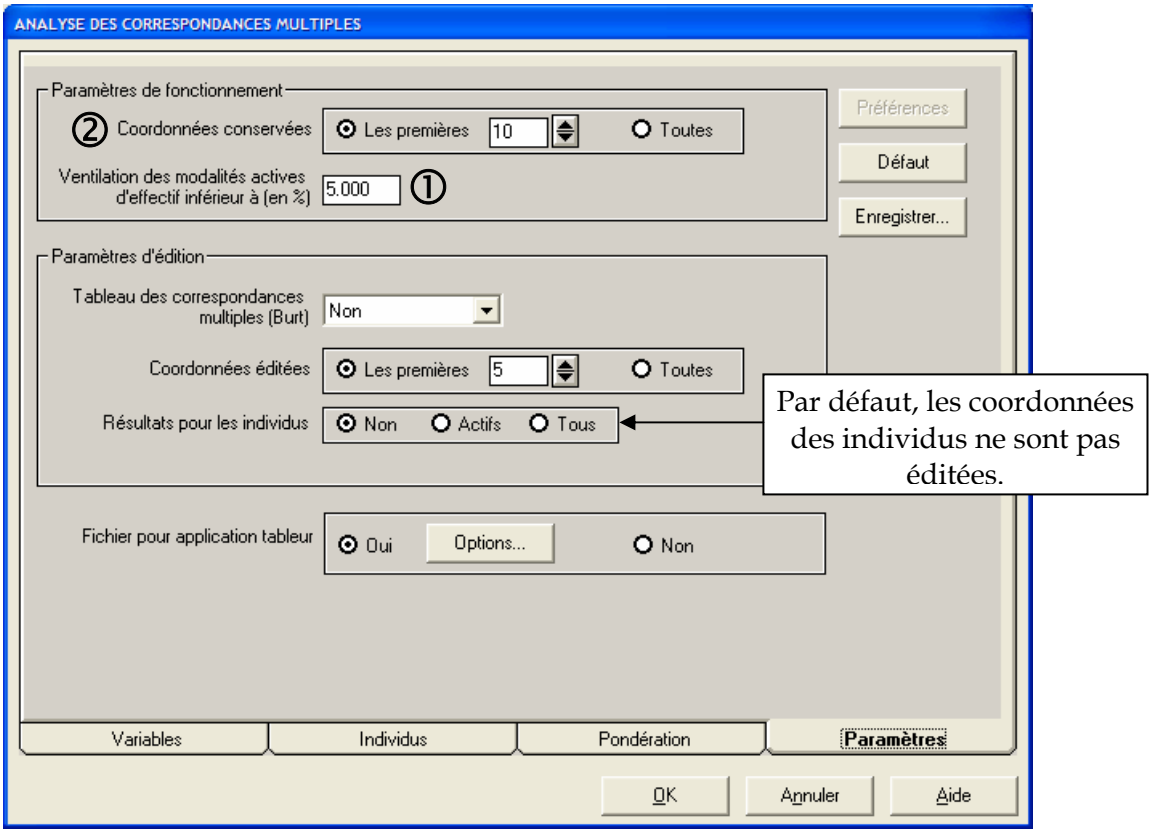
VARIABLES NOMINALES ACTIVES - 7 VARIABLES - 28 MODALITÉS ASSOCIEES		
11 .	Sexe de la personne interrogée	(2 modalités)
29 .	Possédez vous des valeurs mobilières ?	(2 modalités)
39 .	Taille d'agglomération (en nombre d'habitants)	(5 modalités)
49 .	Type d'emploi	(5 modalités)
51 .	Diplôme de l'enquêté(e) en 5 classes	(5 modalités)
52 .	Statut d'occupation du logement en 4 classes	(4 modalités)
53 .	Age de l'enquêté(e) en 5 classes	(5 modalités)
VARIABLES NOMINALES ILLUSTRATIVES - 35 VARIABLES - 152 MODALITÉS ASSOCIEES		
1 .	La famille est le seul endroit où l'on se sente bien	(3 modalités)
2 .	Opinion à propos du mariage	(5 modalités)
3 .	A qui incombent les travaux ménagers et les soins enfants ?	(5 modalités)
4 .	Opinion sur le cadre de vie quotidien	(5 modalités)
5 .	La préservation de l'environnement est une chose ...	(5 modalités)
6 .	Les découvertes scientifiques améliorent-elles la vie ?	(4 modalités)
7 .	Comparée aux personnes de votre âge, votre santé est ...	(4 modalités)
8 .	Evolution du niveau de vie de l'enquêté depuis 10 ans	(7 modalités)
9 .	Opinion sur le fonctionnement de la justice en 1979	(6 modalités)
10 .	La société française a-t-elle besoin de se transformer ?	(3 modalités)
13 .	La crèche est un mode de garde ...	(6 modalités)
14 .	La mère au foyer est un mode de garde ...	(6 modalités)
16 .	La vue sur l'extérieur vous plaît-elle ?	(5 modalités)
17 .	Possession ou usage d'une machine à laver la vaisselle	(2 modalités)
18 .	Possession ou usage d'une télévision couleur	(3 modalités)
21 .	Les dépenses de logement sont pour vous ...	(7 modalités)
22 .	Etes-vous gêné par les bruits ?	(4 modalités)
23 .	Participation à une action de défense de l'environnement	(2 modalités)
25 .	Votre travail présente-t-il des risques pour la santé ?	(4 modalités)
26 .	Avez-vous des conflits (travail et vie personnelle) ?	(3 modalités)
27 .	A souffert de nervosité ces quatre dernières semaines	(3 modalités)
28 .	A souffert d'état dépressif ces quatre dernières semaines	(3 modalités)
30 .	Possédez vous des biens immobiliers ?	(3 modalités)
31 .	Vous imposez-vous régulièrement des restrictions ?	(3 modalités)
32 .	Evolution du niveau de vie des français depuis 10 ans	(7 modalités)
33 .	Vous arrive-t-il d'inviter des amis à déjeuner ?	(3 modalités)
34 .	Faites-vous partie d'une association confessionnelle ?	(2 modalités)
35 .	Regardez-vous la télévision ...	(4 modalités)
36 .	Pour que la société change, faut-il ...	(4 modalités)
38 .	Appartenance à au moins une association	(2 modalités)
40 .	Age et sexe de l'enquêteur	(5 modalités)
41 .	Heure de coucher	(7 modalités)
43 .	L'enquêté(e) s'est-il (elle) montré(e) intéressé(e) ?	(4 modalités)
44 .	Nombre d'enfants considéré comme idéal	(5 modalités)
54 .	Profession de l'enquêté(e) en 7 classes	(8 modalités)
VARIABLES CONTINUES ILLUSTRATIVES - 8 VARIABLES		
15 .	Estimation du salaire mensuel d'un ingénieur	(continue)
19 .	Estimation du revenu mensuel d'un médecin	(continue)
37 .	Age de l'enquêté(e)	(continue)
42 .	Nombre de non-réponses au questionnaire	(continue)
45 .	Age de fin d'étude	(continue)
46 .	Revenu personnel souhaité	(continue)
47 .	Estimation du revenu minimum d'une famille de 2 enfants	(continue)
48 .	Nombre de jours de vacances en été	(continue)

LES PARAMETRES DE LA METHODE **ACM**

L'ONGLET « **VARIABLES** »



L'ONGLET « **PARAMETRES** »



① Ventilation des modalités actives d'effectif inférieur à (en %)

L'apurement permet de s'affranchir des modalités à faibles effectifs qui peuvent avec des effets perturbateurs sur l'analyse.

Les réponses appartenant à ces modalités peu fréquentes seront réparties aléatoirement entre les autres modalités de la variable.

Par défaut, sont ventilés les modalités actives dont l'effectif est inférieur à 2 %.

L'apurement vise à rendre plus robuste l'analyse. Les modalités ne sont pas abandonnées pour autant. Elles seront positionnées en éléments supplémentaires.

② Coordonnées conservées

Ce paramètre indique le nombre d'axes factoriels sur lesquels les coordonnées seront sauvegardées (10 par défaut). Seules les coordonnées sauvegardées sur ces axes factoriels pourront être édités et réutilisés pour les graphiques et pour la classification.

LES RESULTATS DE **ACM**

ANALYSE DES CORRESPONDANCES MULTIPLES

APUREMENT DES MODALITES ACTIVES

```
SEUIL (PCMIN)    :      5.00 %       POIDS:      50.00
AVANT APUREMENT :      7 QUESTIONS ACTIVES         28 MODALITES ASSOCIEES
APRES            :      7 QUESTIONS ACTIVES         27 MODALITES ASSOCIEES
POIDS TOTAL DES INDIVIDUS ACTIFS :   1000.00
```

TRI-A-PLAT DES QUESTIONS ACTIVES

MODALITES	AVANT APUREMENT		APRES APUREMENT		
IDENT LIBELLE	EFF.	POIDS	EFF.	POIDS	HISTOGRAMME DES POIDS RELATIFS
11 . Sexe de la personne interrogée					
masc - masculin	469	469.00	469	469.00	**********
fémi - féminin	531	531.00	531	531.00	**********
29 . Possédez vous des valeurs mobilières ?					
vmo1 - oui	121	121.00	121	121.00	********
vmo2 - non	879	879.00	879	879.00	**********
39 . Taille d'agglomération (en nombre d'habitants)					
agg1 - moins de 2.000	83	83.00	83	83.00	*****
agg2 - 2.000 - 20.000	87	87.00	87	87.00	******
agg3 - 20.000 - 100.000	175	175.00	175	175.00	**********
agg4 - plus de 100.000	329	329.00	329	329.00	**********
agg5 - Paris	326	326.00	326	326.00	**********
49 . Type d'emploi					
emp1 - Ouvriers	263	263.00	276	276.00	**********
emp2 - Employés	335	335.00	344	344.00	**********
emp3 - Cadres	229	229.00	241	241.00	**********
emp4 - Autres	48	48.00	=== VENTILEE ===		
49_ - *Reponse manquante*	125	125.00	139	139.00	*********
51 . Diplôme de l'enquêté(e) en 5 classes					
die1 - Aucun	189	189.00	189	189.00	**********
die2 - CEP ou fin études	321	321.00	321	321.00	**********
die3 - BEPC-BE-BEPS	158	158.00	158	158.00	**********
die4 - Bac - Brevet sup.	182	182.00	182	182.00	**********
die5 - Université,gde école	150	150.00	150	150.00	**********
52 . Statut d'occupation du logement en 4 classes					
slo1 - en accession	120	120.00	120	120.00	********
slo2 - propriétaire	290	290.00	290	290.00	**********
slo3 - locataire	523	523.00	523	523.00	**********
slo4 - logé gratuit, autre	67	67.00	67	67.00	*****
53 . Age de l'enquêté(e) en 5 classes					
agc1 - Moins de 25 ans	150	150.00	150	150.00	**********
agc2 - 25 à 34 ans	284	284.00	284	284.00	**********
agc3 - 35 à 49 ans	209	209.00	209	209.00	**********
agc4 - 50 à 64 ans	188	188.00	188	188.00	**********
agc5 - 65 ans et plus	169	169.00	169	169.00	**********

VALEURS PROPRES

```
APERCU DE LA PRECISION DES CALCULS : TRACE AVANT DIAGONALISATION ..   2.8571
                                     SOMME DES VALEURS PROPRES ....   2.8571
```

HISTOGRAMME DES 20 PREMIERES VALEURS PROPRES

NUM	VALEUR PROPRE	POURCENT	POURCENT CUMULE	
1	0.2703	9.46	9.46	**********
2	0.2369	8.29	17.75	*********
3	0.2084	7.29	25.05	********
4	0.1922	6.73	31.77	********
5	0.1846	6.46	38.23	*******
6	0.1578	5.52	43.76	*******
7	0.1534	5.37	49.13	*******
8	0.1493	5.23	54.35	*******
9	0.1441	5.04	59.40	*******
10	0.1398	4.89	64.29	*******
11	0.1326	4.64	68.93	*******
12	0.1300	4.55	73.48	*******
13	0.1284	4.49	77.97	*******
14	0.1222	4.28	82.25	*******
15	0.1070	3.74	86.00	******
16	0.1015	3.55	89.55	******
17	0.0954	3.34	92.89	******
18	0.0821	2.87	95.76	*****
19	0.0748	2.62	98.38	*****
20	0.0462	1.62	100.00	*****

RECHERCHE DE PALIERS (DIFFERENCES TROISIEMES)

PALIER ENTRE	VALEUR DU PALIER	
5 -- 6	-27.77	**********
14 -- 15	-10.42	********
17 -- 18	-6.67	******
13 -- 14	-5.44	*****
10 -- 11	-3.77	****
2 -- 3	-3.66	****
8 -- 9	-1.53	**

Palier différence 3ième entre 5 et 6 = [($\lambda_6 - \lambda_3$) – 3 * ($\lambda_5 - \lambda_4$)] * 1000

RECHERCHE DE PALIERS ENTRE (DIFFERENCES SECONDES)

PALIER ENTRE	VALEUR DU PALIER	
5 -- 6	22.31	**********
2 -- 3	12.28	******
14 -- 15	9.83	*****
3 -- 4	8.62	*****
1 -- 2	4.94	*****
10 -- 11	4.67	*****
11 -- 12	0.90	**
8 -- 9	0.81	**
6 -- 7	0.40	*

Palier différence 2nde entre 5 et 6 = [($\lambda_7 - \lambda_6$) – ($\lambda_6 - \lambda_5$)] * 1000

Les deux tableaux précédents sont deux versions de l'algorithme du coude (ou critère de Cattel). La procédure cherche à détecter un coude sur l'histogramme des valeurs propres pour faciliter le choix d'axes factoriels à utiliser.
L'algorithme sélectionne les coudes significatifs et les ordonne.

COORDONNEES, CONTRIBUTIONS ET COSINUS CARRES DES MODALITES ACTIVES AXES 1 A 5

MODALITES			COORDONNEES					CONTRIBUTIONS					COSINUS CARRES				
IDEN - LIBELLE	P.REL	DISTO	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
11 . Sexe de la personne interrogée																	
masc - masculin	6.70	1.13	-0.29	0.08	0.43	-0.47	-0.25	2.1	0.2	6.0	7.6	2.3	0.07	0.01	0.16	0.19	0.06
fémi - féminin	7.59	0.88	0.26	-0.07	-0.38	0.41	0.22	1.8	0.2	5.3	6.7	2.0	0.07	0.01	0.16	0.19	0.06
			CONTRIBUTION CUMULEE =					3.9	0.3	11.2	14.4	4.3					
29 . Possédez vous des valeurs mobilières ?																	
vmo1 - oui	1.73	7.26	0.69	1.46	-0.25	-0.23	0.06	3.1	15.5	0.5	0.5	0.0	0.07	0.29	0.01	0.01	0.00
vmo2 - non	12.56	0.14	-0.10	-0.20	0.03	0.03	-0.01	0.4	2.1	0.1	0.1	0.0	0.07	0.29	0.01	0.01	0.00
			CONTRIBUTION CUMULEE =					3.5	17.6	0.6	0.6	0.0					
39 . Taille d'agglomération (en nombre d'habitants)																	
agg1 - moins de 2.000	1.19	11.05	-1.06	0.83	-1.06	0.75	-0.06	5.0	3.4	6.4	3.5	0.0	0.10	0.06	0.10	0.05	0.00
agg2 - 2.000 - 20.000	1.24	10.49	-0.55	0.26	0.28	0.80	-0.61	1.4	0.3	0.5	4.2	2.5	0.03	0.01	0.01	0.06	0.04
agg3 - 20.000 - 100.000	2.50	4.71	-0.27	0.07	-0.17	0.07	-0.12	0.7	0.1	0.3	0.1	0.2	0.02	0.00	0.01	0.00	0.00
agg4 - plus de 100.000	4.70	2.04	-0.04	-0.40	0.05	-0.22	-0.27	0.0	3.2	0.0	1.2	1.9	0.00	0.08	0.00	0.02	0.04
agg5 - Paris	4.66	2.07	0.60	0.08	0.24	-0.22	0.52	6.2	0.1	1.3	1.2	6.7	0.18	0.00	0.03	0.02	0.13
			CONTRIBUTION CUMULEE =					13.3	7.1	8.5	10.1	11.3					
49 . Type d'emploi																	
emp1 - Ouvriers	3.94	2.62	-0.88	-0.47	0.54	-0.66	-0.20	11.2	3.6	5.6	8.9	0.8	0.29	0.08	0.11	0.17	0.01
emp2 - Employés	4.91	1.91	-0.19	-0.20	-0.38	0.67	0.63	0.6	0.8	3.5	11.4	10.5	0.02	0.02	0.08	0.23	0.21
emp3 - Cadres	3.44	3.15	0.80	0.89	0.74	0.02	-0.14	8.2	11.4	9.0	0.0	0.4	0.21	0.25	0.17	0.00	0.01
49_ - *Reponse manquante*	1.99	6.19	0.80	-0.12	-1.41	-0.38	-0.91	4.7	0.1	18.9	1.5	9.0	0.10	0.00	0.32	0.02	0.13
			CONTRIBUTION CUMULEE =					24.8	16.0	36.9	21.8	20.6					
51 . Diplôme de l'enquêté(e) en 5 classes																	
die1 - Aucun	2.70	4.29	-0.70	-0.23	-0.23	-0.93	0.34	5.0	0.6	0.7	12.1	1.7	0.12	0.01	0.01	0.20	0.03
die2 - CEP ou fin études	4.59	2.12	-0.80	0.08	0.05	0.29	-0.07	10.9	0.1	0.1	2.0	0.1	0.30	0.00	0.00	0.04	0.00
die3 - BEPC-BE-BEPS	2.26	5.33	0.23	-0.62	-0.17	0.47	0.56	0.4	3.7	0.3	2.6	3.8	0.01	0.07	0.01	0.04	0.06
die4 - Bac - Brevet sup.	2.60	4.49	0.93	-0.06	-0.32	0.26	-0.95	8.3	0.0	1.3	0.9	12.6	0.19	0.00	0.02	0.01	0.20
die5 - Université,gde école	2.14	5.67	1.23	0.84	0.73	-0.26	0.27	12.1	6.4	5.5	0.8	0.8	0.27	0.13	0.10	0.01	0.01
			CONTRIBUTION CUMULEE =					36.6	10.9	7.9	18.4	19.2					
52 . Statut d'occupation du logement en 4 classes																	
slo1 - en accession	1.71	7.33	-0.31	-0.06	0.85	1.02	-1.30	0.6	0.0	5.9	9.2	15.7	0.01	0.00	0.10	0.14	0.23
slo2 - propriétaire	4.14	2.45	-0.44	1.00	-0.51	-0.07	-0.01	3.0	17.6	5.2	0.1	0.0	0.08	0.41	0.11	0.00	0.00
slo3 - locataire	7.47	0.91	0.27	-0.51	0.15	-0.15	0.33	2.0	8.2	0.8	0.9	4.4	0.08	0.28	0.03	0.03	0.12
slo4 - logé gratuit, autre	0.96	13.93	0.34	-0.25	-0.50	-0.33	-0.20	0.4	0.3	1.2	0.6	0.2	0.01	0.00	0.02	0.01	0.00
			CONTRIBUTION CUMULEE =					6.0	26.0	13.0	10.8	20.3					
53 . Age de l'enquêté(e) en 5 classes																	
agc1 - Moins de 25 ans	2.14	5.67	0.81	-0.98	-0.89	-0.68	-0.80	5.2	8.7	8.2	5.2	7.4	0.12	0.17	0.14	0.08	0.11
agc2 - 25 à 34 ans	4.06	2.52	0.35	-0.45	0.63	0.47	0.41	1.9	3.4	7.8	4.8	3.7	0.05	0.08	0.16	0.09	0.07
agc3 - 35 à 49 ans	2.99	3.78	-0.33	0.36	0.41	0.41	-0.69	1.2	1.6	2.5	2.6	7.6	0.03	0.03	0.05	0.04	0.12
agc4 - 50 à 64 ans	2.69	4.32	-0.51	0.30	-0.42	0.21	0.25	2.6	1.0	2.3	0.6	0.9	0.06	0.02	0.04	0.01	0.01
agc5 - 65 ans et plus	2.41	4.92	-0.34	0.84	-0.32	-0.93	0.59	1.0	7.2	1.2	10.8	4.6	0.02	0.14	0.02	0.17	0.07
			CONTRIBUTION CUMULEE =					11.8	22.0	21.9	23.9	24.2					

P.REL : poids relatif de la modalité. Le poids relatif se calcule de la manière suivante : P.REL = (n_q * 100) / (n * Q) avec n_q l'effectif de la modalité, n l'effectif total et Q le nombre de variables actives.
Par exemple pour la modalité « masculin », P.REL = (469 * 100) / (1000 * 7) = 6.70 .

DISTO : distance à l'origine. Carré de la distance du khi² à l'origine. Cette distance ne dépend en fait que de l'effectif dans la modalité. La formule est la suivante :

$$d^2(j,G) = (n / n_j) - 1$$ avec n_j l'effectif de la modalité j et n l'effectif total

COORDONNEES ET VALEURS-TEST DES MODALITES ACTIVES - AXES 1 A 5

MODALITES			VALEURS-TEST					COORDONNEES					
IDEN - LIBELLE	EFF.	P.ABS	1	2	3	4	5	1	2	3	4	5	DISTO.
11 . Sexe de la personne interrogée													
masc - masculin	469	469.00	-8.6	2.3	12.8	-13.9	-7.5	-0.29	0.08	0.43	-0.47	-0.25	1.13
fémi - féminin	531	531.00	8.6	-2.3	-12.8	13.9	7.5	0.26	-0.07	-0.38	0.41	0.22	0.88
29 . Possédez vous des valeurs mobilières ?													
vmo1 - oui	121	121.00	8.1	17.1	-2.9	-2.7	0.7	0.69	1.46	-0.25	-0.23	0.06	7.26
vmo2 - non	879	879.00	-8.1	-17.1	2.9	2.7	-0.7	-0.10	-0.20	0.03	0.03	-0.01	0.14
39 . Taille d'agglomération (en nombre d'habitants)													
agg1 - moins de 2.000	83	83.00	-10.1	7.9	-10.1	7.2	-0.6	-1.06	0.83	-1.06	0.75	-0.06	11.05
agg2 - 2.000 - 20.000	87	87.00	-5.4	2.5	2.7	7.8	-5.9	-0.55	0.26	0.28	0.80	-0.61	10.49
agg3 - 20.000 - 100.000	175	175.00	-3.9	1.1	-2.4	1.0	-1.7	-0.27	0.07	-0.17	0.07	-0.12	4.71
agg4 - plus de 100.000	329	329.00	-0.9	-8.8	1.0	-4.8	-6.0	-0.04	-0.40	0.05	-0.22	-0.27	2.04
agg5 - Paris	326	326.00	13.2	1.8	5.2	-4.9	11.3	0.60	0.08	0.24	-0.22	0.52	2.07
49 . Type d'emploi													
emp1 - Ouvriers	263	263.00	-16.1	-9.7	10.7	-12.6	-3.5	-0.86	-0.51	0.57	-0.67	-0.18	2.80
emp2 - Employés	335	335.00	-3.6	-5.0	-8.5	15.2	14.2	-0.16	-0.22	-0.38	0.68	0.63	1.99
emp3 - Cadres	229	229.00	14.6	14.9	13.2	0.2	-2.1	0.85	0.86	0.77	0.01	-0.12	3.37
emp4 - Autres	48	48.00	-5.2	5.3	-3.5	-0.2	-3.3	-0.73	0.75	-0.50	-0.03	-0.47	19.83
49_ - *Reponse manquante*	125	125.00	11.4	-2.4	-16.6	-5.0	-10.9	0.96	-0.20	-1.39	-0.42	-0.91	7.00
51 . Diplôme de l'enquêté(e) en 5 classes													
die1 - Aucun	189	189.00	-10.8	-3.5	-3.5	-14.2	5.3	-0.70	-0.23	-0.23	-0.93	0.34	4.29
die2 - CEP ou fin études	321	321.00	-17.4	1.8	1.2	6.3	-1.5	-0.80	0.08	0.05	0.29	-0.07	2.12
die3 - BEPC-BE-BEPS	158	158.00	3.1	-8.5	-2.3	6.5	7.7	0.23	-0.62	-0.17	0.47	0.56	5.33
die4 - Bac - Brevet sup.	182	182.00	13.9	-0.9	-4.8	3.8	-14.1	0.93	-0.06	-0.32	0.26	-0.95	4.49
die5 - Université,gde école	150	150.00	16.4	11.2	9.7	-3.5	3.6	1.23	0.84	0.73	-0.26	0.27	5.67
52 . Statut d'occupation du logement en 4 classes													
slo1 - en accession	120	120.00	-3.6	-0.7	9.9	11.9	-15.2	-0.31	-0.06	0.85	1.02	-1.30	7.33
slo2 - propriétaire	290	290.00	-8.9	20.2	-10.3	-1.4	-0.2	-0.44	1.00	-0.51	-0.07	-0.01	2.45
slo3 - locataire	523	523.00	9.0	-16.9	5.1	-5.0	10.9	0.27	-0.51	0.15	-0.15	0.33	0.91
slo4 - logé gratuit, autre	67	67.00	2.8	-2.1	-4.3	-2.8	-1.7	0.34	-0.25	-0.50	-0.33	-0.20	13.93
53 . Age de l'enquêté(e) en 5 classes													
agc1 - Moins de 25 ans	150	150.00	10.7	-13.0	-11.8	-9.1	-10.6	0.81	-0.98	-0.89	-0.68	-0.80	5.67
agc2 - 25 à 34 ans	284	284.00	7.0	-8.9	12.6	9.5	8.1	0.35	-0.45	0.63	0.47	0.41	2.52
agc3 - 35 à 49 ans	209	209.00	-5.3	5.9	6.7	6.6	-11.2	-0.33	0.36	0.41	0.41	-0.69	3.78
agc4 - 50 à 64 ans	188	188.00	-7.7	4.6	-6.4	3.1	3.9	-0.51	0.30	-0.42	0.21	0.25	4.32
agc5 - 65 ans et plus	169	169.00	-4.8	12.0	-4.5	-13.2	8.4	-0.34	0.84	-0.32	-0.93	0.59	4.92

COORDONNEES ET VALEURS-TEST DES MODALITES ILLUSTRATIVES - AXES 1 A 5

MODALITES			VALEURS-TEST					COORDONNEES					
IDEN - LIBELLE	EFF.	P.ABS	1	2	3	4	5	1	2	3	4	5	DISTO.
1 . La famille est le seul endroit oú l'on se sente bien													
fbi1 - oui	561	561.00	-14.5	4.4	-3.6	0.6	0.5	-0.40	0.12	-0.10	0.02	0.02	0.78
fbi2 - non	431	431.00	14.6	-4.5	3.6	-0.4	-0.8	0.53	-0.16	0.13	-0.02	-0.03	1.32
1_ - *Reponse manquante*	8	8.00	-0.3	0.6	-0.4	-1.0	1.5	-0.11	0.20	-0.13	-0.34	0.54	124.00
2 . Opinion à propos du mariage													
opm1 - union indissoluble	231	231.00	-7.9	4.1	-3.3	-3.2	0.3	-0.46	0.23	-0.19	-0.19	0.02	3.33
opm2 - dissout si pb. grave	342	342.00	-1.8	3.4	-1.8	3.2	-0.7	-0.08	0.15	-0.08	0.14	-0.03	1.92
opm3 - dissout si accord	387	387.00	8.7	-6.4	4.7	-0.2	0.2	0.35	-0.25	0.19	-0.01	0.01	1.58
opm4 - ne sait pas	39	39.00	-0.3	-1.3	0.0	-0.4	0.5	-0.05	-0.21	0.00	-0.06	0.08	24.64
2_ - *Reponse manquante*	1	1.00	0.8	0.8	-0.8	0.1	-0.4	0.79	0.77	-0.81	0.09	-0.42	999.00
3 . A qui incombent les travaux ménagers et les soins enfants ?													
esc1 - incombent à la femme	42	42.00	-3.5	-0.6	-0.9	-0.9	-0.4	-0.52	-0.08	-0.14	-0.14	-0.06	22.81
esc2 - plutôt à la femme	336	336.00	-2.4	4.9	-1.4	-2.3	2.0	-0.11	0.22	-0.06	-0.10	0.09	1.98
esc3 - homme et femme	599	599.00	3.6	-4.3	2.1	2.9	-2.1	0.09	-0.11	0.05	0.07	-0.05	0.67
esc4 - ne sait pas	19	19.00	0.7	-0.3	-2.1	-0.7	0.1	0.15	-0.07	-0.47	-0.15	0.02	51.63
3_ - *Reponse manquante*	4	4.00	0.2	-1.3	1.2	-0.9	1.9	0.11	-0.64	0.62	-0.43	0.93	249.00
4 . Opinion sur le cadre de vie quotidien													
cvi1 - très satisfait	259	259.00	-0.8	5.3	-3.4	1.7	-0.9	-0.04	0.28	-0.18	0.09	-0.05	2.86
cvi2 - satisfait	549	549.00	-0.9	0.1	1.2	0.1	0.2	-0.03	0.00	0.03	0.00	0.00	0.82
cvi3 - peu satisfait	145	145.00	1.9	-4.8	1.3	-1.3	1.1	0.14	-0.37	0.10	-0.10	0.08	5.90
cvi4 - pas du tout satisf.	46	46.00	0.6	-3.3	2.0	-1.6	-0.3	0.08	-0.47	0.29	-0.23	-0.04	20.74
4_ - *Reponse manquante*	1	1.00	0.4	1.5	0.7	-0.6	-0.1	0.35	1.52	0.72	-0.56	-0.12	999.00
5 . La préservation de l'environnement est une chose ...													
env1 - très importante	657	657.00	8.0	0.0	1.8	0.8	-1.4	0.18	0.00	0.04	0.02	-0.03	0.52
env2 - assez importante	298	298.00	-7.1	-0.1	-0.7	0.3	0.6	-0.34	0.00	-0.04	0.02	0.03	2.36
env3 - peu importante	36	36.00	-3.0	-0.1	-2.8	-1.7	2.4	-0.49	-0.01	-0.46	-0.27	0.39	26.78
env4 - pas du tout import.	7	7.00	-0.4	0.1	-0.1	-2.5	-0.3	-0.16	0.05	-0.02	-0.94	-0.13	141.86
5_ - *Reponse manquante*	2	2.00	0.3	0.8	-0.1	-0.3	-0.5	0.20	0.59	-0.10	-0.24	-0.37	499.00
6 . Les découvertes scientifiques améliorent-elles la vie ?													
sci1 - oui, un peu	509	509.00	-1.9	-0.2	0.3	0.5	-0.6	-0.06	0.00	0.01	0.02	-0.02	0.96
sci2 - oui, beaucoup	383	383.00	3.1	1.8	-1.2	0.8	-0.3	0.12	0.07	-0.05	0.03	-0.01	1.61
sci3 - pas du tout	105	105.00	-1.6	-2.3	1.3	-2.0	1.6	-0.15	-0.22	0.12	-0.19	0.15	8.52
6_ - *Reponse manquante*	3	3.00	-1.1	-1.5	0.1	-0.8	-0.6	-0.65	-0.89	0.07	-0.49	-0.36	332.33

MODALITES			VALEURS-TEST					COORDONNEES					
IDEN - LIBELLE	EFF.	P.ABS	1	2	3	4	5	1	2	3	4	5	DISTO.
7 . Comparée aux personnes de votre âge, votre santé est ...													
san1 - très satisfaisante	267	267.00	3.8	0.3	0.4	-1.1	-2.1	0.20	0.02	0.02	-0.06	-0.11	2.75
san2 - satisfaisante	600	600.00	-2.7	0.4	0.4	2.0	2.3	-0.07	0.01	0.01	0.05	0.06	0.67
san3 - peu satisfaisante	115	115.00	-0.6	-0.8	-1.1	-1.3	-1.2	-0.05	-0.07	-0.09	-0.12	-0.10	7.70
san4 - pas du tout satisf.	18	18.00	-1.1	-0.5	-0.2	-0.3	1.3	-0.25	-0.11	-0.05	-0.06	0.30	54.56
8 . Evolution du niveau de vie de l'enquêté depuis 10 ans													
niv1 - beaucoup mieux	102	102.00	1.7	0.9	0.5	1.8	-0.8	0.16	0.08	0.04	0.17	-0.08	8.80
niv2 - un peu mieux	316	316.00	-1.2	-1.5	1.8	4.2	-0.9	-0.05	-0.07	0.08	0.20	-0.04	2.16
niv3 - c'est pareil	250	250.00	0.8	2.3	-2.6	-3.0	-2.1	0.05	0.12	-0.14	-0.16	-0.11	3.00
niv4 - un peu moins bien	190	190.00	-2.2	0.3	0.8	-2.1	3.7	-0.14	0.02	0.05	-0.14	0.24	4.26
niv5 - beaucoup moins bien	114	114.00	0.3	-0.1	1.3	-0.1	1.6	0.03	-0.01	0.12	-0.01	0.14	7.77
niv6 - ne sait pas	26	26.00	2.9	-4.0	-3.2	-1.9	-2.3	0.55	-0.78	-0.61	-0.36	-0.45	37.46
8_ - *Reponse manquante*	2	2.00	-0.7	-0.3	-1.2	-1.8	-1.0	-0.47	-0.23	-0.83	-1.30	-0.73	499.00
9 . Opinion sur le fonctionnement de la justice en 1979													
jus1 - très bon	13	13.00	0.0	1.7	-2.2	-2.1	0.2	0.01	0.47	-0.60	-0.57	0.06	75.92
jus2 - assez bon	243	243.00	-0.8	3.4	-0.1	-0.2	-0.8	-0.05	0.19	-0.01	-0.01	-0.04	3.12
jus3 - assez mauvais	398	398.00	0.6	-1.0	-1.7	1.2	-1.8	0.02	-0.04	-0.06	0.05	-0.07	1.51
jus4 - très mauvais	256	256.00	1.3	-2.9	3.9	-1.3	1.1	0.07	-0.16	0.21	-0.07	0.06	2.91
jus5 - ne sait pas	65	65.00	-3.3	0.5	-2.1	0.0	0.6	-0.40	0.05	-0.26	0.00	0.07	14.38
jus6 - ne veut pas répondre	25	25.00	2.2	-0.1	-0.4	1.9	3.4	0.43	-0.02	-0.09	0.37	0.68	39.00
10 . La société francaise a-t-elle besoin de se transformer ?													
tso1 - oui	759	759.00	1.8	-4.8	3.1	-0.3	-0.4	0.03	-0.08	0.05	-0.01	-0.01	0.32
tso2 - non	170	170.00	-0.6	4.4	-2.3	0.9	-0.6	-0.04	0.31	-0.16	0.06	-0.04	4.88
tso3 - ne sait pas	71	71.00	-2.1	1.5	-1.7	-0.8	1.7	-0.24	0.17	-0.20	-0.09	0.19	13.08
13 . La crèche est un mode de garde ...													
cre1 - très satisfaisant	139	139.00	1.6	-3.6	1.5	1.8	2.3	0.13	-0.28	0.12	0.14	0.18	6.19
cre2 - assez satisfaisant	386	386.00	1.8	2.8	1.8	0.7	-0.7	0.07	0.11	0.07	0.03	-0.03	1.59
cre3 - peu satisfaisant	242	242.00	1.6	0.4	-0.4	-1.3	-1.5	0.09	0.02	-0.02	-0.08	-0.08	3.13
cre4 - pas du tout satisf.	92	92.00	-0.9	-1.8	-0.8	-1.1	0.8	-0.09	-0.18	-0.08	-0.11	0.08	9.87
cre5 - ne sait pas	139	139.00	-5.8	0.6	-2.7	-0.1	0.0	-0.45	0.05	-0.21	-0.01	0.00	6.19
13_ - *Reponse manquante*	2	2.00	2.0	0.5	-1.6	-0.9	-1.4	1.40	0.37	-1.11	-0.66	-0.98	499.00
14 . La mère au foyer est un mode de garde ...													
mèr1 - très satisfaisant	786	786.00	-6.8	2.9	-4.5	3.5	-0.3	-0.11	0.05	-0.07	0.06	-0.01	0.27
mèr2 - assez satisfaisant	129	129.00	6.0	-1.7	1.9	-1.8	-0.7	0.50	-0.14	0.16	-0.14	-0.06	6.75
mèr3 - peu satisfaisant	35	35.00	2.7	-1.5	2.8	-1.5	1.3	0.45	-0.25	0.47	-0.25	0.21	27.57
mèr4 - pas du tout satisf.	20	20.00	2.8	-1.0	2.4	-1.3	1.0	0.63	-0.22	0.53	-0.29	0.21	49.00
mèr5 - ne sait pas	29	29.00	-1.0	-1.5	2.1	-2.3	0.2	-0.19	-0.27	0.38	-0.43	0.03	33.48
14_ - *Reponse manquante*	1	1.00	0.8	0.8	-0.8	0.1	-0.4	0.79	0.77	-0.81	0.09	-0.42	999.00
16 . La vue sur l'extérieur vous plaît-elle ?													
vue1 - beaucoup	516	516.00	-4.8	4.7	-4.1	2.2	0.0	-0.15	0.14	-0.13	0.07	0.00	0.94
vue2 - moyennement	296	296.00	3.2	-0.4	3.7	-0.3	-0.5	0.16	-0.02	0.18	-0.01	-0.02	2.38
vue3 - pas beaucoup	82	82.00	1.3	-2.6	1.8	-1.0	1.5	0.14	-0.27	0.19	-0.10	0.16	11.20
vue4 - pas du tout	104	104.00	1.7	-4.7	-0.4	-2.0	-0.5	0.15	-0.44	-0.03	-0.19	-0.05	8.62
16_ - *Reponse manquante*	2	2.00	1.0	0.3	0.1	-2.4	0.1	0.69	0.23	0.05	-1.68	0.05	499.00
17 . Possession ou usage d'une machine à laver la vaisselle													
lav1 - oui	211	211.00	4.6	7.4	1.0	2.9	-6.0	0.28	0.45	0.06	0.18	-0.37	3.74
lav2 - non	789	789.00	-4.6	-7.4	-1.0	-2.9	6.0	-0.07	-0.12	-0.02	-0.05	0.10	0.27
18 . Possession ou usage d'une télévision couleur													
tco1 - oui	373	373.00	-2.5	3.8	-0.6	0.4	0.2	-0.10	0.16	-0.02	0.02	0.01	1.68
tco2 - non	624	624.00	2.6	-3.7	0.5	-0.4	-0.4	0.06	-0.09	0.01	-0.01	-0.01	0.60
18_ - *Reponse manquante*	3	3.00	-1.0	-0.3	0.8	0.1	1.0	-0.59	-0.17	0.45	0.08	0.59	332.33
21 . Les dépenses de logement sont pour vous ...													
dlo1 - négligeables	113	113.00	0.1	2.8	-4.0	-2.1	0.9	0.01	0.24	-0.36	-0.19	0.08	7.85
dlo2 - pas de gros problème	444	444.00	-2.0	2.6	1.9	-0.4	-1.6	-0.07	0.09	0.07	-0.01	-0.06	1.25
dlo3 - une lourde charge	352	352.00	1.1	-2.9	1.9	2.8	1.9	0.05	-0.12	0.08	0.12	0.08	1.84
dlo4 - très lourde charge	55	55.00	0.2	-3.0	1.1	0.1	0.6	0.03	-0.39	0.14	0.01	0.07	17.18
dlo5 - ne peut pas payer	6	6.00	-0.2	1.2	0.8	-0.8	-0.1	-0.10	0.47	0.32	-0.32	-0.03	165.67
dlo6 - ne sait pas	22	22.00	2.0	-1.2	-4.1	-1.6	-2.3	0.42	-0.25	-0.86	-0.34	-0.48	44.45
21_ - *Reponse manquante*	8	8.00	0.9	-0.1	-3.2	-1.9	-1.8	0.33	-0.05	-1.13	-0.66	-0.62	124.00

CORRELATIONS ENTRE LES VARIABLES CONTINUES ET LES FACTEURS - AXES 1 A 5

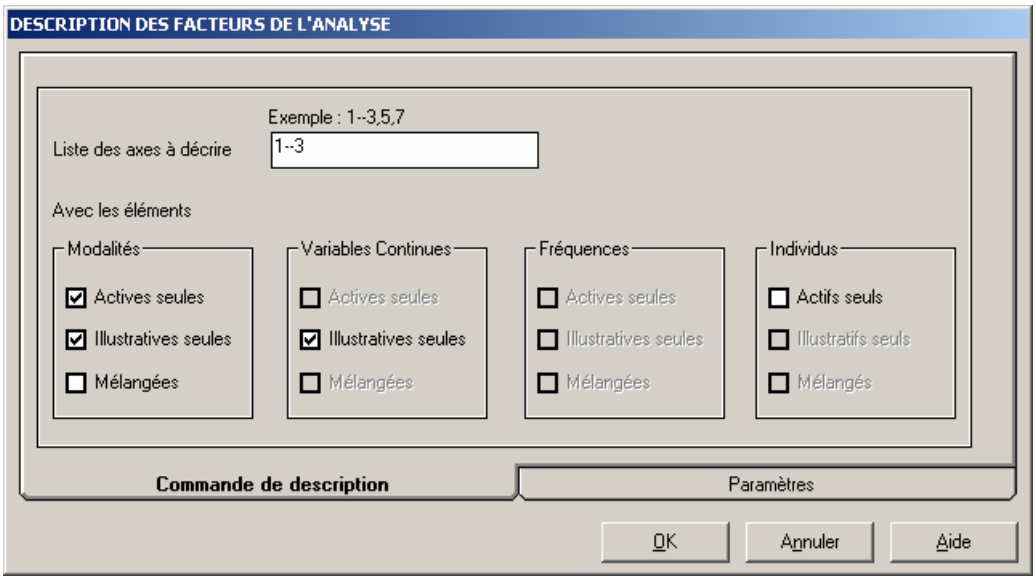
VARIABLES		CARACTERISTIQUES			CORRELATIONS				
NUM . (IDEN) LIBELLE COURT	EFF.	P.ABS	MOYENNE	EC.TYPE	1	2	3	4	5
15 . (ring) Estimation du salair	806	806.00	8478.73	3668.95	-0.04	0.06	0.04	-0.01	0.05
19 . (rmed) Estimation du revenu	713	713.00	19383.85	12608.83	-0.05	0.12	0.02	0.03	-0.06
37 . (âge) Age de l'enquêté(e)	1000	1000.00	42.68	17.50	-0.40	0.55	-0.14	-0.21	0.28
42 . (nrep) Nombre de non-répons	1000	1000.00	4.05	4.19	-0.20	0.12	-0.20	-0.08	0.12
45 . (finé) Age de fin d'étude	997	997.00	17.29	3.88	0.69	0.13	0.24	0.05	-0.11
46 . (rsou) Revenu personnel sou	915	915.00	7244.48	4756.78	0.26	0.21	0.15	0.03	-0.09
47 . (rmin) Estimation du revenu	897	897.00	5561.89	2423.40	0.19	-0.01	0.14	-0.08	0.14
48 . (vaca) Nombre de jours de v	1000	1000.00	18.31	19.37	0.38	0.02	0.03	-0.06	-0.07

DEFAC - Description des axes factoriels

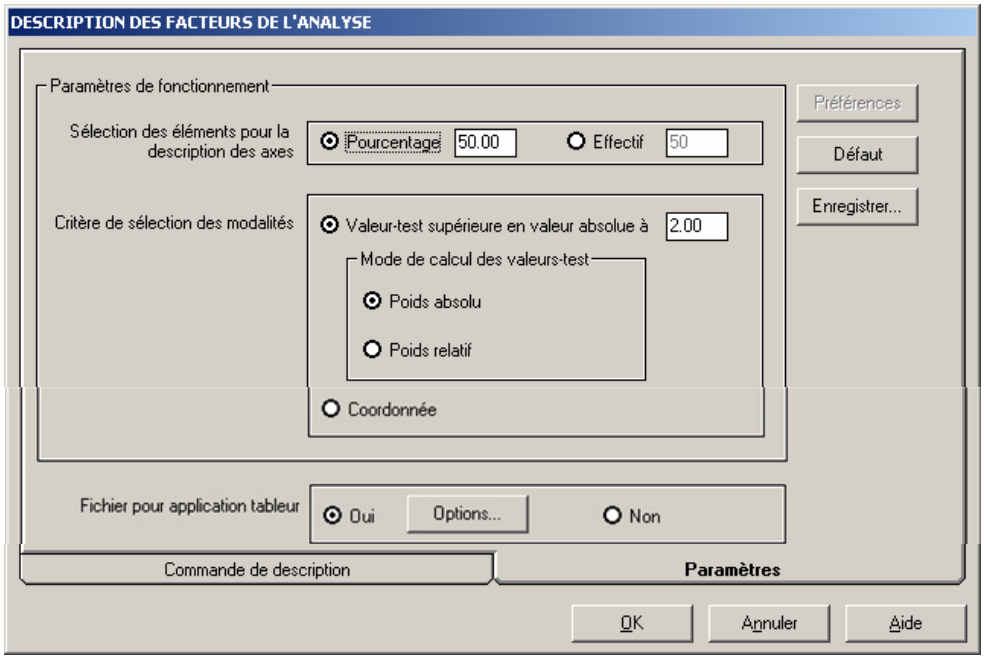
Cette procédure constitue une aide à l'interprétation des facteurs issus d'une procédure d'analyse factorielle. Les items statistiquement caractéristiques sont sélectionnés et rangés en fonction du critère de la valeur-test.

LES PARAMETRES DE LA METHODE **DEFAC**

L'ONGLET « **COMMANDE DE DESCRIPTION** »



L'ONGLET « **PARAMETRES** »



LES RESULTATS DE **DEFAC**

DESCRIPTION DES AXES FACTORIELS

DESCRIPTION DU FACTEUR 1 - PAR LES MODALITES ACTIVES

ID.	V.TEST	LIBELLE MODALITE	LIBELLE DE LA VARIABLE	POIDS	NUMERO
die2	-17.40	CEP ou fin études	Diplôme de l'enquêté(e) en 5 classes	321.00	1
emp1	-16.14	Ouvriers	Type d'emploi	263.00	2
die1	-10.75	Aucun	Diplôme de l'enquêté(e) en 5 classes	189.00	3
agg1	-10.11	moins de 2.000	Taille d'agglomération (en nombre d'habitants)	83.00	4
slo2	-8.88	propriétaire	Statut d'occupation du logement en 4 classes	290.00	5
masc	-8.62	masculin	Sexe de la personne interrogée	469.00	6
vmo2	-8.14	non	Possédez vous des valeurs mobilières ?	879.00	7
		Z O N E C E N T R A L E			
slo3	8.98	locataire	Statut d'occupation du logement en 4 classes	523.00	22
agc1	10.73	Moins de 25 ans	Age de l'enquêté(e) en 5 classes	150.00	23
49_	11.44	*Reponse manquante*	Type d'emploi	125.00	24
agg5	13.22	Paris	Taille d'agglomération (en nombre d'habitants)	326.00	25
die4	13.86	Bac - Brevet sup.	Diplôme de l'enquêté(e) en 5 classes	182.00	26
emp3	14.64	Cadres	Type d'emploi	229.00	27
die5	16.38	Université,gde école	Diplôme de l'enquêté(e) en 5 classes	150.00	28

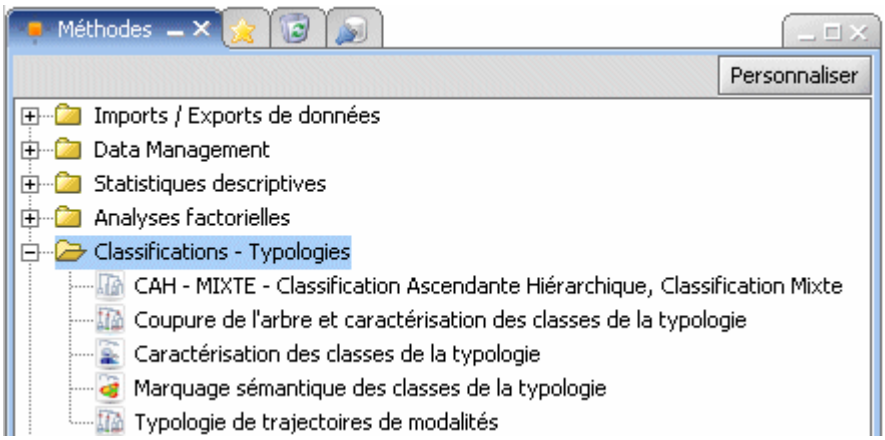
DESCRIPTION DU FACTEUR 2 - PAR LES MODALITES ACTIVES

ID.	V.TEST	LIBELLE MODALITE	LIBELLE DE LA VARIABLE	POIDS	NUMERO
vmo2	-17.08	non	Possédez vous des valeurs mobilières ?	879.00	1
slo3	-16.86	locataire	Statut d'occupation du logement en 4 classes	523.00	2
agc1	-13.04	Moins de 25 ans	Age de l'enquêté(e) en 5 classes	150.00	3
emp1	-9.65	Ouvriers	Type d'emploi	263.00	4
agc2	-8.90	25 à 34 ans	Age de l'enquêté(e) en 5 classes	284.00	5
agg4	-8.83	plus de 100.000	Taille d'agglomération (en nombre d'habitants)	329.00	6
die3	-8.54	BEPC-BE-BEPS	Diplôme de l'enquêté(e) en 5 classes	158.00	7
		Z O N E C E N T R A L E			
agc3	5.86	35 à 49 ans	Age de l'enquêté(e) en 5 classes	209.00	22
agg1	7.86	moins de 2.000	Taille d'agglomération (en nombre d'habitants)	83.00	23
die5	11.21	Université,gde école	Diplôme de l'enquêté(e) en 5 classes	150.00	24
agc5	11.97	65 ans et plus	Age de l'enquêté(e) en 5 classes	169.00	25
emp3	14.87	Cadres	Type d'emploi	229.00	26
vmo1	17.08	oui	Possédez vous des valeurs mobilières ?	121.00	27
slo2	20.24	propriétaire	Statut d'occupation du logement en 4 classes	290.00	28

DESCRIPTION DU FACTEUR 3 - PAR LES MODALITES ACTIVES

ID.	V.TEST	LIBELLE MODALITE	LIBELLE DE LA VARIABLE	POIDS	NUMERO
49_	-16.60	*Reponse manquante*	Type d'emploi	125.00	1
fémi	-12.80	féminin	Sexe de la personne interrogée	531.00	2
agc1	-11.84	Moins de 25 ans	Age de l'enquêté(e) en 5 classes	150.00	3
slo2	-10.29	propriétaire	Statut d'occupation du logement en 4 classes	290.00	4
agg1	-10.05	moins de 2.000	Taille d'agglomération (en nombre d'habitants)	83.00	5
emp2	-8.50	Employés	Type d'emploi	335.00	6
agc4	-6.40	50 à 64 ans	Age de l'enquêté(e) en 5 classes	188.00	7
		Z O N E C E N T R A L E			
agc3	6.74	35 à 49 ans	Age de l'enquêté(e) en 5 classes	209.00	22
die5	9.75	Université,gde école	Diplôme de l'enquêté(e) en 5 classes	150.00	23
slo1	9.87	en accession	Statut d'occupation du logement en 4 classes	120.00	24
emp1	10.73	Ouvriers	Type d'emploi	263.00	25
agc2	12.62	25 à 34 ans	Age de l'enquêté(e) en 5 classes	284.00	26
masc	12.80	masculin	Sexe de la personne interrogée	469.00	27
emp3	13.18	Cadres	Type d'emploi	229.00	28

LA CLASSIFICATION AVEC SPAD



CAH – MIXTE : Classification sur facteurs

PARTI - DECLA : Coupure de l'arbre et description des classes

CLASS - MINER : Caractérisation des classes des typologies

ESCAL : Archivages axes factoriels et partitions

Les techniques de classification ont pour but d'expliciter la structure d'un ensemble de données importantes, permettant ainsi de formuler des hypothèses à vérifier dans une étape ultérieure. Elles sont à distinguer des méthodes de classement qui ont un but explicatif ou prédictif.

CAH / MIXTE - Classification sur Facteurs

JUSTIFICATION DU PASSAGE AUX COORDONNEES FACTORIELLES

La méthode CAH/MIXTE (nommée RECIP/SEMIS dans les éditions de résultats) de SPAD permet d'effectuer une classification à partir de coordonnées factorielles issues d'une analyse préalable.

Il est équivalent d'effectuer une classification des individus à partir d'un ensemble de p variables ou à partir de l'ensemble des p facteurs issus de l'analyse factorielle. En effet, en passant des variables initiales aux facteurs, sans en réduire leur nombre et ce, malgré leur obtention dans l'ordre décroissant de la variance expliquée, on ne perd aucune information. Il s'agit mathématiquement d'un changement de repère des individus (changement de base).

On peut, néanmoins, ne prendre en compte qu'un sous-espace factoriel de dimension q avec q inférieur à p et effectuer une classification sur les q premiers axes factoriels. Cela présente l'avantage d'éliminer des fluctuations aléatoires qui constituent en général l'essentiel de la variance prise en compte par les (p-q) derniers axes.

Le fait d'abandonner les derniers facteurs revient à « lisser » les données, ce qui en général améliore la partition en produisant des classes plus homogènes.

Les axes factoriels qui sont à conserver pour la classification sont ceux qui engendrent un sous-espace dans lequel le nuage des individus à classer est stable. En pratique, on garde généralement un peu plus de la moitié des axes, même si un « coude » apparaît au bout de quelques axes à l'examen de l'histogramme des valeurs propres associées à ces axes.

Dans le paramétrage de cette méthode, vous pouvez définir le nombre de coordonnées factorielles à prendre en compte pour l'agrégation (10 par défaut).

Ainsi, quelque soit le tableau de données initial, on se ramènera toujours à un tableau de données quantitatives à partir duquel sera effectué la classification des individus. Une seule distance, *la distance euclidienne usuelle*, sera utilisée pour calculer les ressemblances entre individus et un seul critère d'agrégation, *la perte d'inertie minimum* (critère de Ward) sera utilisé pour calculer l'écart entre deux sous-ensembles disjoints.

LES TECHNIQUES DE CLASSIFICATION

Les techniques proposées dans SPAD sont la classification ascendante hiérarchique (CAH, RECIP dans SPAD) qui fournit une hiérarchie de partitions et la méthode d'agrégation autour de centres mobiles qui conduit directement à une seule partition.

Une utilisation conjointe de ces deux types de méthodes (classification mixte) permettra de consolider la partition et d'obtenir une partition fiable sinon optimale (SEMIS).

Les deux types de méthode – CAH et centres mobiles – présentent les inconvénients respectifs suivants :

- la CAH fournit un grand nombre de partitions parmi lesquelles on doit en choisir une : il n'est souvent pas aisé de choisir la coupure significative. D'autre part, l'arbre hiérarchique obtenu n'est pas un arbre optimal puisque la partition construite à un niveau donné dépend de la partition obtenue à l'étape précédente.
- dans la méthode des centres mobiles, le nombre de classes doit être fixé au départ, et la partition obtenue dépend du tirage initial des centres provisoires des classes.

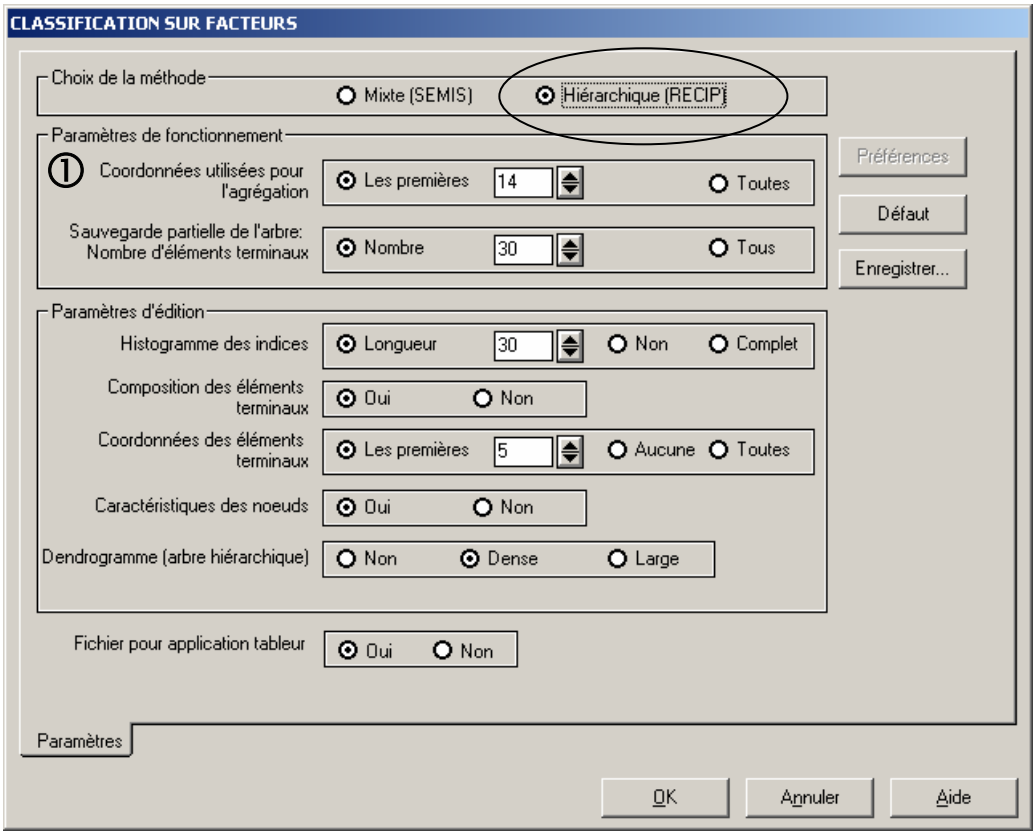
Pour remédier en partie à ces inconvénients et pour essayer de s'approcher le plus possible de la partition optimale si elle existe, on peut avoir recours à l'utilisation conjointe de la CAH et de la CCM : c'est l'objet de la classification mixte appelée SEMIS dans SPAD.

Une première utilisation conjointe des deux techniques de classification est la suivante : on effectue une classification (CCM) autour d'un nombre important de centres mobiles et on construit ensuite un arbre hiérarchique à partir des classes formées dans cette CCM.

Cependant, cette méthode est relativement instable sur des échantillons de petite taille. Nous vous conseillons d'utiliser la procédure RECIP (CAH) sur des échantillons de moins de 20 000 individus. Au delà, la méthode SEMIS permet de réduire les temps d'exécution et fournit des partitions stables.

LES PARAMETRES DE LA METHODE **RECIP / SEMIS**

La méthode « Hiérarchique » (RECIP)



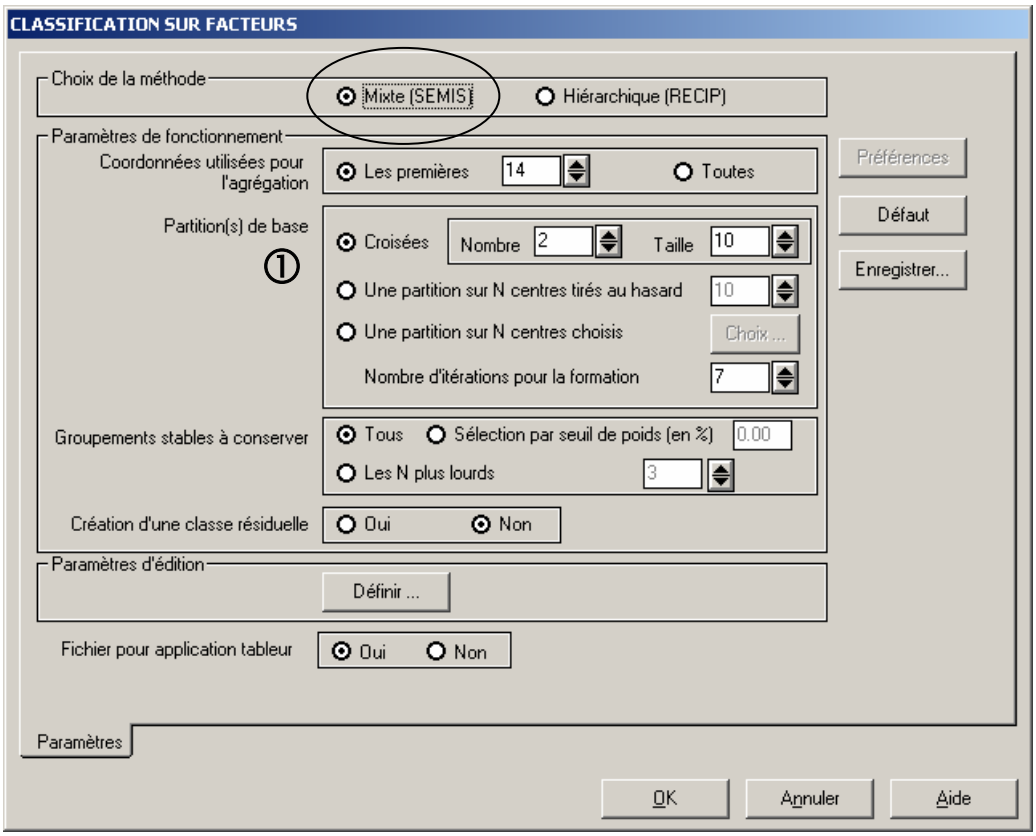
① Coordonnées utilisées pour l'agrégation

Ce paramètre indique le nombre de facteurs pris en compte pour calculer les distances entre les individus et pour effectuer la partition en classes.

C'est au moment du choix de ce paramètre que l'utilisateur devra se référer aux résultats de l'analyse factorielle.

Il n'y a pas de règle simple pour le choix du nombre d'axes. On peut conseiller en général de conserver au moins la moitié des axes et souvent les deux tiers.

La méthode « Mixte » (SEMIS)



① Partition(s) de base

Trois méthodes de classification sont disponibles.

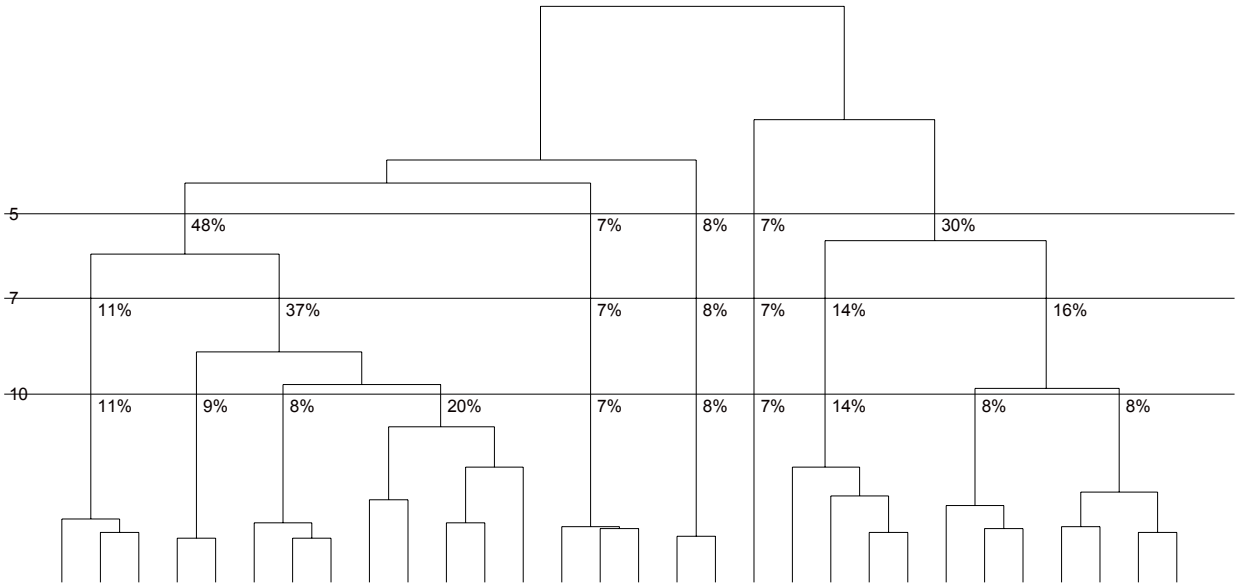
- ✓ La première consiste à chercher les classes stables par croisement de plusieurs partitions de base construites autour d'individus tirés au hasard.
  L'item « Nombre » définit le nombre de partitions construites (2 par défaut) et l'item « Taille » détermine le nombre d'individus tirés au hasard pour chaque partition. Ce sont les centres initiaux de chacune des partitions.
- ✓ Les deux autres consistent à construire une seul partition par l'algorithme des centres mobiles autour de N centres choisis par l'utilisateur ou tirés au hasard dans l'ensemble de la population.

L'EDITEUR DE GRAPHIQUES HIERARCHIQUES

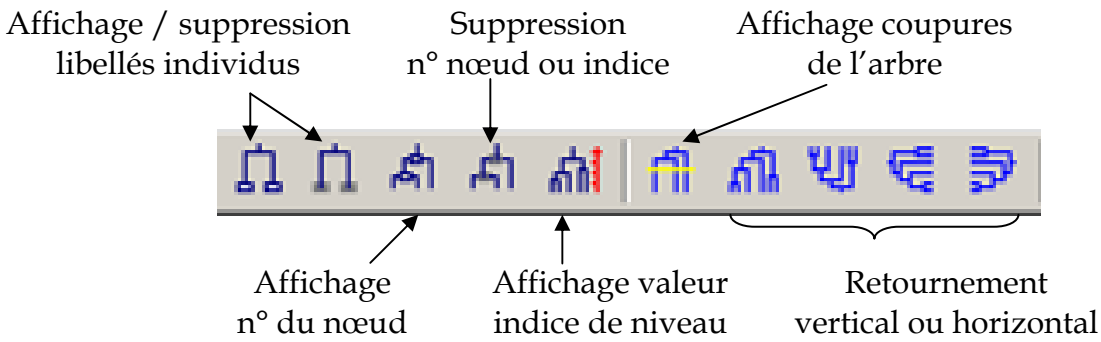
Pour accéder à l'**Editeur hiérarchique**, double cliquez sur l'icône .

LE « DENDROGRAMME »

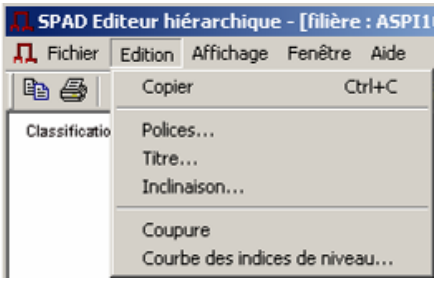
Le dendrogramme est la représentation graphique de la hiérarchie des partitions.
Un intérêt du dendrogramme est de suggérer le nombre de classes qui existent dans l'échantillon. On peut couper l'arbre là où le palier est le plus grand. Le nombre de classes sera égal au nombre de branches coupées.



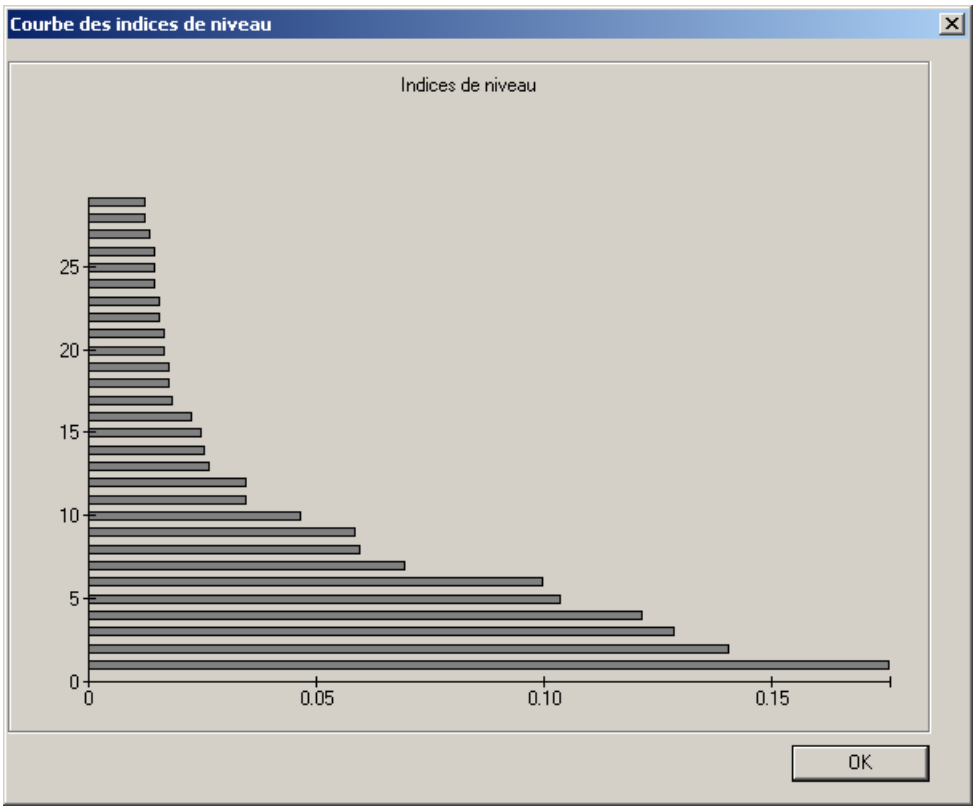
LA BARRE D'OUTILS DE L'EDITEUR HIERARCHIQUE



L'HISTOGRAMME DES INDICES DE NIVEAU



« **Edition** » - « **Courbe des indices de niveau** »



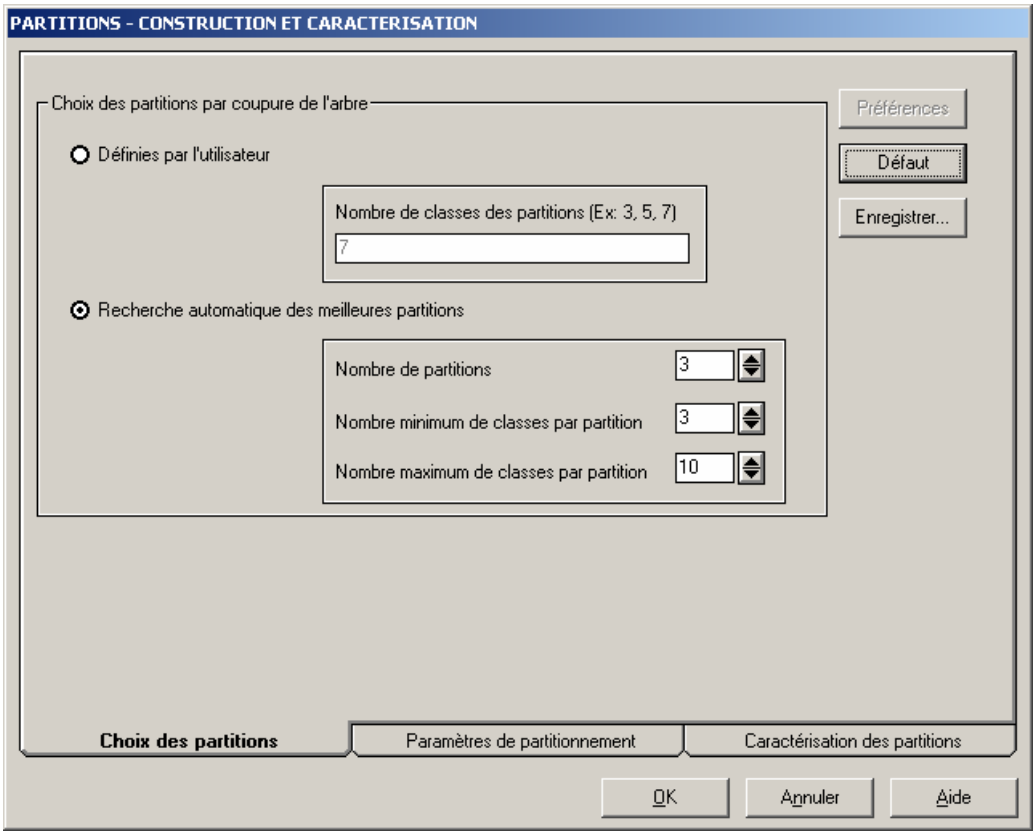
L'indice de niveau est la perte d'inerte inter-classes occasionnée par la formation du nœud.
La barre la plus longue, en bas de l'histogramme, correspond à une coupure de l'arbre en deux classes. En coupant par exemple au niveau de la seconde barre la plus longue de l'histogramme, on génère une partition en trois classes.

PARTI - DECLA - Coupure de l'arbre et description

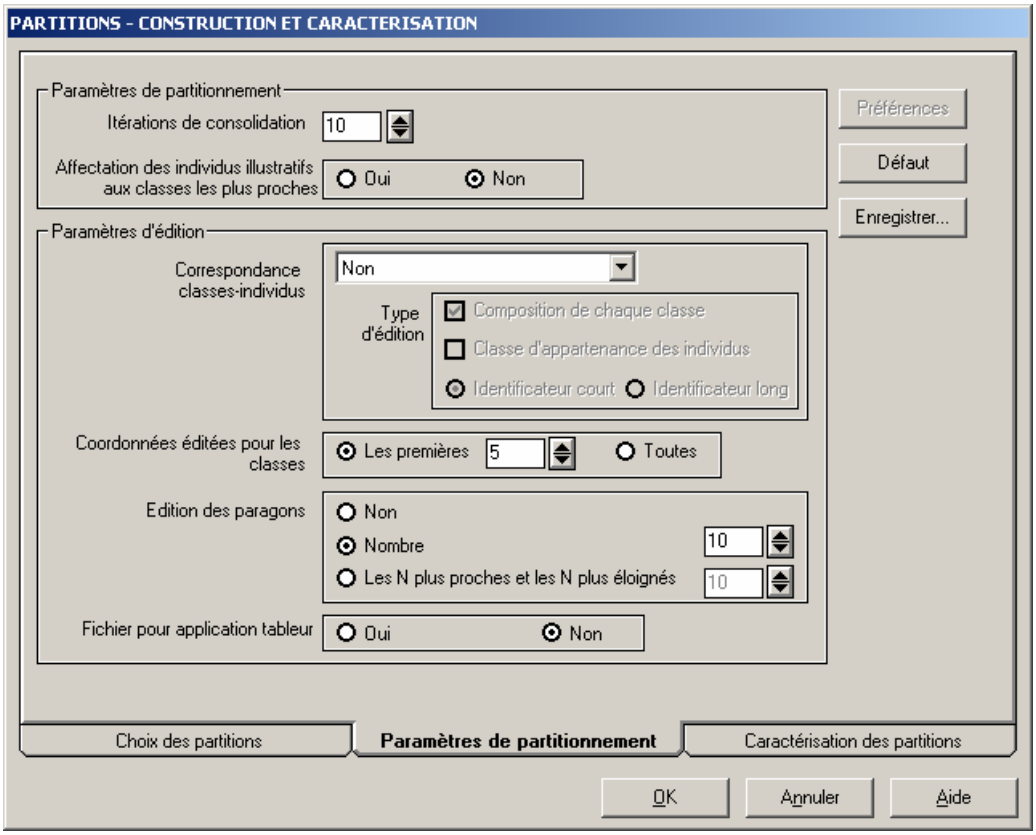
La procédure permet de couper à un niveau convenable l'arbre obtenu par la procédure RECIP ou SEMIS pour construire une partition des individus. Il est possible de construire plusieurs partitions simultanément et de décrire statistiquement les partitions choisies.

LES PARAMETRES DE LA METHODE **PARTI-DECLA**

L'ONGLET « **CHOIX DES PARTITIONS** »



L'ONGLET « PARAMETRES DE PARTITIONNEMENT »



L'ONGLET « CARACTERISATION DES PARTITIONS »

Onglet identique à l'onglet « **Paramètres** » de la méthode **DEMOD**.

LES RESULTATS DE PARTI-DECLA

PARTITION PAR COUPURE D'UN ARBRE HIERARCHIQUE

RECHERCHE DES MEILLEURES PARTITIONS
RECHERCHE DES PALIERS

Ce tableau vise à aider l'utilisateur dans le choix de la partition à adopter en détectant l'existence d'un coude sur l'histogramme des indices de niveau.

```
+--------------+-----------+-----------------------------------------------------+
|    PALIER    | VALEUR DU |                                                     |
|     ENTRE    |   PALIER  |                                                     |
+--------------+-----------+-----------------------------------------------------+
|  1993--  1994|    -39.99 | ************************************************** |
|  1990--  1991|    -19.79 | *************************                          |
|  1995--  1996|    -17.70 | ***********************                            |
+--------------+-----------+-----------------------------------------------------+
```

LISTE DES 3 MEILLEURE(S) PARTITION(S) ENTRE 3 ET 10 CLASSES

```
  1 - PARTITION EN   7 CLASSES
  2 - PARTITION EN  10 CLASSES
  3 - PARTITION EN   5 CLASSES
```

Coupure 'b' de l'arbre en 7 classes
FORMATION DES CLASSES (INDIVIDUS ACTIFS)
DESCRIPTION SOMMAIRE

```
+--------+----------+-----------+-----------+
| CLASSE | EFFECTIF |   POIDS   |  CONTENU  |
+--------+----------+-----------+-----------+
|  bb1b  |    106   |   106.00  |   1 A   3 |
|  bb2b  |    375   |   375.00  |   4 A  13 |
|  bb3b  |     70   |    70.00  |  14 A  16 |
|  bb4b  |     79   |    79.00  |  17 A  18 |
|  bb5b  |     67   |    67.00  |  19 A  19 |
|  bb6b  |    141   |   141.00  |  20 A  23 |
|  bb7b  |    162   |   162.00  |  24 A  30 |
+--------+----------+-----------+-----------+
```

COORDONNEES ET VALEURS-TEST AVANT CONSOLIDATION - AXES 1 A 5

```
+-----------------------------------------+---------------------------+-------------------------------+--------+
|                CLASSES                  |        VALEURS-TEST       |          COORDONNEES          |        |
|-----------------------------------------|---------------------------|-------------------------------|--------|
| IDEN - LIBELLE              EFF.  P.ABS |   1    2    3    4    5   |   1     2     3     4     5   | DISTO. |
+-----------------------------------------+---------------------------+-------------------------------+--------+
|   Coupure 'b' de l'arbre en  7 classes                                                                      |
|                                                                                                             |
| bb1b - CLASSE  1 /  7       106  106.00 |   3.7 -8.7 -0.3  2.3   6.9 |  0.18 -0.39 -0.01  0.09  0.27 |   0.83 |
| bb2b - CLASSE  2 /  7       375  375.00 | -15.2 -7.2  2.3 -6.6   4.2 | -0.32 -0.14  0.04 -0.12  0.07 |   0.22 |
| bb3b - CLASSE  3 /  7        70   70.00 | -10.6  7.0 -9.4  6.8   0.6 | -0.63  0.39 -0.49  0.35  0.03 |   1.75 |
| bb4b - CLASSE  4 /  7        79   79.00 |  -6.3  2.4  3.6  8.2  -4.9 | -0.35  0.13  0.18  0.39 -0.23 |   1.57 |
| bb5b - CLASSE  5 /  7        67   67.00 |   2.8 -2.1 -4.3 -2.8  -1.7 |  0.17 -0.12 -0.23 -0.15 -0.09 |   1.98 |
| bb6b - CLASSE  6 /  7       141  141.00 |  12.2 -1.6 -1.9  2.9 -11.2 |  0.49 -0.06 -0.07  0.10 -0.38 |   0.75 |
| bb7b - CLASSE  7 /  7       162  162.00 |  15.4 13.0  5.8 -4.8   3.5 |  0.58  0.46  0.19 -0.15  0.11 |   0.73 |
+-----------------------------------------+---------------------------+-------------------------------+--------+
```

CONSOLIDATION DE LA PARTITION

L'intérêt de la consolidation est de réaffecter les éléments d'une classe dans une autre classe dont ils sont plus proche de façon à améliorer l'homogénéité interne des classes. Ce processus s'effectue par des itérations successives à centres mobiles. Les centres sont initialement les centres de gravité des classes obtenues par coupure de l'arbre. Ces centres évoluent lorsque les individus passent d'une classe à l'autre.

AUTOUR DES 7 CENTRES DE CLASSES, REALISEE PAR 10 ITERATIONS A CENTRES MOBILES
PROGRESSION DE L'INERTIE INTER-CLASSES

ITERATION	I.TOTALE	I.INTER	QUOTIENT
0	2.35008	0.77272	0.32881
1	2.35008	0.82435	0.35078
2	2.35008	0.82613	0.35153
3	2.35008	0.82630	0.35160
4	2.35008	0.82630	0.35160

ARRET APRES L'ITERATION 4 L'ACCROISSEMENT DE L'INERTIE INTER-CLASSES
PAR RAPPORT A L'ITERATION PRECEDENTE N'EST QUE DE 0.000 %.

DECOMPOSITION DE L'INERTIE

CALCULEE SUR 14 AXES.

	INERTIES		EFFECTIFS		POIDS		DISTANCES	
INERTIES	AVANT	APRES	AVANT	APRES	AVANT	APRES	AVANT	APRES
INTER-CLASSES	0.7727	0.8263						
INTRA-CLASSE								
CLASSE 1 / 7	0.1299	0.1731	106	128	106.00	128.00	0.8283	0.8028
CLASSE 2 / 7	0.6116	0.5710	375	358	375.00	358.00	0.2191	0.2551
CLASSE 3 / 7	0.0930	0.0945	70	72	70.00	72.00	1.7521	1.7687
CLASSE 4 / 7	0.1233	0.1336	79	82	79.00	82.00	1.5661	1.5452
CLASSE 5 / 7	0.1293	0.1293	67	67	67.00	67.00	1.9831	1.9831
CLASSE 6 / 7	0.2054	0.2180	141	149	141.00	149.00	0.7483	0.7707
CLASSE 7 / 7	0.2849	0.2043	162	144	162.00	144.00	0.7286	0.9060
TOTALE	2.3501	2.3501						

QUOTIENT (INERTIE INTER / INERTIE TOTALE) : AVANT ... 0.3288
APRES ... 0.3516

COORDONNEES ET VALEURS-TEST APRES CONSOLIDATION
AXES 1 A 5

CLASSES			VALEURS-TEST					COORDONNEES					
IDEN - LIBELLE	EFF.	P.ABS	1	2	3	4	5	1	2	3	4	5	DISTO.
Coupure 'b' de l'arbre en 7 classes													
bb1b - CLASSE 1 / 7	128	128.00	3.8	-8.6	-1.8	4.1	8.0	0.16	-0.35	-0.07	0.15	0.28	0.80
bb2b - CLASSE 2 / 7	358	358.00	-16.0	-6.4	1.9	-8.1	4.0	-0.35	-0.13	0.04	-0.15	0.07	0.26
bb3b - CLASSE 3 / 7	72	72.00	-10.7	8.3	-9.3	6.5	-0.1	-0.63	0.46	-0.48	0.32	0.00	1.77
bb4b - CLASSE 4 / 7	82	82.00	-5.8	2.7	2.9	7.9	-5.6	-0.32	0.14	0.14	0.37	-0.25	1.55
bb5b - CLASSE 5 / 7	67	67.00	2.8	-2.1	-4.3	-2.8	-1.7	0.17	-0.12	-0.23	-0.15	-0.09	1.98
bb6b - CLASSE 6 / 7	149	149.00	13.3	-1.2	-2.6	2.4	-11.6	0.52	-0.04	-0.09	0.08	-0.38	0.77
bb7b - CLASSE 7 / 7	144	144.00	15.1	11.5	9.4	-4.1	4.3	0.61	0.43	0.33	-0.14	0.14	0.91

PARANGONS

Les parangons sont les individus les plus « caractéristiques » de chaque groupe au sens suivant : ce sont les individus les plus proches du centre de gravité (du point moyen) de la classe.

CLASSE 1/ 7

EFFECTIF : 128

RK	DISTANCE	IDENT.	RK	DISTANCE	IDENT.	RK	DISTANCE	IDENT.
1	0.51034	0980	2	0.56936	0091	3	0.58376	0485
4	0.58376	0619	5	0.62658	0368	6	0.62658	0897
7	0.63989	0704	8	0.66465	0184	9	0.66465	0232
10	0.66465	0238						

CLASSE 2/ 7

EFFECTIF : 358

RK	DISTANCE	IDENT.	RK	DISTANCE	IDENT.	RK	DISTANCE	IDENT.
1	0.66989	0459	2	0.80053	0043	3	0.80753	0322
4	0.86366	0393	5	0.86366	0450	6	0.86366	0780
7	0.86366	0540	8	0.86366	0460	9	0.90535	0082
10	0.91404	0593						

CLASSE 3/ 7

EFFECTIF : 72

RK	DISTANCE	IDENT.	RK	DISTANCE	IDENT.	RK	DISTANCE	IDENT.
1	0.58799	0741	2	0.60470	0940	3	0.61735	0639
4	0.61735	0788	5	0.69764	0789	6	0.70722	0758
7	0.78494	0766	8	0.78494	0806	9	0.82442	0742
10	0.82442	0946						

CLASSE 4/ 7

EFFECTIF : 82

RK	DISTANCE	IDENT.	RK	DISTANCE	IDENT.	RK	DISTANCE	IDENT.
1	0.74814	0156	2	0.98976	0575	3	1.01170	0730
4	1.07622	0569	5	1.12107	0721	6	1.12879	0148
7	1.12879	0660	8	1.12879	0715	9	1.14287	0566
10	1.14460	0360						

CLASSE 5/ 7

EFFECTIF : 67

RK	DISTANCE	IDENT.	RK	DISTANCE	IDENT.	RK	DISTANCE	IDENT.
1	0.97554	0358	2	1.10787	0130	3	1.12353	0328
4	1.27382	0288	5	1.27888	0825	6	1.29654	0165
7	1.30224	0828	8	1.30330	0302	9	1.30330	0326
10	1.34956	0208						

CLASSE 6/ 7

EFFECTIF : 149

RK	DISTANCE	IDENT.	RK	DISTANCE	IDENT.	RK	DISTANCE	IDENT.
1	0.52061	0062	2	0.52061	0240	3	0.55153	0419
4	0.55153	0611	5	0.66158	0991	6	0.70375	0286
7	0.70767	0251	8	0.75757	0497	9	0.77031	0377
10	0.78869	0242						

CLASSE 7/ 7

EFFECTIF : 144

RK	DISTANCE	IDENT.	RK	DISTANCE	IDENT.	RK	DISTANCE	IDENT.
1	0.54714	0141	2	0.58623	0007	3	0.60549	0243
4	0.63791	0200	5	0.64338	0025	6	0.72304	0172
7	0.72691	0004	8	0.74024	0006	9	0.74024	0352
10	0.74024	0343						

DESCRIPTION DE LA Coupure 'b' de l'arbre en 7 classes

CARACTERISATION PAR LES MODALITES DES CLASSES OU MODALITES DE Coupure 'b' de l'arbre en 7 classes

L'objet de cette caractérisation est de déterminer les modalités des variables qui caractérisent le plus la classe à décrire. Cela permet de savoir s'il y a une sur ou une sous représentation de cette modalité dans la classe.
On évalue l'écart entre le pourcentage de la modalité dans la classe (MOD/CLA) et le pourcentage de cette modalité dans la population globale (GLOBAL). Pour chacun des modalités, SPAD calcule la probabilité d'observer un écart au moins aussi grand que celui qui s'est réalisé, dans l'hypothèse où la modalité serait distribuée dans la classe comme dans la population. Cette probabilité évalue l'importance de l'écart entre les deux pourcentages. Plus la probabilité est faible, plus l'écart est jugé « caractéristique ». La valeur-test est le fractile de la loi normale correspondant à la même probabilité.

CLASSE 1 / 7

V.TEST	PROBA	POURCENTAGES CLA/MOD	POURCENTAGES MOD/CLA	POURCENTAGES GLOBAL	MODALITES CARACTERISTIQUES	DES VARIABLES	IDEN	POIDS
				12.80	CLASSE 1 / 7		bb1b	128
24.52	0.000	81.01	100.00	15.80	BEPC-BE-BEPS	Diplôme de l'enquêté(e) en 5 classes	die3	158
4.73	0.000	17.59	71.88	52.30	locataire	Statut d'occupation du logement en 4 classes	slo3	523
4.51	0.000	23.23	35.94	19.80	employés	Profession de l'enquêté(e) en 7 classes	csp4	198
3.56	0.000	15.84	75.00	60.60	souvent	Vous arrive-t-il d'inviter des amis à déjeuner ?	bou1	606
3.10	0.001	18.31	40.63	28.40	25 à 34 ans	Age de l'enquêté(e) en 5 classes	agc2	284
3.08	0.001	17.61	46.09	33.50	Employés	Type d'emploi	emp2	335
2.85	0.002	20.67	24.22	15.00	Moins de 25 ans	Age de l'enquêté(e) en 5 classes	agc1	150
2.73	0.003	17.75	38.28	27.60	aucun risque	Votre travail présente-t-il des risques pour la santé ?	tra3	276
2.69	0.004	16.54	50.00	38.70	dissout si accord	Opinion à propos du mariage	opm3	387
2.52	0.006	14.76	75.78	65.70	très importante	La préservation de l'environnement est une chose ...	env1	657
2.24	0.013	20.37	17.19	10.80	beaucoup moins bien	Evolution du niveau de vie des français depuis 10 ans	nif5	108
2.16	0.015	15.55	52.34	43.10	deux	Nombre d'enfants considéré comme idéal	enf2	431
2.16	0.015	15.55	52.34	43.10	non	La famille est le seul endroit oú l'on se sente bien	fbi2	431
2.05	0.020	15.35	53.13	44.30	entre 22h. et 23h.	Heure de coucher	dod3	443
2.03	0.021	15.67	46.88	38.30	oui, beaucoup	Les découvertes scientifiques améliorent-elles la vie ?	sci2	383

CLASSE 2 / 7

V.TEST	PROBA	POURCENTAGES CLA/MOD	POURCENTAGES MOD/CLA	POURCENTAGES GLOBAL	MODALITES CARACTERISTIQUES	DES VARIABLES	IDEN	POIDS
				35.80	CLASSE 2 / 7		bb2b	358
14.73	0.000	68.54	61.45	32.10	CEP ou fin études	Diplôme de l'enquêté(e) en 5 classes	die2	321
12.34	0.000	67.68	49.72	26.30	ouvriers	Profession de l'enquêté(e) en 7 classes	csp3	263
12.34	0.000	67.68	49.72	26.30	Ouvriers	Type d'emploi	emp1	263
11.58	0.000	73.02	38.55	18.90	Aucun	Diplôme de l'enquêté(e) en 5 classes	die1	189
7.95	0.000	50.12	58.66	41.90	tous les jours	Regardez-vous la télévision ...	tél1	419
6.94	0.000	47.20	61.17	46.40	non	Appartenance à au moins une association	ass2	464
6.67	0.000	44.74	70.11	56.10	oui	La famille est le seul endroit oú l'on se sente bien	fbi1	561
6.09	0.000	49.24	45.25	32.90	plus de 100.000	Taille d'agglomération (en nombre d'habitants)	agg4	329
5.94	0.000	56.00	27.37	17.50	20.000 - 100.000	Taille d'agglomération (en nombre d'habitants)	agg3	175
5.85	0.000	39.02	95.25	87.40	non	Parcipation à une action de défense de l'environnement	déf2	874
5.52	0.000	51.52	33.24	23.10	union indissoluble	Opinion à propos du mariage	opm1	231
5.32	0.000	38.68	94.97	87.90	non	Possédez vous des valeurs mobilières ?	vmo2	879
5.18	0.000	39.80	87.71	78.90	non	Possession ou usage d'une machine à laver la vaisselle	lav2	789
4.88	0.000	56.67	18.99	12.00	jamais	Vous arrive-t-il d'inviter des amis à déjeuner ?	bou3	120
4.41	0.000	46.31	38.55	29.80	assez importante	La préservation de l'environnement est une chose ...	env2	298
4.33	0.000	50.89	24.02	16.90	65 ans et plus	Age de l'enquêté(e) en 5 classes	agc5	169
4.14	0.000	41.87	61.17	52.30	locataire	Statut d'occupation du logement en 4 classes	slo3	523
4.02	0.000	37.58	96.37	91.80	non	Possédez vous des biens immobiliers ?	vim2	918
4.02	0.000	45.99	35.20	27.40	rarement	Vous arrive-t-il d'inviter des amis à déjeuner ?	bou2	274
3.78	0.000	44.05	41.34	33.60	plutôt à la femme	A qui incombent les travaux ménagers et les soins enfants ?	esc2	336
3.32	0.000	54.79	11.17	7.30	21h. ou avant	Heure de coucher	dod1	73
3.04	0.001	51.09	13.13	9.20	pas du tout satisf.	La crèche est un mode de garde ...	cre4	92
2.85	0.002	42.01	39.66	33.80	femme plus 38 ans	Age et sexe de l'enquêteur	enq4	338
2.78	0.003	42.96	32.40	27.00	entre 21h. et 22h.	Heure de coucher	dod2	270
2.72	0.003	57.14	6.70	4.20	incombent à la femme	A qui incombent les travaux ménagers et les soins enfants ?	esc1	42
2.60	0.005	42.39	32.68	27.60	un peu moins bien	Evolution du niveau de vie des français depuis 10 ans	nif4	276
2.54	0.005	44.15	23.18	18.80	50 à 64 ans	Age de l'enquêté(e) en 5 classes	agc4	188
2.53	0.006	47.62	13.97	10.50	pas du tout	Les découvertes scientifiques améliorent-elles la vie ?	sci3	105
2.29	0.011	37.66	82.68	78.60	très satisfaisant	La mère au foyer est un mode de garde ...	mèr1	786
2.25	0.012	43.16	22.91	19.00	un peu moins bien	Evolution du niveau de vie de l'enquêté depuis 10 ans	niv4	190
2.19	0.014	36.73	95.53	93.10	non	Faites-vous partie d'une association confessionnelle ?	asc2	931
2.05	0.020	38.75	58.66	54.20	assez	L'enquêté(e) s'est-il (elle) montré(e) intéressé(e) ?	sop2	542
1.97	0.024	42.13	23.18	19.70	beaucoup	Etes-vous gêné par les bruits ?	bru2	197

CLASSE 3 / 7

V.TEST	PROBA	CLA/MOD	MOD/CLA	GLOBAL	MODALITES CARACTERISTIQUES	DES VARIABLES	IDEN	POIDS
				7.20	CLASSE 3 / 7		bb3b	72
21.01	0.000	86.75	100.00	8.30	moins de 2.000	Taille d'agglomération (en nombre d'habitants)	agg1	83
9.61	0.000	20.34	81.94	29.00	propriétaire	Statut d'occupation du logement en 4 classes	slo2	290
8.27	0.000	50.00	33.33	4.80	Autres	Type d'emploi	emp4	48
7.83	0.000	16.57	77.78	33.80	femme plus 38 ans	Age et sexe de l'enquêteur	enq4	338
7.28	0.000	30.53	40.28	9.50	ex. agr.-art-commer	Profession de l'enquêté(e) en 7 classes	csp1	95
6.42	0.000	11.59	90.28	56.10	oui	La famille est le seul endroit oú l'on se sente bien	fbi1	561
6.24	0.000	22.30	43.06	13.90	ne sait pas	La crèche est un mode de garde ...	cre5	139
5.79	0.000	12.89	75.00	41.90	tous les jours	Regardez-vous la télévision ...	tél1	419
5.70	0.000	11.63	83.33	51.60	beaucoup	La vue sur l'extérieur vous plaît-elle ?	vue1	516
4.50	0.000	8.91	97.22	78.60	très satisfaisant	La mère au foyer est un mode de garde ...	mèr1	786
4.47	0.000	10.07	84.72	60.60	pas du tout	Etes-vous gêné par les bruits ?	bru3	606
4.40	0.000	12.77	56.94	32.10	CEP ou fin études	Diplôme de l'enquêté(e) en 5 classes	die2	321
4.21	0.000	10.61	75.00	50.90	oui, un peu	Les découvertes scientifiques améliorent-elles la vie ?	sci1	509
4.18	0.000	12.35	56.94	33.20	très	L'enquêté(e) s'est-il (elle) montré(e) intéressé(e) ?	sop1	332
3.37	0.000	20.00	18.06	6.50	ne sait pas	Opinion sur le fonctionnement de la justice en 1979	jus5	65
3.19	0.001	11.85	44.44	27.00	entre 21h. et 22h.	Heure de coucher	dod2	270
3.19	0.001	12.56	38.89	22.30	*Reponse manquante*	Pour que la société change, faut-il ...	36_	223
2.94	0.002	12.77	33.33	18.80	50 à 64 ans	Age de l'enquêté(e) en 5 classes	agc4	188
2.91	0.002	11.58	41.67	25.90	très satisfait	Opinion sur le cadre de vie quotidien	cvi1	259
2.74	0.003	8.01	97.22	87.40	non	A souffert d'état dépressif ces quatre dernières semaines :	éta2	874
2.63	0.004	9.00	75.00	60.00	satisfaisante	Comparée aux personnes de votre âge, votre santé est ...	san2	600
2.46	0.007	9.14	68.06	53.60	oui	Appartenance à au moins une association	ass1	536
2.35	0.009	15.49	15.28	7.10	ne sait pas	La société francaise a-t-elle besoin de se transformer ?	tso3	71
2.33	0.010	20.00	9.72	3.50	homme plus 38 ans	Age et sexe de l'enquêteur	enq2	35
2.21	0.014	10.82	34.72	23.10	union indissoluble	Opinion à propos du mariage	opm1	231
2.13	0.017	7.85	95.83	87.90	non	Possédez vous des valeurs mobilières ?	vmo2	879
2.11	0.018	10.07	41.67	29.80	assez importante	La préservation de l'environnement est une chose ...	env2	298

CLASSE 4 / 7

V.TEST	PROBA	CLA/MOD	MOD/CLA	GLOBAL	MODALITES CARACTERISTIQUES	DES VARIABLES	IDEN	POIDS
				8.20	CLASSE 4 / 7		bb4b	82
22.73	0.000	94.25	100.00	8.70	2.000 - 20.000	Taille d'agglomération (en nombre d'habitants)	agg2	87
4.33	0.000	34.29	14.63	3.50	homme plus 38 ans	Age et sexe de l'enquêteur	enq2	35
3.15	0.001	16.67	24.39	12.00	en accession	Statut d'occupation du logement en 4 classes	slo1	120
2.94	0.002	12.05	48.78	33.20	très	L'enquêté(e) s'est-il (elle) montré(e) intéressé(e) ?	sop1	332
2.71	0.003	10.34	70.73	56.10	oui	La famille est le seul endroit oú l'on se sente bien	fbi1	561
2.70	0.003	10.45	68.29	53.60	oui	Appartenance à au moins une association	ass1	536
2.60	0.005	10.07	74.39	60.60	pas du tout	Etes-vous gêné par les bruits ?	bru3	606
2.56	0.005	15.84	19.51	10.10	homme moins 39 ans	Age et sexe de l'enquêteur	enq1	101
2.17	0.015	11.38	40.24	29.00	propriétaire	Statut d'occupation du logement en 4 classes	slo2	290
2.13	0.017	10.08	63.41	51.60	beaucoup	La vue sur l'extérieur vous plaît-elle ?	vue1	516
2.11	0.017	14.74	17.07	9.50	ex. agr.-art-commer	Profession de l'enquêté(e) en 7 classes	csp1	95
2.06	0.020	15.38	14.63	7.80	beaucoup mieux	Evolution du niveau de vie des français depuis 10 ans	nif1	78
2.03	0.021	11.96	30.49	20.90	35 à 49 ans	Age de l'enquêté(e) en 5 classes	agc3	209
1.99	0.023	11.52	34.15	24.30	assez bon	Opinion sur le fonctionnement de la justice en 1979	jus2	243

CLASSE 5 / 7

V.TEST	PROBA	CLA/MOD	MOD/CLA	GLOBAL	MODALITES CARACTERISTIQUES	DES VARIABLES	IDEN	POIDS
				6.70	CLASSE 5 / 7		bb5b	67
21.82	0.000	100.00	100.00	6.70	logé gratuit, autre	Statut d'occupation du logement en 4 classes	slo4	67
7.47	0.000	27.43	46.27	11.30	négligeables	Les dépenses de logement sont pour vous ...	dlo1	113
5.97	0.000	54.55	17.91	2.20	ne sait pas	Les dépenses de logement sont pour vous ...	dlo6	22
3.73	0.000	20.00	20.90	7.00	personnel de service	Profession de l'enquêté(e) en 7 classes	csp6	70
2.09	0.018	9.92	35.82	24.20	peu satisfaisant	La crèche est un mode de garde ...	cre3	242

CLASSE 6 / 7

V.TEST	PROBA	---- POURCENTAGES ---- CLA/MOD	MOD/CLA	GLOBAL	MODALITES CARACTERISTIQUES	DES VARIABLES	IDEN	POIDS
				14.90	CLASSE 6 / 7		bb6b	149
25.06	0.000	80.77	98.66	18.20	Bac - Brevet sup.	Diplôme de l'enquêté(e) en 5 classes	die4	182
7.70	0.000	38.26	38.26	14.90	contremaî-cad. moy.	Profession de l'enquêté(e) en 7 classes	csp5	149
6.50	0.000	23.43	67.79	43.10	non	La famille est le seul endroit oú l'on se sente bien	fbi2	431
5.87	0.000	27.95	42.95	22.90	Cadres	Type d'emploi	emp3	229
5.03	0.000	25.97	40.27	23.10	pas très souvent	Regardez-vous la télévision ...	tél3	231
4.67	0.000	30.40	25.50	12.50	*Reponse manquante*	Profession de l'enquêté(e) en 7 classes	54_	125
4.67	0.000	30.40	25.50	12.50	*Reponse manquante*	Type d'emploi	49_	125
4.23	0.000	27.33	27.52	15.00	Moins de 25 ans	Age de l'enquêté(e) en 5 classes	agc1	150
4.14	0.000	18.65	75.84	60.60	souvent	Vous arrive-t-il d'inviter des amis à déjeuner ?	bou1	606
3.52	0.000	20.86	45.64	32.60	Paris	Taille d'agglomération (en nombre d'habitants)	agg5	326
3.38	0.000	17.66	77.85	65.70	très importante	La préservation de l'environnement est une chose ...	env1	657
3.38	0.000	31.75	13.42	6.30	après minuit	Heure de coucher	dod5	63
3.12	0.001	20.61	40.94	29.60	moyennement	La vue sur l'extérieur vous plaît-elle ?	vue2	296
3.04	0.001	19.38	50.34	38.70	dissout si accord	Opinion à propos du mariage	opm3	387
2.96	0.002	21.80	30.87	21.10	oui	Possession ou usage d'une machine à laver la vaisselle	lav1	211
2.87	0.002	23.88	21.48	13.40	entre 23h. et 24h.	Heure de coucher	dod4	134
2.61	0.005	23.26	20.13	12.90	assez satisfaisant	La mère au foyer est un mode de garde ...	mèr2	129
2.27	0.012	22.50	18.12	12.00	en accession	Statut d'occupation du logement en 4 classes	slo1	120
2.26	0.012	17.35	62.42	53.60	oui	Appartenance à au moins une association	ass1	536
2.25	0.012	22.22	18.79	12.60	oui	Parcipation à une action de défense de l'environnement	déf1	126
2.17	0.015	19.01	36.24	28.40	25 à 34 ans	Age de l'enquêté(e) en 5 classes	agc2	284
2.16	0.015	18.07	47.65	39.30	trois	Nombre d'enfants considéré comme idéal	enf3	393
2.11	0.017	22.77	15.44	10.10	homme moins 39 ans	Age et sexe de l'enquêteur	enq1	101
2.10	0.018	17.16	62.42	54.20	assez	L'enquêté(e) s'est-il (elle) montré(e) intéressé(e) ?	sop2	542

CLASSE 7 / 7

V.TEST	PROBA	---- POURCENTAGES ---- CLA/MOD	MOD/CLA	GLOBAL	MODALITES CARACTERISTIQUES	DES VARIABLES	IDEN	POIDS
				14.40	CLASSE 7 / 7		bb7b	144
24.37	0.000	88.67	92.36	15.00	Université,gde école	Diplôme de l'enquêté(e) en 5 classes	die5	150
12.62	0.000	71.43	41.67	8.40	prof. lib.-cad. sup.	Profession de l'enquêté(e) en 7 classes	csp2	84
11.52	0.000	40.17	63.89	22.90	Cadres	Type d'emploi	emp3	229
7.36	0.000	26.69	60.42	32.60	Paris	Taille d'agglomération (en nombre d'habitants)	agg5	326
6.99	0.000	23.43	70.14	43.10	non	La famille est le seul endroit oú l'on se sente bien	fbi2	431
5.76	0.000	33.88	28.47	12.10	oui	Possédez vous des valeurs mobilières ?	vmo1	121
4.83	0.000	25.59	37.50	21.10	oui	Possession ou usage d'une machine à laver la vaisselle	lav1	211
4.74	0.000	18.11	82.64	65.70	très importante	La préservation de l'environnement est une chose ...	env1	657
4.69	0.000	29.10	27.08	13.40	entre 23h. et 24h.	Heure de coucher	dod4	134
4.21	0.000	18.15	76.39	60.60	souvent	Vous arrive-t-il d'inviter des amis à déjeuner ?	bou1	606
4.12	0.000	27.78	24.31	12.60	oui	Parcipation à une action de défense de l'environnement	déf1	126
3.98	0.000	27.42	23.61	12.40	jamais	Regardez-vous la télévision ...	tél4	124
3.78	0.000	18.44	67.36	52.60	femme moins 39 ans	Age et sexe de l'enquêteur	enq3	526
3.71	0.000	18.28	68.06	53.60	oui	Appartenance à au moins une association	ass1	536
3.07	0.001	18.86	50.69	38.70	dissout si accord	Opinion à propos du mariage	opm3	387
2.89	0.002	28.57	12.50	6.30	après minuit	Heure de coucher	dod5	63
2.86	0.002	34.29	8.33	3.50	peu satisfaisant	La mère au foyer est un mode de garde ...	mèr3	35
2.75	0.003	20.35	32.64	23.10	pas très souvent	Regardez-vous la télévision ...	tél3	231
2.64	0.004	40.00	5.56	2.00	pas du tout satisf.	La mère au foyer est un mode de garde ...	mèr4	20
2.55	0.005	18.13	48.61	38.60	assez satisfaisant	La crèche est un mode de garde ...	cre2	386
2.47	0.007	24.42	14.58	8.60	quatre et plus	Nombre d'enfants considéré comme idéal	enf4	86
2.27	0.012	21.77	18.75	12.40	peu ou pas	L'enquêté(e) s'est-il (elle) montré(e) intéressé(e) ?	sop3	124
2.21	0.014	24.64	11.81	6.90	oui	Faites-vous partie d'une association confessionnelle ?	asc1	69
2.21	0.014	16.83	61.11	52.30	locataire	Statut d'occupation du logement en 4 classes	slo3	523
2.15	0.016	23.46	13.19	8.10	oui	Possédez vous des biens immobiliers ?	vim1	81
2.02	0.022	18.35	34.03	26.70	très satisfaisante	Comparée aux personnes de votre âge, votre santé est ...	san1	267

CARACTERISATION PAR LES CONTINUES DES CLASSES OU MODALITES DE Coupure 'b' de l'arbre en 7 classes

Une variable continue caractérise une classe d'individus si sa moyenne dans la classe diffère notablement de la moyenne générale de la variable.

CLASSE 1 / 7

V.TEST	PROBA	MOYENNES CLASSE	MOYENNES GENERALE	ECARTS TYPES CLASSE	ECARTS TYPES GENERAL	VARIABLES CARACTERISTIQUES NUM.LIBELLE	IDEN
		CLASSE 1 / 7		(POIDS = 128.00		EFFECTIF = 128)	bb1b
-4.27	0.000	36.52	42.68	15.77	17.50	37.Age de l'enquêté(e)	âge

CLASSE 2 / 7

V.TEST	PROBA	MOYENNES CLASSE	MOYENNES GENERALE	ECARTS TYPES CLASSE	ECARTS TYPES GENERAL	VARIABLES CARACTERISTIQUES NUM.LIBELLE	IDEN
		CLASSE 2 / 7		(POIDS = 358.00		EFFECTIF = 358)	bb2b
6.92	0.000	47.81	42.68	17.76	17.50	37.Age de l'enquêté(e)	âge
2.31	0.011	4.46	4.05	4.57	4.19	42.Nombre de non-réponses au questionnaire	nrep
-4.13	0.000	14.92	18.31	17.45	19.37	48.Nombre de jours de vacances en été	vaca
-4.61	0.000	6303.81	7244.48	4430.63	4756.78	46.Revenu personnel souhaité	rsou
-15.00	0.000	14.81	17.29	2.38	3.88	45.Age de fin d'étude	finé

CLASSE 3 / 7

V.TEST	PROBA	MOYENNES CLASSE	MOYENNES GENERALE	ECARTS TYPES CLASSE	ECARTS TYPES GENERAL	VARIABLES CARACTERISTIQUES NUM.LIBELLE	IDEN
		CLASSE 3 / 7		(POIDS = 72.00		EFFECTIF = 72)	bb3b
4.70	0.000	6.29	4.05	4.26	4.19	42.Nombre de non-réponses au questionnaire	nrep
2.56	0.005	47.76	42.68	14.15	17.50	37.Age de l'enquêté(e)	âge
-4.27	0.000	4223.21	5561.89	1412.19	2423.40	47.Estimation du revenu minimum d'une famille de 2 enfants	rmin
-4.95	0.000	15.08	17.29	2.49	3.88	45.Age de fin d'étude	finé
-5.88	0.000	5.37	18.31	8.65	19.37	48.Nombre de jours de vacances en été	vaca

CLASSE 4 / 7

V.TEST	PROBA	MOYENNES CLASSE	MOYENNES GENERALE	ECARTS TYPES CLASSE	ECARTS TYPES GENERAL	VARIABLES CARACTERISTIQUES NUM.LIBELLE	IDEN
		CLASSE 4 / 7		(POIDS = 82.00		EFFECTIF = 82)	bb4b
-2.71	0.003	16.17	17.29	3.30	3.88	45.Age de fin d'étude	finé

CLASSE 5 / 7

V.TEST	PROBA	MOYENNES CLASSE	MOYENNES GENERALE	ECARTS TYPES CLASSE	ECARTS TYPES GENERAL	VARIABLES CARACTERISTIQUES NUM.LIBELLE	IDEN
		CLASSE 5 / 7		(POIDS = 67.00		EFFECTIF = 67)	bb5b
2.77	0.003	9816.67	8478.73	8794.04	3668.95	15.Estimation du salaire mensuel d'un ingénieur	ring

CLASSE 6 / 7

V.TEST	PROBA	MOYENNES CLASSE	MOYENNES GENERALE	ECARTS TYPES CLASSE	ECARTS TYPES GENERAL	VARIABLES CARACTERISTIQUES NUM.LIBELLE	IDEN
		CLASSE 6 / 7		(POIDS = 149.00		EFFECTIF = 149)	bb6b
8.19	0.000	19.69	17.29	2.23	3.88	45.Age de fin d'étude	finé
6.13	0.000	27.29	18.31	22.66	19.37	48.Nombre de jours de vacances en été	vaca
3.81	0.000	8680.15	7244.48	5272.94	4756.78	46.Revenu personnel souhaité	rsou
-3.67	0.000	2.89	4.05	3.44	4.19	42.Nombre de non-réponses au questionnaire	nrep
-5.97	0.000	34.78	42.68	14.47	17.50	37.Age de l'enquêté(e)	âge

CLASSE 7 / 7

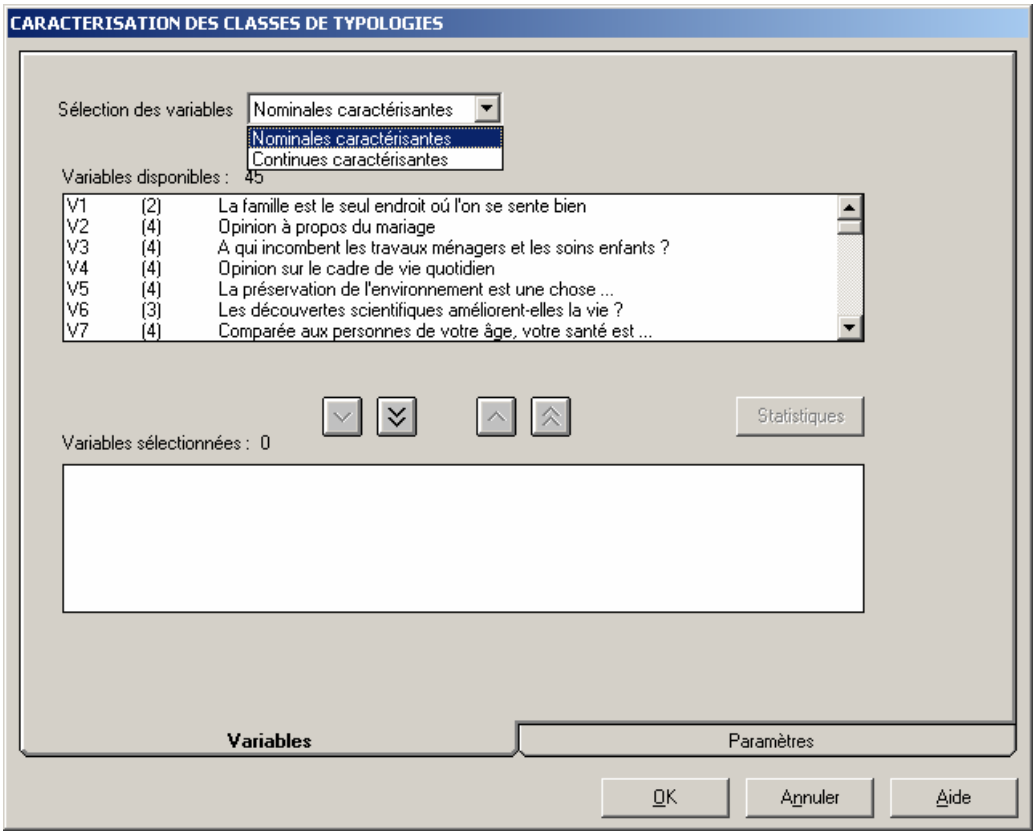
V.TEST	PROBA	MOYENNES CLASSE	MOYENNES GENERALE	ECARTS TYPES CLASSE	ECARTS TYPES GENERAL	VARIABLES CARACTERISTIQUES NUM.LIBELLE	IDEN
		CLASSE 7 / 7		(POIDS = 144.00		EFFECTIF = 144)	bb7b
17.43	0.000	22.51	17.29	3.76	3.88	45.Age de fin d'étude	finé
6.19	0.000	9692.00	7244.48	5711.32	4756.78	46.Revenu personnel souhaité	rsou
4.50	0.000	25.03	18.31	20.30	19.37	48.Nombre de jours de vacances en été	vaca
3.07	0.001	6148.18	5561.89	1957.92	2423.40	47.Estimation du revenu minimum d'une famille de 2 enfants	rmin
-3.20	0.001	3.02	4.05	3.31	4.19	42.Nombre de non-réponses au questionnaire	nrep

CLASS - MINER - Description des classes

Dans la procédure précédente PARTI-DECLA, on caractérise la partition obtenue à partir des variables actives et illustratives sélectionnées dans l'analyse factorielle préalable.
La procédure CLASS-MINER permet de ne sélectionner qu'une partie des variables du fichier pour la caractérisation de la partition. Elle permet à l'utilisateur de bénéficier d'une grande souplesse en terme de choix des variables.

LES PARAMETRES DE LA METHODE **CLASS-MINER**

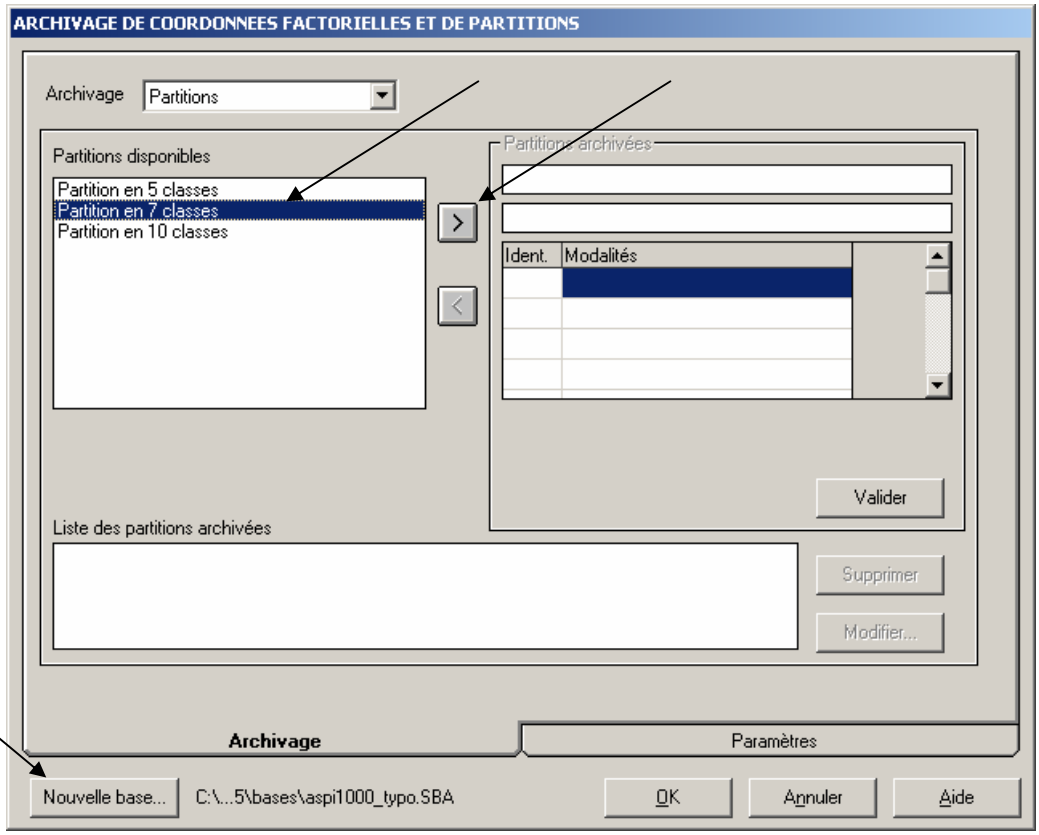
L'ONGLET « **VARIABLES** »

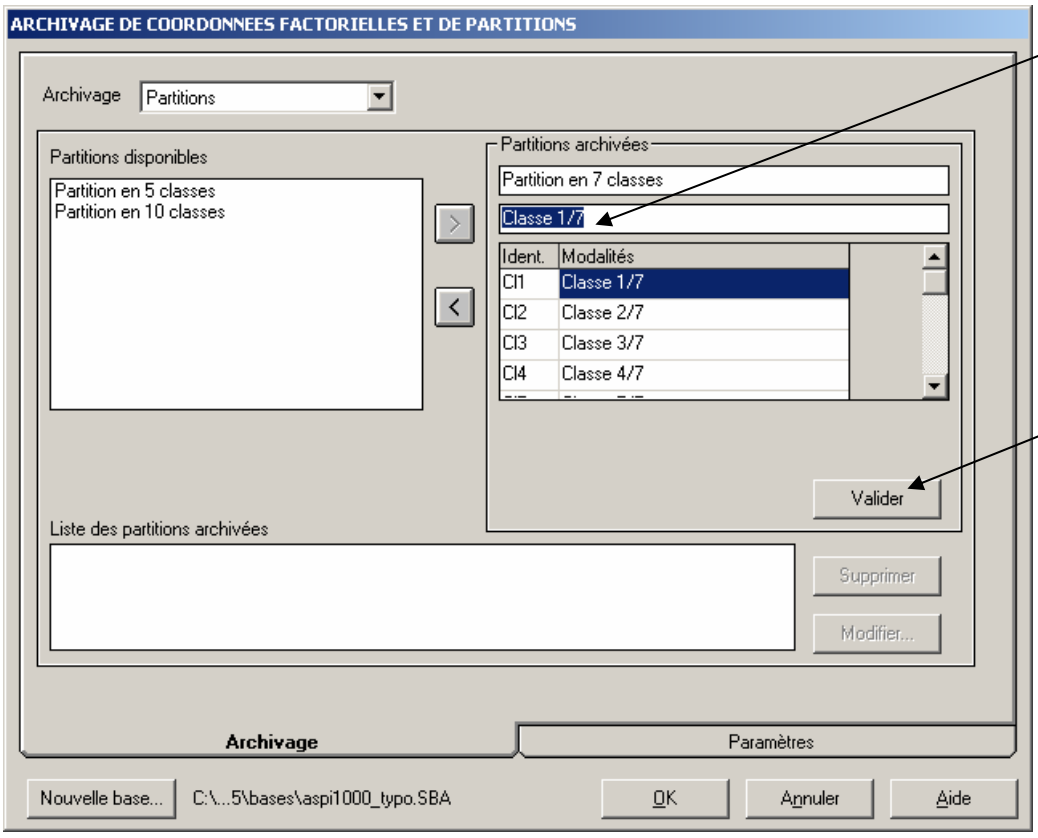


L'ONGLET « **PARAMETRES** »

Onglet identique à l'onglet « **Paramètres** » de la méthode **DEMOD**.

Archivages axes factoriels et partitions





LE MODELE LINEAIRE ET SES EXTENSIONS


Régression et analyse de la variance, Modèle Linéaire Général

---

1.1 Objet

Cette étape gère les calculs et les éditions d'un ajustement des moindres-carrés sur un modèle linéaire comprenant un terme constant. Elle permet d'effectuer les régressions multiples, les analyses de variance et de covariance avec interactions jusqu'à l'ordre 3. A chaque coefficient de la régression est associé le test de nullité, valable dans le contexte classique où le terme aléatoire est supposé engendré par une loi de Laplace-Gauss.

Si les observations sont caractérisées par différents critères nominaux ou "facteurs", le programme exécutera une analyse de la variance pour tester l'existence de l'effet de chacun des facteurs. S'il y a plusieurs critères, on peut introduire dans les modèles et tester d'éventuelles interactions entre couples et triplets de facteurs. Les estimations peuvent prendre en compte les répétitions d'observations dans les plans d'expérience.

On peut demander l'écriture d'un fichier contenant les coefficients de la régression, récupérables pour réaliser en particulier des représentations graphiques des résultats d'analyse de variance. Dans la version micro-ordinateur P.C., ce fichier sera utilisé par le module graphique d'interprétation des analyses de variance.

Le traitement des **données manquantes** est paramétrable tant pour les données continues, que pour les variables nominales.

1.2 Editions

On édite les statistiques sommaires sur les variables du modèle (tri-à-plat des variables nominales, moyenne, écart-type, minimum et maximum des variables continues). L'étape fournit l'identification des coefficients du modèle: coefficients des variables continues, des modalités des facteurs et des interactions éventuelles. Il est ensuite possible d'obtenir l'édition de la matrice des variances-covariances ou celle de la matrice des corrélations.

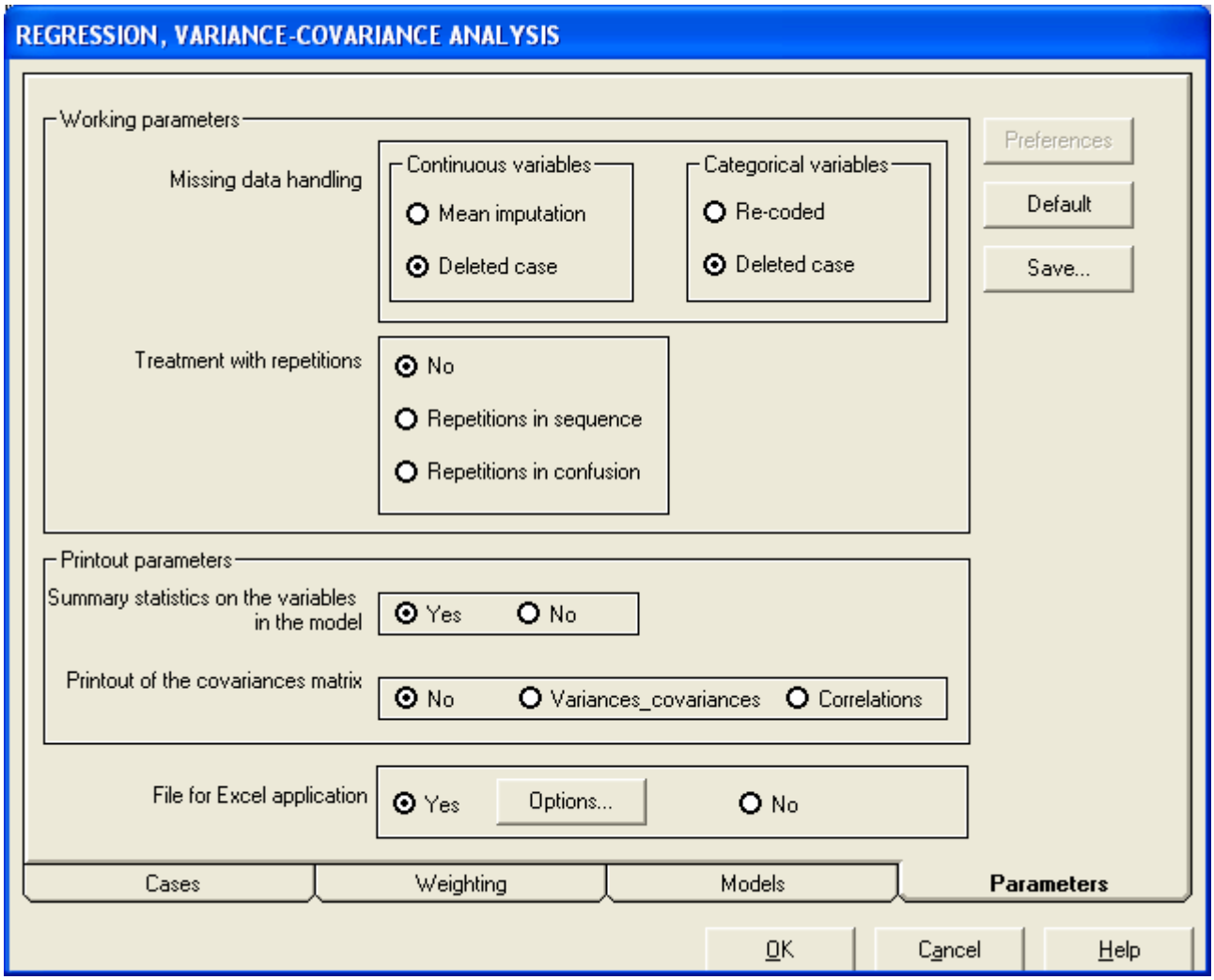
L'étape imprime les coefficients, l'estimation de leur écart-type, la statistique de Student correspondante, la probabilité critique ainsi que la valeur-test associée. On trouve également la somme des carrés des écarts, le coefficient de corrélation multiple, et

l'estimation de la variance commune des résidus. On effectue enfin le test de nullité simultanée de tous les coefficients (test d'une variable endogène "y" constante).

Dans le cas d'une analyse de la variance, on obtient de plus les sommes des carrés d'écarts suivant leur source (résiduelle, critère ou interaction), ainsi que les statistiques de Fisher, les probabilités critiques et valeurs-tests associées. Dans le cas d'observations répétées, on édite la variance "de répétabilité" ainsi que les estimations obtenues en tenant compte de cette variance.

1.3 Paramètres

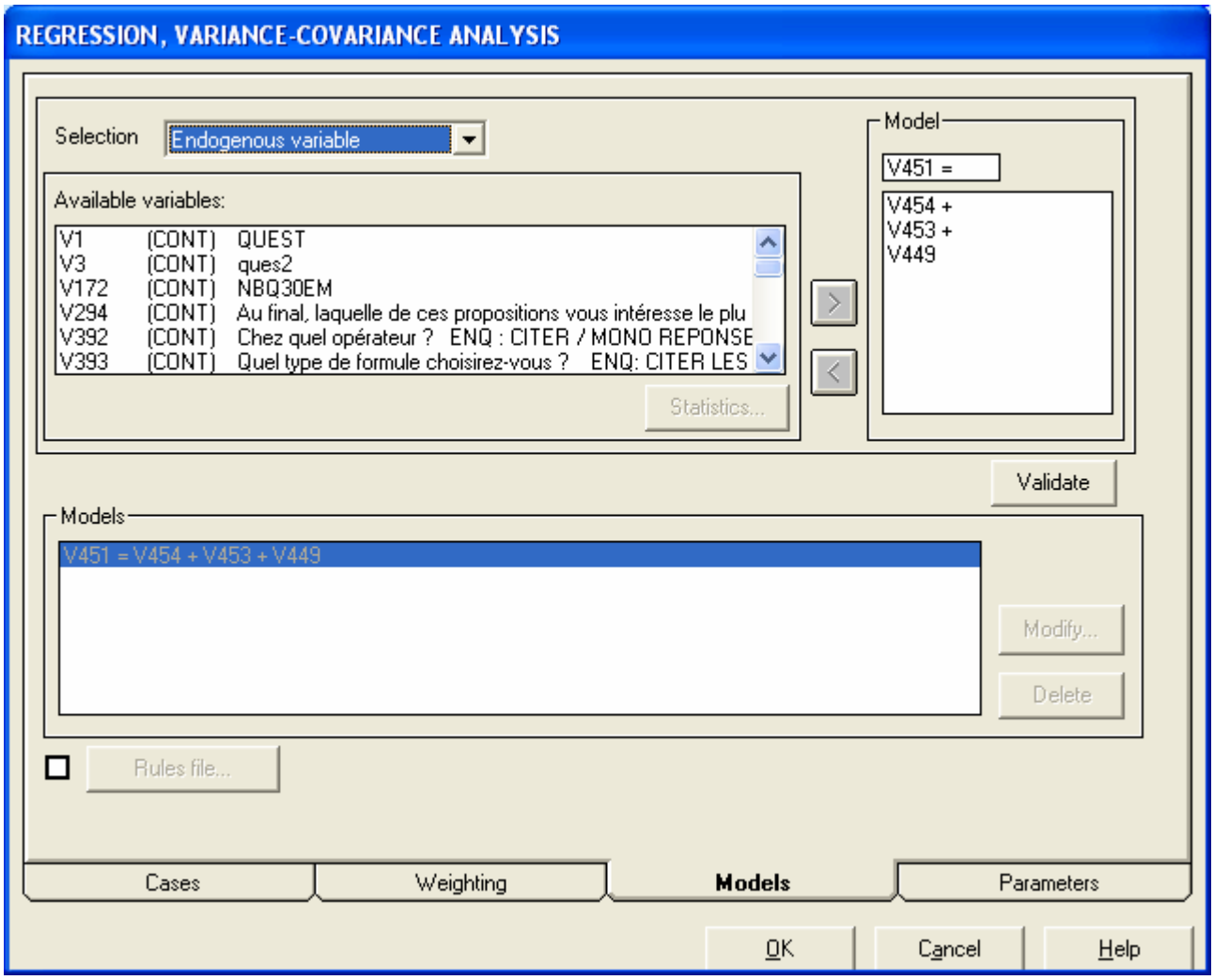
Les paramètres de la procédure VAREG permettent de choisir le mode de traitement des données manquantes (soit l'abandon, soit l'affectation à la moyenne générale). Un paramètre permet de prendre en compte les répétitions.



Notons que sur le plan des méthodes statistiques, des possibilités complémentaires de paramétrisation du modèle linéaire existent dans la méthode MLGEN – modèle linéaire généralisé.

1.4 Les modèles de régression

L'onglet Modèles permet de définir les modèles de régressions :



Une boite de sélection permet de sélectionner la catégorie : Endogène (*continue seulement*), Exogène, Interactions d'ordre 2 ou 3.

Les interactions peuvent être définies entre variables continues, nominales, ou mixant continues et nominales. Elles sont définies avec la souris sur le panneau de sélection des variables exogènes.

L'écriture du fichier externe des coefficients est commandée par une case à cocher et un bouton de sélection du fichier de sortie.

On crée alors un fichier de règles. Il s'agit d'un fichier texte contenant les valeurs des coefficients et les libellés des variables auxquelles ils se rapportent.

Si la procédure VAREG traite plusieurs modèles dans la même exécution, le fichier NCOVA ne contient que les résultats du *dernier* modèle traité.

Présentation détaillée des paramètres

LSUPR **Traitement des données manquantes**

- *valeurs possibles* : Recodé (les individus sont conservés)
  Supprimé (les individus sont retirés de l'analyse)
- *valeur par défaut* : Supprimé

L'option « Supprimé », élimine de l'analyse les individus présentant des données manquantes pour une des variables du modèle (endogène ou exogène).

On peut régler séparément le cas des
- variables continues (dont la variable endogène) : remplacement par la moyenne
- variables nominales : création de la modalité « valeur manquante »

Attention : si vous choisissez de recoder les continues, vérifiez que la variable endogène ne présente pas de données manquantes.

LSTAT **édition des statistiques sommaires sur les variables du modèle**

- *valeurs possibles* : 0 ou NON (pas d'édition)
  1 ou OUI (édition des statistiques)
- *valeur par défaut* : 0 ou NON

Si LSTAT = 1, on obtiendra un tri-à-plat des variables nominales du modèle, ainsi que diverses statistiques concernant les variables continues (moyenne, écart-type, minimum et maximum).

LMAT **édition de la matrice des variances-covariances**

- *valeurs possibles* : Pas d'Edition
  Edition de la matrice de variance- covariance
  Edition de la matrice de corrélations)
- *valeur par défaut* : Pas d'édition

LREP **calcul tenant compte des répétitions**

- *valeurs possibles* : Pas de répétitions
  Répétitions en séquence
  Répétitions en désordre
- *valeur par défaut* : Pas de répétition

Ce paramètre concerne le traitement des plans d'expérience. Lorsqu'il y a des répétitions, la variance des observations peut être estimée sur les répétitions d'observations plutôt que sur l'ensemble des observations. Il n'est pas nécessaire que le nombre de répétitions soit le même partout.

Avec l'option « Répétitions en séquence », les répétitions sont les unes en dessous des autres en lignes du tableau des données. Sinon on codera « Répétitions en désordre ». Cette dernière option est plus coûteuse en mémoire et temps de calcul.

Recherche des régressions optimales

Principes généraux

Cette méthode s'utilise avant de construire un modèle de régression ; elle permet d'explorer l'ensemble des modèles possibles, et donne pour chacun la qualité d'ajustement, le test de validation globale, les variables qui le composent, ainsi que les estimations des coefficients.

Ces éditions sont triées selon le nombre de variables entrant dans le modèle puis selon la qualité d'ajustement. L'utilisateur, dispose ainsi d'un outil de sélection parmi l'ensemble des modèles proposés en fonction de leurs ajustements et de leurs caractéristiques.

La procédure utilise l'algorithme de Furnival et Wilson pour explorer l'ensemble des modèles possibles.

Les données

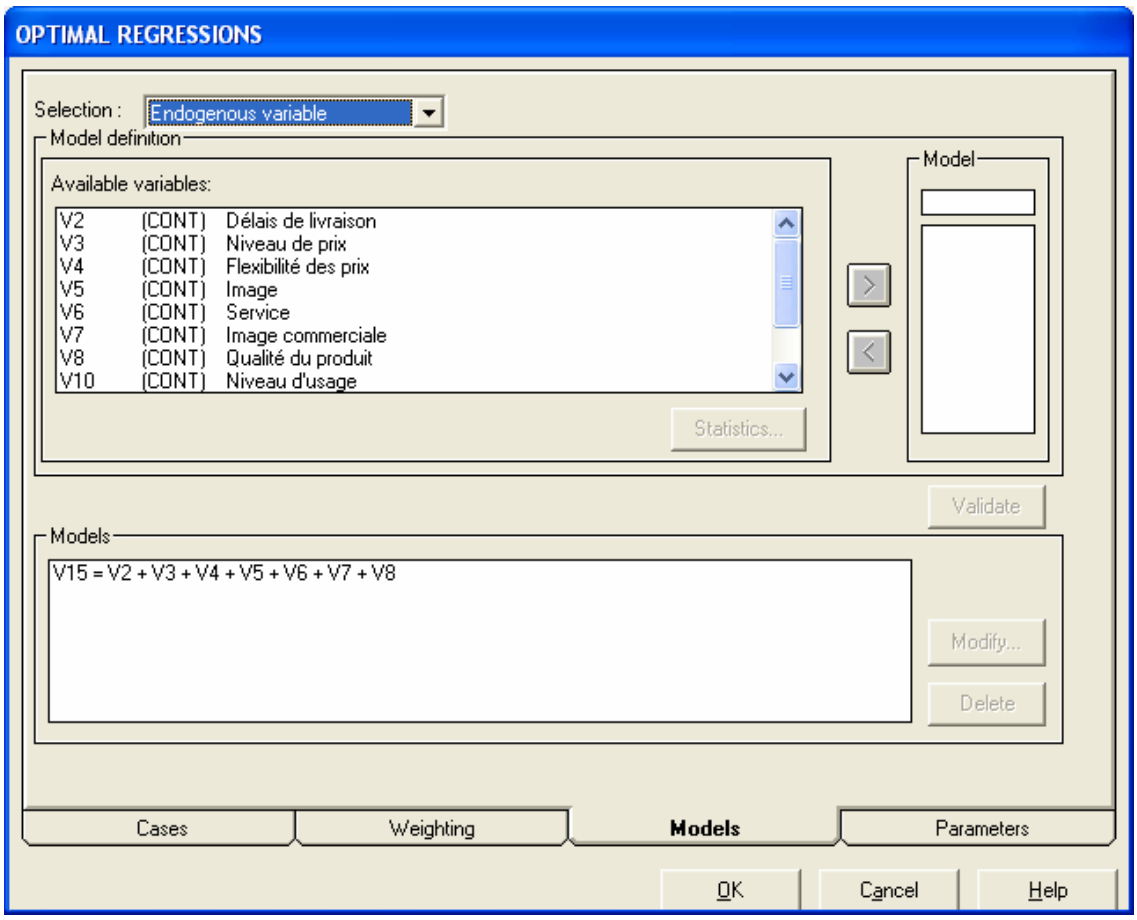
Il s'agit d'un questionnaire administré à 100 personnes pour leur demander de juger leurs fournisseurs. Les critères sont :

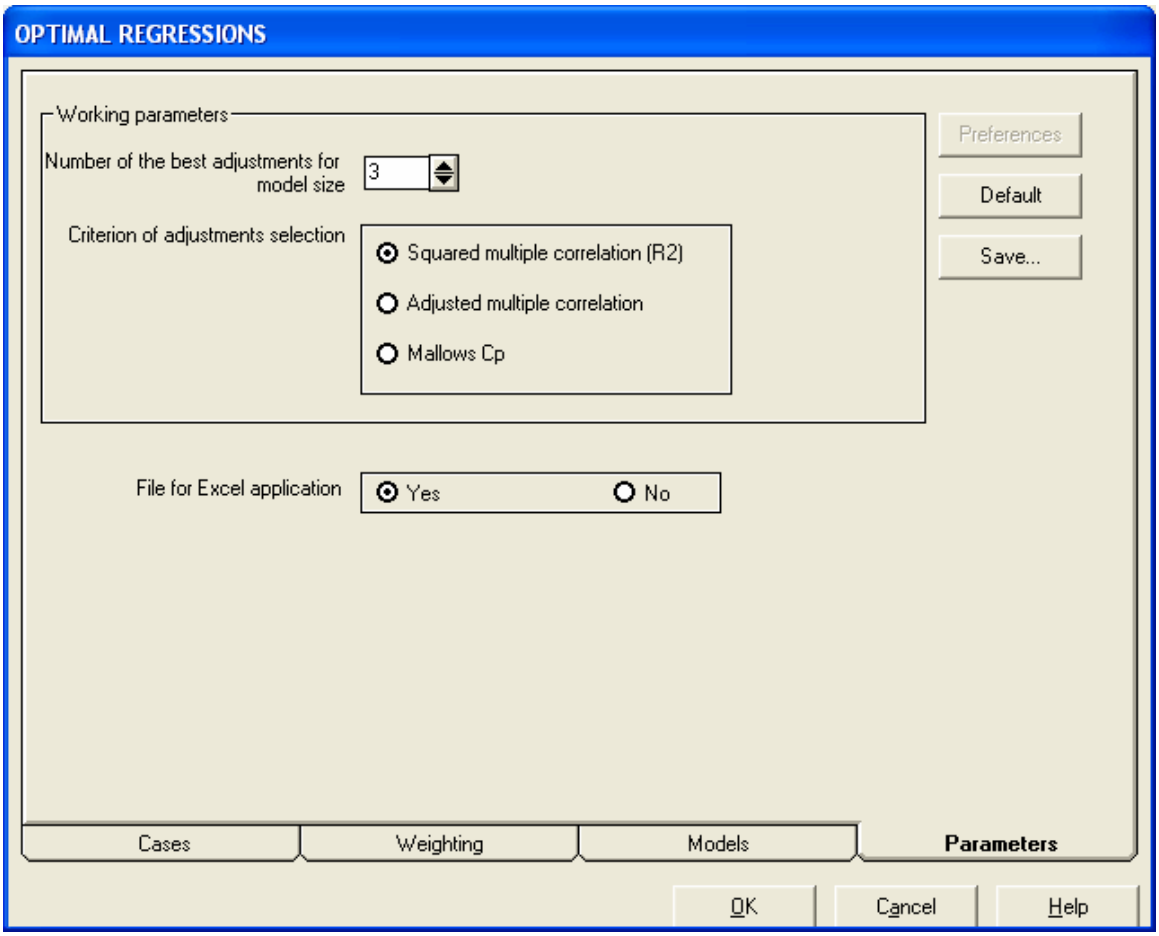
- Les délais de livraison
- Le niveau général des prix
- La flexibilité des prix
- L'image
- Le service
- L'image commerciale
- La qualité du produit

Sur les fournisseurs, nous disposons de plusieurs données dont la taille de l'entreprise en deux catégories : plus ou moins de 50 salariés.

Le but est d'étudier les différences entre les deux groupes.

Id	Taille de l'entreprise	Délais de livraison	Niveau de prix	Flexibilité des prix	Image	Service	Image commerciale	Qualité du produit
1	< 50 salariés	4.1	0.6	6.9	4.7	2.4	2.3	5.2
2	>= 50 salariés	1.8	3	6.3	6.6	2.5	4	8.4
3	>= 50 salariés	3.4	5.2	5.7	6	4.3	2.7	8.2
4	>= 50 salariés	2.7	1	7.1	5.9	1.8	2.3	7.8
5	< 50 salariés	6	0.9	9.6	7.8	3.4	4.6	4.5
6	>= 50 salariés	1.9	3.3	7.9	4.8	2.6	1.9	9.7
7	< 50 salariés	4.6	2.4	9.5	6.6	3.5	4.5	7.6
8	>= 50 salariés	1.3	4.2	6.2	5.1	2.8	2.2	6.9
9	< 50 salariés	5.5	1.6	9.4	4.7	3.5	3	7.6
10	>= 50 salariés	4	3.5	6.5	6	3.7	3.2	8.7
11	< 50 salariés	2.4	1.6	8.8	4.8	2	2.8	5.8
12	< 50 salariés	3.9	2.2	9.1	4.6	3	2.5	8.3
13	>= 50 salariés	2.8	1.4	8.1	3.8	2.1	1.4	6.6
14	< 50 salariés	3.7	1.5	8.6	5.7	2.7	3.7	6.7
15	< 50 salariés	4.7	1.3	9.9	6.7	3	2.6	6.8
16	< 50 salariés	3.4	2	9.7	4.7	2.7	1.7	4.8
17	< 50 salariés	3.2	4.1	5.7	5.1	3.6	2.9	6.2
18	< 50 salariés	4.9	1.8	7.7	4.3	3.4	1.5	5.9
19	< 50 salariés	5.3	1.4	9.7	6.1	3.3	3.9	6.8
20	< 50 salariés	4.7	1.3	9.9	6.7	3	2.6	6.8
21	< 50 salariés	3.3	0.9	8.6	4	2.1	1.8	6.3
22	< 50 salariés	3.4	0.4	8.3	2.5	1.2	1.7	5.2
23	< 50 salariés	3	4	9.1	7.1	3.5	3.4	8.4
24	>= 50 salariés	2.4	1.5	6.7	4.8	1.9	2.5	7.2
25	< 50 salariés	5.1	1.4	8.7	4.8	3.3	2.6	3.8
26	< 50 salariés	4.6	2.1	7.9	5.8	3.4	2.8	4.7
27	>= 50 salariés	2.4	1.5	6.6	4.8	1.9	2.5	7.2
28	< 50 salariés	5.2	1.3	9.7	6.1	3.2	3.9	6.7
29	< 50 salariés	3.5	2.8	9.9	3.5	3.1	1.7	5.4
30	>= 50 salariés	4.1	3.7	5.9	5.5	3.9	3	8.4
31	>= 50 salariés	3	3.2	6	5.3	3.1	3	8
32	< 50 salariés	2.8	3.8	8.9	6.9	3.3	3.2	8.2
33	< 50 salariés	5.2	2	9.3	5.9	3.7	2.4	4.6
34	>= 50 salariés	3.4	3.7	6.4	5.7	3.5	3.4	8.4
35	>= 50 salariés	2.4	1	7.7	3.4	1.7	1.1	6.2
36	>= 50 salariés	1.8	3.3	7.5	4.5	2.5	2.4	7.6
37	>= 50 salariés	3.6	4	5.8	5.8	3.7	2.5	9.3
38	< 50 salariés	4	0.9	9.1	5.4	2.4	2.6	7.3
39	>= 50 salariés	0	2.1	6.9	5.4	1.1	2.6	8.9
40	>= 50 salariés	2.4	2	6.4	4.5	2.1	2.2	8.8
41	>= 50 salariés	1.9	3.4	7.6	4.6	2.6	2.5	7.7
42	< 50 salariés	5.9	0.9	9.6	7.8	3.4	4.6	4.5
43	< 50 salariés	4.9	2.3	9.3	4.5	3.6	1.3	6.2
44	< 50 salariés	5	1.3	8.6	4.7	3.1	2.5	3.7
45	>= 50 salariés	2	2.6	6.5	3.7	2.4	1.7	8.5
46	< 50 salariés	5	2.5	9.4	4.6	3.7	1.4	6.3
47	< 50 salariés	3.1	1.9	10	4.5	2.6	3.2	3.8
48	>= 50 salariés	3.4	3.9	5.6	5.6	3.6	2.3	9.1
49	< 50 salariés	5.8	0.2	8.8	4.5	3	2.4	6.7
50	< 50 salariés	5.4	2.1	8	3	3.8	1.4	5.2
51	< 50 salariés	3.7	0.7	8.2	6	2.1	2.5	5.2
52	>= 50 salariés	2.6	4.8	8.2	5	3.6	2.5	9
53	>= 50 salariés	4.5	4.1	6.3	5.9	4.3	3.4	8.8
54	>= 50 salariés	2.8	2.4	6.7	4.9	2.5	2.6	9.2
55	< 50 salariés	3.8	0.8	8.7	2.9	1.6	2.1	5.6
56	< 50 salariés	2.9	2.6	7.7	7	2.8	3.6	7.7


OPTIMAL REGRESSIONS
Selection :
Endogenous variable
Model definition
Available variables:
V2 (CONT) Délais de livraison
V3 (CONT) Niveau de prix
V4 (CONT) Flexibilité des prix
V5 (CONT) Image
V6 (CONT) Service
V7 (CONT) Image commerciale
V8 (CONT) Qualité du produit
V10 (CONT) Niveau d'usage
Statistics...
Model
Validate
Models
V15 = V2 + V3 + V4 + V5 + V6 + V7 + V8
Modify...
Delete
Cases
Weighting
Models
Parameters
OK
Cancel
Help


OPTIMAL REGRESSIONS
Working parameters
Number of the best adjustments for model size
3
Criterion of adjustments selection
Squared multiple correlation (R2)
Adjusted multiple correlation
Mallows Cp
Preferences
Default
Save...
File for Excel application
Yes
No
Cases
Weighting
Models
Parameters
OK
Cancel
Help

Fuwil – 4

On présente les moyennes générales et par groupe pour chaque variable, ainsi que le décompte du nombre de valeurs manquantes.

La colonne « moyenne intra groupe » donne les moyennes, pour chaque variable dans les groupes 1 et 2. Par défaut, le groupe libellé « 1 » est celui qui a la plus faible valeur dans le codage de V9, le groupe libellé « 2 » est celui qui a la plus forte valeur.
La colonne « moyenne générale » donne les moyennes pour chaque variable sur l'échantillon total.

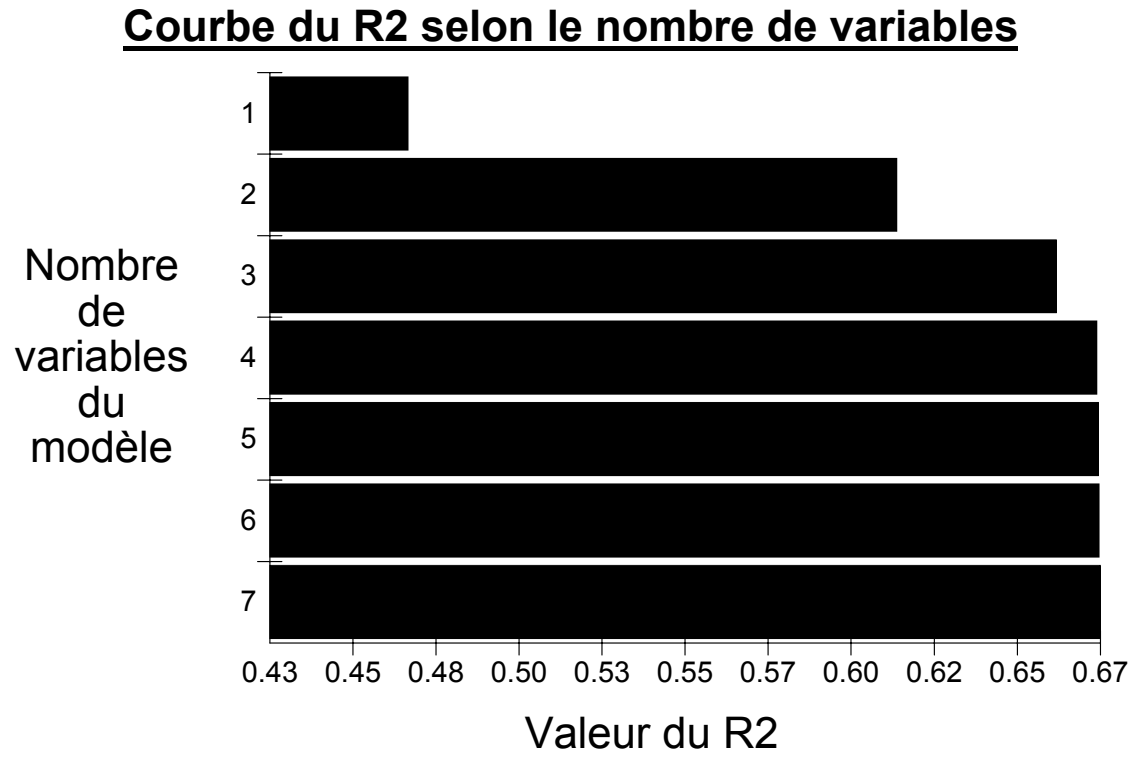
Gestion des données manquantes des variables exogènes
Les valeurs manquantes sont remplacées par les moyennes intra-groupes

Groupe	Libellé de la variable	Moyenne intra groupe	Moyenne générale	Nombre de valeurs manquantes
1	Délais de livraison	4.192	3.515	0
1	Niveau de prix	1.948	2.364	0
1	Flexibilité des prix	8.622	7.894	0
1	Image	5.213	5.248	0
1	Service	3.050	2.916	0
1	Image commerciale	2.692	2.665	0
1	Qualité du produit	6.090	6.971	0
2	Délais de livraison	2.500	3.515	0
2	Niveau de prix	2.988	2.364	0
2	Flexibilité des prix	6.803	7.894	0
2	Image	5.300	5.248	0
2	Service	2.715	2.916	0
2	Image commerciale	2.625	2.665	0
2	Qualité du produit	8.293	6.971	0

CRITERE du R^2
Courbe du R^2 en fonction du nombre de variables

Ce graphique récapitule l'évolution de l'indice R^2 en fonction du nombre de variables rentrées dans les modèles. Plus cet indice est élevé, meilleur est l'ajustement.
Bien sûr, il faut tenir compte du nombre de variables entrées dans le modèle : le R^2 augmente avec le nombre de variables entrées dans le modèle. Nous verrons plus tard d'autres critères de choix disponibles dans SPAD : le R^2 ajusté et le C(p) de Mallow.
Au vu du graphique l'augmentation du critère R^2 est tangible jusqu'à 4 variables ; après l'augmentation est beaucoup plus faible : les variables suivantes sont redondantes et n'apportent finalement rien aux modèles.

Le R^2 peu ici s'interpréter comme la part de variance 'expliquée' par le modèle. Il varie de 0 à 1.



Les sorties **1 var** à **7 vars** présentent les 3 meilleurs ajustements au sens du R^2 pour des modèles de 1 à 7 variables

1 var

Ce tableau présente les 3 meilleurs ajustements au sens du R² pour des modèles avec une seule variable.

Adjustments with 1 variable + constante DDL(Student) = 98
Adjustment 1 (Full printout)
R2 = 0.4680**
Fisher = 86.1999
Probability = 0.0000
Test-Value = 7.845

Variable label	Coefficient	Student	Probability	Test-Value
Qualité du produit	0.2125	9.28	0.000	7.85

Adjustment 2 (Full printout)
R2 = 0.4173**
Fisher = 70.1913
Probability = 0.0000
Test-Value = 7.258

Variable label	Coefficient	Student	Probability	Test-Value
Flexibilité des prix	-0.2294	8.38	0.000	7.26

Adjustment 3 (Full printout)
R2 = 0.3977**
Fisher = 64.7155
Probability = 0.0000
Test-Value = 7.032

Variable label	Coefficient	Student	Probability	Test-Value
Délais de livraison	-0.2351	8.04	0.000	7.03

Le nombre de degré de liberté de l'erreur est de 98.

Le premier ajustement est le meilleur, avec un indice R² de 0.468 ; c'est à dire que la variance *expliquée par le modèle* représente 46.8% de la variance totale. Un modèle qui ne différencierait pas les deux groupes aurait un R² de 0. Cet indice est inférieur à 1.

La statistique de Fisher correspond au test de validation globale du modèle. Cette statistique suit une loi de Fisher à 1 et 98 degrés de liberté. Elle a une valeur de 86.2 correspond une probabilité inférieure à 1/10000 (0.0000). On accepte donc le modèle. Cette probabilité est exprimée en valeur-test, on a ici 7.85.

La colonne coefficient présente l'estimation du coefficient de « Qualité du produit » : le modèle s'écrit : D = constante – 0.4337 x Qualité du produit.

La colonne Student teste la nullité du coefficient pour « Qualité du produit » : cette statistique suit une loi de Student à 98 degrés de liberté ; à cette valeur de 9.28 correspond une probabilité inférieure à 1/10000 (0.0000). Le coefficient est significatif.

Cette probabilité est exprimée en valeur-test, on a ici 7.85. Puisque le modèle a un coefficient, la valeur test associée au coefficient égale celle associée au modèle globalement.

6 vars

Adjustments avec 6 variables + constante DDL(Student) = 93

Adjustment 1 (Full printout)
R2 = 0.6718**
Fisher = 31.7290
Probability = 0.0000
Test-Value = 9.210

Variable label	Coefficient	Student	Probability	Test-Value
Délais de livraison	-0.1472	1.12	0.264	1.12
Niveau de prix	-0.0608	0.45	0.656	0.44
Flexibilité des prix	-0.1185	4.40	0.000	4.18
Service	0.1131	0.45	0.657	0.44
Image commerciale	-0.0743	1.85	0.067	1.83
Qualité du produit	0.1378	5.90	0.000	5.42

Adjustment 2 (Full printout)
R2 = 0.6716**
Fisher = 31.6986
Probability = 0.0000
Test-Value = 9.207

Variable label	Coefficient	Student	Probability	Test-Value
Délais de livraison	-0.0912	3.27	0.002	3.17
Niveau de prix	-0.0034	0.11	0.910	0.11
Flexibilité des prix	-0.1167	4.33	0.000	4.12
Image	0.0161	0.37	0.711	0.37
Image commerciale	-0.0898	1.44	0.152	1.43
Qualité du produit	0.1367	5.87	0.000	5.40

Adjustment 3 (Full printout)
R2 = 0.6716**
Fisher = 31.6925
Probability = 0.0000
Test-Value = 9.206

Variable label	Coefficient	Student	Probability	Test-Value
Délais de livraison	-0.0904	2.35	0.021	2.31
Flexibilité des prix	-0.1160	4.34	0.000	4.13
Image	0.0155	0.36	0.722	0.36
Service	-0.0014	0.02	0.980	0.02
Image commerciale	-0.0897	1.44	0.153	1.43
Qualité du produit	0.1362	5.88	0.000	5.41

Les deux ajustements présentés dans ces tableaux ont chacun 6 variables.

En plus des renseignements déjà vus sur les tableaux précédents, nous avons ici le détail pour chaque variable rentrée dans le modèle : coefficient, test de Student, probabilité associée et valeur test associée.

Pour le modèle 1 par exemple, seules les variables « Flexibilité des prix » et « Qualité du produit » sont significatives à 5% (probabilité que le coefficient associé soit nul inférieure à 5%). On retrouve ces deux critères qui se détachaient déjà dans les modèles à une variable.

3 Vars

Adjustments avec 3 variables + constante DDL(Student) = 96

Adjustment 1 (Full printout)
R2 = 0.6591**
Fisher = 61.8788
Probability = 0.0000
Test-Value = 9.660

Variable label	Coefficient	Student	Probability	Test-Value
Délais de livraison	-0.0995	3.64	0.000	3.51
Flexibilité des prix	-0.1161	4.55	0.000	4.32
Qualité du produit	0.1270	5.79	0.000	5.35

Adjustment 2 (Full printout)
R2 = 0.6392**
Fisher = 56.6932
Probability = 0.0000
Test-Value = 9.378

Variable label	Coefficient	Student	Probability	Test-Value
Flexibilité des prix	-0.1478	6.06	0.000	5.56
Service	-0.1081	2.68	0.009	2.63
Qualité du produit	0.1517	7.12	0.000	6.36

Adjustment 3 (Full printout)
R2 = 0.6338**
Fisher = 55.3919
Probability = 0.0000
Test-Value = 9.303

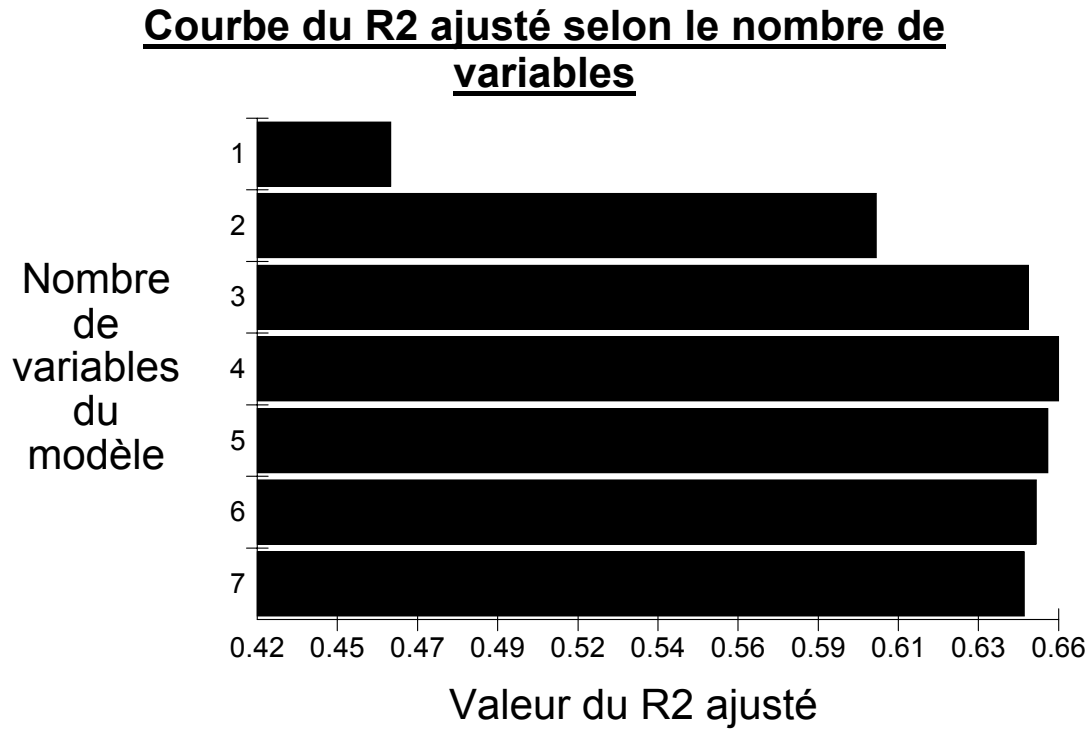
Variable label	Coefficient	Student	Probability	Test-Value
Flexibilité des prix	-0.1478	6.02	0.000	5.53
Image commerciale	-0.0957	2.38	0.019	2.34
Qualité du produit	0.1628	7.46	0.000	6.61

Finalement, les meilleurs ajustements seraient à chercher dans les solutions à 3 au 4 variables, pour lesquelles tous les coefficients sont significatifs et les valeurs–tests associées aux modèles les plus fortes.

Cela se confirme ci-dessous avec les critères du R2 et du Cp de Mallows qui font ressortir l'ajustement à 3 ou 4 variables.

CRITERE DU R² ajusté
Courbe du R² ajusté en fonction du nombre de variables

Le critère R² ajusté pénalise le R² classique en fonction du nombre de variables entrées dans le modèle : pour continuer à augmenter cet indice l'ajout d'une nouvelle variable doit être suffisant (elle ne doit pas être trop redondante avec les variables déjà rentrées, sinon l'indice baisse).
Le graphique indique clairement que les meilleurs modèles se situent dans la zone à 3 ou 4 variables explicatives.



4 vars

Le premier modèle à 4 variables nous donne :

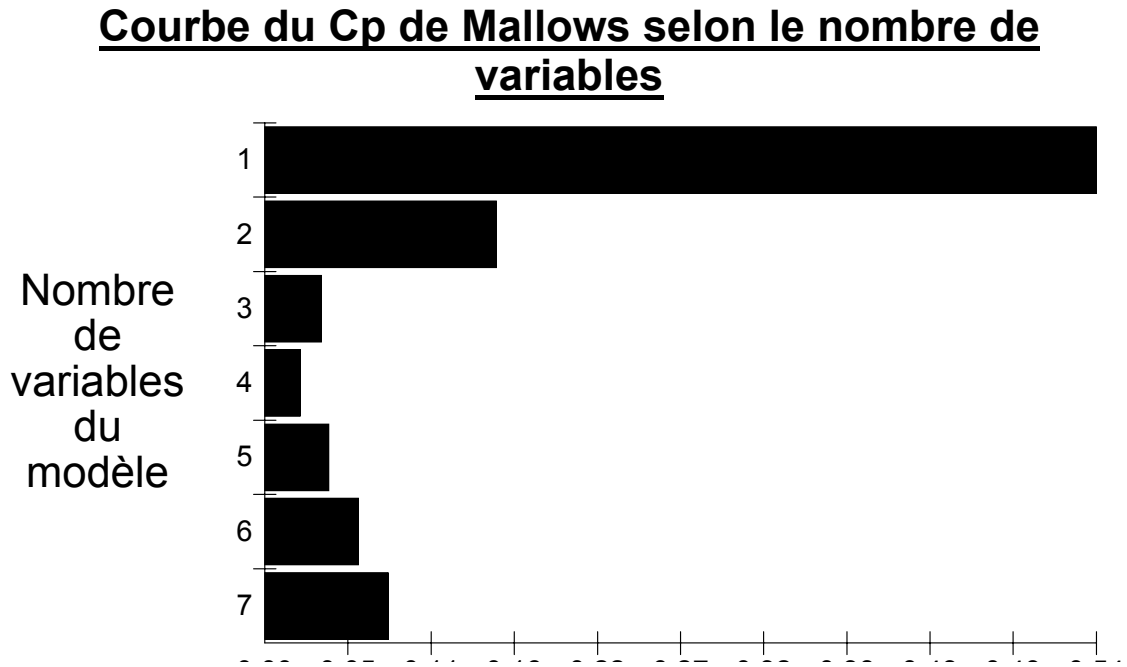
Adjustments avec 4 variables + constante DDL(Student) = 95
Adjustment 1 (Full printout)
R2 = 0.6711**
Fisher = 48.4607
Probability = 0.0000
Test-Value = 9.610

Variable label	Coefficient	Student	Probability	Test-Value
Délais de livraison	-0.0901	3.28	0.001	3.18
Flexibilité des prix	-0.1171	4.64	0.000	4.40
Image commerciale	-0.0723	1.86	0.066	1.84
Qualité du produit	0.1366	6.13	0.000	5.61

Le R² ajusté est de 0.6574 ; contre un R² 'classique' de 0.6711

CRITERE DU Cp de Mallow
Courbe du Cp de Mallow en fonction du nombre de variables

Cet indice est à minimiser : on retrouve comme précédemment que les meilleurs modèles sont dans la zone « 3 à 4 variables ».



4 vars

Ajustements avec 4 variables + constante DDL(Student) = 95
Ajustement 1 (Edition complète)
C(P) = 2.2916
Fisher = 48.4607
Probabilité = 0.0000
Valeur-Test = 9.610

Libellé de la variable	Coefficient	Student	Probabilité	Valeur-Test
Délais de livraison	0.1840	3.28	0.001	3.18
Flexibilité des prix	0.2390	4.64	0.000	4.40
Image commerciale	0.1476	1.86	0.066	1.84
Qualité du produit	-0.2788	6.13	0.000	5.61

Annexes : Formules des indices R², R² ajuste et Cp de Mallow

1. ***R² :***

$$R^2 = 1 - \frac{SSE}{SST}$$

Le R² est la somme des carrés 'expliqués' par le modèle, c'est à dire 1 moins le ratio entre la somme des carrés des erreurs(SSE) et la somme des carrés totaux (SST).

2. ***R² ajusté :***

$$R^2 = 1 - \frac{(n-1)(1-R^2)}{(n-p)}$$

Le R² ajusté corrige le R² en fonction du nombre de variables entrées dans le modèle (p).

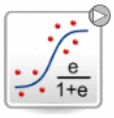
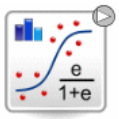
3. ***CP de Mallow C(p) :***

$$C(p) = \frac{SSE}{SST} + 2p - n$$

Le C(p) de Mallow est lié positivement à l'erreur (SSE) et au nombre de variables entrées dans le modèle : un modèle avec beaucoup de variables ou une erreur importante sera pénalisé par cet indice.

Références :

- Furnival, G.M. and Wilson, R.W. (1974), "Regression by Leaps and Bounds" *Technometrics*, 16, 499 -511.

Régression Logistique

La régression logistique étudie la liaison entre une variable nominale à expliquer Y à k modalités et p variables explicatives de nature quelconque X_1, X_2,..., X_p.
Le cas d'une variable à expliquer présentant k = 2 modalités est la situation la plus répandue. Elle correspond à la discrimination entre deux groupes. On se limitera à l'étude de ce cas.

INTRODUCTION DU LOGIT

On a effectué une expérience médicale afin de mesurer les effets secondaires, en particulier les maux d'estomac provoqué par un nouveau traitement.
On dispose d'un échantillon d'individus ayant reçu soit le nouveau traitement ($X_1 = 1$) soit l'ancien ($X_1 = 2$). On sait, de plus, si l'individu a des antécédents de maux d'estomac ($X_2 = 1$) ou non ($X_2 = 2$).
On divise alors l'échantillon en deux groupes: les individus souffrant de maux d'estomac durant le traitement (Y = 1) et les individus n'en souffrant pas (Y = 2).
Les variables X_1 et X_2 sont les variables « explicatives » de l'étude. Ce sont des variables nominales binaires ou dichotomiques. La variable Y est la variable de groupe à deux modalités.
La régression logistique modélise la probabilité d'un événement. Dans l'exemple, on cherche à estimer la probabilité qu'un individu souffre de maux d'estomac connaissant les caractéristiques des variables X_1 et X_2, c'est-à-dire la probabilité a posteriori suivante

$$P(Y=1/X_1,X_2)$$

On note pour l'exemple la relation suivante:

$$P(Y=1/X_1,X_2)+P(Y=2/X_1,X_2)=1$$

Pour modéliser cet événement, pourquoi ne pas choisir la régression elle-même?
Cela reviendrait, par exemple, à poser que la probabilité de souffrir de maux d'estomac pendant le traitement chez un sujet qui possède les caractéristiques X_1 et X_2 peut s'écrire:

$$P[Y=1/X_1,X_2]=\beta_0+\beta_1X_1+\beta_2X_2$$

On note

$$P[Y=1/X_1,X_2]=P$$

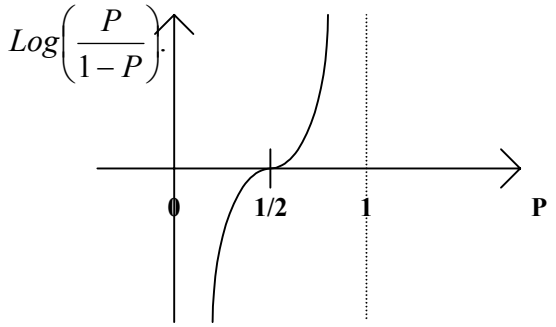
Ce modèle implique que chez les sujets avec antécédents, la probabilité considérée est augmentée de la quantité β_2 par rapport aux sujets sans antécédents et ce, quelles que soient leurs autres caractéristiques.
Si cette approche, qui n'est pas a priori déraisonnable, n'est quasiment pas utilisée, c'est qu'elle présente plusieurs inconvénients.
Ce modèle pourrait prédire des probabilités négatives ou supérieures à 1. En effet le domaine de variation de la combinaison linéaire des X_i est la droite réelle et P est une probabilité, donc variant entre 0 et 1. Cet inconvénient peut être évité si, au lieu de considérer que la probabilité est une fonction linéaire des paramètres, on pose que c'est une certaine transformation de la probabilité qui est fonction linéaire des paramètres.

Le logit de la probabilité P est le logarithme du quotient $\frac{P}{1-P}$:

$$\text{Logit}\,(P) = \text{Log}\left(\frac{P}{1-P}\right) \qquad (1)$$

Il permet de transformer un intervalle de la droite réelle en l'intervalle [0.1] en respectant les axiomes de probabilité.

Graphique 1
Représentation graphique du logit de P



MODELE LOGISTIQUE AVEC DES VARIABLES EXPLICATIVES BINAIRES

Pour l'exemple, le modèle logistique s'écrit:

$$\text{Log}\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \qquad (2)$$

Le logit de la probabilité est une fonction linéaire des caractéristiques des sujets mais la probabilité elle-même est une fonction non linéaire de ces caractéristiques. En effet, d'après (2)

$$P = \frac{\exp\left(\beta_0 + \beta_1 X_1 + \beta_2 X_2\right)}{1 + \exp\left(\beta_0 + \beta_1 X_1 + \beta_2 X_2\right)}$$

Le modèle (IX.2) est un modèle additif pour des variables explicatives X_i *binaires* (les modèles avec variables qualitatives à plus de deux modalités et prise en compte des interactions sont décrits plus loin).
Il indique que la force de la liaison entre chacune des variables explicatives X_i et la variable de groupe Y ne dépend pas des valeurs prises par les autres variables explicatives (différentes de X_i): il s'agit du modèle dont le logit est exprimé en fonction des seuls effets principaux (absence d'interactions) et qui est déterminé par un vecteur de paramètres **b** de longueur 3.

$$\mathbf{b} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

Dans le modèle (2) la variable X_1 représente le traitement. Pour savoir si la fréquence des maux d'estomac diffère entre les deux groupes, il suffit de tester l'hypothèse

$$\beta_1 = 0$$

Un autre modèle permettrait de répondre à la question « le nouveau traitement favorise-t-il la fréquence des maux d'estomac chez les patients déclarant des antécédents mais non chez les autres sujets? ». Il suffirait d'introduire un terme d'interaction entre les variables traitements et antécédents.

MODELE LOGISTIQUE AVEC UNE VARIABLE EXPLICATIVE NOMINALE A PLUS DE DEUX MODALITES

Une variable explicative nominale pour laquelle il n'existe pas de relation d'ordre entre les modalités doit être codée de manière spécifique.

Il est clair que si les modalités sont identifiées par un symbole numérique (1, 2, 3, 4 si la variable présente quatre modalités), ces chiffres n'ont pas de signification quantitative.

La variable explicative à plus de deux modalités doit être recodée avant son introduction dans le modèle en plusieurs variables binaires prenant les valeurs 0 et 1 connues sous le nom de variables "design".
On introduit autant de variables qu'il y a de modalités.

Si la variable explicative est nominale à k modalités le problème suivant se pose: les k variables binaires introduites pour la représenter ne sont pas indépendantes, puisque leur somme vaut 1 quel que soit l'individu i.

Une solution consiste à éliminer une des modalités. Cette modalité non introduite dans le modèle a donc un coefficient égal à 0 par convention. On peut considérer qu'elle représente une situation de référence, par rapport à laquelle on mesure des déviations.

Mathématiquement, le choix de cette situation de référence n'a aucune importance. On peut par exemple choisir comme situation de référence la situation « modale » (modalité ayant le plus grand effectif).

Soit Y la variable réponse à deux modalités 1 et 2. Considérons Z une variable explicative nominale à quatre modalités représentant la race de l'individu
Z = 1: individu de race blanche
Z = 2: individu de race noire
Z = 3: individu hispanique
Z = 4: autres

Si on choisit pour race de référence la race blanche, la matrice D est la suivante:

Les trois colonnes de D (D2, D3, D4) correspondent au codage de Z en variables à inclure dans le modèle.

TABLEAU 1
Construction de la matrice D

RACE (Modalité)	D2	D3	D4
Blancs (1)	0	0	0
Noirs (2)	1	0	0
Hispaniques (3)	0	1	0
Autres (4)	0	0	1

Le modèle logistique s'écrit alors

$$\text{Logit}\begin{vmatrix} P(Y=1/Z=1) \\ P(Y=1/Z=2) \\ P(Y=1/Z=3) \\ P(Y=1/Z=4) \end{vmatrix} = \underbrace{\begin{vmatrix} 1 \\ 1 \\ 1 \\ 1 \end{vmatrix}}_{D_0} \beta_0 + \underbrace{\begin{vmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{vmatrix}}_{D} \begin{vmatrix} \beta_2 \\ \beta_3 \\ \beta_4 \end{vmatrix}$$

Ainsi, si la variable explicative Z est une variable nominale à k modalités, elle sera transformée en (k-1) variables indicatrices.

Ces variables sont notées d_u. Si la première modalité est la modalité de référence, le logit s'écrit:

$$\text{Logit}\left[P(Y=1/Z)\right] = \beta_0 + \sum_{u=2}^{k} \beta_u d_u$$

Pour l'exemple, on obtient

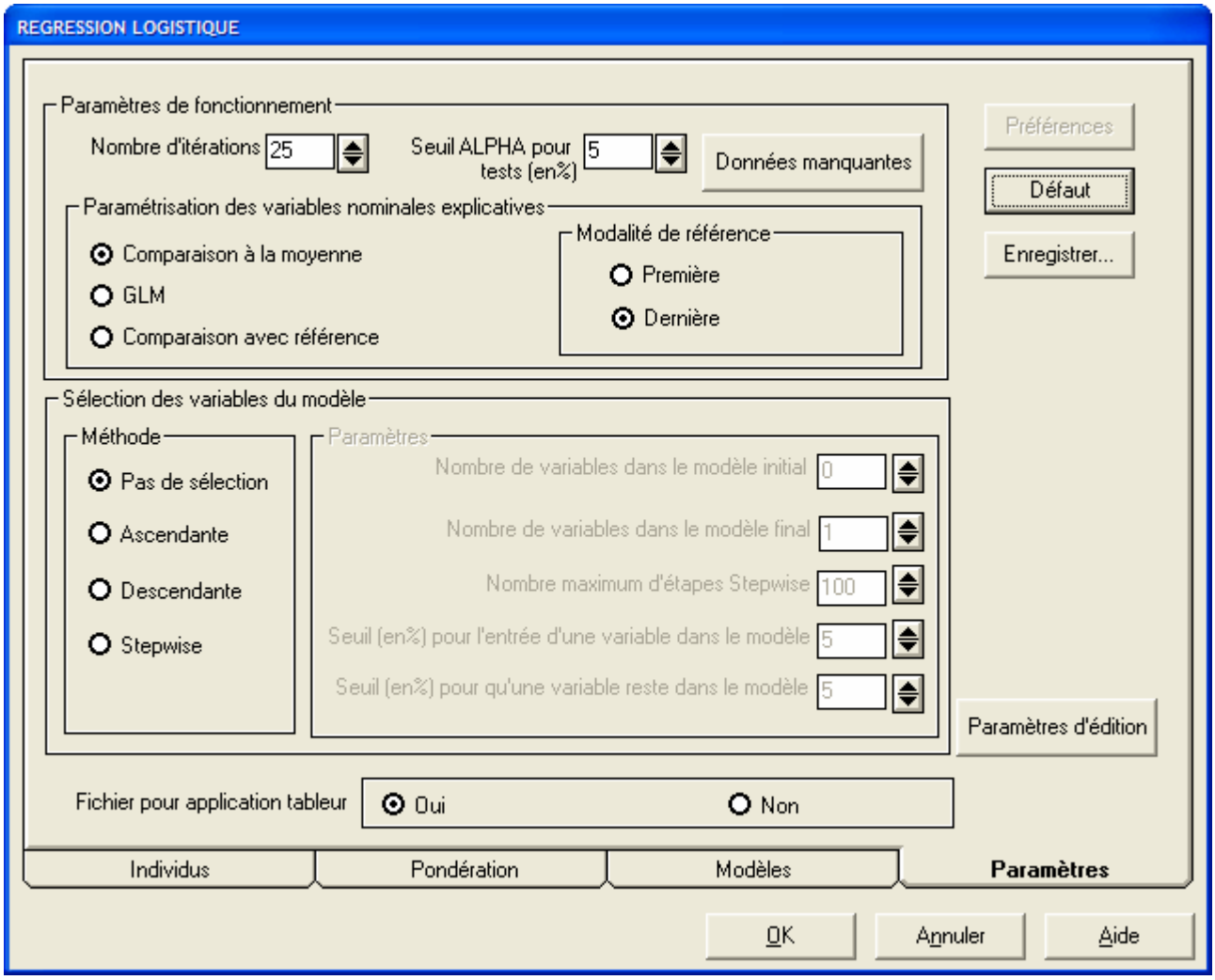
$$\text{Logit}\left[P(Y=1/Z)\right] = \beta_0 + \beta_2 d_2 + \beta_3 d_3 + \beta_4 d_4$$

avec

$d_u = 1$ si $Z = u$

$d_u = 0$ sinon

LA REGRESSION LOGISTIQUE AVEC SPAD



Nombre maximum d'itération :

Pour l'estimation des coefficients de la régression logistique, on fait appel à une recherche de maximum de vraisemblance. Cette procédure de recherche est itérative et s'arrêt lorsque les estimations se stabilisent c'est-à-dire si l'évolution d'une itération à l'autre est inférieure à un seuil fixé à 10^{-8}.

Si le processus est trop long à converger, on peut, grâce à ce paramètre, modifier le nombre maximum de ces itérations.

Seuil alpha pour les tests (en %) :

C'est le seuil exprimé en pourcentage utilisé pour établir l'intervalle de confiance des Odds-ratios estimés. Cet intervalle de confiance est égal à 100-alpha.

Paramétrisation des variables nominales explicatives :

Si le modèle contient des variables nominales (qualitatives) dans les facteurs explicatifs, le nombre de paramètres à estimer est supérieur au nombre de paramètres estimables, c'est-à-dire au nombre de degrés de liberté du modèle.

Pour pallier ce type de problème, trois solutions sont proposées :

Comparaison à la moyenne :

Pour chaque variable nominale explicative du modèle, l'estimation de l'effet de chaque modalité (niveau) à l'exception de la modalité de référence est faire par comparaison à l'effet moyen de tous les niveaux y compris la modalité de référence.

GLM :

On adopte la solution utilisé dans une analyse de la variance qui consiste à fixer arbitrairement à zéro le coefficient associé à la dernière modalité.

Comparaison à une référence :

Pour chaque variable nominale explicative du modèle, l'estimation de l'effet est faite par comparaison au niveau de référence désigné (la modalité de référence est soit la première soit la dernière).

Sélection des variables du modèle :

Le mode de sélection des variables explicatives dans le modèle n'est disponible que si les facteurs explicatifs ne contiennent que des facteurs simples (pas d'interactions).

Pas de sélection

Le modèle est calculé avec toutes les variables, c'est l'option par défaut.

Ascendant

Dans une sélection ascendante, les variables sont introduites une à une dans le modèle dans l'ordre de leur pouvoir explicatif. La sélection s'arrête lorsqu'aucune variable candidate n'est suffisamment explicative ou lorsque le nombre de variables souhaitées dans le modèle final est atteint.

Une variable entre dans le modèle si elle apporte une information significative selon le Khi² résiduel. Par défaut le seuil de la probabilité associé à ce Khi² est de 5%. Ce seuil est modifiable dans le champ « Seuil (en %)pour l'entrée d'une variable dans le modèle ». En l'augmentant on est plus tolérant pour l'entrée d'une variable, en le diminuant on est plus contraignant.

Par défaut, le nombre de variables dans le modèle initial est égal à 0 et le nombre de variables dans le modèle final au nombre de facteurs simples du modèle. Ces paramètres sont modifiables en respectant le fait que le nombre de variables dans le modèle final soit supérieur au nombre de variables dans le modèle initial. Si le nombre de variables dans le modèle initial est supérieur à 0, les variables mises de fait dans le modèle initial sont les premières variables explicatives de la définition du modèle.

Descendant

Dans une sélection descendante, les variables sont retirées à une à une si leur pouvoir explicatif n'est pas suffisant. L'élimination des variables s'arrête quand la variable la moins explicative l'est malgré tout ou si le nombre de variables dans le modèle final est atteint.

Une variable est retirée du modèle si elle apporte une information non significative selon le Khi² résiduel. Par défaut le seuil de la probabilité associé à ce Khi² est de 5% + Epsilon. Ce seuil est modifiable dans le champ « Seuil (en %) pour qu'une variable reste dans le modèle »,.en l'augmentant on est plus tolérant dans le maintien d'une variable dans le modèle, en le diminuant on est plus contraignant (une variable sortira plus vite).

Par défaut le nombre de variables dans le modèle initial est égal au nombre au nombre de facteurs simples et le nombre de variables dans le modèle final à 1. Ces paramètres sont modifiables en respectant le fait que le nombre de variables dans le modèle final soit inférieur au nombre de variables dans le modèle initial.

Stepwise

La sélection stepwise est une sélection ascendante avec après l'introduction d'une nouvelle variable, la remise en cause des variables déjà introduites dans la mesure ou la nouvelle variable peut remettre en cause le pouvoir explicatif de certaines variables déjà présentes. Dans une sélection stepwise les seuils d'entrée et de sortie sont fixés par défaut à 5% et 5% + Epsilon.

Par défaut le nombre de variables dans le modèle initial est égal à 0 et le nombre d'étapes stepwise à 100. Si le nombre de variables dans le modèle initial est supérieur à 0, les variables mises de fait dans le modèle initial sont les premières variables explicatives de la définition du modèle.

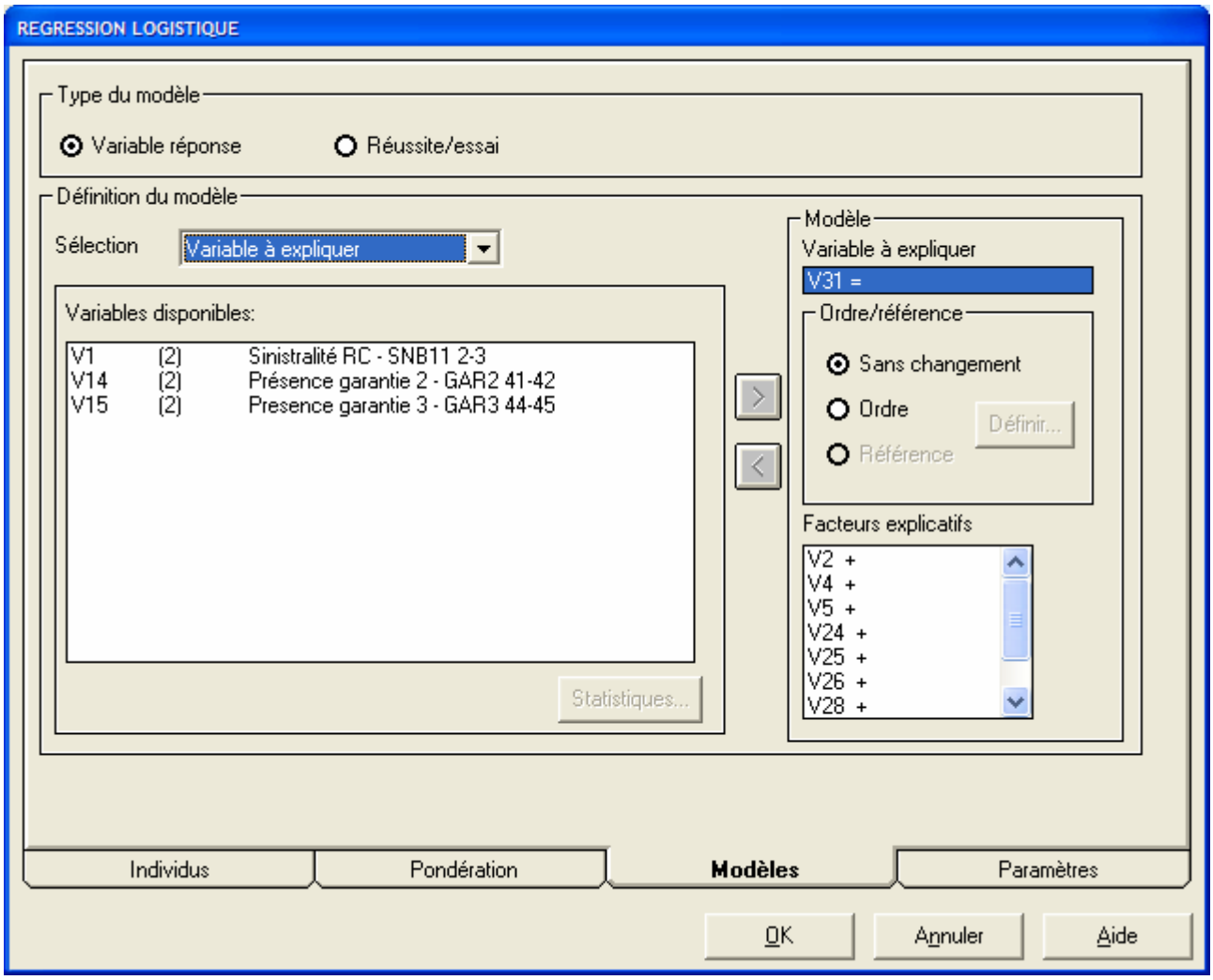
EXEMPLE SUR LA BASE ASSUR.SBA

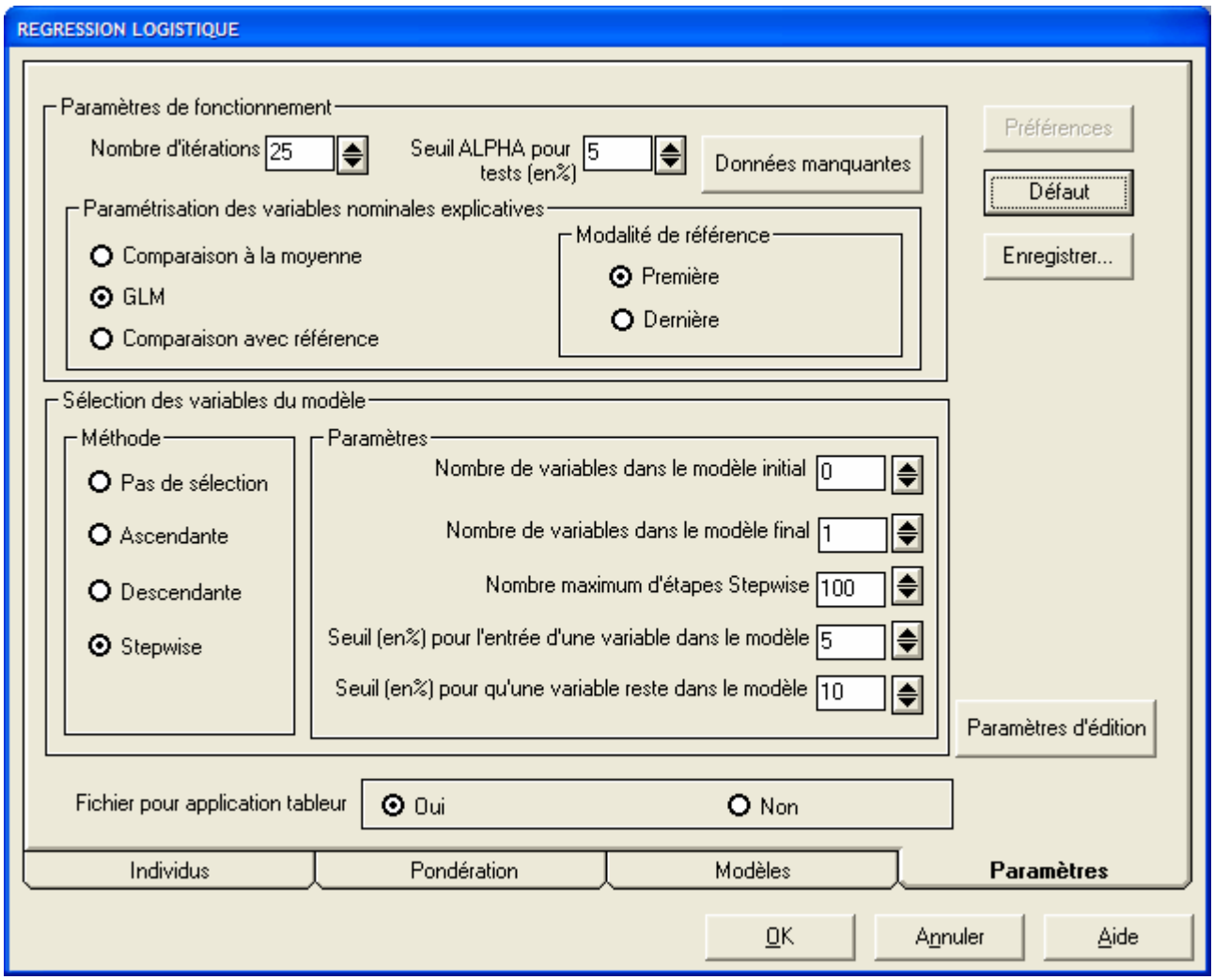
Variable nominale à expliquer :

V 31 . Sinistralité (> 1 sin – 0 sin)	2 modalités

Variables nominales explicatives :

V 2 . Code usage - CUSA 5-6	2 modalités
V 4 . Sexe - SEXE 11-12	3 modalités
V 5 . Code Langue - CLAN 14-15	2 modalités
V 24 . Age de l'assuré (3 mod) - DNAI 8-9	3 modalités
V 25 . Code postal souscripteur (2 mod) - POSS2 17-18	2 modalités
V 26 . Bonus-malus Année -1 (2 mod) - GBM1	2 modalités
V 27 . Date effet Police (2 mod) - DPEP 26-27	2 modalités
V 28 . Puissance du véhicule (2 mod) - PUIS 32-33	2 modalités
V 29 . Année de construction du véhicule (2 mod) - DCOS 38-39	2 modalités




REGRESSION LOGISTIQUE
Paramètres de fonctionnement
Nombre d'itérations 25
Seuil ALPHA pour tests (en%) 5
Données manquantes
Paramétrisation des variables nominales explicatives
Comparaison à la moyenne
GLM
Comparaison avec référence
Modalité de référence
Première
Dernière
Préférences
Défaut
Enregistrer...
Sélection des variables du modèle
Méthode
Pas de sélection
Ascendante
Descendante
Stepwise
Paramètres
Nombre de variables dans le modèle initial 0
Nombre de variables dans le modèle final 1
Nombre maximum d'étapes Stepwise 100
Seuil (en%) pour l'entrée d'une variable dans le modèle 5
Seuil (en%) pour qu'une variable reste dans le modèle 10
Paramètres d'édition
Fichier pour application tableur
Oui
Non
Individus
Pondération
Modèles
Paramètres
OK
Annuler
Aide

REGRESSION LOGISTIQUE

PRESENTATION DU MODELE
DEFINITION DU MODELE

```
  VARIABLE REPONSE ..... : Sinistralité
  NOMBRE DE NIVEAUX DE REPONSE . :      2
  NOMBRE D'OBSERVATIONS ........ :   1106
  FONCTION DE LIEN ..... : LOGIT BINAIRE
  TECHNIQUE D'OPTIMISATION ..... : SCORES DE FISHER
```

PROFIL DE REPONSE

```
  VARIABLE REPONSE : Sinistralité
  ==========================
  ORDRE   REPONSE  FREQUENCE
  --------------------------
      1   > 1 sin        550
      2     O sin        556
  ==========================
CETTE REGRESSION MODELISE LA PROBABILITE QUE : Sinistralité = > 1 sin
PROCEDURE LA SELECTION STEPWISE
```

PROCEDURE DE SELECTION PAS-A-PAS

ETAPE 0 : ENTREE DE L'ORDONNEE A L'ORIGINE

```
LE CRITERE DE CONVERGENCE (.1E-07) EST SATISFAIT

TEST DU KHI2 RESIDUEL
=========================
KHI-CARRE  DDL  PR > KHI2
-------------------------
 708.4096   11   < 0.0001
=========================
```

COEFFICIENTS DE REGRESSION ESTIMES PAR MAXIMUM DE VRAISEMBLANCE

```
================================================================================================================
================================
PARAMETRE                                                DDL    ESTIMATION  ERREUR STAND  KHI2 DE
WALD  PROB > KHI2  EXP(ESTIM.)
----------------------------------------------------------------------------------------------------------------
--------------------------------
Intercept                                                  1       -0.0109        0.0601
0.0325       0.8568       0.9892
================================================================================================================
================================
```

ETAPE 1 : ENTREE DU FACTEUR Bonus-malus Année -1 (2 mod) - GBM1

ANALYSE DE TYPE III DES FACTEURS

```
================================================================================
FACTEUR                                          DDL  KHI2 DE WALD  PROB > CHISQ
--------------------------------------------------------------------------------
Bonus-malus Année -1 (2 mod) - GBM1                1      399.7544      < 0.0001
================================================================================
```

ETAPE 2 : ENTREE DU FACTEUR Année de construction du véhicule (2 mod) - DCOS 38-39

ANALYSE DE TYPE III DES FACTEURS

```
================================================================================
FACTEUR                                          DDL  KHI2 DE WALD  PROB > CHISQ
--------------------------------------------------------------------------------
Bonus-malus Année -1 (2 mod) - GBM1                1      363.2370      < 0.0001
Année de construction du véhicule (2 mod) - DCOS 38-39     1       57.9012      < 0.0001
================================================================================
```

ETAPE 3 : ENTREE DU FACTEUR Age de l'assuré (3 mod) - DNAI 8-9

ANALYSE DE TYPE III DES FACTEURS

FACTEUR	DDL	KHI2 DE WALD	PROB > CHISQ
Bonus-malus Année -1 (2 mod) - GBM1	1	280.1464	< 0.0001
Année de construction du véhicule (2 mod) - DCOS 38-39	1	51.6761	< 0.0001
Age de l'assuré (3 mod) - DNAI 8-9	2	47.0178	< 0.0001

ETAPE 4 : ENTREE DU FACTEUR Code postal souscripteur (2 mod) - POSS2 17-18

ANALYSE DE TYPE III DES FACTEURS

FACTEUR	DDL	KHI2 DE WALD	PROB > CHISQ
Bonus-malus Année -1 (2 mod) - GBM1	1	242.9634	< 0.0001
Année de construction du véhicule (2 mod) - DCOS 38-39	1	54.1006	< 0.0001
Age de l'assuré (3 mod) - DNAI 8-9	2	46.0256	< 0.0001
Code postal souscripteur (2 mod) - POSS2 17-18	1	20.9027	< 0.0001

ETAPE 5 : ENTREE DU FACTEUR Code usage - CUSA 5-6

ANALYSE DE TYPE III DES FACTEURS

FACTEUR	DDL	KHI2 DE WALD	PROB > CHISQ
Bonus-malus Année -1 (2 mod) - GBM1	1	197.1667	< 0.0001
Année de construction du véhicule (2 mod) - DCOS 38-39	1	49.0268	< 0.0001
Age de l'assuré (3 mod) - DNAI 8-9	2	57.1381	< 0.0001
Code postal souscripteur (2 mod) - POSS2 17-18	1	20.0681	< 0.0001
Code usage - CUSA 5-6	1	15.8602	< 0.0001

ETAPE 6 : ENTREE DU FACTEUR Puissance du véhicule (2 mod) - PUIS 32-33

AJUSTEMENT DU MODELE

LE CRITERE DE CONVERGENCE (.1E-07) EST SATISFAIT

TEST DU KHI2 RESIDUEL

KHI-CARRE	DDL	PR > KHI2
7.7837	4	0.0998

	INTERCEPT SEULEMENT	INTERCEPT ET COVARIABLES
CRITERE D'AKAIKE	1535.209	848.583
CRITERE DE SCHWARZ	1540.218	888.651
-2 LOG (L)	1533.209	832.583

TEST GLOBAL DE L'HYPOTHESE NULLE : BETA = 0

	KHI-CARRE	DDL	PROB > KHI2
RAPPORT DE VRAISEMBLANCE	700.6259	7	< 0.0001
WALD	370.1238	7	< 0.0001

ANALYSE DE TYPE III DES FACTEURS

FACTEUR	DDL	KHI2 DE WALD	PROB > CHISQ
Bonus-malus Année -1 (2 mod) - GBM1	1	192.7969	< 0.0001
Année de construction du véhicule (2 mod) - DCOS 38-39	1	44.3886	< 0.0001
Age de l'assuré (3 mod) - DNAI 8-9	2	59.6889	< 0.0001
Code postal souscripteur (2 mod) - POSS2 17-18	1	20.4287	< 0.0001
Code usage - CUSA 5-6	1	13.4376	0.0002
Puissance du véhicule (2 mod) - PUIS 32-33	1	8.7719	0.0031

COEFFICIENTS DE REGRESSION ESTIMES PAR MAXIMUM DE VRAISEMBLANCE

PARAMETRE		DDL	ESTIMATION	ERREUR STAND	KHI2 DE WALD	PROB > KHI2	EXP(ESTIM.)
Intercept		1	1.3390	0.2659	25.3580	< 0.0001	3.8153
Bonus-malus Année -1 (2 mod) - GBM1	1	1	-2.5849	0.1862	192.7969	< 0.0001	0.0754
	2	0	0	.	.	.	.
Année de construction du véhicule (2 mod) - DCOS 38-39	1	1	-1.3871	0.2082	44.3886	< 0.0001	0.2498
	2	0	0	.	.	.	.
Age de l'assuré (3 mod) - DNAI 8-9	1	1	0.5030	0.2136	5.5444	0.0185	1.6536
	2	1	1.7875	0.2336	58.5520	< 0.0001	5.9743
	3	0	0	.	.	.	.
Code postal souscripteur (2 mod) - POSS2 17-18	1	1	0.8601	0.1903	20.4287	< 0.0001	2.3633
	2	0	0	.	.	.	.
Code usage - CUSA 5-6	1	1	0.9056	0.2470	13.4376	0.0002	2.4733
	2	0	0	.	.	.	.
Puissance du véhicule (2 mod) - PUIS 32-33	1	1	-0.7313	0.2469	8.7719	0.0031	0.4813
	2	0	0	.	.	.	.

ESTIMATIONS DES "ODDS RATIO"

FACTEUR				ESTIMATION	INTERVALLE DE CONFIANCE *	
Bonus-malus Année -1 (2 mod) - GBM1	1	VS	2	0.075	0.056	0.102
Année de construction du véhicule (2 mod) - DCOS 38-39	1	VS	2	0.250	0.177	0.352
Age de l'assuré (3 mod) - DNAI 8-9	1	VS	3	1.654	1.164	2.350
	2	VS	3	5.974	4.068	8.773
Code postal souscripteur (2 mod) - POSS2 17-18	1	VS	2	2.363	1.728	3.232
Code usage - CUSA 5-6	1	VS	2	2.473	1.647	3.713
Puissance du véhicule (2 mod) - PUIS 32-33	1	VS	2	0.481	0.321	0.722

* INTERVALLE DE CONFIANCE DE WALD A 90%

FIN DE LA SELECTION DES VARIABLES

AUX SEUILS RESPECTIFS DE 0.050 ET 0.049 IL N'Y A PLUS DE FACTEURS CANDIDATS A LA SORTIE OU A L'ENTREE !

MATRICE DE CONFUSION

FREQUENCES

	ESTIME	0	1	TOTAL
OBSERVE	0	484	66	550
	1	86	470	556
TOTAL		570	536	1106

POURCENTAGES LIGNE

	ESTIME	0	1	TOTAL
OBSERVE	0	88.000	12.000	100.000
	1	15.468	84.532	100.000
TOTAL		51.537	48.463	100.000

POURCENTAGES COLONNE

	ESTIME	0	1	TOTAL
OBSERVE	0	84.912	12.313	49.729
	1	15.088	87.687	50.271
TOTAL		100.000	100.000	100.000

Fonction de Score

Dans SPAD, l'élaboration d'un score se fait en plusieurs temps :

- ✓ Tout d'abord, on détermine les variables les plus discriminantes de l'élément à scorer (à l'aide de la méthode DEMOD)
- ✓ Ensuite, on réalise soit une analyse discriminante sur données qualitatives (méthodes DISQUAL), soit une régression logistique sur données qualitatives (LOGISQUAL) pour calculer la fonction discriminante ou la régression logistique correspondante.
- ✓ Puis on attribue une notre à chaque modalité de chaque variable explicative en fonction de son influence sur le comportement à expliquer.
- ✓ Enfin, chaque individu est noté en fonction de ses caractéristiques.

NB : Les étapes 3 et 4 sont incrémentées dans la méthode SCORE, qui requiert l'exécution de la méthode LOGISQUAL ou DISQUAL au préalable .

La méthode de scoring de SPAD réalise une ACM préalablement à l'analyse discriminante pour les raisons suivantes :

Afin de présenter la démarche méthodologique de la construction d'une fonction de SCORE avec SPAD, nous continuons d'utiliser le fichier **CREDIT.SBA**.

L'objectif est de discriminer des individus à risques et des individus sûrs par rapport à leurs caractéristiques et habitudes bancaires.
L'objectif étant bien entendu de construire des règles de décisions applicables à de nouveaux individus.

Afin de visualiser les variables, nous pouvons ouvrir la base CREDIT.SBA dans Edibase. Pour ce faire, dans la fenêtre principale de SPAD, cliquez sur « Base » puis « Editer base ». Regardez alors la structure de la base et les caractéristiques des variables explicatives.

Extrait de la base CREDIT.SBA :

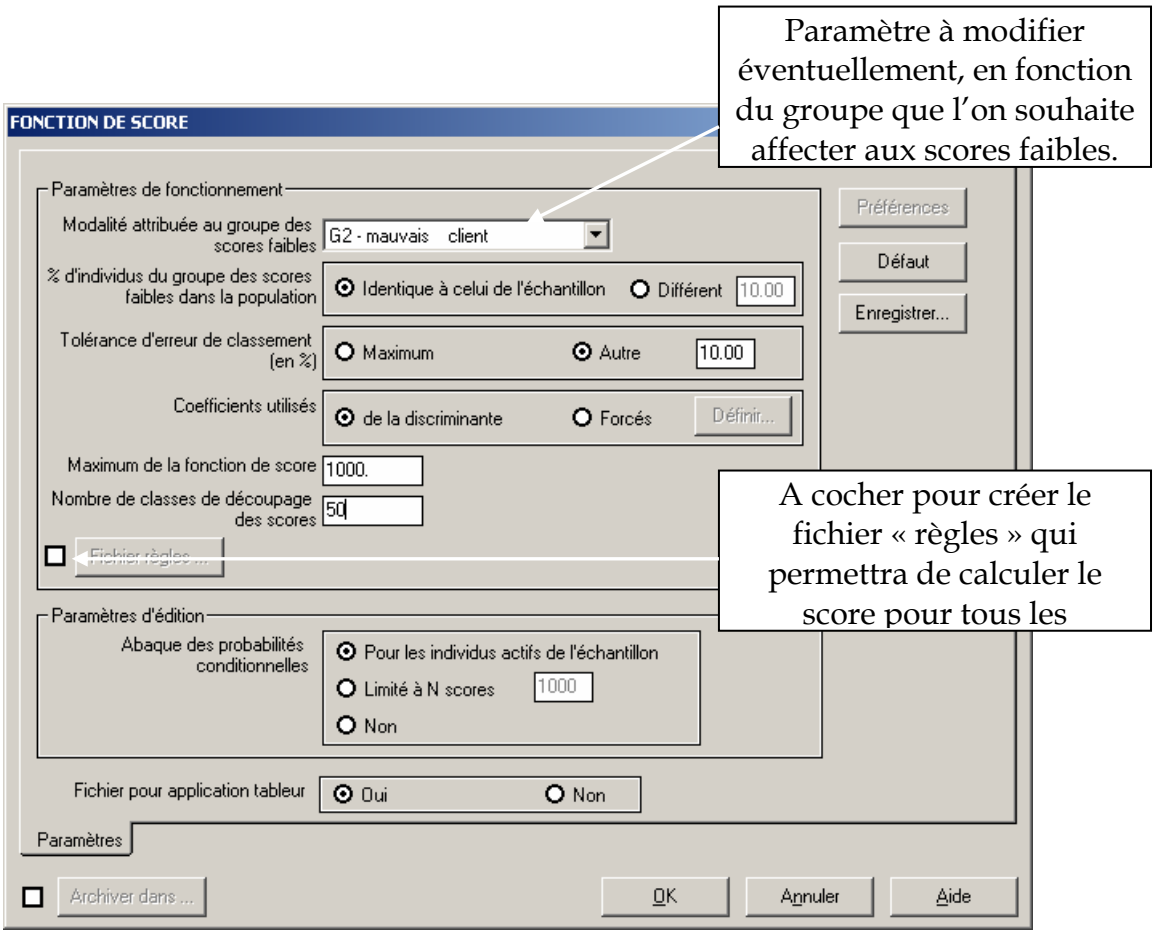
Type de client	Age du client	Situation familiale	Ancienneté	Domiciliation du salaire
bon client	plus de 50 ans	célibataire	anc. plus 12 ans	domicile salaire
bon client	moins de 23 ans	célibataire	anc. 1 an ou moins	domicile salaire
mauvais client	de 23 à 40 ans	veuf	anc. de 6 à 12 ans	domicile salaire
bon client	de 23 à 40 ans	divorcé	anc. de 1 à 4 ans	domicile salaire
bon client	moins de 23 ans	célibataire	anc. de 6 à 12 ans	non dimicile salaire
bon client	de 23 à 40 ans	célibataire	anc. 1 an ou moins	domicile salaire
bon client	plus de 50 ans	marié	anc. de 6 à 12 ans	domicile salaire
bon client	plus de 50 ans	marié	anc. plus 12 ans	domicile salaire
bon client	de 40 à 50 ans	célibataire	anc. de 1 à 4 ans	domicile salaire
bon client	plus de 50 ans	célibataire	anc. de 4 à 6 ans	domicile salaire
bon client	plus de 50 ans	marié	anc. plus 12 ans	domicile salaire
bon client	de 40 à 50 ans	marié	anc. 1 an ou moins	non dimicile salaire
bon client	de 23 à 40 ans	célibataire	anc. de 4 à 6 ans	non dimicile salaire

Le fichier CREDIT.SBA se compose de 468 individus sur lesquels on a relevé 12 caractéristiques.

VARIABLES NOMINALES	
1. Type de client	(2 modalités)
2. Age du client	(4 modalités)
3. Situation familiale	(4 modalités)
4. Ancienneté	(5 modalités)
5. Domiciliation du salaire	(2 modalités)
6. Domiciliation de l'épargne	(4 modalités)
7. Profession	(3 modalités)
8. Moyenne en cours	(3 modalités)
9. Moyenne des mouvements	(4 modalités)
10. Cumul des débits	(3 modalités)
11. Autorisation de découvert	(2 modalités)
12. Interdiction de chéquier	(2 modalités)

Nous sommes bien dans le cadre de la construction de la fonction de SCORE avec SPAD puisque la variable à expliquer « **Type de client** » a 2 modalités et que toutes les variables explicatives sont nominales.

LES PARAMETRES DE LA METHODE SCORE



Après avoir validé par « OK » les paramètres de SCORE, exécutez alors la méthode.

LES RESULTATS DE SCORE

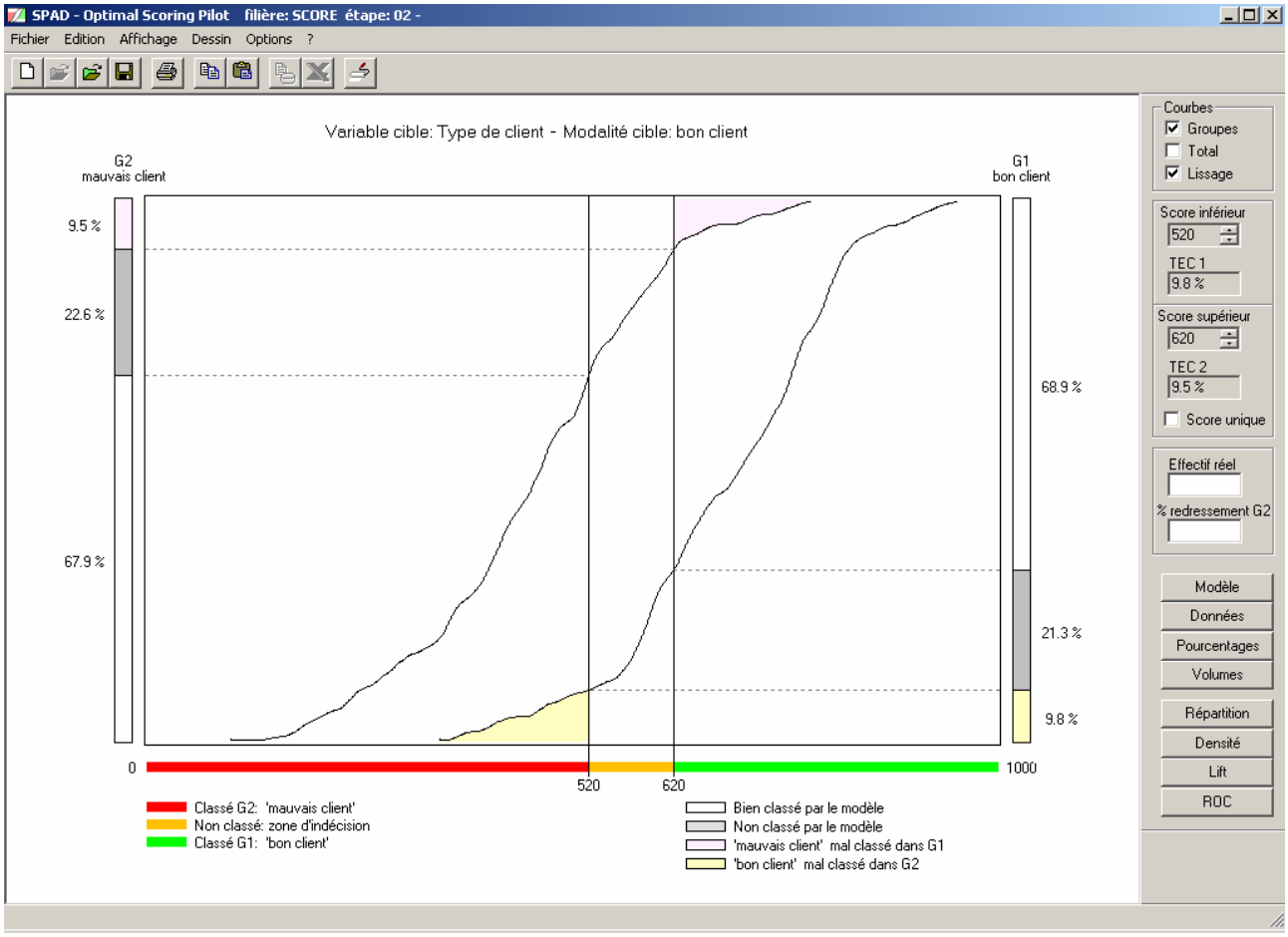
Coefficients des fonctions discriminante et de score

Libellés des modalités	Coefficients de la F.L.D.	Coefficients de la fonction de Score
Age du client		
moins de 23 ans	-6.395	0.00
de 23 à 40 ans	2.830	64.46
de 40 à 50 ans	1.971	58.46
plus de 50 ans	-0.946	38.07
Situation familiale		
célibataire	-0.615	31.95
marié	1.769	48.61
divorcé	-3.335	12.95
veuf	-5.188	0.00
Ancienneté		
anc. 1 an ou moins	-7.076	6.14
anc. de 1 à 4 ans	-7.955	0.00
anc. de 4 à 6 ans	8.594	115.63
anc. de 6 à 12 ans	4.312	85.72
anc. plus 12 ans	10.395	128.22
Domiciliation du salaire		
domicile salaire	5.537	119.12
non dimicile salaire	-11.511	0.00
Domiciliation de l'épargne		
pas d'épargne	-1.771	0.00
moins de 10KF épargn	3.506	36.87
de 10 à 100KF épargn	10.735	87.39
plus de 100KF épargn	13.555	107.09
Profession		
cadre	3.112	64.07
employé	2.925	62.77
autre	-6.057	0.00
Moyenne en cours		
moins de 2KF encours	-15.109	0.00
de 2 à 5 KF encours	2.918	125.96
plus de 5 KF encours	9.386	171.16
Moyenne des mouvements		
moins 10 KF mouvt	-4.056	0.00
de 10 à 30KF mouvt	0.003	28.36
de 30 à 50KF mouvt	0.877	34.47
plus de 50KF mouvt	4.485	59.68
Cumul des débits		
moins de 40 débits	6.608	109.89
de 40 à 100 débits	0.684	68.49
plus de 100 débits	-9.118	0.00
Autorisation de découvert		
découvert autorisé	-0.526	0.00
découvert interdit	0.399	6.47
Interdiction de chéquier		
chéquier autorisé	1.965	121.23
chéquier interdit	-15.384	0.00

EDITEUR GRAPHIQUE OPTIMAL SCORING PILOT

Un double-clic sur cette icône  permet d'accéder à l'éditeur graphique « Optimal Scoring Pilot ».

Faire « **Fichier** » - « **Nouveau** » pour afficher le graphique initial issu du calcul numérique.



La vue « graphique » initiale est affichée pour un taux maximum d'erreur de classement (TEC) de 10% environ pour les 2 groupes.

Ce TEC détermine trois zones : la « zone rouge », la « zone orange », la « zone verte » matérialisées pas des bornes sur l'échelle des scores.

Taux d'erreur de classement (TEC)

Les taux d'erreur de classement (TEC) représentent le pourcentage accepté de « bons » classés « mauvais » par la fonction de score (TEC1) et de « mauvais » classés « bons » par la fonction de score (TEC2).

Pour le graphique initial, ces taux sont fixés à 10% pour les 2 groupes. Ils déterminent les bornes inférieures et supérieures de la zone d'indécision.

Exemple

Le score 520 (borne inférieur) est tel que 9.8% des individus du groupe des scores forts soient mal classés et le score 620 (borne supérieure) tel que 9.5% des individus du groupe des scores faibles soient mal classés.

Les bornes sont modifiables de façon individuelle afin d'obtenir des TEC différents pour les 2 groupes.

La « zone rouge »

Correspond aux individus des 2 groupes déclarés « Mauvais » par la fonction de score (à tort pour le groupe des « bons »).
Dans l'exemple : 67.9% des individus du groupe des « mauvais » sont classés dans cette zone et 9.8% des individus du groupe des « bons ».

La « zone orange »

Correspond aux individus des 2 groupes déclarés « Non classés » par la fonction de score.
Dans l'exemple : 22.6% des individus du groupe des « mauvais » sont classés dans cette zone et 21.3% des individus du groupe des « bons ».

La « zone verte »

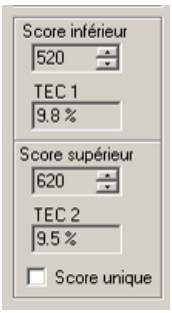
Correspond aux individus des 2 groupes déclarés « Bons » par la fonction de score (à tort pour le groupe des « mauvais »).
Dans l'exemple : 9.5% des individus du groupe des « mauvais » sont classés dans cette zone et 68.9% des individus du groupe des « bons ».

Modifier les bornes de façon interactive sur le graphique

Dans la vue « graphique », pour modifier les bornes qui déterminent la zone d'indécision (zone orange), se positionner sur le trait vertical à l'aplomb d'une borne (le curseur change de forme) puis déplacer le curseur à gauche ou à droite.

Les pourcentages et vues associés sont modifiés dynamiquement.

Modifier les bornes en intervenant sur la partie Score du tableau de bord



Cette partie du tableau de bord permet de modifier manuellement les bornes ou les pourcentages (TEC) qui déterminent les taux acceptés d'individus mal classés par groupe.

La vue « Données » 

Tableau des données (468 individus)

Séquence	Identificateur	Poids	Echant.	Score	Groupe	Affect.	Err. G1	Err. G2
1	0005	1.00	Apprent.	762	G1	G1		
2	0011	1.00	Apprent.	584	G1	nc	x	
3	0018	1.00	Apprent.	586	G2	nc		x
4	0030	1.00	Apprent.	572	G1	nc	x	
5	0048	1.00	Apprent.	821	G1	G1		
6	0054	1.00	Apprent.	428	G1	G2	xx	
7	0066	1.00	Apprent.	779	G1	G1		
8	0072	1.00	Apprent.	612	G1	nc	x	
9	0084	1.00	Apprent.	731	G1	G1		
10	0090	1.00	Apprent.	595	G1	nc	x	
11	0096	1.00	Apprent.	568	G2	nc		x

Les différents champs de la vue « Données » sont les suivants :

Identificateur :

Dans ce tableau, l'identificateur des individus est tronqué à 40 caractères.

Poids :

Le poids est défini dans l'onglet pondération de la méthode DISQUAL ; il est par défaut uniforme et égal à 1.

Echant. - Echantillon :

Les individus de l'échantillon sont soit des individus « d'apprentissage », soir des individus « test ».

Seuls les individus « d'apprentissage » ont participé à la construction de la fonction de score. La définition des individus « test » se fait dans l'onglet « Paramètres » de la méthode DISQUAL.

Score :

On affiche le score de chaque individu.

Groupe :

Affiche le groupe d'origine de l'individu (G1 ou G2), quelque soit son statut (apprentissage ou test).

Affect. - Affectation :

Affiche le groupe d'affectation de l'individu (G1, NC ou G2), quelque soir son statut.

« NC » signifie que l'individu est dans la zone d'indécision (zone orange).

Err. G1 - Erreur groupe 1 :

Si l'individu appartient au groupe 1 (G1) et s'il est affecté :

au groupe 1, il n'y a pas d'erreur d'affectation
à la zone d'indécision, il y a une erreur d'affectation notée (x)
au groupe 2, il y a une erreur d'affectation marquée (xx)

Err. G2 - Erreur groupe 2 :

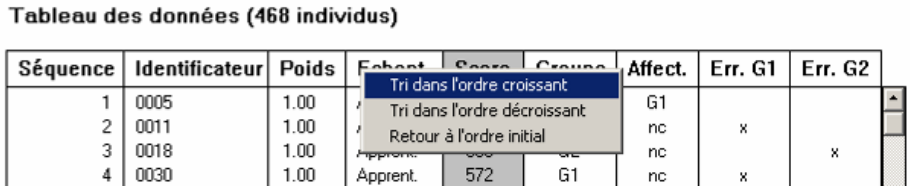
Si l'individu appartient au groupe 2 (G2) et s'il est affecté :

au groupe 2, il n'y a pas d'erreur d'affectation
à la zone d'indécision, il y a une erreur d'affectation notée (x)
au groupe 1, il y a une erreur d'affectation marquée (xx)

Tri des données sur un champ :

Un simple clic sur un intitulé de colonne fait apparaît un menu permettant de :

- trier dans l'ordre croissant
- trier dans l'ordre décroissant
- retourner à l'ordre initial

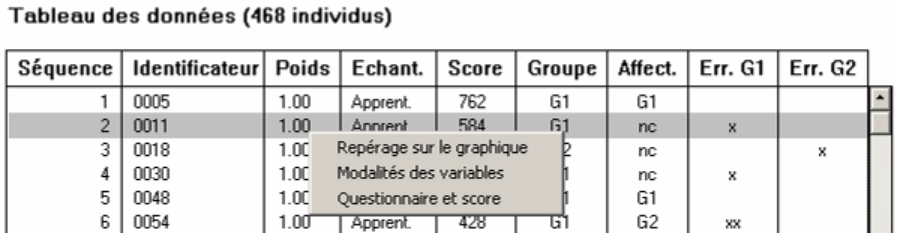


Les réponses individuelles :

En sélectionnant un individu dans la vue « Données », simple clic sur la ligne correspondant à l'individu, il est possible de :

- repérer l'individu sur la vue graphique
  l'individu sélectionné apparaît en rouge avec son identificateur, faire « échap » pour revenir à la vue « Données »
- afficher ses réponses sous forme condensée
- afficher son questionnaire et les scores associés aux réponses

les réponses de l'individu et les scores associés apparaissent en bleu
on dispose également de son score global, de son classement et de son groupe d'origine



Simulation interactives, après avoir choisi « Questionnaire et score » :

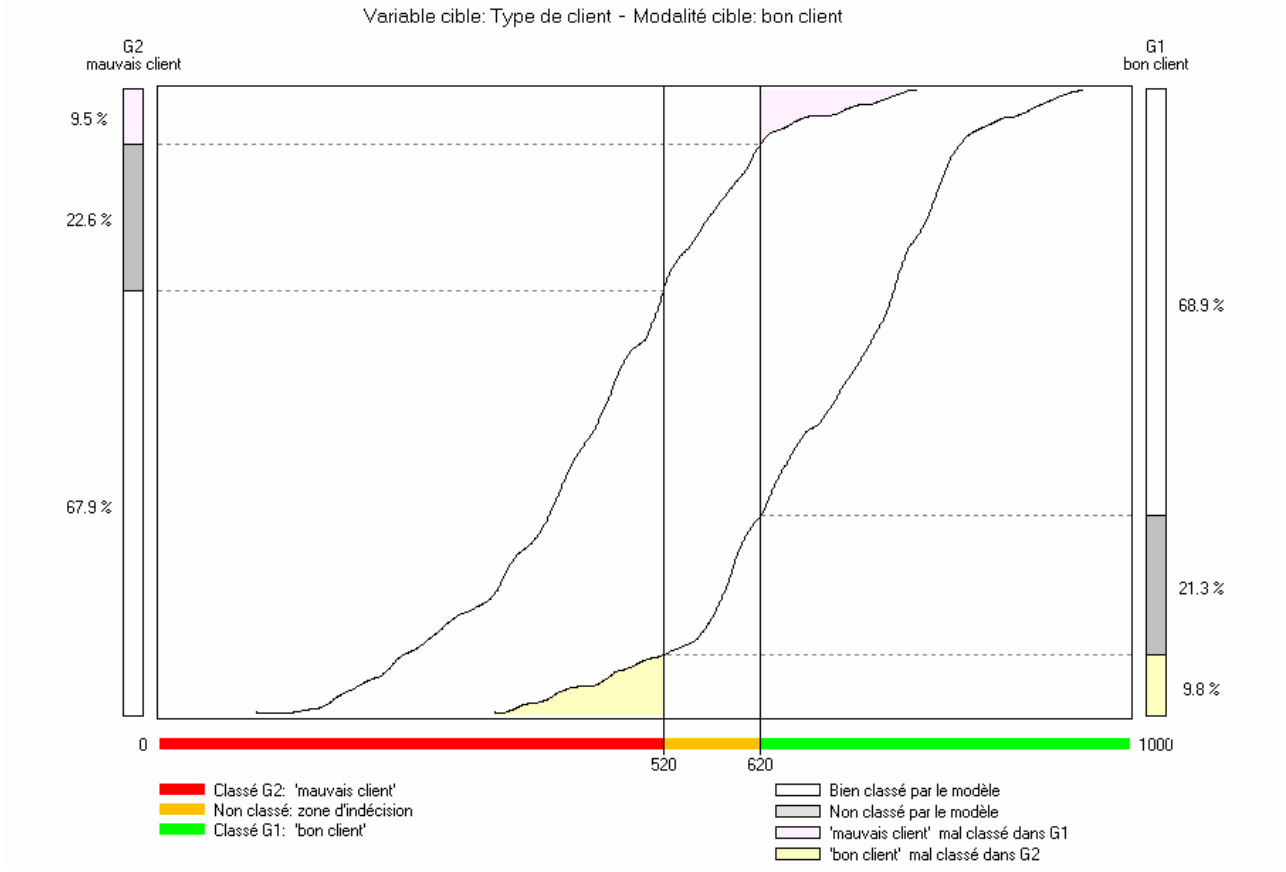
Il est possible de modifier de façon interactive les réponses de l'individu pour faire des simulations.

- cliquez sur les réponses choisies, elles passent en rouge
- le nouveau score et le nouveau classement sont affichés

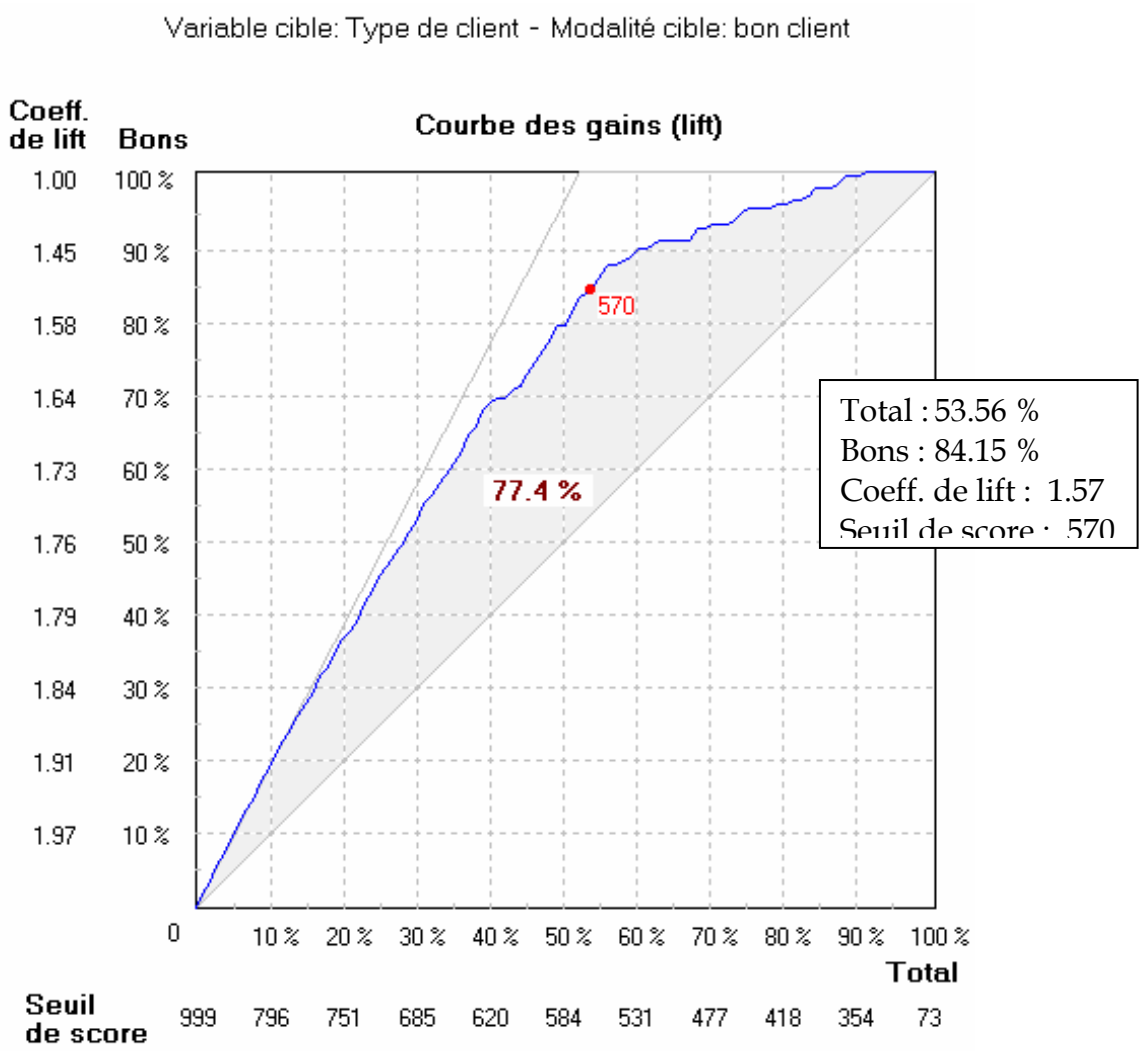
Questionnaire et score		Scores des modalités	Histogrammes des scores
2	**Age du client**		
AGE1	moins de 23 ans	0.00	
AGE2	de 23 à 40 ans	64.46	
AGE3	de 40 à 50 ans	58.46	
AGE4	plus de 50 ans	38.07	
3	**Situation familiale**		
CELB	célibataire	31.95	
MARI	marié	48.61	
DIVO	divorcé	12.95	
VEUF	veuf	0.00	
4	**Ancienneté**		
ANC1	anc. 1 an ou moins	6.14	
ANC2	anc. de 1 à 4 ans	0.00	
ANC3	anc. de 4 à 6 ans	115.63	
ANC4	anc. de 6 à 12 ans	85.72	
ANC5	anc. plus 12 ans	128.22	
5	**Domiciliation du salaire**		
Soui	domicile salaire	119.12	
Snon	non dimicile salaire	0.00	
6	**Domiciliation de l'épargne**		
EPA0	pas d'épargne	0.00	
EPA1	moins de 10KF épargn	36.87	
EPA2	de 10 à 100KF épargn	87.39	

Individu 0011 (groupe G1)
Score: 584 Classement: Non classé
Score: 705 Classement: G1

Courbe de répartition



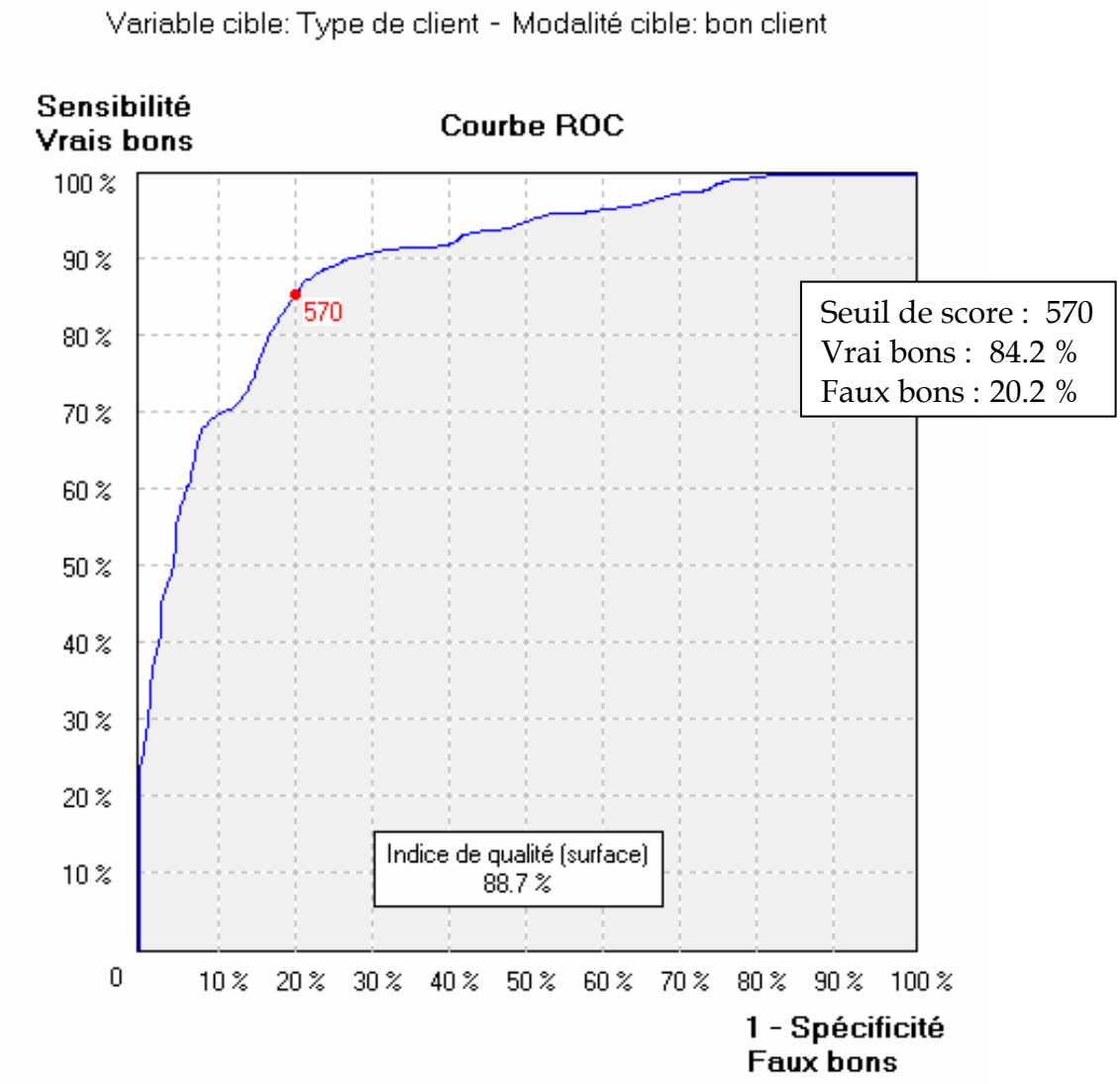
Courbe de Lift (ou d'efficacité de la sélection)



Abscisse : % de tous les individus bons et mauvais ayant un score supérieur à s (seuil de score)
Ordonnée : % de vrais bons ayant un score supérieur à s

La courbe idéale est le segment brisé qui correspond au cas où la distribution des « mauvais » est entièrement inférieure à la distribution des « bons ».

Courbe ROC (Receiver Operating Curve)



Sensibilité : Proportion de vrais bons chez les bons
Spécificité : Proportion de vrais mauvais chez les mauvais
1-Spécificité : Proportion de faux bons chez les mauvais

Plus la courbe est proche de la partie supérieure du carré, meilleure est la séparation.
Lorsque les densités sont identiques, la courbe ROC se confond ave la diagonale du carré.

L'ANALYSE DISCRIMINANTE ET SES METHODES

Discriminante sur variables qualitatives pour fonction de Score

---

Dans SPAD, l'élaboration d'un score se fait en plusieurs temps :

- ✓ Tout d'abord, on détermine les variables les plus discriminantes de l'élément à scorer (à l'aide de la méthode DEMOD)
- ✓ Ensuite, on réalise une Analyse des Correspondances Multiples de l'ensemble des données sélectionnées.
- ✓ On réalise une analyse discriminante sur les coordonnées factorielles issues de l'ACM.
- ✓ Puis on attribue une notre à chaque modalité de chaque variable explicative en fonction de son influence sur le comportement à expliquer.
- ✓ Enfin, chaque individu est noté en fonction de ses caractéristiques.

NB : Les étapes 2 et 3 sont incrémentées dans la première méthode DISQUAL.

La méthode de scoring de SPAD réalise une ACM préalablement à l'analyse discriminante pour les raisons suivantes :

- ✓ L'analyse discriminante est une méthode qui nécessite de n'avoir que des variables continues en entrée.
- ✓ L'ACM travaille sur des variables nominales puis l'analyse discriminante se fait à partir des coordonnées factorielles (combinaisons linéaires des variables d'entrée) de l'ACM.
- ✓ Les variables d'entrée de l'analyse discriminante sont indépendantes par construction, on s'affranchit donc des problèmes de multicolinéarité de la discriminante.
- ✓ Enfin, la sélection des axes les plus informatifs de l'ACM permet d'affiner les résultats.

Afin de présenter la démarche méthodologique de la construction d'une fonction de SCORE avec SPAD, nous allons utiliser le fichier **CREDIT.SBA**.

L'objectif est de discriminer des individus à risques et des individus sûrs par rapport à leurs caractéristiques et habitudes bancaires.
L'objectif étant bien entendu de construire des règles de décisions applicables à de nouveaux individus.

Afin de visualiser les variables, nous pouvons ouvrir la base CREDIT.SBA dans Edibase. Pour ce faire, dans la fenêtre principale de SPAD, cliquez sur « Base » puis « Editer base ». Regardez alors la structure de la base et les caractéristiques des variables explicatives.

Extrait de la base CREDIT.SBA :

Type de client	Age du client	Situation familiale	Ancienneté	Domiciliation du salaire
bon client	plus de 50 ans	célibataire	anc. plus 12 ans	domicile salaire
bon client	moins de 23 ans	célibataire	anc. 1 an ou moins	domicile salaire
mauvais client	de 23 à 40 ans	veuf	anc. de 6 à 12 ans	domicile salaire
bon client	de 23 à 40 ans	divorcé	anc. de 1 à 4 ans	domicile salaire
bon client	moins de 23 ans	célibataire	anc. de 6 à 12 ans	non dimicile salaire
bon client	de 23 à 40 ans	célibataire	anc. 1 an ou moins	domicile salaire
bon client	plus de 50 ans	marié	anc. de 6 à 12 ans	domicile salaire
bon client	plus de 50 ans	marié	anc. plus 12 ans	domicile salaire
bon client	de 40 à 50 ans	célibataire	anc. de 1 à 4 ans	domicile salaire
bon client	plus de 50 ans	célibataire	anc. de 4 à 6 ans	domicile salaire
bon client	plus de 50 ans	marié	anc. plus 12 ans	domicile salaire
bon client	de 40 à 50 ans	marié	anc. 1 an ou moins	non dimicile salaire
bon client	de 23 à 40 ans	célibataire	anc. de 4 à 6 ans	non dimicile salaire

Le fichier CREDIT.SBA se compose de 468 individus sur lesquels on a relevé 12 caractéristiques.

VARIABLES NOMINALES

1.	Type de client	(2 modalités)
2.	Age du client	(4 modalités)
3.	Situation familiale	(4 modalités)
4.	Ancienneté	(5 modalités)
5.	Domiciliation du salaire	(2 modalités)
6.	Domiciliation de l'épargne	(4 modalités)
7.	Profession	(3 modalités)
8.	Moyenne en cours	(3 modalités)
9.	Moyenne des mouvements	(4 modalités)
10.	Cumul des débits	(3 modalités)
11.	Autorisation de découvert	(2 modalités)
12.	Interdiction de chéquier	(2 modalités)

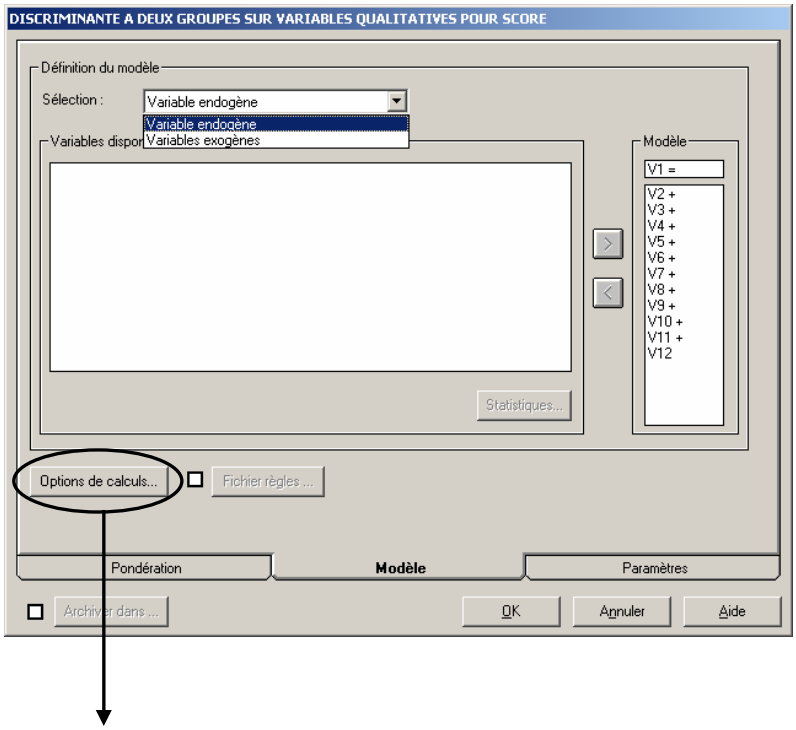
Nous sommes bien dans le cadre de la construction de la fonction de SCORE avec SPAD puisque la variable à expliquer « **Type de client** » a 2 modalités et que toutes les variables explicatives sont nominales.

Avant de construire le score, prenez quelques minutes pour faire les analyses descriptives et la caractérisation des modalités à discriminer (méthode STATS, DEMOD).

LE PARAMETRAGE DE DISQUAL

Le paramétrage de la méthode DISQUAL va permettre d'expliciter la variable que l'on veut discriminer (ou à expliquer ou endogène) et les variables discriminantes (ou explicatives ou exogènes).

On a ainsi :

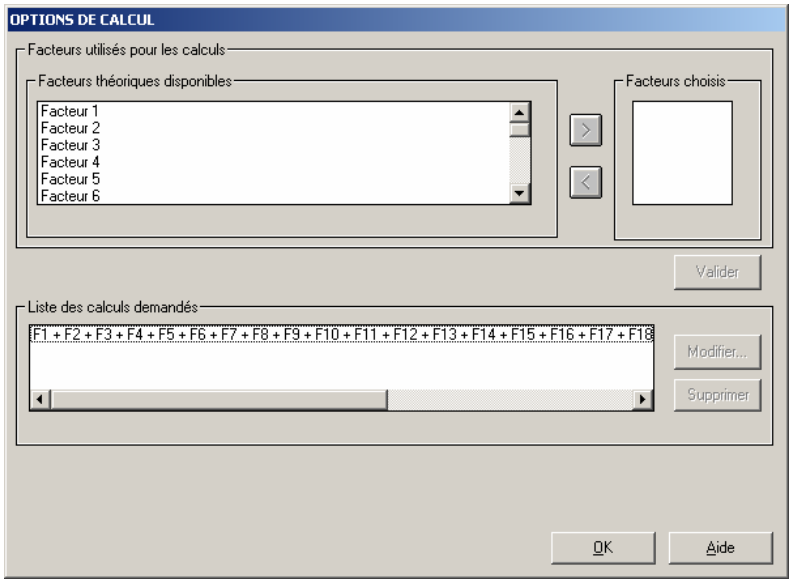


Le modèle défini est le suivant :

V1 = V2 + ... + V12

Dans ce même onglet « **Modèle** », nous allons expliciter le vrai modèle sous-jacent, i.e. celui construit à partir des facteurs issus de l'ACM préalable.

Pour cela, il suffit de cliquer sur le bouton « **Options de calculs** »

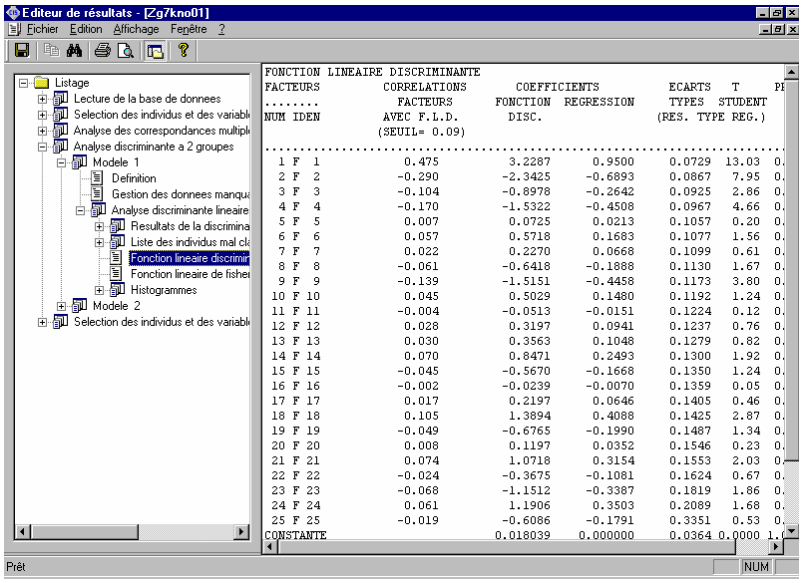


On retient dans un premier temps le **modèle complet**, celui qui prend en compte tous les facteurs en même temps.

Cliquez sur « **OK** » pour revenir dans la fenêtre « Modèle » et à nouveau sur « **Ok** » pour spécifier que le paramétrage de la méthode est terminé.

Exécutez la méthode

La construction du modèle à partir du modèle complet est simplement un artifice de calcul permettant de sélectionner un modèle optimal.



On va simplement ici visualiser et sélectionner les axes permettant de discriminer efficacement la variable de départ.

On s'appuie sur la Proba associée à la valeur du T de Student.

Celle-ci quantifie le risque pris en considérant que le coefficient associé au facteur est différent de 0.

LES RESULTATS DE DISQUAL

Fonction linéaire discriminante

Modèle : V1 = F1 + F2 + F3 + F4 + F5 + F6 + F7 + F8 + F9 + F10 + F11 + F12 + F13 + F14 + F15 + F16 + F17 + F18 + F19 + F20 + F21 + F22 + F23 + F24 + F25

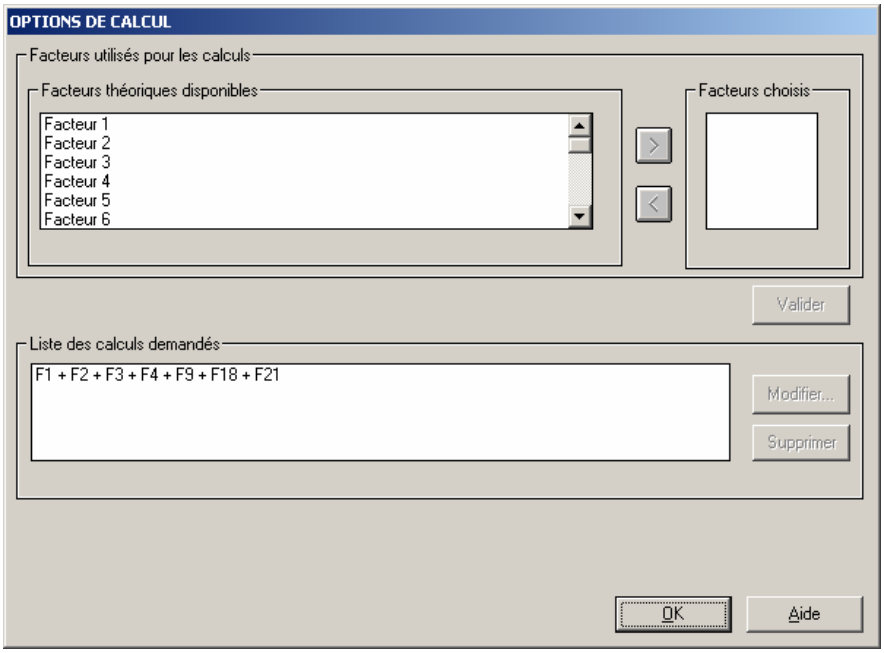
R2 = 0.41398 Fisher = 12.48967 Probabilité = 0.0000
D2 (Mahalanobis) = 2.81410 T2 (Hotelling) = 329.19614 Probabilité = 0.0000

Libellés des facteurs	Corrélations avec la F.L.D. (seuil = 0.093)	Coefficients de la F.L.D.	Coefficients de régression	Ecarts-types (régression)	T de Student (régression)	Probabilité
F 1	-0.4750	-3.2287	-0.9500	0.0729	13.0262	**0.0000**
F 2	0.2898	2.3425	0.6893	0.0867	7.9474	**0.0000**
F 3	0.1041	0.8978	0.2642	0.0925	2.8551	**0.0045**
F 4	0.1700	1.5322	0.4508	0.0967	4.6611	**0.0000**
F 5	-0.0074	-0.0725	-0.0213	0.1057	0.2018	0.8402
F 6	-0.0570	-0.5718	-0.1683	0.1077	1.5617	0.1191
F 7	-0.0222	-0.2270	-0.0668	0.1099	0.6076	0.5437
F 8	0.0609	0.6418	0.1888	0.1130	1.6705	0.0955
F 9	0.1386	1.5151	0.4458	0.1173	3.8017	**0.0002**
F 10	-0.0453	-0.5029	-0.1480	0.1192	1.2411	0.2152
F 11	0.0045	0.0513	0.0151	0.1224	0.1233	0.9020
F 12	-0.0277	-0.3197	-0.0941	0.1237	0.7605	0.4473
F 13	-0.0299	-0.3563	-0.1048	0.1279	0.8197	0.4128
F 14	-0.0699	-0.8471	-0.2493	0.1300	1.9170	0.0559
F 15	0.0451	0.5670	0.1668	0.1350	1.2364	0.2170
F 16	0.0019	0.0239	0.0070	0.1359	0.0518	0.9587
F 17	-0.0168	-0.2197	-0.0646	0.1405	0.4599	0.6458
F 18	-0.1046	-1.3894	-0.4088	0.1425	2.8691	**0.0043**
F 19	0.0488	0.6765	0.1990	0.1487	1.3381	0.1815
F 20	-0.0083	-0.1197	-0.0352	0.1546	0.2279	0.8199
F 21	-0.0740	-1.0718	-0.3154	0.1553	2.0303	**0.0429**
F 22	0.0243	0.3675	0.1081	0.1624	0.6659	0.5058
F 23	0.0679	1.1512	0.3387	0.1819	1.8622	0.0632
F 24	-0.0611	-1.1906	-0.3503	0.2089	1.6768	0.0943
F 25	0.0195	0.6086	0.1791	0.3351	0.5343	0.5934
Constante		0.0180	0.0000	0.0364	0.0000	1.0000

On était mis en **gras** les Proba < 0,05 ainsi que les noms des facteurs associés.
C'est donc à partir de ceux-ci que l'on va construire le modèle optimal.

Il faut donc revenir dans le paramétrage de la première méthode (DISQUAL) et modifier le modèle dans les « **Options de calculs** ».

LE NOUVEAU PARAMETRAGE DE LA METHODE DISQUAL



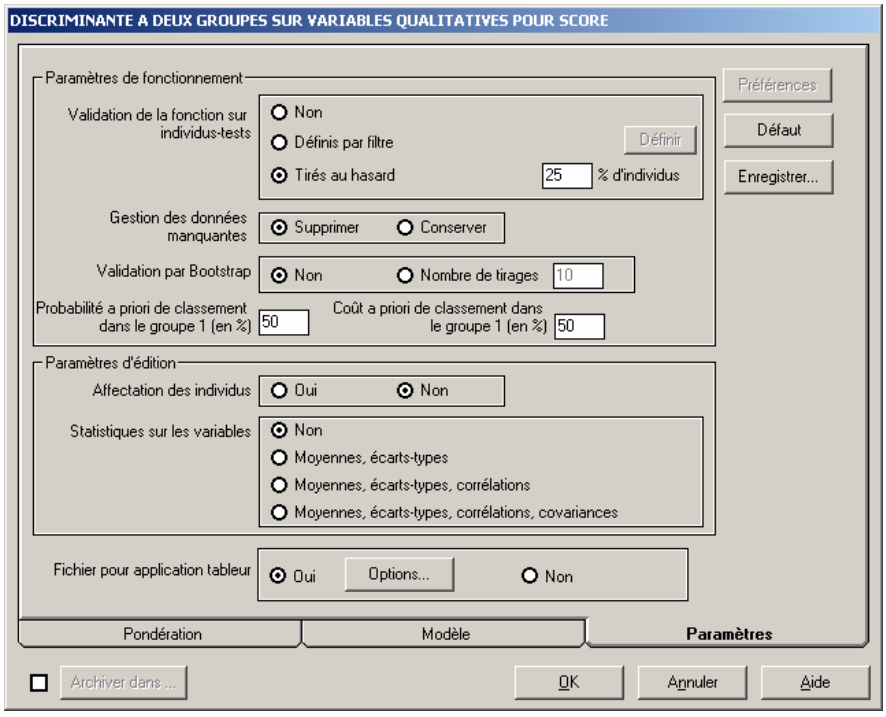
On a donc bien spécifié ici le modèle optimal à partir duquel on va construire la fonction de score.

Le modèle optimal est constitué des facteurs :

F1 à F4, F9, F18 et F21.

Il faut alors ré-exécuter la méthode.

Maintenant que le modèle optimal est construit, il va falloir spécifier l'utilisation d'une partie de la population pour tester le modèle, on parle d'**échantillon test**, onglet « **Paramètres** » de la méthode DISQUAL.



On va utiliser 25% de l'échantillon de départ pour tester le modèle calculé sur les 75 % autres.

Cela permet de tester la reproductibilité du modèle ainsi défini. C'est une phase importante dans la validation de la modélisation.

LES RESULTATS DE DISQUAL

On s'intéresse à présent à la qualité de prédiction du modèle. Pour cela, nous disposons des matrices de classement.

Résultats de la discrimination linéaire de Fisher sur l'**ECHANTILLON de BASE**

Tableau des effectifs par groupes

	Groupe d'affectation : bon client	Groupe d'affectation : mauvais client	Total
Groupe d'origine : bon client	150	28	178
Groupe d'origine : mauvais client	35	138	173

Tableau de classement (Effectifs et pourcentages)

	Bien classés	Mal classés	Total
Groupe d'origine : bon client	150	28	178
	84.27	15.73	100.00
Groupe d'origine : mauvais client	138	35	173
	79.77	20.23	100.00
Total	288	63	351
	82.05	17.95	100.00

Résultats de la discrimination linéaire de Fisher sur l'**ECHANTILLON TEST**

Tableau des effectifs par groupes

	Groupe d'affectation : bon client	Groupe d'affectation : mauvais client	Total
Groupe d'origine : bon client	50	9	59
Groupe d'origine : mauvais client	21	37	58

Tableau de classement (Effectifs et pourcentages)

	Bien classés	Mal classés	Total
Groupe d'origine : bon client	50	9	59
	84.75	15.25	100.00
Groupe d'origine : mauvais client	37	21	58
	63.79	36.21	100.00
Total	87	30	117
	74.36	25.64	100.00

Sur l'échantillon de **BASE**, on a 82.05 % de bien classés.
Sur l'échantillon **TEST**, 74.36 % sont bien classés.

Le modèle construit a donc un pouvoir de prédiction important et ce aussi sur l'échantillon de test. Il ne « colle » donc pas trop aux données de base et a donc un pouvoir de reproductibilité important.

Une autre alternative, pour valider le modèle, serait de faire du bootstrap.

Une fois le modèle validé, nous pouvons nous intéresser aux paramètres de la méthode **SCORE**.

Fonction linéaire discriminante

Modèle : V1=F1+F2+F3+F4+F9+F18+F21

R2 = 0.42370 Fisher = 36.02458 Probabilité = 0.0000
D2 (Mahalanobis) = 2.92462 T2 (Hotelling) = 256.58322 Probabilité = 0.0000

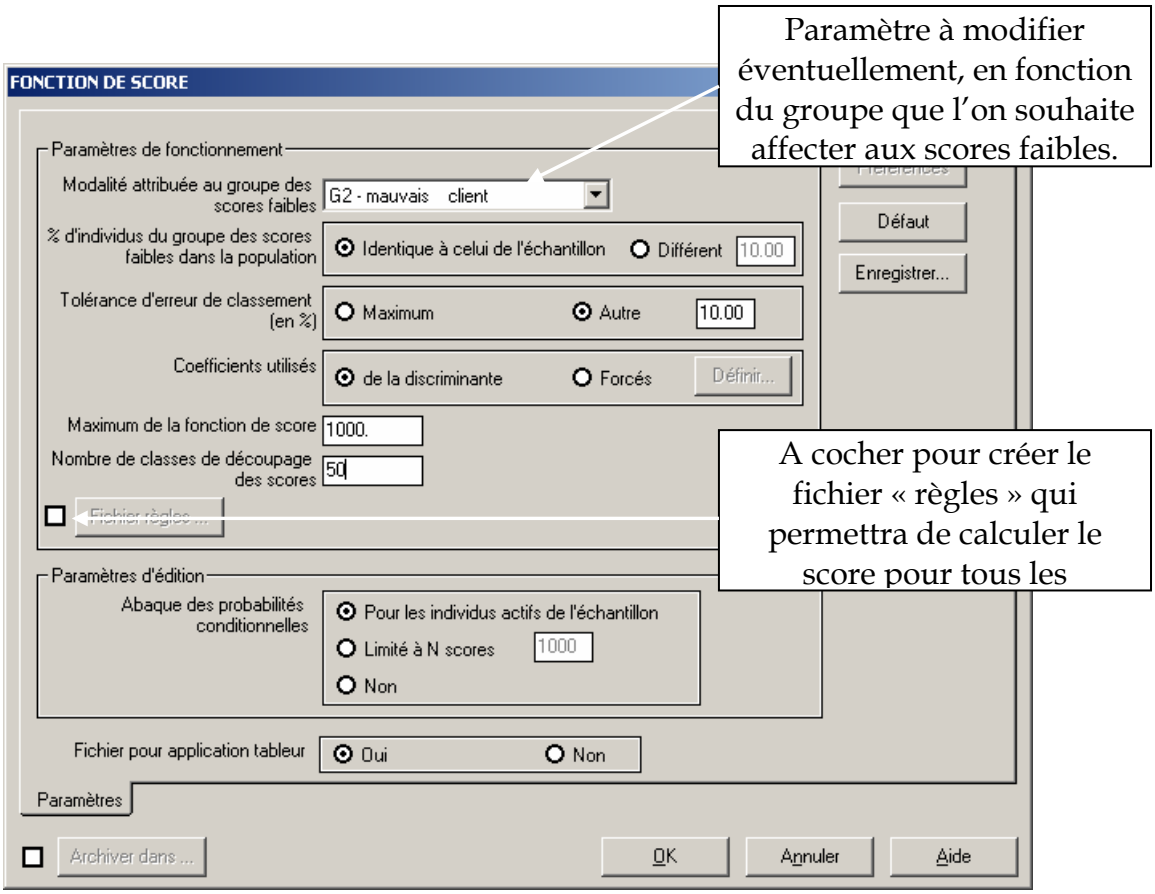
Libellés des facteurs	Corrélations avec la F.L.D. (seuil = 0.107)	Coefficients de la F.L.D.	Coefficients de régression	Ecarts-types (régression)	T de Student (régression)	Probabilité
F 1	-0.512	-3.643340	-1.055740	0.0854	12.3593	0.0000
F 2	0.303	2.422390	0.701946	0.0972	7.2195	0.0000
F 3	0.095	1.208380	0.350157	0.1126	3.1087	0.0020
F 4	0.195	1.638830	0.474890	0.1114	4.2646	0.0000
F 9	0.082	1.055530	0.305865	0.1362	2.2453	0.0254
F 18	-0.100	-1.818270	-0.526888	0.1610	3.2727	0.0012
F 21	-0.084	-1.167080	-0.338190	0.1703	1.9853	0.0479
CONSTANTE		0.146154	0.036316	0.0412	0.8821	0.3783

Fonction linéaire de Fisher reconstituée à partir des variables d'origine

Libellés des variables	Libellés des modalités	Coefficients de la F.L.D.	Coefficients de régression	T de Student (régression)	Probabilité
Age du client	moins de 23 ans	-6.394940	-1.853090	3.4909	0.0005
	de 23 à 40 ans	2.830280	0.820142	2.6505	0.0084
	de 40 à 50 ans	1.970640	0.571041	1.9463	0.0525
	plus de 50 ans	-0.946353	-0.274228	0.6067	0.5445
Situation familiale	célibataire	-0.614705	-0.178125	0.4589	0.6466
	marié	1.768850	0.512566	2.1213	0.0347
	divorcé	-3.334670	-0.966301	4.2334	0.0000
	veuf	-5.187540	-1.503210	1.3833	0.1675
Ancienneté	anc. 1 an ou moins	-7.075750	-2.050370	6.0816	0.0000
	anc. de 1 à 4 ans	-7.954850	-2.305110	5.2288	0.0000
	anc. de 4 à 6 ans	8.593590	2.490200	3.8243	0.0002
	anc. de 6 à 12 ans	4.312050	1.249520	2.5919	0.0100
	anc. plus 12 ans	10.395400	3.012310	4.1175	0.0000
Domiciliation du salaire	domicile salaire	5.536680	1.604390	10.3812	0.0000
	non dimicile salaire	-11.510500	-3.335430	10.3812	0.0000
Domiciliation de l'épargne	pas d'épargne	-1.771050	-0.513204	5.0430	0.0000
	moins de 10KF épargn	3.505570	1.015820	3.0824	0.0022
	de 10 à 100KF épargn	10.735200	3.110770	4.0443	0.0001
	plus de 100KF épargn	13.555100	3.927910	3.0771	0.0023
Profession	cadre	3.111500	0.901630	2.0071	0.0456
	employé	2.925100	0.847618	4.1505	0.0000
	profession autre	-6.057360	-1.755270	7.2684	0.0000
Moyenne en cours	moins de 2KF encours	-15.108900	-4.378170	9.2081	0.0000
	de 2 à 5 KF encours	2.917930	0.845540	4.7339	0.0000
	plus de 5 KF encours	9.386300	2.719910	4.8238	0.0000
Moyenne des mouvements	moins 10 KF mouvt	-4.055940	-1.175300	3.2944	0.0011
	de 10 à 30KF mouvt	0.002863	0.000829	0.0035	0.9972
	de 30 à 50KF mouvt	0.876654	0.254031	0.5584	0.5770
	plus de 50KF mouvt	4.485280	1.299720	4.4456	0.0000
Cumul des débits	moins de 40 débits	6.608100	1.914860	7.1414	0.0000
	de 40 à 100 débits	0.683879	0.198170	0.7778	0.4372
	plus de 100 débits	-9.118310	-2.642250	7.6443	0.0000
Autorisation de découvert	découvert autorisé	-0.525901	-0.152392	0.3488	0.7275
	découvert interdit	0.399366	0.115726	0.3488	0.7275
Interdiction de chéquier	chéquier autorisé	1.964710	0.569321	7.6750	0.0000
	chéquier interdit	-15.384000	-4.457890	7.6750	0.0000
	CONSTANTE	0.146154	0.036316		

Fonction de Score

LES PARAMETRES DE LA METHODE SCORE

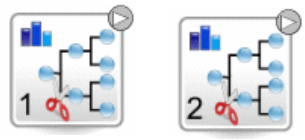


Après avoir validé par « OK » les paramètres de SCORE, exécutez alors la méthode.

LES RESULTATS DE SCORE

Coefficients des fonctions discriminante et de score

Libellés des modalités	Coefficients de la F.L.D.	Coefficients de la fonction de Score
Age du client		
moins de 23 ans	-6.395	0.00
de 23 à 40 ans	2.830	64.46
de 40 à 50 ans	1.971	58.46
plus de 50 ans	-0.946	38.07
Situation familiale		
célibataire	-0.615	31.95
marié	1.769	48.61
divorcé	-3.335	12.95
veuf	-5.188	0.00
Ancienneté		
anc. 1 an ou moins	-7.076	6.14
anc. de 1 à 4 ans	-7.955	0.00
anc. de 4 à 6 ans	8.594	115.63
anc. de 6 à 12 ans	4.312	85.72
anc. plus 12 ans	10.395	128.22
Domiciliation du salaire		
domicile salaire	5.537	119.12
non dimicile salaire	-11.511	0.00
Domiciliation de l'épargne		
pas d'épargne	-1.771	0.00
moins de 10KF épargn	3.506	36.87
de 10 à 100KF épargn	10.735	87.39
plus de 100KF épargn	13.555	107.09
Profession		
cadre	3.112	64.07
employé	2.925	62.77
autre	-6.057	0.00
Moyenne en cours		
moins de 2KF encours	-15.109	0.00
de 2 à 5 KF encours	2.918	125.96
plus de 5 KF encours	9.386	171.16
Moyenne des mouvements		
moins 10 KF mouvt	-4.056	0.00
de 10 à 30KF mouvt	0.003	28.36
de 30 à 50KF mouvt	0.877	34.47
plus de 50KF mouvt	4.485	59.68
Cumul des débits		
moins de 40 débits	6.608	109.89
de 40 à 100 débits	0.684	68.49
plus de 100 débits	-9.118	0.00
Autorisation de découvert		
découvert autorisé	-0.526	0.00
découvert interdit	0.399	6.47
Interdiction de chéquier		
chéquier autorisé	1.965	121.23
chéquier interdit	-15.384	0.00



Arbres de décision interactifs

La procédure IDT produit des arbres de décision à partir d'un ensemble de données, il s'agit d'une procédure discriminante destinée à prédire les valeurs d'une variable nominale (variable à expliquer, comportant K groupes) à partir d'un ensemble de variables explicatives qui peuvent être nominales, ordinales ou continues.

La procédure IDT donne à l'utilisateur le choix entre trois méthodes différentes qui font référence en Data Mining : CHAID, C&RT et C4.5. Le modèle produit par la méthode, l'arbre de décision, peut être évaluée à l'aide d'un échantillon test ou d'une validation croisée. La procédure intègre des informations supplémentaires qui permettent d'affiner les résultats : l'intégration du redressement à travers les probabilités a priori d'appartenance aux groupes et l'introduction d'une matrice de coût de mauvaise affectation.

La procédure IDT offre à l'utilisateur la possibilité de manipuler de manière interactive l'arbre de décision produit par la méthode : élagage de l'arbre à partir d'un sommet, segmentation interactive d'un sommet, modification des propriétés d'une segmentation.
La procédure offre également un mode totalement interactif où la construction de l'arbre repose entièrement sur les intuitions de l'utilisateur, plusieurs outils d'aide (liste des meilleures segmentations, statistiques descriptives…) lui permettent de choisir en connaissance de cause l'arbre qui correspond le mieux au problème à résoudre.

A toutes les étapes de l'étude réalisée par l'utilisateur, il est possible d'éditer des rapports au format HTML, globalement sur l'arbre de décision construit ou localement sur chaque sommet regroupant un sous-ensemble de la base de données analysée.

TRIS A PLATS DES VARIABLES

Type de client

	Effectif	% / Total
bon client	237	50.64
mauvais client	231	49.36
Total	468	100.00

Profession

	Effectif	% / Total
cadre	77	16.45
employé	237	50.64
profession autre	154	32.91
Total	468	100.00

Age du client

	Effectif	% / Total
moins de 23 ans	88	18.80
de 23 à 40 ans	150	32.05
de 40 à 50 ans	122	26.07
plus de 50 ans	108	23.08
Total	468	100.00

Moyenne en cours

	Effectif	% / Total
moins de 2KF encours	98	20.94
de 2 à 5 KF encours	308	65.81
plus de 5 KF encours	62	13.25
Total	468	100.00

Situation familiale

	Effectif	% / Total
célibataire	170	36.32
marié	221	47.22
divorcé	61	13.03
veuf	16	3.42
Total	468	100.00

Moyenne des mouvements

	Effectif	% / Total
moins 10 KF mouvt	154	32.91
de 10 à 30KF mouvt	71	15.17
de 30 à 50KF mouvt	129	27.56
plus de 50KF mouvt	114	24.36
Total	468	100.00

Ancienneté

	Effectif	% / Total
anc. 1 an ou moins	199	42.52
anc. de 1 à 4 ans	47	10.04
anc. de 4 à 6 ans	69	14.74
anc. de 6 à 12 ans	66	14.10
anc. plus 12 ans	87	18.59
Total	468	100.00

Cumul des débits

	Effectif	% / Total
moins de 40 débits	171	36.54
de 40 à 100 débits	161	34.40
plus de 100 débits	136	29.06
Total	468	100.00

Domiciliation du salaire

	Effectif	% / Total
domicile salaire	316	67.52
non dimicile salaire	152	32.48
Total	468	100.00

Autorisation de découvert

	Effectif	% / Total
découvert autorisé	202	43.16
découvert interdit	266	56.84
Total	468	100.00

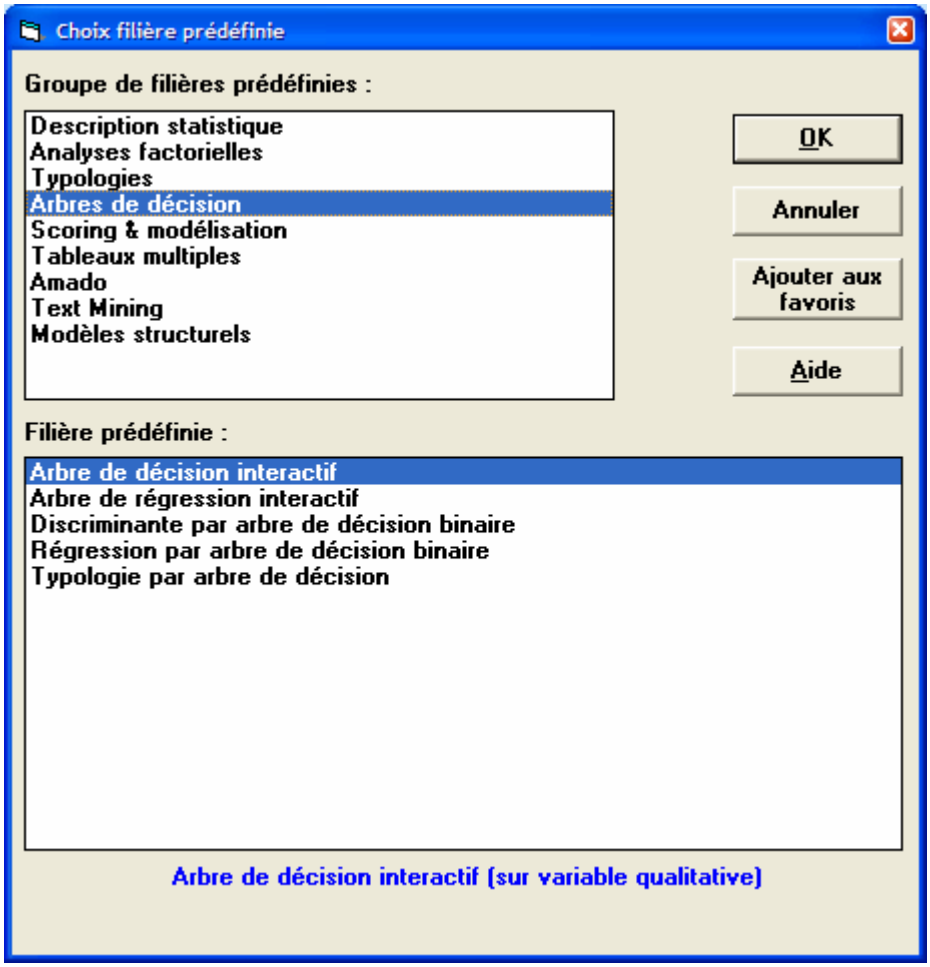
Domiciliation de l'épargne

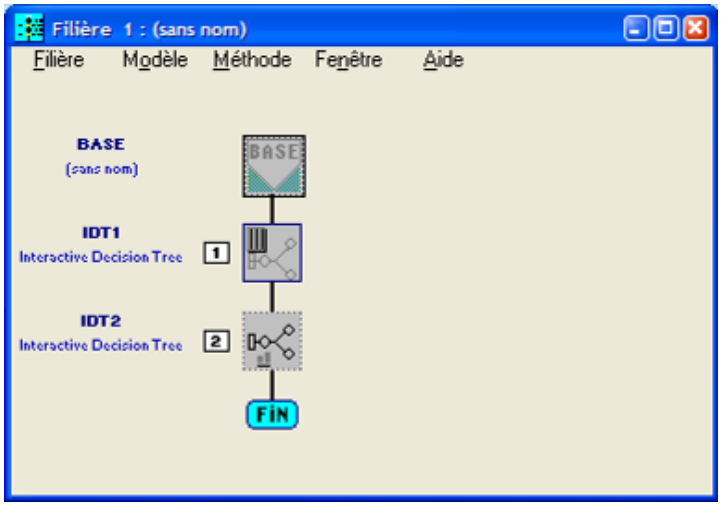
	Effectif	% / Total
pas d'épargne	370	79.06
moins de 10KF épargn	58	12.39
de 10 à 100KF épargn	32	6.84
plus de 100KF épargn	8	1.71
Total	468	100.00

Interdiction de chéquier

	Effectif	% / Total
chéquier autorisé	415	88.68
chéquier interdit	53	11.32
Total	468	100.00

L'ARBRE DE DECISION INTERACTIF



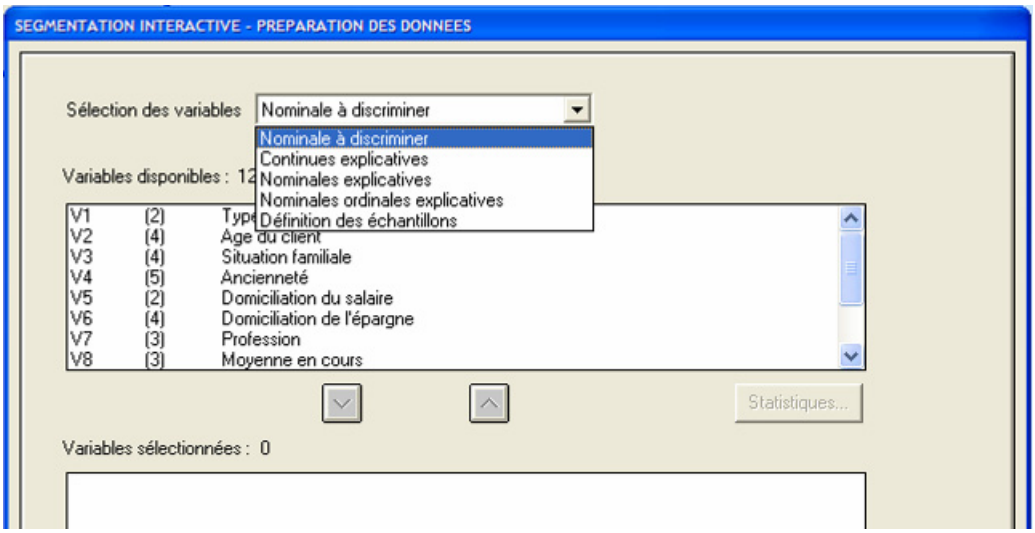


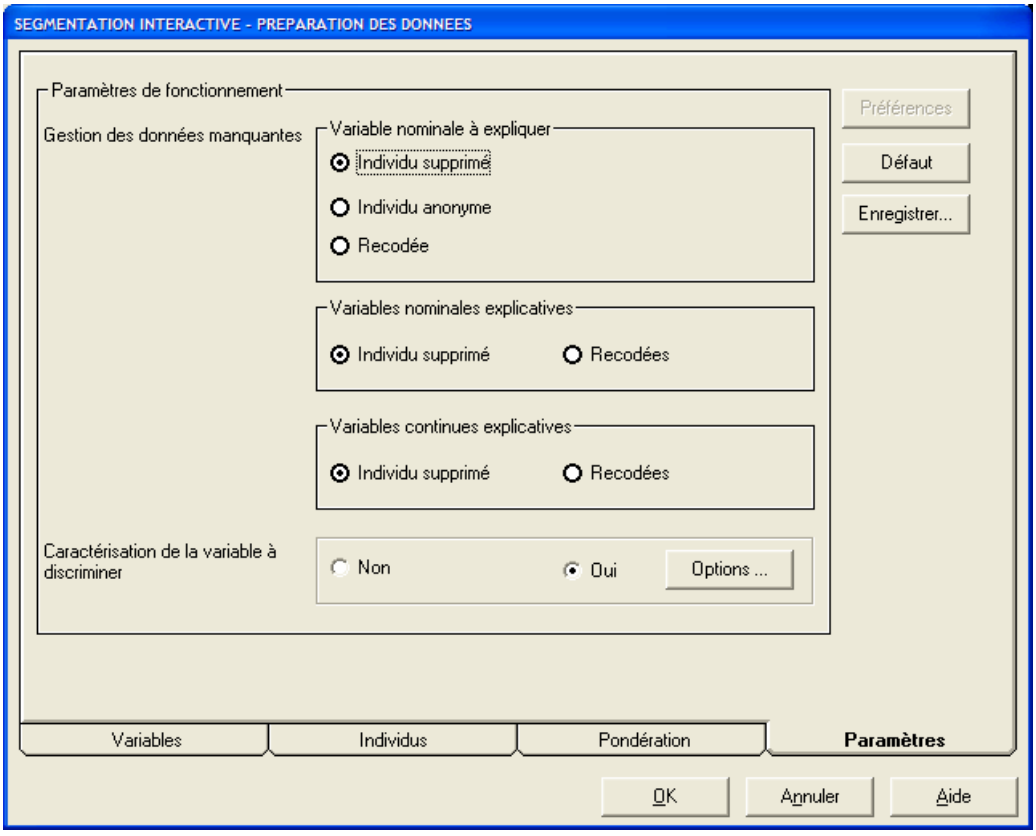
LE PARAMETRAGE DE IDT 1

La procédure IDT1 prépare les données pour la construction de l'arbre (procédure IDT2), en particulier elle gère les données manquantes des variables sélectionnées. La procédure édite un bilan de la gestion des données manquantes.

Vous disposez également par défaut d'une caractérisation automatique de la variable à discriminer par l'ensemble des variables explicatives sélectionnées.

Cette caractérisation vous permet une meilleure sélection des variables explicatives, en retirant par exemple toutes celles qui n'ont aucune liaison avec la variable à discriminer.





LE PARAMETRAGE DE IDT 2

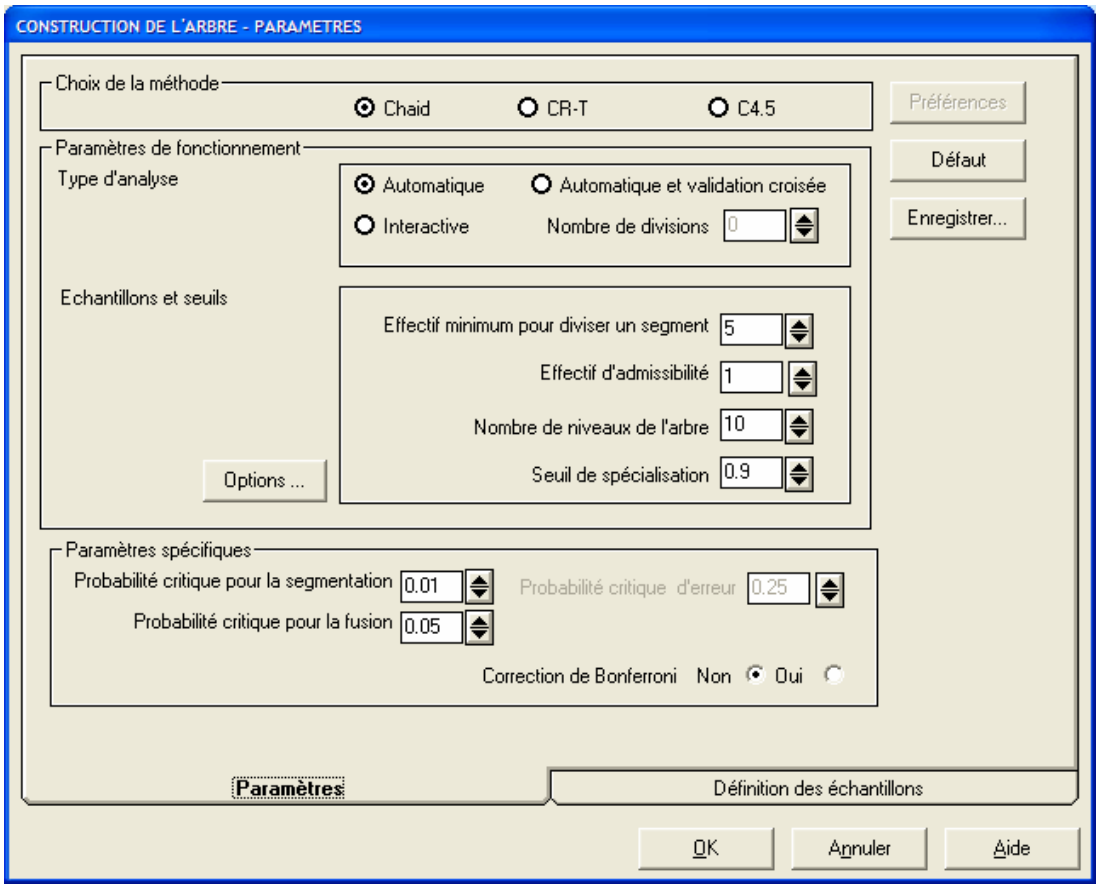
La procédure IDT2 construit l'arbre de segmentation initial en fonction de la méthode choisie (CHAID, CR-T, C4.5) et des paramètres associés.

La méthode CHAID

CHAID est une méthode d'induction d'arbre de décision reposant sur un critère de discrimination statistique, la mesure du Chi-2. Il s'agit vraisemblablement d'une des méthodes les plus anciennes, elle prend ses racines dans des travaux effectués au milieu des années 60. L'algorithme développé dans IDT est dû à Kass, 1980.
Il possède deux particularités par rapport aux autres méthodes d'induction d'arbre :

- ✓ La détermination de la bonne taille de l'arbre s'effectue par pré-élagage, c'est-à-dire lors de l'expansion de l'arbre. La décision de segmenter un sommet dépend d'un test d'indépendance du Chi-2 effectué sur le tableau de contingence associé aux feuilles qui seront produites par la segmentation. Si ce test est négatif, le sommet n'est pas segmenté et devient un sommet terminal.
- ✓ La méthode procède, éventuellement, à un regroupement des modalités de la variable de segmentation. L'arbre produit est donc n-aire, l'algorithme recherche les regroupements les plus appropriés compte tenu des paramètres fixés par l'utilisateur.

La méthode CHAID est particulièrement appropriée si le temps de calcul est un critère important pour l'utilisateur. Elle est indiquée lorsque l'on veut procéder à une première exploration des données.

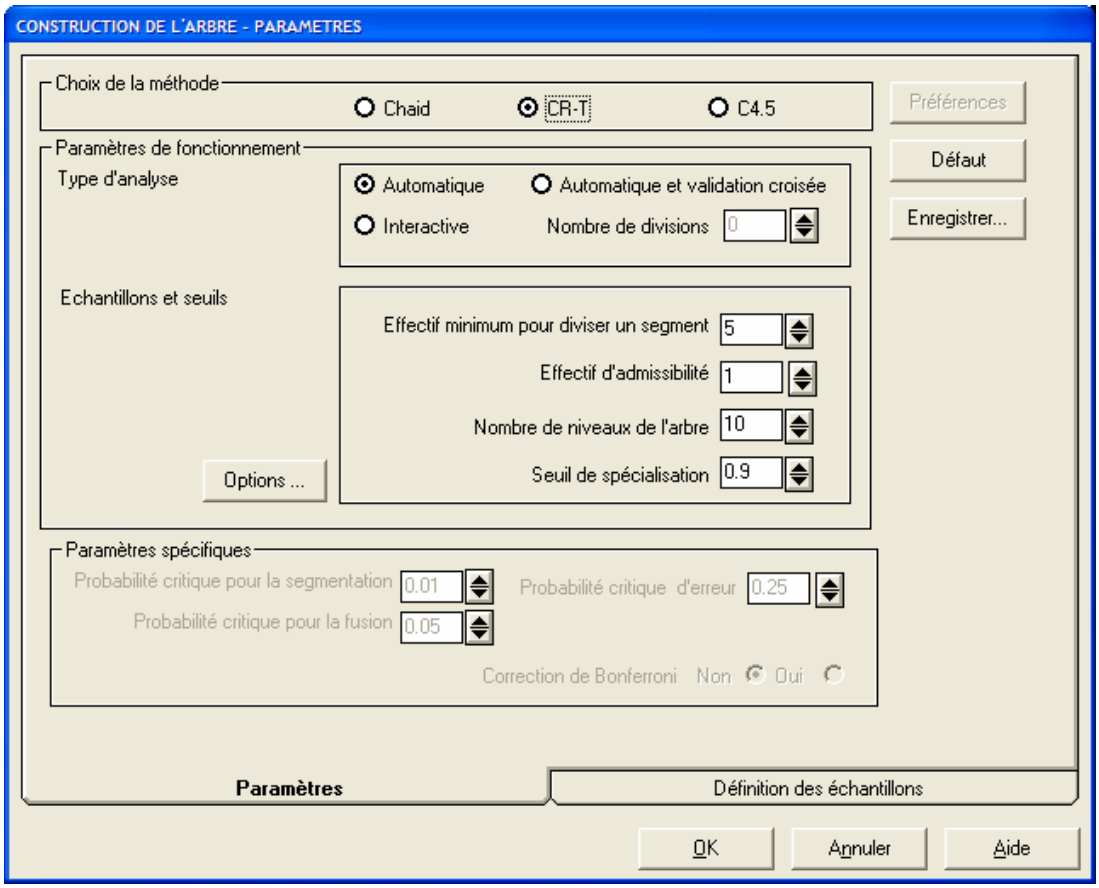


La méthode CR-T

CR-T est issue d'une monographie de Breiman et al. (1984) qui propose une approche unifiée pour traiter les problèmes de discrimination (la variable à prédire est qualitative) et de régression (la variable à prédire est quantitative) à l'aide d'un arbre. Dans le cadre de la discrimination, le critère utilisé repose sur la notion de « pureté », il est également possible de l'interpréter comme une analyse de variance sur données catégorielles.
CR-T possède deux particularités :

- ✓ La détermination de la bonne taille de l'arbre s'effectue par post-élagage, c'est-à-dire l'arbre est dans un premier temps complètement développé avec le critère de pureté sur un premier échantillon, puis, dans un second temps, il est réduit de manière à optimiser le taux de mauvais classement calculé sur un second échantillon dit d'élagage. Lors de cette seconde phase, il est possible d'introduire une matrice de coût de mauvais classement.
- ✓ Sur chaque segmentation, la méthode procède à un regroupement de manière à ce que l'arbre soit systématiquement binaire, càd chaque sommet segmenté ne possède que deux enfants.

CR-T construit généralement des arbres très « compacts », ayant de bonnes capacités de prédiction. Elle est assez lente à cause du dispositif de post-élagage. Du fait de la nécessité de subdiviser l'échantillon initial en échantillon d'apprentissage et d'élagage, cette méthode n'est pas très appropriée lorsque la taille de la base de données est faible.

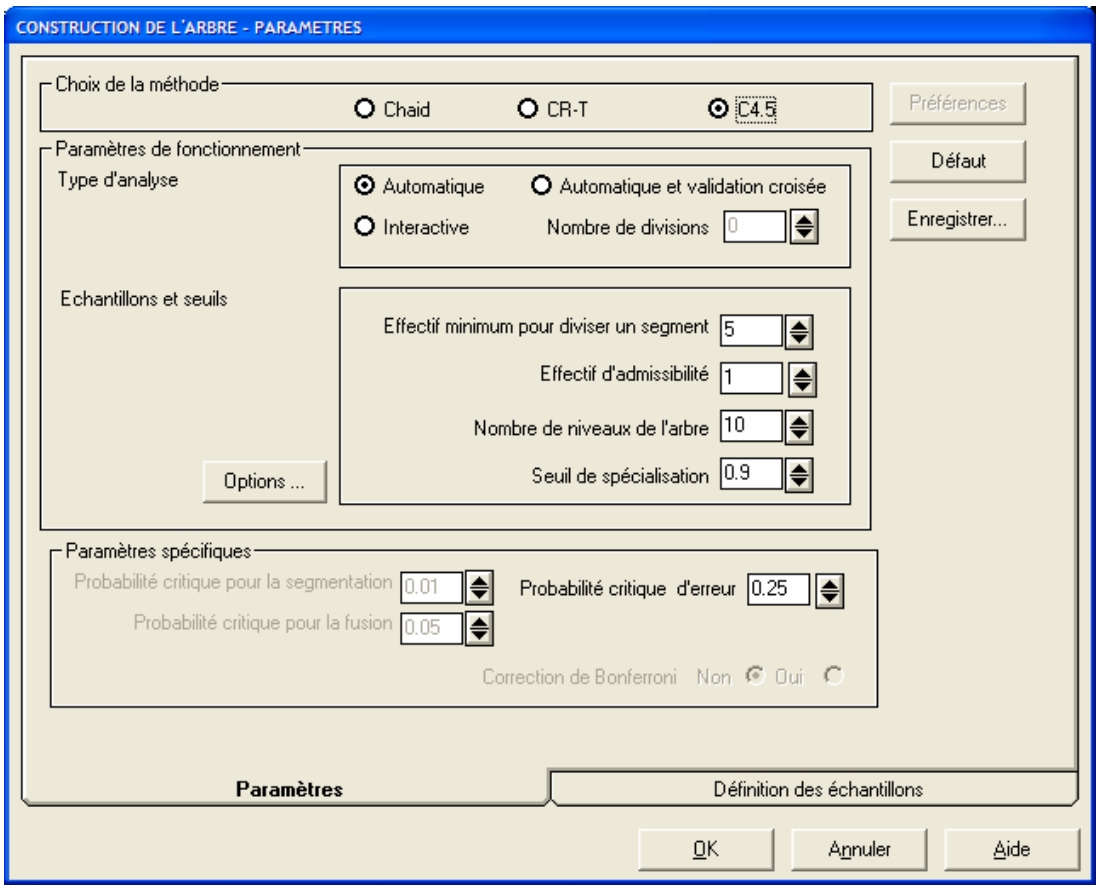


La méthode C4.5

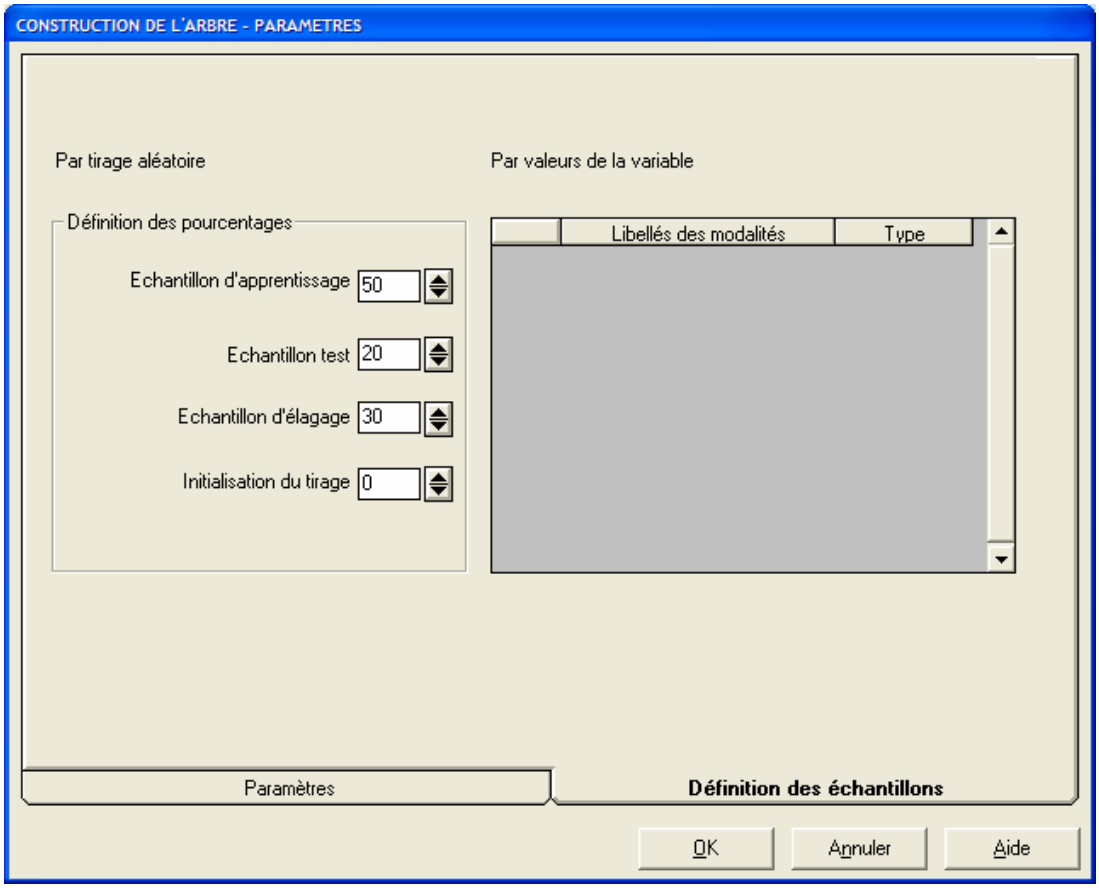
C4.5 est une méthode d'induction d'arbre très répandue au sein de la communauté de l'intelligence artificielle. Elle prend ses racines dans des travaux anciens sur la théorie de l'information, le critère de discrimination utilisé repose sur la notion de gain informationnel.
Cette méthode possède deux particularités :

- ✓ La détermination de la bonne taille de l'arbre s'effectue par post-élagage, c'est-à-dire l'arbre est dans un premier temps complètement développé avec le critère du gain informationnel, puis, dans un second temps, il est réduit de manière à optimiser la capacité à bien classer les observations. Le critère utilisé est alors le taux de mauvais classement.
- ✓ Lorsque la variable de segmentation est catégorielle, chacune de ses modalités correspond à une feuille, même si cette dernière ne recouvre aucun individu.

C4.5 produit des arbres « larges », avec beaucoup de feuilles. Le temps de calcul, à cause du dispositif de post-élagage, est un peu plus élevé.



La définition des échantillons



Par tirage aléatoire :

Si vous n'avez pas choisi de variable « Définition des échantillons » dans la méthode de préparation des données IDT1, le choix des échantillons se fait par tirage aléatoire. Vous devez préciser la taille des échantillons (Apprentissage, Test, Elagage (méthode CR-T)) en pourcentages.

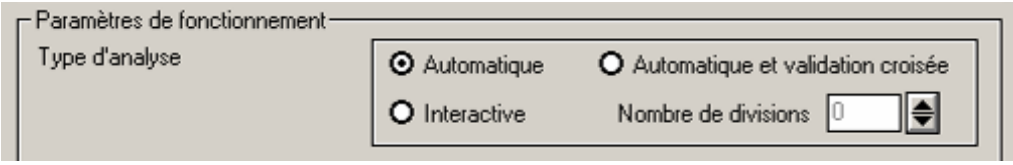
Le paramètre « Initialisation du tirage » vous permet de tirer des échantillons différents.

Par valeur de la variable :

Si vous avez choisi une variable « Définition des échantillons » dans la méthode de préparation des données IDT1, le choix des échantillons se fait en indiquant le type d'échantillon correspondant à chaque modalité.

Paramètres de Fonctionnement

Type d'analyse :



Valeur par défaut : Automatique

Automatique :

L'arbre est construit entièrement en respectant les paramètres sélectionnés par l'utilisateur.

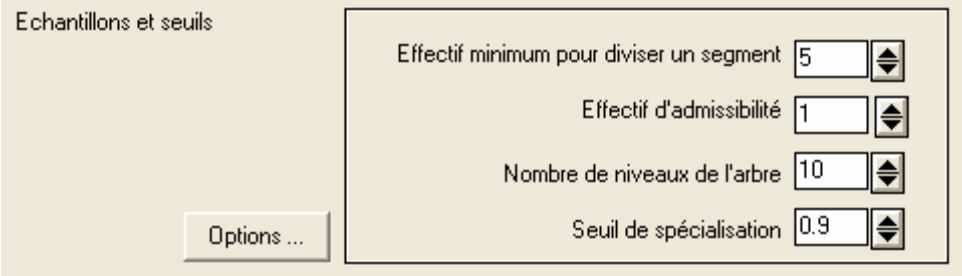
Automatique avec validation croisée :

L'arbre est construit entièrement en respectant les paramètres sélectionnés par l'utilisateur, la procédure en évalue l'erreur en procédant à une validation croisée.
Dans ce cas vous devez spécifier le nombre de divisions pour la validation croisée.

Interactive :

Seule la racine de l'arbre est construite, ce sommet sera le même quelle que soit la méthode mise en œuvre. Le développement de l'arbre se fera dans l'application graphique.

Echantillons et seuils :



Effectif minimum pour diviser un segment : (par défaut : 5)

Ce paramètre désigne le nombre minimal d'individus requis sur un sommet pour effectuer une segmentation. En dessous de ce seuil, le sommet ne sera pas segmenté même si le gain d'information est positif.

En augmentant la valeur de ce paramètre, on réduit la taille de l'arbre.

Effectif d'admissibilité : (par défaut : 1)

Ce paramètre indique la taille minimum d'au moins deux des sommets produits par une segmentation pour que cette dernière soit validée.

Plus grande sera la valeur de ce paramètre, plus petit sera l'arbre de décision produit par la méthode.

Nombre de niveaux de l'arbre : (par défaut : 10)

Ce paramètre limite la profondeur de l'arbre.

Si vous avez choisi « Interactive » pour le type d'analyse, le nombre de niveaux est égal à 1. Le développement de l'arbre se fera dans l'application graphique.

Seuil de spécification : (par défaut : 0.9)

Ce seuil indique la probabilité limite à partir de laquelle une des modalités de la variable à prédire est considérée comme seule présente sur le sommet. Si la probabilité d'occurrence d'une des modalités de la variable à prédire est supérieure à ce seuil, toute segmentation est refusée.

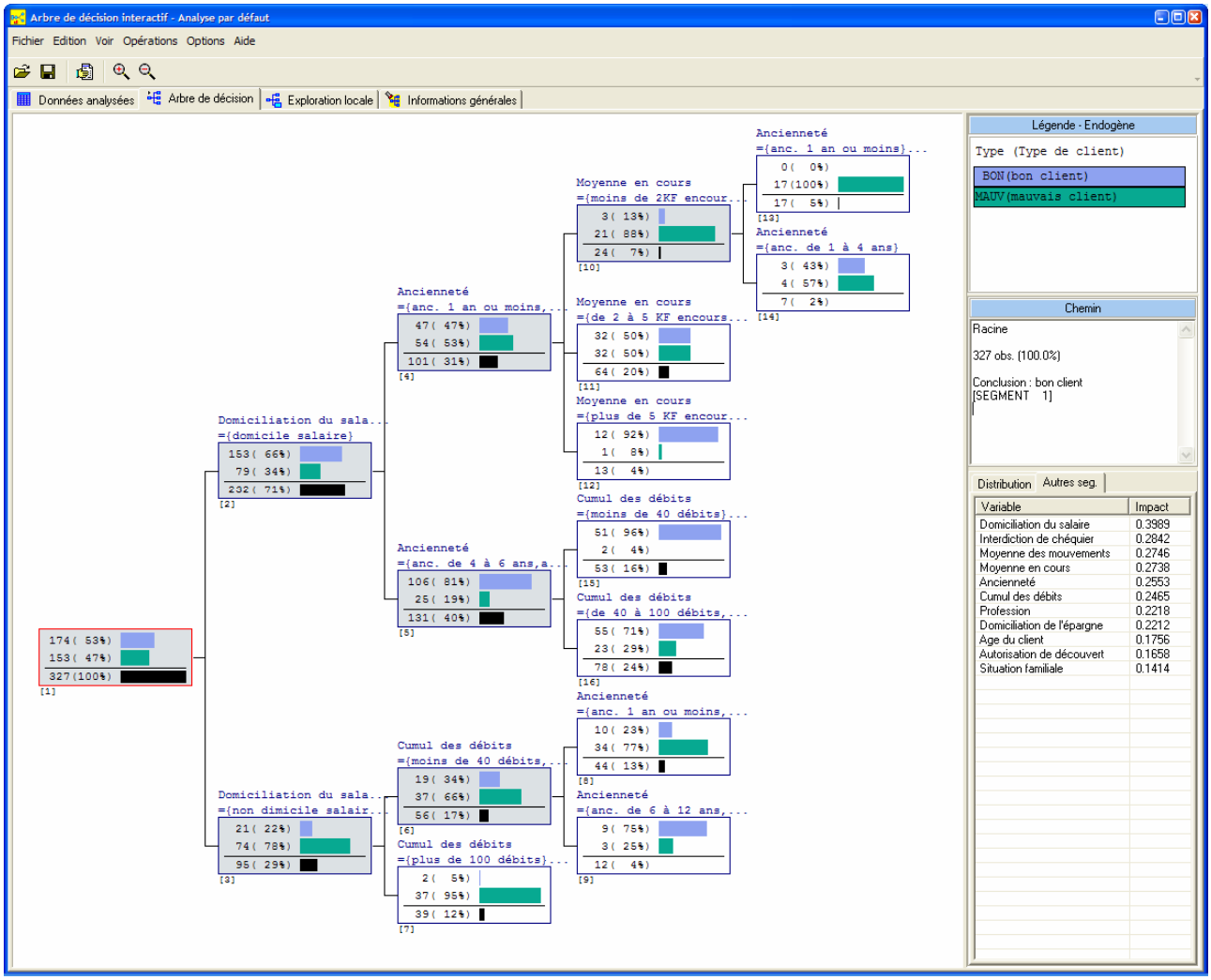
En réduisant la valeur de cette option, on limite la taille de l'arbre.

Après exécution de la procédure, vous disposez dans les résultats d'une sortie intitulée « Interactive Decision Tree » permettant d'accéder à la visualisation graphique de l'arbre.

EDITEUR ARBRE DE DECISION INTERATIF

Fenêtres de résultats

L'outil de visualisation de l'arbre de décision produit par la procédure IDT intègre plusieurs fenêtres, regroupées dans une page à onglets, elles correspondent à différents niveaux d'informations relatives au modèle construit.



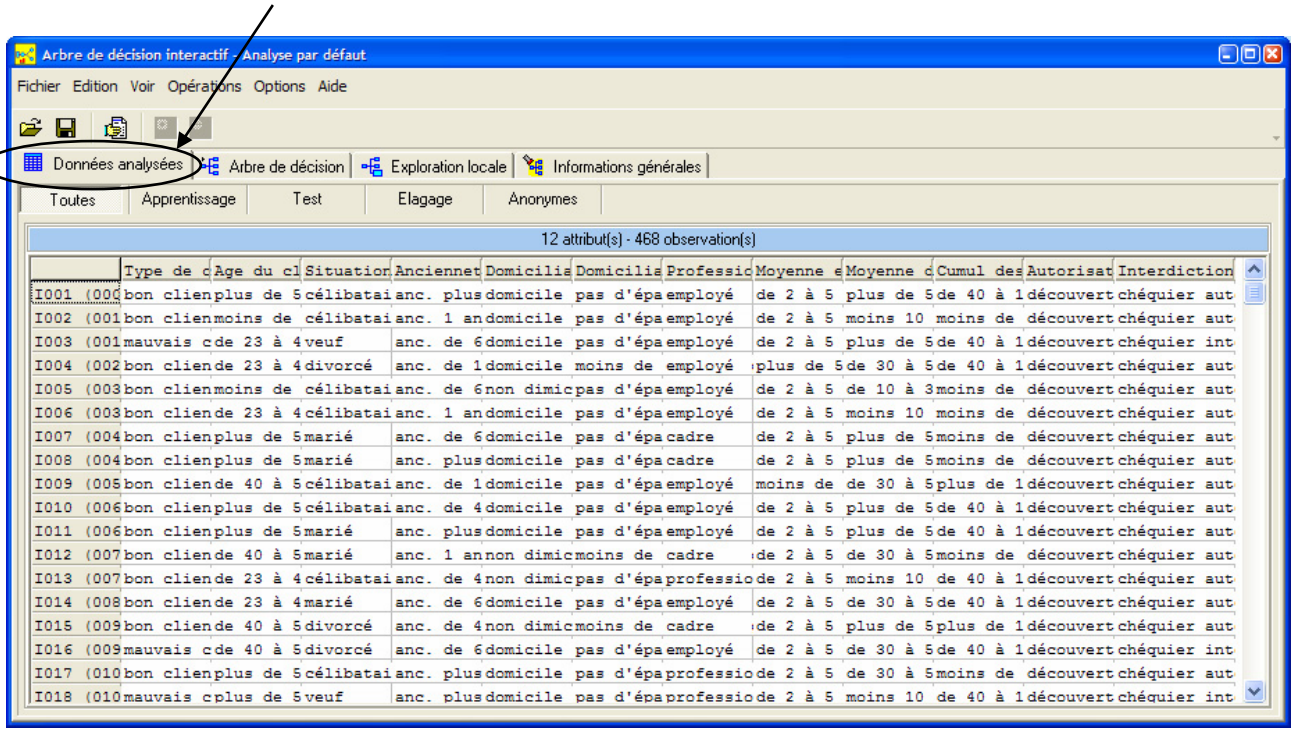
Visualiser les données

Cette fenêtre présente dans une grille l'ensemble de données qui est en cours d'analyse. Utilisé conjointement avec la fenêtre d'information sur les sommets, elle permet de suivre le chemin parcouru par un individu dans l'arbre de décision.

Il est possible de copier le contenu de la grille dans le presse papier, afin d'être collé dans un tableur par exemple.

La grille de données est au format « Individus x Variables » :

- la colonne la plus à gauche, grisée, correspond à l'identifiant des individus;
- la colonne suivante représente la variable à expliquer;
- les colonnes suivantes représentent les variables explicatives;
- la colonne la plus à droite indique le poids associé à chaque individu.



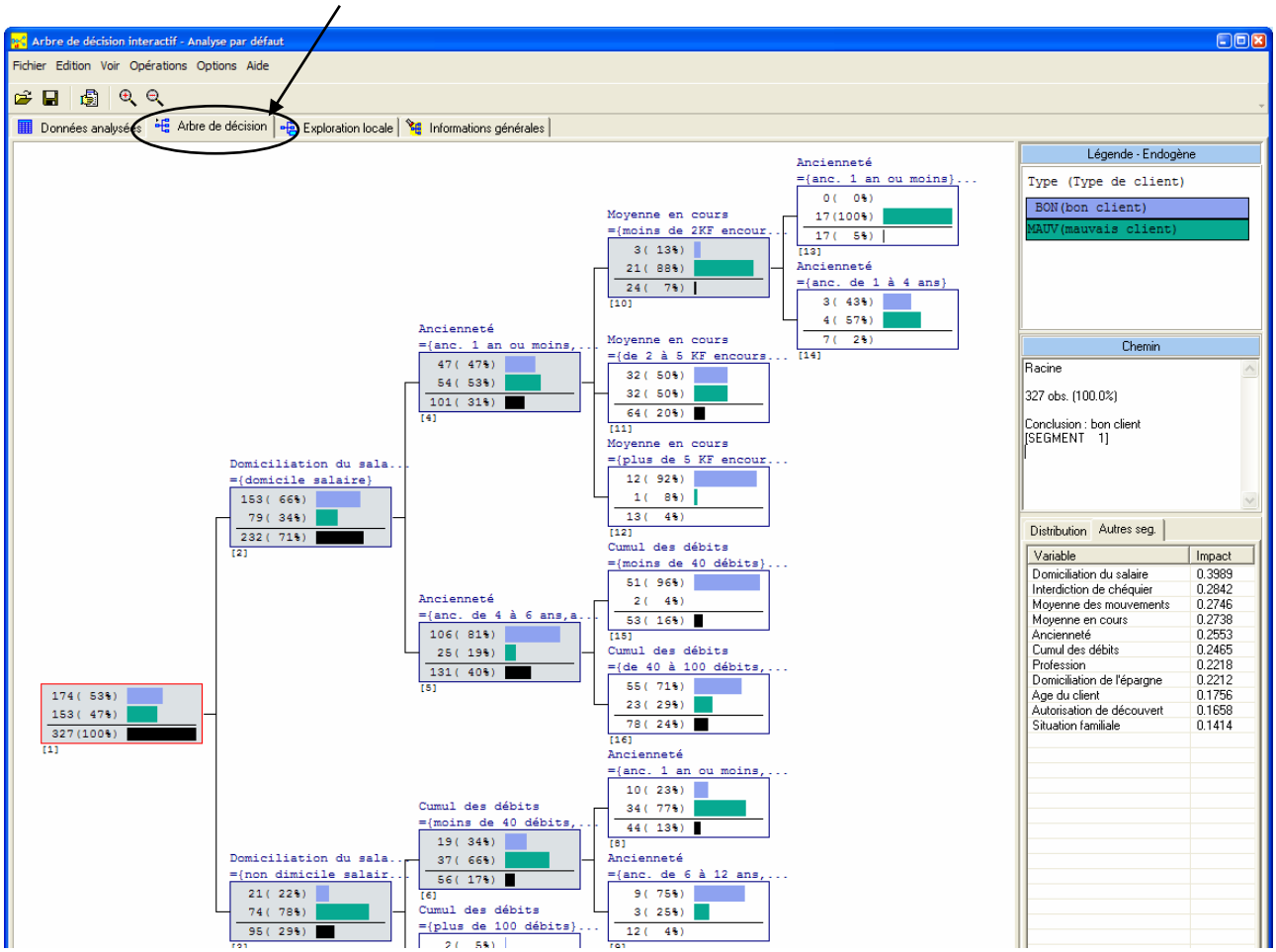
Visualiser l'arbre de décision

Cette fenêtre propose une représentation graphique de l'arbre de décision. Il est possible de moduler l'échelle d'affichage en activant les commandes de zoom (menu Edition - Zoom avant / Zoom arrière, ou en cliquant sur les icônes correspondantes dans la barre d'outils).

L'arbre est représenté transversalement, partant de la racine, à gauche, vers les feuilles, à droite. Chaque sommet propose la distribution de probabilité conditionnelle estimée de la variable à prédire, en absolu (effectifs réels) et en relatif (pourcentages). En haut à droite de la fenêtre, une légende, associant les modalités aux codes de couleurs employés, est disponible. Attention, si un redressement a été demandé, l'outil affiche les probabilités estimées redressées. Dans la partie haute du sommet est affiché la règle de décision (variable - opérateur - valeur) relative à la création du sommet.

En cliquant sur un sommet de l'arbre, il est possible d'obtenir des informations supplémentaires qui sont proposées dans la partie droite de la fenêtre : le chemin complet allant de la racine jusqu'au sommet activé, et la pertinence des variables candidates à la segmentation. Cette dernière peut être triée selon le nom de la variable ou selon la valeur de la qualité de segmentation (cliquer sur l'en-tête de la liste).

Il est également possible d'approfondir l'exploration du sous-ensemble d'individus circonscrit par le sommet ou encore de diriger de manière interactive l'analyse.



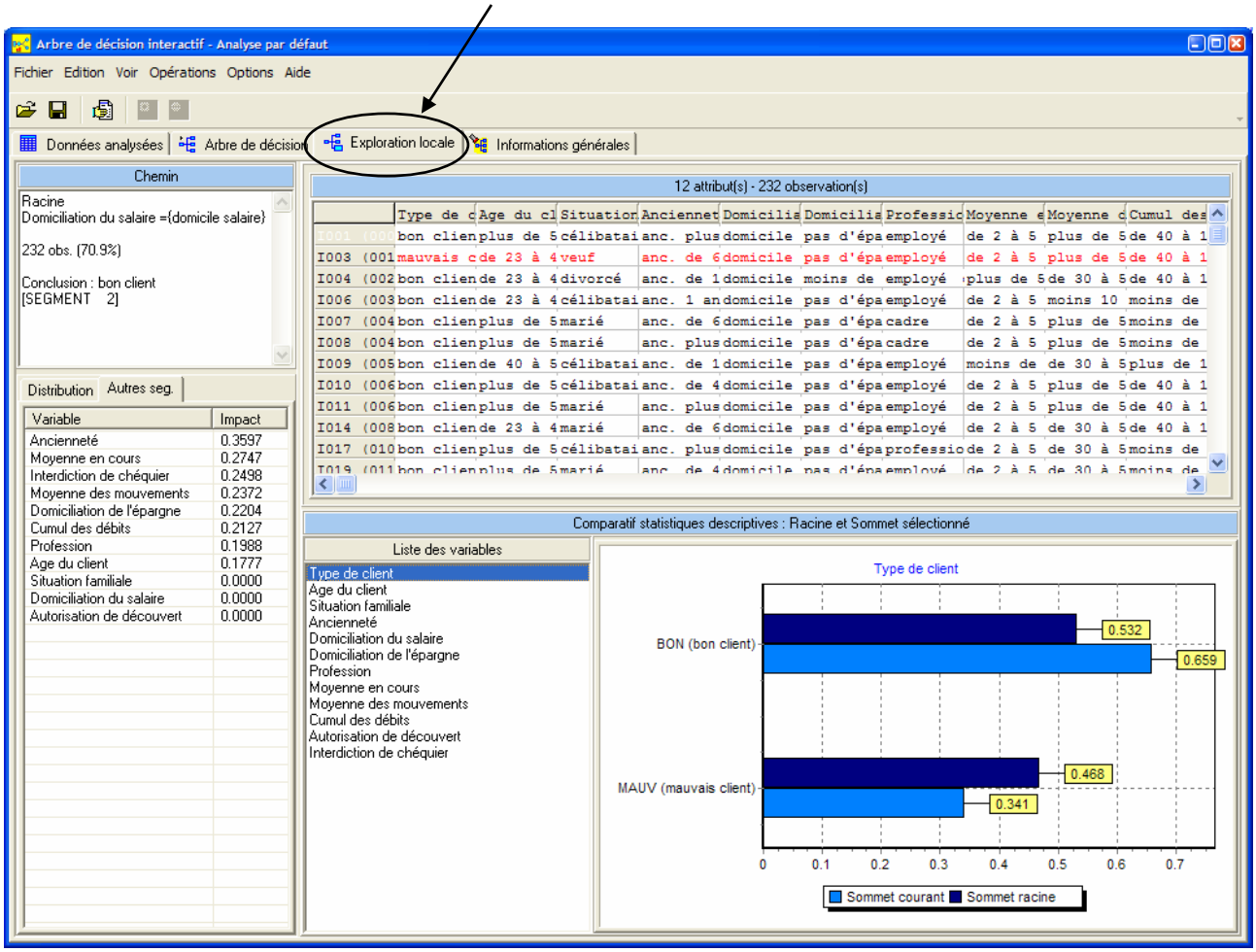
Informations sur les sommets

Lorsque l'on a cliqué sur un sommet particulier, il est possible d'en effectuer une analyse approfondie en basculant sur la fenêtre « Exploration locale » (menu Voir - Exploration locale ou clic sur l'onglet correspondant).

Les informations sur les chemins et la pertinence des variables sont réitérées.

Il est également possible de visionner les individus (et leurs valeurs pour chaque variable de l'analyse) présent sur le sommet sélectionné. Notons qu'à chaque sommet correspond une conclusion affectée par la méthode, ainsi les individus qui ne correspondent pas cette conclusion sont affichés en rouge.

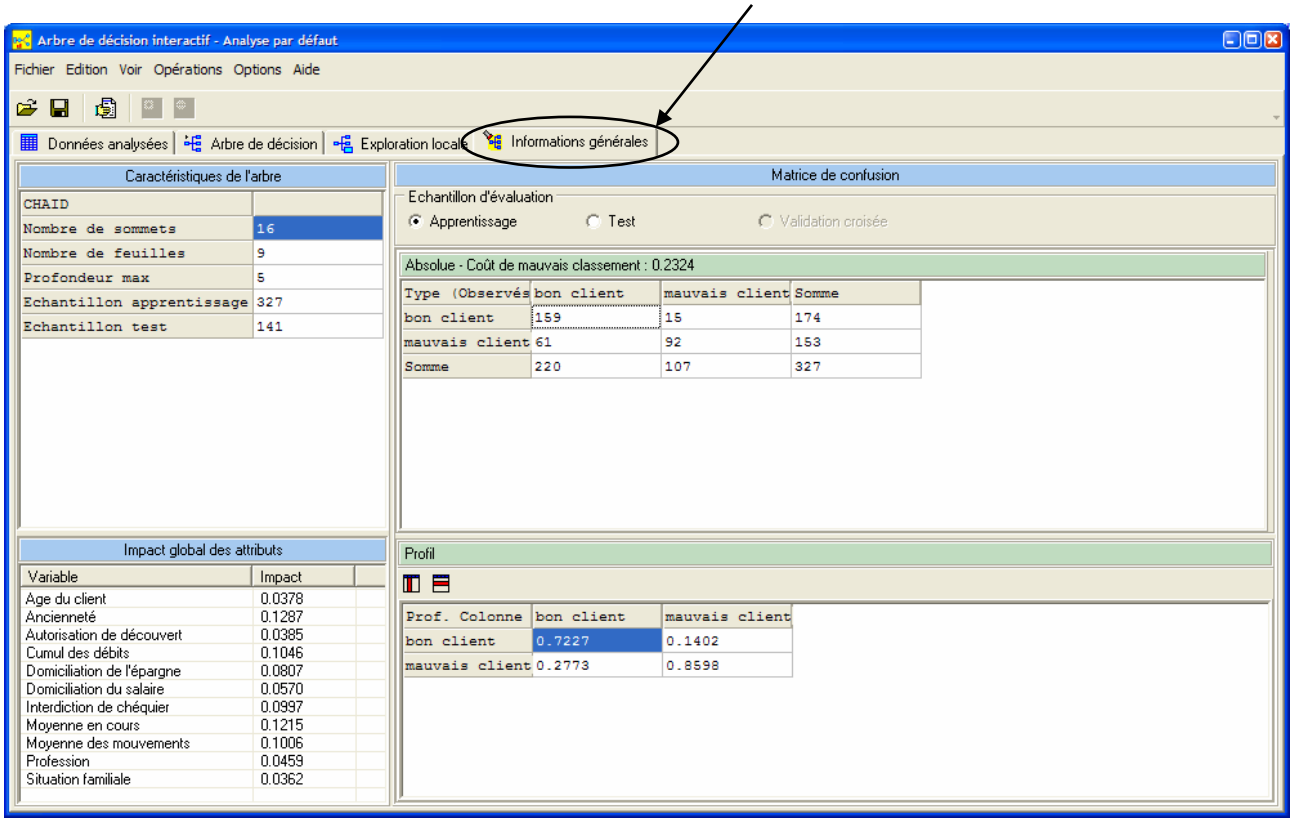
Enfin, il est possible d'approfondir l'analyse du sommet en demandant pour chaque variable, dans la partie basse de la fenêtre, des statistiques descriptives comparatives avec l'ensemble des individus (la racine de l'arbre).



Informations sur l'arbre de décision

Cette fenêtre permet de juger de la qualité de l'arbre de décision. La fenêtre est subdivisée en plusieurs parties :

- **Caractéristiques de l'arbre** : indique les propriétés de l'arbre de décision produit par la méthode. On y recense le nombre de sommets dans l'arbre, le nombre de feuilles, sa profondeur maximum. Y figurent également la taille de l'échantillon utilisé pour l'apprentissage, pour le test, et le cas échéant, pour l'élagage.
- **Impact des attributs** : affiche le rôle de chaque attribut dans l'élaboration de l'arbre. La valeur indiquée représente une moyenne pondérée de l'impact de chaque attribut sur toutes les segmentations candidates. On donne moins d'importance aux impacts mesurés sur les parties basses de l'arbre.
- **Matrice de confusion** : recense la confrontation entre les prédictions de l'arbre et les valeurs observées sur la variable à prédire. La matrice peut être mesurée sur l'échantillon d'apprentissage, sur l'échantillon test ou en validation croisée (ces deux dernières options sont actives si elles ont été demandées lors du paramétrage de la procédure). (De cette matrice est déduit le coût de mauvais classement, qui est un taux d'erreur lorsque la matrice de coût est unitaire.)
- **Profil** : présente la matrice de confusion courante sous forme de profil ligne (pour mesurer les sensibilités) ou sous forme de profil colonne (pour mesurer les spécificités).



Manipuler l'arbre de décision

L'originalité de la procédure IDT repose en grande partie sur la possibilité offerte à l'utilisateur de modeler à sa convenance l'arbre de décision, soit en modifiant un arbre produit par une méthode d'induction, soit en le construisant entièrement en se basant sur ses connaissances d'expert.

Plusieurs outils sont à la disposition de l'utilisateur, ils lui permettent de définir les propriétés des segmentations sur les sommets, ils lui permettent également de réduire les parties de l'arbre qui lui semblent peu pertinentes.

Les opérateurs que l'on peut mettre en œuvre peuvent s'appliquer, soit sur un ensemble de sommets (les feuilles plus généralement), soit sur un sommet que l'on aura au préalable sélectionné.

Manipulations sur un sommet de l'arbre

En effectuant un clic droit sur un sommet, un menu contextuel apparaît. Selon le statut du sommet (feuille ou sommet interne), différentes options sont disponibles, elles permettent à l'utilisateur de définir avec précision l'arbre qui lui convient le mieux pour l'analyse qu'il est en train d'effectuer.

Deux opérateurs principaux sont disponibles : élaguer pour un sommet interne de l'arbre, segmenter pour une feuille de l'arbre.

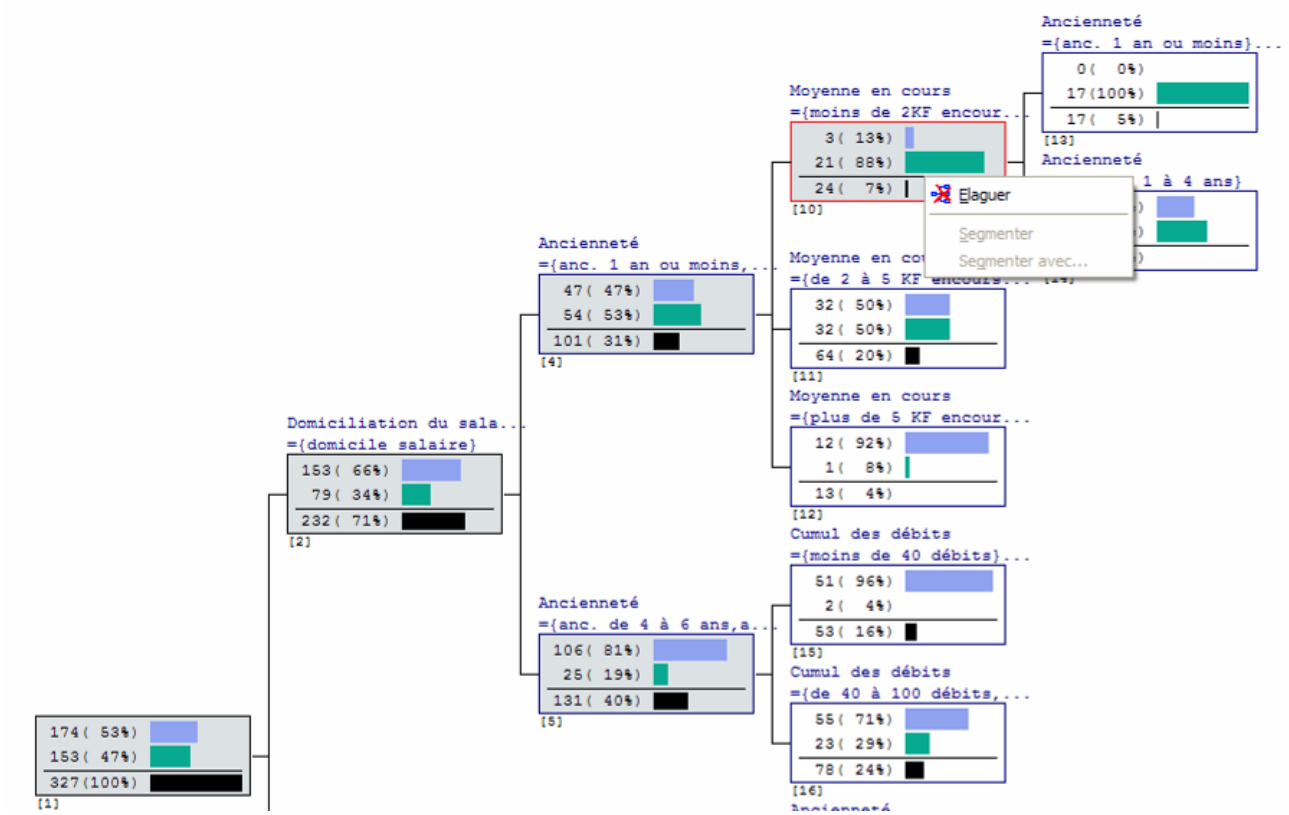
Elaguer un sous-arbre

L'élagage d'un sous-arbre consiste à supprimer les sommets et feuilles situés en deçà d'un sommet que l'on a préalablement sélectionné. Cette opération est nécessaire lorsque l'on considère que le sous-arbre correspondant n'apporte par d'informations supplémentaires dans l'analyse en cours ou lorsque l'on veut induire manuellement une autre segmentation à partir du sommet sélectionné.

Attention, cette opération n'est possible que sur les sommets internes de l'arbre.

Comment procéder ?

1. Sélectionner le sommet à partir duquel vous voulez procéder à l'élagage
2. Effectuer un clic droit - Le menu *Elaguer* est disponible si le sommet n'est pas une feuille
3. Cliquer sur le menu *Elaguer*



Segmenter un sommet de l'arbre

On dispose, sur chaque sommet de l'arbre, de la liste des variables candidates à la segmentation, avec leurs impacts respectifs. L'utilisateur peut à sa convenance trier ces variables selon leur nom ou selon leur pertinence afin de retrouver les variables qui l'intéressent.

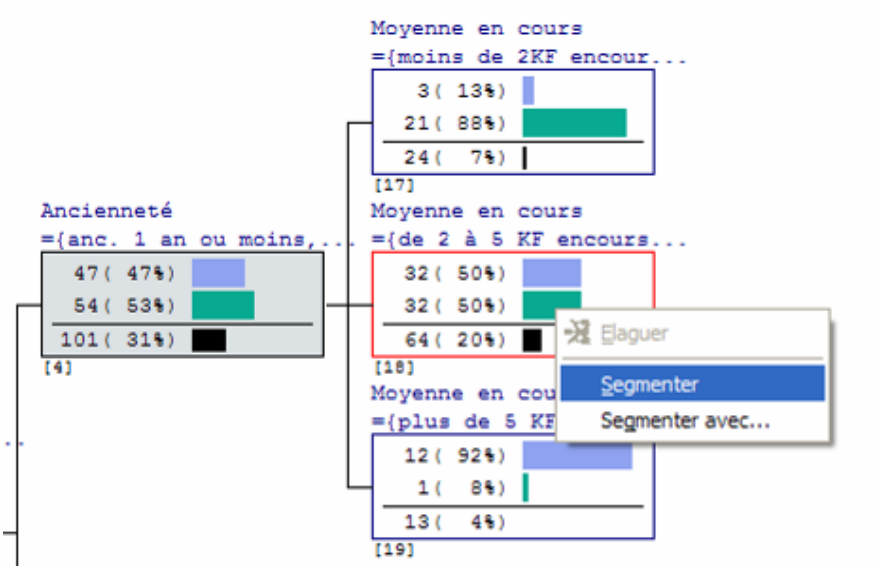
Une première originalité marquante de la procédure IDT est la possibilité donnée à l'utilisateur d'introduire la segmentation qui lui semble la plus pertinente, soit en suivant les propositions de la méthode, soit en choisissant lui même la variable de segmentation.

Une seconde originalité est la possibilité donnée à l'utilisateur de modifier à sa guise les propriétés d'une segmentation, en lui permettant d'introduire, par exemple, la borne de discrétisation pour une variable continue (ex. la méthode propose, pour une segmentation à partir de l'âge, de découper à partir de la borne 17.5, l'utilisateur, en se fondant sur ses propres connaissances du problème, peut décider de modifier cette valeur en introduisant manuellement la borne 18 qui correspond au passage à la majorité).

La segmentation est impossible dans trois cas particuliers :

- le sommet n'est pas une feuille : il a déjà été segmenté et possède déjà des sommets enfants
- le sommet est vide : il n'y a pas d'individus sur le sommet
- le sommet est pur : une seule des modalités de la variable à prédire est représenté sur le sommet, dans ce cas la règle de décision est sans équivoque, il est inutile d'essayer d'approfondir l'analyse

Attention, dans ce cadre, les règles d'arrêt d'expansion de l'arbre deviennent inactives (ex. effectif minimum sur un sommet, seuil de spécialisation, etc.).



Modifier les propriétés d'une segmentation

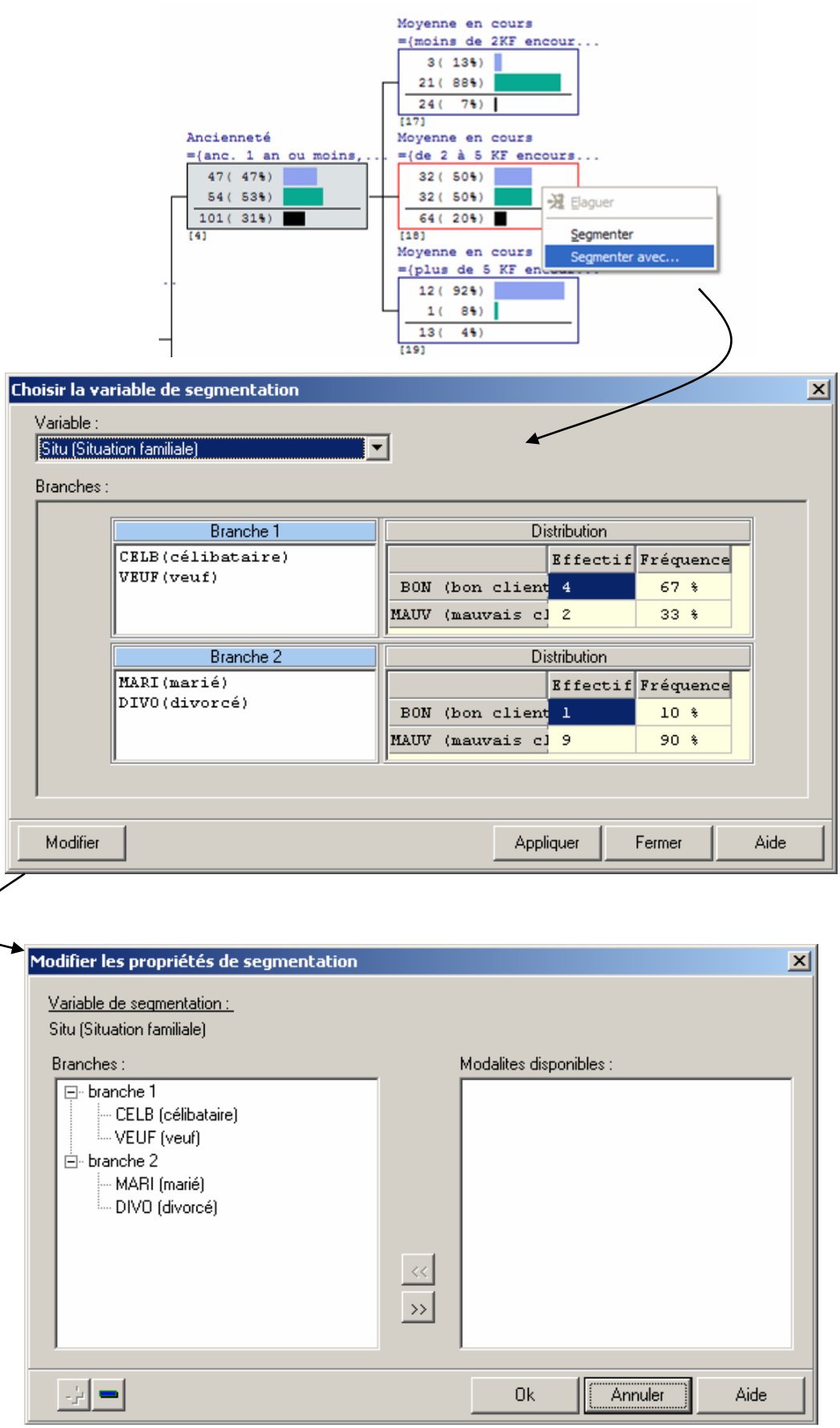
IDT permet à l'utilisateur de sélectionner la variable la plus pertinente pour une segmentation, il lui permet également de modifier les propriétés de la segmentation qu'il a sélectionné.

Selon le type de variable en jeu dans la segmentation, la procédure est différente :

- modification du seuil de discrétisation pour les variables continues,
- modification des regroupements pour les variables catégorielles (nominales ou ordinales).

Comment procéder ?
La procédure de modification des propriétés de la segmentation possède une partie commune avec la procédure de segmentation manuelle.

1. Sélectionner le sommet que l'on veut segmenter
2. Effectuer un clic droit sur ce sommet - Pour que le menu *Segmenter avec...* soit actif, il faut que le sommet soit une feuille et qu'une segmentation est possible
3. Dans la boîte de dialogue qui apparaît, on observe le liste des variables explicatives candidates et la segmentation qu'elles proposent. Le tri des variables respecte le tri demandé dans la fenêtre *Arbre de Décision*
4. Pour modifier les propriétés de la variable en cours de sélection, il faut alors cliquer sur le bouton *Modifier*
5. Selon le type de la variable, deux boîtes de dialogues différentes apparaissent
   - pour les **variables continues**
     - la boîte de dialogue indique la variable sur laquelle nous travaillons et propose la borne de discrétisation qui a été utilisée jusqu'à présent
     - l'utilisateur doit alors entrer son nouveau seuil, attention la zone d'édition n'accepte que les valeurs numériques, et le point décimal est un '.'
     - validez alors votre nouvelle borne en cliquant sur le bouton *Ok*
   - pour les **variables catégorielles** (nominales - ordinales)
     - la boîte de dialogue indique, dans la liste de gauche, les branches (les feuilles issues de la segmentation) en cours d'édition, et dans la liste de droite, les modalités disponibles pour élaborer les branches
     - pour modifier le contenu d'une branche, il faut tout passer son contenu (les modalités de la variable explicative) dans la liste de droite à l'aide du bouton « >> », puis lui les transférer, le cas échéant, vers une autre branche, à l'aide du bouton « << »
     - il est possible d'ajouter ou de supprimer une branche à l'aide des boutons « + » et « - »
     - lorsque les modifications sont terminées, il faut valider la nouvelle segmentation à l'aide du bouton Ok


Moyenne en cours
=(moins de 2KF encour...
3(13%)
21(88%)
24(7%)
[17]
Ancienneté
=(anc. 1 an ou moins,...
47(47%)
54(53%)
101(31%)
[4]
Moyenne en cours
=(de 2 à 5 KF encours...
32(50%)
32(50%)
64(20%)
[18]
Elaguer
Segmenter
Segmenter avec...
Moyenne en cours
=(plus de 5 KF en
12(92%)
1(8%)
13(4%)
[19]
Choisir la variable de segmentation
Variable :
Situ (Situation familiale)
Branches :
Branche 1
CELB (célibataire)
VEUF (veuf)
Distribution
Effectif
Fréquence
BON (bon client
4
67 %
MAUV (mauvais cl
2
33 %
Branche 2
MARI (marié)
DIVO (divorcé)
Distribution
Effectif
Fréquence
BON (bon client
1
10 %
MAUV (mauvais cl
9
90 %
Modifier
Appliquer
Fermer
Aide
Modifier les propriétés de segmentation
Variable de segmentation :
Situ (Situation familiale)
Branches :
Modalites disponibles :
branche 1
CELB (célibataire)
VEUF (veuf)
branche 2
MARI (marié)
DIVO (divorcé)
<<
>>
Ok
Annuler
Aide

Manipuler l'arbre par niveaux

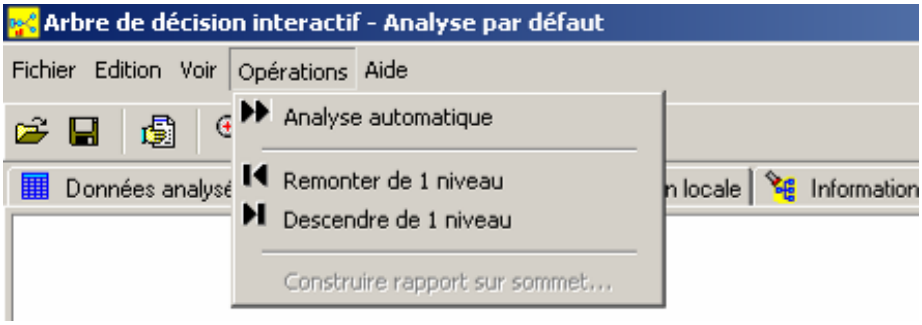
L'utilisateur peut explorer différentes solutions à partir de chaque sommet de l'arbre, il lui est également possible de manipuler interactivement l'arbre de décision en travaillant par niveau.

Dans ce cadre, la procédure effectue l'opération demandée sur toutes feuilles situées sur le niveau le plus profond de l'arbre.

Deux types d'opérations sont disponibles :

- Remonter de 1 niveau : la procédure élague tous les sommets situés sur l'avant dernier niveau de l'arbre
- Descendre de 1niveau : la procédure cherche, pour chaque feuille située sur le dernier niveau de l'arbre, la segmentation la plus efficace. Attention, les règles d'arrêt d'expansion de l'arbre sont désactivées ici.

Comment procéder ?
Selon l'opération demandée, il faut cliquer le menu « *Opérations - Remonter de 1 niveau* ou *Opérations - Descendre de 1 niveau* »



Poursuivre une analyse automatique

A tous les stades de la manipulation interactive de l'arbre, l'utilisateur a la possibilité de demander à la procédure de poursuivre la construction automatique du modèle en utilisant les options spécifiées lors du paramétrage de la méthode. Ainsi, l'utilisateur peut, par exemple, choisir la segmentation qui lui semble la plus intéressante sur le premier sommet, la racine, puis demander au logiciel de poursuivre en cherchant le meilleur arbre à partir de ce premier découpage.

Toutes les options sélectionnées, notamment celles relatives aux règles d'arrêt de l'expansion, sont actives dans ce cadre.

Comment procéder ?

1. Veiller à ce que la fenêtre *Arbre de décision* soit sélectionnée
2. Cliquer sur le menu « *Opérations - Analyse automatique* »

Gestion des sauvegardes

A l'occasion de la première exécution, l'arbre de décision est sauvegardé, il porte le titre « Analyse par défaut ». C'est l'arbre qui est automatiquement visualisé à l'ouverture de la procédure IDT.
L'utilisateur a la possibilité de manipuler à sa guise l'arbre qui lui est proposé, le fruit de son travail peut alors être sauvegardé de deux manières différentes :

- soit il sauvegarde son arbre en écrasant la version précédente;
- soit il sauvegarde une nouvelle version de l'arbre pour le même problème donné, en lui attribuant un titre adéquat.

Il est possible à tout moment de charger dans IDT un arbre de décision que l'on aura sauvegardé, on distingue les différentes versions à partir du titre qui leur a été attribué par l'utilisateur.
Attention, la modification des paramètres d'analyse supprime automatiquement toutes les sauvegardes effectuées pour le problème analysé, si l'utilisateur désire une sauvegarde définitive de son travail, il lui est conseillé de passer plutôt par les rapports ou par les exportations.

Sauvegarder l'arbre courant

Lors de l'exécution de la procédure IDT, un arbre portant le titre « Analyse par défaut » est automatiquement créé, c'est l'arbre qui est visualisé lors de l'ouverture d'IDT. L'utilisateur peut modifier à sa guise cet arbre, il peut alors rendre définitive ses modifications en sauvegardant le fruit de ses manipulations.
De manière générale, il est possible de sauvegarder tout arbre que l'utilisateur est en train de manipuler.
Comment procéder ?

1. Cliquer le menu *« Fichier – Sauver »*
2. IDT écrase alors l'ancienne version de l'arbre pour la remplacer par la nouvelle

Sauvegarder une nouvelle version de l'arbre

Lors de son travail, l'utilisateur peut désirer travailler en parallèle sur différents scénarios d'analyse, correspondant à différentes intuitions. Il a la possibilité de sauvegarder une version particulière d'un arbre de décision en l'enregistrant avec un titre différent.
Comment procéder ?
L'utilisateur désire enregistrer une nouvelle version de l'arbre dans laquelle il a élagué une partie de l'arbre.

1. Procéder aux manipulations permettant d'élaguer une partie de l'arbre
2. Cliquer alors sur le menu *« Fichier - Sauver sous... »*
3. Une boîte de dialogue apparaît, invitant l'utilisateur à attribuer un nouveau titre à cette nouvelle version de l'arbre. Cette opération est obligatoire, les différentes versions sont différenciées par leur titre
4. Cliquer sur le bouton *Ok*

Attention, en cliquant sur *« Fichier – Sauver »*, l'utilisateur écrase la version qui est en mémoire.

Chargement d'un arbre de décision

A tout moment dans IDT, l'utilisateur peut charger en mémoire une version de l'arbre qu'il aura préalablement sauvegardé. A chaque version de l'arbre correspond un titre attribué par l'utilisateur.
Comment procéder ?

1. Cliquer le menu « *Fichier – Ouvrir* »
2. Une boîte liste alors les différentes versions associées au problème courant
3. Sélectionner l'analyse désirée en cliquant sur son titre
4. Et confirmer en cliquant sur le bouton *OK*

Exportation des règles

Tout arbre de décision peut être traduit en base de règles sans pertes d'informations. Une règle correspond au chemin menant de la racine à une feuille donnée, la conclusion associée à la règle correspond à la conclusion associée à la feuille.

La règle est donc de la forme : **Si** condition **Alors** conclusion

IDT produit la liste de règles associée à un arbre au format HTML, SPAD ou SQL.

Comment procéder ?

1. Cliquer sur le menu « *Fichier – Exporter règles au format … »*
2. Une boîte de dialogue apparaît, permettant d'introduire le nom du fichier généré et sa localisation
3. Cliquer alors sur le bouton *Enregistrer*