SPAD 7.0

Data management . Analyse des Données . Data Mining

GUIDE
DE
L'UTILISATEUR

Interface – Projets – Diagrammes
Imports / Exports – Data Management – Industrialisation

Coheris Spad

La défense - 13 bis, rue de l'abreuvoir - 92400 Courbevoie - France
Tél : +33 1 41 88 02 02 - Fax : +33 1 41 88 80 98
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 de l'utilisateur

*Le **logiciel** décrit dans le manuel est diffusé dans le cadre d'un accord de licence d'utilisation et de non divulgation, et ne peut être utilisé ou copié qu'en conformité avec les stipulations de l'accord. Toute copie du programme sur CD-Rom, disque ou autre support à des fins autres que l'usage personnel du programme par le licencié est interdite par la loi. Les informations figurant dans ce **manuel** sont sujettes à révision sans préavis et ne présentent aucun engagement de la part de SPAD.*

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

La défense - 13 bis, rue de l'abreuvoir - 92400 Courbevoie - France
Tél : +33 1 41 88 02 02 - Fax : +33 1 41 88 80 98
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, l'utilisateur 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.

Dans le manuel « Guide du Data Miner », le lecteur découvrira les principales méthodes d'analyse de données et de data mining au travers d'exemples illustrés et commentés.

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

Chapitre

1

Panorama général de SPAD

La notion de projet

La version 7 du logiciel SPAD repose sur la notion de projet.
Un projet regroupe l'ensemble des éléments relatifs aux traitements réalisés, aux données et aux modèles que l'utilisateur a souhaité archiver dans le projet.

L'ouverture d'un projet ou la création d'un nouveau projet se fait à partir de la fenêtre suivante, qui apparaît à chaque ouverture de SPAD :

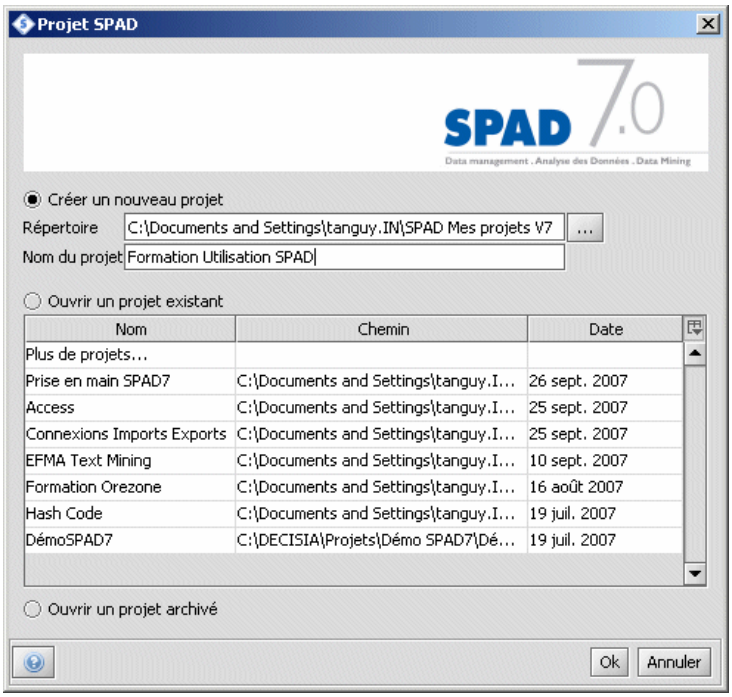


Figure 1-1 : Ouverture d'un projet SPAD

Créer un nouveau projet.
Spécifiez le répertoire dans lequel sauvegarder le projet, donnez-lui un nom et cliquez sur OK pour valider.
A la validation, le projet est créé sous la forme d'un sous-répertoire (portant le nom du projet) au sein du répertoire spécifié.

Le répertoire proposé par défaut peut être modifié dans le menu « Outils – Préférences Générales »

Ouvrir un projet existant
Le tableau liste l'ensemble des projets récemment créés et consultés. Vous pouvez trier les projets par Nom, Chemin et par Date en cliquant sur l'entête de colonne souhaitée.
Pour ouvrir un projet ne figurant pas dans la liste, cliquez sur « Plus de projets » afin d'accéder à l'explorateur.

Ouvrir un projet archivé
Un projet archivé est un projet zippé qui permet de réduire l'espace disque utilisé. Pour ouvrir un projet archivé, cliquez sur le bouton option correspondant afin d'accéder à l'explorateur. Sélectionnez le projet archive et spécifiez ensuite le répertoire dans lequel le projet sera désarchivé.

L'interface générale de SPAD

Par défaut, l'interface générale de SPAD est organisée autour de la notion de projet et se décompose en 4 fenêtres principales :

1. La fenêtre « Projet »
2. La fenêtre « Diagrammes »
3. La fenêtre « Méthodes »
4. La fenêtre « Exécutions »

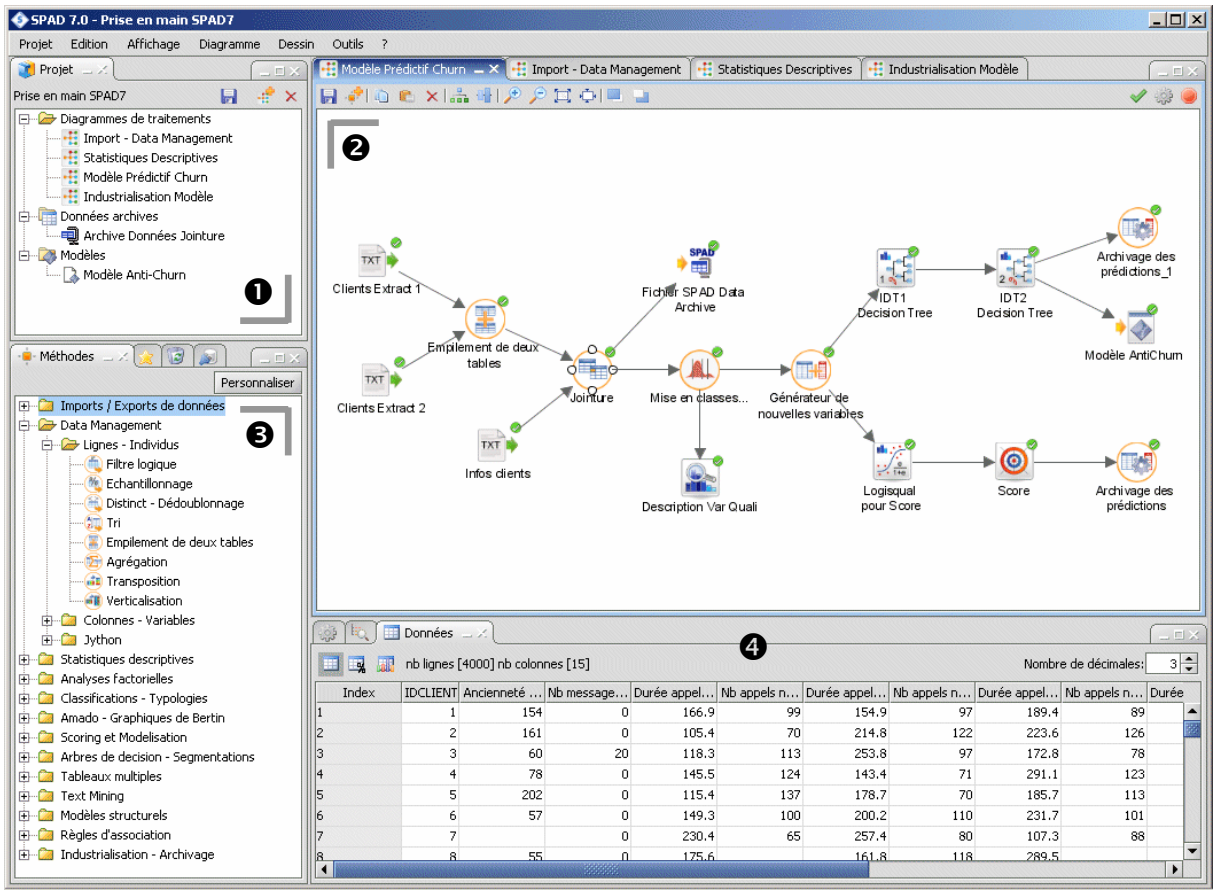


Figure 1-2 : Interface générale du logiciel SPAD

La fenêtre « Projet »

La fenêtre ou la vue « Projet » affiche l'ensemble des éléments relatifs au projet courant et en permet la gestion. C'est le gestionnaire de projet.

Les différents éléments sont regroupés par catégorie et sont organisés dans une arborescence, dans laquelle il est possible de créer des sous-répertoires et de déplacer les différents éléments respectivement par clic droit et glisser/déposer.

Les éléments d'un projet sont organisés en trois répertoires :

- **Diagrammes de traitements**
  Contient l'ensemble des diagrammes (de traitements) du projet. Chaque diagramme représente un traitement particulier des données.

  Par défaut, chaque nouveau projet est créé avec un diagramme nommé « Diagramme ».

  Les diagrammes sont indépendants les uns des autres. Ils peuvent être renommés, dupliqués, supprimés et déplacés dans l'arborescence du répertoire « Diagramme de traitements » par clic droit.
- **Données archives**
  Contient l'ensemble des tables de données que l'utilisateur a archivées directement au sein du projet. Ce répertoire n'apparaît pas si aucune archive n'a été créée dans le projet.
- **Modèles**
  Contient l'ensemble des modèles prédictifs que l'utilisateur a archivés au sein du projet. Ce répertoire n'apparaît pas si aucun modèle n'a été archivé.

Pour **ouvrir un diagramme**, double-cliquez sur le nom du diagramme choisi dans la vue « Projet ». Le diagramme apparaît alors dans la fenêtre dédiée à l'affichage des diagrammes.

La fenêtre « Diagrammes ».

La fenêtre « Diagrammes » est la fenêtre la plus étendue de l'interface de SPAD. Elle constitue l'espace privilégié de l'utilisateur. C'est dans le diagramme que l'utilisateur spécifie les données à importer et les différents types de traitements qu'il souhaite effectuer sur ses données.

Cette fenêtre peut contenir plusieurs diagrammes organisés dans un système d'onglets.

Définition : On entend par diagramme la représentation visuelle d'un ensemble de traitements symbolisés par un graphe orienté dans lequel transitent les données. Les nœuds ou méthodes du diagramme sont symbolisées par des icônes et représentent un traitement effectué sur les données.

- La construction d'un diagramme se fait à l'aide de la souris et des vues « Méthodes », « Favoris » et « Connexions » présentées ci-après.
- L'utilisateur insère successivement par glisser-déposer ou par double-clic des éléments de ces trois vues dans le diagramme.
- A chaque nouvelle insertion, il relie l'une des méthodes préalables à la méthode insérée afin de transmettre les données de la méthode préalable à la méthode insérée. Cette spécification se fait soit manuellement soit automatiquement en déposant la nouvelle méthode sur sa méthode « source ».
- Après insertion d'une méthode dans un chaînage, l'utilisateur double clique sur l'icône pour effectuer le paramétrage. Après validation du paramétrage, la procédure peut s'exécuter soit automatiquement soit en mode manuel selon les options définies dans les préférences générales.

Note. Certaines méthodes comme la jointure ou la juxtaposition requièrent deux méthodes en entrée.

La fenêtre « Méthodes »

La vue « Méthodes » ne dépend pas du projet et donne accès aux différentes méthodes de SPAD (Statistiques et Data Management).
Le bouton « Personnaliser » permet à l'utilisateur de faire le choix des rubriques et des méthodes à afficher ou masquer pour plus de lisibilité.

Juxtaposées à la fenêtre méthodes, on trouve les vues « Favoris », « Historique » et « Connexions ».

La vue « Favoris »

Les favoris sont des enchainements prédéfinis de méthodes, couramment utilisés, et qui évitent de devoir redéfinir le chainage entre les différentes méthodes. L'utilisateur peut enregistrer ses propres favoris à partir de cette vue.

La vue « Historique »

La vue Historique affiche l'ensemble des éléments supprimés et exécutés au sein du projet courant. Chaque élément supprimé peut être restauré au sein d'un diagramme. Il est également possible de revenir au paramétrage précédent d'une méthode à partir de cette vue.

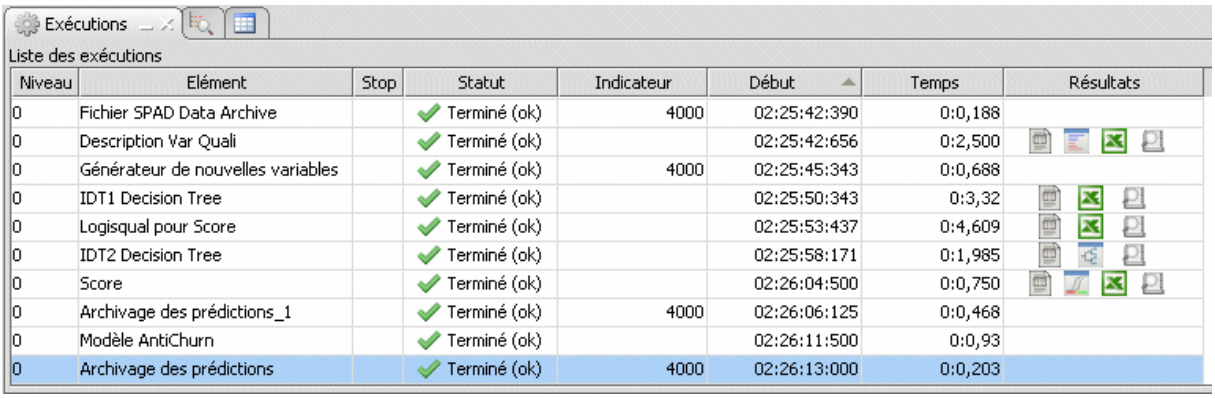
La vue « Connexions »

Une connexion est un lien vers un répertoire contenant des données ou vers une base de données. Les connexions permettent d'importer et d'exporter des données. Toute connexion peut être définie comme permanente (Connexion par défaut) afin qu'elle soit disponible dans chaque nouveau projet.
Pour plus d'information, cf. page 50.

La fenêtre « Exécutions »

La fenêtre « Exécutions » rassemble l'ensemble des résultats relatifs à l'exécution des différents diagrammes et notamment la vue « Données » qui permet pour chaque méthode de data management de visualiser de façon interactive la table de données produite en fin d'exécution.
Cette fonctionnalité est particulièrement appréciable pour le contrôle de cohérence des données et la vérification des traitements produits.

La vue « Exécutions » liste l'ensemble des méthodes du diagramme courant avec leurs statuts, le temps d'exécution, et les liens vers les résultats éventuellement produit par la méthode.
Pour accéder à l'un des résultats d'une méthode, double-cliquez sur l'icône choisie.


Exécutions

Liste des exécutions

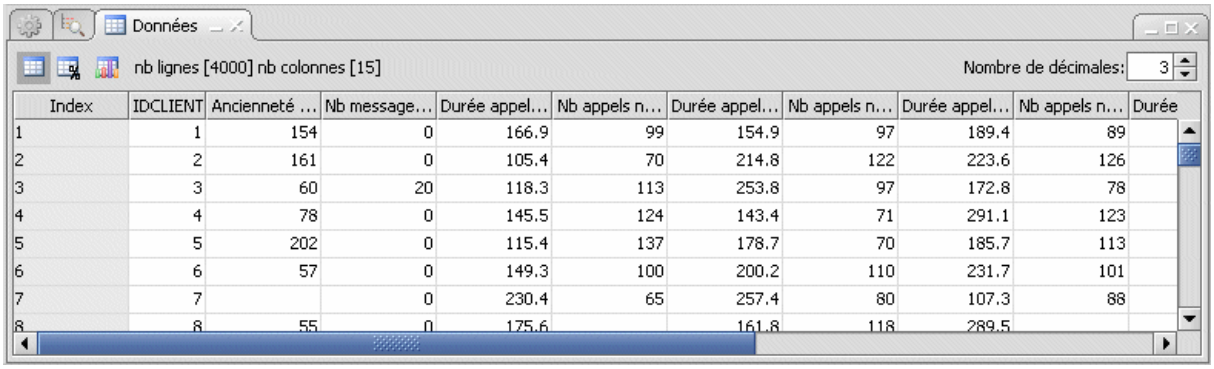
Niveau	Elément	Stop	Statut	Indicateur	Début	Temps	Résultats
0	Fichier SPAD Data Archive		Terminé (ok)	4000	02:25:42:390	0:0,188	
0	Description Var Quali		Terminé (ok)		02:25:42:656	0:2,500	
0	Générateur de nouvelles variables		Terminé (ok)	4000	02:25:45:343	0:0,688	
0	IDT1 Decision Tree		Terminé (ok)		02:25:50:343	0:3,32	
0	Logisqual pour Score		Terminé (ok)		02:25:53:437	0:4,609	
0	IDT2 Decision Tree		Terminé (ok)		02:25:58:171	0:1,985	
0	Score		Terminé (ok)		02:26:04:500	0:0,750	
0	Archivage des prédictions_1		Terminé (ok)	4000	02:26:06:125	0:0,468	
0	Modèle AntiChurn		Terminé (ok)		02:26:11:500	0:0,93	
0	Archivage des prédictions		Terminé (ok)	4000	02:26:13:000	0:0,203	

Figure 1-3 : Vue Exécutions

Lors de l'exécution de méthodes d'import/export ou de data management, la colonne Indicateur affiche toutes les 2 secondes (par défaut) le nombre de lignes traitées afin de suivre l'évolution du traitement.
Pendant l'exécution, un rond rouge apparaît dans la colonne « Stop » pour la méthode en cours d'exécution. Cliquez sur ce rond pour interrompre l'exécution.

La vue « Données »

La vue « Données » est directement liée à la méthode sélectionnée du diagramme courant. Elle affiche la table de données directement produite par la méthode de data management, ou la table de données disponible s'il s'agit d'une méthode statistique.
Elle permet ainsi un contrôle et une validation immédiate des données produites.



Index	IDCLIENT	Ancienneté ...	Nb message...	Durée appel...	Nb appels n...	Durée appel...	Nb appels n...	Durée appel...	Nb appels n...	Durée
1	1	154	0	166.9	99	154.9	97	189.4	89	
2	2	161	0	105.4	70	214.8	122	223.6	126	
3	3	60	20	118.3	113	253.8	97	172.8	78	
4	4	78	0	145.5	124	143.4	71	291.1	123	
5	5	202	0	115.4	137	178.7	70	185.7	113	
6	6	57	0	149.3	100	200.2	110	231.7	101	
7	7		0	230.4	65	257.4	80	107.3	88	
8	8	55	0	175.6		161.8	118	289.5		

Figure 1-4 : Vue Données

Le bouton affiche les caractéristiques de l'ensemble des variables de la table de données (type de stockage, rôle statistique, nombre de valeurs distinctes, nombre de valeurs manquantes, min, max, moyenne, écart-type…).

Le bouton affiche l'ensemble des tris-à-plat des variables nominales de la table sélectionnée.

La vue « Log »

La vue « Log » n'est alimentée qu'en cas d'erreur d'exécution d'une méthode ou d'avertissement.

Personnaliser l'interface de SPAD

Vous pouvez redimensionner et fermer chaque fenêtre de l'application. Le positionnement des différentes fenêtres peut être modifié par glisser/déposer (en mode d'affichage « Fenêtres libres »).

Le menu « Affichage » permet de sauvegarder l'interface personnalisée en tant que vue utilisateur. Il permet également d'afficher ou de masquer les différentes fenêtres du logiciel.

La barre de menus

Menu Projet

Nouveau projet. Créer un nouveau projet.

Ouvrir un projet. Ouvrir un projet existant à partir de la fenêtre suivante :

Sélectionnez un des projets dans la liste ou cliquez sur « Plus de projets… » pour accéder à l'explorateur.

Dans ce cas, spécifiez le projet à ouvrir en sélectionnant le fichier d'extension « *.project », figurant dans le répertoire portant le nom du projet à ouvrir.

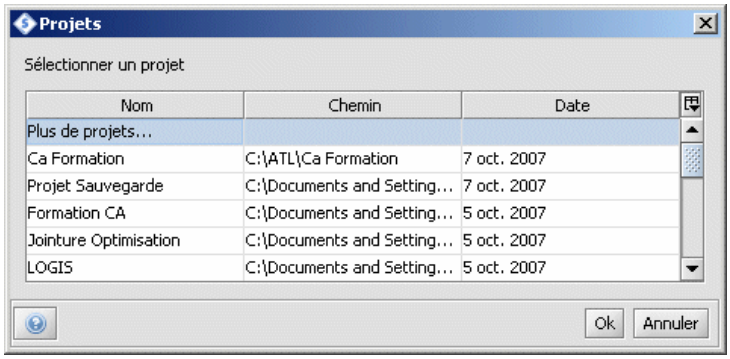


Figure 1-5 : Ouvrir un projet

Ouvrir un projet archivé. Ouvrir un projet précédemment archivé. Spécifiez le **projet archivé** en accédant à l'explorateur en cliquant sur le bouton « … » et spécifiez le **répertoire cible** dans lequel le projet archivé sera désarchivé.

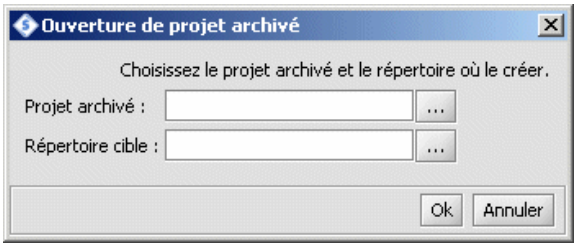


Figure 1-6 : Ouvrir un projet archivé

Enregistrer sous… Enregistrer une copie du projet ouvert. Spécifier le type de copie dans la fenêtre suivante :

- **Archive du projet** : une archive est fichier zippé d'extension « *.spad » du répertoire contenant l'ensemble des éléments du projet.
- **Copie du projet** : une copie est une duplication du projet courant. Spécifiez le répertoire du projet dupliqué.

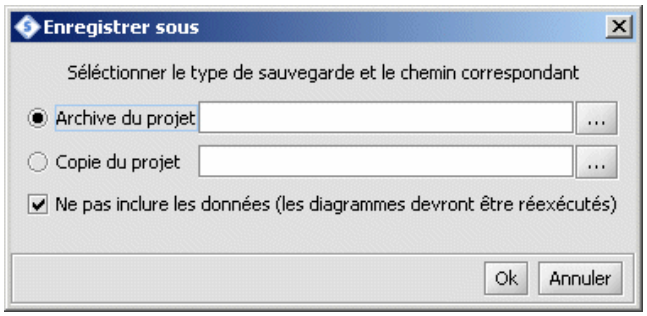


Figure 1-7 : Enregistrer sous…

- **Ne pas inclure les données.** Cette option, cochée par défaut, permet de ne sauvegarder que les éléments figurant dans le gestionnaire de projet : les diagrammes et leur paramétrage, les données et modèles archivés. La base de données interne et l'ensemble des fichiers relatifs aux méthodes statistiques (données et résultats) ne sont pas sauvegardés. Ce qui peut constituer un gain d'espace disque important.

Fermer le projet. Fermer le projet courant.

Fermer le projet sans sauvegarder. Fermer le projet sans sauvegarder les dernières modifications, effectuées après la dernière exécution.

Nettoyer le projet. Supprimer la base de données interne et l'ensemble des fichiers relatifs aux méthodes statistiques (données et résultats). L'ensemble des éléments du gestionnaire de projet (diagrammes et leur paramétrage, données et modèles archivés) sont conservés.

Supprimer des projets. Pour supprimer des projets SPAD, il est conseillé d'utiliser cette commande. Sélectionnez les fichiers à supprimer et cliquez sur OK.

Projets récents. Ouvrir un projet à partir d'une liste de projets récemment ouverts.

Quitter. Ferme l'application SPAD.

Menu Edition

Le menu Edition ne contrôle que les éléments du diagramme courant.

Copier (Ctrl-C)**.** Ajouter les méthodes sélectionnées du diagramme courant dans le presse-papiers.

Coller (Ctrl-V)**.** Coller le contenu du presse-papiers dans le diagramme courant. Spécifiez l'endroit où coller le contenu du presse-papiers à l'aide du curseur de la souris.

Supprimer (Suppr). Supprimer l'ensemble des méthodes sélectionnées.

Sélectionner tout (Ctrl-A)**.** Sélectionner tous les éléments du diagramme courant.

Menu Affichage

Fenêtres fixes. Figer les fenêtres de l'application.

Fenêtres normales. Les fenêtres de l'application peuvent être redimensionnées mais pas déplacées.

Fenêtres libres. Les fenêtres de l'application peuvent être redimensionnées et déplacées.

Vue par défaut. Revenir à la vue par défaut de SPAD.

Vue utilisateur. Revenir à la vue utilisateur.

Enregistrer en tant que vue utilisateur. Enregistrer l'interface courante comme vue utilisateur. La vue utilisateur sera utilisée par défaut.

Projet. Afficher/masquer la fenêtre Projet.
Vue aérienne. Afficher/masquer la fenêtre Vue aérienne.
Méthodes. Afficher/masquer la fenêtre Méthodes.
Exécutions. Afficher/masquer la fenêtre Exécutions.
Log. Afficher/masquer la fenêtre Log.
Favoris. Afficher/masquer la fenêtre Favoris.
Historique. Afficher/masquer la fenêtre Historique.
Description. Afficher/masquer la fenêtre Description.
Données. Afficher/masquer la fenêtre Données.
Connexions. Afficher/masquer la fenêtre Connexions.
Console. Afficher/masquer la fenêtre Console.

Menu Diagramme

Exécuter (F5)**.** Exécuter les méthodes sélectionnées du diagramme courant.

Interrompre l'exécution (F7)**.** Interrompre l'exécution des méthodes en cours d'exécution dans le diagramme courant.

Nouveau diagramme. Créer un nouveau diagramme. Le nouveau diagramme apparaît dans le gestionnaire de projet.

Supprimer diagramme. Supprimer le diagramme courant.

Ajouter une méthode (Insert)**.** Ajouter une méthode à partir de la boite de dialogue Rubriques/Méthodes. Si une méthode est sélectionnée dans le diagramme courant, seules les méthodes admises en aval de la méthode sélectionnée seront sélectionnables.

Menu Dessin

Zoom. Agrandir ou réduire l'échelle de la fenêtre Diagramme.

- **Zoom +** (+): Agrandit la taille des méthodes dans le diagramme.
- **Zoom –** (-): Réduit la taille des méthodes dans le diagramme.
- **Zoom automatique** (*) : Ajuste le zoom pour voir toutes les méthodes du diagramme.
- **Zoom par défaut** (=) : Rétablit le zoom par défaut.

Organiser. Réorganiser les méthodes sélectionnées sous forme d'arbres. Deux types de réorganisation sont proposés : Arbres et Arbres par clan. Si aucune méthode n'est sélectionnée, la réorganisation « Arbres par clan » réorganise le diagramme complet.

- **Arbre vers le bas** (Alt-S) : réorganise les méthodes sélectionnées en arbre avec les feuilles en bas.
- **Arbre vers le haut** (Alt-D) : réorganise les méthodes sélectionnées en arbre avec les feuilles en haut.
- **Arbre vers la gauche** (Alt-Q) : réorganise les méthodes sélectionnées en arbre horizontal avec les feuilles à gauche.
- **Arbre vers la droite** (Alt-F) : réorganise les méthodes sélectionnées en arbre horizontal avec les feuilles à droite.
- **Disposition en grille** (Alt-G) : réorganise le diagramme en positionnant les méthodes en grille.

Aligner. Aligner les méthodes sélectionnées vers le haut, au milieu, vers le bas, à gauche, au centre, à droite.

Mettre au premier plan. Positionne la ou les méthodes sélectionnées au premier plan.

Mettre en arrière-plan. Positionne la ou les méthodes sélectionnées à l'arrière-plan.

Menu Outils

Préférences. Ouvre le gestionnaire des préférences de SPAD (cf. page 18).

Votre licence. Ouvre le gestionnaire de licences (cf. page 21).

Librairies de formats. Ouvre le gestionnaire des librairies de formats. (cf. Gestionnaire des librairies de formats, page 91)

Menu ?

Sommaire de l'aide. Ouvre l'aide en ligne de SPAD.

A propos de SPAD. Affiche le copyright et le numéro de version de SPAD.

Guide de l'utilisateur. Ouvre le manuel « Guide de l'utilisateur » au format pdf.

Guide du Data Miner. Ouvre le manuel « Guide du Data Miner » au format pdf.

Architecture Projet. Affiche les correspondances entre éléments du projet et noms physiques.

Raccourcis clavier

La plupart des opérations réalisées dans la vue « Diagrammes » ont des raccourcis clavier.
Voici la liste des raccourcis :

Raccourci clavier	**Fonction**
Ctrl-A	Sélectionner tous les éléments du diagramme courant.
Ctrl-C	Copier
Ctrl-V	Coller
Insert	Insérer Méthode dans le diagramme courant
+ pavé numérique	Zoom avant
- pavé numérique	Zoom arrière
*****	Zoom automatique
=	Zoom par défaut
F2	Renommer la méthode sélectionnée
F4	Paramétrer la méthode sélectionnée
F5	Exécuter les méthodes sélectionnées
F7	Interrompre l'exécution
Suppr	Supprimer
Alt-X	Arbre vers le bas
Alt-C	Arbre vers le haut
Alt-W	Arbre vers la gauche
Alt-V	Arbre vers la droite
Alt-S	Arbre par clan vers le bas
Alt-D	Arbre par clan vers le haut
Alt-Q	Arbre par clan vers la gauche
Alt-F	Arbre par clan vers la droite

Quelques conseils

Importation de données

Utilisez le typage personnalisé.
Utilisez systématiquement le typage personnalisé lors des imports de données pour un typage affiné des variables importées. Bien paramétré, le typage personnalisé évite les erreurs de conversion de typage comme par exemple la création de variables nominales comportant des milliers de modalités…

N'importez pas les variables dont vous ne vous servirez pas ou mettez à « Exclue » les variables sans intérêt pour l'analyse statistique.
Pour réduire l'espace disque utilisé et pour de meilleures performances, n'importez pas les variables sans intérêt et/ou assignez le rôle « Exclue » à celles qui n'ont pas d'intérêt pour l'analyse statistique ou qui nécessitent un recodage préalable. Les variables exclues seront disponibles pour le data management.

Connexions ou imports directs.
Préférez les connexions aux imports directs si vous souhaitez simplifier l'échange de projets SPAD entre collaborateurs ou pouvoir passer rapidement d'un environnement de test à un environnement de production. Dans ce cas là, il suffit uniquement de modifier les propriétés des connexions utilisées dans le projet.

Connexions natives aux bases de données plutôt que liaisons ODBC.
Pour de meilleures performances, préférez les connexions natives vers les bases de données (lorsqu'elles sont disponibles dans SPAD) aux connexions ODBC.

Testez vos imports.
Dans le cas d'imports de données complexes, testez vos imports sur des échantillons à l'aide des limitations proposées (Tirage aléatoire, nombre de lignes limité).

Archiver vos données.
N'hésitez pas à créer une archive compressée des données que vous importez dans SPAD (c.f. Méthode Export Sda, page 62) surtout lorsque ces données évoluent dans le temps.

Projets

Archivez vos projets.
N'hésitez pas à créer une archive compressée de votre projet pour en faciliter les échanges et le conserver dans un format compact.

Organisez vos projets et nommez vos diagrammes.
N'hésitez pas à créer des dossiers et à utiliser des noms de dossiers et de diagrammes explicites pour une meilleure gestion et un meilleur suivi dans le temps de votre projet. Renommez également les méthodes des diagrammes pour comprendre visuellement l'ensemble des étapes d'un traitement. Utilisez la vue « Description » pour détailler chaque méthode si nécessaire.

Méthodes

Personnalisez la vue « Méthodes ».
N'hésitez pas à simplifier la vue « Méthodes » pour une meilleure lisibilité.

Utilisez les favoris.
Utilisez les favoris pour les typologies, les modèles de scores ou arbres de décision. Créez les vôtres.

Diagrammes

Insérer méthode plutôt qu'Ajouter méthode.
Privilégiez l'insertion de méthode qui permet (par glisser/déposer) de lier immédiatement la méthode source à la méthode insérée.

Dupliquer ou Copier/Coller.
Pour dupliquer une méthode ou un ensemble d'éléments :

1. Faites votre sélection à l'aide de la souris (usage des touches Ctrl et Shift standard pour la sélection multiple)
2. Appuyez sur la touche Ctrl et déplacez la sélection à l'aide la souris en maintenant la touche Ctrl enfoncée. Le curseur  apparaît.
3. Relâchez le clic de la souris à l'endroit où vous souhaitez dupliquer la sélection.

Utilisez les « raccourcis ».
Double-cliquez sur la méthode pour la paramétrer.
Utilisez la touche F5 pour exécuter.
Double-cliquez sur le nom d'une méthode pour la renommer.

Contrôlez toutes les étapes d'import et de data management avec la vue « Données ».
Pour toute étape d'import ou de transformation des données, consultez la vue « Données » pour visualiser les données créées, contrôler et valider.

Le gestionnaire des préférences

Les préférences permettent de définir les préférences de fonctionnement, d'affichage et d'enregistrement de l'application. Elles sont accessibles à partir du menu « Outils/Préférences ».

Préférences générales

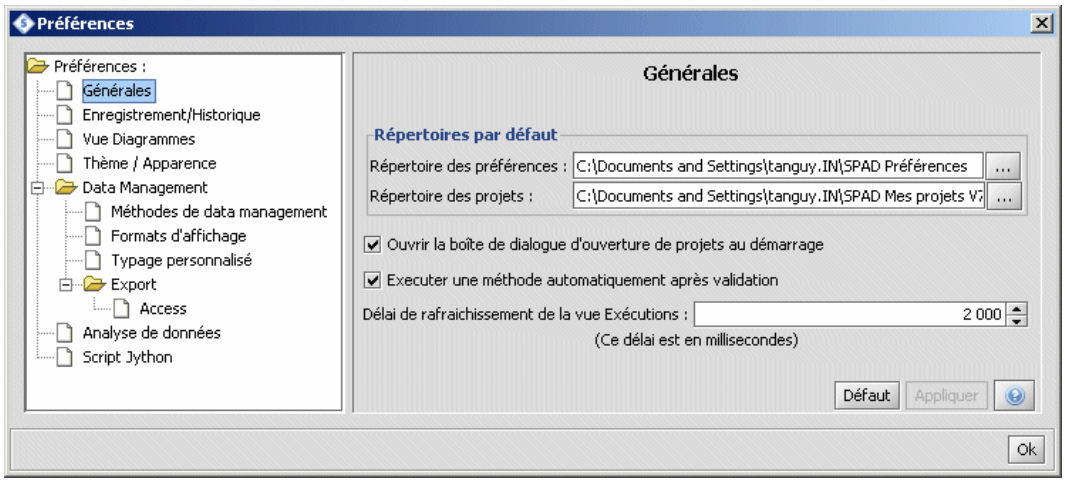


Figure 1-8 : Préférences générales

Répertoires par défaut.

- **Répertoire des préférences.** Répertoire dans lequel seront enregistrées l'ensemble des préférences de SPAD. Ce répertoire contient les préférences spécifiées par l'utilisateur, les connexions par défaut et les favoris.
- **Répertoire des projets.** Répertoire proposé par défaut pour la création d'un nouveau projet.

Ouvrir la boite de dialogue d'ouverture de projets au démarrage. Décocher cette case si vous ne souhaitez pas faire apparaître la fenêtre d'ouverture de projets au démarrage de SPAD.

Exécuter une méthode automatiquement après validation. Décocher cette case si vous souhaitez pouvoir paramétrer une méthode et l'exécuter en deux temps.

Délai de rafraichissement de la vue Exécutions. Spécifiez le délai de rafraichissement en millisecondes des informations de la vue Exécutions. Par défaut, le délai est de 2 secondes.

Enregistrement / Historique

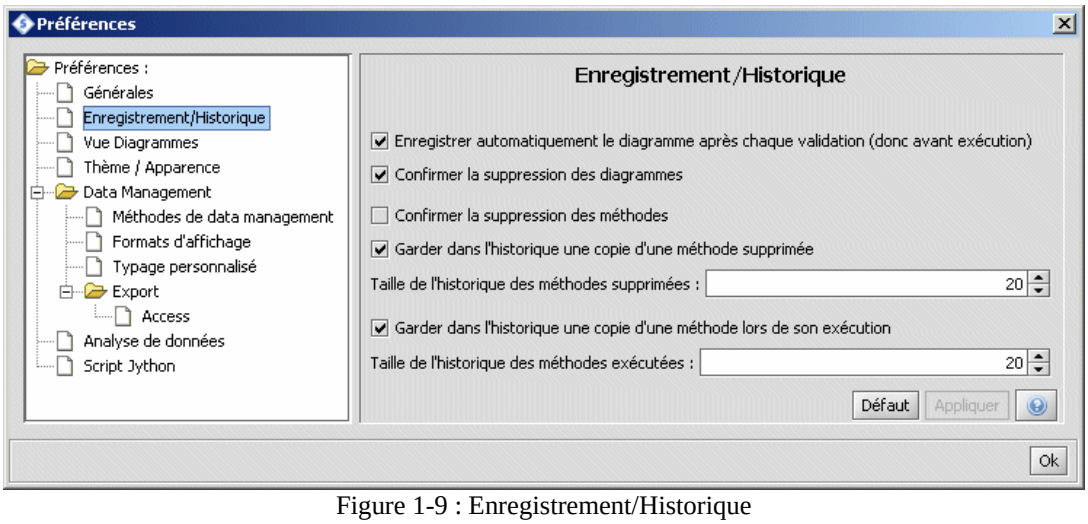


Figure 1-9 : Enregistrement/Historique

Enregistrer automatiquement le diagramme après chaque validation (donc avant exécution). Option par défaut.

Confirmer la suppression des diagrammes. Décochez si vous ne souhaitez pas afficher de demande de confirmation de suppression du diagramme. La suppression d'un diagramme est irréversible.

Confirmer la suppression des méthodes. Décochez si vous ne souhaitez pas afficher de demande de confirmation de suppression.

Garder dans l'historique une copie d'une méthode supprimée. Conserver dans la vue Historique une trace des méthodes supprimées pour restauration éventuelle.

- **Taille de l'historique des méthodes supprimées.** Taille allouée à la sauvegarde des méthodes supprimées.

Garder dans l'historique une copie d'une méthode lors de sons exécution. Conserver dans la vue Historique une trace des paramétrages successifs des méthodes exécutées.

- **Taille de l'historique des méthodes exécutées.** Taille allouée en nombre de méthodes à la sauvegarde des paramétrages des méthodes exécutées.

Vue Diagrammes

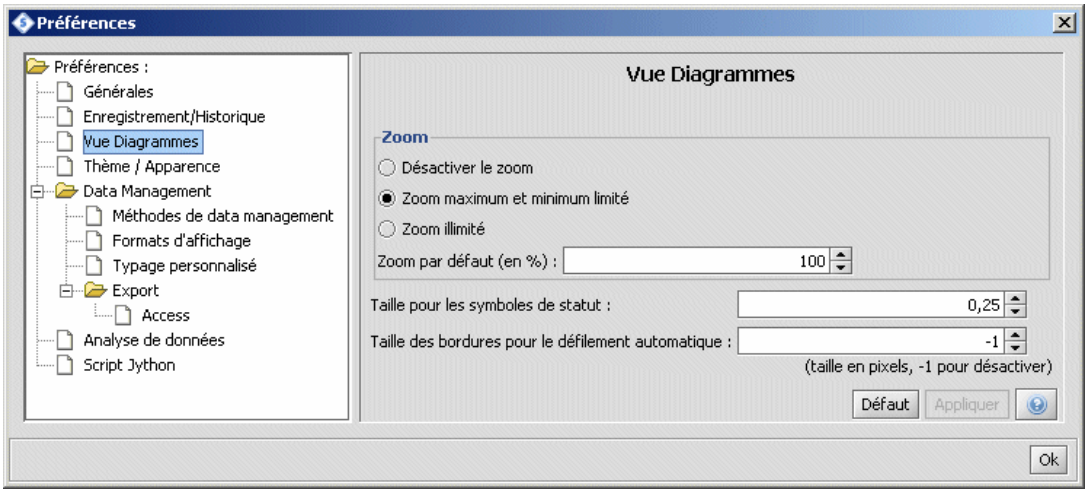


Figure 1-10 : Vue Diagrammes

Zoom. Permet de gérer le zoom dans la vue Diagrammes.

- **Désactiver le zoom.** Cochez la case pour désactiver le zoom.
- **Zoom maximum et minimum limité.** Limite les possibilités de zoom.
- **Zoom illimité.** Zoom sans limite.
- **Zoom par défaut** (en %). Niveau de zoom par défaut pour l'affichage des diagrammes.

Taille pour les symboles de statut. Permet d'agrandir ou de diminuer la taille des symboles de statut des méthodes. (Valeurs comprises entre 0,1 et 1).

Taille des bordures pour le défilement automatique. Taille en nombre de pixels en dessous de laquelle le contenu du diagramme défile lorsqu'on l'utilisateur rapproche un élément d'une des bordures du diagramme.

Afficher la barre d'outils. Décochez la case si vous ne souhaitez pas faire apparaître la barre d'outils dans la vue Diagrammes.

Thème / Apparence

Les différentes options de cette rubrique vous permettent de modifier le thème et l'apparence de SPAD.

Data Management

Méthodes de data management.

- **Taille du buffer interne.** Ce paramètre gère la taille du flot de données, contenu en mémoire, utilisé pour les imports/exports. Cette taille est exprimée en nombre de lignes. L'augmentation de la taille du buffer dépend des capacités de la machine en mémoire. Un mauvais réglage du flot de données peut entraîner l'effet inverse de celui escompté.
- **Stocker le type <Autre> comme une chaîne**. Les variables dont le type de stockage est « Autre » dans les métadonnées sont importées comme des chaines de caractères. Décochez si vous souhaitez conserver le format « Autre » qui permet notamment de stocker tous types d'objets.
- **Avertissement si aucune variable n'est disponible pour les méthodes statistiques**. Cette option affiche un avertissement lorsqu'une méthode d'import ou de data management ne propose aucune variable nominale ou continue. Cet avertissement indique qu'aucune méthode statistique ne pourra être alimentée par cette méthode.

Formats d'affichage.

- **Renommer automatiquement les méthodes d'import.** Cette option permet d'assigner automatiquement à toute méthode d'import le nom du fichier importé.
- **Format des dates.** Format d'affichage des dates dans la vue Données. Choisissez le format d'affichage dans la liste déroulante.
- **Format des nombres – Nombre de décimales.** Nombre de décimales affichées pour les variables de type de stockage « Réel » dans la vue Données.

Typage personnalisé. Voir page 35.

Export.

- **Export Access.** Base de données Access (vide) utilisée comme modèle pour définir le format des bases Access générées par SPAD lors de l'export. Par défaut, les bases Access générées par SPAD correspondent à la version Microsoft Office Access 2003. Cliquez sur le bouton « … » pour sélectionner le modèle de base Access que vous souhaitez générer parmi ceux proposés.

Analyse de données

Répertoire racine de SPAD. Répertoire racine de l'ensemble des répertoires et fichiers relatifs aux méthodes statistiques. Il est conseillé de ne pas modifier le choix par défaut.

Fichier *.xla de XLSTAT. Emplacement du fichier XLSTAT.xla nécessaire au lancement de l'application XLSTAT depuis SPAD. Si vous disposez déjà d'une version XLSTAT Pro, spécifiez l'emplacement de ce fichier pour permettre à SPAD de lancer XLSTAT.

Afficher un avertissement lorsque le paramétrage des méthodes statistiques ne correspond pas aux données sources. Affiche un avertissement lorsque les données sur lesquelles reposent les méthodes statistiques ne correspondent plus au paramétrage courant.

Afficher un avertissement lorsque les noms des variables utilisées dans le paramétrage des méthodes statistiques diffèrent de ceux définis dans les données sources. Cet avertissement permet d'alerter l'utilisateur sur des conflits de paramétrage entre les variables réelles et celles définies dans le paramétrage des méthodes statistiques.

Script Jython

Répertoire d'installation de Jython. Emplacement du répertoire de l'application Jython.

Répertoire des scripts Jython.

Répertoires Jython à importer. Permet de spécifier de nouveaux répertoires contenant des fichiers Jython à importer.

Le gestionnaire de licence

Accessible depuis le menu « Outils/Votre licence », le gestionnaire de licence vous permet d'activer votre licence d'utilisation SPAD et de consulter l'ensemble des modules et méthodes auxquels vous avez droit.
Si vous n'avez pas activé votre licence ou si celle-ci est expirée, vous ne pouvez pas utiliser SPAD

Pour recevoir de la part de SPAD le fichier d'activation de votre licence, renseignez les informations d'enregistrement demandées et cliquez sur le bouton « Obtenir ma licence ».

Toutes les informations demandées ici figurent sur votre « Carte d'enregistrement et bon de garantie » SPAD.
Vous devez renseigner tous les champs.
Si vous ne disposez pas de votre Code client, entrez le nom de votre organisme.

Informations d'enregistrement

Pour recevoir de la part de SPAD le fichier d'activation de votre licence, renseignez les informations d'enregistrement demandées et cliquez sur le bouton « Obtenir ma licence ».

- **Numéro de licence (obligatoire)**. Ce numéro figure sur le contrat de maintenance et d'assistance SPAD Service. Entrez ce numéro de licence.
- **Code client ou Nom de l'organisme**. Entrez le Code client qui vous a été communiqué par SPAD ou le nom l'organisme auquel vous appartenez.
- **Utilisateur.** Entrez le nom et le prénom de l'utilisateur de la licence.
- **Nom du professeur (seulement pour les étudiants)**. Pour les licences de type « Kit étudiant », entrez le nom du professeur responsable de la licence SPAD au sein de votre école.
- **Identifiant machine.** L'identifiant de votre machine est affiché automatiquement. Le bouton situé légèrement sur la droite permet de le copier dans le presse-papiers.
- **Obtenir ma licence.** Cliquez sur ce bouton pour faire apparaître une fenêtre qui vous permettra de copier ces informations dans le presse-papiers. Collez le contenu du presse-papiers dans l'email de demande d'activation de licence à envoyer à l'adresse « licence@spad.eu ».

Contenu de la licence

Le tableau liste l'ensemble des méthodes disponibles dans SPAD, triées par rubrique. Après activation de la licence, les méthodes ayant le statut ✔ sont actives et donc utilisables.

Liste des fichiers d'activation de la licence

Cette partie du gestionnaire de licence affiche l'ensemble des fichiers d'activation relatifs à votre licence SPAD. A réception du fichier d'activation (d'extension *.spadkey) envoyé directement par SPAD, enregistrez ce fichier sur votre machine et suivez la procédure suivante pour activer votre licence :

1. Cliquez sur le bouton ✚ pour spécifier le fichier d'activation de votre licence préalablement enregistré sur votre machine. Ce fichier a pour extension « *.spadkey ».
2. Validez en cliquant sur le bouton ouvrir.
3. Une nouvelle ligne apparaît dans la liste des fichiers d'activation de votre licence.
4. Vérifiez que les méthodes auxquelles vous avez contractuellement droit ont bien le statut ✔ .
5. Vous pouvez désormais utiliser SPAD normalement.

Le bouton ✕ permet de supprimer des fichiers d'activation n'étant plus actifs.

Chapitre

2

Le gestionnaire de projet – La fenêtre Projet

La fenêtre ou la vue « Projet » affiche l'ensemble des éléments relatifs au projet courant et en permet la gestion. C'est le gestionnaire de projet.

Les différents éléments sont regroupés par catégorie et sont organisés dans une arborescence, dans laquelle il est possible de créer des sous-répertoires et de déplacer les différents éléments respectivement par clic droit et glisser/déposer.

Les éléments d'un projet sont organisés en trois répertoires : les diagrammes de traitements, les données et modèles archivés.

La gestion de ces différents éléments se fait majoritairement à partir du menu contextuel (clic droit).

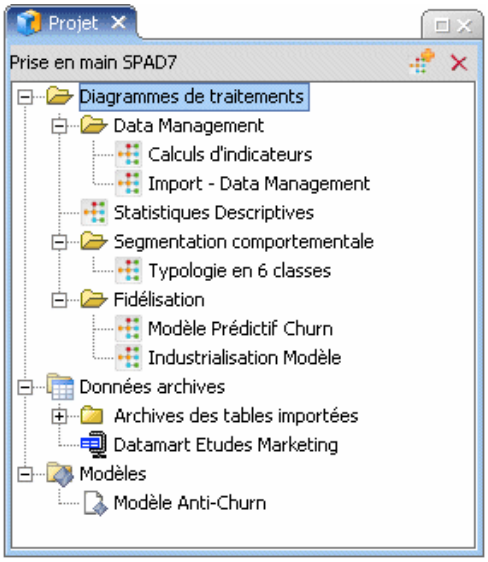


Figure 2-1 : Fenêtre Projet

Diagrammes de traitements

Ce répertoire contient l'ensemble des diagrammes (de traitements) du projet. Chaque diagramme représente un traitement particulier des données.

Par défaut, chaque nouveau projet est créé avec un diagramme nommé « Diagramme ».

Les diagrammes sont indépendants les uns des autres. Ils peuvent être renommés, dupliqués, supprimés et déplacés dans l'arborescence du répertoire « Diagramme de traitements » par clic droit.

Ouvrir un diagramme. Double-cliquez sur le nom du diagramme à ouvrir. Le diagramme apparaît instantanément dans la vue « Diagrammes ».

Activer un diagramme. Pour faire apparaître au premier plan un diagramme dans la vue « Diagrammes », cliquez sur son nom s'il est déjà ouvert ou double-cliquez sinon.

Fermer un diagramme. La fermeture d'un diagramme se fait directement dans la vue « Diagrammes » à l'aide du bouton standard de fermeture d'une fenêtre.

Menu contextuel – Clic droit

La liste des commandes accessibles par clic droit sur l'un des éléments du répertoire « Diagrammes de traitements » est la suivante :

Nouveau dossier. Pour créer un nouveau dossier au sein du répertoire Diagrammes de traitements. Le dossier créé est automatiquement nommé « Nouveau dossier ». Il est possible de le renommer (simple clic sur le nom) et de lui assigner des diagrammes de traitements par glisser/déposer.

Nouveau diagramme. Pour ajouter un nouveau diagramme dans le projet. Si le menu contextuel a été activé à partir d'un dossier, le nouveau diagramme est automatiquement inclus dans le dossier. Cette commande est équivalente à celle du bouton  .

Dupliquer. Pour dupliquer un diagramme existant. Sélectionnez préalablement le diagramme à dupliquer et cliquez sur Dupliquer à partir du clic droit.

Supprimer. Pour supprimer un ou plusieurs éléments préalablement sélectionnés. Cette commande est équivalente à celle du bouton  .

Renommer. Sélectionnez l'élément à renommer (dossier ou diagramme) et cliquez sur Renommer du menu contextuel pour éditer le nom. Tapez sur la touche Entrée pour valider le nouveau nom.

Haut. Déplace l'élément sélectionné vers le haut dans la liste des éléments du même niveau de l'arborescence. Cette manipulation peut se faire par glisser/déposer.

Bas. Déplace l'élément sélectionné vers le bas dans la liste des éléments du même niveau de l'arborescence. Cette manipulation peut se faire par glisser/déposer.

Données archives

Ce répertoire liste l'ensemble des tables de données archivées directement dans le projet par l'utilisateur à partir de la méthode Export Fichier Sda (cf. page 62). Ce répertoire n'apparaît pas si aucune archive n'a été générée dans le projet.

Pour gérer les données archives, utilisez le menu contextuel (clic droit) de la souris et référez-vous à la description des commandes du menu contextuel du répertoire « Diagrammes de traitements », ci-dessus.

Les données archives sont sauvegardées dans un format compressé qui comprend un descriptif des métadonnées.

Réutiliser des données archives dans le projet. Pour réutiliser des données archives dans un diagramme, glissez-déposez l'archive dans le diagramme courant. La méthode d'import SPAD Data Archive apparaît instantanément et pointe directement sur les données archives à importer.

Modèles

Ce répertoire liste l'ensemble des modèles prédictifs que l'utilisateur a archivés au sein du projet à partir de la méthode « Modèle prédictif – Fichier règles » (cf. page 155). Ce répertoire n'apparaît pas si aucun modèle n'a été archivé dans le projet.

Pour gérer les modèles archivés, utilisez le menu contextuel (clic droit) de la souris et référez-vous à la description des commandes du menu contextuel du répertoire « Diagrammes de traitements », ci-dessus.

Appliquer un modèle archivé. Pour appliquer un modèle à un jeu de données, glissez-déposez le modèle sur la méthode contenant les données sources. La méthode « Application d'un modèle prédictif » apparaît instantanément et pointe directement sur le modèle archivé.

Chapitre

3

Les diagrammes SPAD

Définition

L'ergonomie de SPAD est basée sur l'enchaînement visuel de méthodes de data management et de méthodes statistiques, représentées sous forme d'icônes. Ce sont les diagrammes SPAD. Ils ne nécessitent aucune programmation et montrent instantanément le cheminement logique de l'analyse.

Les liens fléchés qui relient les méthodes entre elles au sein du diagramme définissent le sens dans lequel transitent les données ainsi que l'ordre des traitements.

Chaque méthode d'import produit en fin d'exécution une table de données, disponible pour toute nouvelle méthode de data management ou méthode statistique.

Chaque méthode de data management est alimentée par une ou plusieurs tables de données dites « Source(s) » ou « Entrée(s) » et met à disposition une nouvelle table de données en sortie d'exécution. Cette nouvelle table est disponible pour toute nouvelle méthode qui lui serait connectée.

La plupart des méthodes statistiques ne modifient pas la table de données source sur laquelle elles effectuent les calculs. Certaines d'entre-elles le font néanmoins.

Nomenclature des icônes

Toute méthode d'import est symbolisée par une flèche verte, située à droite de l'icône, qui signifie l'entrée dans SPAD. Toute méthode d'import produit une table de données disponible pour les nouvelles méthodes.

Toute méthode d'export de données est symbolisée par une flèche orange située à gauche de l'icône qui signifie l'export. Les méthodes d'export ne peuvent être la source de données d'aucune autre méthode. Hormis l'export « SPAD Data Archive » au sein du projet, les méthodes d'export ne font que produire des données à l'extérieur de SPAD.

Toute méthode de data management est symbolisée par un rond au contour orange qui signifie que les données sont transformées par la méthode.

Toute méthode statistique ne transformant pas les données est symbolisée par un carré au contour noir ou grisé.

Toute méthode statistique qui génère des transformations dans les données se différencie des autres par la couleur orange du contour du carré.

Etapes de la construction d'un diagramme

La construction d'un diagramme se fait à l'aide de la souris et des vues « Méthodes », « Favoris » et « Connexions ». L'utilisateur ajoute successivement de nouvelles méthodes dans le diagramme.

A chaque nouvel ajout, il relie l'une des méthodes préalables à la méthode ajoutée afin de définir la source de données qui alimentera la méthode ajoutée. Cette spécification se fait par la création d'un arc fléché qui relie les deux méthodes. Ensuite, on accède au paramétrage de la méthode insérée par double-clic sur son icône. On valide, on exécute et ainsi de suite.

Exemple de construction d'un nouveau diagramme

1. **Créer un nouveau diagramme.** Sur le dossier « Diagrammes de traitements » de la vue projet, accéder au menu contextuel (clic droit) et cliquer sur « Nouveau diagramme ». Le nouveau diagramme apparaît instantanément dans la vue « Diagrammes ».
2. **Ajouter une méthode d'importation de données** dans le diagramme à partir du répertoire « Imports » de la rubrique « Imports / Exports » de données de la vue « Méthodes ». Double-cliquez sur la méthode d'import choisie ou sélectionnez et glissez-déposez la méthode dans le diagramme. La méthode apparaît alors à l'écran.
   La vue « Connexions » permet également d'ajouter dans le diagramme une table de données à importer, à condition d'avoir préalablement défini une connexion.
3. **Paramétrer la méthode** d'importation des données. Double-cliquer sur la méthode ou cliquer sur « Paramétrer » dans le menu contextuel accessible par clic droit sur la méthode.
   La fenêtre de paramétrage de la méthode d'import apparaît. Modifier le paramétrage si besoin et valider le paramétrage.
4. **Exécuter la méthode**. Selon que l'option « Exécuter automatiquement une méthode après validation » est active ou non (voir dans les préférences générales du menu Outils), exécuter la méthode en cliquant sur « Exécuter » du menu contextuel de la méthode ou sélectionner la méthode et tapez F5.
5. **Contrôler et valider** l'import de données à partir de la vue « Données » et de ses différents éléments. Les données ne sont affichées que si la méthode d'import est sélectionnée.
6. **Ajouter / insérer une nouvelle méthode** dans le diagramme. A partir des vues « Méthodes » ou « Favoris », ajouter une nouvelle méthode dans le diagramme par double-clic ou glisser/déposer.

A ce stade, le paramétrage de la nouvelle méthode ajoutée n'est pas accessible car on ne sait pas quelles sont les données sur lesquelles repose la méthode. Il est pour cela nécessaire de spécifier à la méthode ajoutée sa table de données source (ici la méthode d'import).

7. **Définir une méthode « Source » à la méthode ajoutée** par en reliant la méthode « Source » à la méthode ajoutée. La liaison est matérialisée par une flèche dont le sens indique la relation d'ordre dans la transition des données et l'ordonnancement des traitements.
   Cette notion d'ordre est prépondérante dans la construction d'un diagramme. Dès lors que la flèche apparaît. La nouvelle méthode est paramétrable.
8. **Paramétrer** la nouvelle méthode en double-cliquant sur la méthode.
9. **Exécutez la nouvelle méthode.** (voir étape 4)
10. **Contrôler et valider la nouvelle méthode.** (voir étape 5).
11. **Reproduire les étapes 6 à 10 autant de fois que le nécessitent les traitements à réaliser.**

Note 1. Les étapes 6 et 7 font l'objet du paragraphe suivant.

Note 2. Certaines méthodes de data management nécessitent plusieurs méthodes sources (jointure, juxtaposition...

Ajouter / Insérer une méthode

La différence entre l'ajout et l'insertion d'une méthode réside dans le fait que l'insertion génère le lien entre une méthode existante et la nouvelle méthode ajoutée.

A partir de la vue « Méthodes »

Ajouter une méthode. Double-cliquez sur la méthode choisie. La méthode apparaît instantanément dans le diagramme courant.

Insérer une méthode. Glissez/déposez la méthode choisie sur sa méthode « source » dans le diagramme courant. L'apparition du symbole ✔ indique que l'insertion est autorisée. La nouvelle méthode apparaît alors à droite de la méthode « source » et les deux méthodes sont reliées par une flèche.

Le symbole ⊘ indique que l'insertion est refusée.

A partir de la barre d'outils

Ajouter une méthode. Aucune méthode ne doit être sélectionnée dans le diagramme courant.
Cliquez sur le bouton pour faire apparaître la boite de dialogue suivante :

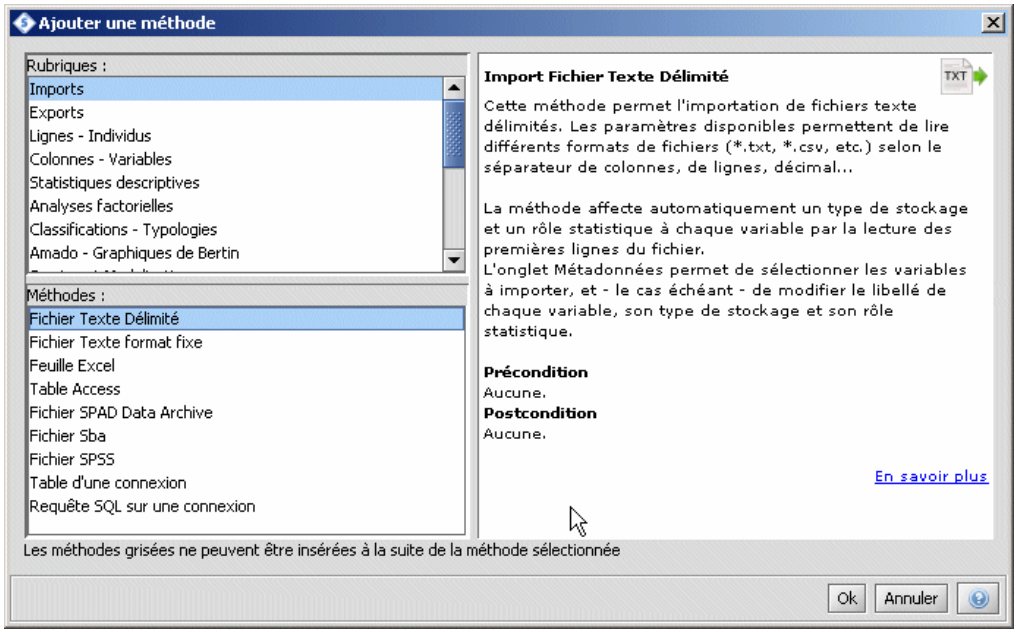


Figure 3-1 : Ajouter une méthode

Sélectionnez la rubrique et la méthode à ajouter et cliquez sur OK. La méthode apparaît instantanément dans le diagramme courant.

Insérer une méthode. Sélectionnez la méthode « source » dans le diagramme après laquelle vous souhaitez insérer une nouvelle méthode.
Cliquez sur le bouton pour faire apparaître la boite de dialogue décrite précédemment. Les méthodes interdites pour l'insertion sont grisées.
Sélectionnez une méthode autorisée et cliquez OK. La nouvelle méthode apparaît alors à droite de la méthode « source » et les deux méthodes sont reliées par une flèche.

Définir un lien entre deux méthodes

La définition d'un lien entre deux méthodes induit une relation d'ordre représentée par le sens de la flèche.
Certaines méthodes nécessitent la présence obligatoire d'autres méthodes en amont de leur insertion pour pouvoir être exécutée. C'est le cas par exemple des méthodes de la rubrique « Classifications – Typologies » qui requièrent la présence d'une analyse factorielle préalable dans l'enchainement.

Nous allons dans l'exemple suivant connecter la méthode « CREDIT » à la méthode « Stats ». Dans cet exemple la flèche ne peut aller que de « CREDIT » vers « Stats » puisque « CREDIT » est un import de données.

Pour connecter ces deux méthodes, il faut :

1. Positionner la souris sur la méthode « CREDIT » à la périphérie de l'icône jusqu'à faire apparaître l'étoile bleue suivante :

2. A l'apparition de cette étoile, cliquer et diriger la souris sur la méthode Stats tout en maintenant le clic gauche appuyé.
3. Si la connexion entre les deux méthodes est licite, le symbole ✔ apparaît. C'est le cas.
4. Relâcher alors le clic pour faire apparaître le lien entre les deux méthodes.

Supprimer un lien entre deux méthodes. Pour supprimer un lien entre deux méthodes, sélectionnez-le à l'aide de la souris et tapez sur la touche Suppr.

Paramétrer une méthode

Pour accéder au paramétrage d'une méthode, double-cliquez sur la méthode ou par cliquez droit sur la méthode et « Paramétrer ». Hormis les méthodes d'import, il est nécessaire que la ou les méthodes « source » de la méthode à paramétrer aient été exécutées préalablement.

Validez le paramétrage en cliquant sur OK.

Exécuter une méthode

Pour exécuter une (ou plusieurs) méthode, sélectionnez-là dans le diagramme courant et tapez sur la touche F5 ou cliquez droit sur la méthode et « Exécuter ».

Pour ré-exécuter une chaine de traitement dont l'une des méthodes sources auraient été modifiées, demander uniquement l'exécution de la dernière méthode de la chaine. SPAD se chargera d'exécuter les méthodes nécessaires à l'obtention du résultat demandé.

La vue « Exécutions » permet de voir l'avancement des traitements. (cf. page 11).

Renommer une méthode

Pour renommer une méthode, double-cliquez sur le nom de la méthode dans le diagramme et entrez le nom choisi.
La touche Entrée permet d'écrire un nom sur plusieurs lignes.
Pour valider le nouveau nom, cliquez à l'extérieur de la case éditée.

SPAD impose l'unicité des noms de méthodes au sein d'un même diagramme.
En cas de conflit, SPAD ajoute en fin de libellé « _1 » si une méthode porte déjà le même nom.

Indicateurs d'état des méthodes - Statut

Toutes les méthodes de SPAD disposent d'un indicateur d'état, présent dans le coin haut droit de leur icône. Cet indicateur suit la nomenclature suivante :

 **Gris** : méthode non paramétrée.

 **Bleu** : méthode paramétrée et validée, non exécutée.

 **Orange** : méthode paramétrée non validée, comporte des erreurs dans le paramétrage.

 **Vert** : méthode exécutée, résultats disponibles.

 **Jaune, marque verte** : méthode exécutée, résultats disponibles, comporte des avertissements.

 **Croix Rouge** : méthode exécutée, résultats non disponibles, comporte des erreurs.

Visualiser les résultats d'une méthode

La visualisation des résultats n'est possible que pour les méthodes dont l'indicateur d'état est Vert ou Jaune marque verte.

Les **méthodes d'import** et de **data management** produisent des résultats sous forme d'une **table de données** rectangulaire, directement accessible à partir de la **vue Données** (cf. page 12).

Les **méthodes statistiques** produisent différents types de résultats :

- Des **listings compacts** accessibles à partir de l'éditeur de résultats,
- Des **graphiques**, modifiables et interactifs,
- Des **sorties Excel** avec mise en forme automatique des résultats.
- Un **rapport de log** listant le compte-rendu des traitements.

On accède à ces résultats par le menu contextuel (clic droit), « Résultats ». Dans l'exemple ci-après, on affiche les résultats disponibles pour une analyse en composantes principales :

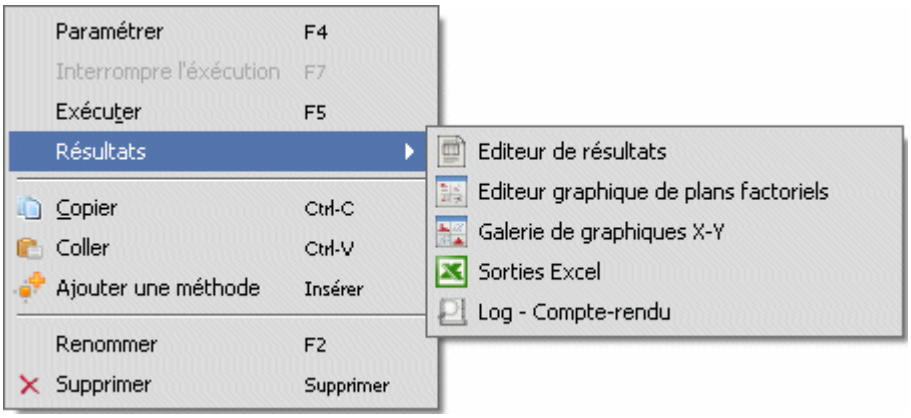


Figure 3-2 : Accéder aux résultats d'une méthode

Les résultats des méthodes statistiques sont également accessibles par la colonne « Résultats » de la vue Exécutions (cf. page 11).

Menu contextuel

Accessible à partir du clic droit sur une méthode, le menu contextuel de la vue « Diagrammes » permet les commandes suivantes.

Paramétrer. Accéder à la fenêtre de paramétrage de la méthode sélectionnée.

Interrompre l'exécution. Interrompre l'exécution de la méthode en cours d'exécution.

Exécuter. Exécuter la méthode sélectionnée.

Résultats. Accéder aux différents résultats de la méthode exécutée.

Copier. Copier la méthode ou les méthodes sélectionnées dans le presse-papiers.

Coller. Coller le contenu du presse-papiers là ou est situé le curseur de la souris.

Ajouter une méthode. Insérer une méthode après la méthode sélectionnée.

Renommer. Renommer la méthode.

Supprimer. Supprimer la méthode.

Barre d'outils

Ajouter / **Insérer méthode** à partir de la boite de dialogue « Ajouter méthode » (cf. Figure 3-1, page 27).

Copier l'ensemble des méthodes sélectionnées dans le presse-papiers.

Coller le contenu du presse-papiers dans le diagramme courant.

Supprimer les éléments sélectionnés dans le diagramme courant.

Réorganiser le diagramme courant sous forme d'arbre.

Aligner les méthodes sélectionnées.

Augmenter le niveau de **zoom**.

 **Diminuer** le niveau de **zoom**.

 **Ajuster le zoom** pour voir toutes les méthodes du diagramme.

 Revenir au niveau de **zoom par défaut**.

 **Mettre au premier plan** les éléments sélectionnés.

 **Mettre à l'arrière-plan** les éléments sélectionnés.

 Faire **valider** le paramétrage des méthodes sélectionnées.

 **Exécuter** l'ensemble des méthodes sélectionnées.

 **Interrompre** l'exécution en cours des méthodes sélectionnées.

Chapitre

4

Importation de données

Panorama général

SPAD permet d'accéder à de nombreuses sources de données hétérogènes au sein d'un même diagramme.

L'accès aux données sources peut se faire soit par la vue « Méthodes » et son répertoire « Imports / Exports » soit par la vue « Connexions » qui concerne plus spécifiquement les bases de données.

Quelle que soit la source de données (fichier texte, table d'une base de données…), SPAD affecte à chaque colonne un nom, un type de stockage et un rôle statistique que l'utilisateur peut modifier soit manuellement soit par typage personnalisé dans les métadonnées présentées ci-après.

Lors de l'exécution de l'import, la vue « Exécution » permet de voir l'état d'avancement de l'import avec le nombre de lignes importées qui s'affiche dans la colonne « Indicateur ».

A l'issue de l'import, les données sont visibles dans la vue « Données » de l'interface.

Les données importées sont stockées dans une base de données interne associée au projet.
Pour pouvoir les réutiliser dans d'autres diagrammes ou d'autres projets, il est nécessaire d'archiver ces données avec la méthode « Fichier SPAD Data Archive » du répertoire « Export » de la vue « Méthodes ».

Les métadonnées

Quel que soit le type de fichier à importer avec SPAD, toute méthode d'import dispose d'un onglet Métadonnées qui permet de spécifier le nom, le type de stockage et le rôle statistique de chaque champ ou variable.

Dans la Figure 4-1 ci-après, l'onglet Métadonnées liste l'ensemble des champs trouvés dans le fichier à importer, dans l'ordre initial.

Par défaut, SPAD affecte à chaque champ disponible un type de stockage (colonne « Stockage ») et un rôle statistique (colonne « Rôle »). Pour les fichiers Texte (format fixe ou délimité), les types par défaut sont déterminés par la lecture des 30 premières lignes du fichier.

Pour les bases de données et l'import Excel, le type est déterminé par le type défini dans la base de données source.

Les stockages et rôles affectés par défaut **peuvent être modifiés** manuellement en **double-cliquant** dans la case voulue, par clic droit ou par **typage personnalisé.**

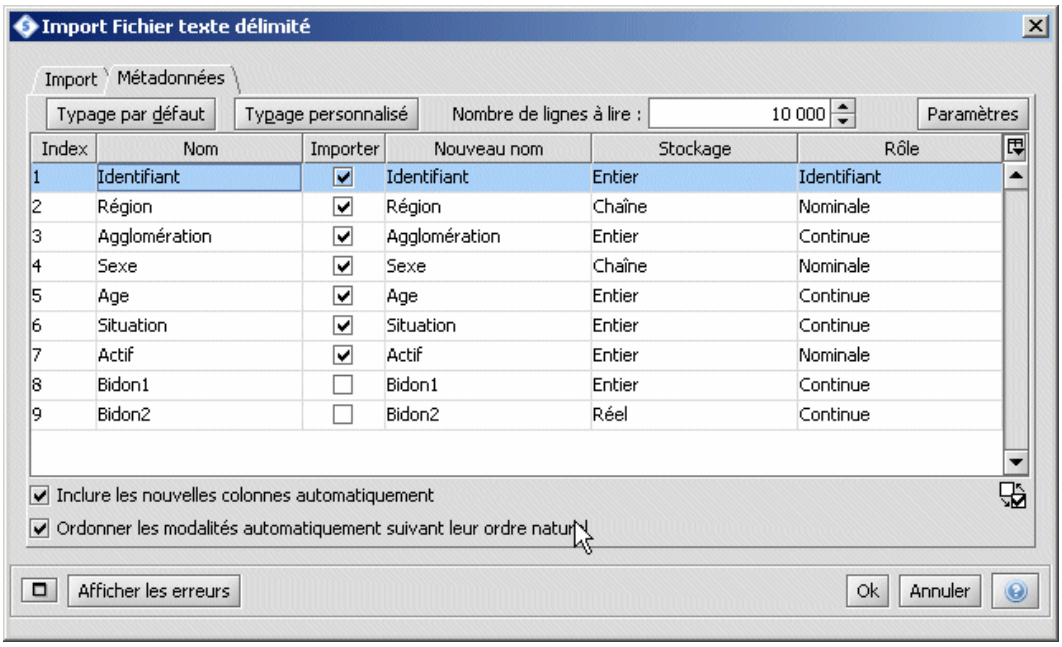


Figure 4-1 : Panneau Métadonnées

Index. Ordre d'apparition de la variable dans le fichier

Nom. Nom du champ trouvé dans l'entête du fichier ou généré automatiquement par SPAD.
Il n'y a pas de limitation sur le nombre de caractères du nom dans le stockage interne de la base de données SPAD. Ce nom est tronqué aux 60 premiers caractères pour les différentes méthodes statistiques.

Importer. Cochée par défaut, le champ sera importé. En décochant la case, le champ n'est pas importé.

Nouveau nom. Par défaut, le « nouveau nom » est identique au « nom ». Il est possible de modifier le nom du champ directement dans la case.

Stockage. Le type de stockage assigné à un champ correspond au format de stockage du champ dans la base de données interne SPAD. Les types disponibles sont : Entier, Réel, Chaine de caractères, Date, Autre. Pour modifier le type, double-cliquez sur la case pour faire apparaître la liste des types ou cliquez sur la case et tapez la première lettre du type choisi. Les différents types de stockage sont décrits ci-après :

- **Entier** : Nombre entier codé sur un ou plusieurs chiffres pouvant être précédé du signe – pour les nombres négatifs. Ex. : 34
- **Réel** : Nombre comportant une partie décimale. La partie entière est séparée de la partie décimale par un séparateur (point ou virgule) Les nombres réels peuvent être précédés du signe – pour les nombres négatifs. Ex. : 12.234
- **Chaine de caractères** : suite de caractères alphanumériques définis par le code ASCII. Les chaines caractères peuvent éventuellement être « entourées » d'un identificateur (' ou ").
- **Date** : Date stockée sous la forme Jour/Mois/Année Heure : Minute : Seconde. L'affichage des dates diffère selon le format d'affichage choisi dans les préférences.
- **Autre** : type de données non reconnu par SPAD.

Rôle. Le rôle assigné à un champ correspond au rôle statistique que jouera la variable dans les différentes méthodes d'analyse de données de SPAD. On retrouve principalement deux types de variables : nominales ou continues. Cette distinction entre variables nominales et continues est importante dans la mesure où l'on n'applique pas les mêmes calculs selon la nature de la variable. L'interface des méthodes statistiques est organisée autour de cette notion de rôle de variable et évite tout risque d'erreur méthodologique.
Pour modifier le rôle, double-cliquez sur la case pour faire apparaître la liste des rôles ou cliquez sur la case et tapez la première lettre du rôle choisi.

Les rôles disponibles sont :

- **Identifiant** : désigne le ou les champs servant à identifier les individus. Si plusieurs champs sont spécifiés comme Identifiant, ceux-ci seront concaténés pour définir un seul identifiant pour chaque individu dans les méthodes statistiques.
  Le libellé est utilisé dans les graphiques, les éditions de certaines méthodes, les outils de gestion des bases, mais ce n'est pas une variable sur laquelle on peut appliquer des méthodes statistiques.
- **Nominale** : désigne une variable nominale ou qualitative prenant des valeurs alphanumériques et/ou numériques que l'on appelle « **modalités** ». Les cases vides seront considérées comme données manquantes.
  Ex. : Sexe de la personne (codée H / F ou codée 1 / 2) est une variable à deux modalités.
- **Continue** : désigne une variable continue ou quantitative (obligatoirement typée Entier ou Réel) admettant un continuum de valeurs entières ou décimales sur lesquelles on pourra calculer des indicateurs tels que la moyenne, l'écart-type, etc. Toute valeur non-numérique sera considérée comme donnée manquante, au même titre que les cases vides. Ex. : Revenu du foyer.
- **Textuelle** (pour l'analyse textuelle) : désigne une variable textuelle admettant des données alphanumériques, destinée à l'analyse de données textuelles disponible dans le module Text Mining de SPAD.
- **Exclue** : désigne une variable qui ne sera jamais prise en compte dans les différentes méthodes statistiques de SPAD. Une variable exclue reste disponible pour l'ensemble des méthodes de data management de SPAD.

Inclure les nouvelles colonnes automatiquement. Ce paramètre concerne la ré-exécution d'une importation. Si entre-temps, le fichier a été modifié et qu'il comporte de nouvelles colonnes ou variables, cette option permet de les importer automatiquement. Si cette option est décochée, les nouvelles variables ne sont pas importées.

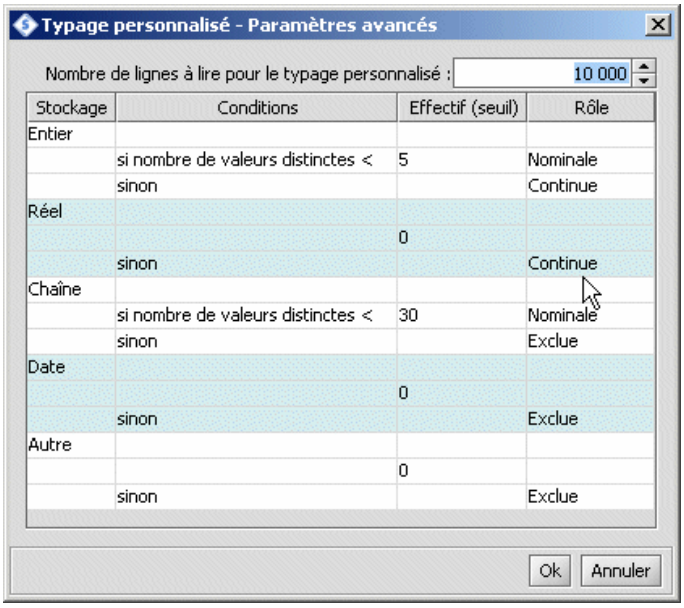
Ordonner les modalités automatiquement suivant leur ordre naturel. Cochée par défaut, cette option trie automatiquement les modalités des variables déclarées nominales dans l'ordre alphabétique si le type de stockage est « Chaine de caractères » et dans l'ordre numérique si le type de stockage est « Entier » ou « Réel ».
Cet ordre sera utilisé pour l'édition des résultats des différentes méthodes statistiques. Cet ordre peut être modifié manuellement avec les méthodes « Format » et « Edition des libellés » du répertoire « Data Management - Colonnes » de la vue « Méthodes ».
Si l'option est décochée, les modalités seront ordonnées selon l'ordre dans lequel elles apparaissent dans le fichier.

Typage par défaut : Affecte un type par défaut à l'ensemble des variables. Pour les fichiers de type texte (format fixe ou délimité), les types par défaut sont déterminés par la lecture des 30 premières lignes du fichier.
Pour les bases de données et l'import Excel, le type est déterminé par le type défini dans la base de données source.

Typage personnalisé : Affecte un type de stockage et un rôle statistique aux variables selon des règles prédéfinies et modifiables dans les paramètres avancés. **Le typage personnalisé est recommandé car il affine le typage par défaut.** Il permet notamment d'éviter la création de variables nominales avec un nombre de modalités très élevé. Ce qui généralement présente un intérêt limité pour les traitements.
Les paramètres du typage personnalisé sont modifiables en cliquant sur le bouton « Paramètres avancés » :

Remarque : pour effectuer un typage par défaut ou un typage personnalisé sur une sélection de variable, sélectionnez la ou les variables concernées et utilisez le menu contextuel du clic droit.

Typage personnalisé – Paramètres avancés



Stockage	Conditions	Effectif (seuil)	Rôle
Entier			
	si nombre de valeurs distinctes <	5	Nominale
	sinon		Continue
Réel			
		0	
	sinon		Continue
Chaîne			
	si nombre de valeurs distinctes <	30	Nominale
	sinon		Exclue
Date			
		0	
	sinon		Exclue
Autre			
		0	
	sinon		Exclue

Figure 4-2 : Paramètres avancés du typage personnalisé

Les paramètres avancés du typage personnalisé ci-dessus se traduisent de la manière suivante :
Les types de stockage (« Stockage ») sont déterminés par la lecture des 10 000 premières lignes du fichier.
Les rôles statistiques sont définis en fonction du type et des caractéristiques de chaque variable :

- Toutes les variables de type de stockage « **Entier** » ou « **Réel** » qui comprennent moins de 10 valeurs distinctes (sur les 1000 premières lignes) sont considérées comme variables nominales. Sinon, ce sont des variables continues.
- Toutes les variables de type de stockage « **Chaine** » qui comprennent moins de 100 valeurs distinctes (sur les 1000 premières lignes) sont considérés comme variables nominales. Sinon, ces champs sont exclus des traitements statistiques.
- Toutes les variables de type de stockage « **Date** » sont exclues des traitements statistiques.
- Toutes les variables de type de stockage « **Autre** » sont exclues des traitements statistiques.

Pour modifier les paramètres avancés, double-cliquez sur les cases concernées des colonnes « Effectif (seuil) » et « Rôle ».

Remarque : les variables qui ne comportent qu'une seule valeur ou aucune valeur sont automatiquement exclues des traitements statistiques.

Correspondances stockage / rôle statistique et limitations sur la taille des libellés

Tableau de correspondance des types de stockage et des rôles statistiques

Ce tableau synthétise l'ensemble des correspondances possibles entre type de stockage et rôles statistiques.

Rôle Statistique SPAD / Stockage SPAD	Nominale	Continue	Textuelle	Exclue	Identifiant
Entier	OK	OK	OK	OK	OK
Réel	OK	OK	OK	OK	OK
Chaine	OK	⃠	OK	OK	OK
Date	OK	⃠	OK	OK	OK
Autre	⃠	⃠	⃠	OK	⃠

⃠ indique que la correspondance entre le stockage et le rôle est interdite.

Remarques :
Le rôle « Exclue » est destiné aux variables dont l'intérêt est limité pour le traitement statistique. Les variables « Exclues » restent disponibles pour les méthodes de data management mais ne sont pas utilisables dans les méthodes statistiques. Cela permet de gagner en clarté et en vitesse d'exécution dans ces méthodes.

Il est recommandé d'assigner le rôle de variable nominale aux variables dont le nombre de valeurs distinctes n'est pas trop élevé (inférieur à 500).

Tableau de correspondance entre stockage d'origine et stockage SPAD

Le tableau suivant permet d'illustrer l'impact d'éventuelles erreurs de correspondances entre le type de stockage observé dans les données sources et celui affecté pour l'import dans SPAD.

Stockage SPAD / Stockage Source	Entier	Réel	Chaine	Date
Entier : 11	11	11.0	11	_Missing_
Réel : 2.5	2	2.5	2.5	_Missing_
Chaine : Oui	_Missing_	_Missing_	Oui	_Missing_
Chaine : 011	11	11.0	011	_Missing_
Date : 12/09/2007	_Missing_	_Missing_	12/09/2007	12/09/2007

Missing indique que la correspondance entre le stockage source et le stockage SPAD produira des données manquantes dans la base de données interne SPAD sur l'exemple choisi.

Remarque :
Si des valeurs numériques sont stockées sous forme de chaines de caractères dans une base de données, il est possible de leur affecter le stockage SPAD « Entier » ou « Réel » au moment de l'import.

Limitations sur les libellés

Il n'y a **aucune limitation sur la taille des libellés dans la base de données interne** SPAD.
Les limitations interviennent pour les procédures statistiques, à la fois lors des calculs et de l'affichage des résultats.
Les limitations sur la taille des libellés pour les procédures statistiques sont les suivantes :

- Libellé de variable : 60 caractères
- Libellé de modalité : 20 caractères
- Identifiant individu : 60 caractères

Lorsque la taille des libellés dépasse le maximum autorisé, les libellés sont tronqués pour respecter la taille limite. Un message d'avertissement accompagne l'exécution de la méthode concernée.

Import de données par la vue Méthodes

Le répertoire « Import / Export » de la vue « Méthodes » permet d'importer les formats de données suivants :

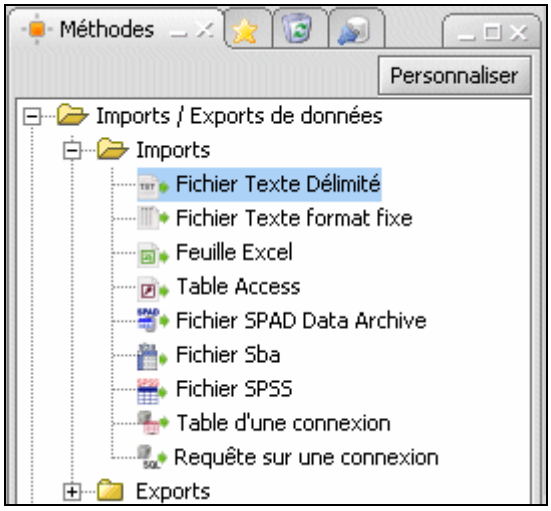


Figure 4-3 : Import de données

Import Txt – Fichier Texte délimité

La méthode « Fichier Texte Délimité » permet de lire tout type de fichier texte au format délimité dont les formats les plus courants sont *.txt et *.csv.

Panneau Import

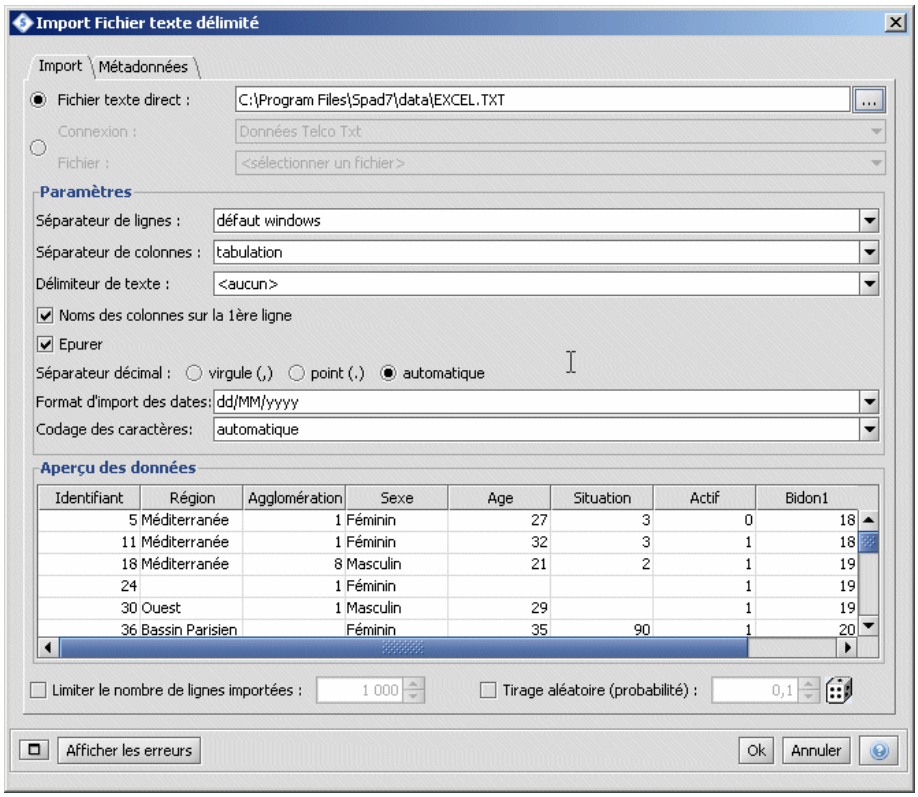


Figure 4-4 : Import Fichier Texte Délimité

Paramétrage pour l'import d'un fichier texte délimité

Fichier texte direct. Spécifiez le nom du fichier. Vous pouvez entrer le nom du fichier à importer ou cliquer sur le bouton « … » pour sélectionner le fichier.

Connexion / Fichier. Si le fichier à importer appartient à une connexion créée à partir de la vue Connexions, sélectionnez-là et spécifiez le fichier à importer.

Séparateur de lignes. Spécifiez le séparateur de lignes. Pour modifier le choix par défaut, cliquez sur le bouton de la liste déroulante pour faire apparaître les séparateurs de lignes disponibles.

Séparateur de colonnes. Spécifiez le séparateur de colonnes utilisé dans le fichier texte à importer. Pour modifier le choix par défaut, cliquez sur le bouton de la liste déroulante pour choisir un autre séparateur (virgule, point-virgule, espace, retour à la ligne…).
Pour spécifier un autre séparateur que ceux proposés, entrez le séparateur choisi directement dans la case.

Délimiteur de texte. Utilisez la liste déroulante pour spécifier le délimiteur de texte (apostrophe ou guillemet) utilisé dans le fichier à importer.

Noms des colonnes sur la première ligne. Cochée par défaut, cette option considère la première ligne du fichier à importer comme les entêtes des colonnes. Si ce n'est pas le cas dans le fichier à importer, décochez cette case. SPAD générera automatiquement un nom pour chaque colonne, tel que Col_1, Col_2, par rapport au positionnement de la colonne dans le fichier.

Epurer. Cochée par défaut, cette option supprime les espaces de début et de fin pour chaque valeur importée.

Séparateur décimal. Par défaut, le séparateur décimal « auto » détermine automatiquement le séparateur décimal utilisé pour chaque colonne numérique. Il permet de gérer dans un même fichier des colonnes numériques utilisant le point et la virgule comme séparateur.
Sélectionnez soit le point (.), soit la virgule (,) pour modifier le choix par défaut.

Format d'import des dates. Utilisez la liste déroulante pour sélectionner le format approprié de lecture des variables que vous déclarerez en type « Date » dans le panneau « Métadonnées ».
Vous pouvez spécifier vous-même le format en suivant la nomenclature suivante :

Lettre	Désignation	Présentation	Exemples
y	Année	Numérique	1996; 96
M	Mois dans l'année	Texte ou numérique	Avril; Avr; 04; 4
w	Semaine dans l'année	Entier	27
W	Semaine dans le mois	Entier	2
D	Jour dans l'année	Entier	189
d	Jour dans le mois	Entier	10
F	Jour de la semaine dans le mois	Entier	2
E	Jour de la semaine dans le mois	Texte	Mardi; Mar
a	Indicateur Am/pm	Texte	PM
H	Heure (0-23)	Entier	0
k	Heure (1-24)	Entier	24
K	Heure pour am/pm (0-11)	Entier	0
h	Heure pour am/pm (1-12)	Entier	12
m	Minute dans l'heure	Entier	30
s	Seconde dans la minute	Entier	55
S	Milliseconde	Entier	978
Z	Zone horaire	Texte	CET

Figure 4-5 : Nomenclature pour le format des dates

Si votre date comporte des séparateurs, les introduire tels quels dans le format.

Si votre fichier contient des dates de différents formats, affectez-leur le type « Chaine » dans le panneau « Métadonnées » et utilisez la méthode « Format » pour assigner à chaque date son format spécifique en reprenant la nomenclature précédente.

Codage des caractères. Spécifiez le codage des caractères utilisé dans le fichier à importer à partir de la liste déroulante. Par défaut (automatique), SPAD choisit le même codage que celui du système d'exploitation de la machine sur laquelle il est installé.

Aperçu des données. L'aperçu du fichier texte à importer est rafraichi dès que le paramétrage est modifié.

Limitation du nombre de lignes importées.
Vous pouvez limiter le nombre de lignes importées, soit en nombre soit par tirage aléatoire. Pour le tirage aléatoire, spécifiez la probabilité qu'aura chaque ligne de figurer dans les données importées.
Exemple : en spécifiant une probabilité de 0.3, les lignes importées représenteront plus ou moins 30% du fichier d'origine.
En cliquant sur  , vous pouvez modifier l'initialisation du tirage pour obtenir un échantillon différent.
Vous pouvez utiliser conjointement le tirage aléatoire et la limitation en nombre : dans ce cas, le nombre de lignes importées sera égal à la première des deux limitations atteinte.

Panneau Métadonnées

Voir en page 32.

Import Fichier Texte Fixe

La méthode « Fichier Texte Format Fixe » permet de lire tout type de fichier texte au format fixe.

Panneau Import

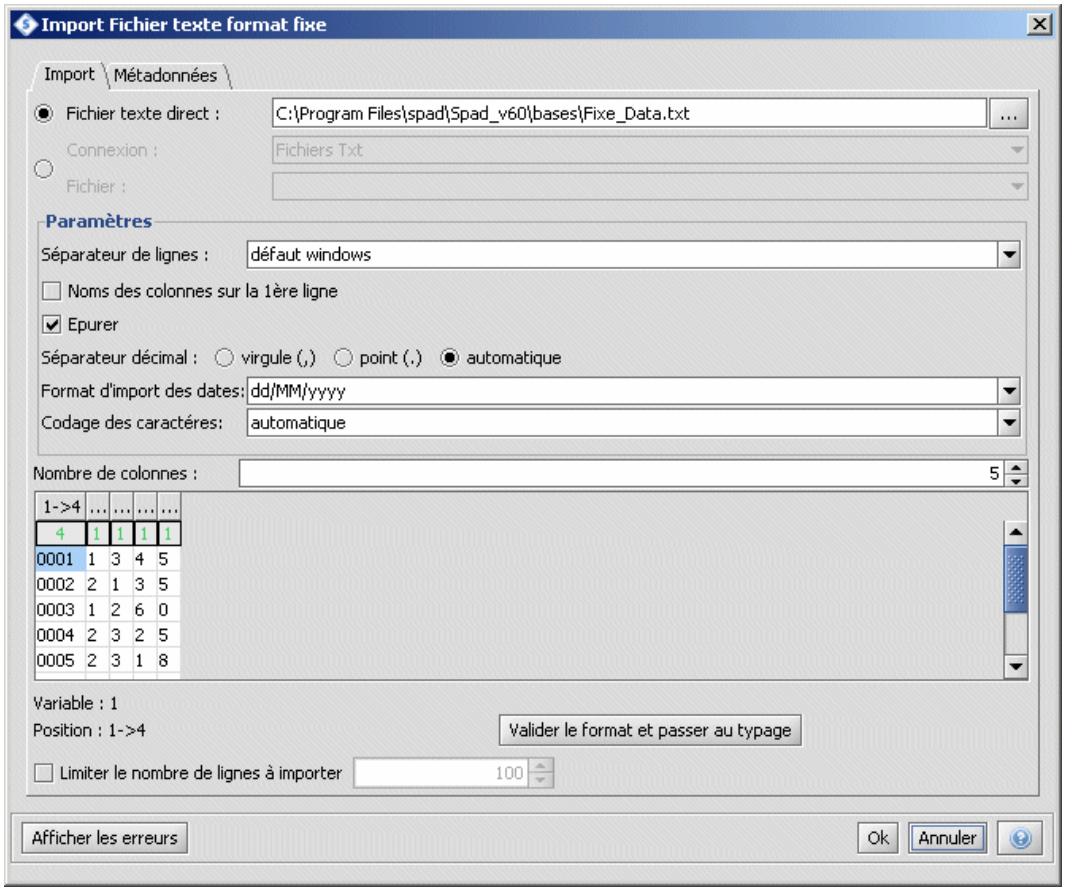


Figure 4-6 : Import Txt format fixe

Paramétrage pour l'import de fichiers texte format fixe

Fichier texte direct. Spécifiez le nom du fichier. Vous pouvez entrer le nom du fichier à importer ou cliquer sur le bouton « … » pour sélectionner le fichier.

Connexion / Fichier. Si le fichier à importer appartient à une connexion créée à partir de la vue Connexions, sélectionnez-là et spécifiez le fichier à importer.

Séparateur de lignes. Spécifiez le séparateur de lignes. Pour modifier le choix par défaut, cliquez sur le bouton de la liste déroulante pour faire apparaître les séparateurs de lignes disponibles.

Noms des colonnes sur la première ligne. Non cochée par défaut. Lorsqu'elle est cochée, cette option considère la première ligne du fichier à importer comme les entêtes des colonnes. Si ce n'est pas le cas dans le fichier à importer, décochez cette case. SPAD générera automatiquement un nom pour chaque colonne, tel que Col_1, Col_2, par rapport au positionnement de la colonne dans le fichier.

Epurer. Cochée par défaut, cette option supprime les espaces de début et de fin pour chaque valeur importée.

Séparateur décimal. Par défaut, le séparateur décimal « auto » détermine automatiquement le séparateur décimal utilisé pour chaque colonne numérique. Il permet de gérer dans un même fichier des colonnes numériques utilisant le point et la virgule comme séparateur.
Sélectionnez soit le point (.), soit la virgule (,) pour modifier le choix par défaut.

Format d'import des dates. Utilisez la liste déroulante pour sélectionner le format approprié de lecture des variables que vous déclarerez en type « Date » dans le panneau « Métadonnées ». Si votre fichier contient des dates de différents formats, affectez-leur le type « Chaine » dans le panneau « Métadonnées » et utilisez la méthode « Format » pour assigner à chaque date son format spécifique.

Codage des caractères. Spécifiez le codage des caractères utilisé dans le fichier à importer à partir de la liste déroulante. Par défaut (automatique), SPAD choisit le même codage que celui du système d'exploitation de la machine sur laquelle il est installé.

Nombre de colonnes. Déclarez le nombre de colonnes du fichier à importer.

Aperçu des données. L'aperçu du fichier à importer permet de choisir la largeur des champs, soit à l'aide de la souris en déplaçant les séparateurs verticaux dans les entêtes de colonnes, soit en double-cliquant sur les cases grisées et en spécifiant la longueur de la variable en nombre de caractères. Dans cet exemple, la première variable est codée sur 4 caractères, les 4 suivantes sur un seul.

En cliquant sur une des cases du tableau, on fait apparaître ses coordonnées : variable concernée et position.
En cliquant droit sur une des cases d'une colonne, on fait apparaître trois options qui permettent de :

- appliquer la taille de la colonne cliquée droit à toutes les colonnes suivantes
- appliquer la taille de la colonne cliquée droit aux colonnes sélectionnées
- répéter les tailles des colonnes sélectionnées aux colonnes suivantes.

Limiter le nombre de lignes à importer. Voir page 40.

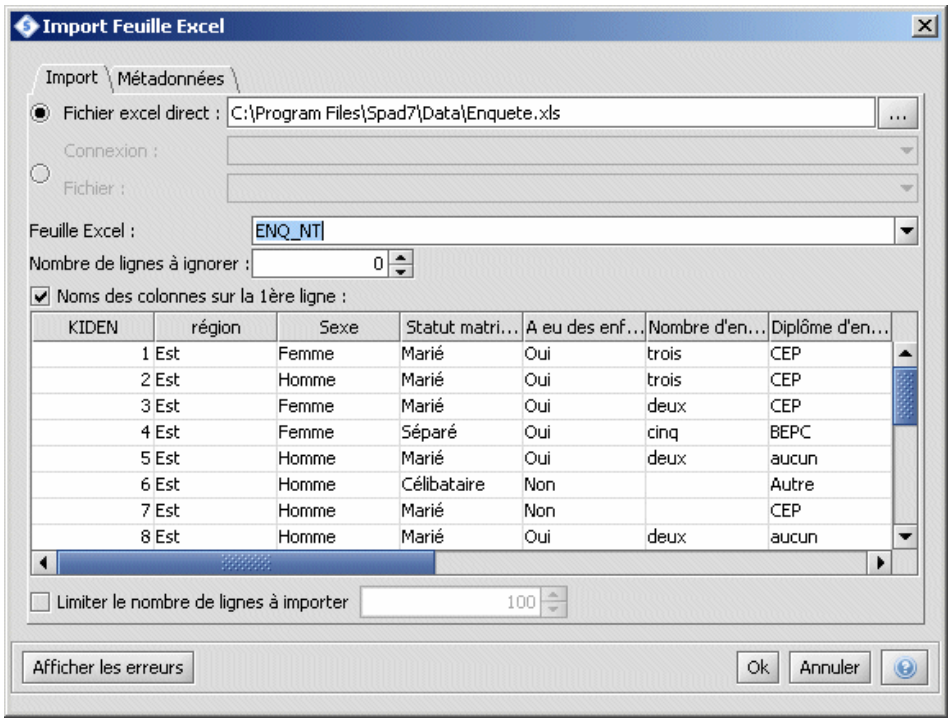
Panneau Métadonnées

Voir en page 32.

Import Feuille Excel

La méthode « Feuille Excel » permet de lire une feuille d'un classeur Excel.

Panneau Import



KIDEN	région	Sexe	Statut matri...	A eu des enf...	Nombre d'en...	Diplôme d'en...
1	Est	Femme	Marié	Oui	trois	CEP
2	Est	Homme	Marié	Oui	trois	CEP
3	Est	Femme	Marié	Oui	deux	CEP
4	Est	Femme	Séparé	Oui	cinq	BEPC
5	Est	Homme	Marié	Oui	deux	aucun
6	Est	Homme	Célibataire	Non		Autre
7	Est	Homme	Marié	Non		CEP
8	Est	Homme	Marié	Oui	deux	aucun

Figure 4-7 : Import Feuille Excel

Paramétrage pour l'import d'une feuille Excel

Fichier Excel direct. Spécifiez le classeur Excel (*.xls) contenant la feuille Excel à importer. Vous pouvez entrer le nom du fichier à importer ou cliquer sur le bouton « … » pour sélectionner le fichier.

Connexion / Fichier. Si le fichier à importer appartient à une connexion créée à partir de la vue Connexions, sélectionnez-là et spécifiez le fichier Excel à importer.

Feuille. Sélectionnez la feuille Excel à importer en cliquant sur le bouton de la liste déroulante.

Nombre de lignes à ignorer. Spécifiez le nombre de ligne à ignorer avant de démarrer la l'import.

Noms des colonnes sur la première ligne. Cochée par défaut, cette option considère la première ligne du fichier à importer comme les entêtes des colonnes. Si ce n'est pas le cas dans le fichier à importer, décochez cette case. SPAD générera automatiquement un nom pour chaque colonne, tel que Col_1, Col_2, par rapport au positionnement de la colonne dans le fichier.

Aperçu des données. L'aperçu des données est rafraichi dès que le paramétrage est modifié.

Limiter le nombre de lignes à importer. Voir page 40.

Panneau Métadonnées

Voir en page 32.

Import Table Access

La méthode « Table Access » permet de lire toute table et toute vue (requête) d'une base de données Access.

Panneau Import

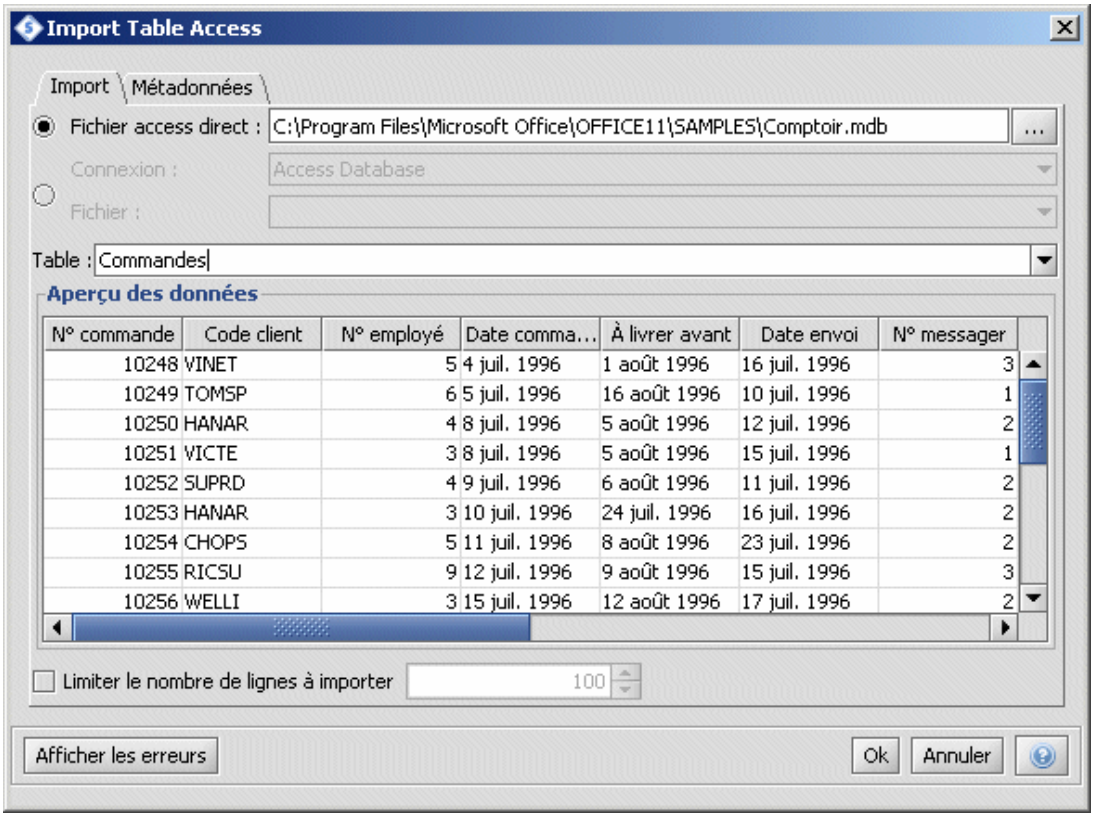


Figure 4-8 : Import Table Access

Fichier Access direct. Spécifiez la base Access (*.mdb) contenant la table ou la vue à importer. Vous pouvez entrer le nom du fichier à importer ou cliquer sur le bouton « … » pour sélectionner le fichier.

Connexion / Fichier. Si la base Access à importer appartient à une connexion créée à partir de la vue Connexions, sélectionnez-là et spécifiez la base Access à importer.

Table. Sélectionnez la table ou la vue à importer en cliquant sur le bouton de la liste déroulante.

Aperçu des données. L'aperçu des données est rafraichi dès que le paramétrage est modifié.

Limiter le nombre de lignes à importer. Voir page 40.

Panneau Métadonnées

Voir en page 32.

Import Fichier Spad Data Archive

Cette méthode permet de lire et d'importer tout fichier au format Sda, nouveau format d'archivage des données dans la version 7. Ce format de fichier présente l'avantage de conserver l'ensemble des informations sur les métadonnées et est particulièrement compact.

Panneau Import

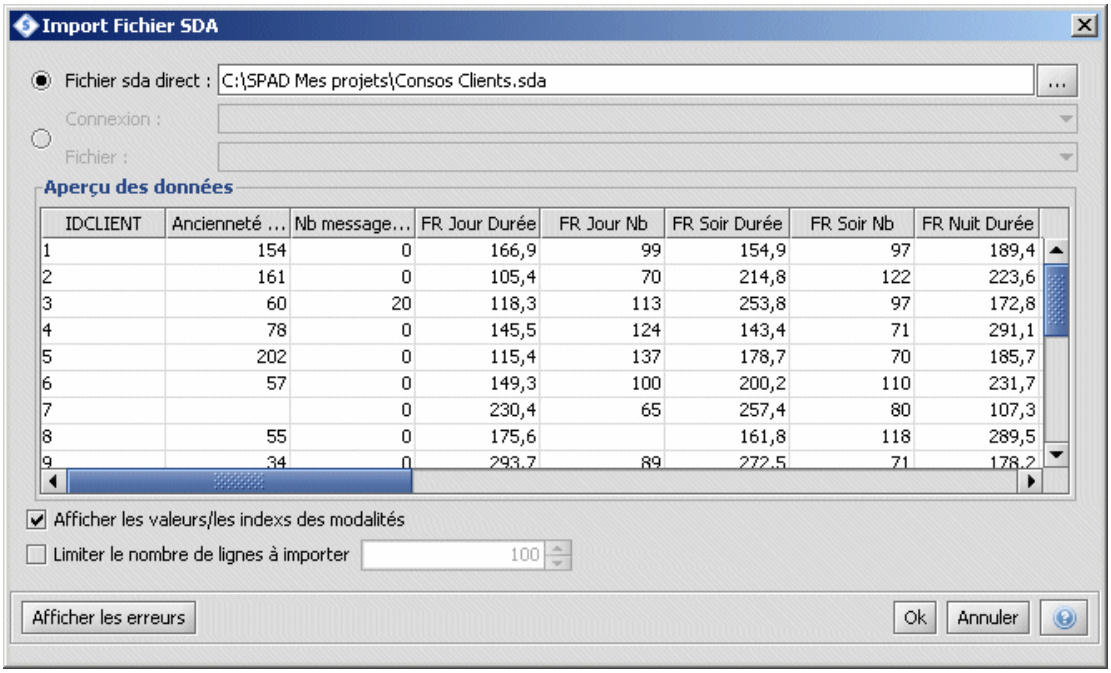


Figure 4-9 : Import Fichier Sda

Fichier SDA direct. Spécifiez le fichier Sda à importer. Vous pouvez entrer le nom du fichier à importer ou cliquer sur le bouton « … » pour sélectionner le fichier.

Connexion / Fichier. Si le fichier Sda à importer appartient à une connexion créée à partir de la vue Connexions, sélectionnez-là et spécifiez le fichier Sda à importer.

Afficher les étiquettes/les valeurs. Cette option permet, pour les variables nominales, de basculer entre les valeurs numériques et les libellés de modalités.

Aperçu des données. L'aperçu des données est rafraichi dès que le paramétrage est modifié.

Limiter le nombre de lignes à importer. Voir page 40.

Remarque : ce nœud d'import de données est le seul avec l'import de fichiers Sba qui ne comporte pas d'onglet Métadonnées, étant donné que toutes les informations sont déjà définies dans le fichier Sda.

Import Fichier SPSS

La méthode « Fichier SPSS » permet de lire tout fichier SPSS d'extension *.sav.

Panneau Import

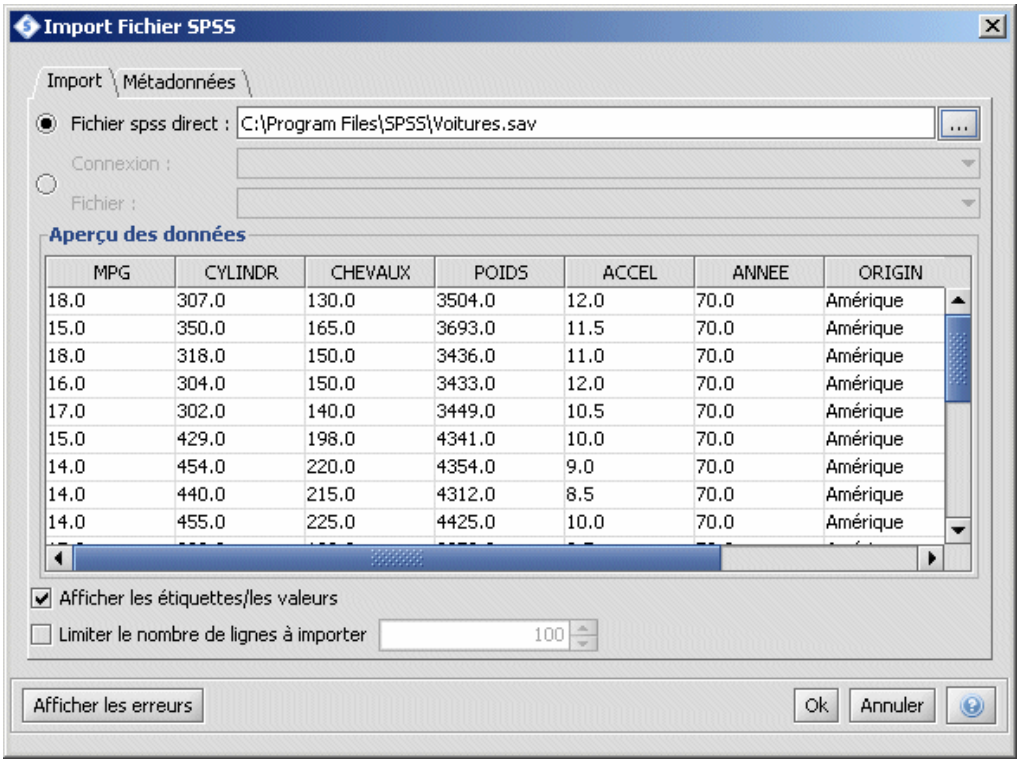


Figure 4-10 : Import Fichier SPSS

Fichier SPSS direct. Spécifiez le fichier SPSS à importer. Vous pouvez entrer le nom du fichier à importer ou cliquer sur le bouton « … » pour sélectionner le fichier.

Connexion / Fichier. Si le fichier SPSS à importer appartient à une connexion créée à partir de la vue Connexions, sélectionnez-là et spécifiez le fichier SPSS à importer.

Afficher les étiquettes/les valeurs. Cette option permet de basculer entre les valeurs de données numériques et les étiquettes de valeurs descriptives définies par l'utilisateur dans la base SPSS.

Aperçu des données. L'aperçu des données est rafraichi dès que le paramétrage est modifié.

Limiter le nombre de lignes à importer. Voir page 40.

Panneau Métadonnées

Voir en page 32.

Remarque : le panneau Métadonnées pour l'import SPSS comporte une colonne supplémentaire intitulée « **Etiquette/Valeur** ». Cette colonne permet, pour chaque variable, de spécifier si l'on importe les valeurs ou les étiquettes de valeurs.
Par ailleurs, pour utiliser les **libellés longs de variables SPSS** comme nouveaux noms des variables dans SPAD, utilisez le menu contextuel (clic droit) pour appliquer le choix à la sélection de variable.

Import Fichier Sba

La méthode « Fichier Sba » permet de lire toute base Sba (format standard des bases SPAD jusqu'en version 6.5).

Panneau Import

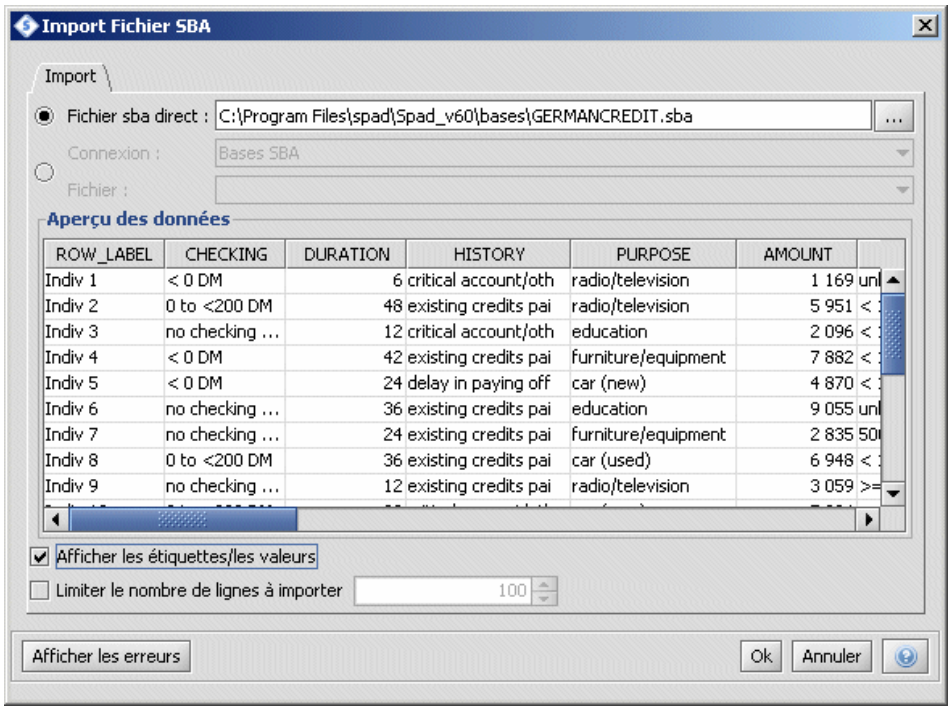


Figure 4-11 : Import Fichier Sba

Fichier Sba direct. Spécifiez le fichier Sba à importer. Vous pouvez entrer le nom du fichier à importer ou cliquer sur le bouton « … » pour sélectionner le fichier.

Connexion / Fichier. Si le fichier Sba à importer appartient à une connexion créée à partir de la vue Connexions, sélectionnez-là et spécifiez le fichier Sba à importer.

Afficher les étiquettes/les valeurs. Cette option permet, pour les variables nominales, de basculer entre les valeurs numériques et les libellés de modalités définis par l'utilisateur dans la base SPAD.

Aperçu des données. L'aperçu des données est rafraichi dès que le paramétrage est modifié.

Limiter le nombre de lignes à importer. Voir page 40.

Remarque : ce nœud d'import de données est le seul qui ne comporte pas d'onglet Métadonnées, étant donné que toutes les informations sont déjà définies dans le fichier Sba.

Table d'une connexion

La méthode « Table d'une connexion » permet de lire toute table d'une connexion à une base de données, définie préalablement dans la vue « Connexions », voir en page 50.

Panneau Import

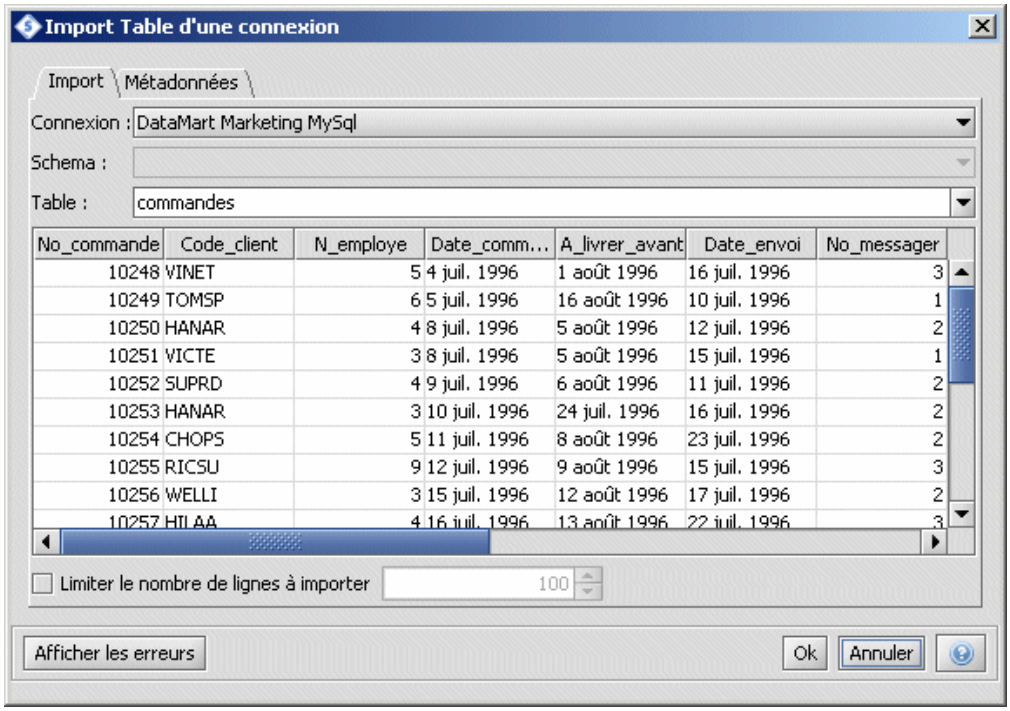


Figure 4-12 : Import Table d'une Connexion

Connexion. Spécifiez la connexion contenant la table ou la vue à importer en cliquant sur la liste déroulante.

Schéma. Spécifiez le schéma associé à la connexion si nécessaire.

Table. Sélectionnez la table ou la vue à importer en cliquant sur le bouton de la liste déroulante.

Aperçu des données. L'aperçu des données est rafraichi dès que le paramétrage est modifié.

Limiter le nombre de lignes à importer. Voir page 40.

Panneau Métadonnées

Voir en page 32.

Requête SQL sur une connexion

Le nœud « Requête SQL sur une connexion » permet de formuler une requête, exécutée directement au sein de la base de données définie dans la connexion, voir en page 50.
Seul le résultat de cette requête est importé dans SPAD.

Panneau Import

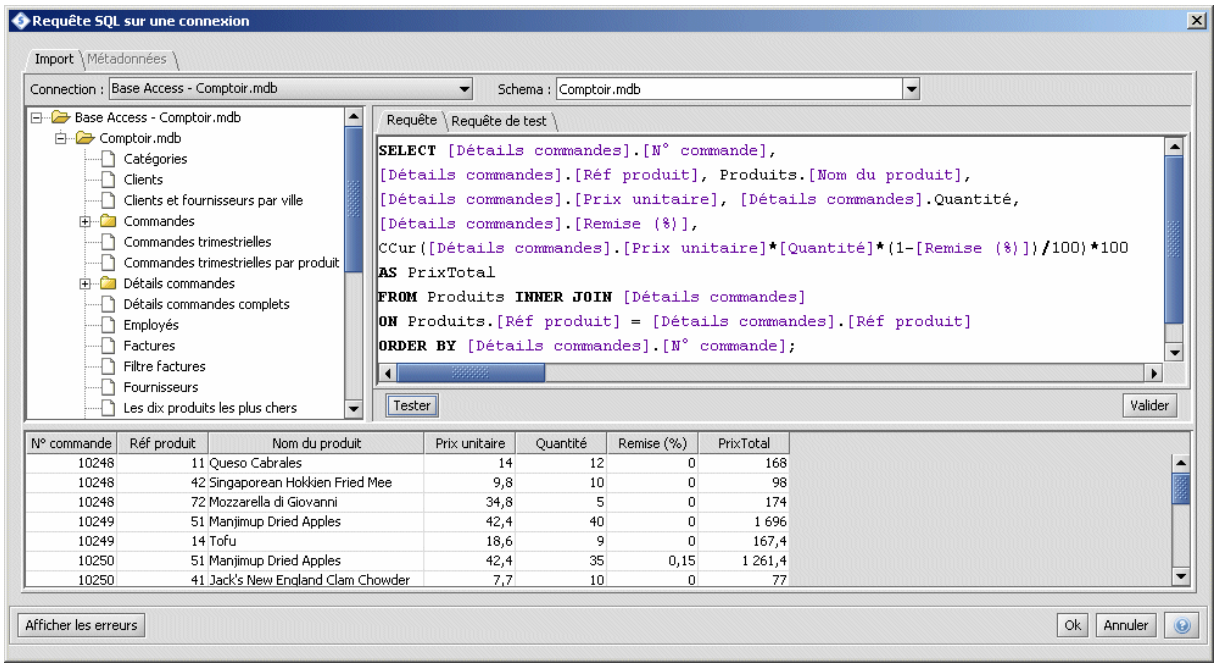


Figure 4-13 : Requête SQL sur une Connexion

Connexion. Spécifiez la connexion contenant la base de données sur laquelle appliquer la requête.

Schéma. Spécifiez le schéma associé à la connexion si nécessaire.

Explorateur. L'explorateur se situe sur la partie gauche de la fenêtre. Il permet d'accéder à chaque table de la base de données et à chacun de ses champs. Lors de la conception de la requête dans la vue requête, l'utilisateur peut insérer dans sa requête un nom de table ou de champ en double-cliquant sur la table ou le champ approprié dans l'explorateur.

Requête. L'onglet Requête est destiné à recevoir la requête sous forme d'instructions en langage SQL. Lors de l'exécution, elle sera exécutée au sein de la base de données. Le langage SQL utilisé doit être compatible avec celui de la base de données.

Tester. Le bouton Tester permet de tester la validité de la requête SQL et d'afficher dans la partie inférieure de la fenêtre un aperçu du résultat de la requête.

Valider. Le bouton Valider permet d'activer l'onglet Métadonnées pour modifier, le cas échéant, les noms de variables, leurs types de stockages et rôles statistiques. Si l'utilisateur clique directement sur OK sans valider, les variables seront typées par défaut selon leur type d'origine dans la base de données. Il est conseillé de vérifier les métadonnées avant d'exécuter l'importation de la requête.

Panneau Métadonnées

Voir en page 32.

Import de données par la vue Connexions

Le panneau « Connexion » permet de se connecter à l'ensemble des tables d'une base de données de type Oracle, DB2, SQL Server, Access, fichiers texte, SPSS, SAS, etc. en utilisant des liens vers des répertoires, des liaisons natives (JDBC) ou des liaisons ODBC.
L'utilisateur peut spécifier la connexion comme permanente pour rendre accessible ses données dans tout nouveau projet.

Pour accéder aux données d'une base de données, suivez les étapes suivantes :

1. Dans la vue « Connexions », cliquez sur le bouton  pour créer une nouvelle connexion,
2. Sélectionnez le type de connexion :
    - Répertoires vers des fichiers Sba, Sda, Texte, Access, Excel, Spss
    - Connexions directes aux bases de données pour pouvoir accéder aux bases de données MySQL, DB2, SAS (serveur), ORACLE, SQLServer
    - Sources ODBC pour accéder aux autres bases de données accessibles via ODBC et aux fichiers SAS.
3. Paramétrez la connexion
4. Testez la connexion
5. Validez

La liste des tables de la base de données apparaît alors dans la vue « Connexion ». Pour accéder aux données d'une table, déplacez la table dans le diagramme par glisser/déposer ou utilisez la méthode « Table d'une base de données » du répertoire Import de la vue « Méthodes ».

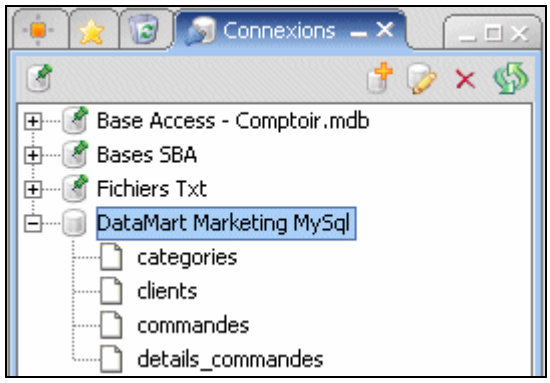
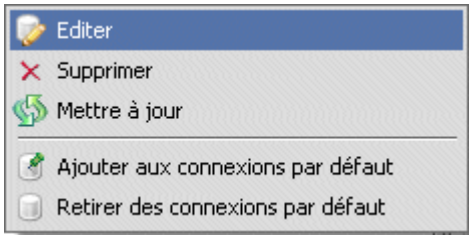


Figure 4-14 : Panneau Connexions

La gestion des connexions se fait par clic droit ou à l'aide de la barre de boutons.
Par clic droit, on accède au menu suivant :



Editer. Edite la connexion sélectionnée. Permet de modifier les attributs de la connexion.
Supprimer. Supprime définitivement la connexion. S'il s'agit d'une connexion par défaut, un message demande si la suppression est locale au projet ou générale.
Mettre à jour. Permet de rafraîchir les données de la connexion.
Ajouter aux connexions par défaut. Rend la connexion disponible pour chaque projet.
Retirer des connexions par défaut. Désactive la disponibilité permanente de la connexion.

Créer une connexion vers des fichiers Sba, Sda, Textes, Access, Excel, Spss

Pour définir une nouvelle connexion vers l'un des formats suivants : Sba, Sda, Textes, Access, Excel, Spss :

1. Cliquer sur le bouton  dans le panneau « Connexions »
2. Choisissez parmi les différents formats de répertoires proposés
3. Renseignez les paramètres suivants dans la fenêtre « Connexions » :

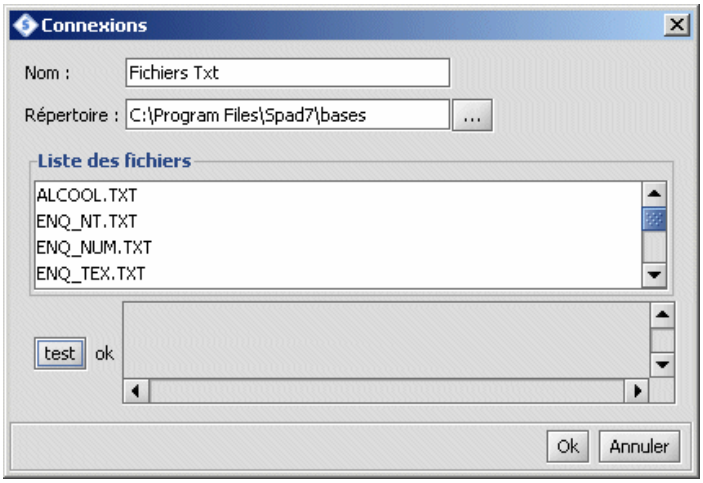


Figure 4-15 : Paramétrage d'une Connexion vers un répertoire

Nom. Nom arbitraire désignant la connexion dans SPAD. Chaque connexion doit avoir un nom différent.

Répertoire. Répertoire contenant les fichiers de la connexion. Vous pouvez taper le chemin du répertoire ou cliquer sur le bouton « … » pour sélectionner le répertoire à travers l'explorateur.
Ce répertoire peut se trouver sur la même machine que celle sur laquelle SPAD est installé, ou sur n'importe quelle autre machine.

Test. Ce bouton permet de tester la validité de la connexion. En cas d'erreur, un message s'inscrit à droite du bouton. En cas de bon fonctionnement, les éléments correspondant au format recherché sont listés à droite du bouton Test.

Créer une connexion directe à une base de données

Plus efficaces et plus sûres que les connexions ODBC, ces connexions utilisent Java et les drivers propriétaires pour se connecter directement aux serveurs de bases de données.

Pour définir une nouvelle connexion vers l'une des bases de données de type MySQL, DB2, SAS, ORACLE, SQLServer :

1. Cliquer sur le bouton  dans le panneau « Connexions »
2. Choisissez « Connexions aux bases de données »
3. Renseignez les paramètres suivants dans la fenêtre « Connexions » :

Hormis SAS qui nécessite un paramétrage spécifique, le paramétrage de ce type de connexion se fait comme suit :

- ***MySql, Oracle, SQLServer, DB2***

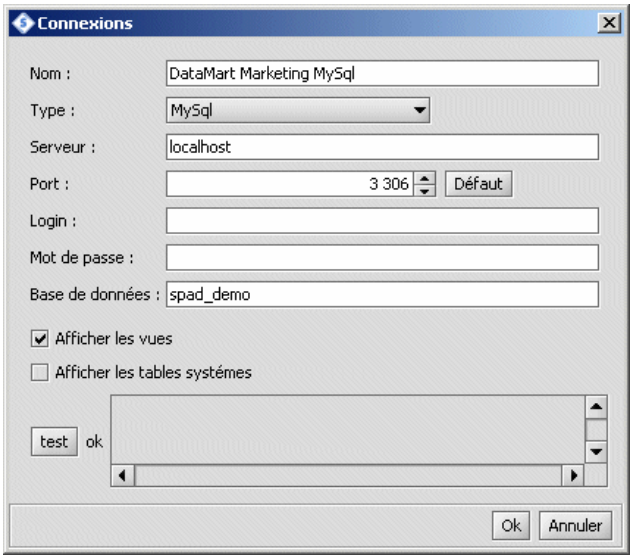


Figure 4-16 : Paramétrage d'une Connexion directe vers une base de données

Nom. Nom arbitraire désignant la connexion dans SPAD. Chaque connexion doit avoir un nom unique.

Type. Sélectionnez le type de serveur qui correspond à la base de données

Serveur. Nom de la machine sur laquelle se trouve le serveur de base de données.

Port. Port utilisé pour se connecter au serveur de base de données. Vous pouvez cliquer sur le bouton « Défaut » pour appliquer le port par défaut au serveur.

Login. Nom d'utilisateur ou login pour se connecter au serveur.
Mot de passe. Mot de passe associé au login pour se connecter au serveur.

Base de données. Nom de la base de données à laquelle cette connexion est liée.

Afficher les vues / les tables systèmes. Par défaut, toutes les vues de la connexion sont accessibles dans la connexion. Les tables systèmes sont par défaut non importables dans SPAD. Cochez la case si vous souhaitez pouvoir les importer.

Test. Ce bouton permet de tester la validité de la connexion. En cas d'erreur, un message s'inscrit à droite du bouton. En cas de bon fonctionnement, les éléments correspondant au format recherché sont listés à droite du bouton Test.

Remarque. Les connexions DB2 nécessitent un paramètre en plus : le type de serveur, à choisir entre DB2 UDB et DB2 Cloudscape.

En cliquant sur OK, la connexion apparaît dans le panneau connexion ci-dessous :

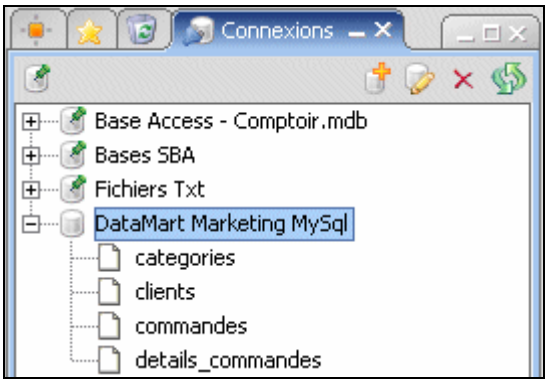


Figure 4-17 : Panneau Connexions

Pour importer une des tables de la connexion, double-cliquez sur la table ou glissez/déposez la table dans le diagramme. Spécifiez les paramètres de l'import tels que décrits dans le chapitre « Import Table d'une connexion », page 48.

Créer une connexion à une Source ODBC

Les connexions ODBC vous permettent d'importer tout type de sources de données dès lors que la source de données ODBC a été définie dans les outils d'administration du panneau de configuration de Windows.
Ce type de connexion est plus lent que les autres types de connexion. Il est à utiliser uniquement lorsque les connexions directes ne permettent pas d'accéder à vos données.

Pour pouvoir définir une connexion ODBC dans SPAD, il est nécessaire d'avoir préalablement défini la source de données ODBC dans le panneau de configuration windows. Si ce n'est pas le cas, procédez comme suit :

Création de la source de données ODBC :

1. Allez dans le « Panneau de configuration » puis « Outils d'administration »,
2. Puis « Sources de données (ODBC) »,
3. Dans l'onglet « Sources de données utilisateurs », cliquez sur « Ajouter »,
4. Choisissez le pilote correspondant au format des données que vous souhaitez importer et cliquez sur « Terminer »,
5. Selon le pilote choisi, renseignez les paramètres de la source ODBC et validez. Le nom de la source de données que vous avez spécifié sera réutilisé.

Création d'une connexion à une source ODBC dans SPAD

Pour définir une nouvelle connexion à une source ODBC :

1. Cliquer sur le bouton  dans le panneau « Connexions »
2. Choisissez « Source ODBC »
3. Renseignez les paramètres suivants dans la fenêtre « Connexions » :

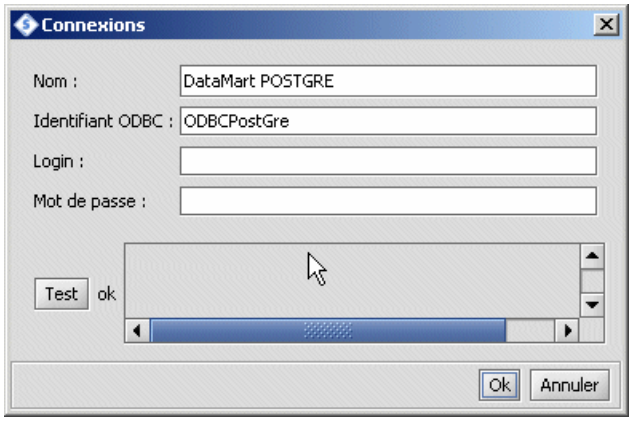


Figure 4-18 : Paramétrage d'une connexion ODBC

Nom. Nom arbitraire désignant la connexion dans SPAD. Chaque connexion doit avoir un nom unique.

Identifiant ODBC. Cet identifiant doit être exactement identique à celui que vous avez choisi lors de la création de la source ODBC.

Login. Selon la source de données ODBC, spécifiez votre login d'utilisateur.
Mot de passe. Selon la source de données ODBC, spécifiez le mot de passe associé au login.

Test. Ce bouton permet de tester la validité de la connexion. En cas d'erreur, un message s'inscrit à droite du bouton. En cas de bon fonctionnement, les éléments correspondant au format recherché sont listés à droite du bouton Test.

En cliquant sur OK, la connexion apparaît dans le panneau connexion ci-dessous :

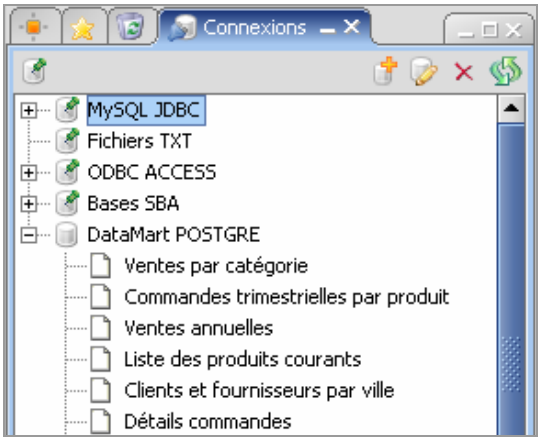


Figure 4-19 : Panneau Connexions

Pour importer une des tables de la connexion, double-cliquez sur la table ou glissez/déposez la table dans le diagramme. Spécifiez les paramètres de l'import tels que décrits dans le chapitre « Import Table d'une connexion », page 48.

Créer une connexion à une Source SAS ODBC – monoposte

L'importation de fichiers SAS dans SPAD requiert les éléments suivants :

- SAS monoposte doit être installé sur le PC
- Le driver SAS ODBC doit être installé sur le PC. Ce driver n'est pas installé par défaut lors de l'installation du système SAS. Il se trouve dans le CD-Rom [Client-Side Components Volume 1].
- Il faut définir une source SAS ODBC dans les options d'administration de votre PC.

Création de la source SAS ODBC :

1. Allez dans le « Panneau de configuration » puis « Outils d'administration »
2. Puis « Sources de données (ODBC) »
3. Dans l'onglet « Sources de données utilisateurs », cliquez sur « Ajouter »
4. Choisissez « SAS » comme nom de pilote
5. Si vous ne voyez pas « SAS » vous devez installer les drivers SAS ODBC. Ils sont disponibles sur le CD d'installation de SAS ou sur leur site web.
6. La fenêtre suivante apparaît :

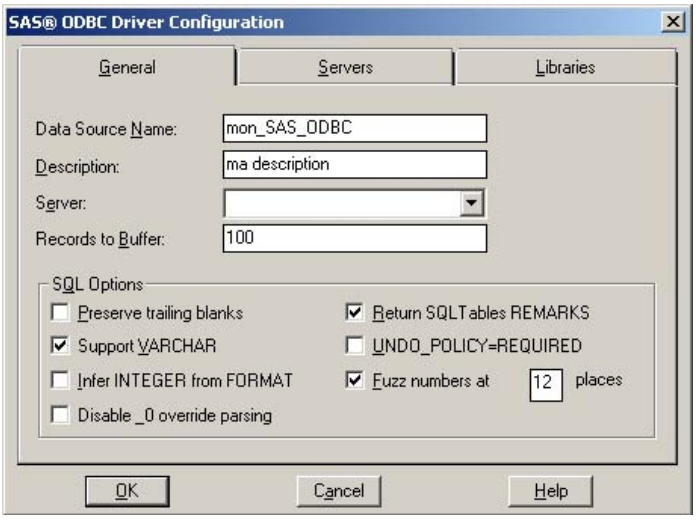


Figure 4-20 : Configuration du driver SAS ODBC - Général

1. Saisissez un nom pour « Data Source Name »
2. Une description si vous le désirez (optionnel)
3. Cochez les mêmes cases que sur l'image dans la partie basse
4. Cliquez ensuite sur l'onglet « Servers »

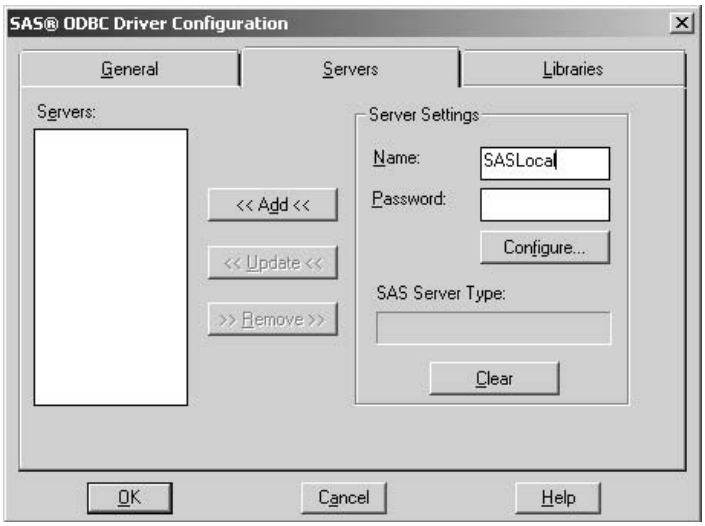


Figure 4-21 : Configuration du driver SAS ODBC – Serveur 1/2

1. Choisissez un nom pour votre source (ici « SASLocal »)
2. Cliquez sur le bouton « Configure »

1. Path : entrez le chemin vers l'exécutable sas.exe sur votre poste
2. Working Directory : saisissez le chemin vers le répertoire
3. Startup Parameters : laissez les paramètres par défaut
4. Timeout : 60
5. Cliquez sur « OK
6. Cliquez sur l'onglet « Libraries »

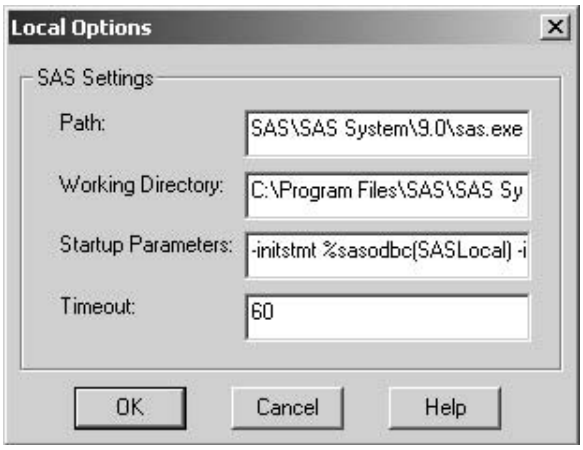


Figure 4-22 : Configuration du driver SAS ODBC – Options locales

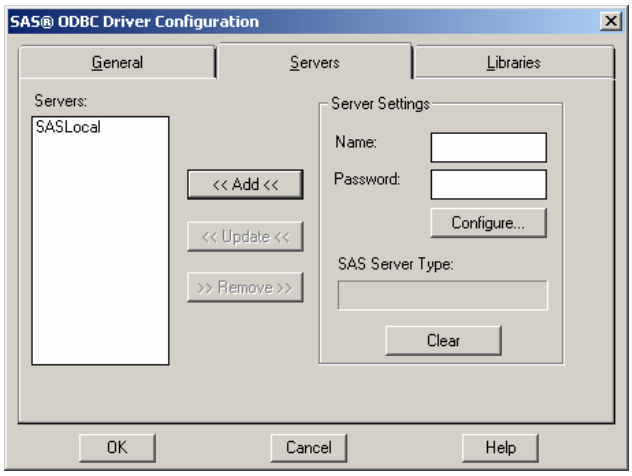


Figure 4-23 : Configuration du driver SAS ODBC – Serveur 2/2

1. Cliquez sur le bouton « Add » pour ajouter le nouveau serveur
2. Cliquez sur l'onglet « Libraries »

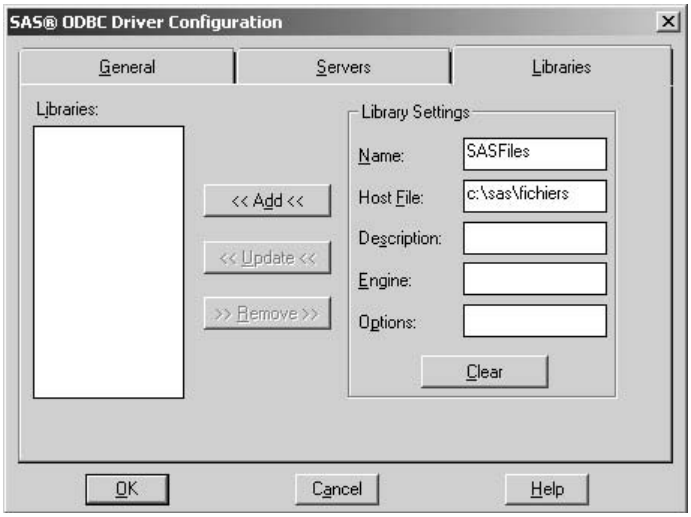


Figure 4-24 : Configuration du driver SAS ODBC – Libraries 1/2

1. Name : le nom de votre librairie SAS. (libref SAS). Cette librairie doit être une librairie permanente.
2. Host File : le chemin vers la librairie SAS contenant les fichiers à importer.
3. Cliquez sur le bouton « Add »

Note. Vous pouvez spécifier plusieurs librairies. Elles apparaîtront sous la forme de répertoires dans la connexion.

1. Cliquez sur le bouton « OK »

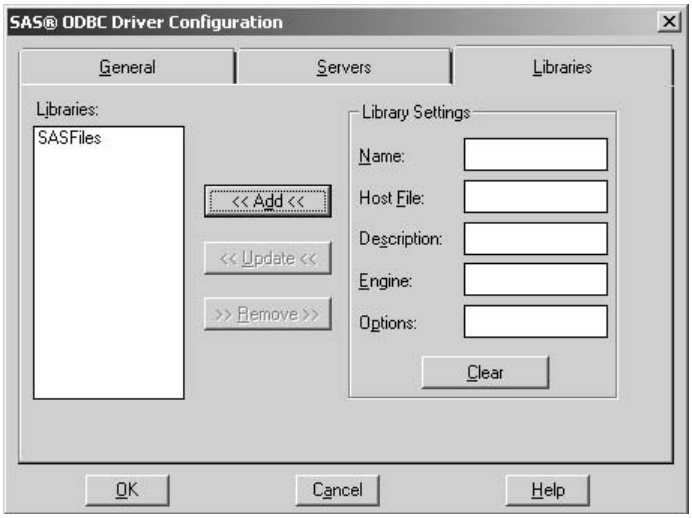


Figure 4-25 : Configuration du driver SAS ODBC – Libraries 2/2

La dernière étape consiste à ajouter une ligne d'information à un fichier système Windows.

1. Editez le fichier C:\Windows\system32\drivers\etc\services dans le bloc-note
2. Vous devez ajouter une ligne semblable aux autres (avant la dernière)

   SASLocal 5010/tcp #local sas server

Le premier mot doit être le nom exact que vous avez choisi pour votre serveur SAS (voir Figure 4-23). Il est possible que vous deviez redémarrer Windows.

Vous avez terminé la configuration de votre source ODBC SAS. Cliquez sur OK.

Création de la connexion à une source SAS ODBC dans SPAD

Pour définir une nouvelle connexion à une source SAS ODBC :

1. Cliquer sur le bouton  dans le panneau « Connexions »
2. Choisissez « Source SAS par ODBC »
3. Renseignez les paramètres suivants dans la fenêtre « Connexions » :

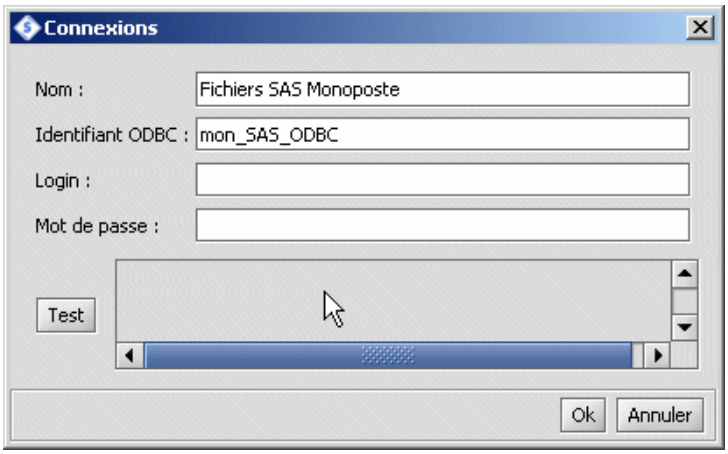


Figure 4-26 : Connexion SAS ODBC

Nom. Nom arbitraire désignant la connexion dans SPAD. Chaque connexion doit avoir un nom unique.

Identifiant ODBC. Cet identifiant doit être exactement identique à celui que vous avez spécifié pour « Data Source Name » (voir Figure 4-20).

Login. Sans objet.
Mot de passe. Sans objet.

En cliquant sur OK, la connexion apparaît dans la vue connexion. Une instance de SAS est ouverte automatiquement afin de permettre la lecture des tables SAS mais vous ne pourrez pas l'utiliser directement. La fermeture de cette instance de SAS entraîne l’impossibilité de lire les tables SAS. Une simple mise à jour de la connexion par clic droit permet de la relancer.

Chapitre

5

Exportation de données

Panorama général

Le répertoire « Export » de la vue « Méthodes » permet d'exporter des données vers les formats de données suivants :

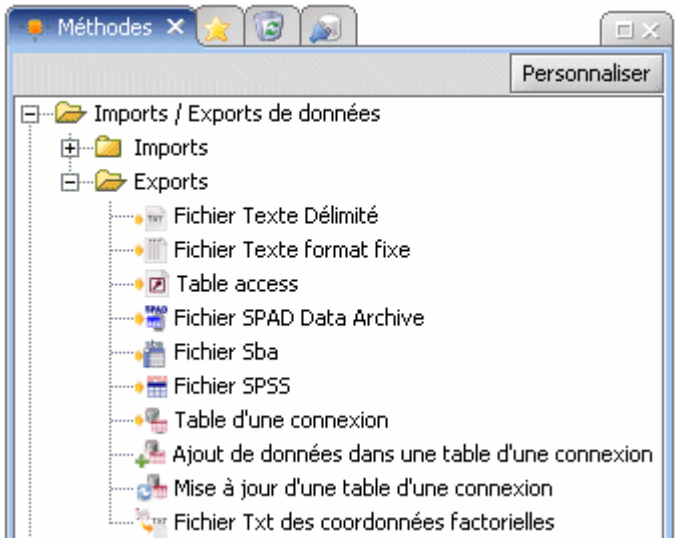


Figure 5-1 : Méthodes d'exportation de données

Les méthodes d'exportation de données constituent la fin d'un enchaînement. On ne peut plus y connecter de nouvelles méthodes.

Fichier texte délimité

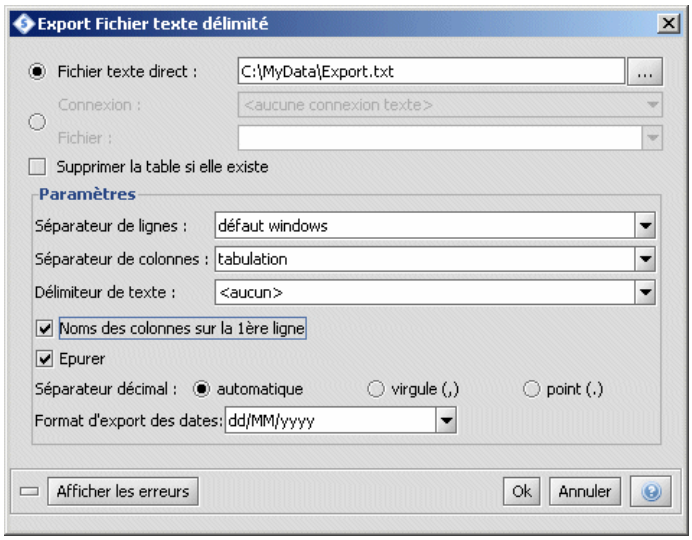


Figure 5-2 Export de Fichier Txt Délimité

Fichier Texte direct. Spécifiez le nom du fichier dans lequel les données seront exportées. Vous pouvez entrer le chemin et le nom du fichier ou cliquer sur le bouton « … » pour accéder à l'explorateur. Dans cette fenêtre, vous pouvez spécifier le type du fichier (*.txt ou *.csv).
Connexion / Fichier. Si vous souhaitez exporter vos données dans une connexion de type texte, sélectionnez-là et spécifiez le nom du fichier (avec l'extension) à créer.

Supprimer le fichier s'il existe. Cochez cette case si vous souhaitez écraser un fichier existant.

Séparateur de lignes. Spécifiez le séparateur de lignes. Pour modifier le choix par défaut, cliquez sur le bouton de la liste déroulante pour faire apparaître les séparateurs de lignes disponibles.

Séparateur de colonnes. Spécifiez le séparateur de colonnes à utiliser dans le fichier exporté. Pour modifier le choix par défaut, cliquez sur le bouton de la liste déroulante pour choisir un autre séparateur (virgule, point-virgule, espace, retour à la ligne…).
Vous pouvez taper un autre séparateur directement dans la case.

Délimiteur de texte. Utilisez la liste déroulante pour spécifier le délimiteur de texte (apostrophe ou guillemet) utilisé dans le fichier à importer.

Noms des colonnes sur la première ligne. Décochez cette option si vous ne souhaitez pas faire figurer les noms de variables dans le fichier exporté.

Epurer. Cochée par défaut, cette option supprime les espaces de début et de fin pour les noms de variables et l'ensemble des valeurs.

Séparateur décimal. Par défaut, le séparateur décimal « automatique » est déterminé en fonction des paramètres du système d'exploitation.
Sélectionnez soit le point (.), soit la virgule (,) pour modifier le choix par défaut.

Format d'export des dates. Utilisez la liste déroulante pour sélectionner le format approprié d'export des variables de type « Date ». Vous pouvez spécifier vous-même le format en vous référant à la nomenclature spécifiée Figure 4-5, page 39.

Export Table Access

Cette méthode d'export permet plusieurs types d'export vers Access :

- Création d'une nouvelle base de données Access,
- Ajout d'une nouvelle table dans une base de données Access,
- Ecrasement d'une table existante dans une base de données Access.

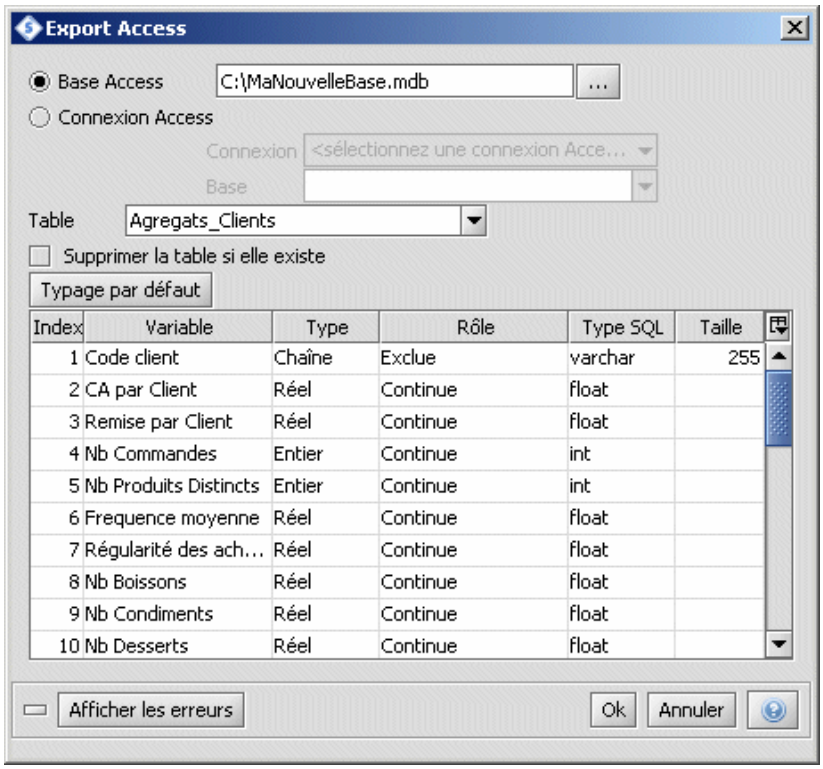


Figure 5-3 : Export Access

Base Access. Spécifiez le nom de la nouvelle base Access à créer ou sélectionnez une base existante en accédant à l'explorateur par le bouton « … ».

Connexion. Spécifiez la connexion Access dans laquelle seront exportées les données.
Base. Spécifiez le nom de la nouvelle base Access à créer dans la connexion ou sélectionnez une des bases existantes dans la liste déroulante.

Table. Entrez le nom de la nouvelle table à créer ou sélectionnez la table à écraser dans la liste déroulante. L'opération d'écrasement est irréversible.

Supprimer la table si elle existe. Cochez cette case si vous souhaitez écraser une table existante.

Typage par défaut. Ce bouton permet de revenir au paramétrage par défaut lorsque des modifications ont été apportées au « Type SQL » et à la « Taille » des variables.

Type SQL et Taille. Par défaut, SPAD assigne un type SQL et une taille (en nombre de caractères uniquement pour le Type SQL « String ») à chaque variable en fonction de son type de stockage SPAD. Vous pouvez modifier le type SQL et la taille en double cliquant dans les cases choisies.

Fichier SPAD Data Archive *.sda

Cette méthode permet d'exporter des données dans une archive SPAD SDA. Ce format de fichier présente l'avantage de conserver l'ensemble des informations sur les métadonnées et est particulièrement compact.

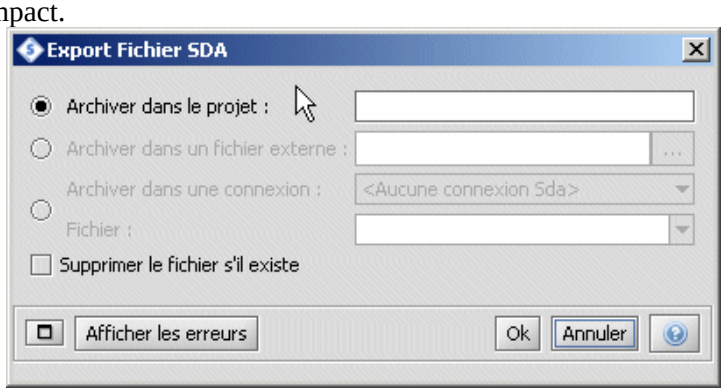


Figure 5-4 : Export Fichier SDA

Archiver dans le projet. Spécifiez le nom de l'archive Sda à créer. A la fin de l'exécution, le fichier apparaît dans le dossier « Données Archives » de la vue projet.
L'archive SDA est enregistrée dans le répertoire du projet SPAD auquel elle appartient. Pour réutiliser cette archive dans le projet, double-cliquez sur l'archive dans la vue « Projet » pour la faire apparaître dans le diagramme et exécutez.

Archiver dans un fichier externe. Spécifiez le nom du fichier SDA à créer en tapant le chemin et le nom du fichier dans la case ou cliquer sur le bouton « … » pour accéder à l'explorateur.

Archiver dans une connexion. Spécifiez la connexion SDA et tapez le nom du fichier à créer dans la connexion. La connexion doit avoir préalablement définie dans la vue « Connexions ».

Supprimer le fichier s'il existe. Cochez cette case si vous souhaitez écraser un fichier existant.

Fichier Sba

Le format SBA est l'ancien format standard de données des versions précédentes de SPAD.
Toutes les variables disponibles sont exportées hormis celles dont le rôle statistique est « Exclue ».
Les libellés de variables et de modalités seront respectivement tronqués aux 60 et 20 premiers caractères.

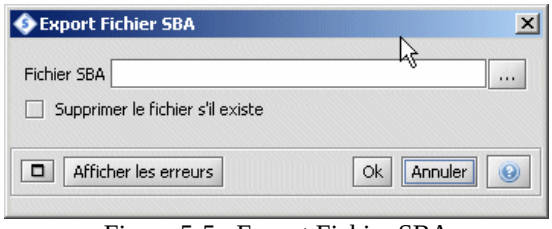


Figure 5-5 : Export Fichier SBA

Fichier SBA. Spécifiez le nom du fichier dans lequel les données seront exportées. Vous pouvez entrer le chemin et le nom du fichier ou cliquer sur le bouton « … » pour accéder à l'explorateur.

Supprimer le fichier s'il existe. Cochez cette case si vous souhaitez écraser un fichier existant.

Fichier SPSS

Cette méthode permet d'exporter les données d'une étape au format SPSS selon les spécifications décrites par SPSS dans la version 15. Néanmoins, les fichiers SPSS générés par cette méthode sont pour la plupart compatibles avec les versions antérieures.
La correspondance de types et rôles statistiques entre SPAD et SPSS est effectuée automatiquement.

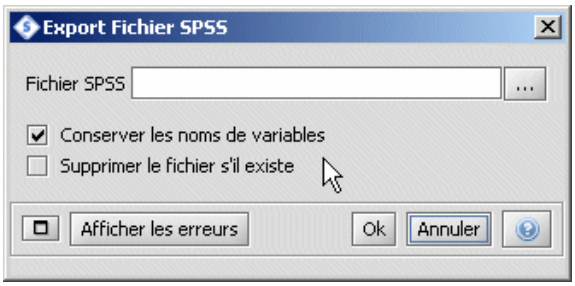


Figure 5-6 : Export Fichier SPSS

Fichier SPSS. Spécifiez le nom du fichier dans lequel les données seront exportées. Vous pouvez entrer le chemin et le nom du fichier ou cliquer sur le bouton « … » pour accéder à l'explorateur.

Conserver les noms de variable. Cochée par défaut, cette option signifie que les noms de variables SPAD seront conservés tels quels dans SPSS dans la mesure où ils respectent la nomenclature SPSS. Les noms « exotiques » seront automatiquement convertis afin de se conformer à la nomenclature.
Si l'option est décochée, les variables seront nommées VAR1, VAR2, etc. dans SPSS
Dans les deux cas, les étiquettes de variables SPSS sont basées sur les noms de variables SPAD.

Supprimer le fichier s'il existe. Cochez cette case si vous souhaitez écraser un fichier existant.

Création d'une nouvelle table dans une connexion BDD – Create Table

Cette méthode d'export permet soit d'ajouter une nouvelle table soit d'écraser une table existante dans une base de données, accessible par une connexion. Elle correspond à la commande SQL « CREATE°TABLE » couramment utilisée dans les bases de données.

Cette opération n'est possible que dans les connexions de type « Bases de données » ou « Sources ODBC ».

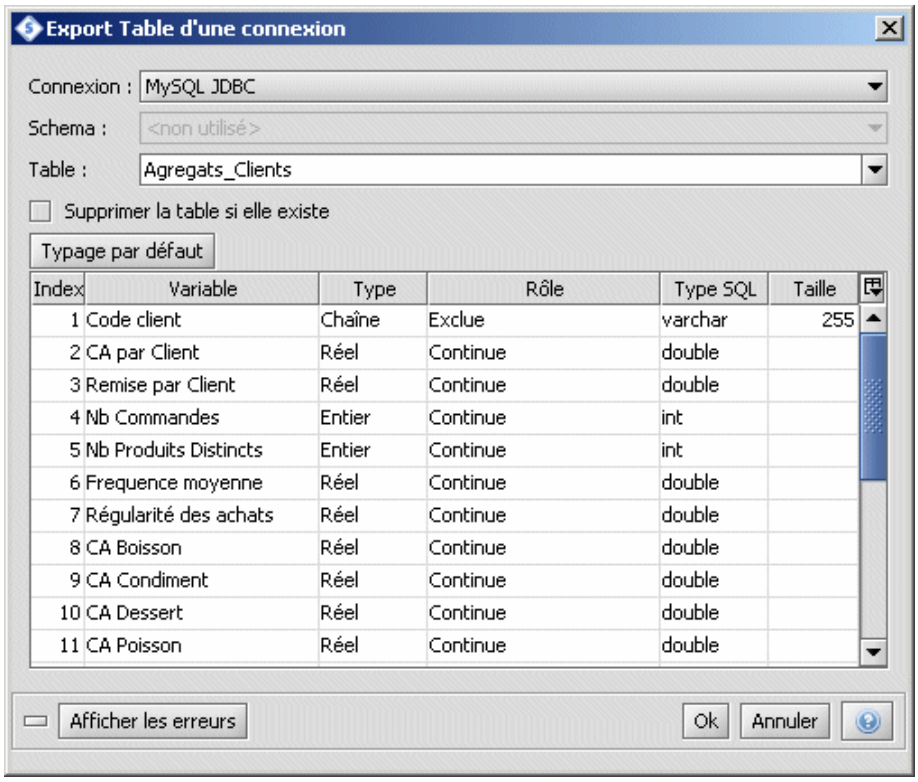


Figure 5-7 : Ajout d'une table dans une connexion

Connexion. Spécifiez la connexion à la base de données dans laquelle ajouter la nouvelle table.
Schéma. Spécifiez le schéma associé à la connexion si nécessaire.
Table. Sélectionnez la table à écraser ou entrez le nom de la nouvelle table à créer.

Supprimer la table si elle existe. Cochez cette case si vous souhaitez écraser une table déjà existante.

Typage par défaut. Ce bouton permet de revenir au paramétrage par défaut lorsque des modifications ont été apportées au « Type SQL » et à la « Taille » des variables.

Type SQL et Taille. Par défaut, SPAD assigne un type SQL et une taille (en nombre de caractères uniquement pour le Type SQL « String ») à chaque variable en fonction de son type de stockage SPAD. Vous pouvez modifier le type SQL et la taille en double cliquant dans les cases choisies.

Exécution. Cliquez sur OK pour valider le paramétrage de la méthode. A l'issue de l'exécution, la base de données dispose d'une nouvelle table.

Mise à jour dans une table d'une connexion BDD – Update Table

Cette méthode d'export permet de mettre à jour les données d'un ou plusieurs enregistrements d'une table appartenant à une connexion. Elle permet également l'ajout de nouvelles variables dans la table. Elle correspond à la commande SQL « UPDATE » couramment utilisée dans les bases de données.

Les mises à jour sont possibles dans les connexions de type « Bases de données » ou « Sources ODBC ».

Table source

Liste des variables disponibles dans les données sources pour la mise à jour.

Opérations

Spécifiez les opérations à réaliser lors de la mise à jour. Pour ajouter une opération, cliquez sur les boutons +.
Glissez-déposez respectivement dans les colonnes Entrée et Sortie les variables des Tables Source et Destination. Choisissez la nature de l'opération à réaliser parmi :

- la définition de la clé de jointure (=),
- la mise à jour d'une variable (▶)
- l'ajout d'une variable dans la table destination (✚)

Table destination

Liste des variables présentes dans la table destination.

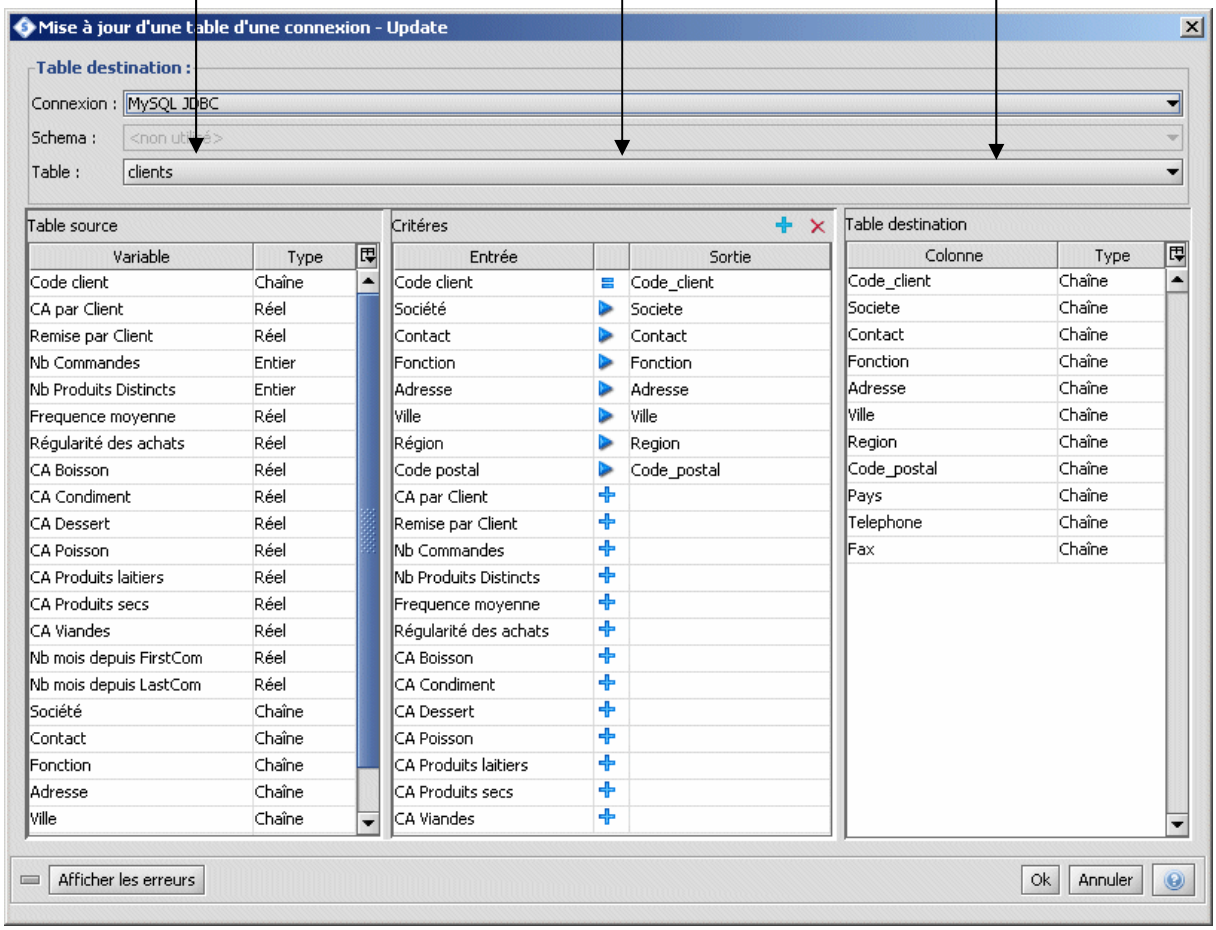


Figure 5-8 : Mise à jour d'une table d'une connexion

Connexion. Spécifiez la connexion contenant la base de données sur laquelle effectuer l'opération de mise à jour.
Schéma. Spécifiez le schéma associé à la connexion si nécessaire.
Table. Sélectionnez la table à mettre à jour.

Opérations. Spécifiez dans cette partie la liste des opérations de mise à jour à effectuer sur la table destination.
Trois types d'opérations sont permis :

- la définition de la clé de jointure : =
- la mise à jour d'une variable : ▶
- l'ajout d'une variable dans la table destination : ✚

La **définition de la clé de jointure** est indispensable pour identifier les correspondances des enregistrements entre la table source et la table destination. Vous pouvez spécifier une clé multiple.

La **mise à jour des valeurs d'une variable** symbolisée par ▶ indique que les valeurs de la variable Destination seront remplacées par les valeurs de la variable Source pour les enregistrements vérifiant la clé de jointure.

L'**ajout d'une variable source** symbolisée par ✚ indique que la table destination comptera une variable supplémentaire à l'issue de l'exécution. Tous les enregistrements de la table Destination présents dans la table Source selon la clé de jointure se verront affecter la valeur de la variable source. Les autres enregistrements de la table Destination se verront affecter la réponse manquante (Null en base de données).

Ajouter/Supprimer une opération. Cliquez respectivement sur les boutons ✚ et ✕ pour ajouter ou supprimer une opération. L'ajout d'une opération fait apparaître une nouvelle ligne vierge dans le tableau des opérations.

Paramétrer une opération. Le paramétrage d'une opération comporte trois étapes.

1. **Spécifiez la variable source** de l'opération en glissant / déposant la variable source dans la case de la colonne du même nom.
2. **Spécifiez le type** de l'opération en cliquant dans la case de la colonne Type jusqu'à faire apparaître le symbole choisi selon les explications ci-dessus.
3. **Spécifiez la variable destination** concernée par l'opération en glissant / déposant la variable destination dans la case de la colonne du même nom. Cette étape est facultative lorsque le type d'opération est un ajout de variable.

Exécution. Cliquez sur OK pour valider le paramétrage de la méthode. A l'issue de l'exécution, la table Destination de la connexion a été mise à jour. Cette opération est irréversible.

Ajout de données dans une table d'une connexion BDD – Insert Into Table

Cette méthode d'export permet d'ajouter de nouvelles lignes dans une table existante dans une base de données, accessible par une connexion.

Elle correspond à la commande SQL "**INSERT INTO TABLE**" couramment utilisée dans les bases de données.

Les paramètres de la méthode permettent de spécifier les correspondances entre les variables de la table Source et de la table Destination...

Table source

Liste des variables disponibles dans les données sources pour la mise à jour.

Opérations

Spécifiez les opérations à réaliser lors de l'insertion de ligne. Pour ajouter une opération, cliquez sur les boutons +.
Glissez-déposez respectivement dans les colonnes Variable source et Variable destination les correspondances entre variables des Tables Source et Destination. Choisissez la nature de l'opération à réaliser parmi :

- Correspondance entre variables pour l'ajout de lignes (=),
- Ajout d'une variable dans la table destination (✚)

Table destination

Liste des variables présentes dans la table destination.

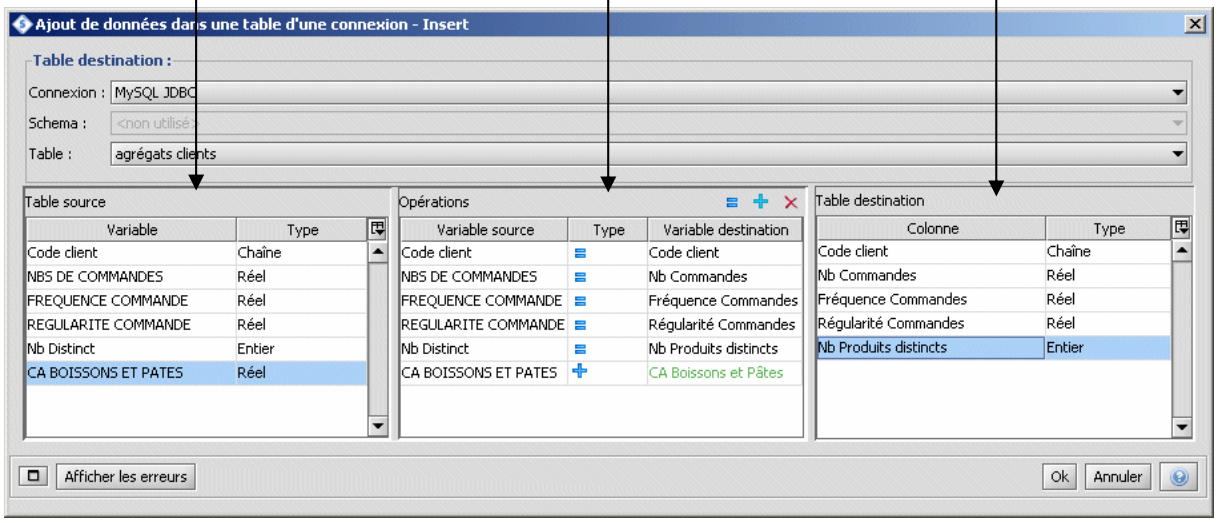


Figure 5-9 : Mise à jour d'une table d'une connexion

Connexion. Spécifiez la connexion contenant la base de données sur laquelle effectuer l'opération d'insertion de lignes.
Schéma. Spécifiez le schéma associé à la connexion si nécessaire.
Table. Sélectionnez la table dans laquelle insérer les lignes.

Opérations. Spécifiez dans cette partie la liste des correspondances entre variables « source » et variables « destination » ainsi que les nouvelles variables de la table source à créer dans la table destination.
Deux types d'opérations sont permis :

- la définition de correspondances entre variables source et destination : =
- l'ajout d'une variable dans la table destination : ✚

La **définition des correspondances** est indispensable pour savoir dans quelles variables de la table Destination écrire les nouvelles valeurs des données sources.

L'**ajout d'une variable source** symbolisée par ✚ indique que la table destination comptera une variable supplémentaire à l'issue de l'exécution. Tous les enregistrements de la table Destination présents avant l'opération d'insertion de lignes se verront affecter la réponse manquante (Null en base de données).

Ajouter/Supprimer une opération. Cliquez respectivement sur les boutons ✚ et ✕ pour ajouter ou supprimer une opération. L'ajout d'une opération fait apparaître une nouvelle ligne vierge dans le tableau des opérations.

Paramétrer une opération. Le paramétrage d'une opération comporte trois étapes.

1. **Spécifiez la variable source** de l'opération en glissant / déposant la variable source dans la case de la colonne du même nom.
2. **Le type d'opération est géré automatiquement par SPAD**.
3. **Spécifiez la variable destination** dans le cas d'une correspondance (Type devient **=**) ou double-cliquez dans la case pour spécifier le nom de la nouvelle variable à créer dans la table Destination (Type devient **+**). Le nom de la nouvelle variable est coloré en vert.

Exécution. Cliquez sur OK pour valider le paramétrage de la méthode. A l'issue de l'exécution, les lignes de la table Source ont été ajoutées à celles de la table Destination. Cette opération est irréversible.

Fichier Txt des coordonnées factorielles

Cette méthode vous permet d'exporter tout ou partie des coordonnées factorielles des individus et/ou des variables issus d'une analyse factorielle préalable sous forme d'un fichier texte.
Cette méthode doit être directement connectée à l'analyse factorielle.

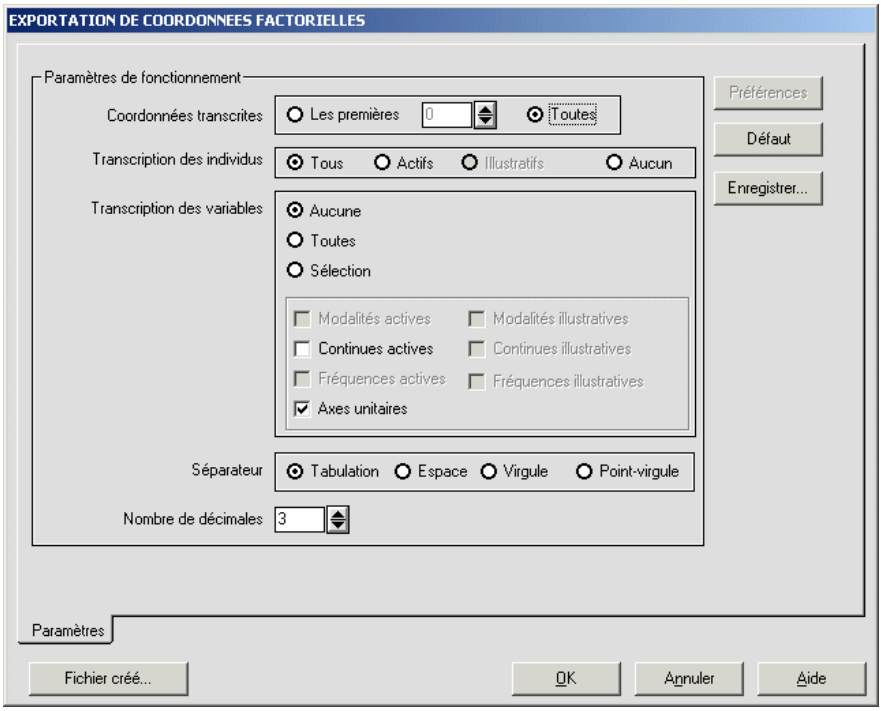


Figure 5-10 : Exportation Txt de coordonnées factorielles

Fichier créé. Spécifiez le nom et l'emplacement du fichier texte à créer en cliquant sur le bouton.

Coordonnées transcrites. Spécifiez le nombre de coordonnées à exporter dans le fichier texte. Elles le sont toutes par défaut.

Transcription des individus. Spécifiez les individus dont les coordonnées seront exportées.

Transcription des variables. Spécifiez les variables dont les coordonnées seront exportées. Le choix dépend de l'analyse factorielle préalable.

Séparateur de colonnes. Spécifiez le séparateur de colonnes à utiliser dans le fichier exporté.

Nombre de décimales. Précision en nombre de décimales des coordonnées factorielles exportées.

Chapitre

6

Data Management

Panorama général

Parmi les principales nouveautés de la version 7, SPAD fournit plusieurs méthodes de data management qui permettent de filtrer, nettoyer, agréger ses données et de créer de nouveaux indicateurs. Les méthodes de data management sont regroupées en sous-catégories :

- **Lignes - Individus**
  - Filtre logique, Echantillonnage par tirage aléatoire, stratifié ou par blocs,
  - Distinct – Dédoublonnage,
  - Tri,
  - Empilement de deux tables,
  - Agrégation,
  - Transposition, Verticalisation
- **Colonnes - Variables :**
  - Formats (normalisation, libellé, type de stockage, rôle statistique…),
  - Edition des libellés,
  - Sélection de colonnes,
  - Juxtaposition,
  - Jointure,
  - Binarisation,
  - Remplacement des données manquantes,
  - Mise en classes, Regroupement de modalités…,
  - Générateur de nouvelles variables,
  - Création d'une variable de pondération
- **Outils antérieurs V6**
  - Recodages - Escal, Croisements de variables
- **Jython**
  - Création de nouvelles variables en Jython,
  - Script Data Management Jython

La vue « Données » de SPAD permet à chaque étape d'un processus de data management de visualiser les données produites en sortie d'exécution. Pour visualiser les données produites par une méthode, il suffit de sélectionner la méthode.

Méthodes de data management pour les lignes

Les méthodes de data management pour les lignes sont accessibles dans la vue « Méthodes » :

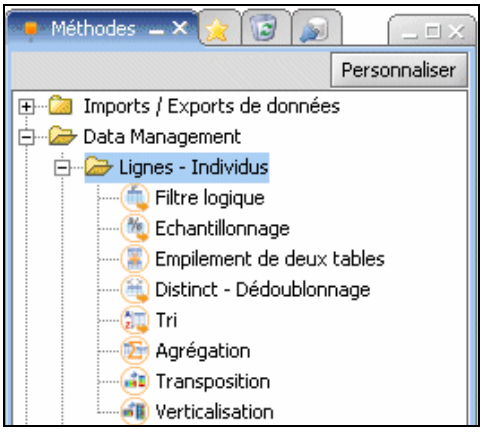


Figure 6-1 : Opérations de data management sur les lignes

Filtre logique

Cette méthode permet d'inclure ou d'exclure (Paramètre Méthode de sélection) des lignes (i.e. individus) par des conditions exprimées sur les variables. Elle ne duplique pas les données sources.

L'interface permet de définir un filtre élémentaire de la manière suivante :

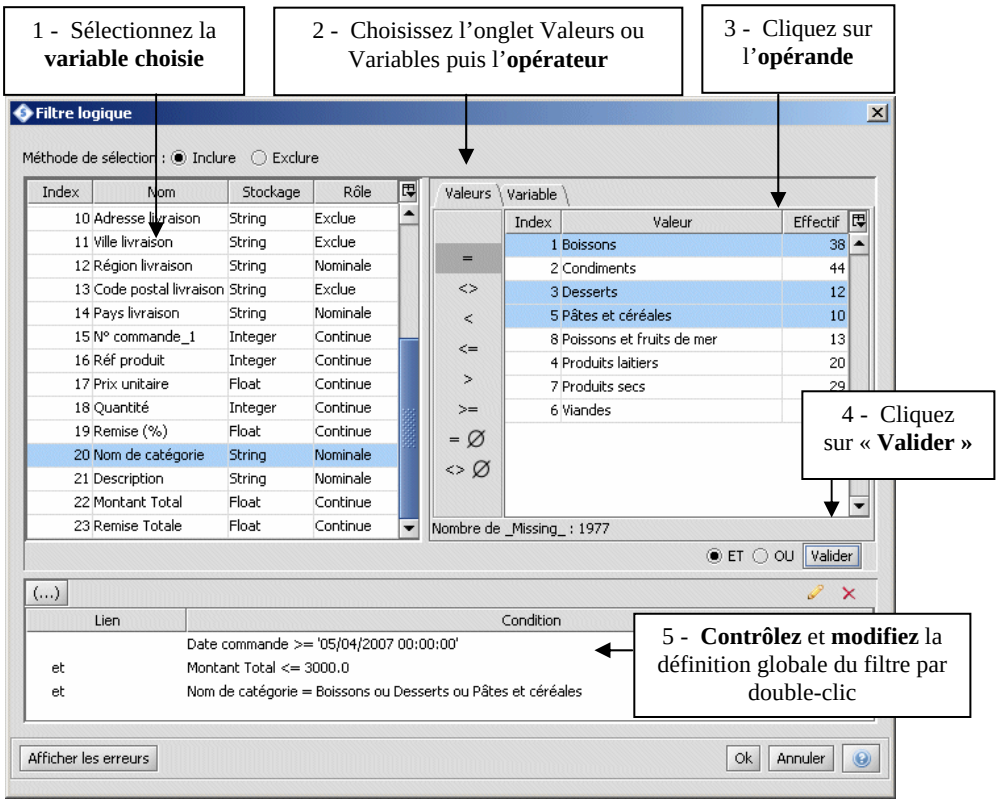


Figure 6-2 : Filtre logique

Après la définition du premier filtre élémentaire, les opérateurs **ET / OU** sont disponibles à gauche du bouton « Valider ». Continuez ainsi de suite pour définir l'ensemble des conditions du filtre global.

L'**étape 2** permet de **comparer la variable sélectionnée dans la liste de gauche** (Etape 1) **à ses propres valeurs ou de la comparer à une autre variable du fichier** en cliquant sur l'onglet variable. Il faut bien entendu comparer des variables « comparables » en termes de types de stockage.
Dans cette étape, le symbole **Ø symbolise les données manquantes**.

L'étape 3 diffère selon le stockage et le rôle statistique de la variable sélectionnée pour le filtre élémentaire. On sélectionnera la ou les modalités pour une variable nominale, une valeur pour une variable continue, un jour pour une variable date, un texte libre ou une expression régulière pour toute autre variable.

Les **textes libres et expressions régulières concernent les variables stockées sous forme de chaine de caractères (« Chaine ») et dont le rôle est exclu**. Dans ce cas là, l'interface pour l'étape 3 se présente sous la forme suivante :

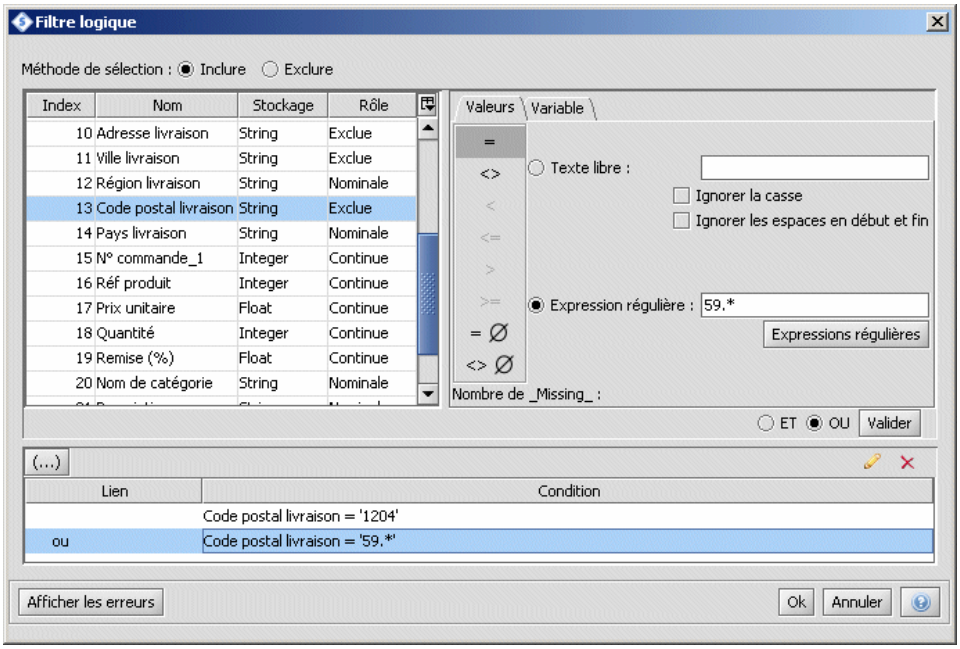


Figure 6-3 : Filtre logique – Textes libres et expressions régulières

Texte libre. La condition sera évaluée par rapport à la chaîne de caractères tapée dans la case correspondante.

Ignorer la casse. Cette option permet d'ignorer la mise en forme Majuscules/Minuscules pour les comparaisons.

Ignorer les espaces de début et de fin. Cette option permet de spécifier que les comparaisons doivent être faîtes indépendamment des espaces présents en début ou en fin de chaine.

Expression régulière. Spécifie que la valeur entrée dans la case correspondante sera considérée comme une expression régulière (voir définition page XXXX). Le bouton « Expressions régulières » permet de tester si l'expression régulière saisie par l'utilisateur a bien le comportement attendu. Dans l'exemple suivant, on constate que la comparaison entre la valeur « 59001 » et l'expression régulière « 59.* » correspond bien. Ce qui n'est pas le cas avec la valeur « 69001 ».

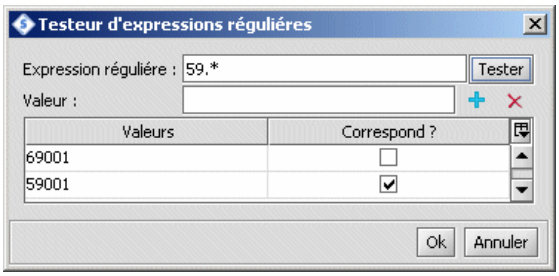


Figure 6-4 : Filtre logique – Test d'une expression régulière

L'étape **5** vous permet d'effectuer les opérations suivantes :

- **Supprimer** un filtre élémentaire : sélectionnez le filtre et cliquez sur le bouton
- **Editer** et **Modifier** un filtre élémentaire en double-cliquant sur le filtre concerné dans la colonne condition ou en cliquant sur le bouton
- **Modifier le lien ET/OU** en double-cliquant sur le lien concerné dans la colonne lien,
- **Ajouter une notion de « groupes de filtres »** en cliquant sur le bouton (...)

Echantillonnage

Cette méthode permet d'échantillonner les données, c'est-à-dire de réduire la taille de l'échantillon pour la suite des traitements.

Elle est particulièrement intéressante pour effectuer des analyses sur des fichiers de plus petites tailles (gain de temps) et aussi pour les techniques statistiques décisionnelles (Scoring, segmentation) nécessitant un échantillon équilibré par rapport à la variable à expliquer.

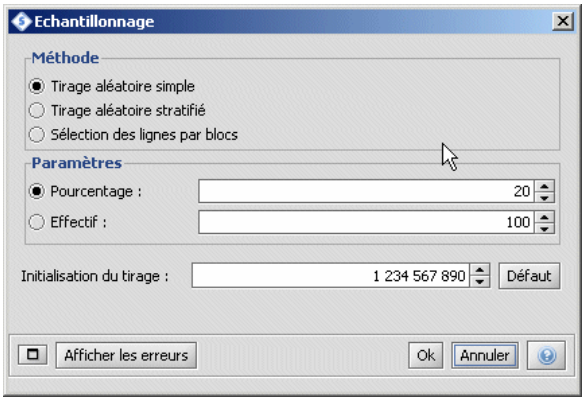


Figure 6-5 : Echantillonnage

Méthode. Choix de la méthode d'échantillonnage. Trois méthodes sont proposées :
Tirage aléatoire simple en % ou en effectif. Pour effectuer des analyses sur des échantillons de plus petites tailles. Ex. : 20% de l'échantillon total.

Tirage aléatoire stratifié, selon une variable de stratification. Pour effectuer des analyses sur des échantillons de plus petites tailles et rééquilibrer certaines sous-populations au sein du nouvel échantillon. Ex. : 90% des hommes et 60% des femmes.

Sélection des lignes par blocs. Pour sélectionner une ou plusieurs plages d'enregistrements. Ex. : de l'enregistrement n°1 au n°100.

Initialisation du tirage. Les méthodes de tirage aléatoire simple et stratifié font appel à un générateur de nombres pseudo-aléatoires qui, pour une initialisation donnée, fournit toujours la même séquence de nombres. Modifier cette valeur permet de modifier la séquence de nombres engendrés. Ceci permet notamment d'engendrer des sous-échantillons différents.

Paramètres de la méthode. Les paramètres de la méthode sont directement liés à la méthode choisie.
Tirage aléatoire simple
Spécifiez en pourcentage ou en effectif la taille de l'échantillon à conserver.

Tirage aléatoire stratifié
Cette méthode est particulièrement recommandée pour l'usage de méthodes décisionnelles sur échantillons déséquilibrés.
Exemple : Vous disposez d'une base contenant 5000 clients dont 500 constituent votre cible. Si vous effectuez un calcul de score ou une segmentation sur la variable nominale « Profil » à deux modalités (Cible / hors cible), les résultats seront mauvais en raison du déséquilibre des deux populations.
Il est préférable dans ce cas d'effectuer un tirage aléatoire stratifié par la variable « Profil » en retenant 100% de votre cible et 10 % des autres.

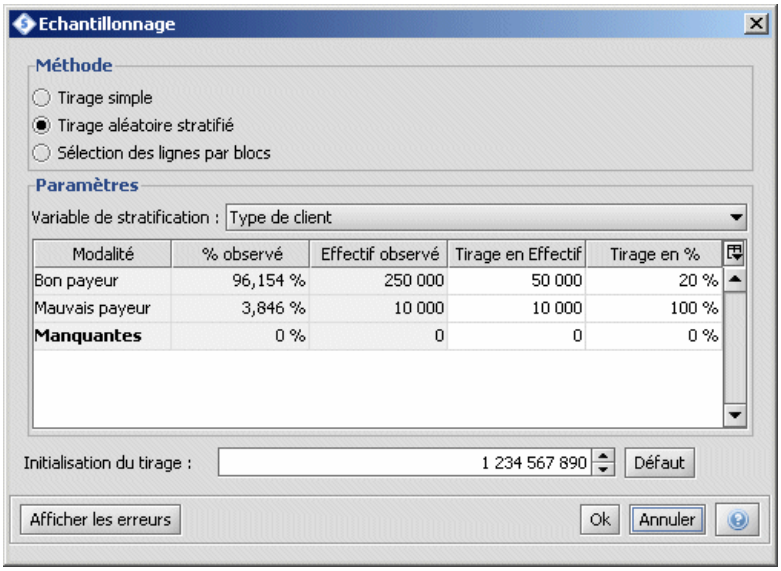


Figure 6-6 : Echantillonnage par tirage aléatoire stratifié

Variable de stratification. Sélectionnez dans la liste déroulante la variable de stratification. Seules les variables nominales sont disponibles.

Tirage en Effectif. Spécifiez le nombre d'individus à conserver dans chaque modalité de la variable de stratification, y compris pour les données manquantes.

Tirage en %. Spécifiez le pourcentage d'individus à conserver dans chaque modalité de la variable de stratification, y compris pour les données manquantes.

Ces deux modes de sélection Effectif / % peuvent être utilisés simultanément.

Sélection des lignes par blocs

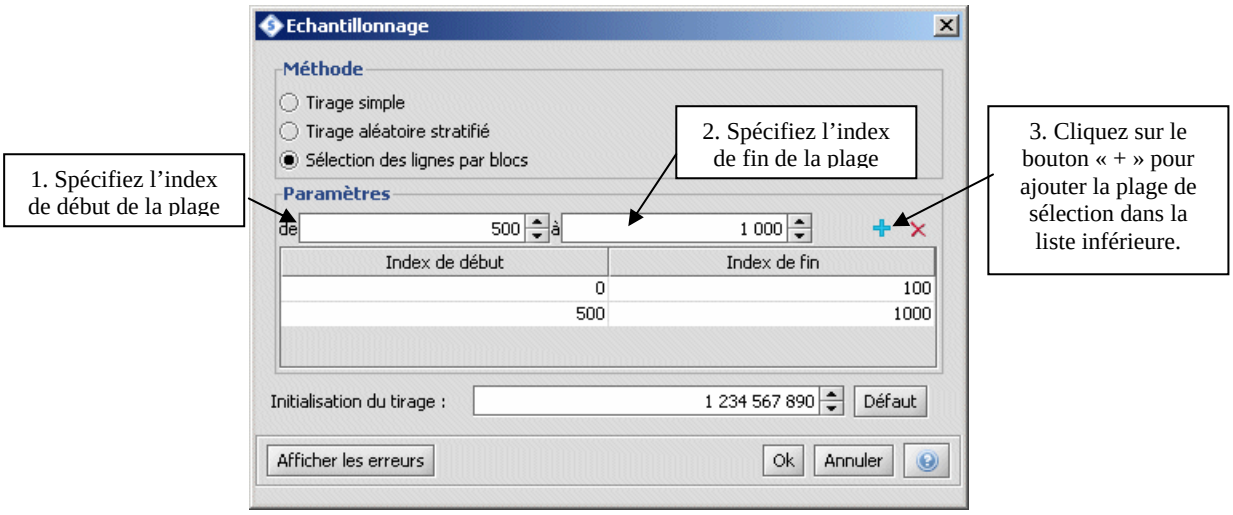


Figure 6-7 : Echantillonnage par blocs

Index de début. Numéro de la ligne définissant le début de la plage.
Index de fin. Numéro de la ligne définissant la fin de la plage.

Les index de début et de fin sont inclus dans la sélection.

Vous pouvez supprimer une ou plusieurs plages en les sélectionnant et en cliquant sur le bouton 

Distinct – Dédoublonnage

Cette méthode peut être utilisée pour supprimer des doublons présents dans les données en ne conservant que le premier individu distinct. Cette opération est utile lorsque l'on veut conserver une seule occurrence de chaque item dans les données.
Cette méthode correspond à l'instruction « Distinct » fréquemment utilisée en langage SQL dans les bases de données.

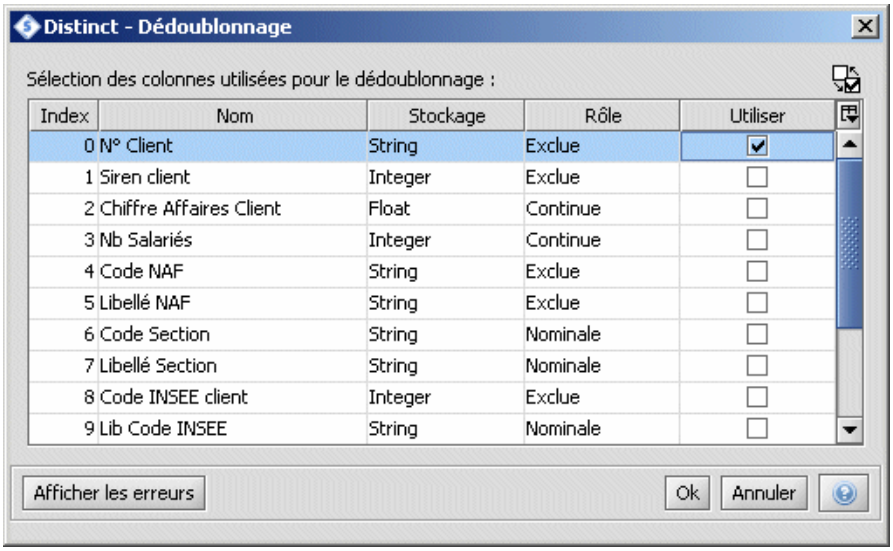


Figure 6-8 : Distinct - Dédoublonnage

Utiliser. Spécifiez la ou les colonnes à utiliser pour la recherche de doublons.

Tri

Cette méthode permet de trier les individus d'une table selon un ou plusieurs critères, en spécifiant pour chaque critère l'ordre Croissant / Décroissant.

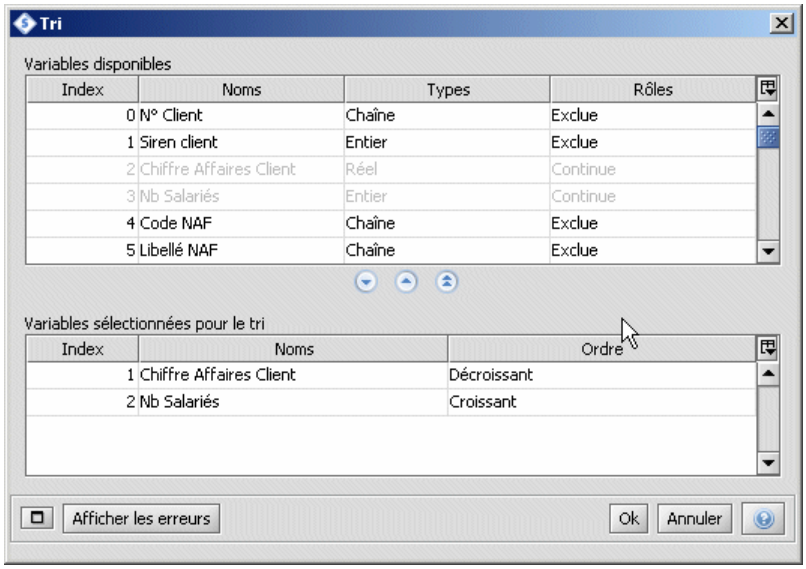


Figure 6-9 : Tri des lignes

1. **Sélectionnez la ou les variables disponibles** dans la liste des variables disponibles,
2. **Validez** votre choix en cliquant sur le bouton .
3. Les variables sélectionnées apparaissent grisées dans la liste supérieure et figurent maintenant dans la liste des variables sélectionnées pour le tri.
4. Pour chaque variable sélectionnée, **choisissez l'ordre de tri Croissant / Décroissant** en cliquant dans la case correspondante.

Pour désélectionner une variable, sélectionnez-la et cliquez sur le bouton . Pour les désélectionner toutes, cliquez sur le bouton .

Ordre de tri. Lorsque plusieurs variables sont sélectionnées pour le tri, SPAD trie d'abord les individus par la première variable de la liste, puis par la seconde…
Dans l'exemple ci-dessus, les lignes sont d'abord triées par Chiffre d'affaires décroissant. Ensuite, à chiffre d'affaires égal, on triera les lignes dans l'ordre croissant du nombre de salariés.

Modifier l'ordre des variables sélectionnées pour le tri. Pour modifier le positionnement d'une variable dans la liste, sélectionnez-la puis glissez/déposez la variable à l'emplacement choisi, matérialisé par une ligne bleue.

Remarque. **Les données triées selon une variable nominale sont triées par groupes de valeurs (i.e. modalités)**. **L'ordre des modalités est arbitraire** et ne respecte pas l'ordre alphabétique (sauf hasard) en raison du mode de stockage interne de ce type de variables pour optimiser l'espace disque.

Empilement de deux tables

Cette méthode réalise l'empilement de deux tables de données.

Les deux tables n'ont pas nécessairement les mêmes variables.
La méthode détecte automatiquement les équivalences de variables (noms identiques) par un prétraitement automatique.

L'utilisateur peut intervenir manuellement pour :

- définir les équivalences,
- sélectionner tout ou partie des variables à inclure dans la table de sortie,
- modifier leurs attributs (Nom et rôle statistique).

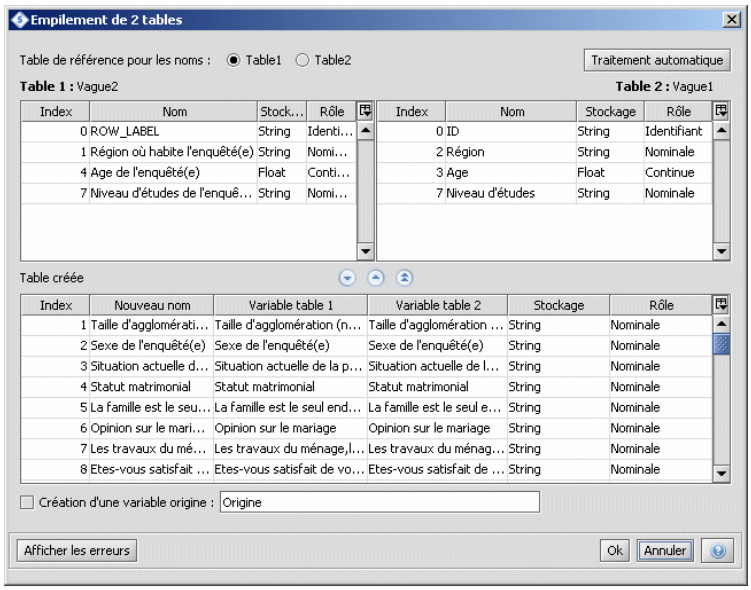


Figure 6-10 : Empilement de deux tables

Table de référence pour les noms. Choix de la table de référence pour les noms de variables qui seront utilisés dans la table créée.

Traitement automatique. Par défaut, SPAD détecte automatiquement les équivalences de variables selon leur nom ou libellé. Les variables concernées sont listée dans la partie inférieure de la fenêtre comme nouvelles variables de la table créée.
La méthode propose 4 prétraitements automatiques :

- **Variables communes** (choix par défaut)
- **Variables communes + variables Table 1** : la table empilée contiendra toutes les variables communes ainsi que les autres variables de la table 1. Ces variables ne seront pas renseignées pour les individus de la table 2.
- **Variables communes + variables Table 2** : la table empilée contiendra toutes les variables communes ainsi que les autres variables de la table 2. Ces variables ne seront pas renseignées pour les individus de la table 1.
- **Toutes les variables** : toutes les variables communes et toutes les autres variables des deux tables. Dans ce cas, pour les individus issus du la table 1 (réciproquement 2), les « autres variables » issues de la table 2 (réciproquement 1) ne seront pas renseignées.

Création d'une variable origine. Non cochée par défaut, cette option permet de garder une trace de la provenance de chaque individu en créant une nouvelle variable nominale nommée « Origine » par défaut à deux modalités du nom des tables 1 et 2.

Définir une équivalence manuelle entre deux variables.

1. **Sélectionnez** la ou les **variables** dans **la Table 1** (usage standard de la souris et des touches),
2. **puis** les variables équivalentes (en nombre égal) dans la **Table 2**,
3. **Validez** vos équivalences en cliquant sur le bouton de transfert .
4. Les équivalences apparaissent dans la liste "Table créée" et les variables concernées disparaissent de leurs tables d'origine.
5. **Procédez ainsi de suite**.

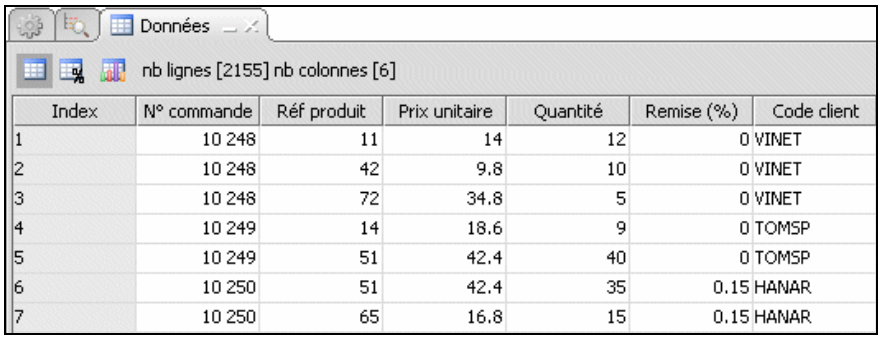
Pour désélectionner une équivalence, sélectionnez-la dans la liste « Table créée » et cliquez sur le bouton . Pour les désélectionner toutes, cliquez sur le bouton .

Remarque : Lorsque les types de stockage ou rôles statistiques diffèrent, SPAD effectue le meilleur compromis pour la variable créée. Néanmoins, vous pouvez modifier le rôle statistique de la variable dans la liste « Table créée » en double-cliquant sur la case de la variable concernée dans la colonne « Rôle ».

Agrégation

Cette méthode permet d'agréger les données disponibles selon une clé simple ou multiple et de calculer de nouveaux agrégats à partir des variables disponibles et des fonctions de la bibliothèque.

Exemple : on dispose d'une table dite transactionnelle contenant l'ensemble des commandes des clients, éclatées par produit. Chaque ligne de la table correspond donc à une commande, un produit et un client donné comme dans l'extrait ci-dessous, issu de la base de données « Comptoir.mdb » :


Données

nb lignes [2155] nb colonnes [6]

Index	N° commande	Réf produit	Prix unitaire	Quantité	Remise (%)	Code client
1	10 248	11	14	12	0	VINET
2	10 248	42	9.8	10	0	VINET
3	10 248	72	34.8	5	0	VINET
4	10 249	14	18.6	9	0	TOMSP
5	10 249	51	42.4	40	0	TOMSP
6	10 250	51	42.4	35	0.15	HANAR
7	10 250	65	16.8	15	0.15	HANAR

Figure 6-11 : Données avant agrégation

L'agrégation de cette table sur la **clé "Code client"** nous permettra de calculer pour chaque client les agrégats suivants :

- Nombre de commandes par client,
- Nombres d'articles différents achetés par client,
- Montant total des achats par client,
- Montant total des remises accordées par client,
- Ratio Remise/CA par client…

Principes de fonctionnement

L'agrégation se fait en 3 étapes successives :

1. **Pré-processing**
   Optionnelle, cette étape permet de générer de nouvelles variables à partir des variables existantes en utilisant toutes les fonctions et outils de la bibliothèque hormis celles de la rubrique « Agrégation ».
   Dans cet exemple, on calculera lors de cette phase le « CA par ligne », la « Remise par ligne ».
   Les fonctions de pré-processing doivent être positionnées au dessus des fonctions d'agrégation.

2. **Agrégation**
   La phase d'agrégation consiste à calculer de nouvelles variables, agrégées selon la clé de regroupement, à partir des données disponibles à l'issue du pré-processing ou des données sources si le pré-processing n'a pas eu lieu.
   Les seules fonctions disponibles pour cette étape sont celles de la rubrique « Agrégation ».
   Toutes les fonctions de cette rubrique sont précédées du préfixe « AG_ » pour marquer leur différence avec les autres fonctions.
   Dans cet exemple, nous allons calculer, pour chaque client, le nombre de commandes, le nombre d'articles différents achetés, le CA total et la remise totale.
   A l'issue de cette étape, les données disponibles comptent autant de lignes qu'il y a de groupes distincts (i.e. clés d'agrégation distinctes).

3. Post-processing

La phase de post-processing permet de calculer de nouveaux indicateurs à partir des données produites par l'étape d'agrégation. Toutes les fonctions de la bibliothèque sont disponibles hormis celles de la rubrique « Agrégation ». Elles n'acceptent que les variables créées par l'étape d'agrégation.

Dans cet exemple, nous allons calculer deux ratios : le taux de remise par client et le montant moyen des commandes par client.

Figure 6-12 : Interface de la méthode d'agrégation

Définir la clé d'agrégation.

La clé simple ou multiple doit être spécifiée dans la liste des « Variables initiales et créées ». Pour cela, il suffit de cliquer dans la colonne « clé/créer » sur la ou les variables initiales concernées. Pour désélectionner une variable utilisée comme clé, cliquez à nouveau.

Fonctions – Aide

Les fonctions sont classées par grandes rubriques dans la partie supérieure gauche de la fenêtre. Pour obtenir de l'aide sur une fonction, il suffit de la sélectionner. La zone de description se met automatiquement à jour. Pour plus d'informations, voir le descriptif complet de la bibliothèque de fonctions page 118.

Ajouter une fonction.
L'ajout d'une fonction dans la liste des transformations successives se fait soit par double-clic sur la fonction soit à l'aide du bouton  .

Comportement du bouton 
Ce bouton permet d'ajouter une ou plusieurs fonctions (usage standard des touches Ctrl Shift) et de renseigner leur premier paramètre si des variables initiales et/ou créées ont été également sélectionnées.
Le comportement diffère selon que la fonction accepte en premier paramètre une liste de valeurs ou non.
La fonction apparaît immédiatement dans la liste des transformations successives. La nouvelle variable, associée à cette fonction, apparaît dans la liste des variables initiales et créées. Son nom a été automatiquement généré et est affiché dans la colonne « Résultats » du tableau des transformations.

Lorsque la création d'une nouvelle variable nécessite l'utilisation successive de plusieurs fonctions, procédez par étape en appliquant la nouvelle fonction au résultat de l'étape précédente.
Le tableau suivant résume les différentes possibilités proposées par ce bouton :

	Type du premier paramètre	P variables sélectionnées
N fonctions sélectionnées	Valeur	N * P
N fonctions sélectionnées	[Liste de valeurs]	N

Dans le cas où on sélectionne P variables et N fonctions admettant comme premier paramètre une valeur et non une liste de valeurs, N * P transformations sont alors ajoutées dans la liste des transformations successives.

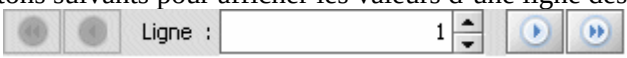
Insérer une fonction.
Pour insérer une fonction à un endroit précis, sélectionnez-la et glissez/déposez à l'endroit choisi.

Paramétrer une fonction.
Les paramètres de chaque fonction sont décrits dans la colonne « Paramètres » du tableau.
Le paramétrage se fait soit par glisser/déposer soit par double-clic dans la case de la colonne « Valeurs ».
Lorsque le paramètre attend une ou plusieurs variables (initiales et/ou créées), sélectionnez la ou les variables concernées et glissez/déposez la sélection dans la case de la colonne « Valeurs » correspondante ou double-cliquez dans la case de la colonne valeur pour accéder à une fenêtre de sélection de variables.
Avant le paramétrage, la case de la colonne valeur indique le type de paramètre attendu (Déposer Réel, Déposer Entier, Déposer Chaine…) et la nature du paramètre (valeur unique ou liste de valeurs).
La présence de crochets dans l'instruction « Déposer [Réel, …] » indique que le paramètre attend une liste de valeurs ou variables.
Le tableau suivant résume les correspondances de types acceptées :

Types acceptés / Type attendu	Entier	Réel	Chaine	Date	Autre
Déposer Entier	OK	⊘	⊘	⊘	⊘
Déposer Réel	OK	OK	⊘	⊘	⊘
Déposer Chaine	OK	OK	OK	OK	OK
Déposer Date	⊘	⊘	⊘	OK	⊘
Déposer	OK	OK	OK	OK	OK

Les outils « Calculatrice » et « Assistant logique » sont présentés respectivement pages 130 et 131.

Tester une fonction.
Cette fonctionnalité n'est active que pour les calculs de pré-processing.
Cliquez sur l'un des boutons suivants pour afficher les valeurs d'une ligne des données sources.



Les valeurs apparaissent dans la colonne « Valeurs » de la liste « Variables initiales et créées ».Pour tester l'ensemble des calculs et fonctions de pré-processing utilisées, cliquez sur le bouton **Tester**.
Les valeurs-résultats apparaissent dans les cases des variables créées.

Déplacer une fonction.
Sélectionnez la première ligne de la fonction dans la liste des transformations successives et glissez/déposez la fonction à l'endroit choisi, matérialisé par une ligne bleue.

Supprimer une fonction.
Sélectionnez la fonction dans la liste des transformations successives et cliquez sur le bouton .

Renommer une variable créée.
Double-cliquez sur le nom de la variable concernée soit dans la liste « Variables initiales et créées » soit dans la colonne « Résultats » de la « liste des transformations successives» et tapez le nouveau nom.

Supprimer une variable créée.
Sélectionnez la variable dans la liste « Variables initiales et créées » et cliquez sur le bouton .

Attention : renommer ou supprimer une variable créée entraîne des erreurs si cette variable est utilisée dans d'autres fonctions. La fonction liée à la variable supprimée n'est pas supprimée.

Modifier le stockage et le rôle d'une variable créée.
Par défaut, SPAD affecte un type de stockage et un rôle à toute variable créée. Pour les modifier, double-cliquez sur le stockage ou sur le rôle de la variable concernée dans la liste « Variables initiales et créées » et choisissez dans la liste déroulante.

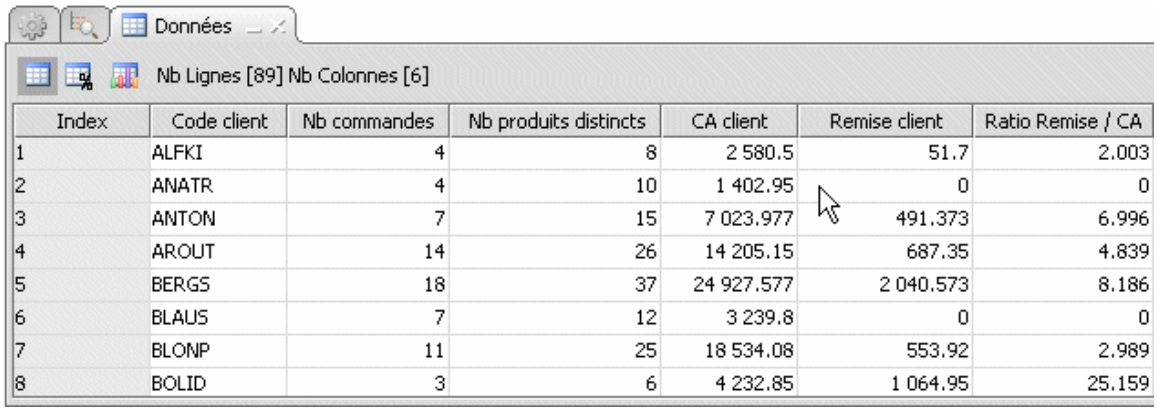
Définir une variable créée comme temporaire.
Si certaines variables, utilisées comme étapes intermédiaires au calcul, n'ont pas d'intérêt pour la suite de l'analyse, décochez la case à cocher de la colonne « Clé/Créer ? » dans la vue « Variables initiales et créées » pour qu'elles n'apparaissent pas dans les données produites.
Seules les variables basées des étapes Agrégation et Post-processing sont concernées.

Conserver la première valeur d'une variable d'origine pour chaque clé de regroupement.
Pour conserver dans la table de données en sortie la première valeur d'une variable d'origine pour chaque clé de regroupement, cliquez deux fois dans la case de la colonne « Clé/Créer » de la variable d'origine pour faire apparaître « 1st ».

Obligation de spécifier une clé.
En décochant cette option, l'agrégation est effectuée sur les données sources en considérant qu'il n'y a qu'un seul groupe. La table produite ne comporte qu'une ligne.

Ordonner les modalités automatiquement selon leur ordre naturel.
Cochée, cette option trie les modalités d'une nouvelle variable nominale dans l'ordre alphanumérique ou numérique selon qu'elle est stockée comme chaine de caractères ou comme entier. Cet ordre est repris dans les méthodes statistiques pour l'affichage des résultats. En décochant cette option, l'ordre obtenu est celui dans lequel les modalités apparaissent dans les données.

Résultats de l'exemple
Voici un extrait issu de l'exécution du paramétrage présenté Figure 6-12 :


Données
Nb Lignes [89] Nb Colonnes [6]

Index	Code client	Nb commandes	Nb produits distincts	CA client	Remise client	Ratio Remise / CA
1	ALFKI	4	8	2 580.5	51.7	2.003
2	ANATR	4	10	1 402.95	0	0
3	ANTON	7	15	7 023.977	491.373	6.996
4	AROUT	14	26	14 205.15	687.35	4.839
5	BERGS	18	37	24 927.577	2 040.573	8.186
6	BLAUS	7	12	3 239.8	0	0
7	BLONP	11	25	18 534.08	553.92	2.989
8	BOLID	3	6	4 232.85	1 064.95	25.159

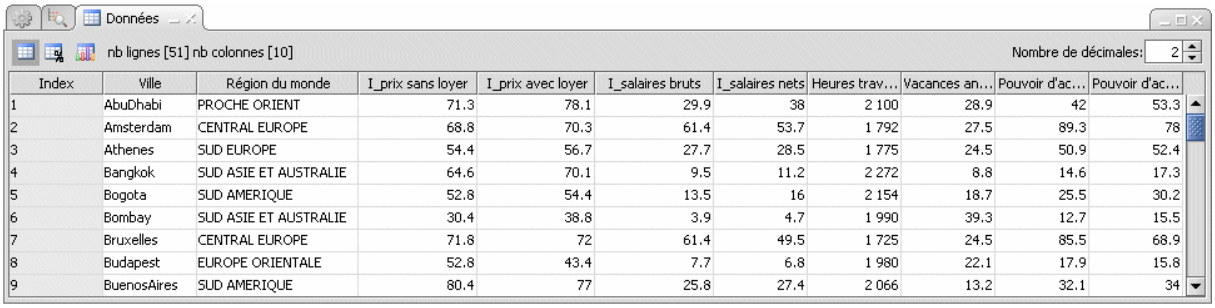
Figure 6-13 : Table produite par la méthode Agrégation

La table initiale comportait 2155 lignes. Après agrégation sur le Code Client, nous obtenons une table de 89 lignes ou 89 clients distincts.
Le client « ALFKI » a effectué 4 commandes, basée sur 8 produits distincts, pour un chiffre d'affaire total de 2 580,5 Euros...

Transposition

Cette méthode permet de transposer les variables numériques d'une table. Les autres types de variables sont ignorés par la méthode. Les variables numériques sélectionnées constituent une plage de données « verticale » qui est renvoyée en plage horizontale par la méthode « Transposition ».

Exemple : on dispose d'une table de données avec des villes du monde en ligne et en colonne un certain nombre d'indices socio-économiques.


Données
nb lignes [51] nb colonnes [10]
Nombre de décimales: 2

Index	Ville	Région du monde	I_prix sans loyer	I_prix avec loyer	I_salaires bruts	I_salaires nets	Heures trav...	Vacances an...	Pouvoir d'ac...	Pouvoir d'ac...
1	AbuDhabi	PROCHE ORIENT	71.3	78.1	29.9	38	2 100	28.9	42	53.3
2	Amsterdam	CENTRAL EUROPE	68.8	70.3	61.4	53.7	1 792	27.5	89.3	78
3	Athenes	SUD EUROPE	54.4	56.7	27.7	28.5	1 775	24.5	50.9	52.4
4	Bangkok	SUD ASIE ET AUSTRALIE	64.6	70.1	9.5	11.2	2 272	8.8	14.6	17.3
5	Bogota	SUD AMERIQUE	52.8	54.4	13.5	16	2 154	18.7	25.5	30.2
6	Bombay	SUD ASIE ET AUSTRALIE	30.4	38.8	3.9	4.7	1 990	39.3	12.7	15.5
7	Bruxelles	CENTRAL EUROPE	71.8	72	61.4	49.5	1 725	24.5	85.5	68.9
8	Budapest	EUROPE ORIENTALE	52.8	43.4	7.7	6.8	1 980	22.1	17.9	15.8
9	BuenosAires	SUD AMERIQUE	80.4	77	25.8	27.4	2 066	13.2	32.1	34

Figure 6-14 : Table de données avant transposition

L'interface de la transposition se présente sous cette forme.

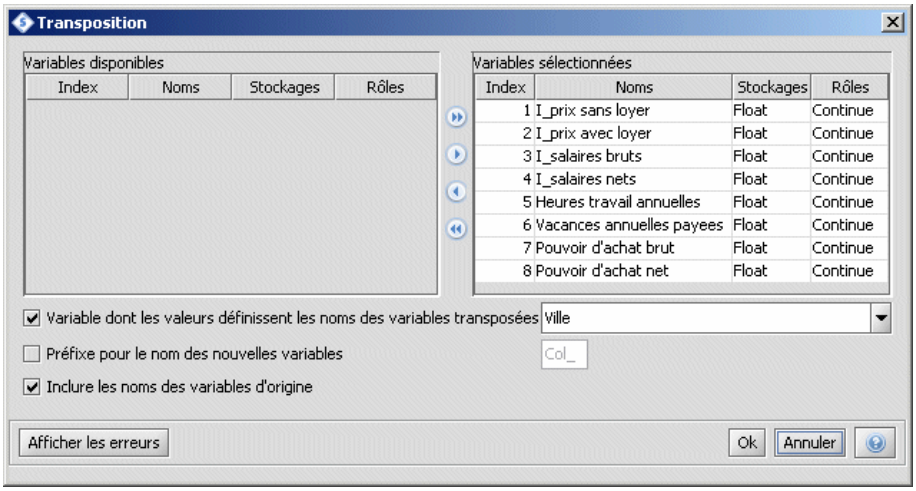


Figure 6-15 : Interface de la méthode Transposition

Variables disponibles. Liste des variables disponibles pour la transposition. Par défaut, toutes les variables numériques sont automatiquement sélectionnées dans la zone de droite. Les variables d'autres types n'apparaissent pas comme disponibles.

Variables sélectionnées. Liste des variables qui seront transposées.

Sélection et transfert de variables. Dans les deux listes, l'usage des touches Shift et Ctrl est standard pour sélectionner les variables. Le double clic sur une variable transfère la variable d'une liste à l'autre.

Transfère toutes les variables disponibles dans la liste « Variables binarisées ».

Transfère les variables disponibles sélectionnées dans la liste « Variables binarisées ».

Transfère les variables binarisées sélectionnées dans la liste « Variables disponibles ».

Transfère toutes les variables binarisées dans la liste « Variables disponibles ».

Variable dont les valeurs définissent les noms des variables transposées. Décochées par défaut, cette option permet de choisir une variable qui sera utilisée pour définir les noms des nouvelles variables obtenues. Tout type de variable peut être sélectionné.

Préfixe pour le nom des nouvelles variables. Cochée par défaut, cette option ajoute le préfixe spécifié (Col_ par défaut) au nom des nouvelles variables générées. Dans le cas ou l'option précédente est décochée, SPAD ajoute un numéro unique (Col_1, Col_2…) pour assurer l'unicité des noms de variables.

Inclure les noms des variables d'origine. Cochée par défaut, cette option permet de transposer les noms des variables d'origine sous forme d'une nouvelle variable dans la table de données transposée. Cette nouvelle variable est automatiquement nommée « ID ».

Résultats de l'exemple

Voici un extrait issu du paramétrage présenté Figure 6-12

Données

nb lignes [8] nb colonnes [52] Nombre de décimales: 2

Index	ID	AbuDhabi	Amsterdam	Athenes	Bangkok	Bogota	Bombay	Bruxelles	Budapest	BuenosAires	Caracas
1	I_prix sans loyer	71.3	68.8	54.4	64.6	52.8	30.4	71.8	52.8	80.4	40.
2	I_prix avec loyer	78.1	70.3	56.7	70.1	54.4	38.8	72	43.4	77	39.
3	I_salaires bruts	29.9	61.4	27.7	9.5	13.5	3.9	61.4	7.7	25.8	4.
4	I_salaires nets	38	53.7	28.5	11.2	16	4.7	49.5	6.8	27.4	5.
5	Heures travail annuelles	2 100	1 792	1 775	2 272	2 154	1 990	1 725	1 980	2 066	1 91
6	Vacances annuelles payees	28.9	27.5	24.5	8.8	18.7	39.3	24.5	22.1	13.2	15.
7	Pouvoir d'achat brut	42	89.3	50.9	14.6	25.5	12.7	85.5	17.9	32.1	12.
8	Pouvoir d'achat net	53.3	78	52.4	17.3	30.2	15.5	68.9	15.8	34	14.

Figure 6-16 : Table produite par la méthode Transposition

Les villes représentent maintenant les nouvelles variables de la table. Les noms de variables de la table d'origine représentent ici l'identificateur des lignes. Les données numériques initialement verticales sont maintenant horizontales.

Verticalisation

Cette méthode permet de verticaliser une table de données. Cette transformation de données ré-agence les données d'origine dans une autre disposition sans les modifier.
Les variables d'origine peuvent être conservées, abandonnées ou verticalisées.
Pour un tableau initial composé de n lignes, p variables à conserver et p' variables à verticaliser :

- Chaque ligne d'origine est éclatée en p' lignes comprenant les variables suivantes :
  - les p variables conservées
  - une colonne intitulée « Variable » par défaut, comprenant le nom de la variable verticalisée,
  - une colonne intitulée « Valeur » par défaut, comprenant la valeur correspondant à la variable verticalisée.
- Le tableau final comprend n*p' lignes et (p + p') colonnes.

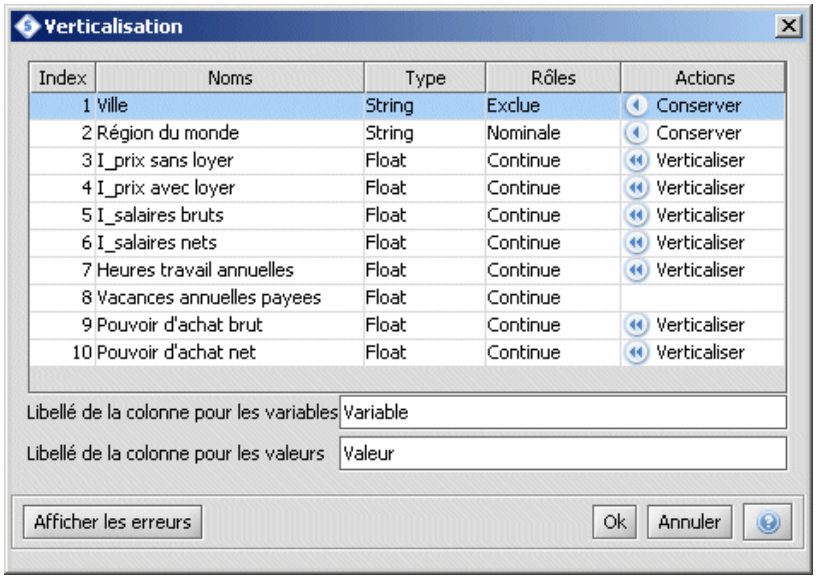


Figure 6-17 : Interface de la verticalisation

Action. La colonne « Action » permet de définir par clics successifs dans la case l'action à effectuer pour chaque variable. Trois actions sont possibles :

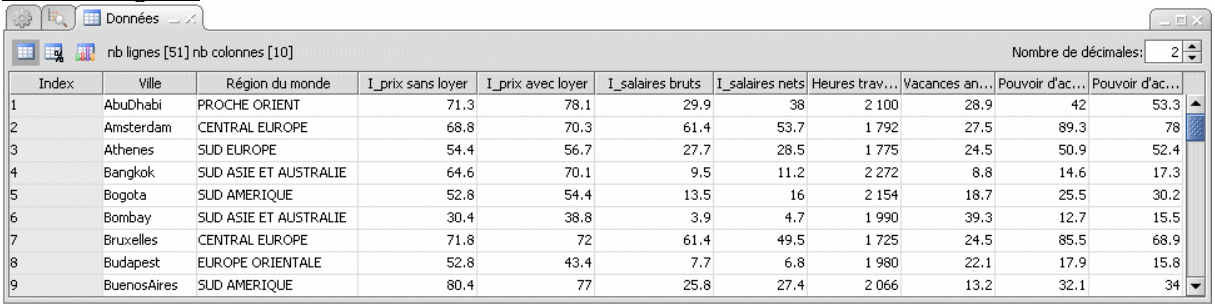
- Conserver la variable d'origine telle quelle
- Verticaliser la variable d'origine
- Abandonner la variable (Case vide)

Libellé de la colonne pour les variables. Libellé de la colonne dans la table de données produite conservant les noms des variables verticalisées.

Libellé de la colonne pour les valeurs. Libellé de la colonne dans la table de données produite conservant les valeurs des variables verticalisées.

Le paramétrage de la Figure 6-17, avec 2 variables à conserver (Ville, Région du monde) et 7 variables à verticaliser produit les résultats suivants :

Table d'origine :

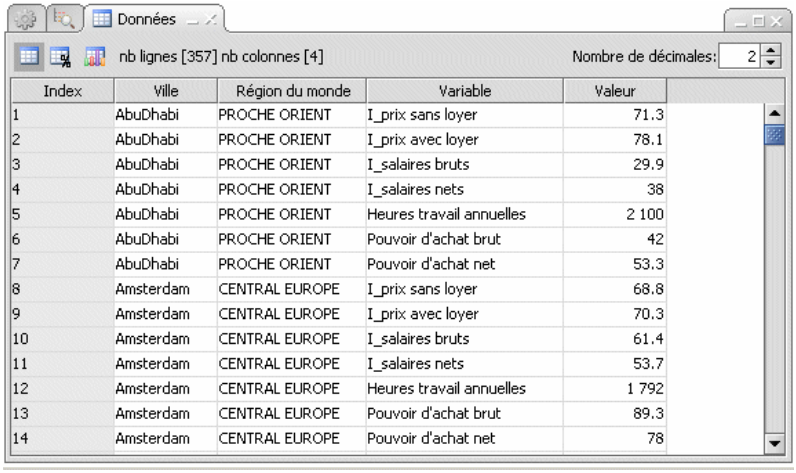

Données

nb lignes [51] nb colonnes [10] Nombre de décimales: 2

Index	Ville	Région du monde	I_prix sans loyer	I_prix avec loyer	I_salaires bruts	I_salaires nets	Heures trav...	Vacances an...	Pouvoir d'ac...	Pouvoir d'ac...
1	AbuDhabi	PROCHE ORIENT	71.3	78.1	29.9	38	2 100	28.9	42	53.3
2	Amsterdam	CENTRAL EUROPE	68.8	70.3	61.4	53.7	1 792	27.5	89.3	78
3	Athenes	SUD EUROPE	54.4	56.7	27.7	28.5	1 775	24.5	50.9	52.4
4	Bangkok	SUD ASIE ET AUSTRALIE	64.6	70.1	9.5	11.2	2 272	8.8	14.6	17.3
5	Bogota	SUD AMERIQUE	52.8	54.4	13.5	16	2 154	18.7	25.5	30.2
6	Bombay	SUD ASIE ET AUSTRALIE	30.4	38.8	3.9	4.7	1 990	39.3	12.7	15.5
7	Bruxelles	CENTRAL EUROPE	71.8	72	61.4	49.5	1 725	24.5	85.5	68.9
8	Budapest	EUROPE ORIENTALE	52.8	43.4	7.7	6.8	1 980	22.1	17.9	15.8
9	BuenosAires	SUD AMERIQUE	80.4	77	25.8	27.4	2 066	13.2	32.1	34

Figure 6-18 : Table d'origine avant verticalisation

Table produite par la verticalisation :


Données

nb lignes [357] nb colonnes [4] Nombre de décimales: 2

Index	Ville	Région du monde	Variable	Valeur
1	AbuDhabi	PROCHE ORIENT	I_prix sans loyer	71.3
2	AbuDhabi	PROCHE ORIENT	I_prix avec loyer	78.1
3	AbuDhabi	PROCHE ORIENT	I_salaires bruts	29.9
4	AbuDhabi	PROCHE ORIENT	I_salaires nets	38
5	AbuDhabi	PROCHE ORIENT	Heures travail annuelles	2 100
6	AbuDhabi	PROCHE ORIENT	Pouvoir d'achat brut	42
7	AbuDhabi	PROCHE ORIENT	Pouvoir d'achat net	53.3
8	Amsterdam	CENTRAL EUROPE	I_prix sans loyer	68.8
9	Amsterdam	CENTRAL EUROPE	I_prix avec loyer	70.3
10	Amsterdam	CENTRAL EUROPE	I_salaires bruts	61.4
11	Amsterdam	CENTRAL EUROPE	I_salaires nets	53.7
12	Amsterdam	CENTRAL EUROPE	Heures travail annuelles	1 792
13	Amsterdam	CENTRAL EUROPE	Pouvoir d'achat brut	89.3
14	Amsterdam	CENTRAL EUROPE	Pouvoir d'achat net	78

Figure 6-19 : Table obtenue après verticalisation

Les variables conservées apparaissent en début de table et leurs valeurs sont répétées autant de fois qu'il y a de variables verticalisées (7 fois dans cet exemple).

Méthodes de data management pour les colonnes

Les méthodes de data management pour les colonnes sont accessibles dans la vue « Méthodes » :

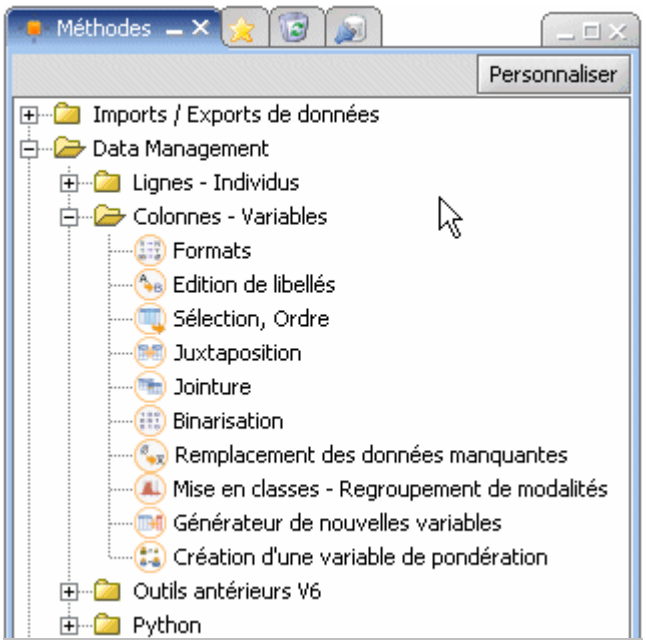


Figure 6-20 : Opérations de data managements sur les colonnes

Formats

La méthode « Format » permet de modifier le format des variables d'une table de données. Concrètement, voici la liste des opérations que l'on peut réaliser à partir de cette méthode :

- Modifier le nom d'une variable,
- Modifier le type de stockage d'une variable (Chaine ➢ date, Chaine ➢ Entier…),
- Modifier le rôle statistique d'une variable,
- Modifier les libellés des modalités d'une variable nominale par une grille d'association,
- Définir un intervalle [Valeur min, Valeur max] d'acceptation des valeurs pour toute variable numérique, les valeurs sortant de cet intervalle étant considérées comme manquantes
- Spécifier les valeurs à considérer comme manquantes et le traitement des données manquantes pour chaque variable.

Vous pouvez sauvegarder les formats dans une bibliothèque pour pouvoir les réutiliser par la suite.

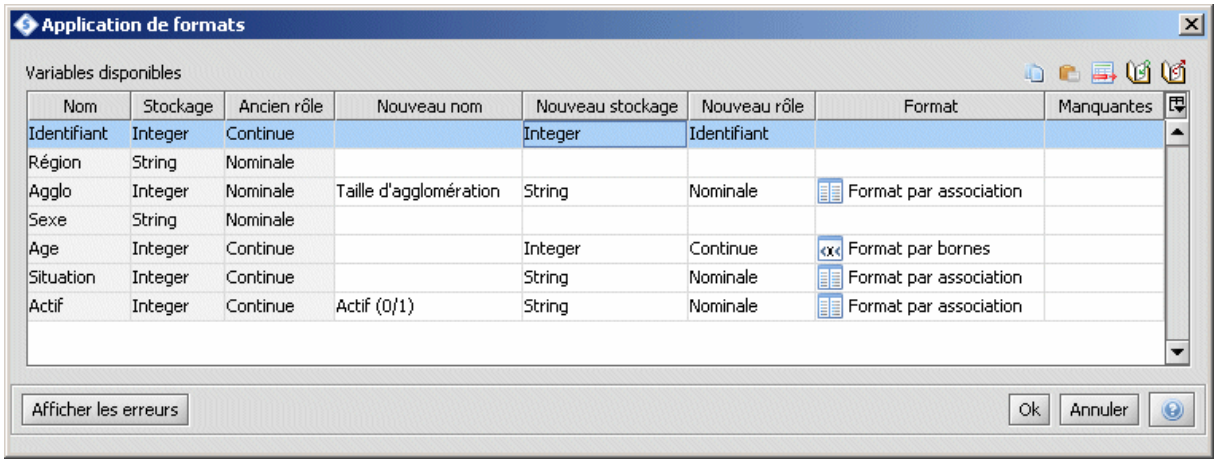


Figure 6-21 : Méthode Format

Le clic droit de la souris permet d'accéder au menu suivant :

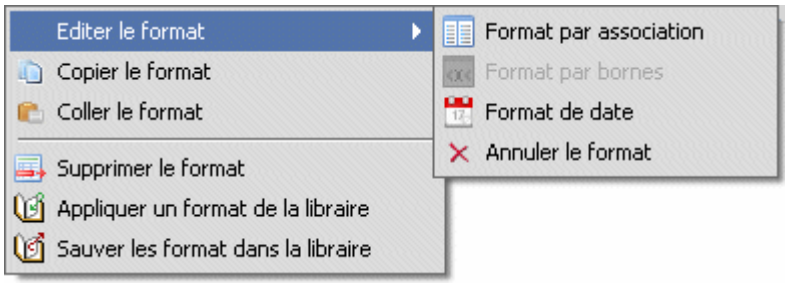


Figure 6-22 : Menu clic droit de l'interface Format

Les différents formats sont présentés dans les pages suivantes.

Par clic-droit, il est possible de copier/coller un format d'une variable à une autre. Le clic droit permet également de rechercher un format dans une librairie pour l'appliquer à la ou les variables sélectionnées et enfin d'enregistrer un format dans une librairie.

Variable, Ancien stockage, Ancien rôle. Ces trois colonnes présentent les noms, les types de stockage et les rôles statistiques des variables disponibles dans la table d'origine.

Nouveau nom. Pour modifier le nom d'une variable, double-cliquez dans la case de la colonne « Nouveau nom » et tapez le nom choisi.

Nouveau stockage. Pour modifier le type de stockage d'une variable, double-cliquez dans la case de la colonne correspondante pour faire apparaître la liste de choix (Entier, Réel, Chaine de caractères, Date, Autre). Si certaines valeurs posent problème lors de la conversion (ex. convertir « Oui » en entier), ces seront remplacées par des données manquantes.
Pour convertir une chaine de caractère représentant une date en stockage « Date », spécifiez le format date associé dans la colonne « Format ».

Nouveau rôle. Pour modifier le rôle statistique des variables pour les méthodes d'analyse de données, double-cliquez dans la case de la colonne « Nouveau rôle » pour faire apparaître la liste de choix (Identifiant, Continue, Nominale, Textuelle, Exclue). Référez vous au chapitre sur les métadonnées en page 32 pour faire votre choix.
Remarque : seules les données stockées comme « Entier » ou « Réel » peuvent avoir le rôle de variable continue.

Format. Selon le type de stockage et le rôle statistique, vous pouvez compléter la définition de la variable par un format spécifique en double-cliquant dans la case de la variable concernée.
Pour les variables nominales, le format par association permet d'associer à chaque modalité d'origine un nouveau libellé ou une nouvelle valeur numérique si la variable nominale est stockée sous forme d'entier.
Pour les variables numériques (Entier ou Réel), vous pouvez spécifier un intervalle d'acception des valeurs pour traiter le problème des valeurs aberrantes.
Pour les chaines de caractères transformées en stockage date, vous pouvez spécifier le format de lecture des dates.

- **Format par association**

Le format par association n'est disponible que pour les variables dont l'ancien rôle ou le nouveau rôle est « Nominale ». A chaque valeur d'origine est associée un rang (Nouvel ordre) et une nouvelle valeur.

Afficher les valeurs. Ce bouton permet d'afficher les valeurs d'origine ou modalités de la variable concernée dans la « Liste des associations ». Cette fonctionnalité n'est disponible que pour les variables dont l'ancien rôle est « Nominale ».

Nouveau stockage. Nouveau type de stockage de la variable nominale. Selon le type numérique (Entier ou réel) ou chaine de caractères, les nouveaux libellés seront respectivement des nombres ou des caractères alphanumériques.

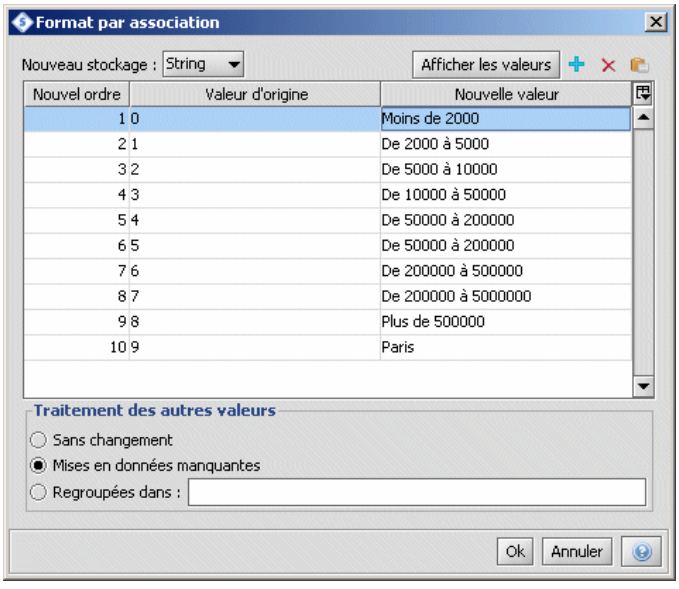


Figure 6-23 : Format des modalités

Liste des associations. Chaque ligne de la liste correspond à une association entre une valeur d'origine d'une part, un ordre et une nouvelle valeur d'autre part. Tous ces éléments sont modifiables.

Nouvel ordre. Cette colonne indique l'ordre d'affichage des modalités qui sera utilisé pour l'ensemble des méthodes statistiques de SPAD et notamment les tris-à-plat. Pour modifier l'ordre des modalités, sélectionnez la ou les modalités à l'aide de la souris et glissez-déposez la sélection à l'endroit choisi.

Valeur d'origine. Cette colonne liste les modalités ou valeurs d'origine de la variable sélectionnée. Si la liste est vide, cliquez sur le bouton « Afficher les valeurs » pour faire apparaître les valeurs de la variable.
Les valeurs d'origine sont modifiables. Cette fonctionnalité est surtout utile lorsque l'on souhaite définir une nouvelle association entre une valeur d'origine absente du fichier courant et une nouvelle valeur.

Nouvelle valeur. La colonne « Nouvelle valeur » présente les nouvelles modalités qui apparaitront dans la table de données en sortie.

Cliquez sur ce bouton pour définir une nouvelle association. Cette nouvelle association apparaît en dernier dans la liste des associations. Il faut alors renseigner la valeur d'origine et la nouvelle valeur.
Cliquez sur ce bouton pour supprimer la ou les associations préalablement sélectionnées.

Traitement pour les autres valeurs. Cette fonctionnalité permet de gérer le cas des valeurs d'origine non listées dans la liste des associations ou des nouvelles valeurs qui pourraient apparaître lors d'une mise à jour des données sources.

- Les autres valeurs restent telles quelles : chaque valeur non listée dans la liste d'association est conservée telle quelle, c'est-à-dire comme une modalité.
- Mises en données manquantes : chaque valeur non listée dans la liste d'association est remplacée par une donnée manquante.
- Regroupées dans : permet de regrouper toutes les valeurs non listées dans la liste d'association dans une seule modalité dont le nom est à spécifier (ou la valeur si le stockage est de type numérique).

- **Format par bornes**

Le format par bornes est uniquement disponible pour les variables numériques (Stockage Entier ou Réel). Ce format permet de spécifier un intervalle d'acception des valeurs pour traiter le problème des valeurs aberrantes.

Statistiques descriptives. Les statistiques descriptives rappellent les principales caractéristiques de la variable sélectionnée.

Borne inférieure. En cochant cette case, toutes les valeurs qui seront strictement inférieures à la borne inférieure définie dans la case située à droite seront transformées en données manquantes.

Borne inférieure. En cochant cette case, toutes les valeurs qui seront strictement supérieures à la borne supérieure définie dans la case située à droite seront transformées en données manquantes.

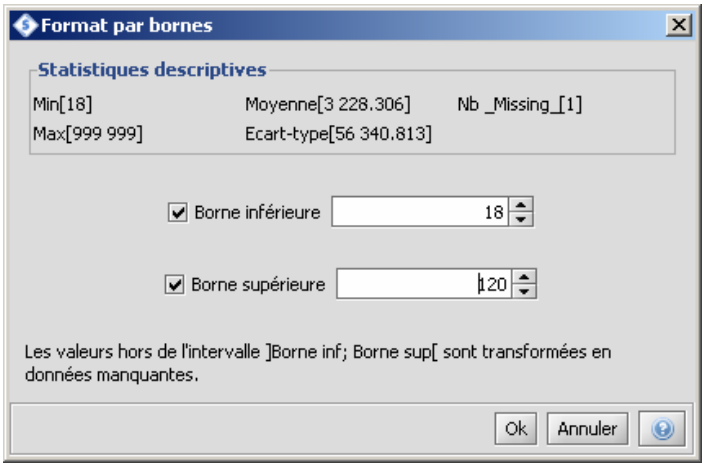


Figure 6-24 : Format par bornes

-

Format Date

Ce format n'est disponible que pour les variables dont l'ancien et le nouveau stockage sont respectivement « Chaine de caractères » et « Date ».
Ce format permet de définir la séquence des éléments caractéristiques d'une date ou date/heure dans la variable d'origine stockée sous forme de chaine de caractères.

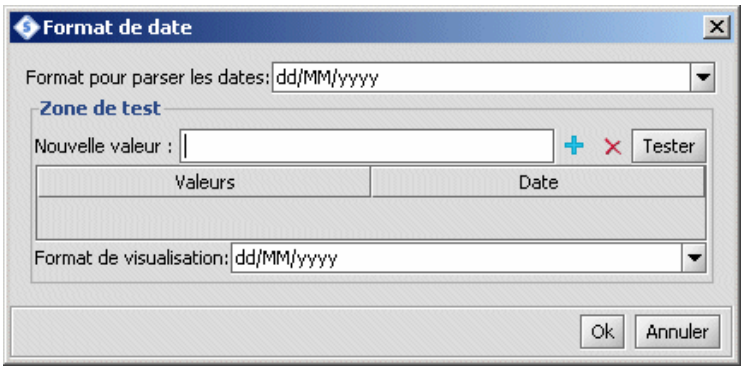


Figure 6-25 : Format Date

Format pour parser les dates. Les lettres « d », « M », « y » représentent respectivement les chiffres qui composent les jours (day), les mois(Month) et les années (year).

Exemples de correspondances de formats et de dates stockées en chaines de caractères :

Chaine de caractères	Format pour parser	Résultats
12-7-87	dd-MM-yy	12/07/1987
12/7/1987	dd/MM/yyyy	12/07/1987
12/07/07	dd/MM/yyyy	12/07/0007

Pour spécifier votre format, suivez la nomenclature suivante :

Lettre	Désignation	Présentation	Exemples
y	Année	Numérique	1996; 96
M	Mois dans l'année	Texte ou numérique	Avril; Avr; 04; 4
w	Semaine dans l'année	Entier	27
W	Semaine dans le mois	Entier	2
D	Jour dans l'année	Entier	189
d	Jour dans le mois	Entier	10
F	Jour de la semaine dans le mois	Entier	2
E	Jour de la semaine dans le mois	Texte	Mardi; Mar
a	Indicateur Am/pm	Texte	PM
H	Heure (0-23)	Entier	0
k	Heure (1-24)	Entier	24
K	Heure pour am/pm (0-11)	Entier	0
h	Heure pour am/pm (1-12)	Entier	12
m	Minute dans l'heure	Entier	30
s	Seconde dans la minute	Entier	55
S	Milliseconde	Entier	978
Z	Zone horaire	Texte	CET

Figure 6-26 : Nomenclature pour le format des dates

Si votre date comporte des séparateurs, les introduire tels quels dans le format.

Si vos dates ne sont pas reconnues, utilisez les fonctions de conversion de dates et notamment la fonction _STRING_TO_DATE_ dans le générateur de nouvelles variables pour convertir votre variable de type « chaine de caractères » en type « date ».

Test. Cette fonctionnalité vous permet de tester que le format pour parser les dates permet bien de retrouver l'ensemble des éléments (Jour, Mois, Année, Heure, Minutes, Secondes) qui composent la date.

Manquantes. Pour définir le traitement des données manquantes pour une variable donnée, cliquez dans la case correspondante de la colonne « Manquantes ». La boite de dialogue ci-après vous permet de spécifier les valeurs à considérer comme manquantes et éventuellement de remplacer au final ces valeurs manquantes par une nouvelle valeur. Cette dernière fonctionnalité est intéressante pour recoder les données manquantes d'une variable nominale en une nouvelle modalité.

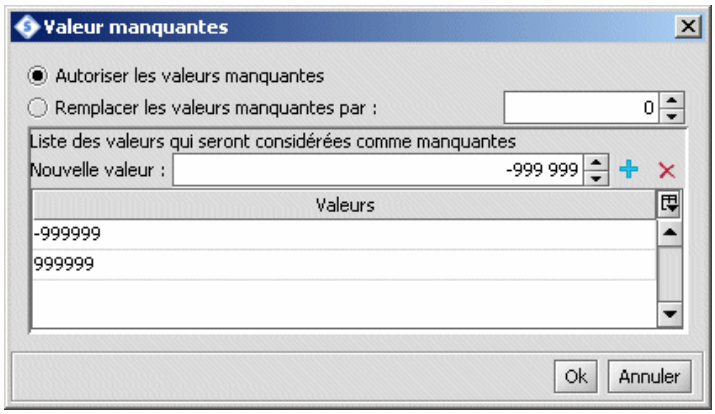


Figure 6-27 : Valeurs manquantes

Ordonner les modalités automatiquement selon leur ordre naturel.
Cochée, cette option trie les modalités d'une nouvelle variable nominale dans l'ordre alphanumérique ou numérique selon qu'elle est stockée comme chaine de caractères ou comme entier. Cet ordre est repris dans les méthodes statistiques pour l'affichage des résultats. En décochant cette option, l'ordre obtenu est celui dans lequel les modalités apparaissent dans les données.

Gestionnaire des librairies de formats

Accessible à partir du menu « Outils / Librairies de formats », le gestionnaire des librairies vous permet de :

- Créer une nouvelle librairie de formats
- Supprimer une librairie existante
- Modifier une librairie existante
- Créer de nouveaux formats
- Supprimer des formats existants
- Modifier des formats existants.

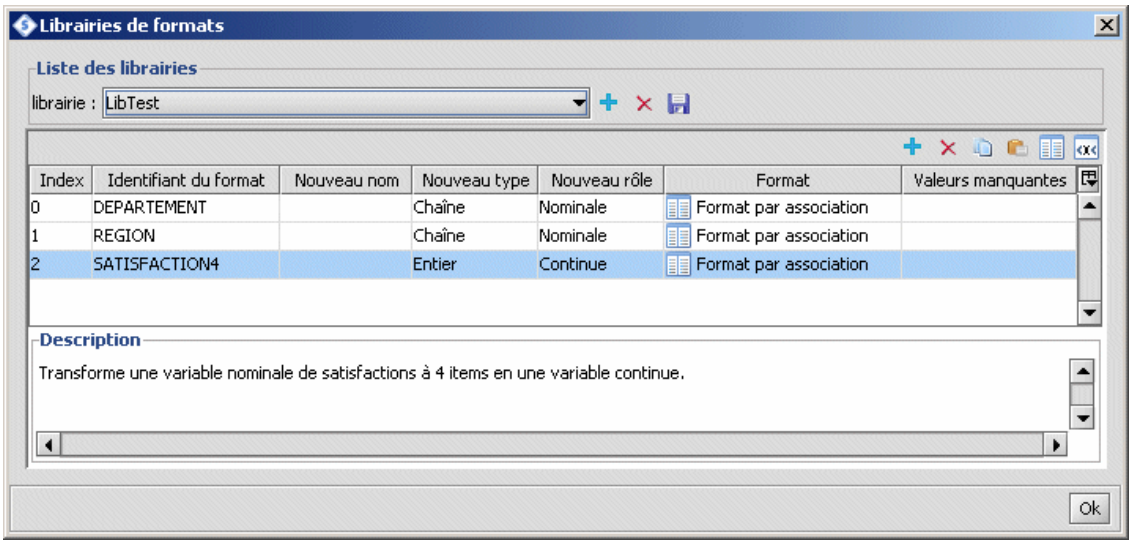


Figure 6-28 : Gestionnaire des librairies de formats

Edition des libellés

Cette méthode permet de modifier les libellés de variables et de modalités de variables nominales. Elle dispose de fonctionnalités avancées pour réordonner, fusionner, uniformiser, générer des libellés de modalités de variables nominales.

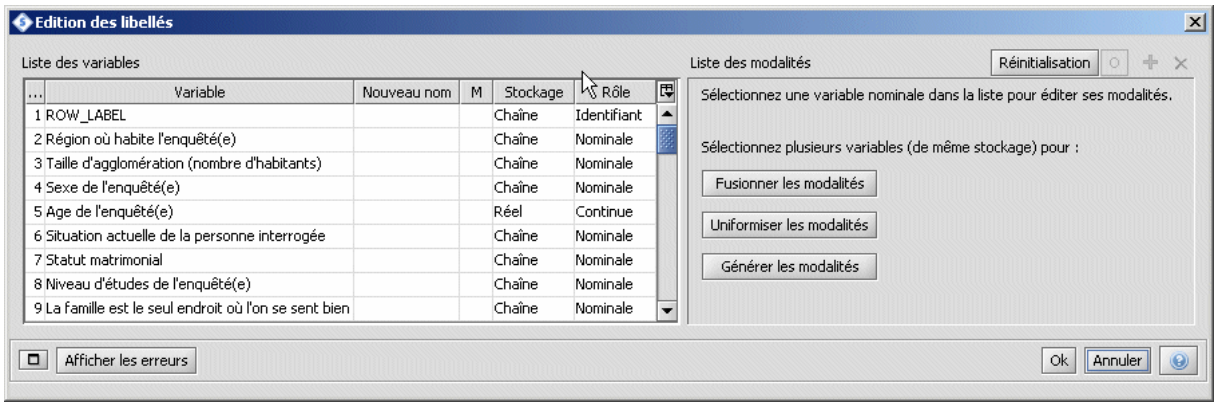


Figure 6-29 : Edition des libellés

A l'ouverture de la méthode vous disposez dans la fenêtre de gauche de l'ensemble des variables avec leur Index, leur nom, une colonne pour le nouveau nom, une colonne notée « M » qui indiquera la modification éventuelle de libellés de modalités, leur stockage et leur rôle.

Modifier le nom d'une variable.
Double-cliquez dans la case « Nouveau nom » de la variable. L'ancien nom est automatiquement dupliqué dans la case pour servir de base éventuelle. Spécifiez le nouveau nom et validez par la touche « Entrée » (touche « Echap » pour annuler la modification en cours).

Modifier les libellés des modalités d'une variable nominale.
Sélectionnez la variable dans la liste, les modalités apparaissent dans la liste de droite. Vous disposez d'une colonne « Valeur » initialisée à l'ancienne valeur. Pour modifier un libellé de modalité double-cliquez dans la case « Valeur » de la modalité, effectuez la modification et validez par la touche « Entrée » (touche « Echap » pour annuler la modification en cours).

Les variables ainsi modifiées sont marquées d'une croix « X » dans la colonne « M » de la liste des variables.
Attention : le fait de modifier une valeur en une valeur existante ne provoque pas un regroupement de modalités, vous aurez seulement deux valeurs identiques pour deux index différents.

Réordonner les libellés des modalités d'une variable nominale.
Sélectionnez une variable dans la liste de gauche, les modalités apparaissent dans la fenêtre de droite. La colonne « Index » indique le nouvel ordre des modalités. A l'initialisation, il correspond à l'ancien ordre (colonne « Ancien »).

- **Réordonner les modalités par glisser – déposer**
  Sélectionnez une ou plusieurs modalités dans la liste, puis en maintenant le clic gauche, faites glisser la sélection à l'endroit voulu et relâcher le clic souris. Dans la colonne « Ancien » , les index modifiés sont marqués en vert, dans la colonne « Index » figure le nouvel ordre.

- **Par validation d'un tri automatique**
  Vous pouvez également effectuer des tris automatiques et validez le nouvel ordre généré. Les tris sont effectués en cliquant sur les entêtes de colonnes. On peut par exemple trier les

modalités par effectifs décroissants.

Note : Si vous effectuez un tri, l'affichage de toutes les variables se fera selon ce tri, pour retrouver l'ordre initial, triez selon la colonne « Index ».
Ce tri est visuel (la colonne « Index » n'est pas triée), pour valider ce tri en tant que nouvel ordre, cliquez sur le bouton  .

La colonne "Index" est triée, le nouvel ordre est validé.
Dans la liste des variables, les variables ainsi modifiées sont marquées « X » dans la colonne « M ».

Uniformiser les libellés des modalités de plusieurs variables nominales
Cette fonctionnalité permet modifier les libellés de plusieurs variables en utilisant les modalités en cours (modifiés ou non) d'une variable comme référence.

1. Sélectionnez les variables dans la liste de gauche, puis par le menu contextuel (Clic droit), ou par le bouton dans la partie droite, accédez à la fonctionnalité « Uniformiser les modalités ».
2. Dans la fenêtre « Uniformiser les modalités », la première variable de la liste est sélectionnée comme variable de référence. Pour changer de variable de référence, double-cliquez sur une des variables de la liste ou accédez à une autre variable dans liste déroulante.
3. Cliquez sur le bouton OK pour valider votre choix.
4. Dans la liste des variables, les variables ainsi modifiées sont marquées « X » dans la colonne « M ».

Fusionner les modalités de plusieurs variables nominales
Pour une liste de variables sélectionnées, l'opération de fusion de modalités conserve les modalités communes et ajoute les modalités non communes à chaque variable.

Exemple : vous disposez de trois variables codées

- Première variable : Oui /Non
- Seconde variable : Oui /Non/Ne sait pas
- Troisième variable : Oui/Non/ne sait pas/Sans Objet

Si vous fusionnez les modalités de ces trois variables, les trois variables auront chacune 4 modalités : Oui / Non / Ne sait pas/ Sans objet.

Pour réaliser une fusion :
1. Sélectionnez les variables dans la liste,
2. Puis dans le menu contextuel accessible par clic droit, choisissez « fusionner les modalités » ou double-cliquez sur le bouton « fusionner les modalités ».
3. Dans la fenêtre de fusion, vous avez un rappel des variables dont on fusionne les modalités, cliquez sur Ok pour valider.
4. Dans la liste des variables, les variables ainsi modifiées sont marquées « X » dans la colonne « M ».

Générer des libellés de façon automatique.
Cette fonctionnalité ne s'applique qu'aux variables nominales de stockage « Chaîne » et permet de générer automatiquement des libellés de modalités par une gestion combinée du nom de la variable et des libellés des modalités.

Cette fonctionnalité utilise les libellés en cours (modifiés ou non) des noms de variables et des modalités.

1. Sélectionnez la ou les variables, puis dans le menu contextuel (Clic droit) sélectionnez l'item « Générer les modalités ». Si vous avez sélectionné plusieurs variables vous pouvez aussi cliquer sur le bouton « Générer les modalités » dans la partie droite.
2. Dans la fenêtre « Générer les modalités », choisissez le mode d'utilisation du libellé de la variable, puis du libellé des modalités et cliquez sur OK pour valider vos choix.
3. Dans la liste des variables, les variables ainsi modifiées sont marquées « X » dans la colonne « M ».

Annuler des modifications
Pour annuler une ou plusieurs modifications des noms de variables, sélectionnez la ou les variables, puis dans le menu contextuel accessible par clic droit, sélectionnez « Annuler les nouveaux noms ».

Pour annuler une ou plusieurs modifications de libellés de modalités, sélectionnez la ou les variables, puis dans le menu contextuel accessible par clic droit sélectionnez « Annuler les modifications sur les modalités ». Après annulation, le signe « X » de la colonne « M » disparaît.

Le bouton « Réinitialisation » annule toutes les modifications effectuées.

Supprimer systématiquement les modalités d'effectif nul
Cette option supprime pour l'ensemble des variables nominales les modalités sans effectif ou d'effectif nul (0).

Les opérations de Copier / coller
Vous disposez de deux modes pour effectuer les opérations de copier / coller.

- **Par associations**
  Le principe est de travailler avec deux colonnes, l'une contient une valeur que l'on recherche et l'autre la valeur de remplacement. C'est le principe adopté par défaut dans cette interface.

  Exemple : Vous disposez de plusieurs variables nominales qui ont par exemple deux valeurs communes « Ne sait pas « et « Sans Objet », les autres valeurs étant différentes ou non. Vous voulez remplacer « Ne Sait pas » par « NSP » et « Sans Objet » par « SO » pour toutes les variables.

  Pour l'une quelconque des variables, procédez au changement de libellés (colonne « Valeur »), puis sélectionnez toutes les modalités (Ctrl A) et copiez les dans le presse-papiers (Ctrl C), positionnez vous sur une des variables à modifier et faites « Coller » (Ctrl V) dans la liste des modalités, le remplacement par association est automatique.

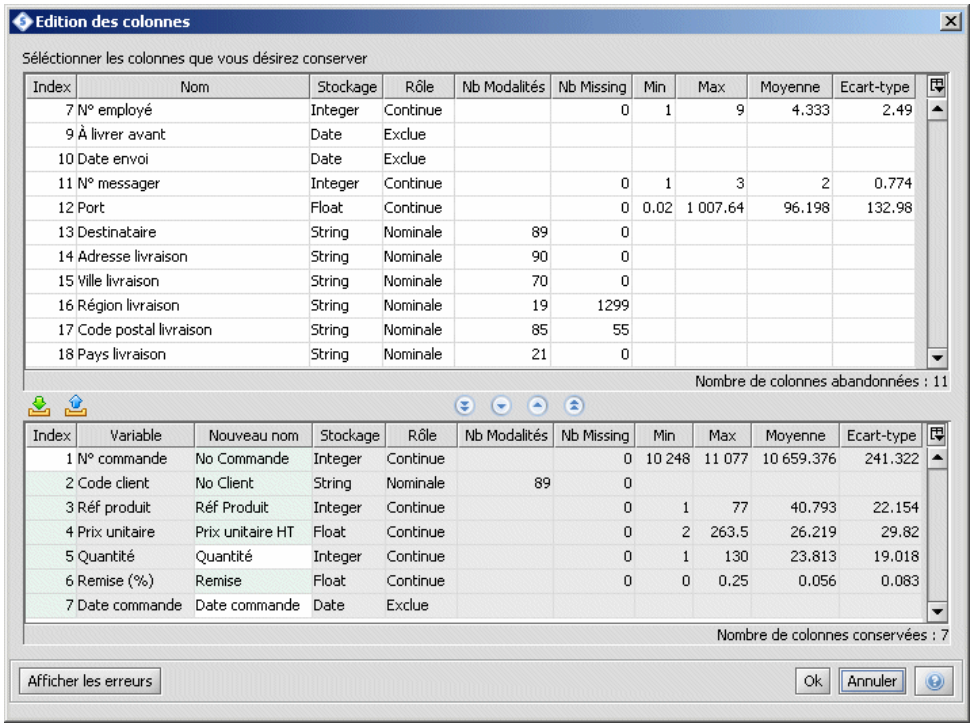
  Vous pouvez également faire ce type de « Copier coller » en utilisant un tableur externe comme Excel. Pour les modalités la première colonne correspond à la colonne « Valeur » et la seconde à « Ancienne valeur. Pour le libellés des variables, la première colonne correspond à la colonne « Variable » et la seconde à la colonne « Nouveau nom ».
- **Le copier coller simple**
  Travaille à partir d'une seule colonne et copie le contenu du presse-papiers dans la colonne « Nouveau nom » si on est positionné dans la liste des variables ou dans la colonne « Valeur » si on est positionné sur une liste de modalités de variables.

Exécution.
Cliquez sur OK pour valider et exécutez. Après exécution de la méthode, dans la vue « Données » de la fenêtre « Exécutions », on peut visualiser les données correspondant à la table construite.

Les données issues de l'édition des libellés sont immédiatement utilisables par toutes les autres méthodes.

Sélection, Ordre

Cette procédure permet filtrer des colonnes et de les réordonner.
Les données issues de cette sélection sont immédiatement utilisables par toutes les autres méthodes.



Index	Nom	Stockage	Rôle	Nb Modalités	Nb Missing	Min	Max	Moyenne	Ecart-type
7	N° employé	Integer	Continue		0	1	9	4.333	2.49
9	À livrer avant	Date	Exclue						
10	Date envoi	Date	Exclue						
11	N° messager	Integer	Continue		0	1	3	2	0.774
12	Port	Float	Continue		0	0.02	1 007.64	96.198	132.98
13	Destinataire	String	Nominale	89	0				
14	Adresse livraison	String	Nominale	90	0				
15	Ville livraison	String	Nominale	70	0				
16	Région livraison	String	Nominale	19	1299				
17	Code postal livraison	String	Nominale	85	55				
18	Pays livraison	String	Nominale	21	0				

Index	Variable	Nouveau nom	Stockage	Rôle	Nb Modalités	Nb Missing	Min	Max	Moyenne	Ecart-type
1	N° commande	No Commande	Integer	Continue		0	10 248	11 077	10 659.376	241.322
2	Code client	No Client	String	Nominale	89	0				
3	Réf produit	Réf Produit	Integer	Continue		0	1	77	40.793	22.154
4	Prix unitaire	Prix unitaire HT	Float	Continue		0	2	263.5	26.219	29.82
5	Quantité	Quantité	Integer	Continue		0	1	130	23.813	19.018
6	Remise (%)	Remise	Float	Continue		0	0	0.25	0.056	0.083
7	Date commande	Date commande	Date	Exclue						

Figure 6-30 : Sélection de colonnes, Ordre

Dans les deux listes, l'usage des touches Shift et Ctrl est standard pour sélectionner les variables.
Le double clic sur une variable transfère la variable d'une liste à l'autre.

Sélection et transfert.

- Transfère toutes les variables dans la liste « Variables retenues ».
- Transfère les variables préalablement sélectionnées de la liste « Variables disponibles » dans la liste « Variables retenues ».
- Transfère toutes les variables dans la liste « Variables disponibles ».
- Transfère les variables préalablement sélectionnées dans la liste « Variables retenues » dans la liste « Variables disponibles ».

Ordre/Positionnement des variables.
Dans la nouvelle base, l'ordre des variables sera celui de la liste des « Variables retenues ».

Modifier l'ordre des variables.
Pour redéfinir le positionnement des variables dans la table créée, sélectionnez la ou les variables à repositionner et glissez/déposer la sélection à l'endroit choisi.

Nouveau nom.
Pour modifier le nom d'une variable sélectionnée, double-cliquez dans la case correspondante et entrez le nouveau nom.

Cliquez sur Ok pour valider votre paramétrage. Après exécution, un aperçu des données est visible dans la vue « Données ».

Juxtaposition

Cette méthode permet de juxtaposer plusieurs tables « côtes à côtes » pour n'en produire qu'une seule en sortie.
Attention : Les lignes des différentes tables sont juxtaposées selon leur rang dans leur table d'origine : la première ligne de la table 1 avec la première ligne de la table 2…
Si les tables d'origine ne sont pas triées de manière identique, il est préférable d'utiliser la méthode « Jointure » qui permet de spécifier une clé de jointure ou de correspondance entre les lignes de deux tables distinctes.

La méthode « Juxtaposition » présente l'avantage d'être extrêmement rapide et de ne pas dupliquer les données de la base de données interne.

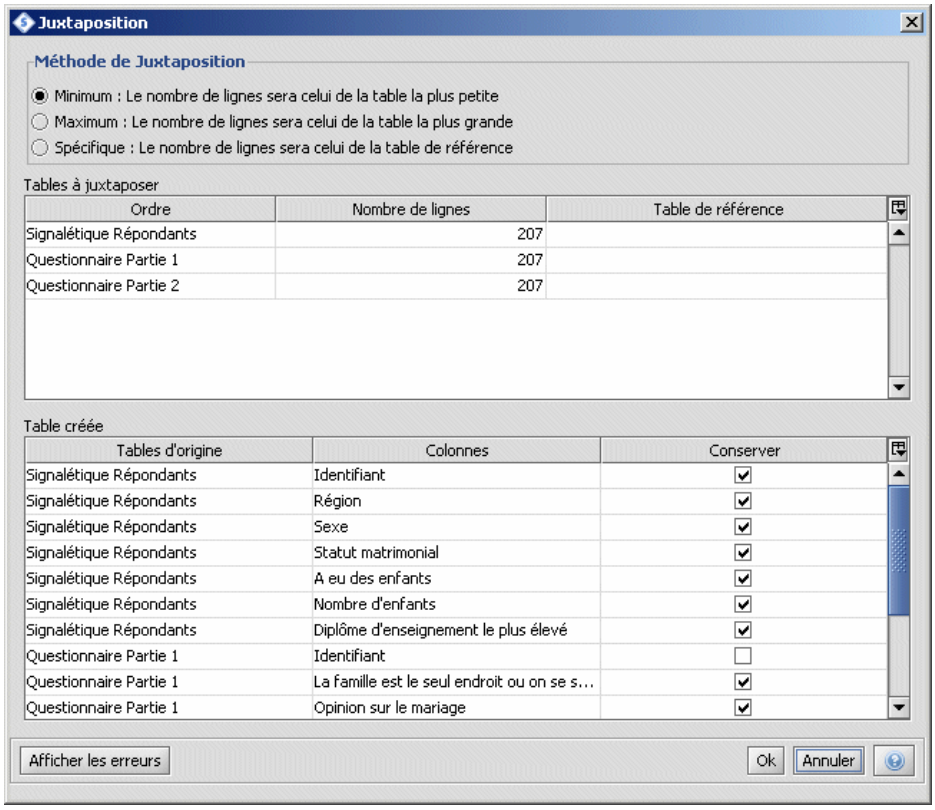


Figure 6-31 : Juxtaposition de tables

Méthode de juxtaposition. Utile si le nombre de lignes dans les tables à juxtaposer diffère.
Minimum. La table produite comportera le nombre de lignes de la plus petite table d'origine.
Maximum. La table produite comportera le nombre de lignes de la plus grande table d'origine.
Spécifique. La table produite comportera le nombre de lignes de la table de référence, définie par l'utilisateur.

Pour les méthodes « Maximum » et « Spécifique », les individus (i.e. lignes) ne figurant pas dans une ou plusieurs tables d'origine contiendront des données manquantes pour les tables concernées.

Ordre. Pour modifier l'ordre d'apparition des tables à juxtaposer dans la table produite, sélectionnez la ou les tables à réordonner dans la zone « Tables à juxtaposer » et glissez-déposez la sélection à l'endroit choisi.

Table de référence. Pour spécifier une table de référence dans la liste des tables à juxtaposer, cliquez dans la case de la colonne « Table de référence » sur la variable choisie. Une case cochée apparaît.

Conserver. Pour faire une sélection de variables, cochez/décochez les cases de la colonne « Conserver » dans la zone « Table créée ».

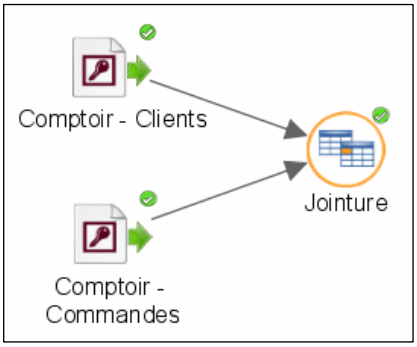
Jointure

Une jointure est une combinaison (produit cartésien) des enregistrements de deux tables disposant de valeurs correspondantes dans une variable déterminée de chacune des deux bases.

Exemple :
Vous disposez d'une table client avec une variable « Code Client » et d'autres informations socio-économiques, comportementales, ou autres. Par ailleurs vous disposez d'une table « Achats » avec pour chaque achat, le « Code client » et la description de l'achat.

Une équijointure entre les deux tables avec comme variable commune de référence le « Code client », permettra d'obtenir une table « Achats » contenant pour chaque achat, les informations de l'achat et les informations du client.

Mise en œuvre
On peut utiliser pour l'exemple la Base Access « Comptoir.mdb » fourni avec SPAD. On procédera tout d'abord à l'importation des tables Clients (91 clients) et Commandes (830 commandes). A noter que 2 clients parmi les 91 n'ont pas effectué de commande.



Dans la vue « Méthodes », accédez à Data Management / Colonnes – Variables, puis double cliquez sur Jointure ou faites glisser l'icône Jointure sur la vue « Diagramme » en cours. Reliez ensuite les deux tables sur lesquelles vous souhaitez réaliser la jointure à la méthode.

Paramétrage
Double cliquez sur l'icône Jointure, les variables disponibles pour les deux tables apparaissent automatiquement. Dans notre exemple la table « Clients » est la table de gauche.

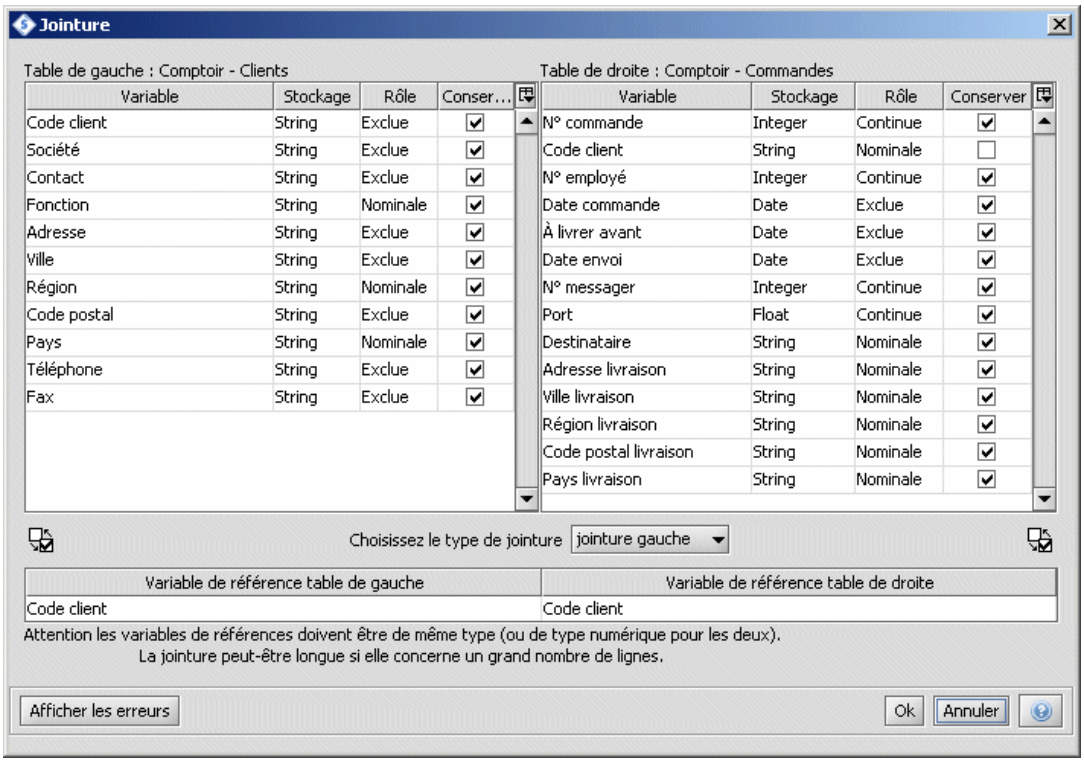


Figure 6-32 : Interface de la jointure

Variables de références

Les deux variables de références (i.e. clés de jointure) sont celles qui permettent de définir le lien existant entre la table de gauche et la table de droite.

Par défaut, à l'ouverture du paramétrage, SPAD recherche la présence d'une même variable (même nom) au sein des deux tables. S'il existe plusieurs variables présentes dans les deux tables, SPAD propose la première équivalence trouvée.
Pour définir la variable de référence de chaque table, sélectionnez la variable de référence, double-cliquez sur celle-ci ou glissez-déposez la variable dans la case « Variable de référence ».

Les deux variables de référence doivent avoir le même type de stockage. Si ce n'est pas le cas, utilisez la méthode « Formats » dans la rubrique « Data Management \ Colonnes – Variables » pour modifier le type de stockage.
Pour une jointure basée sur une clé multiple (i.e. plusieurs variables de référence dans chaque table), utilisez le générateur de nouvelles variables et la fonction « _CONCATENATE_ » de la rubrique « Texte » pour constituer pour chaque table une variable de référence unique par concaténation.

Choix du type de jointure

La liste déroulante permet de choisir entre plusieurs types de jointure. Par défaut, l'équijointure est sélectionnée. Nous allons illustrer les différents types de jointures et résultats obtenus à partir des deux tables suivantes :

Table de Gauche

Department Table

DepartmentID	DepartmentName
31	Sales
33	Engineering
34	Clerical
35	Marketing

Table de Droite

Employee Table

LastName	DepartmentID
Rafferty	31
Jones	33
Steinberg	33
Robinson	34
Smith	34
Jasper	36

Equijointure

L'équijointure ou jointure interne (INNER JOIN en langage SQL) retient les enregistrements communs aux deux tables au travers de la variable de référence. L'enregistrement créé contient les variables retenues des deux bases.
Une équijointure entre les deux tables avec comme variable commune de référence le « DepartmentID », permettra d'obtenir la table suivante :

DepartmentID	DepartmentName	LastName	DepartmentID_1
31	Sales	Rafferty	31
33	Engineering	Jones	33
33	Engineering	Steinberg	33
34	Clerical	Robinson	34
34	Clerical	Smith	34

Jointure gauche

La jointure gauche contiendra tous les éléments présents dans la table de gauche et référencés ou non par la colonne de référence dans la table de droite. Si des enregistrements ne sont pas référencés dans la table de droite, ils seront complétés par des « non réponses » pour ces informations. Dans l'exemple, on obtient la table suivante :

DepartmentID	DepartmentName	LastName	DepartmentID_1
31	Sales	Rafferty	31
33	Engineering	Jones	33
33	Engineering	Steinberg	33
34	Clerical	Robinson	34
34	Clerical	Smith	34
35	Marketing		

Jointure gauche sans doublon

La jointure gauche sans doublon a le même comportement que la jointure gauche si ce n'est que le nombre de lignes de la table produite est égal à celui de la table de gauche. Si une clé présente dans la table de gauche est trouvée plusieurs fois dans la table de droite, c'est la première ligne lue dans la table de droite qui sera jointe à celle de la table de gauche.
Dans l'exemple, on obtient la table suivante :

DepartmentID	DepartmentName	LastName	DepartmentID_1
31	Sales	Rafferty	31
33	Engineering	Jones	33
34	Clerical	Robinson	34
35	Marketing		

Jointure droite

La jointure droite est l'équivalent de la jointure gauche dans l'autre sens. Dans notre exemple, on obtient le résultat suivant :

DepartmentID	DepartmentName	LastName	DepartmentID_1
31	Sales	Rafferty	31
33	Engineering	Jones	33
33	Engineering	Steinberg	33
34	Clerical	Robinson	34
34	Clerical	Smith	34
		Jasper	36

Jointure droite sans doublon

La jointure droite sans doublon est l'équivalent de la jointure gauche sans doublon, dans l'autre sens.

Jointure complète

La jointure complète (FULL OUTER JOINT en langage SQL) applique les règles de la jointure gauche et de la jointure droite.

DepartmentID	DepartmentName	LastName	DepartmentID_1
31	Sales	Rafferty	31
33	Engineering	Jones	33
33	Engineering	Steinberg	33
34	Clerical	Robinson	34
34	Clerical	Smith	34
35	Marketing		
		Jasper	36

Produit cartésien

Cette jointure ne prend pas en compte de variable de référence. Elle croise chaque ligne de la table de gauche avec chaque ligne de la table de droite. La table créée contient autant de lignes que le produit des nombres de lignes des deux tables. Ce type de jointure doit être utilisé avec précaution.

Autres paramétrages

Les cases à cocher des variables des deux tables vous permettent de sélectionner les variables à conserver dans la table créée. Le bouton  permet d'inverser la sélection en cours.

Cliquez sur Ok pour valider votre paramétrage
Après exécution, un aperçu des données est visible dans la vue « Données ».

Les nouvelles données sont immédiatement utilisables par d'autres méthodes statistiques ou de Data Management.

Remarque sur les jointures sans doublon.
Les jointures sans doublon sont très souvent utiles lorsque l'on souhaite ***enrichir une table de référence à partir d'informations provenant d'autres tables****. Dès lors que l'on souhaite* ***conserver exactement le même nombre de lignes dans la table de référence****, il est conseillé d'utiliser une jointure sans doublon.*

Binarisation

L'opération de binarisation pour une variable nominale à N modalités consiste à créer N variables nominales (présence / absence) ou N variables continues codées (0/1) correspondant aux N modalités initiales.
La transformation en N variables continues (0/1) est très pratique pour effectuer des analyses de correspondances simples.
Les données issues de cette transformation sont immédiatement utilisables par toutes les autres méthodes.

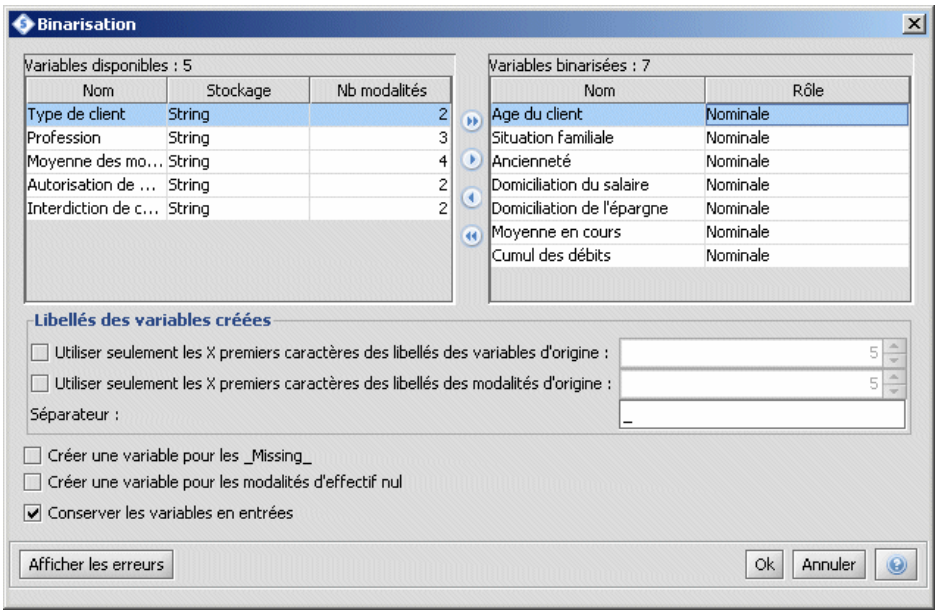


Figure 6-33 : Interface de la binarisation

Variables disponibles. Seules les variables nominales (rôle) sont disponibles pour la binarisation. La colonne « Nb modalités » permet de savoir combien de variables seront générées par la binarisation d'une variable.

Variables binarisées. Liste des variables qui seront respectivement « éclatées » en autant de colonnes qu'elles ont de modalités.

Rôle des variables binarisées. Par défaut, le rôle des variables binarisées est « Nominale ». Pour le modifier, double-cliquez dans la case de la colonne « Rôle » pour faire apparaître la liste déroulante.

Sélection et transfert de variables. Dans les deux listes, l'usage des touches Shift et Ctrl est standard pour sélectionner les variables. Le double clic sur une variable transfère la variable d'une liste à l'autre.
Transfère toutes les variables disponibles dans la liste « Variables binarisées ».
Transfère les variables disponibles sélectionnées dans la liste « Variables binarisées ».

Transfère les variables binarisées sélectionnées dans la liste « Variables disponibles ».
Transfère toutes les variables binarisées dans la liste « Variables disponibles ».

Modifier le rôle des variables binarisées (Nominales/continues). Pour le modifier le rôle des variables binarisées, double-cliquez dans la case de la colonne « Rôle » et choisissez entre Nominale et Continue.

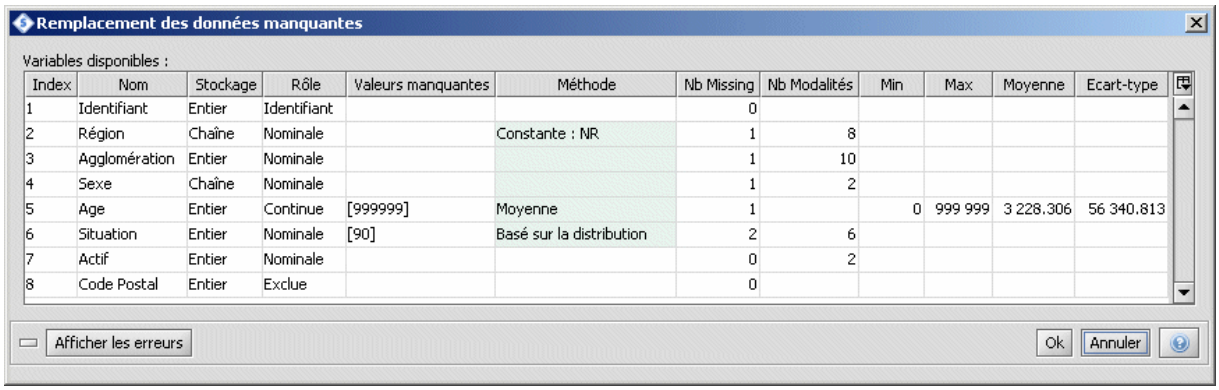
Libellés des variables créées. Par défaut, les libellés des modalités des variables d'origine deviennent les nouveaux de noms de variables.
Les cases à cocher permettent à l'utilisateur de combiner libellés de variables et de modalités d'origine en spécifiant le nombre de caractères à conserver. Le séparateur sera positionné entre le libellé de la variable d'origine et le libellé de la modalité.

Créer une variable pour les _Missing_. Cette option permet de considérer les données manquantes comme une modalité et donc de générer, pour chaque variable binarisée avec des données manquantes, une nouvelle variable binaire prenant la valeur 1 s'il s'agit d'une donnée manquante, 0 sinon.

Créer une variable pour les modalités d'effectif nul. Cette option permet d'inclure ou non les modalités d'effectif nul. Les modalités concernées n'auront que la valeur 0.

Remplacement des valeurs manquantes

Cette méthode permet de remplacer les données manquantes de l'ensemble des variables par des valeurs statistiques ou des valeurs spécifiques.
Les méthodes de remplacement diffèrent selon le type de stockage et le rôle statistique des variables.
Les données issues de cette méthode sont immédiatement utilisables par toutes les autres méthodes SPAD.



Index	Nom	Stockage	Rôle	Valeurs manquantes	Méthode	Nb Missing	Nb Modalités	Min	Max	Moyenne	Ecart-type
1	Identifiant	Entier	Identifiant			0					
2	Région	Chaîne	Nominale		Constante : NR	1	8				
3	Agglomération	Entier	Nominale			1	10				
4	Sexe	Chaîne	Nominale			1	2				
5	Age	Entier	Continue	[999999]	Moyenne	1		0	999 999	3 228.306	56 340.813
6	Situation	Entier	Nominale	[90]	Basé sur la distribution	2	6				
7	Actif	Entier	Nominale			0	2				
8	Code Postal	Entier	Exclue			0					

Figure 6-34 : Remplacement des données manquantes

La méthode édite l'ensemble des variables de la table avec leurs caractéristiques. La colonne « Nb Missing » permet immédiatement de repérer les variables avec des données manquantes : leur case dans la colonne « Méthode » est légèrement teintée en vert.

Tri du tableau. Les lignes de ce tableau peuvent être triées selon les différentes colonnes du tableau, il est particulièrement intéressent de trier ce tableau selon la colonne « Nb Missing » pour disposer immédiatement en haut du tableau des variables ayant des données manquantes. Cette opération de tri ne change pas l'ordre des variables dans la table créée.

Valeurs manquantes. Cette colonne permet de définir pour chaque variable une ou plusieurs valeurs à considérer comme manquantes. Pour accéder à la boite de dialogue de définition des valeurs manquantes, double-cliquez sur la case « Valeurs manquantes » de la variable concernée ou accédez au menu contextuel par clic droit et cliquez sur « Editer (double-clic) ». Dans le cas d'une sélection multiple de variables (obligatoirement de même type de stockage), toutes les variables se verront assigner la même liste de valeurs à considérer comme manquantes.
Pour supprimer une liste de valeurs manquantes, utilisez le clic droit et cliquez sur « Aucune ».

Méthode. Cette colonne permet de choisir pour chaque variable la méthode de remplacement des données manquantes.
Sélectionnez (Clic sur la ligne) une ou plusieurs variables de même type (touches Shift – Ctrl), puis accédez aux possibilités de traitement par clic droit sur la case « Méthode » de l'une des variables de la sélection.

Méthodes disponibles pour toutes les variables :

- **Constante à définir**. Toutes les valeurs manquantes de la ou des variables concernées sont remplacées par une constante définie par l'utilisateur. Après spécification de la valeur, le recodage apparaît sous la forme « Constante : Valeur spécifiée ».

- **Aucune** : Choix par défaut. Les données manquantes sont conservées telles quelles. Permet de supprimer un recodage préalablement défini.

Méthodes uniquement disponibles pour les variables nominales :

- **Mode**. Les données manquantes sont remplacées par la modalité la plus fréquente.

- **Basé sur la distribution** : Les données manquantes sont remplacées par une des modalités existantes, choisie en se basant sur la distribution (tirage aléatoire stratifié).

Méthodes uniquement disponibles pour les variables continues :

- **Remplacement par la moyenne**.

- **Remplacement par la moyenne robuste.** Moyenne calculée sur 95% des observations (les valeurs les plus faibles (2.5%) et les plus élevées (2.5%) sont exclues du calcul).

- **Remplacement par la médiane.**

- **Remplacement par la valeur « Demi-étendue »** : (Min+Max)/2

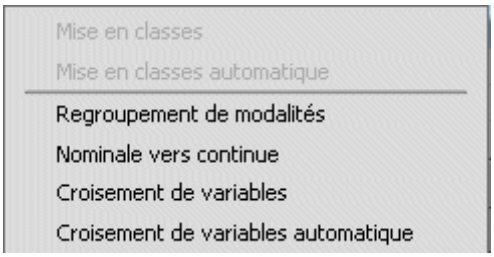
Mises en classes – Regroupements de modalités – Croisements de variables

Cette méthode permet d'effectuer les recodages suivants :

- Mise en classes de variables continues : amplitudes égales, effectifs égaux,
- Regroupement de modalités de variables nominales,
- Transformation de variables nominales en variables continues,
- Croisements de variables

La procédure permet d'effectuer un même recodage à une batterie de variables. L'utilisateur peut spécifier s'il souhaite conserver ou non les variables d'origine.

Généralités
Dans la fenêtre principale, vous disposez des variables initiales, sélectionnez la ou les variables à transformer, accédez au menu contextuel par clic droit et choisissez le type de recodage souhaité selon la nature de la sélection :



Rôle des variables sélectionnées	Type de sélection
Continues	Simple et multiple
Continues	Multiple
Nominales	Simple et multiple
Nominales	Simple et multiple
Nominales	Multiple
Nominales	Multiple

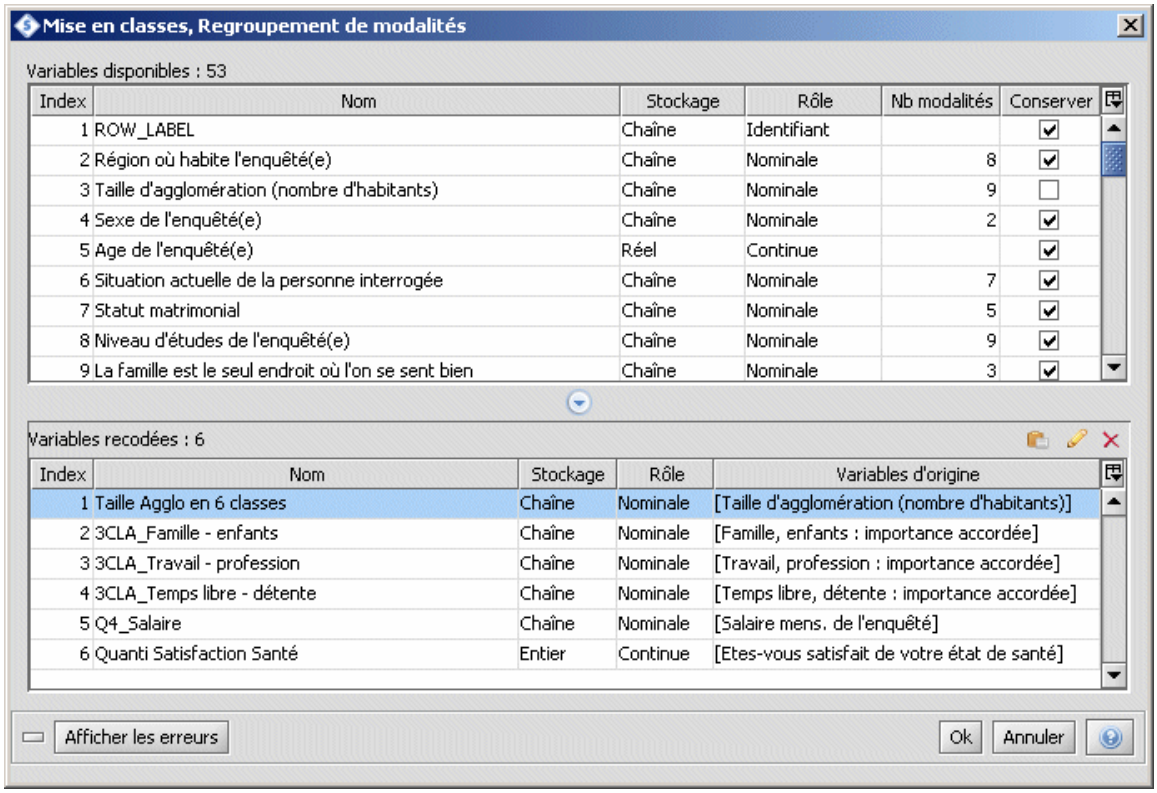


Figure 6-35 : Mises en classes, Regroupements de modalités, Croisements de variables

Une fois le recodage effectué et validé, la ou les nouvelles variables apparaissent dans la liste « Variables recodées ».
Dans cette liste, il est possible de :

- Modifier les noms (double clic dans la colonne « Nom »)
- Retournez si besoin dans le recodage (Sélection de la variable et clic sur le bouton)
- Supprimer une ou plusieurs variables (Sélection et clic sur le bouton)
- Dupliquer une variable (sélection et clic sur le bouton pour utiliser le recodage comme base d'un autre recodage)

Mise en classe

Cette méthode permet de mettre en classes une ou plusieurs variables continues selon l'algorithme des amplitudes égales ou des fréquences égales.

En cas de sélection multiple, une des variables de la sélection sera utilisée comme modèle pour déterminer les bornes inférieures et supérieures de chaque classe. Toutes les variables de la sélection seront alors découpées selon les bornes définies à partir de la variable modèle.
Assurez-vous que les distributions et/ou les échelles de valeurs de ces variables soient bien comparables. Sinon, préférez la méthode « Mise en classes automatique ».

Sélectionnez la variable ou les variables continues dans la liste « Variables disponibles », accédez au menu contextuel par clic droit et choisissez « Mise en classe ».
La fenêtre suivante apparaît :

Etape 1/2 : Paramétrage de la mise en classe

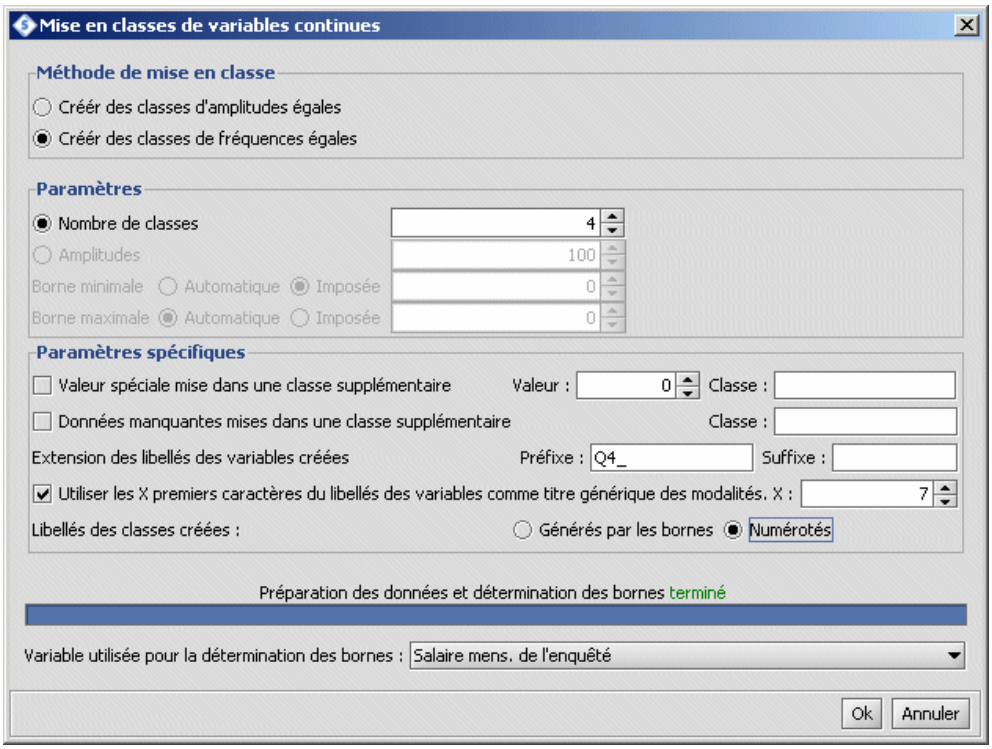


Figure 6-36 : Mise en classes d'une variable continue 1/2

Méthode de mise en classes.

- **Amplitudes égales**.
  Choisissez cette méthode pour créer des classes de même amplitude. Entrez alors le **nombre de classes** ou l'**amplitude**.
  Vous pouvez ensuite spécifier la **borne minimale** et/ou **maximale**, correspondant respectivement à la borne inférieure de la première classe et à la borne supérieure de la dernière classe. Ces valeurs doivent être respectivement inférieure ou égale et supérieure ou égale au minimum et au maximum de la variable. Si elles ne sont pas spécifiées, SPAD choisira par défaut (**Automatique**) le minimum et le maximum de la variable.
- **Fréquences égales** : choisissez cette méthode pour que les classes créées comprennent toutes le même nombre d'observations (dans la mesure du possible). Entrez alors le **nombre de classes** à créer.

Paramètres spécifiques

- **Valeur spéciale mise dans une classe supplémentaire.**
  Cochez cette case si vous souhaitez constituer une classe pour une valeur spéciale. Entrez la valeur et le nom de la classe qui sera créée.
  Cette option est utile lorsque l'on veut éliminer du calcul des fréquences égales une valeur très fréquente ayant une signification particulière.
- **Données manquantes mises dans une classe supplémentaire.**
  Cochez cette case si vous souhaitez constituer une classe pour les données manquantes. Entrez le nom de la classe qui regroupera toutes les données manquantes. Si l'option est décochée, les données manquantes sont conservées telles quelles.
- **Extension des libellés des variables créées.**
  Spécifiez un préfixe ou un suffixe qui sera respectivement inséré au début ou en fin de libellé de chaque variable d'origine pour définir le nom de la variable créée
- **Utiliser les X premiers caractères du libellé des variables comme titre générique des modalités.**
  Cochez cette case et entrez le nombre de caractères désiré si vous souhaitez que les libellés des modalités reprennent les X premiers caractères du libellé de leur variable d'origine. Ces X premiers caractères seront automatiquement insérés avant les libellés des classes présentés ci-dessous.
- **Libellé des classes créées.**
  Générés par les bornes : les libellés des classes sont générés selon le modèle suivant :
  [Borne minimale de la classe ; Borne maximale de la classe [
  Numérotés : les libellés des classes sont numérotés selon l'ordre des classes : 1, 2, 3…

Préparation des données et détermination des bornes.
Afin d'afficher la deuxième fenêtre de paramétrage de la mise en classes pour la modification éventuelle des bornes des classes, SPAD détermine automatiquement les bornes à ce niveau. La barre horizontale indique le niveau d'avancement des calculs.

Variable utilisée pour la détermination des bornes.
Uniquement disponible en cas de sélection multiple, cette méthode vous permet de spécifier la variable à utiliser comme modèle pour la détermination des bornes inférieures et supérieures de chaque classe.

Valider la première étape de la mise en classes.
Cliquez sur le bouton Ok, une deuxième fenêtre de paramétrage apparaît pour permettre la modification interactive du découpage proposé.

Etape 2/2 : Modification interactive du découpage en classes

Dans la fenêtre suivante, l'on retrouve les statistiques descriptives de la variable à mettre en classes (ou de la variable modèle dans le cadre d'une sélection multiple). On visualise dans le tableau le découpage en classes déterminé selon les paramètres de la Figure 6-36.

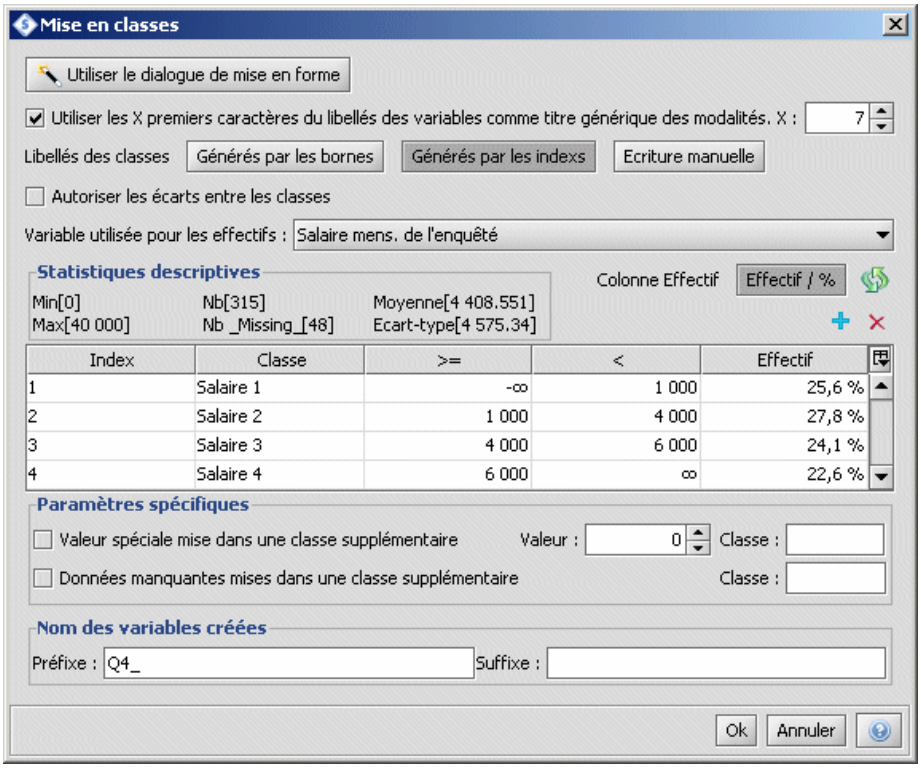
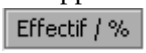


Figure 6-37 : Mise en classes d'une variable continue 2/2

Pour revenir à la fenêtre précédente, cliquez sur le bouton 

Calcul automatique des effectifs / % des classes.
Pour faire apparaître les effectifs ou pourcentages d'observations de chaque classe, cliquez sur le bouton 

pour passer des effectifs aux pourcentages (et réciproquement) puis sur le bouton  pour effectuer les calculs et afficher les résultats dans la colonne « Effectif / % » du tableau.

Modification interactive des libellés des classes.
Les libellés des classes sont affichés dans la colonne « Classe ».
Pour les modifier, on retrouve le paramètre permettant d'insérer au début du libellé de chaque classe les X premiers caractères du libellé de la variable d'origine.
Les libellés peuvent être modifiés automatiquement en cliquant sur les deux boutons suivants. Le tableau ci-dessous représente les différents résultats qui seraient obtenus pour la première classe :

	Utiliser les 7 premiers caractères du libellé de variable	
Libellés des classes	Cochée	Décochée
Générés par les bornes	Salaire [0;1 000[	[0;1 000[
Générés par les indexs	Salaire 1	1

Le bouton 

permet de modifier manuellement chaque libellé en double-cliquant dans les cases de la colonne « Classe ».

Modification interactive du découpage en classes.

- **Autoriser les écarts entre les classes.**
  Cette option permet de définir des écarts entre les bornes maximales et minimales de deux classes contigües ou consécutives.

- **Modification des bornes des classes**
  Pour modifier les bornes des classes, double-cliquez dans les cases des colonnes « >= » et « < » et entrez la valeur souhaitée. Les bornes des classes contigües sont automatiquement mises à jour selon que l'option « Autoriser les écarts entre les classes » a été cochée ou non.

- **Insérer ou supprimer une classe.**
  Pour insérer une nouvelle classe, sélectionnez une des classes existantes et cliquez sur le bouton  pour insérer la nouvelle classe avant la classe sélectionnée. Les bornes de la nouvelle classe et de la classe sélectionnée sont automatiquement recalculées.
  Pour supprimer une classe existante, sélectionnez-la et cliquez sur le bouton  . Les bornes des classes contigües à la classe supprimée sont automatiquement recalculées si l'option « Autoriser les écarts entre les classes » a été cochée.

- **Recalculer les effectifs / % des classes modifiées.**
  Pour faire apparaître les effectifs ou pourcentages des classes modifiées, cliquez sur le bouton 

  pour passer des effectifs aux pourcentages (et réciproquement) puis sur le bouton  pour effectuer les calculs et afficher les résultats dans la colonne « Effectif / % » du tableau.

On retrouve dans cette deuxième fenêtre les paramètres spécifiques déjà présentés page 106.

Valider la deuxième étape de la mise en classes.
Cliquez sur le bouton Ok pour valider.
La ou les variables mises en classes apparaissent alors dans la liste des variables recodées de la fenêtre présentée Figure 6-35.
Dans le cas d'une sélection multiple, toutes les variables de la sélection sont découpées en classes selon les bornes spécifiées précédemment. Assurez-vous que les distributions et/ou les échelles de valeurs de ces variables soient bien comparables.
Sinon, préférez la méthode « Mise en classes automatique ».

Mise en classes automatique

Cette méthode permet de mettre en classes automatiquement plusieurs variables continues selon l'algorithme des amplitudes égales ou des fréquences égales.
Chaque découpage en classes est indépendant d'une variable à l'autre. Cette méthode permet notamment de mettre en classes selon la méthode des fréquences égales une batterie de variables aux échelles de valeurs hétérogènes.
Le paramétrage de la mise en classes automatique correspond à l'étape 1/2 de la « Mise en classes », décrite Figure 6-36, page 105.
L'étape de modification interactive des découpages en classes n'est accessible qu'après validation de la mise en classe automatique.

Valider la mise en classes automatique.
Cliquez sur le bouton Ok pour valider.
SPAD réalise alors l'ensemble des calculs pour déterminer les découpages en classes de chaque variable selon les paramètres demandés.

Les variables mises en classes apparaissent alors dans la liste des variables recodées de la fenêtre présentée Figure 6-35, page 104. Pour visualiser et/ou modifier le découpage en classes d'une variable, sélectionnez-la dans la liste des variables recodées et cliquez sur le bouton  pour éditer la fenêtre de modification interactive du découpage en classes (Figure 6-37, page 107).

Regrouper les modalités d'une variable nominale

Sélectionnez la variable ou les variables dans la liste « Variables disponibles », accédez au menu contextuel par clic droit et choisissez « Regroupement de modalités ».

Dans la fenêtre « Regroupement de modalités », les modalités initiales de la variable apparaissent dans la liste de gauche dans la colonne « Valeur ». On dispose des effectifs ou des pourcentages pour chaque modalité initiale (cliquez sur le bouton « Effectifs / % » pour passer de l'un à l'autre). La colonne « Nouvelle valeur » est vide au démarrage.

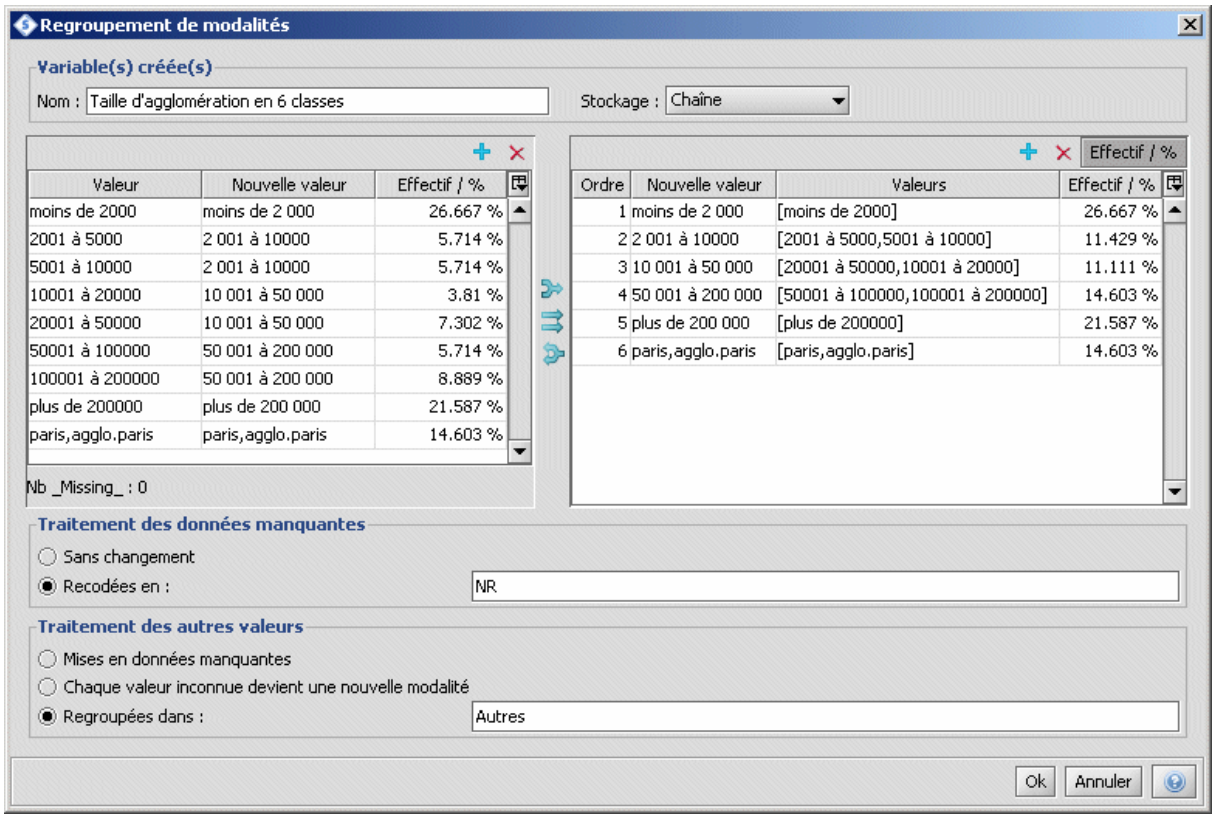


Figure 6-38 : Regroupements de modalités

De façon générale, les fonctions de tri intégrées sur les colonnes (Clic sur le titre de la colonne) des deux listes sont uniquement visuelles et ne modifient pas les données. Pour la sélection dans les listes l'usage des touches Shift et Ctrl est standard.

Regrouper des modalités initiales en une nouvelle modalité.

Sélectionnez les modalités dans la liste de gauche et cliquez sur le bouton  .
La nouvelle modalité créée apparaît dans la liste de droite reprend automatiquement le nom de la première modalité initiale comme « Nouvelle valeur ».

Conserver des modalités initiales à l'identique.

Sélectionnez les modalités dans la liste de gauche et cliquez sur le bouton  .
Les modalités apparaissent telles quelles dans la liste de droite.

Modifier une modalité créée.
Double-cliquez dans la case « Valeurs » de la modalité créée pour éditer la liste des modalités initiales à regrouper. Vous pouvez ajouter ou supprimer des modalités initiales dans la liste.

Supprimer une ou plusieurs modalités créées.

Sélectionnez la ou les modalités dans la liste de droite et cliquez sur le bouton  ou sur le bouton 

Modifier le libellé (valeur) d'une modalité créée.
Dans liste de droite, double cliquez dans la colonne « Nouvelle Valeur » de la modalité et modifiez le libellé.

Modifier l'ordre des modalités créées.
Sélectionnez la modalité à déplacer, puis faites glisser la ligne correspondante (Drag and drop) à l'endroit souhaité en maintenant le clic gauche. L'endroit du déplacement est matérialisé par une ligne verte.

Ajouter une modalité initiale.

Le bouton  de la liste de gauche permet de créer une modalité initiale qui n'existe pas dans le fichier en cours (cas d'un fichier échantillon), mais qui existera dans le fichier complet.

Supprimer la correspondance entre une modalité initiale et une modalité créée.

Le bouton  de la liste de gauche retire du regroupement défini la ou les modalités initiales sélectionnées (Utile en cas de regroupement erroné).

Ajouter / supprimer des modalités créées.

Le bouton  de la liste de droite permet de créer une modalité vide si on souhaite par exemple avoir des codages homogènes.

Le bouton  de la liste de droite supprime la ou les modalités créées sélectionnées. Les modalités initiales composantes ne sont plus liées aux modalités créées (cases « Nouvelle valeur » vides).

Les opérations de Drag and Drop.
Les opérations de Drag and drop permettent d'insérer une ou plusieurs modalités initiales dans une modalité créée existante. Sélectionnez la ou les modalités initiales, puis faites glisser la sélection sur la colonne « Valeurs » de la modalité créée.

Traitement des données manquantes.
Par défaut, les données manquantes le restent dans la variable créée (Sans changement).
Pour créer une nouvelle modalité désignant les données manquantes, choisissez l'option « Recodées en » et entrez le nom de la modalité créée.

Traitement des autres valeurs.
On entend par autres valeurs toutes les valeurs qui ne sont pas associées à l'une des modalités créées.
Les autres valeurs peuvent être mises en données manquantes.
Elles peuvent être conservées telles quelles : chaque autre valeur devenant une nouvelle modalité de la variable créée.
Elles peuvent être regroupées dans une nouvelle modalité dont le nom est à spécifier.

Valider le regroupement de modalités.
Cliquez sur le bouton Ok, la variable créée apparaît dans la liste « Variables recodées » de la fenêtre générale, son nom est modifiable.

Regrouper les modalités de plusieurs variables nominales

Sélectionnez ces variables dans la fenêtre « Variables disponibles » et accédez au menu contextuel par clic droit et choisissez « Regroupement de modalités ».

L'interface recherche toutes les valeurs différentes pour les variables initiales sélectionnées et les affichent dans la liste de gauche, quelle que soit la variable initiale que l'on visualise.

On procède au regroupement pour une variable et il s'applique automatiquement à toutes. En validant par le bouton OK, on dispose de N variables recodées dont on peut modifier les noms.

L'interface permet donc de traiter des variables codées de façon identique (même valeurs) ou non.

Exemple : On dispose de 2 variables codées : Beaucoup, Moyennement, Un peu, Pas du tout, et 2 variables codées : 1, 2, 3, 4

Sachant que d'un point de vue du contenu 1 = Beaucoup, 2 = Moyennement, 3 = Un peu, 4 = Pas du tout
On peut regrouper les modalités de ces variables en seule fois.
Dans liste des modalités, apparaitront toutes les valeurs (8) et il suffira de les regrouper en tenant compte du sens.
Par exemple :
Beaucoup, Moyennement, 1, 2 dans une modalité Oui
Un peu, Pas du tout, 3,4 dans une modalité Non

Ce principe confère une grande puissance à l'interface.

Transformer une variable nominale en variable continue.

Sélectionnez la ou les variables dans la liste « Variables disponibles », accédez au menu contextuel par clic droit et choisissez « Nominale vers continue »

Dans la fenêtre « Regroupement de modalités » les modalités initiales de la ou des variables initiales apparaissent dans la liste de gauche dans la colonne « Valeur ». On dispose des effectifs ou des pourcentages pour chaque modalité initiale (cliquez sur le bouton « Effectifs / % » pour passer de l'un à l'autre). La colonne « Nouvelle valeur » est vide.

Exemple :
Vous disposez d'une variable dont les valeurs sont : Beaucoup, Assez, un peu, pas du tout
Et vous voulez créer une variable continue qui vaut 10 si beaucoup, 8 si Assez, 4 si peu, 0 si pas du tout

Dans la liste de gauche sélectionnez toutes les modalités initiales et cliquez sur le bouton . Les modalités apparaissent dans la liste de droite avec comme nouvelles valeurs : 1 pour Beaucoup, 2 pour Assez, 3 pour Peu et 4 pour Pas dut tout.

Modifiez les nouvelles valeurs selon le codage souhaitée et validez votre recodage par Ok

L'interface et le fonctionnement est identique à un regroupement de modalités, ce qui change c'est que les nouvelles valeurs sont des nombres (Entier ou réel) par défaut de 1 à N et que l'on peut modifier. Si vous entrez des valeurs avec des décimales, vous devez choisir le stockage « Réel ».

Croisement de variables.

Cette méthode vous permet de créer une nouvelle variable nominale par le croisement de plusieurs variables nominales (au moins 2).

Sélectionnez les variables nominales à croiser dans la liste des variables initiales (voir Figure 6-35, page 104), accédez au menu contextuel par clic droit et choisissez « Croisement de variables ».
La fenêtre de paramétrage du croisement apparaît :

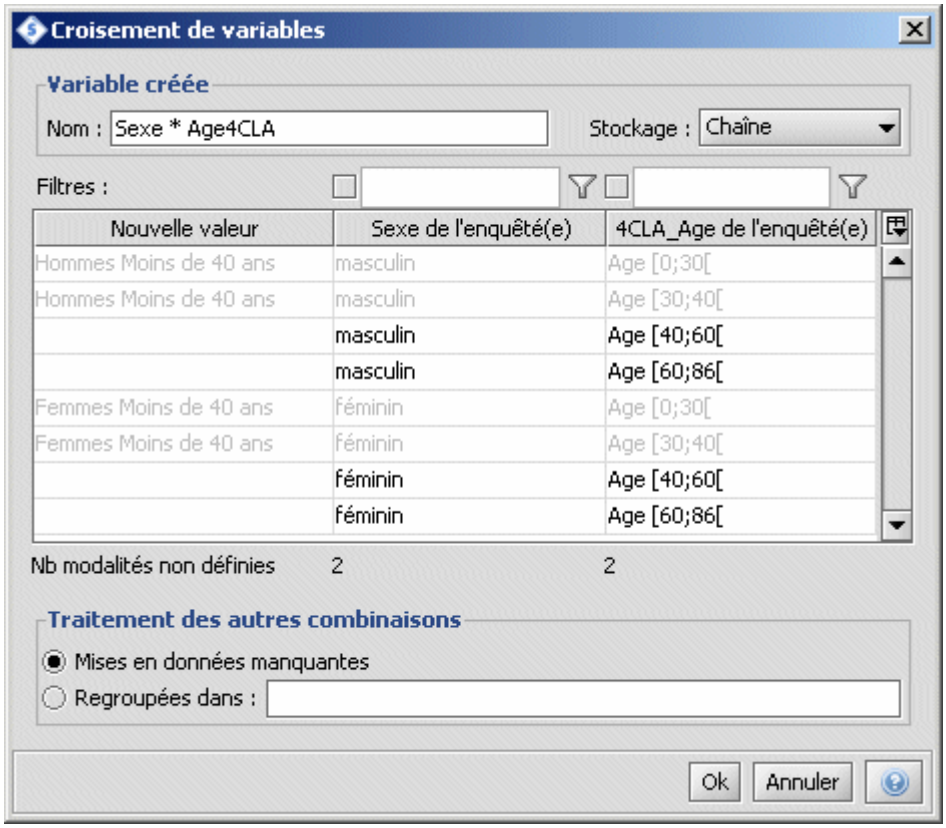


Figure 6-39 : Croisement de variables

Le tableau présenté dispose d'autant de colonnes qu'il y a de variables à croiser et d'une première colonne « Nouvelle valeur », vide au démarrage.
Il y a autant de lignes que de combinaisons distinctes (i.e. croisements) de modalités. Dans l'exemple ci-dessus, nous avons 2 modalités et 4 modalités respectivement pour le sexe de et l'âge de l'enquêté. Il y a donc 8 croisements possibles.
Le principe de l'interface est d'affecter une nouvelle valeur à chaque croisement.
Des filtres permettent de filtrer les croisements par colonne ou selon qu'une nouvelle valeur leur a déjà été assignée.

Variable créée. Nom désignant la nouvelle variable issue du croisement.
Stockage. Type de stockage de la nouvelle variable.

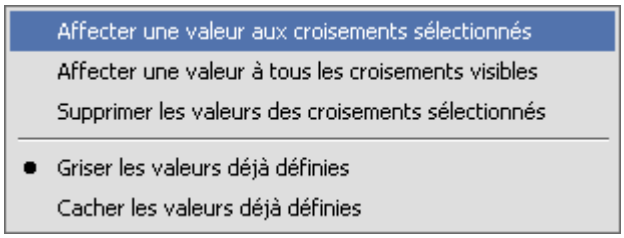
Filtre. Pour filtrer les croisements selon une variable, cochez la case pour activer le filtre et entrez les caractères permettant d'identifier les croisements à filtrer et/ou cliquez sur le bouton pour faire afficher la liste des modalités de la variable. Plusieurs filtres peuvent être utilisés simultanément.
Pour désactiver le filtre, décochez la case.

Nouvelle valeur. Pour chaque croisement, double-cliquez sur la case de la colonne « Nouvelle valeur » pour entrer le nom de la modalité créée.
Pour attribuer la même nouvelle valeur à plusieurs croisements :

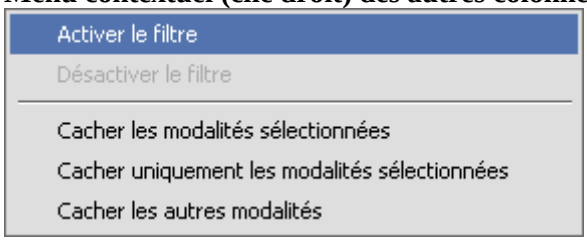
1. Sélectionnez les croisements (usage standard des touches Ctrl Shift)
2. Accéder au menu contextuel en cliquant droit dans la colonne « Nouvelle valeur » et choisissez « Affecter une nouvelle valeur aux croisements sélectionnés ».
3. Entrez la valeur désignant la modalité créée et validez.

Menu contextuel (clic droit) de la colonne « Nouvelle valeur ».

Ce menu contextuel permet de gérer les nouvelles valeurs des différents croisements, de griser ou masquer les valeurs déjà définies pour simplifier l'affichage des croisements non attribués.



Menu contextuel (clic droit) des autres colonnes.



Ce menu permet d'activer / désactiver le filtre d'une des variables du croisement (voir précédemment). Il permet aussi de filtrer selon la ou les modalités des croisements préalablement sélectionnés.

Traitement des autres combinaisons.
On entend par autres combinaisons toutes les combinaisons existantes ou futures qui ne sont pas associées à une nouvelle valeur.
Elles peuvent être mises en données manquantes ou être regroupées dans une nouvelle modalité dont le nom est à spécifier.

Générateur de nouvelles variables

Cette méthode permet de générer de nouvelles variables à partir des variables existantes et des fonctions de la bibliothèque.
L'interface de cette méthode ne nécessite pas de programmation.

Toutes les transformations des plus simples au plus complexes sont possibles.
La méthodologie d'utilisation est séquentielle. Toute transformation complexe est décomposée en une suite de transformations élémentaires, ordonnées, qui permettent d'aboutir au résultat voulu.
Les variables générées peuvent être réutilisées pour calculer de nouvelles variables.

L'utilisateur peut spécifier s'il souhaite conserver ou non les variables d'origines ainsi que tout ou partie des variables générées.

Bibliothèque de fonctions

Double-cliquez sur la fonction choisie pour la faire apparaître dans la partie inférieure ou glissez-déposez pour l'insérer à l'endroit voulu.

Liste des variables initiales et créées

Modifiez les noms, stockages et rôles des variables créées en double-cliquant dans les cases correspondantes.
Cochez-décochez les cases selon que vous souhaitez conserver ou non certaines variables d'origine ou créées.

Métadonnées

La sélection d'une variable initiale dans la liste à gauche affiche dans cette partie les métadonnées de cette variable. L'onglet « Modalités » est activé pour les variables nominales.

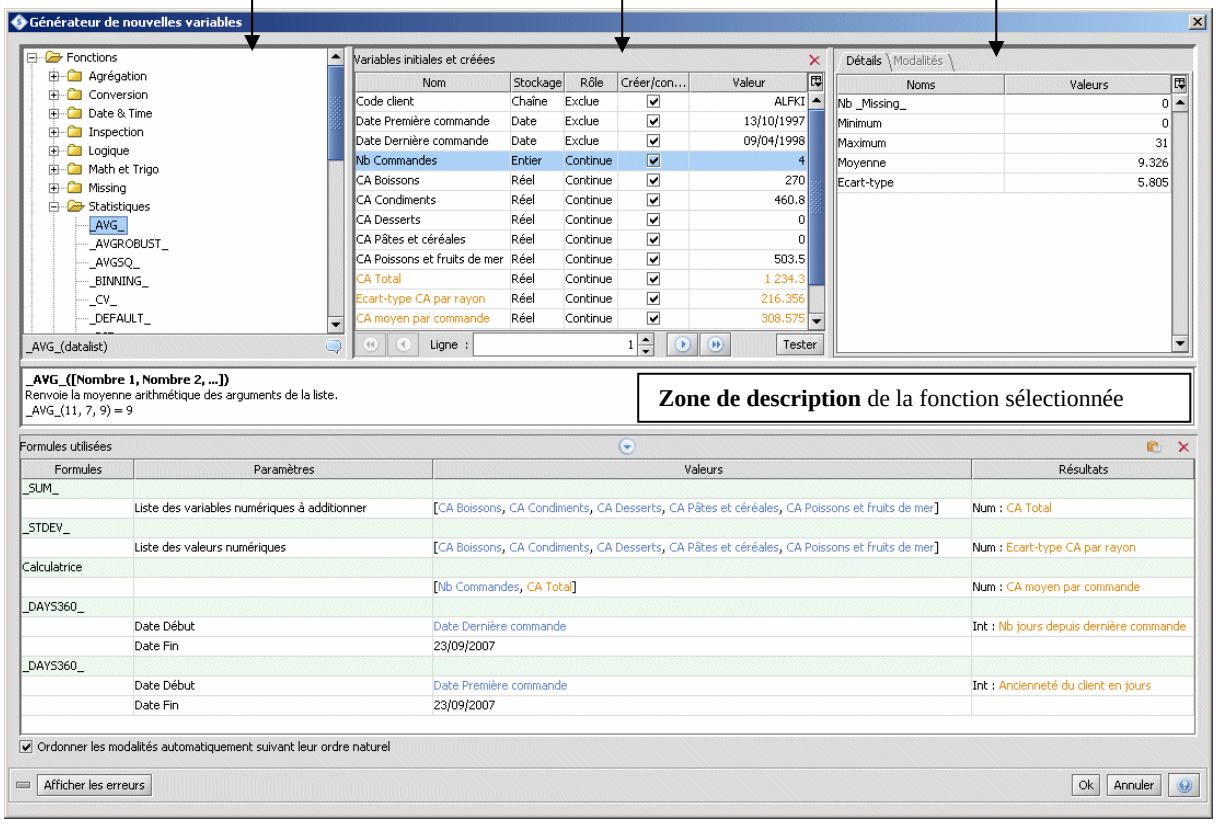


Figure 6-40 : Interface du générateur de nouvelles variables

Fonctions – Aide
Les fonctions sont classées par grandes rubriques dans la partie supérieure gauche de la fenêtre. Pour obtenir de l'aide sur une fonction, il suffit de la sélectionner. La zone de description se met automatiquement à jour. Pour plus d'informations, voir page 118.

Ajouter une fonction.
L'ajout d'une fonction dans la liste des transformations successives se fait soit par double-clic sur la fonction soit à l'aide du bouton  .

Comportement du bouton 
Ce bouton permet également d'ajouter une ou plusieurs fonctions (usage standard des touches Ctrl Shift) et de renseigner leur premier paramètre si des variables initiales et/ou créées ont été également sélectionnées.
Le comportement diffère selon que la fonction accepte en premier paramètre une liste de valeurs ou non.
La fonction apparaît immédiatement dans la liste des transformations successives. La nouvelle variable, associée à cette fonction, apparaît dans la liste des variables initiales et créées. Son nom a été automatiquement généré et est affiché dans la colonne « Résultats » du tableau des transformations.
Le tableau suivant résume les différentes possibilités proposées par ce bouton :

	Type du premier paramètre	P variables sélectionnées
N fonctions sélectionnées	Valeur	N * P
N fonctions sélectionnées	Liste de valeurs	N

Dans le cas où on sélectionne P variables et N fonctions admettant comme premier paramètre une valeur et non une liste de valeurs, N * P transformations sont alors ajoutées dans la liste des transformations successives.

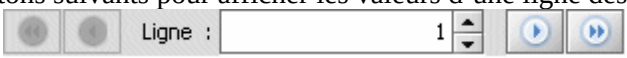
Insérer une fonction.
Pour insérer une fonction à un endroit précis, sélectionnez-la et glissez/déposez à l'endroit choisi.

Paramétrer une fonction.
Les paramètres de chaque fonction sont décrits dans la colonne « Paramètres » du tableau.
Le paramétrage se fait soit par glisser/déposer soit par double-clic dans la case de la colonne « Valeurs ».
Lorsque le paramètre attend une ou plusieurs variables (initiales et/ou créées), sélectionnez la ou les variables concernées et glissez/déposez la sélection dans la case de la colonne « Valeurs » correspondante ou double-cliquez dans la case de la colonne valeur pour accéder à une fenêtre de sélection de variables.
Avant le paramétrage, la case de la colonne valeur indique le type de paramètre attendu (Déposer Réel, Déposer Entier, Déposer Chaine…) et la nature du paramètre (valeur unique ou liste de valeurs).
La présence de crochets dans l'instruction « Déposer [Réel, …] » indique que le paramètre attend une liste de valeurs ou variables.
Dans l'exemple précédent, les fonctions _SUM_ et _STDEV_ admettent une liste de réels comme paramètre.
Le tableau suivant résume les correspondances de types acceptées :

Types acceptés / Type attendu	Entier	Réel	Chaine	Date	Autre
Déposer Entier	OK	⃠	⃠	⃠	⃠
Déposer Réel	OK	OK	⃠	⃠	⃠
Déposer Chaine	OK	OK	OK	OK	OK
Déposer Date	⃠	⃠	⃠	OK	⃠
Déposer	OK	OK	OK	OK	OK

Les outils « Calculatrice » et « Assistant logique » sont présentés respectivement pages 130 et 131.

Tester une fonction.
Cette fonctionnalité vous permet de tester interactivement l'ensemble de vos transformations.
Cliquez sur l'un des boutons suivants pour afficher les valeurs d'une ligne des données sources.



Les valeurs apparaissent dans la colonne « Valeurs » de la liste « Variables initiales et créées ».Pour tester l'ensemble des calculs et fonctions utilisés, cliquez sur le bouton **Tester**.
Les valeurs-résultats apparaissent dans les cases des variables créées.

Déplacer une fonction.
Sélectionnez la première ligne de la fonction dans la liste des transformations successives et glissez/déposez la fonction à l'endroit choisi, matérialisé par une ligne bleue.

Supprimer une fonction.
Sélectionnez la fonction dans la liste des calculs et cliquez sur le bouton .

Renommer une variable créée.
Double-cliquez sur le nom de la variable concernée soit dans la liste « Variables initiales et créées » soit dans la colonne « Résultats » de la « liste des calculs » et tapez le nouveau nom.

Supprimer une variable créée.
Sélectionnez la variable dans la liste « Variables initiales et créées » et cliquez sur le bouton .

Attention : renommer ou supprimer une variable créée entraîne des erreurs si cette variable est utilisée dans d'autres fonctions. La fonction liée à la variable supprimée n'est pas supprimée.

Modifier le stockage et le rôle d'une variable créée.
Par défaut, SPAD affecte un type de stockage et un rôle à toute variable créée. Pour les modifier, double-cliquez sur le stockage ou sur le rôle de la variable concernée dans la liste « Variables initiales et créées » et choisissez dans la liste déroulante.

Définir une variable créée comme temporaire.
Si certaines variables, utilisées comme étapes intermédiaires au calcul, n'ont pas d'intérêt pour la suite de l'analyse, décochez la case à cocher de la colonne « Clé/Créer ? » dans la vue « Variables initiales et créées » pour qu'elles n'apparaissent pas dans les données produites.

Ordonner les modalités automatiquement selon leur ordre naturel.
Cochée, cette option trie les modalités d'une nouvelle variable nominale dans l'ordre alphanumérique ou numérique selon qu'elle est stockée comme chaine de caractères ou comme entier. Cet ordre est repris dans les méthodes statistiques pour l'affichage des résultats. En décochant cette option, l'ordre obtenu est celui dans lequel les modalités apparaissent dans les données.

Création d'une variable de pondération

Cette procédure crée une variable de pondération destinée à "redresser" l'échantillon. Le poids ou coefficient de redressement est calculé pour assurer une certaine distribution (appelée "distribution demandée") dans une ou plusieurs variables nominales du fichier. Le calcul des poids est itératif ; les distributions obtenues sont des approximations des distributions demandées, et la somme des poids est ajustée in fine à la taille de l'échantillon.

La méthode utilisée est une variante de la méthode "RAS".

La procédure fournit en premier lieu les résultats généraux : pour chaque variable utilisée, on obtient les distributions observées, demandées et calculées. De façon optionnelle, on peut obtenir une édition des coefficients calculés pour chaque individu.

L'interface vous permet de définir les distributions (pourcentages) théoriques pour la ou les variables pour lesquelles vous souhaitez ajuster les distributions.
Note: Vous ne pouvez utiliser une variable contenant des données manquantes.

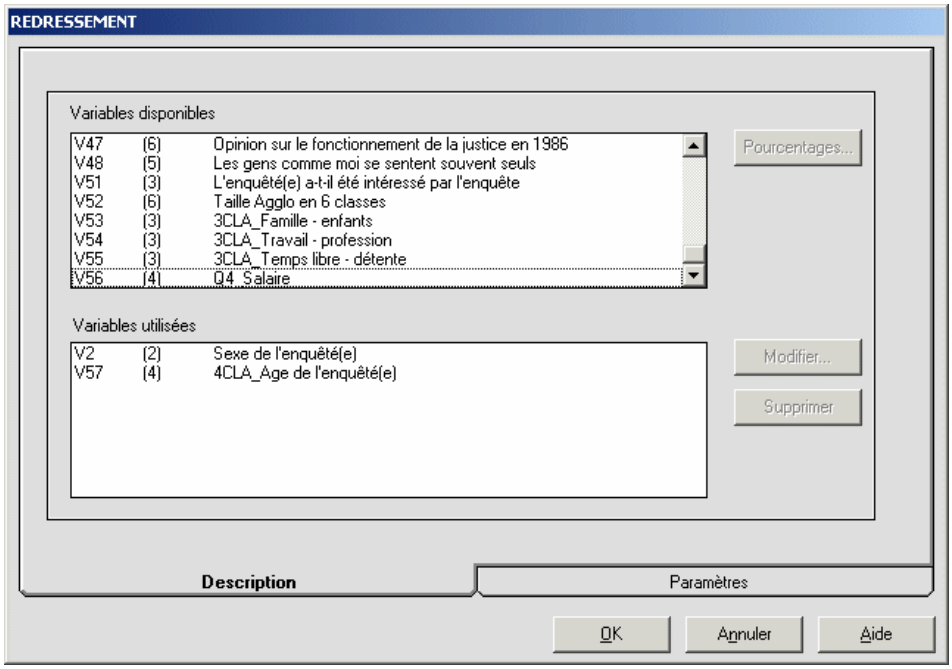


Figure 6-41 : Création d'une variable de pondération

Utiliser une variable pour la pondération.
Dans la liste "Variables disponibles", choisissez la variable, puis cliquez sur le bouton Pourcentages...
La fenêtre "Définition des pourcentages théoriques" s'affiche avec les pourcentages observés pour chaque modalité. Entrez les pourcentages théoriques (strictement supérieurs à 0) de chaque modalité puis validez en cliquant sur le bouton OK.
La variable utilisée apparaît alors dans la liste "Variables utilisées".
Procédez ainsi de suite pour chaque variable.

Supprimer une variable utilisée.
Sélectionnez la variable dans la liste "Variables utilisées", puis cliquez sur le bouton supprimer.

Modifier les pourcentages théoriques d'une variable utilisée.
Sélectionnez la variable dans la liste "Variables utilisées", puis cliquez sur le bouton modifier.

Agrégation – Générateur de variables : liste des fonctions par rubrique

Les fonctions suivantes sont utilisables dans les méthodes « Agrégation » et « Générateur de nouvelles variables ».
Elles sont présentées par rubrique dans l'ordre alphabétique.
Les fonctions de la rubrique Agrégation ne sont utilisables que dans la méthode du même nom.

Agrégation

AG_AVG (Variable numérique)
Renvoie la moyenne arithmétique de Variable numérique pour chaque groupe.

AG_AVGROBUST (Variable numérique, k-left, k-right)
Renvoie la moyenne arithmétique "robuste" de Variable numérique pour chaque groupe. Elle est calculée sur l'ensemble des observations, moins les queues de distribution définies en % (0-1) à gauche k-left et à droite k-right.

AG_CONCATENATE(Variable, Séparateur)
Renvoie la concaténation de l'ensemble des valeurs de Variable, éventuellement séparées par un séparateur, pour chaque groupe.

AG_CORR(X, Y)
Renvoie le coefficient de corrélation entre les variables numériques X et Y pour chaque groupe.

AG_COUNT()
Renvoie le nombre d'observations pour chaque groupe. Les données manquantes sont comptabilisées.

AG_COUNTDISTINCT(Variable)
Renvoie le nombre de valeurs distinctes de Variable pour chaque groupe.

AG_COUNTWHERE(Condition)
Renvoie pour chaque groupe le nombre d'observations qui vérifient la Condition.

AG_COV(X, Y)
Renvoie la covariance, moyenne des produits des écarts pour chaque série d'observations, entre les variables numériques X et Y pour chaque groupe.

AG_CV(Variable numérique)
Renvoie le coefficient de variation de Variable numérique pour chaque groupe.

AG_DATEMAX (Variable Date)
Renvoie la date la plus récente de Variable Date pour chaque groupe.

AG_DATEMIN (Variable Date)
Renvoie la date la plus ancienne de Variable Date pour chaque groupe.

AG_DATE_AVGFREQ(Variable date)
Renvoie la fréquence moyenne en nombre de jours des évènements survenus aux dates consécutives de Variable date pour chaque groupe.

AG_DATE_STDEVFREQ(Variable date)
Renvoie l'écart-type des écarts observés entre les dates consécutives deux-à-deux de Variable date pour chaque groupe.
Un écart-type de 0 indique une régularité parfaite entre toutes les dates consécutives.

AG_EVALWHERE(Condition, Variable)
Renvoie pour chaque groupe la première valeur de Variable qui vérifie la Condition.

AG_FIND(Nième, Condition)
Renvoie l'index du Nième enregistrement dans le groupe qui vérifie la condition.

AG_FIRNOMISS(Variable)
Renvoie la première valeur renseignée (non manquante) de Variable pour chaque groupe.

AG_FREQ(Variable, Valeur)
Renvoie pour chaque groupe le pourcentage d'observations qui présentent la modalité Valeur pour la Variable.

AG_GETQUANTILE(Variable, Nb)
Renvoie sous la forme d'une chaine la liste des Nb quantiles de la Variable séparés par une virgule.
Cette liste peut être utilisée par la fonction _BINNING_ de la rubrique Statistique.

AG_MAX (Variable numérique)
Renvoie la valeur maximale de Variable numérique pour chaque groupe.

AG_MEDIAN (Variable numérique)
Renvoie la médiane de Variable numérique observée dans chaque groupe.

AG_MIN (Variable numérique)
Renvoie la valeur minimale de Variable numérique pour chaque groupe.

AG_MODE(Variable)
Renvoie la valeur de Variable la plus représentée dans chaque groupe.

AG_PERCENTIF(Condition)
Renvoie pour chaque groupe le pourcentage d'observations qui vérifient la Condition.

AG_PERCENTILE (Variable numérique, k)
Renvoie le k-ième percentile de Variable numérique pour chaque groupe.
AG_PERCENTILE(Durée des appels, 0.25) renvoie le 1er quartile de la durée des appels pour chaque groupe.

AG_RANGE (Variable numérique,)
Renvoie l'étendue (Max - Min) de Variable numérique pour chaque groupe.

AG_STDEV (Variable numérique)
Renvoie l'écart-type de Variable numérique pour chaque groupe.

AG_SUM (Variable numérique)
Renvoie la somme des valeurs de Variable numérique pour chaque groupe.

Conversion

TO_FLOAT_(Valeur)
Renvoie la valeur au type Float (Réel) si cela est possible.

TO_INT_(Valeur)
Renvoie la valeur au type Integer (Entier) si cela est possible.

TO_STRING_(Valeur)
Renvoie la valeur au type string (Chaîne de caractères) si cela est possible.

Date & Time

DATE_TO_NBDAYS(Date)
Renvoie la date sous la forme d'un entier représentant le nombre de jours écoulés entre le 1er janvier 1900 et Date.

DATE(Année, Mois, Jour)
Renvoie la date de l'année Année, du mois Mois et du jour Jour.

_DATE_AVGFREQ_([Date1, Date2, Date3...])
Renvoie la fréquence moyenne en nombre de jours des évènements survenus aux dates consécutives [Date1, Date2, Date3...].

_DATE_MAX_([Date1, Date2, Date3...])
Renvoie la date la plus récente de [Date1, Date2, Date3...].

_DATE_MIN_([Date1, Date2, Date3...])
Renvoie la date la plus ancienne de [Date1, Date2, Date3...].

_DATE_RANGE_([Date1, Date2, Date3...])
Renvoie l'étendue en nombre de jours entre la date la plus ancienne et la date la plus récente de [Date1, Date2, Date3...].

_DATE_SPLIT_(Date)
Décompose la date Date et renvoie au choix : un booléen indiquant s'il s'agit d'une date, l'année, le mois, le jour, l'heure, la minute, la seconde, le n° de semaine.

_DATE_STDEVFREQ_([Date1, Date2, Date3...])
Renvoie l'écart-type des écarts observés entre les dates consécutives deux-à-deux [Date1, Date2, Date3...].
Un écart-type de 0 indique une régularité parfaite entre toutes les dates consécutives.

_DATE_TO_NUMBER_(Date, Format_Sortie)
Convertit la Date en nombre selon le format Format_Sortie.
_DATE_TO_NUMBER_(12/06/2007, YMD) = 20070612

_DATE_TO_STRING_(Date, Format_Sortie)
Renvoie la Date comme chaîne de caractères selon le format Format_Sortie.

DAYS360(Date_Début, Date_Fin)
Renvoie le nombre de jours compris entre deux dates sur la base d'une année de 360 jours (12 mois de 30 jours).

DAY(Date)
Renvoie le numéro du jour de la Date.

EDATE(Date_départ, M)
Renvoie la date M mois après Date_départ si M est positif, M mois avant sinon.

EOMONTH(Date, Nb)
Renvoie la date de fin de mois Nb mois avant ou après la Date selon que Nb est négatif ou positif.

HOUR(Date)
Renvoie l'heure de la date/heure Date, comprise entre 0 et 23.

MINUTE(Date)
Renvoie le numéro de la minute de la Date, compris entre 0 et 59.

MONTH(Date)
Renvoie le numéro du mois de la Date, compris entre 1 et 12.

_NBDAYS_TO_DATE_(Nombre de jours depuis 1900)
Renvoie une date exprimée en nombre de jours depuis le 1er janvier 1900sous la forme d'une Date au format date.

NOW()
Renvoie la date et l'heure courantes du système.

_NUMBER_DATE_TO_STRING_(Number_Date, Format_Entrée, Format_Sortie)
Lit la date Number_Date selon le format Format_Entrée et renvoie une date au format Format_Sortie en chaîne de caractères.
_NUMBER_DATE_TO_STRING_(20070512, YMD, %DD% %MMM% %YYYY%) = 01 avr 2007

_NUMBER_TO_DATE_(Number_Date, Format_Entrée)
Lit la date Number_Date stockée comme nombre selon le format Format_Entrée et la renvoie au format Date. En option, renvoie un booléen qui vérifie si Number_Date est conforme à Format_Entrée

QUARTER(Date)
Renvoie le numéro du trimestre de Date.
QUARTER(23/07/2007) = 3

SECOND(Date)
Renvoie le numéro de la seconde de la Date, compris entre 0 et 59.

SEMESTER(Date)
Renvoie le numéro du semestre de Date.
SEMESTER(23/07/2007) = 2

_STRING_DATE_TO_STRING_(String_Date, Format_Entrée, Format_Sortie)
Lit la date String_Date selon le format Format_Entrée et renvoie une date au format Format_Sortie en chaîne de caractères.
_STRING_DATE_TO_STRING_(4-06-07, DMY, %DD%/%MM%/%YYYY%) = 04/06/2007

_STRING_TO_DATE_(String_Date, Format_Entrée)
Lit la chaine de caractères String_Date selon le format Format_Entrée et la renvoie convertie en date. En option, renvoie un booléen indiquant s'il s'agit bien d'une date.
_STRING_TO_DATE_(6 avril 2007, DMY) = 06/04/2007 selon le format d'affichage des dates.

TIMESPAN(J, H, M, S)
Renvoie la durée représentée par le jour J, l'heure H, la minute M et la seconde S comme une fraction de la journée.

TIME(Heure, Minutes, Secondes)
Renvoie le nombre décimal de l'heure représentée par Heure:Minutes:Secondes.
Le nombre décimal renvoyé par la fonction _TIME_ est une valeur comprise entre 0 (zéro) et 0,99999999, qui représente l'heure, de 0:00:00 (12:00:00 AM) à 23:59:59 (11:59:59 PM).

TODAY()
Renvoie la date courante du système.

WEEKDAY(Date)
Renvoie un chiffre compris entre 1 (lundi) et 7 (dimanche) désignant le jour de la semaine.

WEEK(Date)
Renvoie le numéro d'ordre de la semaine dans l'année correspondant à la date Date, selon la norme ISO.

YEARFRAC(Date_Début, Date_Fin)
Renvoie une fraction correspondant au nombre de jours séparant Date_Début de Date_Fin par rapport à une année complète.

YEAR(Date)
Renvoie l'année de la Date.

Inspection

ISEMPTYSTRING(Valeur)
Renvoie 1 si Valeur est une chaine vide, 0 sinon.
ISEMPTYSTRING(' ') = 1

ISMISSING(Valeur)
Renvoie 1 si Valeur vaut _Missing_ (i.e. Manquante), 0 sinon.

ISNONE(Valeur)
Renvoie 1 si Valeur vaut None (i.e. Manquante ou Null), 0 sinon.

ISNOTEMPTYSTRING(Valeur)
Renvoie 1 si Valeur est une chaine non vide, 0 sinon.
ISNOTEMPTYSTRING('Bonjour') = 1

ISNUMBER(Valeur)
Renvoie 1 si Valeur est un nombre, 0 sinon.
ISNUMBER('100') = 1

ISSTRING(Valeur)
Renvoie 1 si Valeur est une chaîne de caractère, 0 sinon.
ISSTRING('Bonjour') = 1

OUTOFRANGE(Valeur numérique, Borne Min, Borne Max)
Renvoie 1 si la Valeur numérique sort de l'intervalle [Borne Min; Borne Max], 0 si Valeur appartient à l'intervalle.
En option, renvoie le type de problème constaté.

Logique

AND(Condition 1, Condition 2, Condition 3, ...)
Renvoie 1 si toutes les conditions sont vérifiées, 0 sinon.
AND([2>1, 'A'=='A', 2

_EQ_(Valeur 1, Valeur 2)
Renvoie 1 si Valeur 1 et Valeur 2 sont égales, 0 sinon.

_GE_(Valeur 1, Valeur 2)
Renvoie 1 si Valeur 1 est supérieure ou égale à Valeur 2, 0 sinon.

_GT_(Valeur 1, Valeur 2)
Renvoie 1 si Valeur 1 est strictement supérieure à Valeur 2, 0 sinon.

_IF_(Condition, Valeur_si_vrai, Valeur_si_faux)
Renvoie Valeur_si_vrai si la Condition est vérifiée, Valeur_si_faux sinon.

_IN_(Valeur cherchée, [Valeur1, Valeur2, ...])
Renvoie 1 si Valeur cherchée est présente dans la liste de valeurs, 0 sinon.

_LE_(Valeur 1, Valeur 2)
Renvoie 1 si Valeur 1 est inférieure ou égale à Valeur 2, 0 sinon.

_LT_(Valeur 1, Valeur 2)
Renvoie 1 si Valeur 1 est strictement inférieure à Valeur 2, 0 sinon.

NAND(Condition 1, Condition 2, Condition 3, ...)
Renvoie 1 si au moins une condition n'est pas vérifiée, 0 sinon.

_NE_(Valeur 1, Valeur 2)
Renvoie 1 si Valeur 1 et Valeur 2 sont différentes, 0 sinon.

NOR(Condition 1, Condition 2, Condition 3, ...)
Renvoie 1 si aucune des conditions n'est vérifiée, 0 sinon.
NOR([2>1, 'A'=='B', 2

NOT(Condition)
Renvoie 1 si Condition vaut 0, 0 sinon.

_OR_(Condition 1, Condition 2, Condition 3, ...)
Renvoie 1 si au moins une condition est vérifiée, 0 sinon.
_OR_([2>1, 'A'=='B', 2

XOR(Condition 1, Condition 2, Condition 3, ...)
Renvoie 1 si une et une seule des conditions est vérifiée, 0 sinon.
XOR([2>1, 'A'=='B', 2

Math & Trigo

ABS(Nombre)
Renvoie la valeur absolue de Nombre.
ABS(-2) = 2

ACOS(Nombre)
Renvoie l'arccosinus de Nombre. L'arccosinus, ou inverse du cosinus, est l'angle dont le cosinus est l'argument Nombre.
L'angle renvoyé, exprimé en radians, est compris entre 0 (zéro) et pi.

ASIN(Nombre)
Renvoie l'arcsinus ou le sinus inverse de Nombre. L'arcsinus est l'angle dont le sinus est l'argument Nombre.
L'angle renvoyé, exprimé en radians, est compris entre -pi/2 et pi/2.

ATAN2(Nombre X, Nombre Y)
Renvoie l'arctangente ou la tangente inverse des coordonnées X et Y spécifiées. L'arctangente est l'angle formé par l'axe des abscisses (x) et une droite passant par l'origine (0, 0) et un point dont les coordonnées sont (Nombre X, Nombre Y).
Cet angle, exprimé en radians, est compris entre -pi et pi, -pi non compris.

ATAN(Nombre)
Renvoie l'arctangente ou la tangente inverse d'un Nombre. L'arctangente est l'angle dont la tangente est l'argument Nombre.
L'angle renvoyé, exprimé en radians, est compris entre -pi/2 et pi/2.

CEILING(Nombre, Précision)
Renvoie Nombre après l'avoir arrondi au multiple de l'argument Précision en s'éloignant de zéro.
CEILING(2.5, 1) = 3

COSH(Nombre)
Renvoie le cosinus hyperbolique de Nombre.

COS(Nombre)
Renvoie le cosinus de l'angle Nombre, exprimé en radians.
Si l'angle est mesuré en degrés, multipliez-le par PI()/180 ou utilisez la fonction _RADIANS_ pour le convertir en radians.

CUMSUM(Variable, By)
Renvoie la somme cumulative de Variable, pour chaque groupe défini par les valeurs de la variable By. Les données doivent être déjà triées selon la variable By.

DEGREES(Angle)
Convertit l'angle exprimé en radians Angle en degrés.

EXP(Nombre)
Renvoie la constante e élevée à la puissance de l'argument Nombre.
La constante e est égale à 2,71828182845904, soit la base du logarithme népérien.

FACT(Nombre)
Renvoie la factorielle de Nombre.
La factorielle de l'argument nombre est égale à 1*2*3*...*Nombre.

FLOOR(Nombre)
Renvoie le plus grand entier immédiatement inférieur ou égal à Nombre.
FLOOR(- 3.78) = - 3

FRAC(Nombre)
Renvoie la partie décimale de Nombre.

INT(Nombre)
Renvoie la partie entière de Nombre.
INT(- 3.78) = - 3

_LN_(Nombre)
Renvoie le logarithme népérien du Nombre. _LN_ est la fonction réciproque de la fonction

EXP(Nombre)
Renvoie la constante e élevée à la puissance de l'argument Nombre.
La constante e est égale à 2,71828182845904, soit la base du logarithme népérien.

LOG10(Nombre)
Renvoie le logarithme en base 10 de Nombre.

LOG(Nombre, Base)
Renvoie le logarithme en base en base Base de Nombre.

MODULUS(Nombre, Diviseur)
Renvoie le reste de la division Nombre par Diviseur. Le résultat est du même signe que Diviseur.
MODULUS(-18, 11) = 4

PERMUT(n, p)
Renvoie le nombre de combinaisons de p éléments que l'on peut former à partir de n éléments.

_PI_
Renvoie la valeur 3,14159265358979, la constante mathématique pi.

POWER(Nombre, Puissance)
Renvoie la valeur de Nombre élevé à la puissance Puissance.
POWER(5, 2) = 25

PRODUCT([Nombre 1, Nombre 2, ...])
Renvoie le produit de l'ensemble des nombres de la liste Nombre 1, Nombre 2, ...
PRODUCT([2, 3, 6]) = 36

QUOTIENT(Numérateur, Dénominateur)
Renvoie la partie entière de la division de Numérateur par Dénominateur.
QUOTIENT(18, 5) = 3

RADIANS(Angle)
Convertit l'angle exprimé en degrés Angle en radians.

ROUND(Nombre, N)
Renvoie l'arrondi de Nombre à N décimales.
ROUND(2.149, 1) = 2.1

SINH(Nombre)
Renvoie le sinus hyperbolique de Nombre.

SIN(Nombre)
Renvoie le sinus de l'angle Nombre, exprimé en radians.
Si l'angle est mesuré en degrés, multipliez-le par PI()/180 ou utilisez la fonction _RADIANS_ pour le convertir en radians.

SQRT(Nombre positif)
Retourne la racine carré du Nombre positif.
SQRT(100) = 10

SUMSQ([Nombre 1, Nombre 2, ...])
Renvoie la somme de l'ensemble des nombres élevés au carré de la liste Nombre 1, Nombre 2, …
SUMSQ([2, 3]) = 13

SUM([Nombre 1, Nombre 2, ...])
Renvoie la somme de l'ensemble des nombres de la liste Nombre 1, Nombre 2, ...
SUM([2, 3, 6]) = 11

TANH(Nombre)
Renvoie la tangente hyperbolique de Nombre.

TAN(Nombre)
Renvoie la tangente de l'angle Nombre, exprimé en radians.
Si l'angle est mesuré en degrés, multipliez-le par PI()/180 ou utilisez la fonction _RADIANS_ pour le convertir en radians.

VECLEN([Nombre 1, Nombre 2, ...])
Renvoie la racine carrée de la somme des nombres élevés au carré de la liste Nombre 1, Nombre 2, ...
VECLEN([2, 3]) = 3.606

Statistique

AVGROBUST([Nombre 1, Nombre 2, ...], %gauche, %droite)
Renvoie la moyenne arithmétique des arguments de la liste en excluant les queues de distribution définies entre 0 et 1 par %gauche et %droite.

AVGSQ([Nombre 1, Nombre 2, ...])
Renvoie la moyenne arithmétique des carrés de ses arguments.
AVGSQ(4, 5, 2) = 15

AVG([Nombre 1, Nombre 2, ...])
Renvoie la moyenne arithmétique des arguments de la liste.
AVG(11, 7, 9) = 9

BINNING(Valeur, [borne1, borne2, ...], Minmax)
Renvoie la classe d'appartenance [borne inf; borne sup[de Valeur à partir des bornes spécifiées. Minmax=1 indique que les bornes minimales et maximales définissent la zone d'acceptation des valeurs. Minmax=0 (défaut) indique qu'il s'agit de bornes intermédiaires.
BINNING(25, [20, 40, 60, 80], 0) = [20; 40[

_COUNT_IF_([Valeur 1, Valeur 2, ...], Critère)
Renvoie le nombre de valeurs qui vérifient le critère spécifié.
Le critère comporte en règle générale deux éléments : un opérateur de comparaison suivi de la valeur.
Les opérateurs de comparaison utilisables sont : =, <>, !=, <, <=, >, >=
L'absence d'opérateur équivaut à l'opérateur =.
_COUNT_IF_(["oui","non","non"], oui) = 1 _COUNT_IF_([25, 45, 65], >=32) = 2

_CV_([Nombre 1, Nombre 2, ...])
Renvoie le coefficient de variation des arguments de la liste.
_CV_(11, 7, 9) = _STDEV_(11, 7, 9) / _AVG_(11, 7, 9) = 0.1814

DEFAULT(Variable, Valeur, liste de valeurs à remplacer)
Remplace les données manquantes et les valeurs de Liste de valeurs à remplacer par Valeur pour le champ numérique Variable.
Les valeurs doivent être séparées par des virgules dans la liste.

DIF(Variable, N, By)
Renvoie la différence entre la valeur courante et la N-ième valeur précédente de variable, dans le groupe défini par By.
Les données doivent être déjà triées selon la variable By.

FIRNOMISS([Var1, Var2, Var3, ...])
Renvoie le rang de la première valeur renseignée (non manquante) de la liste d'arguments.

FIRNOZ([Var1, Var2, Var3, ...])
Retourne le rang de la première valeur non nulle (différente de zéro) de la liste d'arguments.
FIRNOZ(0,0,1,0) = 3

GEOMEAN([Nombre 1, Nombre 2, ...])
Renvoie la moyenne géométrique des nombres positifs de la liste.
GEOMEAN(11, 7, 9) = 8,849

HARMEAN([Nombre 1, Nombre 2, ...])
Renvoie la moyenne harmonique des arguments de la liste.
HARMEAN(11, 2, 9) = 4.273

INDEXBY(By)
Renvoie l'index de la ligne (à partir de 0), dans le groupe défini par By.
Les données doivent être déjà triées selon la variable By.

INDEX()
Renvoie l'index de la ligne en démarrant à 0.

LAG(Variable, N, By)
Renvoie la N-ième valeur précédente de variable, dans le groupe défini par By.
Les données doivent être déjà triées selon la variable By.

MAX([Nombre 1, Nombre 2, ...])
Renvoie la valeur maximale de la liste d'arguments.
MAX(11, 7, 9) = 11

MEDIAN([Nombre 1, Nombre 2, ...])
Renvoie la médiane des arguments de la liste.
MEDIAN(1,2,3,4,5,6) = 3.5

MIN([Nombre 1, Nombre 2, ...])
Renvoie la plus petite valeur de la liste d'arguments.
MIN(11, 7, 9) = 7

MISSING(Valeur)
Renvoie 1 si Valeur est manquante, 0 sinon.

NMISSING([Valeur1, Valeur2, ...])
Renvoie le nombre de données manquantes dans la liste de valeurs.
_NMISSING([10, , 23, 2, , 3]) = 2

_N_([Valeur 1, Valeur 2, ...])
Renvoie le nombre de valeurs non manquantes dans la liste.

PERCENTILE([Nombre 1, Nombre 2, ...], k)
Renvoie le k-ième centile d'une liste de valeurs.
k doit être compris entre 0 et 1.

RANDBETWEEN(Borne Inf, Borne Sup)
Renvoie un nombre uniformément aléatoire sur l'intervalle [Borne Inf, Borne Sup].

RANDOM()
Renvoie un nombre uniformément aléatoire sur l'intervalle [0, 1].

RANGE([Nombre 1, Nombre 2, ...])
Renvoie l'étendue (Max - Min) des arguments de la liste.
RANGE(15, 7, 9) = 8

STANDARDIZE(Valeur, Moyenne, Ecart-type)
Renvoie la Valeur centrée réduite selon les arguments Moyenne et Ecart-type spécifiés.
STANDARDIZE(42, 40, 1.5) = 1.333

STDEV([Nombre 1, Nombre 2, ...])
Renvoie l'écart-type des arguments de la liste, basé sur la population entière.
STDEV(11, 7, 9) = 1.633

VAR([Nombre 1, Nombre 2, ...])
Renvoie la variance des arguments de la liste, basée sur la population entière.
VAR(11, 7, 9) = 2.667

Texte – Chaines de caractères

CHAR(N)
Renvoie le caractère représenté par le code ASCII N.
CHAR(75) = K

CODE(Texte)
Renvoie le code ASCII du premier caractère de la chaîne Texte.

CONCATENATE([Texte 1, Texte2, ...], Séparateur)
Renvoie la concaténation de plusieurs chaines de caractères, éventuellement séparées par un séparateur.
CONCATENATE([17, 02, 2007], /) = 17/02/2007

COUNT(Texte, Texte recherché)
Renvoie le nombre de fois que Texte recherché apparaît dans Texte. _COUNT_ respecte la casse.
COUNT(This is a thistle? Yes, this is a thistle, this) = 3

EXACT(Texte1, Texte2)
Compare deux textes et renvoie la valeur 1 si ils sont identiques, 0 sinon. _EXACT_ respecte la casse.
EXACT(Oui, Non) = 0

FIND(Texte, Texte cherché, Start Nb)
Recherche Texte cherché au sein de Texte et renvoie l'index de départ de Texte cherché, en comptant à partir du premier caractère de Texte.
Start Nb indique la position à partir de laquelle doit commencer la recherche. _FIND_ respecte la casse.
FIND(Britanny, n) = 5

HEXTONUM(Nb hexadecimal)
Renvoie la valeur numérique correspondant au nombre hexadécimal Nombre hexadécimal.
HEXTONUM(4B) = 75

LEFT(Texte à modifier, nb_car)
Renvoie le(s) premier(s) caractère(s) de Texte à modifier en fonction du nombre de caractères spécifié.
LEFT(Prix de vente, 4) = Prix

LENGTH(Texte)
Renvoie le nombre de caractères contenus dans Texte. Les espaces sont comptés comme étant des caractères.
LENGTH(Pays) = 4

LOWERNOACCENT(Texte à modifier)
Convertit Texte à modifier en lettres minuscules. Les accents ne sont pas conservés.
LOWERNOACCENT(MÉDITERRANÉE) = mediterranee

LOWER(Texte à modifier)
Convertit Texte à modifier en lettres minuscules. Les accents sont conservés.
LOWER(Ile de France) = ile de France

MID(Texte, Start pos, Nb car)
Renvoie la chaine de longueur Nb car extraite de Texte à partir de la position Start pos, en comptant à partir de 1.
MID(17FEV2007, 3, 3) = FEV

NUMTOHEX(Nombre)
Renvoie la représentation hexadécimale de la partie entière de Nombre sous la forme d'une chaîne de caractères (String).
NUMTOHEX(1000) = 3e8

PROPER(Texte à modifier)
Renvoie Texte à modifier avec la première lettre de chaque mot mise en majuscule.
Toutes les autres lettres sont converties en lettres minuscules.

REGEX(RegEx, Chaine, Format)
Format = 0 : Applique l'expression régulière RegEx au début de la Chaine et renvoie 1 si le résultat de l'expression régulière matche exactement avec Chaine, 0 sinon. Si Format = 1, renvoie l'objet matché au lieu de 1 et une valeur manquante sinon.

REPEAT(Texte, Nb)
Répète Texte le nombre de fois spécifié Nb.

REPLACEREGEX(Texte à modifier, Expression régulière, Nouveau texte)
Remplace dans la chaine de caractères Texte à modifier le résultat de l'Expression régulière par Nouveau texte.

REPLACE(Ancien texte, Start nb, Nb car, Nouveau texte)
Remplace dans la chaine de caractères Ancien texte les Nb_car caractères à partir de la position Start_nb.
REPLACE(2009, 3, 2, 10) = 2010

RIGHT(Texte à modifier, nb_car)
Renvoie le(s) dernier(s) caractère(s) de Texte à modifier, en fonction du nombre de caractères spécifié.
RIGHT(17FEV2007, 4) = 2007

SLICING(Texte, Index_début, Index_Fin)
Renvoie la sous-chaine de caractères comprise entre Index_début et Index_Fin dans Texte. Des index négatifs sont considérés comme indiquant des positions relatives à la fin de la chaine.
SLICING('Bonjour M. Durand,', 8, -1] = 'M. Durand'

STRLENGTH(Texte 1, Texte 2, ...)
Renvoie la longueur totale de tous les textes.

SUBSTITUTE(Texte à modifier, Ancien texte, Nouveau texte)
Remplace Ancien texte par Nouveau texte dans Texte à modifier.
SUBSTITUTE(Région Ile de France, Ile de France, IDF) = Région IDF

TRIM(Texte)
Supprime tous les espaces de début et de fin qui apparaissent dans Texte.

UPPERNOACCENT(Texte à modifier)
Convertit Texte à modifier en majuscule. Les accents ne sont pas conservés.
UPPERNOACCENT(élève) = ELEVE

UPPER(Texte à modifier)
Convertit Texte à modifier en majuscule. Les accents sont conservés.
UPPER(Méditerranée) = MÉDITERRANÉE

Agrégation – Générateur de variables : Calculatrice

La calculatrice est disponible dans la rubrique « Outils » de l'interface des méthodes « Agrégation » et « Générateur de nouvelles variables ».

La fenêtre « Calculatrice » vous permet de définir une formule de calcul pour générer une variable calculée à partir des variables numériques disponibles.

Pour définir la formule, insérez successivement des variables d'origine, des opérateurs arithmétiques et éventuellement des parenthèses et des fonctions mathématiques.

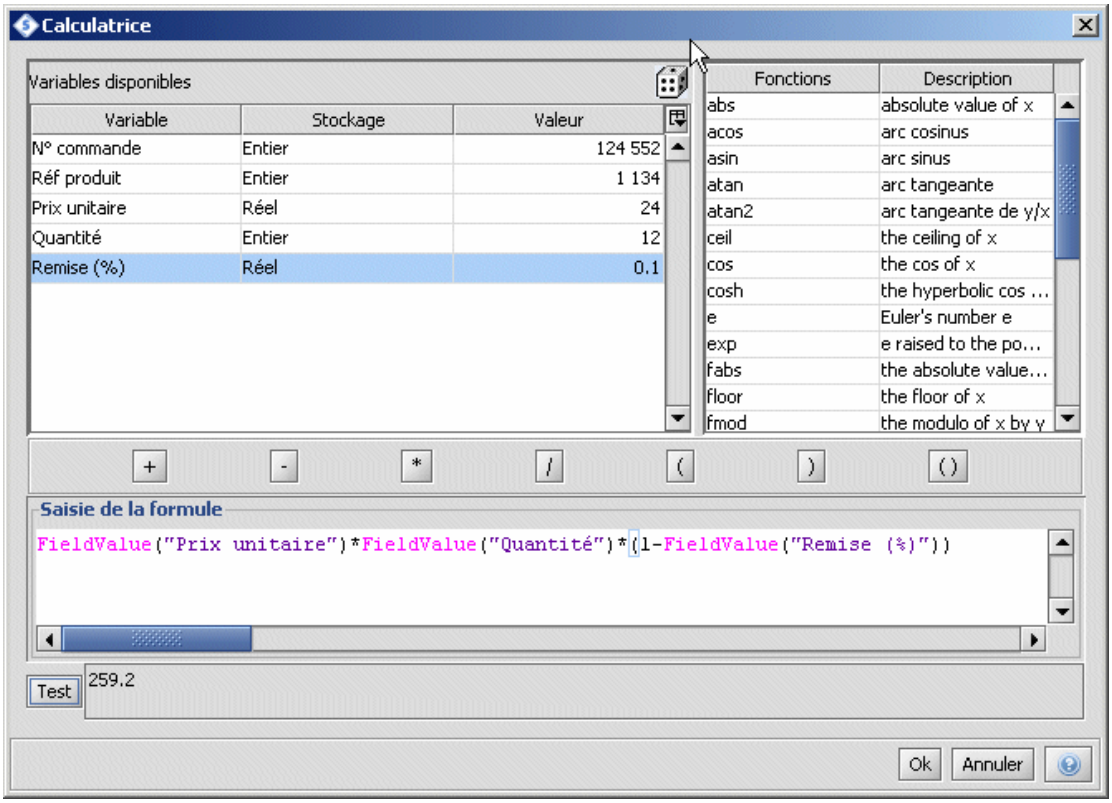


Figure 6-42 : Calculatrice

La formule de calcul doit être saisie dans la zone « Saisie de la formule ».

Introduire une variable numérique dans la formule. Double-cliquez sur la variable dans liste des variables disponibles. La variable apparaît instantanément dans la zone sous la forme suivante : FieldValue(« Variable »). La case de la colonne « Utilisée » dans la liste des variables disponibles est automatiquement cochée.
Ne pas décocher cette case sauf si vous supprimez définitivement la variable de la formule.

Introduire un opérateur, des parenthèses dans la formule. Utilisez les boutons ou le clavier pour introduire les opérateurs +, - , *, / et des parenthèses.

Introduire une fonction dans la formule. Dans la zone de saisie de la formule, positionnez le curseur de la souris à l'endroit souhaité pour introduire la fonction choisie (double-clic). La fonction apparaît instantanément dans la formule.

Supprimer un élément de la formule. Sélectionnez l'élément à supprimer et tapez sur la touche Suppr. Si vous supprimez définitivement une variable de la formule, vous pouvez décocher la case de la colonne « Utilisée » dans la liste des variables disponibles.

Test. Pour tester la formule, cliquez sur le bouton Test. En cas d'erreur, un message apparaît. Sinon, le test est réalisé à partir des valeurs définies arbitrairement par SPAD dans la colonne « Valeur » de la liste des variables disponibles.

Modifier les valeurs pour le test. Pour modifier les valeurs arbitraires proposées par SPAD pour le test, double-cliquez dans les cases de la colonne « Valeur » pour entrer les nouvelles valeurs à tester.

Valider la formule. Cliquez sur le bouton OK.

Agrégation – Générateur de variables : Assistant logique

L'assistant logique est disponible dans la rubrique « Outils » de l'interface des méthodes « Agrégation » et « Générateur de nouvelles variables ».

Cet assistant est l'**équivalent de la fonction _IF_(Test logique, Valeur Si Vrai, Valeur Si Faux)**. Il permet de spécifier un ensemble de conditions à tester renvoyant comme résultat soit Vrai soit Faux. Selon le résultat du test logique, l'utilisateur spécifie la valeur à renvoyer (Valeur définie, Valeur d'une variable d'origine, donnée manquante).

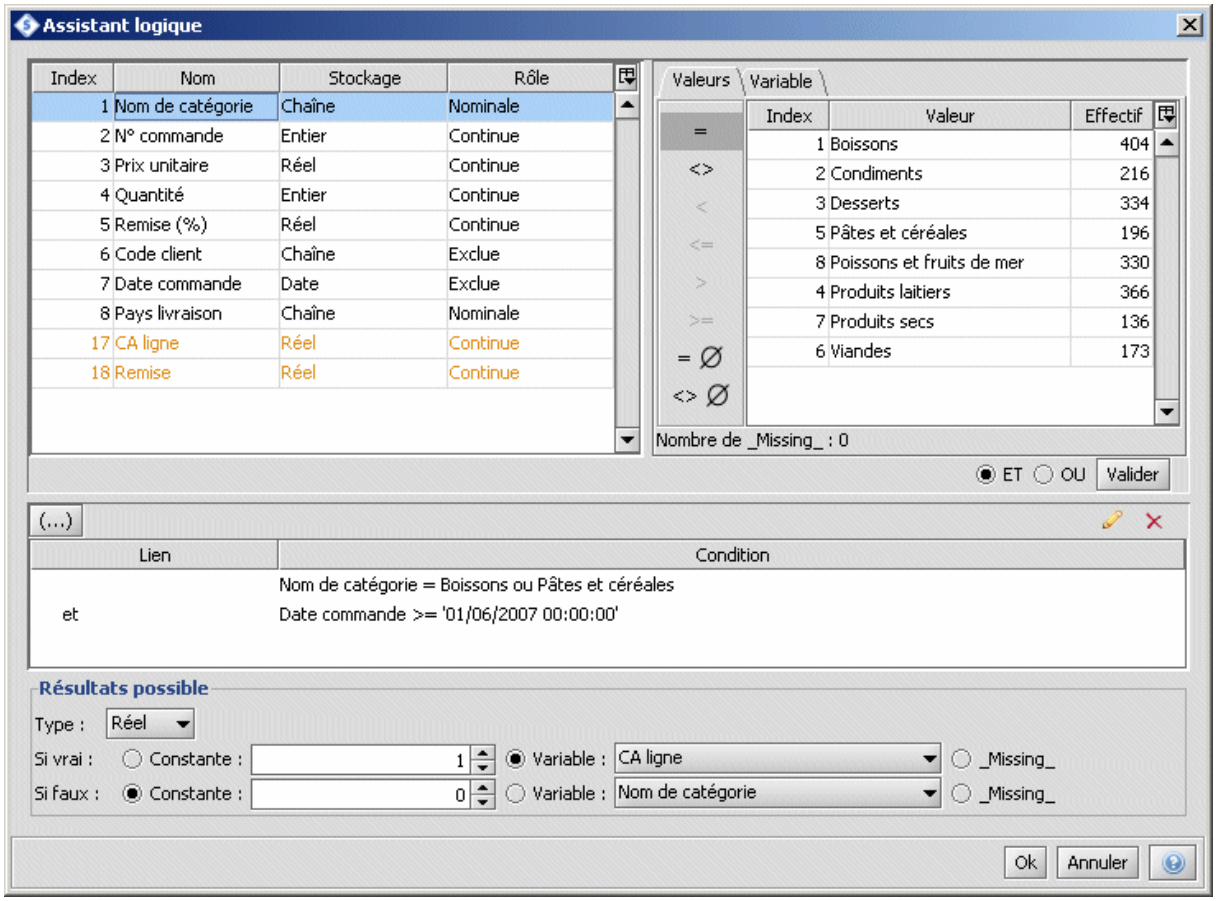


Figure 6-43 : Assistant logique

Conditions – Test logique.
La définition des conditions et des liens (et / ou) les unissant suit exactement la même logique que celle de la méthode « Filtre logique » présentée page 70. Référez-vous à cette méthode pour la définition des conditions.

Résultats.
Spécifiez le type de résultats à retourner (Chaine, Entier, Réel ou Date) selon que test logique renvoie la valeur Vrai ou Faux.
Les résultats à renvoyer peuvent être :

- une constante à définir en entrant la valeur dans la case,
- la valeur d'une variable disponible,
- une valeur manquante (_Missing_).

Valider. Cliquez sur le bouton OK.

Chapitre

7

Méthodes statistiques – Data Mining

Introduction

Les méthodes statistiques figurent dans les rubriques suivantes de la vue « Méthodes » :

- Statistiques descriptives
- Analyses factorielles
- Classifications – Typologies
- Scoring et Modélisation
- Arbres de décision – Segmentations
- Règles d'association
- Tableaux Multiples
- Text Mining

La plupart des méthodes statistiques proposent deux onglets « Individus » et « Pondération » en dehors de leurs onglets de paramétrage spécifiques.
L'onglet « Individus » permet de définir le rôle des individus (Actif/Illustratifs/Exclus) par sélection dans une liste, par filtre logique ou encore par tirage aléatoire.
L'onglet « Pondération » permet de redresser les calculs en spécifiant une variable de pondération ou en définissant cette variable à la volée.

Les principales méthodes statistiques sont présentées dans le manuel « Guide du Data Miner », accessible à partir du menu « ? » de SPAD.

Nous listons ci-après l'ensemble des méthodes avec leur description.

Statistiques descriptives

Icône	Description
	**Tris à plats - Histogrammes - Discrétisation** Cette méthode fournit une **description rapide et automatique des variables nominales et continues** : tris-à-plat, statistiques usuelles (moyenne, écart-type, etc.), histogrammes et discrétisation, croisements de variables deux à deux, boites de dispersion, graphes 3D..
	**Tableaux croisés** Cette procédure permet d'obtenir en une seule fois un nombre illimité de tableaux d'effectifs, de moyennes ou de fréquences. En croisant deux variables nominales, on définit un tableau dont chaque case peut contenir : • Pour les **tableaux d'effectifs** : effectif, % ligne, % colonne. • Pour les **tableaux de moyennes** : l'effectif, la moyenne, l'écart-type. • Pour les **tableaux de fréquences** : effectif, somme des fréquences, % ligne, % colonne.
	**Caractérisation d'une variable nominale** Cette méthode est IDEALE pour caractériser une variable nominale**, caractériser et comparer les groupes définis par ses modalités**, en les **croisant avec le reste de l'information**. Les éléments caractéristiques sont filtrés et classés par ordre d'importance.
	**Caractérisation d'une variable continue** Cette méthode est IDEALE pour caractériser une variable continue **en explorant l'ensemble des liaisons qu'elle entretient avec toutes les autres variables du fichier** (**corrélations** pour les variables continues, **comparaisons de moyennes** pour les variables nominales). Les éléments caractéristiques sont filtrés et classés par ordre d'importance.
	**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 d'un fichier. De façon schématique, les deux variables continues seront les axes x et y du graphique, se coupant au point moyen des variables.
	**Marquage sémantique des modalités** Cette procédure permet de caractériser **les modalités d'une variable nominale** par des marquages sémantiques. C'est une généralisation de la procédure DEMOD existante. Un marquage sémantique est une conjonction logique des modalités des variables caractérisantes. Exemple : les hommes de moins de 25 ans habitant en région parisienne.
	**XLSTAT** Cette procédure permet de sélectionner tout ou partie des données pour effectuer ensuite des **statistiques avec l'application XLSTAT**. Les données sélectionnées sont automatiquement transférées dans **Excel** et un menu XLSTAT vient s'ajouter à la barre de menu classique d'Excel. XLSTAT propose plusieurs méthodes statistiques dont : • les tests d'ajustements, • les tests paramétriques et non paramétriques, • le multi-dimensional scaling...

Analyses factorielles

Icône	Description	Procédure
	**ACP : Analyse en Composantes Principales** Cette procédure effectue l'analyse en composantes principales d'un ensemble d'individus caractérisés par des **variables continues** (actives). L'ACP est une méthode particulièrement puissante pour **résumer et visualiser l'ensemble des liaisons linéaires** entre variables continues actives. L'analyse peut être **normée** (conseillé) ou non normée lorsque les variables actives de l'analyse sont basées sur une échelle comparable. L'utilisateur peut explorer interactivement l'ensemble des résultats à travers l'**éditeur graphique**.	**COPRI**
	**AFC : Analyse Factorielle des Correspondances** Cette procédure effectue l'analyse factorielle des correspondances d'un **tableau de contingence** ou, de façon plus générale, de tout tableau de nombres non négatifs. L'AFC est une méthode particulièrement puissante pour **résumer et visualiser les liaisons entre les profils-lignes et les profils-colonnes** d'un tableau. L'utilisateur peut explorer interactivement l'ensemble des résultats à travers l'**éditeur graphique**.	**CORBI**
	**ACM : Analyse factorielle des Correspondances Multiples** Cette procédure effectue l'analyse des correspondances multiples d'un ensemble d'individus caractérisés par des **variables nominales** (actives). L'ACM est particulièrement adaptée lorsque l'on souhaite **mettre en évidence l'ensemble des liaisons non-linéaires** qui caractérisent un jeu de données. La méthode assure la robustesse des résultats de l'analyse vis-à-vis des modalités de poids faible. L'utilisateur peut explorer interactivement l'ensemble des résultats à travers l'**éditeur graphique**.	**CORMU**
	**Description des facteurs de l'analyse** Cette procédure constitue une **aide à l'interprétation des facteurs** issus d'une analyse factorielle (ACP, ACM, AFC...). Un facteur peut ainsi être rapidement et clairement décrit par les éléments les plus significatifs. Ces éléments pourront être au choix les individus, les modalités, les variables continues, les fréquences, utilisés en tant qu'éléments actifs ou illustratifs dans l'analyse préalable.	**DEFAC**
	**ACM avec choix des modalités actives** Cette procédure effectue l'**analyse des correspondances multiples** d'un ensemble d'individus caractérisés par des variables nominales (actives). Du point de vue de cette procédure plus précisément, les individus sont caractérisés par un **sous-ensemble de modalités** des variables nominales. La procédure permet un **choix contrôlé des modalités actives**. L'utilisateur peut explorer interactivement l'ensemble des résultats à travers l'**éditeur graphique**.	**COREM**
	**Analyse factorielle des correspondances multiples conditionnelles** Cette procédure effectue l'**analyse des correspondances multiples** d'un ensemble d'individus caractérisés par des variables nominales (actives). On s'intéresse ici au cas où les individus sont caractérisés par un sous-ensemble de modalités des variables nominales et **où une variable nominale particulière constitue une contrainte de conditionnement** des résultats. La procédure permet un choix contrôlé des modalités actives. L'utilisateur peut explorer interactivement l'ensemble des résultats à travers l'éditeur graphique.	**CORCO**

Croisement de variables et ACP **TABLE + COPRI**

A partir d'une table de données (N lignes, P colonnes), cette méthode calcule un **tableau de moyennes** et effectue une **analyse en composantes principales de ce tableau**.

Le tableau de moyennes produit devient la **nouvelle table de données source** pour toute nouvelle méthode qui lui serait connectée. Il est visualisable dans la vue « Données ».

Exemple :
Vous pouvez construire un tableau croisant la variable REGION (22 modalités) en ligne et la variable CATEGORIE SOCIO-PROFESSIONNELLE (8 modalités) en colonne en calculant les moyennes de la variable REVENU pour chaque case, puis effectuer l'analyse en composantes principales de ce tableau de moyennes.

Croisement de variables et AFC **TABLE + CORBI**

A partir d'une table de données (N lignes, P colonnes), cette méthode calcule un **tableau de contingence** et effectue une **analyse factorielle des correspondances simples** de ce tableau.

Le tableau de fréquences produit devient la **nouvelle table de données source** pour toute nouvelle méthode qui lui serait connectée. Il est visualisable dans la vue « Données ».

Exemple :
Vous pouvez construire un tableau croisant la variable REGION (en 22 modalités) en ligne et la variable CATEGORIE SOCIO-PROFESSIONNELLE (8 modalités) en colonne, puis effectuer l'analyse des correspondances simples de ce tableau de contingence.
La statistique du KHI-DEUX vous permet de tester l'indépendance de la variable en ligne et de la variable en colonne.

Visualisation de trajectoires de modalités **TYTRA1**

On appelle "trajectoire" d'une modalité la projection sur un espace factoriel des croisements de cette modalité avec une variable privilégiée, qui sera en général le temps.

TYTRA est une procédure permettant d'observer des trajectoires de modalités (TYTRA1), et de préparer une typologie de ces mêmes trajectoires (TYTRA2).

On considère des variables nominales observées par exemple à différentes dates. Leurs modalités (par exemple une catégorie "A" aux dates 1, 2, 3, 4) dessinent avec le temps les trajectoires que l'on retrouve dans les plans d'une analyse factorielle (ACM).

La procédure TYTRA1 permet d'observer les formes et les positions de ces trajectoires dans les plans d'une analyse des correspondances multiples.
Le problème est souvent de repérer les catégories ayant des trajectoires analogues : c'est l'objet de la procédure TYTRA2.

Classifications - Typologies

Icône	Description
	**CAH - Mixte : Classification** **RECIP-SEMIS** Cette procédure effectue la classification des individus au regard de leurs premières coordonnées factorielles issues de l'analyse factorielle préalable (obligatoire). Deux méthodes sont proposées : ▪ La **Classification Ascendante Hiérarchique** (CAH) dont le critère d'agrégation est le critère de **Ward**. Cette méthode est conseillée pour de petits échantillons (**moins de 10 000 individus**). ▪ La **Classification Mixte** (Semis), particulièrement **adaptée aux fichiers volumineux**. Une première classification est obtenue par le croisement de plusieurs partitions de base construites autour de centres mobiles, puis les classes stables ainsi formées sont agrégées par une CAH. Pour chaque méthode, SPAD produit l'arbre d'agrégation, fournit l'ensemble des résultats statistiques ainsi qu'un éditeur graphique de hiérarchies et d'indices de niveaux pour le choix du nombre de classes. **Précondition** Cette procédure ne peut être connectée qu'à une analyse factorielle ou à une description des axes factoriels. **Postcondition** Pour construire une ou plusieurs partitions (ie. typologies) à partir de l'arbre d'agrégation fourni par cette procédure, il suffit de connecter la méthode "Coupure de l'arbre et caractérisation des classes".
	**Coupure de l'arbre - Description des classes** **PARTI-DECLA** Cette procédure **construit des partitions par coupures successives de l'arbre** d'agrégation issu de la méthode de classification "CAH - Mixte". L'utilisateur peut **spécifier le nombre de classes** ou laisser la procédure **rechercher automatiquement les meilleures partitions**, en les améliorant éventuellement par des itérations à centres mobiles (consolidation). On peut **caractériser automatiquement** soit **chaque classe** d'une partition, soit globalement **la partition** elle-même. L'éditeur graphique de plans factoriels permet de représenter les classes dans les différents plans et de colorer les individus selon leur classe d'appartenance. **Précondition** Cette méthode ne peut être connectée qu'à la méthode "CAH - Mixte - Classification Ascendante Hiérarchique, Classification mixte". **Postcondition** Les méthodes "Caractérisation des classes..." et "Marquage sémantique des classes" permettent de compléter la description des classes obtenues. Les partitions peuvent être ajoutées comme nouvelles variables à l'aide de la méthode "Archivage d'axes factoriels et typologies" de la rubrique "Industrialisation - Archivage".
	**Caractérisation des classes** **DECLA** Cette procédure permet de **décrire les partitions créées** par la procédure précédente avec des variables n'ayant pas participé à l'analyse. On peut ainsi sélectionner des variables par thèmes et évaluer leur pouvoir caractérisant des partitions (typologies) construites. Le paramétrage et les éditions sont identiques à celles de la procédure DECLA de l'étape précédente. Les éléments caractéristiques sont classés par ordre d'importance à l'aide d'un critère statistique ("valeur-test") auquel est associé une probabilité : plus la valeur-test est grande, plus la probabilité est faible, plus l'élément est caractéristique. Dans le cas de la description des classes par les modalités des variables nominales, une option permet de classer les modalités caractéristiques soit suivant les valeurs-tests, soit suivant les pourcentages.

	Précondition Cette méthode ne peut être connectée qu'à la méthode "Coupure de l'arbre – Descriptions des classes".
	**Marquage sémantique des classes de la typologie** **MSCLA** Cette procédure permet de caractériser les classes d'une partition par des marquages sémantiques. **C'est une généralisation de la procédure DECLA** existante. La partition est créée par la méthode "PARTI-DECLA", la procédure MSCLA doit être insérée dans la filière après cette méthode. Un marquage sémantique est une conjonction logique des modalités des variables caractérisantes (Exemple: les hommes de moins de 25 ans habitant en région parisienne). Si vous utilisez des variables continues comme variables caractérisantes, elles seront découpées à chaud en N classes d'effectifs égaux et considérées comme des variables nominales. Exemple de la qualité d'un marquage sémantique On caractérise la classe 1 que vous avez qualifié au regard des résultats de l'analyse "Les acheteurs". Si le marquage sémantique "Hommes, moins de 25 ans, habitant en région parisienne" est tel que les individus ainsi définis recouvre 20% des acheteurs et qu'ils soient tous des acheteurs (pas de débordement), c'est un marquage sémantique de qualité. **Précondition** Cette méthode ne peut être connectée qu'à la méthode "Coupure de l'arbre – Descriptions des classes".
	**Typologie de trajectoires de modalités** **TYTRA2** La procédure TYTRA2 doit être exécutée après la procédure TYTRA1. La procédure TYTRA2 récupère l'ensemble des coordonnées factorielles des modalités croisées avec la variable privilégiée représentant le temps. Ces coordonnées factorielles sont issues de l'analyse des correspondances multiples réalisée dans la procédure TYTRA1. Pour réaliser une typologie des modalités, on effectue tout d'abord une analyse en composantes principales sur les coordonnées factorielles définissant les trajectoires des modalités. Deux modalités sont proches lorsque leurs trajectoires ont la même forme, indépendamment de leur position dans l'espace. Les procédures de classification (RECIP) et de coupure de l'arbre (PARTI) permettent ensuite de construire une typologie des trajectoires. Deux modalités sont dans une même classe quand elles ont des trajectoires semblables, quelles que soient leurs positions dans l'espace.

Amado – Graphiques de Bertin

Icône	Description
	**Construction d'une matrice de données et analyse graphique** A partir d'une table de type Individus x Variables, cette méthode vous permet de construire une matrice de données de type tableau de contingence, de moyennes, de cumuls. Après construction (Exécution de la filière contenant la méthode), cette matrice de données sera analysée par la technique dite "Graphique de BERTIN". La matrice de données est définie par le croisement de deux variables nominales. Dans chaque case de la matrice on aura soit les effectifs (tableau de contingence), soit les moyennes ou les cumuls d'une variable continue. Si la matrice construite est un tableau de contingence, la technique "Graphique de BERTIN" est une alternative à la technique des correspondances simples.
	**Sélection des éléments d'une matrice de données pour AMADO** A partir d'une table de contingence, table de moyennes ou de cumuls, cette méthode vous permet de sélectionner les éléments (lignes et colonnes) de la table initiale pour construire une nouvelle matrice de données. Après exécution, cette matrice de données sera analysée par la technique dite "Graphique de BERTIN". Si la matrice construite est un tableau de contingence, la technique "Graphique de BERTIN" est une alternative à la technique des correspondances simples.

Scoring et Modélisation

Icône	Description
	**Discriminantes optimales – Furnival et Wilson** **FUWILD** Cette procédure réalise la **sélection des N "meilleurs" ajustements pour une analyse linéaire discriminante à deux groupes**. Le critère de sélection peut être le "**R2**", le "**R2 ajusté**" ou le "**Cp**" **de Mallows**. Soit N le nombre de meilleurs ajustements demandés et P le nombre des variables "explicatives" du modèle. La procédure édite les N meilleurs ajustements, pour toutes les tailles de modèles, de 1 à P - 1 variables (l'ajustement avec les P variables est unique). La procédure fournit la valeur du critère (le R2 , le R2 ajusté ou le Cp), le F de Fisher associé au R2, la probabilité critique associée à ce F, et la valeur-test correspondante. La liste des variables du modèle est ensuite éditée avec les coefficients estimés, les tests de nullité, la probabilité critique et la valeur-test associée. Enfin un graphique représentant l'évolution du critère en fonction du nombre de variables dans les modèles fournit une synthèse rapide des sélections. Dans le cas d'une analyse discriminante, on édite les résultats de la régression "équivalente" à l'analyse discriminante à 2 groupes. On utilisera cependant cette présentation à titre indicatif seulement, puisque les hypothèses classiques de normalité ne sont pas admissibles dans ce cas. Pour le critère du R2, toutes les sélections imprimées sont optimales. Pour les deux autres critères, les sélections ne sont pas toujours optimales (le R2 ajusté et le Cp de Mallows varient de façon non monotone en fonction du nombre de variables). On repère qu'une sélection n'est pas optimale si la procédure n'édite pas les coefficients des variables (seuls les noms des variables et la valeur du critère sont imprimés). Dans ce cas l'ajustement sélectionné, s'il n'est pas optimal pour le critère, est cependant meilleur que les ajustements qui n'ont pas été calculés.

Discriminante directe sur variables continues **DIS2GD**

Cette procédure effectue une **analyse linéaire discriminante à deux groupes sur variables continues** par la méthode classique de Fisher.

Elle permet de réaliser des "estimations bootstrap" du biais et de la précision des principaux résultats de la discrimination : coefficients, probabilités individuelles de classement, pourcentages globaux de classement. Elle permet de modifier les coûts et les probabilités a priori de classement dans les groupes. Elle assure la gestion d'Individus de base, d'individus-tests et d'individus anonymes.

La procédure permet d'imprimer au préalable les statistiques descriptives sur les variables du modèle dans chacun des 2 groupes. On présente ensuite les résultats de l'analyse discriminante: tableaux de classement, fonction discriminante, résultats de la "régression équivalente", édition des affectations des individus.

Si une validation "bootstrap" est demandée, les résultats de la discrimination sont réédités avec les estimations "bootstrap". En particulier, le biais et la précision des classements globaux sont édités en vis-à-vis des classements directs. Pour les individus-anonymes, la procédure calcule leur probabilité bootstrap d'affectation.

Si une évaluation sur des individus tests est demandée, la procédure réédite les résultats de la discrimination relatifs à ces individus. Si l'affectation d'individus anonymes est demandée, seule l'édition des affectations est fournie.

Discriminante sur facteurs principaux **DIS2GFP**

Cette procédure effectue une **analyse linéaire discriminante à deux groupes sur les coordonnées factorielles issues d'une analyse en composantes principales** NON NORMEE par la méthode classique de Fisher.

Elle permet de réaliser des "estimations bootstrap" du biais et de la précision des principaux résultats de la discrimination : coefficients, probabilités individuelles de classement, pourcentages globaux de classement. Elle permet de modifier les coûts et les probabilités a priori de classement dans les groupes. Elle assure la gestion d'Individus de base, d'individus-tests et d'individus anonymes.

La procédure permet d'imprimer au préalable les statistiques descriptives sur les variables du modèle dans chacun des 2 groupes. On présente ensuite les résultats de l'analyse discriminante: tableaux de classement, fonction discriminante, édition des affectations des individus.

La règle de décision est finalement exprimée en fonction des variables d'origine. Les résultats de la "régression équivalente" sont donnés à titre indicatif ; en effet les hypothèses classiques de "normalité" n'ont pas de sens ici.

Si une validation "bootstrap" est demandée, les résultats de la discrimination sont réédités avec les estimations "bootstrap". En particulier, le biais et la précision des classements globaux sont édités en vis-à-vis des classements directs. Pour les individus-anonymes, la procédure calcule leur probabilité bootstrap d'affectation.

Si une évaluation sur des individus tests est demandée, la procédure réédite les résultats de la discrimination relatifs à ces individus. Si l'affectation d'individus anonymes est demandée, seule l'édition des affectations est fournie.

Discriminante sur variables qualitatives **DIS2GFM**

Cette procédure effectue une **analyse linéaire discriminante à deux groupes sur des coordonnées factorielles** par la méthode classique de Fisher.
On peut ainsi effectuer une discrimination sur des données nominales ou qualitatives, après une analyse des correspondances multiples.

Elle permet de réaliser des "estimations bootstrap" du biais et de la précision des principaux résultats de la discrimination : coefficients, probabilités individuelles de classement, pourcentages globaux de classement. Elle permet de modifier les coûts et les probabilités a priori de classement dans les groupes.

Elle assure la gestion des individus d'apprentissage et des individus-tests.

La procédure permet d'imprimer au préalable les statistiques descriptives sur les variables du modèle dans chacun des 2 groupes. On présente ensuite les résultats de l'analyse discriminante: tableaux de classement, fonction discriminante, édition des affectations des individus.

La règle de décision est finalement exprimée en fonction des modalités d'origine. Les résultats de la "régression équivalente" sont donnés à titre indicatif ; en effet les hypothèses classiques de "normalité" n'ont pas de sens ici.

Si une validation "bootstrap" est demandée, les résultats de la discrimination sont réédités avec les estimations "bootstrap". En particulier, le biais et la précision des classements globaux sont édités en vis-à-vis des classements directs. Pour les individus-anonymes, la procédure calcule leur probabilité bootstrap d'affectation.

Si une évaluation sur des individus tests est demandée, la procédure réédite les résultats de la discrimination relatifs à ces individus. Si l'affectation d'individus anonymes est demandée, seule l'édition des affectations est fournie.

Régression logistique **LOGIS**

La régression logistique tend à **expliquer la probabilité d'occurrence d'un événement binaire**. Cette probabilité, du fait de sa distribution ne peut être expliquée par un modèle de régression classique utilisant la méthode des moindres carrés. On effectue une transformation qualifiée de **LOGIT** dont le traitement relève du modèle linéaire généralisé et qui met en œuvre une méthode basée sur la recherche du maximum de vraisemblance.

Si P est la probabilité que l'on cherche à expliquer, on définit sous le terme ODDS le rapport P/(1-P) et la grandeur qui est en définitive expliquée est le logarithme de ce ODDS.

La régression logistique travaille :

- Soit sur une variable nominale à deux modalités dont on cherche à estimer la probabilité d'occurrence de l'une de ses modalités.
- Soit sur un couple de variables quantitatives, dont l'une indique la fréquence d'un événement qui s'est réalisé (variable nombre de réussites) et l'autre la fréquence ou taille de la population (variable nombre d'essais) dans la laquelle cette fréquence s'est réalisée.

La procédure édite au préalable les statistiques descriptives sur les variables du modèle dans chacun des 2 groupes. On présente ensuite les résultats de la régression logistique : coefficients du modèle, Odds Ratio, tableaux de classement...

Discriminante sur variables qualitatives et fonction de Score **DISCO + SCORE**

L'interface de la méthode DISCO est différente de la méthode DIS2GFM pour assurer une compatibilité complète avec le module graphique de la méthode SCORE.

La procédure SCORE calcule une fonction de score qui est une modification de la fonction discriminante, destinée à faciliter son interprétation et son utilisation. On introduit en particulier des "zones" de décision (rouge, verte, centrale) à partir d'une tolérance d'erreur de classement, et on édite divers résultats permettant d'apprécier les performances de la discrimination.

Dans la suite, les deux groupes à discriminer sont appelés:

- groupe des scores "forts" ou groupe "1"
- groupe des scores "faibles" ou groupe "2".

C'est l'utilisateur qui définit en entrée la correspondance entre les modalités 1 et 2 de la variable de groupe, et les attributs "fort" et "faible" des scores.

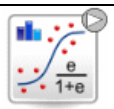
La procédure SCORE transforme les coefficients en utilisant les deux règles suivantes:

Coefficient minimum de chaque variable: pour chaque variable nominale, le plus petit coefficient est mis à la valeur 0. Ainsi le score minimum possible d'un individu est 0. Il est obtenu pour un individu qui, pour chaque variable, possèderait la modalité affectée du coefficient transformé 0.

Maximum possible de la fonction score: la valeur du score maximum possible est un choix de l'utilisateur (par exemple 1000). Ce maximum correspond à la somme des plus grands coefficients transformés de chaque variable.

Le score attribué à un individu s'obtient en additionnant les coefficients transformés associés aux modalités de l'individu. La fonction de score transformée classe les individus de la même façon que la fonction discriminante initiale.

La procédure SCORE fournit un éditeur graphique qui présentent les courbes de densité et de répartition des fonctions de scores pour chaque groupe. Les courbes ROC et LIFT sont également disponibles.




Régression logistique pour fonction de Score **LOGISQUAL + SCORE**

L'interface de la méthode LOGISQUAL est différente de la méthode LOGIS pour assurer une compatibilité complète avec le module graphique de la méthode SCORE. Les variables explicatives sont obligatoirement des variables nominales et les interactions ne sont pas permises.

La procédure SCORE calcule une fonction de score qui est une modification du modèle de régression logistique, destinée à faciliter son interprétation et son utilisation. On introduit en particulier des "zones" de décision (rouge, verte, centrale) à partir d'une tolérance d'erreur de classement, et on édite divers résultats permettant d'apprécier les performances de la discrimination.

Voir la description précédente.

Régressions multiples optimales – Furnival et Wilson **FUWILR**

Cette procédure réalise **la sélection des N "meilleurs" ajustements pour une régression**. Le critère de sélection peut être le "**R2**", le "**R2 ajusté**" ou le **"Cp" de Mallows**.

Soit N le nombre de meilleurs ajustements demandés et P le nombre des variables "explicatives" du modèle. La procédure édite les N meilleurs ajustements, pour toutes les tailles de modèles, de 1 à P - 1 variables (l'ajustement avec les P variables est unique).

La procédure fournit la valeur du critère (le R2 , le R2 ajusté ou le Cp), le F de Fisher associé au R2, la probabilité critique associée à ce F, et la valeur-test correspondante.

La liste des variables du modèle est ensuite éditée avec les coefficients estimés, les tests de nullité, la probabilité critique et la valeur-test associée. Enfin un graphique représentant l'évolution du critère en fonction du nombre de variables dans les modèles fournit une synthèse rapide des sélections.

Pour le critère du R2, toutes les sélections imprimées sont optimales. Pour les deux autres critères, les sélections ne sont pas toujours optimales (le R2 ajusté et le Cp de Mallows varient de façon non monotone en fonction du nombre de variables). On repère qu'une sélection n'est pas optimale si la procédure n'édite pas les coefficients des variables (seuls les noms des variables et la valeur du critère sont imprimés). Dans ce cas l'ajustement sélectionné, s'il n'est pas optimal pour le critère, est cependant meilleur que les ajustements qui n'ont pas été calculés.

Référence: L'algorithme de sélection est une transcription de l'algorithme "leaps and bounds" de Furnival et Wilson. (Technometrics, 174, Vol.16, pp.499-511).

Régression simple et multiple – Anova, Ancova **VAREG**

Cette méthode 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.

Le traitement des données manquantes est assuré par le paramétrage.

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: coefficient 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.

La procédure 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 à expliquer "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.

Régression PLS **REPLS**

La régression PLS contient les régressions communément appelées PLS1 (**Une seule variable à expliquer**) et PLS2 (**Plusieurs variables à expliquer**).

La régression PLS n'utilise pas directement les variables explicatives, mais calcule d'abord les **composantes PLS** sur lesquels est ensuite réalisée la régression.

Les composantes, orthogonales entre elles, sont des combinaisons linéaires des variables explicatives initiales calculées de manière à décrire les plus grandes variations des données et à maximiser la covariance entre les variables explicatives et les variables à expliquer.

La **régression PLS** est une **alternative à la régression linéaire multiple** et **présente un grand intérêt si** :

- **Le nombre d'observations (lignes) est inférieur au nombre de variables** explicatives du modèle.
- Le tableau des variables explicatives présentes de **nombreuses données manquantes**.
- Les variables explicatives présentent de **fortes colinéarités**.

Modèle Linéaire Général **MLGEN**

Cette procédure assure 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 facteurs hiérarchisés** (ou emboîtés), et interactions d'ordre un ou deux.

En option, il est possible d'obtenir le calcul et l'impression des coefficients de régression avec pour chacun, son écart-type et le test de sa nullité, valable dans le contexte où le terme aléatoire est supposé engendré par une loi de Laplace-Gauss.

La procédure réalise une analyse de la variance pour tester l'existence de l'effet de chacun des facteurs du modèle. Pour ces tests, le carré moyen de référence est, par défaut, le carré moyen résiduel. Trois types de sommes des carrés des écarts sont possibles qui considèrent plusieurs conditions d'ajustement des facteurs. Le programme peut aussi tester un facteur par rapport à un autre facteur dont le carré moyen est pris comme référence en lieu et place du carré moyen résiduel.

En option, la procédure permet le calcul d'estimations ou de contrastes, ainsi que celui des moyennes ajustées des niveaux de facteurs du modèle. Ces moyennes ajustées peuvent faire l'objet de comparaisons deux à deux.

Les éditions sont nombreuses et pour la majorité d'entre elles, optionnelles.

Discriminante neuronale sur variables continues **NEURO**

Cette procédure permet de **discriminer une variable nominale à N classes par une méthode neuronale**. Cette méthode est basée sur l'utilisation d'un réseau de neurones multicouche. Elle est particulièrement adaptée à la prise en compte de relations non linéaires.

Un réseau comprend plusieurs couches, chaque couche étant constituée de neurones. Dans la couche d'entrée, il y a autant de neurones que de variables explicatives; dans la couche de sortie, il y a autant de neurones que de classes à discriminer. Les couches intermédiaires appelées couches cachées, sont des paramètres du réseau.

Le nombre de couches cachées et le nombre de neurones des couches cachées sont choisis par l'utilisateur. On prendra rarement plus de deux couches cachées. Un choix judicieux du nombre de neurones par couche cachée évitera la création d'une couche supplémentaire.

Les neurones d'une couche inférieure seront reliés aux neurones d'une couche supérieure par des connexions appelées POIDS et chaque neurone de chaque couche est affectée d'un BIAIS. C'est le calcul de ces poids et biais qui permet l'affectation des individus dans les classes. Les poids et les biais sont pris au hasard lors de la première présentation des individus.

Les poids et les biais seront ajustés alors par un apprentissage utilisant la méthode de rétropropagation du gradient. C'est un algorithme qui se divise en deux parties :

La propagation ou relaxation
On calcule les valeurs de chaque neurone avec le réseau de connexions calculé à l'itération t (on dit aussi au temps t). Les neurones de sortie seront combinaisons linéaires des neurones d'entrée. Poids et biais sont les coefficients de ces combinaisons linéaires.

La rétropropagation
Elle consiste à modifier les poids et les biais en fonction de l'erreur observée afin que les données d'entrée fournissent de meilleurs résultats.

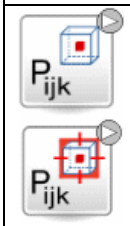
L'estimation des poids est un processus itératif avec correction en fonction de l'erreur rencontrée. L'erreur est appelée coût. Elle est calculée après un certain nombre de présentations de l'échantillon d'apprentissage. Le nombre de présentations du fichier d'apprentissage est fixé par paramétrage. Un paramètre permet également d'arrêter la procédure quand la décroissance du coût devient très faible. L'apprentissage nécessite le choix du nombre de couches du réseau et du nombre de neurones dans ces couches. On peut également ajuster les paramètres epsilon et êta dont dépendent la correction apportée aux poids et biais à chaque itération. Il est possible de permuter les individus de l'échantillon.

Echantillon d'apprentissage, échantillon-test et anonymes

Il est possible de tester la qualité de la fonction discriminante sur des individus-tests n'ayant pas participé au calcul. On compare alors la classe de sortie obtenue avec la classe observée et on édite le pourcentage de bon classement.

Discriminante neuronale sur facteurs principaux **NEUROFP**

Cette procédure permet de **discriminer une variable nominale à N classes par une méthode neuronale** avec comme **variables explicatives** des variables **nominales**. Une analyse des correspondances multiples (**ACM**) préalable est réalisée sur les variables nominales explicatives, ce qui permet ensuite de discriminer la variable à expliquer à partir des coordonnées factorielles issues de l'ACM.
Le paramétrage est ensuite équivalent à celui de la procédure NEURO.



Modèle log-linéaire sur tableau Individus*Variables **LOGLI1**
Modèle log-linéaire sur tableau de contingence **LOGLI2**

L'écriture de la procédure LOGLI est une adaptation du programme d'analyse de modèles log-linéaires réalisé par Noboru Ohsumi, de l'Institut de Mathématique Statistique de Tokyo.

L'analyse log-linéaire est une méthode de modélisation des tableaux de contingence à plusieurs entrées. A partir d'un tableau, la procédure permet d'ajuster plusieurs modèles log-linéaires hiérarchiques (une interaction n'est spécifiée qu'en présence des termes principaux).
Deux types de données sont acceptés en entrée : des données "individus x variables" d'une part et des tables de contingence d'autre part. Dans le cas de données individuelles, on peut utiliser une variable poids (définie dans l'onglet pondération) pour calculer les fréquences.

Plusieurs possibilités sont offertes pour sélectionner les modèles à ajuster :

- Modèle(s) défini(s) par l'utilisateur (le nombre de variables dans chaque modèle est limité à 7).
- Tous les modèles possibles (dans ce cas, le nombre de variables est limité à 4).
- Sélection d'un modèle par une méthode combinatoire fondée sur la minimisation du critère de l'information d'AKAIKE (AIC) ainsi que sur la décomposition de la statistique du CHI2 du rapport de vraisemblance (le nombre de variables est limité à 4).
- Sélection d'un modèle par une méthode "pas à pas" ascendante fondée sur la minimisation du critère de l'information d'AKAIKE (le nombre de variables est limité à 4).

L'édition commence par l'écriture du modèle à estimer, suivie éventuellement par les tris à plat des variables nominales du modèle. Apparaissent ensuite, à la demande de l'utilisateur, le tableau de contingence avec les effectifs estimés par le modèle, les informations sur la convergence de l'algorithme et le nombre d'itérations de l'ajustement. Une synthèse des statistiques d'ajustement est imprimée (estimation du CHI-2 de Pearson, du CHI2 du rapport de vraisemblance ainsi que leurs degrés de liberté et les probabilités associées). Puis la procédure fournit l'identification des coefficients des modalités des facteurs et des interactions éventuelles. Il est également possible d'obtenir la matrice de structure du modèle. Dans le cas d'une sélection automatique d'un modèle, le tableau récapitulatif de tous les modèles traités est imprimé.

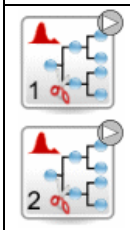
Arbres de décision - Segmentation

Icône	Description
 	**Arbre de décision interactif** **IDT1 – IDT2** Cette méthode permet d'**expliquer et de prédire une variable nominale** par l'ensemble des variables disponibles en formalisant le modèle sous la forme d'un arbre de décision. 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. La procédure IDT2 construit l'arbre de segmentation initial en fonction de la méthode choisie (CHAID, C&RT, C4.5) et des paramètres associés. Après exécution de la procédure, vous disposez dans le menu contextuel de la méthode de l'**outil graphique « Interactive Decision Tree » pour manipuler l'arbre initial**. **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.-à-d. 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. **C&RT** C&RT 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. C&RT possède deux particularités : ▪ La détermination de la bonne taille de l'arbre s'effectue par post-élagage, c.-à-d. 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, c.-à-d. chaque sommet segmenté ne possède que deux enfants. C&RT construit généralement des arbres très « compacts », ayant de bonnes capacités de prédiction. Le calcul est couteux en temps 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. **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.-à-d. 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.



Arbre de régression interactif **IRT1 – IRT2**

Cette méthode permet d'**expliquer et de prédire une variable continue** par l'ensemble des variables disponibles en formalisant le modèle sous la forme d'un arbre de décision.

La procédure IRT1 prépare les données pour la construction de l'arbre (Procédure IRT2), 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.

La procédure IRT2 construit l'arbre de segmentation initial en fonction de la méthode choisie (AID, C&RT) et des paramètres associés.

Après exécution de la procédure, vous disposez dans le menu contextuel de la méthode de l'**outil graphique « Interactive Regression Tree » pour manipuler l'arbre initial**.

AID
AID est une méthode d'induction d'arbre de régression reposant sur un critère de discrimination statistique, le F de Fisher issu de la décomposition de la variance : le rapport variance expliquée et la variance résiduelle. Il s'agit vraisemblablement d'une des méthodes les plus anciennes, elle prend ses racines dans des travaux effectués au milieu des années 60 par Morgan et Sonquist (1963).

AID possède deux particularités par rapport aux autres méthodes d'induction d'arbres de régression :

la détermination de la bonne taille de l'arbre s'effectue par pré-élagage, c.-à-d. lors de l'expansion de l'arbre. La décision de segmenter un sommet dépend d'un test d'indépendance de Fisher effectué à partir de la décomposition de la variance calculée sur les feuilles proposées 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, si celle-ci est catégorielle. 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 AID 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.

Référence : J.N. Morgan and J.A. Sonquist, "Problems in the analysis of survey data and a proposal", JASA, 58, pp. 415-434, 1963.

C&RT
C&RT 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 régression, le critère utilisé repose sur la notion de « pureté » calculée à partir de la variance des sous-groupes produits, elle s'appuie donc également sur une décomposition de la variance pour déterminer la segmentation la plus informative.

Les particularités de C&RT sont décrites dans la méthode précédente IDT1-IDT2.

Typologie par arbre de décision **ICT**

Les typologies dans SPAD

Les typologies ou « classifications non supervisées » segmentent une population en groupes homogènes d'individus (i.e. segments) du point de vue d'un ensemble de variables de telle sorte que :

- Deux individus d'un même segment se ressemblent le plus possible,
- Deux individus de segments différents diffèrent le plus possible.

Dans SPAD, les typologies sont construites sur la base d'analyses factorielles préalables (ACP, ACM, AFC) ou les variables actives constituent le « Thème » pour lequel on cherche à segmenter la population.

Le fait d'utiliser une analyse préalable dans les techniques de classification, permet d'obtenir des typologies plus robustes en n'utilisant que l'information stable et pertinente représentée par les X premiers axes factoriels de l'analyse.

ICT – Interactive Classification Tree

ICT s'inscrit parfaitement dans ce cadre en effectuant une analyse factorielle préalable à la construction de la typologie, mais présente des particularités nouvelles et remarquables :

La segmentation finale (typologie) obtenue est directement formalisée à l'aide d'un arbre de décision interactif.

Dans cet arbre, la création des segments se fait par divisions successives, un segment étant toujours divisé en 2 sous-segments.

L'appartenance à un segment est basée sur des règles logiques de type « SI Condition Alors appartenance au Segment K »

Les typologies obtenues sont donc très facilement interprétables et produisent des règles simples à appliquer sur d'autres populations.

L'interactivité permet d'intervenir dans la construction de la typologie en choisissant les variables de coupures les plus judicieuses, en modifiant les seuils de partitionnement ou le nombre de segments terminaux.

La segmentation finale peut être construite sur la base des variables actives de l'analyse (cas classique) ou sur une sélection des variables (actives et/ou illustratives) de cette analyse

On peut donc construire un arbre en utilisant d'autres variables que les variables actives de l'analyse préalable.

Exemple : On peut construire un arbre de décision avec des variables socio- économiques (Age, sexe, revenu,..) dans un espace déterminé par des variables d'opinions. On obtient donc une typologie où les segments sont définis par des règles sur les variables socio-économiques, mais où le regroupement des individus est fait en fonction de leurs ressemblances sur les opinions (variables actives de l'analyse préalable).

Mise en œuvre – résultats disponibles

Vous devez définir les variables actives et/ou illustratives de l'analyse préalable, précisez si besoin les variables utilisées dans la construction de l'arbre. Vous pouvez choisir de construire un arbre automatique ou un arbre totalement interactif.

Après exécution de la procédure, vous disposez dans le menu contextuel de la méthode (clic droit sur la méthode) de l'outil graphique « Interactive Clustering Tree » pour manipuler l'arbre initial. Vous disposez également l'éditeur de « Plans factoriels » pour de l'analyse préalable.

Tableaux multiples

Analyse Factorielle Multiple **AFM**

Cette procédure effectue l'**analyse factorielle d'un fichier dans lequel un ensemble d'individus est caractérisé par plusieurs groupes de variables**. Les variables peuvent être nominales ou continues.

Le premier intérêt de la méthode est de réaliser une analyse factorielle dans laquelle l'influence des différents groupes de variables est équilibrée a priori. Cet équilibre est nécessaire car les groupes de variables diffèrent toujours selon le nombre de variables et/ou selon la structure de ces variables (i.e. leurs interrelations). En particulier, ceci permet de faire intervenir en tant qu'éléments actifs, simultanément des groupes de variables nominales et des groupes de variables continues.

La procédure fournit, dans un premier temps, le tri-à-plat des variables nominales et les statistiques sommaires des variables continues. Elle édite, pour chaque groupe de variables, l'histogramme des valeurs propres issues de l'analyse factorielle effectuée sur ce groupe de variables. Elle édite aussi, pour chaque groupe de variables continues, la matrice des corrélations (ou des covariances). On édite ensuite la matrice des corrélations entre facteurs partiels (facteurs des analyses séparées).

On trouve ensuite les résultats de l'AFM proprement dite, analyse factorielle globale de l'ensemble des variables : on obtient, l'histogramme des valeurs propres, les coefficients Lg et RV de liaison entre groupes, les corrélations entre les variables canoniques et les facteurs de l'analyse globale, les rapports (inertie inter/inertie totale), les contributions et cosinus carrés de tous les groupes (actifs et illustratifs). Si l'édition en est demandée, le listage comporte les coordonnées, contributions et cosinus carrés des individus et individus partiels associés (actifs et éventuellement illustratifs).

Pour les individus et modalités, on trouve les tableaux triés des éléments ayant les plus fortes contributions, les éléments ayant les plus fortes et les plus faibles inerties intra groupes, les éléments partiels ayant les plus fortes inerties intra groupes. De plus, on édite les coordonnées et valeurs-test de toutes les modalités et de toutes les modalités partielles.

Le listage suivant fournit les coordonnées, les contributions, les cosinus carrés des variables continues actives et illustratives, les corrélations entre les axes et les variables non réduites (actives et illustratives). On obtient enfin les coordonnées, les contributions, les cosinus carrés des axes partiels des analyses (actifs et illustratifs).

Analyse conjointe de tableaux multiples **STTIS**

La procédure STATIS permet d'extraire l'information de données multidimensionnelles recueillies dans des situations différentes.

Prenons l'exemple d'un essai thérapeutique : les patients sont suivis pendant plusieurs mois et lors de chaque visite, le praticien décrit l'état pathologique d'un patient en situant ses performances sur un certain nombre d'échelles. On obtient des données à trois indices : un indice "patient", un indice "performance" et un indice "visite".

L'exploitation de telles données, à trois "voies", est un axe de recherche qui est loin d'être épuisé. En effet, les méthodes proposées à l'heure actuelle (y compris les méthodes exposées ici ...) n'abordent pas le problème dans sa complexité. En particulier, l'ordre dans lequel sont effectuées les visites, est rarement pris en compte. La méthode STATIS, tout comme l'analyse factorielle multiple (procédure AFMUL), privilégie une des "voies" car elle découpe les données en tranches. Ces tranches sont ensuite comparées entre elles à l'aide des outils de géométrie euclidienne habituels en analyse de données.

La procédure édite la situation des tableaux (ou des nuages suivant l'option) les uns par rapport aux autres ; l'interprétation des distances entre tableaux (ou nuages) en termes de contributions des lignes et (ou) des colonnes ; l'étude de la configuration de l'interstructure ; et l'étude du nuage compromis lorsqu'il existe.

Text Mining

Construction du vocabulaire **MOTS**

La procédure **MOTS est obligatoire avant toute analyse textuelle**. Elle **crée le « vocabulaire initial des mots »** sur la question (variable) textuelle choisie. Tous les mots sont pris en compte sans aucun seuil de sélection. Si on sélectionne plusieurs questions, elles sont fusionnées et considérées comme une seule.
La procédure permet de sélectionner tout échantillon par filtre logique, sélection par liste et tirage aléatoire.
Les paramètres permettent d'homogénéiser la casse du texte et de définir les « caractères séparateurs » des mots et des phrases.

La procédure fournit trois tableaux principaux :

- Le tableau des effectifs des mots selon le nombre de lettres
- La liste des mots par ordre alphabétique
- La liste des mots par ordre de fréquence.

On dispose également du nombre total de mots du corpus, du nombre et du pourcentage de mots distincts, ce qui donne une première idée de la richesse du vocabulaire.

Les tableaux de résultats sont immédiatement disponibles et mis en forme dans Excel.

Après exécution de la procédure MOTS, on peut utiliser la **méthode de lemmatisation semi automatique qui permet d'affiner le vocabulaire en effectuant des sélections par seuils, des corrections, des équivalences de mots (lemmatisation), etc.**

Construction des segments répétés **SEGME**

La procédure SEGME recherche les « Segments répétés » dans le corpus. **Un segment répété est une suite de mots (Exemple : la vie est belle)**. Les segments répétés sont construits à partir du « Vocabulaire des mots en cours », c'est à dire le vocabulaire initial ou le vocabulaire modifié s'il existe.

Les paramètres de la procédure permettent une sélection des segments répétés par seuil de fréquence ou de longueur en nombre de mots.

La procédure fournit deux tableaux principaux :

- La liste des segments répétés par ordre alphabétique.
- La liste des segments répétés par ordre de fréquence.

Les tableaux de résultats sont immédiatement disponibles et mis en forme dans Excel.

Après exécution de la procédure SEGME, **le vocabulaire des mots (vocabulaire en cours) et le vocabulaire des segments répétés sont fusionnés** pour former un seul et même vocabulaire, **modifiable par la méthode de lemmatisation** semi automatique qui permet d'affiner le vocabulaire en effectuant des sélections par seuils, des corrections, des équivalences de mots (lemmatisation), etc.

Edition des contextes des mots **CORDA**

La procédure CORDA recherche et édite les contextes des mots. Cette procédure est toujours effectuée à partir du « Vocabulaire initial des mots » créé par la méthode MOTS.

Les phrases (contextes) éditées sont les vraies phrases du corpus. Pour chaque mot sélectionné, la procédure édite tous les contextes.

Les tableaux de résultats sont immédiatement disponibles et mis en forme dans Excel.

Lemmatisation semi-automatique **CORTEX**

La méthode interactive de lemmatisation semi-automatique du vocabulaire est exécutable après paramétrage et exécution des méthodes MOTS et/ou SEGME.

Elle permet d'effectuer de nombreuses modifications sur le vocabulaire en cours :

Vous pouvez faire les opérations suivantes sur les mots et segments :

- **Corriger des mots** et des segments
- **Supprimer des mots** et des segments,
- **Définir des équivalences entre plusieurs mots** et/ou plusieurs segments.

Les modifications sur un vocabulaire en cours sont enregistrées par le Menu – Fichier – Enregistrer. Les méthodes statistiques en aval utiliseront ce nouveau vocabulaire.

Important
Les **modifications effectuées sur un vocabulaire en cours peuvent également être enregistrées sous forme de règles**.

Ces règles peuvent être réutilisées de façon automatique, soit dans un autre diagramme, ou sur un autre vocabulaire.

Vocabulaire spécifique de groupes d'individus **SEGME**

La procédure VOSPEC recherche et **édite les mots et segments répétés caractéristiques de groupes d'individus en fonction de la fréquence**.
On peut par exemple connaître immédiatement le vocabulaire spécifique des femmes et des hommes.

La procédure travaille sur le vocabulaire en cours, qu'il contienne des mots seuls ou des mots et des segments répétés, qu'il s'agisse d'un vocabulaire initial ou d'un vocabulaire modifié.

Les mots et segments répétés sont édités par ordre de « valeurs-test » décroissantes. En tête figurent les mots et les segments répétés plus fréquemment rencontrés dans le groupe d'individus qu'en moyenne (spécificités positives). En fin de liste figurent les mots ou segments répétés moins souvent cités (spécificités négatives).

La procédure édite également les phrases caractéristiques des groupes selon deux critères.

Le premier critère « Moyenne des valeurs-tests » des mots de la phrase a tendance à favoriser les réponses courtes.
Le second critère classe les phrases par ordre de distance croissante à la réponse moyenne du groupe d'individus (au sens de la distance du Khi-2). Ce critère à tendance à favoriser les réponses longues.

Les paramètres permettent de choisir le nombre de mots et segments caractéristiques ainsi que le nombre de phrases courtes (premier critère) ou longues (second critère) à éditer.

Les tableaux de résultats sont immédiatement disponibles et mise en forme dans Excel.

Construction d'un tableau lexical de contingence **TALEX**

La construction et l'analyse d'un tableau lexical de contingence est une des approches très originales du traitement des données textuelles dans SPAD

La procédure TALEX construit une table de contingence avec :

- En ligne les mots et segments répétés du vocabulaire en cours.
- En colonne les groupements définis par les modalités des variables nominales choisies.

Par exemple, on construit **un tableau avec en ligne les mots et segments répétés du vocabulaire en cours et en colonne les classes d'âges des individus**. A l'intersection d'une ligne (les mots et segments) et d'une colonne (groupe d'individus), le tableau contient la fréquence d'emploi du mot ou du segment répété par le groupe.

La procédure permet de choisir plusieurs variables et de créer ainsi un tableau juxtaposé.

Le tableau ainsi construit est édité à la demande et immédiatement disponible et mis en forme dans Excel.
Ce tableau de contingence peut ensuite être soumis à l'analyse des correspondances.

AFC sur Tableau lexical de contingence **CORBIT**

Cette procédure effectue **l'analyse factorielle des correspondances d'un tableau lexical de contingence**.
Les éléments en colonnes seront toujours appelées "fréquences" (les modalités des variables nominales) et les éléments en lignes des "individus" ou des "lignes" (les mots).

La procédure calcule les coordonnées des lignes (mots) des fréquences et des modalités des variables nominales illustratives (centres de gravité des individus).

On dispose de deux modes de calcul des axes factoriels: la méthode classique de diagonalisation en mémoire centrale et une méthode en "lecture directe" réservée aux grands tableaux.

On édite l'histogramme des valeurs-propres de l'analyse puis les coordonnées, contributions à l'inertie et cosinus carrés des fréquences et, à la demande, celles des individus (actifs et éventuellement illustratifs).

On trouve enfin les coordonnées et valeurs-test pour tous les centres de gravité (modalités des variables nominales).

L'éditeur graphique de plans factoriels permet de superposer sur un même plan l'ensemble des mots des fréquences.

Les procédures de classification (RECIP-SEMIS) et de coupure de l'arbre (PARTI-DECLA) permettent ensuite de construire une typologie des mots.

Création d'une table Mots / Segments / Variables **TEXNU**

TEXNU est une procédure de gestion qui permet de fusionner les informations disponibles sur les mots et segments répétés avec les variables nominales et continues sous forme d'une table unique pour disposer de toute la richesse des méthodes du logiciel.

La table construite aura la forme suivante :
- En ligne : les individus
- En colonne : les mots, segments répétés et les variables sélectionnés.

Chaque mot ou segment répété sélectionné est considéré comme une variable continue. A l'intersection d'une ligne (individu) et d'une colonne (mots ou segments) on dispose de la fréquence de citation. Les variables numériques sélectionnées conservent leur rôle (nominale ou continue) comme dans un tableau classique de données numériques.

De façon automatique cette table est immédiatement utilisable par toute autre méthode statistique, de Data Management ou d'exportation des données.

Modèles structurels

Approche PLS – Modélisation d'équations structurelles **PLSX**

PLS Path Modeling ou Approche PLS à été développée dans le cadre du projet Européen N° IST-2000-31071

Vous trouverez sur le site **www.esisproject.com** toutes les informations concernant le projet et les différents partenaires du projet.

Si vous avez acheté cette option la documentation est disponible dans le sous-menu ESIS Project Documentation du menu d'Aide de SPAD.

Chapitre

8

Archivage et industrialisation des modèles

Les méthodes d'archivage et d'industrialisation sont directement liées aux méthodes d'analyse de données (analyses factorielles et classification) et aux méthodes prédictives (analyses discriminantes, régressions, arbres de décision…).

Les méthodes d'archivage permettent d'ajouter aux données sources l'ensemble des résultats, prédictions et indicateurs sous forme de nouvelles variables.

La méthode d'archivage « Modèle prédictif – Fichier Règles » permet d'archiver les résultats d'une méthode prédictive sous la forme d'un fichier de règles, applicable à un autre jeu de données.

C'est la méthode d'industrialisation « Application d'un modèle prédictif » qui permet d'appliquer le fichier de règles ainsi archivé à un nouveau de jeu de données.

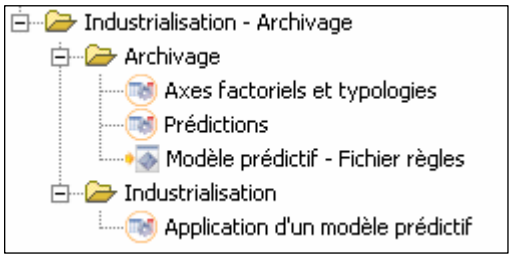


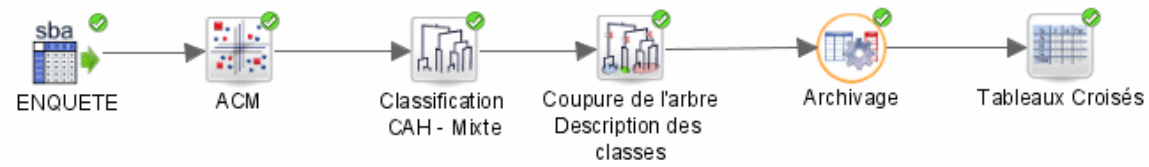
Figure 8-1 : Industrialisation - Archivage

Archivage

Axes factoriels et typologies

Pour archiver des axes factoriels ou des typologies, dans la vue « Méthodes » accédez à la méthode Archivage - Axes factoriels et typologies » et connectez cette méthode à la procédure d'analyse factorielle ou à la procédure de création de typologies (Coupure de l'arbre et description des classes).

Après paramétrage et exécution de la méthode, les données archivées sont automatiquement utilisables par toutes les méthodes de SPAD.



Paramétrage de la méthode.
Double cliquez sur la méthode pour accéder au paramétrage

- **Les axes factoriels à archiver**

1. Dans la liste "Archivage", choisissez Axes factoriels.
   Les axes factoriels disponibles apparaissent dans la liste "Axes factoriels disponibles".
2. Dans la liste "Axes factoriels disponibles", marquez les axes à archiver en cliquant dessus (Usage des touches SHIFT et CTRL standard).
3. Utilisez le bouton pour valider vos choix. Les axes factoriels marqués disparaissent de la liste "Axes factoriels disponibles" et apparaissent dans la liste "Axes factoriels archivés". Utilisez le bouton pour archiver tous les axes factoriels en une seule fois.
4. Modifiez éventuellement les libellés qui vous sont proposés par défaut

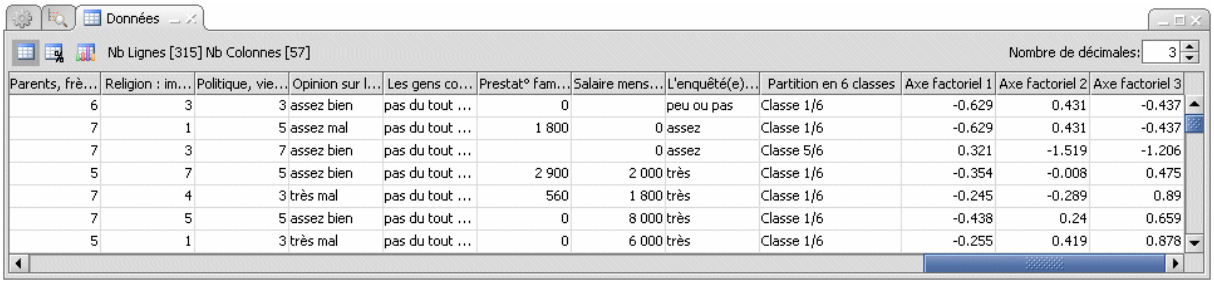
Les boutons permettent de désélectionner des axes factoriels dont on a demandé l'archivage par erreur.

- **Les partitions à archiver**
  1. Dans la liste "Archivage", choisissez Partitions.
     Les partitions disponibles apparaissent dans la liste "Partitions disponibles".
  2. Sélectionnez la partition à archiver, puis cliquez sur le bouton ">" .
     La partition apparaît dans la liste "Partitions archivées" avec des libellés de modalités par défaut.
  3. Modifiez éventuellement les libellés qui vous sont proposés par défaut puis "VALIDER".

  La partition apparaît dans la liste "Liste des partitions archivées".
  Procédez ainsi de suite pour chaque partition que vous souhaitez archiver.

Cliquez sur le bouton Ok pour valider votre paramétrage.

Après exécution de la méthode, dans la vue « Données » de la fenêtre « Exécutions », on peut visualiser les données correspondant à la table construite.


Données

Nb Lignes [315] Nb Colonnes [57] Nombre de décimales: 3

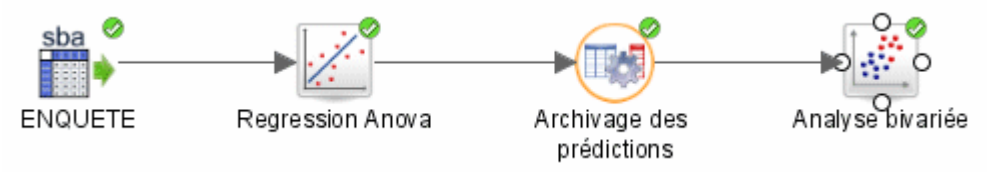
Parents, frè…	Religion : im…	Politique, vie…	Opinion sur l…	Les gens co…	Prestat° fam…	Salaire mens…	L'enquêté(e)…	Partition en 6 classes	Axe factoriel 1	Axe factoriel 2	Axe factoriel 3
6	3	3	assez bien	pas du tout …	0		peu ou pas	Classe 1/6	-0.629	0.431	-0.437
7	1	5	assez mal	pas du tout …	1 800	0	assez	Classe 1/6	-0.629	0.431	-0.437
7	3	7	assez bien	pas du tout …		0	assez	Classe 5/6	0.321	-1.519	-1.206
5	7	5	assez bien	pas du tout …	2 900	2 000	très	Classe 1/6	-0.354	-0.008	0.475
7	4	3	très mal	pas du tout …	560	1 800	très	Classe 1/6	-0.245	-0.289	0.89
7	5	5	assez bien	pas du tout …	0	8 000	très	Classe 1/6	-0.438	0.24	0.659
5	1	3	très mal	pas du tout …	0	6 000	très	Classe 1/6	-0.255	0.419	0.878

Figure 8-2 : Données issues de l'archivage des axes factoriels et des partitions

Prédictions

Les **méthodes prédictives** des rubriques « **Scoring – Modélisation** » et « **Arbres de décision** » produisent des **fichiers règles** permettant de **calculer de nouvelles informations**.

Par exemple la régression linéaire produit les règles permettant de calculer la variable Y estimée et la variable résidu pour chaque individu. Après la méthode de « Régressions – Analyse de variance », on peut connecter la méthode « Archivage des prédictions », qui produit de façon transparente une nouvelle table de données contenant les informations initiales et les colonnes issues de l'application des règles. Les méthodes connectées après la méthode « Archivage des prédictions » pourront utiliser ces nouvelles variables.



L'interface générale de la méthode « Archivage des prédictions » édite les **coefficients** issus du modèle et permet de **choisir les colonnes** que l'on souhaite conserver et de les **renommer**.

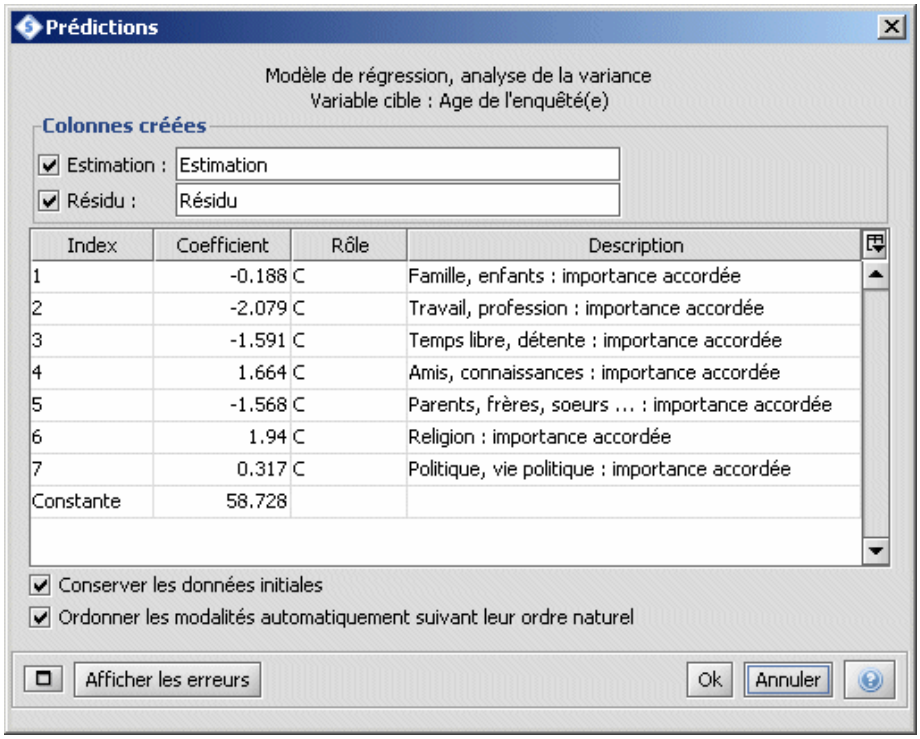


Figure 8-3 : Archivage des prédictions pour la Régression, Analyse de la variance

Vous pouvez **conserver ou non les données initiales** et ordonner ou non les modalités suivant leur ordre naturel lorsque des variables nominales sont générées par l'archivage. Ce qui n'est pas le cas ici.

Dans le tableau des coefficients, la colonne « Rôle » précise si le coefficient est lié à une variable continue (C), nominale (N), voire à une interaction (N*N, C*N, C*C, N*N*N…).

Cliquez sur le bouton Ok pour **valider** votre paramétrage.

Après exécution de la méthode, dans la vue « Données » de la fenêtre « Exécutions », on peut **visualiser les données** correspondant à la table construite.

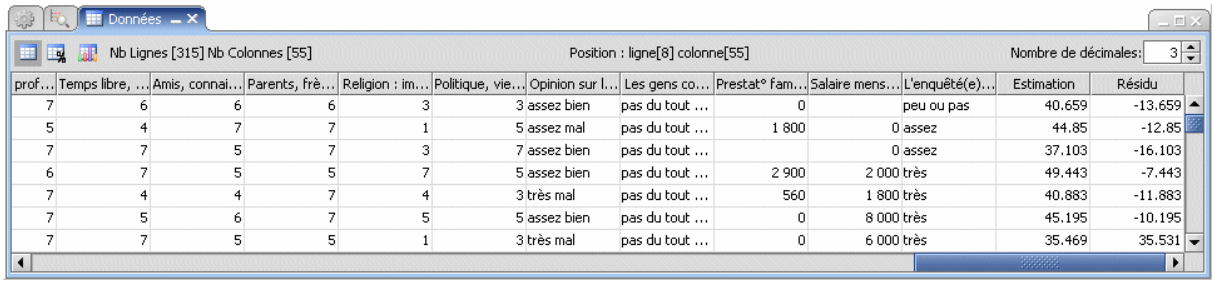


Figure 8-4 : Données issues de l'archivage des prédictions

Liste des prédictions archivables par méthode

- **Régression simple et multiple - Anova, Ancova / Régression PLS / Modèle Linéaire Général**
  - Variable Y estimée
  - Résidu
- **Régression Logistique** :
  - Probabilité d'appartenance à chaque groupe
  - Modalité de classement attribuée
  - Partition en 4 groupes (G1 classé G1, G1 classé G2, G2 classé G1, G2 classé G2)
- **Analyses discriminantes** :
  - Valeur de la fonction discriminante
  - Modalité de classement attribuée
  - Probabilité d'appartenance à chaque groupe
  - Partition en 4 groupes (G1 classé G1, G1 classé G2, G2 classé G1, G2 classé G2)
- **Fonction de Score** :
  - Fonction de Score
- **Arbres de décision (Discriminante - IDT2)** :
  - Numéro du segment d'appartenance
  - Modalité de classement attribuée
  - Probabilité d'appartenance à chaque groupe
- **Arbres de décision (Régression - IRT2)** :
  - Numéro du segment d'appartenance
  - Prédiction de la variable à expliquer
  - Résidu
- **Typologie par arbre de décision** :
  - Numéro du segment d'appartenance
  - Libellé du segment

Modèle prédictif – Fichier règles

Les **méthodes prédictives** des rubriques « **Scoring – Modélisation** » et « **Arbres de décision** » produisent des **modèles prédictifs sous forme de règles** archivables dans un fichier interne au projet en cours ou dans un fichier externe. Après archivage, ces règles pourront ensuite être appliquées à d'autres tables disposant des mêmes variables explicatives du modèle.

Dans l'interface, **cochez l'option d'enregistrement** souhaitée et donnez un nom si vous archivez le modèle dans le projet ou cliquez sur le bouton « … » pour spécifier le répertoire et le nom du fichier règles dans le cas d'une sauvegarde externe.

Après exécution de la méthode, si vous avez enregistré votre modèle dans le projet, le fichier contenant les règles est matérialisé dans le gestionnaire de projet et sa rubrique « Modèles ».

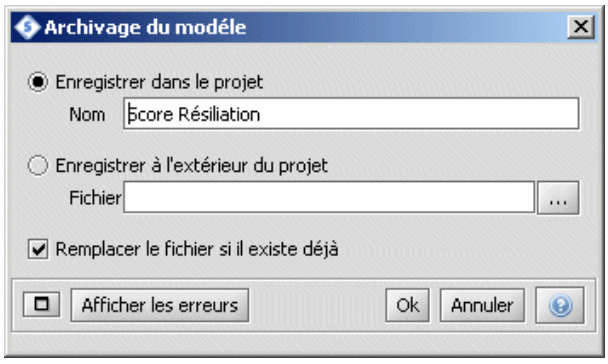


Figure 8-5 : Archivage d'un modèle prédictif

Industrialisation

Application d'un modèle prédictif

Appliquer un modèle préalablement archivé dans le projet en cours.

1. Dans le gestionnaire de projet, accédez au dossier « Modèles », puis glissez-déposez le modèle choisi sur la méthode contenant les données sources.
2. La connexion entre les deux méthodes est insérée automatiquement.
3. Double-cliquez sur la méthode insérée « Application de modèle » pour accéder au paramétrage de la méthode.
4. Cliquez sur le bouton « Charger le modèle » pour faire apparaître la liste de choix des colonnes à créer et les caractéristiques du modèle.
5. Sélectionnez les nouvelles colonnes à créer.
6. Validez avec OK et exécutez.

Après exécution de la méthode, dans la vue « Données » de la fenêtre « Exécutions », on peut **visualiser** les données sources et **les prédictions issues de l'application du modèle**.

Appliquer un modèle préalablement archivé dans un fichier externe au projet en cours.

1. Dans la vue « Méthodes », glissez-déposez la méthode « Industrialisation - Application d'un modèle prédictif » sur la méthode contenant les données sources.
2. La connexion entre les deux méthodes est insérée automatiquement.
3. Double-cliquez sur la méthode insérée « Application de modèle » pour accéder au paramétrage de la méthode.
4. Spécifiez le fichier modèle externe contenant les règles à appliquer aux données sources en cliquant sur le bouton « … ».
5. Cliquez sur le bouton « Charger le modèle » pour faire apparaître la liste de choix des colonnes à créer et les caractéristiques du modèle.
6. Sélectionnez les nouvelles colonnes à créer.
7. Validez avec OK et exécutez.

Après exécution de la méthode, dans la vue « Données » de la fenêtre « Exécutions », on peut **visualiser** les données sources et **les prédictions issues de l'application du modèle**.

Chapitre

9

Annexes

Expressions régulières

Les expressions régulières sont des caractères permettant de coder n'importe quelle entité écrite appartenant à un ensemble (ex: les décimales, les majuscules, les minuscules, ou les mots commençant par k, etc.)

Le principe de ces caractères est de construire un masque codant l'entité cible. Ensuite chaque séquence de caractères est comparée au masque. S'il y a correspondance, la séquence appartient à l'ensemble de l'entité.

Ce mécanisme permet de rechercher et/ou de remplacer des chaines de caractères dans une variable sans savoir où sont placées les chaînes dans les valeurs de la variable.

Les expressions régulières dans SPAD peuvent être utilisées dans le filtre logique (filtre des individus), dans les fonctions _REGEX_ et _REPLACEREGEX_ de la rubrique « Texte » ainsi que dans l’assistant logique des méthodes « Agrégation » et « Générateur de nouvelles variables ».

Format des expressions régulières

Une expression régulière permet de définir un ensemble de chaînes de caractères à partir d'une seule définition. Elles peuvent contenir que des caractères ASCII.

Certains caractères présentent des significations particulières. Voici la liste de ces caractères :

\ | [^ * ? . - %

Si vous voulez inclure un de ces caractères dans la définition d'une expression régulière, vous devez le précéder du caractère \. Lorsque le caractère \ précède certains caractères alphanumériques, ceux-ci présentent une signification différente. Voici la liste des caractères et leurs significations :

Caractère	**Description**
.	tous les caractères ASCII.
+	accepte 1 ou plusieurs fois le caractère précédent. Par exemple : **a+** accepte un ou plusieurs '**a**', c'est-à-dire : • '**a**', '**aa**', '**aaa**' ... sont acceptées,· • '**ab**' n'est pas acceptée,· • ' ' n'est pas acceptée.
?	accepte 0 ou 1 fois le caractère précédent. Par exemple : **a?** accepte 0 ou 1 '**a**', c'est-à-dire : • ' ' est acceptée,· • '**a**' est acceptée,· • '**aa**' n'est pas acceptée.
*****	accepte 0 ou plusieurs fois le caractère précédent. Par exemple : **a*** accepte 0 ou plusieurs '**a**', c'est-à-dire· : • '**a**', '**aa**', '**aaa**' ... sont acceptées· • '**ab**' n'est pas acceptée· • ' ' est acceptée.
[]	définit un ensemble de caractères y compris un intervalle. Un intervalle se définit comme un caractère initial et un caractère final séparés par un tiret. Par exemple **[a-z]**, accepte tous les caractères alpha en minuscule de '**a**' à '**z**'. Si vous voulez inclure dans la liste des caractères le '**-**', vous devez le précéder du \ Par exemple, **[\-ab]** accepte '**-**' , '**a**' et '**b**' **[ac]** accepte le caractère '**a**' ou le caractère '**c**' **[a-c]** accepte l'intervalle des caractères '**a**' jusqu'au caractère '**c**', c'est-à-dire '**a**', '**b**' et '**c**' **[a-cg]** accepte '**a**', '**b**', '**c**' et '**g**' **[a0-4]** accepte '**a**' et '**0**', '**1**', '**2**', '**3**', '**4**'
[^]	le caractère ^ en début de chaîne désigne le complément de l'ensemble des caractères du reste de la chaîne : le caractère désigné est n'importe quel caractère n'appartenant pas à la chaîne. Par exemple, **[^0-3]** accepte **tous** les caractères ASCII **sauf** pour '**0**', '**1**', '**2**', '**3**'.
\a	n'importe quel caractère alpha en incluant des caractères spéciaux comme les accents. • '**a**' ... '**z**' • '**A**' ... '**Z**' • '**_**'· • '**à**' , '**é**', '**è**', '**ç**' ...
\A	tout sauf \a
\d	n'importe quel caractère numérique, ainsi que le point des décimaux • '**0**' ... '**9**'· • '**.**'
\D	tout sauf \d
\e	tous les caractères ASCII.
\E	tout sauf \e
\f	uniquement des caractères en minuscule. • '**a**' ... '**z**'
\F	tout sauf \f
\g	uniquement des caractères en majuscule.'A' ... 'Z'
\G	tout sauf \g

\h	uniquement des caractères alpha en minuscule ou majuscule sans accents.'a' ... 'z' , 'A' ... 'Z'
\H	tout sauf \h
\n	uniquement des caractères numériques'0' ... '9'
\N	tout sauf \n
\s	n'importe quel caractère d'espacement • ' '. • tabulations
\S	tout sauf \s
\w	n'importe quel caractère alphanumérique sauf les caractères d'espacement.
\W	tout sauf \w
\%	cas particuliers : '**d**' pour les dates **(\%d)** et '**n**' pour les formats numériques **(\%n)**. Par exemple : **\%d date** ie:· • **\%dddmm** jour (2) mois (2) • **\%dddmmyy** jour (2) mois (2) année (2)· • **\%dmmyyyy** mois (2) année (4) • **\%n** accepte tous les chiffres 0.00, 85.10
Ponctuation	'**!**' '/' '**:**' '**@**' '**[**' '**`**' '**{**' '**~**' '**-**'
|	Opération "OU" logique entre les expressions régulières. Par exemple : **ab|bc** accepte les chaînes '**ab**' et '**bc**'
	Pour inscrire un point ou d'autres caractères spéciaux, vous devez les précéder d'un anti-slash (barre oblique inverse). Par exemple :· • Pour un point -> **\.**· • Pour un signe d'interrogation -> **\?**· • Pour un astérisque -> *****· • Pour un crochet -> **\[**· • Pour un antislash -> \\
Exemple :	Pour accepter des adresses mail : On accepte (au moins 1) les caractères alpha, les caractères numériques, le '.' (point), le '-' (tiret) et le '_' (trait de soulignement) suivis d'un '@' et puis encore des caractères alpha, numériques, '.', '-' et '_' . A la fin il pourrait y avoir aussi des espaces qu'il faudrait éliminer. Le résultat en expression régulière est : **[\-\h\d_]+@[\-\h\d_]+\s*** Ici, '**d**' correspond aux caractères numériques avec le point des décimaux, '**h**' correspond à tout caractère alpha en minuscule ou majuscule, **\s*** on élimine tous les espaces car **\s** représente tout sauf des espaces.