DataLab 7 ®

DataLab ®

Guide de référence

Septembre 2008

COMPLEX SYSTEMS

COMPLEX SYSTEMS

149 rue Montmartre 75 002 PARIS

Tel : 01 42 21 40 80 – Fax : 01 42 21 40 79

info@complex-systems.fr

www.complex-systems.fr

Sommaire

1. Généralités

1.1 Principe

La phase de recherche de variables discriminantes peut prendre jusqu'à 70% du temps dévolu à la construction d'un modèle.

DataLab a pour objectif d'accélérer, simplifier mais aussi approfondir les analyses statistiques et data mining. Pour ce faire, DataLab privilégie à la fois une automatisation poussée et un contrôle utilisateur de tous les instants pour une appropriation la plus efficace possible.

DataLab est en particulier structuré autour d'un moteur de Data Scanning unique qui automatiquement explore, évalue et valide un grand nombre de variables explicatives d'un comportement ou d'une grandeur. Son principe de fonctionnement repose sur la génération et l'évaluation d'un nombre élevé de transformations et de combinaisons issues des variables disponibles en entrée. Les résultats sont restitués sous la forme la plus lisible qui soit.

DataLab regroupe des fonctionnalités de :

- description : statistiques, profiling, tableaux croisés, segmentation et typologie
- prévision : segment, arbre de décision, modèle de score

Pour la partie prédictive, l'utilisation de DataLab repose sur 4 opération principales :

1. Chargement du fichier contenant la variable à expliquer et les variables potentiellement explicatives
2. Définition éventuelle de l'échantillon sur lequel va être réalisée l'exploration (sélection des enregistrements et des variables, pondération des enregistrements si redressement)
3. Définition de la variable à expliquer
4. Démarrage de l'exploration

Les nouvelles variables évaluées sont issues des transformations et combinaisons détaillées en Table 1. La Figure 1 détaille l'ensemble des traitements réalisés par DataLab.

A partir des variables d'entrée, DataLab crée et évalue un ensemble de variables transformées. Toutes les évaluations se font dans un cadre de régression (linéaire si la cible est continue,

logistique si elle est binaire). Le critère d'évaluation est la part de variance expliquée et la signification statistique est déterminée par un test de variance de Fisher.
DataLab propose deux principaux types de résultats :

- la liste des variables les plus explicatives de la variable à expliquer au niveau individuel, détaillées selon le niveau de complexité (Cf. Figure 1) :
    - variables de base : variables d 'entrée, éventuellement éclatées en variables binaires pour les variables catégoriques
    - variables transformées : variables de base transformées individuellement
    - variables combinées : combinaisons de 2 ou 3 variables (plus pour les variables regroupées en groupe de variables) selon des opérateurs spécifiques.

- la sélection de (c'est à dire, l'ensemble des) variables la plus explicative de la variable à expliquer (modèles)
  La sélection de variables est réalisée de manière séparée en ne tenant compte que des variables d'entrée d'une part (sélection de base), et de l'ensemble des variables d'autre part (sélection DataLab). Ces deux modèles sont directement comparés afin d'évaluer l'apport de DataLab selon plusieurs critères statistiques.
  Les modèles sont réalisés au moyen d'une sélection mixte (avant et arrière alternée).

Table 1. Résumé des transformations mises en œuvre dans DataLab.

Niveau	Variable	Arguments	Transformations	Conditions	Résultat
Intermédiaire	Catégorique	(a)	= valeur manquante	> frequence min	binaire
			= modalité	> frequence min	binaire
			regroupement modalités vs. Cible	> frequence min	binaire
	Numérique	(a)	remplacement valeurs manquantes > moyenne	> frequence min (non manquants)	continue
			= valeur manquante	> frequence min	binaire
			= valeur unique	> frequence min	binaire
			regroupement intervalles vs. Cible	> frequence min	binaire
			Carré(a)	> frequence min (non manquants), ABS(a) < 10 000	continue
			Racine(a)	> frequence min (non manquants), MIN(a) > 0	continue
			Log10(a)	> frequence min (non manquants), MIN(a) > 0	continue
			Inverse(a)	> frequence min (non manquants), MIN(a) > 0	continue
Combiné	Groupes	(a,b,c,d)	Moyenne(a,b,c,d)		continue
			Ecart-type(a,b,c,d)		continue
			Min(a,b,c,d)		continue
			Max(a,b,c,d)		continue
			Nombre(a,b,c,d)		continue
			Pente(a,b,c,d)	Groupe ordonné	continue
			Variation(a,b,c,d)	Groupe ordonné	continue
	Binaire	(a,b)	Et, Et non, Ou, Ou non, Ou exclusif	couples autorisés, > frequence min	binaire
	Numérique	(a,b)	a/b	couples autorisés, MIN(var) > 0	continue
			a+b, a-b, a*c	couples autorisés, DOMAINE(a)<10*DOMAINE(b)	continue
			a<b, a=b, a>b	couples autorisés, MAX(a)<MIN(b) ou MAX(b)<MIN(a)	binaire

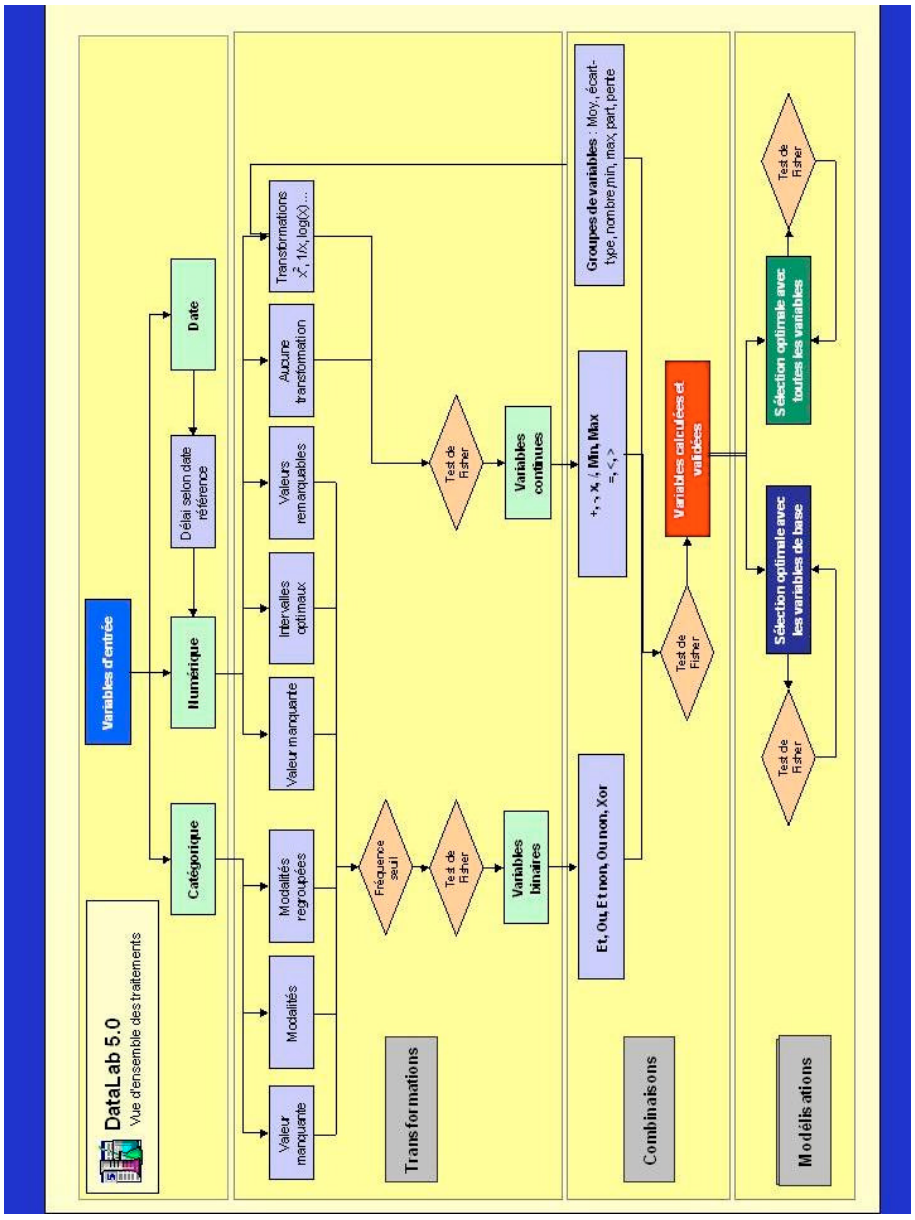


Figure 1. Vue d'ensemble des traitements mis en œuvre dans DataLab

A coté de ces deux résultats phares, DataLab fournit d'autres informations particulièrement utiles :

- la liste exhaustive des valeurs remarquables pour chaque variable, c'est à dire de valeurs supérieures à une fréquence d'occurrence fixée par l'utilisateur et statistiquement significatives (par exemple, un don égal exactement à 20€)
- le découpage optimal de variables numériques en intervalles (ou le regroupement optimal de modalités pour les variables catégoriques) dont le nombre et la nature sont déterminés par un test du Chi2 normé
- l'identification de segments à fort ou faible potentiel, les segments étant l'ensemble des variables binaires. Cette représentation peut aussi être faite sous forme d'arbre de décision.

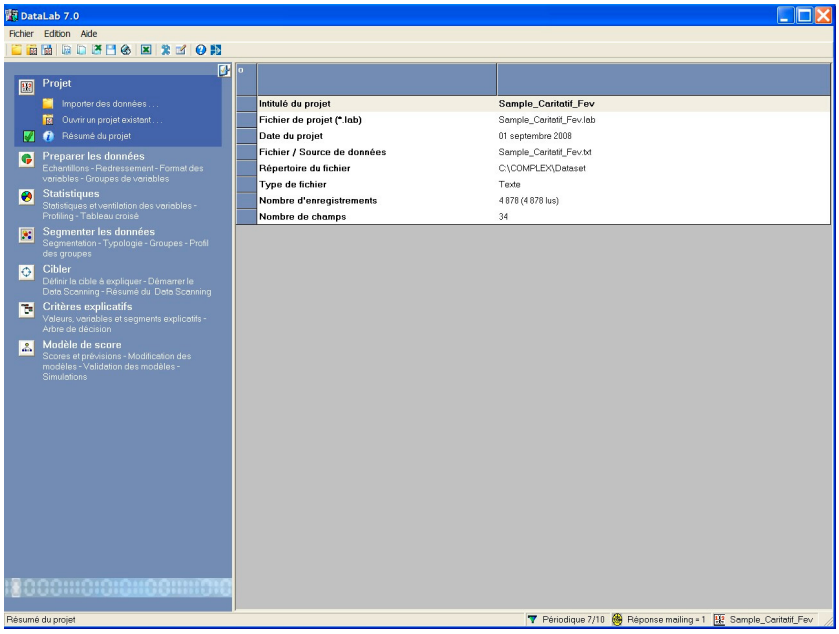
Pour la parie descriptive, DataLab offre les fonctionnalités :

- Statistiques
- Profiling
- Tableau croisé
- Segmentation
- Typologie

1.2 Organisation

DataLab est organisé en plusieurs zones :

- ☑ Un menu
- ☑ Une barre de boutons
- ☑ Une barre de message inférieure
- ☑ Une arborescence des résultats / liste des taches
- ☑ Deux zones de résultats

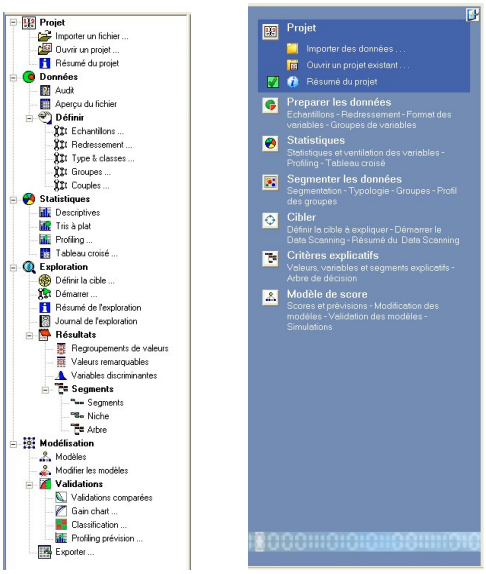


Vue générale de DataLab

Le menu et l'arborescence des résultats / liste des taches permettent d'accéder aux résultats issus de l'analyse. La barre de boutons permet d'accéder rapidement à une partie du menu. Elle permet aussi de filtrer l'affichage de certains résultats selon une variable donnée.

La barre de message inférieure indique l'état de l'application. La zone de résultats présente le résultat courant.

Barre de bouton de DataLab



Arborescence des résultats / Liste des taches

Le menu est organisé selon les thèmes :

- Fichier : permet d'ouvrir un projet ou d'importer un fichier et de quitter l'application
- Edition : permet d'exporter les résultats
- Aide : affiche l'aide et la fenêtre 'A propos de'

L'arborescence des résultats est organisée selon les thèmes :

- Projet : affiche une synthèse du projet courant
- Données : permet de créer un échantillon d'étude (enregistrements et variables), de le redresser, et de modifier les type des variables
- Description : affiche différentes statistiques sur les variables du projet
- Exploration : permet de définir la variable à expliquer, de lancer la recherche des nouvelles variables et d 'analyser les résultats ainsi que créer ou modifier un modèle

La liste des taches (affichage par défaut) est une façon alternative de représenter les tâches à effectuer :

- Projet : affiche une synthèse du projet courant
- Préparer les données : échantillons, redressement, format des variables, groupes de variables
- Statistiques : statistiques, profiling, tableau croisé
- Segmenter les données : segmentation, typologie
- Cibler : définir la cible, démarrer le data scanning
- Critères explicatifs : segments, niche, arbre de décision
- Modèle de score : modèle, modification des modèles, statistiques et simulations

1.3 Exportation des résultats

Tous les résultats obtenus avec DataLab peuvent être copiés, enregistrés, imprimés à partir du menu **Edition** ou des boutons correspondants. Ils peuvent donc être collés ou insérés individuellement dans toute application de bureautique telle que MS Word, Excel, Powerpoint. Les tables sont enregistrées au format texte et peuvent être copiées au format image (BMP) ou au format texte (champs séparés par des TAB rendant les données directement exportables vers un tableur). Chaque résultat peut aussi être exporté vers Excel (Création d'un nouveau classeur).

 : copie le résultat en format texte

 : copie le résultat en format image

 : exporte le résultat vers Excel

 : enregistre le résultat en format texte

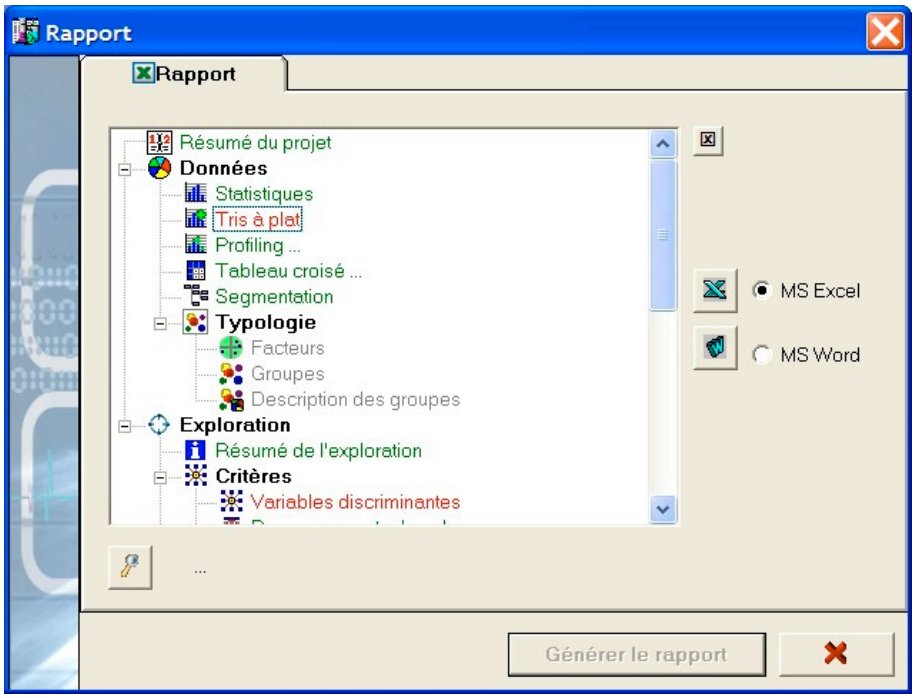
 : imprime le résultat

Lorsque deux fenêtres de résultats sont présentes, la fenêtre active est repérée par le symbole 'o' situé au haut à gauche de la fenêtre. Pour activer une fenêtre, il suffit de cliquer dessus. La fenêtre est alors celle qui sera exportée par défaut.

1.4 Rapport

Un rapport d'analyse est disponible à partir du menu **Edition** ou du bouton correspondant . Ce rapport configurable par l'utilisateur est au format Microsoft Word ou Microsoft Excel et est généré automatiquement.

: affiche la fenêtre permettant de générer le rapport



Fenêtre permettant la génération du rapport

Pour générer un rapport,

- Sélectionner le type de rapport désiré : Word ou Excel (plus rapide et plus souple)
- Sélectionner les résultats à mettre dans le rapport en cliquant sur le nom correspondant dans l'arborescence. Un résultat affiché en vert est sélectionné, en rouge est désélectionné,

en gris est non disponible. L'utilisateur passe du vert au rouge et vice versa en cliquant sur le résultat de la liste

- Entrer le nom du rapport à générer 
- Cliquer sur le bouton **Générer le rapport**

DataLab scanne automatiquement les résultats sélectionnés puis présente à l'utilisateur le rapport généré.

1.5 Caractéristiques techniques

DataLab fonctionne avec l'environnement minimal défini dans la Table 2.

Table 2 : Configuration minimale recommandée

Grandeur	Type recommandé
Système d'exploitation	MS Windows 32 bits (98, NT, 2000, XP, Vista)
Mémoire RAM	512 MB (1 Go recommandé, plus selon données)
Disque dur	50 Mo
Processeur	Pentium III (P4 recommandé)

Les limites relatives aux données traitées sont résumées en Table 3.

Table 3 : Limites relatives aux données

Grandeur	Limite
Nombre d'enregistrements	Aucune *
Nombre de variables en entrée	1000 *
Nombre de modalités pour les variables catégoriques	100
Nombre de groupes	100
Nombre de variables par groupe	30
Nombre de nouvelles variables évaluées	Aucune *
Nombre de variables dans la sélection	30
Nombre de nouvelles variables évaluées	Aucune *

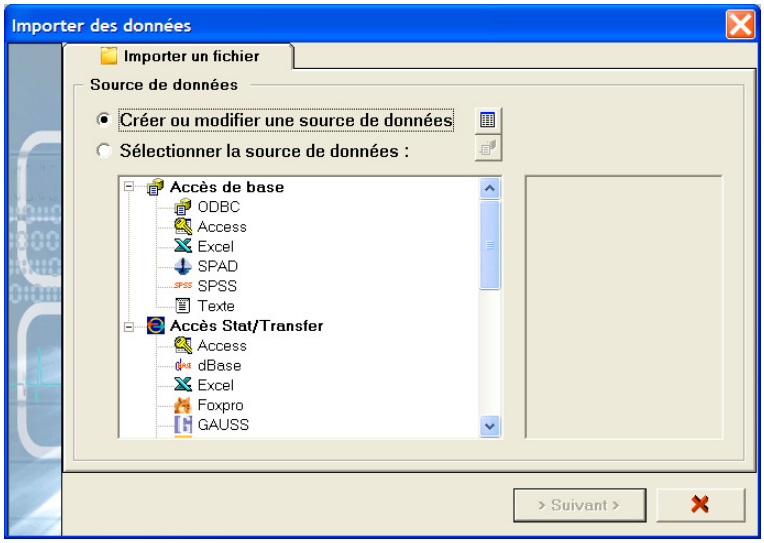
* Selon la mémoire RAM disponible sur l'ordinateur

2. Accès aux sources de données

2.1 Créer ou modifier une table

Le module DataBuilder de DataLab permet de constituer facilement et rapidement la table d'analyse (matrice individus x variables) à mettre en entrée de DataLab. Pour lancer le module à partir de DataLab :

- arborescence **Projet > Importer des données …**
- Cliquer sur l'option **Créer ou modifier une source de données**
- Cliquer sur le bouton  . DataBuilder est automatiquement lancé.



fenêtre d'importation des données

Pour plus de détails, se référer à la rubrique **7.10 Le module DataBuilder**.

2.2 Importer un fichier

DataLab permet l'import de fichiers plats ou tables provenant d'un accès de base aux sources de données standard et d'un accès via la technologie Stat/Transfer ® qui permet une connexion directe à plus de 20 types de fichiers statistiques ainsi qu'à la majorité des tables de SGBD.

Accès de base :

- ☑ Fichier ASCII avec séparateurs de champs : espace, virgule, TAB, point-virgule, pipe (|) (avec ou sans ligne d'intitulé des champs)
- ☑ Fichier dBase
- ☑ Table MS Access
- ☑ Feuille MS Excel
- ☑ Base SPAD (*.sba)
- ☑ Fichier SPSS (*.sav)
- ☑ Source de données ODBC (*)

Accès Stat/Transfer ® :

- ☑ MS Access
- ☑ dBase
- ☑ MS Excel
- ☑ FoxPro
- ☑ Gauss
- ☑ JMP
- ☑ Limdep
- ☑ Lotus
- ☑ Matlab
- ☑ ODBC (*)
- ☑ Osiris
- ☑ Mineset
- ☑ Minitab
- ☑ Paradox
- ☑ QuattroPro
- ☑ SAS
- ☑ SAS Transport
- ☑ S Plus
- ☑ SPSS
- ☑ SPSS Portable
- ☑ Stat
- ☑ Statistica
- ☑ Systat

(*) L'accès effectif aux sources de données ODBC est conditionnée par la disponibilité du connecteur associé nécessaire (pilote ODBC). L'accès Stat/Transfer est recommandé pour la connexion à une table via un pilote ODBC.

Pour importer un fichier :

1. sélectionner **Importer un fichier …** dans le menu **Fichier**
2. sélectionner la source de données
3. cliquer sur **Suivant** puis sélectionner le fichier dans la fenêtre
4. dans le cas d'un accès par pilote ODBC, entrer les paramètres demandés (utilisateur, mot de passe, serveur …).
5. un écran s'affiche qui synthétise la connexion. Dans le cas d'une base de données, sélectionner la table contenant les données
6. éventuellement, définir le ou les caractères associés à un enregistrement manquant ainsi que le format des dates
7. cliquer sur **Importer**

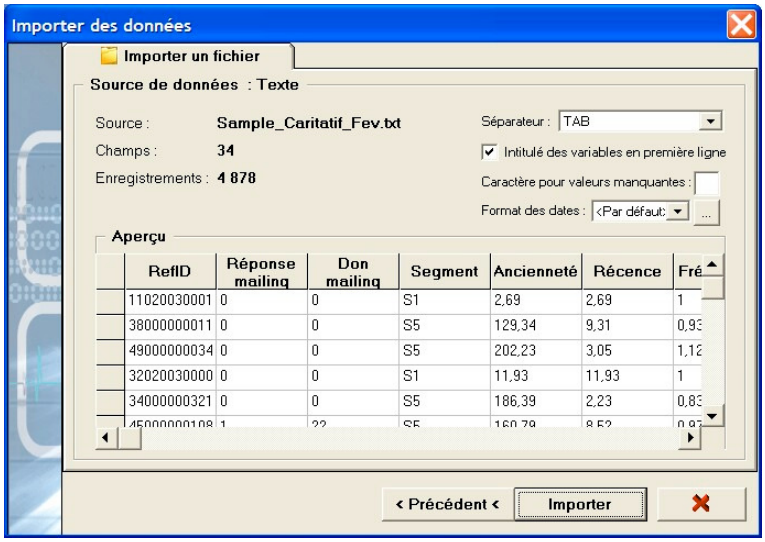
Les champs importés sont automatiquement classés sous les types catégorique, numérique ou date. Dans le dernier cas, toutes les dates du champ sont converties préalablement en nombre de jours jusqu'à la date de référence (par défaut, la date de l'importation). L'unité d'une variable date peut être modifiée ultérieurement (jour, mois, année).
Les formats de date reconnus sont les suivants :

- jjmmaa
- jj/mm/aa
- jj-mm-aa
- jjmmaaaa
- jj/mm/aaaa
- jj-mm-aaaa
- aammjj
- aa/mm/jj
- aa-mm-jj
- aaaammjj
- aaaa/mm/jj
- aaaa-mm-jj
- jjmmmaaaa (Oracle)

Le type (catégorique ou numérique), affecté automatiquement à chaque variable, peut être modifié ultérieurement.

Importation d'un fichier (étape 1)

Par défaut, DataLab reconnaît comme manquant tout enregistrement vide ou composé exclusivement d'un ou plusieurs caractères espace, ou encore composé d'un point (SAS). Un (des) caractère(s) spécifique(s) peuvent être défini(s) lors de l'import. Il sont alors ajoutés automatiquement à ces critères de valeur manquante.



Importation d'un fichier (étape 2)

Il existe une limite sur le nombre maximal de champs en entrée. Il n'existe pas de limites sur le nombre d'enregistrements autres que celle imposée par la mémoire vive de l'ordinateur.

2.3 Ouvrir un projet

DataLab permet d'ouvrir un projet préalablement enregistré. Le fichier projet (*.lab) permet l'accès direct aux données mais aussi aux résultats des analyses effectuées.

Pour ouvrir un projet :

1. sélectionner **Ouvrir un projet …** dans le menu 'Fichier' 
2. cliquer sur **Ouvrir** puis sélectionner le fichier projet dans la fenêtre d'ouverture

Lorsqu'un projet, faisant appel au même fichier que le projet courant, est ouvert, les données ne sont pas rechargées.
Un projet peut être enregistré en sélectionnant **Enregistrer le projet …** dans le menu **Fichier** 

2.4 Appliquer un format

Il est possible d'appliquer à un fichier importé le format d'un projet (*.lab) préalablement enregistré.

Pour appliquer à un format à un fichier importé dans DataLab :

1. sélectionner **Appliquer un format …** dans le menu **Fichier**
2. entrer le nom du fichier projet existant
3. valider

Le format est appliqué automatiquement à toutes les variables communes (présentant le même nom) et porte sur :

- ☑ le type (catégorique ou numérique)
- ☑ le découpage des variables numériques
- ☑ le libellé des classes et modalités
- ☑ les groupes de variables

3. Exporter variables et modèles

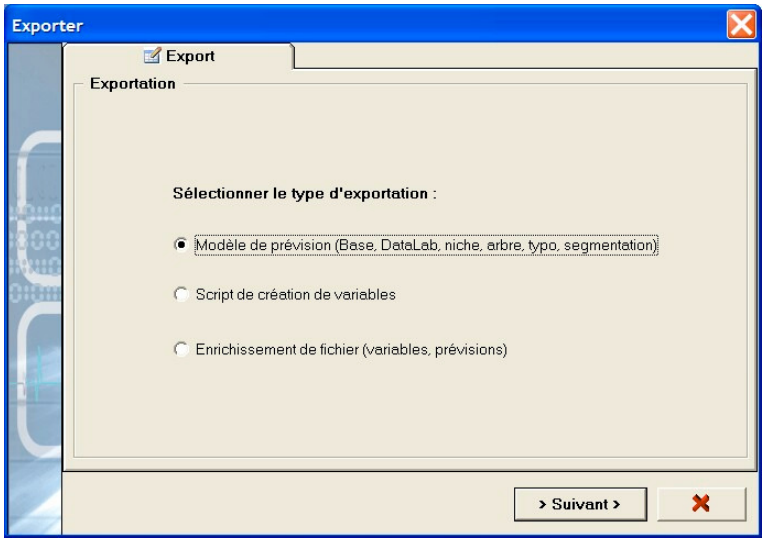
3.1 Exporter un modèle

Il est possible d'exporter un modèle (Base ou DataLab) dans un fichier au format :

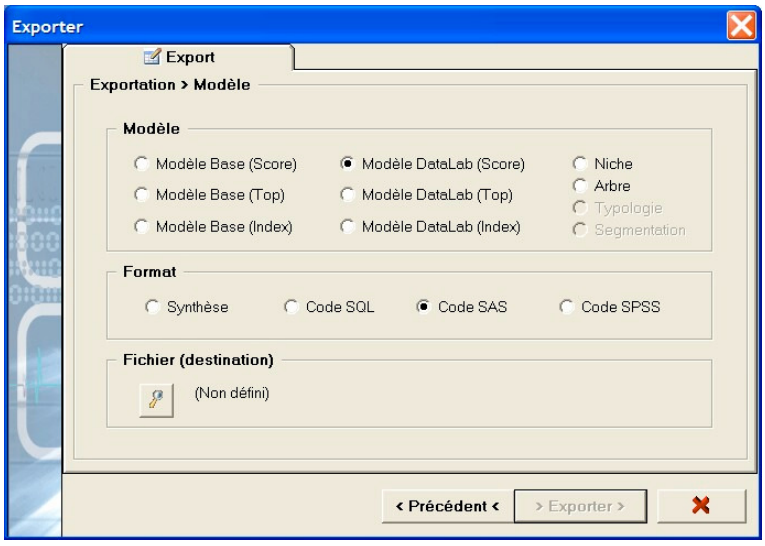
- ASCII (synthèse)
- Code SQL
- Code SAS
- Code SPSS

Les informations fournies comprennent les paramètres du modèle ainsi que les éventuelles transformations préalables permettant de calculer les variables du modèle (écrêtage, remplacement des valeurs manquantes, variables intermédiaires).
Pour exporter un modèle :

1. sélectionner **Exporter …** dans le menu **Fichier** 
2. cocher le bouton d'option **Modèle de prévision**
3. cliquer sur **Suivant**
4. sélectionner le modèle à exporter
5. sélectionner le format du code d'exportation
6. sélectionner le fichier de destination en cliquant sur le bouton 
7. cliquer sur **Exporter**



- Exportation du code d'un modèle (étape 1)



- Exportation du code d'un modèle (étape 2)

Pour plus de détails, se référer à la rubrique **7.8 L'export**.

3.2 Exporter un script de création de variables

Il est possible d'exporter les scripts de création de variables issues de l'analyse pour des analyses complémentaires. Les variables concernées sont :

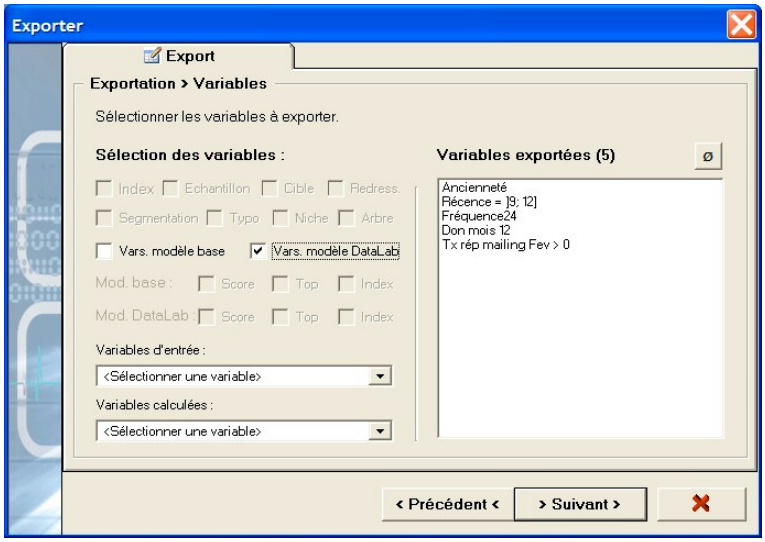
- ☑ Variables de la sélection de base
- ☑ Variables de la sélection DataLab
- ☑ Tout variable d'entrée
- ☑ Tout variable calculée …

Les formats de scripts proposés sont :

- Requête SQL (script de création des tables et script de remplissage des champs) (Figure 4).
- Code SAS
- Code SPSS

Pour exporter un script :

1. sélectionner **Exporter …** dans le menu 'Fichier' 
2. cocher le bouton d'option **Script de création de variables**
3. cliquer sur **Suivant**
4. sélectionner les variables à exporter. Cliquer sur **Suivant**
5. sélectionner le format du script d'exportation (SQL, SAS, SPPS). Cliquer sur **Suivant**
6. sélectionner le fichier de destination en cliquant sur le bouton **Fichier**
7. cliquer sur **Exporter**



Exportation du script des variables (étape 2)

Pour plus de détails, se référer à la rubrique **7.8 L'export**.

3.3 Enrichir un fichier

Un fichier ou une table de base de données externe peuvent être enrichis avec les variables ou les prévisions de DataLab à partir des enregistrements du fichier chargé dans DataLab ou bien d'une source de données externe.

Le format d'export (écriture des enregistrements et lecture des données externes) peut être choisi parmi les mêmes sources que celles disponibles pour l'importation (Accès de base, Accès Stat/Transfer).

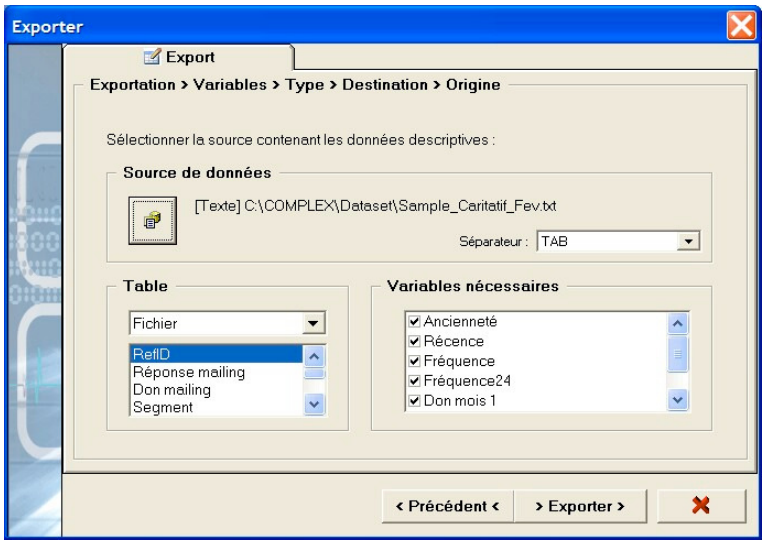
Les variables qu'il est possible d'exporter sont à choisir parmi :

- ☑ Cible
- ☑ Echantillon
- ☑ Poids issus d'un redressement éventuel
- ☑ Prévisions issues des modèles de base et DataLab
- ☑ Segmentation et typologie
- ☑ Niche et arbre
- ☑ Variables de la sélection de base
- ☑ Variables de la sélection DataLab
- ☑ Toute variable d'entrée
- ☑ Toute variable calculée …

Pour exporter des variables :

1. sélectionner **Exporter …** dans le menu 'Fichier' 
2. cocher le bouton d'option **Enrichissement de fichier**
3. cliquer sur **Suivant**
4. sélectionner les variables à exporter. Cliquer sur **Suivant**
5. sélectionner le fichier de destination (fichier exporté) en cliquant sur le bouton **Source de données**  . Cliquer sur **Suivant**

6. sélectionner l'origine des enregistrements (données chargées dans DataLab ou bien provenant d'une source externe). Cliquer sur **Suivant**
7. dans le cas de données présentes dans une base de données externe, sélectionner la source de données externes en cliquant sur le bouton **Source de données**. Sélectionner la table.
8. cliquer sur **Exporter** lorsque les champs nécessaire à l'exportation ont tous été retrouvés dans la table.



Exportation des variables (étape 3)

Pour plus de détails, se référer à la rubrique **7.8 L'export**.

4. Fonctionnalités

4.1 Projet

4.1.1 Résumé du projet

Projet > Résumé du projet

Affiche une synthèse du projet.

Intitulé du projet	**Risque crédit**
Fichier de projet (*.lab)	Risque crédit.lab
Date du projet	24 mars 2006
Fichier	Risque crédit.txt
Répertoire du fichier	C:\COMPLEX\Dataset
Type de fichier	Texte
Nombre d'enregistrements	4 000 (4 000 lus)
Nombre de champs	17

```
Résumé du projet.
```

4.2 Préparer les données

Ce module permet de visualiser les données chargées, créer des échantillons d'apprentissage et de validation, éventuellement redressés. Les fenêtres de cette rubrique permettent aussi à l'utilisateur d'appliquer son expertise aux données importées.

4.2.1 Audit du fichier

Préparer les données > Audit

Affiche un audit du fichier, sous la forme d'une ventilation de chaque champ selon les différentes valeurs trouvées.

4.2.2 Aperçu du fichier

Préparer les données > Aperçu du fichier

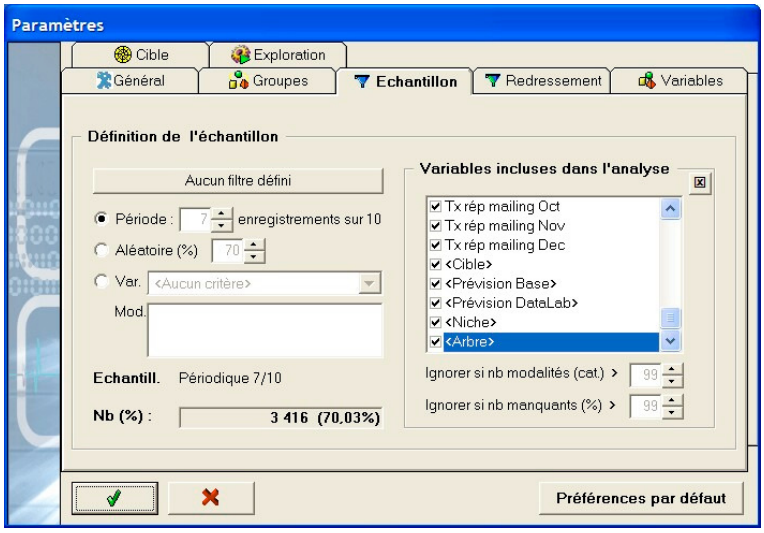
Affiche un aperçu du fichier importé.

4.2.3 Définir un échantillon

Préparer les données > Définir un échantillon

Affiche une fenêtre permettant de créer un échantillon d'étude :

- selon une sélection périodique des enregistrements (séquentiellement, N enregistrements sur 10)
- selon une sélection aléatoire des enregistrements
- selon une ou plusieurs valeurs prise par une variable du fichier d'entrée
- en excluant des enregistrements définis par un ou plusieurs filtre (conditions).



Définir un échantillon.

Les variables à ne pas prendre en compte pour l'analyse peuvent être ignorées en décochant les labels correspondants. Par défaut, les variables sont automatiquement exclues lorsque :

- ☑ elles ne possèdent qu'une valeur unique (exclusion définitive)
- ☑ elles ne présentent aucune valeur en double (considérées comme des identifiants)
- ☑ elles sont catégoriques et possèdent plus de N modalités (par défaut, N=99)
- ☑ elles présentent un taux de valeurs manquantes supérieur à P (par défaut, P=99%)

L'ensemble des statistiques et analyses (hors 'Audit du fichier') porte sur l'échantillon uniquement. Les enregistrements et champs non définis comme faisant partie de l'échantillon sont dans ce cas ignorés des analyses. Les enregistrements hors échantillon sont automatiquement utilisés pour la validation des modèles.

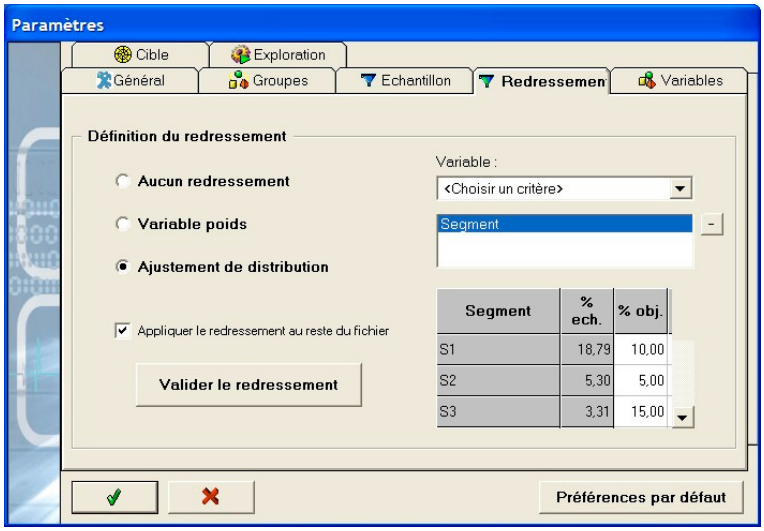
4.2.4 Redresser un échantillon

Préparer les données > Redresser l'échantillon

Affiche une fenêtre permettant de redresser l'échantillon (pondération des enregistrements) selon 2 modes :

- Le vecteur poids est une variable désignée
- Le vecteur poids est déterminé de manière itérative afin de respecter la répartition désirée des critères de redressement (5 au maximum)

L'utilisateur peut aussi appliquer le même redressement sur la partie **hors échantillon**, dans le cas où un échantillon a été défini, en cochant la case **Appliquer le redressement au reste du fichier**. Si cette case est désactivée, chaque enregistrement de la partie complémentaire à l'échantillon se voit affectée un poids unitaire.
L'ensemble des statistiques et analyses (hors **Audit du fichier**) porte sur l'échantillon redressé. Les poids sommant à l'unité, le nombre d'enregistrements est identique au fichier non redressé.



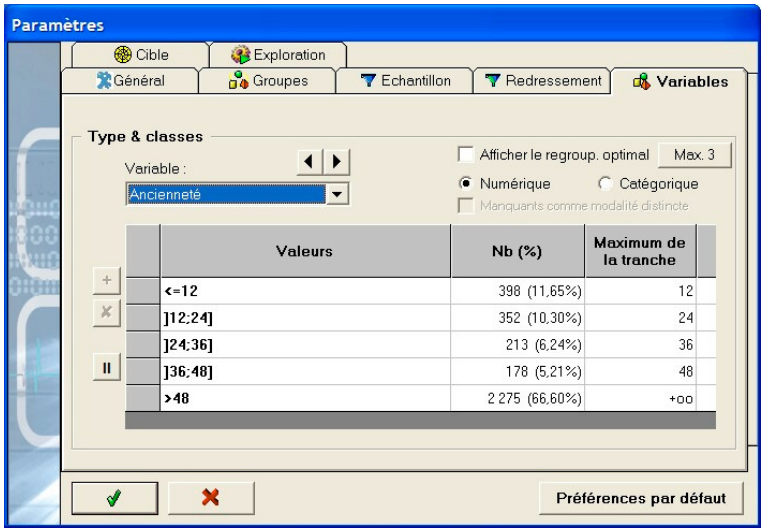
Redresser un échantillon.

Pour plus de détails, se référer à la rubrique **7.2 Le redressement**.

4.2.5 Format des variables

Préparer les données > Format des variables

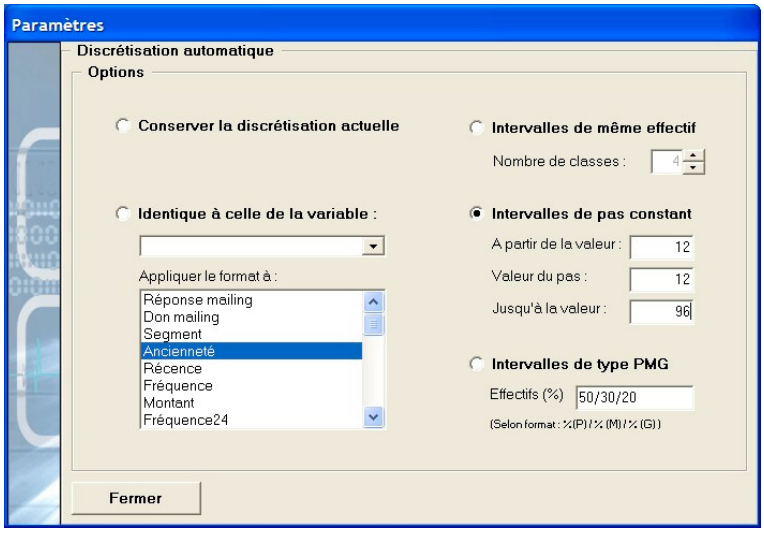
Affiche une fenêtre permettant de modifier les classes de valeurs de chaque variable continue en entrant une nouvelle valeur maximum pour la tranche. Il est aussi possible d'insérer ou de supprimer une classe (respectivement boutons + et - . Pour modifier le nom d'un label, cliquer dessus et entrer le nom.



Définition du format d'une variable

Les classes de valeurs des variables continues peuvent être automatiquement ajustées (intervalles de même effectif ou de même largeur) en cliquant sur le bouton **II** .

Pour plus de détails, se référer à la rubrique **7. 6 Discrétisation et regroupement de modalités**



Discrétisation automatique d'une variable

Il est possible de convertir une variable de type numérique en variable de type catégorique et vice versa. Pour effectuer cette opération, cliquer sur la case à option correspondante.

Cocher la case **Manquants = modalité distincte** convertit les valeurs manquantes de la variable en une modalité supplémentaire. Dans ce cas la variable est automatiquement convertie (si elle ne l'est déjà) en type catégorique.

Cocher la case **Afficher le regroupement optimal** permet d'afficher et de modifier les classes correspondant au regroupement optimal (regroupement de modalité, discrétisation) déterminé au cours de l'exploration. Cette option n'est disponible que lorsque les résultats de l'exploration (Data Scanning) sont disponibles.

Le nombre de regroupements optimaux peut être fixé avant l'exploration en entrant la valeur via le bouton **Max. 3**.

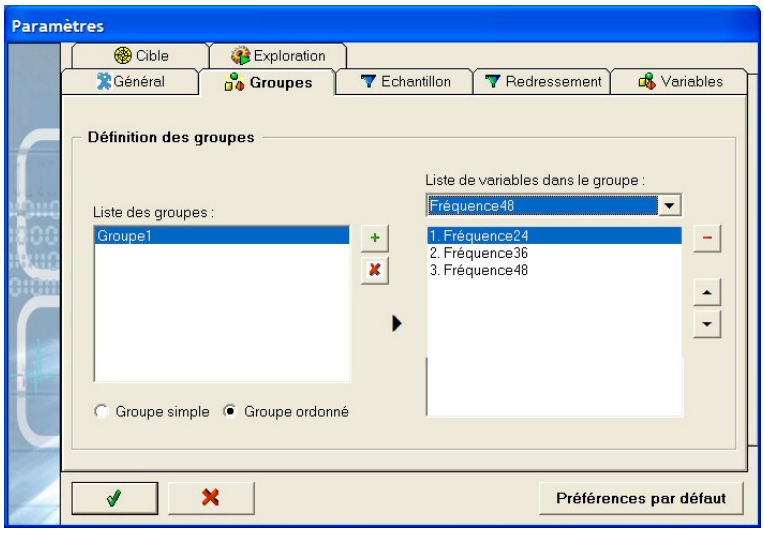
4.2.6 Groupes de variables

Préparer les données > Groupes de variables

Affiche une fenêtre permettant de définir des groupes de variables. Les groupes de variables sont utilisés pour générer de nouvelles variables évaluées lors de l'exploration. Les variables composant un groupe peuvent être numérique ou catégoriques. Dans ce dernier cas, une modalité doit être associée à la variable.

Pour définir un nouveau groupe de variable :

1. Cliquer sur le bouton + et entrer un nom pour le nouveau groupe
2. Sélectionner le type de groupe : simple (ensemble de variable similaires non ordonnées) ou ordonné (ensemble de variable similaires ordonnées en temps)
3. Cliquer sur le nom du groupe dans la liste **Liste des groupes**
4. Sélectionner successivement tous les variables composant le groupe
5. Eventuellement, dans le cas d'un groupe ordonné, ordonner les variables de la plus ancienne à la plus récente



Création d'un groupe de variables

Pour plus de détails, se référer à la rubrique **7. 7 Les groupes**

4.3 Statistiques

Cette rubrique permet de d'obtenir différentes statistiques sur les variables et l'interaction de plusieurs variables.

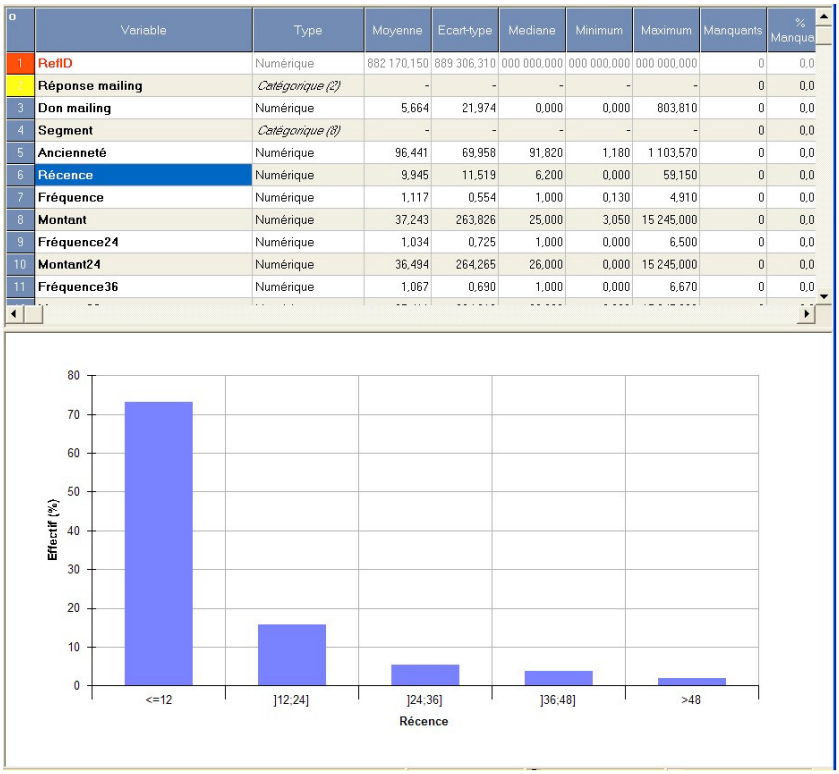
4.3.1 Statistiques des variables

Statistiques > Statistiques des variables

Affiche des statistiques sur les variables du fichier :

- Type (numérique, catégorique, date)
- Moyenne
- Médiane
- Ecart-type
- Minimum
- Maximum
- Nombre de manquants
- Taux de manquants

En cliquant sur un libellé de variable, des statistiques complémentaires sont automatiquement affichées sous forme graphique.
En double-cliquant sur un libellé de variable, la fenêtre de modification **Format des variables** est affichée automatiquement.



	Variable	Type	Moyenne	Ecart-type	Mediane	Minimum	Maximum	Manquants	% Manqua
1	RefID	Numérique	882 170,150	889 306,310	000 000,000	000 000,000	000 000,000	0	0,0
2	Réponse mailing	*Catégorique (2)*	-	-	-	-	-	0	0,0
3	Don mailing	Numérique	5,664	21,974	0,000	0,000	803,810	0	0,0
4	Segment	*Catégorique (8)*	-	-	-	-	-	0	0,0
5	Ancienneté	Numérique	96,441	69,958	91,820	1,180	1 103,570	0	0,0
6	Récence	Numérique	9,945	11,519	6,200	0,000	59,150	0	0,0
7	Fréquence	Numérique	1,117	0,554	1,000	0,130	4,910	0	0,0
8	Montant	Numérique	37,243	263,826	25,000	3,050	15 245,000	0	0,0
9	Fréquence24	Numérique	1,034	0,725	1,000	0,000	6,500	0	0,0
10	Montant24	Numérique	36,494	264,265	26,000	0,000	15 245,000	0	0,0
11	Fréquence36	Numérique	1,067	0,690	1,000	0,000	6,670	0	0,0

▸ Statistiques des variables

4.3.2 Ventilations

Statistiques > Ventilations

Affiche des statistiques pour chaque variable selon des les valeurs ou intervalles observés. Les intervalles peuvent être modifiés (cf. **Préparer les données > Format des variables …**).

4.3.3 Profiling

Statistiques > Profiling

Affiche le profiling (croisement) de l'ensemble des variables avec une variable (par défaut, la variable à expliquer). Pour chaque croisement, le tableau donne l'effectif, le pourcentage et l'indice de sur-representativité. Les sur-représentations sont indiquées sous la forme d'indice base 100. Les indices en vert indiquent un pourcentage significativement plus important que la moyenne, en rouge un pourcentage significativement plus faible que la moyenne, en noir une différence non significative.

Il est possible de stocker jusqu'à 10 profilings dans un projet.

(Réponse mailing)	Ensemble		0			1		
Variables & valeurs	Nb	%	Nb	%	Indice	Nb	%	Indice
Récence								
Moyenne	3 416	9,945	2 933	10,343		483	7,529	
<=12	2 501	100,00%	2 070	82,77%	96	431	17,23%	122
]12;24]	539	100,00%	503	93,32%	109	36	6,68%	47
]24;36]	184	100,00%	174	94,57%	110	10	5,43%	38
]36;48]	127	100,00%	121	95,28%	111	6	4,72%	33
>48	65	100,00%	65	100,00%	116	0	0,00%	100
Fréquence								
Moyenne	3 416	1,117	2 933	1,078		483	1,355	
<=0,83	901	100,00%	829	92,01%	107	72	7,99%	57
]0,83; 1]	930	100,00%	834	89,68%	104	96	10,32%	73

Profiling

Pour plus de détails, se référer à la rubrique **7. 3 Le profiling**

4.3.4 Tableau croisé

Statistiques > Tableau croisé

Affiche le tableau croisé de deux variables avec une variable (par défaut, la variable à expliquer). Plusieurs statistiques sont disponibles : effectif, moyenne, somme, % …

Il est possible de stocker jusqu'à 10 tableaux croisés dans un projet.

o	(Réponse mailing [%])		
	Réponse mailing		
Récence	0	1	TOTAL
<=12	82,77%	17,23%	**100,00%**
]12;24]	93,32%	6,68%	**100,00%**
]24;36]	94,57%	5,43%	**100,00%**
]36;48]	95,28%	4,72%	**100,00%**
>48	100,00%	0,00%	**100,00%**
TOTAL	**85,86%**	**14,14%**	**100,00%**

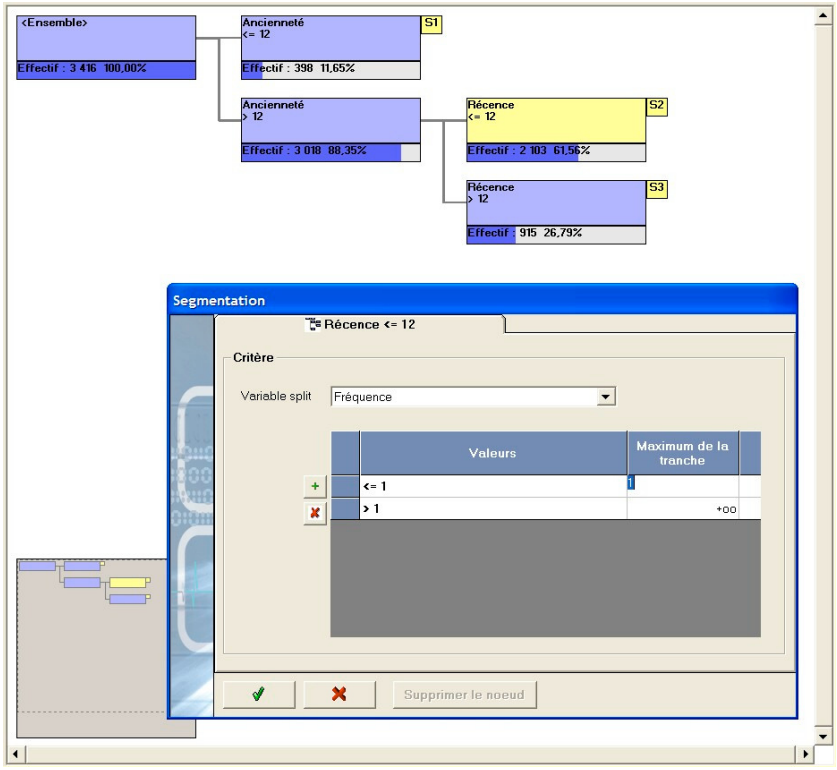
▸ Tableau croisé

4.4 Segmenter les données

4.4.1 Segmentation

Segmenter les données > Segmentation

Affiche une fenêtre permettant de construire une segmentation, c'est à dire un découpage de la base.



Segmentation

Pour chaque nœud,

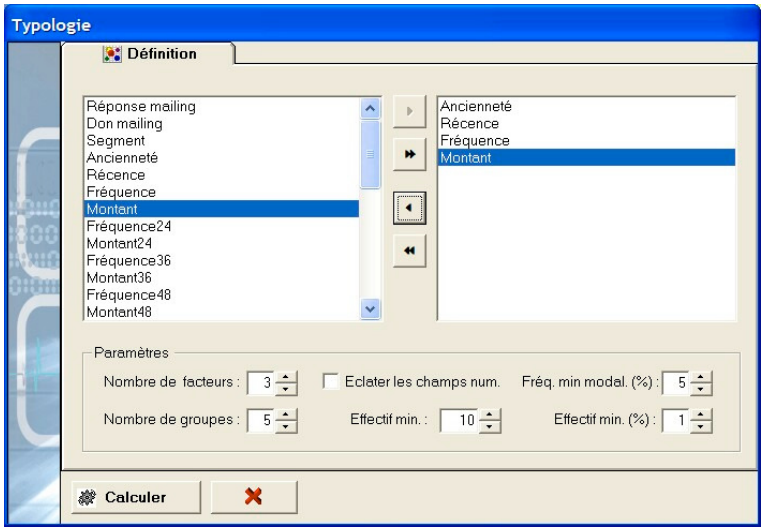
- Double-cliquer sur le nœud
- Dans la fenêtre qui s'affiche, sélectionner la variable à découper
- Ajouter ou supprimer une classe avec les boutons correspondants

- Si la variable est numérique, modifier le seuil de chaque classe dans la colonne **Maximum de la tranche**
- Si la variable est catégorique, sélectionner la modalité dans la colonne **Valeurs**, couper la (CTRL + X), puis coller la dans une autre classe (CTRL + V) sans omettre le point-virgule de séparation des modalités

4.4.2 Typologie

Segmenter les données > Définir une typologie

Affiche une fenêtre permettant de construire une typologie, c'est à dire un ensemble de groupes homogènes selon un ensemble de critères.



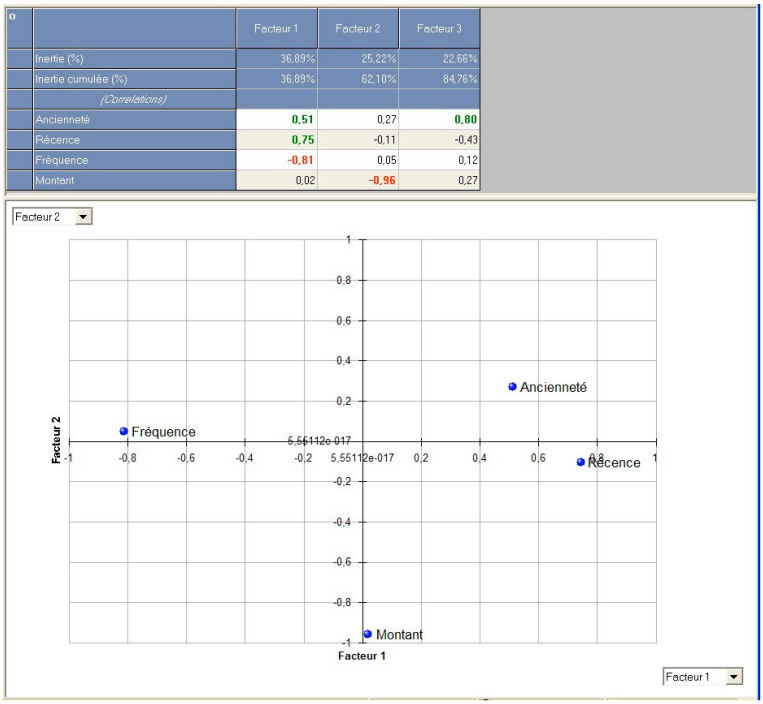
Fenêtre de création de la typologie

Pour plus de détails, se référer à la rubrique **7. 9 La typologie**

4.4.3 Facteurs principaux

Segmenter les données > Facteurs principaux

Affiche une fenêtre permettant de visualiser les corrélations entre les facteurs issus d'une ACP et les critères participant à la typologie.



	Facteur 1	Facteur 2	Facteur 3
Inertie (%)	36,89%	25,22%	22,66%
Inertie cumulée (%)	36,89%	62,10%	84,76%
(Corrélations)			
Ancienneté	**0,51**	0,27	**0,80**
Récence	**0,75**	-0,11	-0,43
Fréquence	**-0,81**	0,05	0,12
Montant	0,02	**-0,96**	0,27

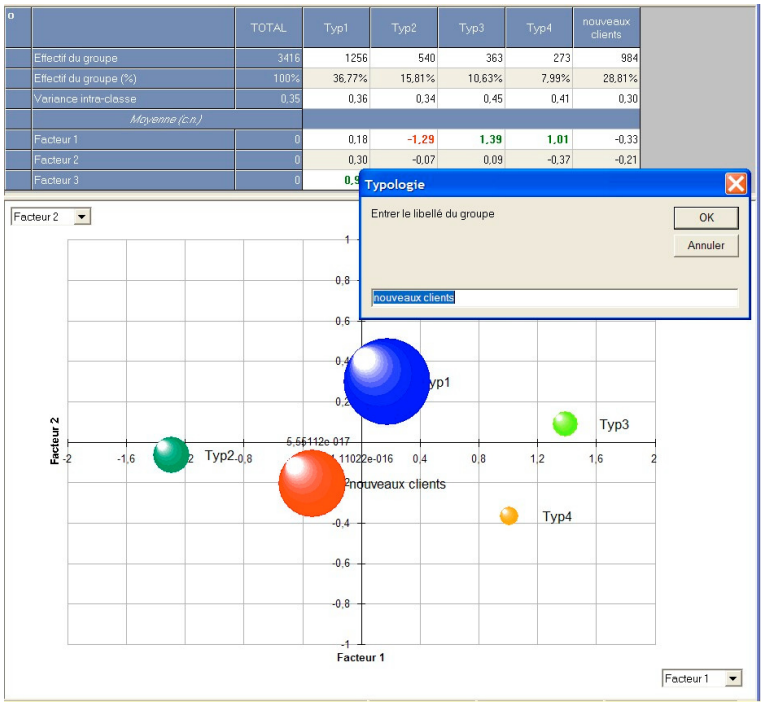
▸ Facteurs principaux

Pour plus de détails, se référer à la rubrique **7. 9 La typologie**

4.4.4 Groupes de la typologie

Segmenter les données > Groupes de la typologie

Affiche une fenêtre permettant de projeter les facteurs principaux sur les groupes de la typologie.



Groupes de la typologie

Pour plus de détails, se référer à la rubrique **7. 9 La typologie**

4.4.5 Description des groupes

Segmenter les données > Description des groupes

Affiche une fenêtre permettant de visualiser des statistiques sur les groupes.

Pour plus de détails consulter la partie **7.9 La typologie**

		Typ1		Typ2		Typ3		Typ4		nouveaux clients	
		Moy.	Ec.Typ.	Moy.	Ec.Typ.	Moy.	Ec.Typ.	Moy.	Ec.Typ.	Moy.	Ec.
	Ancienneté	160,30	126,37	42,20	128,16	158,50	127,43	50,59	124,31	33,64	1
	Récence	5,05	14,33	2,96	12,50	25,04	18,49	26,29	16,33	7,56	
	Fréquence	1,04	1,46	1,84	1,44	0,69	1,41	0,59	1,34	1,03	
	Montant	32,28	70,13	32,94	66,10	29,59	74,45	37,17	98,87	32,57	

Description des groupes

Pour plus de détails, se référer à la rubrique **7. 9 La typologie**

4.5 Cibler

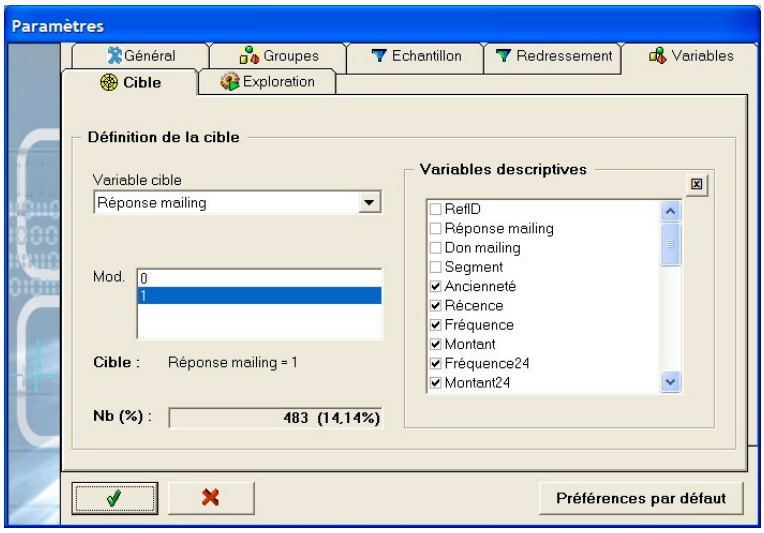
4.5.1 Définir la cible

Cibler > Définir la cible

Affiche une fenêtre permettant de définir le comportement à expliquer.

Pour définir une cible :

1. Sélectionner la variable dans la liste déroulante **Variable**
2. Sélectionner la ou les valeurs (ou plages de valeurs) de la variable



Définition de la cible

Pour considérer l'ensemble des valeurs de la cible (cible continue), cocher la case **Considérer toutes les valeurs**. Pour sélectionner plusieurs modalités dans la liste, maintenir enfoncée la touche CTRL.

Le nombre et le pourcentage d'enregistrements cible résultant est indiqué automatiquement.

L'utilisateur peut aussi sélectionner dans cette fenêtre les variables susceptibles d'être des variables explicatives de la cible. La variable cible est automatiquement et définitivement exclue de la liste des variables explicatives. Les variables exclues de l'analyse (cf. 4.2.3 Définir un échantillon) ne peuvent être sélectionnées.

4.5.2 Démarrer le Data Scanning

Cibler > Démarrer le Data Scanning

Affiche une fenêtre permettant de démarrer l'exploration (Data Scanning). L'utilisateur peut en particulier sélectionner un certain nombre d'options permettant une exploration plus poussée ou plus rapide.

Général

- Niveau de confiance : définition du seuil de confiance
- Fréquence minimum des critères : pourcentage minimum des nouveau critères à évaluer
- Utiliser un modèle de régression logistique : dans le cas d'une cible binaire uniquement, choisir un cadre d'exploration linéaire ou logistique (par défaut)

Etapes

- Transformations : Réalise les opérations de transformations
- Combinaisons : Applique les opérateurs de combinaisons de variables
- Sélections : Calcule les sélections optimales (modèles)

Options détaillées

- **Transformations**
  - Ecréter les valeurs à x écart type : définit la valeur en nombre d'écart type pour l'écrêtage automatique
  - Traitement des valeurs manquantes : Permet à l'utilisateur de définir (variable par variable ou globalement) le type de traitement à appliquer automatiquement pour le remplacement des valeurs manquantes (moyenne, médiane, valeur utilisateur).
  - Variables date : applique les transformations sur les variables date
  - Variables numériques : applique les transformations sur les variables numériques
  - Variables catégorique : applique les transformations sur les variables catégoriques

- ☑ Transformation identité : évalue le pouvoir discriminant de chaque variable d'entrée avec pour seule modification éventuelle un écrêtage et le remplacement des valeurs manquantes par la moyenne, la médiane ou une valeur définie par l'utilisateur.
- ☑ Valeurs remarquables : évalue le pouvoir discriminant des valeurs remarquables pour chaque variable
- ☑ Binarisation : éclate les variables catégoriques et évalue le pouvoir discriminant de chaque modalité
- ☑ Valeurs manquantes : évalue le pouvoir discriminant des valeurs manquantes pour chaque variable
- ☑ Discrétisation optimale : recherche la discrétisation optimale pour chaque variable numérique et évalue le pouvoir discriminant des regroupements obtenus
- ☑ Regroupements optimaaux : recherche les regroupements optimaux pour chaque variable catégorique et évalue le pouvoir discriminant des regroupements obtenus
- ☑ Transformations mathématiques : transforme les variables numérique d'entrée selon des opérateurs mathématiques (Carré, inverse, log10, racine) et évalue le pouvoir discriminant de la variable résultante

- Max. regroupements optimaux : nombre maximal de regroupements attendus
- Init. : nombre maximal de classes initiales pour le calcul des discrétisations optimales
- Conserver les variables non significatives : toutes les variables transformées, même non significatives statistiquement, sont conservée pour analyse ultérieures (combinaisons, sélection, ...)

- **Combinaisons**
  - Groupes de variables : autorise l'évaluation des variables issues des groupes de variables (si définis)
    - ☑ Moyenne : évalue le pouvoir discriminant de la valeur moyenne calculée sur les variables du groupe
    - ☑ Ecart-type : évalue le pouvoir discriminant de l'écart-type calculé sur les variables du groupe
    - ☑ Min : évalue le pouvoir discriminant de la valeur minimum calculée sur les variables du groupe
    - ☑ Max : évalue le pouvoir discriminant de la valeur maximum calculée sur les variables du groupe
    - ☑ Nombre>0 : évalue le pouvoir discriminant du nombre de valeurs positives calculé sur les variables du groupe
    - ☑ Part variable : évalue le pouvoir discriminant de la part de chaque variable dans le groupe
    - ☑ Pente : évalue le pouvoir discriminant de la pente calculée sur les variables du groupe (groupe ordonné)
    - ☑ Variation : évalue le pouvoir discriminant de la variation calculée sur les 2 dernières variables du groupe (groupe ordonné)

- Variables binaires : autorise l'évaluation des combinaisons de variables binaires
  - ☑ Et : évalue le pouvoir discriminant d'une combinaison reposant sur l'opérateur ET
  - ☑ Et Non : évalue le pouvoir discriminant d'une combinaison reposant sur l'opérateur ET NON
  - ☑ Ou : évalue le pouvoir discriminant d'une combinaison reposant sur l'opérateur OU
  - ☑ Ou Non : évalue le pouvoir discriminant d'une combinaison reposant sur l'opérateur OU NON
  - ☑ Ou Excl. : évalue le pouvoir discriminant d'une combinaison reposant sur l'opérateur OU EXCLUSIF
- Variables numériques : autorise l'évaluation des combinaisons de variables numériques
  - ☑ + : évalue le pouvoir discriminant d'une combinaison reposant sur l'opérateur ADDITION
  - ☑ - : évalue le pouvoir discriminant d'une combinaison reposant sur l'opérateur SOUSTRACTION
  - ☑ / : évalue le pouvoir discriminant d'une combinaison reposant sur l'opérateur DIVISION
  - ☑ * : évalue le pouvoir discriminant d'une combinaison reposant sur l'opérateur MULTIPLICATION
  - ☑ Min : évalue le pouvoir discriminant d'une combinaison reposant sur l'opérateur MINIMUM(a,b)
  - ☑ Max : évalue le pouvoir discriminant d'une combinaison reposant sur l'opérateur MAXIMUM(a,b)
  - ☑ < : évalue le pouvoir discriminant d'une combinaison reposant sur l'opérateur INF(a,b)
  - ☑ = : évalue le pouvoir discriminant d'une combinaison reposant sur l'opérateur EGAL(a,b)
  - ☑ > : évalue le pouvoir discriminant d'une combinaison reposant sur l'opérateur SUP(a,b)
- Evaluer les variables triplet : évalue les combinaison de 3 variables
- A partir des vars du top x% : définit le pourcentage des variables les plus discriminantes à combiner 2 à 2 (100% signifie que toute les variables d'entrée et transformées sont combinées). Cette option réduit considérablement le temps de calcul.

- **Sélections**
  - Base : autorise la recherche de la sélection de base (à partir des variables d'entrée uniquement)
  - DataLab : autorise la recherche de la sélection DataLab (à partir de toutes les variables)
  - Nombre maximum de variables : nombre maximal de variables entrant dans les sélections
  - Recherche parmi les meilleures variables : entrer un nombre spécifiant, pour chaque variable d'entrée, les top N variables calculées incluant cette variable qui présentent la

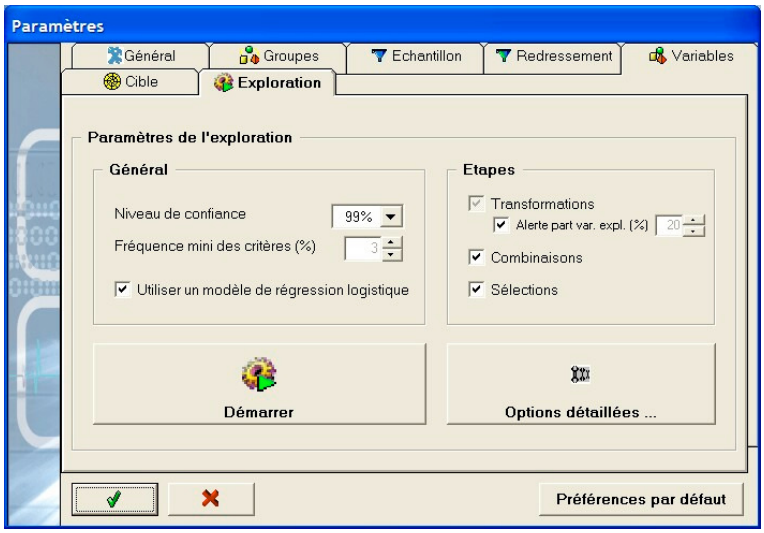
part de variance expliquée la plus élevée et qui seront utilisées pour les sélections. Cette option permet de réduire de manière importante la recherche de la sélection de variables la plus explicative.

- Forcer : sélectionner les variables à forcer dans les sélections

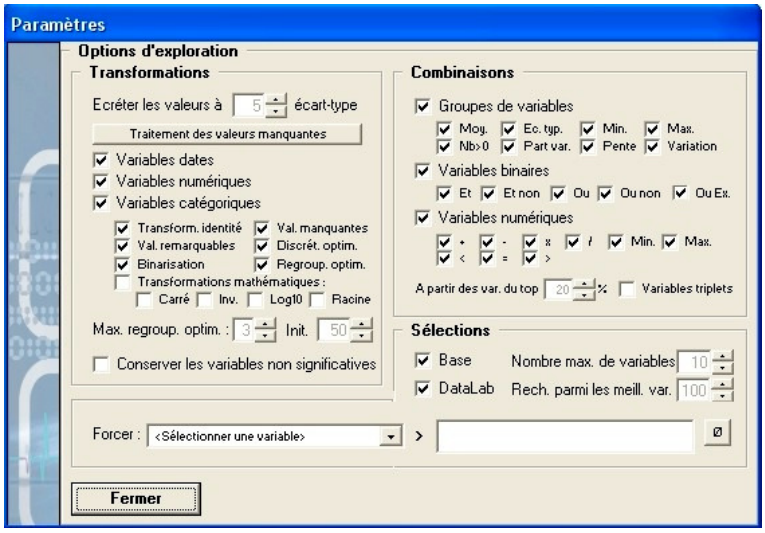
Les options de recherche **Conserver les variables non significatives** et **Evaluer les variables triplet** permettent une recherche plus complète mais aussi significativement plus longue.

L'utilisation d'un modèle de régression logistique, bien que plus adapté à l'explication d'une variable binaire, nécessite des temps de calcul plus importants que la régression linéaire.

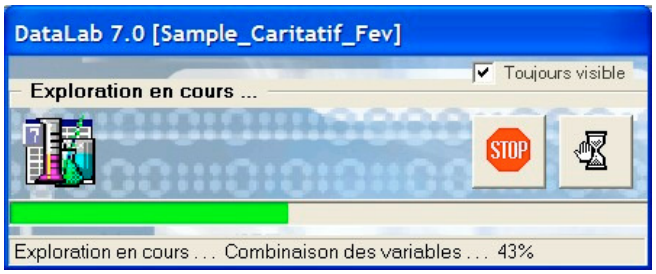
L'exploration est lancée en cliquant sur le bouton **Explorer** symbolisé par l'icône de DataLab. Une fenêtre réduite remplace alors la fenêtre de DataLab et l'exploration démarre. Cette dernière peut être arrêtée définitivement en cliquant sur le bouton **Stop** ou simplement interrompue un instant en cliquant sur le bouton **Pause**.



Paramétrage et lancement de l'exploration



Options détaillées de l'exploration



Fenêtre de progression de l'exploration

Pour plus de détails, se référer à la rubrique **7. 5 L'exploration ou Data Scanning**

4.5.3 Résumé du Data Scanning

Cibler > Résumé du Data Scanning

Affiche une fenêtre synthétisant l'exploration, notamment les principaux paramètres utilisés pour l'exploration et quelques résultats majeurs dont :

- Nombre de variables de base (Evaluées). Sont conservées les variables dépassant le seuil de fréquence minimum et statistiquement significatives (test de Fisher)
- Nombre de variables intermédiaires (Evaluées). Sont conservées les variables dépassant le seuil de fréquence minimum et statistiquement significatives (test de Fisher)
- Nombre de variables combinées (Evaluées). Sont conservées les variables dépassant le seuil de fréquence minimum et statistiquement significatives (test de Fisher)
- Part de variance expliquée de la variable la plus explicative
  - Gain sur la variable de base la plus explicative (gain relatif.)
- Part de variance expliquée de la sélection
  - Gain sur la sélection de base (gain relatif.)

Un journal de l'exploration est disponible à l'issue du Data Scanning tout en bas de la feuille de résumé dans la zone **Journal de l'exploration**. Ce journal détaille toutes les opérations élémentaires réalisées lors de l'exploration : transformations, combinaisons, sélection. Le log donne notamment des informations complémentaires sur l'écrêtage des variables, sur les tests de regroupements optimaux, les différentes combinaisons testées, …

4.6 Critères explicatifs

Cette rubrique regroupe les résultats de l'exploration.

4.6.1 Variables explicatives

Critères explicatifs > Variables explicatives

Cette rubrique regroupe l'ensemble des résultats relatifs aux variables issues de l'exploration. Les variables discriminantes affichées peuvent être filtrées par type : issues des variables de base, transformées, combinées ou liste exhaustive des variables en entrée (discriminantes ou pas). Dans ce dernier cas, lorsque la variable a été trouvée discriminante, le nom de la variable la plus discriminante qui en est issue est indiquée et ses statistiques sont montrées.

Les variables sont caractérisées par des statistiques de base (type, fréquence, chi2, part de variance expliquée). En cliquant sur une variable, l'utilisateur fait apparaître une seconde fenêtre donnant des statistiques complémentaires (moyenne, min, max, table de contingence, coefficient de corrélation, …).

L'onglet **Profiling** sur la seconde fenêtre permet de croiser la variable sélectionnée avec l'ensemble des variables de l'analyse.

Liste : **Variables de base**

Affiche la liste des variables de base triées par pouvoir explicatif décroissant (part de variance expliquée). Les variables de base correspondent aux variables d'entrée, éclatées pour les variables catégoriques.

Liste : **Variables transformées**

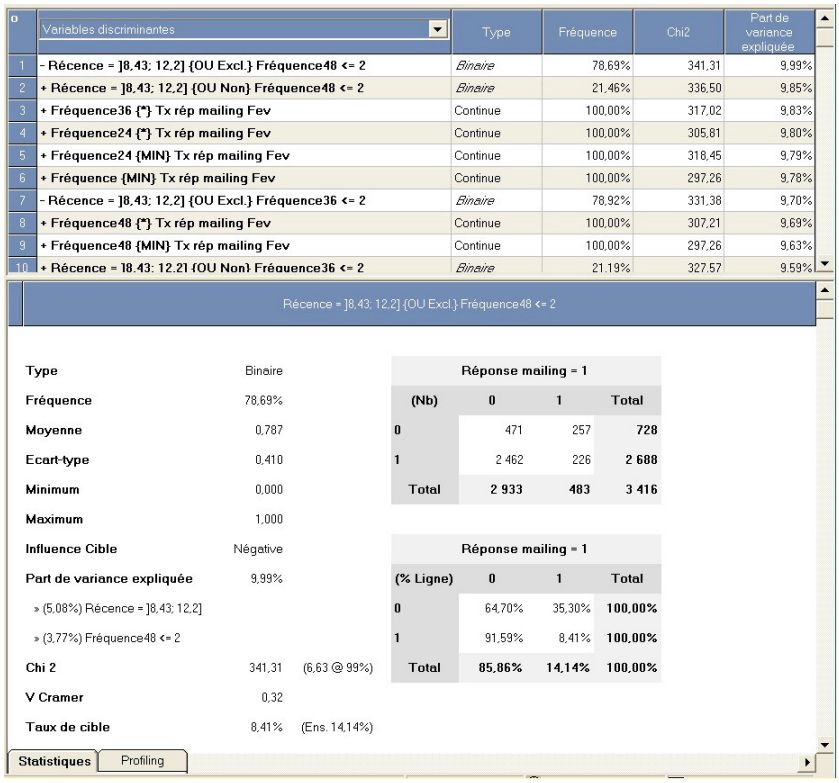
Affiche la liste des variable transformées triées par pouvoir discriminant décroissant (part de variance expliquée). Les variables transformées correspondent aux variables d'entrée transformées individuellement (regroupements de valeurs, valeurs remarquables, …).

Liste : **Variables combinées**

Affiche la liste des variables combinées triées par pouvoir discriminant décroissant (part de variance expliquée). Les variables combinées sont issues de combinaisons de variables 2 à 2, 3 à 3 (optionnel), ou plus (groupes de variables)

Liste : **Variables de base – liste exhaustive**

Affiche la liste de toutes les variables en entrée avec la variable la plus discriminante qui en est issue (transformée ou combinées) ainsi que ses statistiques. Les variables sont triées par pouvoir discriminant décroissant (part de variance expliquée).



	Variables discriminantes	Type	Fréquence	Chi2	Part de variance expliquée
1	- Récence =]8,43; 12,2] {OU Excl.} Fréquence48 <= 2	*Binaire*	78,69%	341,31	9,99%
2	+ Récence =]8,43; 12,2] {OU Non} Fréquence48 <= 2	*Binaire*	21,46%	336,50	9,85%
3	+ Fréquence36 {*} Tx rép mailing Fev	Continue	100,00%	317,02	9,83%
4	+ Fréquence24 {*} Tx rép mailing Fev	Continue	100,00%	305,81	9,80%
5	+ Fréquence24 {MIN} Tx rép mailing Fev	Continue	100,00%	318,45	9,79%
6	+ Fréquence {MIN} Tx rép mailing Fev	Continue	100,00%	297,26	9,78%
7	- Récence =]8,43; 12,2] {OU Excl.} Fréquence36 <= 2	*Binaire*	78,92%	331,38	9,70%
8	+ Fréquence48 {*} Tx rép mailing Fev	Continue	100,00%	307,21	9,69%
9	+ Fréquence48 {MIN} Tx rép mailing Fev	Continue	100,00%	297,26	9,63%
10	+ Récence =]8,43; 12,2] {OU Non} Fréquence36 <= 2	*Binaire*	21,19%	327,57	9,59%

Récence =]8,43; 12,2] {OU Excl.} Fréquence48 <= 2

Type	Binaire	
Fréquence	78,69%	
Moyenne	0,787	
Ecart-type	0,410	
Minimum	0,000	
Maximum	1,000	
Influence Cible	Négative	
Part de variance expliquée	9,99%	
» (5,08%) Récence =]8,43; 12,2]		
» (3,77%) Fréquence48 <= 2		
Chi 2	341,31	(6,63 @ 99%)
V Cramer	0,32	
Taux de cible	8,41%	(Ens. 14,14%)

Réponse mailing = 1

(Nb)	0	1	Total
0	471	257	728
1	2 462	226	2 688
Total	2 933	483	3 416

Réponse mailing = 1

(% Ligne)	0	1	Total
0	64,70%	35,30%	100,00%
1	91,59%	8,41%	100,00%
Total	85,86%	14,14%	100,00%

Statistiques | Profiling

▸Variables discriminantes et statistiques

4.6.2 Regroupement de valeurs

Critères explicatifs > Regroupements de valeurs

Affiche la liste des regroupement optimaux de valeurs obtenus sur chaque variable.

Pour chaque valeur, sont disponibles les informations : nombre, pourcentage, taux ou moyenne de la cible.

Regroupements	Nb	%	% Cible
TOTAL	3 416	100%	14,14%
Ancienneté			Chi2 = 29,02
<= 24,82	778	22,78%	8,23%
> 24,82	2 638	77,22%	15,88%
Récence			Chi2 = 203,59
<= 8,43	1 987	58,17%	13,19%
=]8,43; 12,2]	525	15,37%	32,57%
> 12,2	904	26,46%	5,53%
Fréquence			Chi2 = 129,73
<= 1	1 831	53,60%	9,18%
=]1; 2]	1 382	40,46%	17,51%
> 2	203	5,94%	35,96%
Montant			Chi2 = 19,75
<= 14,42	394	11,53%	8,88%
=]14,42; 15,85]	339	9,92%	20,35%
> 15,85	2 683	78,54%	14,13%

Regroupements optimaux de valeurs

Pour plus de détails, se référer à la rubrique **7. 6 Discrétisation et regroupement de modalités**

4.6.3 Valeurs remarquables

Critères explicatifs > Valeurs remarquables

Affiche la liste des valeurs remarquables statistiquement significatives identifiées pour chaque variable. Pour chaque valeur, sont disponibles les informations : nombre, pourcentage, taux ou moyenne de la cible.

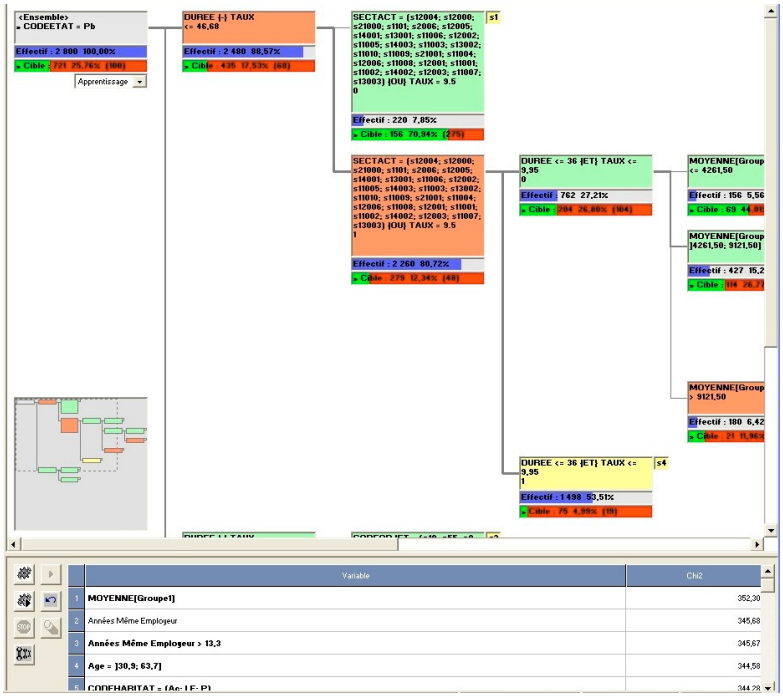
4.6.4 Arbre de décision

Critères explicatifs > Arbre de décision

Affiche une fenêtre permettant de développer un arbre de décision explicatif de la cible choisie. Une imagette de l'arbre permet de se déplacer dans l'arbre en cliquant sur la zone que l'utilisateur veut afficher.

Les nœuds sont représentés en vert (resp. en rouge) lorsque le taux (ou la moyenne) cible est supérieure (resp. inférieure) à celle de l'ensemble de l'échantillon. Pour chaque nœud, on a :

- L'effectif de la branche en nombre et pourcentage
- L'effectif cible de la branche en nombre, pourcentage et indice base 100



▸ Elaboration d'un arbre

A partir d'un nœud sélectionné (coloré en jaune) :

- Le bouton permet d'élaguer le nœud
- Le bouton lance la recherche des variables discriminantes du nœud.
- Le bouton ▸ développe le nœud sélectionné avec la variable sélectionnée dans la liste des variables discriminantes

- Le bouton  permet d'annuler la dernière action
- Le bouton  permet d'interrompre la recherche en cours
- Le bouton  permet d'afficher la fenêtre de paramètre de l'arbre

4.6.5 Segments à potentiels

Critères explicatifs > Segments à potentiel

Un segment est une variable binaire, de base, transformée ou combinée.

Affiche la liste des segments (variables binaires uniquement) à potentiel fort ou faible triés selon leur pouvoir discriminant décroissant et caractérisés par le taux de cible ou la moyenne dans le segment. L'onglet permet d'afficher l'un ou l'autre.

En cliquant sur un segment, l'utilisateur fait apparaître une seconde fenêtre donnant des statistiques complémentaires (moyenne, min, max, table de contingence, coefficient de corrélation, …).

L'onglet 'Profiling' sur la seconde fenêtre permet de croiser la variable sélectionnée avec l'ensemble des variables de l'analyse.

4.6.6 Niche

Critères explicatifs > Niche

Permet de créer une niche constituée de la réunion des segments à partir de segments présentés par pouvoir discriminant décroissant.

4.7 Modèle de score

Cette rubrique regroupe les résultats relatifs à la sélection de variables (modèle) la plus explicative de la cible. Deux modèles sont automatiquement élaborés et montrés : un à partir des variables de base uniquement (Modèle de base), l'autre à partir de l'ensemble de toutes les variables (base, transformées, combinées - Modèle DataLab).

4.7.1 Modèles

Modèle de score > Modèles

Affiche la liste des variables sélectionnées (Modèles de base et DataLab). Ces deux modèles peuvent donc être directement comparées afin d'évaluer l'apport de DataLab.

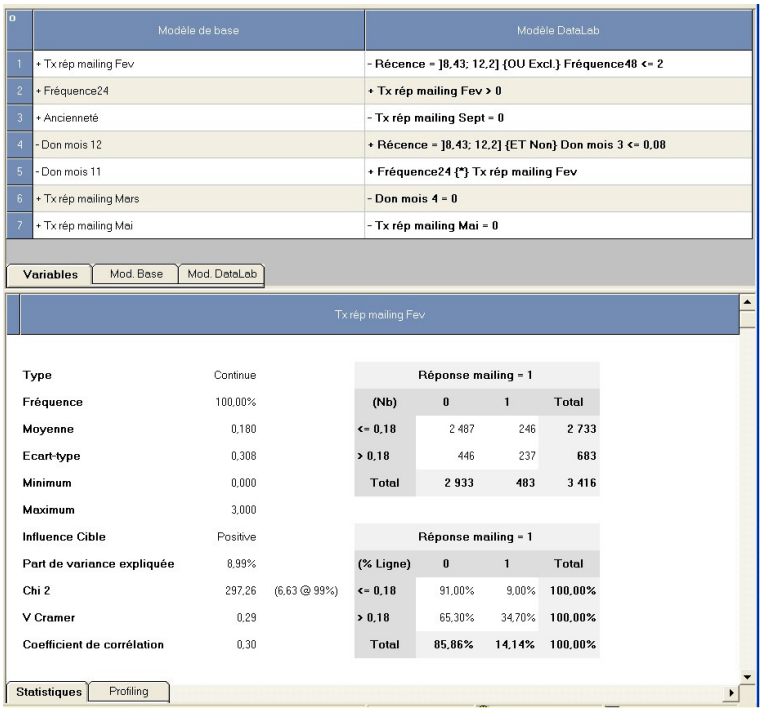
Le signe + devant l'intitulé de la variable indique une contribution positive, le signe – une contribution négative.

Les onglets 'Mod. Base' et 'Mod. DataLab' donnent des informations complémentaires sur, respectivement, les modèles de base et DataLab :

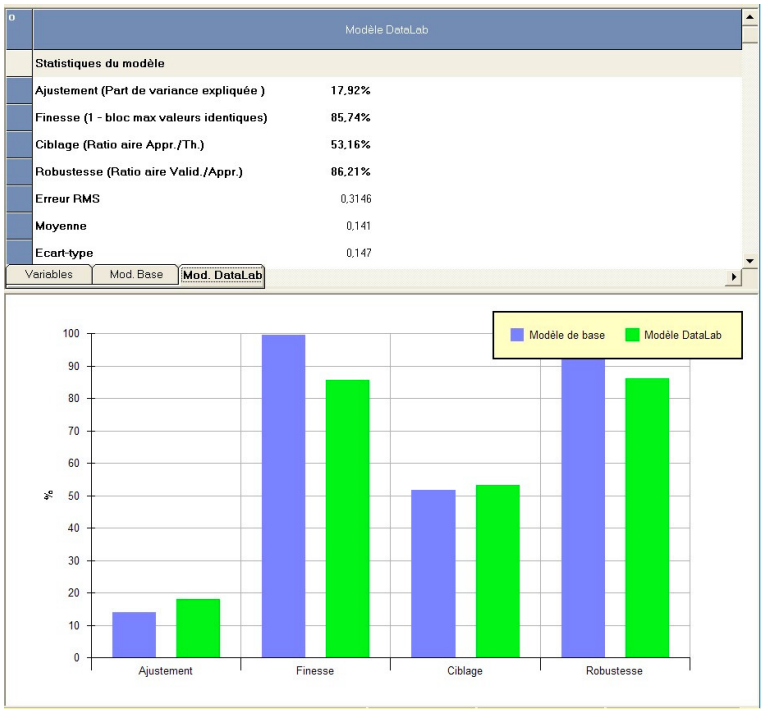
- Statistiques du modèle :
  - Ajustement (part de variance expliquée)
  - Finesse du score (Complément au bloc le plus grand de scores ayant une valeur identique)
  - Ciblage (ratio de l'aire du gain chart définie entre les courbes du modèle et de l'aléatoire, et entre les courbes théorique et de l'aléatoire)
  - Robustesse (ratio de l'aire du gain chart définie entre les courbes du modèle en apprentissage et de l'aléatoire, et entre les courbes du modèle en validation et de l'aléatoire). Cet indicateur n'est disponible que si un échantillon de validation a été défini.
  - erreur RMS
  - moyenne, écart-type

- Statistiques des variables explicatives : moyenne, écart-type
- Paramètres du modèle : coefficients et coefficients avec les variables explicatives centrées normées, VIF
- Corrélations des variables explicatives (V de Cramer pour les variables catégoriques)
- Historique de construction

En cliquant sur une variable, l'utilisateur fait apparaître une seconde fenêtre donnant des statistiques complémentaires (moyenne, min, max, table de contingence, coefficient de corrélation, …).
L'onglet 'Profiling' sur la seconde fenêtre permet de croiser la variable sélectionnée avec l'ensemble des variables de l'analyse.



Modèles de base et DataLab



- Infos sur les modèles

4.7.2 Modifier un modèle

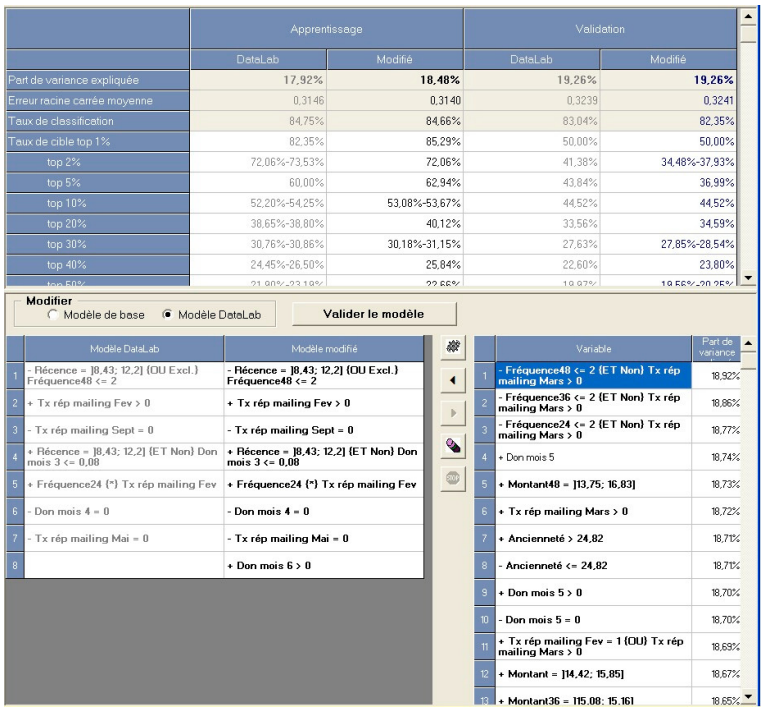
Modèle de score > Modifier un modèle

Affiche une fenêtre permettant de modifier la sélection de variables ainsi qu'une fenêtre de validation des résultats.

Le modèle à modifier peut être sélectionné en cliquant sur le bouton d'option correspondant (Base, DataLab).

Le modèle courant est affiché automatiquement à côté du modèle modifié.

Le bouton permet d'effacer la sélection, le bouton ▸ supprime la variable sélectionnée, le bouton ◂ ajoute la variable sélectionnée, le bouton lance la recherche des variables discriminantes étant donné les variables déjà sélectionnées.



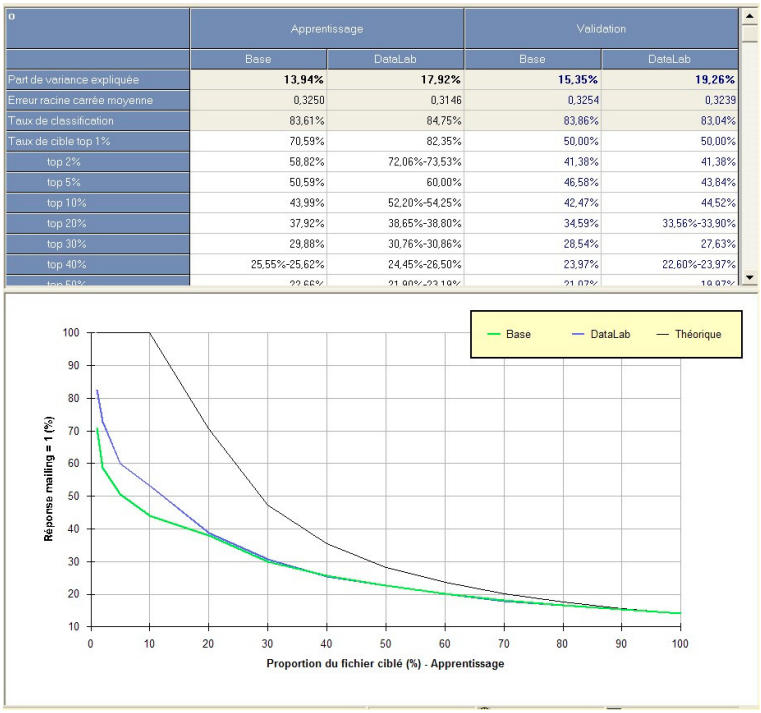
▸ Modification d'un modèle

4.7.3 Comparer les validations

Modèle de score > Comparer les validations

Affiche les statistiques des modèles (Modèle de base et DataLab) selon divers critères de performances. Ces critères sont calculés et présentés séparément sur l'échantillon d'étude défini et sur le reste des enregistrements (hors échantillon) le cas échéant. Les critères mesurant la performance des sélections sont :

- Part de variance expliquée
- Racine de l'erreur carrée moyenne
- Taux de classification (Cible binaire uniquement)
- %cible (Cible binaire) ou moyenne (Cible continue) des top 1, 2, 5, 10, 20, 30, … 90, 100% (fichier trié par prévision décroissante)



	Apprentissage		Validation	
	Base	DataLab	Base	DataLab
Part de variance expliquée	**13,94%**	**17,92%**	**15,35%**	**19,26%**
Erreur racine carrée moyenne	0,3250	0,3146	0,3254	0,3239
Taux de classification	83,61%	84,75%	83,86%	83,04%
Taux de cible top 1%	70,59%	82,35%	50,00%	50,00%
top 2%	58,82%	72,06%-73,53%	41,38%	41,38%
top 5%	50,59%	60,00%	46,58%	43,84%
top 10%	43,99%	52,20%-54,25%	42,47%	44,52%
top 20%	37,92%	38,65%-38,80%	34,59%	33,56%-33,90%
top 30%	29,88%	30,76%-30,86%	28,54%	27,63%
top 40%	25,55%-25,62%	24,45%-26,50%	23,97%	22,60%-23,97%

Statistiques de validation des modèles

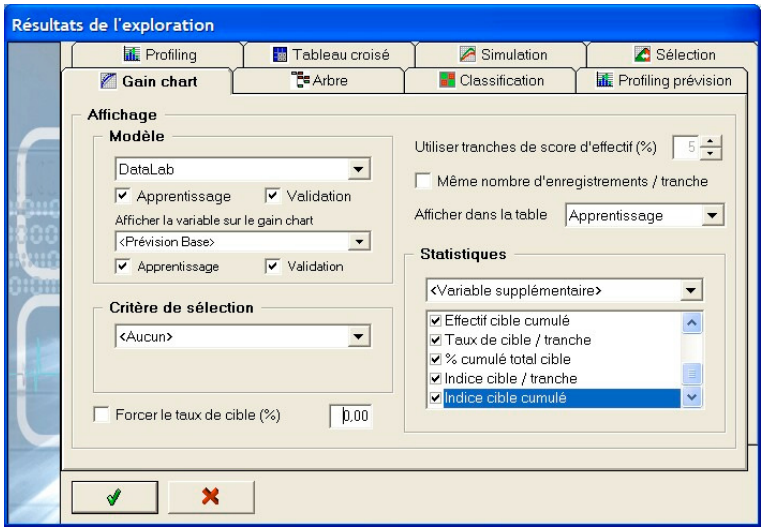
En cliquant sur la zone de résultats, l'utilisateur fait apparaître une seconde fenêtre affichant un graphe présentant le gain chart obtenu sur l'échantillon ou le hors échantillon.

4.7.4 Gain chart

Modèle de score > Gain chart

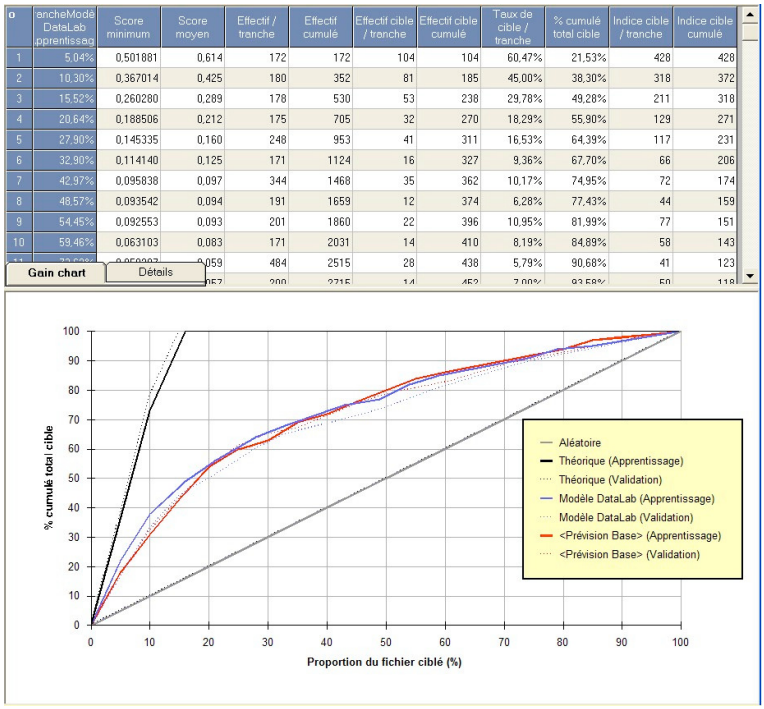
Le gain chart donne pour chaque tranche (dont la taille est paramétrable entre 1% et 20%) le nombre d'enregistrements le nombre de critères cible et différentes statistiques associées.

L'utilisateur peut choisir la population sur laquelle le gain chart est montré (apprentissage, validation ou données externes, critère, taux de cible). Les statistiques d'une variable supplémentaire (cumul, moyenne) peuvent aussi être affichées.



Paramétrage du gain chart

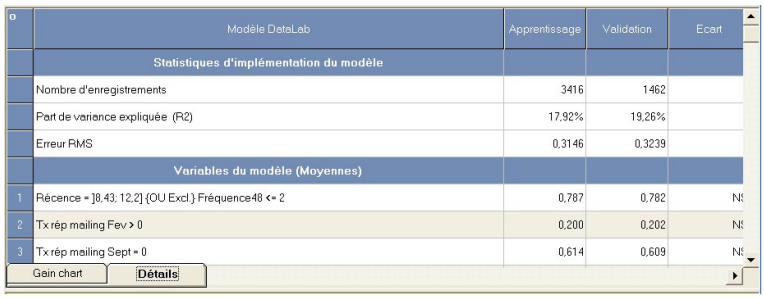
Les résultats sont regroupés sous deux onglets. L'onglet 'Gain chart' présente la performance du ciblage . L'onglet 'Détails' montre une comparaison entre deux échantillons, selon les statistiques du modèle d'une part, selon les moyennes des variables explicatives du modèle sur chaque échantillon d'autre part.



	ancheModè DataLab pprentissag	Score minimum	Score moyen	Effectif / tranche	Effectif cumulé	Effectif cible / tranche	Effectif cible cumulé	Taux de cible / tranche	% cumulé total cible	Indice cible / tranche	Indice cible cumulé
1	5,04%	0,501881	0,614	172	172	104	104	60,47%	21,53%	428	428
2	10,30%	0,367014	0,425	180	352	81	185	45,00%	38,30%	318	372
3	15,52%	0,260280	0,289	178	530	53	238	29,78%	49,28%	211	318
4	20,64%	0,188506	0,212	175	705	32	270	18,29%	55,90%	129	271
5	27,90%	0,145335	0,160	248	953	41	311	16,53%	64,39%	117	231
6	32,90%	0,114140	0,125	171	1124	16	327	9,36%	67,70%	66	206
7	42,97%	0,095838	0,097	344	1468	35	362	10,17%	74,95%	72	174
8	48,57%	0,093542	0,094	191	1659	12	374	6,28%	77,43%	44	159
9	54,45%	0,092553	0,093	201	1860	22	396	10,95%	81,99%	77	151
10	59,46%	0,063103	0,083	171	2031	14	410	8,19%	84,89%	58	143
			0,059	484	2515	28	438	5,79%	90,68%	41	123

Gain chart | Détails

▸ Gain chart



	Modèle DataLab	Apprentissage	Validation	Ecart
	Statistiques d'implémentation du modèle			
	Nombre d'enregistrements	3416	1462	
	Part de variance expliquée (R2)	17,92%	19,26%	
	Erreur RMS	0,3146	0,3239	
	Variables du modèle (Moyennes)			
1	Récence =]8,43; 12,2] {OU Excl.} Fréquence48 <= 2	0,787	0,782	NS
2	Tx rép mailing Fev > 0	0,200	0,202	NS
3	Tx rép mailing Sept = 0	0,614	0,609	NS

Gain chart | Détails

▸ Gain chart (détails)

Pour plus de détails, se référer à la rubrique **7. 1 Le gain chart**

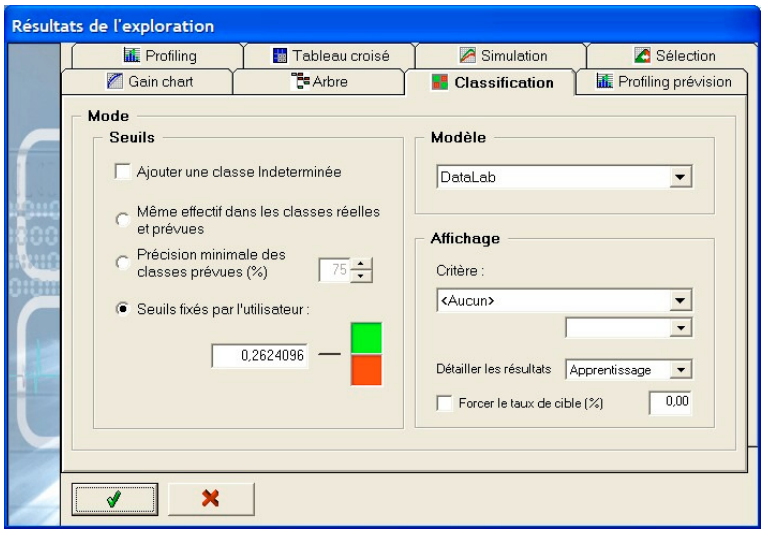
4.7.5 Matrice de classification

Modèle de score > Matrice de classification

Le module de classification permet de déterminer, dans le cas où la variable cible est binaire) le ou les seuils de score transformant le score en classe. 3 modes de classification sont disponibles :

- Automatique avec des effectifs de classe identiques entre le prévu et le réel
- Automatique avec un minimum de précision par classe
- Défini par l'utilisateur

Les statistiques de base qui sont fournies concernent l'effectif en nombre et pourcentage, et le taux de classification.
L'utilisateur peut choisir la population sur laquelle le gain chart est montré (apprentissage ou validation, critère, taux de cible).



Paramétrage de la classification

Cible prévue - Modèle DataLab	Cible observée (Apprentissage) 0	1	TOTAL	TOTAL (%)	Taux de bonnes prévisions
0 (<=0,2624096)	2 669	257	2 926	85,66%	91,22%
1 (>0,2624096)	264	226	490	14,34%	46,12%
TOTAL	2 933	483	3 416	100,00%	84,75%
TOTAL (%)	85,86%	14,14%	100,00%		

Matrice de classification

Pour plus de détails, se référer à la rubrique **7. 4 La classification.**

4.7.6 Profiling du score

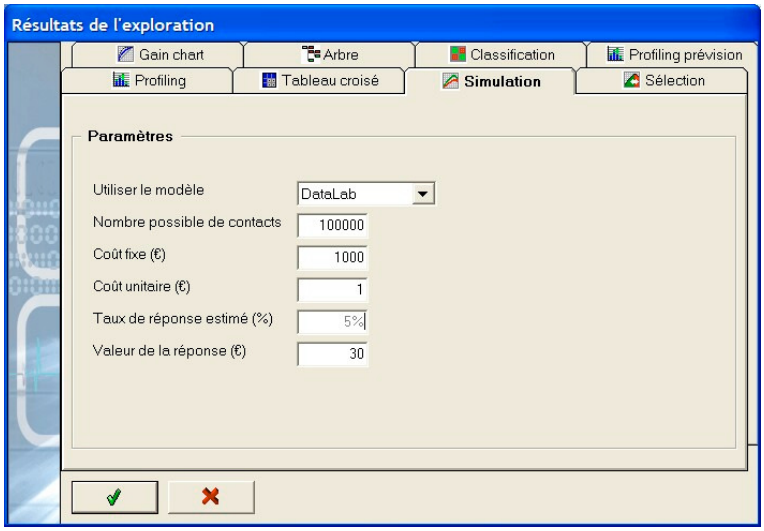
Modèle de score > Profiling du score

Affiche le profiling (croisement) de l'ensemble des variables avec un segment représentant le top score. Pour chaque croisement, le tableau donne l'effectif, le pourcentage et l'indice de sur-representativité. Les sur-représentations sont indiquées sous la forme d'indice base 100. Les indices en vert indiquent un pourcentage significativement plus important que la moyenne, en rouge un pourcentage significativement plus faible que la moyenne, en noir une différence non significative.

4.7.7 Simulation

Modèle de score > Simulation

Propose une simulation marketing directe reposant sur un des modèles. Dans la fenêtre de paramétrage entrer les valeurs de la simulation puis valider.



Paramétrage de la simulation

Les résultats affichés présentent, par tranche de score (identiques à celle définies pour le gain chart), le cumul attendu des réponses, du coût et de la marge nette (valeur – coût). Le maximum de la marge définie ainsi le seuil de score optimum.

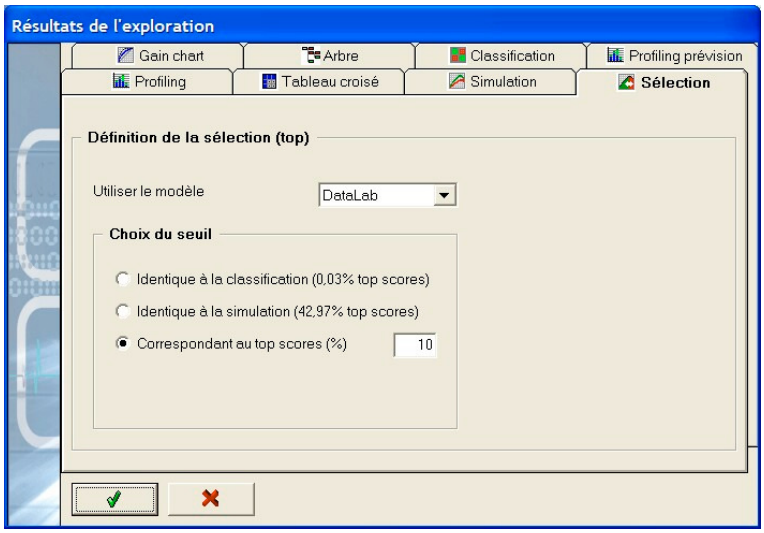
	% Contacts	Contacts	Cumul cible	Taux de cible	Coût	Valeur	Valeur nette
1	5,04%	5035	1077	21,38%	6035	32298	26263
2	10,30%	10304	1915	18,59%	11304	57453	46149
3	15,52%	15515	2464	15,88%	16515	73913	57398
4	20,64%	20638	2795	13,54%	21638	83851	62213
5	27,90%	27898	3219	11,54%	28898	96584	67686
6	32,90%	32904	3385	10,29%	33904	101553	67649
7	42,97%	42974	3747	8,72%	43974	112422	68448
8	48,57%	48566	3872	7,97%	49566	116149	66583
9	54,45%	54450	4099	7,53%	55450	122981	67532
10	59,46%	59456	4244	7,14%	60456	127329	66874
11	73,62%	73624	4534	6,16%	74624	136025	61401

▸ Simulation

4.7.8 Définir une sélection

Modèle de score > Définir une sélection

Affiche une fenêtre permettant de définir la valeur du top score à utiliser pour l'exportation.



▸ Sélection des top scores

5. Mise en œuvre de DataLab

5.1 Exemple Crédit

Le fichier 'Exemple_Crédit.txt' est un fichier réel constitué à partir d'un échantillon de clients ayant souscrits à un prêt. Il contient 4 000 enregistrements (clients) avec un historique de remboursement et un ensemble d'informations descriptives du client et du prêt.
Les variables sont :

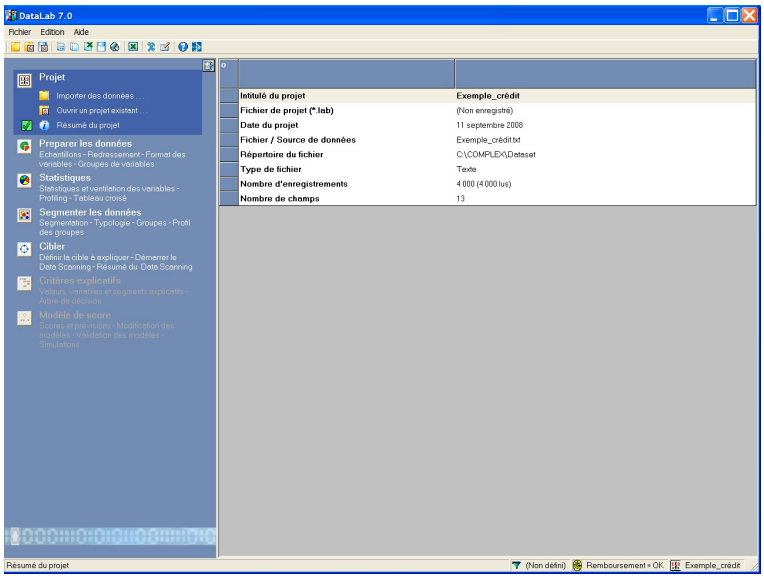
- NUMCLIENT : identifiant client
- Remboursement : issue du remboursement
- CIVILITE : civilité
- Age : âge
- SITUAFAMIL : situation familiale
- NBPACHARGE :nombre de personnes à charge
- Années Même Employeur : nombre d'années passées chez le même employeur
- CODEHABITAT : code habitat
- MONTANT : montant du prêt
- DUREE : durée du prêt
- MENSUALITE : mensualité du prêt
- Charges : charges totales du client
- Revenus : revenus totaux du client

Le principe du test est de prévoir le mieux possible l'issue du prêt avec les informations descriptives du client et du prêt.
La mise en œuvre de DataLab sur ce fichier est détaillée ci-dessous.

1. Importer le fichier

- Sélectionner le menu **Fichier / Importer un fichier …**
- Cliquer sur la source de données **Texte** puis sur le bouton **Suivant…**
- Sélectionner le fichier **Exemple_Crédit.txt**
- Cliquer sur le bouton **Importer**

Une barre de progression montre l'avancement de l'importation à l'issue de laquelle une fenêtre de résumé est affichée.

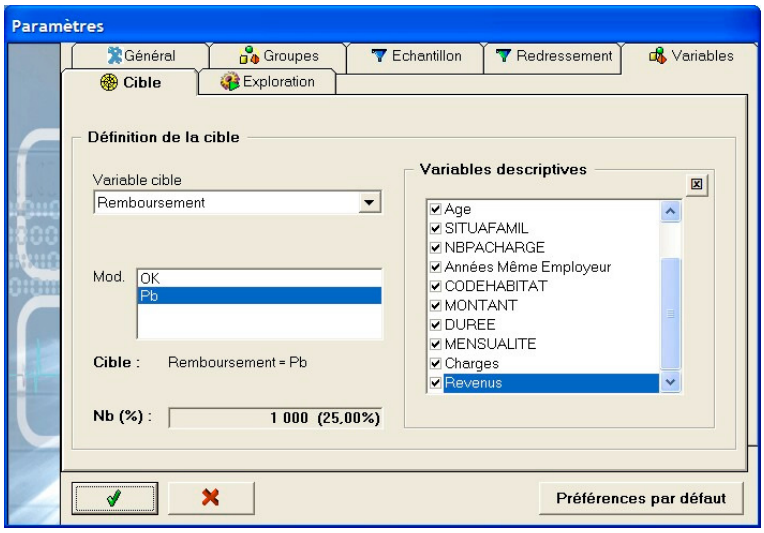


▸ Résumé du projet.

2. Définir la cible

A la fin de l'importation, une fenêtre synthétisant le projet s'affiche.

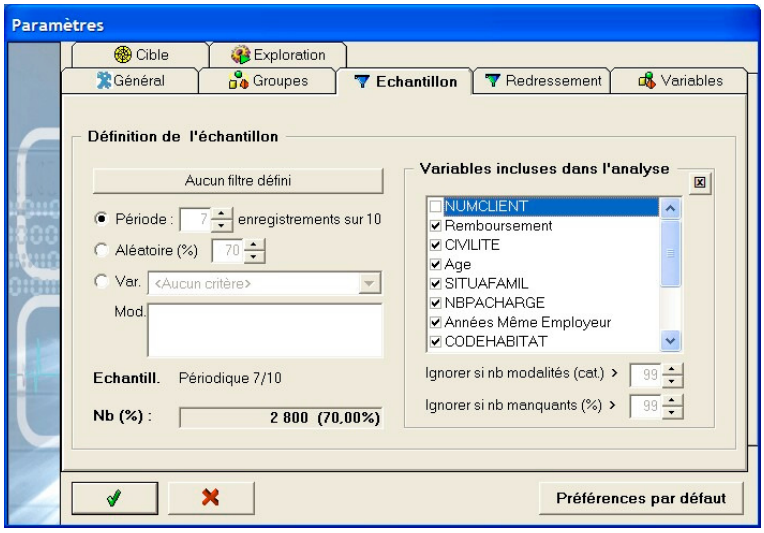
- Sélectionner **Cibler > Définir la cible …** La fenêtre de paramètres s'affiche (onglet **Cible**)
- Sélectionner R**emboursement** dans la liste déroulante **Variable**
- Cliquer sur la modalité **Pb**. La cible est **Remboursement=Pb**. Cette variable est automatiquement désélectionnée de la liste des variables explicatives
- Fermer la fenêtre en cliquant sur **OK**



Définition de la cible.

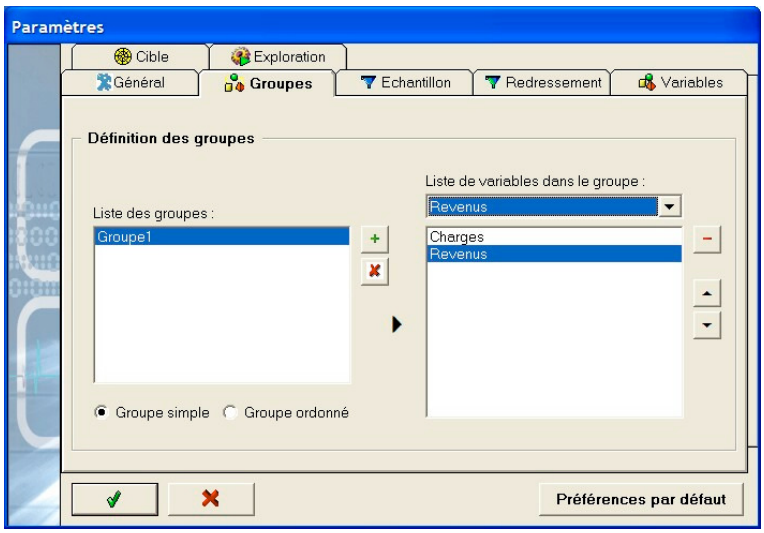
3. Définir un échantillon

- Sélectionner **Préparer les données > Définir un échantillon ….** La fenêtre de paramètres s'affiche (onglet **Echantillon**)
- Cocher **Période** et ajuster 7 enregistrements sur 10. L'échantillon contient 70% des enregistrements du fichier total, soit 2 800.
- Décocher la case **NUMCLIENT** dans la liste **Variables inclues dans l'analyse** puisque ce champ est un identifiant. (en principe, ce champ est reconnu comme tel car il ne comporte aucune valeur doublon et est décoché automatiquement). Fermer la fenêtre



Définition de l'échantillon.

4. Définir des groupes de variables



Définition des groupes de variables.

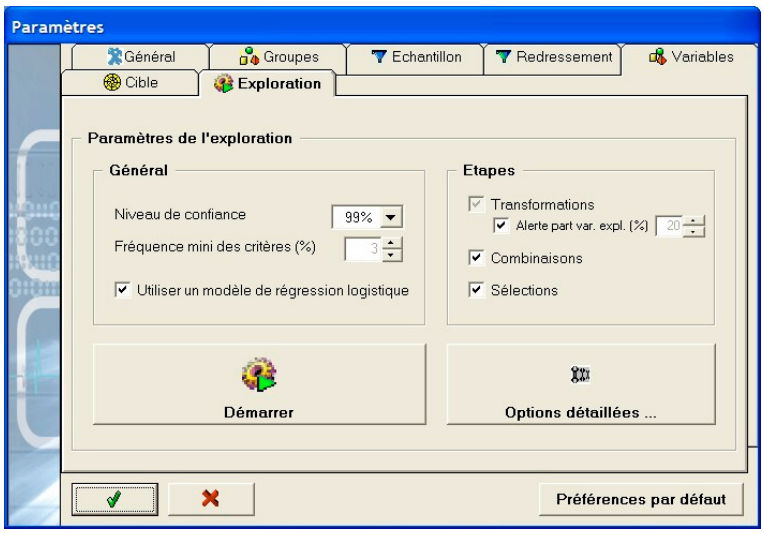
- Sélectionner **Préparer les données > Groupes de variables ….** La fenêtre de paramètres s'affiche (onglet **Groupes**)
- Cliquer sur le bouton **+** et entrer **ChargesRevenus** pour l'intitulé du nouveau groupe
- Cliquer sur le nom du groupe **ChargesRevenus** dans la liste **Liste des groupes**
- Sélectionner successivement les variables **Charges**, **Revenus**

- Cocher le type de groupe **Simple** puisque les variables constitutives sont non ordonnées en temps
- Fermer la fenêtre

5. Démarrer l'exploration

- Sélectionner **Cibler > Démarrer le Data Scanning ….** Une fenêtre de paramètres s'affiche (onglet **Exploration**)
- Lancer l'exploration en cliquant sur le bouton **Explorer** symbolisé par l'icône de DataLab.

La fenêtre d'exploration s'affiche et la barre de progression montre l'avancement de l'exploration. Lorsque l'exploration est terminée une fenêtre est affichée automatiquement synthétisant les résultats obtenus.



Démarrer l'exploration.

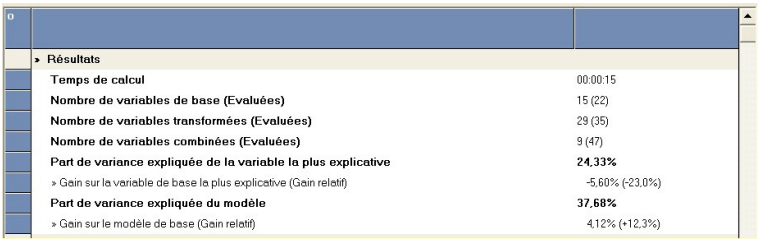
Il est préférable à ce stade d'enregistrer le projet au moyen du menu **Fichier / Enregistrer le projet …**, par exemple sous le nom **Exemple_crédit.lab**.

6. Interpréter les résultats

Les résultats sont détaillés selon deux axes : les nouvelles variables prises individuellement, et la sélection de variables expliquant au mieux la cible choisie.

Sélectionner **Cibler > Résumé du Data Scanning**. Le résumé détaille les principaux paramètres utilisés pour l'exploration ainsi que quelques résultats principaux :

- Temps de calcul : 0'15".
- 15 variables de base significatives sur 22 évaluées.
- 29 variables transformées, sur 35 évaluées.
- 47 variables combinées ont été créées et évaluées et 9 ont été retenues car apportant, individuellement, un gain significatif.
- La variable la plus discriminante apporte un pouvoir explicatif correspondant à une part de variance expliquée de 24.33 % soit un accroissement de –5.60% en valeur absolue (et de –23.0% en valeur relative) par rapport à la variable de base la plus discriminante.
- Par rapport à l'utilisation des seules variables de base, les variables créés et sélectionnés par DataLab (Modèle DataLab) conduisent à une part de variance expliquée de 37.68% soir un accroissement en valeur absolue de 4.12% (et +12.3% en valeur relative).



Résultats	
Temps de calcul	00:00:15
Nombre de variables de base (Evaluées)	15 (22)
Nombre de variables transformées (Evaluées)	29 (35)
Nombre de variables combinées (Evaluées)	9 (47)
Part de variance expliquée de la variable la plus explicative	24,33%
» Gain sur la variable de base la plus explicative (Gain relatif)	-5,60% (-23,0%)
Part de variance expliquée du modèle	37,68%
» Gain sur le modèle de base (Gain relatif)	4,12% (+12,3%)

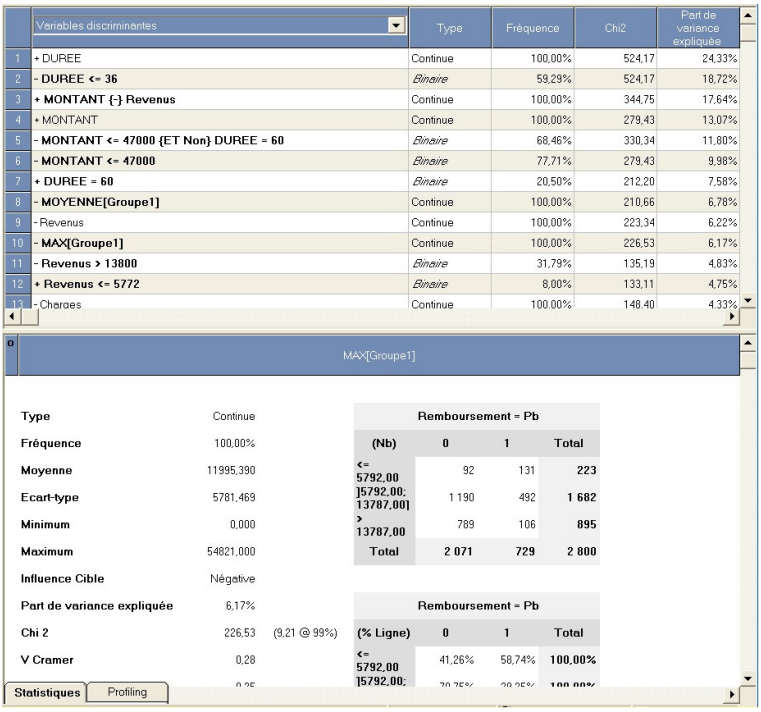
Résumé de l'exploration.

Sélectionner **Critères explicatifs > Variables explicatives**. La fenêtre affiche la liste des variables discriminantes triés par pouvoir discriminant décroissant avec :

- Le nom de la variable
- Le type de variable
- La fréquence d'occurrence de la variable (100% si la variable n'est pas binaire)

- La valeur du chi2 avec la variable cible
- La part de variance expliquée apportée par la variable pour expliquer la cible

Pour la variable **Durée** la part de variance expliquée est égale à 24.33% avec un chi2 égal à 524.17 (seuil de significativité : 6.63 à 99% de confiance).



	Variables discriminantes	Type	Fréquence	Chi2	Part de variance expliquée
1	+ DUREE	Continue	100,00%	524,17	24,33%
2	- DUREE <= 36	*Binaire*	59,29%	524,17	18,72%
3	+ MONTANT {-} Revenus	Continue	100,00%	344,75	17,64%
4	+ MONTANT	Continue	100,00%	279,43	13,07%
5	- MONTANT <= 47000 {ET Non} DUREE = 60	*Binaire*	68,46%	330,34	11,80%
6	- MONTANT <= 47000	*Binaire*	77,71%	279,43	9,98%
7	+ DUREE = 60	*Binaire*	20,50%	212,20	7,58%
8	- MOYENNE[Groupe1]	Continue	100,00%	210,66	6,78%
9	- Revenus	Continue	100,00%	223,34	6,22%
10	- MAX[Groupe1]	Continue	100,00%	226,53	6,17%
11	- Revenus > 13800	*Binaire*	31,79%	135,19	4,83%
12	+ Revenus <= 5772	*Binaire*	8,00%	133,11	4,75%
13	- Charges	Continue	100,00%	148,40	4,33%

MAX[Groupe1]

Type	Continue		
Fréquence	100,00%		
Moyenne	11995,390		
Ecart-type	5781,469		
Minimum	0,000		
Maximum	54821,000		
Influence Cible	Négative		
Part de variance expliquée	6,17%		
Chi 2	226,53	(9,21 @ 99%)	
V Cramer	0,28		

Remboursement = Pb

(Nb)	0	1	Total
<= 5792,00	92	131	223
]5792,00; 13787,00]	1 190	492	1 682
> 13787,00	789	106	895
Total	2 071	729	2 800

Remboursement = Pb

(% Ligne)	0	1	Total
<= 5792,00	41,26%	58,74%	100,00%
]5792,00;			

Statistiques | Profiling

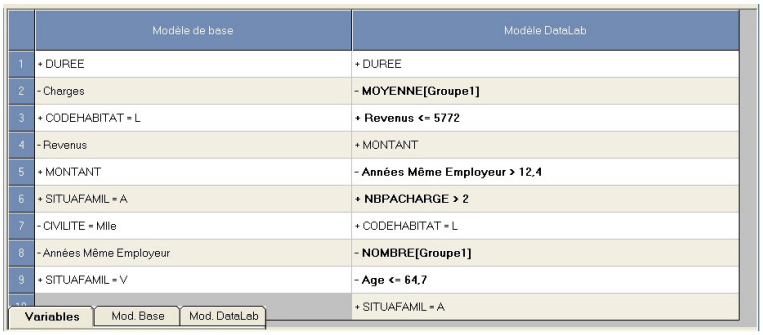
```
Variables discriminantes.
```

Sélectionner **Modèles de score > Modèles**. La fenêtre affiche pour les 2 modèles (modèles de base et modèle DataLab), la liste des variables sélectionnées.

Le modèle de base conduit à 9 variables :

- **DUREE**
- **Charges**
- **CODEHABITAT = L**
- **Années même employeur = n.r.**, ….

Le modèle DataLab donne 10 variables dont la première est aussi **DUREE**, la deuxième MOYENNE[Groupe1].



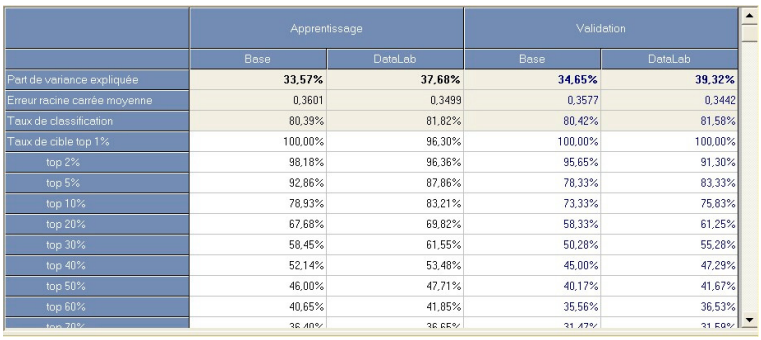
	Modèle de base	Modèle DataLab
1	+ DUREE	+ DUREE
2	- Charges	- **MOYENNE[Groupe1]**
3	+ CODEHABITAT = L	+ **Revenus <= 5772**
4	- Revenus	+ MONTANT
5	+ MONTANT	- **Années Même Employeur > 12,4**
6	+ SITUAFAMIL = A	+ **NBPACHARGE > 2**
7	- CIVILITE = Mlle	+ CODEHABITAT = L
8	- Années Même Employeur	- **NOMBRE[Groupe1]**
9	+ SITUAFAMIL = V	- **Age <= 64,7**
10		+ SITUAFAMIL = A

Variables | Mod. Base | Mod. DataLab

Variables sélectionnées (modèles).

Sélectionner **Modèles de score > Comparer les validations**. La fenêtre affiche, pour la sélection de base et pour la sélection DataLab, différents critères statistiques permettant de comparer les deux sélections et donc de mesurer l'apport de DataLab. Ces statistiques sont disponibles sur l'échantillon défini et le complément (Hors échantillon) le cas échéant (ici, 30%).

- Sur l'échantillon, la part de variance expliquée avec les variables de base uniquement est égale à 33.57% contre 37.68% avec toutes les variables possibles. Sur les enregistrements hors échantillon (validation) ces valeurs sont respectivement égales à 34.65 % et 39.32%. Dans les deux cas, le gain apporté par les nouvelles variables est donc supérieur à 10%.
- En termes de moyenne de cible dans le top 10%, on a respectivement pour l'échantillon 78.93% d'enregistrements cible avec la sélection de base et 83.21% avec le modèle DataLab.



	Apprentissage		Validation	
	Base	DataLab	Base	DataLab
Part de variance expliquée	**33,57%**	**37,68%**	**34,65%**	**39,32%**
Erreur racine carrée moyenne	0,3601	0,3499	0,3577	0,3442
Taux de classification	80,39%	81,82%	80,42%	81,58%
Taux de cible top 1%	100,00%	96,30%	100,00%	100,00%
top 2%	98,18%	96,36%	95,65%	91,30%
top 5%	92,86%	87,86%	78,33%	83,33%
top 10%	78,93%	83,21%	73,33%	75,83%
top 20%	67,68%	69,82%	58,33%	61,25%
top 30%	58,45%	61,55%	50,28%	55,28%
top 40%	52,14%	53,48%	45,00%	47,29%
top 50%	46,00%	47,71%	40,17%	41,67%
top 60%	40,65%	41,85%	35,56%	36,53%
top 70%	36,49%	36,65%	31,47%	31,59%

Validation des modèles

Sélectionner **Critères explicatifs > Regroupements de valeurs**. La fenêtre affiche, pour chaque variable, les regroupements des valeurs réalisés par DataLab caractérisés par :

- Le nombre d'enregistrements concernés
- Le pourcentage d'enregistrements concernés
- La valeur moyenne de la cible (si toutes les valeurs de la cible sont considérées) ou le taux de cible (si la cible est binaire)

	Regroupements	Nb	%	% Cible
	TOTAL	2 800	100%	26,04%
	CIVILITE			
	Age			Chi2 = 20,56
	<= 64,7	2 670	95,36%	25,21%
	> 64,7	130	4,64%	43,08%
	SITUAFAMIL			Chi2 = 60,53
	= VM	84	3,00%	0,00%
	= (M; C; D)	2 455	87,68%	25,34%
	= (S; V; A)	261	9,32%	41,00%
	NBPACHARGE			Chi2 = 21,01
	<= 1	1 782	63,64%	26,26%
	=]1; 2]	616	22,00%	20,62%
	> 2	400	14,29%	33,50%
	Années Même Employeur			Chi2 = 40,50
	<= 12,4	529	18,89%	28,73%
	> 12,4	1 544	55,14%	19,11%

Regroupement des valeurs.

Pour la variable **MONTANT**, le regroupement optimal des valeurs conduit ici à 2 classes : ≤ 47000 (77.71% des enregistrements) et > 47000 (22.29% du total échantillon). Dans la seconde tranche, le taux d'enregistrements cible est égal à 51.92% (contre 26.04% en moyenne). Cette discrétisation optimale de la variable **MONTANT** est associée à une valeur de chi2 de 279.

La discrétisation optimale peut être modifiée (par exemple, pour arrondir le découpage à 50000) au moyen du menu **Préparer les données > Format des variables**, puis en sélectionnant la variable **MONTANT** et en cochant **Afficher le regroupement optimal**.

Sélectionner **Critères explicatifs > Valeurs remarquables**. La fenêtre affiche la liste, pour chaque variable, des valeurs à la fois fréquentes (supérieures au seuil de fréquence fixé) et discriminantes. Ces valeurs sont caractérisés par :

- Le nombre d'enregistrements concernés
- Le pourcentage d'enregistrements concernés

- La valeur moyenne de la cible (si toutes les valeurs de la cible sont considérées) ou le taux de cible (si la cible est binaire)

Pour la variable **MONTANT**, la valeur 10000 montre un taux d'enregistrements cible égal à 8.30% alors qu'il est égal de 26.04% dans l'ensemble de l 'échantillon.

Sélectionner **Critères explicatifs > Segments.**. La fenêtre affiche la liste des segments (i.e. variables calculées binaires) triés par pouvoir discriminant décroissant avec :

- La fréquence d'occurrence du segment
- La valeur moyenne de la cible (si toutes les valeurs de la cible sont considérées) ou le taux de cible (si la cible est binaire)
- L'indice base 100, la valeur 100 correspondant à la valeur moyenne de la cible

Le segment le plus discriminant est le segment 'Revenus <= 5772'. Dans ce segment, qui totalise 8.0% des enregistrements, le taux d'enregistrements cible est égal à 58.48% soit 2.25 fois plus que sur l'ensemble de l'échantillon (indice 225).

6. Support

Pour toute question, merci de contacter le support technique de COMPLEX SYSTEMS par l'un des moyens suivants :

Email : support@complex-systems.fr
Tel : 01 42 21 40 80
Fax : 01 42 21 40 79

7. Focus

7. 1 Le Gain chart

1. Objet

Le gain chart permet d'évaluer les performances d'un score ou d'un modèle de prévision.

Il détaille par tranche, sur le fichier trié selon le modèle, le nombre d'enregistrements, le nombre de critères cible et différentes statistiques associées.

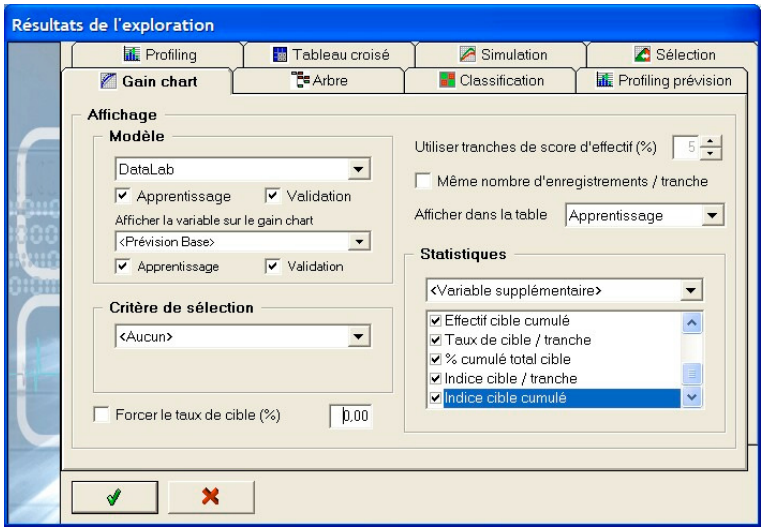
L'utilisateur peut choisir la population sur laquelle le gain chart est calculé.

Cette fiche technique détaille le mode de calcul du gain chart et sa représentation graphique.

2. Détail

2.1 Fenêtre de paramétrage du gain chart

Le gain chart est disponible après validation d'une feuille de paramètres.



a) Calcul du gain chart

- Choix des tranches d'effectifs :
  - entre 1% et 20%
  - possibilité de choisir des tranches comportant exactement le même nombre d'enregistrements. Si cette option n'est pas cochée, la taille des tranches peut être supérieure à la taille paramétrée, du fait de l'existence éventuelle de plages de scores identiques
- Au choix sur :
  - le modèle de base ou le modèle DataLab [zone Modèle de la fenêtre de paramétrage]
  - l'échantillon d'apprentissage, de validation ou des données externes. Dans ce dernier cas, une fenêtre permettant de choisir les données externes est affiché après validation de la fenêtre du gain chart.
- Sur une sous population, par application d'un filtre supplémentaire [zone Critère]:
  - variable catégorique : par exemple « CIVILITE = Mme »
  - variable numérique : par exemple « Age > 40 »

- En redressant la variable cible :
  - le taux de la variable cible peut être modifié par l'utilisateur ; ce paramétrage est particulièrement utile lorsque le modèle est construit sur un échantillon comportant un taux de cible très différent de celui de la réalité (exemple 50/ 50)
- Calcul de statistiques complémentaires [zone Statistiques]:
  - Portant sur une variable catégorique (choix d'une modalité) ou numérique
  - Somme par tranche, somme cumulée, moyenne par tranche

b) Représentation des résultats [zone Modèle]

- Le gain chart peut être représenté graphiquement sur les échantillons d'apprentissage et / ou de validation
- Une variable supplémentaire (choisie dans la liste déroulante : variable de base numérique, variable cible, modèle de base ou modèle DataLab) peut figurer sur la représentation graphique

2.2. Lecture des résultats

Onglet Gain chart

Les résultats sont présentés à l'écran dans 2 zones :

- Gain chart
- Graphique

a) Gain chart

Les résultats sont présentés sur l'échantillon trié par score décroissant :

- sur le modèle choisi
- sur la population paramétrée : échantillon d'apprentissage ou de validation éventuellement croisé par un critère supplémentaire
- avec un taux de cible éventuellement redressé

o	Modèle DataLab pprentissag	Score minimum	Score moyen	Effectif / tranche	Effectif cumulé	Effectif ible/tranch	Effectif cible cumulé	Taux de ible/tranch	% cumulé total cible	Indice cible / tranche	Indice cible cumulé	CIVILITE = Mlle (Somme /	CIVILITE = Mlle (Somme	CIVILITE = Mlle (Moyenne /
1	5,00%	0,663	0,827	140	140	70	70	50,26%	50,29%	1005	1005	5,600	5,600	0,040
2	10,02%	0,477	0,553	140	281	25	96	17,94%	68,29%	359	682	9,667	15,267	0,069
3	15,05%	0,345	0,401	141	421	12	107	8,24%	76,57%	165	509	12,600	27,867	0,089
4	20,07%	0,263	0,298	141	562	8	115	5,41%	82,00%	108	409	14,333	42,200	0,102
5	25,07%	0,198	0,230	140	702	6	121	4,14%	86,14%	83	344	12,667	54,867	0,090
6	30,08%	0,149	0,172	140	842	5	125	3,42%	89,57%	68	298	2,533	57,400	0,018
7	35,09%	0,114	0,130	140	983	2	128	1,57%	91,14%	31	260	14,533	71,933	0,104
8	40,10%	0,090	0,101	140	1123	2	130	1,43%	92,57%	29	231	16,467	88,400	0,118
9	45,14%	0,071	0,080	141	1264	2	131	1,28%	93,86%	26	208	7,600	96,000	0,054
10	50,18%	0,055	0,063	141	1405	2	133	1,28%	95,14%	26	190	8,867	104,867	0,063
11	55,20%	0,045	0,050	141	1546	1	135	0,99%	96,14%	20	174	19,000	123,867	0,135
12	60,22%	0,037	0,041	141	1686	1	136	0,85%	97,00%	17	161	6,333	130,200	0,045
13	65,24%	0,030	0,033	141	1827	1	137	0,85%	97,86%	17	150	11,400	141,600	0,081
14	70,26%	0,025	0,028	141	1967	1	138	0,85%	98,71%	17	140	7,600	149,200	0,054
15	75,30%	0,020	0,022	141	2108	1	139	0,42%	99,14%	8	132	8,867	158,067	0,063
16	80,35%	0,016	0,018	141	2250	1	139	0,42%	99,57%	8	124	7,600	165,667	0,054
17	85,38%	0,011	0,013	141	2391	0	140	0,28%	99,86%	6	117	10,133	175,800	0,072
18	90,41%	0,009	0,010	141	2531	0	140	0,14%	100,00%	3	111	6,333	182,133	0,045
19	95,43%	0,006	0,007	141	2672	0	140	0,00%	100,00%	0	105	8,867	191,000	0,063
20	100,00%	0,002	0,005	128	2800	0	140	0,00%	100,00%	0	100	1,267	192,267	0,010

La table présente les colonnes suivantes :

- effectif en %
- score minimum : valeur minimale du score de la tranche
- score moyen : valeur moyenne du score sur la tranche
- effectif / tranche : nombre d'enregistrements par tranche
- effectif cumulé : nombre d'enregistrements cumulé
- effectif cible/ tranche : nombre d'enregistrements cible par tranche
- effectif cible cumulé : nombre d'enregistrements cible cumulé
- taux de cible / tranche : taux d'enregistrements cible dans la tranche
- % cumulé total cible : proportion cumulée d'enregistrements cible/ total des cibles
- indice cible / tranche : indice du taux de cible dans la tranche. Cet indice est calculé - en base 100 - comme le ratio du taux de cible dans la tranche et du taux de cible global
- indice cible cumulé : indice cumulé des enregistrements cible. Cet indice est calculé – en base 100 - comme le ratio du taux d'enregistrements cible cumulés et de l'effectif des tranches (en %)
- trois colonnes supplémentaires peuvent éventuellement figurer sur le gain chart, si une variable supplémentaire a été définie par l'utilisateur, donnant la :
  - somme par tranche
  - somme cumulée
  - moyenne sur la tranche

Exemple de calcul des indices :

	Taux de cible/tranche	% cumulé total cible	Indice cible / tranche	Indice cible cumulé
20%	62,08%	49,67%	248	248
40%	36,67%	79,00%	147	198
60%	20,00%	95,00%	80	158
80%	5,42%	99,33%	22	124
100%	0,83%	100,00%	3	100
Total	25,00%			

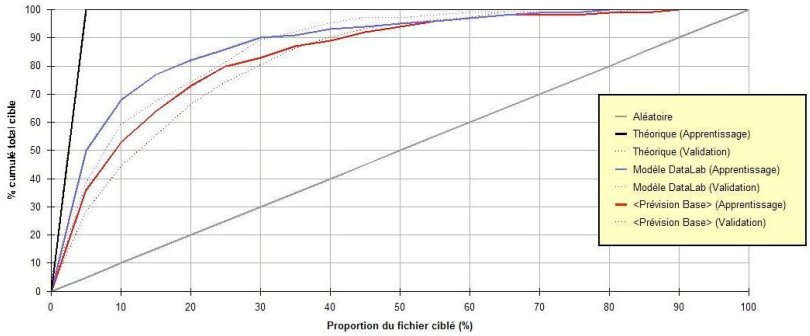
- Indice cible / tranche :
  - Le taux d'enregistrements cible de la tranche 0%-20% est de 62.08%.
  - Le taux d'enregistrements cible global est de 25%
  - L'indice du taux de cible pour cette tranche est égal à (62.08%/25%)*100 = 248
- Indice cible cumulé :
  - La tranche 0%-40% regroupe 79.00% de l'ensemble des enregistrements cible
  - L'indice cible cumulé est égal à (79.00%/40%)*100 = 198

Exemple de lecture du gain chart

Modèle DataLab (Validation)	Score minimum	Score moyen	Effectif / tranche	Effectif cumulé	Effectif cible/tranche	Effectif cible cumulé	Taux de cible/tranche	% cumulé total cible	Indice cible / tranche	Indice cible cumulé
20,00%	0,544	0,771	240	240	177	177	73,75%	59,00%	295	295
40,00%	0,185	0,350	240	480	78	255	32,50%	85,00%	130	213

- les 20% d'enregistrements ayant le taux de cible le plus fort représentent 240 enregistrements
- le score minimum de la tranche est de 0.544
- le score moyen de la tranche est de 0.771
- le nombre d'enregistrements cible dans la tranche est de 177, soit 73.75% et un indice de 295
- le nombre d'enregistrements cible cumulé représente 59.00% de la totalité des enregistrements cible, soit un indice de 295, donc 2.95 fois plus que la moyenne.

b) Représentation graphique



Le graphe présente le taux d'enregistrements cible cumulé (ou le pourcentage cumulé de la valeur cible si la variable cible est continue) en fonction de l'effectif trié selon le modèle choisi (cf. paramétrage du gain chart) et sur la population définie par l'utilisateur.

Les courbes représentées sont :

- La courbe théorique en apprentissage et/ ou validation : courbe obtenue en triant l'échantillon selon les enregistrements cible
- La courbe aléatoire d'équation y=x
- Selon le paramétrage défini :
  - la courbe du modèle DataLab ou de base en apprentissage et / ou validation
  - la courbe de la variable supplémentaire en apprentissage et / ou validation

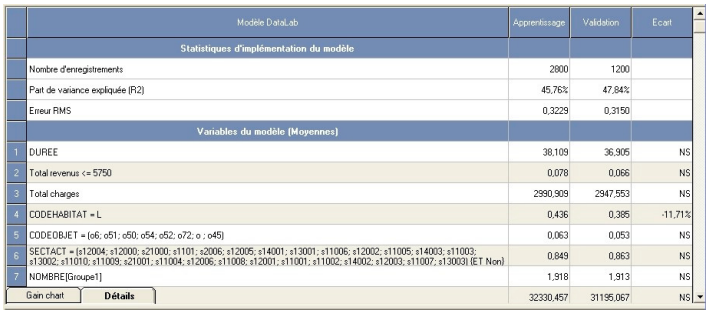
Exemple de lecture :

- 20% des enregistrements choisis au hasard concentrent 20% des enregistrements cible (aléatoire)
- 20% des enregistrements triés selon la valeur cible regroupent 100% des enregistrements cible (théorique)
- 20% des enregistrements triés selon le modèle DataLab concentrent 82% de tous les enregistrements cible (DataLab apprentissage)
- 20% des enregistrements triés selon le modèle de Base concentrent 73% de tous les enregistrements cible (Prévision base apprentissage)

Onglet Détail

La table compare deux échantillons parmi apprentissage, validation et données externes. La comparaison porte d'une part sur les statistiques du modèle choisi d'autre part sur la moyenne de chaque variable explicative du modèle sur chaque échantillon et l'écart relatif si celui ci est significatif à 95% de seuil de confiance..

Exemple de lecture de la comparaison d'échantillon



	Modèle DataLab	Apprentissage	Validation	Ecart
	Statistiques d'implémentation du modèle			
	Nombre d'enregistrements	2800	1200	
	Part de variance expliquée (R2)	45,76%	47,84%	
	Erreur RMS	0,3229	0,3150	
	Variables du modèle (Moyennes)			
1	DUREE	38,109	36,905	NS
2	Total revenus <= 5750	0,078	0,066	NS
3	Total charges	2990,909	2947,553	NS
4	CODEHABITAT = L	0,436	0,385	-11,71%
5	CODEOBJET = (o6; o51; o50; o54; o52; o72; o ; o45)	0,063	0,053	NS
6	SECTACT = (s12004; s12000; s21000; s1101; s2006; s12005; s14001; s13001; s11006; s12002; s11005; s14003; s11003; s13002; s11010; s11009; s21001; s11004; s12006; s11008; s12001; s11001; s11002; s14002; s12003; s11007; s13003) (ET Non)	0,849	0,863	NS
7	NOMBRE[Groupe1]	1,918	1,913	NS
	Gain chart / Détails	32330,457	31195,067	NS

La comparaison porte sur l'implémentation du modèle DataLab sur les échantillons d'apprentissage (2800 enregistrements) et de validation (1200 enr.)

- la part de variance expliquée (R2) est de 45,76% et de 47,84% respectivement sur les échantillons.
- Les variables du modèle DataLab ne présentent pas de différence significative sur les deux échantillons sauf la variable 'Code habitat = L' sous représentée de 12% sur la validation.

7. 2 Le redressement

1. Objet

L'utilisateur peut redresser un échantillon d'apprentissage s'il présente un biais par rapport à la population dont il est extrait.

Il est aussi possible, si un modèle a été élaboré sur un échantillon sur-représenté en enregistrements cible, donc non représentatif de la population initiale, de simuler les résultats du modèle sur la base du taux d'enregistrements cible réel.

Cette fiche technique détaille les différentes fonctionnalités de DataLab faisant appel au redressement.

2. Détail

2.1 Redressement d'un échantillon

L'échantillon d'apprentissage peut être redressé de 2 manières :

- En utilisant une variable de pondération
- Par ajustement de distribution

L'échantillon de validation peut être redressé simultanément en cochant la case « Appliquer le redressement au reste du fichier ».
L'ensemble des statistiques et analyses (hors 'Audit du fichier') porte sur l'échantillon redressé.

a) Variable de pondération

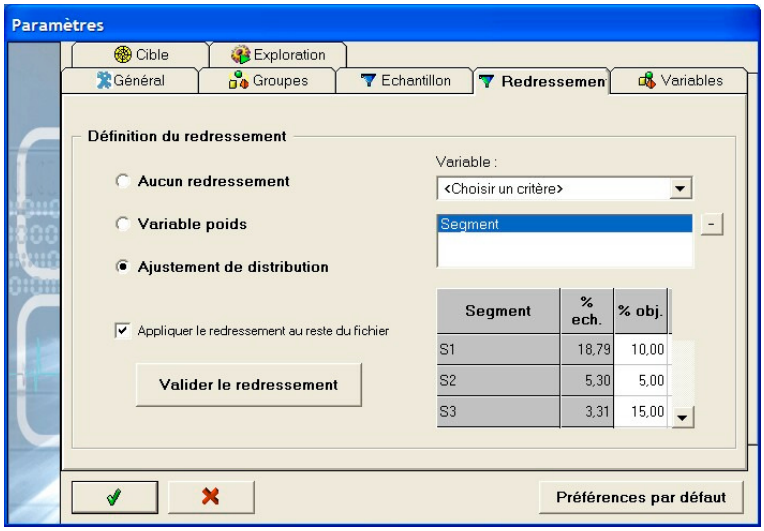
Le vecteur poids est dans ce cas une variable faisant partie de l'échantillon.
Elle doit être numérique et ne comporter aucune valeur manquante. La variable à expliquer ne peut être sélectionnée comme variable de pondération.
Les poids sommant à l'unité, le nombre d'enregistrements est identique à celui du fichier non redressé.

Exemple :

La variable de pondération donne un poids 2 à chaque enregistrement cible et un poids 1 aux autres enregistrements.
L'échantillon d'apprentissage comporte 1000 enregistrements, dont 100 enregistrements cible et 900 non cible.
Chaque enregistrement cible est pondéré par le coefficient 2/1100.
Chaque enregistrement non cible est pondéré par le coefficient 1/1100.

b) Ajustement de distribution

Le vecteur poids est dans ce cas déterminé de manière itérative afin de respecter une répartition sur un ensemble de critères de redressement (au maximum 5).



La fenêtre affiche, pour chaque critère choisi, la répartition de la variable (numérique ou catégorique) dans l'échantillon d'apprentissage. L'utilisateur peut alors entrer manuellement les valeurs souhaitées.

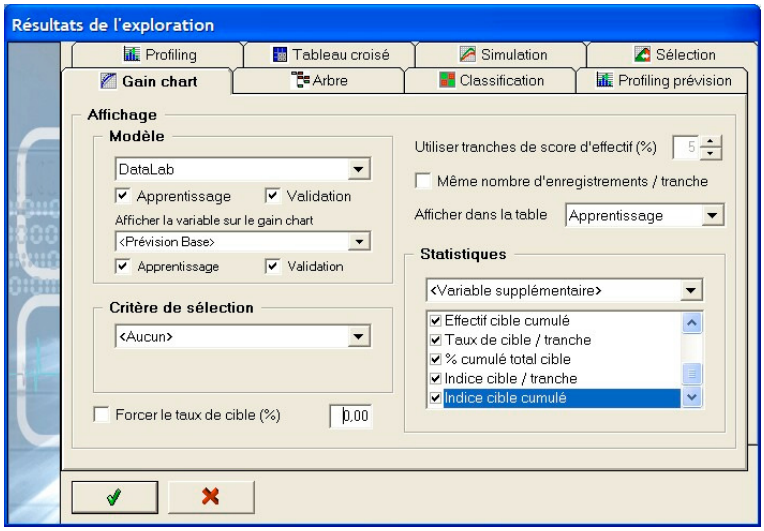
Note : Le découpage des variables numériques (par défaut en quartiles) peut être préalablement modifié en utilisant la fonctionnalité « Type et Classes ».

2.2 Résultats redressés

Lorsque le modèle a été élaboré sur un échantillon sur-représenté en enregistrements cible (par exemple 50/50), il est possible de simuler les résultats sur un échantillon comportant la distribution réelle de variable cible.

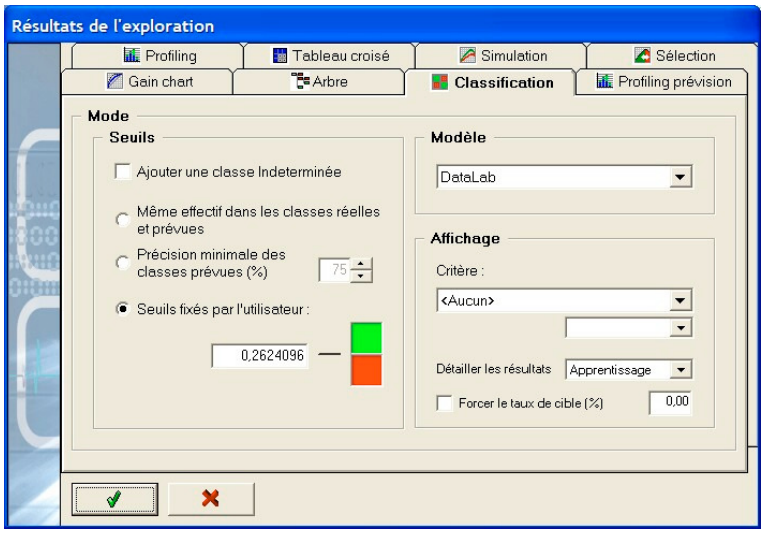
a) Gain chart

Le gain chart peut être paramétré en modifiant le taux de cible. Pour cela il suffit de choisir le taux de cible dans la case « Affecter à la cible le taux de … ».



b) Classification

Le module de classification peut être paramétré en modifiant le taux de cible. Pour cela il suffit de choisir le taux de cible dans la case « Affecter à la cible le taux de … ».



7. 3 Le profiling

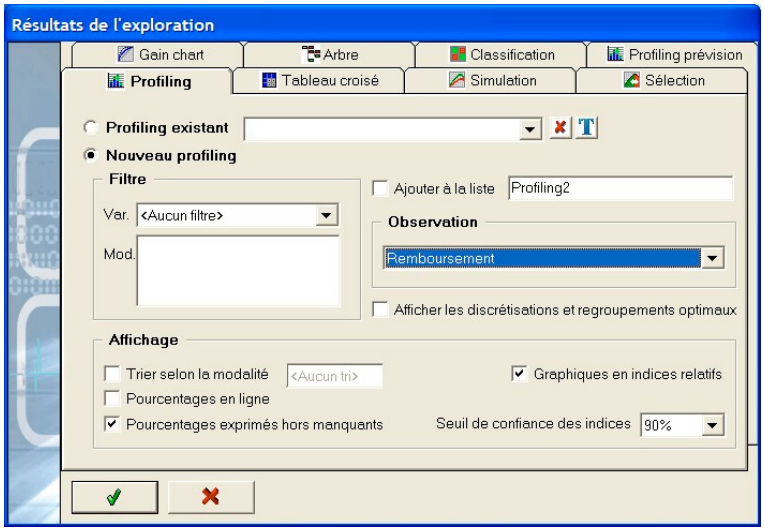
1. Objet

Le profiling permet de croiser automatiquement une variable, en particulier la variable à expliquer, avec toutes les autres. C'est donc une fonctionnalité de base essentielle pour identifier des critères individuellement discriminants.

Cette fiche technique détaille le mode de calcul d'un profiling.

2. Détail

2.1 Paramétrage du profiling



Le paramétrage porte sur :

- La variable à décrire
- La conservation ou non des valeurs manquantes pour le calcul du profiling
- Le seuil de confiance

a) La variable à décrire

Elle peut être discrète ou continue.

Si elle est discrète, toues les modalités de la variable apparaîtront lors du croisement avec l'ensemble des variables.

Si la variable est continue, elle apparaîtra discrétisée dans le profiling, sous la forme sous laquelle elle a été découpée dans le menu « Format des variables » - par défaut découpage en quartiles.

La variable à décrire peut être une des variables du fichier mais aussi les variables prévisions, arbre, typologie, segmentation, niche, cible.

b) Les valeur manquantes

Le principe du profiling consiste à croiser une variable à décrire, la comparaison s'effectuant sur la répartition des modalités de chaque variable. Pour cette raison, il est préférable de ne pas tenir compte des valeurs manquantes qui, si elles sont en proportions différentes selon les variables, biaisent l'observation.

c) Le seuil de confiance

Le seuil de confiance intervient dans le calcul de la significativité de l'indice de confiance des résultats, qui permet plus facilement d'identifier les critères discriminants.

2.2 Lecture des résultats

Les résultats sont présentés à l'écran dans 2 zones :

- Profiling
- Graphique

a) Profiling

Les résultats sont présentés sur l'échantillon d'apprentissage.

o	(CODEETAT)	Ensemble		OK			Pb		
	Variables & valeurs	Nb	%	Nb	%	Indice	Nb	%	Indice
	CODEETAT								
	OK	2 079	74,24%	2 079	100,00%	**135**	0	0,00%	100
	Pb	721	25,76%	0	0,00%	100	721	100,00%	**388**
	CIVILITE								
	M.	1 400	50,00%	999	48,05%	100	401	55,61%	**111**
	Mlle	280	10,00%	235	11,31%	100	45	6,23%	**62**
	Mme	1 120	40,00%	845	40,64%	100	275	38,16%	100
	Age								
	Moyenne	2 800	46,230	2 079	46,198		721	46,323	
	<=35	371	13,25%	261	12,57%	100	110	15,22%	100
	]35,45]	960	34,28%	721	34,67%	100	239	33,16%	100
	]45,55]	960	34,29%	728	35,02%	100	232	32,21%	100
	]55,65]	391	13,98%	301	14,50%	100	90	12,47%	100
	>65	118	4,20%	67	3,24%	**77**	50	6,94%	**165**
	SITUAFAMIL								
	A	97	3,46%	53	2,55%	**74**	44	6,09%	**176**
	C	565	20,16%	414	19,92%	100	151	20,87%	100
	D	426	15,22%	298	14,34%	100	128	17,76%	100
	M	1 459	52,11%	1 121	53,95%	100	338	46,79%	**90**
	S	69	2,48%	44	2,12%	100	25	3,51%	100
	V	93	3,33%	57	2,75%	100	36	4,98%	**150**
	VM	91	3,24%	91	4,37%	**135**	0	0,00%	100

Dans l'exemple ci-dessus, la variable CODETAT est la variable observée. Elle présente 2 modalités : OK et Pb.
Deux types de zones d'information sont présentées :

- La répartition de l'ensemble de la base, en nombre et %, par variable
- La répartition de chacune des modalités de la variable observée, en nombre et %, par variable

Si on prend l'exemple de la variable CIVILITE, la comparaison de la répartition selon les modalités de cette variable :

- De l'ensemble des enregistrements
- Des enregistrements du segment « Pb » de la variable CODETAT (segment des individus à risque)

Permet de mettre en évidence, dans le segment des individus « Pb », donc à risque :

- Une sur-représentation de M. : 51.43% contre 47.61% en moyenne (indice 108)
- Une sous-représentation de Mlle : 3.43% contre 7.05% en moyenne (indice 56)

Les indices sont calculés en base 100. Par exemple :

$$\text{Indice } 108 = (51.43\%/47.61\%)*100$$

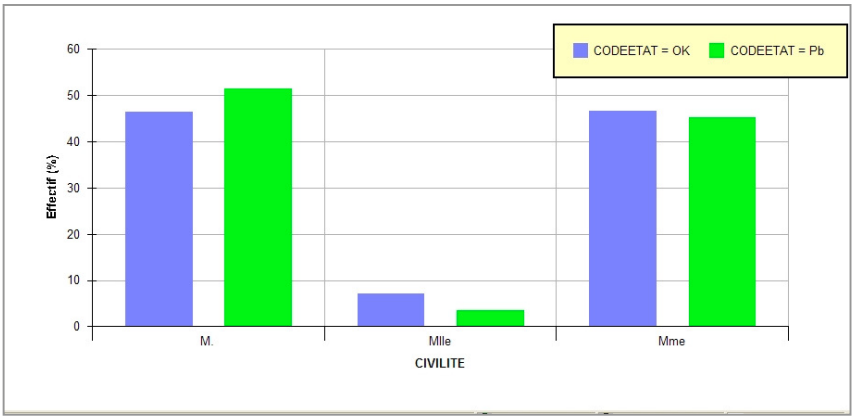
Les indices non significatifs sont égaux à 100 (par exemple indice de la modalité Mme).
Pour plus de lisibilité, les indices indiquant :

- Une sur-représentation significative sont indiqués en vert
- Une sous-représentation significative sont indiqués en rouge

b) Graphique

Le graphe affiché dans la partie inférieure de la fenêtre de résultats correspond à la variable cliquée dans le profiling.

Le graphe représente la comparaison des modalités (ou intervalles) de la variable décrite selon la variable choisie.



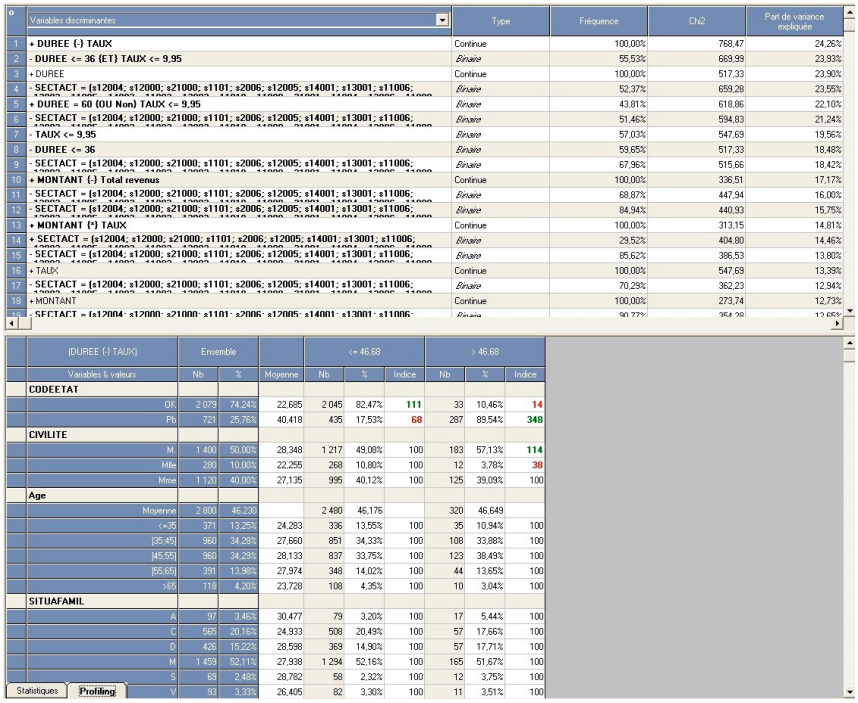
2.3 Utilisation du profiling dans DataLab

Le profiling apparaît aussi dans d'autres fonctionnalités de DataLab :

- **Variables explicatives > Variables discriminantes** : profiling des variables discriminantes
- **Variables explicatives > Segments** : profiling des segments
- **Modèles de score > Modèles** : profiling des variables du modèle
- **Modèles de score > Profiling du score** : profiling de la prévision (Top)

a) Variables discriminantes

On accède au profiling des variables discriminantes en cliquant sur l'onglet 'Profiling' dans la partie inférieure de la fenêtre de résultats.



	Variables discriminantes	Type	Fréquence	Chi2	Part de variance expliquée
1	+ DUREE (-) TAUX	Continue	100,00%	768,47	24,26%
2	- DUREE <= 36 (ET) TAUX <= 9,95	*Binaire*	55,53%	669,99	23,93%
3	+ DUREE	Continue	100,00%	517,33	23,90%
4	- SECTACT = {s12004; s12000; s21000; s1101; s2006; s12005; s14001; s13001; s11006;	*Binaire*	52,37%	659,28	23,55%
5	+ DUREE = 60 (OU Non) TAUX <= 9,95	*Binaire*	43,81%	618,86	22,10%
6	- SECTACT = {s12004; s12000; s21000; s1101; s2006; s12005; s14001; s13001; s11006;	*Binaire*	51,46%	594,83	21,24%
7	- TAUX <= 9,95	*Binaire*	57,03%	547,69	19,56%
8	- DUREE <= 36	*Binaire*	59,65%	517,33	18,48%
9	- SECTACT = {s12004; s12000; s21000; s1101; s2006; s12005; s14001; s13001; s11006;	*Binaire*	67,96%	515,66	18,42%
10	+ MONTANT (-) Total revenus	Continue	100,00%	336,51	17,17%
11	- SECTACT = {s12004; s12000; s21000; s1101; s2006; s12005; s14001; s13001; s11006;	*Binaire*	68,87%	447,94	16,00%
12	- SECTACT = {s12004; s12000; s21000; s1101; s2006; s12005; s14001; s13001; s11006;	*Binaire*	84,94%	440,93	15,75%
13	+ MONTANT (*) TAUX	Continue	100,00%	313,15	14,81%
14	+ SECTACT = {s12004; s12000; s21000; s1101; s2006; s12005; s14001; s13001; s11006;	*Binaire*	29,52%	404,80	14,46%
15	- SECTACT = {s12004; s12000; s21000; s1101; s2006; s12005; s14001; s13001; s11006;	*Binaire*	85,62%	386,53	13,80%
16	+ TAUX	Continue	100,00%	547,69	13,39%
17	- SECTACT = {s12004; s12000; s21000; s1101; s2006; s12005; s14001; s13001; s11006;	*Binaire*	70,29%	362,23	12,94%
18	+ MONTANT	Continue	100,00%	273,74	12,73%

	(DUREE (-) TAUX)	Ensemble			<= 46,68			> 46,68		
	Variables & valeurs	Nb	%	Moyenne	Nb	%	Indice	Nb	%	Indice
	CODEETAT									
	OK	2 079	74,24%	22,685	2 045	82,47%	**111**	33	10,46%	**14**
	Pb	721	25,76%	40,418	435	17,53%	**68**	287	89,54%	**348**
	CIVILITE									
	M	1 400	50,00%	28,348	1 217	49,08%	100	183	57,13%	**114**
	Mlle	280	10,00%	22,255	268	10,80%	100	12	3,78%	**38**
	Mme	1 120	40,00%	27,135	995	40,12%	100	125	39,09%	100
	Age									
	Moyenne	2 800	46,230		2 480	46,176		320	46,649	
	<=35	371	13,25%	24,283	336	13,55%	100	35	10,94%	100
	]35,45]	960	34,28%	27,660	851	34,33%	100	108	33,88%	100
	]45,55]	960	34,29%	28,133	837	33,75%	100	123	38,49%	100
	]55,65]	391	13,96%	27,974	348	14,02%	100	44	13,65%	100
	>65	118	4,20%	23,728	108	4,35%	100	10	3,04%	100
	SITUAFAMIL									
	A	97	3,46%	30,477	79	3,20%	100	17	5,44%	100
	C	565	20,16%	24,933	508	20,49%	100	57	17,66%	100
	D	426	15,22%	28,598	369	14,90%	100	57	17,71%	100
	M	1 459	52,11%	27,938	1 294	52,16%	100	165	51,67%	100
	S	69	2,48%	28,782	58	2,32%	100	12	3,75%	100
	V	93	3,33%	26,405	82	3,30%	100	11	3,51%	100

Statistiques | Profiling

Les variables cible continues apparaissent découpées de manière optimale dans le profiling (découpage issu de la phase d'exploration). Cette discrétisation peut être modifiée par l'utilisateur en double cliquant sur la variable dans la liste figurant dans la partie supérieure de la fenêtre de résultats.

b) Segments

De la manière analogue on dispose d'un profiling sur les segments. Les variables cible du profiling sont alors les variables binaires issues des segments.

c) Modèles

On dispose d'un profiling pour les variables sélectionnées dans les 2 modèles (base et DataLab).

Le principe de fonctionnement est le même que celui du profiling des variables discriminantes.

d) Prévision

On dispose enfin d'un profiling pour le top prévision (pourcentage égal à celui de la cible) après sélection du modèle (base ou DataLab) dans une fenêtre de paramétrage.

Le principe de fonctionnement est le même que celui du profiling des variables discriminantes.

7. 4 La classification

1. Objet

Le module de classification est une fonctionnalité très pratique pour analyser les résultats d'un score dans le cas où la variable cible est binaire.
Il permet de déterminer le ou les seuils de score transformant le score en classes et de disposer automatiquement de la matrice de classification correspondante.
Trois modes de classification sont disponibles :

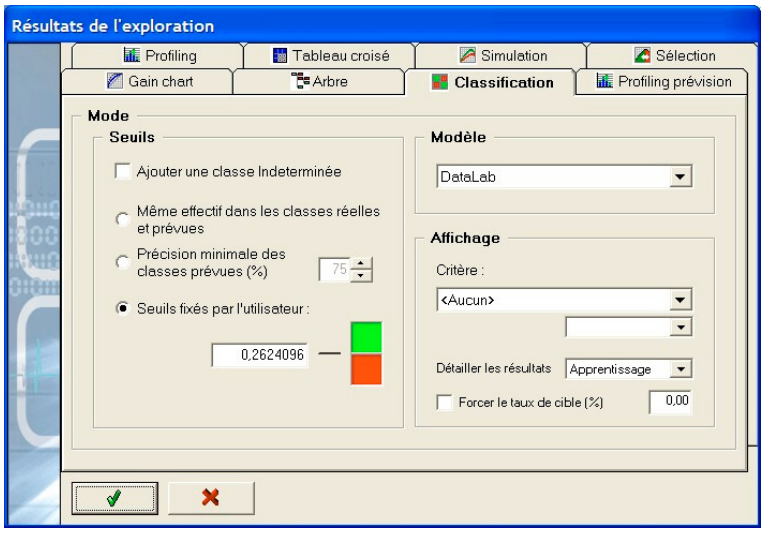
- Automatique avec des effectifs de classe identiques entre le prévu et le réel
- Automatique avec un minimum de précision par classe
- Défini par l'utilisateur

Cette fiche technique détaille les différentes options de classification et les résultats obtenus.

2. Détail

2.1 Paramétrage

Le paramétrage du module classification se fait à partir de la fenêtre Modélisation > Classification, pour une variable binaire uniquement.



a) Détermination des seuils

Il est possible d'utiliser le module de classification en 2 ou 3 classes, en ajoutant alors une classe indéterminée.

Pour cela cliquer « Ajouter une classe indéterminée ».

L'utilisateur a le choix entre 3 possibilités pour transformer le score en classes :

- Classification standard : les effectifs des classes prévues sont identiques à ceux des classes réelles (la classification est alors obligatoirement effectuée en 2 classes)
- Classification optimisée : un seuil de précision minimale est spécifié par l'utilisateur pour toutes les classes
- Classification manuelle : le seuil est spécifié par l'utilisateur

b) Choix du modèle

L'utilisateur choisit le modèle pour lequel il souhaite établir la matrice de classification :

- Modèle DataLab
- Modèle de base

c) Paramètres d'affichage

L'utilisateur peut afficher les résultats (matrice de classification) sur un segment en particulier, spécifié à partir de la liste déroulante « Critère ». Ceci permet de valider les résultats en classification sur divers segments et permet d'identifier, le cas échéant, des segments sur lesquels le modèle serait moins performant.

Les résultats sont affichés au choix sur l'échantillon d'apprentissage ou de validation, spécifié dans « Détailler les résultats ».

Enfin, la variable cible peut être redressée pour l'affichage des résultats. Ainsi, par exemple, il est possible de disposer de la matrice de classification pour un taux de cible réel, alors que le modèle a été élaboré sur un échantillon enrichi en valeurs cible (ex 50/50).

2.2 Résultat

Le résultat consiste en la matrice de classification calculée avec les paramètres choisis par l'utilisateur.

a) Matrice standard

La matrice ci-dessous correspond au cas standard, les effectifs prévus sont très proches des effectifs réels (75% pour la valeur 0 et 25% pour la valeur 1).

0	Cible observée (Ap...				Taux de bonnes prévisions
Cible prévue - Modèle de base	0	1	TOTAL	TOTAL (%)	
0 (<=0,396209)	1 846	254	**2 100**	**75,00%**	87,90%
1 (>0,396209)	254	446	**700**	**25,00%**	63,71%
TOTAL	**2 100**	**700**	**2 800**	**100,00%**	**81,86%**
TOTAL (%)	**75,00%**	**25,00%**	**100,00%**		

Le seuil retenu est de 0.396209.

Le taux de classification est de 81.86%, mais le taux de bonnes prévision de la classe 1 n'est que de 63.71%.

b) Amélioration des taux de bonnes prévisions

Le choix d'un seuil minimum de bonnes prévisions dans chaque classe de 70% donne la matrice de classification suivante :

o	Cible observée (Ap ...				Taux de bonnes prévisions
Cible prévue - Modèle de base	0	1	TOTAL	TOTAL (%)	
0 (<=0,497557)	1 947	341	2 288	81,71%	85,10%
1 (>0,497557)	153	359	512	18,29%	70,12%
TOTAL	2 100	700	2 800	100,00%	82,36%
TOTAL (%)	75,00%	25,00%	100,00%		

Le seuil retenu est de 0.497557, plus élevé que précédemment.

Le taux de classification global est de 82.36%, mais le taux de bonnes prévision de la classe 1 est maintenant de 70.12% et celui de la classe 0 de 85.10%. Cependant la classe 1 prévue ne représente plus que 18.29% (ce qui correspond à l'augmentation du seuil retenu).

c) Ajout d'une classe indéterminée

Le recours à une classe intermédiaire se justifie lorsque les conditions demandées pour le calcul de la matrice de classification ne peuvent être respectées.

Dans l'exemple ci-dessous, on spécifie un taux de bonnes prévision minimal de 85% pour chaque classe. Cette condition ne peut être respectée pour les deux classes : la classe 0 présente un taux de bonnes prévision de 80.75% seulement.

o	Cible observée (Ap ...				Taux de bonnes prévisions
Cible prévue - Modèle de base	0	1	TOTAL	TOTAL (%)	
0 (<=0,715749)	2 064	492	2 556	91,29%	80,75%
1 (>0,715749)	36	208	244	8,71%	85,25%
TOTAL	2 100	700	2 800	100,00%	81,14%
TOTAL (%)	75,00%	25,00%	100,00%		

On introduit donc une classe indéterminée, qui permet de respecter la condition demandée. Cette classe représente 9.29% de la population sur laquelle on décide de ne pas affecter de classe (cf ci-dessous).

o	Cible observée (Ap ...				Taux de bonnes prévisions
Cible prévue - Modèle de base	0	1	TOTAL	TOTAL (%)	
0 (<=0,499779)	1 951	345	2 296	82,00%	84,97%
Indet. ((0,499779,0,715749))	113	147	260	9,29%	-
1 (>0,715749)	36	208	244	8,71%	85,25%
TOTAL	2 100	700	2 800	100,00%	85,00%
TOTAL (%)	75,00%	25,00%	100,00%		

d) Application de la matrice de classification sur un segment

La matrice ci-dessous représente l'application de la matrice de classification standard au segment des plus de 60 ans (à comparer avec matrice du paragraphe A/).

o	Cible observée (Ap...				Taux de bonnes prévisions
Cible prévue - Modèle de base	0	1	TOTAL	TOTAL (%)	
0 (<=0,396209)	125	50	**175**	**70,85%**	71,43%
1 (>0,396209)	24	48	**72**	**29,15%**	66,67%
TOTAL	**149**	**98**	**247**	**100,00%**	**70,04%**
TOTAL (%)	**60,32%**	**39,68%**	**100,00%**		

On observe les performances globalement très dégradées du score sur ce segment :

- Seulement 70.04% de taux de classification contre 81.86% sur la totalité de la population
- En revanche, on a 66.67% de bonnes prévisions sur la classe 1 contre 63.71%

e) Redressement du taux de cible

Si le taux de cible réel n'est pas de 25% mais 5%, il est possible de disposer automatiquement de la matrice en affectant au taux de cible la valeur 5%.

La matrice obtenue est alors la suivante :

o	Cible observée (Ap...				Taux de bonnes prévisions
Cible prévue - Modèle de base	0	1	TOTAL	TOTAL (%)	
0 (<=0,396209)	2 338	51	**2 389**	**85,32%**	97,87%
1 (>0,396209)	322	89	**411**	**14,68%**	21,71%
TOTAL	**2 660**	**140**	**2 800**	**100,00%**	**86,70%**
TOTAL (%)	**95,00%**	**5,00%**	**100,00%**		

7. 5 L'exploration ou Data Scanning

1. Objet

Le module d'exploration est la fonctionnalité centrale de DataLab. Il permet une exploration totalement automatique et quasi exhaustive des données jusqu'à la phase de modélisation.
Ce module dispose d'une feuille de paramétrage avec des options au niveau :

- Des traitements réalisés individuellement sur chaque variable
- Des traitements réalisés en combinant les variables
- De la phase finale de modélisation (ou sélection de variables)

Cette fiche technique détaille le fonctionnement du module d'exploration et son paramétrage.

2. Détail

2.1 Principe du module exploratoire de DataLab

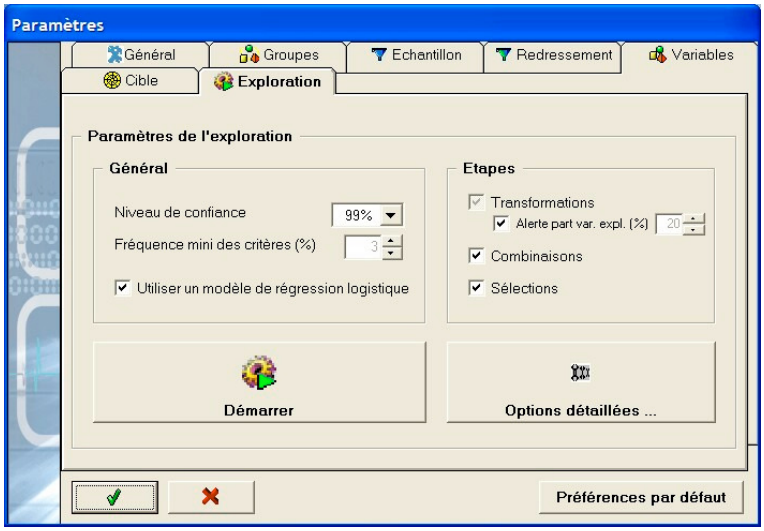
Le module exploratoire de DataLab enchaîne 3 étapes visant :

- D'une part à générer et tester de nouvelles variables explicatives :
  - Transformations sur les variables individuelles
  - Combinaisons entre variables binaires ou continues
- D'autre part à bâtir un modèle de prévision - sélectionnant parmi l'ensemble des variables générées et discriminantes les variables les moins corrélées et les plus explicatives

Chacune de ces 3 étapes (transformations, combinaisons, sélection) peut être paramétrée.

2.2 Paramétrage

La fenêtre de paramétrage comporte 2 niveaux (général et détaillé) et est accessible par la fonctionnalité **Cibler > Démarrer le Data Scanning**



a) Paramétrage général

Les paramètres de l'exploration concernent les points suivants.

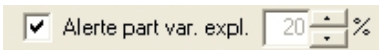
a1) au niveau général

- choix de seuil de confiance de l'ensemble des tests statistiques : 90%, 95%, 99% ou 99.5%
- choix de seuil minimum des critères : par défaut 5% (entre 1% et 10%)

  - les variables binaires dont une modalité présente une fréquence inférieure au seuil choisi ne seront pas retenues
  - un seuil plus élevé garantit une robustesse plus importante du modèle résultant
  - l'abaissement du seuil permet la prise en compte de phénomènes plus rares
- type de modèle : par défaut le modèle est réalisé avec une régression logistique lorsque la variable à expliquer est une variable binaire et linéaire lorsque celle-ci est continue

a2) les 3 étapes

- transformations : étape minimum. La case est impossible à décocher. L'étape des transformations est cependant paramétrable de manière détaillée
- combinaisons : étape qu'il est possible de supprimer. L'exploration se fait alors uniquement au niveau des variables individuelles, sans croisement des variables entre elles.
- alerte :



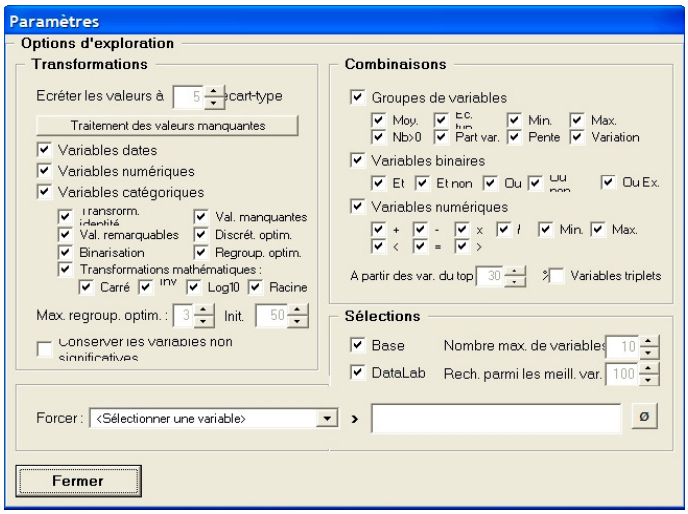
  Cette alerte permet, lorsqu'elle est activée, de lister toutes les variables présentant une part de variance expliquée supérieure au seuil retenu (par défaut 20%), pour validation par l'utilisateur. Cette fonctionnalité permet en particulier d'identifier les variables très corrélées avec la cible et qui auraient dues être exclues (par exemple, dans le cas de l'élaboration d'un score de cross selling, variable montant de l'achat lorsque la variable à expliquer est l'acte d'achat).

- sélections : étape qu'il est possible de supprimer. Il n'y a pas dans ce cas d'étape de modélisation, les résultats disponibles concernent alors uniquement les discrétisations optimales et variables discriminantes, ainsi que l'ensemble des variables discriminantes

b) Paramétrage détaillé

On accède à la fenêtre de paramétrage détaillé en cliquant sur le bouton 

:



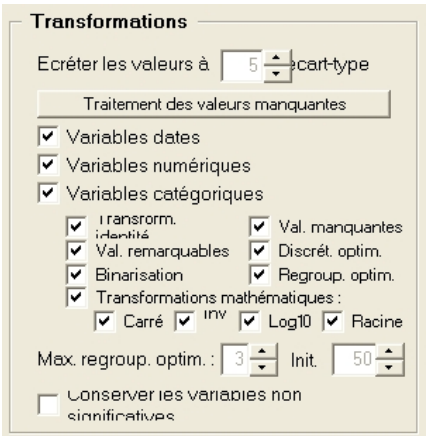
Cette fenêtre est partagée en 3 zones :

- Transformations
- Combinaisons
- Sélections

b1) Transformations

Toutes les variables du fichier importé subissent a priori un ensemble de transformations, visant à générer un ensemble de nouvelles variables qui seront ensuite combinées 2 à 2 ou 3 à 3. L'ensemble des traitements est totalement paramétrable.

Le pouvoir discriminant de toutes les variables générées est évalué (test de Fisher).



Ecrêtage

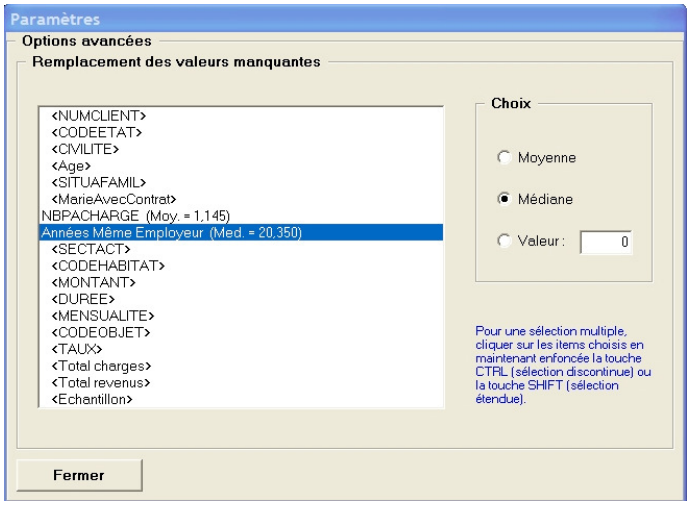
Les variables sont par défaut automatiquement écrêtées à la moyenne +/- 5 écart types. Les enregistrements présentant des valeurs extrêmes ne sont donc pas supprimés, mais tronquées. Le nombre d'écart types retenu pour l'écrêtage peut être fixé entre 1 et 99.

Traitement des valeurs manquantes

Il s'agit ici du mode de remplacement des valeurs manquantes présentes dans les variables continues (la présence d'une valeur manquante dans une variable génère par ailleurs une variable binaire indicatrice de la valeur manquante).

On accède à ce paramétrage en cliquant sur le bouton

Le choix de la valeur remplaçant les valeurs manquantes se fait variable par variable ou globalement. Par défaut celles-ci sont remplacées par la moyenne, mais il est aussi possible de remplacer les valeurs manquantes par la médiane ou par une valeur particulière.



Variables date

Lorsque la case est cochée les transformations s'appliquent sur les variables date (transformation identité, valeurs remarquables, discrétisation optimale, transformations mathématiques)

Variables numériques

Lorsque la case est cochée les transformations s'appliquent sur les variables numériques (transformation identité, valeurs remarquables, discrétisation optimale, transformations mathématiques)

Variables catégoriques

Lorsque la case est cochée les transformations s'appliquent sur les variables catégoriques (binarisation, valeurs remarquables, regroupements optimaux)

Transformation identité

Les variables continues sont traitées sans aucune modification en dehors de l'écrêtage et du remplacement des valeurs manquantes par la moyenne.

Valeurs remarquables

Recherche, parmi les variables binaires ou continues, des valeurs dont la fréquence est supérieure au seuil défini dans le paramétrage général et significatives (test de Fisher).

Binarisation

Binarisation des variables catégoriques, c'est à dire éclatement de la variable catégorique en autant de variables binaires que de modalités.

Valeurs manquantes

Création pour chaque variable catégorique ou continue d'une indicatrice des valeurs manquantes.

Discrétisations optimales

Recherche de la discrétisation optimale par rapport à la variable cible de chaque variable numérique – pour plus d'information sur ce sujet, voir la note « info_technique_N4 »

Regroupements optimaux

Recherche des regroupements optimaux par rapport à la variable cible de chaque variable catégorique – pour plus d'information sur ce sujet, voir la note « info_technique_N4 »

Transformations mathématiques

Génération, à partir des variables numériques ou date ayant subi une transformation identité, de nouvelles variables numériques, en utilisant les opérateurs mathématiques (Carré, inverse, log10, racine)

Paramètres de la fonctionnalité de discrétisation optimale et regroupements optimaux

Ces paramètres concernent le nombre maximum de classes souhaité et le nombre initial de classes.

Ces paramètres sont utilisés à la fois pour la discrétisation optimale des variables continues et le regroupement optimal des modalités des variables catégoriques.
Le nombre maximum de classes (paramètre « Max. regroup. optim. ») peut varier entre 2 et 5, et est égal à 3 par défaut. Le nombre de classes obtenu sera dans ce cas de 2 ou 3.

> ***Note*** *: Etant donné la combinatoire élevée, les temps de traitements augmentent rapidement lorsque le nombre de classes choisi dépasse 4.*

Le nombre initial de classes (paramètre « Init. »)peut varier entre 10 et 100 ; il est égal à 50 par défaut.

Note : pour plus d'information sur ce sujet voir « info_technique_N4 »

Exemples

Une variable catégorique, telle qu'une CSP comportant 8 modalités, génèrera :

- 8 variables binaires correspondant aux modalités de la variables (les 8 CSP)
- une variable indicatrice des valeurs manquantes, s'il existe au moins une valeur manquante dans l'échantillon
- le cas échéant, autant de variables binaires que de valeurs remarquables identifiées (modalité supérieure au seuil de fréquence et significative)
- le cas échéant, autant de variables binaires que de regroupements de modalités (CSP) optimaux

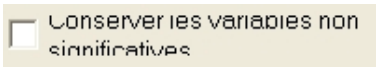
Une variable numérique, telle qu'un âge, génèrera :

- la variable âge écrêtée et dans laquelle les valeurs manquantes sont remplacées par la moyenne (par défaut), la médiane ou une valeur choisie par l'utilisateur
- une variable indicatrice des valeurs manquantes, s'il existe au moins une valeur manquante dans l'échantillon
- le cas échéant, autant de variables binaires que de valeurs remarquables identifiées (modalité supérieure au seuil de fréquence et significative)
- le cas échéant, autant de variables binaires que de classes (tranches d'âge) optimales
- 4 nouvelles variables : Carré (âge), Inverse (âge) – si cette variable ne comporte pas de valeurs nulles, log10 (âge) – si cette variable ne comporte pas de valeurs nulles, racine (âge)

Profondeur d'exploration

Par défaut, seules les variables transformées significatives sont conservées pour les étapes ultérieures de combinaisons et sélections.

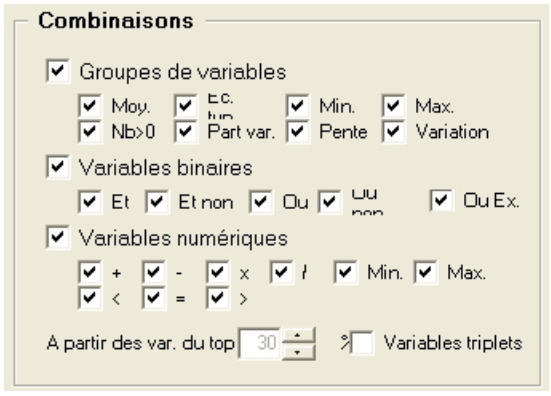
Il est cependant possible de conserver toutes les variables issues de l'étape de transformation – même non significatives - en cochant la case « Conserver les variables non significatives » :



b2) Combinaisons

Par défaut, toutes les variables transformées significatives et satisfaisant au test de fréquence sont ensuite combinées entre elles. Cette étape est totalement paramétrable.

Le pouvoir discriminant de toutes les variables générées est évalué (test de Fisher).



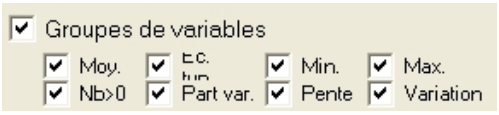
Groupe de variables

La fonctionnalité « Groupe de variables » permet de combiner entre elles un ensemble de variables choisies par l'utilisateur (menu , car de même nature Données > Définir > Groupes)

Exemple : 12 variables correspondant aux 12 CA mensuels de janvier à décembre.

Une fois le groupe défini par l'utilisateur, DataLab génère, sur ce groupe, un ensemble de variables métier.

L'ensemble de cases à cocher suivant permet de paramétrer les traitements retenus sur les groupes de variables définis.



- « Groupes de variables » : autorise l'évaluation des variables issues des groupes de variables (lorsqu'ils ont été définis)
- « Moy » : calcule la variable Moyenne des variables composant le groupe.
- « Ec. Typ. » : calcule la variable Ecart type des variables composant le groupe
- « Min » : calcule la variable Minimum des variables composant le groupe
- « Max » : calcule la variable Maximum des variables composant le groupe
- « Nb>0 » : calcule une variable donnant le Nombre de valeurs strictement positives parmi les variables composant le groupe
- « Part var. » : calcule pour chaque variable composant le groupe sa proportion par rapport à la somme des variables du groupe, (Part(i) étant définie comme le ratio de la variable (i) sur la somme de toutes les variables Var(i)).
- « Pente » : calcule la variable Pente des variables composant le groupe
- « Variation » : calcule la variable Variation des 2 dernières variables du groupe

***Note** : Les variables Pente et Variation ne sont calculées que pour les groupes ordonnés.*

Variables binaires

Le paramétrage concerne les traitements retenus pour les combinaisons entre variables binaires.

- « Variables binaires » : autorise la création et l'évaluation des combinaisons de variables binaires
- Les variables binaires sont combinées entre elles avec les opérateurs : ET, ET NON, OU, OU NON, OU EXCLUSIF

D'une manière générale Variable1 {OPERATEUR} Variable2 vaut 1 si la condition est vraie, et 0 sinon

Exemple :

- Age <40 {ET} CSP=cadre : variable valant 1 si Age <40 et CSP=cadre et 0 sinon
- Age <40 {ET Non} CSP=cadre : variable valant 1 si Age < 40 et CSP = non cadre et 0 sinon
- Age <40 {OU} CSP=cadre : variable valant 1 si Age <40 ou CSP=cadre et 0 sinon
- Age <40 {OU Non} CSP= variable valant 1 si cadre : Age < 40 ou CSP = non cadre et 0 sinon
- Age <40 {OU Excl.} CSP=cadre : variable valant 1 l'une des 2 conditions seulement est vraie, et 0 si les 2 sont vraies et les 2 sont fausses

Variables numériques

Le paramétrage concerne les traitements retenus pour les combinaisons entre variables numériques.

- « Variables numériques » : autorise la création et l'évaluation des combinaisons de variables numériques
- Les variables binaires sont combinées entre elles avec les opérateurs :

  - Numériques : addition, soustraction, multiplication, division
  - Min, Max
  - Relationnel : inférieur, égal, supérieur

Exemple :

- CA1{+}CA2 : variable calculée comme la somme des 2 variables CA1 et CA2
- CA1{-}CA2 : variable calculée comme la différence des 2 variables CA1 et CA2
- CA1{*}CA2 : variable calculée comme le produit des 2 variables CA1 et CA2
- CA1{/}CA2 : variable calculée comme le ratio des 2 variables CA1 et CA2, si CA2 ne comporte aucune valeur nulle
- CA1{Min}CA2 : variable calculée comme le minimum des 2 variables CA1 et CA2
- CA1{Max}CA2 : variable calculée comme le maximum des 2 variables CA1 et CA2
- CA1{<}CA2 : variable binaire valant 1 si la valeur de CA1 est inférieur à celle de CA2 et 0 sinon
- CA1{=}CA2 : variable binaire valant 1 si la valeur de CA1 est égal à la valeur de CA2 et 0 sinon
- CA1{>}CA2 : variable binaire valant 1 si la valeur de CA1 est supérieur à la valeur de CA2 et 0 sinon

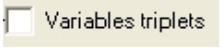
Profondeur d'exploration

Deux paramètres conditionnent la profondeur d'exploration au niveau des combinaisons.

- La proportion de variables transformées utilisées pour générer les variables combinées :

Par défaut, les variables combinées sont créées à partir du top 30% de meilleures variables transformées. Cette valeur est paramétrable et varie entre 5% et 100%.

- Le niveau de combinatoire

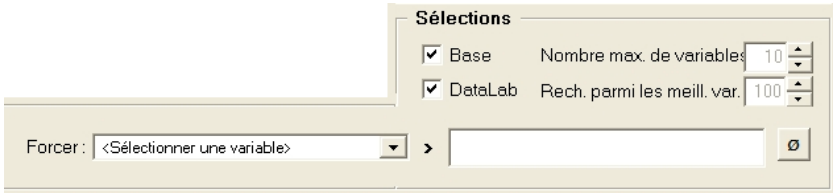


Les variables transformées sont systématiquement combinées 2 à 2, en fonction des paramètres choisis. Il est possible d'augmenter la combinatoire en cochant la case « Variables triplets ». Les variables seront alors combinées 3 à 3.

Note : Les temps de traitement augmentent fortement lorsque cette option est cochée.

b3) Sélections

La dernière étape consiste en une étape de modélisation, à partir de l'ensemble des variables disponibles, - variables d'origine, variables transformées et combinées - et significatives.



Techniques de modélisation

La technique utilisée est une régression (linéaire ou logistique selon le type de variable à expliquer, continue ou binaire) pas à pas, comportant un de test de colinéarité calcul du VIF)

Choix des modèles

Par défaut, DataLab réalise 2 modèles :

- Base : modèle réalisé uniquement avec les variables d'origine (ou variables de base). Les seuls traitements effectués sur les variables sont alors :
  - Binarisation des variables catégoriques
  - Remplacement des valeurs manquantes par la moyenne
- DataLab : modèle réalisé avec toutes les variables générées par DataLab

Il est possible de décocher une des 2 options (ou les 2) pour réaliser alors un seul modèle.

Nombre de variables en entrée du modèle

Le paramètre « Rech. parmi les meill. var. » permet de spécifier le nombre de variables en entrée du modèle : par défaut, ce nombre est fixé à 100, ce qui signifie que pour chaque variable d'origine, le modèle prendra en compte les 100 meilleures variables (y compris celle-ci), c'est à dire celles présentant la part de variance expliquée la plus élevée.

Ce paramètre peut varier entre 50 et 950.
Cette option peut permet de réduire de manière importante la recherche de la sélection de variables la plus explicative, donc les temps de traitements.

Nombre de variables sélectionnées dans le modèle

Le « nombre max. de variables » fixe le nombre maximum de variables acceptées dans chacun des 2 modèles. Par défaut, ce paramètre est fixé à 10. Il peut varier entre 1 et 30. Le nombre de variables retenu par le modèle est potentiellement inférieur au nombre maximal, étant donné la technique de modélisation.

Variable forcée dans le modèle

Il est possible de forcer des variables d'origine dans les 2 modèles (discrétisées dans le cas de variables catégoriques). Les variables sont choisies dans la liste déroulante (au maximum 10 variables)

Note sur les temps de calcul

Certaines options ont un impact important sur les temps de calcul: Les principales sont détaillées ci-dessous.

- Transformations de variables
  - Option de discrétisation optimale : à partir de 4 classes et surtout de 5 classes, le temps de calcul augmente de manière importante. Il est cependant possible, si nécessaire, de fixer ponctuellement (et non pour toutes les variables) une discrétisation en un nombre plus élevé de classes (cf. à ce sujet la note « info_technique_N4 »
  - L'option « conserver les variables non significatives » augmente les temps de calcul de manière importante, pour un gain en général très marginal
- Combinaisons de variables
  - La proportion de variables transformées retenues pour calculer les combinaisons : le paramètre par défaut (Top 20%) est généralement suffisant
  - L'option « combinaisons par triplets » augmente les temps de calcul de manière très importante

7. 6 Discrétisation et regroupement de modalités

1. Objet

La discrétisation des variables continues (ainsi que le regroupement des modalités des variables catégoriques) sont effectuées dans DataLab, de 2 manières différentes :

- par l'utilisateur : cette démarche répond à un but descriptif uniquement (présentation de statistiques à plat, statistiques croisées, profiling)
- par DataLab lors de la phase d'exploration : les découpages de variables sont alors optimaux par rapport à la variable à expliquer, et permettent de prendre en compte les phénomènes non linéaires.

Cette fiche technique détaille le mode de découpage des variables dans DataLab et le regroupement des modalités.

2. Détail

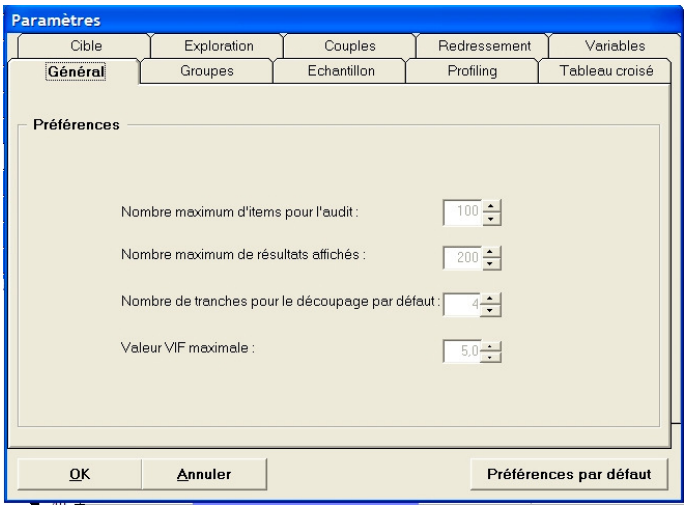
2.1 Discrétisations à but descriptif

Les discrétisations à but descriptif concernent les variables continues de base et les variables transformées ou combinées issues de la phase d'exploration.

a) Discrétisation par défaut des variables de base

Les variables continues de base sont par défaut discrétisées en quartiles. Toutes les valeurs, y compris les valeurs manquantes sont dans ce cas considérées.

Ce paramétrage figure dans l'onglet Type &Classes > Général : « Nombre de tranches pour le découpage par défaut ». Le fichier doit être rechargé pour que le nouveau paramétrage soit pris en compte.



b) Modification de la discrétisation par défaut

La discrétisation par défaut peut être modifiée par l'utilisateur, individuellement pour chaque variable. L'objectif est de pouvoir disposer des statistiques descriptives sur des tranches plus lisibles (par exemple : âge en tranches de 10 ans…).

En particulier, les nouveaux seuils de discrétisation définis par l'utilisateur ne sont pas pris en compte dans la phase d'exploration de DataLab.

La discrétisation d'une variable continue peut être effectuée soit totalement librement soit de manière automatique avec différentes options.

Dans tous les cas, les valeurs manquantes ne sont pas considérées et les résultats sont présentés sur l'échantillon d'apprentissage (hormis les statistiques descriptives présentées à la fois les échantillons d'apprentissage et de validation).

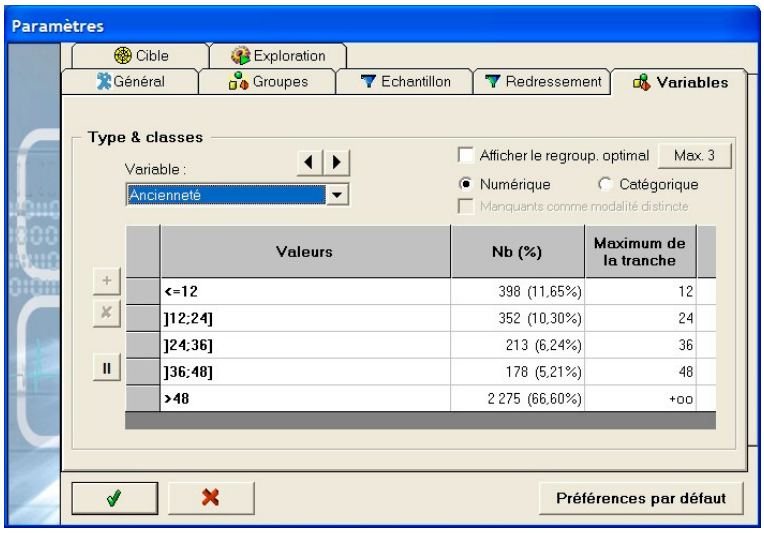
Discrétisation libre

On utilise la fonctionnalité Type & classes > onglet [Variables] :

Le nombre de classes peut être modifié en cliquant sur un ligne puis sur les boutons + ou − pour ajouter ou supprimer une classe.

Les seuils des classes peuvent être modifiés par l'utilisateur dans la colonne « Maximum de la tranche ».

Le libellé des classes est automatiquement mis à jour en fonction des valeurs saisies comme maximum des tranches. Il peut cependant être modifié par l'utilisateur dans la colonne Valeurs.



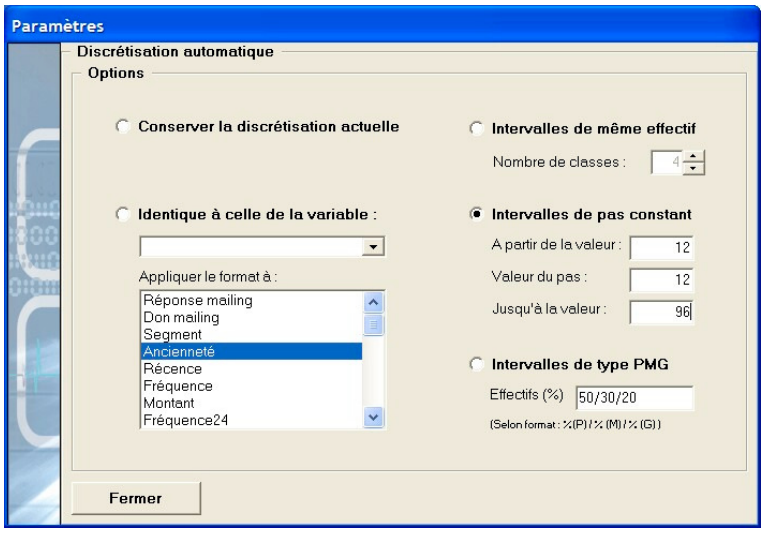
Discrétisations automatiques

Il y a plusieurs modes de discrétisation automatique pour les variables continues. On y accède par la fonctionnalité Type & classes > onglet [Variables] puis en cliquant sur le bouton II.

Différentes options sont alors possibles :

- discrétisation en intervalles de même effectif (par exemple déciles) > saisir le nombre de classes désiré. Dans le cas de plages de valeurs identiques, les intervalles peuvent présenter des effectifs légèrement différents.

- discrétisation en intervalles de pas constant > saisir les valeurs minimales et maximales des classes ainsi que la valeur du pas.
- Discrétisation identique à celle d'une autre variable continue : choisir la variable dans la liste déroulante et, éventuellement, élargir en sélectionnant dans la liste 'Appliquer à' l'application du format à d'autres variables.
- Discrétisation en classes Petit (P), moyen (M), gros (G). Entrer les valeurs respectives séparées par un signe /



2.2 Discrétisations optimales

La discrétisation optimale est à but explicatif, et porte sur les variables numériques ou catégoriques. Elle est réalisée au début de la phase d'exploration. La recherche du nombre de classes et des seuils des classes est ici *optimale par rapport à la variable à expliquer*.

a) Discrétisation des variables de base continues

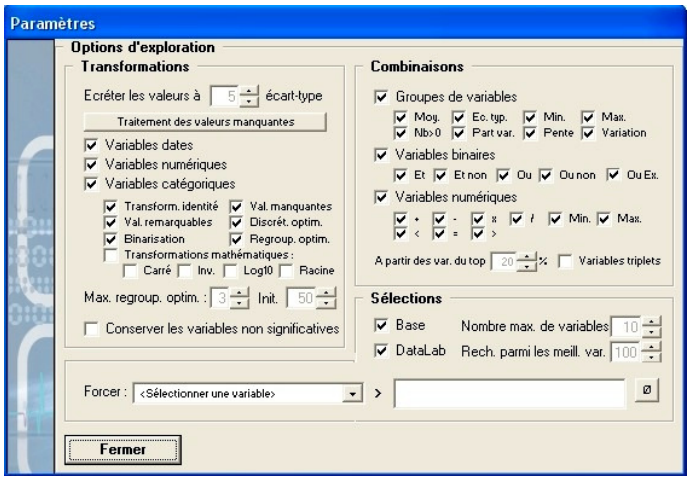
Principe

Par défaut, DataLab cherche à discrétiser la variable en au plus 3 classes (à la fois pour des raisons de temps de traitement et de robustesse), à partir d'une discrétisation initiale en 50 classes. Toutes les classes agrégées possibles sont évaluées automatiquement de manière à maximiser le Chi2 normé.

Paramétrage de la discrétisation optimale

Cette discrétisation peut être paramétrée dans la feuille de paramètres obtenue par la séquence **Cibler > Démarrer le Data Scanning** (+ bouton 

)



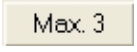
Les 2 paramètres concernés sont :

- « Max. regroup. optim » : nombre maximal de classes pour la discrétisation, qui peut varier de 2 à 5. Par exemple, si on a choisi 4 classes, DataLab cherche une discrétisation en 4 classes, puis 3, puis 2 et retient celle qui maximise le Chi2 normé.
- « Init. » : nombre d'intervalles de départ, qui peut varier entre 10 et 100

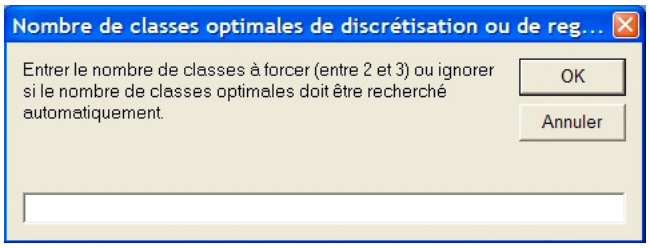
Ces paramètres concernent l'ensemble des variables.

Personnalisation du paramétrage de la discrétisation optimale

Il est possible de forcer DataLab à discrétiser une variable donnée en un nombre exact de classes. Le nombre de classes demandé doit être inférieur ou égal au nombre maximum de classes défini par défaut (par exemple, 2 classes exactement alors que le paramètre par défaut est 3).

Le paramétrage se fait dans la feuille de paramètres obtenue à partir de Type & classes > onglet [Variables] puis en cliquant sur le bouton en haut à droite « Max. n » : 

Il suffit ensuite d'entrer le nombre de classes à forcer. Ce paramétrage est pris en compte à la prochaine exploration.



Résultat de la discrétisation optimale

Les discrétisations optimales résultent de la phase d'exploration et sont disponibles par la fonctionnalité Exploration > Résultats > Regroupements de valeurs.

Regroupements	Nb	%	% cible
TOTAL	2 800	100%	25,76%
CIVILITE			Chi2 = 15,35
= Mlle	280	10,00%	16,05%
= (Mme; M.)	2 520	90,00%	26,84%
Age			Chi2 = 27,58
<= 30,9	173	6,18%	34,08%
=]30,9; 63,7]	2 480	88,57%	24,27%
> 63,7	147	5,25%	41,26%
SITUAFAMIL			Chi2 = 61,66
= VM	91	3,25%	0,00%
= (M; C; D)	2 450	87,50%	25,15%
= (S; V; A)	259	9,25%	40,55%
MarieAvecContrat			
NBPACHARGE			Chi2 = 14,21
<= 3	2 712	96,86%	25,23%
> 3	86	3,07%	43,30%
Années Même Employeur			Chi2 = 42,39
<= 13,3	607	21,68%	27,83%
> 13,3	1 478	52,79%	18,48%
SECTACT			Chi2 = 243,37
= (s12004; s12000; s21000; s1101; s2006;	2 452	87,57%	20,91%
= (s91000; s21002; s21006; s21005; s21003;	348	12,43%	59,99%
CODEHABITAT			Chi2 = 158,50
= (Ac; LF; P)	1 280	45,71%	16,33%

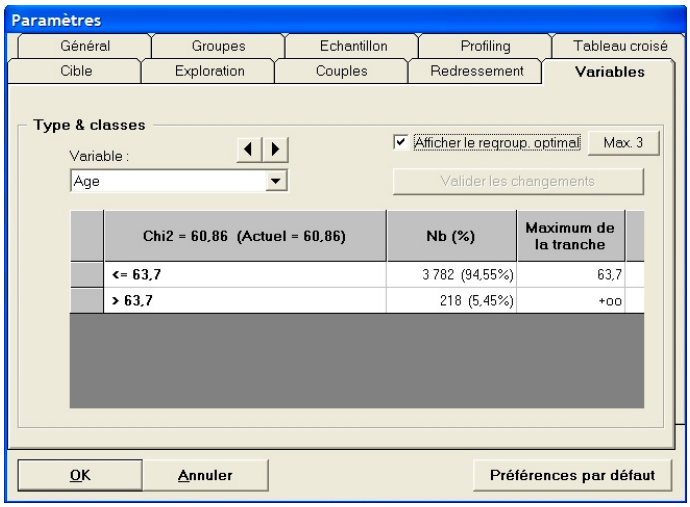
Modification de la discrétisation optimale

Les discrétisations peuvent être modifiées par l'utilisateur – mais pas le nombre de classes, à moins de relancer la phase d'exploration après modification des paramètres de discrétisation.

La modification est réalisée à partir de la fenêtre Type & classes > onglet [Variables], puis en cochant sur la case ☑ Afficher le regroup. optimal. La modification des seuils se fait alors de la même manière que lors d' une discrétisation manuelle.

L'utilisateur peut apprécier l'impact de la modification en comparant le nouveau Chi2 avec le Chi2 optimal.

Après validation - en cliquant sur le bouton Valider les changements - la nouvelle discrétisation remplace la discrétisation optimale initiale. Les nouveaux seuils remplacent les anciens dans tous les cas où ils apparaissaient dans les résultats (variables transformées, combinées). Les paramètres du modèle sont, le cas échéant, aussi mis à jour.



b) Regroupement optimaux des modalités des variables de base catégoriques

Principe

Le principe est similaire à celui de la discrétisation des variables continues.

Paramétrage de la discrétisation optimale

Les paramétrage est également identique. Cependant, le nombre d'intervalles initiaux est remplacé par le nombre de modalités (si celui-ci est inférieur à la valeur Init.)

Personnalisation du paramétrage de la discrétisation optimale

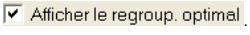
Il est possible de forcer DataLab à discrétiser une variable catégorique donnée en un nombre exact de classes. Le procédé est identique à celui concernant les variables continues.

Résultat de la discrétisation optimale

Les discrétisations optimales résultent de la phase d'exploration et sont disponibles par la fonctionnalité Exploration > Résultats > Regroupements de valeurs.

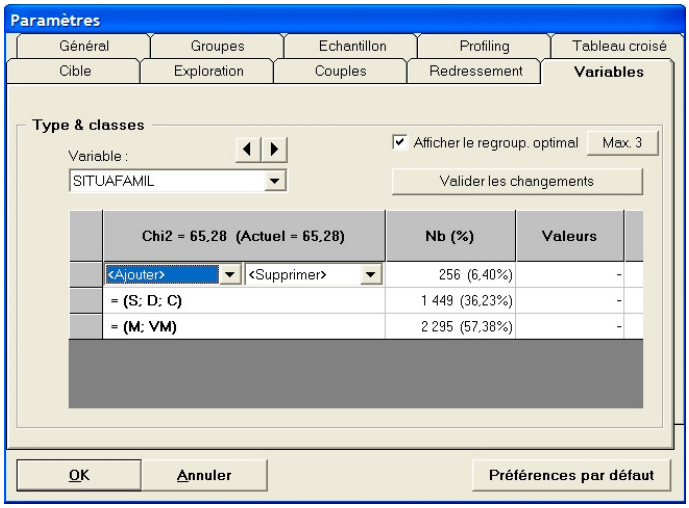
Modification de regroupements de modalités optimaux

Les regroupements optimaux peuvent être modifiées par l'utilisateur – mais pas le nombre de classes, à moins de relancer la phase d'exploration, après modification des paramètres de discrétisation.

La modification est réalisée à partir de la fenêtre Type & classes > onglet [Variables], puis en cochant la case 

.

La modification des regroupements se fait dans la fenêtre de paramétrage Type & classes > onglet [Variables]. Il suffit de cliquer sur une classe à modifier pour ajouter ou supprimer une modalité.

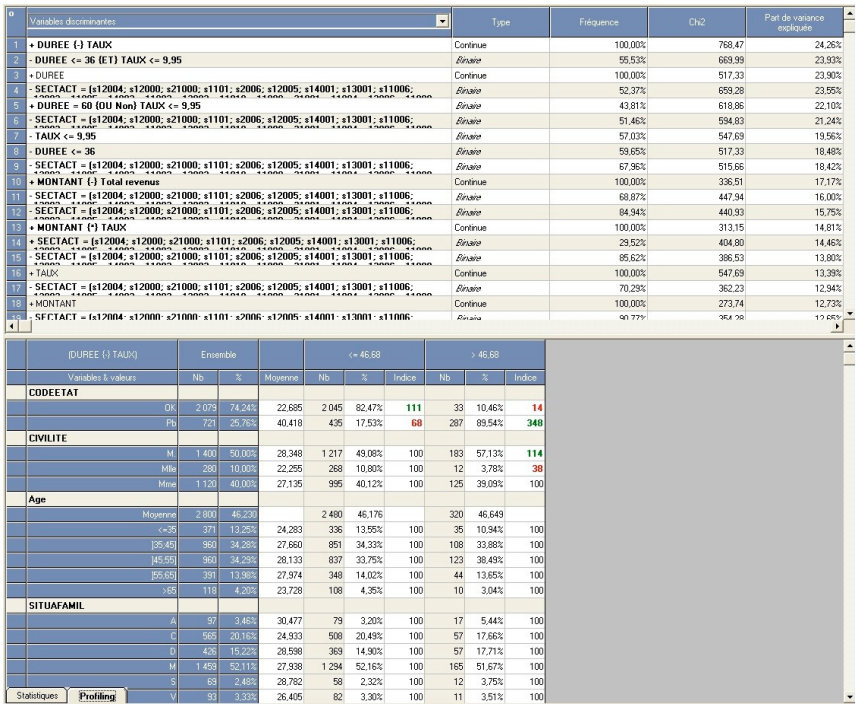


c) Discrétisation des variables issues de la phase d'exploration

A l'issue de la phase d'exploration, un ensemble de variables transformées et combinées sont disponibles (Exploration > Résultats > Variables discriminantes).

La fonctionnalité profiling (onglet [Profiling] de la feuille de résultat de la zone inférieure) permet de croiser n'importe laquelle de ces variables avec l'ensemble des variables de base. Par défaut, les variables continues sont discrétisées en 2 ou 3 classes (discrétisation optimale).

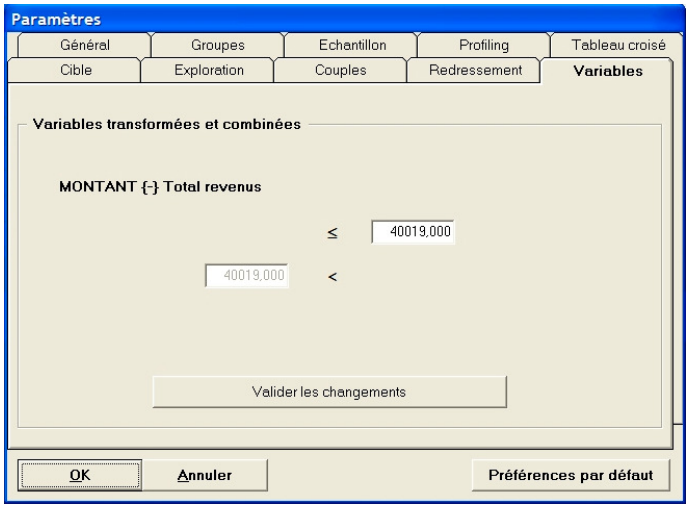
Il est possible de modifier cette discrétisation en double-cliquant sur la variable concernée dans la liste des variables discriminantes affichée dans la zone de résultat supérieure. La fenêtre affichée permet la modification du seuil.



	Variables discriminantes	Type	Fréquence	Chi2	Part de variance expliquée
1	+ DUREE (-) TAUX	Continue	100,00%	768,47	24,26%
2	- DUREE <= 36 (ET) TAUX <= 9,95	*Binaire*	55,53%	669,99	23,93%
3	+ DUREE	Continue	100,00%	517,33	23,90%
4	- SECTACT = {s12004; s12000; s21000; s1101; s2006; s12005; s14001; s13001; s11006;	*Binaire*	52,37%	659,28	23,55%
5	+ DUREE = 60 (OU Non) TAUX <= 9,95	*Binaire*	43,81%	618,86	22,10%
6	- SECTACT = {s12004; s12000; s21000; s1101; s2006; s12005; s14001; s13001; s11006;	*Binaire*	51,46%	594,83	21,24%
7	- TAUX <= 9,95	*Binaire*	57,03%	547,69	19,56%
8	- DUREE <= 36	*Binaire*	59,65%	517,33	18,48%
9	- SECTACT = {s12004; s12000; s21000; s1101; s2006; s12005; s14001; s13001; s11006;	*Binaire*	67,96%	515,66	18,42%
10	+ MONTANT (-) Total revenus	Continue	100,00%	336,51	17,17%
11	- SECTACT = {s12004; s12000; s21000; s1101; s2006; s12005; s14001; s13001; s11006;	*Binaire*	68,87%	447,94	16,00%
12	- SECTACT = {s12004; s12000; s21000; s1101; s2006; s12005; s14001; s13001; s11006;	*Binaire*	84,94%	440,93	15,75%
13	+ MONTANT (*) TAUX	Continue	100,00%	313,15	14,81%
14	+ SECTACT = {s12004; s12000; s21000; s1101; s2006; s12005; s14001; s13001; s11006;	*Binaire*	29,52%	404,80	14,46%
15	- SECTACT = {s12004; s12000; s21000; s1101; s2006; s12005; s14001; s13001; s11006;	*Binaire*	85,62%	386,53	13,80%
16	+ TAUX	Continue	100,00%	547,69	13,39%
17	- SECTACT = {s12004; s12000; s21000; s1101; s2006; s12005; s14001; s13001; s11006;	*Binaire*	70,29%	362,23	12,94%
18	+ MONTANT	Continue	100,00%	273,74	12,73%

(DUREE (-) TAUX)	Ensemble			<= 46,68			> 46,68		
Variables \ Valeurs	Nb	%	Moyenne	Nb	%	Indice	Nb	%	Indice
CODEETAT									
OK	2 079	74,24%	22,685	2 045	82,47%	111	33	10,46%	14
Pb	721	25,76%	40,418	435	17,53%	68	287	89,54%	348
CIVILITE									
M.	1 400	50,00%	28,348	1 217	49,08%	100	183	57,13%	114
Mlle	280	10,00%	22,255	268	10,80%	100	12	3,78%	38
Mme	1 120	40,00%	27,135	995	40,12%	100	125	39,09%	100
Age									
Moyenne	2 800	46,230		2 480	46,176		320	46,649	
<=35	371	13,25%	24,283	336	13,55%	100	35	10,94%	100
]35,45]	960	34,28%	27,660	851	34,33%	100	108	33,88%	100
]45,55]	960	34,29%	28,133	837	33,75%	100	123	38,49%	100
]55,65]	391	13,96%	27,974	348	14,02%	100	44	13,65%	100
>65	118	4,20%	23,728	108	4,35%	100	10	3,04%	100
SITUAFAMIL									
A	97	3,46%	30,477	79	3,20%	100	17	5,44%	100
C	565	20,16%	24,933	508	20,49%	100	57	17,66%	100
D	426	15,22%	28,598	369	14,90%	100	57	17,71%	100
M	1 459	52,11%	27,938	1 294	52,16%	100	165	51,67%	100
S	69	2,46%	28,782	58	2,32%	100	12	3,75%	100
V	93	3,33%	26,405	82	3,30%	100	11	3,51%	100

Statistiques | Profiling



7. 7 Les groupes

1. Objet

La fonctionnalité « groupes » de DataLab permet d'étendre considérablement le champ du module d'exploration, en donnant la possibilité à l'utilisateur de combiner un plus grand nombre de variables qu'avec les combinaisons.

De plus les variables d'un groupe sont des variables de même nature, sur lesquelles les traitements sont d'habitude effectués manuellement par l'utilisateur.

C'est donc une fonctionnalité qui permet des gains de temps considérables.

Cette fiche technique détaille :

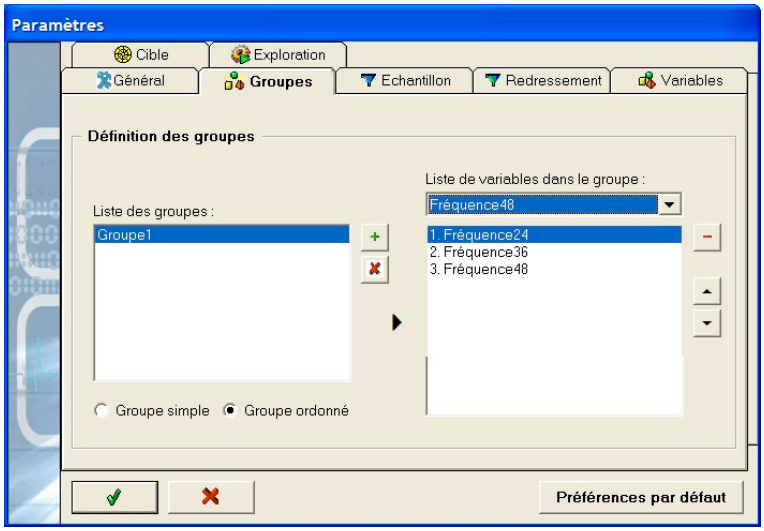
- La définition des groupes
- Le paramétrage des traitements
- Les différents traitements effectués lors de l'étape d'exploration

2. Détail

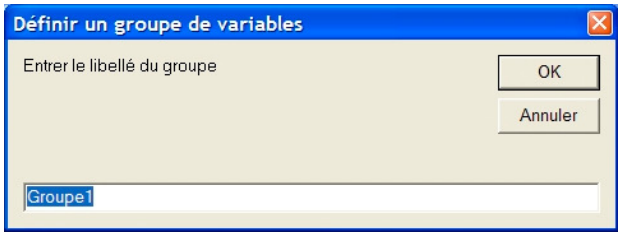
2.1 Définition des groupes

Les groupes se définissent en sélectionnant **Préparer les données > Groupes de variables**

On crée un groupe en cliquant sur  « ajouter un groupe » :



On définit le nom de groupe dans la fenêtre suivante :



On définit ensuite les variables constitutives du groupe. Celles-ci peuvent être continues ou catégoriques. Dans ce dernier cas, l'utilisateur choisit la modalité de la variable catégorique.

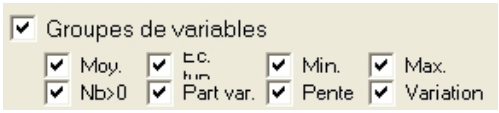
Lorsque toutes les variables du groupe ont été choisies, l'utilisateur peut spécifier si le groupe est ordonné, et définir l'ordre des variables grâce aux boutons  et 

Exemple : Si le groupe est constitué de 4 variables représentant les Chiffres d'affaires des 4 trimestres de l'année, le groupe est ordonné (ordre : CA_trim1, CA_trim2, CA_trim3, CA_trim4)

2.2 Paramétrage des traitements

Un ensemble de traitements est effectué sur les variables du groupe pour créer de nouvelles variables à partir des variables du groupe : calcul de la moyenne, du minimum et du maximum, de l'écart-type, de la pente, de la part de chaque variable dans le groupe, de la pente et de la variation des 2 dernières variables du groupe dans le cas d'un groupe ordonné.

Le paramétrage (choix des variables créées) s'effectue à partir de l'arborescence Exploration > Démarrer > Options détaillées, dans la feuille de paramétrage :



- « Groupes de variables » : autorise l'évaluation des variables issues des groupes de variables (lorsqu'ils ont été définis)
- « Moy » : calcule la variable Moyenne des variables composant le groupe.
  La variable Moyenne est calculée hors valeurs manquantes.
  Si pour un enregistrement donné, toutes les variables du groupe comportent des valeurs manquantes, la variable Moyenne prend la valeur « moyenne de la moyenne des variables composant le groupe ».
- « Ec. Typ. » : calcule la variable Ecart type des variables composant le groupe
  Dans le calcul de l'écart-type, la moyenne utilisée est la valeur « moyenne de la moyenne des variables composant le groupe ».
  La variable Ecart type est calculée hors valeurs manquantes.
  Si pour un enregistrement donné, toutes les variables du groupe comportent des valeurs manquantes, la variable Ecart type prend la valeur « 0 »
- « Min » : calcule la variable Minimum des variables composant le groupe
  La variable est calculée hors valeurs manquantes
- « Max » : calcule la variable Maximum des variables composant le groupe
  La variable est calculée hors valeurs manquantes
- « Nb>0 » : calcule une variable donnant le Nombre de valeurs strictement positives parmi les variables composant le groupe
  La variable est calculée hors valeurs manquantes.
  Si pour un enregistrement donné, toutes les variables du groupe comportent des valeurs manquantes, la variable Nb>0 prend la valeur « 0 »

- « Part var. » : calcule pour chaque variable composant le groupe sa proportion par rapport à la somme des variables du groupe, (Part(i) étant définie comme le ratio de la variable (i) sur la somme de toutes les variables Var(i)).

  Si une variable i est manquante, sa valeur est remplacée par Moyenne de la variable.

  La même règle est appliquée lorsque toutes les variables du groupe comportent des valeurs manquantes.
- « Pente » : calcule la variable Pente des variables composant le groupe.

  La variable est calculée hors valeurs manquantes.

$$\text{Pente } (X_i) = [\Sigma(X_i\text{-moy}(X_i))(i\text{-moy}(i))]/\ \Sigma((i\text{-moy}(i))^2$$

  Où i = le numéro d'ordre de la variable dans le groupe.

  Si pour un enregistrement donné, toutes les variables du groupe comportent des valeurs manquantes, la variable Pente prend la valeur « 0 »
- « Variation » : calcule la variable Variation des 2 dernières variables du groupe

$$\text{Var } (X_n, X_{n-1}) = (X_n\text{-}X_{n-1})/\ [(X_n\text{+}X_{n-1})/\ 2]$$

  Si une ou des deux variables est manquante, sa valeur est remplacée par la Moyenne de la variable.

  La même règle est appliquée lorsque les deux variables du groupe comportent des valeurs manquantes.

Note *: Les variables Pente et Variation ne sont calculées que pour les groupes ordonnés.*

7. 8 L'export

1. Objet

A l'issue de la modélisation, différents éléments peuvent être exportés :

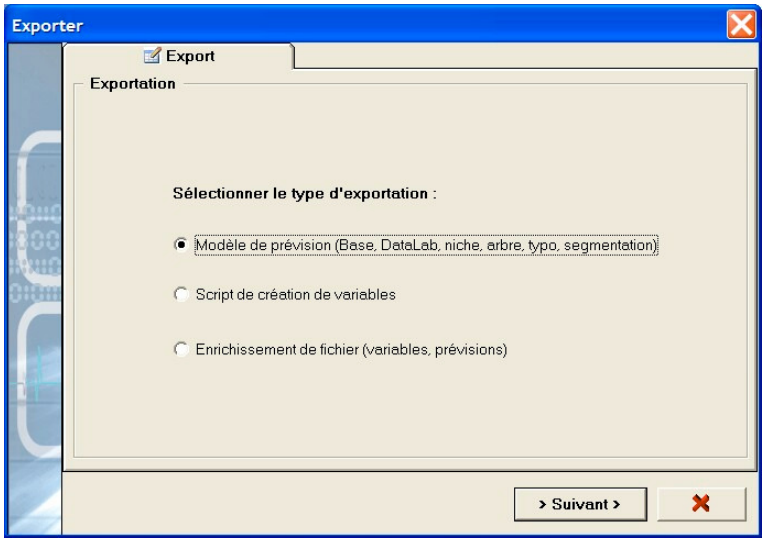
- le code des modèles
- le script de création des variables
- la valeur des variables et / ou prévisions pour un fichier donné

La fonctionnalité Export permet, en particulier, de générer du code SAS ou SPSS, directement utilisable.
Cette fiche technique détaille les différentes options possibles dans chaque catégorie d'exportation.

2. Détail

La fonctionnalité Export est accessible par le menu Fichier / Exporter…
La fenêtre Exportation permet de choisir le type d'export :

- code des modèles
- script de création des variables
- valeur des variables et / ou des prévisions pour un fichier donné



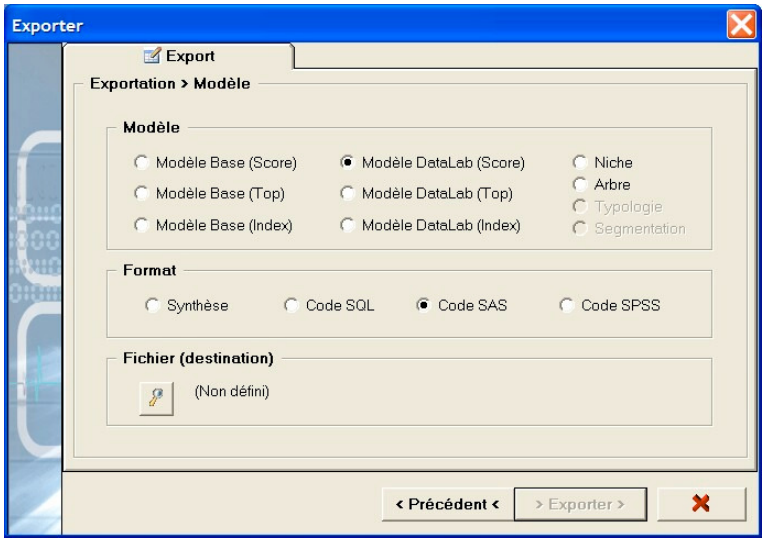
2.1 Export des modèles de prévision

Etape 1

Dans la fenêtre Exportation, sélectionner l'option « Modèle de prévision » et cliquer sur « Suivant ».

Etape 2

Cette étape concerne l'ensemble du paramétrage de l'export :



La fenêtre de paramétrage de l'export permet de sélectionner le type de modèle à exporter :

- modèle de prévision de Base final : élaboré avec les seules variables d'entrée et éventuellement modifié par l'utilisateur
- modèle de prévision DataLab final : élaboré avec les seules variables d'entrée et éventuellement modifié par l'utilisateur, sous forme de score, top, ou tranches de gain chart
- niche : union ou intersection de segments construite par l'utilisateur
- arbre de décision
- typologie
- segmentation

le format :

- synthèse : code en langage clair
- Code SQL : code de la requête au format SQL
- code SAS : code au format SAS
- code SPSS : code au format SPSS

ainsi que le nom du fichier qui contiendra le code, en cliquant sur le bouton […].
Ce fichier est au format texte. Le code SAS (resp. SPSS), SQL peut être directement copié dans un programme SAS (resp. SPSS, SQL).

Le code exporté comporte :

- les traitements préalables (le cas échéant) :
  - remplacement des valeurs manquantes (par la moyenne, médiane ou valeur choisie par l'utilisateur)
  - écrêtage
- la création des variables du modèle
- l'équation du modèle

Etape 3

Réaliser l'export en cliquant sur « Exporter » .

Une fois l'export réalisé, DataLab vous propose de visualiser le fichier.

```
/* DataLab 5.03 */
/* Intitulé du projet : Risque créditv5 */
/* Date du projet : 27 août 2005 */
/* Modèle DataLab (Régression logistique) */

DATA ALL;
INFILE 'Filename.dat' MISSOVER;

/* ECHANTILLON UTILISE POUR BATIR LE MODELE */
/*      Filtres (Exclusions) : Aucun */
/*      Echantillon d'apprentissage : Périodique 7/10 */
/*      Redressement : Non */

DATA ALL;
INFILE 'Filename.dat' MISSOVER;

/* TRAITEMENTS PREALABLES */
IF MONTANT >  155235.228 THEN MONTANT =  155235.228;
IF Total charges >  13565.397 THEN Total charges =  13565.397;
IF Total revenus >  41592.343 THEN Total revenus =  41592.343;

/* CREATION DES VARIABLES */
IF (CODEOBJET IN (o72, o80, o0, o73, o, o42, o61, o5, o2, o30, o43, o11, o20, o41, o40, o70, o32, o60, o13, o53, o9, o12, o31, o90,
ELSE Field1 = 0;
Field2 = MONTANT - Total revenus;
IF SECTACT IN (s2006, s12000, s21000, s1101, s21001, s11005, s14001, s14003, s11003, s13001, s14002, s11009, s13002, s11004, s12002,
ELSE Field3 = 0;
Field4 = DUREE;
IF CODEHABITAT IN (Ac, LF, P) THEN Field5 = 1;
ELSE Field5 = 0;
IF CODEOBJET = o9 THEN Field6 = 1;
ELSE Field6 = 0;
Field7 = (NBPACHARGE)**2;
Field8 = Total charges;
IF CIVILITE = Mlle THEN Field9 = 1;
ELSE Field9 = 0;

/* MODELE */
Prev = Field1 * (-1.188568)
 + Field2 * ( .000016)
 + Field3 * (-2.529033)
 + Field4 * ( .074736)
 + Field5 * (-1.012013)
 + Field6 * (-1.073884)
 + Field7 * ( .076961)
 + Field8 * (-.000202)
 + Field9 * (-.816381)
 + (-.842303);

Prev = 1 - 1 / (1 + Exp(Prev));

Run;
```

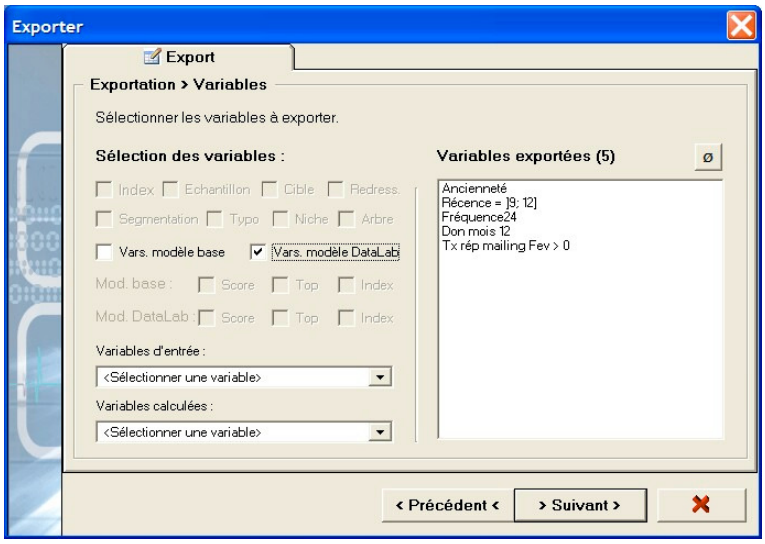
2.2 Export du script de création des variables

Etape 1

Dans la fenêtre Exportation, sélectionner l'option « Script de création des variables» et cliquer sur « Suivant ».

Etape 2

Cette étape concerne la sélection des variables à créer :



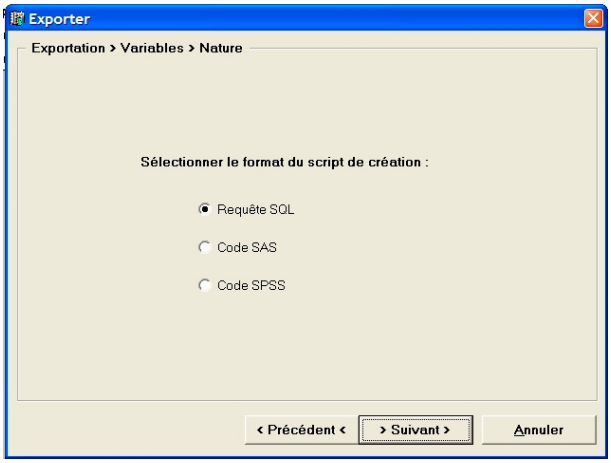
La fenêtre de sélection des variables permet de choisir :

- l'ensemble des variables du modèle de base (sélection base), et / ou
- l'ensemble des variables du modèle DataLab (sélection DataLab)
- des variables complémentaires (variables d'entrée ou variables calculées) à choisir dans les listes déroulantes

Etape 3

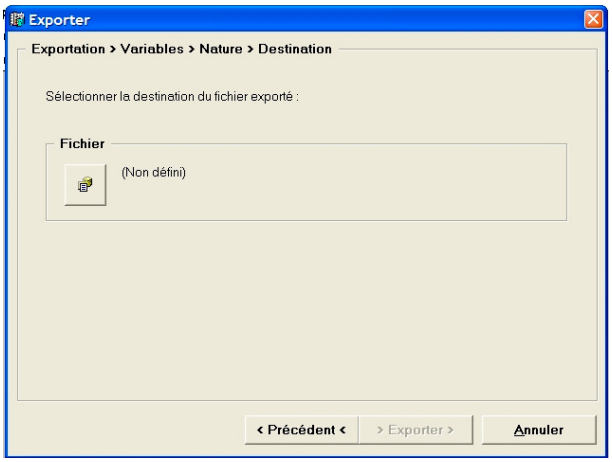
En cliquant sur le bouton « Suivant » on accède à la fenêtre de paramétrage du format de script :

- requête SQL
- code SAS
- code SPSS



Etape 4

La fenêtre suivante permet de choisir le fichier vers lequel exporter le script, en cliquant sur le bouton  :



Etape 5

Réaliser l'export en cliquant sur « Exporter ».

Une fois l'export réalisé, DataLab vous propose de visualiser le fichier.

```
/* DataLab 5.03 */
/* Intitulé du projet : Risque créditv5 */
/* Date du projet : 27 août 2005 */

CREATE TABLE Table_Destination (
'CIVILITE_EQ_Mlle' INTEGER NOT NULL,
'Carré[NBPACHARGE]' DECIMAL(8,2) NOT NULL,
'SECTACT_EQ_(s2006;_s12000;_s21000;_s1101;_s21001;_s11005;_s14001;_s14003;_s11003;_s13001;_s14002;_s11009;_s13002
'CODEHABITAT_EQ_(Ac;_LF;_P)' INTEGER NOT NULL,
'DUREE' DECIMAL(8,2) NOT NULL,
'CODEOBJET_EQ_o9' INTEGER NOT NULL,
'Total_charges' DECIMAL(8,2) NOT NULL,
'MONTANT_-_Total_revenus' DECIMAL(8,2) NOT NULL,
'CODEOBJET_EQ_(o72;_o80;_o0;_o73;_o;_o42;_o61;_o5;_o2;_o30;_o43;_o11;_o20;_o41;_o40;_o70;_o32;_o60;_o13;_o53;_o9;.
)

INSERT INTO Table_Destination (
'CIVILITE_EQ_Mlle',
'Carré[NBPACHARGE]',
'SECTACT_EQ_(s2006;_s12000;_s21000;_s1101;_s21001;_s11005;_s14001;_s14003;_s11003;_s13001;_s14002;_s11009;_s13002
'CODEHABITAT_EQ_(Ac;_LF;_P)',
'DUREE',
'CODEOBJET_EQ_o9',
'Total_charges',
'MONTANT_-_Total_revenus',
'CODEOBJET_EQ_(o72;_o80;_o0;_o73;_o;_o42;_o61;_o5;_o2;_o30;_o43;_o11;_o20;_o41;_o40;_o70;_o32;_o60;_o13;_o53;_o9;.
)

SELECT
(CASE WHEN 'CIVILITE' = 'Mlle' THEN 1 ELSE 0 END),
SQUARE('NBPACHARGE'),
(CASE WHEN 'SECTACT' = '(s2006; s12000; s21000; s1101; s21001; s11005; s14001; s14003; s11003; s13001; s14002; s11
(CASE WHEN 'CODEHABITAT' = '(Ac; LF; P)' THEN 1 ELSE 0 END),
'DUREE',
(CASE WHEN 'CODEOBJET' = 'o9' THEN 1 ELSE 0 END),
'Total charges',
'MONTANT' - 'Total revenus',
(CASE WHEN ('CODEOBJET' = '(o72; o80; o0; o73; o; o42; o61; o5; o2; o30; o43; o11; o20; o41; o40; o70; o32; o60;
FROM 'Table_Origine'
```

2.3 Enrichissement de fichier

Etape 1

Dans la fenêtre Exportation, sélectionner l'option « Enrichissement de fichier» et cliquer sur « Suivant ».

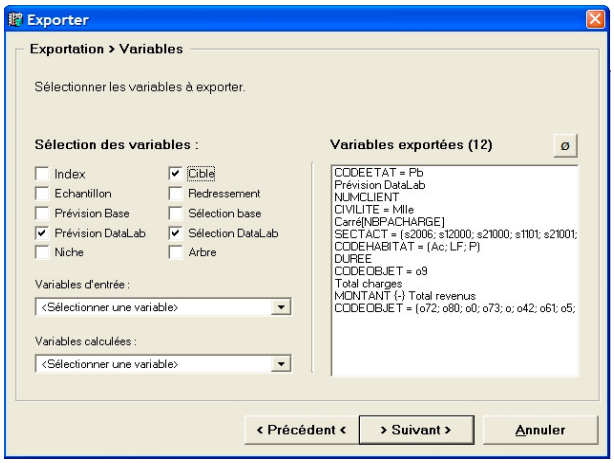
Etape 2

Il est possible d'enrichir un fichier ou une table de base de données externe avec les variables ou les prévisions calculées à partir des enregistrements de ce fichier ou de cette source de données externe.

L'enrichissement concerne :

- Cible
- Echantillon
- Poids issus d'un redressement éventuel
- Prévisions issues des modèles de base et DataLab
- Variables de la sélection de base
- Variables de la sélection DataLab
- Tout variable d'entrée
- Tout variable calculée

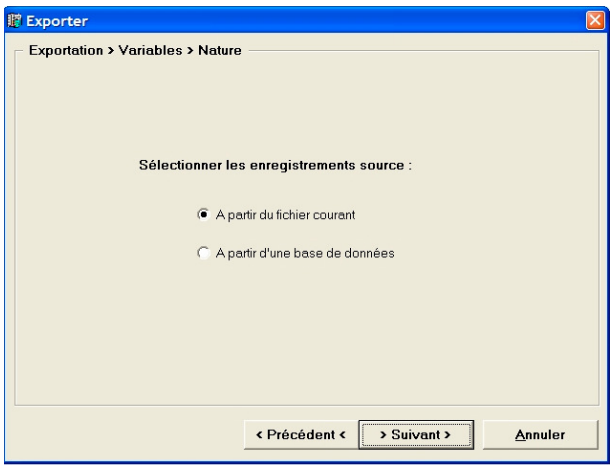
Le paramétrage se fait dans la fenêtre de paramétrage de l'enrichissement en cochant les cases appropriées.



Etape 3

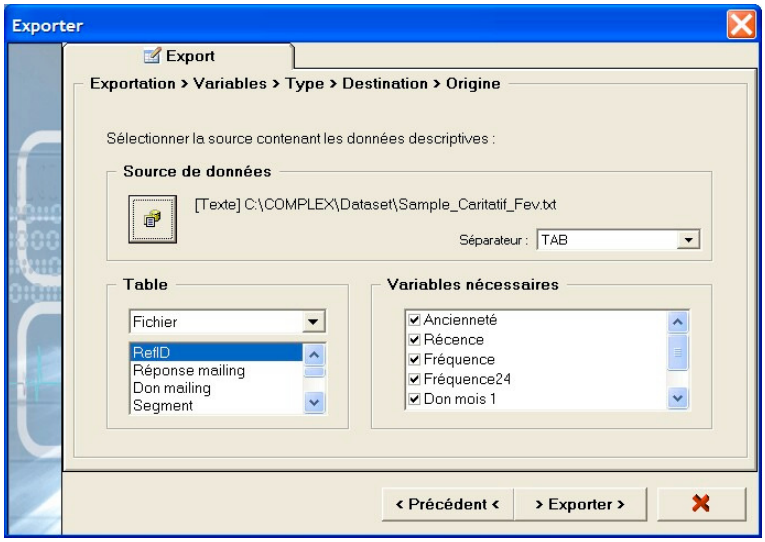
En cliquant sur « Suivant », on accède à la fenêtre de sélection du fichier contenant les données à partir desquelles sera calculée la valeur des variables et prévisions :

- Fichier déjà chargé dans DataLab
- Base de données externe



Etape 4

Sélection du fichier destination, contenant les éléments calculés, en cliquant sur le bouton  . On définit le fichier de destination de la même manière qu'on accède à une source de données. Le fichier de destination peut donc être un fichier texte, une table d'une base de données et est accessible directement ou via un pilote ODBC ou un fournisseur OLE DB .



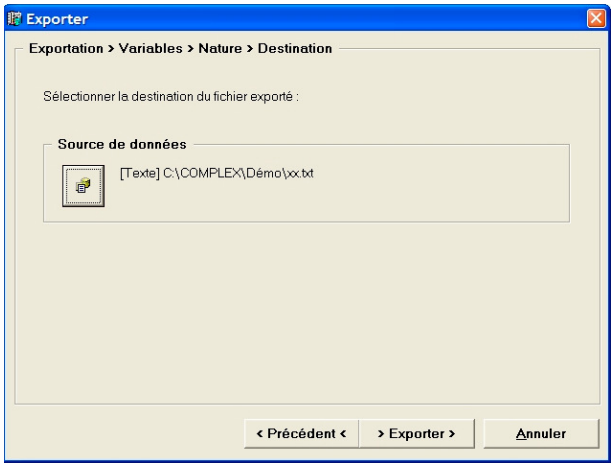
L'étape suivante est différente selon que le fichier source (à partir duquel sera calculée la valeur des variables et prévisions) est :

- Le fichier courant : fichier déjà chargé dans DataLab > étape 5a
- Une base de données > étape 5b

Etape 5a

Le calcul des valeurs des variables et prévisions s'effectue à partir du fichier déjà chargé.

Sélectionner le fichier destination, qui contiendra les éléments calculés, en cliquant sur  , puis cliquer sur « Exporter ».

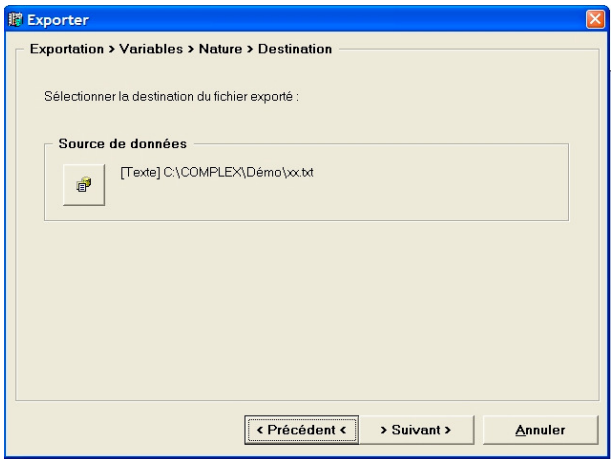


DataLab procède alors à la création du fichier (ou de la table) enrichi. Le fichier ou la table contiennent une ligne d'en-tête pour le nom des champs créés.

Etape 5b

Le calcul des valeurs des variables et prévisions s'effectue à partir d'une base de données, qu'il faudra définir.

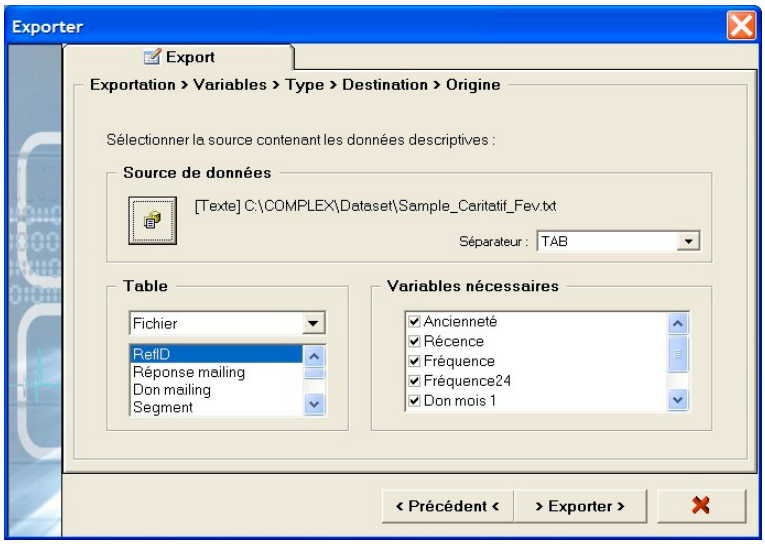
Sélectionner tout d'abord le fichier destination, qui contiendra les éléments calculés, en cliquant sur  (la procédure est la même que pour l'étape 5a), puis cliquer sur « Suivant » pour sélectionner la base de données.



La fenêtre permet de sélectionner

- la base de données
- la table dans laquelle se trouvent les champs nécessaires au calcul des valeurs des variables et prévisions

Les variables nécessaires sont listées dans le cadre inférieur droit, elles sont automatiquement cochées si elles ont été retrouvées dans la table (cadre inférieur gauche). Les champs doivent porter le même nom dans la table de la base et dans le fichier export.



Cliquer ensuite sur « Exporter ».

DataLab procède alors à la création du fichier (ou de la table) enrichi. Le fichier ou la table contiennent une ligne d'en-tête pour le nom des champs créés.

7. 9 La typologie

1. Objet

La partie Description de DataLab offre un module de typologie. Cette note a pour objet de décrire son utilisation. Les points suivants sont abordés :

- Paramètres
- Analyse des facteurs
- Description des groupes

2. Généralités

La typologie est de type Kmeans sur facteurs principaux issus d'une analyse en composantes principales. Les variables utilisées pour définir ces axes sont choisies par l'utilisateur. Les variables catégoriques sont automatiquement binarisées (optionnellement, en classes, dans le cas de variables numériques).

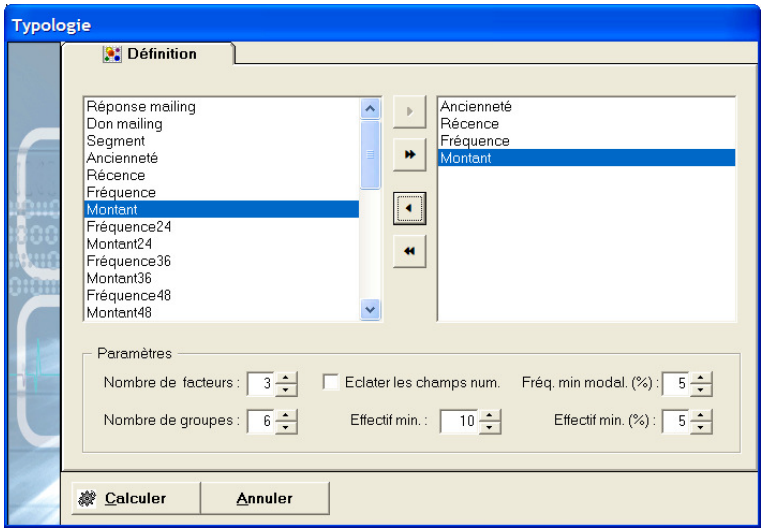
La typologie est réalisée sur l'échantillon d'apprentissage si celui-ci a été défini, sur la totalité dans le cas contraire.

Les groupes sont initialisés à partir des valeurs d'enregistrements choisis aléatoirement. L'initialisation est répétée un grand nombre de fois et la typologie présentant l'inertie inter-groupes cumulée la plus grande est sélectionnée.

Les résultats sont présentés ici avec l'exemple livré 'Sample_Caritatif_Fex.txt' (données issues d'un historique de dons).

3. Définir et calculer la typologie

La fonctionnalité **Typologie** est accessible par l'arborescence via **Description > Typologie > Définir ....** La fenêtre Typologie permet de définir ou valider les paramètres de la typologie.



▸ fenêtre Typologie

Pour définir une typologie :

a. Sélectionner les variables à intégrer dans la typologie. Ces variables doivent être autant que possible différentes :

- Pour ajouter une variable, la sélectionner dans la liste de gauche et cliquer sur le bouton ▸
- Pour ajouter toutes les variables, cliquer sur le bouton ▸▸
- Pour supprimer une variable, la sélectionner dans la liste de droite et cliquer sur le bouton ◂
- Pour supprimer toutes les variables sélectionnées, cliquer sur le bouton ◂◂

b. Définir ou valider les paramètres :

- Nombre de facteurs : Nombre de facteurs de l'analyse en composantes principales. La typologie est réalisée sur ces facteurs et non directement sur les variables
- Eclater les champs num. : Les variables numériques peuvent être utilisées telles quelles (dans ce cas les valeurs manquantes sont remplacées par la valeur moyenne de la variable) ou bien discrétisées (dans ce cas la discrétisation utilisée est la discrétisation descriptive accessible par l'arborescence Description > Définir > Variables…)
- Freq. Min. modal. (%) : Les variables catégoriques binarisées (ou discrétisées pour les variables numériques) sont automatiquement filtrées selon un seuil de fréquence minimal
- Nombre de groupes : Nombre de groupes de la typologie. Le nombre final effectif de groupes peut être inférieur à cette valeur afin de respecter les paramètres relatifs à la taille des groupes
- Effectif min. : Nombre minimum d'enregistrements dans chaque groupe
- Effectif min. (%) : Nombre relatif minimum d'enregistrements dans chaque groupe

c. Générer la typologie en cliquant sur le bouton Calculer. A l'issue du calcul de la typologie les résultats suivants sont disponibles :

- Facteurs : Statistiques sur les facteurs résultant de l'analyse factorielle
- Groupes : Statistiques sur les groupes
- Description des groupes : Statistiques complémentaires sur les groupes

4. Analyse des facteurs

Le résultat **Facteurs** est accessible par l'arborescence via **Description > Typologie > Facteurs**. La fenêtre suivante s'affiche qui décrit chaque facteur selon les variables constitutives de la typologie.

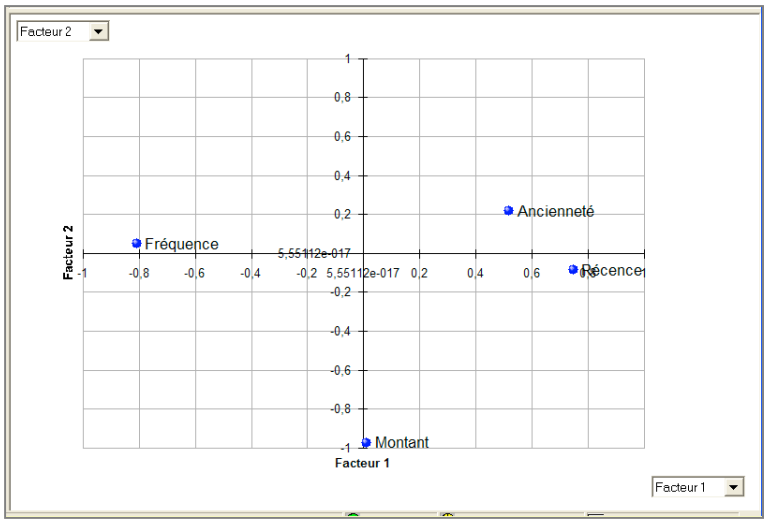
		Facteur 1	Facteur 2	Facteur 3
	Inertie (%)	37,02%	25,15%	22,53%
	Inertie cumulée (%)	37,02%	62,17%	84,69%
	(Corrélations)			
	Ancienneté	**0,52**	0,22	**0,81**
	Récence	**0,75**	-0,09	-0,42
	Fréquence	**-0,81**	0,05	0,14
	Montant	0,01	**-0,97**	0,22

résultat Facteurs

Pour chaque facteur, on dispose des statistiques suivantes :

- Inertie (%) : pourcentage d'inertie expliquée
- Inertie cumulée (%) : pourcentage cumulé d'inertie expliquée
- Corrélations : pour chaque variable coefficient de corrélation

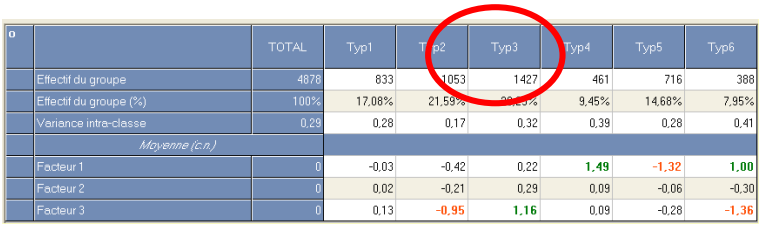
Les valeurs significatives sont colorées en rouge (sous-représentation) ou en vert (sur-représentation).
Ici, le facteur 2 est corrélé négativement presque totalement avec la variable Montant (-0.97) et le facteur 3 corrélé positivement avec l'Ancienneté (+0.81).
Le graphe des corrélations est automatiquement affiché. L'utilisateur peut choisir dans les listes correspondances les facteurs (axes) du plan qu'il souhaite afficher.
Ici on observe l'opposition entre récence et fréquence, c'est à dire plus la récence est importante et moins la fréquence de dons l'est.



- Plan des corrélations Facteurs / Variables

5. Analyse des groupes

Le résultat **Groupes** est accessible par l'arborescence via **Description > Typologie > Groupes**. La fenêtre suivante s'affiche qui décrit chaque groupe selon les facteurs.



	TOTAL	Typ1	Typ2	Typ3	Typ4	Typ5	Typ6
Effectif du groupe	4878	833	1053	1427	461	716	388
Effectif du groupe (%)	100%	17,08%	21,59%	[illegible]	9,45%	14,68%	7,95%
Variance intra-classe	0,29	0,28	0,17	0,32	0,39	0,28	0,41
Moyenne (c.n.)							
Facteur 1	0	-0,03	-0,42	0,22	**1,49**	**-1,32**	**1,00**
Facteur 2	0	0,02	-0,21	0,29	0,09	-0,06	-0,30
Facteur 3	0	0,13	**-0,95**	**1,16**	0,09	-0,28	**-1,36**

résultat Groupes

Chaque groupe peut être renommé en cliquant sur le nom du groupe et en entrant un nouveau libellé.

Pour chaque groupe, on dispose des statistiques suivantes :

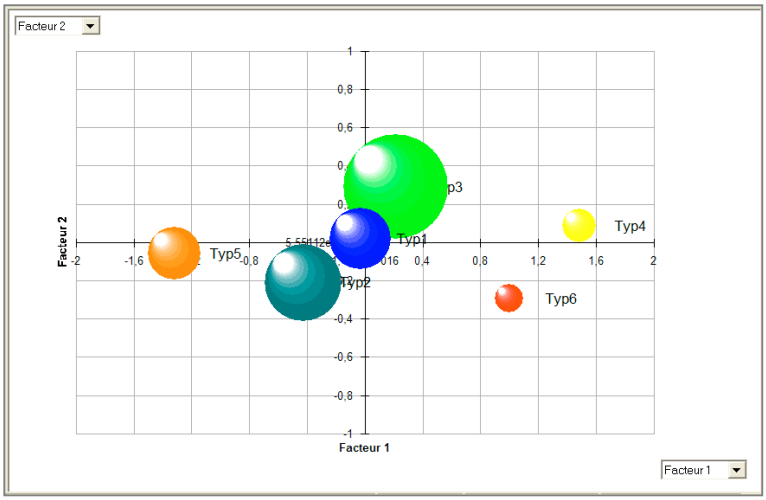
- Effectif : Nombre d'enregistrements
- Effectif (%) : pourcentage d'enregistrements
- Variance intra-classe : variance à l'intérieure de chaque groupe
- Moyenne (c.n.) : moyenne centrée normée de chaque facteur (0 pour le total)

Les valeurs significatives sont colorées en rouge (sous-représentation) ou en vert (sur-représentation).

Ici, le groupe Typ3 présente une valeur moyenne de 1.16 pour le facteur 3.

Les groupes sont représentés automatiquement sur un graphique. L'utilisateur peut choisir dans les listes correspondances les facteurs (axes) du plan qu'il souhaite afficher.

Ici on observe l'opposition entre les groupes Typ4 et Typ5 sur le facteur 1.



- représentation graphique des groupes

6. Description des groupes

Le résultat **Description des groupes** est accessible par l'arborescence via **Description > Typologie > Description des groupes**. La fenêtre suivante s'affiche qui décrit chaque groupe selon les facteurs.

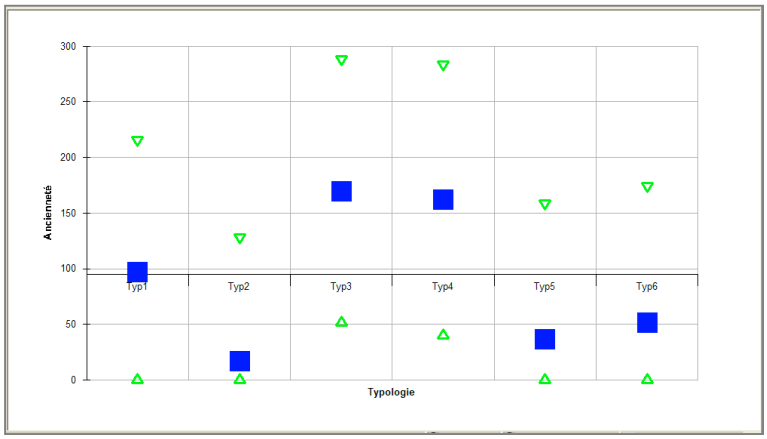
	Typ1		Typ2		Typ3		Typ4		Typ5		Typ6	
	Moy.	Ec.Typ.	Moy.	Ec.Typ.	Moy.	Ec.Typ.	Moy.	Ec.Typ.	Moy.	Ec.Typ.	Moy.	Ec.Typ.
Ancienneté	96,87	119,20	**17,00**	111,44	**169,74**	118,38	**161,91**	121,76	**36,33**	122,31	**51,47**	122,57
Récence	6,62	14,92	7,68	14,97	**4,75**	14,00	**25,94**	17,81	**2,78**	12,14	**26,06**	16,21
Fréquence	1,02	1,42	1,05	1,34	1,07	1,50	**0,67**	1,42	**1,86**	1,44	**0,60**	1,36
Montant	36,56	77,05	30,75	70,90	30,76	61,29	27,39	60,87	33,24	64,10	34,49	82,53

▸ résultat Description des groupes

Pour chaque groupe, on dispose des statistiques suivantes selon chaque variable constitutive de la typologie :

- Moyenne
- Ecart-type

Les valeurs significatives sont colorées en rouge (sous-représentation) ou en vert (sur-représentation).

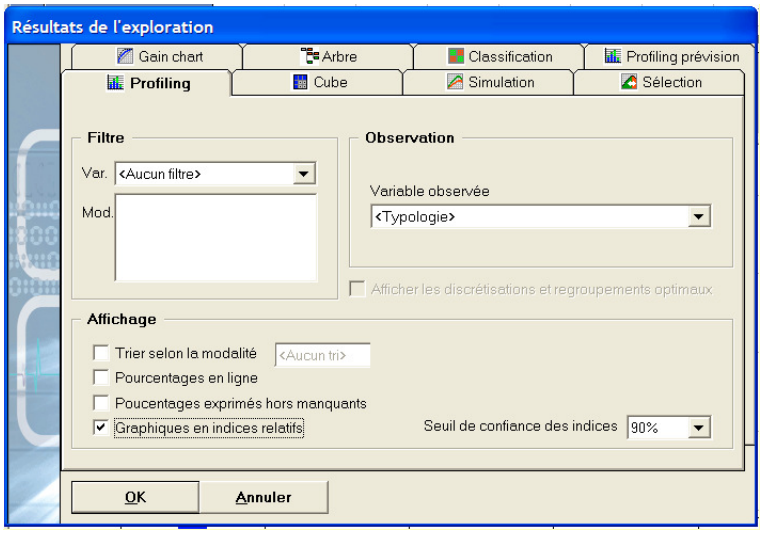


▸ résultat Description des groupes

Ici, le groupe Typ2 présente une valeur moyenne d'ancienneté égale à 17 mois, respectivement 169 pour le groupe Typ3.

Les statistiques sont représentés automatiquement sur un graphique. Les variables à afficher sont sélectionnées par l'utilisateur en cliquant sur la ligne de la table.
Ici on observe la très faible valeur moyenne de l'ancienneté au sein du groupe Typ2.

Les groupes de la typologies peuvent aussi classiquement être décrits plus complètement par la fonctionnalité **Profiling** de DataLab via **Statistiques > Profiling**. Dans la fenêtre **Profiling** sélectionner la variable automatiquement créée **<Typologie>** puis cliquer sur OK. Le profiling est automatiquement calculé sur toutes les variables et affiché (Cf. info_technique_N3.pdf).



Profiling de la typologie

7. 10 Le module DataBuilder

1. Objet

Le module DataBuilder de DataLab permet de constituer simplement le fichier d'analyse à mettre en entrée de DataLab à partir de une ou plusieurs tables de base de données. Cette note a pour objet de décrire son utilisation. Les points suivants sont abordés :

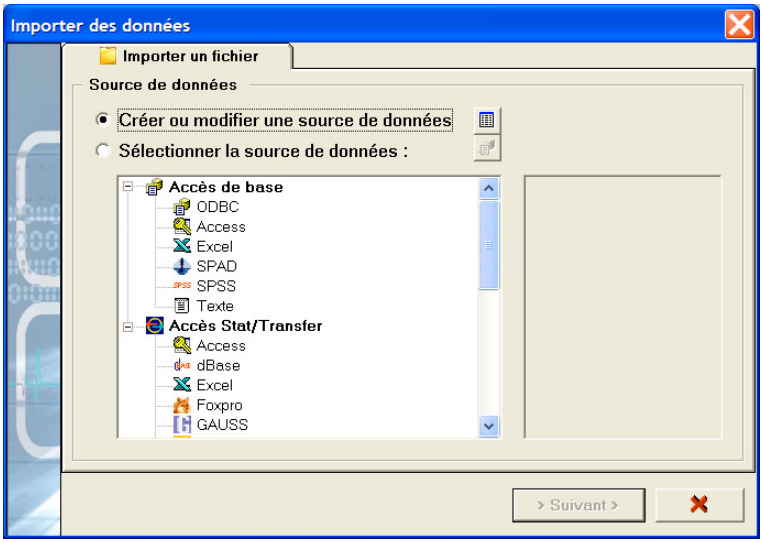
- Connexion à une source de données
- Définition des relations entre les tables
- Constitution de la table d'analyse
  - Création de la table
  - Création de la date de référence
  - Ajout d'un champ
  - Autres fonctions
    - Supprimer un champ de la table d'analyse
    - Mettre un ou plusieurs champs en attente d'ajout
    - Mettre à jour la table d'analyse
    - Importer des requêtes existantes
    - Visualiser des statistiques sur la table d'analyse en cours
    - Gestion du projet et de la source de données
- Utilisation de la table par DataLab

2. Généralités

Le module DataBuilder de DataLab permet de constituer facilement et rapidement la table d'analyse (matrice individus x variables) à mettre en entrée de DataLab.

Pour lancer le module à partir de DataLab :

- arborescence **Projet > Importer des données …**
- Cliquer sur l'option **Créer ou modifier une source de données**
- Cliquer sur le bouton  . DataBuilder est automatiquement lancé.



fenêtre d'importation des données

3. Connexion à une source de données

DataBuilder nécessite la connexion à une base de données comprenant une ou plusieurs tables telles que

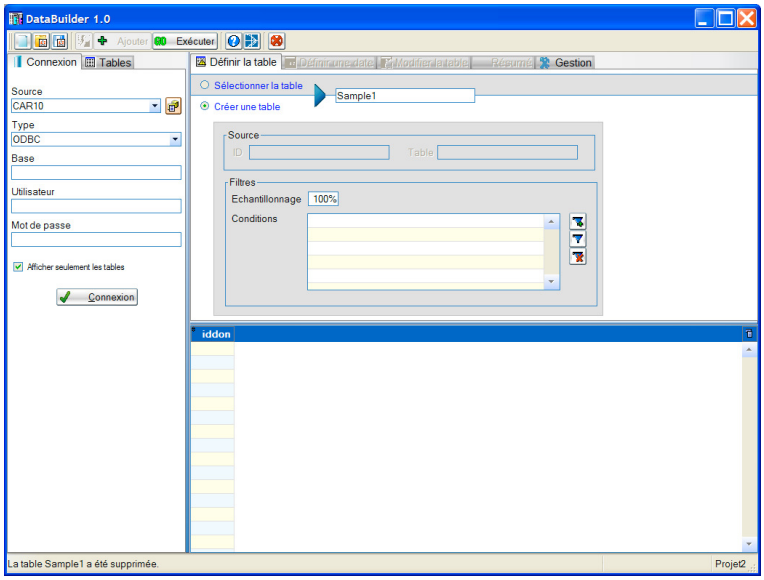
- Clients
- Transactions
- Appels entrants …

A ce jour les bases de données suivantes sont supportées :

- MS Access
- Oracle
- SQLServer

Pour définir la connexion à une base de données :

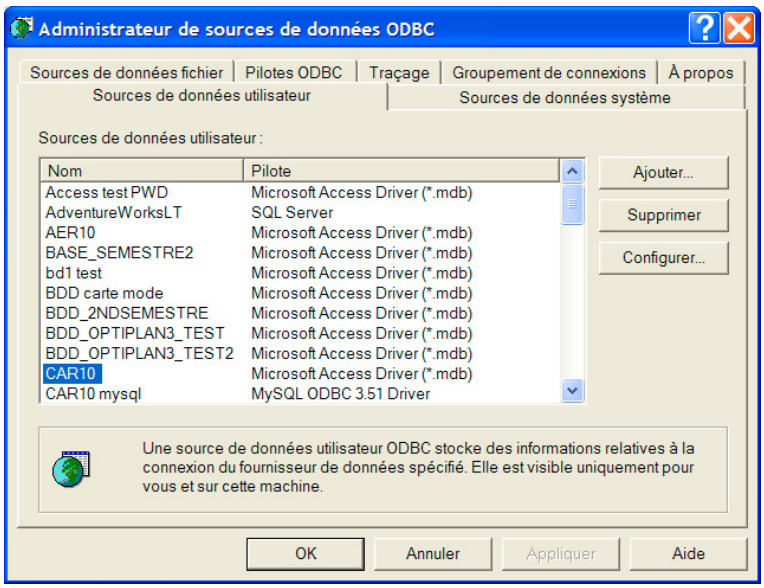
- Cliquer sur l'onglet Connexion
- Sélectionner la source de données ODBC (DSN) dans la liste **Source**. Le cas échéant, définir cette source avec l'aide de l'administrateur de sources de données de Windows en cliquant sur le bouton (Cf. ci-dessous)
- Le cas échéant, entrer le nom d'utilisateur et le mot de passe
- Cliquer sur le bouton **Connexion**. Les tables disponibles s'affichent dans l'onglet **Tables**.
- Double-cliquer sur une table pour obtenir le détail des champs et des types (N : numérique, D : date, C : caractère)



Connexion à la source de données

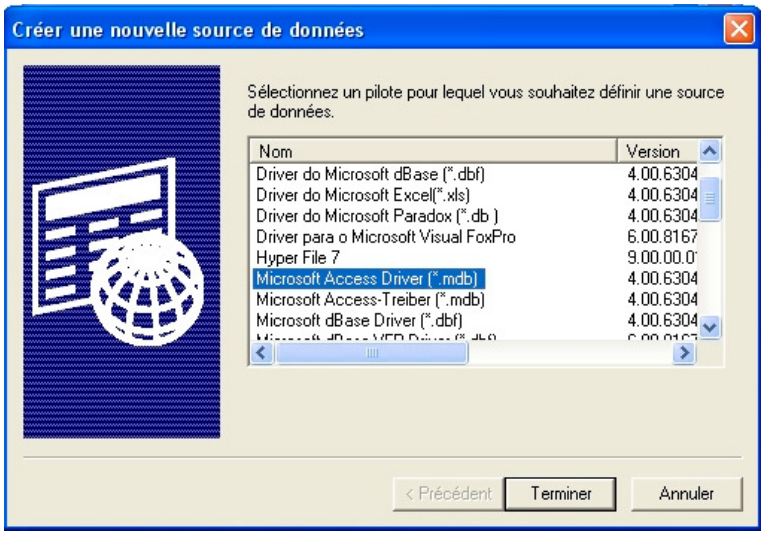
Créer une source de données avec l'administrateur de sources de données ODBC de Windows

a. Dans l'onglet **Connexion**, cliquer sur le bouton **Définir une source ODBC …** . La fenêtre suivante s'affiche .



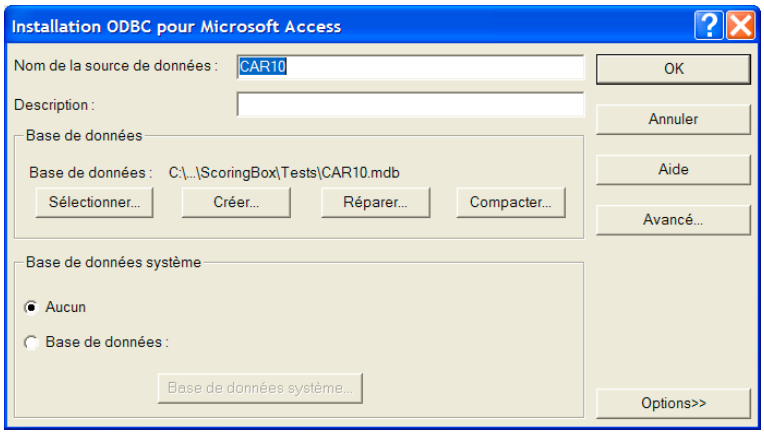
Administrateur de sources de données ODBC de Windows (1/3)

b. Cliquer sur **Ajouter** … . La fenêtre suivante s'affiche .



▸ Administrateur de sources de données ODBC de Windows (2/3)

c. Sélectionner le pilote de la base de données dans la liste, par exemple MS Access Driver pour Access, puis sur Terminer



▸ Administrateur de sources de données ODBC de Windows (3/3)

d. Entrer un nom pour la source de données (bouton Sélectionner …), éventuellement une description , puis sélectionner la base et le cas échéant d'autres paramètres requis dans Options… . Ceci est valable pour une base de type MS Access. Pour d'autres bases la définition peut nécessiter d'entrer le nom du serveur, l'ip si la base est distante, … Dans ce cas, voir le paramétrage avec l'administrateur de la base.

e. Cliquer sur OK pour enregistrer la définition de la nouvelle source et encore sur OK pour fermer l'administrateur de sources de données ODBC de Windows. La nouvelle source de données est automatiquement ajoutée à la liste des sources disponibles.

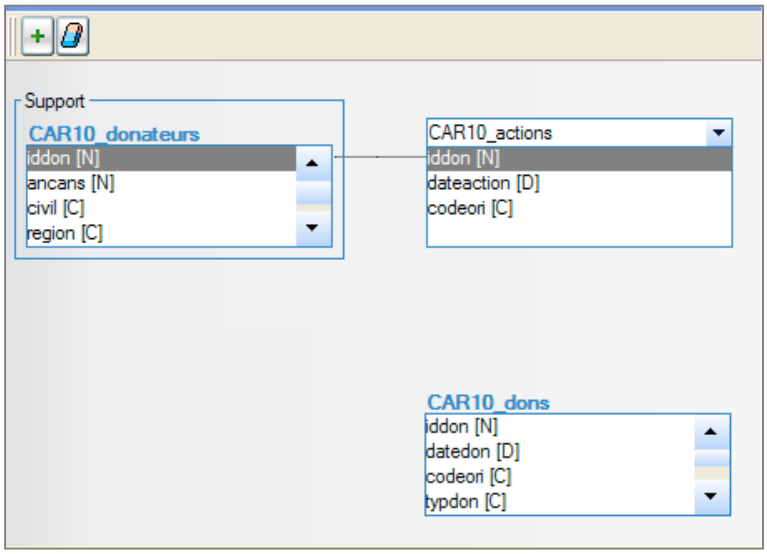
4. Définition des relations entre les tables

Une fois connecté à la source de données ODBC, DataBuilder affiche la liste des tables de la base.

En double-cliquant sur une table l'utilisateur peut obtenir le détail des champs et des types (N : numérique, D : date, C : caractère).

DataBuilder permet la définition de relations entre une table support (table client) et des tables liées (table de description client, de transactions …) selon un lien unique. Des liaisons peuvent aussi être définies entre chaque table et une table de correspondance.

Pour définir les relations entre les tables cliquer sur le bouton  . La fenêtre de définition des relations entre les tables s'affiche.



▸ Définition des relations entre les tables

Définir les relations (jointures)

a. Sélectionner la table 'Support' dans la liste de la zone **Support**. La table 'Support' est la table de référence client comportant au minimum l'identifiant client et, souvent, des informations descriptives. Cette table ne devrait pas comporter de clients doublons et constitue la table centrale des analyses.

b. Pour chaque table à intégrer à l'analyse :
   - Ajouter une table en cliquant sur le bouton +
   - Positionner la table en cliquant droit sur la table puis, après le déplacement, en cliquant droit à nouveau
   - Définir la table en la sélectionnant dans la liste qui apparaît en cliquant sur la zone
c. Pour chaque relation à intégrer à l'analyse :
   - Cliquer sur le champ à lier dans la première table. La curseur de la souris prend la forme d'une main.
   - Cliquer sur le champ à lier dans la deuxième table
   - Un lien est automatiquement dessiné entre les deux tables
d. Vérifier que l'identifiant de référence de la table 'Support' est sélectionné. Cliquer sur Terminer pour enregistrer les changements.

Les liens peuvent être effacés en cliquant sur le bouton  .

Veiller à ce qu'un index soit créé sur chaque champ utilisé pour les jointures. Dans le cas contraire les temps de calcul des requêtes peuvent être anormalement longs. Pour créer un index attaché à un champ se reporter à la documentation de la base de données utilisée.

5. Constitution de la table d'analyse

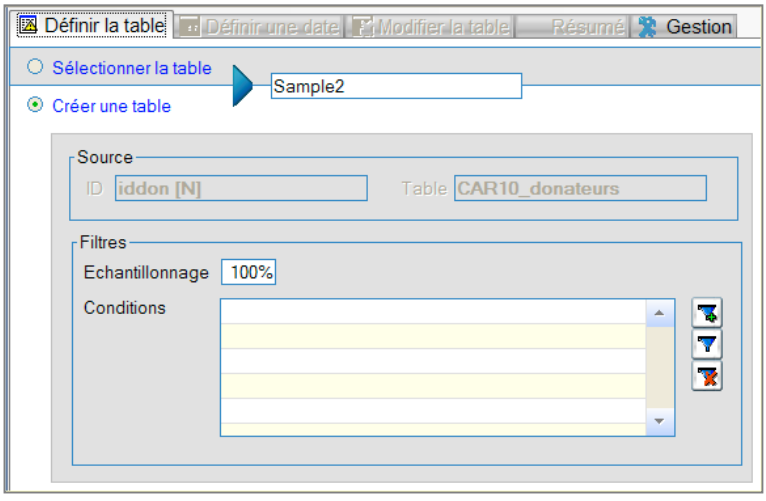
Une fois les tables et les relations entre les tables définies, l'utilisateur peut créer la table d'analyse.

Une table d'analyse créée par DataBuilder est constituée des éléments suivants :

- Un champ **Identifiant de référence** : ce champ (obligatoire) est par construction l'identifiant de référence de la table 'Support' défini dans la partie 4.
- Un champ **Date de référence** : ce champ (obligatoire) permet de définir une date pivot pour la plupart des analyse et sa valeur pourra être fixe ou différente selon les clients
- Des champs issus de la sélection ou d'agrégats des champs des tables utilisées

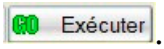
5. 1. Création de la table comportant l'identifiant de référence

a. Cliquer sur l'onglet **Définir la table**
b. Sélectionner l'option **Créer la table**
c. Modifier le cas échéant le nom de la table à créer
d. Modifier le cas échéant le taux d'échantillonnage (par défaut 100%)



▸ Création de la table

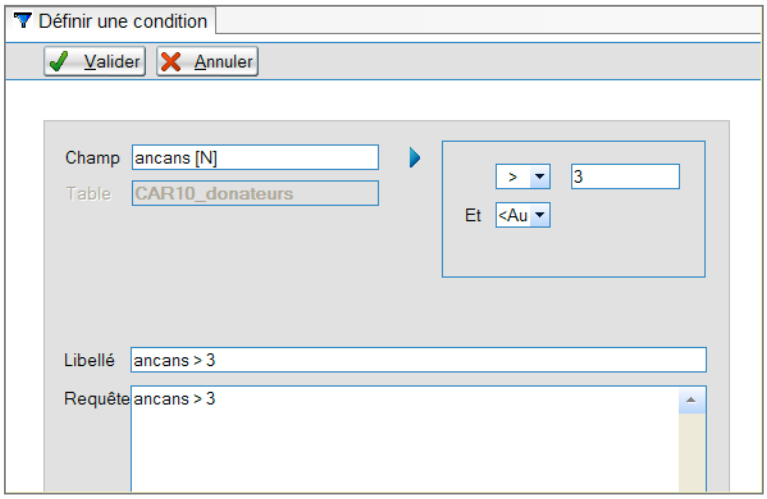
e. Le cas échéant sélectionner une ou plusieurs conditions de sélection des individus en cliquant sur le bouton 

f. Exécuter la requête en cliquant sur le bouton 

. Le champ **Identifiant de référence** défini lors de la définition des relations entre les tables est automatiquement ajouté

Définir une condition de sélection

- Cliquer sur le bouton  pour ajouter une condition
- Dans la liste des Table / champs à gauche (onglet **Tables**) sélectionner avec la souris et (sans relâcher le bouton de la souris) faire glisser le champ à utiliser pour définir la condition jusqu'à la zone **Champ** de la partie droite
- Relâcher le bouton de la souris. Le champ qui spécifie la condition est défini
- Définir dans la zone de droite la condition à appliquer au champ (ici, > 3)
- Cliquer sur le bouton **Valider**. La condition est automatiquement ajoutée à la liste



Définition d'une condition

5. 2. Définition de la date de référence

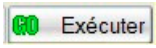
a. Cliquer sur l'onglet **Définir une date**

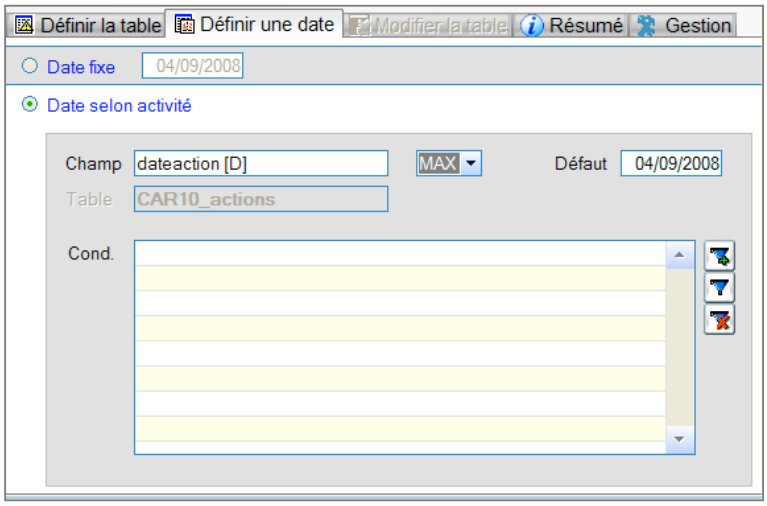
b. Sélectionner l'option **Date fixe** ou **Date selon activité**

c. Si option **Date fixe** entrer la date fixe (par défaut la date du jour)

d. Si option **Date selon activité**

- sélectionner le champ **Date** (repéré par le symbole [D]) d'une des tables et le faire glisser jusqu'à la zone **Champ**
- Choisir la sélection de la date désirée (MIN ou MAX, c'est à dire première ou dernière date trouvée)
- Entrer une date par défaut à utiliser si aucune date n'est trouvée
- Le cas échéant entrer une condition
- Le cas échéant sélectionner une ou plusieurs conditions en cliquant sur le bouton  (Cf. plus haut)

e. Exécuter la requête en cliquant sur le bouton  . Le champ est automatiquement ajouté



▸ Définition de la date de référence (ici l'instant d'observation de l'individu est la plus récente valeur de *dateaction*)

5. 3. Ajout d'un champ

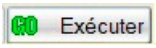
a. Cliquer sur l'onglet **Modifier la table**

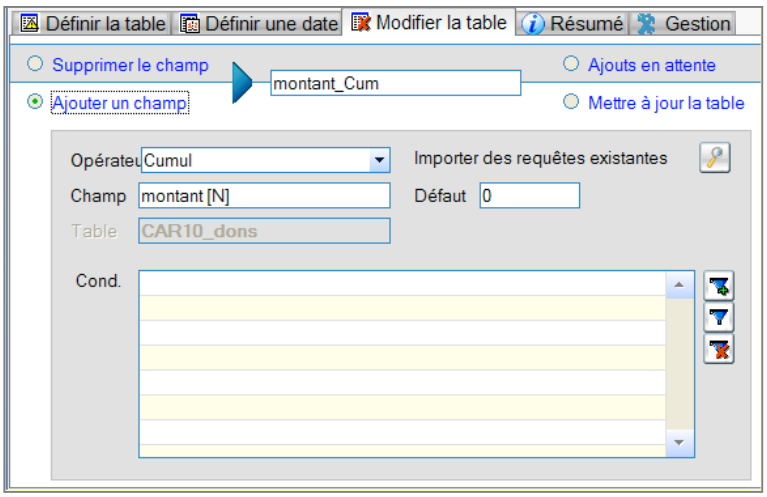
b. Sélectionner l'option **Ajouter un champ**

c. Sélectionner l'un des opérateurs suivants :

- Sélection : sélection d'un champ
- Nombre : comptage du nombre de lignes
- Cumul : somme de la valeur du champ (champ de type Numérique uniquement)
- Min : valeur mini du champ (champ de type Numérique uniquement)
- Max : valeur maxi du champ (champ de type Numérique uniquement)
- Ancienneté : délai maxi du champ par rapport à la date de référence (champ de type Date uniquement)
- Récence : délai mini du champ par rapport à la date de référence (champ de type Date uniquement)
- Combinaison de champs : combinaison de 2 champs avec un opérateur SQL
- Le cas échéant sélectionner une ou plusieurs conditions en cliquant sur le bouton  (Cf. plus bas)

d. Renommer le cas échéant le nom du champ à ajouter (par défaut, système de libellé automatique)

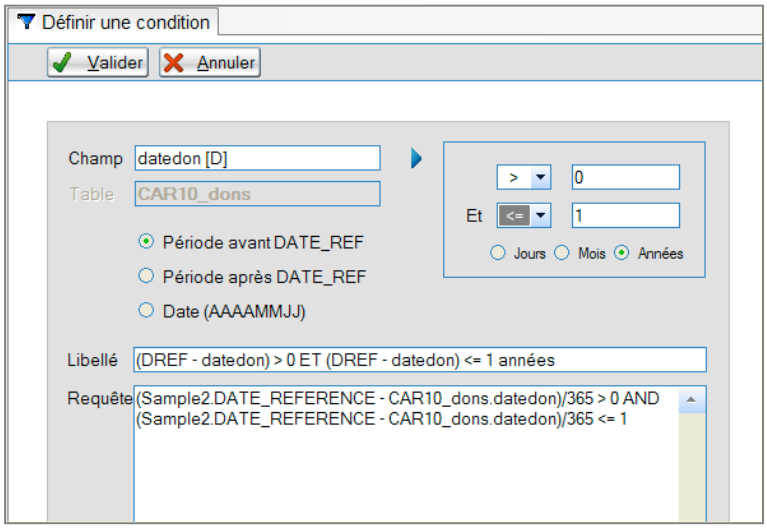
e. Exécuter la requête en cliquant sur le bouton  . Le champ est automatiquement ajouté



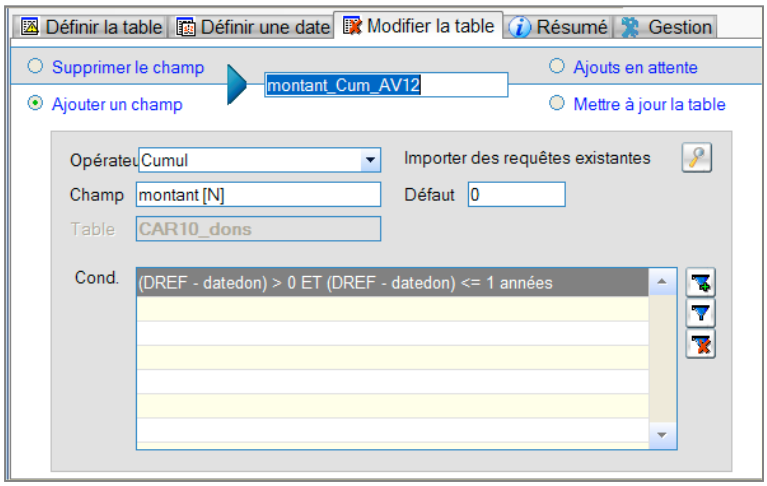
- Ajout d'un champ **Cumul(Montant)**

Définir une condition reposant sur une date

- Cliquer sur le bouton  pour ajouter une condition
- Dans la liste des Table / champs à gauche (onglet Tables) sélectionner avec la souris et (sans relâcher le bouton de la souris) faire glisser le champ Date à utiliser pour définir la condition jusqu'à la zone Champ de la partie droite
- Relâcher le bouton de la souris. Le champ qui spécifie la condition est défini
- Cocher l'option suivante :
  - Période avant DATE_REF : la condition porte sur une période calculée avant la date de référence
  - Période après DATE_REF : la condition porte sur une période calculée après la date de référence
  - Date (AAAMMJJ) : la condition porte sur date dont le format est AAAAMMJJ
- Définir dans la zone de droite la condition ou les conditions à appliquer au champ (ici respectivement, > 0 et <= 1)
- Cliquer sur le bouton **Valider**. La condition est automatiquement ajoutée à la liste

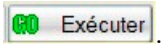


▸ Condition sur un champ Date (Ici, la période de *datedon* à Date de référence est comprise entre 0 et 1 an inclus)

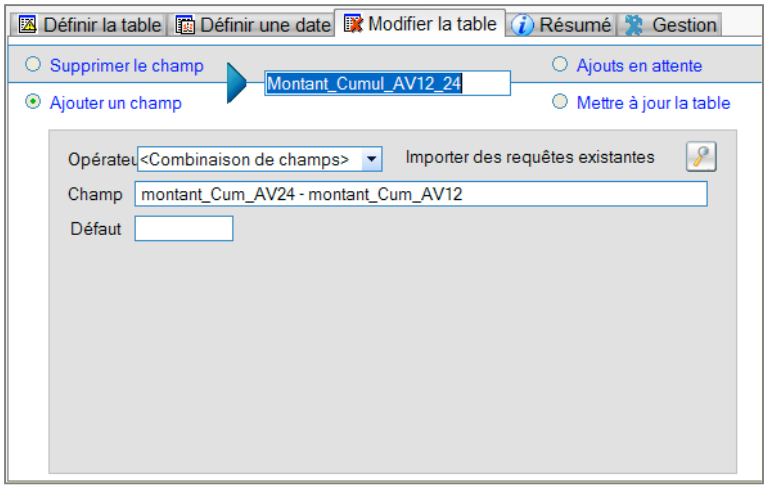


- Création du champ **Cumul(Montant) dans les 12 mois avant Date de référence**

Ajouter un champ avec l'opérateur *Combinaison de champs*

- Sélectionner dans la liste l'opérateur **Combinaison de champs**
- Cliquer, en appuyant sur la touche CTRL, sur le 1er champ dans la table de la partie inférieure de DataBuilder. Le nom du champ apparaît dans la zone **Champ**
- Cliquer, en appuyant sur la touche CTRL, sur le 2eme champ dans la table de la partie inférieure de DataBuilder. Le nom du champ apparaît dans la zone **Champ** à la suite du premier
- Modifier le cas échéant l'opérateur liant les 2 champs en remplaçant l'opérateur par défaut (+) par un des opérateur SQL autorisé (*, /, -, …)
- Entrer une valeur par défaut ou laisser vide si la valeur par défaut doit être la valeur NULL
- Renommer le cas échéant le nom du champ à ajouter (par défaut, système de libellé automatique)
- Exécuter la requête en cliquant sur le bouton 

. Le champ est automatiquement ajouté



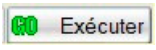
- Création du champ **Cumul(Montant) dans la période 12-24 mois avant Date de référence**

ⓘ **L'opérateur Combinaison de champ permet plus généralement de créer une requête simple (sans SELECT, UPDATE …) portant sur un ou plusieurs champs. Toute requête SQL simple est autorisée, par exemple :**

- **AVG(champ1) (valeur par défaut utilisée si champ2 = NULL)**
- **MIN(champ1)**
- **LTRIM(champ1)**
- **champ1 / champ2 (valeur par défaut utilisée si champ2 = 0)**
- **CASE (champ1 > champ2) THEN 1 ELSE 0**
- **…**

5. 4. Autres fonctions

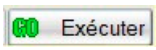
Supprimer un champ de la table d'analyse

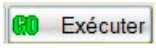
a. Cliquer sur l'onglet **Modifier la table**
b. Sélectionner l'option **Supprimer le champ**
c. Cliquer sur le champ dans la table de la partie inférieure de DataBuilder. Le nom du champ apparaît dans la zone **Champ** à la suite du premier
d. Exécuter la requête de suppression en cliquant sur le bouton 

. Le champ est automatiquement supprimé

Mettre un ou plusieurs champs en attente d'ajout

Il est possible d'ajouter un champ qui a été défini pour une exécution ultérieure (au lieu d'exécuter la requête d'ajout immédiatement après l'avoir défini).
Pour ajouter un champ de cette manière, reprendre les étapes détaillées en 5.3. mais valider (ajouter) en cliquant sur le bouton 

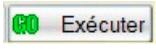
au lieu de cliquer sur le bouton 

.
Les champs en attente sont repérés par une couleur verte.
L'ensemble des ajouts en attente est effectivement ajouté à la table d'analyse en sélectionnant l'option **Ajouts en attente** (onglet **Modifier la table**) puis en cliquant sur le bouton 

.

Mettre à jour la table d'analyse

La totalité de la table d'analyse peut être mise à jour (tous les champs sont recalculés) de la manière suivante :

e. Cliquer sur l'onglet **Modifier la table**
f. Sélectionner l'option **Mettre à jour la table**
g. Exécuter la requête de mise à jour en cliquant sur le bouton 

. Les champs sont automatiquement mis à jours séquentiellement.

Importer des requêtes existantes

Il est possible d'importer des requêtes existantes, enregistrées dans un autre projet DataBuilder (*.dbd).

a. Cliquer sur l'onglet **Modifier la table**
b. Sélectionner l'option **Ajouter un champ**
c. Cliquer sur le bouton **Importer des requêtes existantes** 
d. Dans la fenêtre, ouvrir un projet (*.dbd) en cliquant sur le bouton 
e. Cocher les requêtes à importer
f. Le cas échéant, renommer les requêtes
g. **Valider** pour fermer la fenêtre

Visualiser des statistiques sur la table d'analyse en cours

A tout moment il est possible d'obtenir des statistiques sue la table d'analyse en cours en cliquant sur l'onglet **Résumé**. Les statistiques disponibles sont :

- Nombre d'enregistrements
- Nombre de champs
- Type, moyenne, min, max, nombre de manquants de chaque champ

Gestion du projet et de la source de données

L'ensemble des information relative à la table d'analyse créée est stockée dans un projet (*.dbd). Le nom et l'emplacement du projet peuvent être modifiés via l'interface disponible sur l'onglet **Gestion**.
Il est aussi possible de supprimer une table ou un champ de la source de données sélectionnées à partir de cet onglet.

6. Utiliser la table d'analyse dans DataLab

Lorsque la table est terminée, cette dernière est physiquement écrite dans la source de données ouverte dans le projet DataBuilder et donc disponible pour une connexion avec DataLab.

- Fermer DataBuilder
- Sélectionner **Oui** sur le message affiché par DataLab « Voulez-vous importer la table modifiée par DataBuilder ? »
- La table d'analyse est automatiquement affichée dans DataLab
- Cliquer sur **Importer**