IBM SPSS Data Preparation 19

Note: Before using this information and the product it supports, read the general information under Notices sur p. 149.

This document contains proprietary information of SPSS Inc, an IBM Company. It is provided under a license agreement and is protected by copyright law. The information contained in this publication does not include any product warranties, and any statements provided in this manual should not be interpreted as such.

When you send information to IBM or SPSS, you grant IBM and SPSS a nonexclusive right to use or distribute the information in any way it believes appropriate without incurring any obligation to you.

© Copyright SPSS Inc. 1989, 2010.

Préface

IBM® SPSS® Statistics est un système complet d'analyse de données. Le module complémentaire facultatif Préparation des données fournit les techniques d'analyse supplémentaires décrites dans ce manuel. Le module complémentaire Préparation des données doit être utilisé avec le système central SPSS Statistics auquel il est entièrement intégré.

A propos de SPSS Inc., an IBM Company

SPSS Inc., an IBM Company, est un des leaders dans le domaine des solutions logicielles d'analyse prédictive. Le portfolio complet des produits de la société — Data collection, Statistics, Modeling et Deployment — capture les opinions et les attitudes du public, prédit les résultats des interactions futures des clients, et agit ensuite sur ces données en intégrant les analyses dans les processus commerciaux. Les solutions SPSS Inc. répondent aux objectifs commerciaux interdépendants d'une organisation dans sa totalité en se concentrant sur la convergence des analyses, de l'architecture informatique et des processus commerciaux. Des clients issus du milieu des affaires, du milieu gouvernemental ou du milieu académique, dans le monde entier, font confiance à la technologie SPSS Inc., et la considère comme un atout pour attirer et retenir leurs clients, ou encore augmenter leur nombre, tout en réduisant les fraudes et les risques. SPSS Inc. a été acheté par IBM en octobre 2009. Pour plus d'informations, visitez le site *http://www.spss.com*.

Support technique

Un support technique est disponible pour les clients du service de maintenance. Les clients peuvent contacter l'assistance technique pour obtenir de l'aide concernant l'utilisation des produits SPSS Inc. ou l'installation dans l'un des environnements matériels pris en charge. Pour contacter l'assistance technique, consultez le site Web SPSS Inc. à l'adresse *http://support.spss.com*, ou recherchez votre représentant local à la page *http://support.spss.com/default.asp?refpage=contactus.asp* Votre nom, celui de votre société, ainsi que votre contrat d'assistance vous seront demandés.

Service clients

Si vous avez des questions concernant votre envoi ou votre compte, contactez votre bureau local, dont les coordonnées figurent sur le site Web à l'adresse : *http://www.spss.com/worldwide*. Veuillez préparer et conserver votre numéro de série à portée de main pour l'identification.

© Copyright SPSS Inc. 1989, 2010

Séminaires de formation

SPSS Inc. propose des séminaires de formation, publics et sur site. Tous les séminaires font appel à des ateliers de travaux pratiques. Ces séminaires seront proposés régulièrement dans les grandes villes. Pour plus d'informations sur ces séminaires, contactez votre bureau local dont les coordonnées sont indiquées sur le site Web à l'adresse : *http://www.spss.com/worldwide*.

Documents supplémentaires

Les ouvrages *SPSS Statistics : Guide to Data Analysis*, *SPSS Statistics : Statistical Procedures Companion*, et *SPSS Statistics : Advanced Statistical Procedures Companion*, écrits par Marija Norušis et publiés par Prentice Hall, sont suggérés comme documentation supplémentaire. Ces publications présentent les procédures statistiques des modules SPSS Statistics Base, Advanced Statistics et Regression. Que vous soyez novice dans les analyses de données ou prêt à utiliser des applications plus avancées, ces ouvrages vous aideront à exploiter au mieux les fonctionnalités offertes par IBM® SPSS® Statistics. Pour obtenir des informations supplémentaires y compris le contenu des publications et des extraits de chapitres, visitez le site web de l'auteur : *http://www.norusis.com*

Contenu

Partie I:
Guide de l'utilisateur

Chapitre
1

Introduction à la préparation des données

L'augmentation de la demande d'information est proportionnelle à l'augmentation de la puissance des systèmes informatiques, provoquant la multiplication des données collectées, tout comme celle des observations, des variables et des erreurs de saisie de données. Ces erreurs représentent l'ennemi principal des modèles de prévision, ces derniers servant à entreposer les données, vous devez donc conserver des données « propres ». Cependant, la quantité de données entreposées a augmenté de telle façon qu'il n'est plus possible de vérifier manuellements les observations. Il devient alors primordial d'automatiser les processus de validation des données.

Le module complémentaire Préparation des données vous permet d'identifier les observations inhabituelles et les observations non valides, ainsi que les variables et les valeurs de données dans votre ensemble de données actif, de plus ce module prépare les données pour la modélisation.

Utilisation des procédures de préparation des données

Votre utilisation des procédures de préparation des données dépend de vos besoins. Un processus standard de validation des données, une fois vos données chargées, consiste à :

- **Préparer les métadonnées** Etudiez les variables de votre fichier de données et déterminez leur valeur valide, leur étiquette et leurs niveaux de mesure. Identifiez les combinaisons des valeurs de variables impossibles qui sont couramment mal codées. Définissez les règles de validation en vous basant sur cette information. Cette tâche peut prendre beaucoup de temps, mais elle peut s'avèrer vraiment utile si vous devez régulièrement valider des fichiers de données possédant des attributs similaires.
- **Valider les données** Exécutez des vérifications et des contrôles de base des règles de validation définies afin d'identifier les observations inhabituelles, les variables et les valeurs de données. Une fois les données invalides repérées, déterminez-en la cause et corrigez le problème. Vous devrez peut-être effectuer une étape supplémentaire de préparation des métadonnées.
- **Préparer le modèle** Utilisez une préparation automatique des données afin de transformer les champs d'origine, ce qui va améliorer la construction du modèle. Identifiez les valeurs éloignées statistiques potentielles pouvant être à l'origine de problèmes rencontrés dans de nombreux modèles de prévision. Certaines valeurs éloignées sont dues à des valeurs de variables invalides qui n'ont pas été identifiées. Vous devrez peut-être effectuer une étape supplémentaire de préparation des métadonnées.

Une fois que votre fichier de données est « propre », vous êtes prêt à construire des modèles à partir d'autres modules complémentaires.

© Copyright SPSS Inc. 1989, 2010

Chapitre
2

Règles de validation

Une règle sert à déterminer la validité d'une observation. Il existe deux types de règles de validation :

- **Règles de variable unique.** Les règles de variable unique sont composées d'un ensemble fixe de vérification s'appliquant à une variable unique, telle que les vérifications des valeurs hors plage. Les valeurs valides peuvent être exprimées sous la forme d'un intervalle de valeurs ou d'une liste de valeurs possibles en ce qui concerne les règles de variable unique.
- **Règles de variable croisée.** Les règles de variable croisée sont des règles définies par l'utilisateur qui peuvent être appliquées à une variable unique ou à des variables combinées. Les règles de variable croisée sont définies par une expression logique qui repère les valeurs non valides.

Les règles de validation sont enregistrées dans le dictionnaire de données de votre fichier de données. Vous pouvez ainsi spécifier une règle une fois et la réutiliser ensuite.

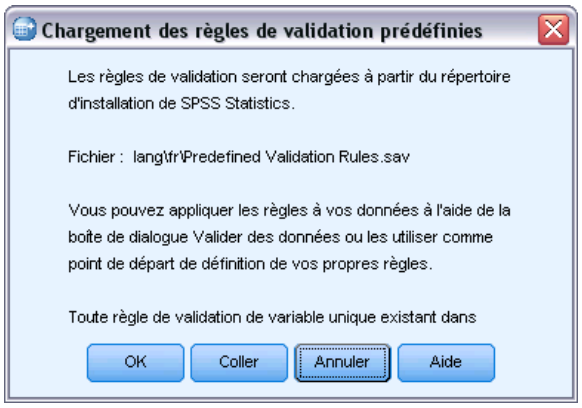
Chargement des règles de validation prédéfinies

Vous pouvez rapidement obtenir un ensemble de règles de validation prêtes à l'emploi en chargeant des règles prédéfinies à partir d'un fichier de données externe inclus dans l'installation.

Pour charger des règles de validation prédéfinies

▶ A partir des menus, sélectionnez :
Données > Validation > Charger des règles prédéfinies...

Figure 2-1
Chargement des règles de validation prédéfinies



© Copyright SPSS Inc. 1989, 2010

Notez que ce processus supprime les règles de variable unique existantes dans l'ensemble de données actif.

Vous pouvez également utiliser l'assistant Copier des propriétés de données pour charger les règles à partir de n'importe quel fichier de données.

Définir des règles de validation

La boîte de dialogue Définir des règles de validation vous permet de créer et d'afficher des règles de validation de variable unique et de variable croisée.

Pour créer et afficher les règles de validation

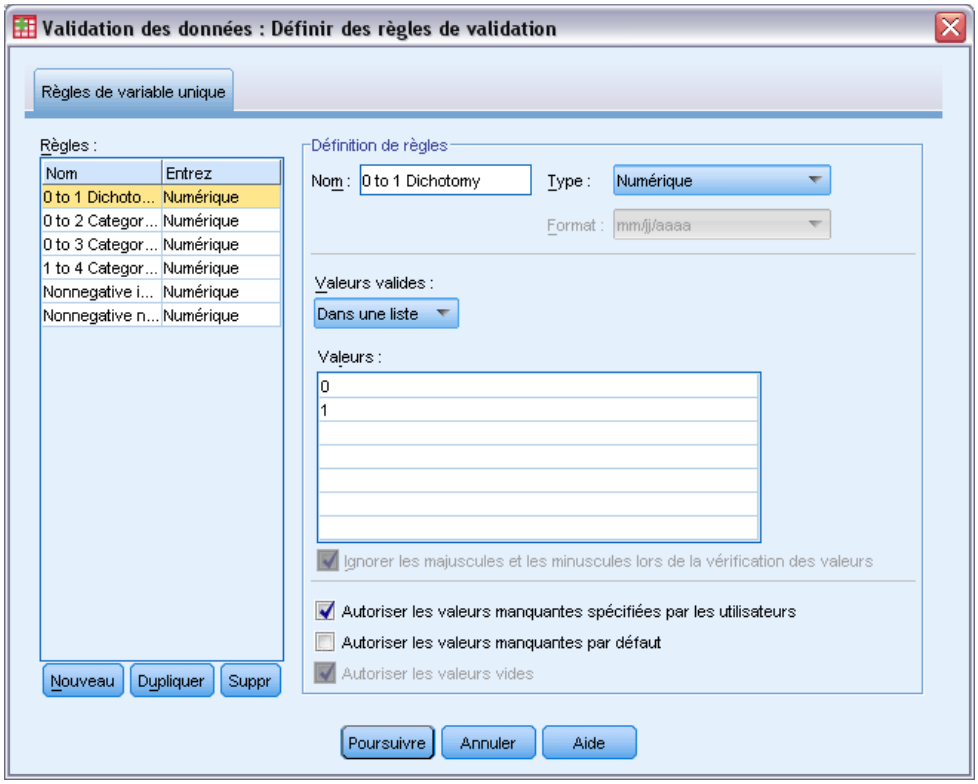
- A partir des menus, sélectionnez :
  Données > Validation > Définir des règles...

  La boîte de dialogue est remplie de règles de validation de variable unique et de variable croisée issues du dictionnaire de données. En l'absence de règles, une nouvelle règle de substitution que vous pouvez modifiez en fonction de vos besoins est créée automatiquement.

- Sélectionnez des règles individuelles dans les onglets Règles de variable unique et Règles de variable croisée pour afficher et modifier leurs propriétés.

Définition des règles de variable unique

Figure 2-2
Boîte de dialogue Définir des règles de validation, onglet Règles des variables uniques



L'onglet Règles de variable unique vous permet de créer, d'afficher et de modifier les règles de validation de variable unique.

Règles. La liste affiche les règles de validation de variable unique par nom et le type de variable auquel la règle peut être appliquée. A l'ouverture de la boîte de dialogue, les règles définies dans le dictionnaire de données s'affichent ou, si aucune règle n'a été définie, une règle de substitution intitulée « Règle de variable unique 1 » apparaît. Les boutons suivants apparaissent au-dessous de la liste Règles :

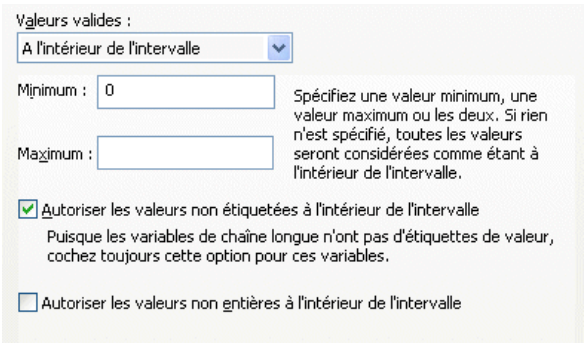
- **Nouveau.** Ajoute une nouvelle entrée au bas de la liste Règles. La règle est sélectionnée et le nom « SingleVarRule *n* » lui est appliqué, *n* correspondant à un nombre entier de sorte que le nom de la nouvelle règle n'ait pas de doublon parmi les règles de variable unique et de variable croisée.
- **Dupliquer.** Ajoute une copie de la règle sélectionnée au bas de la liste Règles. Le nom de la règle est ajusté de sorte qu'il n'y ait pas de doublon parmi les règles de variable unique et de variable croisée. Par exemple, si vous dupliquez « SingleVarRule 1)», le nom de la première règle dupliquée sera « Copy of SingleVarRule 1 » tandis que le nom de la deuxième sera « Copy (2) of SingleVarRule 1)» etc.
- **Supprimer.** Supprime la règle sélectionnée.

Définir la règle. Ces commandes vous permettent d'afficher et de définir les propriétés d'une règle sélectionnée.

- **Nom.** Le nom de la règle doit être unique parmi les règles de variable unique et de variable croisée.
- **Type :** Il s'agit du type de variable auquel une règle est appliquée. Effectuez votre sélection à partir de Numérique, Chaîne et Date.
- **Format :** Le format vous permet de sélectionner le format de date pour les règles pouvant être appliquées à des variables de date.
- **Valeurs valides.** Vous pouvez indiquer les valeurs valides sous la forme d'une plage ou d'une liste de valeurs.

Les commandes de définition de la plage vous permettent de spécifier une plage de valeurs valides. Les valeurs se trouvant à l'extérieur de cette plage sont repérées et considérées comme invalides.

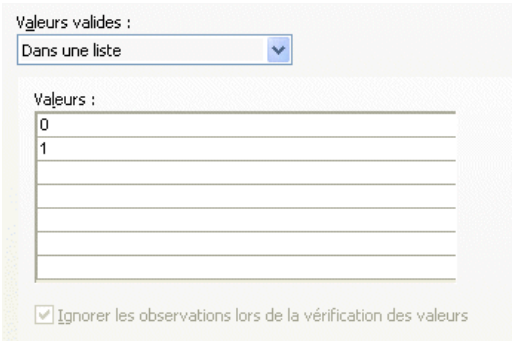
Figure 2-3
Règles de variable unique : Définition de la plage



Entrez la valeur minimale ou la valeur maximale ou bien les deux pour spécifier une plage. Les commandes des cases à cocher vous permettent de repérer les valeurs non étiquetées et non entières à l'intérieur de cette plage.

Les commandes de définition de liste vous permettent de définir une liste de valeurs valides. Les valeurs non comprises dans la liste sont repérées comme invalides.

Figure 2-4
Règles de variable unique : Définition de liste

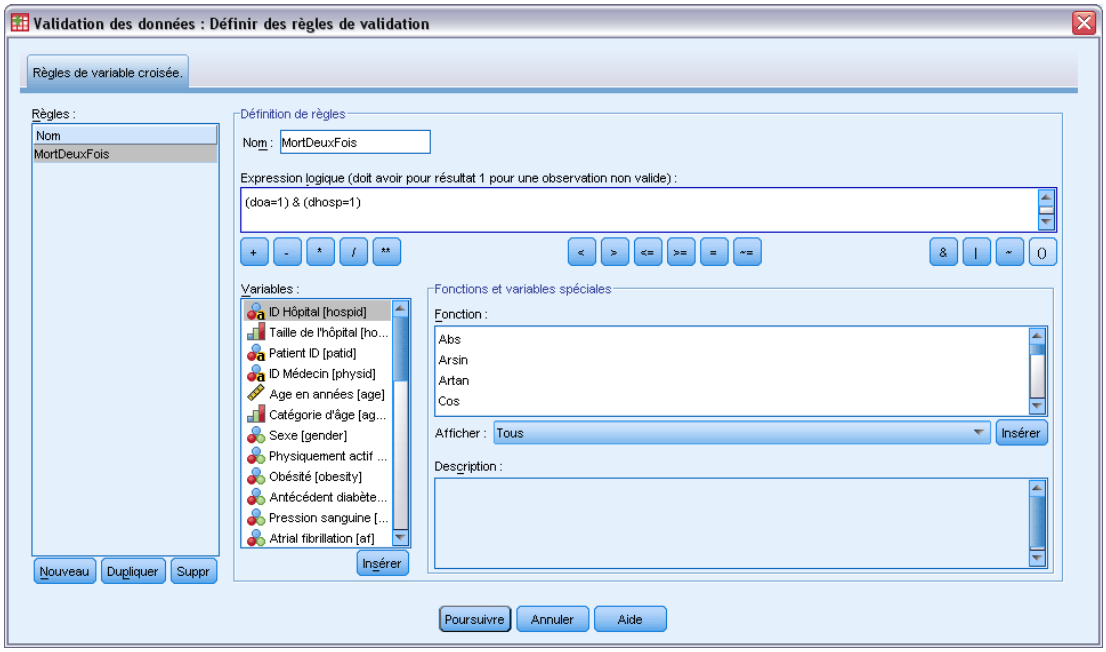


Entrez les valeurs de la liste dans la grille. La case à cocher détermine si les observations sont importantes lorsque les valeurs de données chaîne sont comparées à la liste de valeurs possibles pour vérification.

- **Autoriser les valeurs manquantes spécifiées par les utilisateurs.** Cette fonctionnalité contrôle si les valeurs manquantes spécifiées par l'utilisateur sont repérées comme invalides.
- **Autoriser les valeurs manquantes par défaut.** Cette fonctionnalité contrôle si les valeurs manquantes par défaut sont repérées comme invalides. Elle ne s'applique pas aux types de règle chaîne.
- **Autoriser les valeurs vides.** Cette fonctionnalité contrôle si les valeurs chaîne vides (complètement vides) sont repérées comme invalides. Elle ne s'applique pas aux types de règle non-chaîne.

Définition des règles de variable croisée

Figure 2-5
Boîte de dialogue Définir des règles de validation, onglet Règles des variables croisées



L'onglet Règles de variable croisée vous permet de créer, d'afficher et de modifier les règles de validation de variable croisée.

Règles. La liste affiche les règles de validation de variable croisée par nom. A l'ouverture de la boîte de dialogue, une règle de subsitution intitulée « CrossVarRule 1 » s'affiche. Les boutons suivants apparaissent au-dessous de la liste Règles :

- **Nouveau.** Ajoute une nouvelle entrée au bas de la liste Règles. La règle est sélectionnée et le nom « CrossVarRule *n* » lui est appliqué, *n* correspondant à un nombre entier de sorte que le nom de la nouvelle règle n'ait pas de doublon parmi les règles de variable unique et de variable croisée.

- **Dupliquer.** Ajoute une copie de la règle sélectionnée au bas de la liste Règles. Le nom de la règle est ajusté de sorte qu'il n'y ait pas de doublon parmi les règles de variable unique et de variable croisée. Par exemple, si vous dupliquez « CrossVarRule 1 », le nom de la première règle dupliquée sera « Copy of CrossVarRule 1 » tandis que le nom de la deuxième sera « Copy (2) of CrossVarRule 1 », etc.
- **Supprimer.** Supprime la règle sélectionnée.

Définir la règle. Ces commandes vous permettent d'afficher et de définir les propriétés d'une règle sélectionnée.

- **Nom.** Le nom de la règle doit être unique parmi les règles de variable unique et de variable croisée.
- **Expression logique.** Il s'agit de la définition de règle. Vous pouvez coder l'expression de sorte que les observations invalides aient pour résultat 1.

Construction d'expressions

▶ Pour construire une expression, vous pouvez soit coller les composants dans le champ Expression, soit les saisir directement depuis le clavier.

- Pour coller des fonctions ou des variables système couramment utilisées, sélectionnez un groupe dans la liste Groupe de fonctions, puis, dans la liste Fonctions et variables spéciales, double-cliquez sur la fonction ou la variable voulue (ou sélectionnez-la, puis cliquez sur Insérer). Définissez tous les paramètres indiqués par un point d'interrogation (cette opération ne concerne que les fonctions). Le groupe de fonctions étiqueté Tous répertorie toutes les fonctions et variables système disponibles. Une brève description de la variable ou de la fonction sélectionnée apparaît dans une zone particulière de la boîte de dialogue.
- Les constantes alphanumériques doivent être présentées entre guillemets ou apostrophes.
- Si des valeurs contiennent des chiffres décimaux, utilisez la virgule comme indicateur décimal.

Chapitre
3

Valider des données

La boîte de dialogue Valider des données vous permet d'identifier des observations suspicieuses ou invalides, des variables et des valeurs de données dans l'ensemble de données actif.

Exemple : Un analyste de données doit fournir une enquête de satisfaction client à son client tous les mois. L'analyste doit effectuer une vérification de la qualité des données reçues chaque mois, afin de contrôler qu'il n'y a pas d'ID client incomplet, de valeurs de variables hors plage, de combinaisons de valeurs de variable régulièrement saisies par erreur. Avec la boîte de dialogue Valider des données, l'analyste peut spécifier les variables qui ne servent à identifier que les clients, définir les règles de variable unique pour les plages de variables valides et enfin définir les règles de variable croisée afin de repérer les combinaisons impossibles. La procédure renvoit un rapport sur les observations et les variables posant problèmes. De plus, les données possèdent les mêmes éléments de données chaque mois, ce qui permet à l'analyste d'appliquer les règles au nouveau fichier de données du mois suivant.

Statistiques : La procédure génère des listes de variables, d'observations et de valeurs de données qui n'ont pas passé plusieurs contrôles, des effectifs de violation des règles de variable unique et de variable croisée, ainsi que de simples récapitulatifs descriptifs des variables d'analyse.

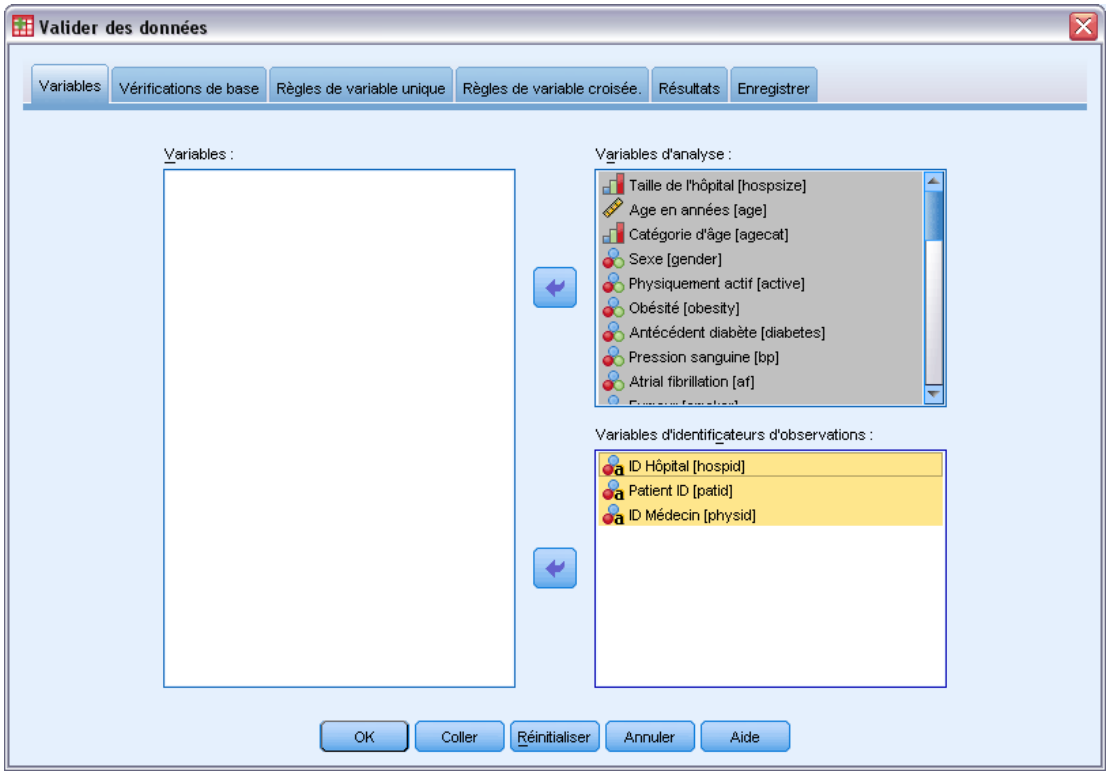
Pondérations. La procédure ignore la spécification de la variable de pondération et la traite comme toute autre variable d'analyse.

Pour valider des données

- A partir des menus, sélectionnez :
  Données > Validation > Valider des données...

© Copyright SPSS Inc. 1989, 2010

Figure 3-1
Boîte de dialogue Valider les données, onglet Variables



- Sélectionnez une ou plusieurs variables d'analyse afin de les faire valider par des vérifications de base des variables ou par des règles de validation de variable unique.

  Vous pouvez également :

- cliquer sur l'onglet Règles de variable croisée et appliquer une ou plusieurs règles de variable croisée.

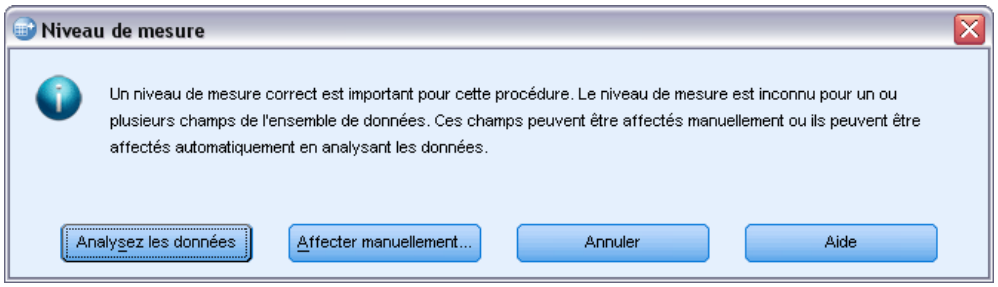
  Sinon, vous pouvez :

  - sélectionner une ou plusieurs variables d'identification d'observations afin de vérifier s'ils existent des ID dupliqués ou incomplets. Les variables d'ID d'observation sont également utilisées pour étiqueter les résultats par observations. Si deux ou plus de deux variables d'ID d'observations sont spécifiées, la combinaison de leurs valeurs est traitée comme un identificateur d'observations.

Champs avec un niveau de mesure inconnu

L'alerte du niveau de mesure apparaît lorsque le niveau de mesure d'une ou plusieurs variables (champs) de l'ensemble de données est inconnu. Le niveau de mesure ayant une incidence sur le calcul des résultats de cette procédure, toutes les variables doivent avoir un niveau de mesure défini.

Figure 3-2
Alerte du niveau de mesure

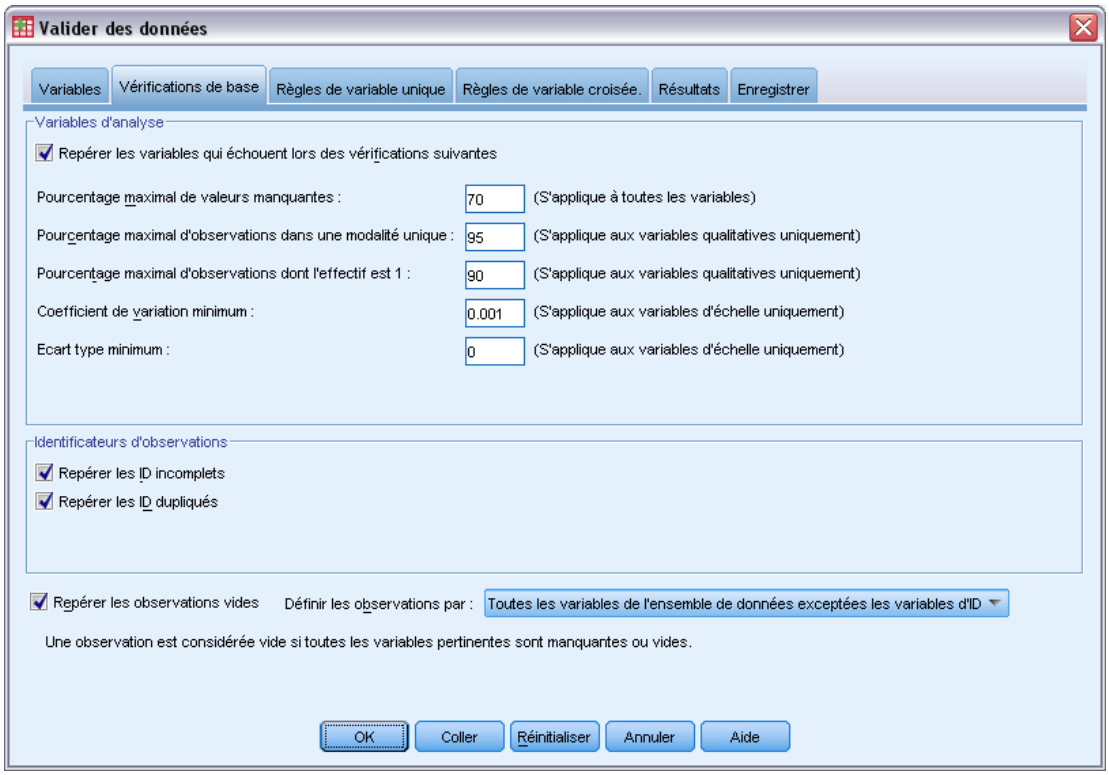


- **Analysez les données.** Lit les données dans l'ensemble de données actifs et attribue le niveau de mesure par défaut à tous les champs ayant un niveau de mesure inconnu. Si l'ensemble de données est important, cette action peut prendre un certain temps.
- **Attribuer manuellement.** Ouvre une boîte de dialogue qui répertorie tous les champs ayant un niveau de mesure inconnu. Vous pouvez utiliser cette boîte de dialogue pour attribuer un niveau de mesure à ces champs. Vous pouvez également attribuer un niveau de mesure dans l'affichage des variables de l'éditeur de données.

Le niveau de mesure étant important pour cette procédure, vous ne pouvez pas accéder à la boîte de dialogue d'exécution de cette procédure avant que tous les champs n'aient des niveaux de mesure définis.

Vérifications de base de validation des données

Figure 3-3
Boîte de dialogue Valider les données, onglet Vérifications de base



L'onglet Vérifications de base vous permet de sélectionner les vérifications de base pour les variables d'analyse, les identificateurs d'observations ainsi que les observations complètes.

Variables d'analyse. Si vous avez sélectionné des variables d'analyse dans l'onglet Variables, vous pouvez sélectionner la ou les vérifications suivantes correspondant à leur validité. La case à cocher vous permet d'activer ou de désactiver les vérifications.

- **Pourcentage maximal de valeurs manquantes.** Répertorie les variables d'analyse dont le pourcentage de valeurs manquantes est supérieur à la valeur indiquée. La valeur indiquée doit être un nombre positif inférieur ou égal à 100.
- **Pourcentage maximal d'observations dans une modalité unique.** Lorsque des variables d'analyse sont qualitatives, cette option répertorie alors les variables d'analyse qualitatives dont le pourcentage d'observations répresentant une modalité unique non manquante est supérieur à la valeur indiquée. La valeur indiquée doit être un nombre positif inférieur ou égal à 100. Le pourcentage est basé sur des observations n'ayant pas de valeur manquante de la variable.
- **Pourcentage maximal de modalités dont l'effectif est 1.** Lorsque des variables d'analyse sont qualitatives, cette option répertorie alors les variables d'analyse qualitatives dont le pourcentage des modalités des variables contenant une seule observation est supérieur à la valeur indiquée. La valeur indiquée doit être un nombre positif inférieur ou égal à 100.

- **Coefficient de variation minimum.** Lorsque des variables d'analyse sont mesurées sur une échelle, cette option répertorie les variables d'analyse d'échelle dont la valeur absolue du coefficient de variation est inférieure à la valeur indiquée. Cette option ne s'applique qu'aux variables dont la moyenne n'est pas nulle. La valeur indiquée doit être un nombre non-négatif. Pour désactiver le coefficient de vérification de la variation, tapez 0.
- **Ecart type minimum.** Lorsque des variables d'analyse sont mesurées sur une échelle, cette option répertorie les variables d'analyse d'échelle dont l'écart-type est inférieur à la valeur indiquée. La valeur indiquée doit être un nombre non-négatif. Pour désactiver la vérification de l'écart-type, tapez 0.

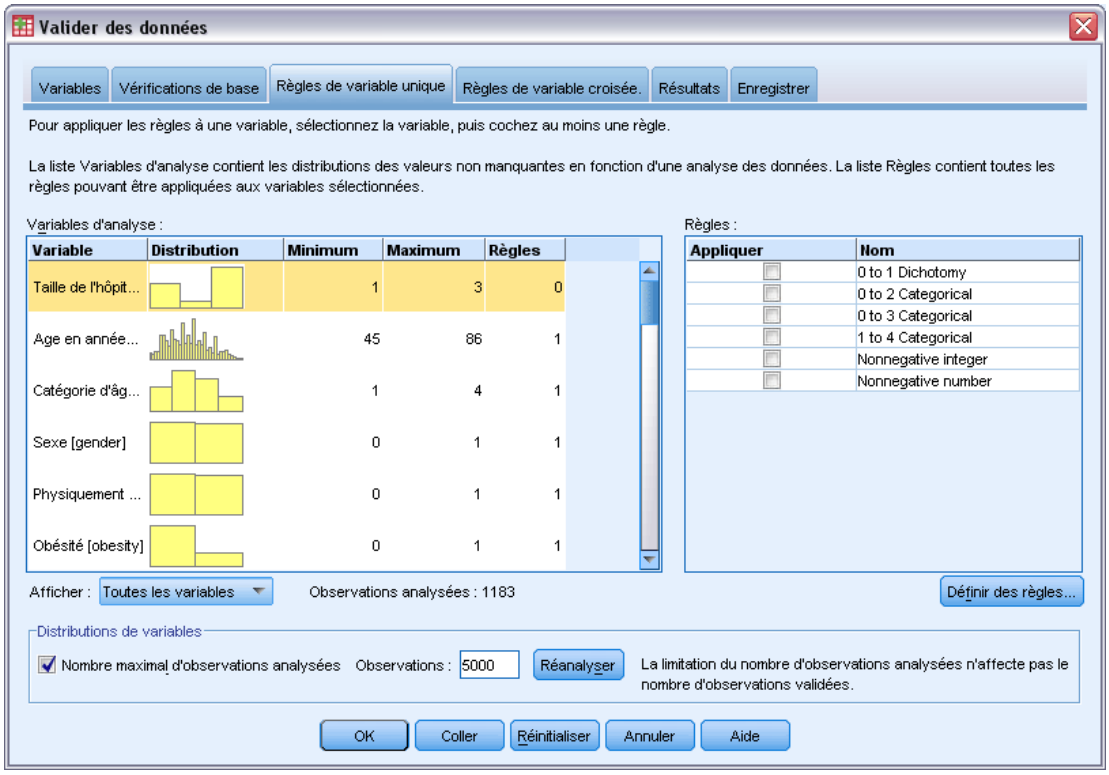
Identificateurs d'observations. Si vous avez sélectionné des variables d'identificateurs d'observations dans l'onglet Variables, vous pouvez sélectionner la ou les vérifications suivantes correspondant à leur validité.

- **Repérer les ID incomplets.** Cette option répertorie les observations dont les identificateurs d'observations sont incomplets. Pour une observation donnée, un identificateur est considéré comme incomplet lorsque la valeur de toute variable ID est vide ou manquante.
- **Repérer les ID dupliqués.** Cette option répertorie les observations dont les identificateurs d'observations sont dupliqués. Les identificateurs incomplets sont exclus de l'ensemble de duplicats possibles.

Repérer les observations vides. Cette option répertorie les observations dont toutes les variables sont vides ou nulles. Pour identifier des observations vides, vous pouvez utiliser toutes les variables du fichier (à l'exception des variables ID) ou seulement les variables d'analyse définies sur l'onglet Variables.

Règles de variable unique de la validation des données

Figure 3-4
Boîte de dialogue Valider les données, onglet Règles des variables uniques



L'onglet Règles de variable unique affiche les règles de validation de variable unique disponibles et vous permet de les appliquer aux variables d'analyse. Pour définir d'autres règles de variable unique, cliquez sur Définir des règles. Pour plus d'informations, reportez-vous à la section Définition des règles de variable unique dans le chapitre 2 sur p. 4.

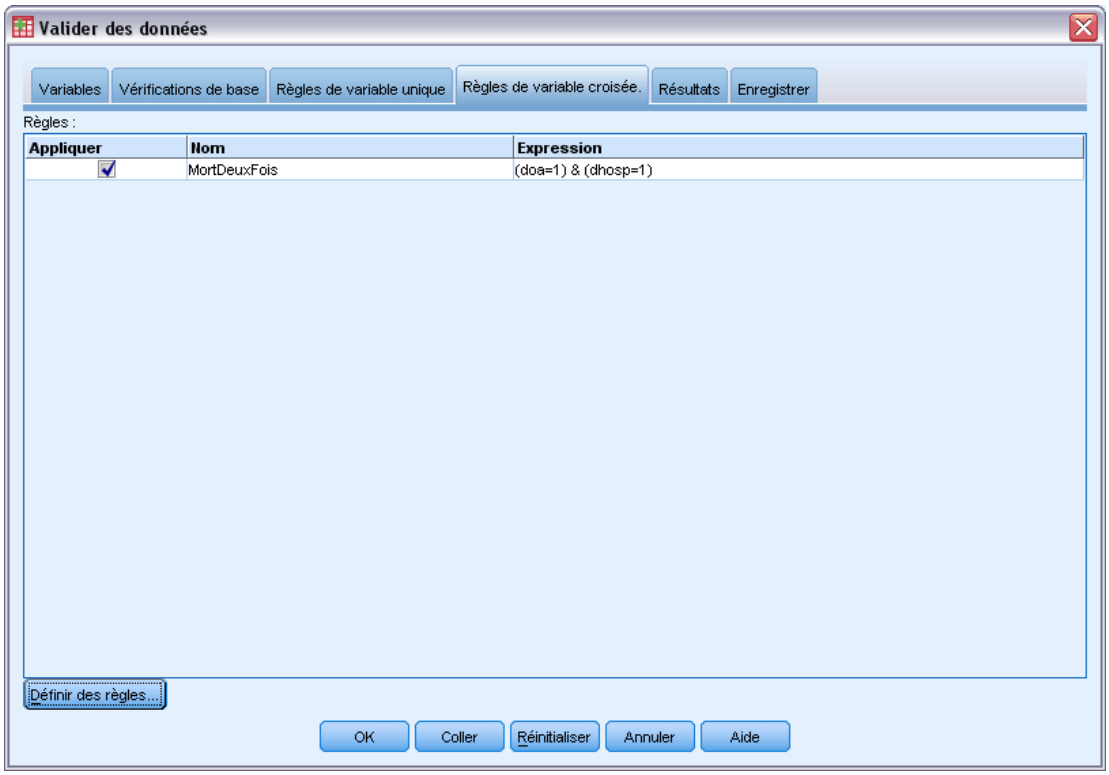
Variables d'analyse. La liste affiche les variables d'analyse, récapitule leurs distributions et indique également le nombre de règles appliqué à chaque variable. Notez que les valeurs manquantes définies par l'utilisateur et par le système ne sont pas incluses dans les récapitulatifs. La liste déroulante Afficher contrôle l'affichage des variables. Vous pouvez sélectionner les affichages suivants : Toutes les variables, Variables numériques, Variables chaîne et Variables de date.

Règles. Pour appliquer des règles à des variables d'analyse, sélectionnez une ou plusieurs variables et vérifiez toutes les règles que vous voulez appliquer dans la liste Règles. La liste Règles n'affiche que les règles appropriées aux variables d'analyse sélectionnées. Si, par exemple, vous sélectionnez des variables d'analyse numériques, seules les règles numériques s'affichent. Si vous sélectionnez une variable chaîne, seules les règles chaîne s'affichent. Si vous n'avez sélectionné aucune variable d'analyse ou si les types de données ont été mélangés, aucune règle ne s'affiche.

Distributions de variables. Les récapitulatifs de distribution affichés dans la liste Variables d'analyse peuvent être basés sur l'ensemble des observations ou sur une analyse des premières observations *n*, comme indiqué dans la zone de texte Observations. Pour mettre à jour les récapitulatifs de distribution, cliquez sur Réanalyser.

Règles de variable croisée de la validation des données

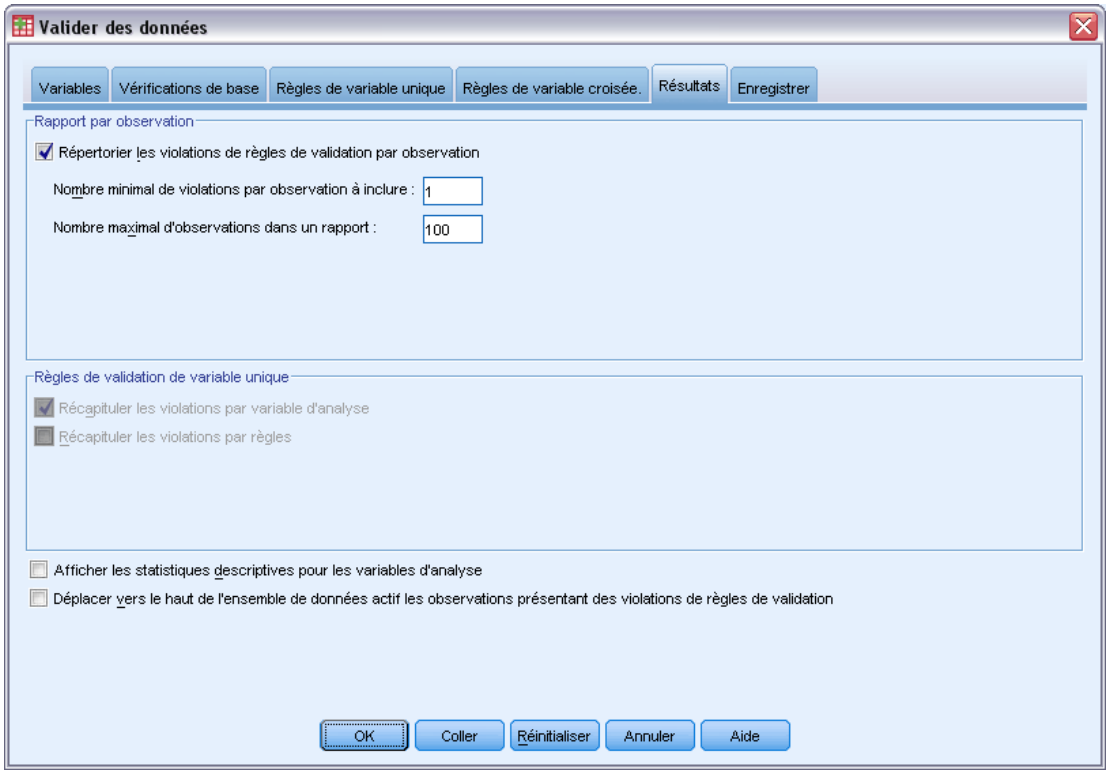
Figure 3-5
Boîte de dialogue Valider les données, onglet Règles des variables croisées



L'onglet Règles de variable croisée affiche les règles de variable croisée disponibles et vous permet de les appliquer aux données. Pour définir d'autres règles de variable croisée, cliquez sur Définir des règles. Pour plus d'informations, reportez-vous à la section Définition des règles de variable croisée dans le chapitre 2 sur p. 6.

Résultats de la validation des données

Figure 3-6
Boîte de dialogue Valider les données, onglet Résultat



Rapport par observation. Si vous avez appliqué des règles de validation de variable unique ou de variable croisée, vous pouvez demander un rapport répertoriant les violations des règles de validation pour les observations individuelles.

- **Nombre minimum de violations.** Cette option indique le nombre minimum de violations de règles nécessaires à l'intégration d'une observation au rapport. Spécifiez un nombre entier positif.
- **Nombre maximum d'observations.** Cette option indique le nombre maximum d'observations incluses dans le rapport d'observations. Entrez un nombre entier positif inférieur ou égal à 1000.

Règles de validation de variable unique. Si vous avez appliqué des règles de validation de variable unique, vous pouvez sélectionner le mode d'affichage et les résultats à afficher.

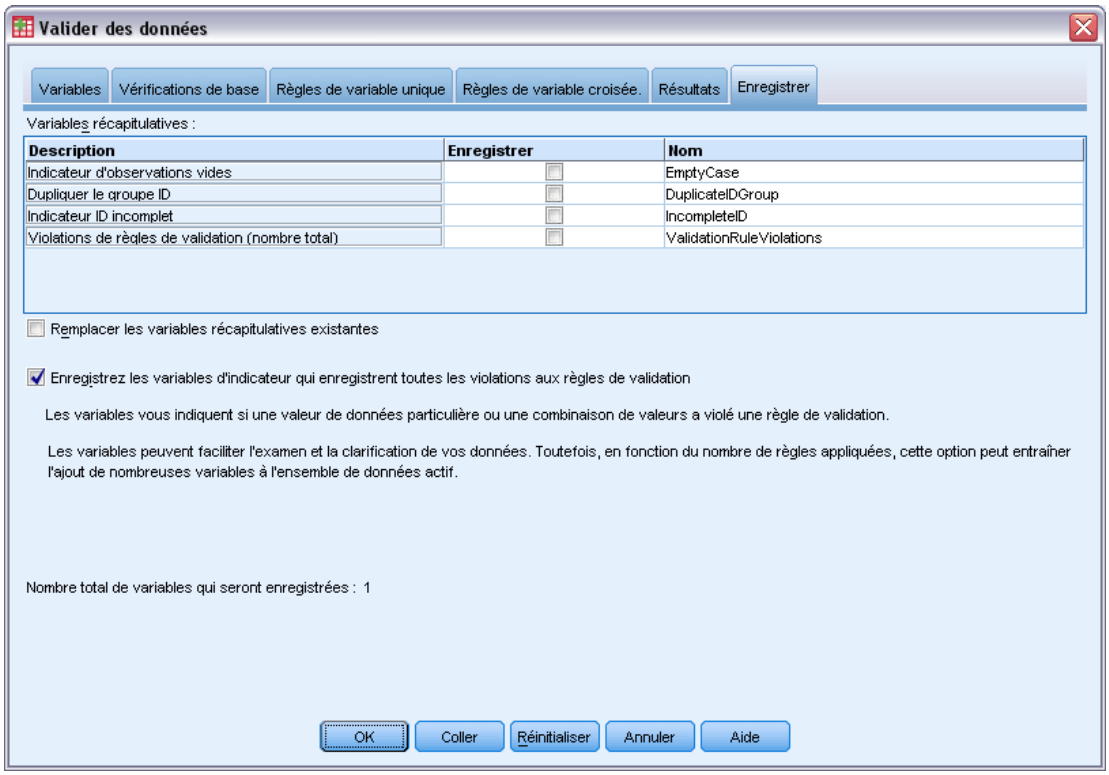
- **Récapituler les violations par variable d'analyse.** Pour chaque variable d'analyse, cette option affiche toutes les règles de validation de variable unique violées et le nombre de valeurs ayant violé chaque règle. Elle répertorie également le nombre total de violations de règles de variable unique pour chaque variable.
- **Récapituler les violations par règles.** Pour chaque règle de validation de variable unique, cette option affiche les variables ayant violé la règle et le nombre de valeurs non valides par variable. Elle répertorie également le nombre total de valeurs ayant violé chaque règle dans l'ensemble des variables.

Afficher les statistiques descriptives. Cette option vous permet de demander les statistiques descriptives pour les variables d'analyse. Un tableau de fréquences est généré pour chaque variable qualitative. Un tableau de statistiques récapitulatives, comprenant la moyenne, l'écart-type, les valeurs mininum et maximum, est généré pour les variables d'échelle.

Déplacer les observations à l'aide des violations des règles de validation. Cette option permet de déplacer les observations contenant des violations de règles de variable unique ou de variable croisée au haut de l'ensemble de données actif pour faciliter la lecture.

Enregistrer la validation des données.

Figure 3-7
Boîte de dialogue Valider les données, onglet Enregistrer



L'onglet Enregistrer vous permet d'enregistrer les variables qui stockent les violations de règles dans l'ensemble de données actif.

Variables récapitulatives. Ces variables individuelles peuvent être enregistrées. Cochez une case pour enregistrer la variable. Les noms des variables par défaut sont fournis, vous pouvez les modifier.

- **Indicateur d'observations vides.** La valeur 1 est attribuée aux observations vides. Toutes les autres observations sont codées 0. Les valeurs de la variable reflètent le champ d'application indiqué sur l'onglet Vérifications de base.

- **Dupliquer le groupe ID.** Le même numéro de groupe est attribué aux observations disposant du même identificateur d'observations (sauf les observations possédant des identificateurs incomplets) Les observations disposant d'identificateurs uniques ou incomplets sont codées 0.
- **Indicateur ID incomplet.** La valeur 1 est attribuée aux observations disposant d'identificateurs vides ou incomplets. Toutes les autres observations sont codées 0.
- **Violations d'une règle de validation.** Il s'agit de l'effectif total par observation de violations des règles de validation de variable unique et de variable croisée.

Remplacer les variables récapitulatives existantes. Les variables enregistrées dans un fichier de données doivent avoir des noms identiques ou remplacer les variables de même nom.

Enregistrer les variables indicatrices. Cette option vous permet d'effectuer un enregistrement complet des violations des règles de validation. Chaque variable correspond à l'application d'une règle de validation et dispose d'une valeur de 1 si l'observation viole la règle et d'une valeur de 0 dans le cas contraire.

Chapitre
4

Préparation automatique des données

La préparation des données pour l'analyse est une des étapes les plus importantes des projets—et généralement, l'une de celles qui prend le plus de temps. La préparation automatique des données (ADP) s'occupe de cette tâche à votre place, analyse vos données, identifie les corrections, supprime les champs problématiques ou inutiles, dérive de nouveaux attributs si nécessaire et améliore les performances grâce à des techniques d'analyse intelligentes. Vous pouvez utiliser l'algorithme en mode complètement **automatique**, le laissant choisir et appliquer les corrections ou vous pouvez utiliser son mode **interactif** qui prévoit les modifications avant qu'elles ne soient effectuées vous laissant libre de les accepter ou de les refuser.

L'utilisation de l'ADP vous permet de préparer facilement et rapidement vos données pour la création de modèle, sans qu'il soit nécessaire de maîtriser les concepts de statistiques utilisés. Les modèles seront alors créés et les scores déterminés plus rapidement ; de plus, l'utilisation de l'ADP améliore la robustesse des processus de modélisation automatique.

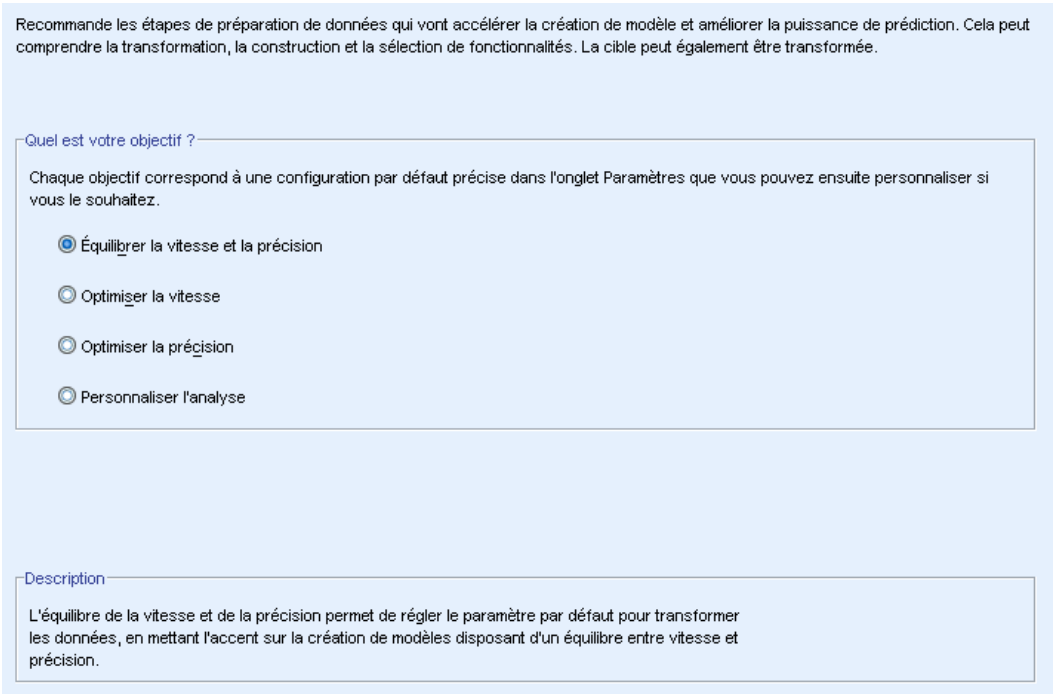
Remarque : lorsque la préparation automatique des données prépare un champ pour l'analyse, elle crée un nouveau champ contenant les ajustements ou les transformations, au lieu de remplacer les valeurs et les propriétés existantes de l'ancien champ. L'ancien champ n'est pas utilisé pour l'analyse, son rôle est défini sur Aucun. Veuillez aussi noter que toute information de valeur manquante spécifiée par l'utilisateur n'est pas transférée dans ces champs nouvellement créés, et que toutes les valeurs manquantes du nouveau champ sont manquantes par défaut.

Exemple : Une compagnie d'assurances disposant de ressources restreintes pour enquêter sur les demandes de remboursement des propriétaires de biens immobiliers, souhaite construire un modèle pour signaler des réclamations suspectes et potentiellement frauduleuses. Avant de construire le modèle, il est nécessaire de préparer les données à l'aide de la préparation automatique des données. La compagnie souhaitant être capable de consulter et modifier les transformations avant de les appliquer, elle utilise la préparation automatique des données de manière interactive. Pour plus d'informations, reportez-vous à la section Utilisation interactive de la préparation automatique des données dans le chapitre 8 sur p. 84.

Un groupe automobile suit les ventes de véhicules automobiles personnels divers. Afin d'être en mesure d'identifier les modèles dont les ventes sont très satisfaisantes et ceux pour lesquels elles le sont moins, des responsables du groupe souhaitent établir une relation entre les ventes de véhicules et les descriptives des véhicules. Ils utilisent la préparation automatique des données pour cette analyse afin de construire des modèles à l'aide des données " avant" et " après " la préparation et de pouvoir en comparer les résultats. Pour plus d'informations, reportez-vous à la section Utilisation automatique de la préparation automatique des données dans le chapitre 8 sur p. 95.

© Copyright SPSS Inc. 1989, 2010

Figure 4-1
Onglet Objectif de la préparation automatique des données



Quel est votre objectif ? La préparation automatique des données recommande des étapes de préparation de données qui amélioreront la vitesse de création de modèles par les autres algorithmes et la puissance de prédiction de ces modèles. Cela peut comprendre la transformation, la construction et la sélection de fonctionnalités. La cible peut également être transformée. Vous pouvez spécifier les priorités de création de modèle sur lesquelles le processus de préparation des données doit se concentrer.

- **Équilibrer la vitesse et la précision.** Cette option prépare les données à accorder la même importance à la vitesse à laquelle les données sont traitées par les algorithmes de création de modèle et à la précision des prévisions.
- **Optimiser la vitesse.** Cette option prépare les données à accorder la priorité à la vitesse à laquelle les données sont traitées par les algorithmes de création de modèle. Lorsque vous travaillez avec de très grands ensembles de données ou que vous recherchez une réponse rapide, sélectionnez cette option.
- **Optimiser la précision.** Cette option prépare les données à accorder la priorité à la précision des prédictions produites par les algorithmes de création de modèle.
- **Analyse personnalisée.** Lorsque vous souhaitez modifier manuellement l'algorithme dans l'onglet Paramètres, sélectionnez cette option. Veuillez noter que ce paramètre est automatiquement sélectionné si vous modifiez ensuite des options dans l'onglet Paramètres qui ne sont pas compatibles avec l'un des autres objectifs.

Pour obtenir une préparation automatique des données

A partir des menus, sélectionnez :
Transformer > Préparer les données pour la modélisation > Automatique...

▶ Cliquez sur Exécuter.

Sinon, vous pouvez :

- Spécifiez un objectif dans l'onglet Objectif.
- spécifiez les affectations de champ dans l'onglet Champs.
- spécifiez les paramètres d'expert dans l'onglet Paramètres.

Pour obtenir une préparation interactive des données

A partir des menus, sélectionnez :
Transformer > Préparer les données pour la modélisation > Interactif...

▶ Cliquez sur Analyser dans la barre d'outils au-dessus de la boîte de dialogue.

▶ Cliquez sur l'onglet Analyse pour consulter les étapes conseillées de préparation des données.

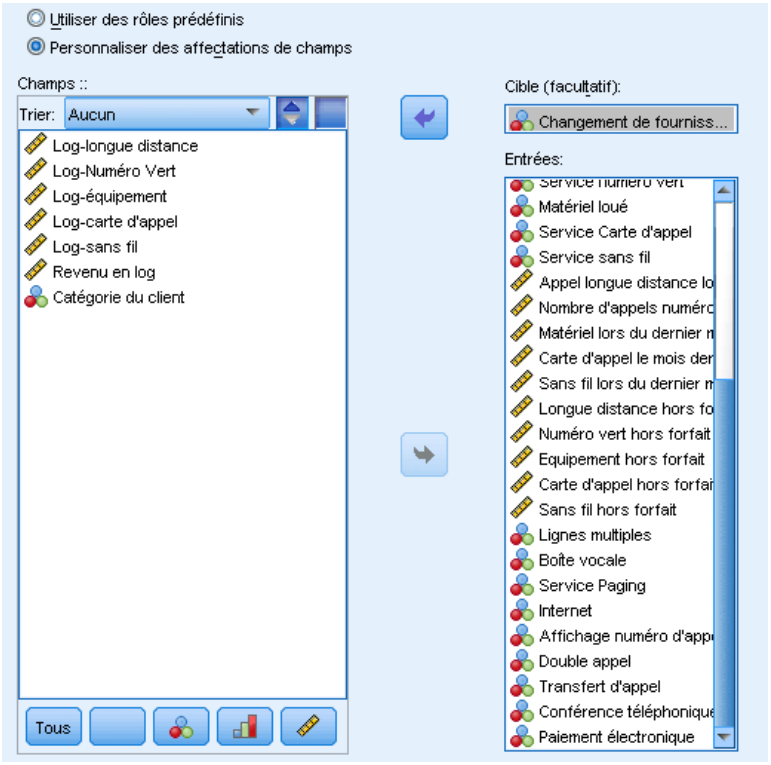
▶ Si elles vous conviennent, cliquez sur Exécuter. Sinon, cliquez sur Effacer l'analyse, modifiez les paramètres souhaités et cliquez sur Analyser.

Sinon, vous pouvez :

- Spécifiez un objectif dans l'onglet Objectif.
- spécifiez les affectations de champ dans l'onglet Champs.
- spécifiez les paramètres d'expert dans l'onglet Paramètres.
- enregistrez les étapes de préparation des données conseillées dans un fichier XML en cliquant sur Enregistrer XML.

Onglet Champs

Figure 4-2
Onglet Champs de la préparation automatique des données



L'onglet Champs indique les champs à préparer pour une analyse ultérieure.

Utiliser des rôles prédéfinis. Cette option utilise des informations sur des champs existants. S'il n'existe qu'un champ avec le rôle Cible, il sera utilisé comme cible ; dans le cas contraire, il n'y aura pas de cible. Tous les champs avec un rôle prédéfini d'Entrée seront utilisés comme entrées. Au moins un champ d'entrée est requis.

Utiliser des affectations de champs personnalisées. Lorsque vous remplacez des rôles de champs en les déplaçant de leur listes par défaut, la boîte de dialogue sélectionne automatiquement cette option. Lors des affectations personnalisées, spécifiez les champs suivants :

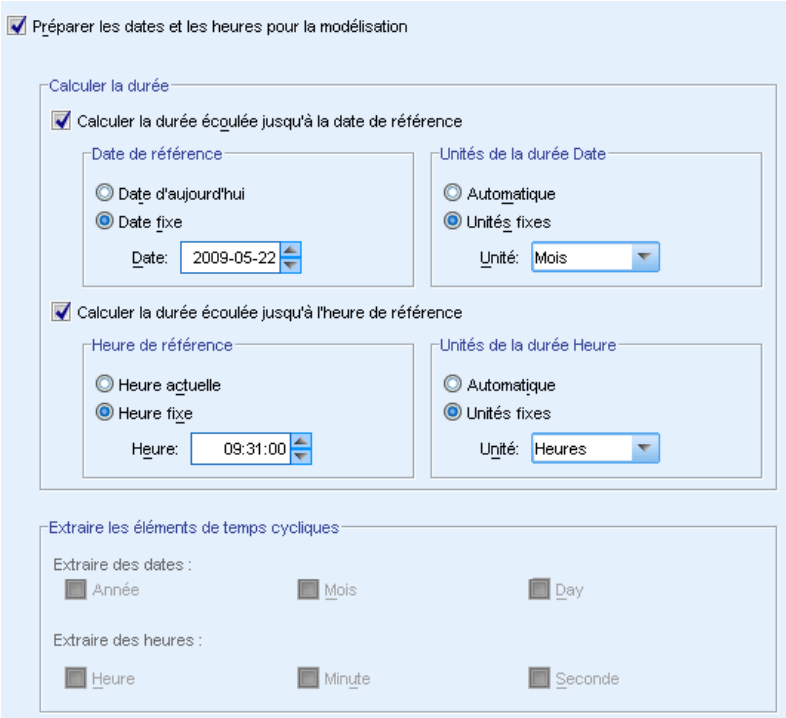
- **Cible (facultative).** Si vous souhaitez créer des modèles nécessitant une cible, sélectionnez le champ cible. Il s'agit de la même action que lorsque l'on définit le rôle du champ sur Cible.
- **Entrées.** Sélectionnez un ou plusieurs champs d'entrée. Il s'agit de la même action que lorsque l'on définit le rôle du champ sur Entrée.

Onglet Paramètres

L'onglet Paramètres contient plusieurs groupes de paramètres différents que vous pouvez modifier pour affiner le traitement des données par l'algorithme. Si vous modifiez les paramètres par défaut et que ces modifications sont incompatibles avec les autres objectifs, l'onglet Objectif est automatiquement mis à jour pour sélectionner l'option Personnaliser l'analyse.

Préparer les dates & les heures

Figure 4-3
Paramètres Dates & Heures de la préparation automatique des données



De nombreux algorithmes de modélisation ne peuvent pas traiter directement les informations sur la date et l'heure. Ces paramètres vous permettent de calculer de nouvelles données de durée qui peuvent être utilisées comme entrées de modèle à partir des dates et des heures de vos données existantes. Les champs contenant les dates et les heures doivent être prédéfinis à l'aide des types de stockage de dates et d'heures. Il n'est pas recommandé de définir les champs de date et d'heure d'origine comme entrées de modèle après la préparation automatique des données.

Préparer les dates et les heures pour la modélisation. En déselectionnant cette option, vous désactivez tous les autres contrôles Préparer les dates et les heures, tout en conservant les sélections.

Calculer la durée écoulée jusqu'à la date de référence. Cette option génère le nombre d'années/mois/jours depuis une date de référence pour chaque variable qui contient des dates.

- **Date de référence.** Spécifier la date à partir de laquelle la durée sera calculée en fonction des informations sur la date dans les données d'entrée. Sélectionner Date d'aujourd'hui signifie que la date du système actuelle est toujours utilisée lorsque l'ADP est exécuté. Pour utiliser une date spécifique, sélectionnez Date fixe et saisissez la date désirée.
- **Unités de la durée Date.** Spécifier si l'ADP doit décider automatiquement de l'unité de la durée Date ou choisir dans les unités fixes des Années, Mois ou Jours.

Calculer la durée écoulée jusqu'à l'heure de référence. Cette option génère le nombre d'heures/minutes/secondes depuis une heure de référence pour chaque variable qui contient des heures.

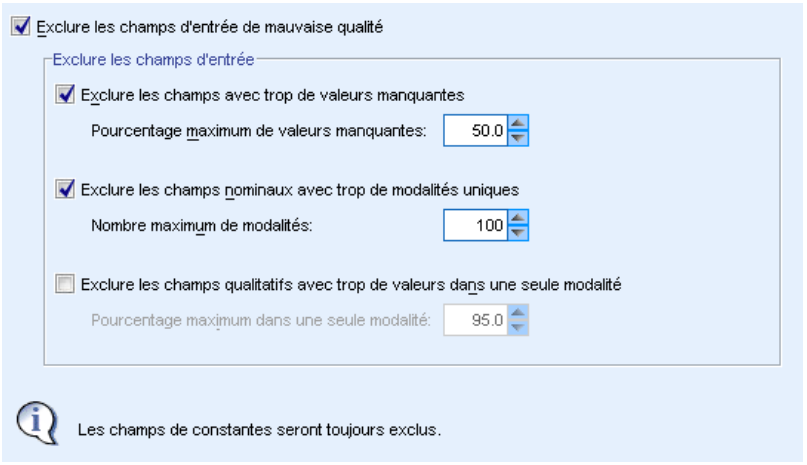
- **Heure de référence.** Spécifier l'heure à partir de laquelle la durée sera calculée en fonction des informations sur l'heure dans les données d'entrée. Sélectionner Heure actuelle signifie que l'heure du système actuelle est toujours utilisée lorsque l'ADP est exécuté. Pour utiliser une heure spécifique, sélectionnez Heure fixe et saisissez l'heure désirée.
- **Unités de la durée Heure.** Spécifier si l'ADP doit décider automatiquement de l'unité de la durée Heure ou choisir dans les unités fixes des Heures, Minutes ou Secondes.

Extraire les éléments de temps cycliques. Utilisez ces paramètres pour scinder un champ de date ou d'heure en un ou plusieurs autres champs. Par exemple, si vous sélectionnez les trois cases de date, le champ de date d'entrée "1954-05-23" est divisé en trois champs : 1954, 5 et 23, chacun utilisant le suffixe défini dans le panneau Noms des champ et le champ de date d'origine est ignoré.

- **Extraire des dates.** Pour chaque entrée de date, spécifiez si vous souhaitez extraire des années, des mois, des jours ou une des combinaisons possibles.
- **Extraire des heures.** Pour chaque entrée de date, spécifiez si vous souhaitez extraire des heures, des minutes ou des secondes ou une des combinaisons possibles.

Exclure les champs

Figure 4-4
Paramètres Exclure les champs de la préparation automatique des données



Les données de mauvaise qualité peuvent affecter la précision de vos prédictions. Par conséquent, vous pouvez spécifier le niveau de qualité acceptable des descriptives d'entrée. Tous les champs constants ou avec 100% de valeurs manquantes sont automatiquement exclus.

Exclure les champs d'entrée de mauvaise qualité. En déselectionnant cette option, vous désactivez tous les autres contrôles Exclure les champs, tout en conservant les sélections.

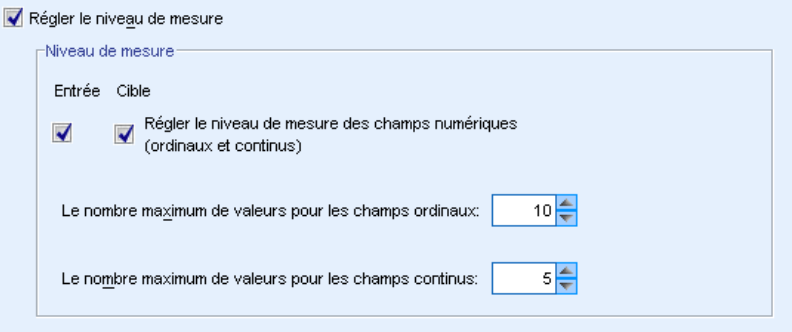
Exclure les champs avec trop de valeurs manquantes. Les champs ayant plus que le pourcentage spécifié de valeurs manquantes sont supprimés de l'analyse. Définissez une valeur supérieure ou égale à 0, ce qui revient à déselectionner cette option, et inférieure ou égale à 100, puisque les champs qui ne contiennent que des valeurs manquantes sont exclus automatiquement. La valeur par défaut est 50.

Exclure les champs nominaux avec trop de modalités uniques. Les champs nominaux ayant plus que le nombre spécifié de modalités sont supprimés de l'analyse. Spécifiez un nombre entier positif. La valeur par défaut est 100. Cette option est utile pour supprimer automatiquement de la modélisation les champs contenant des informations d'enregistrement unique, tels que l'ID, l'adresse ou le nom.

Exclure les champs qualitatifs avec trop de valeurs dans une seule modalité. Les champs ordinaux et nominaux avec une modalité contenant plus que le pourcentage spécifié d'enregistrements sont supprimés de l'analyse. Définissez une valeur supérieure ou égale à 0, ce qui revient à déselectionner cette option, et inférieure ou égale à 100, puisque les champs constants sont exclus automatiquement. La valeur par défaut est 95.

Régler les mesures

Figure 4-5
Paramètres Régler les mesures de la préparation automatique des données



Régler le niveau de mesure. En déselectionnant cette option, vous désactivez tous les autres contrôles Régler les mesures, tout en conservant les sélections.

Niveau de mesure. Spécifier si le niveau de mesure des champs continus avec « trop peu » de valeurs peut être réglé sur ordinal et si les champs ordinaux avec « trop » de valeurs peuvent être réglés sur continu.

- **Le nombre maximum de valeurs pour les champs ordinaux.** Les champs ordinaux ayant plus que le nombre spécifié de modalités sont reconvertis en champs continus. Spécifiez un nombre entier positif. La valeur par défaut est 10. Cette valeur doit être supérieure ou égale au nombre minimum de valeurs pour les champs continus.
- **Le nombre minimum de valeurs pour les champs continus.** Les champs continus ayant moins que le nombre spécifié de valeurs uniques sont reconvertis en champs ordinaux. Spécifiez un nombre entier positif. La valeur par défaut est 5. Cette valeur doit être inférieure ou égale au nombre maximum de valeurs pour les champs ordinaux.

Améliorer la qualité des données

Figure 4-6
Paramètres Améliorer la qualité des données de la préparation automatique des données

Préparer les champs pour améliorer la qualité des données. En déselectionnant cette option, vous désactivez tous les autres contrôles Améliorer la qualité des données, tout en conservant les sélections.

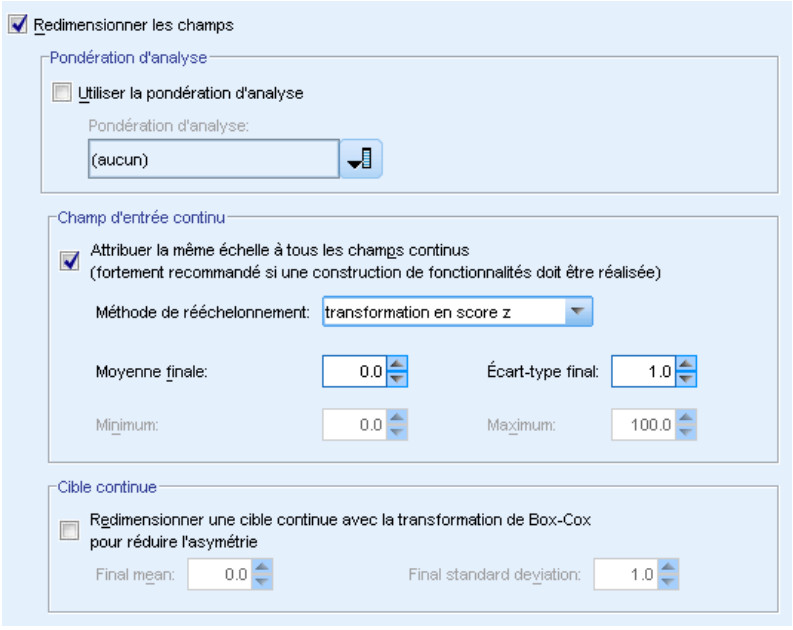
Traitement des valeurs éloignées. Spécifier s'il faut remplacer les valeurs éloignées des entrées et des cibles. Si oui, spécifier un critère de césure des valeurs éloignées, mesuré en écarts-types et une méthode de remplacement des valeurs éloignées. Les valeurs éloignées peuvent être remplacées soit en les tronquant (définies sur la valeur de césure) ou en les définissant comme valeurs manquantes. Les valeurs éloignées définies comme valeurs manquantes suivent les paramètres de traitement des valeurs manquantes sélectionnées ci-dessous.

Remplacer les valeurs manquantes. Spécifier s'il faut remplacer les valeurs manquantes des champs continus, nominaux ou ordinaux.

Réorganiser les champs nominaux. Sélectionner cette option pour recoder les valeurs des champs nominaux (ensemble) de la plus petite modalité (la moins utilisée) à la plus grande (la plus utilisée). Les valeurs des nouveaux champs démarrent à 0, 0 étant la modalité la moins fréquente. Remarque : le nouveau champ doit être numérique même si le champ d'origine est une chaîne. Par exemple, si les valeurs d'un champ nominal sont "A", "A", "A", "B", "C", "C", la préparation automatique des données recodent "B" en 0, "C" en 1, et "A" en 2.

Rééchelonner les champs

Figure 4-7
Paramètres Rééchelonner les champs de la préparation automatique des données



Rééchelonner les champs. En déselectionnant cette option, vous désactivez tous les autres contrôles Rééchelonner les champs, tout en conservant les sélections.

Pondération d'analyse. Cette variable contient des pondérations (de régression ou d'échantillon) d'analyse. Les pondérations d'analyse sont utilisées pour représenter les différences de variance dans les niveaux du champ cible. Sélectionnez un champ continu.

Champs d'entrée continus. Cela normalisera les champs d'entrée continus avec une transformation en score z ou une transformation min/max. Le rééchelonnement des entrées est particulièrement utile lorsque vous sélectionnez l'option Exécuter la construction des fonctionnalités dans les paramètres Sélectionner et Construire.

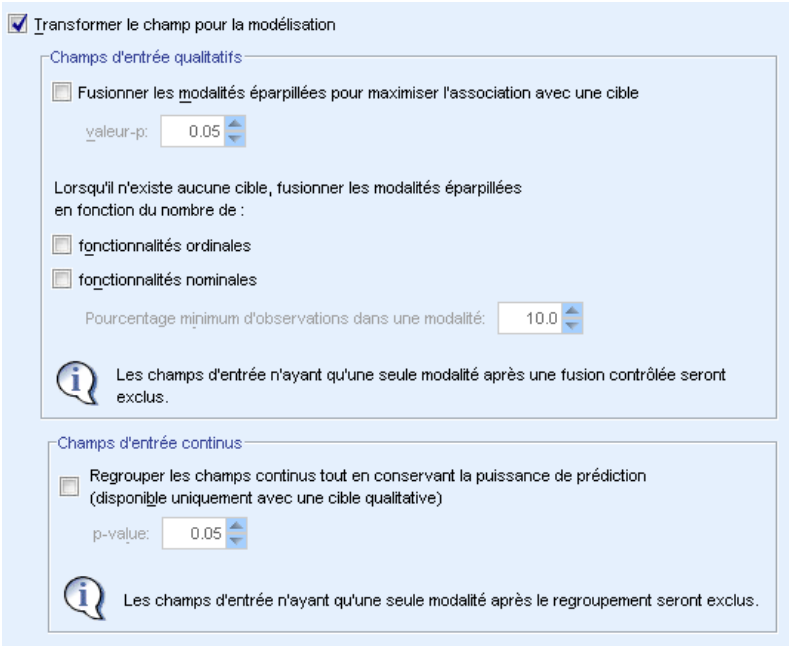
- **Transformation en score z.** Avec la moyenne et l'écart-type observés utilisés comme estimations des paramètres de population, les champs sont standardisés puis les scores z sont mappés aux valeurs correspondantes d'une distribution normale avec la moyenne finale et l'écart-type final spécifiés. Spécifiez un nombre pour la moyenne finale et un nombre positif pour l'écart-type final. Les valeurs par défaut sont 0 et 1 respectivement, ce qui correspond au rééchelonnement standardisé.
- **Transformation min/max.** Avec la transformation minimum et maximum observée qui est utilisée comme estimations des paramètres de population, les champs sont mappés aux valeurs correspondantes d'une distribution uniforme avec la transformation Minimum et Maximum spécifiée. Spécifiez les nombres avec la transformation Maximum supérieure à la transformation Minimum.

Cible continue. Cela transforme une cible continue utilisant la transformation de Box-Cox en un champ ayant une distribution à peu près normale avec la moyenne finale et l'écart-type final spécifiés. Spécifiez un nombre pour la moyenne finale et un nombre positif pour l'écart-type final. Les valeurs par défaut sont 0 et 1 respectivement.

Remarque : Si une cible a été transformée par l'ADP, les modèles en résultant créés à l'aide de la cible transformée évaluent les unités transformées. Afin d'interpréter et d'utiliser les résultats, vous devez reconvertir la valeur observée dans son échelle d'origine. Pour plus d'informations, reportez-vous à la section Rétablir les scores sur p. 45.

Transformer les champs

Figure 4-8
Paramètres Transformer les champs de la préparation automatique des données



Pour améliorer la puissance de prédiction de vos données, vous pouvez transformer les champs d'entrée.

Transformer le champ pour la modélisation. En déselectionnant cette option, vous désactivez tous les autres contrôles Transformer les champs, tout en conservant les sélections.

Champs d'entrée qualitatifs

- **Fusionner les modalités éparpillées pour optimiser l'association avec une cible.** Sélectionnez cette option pour créer un modèle plus petit en réduisant le nombre de champs à traiter en association avec la cible. Les modalités similaires sont identifiées en fonction de la relation entre l'entrée et la cible. Les modalités ne différant pas de manière significative, c'est-à-dire ayant une valeur *p* supérieure à la valeur spécifiée, sont fusionnées. Spécifiez une valeur supérieure à 0 et inférieure ou égale à 1. Si toutes les modalités sont fusionnées en une

modalité, les versions d'origine et dérivées du champ sont exclues d'une analyse ultérieure car elles n'ont pas de valeur de variable prédite.

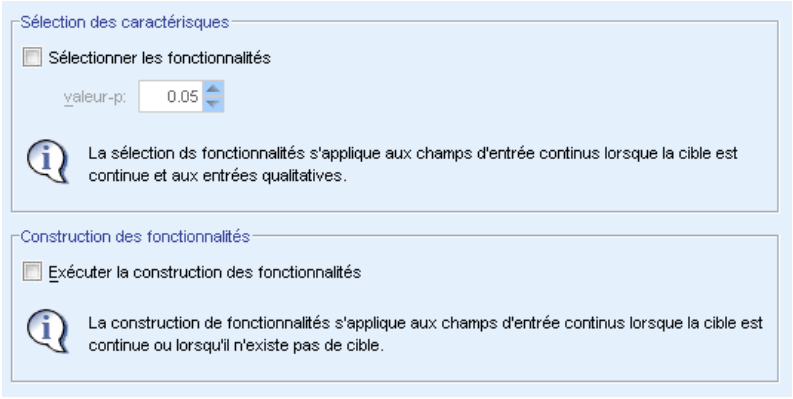
- **Lorsqu'il n'existe aucune cible, fusionner les modalités éparpillées en fonction de leur nombre.** Si l'ensemble de données n'a pas de cible, vous pouvez choisir de fusionner les modalités éparpillées des champs ordinaux et nominaux. La méthode d'effectifs égaux est utilisée pour fusionner les modalités ayant moins que le pourcentage minimum spécifié du nombre total d'enregistrements. Spécifiez une valeur supérieure ou égale à 0 et inférieure ou égale à 100. La valeur par défaut est 10. La fusion s'arrête lorsqu'il n'y a plus de modalités avec moins que le pourcentage d'observations minimum spécifié ou lorsqu'il ne reste que deux modalités.

Champs d'entrée continus. Si l'ensemble de données comprend une cible qualitative, vous pouvez regrouper les entrées continues ayant de fortes associations pour améliorer les performances du traitement. Les regroupements sont créés en fonction des propriétés des « sous-ensembles homogènes » qui sont identifiés avec la méthode de Scheffé qui utilise la valeur de *p* comme valeur alpha de la valeur critique pour déterminer les sous-ensembles homogènes. Définissez une valeur supérieure à 0 et inférieure ou égale à 1. La valeur par défaut est 0,05. Si l'opération de regroupement génère un regroupement unique pour un champ spécifique, les versions d'origine et regroupées du champ sont exclues car elles n'ont pas de valeur de variable prédite.

Remarque : Le regroupement dans l'ADP est différent du regroupement optimal. Le regroupement optimal utilise des informations d'entropie pour convertir un champ continu en un champ qualitatif ; il doit trier les données et les stocker dans la mémoire. L'ADP utilise des sous-ensembles homogènes pour regrouper un champ continu. Cela signifie que le regroupement ADP n'a pas besoin de trier les données et ne stocke pas toutes les données dans une mémoire. L'utilisation de la méthode des sous-ensembles homogènes pour regrouper un champ continu signifie que le nombre de modalités après le regroupement est toujours inférieur ou égal au nombre de modalités dans la cible.

Sélectionner et Construire

Figure 4-9
Paramètres Sélectionner et Construire de la préparation automatique des données



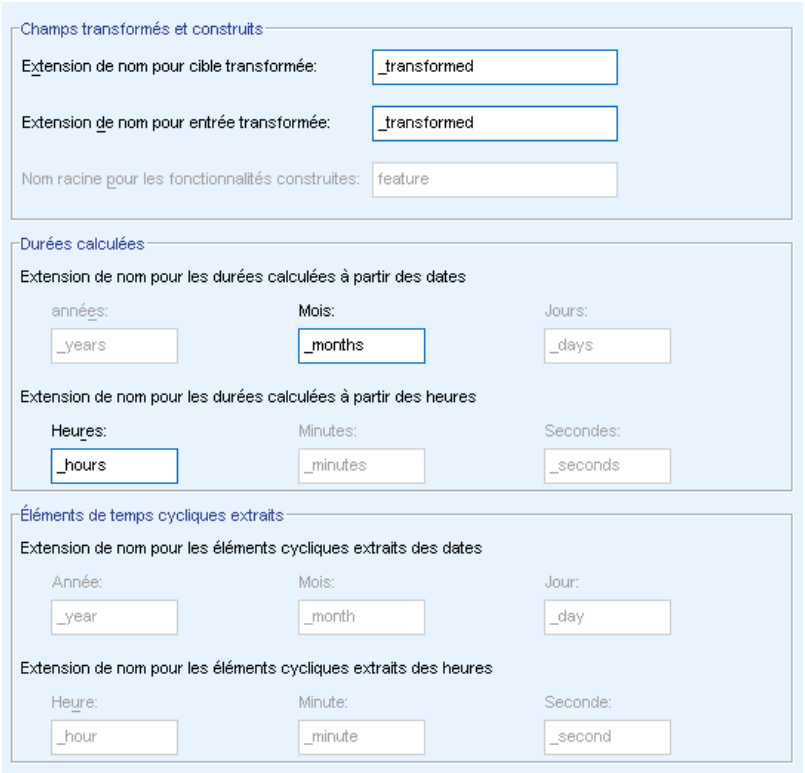
Pour améliorer la puissance de prédiction de vos données, vous pouvez construire de nouveaux champs basés sur les champs existants.

Exécuter la sélection des descriptives. Une entrée continue est supprimée de l'analyse si la valeur de *p* pour sa corrélation avec la cible est supérieure à la valeur de *p* spécifiée.

Exécuter la construction des descriptives. Sélectionnez cette option pour dériver de nouvelles descriptives d'une combinaison de plusieurs descriptives existantes. Les anciennes descriptives ne sont pas utilisées dans l'analyse ultérieure. Cette option s'applique uniquement aux descriptives d'entrée continues où la cible est continue ou lorsqu'il n'y a pas de cible.

Nom de champs

Figure 4-10
Paramètres Nommer les champs de la préparation automatique des données



Pour identifier facilement les descriptives nouvelles et transformées, l'ADP crée et applique de nouveaux noms, préfixes ou suffixes de base. Vous pouvez modifier ces noms pour qu'ils soient plus adaptés à vos propres besoins et données.

Champs transformés et construits. Spécifiez les extensions de nom à appliquer aux champs cibles et d'entrées transformés.

En outre, spécifiez le nom du préfixe à appliquer aux descriptives construites à l'aide des paramètres Sélectionner et Construire. Le nouveau nom est créé en ajoutant un suffixe numérique à ce nom racine du préfixe. Le format du nombre dépend du nombre de nouvelles descriptives dérivées, par exemple :

- si 1 à 9 descriptives sont construites, elles seront nommées : descriptive1 à descriptive9.

- si 10 à 99 descriptives sont construites, elles seront nommées : descriptive01 à descriptive99.
- si 100 à 999 descriptives sont construites, elles seront nommées : descriptive001 à descriptive999, etc.

Cela permet que les descriptives construites soient triées dans un ordre cohérent quel que soit leur nombre.

Durée calculée à partir des dates et des heures. Spécifier les extensions de nom à appliquer aux durées calculées à partir des dates et des heures.

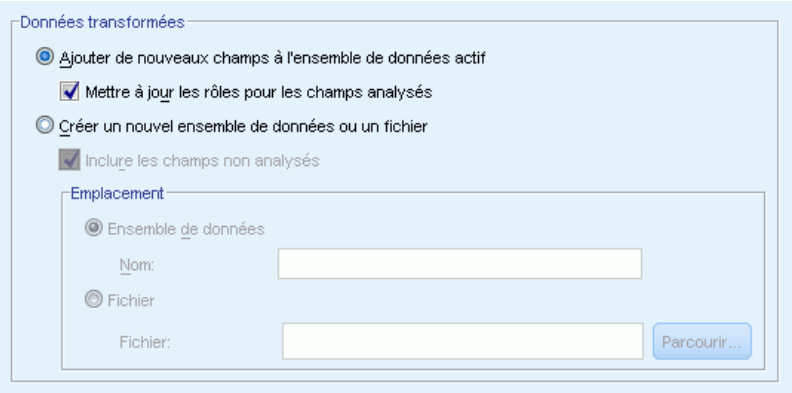
Éléments cycliques extraits de dates et des heures. Spécifier les extensions de nom à appliquer aux éléments cycliques extraits des dates et des heures.

Appliquer et enregistrer les transformations

Selon que vous utilisez la boîte de dialogue de préparation automatique ou interactive des données, les paramètres d'application et d'enregistrement des transformations des données diffèrent légèrement.

Paramètres Appliquer les transformations de la préparation automatique des données

Figure 4-11
Paramètres Appliquer les transformations de la préparation automatique des données



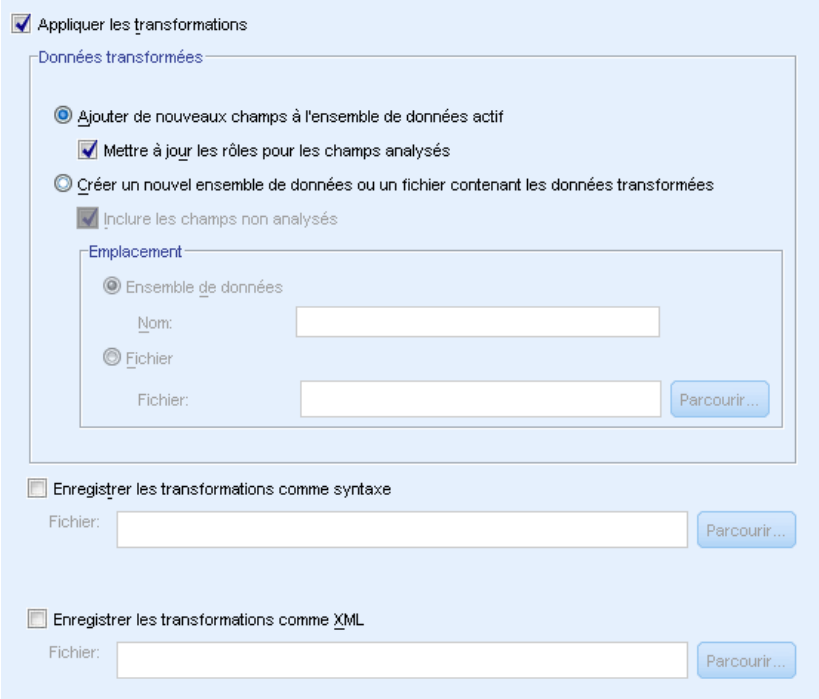
Données transformées. Ces paramètres spécifient l'emplacement de l'enregistrement des données transformées.

- **Ajouter de nouveaux champs à l'ensemble de données actif.** Tous les champs créés par la préparation automatique des données sont ajoutés comme nouveaux champs à l'ensemble de données actif. Mettre à jour les rôles pour les champs analysés définira le rôle sur Aucun pour tous les champs exclus d'une analyse ultérieure par la préparation automatique des données.
- **Créer un nouvel ensemble de données ou un fichier contenant les données transformées.** Les champs recommandés par la préparation automatique des données sont ajoutés à un nouvel ensemble de données ou à un fichier. Inclure les champs non analysés ajoute les champs dans l'ensemble de données d'origine qui n'ont pas été spécifiés dans l'onglet Champs du nouvel ensemble de données. Cette option est utile pour transférer vers le nouvel ensemble de

données les champs contenant des informations non utilisées dans la modélisation, telles que l'ID, l'adresse ou le nom.

Paramètre Appliquer et Enregistrer de la préparation automatique des données

Figure 4-12
Paramètre Appliquer et Enregistrer de la préparation automatique des données



Le groupe des données transformées est le même que celui de la préparation interactive des données. Les options supplémentaires suivantes sont disponibles pour la préparation automatique des données :

Appliquer les transformations. Dans les boîtes de dialogue de la Préparation automatique des données, déselectionner cette option revient à désactiver tous les autres contrôles Appliquer et Enregistrer, tout en conservant les sélections.

Enregistrer les transformations comme syntaxe. Cette option enregistre les transformations recommandées comme syntaxe de commande dans un fichier externe. La boîte de dialogue Préparation interactive des données ne contient pas ce contrôle car elle collera les transformations comme syntaxe de commande dans la fenêtre de syntaxe si vous cliquez sur Coller.

Enregistrer les transformations comme XML. Cette option enregistre les transformations recommandées au format XML dans un fichier externe, qui peut être fusionné avec le modèle PMML à l'aide de la commande `TMS MERGE` ou appliqué à un autre ensemble de données à l'aide de la commande `TMS IMPORT`. La boîte de dialogue Préparation interactive des données ne contient pas ce contrôle car elle enregistrera les transformations au format XML si vous cliquez sur Enregistrer XML dans la barre d'outils au-dessus de la boîte de dialogue.

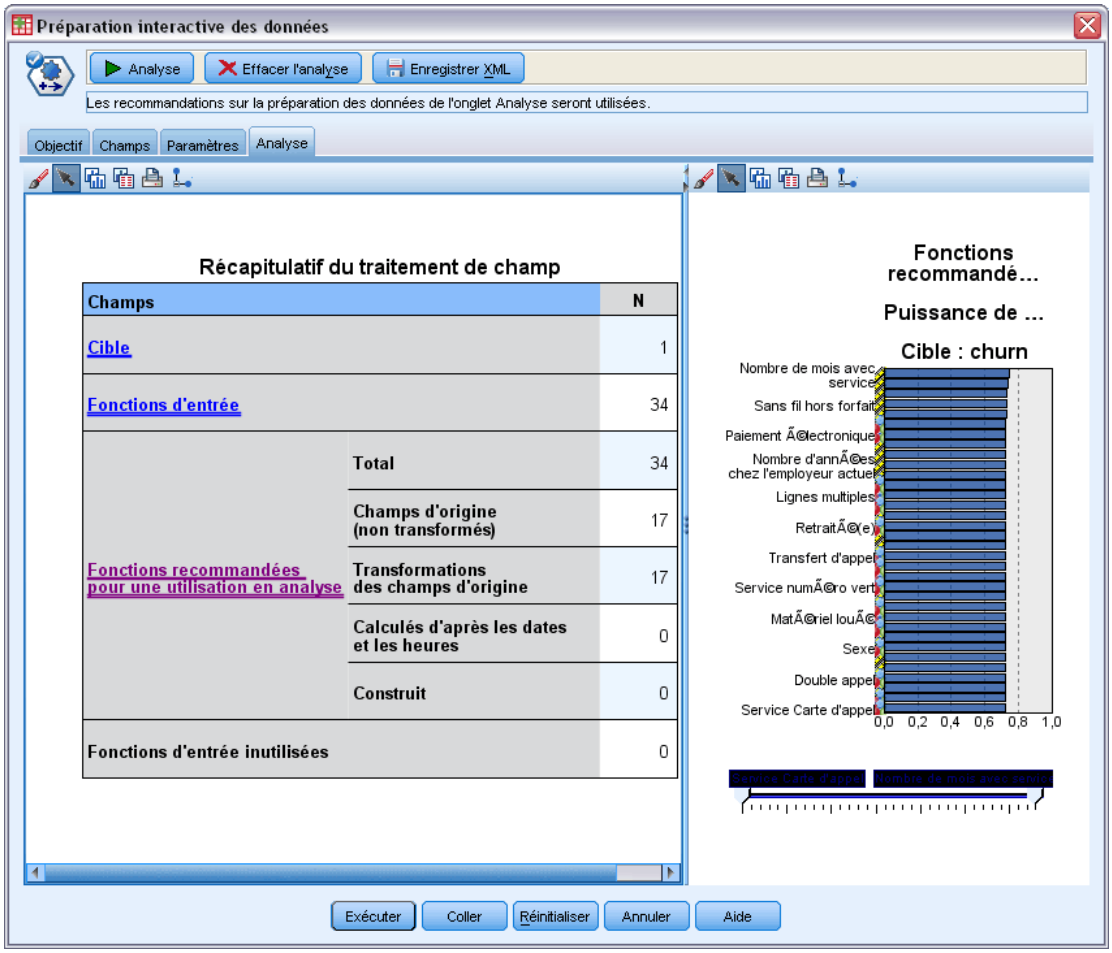
Onglet Analyse

Remarque : L'onglet Analyse est utilisé dans la boîte de dialogue Préparation interactive des données pour vous permettre de passer en revue les transformations recommandées. La boîte de dialogue de préparation automatique des données ne comprend pas cette étape.

- Lorsque les paramètres d'ADP vous conviennent, y compris les modifications effectuées dans les onglets Objectif, Champs et Paramètres, cliquez sur Analyser les données. L'algorithme applique les paramètres aux entrées de données et affiche les résultats dans l'onglet Analyse.

L'onglet Analyse contient à la fois des résultats en tableaux et des résultats graphiques qui résument le traitement de vos données et affichent les recommandations sur la façon de modifier ou d'améliorer les données pour l'évaluation. Vous pouvez ensuite revoir puis accepter ou refuser ces recommandations.

Figure 4-13
Onglet Analyse de la préparation automatique des données



L'onglet Analyse est composé de deux panneaux, la vue principale à gauche et la vue liée, ou auxiliaire, à droite. Il existe trois vues principales :

- Récapitulatif de traitement des champs (par défaut). Pour plus d'informations, reportez-vous à la section Récapitulatif de traitement des champs sur p. 34.
- Champs. Pour plus d'informations, reportez-vous à la section Champs sur p. 35.
- Récapitulatif des actions. Pour plus d'informations, reportez-vous à la section Récapitulatif des actions sur p. 37.

Il existe quatre vues liées/auxiliaires :

- Puissance de prédiction (par défaut). Pour plus d'informations, reportez-vous à la section Puissance de prédiction sur p. 38.
- Tableau des champs. Pour plus d'informations, reportez-vous à la section Tableau des champs sur p. 39.
- Détails des champs. Pour plus d'informations, reportez-vous à la section Détails des champs sur p. 40.
- Détails des actions. Pour plus d'informations, reportez-vous à la section Détails des actions sur p. 42.

Liens entre les vues

Dans la vue principale, le texte souligné dans les tableaux contrôle ce qui apparaît dans la vue liée. Si vous cliquez sur ces parties de texte, vous obtenez des détails sur un champ, un ensemble de champs ou une étape de traitement spécifique. Le lien que vous avez sélectionné en dernier apparaît en une couleur plus foncée qui permet d'identifier la connection entre les contenus des deux panneaux de la vue.

Réinitialisation des vues

Pour afficher de nouveau les recommandations d'analyse d'origine et abandonner les modifications effectuées sur les vues Analyse, cliquez sur Réinitialiser au bas du panneau de la vue principale.

Récapitulatif de traitement des champs

Figure 4-14
Récapitulatif de traitement des champs

Récapitulatif de traitement des champs

Champs		N
Cible		1
Caractéristiques d'entrée		9
Caractéristiques recommandées pour l'analyse	Total	8
	Champs d'origine (non transformés)	1
	Transformations des champs d'origine	7
	Calculé à partir des dates et des heures	0
	Construits	0
Caractéristiques d'entrée non utilisées		1

Le tableau Récapitulatif de traitement des champs fournit un instantané de l'impact du traitement général projeté, y compris les modifications de l'état des descriptives et le nombre de descriptives construites.

Veuillez noter que le modèle est bien construit, et que par conséquent il n'y a pas de mesure ou de diagramme de la modification de la puissance prédictive générale avant et après la préparation des données. Par contre, vous pouvez afficher les diagrammes de la puissance de prédiction des variables indépendantes prédites recommandées.

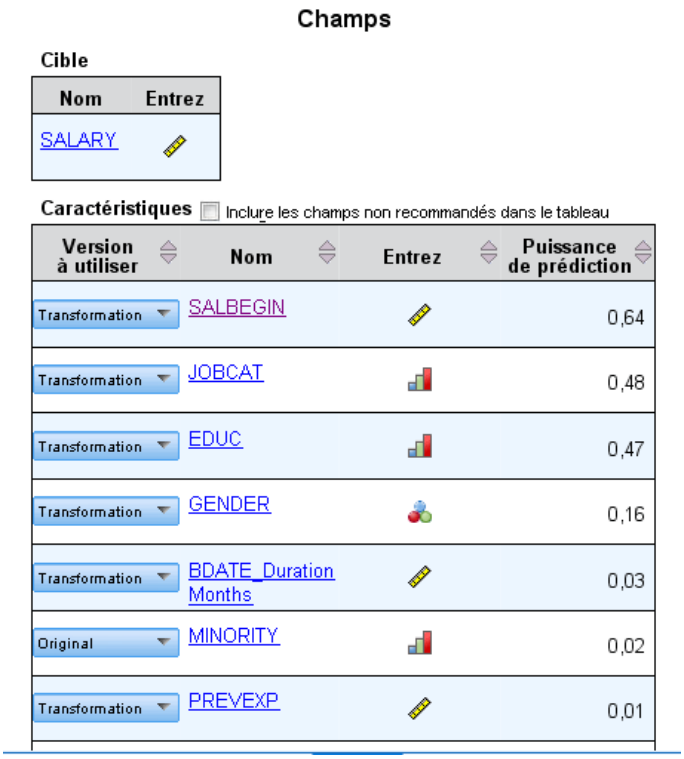
Le tableau affiche les informations suivantes :

- le nombre de champs cibles.
- Le nombre de variables prédites (d'entrée) d'origine.
- Les valeurs prédites recommandées pour l'analyse et la modélisation. Cela comprend le nombre total de champs recommandés ; le nombre de champs d'origine non transformés recommandés ; le nombre de champs transformés recommandés (sans les versions intermédiaires des champs, champs dérivés des valeurs prédites de date/heure et valeurs prédites construites) ; le nombre de champs dérivés recommandés des champs date/heure ; et le nombre de valeurs prédites construites.
- Le nombre de valeurs prédites d'entrée non recommandées quelle que soit leur forme, que ce soit sous leur forme d'origine, comme champ dérivé, ou comme entrée d'une valeur prédite construite.

Lorsque des informations sur les champs sont soulignées, cliquez pour afficher plus de détails dans une vue liée. Les détails de la Cible, des Descriptives d'entrée, et des Descriptives d'entrée non utilisées apparaissent dans la vue liée Tableau des champs. Pour plus d'informations, reportez-vous à la section Tableau des champs sur p. 39. Les Descriptives recommandées pour l'analyse apparaissent dans la vue liée Puissance de prédiction. Pour plus d'informations, reportez-vous à la section Puissance de prédiction sur p. 38.

Champs

Figure 4-15
Champs



La vue principale Champs affiche les champs traités et si l'ADP recommande de les utiliser dans les modèles en aval. Vous pouvez ignorer les recommandations pour n'importe quel champ ; par exemple, exclure les descriptives construites ou inclure les descriptives que l'ADP recommande d'exclure. Si un champ a été transformé, vous pouvez décider d'accepter ou non la transformation suggérée ou d'utiliser ou non la version d'origine.

La vue Champs est composée de deux tableaux, un pour la cible et un pour les valeurs prédites qui ont été traitées ou créées.

Tableau Cible

Le tableau Cible n'apparaît que si une cible est définie dans les données.

Ce tableau contient deux colonnes :

- **Nom.** C'est le nom ou l'étiquette du champ cible ; le nom d'origine est toujours utilisé, même si le champ a été transformé.
- **Niveau de mesure.** Ceci affiche l'icône représentant le niveau de mesure. Placez la souris sur l'icône pour afficher une étiquette (continu, ordinal, nominal, etc.) qui décrit les données.

  Si la cible a été transformée, la colonne Niveau de mesure reflète la version transformée finale. *Remarque* : vous ne pouvez pas désactiver les transformations pour la cible.

Tableau des valeurs prédites

Le tableau Valeurs prédites est affiché en permanence. Chaque ligne du tableau représente un champ. Les lignes sont triées par défaut dans l'ordre décroissant de la puissance de prédiction.

Pour les descriptives ordinaires, le nom d'origine est toujours utilisé comme nom de ligne. Les versions d'origine et dérivée des champs date/heure apparaissent dans le tableau (dans des lignes séparées) ; le tableau contient également les valeurs prédites construites.

Veuillez noter que les versions transformées des champs apparaissant dans le tableau représentent toujours les versions finales.

Par défaut, seuls les champs recommandés sont affichés dans le tableau des valeurs prédites. Pour afficher les champs restants, sélectionnez la boîte de dialogue Inclure les champs non recommandés dans le tableau au-dessus du tableau ; ces champs sont ensuite affichés au bas du tableau.

Le tableau contient les colonnes suivantes :

- **Version à utiliser.** Affiche une liste déroulante qui contrôle l'utilisation d'un champ en aval et s'il faut utiliser les transformations recommandées. Par défaut, la liste déroulante reflète les recommandations.

  Pour les valeurs prédites ordinaires qui ont été transformées, la liste déroulante contient trois choix : Transformée, D'origine, et Ne pas utiliser.

  Pour les valeurs prédites non transformées ordinaires, les choix sont : D'origine et Ne pas utiliser.

  Pour les champs dérivés date/heure et les valeurs prédites construites, les choix sont : Transformée et Ne pas utiliser.

  Pour les champs de date d'origine, la liste déroulante est désactivée et définie sur Ne pas utiliser.

  Remarque : Pour les valeurs prédites contenant à la fois les versions d'origine et transformées, passer des versions d'origine aux versions transformées met automatiquement à jour les paramètres Niveau de mesure et Puissance de prédiction pour ces descriptives.
- **Nom.** Chaque nom de champ est un lien. Cliquez sur un nom pour afficher plus d'informations sur le champ dans la vue liée. Pour plus d'informations, reportez-vous à la section Détails des champs sur p. 40.

- **Niveau de mesure.** Affiche l'icône représentant le type de données ; passez la souris sur l'icône pour afficher une étiquette (continu, ordinal, nominal, etc.) qui décrit les données.
- **Puissance de prédiction.** La puissance de prédiction est affichée uniquement pour les champs recommandés par l'ADP. Cette colonne n'apparaît pas si aucune cible n'est définie. La puissance de prédiction est comprise entre 0 et 1, les valeurs les plus élevées, indiquant des variables prédites de "meilleure" qualité. En général, la puissance de prédiction est utile pour comparer les variables prédites dans une analyse ADP, mais les valeurs de la puissance de prédiction ne peuvent être comparées entre des analyses différentes.

Récapitulatif des actions

Figure 4-16
Récapitulatif des actions

Récapitulatif des actions

Action
Champs de texte
Caractéristiques de date et d'heure
Filtrage des caractéristiques
Vérifier le type
valeurs éloignées
Valeurs manquantes
Cible
Caractéristiques qualitatives
Caractéristiques continues

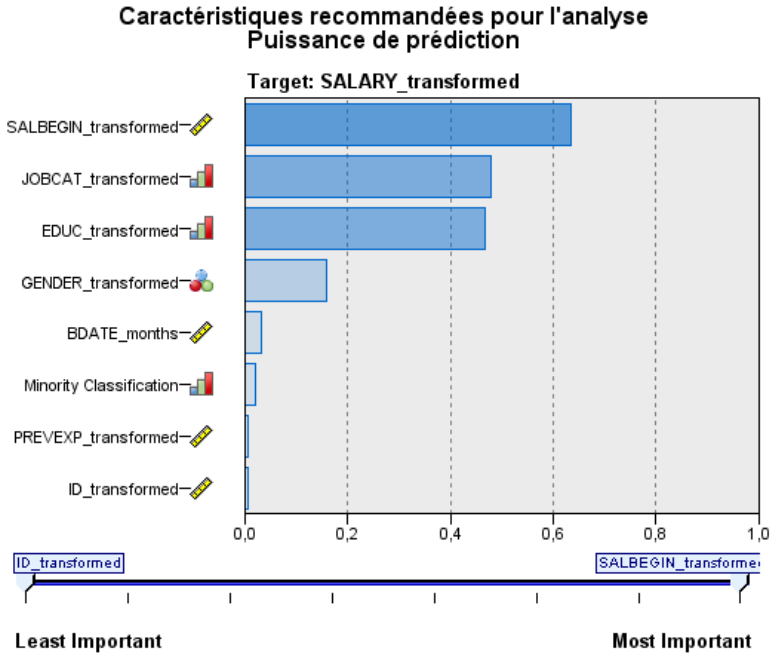
Pour chaque action effectuée par la préparation automatique des données, les vvaleurs prédites d'entrée sont transformées et/ou supprimées ; les champs qui survivent à une action sont utilisés à la suivante. Les champs qui survivent jusqu'à la dernière étape sont ensuite recommandés pour la modélisation, alors que les entrées des valeurs prédites transformées et construites sont supprimées.

Le Récapitulatif des actions est un simple tableau qui répertorie les actions effectuées par l'ADP. Lorsqu'une Action est soulignée, vous pouvez cliquer dessus pour afficher plus de détails sur les actions effectuées dans une vue liée. Pour plus d'informations, reportez-vous à la section Détails des actions sur p. 42.

Remarque : Seules les versions d'origine et transformées finales de chaque champ sont affichées, et pas les versions intermédiaires utilisées pendant l'analyse.

Puissance de prédiction

Figure 4-17
Puissance de prédiction



Affichée par défaut au début de l'analyse ou lorsque vous sélectionnez Valeurs prédites recommandées pour l'analyse dans la vue principale Récapitulatif du traitement des champs, le diagramme affiche la puissance de prédiction des valeurs prédites recommandées. Les champs sont triés par puissance de prédiction, avec le champ ayant la plus haute valeur apparaissant en premier.

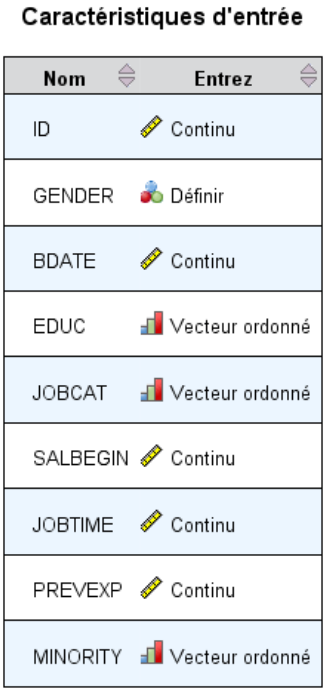
Pour les versions transformées des valeurs prédites ordinaires, le nom des champs reflète votre choix de suffixe dans le panneau Noms de champ de l'onglet Paramètres ; par exemple : *_transformed*.

Les icônes de niveau de mesure sont affichées après les noms de champ individuels.

La puissance de prédiction de chaque valeur prédite recommandée est calculée à partir d'une régression linéaire ou d'un modèle de Naïve Bayes selon que la cible est continue ou qualitative.

Tableau des champs

Figure 4-18
Tableau des champs



La vue Tableau des champs est un simple tableau qui répertorie les descriptives importantes et qui apparaît lorsque vous cliquez sur Cible, Valeurs prédites, ou Valeurs prédites non utilisées dans la vue principale Récapitulatif du traitement des champs.

Ce tableau contient deux colonnes :

- **Nom.** Le nom de la valeur prédite.

  Pour les cibles, l'étiquette ou le nom d'origine du champ est utilisé, même si la cible a été transformée.

  Pour les versions transformées des valeurs prédites ordinaires, le nom reflète votre choix de suffixe dans le panneau Noms de champ de l'onglet Paramètres ; par exemple : *_transformed*.

  Pour les champs dérivés des dates et des heures, le nom de la version transformée finale est utilisé ; par exemple, *bdate_years*.

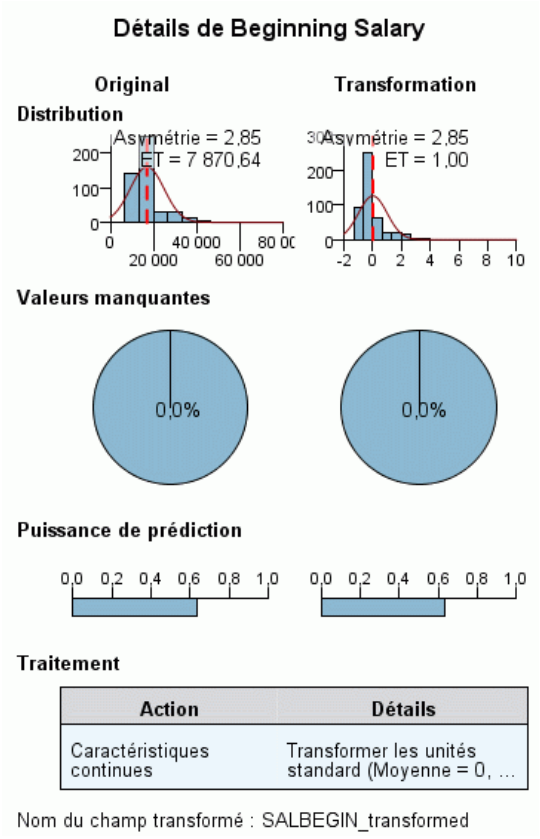
  Pour les valeurs prédites construites, le nom de la valeur prédite construite est utilisée ; par exemple : *Valeurprédite1*.

- **Niveau de mesure.** Affiche l'icône représentant le type de données.

  Pour la cible, le Niveau de mesure reflète toujours la version transformée (si la cible a été transformée), par exemple, changée d'ordinale (ensemble ordonné) à continue (plage, échelle) et vice versa.

Détails des champs

Figure 4-19
Détails des champs



La vue Détails des champs contient les diagrammes de distribution, des valeurs manquantes et de la puissance de prédiction (le cas échéant) pour le champ sélectionné et s'affiche lorsque vous cliquez sur un Nom de la vue principale Champs. De plus, l'historique du traitement pour le champ et le nom du champ transformé apparaissent également (le cas échéant).

Pour chaque ensemble de diagrammes, deux versions apparaissent côte à côte pour comparer le champ avec et sans transformations appliquées ; si aucune version transformée du champ n'existe, un diagramme apparaît pour la version d'origine uniquement. Pour les champs de date ou d'heure dérivés et les valeurs prédites construites, les diagrammes n'apparaissent que pour la nouvelle valeur prédite.

Remarque : Si un champ est exclu parce qu'il contient trop de modalités, seul l'historique de traitement apparaît.

Diagramme de distribution

La distribution des champs continus apparaît dans un histogramme, avec une courbe normale superposée et une ligne de référence verticale pour la valeur moyenne ; les champs qualitatifs apparaissent sous forme de diagramme en bâtons.

Les histogrammes sont étiquetés pour montrer l'écart-type et l'asymétrie, toutefois l'asymétrie n'apparaît pas si le nombre des valeurs est inférieur ou égal à 2 ou si la variance du champ d'origine est inférieure à 10-20.

Passez la souris sur le diagramme pour afficher la moyenne des histogrammes ou le nombre et le pourcentage du nombre total d'enregistrements des modalités dans les diagrammes en bâtons.

Diagramme des valeurs manquantes

Les diagrammes en secteurs comparent le pourcentage des valeurs manquantes avec et sans transformations appliquées ; les étiquettes de diagramme indiquent le pourcentage.

Si l'ADP traite les valeurs manquantes, le diagramme en secteurs après la transformation comprend la valeur de remplacement comme étiquette, c'est-à-dire la valeur utilisée à la place des valeurs manquantes.

Passez la souris sur le diagramme pour afficher le nombre des valeurs manquantes et le pourcentage du nombre total d'enregistrements.

Diagramme de puissance de prédiction

Pour les champs recommandés, les diagrammes en bâtons affichent la puissance de prédiction avant et après la transformation. Si la cible a été transformée, la puissance de prédiction calculée tient compte de la cible transformée.

Remarque : Les diagrammes de puissance de prédiction ne sont pas affichés si aucune cible n'est définie, ou si la cible est atteinte depuis le panneau de la vue principale.

Passez la souris sur le diagramme pour afficher la valeur de la puissance de prédiction.

Tableau des historiques du traitement

Ce tableau indique la façon dont la version transformée d'un champ a été dérivée. Les actions entreprises par l'ADP sont répertoriées dans l'ordre dans lequel elles ont été exécutées ; mais, pour certaines étapes, plusieurs actions ont pu être exécutées pour un champ particulier.

Remarque : Ce tableau n'apparaît pas pour les champs qui n'ont pas été transformés.

Les informations du tableau sont divisées en deux ou trois colonnes :

- **Action.** Le nom de l'action. Par exemple, Valeurs préditess continues. Pour plus d'informations, reportez-vous à la section Détails des actions sur p. 42.
- **Détails.** La liste des traitements effectués. Par exemple, Transformer en unités standard.
- **Fonction.** Apparaît uniquement pour les valeurs prédites construites et affiche la combinaison linéaire de champs d'entrée, par exemple, 0,06*âge + 1,21*hauteur.

Détails des actions

Figure 4-20
Analyse ADP - Détails des actions

Étape 9 : Caractéristiques continues

Transformation	Nombre de caractéristiques	Critères	
		Moyenne	SD
Transformer en unités standard	5	0	1

Construction d'espace de caractéristiques	N
Caractéristiques construites	0
Caractéristiques exclues en raison d'une faible association avec une cible	1
Caractéristiques exclues parce qu'elles étaient constantes après le regroupement	0

La vue liée Détails des actions apparaît lorsque vous cliquez sur Action dans la vue principale Récapitulatif des actions.La vue liée Détails des actions affiche des informations relatives aux actions et des informations communes pour chaque étape de traitement effectuée. Les détails relatifs à chaque action spécifique apparaissent d'abord.

La description de chaque action est utilisée comme titre en haut de la vue liée. Les détails relatifs à chaque action sont affichés sous le titre, et peuvent contenir des détails sur le nombre de valeurs prédites dérivées, de champs reconvertis, de transformations de cible, de modalités fusionnées ou réorganisées et de valeurs prédites construites ou exclues.

Au cours du traitement des actions, le nombre de valeurs prédites utilisées pour le traitement peut varier, par exemple lorsque des valeurs prédites sont exclues ou fusionnées.

Remarque : Si une action est désactivée ou qu'aucune cible n'est spécifiée, un message d'erreur apparaît à la place des détails de l'action lorsque vous cliquez sur l'action dans la vue principale Récapitulatif des actions.

Il existe neuf actions possibles, toutefois, toutes ne sont pas nécessairement actives pour chaque analyse.

tableau Champs de texte

Ce tableau affiche le nombre :

- Valeurs prédites exclues de l'analyse.

Tableau Valeurs prédites de date et d'heure

Ce tableau affiche le nombre :

- Durées dérivées des valeurs prédites de date et d'heure.
- d'éléments Date et heure.
- Valeurs prédites de date et d'heure dérivées, au total.

La date ou heure de référence est affichée comme note de bas de page si des durées de date ont été calculées.

Tableau Filtrage des valeurs prédites

Ce tableau affiche le nombre des valeurs prédites suivantes exclues du traitement :

- constantes.
- Valeurs prédites avec trop de valeurs manquantes.
- Valeurs prédites avec trop d'observations dans une seule modalité.
- Champs nominaux (ensembles) avec trop de modalités.
- Valeurs prédites supprimées, au total.

Tableau Vérifier le niveau de mesure

Ce tableau affiche le nombre de champs reconvertis, répartis selon les catégories suivantes :

- Champs ordinaux (ensembles ordonnés) reconvertis en champs continus.
- Champs continus reconvertis en champs ordinaux.
- Nombre total des champs reconvertis.

Si aucun champ d'entrée (cible ou de valeurs prédites) n'est un ensemble continu ou ordinal, cela apparaît en note de bas de page.

tableau Valeurs éloignées

Ce tableau affiche le nombre de valeurs éloignées traitées.

- soit le nombre de champs continus pour lesquels des valeurs éloignées ont été recherchées et éliminées, ou le nombre de champs continus pour lesquels les valeurs éloignées ont été recherchées et définies sur manquantes, en fonction de vos paramètres dans le panneau Préparer les entrées & la cible dans l'onglet Paramètres.
- le nombre de champs continus exclus parce qu'ils étaient constants après le traitement des valeurs éloignées.

Une note de bas de page indique la valeur de césure des valeurs éloignées et une autre note de bas de page apparaît si aucun champ d'entrée (cible ou de valeurs prédites) n'est continu.

tableau Valeurs manquantes

Ce tableau affiche le nombre de champs qui contenaient des valeurs manquantes remplacées, selon les catégories suivantes :

- Cible. Cette ligne n'apparaît pas si aucune cible n'est spécifiée.
- Valeurs prédites. Elles sont divisées en nombre de champs nominaux (ensemble), ordinaux (ensemble ordonné) et continus.
- Le nombre total de valeurs manquantes remplacées.

Tableau Cible

Ce tableau indique si la cible a été transformée :

- transformation de Box-Cox en normalité. Cette catégorie est elle-même divisée en colonnes qui indiquent le critère spécifié (moyenne et écart-type) et le Lambda.
- modalités cibles réorganisées pour améliorer la stabilité.

Tableau valeurs prédites qualitatives

Ce tableau affiche le nombre de valeurs prédites qualitatives :

- dont les modalités ont été réorganisées de la plus faible la plus élevée pour améliorer la stabilité.
- dont les modalités ont été fusionnées pour optimiser l'association avec la cible.
- dont les modalités ont été fusionnées pour traiter les modalités éparpillées.
- exclues en raison d'une faible association avec la cible.
- exclues parce qu'elles étaient constantes après la fusion.

Une note de bas de page apparaît si aucune valeur prédite qualitative n'existe.

Tableau Valeurs prédites continues

Il existe deux tableaux. Le premier affiche une des transformations suivantes :

- les valeurs des variables prédites transformées en unités standard. De plus, il indique le nombre de valeurs prédites transformées, la moyenne spécifiée et l'écart-type.
- Les valeurs des variables prédites mappées sur un intervalle commun. De plus, il indique le nombre de valeurs prédites transformées utilisant une transformation min-max, ainsi que les valeurs minimum et maximum spécifiées.
- les valeurs des variables prédites et le nombre de variables prédites regroupées.

Le deuxième tableau affiche les détails de construction de l'espace des valeurs prédites, sous la forme du nombre de valeurs prédites :

- construites.
- exclues en raison d'une faible association avec la cible.

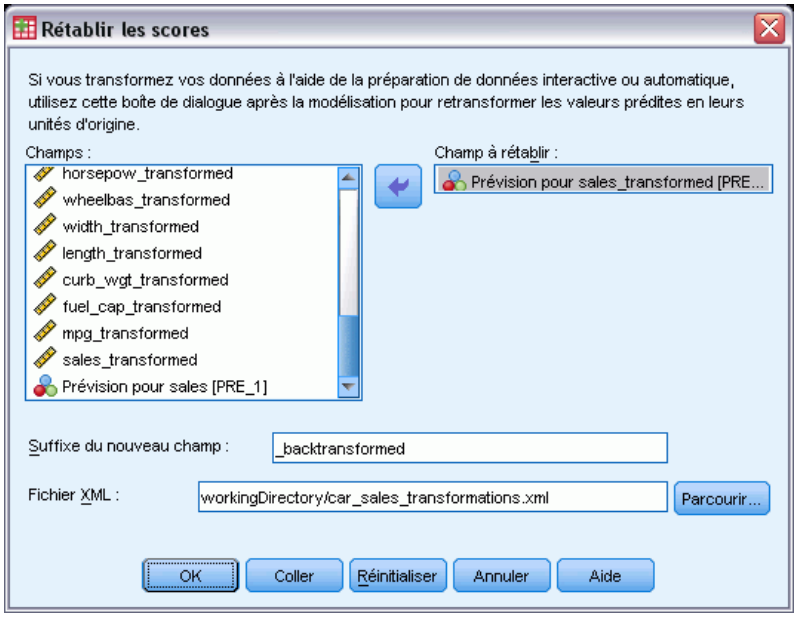
- exclues parce qu'elles étaient constantes après le regroupement.
- exclues parce qu'elles étaient constantes après la construction.

Une note de bas de page apparaît si aucune valeur prédite continue n'a été saisie.

Rétablir les scores

Si une cible a été transformée par l'ADP, les modèles en résultant créés à l'aide de la cible transformée évaluent les unités transformées. Afin d'interpréter et d'utiliser les résultats, vous devez reconvertir la valeur observée dans son échelle d'origine.

Figure 4-21
Rétablir les scores



Pour rétablir les scores, dans les menus, choisissez :

Transformer > Préparer les données pour la modélisation > Rétablir les scores...

▶ Sélectionnez un champ à rétablir. Ce champ doit contenir des valeurs prévues par le modèle de la cible transformée.

▶ Spécifiez un suffixe pour le nouveau champ. Ce nouveau champ contiendra des valeurs prévues par le modèle à l'échelle d'origine de la cible non transformée.

▶ Spécifiez l'emplacement du fichier XML contenant les transformations de l'ADP. Ce doit être un fichier enregistré à partir des boîtes de dialogue Préparation automatique ou interactive des données. Pour plus d'informations, reportez-vous à la section Appliquer et enregistrer les transformations sur p. 30.

Chapitre
5

Identification des observations inhabituelles

La procédure de détection des anomalies vise à repérer les observations inhabituelles en se basant sur les écarts par rapport aux normes de leurs groupes. La procédure est destinée à détecter rapidement les observations inhabituelles afin de vérifier les données à l'étape d'analyse exploratoire des données, avant d'effectuer toute sorte d'analyse inférentielle de ces mêmes données. Cet algorithme sert à détecter des anomalies générales. Il est vrai que la définition d'une observation anormale ne s'applique pas à tous les secteurs. Par exemple, la définition d'une anomalie peut être clairement définie lorsqu'il s'agit de détecter des moyens de paiements inhabituels dans l'industrie pharmaceutique ou du blanchissement d'argent dans l'industrie bancaire.

Exemple : Un analyste de données employé pour construire des modèles capables de prédire les résultats obtenus suite au traitement d'attaques cardiaques cherche des données de qualité, car de tels modèles sont sensibles aux observations inhabituelles. Certaines de ces observations éloignées sont des observations tout à fait uniques et s'avèrent donc inexploitables en matière de prédiction, alors que d'autres sont dues à des erreurs de saisie de données dans lesquelles les valeurs sont techniquement « correctes » sans pouvoir toutefois être prises en compte par les procédures de validation de données. La procédure d'identification des observations inhabituelles sert à identifier ces valeurs éloignées et à en dresser la liste afin que l'analyste puisse décider de la manière de les traiter.

Statistiques : La procédure génère des groupes de pairs, des normes de groupes de pairs pour des variables continues et qualitatives, des indices d'anomalies basés sur les écarts par rapport aux normes de groupes de pairs, ainsi que des valeurs d'impact de variables pour les variables contribuant le plus à une observation considérée comme inhabituelle.

Analyse des données

Données : Cette procédure fonctionne avec des variables continues et qualitatives. Chaque ligne représente une observation distincte tandis que chaque colonne représente une variable différente sur laquelle les groupes de pairs sont basés. Une variable d'identification d'observations est disponible dans le fichier de données pour marquer les résultats, mais elle ne sera pas utilisée dans l'analyse. Les valeurs manquantes sont autorisées. La variable de pondération est ignorée, si indiquée auparavant.

© Copyright SPSS Inc. 1989, 2010

Le modèle de détection peut être appliqué à un nouveau fichier de données de test. Les éléments des données du test doivent être identiques aux éléments contenus dans les données de formation. Et, en fonction des paramètres d'algorithme, le traitement de la valeur manquante utilisé pour créer le modèle doit être appliqué au fichier de données de test avant d'effectuer une notation.

Tri par observation. Notez que la solution peut dépendre de l'ordre des observations. Pour réduire les effets de tri, classez les observations de manière aléatoire. Pour vérifier la stabilité d'une solution donnée, vous pouvez obtenir différentes solutions dans lesquelles les observations sont triées de différentes manières aléatoires. Si les fichiers sont très volumineux, vous pouvez effectuer plusieurs fois l'opération sur un échantillon des observations triées de différentes manières aléatoires.

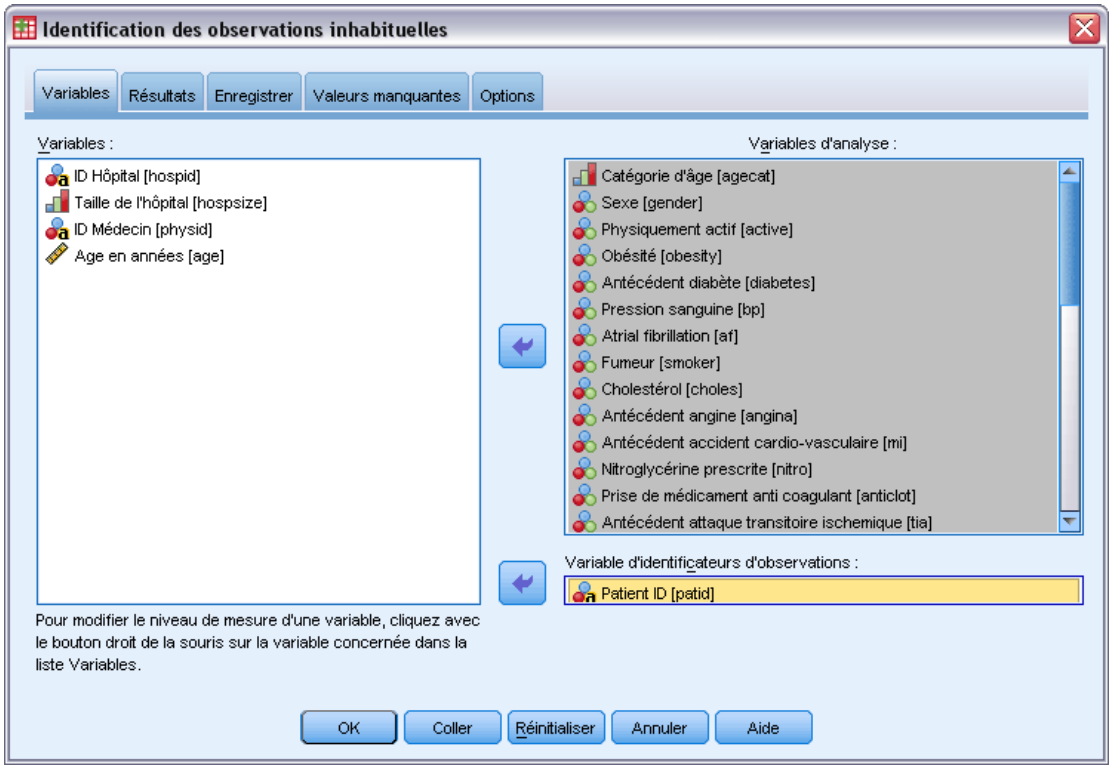
Hypothèses : L'algorithme suppose que toutes les variables sont non constantes et indépendantes, et qu'aucune observation ne possède de valeur manquante pour les variables d'entrée. Chaque variable continue est considérée comme ayant une distribution normale (gaussienne) et chaque variable qualitative comme ayant une distribution multinomiale. Des tests internes empiriques indiquent que la procédure est assez résistante aux violations de l'hypothèse d'indépendance et des hypothèses de distribution, mais vous devez savoir comment ces hypothèses sont vérifiées.

Pour identifier les observations inhabituelles

▶ A partir des menus, sélectionnez :

Données > Identifier les observations inhabituelles...

Figure 5-1
Boîte de dialogue Identifier les observations inhabituelles, onglet Variables

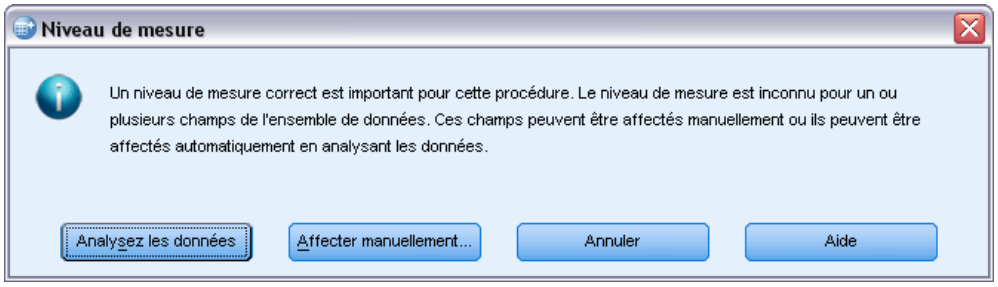


- Sélectionnez au moins une variable d'analyse.
- Vous pouvez également sélectionner une variable d'identificateur d'observation à utiliser pour l'étiquetage du résultat.

Champs avec un niveau de mesure inconnu

L'alerte du niveau de mesure apparaît lorsque le niveau de mesure d'une ou plusieurs variables (champs) de l'ensemble de données est inconnu. Le niveau de mesure ayant une incidence sur le calcul des résultats de cette procédure, toutes les variables doivent avoir un niveau de mesure défini.

Figure 5-2
Alerte du niveau de mesure

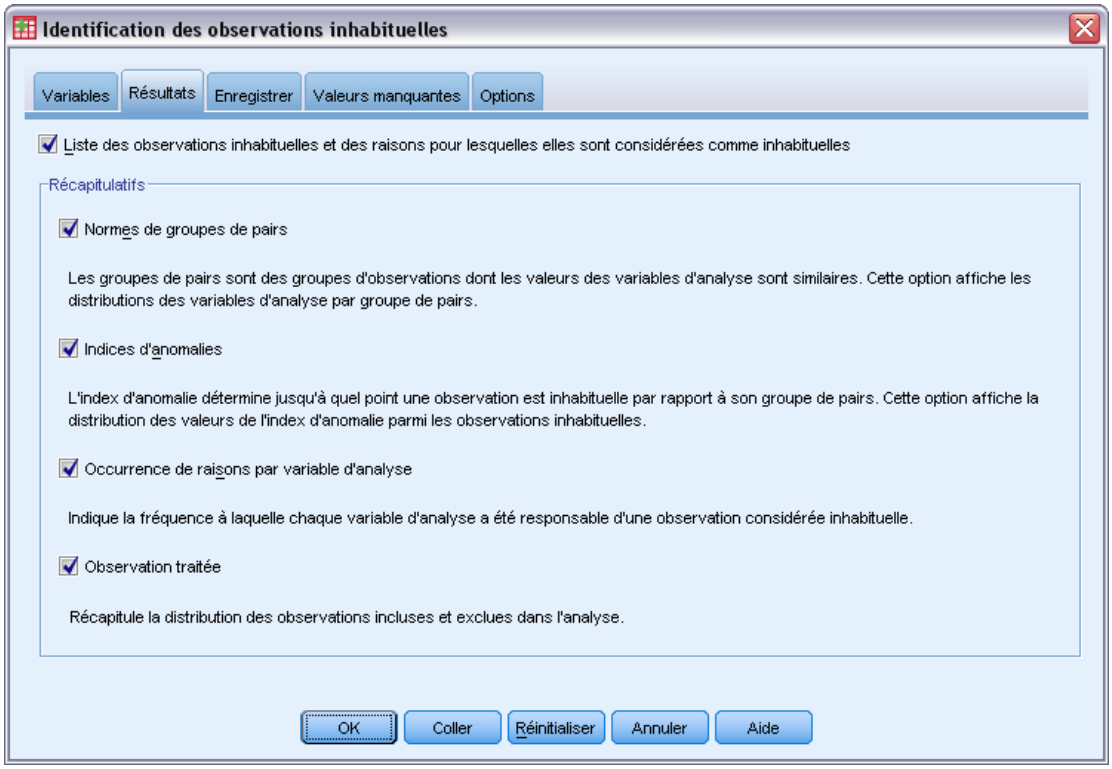


- **Analysez les données.** Lit les données dans l'ensemble de données actifs et attribue le niveau de mesure par défaut à tous les champs ayant un niveau de mesure inconnu. Si l'ensemble de données est important, cette action peut prendre un certain temps.
- **Attribuer manuellement.** Ouvre une boîte de dialogue qui répertorie tous les champs ayant un niveau de mesure inconnu. Vous pouvez utiliser cette boîte de dialogue pour attribuer un niveau de mesure à ces champs. Vous pouvez également attribuer un niveau de mesure dans l'affichage des variables de l'éditeur de données.

Le niveau de mesure étant important pour cette procédure, vous ne pouvez pas accéder à la boîte de dialogue d'exécution de cette procédure avant que tous les champs n'aient des niveaux de mesure définis.

Identification du résultat d'observations inhabituelles

Figure 5-3
Boîte de dialogue Identifier les observations inhabituelles, onglet Résultat



Liste des observations inhabituelles et des raisons pour lesquelles elles sont considérées comme inhabituelles. Cette option propose trois tableaux :

- La liste d'index des observations présentant une anomalie affiche les observations identifiées comme étant inhabituelles, ainsi que leur valeur d'index d'anomalie correspondante.
- La liste d'ID des pairs d'observation présentant une anomalie affiche les observations inhabituelles ainsi que les informations relatives à leur groupe de pairs correspondant.
- La liste des raisons expliquant les anomalies affiche le numéro de l'observation, la variable de raison, la valeur d'impact de la variable, la valeur de la variable et la norme de la variable pour chaque raison.

Tous les tableaux sont triés par index d'anomalie en ordre décroissant. De plus, les ID des observations ne sont affichés que si la variable d'identificateur de l'observation est indiquée dans l'onglet Variables.

Principales statistiques : Les commandes de ce groupe génèrent des récapitulatifs de distribution.

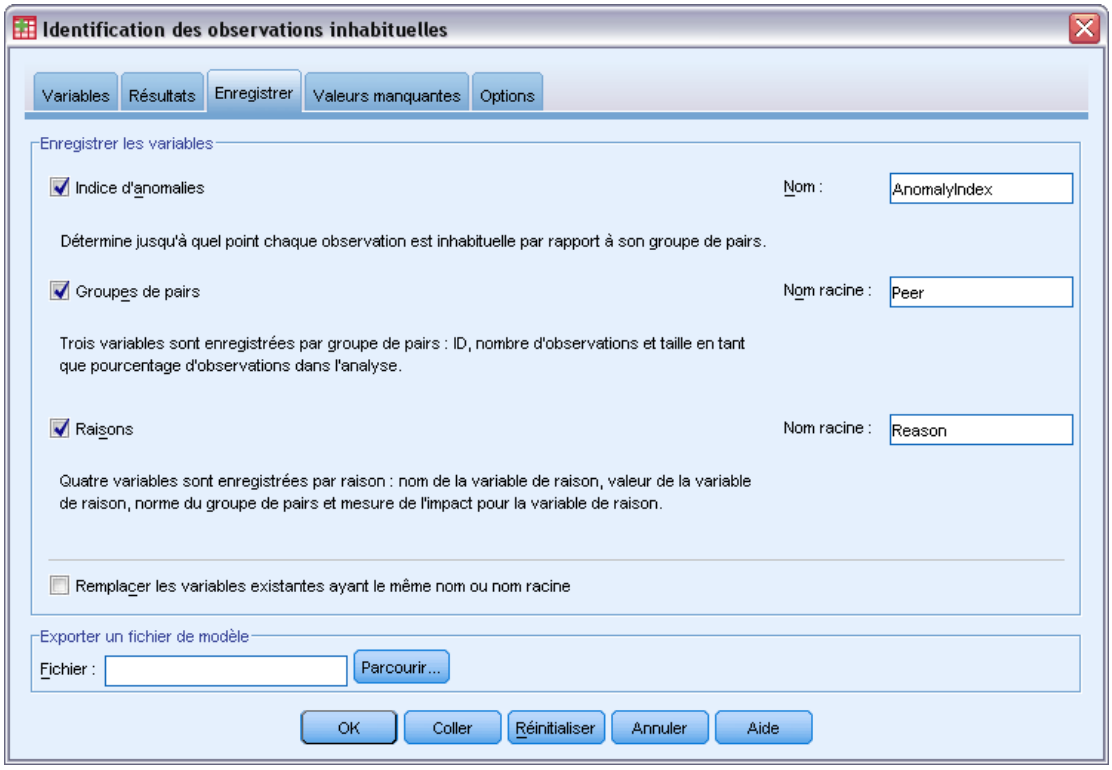
- **Normes de groupes de pairs.** Cette option affiche le tableau des normes de variables continues (en cas d'utilisation de variables continues dans l'analyse) et le tableau des normes de variables qualitatives (en cas d'utilisation de variables qualitatives dans l'analyse). Le tableau des normes de variables continues affiche la moyenne et l'écart-type de chaque variable continue pour chaque groupe de pairs. Le tableau des normes de variables qualitatives affiche

le mode (modalité la plus utilisée), sa fréquence et le pourcentage de fréquence de chaque variable qualitative pour chaque groupe de pairs. La moyenne d'une variable continue et le mode d'une variable qualitative sont utilisés comme les valeurs standard dans l'analyse.

- **Indices d'anomalies.** Le récapitulatif de l'index d'anomalies affiche les statistiques descriptives pour l'index d'anomalies des observations identifiées comme étant les plus inhabituelles.
- **Occurrence de raisons par variable d'analyse.** Pour chaque raison, le tableau affiche la fréquence et le pourcentage de fréquence de chaque occurrence de variable exprimé sous la forme d'une raison. Le tableau indique également les statistiques descriptives de l'observation de chaque variable. Si le nombre de raisons maximal est défini sur 0 dans l'onglet Options, cela signifie que cette option n'est pas disponible.
- **Observations traitées.** Le récapitulatif du traitement des observations affiche les effectifs et les pourcentages d'effectif pour toutes les observations dans un fichier de travail, les observations incluses et exclues de l'analyse, et les observations de chaque groupe de pairs.

Identification des enregistrements d'observations inhabituelles

Figure 5-4
Boîte de dialogue Identifier les observations inhabituelles, onglet Enregistrer



Enregistrer les variables. Les commandes de ce groupe vous permettent d'enregistrer des variables de modèle dans le fichier de travail. Vous pouvez également choisir de remplacer les variables existantes dont le nom est en conflit avec les variables à enregistrer.

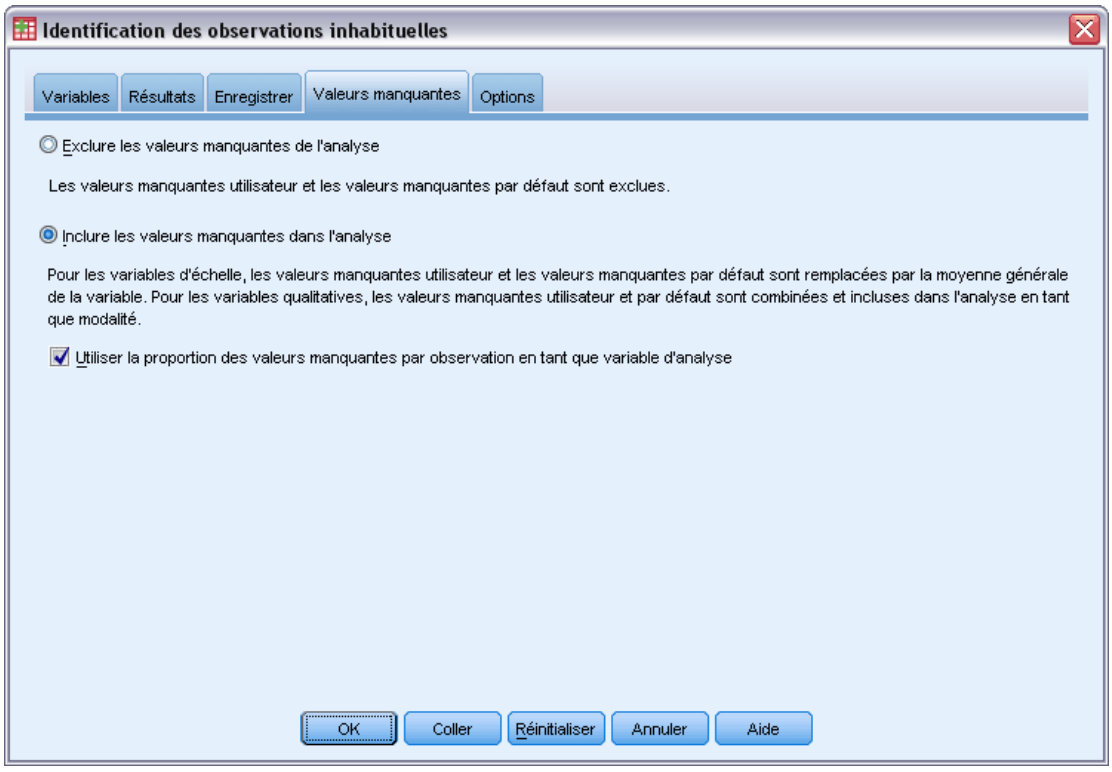
- **Indices d'anomalies.** Enregistre la valeur de l'index d'anomalie pour chaque observation dans une variable portant le nom indiqué.

- **Groupes de pairs.** Enregistre l'ID du groupe de pairs, le nombre d'observations et la taille en tant que pourcentage pour chaque observation dans les variables portant le nom de racine spécifié. Par exemple, si le nom racine *Peer* est spécifié, les variables *Peerid*, *PeerSize* et *PeerPctSize* sont générées. *Peerid* est l'ID du groupe de pairs de l'observation, *PeerSize* la taille du groupe et *PeerPctSize* la taille du groupe exprimée en pourcentage.
- **Raisons.** Enregistre les ensembles de variables de raison portant le nom racine spécifié. Un ensemble de variables de raison est composé du nom de la variable en tant que raison, de sa mesure de l'impact de la variable, de sa propre valeur et de la valeur standard. Le nombre d'ensembles dépend du nombre de raisons demandées dans l'onglet Options. Par exemple, si le nom racine *Reason* est spécifié, les variables *ReasonVar_k* , *ReasonMeasure_k*, *ReasonValue_k* et *ReasonBorm_k* sont alors générées, *k* correspondant à la raison *k*th. Cette option n'est pas disponible si le nombre des raisons est défini sur 0.

Exporter un fichier de modèle Cette option vous permet d'enregistrer le modèle au format XML.

Identification des valeurs manquantes des observations inhabituelles

Figure 5-5
Boîte de dialogue Identifier les observations inhabituelles, onglet Valeurs manquantes

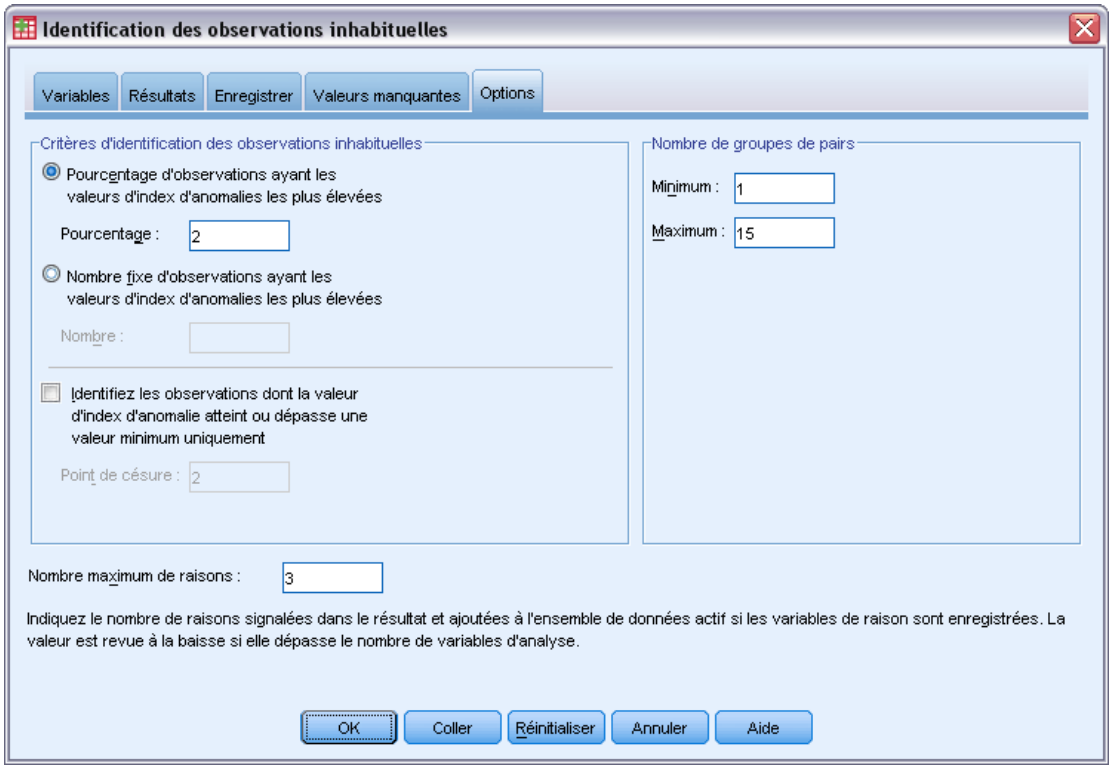


L'onglet Valeurs manquantes sert à contrôler le traitement des valeurs manquantes spécifiées par l'utilisateur et les valeurs manquantes spécifiées par le système.

- **Exclure les valeurs manquantes de l'analyse.** Les observations contenant des valeurs manquantes sont exclues de l'analyse.
- **Inclure les valeurs manquantes dans l'analyse** Les valeurs manquantes des variables continues sont remplacées par leur moyenne générale correspondante, et les modalités manquantes des variables qualitatives sont groupées et traitées en tant que modalité valide. Les variables traitées sont ensuite utilisées dans l'analyse. Vous pouvez également demander la création d'une variable supplémentaire représentant la proportion de variables manquantes dans chaque observation et utiliser cette variable dans l'analyse.

Options d'identification des observations inhabituelles

Figure 5-6
Boîte de dialogue Identifier les observations inhabituelles, onglet Options



Critères d'identification des observations inhabituelles. Ces sélections déterminent le nombre d'observations à inclure dans la liste d'anomalies.

- **Pourcentage d'observations ayant les valeurs d'index d'anomalies les plus élevées.** Indiquez un nombre positif inférieur ou égal à 100.
- **Nombre fixe d'observations ayant les valeurs d'index d'anomalies les plus élevées.** Indiquez un entier positif inférieur ou égal au nombre total d'observations contenues dans le fichier de travail et utilisées dans l'analyse.
- **Identifiez les observations dont la valeur d'index d'anomalie atteint ou dépasse une valeur minimum uniquement.** Spécifiez un nombre non négatif. Une observation est considérée comme anormale si la valeur d'index d'anomalie est supérieure ou égale à la limite d'inclusion

spécifiée. Cette option est employée avec les options Pourcentage d'observations et Nombre fixe d'observations. Par exemple, si vous spécifiez un nombre fixe de 50 observations et une valeur de césure de 2, la liste d'anomalie sera composée de 50 observations au moins, chaque observation aura une valeur d'index d'anomalie supérieure ou égale à 2.

Nombre de groupes de pairs. La procédure cherche le meilleur nombre de groupes de pairs compris entre la valeur minimum et la valeur maximum spécifiées. Les valeurs doivent être des entiers positifs dont la valeur minimum ne doit pas dépasser la valeur maximum. Lorsque les valeurs spécifiées sont égales, la procédure part du principe que le nombre de groupes de pairs est fixe.

Remarque : En fonction de la variance de vos données, il peut arriver que le nombre de groupes de pairs pris en charge par les données soit inférieur au nombre spécifié comme valeur minimum. Dans une telle situation; la procédure risque d'engendrer un nombre de groupes de pairs plus petit.

Nombre maximum de raisons. Une raison est constituée de la mesure de l'impact d'une variable, du nom de la variable pour cette raison, de la valeur de la variable et de la valeur du groupe de pairs correspondant. Spécifiez un nombre entier non-négatif. Si cette valeur égale ou dépasse le nombre de variables traitées qui sont ensuite utilisées dans l'analyse, les variables sont alors affichées.

Fonctionnalités supplémentaires de la commande DETECTANOMALY

Le langage de syntaxe de commande vous permet aussi de :

- Omettre de l'analyse quelques variables du fichier de travail sans indiquer de façon explicite toutes les variables d'analyse (à l'aide de la sous-commande `EXCEPT`).
- spécifier un ajustement pour équilibrer l'influence des variables continues et qualitatives (à l'aide du mot-clé `MLWEIGHT` de la sous-commande `CRITERIA`).

Reportez-vous à la *Référence de syntaxe de commande* pour une information complète concernant la syntaxe.

Chapitre
6

Regroupement par casiers optimal

La procédure Regroupement optimal discrétise une ou plusieurs variables d'échelle (désormais appelées **variables d'entrée de regroupement**) en distribuant les valeurs de chaque variable dans des casiers. La formation de casiers est optimale par rapport à une variable guide catégorielle qui "supervise" le regroupement par casiers. Les casiers peuvent ensuite être utilisés à la place des valeurs de données d'origine pour de plus amples analyses.

Exemples : La réduction du nombre de valeurs distinctes que prend une variable a un certain nombre d'utilisations :

- Les données requises d'autres procédures. Les variables discrétisées peuvent être traitées comme catégorielles lors d'une utilisation dans des procédures faisant appel à ce type de variable. Par exemple, la procédure Tableaux croisés nécessite que toutes les variables soient catégorielles.
- Confidentialité des données. Signaler des valeurs regroupées par casiers au lieu des valeurs réelles aide à protéger la confidentialité de vos sources de données. La procédure Regroupement optimal peut guider le choix des casiers.
- Performances en matière de vitesse. Certaines procédures sont plus efficaces lorsque vous travaillez avec un nombre réduit de valeurs distinctes. Par exemple, la vitesse de la régression logistique multinomiale peut être améliorée grâce à l'utilisation de variables discrétisées.
- Révélation de la séparation complète ou quasi complète des données.

Recodage supervisé optimal et regroupement visuel. Les boîtes de dialogue Regroupement visuel proposent plusieurs méthodes automatiques de création de casiers sans utiliser de variable guide. Ces règles "non supervisées" sont utiles pour générer des statistiques descriptives, telles que des tableaux d'effectifs, mais le recodage supervisé optimal donne de meilleurs résultats si votre objectif final est de générer un modèle de prévision.

Résultats. La procédure génère des tableaux de divisions pour les casiers et les statistiques descriptives de chaque variable d'entrée de regroupement. En outre, vous pouvez enregistrer de nouvelles variables dans l'ensemble de données actif contenant les valeurs regroupées par casiers des variables d'entrée de regroupement et enregistrer les règles de regroupement comme syntaxe de commande pour les utiliser dans la discrétisation de nouvelles données.

Données : Cette procédure exige que les variables d'entrée de regroupement soient des variables d'échelle numériques. La variable guide doit être catégorielle et peut être chaîne ou numérique.

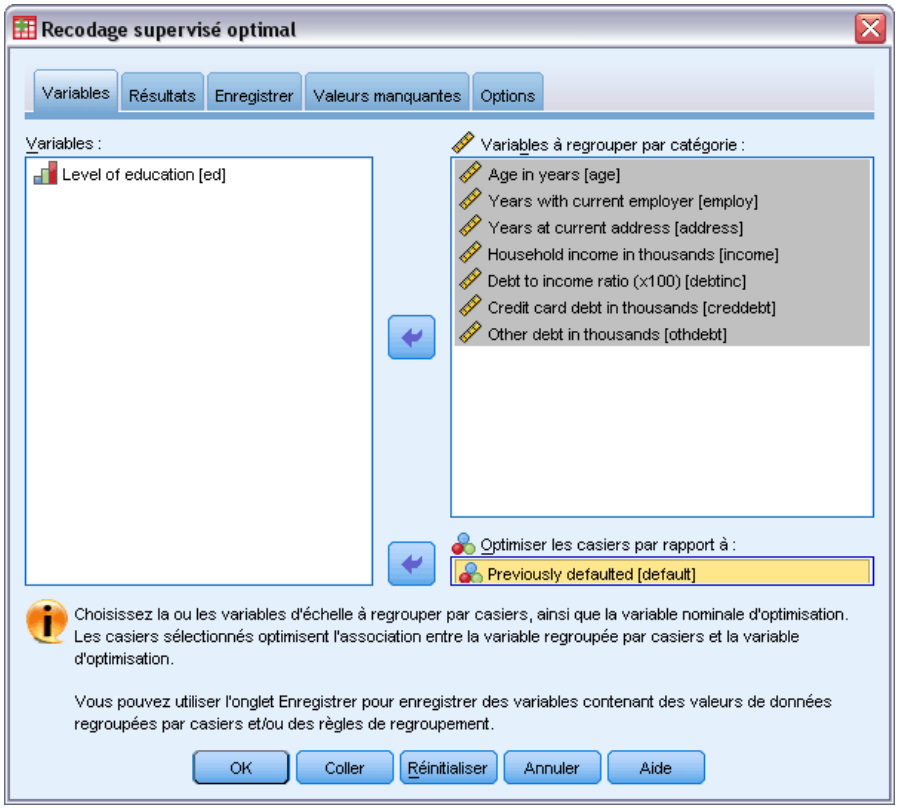
Pour obtenir un recodage supervisé optimal

A partir des menus, sélectionnez :

Transformer > Recodage supervisé optimal…

© Copyright SPSS Inc. 1989, 2010

Figure 6-1
Boîte de dialogue Regroupement optimal, onglet Variables

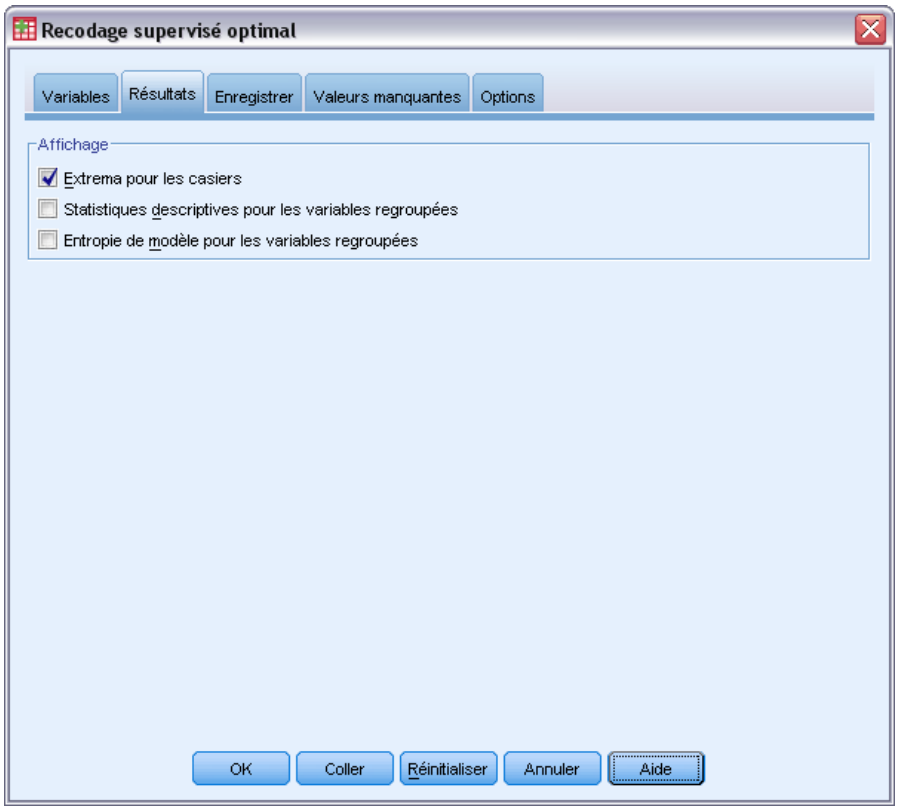


- Sélectionnez une ou plusieurs variables d'entrée de regroupement.
- Sélectionnez une variable guide.

Les variables contenant les valeurs des données regroupées par casiers ne sont pas générées par défaut. Utilisez l'onglet Enregistrer pour enregistrer ces variables.

Résultats du recodage supervisé optimal

Figure 6-2
Boîte de dialogue Regroupement optimal, onglet Résultat

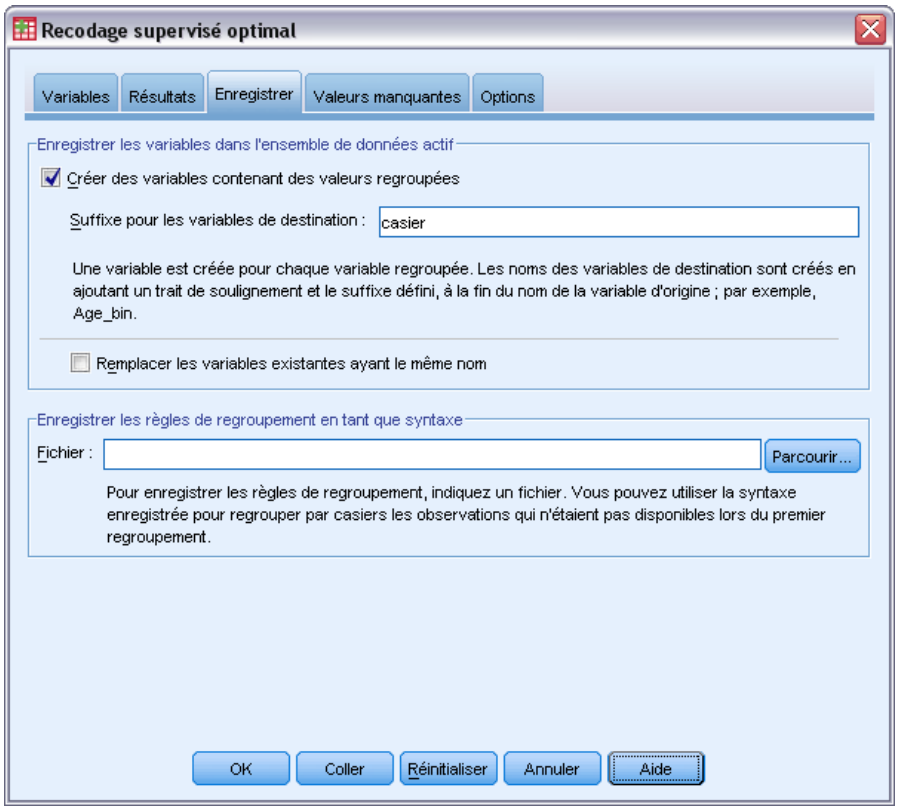


L'onglet Résultat contrôle l'affichage des résultats.

- **Extrema pour casiers.** Affiche l'ensemble d'extrema pour chaque variable d'entrée de regroupement.
- **Statistiques descriptives pour variables regroupées.** Pour chaque variable d'entrée de regroupement, cette option affiche le nombre d'observations dotées de valeurs valides, le nombre d'observations dotées de valeurs manquantes, le nombre de valeurs valides distinctes, et les valeurs minimale et maximale. Pour la variable guide, cette option affiche la distribution de classe pour chaque variable d'entrée de regroupement liée.
- **Entropie de modèle pour variables regroupées.** Pour chaque variable d'entrée de regroupement, cette option affiche une mesure de l'exactitude des prévisions de la variable par rapport à la variable guide.

Enregistrement du recodage supervisé optimal

Figure 6-3
Boîte de dialogue Regroupement optimal, onglet Enregistrer

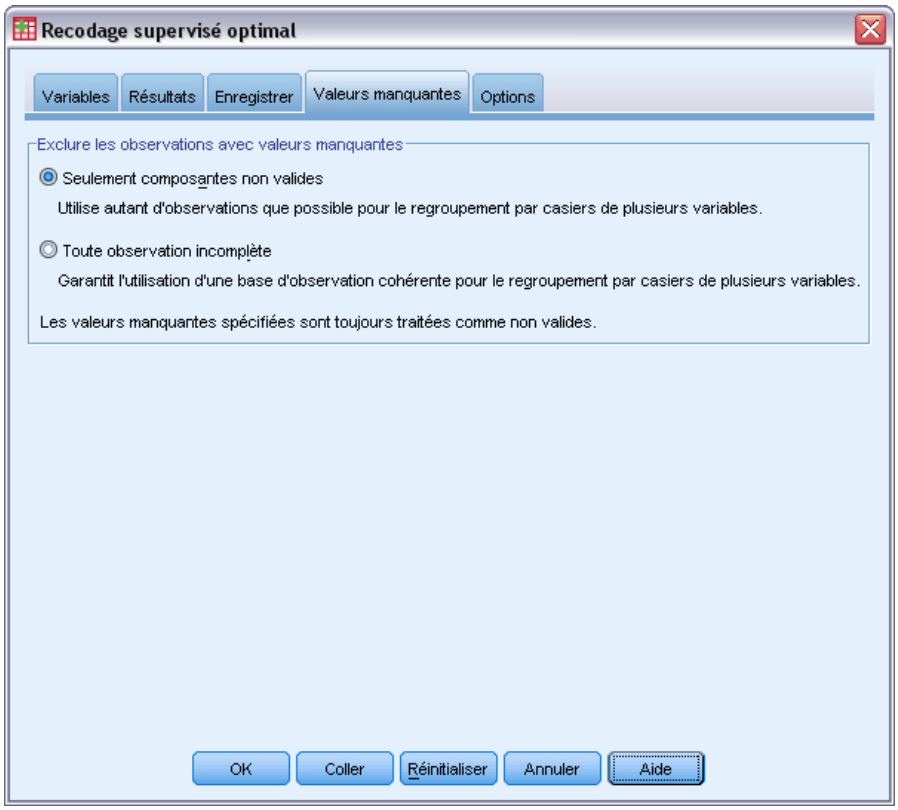


Enregistrer les variables dans le fichier de travail. Les variables contenant les valeurs de données regroupées par casiers peuvent se substituer aux variables d'origine pour une analyse ultérieure.

Enregistrer les règles de regroupement en tant que syntaxe . Génère une syntaxe de commande qui peut être utilisée pour regrouper d'autres ensembles de données par casiers. Les règles de recodage sont basées sur les divisions déterminées par l'algorithme de regroupement par casiers.

Valeurs manquantes de recodage supervisé optimal

Figure 6-4
Boîte de dialogue Regroupement optimal, onglet Valeurs manquantes

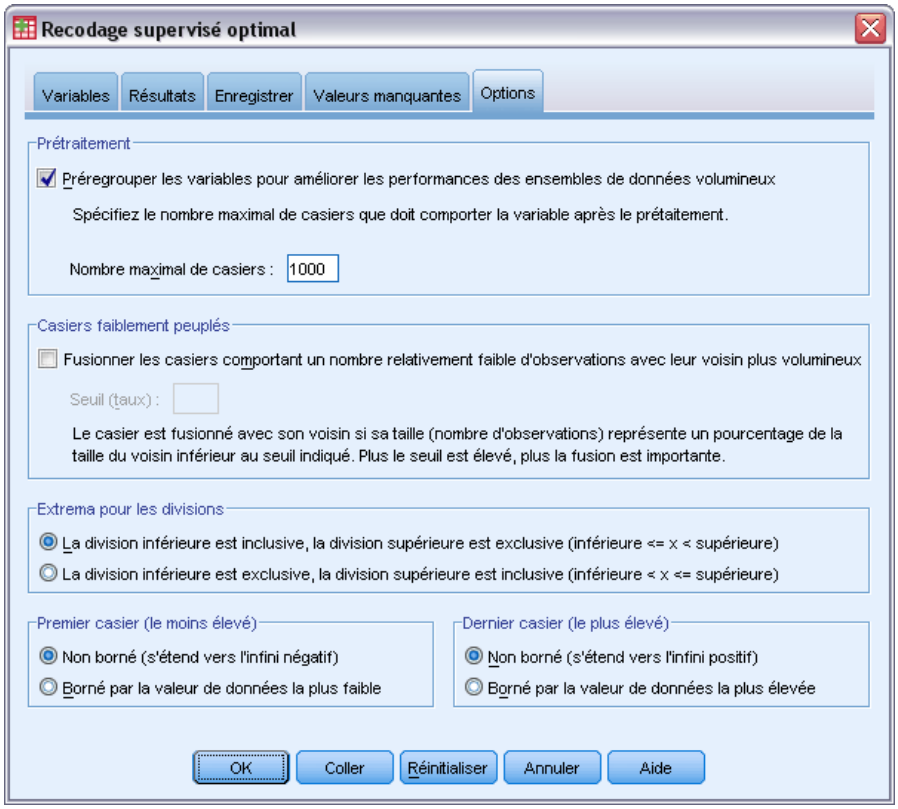


L'onglet Valeurs manquantes spécifie si les valeurs manquantes sont gérées par la suppression des observations incomplètes ou des composantes non valides seulement. Les valeurs manquantes spécifiées sont toujours traitées comme non valides. Lors du recodage des valeurs de variable d'origine en nouvelle variable, les valeurs manquantes spécifiées sont converties en valeurs manquantes par défaut.

- **Paires.** Cette option concerne chaque paire de variables guide et d'entrée de regroupement. La procédure utilise toutes les observations ayant des valeurs non manquantes sur la variable guide et d'entrée de regroupement.
- **Toute observation incomplète** Cette option concerne toutes les variables spécifiées dans l'onglet Variables. Si une variable est manquante pour une observation, l'observation est intégralement exclue.

Options Regroupement optimal

Figure 6-5
Boîte de dialogue Regroupement optimal, onglet Options



Prétraitement. Les variables d'entrée de "pré-regroupement" dotées de nombreuses valeurs distinctes améliorent le temps de traitement sans altérer la qualité des casiers finaux. Le nombre maximal de casiers fournit la limite supérieure du nombre de casiers créés. Ainsi, si vous spécifiez 1 000 comme maximum, mais qu'une variable d'entrée de regroupement possède moins de 1 000 valeurs distinctes, le nombre de casiers prétraités créés pour la variable d'entrée de regroupement sera égal au nombre de valeurs distinctes dans cette variable.

Casiers faiblement peuplés. La procédure génère parfois des casiers avec très peu d'observations. La stratégie suivante supprime ces pseudo-divisions :

- Pour une variable donnée, supposons que l'algorithme ait trouvé n divisions $_{\text{finales}}$ et donc n+1 casiers$_{\text{finaux}}$. Pour les casiers $i = 2, ..., n_{\text{final}}$ (compris entre le deuxième casier avec la plus faible valeur et le deuxième casier avec la valeur la plus élevée), calculez

$$\frac{sizeof(b_i)}{\min(sizeof(b_{i-1}),sizeof(b_{i+1}))}$$

où *taillede(b)* est le nombre d'observations dans le casier.

- Lorsque cette valeur est inférieure au seul de fusion spécifié, b_i est considérée comme faiblement peuplée et est fusionnée avecb_{i-1} ou b_{i+1}, selon la valeur avec la plus faible entropie d'informations de classe.

La procédure effectue un seul passage à travers les casiers.

Extrema des casiers. Cette option indique comment la limite inférieure d'un intervalle est définie. Puisque la procédure détermine automatiquement les valeurs des divisions, il s'agit surtout d'une question de préférence.

Premier casier (le moins élevé)/Dernier casier (le plus élevé). Ces options spécifient comment les divisions minimal et maximal de chaque variable d'entrée de regroupement sont définis. Généralement, la procédure suppose que les variables d'entrée de regroupement peuvent prendre n'importe quelle valeur sur la ligne des nombres réels, mais si vous avez une raison théorique ou pratique de limiter l'intervalle, vous pouvez le faire sur la base des valeurs les plus faibles/les plus élevées.

Fonctionnalités supplémentaires de la commande OPTIMAL BINNING

Le langage de syntaxe de commande vous permet aussi de :

- Effectuez un recodage non supervisé à l'aide de la méthode d'effectifs égaux (via la sous-commande `CRITERIA`).

Pour obtenir des renseignements complets sur la syntaxe, reportez-vous au manuel *Command Syntax Reference*.

Partie II: Exemples

Chapitre
7

Valider des données

La procédure Valider des données identifie les observations, variables et valeurs de données suspectes ou invalides.

Validation d'une base de données médicale

Un analyste engagé par un groupe médical doit maintenir la qualité des informations dans le système. Ce procédé implique la vérification des valeurs et des variables ainsi que la préparation d'un rapport pour le chef de l'équipe chargée d'entrer les données.

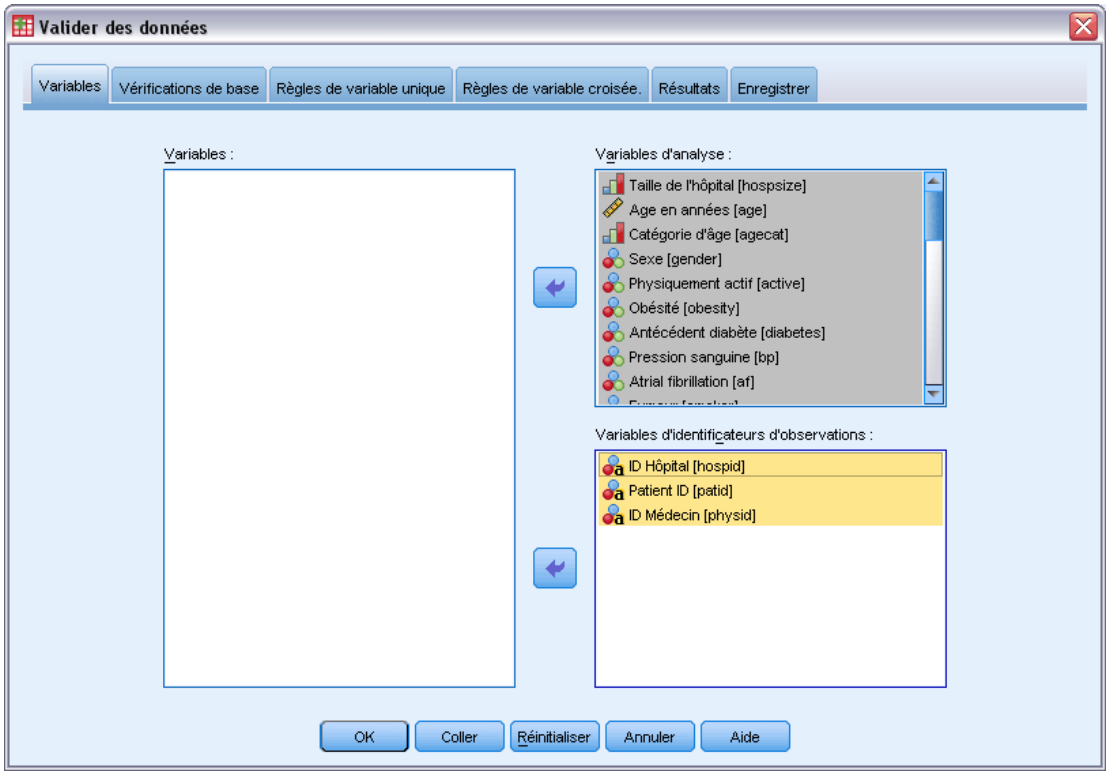
L'état le plus récent de la base de données se trouve dans *stroke_invalid.sav*. Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A sur p. 138. Utilisez la procédure Valider des données pour obtenir les informations nécessaires à la production de ce rapport. La syntaxe servant à produire ces analyses se trouve dans *validatedata_stroke.sps*.

Vérifications de base

- Pour exécuter une analyse de validation des données, sélectionnez à partir des menus :
  Données > Validation > Valider des données...

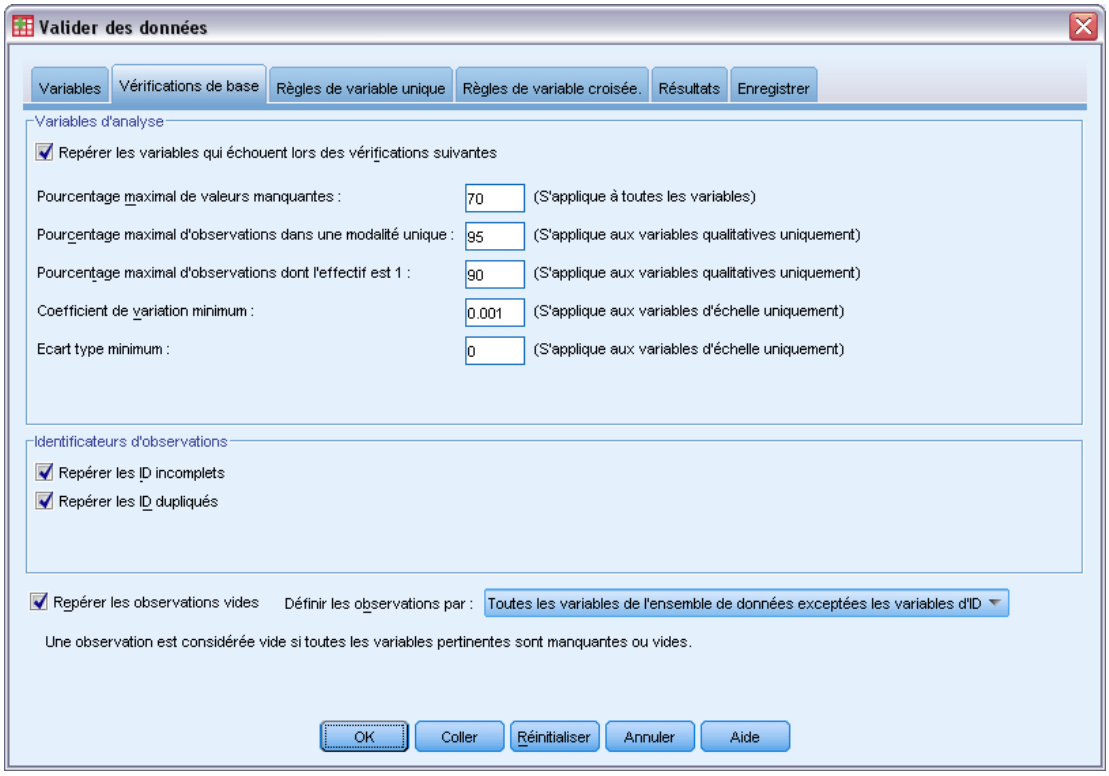
© Copyright SPSS Inc. 1989, 2010

Figure 7-1
Boîte de dialogue Valider les données, onglet Variables



- Sélectionnez *Taille de l'hôpital* et *Age en années* dans *Index de Barthel à 6 mois recodé* comme variables d'analyse.
- Sélectionnez *ID de l'hôpital*, *ID du patient* et *ID du médecin traitant* comme variables d'identificateurs d'observations.
- Cliquez sur l'onglet Vérifications de base.

Figure 7-2
Boîte de dialogue Valider les données, onglet Vérifications de base

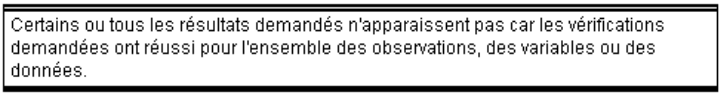


Les paramètres par défaut sont les paramètres que vous souhaitez exécuter.

▶ Cliquez sur OK.

Warnings

Figure 7-3
Warnings



Les variables d'analyse ont effectué les vérifications de base et il n'y a pas d'observations vides : un avertissement apparaît pour expliquer pourquoi aucun résultat ne correspond à ces vérifications.

Identificateurs incomplets

Figure 7-4
Identificateurs d'observations incomplets

	Identificateur		
Observation	hospid	patid	physid
288	OZN		125304
573		6137798782	790697
774		2322241867	176466

S'il manque des valeurs dans les variables d'identification d'observations, l'observation ne peut pas être correctement identifiée. Dans ce fichier de données, la valeur *ID du patient* manque à l'observation 288 et la valeur *ID de l'hôpital* manque aux observations 573 et 774.

Identificateurs dupliqués

Figure 7-5
Identificateurs d'observations dupliqués (affichage des 11 premiers)

Groupe d'identificateurs en double	Nombre de doublons	Observations ayant des identificateurs en double	Identificateur		
			hospid	patid	physid
1	2	10, 11	PBW	1406462419	355184
2	2	14, 15	PBW	2191527525	355184
3	2	21, 22	PBW	7237535360	616528
4	2	28, 29	NHV	4592215163	942982
5	2	30, 31	NHV	7628592330	371884
6	2	64, 65	NHV	0300750006	371884
7	2	83, 84	QWS	4590625286	215041
8	2	86, 87	QWS	6272818258	817329
9	2	96, 97	QWS	1959349605	215041
10	3	100, 101, 102	QWS	5856145337	817329
11	3	104, 105, 106	QWS	1543897849	817329

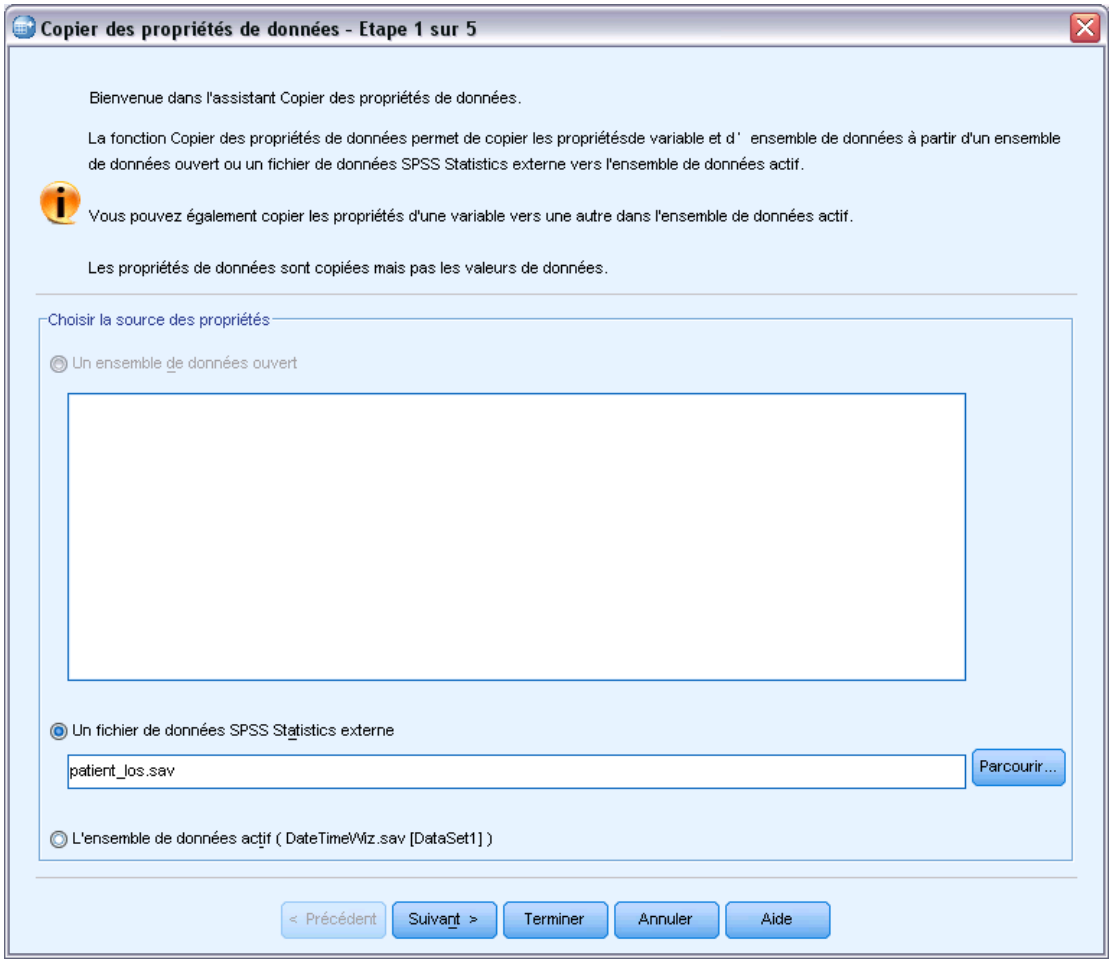
Une observation devrait être uniquement identifiée en fonction de la combinaison de valeurs des variables d'identificateurs. Les 11 premières entrées du tableau des identificateurs dupliqués apparaissent ici. Ces identificateurs dupliqués sont des patients ayant souffert de plusieurs troubles, chacun de ces troubles ayant fait l'objet d'une observation séparée. Ces informations pouvant être collectées sur une même ligne, les observations devraient être nettoyées.

Copie et utilisation de règles provenant d'un autre fichier

L'analyste remarque que les variables de ce fichier de données sont identiques à celles d'un autre projet. Les règles de validation qui sont définies pour ce projet sont enregistrées en tant que propriétés du fichier de données associé et peuvent être appliquées à ce fichier de données en copiant les propriétés de données du fichier.

▶ Pour copier des règles provenant d'un autre fichier, choisissez dans les menus :
Données > Copie des propriétés de données...

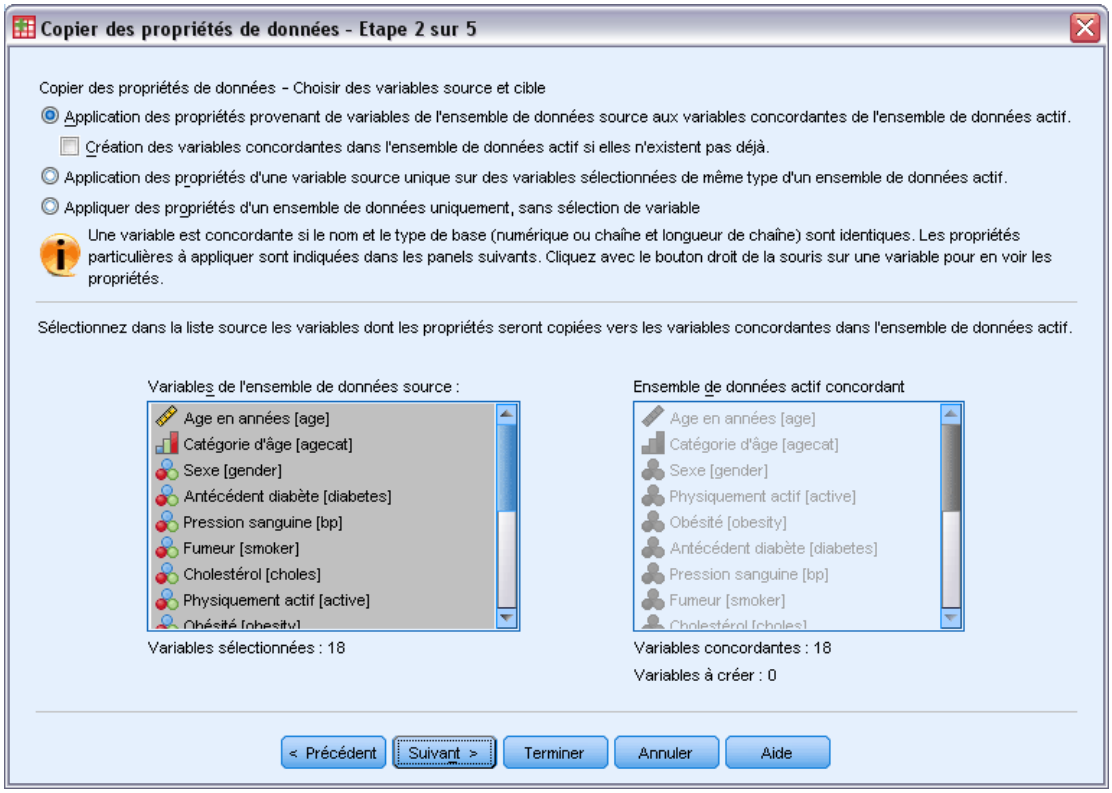
Figure 7-6
Copier des propriétés de données, étape 1 (bienvenue)



▶ Choisissez de copier les propriétés à partir du fichier de données IBM® SPSS® Statistics externe, *patient_los.sav*. Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A sur p. 138.

▶ Cliquez sur Suivant.

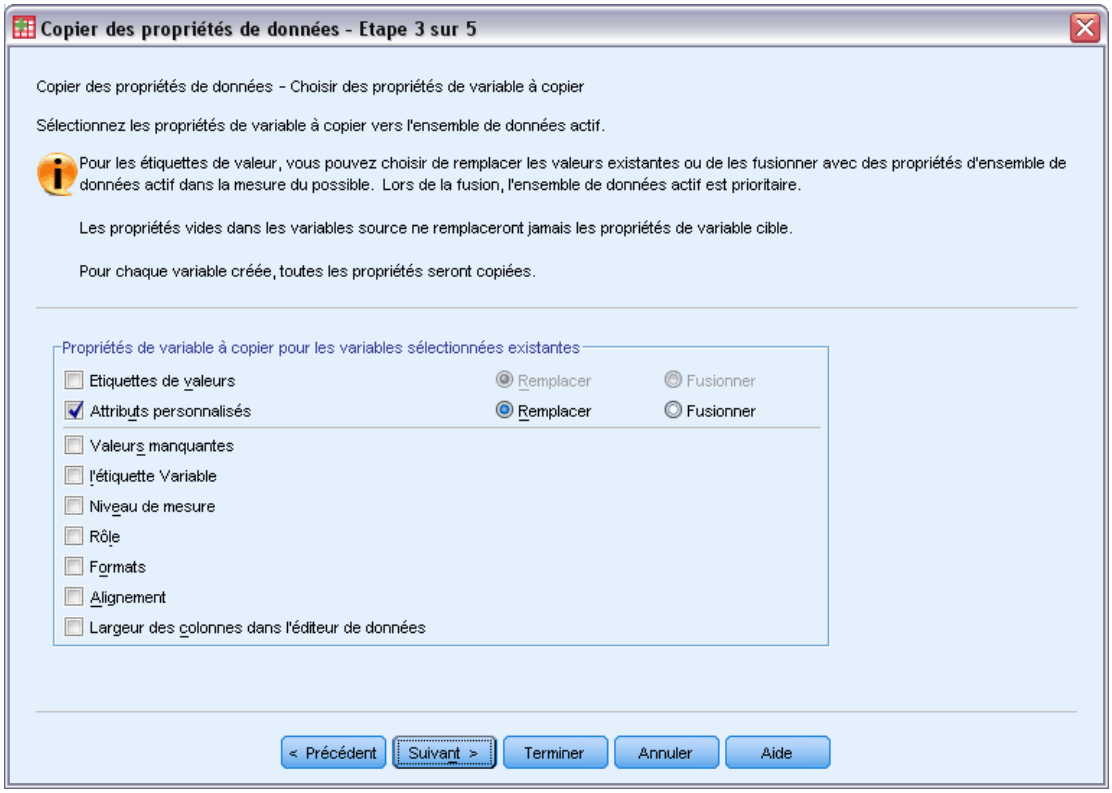
Figure 7-7
Copier des propriétés de données, étape 2 (sélection des variables)



Voici les variables dont vous voulez copier les propriétés à partir de *patient_los.sav* dans les variables correspondantes du fichier *stroke_invalid.sav*.

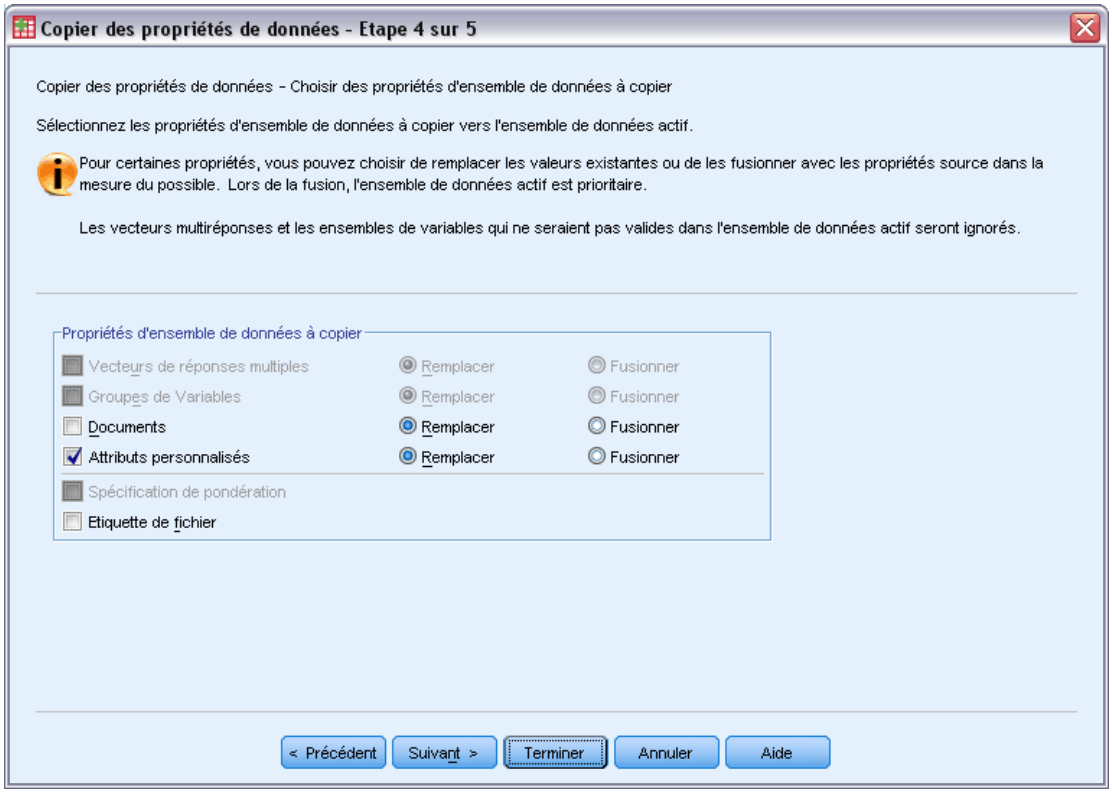
► Cliquez sur Suivant.

Figure 7-8
Copier des propriétés de données, étape 3 (sélection des propriétés des variables)



- Désélectionnez toutes les propriétés excepté Attributs Personnalisés.
- Cliquez sur Suivant.

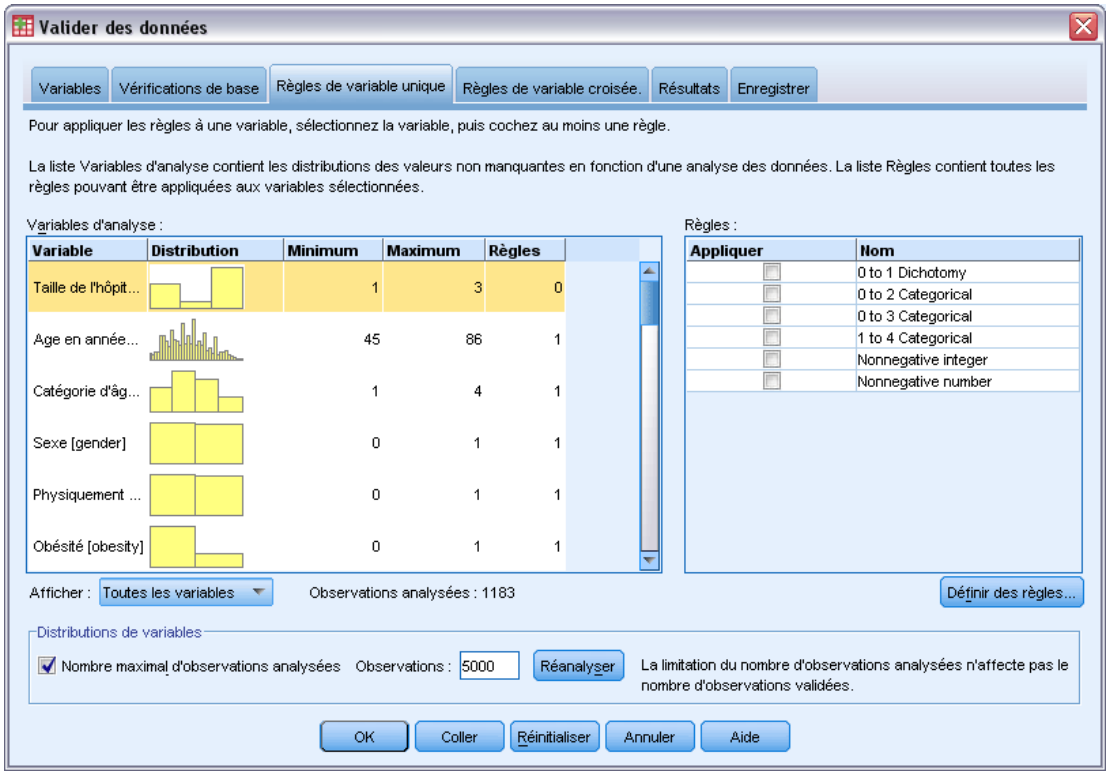
Figure 7-9
Copier des propriétés de données, étape 4 (sélection des propriétés du fichier de données)



- Sélectionnez Attributs Personnalisés.
- Cliquez sur Terminer.

Vous pouvez désormais réutiliser les règles de validation.

Figure 7-10
Boîte de dialogue Valider les données, onglet Règles des variables uniques



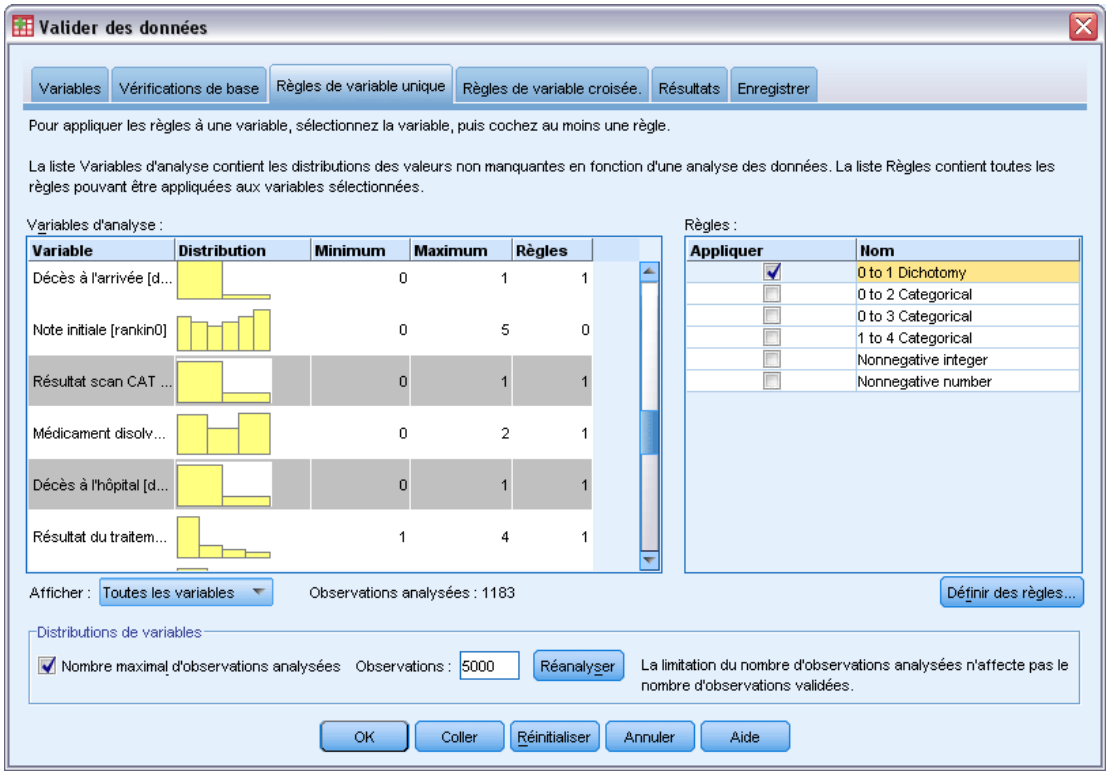
▶ Pour valider les données de *stroke_invalid.sav* à l'aide des règles copiées, cliquez sur le bouton de la barre d'outils Rappeler boîte de dialogue et choisissez Valider des données.

▶ Cliquez sur l'onglet Règles de variable unique.

La liste Variables d'analyse affiche les variables qui sont sélectionnées dans l'onglet Variables, des informations récapitulatives concernant leur distribution et le nombre de règles attachées à chaque variable. Les variables dont les propriétés ont été copiées à partir de *patient_los.sav* ont des règles qui leur sont attachées.

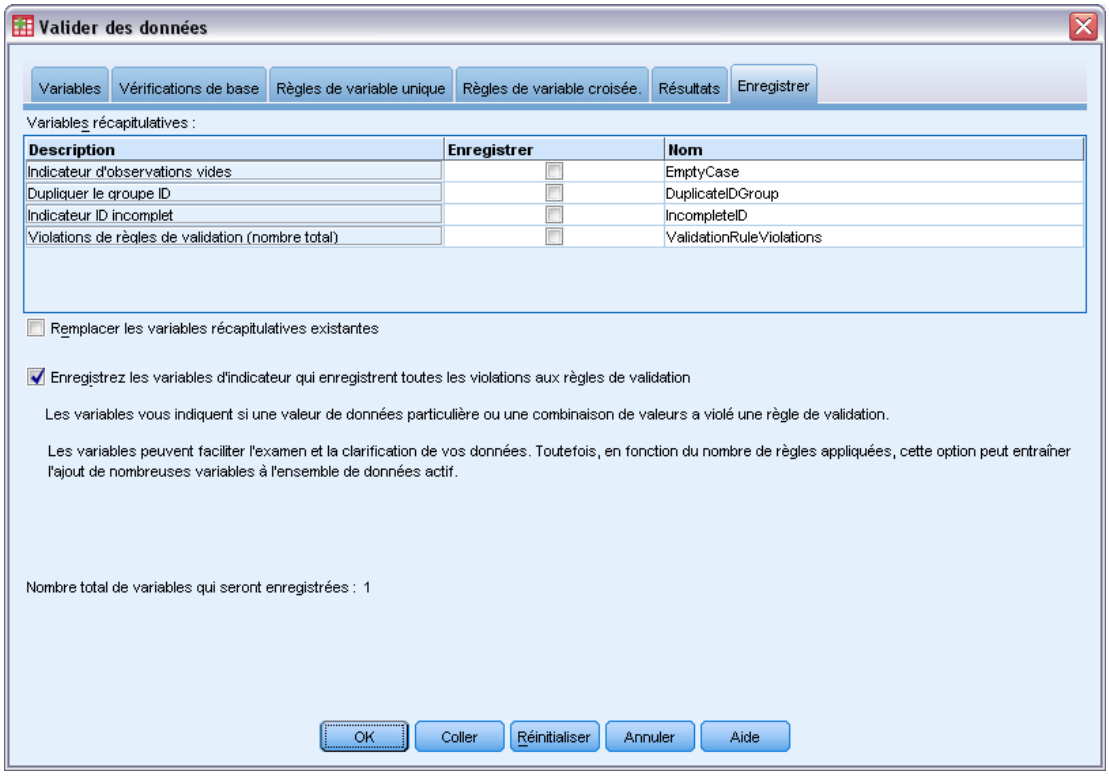
La liste Règles affiche les règles de validation de variable unique disponibles dans le fichier de données. Ces règles ont toutes été copiées à partir de *patient_los.sav*. Notez que certaines de ces règles sont applicables à des variables qui n'ont pas d'équivalents exacts dans l'autre fichier de données.

Figure 7-11
Boîte de dialogue Valider les données, onglet Règles des variables uniques



- Sélectionnez *Fibrillation auriculaire*, *Antécédents d'accident ischémique transitoire*, *Résultats du scanner X* et *Décédé à l'hôpital*, puis appliquez la règle Dichotomie de 0 à 1.
- Appliquez Qualitatif de 0 à 3 à *Rééducation suite à l'événement*.
- Appliquez Qualitatif de 0 à 2 à *Chirurgie préventive suite à l'événement*.
- Appliquez Entier non négatif à *Durée du séjour nécessaire à la rééducation*.
- Appliquez Qualitatif de 1 à 4 à *Index de Barthel à 1 mois recodé* via *Index de Barthel à 6 mois recodé*.
- Cliquez sur l'onglet Enregistrer.

Figure 7-12
Boîte de dialogue Valider les données, onglet Enregistrer



- Sélectionnez Enregistrez les variables d'indicateur qui enregistrent toutes les violations aux règles de validation. Ce processus simplifiera la connexion de l'observation et de la variable causant les violations des règles à variable unique.
- Cliquez sur OK.

Descriptions des règles

Figure 7-13
Descriptions des règles

Règle	Description
Nonnegative integer	Type: Numeric Domain: Range Flag user-missing values: No Flag system-missing values: Yes Minimum: 0 Flag unlabeled values within range: No Flag noninteger values within range: Yes $VD.SRule[5]: Rule
0 to 1 Dichotomy	Type: Numeric Domain: List Flag user-missing values: No Flag system-missing values: Yes List: 0; 1 $VD.SRule[1]: Rule
1 to 4 Categorical	Type: Numeric Domain: List Flag user-missing values: No Flag system-missing values: Yes List: 1; 2; 3; 4 $VD.SRule[4]: Rule

Les règles ayant subi au moins une violation apparaissent.

Le tableau de description des règles affiche des explications sur les règles n'ayant pas été respectées. Cette fonction est très utile pour garder en mémoire de nombreuses règles de validation.

Récapitulatif de variables

Figure 7-14
Récapitulatif de variables

	Règle	Nombre de violations
agecat	0 to 1 Dichotomy	
	1 to 4 Categorical	1
	Nonnegative integer	
	Total	1
gender	0 to 1 Dichotomy	1
	1 to 4 Categorical	
	Nonnegative integer	
	Total	1
angina	0 to 1 Dichotomy	1
	1 to 4 Categorical	
	Nonnegative integer	
	Total	1
time	0 to 1 Dichotomy	
	1 to 4 Categorical	
	Nonnegative integer	2
	Total	2
doa	0 to 1 Dichotomy	1
	1 to 4 Categorical	
	Nonnegative integer	
	Total	1

Ce tableau récapitulatif de variables recense les variables qui n'ont pas respecté au moins une règle de validation, ainsi que les règles qui n'ont pas été respectées et le nombre de violations par règle et par variable.

Rapport d'observations

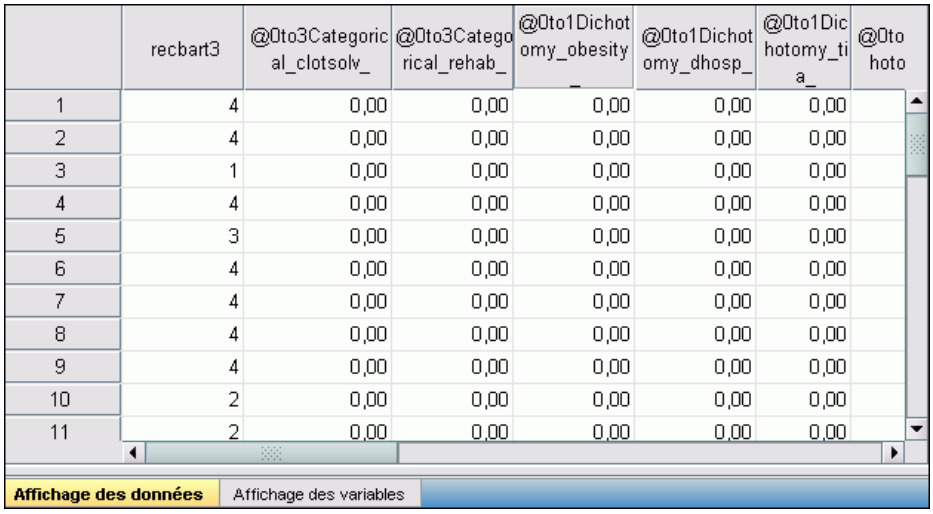
Figure 7-15
Rapport d'observations

	Violations de règle de validation		Identificateur		
Observation	Variable unique[a]	Variable croisée	hospid	patid	physid
175	0 to 1 Dichotomy (1)		OZN	3536728457	615087
274	0 to 1 Dichotomy (1)		OZN	7812188123	355184
310	Nonnegative integer (1)		OZN	6126898743	355184
437	0 to 1 Dichotomy (1)		PBW	7118230827	616528
752	Nonnegative integer (1)		PBW	1646065475	615087
1173	1 to 4 Categorical (3)		PBW	9776111618	616528

a. Le nombre de variables n'ayant pas respecté la règle suit chaque règle.

Le rapport d'observations recense les observations (par le biais de leur numéro et de leur identificateur) qui n'ont pas respecté au moins une règle de validation, les règles n'ayant pas été respectées et le nombre de fois que la règle n'a pas été respectée par l'observation. Les valeurs non valides s'affichent dans l'éditeur de données.

Figure 7-16
Editeur de données avec les indicateurs des violations de règles enregistrés

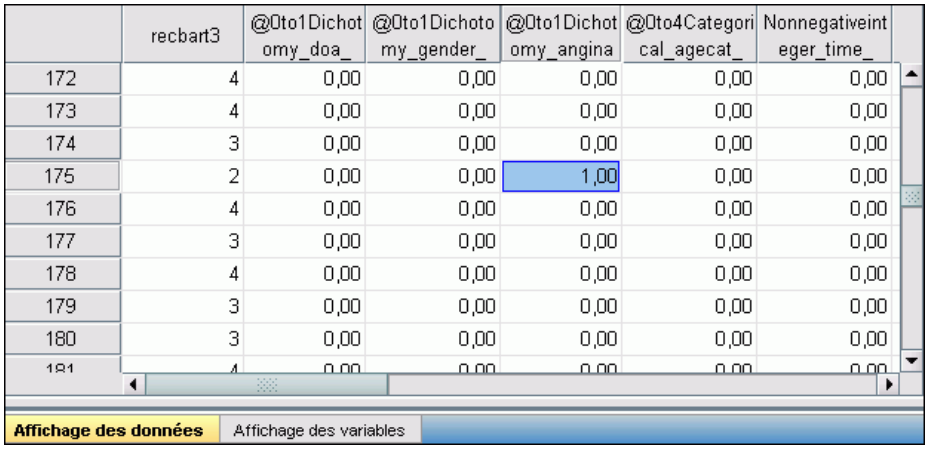


	recbart3	@0to3Categorical_clotsolv_	@0to3Categorical_rehab_	@0to1Dichotomy_obesity_	@0to1Dichotomy_dhosp_	@0to1Dichotomy_tia_	@0toho
1	4	0,00	0,00	0,00	0,00	0,00	
2	4	0,00	0,00	0,00	0,00	0,00	
3	1	0,00	0,00	0,00	0,00	0,00	
4	4	0,00	0,00	0,00	0,00	0,00	
5	3	0,00	0,00	0,00	0,00	0,00	
6	4	0,00	0,00	0,00	0,00	0,00	
7	4	0,00	0,00	0,00	0,00	0,00	
8	4	0,00	0,00	0,00	0,00	0,00	
9	4	0,00	0,00	0,00	0,00	0,00	
10	2	0,00	0,00	0,00	0,00	0,00	
11	2	0,00	0,00	0,00	0,00	0,00	

Affichage des données | Affichage des variables

Une variable d'indicateur séparée est produite pour chaque application d'une règle de validation. Ainsi, *@0to3Categorical_clotsolv_* est l'application de la règle de validation à variable unique Qualitatif de 0 à 3 à la variable *Anticoagulants*. Pour une observation donnée, la façon la plus simple de trouver quelle valeur de variable est non valide consiste à analyser les valeurs des indicateurs. La valeur 1 signifie que la valeur de la variable associée est non valide.

Figure 7-17
Editeur de données avec un indicateur de violation de règle pour l'observation 175

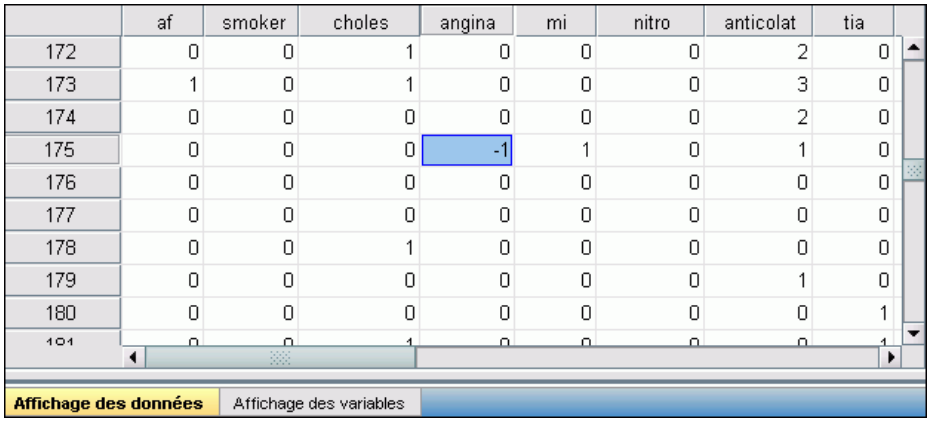


	recbart3	@0to1Dichotomy_doa_	@0to1Dichotomy_gender_	@0to1Dichotomy_angina	@0to4Categorical_agecat_	Nonnegativeinteger_time_
172	4	0,00	0,00	0,00	0,00	0,00
173	4	0,00	0,00	0,00	0,00	0,00
174	3	0,00	0,00	0,00	0,00	0,00
175	2	0,00	0,00	1,00	0,00	0,00
176	4	0,00	0,00	0,00	0,00	0,00
177	3	0,00	0,00	0,00	0,00	0,00
178	4	0,00	0,00	0,00	0,00	0,00
179	3	0,00	0,00	0,00	0,00	0,00
180	3	0,00	0,00	0,00	0,00	0,00
181	4	0,00	0,00	0,00	0,00	0,00

Affichage des données | Affichage des variables

Etudiez l'observation 175, la première observation avec une règle violée. Pour accélérer votre recherche, regardez les indicateurs associés aux variables dans le tableau récapitulatif de variables. Il est facile de voir que c'est *Antécédents d'angine* qui comporte la valeur non valide.

Figure 7-18
Editeur de données avec valeur non valide pour Antécédents d'angine



	af	smoker	choles	angina	mi	nitro	anticolat	tia
172	0	0	1	0	0	0	2	0
173	1	0	1	0	0	0	3	0
174	0	0	0	0	0	0	2	0
175	0	0	0	-1	1	0	1	0
176	0	0	0	0	0	0	0	0
177	0	0	0	0	0	0	0	0
178	0	0	1	0	0	0	0	0
179	0	0	0	0	0	0	1	0
180	0	0	0	0	0	0	0	1

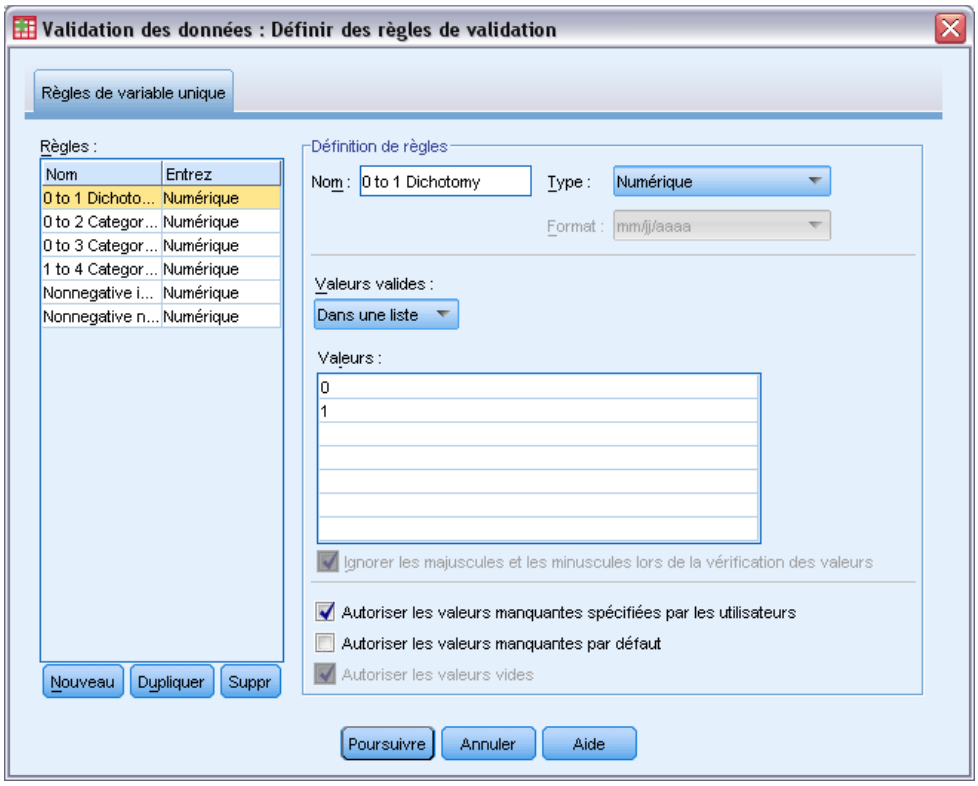
Antécédents d'angine a une valeur de –1. Bien que cette valeur soit une valeur manquante valide pour les variables de traitement et de résultats dans le fichier de données, elle est non valide dans ce cas parce que les valeurs des antécédents du patient n'ont pas de valeurs manquantes utilisateur définies.

Définition de vos propres règles

Les règles de validation copiées à partir de *patient_los.sav* ont été très utiles, mais il vous faut définir quelques règles supplémentaires pour finir le travail. Par ailleurs, il arrive que certains patients déjà décédés à leur arrivée soient accidentellement marqués comme étant décédés à l'hôpital. Les règles de validation de variable unique ne peuvent pas prendre en compte cette situation : vous devez donc définir une règle de variable croisée pour gérer cette situation.

- Cliquez sur le bouton de la barre d'outils Rappeler boîte de dialogue et choisissez Valider des données.
- Cliquez sur l'onglet Règles de variable unique. (Vous devez définir des règles pour la *Taille de l'hôpital*, les variables mesurant les scores de Rankin et celles correspondant aux index de Barthel non recodés).
- Cliquez sur Définir des règles.

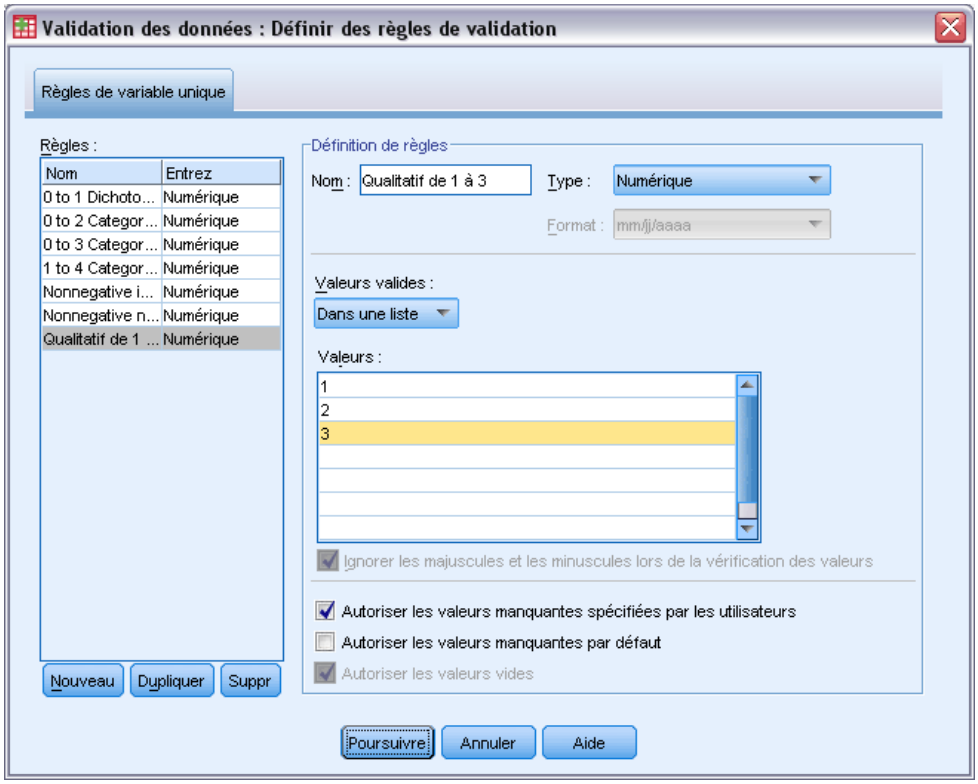
Figure 7-19
Boîte de dialogue Définir des règles de validation, onglet Règles des variables uniques



Les règles actuellement définies sont affichées avec une Dichotomie de 0 à 1 sélectionnée dans la liste des règles et les propriétés de la règle affichées dans le groupe Définition de règles.

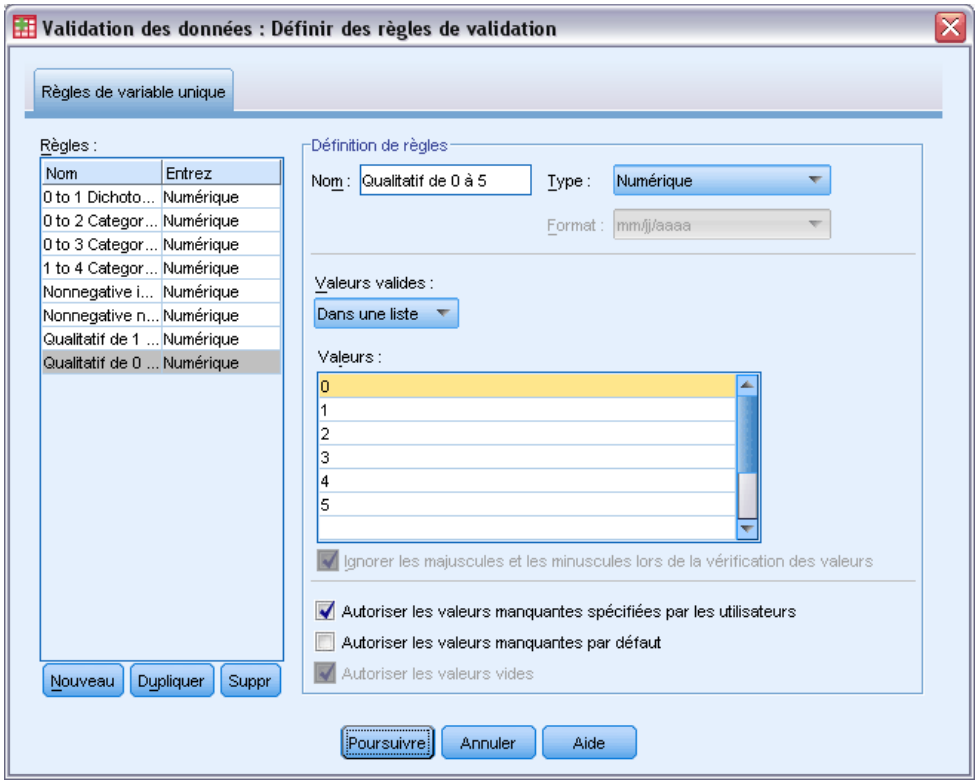
▶ Pour définir une règle, cliquez sur Nouveau.

Figure 7-20
Boîte de dialogue Définir des règles de validation, onglet Règles des variables uniques (Qualitatif de 1 à 3)



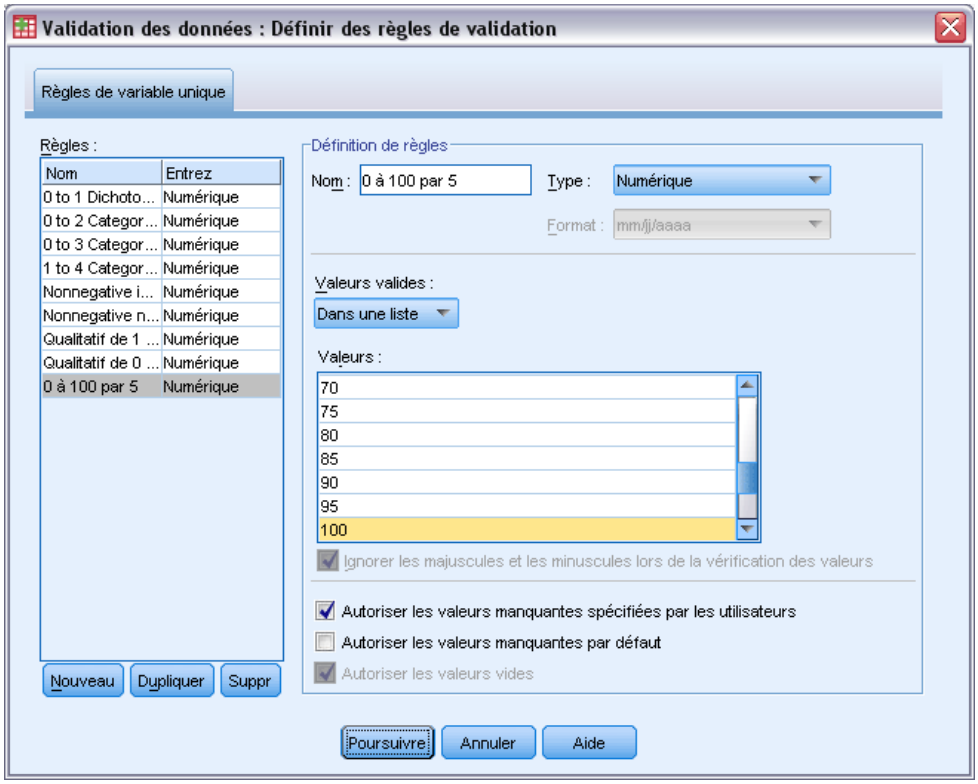
- Entrez Qualitatif de 1 à 3 comme nom de règle.
- Pour Valeurs valides, choisissez Dans une liste.
- Entrez 1, 2 et 3 comme valeurs.
- Désélectionnez Autoriser les valeurs manquantes par défaut.
- Pour définir la règle concernant les scores de Rankin, cliquez sur Nouveau.

Figure 7-21
Boîte de dialogue Définir des règles de validation, onglet Règles des variables uniques (Qualitatif de 0 à 5)



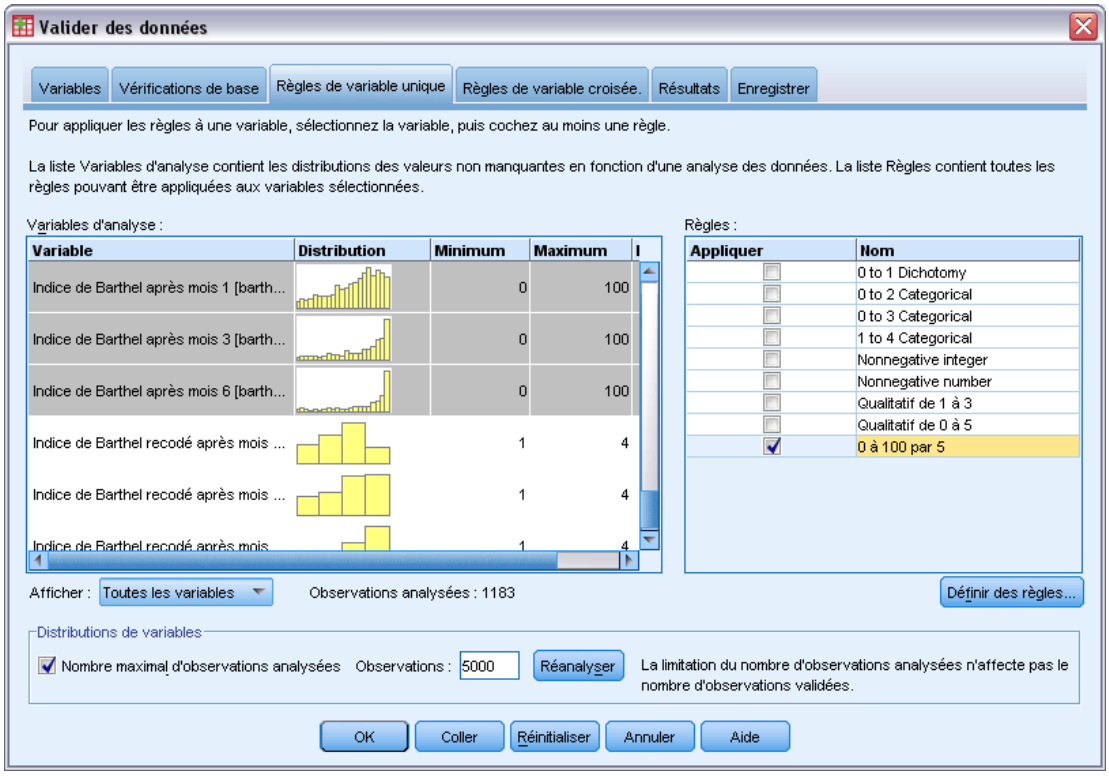
- Entrez Qualitatif de 0 à 5 comme nom de règle.
- Pour Valeurs valides, choisissez Dans une liste.
- Entrez 0, 1, 2, 3, 4 et 5 comme valeurs.
- Désélectionnez Autoriser les valeurs manquantes par défaut.
- Pour définir la règle concernant les index de Barthel, cliquez sur Nouveau.

Figure 7-22
Boîte de dialogue Définir des règles de validation, onglet Règles des variables uniques (0 à 100 par 5 définis)



- Entrez 0 à 100 par 5 comme nom de règle.
- Pour Valeurs valides, choisissez Dans une liste.
- Entrez 0, 5, ... et 100 comme valeurs.
- Désélectionnez Autoriser les valeurs manquantes par défaut.
- Cliquez sur Poursuivre.

Figure 7-23
Boîte de dialogue Valider les données, onglet Règles des variables uniques (0 à 100 par 5 définis)



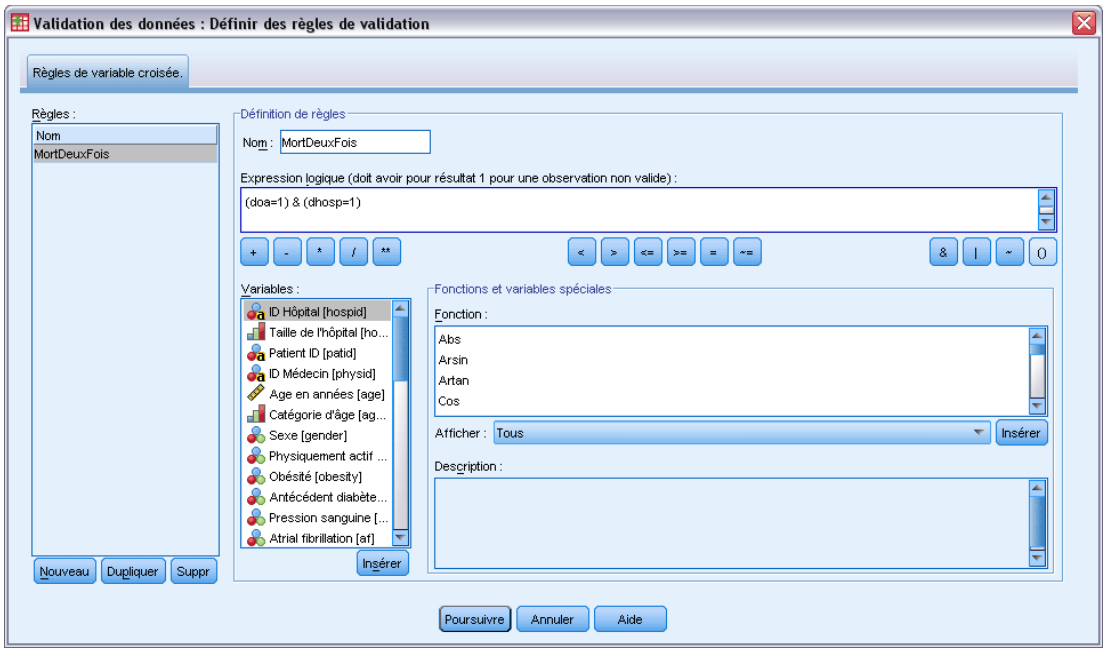
Vous devez maintenant appliquer les règles définies aux variables d'analyse.

- Appliquez Qualitatif de 1 à 3 à *Taille de l'hôpital*.
- Appliquez Qualitatif de 1 à 5 à *Score de Rankin initial* et *Score de Rankin à 1 mois* via *Score de Rankin à 6 mois*.
- Appliquez 0 à 100 par 5 à *Index de Barthel à 1 mois* via *Index de Barthel à 6 mois*.
- Cliquez sur l'onglet Règles de variable croisée.

Il n'y a pas de règles définies actuellement.

- Cliquez sur Définir des règles.

Figure 7-24
Boîte de dialogue Définir des règles de validation, onglet Règles des variables croisées



Lorsqu'il n'y a pas de règle, une nouvelle règle de substitution est automatiquement créée.

- Entrez DiedTwice (MortDeuxFois) comme nom de règle.
- Entrez (doa=1) & (dhosp=1) comme expression logique. La valeur 1 apparaîtra si le patient est enregistré comme étant décédé à la fois avant son arrivée à l'hôpital et à l'hôpital.
- Cliquez sur Poursuivre.

  La règle nouvellement définie est automatiquement sélectionnée dans l'onglet Règles de variable croisée.
- Cliquez sur OK.

Règles de variable croisée.

Figure 7-25
Règles de variable croisée

Règle	Nombre de violations	Expression de règle
Mort deux Fois	27	(doa = 1) & (dhosp = 1)

Le récapitulatif de règles de variable croisée recense les règles de variables croisées n'ayant pas été respectées au moins une fois, le nombre de violations et une description de chaque règle non respectée.

Rapport d'observations

Figure 7-26
Rapport d'observations

	Violations de règle de validation		Identificateur		
Observation	Variable unique[a]	Variable croisée	hospid	patid	physid
20		Mort deux Fois	PBW	1192970826	355184
49		Mort deux Fois	NHV	8717862852	237418
129		Mort deux Fois	QWS	6901932085	215041
138		Mort deux Fois	RLD	1205005069	695521
162		Mort deux Fois	OZN	5546809538	125304
175	0 to 1 Dichotomy (1)		OZN	0333204686	883285
274	0 to 1 Dichotomy (1)		OZN	1038840465	103254
310	Nonnegative integer (1)		OZN	2090290204	883285
414		Mort deux Fois	WPA	3351107142	462020
437	0 to 1 Dichotomy (1)		WPA	2349729006	723384
447		Mort deux Fois	WPA	7163481282	519548
458		Mort deux Fois	WPA	9159094175	652070
462		Mort deux Fois	WPA	2137520354	723384
537		Mort deux Fois	SLB	5246122506	928076
544		Mort deux Fois	SLB	1605957462	506108
620		Mort deux Fois	GFG	8141858966	828754
629		Mort deux Fois	GFG	3397891610	539412
630		Mort deux Fois	GFG	3397891610	539412
639		Mort deux Fois	GFG	3962622031	327422
644		Mort deux Fois	GFG	4271782383	749432
649		Mort deux Fois	GFG	0950686750	618069
653		Mort deux Fois	GFG	0663642766	001448
722		Mort deux Fois	GFG	0418125590	877354
748		Mort deux Fois	GFG	8744721380	539412
752	Nonnegative integer (1) 0 to 1 Dichotomy (3)		GFG	4993307441	828754
868		Mort deux Fois	WWL	9714672452	237547
881		Mort deux Fois	WWL	6613279456	574275
915		Mort deux Fois	EFX	2575793702	501318
933		Mort deux Fois	IZO	2807437472	680253
1010		Mort deux Fois	BLA	5284009939	657638

Le rapport d'observations inclut à présent les observations qui ont violé une règle de variable croisée ainsi que les observations déjà recensées comme ayant violé des règles de variable unique. Ces observations doivent toutes être consignées pour être corrigées par l'équipe chargée de la saisie de données.

Récapitulatif

L'analyste possède toutes les informations nécessaires pour rendre un rapport préliminaire au chef d'équipe chargé de la saisie des données.

Procédures apparentées

La procédure Valider des données est un outil utile au contrôle de la qualité des données.

- La procédure Identification des observations inhabituelles analyse les modèles dans vos données et identifie les observations ayant quelques valeurs significatives variant selon le type.

Chapitre
8

Préparation automatique des données

La préparation des données pour l'analyse est une des étapes les plus importantes des projets—et généralement, l'une de celles qui prend le plus de temps. La préparation automatique des données (ADP) s'occupe de cette tâche à votre place, analyse vos données, identifie les corrections, supprime les champs problématiques ou inutiles, dérive de nouveaux attributs si nécessaire et améliore les performances grâce à des techniques d'analyse intelligentes. Vous pouvez utiliser l'algorithme en mode complètement **automatique**, le laissant choisir et appliquer les corrections ou vous pouvez utiliser son mode **interactif** qui prévoit les modifications avant qu'elles ne soient effectuées vous laissant libre de les accepter ou de les refuser.

L'utilisation de l'ADP vous permet de préparer facilement et rapidement vos données pour la création de modèle, sans qu'il soit nécessaire de maîtriser les concepts de statistiques utilisés. Les modèles seront alors créés et les scores déterminés plus rapidement ; de plus, l'utilisation de l'ADP améliore la robustesse des processus de modélisation automatique.

Utilisation interactive de la préparation automatique des données

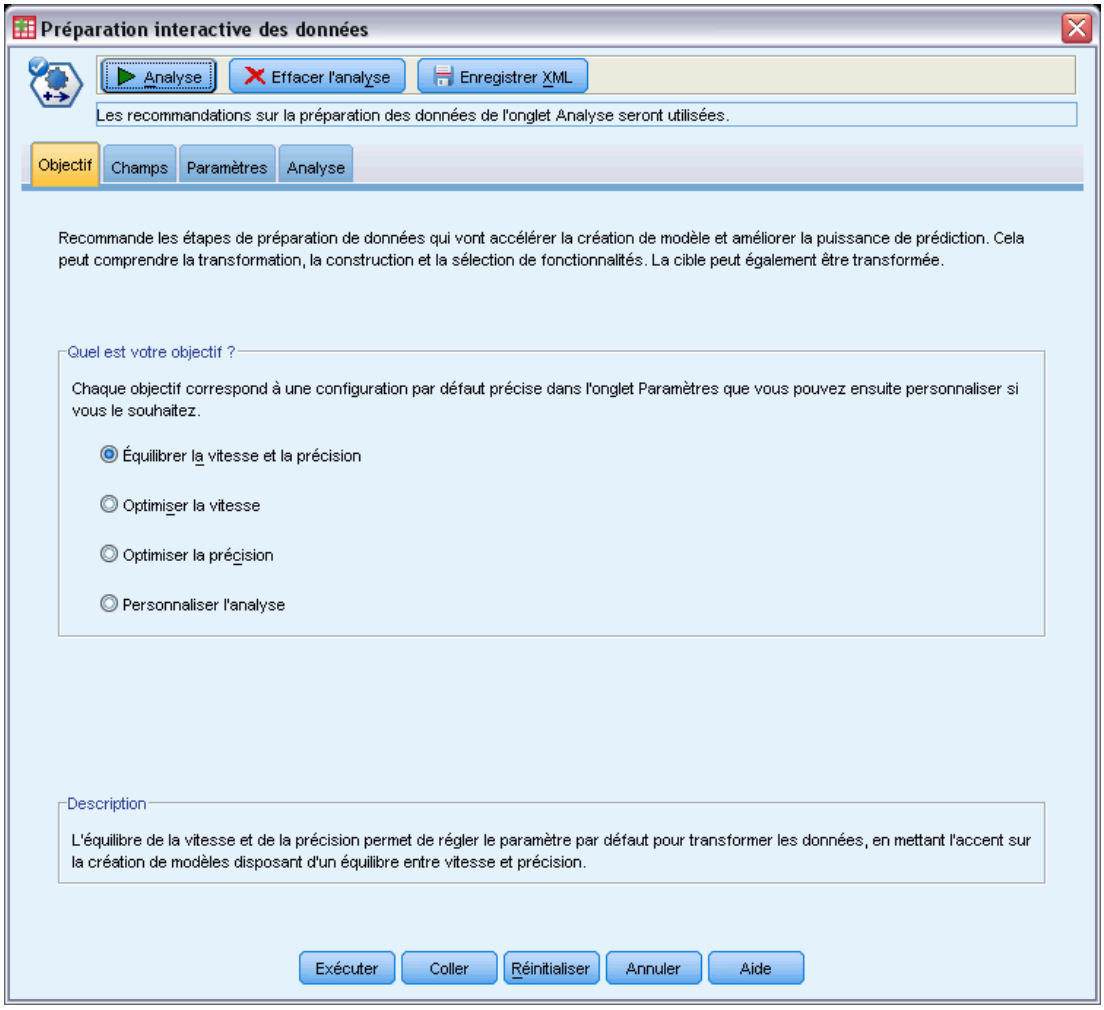
Une compagnie d'assurances disposant de ressources restreintes pour enquêter sur les demandes de remboursement des propriétaires de biens immobiliers, souhaite construire un modèle pour rechercher les demandes suspectes et potentiellement frauduleuses. La compagnie dispose d'informations provenant de demandes précédentes dans le fichier *insurance_claims.sav*. Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A sur p. 138. Avant de construire le modèle, il est nécessaire de préparer les données à l'aide de la préparation automatique des données. La compagnie souhaitant être capable de consulter et modifier les transformations avant de les appliquer, elle utilise la préparation automatique des données de manière interactive.

Choix des objectifs

- Pour exécuter la préparation automatique des données interactive, sélectionnez à partir des menus : Transformer > Préparer les données pour la modélisation > Interactif...

© Copyright SPSS Inc. 1989, 2010

Figure 8-1
Onglet Objectif



Le premier onglet requiert un objectif qui contrôle les paramètres par défaut. Mais quel est la différence entre les différents objectifs ? Lorsque la procédure est exécutée à l'aide de chacun des objectifs, il est possible de voir la manière dont les résultats diffèrent.

► Assurez-vous que l'option Équilibrer la vitesse et la précision est sélectionnée et cliquez sur Analyser.

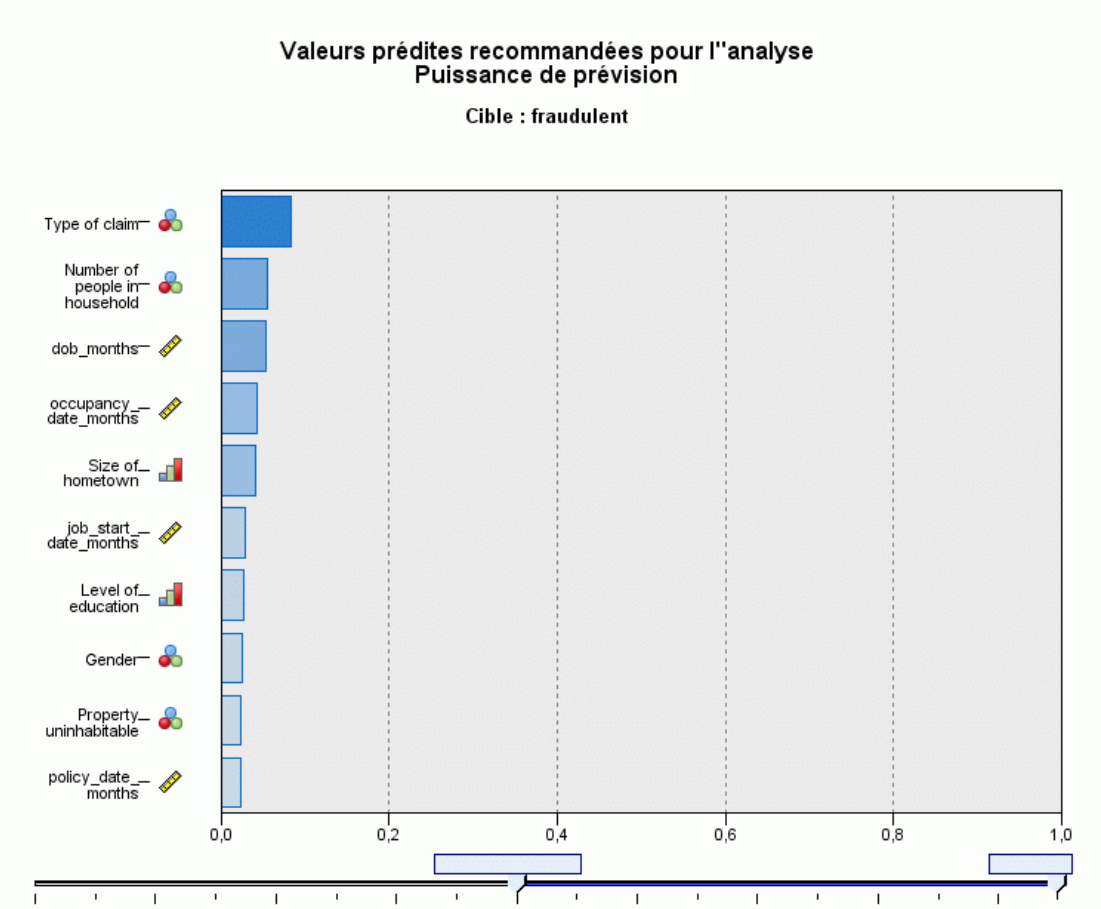
Figure 8-2
Onglet Analyse, récapitulatif de traitement des champs pour objectif équilibré

Récapitulatif du traitement de champ

Champs		N
Cible		1
Valeurs prédites		5
Valeurs prédites recommandées pour une utilisation dans l"analyse	Total	5
	Champs d"origine (non transformés)	4
	Transformations des champs d"origine	0
	Calculés d"après les dateset les heures	1
	Construit	0
Valeurs prédites non utilisées		0

Le focus est automatiquement dirigé sur l'onglet Analyse lors du traitement des données par la procédure. La vue principale par défaut est celle du Récapitulatif de traitement des champs. Elle vous offre une vue générale de la manière dont les champs ont été traités par la préparation automatique des données. Il existe une cible unique, 18 entrées et 18 champs recommandés pour la construction de modèle. Parmi les champs recommandés pour la modélisation, 9 sont des champs d'entrées d'origine, 4 sont des transformations de champs d'entrées d'origine, et 5 sont dérivés de champs de date ou d'heure.

Figure 8-3
Onglet Analyse, puissance de prédiction pour objectif équilibré



La vue auxiliaire par défaut est celle de la Puissance de prédiction. Elle vous offre une idée rapide des champs recommandés les plus utiles à la construction du modèle. Remarque : alors que 18 variables prédites sont recommandées pour l'analyse, seules les 10 premières sont affichées par défaut dans le diagramme de la Puissance de prédiction. Pour afficher un nombre plus élevé ou moins élevé de champs, utilisez le curseur situé en dessous du diagramme.

Avec Équilibrer la vitesse et la précision comme objectif, *Type de réclamation* est identifié comme la meilleure variable prédite, suivi par le *Nombre de personnes dans le ménage* et l'âge de l'assuré en mois (calculé sur la base de la date de naissance et de la date actuelle).

- Cliquez sur Effacer l'analyse, puis sur l'onglet Objectif.
- Sélectionnez Optimiser la vitesse, puis cliquez sur Analyser.

Figure 8-4
Onglet Analyse, récapitulatif de traitement des champs lors de l'optimisation de la vitesse

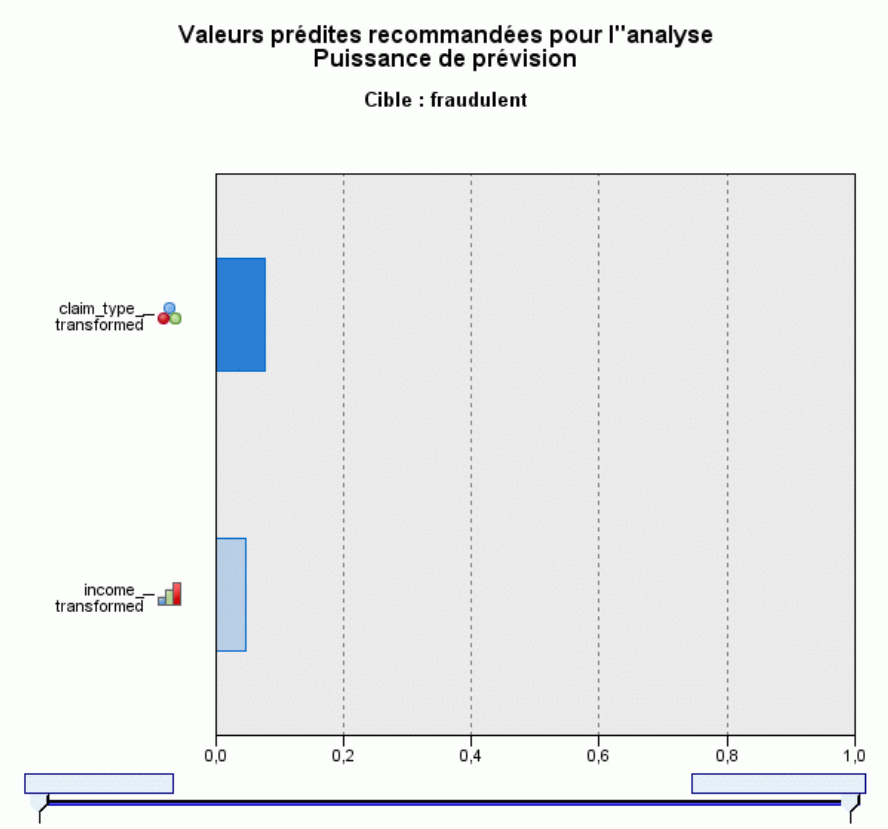
Récapitulatif du traitement de champ

Champs		N
Cible		1
Valeurs prédites		18
Valeurs prédites recommandées pour une utilisation dans l"analyse	Total	2
	Champs d"origine (non transformés)	0
	Transformations des champs d"origine	2
	Calculés d"après les dateset les heures	0
	Construit	0
Valeurs prédites non utilisées		16

- Aucune valeur prédite utile n"a pu être construite. Les raisons les plus courantes sont : trop peu de valeurs prédites continues étaient fortement associées avec la cible ou toutes les valeurs prédites continues étaient indépendantes.

Le focus est de nouveau automatiquement dirigé sur l'onglet Analyse lors du traitement des données par la procédure. Dans ce cas, seuls 2 champs sont recommandés pour la construction du modèle et tous deux sont des transformations de champs d'origine.

Figure 8-5
Onglet Analyse, puissance de prédiction lors de l'optimisation de la vitesse



Avec l'option Optimiser la vitesse comme objectif, *claim_type_transformed* est identifié comme la meilleure variable prédite, suivi par *income_transformed*.

- Cliquez sur Effacer l'analyse, puis sur l'onglet Objectif.
- Sélectionnez Optimiser la précision, puis cliquez sur Analyser.

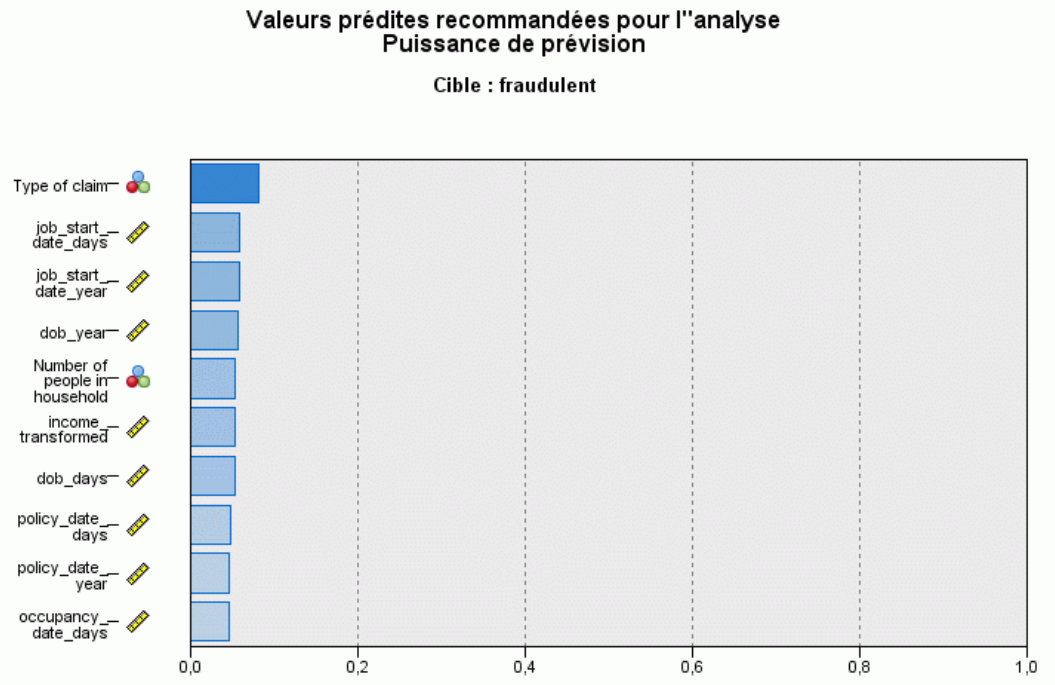
Figure 8-6
Onglet Analyse, puissance de prédiction lors de l'optimisation de la précision

Récapitulatif du traitement de champ

Champs		N
Cible		1
Valeurs prédites		18
Valeurs prédites recommandées pour une utilisation dans l"analyse	Total	32
	Champs d"origine (non transformés)	9
	Transformations des champs d"origine	4
	Calculés d"après les dateset les heures	19
	Construit	0
Valeurs prédites non utilisées		0

Avec Optimiser la précision comme objectif, 32 champs sont recommandés pour la construction du modèle, car davantage de champs sont dérivés de dates et d'heures, à cause de l'extraction des jours, mois et années des dates et de l'extraction des heures, minutes et secondes des heures.

Figure 8-7
Onglet Analyse, puissance de prédiction lors de l'optimisation de la précision



Type de réclamation est identifié comme la meilleure variable prédite suivi par le nombre de jours depuis que l'assuré a commencé son dernier emploi (calculé sur la base de la date de début de l'emploi et la date en cours) et l'année durant laquelle l'assuré a commencé son emploi actuel (donnée extraite de la date de début de l'emploi).

Pour récapituler :

- Équilibrer la vitesse et la précision crée des champs utilisables dans la modélisation à partir de dates, et permet de transformer des champs continus comme *reside* pour rendre leur distribution plus normale.
- Optimiser la précision crée des champs supplémentaires à partir de dates (et vérifie s'il existe des valeurs éloignées, et si la cible est continue, permet de la transformer pour rendre sa distribution plus normale).
- Optimiser la vitesse ne prépare pas de dates et ne redimensionne pas de champs continus, mais fusionne les catégories des variables prédites qualitatives et des variables continues lorsque la cible est qualitative (et réalise la sélection et la construction des caractéristiques si la cible est continue).

La compagnie d'assurance décide d'explorer les résultats de l'objectif Optimiser la précision plus avant.

▶ Sélectionnez Champs dans la liste déroulante de la vue principale.

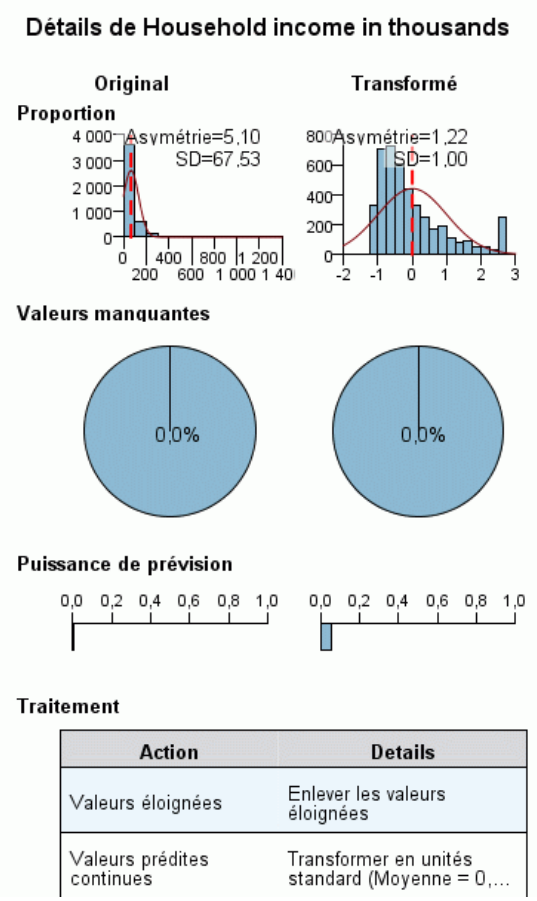
Champs et Détails des champs

Figure 8-8
Champs

La vue Champs affiche les champs traités et si l'ADP recommande de les utiliser dans la construction du modèle. Cliquez sur un nom de champ pour afficher plus d'informations sur le champ dans la vue associée.

▶ Cliquez sur income.

Figure 8-9
Détails du champ Revenu du ménage en milliers

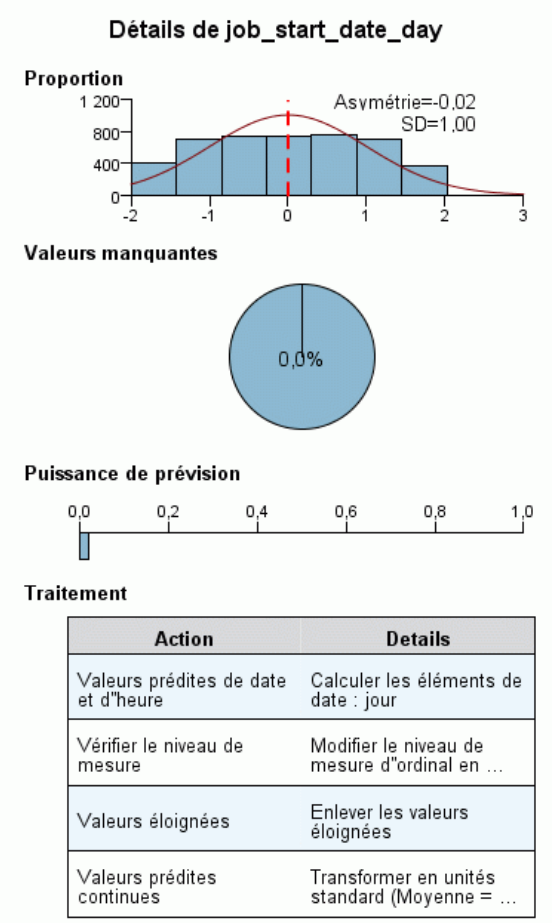


La vue Détails du champ affiche les distributions du champ *Revenu du ménage en milliers* d'origine et transformé. Selon le tableau du traitement, les enregistrements identifiés comme valeurs éloignées ont été éliminés (en définissant leurs valeurs sur la valeur de césure des valeurs éloignées) et le champ a été standardisé pour avoir une moyenne de 0 et un écart-type 1. La " bosse " à l'extrême droite de l'histogramme des champs transformés montre qu'un certain nombre d'enregistrements, probablement plus de 200, ont été identifiés comme des valeurs éloignées. Le revenu a une distribution très asymétrique, il peut s'agir d'un cas pour lequel la césure par défaut est trop agressive dans la définition des valeurs éloignées.

Veuillez noter l'augmentation de la puissance de prédiction du champ transformé par rapport au champ d'origine. Il apparaît que la transformation est ici utile.

▶ Dans la vue Champs, cliquez sur job_start_date_day. (remarque : ce champ est différent de *job_start_date_days*)

Figure 8-10
Détails du champ " job_start_date_day "



Le champ *job_start_date_day* correspond au jour extrait de la *Date d'embauche [job_start_date]*. Il est très peu probable que ce champ ait une quelconque incidence sur le fait qu'une réclamation soit frauduleuse, par conséquent la compagnie souhaite ne pas le prendre en compte dans la construction du modèle.

Figure 8-11
Détails du champ Revenu du ménage en milliers

- Dans la vue Champs, sélectionnez Ne pas utiliser dans la liste déroulante de la version à utiliser de la ligne job_start_date_day. Effectuez de même pour tous les champs comprenant les suffixes *_day* et *_month* .

- Pour appliquer les transformations, cliquez sur Exécuter.

L'ensemble de données est maintenant prêt pour la construction du modèle, du fait que toutes les variables prédites recommandées (aussi bien anciennes que nouvelles) ont leur rôle défini sur Entrée et que toutes les variables prédites non recommandées ont leur rôle défini sur Aucune. Pour créer un ensemble de données uniquement à l'aide des variables prédites recommandées, utilisez les paramètres Appliquer les transformations dans la boîte de dialogue.

Utilisation automatique de la préparation automatique des données

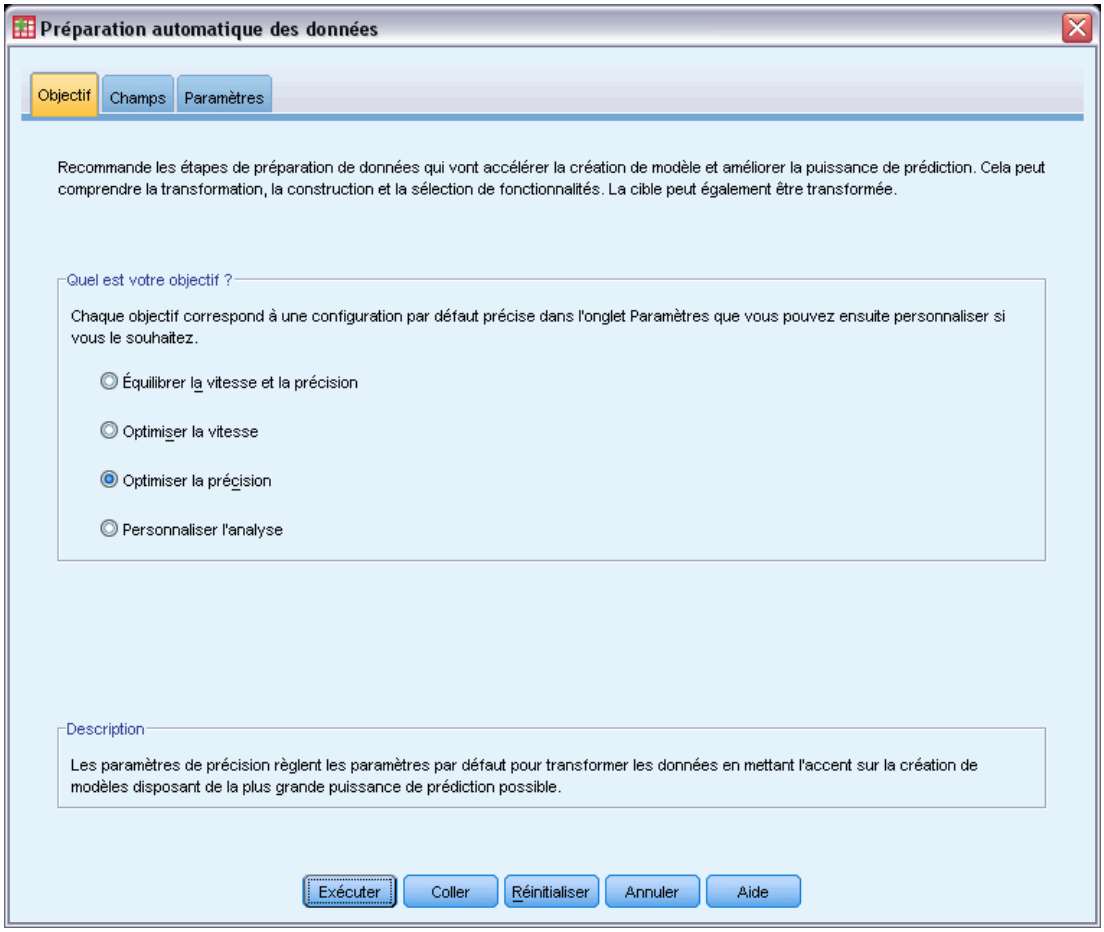
Un groupe automobile suit les ventes de véhicules automobiles personnels divers. Afin d'être en mesure d'identifier les modèles dont les ventes sont très satisfaisantes et ceux pour lesquels elles le sont moins, vous voulez établir une relation entre les ventes de véhicules et les caractéristiques des véhicules. Ces informations sont rassemblées dans le fichier *car_sales_unprepared.sav*. Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A sur p. 138. Utilisez la préparation automatique des données pour préparer les données à analyser. Vous pouvez également créer des modèles à l'aide des données " avant" et " après " la préparation, afin de pouvoir comparer les résultats.

Préparation des données

- Pour exécuter la préparation automatique des données de manière automatique, sélectionnez à partir des menus :

  Transformer > Préparer les données pour la modélisation > Automatique...

Figure 8-12
Onglet Objectif

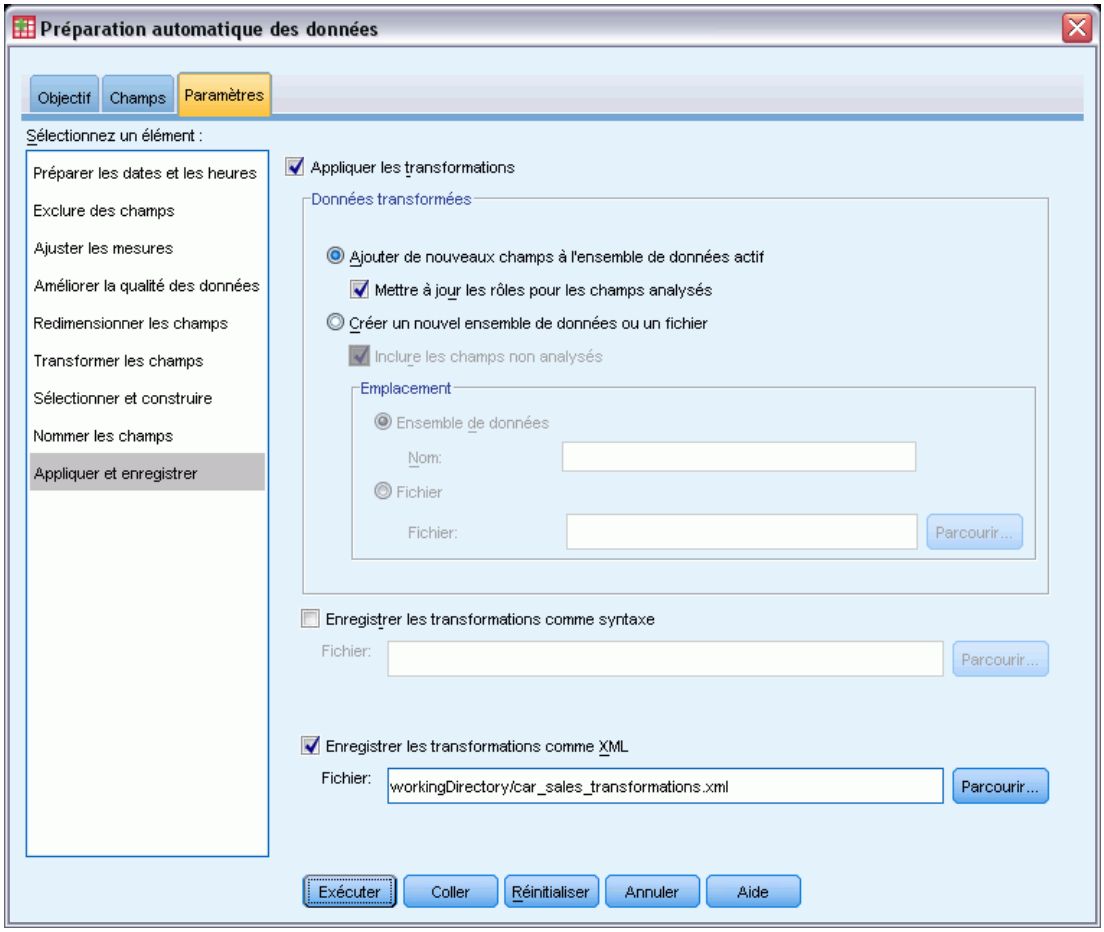


▶ Sélectionnez Optimiser la précision.

Le champ cible *Ventes en milliers* étant un champ continu qui peut être transformé lors de la préparation automatique des données, vous voulez enregistrer les transformations dans un fichier XML afin d'utiliser la boîte de dialogue Rétablir les scores et convertir les valeurs prédites de la cible transformée à leur échelle d'origine.

▶ Cliquez sur l'onglet Paramètres, puis sur Appliquer et enregistrer.

Figure 8-13
Paramètres Appliquer et enregistrer



- Sélectionnez Enregistrer les transformations comme XML et cliquez sur Parcourir pour accéder au fichier workingDirectory/car_sales_transformations.xml, où workingDirectory est le répertoire de destination du fichier à enregistrer.
- Cliquez sur Exécuter.

Ces sélections génèrent la syntaxe de commande suivante :

```
*Automatic Data Preparation.
ADP
  /FIELDS TARGET=sales INPUT=resale type price engine_s horsepow wheelbas width length
    curb_wgt fuel_cap mpg
  /PREPDATETIME DATEDURATION=YES(REFERENCE=YMD('2009-06-04') UNIT=AUTO)
    TIMEDURATION=YES(REFERENCE=HMS('08:43:35') UNIT=AUTO) EXTRACTYEAR=YES(SUFFIX='_year')
    EXTRACTMONTH=YES(SUFFIX='_month') EXTRACTDAY=YES(SUFFIX='_day')
    EXTRACTHOUR=YES(SUFFIX='_hour') EXTRACTMINUTE=YES(SUFFIX='_minute')
    EXTRACTSECOND=YES(SUFFIX='_second')
   /SCREENING PCTMISSING=YES(MAXPCT=50) UNIQUECAT=YES(MAXCAT=100) SINGLECAT=NO
  /ADJUSTLEVEL INPUT=YES TARGET=YES MAXVALORDINAL=10 MINVALCONTINUOUS=5
  /OUTLIERHANDLING INPUT=YES TARGET=NO CUTOFF=SD(3) REPLACEWITH=CUTOFFVALUE
  /REPLACEMISSING INPUT=YES TARGET=NO
  /REORDERNOMINAL INPUT=YES TARGET=NO
```

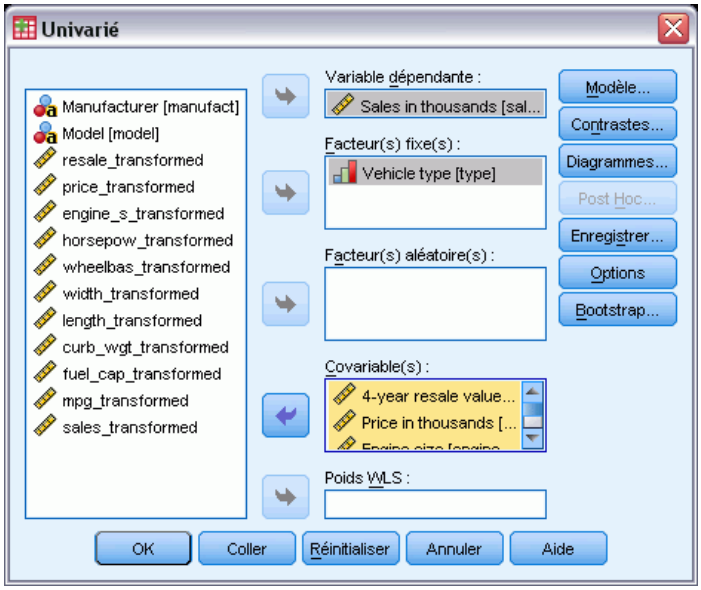
```
  /RESCALE INPUT=ZSCORE(MEAN=0 SD=1) TARGET=BOXCOX(MEAN=0 SD=1)
  /TRANSFORM MERGESUPERVISED=NO MERGEUNSUPERVISED=NO BINNING=NONE SELECTION=NO
    CONSTRUCTION=NO
  /CRITERIA SUFFIX(TARGET='_transformed' INPUT='_transformed')
  /OUTFILE PREPXML='/workingDirectory/car_sales_transformations.xml'.
TMS IMPORT
  /INFILE TRANSFORMATIONS='/workingDirectory/car_sales_transformations.xml'
    MODE=FORWARD (ROLES=UPDATE)
  /SAVE TRANSFORMED=YES.
EXECUTE.
```

- La commande ADP prépare le champ cible *sales* et les champs d'entrée de *resale* jusqu'à *mpg*.
- La sous-commande PREPDATETIME est spécifiée mais non utilisée car aucun des champs n'est un champ de date ou d'heure.
- La sous-commande ADJUSTLEVEL convertit les champs ordinaux avec plus de 10 valeurs en champs continus et les champs continus avec moins de 5 valeurs en champs ordinaux.
- La sous-commande OUTLIERHANDLING remplace les valeurs des entrées continues (pas celle de la cible) dont l'écart-type de la moyenne est supérieur à 3 par la valeur dont l'écart-type de la moyenne est égal à 3.
- La sous-commande REPLACEMISSING remplace les valeurs des entrées manquantes (pas celle de la cible).
- La sous-commande REORDERNOMINAL recode les valeurs des entrées nominales de la moins fréquente à la plus fréquente.
- La sous-commande RESCALE standardise les entrées continues à l'aide d'une transformation en score z de manière leur donner une moyenne de 0 et un écart-type de 1, et standardise la cible continue à l'aide d'une transformation de Box-Cox de manière lui donner une moyenne de 0 et un écart-type de 1.
- La sous-commande TRANSFORM désactive toutes les opérations spécifiées par défaut par cette sous-commande.
- La sous-commande CRITERIA spécifie les suffixes par défaut pour les transformations des entrées et de la cible.
- La sous-commande OUTFILE spécifie que les transformations doivent être enregistrées à l'emplacement */workingDirectory/car_sales_transformations.xml*, où */workingDirectory* est le répertoire de destination dans lequel vous souhaitez enregistrer le fichier *car_sales_transformations.xml*.
- La commande TMS IMPORT lit les transformations du fichier *car_sales_transformations.xml* et les applique à l'ensemble de données actif, mettant à jour les rôles des champs existants qui sont transformés.
- La commande EXECUTE génère le traitement des transformations. Lorsqu'elle est utilisée comme partie d'une syntaxe plus longue, vous pouvez retirez la commande EXECUTE afin de gagner du temps sur la durée d'exécution.

Création d'un modèle sur les données non préparées

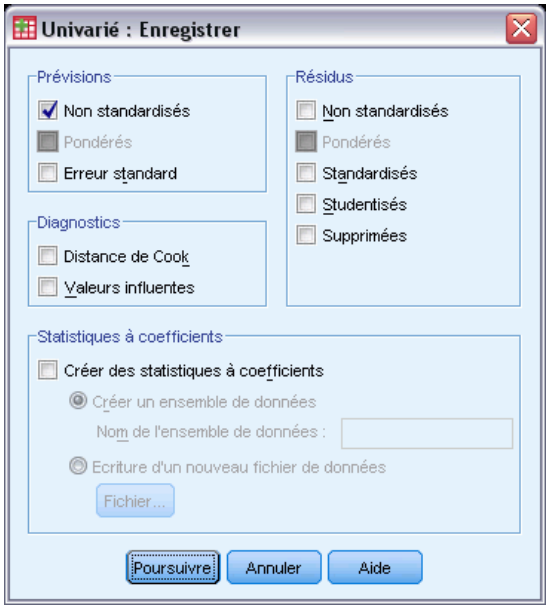
▶ Pour construire un modèle sur les données non préparées, sélectionnez à partir des menus :
Analyse > Modèle linéaire général > Univarié

Figure 8-14
Boîte de dialogue GLM Univarié



- Sélectionnez *Ventes en milliers [sales]* comme variable dépendante.
- Sélectionnez *Type de véhicule [type]* comme facteur fixe.
- Sélectionnez de *Valeur d'occasion après 4 ans [resale]* jusqu'à *Rendement énergétique [mpg]* comme covariables.
- Cliquez sur Enregistrer.

Figure 8-15
Boîte de dialogue Enregistrer



- Sélectionnez l'option Non standardisés dans le groupe Prévisions.
- Cliquez sur Poursuivre.
- Cliquez sur le bouton OK dans la boîte de dialogue GLM Univarié.

Ces sélections génèrent la syntaxe de commande suivante :

```
UNIANOVA sales BY type WITH resale price engine_s horsepow wheelbas width length
    curb_wgt fuel_cap mpg
  /METHOD=SSTYPE(3)
  /INTERCEPT=INCLUDE
  /SAVE=PRED
  /CRITERIA=ALPHA(0.05)
  /DESIGN=resale price engine_s horsepow wheelbas width length curb_wgt fuel_cap
    mpg type.
```

Figure 8-16
Effets inter-sujets pour le modèle basé sur les données non préparées

Variable dépendante:Sales in thousands

Source	Somme des carrés de type III	ddl	Moyenne des carrés	D	Sig.
Modèle corrigé	226123.658[a]	11	20556.696	5.050	.000
Ordonnée à l'origine	12227.688	1	12227.688	3.004	.086
resale	50.702	1	50.702	.012	.911
price	471.630	1	471.630	.116	.734
engine_s	19872.712	1	19872.712	4.882	.029
horsepow	9644.486	1	9644.486	2.369	.127
wheelbas	29824.272	1	29824.272	7.327	.008
width	263.465	1	263.465	.065	.800
length	1374.525	1	1374.525	.338	.562
curb_wgt	32762.692	1	32762.692	8.049	.005
fuel_cap	1124.237	1	1124.237	.276	.600
mpg	337.585	1	337.585	.083	.774
type	17668.779	1	17668.779	4.341	.040
Erreur	427402.183	105	4070.497		
Total	1062354.955	117			
Total corrigé	653525.841	116			

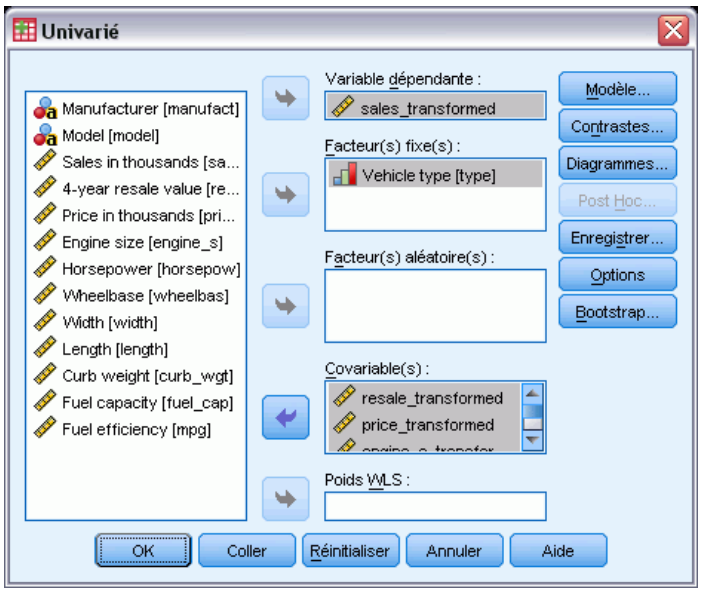
a. R deux = .346 (R deux ajusté = .277)

Le résultat GLM Univarié par défaut inclut les effets inter-sujets, qui est un tableau d'analyse de la variance. Chaque terme du modèle, ainsi que le modèle dans son ensemble, est testé pour connaître sa capacité à rendre compte des variations de la variable dépendante. Remarque : les étiquettes des variables ne sont pas affichées dans ce tableau.

Les variables prédites affichent des niveaux de signification différents. Celles dont la valeur du niveau de signification est inférieure à 0,05 sont considérées être utiles au modèle.

Création d'un modèle sur les données préparées

Figure 8-17
Boîte de dialogue GLM Univarié



- Pour construire le modèle sur les données préparées, ouvrez de nouveau la boîte de dialogue GLM Univarié.
- Désélectionnez *Ventes en milliers [sales]* et sélectionnez *sales_transformed* comme variable dépendante.
- Désélectionnez de *Valeur d'occasion après 4 ans [resale]* jusqu'à *Rendement énergétique [mpg]* et sélectionnez de *resale_transformed* à *mpg_transformed* comme covariables.
- Cliquez sur OK.

Ces sélections génèrent la syntaxe de commande suivante :

```
UNIANOVA sales_transformed BY type WITH resale_transformed price_transformed
    engine_s_transformed horsepow_transformed wheelbas_transformed width_transformed
    length_transformed curb_wgt_transformed fuel_cap_transformed mpg_transformed
  /METHOD=SSTYPE(3)
  /INTERCEPT=INCLUDE
  /SAVE=PRED
  /CRITERIA=ALPHA(0.05)
  /DESIGN=resale_transformed price_transformed engine_s_transformed horsepow_transformed
    wheelbas_transformed width_transformed length_transformed curb_wgt_transformed
    fuel_cap_transformed mpg_transformed type.
```

Figure 8-18
Effets inter-sujets pour le modèle basé sur les données préparées

Variable dépendante:sales_transformed

Source	Somme des carrés de type III	ddl	Moyenne des carrés	D	Sig.
Modèle corrigé	79.327[a]	11	7.212	13.638	.000
Ordonnée à l'origine	2.436	1	2.436	4.606	.034
resale_transformed	.954	1	.954	1.804	.181
price_transformed	9.271	1	9.271	17.533	.000
engine_s_transformed	2.885	1	2.885	5.456	.021
horsepow_transformed	.034	1	.034	.064	.801
wheelbas_transformed	1.213	1	1.213	2.293	.132
width_transformed	.037	1	.037	.071	.791
length_transformed	.265	1	.265	.501	.480
curb_wgt_transformed	.103	1	.103	.194	.660
fuel_cap_transformed	.132	1	.132	.249	.618
mpg_transformed	3.390	1	3.390	6.411	.012
type	4.007	1	4.007	7.579	.007
Erreur	76.673	145	.529		
Total	156.000	157			
Total corrigé	156.000	156			

a. R deux = .509 (R deux ajusté = .471)

On peut remarquer de petites différences entre les effets inter-sujets du modèle basé sur les données non préparées et ceux du modèle basé sur les données préparées. Tout d'abord, les degrés de liberté ont augmenté, du fait que les valeurs manquantes ont été remplacées par des valeurs imputées lors de la préparation automatique des données, et par conséquent les enregistrements incomplets supprimés du premier modèle sont disponibles pour le second. De manière notable, la signification de certaines variables prédites a changé. Alors que les deux modèles s'accordent sur le fait que la taille du moteur [*engine_s*] et le type de véhicule [*type*] sont utiles au modèle, l'empattement [*wheelbas*] et le poids à vide [*curb_wgt*] ne sont plus significatifs, et le prix du véhicule [*price_transformed*] et le rendement énergétique [*mpg_transformed*] sont devenus significatifs.

Pourquoi ce changement a-t-il eu lieu ? Les ventes ont une distribution asymétrique, donc il se peut que l'empattement et le poids à vide comportaient quelques enregistrements influents qui sont devenus non influents une fois que les ventes ont été transformées. Une autre possibilité est que les observations supplémentaires devenues disponibles du fait du remplacement des valeurs manquantes, aient changé la signification statistique de ces variables. Dans tous les cas, cela requiert une investigation plus poussée que nous ne réaliserons pas ici.

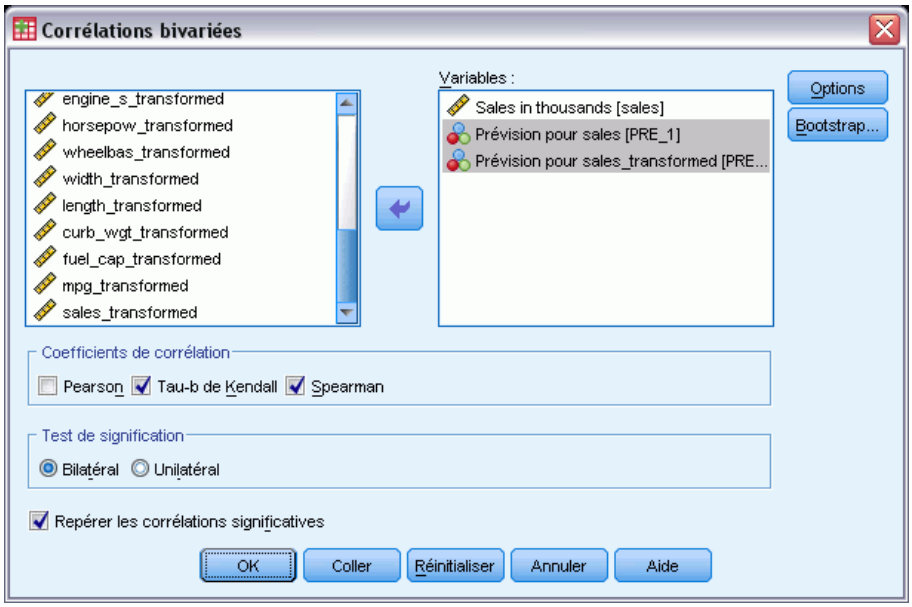
Remarque : le R-deux est plus élevé pour le modèle construit sur les données préparées, mais du fait que les ventes ont été transformées, cette mesure n'est peut-être pas la meilleure pour effectuer la comparaison de la performance de chaque modèle. A la place, vous pouvez calculer les corrélations non paramétriques entre les valeurs observées et les deux ensembles de valeurs prédites.

Comparaison des prévisions

- Pour obtenir les corrélations des prévisions à partir des deux modèles, sélectionnez à partir des menus :

  Analyse > Corrélation > Bivariée

Figure 8-19
Boîte de dialogue Corrélations bivariées



- Sélectionnez *Ventes en milliers [sales]*, *Prévision des ventes [PRE_1]*, et *Prévisions pour sales_transformed [PRE_2]* comme variables d'analyse.

- Désélectionnez Pearson et sélectionnez Tau-b de Kendall et Spearman dans le groupe Coefficients de corrélation.

  Remarque : les *Prévisions pour sales_transformed [PRE_2]* peuvent être utilisées pour calculer les corrélations non paramétriques sans devoir rétablir l'échelle d'origine, car ce rétablissement ne modifie pas l'ordre du classement des prévisions.

- Cliquez sur OK.

  Ces sélections génèrent la syntaxe de commande suivante :

```
NONPAR CORR
  /VARIABLES=sales PRE_1 PRE_2
  /PRINT=BOTH TWOTAIL NOSIG
  /MISSING=PAIRWISE.
```

Figure 8-20
Corrélations non paramétriques

			Sales in thousands	Prévision pour sales	Prévision pour sales_ transformed
Tau-B de Kendall	Sales in thousands	Coefficient de corrélation	1.000	.376**	.484**
		Sig. (bilatérale)	.	.000	.000
		N	157	117	157
	Prévision pour sales	Coefficient de corrélation	.376**	1.000	.655**
		Sig. (bilatérale)	.000	.	.000
		N	117	117	117
	Prévision pour sales_transformed	Coefficient de corrélation	.484**	.655**	1.000
		Sig. (bilatérale)	.000	.000	.
		N	157	117	157
Rho de Spearman	Sales in thousands	Coefficient de corrélation	1.000	.530**	.666**
		Sig. (bilatérale)	.	.000	.000
		N	157	117	157
	Prévision pour sales	Coefficient de corrélation	.530**	1.000	.831**
		Sig. (bilatérale)	.000	.	.000
		N	117	117	117
	Prévision pour sales_transformed	Coefficient de corrélation	.666**	.831**	1.000
		Sig. (bilatérale)	.000	.000	.
		N	157	117	157

**. La corrélation est significative au niveau 0,01 (bilatéral).

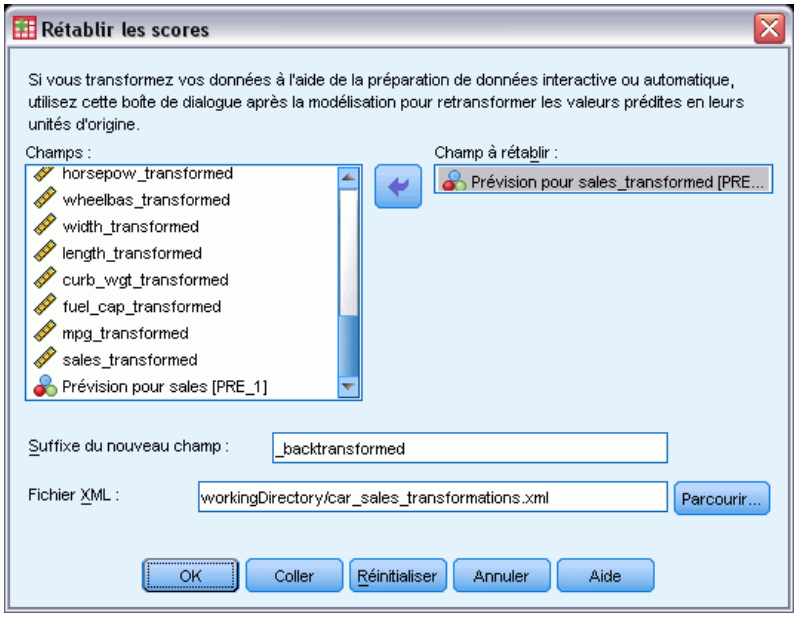
La première colonne montre que les prévisions du modèle construit à l'aide des données préparées sont plus fortement corrélées aux valeurs observées par les mesures du tau-b de Kendall et du rho de Spearman. Cela suggère que la préparation automatique des données a amélioré le modèle.

Rétablir les prévisions

- Les données préparées incluent une transformation des ventes, ainsi les prévisions provenant de ce modèle ne peuvent être utilisées directement en tant que scores. Pour transformer les prévisions selon l'échelle d'origine, à partir des menus, sélectionnez :

  Transformer > Préparer les données pour la modélisation > Rétablir les scores...

Figure 8-21
Boîte de dialogue Rétablir les scores



- Sélectionnez la *Prévision pour sales_transformed [PRE_2]* comme le champ à rétablir.
- Entrez _backtransformed comme suffixe du nouveau champ.
- Entrez workingDirectory\car_sales_transformations.xml comme emplacement du fichier XML contenant les transformations, où workingDirectory est le répertoire de destination du fichier XML.
- Cliquez sur OK.

Ces sélections génèrent la syntaxe de commande suivante :

```
TMS IMPORT
  /INFILE TRANSFORMATIONS='workingDirectory/car_sales_transformations.xml'
    MODE=BACK (PREDICTED=PRE_2 SUFFIX='_backtransformed').
EXECUTE.
```

- La commande `TMS IMPORT` lit les transformations dans *car_sales_transformations.xml* et applique la rétrotransformation à *PRE_2*.
- Le nouveau champ contenant les valeurs rétablies est appelé *PRE_2_backtransformed*.
- La commande `EXECUTE` entraîne le traitement des transformations. Lorsque ceci est utilisé en tant que partie d'un flux de syntaxe plus long, vous pouvez supprimer la commande `EXECUTE` afin de gagner du temps de traitement.

Récapitulatif

A l'aide de la préparation automatique des données, vous pouvez obtenir rapidement des transformations des données qui vous aident à améliorer votre modèle. Si le champ cible est transformé, vous pouvez enregistrer les transformations dans un fichier XML et utiliser la boîte

de dialogue Rétablir les scores pour convertir à l'échelle d'origine, les prévisions du champ cible transformé.

Chapitre
9

Identification des observations inhabituelles

La procédure de détection des anomalies vise à repérer les observations inhabituelles en se basant sur les écarts par rapport aux normes de leurs groupes. La procédure est destinée à détecter rapidement les observations inhabituelles afin de vérifier les données à l'étape d'analyse exploratoire des données, avant d'effectuer toute sorte d'analyse inférentielle de ces mêmes données. Cet algorithme sert à détecter des anomalies générales. Il est vrai que la définition d'une observation anormale ne s'applique pas à tous les secteurs. Par exemple, la définition d'une anomalie peut être clairement définie lorsqu'il s'agit de détecter des moyens de paiements inhabituels dans l'industrie pharmaceutique ou du blanchissement d'argent dans l'industrie bancaire.

Algorithme d'identification des observations inhabituelles

Cet algorithme se divise en trois étapes :

Modélisation. Cette procédure crée un modèle de classement qui explique les groupements naturels (ou classes) au sein d'un ensemble de données là où ils seraient normalement inapparents. Le classement se base sur un ensemble de variables d'entrées. Le modèle de classement obtenu ainsi que des statistiques en quantité suffisante pour calculer les normes du groupe sont stockés en vue d'un usage ultérieur.

Evaluation. Le modèle est appliqué à chaque observation pour identifier son groupe et des indices sont créés pour chaque observation afin d'en mesurer la singularité par rapport au reste de son groupe. Toutes les observations sont triées en fonction des valeurs des indices d'anomalies. Les anomalies sont classées en haut de la liste d'observations.

Détermination des raisons. Pour chaque observation anormale, les variables sont triées en fonction de leurs indices de déviation correspondants. Les variables en haut de la liste, leurs valeurs et les valeurs standard correspondantes sont considérées comme les raisons pour lesquelles une observation est identifiée comme une anomalie.

© Copyright SPSS Inc. 1989, 2010

Identification des observations inhabituelles dans une base de données médicale

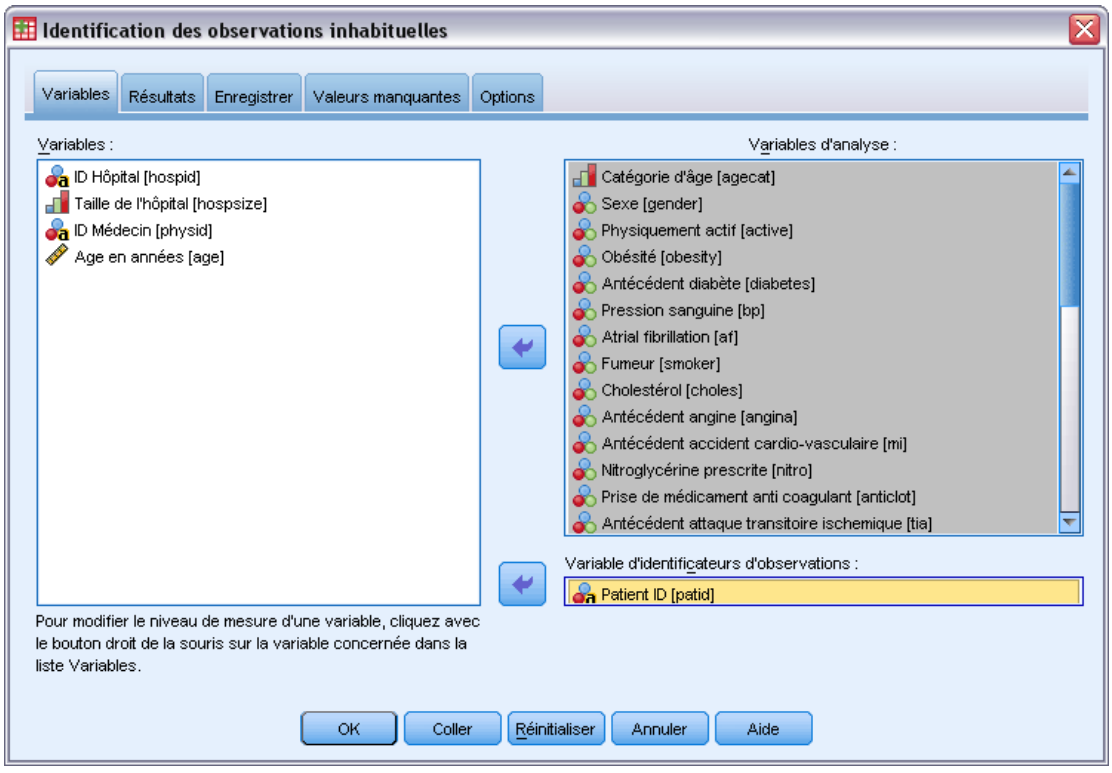
Un analyste de données employé pour construire des modèles capables de prédire les résultats obtenus suite au traitement d'attaques cardiaques cherche des données de qualité, car de tels modèles sont sensibles aux observations inhabituelles. Certaines de ces observations éloignées sont des observations tout à fait uniques et s'avèrent donc inexploitables en terme de prédiction, alors que d'autres sont dues à des erreurs de saisie de données dans lesquelles les valeurs sont techniquement "correctes" sans pouvoir toutefois être prises en compte par les procédures de validation de données.

Ces informations sont rassemblées dans le fichier *stroke_valid.sav*. Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A sur p. 138. Pour nettoyer le fichier de données, utilisez la procédure Identifier les observations inhabituelles. La syntaxe servant à reproduire ces analyses se trouve dans *detectanomaly_stroke.sps*.

Exécution de l'analyse

- Pour identifier les observations inhabituelles, à partir des menus, choisissez :
  Données > Identifier les observations inhabituelles...

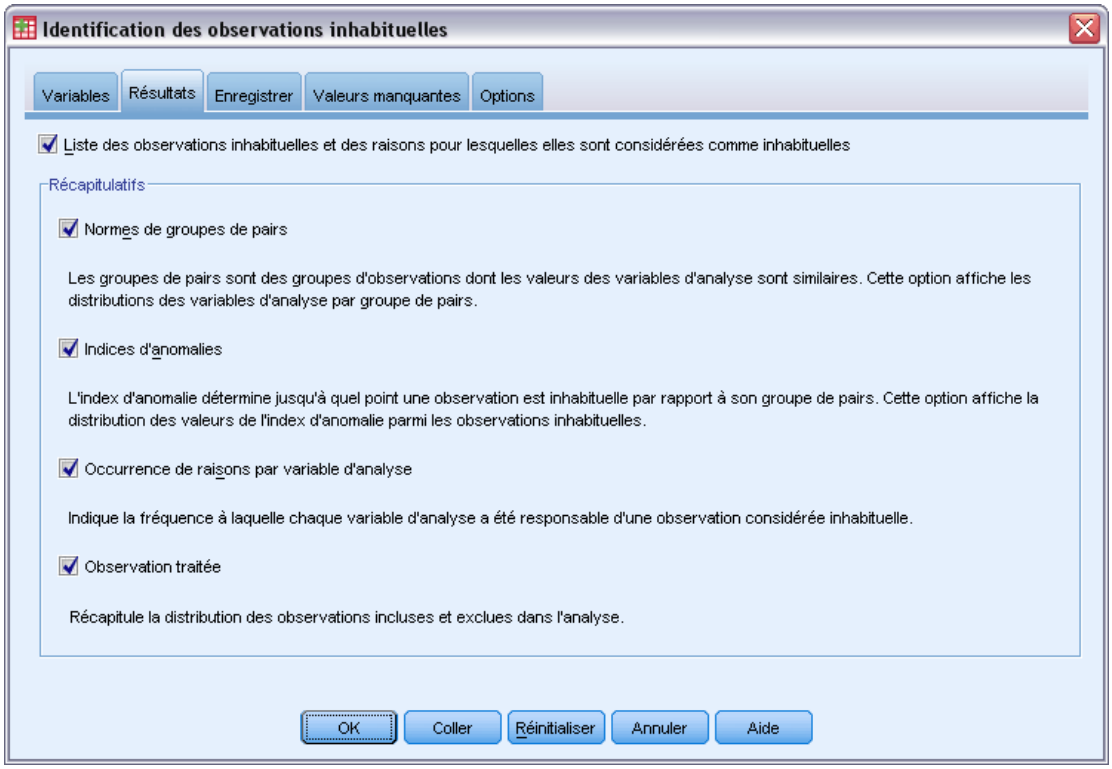
Figure 9-1
Boîte de dialogue Identifier les observations inhabituelles, onglet Variables



- Sélectionnez *Tranche d'âge* via *Attaque cardiaque entre 3 et 6 mois* comme variables d'analyse.

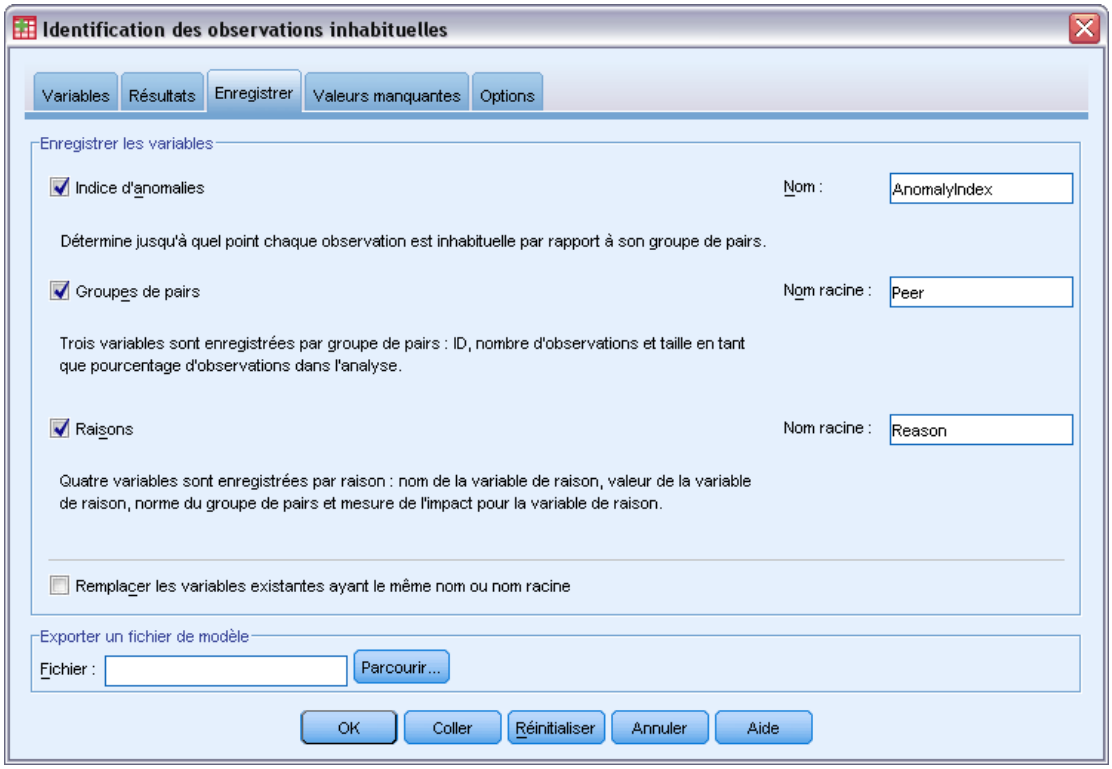
- Sélectionnez *ID du patient* comme variable d'identificateur de l'observation.
- Cliquez sur l'onglet Résultats.

Figure 9-2
Boîte de dialogue Identifier les observations inhabituelles, onglet Résultat



- Sélectionnez Normes de groupes de pairs, Indices d'anomalie, Occurrence de raisons par variable d'analyse et Observations traitées.
- Cliquez sur l'onglet Enregistrer.

Figure 9-3
Boîte de dialogue Identifier les observations inhabituelles, onglet Enregistrer

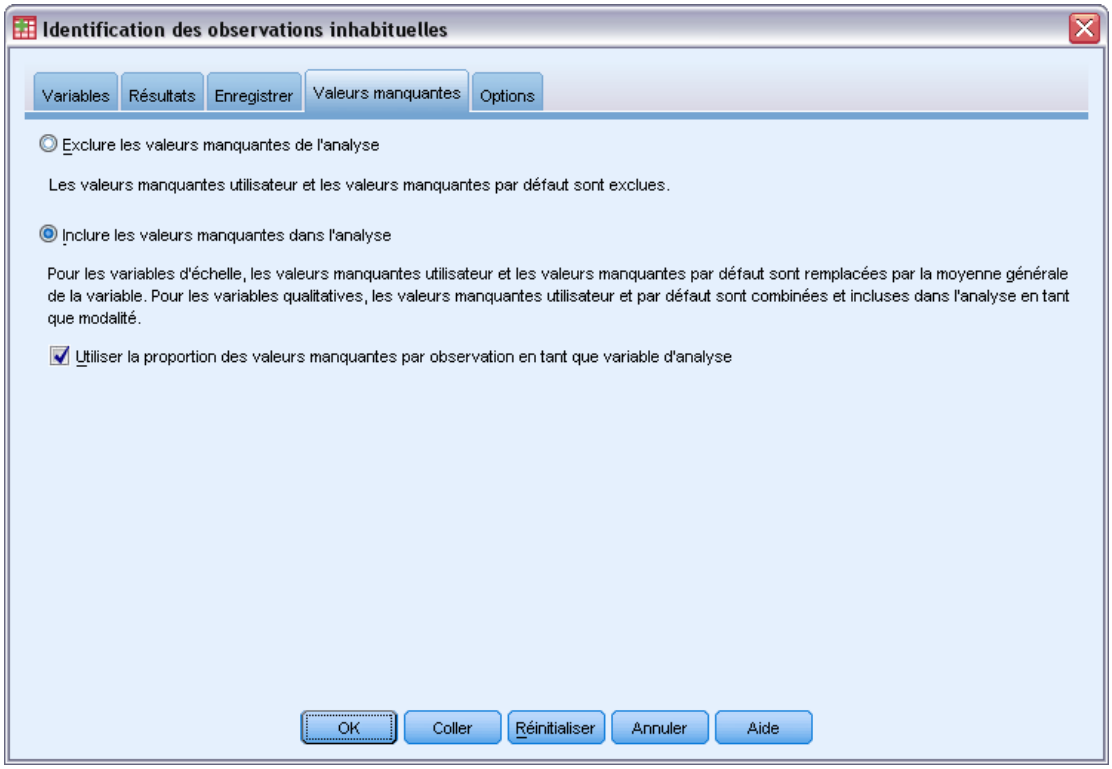


▶ Sélectionnez Index d'anomalie, Groupes de pairs et Raisons.

L'enregistrement de ces résultats vous permet de produire un diagramme de dispersion très utile qui récapitule les résultats.

▶ Cliquez sur l'onglet Valeurs manquantes.

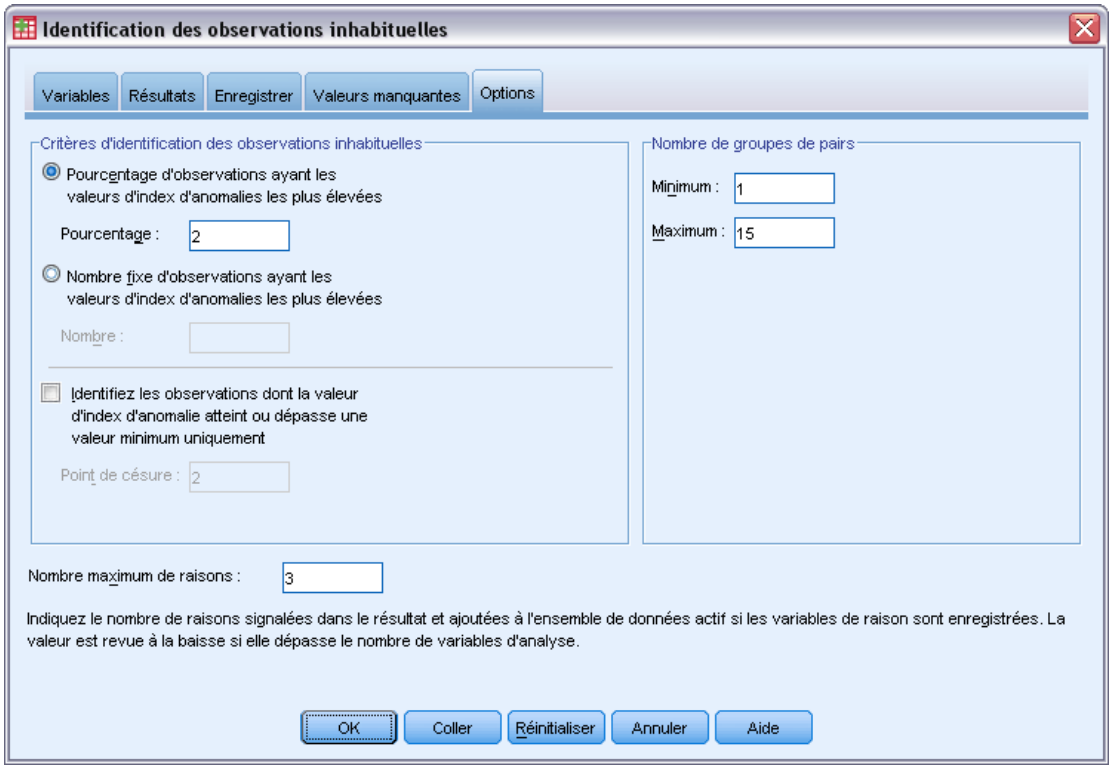
Figure 9-4
Boîte de dialogue Identifier les observations inhabituelles, onglet Valeurs manquantes



- Sélectionnez Inclure les valeurs manquantes dans l'analyse. Ce procédé est indispensable car il existe un grand nombre de valeurs manquantes utilisateur pour traiter des patients qui sont décédés avant ou pendant le traitement. Une variable supplémentaire mesurant la proportion des valeurs manquantes par observation est ajoutée à l'analyse en tant que variable d'échelle.

- Cliquez sur l'onglet Options.

Figure 9-5
Boîte de dialogue Identifier les observations inhabituelles, onglet Options



- Entrez 2 comme pourcentage d'observations à considérer comme anormales.
- Déselectionnez Identifiez les observations dont la valeur d'index d'anomalie atteint ou dépasse une valeur minimum uniquement.
- Entrez 3 comme nombre maximum de raisons.
- Cliquez sur OK.

Récapitulatif de traitement des observations

Figure 9-6
Récapitulatif du traitement des observations

		N :	% des valeurs combinées	% du total
ID d'homologue	1	710	67,7%	67,7%
	2	90	8,6%	8,6%
	3	248	23,7%	23,7%
Combinée		1048	100,0%	100,0%
Total		1048		100,0%

Chaque observation est classée dans un groupe de pairs d'observations similaires. Le récapitulatif de traitement des observations affiche le nombre de groupes de pairs créés ainsi que le nombre et le pourcentage des observations dans chaque groupe.

Liste d'index des observations présentant une anomalie

Figure 9-7
Liste d'index des observations présentant une anomalie

Observation	patid	Index d'anomalie
843	7840326167	2,837
510	0714726620	2,022
623	6553808330	2,014
501	6461046805	2,002
607	1077125669	1,897
884	2260043998	1,889
614	4030164769	1,869
241	1038840465	1,865
13	2191527525	1,826
172	4458028382	1,786
705	1336411777	1,778
651	4103977868	1,767
384	2247641363	1,767
839	0437454972	1,766
861	9746101913	1,757
19	7237535360	1,756
806	4391632997	1,756
871	6961938294	1,739
239	7315965190	1,738
887	6044244232	1,737
245	0816869249	1,736

L'index d'anomalie est une mesure qui reflète l'incohérence d'une observation en fonction de son groupe. Les 2 % d'observations présentant les valeurs les plus élevées dans l'index d'anomalie sont affichées avec leur numéro et ID. Vingt et une observations sont recensées, leur valeur s'échelonnant de 1,736 à 2,837. La différence de valeur relativement importante dans l'index d'anomalie entre la première et la seconde observation de la liste suggère que l'observation 843 comporte probablement une anomalie. Les autres observations devront être évaluées au cas par cas.

Liste d'ID des pairs d'observation présentant une anomalie

Figure 9-8
Liste d'ID des pairs d'observation présentant une anomalie

Observation	patid	ID d'homologue	Taille d'homologue	Taille d'homologue (%)
843	7840326167	3	248	23,7%
510	0714726620	3	248	23,7%
623	6553808330	3	248	23,7%
501	6461046805	3	248	23,7%
607	1077125669	3	248	23,7%
884	2260043998	3	248	23,7%
614	4030164769	3	248	23,7%
241	1038840465	3	248	23,7%
13	2191527525	3	248	23,7%
172	4458028382	3	248	23,7%
705	1336411777	1	710	67,7%
651	4103977868	1	710	67,7%
384	2247641363	3	248	23,7%
839	0437454972	3	248	23,7%
861	9746101913	3	248	23,7%
19	7237535360	1	710	67,7%
806	4391632997	1	710	67,7%
871	6961938294	1	710	67,7%
239	7315965190	3	248	23,7%
887	6044244232	1	710	67,7%
245	0816869249	3	248	23,7%

Les observations potentiellement anormales sont affichées avec les informations d'appartenance de leur groupe de pairs. 15 observations au total, dont les 10 premières, appartiennent au groupe de pairs 3 ; l'observation restante appartient au groupe de pairs 1.

Liste des raisons expliquant une anomalie

Figure 9-9
Liste des raisons expliquant une anomalie

Cause: 1

Observation	patid	Variable de raison	Impact de variable	Valeur de variable	Norme de variable
843	7840326167	cost	,411	200,51	19,83
510	0714726620	cost	,120	96,59	19,83
623	6553808330	cost	,175	114,01	19,83
501	6461046805	barthel1	,084	80	(Valeur Manquante)
607	1077125669	cost	,126	96,11	19,83
884	2260043998	cost	,138	99,73	19,83
614	4030164769	barthel1	,085	45	(Valeur Manquante)
241	1038840465	barthel1	,115	25	(Valeur Manquante)
13	2191527525	barthel1	,118	40	(Valeur Manquante)
172	4458028382	barthel1	,120	100	(Valeur Manquante)
705	1336411777	cost	,244	198,25	42,47
651	4103977868	barthel1	,064	30	95
384	2247641363	barthel1	,122	20	(Valeur Manquante)
839	0437454972	barthel1	,109	95	(Valeur Manquante)
861	9746101913	barthel1	,102	70	(Valeur Manquante)
19	7237535360	barthel3	,080	5	100
806	4391632997	barthel2	,088	10	100
871	6961938294	barthel1	,094	5	95
239	7315965190	barthel1	,092	45	(Valeur Manquante)
887	6044244232	barthel1	,066	40	95
245	0816869249	barthel1	,124	5	(Valeur Manquante)

Les variables de raison sont celles qui influent le plus sur la classification d'une observation dans la catégorie inhabituelle. La variable de raison principale s'affiche pour chaque observation présentant une anomalie, ainsi que l'impact, la valeur et la norme du groupe de pairs de l'observation. La norme du groupe de pairs *(Valeur manquante)* pour une variable qualitative indique que la pluralité des observations dans le groupe de pairs n'avait pas de valeur définie pour cette variable.

La statistique de l'impact d'une variable contribue de façon proportionnelle à l'écart de l'observation de la variable de raison par rapport à son groupe de pairs. Avec 38 variables dans l'analyse, y compris la variable de proportion manquante, l'impact attendu d'une variable serait 1/38 = 0,026. L'impact du *coût* de la variable dans l'observation 843 est 0,411, ce qui est relativement important. La valeur du *coût* pour l'observation 843 est 200,51, comparée à la moyenne de 19,83 pour les observations du groupe de pairs 3.

Les sélections de la boîte de dialogue requéraient des résultats pour les trois principales raisons.

- Pour visualiser les résultats des autres raisons, double-cliquez sur le tableau pour l'activer.
- Déplacez *Raison* de la dimension de strate à la dimension de ligne.

Figure 9-10
Liste des raisons expliquant une anomalie (8 premières observations)

Observation	Cause	patid	Variable de raison	Impact de variable	Valeur de variable	Norme de variable
843	1	7840326167	cost	,411	200,51	19,83
	2	7840326167	barthel1	,076	65	(Valeur Manquante)
	3	7840326167	rankin1	,044	2	(Valeur Manquante)
510	1	0714726620	cost	,120	96,59	19,83
	2	0714726620	barthel1	,083	80	(Valeur Manquante)
	3	0714726620	rehab	,068	3	(Valeur Manquante)
623	1	6553808330	cost	,175	114,01	19,83
	2	6553808330	surgery	,089	2	(Valeur Manquante)
	3	6553808330	barthel1	,089	70	(Valeur Manquante)
501	1	6461046805	barthel1	,084	80	(Valeur Manquante)
	2	6461046805	rehab	,068	3	(Valeur Manquante)
	3	6461046805	rankin1	,063	1	(Valeur Manquante)
607	1	1077125669	cost	,126	96,11	19,83
	2	1077125669	barthel1	,094	85	(Valeur Manquante)
	3	1077125669	rehab	,072	3	(Valeur Manquante)
884	1	2260043998	cost	,138	99,73	19,83
	2	2260043998	barthel1	,114	65	(Valeur Manquante)
	3	2260043998	rehab	,072	3	(Valeur Manquante)
614	1	4030164769	barthel1	,085	45	(Valeur Manquante)
	2	4030164769	rankin1	,085	3	(Valeur Manquante)
	3	4030164769	recbart1	,062	2	(Valeur Manquante)

Cette configuration simplifie la comparaison entre les contributions relatives des trois premières raisons pour chaque observation. Comme supposé, l'observation 843 est considérée comme anormale en raison de la valeur singulièrement importante de son *coût*. A titre comparatif, pas une seule raison ne contribue pour plus de 0,10 à la singularité de l'observation 501.

Normes de variables d'échelle

Figure 9-11
Normes de variables d'échelle

		ID d'homologue			
		1	2	2	Combinée
Durée de séjour pour réhabilitation	Moyenne	16,55	16,39	15,91	16,39
	Ecart type	12,596	,000	6,834	10,887
Coût total pour traitement et réhabilitation en milliers ($)	Moyenne	42,4673	3,5089	19,8273	33,7641
	Ecart type	26,45401	,50997	20,17309	27,31266
Proportion manquante	Moyenne	,006	,541	,354	,134
	Ecart type	,021	2,9E-016	,083	,197

Les normes de variables d'échelle indiquent la moyenne et l'écart-type de chaque variable pour chaque groupe de pairs et pour la totalité des groupes. La comparaison de valeurs fournit quelques indications sur les variables contribuant à la formation de groupe de pairs.

Par exemple, la moyenne de la *Durée du séjour nécessaire à la rééducation* est relativement constante dans les trois groupes de pairs, ce qui signifie que cette variable ne contribue pas à leur formation. En revanche, le *Coût total du traitement et de la rééducation en milliers* et la *Proportion manquante* donnent une idée de l'appartenance à un groupe de pairs. Le groupe de pairs 1 présente le coût moyen le plus élevé et le moins de valeurs manquantes. Le groupe de pairs 2 présente des coûts très bas et un grand nombre de valeurs manquantes. Le groupe de pairs 3 présente un nombre de valeurs manquantes et des coûts moyens.

Cette organisation suggère que le groupe de pairs 2 est composé de patients déjà décédés à leur arrivée, engendrant ainsi des coûts très faibles et une absence de variables de traitement et de rééducation. Le groupe de pairs 3 se compose d'un grand nombre de patients décédés en cours de traitement, engendrant ainsi des coûts de traitement mais aucun coût de rééducation, d'où l'absence de cette variable. Selon toute probabilité, le groupe de pairs 1 se compose presque exclusivement de patients qui ont survécu au traitement et à la rééducation, engendrant ainsi les coûts les plus élevés.

Normes de variables qualitatives

Figure 9-12
Normes de variables qualitatives (10 premières variables)

		ID d'homologue			
		1	2	3	Combinée
Catégorie d'âge	Modalité la plus utilisée	2	2	2	2
	Effectif	209	215	215	424
	Pourcentage :	38,8%	33,4%	33,4%	35,8%
Sexe	Modalité la plus utilisée	0	0	1	0
	Effectif	275	328	328	592
	Pourcentage :	51,0%	50,9%	50,9%	50,0%
Physiquement actif	Modalité la plus utilisée	1	0	0	0
	Effectif	285	342	342	596
	Pourcentage :	52,9%	53,1%	53,1%	50,4%
Obésité	Modalité la plus utilisée	0	0	0	0
	Effectif	422	471	471	893
	Pourcentage :	78,3%	73,1%	73,1%	75,5%
Antécédent diabète	Modalité la plus utilisée	0	0	0	0
	Effectif	512	549	549	1061
	Pourcentage :	95,0%	85,2%	85,2%	89,7%
Pression sanguine	Modalité la plus utilisée	1	1	1	1
	Effectif	350	362	362	712
	Pourcentage :	64,9%	56,2%	56,2%	60,2%
Atrial fibrillation	Modalité la plus utilisée	0	0	0	0
	Effectif	488	571	571	1059
	Pourcentage :	90,5%	88,7%	88,7%	89,5%
Fumeur	Modalité la plus utilisée	0	0	0	0
	Effectif	457	454	454	911
	Pourcentage :	81,4%	76,7%	72,2%	78,8%
Cholestérol	Modalité la plus utilisée	0	0	0	0
	Effectif	302	367	367	669
	Pourcentage :	56,0%	57,0%	57,0%	56,6%

Les normes de variables qualitatives servent essentiellement le même objectif que les normes d'échelle, mais elles indiquent la catégorie modale (la plus courante) ainsi que le nombre et le pourcentage d'observations dans les groupes de pairs appartenant à cette catégorie. La comparaison des valeurs peut s'avérer un peu plus compliquée ; par exemple, au premier coup d'œil, on peut penser que le *Sexe* contribue plus à la formation de classes que la variable *Fumeur*, car la catégorie modale de *Fumeur* est identique pour les trois groupes de pairs, tandis que la catégorie modale de *Sexe* est différente pour le groupe de pairs 3. Toutefois, la variable *Sexe* ne comportant que deux valeurs, vous pouvez en conclure que 49,2 % des observations du groupe 3 présentent une valeur de 0, ce qui est très similaire aux pourcentages des autres groupes. En revanche, les pourcentages pour la variable *Fumeur* s'échelonnent de 72,2 % à 81,4 %.

Figure 9-13
Normes de variables qualitatives (variables sélectionnées)

		ID d'homologue			
		1	2	3	Combinée
Durée de séjour à l'hôpital	Modalité la plus utilisée	3	4	3	2
	Effectif	202	164	339	424
	Pourcentage :	37,5%	25,5%	28,7%	35,8%
Décès à l'arrivée	Modalité la plus utilisée	0	(Valeur Manquante)	0	0
	Effectif	511	573	1084	592
	Pourcentage :	94,8%	89,0%	91,6%	50,0%
Note initiale	Modalité la plus utilisée	0	(Valeur Manquante)	5	0
	Effectif	171	230	254	596
	Pourcentage :	31,7%	35,7%	21,5%	50,4%
Résultat scan CAT	Modalité la plus utilisée	0	(Valeur Manquante)	0	0
	Effectif	492	463	955	893
	Pourcentage :	91,3%	71,9%	80,7%	75,5%
Médicament disolvant-caillot	Modalité la plus utilisée	2	(Valeur Manquante)	2	0
	Effectif	295	368	453	1061
	Pourcentage :	54,7%	57,1%	38,3%	89,7%
Décès à l'hôpital	Modalité la plus utilisée	0	(Valeur Manquante)	0	1
	Effectif	510	451	961	712
	Pourcentage :	94,6%	70,0%	69,0%	60,2%
Résultat du traitement	Modalité la plus utilisée	1	(Valeur Manquante)	1	0
	Effectif	490	251	741	1059
	Pourcentage :	90,9%	39,0%	62,6%	89,5%
Chirurgie préventive Post-événement	Modalité la plus utilisée	1	(Valeur Manquante)	(Valeur Manquante)	0
	Effectif	243	344	579	911
	Pourcentage :	45,1%	53,4%	48,9%	77,0%
Réhabilitation Post-événement	Modalité la plus utilisée	0	(Valeur Manquante)	(Valeur Manquante)	0
	Effectif	220	193	410	669
	Pourcentage :	40,8%	30,0%	34,7%	56,6%

Les suppositions basées sur les normes de variables d'échelle sont confirmées plus bas dans le tableau des normes qualitatives. Le groupe de pairs 2 est entièrement composé des patients décédés à l'arrivée, d'où l'absence des variables de traitement et de rééducation. La plupart des patients dans le groupe de pairs 3 (69 %) sont décédés en cours de traitement, ce qui explique pourquoi la catégorie modale pour les variables de rééducation indique *(Valeur manquante)*.

Récapitulatif de l'index d'anomalie

Figure 9-14
Récapitulatif de l'index d'anomalie

	N dans la liste d'anomalies	Minimum	Maximum	Moyenne	Ecart type
Index d'anomalie	21	1,736	2,837	1,872	,240

N dans la liste d'anomalies est déterminé par la spécification : pourcentage d'anomalies : 2%")

Le tableau fournit des statistiques récapitulatives pour les valeurs d'index d'anomalie dans les observations de la liste d'anomalies.

Récapitulatif des raisons

Figure 9-15
Récapitulatif des raisons (variables de traitement et de rééducation)

	Occurrence utilisée comme raison		Statistiques d'impact de variable			
	Effectif	Pourcentage :	Minimum	Maximum	Moyenne	Ecart type
Catégorie d'âge	0	,0%	.	.	.	.
Sexe	0	,0%	.	.	.	.
Physiquement actif	0	,0%	.	.	.	.
Obésité	0	,0%	.	.	.	.
Antécédent diabète	1	4,2%	,069	,069	,069	.
Pression sanguine	0	,0%	.	.	.	.
Atrial fibrillation	0	,0%	.	.	.	.
Fumeur	0	,0%	.	.	.	.
Cholestérol	0	,0%	.	.	.	.
Antécédent angine	0	,0%	.	.	.	.
Antécédent accident cardio-vasculaire	0	,0%	.	.	.	.
Nitroglycérine prescrite	0	,0%	.	.	.	.
Prise de médicament anti coagulant	0	,0%	.	.	.	.
Antécédent attaque transitoire ischemique	0	,0%	.	.	.	.
Durée de séjour à l'hôpital	0	,0%	.	.	.	.
Décès à l'arrivée	0	,0%	.	.	.	.
Indexe Barthel à 1 mois	13	61,9%	,064	,124	,110	,021
Résultat scan CAT	0	,0%	.	.	.	.
Médicament disolvant-caillot	0	,0%	.	.	.	.
Décès à l'hôpital	0	,0%	.	.	.	.
Résultat du traitement	1	4,2%	,110	,110	,110	.
Chirurgie préventive Post-événement	0	,0%	.	.	.	.
Réhabilitation Post-événement	0	,0%	.	.	.	.
Proportion manquante	0	,0%	.	.	.	.
Global	14	100,0%	,069	,136	,094	,021

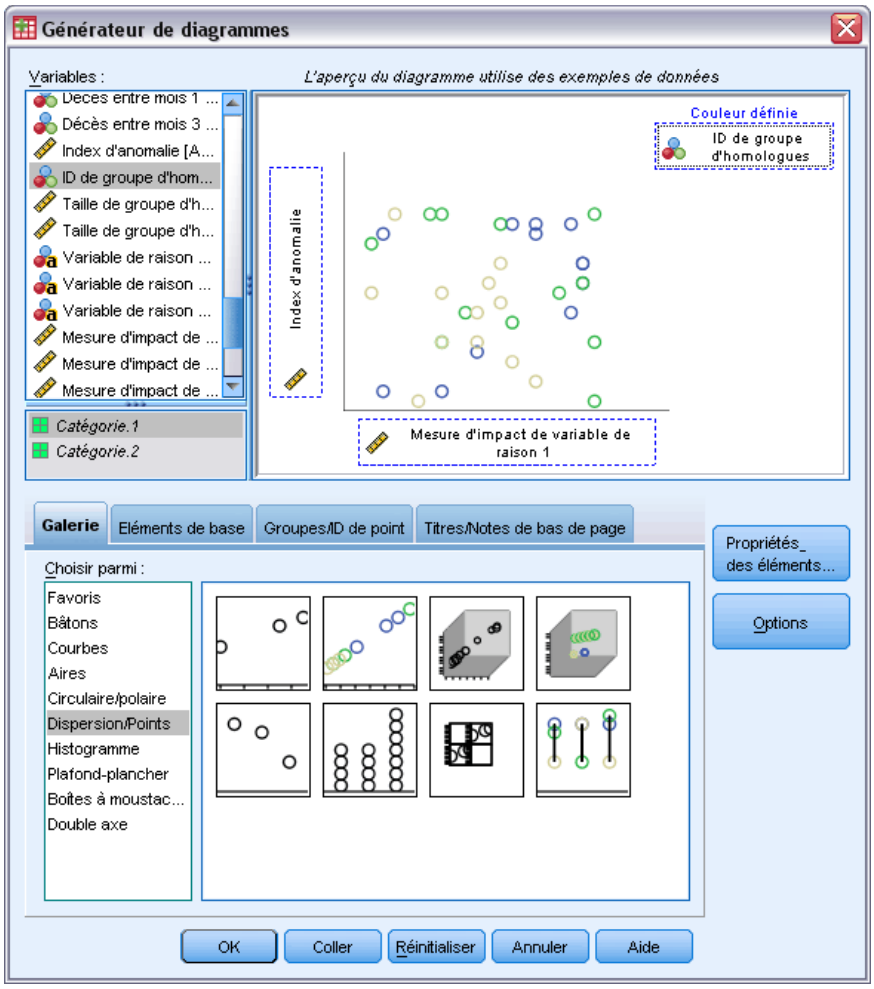
Pour chaque variable dans l'analyse, le tableau récapitule le rôle de la variable en tant que raison principale. La plupart des variables, telles que *Décédé à l'arrivée* ou *Rééducation suite à l'événement*, ne constituent pas la raison principale pour laquelle les observations se trouvent dans la liste d'anomalies. *L'Index de Barthel au premier mois* constitue la raison la plus fréquente, suivi de *Coût total du traitement et de la rééducation en milliers*. Les statistiques d'impact des variables sont récapitulées, avec l'impact minimum, maximum et moyen observé pour chaque variable, ainsi que l'écart-type pour les variables qui constituaient la raison dans plus d'une observation.

Diagramme de dispersion de l'index d'anomalie en fonction de l'impact de variables

Les tableaux contiennent de nombreuses informations utiles, bien qu'il soit parfois difficile de comprendre les relations. A l'aide des variables enregistrées, vous pouvez créer un diagramme qui simplifie ce processus.

▶ Pour produire ce diagramme de dispersion, dans les menus, choisissez :
Graphes > Générateur de diagrammes...

Figure 9-16
Boîte de dialogue Générateur de diagrammes



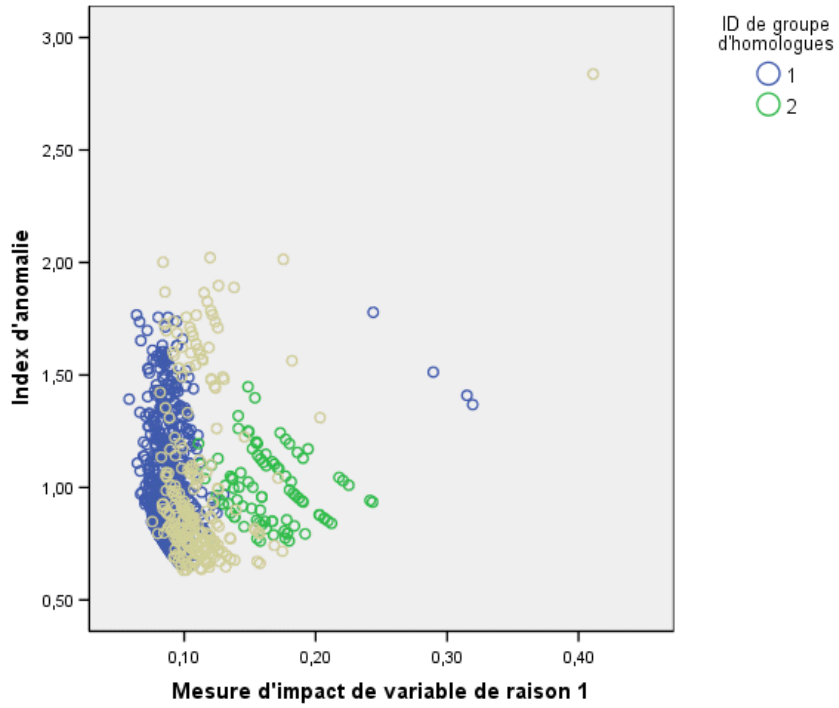
▶ Sélectionnez la galerie Dispersion/Points et faites glisser l'icône Diagramme de dispersion regroupé sur le canevas.

▶ Sélectionnez *Index d'anomalie* pour la variable *y* et *Mesure de l'impact de la variable de raison 1* pour la variable *x*.

- Sélectionnez *ID du groupe de pairs* en tant que variable Définir les couleurs par.
- Cliquez sur OK.

Ces sélections produisent le diagramme de dispersion.

Figure 9-17
Diagramme de dispersion de l'index d'anomalie en fonction de la mesure de l'impact de la première variable de raison.



L'étude du diagramme aboutit à certaines observations :

- L'observation à droite de la partie supérieure appartient au groupe de pairs 3 ; elle est à la fois l'observation la plus anormale et celle sur laquelle une seule variable a le plus d'effet.
- Si l'on suit l'axe *y*, on s'aperçoit que trois observations appartenant au groupe de pairs 3 présentent des valeurs d'index d'anomalie juste au-dessus de 2,00. L'anomalie potentielle de ces observations devrait être étudiée de plus près.
- Si l'on suit à présent l'axe *X*, on s'aperçoit que quatre observations appartiennent au groupe de pairs 1, avec des mesures d'impact de variables se situant approximativement entre 0,23 et 0,33. Ces observations devraient être étudiées plus rigoureusement car ces valeurs les séparent du peloton de points du diagramme.
- Le groupe de points 2 semble relativement homogène dans le sens où ses valeurs d'index d'anomalie et d'impact de variables ne s'éloignent pas trop du peloton central.

Récapitulatif

La procédure Identification des observations inhabituelles vous a permis d'isoler certaines observations qui nécessitent un examen plus approfondi. Ces observations seraient passées inaperçues avec d'autres procédures de validation, car ce sont les relations entre les variables (et pas seulement les valeurs des variables elles-mêmes) qui déterminent les observations anormales.

Il est quelque peu décevant que les groupes de pairs soient en grande partie formés en fonction de deux variables : *Décédé à l'arrivée* et *Décédé à l'hôpital*. Lors d'une analyse plus approfondie, vous pourriez étudier les effets causés par la création forcée d'un grand nombre de groupes de pairs ou vous pourriez construire une analyse n'incluant que les patients qui ont survécu à leur traitement.

Procédures apparentées

La procédure Identification des observations inhabituelles est un outil utile pour la détection d'observations anormales dans votre fichier de données.

- La procédure Valider les données identifie les observations, variables et valeurs de données suspectes ou invalides dans un ensemble de données actif.

Chapitre
10

Recodage supervisé optimal

La procédure Recodage supervisé optimal discrétise une ou plusieurs variables d'échelle (qualifiées de variables d'**entrée de regroupement**) en distribuant les valeurs de chaque variable dans des casiers. La formation de casiers est optimale par rapport à une variable guide qualitative qui « supervise » le regroupement par casiers. Les casiers peuvent être utilisés à la place des valeurs de données d'origine en vue d'analyses approfondies dans les procédures dans lesquelles des variables catégorielles sont nécessaires ou préférables.

Algorithme Recodage supervisé optimal

Les étapes de base de l'algorithme Recodage supervisé optimal présentent les caractéristiques suivantes :

Prétraitement (facultatif). La variable d'entrée de regroupement est divisée en *n* casiers (*n* étant spécifié par vous) et chaque casier contient le même nombre d'observations ou un nombre d'observations aussi proche que possible.

Identification des points de césure potentiels. Chaque valeur distincte de l'entrée de regroupement n'appartenant pas à la même catégorie de variable guide que la plus grande valeur distincte suivante de la variable d'entrée de regroupement constitue un point de césure potentiel.

Sélection des points de césure. Le point de césure potentiel qui génère le gain d'informations le plus grand est évalué par le critère d'acceptation MDLP. Répétez la procédure jusqu'à ce qu'aucun point de césure potentiel ne soit accepté. Les points de césure acceptés définissent les extrema des casiers.

Utilisation du recodage supervisé optimal pour discrétiser les données relatives aux demandeurs de prêt

Dans le cadre des efforts mis en œuvre par une banque pour réduire le taux de défaut de paiement, un responsable des prêts a recueilli des informations financières et démographiques sur les clients passés et actuels, en vue de créer un modèle permettant de prévoir la probabilité de défaut de paiement. Plusieurs variables indépendantes potentielles sont des variables d'échelle, mais le responsable des prêts souhaite pouvoir prendre en compte les modèles fonctionnant au mieux avec les variables indépendantes catégorielles.

Des informations sur 5 000 clients passés sont recueillies dans le fichier *bankloan_binning.sav*. Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A sur p. 138. Utilisez la procédure Recodage supervisé optimal afin de générer des règles de

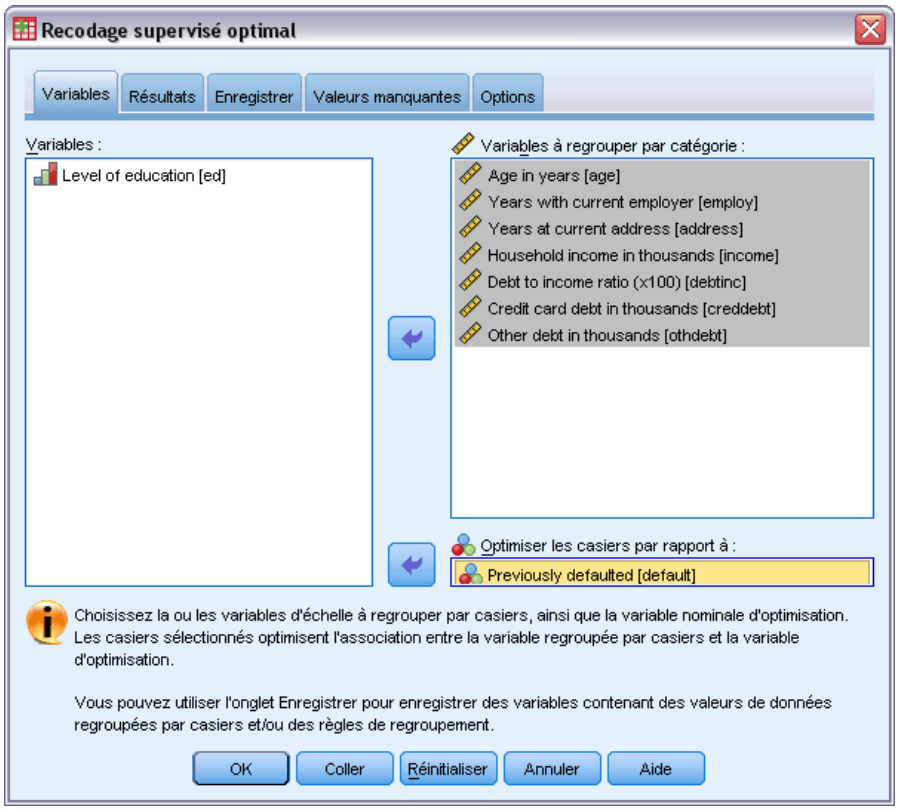
© Copyright SPSS Inc. 1989, 2010

regroupement pour les variables indépendantes d'échelle, puis utilisez ces règles pour traiter le fichier *bankloan.sav*. Vous pouvez ensuite utiliser le fichier de données pris en compte dans l'analyse pour créer un modèle de prévision.

Exécution de l'analyse

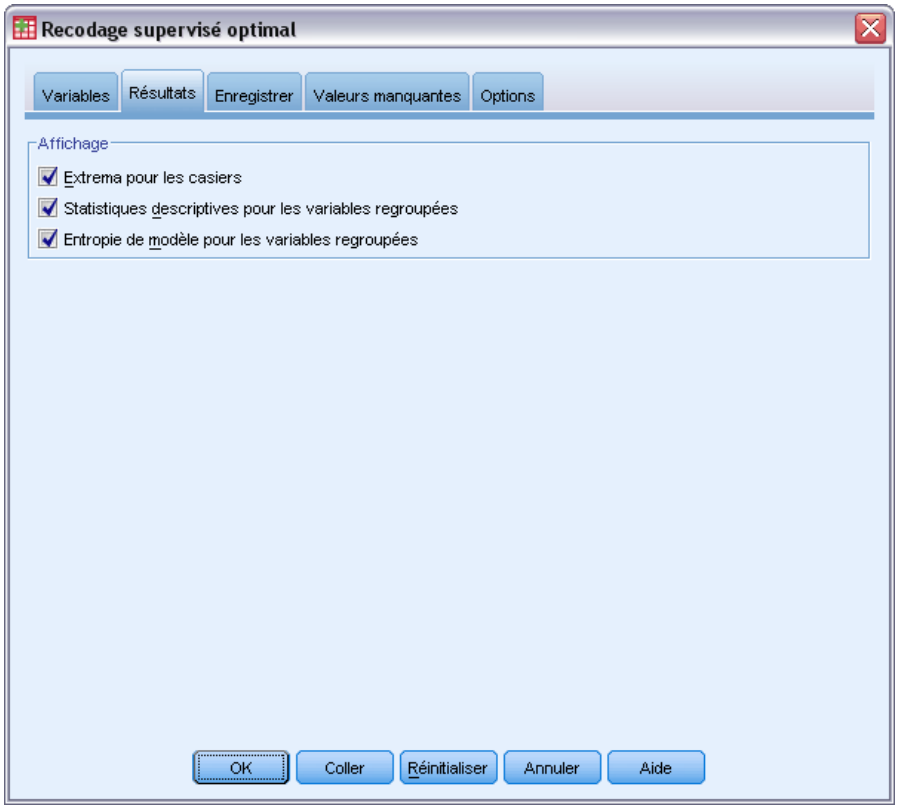
- Pour exécuter une analyse de recodage supervisé optimal, à partir des menus, sélectionnez : Transformer > Recodage supervisé optimal…

Figure 10-1
Boîte de dialogue Regroupement optimal, onglet Variables



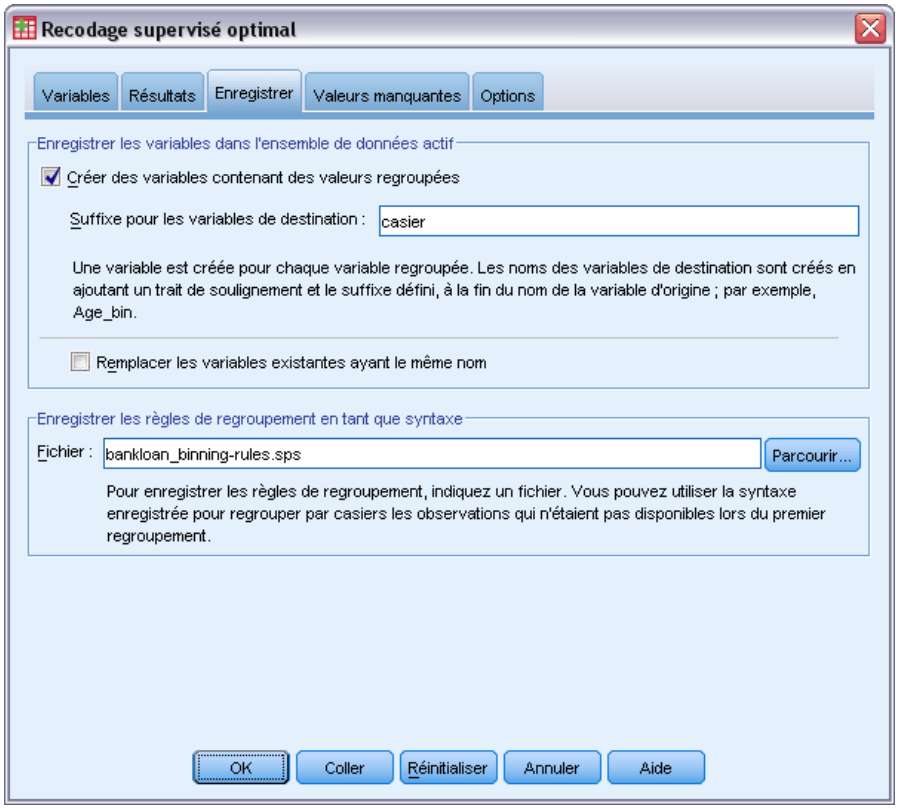
- Sélectionnez les options de *Age en années* et *Nb d'années avec l'employeur actuel* à *Autres dettes en milliers* comme variables à regrouper.
- Sélectionnez *Manquement précédent* comme variable guide.
- Cliquez sur l'onglet Résultats.

Figure 10-2
Boîte de dialogue Regroupement optimal, onglet Résultat



- Sélectionnez les options Statistiques descriptives et Entropie du modèle pour les variables regroupées.
- Cliquez sur l'onglet Enregistrer.

Figure 10-3
Boîte de dialogue Regroupement optimal, onglet Enregistrer



- Sélectionnez Créer des variables contenant des valeurs regroupées.
- Entrez le chemin et le nom du fichier de syntaxe qui doit contenir les règles de regroupement générées. Dans cet exemple, nous utilisons */bankloan_binning-rules.sps*.
- Cliquez sur OK.

Ces sélections génèrent la syntaxe de commande suivante :

```
* Recodage supervisé optimal.
OPTIMAL BINNING
  /VARIABLES GUIDE=default BIN=age employ address income debtinc creddebt
  othdebt SAVE=YES (INTO=age_bin employ_bin address_bin income_bin debtinc_bin
  creddebt_bin othdebt_bin)
  /CRITERIA METHOD=MDLP
   PREPROCESS=EQUALFREQ (BINS=1000)
   FORCEMERGE=0
   LOWERLIMIT=INCLUSIVE
   LOWEREND=UNBOUNDED
   UPPEREND=UNBOUNDED
  /MISSING SCOPE=PAIRWISE
  /OUTFILE RULES='/bankloan_binning-rules.sps'
  /PRINT ENDPOINTS DESCRIPTIVES ENTROPY.
```

- La procédure discrétise les variables d'entrée de regroupement *âge* , *emploi*, *adresse*, *revenu*, *dettrev*, *dettcred* et *autrdettes* à l'aide de la variable guide de regroupement MDLP *default*.

- Les valeurs discrétisées de ces variables sont stockées dans les nouvelles variables *âge_bin*, *emploi_bin*, *adresse_bin*, *revenu_bin*, *dettrev_bin*, *dettcred_bin* et *autrdettes_bin*.
- Si une variable d'entrée de regroupement présente plus de 1 000 valeurs distinctes, la méthode d'effectifs égaux réduit ce nombre à 1 000 avant d'exécuter le regroupement MDLP.
- La syntaxe de la commande représentant les règles de regroupement est enregistrée dans le fichier *c:\bankloan_binning-rules.sps*.
- Les extrema des casiers, les statistiques descriptives et les valeurs d'entropie de modèle sont demandés pour les variables d'entrée de regroupement.
- Les valeurs par défaut des autres critères de regroupement sont définies.

Statistiques descriptives

Figure 10-4
Statistiques descriptives

	N	Minimum	Maximum	Nombre de valeurs distinctes	Nombre de casiers
Age en années	5000	20	58	39	2
Nb d'années avec l'employeur actuel	5000	0	38	39	4
Nb d'années à l'adresse actuelle	5000	0	37	38	3
Revenu du foyer en milliers	5000	12.10	2461.70	1100	2
Ratio Débit/Crédit (x100)	5000	,08	44,62	2060	5
Débit carte de crédit en milliers	5000	,01	139.58	5000	4
Autre dette en milliers	5000	,01	416.52	4999	2

Le tableau de statistiques descriptives fournit des informations récapitulatives sur les variables d'entrée de regroupement. Les quatre premières colonnes concernent les valeurs pré-regroupées.

- N est le nombre d'observations utilisées dans l'analyse. Lorsque la suppression des observations incomplètes des valeurs manquantes est utilisée, cette valeur doit être constante d'une variable à l'autre. Lorsque le traitement des valeurs manquantes appariées est utilisé, cette valeur peut ne pas être constante. Ce fichier de données ne comportant aucune valeur manquante, la valeur est simplement le nombre d'observations.
- Les colonnes Minimum et Maximum affichent les valeurs minimale et maximale (de pré-regroupement) contenues dans le fichier de données pour chaque variable d'entrée de regroupement. Non seulement ces colonnes donnent un sens à l'intervalle de valeurs observé pour chaque variable, mais elles peuvent également s'avérer utiles pour repérer les valeurs situées en dehors de l'intervalle théorique.
- L'option Nombre de valeurs distinctes indique les variables ayant été prétraitées à l'aide de l'algorithme d'effectifs égaux. Par défaut, les variables comportant plus de 1 000 valeurs distinctes (*Revenu du ménage en milliers* à *Autres dettes en milliers*) sont préregroupées dans

1 000 casiers distincts. Ces casiers prétraités sont alors regroupés par rapport à la variable guide via MDLP. Vous pouvez contrôler la fonction de prétraitement dans l'onglet Options.

- Le nombre de casiers est le nombre final de casiers générés par la procédure, largement inférieur au nombre de valeurs distinctes.

Entropie de modèle

Figure 10-5
Entropie de modèle

	Entropie du modèle
Age en années	.788
Nb d'années avec l'employeur actuel	.754
Nb d'années à l'adresse actuelle	.781
Revenu du foyer en milliers	.803
Ratio Débit/Crédit (x100)	.711
Débit carte de crédit en milliers	.776
Autre dette en milliers	.801

Une plus petite entropie de modèle indique une plus grande exactitude des prévisions de la variable regroupée par casiers sur la variable guide Antécédent découvert.

L'entropie de modèle donne une idée de l'utilité possible de chaque variable dans un modèle de prévision de probabilité de défaut.

- La meilleure variable indépendante possible est une variable qui, pour chaque casier généré, contient des observations comportant la même valeur que la variable guide. Par conséquent, la variable guide peut être parfaitement prévue. Une telle variable indépendante présente une entropie de modèle non définie. Ceci ne se produit généralement pas dans des situations réelles et peut révéler des problèmes de qualité des données.
- La pire variable indépendante possible est une variable qui ne permet que de deviner ; la valeur de son entropie de modèle dépend des données. Dans ce fichier de données, 1 256 (ou 0,2512) des 5 000 clients totaux ne parviennent pas à rembourser leur emprunt et 3 744 (ou 0,7488) y parviennent. Par conséquent, la pire variable indépendante possible comporterait une entropie de modèle de $-0{,}2512 \times \log_2(0{,}2512) - 0{,}7488 \times \log_2(0{,}7488) = 0{,}8132$.

La conclusion la plus probante possible est celle selon laquelle les variables comportant des valeurs d'entropie de modèle inférieures sont de meilleures variables indépendantes, car ce qui constitue une bonne valeur d'entropie de modèle dépend des applications et des données. Dans ce cas, il s'avère que les variables avec un nombre élevé de casiers générés, par rapport au nombre de catégories distinctes, comportent des valeurs d'entropie de modèle inférieures. Une évaluation approfondie de ces variables d'entrée de regroupement en tant que variables indépendantes doit être effectuée à l'aide des procédures de modélisation de prévision, qui comportent des outils plus complets pour la sélection des variables.

Récapitulatifs de regroupement par casiers

Le récapitulatif de regroupement par casiers rend compte des limites des casiers générés et de l'effectif de chaque casier en fonction des valeurs de la variable guide. Un tableau récapitulatif de regroupement distinct est généré pour chaque variable d'entrée de regroupement.

Figure 10-6
Récapitulatif de regroupement pour la variable Age en années

	Point final		Nombre d'observations par niveau de Antécédent découvert		
Casier	Inférieur	Supérieur	Non	Oui	Total
1	[a]	32	1129	639	1768
2	32	[a]	2615	617	3232
Total			3744	1256	5000

Chaque casier est calculé selon la formule Inférieur <=Age en années < Supérieur.

a. Non bornée

Le récapitulatif de la variable *Age en années* indique que 1 768 clients, tous âgés de 32 ans maximum, sont placés dans le casier 1, tandis que les 3 232 clients restants, tous âgés de plus de 32 ans, sont placés dans le casier 2. La proportion de clients ayant précédemment manqué à leur engagement de remboursement est largement plus élevée dans le casier 1 (639/1 768 = 0,361) que dans le casier 2 (617/3 232 = 0,191).

Figure 10-7
Récapitulatif de regroupement pour la variable Revenu du ménage en milliers

	Point final		Nombre d'observations par niveau de Antécédent découvert		
Casier	Inférieur	Supérieur	Non	Oui	Total
1	[a]	26.70	1054	513	1567
2	26.70	[a]	2690	743	3433
Total			3744	1256	5000

Chaque casier est calculé selon la formule Inférieur <=Revenu du foyer en milliers < Supérieur.

a. Non bornée

Le récapitulatif de la variable *Revenu du ménage en milliers* indique une tendance similaire, avec un point de césure unique à 26,70 et une proportion supérieure de clients ayant précédemment été en défaut de paiement dans le casier 1 (513/1 567 = 0,327) par rapport au casier 2 (743/3 433 = 0,216). Comme les statistiques d'entropie de modèle le laissent présager, la différence de ces proportions n'est pas aussi marquée que celle des proportions relatives à la variable *Age en années*.

Figure 10-8
Récapitulatif de regroupement pour la variable Autres dettes en milliers

	Point final		Nombre d'observations par niveau de Antécédent découvert		
Casier	Inférieur	Supérieur	Non	Oui	Total
1	[a]	2,19	2161	539	2700
2	2,19	[a]	1583	717	2300
Total			3744	1256	5000

Chaque casier est calculé selon la formule Inférieur <= Autre dette en milliers < Supérieur.

a. Non bornée

Le récapitulatif de la variable *Autres dettes en milliers* indique une tendance inverse, avec un point de césure unique à 2,19 et une proportion inférieure de clients ayant précédemment été en défaut de paiement dans le casier 1 (539/2 700 = 0,200) par rapport au casier 2 (717/2 300 = 0,312). Là encore, comme les statistiques d'entropie de modèle le laissent entrevoir, la différence de ces proportions n'est pas aussi marquée que celle des proportions relatives à la variable *Age en années*.

Figure 10-9
Récapitulatif de regroupement pour la variable Nb d'années avec l'employeur actuel

	Point final		Nombre d'observations par niveau de Antécédent découvert		
Casier	Inférieur	Supérieur	Non	Oui	Total
1	[a]	3	629	458	1107
2	3	8	1066	461	1527
3	8	18	1471	268	1739
4	18	[a]	578	49	627
Total			3744	1256	5000

Chaque casier est calculé selon la formule Inférieur <= Nb d'années avec l'employeur actuel < Supérieur.

a. Non bornée

Le récapitulatif de la variable *Nb d'années avec l'employeur actuel* indique une tendance dans laquelle les proportions de clients en défaut de paiement diminuent à mesure que les nombres de casiers augmentent.

Casier	**Proportion de clients en défaut de paiement**
1	0.432
2	0.302
3	0.154
4	0.078

Figure 10-10
Récapitulatif de regroupement pour la variable Nb d'années à la même adresse (adresse)

	Point final		Nombre d'observations par niveau de Antécédent découvert		
Casier	Inférieur	Supérieur	Non	Oui	Total
1	[a]	7	1652	829	2481
2	7	14	1184	313	1497
3	14	[a]	906	114	1022
Total			3744	1256	5000

Chaque casier est calculé selon la formule Inférieur <= Nb d'années à l'adresse actuelle < Supérieur.

a. Non bornée

Le récapitulatif de la variable *Nb d'années à la même adresse (adresse)* indique une tendance similaire. Comme les statistiques d'entropie de modèle le laissent présager, les différences entre les casiers dans la proportion de clients en défaut de paiement est plus marquée pour la variable *Nb d'années avec l'employeur actuel* que pour la variable *Nb d'années à la même adresse (adresse)*.

Casier	**Proportion de clients en défaut de paiement**
1	0.334
2	0.209
3	0.112

Figure 10-11
Récapitulatif de regroupement pour la variable Dette carte de crédit (en milliers)

	Point final		Nombre d'observations par niveau de Antécédent découvert		
Casier	Inférieur	Supérieur	Non	Oui	Total
1	[a]	,97	2169	466	2635
2	,97	1,91	848	307	1155
3	1,91	6,05	643	352	995
4	6,05	[a]	84	131	215
Total			3744	1256	5000

Chaque casier est calculé selon la formule Inférieur <= Débit carte de crédit en milliers < Supérieur.

a. Non bornée

Le récapitulatif de la variable *Dette carte de crédit (en milliers)* indique la tendance inverse, avec des proportions de clients en défaut de paiement en augmentation à mesure que les nombres de casiers augmentent. Les variables *Nb d'années avec l'employeur actuel* et *Nb d'années à la même adresse (adresse)* identifient mieux les clients présentant une probabilité élevée de rembourser leur emprunt, tandis que la variable *Dette carte de crédit (en milliers)* identifie mieux les clients présentant une probabilité élevée de ne pas le rembourser.

Casier	**Proportion de clients en défaut de paiement**
1	0.177
2	0.266

Casier	Proportion de clients en défaut de paiement
3	0.354
4	0.609

Figure 10-12
Récapitulatif de regroupement pour la variable Rapport dette/revenu (x 100)

	Point final		Nombre d'observations par niveau de Antécédent découvert		
Casier	Inférieur	Supérieur	Non	Oui	Total
1	[a]	4,39	912	88	1000
2	4,39	12,09	2006	437	2443
3	12,09	18,71	625	386	1011
4	18,71	31,00	198	303	501
5	31,00	[a]	3	42	45
Total			3744	1256	5000

Chaque casier est calculé selon la formule Inférieur <= Ratio Débit/Crédit (x100) < Supérieur.

a. Non bornée

Le récapitulatif de la variable *Rapport dette/revenu (x 100)* indique une tendance similaire à la variable *Dette carte de crédit (en milliers)*. Cette variable affiche la valeur d'entropie de modèle la plus faible et constitue donc la meilleure variable indépendante potentielle de la probabilité de non-paiement. Elle classe mieux les clients présentant une probabilité élevée d'être en défaut de paiement que la variable *Dette carte de crédit (en milliers)*, et classe presque aussi bien les clients présentant une faible probabilité d'être en défaut de paiement que la variable *Nb d'années avec l'employeur actuel*.

Casier	Proportion de clients en défaut de paiement
1	0.088
2	0.179
3	0.382
4	0.605
5	0.933

Variables regroupées

Figure 10-13
Variables regroupées du fichier bankloan_binning.sav dans l'éditeur de données

	default	age_bin	employ_bin	address_bin	income_bin	debtinc_bin	creddebt_bin	othdebt_bin
1	0	2	3	2	2	2	4	2
2	0	1	3	1	2	3	2	2
3	0	2	3	3	2	2	1	1
4	0	2	3	3	2	1	3	1
5	0	1	1	1	2	3	2	2
6	0	2	2	1	1	2	1	1
7	1	2	4	2	2	4	3	2
8	0	2	3	2	2	1	1	1
9	0	1	2	1	1	4	2	2
10	0	2	1	2	1	4	3	1
11	0	1	1	1	1	1	1	1
12	1	1	2	1	1	2	1	1
13	0	2	4	3	2	2	3	2
14	1	2	2	2	2	3	2	2
15	0	2	4	3	2	2	3	2

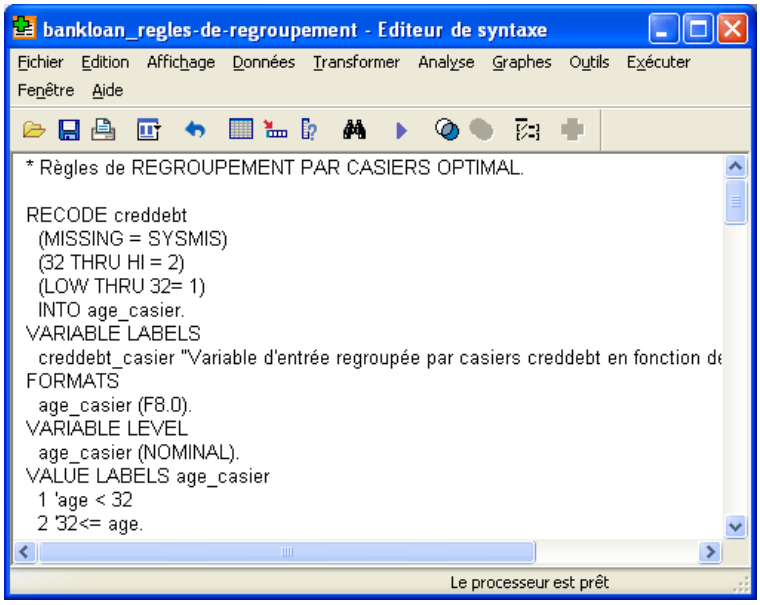
Les résultats du processus de regroupement dans ce fichier de données sont évidents dans l'éditeur de données. Ces variables regroupées sont utiles pour produire des récapitulatifs personnalisés des résultats de regroupement à l'aide de procédures descriptives ou de rapport, mais il est déconseillé d'utiliser ce fichier de données pour construire un modèle de prévision car les règles de regroupement ont été générées à l'aide de ces observations. Une meilleure méthode consiste à appliquer les règles de regroupement à un autre fichier de données contenant des informations sur d'autres clients.

Application de règles de regroupement de syntaxe

Pendant l'exécution de la procédure Recodage supervisé optimal, vous avez demandé à ce que les règles de regroupement générées par la procédure soient enregistrées sous forme de syntaxe de commande.

- Ouvrez le fichier *bankloan_binning-rules.sps*.

Figure 10-14
Fichier de règles de syntaxe



Pour chaque variable d'entrée de regroupement, il existe un bloc de syntaxe de commande qui effectue le regroupement, définit l'étiquette, le format et le niveau de variable, et définit les étiquettes de valeur des casiers. Ces commandes peuvent être appliquées à un fichier de données avec les mêmes variables que *bankloan_binning.sav*.

- Ouvrez le fichier *bankloan.sav*. Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A sur p. 138.
- Revenez à la vue d'éditeur de syntaxe du fichier *bankloan_binning-rules.sps*.

▶ Pour appliquer les règles de regroupement, dans les menus de l'éditeur de syntaxe, choisissez : Exécuter > Tous

Figure 10-15
Variables regroupées du fichier bankloan.sav dans l'éditeur de données

	preddef3	age_bin	employ_bin	address_bin	income_bin	debtinc_bin	creddebt_bin	othdebt_bin
1	0,21304	2	3	2	2	2	4	2
2	0,43690	1	3	1	2	3	2	2
3	0,14102	2	3	3	2	2	1	1
4	0,10442	2	3	3	2	1	3	1
5	0,43690	1	1	1	2	3	2	2
6	0,23358	2	2	1	1	2	1	1
7	0,81709	2	4	2	2	4	3	2
8	0,11336	2	3	2	2	1	1	1
9	0,66390	1	2	1	1	4	2	2
10	0,51553	2	1	2	1	4	3	1
11	0,09055	1	1	1	1	1	1	1
12	0,13631	1	2	1	1	2	1	1
13	0,22890	2	4	3	2	2	3	2
14	0,40484	2	2	2	2	3	2	2
15	0,20866	2	4	3	2	2	3	2

Les variables figurant dans le fichier *bankloan.sav* ont été regroupées en fonction des règles générées par l'exécution de la procédure Recodage supervisé optimal sur le fichier *bankloan_binning.sav*. Ce fichier de données peut à présent être utilisé pour construire des modèles de prévision préférant ou nécessitant des variables catégorielles.

Récapitulatif

Avec la procédure Recodage supervisé optimal, nous avons généré des règles de regroupement pour des variables d'échelle constituant des variables indépendantes potentielles de probabilité de défaut de paiement, que nous avons appliquées à un fichier de données distinct.

Au cours du processus de regroupement, vous avez constaté que les valeurs *Nb d'années avec l'employeur actuel* et *Nb d'années à la même adresse (adresse)* regroupées identifient mieux les clients présentant une probabilité élevée de rembourser leur emprunt, tandis que la variable *Dette carte de crédit (en milliers)* identifie mieux les clients présentant une probabilité élevée de ne pas le rembourser. Cette observation intéressante offre un éclairage supplémentaire lors de la construction de modèles de prévision pour la probabilité de défaut de paiement. Si éviter les mauvaises dettes est votre principale préoccupation, la variable *Dette carte de crédit (en milliers)* est plus importante que les variables *Nb d'années avec l'employeur actuel* et *Nb d'années à la même adresse (adresse)*. Si le développement de votre base client est votre priorité, les variables *Nb d'années avec l'employeur actuel* et *Nb d'années à la même adresse (adresse)* sont plus importantes.

Annexe

A

Fichiers d'exemple

Les fichiers d'exemple installés avec le produit figurent dans le sous-répertoire *Echantillons* du répertoire d'installation. Il existe un dossier distinct au sein du sous-répertoire Echantillons pour chacune des langues suivantes : Anglais, Français, Allemand, Italien, Japonais, Coréen, Polonais, Russe, Chinois simplifié, Espangol et Chinois traditionnel.

Seuls quelques fichiers d'exemples sont disponibles dans toutes les langues. Si un fichier d'exemple n'est pas disponible dans une langue, le dossier de langue contient la version anglaise du fichier d'exemple.

Descriptions

Voici de brèves descriptions des fichiers d'exemple utilisés dans divers exemples à travers la documentation.

- **accidents.sav.**Ce fichier de données d'hypothèse concerne une société d'assurance qui étudie les facteurs de risque liés à l'âge et au sexe dans les accidents de la route survenant dans une région donnée. Chaque observation correspond à une classification croisée de la catégorie d'âge et du sexe.
- **adl.sav.** Ce fichier de données d'hypothèse concerne les mesures entreprises pour identifier les avantages d'un type de thérapie proposé aux patients qui ont subi une attaque cardiaque. Les médecins ont assigné de manière aléatoire les patients du sexe féminin ayant subi une attaque cardiaque à un groupe parmi deux groupes possibles. Le premier groupe a fait l'objet de la thérapie standard tandis que le second a bénéficié en plus d'une thérapie émotionnelle. Trois mois après les traitements, les capacités de chaque patient à effectuer les tâches ordinaires de la vie quotidienne ont été notées en tant que variables ordinales.
- **advert.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un détaillant pour examiner la relation existant entre l'argent dépensé dans la publicité et les ventes résultantes. Pour ce faire, il collecte les chiffres des ventes passées et les coûts associés à la publicité.
- **aflatoxin.sav.** Ce fichier de données d'hypothèse concerne le test de l'aflatoxine dans des récoltes de maïs. La concentration de ce poison varie largement d'une récolte à l'autre et au sein de chaque récolte. Un processeur de grain a reçu 16 échantillons issus de 8 récoltes de maïs et a mesuré les niveaux d'alfatoxine en parties par milliard (PPB).
- **aflatoxin20.sav.** Ce fichier de données contient les mesures d'aflatoxine de chacun des 16 échantillons des récoltes 4 et 8 du fichier de données *aflatoxin.sav*.
- **anorectic.sav.** En cherchant à développer une symptomatologie standardisée du comportement anorexique/boulimique, des chercheurs(Van der Ham, Meulman, Van Strien, et Van Engeland, 1997) ont examiné 55 adolescents souffrant de troubles alimentaires. Chaque patient a été

© Copyright SPSS Inc. 1989, 2010

observé quatre fois sur une période de quatre années, soit un total de 220 observations. A chaque observation, les patients ont été notés pour chacun des 16 symptômes. En raison de l'absence de scores de symptôme pour le patient 71/visite 2, le patient 76/visite 2 et le patient 47/visite 3, le nombre d'observations valides est de 217.

- **autoaccidents.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un analyste en assurances pour modéliser le nombre d'accidents de la route par conducteur tout en prenant en compte l'âge et le sexe du conducteur. Chaque observation représente un conducteur distinct et enregistre son sexe, son âge et le nombre d'accidents de la route au cours des cinq dernières années.
- **band.sav.** Ce fichier de données contient les chiffres de ventes hebdomadaires hypothétiques de CD musicaux d'un groupe. Les données relatives à trois variables explicatives possibles sont également incluses.
- **bankloan.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend une banque pour réduire le taux de défaut de paiement. Il contient des informations financières et démographiques sur 850 clients existants et éventuels. Les premières 700 observations concernent des clients auxquels des prêts ont été octroyés. Les 150 dernières observations correspondant aux clients éventuels que la banque doit classer comme bons ou mauvais risques de crédit.
- **bankloan_binning.sav.** Ce fichier de données d'hypothèse concerne des informations financières et démographiques sur 5 000 clients existants.
- **behavior.sav.** Dans un exemple classique (Price et Bouffard, 1974), on a demandé à 52 étudiants de noter les combinaisons établies à partir de 15 situations et de 15 comportements sur une échelle de 0 à 9, où 0 = « extrêmement approprié » et 9 = « extrêmement inapproprié ». En effectuant la moyenne des résultats de l'ensemble des individus, on constate une certaine différence entre les valeurs.
- **behavior_ini.sav.** Ce fichier de données contient la configuration initiale d'une solution bidimensionnelle pour *behavior.sav*.
- **brakes.sav.** Ce fichier de données d'hypothèse concerne le contrôle qualité effectué dans une usine qui fabrique des freins à disque pour des voitures haut de gamme. Le fichier de données contient les mesures de diamètre de 16 disques de 8 machines de production. Le diamètre cible des freins est de 322 millimètres.
- **breakfast.sav.** Au cours d'une étude classique (Green et Rao, 1972), on a demandé à 21 étudiants en MBA (Master of Business Administration) de l'école de Wharton et à leurs conjoints de classer 15 aliments du petit-déjeuner selon leurs préférences, de 1= « aliment préféré » à 15= « aliment le moins apprécié ». Leurs préférences ont été enregistrées dans six scénarios différents, allant de « Préférence générale » à « En-cas avec boisson uniquement ».
- **breakfast-overall.sav.** Ce fichier de données contient les préférences de petit-déjeuner du premier scénario uniquement, « Préférence générale ».
- **broadband_1.sav.** Ce fichier de données d'hypothèse concerne le nombre d'abonnés, par région, à un service haut débit. Le fichier de données contient le nombre d'abonnés mensuels de 85 régions sur une période de quatre ans.
- **broadband_2.sav.** Ce fichier de données est identique au fichier *broadband_1.sav* mais contient les données relatives à trois mois supplémentaires.

- **car_insurance_claims.sav.** Il s'agit d'un ensemble de données présenté et analysé ailleurs (McCullagh et Nelder, 1989) qui concerne des actions en indemnisation pour des voitures. Le montant d'action en indemnisation moyen peut être modelé comme présentant une distribution gamma, à l'aide d'une fonction de lien inverse pour associer la moyenne de la variable dépendante à une combinaison linéaire de l'âge de l'assuré, du type de véhicule et de l'âge du véhicule. Le nombre d'actions entreprises peut être utilisé comme pondération de positionnement.
- **car_sales.sav.** Ce fichier de données contient des estimations de ventes hypothétiques, des barèmes de prix et des spécifications physiques concernant divers modèles et marques de véhicule. Les barèmes de prix et les spécifications physiques proviennent tour à tour de *edmunds.com* et des sites des constructeurs.
- **car_sales_uprepared.sav.** Il s'agit d'une version modifiée de *car_sales.sav* qui n'inclut aucune version transformée des champs.
- **carpet.sav.** Dans un exemple courant (Green et Wind, 1973), une société intéressée par la commercialisation d'un nouveau nettoyeur de tapis souhaite examiner l'influence de cinq critères sur la préférence du consommateur : la conception du conditionnement, la marque, le prix, une étiquette *Economique* et une garantie satisfait ou remboursé. Il existe trois niveaux de critère pour la conception du conditionnement, suivant l'emplacement de l'applicateur, trois marques (*K2R*, *Glory* et *Bissell*), trois niveaux de prix et deux niveaux (non ou oui) pour chacun des deux derniers critères. Dix consommateurs classent 22 profils définis par ces critères. La variable *Préférence* indique le classement des rangs moyens de chaque profil. Un rang faible correspond à une préférence élevée. Cette variable reflète une mesure globale de préférence pour chaque profil.
- **carpet_prefs.sav.** Ce fichier de données repose sur le même exemple que celui décrit pour *carpet.sav*, mais contient les classements réels issus de chacun des 10 clients. On a demandé aux consommateurs de classer les 22 profils de produits, du préféré au moins intéressant. Les variables *PREF1* à *PREF22* contiennent les identificateurs des profils associés, tels qu'ils sont définis dans *carpet_plan.sav*.
- **catalog.sav.** Ce fichier de données contient des chiffres de ventes mensuelles hypothétiques relatifs à trois produits vendus par une entreprise de vente par correspondance. Les données relatives à cinq variables explicatives possibles sont également incluses.
- **catalog_seasfac.sav.** Ce fichier de données est identique à *catalog.sav* mais contient en plus un ensemble de facteurs saisonniers calculés à partir de la procédure de désaisonnalisation, ainsi que les variables de date correspondantes.
- **cellular.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un opérateur téléphonique pour réduire les taux de désabonnement. Des scores de propension au désabonnement sont attribués aux comptes, de 0 à 100. Les comptes ayant une note égale ou supérieure à 50 sont susceptibles de changer de fournisseur.
- **ceramics.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un fabricant pour déterminer si un nouvel alliage haute qualité résiste mieux à la chaleur qu'un alliage standard. Chaque observation représente un test séparé de l'un des deux alliages ; le degré de chaleur auquel l'alliage ne résiste pas est enregistré.

- **cereal.sav.** Ce fichier de données d'hypothèse concerne un sondage de 880 personnes interrogées sur leurs préférences de petit-déjeuner et sur leur âge, leur sexe, leur situation familiale et leur mode de vie (actif ou non actif, selon qu'elles pratiquent une activité physique au moins deux fois par semaine). Chaque observation correspond à un répondant distinct.
- **clothing_defects.sav.** Ce fichier de données d'hypothèse concerne le processus de contrôle qualité observé dans une usine de textile. Dans chaque lot produit à l'usine, les inspecteurs prélèvent un échantillon de vêtements et comptent le nombre de vêtements qui ne sont pas acceptables.
- **coffee.sav.** Ce fichier de données concerne l'image perçue de six marques de café frappé (Kennedy, Riquier, et Sharp, 1996). Pour chacun des 23 attributs d'image de café frappé, les personnes sollicitées ont sélectionné toutes les marques décrites par l'attribut. Les six marques sont appelées AA, BB, CC, DD, EE et FF à des fins de confidentialité.
- **contacts.sav.** Ce fichier de données d'hypothèse concerne les listes de contacts d'un groupe de représentants en informatique d'entreprise. Chaque contact est classé selon le service de l'entreprise où il travaille et le classement de son entreprise. Sont également enregistrés le montant de la dernière vente effectuée, le temps passé depuis la dernière vente et la taille de l'entreprise du contact.
- **creditpromo.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un grand magasin pour évaluer l'efficacité d'une promotion récente de carte de crédit. A cette fin, 500 détenteurs de carte ont été sélectionnés au hasard. La moitié a reçu une publicité faisant la promotion d'un taux d'intérêt réduit sur les achats effectués dans les trois mois à venir. L'autre moitié a reçu une publicité saisonnière standard.
- **customer_dbase.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend une société pour utiliser les informations figurant dans sa banque de données et proposer des offres spéciales aux clients susceptibles d'être intéressés. Un sous-groupe de la base de clients a été sélectionné au hasard et a reçu des offres spéciales. Les réponses des clients ont été enregistrées.
- **customer_information.sav.** Un fichier de données d'hypothèse qui contient les informations postales du client, telles que le nom et l'adresse.
- **customer_subset.sav.** Un sous-ensemble de 80 observations de *customer_dbase.sav*.
- **customers_model.sav.** Ce fichier de données d'hypothèse concerne les personnes ciblées par une campagne de marketing. Ces données incluent des informations démographiques, un récapitulatif de l'historique d'achat et indiquent si chaque personne a répondu ou non à la campagne. Chaque observation représente une personne distincte.
- **customers_new.sav.** Ce fichier de données d'hypothèse concerne les personnes constituant des cibles potentielles pour une campagne de marketing. Ces données incluent des informations démographiques et un récapitulatif de l'historique d'achat pour chaque personne. Chaque observation représente une personne distincte.
- **debate.sav.** Ce fichier de données d'hypothèse concerne des réponses appariées à une enquête donnée aux participants à un débat politique avant et après le débat. Chaque observation représente un répondant distinct.
- **debate_aggregate.sav.** Il s'agit d'un fichier de données d'hypothèse qui rassemble les réponses dans le fichier *debate.sav*. Chaque observation correspond à une classification croisée de préférence avant et après le débat.

- **demo.sav.** Ce fichier de données d'hypothèse concerne une base de données clients achetée en vue de diffuser des offres mensuelles. Les données indiquent si le client a répondu ou non à l'offre et contiennent diverses informations démographiques.
- **demo_cs_1.sav.** Ce fichier de données d'hypothèse concerne la première mesure entreprise par une société pour compiler une base de données contenant des informations d'enquête. Chaque observation correspond à une ville différente. La région, la province, le quartier et la ville sont enregistrés.
- **demo_cs_2.sav.** Ce fichier de données d'hypothèse concerne la seconde mesure entreprise par une société pour compiler une base de données contenant des informations d'enquête. Chaque observation correspond à un ménage différent issu des villes sélectionnées à la première étape. La région, la province, le quartier, la ville, la sous-division et l'identification sont enregistrés. Les informations d'échantillonnage des deux premières étapes de la conception sont également incluses.
- **demo_cs.sav.** Ce fichier de données d'hypothèse concerne des informations d'enquête collectées via une méthode complexe d'échantillonnage. Chaque observation correspond à un ménage différent et diverses informations géographiques et d'échantillonnage sont enregistrées.
- **dmdata.sav.** Ceci est un fichier de données d'hypothèse qui contient des informations démographiques et des informations concernant les achats pour une entreprise de marketing direct. *dmdata2.sav* contient les informations pour un sous-ensemble de contacts qui ont reçu un envoi d'essai, et *dmdata3.sav* contient des informations sur les contacts restants qui n'ont pas reçu l'envoi d'essai.
- **dietstudy.sav.** Ce fichier de données d'hypothèse contient les résultats d'une étude portant sur le régime de Stillman(Rickman, Mitchell, Dingman, et Dalen, 1974). Chaque observation correspond à un sujet distinct et enregistre son poids en livres avant et après le régime, ainsi que ses niveaux de triglycérides en mg/100 ml.
- **dvdplayer.sav.** Ce fichier de données d'hypothèse concerne le développement d'un nouveau lecteur DVD. A l'aide d'un prototype, l'équipe de marketing a collecté des données de groupes spécifiques. Chaque observation correspond à un utilisateur interrogé et enregistre des informations démographiques sur cet utilisateur, ainsi que ses réponses aux questions portant sur le prototype.
- **german_credit.sav.** Ce fichier de données provient de l'ensemble de données « German credit » figurant dans le référentiel Machine Learning Databases (Blake et Merz, 1998) de l'université de Californie, Irvine.
- **grocery_1month.sav.** Ce fichier de données d'hypothèse est le fichier de données *grocery_coupons.sav* dans lequel les achats hebdomadaires sont organisés par client distinct. Certaines variables qui changeaient toutes les semaines disparaissent. En outre, le montant dépensé enregistré est à présent la somme des montants dépensés au cours des quatre semaines de l'enquête.
- **grocery_coupons.sav.** Il s'agit d'un fichier de données d'hypothèse qui contient des données d'enquête collectées par une chaîne de magasins d'alimentation qui chercher à déterminer les habitudes de consommation de ses clients. Chaque client est suivi pendant quatre semaines et chaque observation correspond à une semaine distincte. Les informations enregistrées concernent les endroits où le client effectue ses achats, la manière dont il les effectue, ainsi que les sommes dépensées en provisions au cours de cette semaine.

- **guttman.sav.** Bell (Bell, 1961) a présenté un tableau pour illustrer les groupes sociaux possibles. Guttman (Guttman, 1968) a utilisé une partie de ce tableau, dans lequel cinq variables décrivant des éléments tels que l'interaction sociale, le sentiment d'appartenance à un groupe, la proximité physique des membres et la formalité de la relation, ont été croisées avec sept groupes sociaux théoriques, dont les foules (par exemple, le public d'un match de football), l'audience (par exemple, au cinéma ou dans une salle de classe), le public (par exemple, les journaux ou la télévision), les bandes (proche d'une foule, mais qui serait caractérisée par une interaction beaucoup plus intense), les groupes primaires (intimes), les groupes secondaires (volontaires) et la communauté moderne (groupement lâche issu d'une forte proximité physique et d'un besoin de services spécialisés).
- **health_funding.sav.** Ce fichier de données d'hypothèse concerne des données sur le financement des soins de santé (montant par groupe de 100 individus), les taux de maladie (taux par groupe de 10 000 individus) et les visites chez les prestataires de soins de santé (taux par groupe de 10 000 individus). Chaque observation représente une ville différente.
- **hivassay.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un laboratoire pharmaceutique pour développer une analyse rapide de détection d'infection HIV. L'analyse a pour résultat huit nuances de rouge, les nuances les plus marquées indiquant une plus forte probabilité d'infection. Un test en laboratoire a été effectué sur 2 000 échantillons de sang, la moitié de ces échantillons étant infectée par le virus HIV et l'autre moitié étant saine.
- **hourlywagedata.sav.** Ce fichier de données d'hypothèse concerne les salaires horaires d'infirmières occupant des postes administratifs et dans les services de soins, et affichant divers niveaux d'expérience.
- **insurance_claims.sav.** Il s'agit d'un fichier de données hypothétiques qui concerne une compagnie d'assurance souhaitant développer un modèle pour signaler des réclamations suspectes, potentiellement frauduleuses. Chaque observation correspond à une réclamation distincte.
- **insure.sav.** Ce fichier de données d'hypothèse concerne une compagnie d'assurance qui étudie les facteurs de risque indiquant si un client sera amené à déclarer un incident au cours d'un contrat d'assurance vie d'une durée de 10 ans. Chaque observation figurant dans le fichier de données représente deux contrats, l'un ayant enregistré une réclamation et l'autre non, appariés par âge et sexe.
- **judges.sav.** Ce fichier de données d'hypothèse concerne les scores attribués par des juges expérimentés (plus un juge enthousiaste) à 300 performances de gymnastique. Chaque ligne représente une performance distincte ; les juges ont examiné les mêmes performances.
- **kinship_dat.sav.** Rosenberg et Kim (Rosenberg et Kim, 1975) se sont lancés dans l'analyse de 15 termes de parenté (cousin/cousine, fille, fils, frère, grand-mère, grand-père, mère, neveu, nièce, oncle, père, petite-fille, petit-fils, sœur, tante). Ils ont demandé à quatre groupes d'étudiants (deux groupes de femmes et deux groupes d'hommes) de trier ces termes en fonction des similarités. Deux groupes (un groupe de femmes et un groupe d'hommes) ont été invités à effectuer deux tris, en basant le second sur un autre critère que le premier. Ainsi, un total de six "sources" a été obtenu. Chaque source correspond à une matrice de proximité 15×15, dont le nombre de cellules est égal au nombre de personnes dans une source moins le nombre de fois où les objets ont été partitionnés dans cette source.
- **kinship_ini.sav.** Ce fichier de données contient une configuration initiale d'une solution tridimensionnelle pour *kinship_dat.sav*.

- **kinship_var.sav.** Ce fichier de données contient les variables indépendantes *sexe*, *génér*(ation) et *degré* (de séparation) permettant d'interpréter les dimensions d'une solution pour *kinship_dat.sav*. Elles permettent en particulier de réduire l'espace de la solution à une combinaison linéaire de ces variables.
- **marketvalues.sav.** Ce fichier de données concerne les ventes de maisons dans un nouvel ensemble à Algonquin (Illinois) au cours des années 1999–2000. Ces ventes relèvent des archives publiques.
- **nhis2000_subset.sav.**Le NHIS (National Health Interview Survey) est une enquête de grande envergure concernant la population des Etats-Unis. Des entretiens ont lieu avec un échantillon de ménages représentatifs de la population américaine. Des informations démographiques et des observations sur l'état de santé et le comportement sanitaire sont recueillies auprès des membres de chaque ménage. Ce fichier de données contient un sous-groupe d'informations issues de l'enquête de 2000. National Center for Health Statistics. National Health Interview Survey, 2000. Fichier de données et documentation d'usage public. *ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NHIS/2000/*. Accès en 2003.
- **ozone.sav.** Les données incluent 330 observations portant sur six variables météorologiques pour prévoir la concentration d'ozone à partir des variables restantes. Des chercheurs précédents (Breiman et Friedman, 1985), (Hastie et Tibshirani, 1990), ont décelé parmi ces variables des non-linéarités qui pénalisent les approches standard de la régression.
- **pain_medication.sav.** Ce fichier de données d'hypothèse contient les résultats d'un essai clinique d'un remède anti-inflammatoire traitant les douleurs de l'arthrite chronique. On cherche notamment à déterminer le temps nécessaire au médicament pour agir et les résultats qu'il permet d'obtenir par rapport à un médicament existant.
- **patient_los.sav.** Ce fichier de données d'hypothèse contient les dossiers médicaux de patients admis à l'hôpital pour suspicion d'infarctus du myocarde suspecté (ou « attaque cardiaque »). Chaque observation correspond à un patient distinct et enregistre de nombreuses variables liées à son séjour à l'hôpital.
- **patlos_sample.sav.** Ce fichier de données d'hypothèse contient les dossiers médicaux d'un échantillon de patients sous traitement thrombolytique après un infarctus du myocarde. Chaque observation correspond à un patient distinct et enregistre de nombreuses variables liées à son séjour à l'hôpital.
- **polishing.sav.** Il s'agit du fichier de données du « Nambeware Polishing Times » de la Data and Story Library. Il concerne les mesures qu'entreprend un fabricant de vaisselle en métal (Nambe Mills, Santa Fe, Nouveau-Mexique) pour planifier sa production. Chaque observation représente un article différent de la gamme de produits. Le diamètre, le temps de polissage, le prix et le type de produit sont enregistrés pour chaque article.
- **poll_cs.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un enquêteur pour déterminer le niveau de soutien du public pour un projet de loi avant législature. Les observations correspondent à des électeurs enregistrés. Chaque observation enregistre le comté, la ville et le quartier où habite l'électeur.
- **poll_cs_sample.sav.** Ce fichier de données d'hypothèse contient un échantillon des électeurs répertoriés dans le fichier *poll_cs.sav*. L'échantillon a été prélevé selon le plan spécifié dans le fichier de plan *poll.csplan* et ce fichier de données enregistre les probabilités d'inclusion et les pondérations d'échantillon. Toutefois, ce plan faisant appel à une méthode d'échantillonnage de probabilité proportionnelle à la taille (PPS – Probability-Proportional-to-Size), il existe

également un fichier contenant les probabilités de sélection conjointes (*poll_jointprob.sav*). Les variables supplémentaires correspondant à la répartition démographique des électeurs et à leur opinion sur le projet de loi proposé ont été collectées et ajoutées au fichier de données une fois l'échantillon prélevé.

- **property_assess.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un contrôleur au niveau du comté pour maintenir les évaluations de valeur de propriété à jour sur des ressources limitées. Les observations correspondent à des propriétés vendues dans le comté au cours de l'année précédente. Chaque observation du fichier de données enregistre la ville où se trouve la propriété, l'évaluateur ayant visité la propriété pour la dernière fois, le temps écoulé depuis cette évaluation, l'évaluation effectuée à ce moment-là et la valeur de vente de la propriété.
- **property_assess_cs.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un contrôleur du gouvernement pour maintenir les évaluations de valeur de propriété à jour sur des ressources limitées. Les observations correspondent à des propriétés de l'état. Chaque observation du fichier de données enregistre le comté, la ville et le quartier où se trouve la propriété, le temps écoulé depuis la dernière évaluation et l'évaluation alors effectuée.
- **property_assess_cs_sample.sav.** Ce fichier de données d'hypothèse contient un échantillon des propriétés répertoriées dans le fichier *property_assess_cs.sav*. L'échantillon a été prélevé selon le plan spécifié dans le fichier de plan *property_assess.csplan* et ce fichier de données enregistre les probabilités d'inclusion et les pondérations d'échantillon. La variable supplémentaire *Valeur courante* a été collectée et ajoutée au fichier de données une fois l'échantillon prélevé.
- **recidivism.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend une agence administrative d'application de la loi pour interpréter les taux de récidive dans la juridiction. Chaque observation correspond à un récidiviste et enregistre les informations démographiques qui lui sont propres, certains détails sur le premier délit commis, ainsi que le temps écoulé jusqu'à la seconde arrestation si elle s'est produite dans les deux années suivant la première.
- **recidivism_cs_sample.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend une agence administrative d'application de la loi pour interpréter les taux de récidive dans la juridiction. Chaque observation correspond à un récidiviste libéré suite à la première arrestation en juin 2003 et enregistre les informations démographiques qui lui sont propres, certains détails sur le premier délit commis et les données relatives à la seconde arrestation, si elle a eu lieu avant fin juin 2006. Les récidivistes ont été choisis dans plusieurs départements échantillonnés conformément au plan d'échantillonnage spécifié dans *recidivism_cs.csplan*. Ce plan faisant appel à une méthode d'échantillonnage de probabilité proportionnelle à la taille (PPS - Probability proportional to size), il existe également un fichier contenant les probabilités de sélection conjointes (*recidivism_cs_jointprob.sav*).
- **rfm_transactions.sav.**Un fichier de données d'hypothèse qui contient les données de transaction d'achat, y compris la date d'achat, le/les élément(s) acheté(s) et le montant monétaire pour chaque transaction.
- **salesperformance.sav.** Ce fichier de données d'hypothèse concerne l'évaluation de deux nouveaux cours de formation en vente. Soixante employés, divisés en trois groupes, reçoivent chacun une formation standard. En outre, le groupe 2 suit une formation technique et le groupe 3 un didacticiel pratique. A l'issue du cours de formation, chaque employé est testé et

sa note enregistrée. Chaque observation du fichier de données représente un stagiaire distinct et enregistre le groupe auquel il a été assigné et la note qu'il a obtenue au test.

- **satisf.sav.** Il s'agit d'un fichier de données d'hypothèse portant sur une enquête de satisfaction effectuée par une société de vente au détail au niveau de quatre magasins. Un total de 582 clients ont été interrogés et chaque observation représente la réponse d'un seul client.
- **screws.sav.**Ce fichier de données contient des informations sur les descriptives des vis, des boulons, des écrous et des clous.(Hartigan, 1975).
- **shampoo_ph.sav.** Ce fichier de données d'hypothèse concerne le processus de contrôle qualité observé dans une usine de produits capillaires. A intervalles réguliers, six lots de sortie distincts sont mesurés et leur pH enregistré. La plage cible est 4,5–5,5.
- **ships.sav.** Il s'agit d'un ensemble de données présenté et analysé ailleurs (McCullagh et al., 1989) et concernant les dommages causés à des cargos par les vagues. Les effectifs d'incidents peuvent être modélisés comme des incidents se produisant selon un taux de Poisson en fonction du type de navire, de la période de construction et de la période de service. Les mois de service totalisés pour chaque cellule du tableau formé par la classification croisée des facteurs fournissent les valeurs d'exposition au risque.
- **site.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend une société pour choisir de nouveaux sites pour le développement de ses activités. L'entreprise a fait appel à deux consultants pour évaluer séparément les sites. Ces consultants, en plus de fournir un rapport approfondi, ont classé chaque site comme constituant une éventualité « bonne », « moyenne » ou « faible ».
- **smokers.sav.** Ce fichier de données est extrait de l'étude National Household Survey of Drug Abuse de 1998 et constitue un échantillon de probabilité des ménages américains. (*http://dx.doi.org/10.3886/ICPSR02934*) Ainsi, la première étape dans l'analyse de ce fichier doit consister à pondérer les données pour refléter les tendances de population.
- **stroke_clean.sav.** Ce fichier de données d'hypothèse concerne l'état d'une base de données médicales une fois celle-ci purgée via des procédures de l'option Validation de données.
- **stroke_invalid.sav.** Ce fichier de données d'hypothèse concerne l'état initial d'une base de données médicales et comporte plusieurs erreurs de saisie de données.
- **stroke_survival.** Ce fichier de données d'hypothèse concerne les temps de survie de patients qui quittent un programme de rééducation à la suite d'un accident ischémique et rencontrent un certain nombre de problèmes. Après l'attaque, l'occurrence d'infarctus du myocarde, d'accidents ischémiques ou hémorragiques est signalée, et le moment de l'événement enregistré. L'échantillon est tronqué à gauche car il n'inclut que les patients ayant survécu durant le programme de rééducation mis en place suite à une attaque.
- **stroke_valid.sav.** Ce fichier de données d'hypothèse concerne l'état d'une base de données médicales une fois les valeurs vérifiées via la procédure Validation de données. Elle contient encore des observations anormales potentielles.
- **survey_sample.sav.** Ce fichier de données concerne des informations d'enquête dont des données démographiques et des mesures comportementales. Il est basé sur un sous-ensemble de variables de la 1998 NORC General Social Survey, bien que certaines valeurs de données aient été modifiées et que des variables supplémentaires fictives aient été ajoutées à titre de démonstration.

- **telco.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend une société de télécommunications pour réduire les taux de désabonnement de sa base de clients. Chaque observation correspond à un client distinct et enregistre diverses informations démographiques et d'utilisation de service.
- **telco_extra.sav.** Ce fichier de données est semblable au fichier de données *telco.sav* mais les variables de permanence et de dépenses des consommateurs transformées log ont été supprimées et remplacées par des variables de dépenses des consommateurs transformées log standardisées.
- **telco_missing.sav.** Ce fichier de données est un sous-ensemble du fichier de données *telco.sav* mais certaines des valeurs de données démographiques ont été remplacées par des valeurs manquantes.
- **testmarket.sav.** Ce fichier de données d'hypothèse concerne une chaîne de fast foods et ses plans marketing visant à ajouter un nouveau plat à son menu. Trois campagnes étant possibles pour promouvoir le nouveau produit, le nouveau plat est introduit sur des sites sur plusieurs marchés sélectionnés au hasard. Une promotion différente est effectuée sur chaque site et les ventes hebdomadaires du nouveau plat sont enregistrées pour les quatre premières semaines. Chaque observation correspond à un site-semaine distinct.
- **testmarket_1month.sav.** Ce fichier de données d'hypothèse est le fichier de données *testmarket.sav* dans lequel les ventes hebdomadaires sont organisées par site distinct. Certaines variables qui changeaient toutes les semaines disparaissent. En outre, les ventes enregistrées sont à présent la somme des ventes réalisées au cours des quatre semaines de l'enquête.
- **tree_car.sav.** Ce fichier de données d'hypothèse concerne des données démographiques et de prix d'achat de véhicule.
- **tree_credit.sav.** Ce fichier de données d'hypothèse concerne des données démographiques et d'historique de prêt bancaire.
- **tree_missing_data.sav** Ce fichier de données d'hypothèse concerne des données démographiques et d'historique de prêt bancaire avec un grand nombre de valeurs manquantes.
- **tree_score_car.sav.** Ce fichier de données d'hypothèse concerne des données démographiques et de prix d'achat de véhicule.
- **tree_textdata.sav.** Ce fichier de données simples ne comporte que deux variables et vise essentiellement à indiquer l'état par défaut des variables avant affectation du niveau de mesure et des étiquettes de valeurs.
- **tv-survey.sav.** Ce fichier de données d'hypothèse concerne une enquête menée par un studio de télévision qui envisage de prolonger la diffusion d'un programme ou de l'arrêter. On a demandé à 906 personnes si elles regarderaient le programme dans diverses situations. Chaque ligne représente un répondant distinct et chaque colonne une situation distincte.
- **ulcer_recurrence.sav.** Ce fichier contient des informations partielles d'une enquête visant à comparer l'efficacité de deux thérapies de prévention de la récurrence des ulcères. Il fournit un bon exemple de données censurées par intervalle et a été présenté et analysé ailleurs (Collett, 2003).

- **ulcer_recurrence_recoded.sav.** Ce fichier réorganise les informations figurant dans le fichier *ulcer_recurrence.sav* pour que vous puissiez modéliser la probabilité d'événement pour chaque intervalle de l'enquête plutôt que la probabilité d'événement de fin d'enquête. Il a été présenté et analysé ailleurs (Collett et al., 2003).
- **verd1985.sav.** Ce fichier de données concerne une enquête (Verdegaal, 1985). Les réponses de 15 sujets à 8 variables ont été enregistrées. Les variables présentant un intérêt sont divisées en trois ensembles. Le groupe 1 comprend l'*âge* et la *situation familiale*, le groupe 2 les *animaux domestiques* et la *presse*, et le groupe 3 la *musique* et l'*habitat*. A la variable *animal domestique* est appliqué un codage nominal multiple et à *âge*, un codage ordinal ; toutes les autres variables ont un codage nominal simple.
- **virus.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un fournisseur de services Internet pour déterminer les effets d'un virus sur ses réseaux. Il a suivi le pourcentage (approximatif) de trafic de messages électroniques infectés par un virus sur ses réseaux sur la durée, de la découverte à la circonscription de la menace.
- **wheeze_steubenville.sav.** Il s'agit d'un sous-ensemble d'une enquête longitudinale des effets de la pollution de l'air sur la santé des enfants (Ware, Dockery, Spiro III, Speizer, et Ferris Jr., 1984). Les données contiennent des mesures binaires répétées de l'état asthmatique d'enfants de la ville de Steubenville (Ohio), âgés de 7, 8, 9 et 10 ans, et indiquent si la mère fumait au cours de la première année de l'enquête.
- **workprog.sav.** Ce fichier de données d'hypothèse concerne un programme de l'administration visant à proposer de meilleurs postes aux personnes défavorisées. Un échantillon de participants potentiels au programme a ensuite été prélevé. Certains de ces participants ont été sélectionnés au hasard pour participer au programme. Chaque observation représente un participant au programme distinct.

Annexe
B

Notices

Licensed Materials – Property of SPSS Inc., an IBM Company. © Copyright SPSS Inc. 1989, 2010.

Patent No. 7,023,453

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: SPSS INC., AN IBM COMPANY, PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. SPSS Inc. may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-SPSS and non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this SPSS Inc. product and use of those Web sites is at your own risk.

When you send information to IBM or SPSS, you grant IBM and SPSS a nonexclusive right to use or distribute the information in any way it believes appropriate without incurring any obligation to you.

Information concerning non-SPSS products was obtained from the suppliers of those products, their published announcements or other publicly available sources. SPSS has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-SPSS products. Questions on the capabilities of non-SPSS products should be addressed to the suppliers of those products.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to SPSS Inc., for the purposes of developing,

using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. SPSS Inc., therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided "AS IS", without warranty of any kind. SPSS Inc. shall not be liable for any damages arising out of your use of the sample programs.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks of IBM Corporation, registered in many jurisdictions worldwide. A current list of IBM trademarks is available on the Web at *http://www.ibm.com/legal/copytrade.shmtl*.

SPSS is a trademark of SPSS Inc., an IBM Company, registered in many jurisdictions worldwide.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Java and all Java-based trademarks and logos are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.

This product uses WinWrap Basic, Copyright 1993-2007, Polar Engineering and Consulting, *http://www.winwrap.com*.

Other product and service names might be trademarks of IBM, SPSS, or other companies.

Adobe product screenshot(s) reprinted with permission from Adobe Systems Incorporated.

Microsoft product screenshot(s) reprinted with permission from Microsoft Corporation.

Bibliographie

Bell, E. H. 1961. *Social foundations of human behavior: Introduction to the study of sociology*. New York: Harper & Row.

Blake, C. L., et C. J. Merz. 1998. "UCI Repository of machine learning databases." Available at http://www.ics.uci.edu/~mlearn/MLRepository.html.

Breiman, L., et J. H. Friedman. 1985. Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80, .

Collett, D. 2003. *Modelling survival data in medical research*, 2 éd. Boca Raton: Chapman & Hall/CRC.

Green, P. E., et V. Rao. 1972. *Applied multidimensional scaling*. Hinsdale, Ill.: Dryden Press.

Green, P. E., et Y. Wind. 1973. *Multiattribute decisions in marketing: A measurement approach*. Hinsdale, Ill.: Dryden Press.

Guttman, L. 1968. A general nonmetric technique for finding the smallest coordinate space for configurations of points. *Psychometrika*, 33, .

Hartigan, J. A. 1975. *Clustering algorithms*. New York: John Wiley and Sons.

Hastie, T., et R. Tibshirani. 1990. *Generalized additive models*. Londres: Chapman and Hall.

Kennedy, R., C. Riquier, et B. Sharp. 1996. Practical applications of correspondence analysis to categorical data in market research. *Journal of Targeting, Measurement, and Analysis for Marketing*, 5, .

McCullagh, P., et J. A. Nelder. 1989. *Generalized Linear Models*, 2nd éd. Londres: Chapman & Hall.

Price, R. H., et D. L. Bouffard. 1974. Behavioral appropriateness and situational constraints as dimensions of social behavior. *Journal of Personality and Social Psychology*, 30, .

Rickman, R., N. Mitchell, J. Dingman, et J. E. Dalen. 1974. Changes in serum cholesterol during the Stillman Diet. *Journal of the American Medical Association*, 228, .

Rosenberg, S., et M. P. Kim. 1975. The method of sorting as a data-gathering procedure in multivariate research. *Multivariate Behavioral Research*, 10, .

Van der Ham, T., J. J. Meulman, D. C. Van Strien, et H. Van Engeland. 1997. Empirically based subgrouping of eating disorders in adolescents: A longitudinal perspective. *British Journal of Psychiatry*, 170, .

Verdegaal, R. 1985. *Meer sets analyse voor kwalitatieve gegevens (en néerlandais)*. Leiden: Department of Data Theory, University of Leiden.

Ware, J. H., D. W. Dockery, A. Spiro III, F. E. Speizer, et B. G. Ferris Jr.. 1984. Passive smoking, gas cooking, and respiratory health of children living in six cities. *American Review of Respiratory Diseases*, 129, .

© Copyright SPSS Inc. 1989, 2010

Index