

# Coheris Spad 7.4

## Release Notes

---

Version : 2.0

Date d'édition : 02/05/2011

*La maîtrise des modèles prédictifs*

# Sommaire

<b>1 - Introduction</b>	<b>3</b>
1.1 - Objet du document	3
<b>2 - Nouveautés de Coheris SPAD 7.4</b>	<b>4</b>
<b>3 - Date de disponibilité de Coheris SPAD 7.4</b>	<b>5</b>
<b>4 - Recodages supervisés</b>	<b>6</b>
4.1 - Créer de l'information intelligente	6
4.2 - Dans une implémentation orientée « métier »	7
<b>5 - Sélection automatique de variables explicatives pour la construction d'un modèle supervisé</b>	<b>8</b>
5.1 - La sélection automatique : Une étape indispensable	8
<b>6 - Statistiques de base</b>	<b>10</b>
<b>7 - Quelques Exemples</b>	<b>11</b>

# 1 - Introduction

---

## 1.1 - Objet du document

Ce document présente la liste des nouveautés et des améliorations de Coheris SPAD 7.4 par rapport à la précédente version (7.3).

Les fonctionnalités de Deployment Server ne sont pas décrites dans ce document, car il s'agit d'un nouveau produit à part entière. Pour plus d'informations, veuillez vous rapprocher de votre contact habituel Coheris.

## 2 - Nouveautés de Coheris SPAD 7.4

---

Dans ce document, vous trouverez une description synthétique des principales nouveautés de Coheris SPAD 7.4 :

- Création d'une information plus intelligente et mieux adaptée pour la construction des modèles statistiques (pour plus d'informations, voir le paragraphe 4 - ).
- Sélection automatique de l'information pertinente en entrée des modèles statistiques (pour plus d'informations, voir le paragraphe 5 - ).
- Nouvelle ergonomie et nouvelles sorties graphiques pour les « Statistiques de base », dans l'explorateur des données (pour plus d'informations, voir le paragraphe 6 - ).
- Autres améliorations de Coheris SPAD 7.4 :
  - Nouvelle procédure « Régression Logistique » plus puissante.
  - Il existe d'autres améliorations mineures et diverses corrections qui seront citées dans l'aide en ligne.

### **3 - Disponibilité de Coheris SPAD 7.4**

---

Coheris SPAD 7.4 sera mis à disposition des clients SPAD qui le souhaitent et dont la maintenance est à jour à partir du lundi 16 Mai 2011.

Cette mise à disposition se présentera sous la forme d'une mise à jour qui nécessite une nouvelle activation de la licence.

Cette mise à jour ne comprend pas la partie Deployment Server qui est un nouveau produit à part entière.

## 4 - Recodages supervisés

---

### 4.1 - Créer de l'information intelligente

Aussi sophistiqué qu'il soit, un modèle ne restitue que l'information des variables qui participent à son calcul, il est donc primordial de disposer de données intelligentes par rapport à l'objectif « métier » recherché.

La notion de « données intelligentes » est ici une notion de codage de l'information. En effet, à quoi cela sert-il de faire intervenir dans un modèle censé prédire l'attrition de clients, une variable « Age en classes », si les bornes des classes ne sont pas calculées en fonction de l'objectif : prédire l'attrition. De façon analogue, à quoi cela sert-il de faire intervenir dans un modèle, une variable qualitative dont les différentes modalités ne sont pas regroupées de façon optimale par rapport à ce même objectif. La réponse est évidente, ces informations « non intelligentes » vont, au mieux polluer le modèle, au pire le rendre inopérant.

Pour répondre à cette problématique, la version 7.4 de Coheris SPAD dispose d'une nouvelle méthode appelée « Recodages supervisés » qui permet de créer de l'information intelligente en fonction d'un objectif « métier ».

Deux algorithmes sont disponibles : le premier permet de mettre en classes une variable quantitative avec des bornes calculées de façon optimale pour prédire une variable dite « Variable cible ». L'algorithme MDLPC implanté est une méthode de discrétisation supervisée de variables continues utilisant la notion de gain d'entropie. La référence scientifique est : U. Fayyad et K. Irani, « Multi-interval discretization of continuous valued attributes for classification learning », In proceedings of the 13th International Joint Conference on Artificial Intelligence, pages 1022-1027, Morgan Kaufmann, 1993.

Le second permet de regrouper de façon optimale par rapport à une variable cible, les modalités d'une variable qualitative. L'algorithme est basé sur la mesure de proximité de profils et utilise le test d'équivalence distributionnelle du Khi2. La référence scientifique est l'article de G. Kass, « An exploratory technique for investigating large quantities of categorical data », in Applied Statistics, 29(2) :119 -127, 1980.

## 4.2 - Dans une implémentation orientée « métier »

L'interface de commande respecte bien sûr tous les critères qui font le succès de SPAD : convivialité de l'interface, paramétrage par défaut modifiable, édition de rapport au format Excel ou Html, création et exploitation immédiate des données créées, mais plus important encore, elle intègre l'aspect « métier ».

En effet, l'utilisateur a la possibilité d'opter soit pour une exécution dynamique, soit pour une exécution figée après la première exécution.

Dans le premier cas, les bornes de classes ou les regroupements sont recalculés à chaque exécution, en particulier si les données changent en amont. Ce mode d'utilisation correspond plutôt à une période de mise au point.

Dans le second cas, les bornes et les regroupements sont figés après le calcul initial et même si les données changent en amont, elles seront recodées selon les bornes ou regroupements ainsi figés. Ce mode correspond alors à une période d'exploitation ou de déploiement de modèles. Plus intéressant encore dans ce mode, l'utilisateur peut intervenir après le calcul statistique initial pour modifier manuellement les bornes et les regroupements initiaux en fonction de contraintes « métier ».

## 5 - Sélection automatique de variables explicatives pour la construction d'un modèle supervisé

---

### 5.1 - La sélection automatique : Une étape indispensable

Face à un volume d'information croissant et des outils de gestion et de transformation des données puissants, la construction d'un modèle peut se voir pénalisée en termes de temps de calcul ou d'interprétation, voir biaisée par un trop grand nombre de variables prédictives. La présélection intelligente et automatique des variables qui serviront à l'élaboration du modèle au regard d'une cible apporte la réponse à ces problèmes.

Les algorithmes de sélection sont basés sur des mesures de dépendance. L'utilisateur pourra mettre en œuvre des stratégies simples de type « ranking » qui permettent d'effectuer une sélection rapide et ordonnée, ou des stratégies plus évoluées qui analysent la pertinence de la sélection d'une variable en tenant compte de la redondance avec les variables déjà sélectionnées.

Pour la sélection de variables continues, la procédure propose deux choix :

- La méthode « Ranking par corrélation », basée sur le schéma classique de l'analyse de la variance, est particulièrement efficace lorsque les distributions conditionnelles des variables prédictives sont unimodales.
- Ce choix peut être complété par une analyse de séquences, méthode adaptée à des données moins usuelles (distributions conditionnelles non unimodales). (Référence scientifique : A. Mood, "The distribution Theory of Runs", in Annals of Mathematical Statistics, 11:367-392, 1940).

Pour la sélection des variables nominales, la procédure choisit l'algorithme le plus approprié en fonction de vos données. Cependant, l'utilisateur pourra toujours piloter les différents choix possibles :

- La méthode « Ranking » basée sur l'information mutuelle retient les variables liées individuellement à la variable cible.
- La méthode CFS (Correlation feature selection) prend en compte la redondance des variables sélectionnées. Elle repose sur un algorithme d'optimisation d'un « Critère de mérite » (Référence scientifique : M. Hall, S. Lloyd, « Feature subset selection : a correlation based filter approach », in 1997 Int. Conf. On Neural Information Processing and Intelligent Information Systems, pp/ 855-858, Springer, 1997).

- La méthode FCBF (Fast correlation based filter solution) prend en compte cette même notion de redondance. Son fonctionnement, similaire à la méthode CFS est basée sur le concept de « prédominance », plus strict en matière de sélection ce qui permet d'analyser plusieurs milliers de variables. (Référence scientifique : L. Yu and H. Liu. "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution". In Proceedings of The Twentieth International Conference on Machine Learning (ICML-03), pp 856-863, Washington, D.C., August 21-24, 2003.)

L'interface de commande respecte bien sûr tous les critères qui font le succès de SPAD : convivialité de l'interface, paramétrage par défaut modifiable, édition de rapport au format Excel ou Html, création et exploitation immédiate des données créées. Elle intègre également l'aspect métier, en permettant à l'utilisateur d'imposer des variables dans la sélection et d'influencer ainsi la sélection automatique. (Pour la méthode CFS)

## 6 - Statistiques de base

---

Totalement repensée, cette nouvelle méthode vous donnera accès en quelques clics à une vision globale de vos données.

Pour les variables nominales, vous obtiendrez les tris à plats avec de nombreuses options d'édition, en particulier la notion de tris groupés très utile pour comparer les distributions de modalités présentes dans différentes variables.

Pour les variables continues, dans des tableaux dont vous piloterez le design, vous obtiendrez les statistiques classiques : effectif, moyenne, écart type, minimum, maximum, médiane, quartiles, quantiles, coefficient de variation, somme, Kurtosis, Skewness pour ne citer que les principaux.

Pour des variables continues choisies, vous pourrez aussi les « discrétiser », c'est à dire obtenir les distributions de toutes les valeurs effectives, ou construire et éditer la matrice des corrélations et des p-values associées (Pearson et Spearman), avec une mise en évidence automatique des corrélations significatives par un jeu de couleurs.

Bien sûr, que ce soit pour les variables nominales ou pour les variables continues, vous disposerez à la demande, de graphiques associés dont vous piloterez les options selon vos besoins.

Cette méthode intègre bien sûr la possibilité de filtrer les individus (filtre logique, tirage aléatoire) et d'utiliser une variable de pondération.

Le plus grand soin a été apporté aux sorties qui seront automatiquement exportées sous Excel ou visualisables par votre éditeur HTML (Internet explorer, Firefox).

## 7 - Quelques Exemples

<b>Tris à plat</b>			
<b>Type de client</b>			
Modalités	Effectifs	Pourcentages	% sur exprimés
bon client	237	50,641	50,641
mauvais client	231	49,359	49,359
Ensemble	468	100,000	100,000
<b>Age du client</b>			
Modalités	Effectifs	Pourcentages	% sur exprimés
moins de 23 ans	88	18,803	18,803
de 23 à 40 ans	150	32,051	32,051
de 40 à 50 ans	122	26,068	26,068
plus de 50 ans	108	23,077	23,077
Ensemble	468	100,000	100,000

Figure 1 : Tri à plat exporté dans un tableur

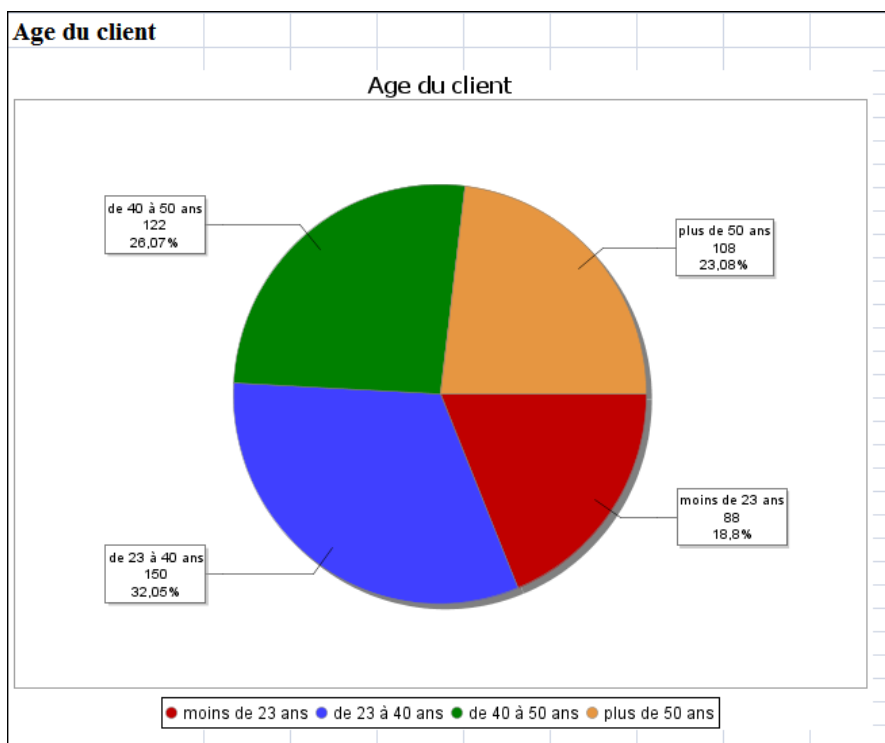


Figure 2 : Export sous un tableur d'une discrétisation simple

Statistiques sur variables continues							
Tableau							
Variabile	Moyenne	Ecart-type (N-1)	Minimum	Maximum	Médiane	Kurtosis	Skewness
Cylindre	1906,125	527,909	1116,000	2986,000	1972,500	-0,307	0,495
Puissance	113,667	38,784	50,000	188,000	107,500	-0,538	0,592
Vitesse	183,083	25,215	135,000	226,000	183,000	-0,365	-0,042
Poids	1123,333	248,433	730,000	1600,000	1162,500	-1,016	0,095
Longueur	421,583	41,340	350,000	473,000	437,500	-1,322	-0,430
Largeur	168,833	7,654	155,000	184,000	169,500	-0,490	-0,126

Figure 3 : Tableau statistique sur des variables continues

Matrice des corrélations						
Matrice des corrélations de Pearson						
Variables	Cylindre	Puissance	Vitesse	Poids	Longueur	Largeur
Cylindre	1,000					
Puissance	0,861	1,000				
Vitesse	0,693	0,894	1,000			
Poids	0,897	0,769	0,507	1,000		
Longueur	0,864	0,689	0,532	0,863	1,000	
Largeur	0,709	0,552	0,363	0,700	0,864	1,000

Matrice des p-values						
Variables	Cylindre	Puissance	Vitesse	Poids	Longueur	Largeur
Cylindre	0,000					
Puissance	0,000	0,000				
Vitesse	0,000	0,000	0,000			
Poids	0,000	0,000	0,011	0,000		
Longueur	0,000	0,000	0,007	0,000	0,000	
Largeur	0,000	0,005	0,081	0,000	0,000	0,000

Figure 4 : Matrice des corrélations et des p-values exporté dans un tableur

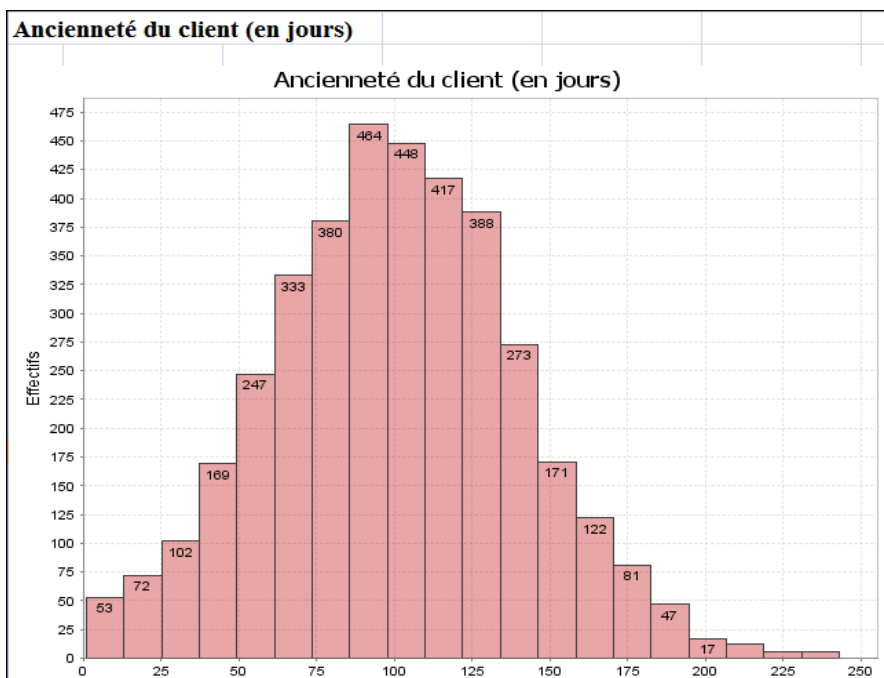


Figure 5 : Une des nombreuses représentations graphiques de Coheris SPAD