LES ANALYSES EN CORRESPONDANCES MULTIPLES DANS SAS® ENTERPRISE GUIDE® 4.3

Le public francophone demande souvent s'il possible de faire une **ACM** (**A**nalyse en **C**orrespondances **M**ultiples) avec SAS Enterprise Guide. Ce type d'analyse n'existe pas en mode standard, dans une tâche. Néanmoins, une tâche expérimentale et téléchargeable sur notre site Internet peut être installée dans Enterprise Guide 4.3. Elle permet de réaliser ce type d'étude statistique très facilement en quelques clics.

Caractéristiques :

Catégorie : Enterprise Guide
OS : Windows
Version : 4.3
Vérifié en mars 2012

Nous allons débuter par une présentation rapide du concept de l'ACM. Ensuite nous expliquerons comment installer l'add-in pour réaliser ce type d'analyse dans Enterprise Guide. Enfin, nous illustrerons son fonctionnement au travers de deux exemples d'applications, et nous étudierons le code SAS qui s'exécute en arrière-plan.

Qu'est-ce qu'une ACM ?

L'ACM est une technique d'analyse de données, faisant partie de la famille de l'AFC (Analyse Factorielle des Correspondances), qui s'appuie sur des données catégorielles nominales. Les données peuvent se présenter sous forme d'un tableau de contingence : des variables en lignes, des variables en colonnes et les effectifs correspondants dans le tableau (par exemple, CSP versus lieu de vacances, ou encore lieu de résidence versus décennie) ou d'un tableau disjonctif complet (tableau de Burt), entre autres. Le premier type de tableau sera retenu pour l'étude qui va suivre. Généralement, les modalités dont le poids est inférieur à 5% sont regroupées avec d'autres (significativité statistique). Les objectifs de l'ACM sont de mettre en évidence:
- les relations entre les différentes modalités des variables,
- hypothétiquement, les relations entre les individus statistiques.

Pour illustrer la partie plus théorique qui va suivre, cette section s'appuie sur le cours de Thierry FOUCART (maître de conférences à la retraite de l'Université de Poitiers) qui étudie le comportement des fumeurs par rapport à leur sexe et leur majorité.

Exemple de tableau de contingence:

X : 3 modalités	Y : 4 modalités
brunes : fumeur de brunes	mm : mineur masculin
blondes : fumeur de blondes	mf : mineur féminin
non fumeur	MF majeur féminin
	MM : majeur masculin

		Y			
		mm	mf	MF	MM
X	brunes	63	37	41	47
	blondes	36	55	39	38
	non fumeur	34	27	72	38

Dans ce type d'analyse, on cherche les différences entre profil de lignes (ou colonnes, valable pour toutes les références suivantes au profil de lignes) plutôt que des différences individuelles.

Un profil de ligne est la répartition suivant les modalités en colonnes des individus données en lignes.

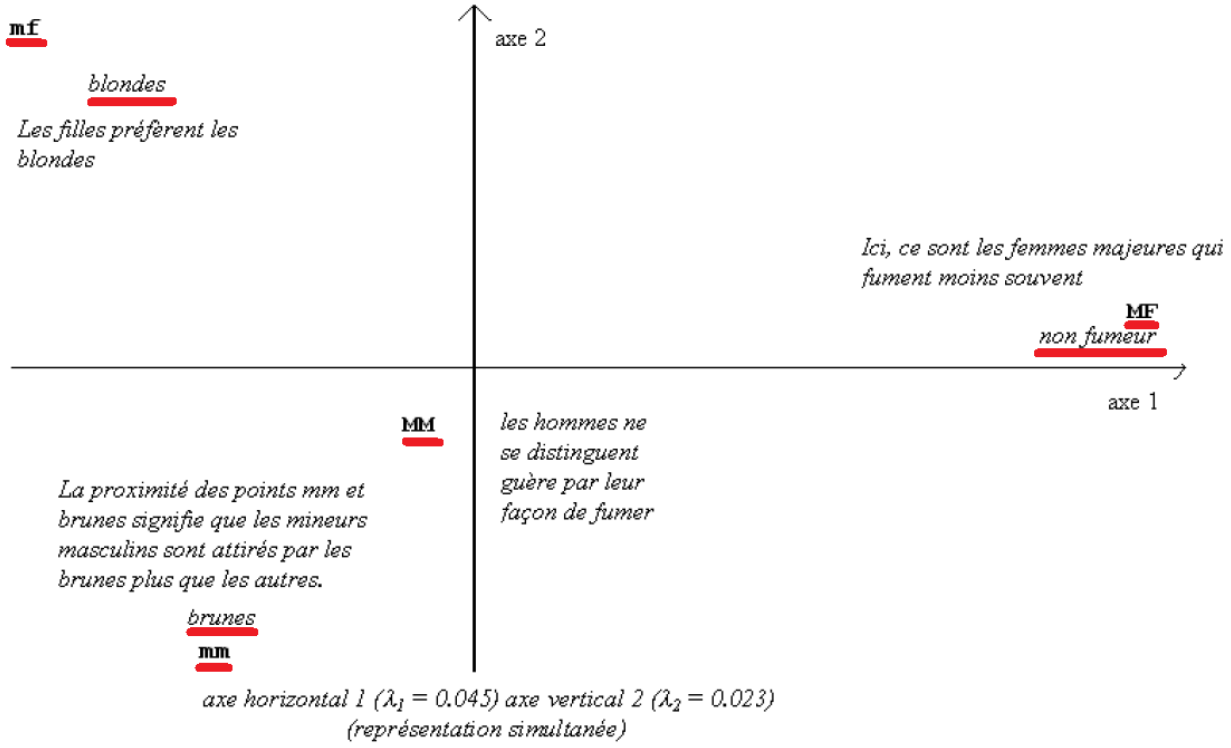
	mm	mf	MF	MM	total
profil brunes	0.335	0.197	0.218	0.250	1
profil blondes	0.214	0.327	0.232	0.226	1
profil non fumeur	0.199	0.158	0.421	0.222	1
centre de gravité P_J	0.252	0.226	0.288	0.233	1

profils lignes P_J^i

Une notion importante est $P_{.j}$, par exemple 0,252 pour « mm » qui représente la part des mineurs mâles dans l'échantillon, ce qui correspondant aux profils colonnes (sans tenir compte de leur comportement vis-à-vis de la cigarette). Comme il a été dit plus haut, le calcul des différences se fait entre les profils. Ainsi, la différence entre deux profils de lignes s'écrit :

$$d^2(i,i') = \sum_{j=1}^{q} [\, p_j^i - p_j^{i'}]^2 / p_{\bullet j}$$

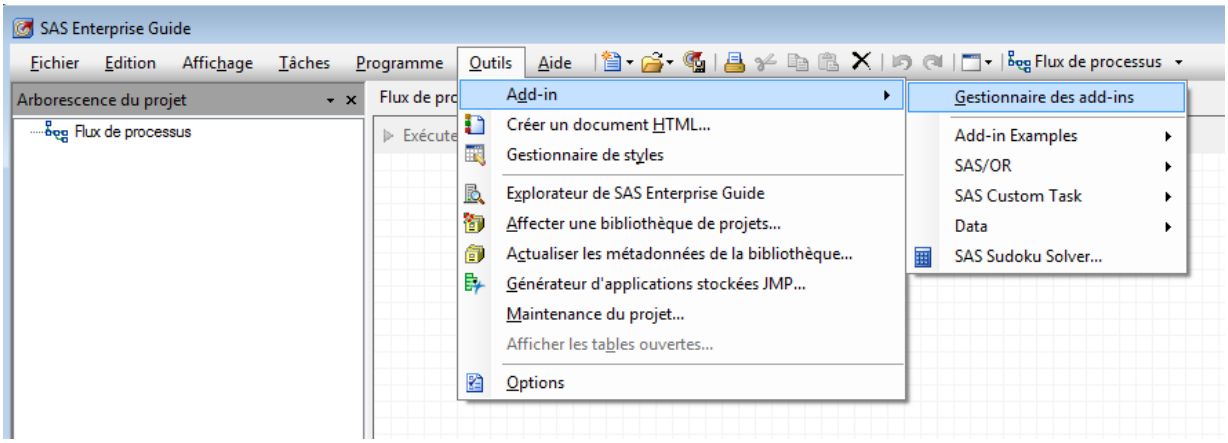
Les associations entre variables vont ensuite être découvertes en calculant la distance du chi-deux (présentée ci-dessus). Ces associations sont représentées graphiquement pour faciliter l'interprétation des structures dans les données. Les oppositions entre lignes et colonnes sont ensuite maximisées, pour découvrir les dimensions sous-jacentes les plus aptes à décrire les oppositions centrales dans les données. Comme en AFC ou en ACP, les axes d'analyses sont classés par ordre d'importance (en fonction de la quantité de variance, inertie du nuage, expliquée). Pour déterminer le nombre d'axes qu'il est pertinent d'étudier, plusieurs règles existent dans la littérature (100%/nombre d'axes=seuil de pertinence d'un axe). Une fois les axes pertinents choisis, il ne reste plus qu'à les analyser, au travers de différents critères (contributions, cosinus,…) qui ne seront pas détaillés dans ce papier. Voici le type de graphique, et quelques conclusions, qui sont obtenus :



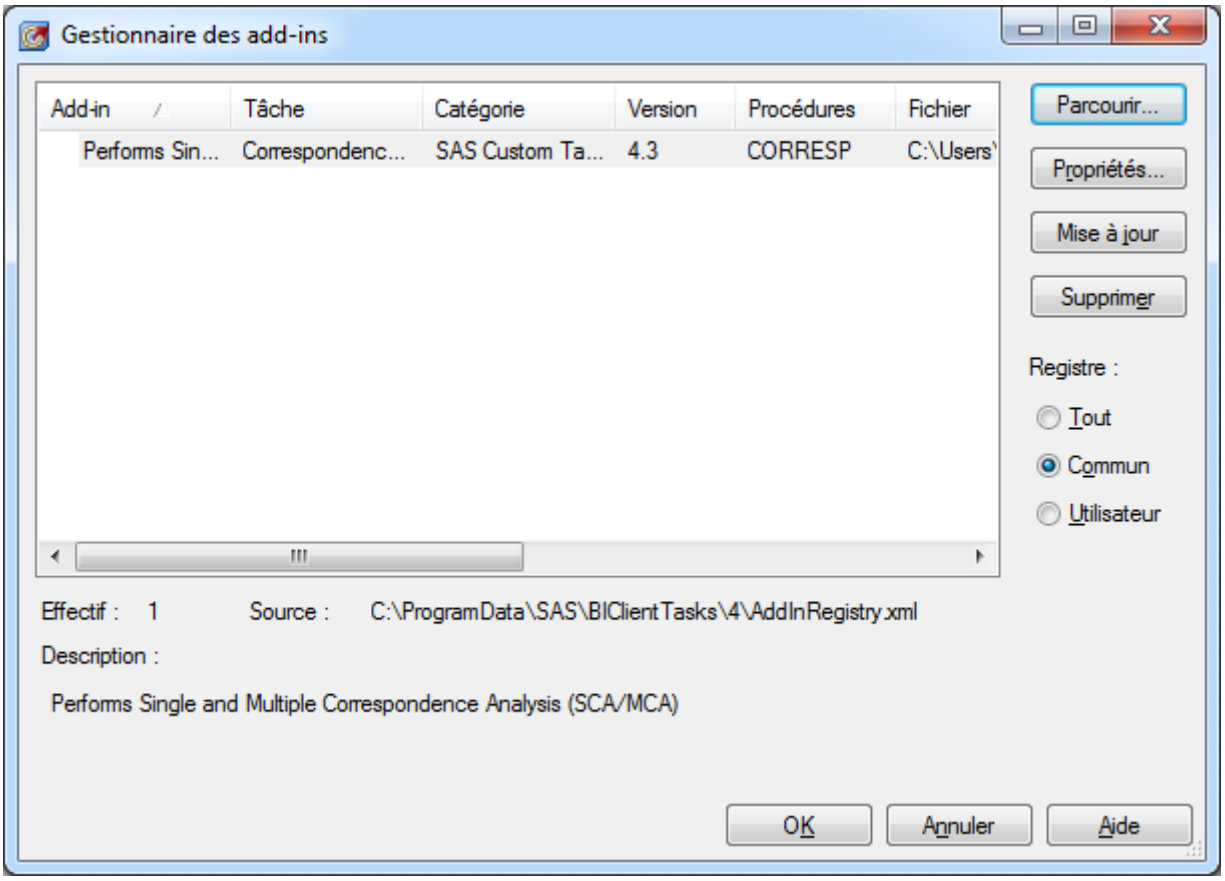
Installation de l'Add-in

Pour réaliser une Analyse en Correspondances Multiples, il vous faudra installer un Add-in (cette add-in, comme tous les add-in, n'est ni traduit, ni supporté, à cause de son statut expérimental), Correspondence Analysis, développé par Audimar Bangi (développeur SAS). La manipulation est simple : il suffit de télécharger le fichier zip en cliquant sur ce lien. Vous y trouverez un fichier texte, deux tables d'exemple et un fichier .dll. La procédure d'installation est la suivante :

1. Copier le fichier SAS.Tasks.CorrespondenceAnalysis.dll dans le répertoire de votre choix, par exemple C:\addinEG
2. Démarrer Enterprise Guide **4.3** (attention : un add-in est spécifique à une version donnée)
3. Cliquer sur « Outils »→ « Add-in » → « Gestionnaire des Add-in »



Cliquer sur « Parcourir », sélectionner le fichier .dll et cliquer sur « OK ».



L'add-in est ainsi installé et activé. Il est prêt à être utilisé, à partir du menu « Outils » → « Add-in » → « Correspondence Analysis… » comme nous allons le voir dans les exemples qui vont suivre.

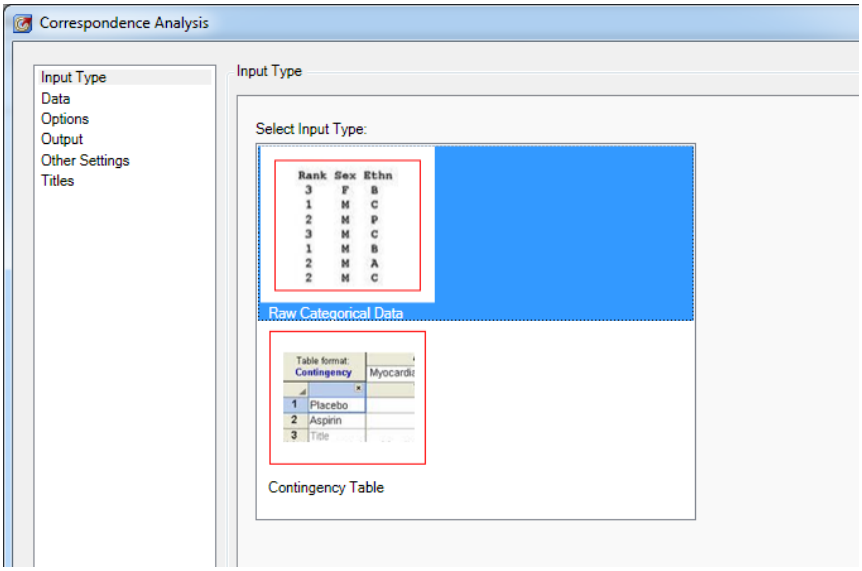
Exemples d'application de l'Add-in d'Analyse en Correspondances Multiples

Nous allons maintenant explorer quelques fonctionnalités offertes par cet outil au travers de plusieurs exemples d'applications, se basant sur les tables d'exemple téléchargées avec l'add-in).

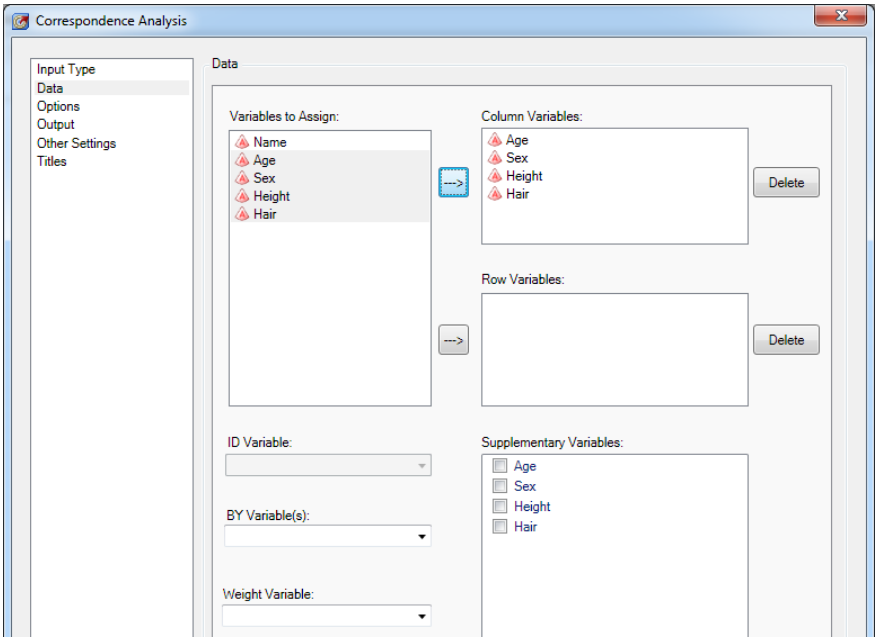
Table avec des variables de classes

La première application se base sur la table categoricaldata.

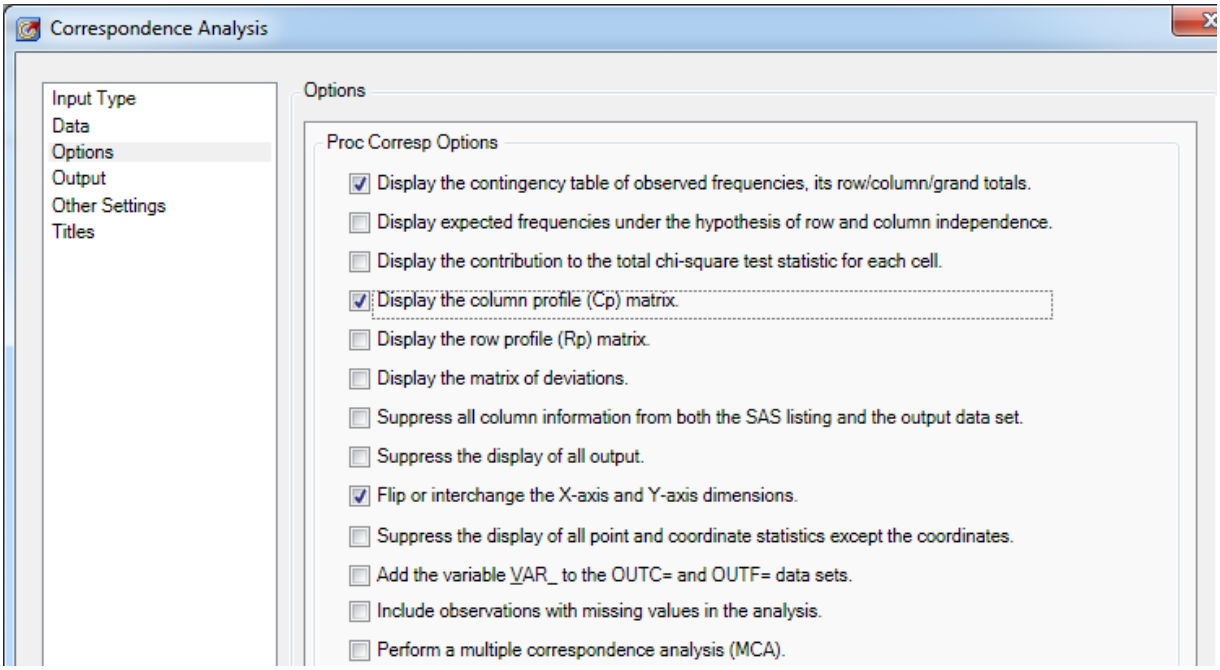
- L'ouvrir dans SAS Enterprise Guide 4.3.
- Sélectionner la table et lancer l'analyse grâce au menu (« Outils » → « Add-in » → « Correspondence Analysis… ») et commencer par sélectionner le type d'entrée, par exemple « Raw Categorical Data » (Données catégorielles brutes).



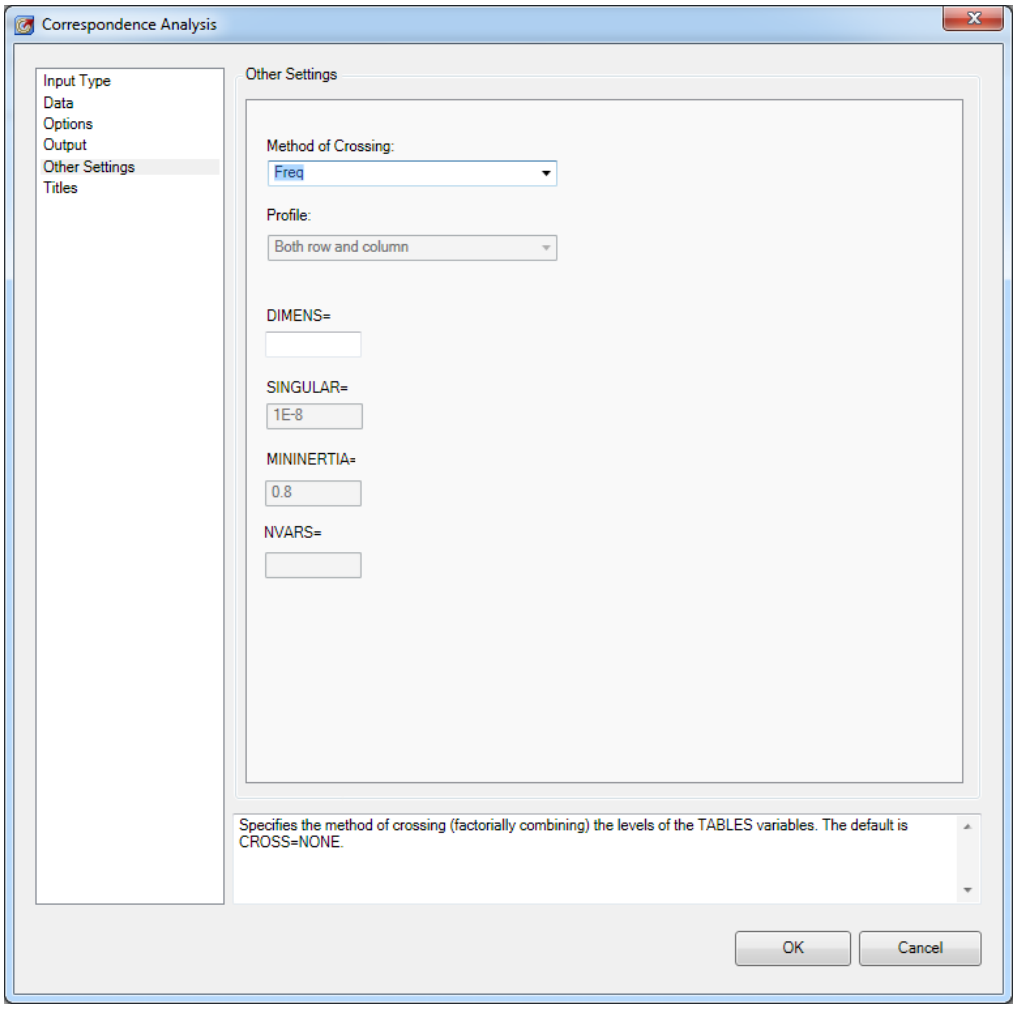
L'exemple s'appuie sur les variables Age, Sex, Height et Hair.



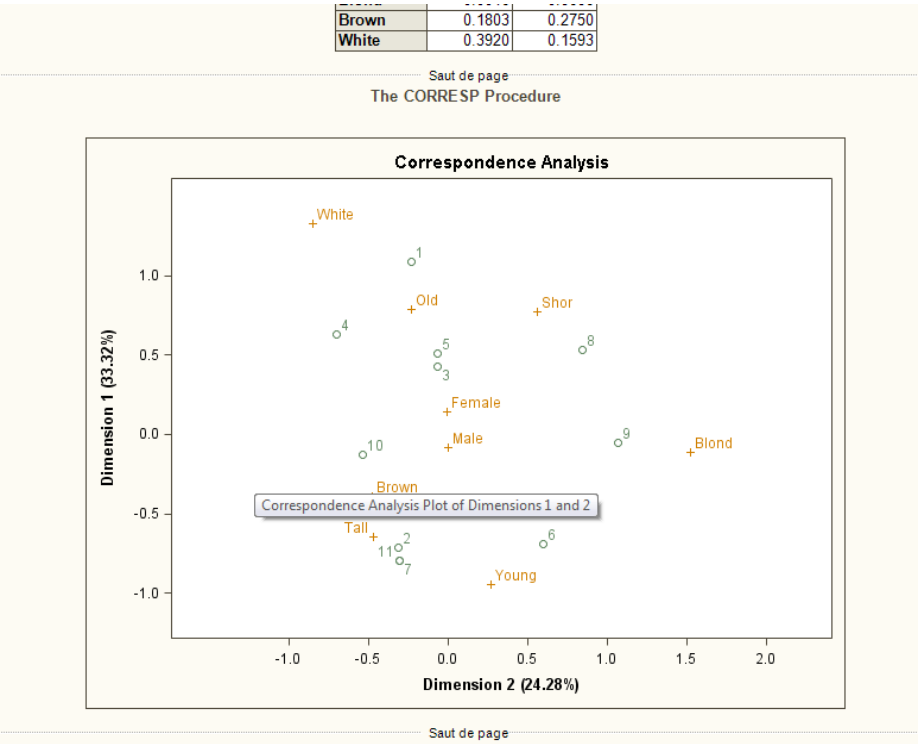
Dans l'onglet "Options", choisir les trois options suivantes (dont voici l'équivalence en Français) :
- affichage de la table de contingence des fréquences observées, avec les totaux en lignes, colonnes et global.
- affichage de la matrice de profil de colonne (Cp).
- retourne ou échange les dimensions des axes X et Y.



Dans l'onglet « Other Settings », choisir « Freq » comme méthode de croisement (« Method of Crossing ») et cliquer sur « OK » pour lancer l'analyse.



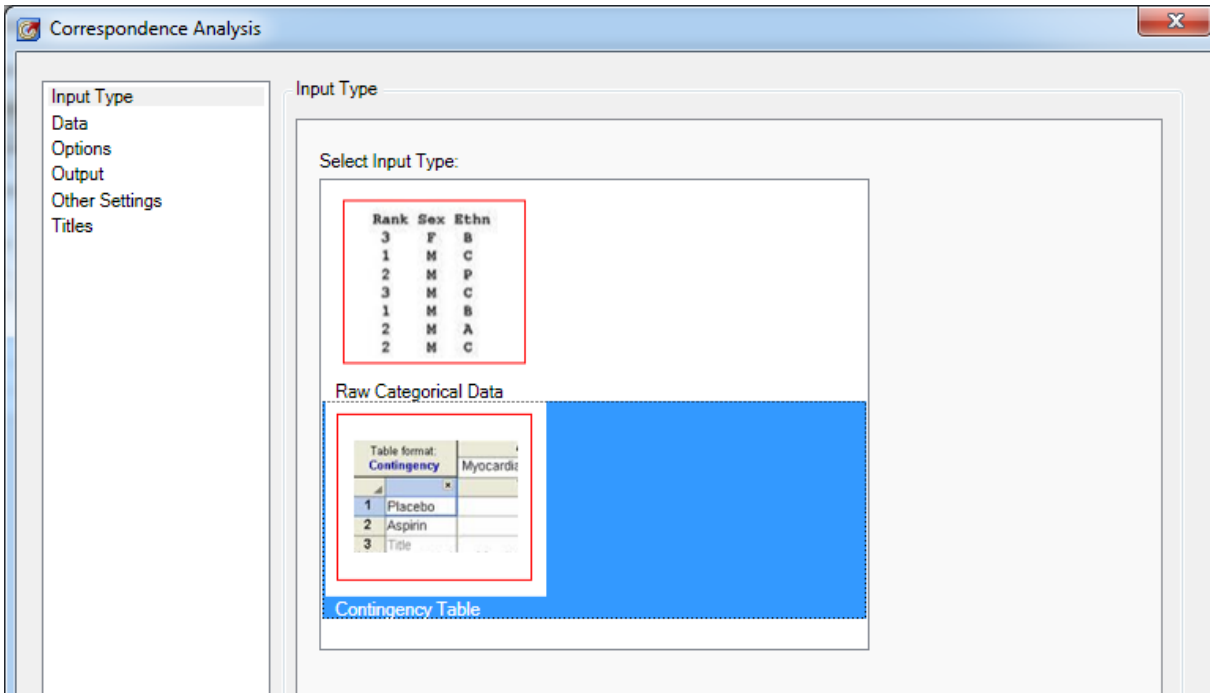
Vous trouverez ci-dessous, un extrait des résultats obtenus en sortie de cette analyse, représentant nos individus et les différentes modalités de nos variables sur un même graphique, où les deux premiers axes sont les deux plus représentatifs.



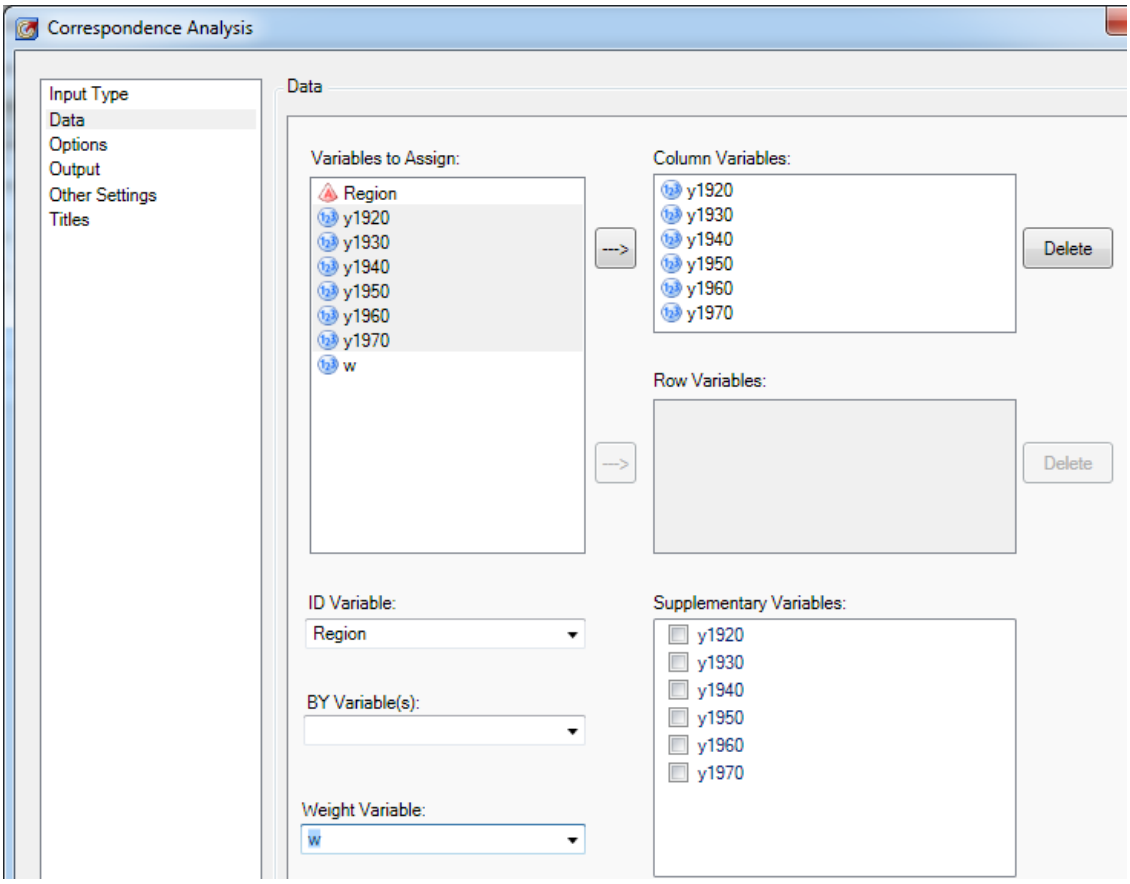
On note sur le graphique ci-dessus, une sur-représentation des personnes agées (OLD) avec des cheveux blancs chez les individus classés dans les groupes 1 et 4 alors que le groupe 9 contient une sur-réprésentation par rapport à la moyenne des blonds.

Application avec une table de contingence

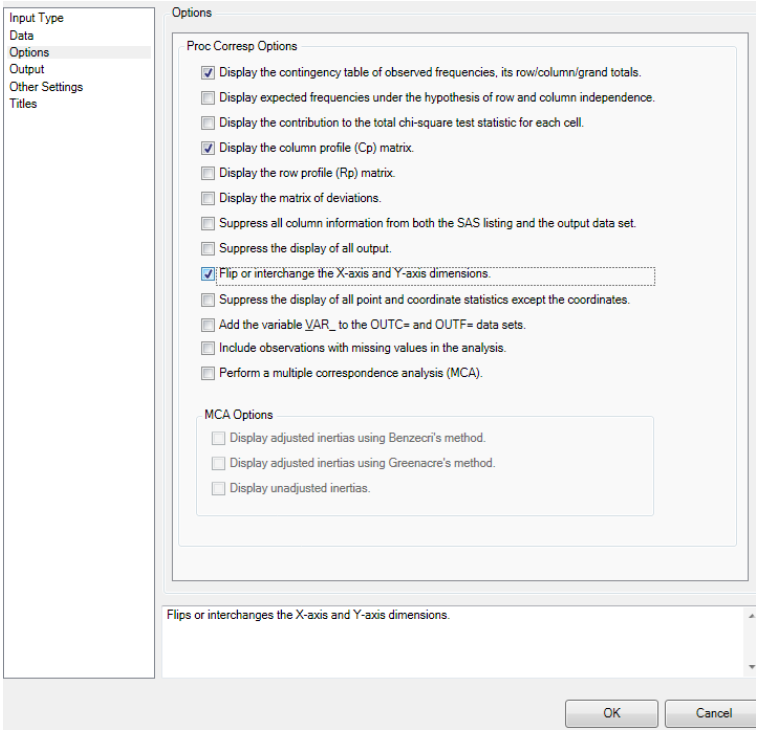
Pour cette seconde analyse, la table contingencytable.sas7bdat sera la table d'étude. Choisir « Table de Contingence » (« Contingency Table ») comme type d'entrée.



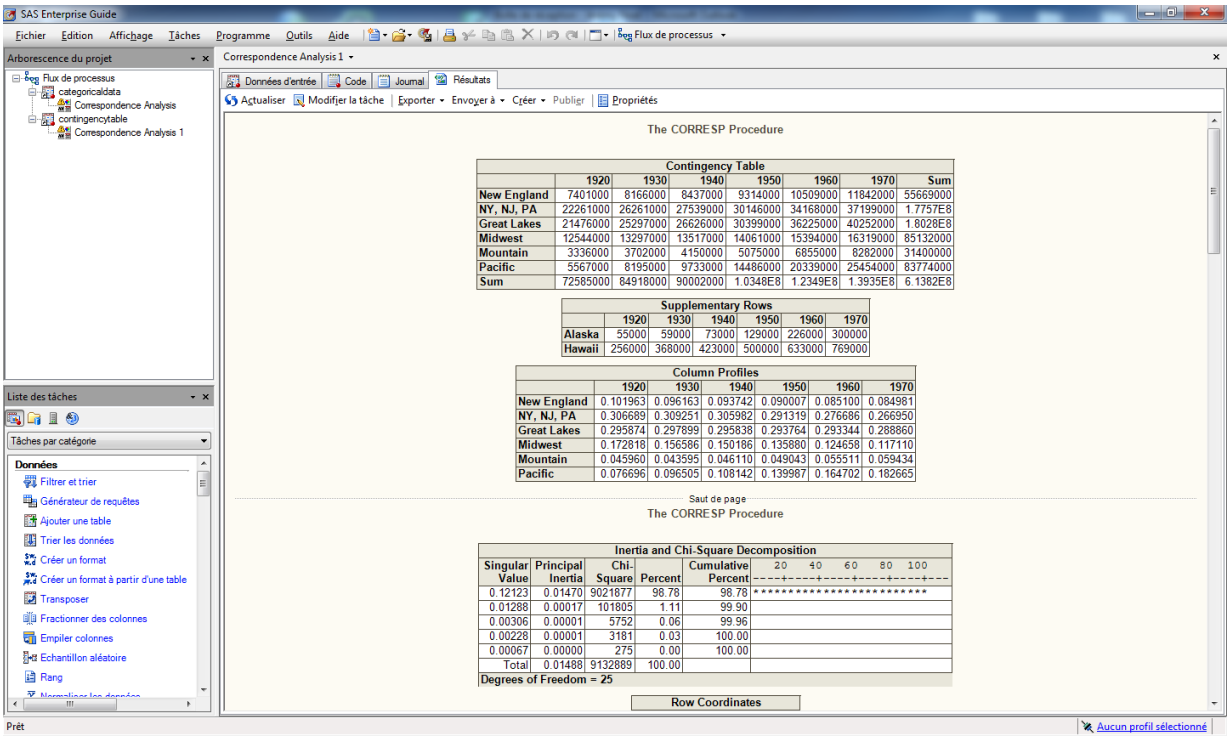
Pour cette analyse, sélectionner les six variables du type y19xx comme variables colonnes, « Region » comme « ID variable » (variable d'identification) et w en tant que variable de pondération (« Weight Variable »).



Le paramétrage des options est identique à celui de la première analyse. Cliquer sur « OK ».



Les deux captures d'écrans ci-dessous montrent une partie des résultats (le début et la fin). Ils mettent en évidence que la quasi-totalité de l'inertie du nuage (98.78%) est captée par le premier axe d'analyse, que ce soit à travers le tableau ou le graphique.



The CORRESP Procedure

Contingency Table

	1920	1930	1940	1950	1960	1970	Sum
New England	7401000	8166000	8437000	9314000	10509000	11842000	55669000
NY, NJ, PA	22261000	26261000	27539000	30146000	34168000	37199000	1.7757E8
Great Lakes	21476000	25297000	26626000	30399000	36225000	40252000	1.8028E8
Midwest	12544000	13297000	13517000	14061000	15394000	16319000	85132000
Mountain	3336000	3702000	4150000	5075000	6855000	8282000	31400000
Pacific	5567000	8195000	9733000	14486000	20339000	25454000	83774000
Sum	72585000	84918000	90002000	1.0348E8	1.2349E8	1.3935E8	6.1382E8

Supplementary Rows

	1920	1930	1940	1950	1960	1970
Alaska	55000	59000	73000	129000	226000	300000
Hawaii	256000	368000	423000	500000	633000	769000

Column Profiles

	1920	1930	1940	1950	1960	1970
New England	0.101963	0.096163	0.093742	0.090007	0.085100	0.084981
NY, NJ, PA	0.306689	0.309251	0.305982	0.291319	0.276686	0.266950
Great Lakes	0.295874	0.297899	0.295838	0.293764	0.293344	0.288860
Midwest	0.172818	0.156586	0.150186	0.135880	0.124658	0.117110
Mountain	0.045960	0.043595	0.046110	0.049043	0.055511	0.059434
Pacific	0.076696	0.096505	0.108142	0.139987	0.164702	0.182665

Saut de page

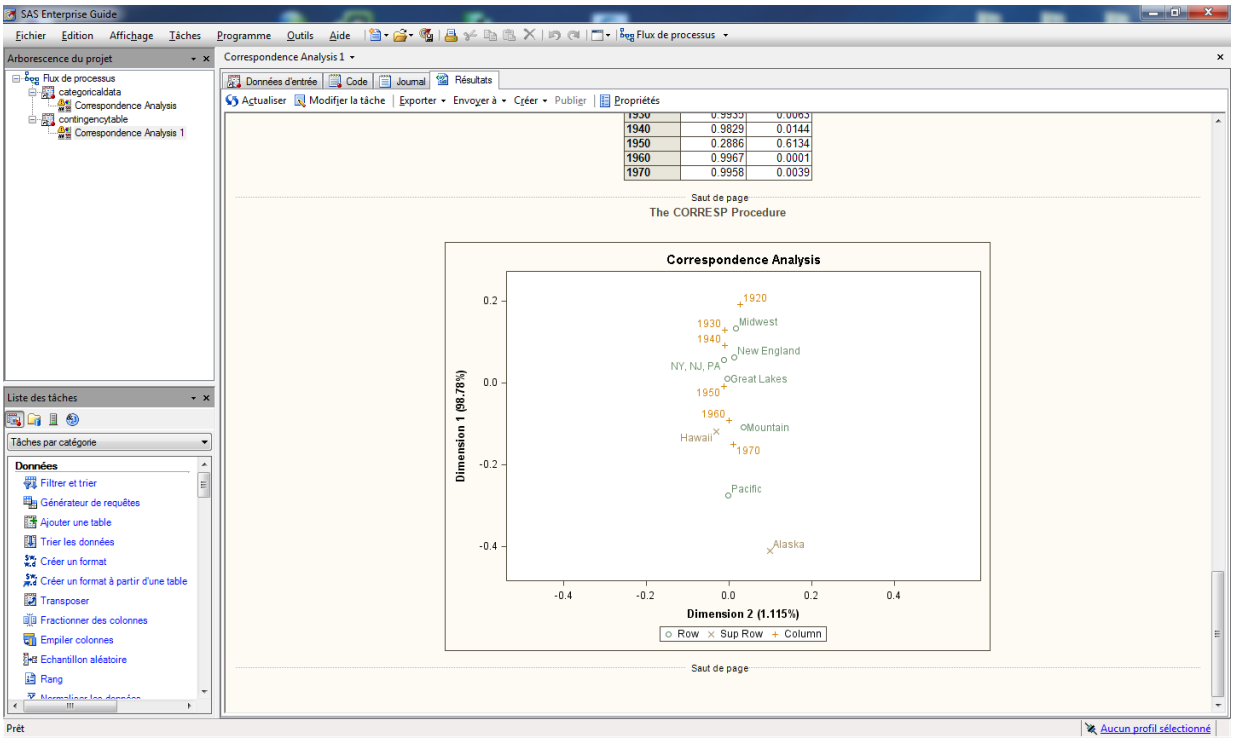
The CORRESP Procedure

Inertia and Chi-Square Decomposition

Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent	20 40 60 80 100 ----+----+----+----+----+----
0.12123	0.01470	9021877	98.78	98.78	**********
0.01288	0.00017	101805	1.11	99.90	
0.00306	0.00001	5752	0.06	99.96	
0.00228	0.00001	3181	0.03	100.00	
0.00067	0.00000	275	0.00	100.00	
Total	0.01488	9132889	100.00		

Degrees of Freedom = 25

Row Coordinates



1940	0.9829	0.0144
1950	0.2886	0.6134
1960	0.9967	0.0001
1970	0.9958	0.0039

Saut de page

The CORRESP Procedure

Saut de page

Le code SAS traditionnel

Dans ce nouvel Add-in, comme derrière chacune des tâches d'Enterprise Guide, du code SAS est créé et compilé à chaque exécution. La procédure utilisée pour la mise en œuvre de cette analyse est la PROC CORRESP, dont la syntaxe va être détaillée dans cette section. La structure générale est la suivante :

```
PROC CORRESP <options> ;
        TABLES <row-variables,> column-variables ;
        VAR variables ;
        BY variables ;
        ID variable ;
        SUPPLEMENTARY variables ;
        WEIGHT variable ;
        run ;
```

Attention les instructions TABLES et VAR sont ressemblantes, il faut en spécifier l'une ou l'autre mais en aucun cas les deux en même temps.
Dans l'instruction PROC CORRESP, en plus de la table d'entrée et des éventuelles tables de sorties (coordonnées et fréquences), il faut, si l'on souhaite faire une ACM, rajouter l'option MCA. Cette option requiert la présence d'un tableau de Burt en entrée. Il en ressort que ces études très francophones sont reprises dans SAS, au travers d'options, puisque ce sont des analyses plus proches de l'exploratoire que du protocole (raison de la présence des ACM dans JMP® ou SAS/IML® Studio, plutôt qu'en standard dans SAS Enterprise Guide).

Les deux exemples d'applications vues précédemment utilisent les mêmes options qui sont :

- observed : affiche une table de contingence des fréquences observées
- cp : affiche la matrice des profils colonnes
- binary: permet la création d'une table binaire facilement
- outc : spécifie le nom de la table de sortie contenant les coordonnées
- outf: spécifie le nom de la table de sortie contenant les fréquences, profils de lignes,…
- print : permet de choisir entre l'affichage en pourcentage ou en fréquence (effectifs, par défaut)
- Plots(flip)=all: produit tous les graphiques appropriés

Vous pouvez vous reporter à l'aide en ligne sur cette procédure pour plus d'information.

Conclusion

Il est désormais possible de faire une ACM (Analyse en Correspondances Multiples) en mode clic-bouton sous SAS Enterprise Guide, grâce à l'Add-in expérimental que nous venons de présenter. Cependant, il ne faut pas oublier que derrière ce module implémenté, et les fenêtres de SAS Enterprise Guide qui s'en chargent (onglet « Code », « Journal » et « Résultats »), du code SAS tourne en arrière plan avec toutes les spécificités que cela engendre (personnalisation, imbriquation,…).

Jérémy NOEL
Consultant Support Clients SAS France