

1 Objectif

Statistiques avec [WinIDAMS 1.3](#).

WinIDAMS (Internationally Developed **D**ata **A**nalysis and **M**anagement **S**oftware Package) est un logiciel de statistique développée sous l'égide de l'UNESCO. Le projet prend ses sources dans les années 70. Mais la première mouture réellement estampillée IDAMS date de la fin des années 80. Deux versions sont développées en parallèle : l'une pour les ordinateurs IBM Mainframe, l'autre pour les PC sous MS-DOS¹. L'idée est de fédérer (comme Roger du même nom) les spécialistes de différents pays pour développer un outil qui exprime la quintessence du savoir statistique. J'avoue avoir eu le vertige lorsque j'ai consulté pour la première fois la liste des contributeurs. Cornaqué par un tel aréopage d'experts internationaux, l'outil devrait présenter de très grandes qualités.

Ce tutoriel décrit la mise en œuvre de WinIDAMS sur un fichier exemple. Nous porterons une attention particulière à l'importation des données car le logiciel procède de manière assez singulière. Puis, nous effectuerons une rapide découverte de quelques méthodes exploratoires en précisant pour chacune d'elles le paramétrage et la lecture (d'une partie) des résultats. Nous mettrons en parallèle les sorties d'autres logiciels tels que Tanagra et SAS.

2 Données

Notre fichier provient du serveur [DASL](#) de StatLib. Il décrit la composition de $n = 26$ poteries collectés sur 4 sites différents. Voici les 5 premières observations de la base au format Excel.

ID	Al	Fe	Mg	Ca	Na	Site	SiteNum
1	14.4	7	4.3	0.15	0.51	L	1
2	13.8	7.08	3.43	0.12	0.17	L	1
3	14.6	7.09	3.88	0.13	0.2	L	1
4	11.5	6.37	5.64	0.16	0.14	L	1
5	13.8	7.06	5.34	0.2	0.2	L	1

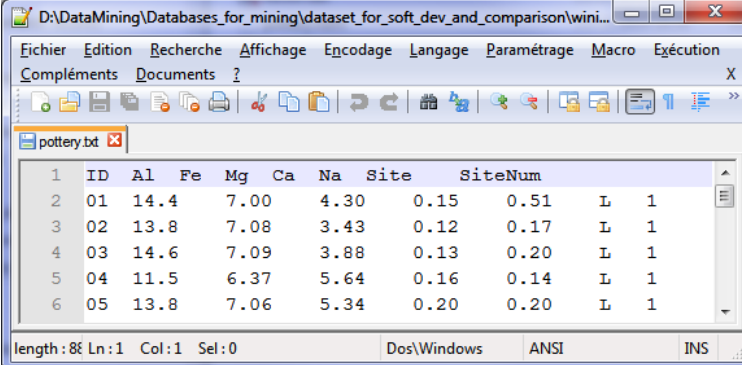
ID est un identifiant. La variable SITE = {L, C, I, A} a été recodée en numérique SITENUM = {1, 2, 3, 4} pour que WinIDAMS puisse le manipuler. Un tel recodage semble possible avec les commandes internes du logiciel. J'avoue être allé au plus simple en procédant en amont.

3 Importation des données

Transformation en fichier texte (.txt). WinIDAMS ne sait pas importer directement les fichiers au format Excel (.xls ou .xlsx). Il nous faut le transformer en fichier texte avec

¹ UNESCO, « [WinIDAMS Reference Manual \(release 1.3\)](#) », Avril 2008.

séparateur tabulation (.txt) en actionnant l'option FICHIER / ENREGISTRER SOUS / TYPE : Texte (séparateur : tabulation) (*.txt) sous Excel. Voici les 5 premières lignes de la base chargée dans un éditeur de texte.

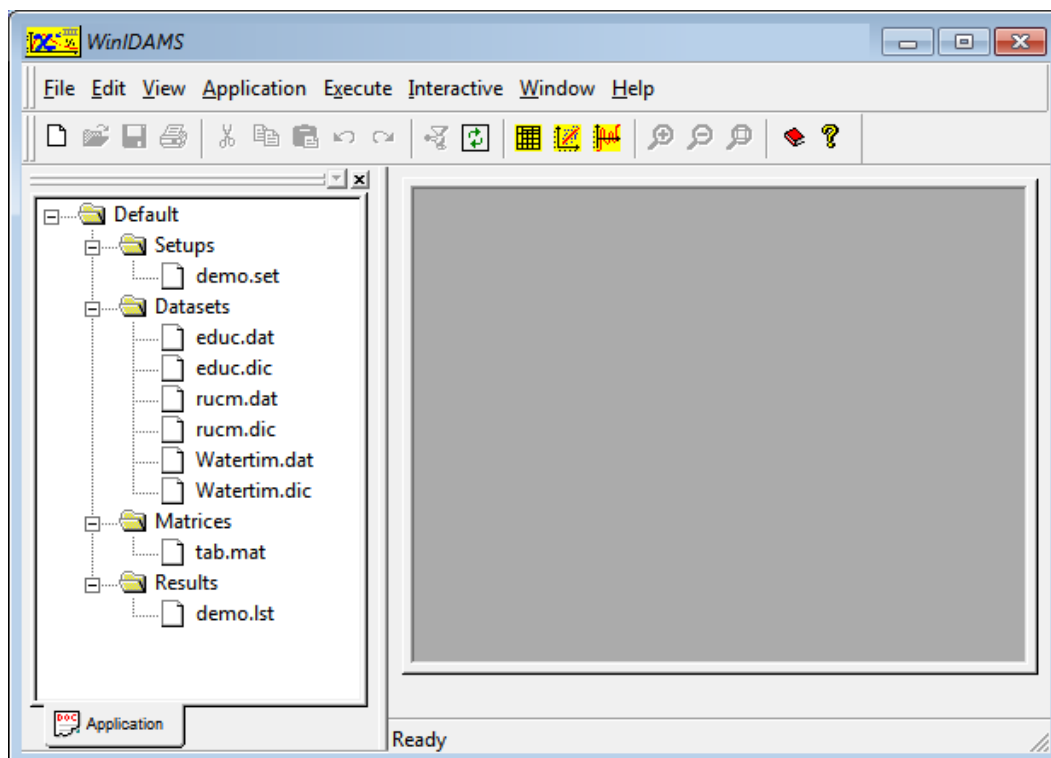


The screenshot shows a text editor window with the following content:

	ID	Al	Fe	Mg	Ca	Na	Site	SiteNum		
1	01	14.4		7.00		4.30	0.15	0.51	L	1
2	02	13.8		7.08		3.43	0.12	0.17	L	1
3	03	14.6		7.09		3.88	0.13	0.20	L	1
4	04	11.5		6.37		5.64	0.16	0.14	L	1
5	05	13.8		7.06		5.34	0.20	0.20	L	1

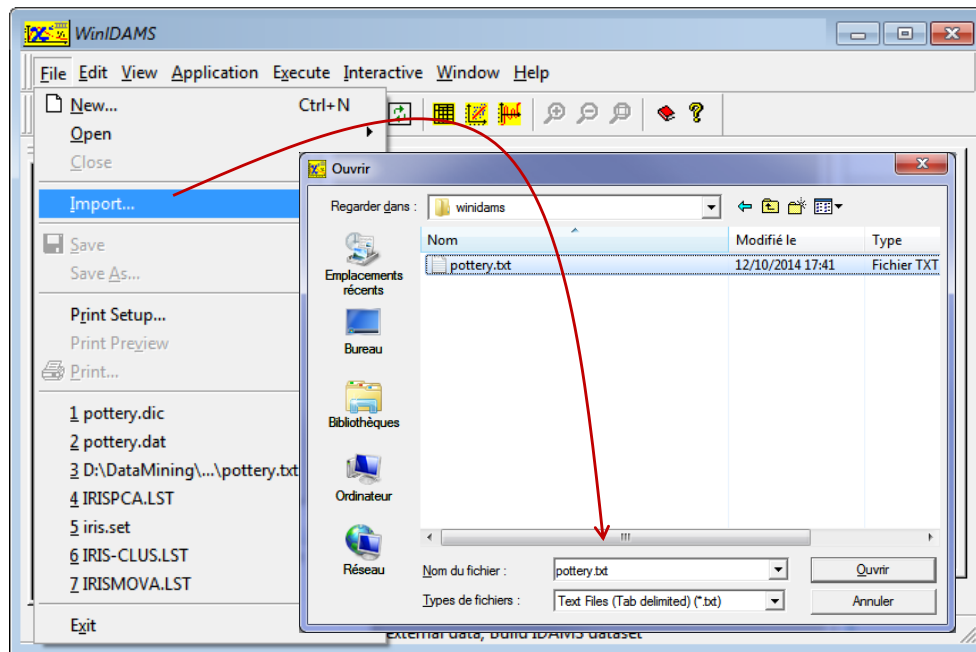
Remarque : Le décalage entre les noms de variables et les colonnes de valeurs n'est qu'apparent.

Importation du fichier « pottery.txt ». Au démarrage du logiciel, la fenêtre principale se présente comme suit². Nous observons pour l'instant les fichiers exemples livrés avec WinIDAMS dans l'onglet « Application ».

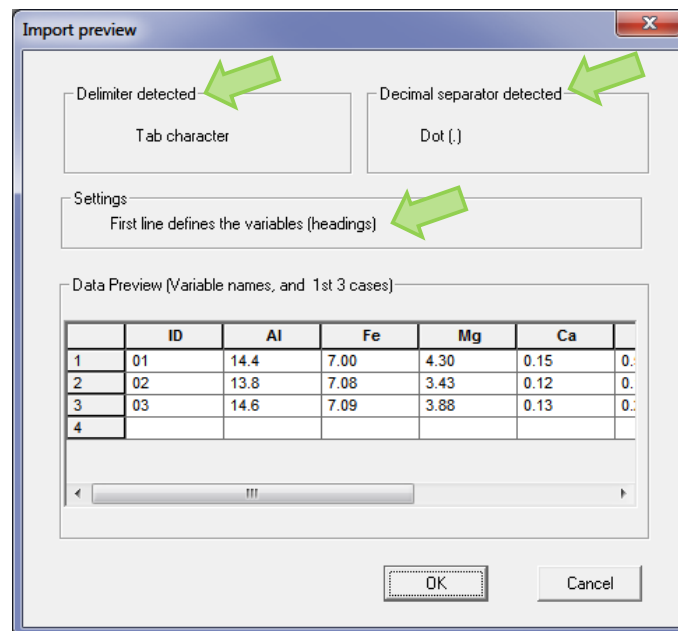


Pour accéder au fichier « pottery.txt », nous actionnons le menu FILE / IMPORT

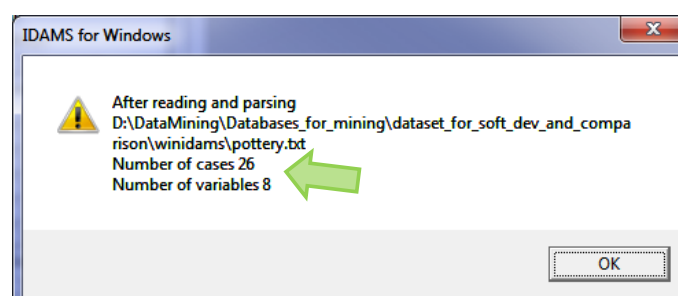
² Une version française est disponible. J'ai choisi le logiciel en anglais pour les copies d'écran parce que ce tutoriel sera par la suite traduit pour le site <http://data-mining-tutorials.blogspot.fr/>



Une boîte de dialogue permet de vérifier les paramètres d'importation (séparateur de colonnes, point décimal, noms des variables).



Nous cliquons sur OK. Un panneau indique les caractéristiques de la base traitée (nombre d'observations et de variables).



Les données apparaissent dans la grille d'affichage « External Data ».

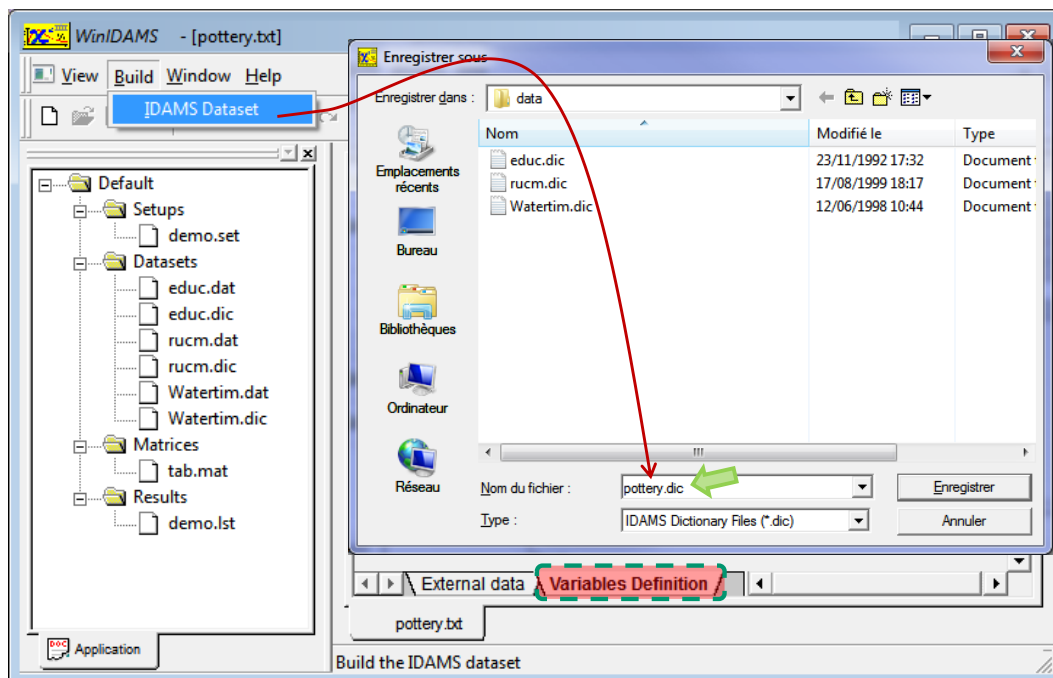
	ID	Al	Fe	Mg	Ca	Na	Site	SiteNum
1	01	14.4	7.00	4.30	0.15	0.51	L	1
2	02	13.8	7.08	3.43	0.12	0.17	L	1
3	03	14.6	7.09	3.88	0.13	0.20	L	1
4	04	11.5	6.37	5.64	0.16	0.14	L	1
5	05	13.8	7.06	5.34	0.20	0.20	L	1
6	06	10.9	6.26	3.47	0.17	0.22	L	1
7	07	10.1	4.26	4.26	0.20	0.18	L	1
8	08	11.6	5.78	5.91	0.18	0.16	L	1
9	09	11.1	5.49	4.52	0.29	0.30	L	1
10	10	13.4	6.92	7.23	0.28	0.20	L	1
11	11	12.4	6.13	5.69	0.22	0.54	L	1
12	12	13.1	6.64	5.51	0.31	0.24	L	1
13	13	12.7	6.69	4.45	0.20	0.22	L	1
14	14	12.5	6.44	3.94	0.22	0.23	L	1
15	15	11.8	5.44	3.94	0.30	0.04	C	2
16	16	11.6	5.39	3.77	0.29	0.06	C	2
17	17	18.3	1.28	0.67	0.03	0.03	I	3
18	18	15.8	2.39	0.63	0.01	0.04	I	3
19	19	18.0	1.5	0.67	0.01	0.06	I	3
20	20	18.0	1.88	0.68	0.01	0.04	I	3
21	21	20.8	1.51	0.72	0.07	0.10	I	3
22	22	17.7	1.12	0.56	0.06	0.06	A	4
23	23	18.3	1.14	0.67	0.06	0.05	A	4
24	24	16.7	0.92	0.53	0.01	0.05	A	4
25	25	14.8	2.74	0.67	0.03	0.05	A	4
26	26	19.1	1.64	0.60	0.10	0.03	A	4

Le dictionnaire de données apparaît dans « Variables Definition ».

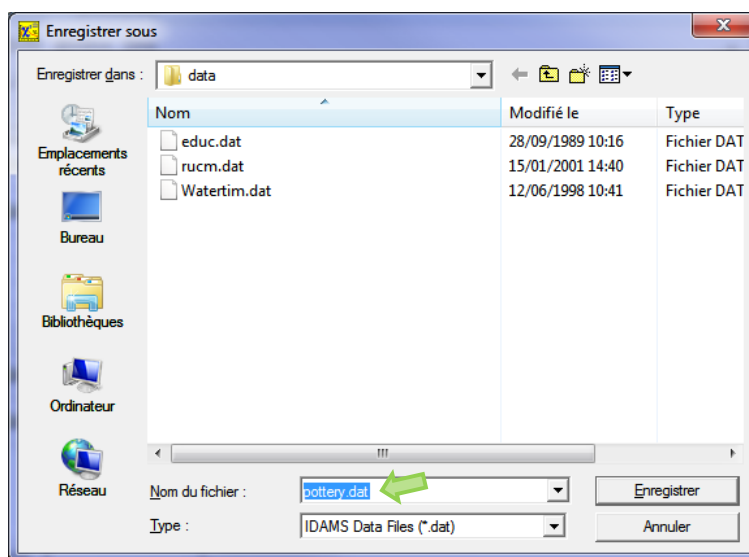
	Description	Type	MaxWidht	NumD	Md1	Md2	Recodin
1	ID	Numeric	2	0			
2	Al	Numeric	4	1			
3	Fe	Numeric	4	2			
4	Mg	Numeric	4	2			
5	Ca	Numeric	4	2			
6	Na	Numeric	4	2			
7	Site	Alphabetic	1				
8	SiteNum	Numeric	1	0			

Toutes les colonnes sont numériques, mis à part « Site ». WinIDAMS sait gérer les éventuelles valeurs manquantes.

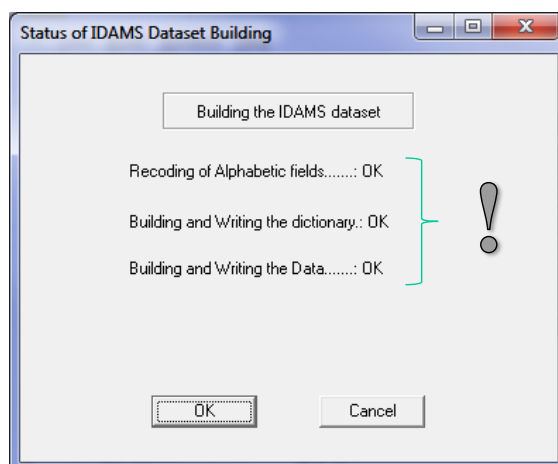
Construction d'une base WinIDAMS. Pour pouvoir exploiter la base, nous devons le convertir au format IDAMS, composé de deux fichiers : « .dat » pour le stockage des valeurs ; « .dic » pour le dictionnaire des données. **L'onglet « Variables Definition » doit être actif pour accéder au menu BUILD.** Nous actionnons l'item « IDAMS DATASET ».



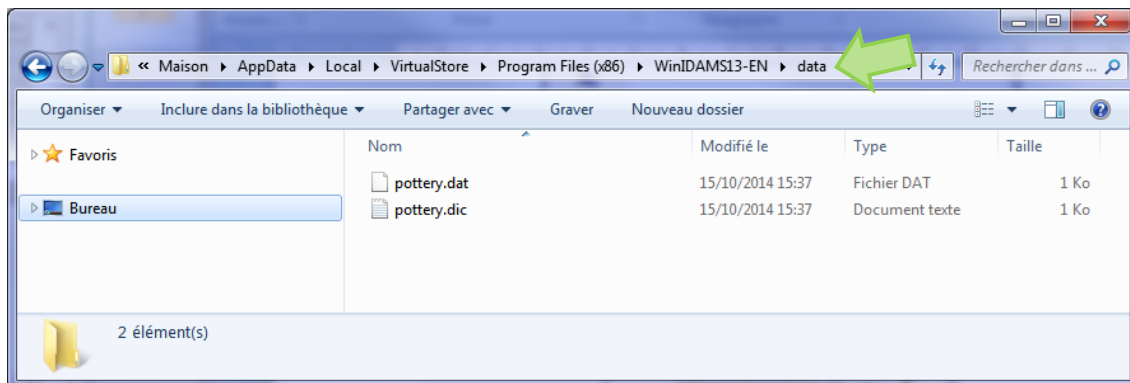
Nous spécifions le nom du dictionnaire « pottery.dic », puis celui du fichier des valeurs.



Une boîte de dialogue affiche le bilan des opérations.



Emplacement des fichiers sur Windows 7. A priori, les fichiers sont enregistrés dans le dossier « C:\Program Files (x86)\WinIDAMS13-EN\data », c'était le chemin que j'avais spécifié en tous les cas. Pourtant, à ma grande surprise, je ne les y ai pas trouvés dans l'explorateur Windows. Ils sont en réalité entreposés dans « C:\Users\Nom_Utilisateur\AppData\Local\VirtualStore\Program Files (x86)\WinIDAMS13-EN\data », Comme nous pouvons le voir dans la copie d'écran ci-dessous.

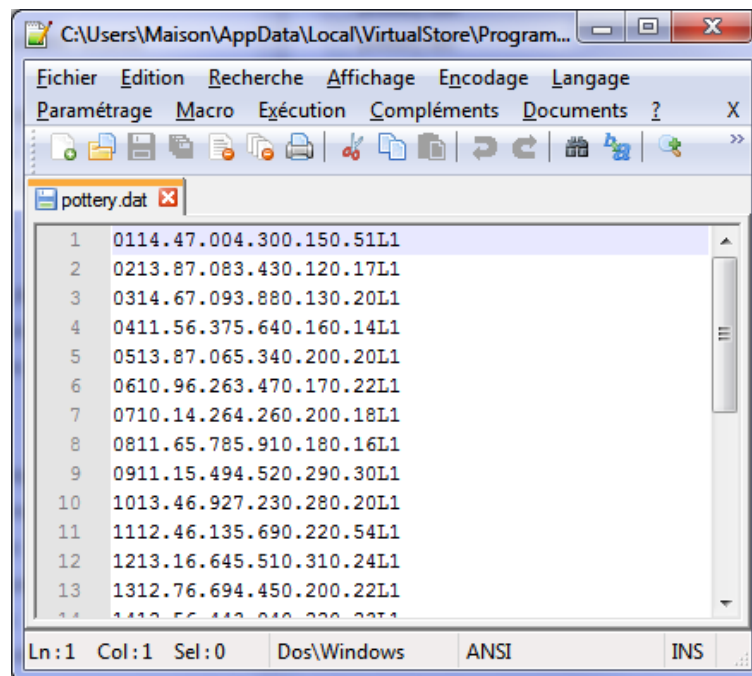


Après coup, cela se comprend. Le dossier « Program Files » est normalement protégé. Windows déporte les écritures sur un autre emplacement où les accès sont autorisés. Pour WinIDAMS, l'opération est complètement transparente. Il « voit » bien les fichiers dans le répertoire dédié aux données.

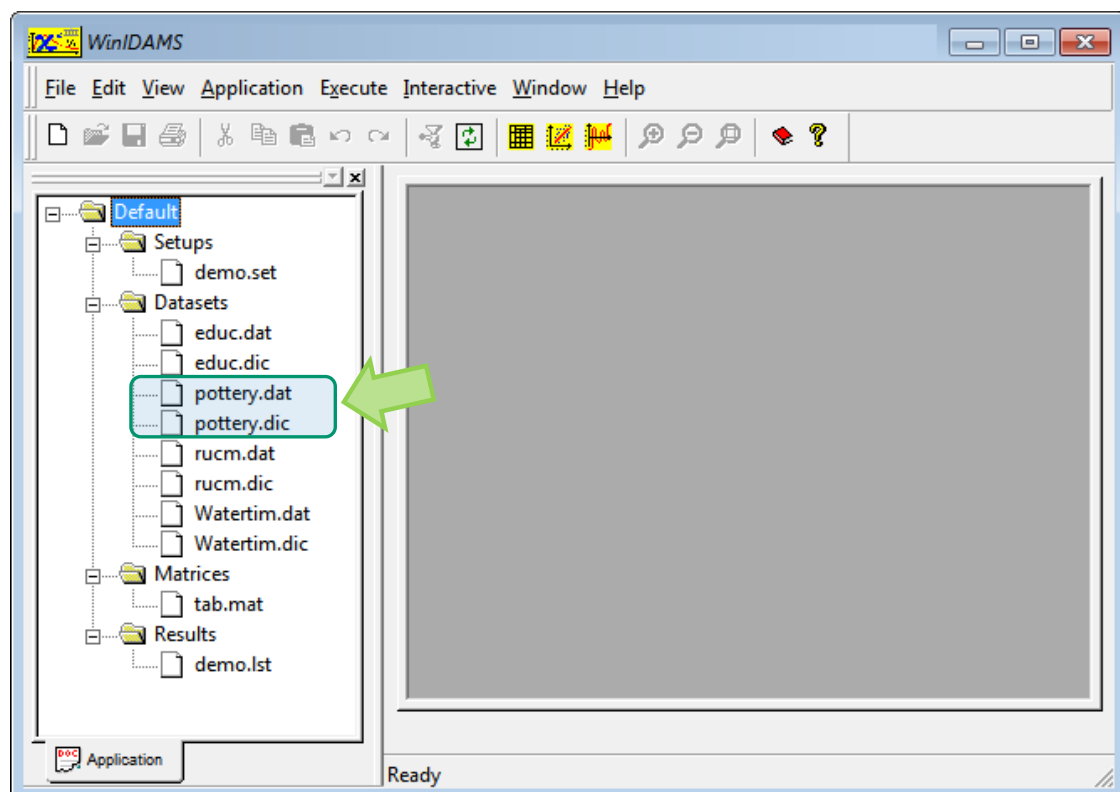
Contenu du dictionnaire des données. Nous y trouvons (fichier « .dic ») la liste des variables (numérotées de 1 à 8 ; la première colonne, l'identifiant, correspond au n°1) et leurs caractéristiques. La largeur de chaque colonne est spécifiée.

1	3	19999	1	1			
2	T	1	ID		1	20	
3	T	2	Al		3	41	
4	T	3	Fe		7	42	
5	T	4	Mg		11	41	
6	T	5	Ca		15	41	
7	T	6	Na		19	42	
8	T	7	Site		23	101	
9	T	8	SiteNum		24	10	

Contenu du fichier des valeurs. Les données sont énumérées dans le fichier « .dat ». Il y a une ligne par individu. WinIDAMS s'appuie sur les largeurs de colonnes pour repérer les valeurs associées aux variables.



Chargement des données « pottery » dans WinIDAMS. Le plus simple pour faire apparaître nos données dans l'onglet « Application » est de fermer puis de redémarrer le logiciel. Les fichiers étant dans le répertoire « Data » spécifique au logiciel, il les reconnaît immédiatement. Je suis allé au plus simple, il doit sûrement y avoir une autre manière, plus directe, pour charger un fichier au format reconnu par WinIDAMS.



Nous sommes maintenant prêts pour lancer les traitements statistiques.

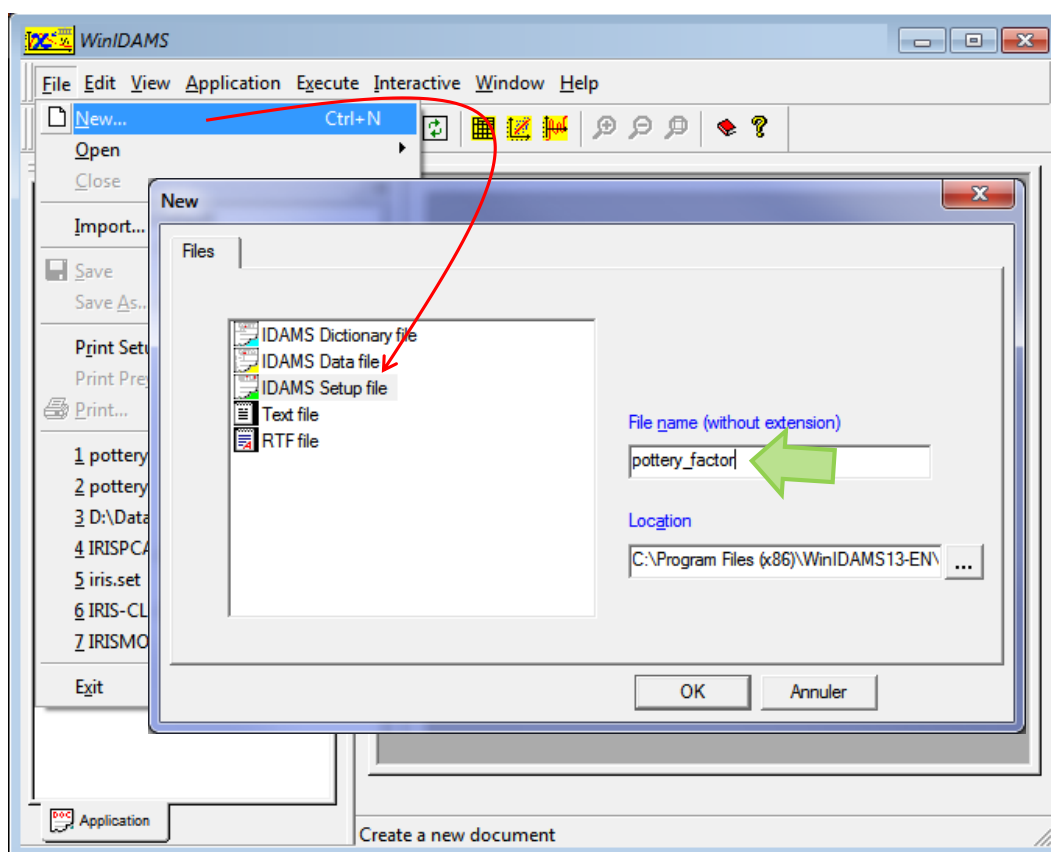
4 Analyses statistiques avec WinIDAMS

Pour définir les traitements, nous devons créer un fichier SETUP « .set ». Pour simplifier, nous créerons un fichier par analyse dans ce tutoriel. Chaque méthode dispose de très nombreuses options. Il faut scruter dans le détail la documentation³ pour en cerner l'étendue. Nous nous en tiendrons à des analyses très simples dans notre document. Il faudra garder à l'esprit que certains paramètres, dont les valeurs sont fixées par défaut, pèsent sur les résultats.

4.1 Analyse factorielle (FACTOR)

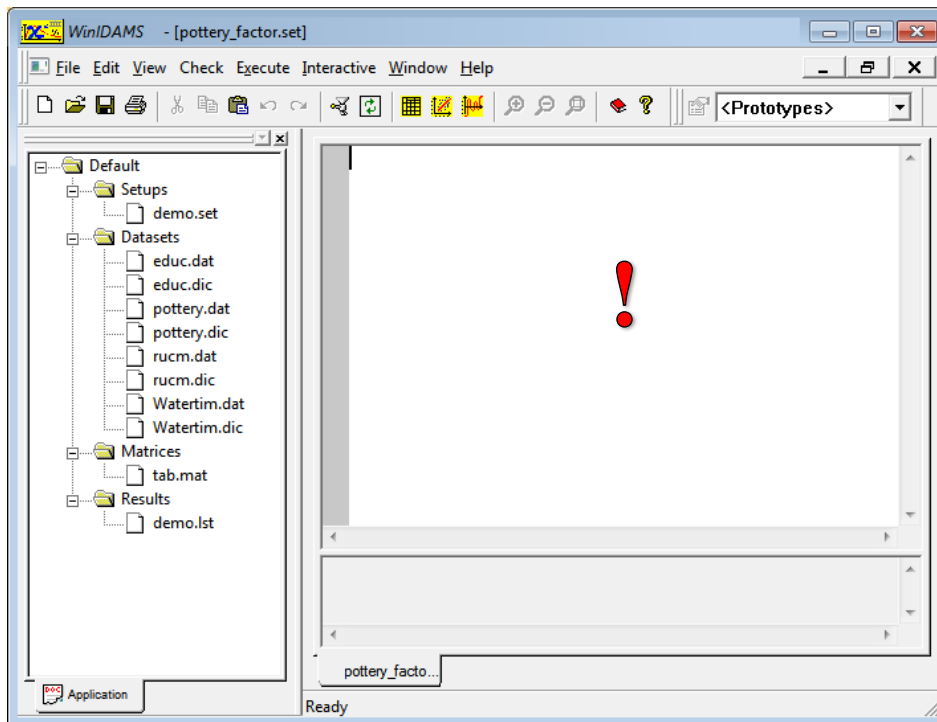
Définition des traitements. La méthode FACTOR recouvre un ensemble de méthodes relatives à l'analyse des correspondances et à l'analyse factorielle. Elle s'appuie sur la diagonalisation d'une matrice calculée à partir des données. Dans le cas de la matrice de corrélation, nous devrions obtenir les résultats de l'analyse en composantes principales.

Pour créer un fichier setup, nous actionnons le menu FILE / NEW. Nous sélectionnons l'item « IDAMS Setup file » dans la boîte de paramétrage. Nous attribuons le nom de fichier « pottery_factor » à notre analyse. Nous conservons l'emplacement par défaut (Location).

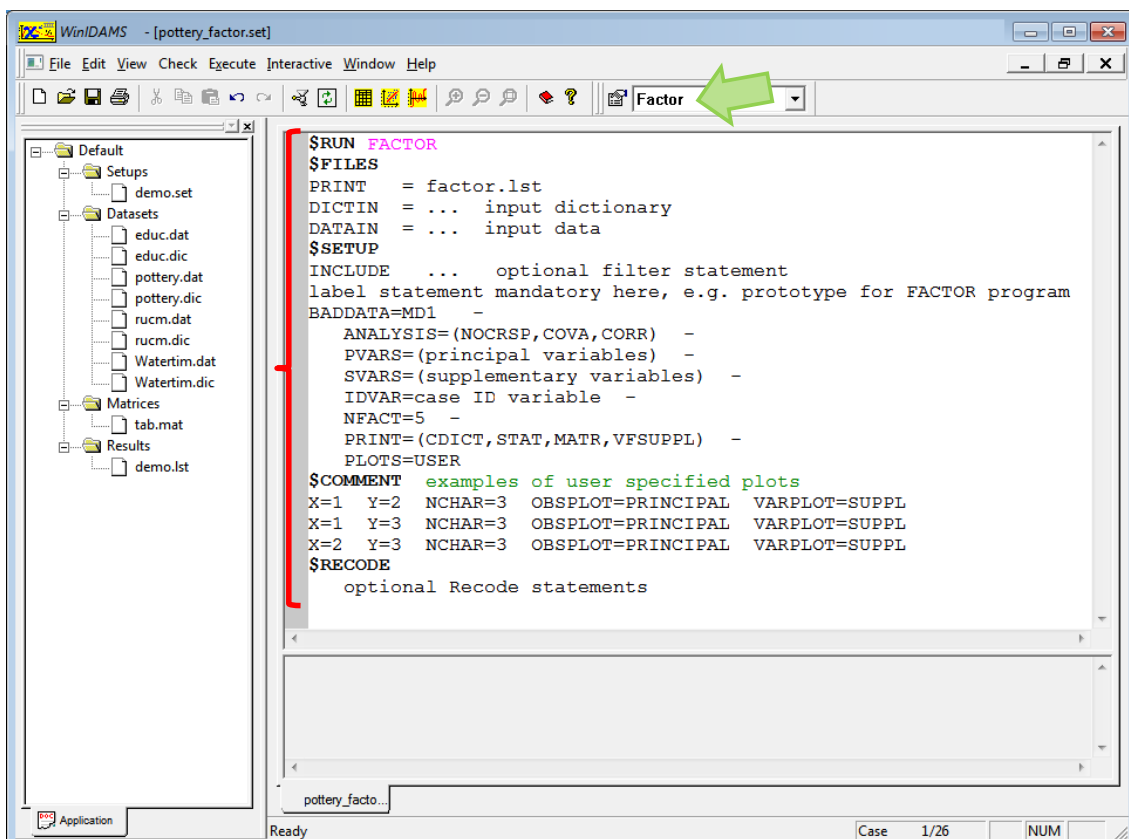


Un éditeur de texte apparaît. Nous pouvons y saisir les spécifications du traitement à réaliser.

³ <http://www.unesco.org/webworld/portal/idams/html/english/TOC.htm>



Bien sûr, nous pouvons saisir chaque section du fichier setup. **Il faut se référer à la documentation pour définir les options adéquates. La liste est longue.** Heureusement, WinIDAMS peut générer automatiquement un prototype du code pour chaque traitement. Nous sélectionnons « Factor » dans la liste « **Prototypes** » en haut à droite de la fenêtre principale. La maquette suivante est générée, libre à nous de la modifier à notre guise.



Nous spécifions le code suivant dans le fichier « **pottery_factor.set** »

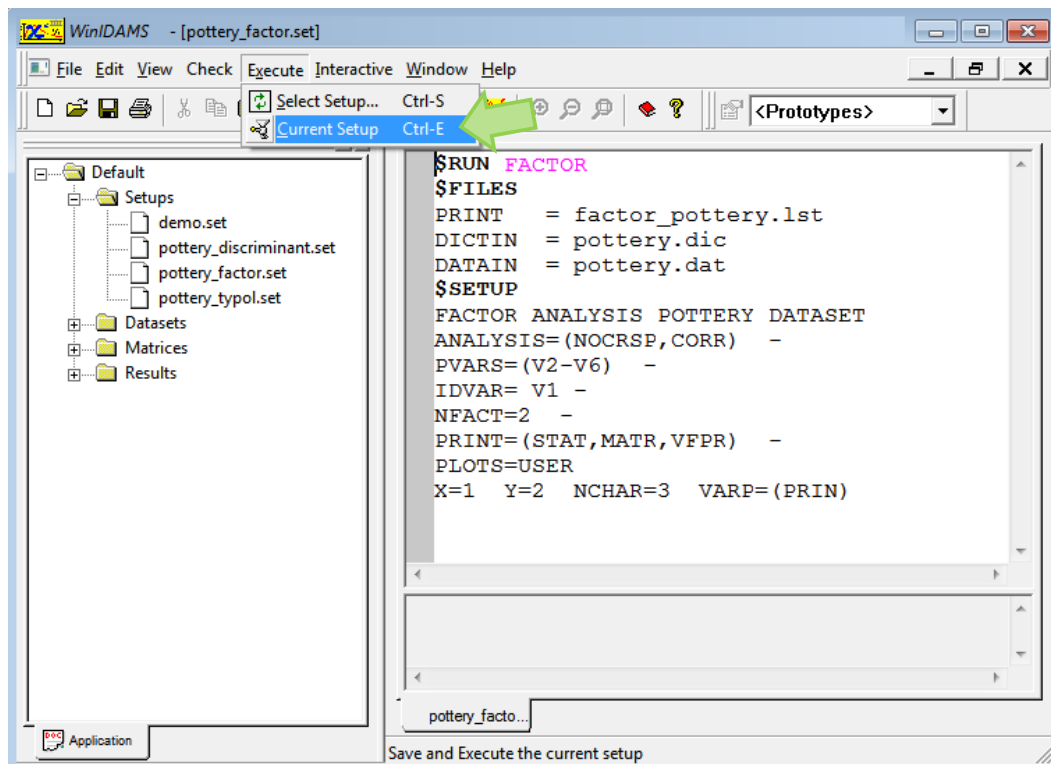
```

$RUN FACTOR
$FILES
PRINT    = factor_pottery.lst
DICTIN   = pottery.dic
DATAIN   = pottery.dat
$SETUP
FACTOR ANALYSIS POTTERY DATASET
ANALYSIS=( NOCRSP, CORR)  -
PVAR= (V2-V6)  -
IDVAR= V1  -
NFACT=2  -
PRINT=( STAT, MATR, VFPR)  -
PLOTS=USER
X=1  Y=2  NCHAR=3  VARP=(PRIN)

```

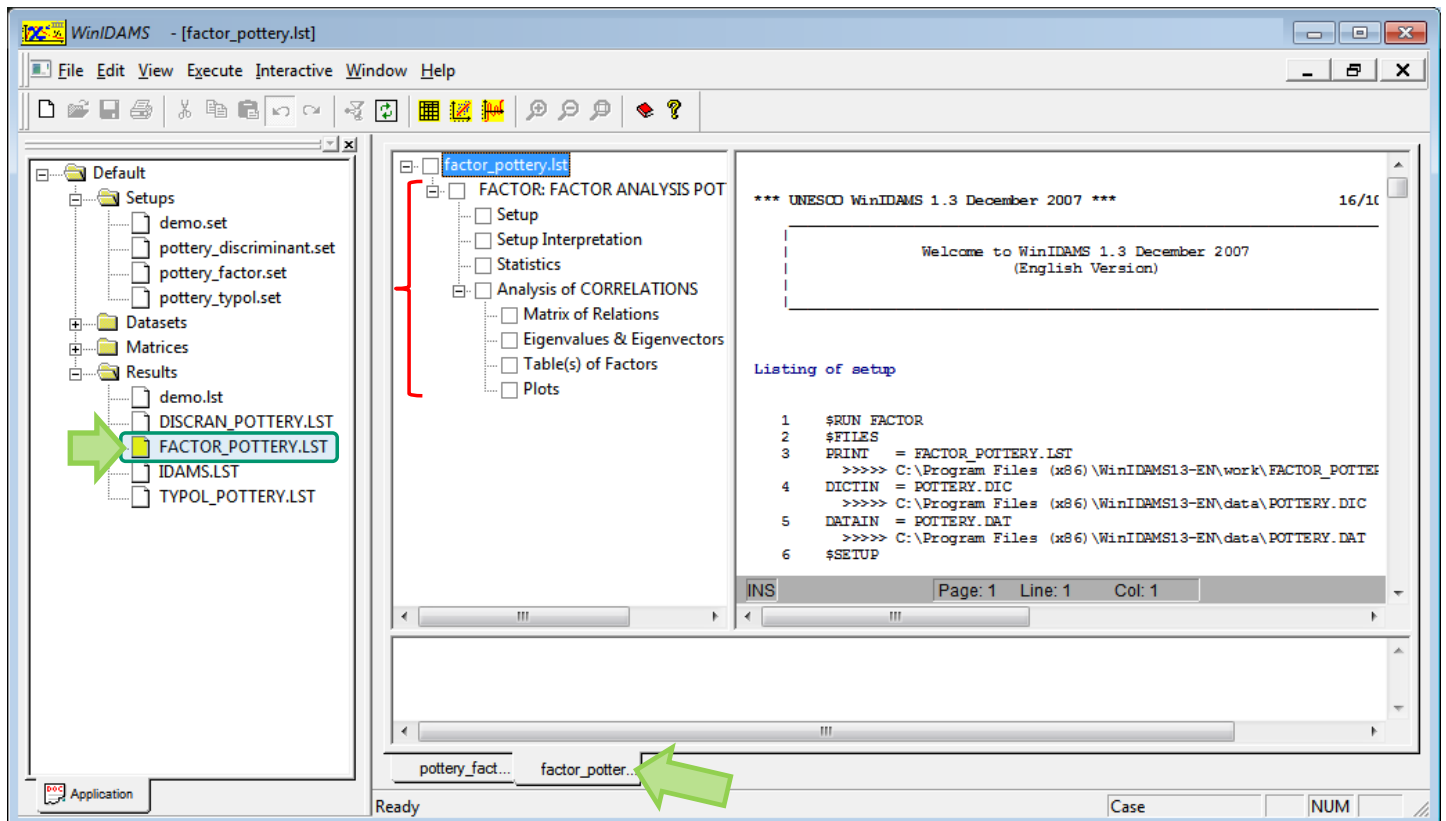
Je ne vais pas rentrer dans les détails. La documentation est très complète⁴. Je me bornerai simplement à préciser que l'option **ANALYSIS** spécifie le type de traitement. Nous ne réalisons pas une analyse de correspondance (NOCRSP), nous diagonalisons la matrice des corrélations (CORR) c.-à-d. nous effectuons une analyse en composantes principales.

Lancement des traitements. Pour exécuter l'analyse, nous actionnons le menu EXECUTE / CURRENT SETUP (raccourci CTRL + E).



⁴ http://www.unesco.org/webworld/portal/idams/html/english/E1factor.htm#Ktw_7

Lecture des résultats. Les sorties sont insérées dans le fichier « factor_pottery.lst » qui est automatiquement affiché. Les résultats sont organisés en sections hiérarchiques, un peu à la manière de SAS et SPSS.



Ci-dessous, nous reprenons les grandes lignes du fichier « factor_pottery.lst » en y insérant quelques commentaires. Une description très détaillée des sorties de la méthode FACTOR est accessible en ligne⁵.

```

*** UNESCO WinIDAMS 1.3 December 2007 ***                               16/10/2014  11: 1:20

|-----|
|               Welcome to WinIDAMS 1.3 December 2007               |
|               (English Version)                                     |
|-----|

After filtering,           26 cases read from the input data file
Cases kept:           26;   Weight of principal cases:           26.00

CORRELATIONS      factor analysis
FACTOR ANALYSIS POTTERY DATASET

```

⁵ « 6(1) Example of Factor Analysis of Correlations », <http://www.unesco.org/webworld/idams/advguide/ex6.htm>

```

# La matrice des corrélations qui est diagonalisée
Core matrix ( multiplied by 10000 )

      Al   Fe   Mg   Ca   Na
Al  10000
Fe -788710000
Mg -7983 900710000
Ca -7634 7652 842010000
Na -4725 6617 6427 481510000

Trace= 0.50000E+01

# Tableau des valeurs propres et des vecteurs propres
# Rappelons qu'en multipliant les valeurs des vecteurs propres
# avec la racine carrée de la valeur propre, nous obtenons
# les corrélations des variables avec les axes factoriels

Eigenvalues and eigenvectors
=====
NO | EIGVAL 1 | EIGVAL 2 |
   | 3.8778 | 0.6088 |
----+-----+-----+
 1| -0.4454 | 0.3565 |
 2| 0.4781 | 0.0412 |
 3| 0.4865 | -0.0496 |
 4| 0.4491 | -0.3441 |
 5| 0.3669 | 0.8662 |

# Corrélation des variables avec les facteurs
# Cos² = (Corrélation)²
# Contribution = Cos² / Valeur propre

# QLT indique la proportion de la variance de la variable
# restituée par les 2 premiers facteurs

CORRELATIONS      factors      FACTOR ANALYSIS POTTERY DATASET
Table of principal variable factors
---*---*-----*-----*-----*-----*
| JPR | QLT WEIG INR| 1#F COS2 CPF| 2#F COS2 CPF|
---*---*-----*-----*-----*-----*
 1| 2| 847 647 200| -877 769 198| 278 77 127|
 2| 3| 888 199 200| 942 887 229| 32 1 2|
 3| 4| 919 140 200| 958 918 237| -39 1 2|
 4| 5| 854 7 200| 884 782 202| -269 72 118|
 5| 6| 979 7 200| 722 522 135| 676 457 750|
---*---*-----*-----*-----*-----*
| | 26.0 1000| 1000| 1000|
---*---*-----*-----*-----*-----*

```

A titre de comparaison, voici les résultats fournis par SAS (PROC FACTOR)⁶ et Tanagra.

The SAS System

The FACTOR Procedure
Initial Factor Method: Principal Components

Prior Communality Estimates: ONE

Eigenvalues of the Correlation Matrix: Total = 5 Average = 1			
	Eigenvalue	Difference	Proportion
1	3.87779699	3.26898715	0.7756
2	0.60880984	0.36461766	0.1218
3	0.24419218	0.05928276	0.0488
4	0.18490942	0.10061786	0.0370
5	0.08429156		0.0169

2 factors will be retained by the NFACTOR criterion.

Factor Pattern		
	Factor1	Factor2
Al	-0.87715	0.27818
Fe	0.94154	0.03213
Mg	0.95810	-0.03871
Ca	0.88428	-0.26852
Na	0.72248	0.67586

Variance Explained by Each Factor	
Factor1	Factor2
3.8777970	0.6088098

Eigenvalue table - Test for significance

Eigenvalues - Significance		
Axis	Eigenvalue	Broken-stick critical values
1	3.877797	2.283333
2	0.608810	1.283333
3	0.244192	0.783333
4	0.184909	0.450000
5	0.084292	0.200000

Factor Loadings [Communality Estimates]

Attribute	Axis_1		Axis_2	
	Corr.	% (Tot. %)	Corr.	% (Tot. %)
Al	-0.87715	77 % (77 %)	0.27818	8 % (85 %)
Fe	0.94154	89 % (89 %)	0.03213	0 % (89 %)
Mg	0.95810	92 % (92 %)	-0.03871	0 % (92 %)
Ca	0.88428	78 % (78 %)	-0.26852	7 % (85 %)
Na	0.72248	52 % (52 %)	0.67586	46 % (98 %)
Var. Expl.	3.87780	78 % (78 %)	0.60881	12 % (90 %)

Factor Score Coefficients

Attribute	Mean	Std-dev	Axis_1	Axis_2
Al	14.4923077	2.9345321	-0.4454340	0.3565238
Fe	4.4676923	2.3629550	0.4781318	0.0411751
Mg	3.1415385	2.1373972	0.4865413	-0.0496072
Ca	0.1465385	0.0992643	0.4490540	-0.3441465
Na	0.1584615	0.1326561	0.3668876	0.8661973

SAS – PROC FACTOR

TANAGRA – PRINCIPAL COMPONENT ANALYSIS

Sans surprise aucune, les 3 outils sont complètement en phase parce qu'ils diagonalisent bien la même matrice.

4.2 Analyse discriminante (DISCRAN)

L'analyse discriminante DISCRAN⁷ recouvre à la fois les aspects descriptifs (analyse factorielle) et prédictifs (sélection de variables, évaluation sur un échantillon test, matrice de confusion).

A l'aide du même cheminement que précédemment, nous avons généré le fichier setup « **pottery_discriminant.set** ».

```

6 proc factor data = mesdata.pottery
  method = principal
  nfactors = 2
  corr;
  var Al Fe Mg Ca Na;
  run;

```

⁷ http://www.unesco.org/webworld/portal/idams/html/english/E1discra.htm#Ktw_0

```

$RUN DISCRAN
$FILES
PRINT = discran_pottery.lst
DICTIN = pottery.dic
DATAIN = pottery.dat
$SETUP
DISCRIMINANT ANALYSIS POTTERY DATASET
VARS=(V2-V6) -
IDVAR=V1 -
PRINT=(DATA,GROU) -
GRVAR=V8 GR01=1 GR02=2 GR03=3 GR04=4

```

Nous obtenons *grasso modo*⁸ :

```

# Pas d'échantillon test. Matrice de confusion en resubstitution

      Number of cases in samples
Basic:   26   Test:    0   Anonymous:    0

      Revised number of cases in samples
Basic:   26   Test:    0   Anonymous:    0

# Moyennes des variables conditionnellement aux groupes d'appartenance

Table of Means

Variable      GR01      GR02      GR03      GR04      TOT.
   2          12.5643   11.7000   18.1800   17.3200   14.4923
   3           6.3721    5.4150    1.7120    1.5120    4.4677
   4           4.8264    3.8550    0.6740    0.6060    3.1415
   5           0.2021    0.2950    0.0260    0.0520    0.1465
   6           0.2507    0.0500    0.0540    0.0480    0.1585

# Première étape : sélection de la 1ère meilleure variable (V3 = Fe)
# Sélection FORWARD

Step number    1
Variables entered:      3

# Matrice de confusion avec V3 comme seule variable prédictive

# Ligne : groupe observé
# Colonne : groupe prédit

      Classification table for basic sample
      *****

```

⁸ Cf. <http://www.unesco.org/webworld/idams/advguide/ex9.htm> pour une analyse approfondie des sorties de DISCRAN.

```

                Allocated group
                GR01  GR02  GR03  GR04
Original group
    GR01         11     3     0     0
    GR02          0     2     0     0
    GR03          0     0     2     3
    GR04          0     0     2     3

# Taux de bon classement = 1 - Taux d'erreur

Per cent of correctly classified cases:  69.23

*****

# 2ème étape : sélection de la 2ème meilleure variable (V3 étant validé)

Step number      2
Variables entered:      3      5

# etc.

# Dernière étape : finalement les 5 variables auront été sélectionnées

Step number      5
Variables entered:      3      5      6      4      2

Classification table for basic sample
*****

                Allocated group
                GR01  GR02  GR03  GR04
Original group
    GR01         14     0     0     0
    GR02          0     2     0     0
    GR03          0     0     4     1
    GR04          0     0     0     5

# Taux de bon classement final en resubstitution
# Un seul individu mal classé
Per cent of correctly classified cases:  96.15

# Groupe attribué et distance aux centres de groupes
# Ce tableau permet de détecter les individus mal classés

```

```
# et surtout l'importance de l'erreur
```

```
Allocation and distances of cases in the basic sample
(In parentheses group number under consideration)
```

```
# Individus appartenant au 1er groupe
```

```
# Allocation = groupe d'affectation
```

```
# Distance aux centres de classes (affectation à la valeur minimum)
```

Group	1	Allocation	Distances to each group			
GR01						
1	1	1	8.277 (1)	25.719 (2)	13.095 (3)	14.225 (4)
2	1	1	5.833 (1)	13.868 (2)	8.745 (3)	10.383 (4)
3	1	1	4.675 (1)	14.449 (2)	7.436 (3)	9.646 (4)
4	1	1	5.116 (1)	15.901 (2)	9.334 (3)	10.414 (4)
5	1	1	2.283 (1)	10.971 (2)	6.108 (3)	7.883 (4)
6	1	1	4.194 (1)	8.847 (2)	8.871 (3)	7.436 (4)
7	1	1	7.421 (1)	11.676 (2)	11.453 (3)	8.601 (4)
8	1	1	5.825 (1)	16.652 (2)	9.914 (3)	10.528 (4)
9	1	1	5.762 (1)	6.612 (2)	11.180 (3)	8.077 (4)
10	1	1	8.561 (1)	16.966 (2)	13.280 (3)	14.961 (4)
11	1	1	9.632 (1)	26.388 (2)	15.878 (3)	15.226 (4)
12	1	1	3.892 (1)	5.191 (2)	8.748 (3)	8.138 (4)
13	1	1	0.641 (1)	6.623 (2)	4.857 (3)	5.014 (4)
14	1	1	1.951 (1)	4.903 (2)	6.226 (3)	5.368 (4)

```
# Individus appartenant au 2ème groupe
```

Group	2	Allocation	Distances to each group			
GR02						
15	2	2	8.293 (1)	0.036 (2)	11.720 (3)	9.149 (4)
16	2	2	7.593 (1)	0.036 (2)	11.141 (3)	8.389 (4)

```
# Individus appartenant au 3ème groupe
```

```
# L'individu n°18 est mal classé
```

```
# Mais on se rend compte que l'erreur tient à peu de choses
```

```
# L'individu est presque aussi proche du 3ème que du 4ème groupe
```

Group	3	Allocation	Distances to each group			
GR03						
17	3	3	4.997 (1)	11.279 (2)	0.197 (3)	0.538 (4)
18	3	4	5.264 (1)	11.975 (2)	2.485 (3)	2.326 (4)
19	3	3	4.638 (1)	13.112 (2)	0.202 (3)	0.733 (4)
20	3	3	4.521 (1)	12.381 (2)	0.191 (3)	1.021 (4)
21	3	3	9.569 (1)	15.408 (2)	4.108 (3)	5.222 (4)

```
# Individus appartenant au 4ème groupe
```



```

Group 4 Allocation Distances to each group
GR04
22 4 5.212 ( 1) 9.316 ( 2) 0.988 ( 3) 0.236 ( 4)
23 4 5.381 ( 1) 9.806 ( 2) 0.784 ( 3) 0.505 ( 4)
24 4 6.171 ( 1) 13.661 ( 2) 2.063 ( 3) 1.413 ( 4)
25 4 5.970 ( 1) 10.967 ( 2) 4.144 ( 3) 3.311 ( 4)
26 4 7.630 ( 1) 8.196 ( 2) 2.958 ( 3) 2.815 ( 4)

# Analyse factorielle discriminante
# Discriminant power = carré de la corrélation canonique
# Sum of eigenvalues = somme des « discriminant power »

Discriminant factor analysis at step 5
*****

Sum of eigenvalues = 1.55394

Discriminant power of first factor: 0.97156
Discriminant power of second factor: 0.55558
Discriminant power of third factor: 0.02680

# Vecteurs propres
# On ne voit pas très bien dans quel ordre sont les variables
# Selon leur position dans la base de données ?
# Selon le processus de sélection ?

First eigenvector
-0.22642 -1.61651 -0.40478 -0.04563 0.08223

Second eigenvector
-0.04214 -14.47849 4.20584 0.59706 0.09015

```

Ces sorties aident à la compréhension des résultats. Entre autres, le tableau des distances aux centres de classes permettent de situer la fiabilité de l'affectation.

Il n'en reste pas moins que des informations importantes manquent. Par exemple, le tableau des tests de significativité des facteurs manquent. Pourtant elle est primordiale dans le choix du nombre d'axes à retenir. Avec SAS (PROC DISCRIM) et Tanagra, on se rend compte que le dernier facteur n'est pas pertinent, il peut être éliminé.

SAS – PROC DISCRIM

	Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation	Eigenvalues of Inv(E)'H = CanRsqr/(1-CanRsqr)				Test of H0: The canonical correlations in the current row and all that follow are zero				
					Eigenvalue	Difference	Proportion	Cumulative	Likelihood Ratio	Approximate F Value	Num DF	Den DF	Pr > F
1	0.985677	0.982731	0.005688	0.971559	34.1611	32.9110	0.9639	0.9639	0.01230091	13.09	15	50.091	<.0001
2	0.745369	0.704663	0.088885	0.555575	1.2501	1.2226	0.0353	0.9992	0.43251355	2.47	8	38	0.0290
3	0.163712	-.065337	0.194640	0.026802	0.0275		0.0008	1.0000	0.97319849	0.18	3	20	0.9063

TANAGRA – CANONICAL DISCRIMINANT ANALYSIS
Roots and Wilks' Lambda

Root	Eigenvalue	Proportion	Canonical R	Wilks Lambda	CHI-2	d.f.	p-value
1	34.16112	0.96395	0.985677	0.012301	90.1607	15	0.000000
2	1.25010	0.99922	0.745369	0.432514	17.1819	8	0.028270
3	0.02754	1.00000	0.163712	0.973199	0.5569	3	0.906218

Test of H0: The canonical correlation in the current row and all that follow are zero (Bartlett's chi-square approximation)

Il en est de même concernant d'autres aides à l'interprétation (ex. les structures canoniques). Il ne s'agit certainement pas d'accabler WinIDAMS. Le plus important est de bien cerner les informations dont nous pourrions avoir besoin pour une interprétation approfondie, et d'avoir conscience de celles qui pourraient nous faire défaut.

Plus ennuyeux en revanche, la matrice de confusion en resubstitution ne coïncide pas avec celles de SAS et Tanagra. Je ne m'explique pas cette différence.

The DISCRIM Procedure
 Classification Summary for Calibration Data: MESDATA.POTTERY
 Resubstitution Summary using Linear Discriminant Function

Number of Observations and Percent Classified into Site					
From Site	A	C	I	L	Total
A	4	0	1	0	5
	80.00	0.00	20.00	0.00	100.00
C	0	2	0	0	2
	0.00	100.00	0.00	0.00	100.00
I	1	0	4	0	5
	20.00	0.00	80.00	0.00	100.00
L	0	0	0	14	14
	0.00	0.00	0.00	100.00	100.00
Total	5	2	5	14	26
	19.23	7.69	19.23	53.85	100.00
Priors	0.19231	0.07692	0.19231	0.53846	

0.0769

Confusion matrix

	L	C	I	A	Sum
L	14	0	0	0	14
C	0	2	0	0	2
I	0	0	4	1	5
A	0	0	1	4	5
Sum	14	2	5	5	26

SAS – PROC DISCRIM

TANAGRA – LINEAR DISCRIMINANT ANALYSIS

Les deux individus mal classés par SAS et Tanagra sont le n°18 (on sait pourquoi) et le n°24 (WinIDAMS l'a affecté correctement dans les sorties ci-dessus – cf. distance aux classes).

4.3 Classification automatique (TYPOL)

Deux procédures sont dévolues à la classification automatique (clustering en anglais) dans WinIDAMS : [CLUSFIND](#) et [TYPOL](#). La première propose différents algorithmes de partitionnement. La seconde est basée sur une approche hiérarchique.

TYPOL agit en deux temps : un premier regroupement est d'abord formé, il peut être spécifié par l'utilisateur à l'aide d'une variable qualitative ; l'algorithme opère alors une classification ascendante hiérarchique (CAH) à partir de ce pré-regroupement. Il s'agit ni plus ni moins que d'une CAH Mixte⁹. Elle est particulièrement adaptée au traitement des gros volumes car elle permet de dépasser la complexité inhérente à la CAH. Mais, dans le même temps, nous bénéficions quand même de l'aide à la détection du nombre groupes illustrée par le dendrogramme. L'approche n'est en rien pénalisante puisque nous nous en tenons à un nombre réduit de groupes à la sortie.

Voici le contenu du fichier setup « **pottery_typol.set** ».

```

$RUN TYPOL
$FILES
PRINT = typol_pottery.lst
DICTIN = pottery.dic
DATAIN = pottery.dat
$SETUP
TYPOL POTTERY DATASET
AQNTVARS=(V2-V6) -
REDUCE -
DTYPE=EUCLID -
INIGROUP=5 -
FINGROUP=3 -
PRINT=(GRAP)

```

Les variables sont standardisées (REDUCE). L'algorithme s'appuie sur la distance euclidienne. Le nombre initial de groupes lors du pré-regroupement est de 5 (INIGROUP). Nous souhaitons obtenir 3 groupes au final (FINGROUP).

WinIDAMS détaille le processus de regroupement en démarrant à 5 groupes. On croque sous les tableaux de chiffres à vrai dire. Nous n'affichons que le résultat final ci-dessous c.-à-d. les caractéristiques de la partition en 3 groupes.

```

# 3 groupes ont été générés
# avec respectivement 53.8% (14 individus), 38.4% (10) et 7.6% (2) de l'échantillon

      Total cases      1000
      Group number      1      4      5
      Proportion of cases 538  384  76

# pour chaque variable, nous disposons :
# - de la proportion de variance expliquée c.-à-d. le carré du rapport de corrélation
# - de la moyenne globale
# - des moyennes conditionnelles
# - des écarts-type conditionnels

```

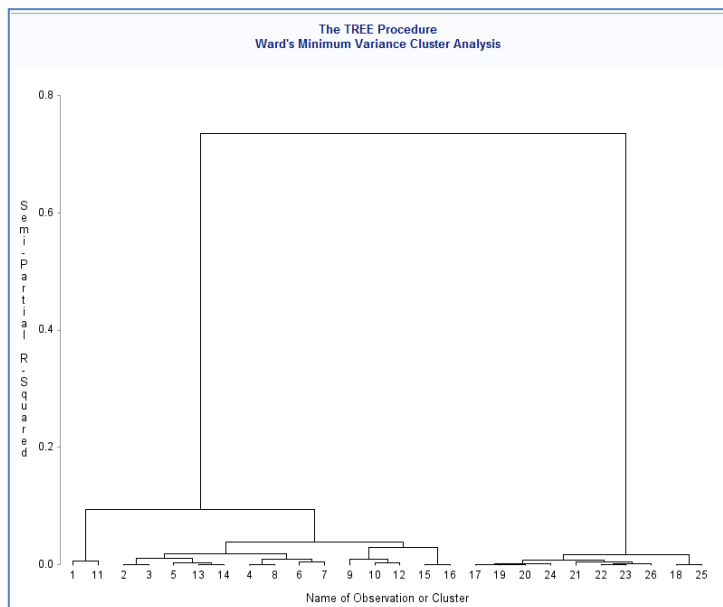
⁹ Tutoriel Tanagra, « [Traitement de gros volumes – CAH Mixte](#) », octobre 2008 ; Tutoriel Tanagra, « [CAH Mixte – Le fichier IRIS de Fisher](#) », mars 2008.

	Explained variance	Grand mean				
1	779 *****	14.49	Al	12.32	17.75	13.40
				1.25	1.60	1.00
2	914 *****	4.47	Fe	6.21	1.61	6.57
				0.80	0.55	0.43
3	858 *****	3.14	Mg	4.66	0.64	4.99
				1.07	0.06	0.70
4	741 *****	0.15	Ca	0.22	0.04	0.18
				0.06	0.03	0.04
5	858 *****	0.16	Na	0.18	0.05	0.53
				0.07	0.02	0.02

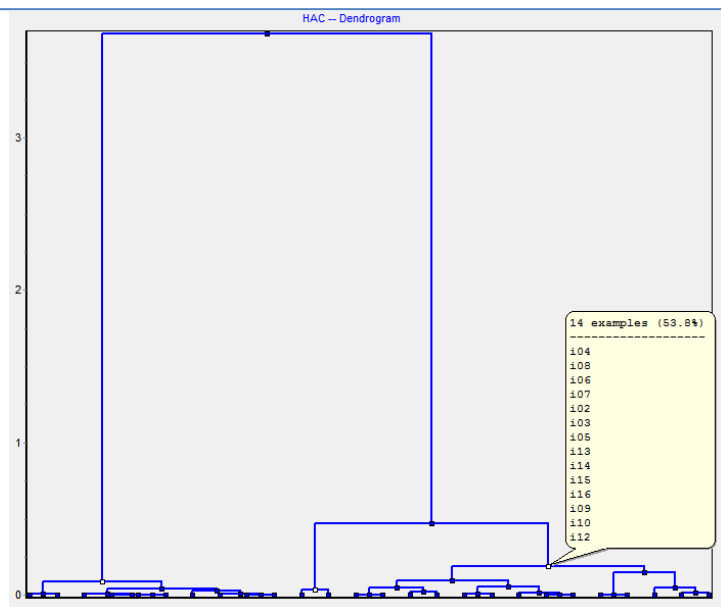
proportion d'inertie expliquée par la partition

Mean variance explained by active variables = 830

A titre de comparaison, voici les dendrogrammes obtenus sous SAS (PROC CLUSTER & TREE) et Tanagra¹⁰ avec la méthode de Ward appliquée sur variables standardisées. L'algorithme prend pour point de départ la partition initiée par les individus¹¹.



SAS – PROC CLUSTER & TREE



TANAGRA – HAC

Lorsque l'on s'intéresse à la solution en 3 groupes, nous constatons que les résultats rejoignent ceux de WinIDAMS. Les effectifs et la proportion d'inerte expliquée (83%) sont les mêmes.

¹⁰ L'ordonnancement graphique est différent, mais nous avons bien la même hiérarchie de partition.

¹¹ Une stratégie analogue – CAH Mixte - à celle de WinIDAMS est tout à fait envisageable sous Tanagra et SAS (cf. PROC FASTCLUS + PROC CLUSTER).

Clustering results		
Clusters	From the dendrogram	After one-pass relocation
cluster n° 1	10	10
cluster n° 2	2	2
cluster n° 3	14	14

Best cluster selection		
Clusters	BSS ratio	Gap
1	0.0000	0.0000
2	0.7366	3.2155
3	0.8300	0.2771

Et surtout nous avons exactement les mêmes regroupements d'individus comme en atteste le calcul des moyennes et écarts-type conditionnels des variables à partir de la solution en 3 classes de SAS ou de Tanagra.

Valeurs	Étiquett <input type="text" value="c_hac_3"/>			Global
	c_hac_3	c_hac_1	c_hac_2	
Moyenne de Al	12.32	17.75	13.40	14.49
Écartypep de Al	1.25	1.60	1.00	2.93
Moyenne de Fe	6.21	1.61	6.57	4.47
Écartypep de Fe	0.80	0.55	0.43	2.36
Moyenne de Mg	4.66	0.64	5.00	3.14
Écartypep de Mg	1.07	0.06	0.69	2.14
Moyenne de Ca	0.22	0.04	0.19	0.15
Écartypep de Ca	0.06	0.03	0.03	0.10
Moyenne de Na	0.18	0.05	0.53	0.16
Écartypep de Na	0.07	0.02	0.02	0.13

4.4 Autres méthodes

WinIDAMS propose plusieurs autres méthodes (régression linéaire, analyse de variance, etc.). Il serait fastidieux de tous les énumérer, encore plus de les analyser individuellement. Le lecteur désireux d'approfondir l'usage du logiciel pourra se référer à la documentation en ligne¹² : <http://www.unesco.org/webworld/portal/idams/html/english/TOC.htm>

5 Capacités et performances

WinIDAMS peut gérer des fichiers allant jusqu'à 1000 variables. Le nombre de lignes est uniquement contraint par la représentation interne des données et les capacités de la machine ([Data in IDAMS](#)). Chaque méthode est limitée par ses propres restrictions. Pour

¹² Pour ceux qui souffrent d'une aversion à l'anglais : <http://www.unesco.org/webworld/portal/idams/html/french/TOC.htm>

l'analyse discriminante par exemple ([DISCRAN](#)), les capacités maximales sont de 20 groupes et 99 variables ; pour l'analyse factorielle ([FACTOR](#)), nous pouvons traiter jusqu'à 80 variables actives ; etc. La documentation est très précise, il n'y a pas de mauvaises surprises.

6 Conclusion

Malgré d'évidentes qualités (procédures pour la manipulation des données, présence des méthodes phares de statistique exploratoire, assise scientifique, etc.), WinIDAMS n'a pas eu le succès escompté. Peut-être justement parce qu'il était trop ambitieux. S'appuyer sur un collège d'experts semble séduisant au premier abord. Mais dans les faits, faire collaborer des scientifiques dont le développement du logiciel n'est pas l'activité principale – ils sont donc très occupés par ailleurs – est très compliqué. Sans compter qu'il fallait pour chaque méthode trouver un point d'accord entre des personnes qui ne se connaissent pas, et qui souvent ont un caractère bien trempé. Le chef de projet a du se faire pas mal de cheveux blancs.

La seconde raison qui peut expliquer la contre-performance du logiciel est son ergonomie très perfectible. Son utilisation obéit à une logique parfaitement cohérente. Mais il faut savoir l'appréhender. Les options relatives aux méthodes sont très nombreuses. Il faut les connaître. Les utilisateurs sont très sensibles à cette barrière à l'entrée. Le découragement gagne vite lorsque, par exemple, on a du mal à charger simplement ses données pour initier les premiers traitements. Hélas, les tutoriels consacrés à WinIDAMS ne sont pas nombreux sur le web. Cette rareté joue en défaveur du logiciel.

Enfin, troisième raison, un peu plus ennuyeuse peut-être, nous avons parfois du mal à comprendre la teneur des résultats. En les rapprochant avec d'autres outils tels que SAS ou Tanagra, nous n'obtenons pas toujours les mêmes sorties. La solidité scientifique n'est pas en cause, mais le mode de présentation adopté n'est pas toujours accessible. Il faut se plonger dans la documentation technique « [WinIDAMS Reference Manual \(release 1.3\)](#) – Partie VI » pour identifier les différents éléments proposés. Pour ma part, comme il s'agissait avant tout d'une prise de contact, je n'ai pas eu le courage de disséquer les formules du manuel de référence, fort détaillées au demeurant.

Avec le recul dont nous disposons aujourd'hui, on pense immédiatement à R dès que l'on parle de projet collaboratif de développement d'un logiciel de statistique. Force est de constater que la comparaison est difficile pour WinIDAMS. Le modèle de R est autrement plus efficace et lui assure une richesse, une popularité et une pérennité qui se situe à un tout autre niveau. Résumé en quelques mots, le développement de R s'appuie sur deux points

essentiels : le cœur du logiciel est développé et maintenu par une équipe relativement réduite, le « R Development Core Team » ; elle a mis en place un système – les packages – suffisamment simple et solide permettant aux contributeurs d’enrichir à l’infini **de manière indépendante** la bibliothèque de méthodes¹³.

Le projet WinIDAMS¹⁴ semble au point mort. La dernière version (1.3) date de Mai 2008. C’est un peu dommage au vu de la puissance organisationnelle et de l’excellence scientifique qui la sous-tendait. Il y avait certainement matière à faire des choses très intéressantes.

¹³ Sans qu’il y ait un quelconque processus de validation d’ailleurs, impossible à mener compte tenu de leur très grand nombre. De fait, je dis toujours à mes étudiants d’être très prudents et de comparer systématiquement les résultats avec les autres logiciels lorsqu’ils utilisent des librairies plus ou moins reconnus. Bon, évitons d’être alarmistes quand même, la plupart des packages sont développés par des chercheurs qui font sérieusement leur boulot. De plus, le processus de validation devient effectif dès lors que la méthode est publiée dans « [The R Journal](#) ».

¹⁴ Une version « Open Source » [OPENIDAMS](#) semble avoir été initié. Mais là également, on ne sait pas vraiment à quel stade est le projet à ce jour.