1 Introduction

L'intégration de Tanagra dans un tableur, que ce soit Excel (http://tutoriels-data-mining.blogspot.com/2008/03/importation-fichier-xls-excel-macro.html) ou Open Office Calc (**OOCalc)** (http://tutoriels-data-mining.blogspot.com/2008/03/connexion-open-office-calc.html), via le système des Add-Ons, est certainement un des principaux facteurs de diffusion du logiciel. Sans connaissances particulières concernant la manipulation de fichiers, un utilisateur peut envoyer directement ses données à partir d'un environnement auquel il est familiarisé, le tableur, vers un logiciel spécialisé de Data Mining.

Les macros ont été initialement développées pour l'environnement Windows. Je me suis intéressé depuis peu au fonctionnement de Tanagra sous Linux via Wine (http://tutoriels-data-mining.blogspot.com/2009/01/tanagra-sous-linux.html). Je me suis rendu compte que le logiciel était pleinement fonctionnel sans que l'utilisateur n'ait besoin de procéder à des tripatouillages compliqués du système. Il ne restait plus qu'à établir une connexion entre le tableur phare sous Linux (OOCalc) et Tanagra.

M. Thierry Leiber a réalisé ce travail pour la version 1.4.31 de Tanagra. Il a étendu la macro complémentaire initialement destinée à la version d'Open Office sous Windows. En résumant un peu, le code consiste à tester le système en vigueur, former la commande adéquate pour lancer Tanagra, et transférer à ce dernier les données via le presse papier. De fait, **l'Add-On est maintenant opérationnel que ce soit sous Windows ou sous Linux**. Il a été testé en tous les cas dans les configurations suivantes : Windows XP + Open Office 3.0.0 ; Windows Vista + Open Office 3.0.1 ; Ubuntu 8.10 + Open Office 2.4 ; Ubuntu 8.10 + Open Office 3.0.1.

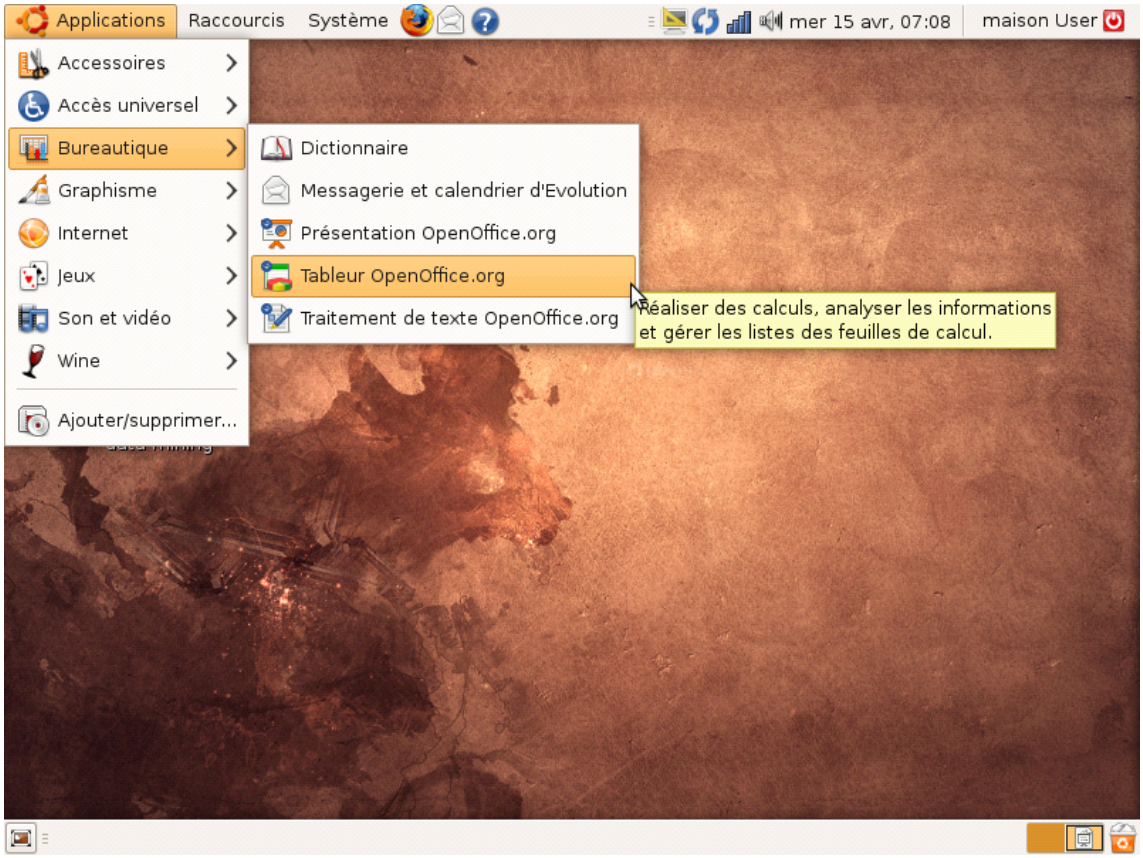
Ce document reprend donc un de nos anciens tutoriels (http://tutoriels-data-mining.blogspot.com/2008/03/connexion-open-office-calc.html). La nouveauté ici est que nous travaillons sous Linux (distribution **Ubuntu 8.10**).

2 Données

Nous utilisons les données CEREALS.XLS dans ce didacticiel (http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/cereals.xls). Elles décrivent les caractéristiques de 76 produits, des céréales, que l'on consomme habituellement au petit déjeuner. Le fichier est au format Excel, Open Office sait le gérer. Nous réaliserons une petite analyse statistique pour illustrer notre propos. Mais notre principal objectif est bien de montrer le portage de la connexion sous Linux.

3 Installation de l'Add-On sous Open Office

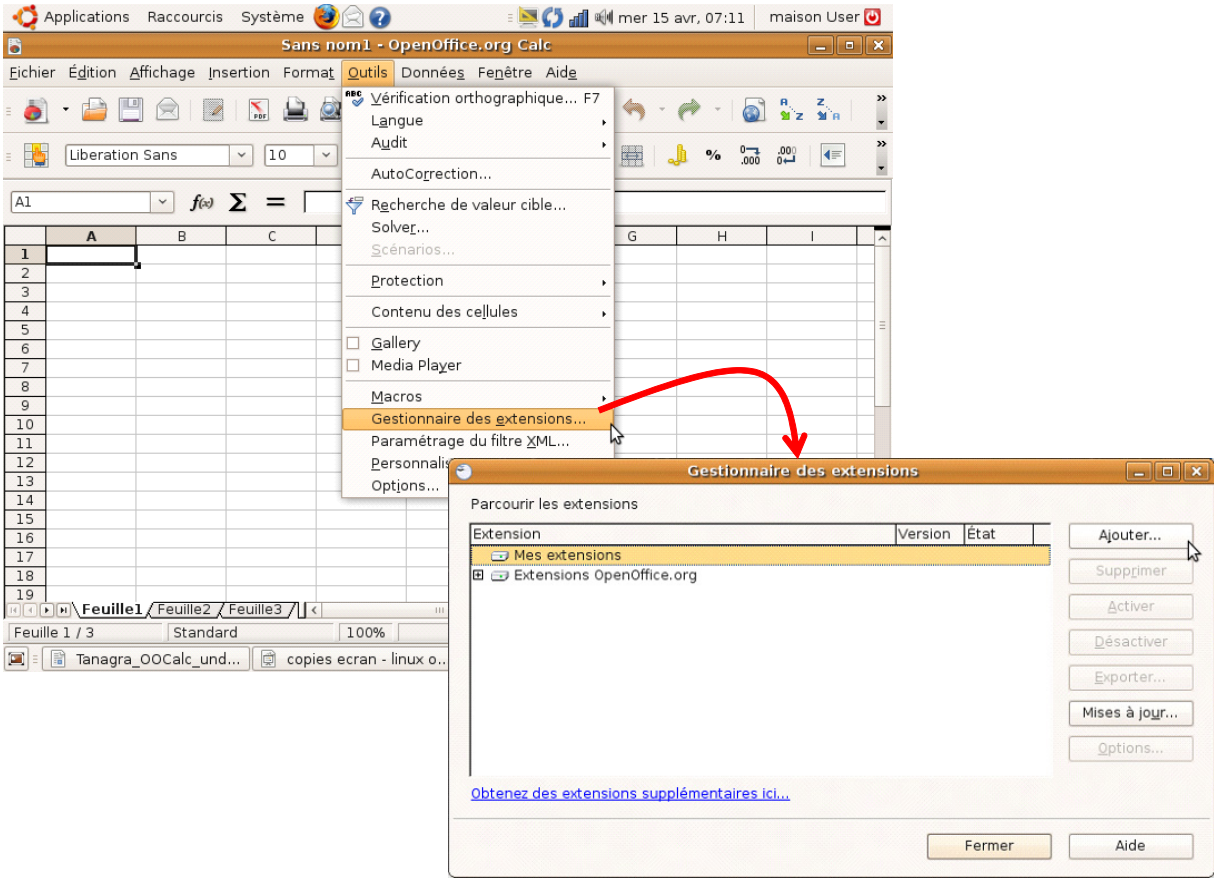
Nous devons avoir installé Tanagra sous Linux, puis nous être assurés qu'il est opérationnel via Wine (voir http://tutoriels-data-mining.blogspot.com/2009/01/tanagra-sous-linux.html)[1]. Nous pouvons alors démarrer Open Office Calc en actionnant le menu APPLICATIONS / BUREAUTIQUE / TABLEUR OPEN OFFICE.ORG de l'espace de travail.



Open Office Calc est démarré. Nous utilisons la version 2.4 dans ce didacticiel.

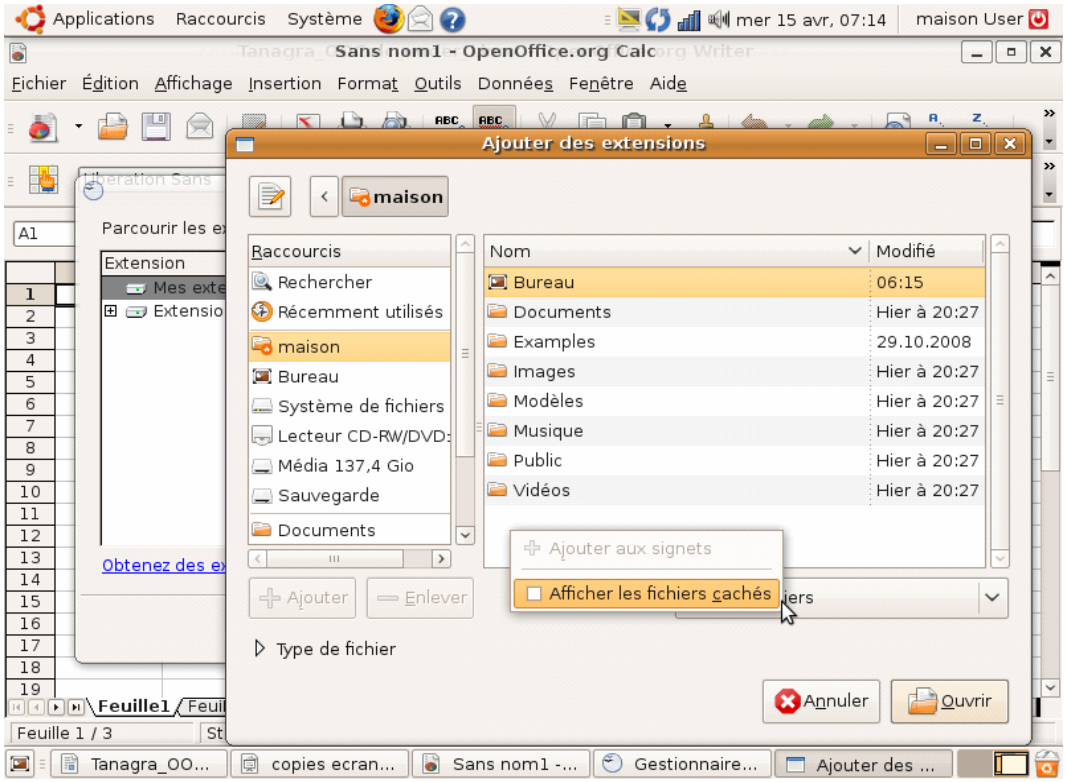
Pour installer l'Add-on, Nous actionnons le menu OUTILS / GESTIONNAIRE DES EXTENSIONS. Une boîte de dialogue apparaît, nous nous assurons que l'option « Mes Extensions » est bien sélectionnée dans la liste.

1 **ATTENTION, sur mon UBUNTU 8.10 Intrepid Ibex, seules les versions antérieures à 1.1.18 de WineHQ sont opérationnelles avec Tanagra**. J'ai rencontré quelques soucis avec la dernière mise à jour 1.1.19 (au 14 avril 2009), je n'arrive pas à savoir pourquoi. Consultez le site suivant pour « downgrader » (c'est possible, j'ai ainsi passé au crible les différentes versions allant de la 1.0.1 à la 1.1.18) votre installation de Wine en cas de problème (http://wine.budgetdedicated.com/archive/index.html).

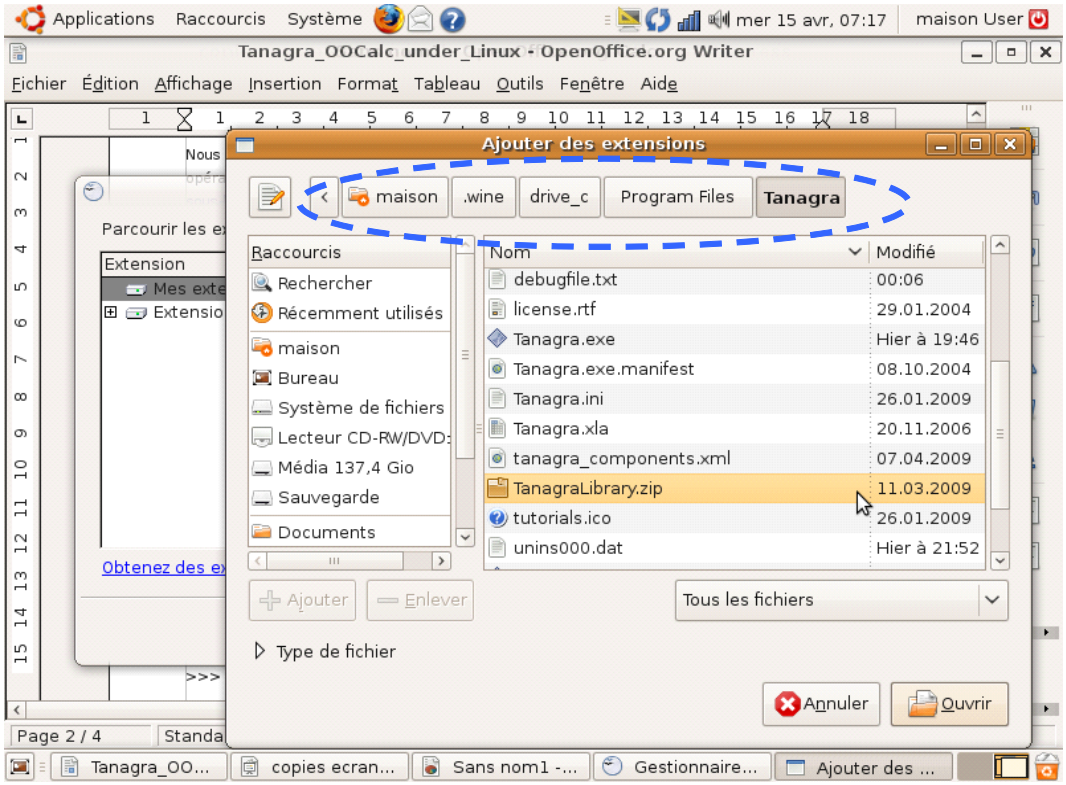


Nous cliquons sur le bouton AJOUTER. Le répertoire principal de l'utilisateur courant est affiché. Mais nous ne pouvons pas accéder pour l'instant au répertoire d'installation de Tanagra.

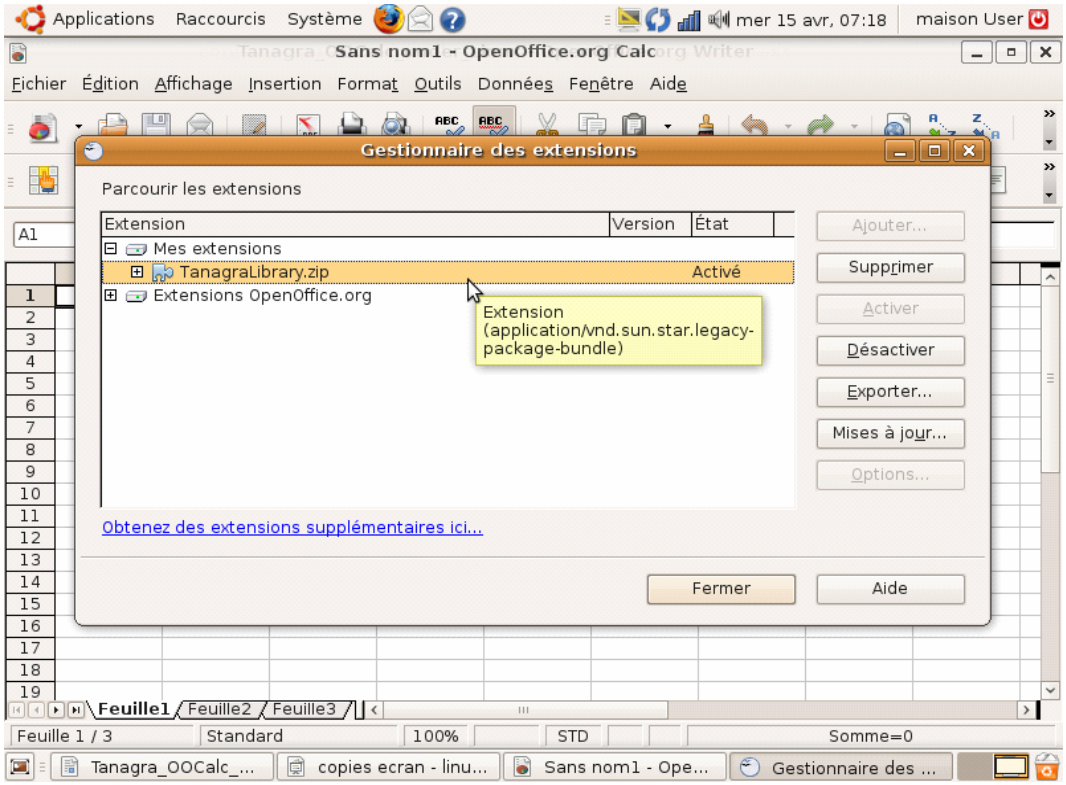
Pour cela, nous devons tout d'abord accéder au répertoire de Wine. Il faut le rendre visible. Nous actionnons le menu contextuel AFFICHER LES FICHIERS CACHES.



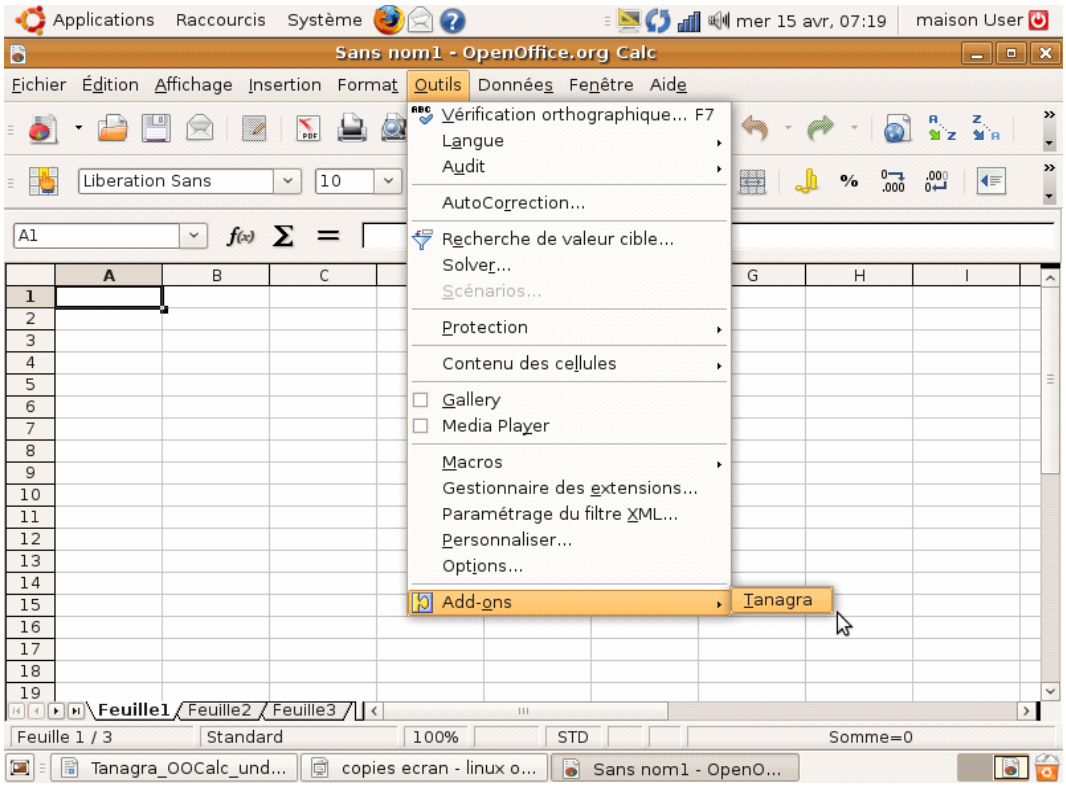
Nous pouvons dès lors atteindre le répertoire d'installation de Tanagra en suivant le chemin .WINE/DRIVE_C/PROGRAM FILES/TANAGRA. Nous sélectionnons le fichier « TanagraLibrary.zip ».



L'extension est maintenant installée. Nous fermons la boîte de dialogue de configuration des extensions.



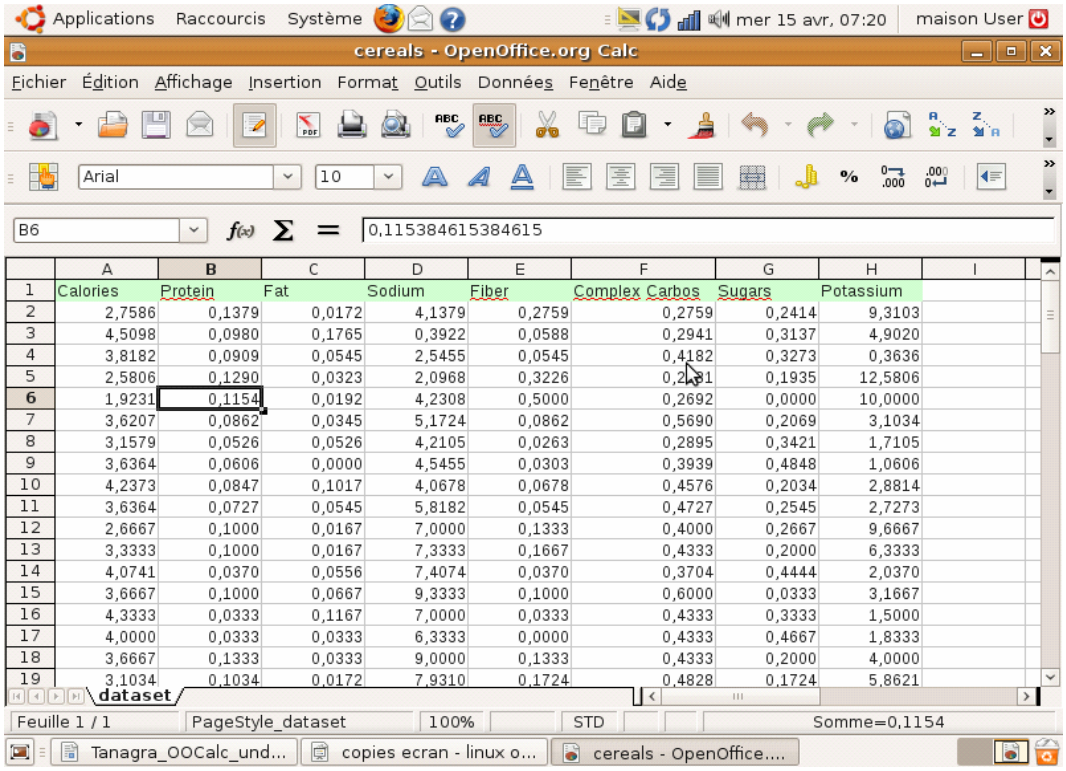
Pour l'instant, même si l'Add-on est installée, elle n'est pas utilisable. **Nous devons fermer Open Office Calc puis le re-démarrer**. Nous verrons alors apparaître une branche ADD-ONS associée à TANAGRA dans le menu OUTILS.



4 Analyse en composantes principales

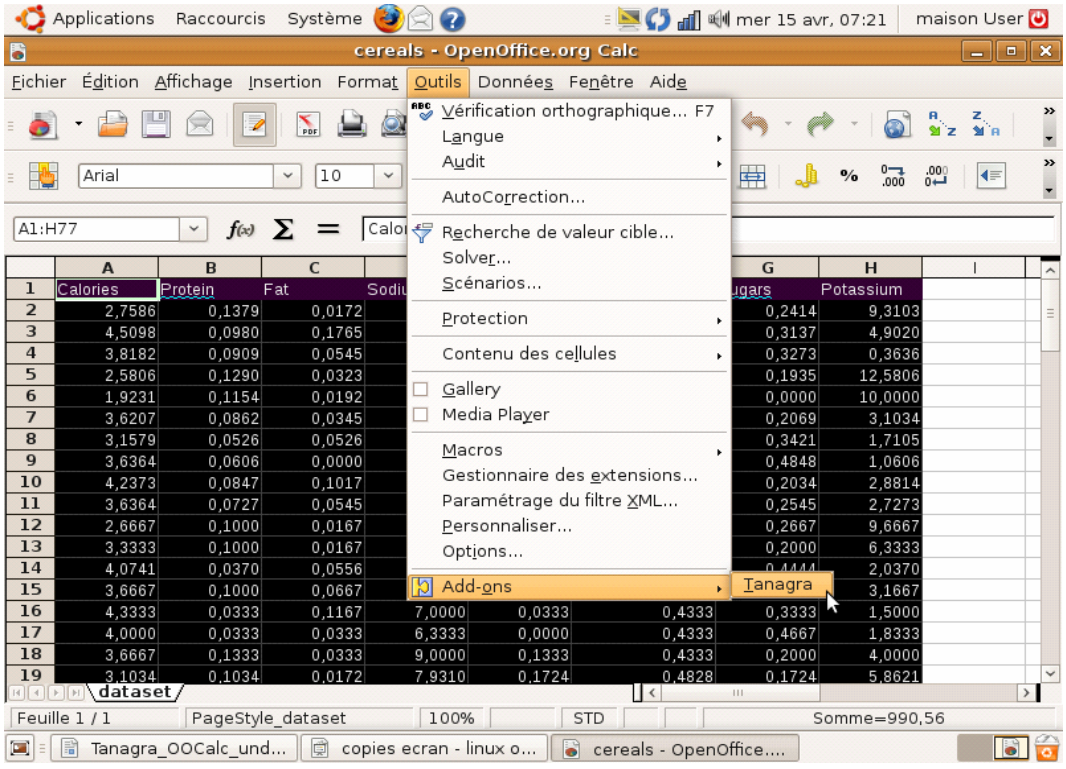
4.1 Démarrer Tanagra à partir de OOCalc

Nous pouvons charger le fichier CEREALS.XLS en cliquant sur le menu FICHIER / OUVRIR de OOCalc.



	A	B	C	D	E	F	G	H	I
1	Calories	Protein	Fat	Sodium	Fiber	Complex Carbos	Sugars	Potassium	
2	2,7586	0,1379	0,0172	4,1379	0,2759	0,2759	0,2414	9,3103	
3	4,5098	0,0980	0,1765	0,3922	0,0588	0,2941	0,3137	4,9020	
4	3,8182	0,0909	0,0545	2,5455	0,0545	0,4182	0,3273	0,3636	
5	2,5806	0,1290	0,0323	2,0968	0,3226	0,2581	0,1935	12,5806	
6	1,9231	0,1154	0,0192	4,2308	0,5000	0,2692	0,0000	10,0000	
7	3,6207	0,0862	0,0345	5,1724	0,0862	0,5690	0,2069	3,1034	
8	3,1579	0,0526	0,0526	4,2105	0,0263	0,2895	0,3421	1,7105	
9	3,6364	0,0606	0,0000	4,5455	0,0303	0,3939	0,4848	1,0606	
10	4,2373	0,0847	0,1017	4,0678	0,0678	0,4576	0,2034	2,8814	
11	3,6364	0,0727	0,0545	5,8182	0,0545	0,4727	0,2545	2,7273	
12	2,6667	0,1000	0,0167	7,0000	0,1333	0,4000	0,2667	9,6667	
13	3,3333	0,1000	0,0167	7,3333	0,1667	0,4333	0,2000	6,3333	
14	4,0741	0,0370	0,0556	7,4074	0,0370	0,3704	0,4444	2,0370	
15	3,6667	0,1000	0,0667	9,3333	0,1000	0,6000	0,0333	3,1667	
16	4,3333	0,0333	0,1167	7,0000	0,0333	0,4333	0,3333	1,5000	
17	4,0000	0,0333	0,0333	6,3333	0,0000	0,4333	0,4667	1,8333	
18	3,6667	0,1333	0,0333	9,0000	0,1333	0,4333	0,2000	4,0000	
19	3,1034	0,1034	0,0172	7,9310	0,1724	0,4828	0,1724	5,8621	

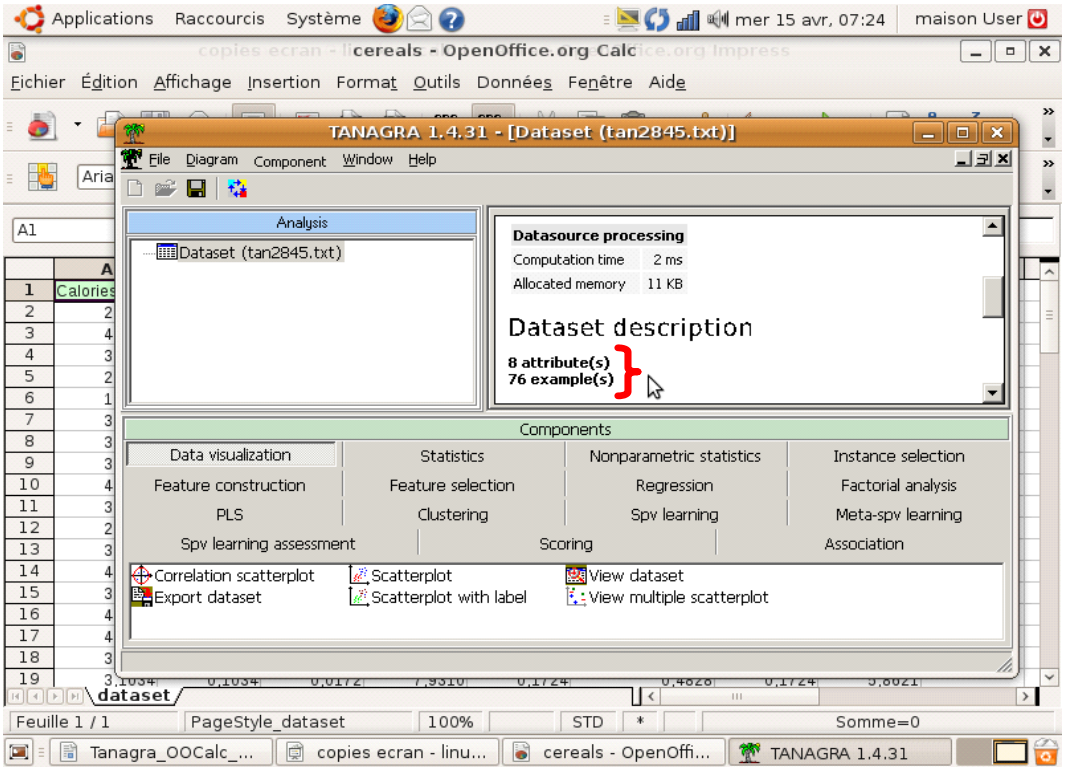
Nous sélectionnons la plage de données, y compris la première colonne correspondant aux noms de variables. Nous actionnons le menu OUTILS / ADD-ONS / TANAGRA.



	A	B	C	G	H
1	Calories	Protein	Fat	ugars	Potassium
2	2,7586	0,1379	0,0172	0,2414	9,3103
3	4,5098	0,0980	0,1765	0,3137	4,9020
4	3,8182	0,0909	0,0545	0,3273	0,3636
5	2,5806	0,1290	0,0323	0,1935	12,5806
6	1,9231	0,1154	0,0192	0,0000	10,0000
7	3,6207	0,0862	0,0345	0,2069	3,1034
8	3,1579	0,0526	0,0526	0,3421	1,7105
9	3,6364	0,0606	0,0000	0,4848	1,0606
10	4,2373	0,0847	0,1017	0,2034	2,8814
11	3,6364	0,0727	0,0545	0,2545	2,7273
12	2,6667	0,1000	0,0167	0,2667	9,6667
13	3,3333	0,1000	0,0167	0,2000	6,3333
14	4,0741	0,0370	0,0556	0,4444	2,0370
15	3,6667	0,1000	0,0667	0,0333	3,1667
16	4,3333	0,0333	0,1167	7,0000 0,0333 0,4333 0,3333	1,5000
17	4,0000	0,0333	0,0333	6,3333 0,0000 0,4333 0,4667	1,8333
18	3,6667	0,1333	0,0333	9,0000 0,1333 0,4333 0,2000	4,0000
19	3,1034	0,1034	0,0172	7,9310 0,1724 0,4828 0,1724	5,8621

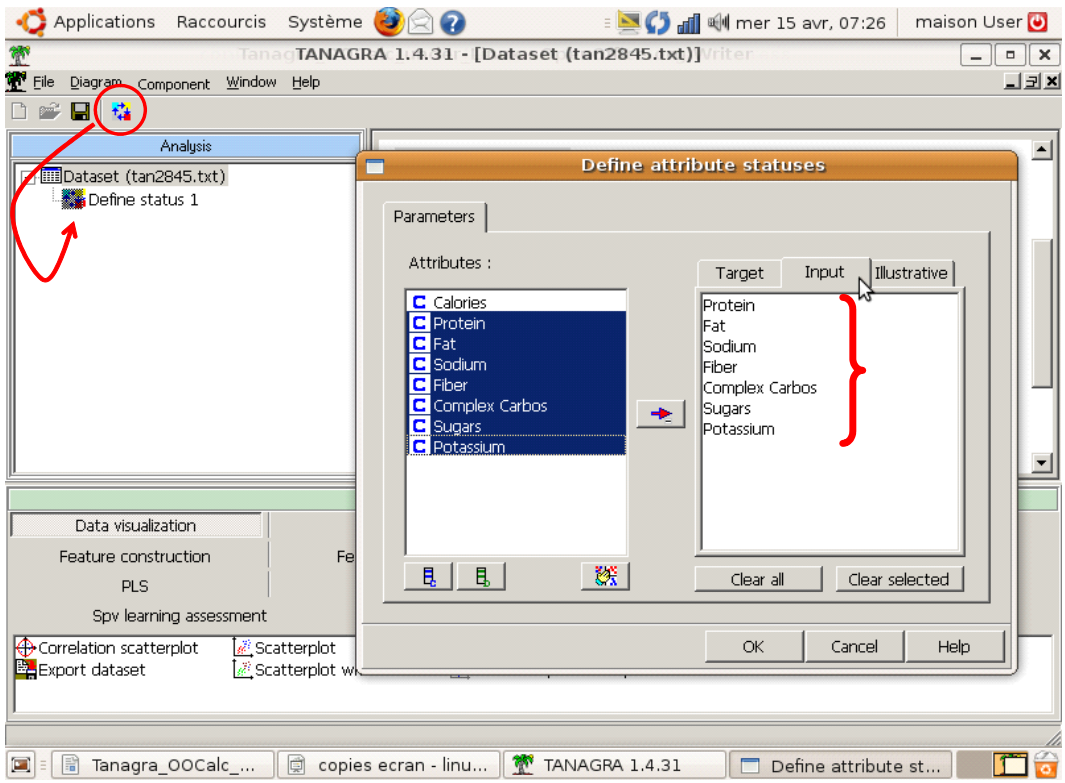
Tanagra est automatiquement démarré, un diagramme est créé et les données sont chargées. Le

fichier comporte 8 variables et 76 observations.

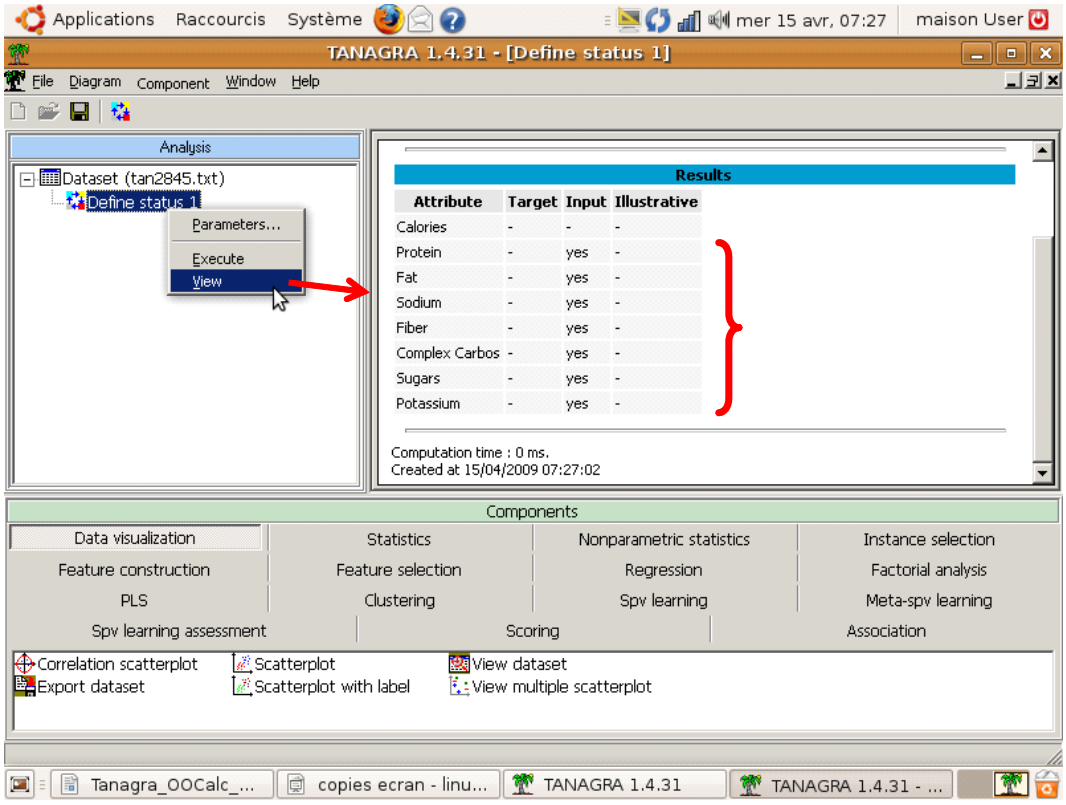


4.2 ACP sous Tanagra

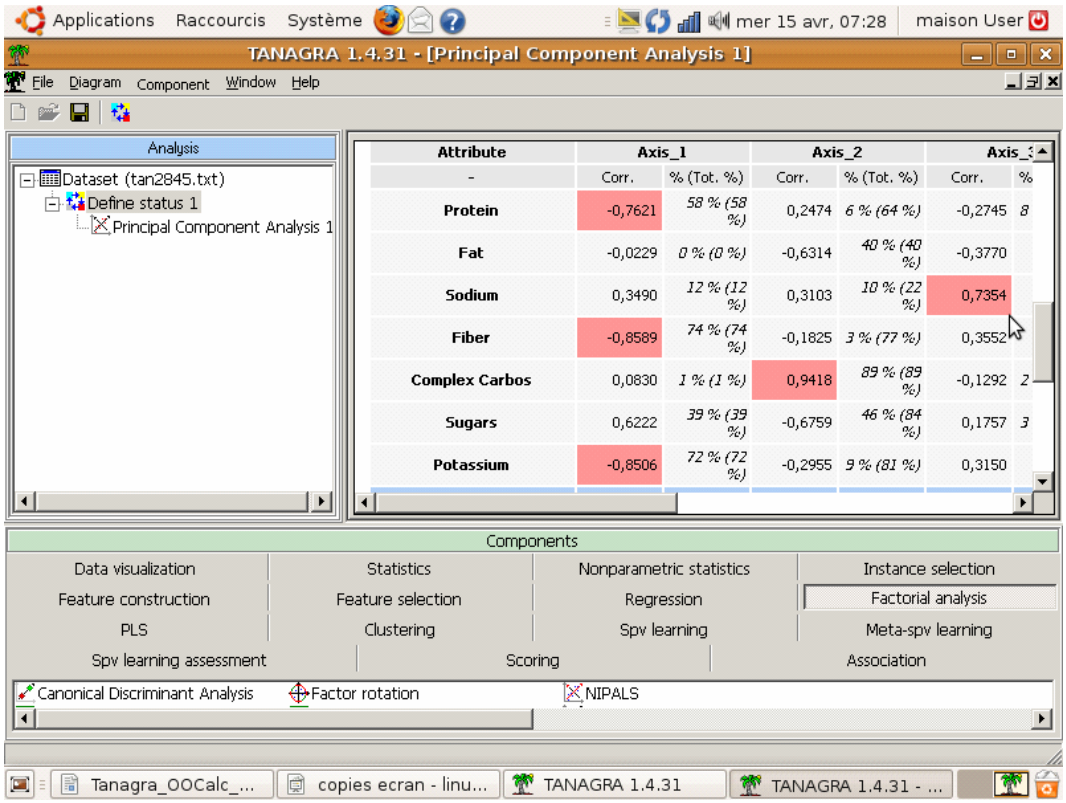
Pour réaliser une ACP, nous devons d'abord indiquer au logiciel les variables actives. Nous utilisons le composant DEFINE STATUS vie le raccourci dans la barre d'outils. Nous plaçons en INPUT toutes les variables, sauf CALORIES.



Nous cliquons sur VIEW pour nous assurer que les spécifications ont bien été prises en compte.



Nous insérons ensuite le composant PRINCIPAL COMPONENT ANALYSIS situé dans l'onglet FACTORIAL ANALYSIS. Nous cliquons directement sur VIEW pour obtenir les résultats.

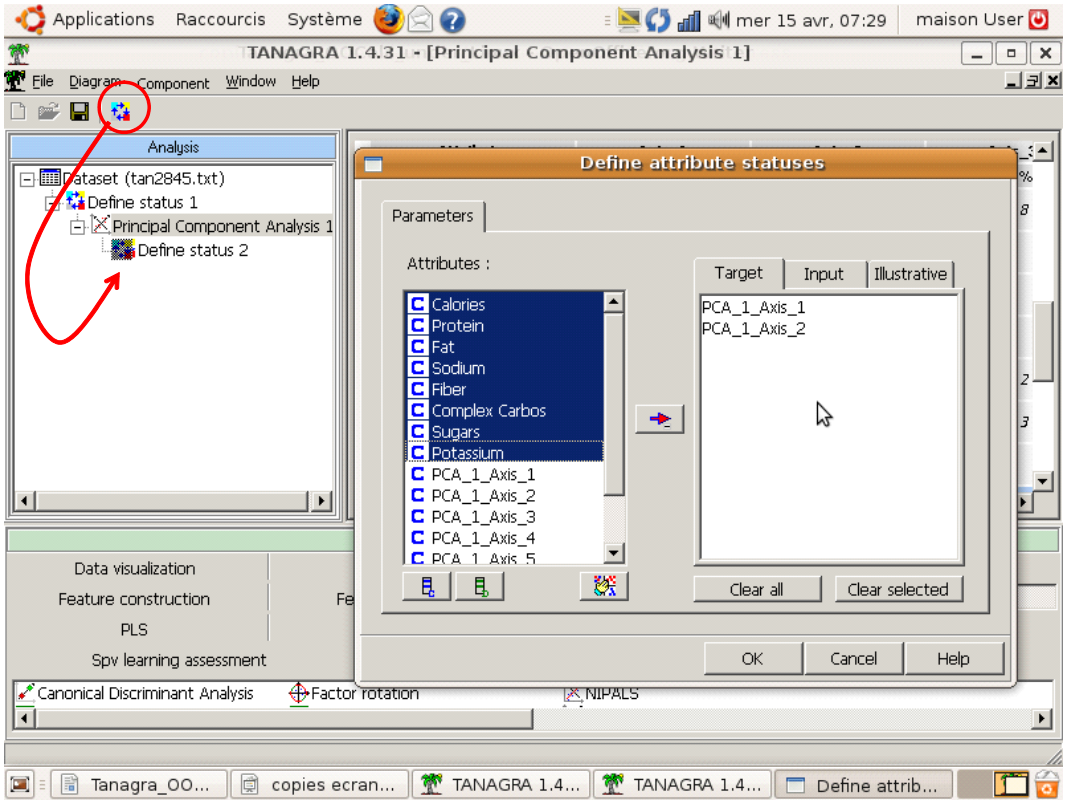


Les 3 premiers axes traduisent 80% de l'information disponible. De manière très succincte, nous

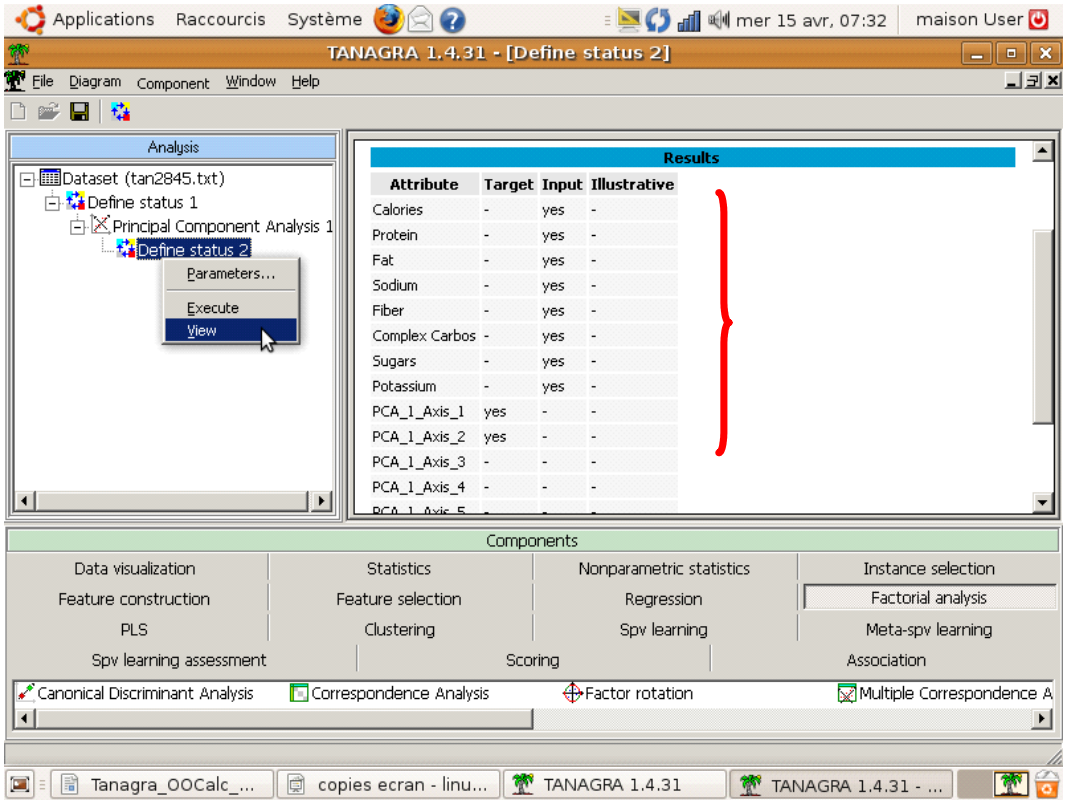
dirons que : (1) sur le premier axe se démarque avant tout les produits à la fois riches en fibres, protéines et potassium, elles sont opposées aux produits riches en sucres rapides ; (2) sur le second axe s'opposent les produits riches en sucres rapides et en graisses, d'une part, et les produits riches en sucres lents (complexes carbos), d'autre part; (3) le sel (sodium) semble une dernière caractéristique marquante, peu liée avec les deux premières dimensions.

4.3 Cercle des corrélations incluant la variable illustrative

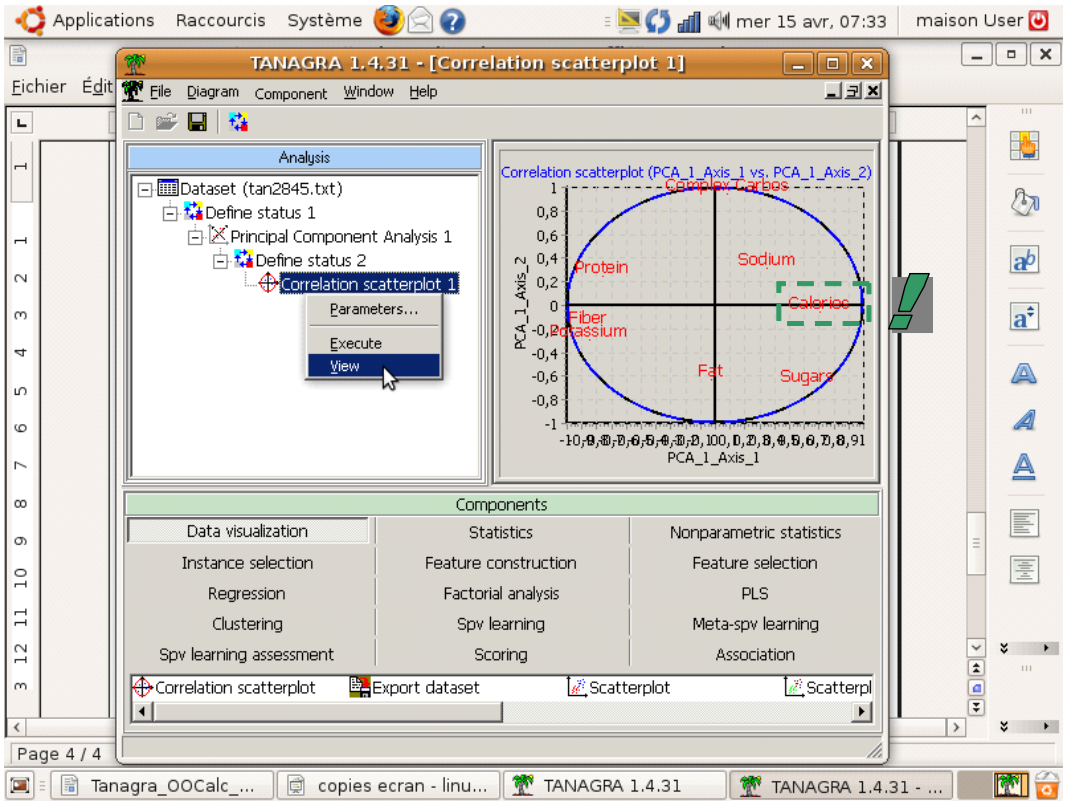
Voyons comment se situe la variable CALORIES par rapport aux deux premières caractéristiques. Pour cela, nous utilisons le cercle des corrélations. Nous insérons le composant DEFINE STATUS, nous plaçons en TARGET les deux premiers axes élaborés automatiquement par l'ACP (PCA_1_Axis_1 et PCA_1_Axis_2). En INPUT, nous plaçons toutes les variables, **y compris la variable illustrative CALORIES**.



Cliquons sur VIEW toujours pour vérifier la sélection.



Nous pouvons insérer maintenant le composant CORRELATION SCATTERPLOT (onglet VISUALIZATION).



Nous pouvons situer dès lors le positionnement de CALORIES par rapport aux autres variables de l'étude.

5 Conclusion

OOCalc présente un triple avantage : il propose des fonctionnalités tout à fait comparables à celles d'Excel, la Rolls des tableurs de ces quinze dernières années ; il est totalement gratuit ; il est opérationnel sous différents systèmes d'exploitation.

Faciliter le transfert des données de OOCalc vers Tanagra via le système des Add-Ons, sous Windows, et maintenant sous Linux, ne peut que faciliter la vie des très nombreux Data Miners qui utilisent principalement un tableur pour manipuler leurs données. Et ils sont nombreux (http://www.kdnuggets.com/polls/2008/tools-languages-used-data-cleaning.htm).