Objectif

Montrer l'utilisation de la macro complémentaire TANAGRA.XLA dans le tableur EXCEL.

De nombreux utilisateurs s'appuient sur EXCEL pour la gestion de leurs données. C'est un outil relativement efficace et largement diffusé. Calculer des statistiques intermédiaires, créer de nouvelles variables, sont des opérations qui peuvent être réalisées très simplement, sans connaissances préalables mirobolantes sur la manipulation de données. L'enjeu par la suite est de pouvoir faire le pont entre le tableur, un environnement familier des utilisateurs, vers un logiciel de Data Mining, moins courant mais absolument nécessaire dès lors que l'on veut réaliser des études plus sophistiquées.

La première solution consiste à importer les fichiers au format XLS. Proposée par de nombreux logiciels, cette option comporte un inconvénient : une fois le fichier importé, nous ne disposons plus des outils de manipulation de données d'EXCEL. Plus ennuyeux pour les développeurs, il faut pouvoir suivre les différentes versions du format XLS. Elles ne sont pas toujours disponibles. Dans TANAGRA, nous avons la garantie que l'importation fonctionne pour les versions 97, 2000 et 2003 d'EXCEL. Au-delà, tout dépend de l'évolution du format.

Une autre solution consiste à programmer les méthodes de Data Mining sous forme de macros complémentaires. Elles deviennent donc de nouvelles fonctionnalités du tableur. Plusieurs logiciels commerciaux s'appuient sur ce schéma. Malheureusement, étant principalement pilotés par menu, ces logiciels ne nous permettent pas d'enchaîner automatiquement les traitements, ni de disposer d'une trace de la succession d'opérations réalisées. Pour le développeur, même si les algorithmes de traitement peuvent être implémentés dans des DLL compilées, il faut quand même beaucoup d'investissement en VBA pour la définition des interfaces de sélection des données, le paramétrage des méthodes, etc.

Enfin, la dernière solution recensée serait d'intégrer le tableur comme une partie du logiciel de Data Mining. Nous avons exploré cette option via la technologie OLE. L'idée semble viable. Nous l'avons mise en œuvre (voir XL-SIPINA, http://eric.univ-lyon2.fr/~ricco/sipina.html). Mais, à cause de la technologie utilisée, peut-être aussi parce que nous la maîtrisons de manière approximative, le système obtenu est relativement lent et peu fiable. Ne voulant pas investir trop de temps de développement dans ce qui n'est qu'un exercice de style, nous n'avons pas voulu aller plus loin.

Bref, la jonction entre EXCEL et TANAGRA restait à ce jour une question délicate. Certes, il était déjà possible d'importer des fichiers XLS dans TANAGRA. Mais seule la première feuille de calcul était accessible. De plus, l'obligation de fermer EXCEL, qui verrouille le fichier, avant d'importer les données était une source d'erreur fréquente, sans parler des incertitudes concernant les versions de fichiers.

Nous avons donc ajouté une nouvelle fonctionnalité faisant le pont entre EXCEL et TANAGRA, indépendamment de la version du fichier XLS et sans avoir à fermer la session de travail sous EXCEL. Toujours en accord avec notre philosophie, nous avons opté pour une approche simplifiée à l'extrême. Elle passe par une **macro complémentaire (TANAGRA.XLA)**, dont le rôle consiste à définir la sélection de l'utilisateur, puis exécuter automatiquement TANAGRA. La transmission des données, qui est la phase critique, emprunte un canal inédit : le presse-papiers. Les expérimentations montrent que ce dispositif est fiable et performant. L'utilisateur, qui travaille sous EXCEL, peut à tout moment lancer une session de DATA MINING en activant un nouveau menu. Toutes les opérations de préparation et de transfert sont transparentes. Il se retrouve instantanément dans l'environnement de TANAGRA avec un nouveau diagramme. Il dispose alors de toutes les fonctionnalités d'un logiciel de Data Mining, notamment la possibilité d'enchaîner les traitements en les traçant sous forme de diagramme.

Dans ce didacticiel, nous montrons comment installer cette nouvelle macro complémentaire et réaliser un traitement sur un fichier de données. Cette fonctionnalité est **disponible depuis la version 1.4.11 de TANAGRA**.

Installer la macro complémentaire dans EXCEL

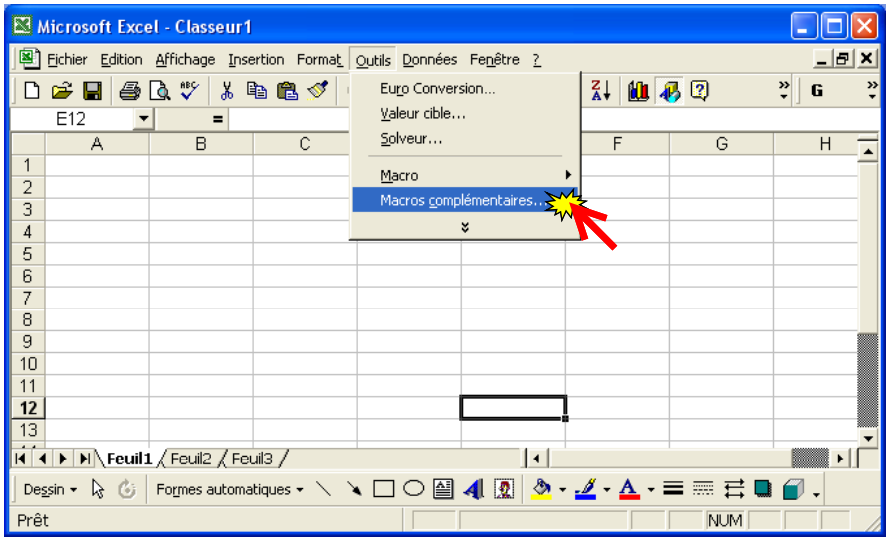
Vérifier la présence de la macro complémentaire

Tout d'abord, nous devons nous assurer que la version installée de TANAGRA possède bien la fonctionnalité voulue. Le plus simple est de vérifier la présence de la macro complémentaire **TANAGRA.XLA** dans le répertoire d'installation du logiciel (la plupart du temps, il s'agira de « *c:\program files\tanagra* »).

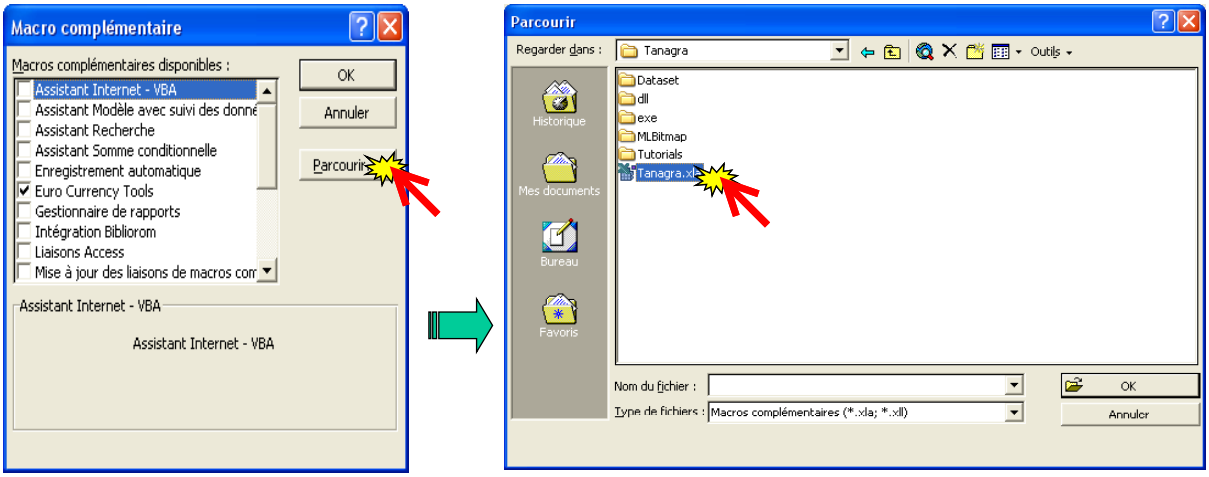
Il est important de **ne pas déplacer ce fichier**, il cherchera l'exécutable lors de son activation.

Installer la macro complémentaire dans EXCEL

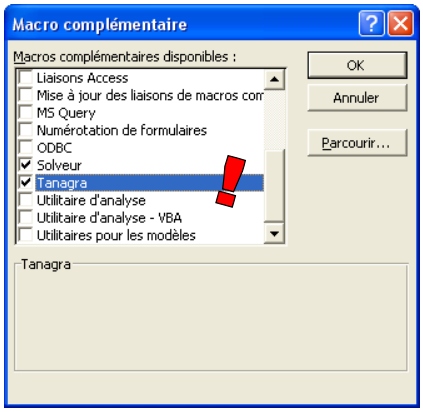
L'étape suivante consiste à lancer le tableur EXCEL. Pour installer la macro complémentaire, nous activons le menu OUTILS/MACRO COMPLEMENTAIRES.



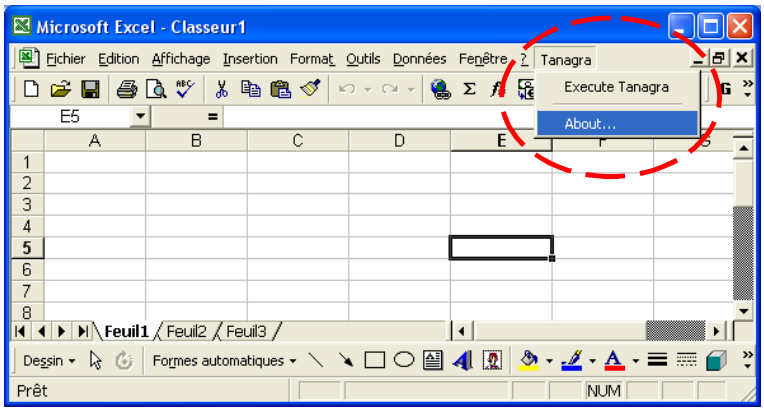
Une boîte de dialogue apparaît, nous devons alors chercher le fichier TANAGRA.XLA dans le répertoire d'installation de TANAGRA.



La macro complémentaire est alors chargée dans EXCEL, nous devons veiller à ce qu'elle soit activée.



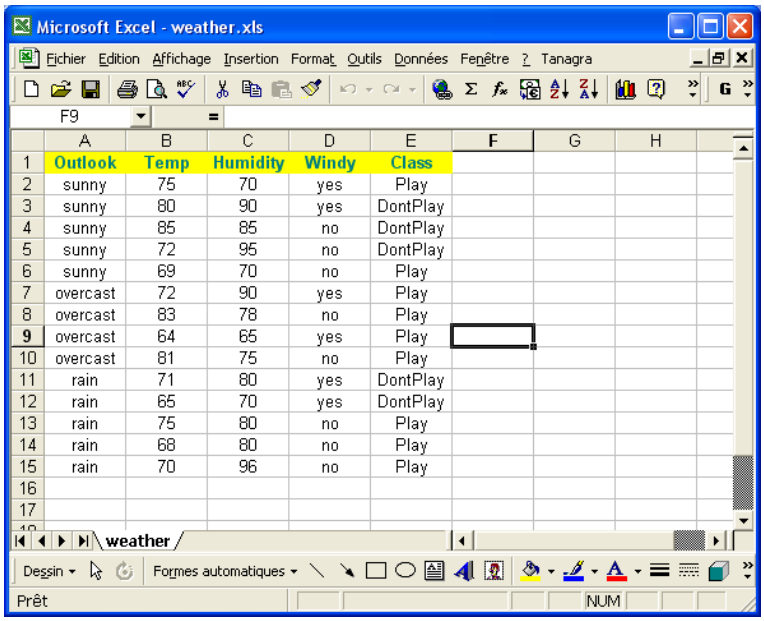
Après avoir validé, nous constatons qu'un nouveau menu est disponible dans EXCEL.



A partir de maintenant, tant que nous n'avons pas désactivé la macro complémentaire, **ce nouveau menu sera disponible à chaque démarrage du tableur EXCEL**.

Travailler sur un fichier

Pour illustrer le fonctionnement du package, nous chargeons le fichier WEATHER.XLS de Quinlan (1993).

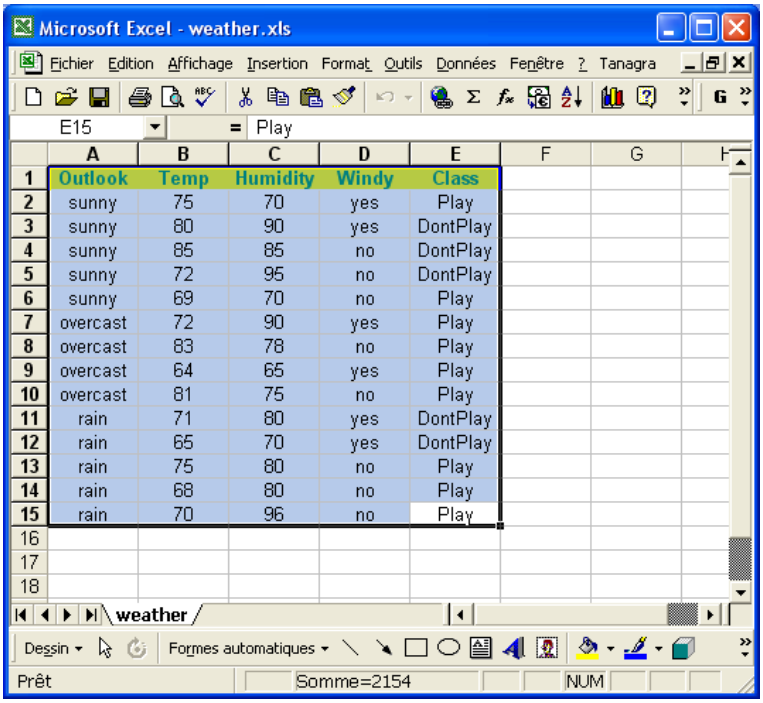


	A	B	C	D	E
1	Outlook	Temp	Humidity	Windy	Class
2	sunny	75	70	yes	Play
3	sunny	80	90	yes	DontPlay
4	sunny	85	85	no	DontPlay
5	sunny	72	95	no	DontPlay
6	sunny	69	70	no	Play
7	overcast	72	90	yes	Play
8	overcast	83	78	no	Play
9	overcast	64	65	yes	Play
10	overcast	81	75	no	Play
11	rain	71	80	yes	DontPlay
12	rain	65	70	yes	DontPlay
13	rain	75	80	no	Play
14	rain	68	80	no	Play
15	rain	70	96	no	Play

Sélectionner les données

Avant de lancer la macro TANAGRA, il est conseillé de **sélectionner les données de travail**. Nous pouvons modifier cette sélection par la suite mais il est plus facile de le faire au préalable.

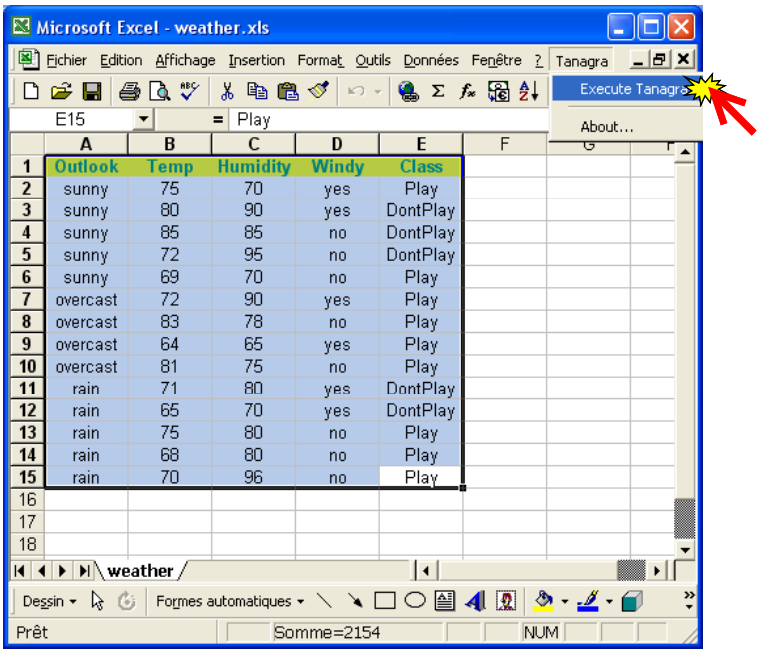
Attention, **la première ligne de la sélection doit correspondre au nom des attributs**. Le typage utilise une règle très simple : si la première donnée de la colonne (la deuxième ligne de la sélection) est numérique, la variable est considérée continue ; elle est définie catégorielle dans le cas contraire.



	A	B	C	D	E
1	Outlook	Temp	Humidity	Windy	Class
2	sunny	75	70	yes	Play
3	sunny	80	90	yes	DontPlay
4	sunny	85	85	no	DontPlay
5	sunny	72	95	no	DontPlay
6	sunny	69	70	no	Play
7	overcast	72	90	yes	Play
8	overcast	83	78	no	Play
9	overcast	64	65	yes	Play
10	overcast	81	75	no	Play
11	rain	71	80	yes	DontPlay
12	rain	65	70	yes	DontPlay
13	rain	75	80	no	Play
14	rain	68	80	no	Play
15	rain	70	96	no	Play

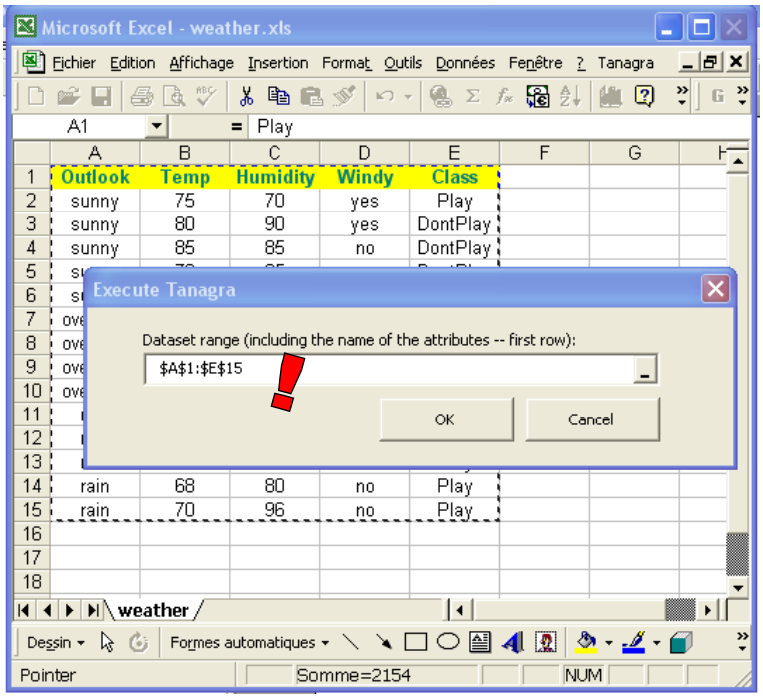
Activer le menu TANAGRA / EXECUTE TANAGRA

Nous activons alors le nouveau menu TANAGRA / EXECUTE TANAGRA dans EXCEL.

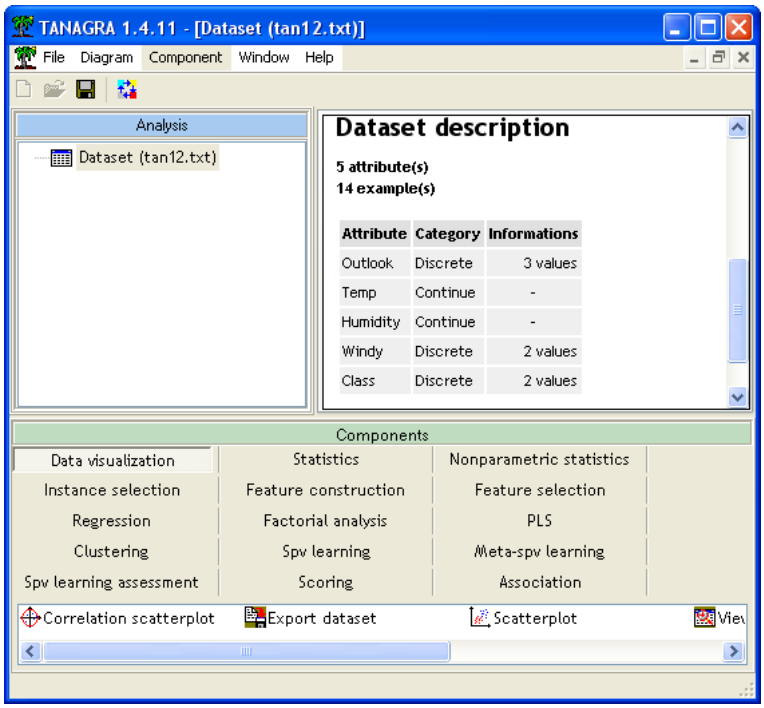


	A	B	C	D	E
1	Outlook	Temp	Humidity	Windy	Class
2	sunny	75	70	yes	Play
3	sunny	80	90	yes	DontPlay
4	sunny	85	85	no	DontPlay
5	sunny	72	95	no	DontPlay
6	sunny	69	70	no	Play
7	overcast	72	90	yes	Play
8	overcast	83	78	no	Play
9	overcast	64	65	yes	Play
10	overcast	81	75	no	Play
11	rain	71	80	yes	DontPlay
12	rain	65	70	yes	DontPlay
13	rain	75	80	no	Play
14	rain	68	80	no	Play
15	rain	70	96	no	Play

Une boîte de dialogue apparaît, elle vous permet de vérifier si la sélection convient, et de la corriger le cas échéant.



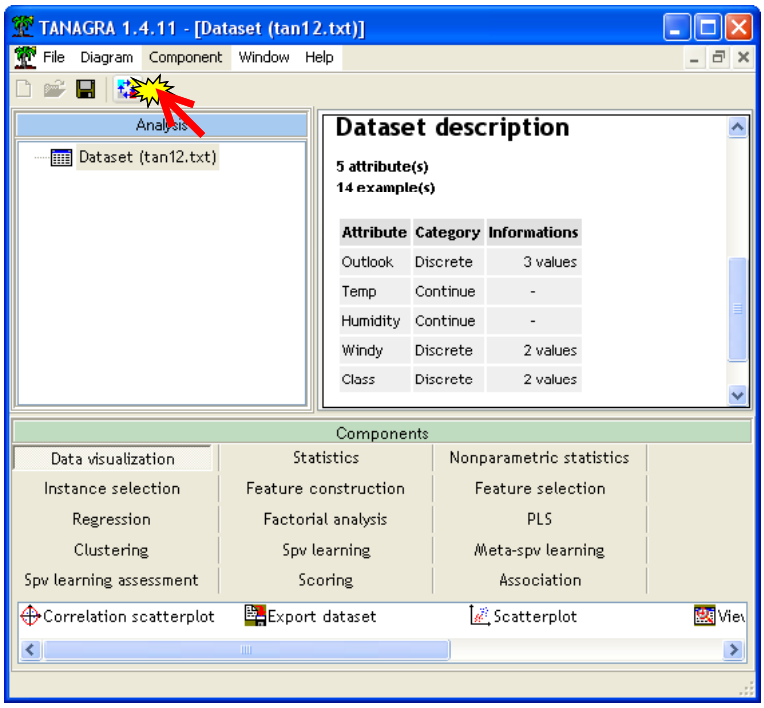
Tout va bien dans notre exemple, il ne nous reste plus alors qu'à **valider la manipulation en cliquant sur le bouton OK**. Le logiciel TANAGRA est alors automatiquement exécuté avec les données sélectionnées.



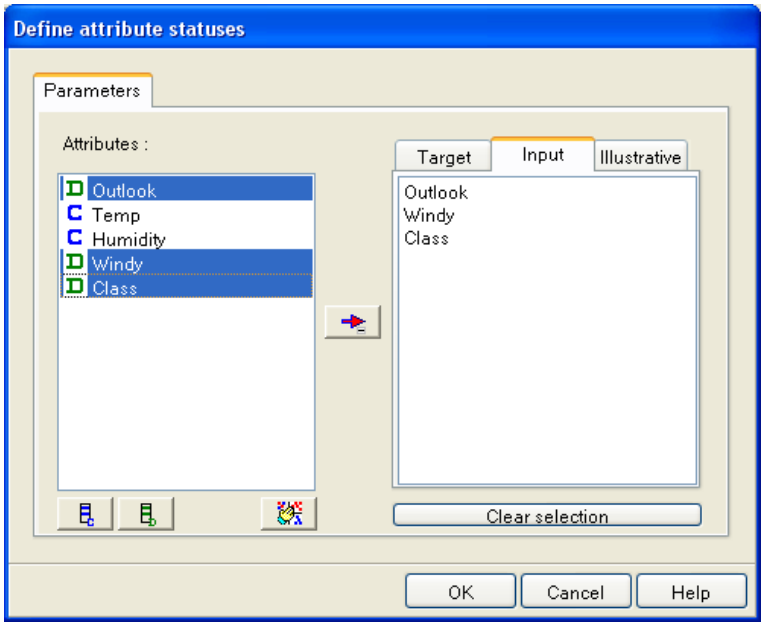
Nous constatons que les données ont été exportées (14 observations et 5 attributs). Les variables ont été automatiquement typées.

Travailler dans TANAGRA

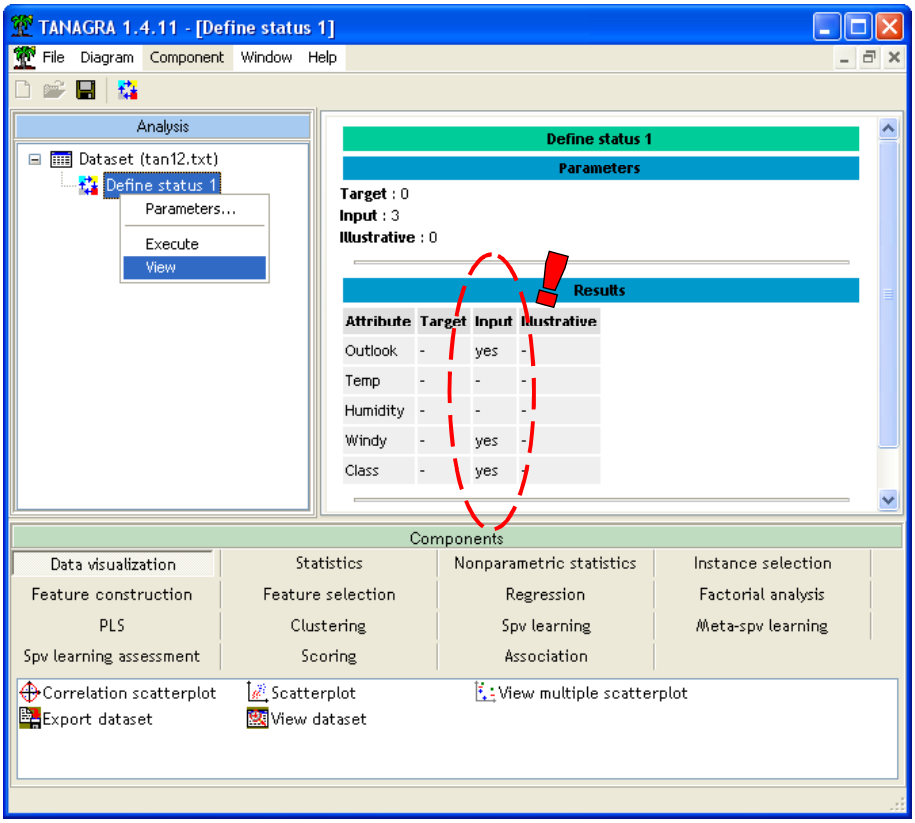
Nous voulons calculer quelques statistiques descriptives sur les variables discrètes. Dans un premier temps, nous devons **spécifier les variables de travail**. Le composant DEFINE STATUS est tout indiqué pour cela, nous pouvons le placer automatiquement dans le diagramme en utilisant le raccourci dans la barre d'outil.



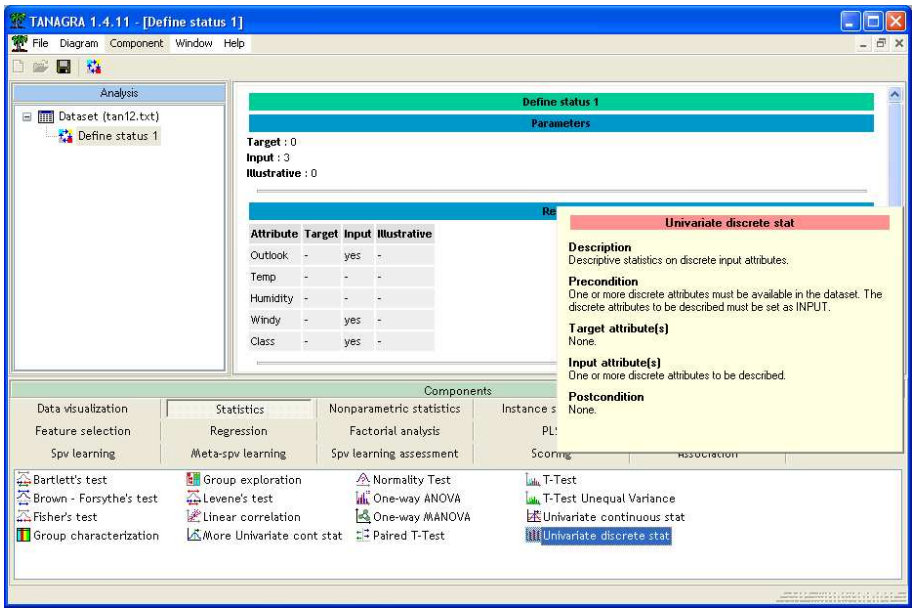
Le composant est ajouté dans le diagramme et la boîte de paramétrage apparaît. Nous plaçons toutes les variables discrètes en INPUT et nous validons.



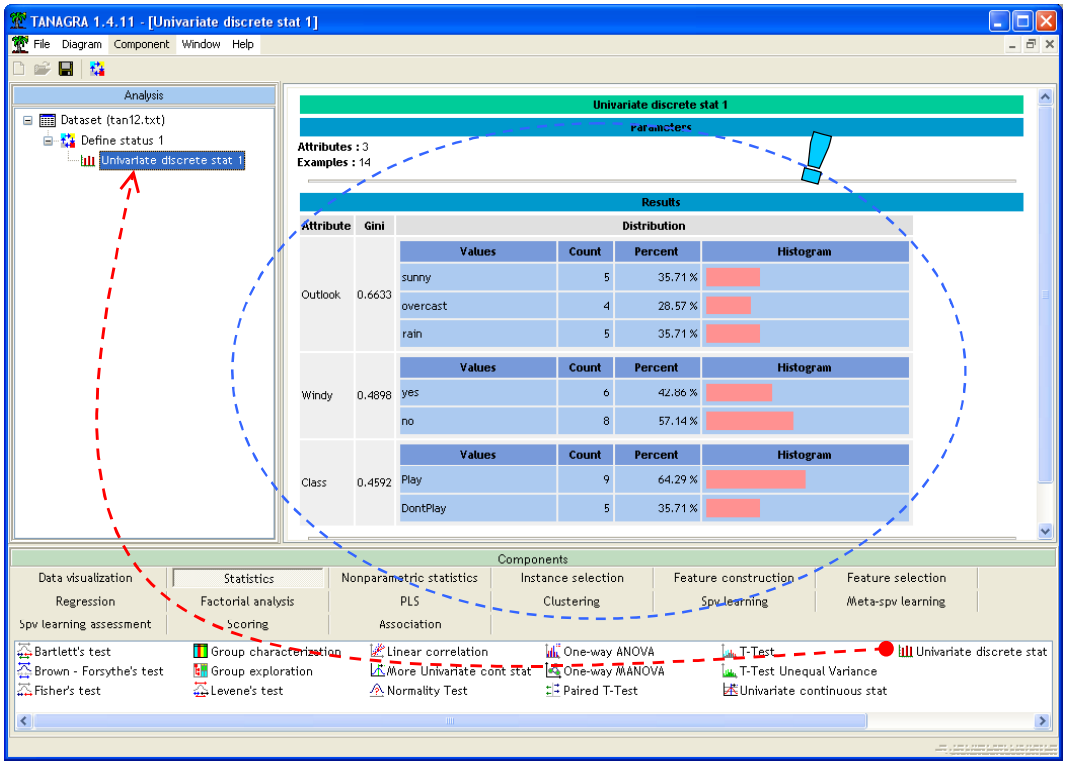
Pour visualiser les résultats de la manipulation, nous cliquons sur le menu contextuel VIEW du composant. La sélection est clairement indiquée.



Enfin, dernière étape dans TANAGRA, nous **insérons le composant de calcul dans le diagramme**. Dans notre exemple, il s'agit du composant UNIVARIATE DISCRETE STAT situé dans l'onglet STATISTICS.

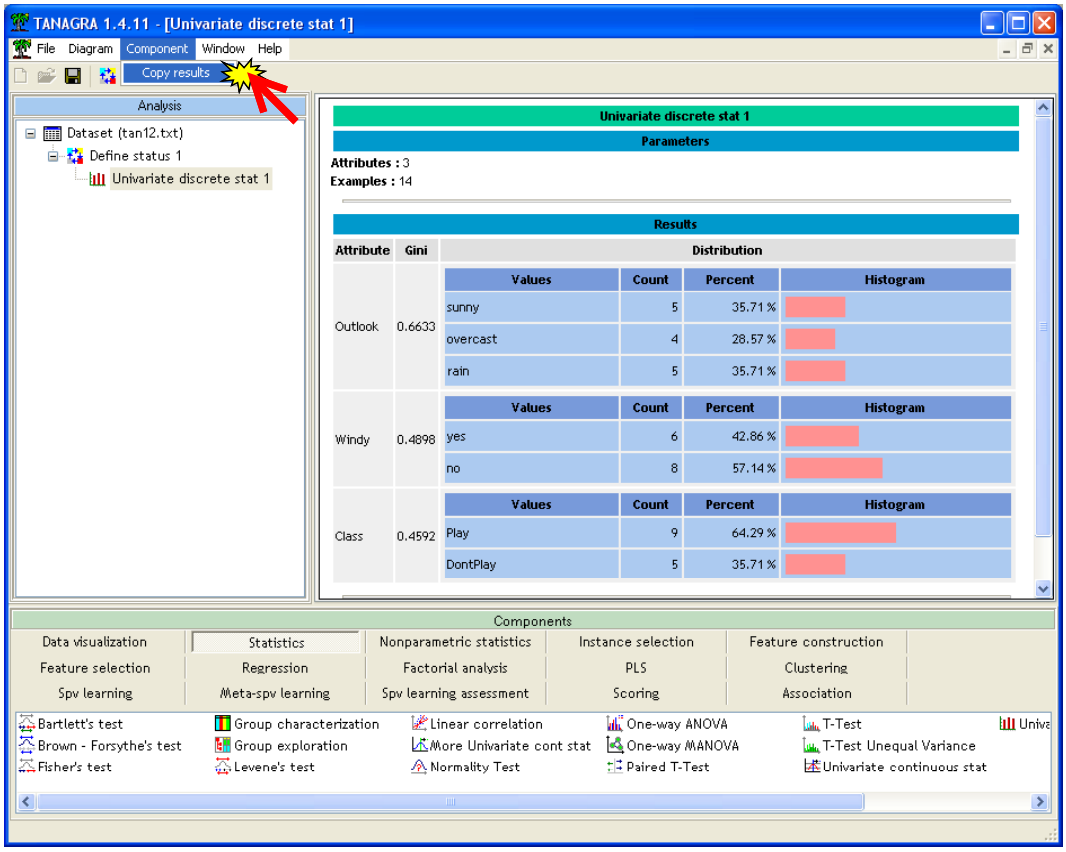


Nous le sélectionnons, puis nous le plaçons sur le composant DEFINE STATUS. Les résultats sont affichés lorsque nous cliquons sur le menu contextuel VIEW.

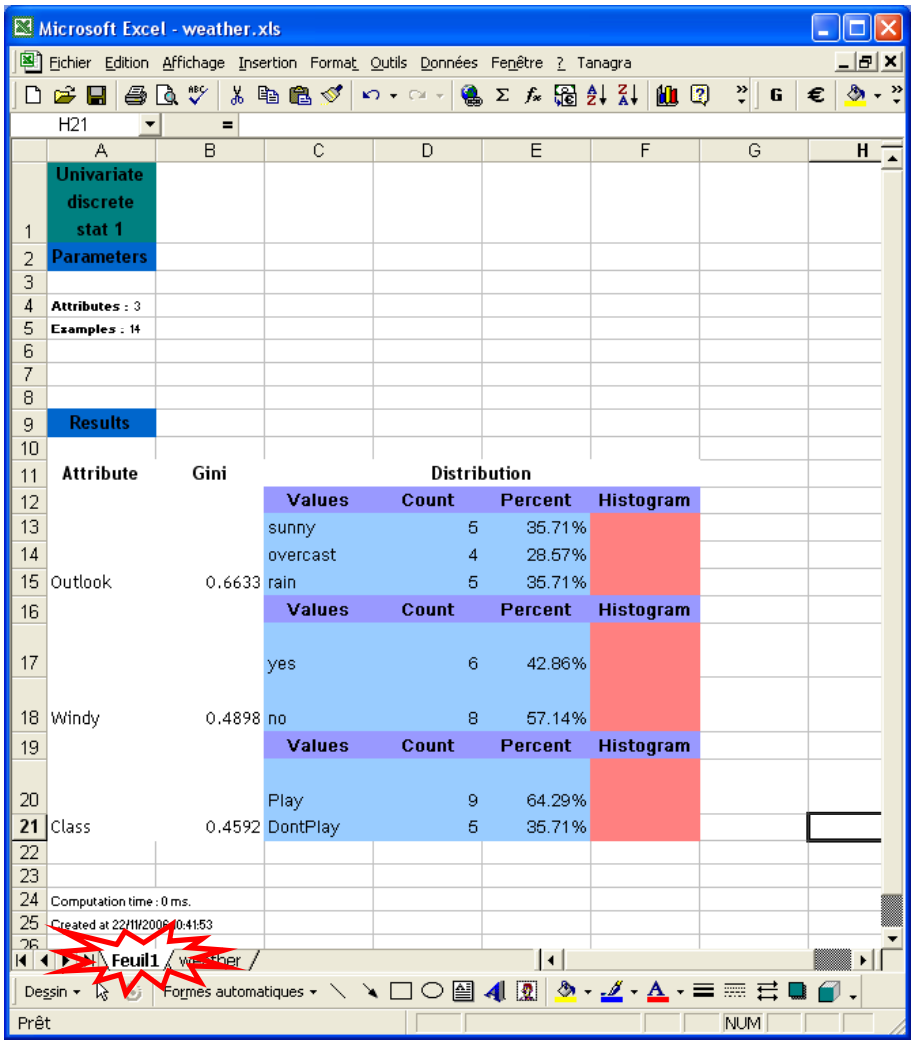


Récupérer les résultats dans EXCEL

A tout moment, il est possible de récupérer les résultats, au format HTML, dans une feuille du tableur EXCEL. Pour ce faire, nous activons le menu COMPONENT / COPY RESULTS.



Puis dans le classeur EXCEL, après avoir ajouté une feuille de calcul dans le classeur courant, nous collons les résultats. Selon le cas, le formatage est plus ou moins respecté mais l'essentiel y est.



Conclusion -- Evaluation des performances

Une des questions clés de cette nouvelle fonctionnalité est la rapidité du passage d'EXCEL à TANAGRA. Si l'utilisation du presse-papiers WINDOWS est lente et trop gourmande en ressources, les aspects de temps de calcul prennent le pas sur le côté pratique de la chose. Dans ce cas, il serait plus indiqué d'exporter le fichier EXCEL et de l'importer dans TANAGRA via le dispositif habituel.

Pour évaluer cela, et surtout pour se donner une idée de la taille critique à partir de laquelle il devient plus judicieux de passer par un système d'exportation/importation de fichier, nous avons testé notre procédé sur plusieurs fichiers, dont le fichier SHUTTLE.XLS comportant 58000 observations et 10 variables distribué avec ce didacticiel.

Vous pourrez reproduire l'expérience chez vous, le passage d'EXCEL à TANAGRA dure quelques secondes. Nous avons constaté que dans la plupart des cas, pour les tailles de fichiers acceptées par EXCEL, le passage se fait quasiment instantanément.