

Objectif

Mesurer l'importance de la relation entre deux variables nominales.

Pour quantifier le lien existant entre deux variables continues, nous utilisons généralement le coefficient de corrélation. Cet indicateur est très largement répandu, ses défauts et ses qualités sont largement connus.

Lorsque nous voulons traiter deux variables catégorielles (variables nominales), les indicateurs sont moins connus. Le point de départ est le tableau croisant les deux variables, le tableau de contingence, il recense les effectifs pour chaque combinaison de valeurs des variables.

A partir de ce tableau, plusieurs indicateurs peuvent être calculés. Ils permettent de caractériser, de différentes manières, les liens -- les associations -- existant entre les deux variables. Nous verrons dans ce didacticiel comment calculer ces différents indicateurs avec TANAGRA.

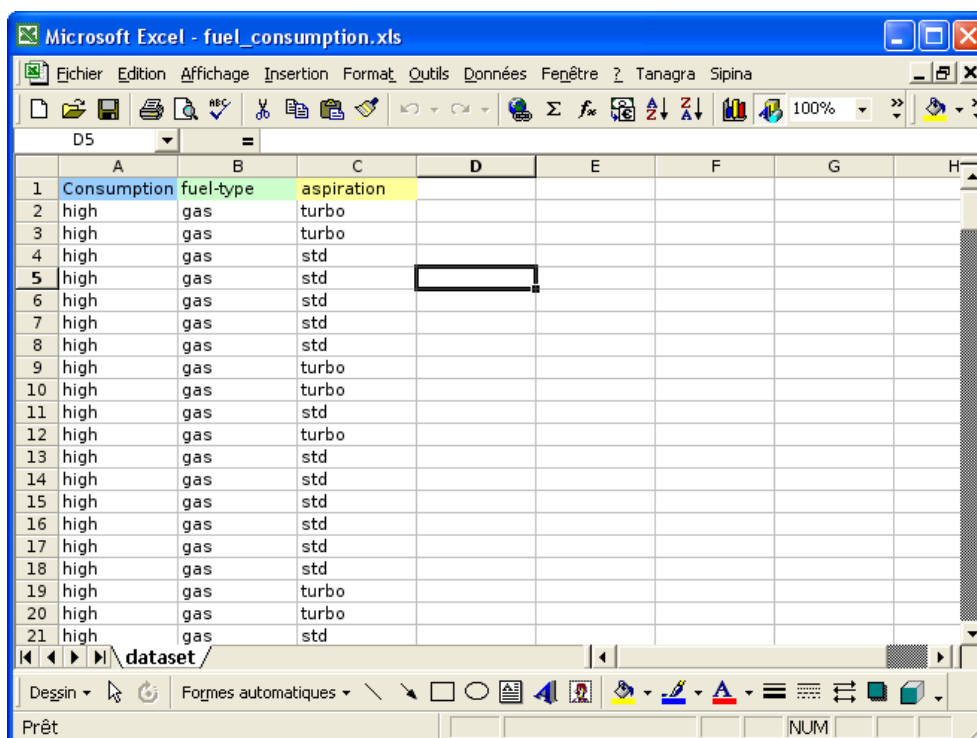
La description des mesures présentées dans ce didacticiel est disponible sur les sites suivants :

- http://www.georgetown.edu/faculty/ballc/webtools/web_chi_tut.html
- <http://v8doc.sas.com/sashtml/stat/chap28/sect20.htm>
- <http://www2.chass.ncsu.edu/garson/PA765/assocnominal.htm>

Données

Le fichier FUEL_CONSUMPTION.XLS recense la consommation (CONSUMPTION), le type de carburant (FUEL_TYPE) et le type d'alimentation (ASPIRATION) de 205 véhicules.

Nous affichons les 20 premières observations du fichier.



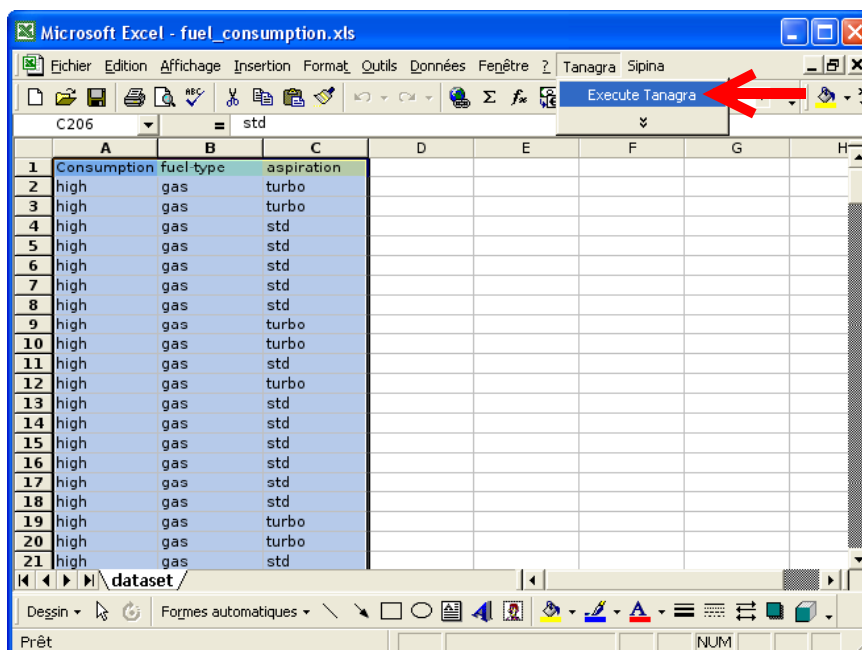
	A	B	C	D	E	F	G	H
1	Consumption	fuel-type	aspiration					
2	high	gas	turbo					
3	high	gas	turbo					
4	high	gas	std					
5	high	gas	std					
6	high	gas	std					
7	high	gas	std					
8	high	gas	std					
9	high	gas	turbo					
10	high	gas	turbo					
11	high	gas	std					
12	high	gas	turbo					
13	high	gas	std					
14	high	gas	std					
15	high	gas	std					
16	high	gas	std					
17	high	gas	std					
18	high	gas	std					
19	high	gas	turbo					
20	high	gas	turbo					
21	high	gas	std					

L'objectif est d'évaluer le lien existant entre la consommation, d'une part, et les autres variables d'autre part.

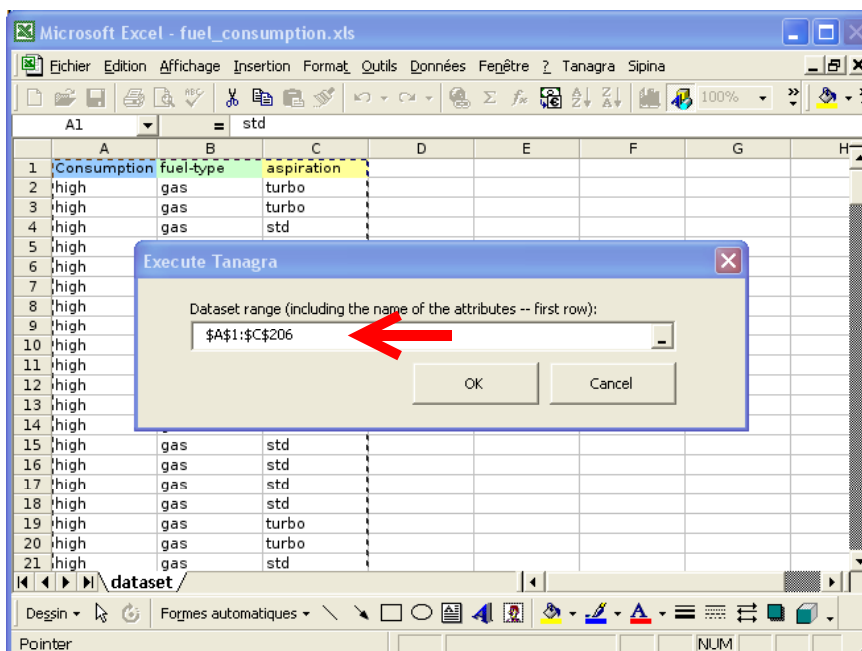
Ecart à l'indépendance – Test du KHI-2

Création du diagramme

Le plus simple est de charger les données dans le tableur EXCEL. Nous sélectionnons les données, puis, nous lançons TANAGRA en activant le menu TANAGRA/EXECUTE TANAGRA installé par la macro complémentaire TANAGRA.XLA¹.



Nous validons la sélection.



¹ Cette macro complémentaire est disponible depuis la version 1.4.11 de TANAGRA. Un didacticiel disponible sur le site web indique comment l'activer dans votre tableur EXCEL.

TANAGRA est automatiquement démarré, nous vérifions que nous disposons bien de 3 variables et 205 observations.

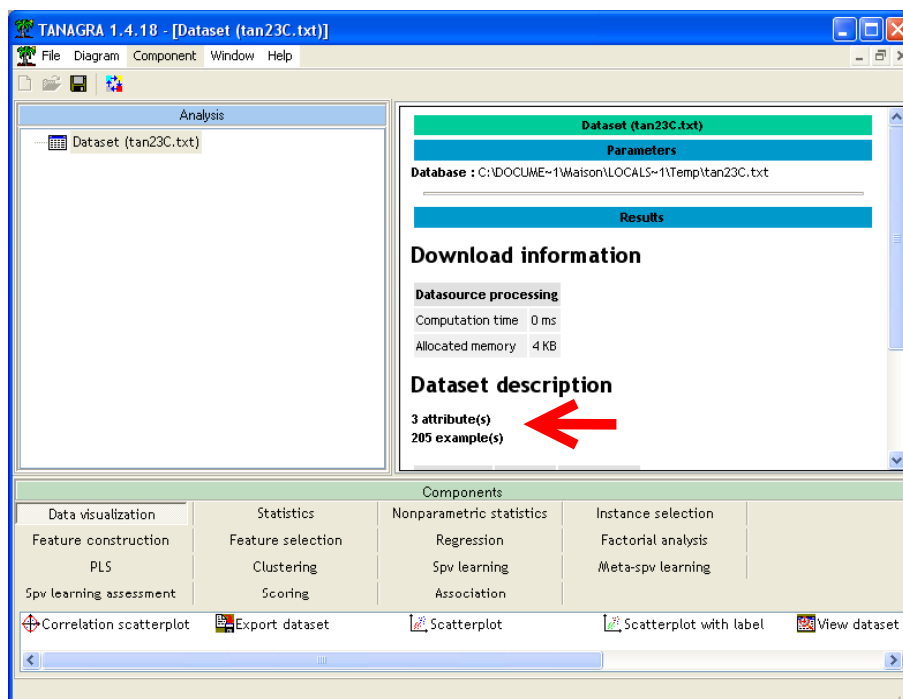
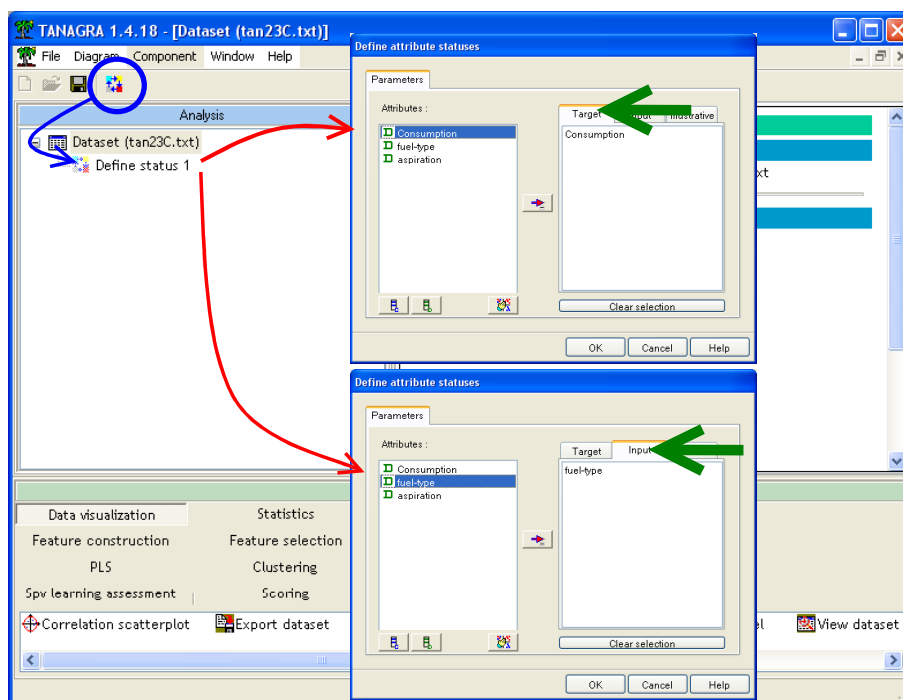


Tableau croisé et test du KHI-2 d'indépendance

Nous plaçons le composant DEFINE STATUS en utilisant le raccourci situé dans la barre d'outils. Dans un premier temps, nous voulons analyser le lien entre, en TARGET, la consommation, et en INPUT, le type de carburant.



Nous insérons alors dans le diagramme le composant CONTINGENCY CHI-SQUARE situé dans l'onglet NONPARAMETRIC STATISTICS. Nous cliquons sur le menu contextuel VIEW pour accéder aux résultats.

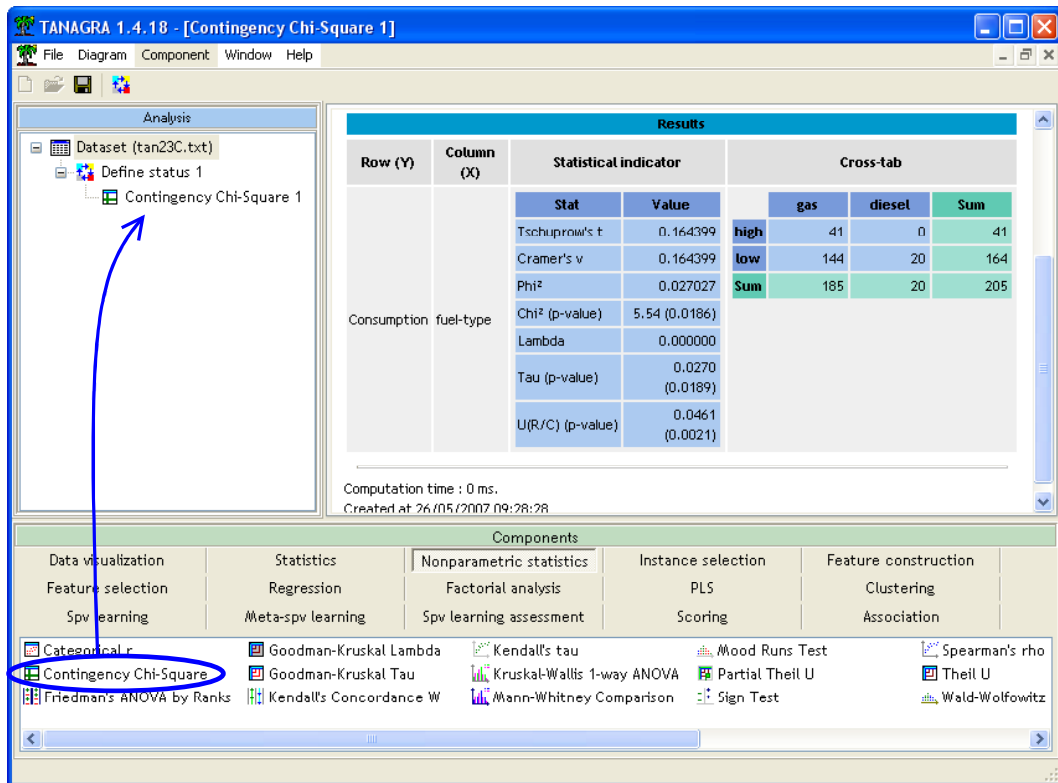


Tableau de contingence. Plusieurs informations sont disponibles. Tout d'abord, le tableau de contingence (CROSS-TAB) croisant les couples de variables indique les co-occurrences entre les différentes valeurs des variables. Pour le premier tableau par exemple, nous constatons qu'il y a 20 véhicules « diesel », ils présentent tous une consommation basse « low ». Par ailleurs, 185 véhicules roulent à l'essence « gas », 41 d'entre eux ont une consommation élevée « high ». Le nombre total de véhicules est bien égal à 205.

Test du KHI-2. Différentes statistiques (STATISTICAL INDICATOR) caractérisent le degré de liaison entre les variables. L'indicateur le plus usuel est certainement la statistique du KHI-2 (CHI-2 en anglais, que l'on utilisera par la suite pour être cohérent avec les affichages du logiciel), il quantifie l'écart entre le tableau construit sur les données et le tableau que l'on aurait obtenu si l'hypothèse d'indépendance entre les variables était avérée. Il est égal à 5.54 dans notre exemple.

La *p-value* de la statistique du CHI-2 permet de déterminer s'il y a lieu de rejeter ou non l'hypothèse d'indépendance (H0). Si elle est inférieure au niveau de signification (généralement 5%), nous rejetons H0. C'est le cas ici puisque la *p-value* obtenue est de 0.0189.

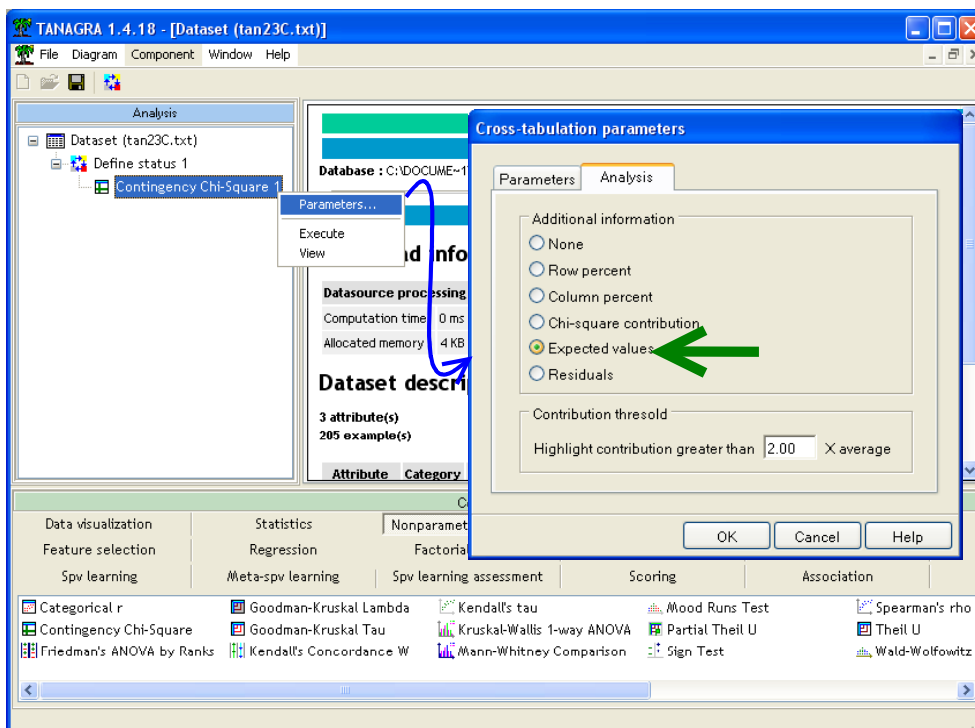
A titre de comparaison, voici les résultats fournis par le logiciel STATISTICA.

STAT. ELEMENT.	Chi ²	dl	p
Chi ² de Pearson	5.540541	df=1	p=.01856
Chi ² du NV	9.452478	df=1	p=.00211
Phi des tables 2 x 2	.1643990		
Corrél. tétrachorique	.5125254		
Coef. de contingence	.1622214		

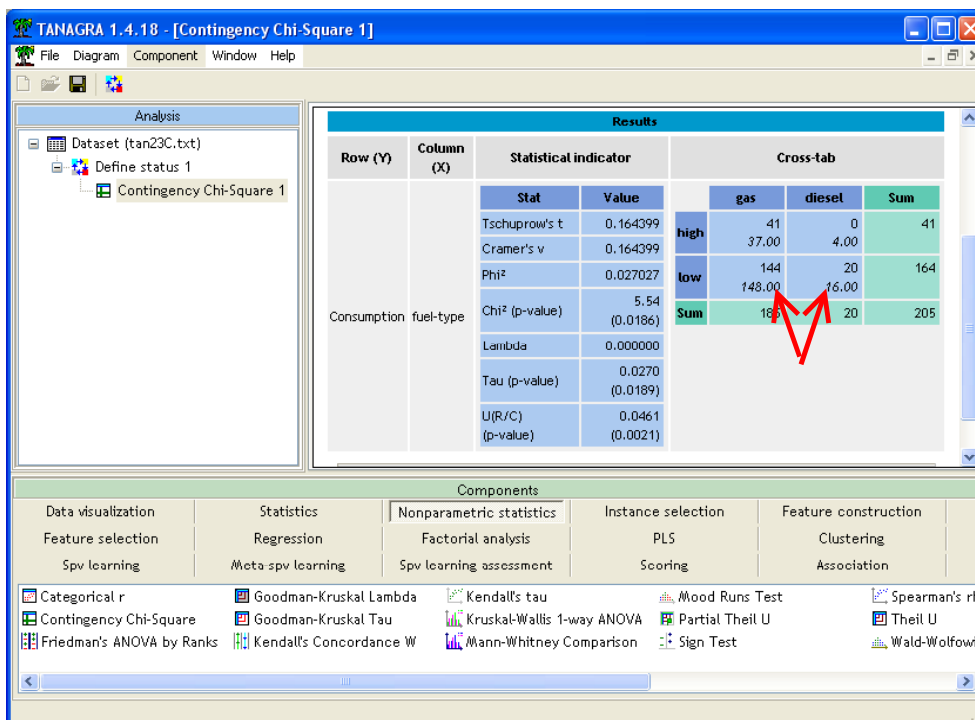
STAT. ELEMENT.	Effect. cellules marquées >10		
CONSUMPT	FUEL_TYP gas	FUEL_TYP diesel	Totaux Lignes
high : high	41	0	41
low : low	144	20	164
Tot. Colonnes	185	20	205

La statistique du CHI-2 varie de 0 à +∞, son interprétation peut se révéler malaisée. TANAGRA propose d'autres mesures dérivées, le PHI² (0.027 ; PHI = SQRT(PHI²) = 0.1643), le t de TSCHUPROW (0.164) et le v de CRAMER (0.164). Ces deux derniers indicateurs sont intéressants car ils varient de 0 à 1 quel que soit le nombre de modalités des variables en jeu.

Contributions et détail des calculs. Il est possible d'obtenir le détail des calculs dans le tableau de contingence. Nous pouvons par exemple vouloir visualiser le tableau que nous aurions sous l'hypothèse d'indépendance. Pour ce faire, nous activons le menu contextuel PARAMETERS, et dans l'onglet ANALYSIS nous sélectionnons l'option EXPECTED VALUES.

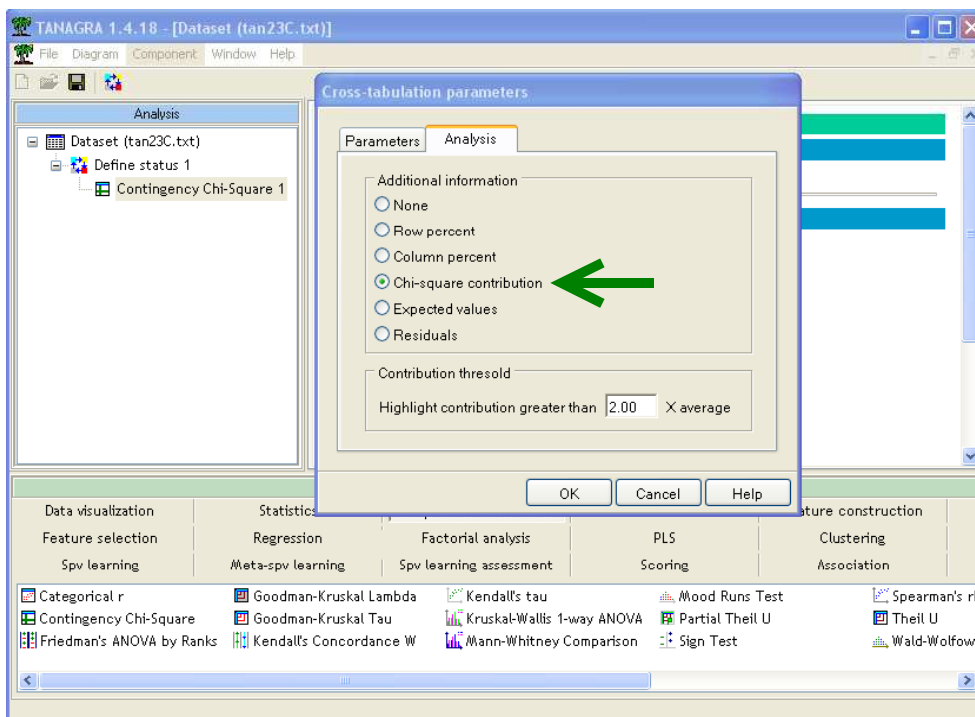


Nous cliquons de nouveau sur le menu VIEW.

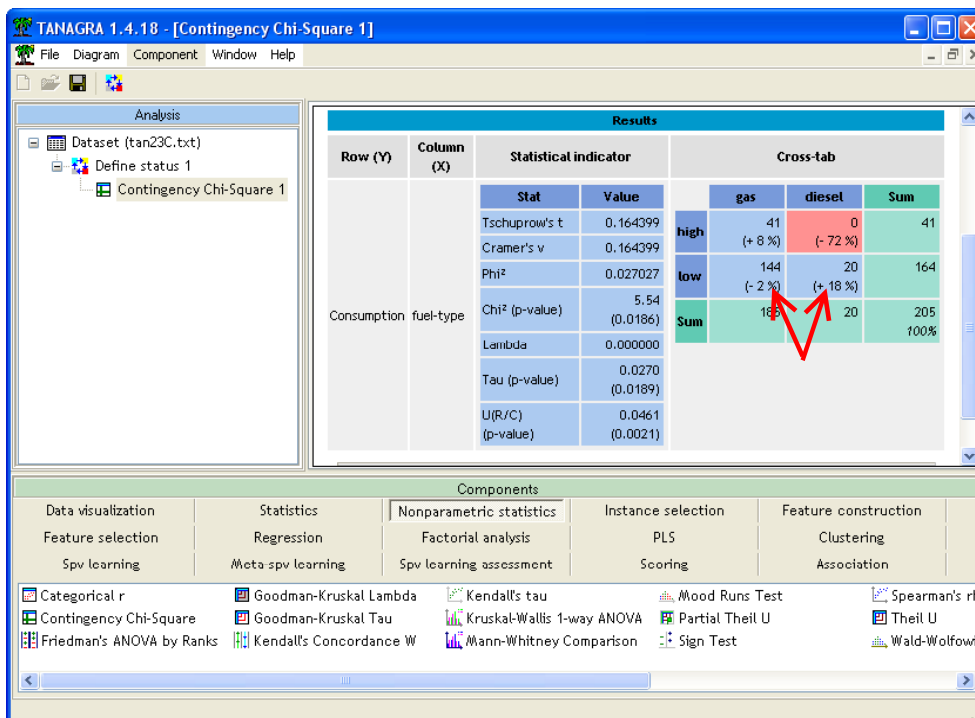


En dessous des effectifs observés sont maintenant affichés les effectifs théoriques sous l'hypothèse d'indépendance.

Le CHI-2 est un critère additif. Il peut être intéressant de détecter quelles sont les cases qui s'écartent le plus de la situation d'indépendance ou, en d'autres termes, qui contribuent le plus au CHI-2. Nous actionnons le menu PARAMETERS, dans l'onglet ANALYSIS nous choisissons l'option CHI-SQUARE CONTRIBUTION.



Nous obtenons les résultats suivants après validation.



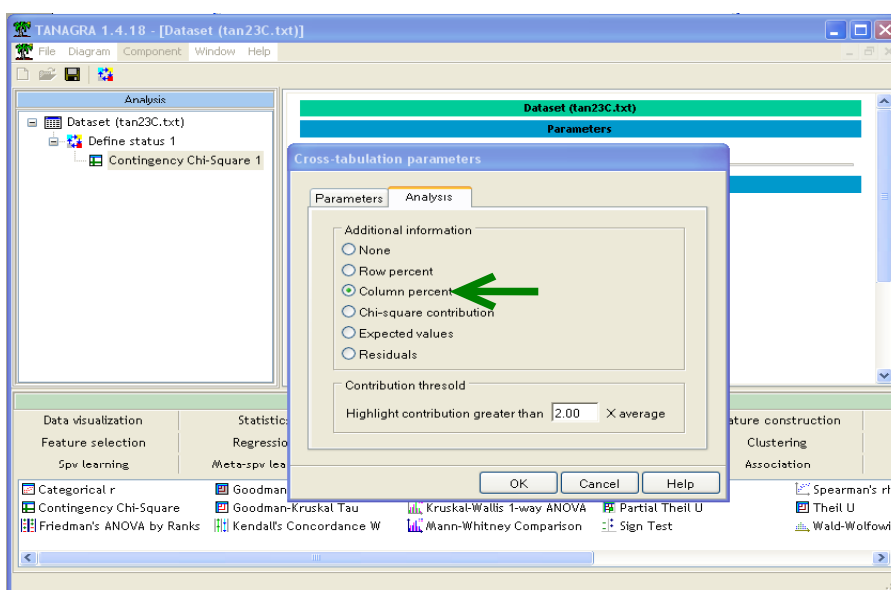
Les contributions sont indiquées en pourcentage du total (qui est CHI-2=5.54). Le signe permet de déterminer s'il s'agit d'une attraction ou une répulsion entre les caractéristiques étudiées. Lorsque la contribution d'une case est 2 (paramétrable) fois plus élevée que la contribution moyenne, elle est surlignée en rouge.

Dans notre premier tableau de résultats, nous constatons que la liaison entre CONSUMPTION et FUEL-TYPE repose avant tout sur une forte répulsion (-72 %) entre le carburant diesel et la consommation élevée.

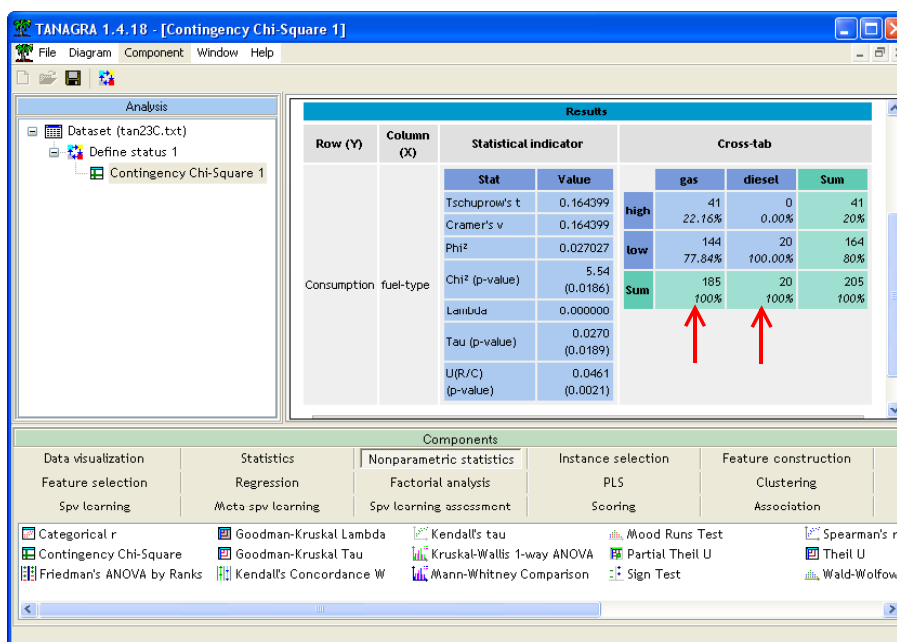
Bien entendu, ce type d'analyse n'a de sens que si le CHI-2 est significatif, indiquant une liaison avérée entre les variables étudiées.

Association asymétrique

Bien souvent, les variables ne jouent pas un rôle symétrique dans une analyse. Dans notre cas, nous voulons en réalité déterminer dans quelle mesure le type de carburant influe sur la consommation. Un premier type d'analyse serait de comparer les proportions des véhicules à forte et faible consommation selon le type de carburant qu'ils utilisent. Nous pouvons obtenir ces « profils colonnes » dans notre tableau de contingence en activant le menu PARAMETERS, puis l'option COLUMN PERCENT dans l'onglet ANALYSIS.



Nous observons les résultats suivants.



Nous constatons effectivement que la situation est très contrastée selon que le véhicule utilise de l'essence (GAS) ou du gazole (DIESEL). Dans ce second cas, tous les véhicules (100%) présentent une consommation basse « low ». Cette proportion est moins élevée chez les véhicules à essence (77.84%). Le fait de travailler sur des pourcentages rend les colonnes comparables, ce qui autorise ce type de lecture.

Mesures PRE (Proportionate Reduction in Error)

Le CHI-2 ne convient pas dans ce type d'analyse. En effet, nous obtiendrons la même valeur si nous transposons notre tableau de contingence. Ce qui est gênant puisque l'analyse serait diamétralement différente dans ce cas : la consommation pèserait sur le carburant utilisé par les voitures ?

Il nous faut donc des mesures qui tiennent compte du rôle asymétrique que jouent les variables. Les mesures PRE indiquent dans quelle proportion la connaissance de la variable INPUT (FUEL-TYPE) permet de réduire la prédiction de la variable TARGET (CONSUMPTION). Trois mesures différentes sont proposées dans TANAGRA : le LAMBDA de Goodman & Kruskal ; le TAU des mêmes auteurs ; le U de Theil, connu également sous l'appellation « uncertainty coefficient ». Les deux derniers indicateurs sont accompagnés de la p-value du test d'absence d'association prédictive entre les variables. Dans le tableau du composant précédent, nous lisons respectivement les valeurs 0.0 ; 0.027 (0.0189) et 0.061 (0.0021). Les deux derniers indicateurs confirment l'existence d'une association entre le type de carburant et le niveau de consommation des véhicules au risque de 5%.

Il est possible d'obtenir le détail des calculs pour chacun de ces indicateurs, notamment pour mieux comprendre l'obtention des probabilités critiques (*p-value*) des tests qui ont été mis en œuvre pour évaluer l'absence d'association. Il nous faut introduire de nouveaux composants, situés dans l'onglet NONPARAMETRIC STATISTICS, dans le diagramme de traitements.

Tau de Goodman et Kruskal. Nous insérons le composant GOODMAN-KRUSKAL Tau dans notre diagramme, en dessous du DEFINE STATUS 1. Nous cliquons sur le menu VIEW pour obtenir les résultats.

The screenshot shows the TANAGRA 1.4.18 software interface. The main window displays the configuration of the Goodman-Kruskal Tau component. The 'Analysis' pane on the left shows a tree structure with 'Dataset (tan23C.txt)', 'Define status 1', 'Contingency Chi-Square 1', and 'Goodman-Kruskal Tau 1'. A blue arrow points from the 'Goodman-Kruskal Tau 1' component in the tree to the 'Goodman-Kruskal Tau' component in the 'Components' pane at the bottom. The 'Results' pane on the right displays the following table:

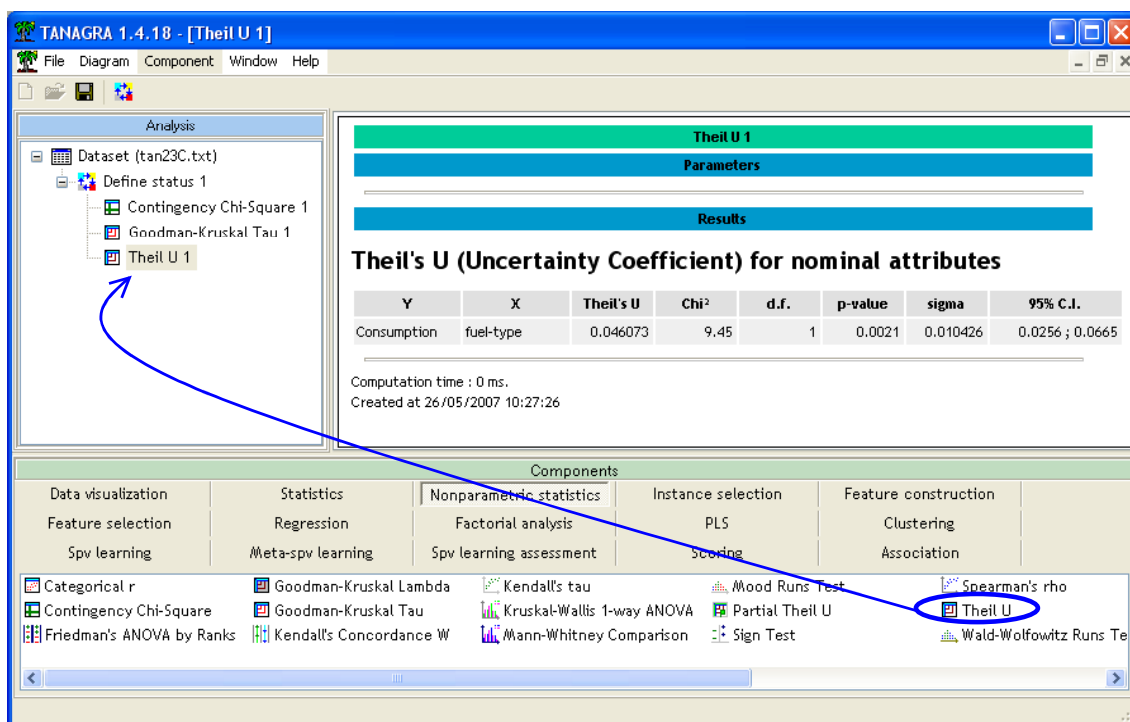
Goodman & Kruskal's Tau for nominal attributes					
Y	X	Tau	Chi ²	d.f.	p-value
Consumption	fuel-type	0.027027	5.51	1	0.0189

Below the table, it states: 'Computation time : 0 ms. Created at 26/05/2007 10:23:51'. The 'Components' pane at the bottom lists various statistical methods, with 'Goodman-Kruskal Tau' circled in blue.

Le Tau est le même que celui précédemment affiché dans le composant CONTINGENCY CHI-SQUARE. Il est égal à 0.0270 pour l'association entre CONSUMPTION et FUEL-TYPE.

Le CHI-2 calculé ici correspond à une transformation du Tau proposé par Light & Margolin (1971). Il est égal à 5.51. Le nombre de degrés de liberté est le même que pour le test d'indépendance du CHI-2. Nous obtenons ainsi les probabilités critiques (p-value) permettant de conclure à une association prédictive significative (ou non) entre les variables. L'association est significative à 5%, mais pas à 1%.

U de Theil. Nous insérons maintenant le composant THEIL U dans notre diagramme.



Le U de Theil est de 0.046 pour le couple de variables que nous étudions. Il existe également une transformation qui permet d'aboutir à une statistique suivant la loi du CHI-2, appelé CHI-2 du Maximum de Vraisemblance dans les autres logiciels (cf. SPSS, STATISTICA, etc.). De même, nous observons les degrés de libertés et des probabilités critiques (p-value) des tests. Les conclusions sont cohérentes avec les indications du Tau de Goodman.

A titre de comparaison, STATISTICA fournit les valeurs suivantes pour le calcul de l'association entre CONSUMPTION et FUEL-TYPE.

STAT. ELEMENT.	Chi ²	dl	p
Chi ² de Pearson	5.540541	df=1	p=.01858
Chi ² du MV	9.452478	df=1	p=.00211
Coeff. d'incertitude	X=.0460726	Y=.0721161	X Y=.05622

Nous disposons d'informations supplémentaires avec le U de Theil. En effet, TANAGRA calcule l'écart type asymptotique de la statistique, il est égal à 0.010426 dans notre étude. L'intérêt de cette nouvelle information est que nous pouvons ainsi calculer des intervalles de confiance en nous appuyant sur l'approximation normale de la distribution de U. TANAGRA propose par défaut les

intervalles à 95%. Mais il est aisé de produire d'autres intervalles en reprenant les valeurs (U et écart type de U) du tableau de résultats.

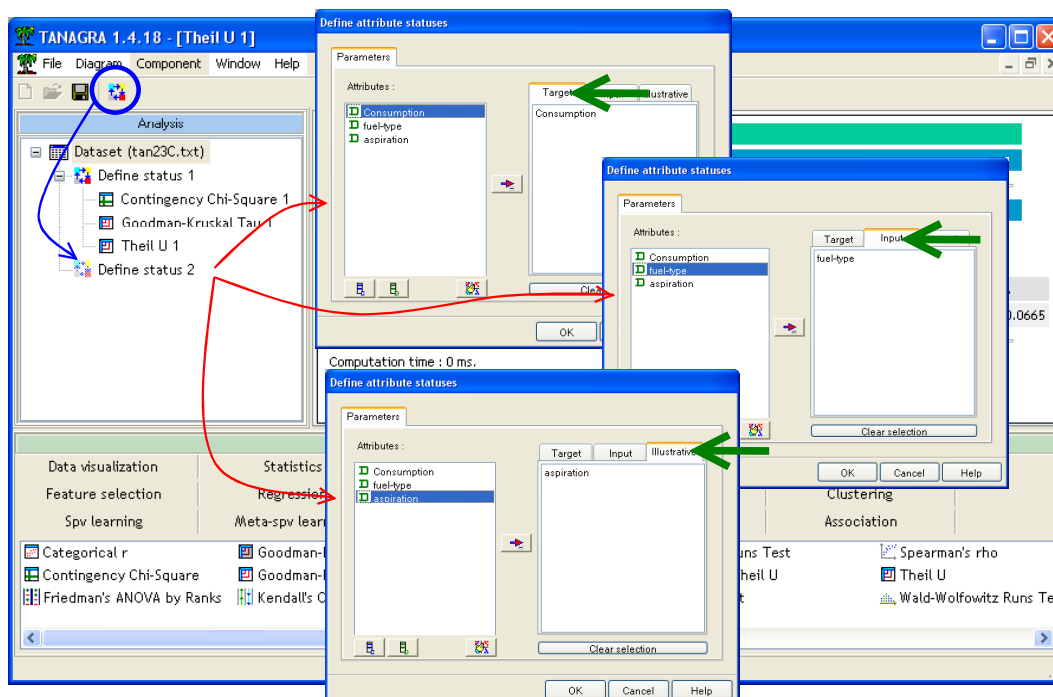
Association partielle

Dans certains cas, il peut être intéressant d'évaluer le rôle d'une tierce variable (E) dans l'étude de l'association entre deux variables A et B. La variable E joue le rôle de variable de contrôle, l'idée est de mesurer l'association entre A et B conditionnellement aux valeurs prises par E. Plusieurs types de résultats peuvent apparaître. Dans certains cas, cette troisième variable peut occulter la relation entre A et B, par exemple lorsqu'elle détermine à la fois les valeurs de A et B. Dès lors la relation directe qui semblait exister entre A et B n'en est que la résultante, il s'agit d'un artefact. On dit alors que E est un facteur confondant. Dans d'autres cas, l'introduction de la variable E peut exacerber la relation entre A et B, c'est le cas lorsque la relation entre A et B est inversée selon les valeurs prises par E.

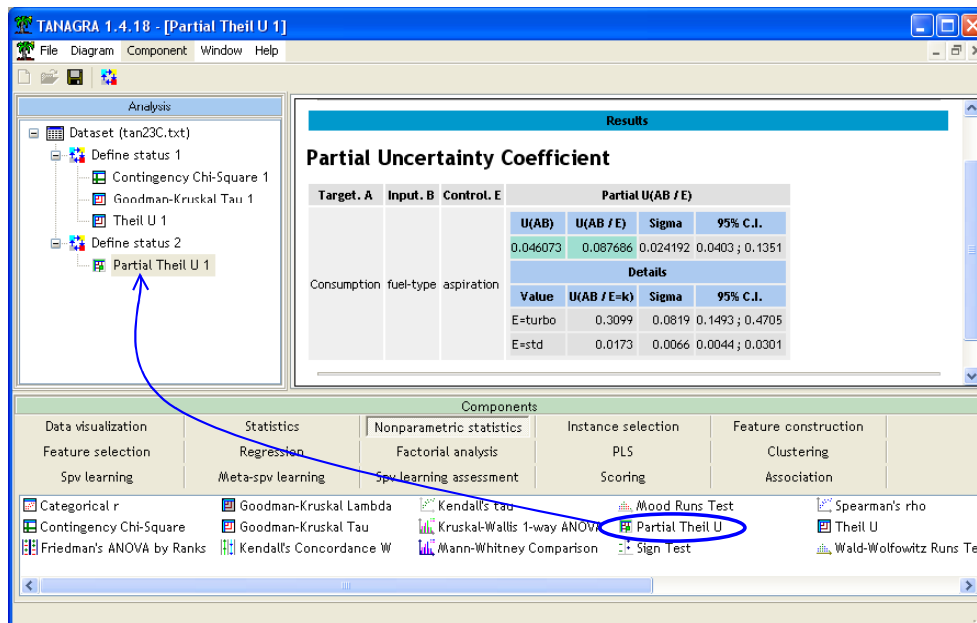
Pour plus de détails sur l'étude de ces effets, nous conseillons la lecture du site <http://www2.chass.ncsu.edu/garson/PA765/association.htm#spss> pour l'association entre variables nominales, et le site <http://www2.chass.ncsu.edu/garson/pa765/partialr.htm> pour l'étude des corrélations partielles entre variables quantitatives.

Dans notre cas, une troisième variable peut jouer un rôle important dans la détermination de la consommation, le type d'alimentation du véhicule (ASPIRATION). Nous voulons déterminer l'association entre le type de carburant et la consommation selon le type d'aspiration.

U de Theil Partiel. Nous introduisons le composant DEFINE STATUS à la racine du diagramme, nous mettons : en TARGET la variable CONSUMPTION ; en INPUT, la variable FUEL-TYPE ; et, en ILLUSTRATIVE, la variable ASPIRATION.

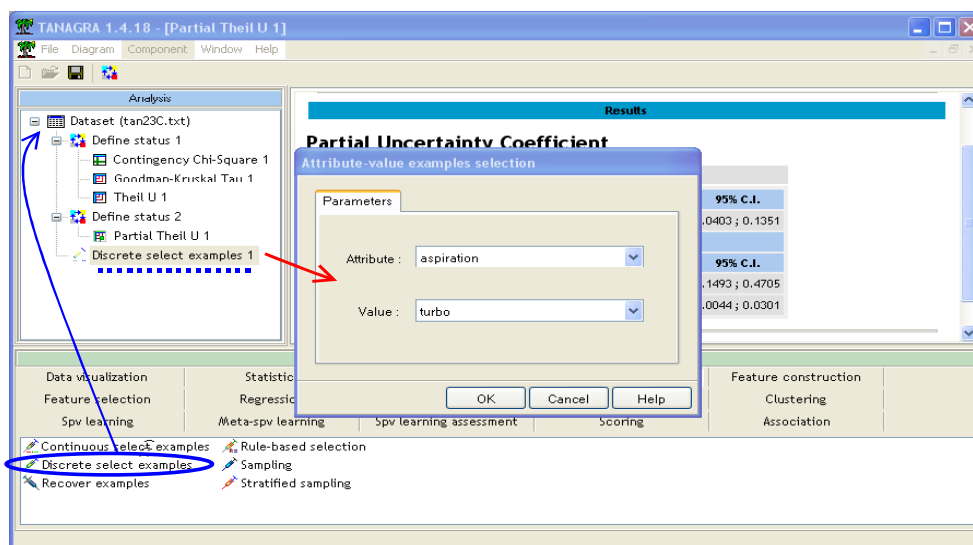


Nous plaçons le composant PARTIAL THEIL U qui calcule les associations partielles à partir du U de Theil (onglet NONPARAMETRIC STATISTICS).



D'emblée nous constatons que l'association partielle $U(AB/E) = 0.0876$ est plus élevée que l'association brute $U(AB) = 0.0460$. La différence est essentiellement due à une association CONSUMPTION/FUEL-TYPE nettement exacerbé chez les véhicules ASPIRATION=TURBO, $U(AB/E=TURBO) = 0.3099$. TANAGRA fournit automatiquement les écarts type asymptotiques et les intervalles de confiance à 95%. Enfin, le coefficient partiel est bien une moyenne pondérée des coefficients conditionnels. La pondération tient compte des poids respectifs des modalités de E, mais également de la structure de l'erreur de prédiction².

Détail des tableaux conditionnels. Nous voulons approfondir l'étude de la relation, très significative semble-t-il, entre consommation et type de carburant chez les véhicules TURBO. Il nous faut donc restreindre les calculs à ce type de véhicules en filtrant les données de départ. Pour ce faire, nous revenons à la racine du diagramme et nous plaçons le composant DISCRETE SELECT EXAMPLES (onglet INSTANCE SELECTION). Nous le paramétrons en spécifiant l'attribut (ASPIRATION) et la modalité (TURBO) de filtrage.



² M. OLSZAK, G. RITSCHARD (1995) « The behavior of nominal and ordinal partial association measure », in The Statistician, vol.44, n°2, pp.195-212.

Ensuite, nous insérons la séquence DEFINE STATUS et CONTINGENCY CHI-SQUARE, en veillant à mettre en TARGET CONSUMPTION, et en INPUT FUEL-TYPE. Nous demandons l'affichage des profils colonnes dans le tableau de contingence.

Row (Y)	Column (X)	Statistical indicator		Cross-tab		
		Stat	Value	gas	diesel	Sum
Consumption	high	Tschuprow's t	0.541667	13	0	13
		Cramer's v	0.541667	54.17%	0.00%	35%
	low	Phi²	0.293403	11	13	24
		Chi² (p-value)	10.86 (0.0010)	45.83%	100.00%	65%
	Sum	Lambda	0.153846	24	13	37
		Tau (p-value)	0.2934 (0.0012)	100%	100%	100%
		U(R/C) (p-value)	0.3099 (0.0001)			

Il y a 37 véhicules TURBO dans la base. Nous constatons que tous (100%) les véhicules DIESEL consomment peu « low ». En revanche, il y a une majorité (54.17%) de véhicules à consommation élevée « high » chez les véhicules à essence (GAS).

L'association est significative au seuil de 1% (p-value du U de Theil → 0.0001).

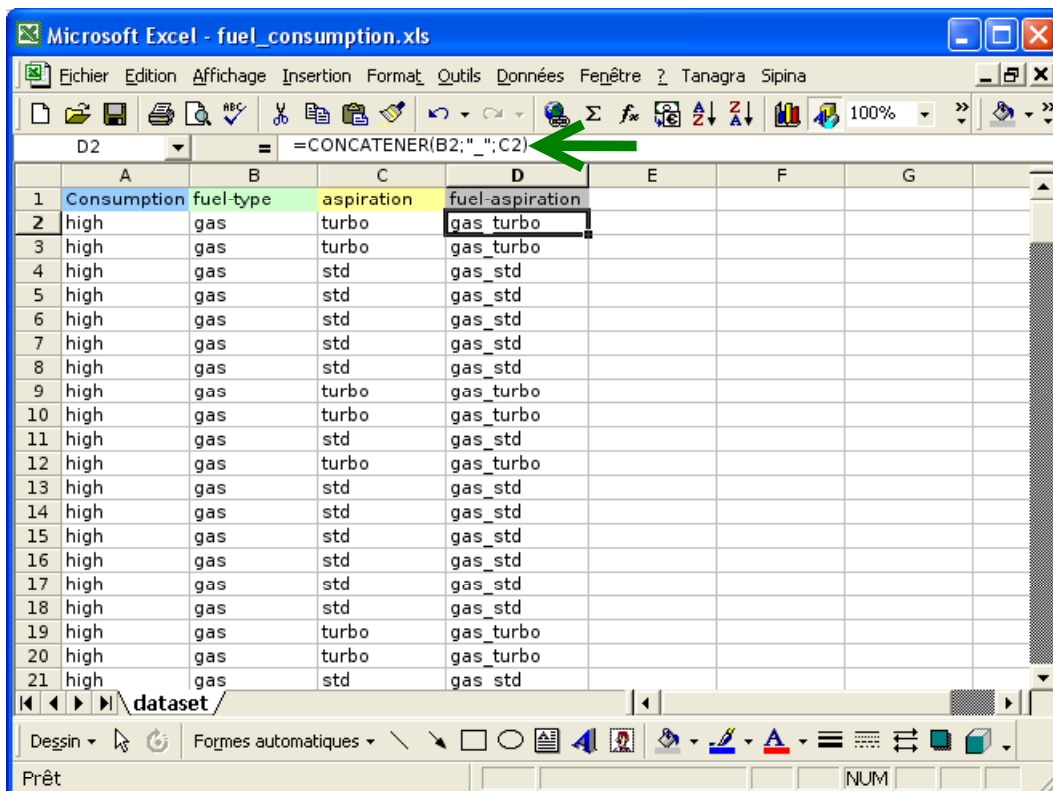
Combinaison de variables

Une autre manière de procéder est d'évaluer non plus les effets conditionnels mais l'effet conjoint des variables FUEL-TYPE et ASPIRATION sur la consommation, introduisant ainsi l'étude des interactions.

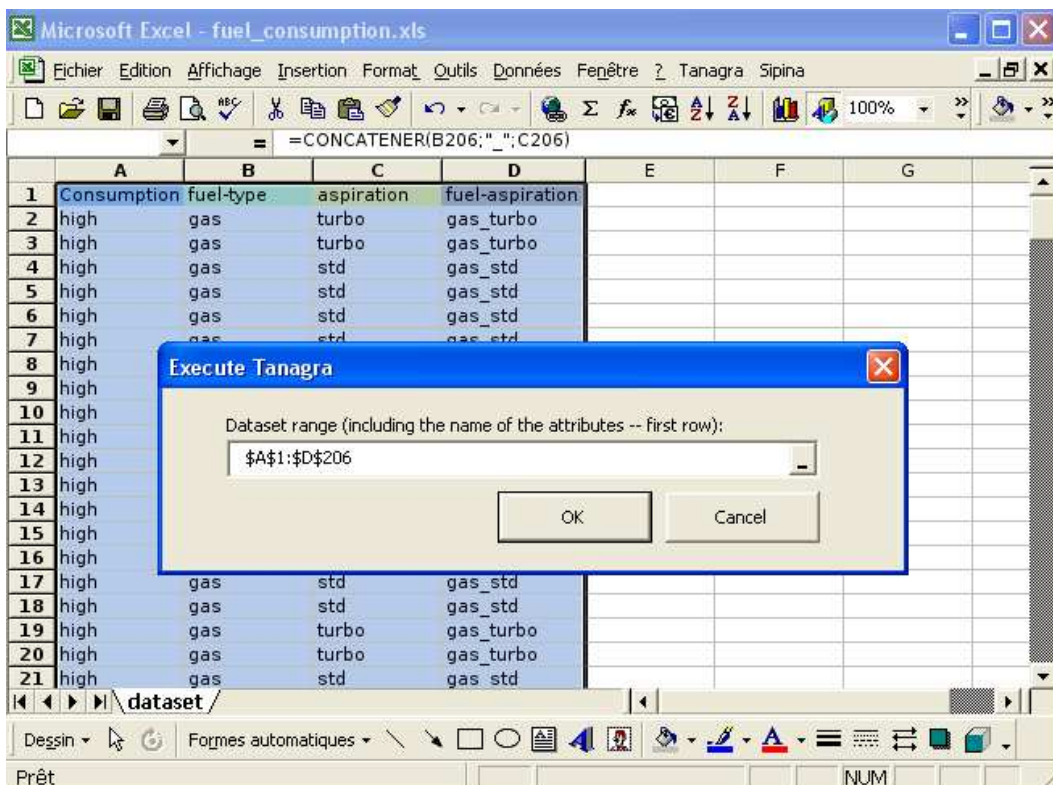
Le plus simple dans ce cas est de construire une variable intermédiaire correspondant aux couples de valeurs prises par nos deux variables ci-dessus. Travailler en osmose avec un tableur s'avère très avantageux dans ce cas, les possibilités de traitements sont très étendues.

Création de la variable composite. Nous fermons donc TANAGRA pour revenir au tableur EXCEL. Nous créons une nouvelle colonne FUEL-ASPIRATION. Nous complétons la colonne en concaténant les valeurs des deux variables³.

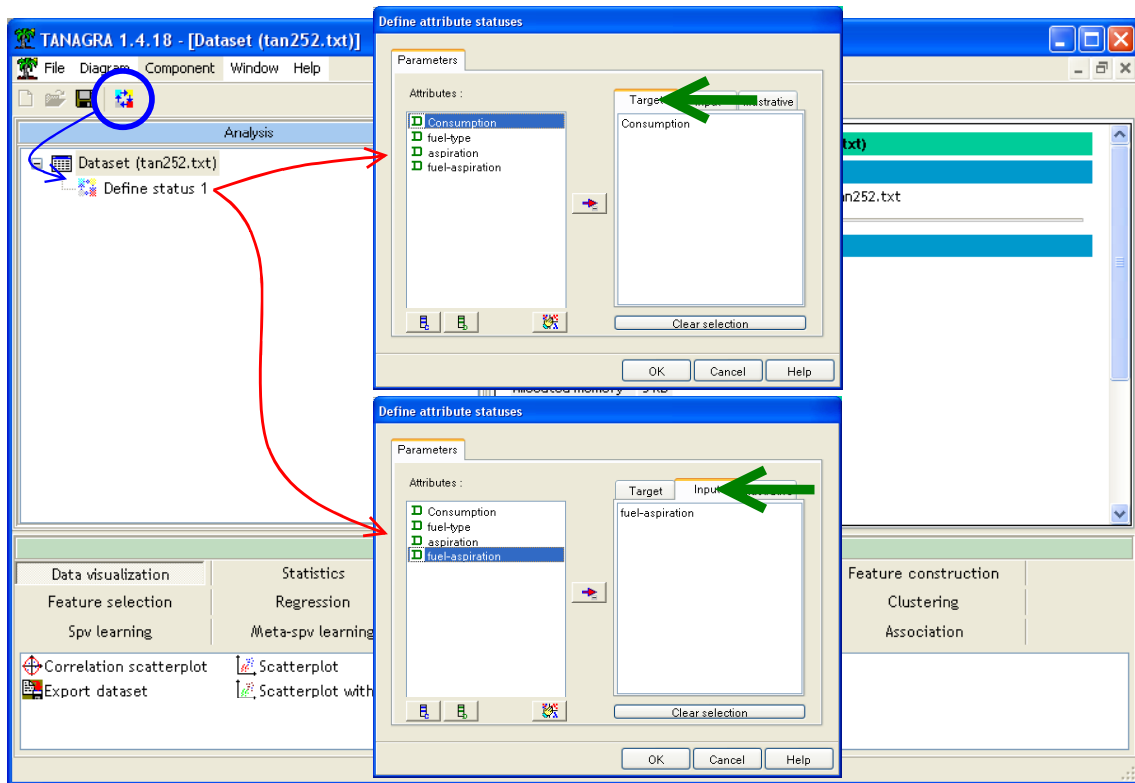
³ Dans la version anglaise d'EXCEL, la fonction à utiliser est =CONCATENATE(...).



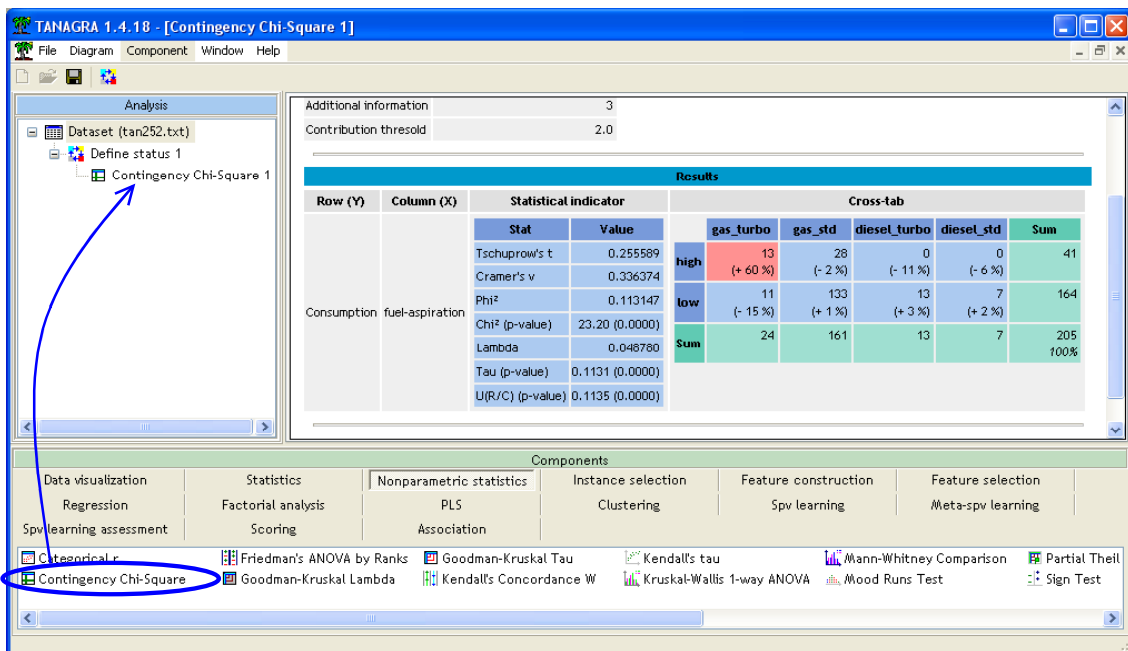
Nouveau diagramme de traitements. Il ne nous reste plus qu'à re-sélectionner les données en incluant cette nouvelle colonne et à lancer TANAGRA via le menu TANAGRA/EXECUTE TANAGRA installé par la macro complémentaire.



Dans la nouvelle session de travail, nous insérons le composant DEFINE STATUS, nous plaçons en TARGET toujours la variable CONSUMPTION, et en INPUT la nouvelle variable FUEL-ASPIRATION.



Puis nous insérons le composant CONTINGENCY CHI-SQUARE en demandant l'affichage des contributions.



Si l'on se réfère au CHI-2 (23.20), la liaison est très significative (p-value < 0.0000...). La principale information véhiculée par la liaison est une forte consommation des véhicules turbo-essence (attraction : +60%).

Conclusion

Le test d'indépendance du KHI-2 pour évaluer le lien entre deux variables nominales est très souvent décrit dans la littérature. Nous montrons comment le mettre en œuvre dans ce didacticiel.

Le second intérêt de ce didacticiel est d'introduire d'autres indicateurs, moins répandus, mais au moins aussi intéressants car ils permettent de caractériser une relation asymétrique, donc plus proche de l'idée de causalité... même s'il faut rester prudent : caractériser une causalité uniquement à l'aide d'outils purement numériques est pour le moins illusoire.