1 Objectif

Détection et traitement des points aberrants avec Tanagra (à partir de 1.4.24).

Dans le processus Data Mining, la détection et le traitement des points aberrants sont incontournables lors de la préparation des données, ou même après coup, pour analyser et valider les résultats.

On parle de point aberrant (point atypique) lorsque qu'un individu prend une valeur exceptionnelle sur une variable (ex. un client d'une banque aurait 158 ans) ou sur des combinaisons de variables (ex. un athlète de 12 ans aurait effectué le 100 m en 10 secondes). Ces points sont problématiques car ils peuvent biaiser les résultats, notamment pour les méthodes basées sur des distances entre individus, ou plus dramatiquement encore, des distances par rapport à des barycentres. Il importe donc d'identifier ces individus et de les considérer attentivement.

Dans ce didacticiel, nous présentons le composant **UNIVARIATE OUTLIER DETECTION** destiné à **détecter les points atypiques sur chacune des variables, prises individuellement**.

Les techniques intégrées dans ce composant sont largement inspirées du texte sur le site de NIST (http://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm). Nous avons implémenté :

- Le test de Grubbs (http://www.itl.nist.gov/div898/handbook/eda/section3/eda35h.htm). Ce test repose sur la normalité de la distribution. On devrait donc tester préalablement la crédibilité de cette hypothèse. Mais lorsque l'on se rend compte que les tests de normalité eux mêmes sont sensibles aux points aberrants, on ne s'en sort plus. Voilà pourquoi on se contente au préalable de techniques graphiques simples destinées à se faire une idée de la répartition des données.

- La règle de « x » – sigmas. Elle consiste à déclarer comme atypique les observations s'écartant de « x » écarts types autour de la moyenne. C'est une règle très fruste. Elle est aussi basée sur une normalité sous jacente des données. On sait par exemple que pour la loi normale, 99.73% des observations sont situées dans l'intervalle $[m-3\times\sigma; m-3\times\sigma]$. Toute observation qui sort de cet intervalle a une très faible probabilité d'apparaître. Il faut savoir pourquoi elle est présente dans les données.

- La règle de la boîte de Tukey (http://en.wikipedia.org/wiki/Box_plot). La boîte à moustaches (BOXPLOT) permet de représenter graphiquement la distribution d'une variable. On peut mettre en évidence les points extrêmes en utilisant une règle simple. Nous calculons le 1er quartile Q1 et le 3ème quartile Q3, nous en déduisons l'intervalle interquartile IQ = Q3 – Q1. On dit qu'une observation est moyennement atypique (mild outlier) s'il est en deçà de LIF = Q1 – 1.5 * IQ ou au delà de UIF = Q3 + 1.5 * IQ (LIF : lowr inner fence, UIF : upper inner fence). Elle est extrêmement atypique si elle en deçà de LOF = Q1 – 3 * IQ ou au delà de UOF = Q3 + 3 * IQ (LOF : lower outer fence, UOF : upper outer fence).

La relation entre la règle des « x » sigmas et la règle de Tukey, lorsque la distribution des données est normale, peut être résumée graphiquement (Figure 1).

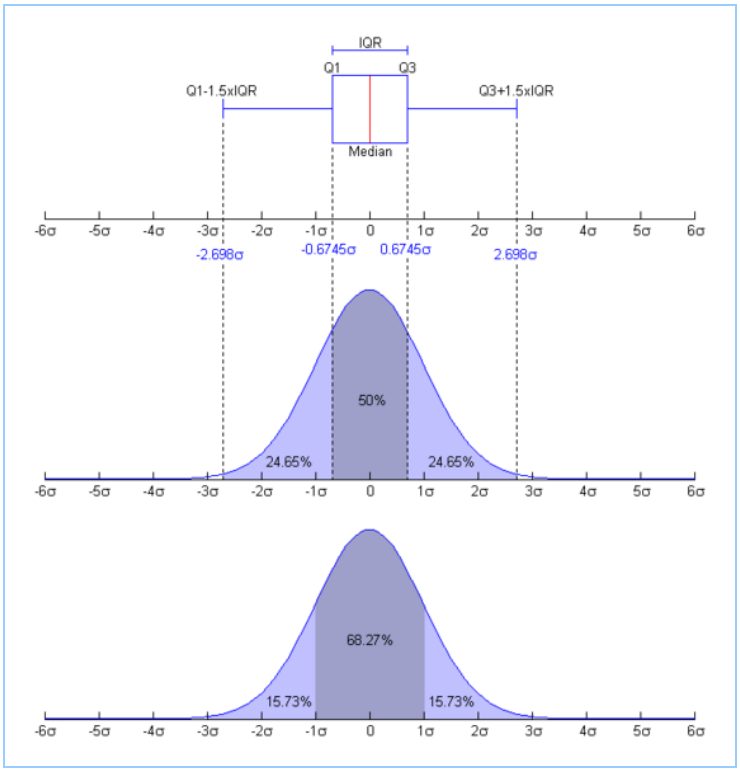


Figure 1 – Lien entre les règles de détection pour la distribution normale (http://en.wikipedia.org/wiki/Image:Boxplot_vs_PDF.png)

Nous essayerons de les combiner au mieux avec les statistiques descriptives dans ce document. On se rendra vite compte que des stratégies simples, notamment les approches graphiques, sont au moins aussi intéressantes finalement. Les techniques numériques ci-dessus ne sont réellement décisives que dans le cadre du traitement automatisé de fichiers comportant de très nombreuses colonnes. Dans ce cas, leurs indications nous permettent de nous orienter rapidement vers les variables à problèmes.

2 Données

Notre fichier de données body_mass_index.xls[1] comporte 50 observations. Les caractéristiques mesurées sont le poids en kg (WEIGHTKG), la taille en mètres (HEIGHTM) et l'indice de masse corporelle (BODYMASS = WEIGHTKG / HEIGHTM^2). L'objectif est de vérifier si des observations se détachent des autres selon au moins une de ces variables.

3 Détection des points aberrants avec TANAGRA

3.1 Création du diagramme – Importation des données

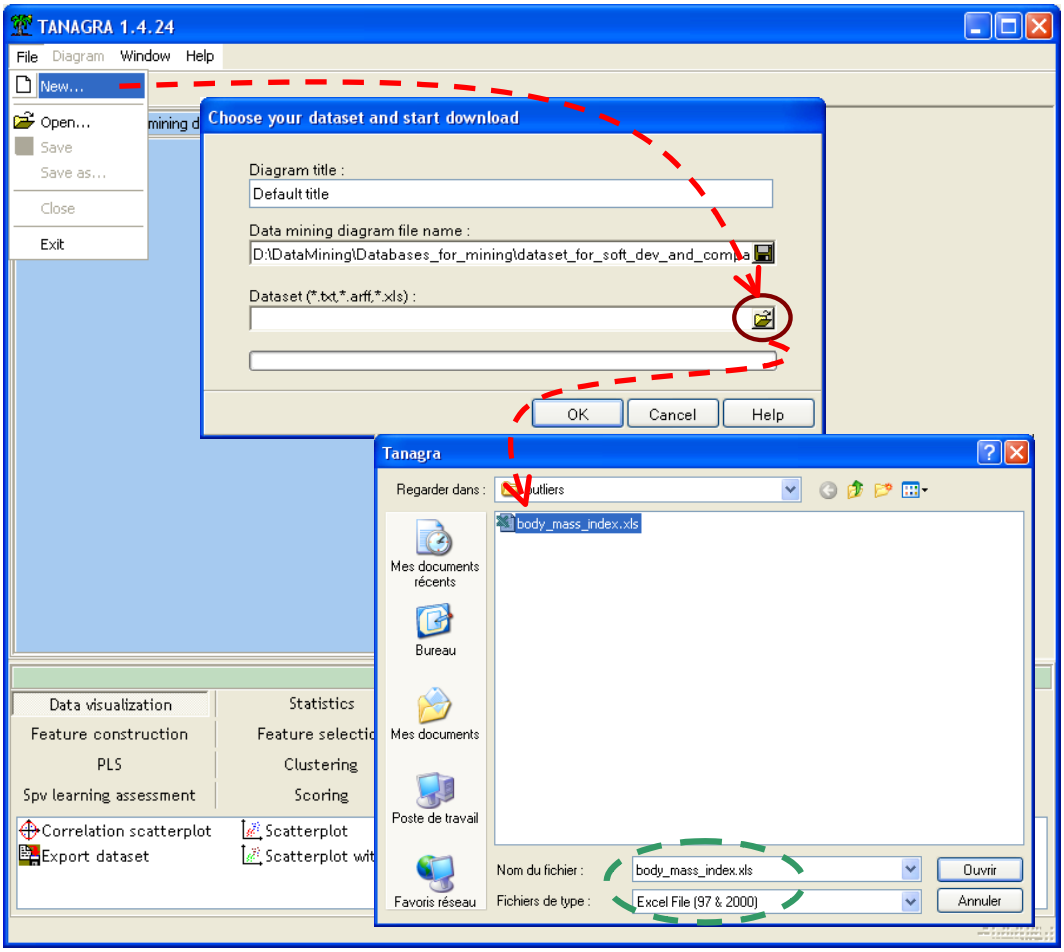
Il existe plusieurs manières de charger les données au format XLS dans TANAGRA. Nous choisissons l'importation directe[2]. Elle présente l'avantage de ne pas requérir la présence du logiciel EXCEL sur

[1] http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/body_mass_index.xls

[2] L'autre possibilité d'importation est d'ouvrir le fichier dans le tableur. Puis à l'aide du nouveau menu TANAGRA dans EXCEL, inséré via la macro complémentaire TANAGRA.XLA, nous transférons les données. Voir : http://tutoriels-data-mining.blogspot.com/2008/03/importation-fichier-xls-excel-macro.html

la machine (voir : http://tutoriels-data-mining.blogspot.com/2008/03/importation-fichier-xls-excel-mode.html). Il faut en revanche que les données soient dans la première feuille de calcul, alignées en haut à gauche, la première ligne correspondant aux noms des variables. Notre configuration respecte ces spécifications. Attention, il ne faut pas que le fichier soit en cours d'édition lors de l'importation.

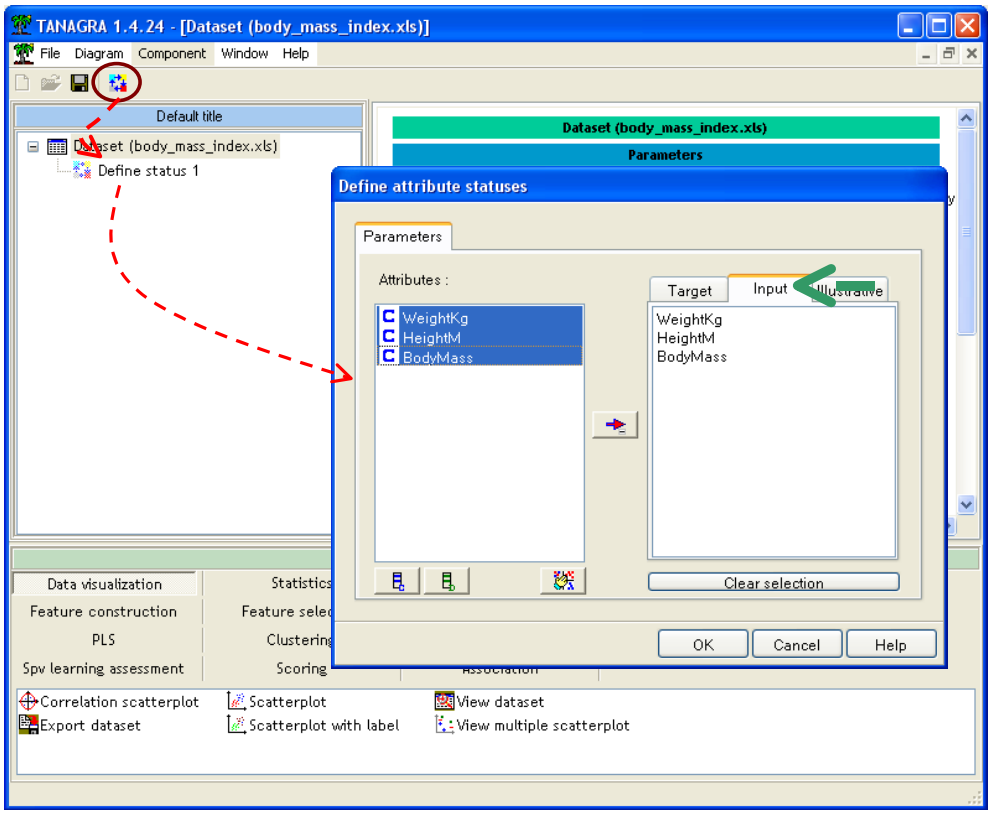
Après avoir démarré TANAGRA, nous activons le menu FILE / NEW pour créer un nouveau diagramme. Dans la boîte de sélection, nous spécifions le nom du fichier de données (body_mass_index.xls) et le nom du fichier diagramme.



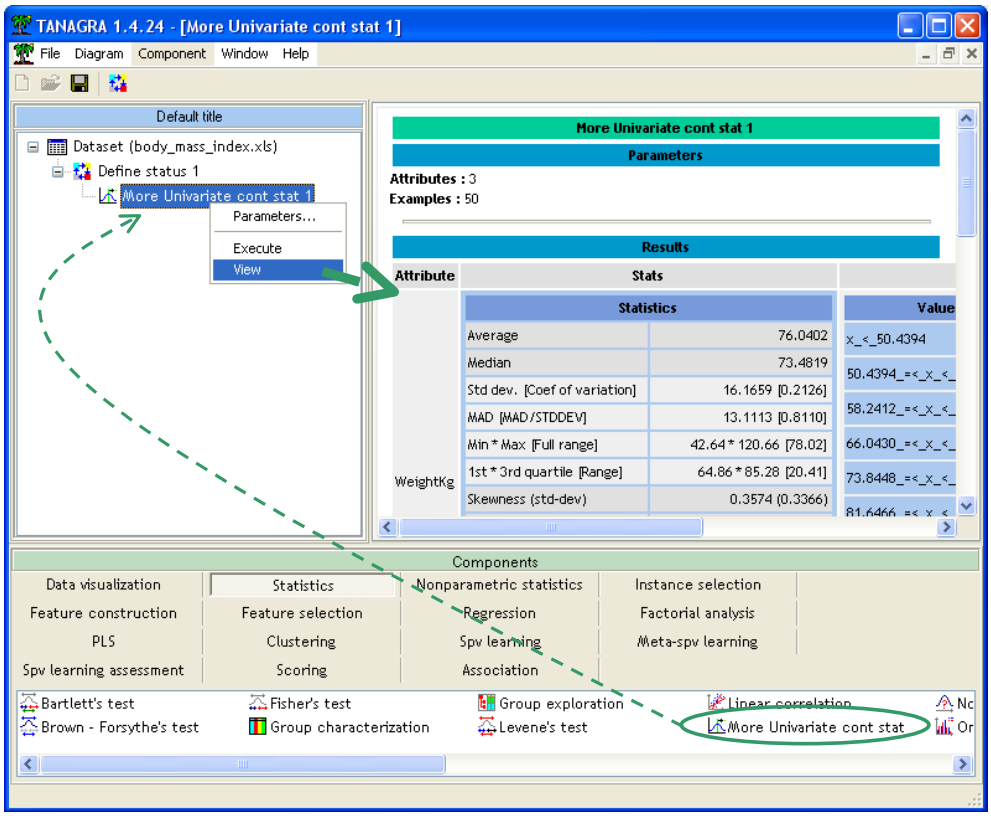
Le fichier comporte bien 50 individus et 3 variables.

3.2 Statistiques descriptives

Première étape, systématique dans les études, résumer les données à l'aide des outils de la statistique descriptive. Nous sélectionnons les variables à analyser à l'aide du composant DEFINE STATUS, accessible via le raccourci dans la barre d'outils.



Puis, nous branchons le composant MORE UNIVARIATE CONT STAT (onglet STATISTICS). Quelques indicateurs usuels et l'histogramme de fréquences sont calculés (Tanagra crée automatiquement 10 intervalles de largeur égales).



Nous résumons dans le tableau suivant les indicateurs et leurs significations.

Indicateur	Description
Average	Moyenne
Median	Médiane
Std.Dev. [Coef of variation]	Ecart type (échantillon) et coefficient de variation (rapport entre l'écart type et la moyenne, permet la comparaison de la dispersion de variables mesurées sur des unités différentes)
MAD [MAD / STDDEV]	Ecart absolu moyen[3]. Rapport entre l'écart absolu moyen et l'écart type. Lorsque la distribution est normale, ce rapport est proche de 0.8.
Min, Max [Full Range]	Minimum, maximum, étendue
1st * 3rd quartile [Range]	1er et 3ème quartile ; intervalle inter quartile
Skewness (std dev)	Coefficient d'asymétrie et son écart type. Lorsque la distribution est normale, skewness = 0
Kurtosis (std dev)	Coefficient d'aplatissement et son écart type. Lorsque la distribution est normale, kurtosis = 0

WEIGHTKG.

Attribute	Stats		Histogram			
	Statistics		Values	Count	Percent	Histogram
WeightKg	Average	76.0402	x_<_50.4394	2	4.00%	
	Median	73.4819	50.4394_=<_x_<_58.2412	4	8.00%	
	Std dev. [Coef of variation]	16.1659 [0.2126]	58.2412_=<_x_<_66.0430	8	16.00%	
	MAD [MAD/STDDEV]	13.1113 [0.8110]	66.0430_=<_x_<_73.8448	11	22.00%	
	Min * Max [Full range]	42.64 * 120.66 [78.02]	73.8448_=<_x_<_81.6466	7	14.00%	
	1st * 3rd quartile [Range]	64.86 * 85.28 [20.41]	81.6466_=<_x_<_89.4483	7	14.00%	
	Skewness (std-dev)	0.3574 (0.3366)	89.4483_=<_x_<_97.2501	8	16.00%	
	Kurtosis (std-dev)	0.1363 (0.6619)	97.2501_=<_x_<_105.0519	1	2.00%	
			105.0519_=<_x_<_112.8537	1	2.00%	
			x>=_112.8537	1	2.00%	

HEIGHTM.

	Statistics		Values	Count	Percent	Histogram
HeightM	Average	1.6581	x_<_1.4902	2	4.00%	
	Median	1.6510	1.4902_=<_x_<_1.5352	3	6.00%	
	Std dev. [Coef of variation]	0.1047 [0.0632]	1.5352_=<_x_<_1.5801	9	18.00%	
	MAD [MAD/STDDEV]	0.0901 [0.8608]	1.5801_=<_x_<_1.6251	8	16.00%	
	Min * Max [Full range]	1.45 * 1.89 [0.45]	1.6251_=<_x_<_1.6701	5	10.00%	
	1st * 3rd quartile [Range]	1.58 * 1.74 [0.17]	1.6701_=<_x_<_1.7150	5	10.00%	
	Skewness (std-dev)	0.0646 (0.3366)	1.7150_=<_x_<_1.7600	8	16.00%	
	Kurtosis (std-dev)	-0.8721 (0.6619)	1.7600_=<_x_<_1.8049	8	16.00%	
			1.8049_=<_x_<_1.8499	1	2.00%	
			x>=_1.8499	1	2.00%	

[3] http://en.wikipedia.org/wiki/Absolute_deviation

BODYMASS.

	Statistics	
BodyMass	Average	27.6806
	Median	26.9761
	Std dev. [Coef of variation]	5.8125 [0.2100]
	MAD [MAD/STDDEV]	4.0471 [0.6963]
	Min * Max [Full range]	17.63 * 48.49 [30.86]
	1st * 3rd quartile [Range]	24.02 * 29.65 [5.63]
	Skewness (std-dev)	1.5480 (0.3366)
	Kurtosis (std-dev)	4.3365 (0.6619)

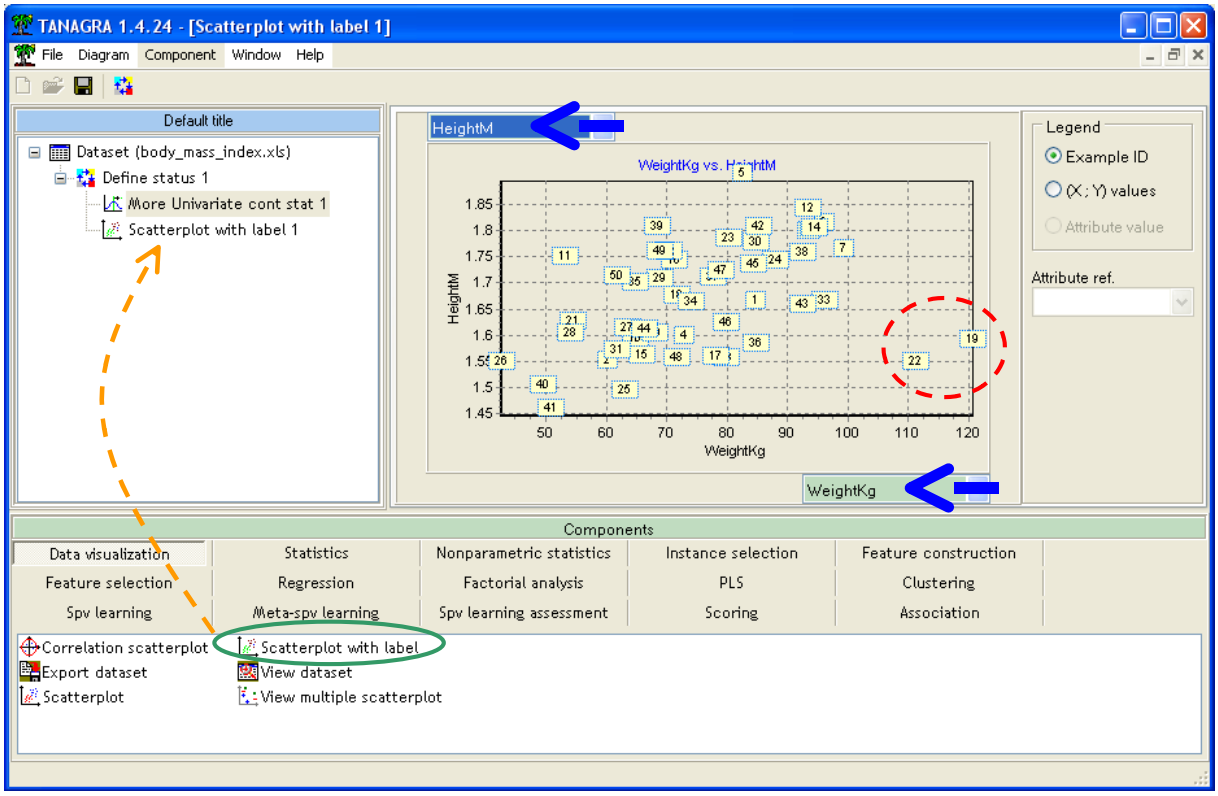
Values	Count	Percent	Histogram
x_<_20.7198	2	4.00%	
20.7198_=<_x_<_23.8059	9	18.00%	
23.8059_=<_x_<_26.8920	12	24.00%	
26.8920_=<_x_<_29.9782	15	30.00%	
29.9782_=<_x_<_33.0643	6	12.00%	
33.0643_=<_x_<_36.1504	4	8.00%	
36.1504_=<_x_<_39.2366	0	0.00%	
39.2366_=<_x_<_42.3227	0	0.00%	
42.3227_=<_x_<_45.4088	0	0.00%	
x>=_45.4088	2	4.00%	

Les distributions de WEIGHTKG et HEIGHTM n'appellent pas de commentaires particuliers. Il y a certes un étalement à droite, avec deux observations qui semblent se démarquer pour les deux variables (1 observation dans les deux dernières barres). Cela devient patent avec la variable BODYMASS, 2 observations à valeurs élevées s'écartent réellement des autres. On ne sait pas s'il s'agit des mêmes observations dans les 3 situations.

3.3 Représentation graphique

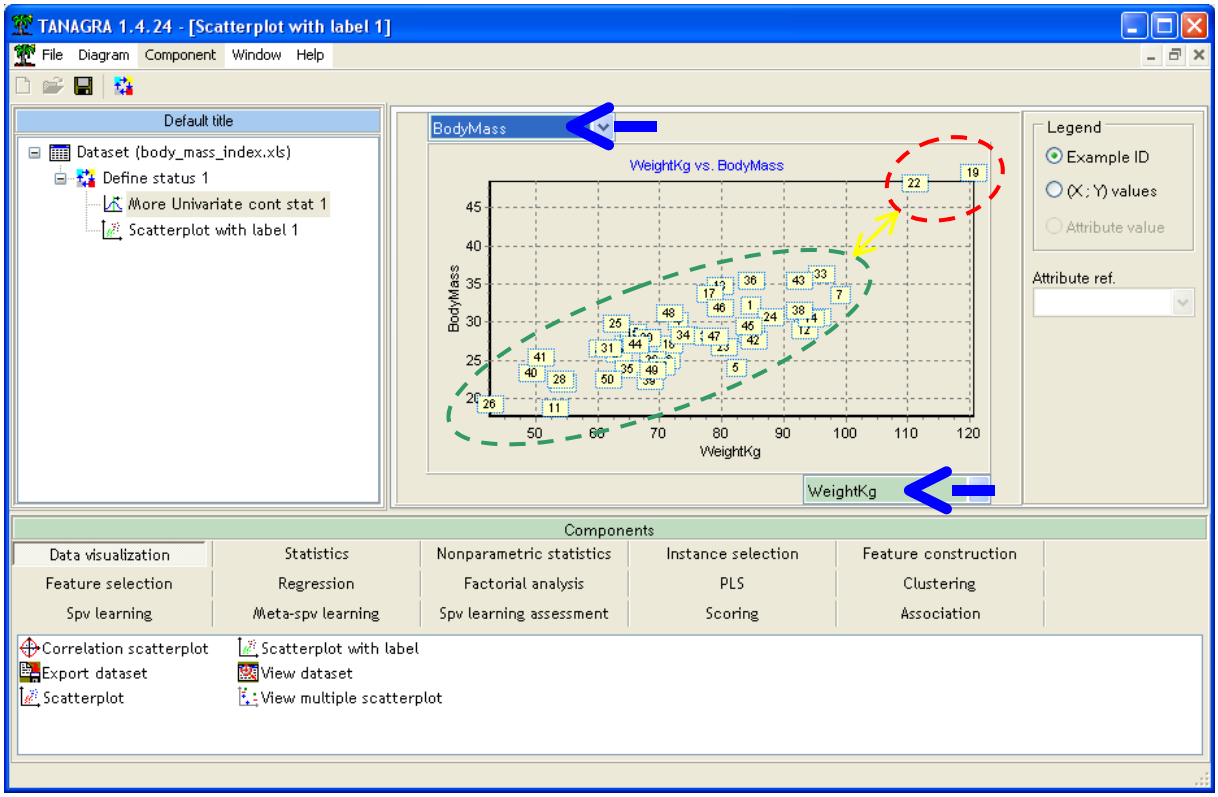
Une autre manière de visualiser les anomalies est de projeter les individus dans le plan, en croisant les variables 2 à 2. Des logiciels comme R (http://www.r-project.org/) le font très bien avec la commande **pairs()**. Notons que nous faisons déjà un pas vers le traitement multivarié dans ce cas, nous appréhendons le rôle conjoint de deux variables.

Dans Tanagra, nous insérons le composant SCATTERPLOT WITH LABEL (onglet DATA VISUALIZATION). Nous croisons tout d'abord les variables WEIGHTKG et HEIGHTM.



En un coup d'œil, il apparaît que les observations n°19 et n°22 sont douteuses si l'on considère le nuage de points. L'écartement est surtout imputable à la variable WEIGHTKG, il y a des individus qui pèsent lourd dans l'échantillon. Nous savons maintenant, par rapport à notre interrogation précédente (section 3.2), les 2 individus qui sont plus grands que les autres (HEIGHTM, dans les deux dernières barres de l'histogramme) ne sont pas ceux qui sont plus corpulents que les autres (dans les deux dernières barres de l'histogramme de la variable WEIGHTKG).

Croisons maintenant les variables WEIGHTKG et BODYMASS.

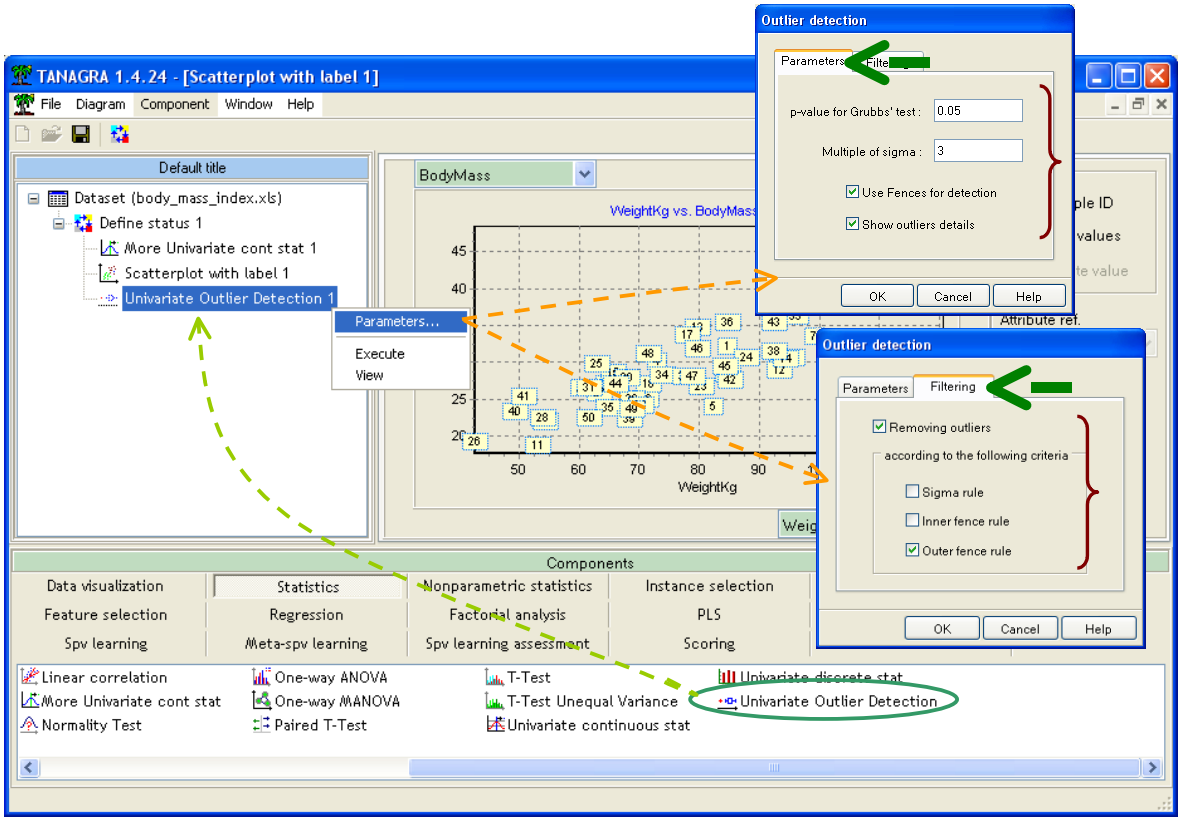


Il n'y a plus de doute, ces deux individus (n°19 et n°22) sont singulièrement dodus, surtout relativement à leur taille.

3.4 Détection et traitement des points aberrants

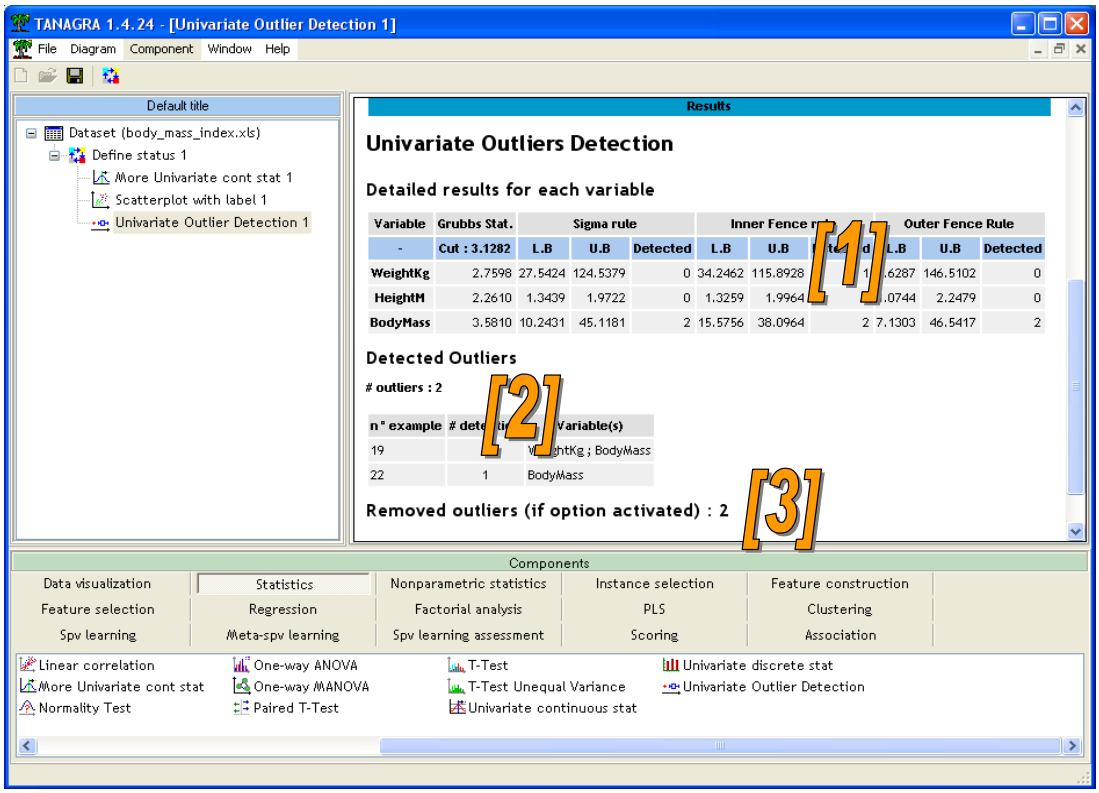
Le composant UNIVARIATE OUTLIER DETECTION identifie les observations qui s'écartent « significativement » des autres pour chaque variable. Il s'appuie sur les différents critères présentés en introduction. Il produit un tableau listant les observations incriminées. Il peut les exclure de l'ensemble des données selon le ou les combinaisons de critères que l'on choisit d'activer.

Nous insérons le composant dans le diagramme. Nous activons le menu contextuel PARAMETERS pour spécifier les paramètres du traitement.



Dans l'onglet PARAMETERS, nous choisissons d'afficher les individus détectés dans le rapport d'exécution. Dans l'onglet FILTERING, nous choisissons de supprimer de l'ensemble de données les individus atypiques, en nous basant uniquement sur le critère OUTER FENCE c.-à-d. retirer des données les individus extrêmement atypiques (voir section 1).

Nous validons ces paramètres et nous activons le menu VIEW pour accéder aux résultats.



Dans la première partie du rapport [1], nous observons les valeurs limites utilisées et le nombre d'observations atypiques détectées pour chaque critère.

Detailed results for each variable

Variable	Grubbs Stat.	Sigma rule			Inner Fence rule			Outer Fence Rule		
-	Cut : 3.1282	L.B	U.B	Detected	L.B	U.B	Detected	L.B	U.B	Detected
WeightKg	2.7598	27.5424	124.5379	0	34.2462	115.8928	1	3.6287	146.5102	0
HeightM	2.2610	1.3439	1.9722	0	1.3259	1.9964	0	1.0744	2.2479	0
BodyMass	3.5810	10.2431	45.1181	2	15.5756	38.0964	2	7.1303	46.5417	2

- Le test de Grubbs nous dit qu'au risque de 5%, la valeur la plus extrême de BODYMASS peut être considérée comme atypique.
- Selon la règle des 3-sigmas, nous détectons 2 données atypiques pour la variable BODYMASS.
- Selon la règle INNER FENCE, il y a 1 individu atypique pour WEIGHTKG, 2 pour BODYMASS.
- La règle OUTER FENCE produit le même résultat que 3-sigmas.

Dans la seconde partie [2], un tableau énumère les observations incriminées sur l'ensemble des critères.

Detected Outliers

 # outliers : 2

n° example	# detection	Variable(s)
19	2	WeightKg ; BodyMass
22	1	BodyMass

L'observation n°19 est atypique selon les variables WEIGHTKG et BODYMASS. Ce qui n'est guère étonnant lorsque l'on se remémore son positionnement dans les graphiques ci-dessus.

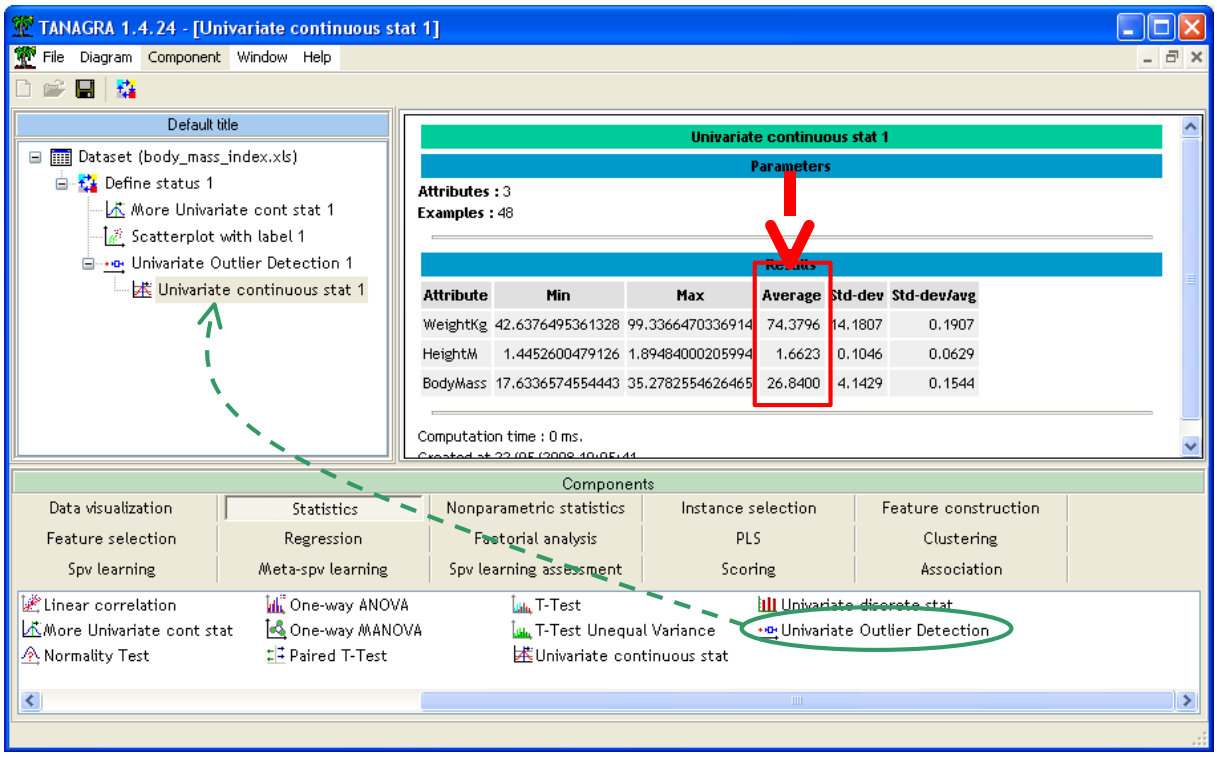
L'observation n°22 en revanche a été détectée uniquement pour la variable BODYMASS.

Enfin, dans la troisième partie du rapport [3], Tanagra nous indique qu'au final 2 individus ont été exclus selon les critères définis pour le filtrage, en l'occurrence seule la règle OUTER FENCE a été activée ici.

Removed outliers (if option activated) : 2

3.5 Statistiques descriptives (bis)

Pour évaluer l'influence des observations n°19 et n°22, recalculons les statistiques descriptives sur les observations restantes. Pour ce faire, nous insérons le composant UNIVARIATE CONTINUOUS STAT (onglet STATISTICS, il est plus rapide mais moins détaillé que celui utilisé plus haut).



Pour chaque variable, comparons la moyenne, indicateur sensible aux points atypiques, sur les 50 et 48 observations. La colonne « écart » nous indique que la présence de ces 2 observations affecte manifestement les résultats, surtout en ce qui concerne la variable BODYMASS.

Variable	Moyenne pour 50 obs.	Moyenne pour 48 obs. (sans n°19 et n°22)	Ecart (en %)
WEIGHTKG	76.0402	74.3796	+2.23 %
HEIGHTM	1.6581	1.6623	-0.25 %
MODYMASS	27.6806	26.8400	+3.13 %

4 Conclusion - Traitement des points aberrants

Notre composant choisit d'exclure les observations atypiques. C'est une solution possible mais ce n'est certainement pas la panacée. Il y a d'autres stratégies : la transformation des données, en rendant symétrique la distribution, on atténue l'écartement des queues de distribution ; une transformation plus radicale encore, le passage au rangs ; l'utilisation de techniques appropriées, peu sensibles aux points aberrants (ex. dans le data mining, plutôt qu'une analyse discriminante, on préfèrera les arbres de décision s'il y a profusion de points atypiques) ; etc[4].

Les techniques présentées dans ce document sont univariées, indépendantes du traitement statistique réalisé en aval. La situation devient plus complexe lorsque l'on veut tenir compte : (a) du rôle conjoint de plusieurs variables ; (b) évaluer l'effet de ces points sur la technique statistique mise en œuvre. On trouvera en ligne un exemple de traitement dans le cadre de la régression linéaire multiple[5] sous Tanagra.

[4] http://cc.uoregon.edu/cnews/spring2000/outliers.html

[5] http://tutoriels-data-mining.blogspot.com/2008/04/points-aberrants-et-influents-dans-la.html