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 (<u>http://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm</u>). Nous avons implémenté :

- Le test de Grubbs (<u>http://www.itl.nist.gov/div898/handbook/eda/section3/eda35h.htm</u>). 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×σ;m-3×σ]. 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 (<u>http://en.wikipedia.org/wiki/Box_plot</u>). 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 1^{er} quartile Q1 et le 3^{ème} 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 (<u>http://en.wikipedia.org/wiki/Image:Boxplot_vs_PDF.png</u>)

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.xls¹ 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 directe². Elle présente l'avantage de ne pas requérir la présence du logiciel EXCEL sur

¹ <u>http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/body_mass_index.xls</u>

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 : <u>http://tutoriels-data-mining.blogspot.com/2008/03/importation-fichier-xls-excel-macro.html</u>

la machine (voir : <u>http://tutoriels-data-mining.blogspot.com/2008/03/importation-fichier-xls-excel-mode.html</u>). 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.

Fie Dogam Window Help Otherwindow Help Open Save as Close Exit Diagram tile : Default tile Dataset (*txt anff*xls) : OK Cancel Help OK Cancel Help Platest (*txt anff*xls) : OK Cancel Help Platest (*txt anff*xls) : OK Cancel Help Platest (*txt anff*xls) : Dataset construction Feature scontruction Feature scontruction Feature scontruction Feature scontruction Fisture Scoring Correlation scatterplot Correlation scatterplot Scotte	TANAGRA 1.4.24									
Data visualization Statistics Feature construction PLS Clustering Source Science Data visualization Statistics Feature construction PLS Clustering Source Science Custering Spulearning assessment Scorring Scorring Custering Spulearning assessment Scorring Scorring Provisit recevul Non du fichier: Deta visualization Statistics Feature construction PLS Custering Spulearning assessment Scorring Scorring Provisit recevul Non du fichier: Deta visualization Statistics Feature construction Pate de traval Feature construction PLS Feature selectifi Correlation scatterplot Scatterplot Scatterplot Provisit recevul Feiner de type: Fervoits recevul Feiner de type: Feiner de type: Index recevul Feiner de type: Index recevul	File Diagram Window Help)								
Choose your dataset and start download Save as Close Ext Diagram tile : Default tile Data mining d Garam file name : Ext DiabateMiningDatabases_for_miningUdataset_for_soft_dev_and_com Dataset(*bd.* arfi.*xls) : CK Cancel Help Tanagra Regarder dans: Mes documents Wes documents Construction Feature construction Feature selecti PLS Clustering Spv learning assessment Scoring CCrrelation scatterplot Scatterpl	🗋 New 🗕 🗕 🗕			_						
Save as Close Ext Default title Default titl	🛱 Open	Choose your dataset a	nd start down	load						
Save as Close Exit Diagram title : Default title Data mining diagram file name : D/D ataMining/D atabases_for_mining/dataset_for_soft_dev_and_comes Dataset("t.ct."arff."xls) : CK Cancel Help Regarder dans: Regarder dans: Mes documents feature construction PLS Clustering Spv learning assessment Scotter plot Correlation scatter plot Scatter plot with Export dataset Scatter plot with Eavoirs reseaul Correlation scatter plot Scatter plot with Eavoirs reseaul Correlation scatter plot Scatter plot with Eavoirs reseaul Correlation scatter plot Scatter plot with Eavoirs reseaul Correlation scatter plot with Correlation scatter plot	Save									
Close Exit Data mining diagram file name : D'UbataMining/Databases_for_mining/dataset_for_soft_dev_and_com/lait Dataset ("txt".artf,"xls) : OK Cancel Help OK Cancel Help Regarder dans: Mes documents récents Bureau PL5 Clustering Spv learning assessment Scorte Poté de havait Poté de havait Export dataset Nem du lichies: body_mass_index.xls Durit Favoris réseau	Save as	Diagram title :								
Exit Data mining diagram file name : D:DataMining:Databases_for_mining:dataset_for_soft_dev_and_com/s Dataset (*bct*.aff(*xis) : OK Cancel Help Image: Concellent of the second of the	Close	Default title								
Exit D:DataMining/Databases_for_mining/dataset_for_soft_dev_and_com/a Dataset (*.txt*.arft*.xls): OK Cancel Help Pegader dars: Pegader		Data mining diagra	im file name :		No.					
Dataset (*.txt*.arff,*.xls): OK Cancel Help Regarder dars: Pagarder dars: Person of the second decomments Mes documents PL5 Clustering Spv learning assessment Scoring Poste de travel Poste de travel <t< td=""><td>Exit</td><td>D:\DataMining\Dat</td><td>tabases_for_mi</td><td>ning\dataset_for</td><td>_soft_dev_and_coma</td><td>3</td><td></td></t<>	Exit	D:\DataMining\Dat	tabases_for_mi	ning\dataset_for	_soft_dev_and_coma	3				
Data visualization Statistics Periode dare: Image: a construction Peature construction Feature construction Feature selection PLS Custering Spv learning assessment Scoring Sc		D-4			N					
OK Cancel Help OK Cancel Help Image a Image a Regarder dan: Image a Mes documents récents Bureau Data visualization Statistics Feature construction Feature selectri PL5 Clustering Spv learning assessment Scoring Poste de traval Image a Image a Image a Image a Image a Image a Image a		Dataset (".txt,".am,"	.xis) :			2				
OK Cancel Help Regarder dans: Periody_mass_index.xds Mes documents récents Bureau Data visualization Statistics Feature construction Feature selection Pus Clustering Spv learning assessment Scoring Poste de traval P					G	2				
OK Cancel Help Tanagra Regarder dans: Regarder dans: Wes documents récents Bureau Data visualization Statistics Peature construction Feature selectic PLS Clustering Spv learning assessment Scoring Poste de traval Poste de traval Poste de traval Feature scatterplot Scatterplot wit Scatterplot wit Feoris réseau Fichiers de type: Export dataset						-				
OK Cancel Help Image: Cancel Image: Cancel Image: Cancel Image: Cancel Image: Cancel Image: Cancel Image: Cancel Image: Cancel Image: Cancel Image: Cancel Image: Cancel Image: Cancel Image: Cancel Image: Cancel Image: Cancel Image: Cancel Image: Cancel Image: Cancel Image: Cancel Image: Cancel Image: Cancel Image: Cancel Image: Cancel Image: Cancel Image: Cancel Image: Cancel Image: Cancel Image: Cancel Image: Cancel Image: Cancel<	-									
Tanagra Regarder dans: Regarder dans: Wes documents Bureau Data visualization Statistics PLS Clustering Spv learning assessment Scoring Poste de traval Poste de traval Poste de traval Feature construction Feature selectio PLS Clustering Spv learning assessment Scoring Poste de traval Poste de traval Feature dataset Imagera Poste de traval Features teleplot Export dataset Imagera				OK OK	Cancel H	elp				
Regarder dans: Image: second se			Tanagra				? 🛛			
Data visualization Statistics Peature construction Feature selectici PLS Clustering Spv learning assessment Scoring Posle de travail Posle de travail Posle de travail Feature to not dichier: body_mass_index.xls Durin: Feature construction Feature construction Feature selectici Posle de travail Posle de travail Fosle de travail Feature dataset Image: Seature plot wite Feature selection Feature construction Feature selection Posle de travail Posle de travail Fosle de travail Feature selection Feature selection Feature selection Posle de travail Fosle de travail Feature selection Feature selection Feature selection Feature selection Feature selection Fosle de travail Fosle de travail Feature selection Feature selection </td <td></td> <td></td> <td>Regarder dans :</td> <td>utliers</td> <td></td> <td>🕑 🕝 🤣 📂 🛙</td> <td>-</td>			Regarder dans :	utliers		🕑 🕝 🤣 📂 🛙	 -			
Mes documents Mes documents Bureau Data visualization Statistics Feature construction Feature selection PLS Clustering Spv learning assessment Scoring Poste de travail Poste de travail Poste de travail Feature to not dataset Scorting Poste de travail Fordits réseau Fichiers de type: Excel Fie (37 & 2000)				Sody mass in	ndex.xls					
Mes documents récents Data visualization Statistics PLS Clustering Spv learning assessment Scoring Poste de travai Poste de travai Feature dataset Scotterplot wit Foring récents Foring récents Scotterplot wit Foring récents Foring récents Scotterplot wit Foring récents Foring récents Foring récents Scotterplot wit Foring récents Foring récents Foring récents Scotterplot wit Foring récents Scotterplot wit Foring récents Staterplot wit Foring récents Staterplot wit Foring récents Foring récents Foring récents Foring récents Foring récents Foring récents Foring réce										
Data visualization Statistics Data visualization Statistics Feature construction Feature selectici PLS Clustering Spv learning assessment Scoring Poste de travail Poste de travail Poste de travail Fichiers de type : Excel File (37 & 2000)			Mes documents récents							
Data visualization Statistics Data visualization Statistics Feature construction Feature selectric PLS Clustering Spv learning assessment Scoring Poste de travail Poste de travail Poste de travail Fichiers de type : Excel File (37 & 2000)										
Data visualization Statistics Data visualization Statistics Feature construction Feature selectic PLS Clustering Spv learning assessment Scoring Poste de travail Poste de travail Export dataset Scatterplot Scatterplot Scatterplot wit Favoris réseau Fichiers de type: Excel File (37 & 2000) Annuler										
Data visualization Statistics Data visualization Statistics Feature construction Feature selectic PLS Clustering Spv learning assessment Scoring Poste de travail Poste de travail Poste de travail Nom du fichier : body_mass_index.xls Poste de travail Fichiers de type : Excel File (37 & 2000)	l		Bureau							
Data visualization Statistics Feature construction Feature selectic PLS Clustering Spv learning assessment Scoring Correlation scatterplot Scatterplot Socatterplot Scatterplot Scatterplot Scatterplot wit Favoris réseau Fichiers de type: Excel File (37 & 2000) Annuler		Charlest -								
Peature construction Feature selectic Mes documents PLS Clustering Spv learning assessment Scoring Correlation scatterplot Scatterplot Spot learning assessment Scotterplot Scoring Poste de travail Poste de travail Nom du fichier : body_mass_index.xls Power dataset Duvrir Favoris réseau Fichiers de type : Excel File (37 & 2000)	Data visualization	Statistics								
PLS Clustering Spv learning assessment Scoring Correlation scatterplot Scatterplot Sport dataset Scatterplot wit Favoris réseau Foibre de travail Foibre de travail Nom du fichier : body_mass_index.xls Duvrir Annuler	Feature construction	Feature selectio	Mes documents							
Spv learning assessment Scoring Correlation scatterplot Scatterplot Export dataset Scatterplot wit Favoris réseau Foite de travail Foite de travail Nom du fichier : body_mass_index.xls Ouvrir Favoris réseau Excel File (37 & 2000)	PLS	Clustering								
Correlation scatterplot Export dataset Scatterplot wit Favoris réseau Foste de travail Foste de travail Nom du fichier : body_mass_index.xls Duvrir Excel File (97 & 2000) Annuler	Spv learning assessment	Scoring	_ 							
Export dataset 🖉 Scatterplot wit	Correlation scatterplot	🧖 Scatterplot	Poste de travail							
Favoris réseau Fichiers de type : Excel File (97 & 2000) Annuler	Export dataset	🔣 Scatterplot wit	S	Nom du fichier :	hodu mass index vis		Ouvrir			
Pavois ieseau i nuiveis de gype . Caxel nie (37 & 2000) Anntole				Fichiers de tuno :	Eucol File (97 t 2000)					
			r avoris reseau	nomers de type :	Excel File (97 & 2000)	-				

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]	1 ^{er} et 3 ^{ème} 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				
	Stati	stics	Values	Count	Percent	Histogram	
	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]			44 0004		
	MAD [MAD/STDDEV]	13.1113 [0.8110]	58.2412_=<_x_<_66.0430	8	16.00%		
	Min*Max [Full range]	42.64 * 120.66 [78.02]	66.0430_=<_x_<_73.8448	11	22.00%		
WeightKg	1st * 3rd quartile [Range]	64.86 * 85.28 [20.41]	73.8448_=<_x_<_81.6466	7	14.00%		
	Skewness (std-dev)	0.3574 (0.3366)	81 6466 =< v < 89 4483	7	14 00%		
	Kurtosis (std-dev)	0.1363 (0.6619)	01.0400_41_X_1_07.4400	,	14.00%	_	
			89.4483_=<_x_<_97.2501	8	16.00%		
			97.2501_=<_x_<_105.0519	1	2.00%		
			105.0519_=<_x_<_112.8537	1	2.00%		
			x>=_112.8537	1	2.00%		

HEIGHTM.

	Stati	stics	Values	Count	Percent	Histogram
	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]				
	MAD [MAD/STDDEV]	0.0901 [0.8608]	1.5352_=<_x_<_1.5801	9	18.00%	
	Min*Max (Full range)	1.45 * 1.89 [0.45]	1.5801_=<_x_<_1.6251	8	16.00%	
HeightM	1st * 3rd quartile [Range]	1.58 * 1.74 [0.17]	1.6251_=<_x_<_1.6701	5	10.00%	
-	Skewness (std-dev)	0.0646 (0.3366)	1 4701 - × ×	F	10.00%	
	Kurtosis (std-dev)	-0.8721 (0.6619)	1.0/01<_X_<_1./150		10.00%	
			1.7150_=<_x_<_1.7600	8	16.00%	
			1.7600_=<_x_<_1.8049	8	16.00%	
			1.8049_=<_x_<_1.8499	1	2.00%	
			x>=_1.8499	1	2.00%	I

³ <u>http://en.wikipedia.org/wiki/Absolute_deviation</u>

BODYMASS.

	Statistics		Values	Count	Percent	Histogram
	Average	27.6806	x_<_20.7198	2	4.00%	
	Median	26.9761	20.7198 =< x < 23.8059	9	18.00%	
	Std dev. [Coef of variation]	5.8125 [0.2100]				
	MAD [MAD/STDDEV]	4.0471 [0.6963]	23.8059_=<_x_<_26.8920	12	24.00%	
	Min*Max [Full range]	17.63 * 48.49 [30.86] 24.02 * 29.65 [5.63]	26.8920_=<_x_<_29.9782	15	30.00%	
BodyMass	1st * 3rd quartile [Range]		29.9782_=<_x_<_33.0643	6	12.00%	
	Skewness (std-dev)	1.5480 (0.3366)	22.0642 26.4504		0.00%	
	Kurtosis (std-dev)	4.3365 (0.6619)	33.0043_=K_X_K_30.1504	4	0.00%	
			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 (<u>http://www.r-project.org/</u>) 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 rul	e	Inr	ner Fence	rule	Ou	ter 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.



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) ; etc⁴.

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 multiple⁵ sous Tanagra.

⁴ <u>http://cc.uoregon.edu/cnews/spring2000/outliers.html</u>

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