SY09 - TP01
Statistique descriptive, Analyse en composantes principales

Cristian GARRIDO
Avril 3, 2014

OBJETIF

L'objectif de ce TP est de manipuler des données afin de mettre en évidence les principaux éléments qui ont une influence sur des variables (temps de gestation, poids des bébés, taille de crabes). La méthode de l'ACP nous aidera dans l'analyse des tableaux de données.

1 STATISTIQUE DESCRIPTIVE

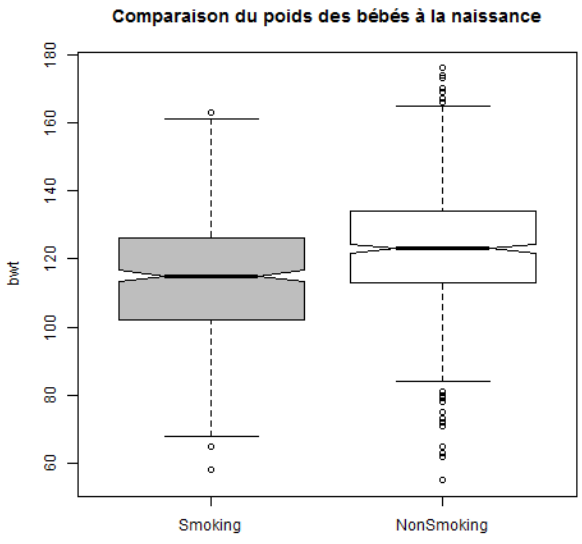
1.1 DONNES BABIES

Dans cet exercice, nous disposons d'un jeu de données babies constitué de 1236 bébés décrits par 23 variables. Dans notre tp, nous n'utilisons que 8 variables : 5 quantitatives (le poids à la naissance, la durée de gestation, le nombre de grossesses précédentes, la taille de la mère et le poids de la mère) et 3 qualitatives (l''âge de la mère, si la mère fume ou non et le niveau d'éducation de la mère).

1.1.1 Quelle est la différence de poids entre les bébés nés de mères qui fumaient durant leur grossesse et celles qui ne fumaient pas ?

Pour observer cette différence, nous pouvons étudier le résumé des valeurs du poids des bébés en fonction du fait que leur mère fumait ou non :

	Min	1er Quartile	Médiane	Moyenne	3e Quartile	Max
Mère fumeuse	55	113	123	123	134	176
Mère non fumeuse	58.0	102.0	115.0	114.1	126.0	163.0



On remarque que les bébés issus de mères non fumeuses ont en moyenne un poids supérieur aux bébés issus de mères fumeuses; comme le montre la figure ci-contre. On réalise un test de Student pour vérifier l'hypothèse selon laquelle les deux moyennes sont différentes, on obtient les résultats suivants :

t = -8.5813, df = 1003.19, p-value < 2.2e-16

95 percent confidence interval:

-10.98148, -6.89385

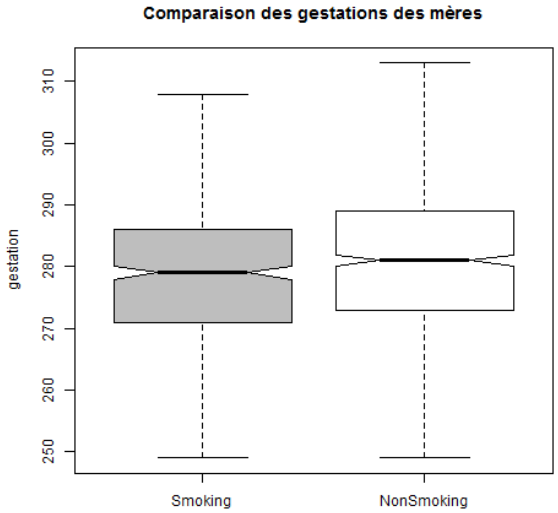
Sample estimates: mean of x: **114.1095** and mean of y: **123.047**

Donc, on peut rejeter l'hypothèse selon laquelle les deux moyennes sont égales à 95%.

1.1.2 Est-ce qu'une mère qui fume durant sa grossesse est enclin à avoir un temps de gestation plus court qu'une mère qui ne fume pas?

On remarque ici que le temps de gestation est plus élevé chez les mères non fumeuses que chez les mères fumeuses, cette différence reste tout de même très légère. Grâce au test de Student, on obtient les résultats suivants :

	Min	1er Quartile	Médiane	Moyenne	3e Quartile	Max
Mère fumeuse	223	271	279	278	286	330
Mère non fumeuse	148.0	273	281.0	280.2	289.0	353.0



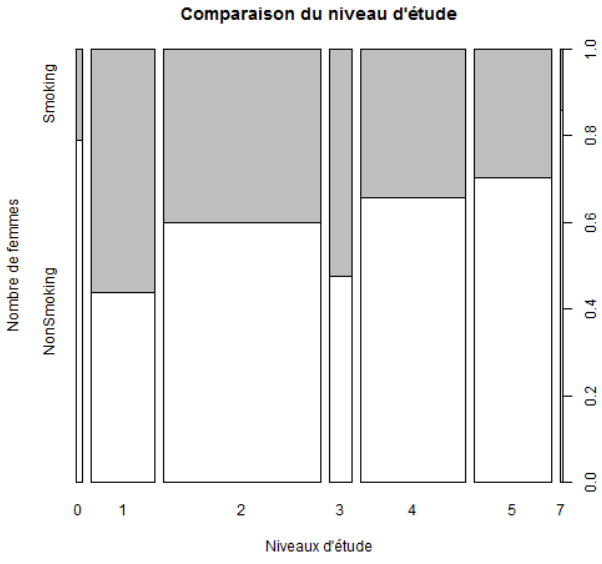
t = -2.3939, df = 1092.553, p-value = 0.01684
alternative hypothesis**: true difference in means is not equal to 0**

95 percent confidence interval: -4.0173229
-0.3981501

sample estimates: mean of x **277.9792**
mean of y **280.1869**

Ceci nous permet de refuser l'hypothèse selon laquelle les deux moyennes sont égales.

1.1.3 Le niveau d'étude a-t-il une influence sur le fait que la mère soit fumeuse?



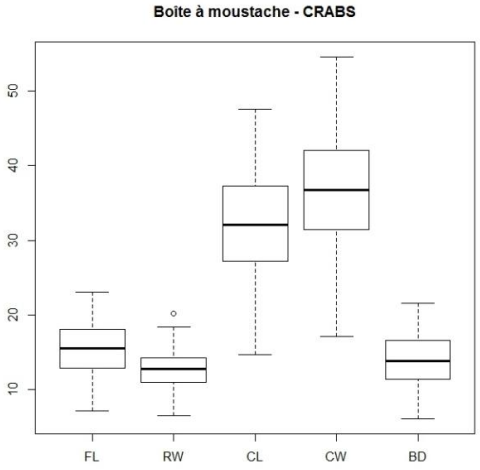
On constate sur cette figure que la proportion de femmes non fumeuse augmente avec le niveau d'éducation. Au niveau 1, on a plus de femmes fumeuse, mais en évoluant jusqu'au niveau 5, la proportion de femmes non fumeuses est plus élevée.

L'étude présentée nous explique que les femmes qui fument sont plus propices à avoir des bébés avec un faible poids à la naissance, ce qui correspond à nos analyses. Par contre, il explique également que le fait de fumer n'influe pas sur le temps de gestation de la femme. On a pu noter dans notre analyse qu'il y avait une très légère différence au niveau du temps de gestation et grâce à notre test de Student, il ressort que le temps de gestation et le statut de fumeuse sont très liés, ce qui ne contredit l'analyse faite dans le texte. Cela pourrait être lié à notre échantillon qui reste très limité et qui ne peut pas véritablement traduire et caractériser un traitement général.

1.2 DONNÉES CRABS

Le jeu de données proposé est constitué de 200 crabes décrits par huit variables dont cinq sont quantitatives.

1.2.1 Analyse descriptive des données. Existe-t-il des différences de caractéristiques morphologiques selon l'espèce ou le sexe ? Semble-t-il possible d'identifier l'espèce ou le sexe d'un crabe à partir d'une ou plusieurs mesures de ces caractéristiques ?



Summary CRABS

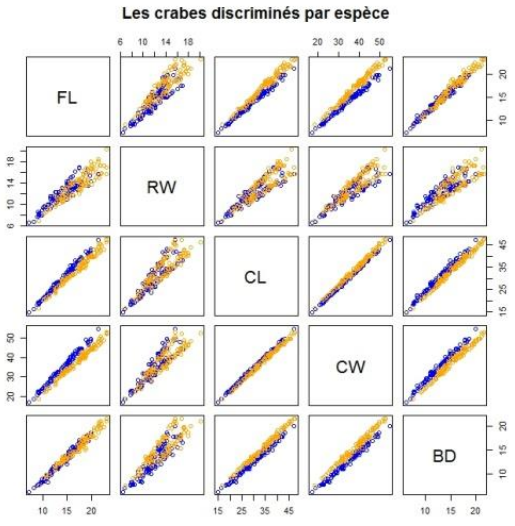
	FL	RW	CL	CW	BD
Min	7.20	6.50	14.70	17.10	6.10
1er Qu.	12.90	11.00	27.27	31.50	11.40
Médiane	15.55	12.80	32.10	36.80	13.90
Moyenne	15.58	12.74	32.11	36.41	14.03
3e Qu.	18.05	14.30	37.23	42.00	16.60
Max	23.10	20.20	47.60	54.60	21.60

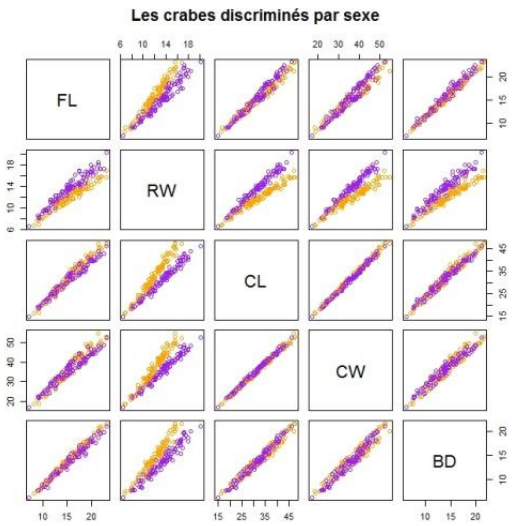
Dans le cas présent, on étudie une population de 200 crabes discriminée par 7 variables dont deux sont qualitatives : le sexe et l'espèce. Effectuons l'analyse descriptive des données.

Le diagramme en boîte montre clairement que toutes les mesures de FL (lobe frontal) et BD (profondeur du corps) sont très proches. Il en va de même pour CW (largeur de la carapace) et CL (longueur de la carapace). De plus, il semble que toutes les variables soient fortement corrélées entre elles.

Corrélation

	FL	RW	CL	CW	BD
FL	1.0000				
RW	0.9069	1.0000			
CL	0.9788	0.8927	1.0000		
CW	0.9649	0.9004	0.9950	1.0000	
BD	0.9876	0.8892	0.9832	0.9678	1.0000





Il n'existe aucune différence morphologique selon l'espèce : les nuages de points sont trop proches pour en observer.

En revanche, pour RW et CL, on distingue les mâles des femelles. De même pour le couple de variables CW et RW.

1.2.2 Corrélation entre les différentes variables. Quelle en est vraisemblablement la cause? Quel traitement est-il possible d'appliquer aux données pour s'affranchir de ce phénomène?

Les variables sont très corrélées positivement entre elles (cf. le tableau des corrélations ci-dessus). Il semble donc que "l'effet taille" pollue notre analyse. Pour s'affranchir de ce problème, il convient de supprimer une variable. Pour ce faire, on norme les variables restantes par la variable que l'on supprime.

2 L'ANALYSE EN COMPOSANTES PRINCIPALES (ACP)

2.1. EXERCICE THÉORIQUE

On dispose de la matrice suivante X. A l'aide de la fonction « Centre.R » définie plus tôt, nous allons centrer notre matrice :

$$X = \begin{pmatrix} 3 & 4 & 3 \\ 1 & 4 & 3 \\ 2 & 3 & 6 \\ 2 & 1 & 4 \end{pmatrix}$$ Après avoir centré la matrice, on obtient $$Xc = \begin{pmatrix} 1 & 1 & -1 \\ -1 & 1 & -1 \\ 0 & 0 & 0 \\ 0 & -2 & 0 \end{pmatrix}$$

Nous allons, par la suite, travailler sur notre matrice centrée.

2.1.1 Axes factoriels – Inerties expliquées

- Calculer les axes vectoriels revient à calculer les vecteurs propres de la matrice de covariance $S = \frac{1}{n} Xc'Xc$ où n représente le nombre d'individus de nos données ; n = 4. Après calcul, on obtient donc la matrice de covariance suivante :

$$S = \begin{pmatrix} 0.5 & 0 & 0 \\ 0 & 1.5 & -0.5 \\ 0 & -0.5 & 1.5 \end{pmatrix}$$

Nous allons maintenant utiliser la commande « eigen » de R pour trouver les vecteurs propres qui correspondent aux axes factoriels de cette matrice, on obtient :

$$U = \begin{pmatrix} 0.0000000 & 0.0000000 & 1 \\ -0.7071068 & 0.7071068 & 0 \\ 0.7071068 & 0.7071068 & 0 \end{pmatrix}$$

- Pour calculer les inerties expliquées par chaque axe factoriel, nous allons utiliser les valeurs propres de la matrice S. On a le vecteur suivant $L = (2, 1, 0.5)$ qui correspond au vecteur des valeurs propres de la matrice.

$$I1 = \frac{2}{2+1+0.5} * 100 = 57.1428 \quad I2 = \frac{1}{2+1+0.5} * 100 = 28.57143$$
$$I3 = \frac{0.5}{2+1+0.5} * 100 = 14.28571$$

	1er axe factoriel	2ème axe factoriel	3ème axe factoriel
Inertie expliquée	57.14286	28.57143	14.28571
Inertie expliquée cumulée	57.14286	85.71429	100

On peut noter ici que le premier axe factoriel est le meilleur car son inertie expliquée est la plus grande. A travers l'inertie cumulée, on peut sélectionner les deux premiers axes qui représentent 86 % de l'information.

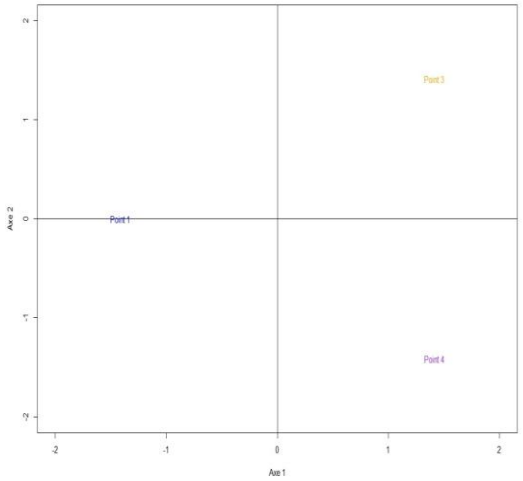
2.1.2 Composantes principales – Représentation dans le premier plan factoriel

Les composantes principales sont obtenues à l'aide de la formule ci-dessous :

$$C = XcMU$$

Où Xc est la matrice centrée, M est la métrique qui correspond au vecteur identité et U est le vecteur des vecteurs propres. On obtient ainsi

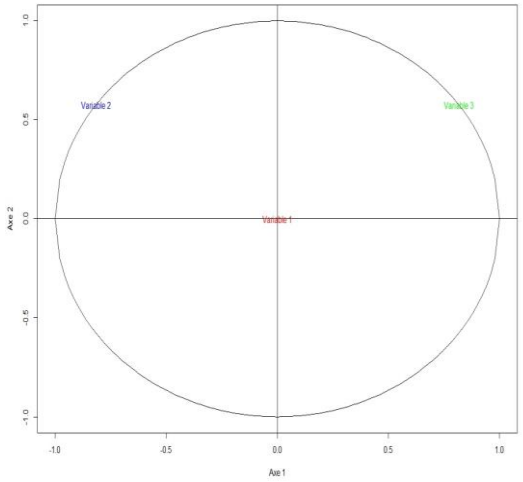
$$C = \begin{pmatrix} -1.4142140 & 2.220446e-16 & 1 \\ -1.414214 & 2.220446e-16 & -1 \\ 1.414214 & 1.414214e+00 & 0 \\ 1.414214 & -1.414214e+00 & 0 \end{pmatrix}$$



On peut voir ici que le point 1 et le point 2 sont confondus, ce qui est normal car ils ont les mêmes coordonnées dans le nouveau repère. La seule manière de différencier ces deux individus est d'utiliser le troisième axe factoriel.
Par ailleurs, les deux autres individus sont symétriques.

2.1.3 Représentation des variables dans le premier plan factoriel

Nous allons construire la matrice de corrélation en utilisant la commande « cor » entre la matrice Xc et la matrice C, nous obtenons la figure ci-dessous.



On remarque que les variables **Variable2** et **Variable 3** sont placées sur le cercle, ce qui montre qu'elles sont bien représentées dans ce plan. A l'inverse, on perd toutes les informations relatives à la **Variable1**.

2.1.4 Calcul de l'expression $\sum_{\alpha}^{k} c_{\alpha} u'_{\alpha}$ pour les valeurs k=1,2 et 3

K=1	K=2	K=3
$\boldsymbol{X} = \begin{pmatrix} \mathbf{0} & \mathbf{1} & \mathbf{-1} \\ \mathbf{0} & \mathbf{1} & \mathbf{-1} \\ \mathbf{0} & \mathbf{-1} & \mathbf{1} \\ \mathbf{0} & \mathbf{-1} & \mathbf{1} \end{pmatrix}$	$X = \begin{pmatrix} 0 & 1 & -1 \\ 0 & 1 & -1 \\ 0 & 0 & 2 \\ 0 & -2 & 0 \end{pmatrix}$	$X = \begin{pmatrix} 1 & 1 & -1 \\ -1 & 1 & -1 \\ 0 & 0 & 2 \\ 0 & -2 & 0 \end{pmatrix}$

On remarque que pour k=3, on reconstitue la matrice Xc d'origine.

2.2 UTILISATION DES OUTILS R.

En utilisant les fonctions de R, on retrouve facilement les résultats de l'ACP:

- ACP <- princomp(notes) réalise l'ACP et nous retourne l'écart-type de chaque composante principale
- ACP$loadings retourne les axes factoriels (vecteurs propres de la matrice de variance)
- APC$scores retourne les composantes principales
- biplot(ACP) projette les données et les caractères dans le plan principal formé des deux axes factoriels de poids maximal.
- plot(ACP) représente l'histogramme des variances des composantes principales.

plot (ACP) nous permet de constater graphiquement que quasiment toutes les données sont représentées sur le premier plan factoriel. En effet, la variance cumulée de ces deux plans atteint plus de 90%. Ceci est suffisant pour s'assurer que notre analyse de données sera solide.

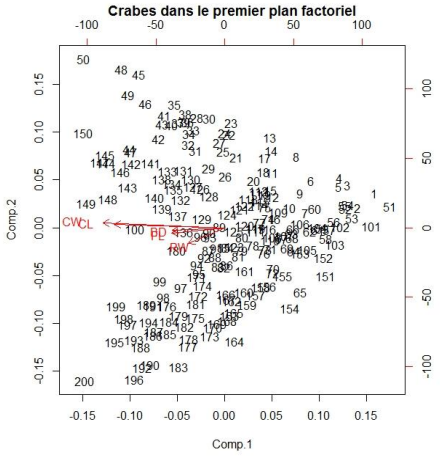
Biplot(ACP) représente tous les individus et les variables (vecteurs rouges) dans le premier plan factoriel. Ceci nous permet d'évaluer graphiquement les corrélations entre variables (angle formé par les deux vecteurs): les vecteurs de deux variables indépendantes forment un angle de 90°.

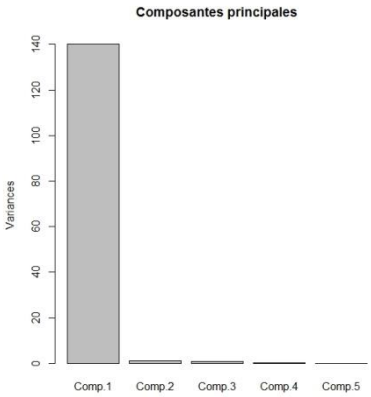
2.3 TRAITEMENT DES DONNÉES CRABS

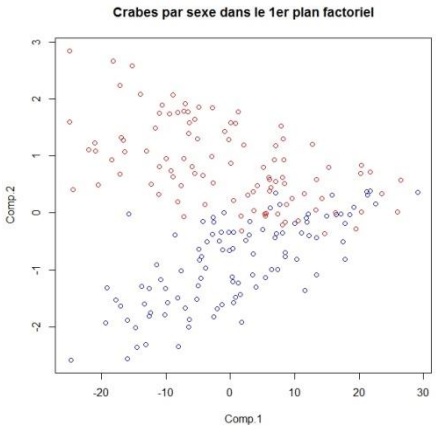
Cette étude vise à utiliser l'ACP pour trouver une représentation des CRABS qui permettent de distinguer visuellement différents groupes, liés à l'espèce et au sexe.

2.3.1 Tester tout d'abord l'ACP sur **crabsquant,** sans traitement préalable. Que constatez-vous?

Sur cette représentation des crabes dans le premier plan factoriel, nous constatons que toutes les variables sont de même signe sur le premier axe. Les variables sont donc très corrélées entre elles. Par ailleurs, la quasi-totalité de l'information peut être représentée sur le premier axe factoriel. On observe clairement " l'effet taille " mentionné précédemment.

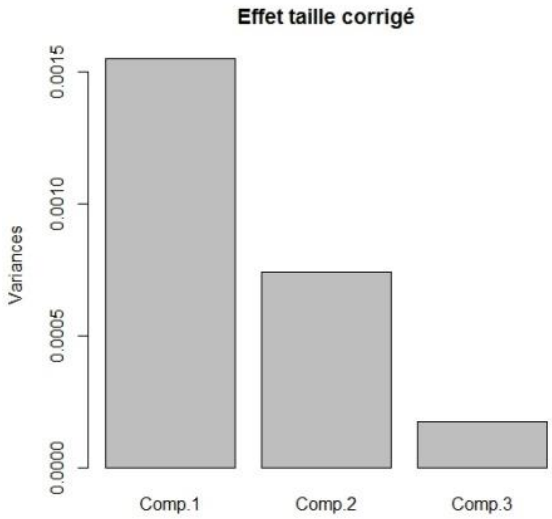






En séparant les crabes selon leur genre, on distingue deux groupes. Cette ACP nous permet donc tout de même de séparer les crabes selon leur sexe. Toutefois, l'analyse n'est pas concluante lorsque l'on cherche à mettre en évidence les deux espèces.

2.3.2 Trouver une solution pour améliorer la qualité de votre représentation en termes de visualisation des différents groupes.

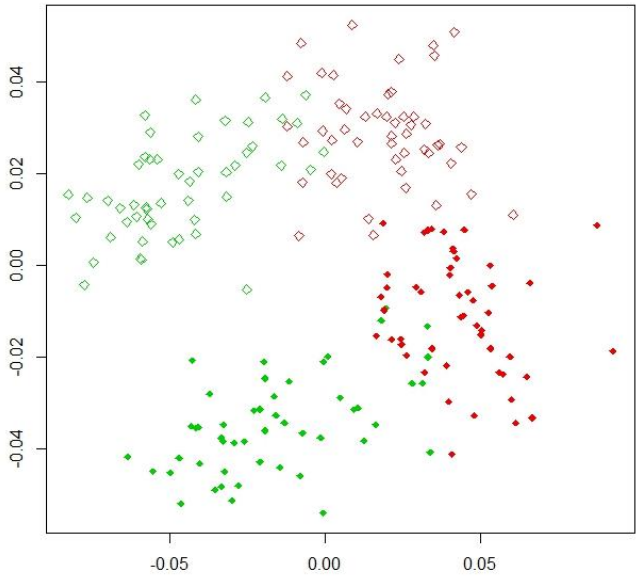


A cause de "l'effet taille", l'ACP réalisée ne nous permet pas d'analyser les données. Pour estomper cet "effet taille", il suffit de définir de nouvelles variables de la façon suivante: supprimer une variable puis normer les variables restantes.

Dans notre cas, on supprime CL qui est redondante par rapport à CW. Ainsi, les données pourront être correctement représentées dans le premier plan factoriel. De plus, BD et FL sont très corrélées, on peut supprimer FL. Il nous reste donc 3 variables: RW, CW et BD.

Grâce à cette solution, l'information est mieux représentée selon les nouvelles composantes.

Ainsi, on observe 4 groupes distincts de crabes sur le premier plan factoriel. On peut clairement séparer les crabes selon leur espèce et leur sexe.



Représentation des crabes dans le 1er plan factoriel (effet taille corrigé)

3 CONCLUSION

L'ACP est une méthode factorielle très puissante qui permet de visualiser des données très denses. Cependant, on a vu dans ce TP que la méthode est faillible. Il convient donc de l'utiliser avec prudence et d'analyser intelligemment les résultats obtenus.