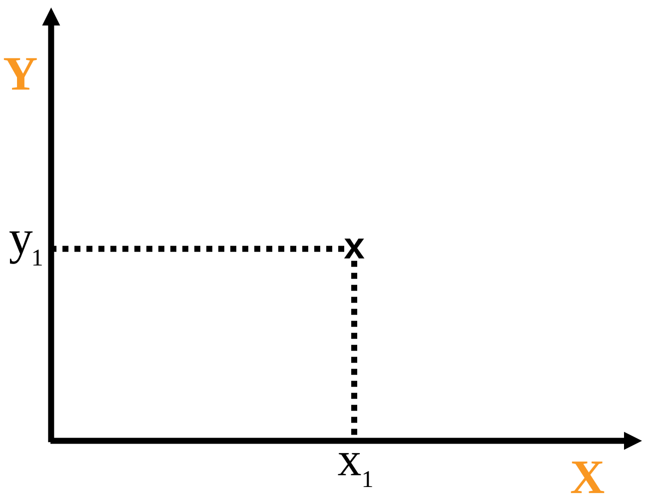
ANALYSES PREALABLES
A UNE ACP

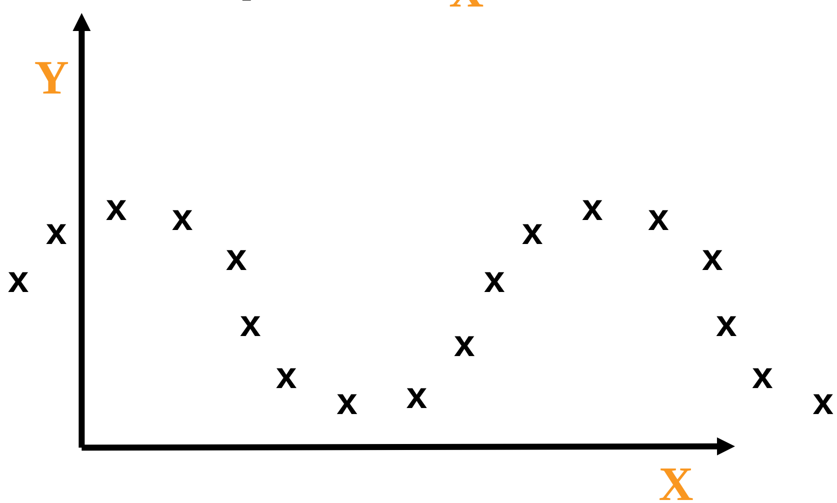
Pierre-Louis GONZALEZ

I COVARIANCE - CORRELATION

1 Nuage de points

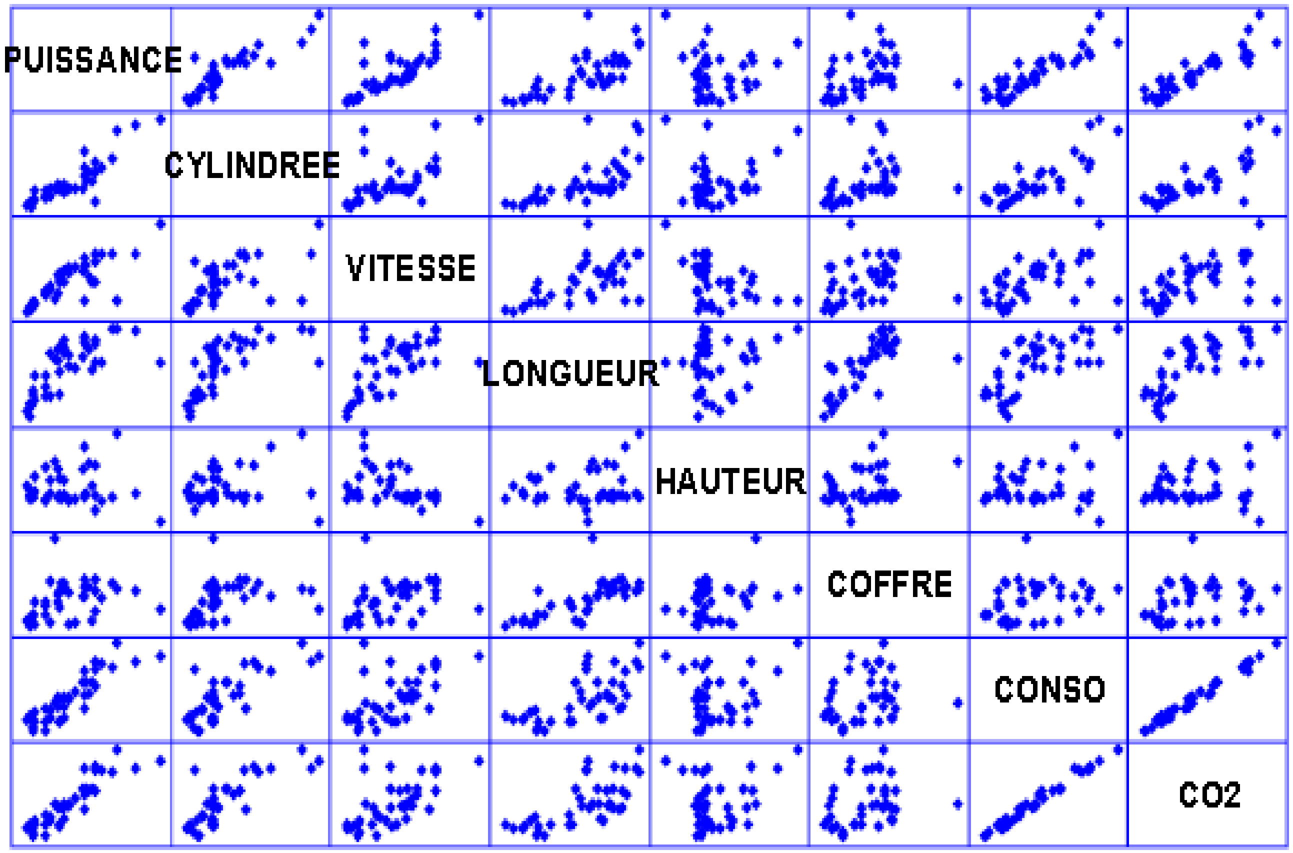


Liaison nulle
ou
indépendance

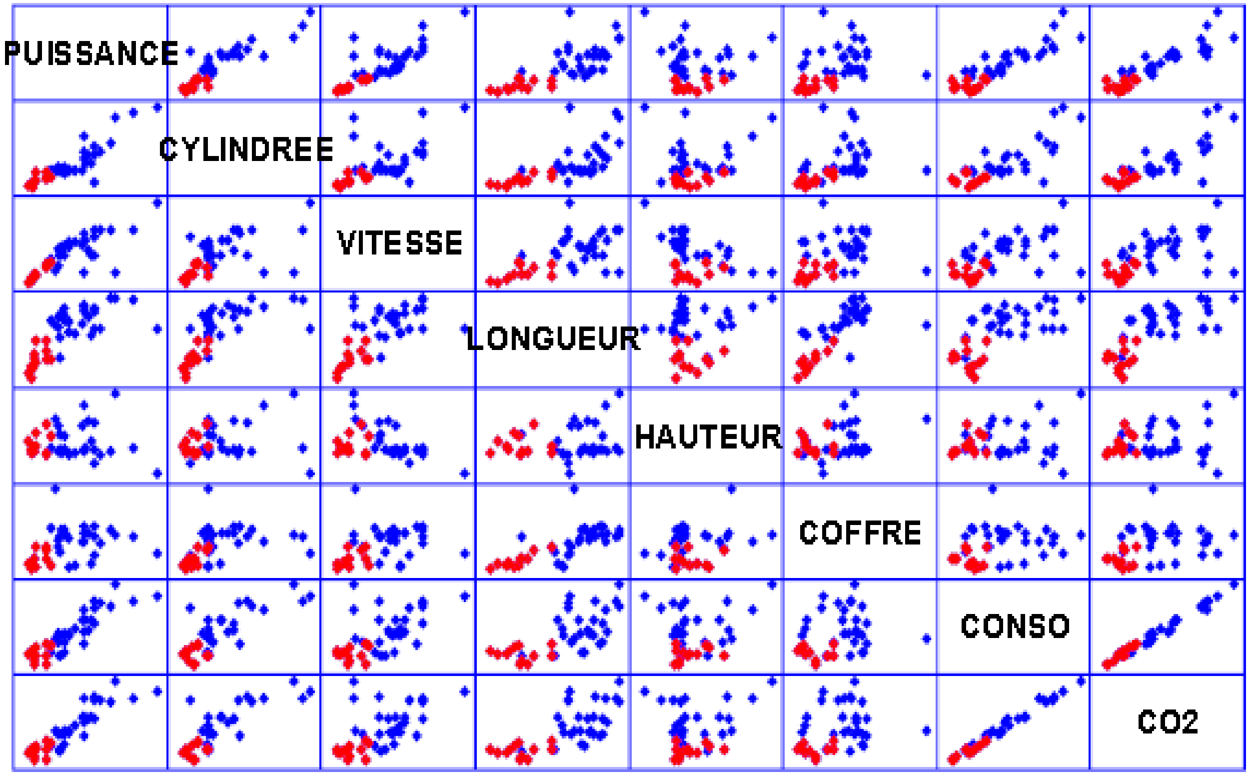


Dépendance fonctionnelle
de y en x
(mais pas de x en y)

Galerie de nuages de points



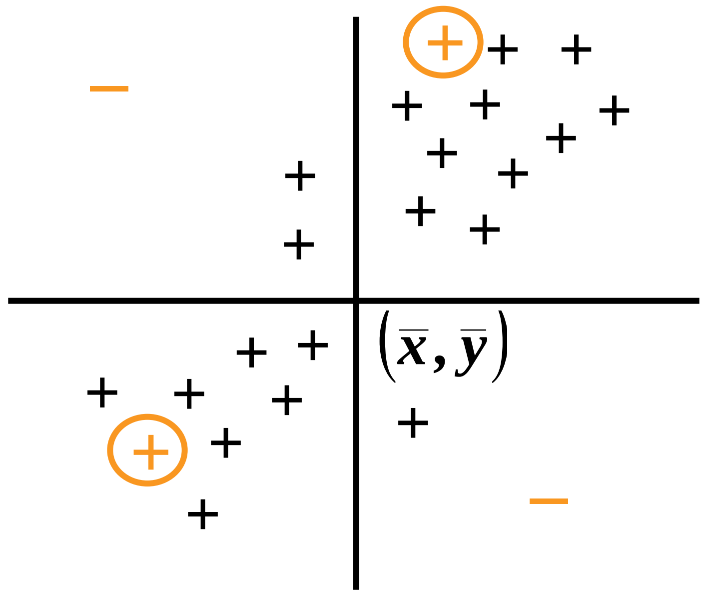
Galerie de nuages de points; identification des véhicules dont le prix est inférieur à 22000 €



2 La covariance

$$\mathrm{Cov}\,(X,Y) = \frac{1}{n}\sum\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)$$

$$\mathrm{Cov}\,(X,Y) = \left[\frac{1}{n}\sum x_i y_i\right] - \bar{x}\,\bar{y}$$



Propriétés

- La covariance est un indicateur de la dispersion des points autour du point moyen $(\bar{x}, \bar{y})$.

  Si X et Y ont tendance à varier dans le même sens :

  → $Cov(X,Y) > 0$

  Si X et Y ont tendance à varier dans le sens contraire :

  → $Cov(X,Y) < 0$

- $Cov(X,X) = \text{Variance de } X$

3 Coefficient de corrélation linéaire

$$r_{XY} = \frac{\frac{1}{n}\sum\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{s_X s_Y}$$

Remarque : Le numérateur est la covariance de ***X*** et ***Y***.

ATTENTION ***r*** ne mesure que la linéarité de la liaison entre ***X*** et ***Y***.

L’absence de liaison linéaire n’exclut pas d’autre forme de liaison.

Ex : $y = x^2$

Propriétés

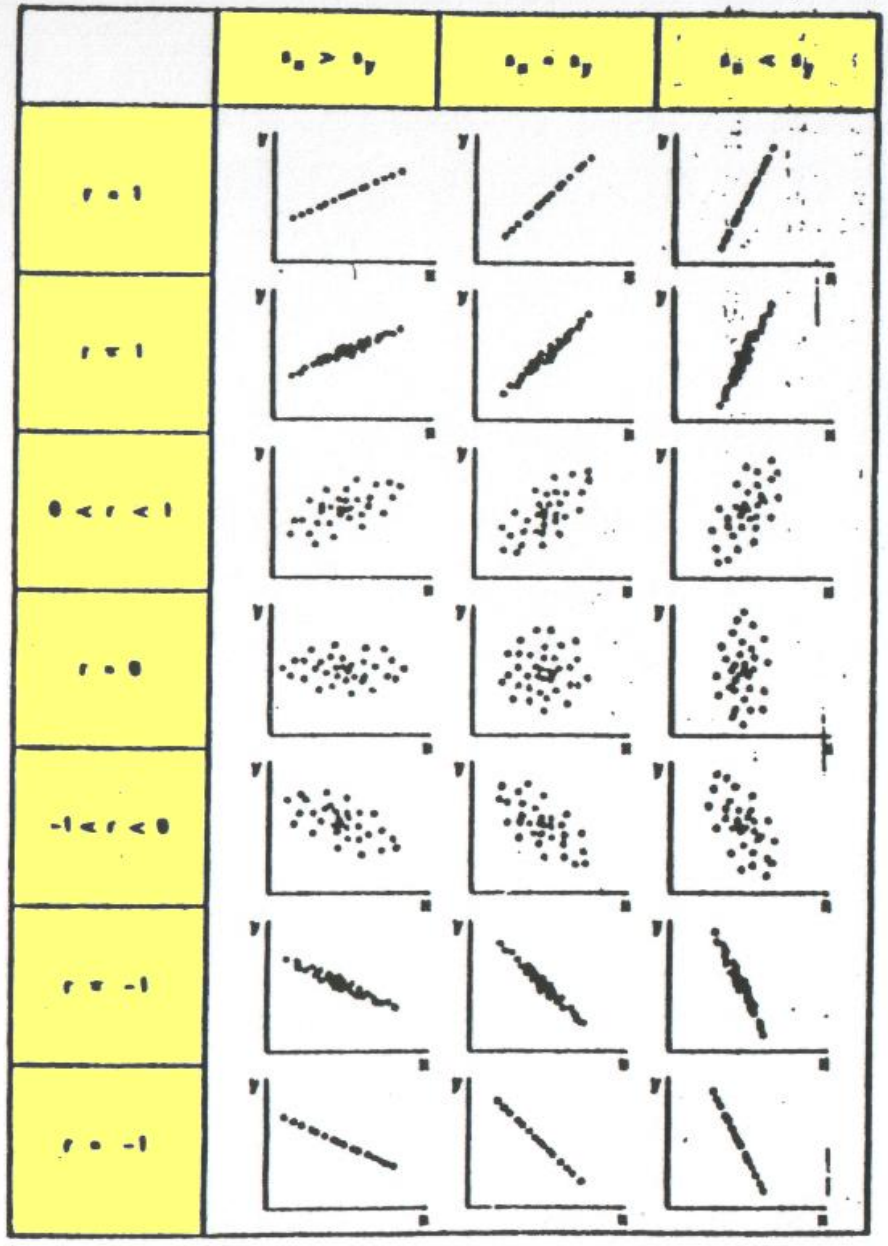
- $-\boldsymbol{1} \leq \boldsymbol{r} \leq +\boldsymbol{1}$

- $\boldsymbol{r} = \boldsymbol{0} \Rightarrow$ pas de liaison linéaire
  (cela ne signifie pas que les variables ***X*** et ***Y*** sont indépendantes)

- Par contre, l'indépendance entraîne la non corrélation linéaire.

Exemple 1

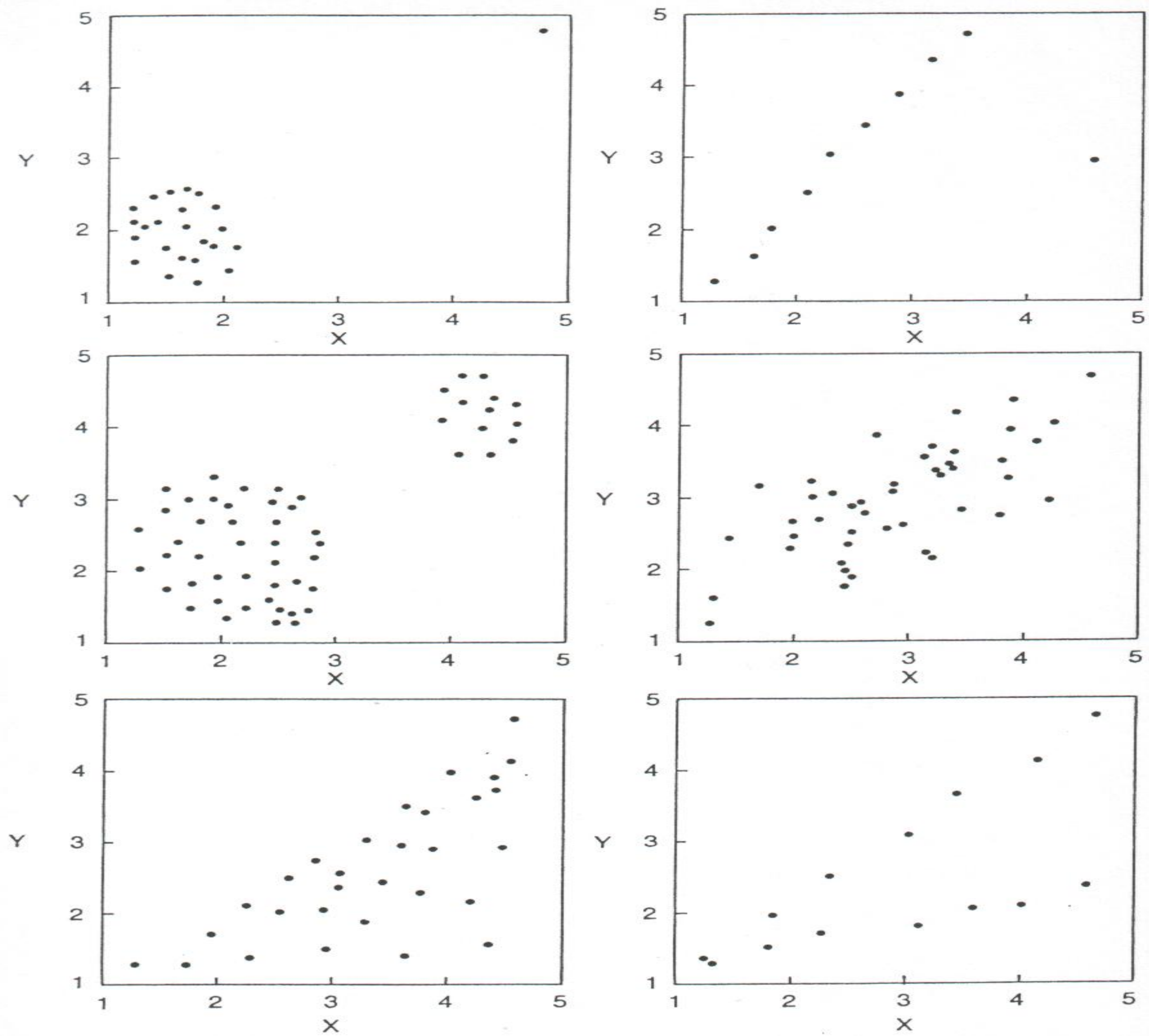
1 La figure 1 donne quelques formes typiques de nuages de points en relations avec les caractéristiques s_x et s_y (écarts-types des caractères) et r (coefficient de corrélation entre les caractères). Mais on ne doit pas en déduire que r résume de façon exhaustive la liaison entre deux caractères quantitatifs.



sx = sy
sx < sy
r = 1
r ≈ 1
0 < r < 1
r = 0
-1 < r < 0
r ≈ -1
r = -1

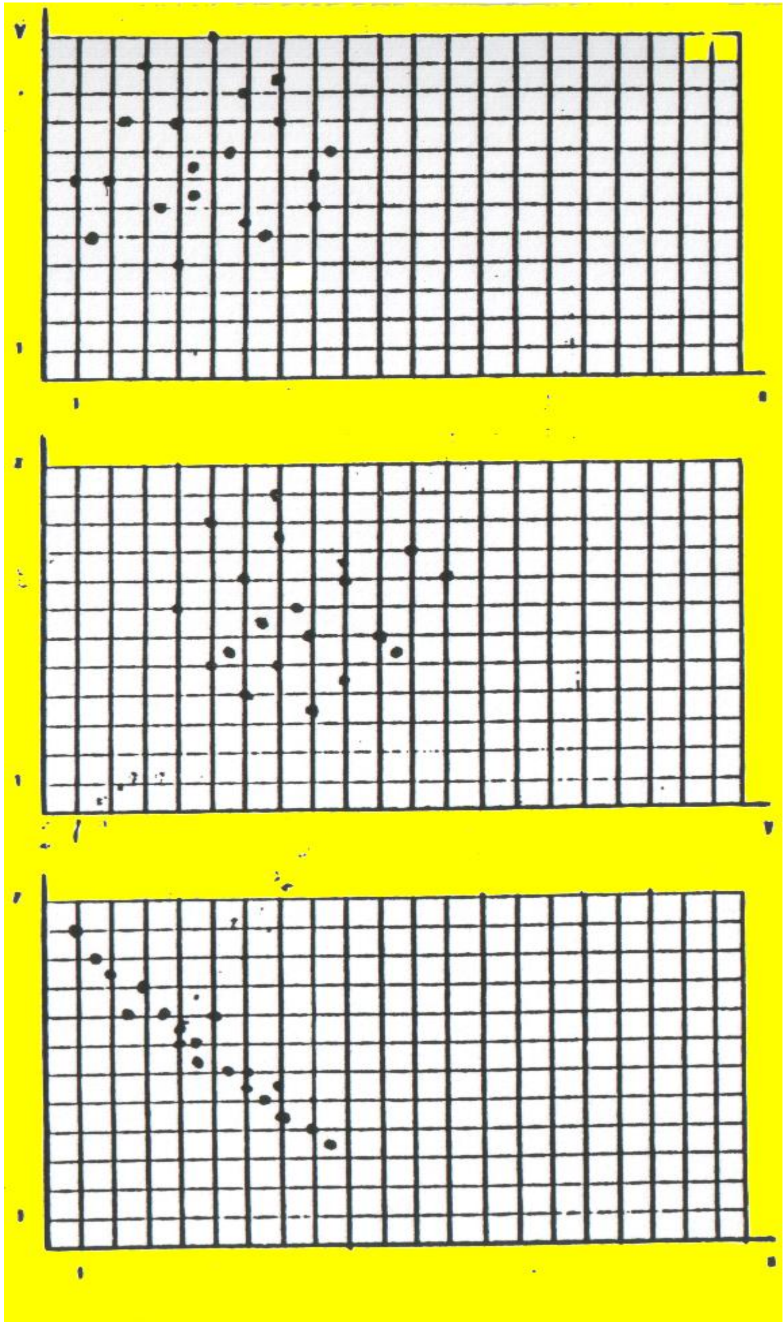
Exemple 2 Méfiez-vous de la corrélation

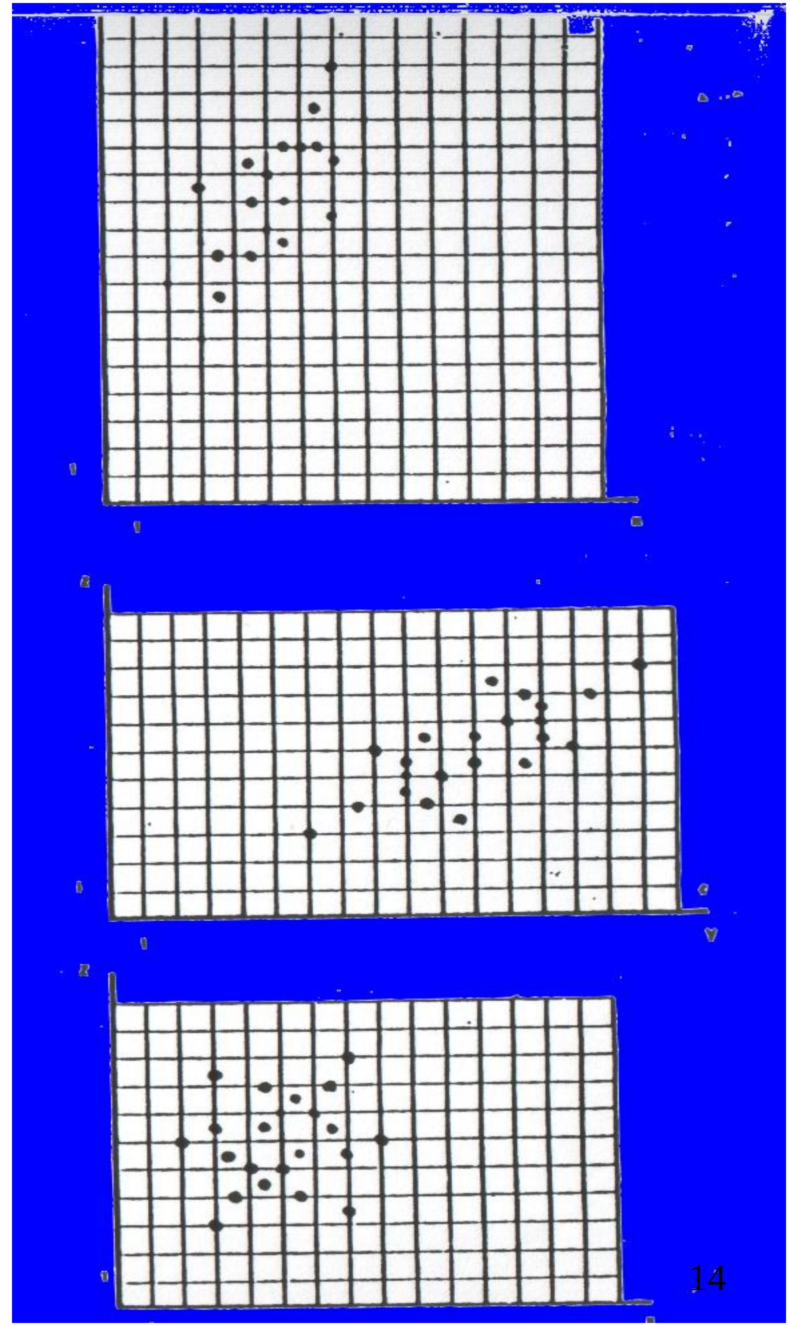
A titre illustratif essayez, sans faire aucun calcul, d'appréhender la valeur du coefficient de corrélation dans les cas suivants :


Y
X
1 2 3 4 5

Exemple 3

La corrélation n’est pas en général une propriété transitive. Pour s’en convaincre, essayez d’appréhender les corrélations entre les couples formés par trois caractères quantitatifs dans les cas proposés aux figures 7 et 8. Là encore, on essaiera de dégager une « morale » de chacun des exemples proposés.





4 Caractère significatif d'un coefficient de corrélation

Principe Si les $\boldsymbol{n}$ observations avaient été prélevées au hasard dans une population où $\boldsymbol{X}$ et $\boldsymbol{Y}$ sont indépendants $(\Rightarrow \rho = 0)$ quelles seraient les valeurs possible de $\boldsymbol{r}$?

Lorsque $\rho = 0$ et que les observations proviennent d'un couple Gaussien :

$$\boxed{\frac{\boldsymbol{R}\sqrt{\boldsymbol{n}-\boldsymbol{2}}}{\sqrt{\boldsymbol{1}-\boldsymbol{R}^2}} \quad \text{suit une loi } \boldsymbol{T}_{\boldsymbol{n}-\boldsymbol{2}} \quad \text{(loi de Student à } \boldsymbol{n}-\boldsymbol{2} \text{ d.d.l.)}}$$

Approximation

$\boldsymbol{R}$ significativement $\neq \boldsymbol{0}$

si $\quad |\boldsymbol{R}| > \dfrac{\boldsymbol{2}}{\sqrt{\boldsymbol{n}+\boldsymbol{2}}}$ au seuil $\quad \boldsymbol{\alpha} = \boldsymbol{5}\,\%$

valable si $\quad \boldsymbol{n} \geq \boldsymbol{30}$

Exemple de corrélations

Corrélations

	PUISSANCE	CYLINDREE	VITESSE	LONGUEUR	HAUTEUR
PUISSANCE		0,8862 (40) 0,0000	0,8041 (40) 0,0000	0,7106 (40) 0,0000	0,0025 (40) 0,9879
CYLINDREE	0,8862 (40) 0,0000		0,5583 (40) 0,0002	0,6594 (40) 0,0000	0,2109 (40) 0,1915
VITESSE	0,8041 (40) 0,0000	0,5583 (40) 0,0002		0,6090 (40) 0,0000	-0,4450 (40) 0,0040
LONGUEUR	0,7106 (40) 0,0000	0,6594 (40) 0,0000	0,6090 (40) 0,0000		0,1820 (40) 0,2609
HAUTEUR	0,0025 (40) 0,9879	0,2109 (40) 0,1915	-0,4450 (40) 0,0040	0,1820 (40) 0,2609	
COFFRE	0,3230 (40) 0,0420	0,3408 (40) 0,0314	0,3035 (40) 0,0569	0,7062 (40) 0,0000	0,2722 (40) 0,0893
CONSO	0,8905 (40) 0,0000	0,7863 (40) 0,0000	0,5808 (40) 0,0001	0,6644 (40) 0,0000	0,1827 (40) 0,2591
CO2	0,9024 (40) 0,0000	0,8100 (40) 0,0000	0,5803 (40) 0,0001	0,7140 (40) 0,0000	0,2278 (40) 0,1575

5 Interprétation d'un coefficient de corrélation linéaire

Lorsque l'on observe une corrélation entre deux variables, élevée, on peut être en présence de l'un des quatre cas suivants.

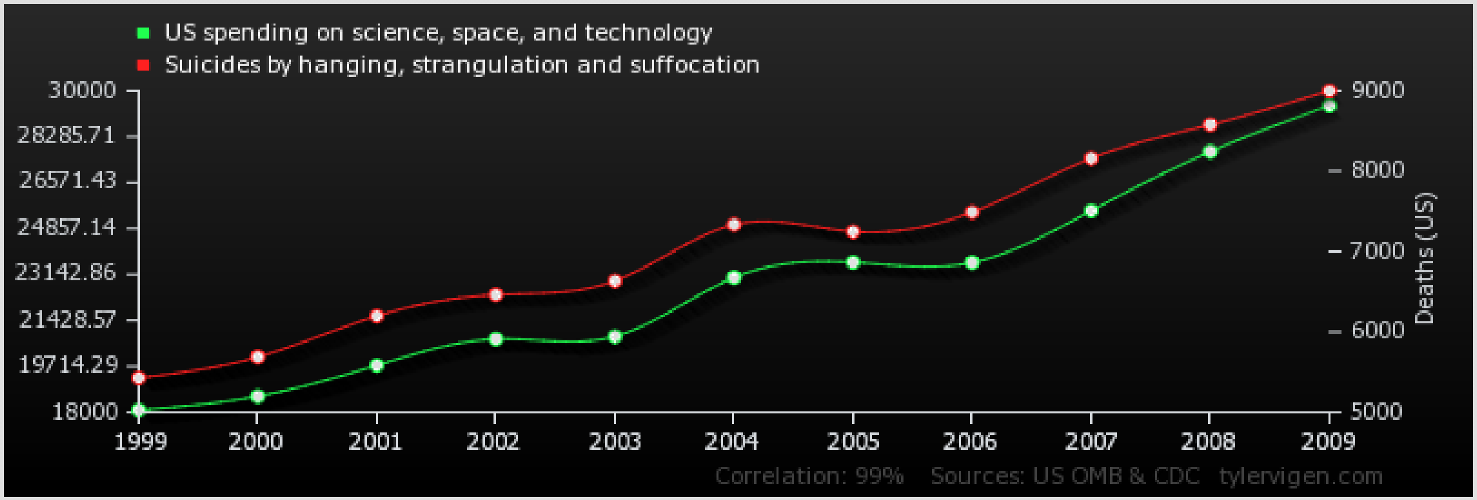
- a - relation de cause à effet.
  X implique ***Y***
- b - relation simultanée (réciproque) entre ***X*** et ***Y***
- c - les variations concomitantes de ***X*** et ***Y*** sont induites par une troisième variable
- d - corrélation due au hasard

http://www.tylervigen.com/

US spending on science, space, and technology

correlates with

Suicides by hanging, strangulation and suffocation



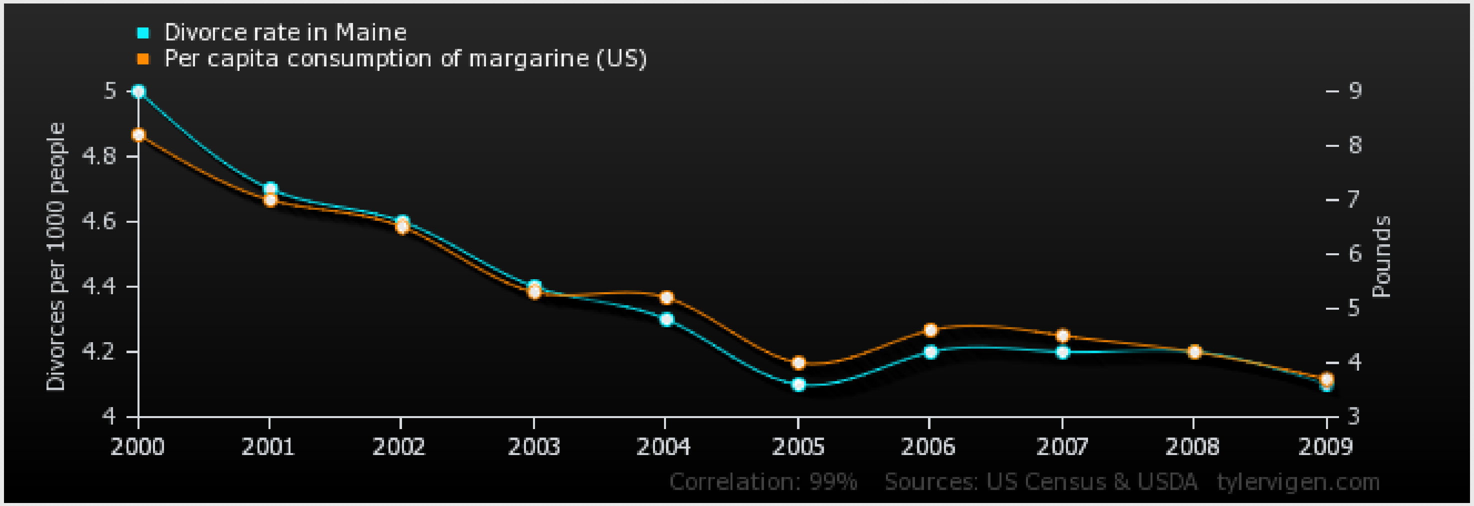
	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
US spending on science, space, and technology *Millions of todays dollars (US OMB)*	18,079	18,594	19,753	20,734	20,831	23,029	23,597	23,584	25,525	27,731	29,449
Suicides by hanging, strangulation and suffocation *Deaths (US) (CDC)*	5,427	5,688	6,198	6,462	6,635	7,336	7,248	7,491	8,161	8,578	9,000

Correlation: 0.992082

Divorce rate in Maine

correlates with

Per capita consumption of margarine (US)



	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Divorce rate in Maine *Divorces per 1000 people (US Census)*	5	4.7	4.6	4.4	4.3	4.1	4.2	4.2	4.2	4.1
Per capita consumption of margarine (US) *Pounds (USDA)*	8.2	7	6.5	5.3	5.2	4	4.6	4.5	4.2	3.7

Correlation: 0.992558

Coefficients de corrélation des rangs

Coefficient de Spearman

X	x_1	x_2	x_3 …… .x_i ……. x_n
Rang	r_{x1}	r_{x2}	r_{x3} …… r_{xi} ……. r_{xn}
Y	y_1	y_2	y_3 …… .y_i ……. y_n
Rang	r_{y1}	r_{y2}	r_{y3} …… r_{yi} ……. R_{yn}

$$R_S(X,Y) \;= \text{Corr}\,(R_X, R_Y)$$

La définition de R_S comme coefficient de corrélation linéaire sur des rangs implique que :

$R_S = 1$ Les 2 classements sont identiques

$R_S = -1$ Les 2 classements sont inverses l’un de l’autre

$R_S = 0$ Les 2 classements sont indépendants

Des tests sur coefficient de corrélation de Spearman indiquent, en fonction de la taille de l’échantillon , à partir de quelle taille il y a concordance ou discordance.

Remarque: En cas d’ex-aequo on utilise la règle du rang moyen.

Coefficient de Kendall

Quand deux juges i et j comparent deux objets A et B:

Si $A < B$ pour le juge i ou $A > B$ pour le juge i
et $A < B$ pour le juge j et $A > B$ pour le juge j
on dit qu'il y a concordance

Si $A < B$ pour le juge i ou $A > B$ pour le juge i
et $A > B$ pour le juge j et $A < B$ pour le juge j
on dit qu'il y a discordance

$$\boxed{R_K = \frac{2}{n(n-1)}(C-D)}$$

où C = nombre de concordances
et D = nombre de discordances

Propriétés

1. Comme le coefficient de Spearman, le coefficient de Kendall est un **coefficient de dépendance monotone**, compris entre -1 et +1

2. On démontre que si l'hypothèse d'indépendance entre les deux variables est vraie, la distribution d'échantillonnage de R_K est approximativement normale:

$$R_K \rightarrow N(0, \sqrt{\frac{2(2n+5)}{9n(n-1)}})$$

   Ceci est vrai dès que n>8, ce qui est un avantage pratique sur le coefficient de Spearman.

3. **Comparaison des coefficients de corrélation des rangs et du coefficient de Pearson**

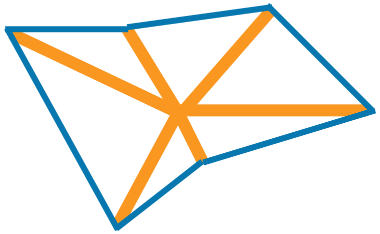
   Si R_P est nettement supérieur à R_S et R_K la corrélation linéaire risque d'être due à des point atypiques.

   Si au contraire R_S ET R_K sont nettement supérieurs à R_P, la relation entre les deux variables est non linéaire

II REPRÉSENTATION DES INDIVIDUS SOUS FORME D'ICÔNES

- diagrammes en étoiles
- diagrammes en rayons de soleil
- diagrammes en bâtons
- diagrammes circulaires
- Visages de Chernoff
- profils
- coordonnées parallèles

1 Diagramme en étoiles



Chaque individu est représenté par une étoile dont les rayons sont proportionnels aux valeurs de l'individu pour les différentes variables.

Pour la variable ***j***, le calcul de la longueur du rayon pour l'individu ***i*** se fait selon l'expression :

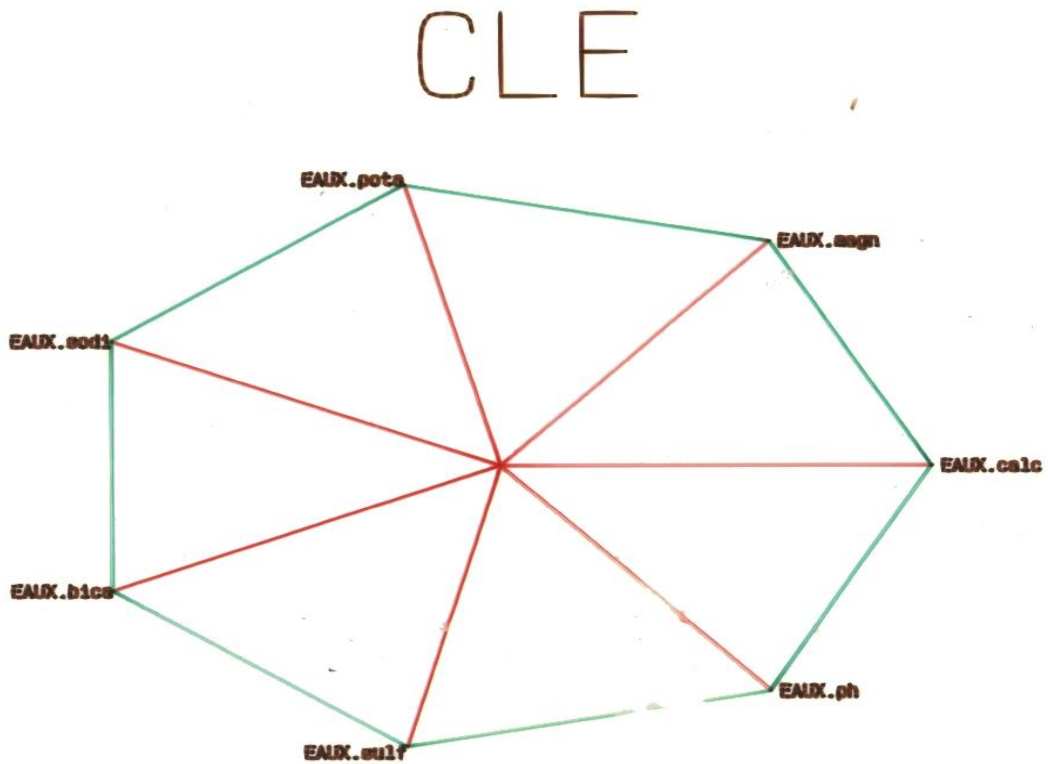
$$\boxed{x_{ij}^{*} = \left(1 - c\right)\frac{x_{ij} - \min\limits_{i} x_{ij}}{\text{étendue var}\left(j\right)} + c}$$

$$\underset{var(j)}{\text{étendue}} = \max_{i} x_{ij} - \min_{i} x_{ij} \quad j \text{ fixé}$$

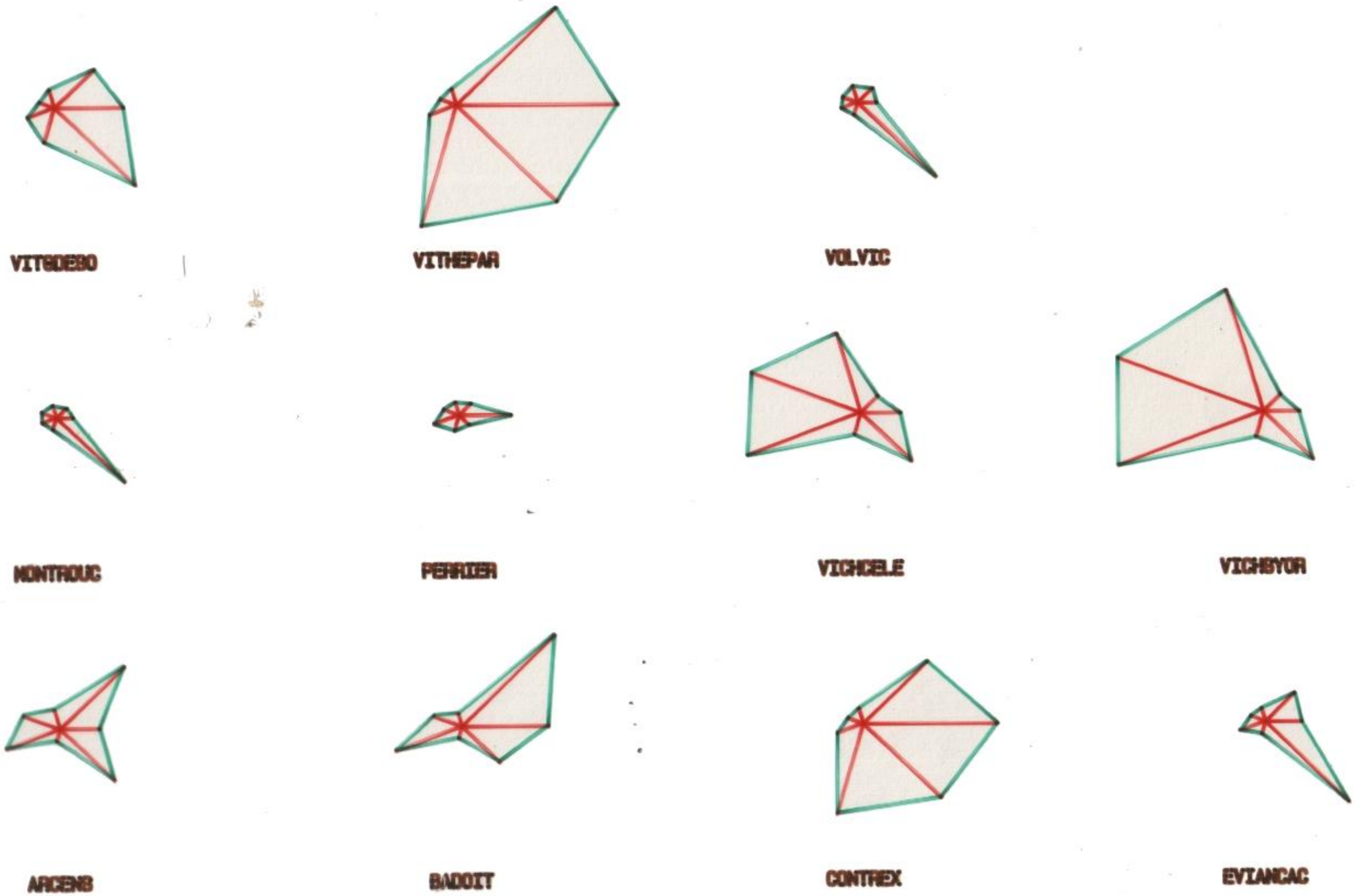
c est une constante que l'on choisit entre ***0*** et ***1***

valeur conseillée : $c = 0{,}1 \text{ ou } 0{,}2$

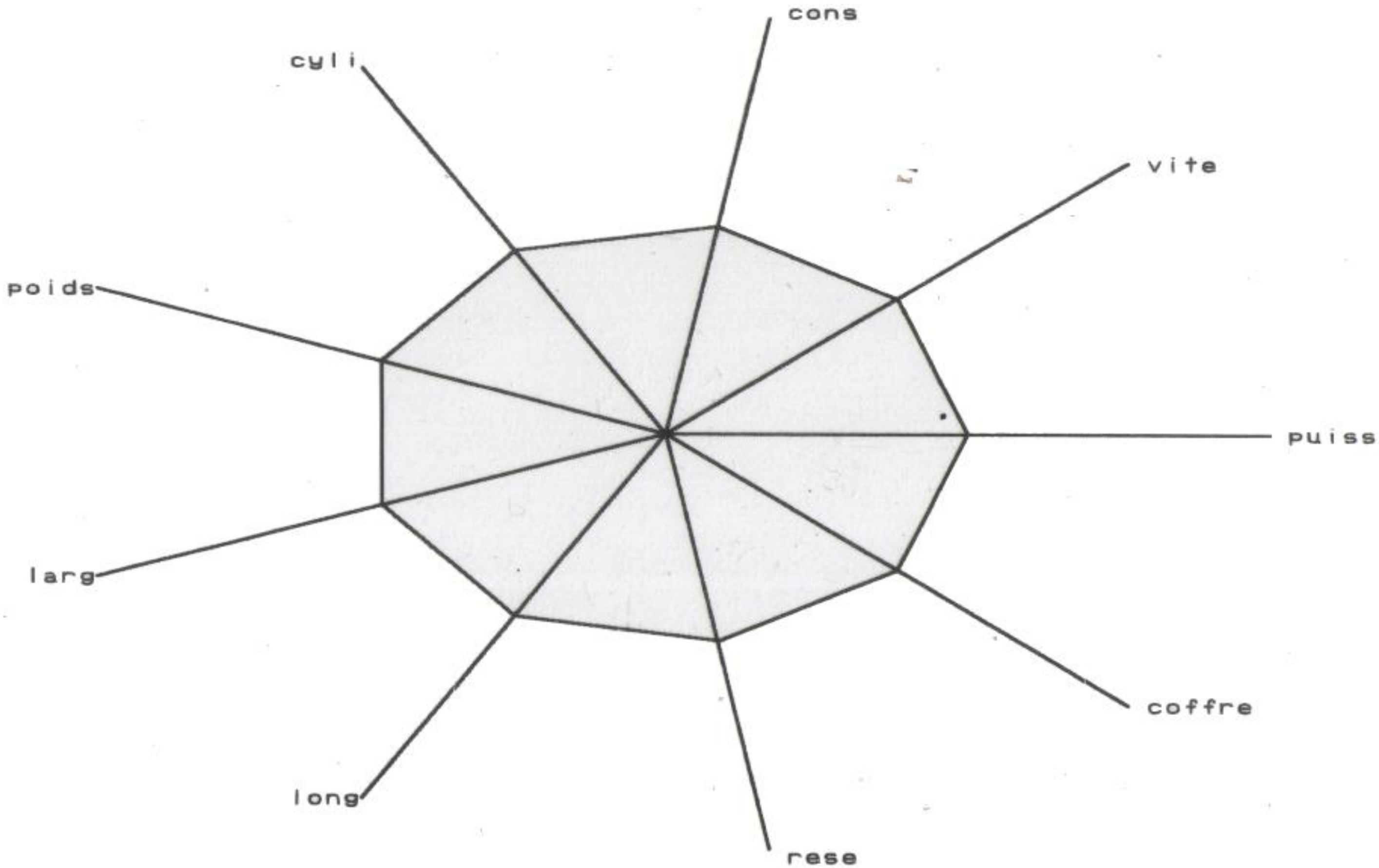
Exemple: Eaux minérales

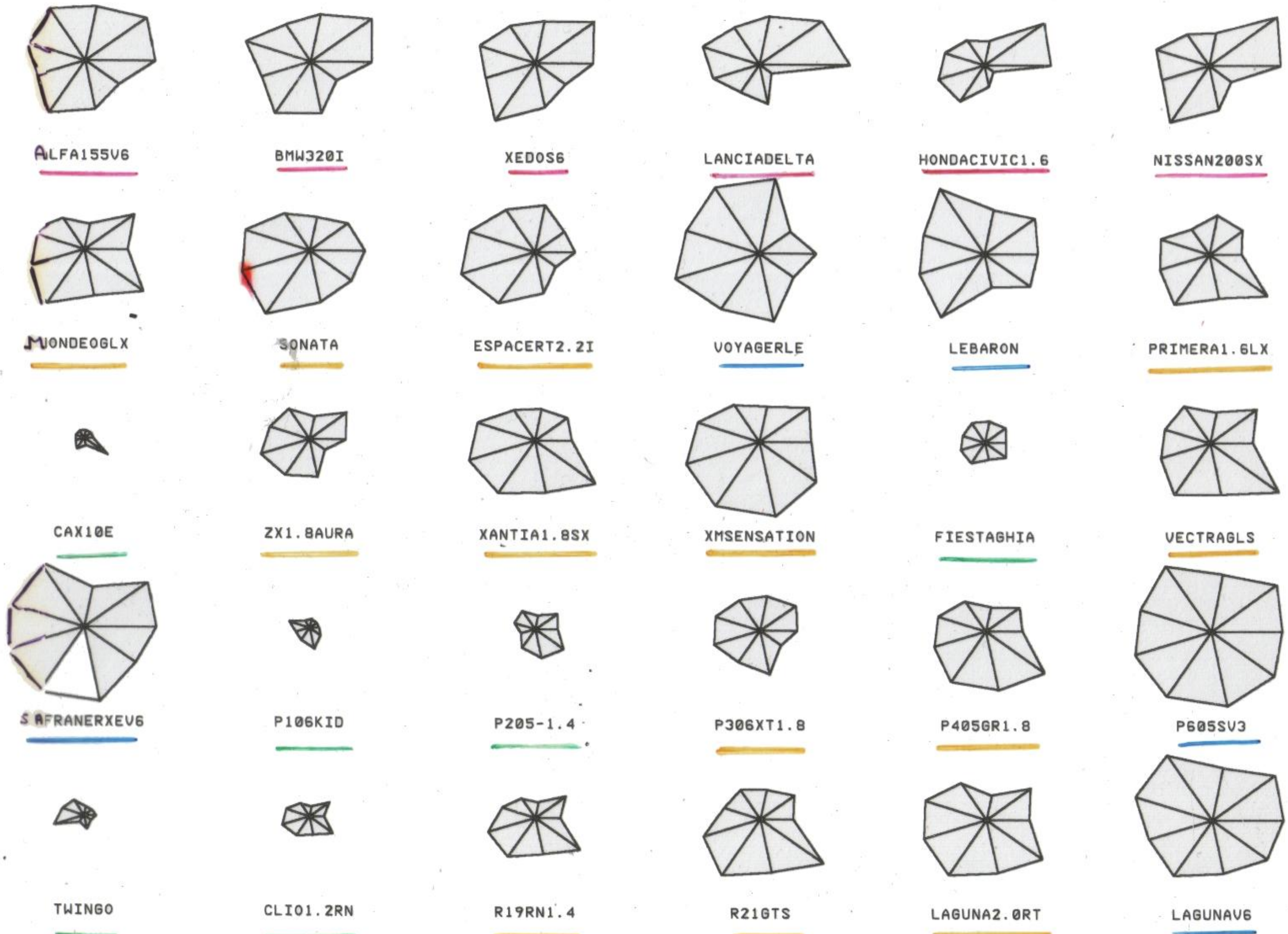


DIAGRAMMES EN ETOILES

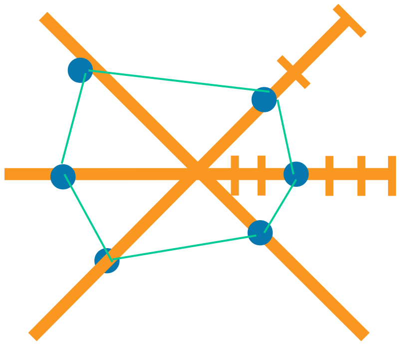


Exemple: Voitures




ALFA155V6
BMW320I
XEDOS6
LANCIADELTA
HONDACIVIC1.6
NISSAN200SX
MONDEOGLX
SONATA
ESPACERT2.2I
VOYAGERLE
LEBARON
PRIMERA1.6LX
CAX10E
ZX1.8AURA
XANTIA1.8SX
XMSENSATION
FIESTAGHIA
VECTRAGLS
SAFRANERXEV6
P106KID
P205-1.4
P306XT1.8
P405GR1.8
P605SV3
TWINGO
CLIO1.2RN
R19RN1.4
R21GTS
LAGUNA2.0RT
LAGUNAV6

2 Diagramme en rayons de soleil

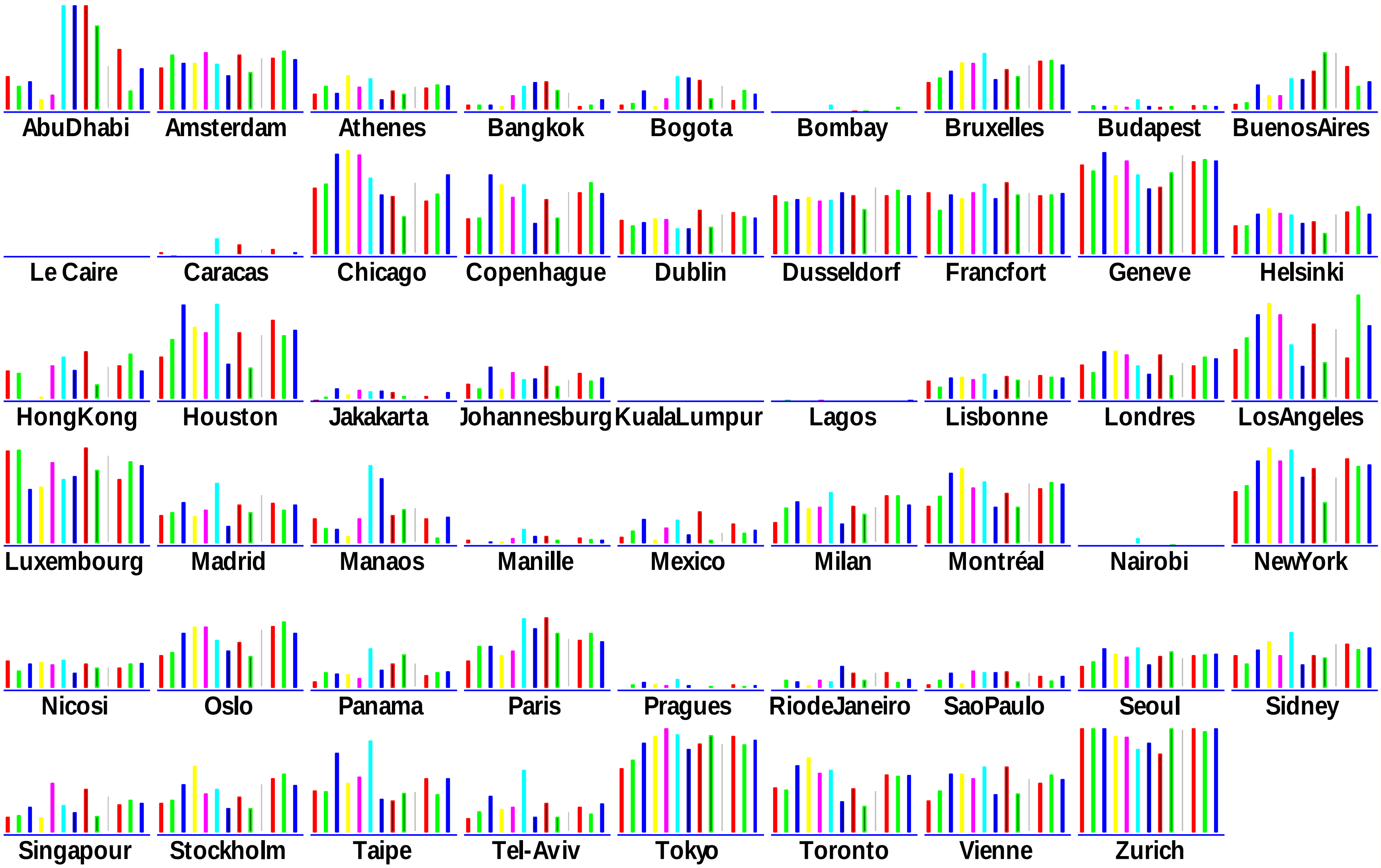


Le milieu de chaque segment représente la valeur moyenne de la variable pour le fichier considéré.
La longueur de chaque rayon est un multiple de ***2k*** σ (σ = écart-type)

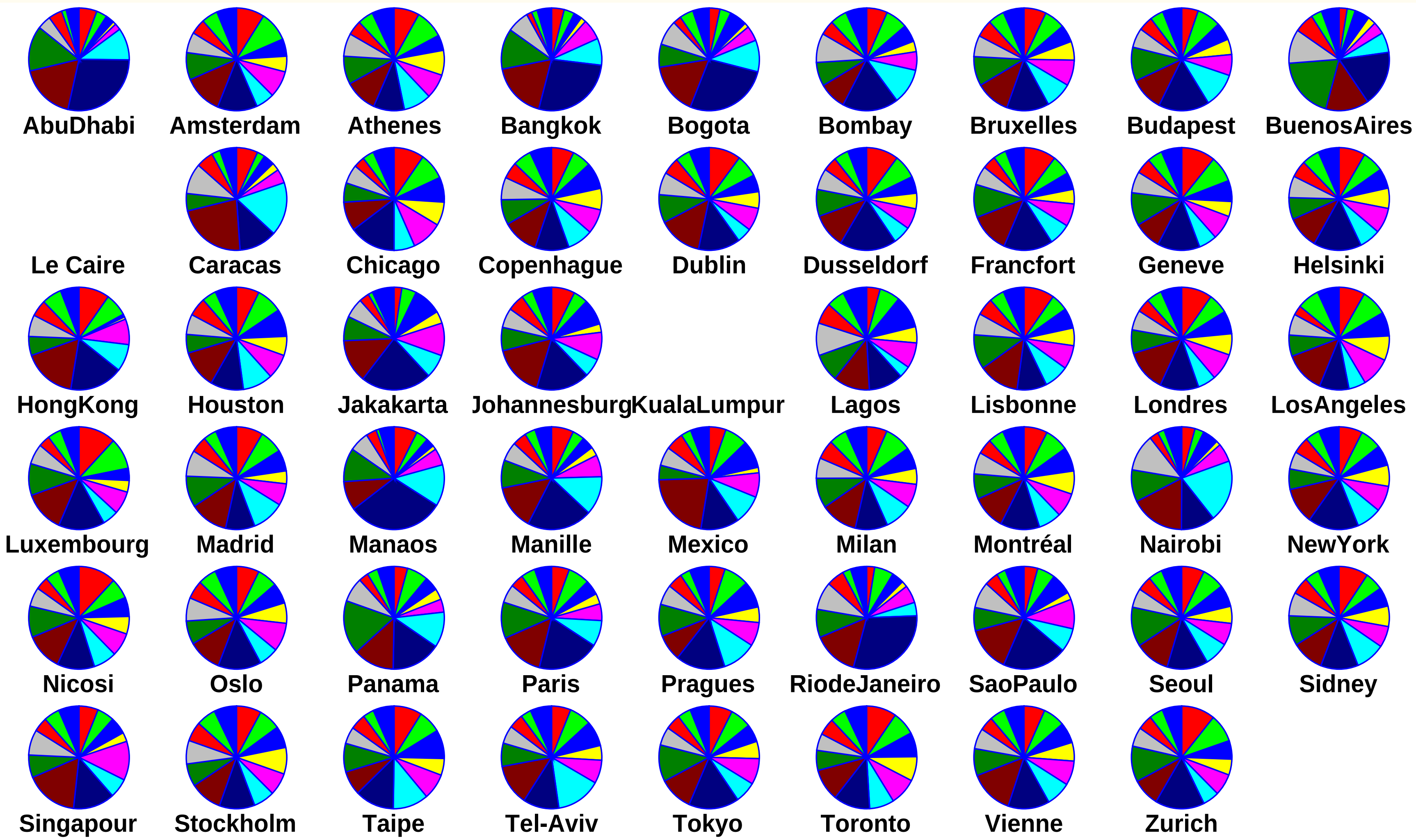
Remarque : Dans le cas du diagramme en rayon de soleil, l'indicateur de dispersion est l'écart-type σ, alors que dans un diagramme en étoiles, on utilise l'étendue.

Le diagramme en étoiles est néanmoins plus lisible.

Tracé de Figures (VILLES94.STA 41v*53c)

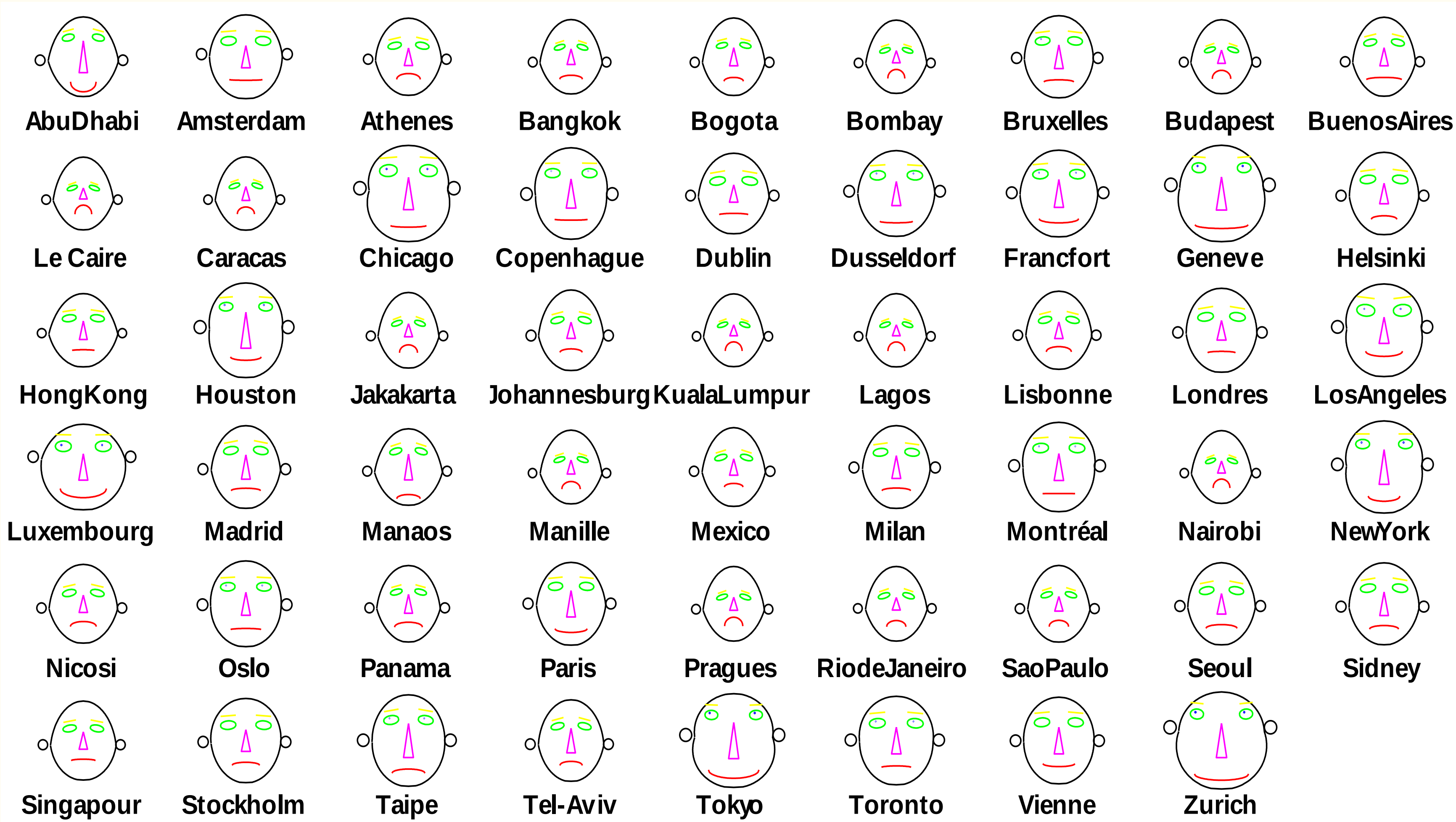


LEGENDE (gauche à droite) : INSTIT, CHAUFBUS, MECANIC, MANOEUVR, TOURNEUR, CUISINIE, CHERSERV, INGENIEU, CBANQUE, SECDIR, VENDEUSE, OUVRTEXT, REVANN

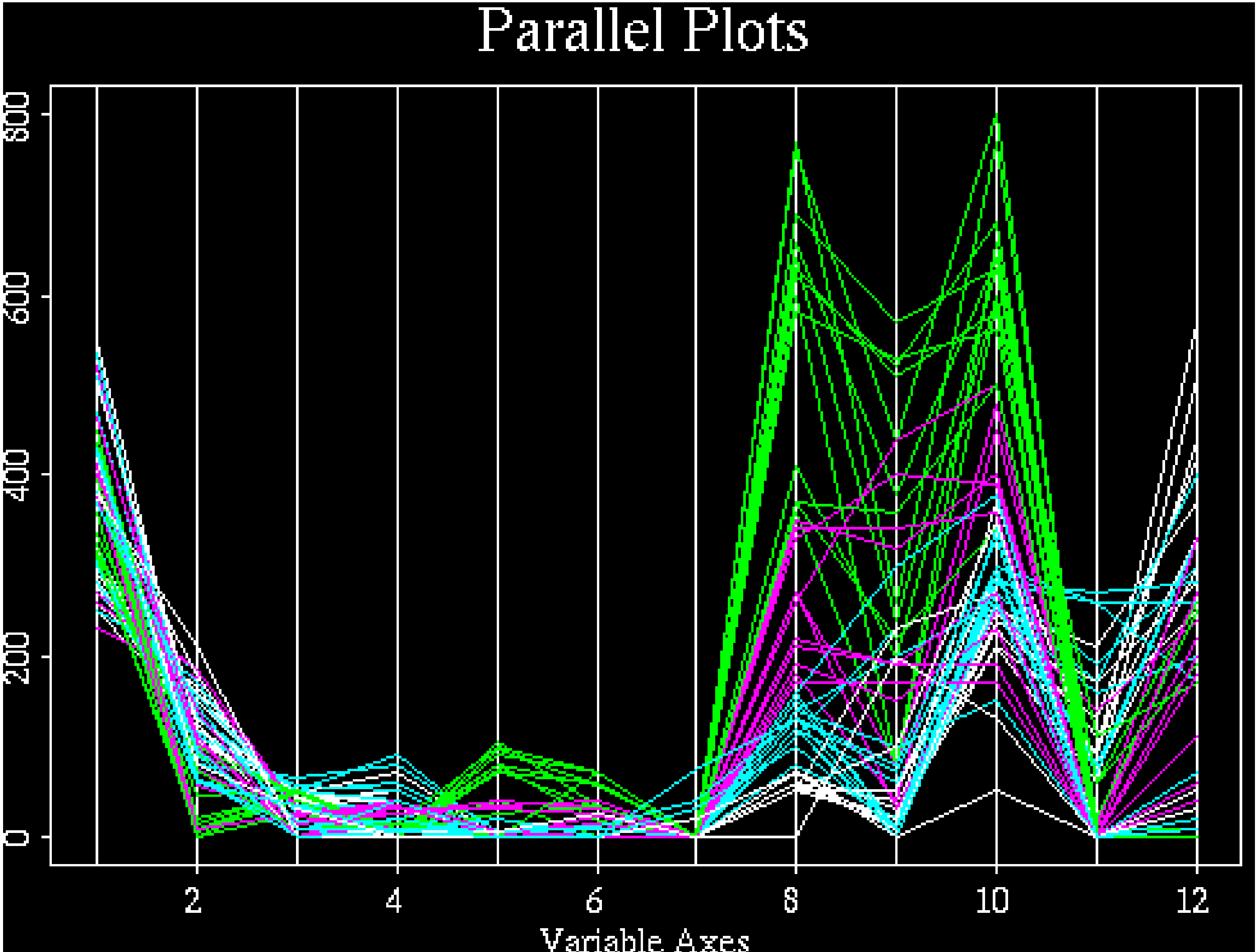


LEGENDE (dans le sens des aiguilles d'une montre): INSTIT, CHAUFBUS, MECANIC, MANOEUVR, TOURNEUR, CUISINIE, CHERSERV, INGENIEU, CBANQUE, SECDIR, VENDEUSE, OUVRTEXT, REVANN,

Tracé de Figures (VILLES94.STA 41v*53c)



LEGENDE: visage/larg. = INSTIT, oreille/niv. = CHAUFBUS, moitié du visage/haut. = MECANIC, haut du visage/exc. = MANOEUVR, bas du visage/exc. = TOURNEUR, nez/long. = CUISINIE, bouche/centr. = CHERSERV, bouche/courb. = INGENIEU, bouche/long. = CBANQUE, yeux/haut. = SECDIR, yeux/écart. = VENDEUSE, yeux/incl. = OUVRTEXT, yeux/exc. = REVANN,


Parallel Plots
800
600
400
200
0
2
4
6
8
10
12
Variable Axes