Analyse Exploratoire Multivariée

Introduction Générale

Pr Roch Giorgi

roch.giorgi@univ-amu.fr

SESSTIM, Faculté de Médecine, Aix-Marseille Université, Marseille, France
http://sesstim.univ-amu.fr/

Introduction (1)

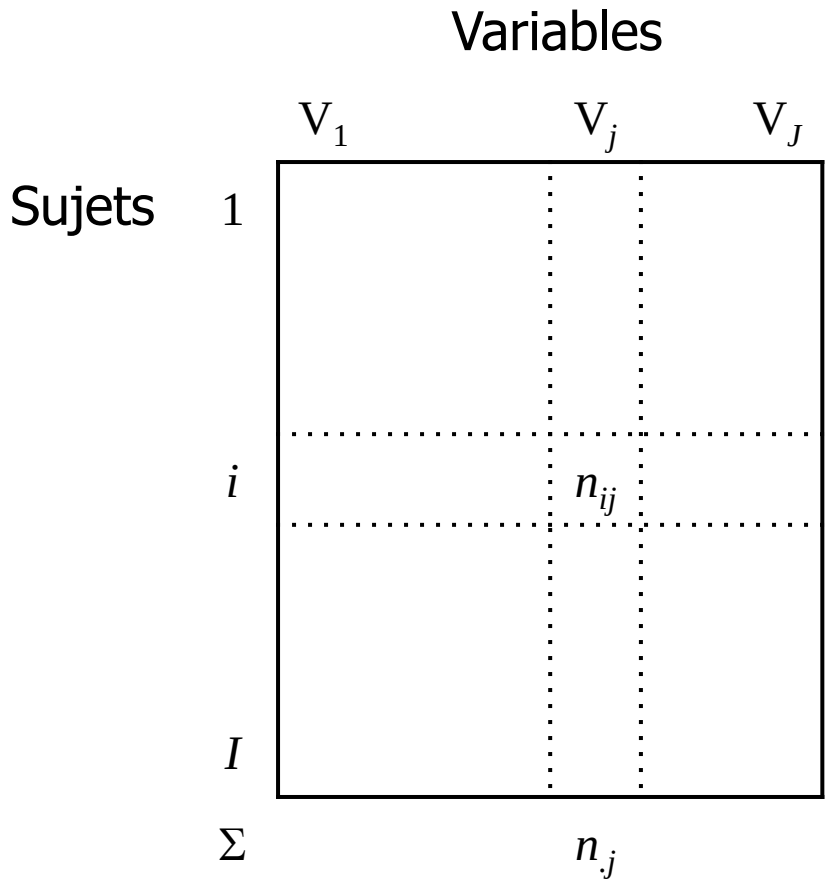
- Mesure de plusieurs variables pour chaque « unité » de l'échantillon
- Besoins d'une analyse globale
  - ✓ Intérêt d'étudier séparément les variables prises de manière isolée peu fréquent
  - ✓ Intérêt de comprendre la structure multidimensionnelle des données $\Rightarrow$ variables reliées entre elles
  - ✓ Nécessite d'analyser les variables simultanément afin de découvrir les tendances et les principales caractéristiques contenues dans ces données

Introduction (2)

- Méthodes
  - ✓ Descriptives
  - ✓ Inférentielles

- Objectifs
  - ✓ Résumer, visualiser les données
  - ✓ Pour
    - Afficher ou extraire n'importe quel signal dans les données en présence de bruit
    - Découvrir ce que les données ont à nous dire

Types de Données (1)



Types de Données (2)

Data = ex1

Variables

Sujets

ID	Sexe	Age (an)	QI	Dépres.	Santé	Poids (kg)
1	Homme	21	120	Oui	Très bon	70
2	Homme	43	uk	Non	Très bon	84
3	Femme	22	135	Non	Moyen	63
4	Homme	86	150	Non	Très mauv.	78
5	Femme	uk	92	Oui	Bon	59
6	Femme	60	130	Oui	Bon	70
7	Homme	42	uk	Non	Très bon	85

uk = unknown.

Types de Données (3)

- Types de variable

  Sexe : Qualitative

  Age (an) : Quantitative

  QI : Quantitative

  Dépress. : Qualitative

  Santé : Qualitative ordinale

  Poids (kg) : Quantitative

- Dans R

```
> str(ex1)
'data.frame':   7 obs. of  7 variables:
 $ ID      : int  1 2 3 4 5 6 7
 $ Sexe    : Factor w/ 2 levels "Femme ","Homme ": 2 2 1 2 1 1 2
 $ Age     : int  21 43 22 86 NA 60 42
 $ QI      : int  120 NA 135 150 92 130 NA
 $ Dépres.: Factor w/ 2 levels "Non ","Oui ": 2 1 1 1 2 2 1
 $ Santé   : Factor w/ 5 levels "Bon ","Moyen ",..: 4 4 2 5 1 1 3
 $ Poids   : int  70 84 63 78 59 70 85
```

Types et Gestion de Données

- Qualitative
  - ✓ Regroupement de modalités ?
- Quantitative
  - ✓ Transformation ?
- Données manquantes
  - ✓ Remplacement ?

Description des Données – Univariée (1)

- Chaque variable de manière séparée

Sexe	`table(ex1$Sexe)` `prop.table(table(ex1$Sexe))`

```
> table(ex1$Sexe)

Femme Homme
    3     4
> prop.table(table(ex1$Sexe))

    Femme     Homme
0.4285714 0.5714286
>
```

Description des Données – Univariée (2)

- Chaque variable de manière séparée

Sexe	`table(ex1$Sexe)` `prop.table(table(ex1$Sexe))`

Age	`mean(ex1$Age, na.rm=T)`
	`sd(ex1$Age, na.rm=T)^2`
	`summary(ex1$Age)`

Dépres.	`table(ex1$Dépres.)` `prop.table(table(ex1$Dépres.))`

Santé	`levels(ex1$Santé) <- c("Très mauv.", "Bon", "Moyen", "Très bon")`
	`table(ex1$Santé)`
	`prop.table(table(ex1$Santé))`

Poids	`mean(ex1$Poids)`
	`sd(ex1$Poids)^2`
	`summary(ex1$Poids)`

Description des Données – Bivariée (1)

- Variables 2 à 2

Sexe et Dépres.

```
table(ex1$Sexe, ex1$Dépres.)
prop.table(table(ex1$Sexe, ex1$Dépres.), 1)
prop.table(table(ex1$Sexe, ex1$Dépres.), 2)
```

```
        Non  Oui
Femme    1    2
Homme    3    2

          Non    Oui
Femme   0.333  0.667
Homme   0.750  0.250

          Non    Oui
Femme   0.250  0.667
Homme   0.750  0.333
```

Description des Données – Bivariée (2)

- Variables 2 à 2

Sexe et Dépres.	`table(ex1$Sexe, ex1$Dépres.)`
	`prop.table(table(ex1$Sexe, ex1$Dépres.), 1)`
	`prop.table(table(ex1$Sexe, ex1$Dépres.), 2)`

Sexe et Age	`by(ex1$Age, ex1$Sexe, FUN=mean, na.rm=T)`
	`by(ex1$Age, ex1$Sexe, FUN=sd, na.rm=T)^2`
	`by(ex1$Age, ex1$Sexe, FUN=summary)`

Age et Poids	`var(ex1[, c("Age", "Poids")], na.rm=T)`
	`cor(ex1[, c("Age", "Poids")], use="complete.obs")`
	`dist(ex1[, c("Age", "Poids")], diag=T)`

```
            Age Poids
Age     604,27  79,2
Poids    79,2   76,8
```

Description des Données – Bivariée (2)

- Variables 2 à 2

Sexe et Dépres.	`table(ex1$Sexe, ex1$Dépres.)`
	`prop.table(table(ex1$Sexe, ex1$Dépres.), 1)`
	`prop.table(table(ex1$Sexe, ex1$Dépres.), 2)`

Sexe et Age	`by(ex1$Age, ex1$Sexe, FUN=mean, na.rm=T)`
	`by(ex1$Age, ex1$Sexe, FUN=sd, na.rm=T)^2`
	`by(ex1$Age, ex1$Sexe, FUN=summary)`

Age et Poids	`var(ex1[, c("Age", "Poids")], na.rm=T)`
	`cor(ex1[, c("Age", "Poids")], use="complete.obs")`
	`dist(ex1[, c("Age", "Poids")], diag=T)`

```
            Age  Poids
Age           1   0,37
Poids      0,37      1
```

Description des Données – Bivariée (2)

- Variables 2 à 2

Sexe et Dépres.	`table(ex1$Sexe, ex1$Dépres.)`
	`prop.table(table(ex1$Sexe, ex1$Dépres.), 1)`
	`prop.table(table(ex1$Sexe, ex1$Dépres.), 2)`

Sexe et Age	`by(ex1$Age, ex1$Sexe, FUN=mean, na.rm=T)`
	`by(ex1$Age, ex1$Sexe, FUN=sd, na.rm=T)^2`
	`by(ex1$Age, ex1$Sexe, FUN=summary)`

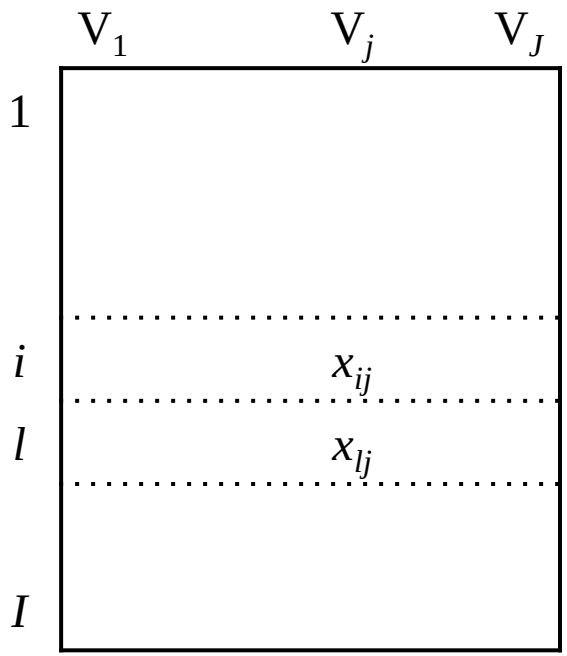
Age et Poids	`var(ex1[, c("Age", "Poids")], na.rm=T)`
	`cor(ex1[, c("Age", "Poids")], use="complete.obs")`
	`dist(ex1[, c("Age", "Poids")], diag=T)`

```
       1     2     3     4     5     6     7
1   0.00
2  26.08  0.00
3   7.07 29.70  0.00
4  65.49 43.42 65.73  0.00
5  15.56 35.36  5.66 26.87  0.00
6  39.00 22.02 38.64 27.20 15.56  0.00
7  25.81  1.41 29.73 44.55 36.77 23.43  0.00
```

Concept de Distance entre Observations (1)

- Distance au sens d'une certaine métrique

- Distance Euclidienne



$$d_{il} = \left[\sum_{j=1}^{J} \left(x_{ij} - x_{lj} \right)^2 \right]^{1/2}$$

Concept de Distance entre Observations (2)

- Distance au sens d'une certaine métrique

- Distance Euclidienne
- Distance maximale
- …
- Différentes métriques selon intérêt / méthode
- Selon intérêt, besoin de standardiser les variables si sont mesurées sur des échelles différentes

Distances Standardisées

- Variables 2 à 2

Age et Poids

```
# sd de chaque variable
std <- sapply(ex1[, c("Age", "Poids")], sd,
  na.rm=T)
# Division des valeurs d'une variable par son sd
AgePoids.std <- sweep(ex1[, c("Age", "Poids")], 2,
  std, FUN="/")
# Matrice des distances standardisées
dist(AgePoids.std, diag=T)
```

```
      1     2     3     4     5     6     7
1  0.00
2  1.66  0.00
3  0.70  2.26  0.00
4  2.76  1.85  3.00  0.00
5  1.55  3.53  0.56  2.68  0.00
6  1.59  1.56  1.70  1.32  1.55  0.00
7  1.72  0.11  2.34  1.92  3.67  1.67  0.00
```

Description des Données – Multivariée

- Données tabulées
- Matrice de statistique
  - ✓ Variances-covariances
  - ✓ Corrélations
  - ✓ Distances Euclidienne
  - ✓ …
- Représentations graphiques
  - ✓ Diagrammes de dispersion
  - ✓ Matrice des diagrammes de dispersion
  - ✓ Graphiques à 3 dimensions
  - ✓ Treillis
  - ✓ …

Exemple Conducteur

- Données de 60 zones urbaines aux Etats-Unis

Region	Zones urbaines aux EU
Rainfall	Moyenne annuelle des précipitations en pouces
Educ	Années scolaires médianes effectuées pour les plus de 25 ans en 1960
Popden	Population / mile carré dans les zones urbaines en 1960
Nonwhite	Pourcentage de non-blanc dans les zones urbaines
NOX	Potentiel de pollution relatif des oxydes d'azote
SO2	Potentiel de pollution relatif des dioxydes de souffre
Mortality	Taux de mortalités ajustés sur l'age, exprimés en décès pour 100 000

Data : `airpoll`

Source : *in Brian S. Everitt. An R and S-PLUS Companion to Multivariate Analysis. Springer 2005.*

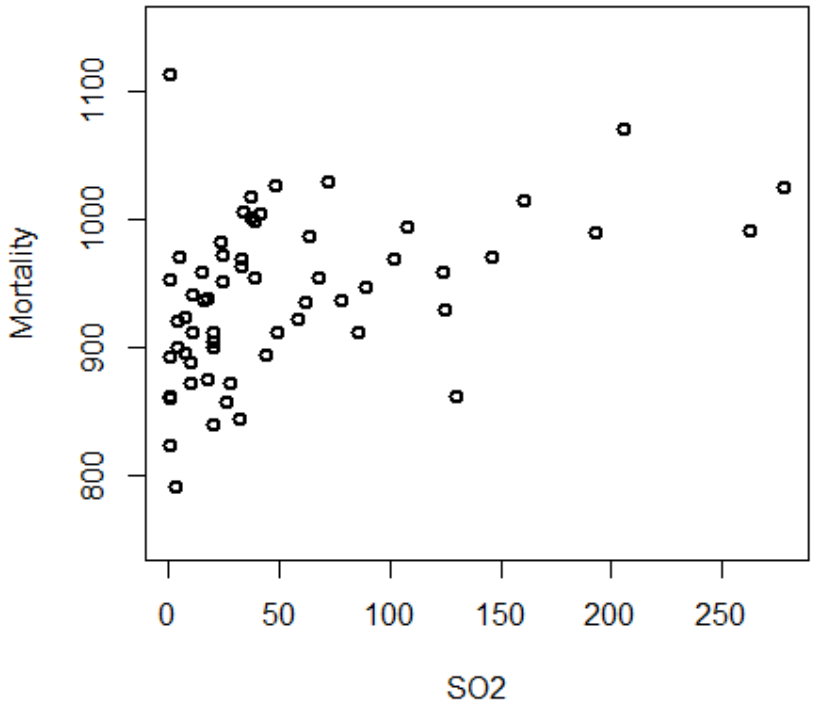
Données `airpoll`

Region	Rainfall	Educ	Popden	Nonwhite	NOX	SO2	Mortality
AkronOH	36	11,4	3243	8,8	15	59	921,9
AlbanyNY	35	11,0	4281	3,5	10	39	997,9
AllenPA	44	9,8	4260	0,8	6	33	962,4
AtlantGA	47	11,1	3125	27,1	8	24	982,3
BaltimMD	43	9,6	6441	24,4	38	206	1071,0
BirmhmAL	53	10,2	3325	38,5	32	72	1030,0
BostonMA	43	12,1	4679	3,5	32	62	934,7
BridgeCT	45	10,6	2140	5,3	4	4	899,5
BufaloNY	36	10,5	6582	8,1	12	37	1002,0
CantonOH	36	10,7	4213	6,7	7	20	912,3
ChatagTN	52	9,6	2302	22,2	8	37	1018,0
...	...	...	...	...	...	...	...

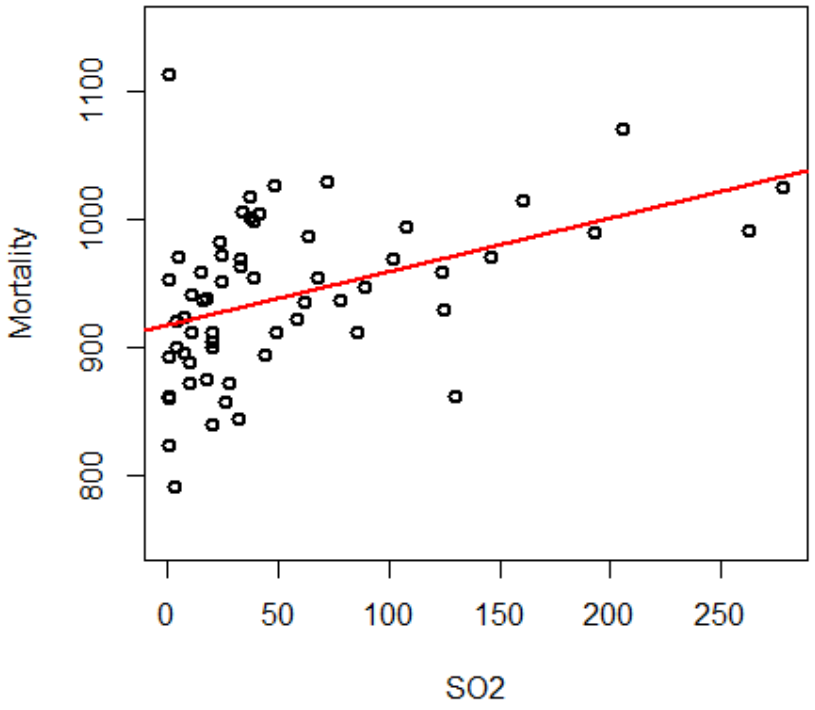
Comment SO2 est liée à la mortalité ?

Scatterplot (1)

Scatterplot simple

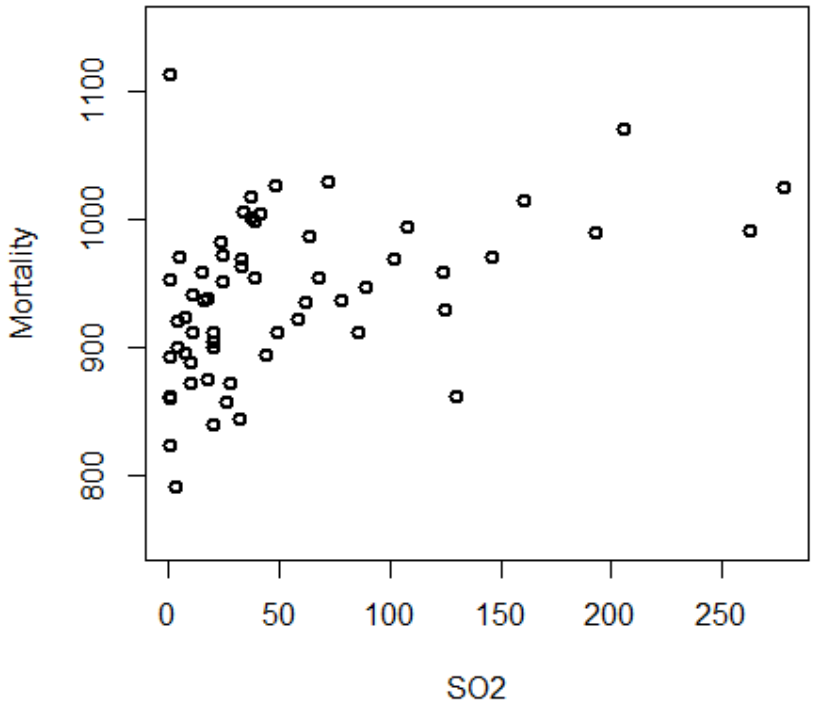


Tendance

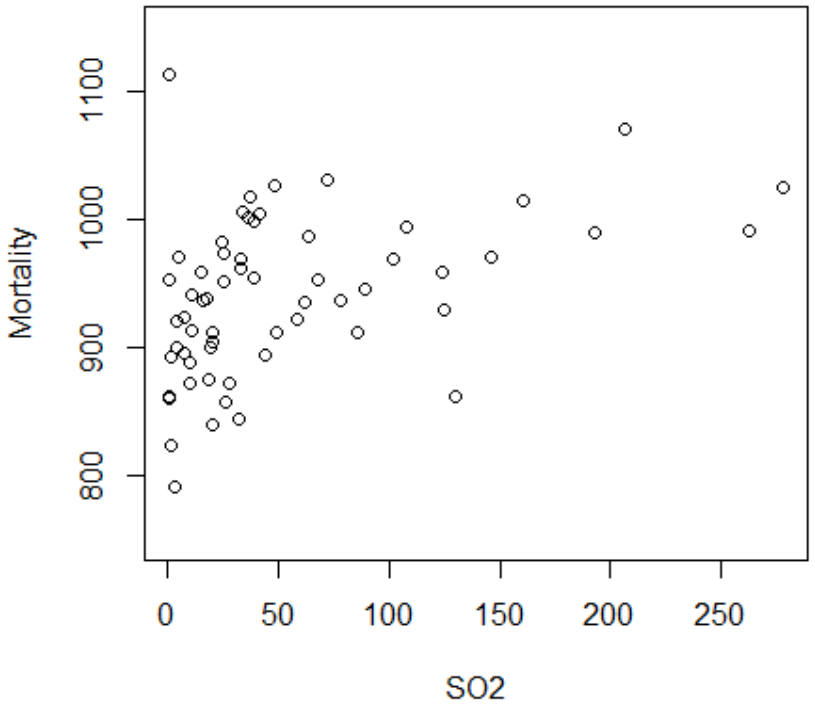


Scatterplot (2)

Scatterplot simple

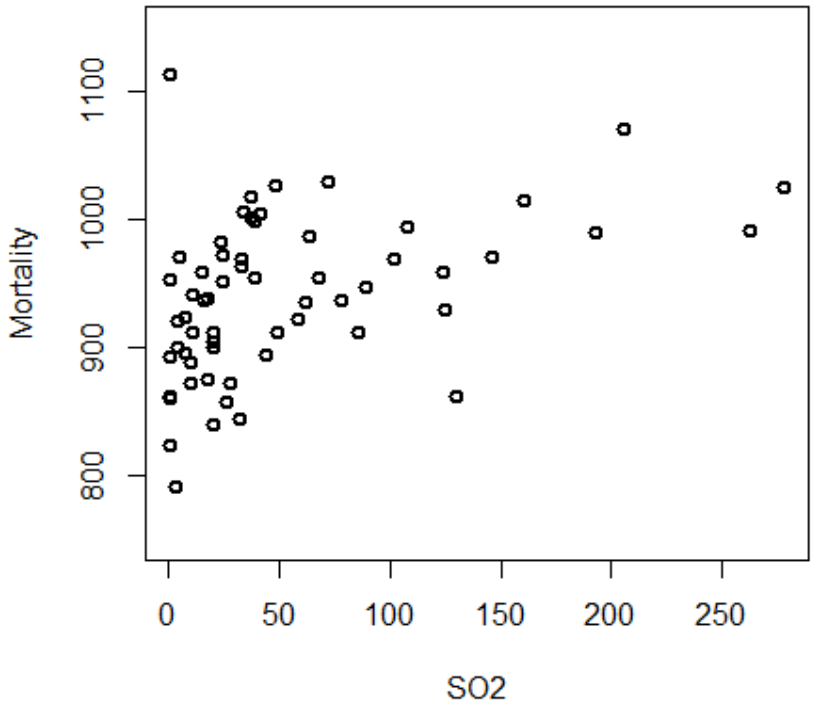


« Ecartement » des données

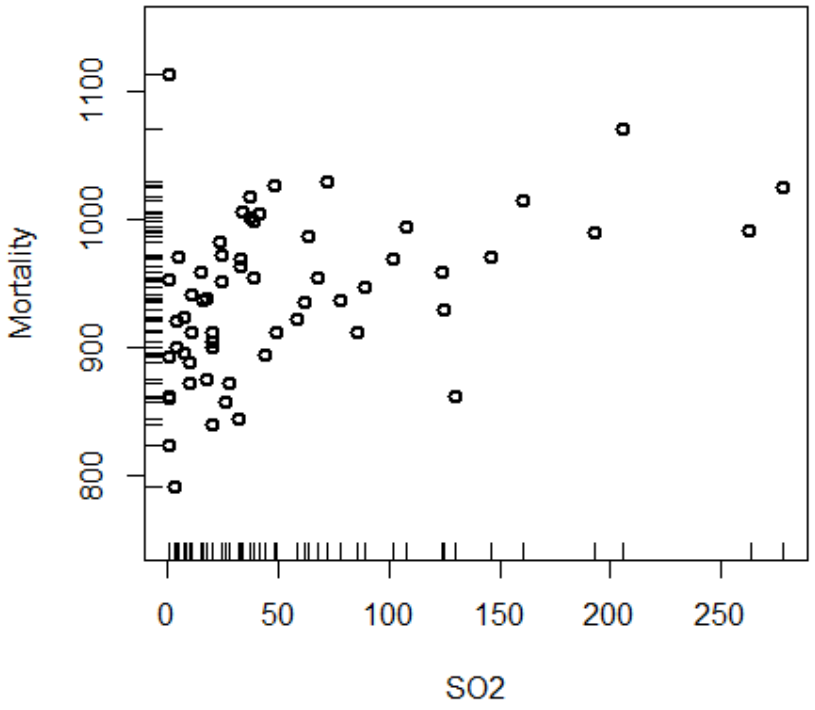


Scatterplot (3)

Scatterplot simple

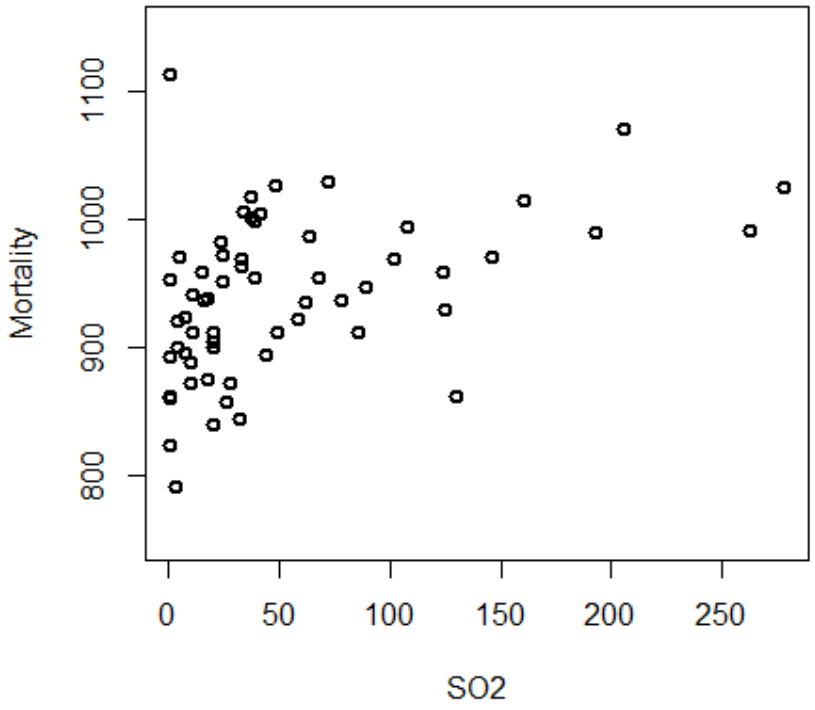


Distributions marginales

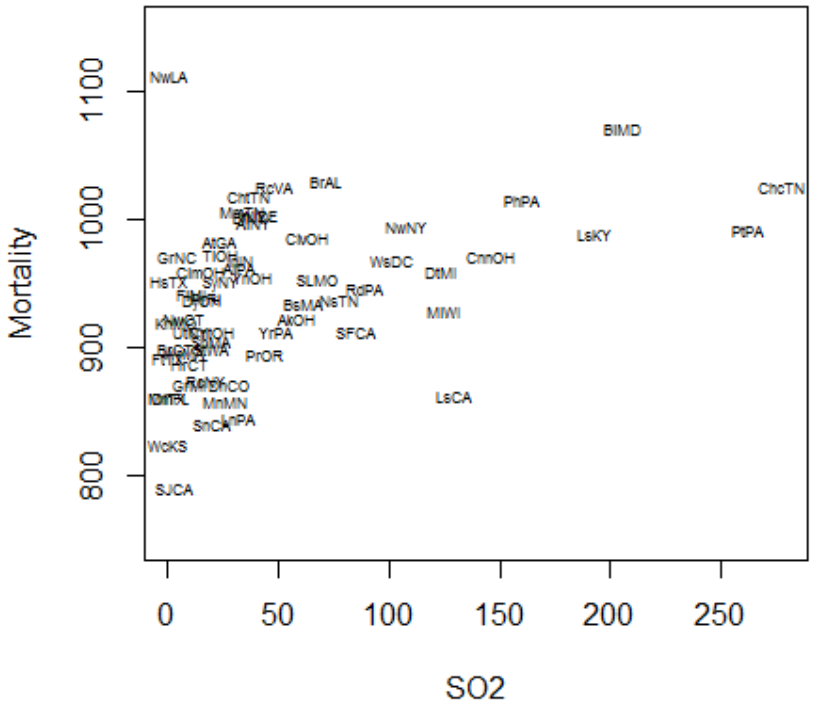


Scatterplot (4)

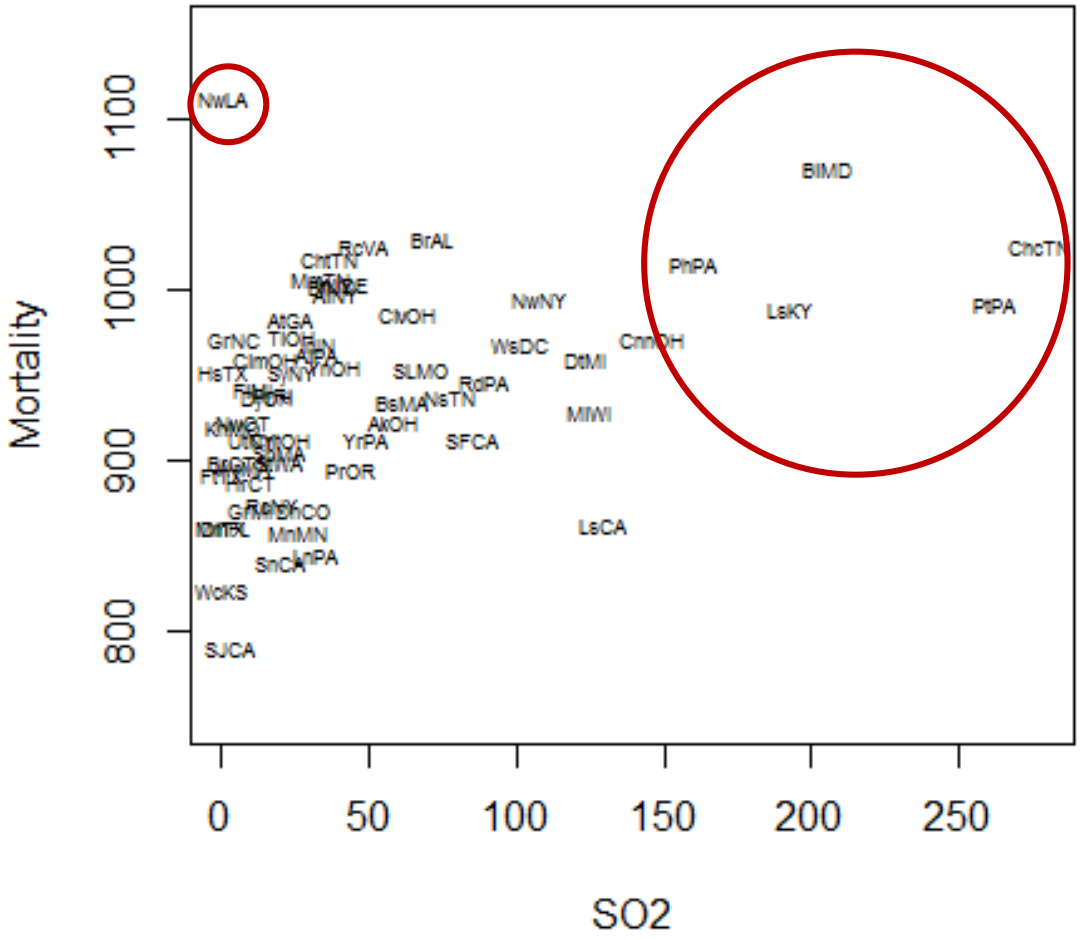
Scatterplot simple



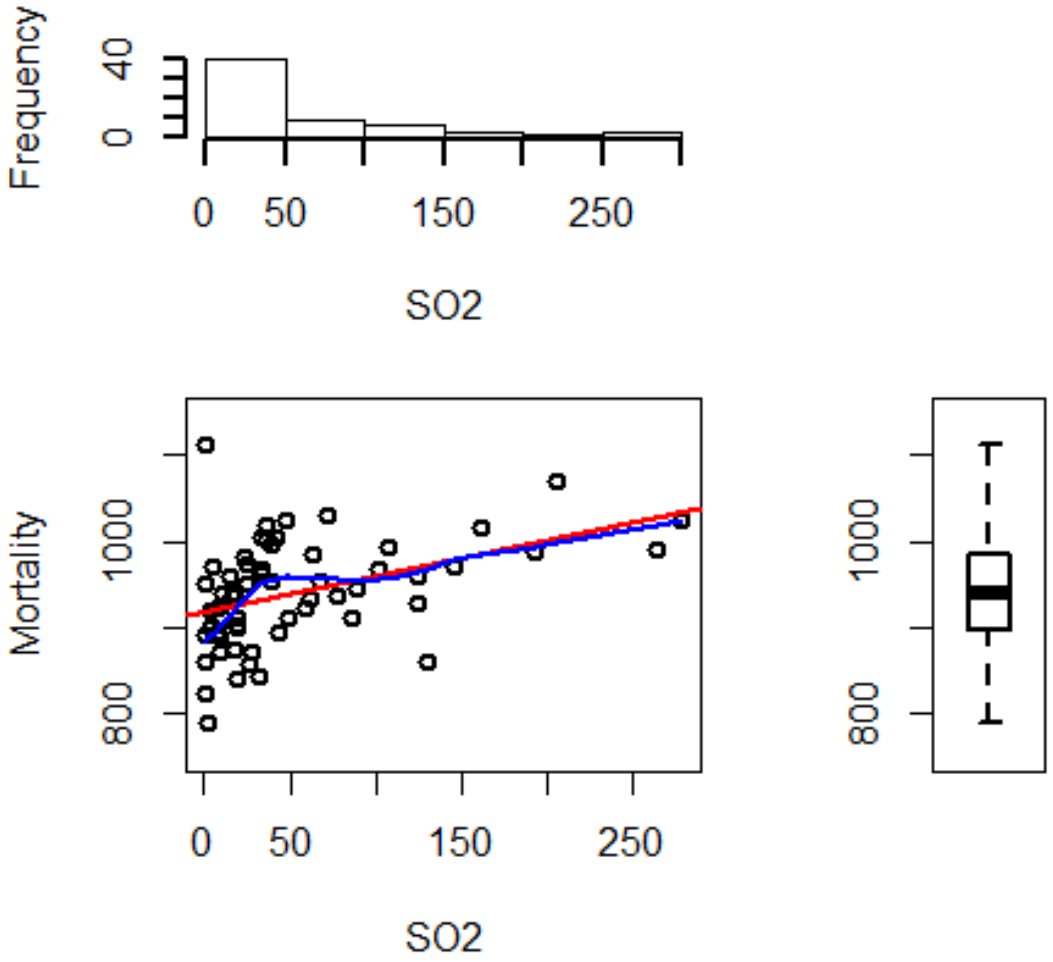
Régions concernées



Scatterplot (5)



Scatterplot (6)

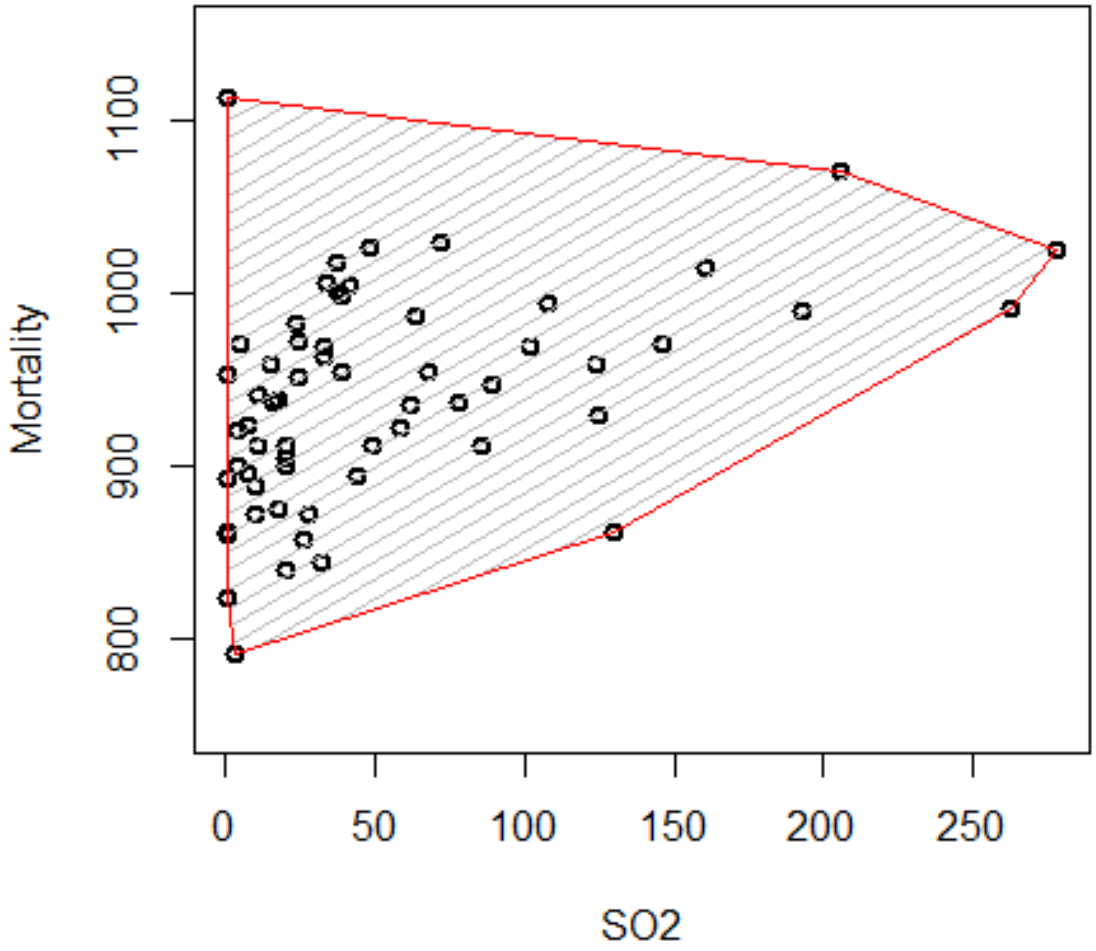


Parage Convexe (1)

- Analyse bivariée de 2 variables continues
  - ✓ Coefficient de corrélation
  - ✓ Impacté par les outliers
  - ✓ Scatterplot permet de les identifier

- En présence d'outliers
  - ✓ Elimination simple, puis estimation de la corrélation
  - ✓ Elimination par approche du parage convexe, donnant une estimation robuste
    - Détermination de l'enveloppe convexe de l'espace étudié
    - Elimination des points se trouvant sur cette enveloppe

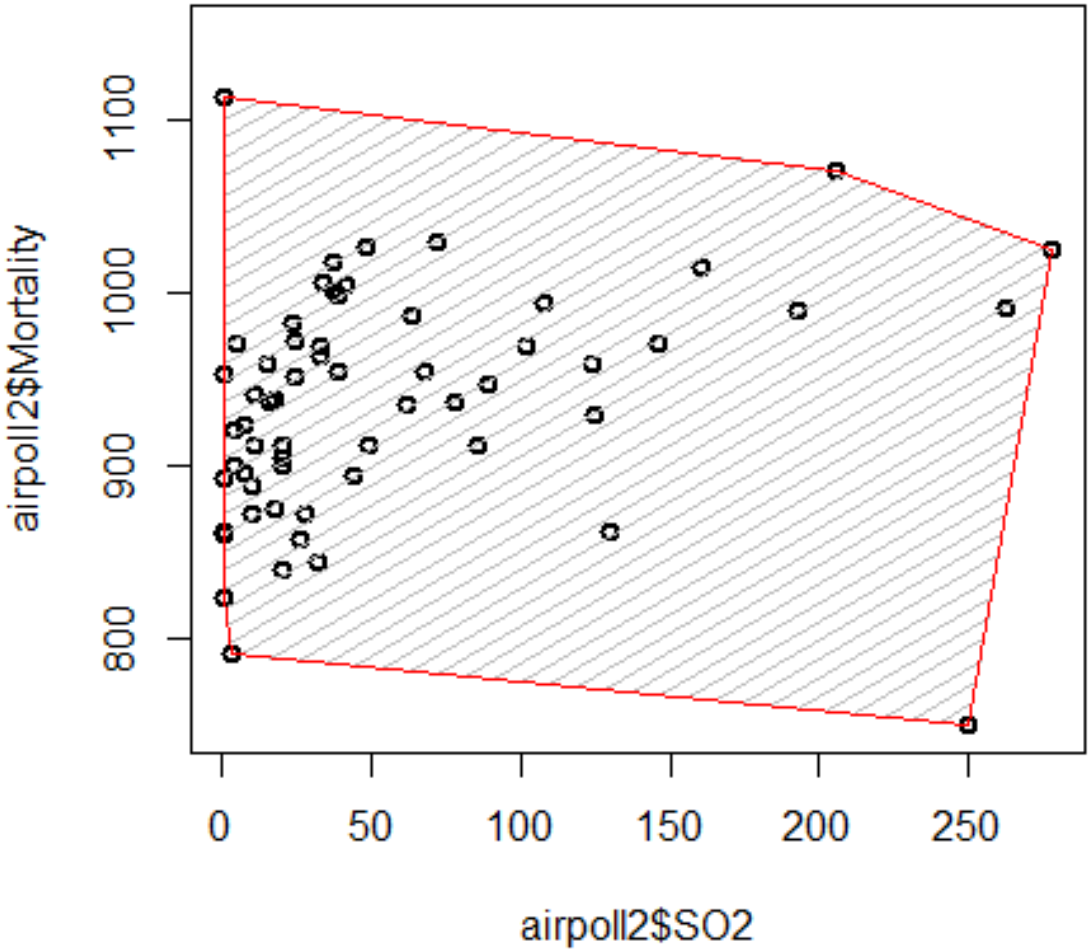
Parage Convexe (2)



Corrélation initiale = 0,427

Corrélation sans les points sur l'enveloppe = 0,439

Parage Convexe (3)



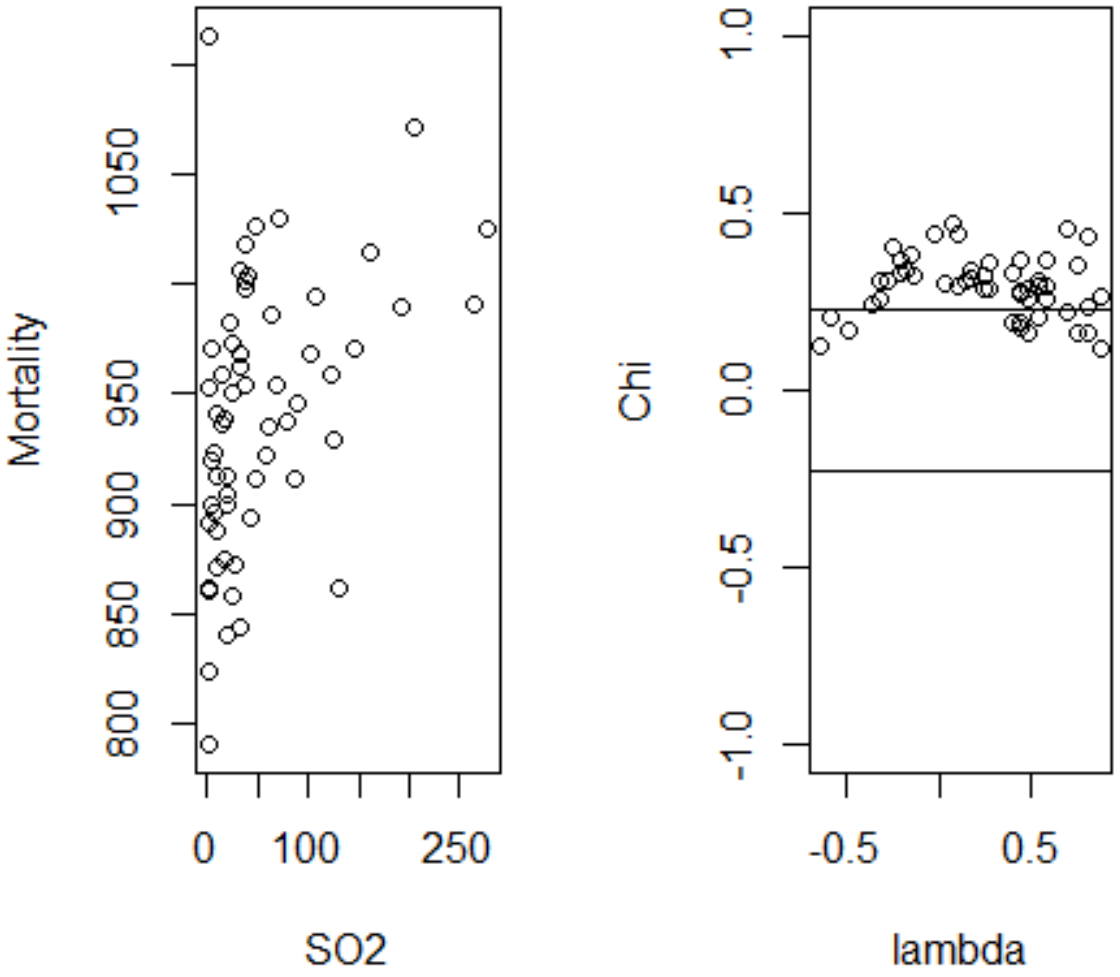
Corrélation initiale = 0,230

Corrélation sans les points sur l'enveloppe = 0,380

Indépendance (1)

- Scatterplot
  - ✓ Étude de la relation entre une paire de variables continues
  - ✓ Indépendance ?
- Chi-plot
  - ✓ Pour évaluer la dépendance
  - ✓ Scatterplot des paires (λ_i, χ_i)
    - λ_i : [-1 , 1], mesure d'une distance entre (x_i, y_i)
    - χ_i : [-1 , 1], mesure de corrélation de la dépendance entre x_i et y_i
    - 99 % des paires (λ_i, χ_i) sont comprises dans une région centrale si X et Y sont indépendantes

Indépendance (2)

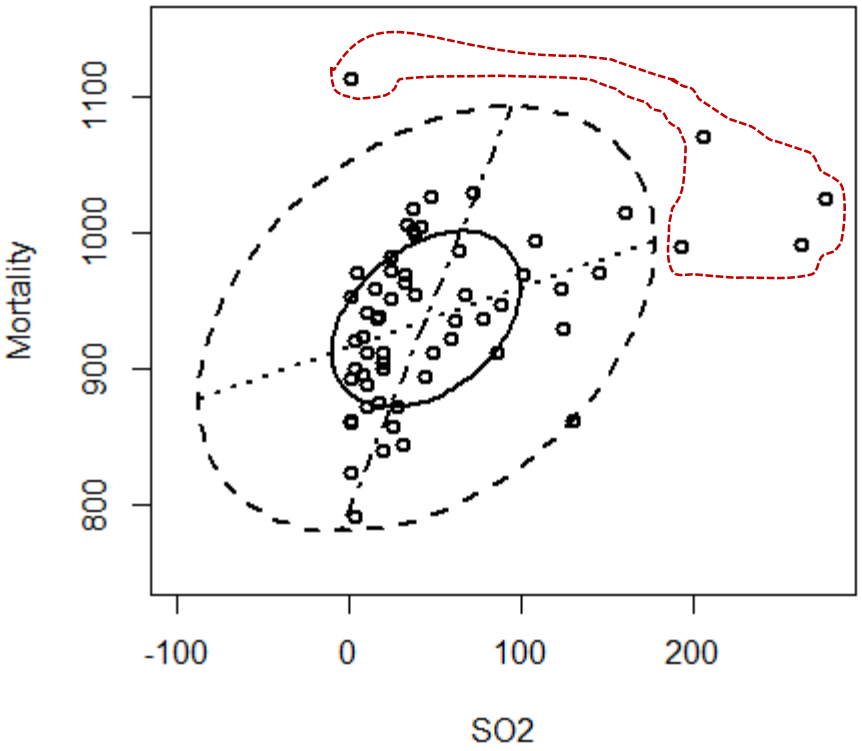


Boxplot Bivarié (1)

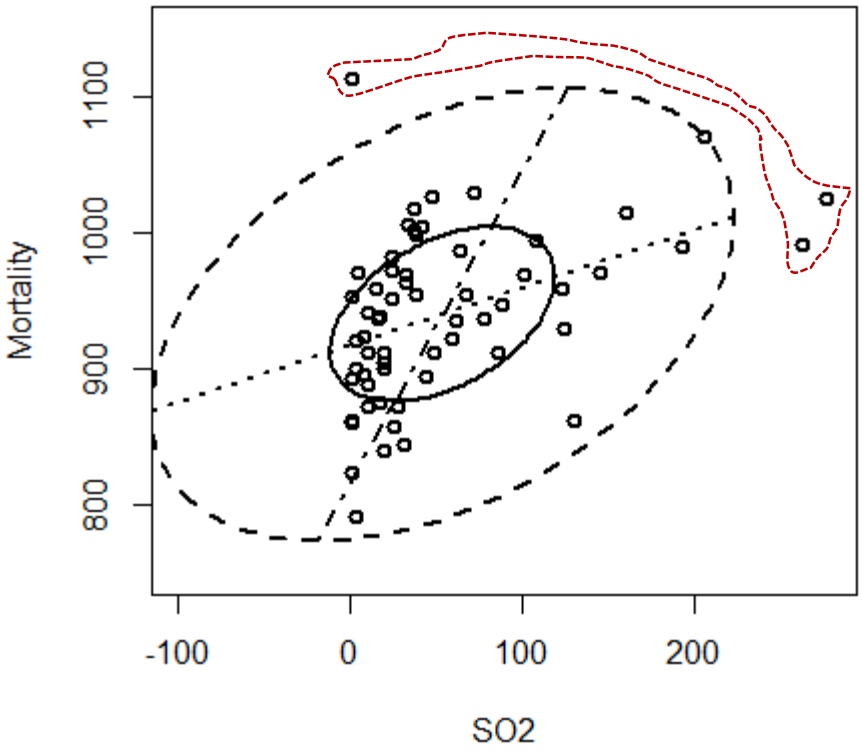
- Utile pour donner information sur les propriétés de distribution des données et pour identifier d'éventuels outliers
- Paire d'ellipses concentriques
  - ✓ L'une comprend 50 % des données
  - ✓ L'autre qui délimite les éventuels outliers
- Ajout des droites de régression de y sur x et de x sur y avec leur intersection
  - ✓ L'angle aigu entre les droites de régression sera faible pour une valeur absolue des corrélations grande
  - ✓ Et inversement

Boxplot Bivarié (2)

Avec estimateur robuste de moyenne, variance, corrélation



Avec estimateur classique de moyenne, variance, corrélation

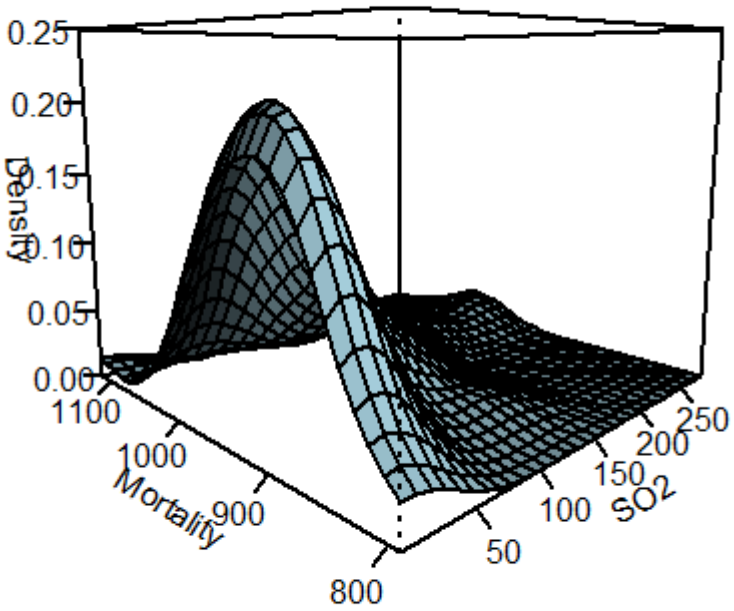


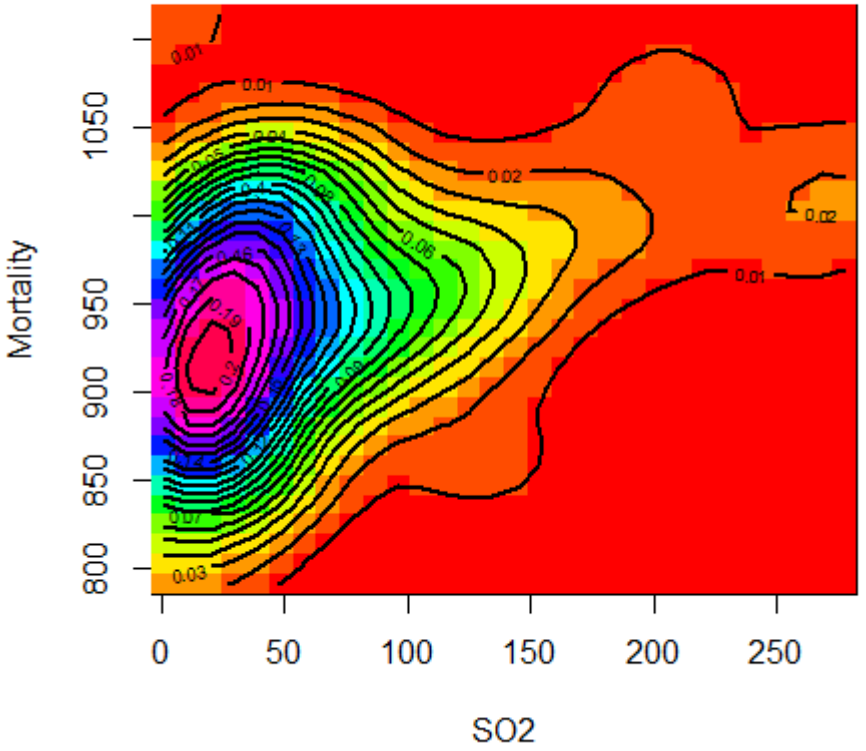
Densité Bivarié (1)

- Ajout d'informations sur la densité peut être plus utile que scatterplot pour identifier outliers, cluster,…

- Approches non paramétriques préférées à celles paramétriques
  - ✓ Absence d'hypothèse sur les distributions
  - ✓ Estimation de la densité bivariée par la densité de Kernel

Densité Bivarié (2)



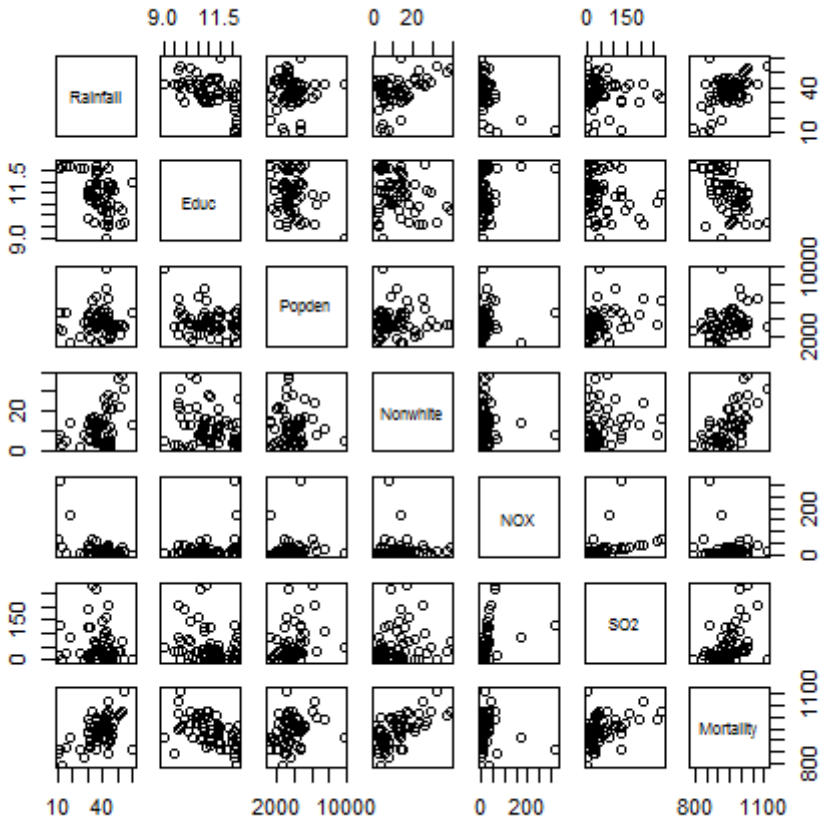


Matrice des Diagrammes de Dispersion (1)

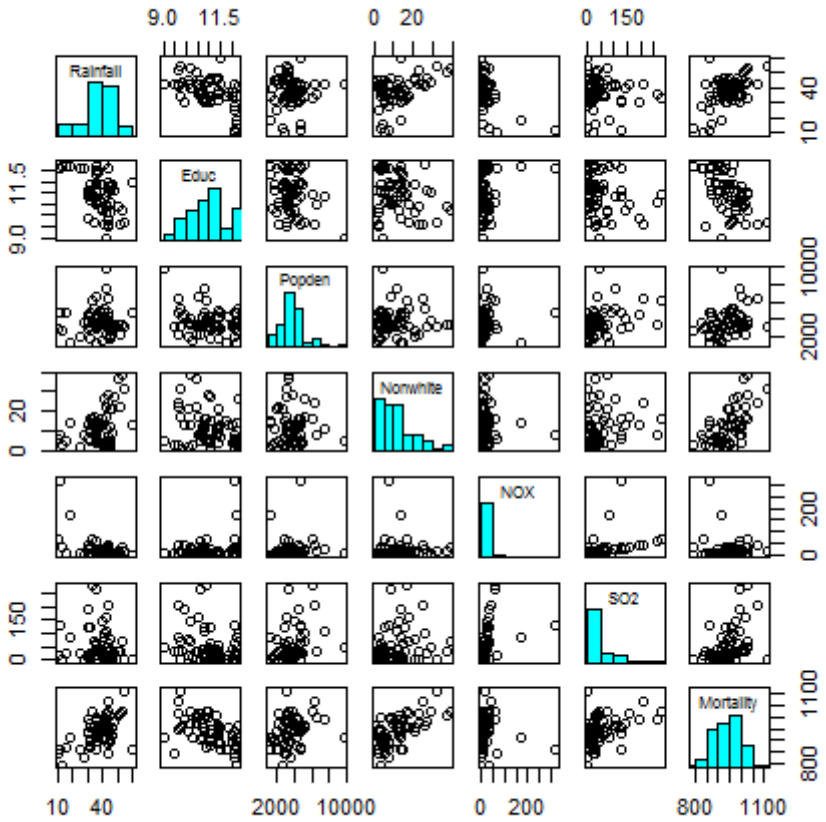
- Matrice carrée, grille symétrique de scatterplots bivariés

- Matrice de q lignes et colonnes, chacune correspondant à une variable

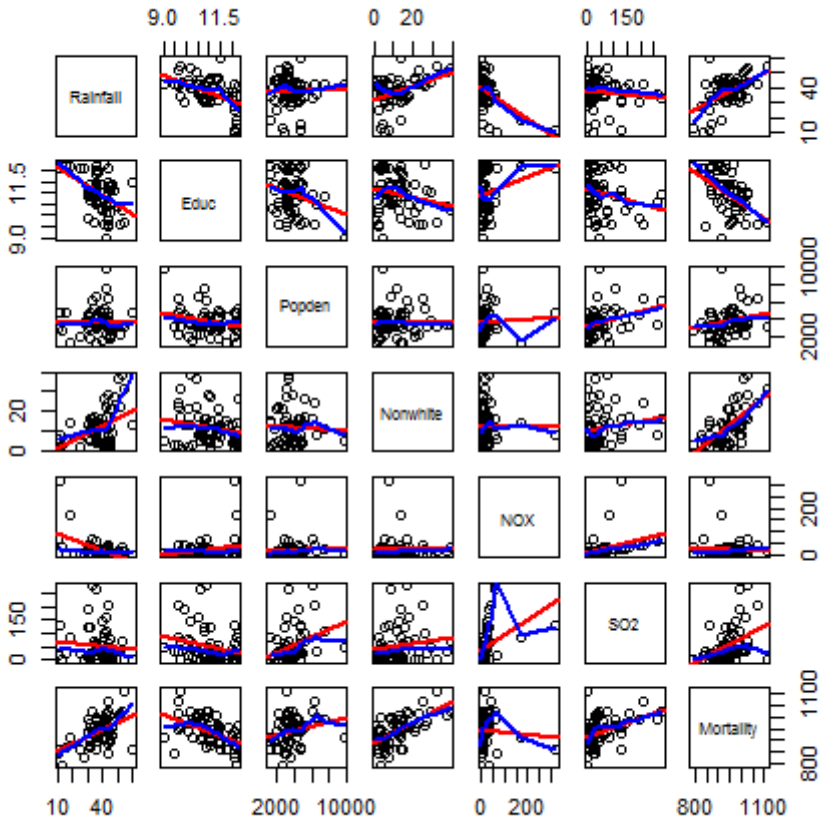
Matrice des Diagrammes de Dispersion (2)



Matrice des Diagrammes de Dispersion (3)



Matrice des Diagrammes de Dispersion (4)



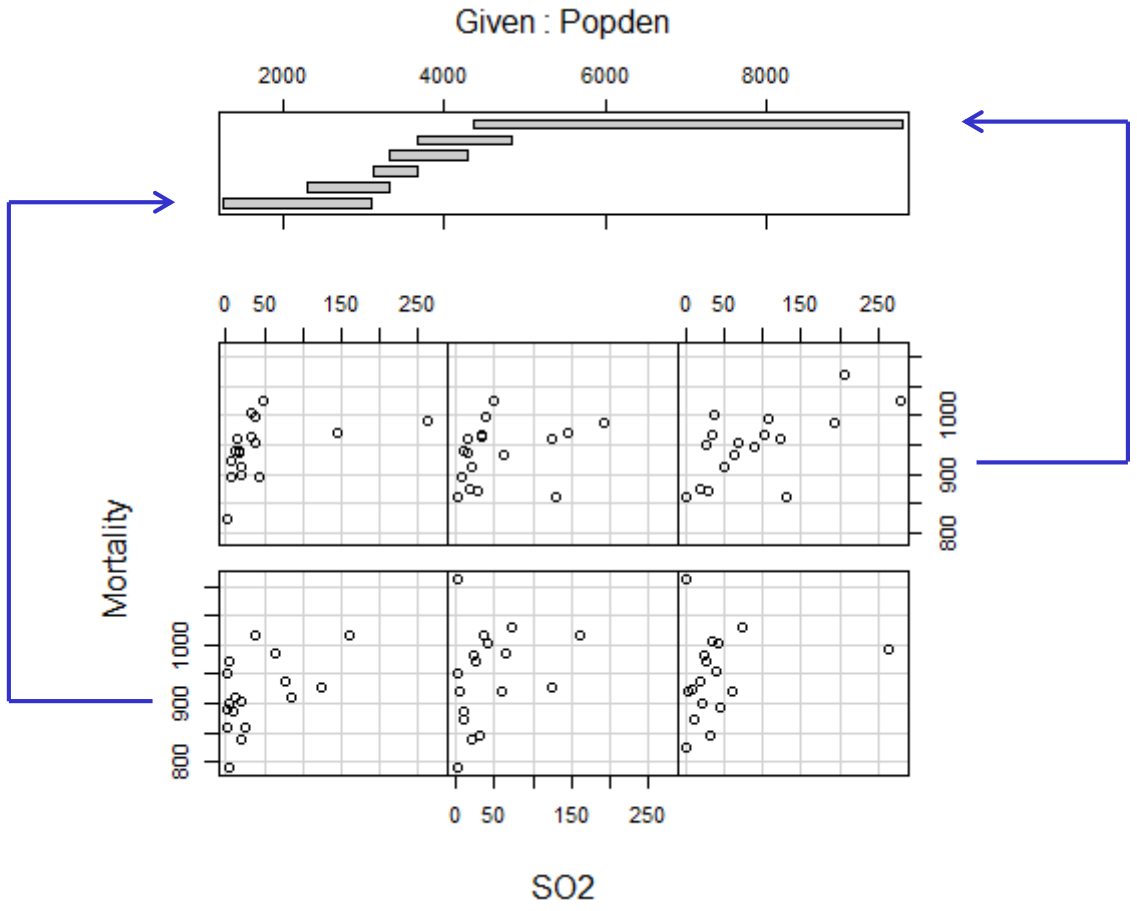
Graphiques Conditionnels (1)

- Exploration de données bivariées conditionnellement aux valeurs d'une ou plusieurs autres variables

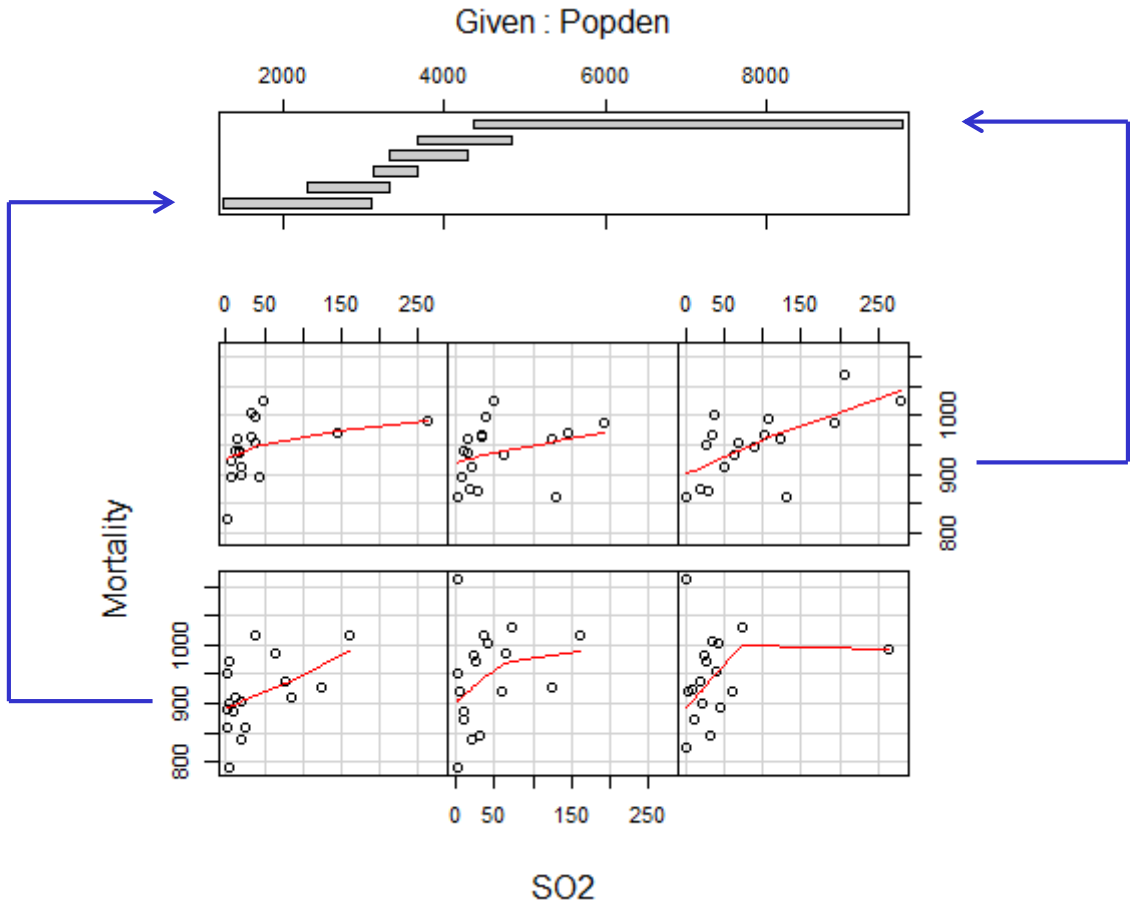
- Peut mettre en évidence interactions entre variables

- Exemple
  - ✓ SO2 et Mortality ⇒ variables dépendantes
  - ✓ En fonction de Popden ⇒ variable donnée

Graphiques Conditionnels (2)



Graphiques Conditionnels (3)



Après la Description, l'Analyse

- Plusieurs variables quantitatives
  - ✓ Analyse en composante principale
  - ✓ Analyse factorielle
- Deux variables qualitatives
  - ✓ Analyse factorielle des correspondances simples
- Plusieurs variables qualitatives
  - ✓ Analyse factorielle des correspondances multiples
- Mais également des méthodes de classification supervisées ou non supervisées
- …

Sources

- Brian S. Everitt. An R and S-PLUS Companion to Multivariate Analysis. Springer 2005.