Graphisme en statistique : quelques bases …

Motivation et buts du cours

Motivation

Le graphique est un élément clef de la communication des résultats d'une analyse statistique. La plupart des observations que l'on peut faire sur des séries de données peuvent en général être illustrées sur la base de graphiques et les "clients" de la statistique sont de plus en plus demandeurs de ce type de résultat.

Buts du cours

- Présenter les différents types de graphiques statistique de base les plus courament utilisés.
- Montrer comment les utiliser, les combiner et les mettre en oeuvre dans différents software
- Quelques exemples de graphiques dynamiques (treillis graphs, XYZ…)
- Donner quelques recommandations dans la réalisation de graphiques statistiques.
- Donner quelques exemples d'horreurs produites par les logiciels…

Graphiques de base

Il existe quelques graphiques génériques qui combinés astucieusement forment des outils puissants de présentation de données et de résultats d'analyses statistique.

- Graphique temporel
- Graphe X-Y
- Diagramme en points, histogramme, box plot
- Diagramme en barre à 2 ou 3 dimensions, diagramme en tarte
- Surfaces de réponse, courbes de niveaux
- QQ plot
- Graphe d'autocorrélation

Les données

environ.txt

Etude de la qualité de l'eau d'une rivière canadienne.

Trois variables sont mesurées 1 fois par semaine durant 3 ans.

Les variables :

- Semaine : no de la semaine
- temp : la température de l'eau,
- DO : quantité d'oxygène dissoud dans l'eau
- secchi : clarté de l'eau
- saison : saison de la mesure

Memoire.txt

Comparaison de cinq méthodes de mémorisation d'une liste de mots. 50 sujets sont regroupés en 5 groupes. Chacun est confronté à la liste de mots avec un méthode donnée.

Variables :

- methode : type de méthode de mémorisation utilisée (…)
- mots : nombre de mots retenus.

Les données (suite)

stress.txt

Enquête sur le lien entre stress, cigarette et mode de transport dans une entreprise. 144 personnes interrogées.

Variables :

- stress : niveau de stress (peu, moyen, beaucoup),
- trajet : mode de transport domicile-travail (piedvelo, transpcom, voiture),
- fumeur : type de fumeur (non, peu, beaucoup)
- cigarettes : nombre moyen de cigarettes fumées par jour

pubsplus.txt

Etude de la relation entre la publicité faite pour une chaine de magasin et le chiffre d'affaire dans les 3 région de Belgique (78 magasins).

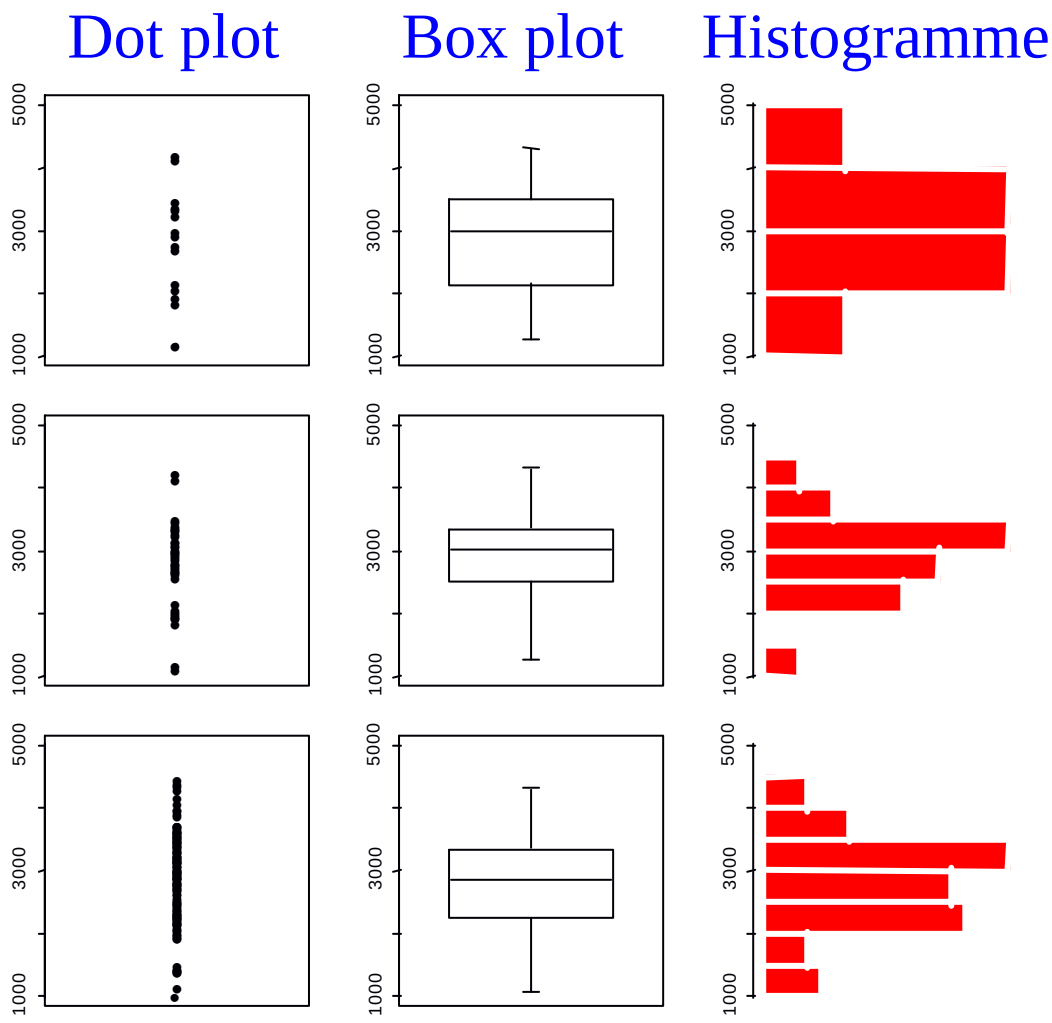
Variables

- Region : region du pays (1=bxl, 2=wal, 3=fl)
- Regiont : region sous forme texte
- Pub : montant (en Euro) utilisé pour la campagne publicitaire
- Ventes : ventes (en Euros) durant le mois après la campagne publicitaire

Représentation de la distribution d'UNE variable quantitative

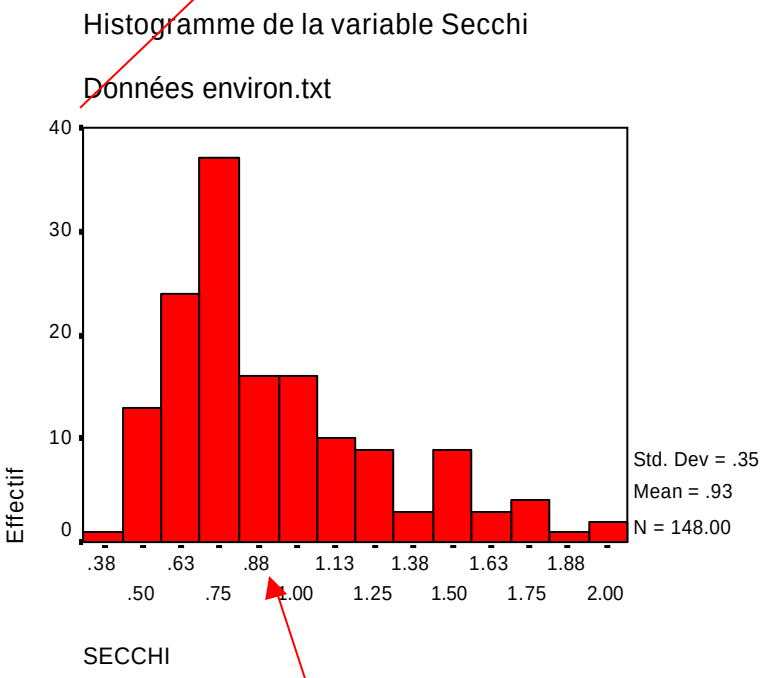
Une variable quantitative

Le choix dépend du nombre de données



Histogramme : définition et recommandations

Nb d'observations par classe ou fréquence relative



Classes

Permet de visualiser la forme de la distribution statistique d'une variable quantitative.

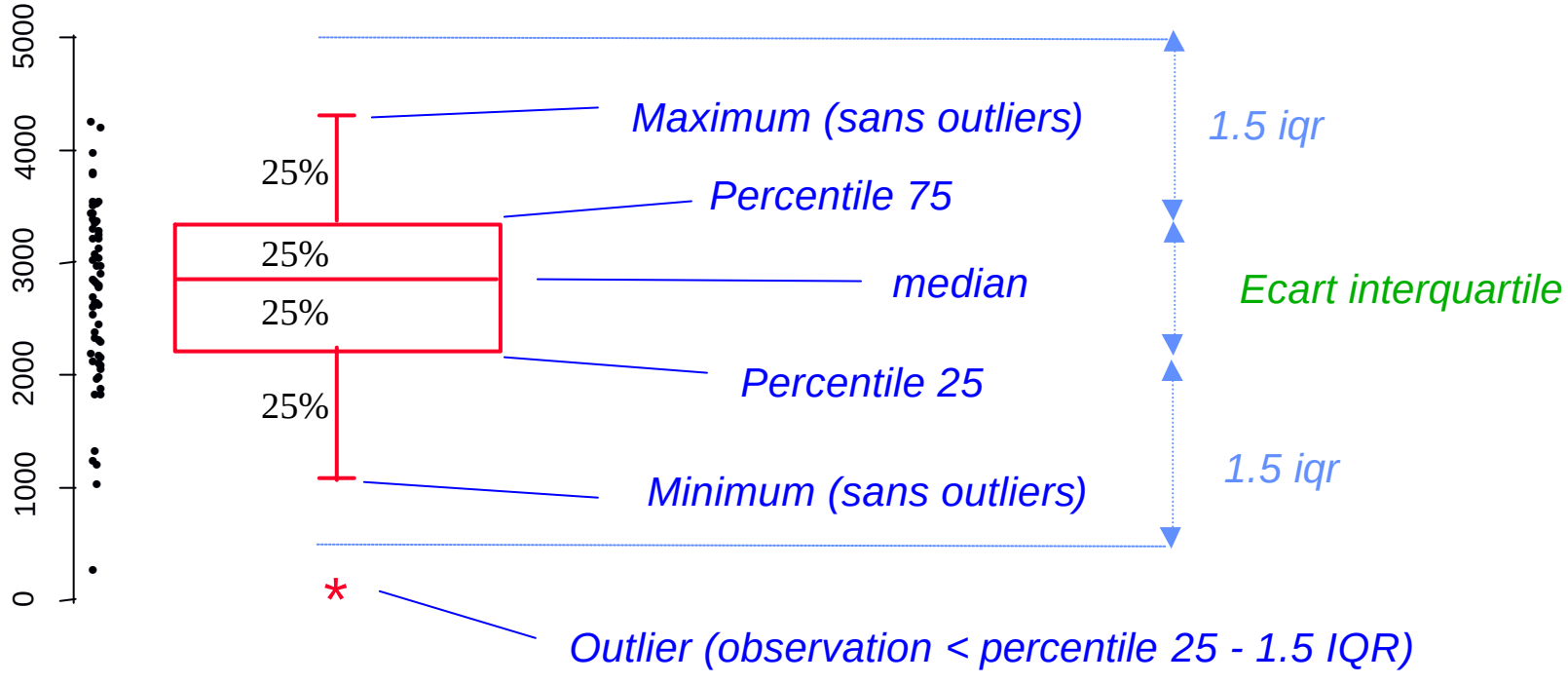
Sa forme peut varier très fort quand on modifie les limites et le nombre de classes.

Prendre un nombre de classes proche de la racine carrée de n

Ne pas utiliser quand n<50

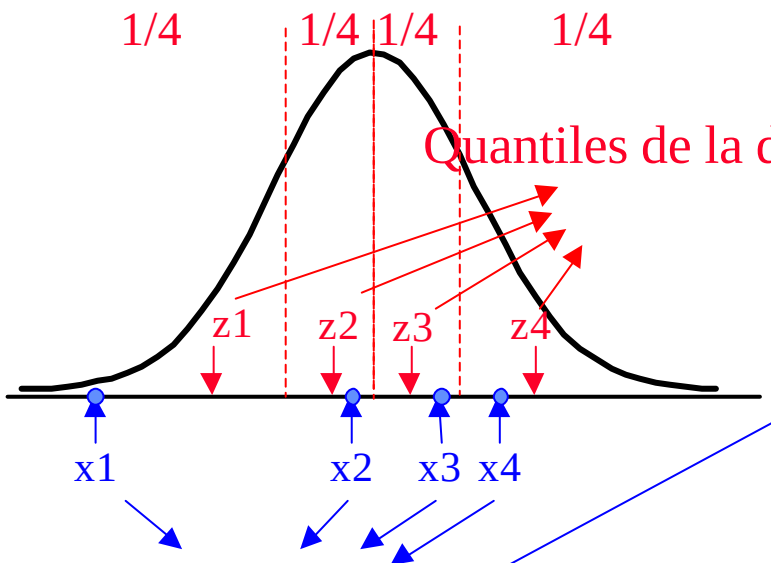
Box plot : définition et construction

Le box plot donne une idée de la distribution d'une variable même quand le nombre de données est faible. Il permet de repérer des valeurs aberrantes.
Méthode : ordonner les données et les couper en 4 groupes de 25% d'observations.

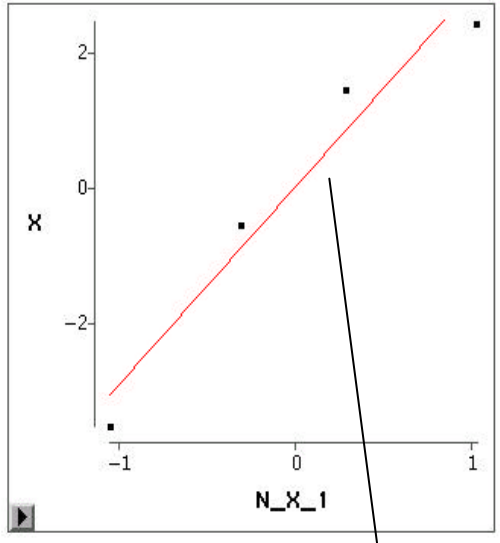


QQ plot pour vérifier la normalité de données

Un QQ plot consiste à comparer les données observées aux données que l'on devrait avoir si elles suivaient « parfaitement » une distribution normale. Les valeurs observées et « idéales » sont comparées sur un graphe X-Y qui doit montrer une tendance linéaire en cas de normalité.

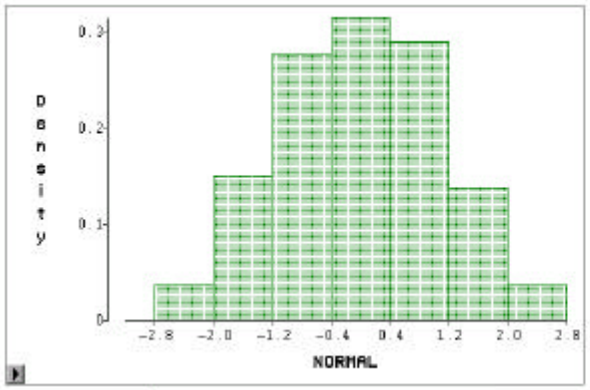


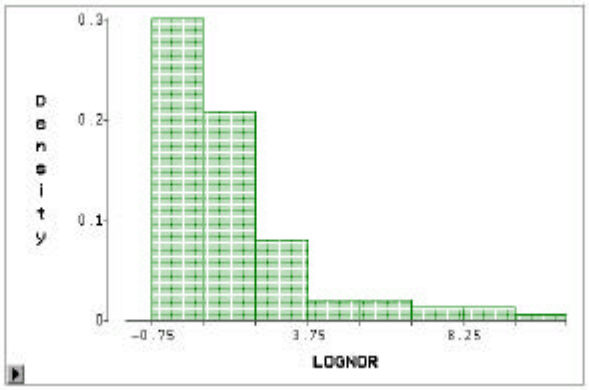
QQ Plot

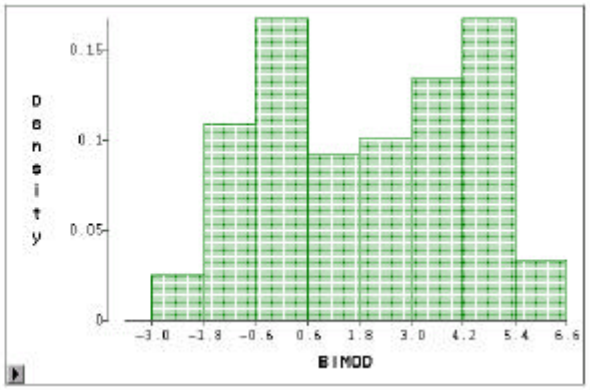


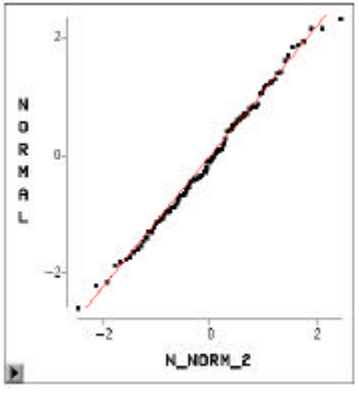
Ligne de référence

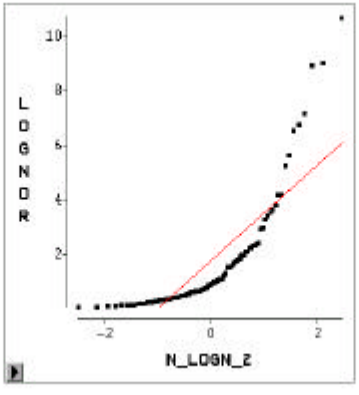
QQ plots typiques

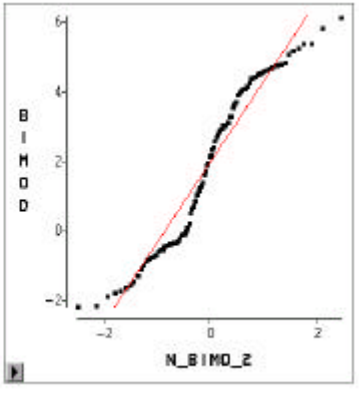










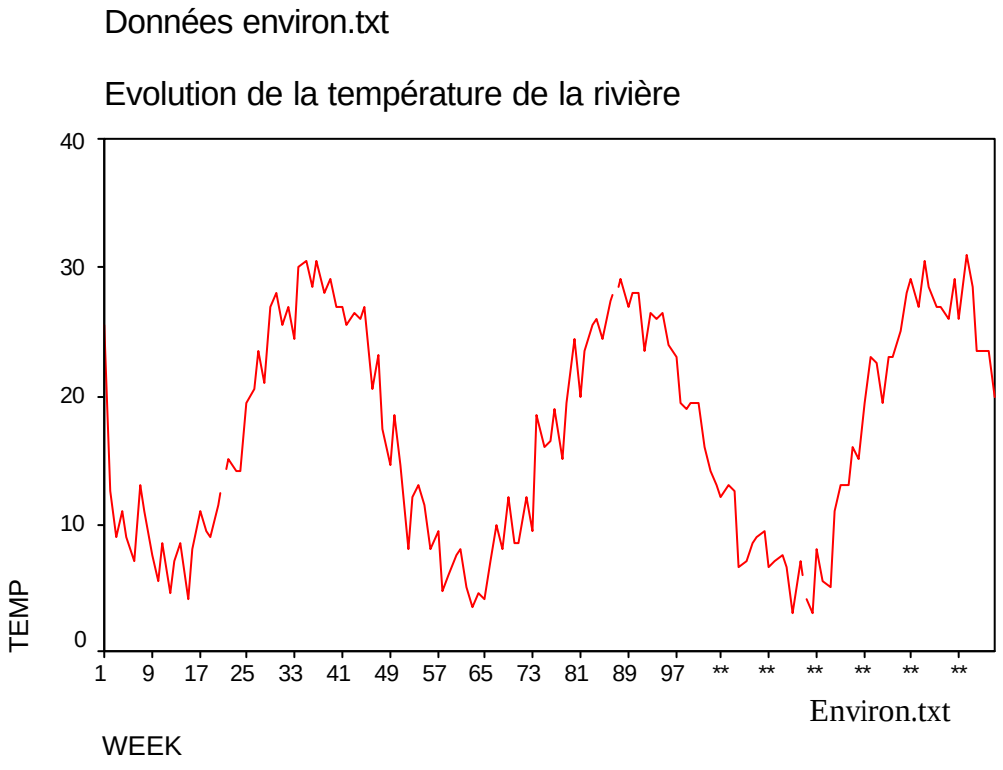


Attention à ce qui est mis en X et Y, cela dépend du software !

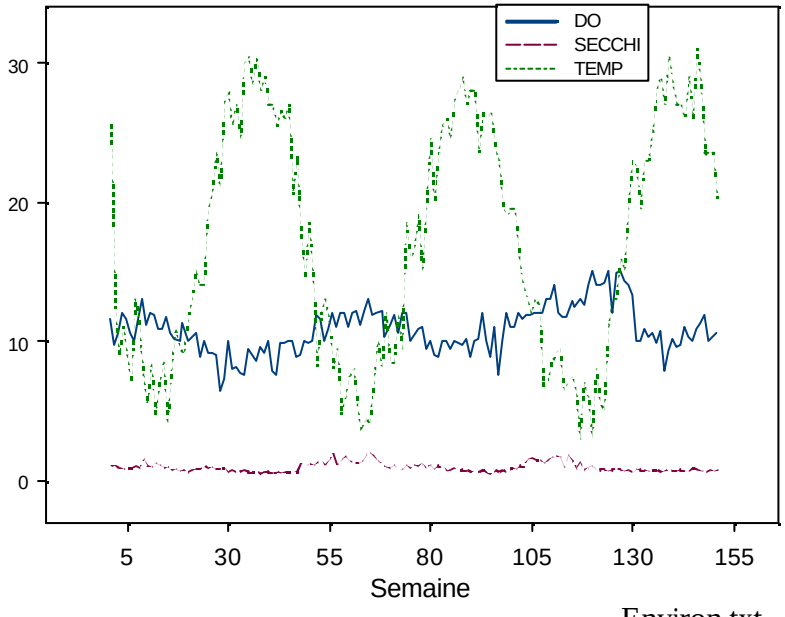
Le qqplot peut s'utiliser pour comparer deux distributions quelconques !

Diagramme temporel ou «time sequence plot»

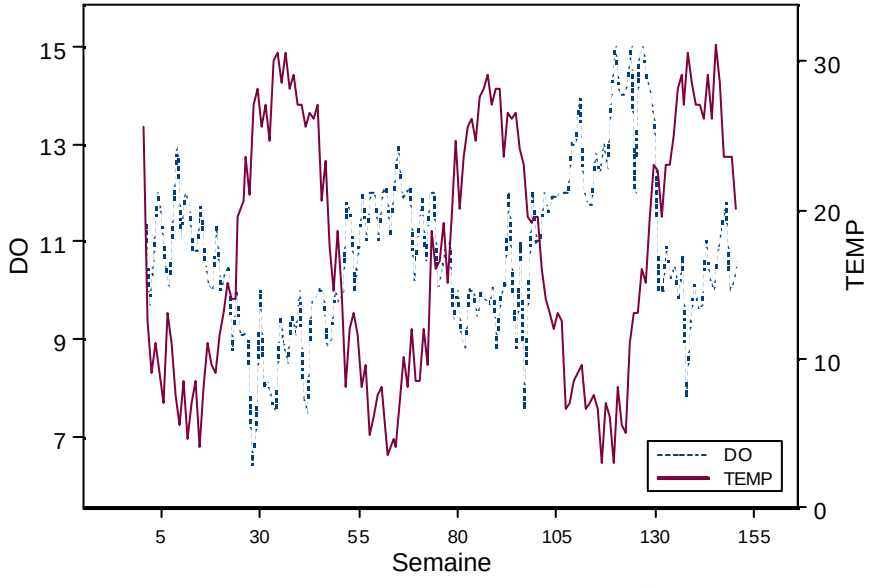
Un diagramme temporel est une représentation graphique d'une série de données quantitatives en fonction de l'ordre dans lequel elles ont été récoltées. Il permet de visualiser la valeur centrale et la variabilité des données ainsi que des tendances ou cycles.



Comparaison des tendances de plusieurs variables



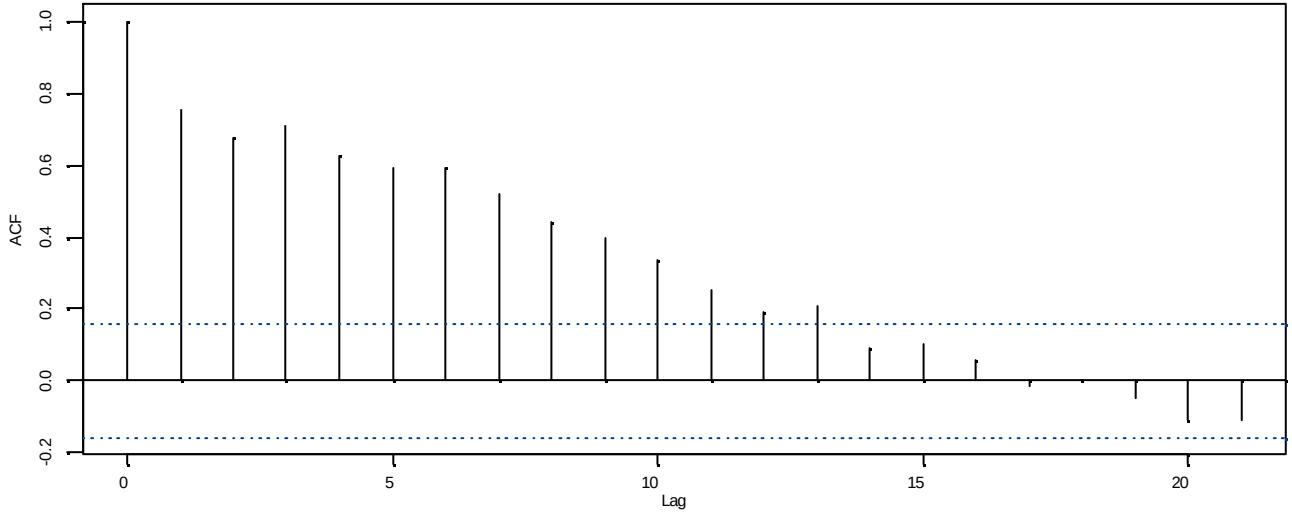
Environ.txt



Environ.txt

Graphique d'autocorrélation

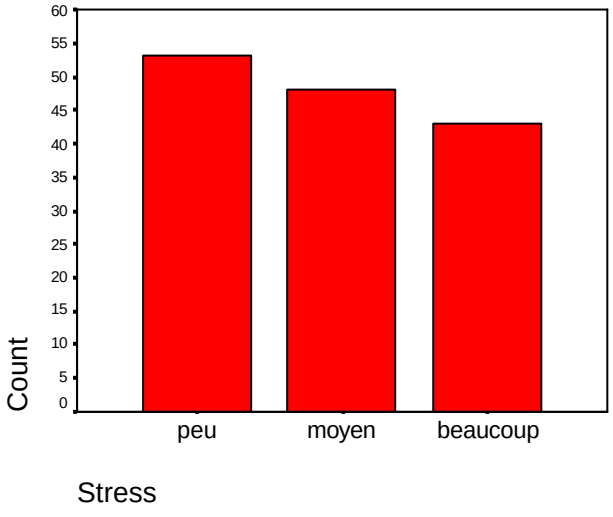
Le graphique d 'autocorrélation présente les autocorrélations d 'ordre 1 à k pour une série de données. C'est un outil qui permet de vérifier l'indépendance entre les observations de la série. Le même type de graphique se réalise pour les autocorrélations partielles.



Environ.txt, variable temp

UNE variable qualitative : Diagramme en barre et en tarte

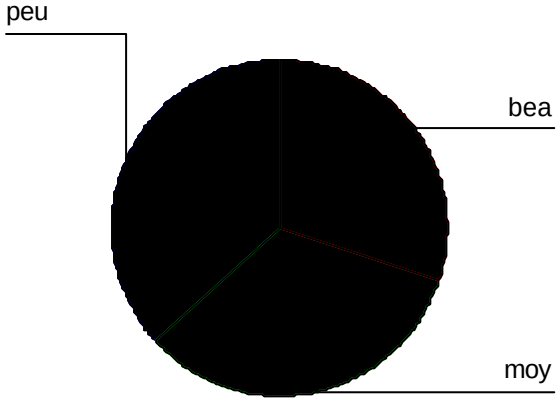
stress.txt, variable stress

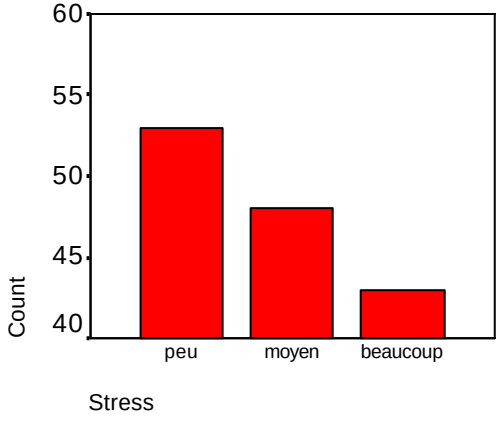


Le diagramme en barre ou en tarte (moins utile) permet de présenter les fréquences des niveaux d 'une variable catégorielle.

Attention !!!

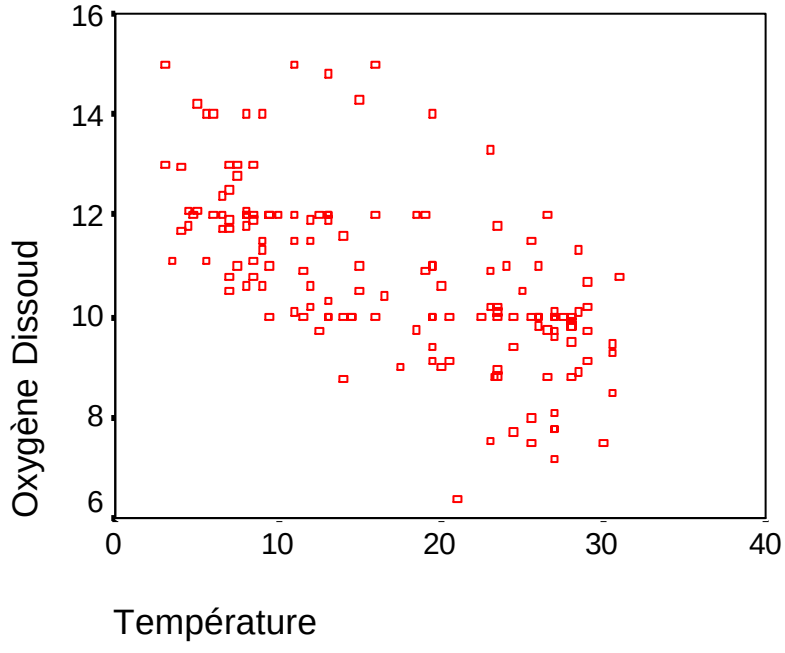
- Eviter de l 'utiliser pour des variables quantitatives.
- Mettre les niveaux dans l'ordre logique
- Se méfier absolument des diagrammes en barre avec l 'axe des Y ne commencant pas à 0





DEUX variables quantitatives : graphe X-Y

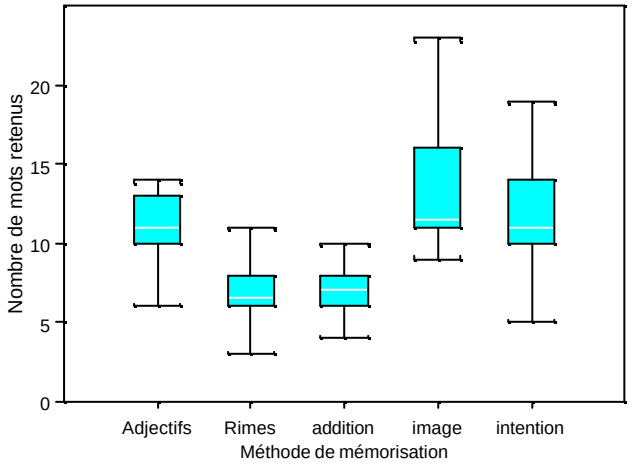
La graphe XY (ou scatter diagram) permet de visualiser la relation entre deux variables quantitatives



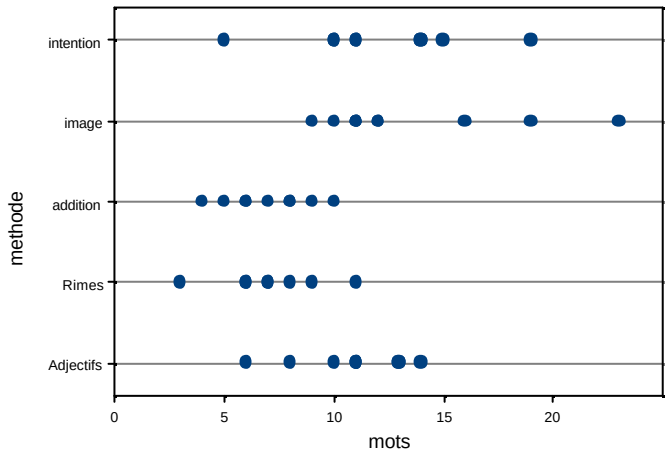
Environ.txt, variables temp et DO

Une variable quantitative et une qualitative

Box plot par catégories

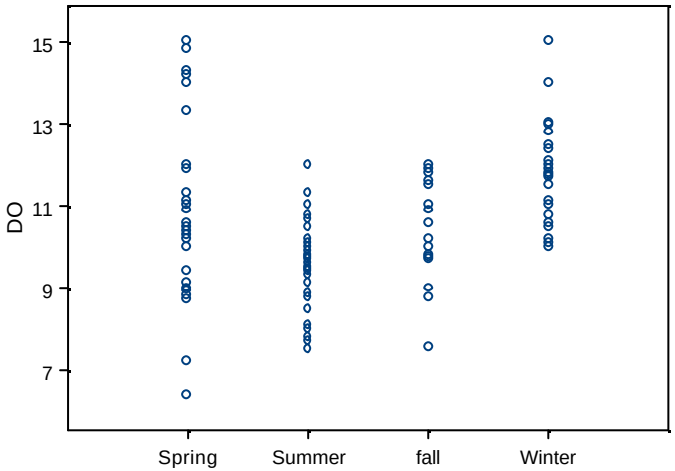


Graphe en points par catégories

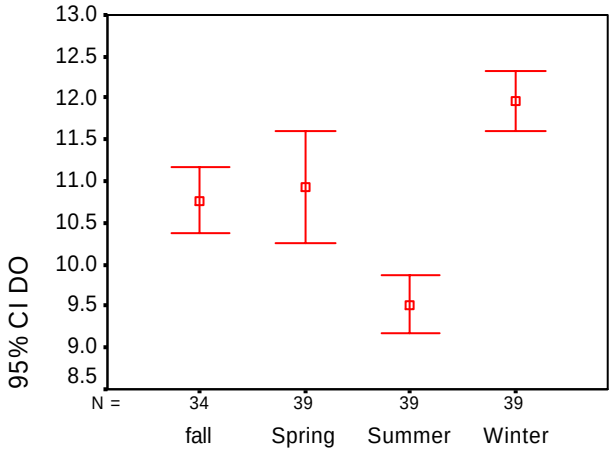


memoire.txt,
Variable mots par methode

Graphe en points par catégories

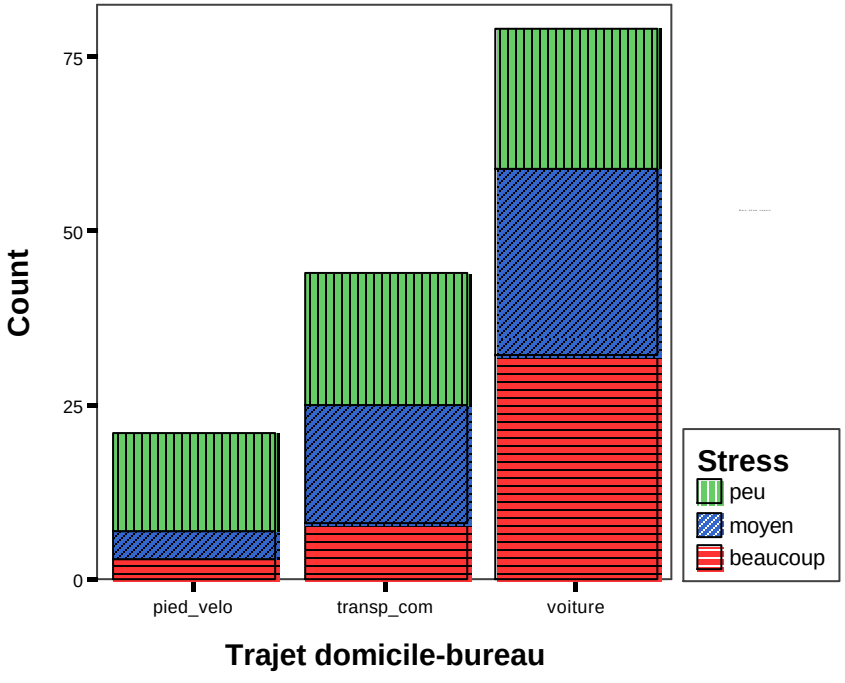


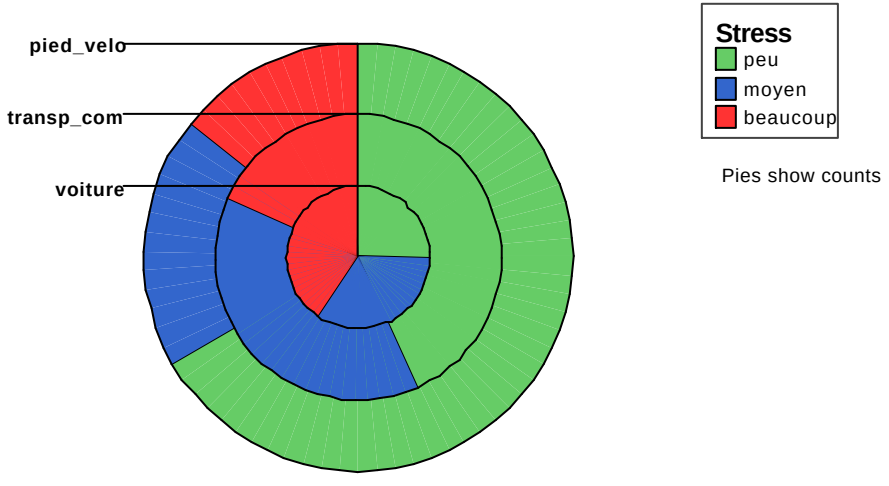
Valeurs centrales et barres d'erreur



Environ.txt,
Variable DO par saison

Deux variables qualitatives



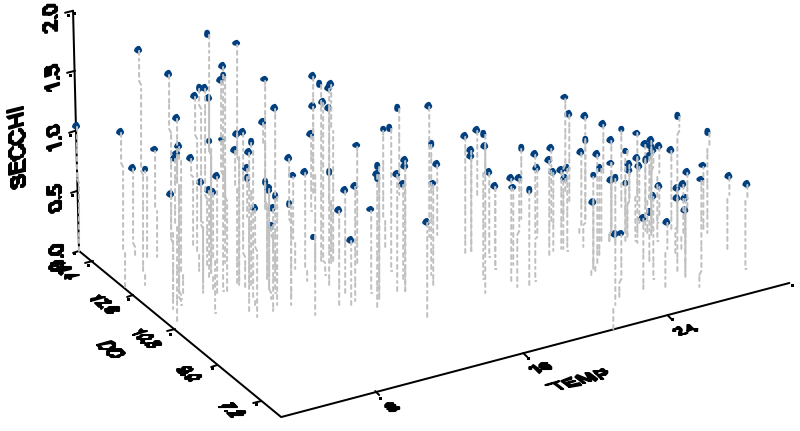


stress.txt,

Variables trajet et stress

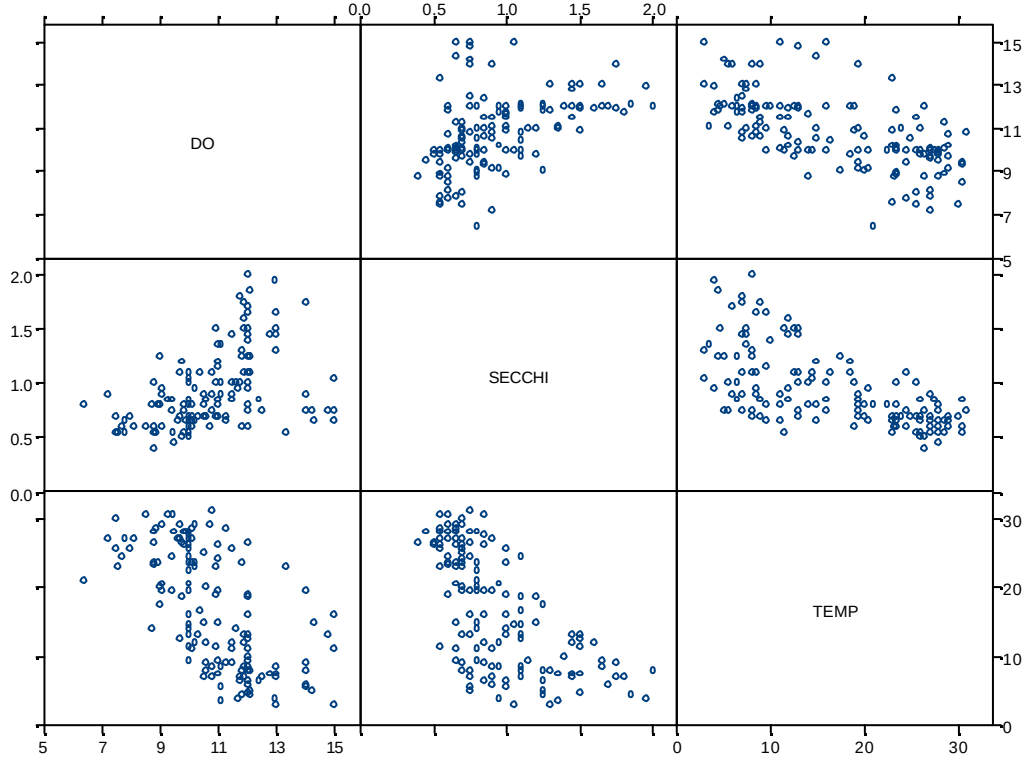
Trois variables quantitatives

Graphe X-Y-Z

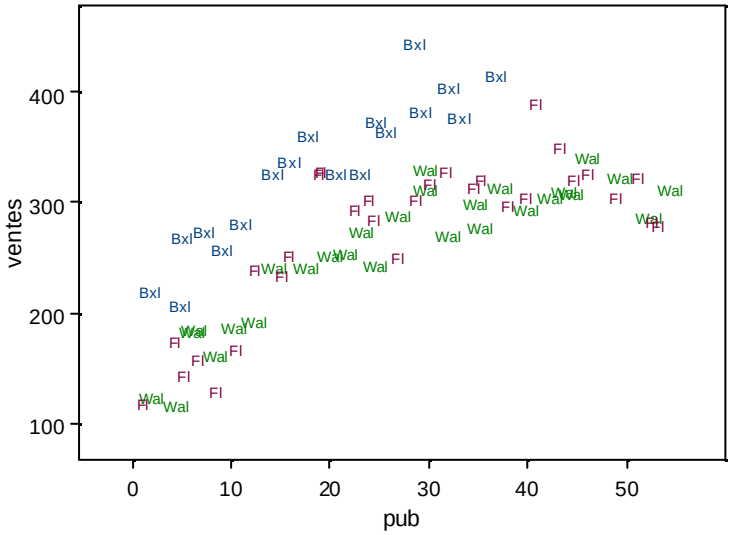


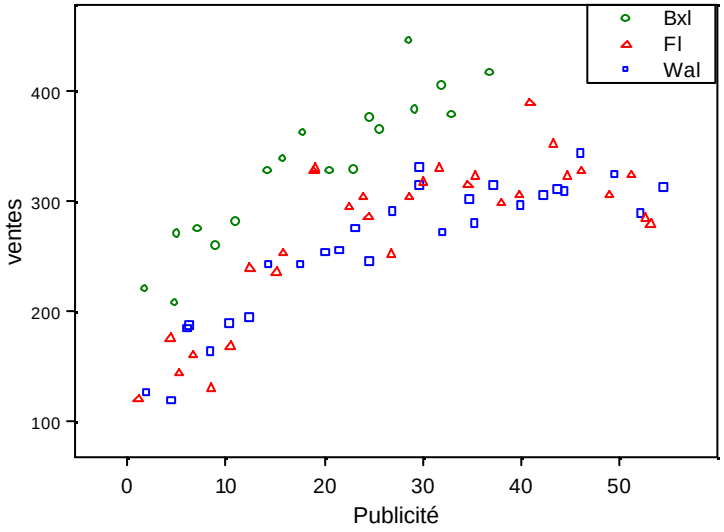
Environ.txt,
Variable DO, Temp, Secchi

Scatter matrix



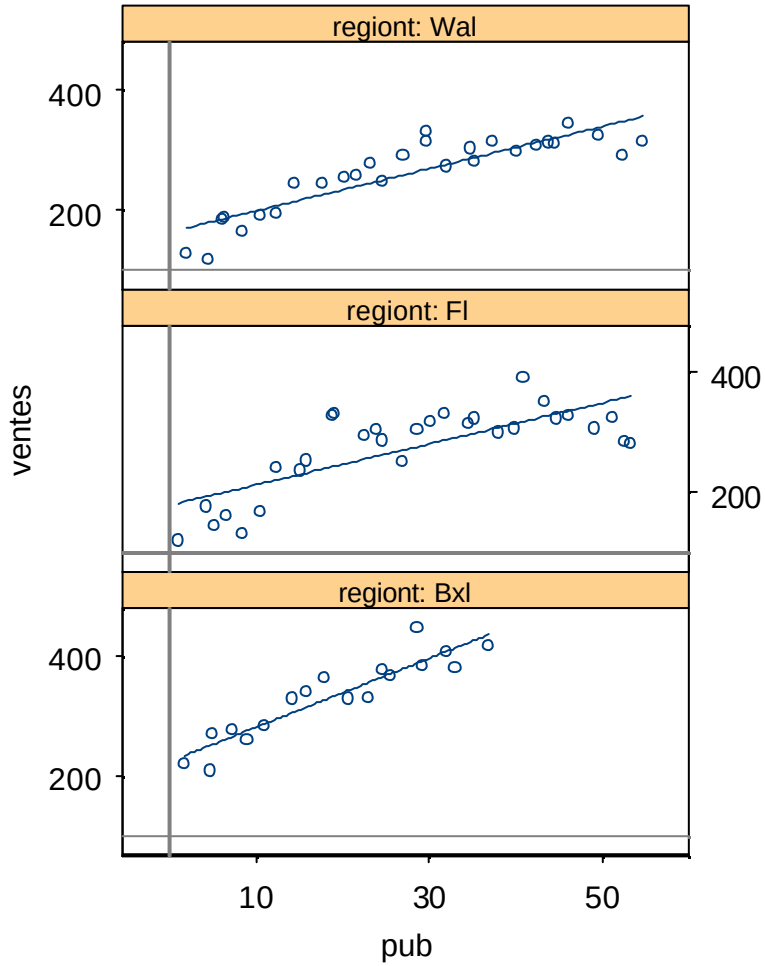
Deux variables quantitatives, une qualitative



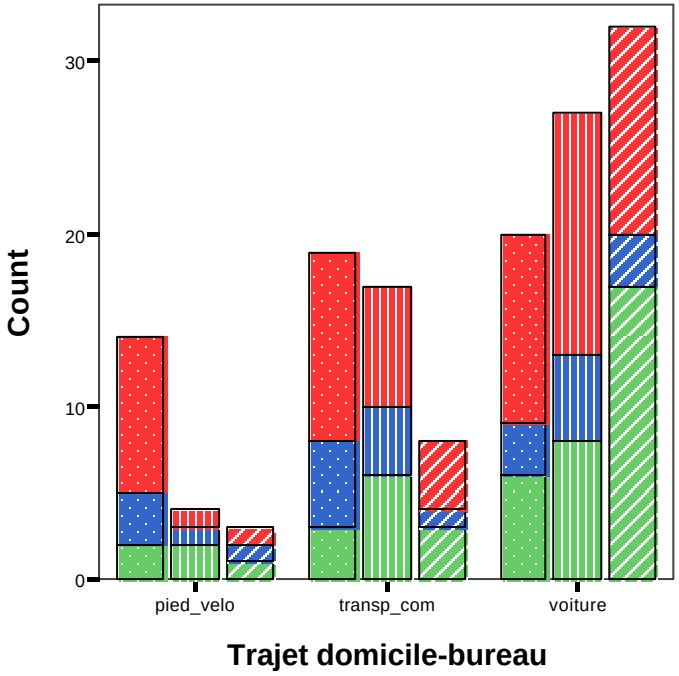


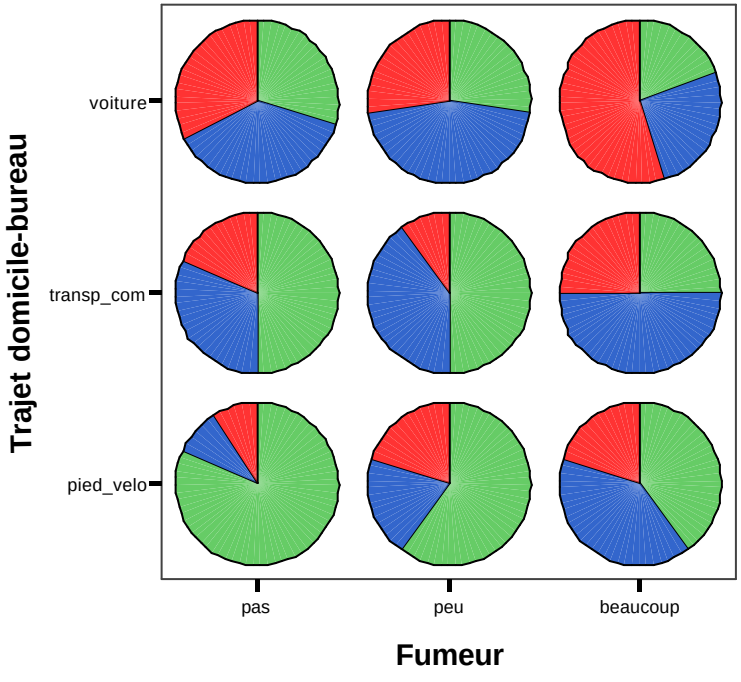
pubsplus.txt, Variables ventes et pub par region

Graphique en « treilli »



Trois variables qualitatives

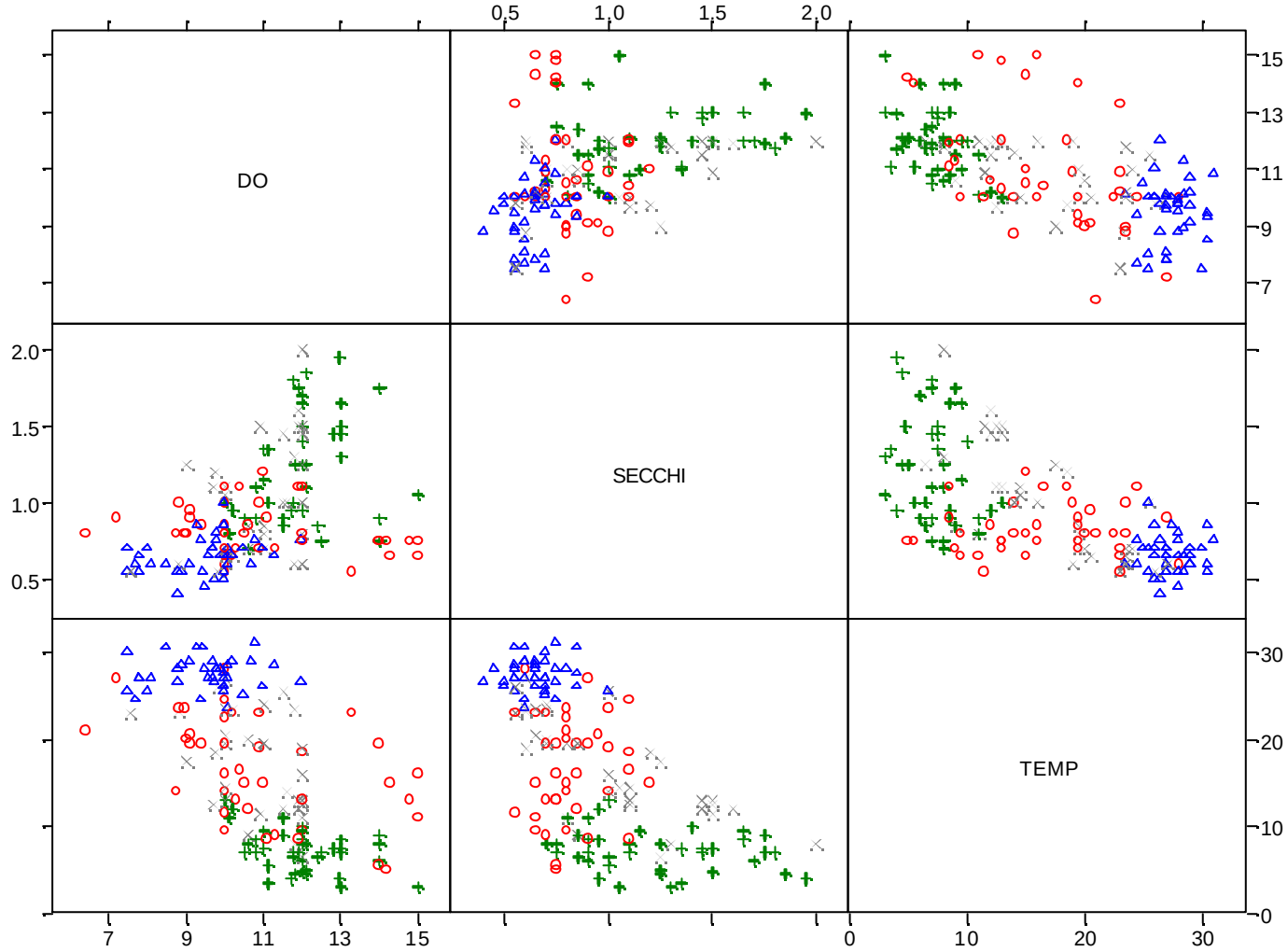




stress.txt,

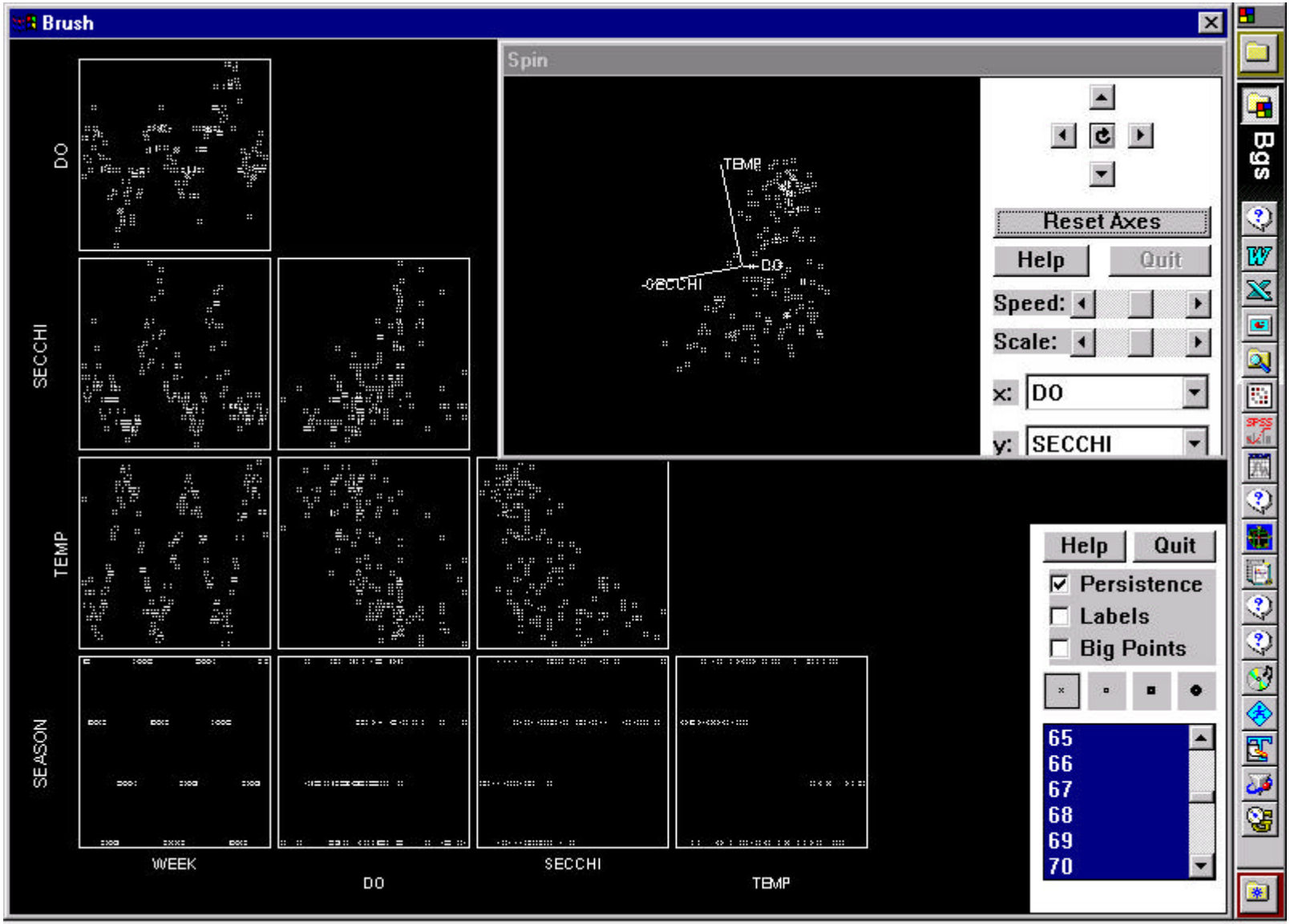
Variables trajet, fumeur et stress

Trois variables quantitatives et une qualitative



Environ.txt. Variable DO, Temp, Secchi par saison

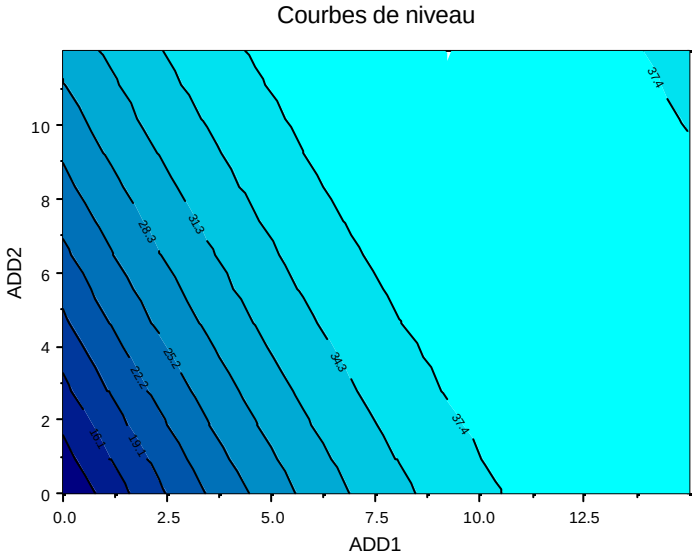
Visualisation dynamique 3D (brush and spin)

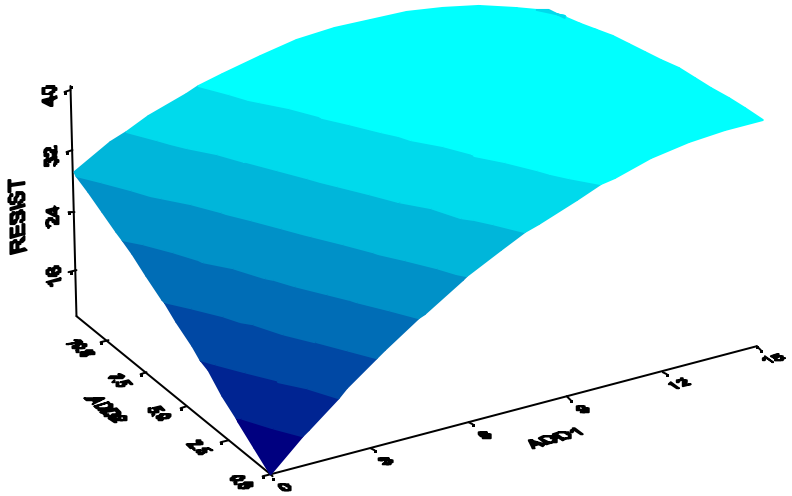


Courbes de niveau et surfaces de réponse.

Les graphiques en courbes de niveaux (contour plot) ou surface de réponse permettent de représenter une fonction z=f(x,y).

Il sont utiles pour représenter par exemple l'équation d'un modèle estimé ou un fonction à optimiser (ex. fonction de vraissemblance ou des moindres carrés…)





Recommandations pour la réalisation de « bons » graphiques

- Dans la présentation d'une série de données essayer de présenter chaque observation individuelle au moins une fois (pas uniquement des résumés)
- Mettre tous les résultats importants d'un travail statistique sous forme graphique.
- Ne pas trop remplir un graphique, les données doivent être l'information la plus visible.
- Choisir des limites pour les axes les plus proches possibles des intervalles de variation des données mais inclure le 0 quand c'est nécessaire (comptage).
- Choisir des échelles pour les axes qui permettent de visualiser au mieux les données (ex. Log.). Mettre dans ce cas si possible l'échelle réelle sur les axes.
- Quand deux graphiques doivent être comparés, utiliser les mêmes échelles.
- Utiliser un ligne ou des référence (ex. Moyenne) si utile. Entourer le graphe par un rectangle.
- Méfiance des graphiques 3D, ils sont difficile à interpréter.
- Libellez clairement les axes (avec les unités des variables), donner un titre, mettre une légende pour les symboles et couleurs.
- Attention aux couleurs, elle disparaissent à la reproduction...
- Expliquer clairement ce que sont les barres d 'erreurs quand il y en a
- La préparation d'un graphique est un travail itératif qui prend du temps et vient avec l'expérience...