MES DONNEES-DESCRIPTION : ANALYSE DES DONNEES

Table des matières

ACP

Données : Performances de 33 athlètes masculins au decathlon lors des JO de 1988.
Domaine d'application :
Thème et description des données : Les variables sont :
Pour decathlon.txt
{ 100m longueur poids hauteur 400m 110m disque perc javelot 1500m score}
Pour olympic$tab dans `ade4 sous R`
dossard : numero du dossard
m100 : course 100 metres
long : saut en longueur
poid : lancer du poids
haut : saut en hauteur
m400 : course 400 metres
m110 : course 110 metres
disq : lancer du disque
perc : saut a la perche
jave : lancer du javelot
m1500 : course de 1500 metres
et
dans **olympic$score** : vector of the final points scores of the competition
Nom du fichier : decathlon.txt ou olympic dans la library ade4 (?* ?)
Objectif(s) : comparer différents cidres.

Origine des données : Example 357 in:
Hand, D.J., Daly, F., Lunn, A.D., McConway, K.J. and Ostrowski, E. (1994) *A handbook of small data sets*, Chapman & Hall, London. 458 p.

Méthode(s) statistique(s) possible(s) : ACP, régression linéaire

Données : Etude olfacto-gustative de cidres
Domaine d'application :
Thème et description des données : Plusieurs caractéristiques du cidre ont été mesurées sur 10 cidres différents. Les variables sont :
Cidre : N° 1 à 10
Odeur : note entre 1 et 10
Sucre : note entre 1 et 10
Acide : note entre 1 et 10
Amer : note entre 1 et 10
Astringence : note entre 1 et 10
Suffocante : note entre 1 et 10
Piquante : note entre 1 et 10
Alcool : note entre 1 et 10

Parfum : note entre 1 et 10
Fruitée : note entre 1 et 10
Nom du fichier : cidre.txt (10*10)
Objectif(s) : comparer différents cidres.
Origine des données : J.-M. Labatte http://www.math.univ-angers.fr/~labatte
Méthode(s) statistique(s) possible(s) : ACP.

Données : Etude de différents laits
Domaine d'application :
Thème et description des données : Plusieurs caractéristiques du lait ont été mesurés sur le lait de 16 mammifères. Les variables sont :
id : identifiant mammifère
proteine
graisse
lactose
Nom du fichier : lait.txt (16*4)
Objectif(s) : comparer différents cidres.
Origine des données : J.-M. Labatte http://www.math.univ-angers.fr/~labatte
Méthode(s) statistique(s) possible(s) : ACP.

Données : Données techniques sur 62 véhicules année modèle 1994.
Domaine d'application :
Thème et description des données : les variables sont :
Puissance : en chevaux fiscaux
Cylindree : en cm**3**
Longueur : longueur de la voiture
Largeur : largeur de la voiture
Surface : surface de la voiture
Poids : poids total en Kg
Vitesse : vitesse maximum en km/h
DepArret : Temps, en secondes, pour parcourir 1000 m, départ arrêté.
Conso : Consommation moyenne aux 100 Km, en litres (essence ou gazole).
Nom du fichier : voit94 (62*10)
Objectif(s) : comparer différents modèles de voitures et ajouter en individus supplémentaires les 10 modèles de voitures (fichier **donsup**) dont les données manquantes ont été remplacées par des valeurs grossièrement estimées (données peu fiables à introduire donc en supplémentaire dans l'analyse).
Origine des données : André Bouchier (INRA)
Méthode(s) statistique(s) possible(s) : ACP.

Données : Comparaison Charollais-Zébus
Domaine d'application :
Thème et description des données : This data set gives six different weights of 23 charolais and zebu oxen. Les variables sont :
vif
carc
qsup
tota
gras
os
race
Nom du fichier : zebu.txt (23*7) ou chazeb dans la library `ade4`
Objectif(s) : comparer charollais et zébus.
Origine des données : Tomassone, R., Danzard, M., Daudin, J. J. and Masson J. P. (1988) *Discrimination et classement*, Masson, Paris. p. 43
Méthode(s) statistique(s) possible(s) : ACP, AFD, …

Données : Comparaison variétés de pommes
Domaine d'application :
Thème et description des données : Treize dégustateurs (codés de "a" à "m") ont dégusté deux fois trois mêmes variétés de pommes (Granny, Jonagored et Canada) au cours de la même séance afin d'évaluer leur niveau de répétabilité. Les dégustations ont eu lieu dans le laboratoire normalisé du GRAPPE (ESA, Angers). Les pommes sont servies l'une après l'autre. Les experts les caractérisent en notant les 16 descripteurs qu'ils utilisent habituellement. Ces descripteurs sont des variables quantitatives correspondant chacun à une note comprise entre 0 et 10. Le tableau de mesure analysé comporte 16 descripteurs en colonne et 78 lignes (13 dégustateurs * 3 variétés de pomme * 2 répétitions). Les variables (descripteurs) sont :
PRODJUGE : identifiant - *exemple : **a_CAN1***
***a** = identification du juge : 13 juges notés de a à m*
***_CAN** = code de la pomme dégustée (CAN = Canada ;JON = Jonagored ; GRA = Granny)*
***1** = répétition : la 1ère pomme = 1 et la 2ème = 2*
O_GLO = Odeur globale
O_TER = Odeur de terre
O_CAV = Odeur de cave
O_FF = Odeur de feuilles fraîches
FLET = Fletri
RUG = Rugueux
RESI = Résistance au toucher
CROQ = Croquant
JUTE = Juteux
FARI = Farineux
MASTI = Mastication
SUCR = Sucré
ACID = Acide
I ARO = Intensité aromatique
FOND = Fondant
ASTRI = Astringence
Nom du fichier : pomme.txt (78*16)
Objectif(s) : caractériser les trois variétés de pomme et de savoir si le panel de dégustateurs est répétable.
Origine des données : laboratoire GRAPPE, ESA, Angers
Méthode(s) statistique(s) possible(s) : ACP, AFD, …

Données : Température mensuelle de 37 villes du monde.
Domaine d'application : Climatologie
Thème et description des données : les données sont les températures mensuelles (en °C) de 37 villes du monde.
Nom du fichier : Climat (37*12)
Objectif(s) : comparer différentes façons d'utiliser les valeurs mensuelles (les 12 valeurs, les 11 différences entre deux mois successifs) pour tenter de définir les différents climats de la terre (cf. par exemple Biométrie (p364-367). Ajouter en individus supplémentaires les 10 modèles de voitures (fichier **Climat--villesSup**)
Origine des données :
Méthode(s) statistique(s) possible(s) : ACP, AFC.

Données : Typologie de la consommation alimentaire des régions françaises
Domaine d'application : Nutrition
Thème et description des données : l'Institut National des Statistiques Economiques publie chaque année des tableaux concernant l'alimentation de différentes régions françaises. Ils peuvent être intéressants pour étudier la variabilité de la consommation alimentaire des français. Ces tableaux (de l'année 1986) sont de deux types pour divers produits :

1. Les quantités (ConsoF1) :

- Code : code INSEE
- Aliment : nature de l'aliment
- Unit : de trois types en poids (kg), en litre (l), en unité (n)
- Paris : Région parisienne
- Bassin : Bassin Parisien
- Nord : Nord
- Est : Est
- Ouest : Ouest
- SudOu : Sud-Ouest
- CentEst : Centre-Est
- Medit : Méditerranée
- France : ensemble de la France

2. Les dépenses (ConsoF2) :

même présentation, l'unité est le franc (FF). Toutefois, certains produits n'ont pas d'équivalent en quantité et la valeur de Unit est notée zFF.
Nom du fichier : ConsoF1.txt(76*11), ConsoF2.txt (109*11)
Objectif(s) : tenter une typologie des régions, noter que les différentes unités pour les quantités mérite réflexion.
Origine des données : Collection INSEE Série M, n°117 (1986), 95-97 & 101-102.
Méthode(s) statistique(s) possible(s) : ACP, AFC.

Données : Evolution de la production de céréales en Asie (1972-1982)
Domaine d'application : Agronomie
Thème et description des données : la Commission Economique et Sociale pour l'Asie et le pacifique, organisme dépendant des Nations Unies, publie un certain nombre de tableaux sur les statistiques agricoles. Celui fourni concerne la production par habitant de l'ensemble des céréales et du riz pour une période allant de 1972 à 1982. Il est constitué de la manière suivante :

- Pays : nom du pays
- A72, A73, A74, A75, A76, A77, A78, A79, A80, A81, A82 : valeur de la production par habitant
- Production : Cereale ou Riz

Nom du fichier : Cereale (54*13)
Objectif(s) : quelle information utile peut-on extraire de tels tableaux ?
Origine des données : ONU
Méthode(s) statistique(s) possible(s) : ACP, Classification, …

Données : U.S crime data
Domaine d'application : Sociologie
Thème et description des données : This is a data set consisting of 50 measurements of 7 variables. It states for one year (1985) the reported number of crimes in the 50 states of the U.S. classified according to 7 categories (X3-X9-).. Il est constitué de la manière suivante :

- X1 : land area (land)
- X2 : population 1985 (popu 1985)
- X3 : murder (murd)
- X4 : rape
- X5 : robbery (robb)
- X6: assault (assa)
- X7 : burglary (burg)

- X8 : larcery (larc)
- X9 autothieft (auto)
- X10: US states region number (reg)
- X11: US states division number (div)

division numbers	*region numbers*
New England 1 Mid Atlantic 2 E N Central 3 W N Central 4 S Atlantic 5 E S Central 6 W S Central 7 Mountain 8 Pacific 9	Northeast 1 Midwest 2 South 3 West 4

Nom du fichier : uscrime (50*12)
Objectif(s) : Quelle information utile peut-on extraire d'un tel tableau ?
Origine des données :
Méthode(s) statistique(s) possible(s) : ACP, Classification, …

Données : U.S health data
Domaine d'application : Sociologie
Thème et description des données : This is a data set consisting of 50 measurements of 13 variables. It states for one year (1985) the reported number of deaths in the 50 states of the U.S. classified according to 7 categories. Il est constitué de la manière suivante :

- X1 : land area (land)
- X2 : population 1985 (popu 1985)
- X3 : accident (acc)
- X4: cardiovascular (card)
- X5:: cancer (canc)
- X6: pulmonar (pul)
- X7: pneumonia flu (pnue)
- X8: diabetis (diab)
- X9: liver (liv)
- X10: Doctors (doc)
- X11: Hospitals (hosp)
- X12: U.S. states region number (r)
- X13: U.S. states division number (d)

division numbers	*region numbers*
New England 1 Mid Atlantic 2 E N Central 3 W N Central 4 S Atlantic 5 E S Central 6 W S Central 7 Mountain 8 Pacific 9	Northeast 1 Midwest 2 South 3 West 4

Nom du fichier : ushealth (50*13)
Objectif(s) : Quelle information utile peut-on extraire d'un tel tableau ?
Origine des données :
Méthode(s) statistique(s) possible(s) : ACP, Classification, …

Données : Profil météorologique mensuel moyen de 34 villes françaises.
Domaine d'application : Météorologie
Thème et description des données : les données sont les températures mensuelles (en °C) de 37 villes du monde.
Nom du fichier : PluieFrance.csv (34*34)
Travail à effectuer : Le but de l'´etude est de comparer les profils météorologiques de 34 villes françaises.
- **Objectif(s) :** Quelle information utile peut-on extraire d'un tel tableau ? Réaliser une typologie des villes ; proposer un bilan des liaisons entre les variables ; étudier si les ressemblances ou les dissemblances correspondent à des proximités (ou des éloignements) géographiques ; comparer différentes façons d'utiliser les valeurs mensuelles de pluie (les 12 valeurs, les 11 différences entre deux mois successifs) pour tenter de définir les différents climats observés en France métropolitaine

Origine des données :
Méthode(s) statistique(s) possible(s) : ACP, Classification, …

AFC et AFCM

Données : Elections présidentielles de 2002 en France
Domaine d'application :
Thème et description des données : Dans un sondage, on a interrogé les lecteurs de 12 périodiques sur leur vote au premier tour des présidentielle 2002 (100 lecteurs par périodiques). La table croise les deux variables « candidat à l'élection) et périodique.
Nom du fichier : election2002.txt (16*12)
Objectif(s) : analyser le profil des lecteurs.
Origine des données :
Méthode(s) statistique(s) possible(s) : AFC

Données : Consommation annuelle pour différentes denrées alimentaires en franc d'un ménage français en 1972
Domaine d'application :
Thème et description des données : consommation annuelle en francs d'un ménage pour différentes denrées alimentaires en 1972. MA, EM, CA indiquent la catégorie socio-professionnelles et 2, 3, 4, 5 la taille du foyer.
Nom du fichier : csp.txt (13*7)
Objectif(s) : analyser le profil des consommateurs.
Origine des données :
Méthode(s) statistique(s) possible(s) : AFC

Données : Poissons de la baie de Cocody
Domaine d'application : Hydrobiologie
Thème et description des données : Dans une étude, effectuée en 1965, sur le peuplement de poissons de la baie de Cocody on a prélevé, tous les mois, des échantillons de 42 espèces.
Nom du fichier : Cocody.txt (42*12)
Objectif(s) : Pour analyser le tableau de données la difficulté principale est celle du traitement des espèces peu représentées.
Origine des données : Daget, J. (1976) *Les modèles mathématiques en écologie.* Masson, Paris.
Méthode(s) statistique(s) possible(s) : AFC, ACP

Données : Critère de confiance pour investissements
Domaine d'application : Finance
Thème et description des données : les banques demandent à des experts de donner une note de confiance à des experts sur 15 critères. Ces notes s'étalent entre 0 et 4 (0 : conditions inacceptables ; 1 : mauvaises conditions ; 2 : conditions médiocres ; 3 : conditions acceptables ; 4 : excellentes conditions). Les critères sont les suivants :
- STAB : Stabilité politique

- ATTI : Attitude envers les investisseurs étrangers
- NATI : Nationalisation
- INFI : Inflation monétaire
- BALA : Balance des paiements
- RETA : Retards administratifs
- CROI : Croissance économique
- CONV : Convertibilité de la monnaie
- EXEC : Exécution des contrats
- COUT : Coût et rendement de la main-d'œuvre
- SERV : Services professionnels et contractuels
- COMM : Communication (télex, téléphone, courtier)
- AGLO : Agence locales et réseaux commerciaux
- CRCT : Crédit à court terne
- CRET : Emprunts à long terme, Risques financiers

Le tableau de données est constitué par les notes moyennes de 15 experts pour 43 pays et pour les 15 quinze critères.

Nom du fichier : Confiance.txt (43*16)
Objectif(s) : conseiller les clients des banques qui souhaitent faire des investissements à l'étranger.
Origine des données : Cahier de l'Analyse des Données
Méthode(s) statistique(s) possible(s) : AFC

Données : Recettes de cuisine
Domaine d'application : Nutrition
Thème et description des données : les auteurs de recettes de cuisine choisissent des types de recettes fort différents ; ainsi pour les recettes de 30 catégories de plats, 13 auteurs ont fait le choix décrit dans le tableau de données avec en ligne les recettes et en colonnes les auteurs. Les auteurs ou les régions concernées sont les suivants :

- A1 : Escoffier
- A2 : Pellaprat
- A3 : Bretagne
- A4 : Occitane
- A5 : Périgord
- A6 : Bocuse cuisine
- A7 : Bocuse marché
- A8 : Guérard gourmande
- A9 : Guérard minceur
- A10 : Olympe
- A12 : Troisgros
- A13 : Vergé.

Nom du fichier : Cuisine.txt (29*13)
Objectif(s) : établir une typologie des catégories et des livres
Origine des données :
Méthode(s) statistique(s) possible(s) : AFC

Données : Criminalité aux Etats-Unis en 1977
Domaine d'application : Sociologie
Thème et description des données : pour les 50 états des Etats-Unis on donne, pour l'année 1977, les taux criminels pour 100000 habitants pour les sept types de crimes. Le fichier est constitué de la façon suivante :

- Etat : nom de l'état
- Meurtre : meurtre
- Viol : viol
- Vol : vol avec intimidation
- Agression : agression
- Cambriolage : cambriolage
- Larcin : larcin « furtif »
- VolSimul : simulation de vol

Nom du fichier : CrimeUS.txt (50*7)
Objectif(s) : faire une typologie des 50 états
Origine des données : Young, F. (1989) Visualizing Six-Dimensional Structure with Dynamic Statistical Graphics. *Chance*, **2**, n°1.
Méthode(s) statistique(s) possible(s) : AFC

Données : Comparaison de l'avifaune en Provence et en Corse
Domaine d'application : Zoologie
Thème et description des données : deux successions sont définies par les caractéristiques communes de l'habitat ; ces successions représentent différents types de physionomie de la végétation que l'on peut assimiler à un gradient. Ces types vont de 1 à 6 :

- de 1 : milieu matorral bas (formation végétale des pays méditerranéens, plus ouverte que le maquis, la hauteur ne dépasse pas 50 cm.)
- à 6 : forêt de chêne vert (*Quercus ilex*).

Les deux gradients ne sont pas strictement équivalents : on peut y trouver des différences tant du point de vue floristique, les stades 1 à 5 sont sur sol calcaire en Provence, sur sol siliceux en Corse, que du point de vue de la physionomie. Les localisations sont indiquées sur la figure ci-jointe. On a observé l'abondance spécifique de 60 espèces d'oiseaux par stade et par région.
Nom du fichier : Avifaune.txt (61*13)
Objectif(s) : On demande d'analyser ces données par une analyse factorielle des correspondances. Existence de différences entre les successions entre les deux régions ? Peut-on déceler une ressemblance dans certains stades ?
Origine des données : Blondel, J. (1986) *Biogéographie évolutive*, Masson, Paris.
Méthode(s) statistique(s) possible(s) : AFC
Figure : Avifaune.JPG (52Ko)

Données : Blé dur
Domaine d'application :
Thème et description des données : 50 observations et 11 variables. Il contient les résultats d'un suivi agronomique sur 50 parcelles de blé dur.
RDT Rendement en grains
PLM Nb de plantes par m²
ZON Zone géographique
ARG Taux d'argile de la parcelle
LIM Taux de limon de la parcelle
SAB Taux de sable de la parcelle
VRT Variété cultivée
PGM Poids de 1000 grains
MST Matière sèche totale à la récolte
AZP Azote dans la plante à la récolte
VRTC Variété cultivée (codée en 3 classes)
Nom du fichier : bledur.txt (50*11)
Objectif(s) : analyser le profil des blés durs.
Origine des données :
Méthode(s) statistique(s) possible(s) : AFCM

Données : Etude du temps moyen passé à différentes activités.
Domaine d'application : Sociologie
Thème et description des données : Le tableau de données est un tableau croisant 28 catégories de personnes (l'ensemble I des lignes) et 10 activités (l'ensemble J des colonnes). Le nombre figurant dans la case (i,j) est le temps moyen (en centièmes d'heures) passé par la catégorie i à réaliser l'activité j. Les catégories sont obtenues en croisant 4 pays ou groupes de pays par 7 types d'individus (soit 28 lignes). Le code sur 3 caractères de chaque population reproduit donc dans ses 2 premières lettres le type d'individus, soit :

- HA : pour les hommes actifs
- FA : pour les femmes actives
- FN : pour les femmes non actives

- HM : pour les hommes mariés
- FM : pour les femmes mariées
- HC : pour les hommes célibataires
- FC : pour les femmes célibataires

La 3e et dernière lettre indique le code des 4 pays ou groupe de pays, soit :

- U : pour les USA
- W : pour 4 pays de l'Ouest
- E : pour 2 pays de l'Est
- Y : pour la Yougoslavie

Les activités (en colonnes) sont codées de la manière suivante :

- PRO : pour le travail professionnel
- TRA : pour les transports liés au travail
- MEN : pour le travail ménager
- ENF : pour les occupations liées aux enfants
- COU : pour les courses
- TOI : pour les soins personnels
- REP : pour les repas
- SOM : pour le sommeil
- TEL : pour la télévision
- LOI : pour les loisirs

Nom du fichier : BudgetTemps (28*11)

Objectif(s) : Les données proviennent d'une Recherche Coopérative Internationale sur les budgets-temps, menée dans différents pays sous l'égide de l'UNESCO en 1965-66. Le but de cette recherche était d'étudier le temps moyen passé par des populations adultes à différentes activités (travail, loisirs ...).

Origine des données : JAMBU M. *Classification automatique pour l'analyse des données*, DUNOD. 1978.

Méthode(s) statistique(s) possible(s) : AFCM

 # Autres données

- "**heberg.txt**" **:** Ce fichier est un tableau de contingence croisant les catégories socio-professionnelles avec des modes de résidence en vacances. (source: M. Goguel 1965). Les catégories professionnelles sont :

AGRI : agriculteurs
SALAG : salariés agricoles
PATR : patrons
CADSUP : cadres supérieurs & prof libérales
CADMOY : cadres moyens
EMPL : employés
OUVR : ouvriers
SERV : personnels de service
AUTRE : autres actifs
NONACT : non actif
Les modes de résidence sont :
HOTEL : Hôtel, pension de famille
LOCAT : Maison louée chez l'habitant
MAISON : Maison en propriété
PARENT : chez des parents
AMIS : chez des amis
CAMPING : Tente, caravane
VVF : Village de vacances
DIVERS : autres modes

- Proximité entre couleur des yeux et couleur des cheveux. Vous pouvez directement saisir les données dans votre logiciel.

Couleur des yeux Couleur des cheveux

Y/X	CBlond	CBrun	CNoir	CRoux
YBleu	1768	807	189	47
YGrisVert	946	1387	746	53
YBrun	115	438	288	16

- "**mariage.txt**" **:** Ce fichier est un tableau de contingence décrivant les mariages entre catégories professionnelles. Il croise pour les hommes et les femmes les catégories professionnelles suivantes :

agri : agriculture
ouva : ouvrier agricole
pat : patron
sup : cadre supérieur
moy : cadre moyen
emp : employé
ouv : ouvrier
serv : services
aut : autre

- "**psysoc.txt**" **:** Pour 19 pays, ce fichier présente le nombre de morts en fonction des causes. Les variables sont :

Pays : nom du pays
Suici : nombre de suicides
Omic : nombre d'homicides
Arout : nombre d'accidents de la route
Aindu : nombre d'accidents industriels
Aautr : autres type de mort
Cirfo : nombre de ciroses du foi

- "**science.txt**" **:** Sur six années consécutives, ce fichier présente les choix de types d'études suivis par les étudiants de grandes écoles.
- "**house**"

This data frame contains four columns : wife, altern, husband and jointly. Each column is a numeric vector.
Les tâches ménagères à assumer sont :
• Laundry Main_meal Dinner Breakfeast
• Tidying Dishes Shopping Official Driving
• Finances Insurance Repairs Holidays

AFC multiple

- **"bledur" :** Vous trouverez dans le fichier "**bledur**" les variables suivantes (dans l'ordre) :

RDT='Rendement en Qt/ha'
PLM='Nb de pieds levés par m2'
ZON='Numero de zone'
ARG='Argile du sol'
LIM='Limon du sol'
SAB='Sable du sol'
VRT='Variete cultivee'
PGM='Poids de 1000 grains'
MST='Matière sèche totale'
AZP='Azote dans la plante'
Comme le découpage en classe des données est souvent une activité fastidieuse, on utilisera le fichier "**bledurCD**" dans lesquels les données ont déjà été codées en classes.
Effectuez une AFC multiple sur ces données, caractérisez les axes 1 et 2 du plan factoriel.

- **"sencd" :**Le fichier **sencd** est issu d'une enquête sur la traction animale. Les variables sont :

EX : numéro d'exploitation
QU : numéro de quartier (village)
AC : nombre d'actifs dans la famille
SP : surface possédée
SU : surface agricole utile
AT : nombre d'ânes de traction
CT : nombre de chevaux de trait
BT : nombre de paires de boeux de trait
VT : nombre de paires de vaches de trait
BV : nombre de bovins hors exploitation
OV : nombre d'ovins
CP : nombre de caprins
Les variables codées en classes sont : ACcd, SPcd, ..., OVcd, CPcd

Classification

Données : Taille des particules d'un sol
Domaine d'application : Agronomie
Thème et description des données : Pour déterminer la distribution de la taille de particules dans des profils de sol, une expérience a été réalisée à l'Université de Californie. Vingt parcelles, de forme carrée, ont été aléatoirement sélectionnées sur un site de 150 hectares. Les pourcentages de sable, limon et argile ont été déterminés à 12 profondeurs (de 15.2 en 15.2 cm.). Le fichier de données est constitué de la façon suivante :

- Profond : profondeur de la mesure (1à 12 à partir du niveau du sol)
- Sable : % sable
- Limon : % limon
- Argile : % argile
- Parc : n° de la parcelle (P1 à P20)

Nom du fichier : SolParticule.txt (240*5)
Objectif(s) : analyser la distribution des tailles de particules et en faire la cartographie.
Origine des données : Nielsen, D.R., Biggar, J.W., Erh, K.T. (1973) Spatial variability of field-measured soil-water properties. *Hilgardia,* **42**, 215-259.
Méthode(s) statistique(s) possible(s) : Classification
Figure : SolParticule.JPG (22Ko)

Données : Pomme de terre
Domaine d'application : Agronomie
Thème et description des données : dans la revue de l'Institut Technique de la Pomme de Terre voici une présentation : « *Sautée, rissolée, en salade, frite ou en purée, au four, à la cendre, à la vapeur, oui, nous aimons la pomme de terre. Un Français sur dix en mange d'ailleurs plusieurs fois par semaine et un sur trois, tous les jours. Aliment de base presque parfait, équilibré, il combine à merveille protéines (0,23 g pour 100 g), lipides (0.07 g pour 100 g), minéraux (0,90 g pour 100 g), vitamine C (10 mg pour 100 g), sucres réducteurs (0,06 g pour 100 g), amidon (13,5 g pour 100 g), glucides (13,3 g pour 100 g), matière sèche (13,1 g pour 100 g)* ».

Parmi les 1200 variétés de pomme de terre voici quelques descriptions de six d'entre elles :

- BINTJE : sur le marché, sauf pendant la saison des primeurs. Assez bonne tenue à la cuisson qualité gustative satisfaisante.
- KER PONDY : arrive sur le marche en octobre. Se conserve facilement jusqu'en mai-juin. Bonne tenue à la cuisson, chair assez fine.
- BF 15 : arrive sur le marché en juillet-août. Reste disponible jusqu'en mars. Très bonne tenue à la cuisson, ne se délite pas ; chair fine et ferme, bonne qualité gustative.
- ROSA : arrive sur le marché en septembre-octobre. Se conserve longtemps. Très bonne tenue à la cuisson, chair très fine et ferme, très bonne qualité gustative.
- BELLE DE FONTENAY : arrive sur le marché en juin, en primeur. On peut en trouver jusqu'à la fin mars. Très bonne tenue à la cuisson, ne se délite pas ; chair très fine et ferme, très bonne qualité gustative.
- APOLLO. Peut arriver sur le marche dès le mois de mai, en primeur. Comme tout primeur, est à consommer dans un délai d'une semaine. Bonne tenue de cuisson. Qualité gustative moyenne.

Le fichier de 78 variétés est constitué de la façon suivante :

- Variete : nom de la variété
- Gros : grosseur (1 : très petit, 3 : petit, 5 : moyen, 7 : gros, 9 : très gros)
- Rend : rendement commercial. Ce rendement est exprimé en % de celui de la variété BINTJE, sauf pour les variétés du type P4 où elle est exprimée en % de la variété ACKERSEGEN ; cette variable n'est donc pas homogène dans le fichier.
- Cons : aptitude à la conservation de 1 (très mauvaise) à 9 (très bonne)
- MS : teneur en matière sèche : 1 (très faible 16.5%), 3 : faible, 5 : moyenne, 7 : élevée, 9 : très élevée (27%)
- Cuis : tenue à la cuisson : 1 (moyenne), 2 : assez bonne, 3 (bonne), 4 (très bonne)

- Noir : noircissement après cuisson : 1 (nul), 2 (léger), 3 (léger à moyen), 4 (assez marqué), 5 (marqué)
- Type : 5 niveaux : Ferme (consommation à chair ferme), P1 (consommation précoces à 1/2 précoces), P2 (consommation 1/2 précoces à moyennes), P3 (consommation moyennes à 1/2 tardives), P4 (consommation 1/2 tardives à tardives)

Nom du fichier : PommedeTerre*txt (78*7)
Objectif(s) : établir des groupes de variétés et vérifier si le regroupement défini par le facteur type a un sens.
Origine des données : Institut technique de la pomme de terre (1977)
Méthode(s) statistique(s) possible(s) : Classification

Données : Les pays du Monde (1981)
Domaine d'application : Démographie
Thème et description des données : Ce fichier de données a été publié par le *Population Reference Bureau,* de Washington DC, sous le titre *World Population Data Sheet,* et repris dans *Population et Sociétés* (un état précédent avait été reproduit en juillet-août 1979 (n° 126). Par rapport au tableau de 1979, on a supprimé la colonne « mortalité infantile », qui était mal connue quand elle était élevée. On l'a remplacée par une colonne « indice synthétique de fécondité », qui n'est à vrai dire guère mieux connu, mais dont les estimations illustrent le thème du numéro de décembre 1980 (n° 142) à savoir qu'il y a beaucoup de pays (27 sur 90 de plus de 5 millions d'habitants) où la fécondité est inférieure à 2,5 enfants par femme, qu'il y en a beaucoup d'autres (46 sur 90) où elle est supérieure à 5 enfants par femme, mais qu'il y en a relativement peu (17 sur 90) dans la situation intermédiaire, qui correspond à la phase critique de la transition démographique.

La comparaison de ce tableau et de ces cartes à leurs états antérieurs ne doit pas être poussée trop loin. En effet, les modifications des chiffres traduisent plutôt les hésitations, remords et corrections des organismes statistiques internationaux que des phénomènes réels. Pour un pays aussi développé que les Etats-Unis, l'estimation de la population est passée de 222 à 226 millions d'habitants à la suite du recensement de 1980. C'est dire que la précision des chiffres reportés ici est très variable, et va de l'estimation vraisemblable à la conjecture. La population de nombreux pays africains reste très mal connue, le PNB des pays pauvres et des pays à économie socialiste est très conventionnel.

On souhaiterait d'ailleurs à ce propos que les organismes producteurs de ces chiffres, la Division de la Population des Nations Unies pour les statistiques démographiques et la Banque mondiale pour les statistiques économiques commentent avec quelque recul les principales modifications intervenues depuis cinq ou dix ans, pour faire la part des progrès de la connaissance et celle des phénomènes démographiques et économiques proprement dits.

Le fichier est ainsi constitué :

- Super : Superficie (milliers km2)
- P81 : Population Estimation 1981 (millions)
- Tauxnata : Taux de natalité (pour 1 000 habitants)
- Fecond : Indice synthétique de fécondité (enfants par femme)
- Morta : Taux de mortalité pour 1 000 habitants)
- Jeune : Population de moins de 15 ans (%)
- Vieux : Population de 65 ans ou plus (%)
- Est2000 : Projection 2 000 (millions)
- PNB : P.N.B. total 1979 (milliards de dollars)
- Pays : nom du pays
- Zone : avec les codes suivant :
  - AF (AFRIQUE), AFA (AFRIQUE AUSTRALE), AFE (AFRIQUE DE L'EST), AFN (AFRIQUE DU NORD), AFO (AFRIQUE DE L'OUEST), AFC (AFRIQUE CENTRALE)
  - AMCA (CARAIBES), AMCE (AMERIQUE CENTRALE), AMLA (AMERIQUE LATINE), AMSE (AMERIQUE SEPTENTRIONALE), AMTE (AMERIQUE DU SUD TEMPEREE), AMTR (AMERIQUE DU SUD TROPICALE)
  - AS (ASIE (sans URSS)), ASE (ASIE DE L'EST), ASS (ASIE DU SUD), ASSE (ASIE DU SUD-EST), ASSO (ASIE DU SUD-OUEST)

  - EU (EUROPE (sans l'URSS)), EUE (EUROPE DE L'EST), EUN (EUROPE DU NORD), EUO (EUROPE DE L'OUEST), EUS (EUROPE DU SUD)
  - OC (OCEANIE (SANS Hawai))
  - Monde (MONDE)
- Resum : 0 (si c'est une zone), 1 (pays proprement dit)

Nom du fichier : Monde81.txt (189*11)
Objectif(s) : essayer d'extraire une information de ce type de tableau officiel.
Origine des données : Populations et SOCIETES, Septembre 1981, n°150 (Institut National d'Etudes Démographiques, 27 rue du Commandeur, 75675 Paris Cedex 14)
Méthode(s) statistique(s) possible(s) : Classification

Données : Demande en céréales de pays importateurs
Domaine d'application : Economie
Thème et description des données : dans une étude entreprise sur la demande en céréales des 140 pays importateurs du monde, on a extrait un certain nombre de pays. On s'intéresse à une variable CONSO définie de la façon suivante :

CONSO = DISPONIBLE / POPULATION

avec DISPONIBLE = IMPORTATION + PRODUCTION - EXPORTATION

Les données sont fournies annuellement pour 17 années allant de 1969 à 1985, pour 22 pays dans un fichier constitué de la façon suivante :

- Pays : 22 pays
- An69, An70, An71, An72, An73, An74, An75, An76, An77, An78, An79, An80, An81, An82, An83, An84, An85 : valeur de CONSO (en kg/habitant)

Type : Cereale ou Ble
Nom du fichier : ConsoCere.txt (44*18)
Objectif(s) : étudier l'évolution de CONSO tant pour le blé que pour l'ensemble des céréales des 22 pays au cours des 17 années. Regrouper les pays qui ont des caractéristiques communes et, si possible, décrire l'évolution par un modèle simple CONSO = f{t}, où t représente l'année.
Origine des données : Bureau des Etudes Economiques de l'ONIC (Office National Interprofessionnel des Céréales, 1987).
Méthode(s) statistique(s) possible(s) : Classification

Données : Mortalité due à l'appareil respiratoire en Italie.
Domaine d'application : Epidémiologie
Thème et description des données : Dans une étude sur les causes de mortalité provenant de l'appareil respiratoire, on a relevé en 1971 dans les 18 régions italiennes 10 causes importantes possibles. On dispose de deux tableaux :

- l'un fournit le taux brut pour 100 000 habitants,
- l'autre le taux corrigé pour tenir compte de l'âge, toujours pour 100 000 habitants.

Le tableau de données contient l'ensemble des valeurs pour les deux taux et pour l'ensemble de l'Italie pour les taux bruts.

- Region : les 18 régions italiennes
- Tuber : Tuberculose
- Tumeur : Tumeur cancéreuse
- Coeur : Cardio-pulmonaire
- Bronchite : Bronchite aiguë
- Grippe : Grippe
- PneuVirale : Pneumonie virale
- Pneumonie : Autre pneumonie
- Emphyseme : Emphysème pulmonaire
- Abces : Abcès pulmonaire
- Autres : Autres causes
- Taux : Taux brut ou Taux corrige

Nom du fichier : Respiration.txt (37*12)
Objectif(s) : On demande d'établir une typologie des 18 régions et d'examiner les éventuelles différences entre les deux taux.
Origine des données : d' ALFONSO , G., d' AMBRA , L., LAURO, N . (1978) Determinanti ambientali e sociali nella distribuzione regionale delle cause di morte per malattie dell'apparato respiratorio alla luce di una analisi fattoriale. *Defesa Sociale* **2**, 39-56.
Méthode(s) statistique(s) possible(s) : Classification

Données : Lacs d'altitude
Domaine d'application : Hydrobiologie
Thème et description des données : Le patrimoine lacustre des massifs montagneux français comprend environ 3000 masses d'eau, parmi lesquelles 1650 sont des lacs. L'amélioration de la connaissance générale de ces lacs permet la mise en œuvre d'une stratégie de gestion. Cette étude a porté sur 30 lacs (Fichier Lac30) puis, plus complète, sur les 17 premiers (Lac17).

1) : Dans la première 17 descripteurs géographiques et morphométriques ont été relevés :

- ALT : Altitude du plan d'eau (m)
- AMX : Altitude maximum du bassin versant (m)
- AMY : Altitude moyenne du bassin versant (m)
- EXP : Exposition (N=1, NE=2, E=3, SE=4, S=5, S0=6, 0=7, N0=8)
- NEG : Neiges persistantes (%surface)
- Ad : Surface du bassin versant (ha)
- FF : Forêt feuillue (%surface)
- FC : Forêt de conifères (%surface)
- LA : Lande alpine (%surface)
- PL : Pelouse alpine (%surface)
- RN : Roches nues (%surface)
- LTH : Lithologie (7 : calcaire, 2 : gneiss, granite, migmatite, 3 : schiste non calcaire, 4 : schiste calcaire, 5 : grès non calcaire, 7 : calcaire)
- GEL : Durée du gel lacustre (mois)
- AO : Surface lacustre (ha)
- PER : Périmètre lacustre (m)
- ZM : Profondeur maximum (m)
- MNG : Marnage (m)
- LAC : nom du lac
- sig : sigle du lac (3 lettres)

2) Dans la seconde, deux groupes de variables ont été relevés :

2a) le premier concerne les matières organiques :

- MOU : Mousses présence (1)/absence (0)
- MAC : Macrophytes présence (1)/absence (0)
- GYM : Débris (gymnospermes) présence (1)/absence (0)
- VAS : Vase présence (1)/absence (0)
- SAB : Sable présence (1)/absence (0)
- GRA : Gravier présence (1)/absence (0)
- PIE : Pierres présence (1)/absence (0)
- MO : teneur en matière organique du substrat littoral (%)
- Mop : teneur en matière organique du substrat profond (%)

2b) le second concerne 13 caractéristiques physico-chimiques et la teneur en chlorophylle :

- ZDS : Transparence
- CDV : Conductivité
- P04 : Orthophosphate
- S04 : Sulfate
- NH4 : Ammoniaque
- N02 : Nitrite
- N03 : Nitrate
- Fer : Fer

- Si02 : Silicate
- TAC : Alcalinité
- TH : Dureté totale
- Ca : Calcium
- Mg : Magnésium
- Chla : Chlorophylle-a
- LAC : nom du lac
- sig : sigle du lac (3 lettres)

Nom du fichier : Lac30.txt (30*18), Lac17.txt (17*24)
Objectif(s) : étudier les relations entre les diverses variables et établir une typologie des lacs.
Origine des données : CHACORNAC J.M. (1986) *Les lacs d'altitude : métabolisme oligotrophe et approche typologique des écosystèmes*. Thèse Univ. Claude Bernard, Lyon, 248 p.
Méthode(s) statistique(s) possible(s) : Classification

Données : Analyse d'eaux minérales
Domaine d'application : Nutrition
Thème et description des données : le fichier a déjà été partiellement analysé dans l'ouvrage cité. Il a été complété par de nouvelles données récoltées au cours des voyages d'un des auteurs ; il est constitué de la façon suivante :

- HCO3 : bicarbonates (en mg/l)
- SO4 : sulfates (en mg/l)
- Cl : Chlore (en mg/l)
- Ca : Calcium (en mg/l)
- Mg : Magnesium (en mg/l)
- Na : Sodium (en mg/l)
- Origine : nom de l'eau minérale
- Sigle : trois caractères
- Pays : 1 (France), 2 (Italie), 3 (Espagne), 4 (Belgique), 5 (Hongrie), 6 (Portugal), 7 (Cuba), 8 (Tunisie)
- Nature : plat (eau plate), gaz (eau gazeuse)

Nom du fichier : `Eaux.txt (38*9)`
Objectif(s) : classer les eaux en fonction de la ressemblance de leur composition.
Origine des données : Tomassone, R., Dervin, C., Masson, J-P. (1993) *Biométrie, Modélisation de phénomènes biologiques.* Masson, paris (2ème éd.)
Méthode(s) statistique(s) possible(s) : Classification, ACP

Données : Populations d'abeilles du Nord-Ouest de l'Espagne
Domaine d'application : Zoologie
Thème et description des données : la race d'abeilles de la Péninsule Ibérique *(Apis mellifera iberica)* appartient au groupe des populations du nord-ouest de l'Europe, et elle représente un tronc d'évolution morphogénétique très différent de celui des populations du sud-est de l'Europe. Une étude a été faite sur des populations provenant de régions géographiques situées des deux côtés de la Cordillère Cantabrique, ainsi que sur cette Cordillère (Fig.1). 53 populations ont été étudiées ; chacune d'elles à partir de mesures sur 20 abeilles, ce sont les moyennes de ces mesures qui sont fournies et qui constituent donc les données de base.

16 variables prises sur les ailes ont été mesurées : 13 angles, une longueur, une largeur et un index cubital classique en apidologie. La signification de ces variables est fournie sur le diagramme standard d'une aile d'abeille (Fig.2). Le fichier est constitué de la manière suivante :

- A1, A4, B4, E9, G7, G18, H12, J10, J16, M17, N23, O26, Q21 : angles comme définis sur la figure 2
- LG : largeur
- A0 : longueur
- Icub : index cubital

Nom du fichier : AbeilleSp.txt (53*16)
Objectif(s) : établir une typologie des populations pour déterminer les éventuels écotypes..
Origine des données : Santiago E. et al. (1986) , *Apidologie* **17**,*2*,79-92.
Méthode(s) statistique(s) possible(s) : Classification

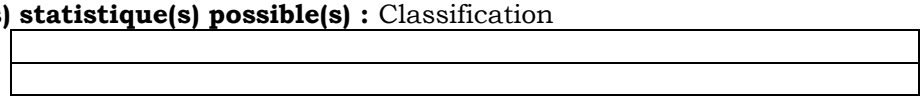

Figure : AbeilleSp-1.JPG (26Ko),AbeilleSp-2.JPG (20Ko)

Données : Attaque de tavelure sur le poirier
Domaine d'application : Agronomie
Thème et description des données : Une étude a été réalisée sur la sensibilité d'une collection de 89 variétés de poiriers à l'attaque de tavelure *(Venturia pirima L.).* A la récolte on a prélevé, deux années successives (1964 et 1965), un lot d'un cinquantaine de fruits sur chaque variété. Sur chaque fruit on a observé le niveau d'attaque quantifié de la façon suivante :

note	signification
0 :	aucune tâche
1 :	petites tâches, sans dépréciation du fruit
2,3,4 :	notes intermédiaires
5 :	fruits très attaqués

Chaque variété est donc caractérisée par une distribution de fréquences dans chacune des six classes pour les deux années. Le fichier se présente de la façon suivante :

- A64t0, A64t1, A64t2, A64t3, A64t4, A64t5 : distribution pour 1964
- A65t0, A65t1, A65t2, A65t3, A65t4, A65t5 : distribution pour 1965
- Var : numéro (1 à 89) de la variété

Nom du fichier : Tavelure.txt (89*13)
Objectif(s) : Les niveaux d'attaque sont-ils voisins pour les 89 variétés le deux années ? Peut-on regrouper les 89 variétés en groupes homogènes ?
Origine des données : M. Brian, INRA, Angers
Méthode(s) statistique(s) possible(s) : Classification, Anaca

Multi-tableaux

Données : Evolution démographique des départements français.
Domaine d'application : Démographie
Thème et description des données : on dispose pour l'ensemble des départements français métropolitains de recensements pas classe d'âge pour quatre années (1968, 1975, 1982 et 1990). Ces informations sont contenues dans trois fichiers, de structure quasiment identique :

- n : le code du département (1à 95)
- A5, A10, A15, A20, A25, A30, A35, A40, A45, A50, A55, A60, A65, A70, A75, A80, A85, A90, A95 : proportion (en millième dans la classe d'âge ≤ Ai)
- Popu : la population du département
- Departement : le nom du département
- An : l'année (1968, 1975, 1982, 1990)

En 1968 et en 1975, la valeur A95 n'existe pas. Le département 20 est la Corse ; en 1982 et 1990, la Corse est représentée par les deux départements 2A et 2B.

Nom du fichier : PopF68.txt (95*21), PopF75.txt (95*21), PopF82.txt (96*22), PopF90.txt (96*22)
Objectif(s) : étudier l'évolution des structures d'âge dans les département français au cours des quatre années.
Origine des données :
Méthode(s) statistique(s) possible(s) : Multi-tableaux

Données : Qualité de service de France-Télécom.
Domaine d'application : Marketing
Thème et description des données : Afin d'expliquer la qualité du service, FranceTélécom a mis en place les 7 indicateurs destinés à mesurer la qualité du service fourni

- IGQS : indice global (synthèse des indicateurs 2 à 8)
- IZAA : % appels efficaces à l'intérieur de la zone d'autonomie d'acheminement ou à la zone urbaine
- EZAA : % appels efficaces à l'extérieur de la zone d'autonomie d'acheminement (étranger non pris en considération)
- TSI : taux de signalisation des dérangements
- VR2 : vitesse de relevé des dérangements dans les 2 fours
- TCR : taux d'efficacité à l'arrivée des centres de renseignements ; % d'appels efficaces par rapport au nombre des appels au service de renseignement
- TCOM : temps d'établissement des communications (en secondes)

Le tableau de données est constitué par ces critères et par les trois années d'observation (An : 1984, 1985 et 1986), pour les 22 directions régionales françaises (Direction), dans l'ordre suivant :

"Direction" "IGQS" "IZAA" "EZAA" "TSI" "VR2" "TCR" "TCOM" "An"

En outre, des données technico-économiques de l'année 1984 pourrait permettre d'expliquer ces indicateurs :

- ACLP : accroissement brut du parc de lignes principales
- DLAI : délai moyen de raccordement (en mois)
- QRAC : indice de qualité de raccordement
- SIXM : demande en instance inférieure à six mois
- UNAN : demande en instance inférieure à un an
- DEM1 : délai de réalisation inférieur à 15 jours
- DEM2 : délai de réalisation inférieur à 30 jours
- TFPB : trafic financier des postes publiés

Nom du fichier : Telecom1.txt (66*9) : indicateurs
Telecom2.txt (22.8) : données technico-économiques
Objectif(s) : Etudier l'évolution de la qualité du service de France-Télécom, relation avec des indicateurs technico-économiques.
Origine des données : *Annuaire statistique des télécommunications*
Méthode(s) statistique(s) possible(s) : Multi-tableaux

Anova & Discrimination

Données : Résistance à l'avancement de versoirs de labour.
Domaine d'application : Agronomie
Thème et description des données : la résistance à l'avancement (c'est-à-dire l'effort de traction) de douze types de versoir a été étudiée en fonction de la vitesse de tracteur. Chaque mesure a été répétée quatre fois à l'occasion de deux essais. Le fichier est constitué de la manière suivante :

- Versoir : marque du versoir
- V1 : vitesse 1.10m/s
- V2 : vitesse 1.61m/s
- V3 : vitesse 2.08m/s
- V4 : vitesse 2.50/s
- Essai : I ou II

Les quatre répétitions des mesures sont placées en séquence. De plus 15 caractéristiques mécaniques des versoirs sont indiquées dans un second fichier (Labour1) :

- X1 : angle d'attaque (degré)
- X2 : longueur de la partie tranchante du soc (cm)
- X3 : angle du versoir avec le plan horizontal (degré)
- X4 : angle du versoir avec le fond du sillon (degré)
- X5 : angle moyen du bord supérieur avec le plan horizontal (degré)

- X6 : angle moyen du bord inférieur avec le plan horizontal (degré)
- X7 : longueur totale du versoir (cm)
- X8 : longueur du bord supérieur (cm)
- X9 : angle total de la pointe du soc à l'extérieur du versoir avec le plan horizontal (degré)
- X10 : angle total de la pointe du soc à l'extérieur du versoir avec le fond du sillon (degré)
- X11 : angle du trajet de la terre avec le sillon à vitesse basse (degré)
- X12 : déport latéral (cm)
- X13 : longueur totale de la pointe à l'extrémité du versoir (cm)
- X14 : hauteur du versoir (cm)
- X15 : surface du corps (soc + versoir) (cm2)

Nom du fichier : Labour.txt (96*6), Labour1.txt (12*15)
Objectif(s) : étudier les effets des deux facteurs (Versoir et Vitesse), décomposer les effets de la Vitesse et de l'interaction Vitesse*Versoir en ses trois composantes (linéaire, quadratique, cubique). Peut-on relier ces effets aux caractéristiques mécaniques des versoirs ?
Origine des données : M. Aubineau, chaire d'Agronomie de l'Institut National Agronomique.
Méthode(s) statistique(s) possible(s) : Anova

Données : Fertilisation de blé argentin
Domaine d'application : Agronomie
Thème et description des données : dans une expérience réalisée en Argentine sur la fertilisation du blé, on a étudié deux facteurs dans un dispositif expérimental de trois blocs complets. Le fichier se présente de la façon suivante :

- Fert : la fertilisation à 6 niveaux T (témoin), PC100 (superphosphate triple de calcium à 100 kg/ha), PA40 (phosphate d'ammonium à 40 kg/ha), PA100 (phosphate d'ammonium à 100 kg/ha), PU60 (PA100 + urée à 60 kg/ha), PU120 (PA100 + urée à 120 kg/ha)

- Densite : la densité des semis à 2 niveaux D99 (99 kg/ha), D42 (42 kg/ha)
- Bloc : B1, B2, B3
- PLM2 : nombre de plantes par m^2
- EPIM2 : nombre d'épis par m^2
- GRM2 : nombre de grains par m^2
- PMG : poids de 1000 grains
- EPIPL : nombre d'épis par plante
- GREPI : nombre de grains par épi
- RDT : le rendement mesuré, obtenu par récolte de toute la parcelle (de dimension 50*16 m) avec moissonneuse
  En outre, une variable supplémentaire, largement utilisée par les agronomes, peut être déduite des précédentes, RTHEO le rendement théorique calculé par :

$$RTHEO = GRM2 * PMG / 100$$

Nom du fichier : ArgentineBle.txt (36*10)
Objectif(s) :
Origine des données : chaire d'Agronomie de l'Institut National Agronomique
Méthode(s) statistique(s) possible(s) : Anova

Données : Variétés de maïs d'ensilage
Domaine d'application : Agronomie
Thème et description des données : dans une étude sur le maïs destiné à l'ensilage on souhaite comparer 56 variétés, dans un plan d'expérience à 2 blocs. Sept variables ont été mesurées et le fichier a la forme suivante :

- `Bloc : B1 ou B2`

- Ht : hauteur totale,
- He : hauteur de l'épi,
- Lf : longueur de feuille,
- Rf : rendement frais,

- Rs : rendement sec,
- Ms : matière sèche,
- Df : date de floraison (une valeur faible caractérise une plus grande précocité).

Il devrait y avoir 2*56 = 112 observations, mais la variété 26 n'est représentée que dans le bloc n°2. Avant toute analyse, il sera donc bon d'estimer les 7 valeurs manquantes pour avoir l'avantage d'analyser ensuite un dispositif équilibré.

On a déjà a priori les informations suivantes :

- les variétés 55 et 56 sont réputées pour avoir le meilleur rendement
- les variétés 53 et 54 sont connues pour être les plus précoces

Nom du fichier : EnsilageMais.txt (111*9)
Objectif(s) : classer les différentes variétés et choisir les meilleures variables pour le faire.
Origine des données : chaire d'Agronomie de l'Institut National Agronomique
Méthode(s) statistique(s) possible(s) : Anova

Données : Détermination d'acides aminés dans des protéines pures
Domaine d'application : Biochimie
Thème et description des données : Afin de comparer deux méthodes de détermination d'acides aminés dans des protéines pures, dans des aliments ou dans des fourrages, on a réalisé des mesures d'acides aminés essentiels (EAA : Cys, Phe, Ile, Leu, Lys, Met, Tyr, Thr, Val) et d'acides aminés non essentiels (NEAA : Asp, Glx, Ala, Arg, Gly, His, Pro, Ser). AA (=EAA+NEAA) représente l'ensemble des acides aminés. On a réalisé les déterminations soit avec la méthode classique d'hydrolyse acide à 110°C pour 24h., soit avec une nouvelle technique 145°C en 4h.

Les analyses ont été faites sur trois groupes d'échantillons :

- G1 : protéines pures (> 75%)
- G2 : protéines concentrées
- G3 : végétaux avec une concentration élevée de carbohydrate

Les trois fichiers sont constitués de la façon suivante avec en première colonne le nom de l'acide aminé, les deux méthodes d'hydrolyse 110°C et 145°C pour chaque échantillon, et en dernière colonne le nom du groupe (G1, G2, G3) :

- Pour G1, les protéines pures sont :
  - Amino Acide : nom en trois lettres de l'acide aminé
  - Ge110, Ge145 : gélatine
  - Oe110, Oe145 : albumine de l'œuf
  - Ca110, Ca145 : caséine
  - He110, He145 : hemisphaericine
  - Ly110, Ly145 : lysozyme
  - Groupe : G1
- Pour G2, protéines concentrées sont :

Whole egg T.irta E.breviflora Sesame seed Cotton seed

  - Amino Acide : comme G1
  - To110, To145 : œuf entier
  - Ti110, Ti145 : T.irta
  - Eb110, Eb145 : E. breviflora
  - Se110, Se145 : graine de sésame
  - Co110, Co145 : graine de coton
  - Groupe : G2
- Pour G3, les végétaux sont :
  - Amine Acide : comme G1
  - Ri110, Ri145 : riz
  - Bl110, Bl145 : blé
  - Bp110, Bp145 : B.purpurea
  - Ma110, Ma145 : maïs
  - Tp110, Tp145 : T.pavonia
  - Groupe : G3

Nom du fichier : Amino1(20*11), Amino2(20*11), Amino3(20*11)
Objectif(s) : comparer les deux méthodes de détermination des acides aminés.
Origine des données : LUCAS, B. et SOTELO, A. (1982) Amino-Acic Determination in pure proteins, foods and feeds using two different acide hydrolysis methods. *Analytical Biochemistry,* **133**, 349-356.
Méthode(s) statistique(s) possible(s) : Anova

Données : Modalités d'infestation du loup par les copépodes
Domaine d'application : Zoologie
Thème et description des données : dans une étude sur l'infestation du loup, poisson marin, par les copépodites de Caligus minimum on a réalisé une expérience à température constante (18°C) dans l'obscurité, et on a mesuré les positions atteintes par les parasites dans leur déplacement. Pour cela on a contrôlé deux facteurs :

- la durée de l'expérience, qui est le temps laissé aux parasites pour se déplacer. Ce facteur a six niveaux, les durées qui sont échelonnées de 10 mn à 60 mn ; ces niveaux sont numérotés de 1 à 6.
- le type de mucus

On dispose de 30 observations par combinaison "Durée*Mucus", réparties dans 12 classes correspondant aux positions atteintes par les parasites. Le fichier est constitué de la façon suivante :

- P1, P2, P3, P4, P5, P6, P7, P8, P9, P10, P11, P12 : position atteinte
- Duree : durée de l'expérience (1 à 6 ; unité 10 minutes)
- `Mucus : facteur à 7 niveaux avec la signification suivante :`
  - T : pas de mucus
  - EcaM : écaille de muge
  - EcaL : écaille de loup
  - Mlpa : mucus de loup en pastille
  - ML30 : mucus de loup lyophilisé à 30mg
  - ML40 : mucus de loup lyophilisé à 40mg
  - ML50 : mucus de loup lyophilisé à 50mg.

Nom du fichier : Caligus.txt (42*14)
Objectif(s) : étudier l'influence des deux facteurs contrôlés sur la répartition des copépodites. Pour cela on peut faire une analyse de variance sur les 30 moyennes des combinaisons "Durée*Mucus" ; on étudiera tout particulièrement les résidus les plus importants. De plus, une Analyse Factorielle des Correspondances effectuée sur ce tableau peut aussi donner une autre vision des résultats expérimentaux. On regardera la position sur les axes factoriels conservés des résidus importants.
Origine des données : RAIBAUT, A. (1985) Les cycles évolutifs des copépodes parasites et les modalités d'infestation. *Ann. Biol. XXIV***, 3**, 16.
Méthode(s) statistique(s) possible(s) : Anova, AFC

Données : Greffons et porte-greffes du pommier
Domaine d'application : Agronomie
Thème et description des données : Un pommier commercial est constitué de deux parties : la partie supérieure, le scion ou greffon, détermine les caractéristiques des fruits et des feuilles, alors que la partie inférieure, le porte-greffe *(rootstock)* détermine la taille et le développement de l'arbre. Au début du siècle, on pensait qu'un porte-greffe reproduit végétativement donnait un arbre nain, alors qu'avec une reproduction sexuelle il fournissait un grand arbre.

Avec l'amélioration des connaissances génétiques, cette hypothèse est apparue fort peu plausible. Une expérience a été faite à la station de recherche anglaise d'arboriculture à East Malling. Des porte-greffes d'origine européenne, numérotés I à IX, ont d'abord été étudiés ; on a ajouté ensuite d'autres porte-greffes d'origine allemande, numérotés de X à XVI. Dans l'expérience un même scion (Worcester Pearmain) a été greffé sur chaque porte-greffe. Les porte-greffe VIII, XI et XIV n'ayant pas fourni de données, seuls 13 porte-greffes sont donc analysés. Les variables sont fournies sous la forme suivante :

- PG : nom du porte-greffe
- n :

- Circ4 : circonférence du tronc au-dessus de la greffe *(trunk girth)* à 4 ans (en mm)
- Croi4 : croissance *(extension growth)* à 4 ans (en cm)
- Circ15 : circonférence du tronc à 15 ans (en mm)
- Poids15 : poids de l'arbre au dessus du sol à 15 ans (en livres)

Alors que la circonférence peut caractériser l'activité du cambium, les deux autres peuvent caractériser l'activité mérismatique à l'apex des scions.

Nom du fichier : Greffon.txt (104*6)

Objectif(s) : peut déceler des différences entre les porte-greffes ?

Origine des données : Pearce S.C. University of Kent at Canterbury, East Malling Research Station.

Méthode(s) statistique(s) possible(s) : Anova, Discrimination

Données : Existence de différences entre populations Pygmées, Saras et Touaregs à l'aide de caractéristiques sanguines

Domaine d'application : Biologie humaine

Thème et description des données : Au cours d'une étude faite en République Centrafricaine et au sud du Sahara, on a relevé 13 caractéristiques sanguines de Pygmées, de Saras et de Touaregs selon le sexe et l'âge. Le fichier est constitué par les moyennes d'un nombre inconnu de personnes (vraisemblablement de l'ordre de la dizaine) de la manière suivante :

- Li : lipoprotéines
- Ch : cholestérol
- Ur : urée
- Au : acide urique
- Pt : protides totaux
- Ag : albumine (globulines)
- A1 : albumine
- alpha1 : αl-globulines
- alpha2 : α2-globulines
- beta : β-globulines
- gamma : γ-globulines
- Vs : vitesse de sédimentation
- Ht : hématocrite
- Age : la classe d'âge à 8 niveaux (2 : moins de 3 ans, 5 : entre 3 et 8 ans, 10 : entre 9 et 12 ans, 15 : entre 13 et 17 ans, 20 : entre 18 et 22 ans, 27 : entre 23 et 30 ans, 35 : entre 31 et 40 ans, 40 : plus de 41 ans).
- Sexe : H pour les hommes, F pour les femmes.
- Pop : Pyg pour les Pygmées, Sar pour les Saras, Tou pour les Touaregs.

Le fichier est un tableau à 48 lignes et 16 colonnes.

Nom du fichier : Pygmees1.txt (48*16)

Objectif(s) : étudier s'il existe des différences entre les caractéristiques sanguines des trois populations vivant dans des conditions différentes, en tenant de l'évolution possible due à l'âge et du sexe.

Origine des données : Dr. Jaeger (CEABH, 1980) dans le cadre d'une mission financée par la DGRST (Délégation Générale à la Recherche Scientifique et Technique).

Méthode(s) statistique(s) possible(s) : elles sont multiples de l'Analyse de Variance classique à des méthodes d'analyse multidimensionnelle.

Données : Existence de différences entre populations Pygmées et Saras à l'aide de caractéristiques sanguines.

Domaine d'application : Biologie humaine

Thème et description des données : Au cours d'une étude faite en République Centrafricaine et au sud du Sahara, on a relevé la taille et le poids, la tension et le pouls de Pygmées et de Saras selon le sexe et l'âge. Le fichier est constitué par les moyennes de ces caractéristiques :

- Age : la classe d'âge à 8 niveaux (2 : moins de 3 ans, 5 : entre 3 et 8 ans, 10 : entre 9 et 12 ans, 15 : entre 13 et 17 ans, 20 : entre 18 et 22 ans, 27 : entre 23 et 30 ans, 35 : entre 31 et 40 ans, 40 : plus de 41 ans).

- Sexe : H pour les hommes, F pour les femmes.
- Pop : Pyg pour les Pygmées, Sar pour les Saras.
- Tmax : tension artérielle maximale (*10)
- Tmin : tension artérielle minimale (*10)
- Tdb : taille debout (cm)
- Tas : taille assis (cm)
- Pds : poids (kg)
- pdb : pouls debout
- pas : pouls assis

Le fichier est un tableau à 32 lignes et 10 colonnes. Ce fichier est complété par deux fichiers, celui des écarts-types (Pygmee2s) et celui du nombre de personnes (Pygmees2n)

Nom du fichier : Pygmees2m.txt (32*10), Pygmees2s.txt (32*10), Pygmees2n.txt (32*10)
Objectif(s) : étudier s'il existe des différences entre les deux populations entre la taille et le poids, la tension ou le pouls des personnes en fonction de l'âge, du sexe.
Origine des données : Dr. Jaeger (CEABH, 1980) dans le cadre d'une mission financée par la DGRST (Délégation Générale à la Recherche Scientifique et Technique).
Méthode(s) statistique(s) possible(s) : elles sont multiples de l'Analyse de Variance classique à des méthodes d'analyse multidimensionnelle.

Données : Existence de différences entre types de pommiers.
Domaine d'application : Arboriculture
Thème et description des données : quatre types parentaux de pommiers sont définis par deux facteurs croisés et correspondent à 4 groupes de huit arbres chacun ; sur chaque arbre on a mesuré trois caractéristiques des rameaux à un an puis à deux ans. Le fichier de 32 unités (les arbres) est constitué de la manière suivante :

- Infl1 : nombre d'inflorescences sur les rameaux de 1 an.
- Long1 : longueur (mm) des rameaux de 1 an.
- Diam1 : diamètre (mm) des rameaux de 1 an.
- Infl2 : nombre d'inflorescences sur les rameaux de 2 ans.
- Long2 : longueur (mm) des rameaux de 2 ans.
- Diam2 : diamètre (mm) des rameaux de 2 ans.
- Groupe : types parentaux correspondant à deux facteurs : Delicious (D) ou Golden (G) et Normal (N) ou Spur (S). Les types sont codés : DN, DS, GN, GS.

Le fichier est un tableau à 32 lignes et 7 colonnes. Il est complété par quatre arbres dont le groupe est inconnu (la valeur du groupe est notée NA) ; il faudra les reclasser dans le groupe le plus vraisemblable.

Nom du fichier : Pommier.txt (36*7)
Objectif(s) : Reconnaissance de type de pommiers
Origine des données : INRA
Méthode(s) statistique(s) possible(s) : Anova, Discrimination, Classement.

Données : Composition minérale et maladies du pommier
Domaine d'application : Agronomie
Thème et description des données : L'incidence de certaines maladies des pommes, comme le « *bitter pit* » peut être liée à certaines déficiences dans la composition minérale. L'expérience est constituée par 48 arbres soumis à quatre traitements, dans quatre blocs de quatre parcelles chacun. Chaque parcelle est formée par trois arbres. Six arbres n'ont pas porté de fruits. Le fichier est constitué de la façon suivante :

- TN : Azote total (ppm)
- PN : Azote assimilable (ppm)
- P : Phosphore (ppm)
- K : Potassium (ppm)
- Ca : Calcium (ppm)
- Mg : Magnesium (ppm)
- Poids : poids moyen des fruits (g)
- Bitter-pit : incidence du « *bitter pit* » (%)

- Traitement : T (Contrôle), Uree (Urée), NO3CaK (Nitrates de Calcium et de Potassium), SO4 (Ammonium et sulfate)
- Bloc : B1, B2, B3, B4

Nom du fichier : Bitterpit.txt (42*10)
Objectif(s) : étudier l'effet des trois traitements sur la composition minérale, le poids et l'incidence du « *bitter pit* ».
Origine des données : Ratkowsky,D.A. and Martin,D. (1974). The use of multivariate analysis in identifying relationships among disorder and mineral content in apples. *Aust.J.Agric.Res.* **25**, 783-790.
Méthode(s) statistique(s) possible(s) : Anova, GLM

Données : Calculs dans les urines
Domaine d'application : Biologie humaine
Thème et description des données : Pour étudier le lien qui peut exister entre la formation de calculs d'oxalate de calcium et certaines caractéristiques physiques, 79 échantillons d'urine ont été analysés : 45 provenaient de personnes sans calcul, 34 de personnes avec calculs. Six caractéristiques physiques ont été mesurées et le fichier est constitué de la façon suivante :

- Patient: numéro du patient
- Calculs: absence (sans) ou présence (avec)
- Densite: densité de l'urine par rapport à l'eau
- pH:
- mOsm: quantité proportionnelle à la concentration moléculaire en solution (*osmolarity = mOsm*)
- mMho: conductivité, mesure proportionnelle à la concentration d'ions chargés dans la solution
- Uree: urée, concentration en millimoles/litre
- Ca : calcium, concentration en millimoles/litre

Nom du fichier : UrineCalcul.txt (79*8)
Objectif(s) : Peut-on, avec ces observations, orienter un diagnostic médical ?
Origine des données : James S.Elliot M.D., Urology Section, Stanford University School of Medecine, Stanford, U.S.A.
Méthode(s) statistique(s) possible(s) : Discrimination

Données : Classement de 19 localisations de nématodes *Xiphinema elongatum.*.
Domaine d'application : Zoologie
Thème et description des données : Variabilité de l'espèce de nématode *Xiphinema elongatum.*

Les nématodes sont des vers vivant dans le sol ou en parasite de l'hommes et des mammifères ; ils jouent un rôle important en agriculture.

Dans une étude portant sur *Xiphinema elongatum*, on a mesuré sept caractéristiques classiques de la morphologie des nématodes correspondant au schéma ci-dessous :

Sigle	Caractéristique	Schéma
BOL :	longueur du corps	
DMA :	Coefficient de Man « a » = BOL/DV	
DMV :	Coefficient de Man « c » = 100*(AV/BOL)	
TAL :	Longueur de la queue	
TLC :	Coefficient C = TAL/DV	
ODS	Position du stylet (1)	
ODP	Position du stylet (2)	
Pays	Localisations (19)	
n°	Numéro d'échantillon	

Pour chaque localisation il existe plusieurs échantillons. En général une localisation est associée à un pays ; toutefois certains pays sont représentés par plusieurs localisations.

Nom du fichier : Nematodes.txt (222*9)
Objectif(s) : Classer 19 localisations de nématodes à l'aide de 7 mesures.
Origine des données : Michel Luc (Muséum National d'Histoire Naturelle & ORSTOM).
Méthode(s) statistique(s) possible(s) : a priori une méthode de discrimination paraît adaptée à l'analyse de ce corpus.
Figure : Nématodes.jpg (21Ko)

Données : Dimorphisme sexuel de la tortue peinte.
Domaine d'application : Zoologie
Thème et description des données : On souhaite étudier le dimorphisme sexuel chez la tortue peinte (*Chrysemis picta marginata*) à l'aide de trois mesures sur la carapace. Le fichier de données est constitué de la manière suivante :

- Long : longueur
- Larg : largeur
- Haut : hauteur
- Sexe : M ou F
- n : numéro de l'échantillon (A à X pour les deux sexes)

Il est recommandé de vérifier si une transformation logarithmique n'est pas utile.

Nom du fichier : Tortue (48*5)
Objectif(s) : étudier le dimorphisme sexuel.
Origine des données : Jolicoeur, P. & Mosimann, J.E. (1960) Size and shape variation in painted turtle. A principal component analysis. *Growth* **24**, 339-354.
Méthode(s) statistique(s) possible(s) : Discrimination

Données : Anatomie de squelettes de Kangourous
Domaine d'application : Zoologie
Thème et description des données : Des mesures ont été faites sur 18 caractéristiques du squelette de kangourous appartenant à trois espèces pour des mâles et des femelles :
Mg : *Macropus giganteus* (25 mâles, 25 femelles)
Mm :*Macropus fuliginosus melanops* (23 mâles, 25 femelles)
Mf :*Macropus fuliginosus fuliginosus* (25 mâles, 25 femelles)

On dispose en outre de mesures incomplètes sur trois spécimens « historiques » se trouvant dans des musées européens ; ainsi, celui de Paris a été capturé en 1803. Le fichier est constitué de la manière suivante :

- n : identification numérique de l'échantillon. Pour les trois spécimens, l'origine est : A (British Museum of Natural History, Londres, male, Mg), B (Muséum National d'Histoire Naturelle, Paris, male, Mg), C (Rijksmuseum van Natuurlijkee, Leiden, femelle, Mf)
- Sexe : M ou F
- Espece : Mg, Mm et Mf
- Caractéristiques : X1, X2, X3, X4, X5, X6, X7, X8, X9, X10, X11, X12, X13, X14, X15, X16, X17, X18. Toutes les mesures sont exprimées en millimètre multiplié par 10.

Certaines valeurs sont manquantes (notées NA)

Nom du fichier : Kangourou.txt (151*21)
Objectif(s) : discriminer les trois races et le sexe, vérifier les classement des trois spécimens historiques.
Origine des données : Poole, W.E. (1976) Breeding biology and current status of the grey kangaroo, Macropus *fuliginosus fuliginosus* , of Kangaroo lsland, South Australia. *Aust.J.Zool.* **24**, 169-187.
Méthode(s) statistique(s) possible(s) : Discrimination

Données : Mésofaune de hêtraie en forêt de Fontainebleau
Domaine d'application : Zoologie
Thème et description des données : Le rôle des microarthropodes dans la décomposition de la matière organique d'un sol et sur les mécanismes de cette décomposition sont importants. Ce constat a conduit à étudier les différents grands groupes faunistiques du sol dans des parcelles forestières de la forêt de Fontainebleau exploitées

différemment. Les observations correspondent à 35 échantillons, répartis en quatre types de parcelles

- celles du type NT (Gros Fouteau) ne sont exploitées que dans un but d'entretien
- celles du type FFT (a Tillaie) constituent une réserve biologique, dans laquelle aucune intervention humaine n'a eu lieu depuis le XVII^{ème} siècle
- celles du type SNA (Semis Naturel Assisté) ont été exploitées par coupes d'éclaircie en 1979, puis mise en régénération naturelle assistée avec labour de surface et application d'herbicide la première année, puis désherbage manuel les années suivantes
- celles du type SL (Semis en Ligne) ont été exploitées par coupes d'éclaircie en 1979, mais mise en régénération par semis en ligne avec labour de surface et désherbage chimique chaque année

Les échantillons sont constitués par des carottes de 5cm de diamètre dont on extrait la faune. Sept groupes sont pris en compte : les Oribates (acariens), les Acténides (acariens), les Gamasides (acariens), les Enchytracides (annélides), les Collembolles Symphypléones (insectes), les Collembolles Arthropléones (insectes) et un groupe indifférencié appelé Autres. Le fichier est constitué de la façon suivante :

- num : numéro d'échantillon (1 à 35)
- Exploit : 4 types de parcelles (NT, FFT, SNA, SL)
- Orib : dénombrement des Oribates
- Acti : dénombrement des Acténides
- Gama : dénombrement des Gamasides
- Enc : dénombrement des Enchytracides
- Symp : dénombrement des Collembolles Symphypléones
- Arth : dénombrement des Collembolles Arthropléones
- Autres : dénombrement des microarthropodes indifférenciés

Nom du fichier : Mesofaune.txt (35*9)
Objectif(s) : étudier le rôle du type d'exploitation sur la composition du sol en microarthropodes.
Origine des données : Massot C., Cancela da Fonseca (1986) *Rev.Ecot.BioZ.Sot* **23,***1,* 19-27.
Méthode(s) statistique(s) possible(s) : Discrimination

Données : Ombre commun dans le versant supérieur du Rhône
Domaine d'application : Zoologie
Thème et description des données : L'ombre (*Thymallus thymallus*) est un salmonidé autochtone dans les parties supérieures du bassin versant du Rhône, qui a fait l'objet d'introductions nombreuses de la part des pêcheurs. Les individus introduits sont d'origine assez diverses ; 5 populations, au sens géographique du terme. ont été étudiées. Le nombre de mesures porte sur un échantillon de 120 ombres, répartis selon les populations suivantes :

Population : nombre d'ombres

1. AinH (Haute rivière de l'Ain) : 41
2. Bienne (Bienne) : 18
3. Loue (Loue) : 20
4. AinB (Basse rivière de l'Ain) : 23
5. Loire (Haute Loire) 18

Dans un premier fichier (Ombre.txt) sept variables méristiques ont été mesurées et le fichier est constitué de la façon suivante :

- Rsimp : Nombre rayons simples nageoire dorsale
- Rrami : Nombre rayons ramifiés nageoire dorsale
- Rpect : Nombre rayons pectorale
- Rvent : Nombre rayons ventrale
- Ranal : Nombre rayons anale
- Ecail : Nombre d'écaille ligne latérale
- Verte : Nombre vertèbres
- Pop : AinB, AinH, Bienne, Loire, Loue
- n : 1 à 120

Dans un second fichier (OmbreMorpho), treize variables morphométriques ont été mesurées, elles correspondent au schéma de la figure (OmbreMorpho.jpg). Le fichier est constitué de la façon suivante :

- (1) Longtot : Longueur totale du corps
- (2) Ddorsale : Distance pré-dorsale
- (3) Longtete : Longueur de la tête
- (4) Danalcaud : Distance anale caudale
- (5) Horbit : Hauteur orbitale
- (6) Hoccip : Hauteur occipitale
- (7) DiamOeil : Diamètre de l'œil
- (8) Longmach : Longueur mâchoire
- (9) Longmaxil : Longueur maxillaire
- (10) Largmaxil : Largeur maxillaire
- (11) Hmaxi : Hauteur maximale du corps
- (12) Lnagcaud : Longueur nageoire caudale
- (13) Lnaganal : Longueur nageoire anale
- Pop :
- n : 1 à 120 puis s1 à s21 pour les échantillons supplémentaires à classer

En outre, on dispose d'un échantillon de 21 observations d'origines géographiques diverses (1-5 : Arve suisse, Aire et Alondon; 6, 19 et 20 : Lac Léman; 7 et 8 : Rhône genevois; 9 et 10 : Allondon; 11-13 : Rhône lyonnais; 14 : Tessin (Italie); 15 : Arve français; 16-18 : Guiers; 21 : Doubs).

Nom du fichier : Ombre.txt (120*9), OmbreMorpho.txt(141*14)

Objectif(s) : Etudier les cinq populations pour les deux types de variables, puis classer les 21 observations supplémentaires dans l'une des 5 populations à l'aide des variables morphométriques.

Origine des données : Persat H. (1978) Ecologie de l'ombre commun. *Bull. Fr. Piscic*, **206**, 11-20.

Méthode(s) statistique(s) possible(s) : Discrimination

Figure : OmbreMorpho.jpg

Données : Existence de différences entre provenances d'*épicéas* et relation avec leur localisation.

Domaine d'application : Foresterie

Thème et description des données : étude portant sur 33 provenances d'*épicéas de Sitka*, répartis sur 5 zones géographiques d'Amérique du Nord. Cinq mesures biologiques ont été faites sur quinze arbres par provenance dont les localisations géographiques ont été relevées. Les données sont les moyennes des mesures des quinze arbres. Le fichier est constitué de la manière suivante :

- Lieu : A : partie orientale de l'île de Vancouver et plaine côtière ; B : partie occidentale de l'île de Vancouver ; C : Îles de la reine Charlotte ; D : partie méridionale de l'Alaska ; E : bassins des rivières Stikine et Nass.
- Lat : latitude (en degré et 1/100 de degré)
- Long : longitude (en degré et 1/100 de degré)
- Alt : altitude (en pieds)
- Lonaig : longueur des aiguilles (mm)
- Laraig : largeur des aiguilles (mm)
- Longra : longueur des graines (mm)
- Largra : largeur des graines (mm)
- Loncon : longueur des cônes (mm)

Nom du fichier : Epicea.txt (33*9)

Objectif(s) : étudier la variabilité des cinq caractéristiques biologiques des 33 provenances, l'homogénéité des régions d'où elles proviennent, la relation avec les localisations géographiques.

Origine des données : *Sylvae Genetica* **27** (1)

Méthode(s) statistique(s) possible(s) : Discrimination, Anaca

Données : Comparaison crânes loup/chien.

Domaine d'application : Zoologie

Thème et description des données : les mensurations des crânes de loup et de chien sont relativement voisines. Le fichier est constitué de la manière suivante :

- n : identification
- Pop : Chien, Loup, ? (crâne fossile)
- LCB : longueur condylo-basale
- LSM : longueur de la mâchoire supérieure
- LBM : largeur bi-maxillaire
- LP : longueur de la carnassière supérieure
- LM : longueur de la première molaire supérieure
- LAM : largeur de la première molaire supérieure

Nom du fichier : Loup (43*8)
Objectif(s) : discriminer les loups des chiens à partir des six mensurations crâniennes. Le crâne fossile est-il un loup ou un chien ?
Origine des données : Cahier de l'Analyse des Données
Méthode(s) statistique(s) possible(s) : Discrimination, Classement
Figure : Loup.JPG (490Ko)

Données : Détection de l'obésité chez les hommes.
Domaine d'application : Biologie humaine
Thème et description des données : Dans une étude sur le poids d'hommes, une équipe de médecins a cherché à voir si celui-ci pouvait être relié à des résultats d'analyses d'urine. Pour y parvenir, 45 hommes avaient été regroupés en quatre groupes de poids.

Groupe	Poids moyen (kg)	Taille de l'échantillon
G4	119	12
G3	100	14
G2	65	11
G1	54	8

Le tableau de données a la structure suivante :

- pH
- Ccrea : coefficient de créatinine
- Pcrea : pigmentation de créatinine
- Phosphate : phosphate (mg/ml)
- Ca : calcium (mg/ml)
- Phosphore : phosphore (mg/ml)
- Crea : créatinine (mg/ml)
- Chlore : chlore (mg/ml)
- Bore : bore (µg/ml)
- Choline : choline (µg/ml)
- Cuivre : cuivre (µg/ml)
- Groupe : G1, G2, G3, G4

Nom du fichier : Obese.txt (45*12)
Objectif(s) : étudier les différences entre les quatre groupes pour chaque variable et pour l'ensemble des variables.
Origine des données : SMITH, H., GNANADESIKAN, R. & HUGHES, J.B. (1962). Multivariate Analysis of Variance (Manova). *Biometrics,* **18**(1), 22-41.
Méthode(s) statistique(s) possible(s) : Discrimination, MANOVA

Régression et chroniques

Données : Production porcine dans le Royaume Uni
Domaine d'application : Economie
Thème et description des données : une description de l'état de la production porcine du Royaume Uni est fournie par 5 indicateurs trimestriels de 1967 à 1978 fournis dans le fichier suivant :

- An : 1967 à 1978
- Trim : numéro du trimestre
- Truies : nombre de truies entrant en élevage
- Profit : index de profit
- Parite : rapport truie/verrat
- Boucher : nombre de porcs de boucherie rapporté au nombre de porcs d'élevage
- DimElev : dimension des élevages

Nom du fichier : PorcinGB.txt (48*7)
Objectif(s) : étudier la coévolution de ces 5 indicateurs
Origine des données : Ministère de l'Agriculture de GrandeBretagne, G.Tunnicliffe Wilson, University of Lancaster.
Méthode(s) statistique(s) possible(s) : Chronique

Données : Piégeages de lynx canadiens
Domaine d'application : Zoologie
Thème et description des données : le nombre de lynx capturés dans le district de la rivière MacKenzie au Canada a été utilisé pour illustrer un grand nombre de méthodes liées à l'analyse de séries chronologiques. On dispose de deux fichiers :

1) Le premier (LynxCap) fournit le nombre de lynx capturés par année de 1821 à 1932 dans le district de la rivière MacKenzie :

- An : l'année de 1821 à 1934
- Lynx : le nombre de lynx capturés

2) Le second (LynxPeaux) fournit le nombre de peaux de lynx vendues par la Compagnie de la Baie d'Hudson, ainsi que le prix payé, par peau, pour une période allant de 1857 à 1911. Ces valeurs sont largement représentatives des captures dans toute l'Amérique du Nord :

- An : l'année de 1857 à 1911
- Peaux : nombre de peaux vendues
- Shilling : le prix en shilling
- Pence : le complément en pence (deux erreurs sûres apparaissent dans les données en 1890 et en 1980)

Nom du fichier : LynxCap (114*2), LynxPeaux (55*4)
Objectif(s) : la série de rivière MacKenzie peut être analysée pour elle-même ; mais i1 peut aussi être intéressant de voir si les variations dans le district de la rivière MacKenzie sont voisines de celles des autres régions. Enfin peut-on trouver une relation entre le prix des peaux et la taille des captures, éventuellement avec un délai ?
Origine des données : Elton, C. and Nicholson,M. (1942) The ten-year cycle in numbers of the lynx in Canada. *J.Animal Ecology,* **11,** 215-244.
Méthode(s) statistique(s) possible(s) : Chronique

Données : Courses aux Jeux Olympiques
Domaine d'application : Divers
Thème et description des données : les temps des vainqueurs de courses (du 100 mètres au 800 mètres) sont connus de 1900 à 1976 ; ils sont présentés dans le fichier suivant :

- An : année de 1900 à 1976
- 100m : vitesse du vainqueur du 100m
- 200m : vitesse du vainqueur du 200m
- 400m : vitesse du vainqueur du 400m
- 800m : vitesse du vainqueur du 800m
- `Altitude : altitude de la ville d'accueil (en pieds)`

Nom du fichier : `JOCourses.txt (17*6)`

Objectif(s) : établir un modèle pour chaque course, étudier l'effet de l'altitude, prédire de futures performances.

Origine des données : Chatterjee, S., Chatterjee, S. (1982) New Lamps for Old: An Exploratory Analysis of Running Times in Olympic Games. *Appl. Statistics.* **31**(1), 14-22.

Méthode(s) statistique(s) possible(s) : Non Linéaire

Données : Comparaison de modèles d'administration du lithium dans le sang.

Domaine d'application : Biologie humaine

Thème et description des données : L'administration du lithium à des malades ayant des troubles psychiques peut être faite par deux traitements l'un sous forme de sulfate, l'autre sous forme de carbonate. Pour étudier l'élimination du Lithium, on fait des prises de sang à des malades ainsi soignés à des intervalles de temps de 2, 4, 6, 8, 12, 16, 24 et 48 heures après l'administration. Les deux traitements ont été appliqués à deux groupes de 14 patients chacun. Les données sont donc représentées par 28 courbes du taux de lithium en fonction du temps. Le fichier est constitué de la manière suivante :

- h2, h4, h6, h8, h12, h16, h24, h36, h48 : taux de lithium (en 100*meq/litre) aux différentes heures.
- Age : âge des patients
- Sexe : H (masculin) ou F (féminin)
- Traitement : Sulfate ou Carbonate

Nom du fichier : Lithium.txt

Objectif(s) : trouver un modèle plausible du taux de lithium en fonction du temps t ; comparer les deux traitements ; voir si l'âge et le sexe peuvent influencer le modèle.

Origine des données : RUTIGLIANO, G., CHIEPPA, M. & DE PALMA, B. (1973). Rate of decline of serum lithium concentration after single oral dose. Comparison between carbonate and slow release sulphate. *Folia Neuropsychiatrica* 16(1), 437-46.

Méthode(s) statistique(s) possible(s) :

Données : Facteurs influençant la possession d'une voiture

Domaine d'application : Economie

Thème et description des données : au cours de l'année 1977, l'OCDE a fourni des données pour tenter d'expliquer le nombre de voitures par personne dans différents pays du monde. Le fichier est constitué de la façon suivante :

- Pays : 24 pays
- Nvoiture : nombre de voiture par personne
- Pop : population du pays en millions d'habitants
- Densite : densité de la population
- Revenu : revenu moyen exprimé en milliers de $ US
- Prix : le prix du carburant exprimé en cents US
- Conso : consommation de carburant par voiture et par an (en T)
- `Commun : utilisateurs de transports en commun (train, autobus) en milliers de passagers-kilomètres`

Nom du fichier : `Voiture.txt (24*7)`

Objectif(s) : explication du nombre de voitures.

Origine des données : Sen, A. & Srivastava, M. (1990). *Regression Analysis ; Theory, Methode and Applications.* Springer Verlag, New York.

Méthode(s) statistique(s) possible(s) : Régression

Données : Difficultés pour s'assurer à Chicago

Domaine d'application : Marketing, Sociologie

Thème et description des données : Afin d'étudier les difficultés rencontrées dans certains quartiers de Chicago pour trouver une compagnie acceptant d'assurer des demandeurs, une étude a été faite par une commission. Elle a consisté, par quartier, à relever quelques indices, décrits dans le fichier de la manière suivante :

- Quartier : code des quartiers de Chicago (coir carte ci-dessous)
- Minor : % des minorités
- Arme : nombre d'armes pour 1000 unités d'habitation
- Vol : vols pour 1000 habitants

- Hab39 : % d'unités d'habitation construites en 1939 ou avant
- AssuVol : nouvelles assurances pour des propriétaires de logement plus les renouvellements, moins les suppressions et les non-renouvellements par 100 unités d'habitation
- AssuInvol : nouvelles assurances avantageuses pour des propriétaires de logement plus les renouvellements par 100 unités d'habitation
- Impot : valeur médiane de l'impôt familial

Nom du fichier : Chicago.txt (47*8)

Objectif(s) : analyser les raisons pour les assureurs d'accepter ou non les personnes vivant dans certains quartiers de la ville de Chicago.

Origine des données : US Commission on Civil Rights (1979) *Insurance Redlining : Fact not Fiction*. A report prepared by the Illinois, Indiana, Michigan, Minnesota, Ohio and Wisconsin Advisory Committees to the U.S. Commission on Civil Rights, Washington, D.C.

Méthode(s) statistique(s) possible(s) : Régression

Figure : Chicago.JPG (420Ko)

Code des 47 quartiers de Chicago

Données : Etude de la qualité de melons

Domaine d'application : Production végétale

Thème et description des données : Pour étudier la qualité de melons de Cavaillon, on a retenu l'indice de réfraction Y, comme indice de qualité. On souhaite relier Y à d'autres caractéristiques qui sont, contrairement à Y, des mesures non destructrices
En outre, on dispose de 4 variétés de melons.

- Variete : A, B, C, D
- Lot : numéro de l'échantillon dans la variété
- Poids : poids
- Diam : diamètre équatorial
- Hauteur : diamètre dans le sens de la hauteur
- Calice : diamètre à la cicatrice du calice
- Fermete : fermeté de la chair
- Indice : indice de réfraction (Y)

Nom du fichier : Melon.txt (32*8)

Objectif(s) : Peut-on prédire Y par un modèle le plus simple possible ? Ce modèle peut-il être commun aux quatre variétés ?

Origine des données :

Méthode(s) statistique(s) possible(s) : Régression

Données : Criminalité aux Etats-Unis

Domaine d'application : Sociologie

Thème et description des données : dans une enquête sur la criminalité, réalisée dans 47 états des Etats-Unis, on a relevé 14 critères ; le fichier de données est constitué de la manière suivante (toutes les valeurs des critères ont été multipliées par 10.) :

- R : le taux de criminalité, exprimé par le nombre d'actes de violence connus pour 1000000 d'habitants
- Âge : la proportion (%) d'hommes d'âge compris entre 14 et 24 ans
- S : vaut 1 si c'est un état du sud, et 0 sinon
- Ed : le taux de scolarisation, exprimé par le nombre d'années d'école (* 10) pour les personnes de moins de 25 ans
- Ex0 : la dépense (en $) par habitant en 1959
- Ex1 : la dépense (en $) par habitant en 1960
- LF : la proportion (%) d'hommes citadins, d'âge compris entre 14 et 24 ans, ayant un travail
- M : nombre d'hommes pour 1000 femmes
- N : population de l'état (en 100000 habitants)
- NW : proportion (%) de la population de race non blanche

- U1 : la proportion (%) d'hommes citadins, d'âge compris entre 14 et 24 ans, au chômage
- U2 : la proportion d'hommes citadins, d'âge compris entre 35 et 39 ans, au chômage
- W : la richesse mesurée par la valeur médiane des biens mobiliers, des actifs ou du revenu (unité 10$)
- X : inégalité de revenu exprimée par le nombre de familles (pour 1000) gagnant moins de la moitié du revenu médian

Nom du fichier : Criminalite.txt (47*14)
Objectif(s) : rechercher les critères qui peuvent le mieux expliquer le taux de criminalité R. Essayer de trouver le modèle le plus simple. Certains états ont-ils des caractéristiques particulières ?
Origine des données :
Méthode(s) statistique(s) possible(s) : Régression

Données : Revenus de PDG (Chicago)
Domaine d'application : Sociologie
Thème et description des données : afin d'étudier les facteurs pouvant influencer les salaires de deux années successives Y1984 et Y1983 de directeurs des 50 entreprises les plus importantes du district de Chicago, on a relevé un certain nombre de variables explicatives possibles. Le fichier est constitué de la façon suivante :

- R84 : salaires de 1984 (en $)
- R83 : salaires de 1983 (en $)
- Titres : nombre de titres possédés par le directeur
- Revenus : revenus totaux de la compagnie
- Taxes : taxes totales de la compagnie
- AgePDG : âge du directeur

Nom du fichier : RevenusPDG.txt (50*6)
Objectif(s) : explication du revenu des PDG.
Origine des données : Sen, A. & Srivastava, M. (1990) *Regression Analysis Theory, Methods and Applications*. Springer Verlag, New York, d'après les résultats provenant d'une enquête publiée le *13* mai *1985* dans le numéro du *Crain Chicago Business.*
Méthode(s) statistique(s) possible(s) : Régression

Données : Explication du nombre d'espèces et d'espèces endémiques des îles Galapagos en fonction de leurs caractéristiques géographiques.
Domaine d'application : Ecologie
Thème et description des données : Dans une étude faite sur un échantillon de 23 îles de l'archipel des Galapagos, on a relevé les 7 variables suivantes :

- Nespece : nombre total d'espèces différentes
- Nendem : nombre d'espèces endémiques
- Surf : superficie de l'île (en km^2)
- Alt : altitude (en m)
- Dmin : distance à l'île la plus proche (en km)
- DSCruz : distance à l'île de Santa-Cruz
- Surfvois : superficie de l'île la plus proche (en km^2)
- Ile : nom de l'île.

Nom du fichier : Galapagos.txt
Objectif(s) : comment expliquer par un modèle simple le nombre d'espèces et d'espèces endémiques d'une île en fonction de ses caractéristiques géographiques ?
Origine des données : JOHNSON, M . P . & RAVEN, P . H . (1972). Species Number Endemism : the Galapagos Archipelago Revisited. *Science* **179***,* 893-895.
Méthode(s) statistique(s) possible(s) : Régression, mesures d'influence
Figure : Galapagos.jpg (209Ko)

GLM

Données : Etude physico-chimique du lac du Brévent
Domaine d'application : Hydrobiologie
Thème et description des données : Au cours d'un travail sur le lac du Brévent (commune de Chamonix, Haute-Savoie), ont été analysées au cours d'une période allant du 30/03/83 au 18/07/84 un ensemble de caractéristiques physico-chimiques mesurées à l'aplomb du point le plus profond du lac à 9 profondeurs. Le fichier est constitué de la façon suivante :

- Prof : Profondeur de la mesure 9 valeurs en mètres (0.3, 2.5, 5.0, 7.5, 10.0, 12.5, 15.0, 17.5, 19.0)
- Temp : Température (°C)
- Cond : Conductivité (µS/cm à 20°C)
- pH :
- OxCo : Concentration en oxygène (mg/l)
- OxSa : Saturation en oxygène (%)
- Clib : Gaz carbonique libre (mg/l)
- Alca : Alcalinité (10^{-1} meq/l)
- Calc : Calcium (mg/l)
- Magn : Magnésium (mg/l)
- Ammo : Ammoniaque (µg-N/l)
- Niti : Nitrites (µg-N/l)
- Nita : Nitrates (µg-N/l)
- Sili : Silice (µg-N/l)
- Sulf : Sulfates (µg-N/l)
- Orph : Orthophosphates (µg-N/l)
- Date : T1 à T18 (respectivement 30/03/83, 08/06/83, 25/06/83, 08/07/83, 21/07/83, 01/08/83, 28/08/83, 24/09/83, 10/10/83, 24/10/83, 15/11/83, 15/12/83, 16/02/84, 15/03/84, 27/04/84, 06/06/84, 25/06/84, 18/07/84)

Certaines valeurs sont manquantes (notées NA) ; toutefois les valeurs manquantes de la profondeur 0.3 sont dues au fait que le lac était pris par les glaces. En outre, la saturation en oxygène (OxSa) ne peut être analysée conjointement aux autres variables.

Nom du fichier : LacBrevent (162*17)
Objectif(s) : étudier l'évolution temporelle et en profondeur.
Origine des données : CHACORNAC J.M. (1986) *Les lacs d'altitude : métabolisme oligotrophe et approche typologique des écosystèmes*. Thèse Univ. Claude Bernard, Lyon, 248 p.
Méthode(s) statistique(s) possible(s) : GLM
Figure : LacBrevent.JPG (66Ko)

Données : Démence ou dépression ?
Domaine d'application : Santé
Thème et description des données : afin d'établir un diagnostic entre la simple dépression ou l'état de démence on a relevé trois caractéristiques qui sont regroupées dans le fichier de la manière suivante :

- Gr : Groupe (Gr) : Dépression : A (95) ; Démence : B (96)
- Age : en années
- Sexe : 1 : homme ; 2 : Femme
- Ed : quatre niveaux d'éducation

Nom du fichier : DemenceDepression.txt (191*4)
Objectif(s) : établir un diagnostic
Origine des données : Everitt B.S. (1992) *The Analysis of Contingency Tables*. (2ed.) Chapman & Hill, Londres.
Méthode(s) statistique(s) possible(s) : GLM

Données : Prévalence de tumeur chez des rats
Domaine d'application : Santé

Thème et description des données : les auteurs de cette expérience sur des rats veulent comparer l'effet de différentes doses d'un traitement sur la prévalence d'une tumeur non létale dont la détection n'est possible qu'après la mort de l'animal. Comme la tumeur n'est pas létale, l'âge de la mort n'est pas d'un intérêt direct pour la comparaison des doses pour le développement de la tumeur ; néanmoins, la probabilité de ce développement peut être liée à l'âge. Le fichier est constitué de la façon suivante :

- Dose : le traitement est du biphénil polybromuré à 6 niveaux d'une échelle logarithmique (0 : 0 ; 1 : 0.1 ; 2 (0.3) ; 3 (1.0) ; 4 (3.0) ; 5 (10.0)
- Sexe : F (femelle) ou M (mâle)
- Poids : poids initial de l'animal (g)
- Etage : numéro (1 à 5) de l'étage dans la cage où se trouve l'animal
- Survie : l'âge à la mort de l'animal (semaine)
- Hyperplasie : présence (1) ou absence (0)

Chaque rat a été observé jusqu'à sa mort naturelle, ou intentionnellement sacrifié à la fin de la durée de l'étude (124 semaines)

Nom du fichier : TumeurRat.txt (319*6)

Objectif(s) : étudier l'effet de la dose, du sexe, de leur possible interaction. S'assurer que le poids et l'étage sont sans effet.

Origine des données : Dinse, G.E., Lagakos, S.W. (1983) Regression Analysis of Tumour Prevalence Data. *Appl. Statist.* **32**(3), 236-248.

Méthode(s) statistique(s) possible(s) : GLM

Données : Traitement des mammites

Domaine d'application : Zootechnie

Thème et description des données : Pour étudier comment traiter les mammites, des vaches ont été soumises à huit traitements ; elles appartiennent à seize élevages. L'état d'infection de 1a. mamelle a été relevé sur chacun des quatre quarts de la mamelle avant et après traitement. On aussi noté la qualité sanitaire de l'élevage et le type de traite. On dispose des données suivantes :

- Vache : identification de la vache
- Elevage : identification de l'élevage (1 à 16)
- Traitement :
  - T : aucun traitement
  - P1 : Pénicilline à 100000 unités
  - P2 : Pénicilline à 200000 unités
  - P4 : Pénicilline à 400000 unités
  - N4 : 400mg de Novobiocine
  - N6 : 600mg de Novobiocine
  - P2N4 : Pénicilline à 200000 unités + 400mg de Novobiocine
  - P4N4 : Pénicilline à 400000 unités + 400mg de Novobiocine
  - P1N4 : Pénicilline à 100000 unités + 400mg de Novobiocine
- Q11, Q21, Q31, Q41, Q12, Q22, Q32, Q42 : l'état d'infection (présence d'un organisme) du quartier de pis selon la position sur le pis codé de la façon suivante :
  - 0 : aucune infection
  - 1 : staphylococcus aureus
  - 2 : staphylococcus epidermis
  - 3 : streptococcus agalactiae
  - 6 : coliforme
  - 7 : autres

  La position sur le pis est codée de la façon suivante :

Position sur le pis	gauche		droite	
	avant	arrière	avant	arrière
avant traitement	Qgav1	Qgar1	Qdav1	Qdar1
après traitement	Qgav2	Qgar2	Qdav2	Qdar2

- Sanitaire : qualité sanitaire de l'élevage à quatre niveaux (mauvais, acceptable, bonne, excellent)

- Traite : le type de traite à de deux niveaux à l'attache dans l'étable (Attache) et en salle de traite (Salle)

Nom du fichier : Mammite.txt (274*13)

Objectif(s) : rechercher la meilleure combinaison de médicaments pour avoir un nombre minimum de mammites

Origine des données : Koch, G.G., Grizzle, J.E., Semenya, K., Sen, P.K. (1978) *Statistical methods for evealution of mastitis treatment data*. Univrsity of North Carolina, Chapel Hill, Institute of Statistics, Mimeo series n°1156.

Méthode(s) statistique(s) possible(s) : GLM

Nouveaux jeux

Données : Les stations thermales françaises
Domaine d'application : Hydrobiologie
Thème et description des données : les stations thermales françaises sont classées selon neuf types de maladies qu'elles permettent de soigner :

- GYP : maladies gynécologiques
- NER : maladies du système nerveux
- DER : maladies dermatologiques
- REI : maladies des reins
- VRE : maladies des voies respiratoires
- RHU : maladies rhumatismales
- NUT : maladies liées à la nutrition
- DIG : maladies du système digestif
- CIR : maladies de la circulation
- Station : nom de la station thermale
- Dep : numéro du département
- Reg : 5 classes 1 : Sud-Ouest, Pyrénées, Languedoc ; 2 : Sud-Est, Rhône, Alpes, Provence ; 3 : Centre, Auvergne ; 4 : Est, Vosges, Jura ; 5 : Normandie, Région parisienne et autres.

Nom du fichier : CureThermale.txt

Objectif(s) : Exemple simple mais scolaire qui peut permettre de classer les différentes stations.
Origine des données : ?
Méthode(s) statistique(s) possible(s) : ACP, AFC, Classification
Figure : CureThermale.jpg

Données : Etude de rendement fromager
Domaine d'application : Industrie alimentaire
Thème et description des données : le rendement et la qualité d'un processus de fabrication d'un fromage est naturellement lié à la composition des laits qui permettent de le fabriquer. Les technologues ont défini un rendement fromager (RFESC) et ils souhaitent connaître l'influence de différentes caractéristiques du lait dans le cas de la fabrication de fromages à pâte pressée. Ces caractéristiques sont définies par trois types de mesures : a) celles que toute usine de fabrication utilise ; b) celles qui demandent de faire appel à un laboratoire spécialisé ; c) celles qui permettent de faire des mesures en continu dites rhéologiques. Le fichier est constitué de la façon suivante :

- A) mesures de routine (7) :
  - MAT : matière azotée totale (g/l)
  - CNE : concentration en caséines (g/l)
  - NPN : azote non protéique (g/l)
  - CAT : concentration en calcium total (mg/l)
  - CAS : concentration en calcium soluble (mg/l)
  - CAI : concentration en calcium ionique (mg/l)
  - ES : extrait sec du lait (%) standardisé à 25g MG/l
- B) mesures de laboratoire (6) :
  - Ktot : proportion de caséine K totale (%)
  - Astot : proportion de caséine α_s (%)
  - DMM : diamètre moyen des micelles (nm)
  - D10 : 10% des micelles ont un diamètre inférieur (nm)
  - D90 : 10% des micelles ont un diamètre supérieur (nm)
  - CIS : concentration en citrate soluble (mM)
- C) mesures rhéologiques :
  - TPR : temps de prise (mn)
  - VRG : vitesse de raffermissement du gel présure (mV/mn)
  - FMG : fermeté maximale du gel présure (mV)

- Enfin le rendement fromager :
  - RFESC

Nom du fichier : RdtFromage.txt (41*17)
Objectif(s) : bâtir un modèle pour prédire RFESC en fonction des mesures de type A ; voir si les mesures de type B ou C valent la peine d'être utilisées pour améliorer la qualité du modèle et, ultérieurement, devenir des mesures de routine. S'assurer que certains échantillons n'ont pas une trop grande influence sur les paramètres du modèle choisi.
Origine des données : Laboratoire de technologie laitière de l'INRA à Grignon dans le cadre d'un contrat ARILAIT (1985)
Méthode(s) statistique(s) possible(s) : Régression
Figure :

Données : Comparaison de variétés de betteraves sucrières
Domaine d'application : Agronomie
Thème et description des données : Afin d'établir un palmarès de variétés de betteraves sucrières, on a réalisé un plan d'expérience. Le plan retenu est un lattice carré (5 lignes et 5 colonnes) permettant de comparer 25 variétés avec 6 répliques. Les parcelles sont de 6 rangs de 12 mètres de long, mais pour éliminer d'éventuels effets de compétitions entre variétés seuls les 4 rangs centraux ont été conservés. Toutes les variétés sont soumises aux mêmes traitements culturaux. Cette expérience a été refaite de façon identique dans cinq lieux différents. Le fichier de données est constitué de la façon suivante :

- n : numéro de la parcelle (1 à 750)
- exp : lieu à 5 niveaux (Ex1 à Ex5)
- r : numéro de la réplique du lattice carré dans le lieu (1 à 6)
- l : numéro de la ligne du lattice carré dans le lieu (1 à 5)
- c : numéro de la colonne du lattice carré dans le lieu (1 à 5)
- tra : numéro de la variété (1 à 25)
- Npied : Peuplement (en milliers de pieds/hectare)
- Racine : Rendement brut des racines (t/ha)
- Sucre : Teneur en sucre (%)
- RdtFi : Rendement net ou financier, soit Npied * Prix
- Rdtext : Rendement en sucre extractible (t/ha)

En outre, on doit analyser la « Richesse en sucre », variable importante pour l'agronome qui est le produit de Sucre par Rdtext.

Nom du fichier : Betterave(750*11)
Objectif(s) :
Origine des données : établir un palmarès des variétés pour les critères Npied, RdtFi, pour le couple {RdtFi,Rdtext}.
Méthode(s) statistique(s) possible(s) : Anova, Discrimination
Figure :

Données : Comparaison de provenances de teck
Domaine d'application : Forêt
Thème et description des données : Le teck est largement utilisé dans les travaux de menuiserie. Les meilleurs arbres sont ceux qui ont le meilleur rendement en bois de placage: volume grand, fourche élevée, peu de bosses de faible grosseur.

Les provenances présentent souvent d'assez grandes variabilités, il est donc intéressant de les comparer. Une expérience portant sur 11 provenances a été réalisée dans un dispositif en blocs incomplets équilibré. Chaque unité expérimentale est un placeau (petite parcelle contenant 10 arbres); les 5 variables analysées sont donc les moyennes d'un placeau. Le fichier est constitué de la façon suivante :

- Hauteur : hauteur (dm)
- Circ : circonférence (dm)
- Bosses : nombre de bosses
- Fourche : hauteur de la fourche (dm)
- Grosseur : grosseur des bosses (dm)

- B : numéro du bloc (11 niveaux)
- T : numéro de la provenance (11 niveaux)

Le volume de l'arbre est une quantité importante dans toute analyse d'une production forestière, il se calcule par :

$$\text{Volume} = \text{Hauteur} * (\text{Circ})^2$$

Nom du fichier : Teck.txt (55*7)
Objectif(s) : comparer les onze provenances et rechercher celles qui ont simultanément le plus grand volume au dessous de la fourche et le moins de bosses de faible grosseur.
Origine des données : F. Caillez (Centre Technique Forestier Tropical)
Méthode(s) statistique(s) possible(s) : Anova, Discrimination
Figure :

Données : Action de fumures sur la production de betterave sucrière
Domaine d'application : Agronomie
Thème et description des données : La production de betterave sucrière est influencée par des apports de fumure minérale et de fumure organique; une expérience a été réalisée pour étudier l'influence de ces deux apports. Il s'agit d'une expérience en "criss-cross" en 4 blocs. L'unité expérimentale est constituée par les deux rangs centraux d'une parcelle de terrain de 4 rangs de 10 mètres chacun; le schéma du dispositif complet est le suivant :

Dimension									
10m	B2	N75 Avec	N200 Avec	N150 Avec		N150 Sans	N200 Sans	N75 Sans	B4
10m		N75 Sans	N200 Sans	N150 Sans		N150 Avec	N200 Avec	N75 Avec	
10m	B1	N150 Sans	N75 Sans	N200 sans		N75 Avec	N150 Avec	N200 Avec	B3
10m		N150 Avec	N75 Avec	N200 sans		N75 Sans	N150 Sans	N200 Sans	
# rangs :	1	2	2	2	4	2	2	2	1

Le fichier est constitué de la façon suivante :

- Forg : fumure organique fumier à deux niveaux : Avec (40T/ha avant le labour précédant la culture de betterave) ; Sans
- Fmin : fumure minérale (sous forme d'ammonitrate au semis) à trois niveaux N75 (75 kg N/ha), N150 (150 kg N/ha), N200 (200 kg N/ha)
- Bloc : 4 niveaux (B1, B2, B3, B4)
- Rep : numéro du rang
- Densite : densité du peuplement (nombre de milliers de racines/ha)
- Rdt : rendement en racines (T/ha)
- Saccharine : richesse en saccharine

Le rendement en sucre (en q/ha) est le produit Rdt * Saccharine.

Nom du fichier : FumureBetterave.txt (48*7)
Objectif(s) : prendre un modèle tenant compte des défauts de randomisation du dispositif. La densité de peuplement (X1, variable a priori indépendante des traitements) peut être considérée comme une covariable dans l'analyse de Rdt et du rendement en sucre.
Origine des données : Chaire d'Agronomie INA.
Méthode(s) statistique(s) possible(s) : Anova, Discrimination
Figure :

Données : Engraissement d'agneaux
Domaine d'application : Zootechnie
Thème et description des données : au cours d'une expérimentation précédant l'actuelle étude, on avait remarqué que les agneaux dont le poids à la naissance était faible avaient, au moment de l'abattage à 35 kg, un état d'engraissement supérieur. On a donc réalisé une nouvelle expérience, afin de déterminer si une limitation de l'alimentation au cours des trois premières semaines avait le même effet qu'un faible poids à la naissance. L'expérience porte sur deux agnelages : pour le premier, il n'y avait que des mâles; pour le second, des mâles et des femelles. Pour chaque agnelage deux facteurs sont croisés :

- le poids à la naissance à deux niveaux : gros (poids moyen de 4kg) et petit (poids moyen de 2.5kg)
- le niveau alimentaire au cours des trois premières semaines, à 2 niveaux : haut et bas (75% du niveau haut).

Le fichier est constitué de la façon suivante :

- Agnel : le numéro de l'agnelage (1 ou 2)
- Sexe : M ou F
- Lot : poids à la naissance par lot à deux niveaux p (petit) ou g (gros)
- Alim : alimentation début vie postnatale à deux niveaux ad (ad libitum) ou re (restreint)
- Pnais : Poids à la naissance (kg*10)
- Pabat : Poids vif à l'abattage (kg*10)
- Pvide : Poids vif vide (kg*100)
- Pcarc : Poids de la carcasse (kg*100)
- Ptoil : Poids de la toilette (g)
- Progn : Poids du gras de rognon (g)
- Larddos : Epaisseur du gras dorsal (en 1/10 mm)
- Longigot : Mensuration de carcasse, longueur du gigot (mm)
- Largigot : largeur du gigot (mm)
- Loncarc : longueur de la carcasse (mm)
- Larpoit : largeur de la poitrine (mm)
- Propoit : profondeur de la poitrine (mm)

Nom du fichier : Agneaux.txt (54*16)
Objectif(s) : étudier plus particulièrement l'effet des différents facteurs sur les différentes variables d'abattage. Retrouve-t-on les résultats de l'expérience précédente ? Peut-on dire que l'effet du poids à la naissance est similaire à l'effet d'une restriction alimentaire après la naissance ?

L'utilisation dans l'analyse, de la variable "poids à la naissance" a-t-elle un intérêt ? Le plan d'expérience utilisé est-il critiquable. Comment pourrait-on l'améliorer ?

Origine des données : Chaire de Zootechnie de l'INA.
Méthode(s) statistique(s) possible(s) : Anova, Discrimination
Figure :

Données : Etude de lignées de colza de printemps
Domaine d'application : Agronomie
Thème et description des données : 10 lignées de colza de printemps ont été étudiées dans un dispositif expérimental en blocs complets (4 blocs). Les 9 variables suivantes ont été relevées sur les 40 unités expérimentales. Le fichier est constitué de la façon suivante :

- Ligne : numéro de la lignée (L1 à L10)
- Bloc : numéro du bloc (B1 à B4)
- Vigueur : note de vigueur
- Floraison : nombre de jours de floraison après la variété la plus précoce
- Hauteur : hauteur
- Feuille : couverture foliaire
- Verse : note de verse
- Hmature : hauteur à maturité
- Rdt : rendement (q/ha)
- Psilique : poids de grains par silique
- P1000 : poids de 1000 grains.

Nom du fichier : Colza(40*11)
Objectif(s) : classer les lignées
Origine des données : Chaire d'Agronomie INA.
Méthode(s) statistique(s) possible(s) : Anova, Discrimination
Figure :

Données : Fertilisation de blé : une expérience argentine (1984)
Domaine d'application : Agronomie

Thème et description des données : dans une expérience de fertilisation de blé, on a étudié 2 facteurs dans un dispositif en blocs complets de 3 blocs :

- La fertilisation à 6 niveaux :
  1) Témoin
  2) SPT 100 (Superphosphate triple de Calcium - 100 kg/ha)
  3) PDA 40 (Phosphate d'Ammonium 40 kg)
  4) PDA 100 (Phosphate d'Ammonium 100 kg)
  5) PDA 100 U60 (Phosphate d'Ammonium 100 kg + 60 kg Urée/ha)
  6) PDA 100 U120 (Phosphate d'Ammonium 100 kg + 120 kg Urée/ha)
- La densité des semis à deux niveaux : 99kg/ha et 42 kg/ha

Le fichier est constitué de la façon suivante :

- Fert : niveau de fertilisation dans l'ordre ci-dessus (T, SCa100, PAa4, PAb, PAbU1, PAbU2)
- Dens : densité des semis (D42, D99)
- Bloc : numéro du bloc (B1, B2, B3)
- PLM2 : nombre de plantes par m^2
- EPIM2 : nombre d'épis par m^2
- GRM2 : nombre de grains par m^2
- P1000 : poids de 1000 grains
- EPL : nombre d'épis par plante
- GREPI : nombre de grains par épi
- Rdt : rendement obtenu par récolte de toute la parcelle (50 x 16m) avec moissonneuse

Une variable supplémentaire doit être introduite : le rendement théorique calculé par multiplication du nombre de grains par m2 et le poids d'un grain (GRM2*P1000/100) qu'on pourra comparer à Rdt.

Les valeurs des composantes de rendement sont les moyennes des 15 placettes (2 rangs x 1,34m) par parcelle.

Nom du fichier : FertilisationBle.txt (36*10)
Objectif(s) : rechercher les différences entre les combinaisons des deux niveaux de chaque facteur.
Origine des données : Chaire d'Agronomie INA
Méthode(s) statistique(s) possible(s) : Anova, Discrimination
Figure :