Analyse des données Licence Pro 2024-2025

Cours n°1- Les données en statistique

Florian Bayer

Objectifs du cours

Savoir caractériser et organiser les données est une étape essentielle de toute étude scientifique. Outre la méthode de représentation graphique, le type de données est très important en statistique :

les **méthodes** ne seront pas les mêmes pour

- **caractériser** le prénom le plus fréquemment données en 2019
- ou pour déterminer l'âge moyen de la population française.

Les objectifs de ce cours sont donc :

- de réviser le **vocabulaire** portant sur les données en cartographie et en statistique
- d'apprendre à **reconnaître** les différents types de données et de tableaux

L'information géographique

Le géographe a pour particularité de s'intéresser aux **lieux** auxquels sont rattachés les données.

- Une information non localisable a donc peu d'intérêt pour le géographe

Le cartographe utilise donc de l'**information géographique**, c'est-à-dire localisable dans l'espace :

- Par des coordonnées
- Par une appartenance à un lieu, à un maillage

Ces appartenances ont un intérêt si elles peuvent être **caractériser** :

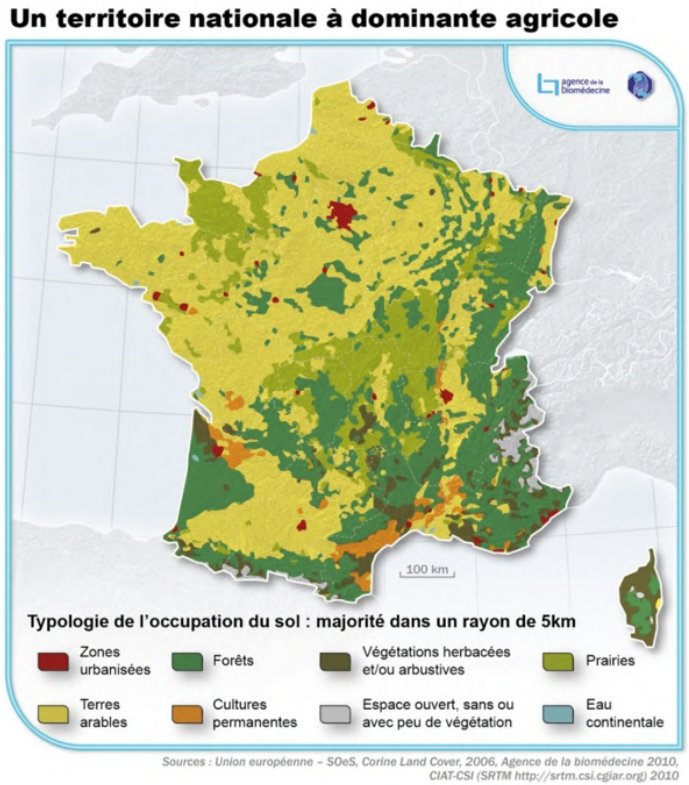
- À une commune peut être associée sa population totale ou sa densité

Si la collecte de l'information géographique peut parfois être laborieuse, son analyse peut se faire avec les mêmes outils qu'en statistique.

Les données en cartographie

Les données utilisées en cartographie et statistique proviennent de multiples sources (*recensement, sondage,images satellites etc.*) et peuvent être **caractérisées** (première partie du cours) :

Données qualitatives



Données quantitatives

Les données en cartographie

Les données peuvent être récupérées sous forme de tableaux (*i.e. INSEE*), ou bien issues de différentes sources (livres, articles, pixels).

Il est nécessaire de **restructurer** ces données, le plus souvent sous la forme d'un nouveau tableau. Les différents types de tableaux seront abordés dans la seconde partie du cours.

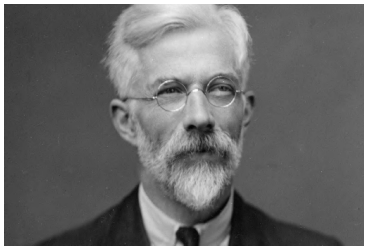
1- Les données en cartographie

Rappels de vocabulaire : statistique

Avant de définir les grands types de données, il est nécessaire de rappeler le vocabulaire commun aux données

La statistique : l'ensemble de techniques et d'outils mathématiques permettant d'analyser des données

« L'objet de la méthode statistique est la réduction des données. Une masse de données doit être remplacée par un petit nombre de quantités représentant correctement cette masse, et contenant autant que possible la totalité de l'information pertinente contenue dans les données d'origine. » - Sir Ronald Aymler Fisher



Les statistiques : les données textes ou chiffres (alphanumériques) décrivant une population, un ensemble

Rappels de vocabulaire : population

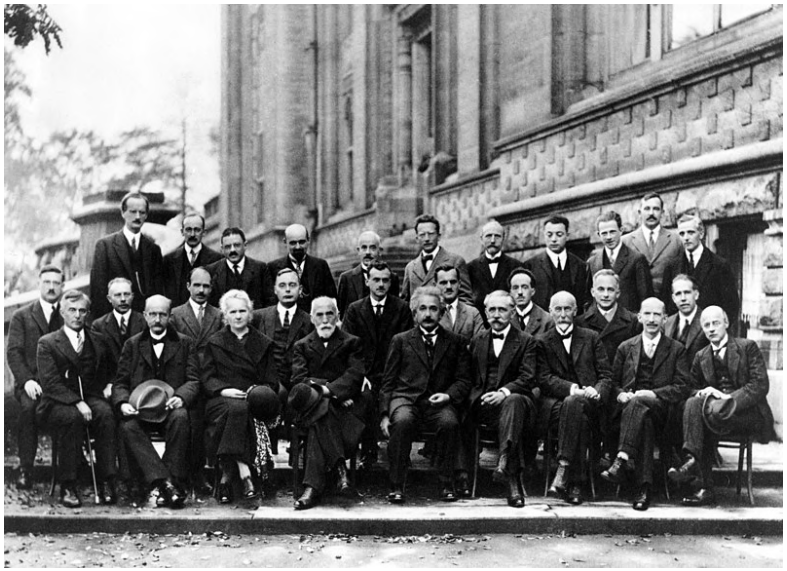
La **population** ou **l'ensemble** :

- La collection (l'ensemble) des données qui vont servir à créer votre carte.
- En géographie cet ensemble est très rarement infini
- On parlera souvent de série statistique pour les données quantitatives

---

Ensemble des unités statistiques étudiées : les pays européens

Mais la population peut aussi correspondre à des personnes :



Rappels de vocabulaire : élément

L'**élément** ou l'individu

- Un objet constitutif de l'ensemble

La Belgique est un élément de l'ensemble des pays européens

iso_a3	Nom	Continent
ALB	Albania	Europe
AUT	Austria	Europe
BEL	**Belgium**	**Europe**
BGR	Bulgaria	Europe
BIH	Bosnia and Herz.	Europe

Marie Curie est un individu de l'ensemble des participants du Congrès de Solvay

Rappels de vocabulaire : caractère

Le **caractère**

- Les éléments d'un ensemble sont décrits par un caractère.

Chaque pays peut-être caractérisé par son code, son nom, sa superficie, sa population

iso_a3	Nom	Continent	Superficie	Population

De même que des personnes (Nom, prénom, age, sexe, adresse etc.)

Nom	Prénom	Sexe	Adresse

Rappels de vocabulaire : modalité / valeur

La **modalité**, la **valeur**

- La valeur descriptive du caractère
  - modalité pour les données qualitatives
  - valeur pour les données quantitatives

La valeur de la population Belge est de 10,4 millions d'habitants. La modalité de son code iso est BEL

iso_a3	Nom	Continent	Superficie	Population
BEL	Belgium	Europe	30280 [km^2]	10414336

Marie Curie est une femme née en 1867. Elle a résidé au 36 quai de Béthune, 75004 Paris.

Nom	Prénom	Sexe	adresse
Curie	Marie	F	36 quai de Béthune, 75004 Paris

Données quantitatives et qualitatives

On peut caractériser les données en deux grands types, eux-mêmes disposant de sous-caractéristiques.

Les données **qualitatives** : caractérise la nature de ce qui est décrit et non la quantité.

- Un nom
- Une couleur
- Le type de sol

Les données **quantitatives** : Caractérise une quantité, par définition mesurable

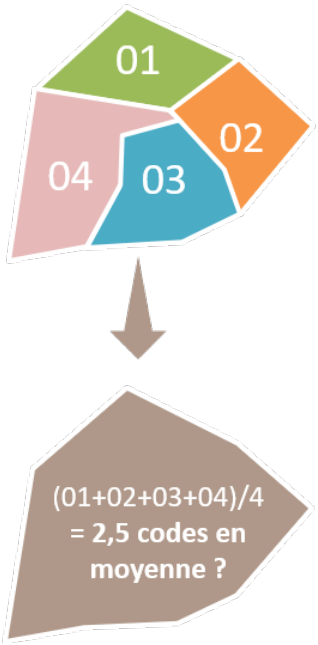
- Une population
- Un taux de chômage
- Une densité
- L'IDH

Quantitatives et qualitatives ?

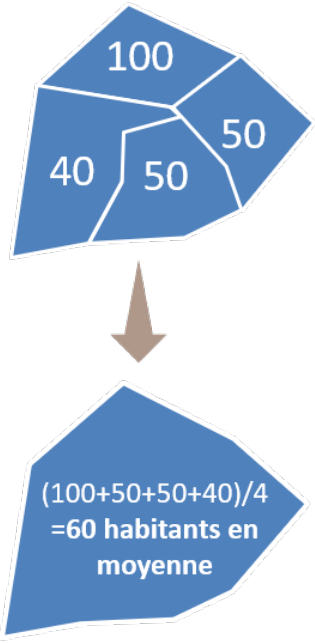
Il est important de pouvoir justifier le type de données :

- Si la moyenne est impossible ou absurde : qualitatif (code départementaux, numéro de téléphone)
- Si la moyenne a un sens : quantitatif (population, température)

Les **codes** de 4 communes :



La **population** de 4 communes :



Les sous types de données qualitatives

Les données qualitatives peuvent avoir d'autres propriétés, importantes en cartographie et en statistique

Qualitatif **nominal** : contient une notion de différence, aucun ordre

- Codes départementaux
- Des prénoms
- Des numéros étudiants *v.s.*

Qualitatif **ordinal**: contient une notion d'ordre sans être mesurable

- Une classification : grand > moyen > petit

Qualitatif **discret** : il y a moins de modalités que d'éléments.

- Le statut des hôpitaux : CHU, CHR, CH (3 statuts, 6 000 hôpitaux)
- Le statut des communes : Capitale, préfecture, sous préfecture

v.s.

Qualitatif **exhaustif** : il y a autant de modalités que d'éléments

- Le nom des pays, des régions
- Un code

Les sous types de données quantitatives

Les données **quantitatives** peuvent aussi avoir d'autres propriétés, toutes aussi importantes en cartographie et en statistique

Quantitatif de **stock** : une quantité brute,
un effectif.

- La population
- Une production en tonne

v.s

Quantitatif de **taux** : un rapport, un indice.

- La densité de population
- Le taux de chômage
- L'IDH

Stock ou taux ?

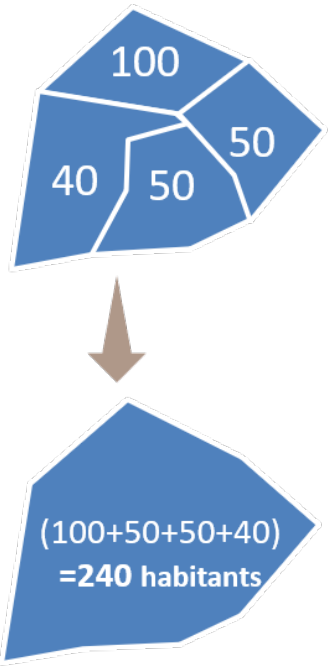
Comment faire la différence entre stock et taux ?

- Si la somme a une signification : stock (la somme de la population des pays du monde = la population mondiale)
- Si la somme n'a pas de sens : taux (la somme du taux de chômage des pays du monde ne correspond pas au taux de chômage mondial)

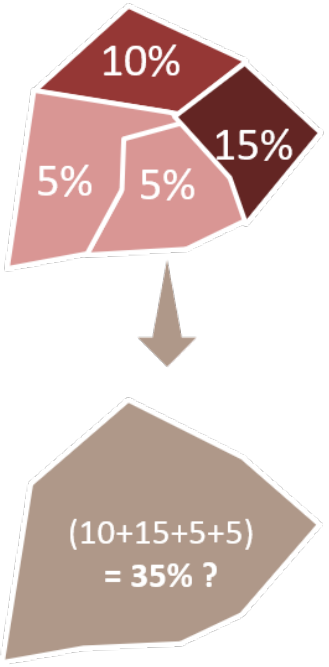
Attention, ce n'est pas parce que la valeur contient une virgule qu'il s'agit d'une données de taux :

- la France à une population de 66,6 millions d'habitants en 2016

La **population** de 4 communes :



Le **taux de chômage** de 4 communes :



Les sous types de données quantitatives

Quantitatif **repérable** : le zéro est conventionnel

- L'altitude

v.s.

Quantitatif **mesurable**: le zéro signifie l'absence concrète

- Le taux de chômage
- La population

Quantitatif **discret** : une variable discrète a un nombre fini ou dénombrable (qu'on peut compter) de valeurs possibles. Elles sont distinctes et séparées, aucune valeur intermédiaire n'est possible :

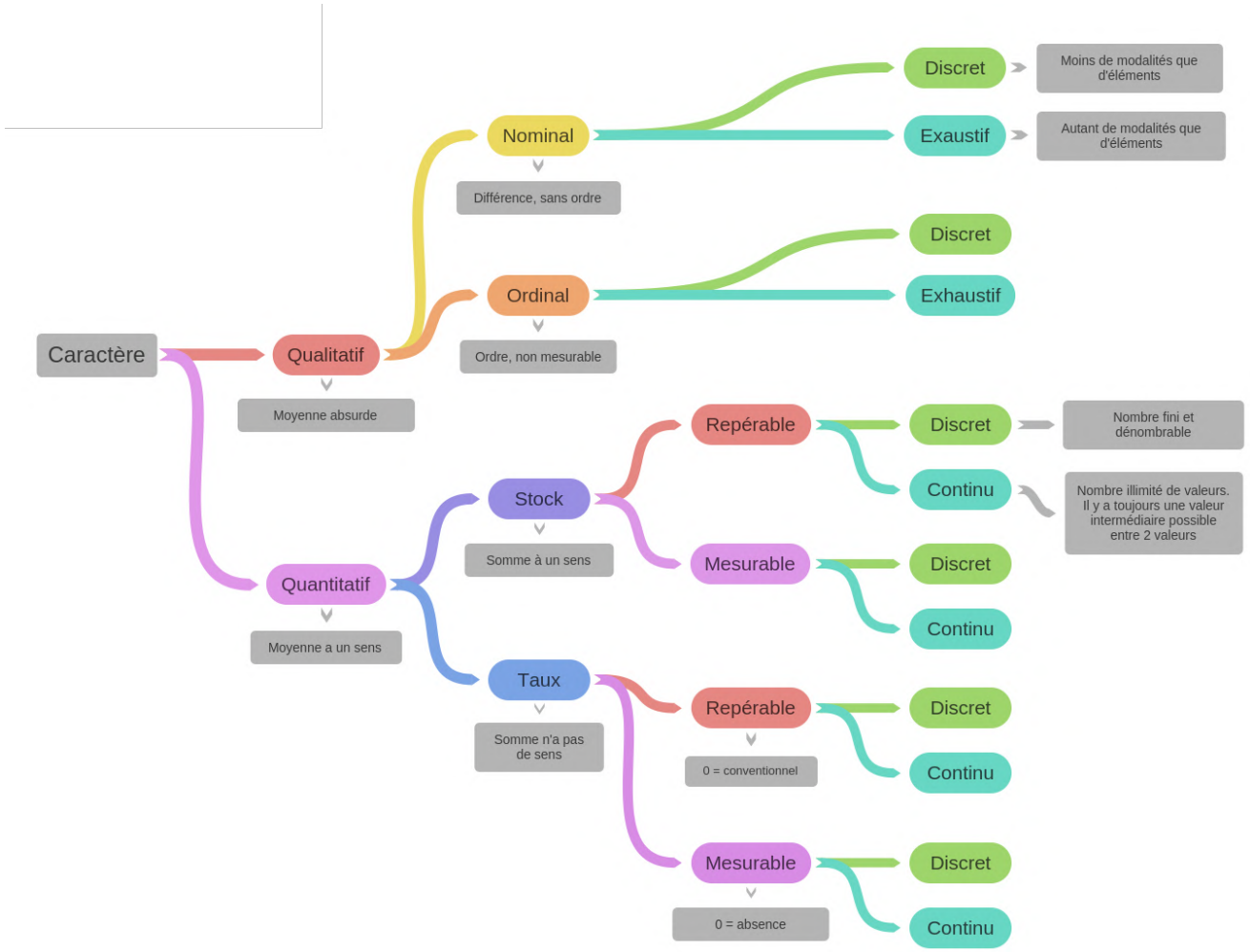
- Le nombre d'étudiants dans cette salle de cours à ce moment de la journée
- Les pointures de chaussures

v.s.

Quantitatif **continu**: il y a un nombre illimité de valeurs (non dénombrables). Entre deux valeurs distinctes, il y a toujours une valeur intermédiaire possible :

- Le taux de chômage
- La vitesse du vent

En résumé


Caractère
Qualitatif
Moyenne absurde
Nominal
Différence, sans ordre
Discret
Moins de modalités que d'éléments
Exaustif
Autant de modalités que d'éléments
Ordinal
Ordre, non mesurable
Discret
Exhaustif
Quantitatif
Moyenne a un sens
Stock
Somme à un sens
Repérable
Discret
Nombre fini et dénombrable
Continu
Nombre illimité de valeurs. Il y a toujours une valeur intermédiaire possible entre 2 valeurs
Mesurable
Discret
Continu
Taux
Somme n'a pas de sens
Repérable
0 = conventionnel
Discret
Continu
Mesurable
0 = absence
Discret
Continu

2- Les différents types de tableaux

Rappels

Il existe de nombreux type de tableaux en statistique. Leur forme peut dépendre entre autre:

- Du type de données en amont (comment ont-elles été recueillies ?)
- De la manière dont vous souhaitez analyser vos données (regroupements ?)
- La manière de mettre en forme les tableaux et les variables est un métier à part dans l'entreprise (data manager)

Tableau élémentaire de données

C'est le cas le plus courant. Il décrit un ensemble d'éléments (lignes du tableau) à l'aide d'un ensemble de caractères (colonnes du tableau).

- La première colonne est généralement réservée à un caractère servant d'identifiant.
- On note *i* un élément quelconque du tableau et *Xi* la " modalité prise par l'élément *i* pour le caractère *X*. "
- En géographie on parlera de **tableau d'information géographique**

Etudiant	groupe	UFR
A	1	GEO
B	1	GEO
C	2	GEO
D	2	HIST
E	3	HIST
F	3	GEO
G	3	HIST

Tableau de contingence

C'est un cas particulier de tableau élémentaire :

- les lignes et les colonnes jouent un rôle symétrique
- le contenu des cases correspond à des effectifs qui peuvent être sommés en ligne et en colonne.
- On peut parfois calculer des moyennes ou tout autre indicateur statistique si les données le permettent

Tout tableau de contingence est le résultat de la transformation d'un tableau élémentaire constitués de deux caractères discrets X et Y décrivant le même ensemble E

Le nombre de ligne d'un tableau de contingence (k) correspond au nombre de modalités du premier caractère discret (X) et le nombre de colonnes (p) correspond au nombre de modalités du second caractère discret (Y)

L'effectif d'une case, noté Nij, correspond au " nombre d'éléments du tableau élémentaire E qui prennent simultanément la modalité i de X et la modalité j de Y ".

Tableau de contingence

Tableau élémentaire de données

Etudiant	groupe	UFR
A	1	GEO
B	1	GEO
C	2	GEO
D	2	HIST
E	3	HIST
F	3	GEO
G	3	HIST

Transformé en tableau de contingence

UFR.x.Grp	GEO	HIST	Total
1	2	0	2
2	1	1	2
3	1	2	3
Total	4	3	7

Tableau disjonctif complet

C'est le résultat de l'éclatement d'un tableau élémentaire contenant des modalités

- Les variables sont codées en 0 ou 1 pour l'absence/présence d'un caractère
- On l'utilise pour certaines analyses spécifiques (analyses factorielles)
- Ils sont de retour en grâce avec le machine learning

Tableau élémentaire de données

Etudiant	groupe	UFR
A	1	GEO
B	1	GEO
C	2	GEO
D	2	HIST
E	3	HIST
F	3	GEO
G	3	HIST

Transformé en tableau disjonctif complet

	groupe.1	groupe.2	groupe.3	UFR.GEO	UFR.HIST
A	1	0	0	1	0
B	1	0	0	1	0
C	0	1	0	1	0
D	0	1	0	0	1
E	0	0	1	0	1
F	0	0	1	1	0
G	0	0	1	0	1

Tableau d’échanges

On parle aussi de matrice de flux

- Il contient des individus géographiques en ligne et en colonne, qui peuvent ou non être identiques
- Un tableau ne peut représenter qu'un seul caractère, par exemple les flux de population entre les individus
- Si le tableau d'échanges n'est pas symétrique, il se lit de la ligne vers la colonne

Aéroports	Paris	Berlin	Londres
Paris	-	10	5
Berlin	15	-	10
Londres	5	20	-

3- Les logiciels ?

Les logiciels en statistique

Il existe de **nombreux** outils informatiques pour réaliser des calculs statistiques et faire des graphiques.

Ils vous **aident dans les calculs** et permettent d'automatiser des tâches de manière efficace et **reproductible.**

Ils ne servent néanmoins à rien si vous ne connaissez pas les méthodes utilisées par ces logiciels :

- Réaliser une ACP avec Philcarto sans maitriser les principes de cette méthode va forcément aboutir à un résultat faux, malgré une jolie carte

Les logiciels en statistique

Pour éviter cet effet **boite noire**, nous verrons durant le semestre le fonctionnement et les principes de chaque méthode statistique, sans pour autant rentrer dans le détail des définitions mathématiques.

Retenez que votre plus value dans votre parcours universitaire **n'est pas de savoir appuyer sur un bouton** pour lancer un calcul, mais :

- de choisir la méthode la plus appropriée pour tester une hypothèse
- d'adapter vos outils au contexte de l'étude et du besoin initial
- de recueillir les résultats
- de les analyser et de les interpréter
- de les restituer de façon intelligible à votre public

Les diapositives suivantes listent quelques outils qui pourraient vous aider en statistique.

Les tableurs : Excel, Libre Office, Open Office

Ils permettent de mettre en page des données et de réaliser quelques analyses statistiques, essentiellement univariées (analyse d'une variable à la fois), mais pas seulement.

- Créer et déplacer des colonnes d'un tableau élémentaire de données
- Faire des calculs simple (Colonne C = Colonne A / Colonne B * 100)
- Faire des fonctions plus complexe (si, alors, jointure)
- Passer d'un tableau élémentaire à un tableau de contingence
- Faire des graphiques
- Quelques outils bivariées et multivariées
- Quelques tests statistiques

Ils sont très **communs** et **faciles à manipuler**, mais ne permettent pas de traiter de gros volumes de données (sans devenir des usines à gaz). Ils sont néanmoins très pratique pour mettre en page un petit volume de données.

Ils proposent souvent un langage de programmation pour **automatiser** certaines tâches, mais la reproductibilité d'un calcul par exemple n'est pas toujours évidente si vous utilisez uniquement le clic – bouton.

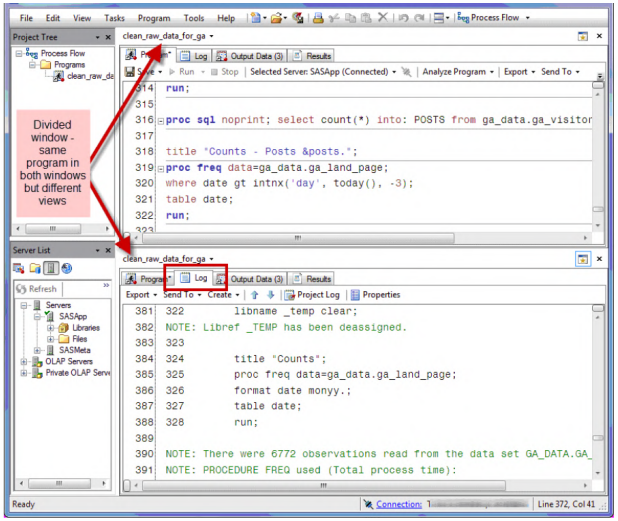
SAS

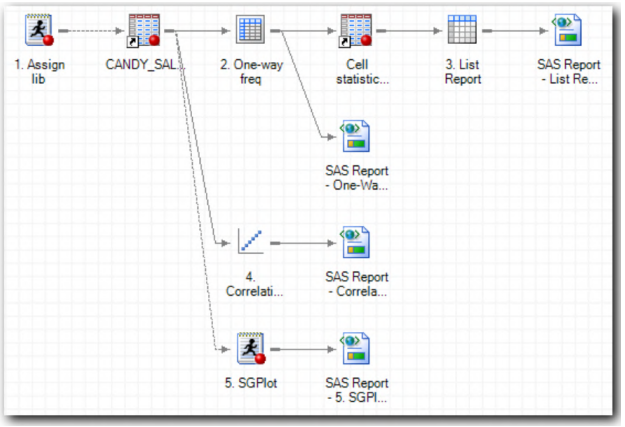
SAS est un logiciel très populaire en entreprise et dans le milieu scientifique pour réaliser des analyses statistiques.

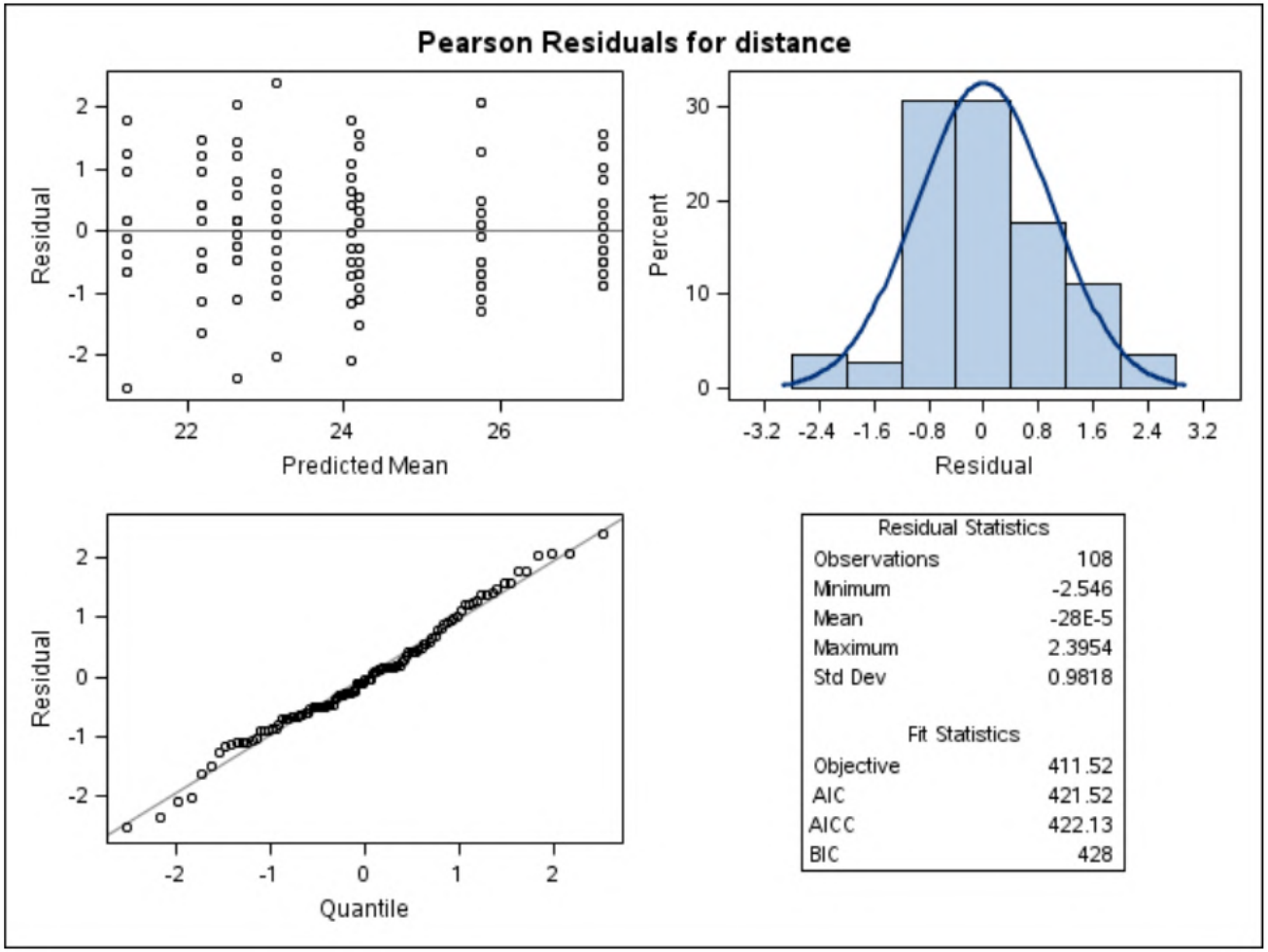
Outre des analyses statitiques, il permet de faire du **data management** de manière efficace et rapide (traitement des données avant les études, vérification de la qualité des données, imputation de données manquantes etc.).

Il est basé sur un langage qui lui est propre, mais loin des standards de 2020. Cela permet néanmoins de relancer facilement une analyse sur différent jeux de données. Une nouvelle version, **SAS Guide**, permet d'utiliser SAS sans maitriser les bases de la programmation informatique.

Les méthodes proposées dans SAS sont toujours validées par des scientifiques, mais parfois plusieurs années après leur publication.






Pearson Residuals for distance
Residual
Predicted Mean
Percent
Residual
Residual
Quantile
Residual Statistics
Observations 108
Minimum -2.546
Mean -28E-5
Maximum 2.3954
Std Dev 0.9818
Fit Statistics
Objective 411.52
AIC 421.52
AICC 422.13
BIC 428

R

R est un autre logiciel très populaire. Contrairement à SAS, il est gratuit et possède une communauté très développée, qui offre régulièrement de nouvelle fonctionnalité téléchargeable (**packages**).

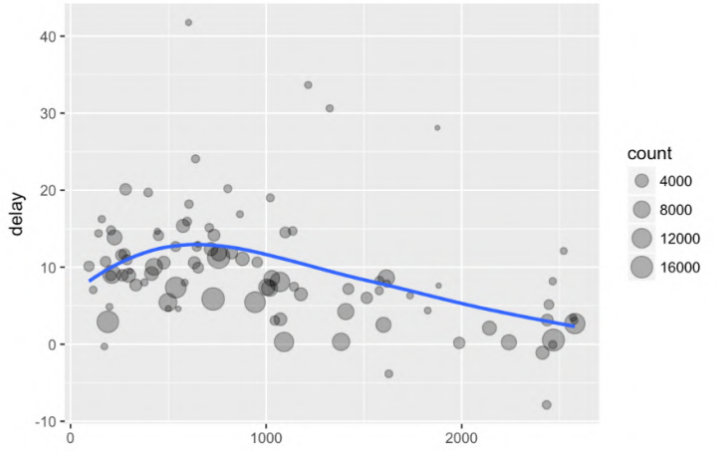
Cela peut poser problème lorsqu'un package que vous utilisez depuis longtemps n'est plus maintenu ou s'avère contenir une erreur (rare). Néanmoins pour une utilisation un peu plus avancée, il ne faut pas avoir peur de sonder des forums et installer, désinstaller, réinstaller des bibliothèques spéciques sur vos ordinateurs...

Utiliser R nécessite de "programmer". Ce langage à l'avantage d'être simple d'accès mais est une hérésie pour certains utilisateurs (dont votre enseignant aujourd'hui). A noter que vous pouvez faire du SIG, des cartes ou encore écrire votre mémoire avec R. Cette présentation est faite avec R.

Multiple steps

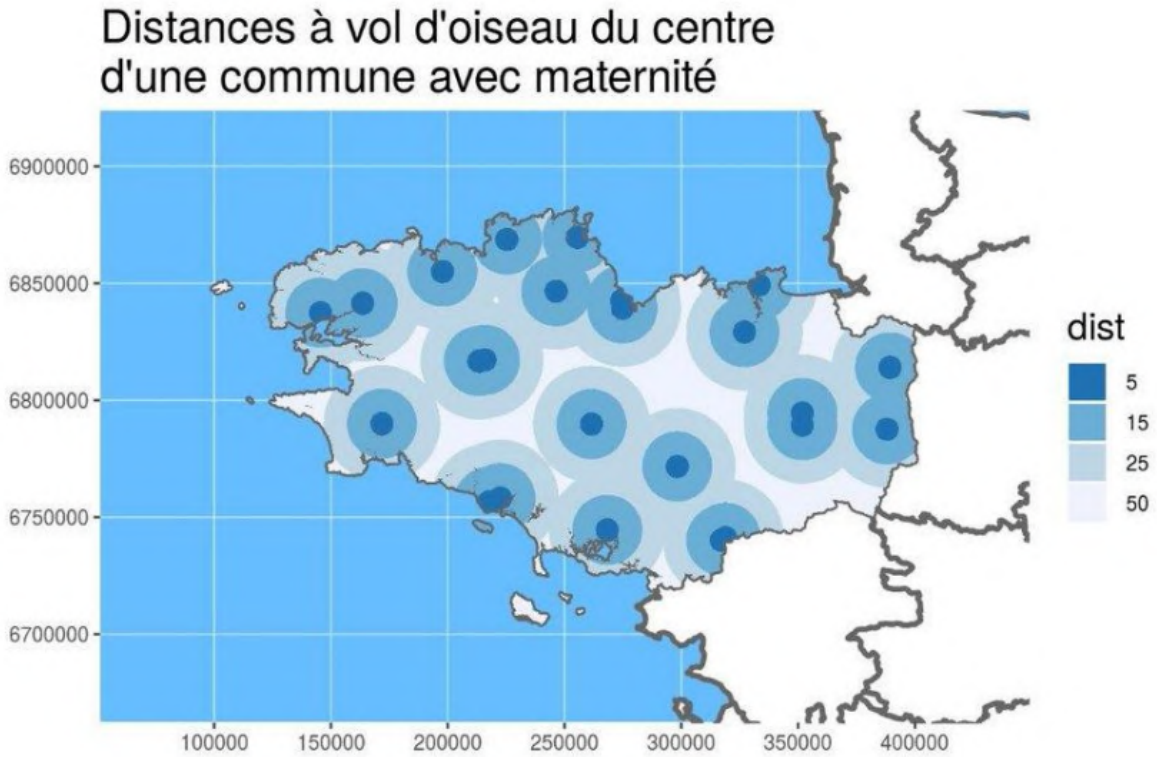
Imagine that we want to explore the relationship between the distance and average delay for each destination in `flights`. Using what you know about dplyr, you might write code like this:

```
by_dest <- group_by(flights, dest)
delay <- summarise(by_dest,
  count = n(),
  dist = mean(distance, na.rm = TRUE),
  delay = mean(arr_delay, na.rm = TRUE)
)
delay <- filter(delay, count > 20, dest != "HNL")

ggplot(data = delay, mapping = aes(x = dist, y = delay)) +
  geom_point(aes(size = count), alpha = 1/3) +
  geom_smooth(se = FALSE)
```



Maps with {ggplot2}

- `geom_sf` : recognizes geometry
- `coord_sf` : axis limits and CRS
- Several layers can be added by specifying `geom_sf(data = ...)`



src: @statnmap

Python

Il ne s'agit pas d'un logiciel de statistique en soit, mais d'un langage de programmation très populaire. Il est utilisé dans l'automatisation de tâches en informatique, pour la domotique, mais aussi en statistique. C'est un langage **facile à apprendre**, souvent enseigné au collège/lycée pour apprendre la programmation.

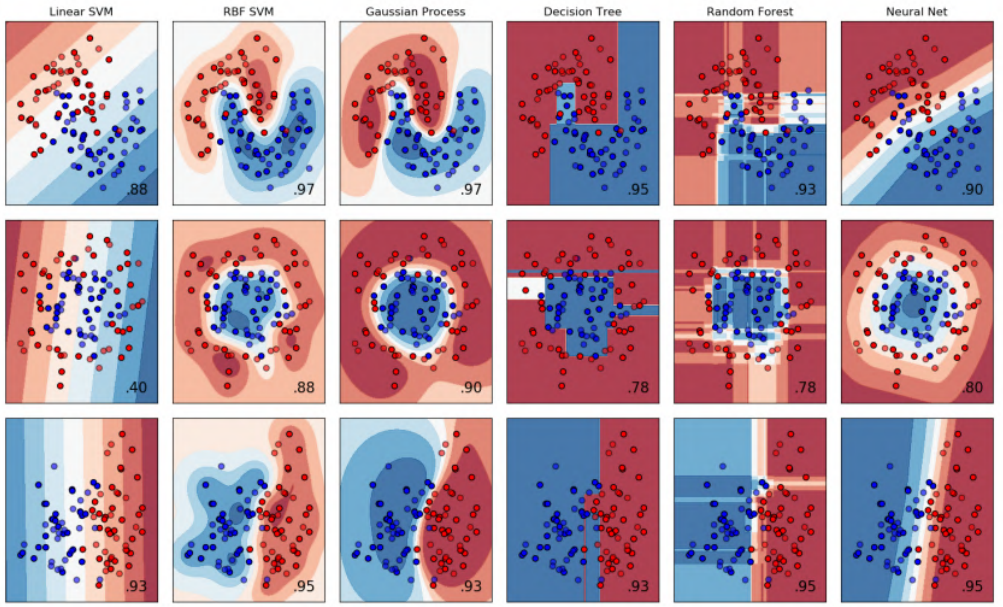
Comme R, des librairies peuvent être installées afin de réaliser certaines tâches spécifiques. Par exemple **numpy** pour les traitements mathématique, **matplotlib** pour les graphiques ou **pandas** pour le traitement de tableau de données. Mais certaines librairies, notamment cartographiques, fonctionnent très mal sous Windows ou Mac OS. Linux est l'ami de Python... Si vous voulez tester Python, installez Anaconda qui permet de gérer facilement les dépendances entre librairies. Il est même fourni avec des outils graphiques pour faire des statistiques (et sans coder) comme Orange 3.

C'est également un langage également très populaire en **machine learning**, grâce à la librairie scikitlearn.

A noter que ce cours d'analyse de données va vous apprendre la régression, l'analyse en composante principale et la classification, qui sont des outils majeurs en machine learning. Nous n'utiliserons toutefois pas Python pour les réaliser.

Python

```python
import numpy as np
import matplotlib.pyplot as plt
from matplotlib.colors import ListedColormap
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.datasets import make_moons, make_circles, make_classification
from sklearn.neural_network import MLPClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.svm import SVC
from sklearn.gaussian_process import GaussianProcessClassifier
from sklearn.gaussian_process.kernels import RBF
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier, AdaBoostClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.discriminant_analysis import QuadraticDiscriminantAnalysis

h = .02  # step size in the mesh

names = ["Nearest Neighbors", "Linear SVM", "RBF SVM", "Gaussian Process",
         "Decision Tree", "Random Forest", "Neural Net", "AdaBoost",
         "Naive Bayes", "QDA"]

classifiers = [
    KNeighborsClassifier(3),
    SVC(kernel="linear", C=0.025),
    SVC(gamma=2, C=1),
    GaussianProcessClassifier(1.0 * RBF(1.0)),
    DecisionTreeClassifier(max_depth=5),
    RandomForestClassifier(max_depth=5, n_estimators=10, max_features=1),
    MLPClassifier(alpha=1),
    AdaBoostClassifier(),
    GaussianNB(),
    QuadraticDiscriminantAnalysis()]

X, y = make_classification(n_features=2, n_redundant=0, n_informative=2,
                           random_state=1, n_clusters_per_class=1)
rng = np.random.RandomState(2)
X += 2 * rng.uniform(size=X.shape)
linearly_separable = (X, y)

datasets = [make_moons(noise=0.3, random_state=0),
            make_circles(noise=0.2, factor=0.5, random_state=1),
            linearly_separable
            ]
```



Pour ce cours ?

Nous utiliserons les outils suivants :

- Orange pour l'analyse de donénes
- Magrit pour la cartographie

Conclusion

Concepts-clés

Les géographes utilisent de **l'information géographique**, localisable dans l'espace.

- Il existe un vocabulaire propre aux données statistiques, qu'il est nécessaire de connaître.
- Les données peuvent être regroupées selon plusieurs **propriétés**
- Qu'il faut maîtriser, car les règles de la sémiologie graphique et les outils statistiques en **dépendent** !

Plusieurs types de tableaux existent et il est nécessaire de les reconnaître

- Les tableaux élémentaires de données sont les plus courants, avec en ligne les individus et en colonnes leurs caractères
- Les tableaux de contingence permettent de croiser des caractères
- Les tableaux d'échanges sont très appréciées des géographes