

Analyse des données Licence Pro 2025-2026

Cours n°1- Les données en statistique

Florian Bayer

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

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 :

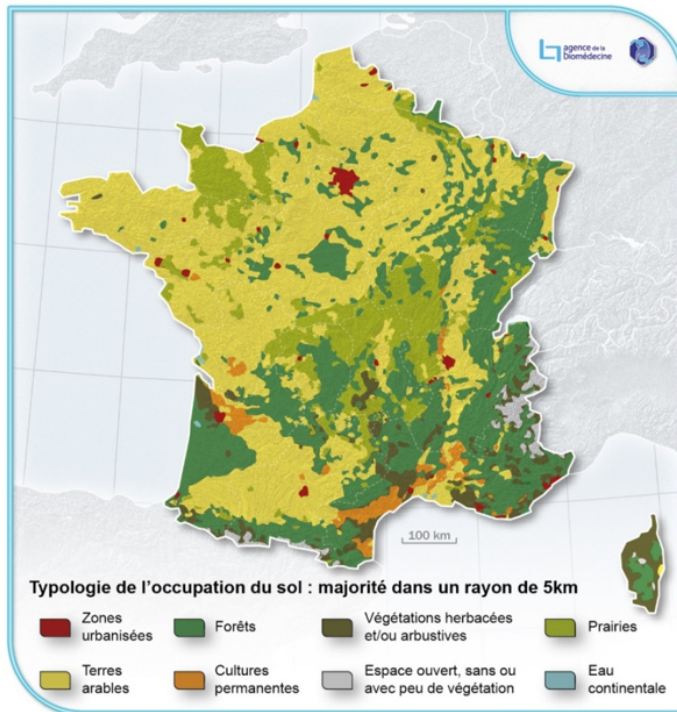
- À une commune peut être associée sa population totale ou sa densité

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

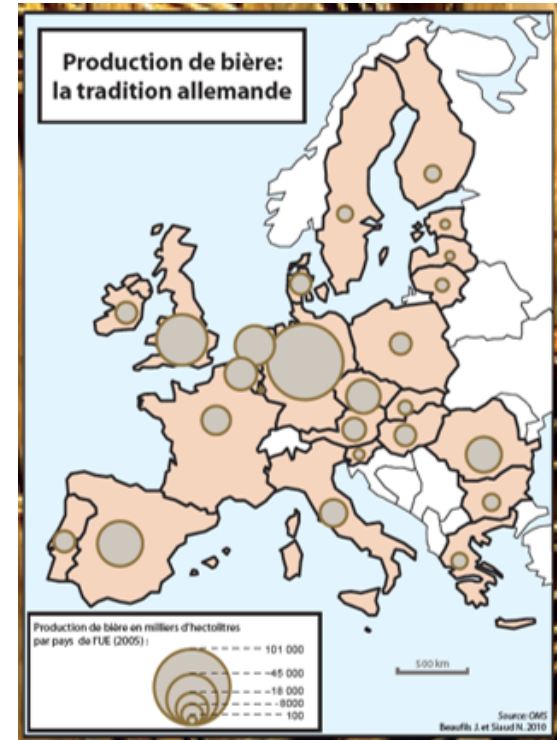
Données qualitatives

Un territoire nationale à dominante agricole



Sources : Union européenne – SOeS, Corine Land Cover, 2006, Agence de la biomédecine 2010, CIAT-CSI (SRTM <http://srtm.csi.cgiar.org>) 2010

Données quantitatives



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

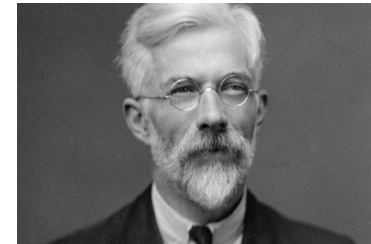
Il est nécessaire de restructurer ces données, le plus souvent sous la forme d'un nouveau tableau. Les différents types de tableaux seront abordés dans la seconde partie du cours.

1- Les données en cartographie

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

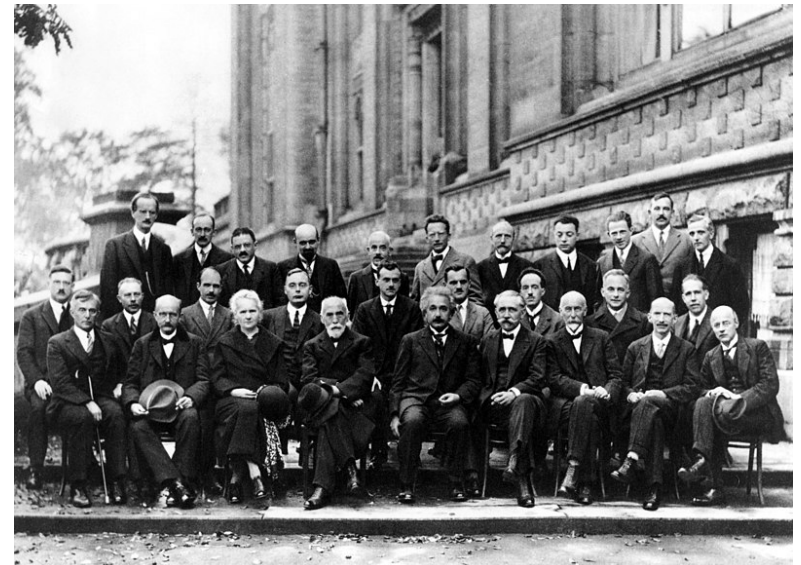
La population ou l'ensemble :

- 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 :



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
BLR	Belarus	Europe
BEL	Belgium	Europe
BIH	Bosnia and Herz.	Europe

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



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
-----	--------	------	---------

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	30561.53 [km^2]	11484055

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

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

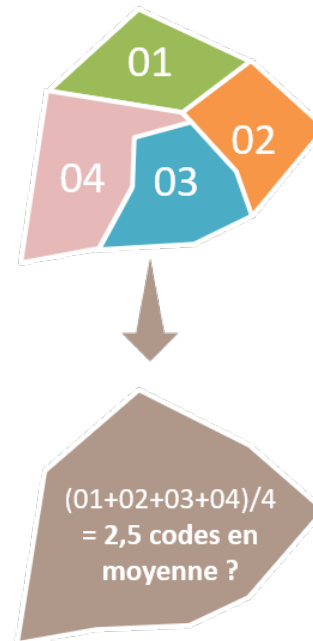
- Une population
- Un taux de chômage
- Une densité
- L'IDH

Quantitatives et qualitatives ?

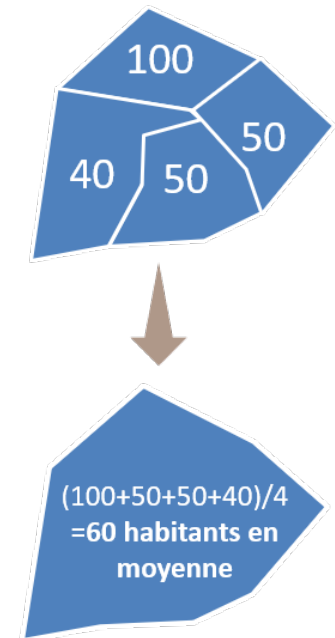
Il est important de pouvoir justifier le type de données :

- 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 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 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

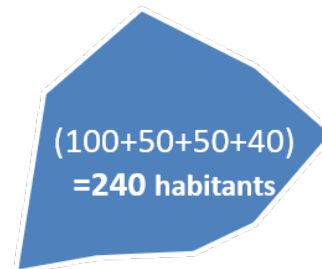
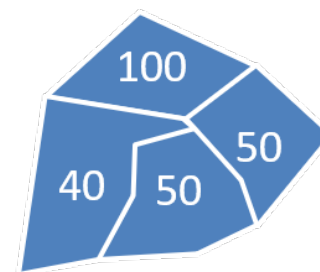
Comment faire la différence entre stock et taux ?

- 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)

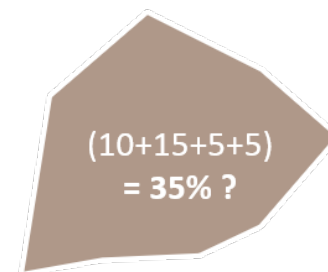
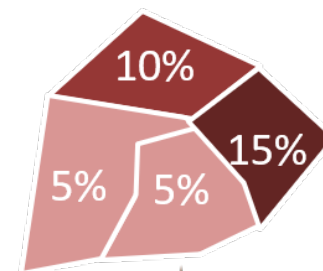
Attention, ce n'est pas parce que la valeur contient une virgule qu'il s'agit d'une données de taux :

- 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 :



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



2- Les différents types de tableaux

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)

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).

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

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

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 N_{ij} , correspond au " nombre d'éléments du tableau élémentaire E qui prennent simultanément la modalité i de X et la modalité j de Y ".

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

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

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	-

Conclusion

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