Introduction aux statistiques et cartographie

Serge Lhomme

Maître de conférences en géographie
http://sergelhomme.fr/
serge.lhomme@u-pec.fr

Généralités sur l'information géographique

Définition

Définition

L'information géographique est la représentation d'un objet ou d'un phénomène localisé dans l'espace à un moment donné.

Il s'agit d'un type d'information très répandu, décrivant des objets, phénomènes, êtres vivants ou sociétés, dès lors qu'ils sont reliés à un territoire.

Généralités sur l'information géographique

Les deux composantes de l'information géographique

Les deux principales composantes de l'information géographique sont :

- Le niveau sémantique qui correspond à l'information relative à un objet. L'ensemble des caractéristiques de l'objet forme ses attributs (comme par exemple : le numéro d'une parcelle cadastrale, le nom d'une route, le nombre d'habitants d'une commune. . .).
- Le niveau géométrique qui correspond à la forme et à la localisation de l'objet sur la surface terrestre. Un système de coordonnées peut être valable sur tout ou partie de la surface terrestre. On peut aussi définir un système de « coordonnées relatives » par rapport à un point d'origine quelconque, comme c'est souvent le cas pour les relevés topographiques.

Les fondements de la cartographie

La subjectivité

On appelle carte toute représentation graphique partielle ou complète dans le plan d'un objet plus complexe. Dans les domaines de la géographie et de l'aménagement, cet objet est un territoire représenté selon une "vue de dessus".

Cette représentation graphique est établie par un auteur, à un moment donné, sur un espace donné. Ce caractère interprétatif (voire subjectif) de la conception cartographique est largement accepté par les géographes.

D'autres publics imaginent parfois que l'image du territoire représentée sur une carte est véridique et même objective.

Naturellement, aucun cartographe digne de ce nom ne cherche à tromper ses lecteurs, mais force est de constater que le travail du cartographe résulte de multiples choix d'application, de conventions, de plaisirs esthétiques qui ne sont pas toujours explicités, ni justifiés.

Les fondements de la cartographie

L'art cartographique

La cartographie peut être définie comme l'« ensemble des études et des opérations scientifiques, artistiques et techniques intervenant à partir des résultats d'observations directes ou de l'exploitation d'une documentation, en vue de l'élaboration de cartes et autres modes d'expression, ainsi que dans leur utilisation ».

A l'instar des mathématiques, la cartographie tend à être un langage universel. Ainsi quelle que soit la langue utilisée, une carte est théoriquement compréhensible par tout le monde, même si la légende est parfois nécessaire pour saisir des points de détails.

Pour cela, ce langage doit respecter les règles de lisibilité, de clarté, d'intelligibilité et d'enchaînement logique, inhérentes à tout langage humain.

Les fondements de la cartographie

L'objectif du cartographe

Outil de communication par l'image, la carte doit être perçue avec un minimum de biais, dans la mesure où le concepteur a su prendre en compte les lois de la perception visuelle, du pouvoir intégrateur et séparateur de l'œil, des contrastes de couleurs, et des règles typographiques concernant les écritures.

Entre la subjectivité de l'objet qu'il réalise et la possibilité d'en faire un objet d'art, le cartographe ne doit jamais oublier que sa première mission est de se faire comprendre pour apporter l'information souhaitée.

Comme il est toujours difficile de se faire comprendre, il ne faut jamais oublier de légender sa carte.

Les fondements de la cartographie

Qu'est-ce que dessiner carte ?

Dessiner une carte, c'est tracer la forme d'éléments cartographiques (des objets géographiques) en fonction d'une échelle choisie de telle manière que le document final puisse être tenu en main.

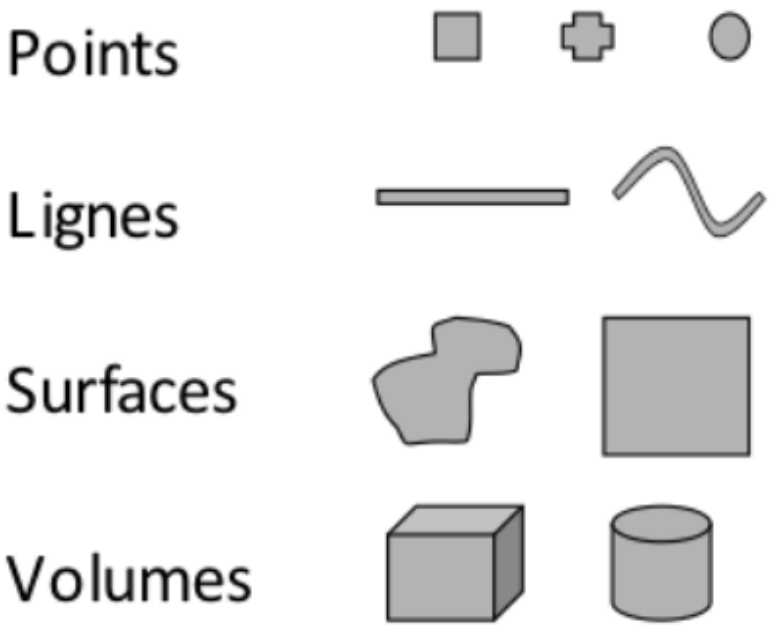
Ces éléments cartographiques sont des représentations d'objets géographiques matériels (routes, maisons...) ou immatériels (limites communales, frontières nationales), visibles (aéroports) ou invisibles (réseaux enterrés), immobiles ou mobiles (camions).

La représentation des éléments cartographiques fait appel à trois types de tracés (ou primitives graphiques, ou encore structures visuelles) : le point, le segment (la ligne ou la polyligne), le périmètre fermé et sa surface associée (un polygone).

Les fondements de la cartographie

Primitives, structures visuelles

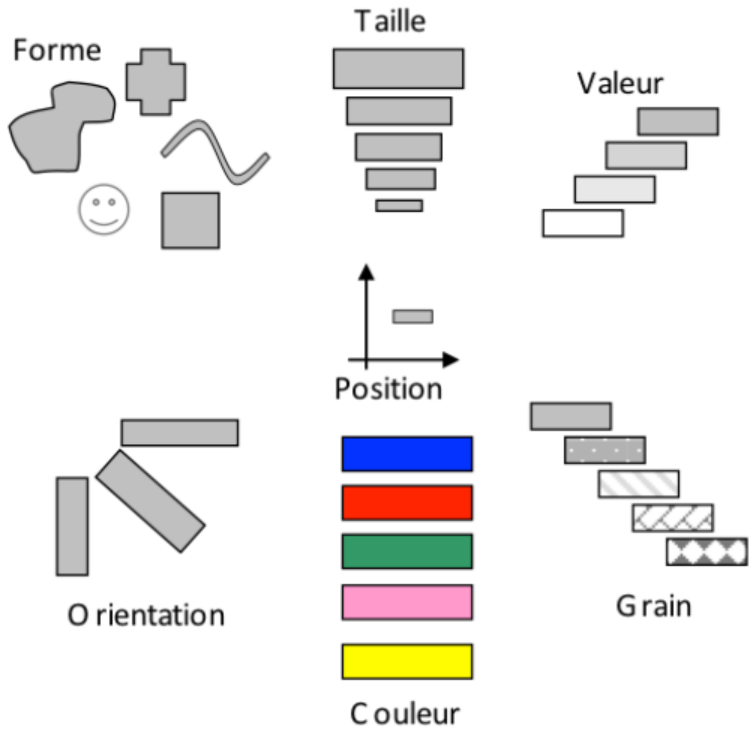
Un objet géographique matériel prend la forme de telle ou telle autre primitive graphique en fonction de l'échelle de représentation : à l'échelle du 1 : 1 000 000 une ville est représentée par un point, alors qu'à l'échelle du 1 : 100 000, elle occupe un périmètre plus ou moins nettement délimité.



Les fondements de la cartographie

Coder les informations

Pour coder des informations, il est possible de faire varier certaines propriétés graphiques de ces structures visuelles (par exemple la forme). Les variations possibles sur les structures visuelles sont regroupées par type et sont nommées « variables visuelles ».



Les fondements de la cartographie

La carte est un ensemble de calques

Du point de vue de l'infographie, une carte thématique peut être considérée comme l'empilement d'un ensemble de calques, d'un ensemble de couches.

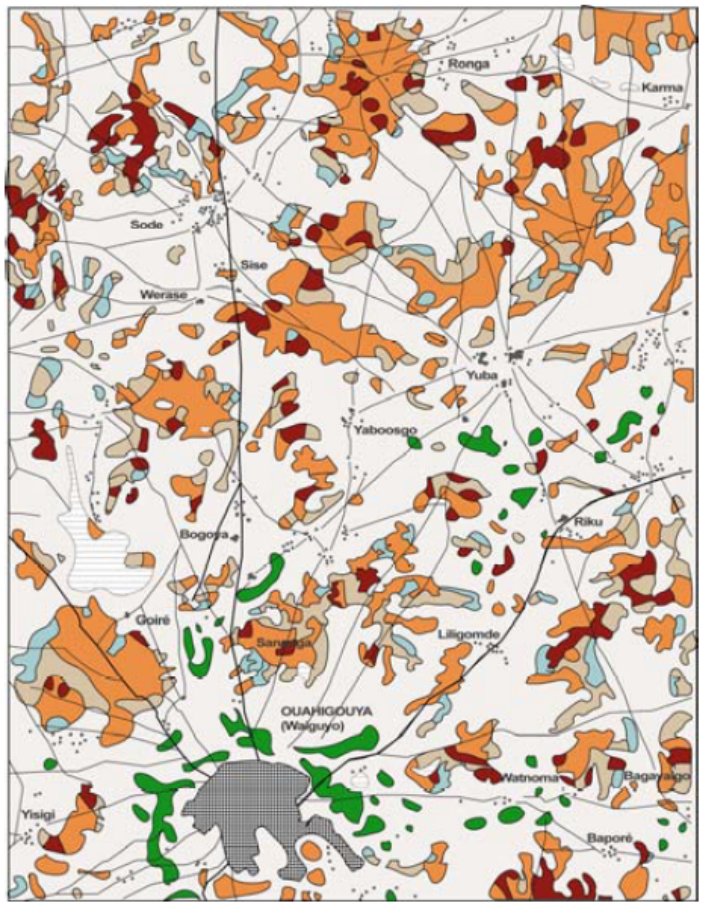
Chacun de ces calques contient un type d'éléments cartographiques et un seul.

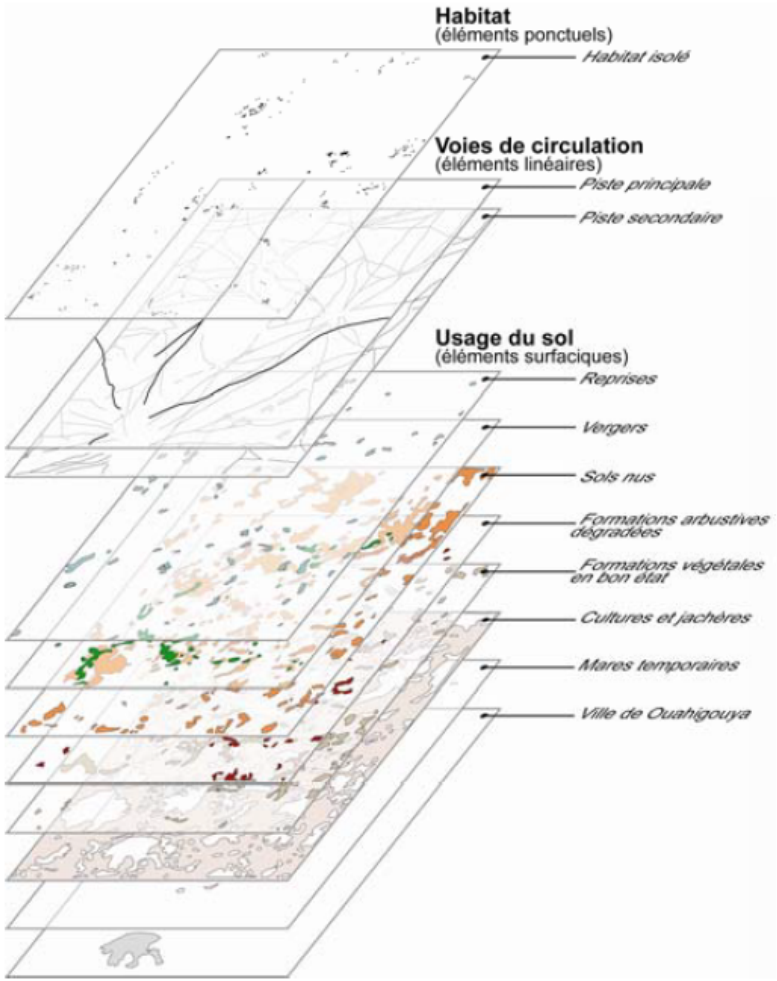
Contrairement au lecteur, le cartographe doit voir la carte comme un ensemble de calques et non comme un ensemble d'objets.

Le parallèle entre la décomposition de la carte en différents calques et la lecture de la légende est alors souvent évident. En effet, une structure visuelle combinée à des variables visuelles correspond à un type d'objets souvent souvent associé à un calque.

Les fondements de la cartographie

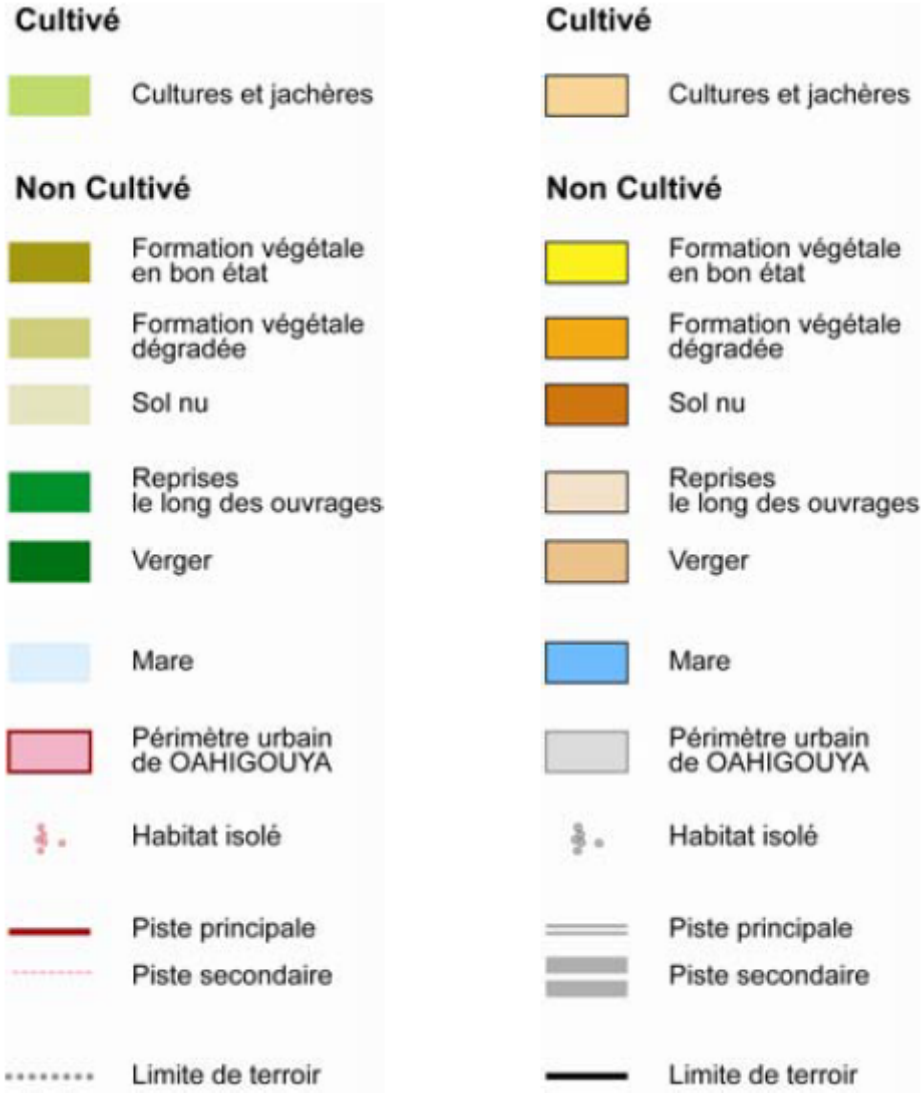
La carte est un ensemble de calques





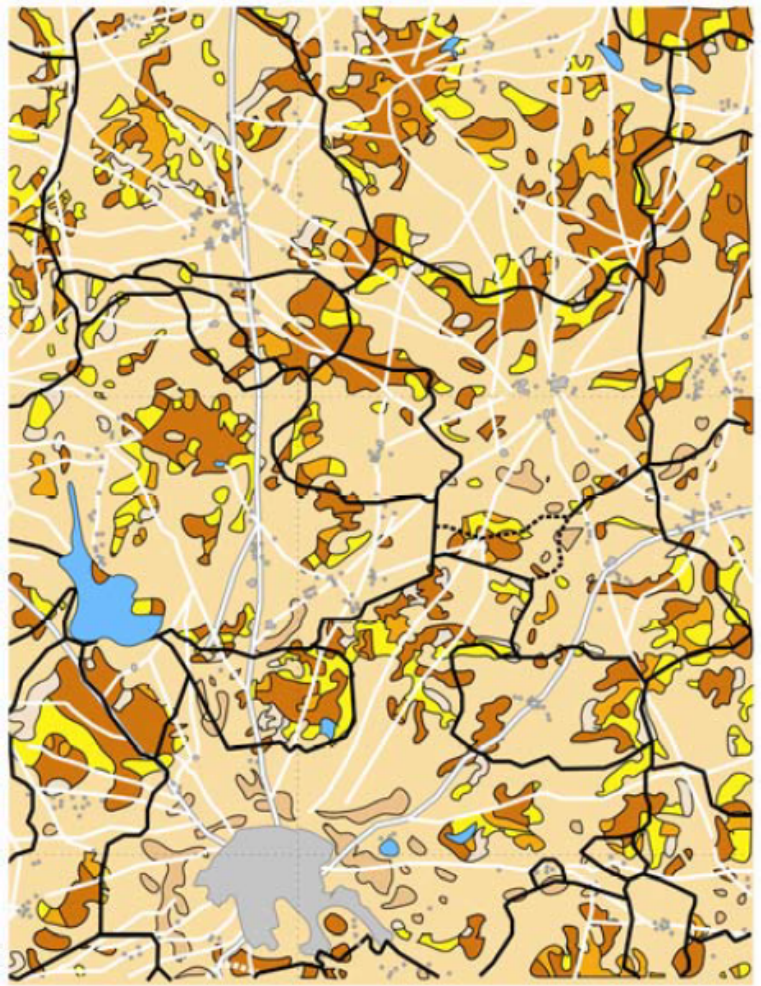
Les fondements de la cartographie

La carte parfaite n'existe pas



Les fondements de la cartographie

La carte parfaite n'existe pas



Les fondements de la cartographie

Faire une carte ne se limite pas à produire une simple image artistique d'un espace. La carte doit transmettre un message. Pour cela, il faut :

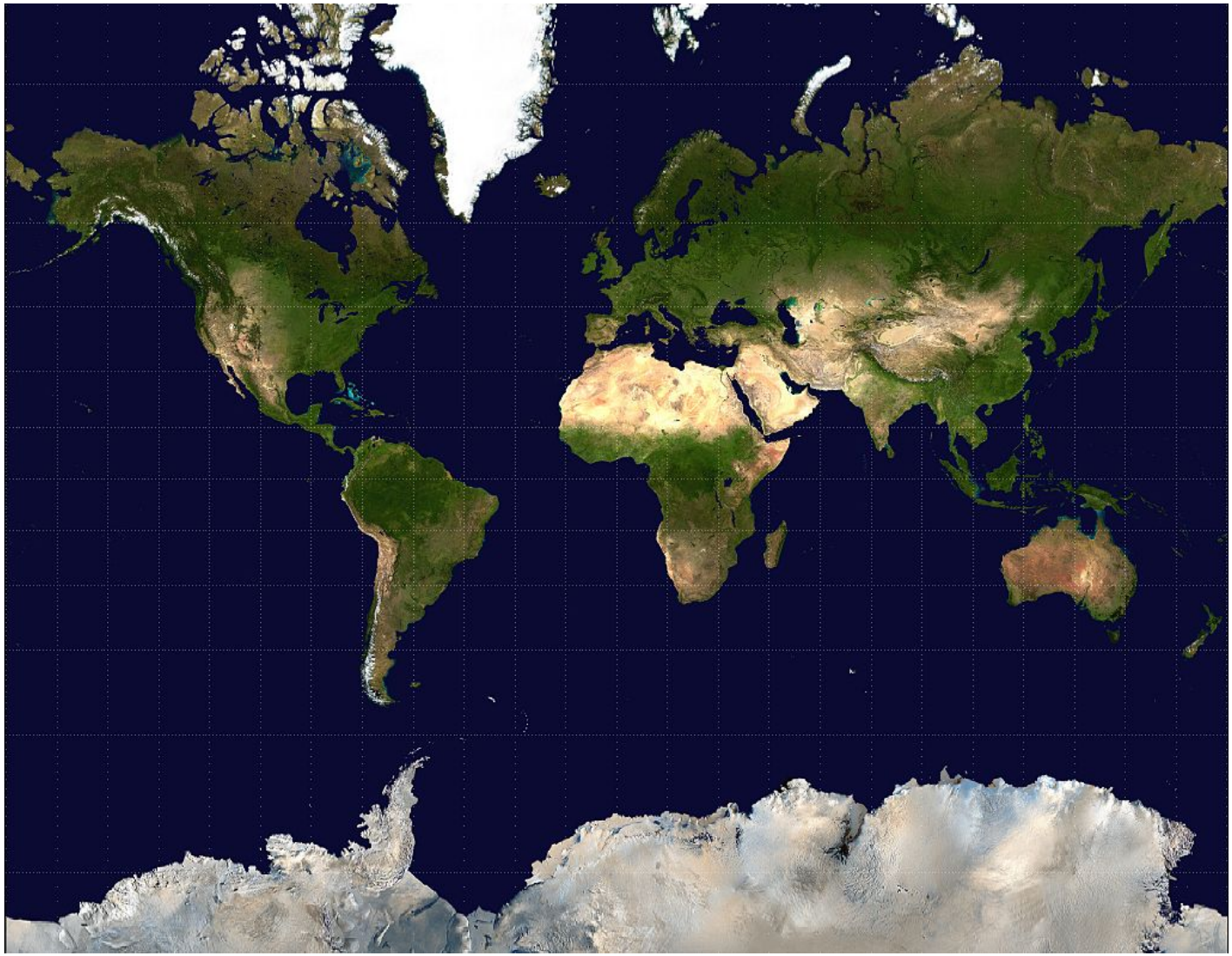
- Identifier l'objectif de la carte ;
- Identifier la cible ;
- Identifier l'information à cartographier (collecte et traitement) ;
- Adapter le fond de carte (projection, généralisation) ;
- Choisir le langage cartographique.

Il ne faut pas oublier :

- D'indiquer le nord ;
- De préciser l'échelle ;
- De mettre une légende ;
- De préciser les sources ;
- De mettre un titre.

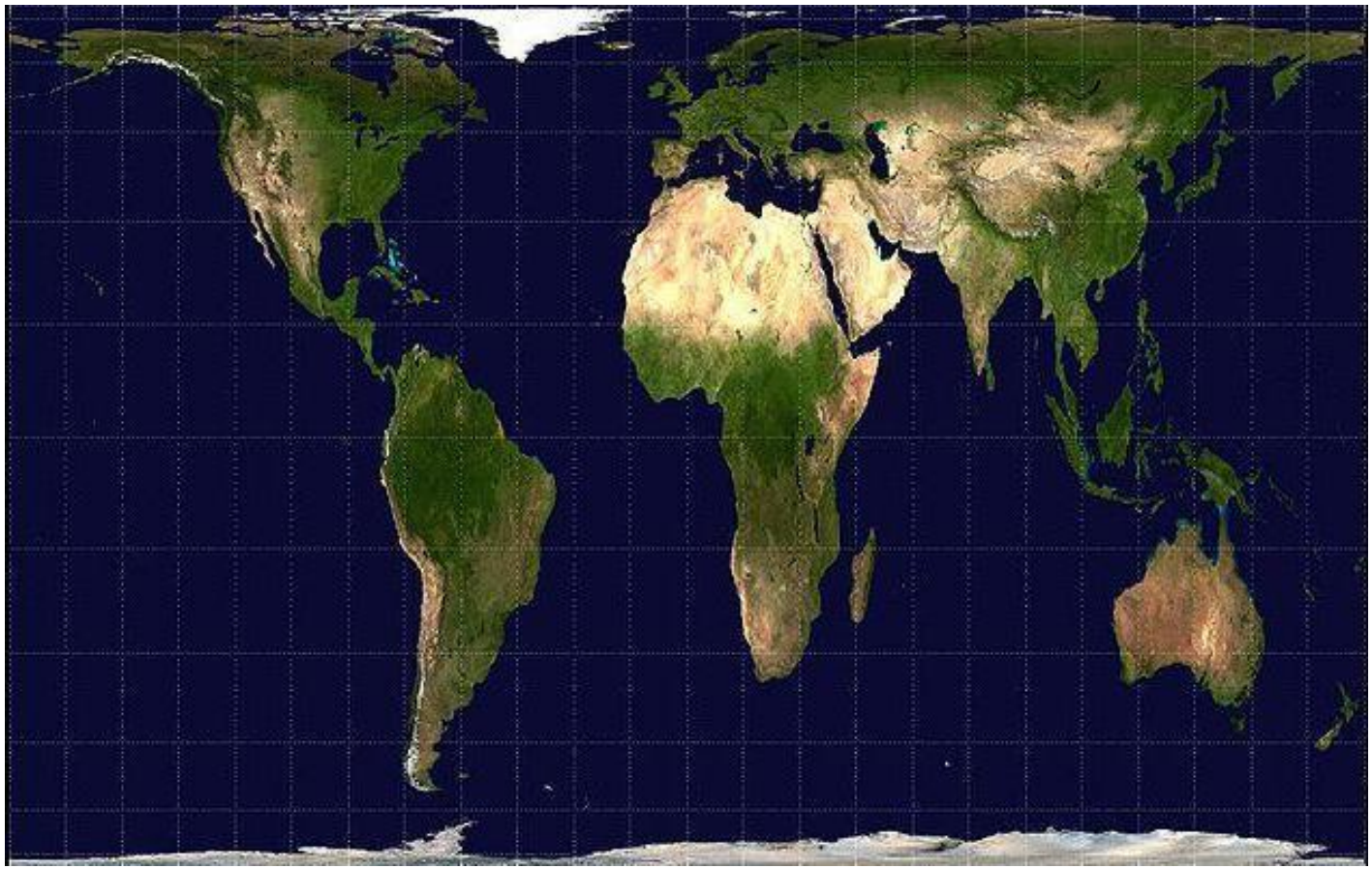
Toutes les cartes sont fausses

La projection Mercator (1569)



Toutes les cartes sont fausses

La projection de Peters (1967)



Produire des cartes thématiques quantitatives

Enrichir l'information géographique

Réaliser une carte, ce n'est pas toujours essayer de représenter au mieux les territoires qui nous entourent, de les dessiner, comme dans les cartes topographiques et les cartes d'inventaire.

Réaliser une carte, c'est parfois mettre en évidence des caractéristiques que l'on ne voit pas forcément.

Pour cela, il convient de ne pas se contenter de représenter l'information géographique, il convient plutôt de l'analyser et de la rendre intelligible.

Il existe deux grandes méthodes d'analyse des données géographiques : l'analyse quantitative et l'analyse qualitative.

On va se concentrer ici sur l'analyse quantitative. En effet, c'est l'analyse la plus simple. Pour cela, on s'appuiera sur la sémantique de l'information géographique : le tableau d'information géographique.

Produire des cartes thématiques quantitatives

Le tableau d'information géographique : la sémantique de l'information géographique

Variables Caractères

Entités
Individus
Objets
Unités

	AGRI	ARTI	CADRE	PRO INT	EMPLOYE	OUVRIER	RETRAITE	AUTRES
0	4067	16745	36426	70663	77349	78998	30417	29910
1	5201	11414	18629	48889	71578	78906	32063	42108
2	6159	9396	11842	31102	46196	42101	24218	21344
3	2027	6387	6951	16140	20885	15560	10687	10278
4	2040	5286	5883	15687	19707	12415	8865	7222
5	1918	39514	73884	117652	160574	87227	54801	68990
6	4132	10758	12461	31859	39088	37199	21660	19063
7	3362	6520	9771	25487	36326	42618	16464	23226
8	2348	4888	5537	14442	20788	16635	10379	9741
9	4786	7198	12499	28889	40121	43224	18616	18373
10	5474	12045	13486	31504	47297	34501	25584	25482
11	11100	10116	9937	25717	33970	29390	18550	13039
12	4508	52145	135584	225526	268192	168920	100790	155952
13	6081	18540	36155	72740	94384	82643	41776	34559
14	7431	4976	4547	12517	19508	17183	10386	7786
15	5756	10236	14308	32982	46169	46763	24653	20416
16	9061	21025	24906	57199	84800	65577	46140	35151
17	4025	8183	12967	29897	42604	39209	22057	18361
18	4930	7553	9273	23538	33050	28048	16824	12970
19	5352	13503	32185	62097	71854	61993	30098	23303
20	12517	17359	24766	54660	70213	68805	43023	29034
21	5259	3696	3669	9387	16039	12764	9811	6780
22	7694	15460	14125	34308	54103	47943	31651	24558
23	4123	12196	29138	57994	65018	77588	28449	27692
24	6106	14998	23027	52929	59618	57491	28522	30398
25	4064	14930	28312	62716	74944	85833	34937	33452
26	3848	10110	22675	47678	58632	57416	25285	22370
27	11054	23043	46057	94797	116573	101817	59116	46699

Valeur
Modalité

Produire des cartes thématiques quantitatives

Variables et valeurs

Les variables d'un tableau d'information géographique peuvent être quantitatives ou qualitatives.

Une variable qualitative est composée de valeurs qui ne sont pas des quantités (ou des rapports de quantités). Les valeurs prises par cette variable ne peuvent pas faire l'objet d'opérations mathématiques simple comme une addition (en tout cas ça n'a pas de sens). Ces valeurs peuvent être simplement nominales (lettres, mots, chiffres) ou ordinales c'est-à-dire qu'on peut les ordonner par ordre croissant (les classements : 1er, 2ème, 3ème, ... ; les jugements : bons, moyens, mauvais...).

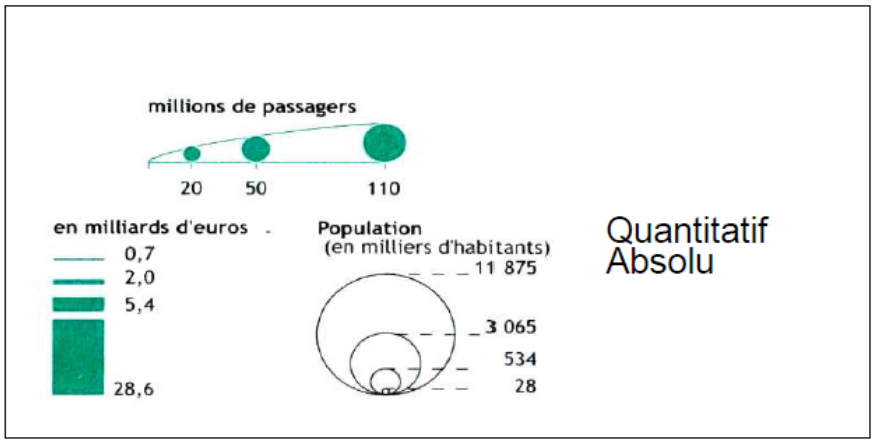
On va dans ce cours plutôt se focaliser sur des variables quantitatives. Les valeurs sont des quantités (des nombres d'individus). Toutes les opérations peuvent être opérés sur les valeurs d'une variable.

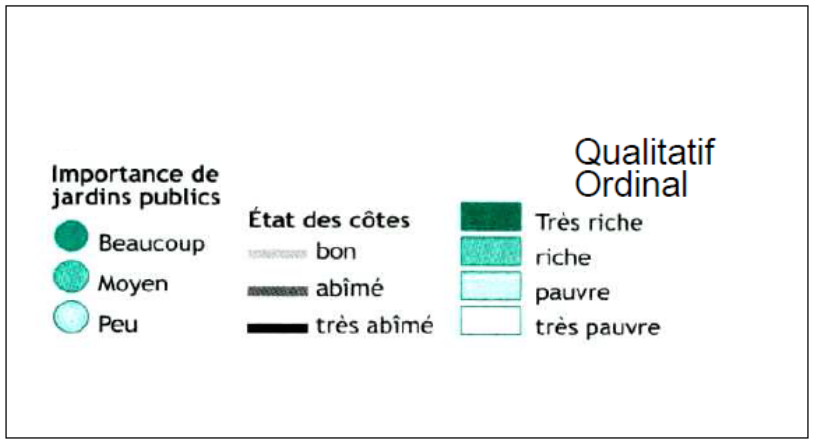
Produire des cartes thématiques quantitatives

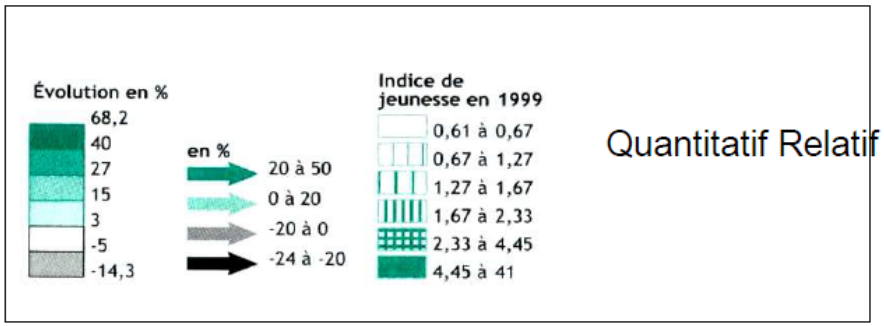
Résumé

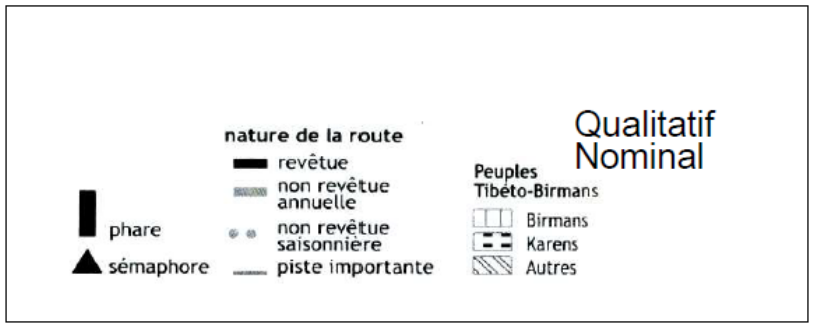
En cartographie, on différencie notamment les variables (quantitatives) de stock et celle de taux.

On distinguera aussi les variables de flux (tableau d'échange).









Discrétisation

Simplifier l'information pour faciliter le message à faire passer

La discrétisation consiste à découper des données en classes homogènes.

C'est une opération de simplification.

Pour réaliser une discrétisation, il faut déterminer le nombre de classes et les bornes des classes.

Pour réaliser une bonne discrétisation, il faut justifier à la fois le nombre de classes et les bornes de ces classes.

Intuitivement, un bon découpage correspond à des classes homogènes et séparées, ce qui correspond respectivement aux notions statistiques de faible variance intraclasse et de forte variance interclasse.

Et oui, pour réaliser une simple carte représentant une seule et pauvre variable, il faut des connaissances statistiques de base solides !

Discrétisation

Les résumés statistiques classiques

La moyenne :

$$\bar{X} = \frac{1}{N} \times \sum_{i=1}^{N} X_i$$

L'écart-type :

$$\sigma_D = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (X_i - \bar{X})^2}$$

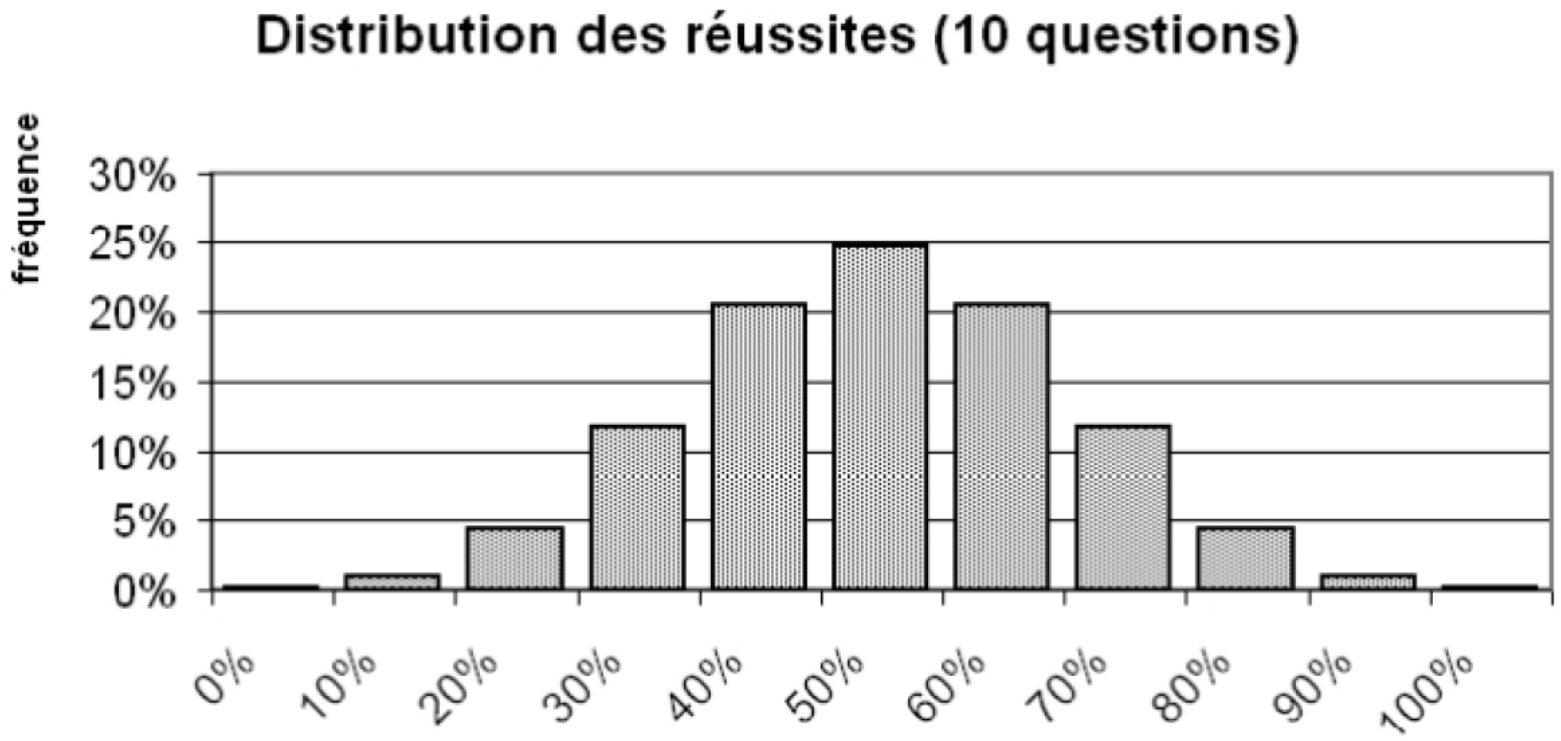
Les caractéristiques de position fondées sur l'ordonnancement (les modes) :

Minimum ; Maximum ; Médiane ; Quantiles ; Déciles...

Discrétisation

L'arme absolue : la distribution statistique

La distribution statistique (distribution des fréquences) est un tableau qui associe des classes de valeurs obtenues lors d'une expérience à leurs fréquences d'apparition. Ce tableau de valeurs est modélisé en théorie des probabilités par une loi de probabilité. Il peut notamment se représenter sous la forme d'un histogramme.



Discrétisation

Les règles

Le nombre de classes est en théorie proportionnel au nombre d'individus observés. Ainsi il peut paraitre excessif de créer 8 classes pour seulement 15 pays. Il existe un indice permettant de connaitre le nombre idéal de classes.

$$N(cl) = 1 + 3,3log(N) \text{ Huntsberger}$$

L'objectif est bien souvent de conserver la forme de la distribution et donc de conserver au mieux la structure interne : Jenks ou Seuils naturels.

Pour illustrer la dispersion des variables étudiées, il faut tout simplement choisir une classification par amplitude égale (qui permet en fait d'obtenir une carte de la distribution statistique).

Pour comparer la position de certains lieux en fonction de différentes variables (sur plusieurs cartes), il faudra utiliser des méthodes faisant référence à des paramètres statistiques (moyenne, médiane, EcT...).

TP1

Produire des cartes thématiques quantitatives

1. A l'aide de QGIS, représentez la variable cadre. Quelle variable visuelle choisir pour représenter cette variable ? Quelles sont les qualités et les défauts de cette carte ?

2. Enregistrez cette représentation aux formats PNG et SVG.

3. A l'aide de QGIS, représentez la variable Tx_Cadre ou Tx_Profession intermédiaire. Quelle variable visuelle choisir pour représenter ces variables ?

4. Comparez les représentations obtenues avec 4 discrétisations différentes comportant toutes 5 classes. Enregistrez les représentations aux formats SVG et PNG. Décrivez les principes de chaque discrétisation. Quelle est la meilleure représentation ?

TP1

Distribution et discrétisation

Dans QGIS, on a testé différentes représentations obtenues par différentes discrétisations des variables suivantes : Tx_Cadre ou Tx_Profession intermédiaire. La question laissée en suspens est la suivante : quelle est la meilleure représentation pour chaque variable ?

1. Réalisez des histogrammes de distribution statistique pour ces deux variables. Testez pour cela plusieurs amplitudes de classe. Quelles sont les formes de ces deux distributions statistiques ?

2. A partir de l'analyse des distributions obtenues, choisissez la meilleure méthode de discrétisation, puis choisissez le bon nombre de classes afin d'obtenir la meilleure représentation possible pour chaque variable.

3. Finalisez cette carte à l'aide d'Inkscape.

TP2

Résumé statistique

Vous disposez d'un fichier Excel décrivant la répartition de différentes catégories socioprofessionnelles au sein des départements de France métropolitaine.

1. Ouvrez le fichier, puis calculez la somme, la moyenne et la médiane de chaque CSP.
2. Déterminez la variance, puis l'écart-type pour la variable Ouvrier.
3. Figez la première ligne afin de la visualiser de manière permanente.
4. Créez un tableau. Quel changement s'est opéré au niveau des formules si vous calculez par exemple le minimum ou le maximum de chacune des variables ?

Tableau de contingence

Poids des entités géographiques et des variables

Effectifs observés (Nij)									
1963	*ALIM*	*TEXT*	*BOIS*	*EDIT*	*CHIM*	*CONS*	*META*	*EQUIP*	**Total**
BULGARIE	130	128	39	14	29	47	21	151	**559**
HONGRIE	144	241	53	28	77	61	91	423	**1118**
POLOGNE	380	612	164	84	222	199	147	881	**2689**
R.D.A.	206	451	119	118	308	142	109	1056	**2509**
ROUMANIE	136	305	244	41	76	114	106	366	**1388**
TCHECO.	185	412	130	63	139	151	177	883	**2140**
YOUGOSL.	126	223	132	58	76	78	69	307	**1069**
Total	**1307**	**2372**	**881**	**406**	**927**	**792**	**720**	**4067**	**11472**

Tableau de contingence

Poids des entités géographiques et des variables

Candidats	Agriculteurs	Artisans, Commerçants et chefs d'entreprise	Professions libérales, Cadres Supérieurs	Professions intermédiaires	Employés	Ouvriers	Étudiants	Chômeurs	Total
Schivardi	0	0	0	0	12	10	0	0	21*
Laguiller	0	0	0	9	23	29	0	4	65
Besancenot	0	0	4	53	58	78	24	32	249
Buffet	5	0	4	9	23	20	3	24	87
Bové	5	0	7	18	12	10	8	4	64
Royal	14	40	110	276	289	205	85	129	1148
Voynet	4	0	7	36	12	10	5	4	77
Nihous	0	0	0	9	12	20	0	4	44
Bayrou	32	64	103	178	185	157	59	89	865
Sarkozy	64	117	103	231	335	205	56	76	1189
Villiers	20	5	7	18	35	10	5	0	100
Le Pen	34	40	11	53	162	225	21	36	582
Total	178	267	356	889	1156	978	267	400	4492

* Avertissement : Le tableau donne les effectifs de vote aux dix millièmes (4492 au lieu de 44.920.000 individus). Les effectifs sont exprimés sans aucune décimale ce qui conduit à des approximations quant aux calculs des effectifs marginaux. Par exemple, le nombre de votes pour le candidat Schivardi a été estimé à 21 (soit 210.000) électeurs et non à 22 (soit 22.000) électeurs (12+10). Ce constat est généralisable à l'ensemble des tableaux de résultats. Cette approximation n'interfère en aucun cas sur le résultat de l'AFC.

Tableau de contingence

Poids des entités géographiques et des variables

	Chimie	Economie	Littérature	Medecine	Paix	Physique	Somme
Allemagne	24	1	8	18	5	24	80
Canada	4	3	2	4	1	4	18
France	8	3	11	12	10	9	53
GB	23	6	7	26	11	20	93
Italie	1	1	6	5	1	5	19
Japon	6	0	2	3	1	11	23
Russie	4	3	5	2	3	10	27
USA	51	43	8	70	19	66	257
Somme	121	60	49	140	51	149	570

Tableau de contingence

Deux profils possibles : le profil en ligne

Profils en ligne (Nij/Ni.)									
1963	*ALIM*	*TEXT*	*BOIS*	*EDIT*	*CHIM*	*CONS*	*META*	*EQUIP*	**Total**
BULGARIE	23%	23%	7%	3%	5%	8%	4%	27%	**100%**
HONGRIE	13%	22%	5%	3%	7%	5%	8%	38%	**100%**
POLOGNE	14%	23%	6%	3%	8%	7%	5%	33%	**100%**
R.D.A.	8%	18%	5%	5%	12%	6%	4%	42%	**100%**
ROUMANIE	10%	22%	18%	3%	5%	8%	8%	26%	**100%**
TCHECO.	9%	19%	6%	3%	6%	7%	8%	41%	**100%**
YOUGOSL.	12%	21%	12%	5%	7%	7%	6%	29%	**100%**
Total	**11%**	**21%**	**8%**	**4%**	**8%**	**7%**	**6%**	**35%**	**100%**

Tableau de contingence

Deux profils possibles : le profil en colonne

Profils en colonne (Nij/N.j)									
1963	*ALIM*	*TEXT*	*BOIS*	*EDIT*	*CHIM*	*CONS*	*META*	*EQUIP*	**Total**
BULGARIE	10%	5%	4%	3%	3%	6%	3%	4%	**5%**
HONGRIE	11%	10%	6%	7%	8%	8%	13%	10%	**10%**
POLOGNE	29%	26%	19%	21%	24%	25%	20%	22%	**23%**
R.D.A.	16%	19%	14%	29%	33%	18%	15%	26%	**22%**
ROUMANIE	10%	13%	28%	10%	8%	14%	15%	9%	**12%**
TCHECO.	14%	17%	15%	16%	15%	19%	25%	22%	**19%**
YOUGOSL.	10%	9%	15%	14%	8%	10%	10%	8%	**9%**
Total	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**

Profil en ligne

Indice de spécialisation

	Chimie	Economie	Littérature	Medecine	Paix	Physique	Somme
Allemagne	24	1	8	18	5	24	80
Canada	4	3	2	4	1	4	18
France	8	3	11	12	10	9	53
GB	23	6	7	26	11	20	93
Italie	1	1	6	5	1	5	19
Japon	6	0	2	3	1	11	23
Russie	4	3	5	2	3	10	27
USA	51	43	8	70	19	66	257
Somme	121	60	49	140	51	149	570

Profil en ligne

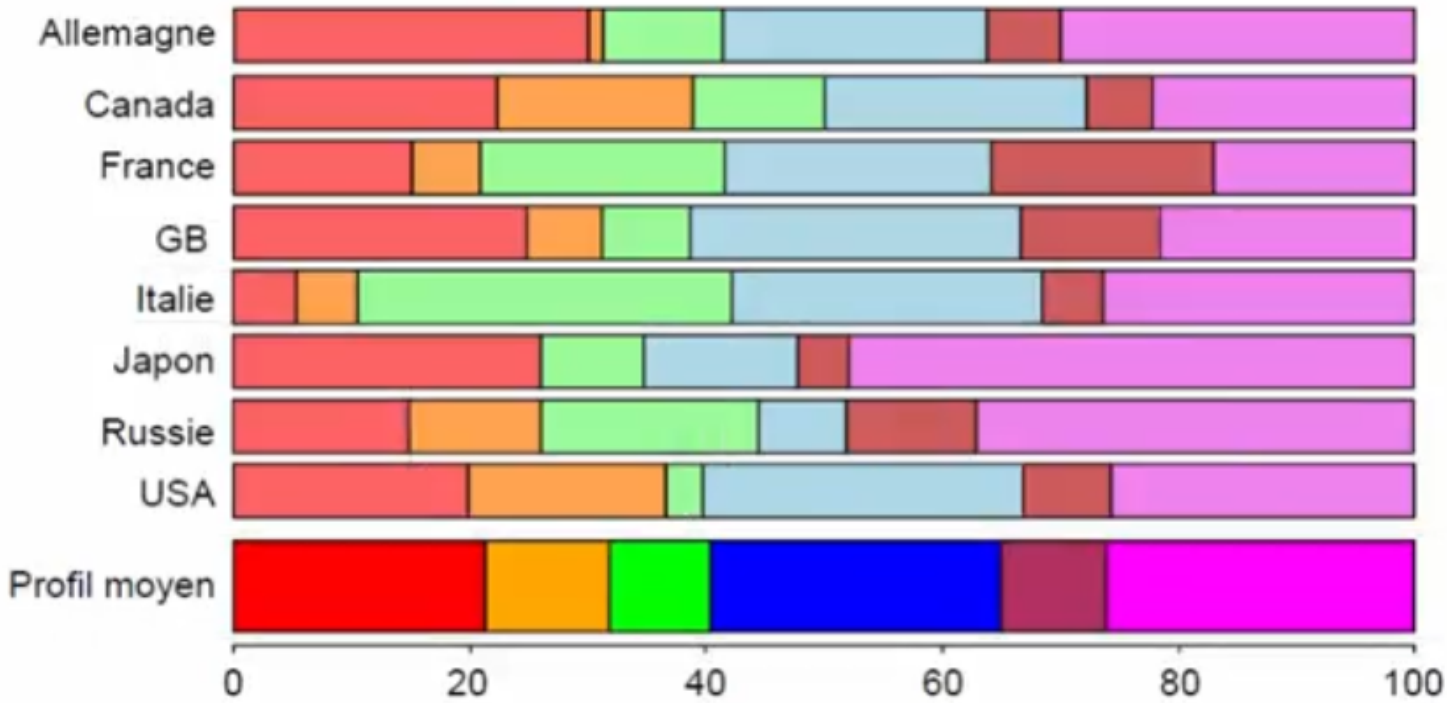
Indice de spécialisation

	Chimie	Economie	Littérature	Medecine	Paix	Physique	**Somme**
Allemagne	0,300	0,013	0,100	0,225	0,063	0,300	**1**
Canada	0,222	0,167	0,111	0,222	0,056	0,222	**1**
France	0,151	0,057	0,208	0,226	0,189	0,170	**1**
GB	0,247	0,065	0,075	0,280	0,118	0,215	**1**
Italie	0,053	0,053	0,316	0,263	0,053	0,263	**1**
Japon	0,261	0,000	0,087	0,130	0,043	0,478	**1**
Russie	0,148	0,111	0,185	0,074	0,111	0,370	**1**
USA	0,198	0,167	0,031	0,272	0,074	0,257	**1**
Somme	**0,212**	**0,105**	**0,086**	**0,246**	**0,089**	**0,261**	**1**

Profil en ligne

Indice de spécialisation

	Chimie	Eco	Lit.	Médecine	Paix	Physique	Somme
Allemagne	30.0	1.2	10.0	22.5	6.2	30.0	100
Canada	22.2	16.7	11.1	22.2	5.6	22.2	100
France	15.1	5.7	20.8	22.6	18.9	17.0	100
GB	24.7	6.5	7.5	28.0	11.8	21.5	100
Italie	5.3	5.3	31.6	26.3	5.3	26.3	100
Japon	26.1	0.0	8.7	13.0	4.3	47.8	100
Russie	14.8	11.1	18.5	7.4	11.1	37.0	100
USA	19.8	16.7	3.1	27.2	7.4	25.7	100
Profil moyen	21.2	10.5	8.6	24.6	8.9	26.1	100



Profil en ligne

Indice de spécialisation

	Chimie	Economie	Littérature	Medecine	Paix	Physique	**Somme**
Allemagne	0,300	0,013	0,100	0,225	0,063	0,300	**1**
Canada	0,222	0,167	0,111	0,222	0,056	0,222	**1**
France	0,151	0,057	0,208	0,226	0,189	0,170	**1**
GB	0,247	0,065	0,075	0,280	0,118	0,215	**1**
Italie	0,053	0,053	0,316	0,263	0,053	0,263	**1**
Japon	0,261	0,000	0,087	0,130	0,043	0,478	**1**
Russie	0,148	0,111	0,185	0,074	0,111	0,370	**1**
USA	0,198	0,167	0,031	0,272	0,074	0,257	**1**
Somme	**0,212**	**0,105**	**0,086**	**0,246**	**0,089**	**0,261**	**1**

On appelle INDICE DE SPECIALISATION (Si) l'écart entre le profil d'une unité spatiale et le profil moyen de l'ensemble de référence.

$$S_i = \sum_{j=1}^{n} \mid \frac{X_{ij}}{X_{i.}} - \frac{X_{.j}}{X_{..}} \mid = \sum_{j=1}^{n} \mid PO_{ij} - PM_{ij} \mid$$

Profil en ligne

Indice de spécialisation

	Chimie	Economie	Littérature	Medecine	Paix	Physique	**Somme**
Allemagne	0,300	0,013	0,100	0,225	0,063	0,300	**1**
Canada	0,222	0,167	0,111	0,222	0,056	0,222	**1**
France	0,151	0,057	0,208	0,226	0,189	0,170	**1**
GB	0,247	0,065	0,075	0,280	0,118	0,215	**1**
Italie	0,053	0,053	0,316	0,263	0,053	0,263	**1**
Japon	0,261	0,000	0,087	0,130	0,043	0,478	**1**
Russie	0,148	0,111	0,185	0,074	0,111	0,370	**1**
USA	0,198	0,167	0,031	0,272	0,074	0,257	**1**
Somme	**0,212**	**0,105**	**0,086**	**0,246**	**0,089**	**0,261**	**1**

$$S_{(Allemagne)} = | 0.300 - 0.212 | + | 0.013 - 0.105 | + | 0.100 - 0.086 | + | 0.225 - 0.246 | + | 0.063 - 0.089 | + | 0.300 - 0.261 | = 0.280$$

Profil en ligne

Indice de spécialisation

Allemagne	0,281
Canada	0,193
France	0,442
GB	0,196
Italie	0,498
Japon	0,533
Russie	0,471
USA	0,178

Profil en colonne

Indice de localisation

	Chimie	Economie	Littérature	Medecine	Paix	Physique	Somme
Allemagne	24	1	8	18	5	24	80
Canada	4	3	2	4	1	4	18
France	8	3	11	12	10	9	53
GB	23	6	7	26	11	20	93
Italie	1	1	6	5	1	5	19
Japon	6	0	2	3	1	11	23
Russie	4	3	5	2	3	10	27
USA	51	43	8	70	19	66	257
Somme	121	60	49	140	51	149	570

Profil en colonne

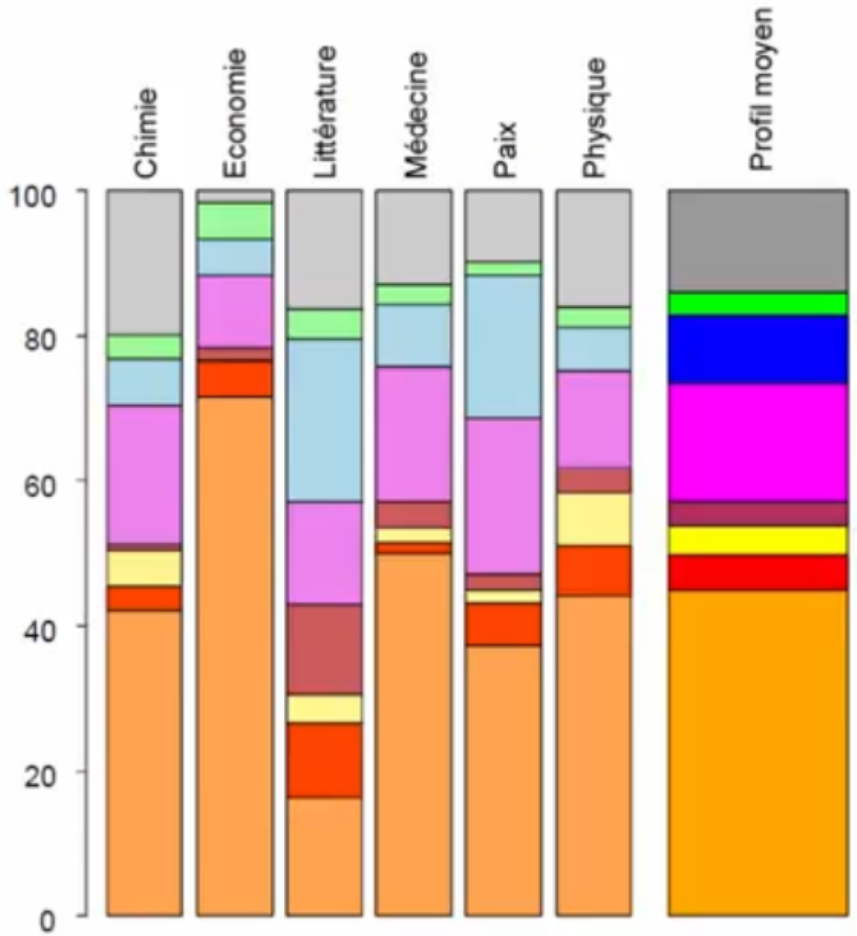
Indice de localisation

	Chimie	Economie	Littérature	Medecine	Paix	Physique	**Somme**
Allemagne	0,198	0,017	0,163	0,129	0,098	0,161	**0,140**
Canada	0,033	0,050	0,041	0,029	0,020	0,027	**0,032**
France	0,066	0,050	0,224	0,086	0,196	0,060	**0,093**
GB	0,190	0,100	0,143	0,186	0,216	0,134	**0,163**
Italie	0,008	0,017	0,122	0,036	0,020	0,034	**0,033**
Japon	0,050	0,000	0,041	0,021	0,020	0,074	**0,040**
Russie	0,033	0,050	0,102	0,014	0,059	0,067	**0,047**
USA	0,421	0,717	0,163	0,500	0,373	0,443	**0,451**
Somme	**1**	**1**	**1**	**1**	**1**	**1**	**1**

Profil en colonne

Indice de localisation

	Chimie	Eco	Lit	Méd	Paix	Phys	Profil moyen
Allemagne	19.8	1.7	16.3	12.9	9.8	16.1	14.0
Canada	3.3	5.0	4.1	2.9	2.0	2.7	3.2
France	6.6	5.0	22.4	8.6	19.6	6.0	9.3
GB	19.0	10.0	14.3	18.6	21.6	13.4	16.3
Italie	0.8	1.7	12.2	3.6	2.0	3.4	3.3
Japon	5.0	0.0	4.1	2.1	2.0	7.4	4.0
Russie	3.3	5.0	10.2	1.4	5.9	6.7	4.7
USA	42.1	71.7	16.3	50.0	37.3	44.3	45.1
Somme	100	100	100	100	100	100	100



Profil en colonne

Indice de localisation

	Chimie	Economie	Littérature	Medecine	Paix	Physique	**Somme**
Allemagne	0,198	0,017	0,163	0,129	0,098	0,161	**0,140**
Canada	0,033	0,050	0,041	0,029	0,020	0,027	**0,032**
France	0,066	0,050	0,224	0,086	0,196	0,060	**0,093**
GB	0,190	0,100	0,143	0,186	0,216	0,134	**0,163**
Italie	0,008	0,017	0,122	0,036	0,020	0,034	**0,033**
Japon	0,050	0,000	0,041	0,021	0,020	0,074	**0,040**
Russie	0,033	0,050	0,102	0,014	0,059	0,067	**0,047**
USA	0,421	0,717	0,163	0,500	0,373	0,443	**0,451**
Somme	**1**	**1**	**1**	**1**	**1**	**1**	**1**

On appelle INDICE DE LOCALISATION (Lj) l'écart entre le profil d'une catégorie et le profil moyen de l'ensemble de référence.

$$L_j = \sum_{i=1}^{n} \mid \frac{N_{ij}}{N_{.j}} - \frac{N_{i.}}{N_{..}} \mid = \sum_{i=1}^{n} \mid PO_{ij} - PM_{ij} \mid$$

Profil en colonne

Indice de localisation

	Chimie	Economie	Littérature	Medecine	Paix	Physique	**Somme**
Allemagne	0,198	0,017	0,163	0,129	0,098	0,161	**0,140**
Canada	0,033	0,050	0,041	0,029	0,020	0,027	**0,032**
France	0,066	0,050	0,224	0,086	0,196	0,060	**0,093**
GB	0,190	0,100	0,143	0,186	0,216	0,134	**0,163**
Italie	0,008	0,017	0,122	0,036	0,020	0,034	**0,033**
Japon	0,050	0,000	0,041	0,021	0,020	0,074	**0,040**
Russie	0,033	0,050	0,102	0,014	0,059	0,067	**0,047**
USA	0,421	0,717	0,163	0,500	0,373	0,443	**0,451**
Somme	**1**	**1**	**1**	**1**	**1**	**1**	**1**

$$L_{(Chimie)} = | 0.198 - 0.140 | + | 0.033 - 0.032 | + | 0.066 - 0.093 | + | 0.190 - 0.163 | + | 0.008 - 0.033 | + | 0.050 - 0.040 | + | 0.033 - 0.047 | + | 0.421 - 0.451 | = 0.192$$

Profil en colonne

Indice de localisation

Chimie	Economie	Littérature	Medecine	Paix	Physique
0,191	0,574	0,616	0,148	0,334	0,148

Quotient de localisation

Explication

Le quotient de localisation est un indicateur de « concentration », de spécialisation.

Il donne une mesure de l'importance relative d'un effectif pour une unité spatiale, comparée à son poids dans les autres unités spatiales.

Le quotient de localisation est un outil d'analyse spatiale, car il permet de caractériser le degré de concentration d'une sous-population dans une unité spatiale en le comparant à toutes les autres unités spatiales d'un même ensemble territorial.

Il permet de mener cette comparaison en faisant abstraction des inégalités de poids entre les unités spatiales et les différentes catégories.

Quotient de localisation

Présentation

	C1	C2	C3	Total
B1	x11	x12	x13	x1●
B2	x21	x22	x23	x2●
B3	x31	x32	x33	x3●
Total	x●1	x●2	x●3	x●●

$$Q(x[ij]) = (x[ij]/x[.j])/(x[i.]/x[..]) = (x[ij] \times x[..])/(x[i.] \times x[.j])$$

Quotient de localisation

Exemple

Branche / Zone	B1	B2	B3	Total
Z1	48	325	287	660
Z2	27	185	148	360
Z3	45	90	45	180
Total	120	600	480	1200

Quotient de localisation

Exemple

Branche / Zone	B1	B2	B3	Total
Z1	48	325	287	660
Z2	27	185	148	360
Z3	45	90	45	180
Total	120	600	480	1200

$$Q(X[z2b1]) = \frac{(27/120)}{(360/1200)} = \frac{0{,}225}{0{,}300} = 0,75$$

Quotient de localisation

Exemple

Branche / Zone	B1	B2	B3
Z1	0,727	0,985	1,087
Z2	0,750	1,028	1,028
Z3	2,500	1,000	0,625

Alternative

Et quand le tableau n'est pas un tableau de contingence ? La standardisation

Name	DEM1	DEM2	DEM3	ECO1	ECO2	ENV1	ENV2
France	107	303	12	26200	19100	6.6	0.35
Italy	192	529	9	20100	18000	7.6	0.42
Spain	78	205	9	14500	14100	6.4	0.45
Algeria	13	385	30	1500	3000	3.3	1.12
Libya	3	238	28	2000	4800	8.8	1.83
Morocco	63	294	23	1300	3600	1.1	0.29
Tunisia	58	194	22	2100	5300	1.8	0.34
West. Medit.	**38**	**310**	**15**	**14800**	**13000**	**5.5**	**0.42**

Definition of variables	
DEM1	Gross population density in inh/km2 (POP/SUP)
DEM2	Net population density in inh/km2 (POP/AGR)
DEM3	Birth rate (BIR/POP)
ECO1	GNP in $ per inhabitant (GNP/POP)
ECO2	GDP in p.p.a per inhabitant (GDP/POP)
ENV1	CO2 in tons per inhabitant (CO2/POP)
ENV2	CO2 in kg per $ of GDP (CO2/GDP)

Alternative

Et quand le tableau n'est pas un tableau de contingence ? La standardisation

$$X'(i) = (X(i) - \bar{X}(i))/\sigma(x)$$

Name	DEM1	DEM2	DEM3	ECO1	ECO2	ENV1	ENV2
France	0.6	-0.1	-0.4	1.2	0.9	0.4	-0.1
Italy	2.0	2.0	-0.7	0.5	0.8	0.8	0.0
Spain	0.1	-1.0	-0.7	0.0	0.2	0.3	0.1
Algeria	-1.0	0.7	1.8	-1.4	-1.5	-0.8	1.3
Libya	-1.2	-0.7	1.6	-1.3	-1.2	1.2	2.6
Morocco	-0.2	-0.1	1.0	-1.4	-1.4	-1.6	-0.2
Tunisia	-0.3	-1.1	0.8	-1.3	-1.2	-1.3	-0.1
moyenne	**0**	**0**	**0**	**0**	**0**	**0**	**0**
écart-type	**1**	**1**	**1**	**1**	**1**	**1**	**1**

TP3

Quotient de localisation

Vous disposez d'un fichier Excel qui contient le nombre d'habitants appartenant à une catégorie socioprofessionnelle donnée pour chaque département français. On est alors en droit de se poser des questions du type : Quelles sont les « spécialités »de chacun des départements ? Le département du Nord est-il un département « ouvrier » ? Paris est-elle une ville de cadres ?

1. Pour chaque département, calculez les quotients de localisation pour les ouvriers et pour les cadres.
2. A l'aide d'un profil en ligne, déterminez le département français le plus "spécialisé".
3. A l'aide d'un profil en colonne, déterminez la CSP la plus "localisée".

Corrélation et régression linéaire

Les statistiques multivariées

Définition

En statistiques, les analyses multivariées ont pour caractéristique de s'intéresser à la distribution conjointe de plusieurs variables. Les analyses bivariées sont des cas particuliers à deux variables.

Les analyses multivariées sont très diverses selon l'objectif recherché ou la nature des variables. On peut identifier deux grandes familles :

- celle des méthodes descriptives visant à structurer et résumer l'information ;
- celle des méthodes explicatives visant à expliquer une ou des variables dites "dépendantes" (variables à expliquer) par un ensemble de variables dites "indépendantes" (variables explicatives).

Corrélation et régression linéaire

Définitions

Corrélation

Etudier la corrélation entre deux ou plusieurs variables, c'est étudier l'intensité de la liaison qui peut exister entre ces variables.

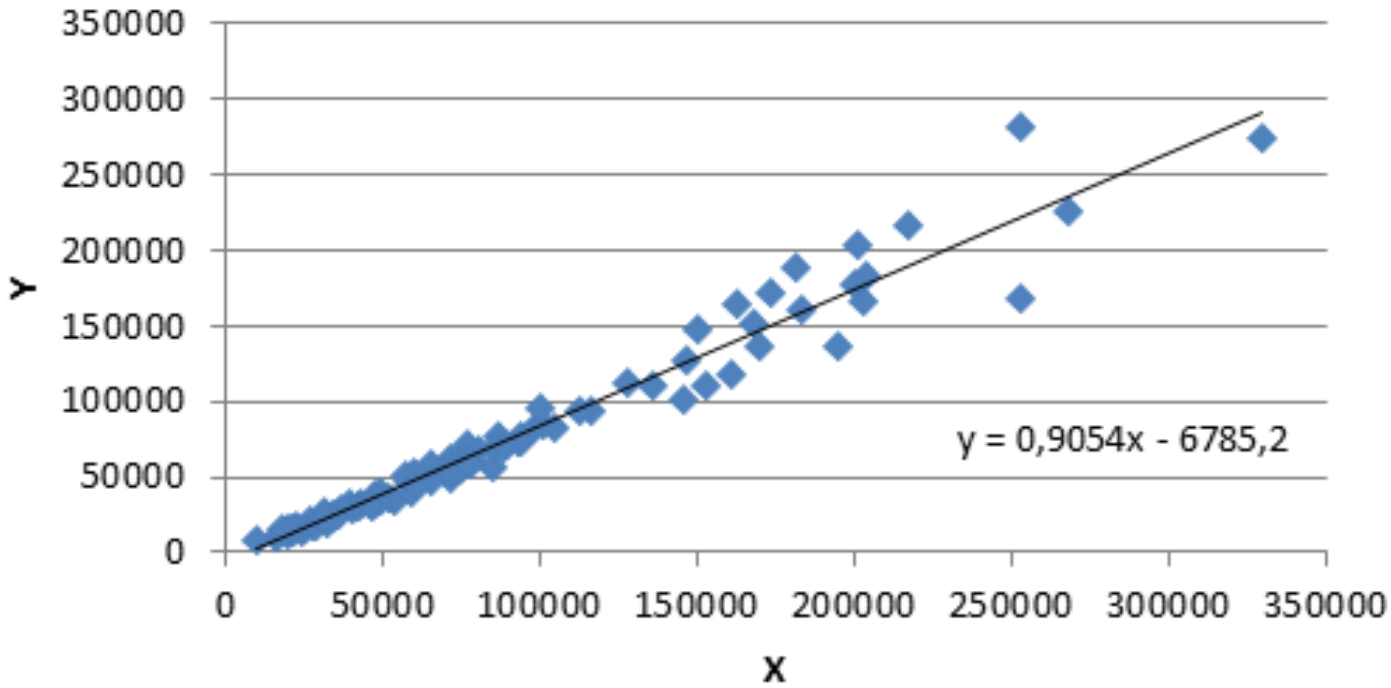
Régression linéaire

La régression est un ensemble de méthodes statistiques très utilisées pour analyser la relation d'une variable par rapport à une ou plusieurs autres.

Corrélation et régression linéaire

Des graphiques plutôt que des définitions ou des calculs pour comprendre

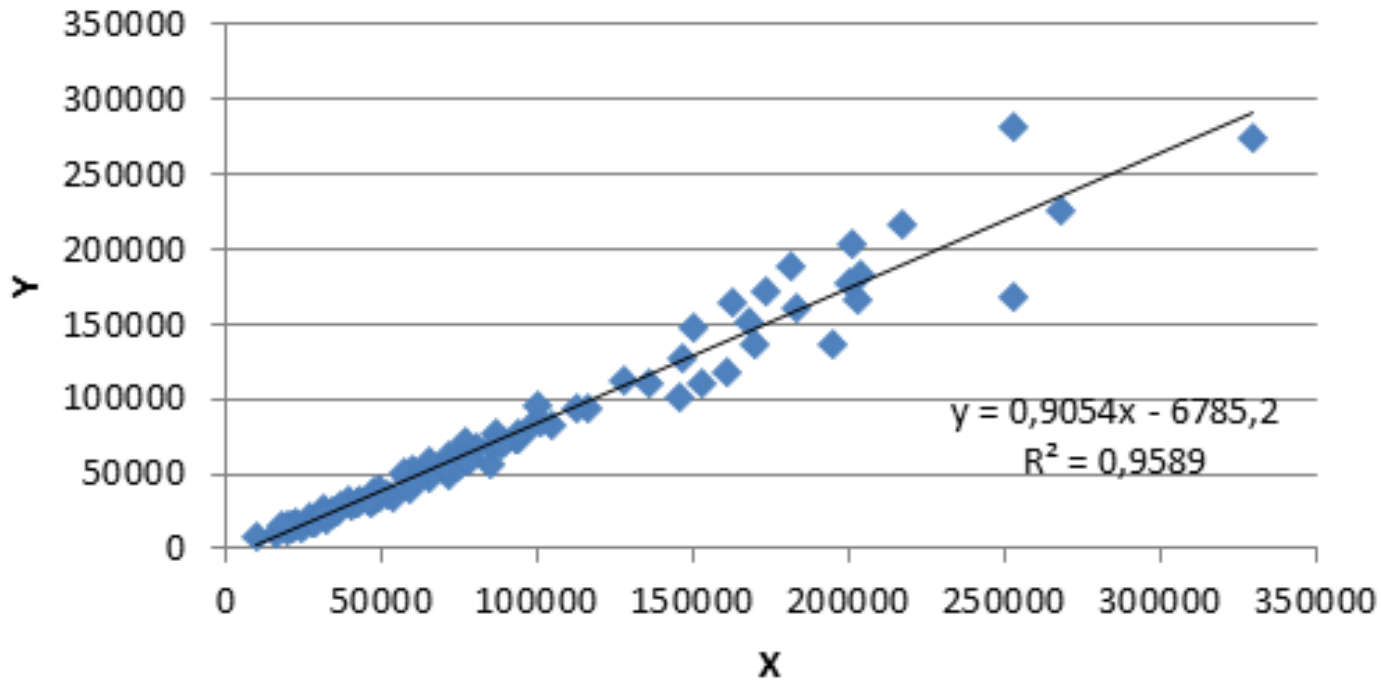
Pour faire simple, lorsque l'on étudie deux variables quantitatives, on peut produire un "nuage de points", une régression linéaire vise alors à résumer ce nuage de points par une forme plus simple à interpréter : une droite.



Corrélation et régression linéaire

Des graphiques plutôt que des définitions ou des calculs pour comprendre

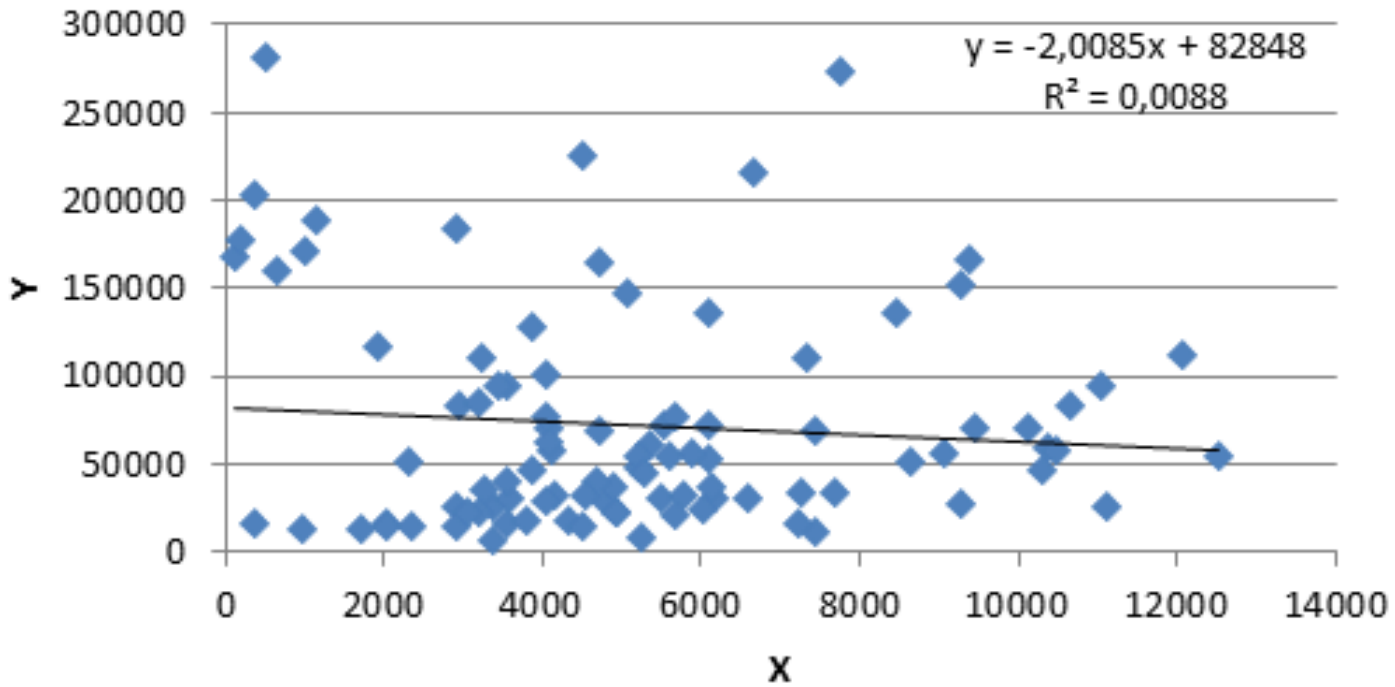
C'est le coefficient de corrélation ou le coefficient de détermination qui nous permet de dire si cette régression est "juste" :



Corrélation et régression linéaire

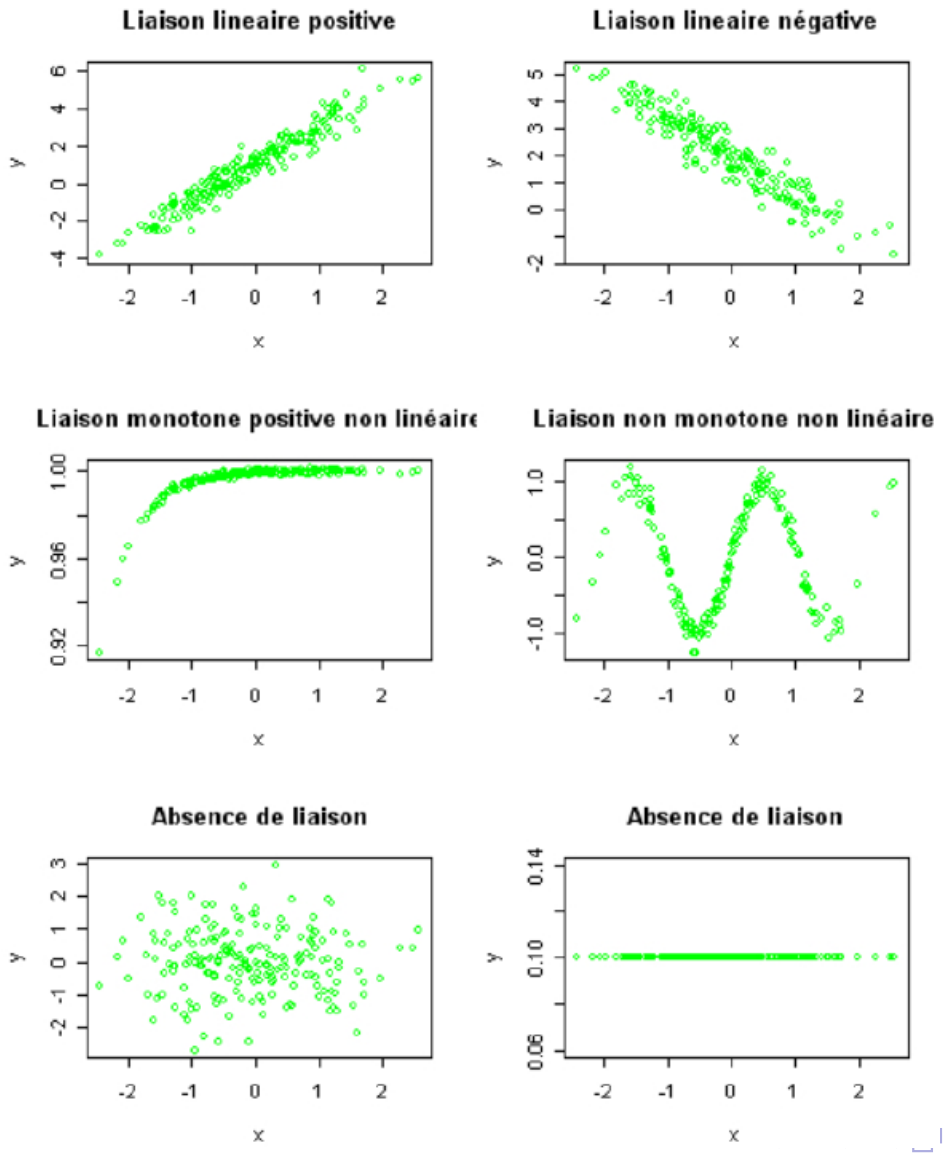
Des graphiques plutôt que des définitions ou des calculs pour comprendre

C'est le coefficient de corrélation ou le coefficient de détermination qui nous permet de dire si cette régression est "juste" ou pas du tout :



Corrélation et régression linéaire

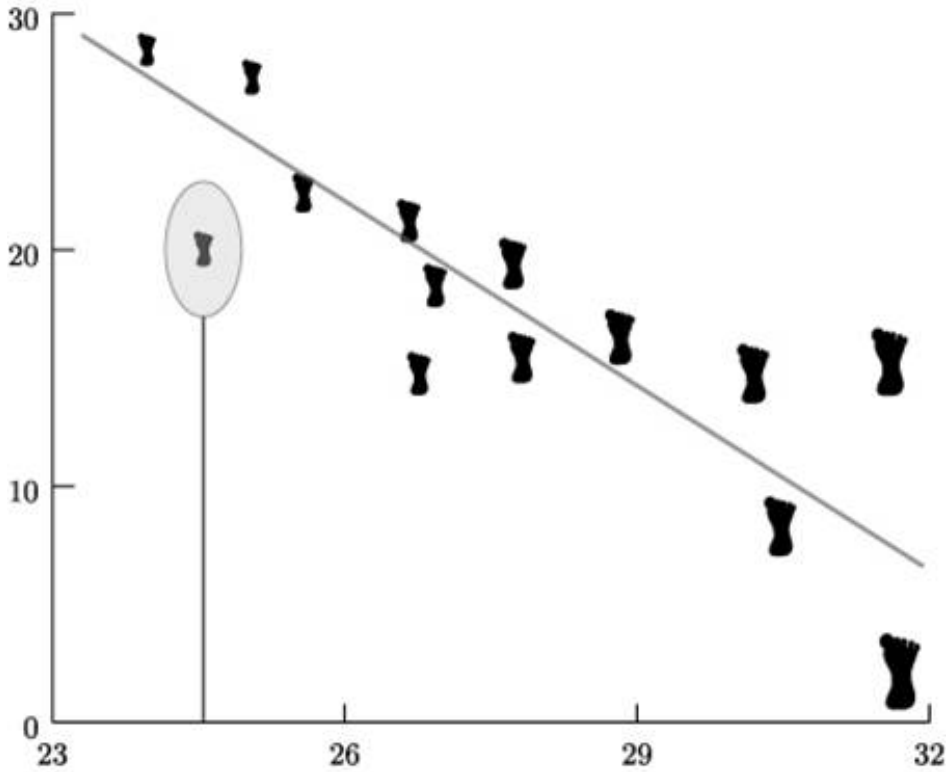
Différentes formes de nuages de points



Corrélation et régression linéaire

Les pièges à éviter : Des relations problématiques

Nombre de fautes d'orthographe en fonction de la pointure. Les élèves ayant les plus grands pieds font moins de fautes.



Corrélation et régression linéaire

Les pièges à éviter : Attention à l'erreur écologique

En géographie, l'étude des corrélations se fait souvent à travers l'analyse d'un ensemble de lieux.

Lorsque les variables décrivant ces lieux sont des attributs sociaux décrivant les habitants, il faut toujours faire attention au fait qu'une corrélation établie au niveau des lieux n'implique pas forcément une corrélation au niveau des individus.

Une étude menée au niveau des individus (sociologique) peut montrer que le taux de criminalité est plus élevé chez les autochtones que chez les étrangers.

Pourtant, cette étude au niveau des quartiers (géographique) peut montrer une corrélation parfaite entre la proportion d'étrangers des quartiers et leur taux de criminalité.

TP4

Corrélation et régression linéaire

1. Calculez la corrélation entre la variable Employé et celle de Profession Intermédiaire. Faites de même entre la variable Employé et celle d'Agriculteur. Commentez les résultats obtenus.
2. Calculez les résidus de la relation Employé, Profession Intermédiaire.
3. Représentez ces résidus sur une carte après les avoir interprétés.
4. Calculez la corrélation entre la variable Tx_Employé et celle de Tx_Profession Intermédiaire. Commentez les résultats obtenus.

Tableau de contigence et valeurs théoriques

Association et indépendance plutôt que corrélation

Lorsque l'on étudie des variables qualitatives, on comprend bien qu'il sera difficile, voire impossible, de produire un nuage de points et par conséquent de calculer des corrélations et des régressions linéaires.

Néanmoins, on peut aussi se dire qu'il faut quand même différencier les variables qualitatives nominales, de celles qui sont ordinales.

On parlera davantage d'association et d'indépendance dans le cas de variables quantitatives.

Entre deux variables qualitatives, il est par exemple possible de compter les effectifs qui correspondent aux associations (conjonctions) possibles entre les deux variables.

On parle de tableau de contingence. La notion de tableau croisé dynamique, proposée par les tableurs, est une généralisation du tableau de contingence.

Tableau de contigence et valeurs théoriques

Présentation

Les cases du tableau correspondent aux effectifs associés conjointement à une modalité de X et une modalité de Y.

Toutes les modalités de X et de Y y sont représentées.

Il est possible de calculer les valeurs totales du tableau, en ligne et en colonne, qui correspondent aux effectifs marginaux. La somme totale des effectifs correspond à l'effectif global.

A partir des effectifs et des effectifs marginaux, il est possible de calculer des proportions pour chaque ligne (profil en ligne) ou pour chaque colonne (profil en colonne).

La lecture du tableau de contingence sur la base des profils est très instructive, mais en tant que statisticien, il convient de caractériser la force du lien à l'aide d'indicateurs numériques et éventuellement tester si elle est significative.

Tableau de contigence et valeurs théoriques

Présentation

Y × X	x_1	$\cdots$	x_c	$\cdots$	x_C	Total
y_1	n_{11}	$\cdots$	n_{1c}	$\cdots$	n_{1C}	$n_{1.}$
$\vdots$		$\cdots$				$\vdots$
y_l	n_{l1}	$\cdots$	n_{lc}	$\cdots$	n_{lC}	$n_{l.}$
$\vdots$		$\cdots$				$\vdots$
y_L	n_{L1}	$\cdots$	n_{Lc}	$\cdots$	n_{LC}	$n_{L.}$
Total	$n_{.1}$	$\cdots$	$n_{.c}$	$\cdots$	$n_{.C}$	$n = n_{..}$

Tableau de contigence et valeurs théoriques

Exemple

Effectifs observés (Nij)									
1963	*ALIM*	*TEXT*	*BOIS*	*EDIT*	*CHIM*	*CONS*	*META*	*EQUIP*	**Total**
BULGARIE	130	128	39	14	29	47	21	151	**559**
HONGRIE	144	241	53	28	77	61	91	423	**1118**
POLOGNE	380	612	164	84	222	199	147	881	**2689**
R.D.A.	206	451	119	118	308	142	109	1056	**2509**
ROUMANIE	136	305	244	41	76	114	106	366	**1388**
TCHECO.	185	412	130	63	139	151	177	883	**2140**
YOUGOSL.	126	223	132	58	76	78	69	307	**1069**
Total	**1307**	**2372**	**881**	**406**	**927**	**792**	**720**	**4067**	**11472**

Tableau de contigence et valeurs théoriques

Deux profils possibles : le profil en ligne

Profils en ligne (Nij/Ni.)									
1963	*ALIM*	*TEXT*	*BOIS*	*EDIT*	*CHIM*	*CONS*	*META*	*EQUIP*	**Total**
BULGARIE	23%	23%	7%	3%	5%	8%	4%	27%	**100%**
HONGRIE	13%	22%	5%	3%	7%	5%	8%	38%	**100%**
POLOGNE	14%	23%	6%	3%	8%	7%	5%	33%	**100%**
R.D.A.	8%	18%	5%	5%	12%	6%	4%	42%	**100%**
ROUMANIE	10%	22%	18%	3%	5%	8%	8%	26%	**100%**
TCHECO.	9%	19%	6%	3%	6%	7%	8%	41%	**100%**
YOUGOSL.	12%	21%	12%	5%	7%	7%	6%	29%	**100%**
Total	**11%**	**21%**	**8%**	**4%**	**8%**	**7%**	**6%**	**35%**	**100%**

Tableau de contigence et valeurs théoriques

Deux profils possibles : le profil en colonne

Profils en colonne (Nij/N.j)									
1963	*ALIM*	*TEXT*	*BOIS*	*EDIT*	*CHIM*	*CONS*	*META*	*EQUIP*	**Total**
BULGARIE	10%	5%	4%	3%	3%	6%	3%	4%	**5%**
HONGRIE	11%	10%	6%	7%	8%	8%	13%	10%	**10%**
POLOGNE	29%	26%	19%	21%	24%	25%	20%	22%	**23%**
R.D.A.	16%	19%	14%	29%	33%	18%	15%	26%	**22%**
ROUMANIE	10%	13%	28%	10%	8%	14%	15%	9%	**12%**
TCHECO.	14%	17%	15%	16%	15%	19%	25%	22%	**19%**
YOUGOSL.	10%	9%	15%	14%	8%	10%	10%	8%	**9%**
Total	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**

Tableau de contigence et valeurs théoriques

Valeurs théoriques

Avec un tableau de contingence, on peut donc obtenir la valeur totale des effectifs concernés. $E = 11472$.

On peut aussi obtenir la taille d'une modalité vis-à-vis des autres pour les colonnes. $ALIM = 1307 / 11472 = 0.11$

On peut aussi obtenir la taille d'une modalité vis-à-vis des autres pour les lignes. $BULGARIE = 559 / 11472 = 0.05$

Si l'on multiplie l'ensemble de ces valeurs, on obtient une valeur théorique, qui correspond à ce que l'on pourrait obtenir si les deux variables étaient indépendantes. $11472 \times 0.11 \times 0.05 = 63$

Cette valeur correspond à ce que l'on pourrait s'attendre à obtenir si la situation était « simple » : sans dépendance, sans sur-représentation, sans sous-représentation, sans spécificité locale...

Le rapport entre la valeur réelle (130) et la valeur théorique (63), c'est ce que mesure le quotient de localisation.

Tableau de contigence et valeurs théoriques

Valeurs théoriques

Valeur Théorique	ALIM	TEXT	BOIS	EDIT	CHIM	CONS	META	EQUIP
BULGARIE	63,69	115,58	42,93	19,78	45,17	38,59	35,08	198,17
HONGRIE	127,37	231,16	85,86	39,57	90,34	77,18	70,17	396,35
POLOGNE	306,36	555,99	206,50	95,17	217,29	185,64	168,77	953,29
R.D.A.	285,85	518,77	192,68	88,79	202,74	173,22	157,47	889,48
ROUMANIE	158,13	286,99	106,59	49,12	112,16	95,82	87,11	492,07
TCHECO.	243,81	442,48	164,34	75,74	172,92	147,74	134,31	758,66
YOUGOSL.	121,79	221,03	82,09	37,83	86,38	73,80	67,09	378,98

Test du Chi-2

Principes

L'idée du chi-2 (χ^2) de Pearson est de comparer les effectifs réellement observés (o_k) avec les effectifs théoriques (e_k) si les variables X et Y étaient indépendantes.

Pour cela, cette technique s'appuie sur une mesure, appelée mesure du χ^2. La statistique du χ^2 quantifie l'écart (la distance) entre tous les effectifs observés et tous les effectifs théoriques.

$$\chi^2 = \sum_{k}^{K} \frac{(o_k - e_k)^2}{e_k}$$

Dans notre cas, la première valeur de ce calcul du χ^2 est :

$$(130 - 63,69)^2/63,69 = 69$$

Test du Chi-2

Calcul

Chi-2	ALIM	TEXT	BOIS	EDIT	CHIM	CONS	META	EQUIP
BULGARIE	69,05	1,33	0,36	1,69	5,79	1,83	5,65	11,23
HONGRIE	2,17	0,42	12,57	3,38	1,97	3,39	6,19	1,79
POLOGNE	17,70	5,64	8,75	1,31	0,10	0,96	2,81	5,48
R.D.A.	22,31	8,85	28,18	9,61	54,65	5,63	14,92	31,17
ROUMANIE	3,10	1,13	177,13	1,34	11,66	3,45	4,09	32,30
TCHECO.	14,19	2,10	7,18	2,14	6,66	0,07	13,57	20,38
YOUGOSL.	0,15	0,02	30,34	10,75	1,25	0,24	0,05	13,67
							Total	**703,82**

Cette valeur totale peut alors faire l'objet d'un test d'indépendance en s'appuyant sur une table du χ^2. Il faut pour cela définir un niveau de risque. Pour déterminer le nombre de degrés de liberté, il faut effectuer le calcul suivant où Nc est le nombre de colonnes et NI le nombre de lignes :

$$DL = (Nc - 1) \times (Nl - 1)$$

Si la valeur du χ^2 est supérieure à celle du tableau alors les deux variables sont liées. Les logiciels fournissent souvent la p-value.

Test du Chi-2

Table du χ^2

dl	.995	.990	.975	.950	.900	.750	.500	.250	.100	.050	.025	.010	.005
1	0.00	0.00	0.00	0.00	0.02	0.10	0.45	1.32	2.71	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	0.21	0.58	1.39	2.77	4.61	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	0.35	0.58	1.21	2.37	4.11	6.25	7.82	9.35	11.35	12.84
4	0.21	0.30	0.48	0.71	1.06	1.92	3.36	5.39	7.78	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	1.61	2.67	4.35	6.63	9.24	11.07	12.83	15.09	16.75
6	0.68	0.87	1.24	1.64	2.20	3.45	5.35	7.84	10.64	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	2.83	4.25	6.35	9.04	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34	10.22	13.36	15.51	17.54	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	5.90	8.34	11.39	14.68	16.92	19.02	21.66	23.59
10	2.15	2.56	3.25	3.94	4.87	6.74	9.34	12.55	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	7.58	10.34	13.70	17.28	19.68	21.92	24.72	26.75
12	3.07	3.57	4.40	5.23	6.30	8.44	11.34	14.85	18.55	21.03	23.34	26.21	28.30
13	3.56	4.11	5.01	5.89	7.04	9.30	12.34	15.98	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	10.17	13.34	17.12	21.06	23.69	26.12	29.14	31.31
15	4.60	5.23	6.26	7.26	8.55	11.04	14.34	18.25	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	11.91	15.34	19.37	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	12.79	16.34	20.49	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.86	13.68	17.34	21.60	25.99	28.87	31.53	34.81	37.15
19	6.84	7.63	8.91	10.12	11.65	14.56	18.34	22.72	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	15.45	19.34	23.83	28.41	31.41	34.17	37.56	40.00
21	8.03	8.90	10.28	11.59	13.24	16.34	20.34	24.93	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	17.24	21.34	26.04	30.81	33.93	36.78	40.29	42.80
23	9.26	10.19	11.69	13.09	14.85	18.14	22.34	27.14	32.01	35.17	38.08	41.64	44.18
24	9.88	10.86	12.40	13.85	15.66	19.04	23.34	28.24	33.20	36.42	39.37	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	19.94	24.34	29.34	34.38	37.65	40.65	44.32	46.93
26	11.16	12.20	13.84	15.38	17.29	20.84	25.34	30.43	35.56	38.89	41.92	45.64	48.29
27	11.80	12.88	14.57	16.15	18.11	21.75	26.34	31.53	36.74	40.11	43.20	46.96	49.64
28	12.46	13.56	15.31	16.93	18.94	22.66	27.34	32.62	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	23.57	28.34	33.71	39.09	42.56	45.72	49.59	52.34
30	13.78	14.95	16.79	18.49	20.60	24.48	29.34	34.80	40.26	43.77	46.98	50.89	53.67
40	20.67	22.14	24.42	26.51	29.06	33.67	39.34	45.61	51.80	55.75	59.34	63.71	66.80
50	27.96	29.68	32.35	34.76	37.69	42.95	49.34	56.33	63.16	67.50	71.42	76.17	79.52
60	35.50	37.46	40.47	43.19	46.46	52.30	59.34	66.98	74.39	79.08	83.30	88.40	91.98
70	43.25	45.42	48.75	51.74	55.33	61.70	69.34	77.57	85.52	90.53	95.03	100.44	104.24
80	51.14	53.52	57.15	60.39	64.28	71.15	79.34	88.13	96.57	101.88	106.63	112.34	116.35
90	59.17	61.74	65.64	69.13	73.29	80.63	89.33	98.65	107.56	113.14	118.14	124.13	128.32
100	67.30	70.05	74.22	77.93	82.36	90.14	99.33	109.14	118.49	124.34	129.56	135.82	140.19

Test du Chi-2

Conclusion

Dans notre exemple, le nombre de degrés de liberté est de : (8-1) $\times$ (7-1) = 42. D'après la table du χ^2, pour un risque de 5 % et un nombre de degrés de liberté de 42, la valeur de référence est comprise entre 55,75 et 67,50.

La valeur du χ^2 est donc très largement supérieure à la valeur de référence. La localisation et la production sont liées.

Attention

Facile à utiliser le test du χ^2 doit en théorie remplir certaines conditions d'application : un effectif global suffisant (>20), peu d'effectifs faibles (80 % des cases $>$ 5).

Lorsque les effectifs sont très élevés, le test du χ^2 aboutit presque systématiquement au rejet de l'hypothèse d'indépendance. Un petit écart, aussi infime soit-il, se répercute fortement sur la statistique.

TP5

Test du Chi-2

A partir des données Excel des CSP des départements de France métropolitaine, déterminez s'il y a dépendance ou indépendance entre la répartition des CSP et leur localisation à l'aide d'un test du chi-2.

La classification ascendante hiérarchique

Présentation

La classification ascendante hiérarchique (CAH) est une méthode de classification multivariée itérative dont le principe est simple :

- On commence par calculer la dissimilarité entre les objets.
- Puis on regroupe les deux objets dont le regroupement minimise un critère d'agrégation donné (ceux qui se ressemblent le plus), créant ainsi une classe comprenant ces deux objets.
- On calcule ensuite la dissimilarité entre cette classe et les autres objets en utilisant un critère d'agrégation.
- Puis on regroupe les deux objets ou classes d'objets dont le regroupement minimise le critère d'agrégation.

On continue ainsi jusqu'à ce que tous les objets soient regroupés.

La classification ascendante hiérarchique

Prenons un exemple simple

	Xi (km)	Yi (km)
Paris	600	2428
Marseille	846	1815
Saint-Etienne	760	2050
Bordeaux	369	1986
Reims	723	2474
Lyon	794	2087

$$Distance(euclidienne) = \sqrt{(X_1 - X_2)^2 + (Y_1 - Y_2)^2}$$

$$Dist_{(Paris-Marseille)} = \sqrt{(600 - 846)^2 + (2428 - 1815)^2} = 660$$

La classification ascendante hiérarchique

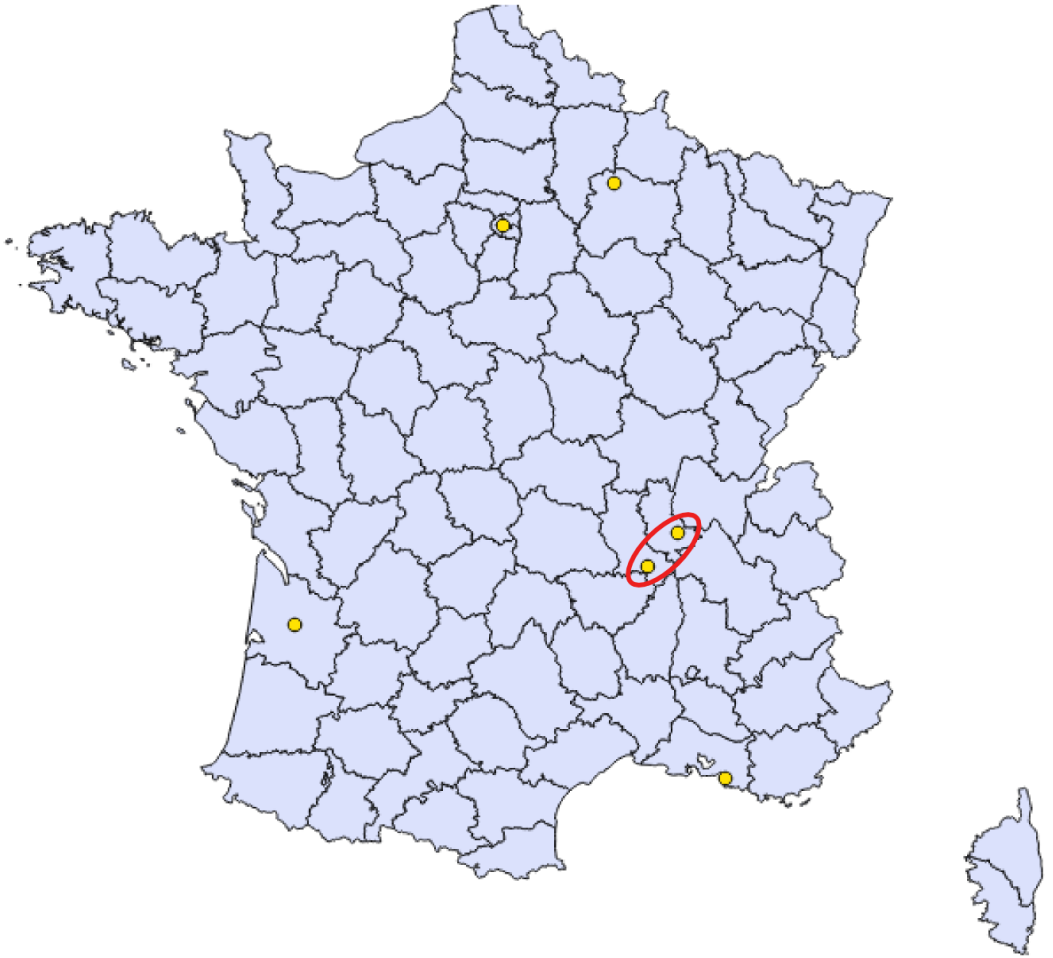
Tableau de dissimilarité (Tableau de distance)

	Paris	Marseille	Saint-Etienne	Bordeaux	Reims	Lyon
Paris	0	660	410	498	131	392
Marseille	660	0	250	506	670	276
Saint-Etienne	410	250	0	396	425	50
Bordeaux	498	506	396	0	602	436
Reims	131	670	425	602	0	393
Lyon	392	276	50	436	393	0

La classification ascendante hiérarchique

Regrouper les éléments qui sont proches

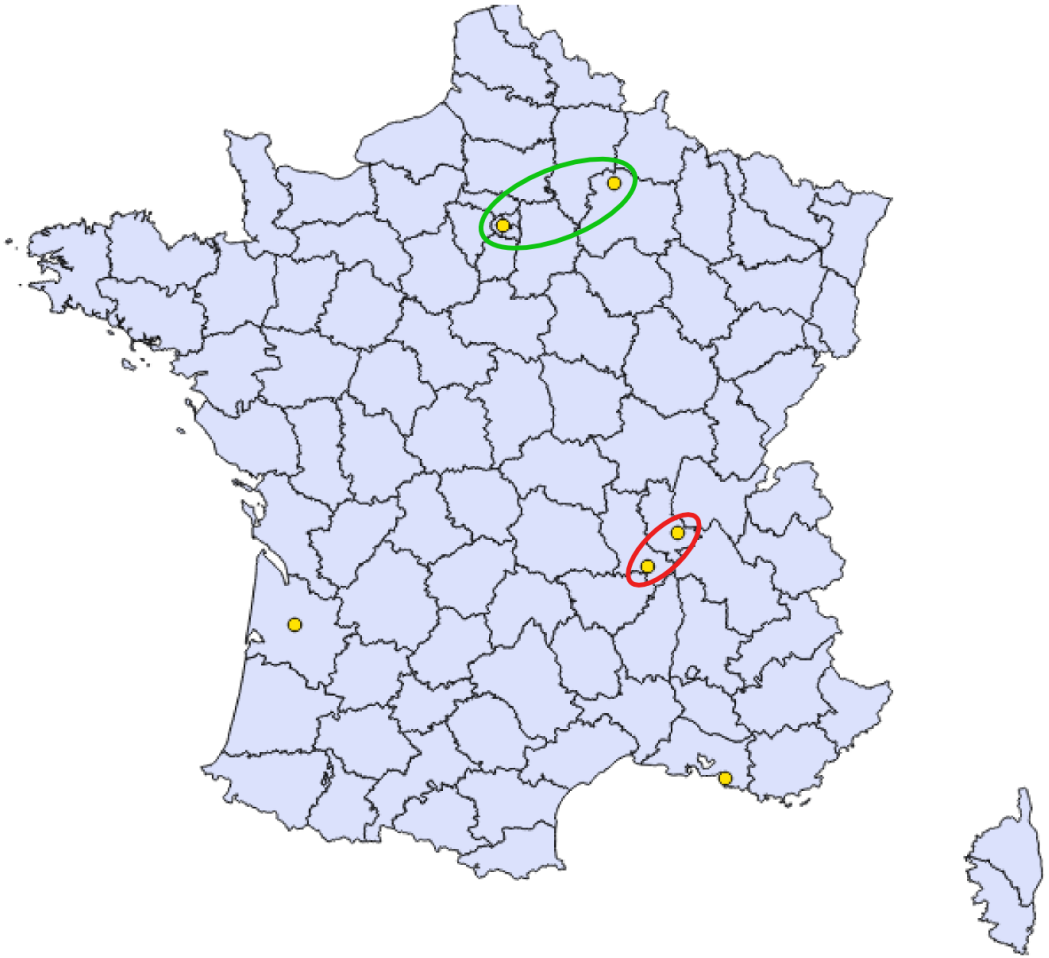
	Paris	Marseille	Saint-Etienne	Bordeaux	Reims	Lyon
Paris	0	660	410	498	131	392
Marseille	660	0	250	506	670	276
Saint-Etienne	410	250	0	396	425	50
Bordeaux	498	506	396	0	602	436
Reims	131	670	425	602	0	393
Lyon	392	276	50	436	393	0



La classification ascendante hiérarchique

Regrouper les éléments qui sont proches

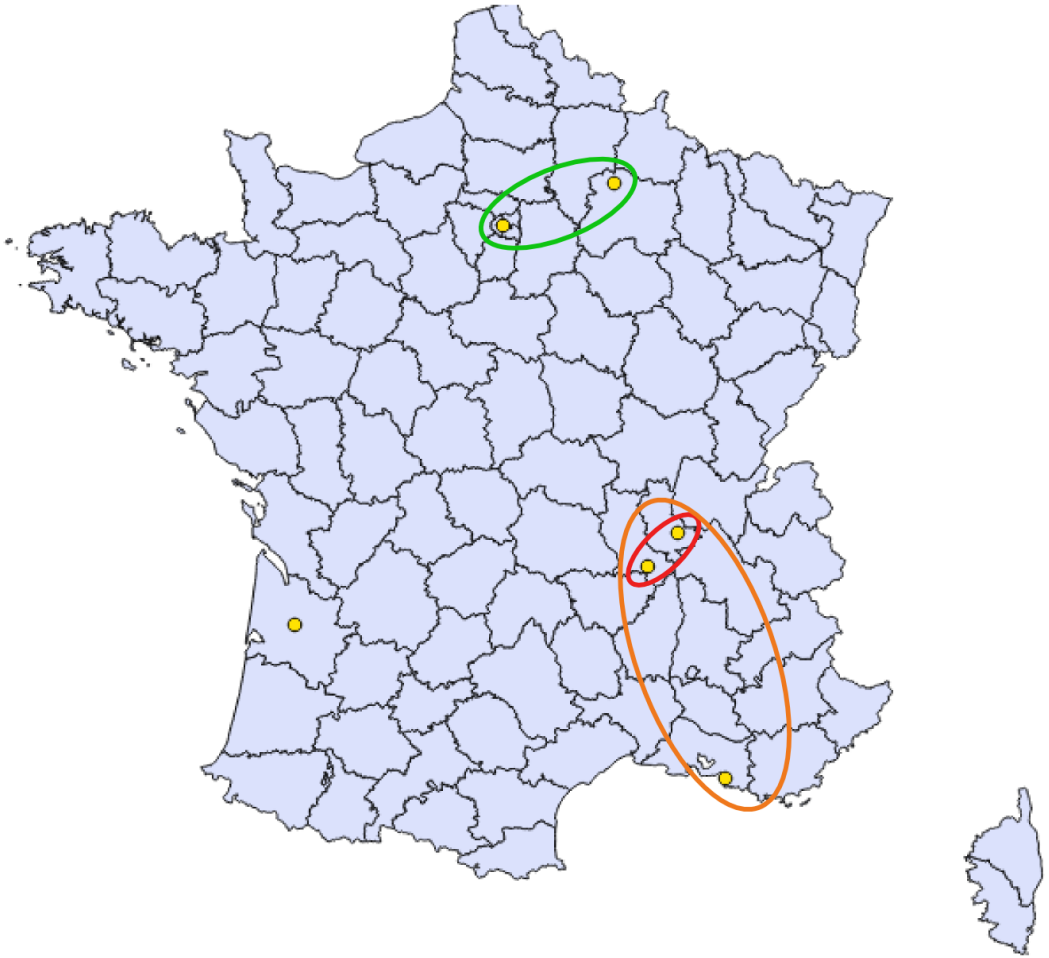
	Paris	Marseille	Saint-Etienne	Bordeaux	Reims	Lyon
Paris	0	660	410	498	131	392
Marseille	660	0	250	506	670	276
Saint-Etienne	410	250	0	396	425	50
Bordeaux	498	506	396	0	602	436
Reims	131	670	425	602	0	393
Lyon	392	276	50	436	393	0



La classification ascendante hiérarchique

Regrouper les éléments qui sont proches

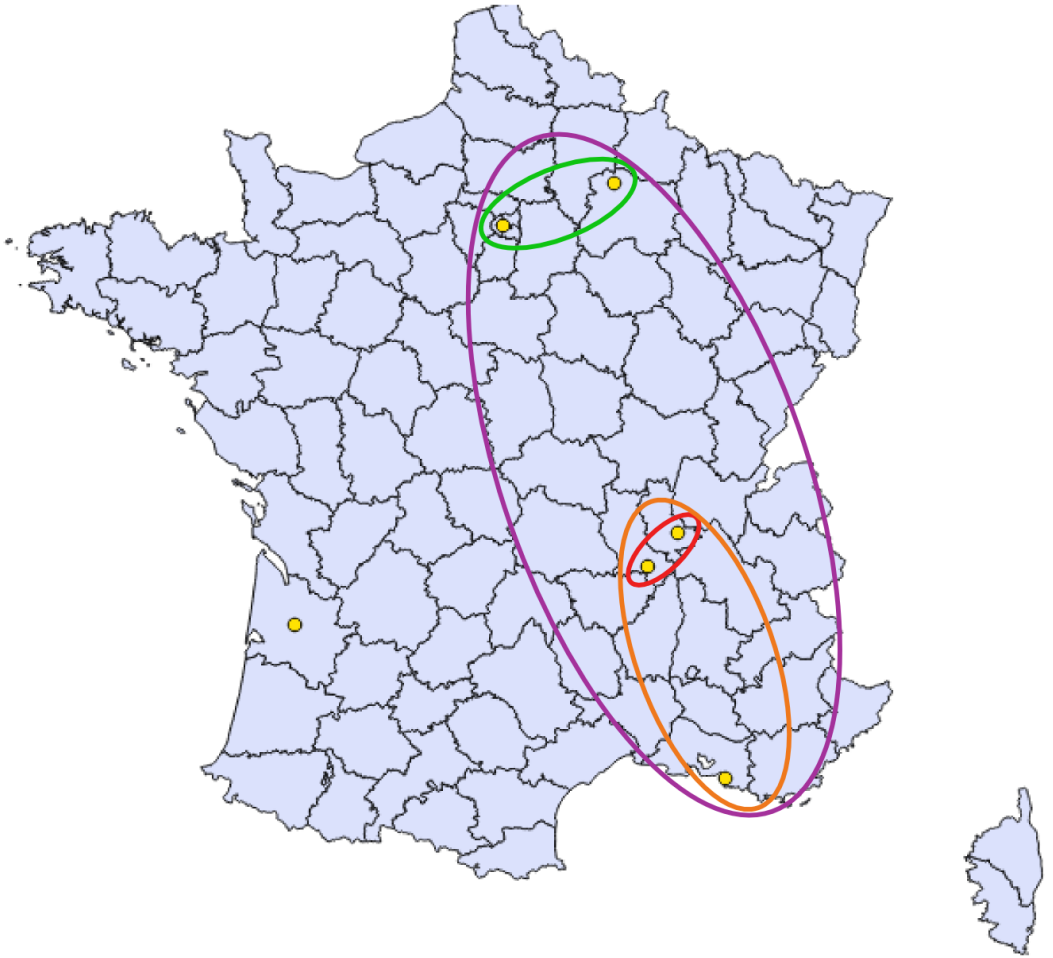
	Paris	Marseille	Saint-Etienne	Bordeaux	Reims	Lyon
Paris	0	660	410	498	131	392
Marseille	660	0	250	506	670	276
Saint-Etienne	410	250	0	396	425	50
Bordeaux	498	506	396	0	602	436
Reims	131	670	425	602	0	393
Lyon	392	276	50	436	393	0



La classification ascendante hiérarchique

Regrouper les éléments qui sont proches

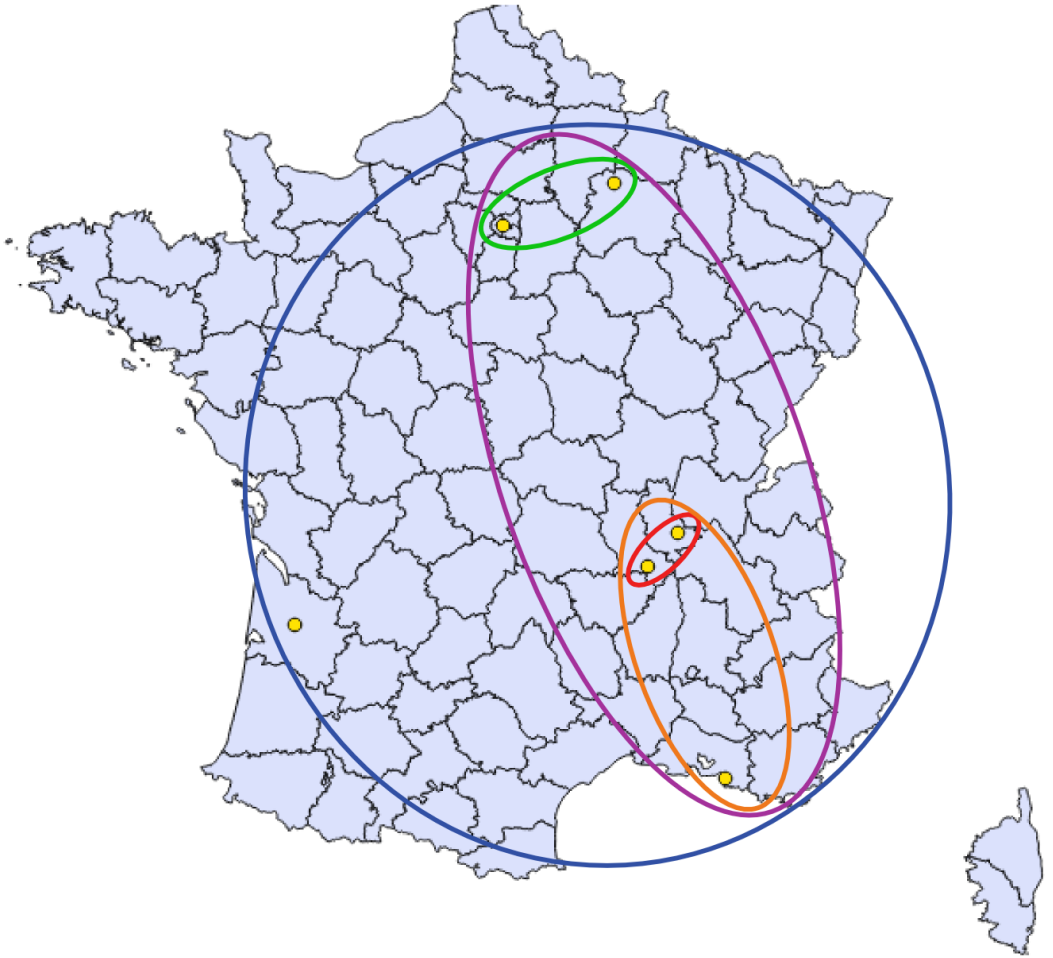
	Paris	Marseille	Saint-Etienne	Bordeaux	Reims	Lyon
Paris	0	660	410	498	131	392
Marseille	660	0	250	506	670	276
Saint-Etienne	410	250	0	396	425	50
Bordeaux	498	506	396	0	602	436
Reims	131	670	425	602	0	393
Lyon	392	276	50	436	393	0



La classification ascendante hiérarchique

Regrouper les éléments qui sont proches

	Paris	Marseille	Saint-Etienne	Bordeaux	Reims	Lyon
Paris	0	660	410	498	131	392
Marseille	660	0	250	506	670	276
Saint-Etienne	410	250	0	396	425	50
Bordeaux	498	506	396	0	602	436
Reims	131	670	425	602	0	393
Lyon	392	276	50	436	393	0



La classification ascendante hiérarchique

Le dendrogramme

Un dendrogramme est la représentation graphique d'une classification ascendante hiérarchique.

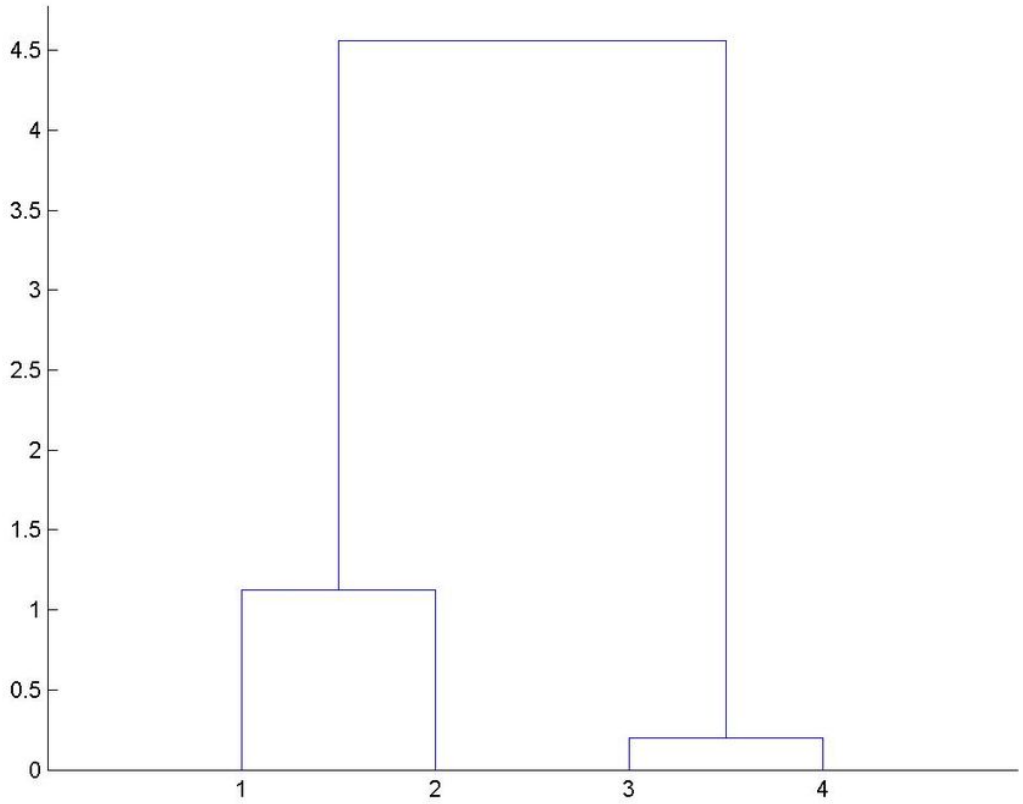
Il se présente souvent comme un arbre binaire dont les feuilles sont les individus alignés sur l'axe des abscisses.

Lorsque deux classes ou deux individus se rejoignent avec l'indice d'agrégation, des traits verticaux sont dessinés de l'abscisse des deux classes jusqu'à l'ordonnée, puis ils sont reliés par un segment horizontal.

À partir d'un indice d'agrégation, on peut tracer une droite d'ordonnée qui permet de voir une classification sur le dendrogramme.

La classification ascendante hiérarchique

Le dendrogramme : choisir un niveau de proximité pour obtenir un nombre de classes



La classification ascendante hiérarchique

Quand on a plus de deux variables c'est pas plus compliqué

	Variable 1	Variable 2	Variable 3	Variable 4
Objet 1	5	2	6	4
Objet 2	2	5	2	4

$$Dist_{(Objet1-Objet2)} = \sqrt{(5-2)^2 + (2-5)^2 + (6-2)^2 + (4-4)^2}$$

La classification ascendante hiérarchique

Plusieurs possibilités pour calculer la distance entre Paris et Marseille

Distance euclidienne : $\sqrt{(X_1 - X_2)^2 + (Y_1 - Y_2)^2}$

$De_{(P-M)} = \sqrt{(600 - 846)^2 + (2428 - 1815)^2} = 660$

Distance de Manhattan : $\mid X_1 - X_2 \mid + \mid Y_1 - Y_2 \mid$

$Dm_{(P-M)} = \mid 600 - 846 \mid + \mid 2428 - 1815 \mid = 246 + 613 = 859$

Distance de Tchebychev : $Max[(X_1 - X_2); (Y_1 - Y_2)]$

$Dt_{(P-M)} = Max[(600 - 846); (2428 - 1815)] = Max[246; 613] = 613$

La classification ascendante hiérarchique

Une précaution importante : la standardisation

Name	DEM1	DEM2	DEM3	ECO1	ECO2	ENV1	ENV2
France	107	303	12	26200	19100	6.6	0.35
Italy	192	529	9	20100	18000	7.6	0.42
Spain	78	205	9	14500	14100	6.4	0.45
Algeria	13	385	30	1500	3000	3.3	1.12
Libya	3	238	28	2000	4800	8.8	1.83
Morocco	63	294	23	1300	3600	1.1	0.29
Tunisia	58	194	22	2100	5300	1.8	0.34
West. Medit.	**38**	**310**	**15**	**14800**	**13000**	**5.5**	**0.42**

Definition of variables

DEM1	Gross population density in inh/km2 (POP/SUP)
DEM2	Net population density in inh/km2 (POP/AGR)
DEM3	Birth rate (BIR/POP)
ECO1	GNP in $ per inhabitant (GNP/POP)
ECO2	GDP in p.p.a per inhabitant (GDP/POP)
ENV1	CO2 in tons per inhabitant (CO2/POP)
ENV2	CO2 in kg per $ of GDP (CO2/GDP)

La classification ascendante hiérarchique

Une précaution importante : la standardisation

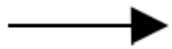
Name	DEM1	DEM2	DEM3	ECO1	ECO2	ENV1	ENV2
France	0.6	-0.1	-0.4	1.2	0.9	0.4	-0.1
Italy	2.0	2.0	-0.7	0.5	0.8	0.8	0.0
Spain	0.1	-1.0	-0.7	0.0	0.2	0.3	0.1
Algeria	-1.0	0.7	1.8	-1.4	-1.5	-0.8	1.3
Libya	-1.2	-0.7	1.6	-1.3	-1.2	1.2	2.6
Morocco	-0.2	-0.1	1.0	-1.4	-1.4	-1.6	-0.2
Tunisia	-0.3	-1.1	0.8	-1.3	-1.2	-1.3	-0.1
moyenne	**0**	**0**	**0**	**0**	**0**	**0**	**0**
écart-type	**1**	**1**	**1**	**1**	**1**	**1**	**1**

La classification ascendante hiérarchique

Une précaution importante : la standardisation

Données

Name	DEM1	DEM3
France	0.6	-0.4
Italy	2.0	-0.7
Spain	0.1	-0.7
Algeria	-1.0	1.8
Libya	-1.2	1.6
Morocco	-0.2	1.0
Tunisia	-0.3	0.8



Distance euclidienne sur variables normées

	Fra	Ita	Spa	Alg	Lib	Mor	Tun
Fra	0.0	1.5	0.6	2.7	2.6	1.5	1.5
Ita	1.5	0.0	1.9	4.0	4.0	2.8	2.8
Spa	0.6	1.9	0.0	2.8	2.6	1.7	1.6
Alg	2.7	4.0	2.8	0.0	0.3	1.2	1.2
Lib	2.6	4.0	2.6	0.3	0.0	1.2	1.2
Mor	1.5	2.8	1.7	1.2	1.2	0.0	0.1
Tun	1.5	2.8	1.6	1.2	1.2	0.1	0.0

TP6

La classification ascendante hiérarchique

1. A l'aide de Philcarto, produisez une CAH sur les CSP des départements français.
2. Commencez par interpréter les résultats obtenus pour une partition en deux classes. Allez comme cela jusqu'à une partition en six classes. Enregistrez la carte au format ai.
3. Finalisez la carte sous Illustrator.