Statistique et cartographie 2 : statistique multivariée

Serge Lhomme

Maître de conférences en Géographie
http://serge.lhomme.pagesperso-orange.fr/
serge.lhomme@u-pec.fr

19 avril 2017

Introduction

Tableau d'information géographique

Variables Caractères

	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

Entités
Individus
Objets
Unités

Valeur
Modalité

Introduction

Définition

Définition

En statistique, 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).

Introduction

Pourquoi faire de l'analyse multivariée ? Enrichir les descriptions

	Moyenne générale
Anselme	16
Adama	12
Eloise	12
Mourad	12
Ezequiel	12
Océane	8

Introduction

Pourquoi faire de l'analyse multivariée ? Enrichir les descriptions

	LV1	Math
Adama	16	8
Eloïse	8	16
Mourad	12	12
Ezequiel	12	12

Introduction

Pourquoi faire de l'analyse multivariée ? Enrichir les descriptions

	Moyenne Stat-carto2
Anselme	16
Adama	12
Eloise	12
Mourad	12
Ezequiel	12
Océane	8

Introduction

Pourquoi faire de l'analyse multivariée ? Enrichir les descriptions

	Contrôle 1 (Régression)	Contrôle 2 (CAH)
Adama	16	8
Eloïse	8	16
Mourad	12	12
Ezequiel	12	12

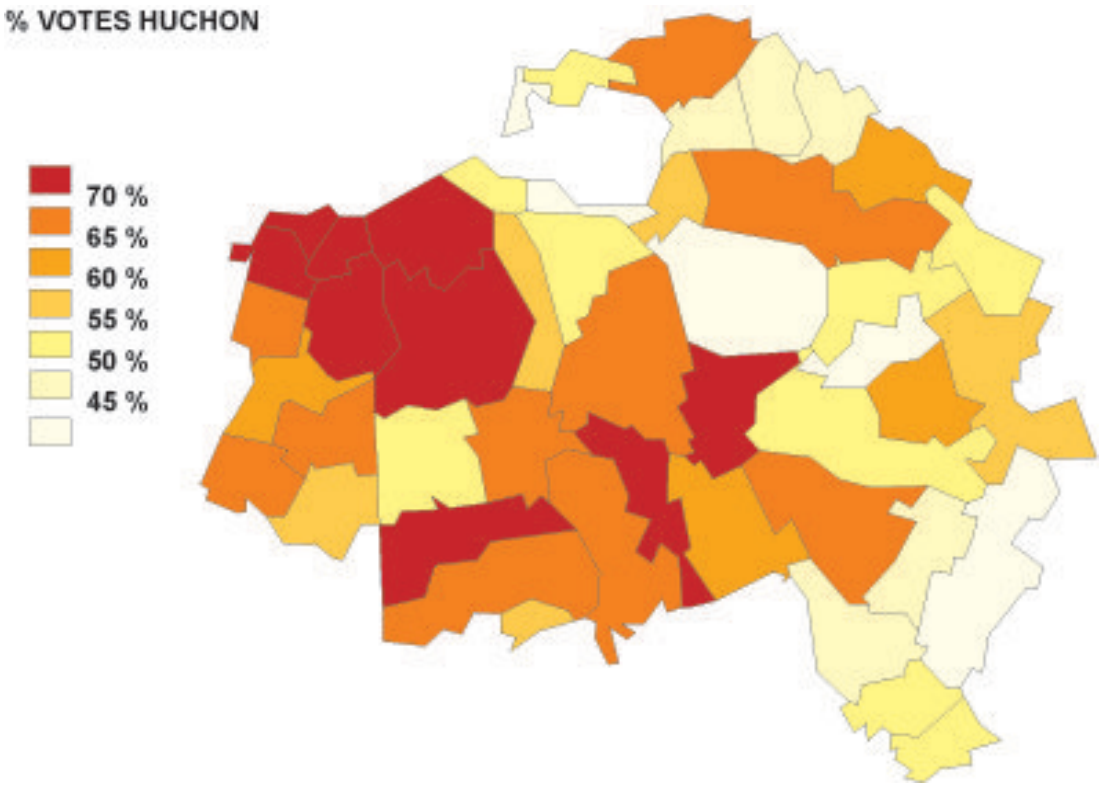
Introduction

Pourquoi faire de l'analyse multivariée ? Enrichir les descriptions

Département	Code de la commune	Libellé de la commune	Inscrits	Abstentions	% Abs/Ins	Verts	% Verts	RBM	% RBM	UMP	% UMP	FG	% FG	Centre	% Centre	PS	% PS
1	1	L'Abergement-Clémenciat	592	84	14,19	13	2,61	126	25,25	159	31,86	25	5,01	54	10,82	112	22,44
1	2	L'Abergement-de-Varey	215	36	16,74	10	5,75	38	21,84	45	25,86	15	8,62	22	12,64	39	22,41
1	4	Ambérieu-en-Bugey	8205	1698	20,69	124	1,94	1359	21,3	1603	25,12	778	12,19	543	8,51	1691	26,5
1	5	Ambérieux-en-Dombes	1152	170	14,76	16	1,66	238	24,69	323	33,51	57	5,91	97	10,06	195	20,23
1	6	Ambléon	105	17	16,19	3	3,45	25	28,74	19	21,84	9	10,34	8	9,2	21	24,14
1	7	Ambronay	1702	222	13,04	26	1,79	330	22,74	371	25,57	190	13,09	141	9,72	332	22,88
1	8	Ambutrix	549	68	12,39	8	1,69	112	23,63	109	23	51	10,76	55	11,6	125	26,37
1	9	Andert-et-Condon	269	40	14,87	5	2,22	40	17,78	69	30,67	33	14,67	27	12	38	16,89
1	10	Anglefort	681	99	14,54	7	1,23	155	27,19	130	22,81	68	11,93	47	8,25	137	24,04
1	11	Apremont	255	30	11,76	4	1,83	50	22,94	62	28,44	34	15,6	22	10,09	37	16,97
1	12	Aranc	287	55	19,16	4	1,78	42	18,67	54	24	46	20,44	19	8,44	52	23,11
1	13	Arandas	140	16	11,43	2	1,64	31	25,41	40	32,79	9	7,38	18	14,75	20	16,39
1	14	Arbent	2163	434	20,06	21	1,23	273	16,01	499	29,27	190	11,14	162	9,5	499	29,27
1	15	Arbignieu	388	54	13,92	6	1,82	60	18,18	102	30,91	41	12,42	42	12,73	69	20,91
1	16	Arbigny	309	71	22,98	3	1,29	60	25,86	71	30,6	18	7,76	28	12,07	41	17,67
1	17	Argis	296	50	16,89	8	3,31	59	24,38	51	21,07	41	16,94	15	6,2	49	20,25
1	19	Armix	37	2	5,41	3	8,57	6	17,14	9	25,71	5	14,29	2	5,71	8	22,86
1	21	Ars-sur-Formans	896	71	7,92	7	0,88	173	21,68	326	40,85	49	6,14	89	11,15	124	15,54
1	22	Artemare	813	144	17,71	13	1,96	167	25,19	167	25,19	77	11,61	55	8,3	146	22,02
1	23	Asnières-sur-Saône	48	3	6,25	0	0	15	34,09	14	31,82	3	6,82	3	6,82	8	18,18
1	24	Attignat	2052	249	12,13	30	1,7	400	22,66	523	29,63	164	9,29	224	12,69	354	20,06
1	25	Bâgé-la-Ville	2117	303	14,31	32	1,83	426	24,3	451	25,73	167	9,53	162	9,24	435	24,81
1	26	Bâgé-le-Châtel	562	101	17,97	6	1,35	76	17,04	164	36,77	37	8,3	46	10,31	97	21,75
1	27	Balan	1071	145	13,54	14	1,54	231	25,47	298	32,86	64	7,06	83	9,15	179	19,74
1	28	Baneins	434	84	19,35	9	2,67	92	27,3	115	34,12	29	8,61	22	6,53	60	17,8
1	29	Beaupont	406	76	18,72	4	1,23	88	27,16	86	26,54	34	10,49	31	9,57	68	20,99
1	30	Beauregard	520	96	18,46	10	2,4	111	26,62	122	29,26	39	9,35	33	7,91	82	19,66
1	31	Bellignat	2085	525	25,18	25	1,63	342	22,32	463	30,22	162	10,57	122	7,96	362	23,63
1	32	Béligneux	2015	368	18,26	29	1,8	453	28,12	422	26,19	137	8,5	143	8,88	358	22,22
1	33	Bellegarde-sur-Valserine	6046	1343	22,21	108	2,34	801	17,32	1029	22,25	622	13,45	386	8,35	1490	32,22
1	34	Belley	5474	986	18,01	90	2,06	883	20,18	1191	27,22	539	12,32	382	8,73	1137	25,99
1	35	Belleydoux	268	45	16,79	7	3,18	65	29,55	44	20	29	13,18	30	13,64	32	14,55
1	36	Belmont-Luthézieu	414	58	14,01	12	3,44	66	18,91	107	30,66	48	13,75	34	9,74	68	19,48

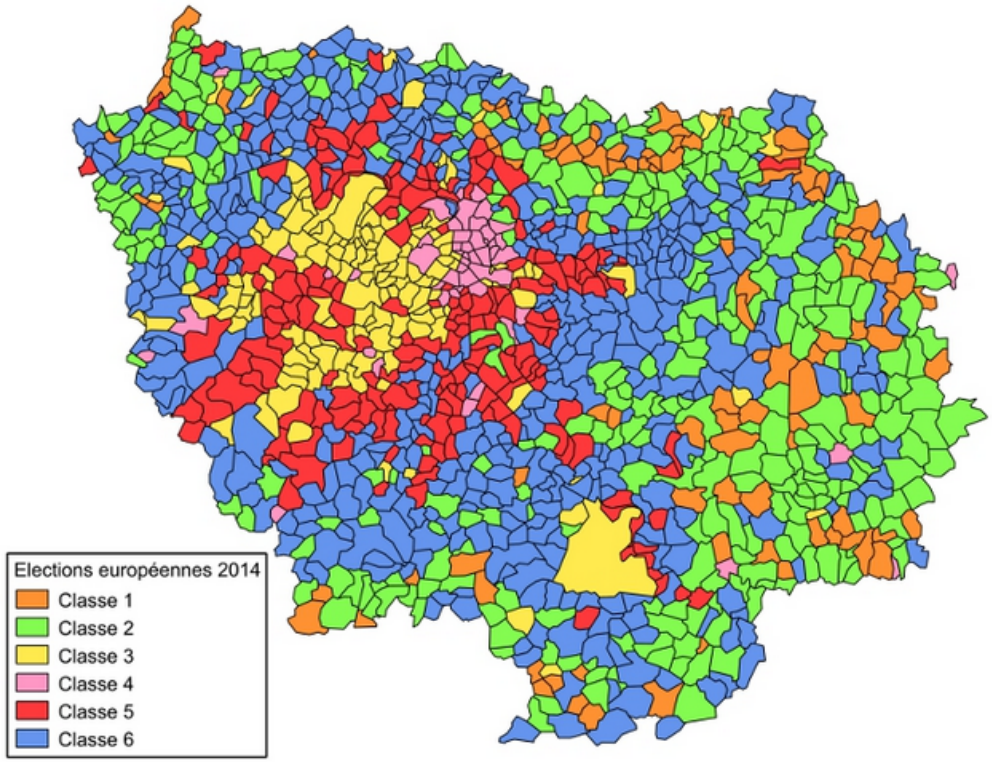
Introduction

Pourquoi faire de l'analyse multivariée ? Enrichir les descriptions



Introduction

Pourquoi faire de l'analyse multivariée ? Enrichir les descriptions

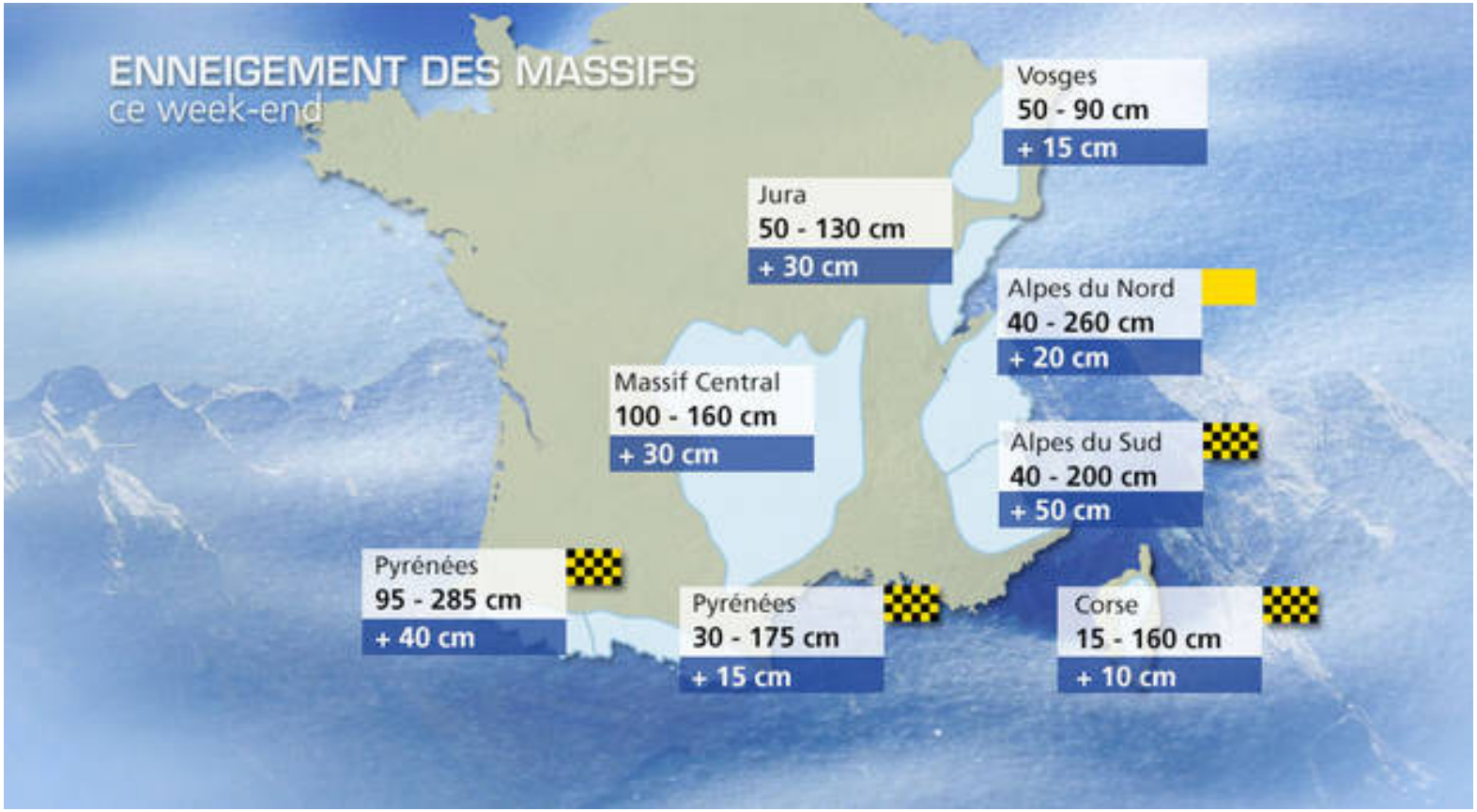


Résultat des élections européennes 2014

Classe 1 (orange) : Très gros score du FN (RBM). L'ensemble de la gauche et le centre écrasés (environ 20% des voix au total).
Classe 2 (vert) : Fort score du FN (RBM).
Classe 3 (jaune) : Fort score de l'UMP et bon score du centre.
Classe 4 (rose) : Résultat « acceptable » pour l'ensemble de la gauche grâce notamment au PG.
Classe 5 (rouge) : Résultat proche de la moyenne avec cependant le FN (RBM) limité à 20% et les « grands » partis qui n'en bénéficient pas.
Classe 6 (bleu) : Résultat proche de la moyenne avec cependant un faible total pour l'ensemble de la gauche dont le FN (RBM) profite.

Introduction

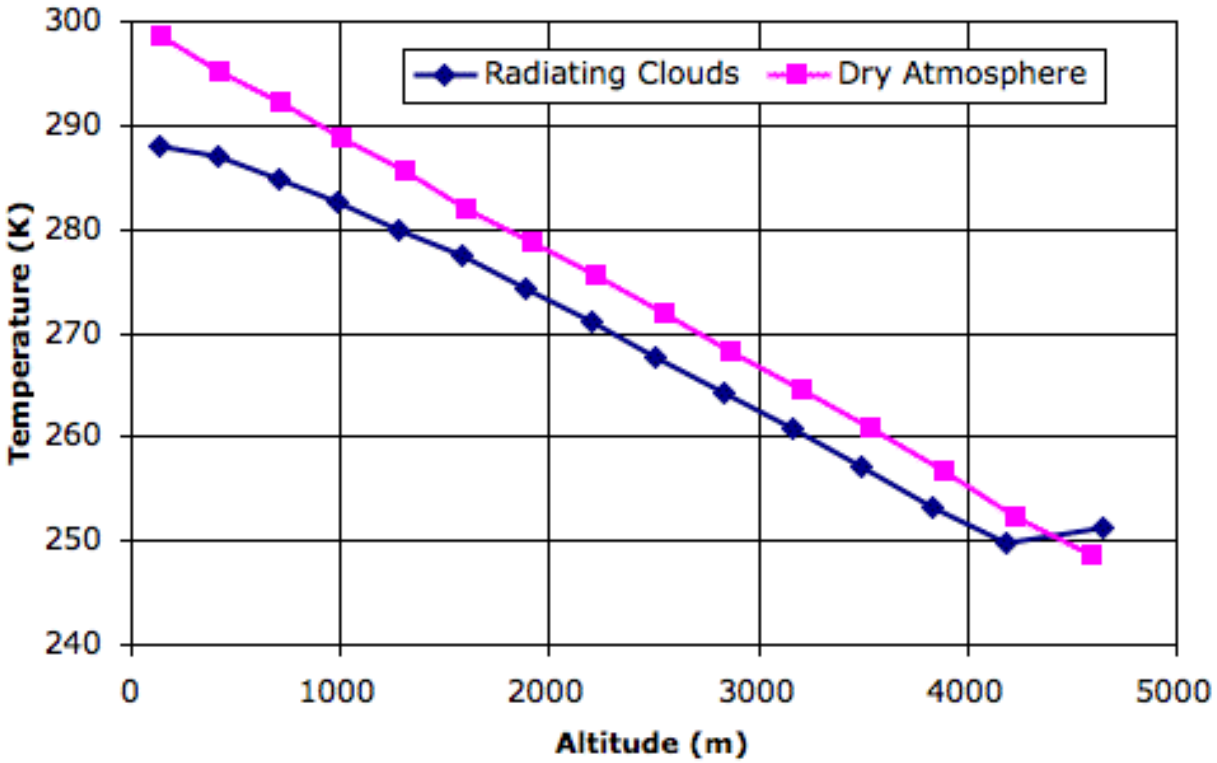
Pourquoi faire de l'analyse multivariée ? Trouver des variables explicatives



Existe-t-il une relation entre la quantité de neiges et l'altitude ?

Introduction

Pourquoi faire de l'analyse multivariée ? Trouver des variables explicatives



Existe-t-il un lien entre altitude et température ?

TP1

Étudions différentes variables indépendamment

1. Ouvrez le fichier Excel du TP1. Que contient ce fichier ?
2. Pour les variables "Agriculteur", "Profession intermédiaire" et "Employé", produisez trois cartes choroplèthes (un aplat de couleurs, un dégradé de couleurs) à l'aide de discrétisations par effectifs égaux comportant 5 classes. Attention, il faut "tricher" sur les types des variables, mais c'est pour la bonne cause !
3. Comparez les trois cartes obtenues. Que constatez vous ?
4. Quelles questions est-on en droit de se poser ?

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. Le type le plus simple de liaison est la relation affine (une droite).

Corrélation et régression linéaire

Des graphiques plutôt que des définitions ou des calculs pour comprendre

Imaginons un territoire découpé en plusieurs unités géographiques.

On connait pour chaque unité géographique les valeurs de deux variables quantitatives nommées X et Y.

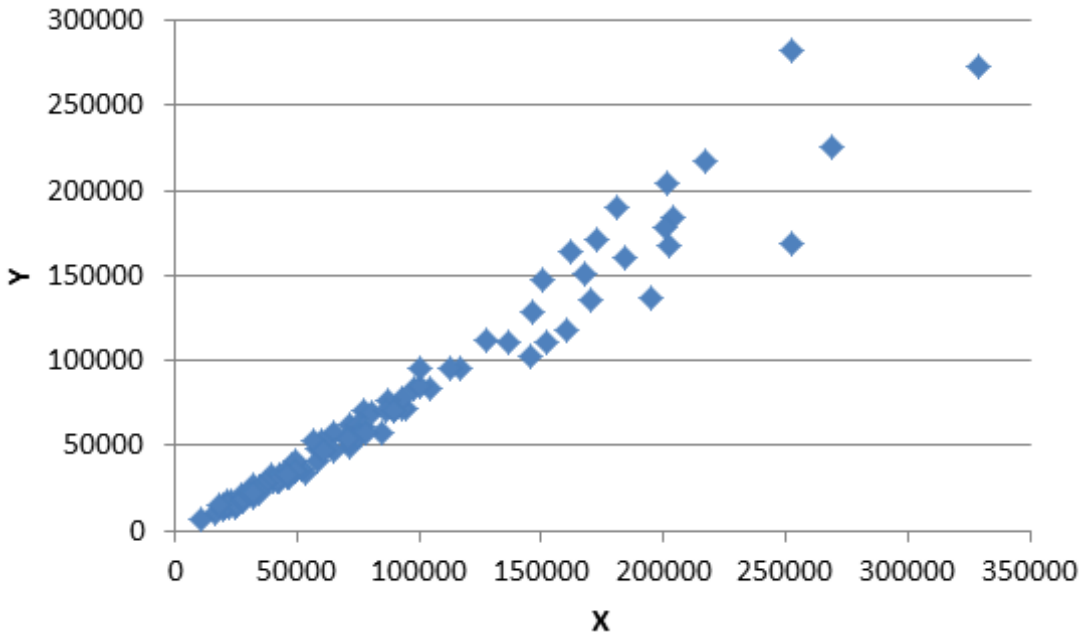
Pour connaitre le lien entre ces deux variables, on peut représenter une variable en fonction de l'autre. Par exemple, on peut représenter Y en fonction de X.

On obtient alors un nuage de points.

Corrélation et régression linéaire

Des graphiques plutôt que des définitions ou des calculs pour comprendre

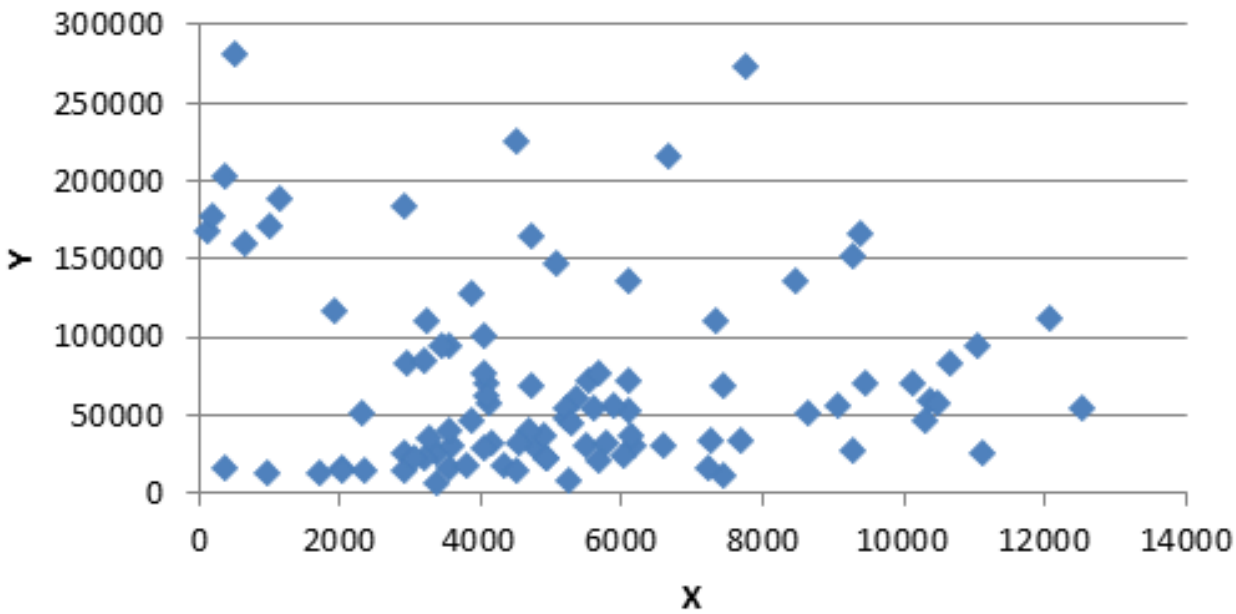
Parfois, la forme du nuage de points semble pouvoir s'apparenter à une droite :



Corrélation et régression linéaire

Des graphiques plutôt que des définitions ou des calculs pour comprendre

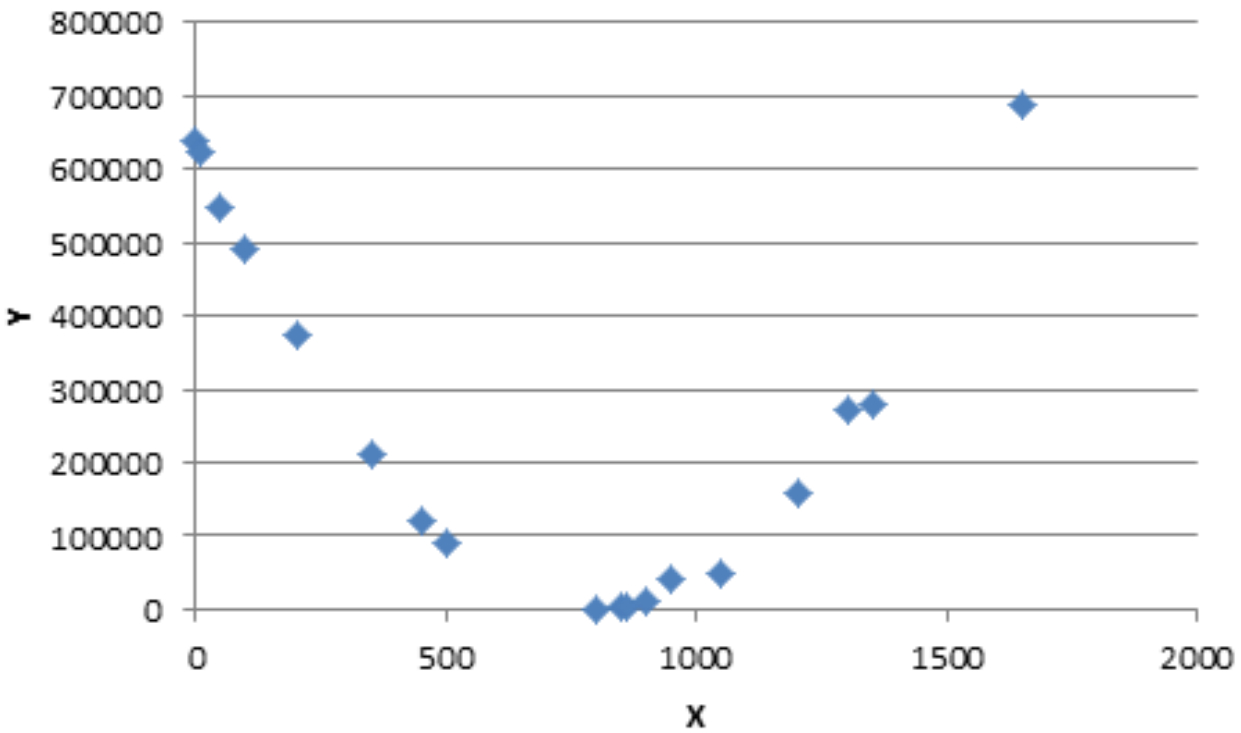
Parfois, ça ressemble à rien :



Corrélation et régression linéaire

Des graphiques plutôt que des définitions ou des calculs pour comprendre

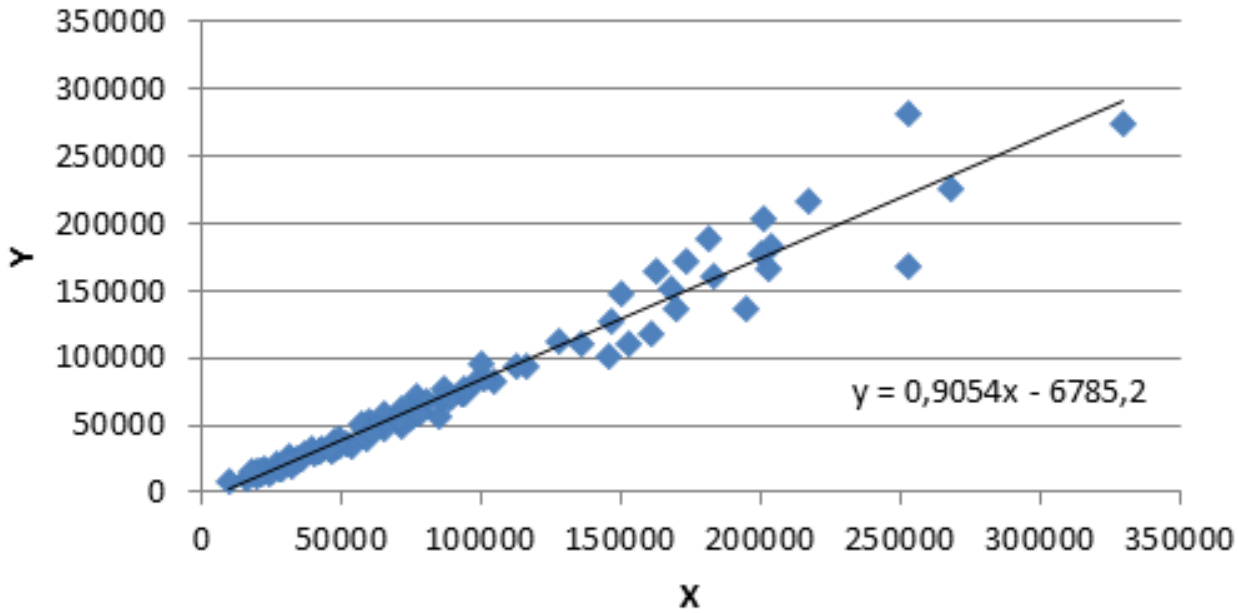
Des fois, ça ressemble plutôt à autre chose :



Corrélation et régression linéaire

Des graphiques plutôt que des définitions ou des calculs pour comprendre

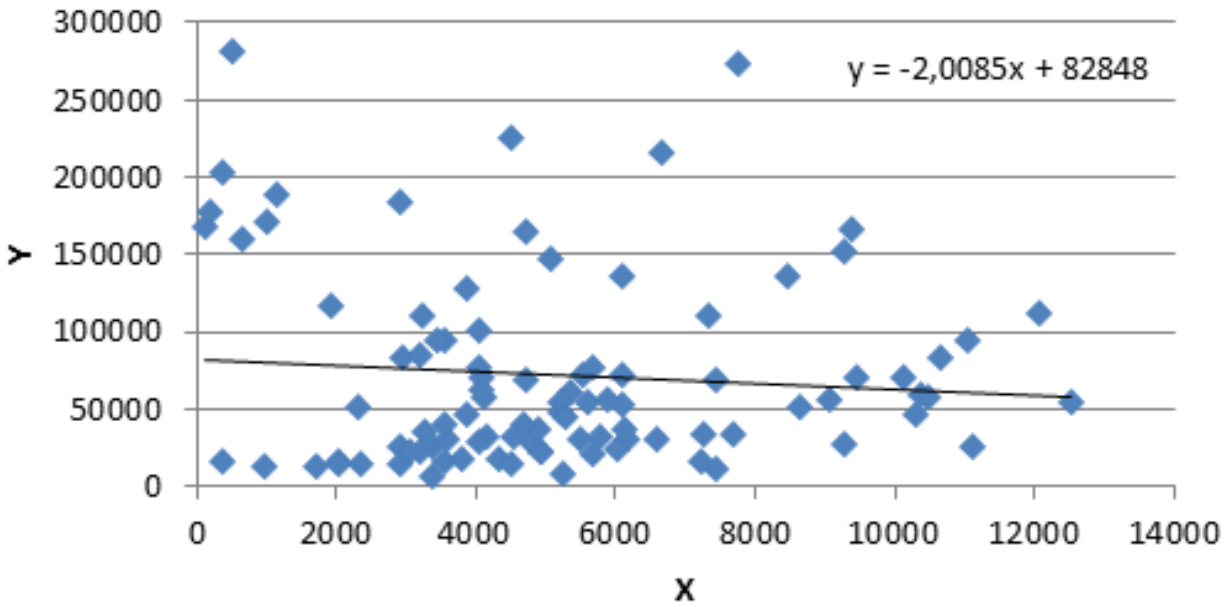
Quoi qu'il en soit, on peut toujours réduire un nuage de points sous la forme d'une droite. Quand ça marche :



Corrélation et régression linéaire

Des graphiques plutôt que des définitions ou des calculs pour comprendre

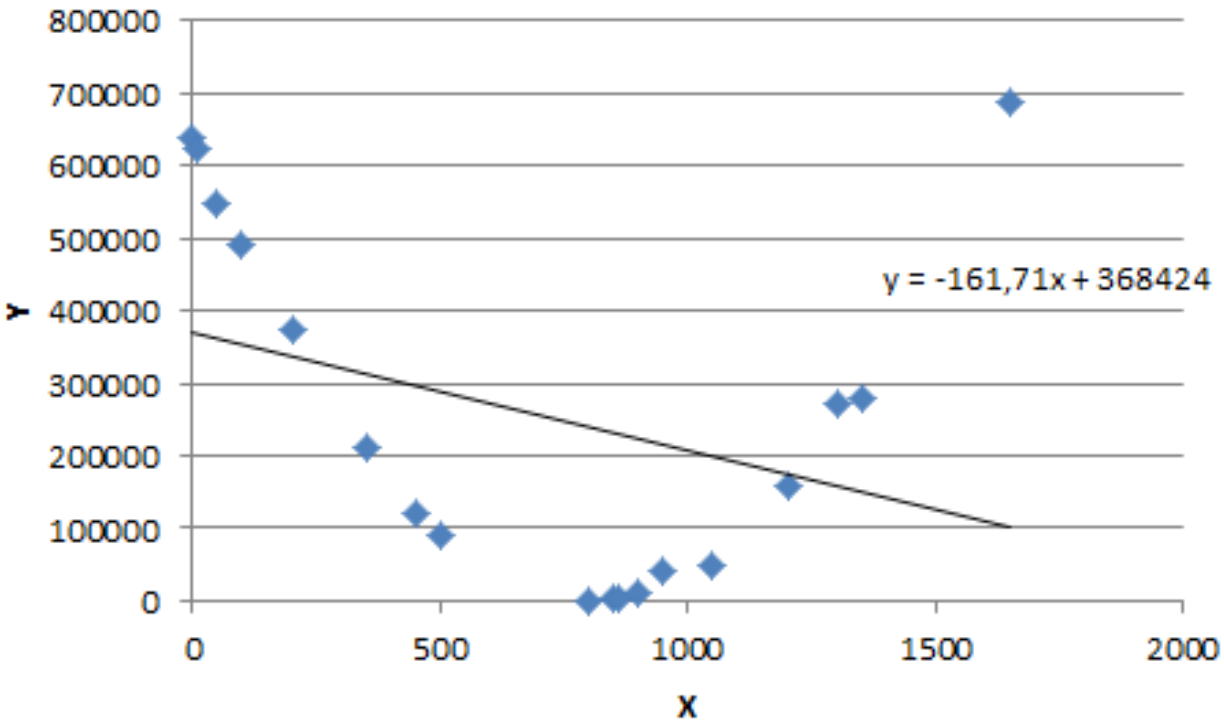
Quand ça ne marche pas :



Corrélation et régression linéaire

Des graphiques plutôt que des définitions ou des calculs pour comprendre

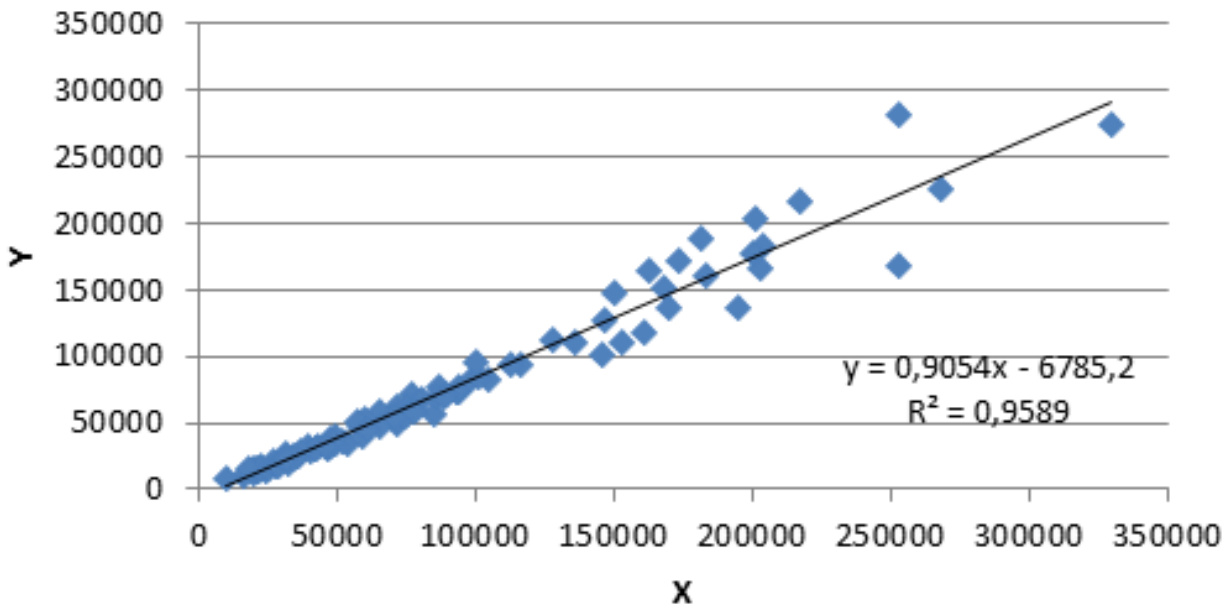
Quand il ne s'agit pas d'une droite :



Corrélation et régression linéaire

Des graphiques plutôt que des définitions ou des calculs pour comprendre

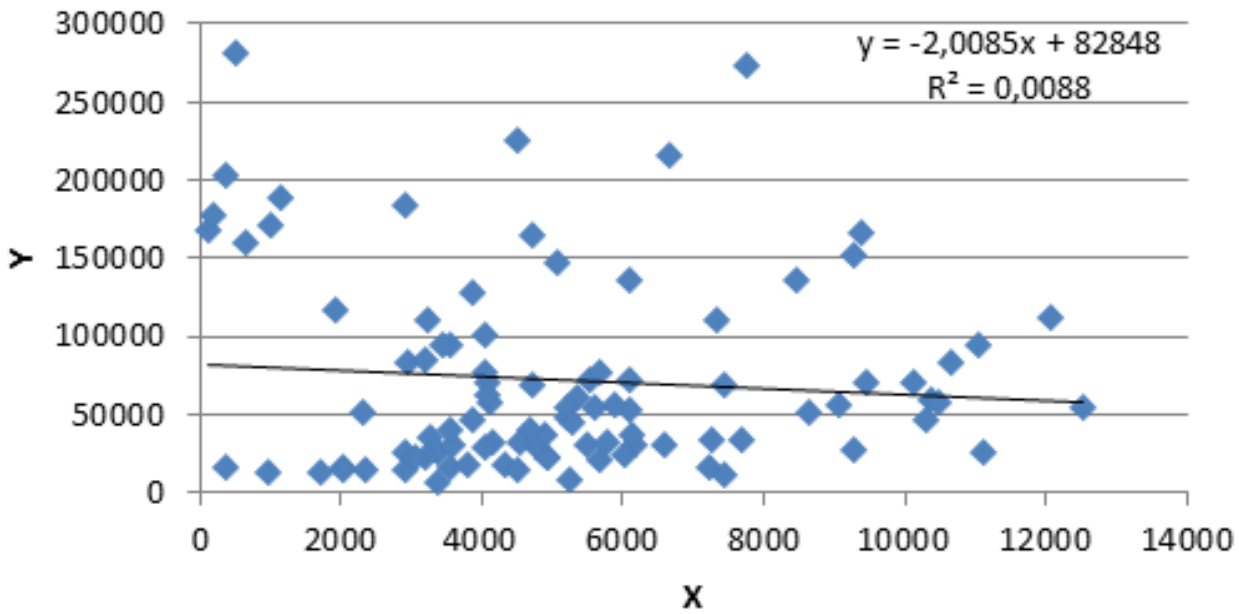
C'est le coefficient de corrélation qui nous permet de dire si cette régression est "juste" :



Corrélation et régression linéaire

Des graphiques plutôt que des définitions ou des calculs pour comprendre

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



TP2

Les variables "AGRI", "PRO INT" et "EMPLOYE" sont-elles corrélées ?

1. Ouvrez le fichier Excel du TP2, puis enregistrez le sous un nouveau nom : CSP-correlation.
2. Produisez un nuage de points avec en ordonnée la variable "Profession intermédiaire" et en abscisse la variable "Employé".
3. Ajoutez une "courbe de tendance" de type linéaire à ce nuage de points.
4. Affichez l'équation de la droite obtenue et le coefficient de corrélation.
5. Recommencez pour les variables "Employé" et "Agriculteur" et pour les variables "Agriculteur" et "Profession intermédiaire".

Corrélation et régression linéaire

Formules pour la régression linéaire

L'équation d'une droite est de type :

$$Y = aX + b$$

On obtient a (le coefficient directeur de la droite) à l'aide de la formule suivante :

$$a = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Et on obtient b à l'aide de la formule suivante :

$$b = \bar{Y} - a \times \bar{X}$$

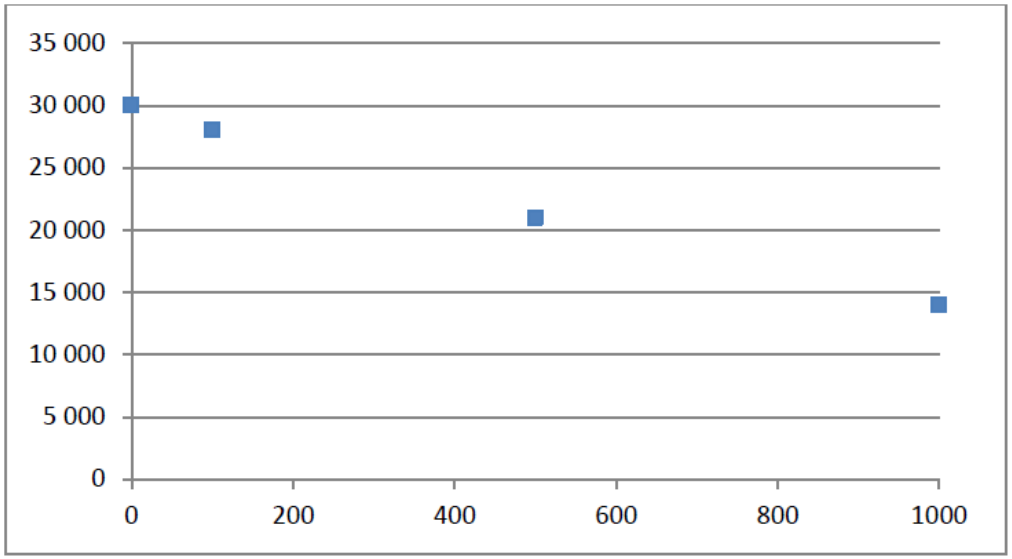
$\bar{X}$ et $\bar{Y}$ représentent respectivement les moyennes de X et de Y.

Corrélation et régression linéaire

La régression linéaire pas à pas

Pour avoir une application, prenons un cas théorique où l'on cherche à déterminer la droite qui approxime le mieux le prix médian des appartements vis-à-vis de leur distance au centre d'une ville. Il semble en effet qu'il existe une relation linaire entre ces deux variables.

Prix médian (Y)	30 000	28 000	21 000	14 000
Distance au centre (X)	0	100	500	1000



Corrélation et régression linéaire

La régression linéaire pas à pas

Xi	Yi
0	30 000
100	28 000
500	21 000
1 000	14 000

Corrélation et régression linéaire

La régression linéaire pas à pas

	Xi	Yi
	0	30 000
	100	28 000
	500	21 000
	1 000	14 000
Somme	1 600	63 000
Moyenne	$\bar{X} = 400$	$\bar{Y} = 23\,250$

Corrélation et régression linéaire

La régression linéaire pas à pas

	Xi	Yi	$Xi - \bar{X}$	$Yi - \bar{Y}$
	0	30 000	- 400	6 750
	100	28 000	- 300	4 750
	500	21 000	100	- 2 250
	1 000	14 000	600	- 9 250
Somme	1 600	63 000		
Moyenne	$\bar{X} = 400$	$\bar{Y} = 23\,250$		

Corrélation et régression linéaire

La régression linéaire pas à pas

	X_i	Y_i	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$	$(X_i - \bar{X})(Y_i - \bar{Y})$
	0	30 000	- 400	6 750	160 000		- 2 700 000
	100	28 000	- 300	4 750	90 000		- 1 425 000
	500	21 000	100	- 2 250	10 000		- 225 000
	1 000	14 000	600	- 9 250	360 000		- 5 550 000
Somme	1 600	63 000			**620 000**		**- 9 900 000**
Moyenne	$\bar{X} = 400$	$\bar{Y} = 23\,250$					

Corrélation et régression linéaire

La régression linéaire pas à pas

	Xi	Yi	Xi - $\bar{X}$	Yi - $\bar{Y}$	$(Xi - \bar{X})^2$	$(Yi - \bar{Y})^2$	$(Xi - \bar{X})(Yi - \bar{Y})$
	0	30 000	- 400	6 750	160 000		- 2 700 000
	100	28 000	- 300	4 750	90 000		- 1 425 000
	500	21 000	100	- 2 250	10 000		- 225 000
	1 000	14 000	600	- 9 250	360 000		- 5 550 000
Somme	1 600	63 000			**620 000**		**- 9 900 000**
Moyenne	$\bar{X} = 400$	$\bar{Y} = 23\,250$					

$a = - 9\,900\,000 / 620\,000 = 15.9677$

$b = 23\,250 - 15.9677 \times 400 = 29636.8$

Corrélation et régression linéaire

Formules coefficient de corrélation linéaire r

$$Cor(X, Y) = \frac{Cov(X,Y)}{\sigma_X \times \sigma_Y} = r$$

$$Cov(X, Y) = \frac{1}{N} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$$

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

$$\sigma_Y = \sqrt{\frac{1}{N} \sum_{i=1}^{n} (Y_i - \bar{Y})^2}$$

Corrélation et régression linéaire

Le coefficient de corrélation linéaire r pas à pas

	X_i	Y_i	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$	$(X_i - \bar{X})(Y_i - \bar{Y})$
	0	30 000	- 400	6 750	160 000	45 562 500	- 2 700 000
	100	28 000	- 300	4 750	90 000	22 562 500	- 1 425 000
	500	21 000	100	- 2 250	10 000	5 056 250	- 225 000
	1 000	14 000	600	- 9 250	360 000	85 562 500	- 5 550 000
Somme	1 600	63 000			620 000	158 743 750	- 9 900 000
Moyenne	$\bar{X} = 400$	$\bar{Y} = 23\,250$			155 000	39 685 937,5	- 2 475 000

Corrélation et régression linéaire

Le coefficient de corrélation linéaire r pas à pas

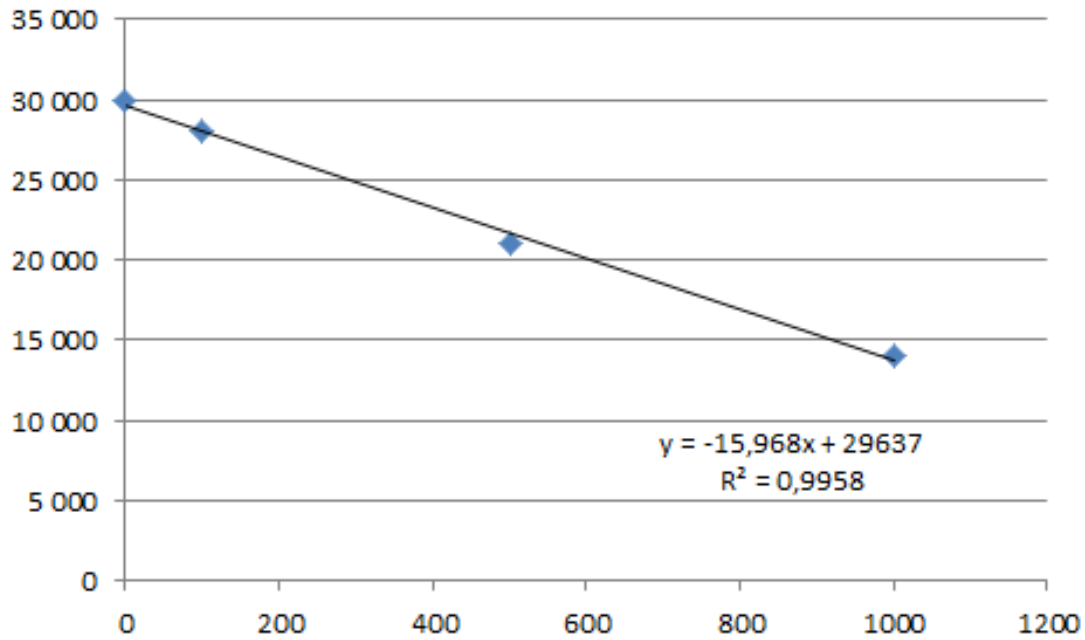
	Xi	Yi	$Xi - \bar{X}$	$Yi - \bar{Y}$	$(Xi - \bar{X})^2$	$(Yi - \bar{Y})^2$	$(Xi - \bar{X})(Yi - \bar{Y})$
	0	30 000	- 400	6 750	160 000	45 562 500	- 2 700 000
	100	28 000	- 300	4 750	90 000	22 562 500	- 1 425 000
	500	21 000	100	- 2 250	10 000	5 056 250	- 225 000
	1 000	14 000	600	- 9 250	360 000	85 562 500	- 5 550 000
Somme	1 600	63 000			620 000	158 743 750	- 9 900 000
Moyenne	$\bar{X} = 400$	$\bar{Y} = 23\,250$			155 000	39 685 937,5	- 2 475 000

$$r = \frac{-2475000}{\sqrt{155000} x \sqrt{39685937,5}} = 0,9979$$

$$r^2 = 0,9958$$

Corrélation et régression linéaire

Le coefficient de corrélation linéaire r pas à pas



$a = -\ 9\ 900\ 000\ /\ 620\ 000 = 15{,}9677$

$b = 23\ 250 - 15{,}9677 \times 400 = 29636{,}8$

$r^2 = 0,9958$

TP3

Excel ne s'est-il pas trompé ?

Reprenez votre fichier Excel CSP, sauvegardez le sous le nom CSP-corrcalcul. "EMPLOYE" sera la variable X et "PRO INT" la variable Y. Mettez le fichier Excel sous la forme suivante :

	X_i	Y_i	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$	$(X_i - \bar{X})(Y_i - \bar{Y})$
Somme							
Moyenne							

1. Complétez ce tableau
2. Déterminez les coefficients a et b de la régression linéaire.
3. Calculez le coefficent de corrélation linéaire.

Les résidus

Définition et formules

Définition

Un résidu est dans une régression le terme qui n'est pas expliqué par les autres variables.

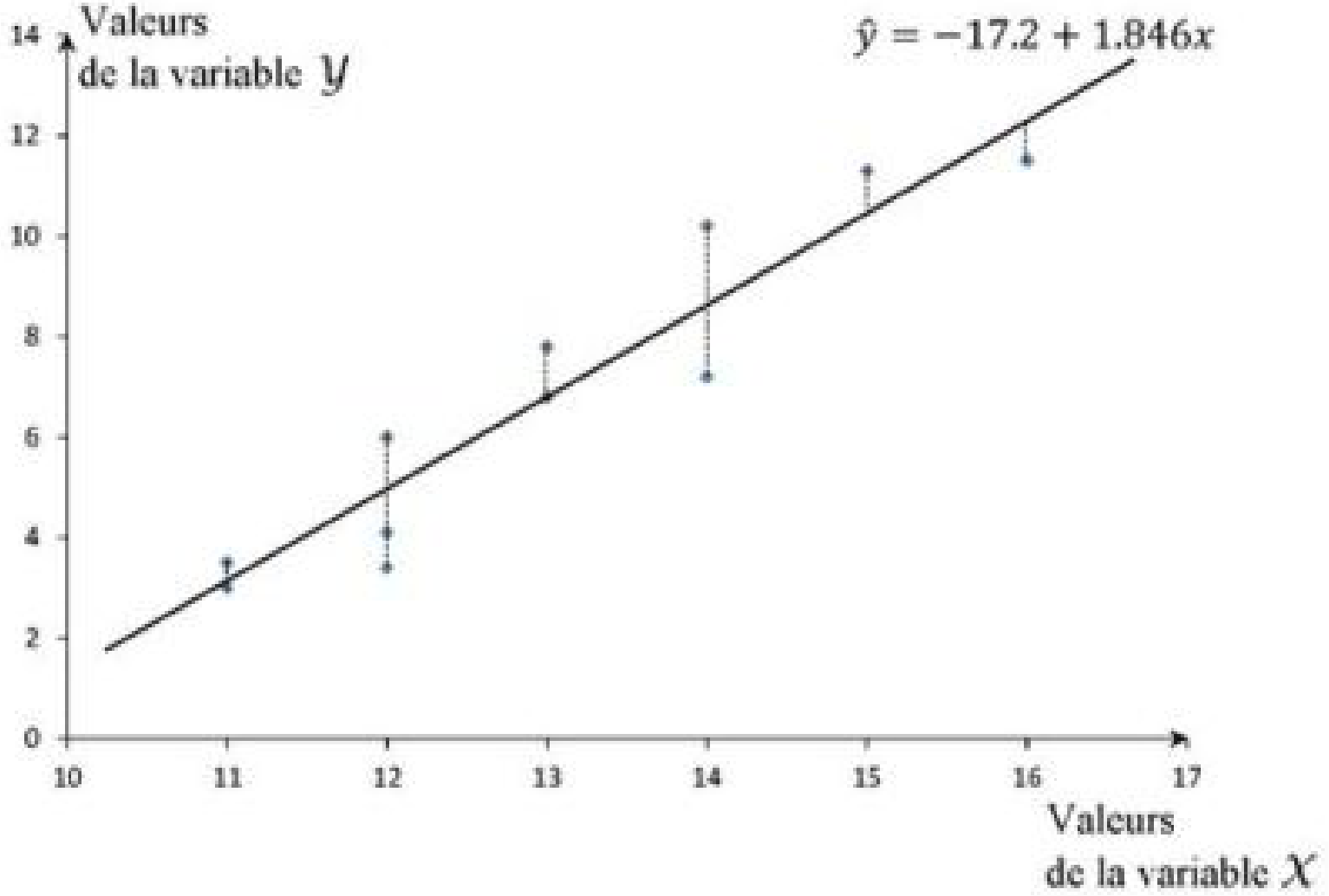
Il se calcule simplement en calculant l'écart entre la valeur réelle de y et la valeur théorique de y (obtenue à partir de l'équation déterminée par la régression linéaire) :

$$e_i = Yi - \hat{Y}i$$

$$\hat{Y}i = aXi - b$$

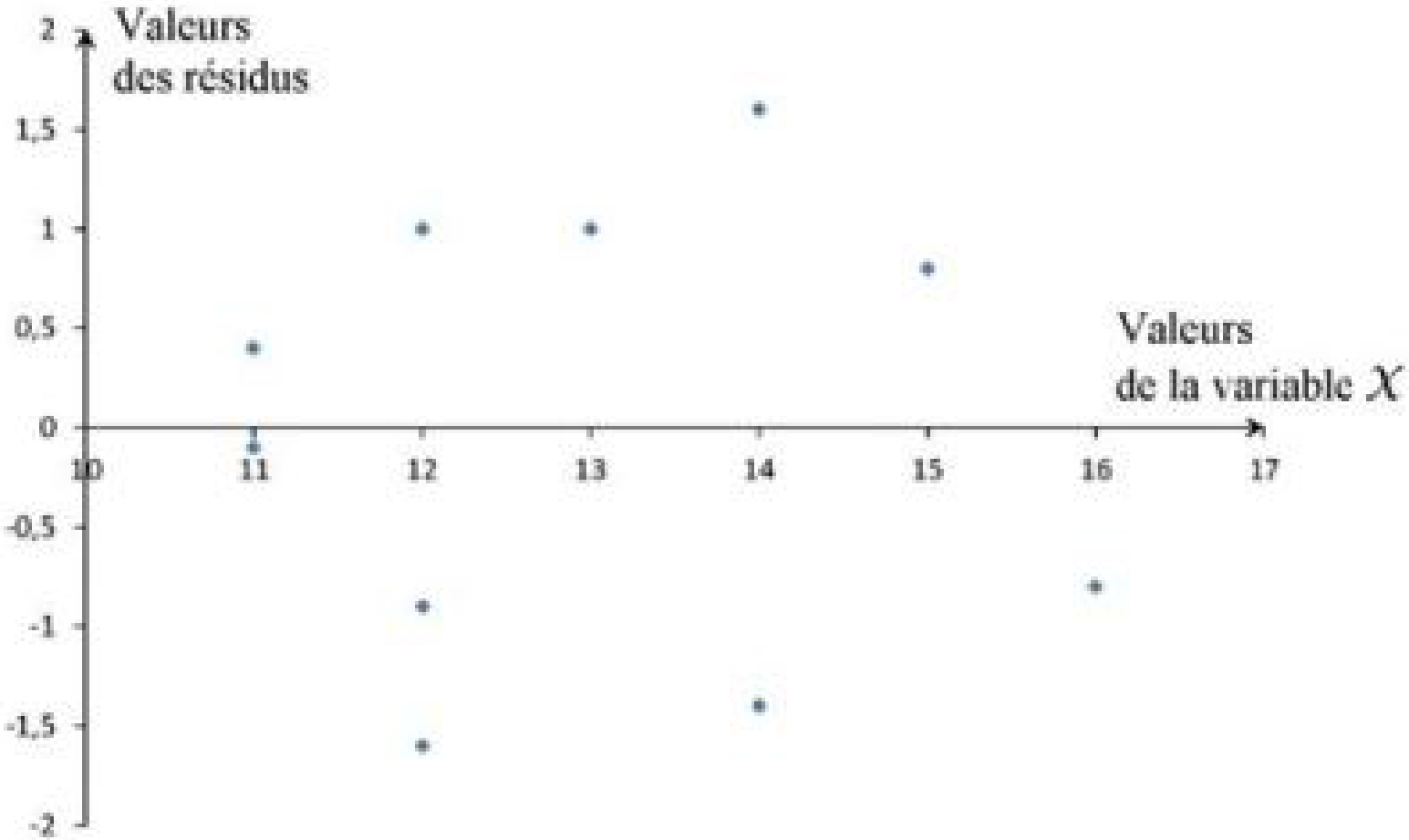
Les résidus

Interprétation graphique



Les résidus

Interprétation graphique



Les résidus

Exemple

Xi	Yi	- 15,968 × Xi + 29637
0	30 000	
100	28 000	
500	21 000	
1 000	14 000	

Les résidus

Exemple

Xi	Yi	- 15,968 × Xi + 29637
0	30 000	29 637
100	28 000	28 040
500	21 000	21 653
1 000	14 000	13 669

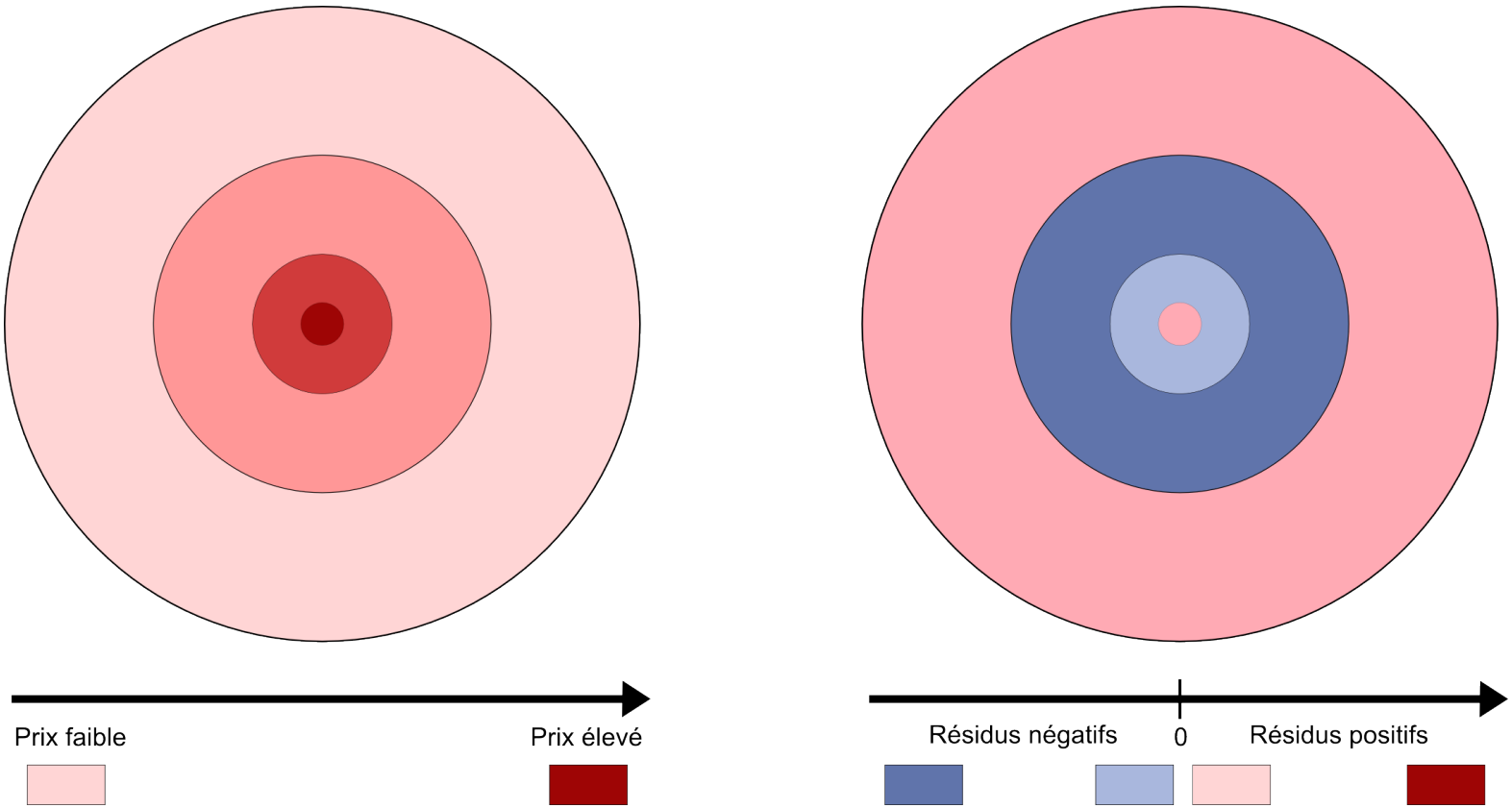
Les résidus

Exemple

Xi	Yi	- 15,968 × Xi + 29637	Yi - 15,968 × Xi + 29637
0	30 000	29 637	363
100	28 000	28 040	- 40
500	21 000	21 653	- 653
1 000	14 000	13 669	331

Les résidus

Exemple



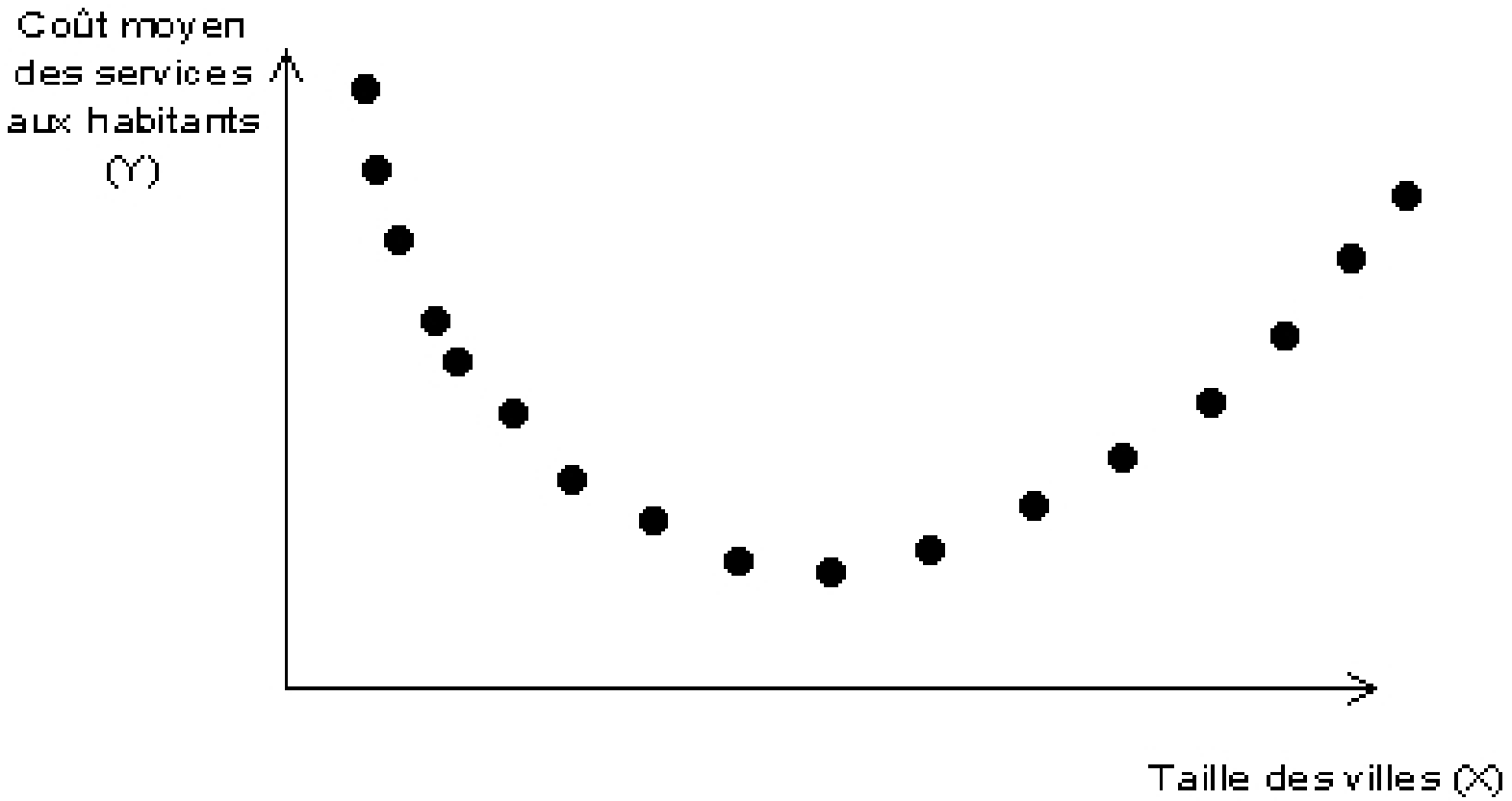
TP4

Analyse des résidus

1. Reprenez le fichier Excel des CSP et conservez uniquement les colonnes correspondant à "Profession intermédiaire" et "Employé"
2. Créez une colonne permettant de calculer pour chaque département les valeurs théoriques de la variable "Profession intermédiaire" à partir des valeurs de la variable "Employé". Pour rappel, l'équation de la régression linéaire est : y = 0,9054x - 6785,2
3. Pour chaque département, calculez les écarts entre les valeurs réelles et les valeurs théoriques de "Profession intermédiaire"
4. Créez un nuage de points avec en abscisse la variable "Employé" et en ordonné les résidus.
5. Utilisez ce fichier Excel sous Philcarto afin de cartographier les résidus issus de la relation entre "Profession intermédiaire" et "Employé".

Les pièges à éviter

Des relations de dépendance pas toujours symétriques



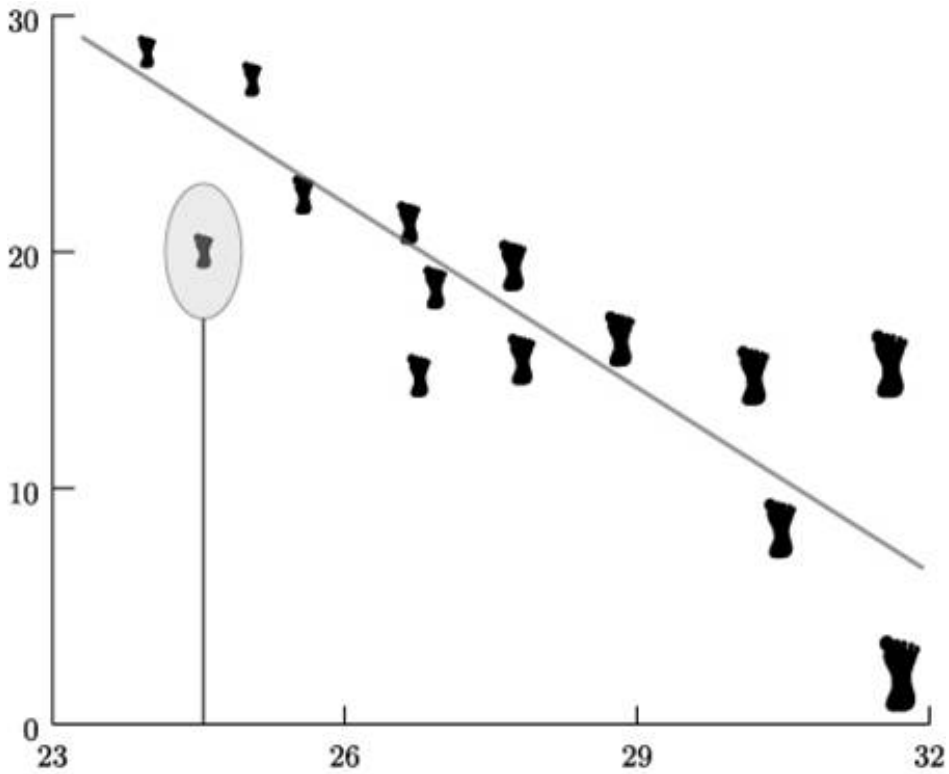
Les pièges à éviter

Des relations de dépendance symétriques toujours problématiques

Les pièges à éviter

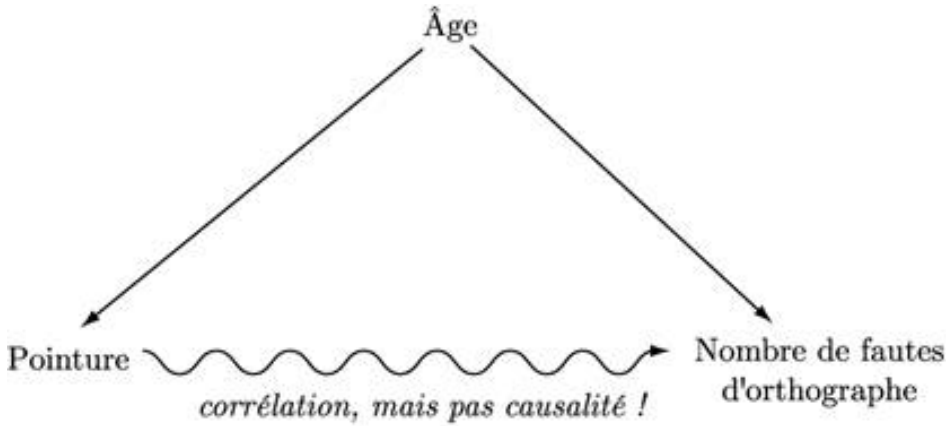
Des relations problématiques

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



Les pièges à éviter

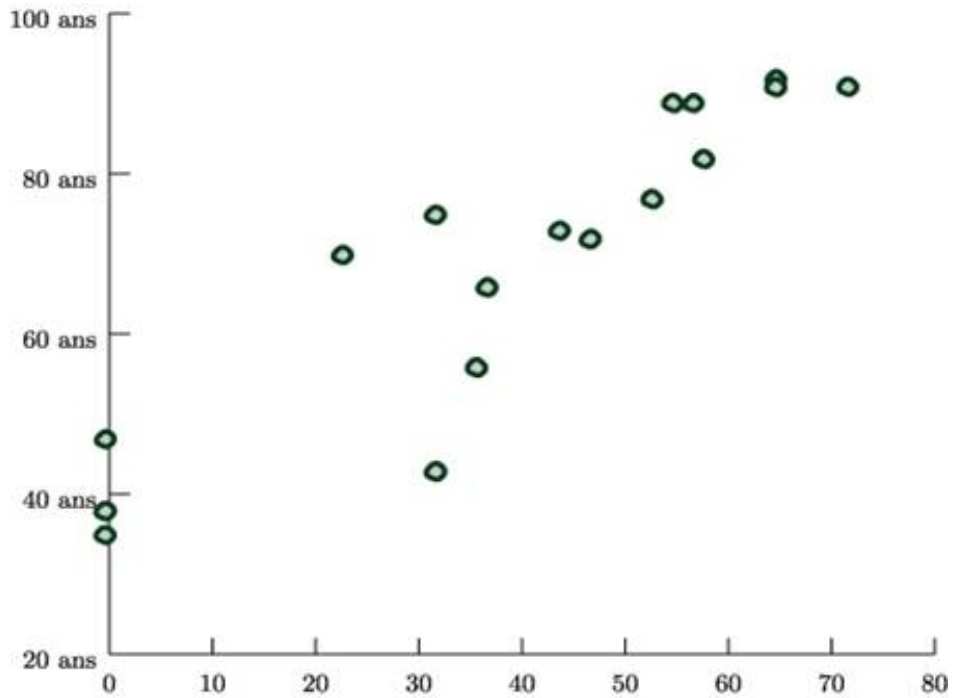
Des relations problématiques



Les pièges à éviter

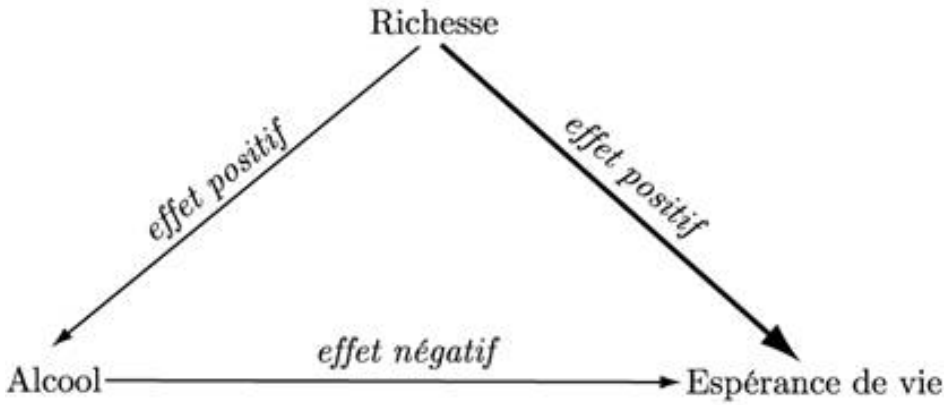
Des relations problématiques

Espérance de vie à la naissance en fonction de la consommation d'alcool par pays



Les pièges à éviter

Des relations problématiques



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.

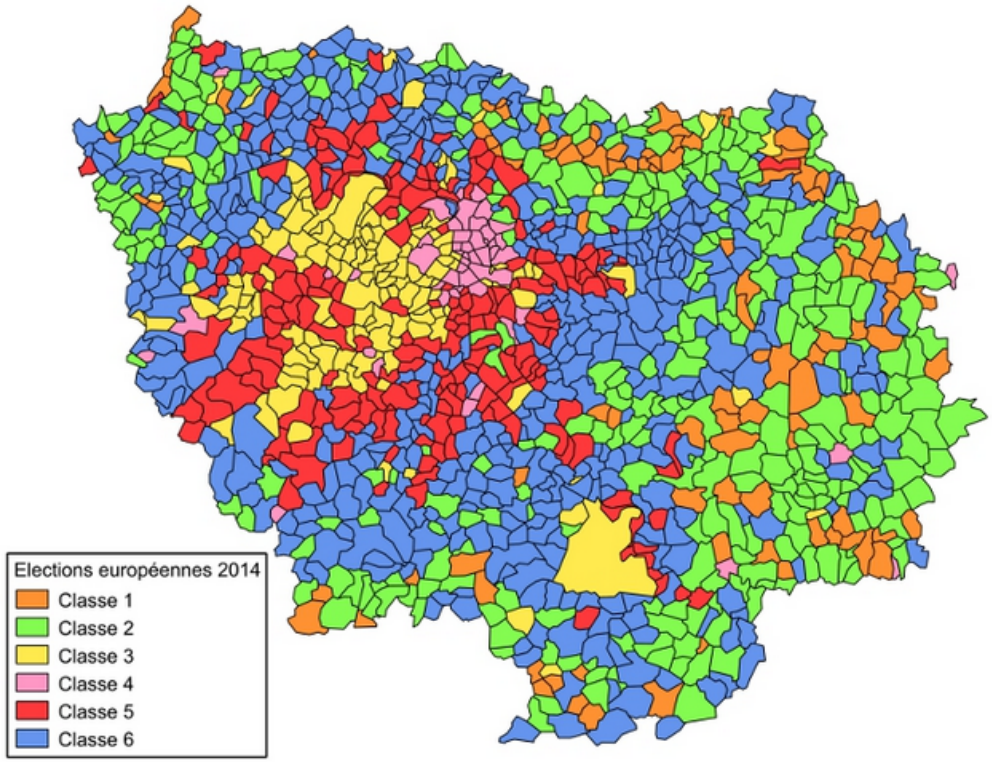
Lorsque les variables décrivant ces lieux sont des attributs sociaux décrivant les habitants, il faut toujours faire attention au fait qu'une corrélation établie au niveau des lieux n'implique pas forcément une corrélation au niveau des individus.

Ainsi, 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, dans le même temps, cette même étude au niveau des quartiers (géographique) peut montrer une corrélation parfaite entre la proportion d'étrangers des quartiers et le taux de criminalité. Il n'y a pas de contradiction, il faut juste faire attention à l'interprétation que l'on fait des résultats au niveau des quartiers...

La classification ascendante hiérarchique

Comprendre la problématique



Résultat des élections européennes 2014

Classe 1 (orange) : Très gros score du FN (RBM). L'ensemble de la gauche et le centre écrasés (environ 20% des voix au total).
Classe 2 (vert) : Fort score du FN (RBM).
Classe 3 (jaune) : Fort score de l'UMP et bon score du centre.
Classe 4 (rose) : Résultat « acceptable » pour l'ensemble de la gauche grâce notamment au PG.
Classe 5 (rouge) : Résultat proche de la moyenne avec cependant le FN (RBM) limité à 20% et les « grands » partis qui n'en bénéficient pas.
Classe 6 (bleu) : Résultat proche de la moyenne avec cependant un faible total pour l'ensemble de la gauche dont le FN (RBM) profite.

La classification ascendante hiérarchique

Comment mesurer la similarité quand on a plusieurs variables ?

	Superficie (km^2)	Habitants (Mio hab.)	Nombre de sièges au Parlement européen
Allemagne	356 900	80,780	96
Autriche	83 900	8,508	18
Belgique	30 500	11,204	21
Bulgarie	110 910	7,246	17
Croatie	56 642	4,246	11
Danemark	43 100	5,627	13
Espagne	504 800	46,508	54
Estonie	45 227	1,316	6
Finlande	337 100	5,451	13
France	544 000	65,857	74

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

Quand on a plus de deux variables ce n'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 : $| X_1 - X_2 | + | Y_1 - Y_2 |$

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

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

Procédure : regrouper les éléments qui sont proches

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

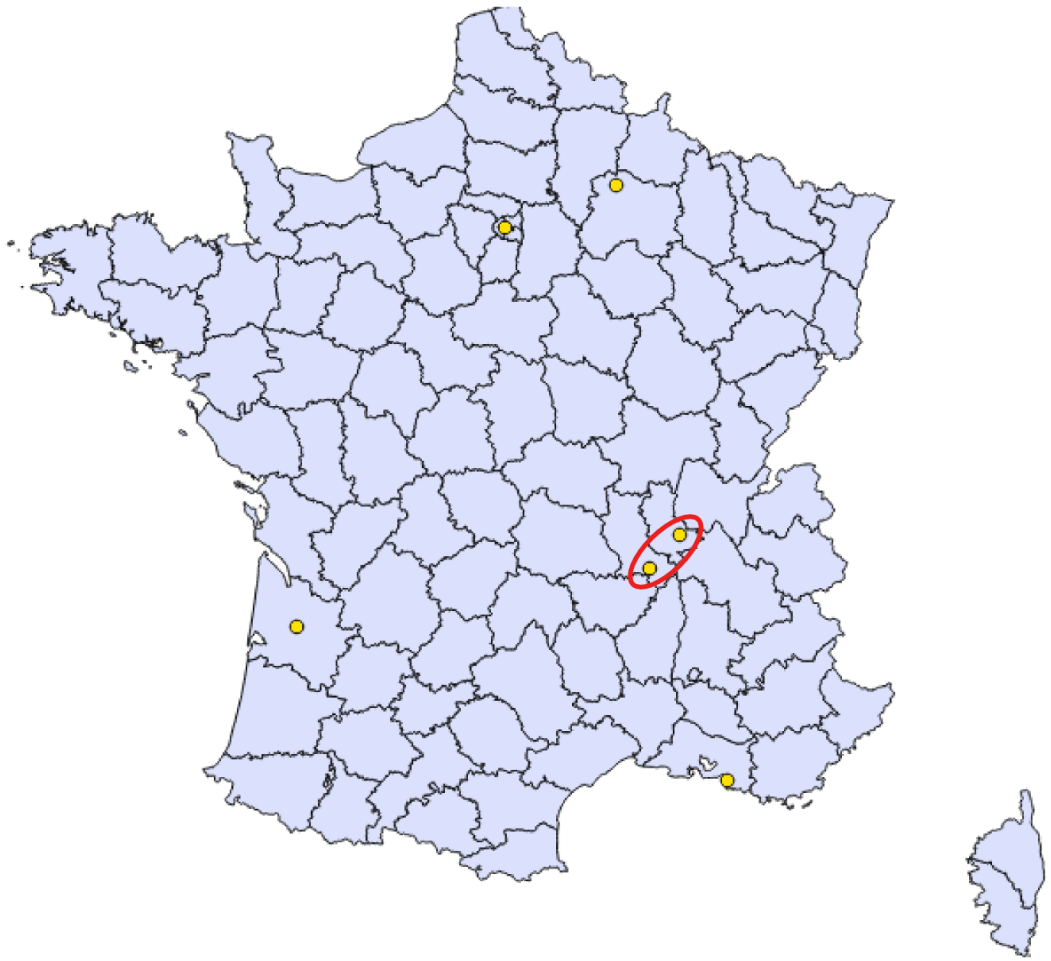
- On commence par calculer la dissimilarité entre les N objets.
- Puis on regroupe les deux objets dont le regroupement minimise un critère d'agrégation donné, créant ainsi une classe comprenant ces deux objets.
- On calcule ensuite la dissimilarité entre cette classe et les N-2 autres objets en utilisant 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

Procédure : regrouper les éléments qui sont proches

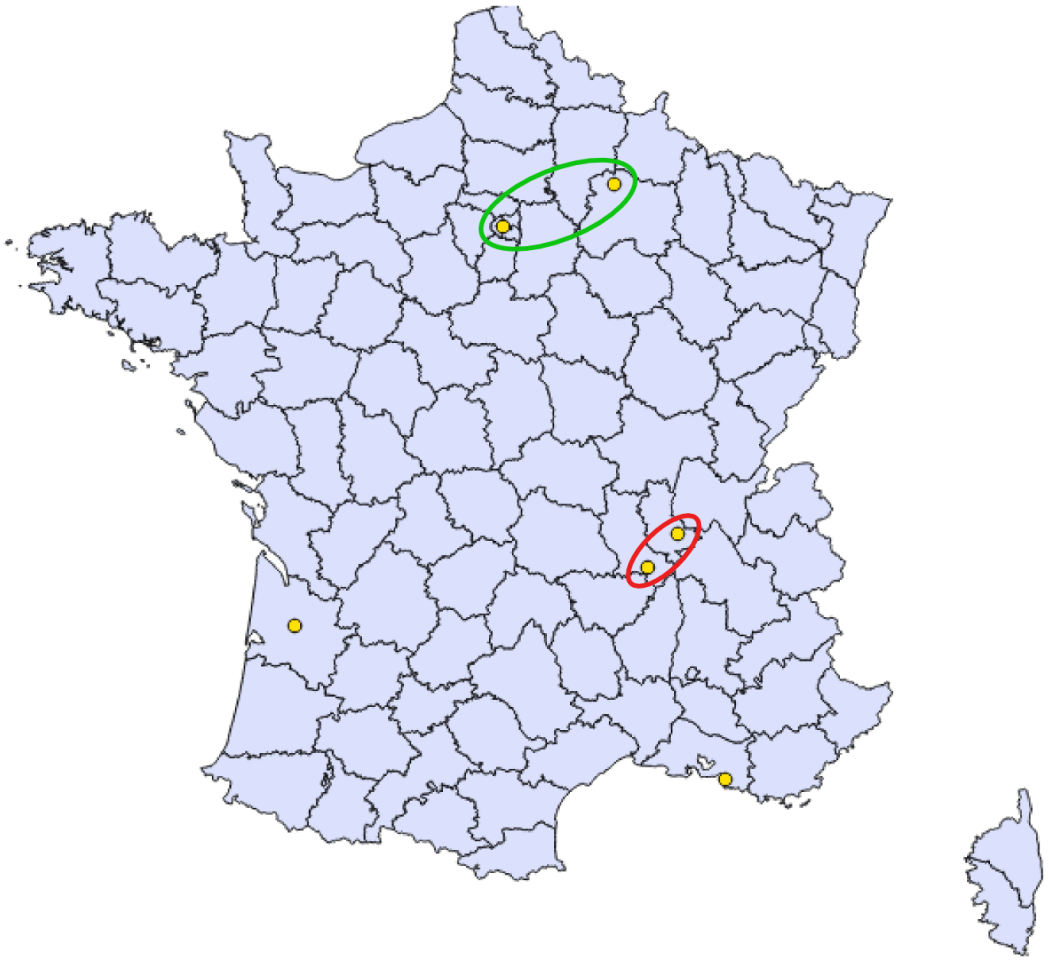
	Paris	Marseille	Saint-Etienne	Bordeaux	Reims	Lyon
Paris	0	660	410	498	131	392
Marseille	660	0	250	506	670	276
Saint-Etienne	410	250	0	396	425	50
Bordeaux	498	506	396	0	602	436
Reims	131	670	425	602	0	393
Lyon	392	276	50	436	393	0



La classification ascendante hiérarchique

Procédure : regrouper les éléments qui sont proches

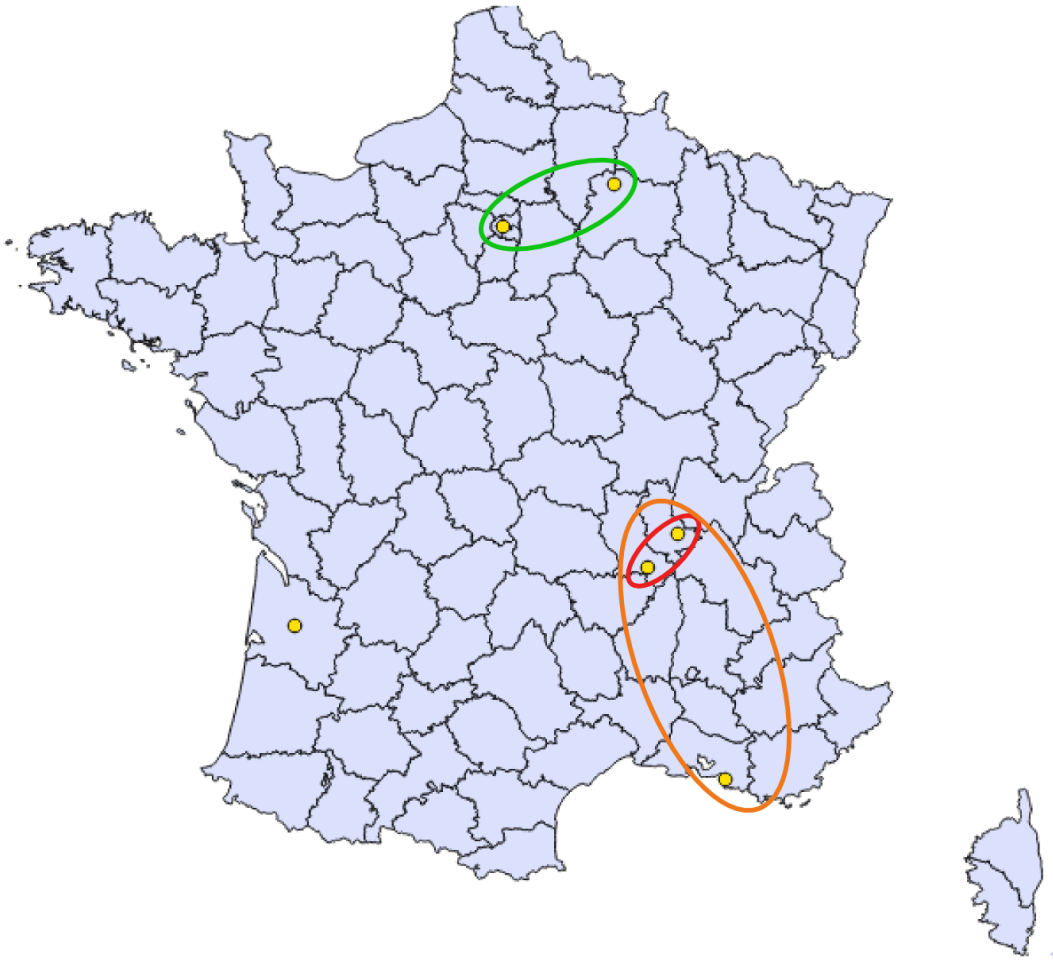
	Paris	Marseille	Saint-Etienne	Bordeaux	Reims	Lyon
Paris	0	660	410	498	131	392
Marseille	660	0	250	506	670	276
Saint-Etienne	410	250	0	396	425	50
Bordeaux	498	506	396	0	602	436
Reims	131	670	425	602	0	393
Lyon	392	276	50	436	393	0



La classification ascendante hiérarchique

Procédure : regrouper les éléments qui sont proches

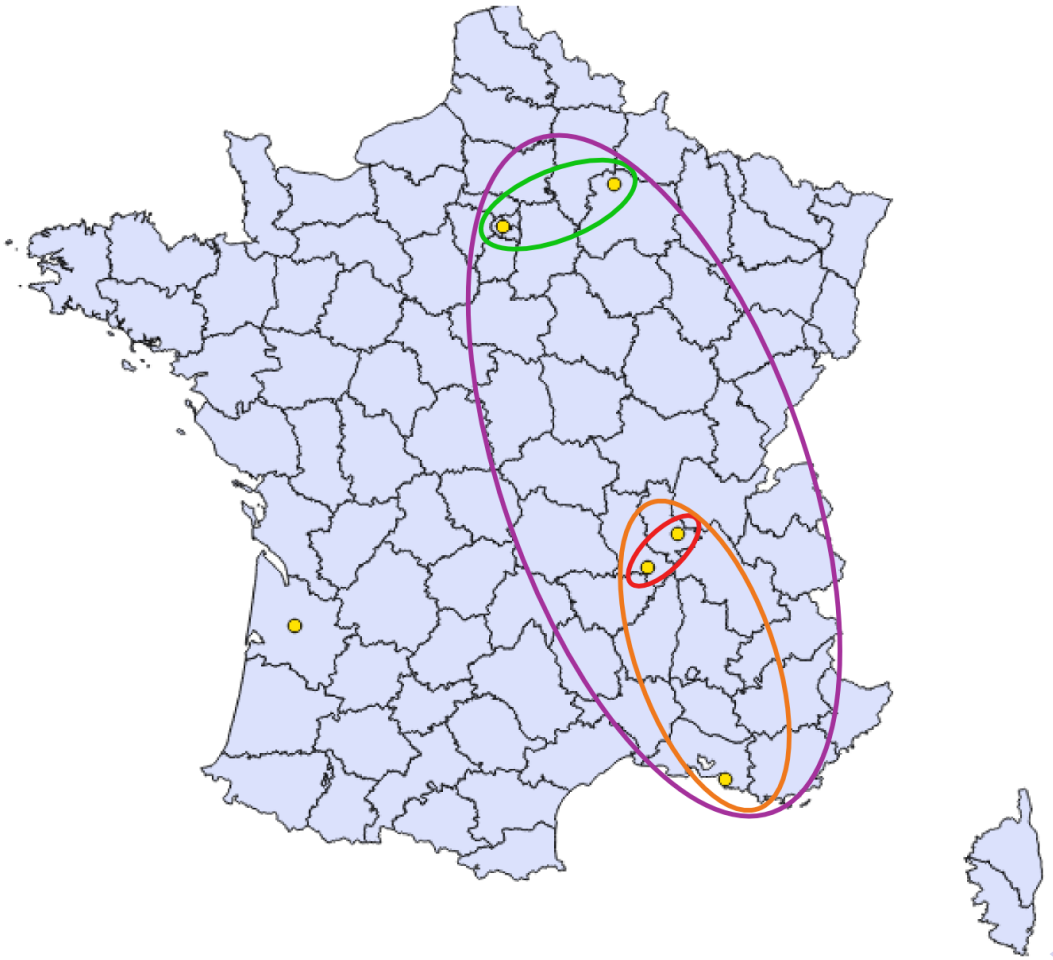
	Paris	Marseille	Saint-Etienne	Bordeaux	Reims	Lyon
Paris	0	660	410	498	131	392
Marseille	660	0	250	506	670	276
Saint-Etienne	410	250	0	396	425	50
Bordeaux	498	506	396	0	602	436
Reims	131	670	425	602	0	393
Lyon	392	276	50	436	393	0



La classification ascendante hiérarchique

Procédure : regrouper les éléments qui sont proches

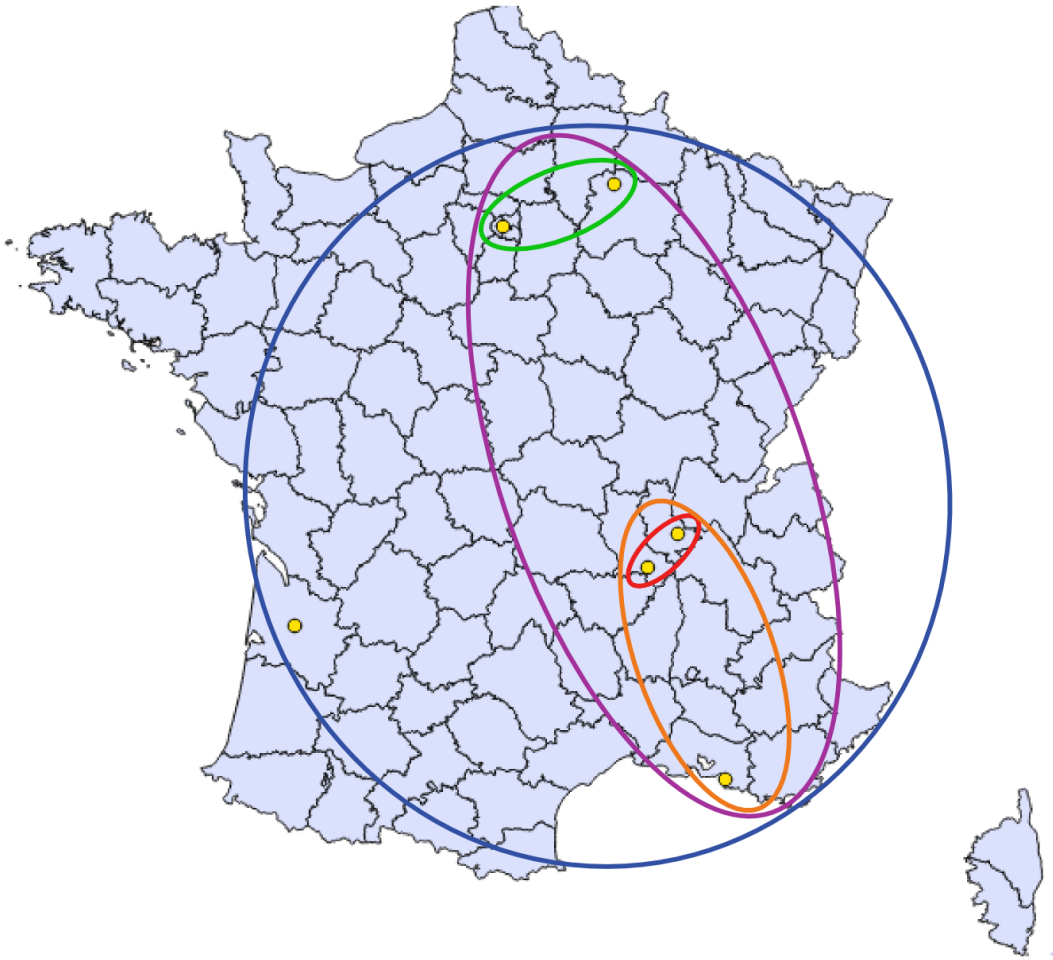
	Paris	Marseille	Saint-Etienne	Bordeaux	Reims	Lyon
Paris	0	660	410	498	131	392
Marseille	660	0	250	506	670	276
Saint-Etienne	410	250	0	396	425	50
Bordeaux	498	506	396	0	602	436
Reims	131	670	425	602	0	393
Lyon	392	276	50	436	393	0



La classification ascendante hiérarchique

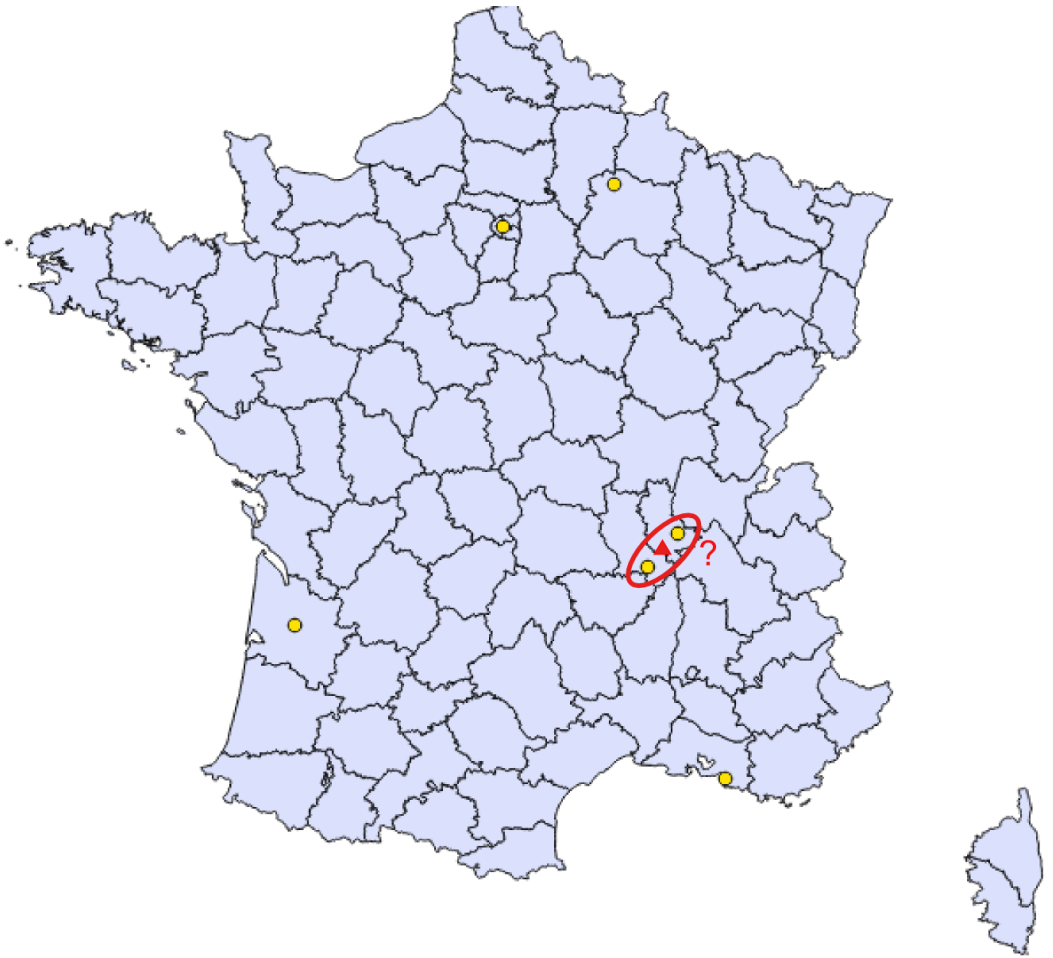
Procédure : regrouper les éléments qui sont proches

	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

Mesure de dissimilarité inter-classe



La classification ascendante hiérarchique

Mesure de dissimilarité inter-classe : différents critères

Le saut minimum retient le minimum des distances entre individus de C1 et C2. C'est ce critère que l'on a appliqué précédemment.

Le saut maximum s'appuie sur la dissimilarité des individus de C1 et C2 les plus éloignés.

Le lien moyen consiste à calculer la moyenne des distances entre les individus de C1 et C2.

La distance de Ward vise à maximiser l'inertie inter-classe.

La classification ascendante hiérarchique

Le dendrogramme

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

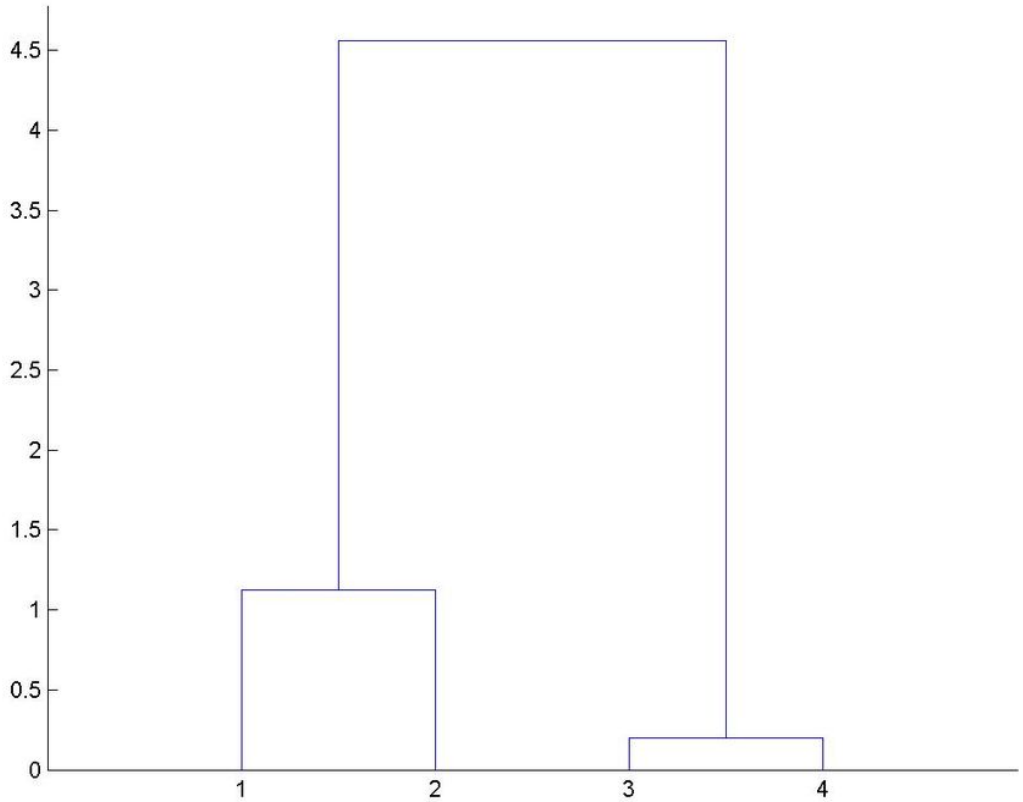
Il se présente souvent comme un arbre binaire dont les feuilles sont les individus alignés sur l'axe des abscisses.

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

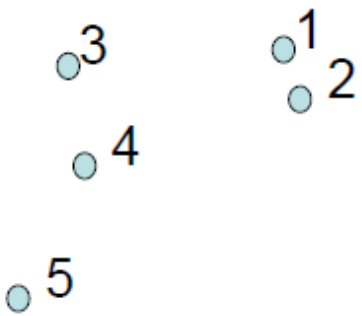
Le dendrogramme : choisir un niveau de proximité pour obtenir un nombre de classes



La classification ascendante hiérarchique

Calculer les distances (1/5)

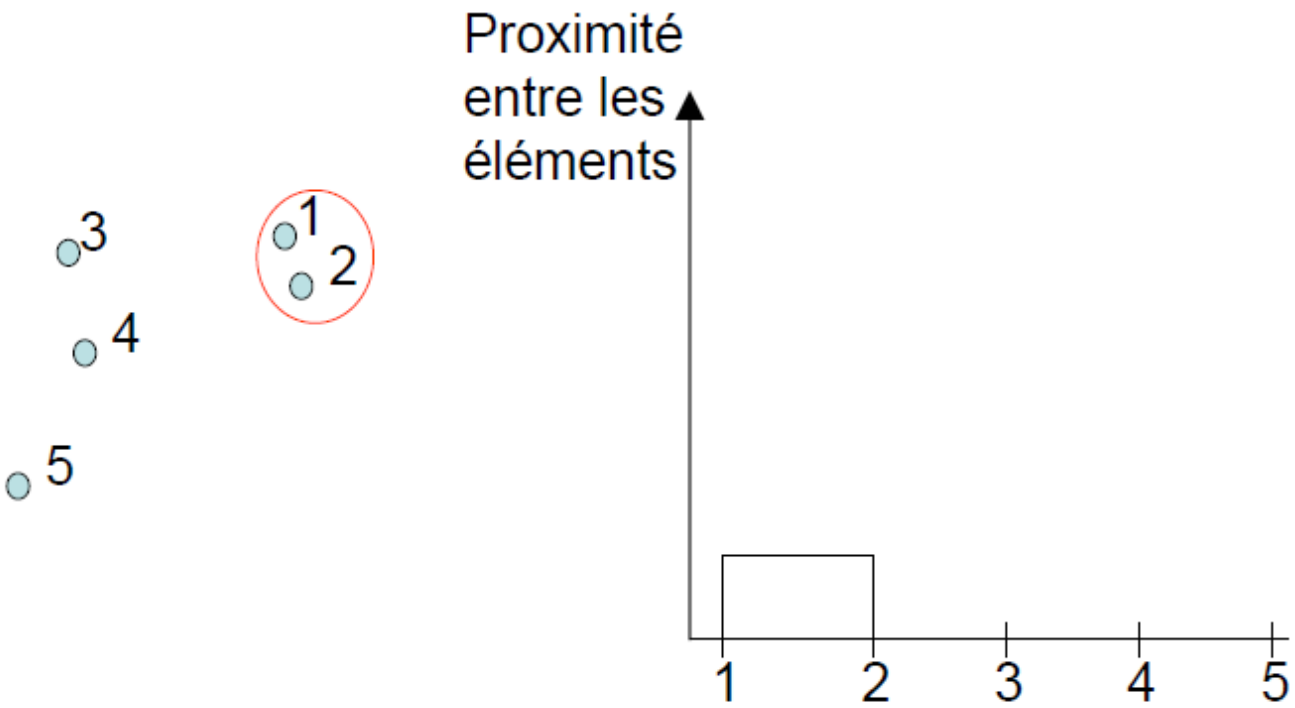
Etape 1 : n individus / n classes



La classification ascendante hiérarchique

Regrouper les éléments les plus proches (2/5)

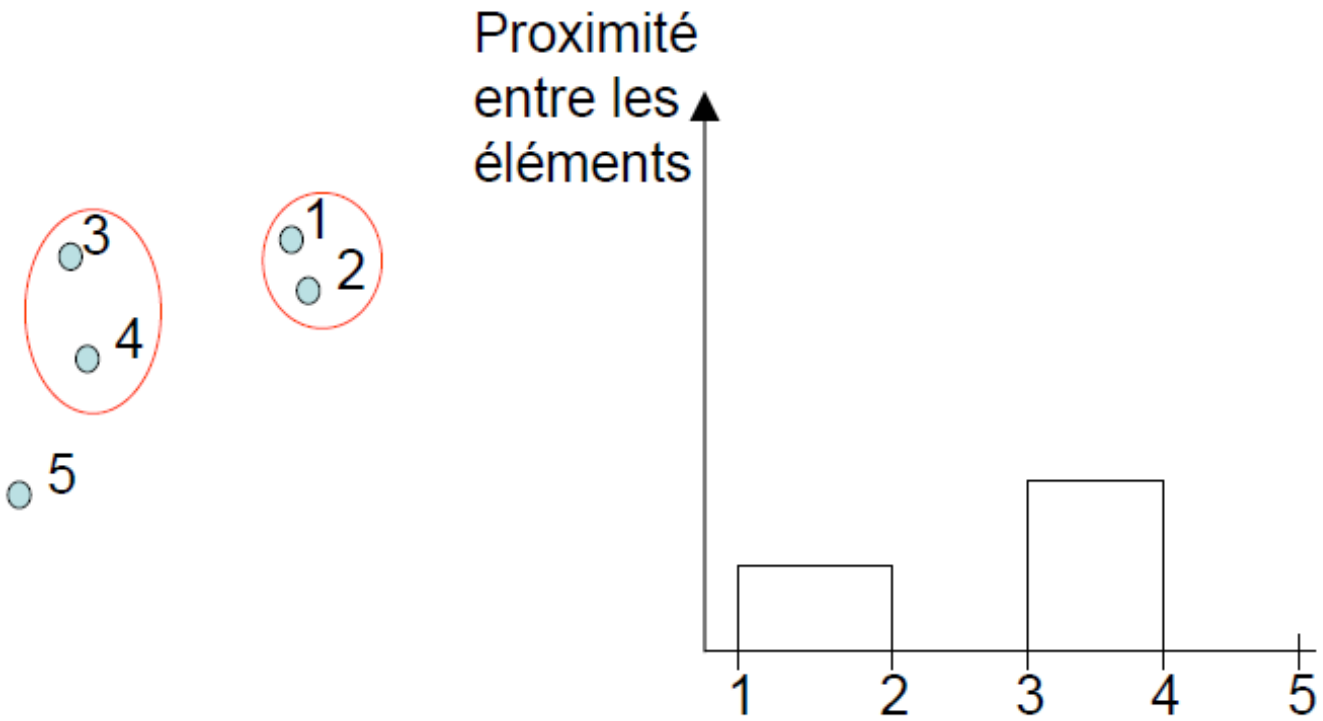
Etape 2 : n -1 classes



La classification ascendante hiérarchique

Regrouper de nouveau à l'aide d'un critère (3/5)

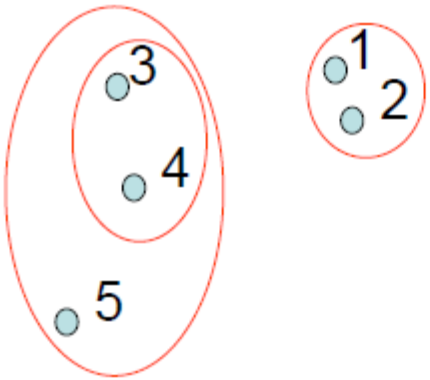
Etape 3 : n -2 classes

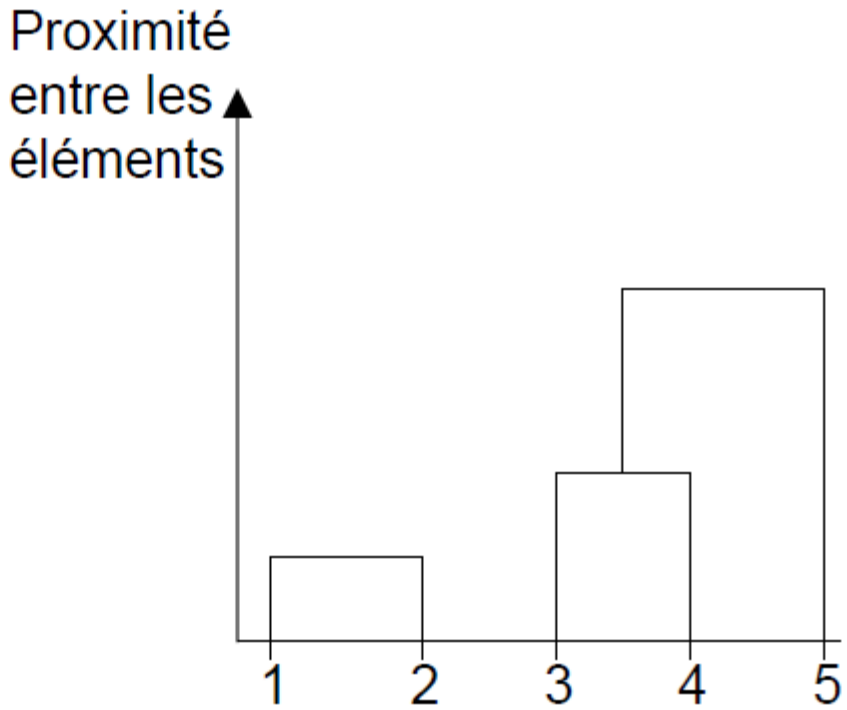


La classification ascendante hiérarchique

Regrouper encore (4/5)

Etape 4 : n -3 classes

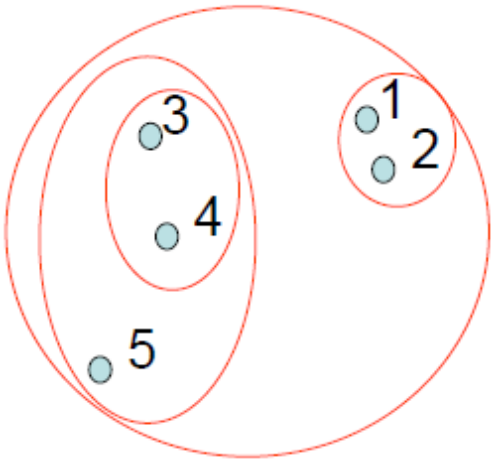


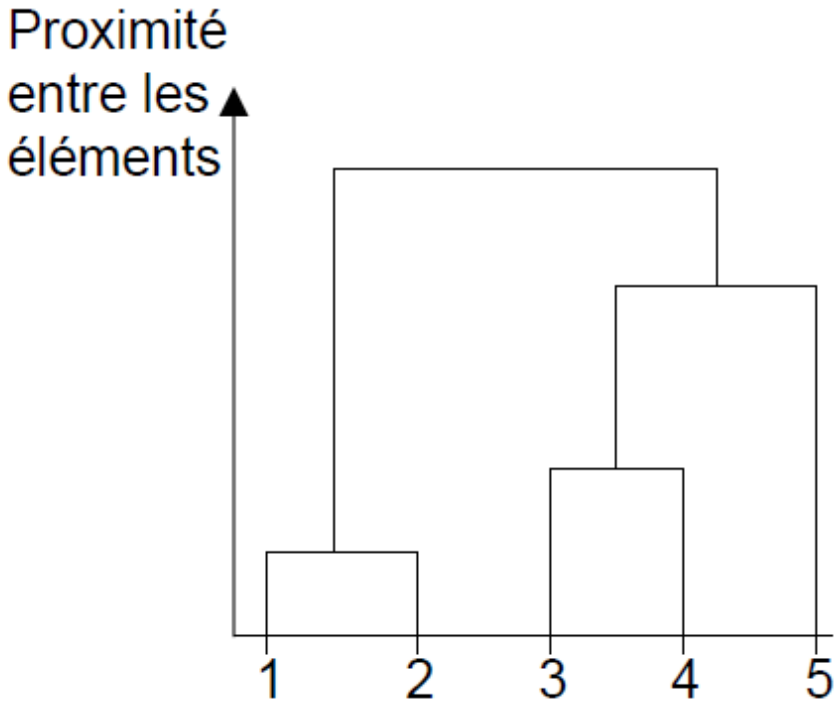


La classification ascendante hiérarchique

Regrouper toujours (5/5)

Etape 5 : n -4 = 1 classe





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

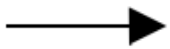
Name	DEM1	DEM2	DEM3	ECO1	ECO2	ENV1	ENV2
France	0.6	-0.1	-0.4	1.2	0.9	0.4	-0.1
Italy	2.0	2.0	-0.7	0.5	0.8	0.8	0.0
Spain	0.1	-1.0	-0.7	0.0	0.2	0.3	0.1
Algeria	-1.0	0.7	1.8	-1.4	-1.5	-0.8	1.3
Libya	-1.2	-0.7	1.6	-1.3	-1.2	1.2	2.6
Morocco	-0.2	-0.1	1.0	-1.4	-1.4	-1.6	-0.2
Tunisia	-0.3	-1.1	0.8	-1.3	-1.2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

La classification ascendante hiérarchique

TP5

1. Reprenez le dossier de départ du TP1 et ouvrez le sous Philcarto. Conservez les types de variables par défaut, puis cliquez sur l'icône "MULTIV", puis sur "CAH mesures". Quelles variables pouvez-vous sélectionner ?
2. Sélectionnez toutes les variables, puis cliquez sur "Calculer".
3. Produisez une carte avec deux classes, puis trois, puis quatre, puis cinq.
4. Reproduisez cette analyse avec cette fois un nombre de 10 classes.
5. Que peut-on faire avec Philcarto une fois ces analyses produites.

L'effet de taille : indices de localisation et de spécialisation

Présentation : tableau de contingence

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	128	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.

L'effet de taille : indices de localisation et de spécialisation

Présentation : tableau de contingence

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*

L'effet de taille : indices de localisation et de spécialisation

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%

Suppression de l'effet de taille des entités géographiques. Mise en valeur de la taille des variables.

L'effet de taille : indices de localisation et de spécialisation

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%

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

$$S_i = \sum_{j=1}^{n} \left| \frac{N_{ij}}{N_{i.}} - \frac{N_{.j}}{N_{..}} \right|$$

L'effet de taille : indices de localisation et de spécialisation

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%

$$S_{(BULGARIE)} = | 0.23 - 0.11 | + | 0.23 - 0.21 | + | 0.07 - 0.08 | + | 0.03 - 0.04 | + | 0.05 - 0.08 | + | 0.08 - 0.07 | + | 0.04 - 0.06 | + | 0.03 - 0.04 | + | 0.27 - 0.35 | = 0.30$$

L'effet de taille : indices de localisation et de spécialisation

Deux profils possibles : le profil en ligne

BULGARIE	30%
HONGRIE	15%
POLOGNE	11%
R.D.A.	24%
ROUMANIE	28%
TCHECO.	17%
YOUGOSL.	13%

L'indice de spécialisation est pertinent d'un point de vue cartographique.

L'effet de taille : indices de localisation et de spécialisation

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%**

Suppression de l'effet de taille des variables. Mise en valeur de la taille des entités géographiques.

L'effet de taille : indices de localisation et de spécialisation

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%**

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

$$L_j = \sum_{j=1}^{n} \left| \frac{N_{ij}}{N_{.j}} - \frac{N_{i.}}{N_{..}} \right|$$

L'effet de taille : indices de localisation et de spécialisation

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%**

$$L_{(ALIM)} = \mid 0.10 - 0.05 \mid + \mid 0.11 - 0.10 \mid + \mid 0.29 - 0.23 \mid + \mid 0.16 - 0.22 \mid + \mid 0.10 - 0.12 \mid + \mid 0.14 - 0.19 \mid + \mid 0.10 - 0.09 \mid = 0.26$$

L'effet de taille : indices de localisation et de spécialisation

Deux profils possibles : le profil en colonne

ALIM	TEXT	BOIS	EDIT	CHIM	CONS	META	EQUIP
26%	9%	43%	24%	25%	12%	25%	13%

L'indice de localisation n'est pas pertinent d'un point de vue cartographique, mais a du sens d'un point de vue géographique.

TP6

Indices de spécialisation et de localisation

1. A partir des données du TP6, calculez les taux de "Profession intermédiaire" et d'"Employé" par département afin de supprimer l'effet de taille des entités géographiques (profil en ligne). Ces deux taux sont-ils corrélés ?
2. Calculez les écarts entre les taux calculés et les taux nationaux.
3. Calculez les taux de "Profession intermédiaire" et d'"agriculteur" afin de supprimer l'effet de taille des variables (profil en colonne).
4. Calculez l'écart entre le profil des variables et le profil de l'ensemble.
5. Calculez les indices de spécialisation de chaque département, puis les indices de localisation de chaque CSP.