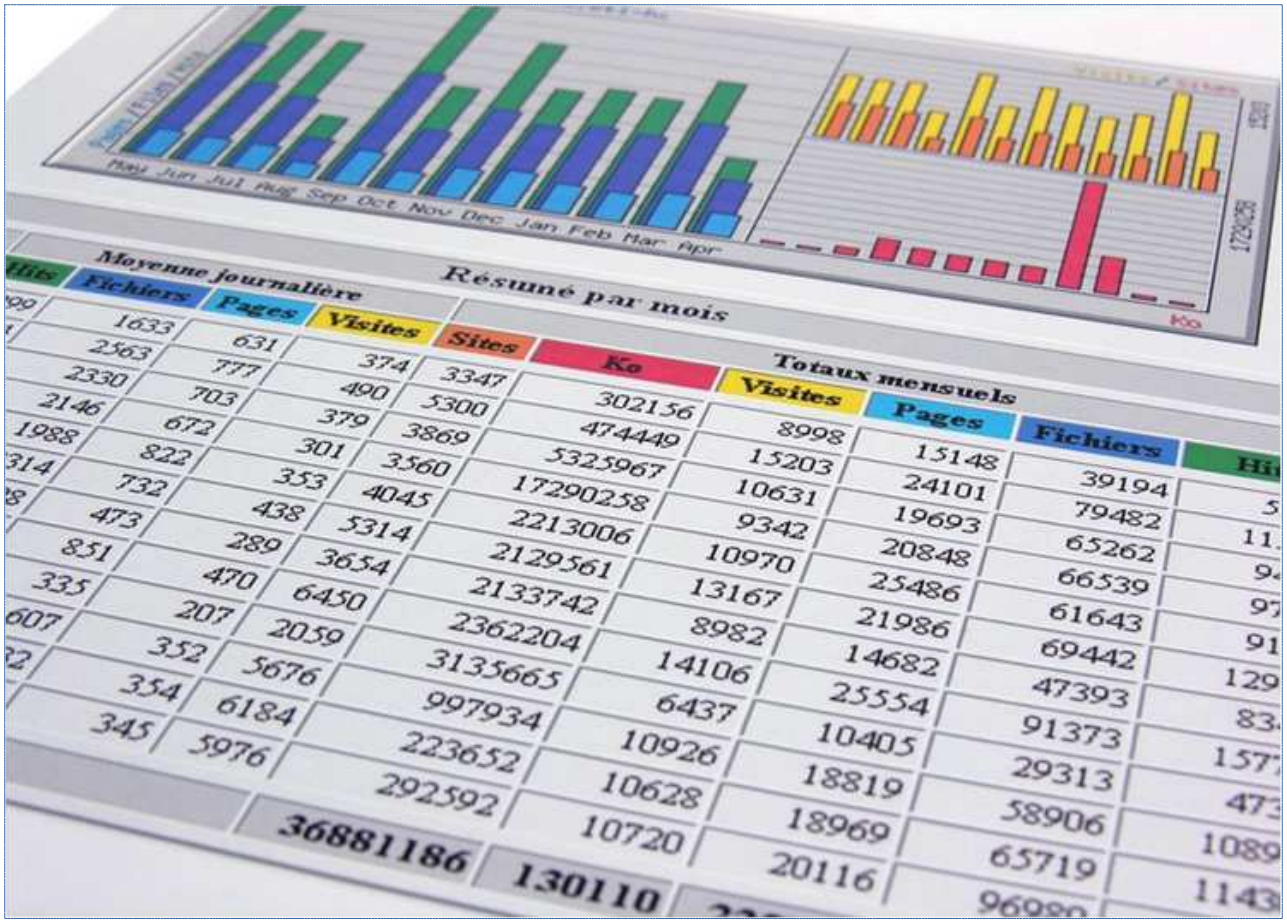
Cerema

La statistique, c’est pas automatique



Peggy Mertiny
Cerema Hauts-de-France

La statistique ?

Faits

Données

Analyse

Pourquoi user de la statistique ?

La donnée c’est des métiers

Manier la statistique pour réaliser des études et analyser des phénomènes de manière pertinente requiert plus qu’un savoir faire technique. Il est utile de disposer à la fois de :

- la connaissance des outils statistiques mobilisables pour traiter les données
  - → rôle de **statisticien**
- la connaissance de la thématique abordée, des sources à exploiter
  - → rôle de **thématicien**
- la connaissance des outils de mise en valeur, d’illustration de la donnée
  - → rôle de **géomaticien**

Travailler au **Cerema** c’est **être, connaître, travailler** avec tout cela à la fois.

Pourquoi user de la statistique ?

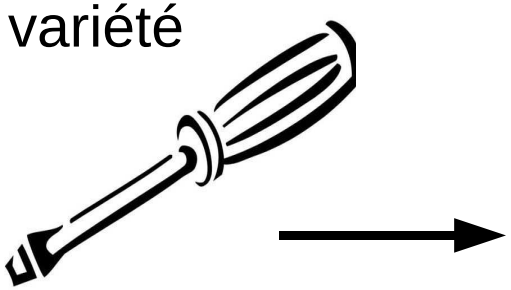
Des outils pour comprendre les données

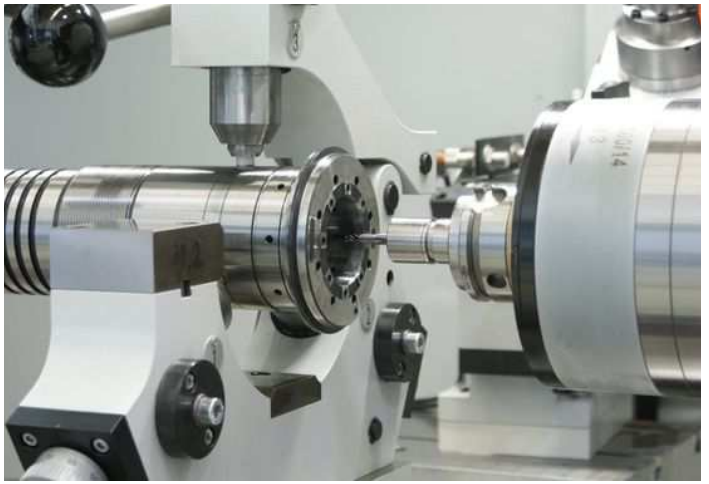
La statistique est avant toute chose un **OUTIL**.

Or, des outils, il en existe une variété incalculable (infinie?),

… du plus simple,

… au plus perfectionné.





Il y a néanmoins une constante : qu'importe son degré de complexité, il est important de savoir s'en servir au risque de se tromper lourdement.

Mais, user de la statistique implique non seulement une connaissance technique, mais également de savoir s'en servir à bon escient

→ l'outil ne fait pas l'artisan.

La maîtrise du domaine, de ses enjeux, de ses usages, de ses ramifications apporte un éclairage indispensable à la description des phénomènes.

Pourquoi user de la statistique ?

La donnée c’est des faits

Mark Twain

“Les faits sont têtus.
Il est plus facile de s'arranger
avec les statistiques.”

Pourquoi user de la statistique ?

Des outils pour montrer l’exemple

- La moyenne ou la médiane ?
- Le paradoxe de Sympson
- L’analyse en composantes principales

Moyenne ou médiane ?

	Moyenne	Médiane
Définition	La moyenne d'une série est égale à la somme de toutes les valeurs divisée par l'effectif total de la série.	La médiane est la valeur qui partage cette série en deux séries de même effectif : autant de valeurs inférieures et supérieures.
Quand l'applique-t-on ?	La moyenne est utilisée pour des distributions normales, ayant un faible nombre de valeurs aberrantes.	La médiane est généralement utilisée pour retourner la tendance centrale des distributions asymétriques.
Ex : distribution Normale	2, 3, 3, 5, 8, 10, 11 AVG = 6	2, 3, 3, 5, 8, 10, 11 MED = 5
Ex : distribution asymétrique	2, 2, 3, 3, 5, 7, 8, 130 (2+2+3+3+5+7+8+130)/8= 20 AVG = 20	2, 2, 3, 3, 5, 7, 8, 130 (3+5)/2=4 MED = 4

Si les **données** que vous comparez sont **uniformes**

→ vous pouvez utiliser en toute sécurité l'agrégateur **Moyenne**.

Si votre ensemble de chiffres contient quelques **valeurs extrêmes**,

→ vous devez alors préférer l'utilisation de **Médiane**, plus à même d'indiquer le **« centre » de votre distribution.**

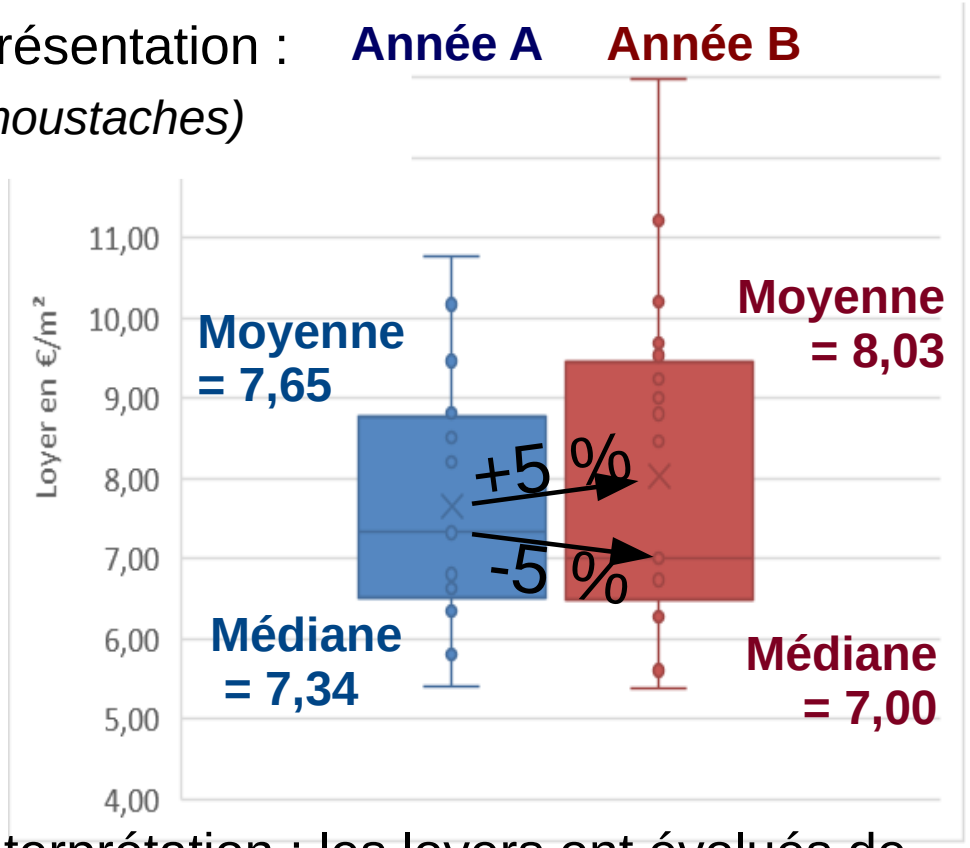
La preuve par l'exemple

Moyenne ou médiane appliquée à l'évolution des loyers

Les données :

Montant du loyer en €/m2	Année A	Année B	Evolutio
Logement 6	5,41	5,60	3,5%
Logement 7	5,80	5,37	-7,4%
Logement 12	6,34	6,84	7,9%
Logement 19	6,47	7,00	8,2%
Logement 20	6,47	7,00	8,2%
Logement 1	6,62	6,31	-4,7%
Logement 10	6,72	6,73	0,1%
Logement 11	6,80	6,39	-6,0%
Logement 3	6,81	6,28	-7,8%
Logement 9	7,32	7,00	-4,4%
Logement 16	7,35	7,00	-4,8%
Logement 17	7,46	8,46	13,4%
Logement 4	8,21	9,00	9,6%
Logement 2	8,51	8,80	3,4%
Logement 15	8,58	9,69	12,9%
Logement 13	8,82	9,53	8,0%
Logement 5	8,88	9,23	3,9%
Logement 8	9,46	10,20	7,8%
Logement 18	10,17	11,22	10,3%
Logement 14	10,76	12,99	20,7%

Leur représentation : *(boîte à moustaches)*



→ L'interprétation : les loyers ont évolués de manière hétérogènes et si les plus élevés se sont envolés, d'autres sont restés faibles voire, se sont dépréciés, les rendant globalement plus accessibles.

Le paradoxe de SIMPSON

Le paradoxe de Simpson est un paradoxe statistique décrit par **Edward Simpson** en 1951, dans lequel un phénomène observé de plusieurs groupes semble s'inverser lorsque les groupes sont combinés.

Ce résultat, impossible au premier abord mais bien présent dans la réalité, est lié à des éléments qui ne sont pas pris en compte :

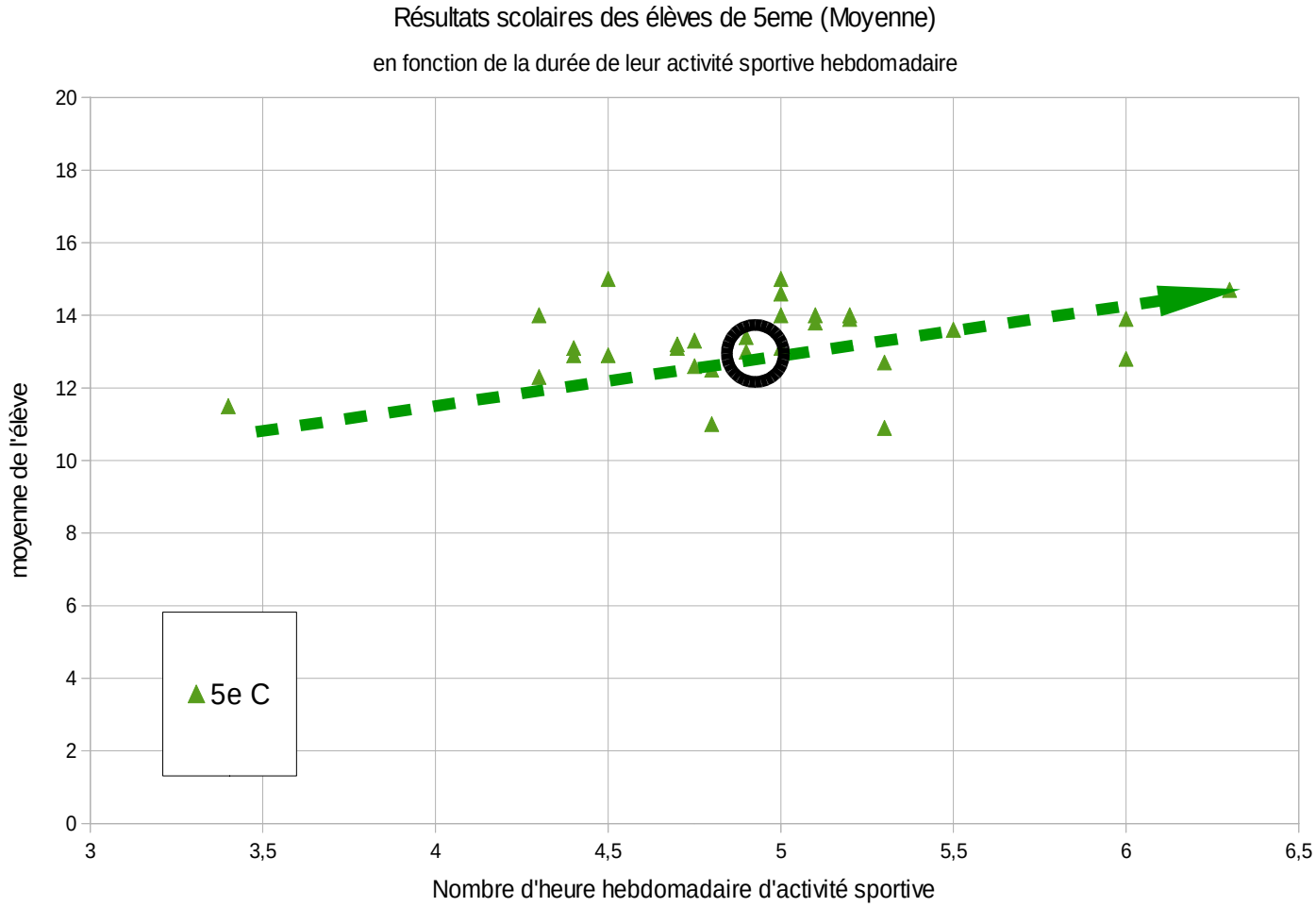
- présence de variables non-indépendantes
- ou différences d'effectifs entre les groupes, etc.

Cet élément est appelé **« facteur de confusion »**.

La preuve par l'exemple

Le paradoxe de SIMPSON appliqué à mon fils

Question : la pratique du sport améliore t'elle les résultats scolaires ?



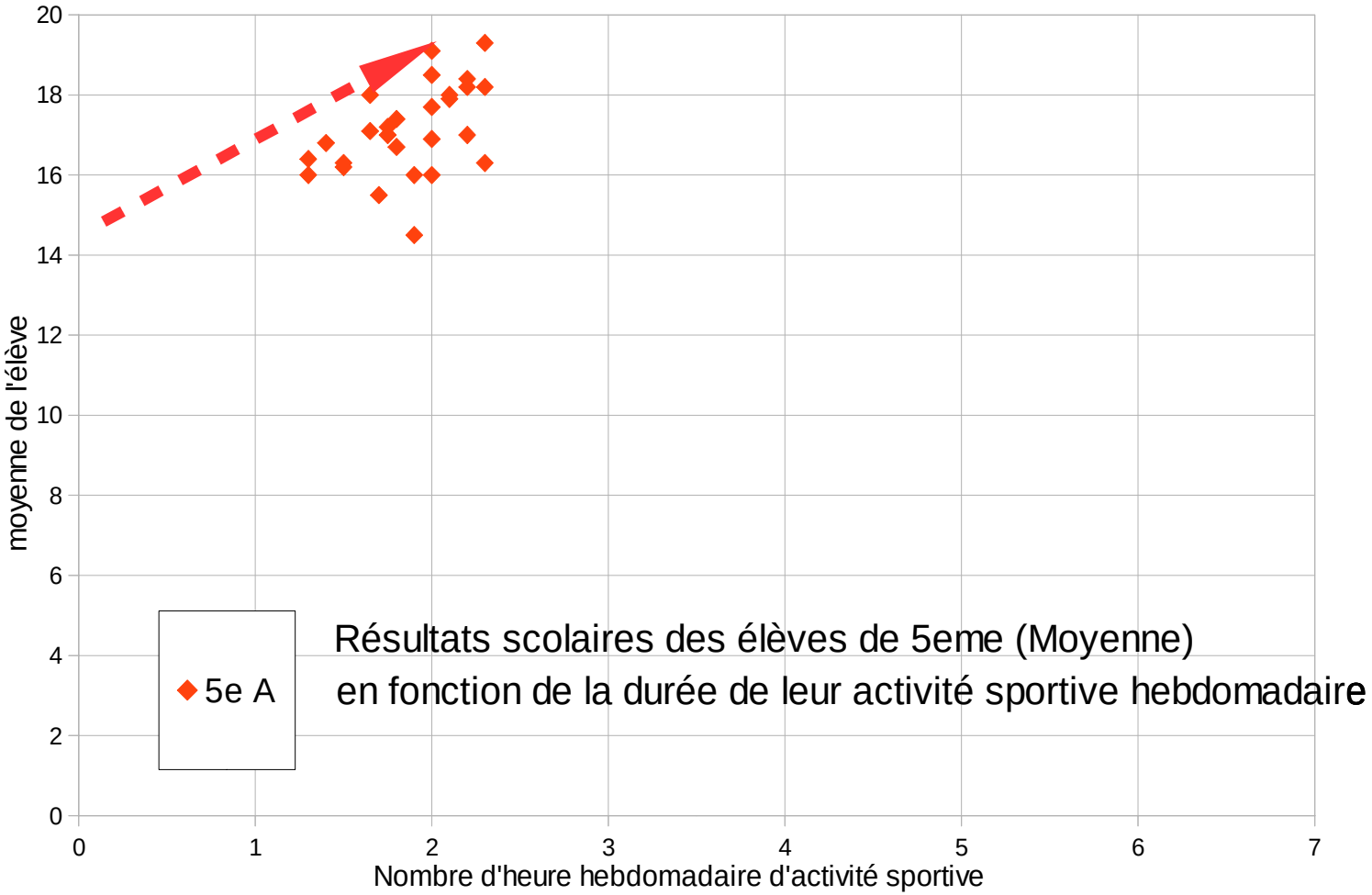
Si je considère que la moyenne scolaire est un bon indicateur, regardons comment celle-ci évolue en fonction du temps à pratiquer le sport.

Voici les élèves de 5e C représentés en nombre d'heures d'activité sportive hebdomadaire (ordonnée) et leur note moyenne (abscisse).

Observation : **plus l'élève fait du sport, meilleurs sont ses résultats scolaires**

La preuve par l’exemple

Le paradoxe de SIMPSON appliqué à mon fils

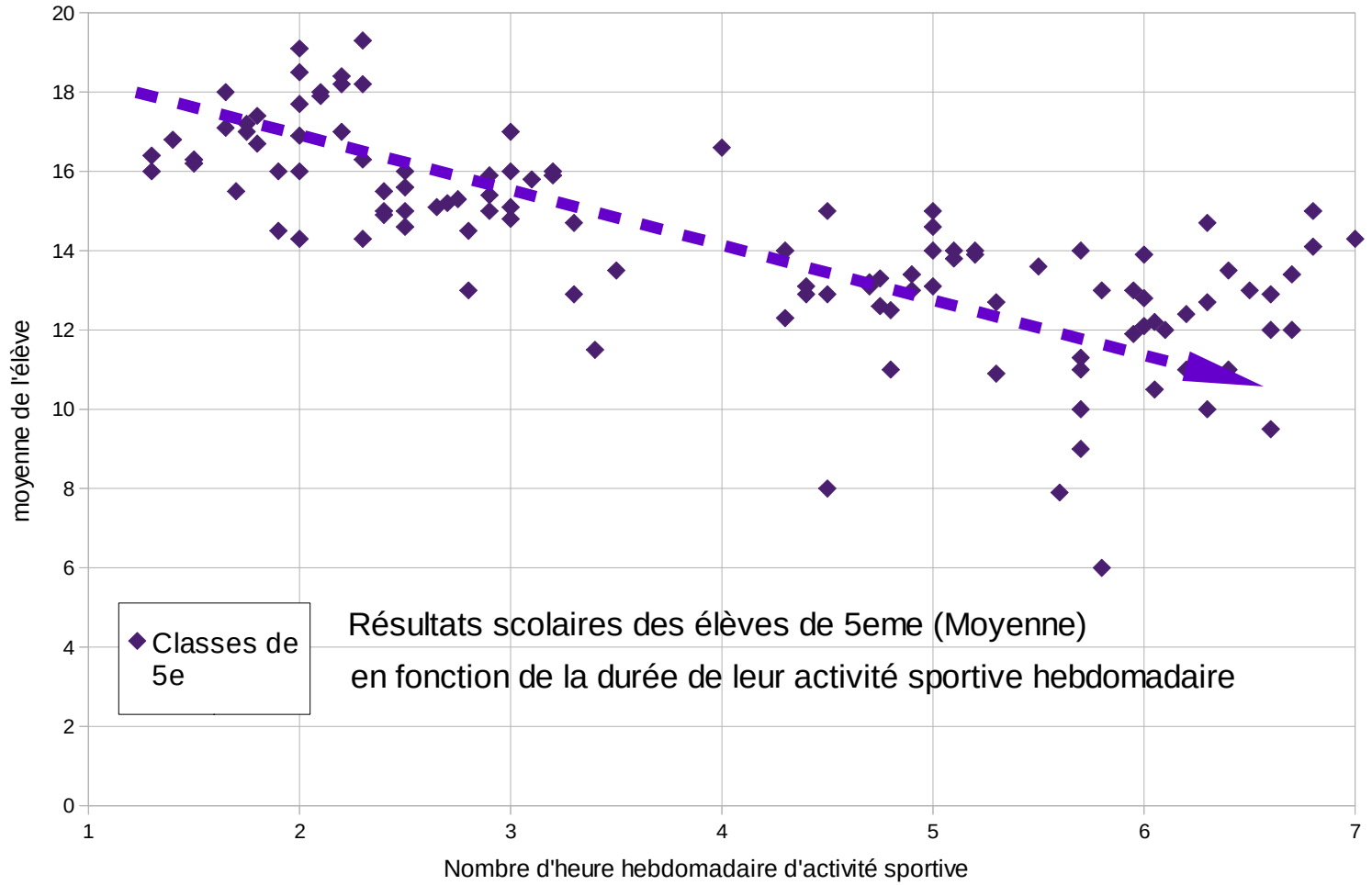


Pour chaque classe succesivement, on fait la même observation

→ **plus l’élève fait du sport, meilleur sont ses résultats.**

La preuve par l'exemple

Le paradoxe de SIMPSON appliqué à mon fils



Mais lorsque l'on représente l'ensemble des classes, l'image est tout autre...

Ici les **facteurs de confusion** sont

- le niveau scolaire
- et la classe

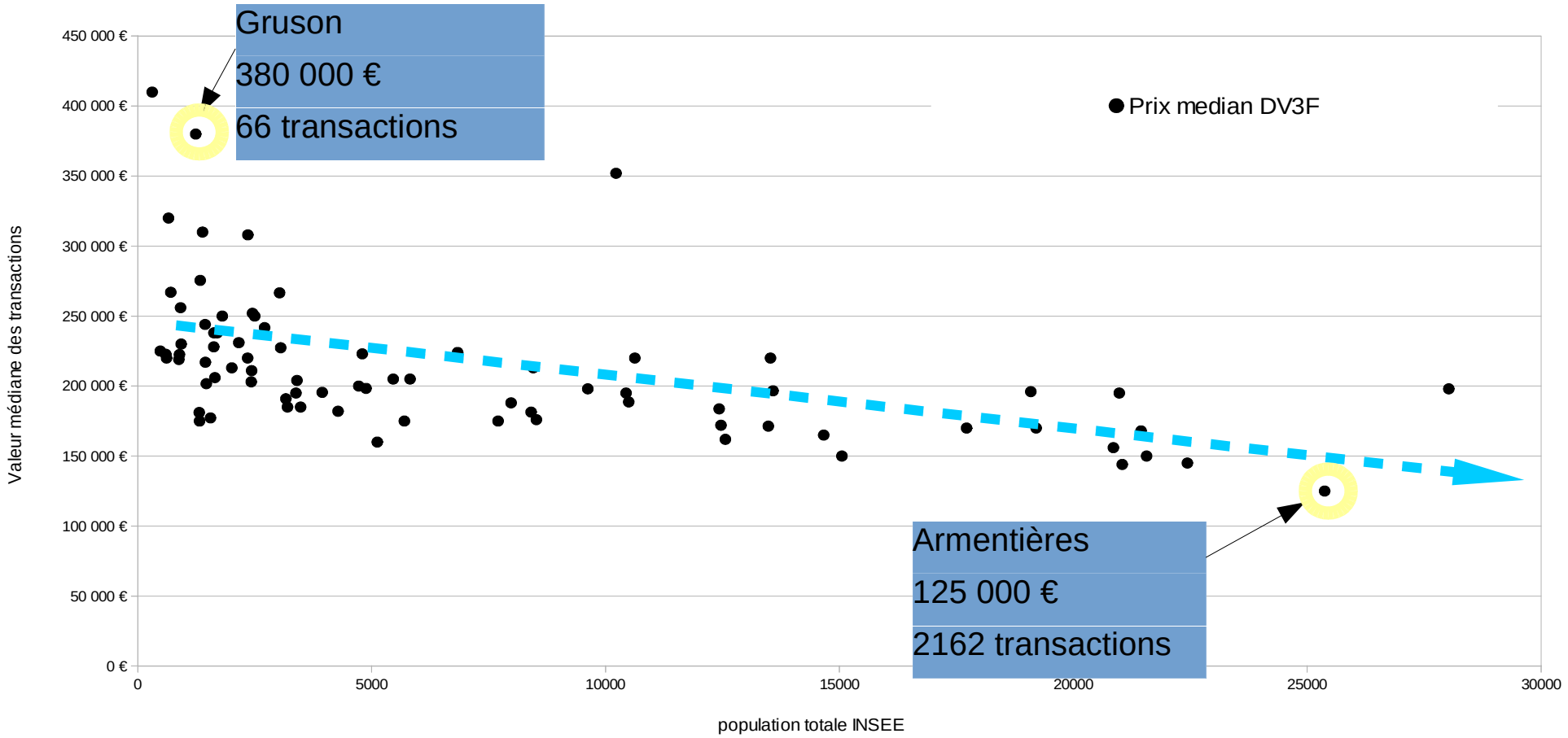
de l'élève qui vont influer à la fois sur les notes et sur le temps passé à la pratique du sport : l'interprétation est « biaisée »

Préconisation: **le meilleur résultat s'observe lorsque le groupe est homogène, donc au niveau de chaque classe !**

La preuve par l'exemple

Le paradoxe de SIMPSON appliqué au marché immobilier de la MEL

Prix médian des transactions honereuses de logements entre entre particuliers (2010-2017)
sur la Métropole européenne de Lille pour les communes < 30 000 habitants, sources DV3F et INSEE



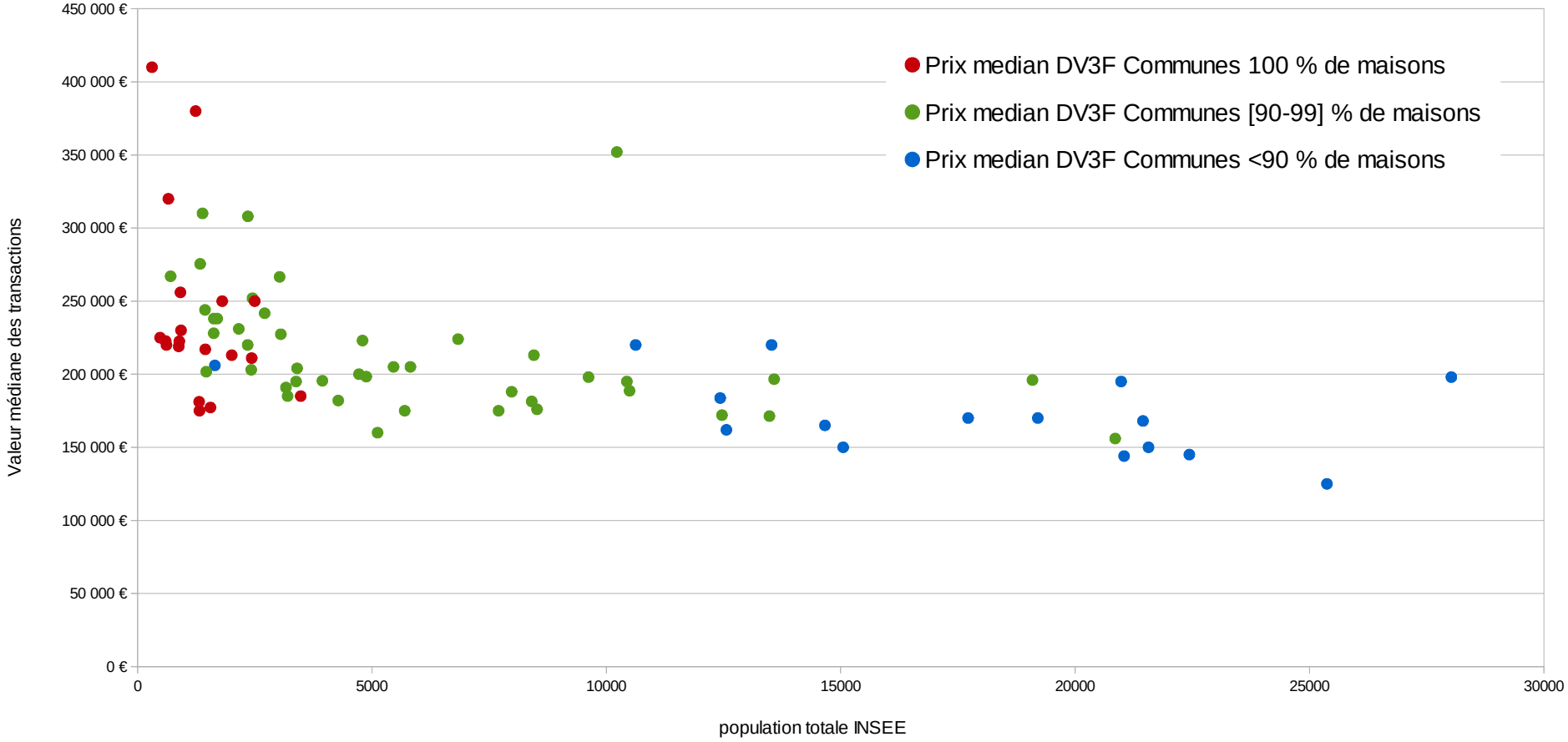
Tendance observée : plus la population augmente, plus les prix baissent.

La preuve par l'exemple

Le paradoxe de SIMPSON appliqué au marché immobilier de la MEL

Prix médian des transactions honereuses de logements entre entre particuliers (2010-2017)

sur la Métropole européenne de Lille pour les communes < 30 000 habitants, sources DV3F et INSEE



Ici les « facteurs de confusion » sont :

- l'effet de structure : les maisons valent globalement plus cher
- les effectifs : bien moindre dans les petites communes

Les territoires d’analyse

Quelle maille pour quelles données ?

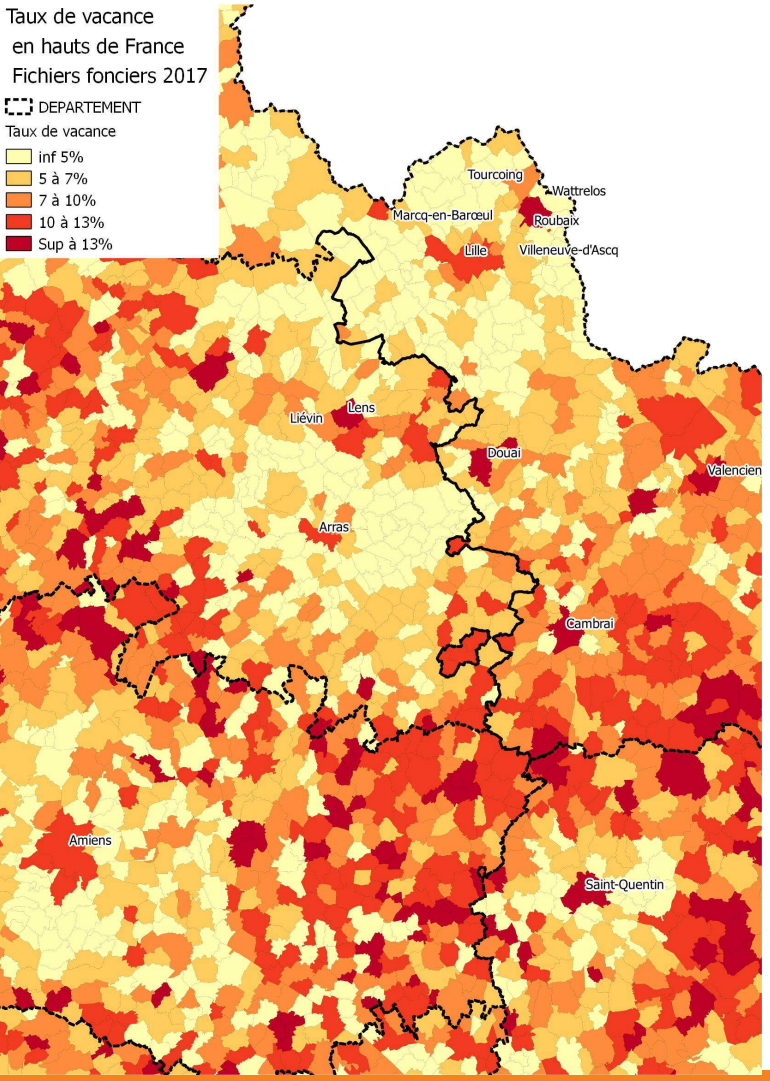
Chaque donnée aura une réalité différente en fonction de l’échelle à laquelle elle est traitée.

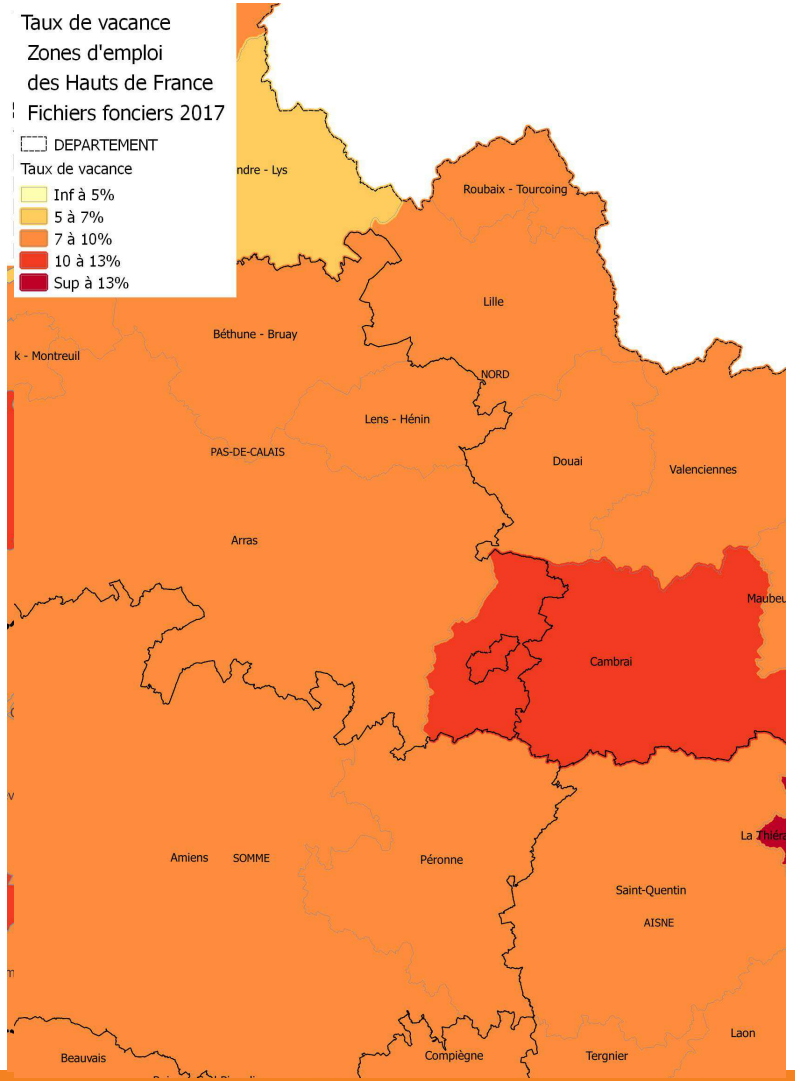
La connaissance de la thématique permettra de choisir des indicateurs différents pour des périmètres différents et ainsi éviter :

- un effet de dilution entre une donnée à maille communale et une donnée à maille plus macro
- Un effet d’occultation ou d’exacerbation de certains indicateurs.

La preuve par l’exemple

Les territoires et l’analyse appliqué au taux de vacance





L’effet de dilution entre une donnée à maille communale et une donnée à maille plus macro

L’analyse en composantes principales (ACP)

DESCRIPTION DU FACTEUR 2
PAR LES VARIABLES CONTINUES ACTIVES

COORD.	POIDS	LIBELLE DE LA VARIABLE	MOYENNE	ECART-TYPE	NUMERO
-0.75	135.00	SG_100_Tx arti 9005	0.55	0.72	1
-0.58	134.00	D_46_Prix terrains ind 94-02 r	19232.46	5716.96	2
-0.57	135.00	F_32_moy revenus ucm 2009	16678.01	3229.59	3
-0.57	135.00	F_37_moy rev nvx emm ucm	15475.59	3933.61	4
-0.54	135.00	SG_96_Part arti surf com 9098	0.44	1.04	5
-0.54	135.00	F_33_méd revenus ucm 2009	14346.78	2343.51	6
-0.54	135.00	F_38_méd rev nvx emm ucm	13821.75	2965.67	7
-0.52	135.00	F_31_%mges sup100%	38.72	10.95	8
-0.52	135.00	SG_97_Tx arti 9098	0.32	1.21	9
-0.49	135.00	S_22_Taux croiss rés 9908	12.29	8.13	10
-0.49	135.00	SG_93_Evarti% 90-98	3.07	13.94	11
-0.49	135.00	SG_83_Evart% 9805	6.07	7.49	12
-0.47	135.00	SG_87_Tx arti 9805	0.82	0.97	13
-0.47	135.00	SG_119_Surfcult arti 9805	3.98	8.26	14
-0.45	135.00	I_08_EvRP9906	7.78	5.68	15
-0.44	135.00	SG_120_Surf prairies arti 9806	3.35	3.79	16
0.43	135.00	SG_114_Surf habitat 9005	36525.57	34043.43	17
-0.43	135.00	S_23_SHON activité 9908	12007.62	33347.43	18
-0.43	135.00	SG_90_Evarti 9098	3.18	16.47	19
-0.41	135.00	SIG_86_Part arti surf com 9805	1.09	1.81	20
		ZONE CENTRALE			
0.02	135.00	SG_68_Part arti 05	22.34	17.56	61
0.03	135.00	SG_73_Natu 05	99.48	280.17	62
0.07	135.00	I_124_Part actifs Cambrai	1.81	3.75	63
0.07	135.00	SG_70_Part natu 05	9.51	13.28	64
0.09	135.00	I_123_Part actifs Douai	1.41	2.48	65
0.14	135.00	S_128_Part ind gr 9908	17.16	21.14	66
0.15	135.00	S_129_Part coll 9908	10.30	18.57	67
0.15	135.00	I_65_Part TC sortants	5.69	4.14	68
0.18	135.00	I_64_Part TC	4.69	3.20	69
0.28	135.00	SG_106_Denspoparti 06	1568.63	680.55	70
0.32	135.00	SG_91_Evagri 9098	-8.54	19.77	71
0.32	135.00	SG_94_Evagri% 9098	-1.81	3.08	72
0.32	135.00	SG_98_Txagri 9098	-0.23	0.47	73
0.33	135.00	SG_84_Evagri% 9805	-3.01	4.34	74
0.34	135.00	SG_88_Tx agri 9805	-0.45	0.68	75
0.35	135.00	F_30_%mges 60-100%	29.96	3.75	76
0.38	135.00	I_50_Tx non moto	16.70	7.46	77
0.41	135.00	SG_81_Evagri 9805	-9.83	11.24	78
0.43	135.00	I_43_TCSP issus UUVal	1.51	1.27	79
0.48	135.00	F_29_%mges 0-60%	1.32	9.26	80

L'Analyse en Composantes Principales (ACP) est l'une des méthodes d'analyse de données multivariées les plus utilisées.

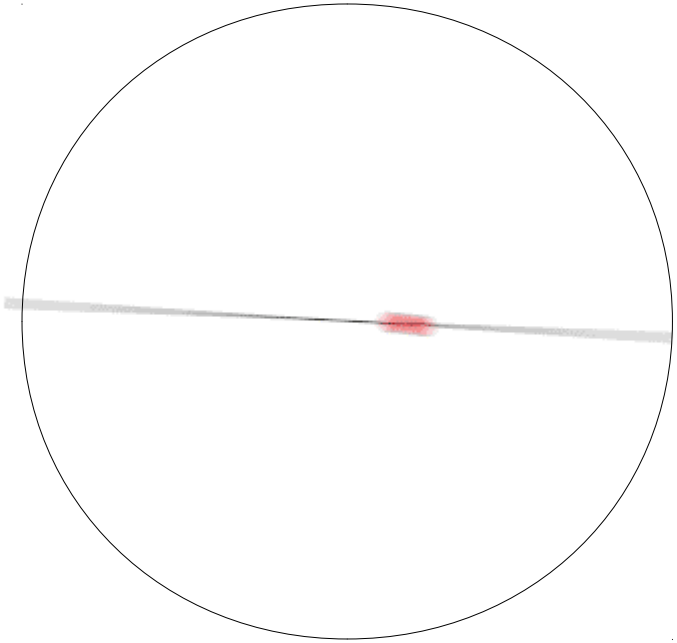
Elle permet d’exploiter de très grands jeux de données multidimensionnels constitués de variables quantitatives.

Qu’est-ce ?

La preuve par l’exemple

L’analyse en composante principale (ACP)

cercle des corrélations



Il s'agit d'une approche à la fois géométrique et statistique.

Cet outil statistique va nous permettre de travailler l’ensemble de nos données en représentant les N variables de notre tableau, au travers autant d’axes.

Afin de représenter l’ensemble des ces axes, imaginez une sphère au sein de laquelle chaque axe représente une variable : c’est le **cercle des corrélations**.

Rendre intelligible le big data

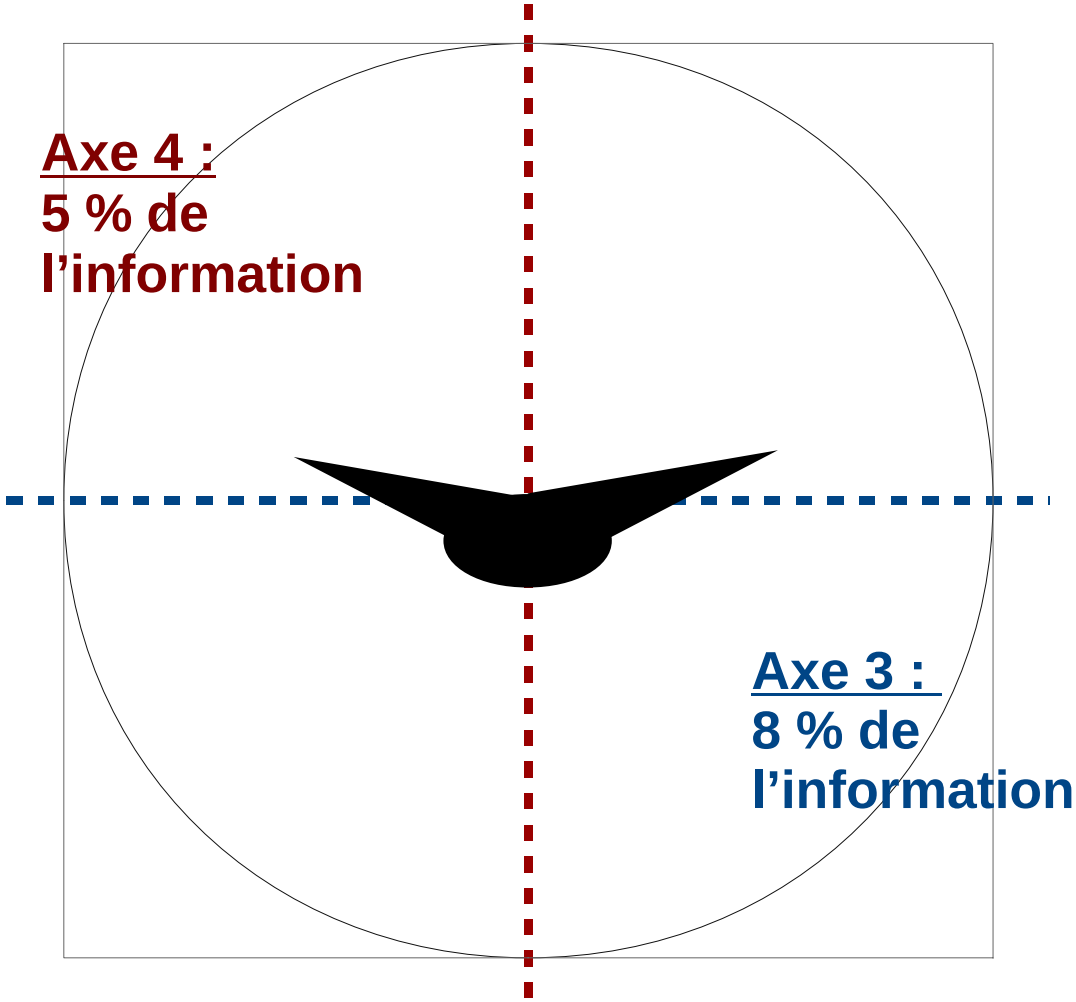
L’analyse en composantes principales (ACP)

L'analyse en composantes principales vous donnera :

- La meilleure représentation de l’ensemble de ces données, au travers un % d’explication de l’ensemble
- ainsi qu’une clé de lecture via les variables les plus explicatives, du comportement de chacune de ses composantes, puisque chaque élément du tableau pourra être positionné dans cet espace.

La preuve par l’exemple

L’ACP appliquée à ?



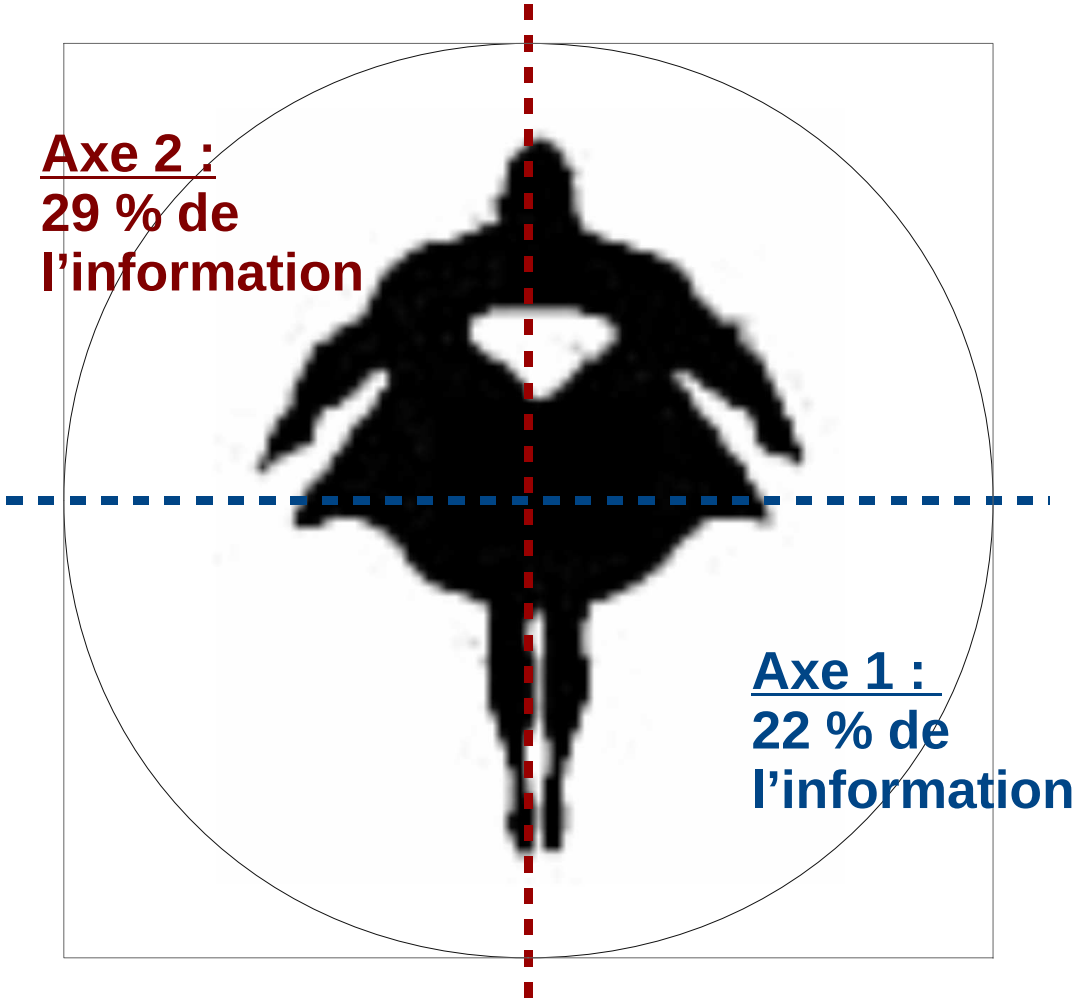
Qu’est-ce ?

Ici sont représentés 13 % de l’information contenue dans le tableau de données, c’est peu ...

- **Un oiseau ?**
- **Une fusée ?**
- **Un avion ?**

La preuve par l’exemple

L’ACP appliquée à la pop culture



L’ACP révèlera le meilleur angle d’observation des individus et l’illustrera par ses composantes principales.

→ Ici sont représentés 51 % de l’information contenue dans le tableau de données, c’est le maximum possible sur 2 axes mais c’est suffisant pour faire comprendre de quoi il s’agit !

La preuve par l'exemple

L'ACP appliquée au besoin en logements