

Section 5: Manipulation et représentation de données : éviter des écueils classiques

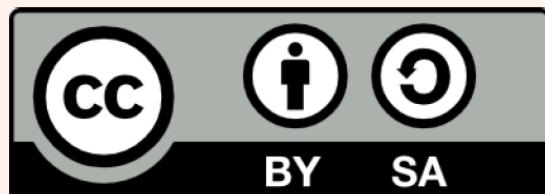
Culture générale des données

Dataactivist, 2018-2019

Ces slides en ligne : <https://dataactivist.coop/SPoSGL/sections/section5.html>

Sources : <https://github.com/dataactivist/SPoSGL/>

Les productions de Dataactivist sont librement réutilisables selon les termes de la licence [Creative Commons 4.0 BY-SA](#).



Plan du cours

1. Les pièges statistiques récurrents

Lire l'article d'Hervé le Bras "La France inégale : Qui vote FN ? Pas forcément ceux à qui l'on pense"

2. Représentations graphique et spatiale, attention danger !

Jouer avec le générateur aléatoire de corrélations absurdes

3. Les limites inhérentes aux indicateurs

Lire l'article du Monde "Pourquoi les chiffres sur la délinquance sont à prendre avec précaution"

Bibliographie

quiz section 5

En guise d'introduction... :)

Rappel sections 1 et 2 : les données sont rarement "neutres"...



© Xavier Gorce

1. Les pièges statistiques récurrents

Médiane ou Moyenne ?

- Salaire mensuel **moyen** net en France en 2014 (secteur privé) : **2 225€**
- Salaire mensuel **median** net en France en 2014 (secteur privé) : **1 783€** Soit une différence de près de 450€ !

Déciles	Ensemble		Hommes		Femmes	
	2014	Évolution (%)	2014	Évolution (%)	2014	Évolution (%)
D1	1 206	0,1	1 257	-0,2	1 164	0,4
D2	1 349	0,1	1 419	-0,2	1 279	0,3
D3	1 480	0,1	1 565	-0,1	1 386	0,4
D4	1 620	0,1	1 717	0,0	1 500	0,5
D5 ou Médiane	1 783	0,1	1 893	0,0	1 636	0,5
D6	1 988	0,2	2 113	0,1	1 812	0,5
D7	2 264	0,4	2 425	0,3	2 051	0,5
D8	2 716	0,8	2 955	0,7	2 402	0,9
D9	3 599	1,0	3 940	0,8	3 100	1,5
C95	4 589	0,9	5 089	0,7	3 839	1,8
C99	8 163	0,8	9 375	0,9	6 183	1,7
Moyenne	2 225	0,5	2 410	0,4	1 962	0,9

Médiane ou Moyenne ?

Rappelez-vous que **la moyenne est très sensible aux valeurs extrêmes...**



A screenshot of a Twitter post. The profile picture is a cartoon drawing of a person's head. The username is **Pemv #GiletsJaunes #BotRusse #Mat...** and the handle is **@numerobis21**. There is a red "Suivre" button and a dropdown arrow. The tweet text is:
Vous savez ce qui est bien avec le salaire moyen?
C'est que si un milliardaire entre dans un PMU, les piliers de bars deviennent automatiquement millionnaire "en moyenne"
:)

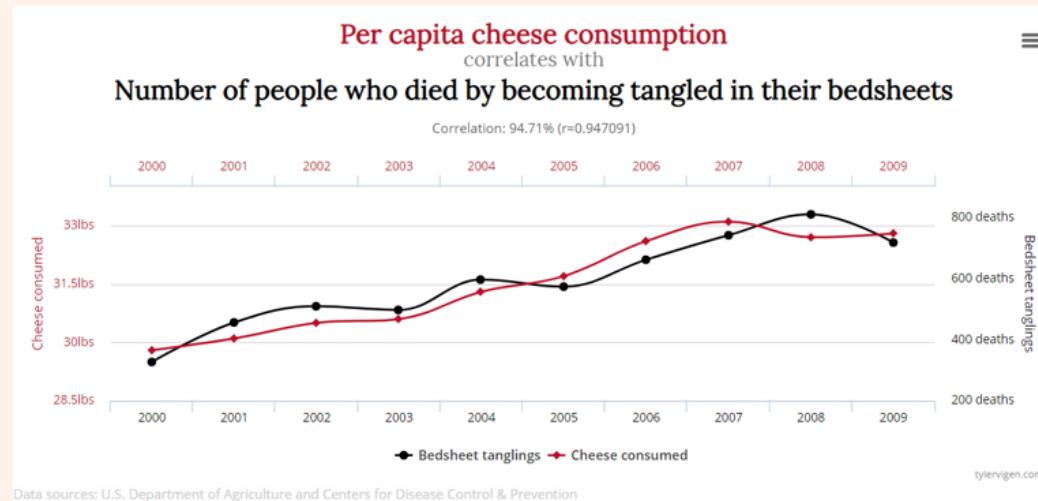
In reply to **L'Est Républicain** (@lestrepublicain) at 05:08 - 30 août 2018:

2 505 euros nets par mois, c'est le salaire moyen d'un fonctionnaire bit.ly/2BZ50kR

Corrélation ou causalité ?

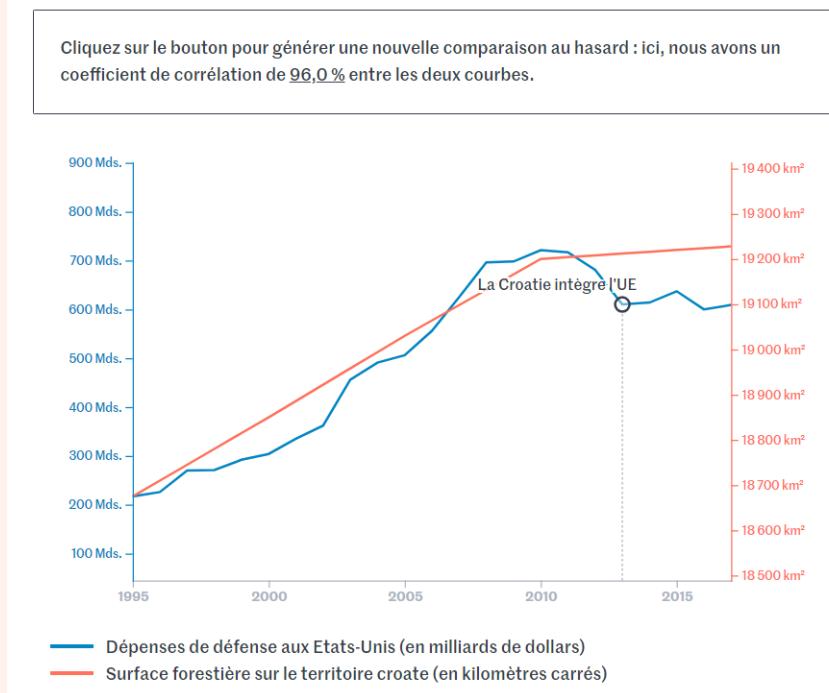
Rappel : une corrélation fortement positive, avec un coefficient de corrélation (r) $> 0,5$, signifie seulement que deux variables évoluent dans le même sens. Cela ne dit **rien** sur le possible lien entre elles

- Faites attention aux **corrélations fallacieuses** !
- Exemple ici avec la corrélation quasi parfaite entre la consommation de fromage par personne et le nombre de personnes qui décèdent étranglées dans leurs draps



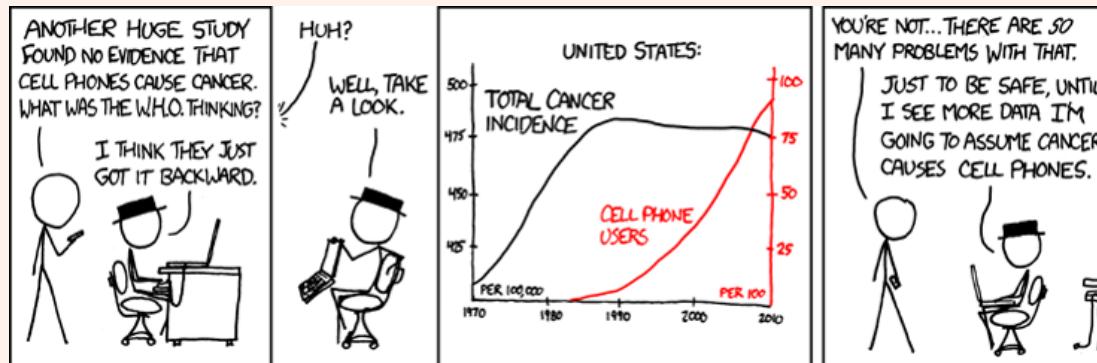
Corrélation ou causalité ?

A l'instar du site parodique "["Spurious Correlations"](#)" dont le graphique précédent est issu, les Décodeurs du monde ont récemment lancé un [générateur aléatoire de comparaisons](#). Un bon moyen pour ne plus jamais faire la confusion entre corrélation ou causalité !



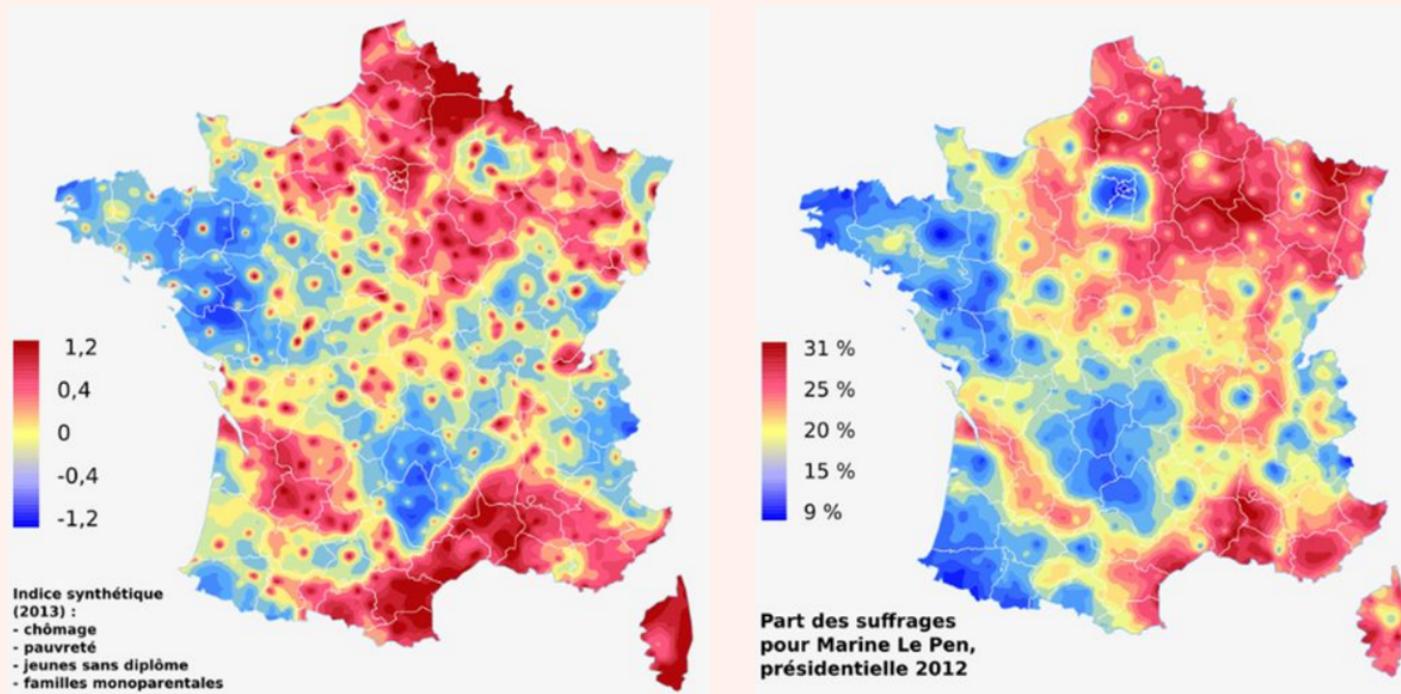
Corrélation ou causalité ?

- En ce qui concerne les corrélations, il faut être attentif au **nombre d'unités observées** : moins il y a d'observations, plus la corrélation est facilement élevée. Ce n'est pas pareil de regarder la corrélation entre 2 variables sur 13 régions ou sur 35 000 villes



- Il faut également être prudent sur l'échelle spatiale utilisée : la corrélation entre le % d'immigrants et le vote FN est positif à l'échelle du département. Mais à l'intérieur de ces départements, **ce sont dans les zones avec la plus faible proportion d'immigrants que les personnes votent FN**. Ainsi, au niveau micro la corrélation peut être inversée que ce qui était observé au niveau macro

Corrélation ou causalité ?



Credits Hervé Le Bras

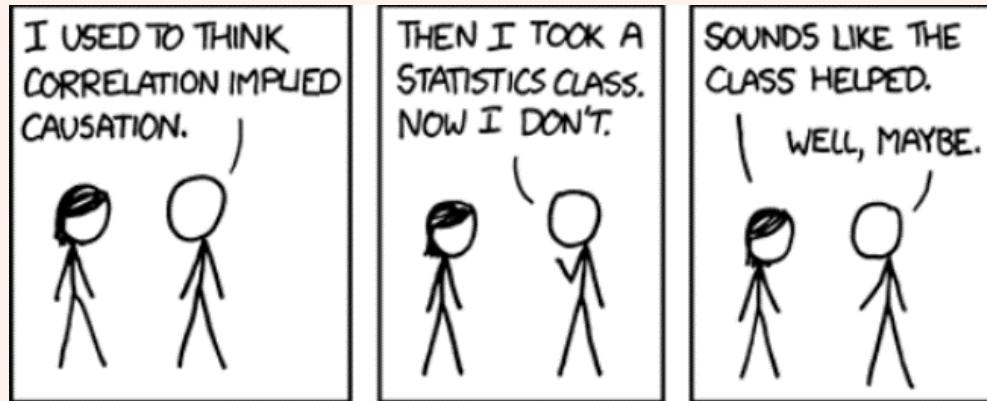
Corrélation ou causalité ?

- Les zones combinant le plus de difficultés sont corrélées avec celles ayant une part des suffrages élevée pour Marine Le Pen à la présidentielle de 2012
- Est-ce que cela signifie que les personnes qui ont le plus de difficultés votent FN ?
 - Pas forcément, pourquoi ?

Il serait cependant inexact d'en déduire que ce vote est celui des pauvres et des laissés pour compte. Ces derniers s'abstiennent le plus souvent. On doit plutôt constater que c'est le vote des régions pauvres, celles où beaucoup craignent les accidents de la vie car ils voient leurs proches atteints par eux.

Hervé Le Bras, dans son article pour The Conversation "La France inégale : Qui vote FN ? Pas forcément ceux à qui l'on pense" (**obligatoire**)

Corrélation ou causalité ?



Credits XKCD

Les facteurs cachés

Il existe des facteurs qu'on ne voit pas ou auxquels on ne pense généralement pas mais qui peuvent fortement influer sur des résultats.

- Exemple ici avec la forte baisse du nombre de décès sur les routes en France entre 2003 et 2014

(**) Taux de variation annuel moyen												
	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Période 2000-2014 : Nombre de décès par accident de la circulation												
<i>Unité : aucune Zone géographique : France entière (hors Mayotte)</i>												
Total	5913	5277	5259	4718	4514	4236	4281	3938	3708	3465	2988	3052

- Est-ce que les politiques actives de lutte contre l'insécurité routière mises en place dans les années 2000 sont la seule explication de cette baisse ? D'autres facteurs "cachés" peuvent aussi avoir eu un rôle significatif

Les facteurs cachés

Quelques exemples :

- Le **taux d'occupation des voitures** a fortement baissé (donc mécaniquement moins de morts à accidents comparables)
- Les **conducteurs "novices"**, i.e avec moins de 12 années de pratique, sont de moins en moins nombreux
- Les personnes blessées dans des accidents de la route sont **mieux pris en charge par les services d'urgence** que par le passé
- La **qualité des voitures** s'est fortement améliorée, et notamment l'aspect sécurité (airbags)

Le périmètre du dénominateur

Lorsqu'il est question de **ratio**, il est utile de se demander ce que comprend le dénominateur de ce ratio. Le périmètre du dénominateur influence évidemment le résultat obtenu.

- Exemple avec le taux de chômage des jeunes en France : à la fin 2017, le taux de chômage des 15-24 ans atteignait 23%
- Est-ce que cela signifie que près d'un jeune sur 4 est au chômage ?



Source

Le périmètre du dénominateur

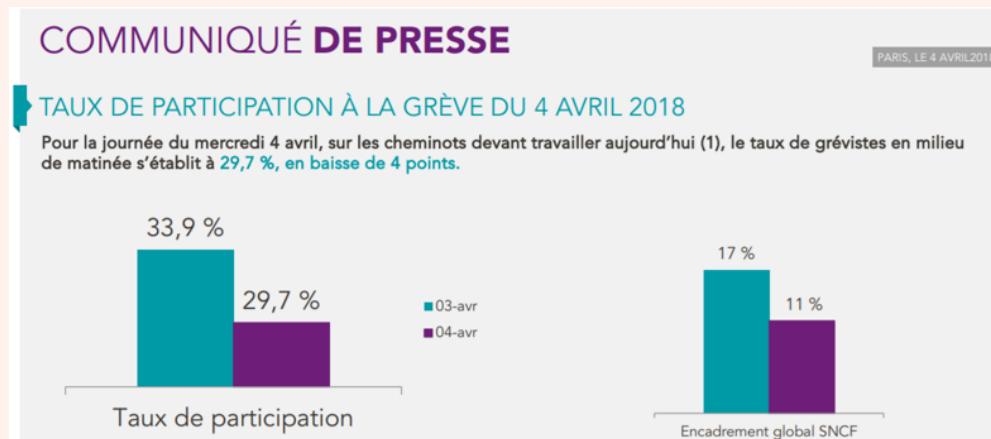
- Est-ce que cela signifie que près d'un jeune sur 4 est au chômage ?
- Non ! Le dénominateur prend en compte les "**personnes actives**", c'est-à-dire ayant un emploi ou à la recherche d'un emploi
- Mais **seulement 37% des 15-24 ans sont actifs** ! Beaucoup des autres sont encore en études secondaires ou tertiaires
- Et ceux qui sont déjà sur le marché du travail sont souvent ceux qui ont quitté tôt le système scolaire et donc **plus susceptible d'éprouver des difficultés pour trouver un travail**
- Si on rapporte le nombre de 15-24 ans à la recherche d'un emploi sur l'ensemble des 15-24 ans (y compris ceux en études), le taux de chômage tombe à 8,5%, soit près d'un jeune sur 12

Regarder la vidéo "[Un jeune sur quatre au chômage ?](#)"

2. Les représentations graphiques et spatiales problématiques

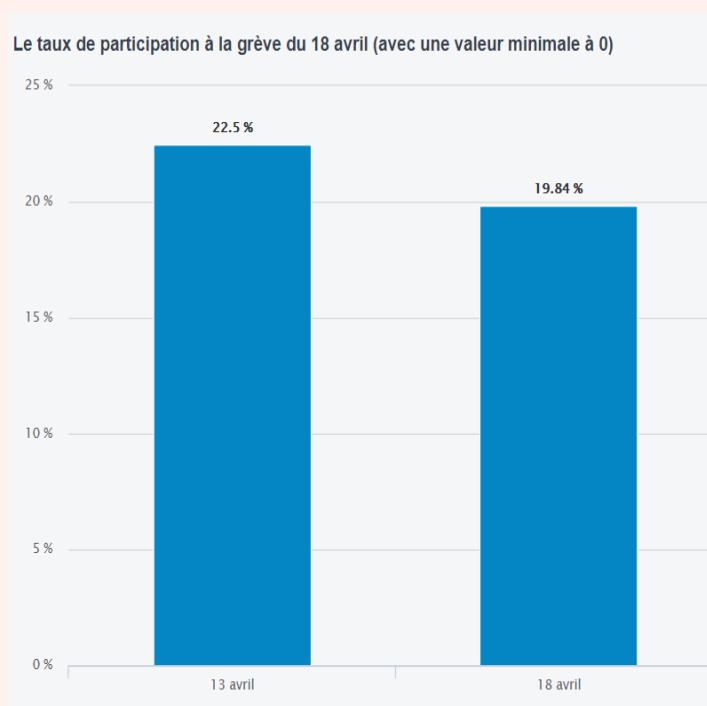
Les représentations graphiques problématiques

- Avoir une **échelle discontinue sur les histogrammes** : il est fréquent de trouver des histogrammes dont l'axe des ordonnées ne démarre pas à 0. Cela a pour conséquence de "tronquer" une partie du graphique et de donner une **présentation biaisée des données**. Cela est souvent réalisé consciemment pour faire passer un message
 - Exemple avec **le communiqué de presse de la SNCF** durant la grève des cheminots (avril 2018) sur le taux de grévistes :



Les représentations graphiques problématiques

Voici ce que donne le même graphique mais avec l'axe des ordonnées démarrant à 0 :

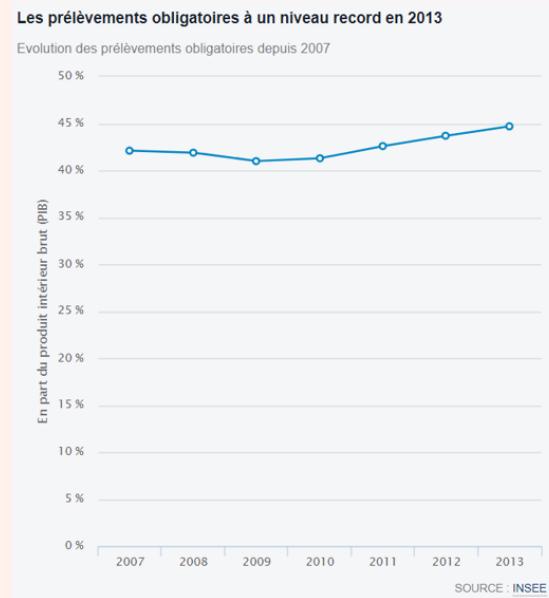


Pas le même rendu, n'est-ce pas ? ;)

NB: de nombreux exemples cités dans cette partie proviennent de cet article des Décodeurs du Monde

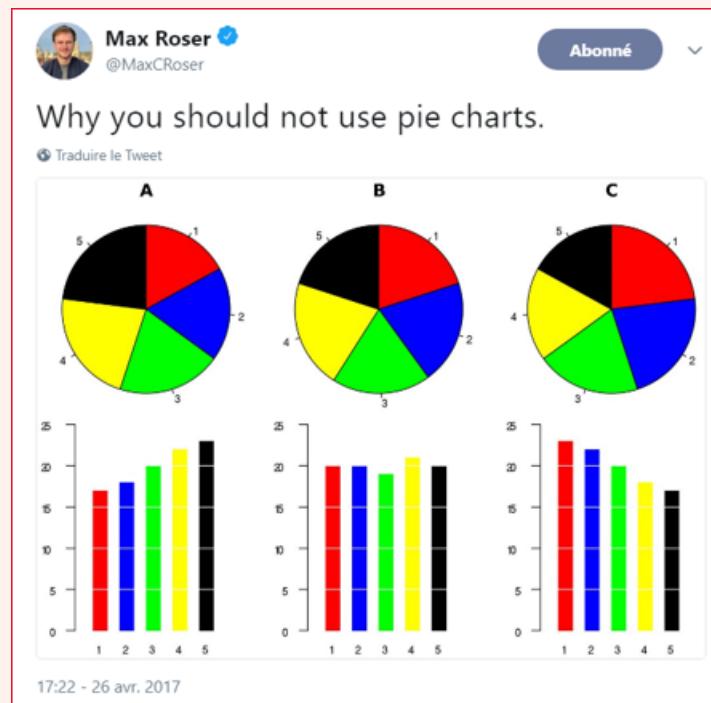
Les représentations graphiques problématiques

- Avoir une **échelle discontinue sur les graphiques en courbe** : même problème que pour les histogrammes. Un autre exemple parlant avec un tract du FN (à gauche) et le même graphique, réalisé par Les Décodeurs, avec l'axe des ordonnées partant de 0 (à droite) :



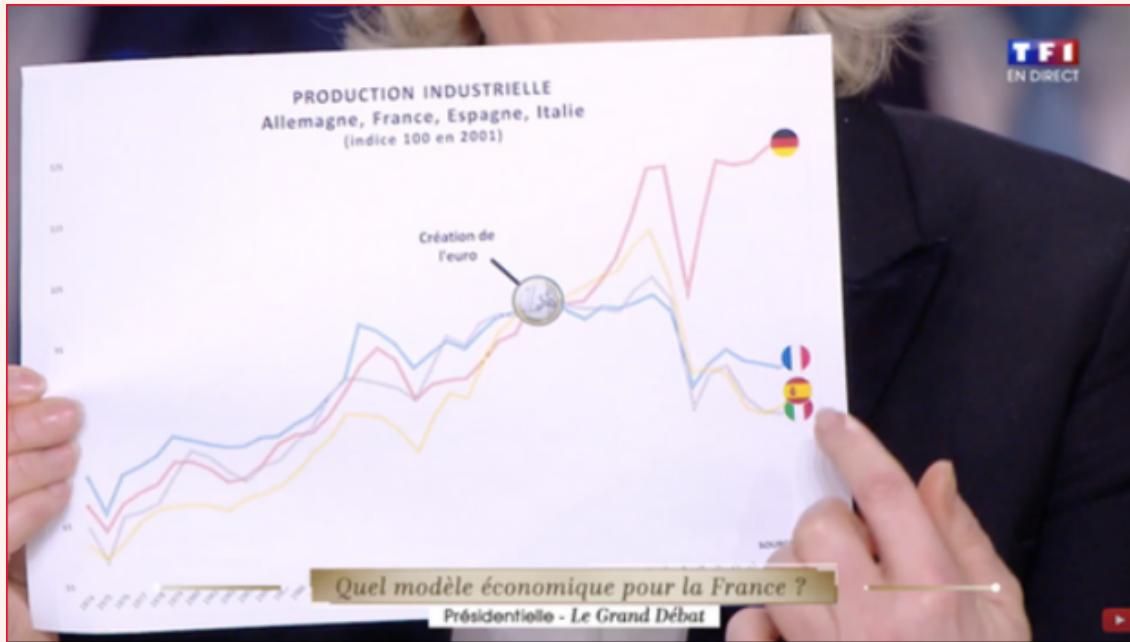
Les représentations graphiques problématiques

- Les **graphiques en camembert** (pie chart) sont à éviter lorsqu'on souhaite représenter des **proportions** :



Les représentations graphiques problématiques

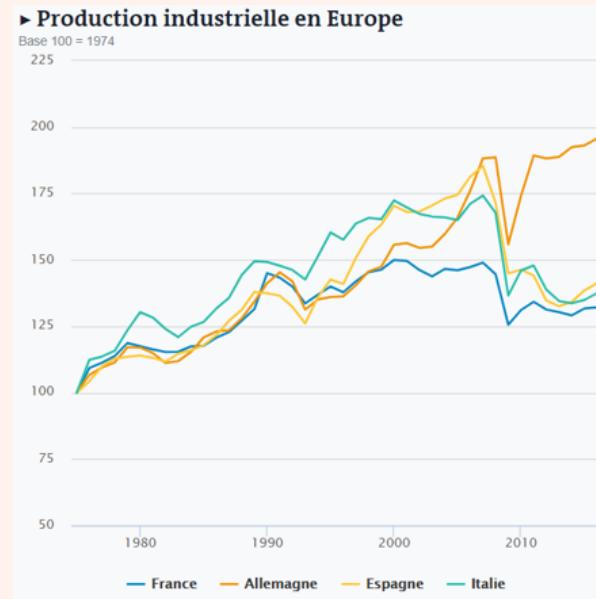
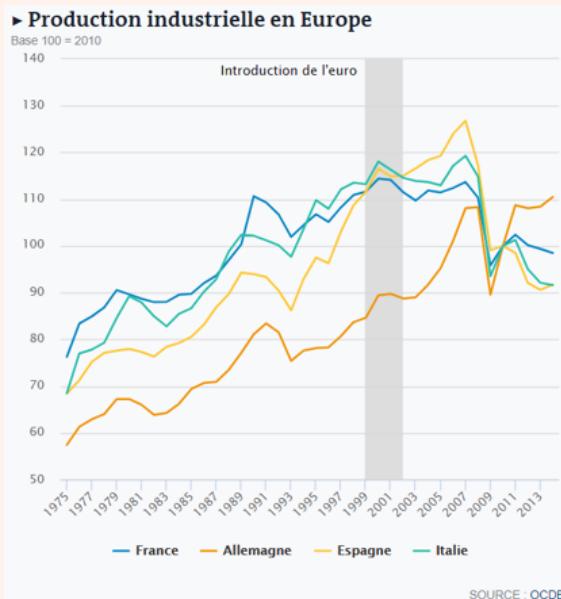
- Les **graphiques avec une "Base 100"** : en fonction de l'année sélectionnée pour être la base 100, les évolutions peuvent fortement changer



Marine Le Pen montre un graphique pendant le 1er débat TV de la présidentielle 2017

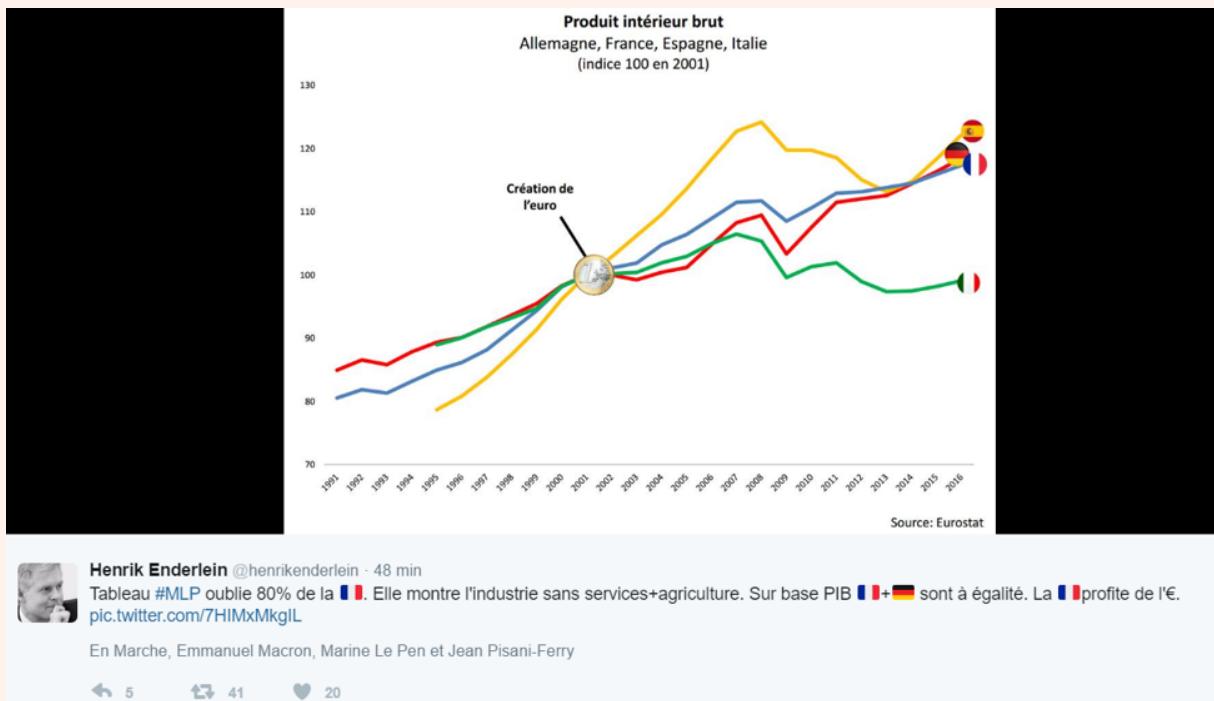
Les représentations graphiques problématiques

- Les **graphiques avec une "Base 100"** : en fonction de l'année sélectionnée pour être la base 100, les évolutions peuvent fortement changer
 - A gauche, Production industrielle en Europe, base 100 = **2010**
 - A droite, Production industrielle en Europe, base 100 = **1974**



Les représentations graphiques problématiques

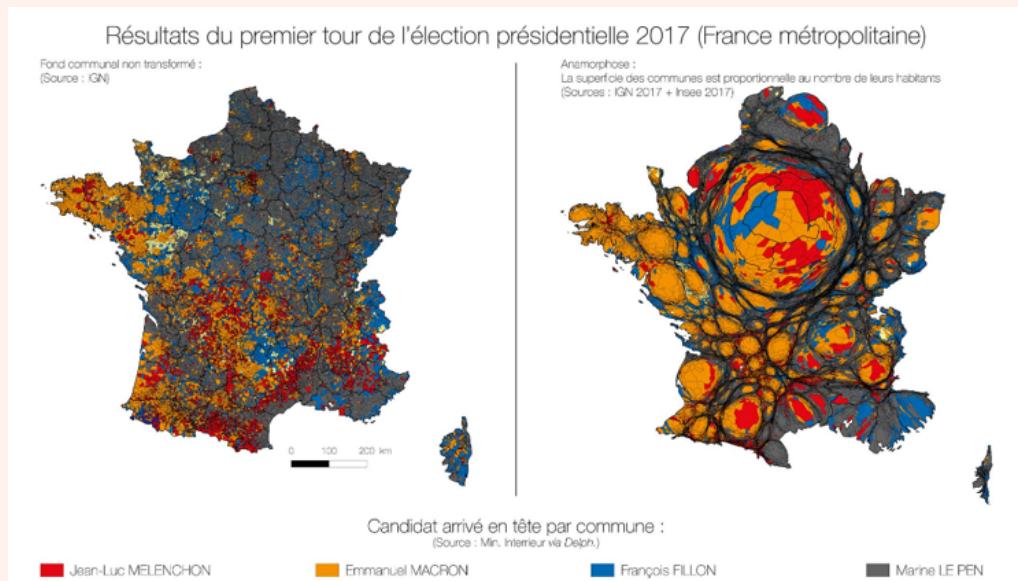
Et si on ne prend pas en compte que l'évolution de la production industrielle, mais celle de l'ensemble du PIB, le graphique évolue significativement



Les représentations spatiales

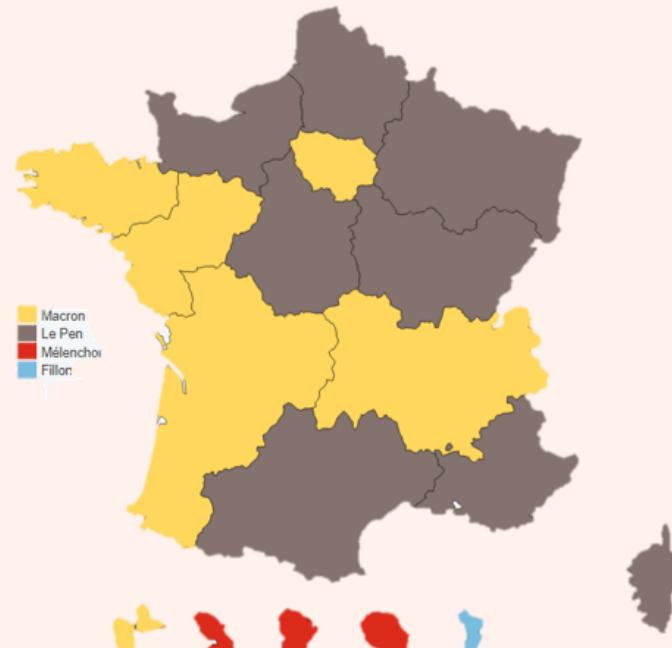
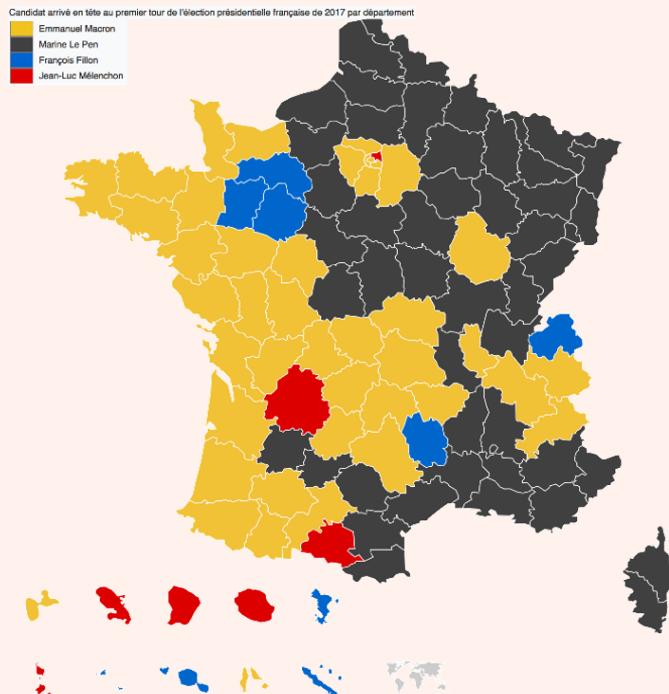
En fonction de l'**unité spatiale** retenue pour représenter graphiquement les données, les résultats peuvent différer fortement d'une carte à une autre...

- Exemple avec la représentation des résultats du 1er tour de l'élection présidentielle française de 2017. A gauche les résultats par commune, à droite une carte à anamorphose où la superficie des communes est proportionnelle au nombre de leurs habitants



Les représentations spatiales

- A gauche les résultats par département, à droite les résultats par région

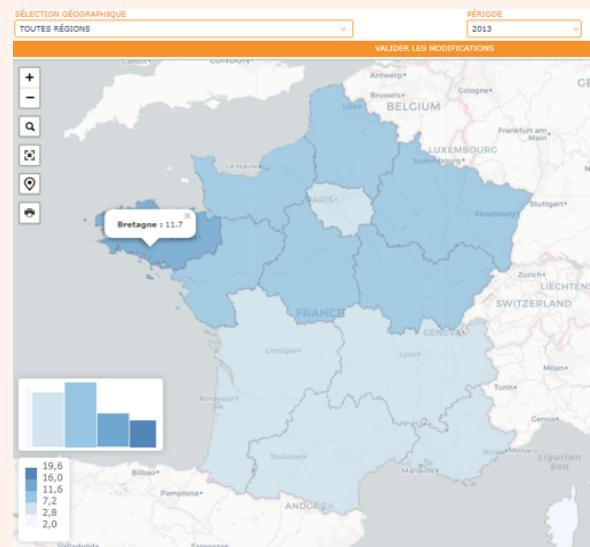
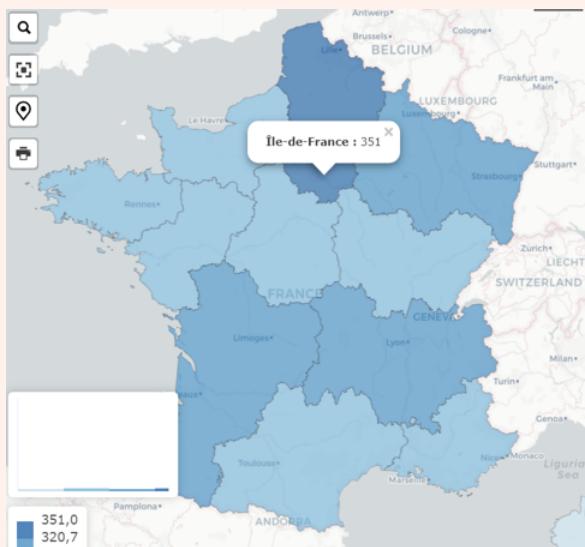


source

Les représentations spatiales

Lorsqu'elles sont représentées sur des cartes, les données doivent souvent être représentées **relativement à la population**, et non pas en valeur absolue

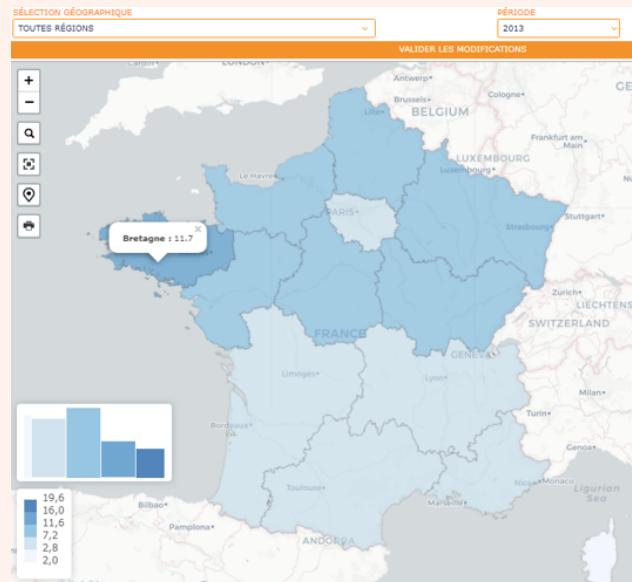
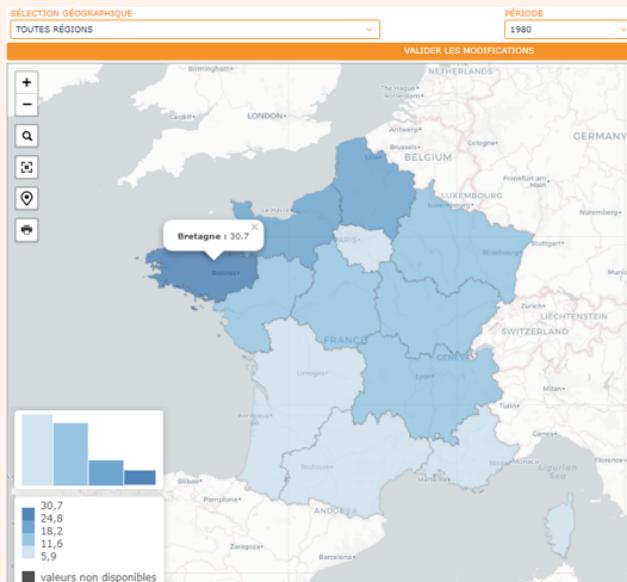
- Exemple ici avec, à gauche, le nombre de décès liés à l'alcoolisme en France en 2013. Et à droite, les mêmes données représentées relativement à la population de chaque région



source

Les représentations spatiales

Soyez attentifs aux couleurs choisies automatiquement par l'outil de visualisation utilisé. Par défaut, **ils adaptent les échelles et les couleurs associées aux nouvelles données**, ce qui peut brouiller les évolutions d'une année à une autre. Ici le taux de décès liés à l'alcoolisme en Bretagne a été divisé par 3 entre 1980 et 2013



source

3. Les limites inhérentes aux indicateurs

Les indicateurs sont-ils neutres ?

- Les indicateurs statistiques sont toujours le fruit de conventions humaines.
Il n'y a pas d'indicateurs "purs" ou neutres
- Un exemple classique est le PIB. Son calcul provient d'une **convention statistique** pour estimer la valeur économique créée par un pays. Quelles limites ?
 - L'impact sur l'environnement ou les inégalités n'est pas pris en compte
 - **La façon dont il est calculé peut changer dans le temps.** Par exemple les revenus tirés de l'activité clandestine de drogues ou de prostitution sont pris en compte dans le calcul du PIB italien depuis 2014, ce qui, mécaniquement, l'a tiré vers le haut
 - L'insee vient d'accepter d'intégrer le trafic de stupéfiants dans le calcul du PIB français

Les indicateurs ne sont pas neutres

Le trafic de drogue va bientôt entrer dans le calcul du PIB français

Par  Le figaro.fr,  AFP agence | Mis à jour le 31/01/2018 à 15:06 / Publié le 30/01/2018 à 20:18



Source

Les indicateurs ne sont pas neutres

Exemple avec les **statistiques sur la criminalité**. En France, il y a une base de données dénommée "Etat 4001" qui agrège les faits connus par la police et la gendarmerie, soit parce qu'il y a eu un dépôt de plainte, soit parce que le fait a été décelé grâce à l'action des forces de l'ordre.

Peut-on déduire l'évolution de la criminalité à partir de cette base de données ?

Non ! Plusieurs raisons pour cela :

- Ce n'est pas une base de données "à jour" : **les faits sont datés en fonction du moment où ils ont été enregistrés dans Etat 4001 et non quand ils ont été commis**. Si quelqu'un dépose une plainte en 2019 pour un fait commis en 2018, le fait sera daté 2019, ce qui brouille évidemment les comparaisons annuelles..

Lire l'article du Monde "[Pourquoi les chiffres sur la délinquance sont à prendre avec précaution](#)" (**obligatoire**)

Les indicateurs ne sont pas neutres

- Par ailleurs, qu'en est-il de tous les crimes ou délits commis qui n'aboutissent pas à un dépôt de plainte ?
- Selon une [étude d'Interstats](#), le service statistique du ministère de l'intérieur, **seulement une victime sur 12 effectue un signalement auprès des forces de sécurité.**



- En conclusion, l'Etat 4001 reflète davantage les évolutions de l'activité des forces de l'ordre que de la criminalité à proprement parler

La loi de Goodhart

La **loi de Goodhart** est un concept qui met en lumière la difficulté de concevoir et mesurer des indicateurs fiables qui sont associés à des enjeux politiques, financiers ou sociaux. Ainsi, **lorsqu'une mesure devient un objectif, elle cesse d'être une bonne mesure**

Tout indicateur statistique cesse d'être un indicateur statistique fiable dès lors qu'il fait l'objet d'enjeux car il devient sujet à des manipulations

Charles Goodhart, économiste

Connaissez-vous des exemples d'indicateurs où la loi de Goodhart peut être observée ?

La loi de Goodhart

La **loi de Goodhart** est un concept qui met en lumière la difficulté de concevoir et mesurer des indicateurs fiables qui sont associés à des enjeux politiques, financiers ou sociaux. Ainsi, **lorsqu'une mesure devient un objectif, elle cesse d'être une bonne mesure**

Tout indicateur statistique cesse d'être un indicateur statistique fiable dès lors qu'il fait l'objet d'enjeux car il devient sujet à des manipulations

Charles Goodhart, économiste

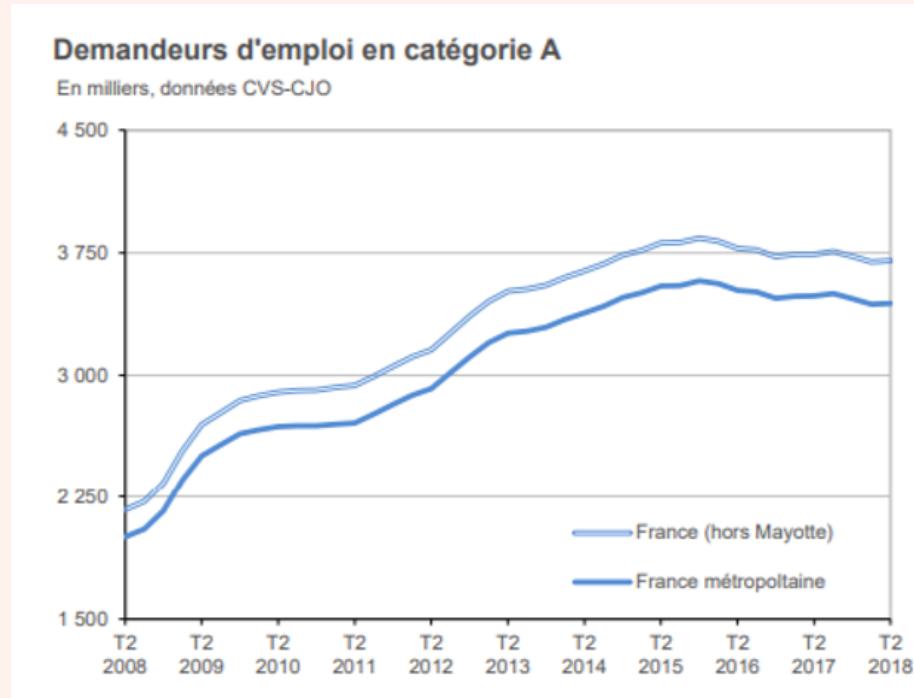
Connaissez-vous des exemples d'indicateurs où la loi de Goodhart peut être observée ?

Le taux de chômage ! On se concentre la plupart du temps sur **le nombre d'inscrits en catégorie A** (et ce sont ces chiffres qui font les gros titres de la presse généralement). Mais est-ce bien pertinent ?

La loi de Goodhart

Exemple : le taux de chômage

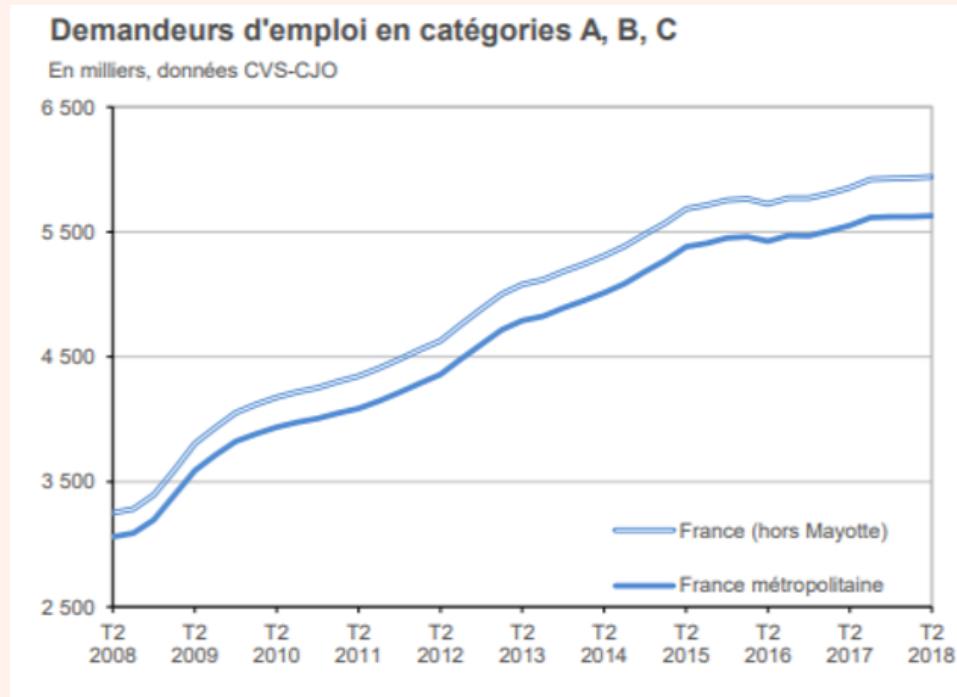
- En fonction de la catégorie à laquelle on s'intéresse, le taux de chômage peut prendre des directions très différentes... Ici focus **catégorie A**



La loi de Goodhart

Exemple : le taux de chômage

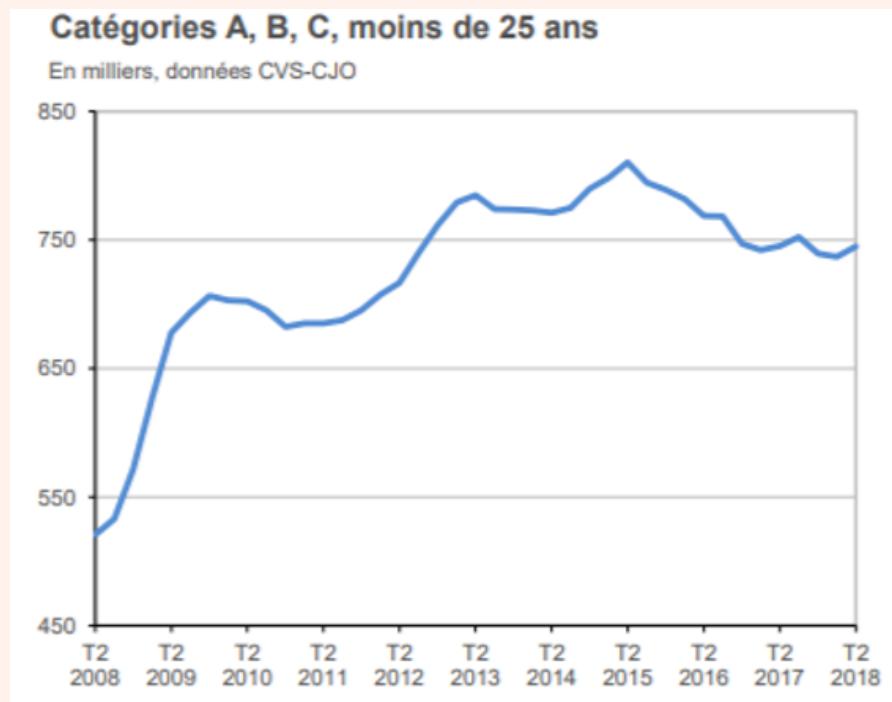
- En fonction de la catégorie à laquelle on s'intéresse, le taux de chômage peut prendre des directions très différentes... Ici focus **catégorie A/B/C**



La loi de Goodhart

Exemple : le taux de chômage

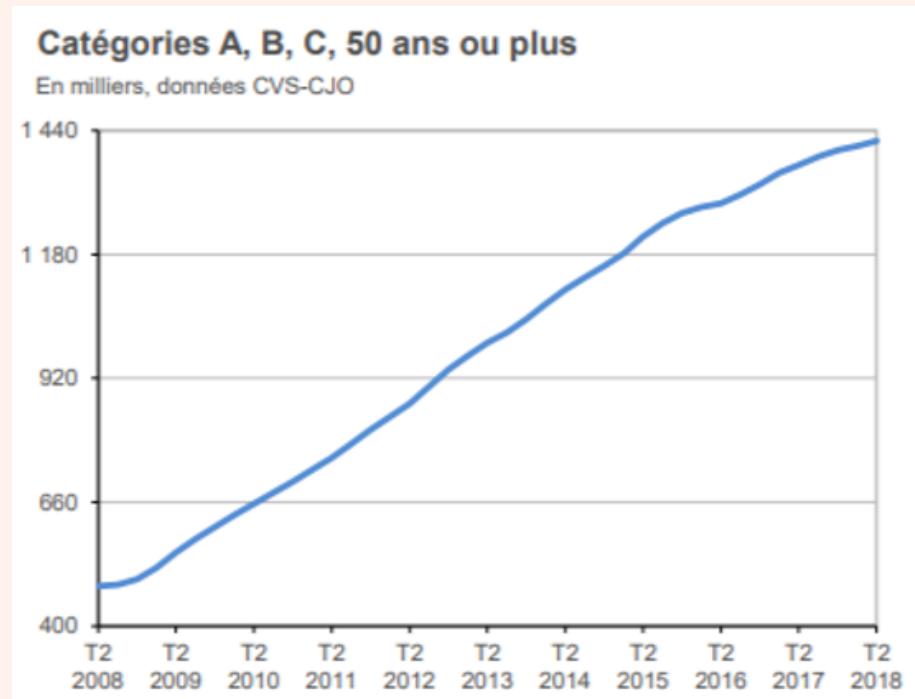
- En fonction de la catégorie à laquelle on s'intéresse, le taux de chômage peut prendre des directions très différentes... Ici focus **catégorie A/B/C, moins de 25 ans**



La loi de Goodhart

Exemple : le taux de chômage

- En fonction de la catégorie à laquelle on s'intéresse, le taux de chômage peut prendre des directions très différentes... Ici focus **catégorie A/B/C, 50 ans ou plus**



La loi de Goodhart

Exemple : le taux de chômage

- Qu'en est-il du **taux d'activité** ? Les chiffres du chômage ne prennent pas en compte les personnes qui sont démotivées et ne recherchant pas ou plus de travail
- **Il peut y avoir des situations où le taux de chômage baisse et le taux d'activité également.** Cela signifie qu'il y a moins de demandeurs d'emplois "officiels" mais de plus en plus de personnes en âge de travailler qui sont inactives
- C'est la situation que connaît les Etats-Unis : Entre 2009 et 2017, le taux de chômage est passé de 9,2% à 4,4% mais parallèlement le taux d'activité est passée de 70% à 67,5%

La loi de Goodhart

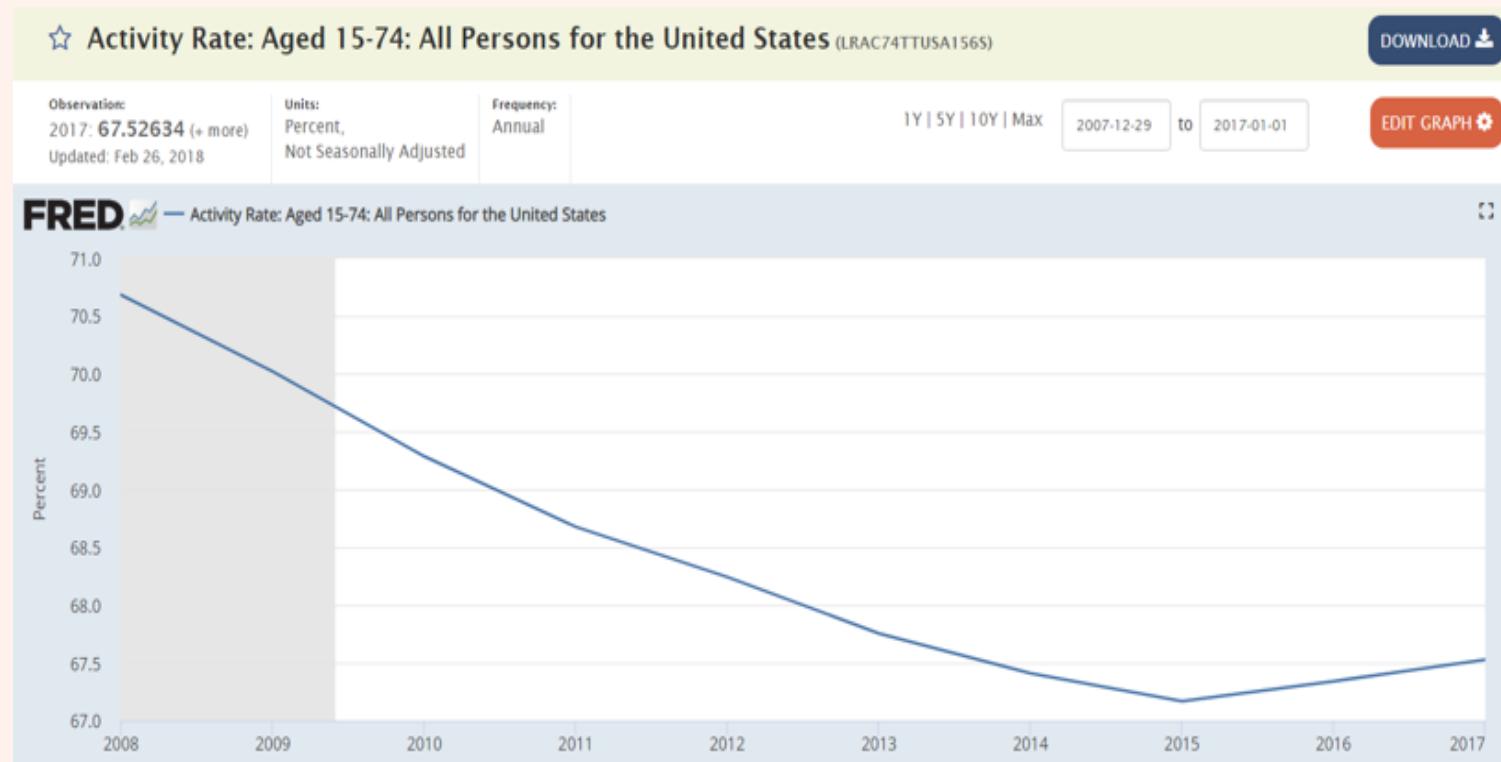
Exemple : le taux de chômage



Taux de chômage US entre 2009 et 2017

La loi de Goodhart

Exemple : le taux de chômage



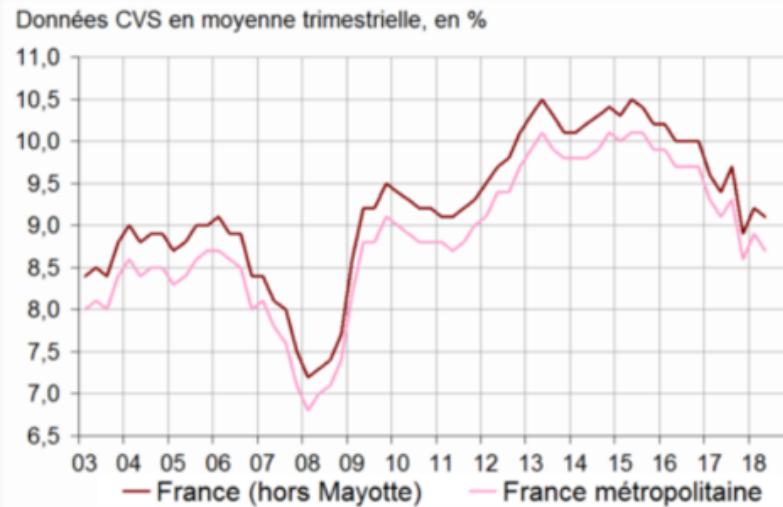
Taux d'activité US entre 2009 et 2017

La loi de Goodhart

Exemple : le taux de chômage

- Au-delà du chômage, qu'en est-il du **halo autour du chômage** ? Comprend le travail occasionnel, le sous-emploi ou les personnes inactives

Graphique1 – Taux de chômage au sens du BIT



Le taux de chômage (au sens du BIT) en France a plutôt eu tendance à baisser depuis 2015

La loi de Goodhart

Exemple : le taux de chômage

- Au-delà du chômage, qu'en est-il du **halo autour du chômage** ? Comprend le travail occasionnel, le sous-emploi ou les personnes inactives

Graphique2 – Personnes dans le halo autour du chômage



Alors que le Halo du chômage en France a lui fortement augmenté sur la même période..

Bibliographie

Bibliographie

- Desrosières, Alain, 2008, L'argument statistique II. Gouverner par les nombres, Paris, Mines ParisTech, les Presses.
- La statistique dans la cité n°8, Société Française de Statistique, Février 2018
- Hervé le Bras, The Conversation "[La France inégale : Qui vote FN ? Pas forcément ceux à qui l'on pense](#)" mis en ligne le 30/01/2018
- Le Figaro "[Le trafic de drogue va bientôt entrer dans le calcul du PIB français](#)" mis en ligne le 31/01/2018
- Le Monde "[Pourquoi les chiffres sur la délinquance sont à prendre avec précaution](#)" mis en ligne le 27/08/2018
- Le Monde "[Présidentielle 2017 : les résultats du premier tour région par région](#)" mis en ligne le 24/04/2017
- Le Monde "[Corrélation ou causalité ? Brillez en société avec notre générateur aléatoire de comparaisons absurdes!!!](#)"

Quiz section 5 : rdv sur votre espace e-campus !

Merci !

Contact : timothee@datactivi.st