Méthodes quantitatives des sciences sociales

2. La fabrique des chiffres : les sources des données quantitatives en sciences sociales et leurs usages

Sciences Po Saint-Germain-en-Laye, 1ère année

2016-2017

Introduction

- Omniprésence des chiffres dans l'espace public: données économiques (PIB, inflation, ratio dette sur PIB, taux de chômage…), sociales (taux de délinquance, indicateurs d'inégalités scolaires…), politiques (résultats d'élections, sondages préélectoraux, sondages d'opinion…)…
- Des sources très diverses: publiques / privées, administratives / scientifiques, globales/ nationales / locales, en ligne / peu accessibles, etc.
- Des usages également très divers: discours politique, expertises, démarche scientifique (« faire preuve »)…

Plan de la séance

- 1. Données de registre ou données d'enquêtes par questionnaire: un débat récurrent.
- 2. Des données alternatives pour les sciences sociales: expériences, biographies collectives, réseaux, discours…
- 3. Les nouvelles sources de données à l'ère numérique.
- 4. Où « trouver » les données ?

Bibliographie

- Patrick Champagne, Remi Lenoir, Dominique Merllié, Louis Pinto, *Initiation à la pratique sociologique*, Paris, Dunod, 1989.
- Corinne Eyraud, *Les données chiffrées en sciences sociales. Du matériau brut à la compréhension des phénomènes sociaux*, Paris, Armand Colin, 2008.
- François Héran, « L'assise statistique de la sociologie », *Economie & Statistique*, 1984, 168, p.23-35.
- Patrick Lehingue, *Le vote. Approches sociologiques de l'institution et des comportements électoraux*, Paris, La Découverte, 2011.
- Laurent Muchielli, « Dix ans d'évolution des délinquances en France », *Regards sur l'actualité*, 336, p.5-15.
- François de Singly, *L'enquête et ses méthodes: le questionnaire*, Paris, Nathan, 1992.

Webographie

- Le collectif pour d'autres chiffres du chômage (ACDC): http://acdc2007.free.fr/
- Le site de Laurent Mucchielli: http://www.laurent-mucchielli.org/index.php?pages/Liens
- Le site de Joël Gombin: http://www.joelgombin.fr/
- Le site de l'équipe OSCJ du CESDIP: http://oscj.cesdip.fr/

1. Données de registre ou données d'enquêtes par questionnaire: un débat récurrent

- Une grande partie des données publiques ne sont pas produites par les chercheurs universitaires, mais par des institutions telles que les administrations, les entreprises, les associations et bien sûr les instituts de statistique publique (en France, l'INSEE, en premier lieu).
- Dans ce cadre, on distingue couramment deux types de sources: les **données de registre** et les **données d'enquêtes par questionnaires**.
- On va donc examiner dans un premier moment les avantages et les inconvénients de ces deux types de sources statistiques.
- D'un point de vue statistique, cela nous conduira à réfléchir sur les notions d'**exhaustivité** et de **représentativité**, sur l'existence de **biais** dans le recueil de l'information.

1. Données de registre ou données d'enquêtes par questionnaire: un débat récurrent

- **Définitions**:
- **Données de registres** = données issues d'un enregistrement administratif (exemple: l'état civil, les données nationales sur les étudiants issues des fiches d'inscription, etc.) ;
- **Données d'enquêtes par questionnaire** = données issues du recueil d'informations à l'aide de questionnaires sur une population exhaustive (recensement) ou sur échantillons (exemple: **Labour Force Survey**).
- Des formes « hybrides » comme les données de la comptabilité nationale ou de certaines enquêtes internationales (EU-SILC sur la pauvreté et l'exclusion).

1. Données de registre ou données d'enquêtes par questionnaire: un débat récurrent

- Seront présentés trois exemples ou cas qui posent des problèmes spécifiques:
- **Exemple 1**: les chiffres de l'emploi et du chômage
- **Exemple 2**: les chiffres de la délinquance et de la criminalité
- **Exemple 3**: les données électorales et politiques

1. Données de registre ou données d'enquêtes par questionnaire: un débat récurrent

- **Exemple 1. Les données sur l'emploi**. En France, elles sont issues de deux sources: l'enquête sur l'emploi (**Labour Force Survey**) et les données des services de l'emploi (**Pôle Emploi**).
- L'existence de ces deux sources explique la diffusion de chiffres révélant des tendances parfois (apparemment) contradictoires. Produites différemment, elles ne mesurent en fait pas la même chose.
- Pôle emploi: les demandeurs d'emploi en fin de mois (DEFM), avec différentes catégories, l'enquête emploi mesure le chômage dit « au sens du BIT ».
- Pôle Emploi: des séries mensuelles / publication trimestrielle pour l'enquête Emploi (depuis 2007).
- Effectifs par catégories, fréquences par catégories, taux d'accroissement.

1. Données de registre ou données d'enquêtes par questionnaire: un débat récurrent

Définition des demandeurs d'emploi en fin de mois (DEFM)*:*

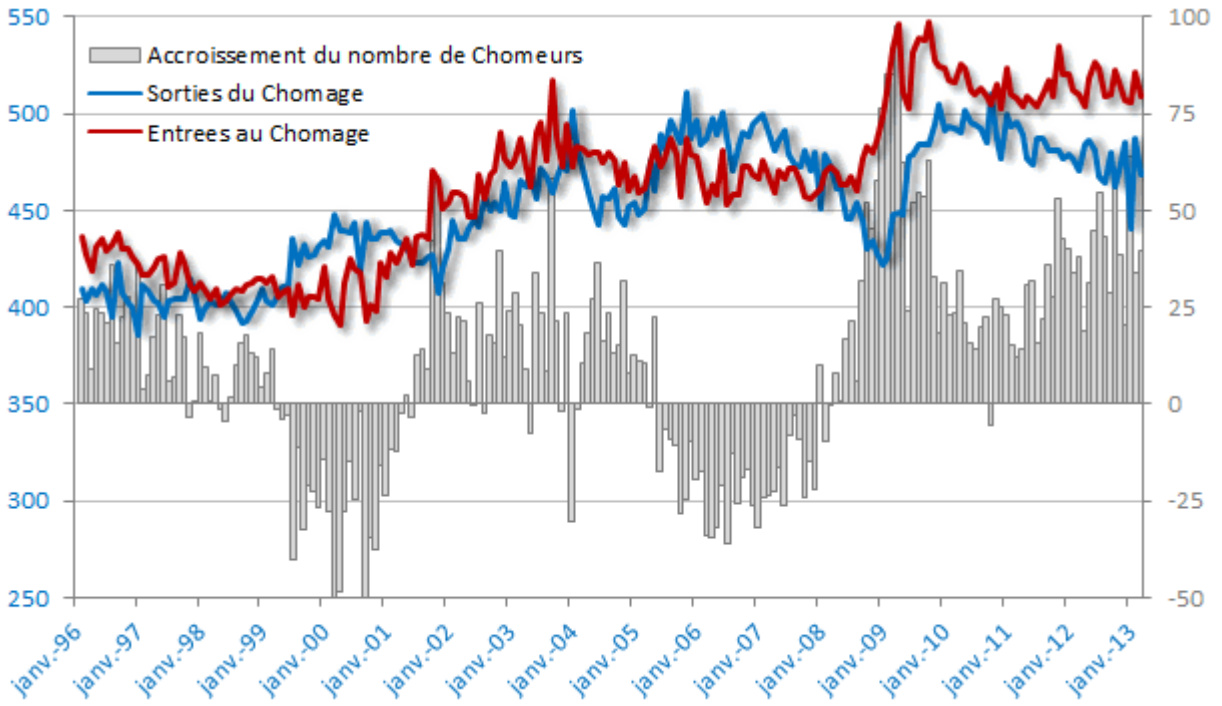
« La plupart des demandeurs d'emploi inscrits à Pôle Emploi sont tenus de faire des actes positifs de recherche d'emploi :
Catégorie A : sans emploi
Catégorie B : exercent une activité réduite courte, d'au plus 78 heures au cours du mois
Catégorie C : exercent une activité réduite longue, de plus de 78 heures au cours du mois
Inscrits à Pôle emploi non tenus de faire des actes positifs de recherche d'emploi :
Catégorie D : sans emploi et non immédiatement disponibles
Catégorie E : pourvus d'un emploi »

Source: DARES.

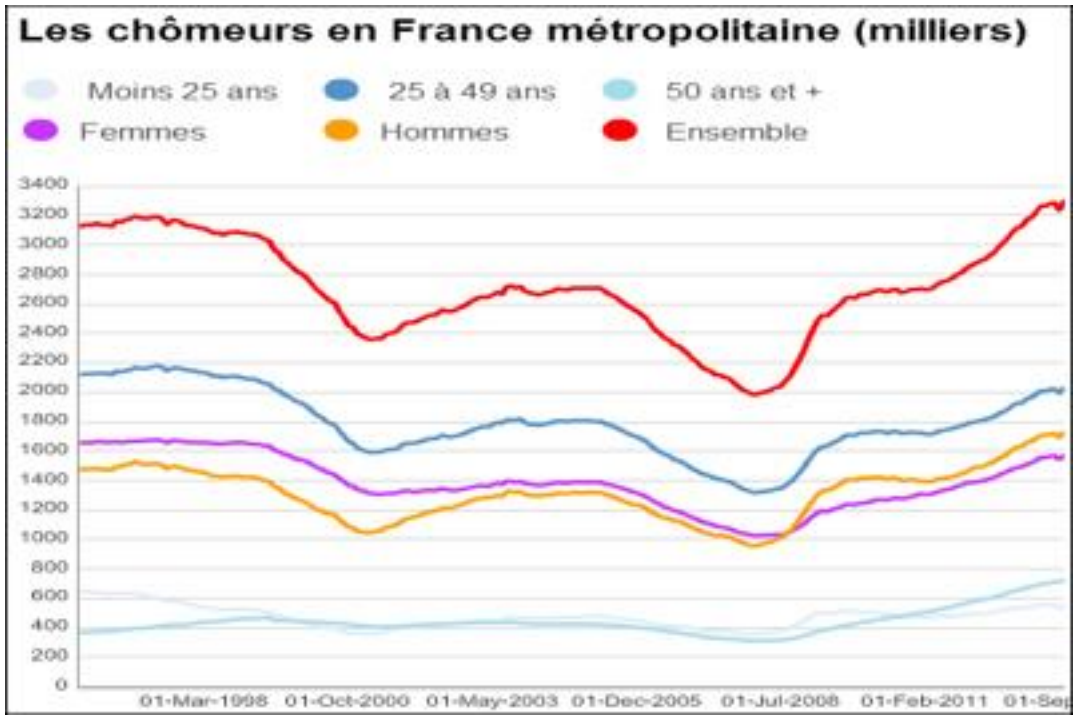
Un ensemble de données complexes, souvent simplifiées par la présentation mensuelle. Voir les débats en 2007 (« collectif ACDC ») autour des radiations, de la simplification dans la présentation, etc.

1. Données de registre ou données d'enquêtes par questionnaire: un débat récurrent

- Evolution du nombre de DEFM de catégories A, B et C



- Catégories A (fin novembre 2015, métropole):3 574 800; A+B+C: 5 442 500



1. Données de registre ou données d'enquêtes par questionnaire: un débat récurrent

- **Définition du chômage**

« En application de la définition internationale adoptée en 1982 par le Bureau international du travail (BIT), un chômeur est une personne en âge de travailler (15 ans ou plus) qui répond simultanément à trois conditions :

- être sans emploi, c'est à dire ne pas avoir travaillé au moins une heure durant une semaine de référence ;
- être disponible pour prendre un emploi dans les 15 jours ;
- avoir cherché activement un emploi dans le mois précédent ou en avoir trouvé un qui commence dans moins de trois mois.

Remarque. Un chômeur au sens du BIT n'est pas forcément inscrit à Pôle Emploi (et inversement). » (Source: INSEE).

- Taux de chômage: ratio du nombre de chômeurs sur le nombre d'actifs.
- Commentaire: une définition restrictive et standardisée à des fins de comparaison internationale. But: éviter les biais liés à des définitions juridiques nationales.

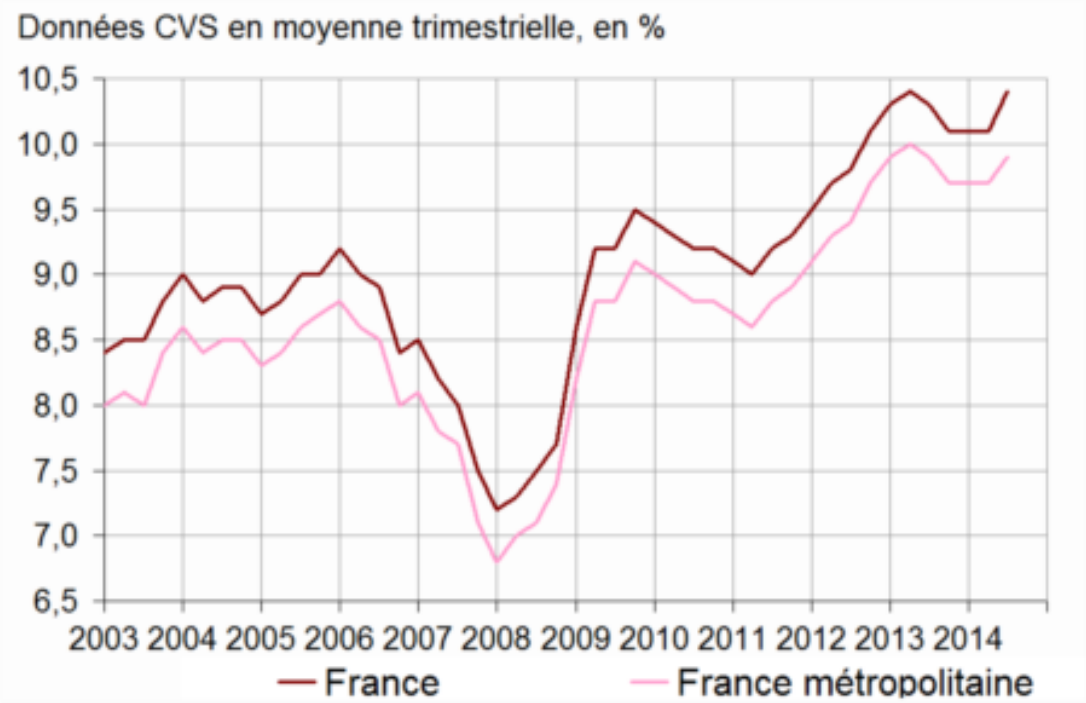
« L'enquête Emploi en continu est une enquête auprès des ménages, portant sur **toutes les personnes de 15 ans et plus vivant en France métropolitaine**. C'est une enquête trimestrielle dont la collecte a lieu en continu durant toutes les semaines de chaque trimestre. Environ 67 000 ménages ordinaires sont enquêtés chaque trimestre (c'est-à-dire les habitants de 67 000 logements, à l'exception des communautés : foyers, hôpitaux, prisons), soit autour de 108 000 personnes de 15 ans ou plus. Cet échantillon est partiellement renouvelé chaque trimestre. L'enquête en continu est prolongée par une enquête postale auprès des non-répondants, dont les résultats sont disponibles plus tardivement ».

« Au premier trimestre 2013, le questionnaire de l'enquête Emploi a été rénové, en particulier pour faciliter le déroulement de l'enquête sur le terrain grâce à des questions aux formulations plus simples. Certaines reformulations du nouveau questionnaire ont modifié la teneur des réponses d'une petite proportion de la population enquêtée. Ceci a un impact sur la mesure en niveau des principaux indicateurs. À partir de la publication de mars 2014 relative aux résultats de l'enquête Emploi au quatrième trimestre 2013, l'Informations Rapides présente les résultats observés avec le questionnaire rénové. Les séries longues publiées avec l'Informations Rapides ont été rétropolées pour les rendre cohérentes avec ce questionnaire ».

In: « Chômage au sens du BIT et indicateurs sur le marché du travail », Note Méthodologique, INSEE, Mars 2014.

1. Données de registre ou données d'enquêtes par questionnaire: un débat récurrent

- Taux de chômage au sens du BIT: évolution depuis 2003 (au 3e trimestre 2015: 10,6%, métro+DOM)



1. Données de registre ou données d'enquêtes par questionnaire: un débat récurrent

- Deux logiques différentes:

- une logique administrative directement liée aux politiques d'emploi et au fonctionnement des services de Pôle Emploi (radiations, contrôles plus ou moins stricts…) => des **effectifs** et des **taux de variation par catégories** ;

- Une logique d'enquête standardisée, visant à la comparaison internationale, et étendue à d'autres indicateurs que le taux de chômage: taux d'activité, part de l'emploi en CDI à temps complet, emploi atypique, caractéristiques des chômeurs (âge, sexe…), etc. Elle ne porte que sur un **échantillon** de la population => des taux de chômage standardisés et leur évolution.

- Les deux ont des limites évidentes: les chômeurs découragés disparaissent de la statistique dans les deux cas ; il faut aussi regarder le **taux d'activité** et le **taux d'emploi** ; regarder le marché du travail du point de vue des emplois créés et détruits (autres sources utiles: caisses de cotisations sociales).

1. Données de registre ou données d'enquêtes par questionnaire: un débat récurrent

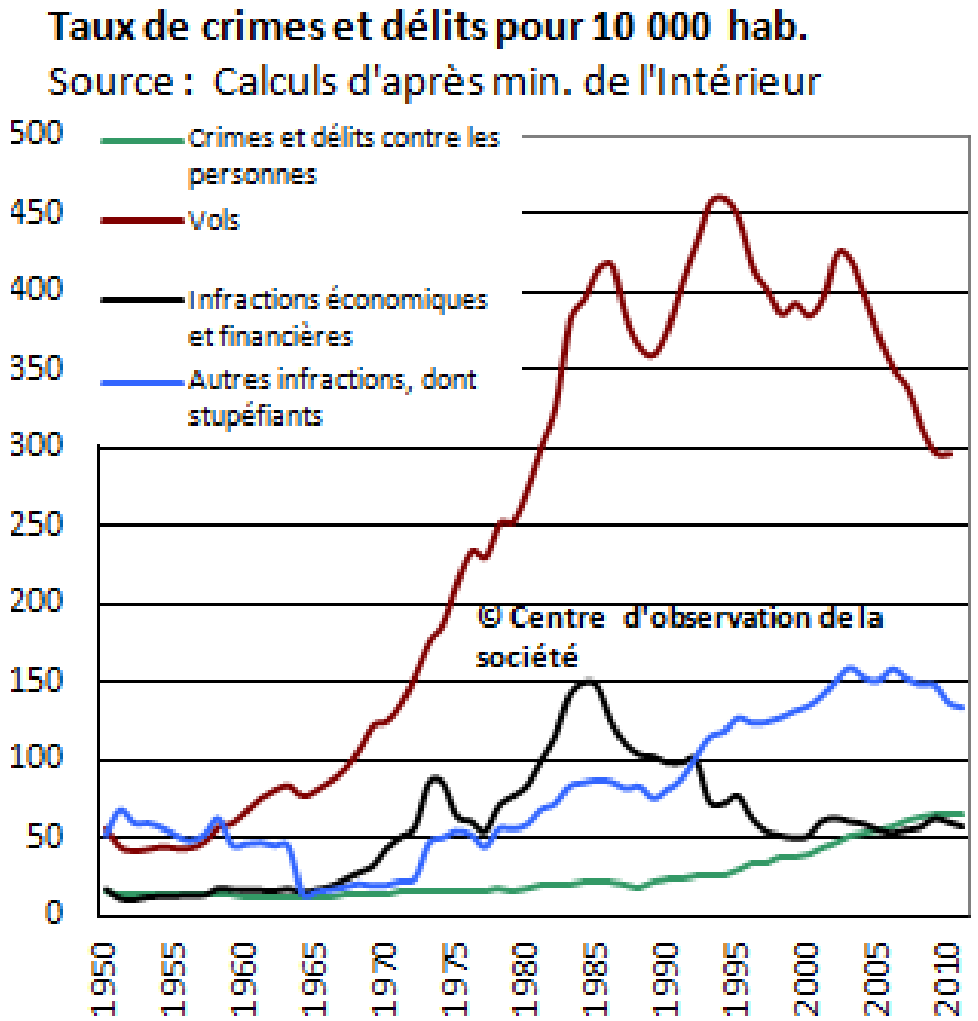
- **Exemple 2.** Les données sur la criminalité et la délinquance: un débat classique en sociologie, entre données des services de justice et de police, et données issues d'enquêtes par questionnaire (dites de « victimation »).
- Un enjeu politique central depuis que l'évolution de la délinquance et la sécurité sont constitués comme débat politique majeur (« insécurité »). De nombreuses polémiques, y compris entre sociologues, statisticiens et politiques.
- Les données policières et judiciaires sont désormais centralisées par l'Observatoire national de la délinquance et des réponses pénales (ONDRP) et publiées régulièrement (mensuellement et annuellement).

1. Données de registre ou données d'enquêtes par questionnaire: un débat récurrent

- L'Observatoire national de la délinquance et des réponses pénales (ONDRP) est un département de l'Institut national des hautes études de la sécurité et de la Justice. Il est doté d'un conseil d'orientation chargé d'assurer l'indépendance de ses travaux. Il a comme activité principale la production et la diffusion de statistiques sur la criminalité et la délinquance. L'ONDRP inscrit ses travaux dans le cadre de la statistique publique et du code des bonnes pratiques de la statistique européenne.
- L'ONDRP a notamment pour mission de recueillir les données statistiques relatives à la délinquance auprès de tous les départements ministériels et organismes publics ou privés ayant à connaître directement ou indirectement de faits ou de situations d'atteinte aux personnes ou aux biens. A ce titre, il analyse et diffuse les données sur les crimes et délits enregistrés par les services de police et les unités de la gendarmerie nationales.
- Quatre principaux indicateurs : les atteintes volontaires a l'intégrité physique, les atteintes aux biens, les infractions révélées par l'action des services, et les infractions économiques, financières et escroqueries.

1. Données de registre ou données d'enquêtes par questionnaire: un débat récurrent

- L'ONDRP prévient que « toute personne qui commente les chiffres sur la délinquance enregistrée doit, pour éviter une telle confusion sur sa nature, toujours préciser qu'il s'agit de « faits constatés » ou d'« infractions enregistrées » par la police ou la gendarmerie ».
- Depuis les années 1960, les sociologues de la délinquance (par exemple: Howard Becker) ont mis en évidence le fait que les statistiques officielles mesurent plus l'activité des services de police (et de la justice) et la propension des individus à porter plainte que le « vrai chiffre » (« chiffre noir ») de la délinquance et de la criminalité.
- Selon les délits, le taux de plainte varie fortement (16% agressions verbales, 91% agressions physiques graves (ITT>8j); selon les objectifs de la police, certains chiffres « gonflent » et d'autres restent peu visibles… Délits liés à l'immigration irrégulière.


Taux de crimes et délits pour 10 000 hab.
Source : Calculs d'après min. de l'Intérieur
Crimes et délits contre les personnes
Vols
Infractions économiques et financières
Autres infractions, dont stupéfiants
© Centre d'observation de la société
500
450
400
350
300
250
200
150
100
50
0
1950
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005
2010

1. Données de registre ou données d'enquêtes par questionnaire: un débat récurrent

- Les enquêtes « en population générale »: la délinquance et la criminalité vues du côté des victimes. Il s'agit d'enquêtes sur échantillons représentatifs de la population générale. Les questions posées portent sur l'ensemble des faits de délinquance ou de criminalité dont les personnes ont été les victimes (« victimation »).
- Les résultats font apparaître d'assez gros écarts avec les statistiques administratives, surtout pour certains faits qui sont sous-déclarés pour diverses raisons (lourdeur et coût de la procédure, mise en jeu de l'intimité et des normes sociales, assurance, etc.).
- Les tendances ne sont pas toujours les mêmes: déclin des vols et des violences aux personnes entre 1996 et 2006 selon diverses enquêtes de victimation.
- Progression des « petites agressions »: un changement de sensibilité à la violence ? Une « civilisation des mœurs » (Elias) ?

1. Données de registre ou données d'enquêtes par questionnaire: un débat récurrent

« Réalisée par l'Insee depuis 2007, l'enquête de victimation - Cadre de vie et sécurité (CVS) - vise à connaître les faits de délinquance dont les ménages et leurs membres ont pu être victimes dans les deux années précédant l'enquête. Elle porte sur les cambriolages, les vols ou dégradations de véhicules ou du logement, que ces délits aient fait ou non l'objet d'une plainte. Elle porte également sur les vols personnels, les violences physiques, les menaces ou les injures ainsi que l'opinion des personnes concernant leur cadre de vie et la sécurité.

L'enquête est menée chaque année auprès d'environ 25 500 ménages résidants en France métropolitaine.

Auparavant, l'Insee mesurait la délinquance subie par la population ainsi que le sentiment d'insécurité à l'aide des enquêtes permanentes des Conditions de vie (EPCV). » (Source: INSEE).

1. Données de registre ou données d'enquêtes par questionnaire: un débat récurrent

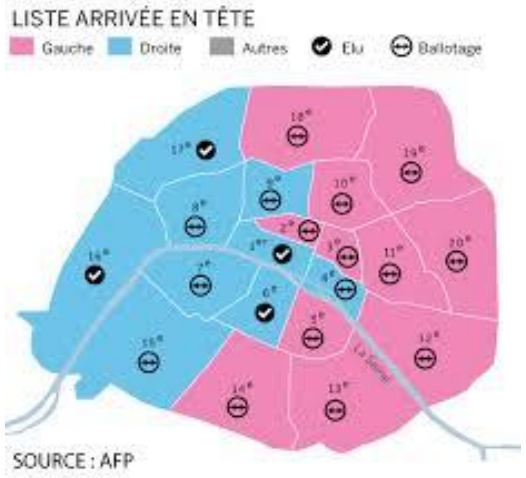
- Des constats qui éclairent des aspects complémentaires de l'évolution de la criminalité et de la délinquance. La sous-déclaration affecte en général plus la délinquance enregistrée que les enquêtes (anonymat, absence d'effet, etc.).
- Les tendances: durant la période récente, les variations sont assez faibles d'une année sur l'autre quelle que soit la source. Les vols sont beaucoup plus nombreux que les atteintes aux personnes, et ont plutôt tendance à diminuer, contrairement aux atteintes aux personnes (selon les statistiques policières) mais non seulement les enquêtes de victimation qui les font apparaître stables.
- Très peu de fiabilité des deux sources s'agissant de la délinquance économique (exemple: fraude fiscale), dont la victime est la collectivité. « Délinquance en col blanc » de plus en plus visible dans l'espace public (cf. affaire Cahuzac, etc.).

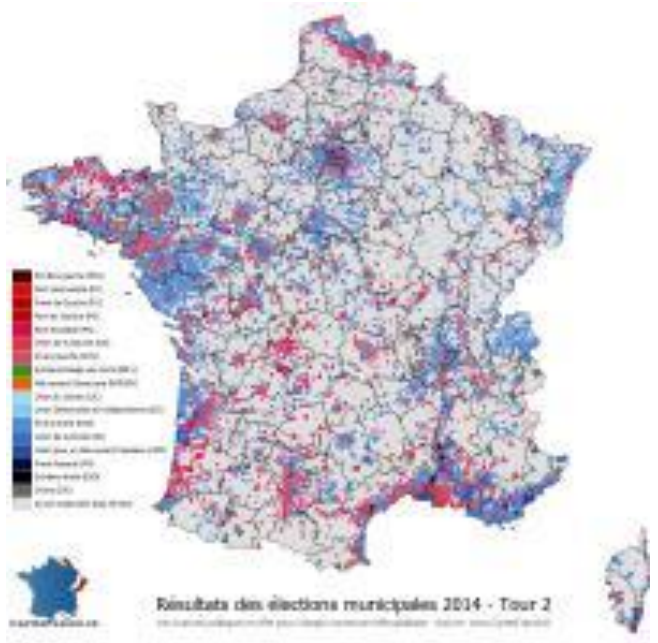
1. Données de registre ou données d'enquêtes par questionnaire: un débat récurrent

- **Exemple 3**: **les données électorales et politiques**. Un cas un peu différent. Les données de registre, ici, ne descendent pas au niveau individuel, mais à des échelles territoriales administratives (bureau de vote, commune, etc.). Ces sont les **résultats électoraux**, qui renseignent sur l'état « politique » d'une unité territoriale: abstention, rapports de force électoraux…
- On oppose ici ces données « objectives » (sauf erreurs de comptage) aux données « subjectives » (voir de Singly) issues des enquêtes d'opinion, qui sont plus délicates à interpréter mais ont l'avantage de « descendre » au niveau de l'électeur et de ses attitudes déclarées.
- Des « faits de vote » bruts enregistrés d'un côté, des déclarations d'attitude de l'autre.

1. Données de registre ou données d'enquêtes par questionnaire: un débat récurrent

- Les données issues des résultats électoraux font apparaître de forts contrastes géographiques, bien connus de la géographie électorale.





1. Données de registre ou données d'enquêtes par questionnaire: un débat récurrent

- Depuis André Siegfried et son *Tableau politique de la France de l'Ouest (1913)*, il est devenu classique de mettre en relation les caractéristiques des territoires (en particulier, sociales, économiques mais aussi culturelles, religieuses…) et les résultats électoraux. Siegfried opposait « calcaire » et « granit » (G/D), mais les variables sous-jacentes sont sociales: dépendance sociale, intensité des convictions politiques.
- Exemple: taux de chômage, proportion d'ouvriers dans la structure sociale et vote Front National (voir introduction).
- Structures sociales et démographiques apparaissent **corrélées** aux votes.
- NB: en ce cas, les territoires sont caractérisés par des enquêtes sur des données « objectives » (recensement, enquête emploi, etc.).

1. Données de registre ou données d'enquêtes par questionnaire: un débat récurrent

- A l'opposé, on dispose de données issues de **sondages d'opinion**, de toutes sortes.
- Un marché très développé, qui produit une bonne partie des chiffres circulant dans l'espace public. Exemple: positions des électeurs de l'UMP sur la législative partielle dans le Doubs.
- Des données reposant sur la technique de l'échantillonnage (aléatoire, par quotas): à partir d'un petit échantillon d'environ $n = 1000$ enquêtés, on « extrapole » à la population générale. C'est l'**inférence statistique**.
- On observe une fréquence particulière f_k dans l'échantillon (30% des ouvriers qui ont voté disent avoir voté pour le PS) et on l'infère à la population dans son ensemble (F_k).
- Notion de « marge d'erreur »: intervalle de confiance. Ensemble des valeurs du paramètre compatibles avec les données. Risque d'inférence erronée.
- Plus l'échantillon est petit, plus l'intervalle est grand.

1. Données de registre ou données d'enquêtes par questionnaire: un débat récurrent

- Les **biais** classiques dans les sondages:
- - Critique de l'idée d' « opinion publique » (P.Bourdieu): des rapports subjectifs au sujet en fait très variables ;
- - L'imposition de problématique dans les questionnaires ;
- - Des échantillons… **représentatifs** ? Le problème des refus de réponse systématiques. Idem sur Internet ;
- - Sous-déclaration de certains votes dans les sondages d'intention de vote ou « sortie des urnes » (d'où l'utilisation de **coefficients de redressement** non rendus publics) ;
- - Des sous-populations parfois très petites sur lesquelles on tire trop vite des conclusions (exemple : vote des ouvriers qualifiés, des 16-25 ans, de tel électorat…).

1. Données de registre ou données d'enquêtes par questionnaire: un débat récurrent

- A nouveau, deux sources complémentaires plus qu'opposées.
- Les données de registre ont une solidité supérieure, mais elles ne descendent pas au niveau individuel. Exemple du vote des gendarmes. « Tous les bureaux abritant une caserne de la gendarmerie mobile affichent un vote pour Marine Le Pen à la présidentielle très nettement supérieur à la moyenne de leur ville » (IFOP, 2014). Cas du bureau n°10 de Versailles, à proximité du camp de Satory. Les électeurs inscrits à ce bureau sont à 100% des gendarmes et des membres de leurs familles, sans présence du reste de la population de la ville. Marine Le Pen à 46,1%, Nicolas Sarkozy à 22,9%, François Bayrou à 11,7% et François Hollande à 11%.
- Les enquêtes d'opinion font apparaître des liaisons entre variables à un niveau plus fin, individuel, avec un plus grand nombre d'indicateurs. A condition, bien sûr, de faire l'objet de traitements approfondis prenant en compte tous les biais possibles.

1. Données de registre ou données d'enquêtes par questionnaire: un débat récurrent

- Conclusions de cette première partie:
- Les deux grandes sources d'informations sont complémentaires et ont des qualités et des défauts respectifs, qui varient selon les domaines.
- **Exhaustivité** apparente dans le cas des données administratives, mais dépendance à des définitions juridico-administratives.
- Limites de l'**échantillonnage** lorsque l'on veut travailler sur des sous-populations de taille réduite, et divers **biais** dans les réponses déclaratives issues d'un questionnement direct.
- Il faut distinguer entre des types d'enquêtes par questionnaire très divers, selon plusieurs critères: nature de l'organisme, **taille de l'échantillon**, enquête rétrospective ou plus ponctuelle, enquête longitudinale (panel), données factuelles ou subjectives (et les problèmes posés par la formulation des questions)…

1. Données de registre ou données d'enquêtes par questionnaire: un débat récurrent

- Quelques exemples d'**enquêtes par questionnaires** importantes pour les sciences sociales:
- Recensements
- Enquêtes de conjoncture
- Labour Force Survey: enquête emploi
- EU-SILC: enquête sur les conditions de vie en Europe
- World Values Survey: l'étude des valeurs et attitudes
- European Social Survey: enquête sociale européenne
- Enquête Pratiques culturelles des Français
- Enquêtes Emploi du temps
- Enquête Histoire de Vie : exemple d'enquête biographique (INSEE, 2003)
- Enquête Elfe: panel de 20000 enfants suivi sur plusieurs années.

2. Des données alternatives pour les science sociales: expériences, biographies collectives, réseaux, discours…

- Ce qui précède conduit à ce que l'on appelle l'**analyse secondaire des données**: le chercheur utilise des données préexistantes produites par d'autres et les étudie avec les méthodes statistiques appropriées.
- Une autre stratégie consiste à produire des données soi-même en mettant en place des **protocoles** d'enquêtes. Cf. Lazarsfeld et son équipe à Marienthal, et encore plus par la suite à Columbia, au sein du Bureau of Applied Social Research (*The People's Choice, 1944*).
- Plusieurs démarches: l'enquête par questionnaire (on revient en quelque sorte au 1., mais avec plus de latitude… et moins de moyens) ; la démarche expérimentale (économie expérimentale, psychologie sociale…) ; la collecte de données biographiques (« prosopographie ») ; l'étude de réseaux ; l'analyse quantitative des discours.

2. Des données alternatives pour les science sociales: expériences, biographies collectives, réseaux, discours…

- L'expérimentation en sciences sociales: une pratique de plus en plus développée et légitime.
- Le cas de **l'économie expérimentale**. « L'économie expérimentale est une méthode scientifique permettant de reconstituer en laboratoire des contextes de décision. Pour participer à une expérience, aucune qualification n'est nécessaire. La participation aux expériences est **volontaire** et permet de gagner une somme d'argent qui dépend des décisions prises ainsi que des décisions des autres participants » (site du Laboratoire d'économie expérimentale de Paris: http://leep.univ-paris1.fr/accueil.htm).
- La psychologie (sociale) expérimentale. Exemple: l'expérience de Milgram sur l'autorité.
- Les évaluations aléatoires du J-PAL (Esther Duflo).

2. Des données alternatives pour les science sociales: expériences, biographies collectives, réseaux, discours…

- La collecte des données biographiques (« prosopographie »).
- Une démarche issue de l'histoire antique et médiévale: identifier des individus (par exemple sur des monuments) et collecter des informations les concernant.
- En sociologie contemporaine: comptages à partir de données biographiques. L'exemple du *Who's who in France* ? Étudié par F.Denord, P.Lagneau-Ymonet et S.Thine (2011). Autres exemples: les patrons des plus grosses sociétés cotées ; présence des différentes catégories socio-professionnelles dans les séries télévisées

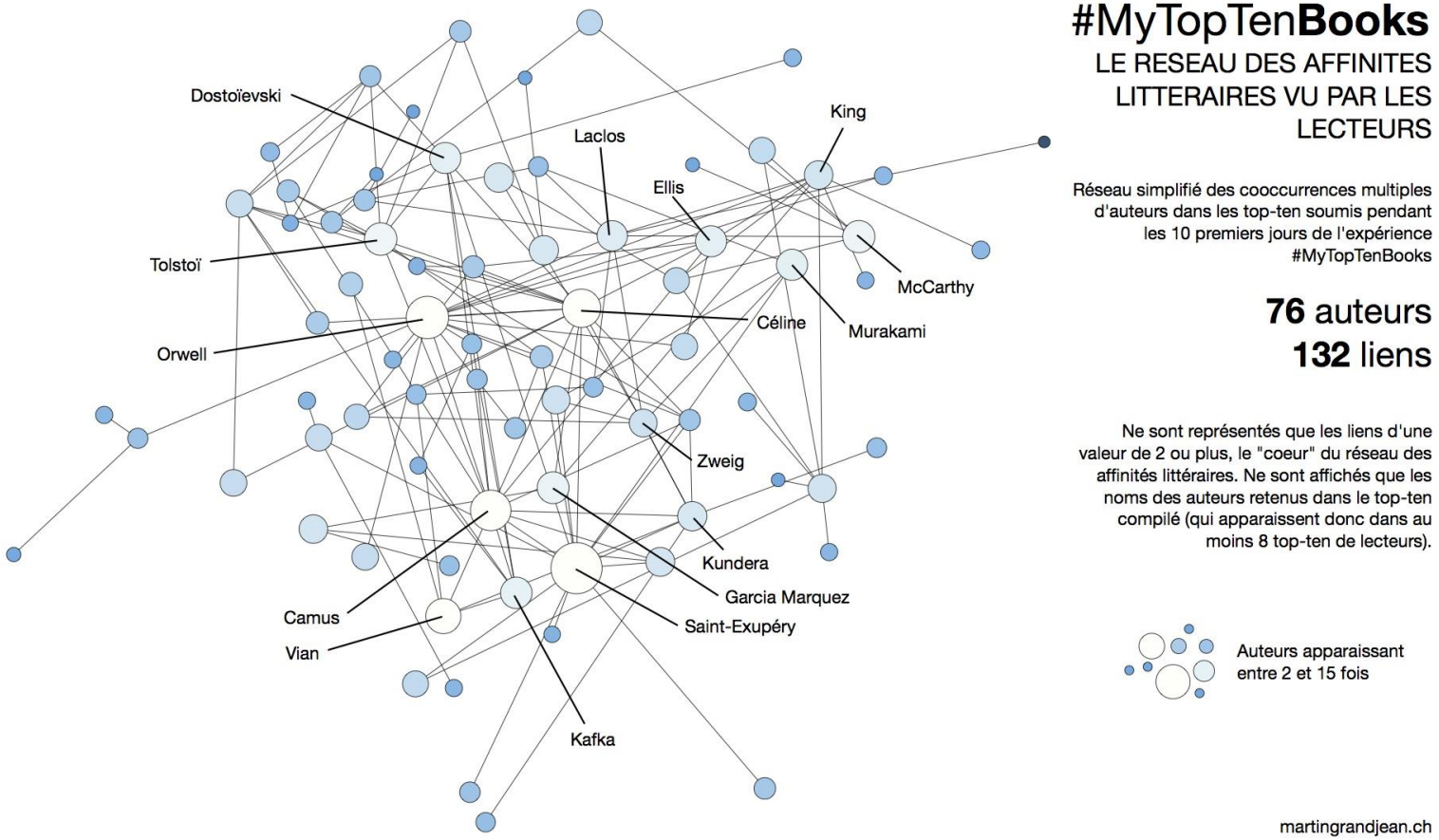
2. Des données alternatives pour les science sociales: expériences, biographies collectives, réseaux, discours…

	1969	2009
Sexe		
Hommes	98,0	86,3
Femmes	2,0	13,7
Forme de la famille		
NSP	1,8	12,0
Célibat, divorce, union libre	16,6	19,9
Marié ou veuf	81,6	68,1
Profession du père		
Agriculteur	5,8	4,2
Cadre privé	6,7	14,6
Employé, ouvrier	21,4	12,4
Haut fonctionnaire	17,2	11,8
Patron	27,9	33,6
Profession intellectuelle	8,0	11,1
Profession libérale	13,1	12,3
Scolarité		
Pas d'études supérieures	17,0	10,0
Études supérieures	83,0	90,0
Secteurs d'activités		
Art	17,0	11,2
Justice	5,0	4,5
Administration	13,0	16,0
Enseignement	10,0	8,4
Santé	6,0	3,5
Finances	5,0	10,7
Industrie	34,0	23,2
Commerce	5,0	15,2
Armée, agriculture, association	5,0	7,2

2. Des données alternatives pour les science sociales: expériences, biographies collectives, réseaux, discours…

- Les données de **réseaux**: on s'intéresse aux liens entre individus ou entités. Exemple : participation à un même conseil d'administration.
- L'objectif général est de décrire et visualiser la structure des liens entre les individus étudiés.
- Pour cela: construire un tableau des liens entre les individus étudiés et dessiner le graphe des relations, qui permet de repérer des individus plus « centraux », d'autres isolés, d'autres périphériques, etc.

2. Des données alternatives pour les science sociales: expériences, biographies collectives, réseaux, discours…



2. Des données alternatives pour les science sociales: expériences, biographies collectives, réseaux, discours…

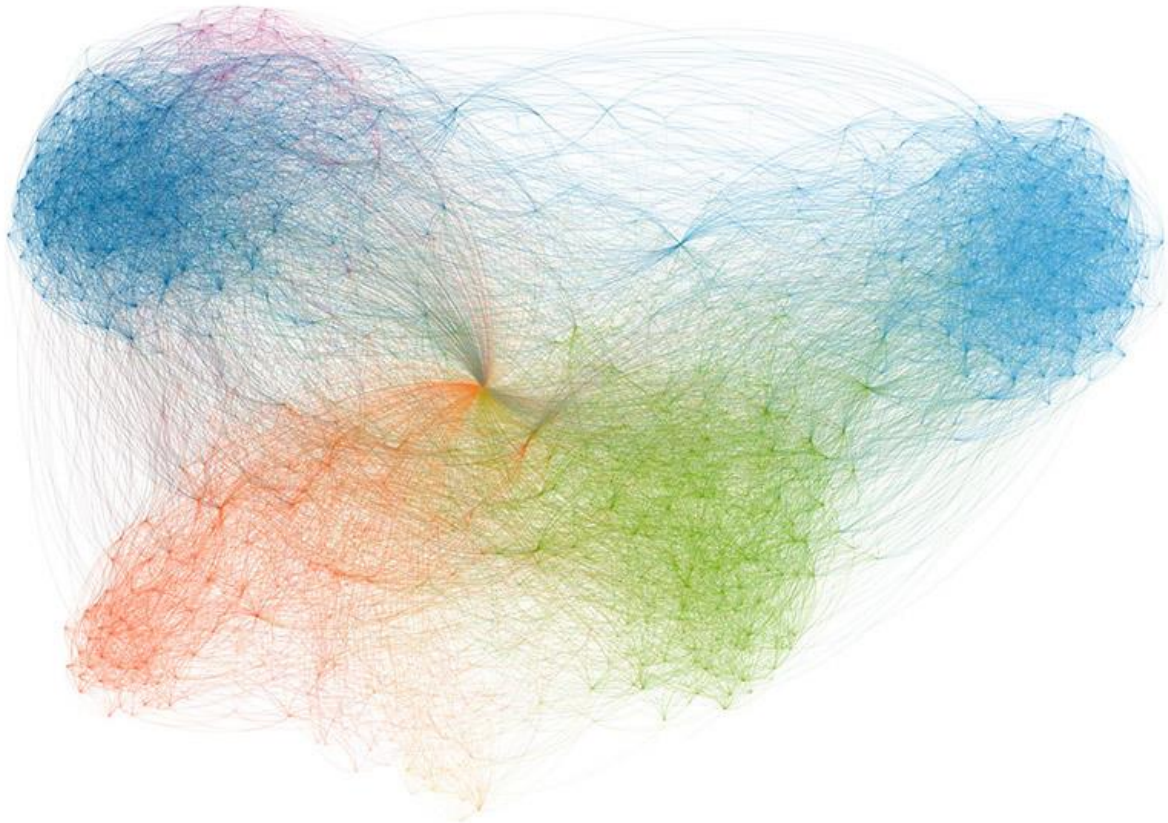
- Les données langagières: une source quasi-infinie de données de plus en plus facilement accessibles.
- Il s'agit de compter les mots utilisés. D'où des tableaux de fréquences des mots les plus utilisés. Mots-outils / mots-pleins.
- Permet de caractériser les stratégies discursives et argumentatives.
- Exemple: les discours de Nicolas Sarkozy (exemple: discours « de rentrée » du 25/09/2014).

2. Des données alternatives pour les science sociales: expériences, biographies collectives, réseaux, discours…

3. Les nouvelles sources de données à l'ère numérique.

- Montée en puissance des discours autour du « BigData »: profusion de données fortement hétérogènes, traitées par des algorithmes de **data mining**.
- Des données liées au temps passé par les individus sur Internet: un fort **biais de sélection**.
- Des données de consommation et d'usage, dont l'intérêt est qu'elles sont interconnectées: temps passé sur des sites, téléchargements, achats, etc.
- La production de données via les réseaux sociaux se développe également: Facebook, Twitter, LinkedIn, etc. Ex: Données langagières en particulier (exemple: #jesuischarlie).
- Développement de nouvelles techniques de « data visualisation ». Interprétation complexe pour les chercheurs.

3. Les nouvelles sources de données à l'ère numérique.



4. Où trouver les données ?

- De plus en plus de données sont directement accessibles en ligne, sur les sites institutionnels. Souvent au format Tableur. Exemple: séries chronologiques de chômage sur le site de Pôle Emploi.

http://www.pole-emploi.org/statistiques-analyses/

- Un effort notable se développe pour rendre accessibles les données publiques.
- https://www.data.gouv.fr/fr/
- L'accès aux **micro-données** est plus difficile et plus contrôlé.
- http://www.reseau-quetelet.cnrs.fr/spip/
- http://cdsp.sciences-po.fr/
- La production directe de données à partir d'Internet (Webscraping) se développe.

Conclusions

- Conclusions:
- Profusion de sources de natures assez diverses, souvent complémentaires.
- Croiser et comparer les sources est toujours nécessaire.
- Nécessité d'un regard critique sur la nature des opérations en jeu dans la **construction statistique**.
- Etre conscient du caractère construit des chiffres n'implique pas un refus de leur utilisation, au contraire…
- Mais cela conduit à redoubler de prudence dans leur interprétation…
- Et à adapter les techniques statistiques et les interprétations en conséquence.