L4. Admin administrative data

Olivier Godechot
Sciences Po – Campus Reims
Inquiries in Sociology

The state collects and produces data

- Governmentality (Foucault). Power and knowledge are not independent
- Statistics describing the State and its resources
- What's a (modern) State ? (Elias)
  - A permanent army
  - A permanent taxation system
- Permanent army and taxation system appeared simultaneously in the end of 14th Century (Elias) in France & England
- Early statistics and surveys
  - Counting people (CENSUS) ➔ military and fiscal purposes
  - Estimating income ➔ fiscal statistics

State data serves multiple purposes

- Not only scientific ones
- Collect data on which the State wants to act. Goal oriented.
- Decisions on individuals depend on their individual answer
- Matter of incentives / negotiation / manipulation
- Action purposes might bias knowledge purposes

- Administration of people also creates data
  - Recruitment, wages, etc.
- Example: fiscal data
  - Fiscal data needed in order to evaluate fiscal return of various taxes
    - Example next slide
  - Fiscal fraud. Tax evasion
  - Incentives for fiscal employee
- It can serve to evaluate inequality
  - Cf. Kuznets transformation of income by Piketty

DGFiP - Bureau des études statistiques en matière fiscale

IRCOM revenus 2015

NATIONAL

Montants en milliers d'euros (exceptés ceux des RFR par tranche)

Revenu fiscal de référence par tranche (en euros)	Nombre de foyers fiscaux	Revenu fiscal de référence des foyers fiscaux	Impôt net (total)*	Nombre de foyers fiscaux imposés	Revenu fiscal de référence des foyers fiscaux imposés	Traitements et salaires		Retraites et pensions	
						Nombre de foyers concernés	Montant	Nombre de foyers concernés	Montant
0 à 10 000	8,779,578	37,017,353	-120,471	64,840	286,206	3,847,734	25,255,051	2,495,406	21,449,278
10 001 à 12 000	2,141,456	23,577,329	-52,668	5,831	63,719	1,300,670	16,047,391	880,353	11,767,189
12 001 à 15 000	3,415,487	46,459,161	-97,920	236,299	3,491,606	2,511,732	38,311,165	986,124	14,808,065
15 001 à 20 000	5,907,523	102,764,312	1,757,683	3,105,770	53,996,313	3,993,610	75,874,898	2,131,570	39,946,616
20 001 à 30 000	6,830,792	167,947,232	5,647,709	3,927,006	96,705,054	4,715,807	119,936,929	2,607,704	61,870,041
30 001 à 50 000	6,553,656	250,560,654	13,309,362	5,134,397	199,437,842	4,657,642	178,508,066	2,595,405	83,917,149
50 001 à 100 000	3,305,940	217,391,802	20,630,441	3,121,015	205,724,469	2,517,357	155,799,009	1,131,280	51,660,986
Plus de 100 000 dont:	749,163	140,216,578	28,027,792	727,817	136,965,445	590,890	73,933,363	207,206	12,208,835
100 001 à 200 000	*597,946*	*78,515,622*	*12,812,674*	*579,100*	*76,088,011*	*471,765*	*49,173,701*	*168,217*	*9,660,755*
200 001 à 300 000	*88,183*	*21,094,944*	*4,697,851*	*86,452*	*20,685,165*	*69,118*	*10,664,544*	*22,488*	*1,365,351*
300 001 à 400 000	*28,243*	*9,670,624*	*2,422,619*	*27,854*	*9,537,660*	*22,080*	*4,371,158*	*7,231*	*470,991*
400 001 à 500 000	*12,523*	*5,567,586*	*1,464,665*	*12,362*	*5,495,321*	*9,933*	*2,383,452*	*3,135*	*213,227*
500 001 à 600 000	*6,552*	*3,572,205*	*965,370*	*6,477*	*3,531,639*	*5,180*	*1,428,637*	*1,724*	*122,400*
600 001 à 700 000	*3,889*	*2,512,860*	*685,321*	*3,845*	*2,484,572*	*3,107*	*966,438*	*1,047*	*75,907*
700 001 à 800 000	*2,456*	*1,833,359*	*511,724*	*2,433*	*1,816,159*	*1,963*	*679,324*	*665*	*52,117*
800 001 à 900 000	*1,709*	*1,445,836*	*405,219*	*1,697*	*1,435,615*	*1,404*	*537,151*	*493*	*42,179*
900 001 à 1 000 000	*1,247*	*1,181,172*	*330,407*	*n.c.*	*n.c.*	*1,007*	*422,670*	*347*	*29,551*
1 000 001 à 2 000 000	*4,463*	*6,045,095*	*1,605,428*	*4,420*	*5,981,102*	*3,670*	*1,820,253*	*1,296*	*120,239*
2 000 001 à 3 000 000	*978*	*2,349,457*	*618,766*	*967*	*2,324,111*	*816*	*601,646*	*277*	*27,882*
3 000 001 à 4 000 000	*389*	*1,339,664*	*332,044*	*389*	*1,339,664*	*337*	*284,754*	*116*	*11,466*
4 000 001 à 5 000 000	*177*	*792,644*	*192,017*	*n.c.*	*n.c.*	*150*	*137,168*	*48*	*4,976*
5 000 001 à 6 000 000	*106*	*582,664*	*145,107*	*n.c.*	*n.c.*	*94*	*100,646*	*31*	*3,161*
6 000 001 à 7 000 000	*62*	*399,427*	*108,241*	*62*	*399,427*	*53*	*54,750*	*19*	*1,672*
7 000 001 à 8 000 000	*41*	*310,129*	*84,401*	*41*	*310,129*	*38*	*51,430*	*n.c.*	*n.c.*
8 000 001 à 9 000 000	*36*	*302,012*	*65,297*	*36*	*302,012*	*31*	*47,320*	*n.c.*	*n.c.*
Plus de 9 000 000	*163*	*2,701,278*	*580,642*	*163*	*2,701,278*	*144*	*208,320*	*48*	*3,981*
Total	**37,683,595**	**985,934,421**	**69,101,927**	**16,322,975**	**696,670,654**	**24,135,442**	**683,665,874**	**13,035,048**	**297,628,161**
Dont non-résidents (DRESG)	**225,089**	**4,448,380**	**762,566**	**118,802**	**3,899,778**	**68,068**	**2,593,683**	**38,065**	**822,638**

* Impôt sur le revenu émis par voie de rôle, hors prélèvement forfaitaire obligatoire sur les revenus de capitaux mobiliers et crédit d'impôt correspondant, hors prélèvements libératoires (sur revenus de capitaux mobiliers, sur revenus des auto-entreprene

But administrative data has many advantages

Questionnaires and surveys less and less reliable

- Refusal to answer increases
- => Incentives
- => Both source of responding bias

- Recurrent (Yearly / Monthly / Decades (census))
- Exhaustive / Full population
- Compulsory
  - Authority of the State
  - Punishment of non-respondents &/or false declaration
- Enables contextual approach
  - Firms/Towns/Area

The boom in administrative data

- https://www.casd.eu/
- Classical Data:
  - DADS : Social security data (population)
  - Census : 1/4th
  - Fiscal data
  - Vote (vote by polling station)
  - Education administrative files
- Open data movement

https://www.data.gouv.fr/fr/

The case of suicide as an administrative data

- State records deaths and causes of deaths
  - Monitor the population as a resource (army, fiscal)
  - Monitor social problems (crime, suicide) on which to act
- Recording of deaths as suicide as a case of negotiation

Douglas, Jack. 1967. *The social meanings of suicide*

- No universal definition of suicide
  - Ex. Denial of medical treatment, euthanasia, madness
- Suicide is a matter of interpretation (accident or suicide : drowning, overmedication)
- People negotiate on qualifications of suicide. Some families don't want death to be recognized as suicide (Religion / Insurance)
- Administrative and medical data don't coincide

Not perfect… but not so bad

- In Durkheim (and following work using such data), gap between theoretical research design and empirical data
- Much better than data on crime… (which exists through complaints)
- At least, all deaths are counted

Baudelot, Christian and Roger Establet. 1982. *Durkheim et le suicide*. Presses Universitaires de France.

➔ Not perfect, but data of good quality

Not perfect… but not so bad (Baudelot & Establet, 1982)

- Death must be declared in 24h to a registrar (Without causes of death)
- The registrar mandates a doctor to establish a death certificate.
  - First part. For civil register
  - Second part. Sealed. For Departmental health authority. Causes of death.
  - Individual anonimized and suicide reported in database
- If death violent: necessary to have the police and the judges, eventually fire workers.
- Difficult to hide suicide.
- In case of ambiguity between accident and suicide => police investigation. Eventual autopsy
- Doctors don't protect more females' honors by over-declaring drowning and overmedication as accidents rather than suicide
- Correcting for known mismatches don't change much results

Following Durkheim in the use of administrative Data

- Administrative data advantage and limits
  - Systematic collection
  - Exhaustive
  - (Relatively reliable)
  - But nor organized along scientific research questions
- Solution: ➔inventive use of administrative data
- Find indirect proxies of key variables

Proxy as an identifying solution

- Introduction. Hypothesis.

Suicide is a social phenomenon

- Table 1. Suicide is stable
  - stability = object of science
- Table 2. On the short term, suicide is more stable than death
  - stability is not a "natural" one ➔ social

Table I Stability of suicide in the principal European countries (absolute figures)

Years	France	Prussia	England	Saxony	Bavaria	Denmark
1841	2,814	1,630		290		377
1842	2,866	1,598		318		317
1843	3,020	1,720		420		301
1844	2,973	1,575		335	244	285
1845	3,082	1,700		338	250	290
1846	3,102	1,707		373	220	376
1847	(3,647)	(1,852)		377	217	345
1848	(3,301)	(1,649)		398	215	(305)
1849	3,583	(1,527)		(328)	(189)	337

- Table 3. Suicide is increasing on the long term
  - => suicide ➔ social issues of the 19th century

Table II Comparative variations of the rate of mortality by suicide and the rate of general mortality

A. ABSOLUTE FIGURES

Period 1841–46	*Suicides per 100,000 inhabitants*	*Deaths per 1,00 inhabitants*	*Period 1849–55*	*Suicides per 100,000 inhabitants*	*Deaths per 1,000 inhabitants*	*Period 1856–60*	*Suicides per 100,000 inhabitants*	*Deaths per 1,000 inhabitants*
1841	8.2	23.2	1849	10.0	27.3	1856	11.6	23.1
1842	8.3	24.0	1850	10.1	21.4	1857	10.9	23.7
1843	8.7	23.1	1851	10.0	22.3	1858	10.7	24.1
1844	8.5	22.1	1852	10.5	22.5	1859	11.1	26.8
1845	8.8	21.2	1853	9.4	22.0	1860	11.9	21.4
1846	8.7	23.2	1854	10.2	27.4			
			1855	10.5	25.9			
Averages	8.5	22.8	Averages	10.1	24.1	Averages	11.2	23.8

B. ANNUAL RATE RELATED TO THE AVERAGE IN PERCENTAGE FORM

1841	96	101.7	1849	98.9	113.2	1856	103.5	97
1842	97	105.2	1850	100	88.7	1857	97.3	99.3
1843	102	101.3	1851	98.9	92.5	1858	95.5	101.2
1844	100	96.9	1852	103.8	93.3	1859	99.1	112.6
1845	103.5	92.9	1853	93	91.2	1860	106.0	89.9
1846	102.3	101.7	1854	100.9	113.6			
			1855	103	107.4			

Table III Rate of suicides per million inhabitants in the different European countries

	Period			*Numerical position in the*		
	1866–70	*1871–75*	*1874–78*	*1 period*	*2 period*	*3 period*
Italy	30	35	38	1	1	1
Belgium	66	69	78	2	3	4
England	67	66	69	3	2	2
Norway	76	73	71	4	4	3
Austria	78	94	130	5	7	7
Sweden	85	81	91	6	5	5
Bavaria	90	91	100	7	6	6
France	135	150	160	8	9	9
Prussia	142	134	152	9	8	8
Denmark	277	258	255	10	10	10
Saxony	293	267	334	11	11	11

Egoistic suicide from hypothesis to test

- Hypothesis. Egoistic suicide => "lack of integration of the individual to society".

  H1: Suicide varies inversely with the degree of integration of society

- Proxy for integration. 3 subdomains
  - Religious society / Domestic society / Political society
- Religious society
  - Problem: No individual data neither on the degree of religious practice nor on the religion
  - Solution: Comparing suicide between regions where religious domination varies.
- Other domains
  - Domestic society:
    - Suicide decreases with marriage
  - Political society
    - Suicide decreases with 1848 revolutions in Europe, with war declarations, hot electoral moments

A scientific method (inspired by Stinchcomb)

- H1 implies A1, A2, A3 (correlation between religion and suicide)
  - H1 implies A1 (Correlation between religion (Protestants vs Catholics) and suicide in Germany)
    - H1 Credible
  - H1 implies also A2 (Jews versus Christians) and A3
    - H1 more credible
- H1 implies B1, B2, B3 (correlation between family and suicide in different location)
    - H1 even more credible
- Alternative hypothesis (HA mental health) does not imply A1, A2, A3
- H1 offers a broader explanation than HA. H1 is the most convincing

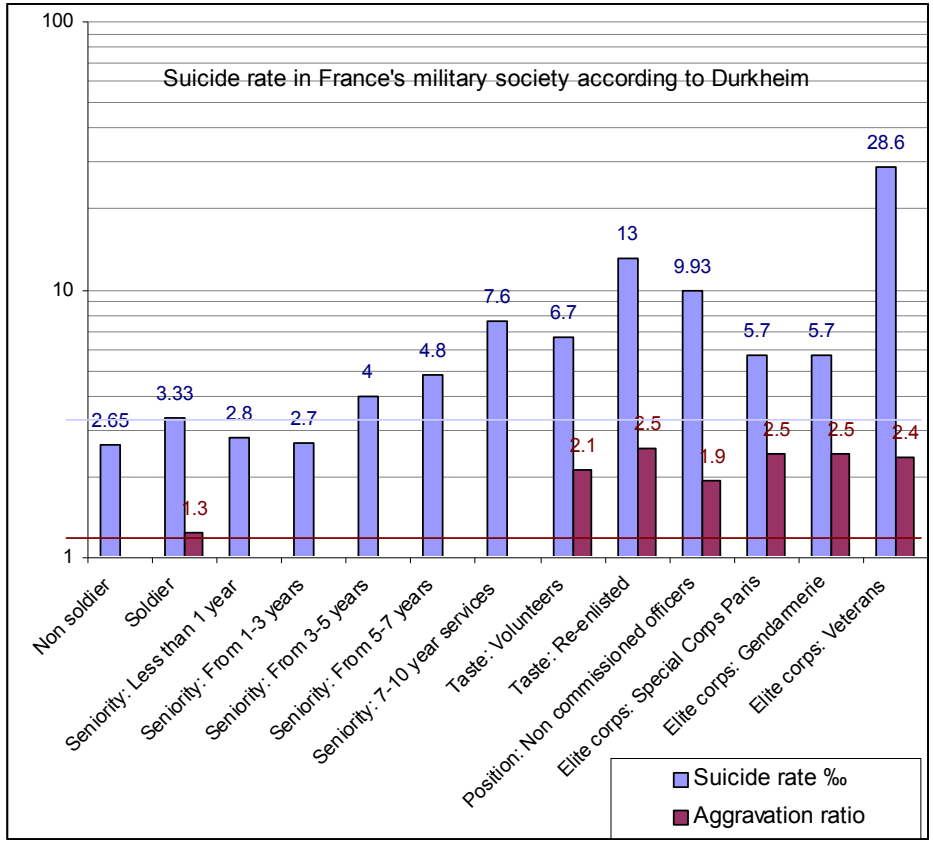
Prussian provinces (1883–90)

Provinces with more than 90% Protestant	*Suicides per million inhabitants*	*Provinces with from 89 to 68% Protestant*	*Suicides per million inhabitants*
Saxony	309.4	Hanover	212.3
Schleswig	312.9	Hesse	200.3
Pomerania	171.5	Brandenburg and Berlin	296.3
		E. Prussia	171.3
Average	264.6	Average	220.0

Provinces with from 40 to 50% Protestant	*Suicides per million inhabitants*	*Provinces with from 32 to 28% Protestant*	*Suicides per million inhabitants*
W. Prussia	123.9	Posen	96.4
Silesia	260.2	Rhineland	100.3
Westphalia	107.5	Hohenzollern	90.1
Average	163.6	Average	95.6

Altruistic suicide

- Suicide due to society's normative pressure
  - Example: Hara-Kiri. Difficult to capture in contemporary society
- Solution difference in suicide between soldiers and rest of society
- Method for proving
  - Common knowledge and examples
    - Military society => higher normative pressure
  - Established facts
    - Military society => higher suicide
    - Higher integration (seniority, hierarchical position, voluntary) in military society ➔ higher suicide



  - Suggested plausible (not totally proven) conclusion
    - higher normative pressure➔ higher suicide

Why exhaustiveness can be critical?

- The potentiality of surveys
  - Sample
  - Central limit theorem
  - No need for exhaustive data
  - With a survey on 10,000, good measure of individual measures
    - Height
    - Vote
    - Individual behavior
- But survey measures
  - Poor reconstruction of the context of individual data
  - Poor reconstruction of the network
- Exhaustive Administrative data brings
  - Precision
  - Enable contextual effects
  - Full network

An example: study academic inbreeding

- Godechot, Louvet 2010, « Academic Inbreeding: An Evaluation », *La vie des idées*
- A classical phenomenon in France (and other countries) : Academic inbreeding
  - Preferential recruitment of university's former PhD students
  - Favoritism ?

- Classical administrative measures
  - Proportion of inbred scholars
  - Bias: we don't know the competitors
- Solution: administrative database on PhD defended
  - Name of the PhD student
  - Year of defense
  - Discipline
  - University
  - Name of the PhD Supervisor
- How?

Academic inbreeding proxy

- Proxy for recruitment
  - PhD student becomes PhD Supervisor
- Proxy for inbred/outbred recruitment
  - PhD supervisor supervises its first PhD in the same/different university than its own PhD
- Proxy for competition
  - PhDs defended the same year in the same discipline

Table 6. Academic inbreeding in the recruitment of doctoral graduates who defended their thesis between 1972 and 1996, as observed in four-year periods

Period of thesis defence	Inbred candidates recruited	'Expected' inbred candidates recruited	Inbred applications	External candidates recruited	External applications	Proportion of inbred candidates out of those recruited	Mantel-Haenszel odds ratio
1972-1976	944	191.04	11,692	854	177,410	52.50%	12.49
1977-1981	1525	233.95	19,243	1254	376,652	54.88%	17.50
1982-1986	914	111.99	14,202	713	301,682	56.18%	20.83
1987-1991	945	94.95	18,226	776	479,248	54.91%	24.49
1992-1996	221	21.65	8894	183	211,194	54.70%	24.49

A plea for exhaustiveness

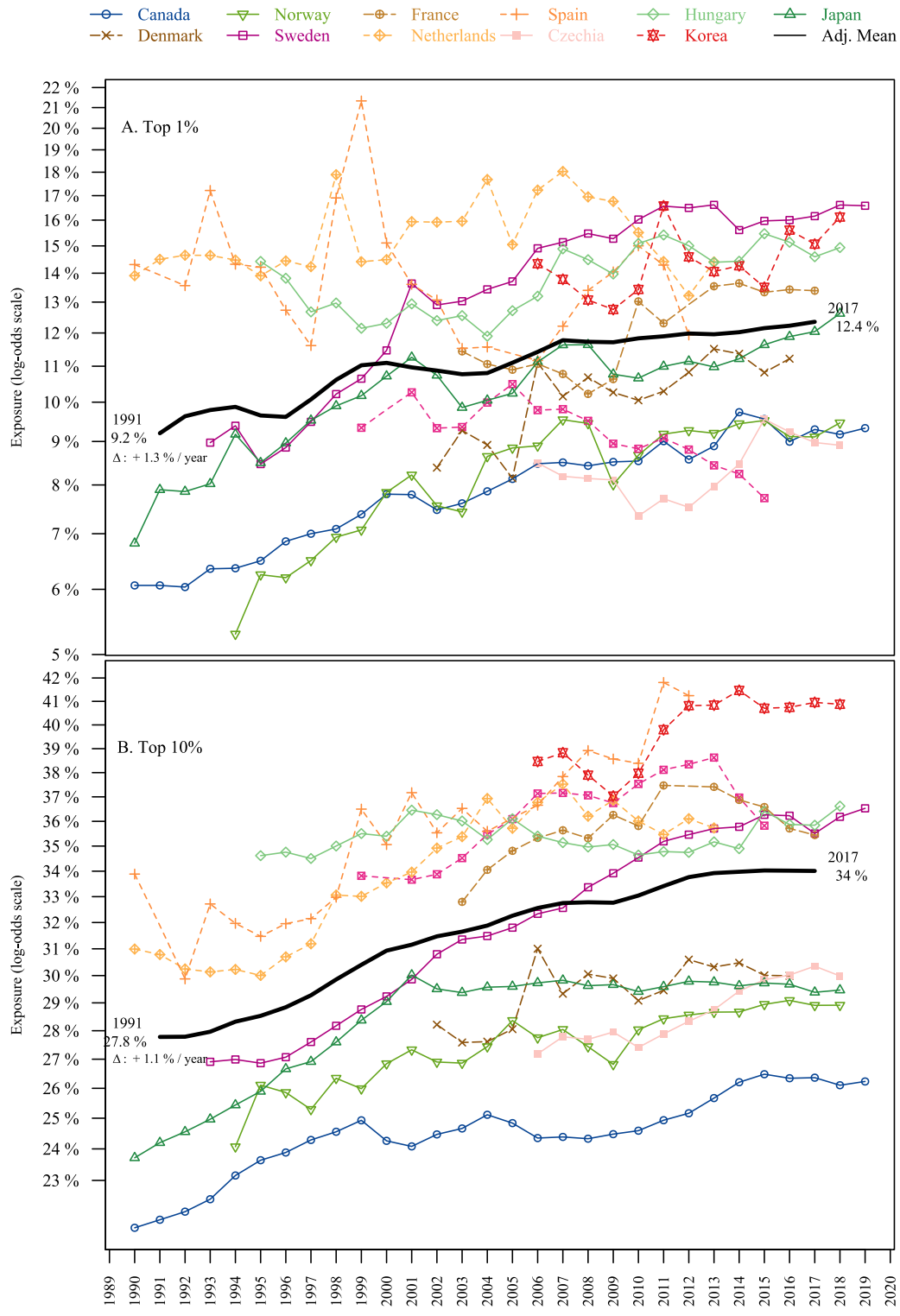
- Exhaustiveness brings
  - Precision
  - And enable contextual effects
  - Full network

Godechot, Louvet, 2010. "Comment les docteurs…"

  - Competition between PhD students of a same supervisor
    - "Eldest" vs "Youngest"
  - Difficult/Impossible to measure properly with survey data

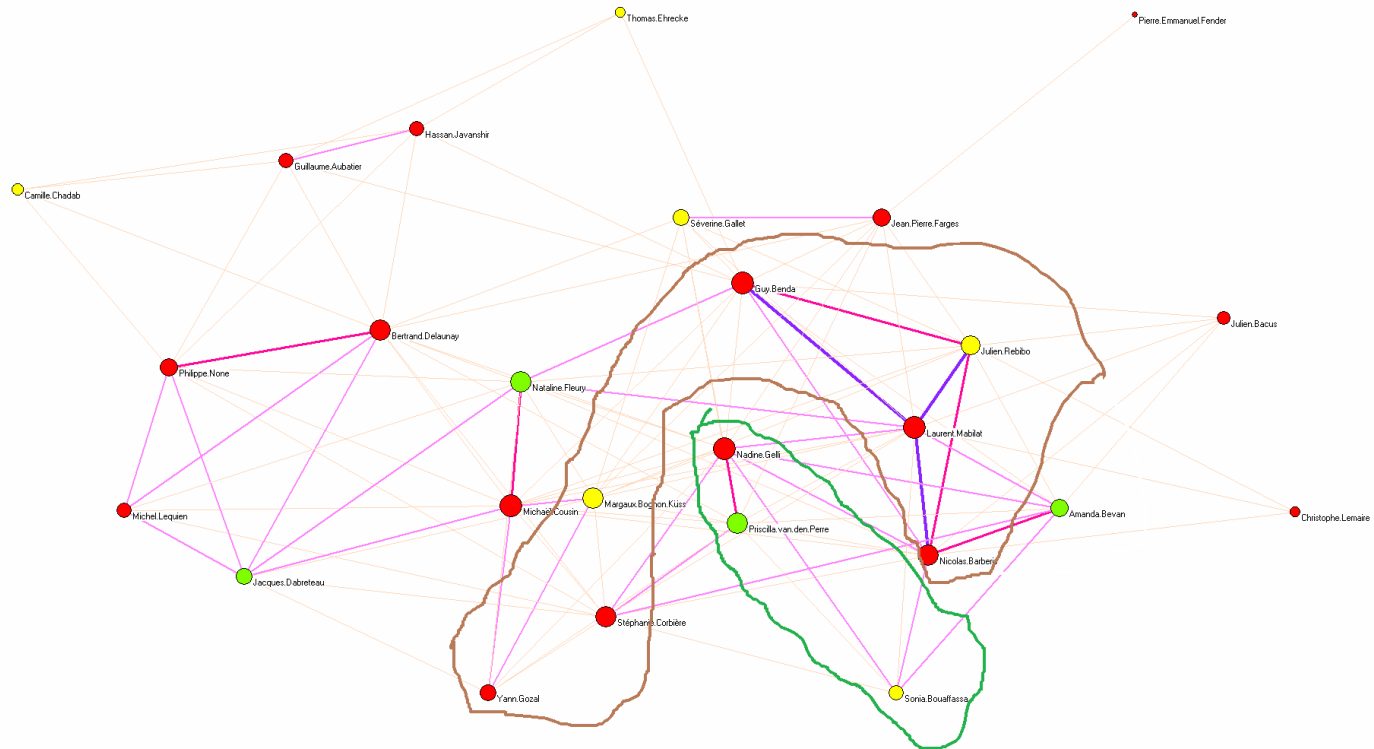
Godechot et. al. , 2020. *The great separation.*

  - Local exposure in each establishments to top1% workers
  - Unstable with survey data


Canada
Denmark
Norway
Sweden
France
Netherlands
Spain
Czechia
Hungary
Korea
Japan
Adj. Mean
A. Top 1%
Exposure (log-odds scale)
22 %
21 %
20 %
19 %
18 %
17 %
16 %
15 %
14 %
13 %
12 %
11 %
10 %
9 %
8 %
7 %
6 %
5 %
2017
12.4 %
1991
9.2 %
Δ : + 1.3 % / year
B. Top 10%
Exposure (log-odds scale)
42 %
41 %
40 %
39 %
38 %
37 %
36 %
35 %
34 %
33 %
32 %
31 %
30 %
29 %
28 %
27 %
26 %
25 %
24 %
23 %
2017
34 %
1991
27.8 %
Δ : + 1.1 % / year
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020

A plea for exhaustiveness

- Collective moves in Finance and Law Firms
- Well collected only if we have information on all members



The limits of administrative data

- Data serves as a policy tool
  - Internalization of the consequences of the policy
  - Respondents' fraud
  - Manipulation
- Respondents' fraud. Example: tax data
  - Tax avoidance / evasion / under-declaration

Alstadsæter, A., N. Johannesen, and G. Zucman. 2019. "Tax evasion and inequality." *American Economic Review* 109 (6): 2073-2103.

  - Norway, Sweden, Denmark. Exploitation of HSBC & Panama leaks
    - Top 0.01% ➔ 25% of wealth evaded.
  - Reestimation of inequality.
    - Top 0.01% : 5 to 6%.
    - Top 0.1%: 8 to 10%

The limits of administrative data

- State manipulation
  - USSR statistics on production
- Crime statistics
  - Dependent on electoral cycles
  - Tough on crime. Increase police activity/investigation
  - Success on crime. Decrease police activity/investigation
  - Evolution of crime statistics => politicized police activity more than crime evolution
    - But depends on the type of crime:
    - drugs/robbery/sex violence versus homicides
  - Need to complement with victimization surveys

Administrative data in short

- Advantage
  - Compulsory (little problem of survey refusal)
  - Exhaustive
    - Precision
    - Contextual effects / Network type of data
- Limits
  - Policy implication of the data might distort the quality

Administrative data in perspective

- Quanti/quali Classical division
  - Quanti: surveys, administrative data, experiments
  - Quali: observation, interviews
  - Mixed: archives
- Other opposition in terms of researcher's intervention
  - Strong: data produced at the initiative of the researching according to her research design
    - interviews, questionnaire, experiments
  - Minimal: data collected for other purposes and collected ex-post by the researcher for her own research design
    - observation, archives, administrative data