Linking Survey Data with Administrative Record : The French Experience

Anne Laferrère
INSEE

SHARE Linkage workshop

22-23 November 2012, Berlin

Why link? a triple motivation

- **Financial**: save on expensive surveys by using existing administrative data (more and more are available).
- **Scientific:** low or declining response rates in surveys. Data quality better in administrative data (memory, interpretation, avoidance issues…); more data.
- **Ethical:** Reduce survey burden for individuals, households or firms by using what they have previously provided to various administrations.
- … but also ethical and scientific issues!

Outline

1. The law and what is done in France for SHARE
2. The “hashing” of identifier technique
3. Health care data
4. Income data
5. What could be done on pensions rights
6 Conclusion

1. The law and what is done in France for SHARE

Law of 1951 on obligation, coordination and secret in matters of statistics.

- Created *INSEE, CNIS* (National Council for Statistical Information) to define statistical programs.
- Modified in 2008 to create a *Comité du secret statistique*, a committee of Statistical privacy.
- And in 2004, to require the transmission of administrative data to official statistical agencies.
- Confidentiality and ethical issues handled by an independent agency, the *Commission Nationale de l'Informatique et des Libertés* (CNIL), created in 1978.

1. The law and what is done in France for SHARE

- **Law of 1951 on obligation, coordination and secret in matters of statistics**.
- Created *INSEE, CNIS* (National Council for Statistical Information)
- Modified in 2008 to create a *Comité du secret statistique*, a committee of Statistical privacy.
- And in 2004, to require the transmission of administrative data to official statistical agencies.
- Confidentiality and ethical issues handled by an independent agency, the *Commission Nationale de l'Informatique et des Libertés* (CNIL).
- Linkage with administrative records can be done. If the matching requires the use of the national identification number (NIR), then authorization has to be given by a decree at the *Conseil d'Etat* level.

The law and what is done in France for SHARE

- Each wave has to be approved by the CNIS.
- The detailed questionnaire and procedure are presented to a **label Committee** to get a visa of "general public interest". Burdensome but helpful in dealing with households and getting their consent.
- The CNIL has to approve data privacy protocol (sensitive questions, "identifying" data, linkage).
- Finally the *Comité du secret* deals with the procedures of the SHARE sample transmission from INSEE to a private agency.

The law and what is done in France for SHARE

- **Data linkage** : authorized by CNIL for death register; and income tax returns data. No income linkage was completed.
- **The respondent's consent is implicit**. As for all INSEE surveys a flyer is given to each household:
- "The use by INSEE of the individual data will conform to the law of June 7 1951 on obligation, coordination and privacy in matters of statistics. The dispositions of its article 6 forbid all use aimed at tax control. The data of this survey could be linked to administrative information to which INSEE has access, in the same conditions of security and same guarantee of confidentiality."

The law and what is done in France for SHARE:

Ascertaining the death of the SHARE respondents

- National death register (RNIPP, *répertoire national des personnes physiques*)
- Need the individual *Etat-civil* (first name, birth name, date of birth, birthplace).
- Advice has to be asked to the CNIL.
- If only demographics are used and if the consultation is “manual”, an advice of the CNIL (art.27.II.1) and a ministry order (art.27. II.1) are sufficient (a decree of the *Conseil d’Etat* would be necessary if the NIR was used).

SHARE Survey of Health, Ageing and Retirement in Europe 50+ in Europe

The law and what is done in France for SHARE

Ascertaining the death of the SHARE respondents

INSEE

- **Wave 2**: interviewers asked to write birth name on paper, few did it (not in the CAPI).
- **Wave 3** (Sharelife): municipality of birth asked using a remark, after the question on home at birth, and birth name was to be asked in the CV (SMS), because the interviewers' computers and the transmission of the data were securitized.
- **Wave 4**: used SMS for birth name, and the « German » question on East Germany for municipality, but not very convenient.

✓ Municipality could only be asked to refresher respondents.

✓ Better to insert the questions within the CAPI (important that they are asked in a natural way and only when the information is not already known → preload).

✓ *A precise description of how privacy and security rules are met by SHARE would be useful in discussions with the privacy authorities.*

- The linkage is not very successful.

2. NIR and « hashing »

- **NIR : *Numéro d'Identification au Répertoire national des personnes physiques (*social security number).**
- The NIR is unique and made from the concatenation of sex/ the last two digits of birth year/ month of birth /code of municipality of birth /rank in the birth of that day in the municipality).

Ex : 1 47 08 75 014 081

- It can be partially reconstructed from demographic data, and from a NIR one could identify a person.
- → "hashing" technique
- The NIR is "hashed", then can be linked to as many data files as needed (coding system and keys being securely stored), providing the NIR was used in those files. Then the hashed NIR is scrambled (scrambling, contrary to hashing, reversible) for usage by scientists.
- For scientific use "hashing" would turn the NIR into a non identifying id number

2. NIR and « hashing »

Two sorts of linkage:

1. The linkage of administrative files without any direct face to face interview.

Ex: work on PACS (registered partnership) and same sex couples.

Cohorts (eliminate attrition)

Ex. : linkage between unemployment register and employers social declaration. A double blind procedure + restricted securitized access centers for researchers .

2. The linkage of survey data with administrative files to enrich the survey data.

Ex: SRCV (the French SILC) where households get less questions on their income when they allow INSEE to access their tax returns.

2. NIR and « hashing »

French Administrative data: some examples

EPAS *échantillon permanent des assurés sociaux*. A representative *sample* of workers. Not exhaustive, hence cannot be linked to a survey like SHARE.

RNIAM : *assurance maladie*.

SNIIR-AM : *système national inter-régimes de l'assurance maladie* ; All reimbursements by the various health insurance regimes made to an affiliated person (an insured, *ayant-droit*).

PMSI : *programme de médicalisation du système d'information* ; diagnostic of long term affections and information on hospitalizations.

SNIIR-AM and PMSI can be linked by an anonymous alphanumeric identifier built by "hashing", twice non reversible encrypting of the NIR.

3. Ex. 1: Linkage to health care data

Montaut et al. (2012) .

Linkage of the "Health and Disability Survey" to SNIIR-AM, the *système national inter-régimes de l'assurance maladie*.

"ordinary households" + individuals drawn in institutions.

The SNIIR-AM data : all reimbursements by health insurance (+ death registers) .

- Reduced survey time by 25 mn,
- allowed getting more detailed data on care;
- reduced memory bias, interpretation problems of the question (ex: what is a specialist), and desirability bias (60% of women aged 20-39 declare they saw a dentist, 40% did so).
- **But** SNIIR-AM includes reimbursement, not consumption.
- Populations covered by special health care regimes are excluded.
- The reference periods of survey and admin data do not coincide.

3. Ex. 1: Linkage to health care data

The linkage involved lots of paperwork , raised technical issues.

CNIL to authorize individual level linkage. *Décret en Conseil d'Etat* to use the NIR.

Contract between INSEE, CNAM-TS and the CNIO (acting as a *tiers de confiance*/third party).

Information on those who get their health insurance through a spouse or parent also to be collected.

The NIR was finally reconstructed from demographic information collected separately in various parts of the survey because the décret by the Conseil d'Etat arrived too late.

Check for errors by looking at *Répertoire national des personnes physiques* .

Reconstructed encrypted NIR are sent to CNAM-TS who added health care consumption and replaced the NIR by the survey ids. Sent info to CNIO who links to survey data. Administrative health data finally recoded and synthesized to be used by researchers.

22,400 reconstructed NIR, data on health care linked for 19,250.

1,450 are considered to have had no consumption; hence 20,700 persons have linked data, a success rate of 70% of the initial 29,000 initial interviews.

3. Ex. 1: Linkage to health care data

The linkage involved lots of paperwork , raised technical issues.

Were the non-linked individuals real non consumers or “non responses”?

A logistic model of non consumption was estimated from another survey (ESPS, 2006). Among the 1,450 non consumers mentioned above 800 were predicted non consumers.

Non response is not random: women, young, inactive persons are more likely not to be linked. They are more often indirectly insured *ayant-droit* (insured through a relative) than directly insured (*ouvrant-droit*, insured for themselves).

3. Ex. 1: Linkage to health care data

Lessons :

1. Permission to use the NIR hard to get. Allow at least one year.
2. Anticipate difficulties and ask for demographic information, incl. birth name of women (used in checking).
3. Do not overrate the reluctance of respondents to provide the NIR. Once the aim is explained, most cooperate and even would rather give their NIR than the data used to build it.
4. Limit the number of actors. Privacy issues lead to multiple partners with risk of having "no pilot in the plane". Hashing should allow to get rid of a third party.
5. One person should be allowed access to the data, at least for a limited time, to do some checks.
6. Describe in detail the data roadmap
7. Plan ahead a manual *reprise* for non linked NIR.

"Linkage is possible but is a long term, costly enterprise, whose stakes, costs and advantages have to be carefully weighted beforehand".

4. Ex. 2 : Linkage to income data for SILC

- Linkage of SRCV (SILC) to fiscal income data: more complex because it does not use the NIR, nor the family name.
- Routinely conducted by INSEE under strict confidentiality rules, it is easier than the health care data experiment (Burricand, 2012).
- Demographic data and address are combined in rounds of identification, progressively relaxing the matching constraints.

4. Ex. 2 : Linkage to income data for SILC

- *Enquêtes Revenus Fiscaux* (Taxed income surveys) in 1956, 62, 65, 70, 75, 79, 84 and 90: the Tax Administration completed a questionnaire for a sample of taxpayers that was then linked to census data by INSEE.
- Since 1996, the Labour Force Surveys is linked to tax files.
- **Two tax files are used**, the local residence tax files (*taxe d'habitation*) for all occupied dwellings and the income return tax files.
- Since 2005, the annual "Taxed income survey" also linked to social benefits data.
- **EU-SILC (Statistics on Income and Living conditions)** survey was started to deliver richer annual cross-sectional and longitudinal data.

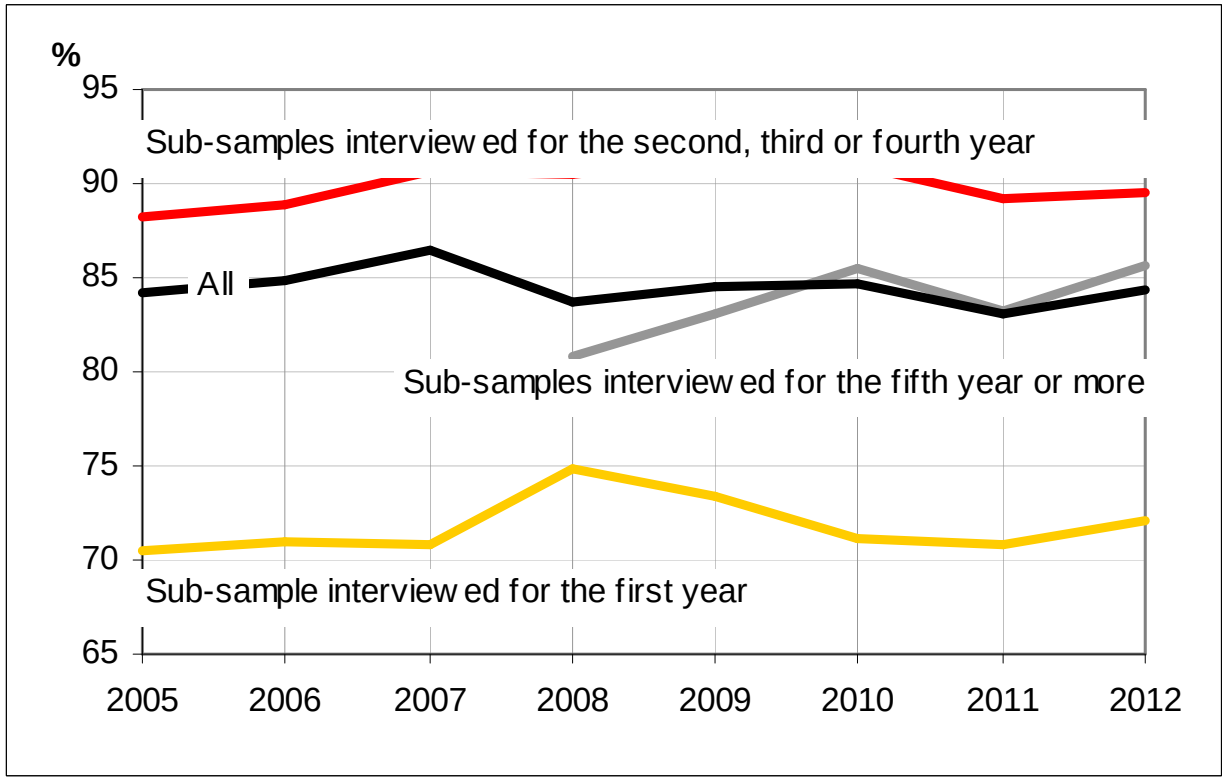
4. Ex. 2 : Linkage to income data for SILC

- 1st SILC survey in 2004: face-to-face interview.
- Getting legal authorization for linkage with income tax files and obligation **to inform** the respondents delayed the linkage to 2008.
- Principle of "fair data collection": a consequence of the 1978 "Informatique et libertés" law.
- **Respondents are notified of the linkage** in the survey presentation brochure or in the advance letter.
- They are also informed during the interview, as follows:

"A goal of this survey is to measure your income. For this purpose, the survey data will be complemented, with all guarantees of confidentiality, with administrative data. The questionnaire is therefore limited to some income components (alimony, veteran pensions...) and does not cover those that INSEE can collect through other sources (such as wages for example)".

4. Ex. 2 : Linkage to income data for SILC

Fig. 1 : Evolution of the household response rate in SILC



Source : SILC, Insee, *France (Burricand, 2012)*

NB. Ratio of the number of household interviews to the number of eligible households.

4. Ex. 2 : Linkage to income data for SILC

- A unified income concept among the population,
- an exhaustive source of data (everyone declares, even if not taxable).
- a simplified questionnaire.
- The questionnaire has to be adapted to the individual situation and on the type of income (the CAPI has to be fine-tuned). The list of various types of income is asked to all (to be able to impute income for those who cannot be matched), but the amount of each type of income is no more asked. It reduced interview duration by **10 minutes** on average.
- Tax data do not cover all income components.
- Some (e.g. the young aged between 18 and 25 who obey specific rules) are impossible to match.

4. Ex. 2 : Linkage to income data for SILC

- The process uses matching keys, a set of common variables in the two data sources (first name, gender, address, date and department of birth).
- Difference between a tax unit and a household. Thanks to the local residence tax file a household with two or more income tax returns but who pays only one residence tax can be identified.
- Quality of the variables used for matching: errors in foreign names or date of birth. Using several key variables and an iterative process reduces the number of failures. Moreover after the 1st linkage experiment Step 0 uses the last year of survey identifier, which improves efficiency.

4. Ex. 2 : Linkage to income data for SILC

Description of the record linkage process :

- **Step 0** Link the identifier of the panel households, with last year table of identifiers.
- **Step 1** key variables : address, year, month and day of birth, sex, department of birth and first name
- **Step 2** key variables : address, year and month of birth, sex and first name
- **Step 3** key variables : address, year, month and day of birth, sex, department of birth
- **Step 4** key variables : municipalities (where the dwelling is), year, month and day of birth for married couple
- **Step 5** key variables : address, year of birth, sex, first name
- **Step 6** key variables : year, month and day of birth, sex, first name
- **Step 7** Manual research for those with probability of identification less than one

4. Ex. 2 : Linkage to income data for SILC

Tax payers found in the income tax file at each step of the linkage process

Steps	%	cumulated %
0	75.3	75.3
1	19.5	94.8
2	1.4	96.2
3	0.4	96.6
4	0.4	97.0
5	0.4	97.4
6	1.3	98.7
7	1.3	100.00

Source : SILC 2011, Insee, France (Burricand, 2012)

4. Ex. 2 : Linkage to income data for SILC

Presence of the person in the two sources

Presence	%	Cumulated %
Income tax returns	1.7	1.7
Survey	2.9	4.6
Survey and income tax revenue	95.4	100,0

Source : SILC 2011, Insee, France (Burricand, 2012)

4. Ex. 2 : Linkage to social security data for SILC

- SILC also linked with social files from the family and the elderly branches of the social security system:
- - CNAF (*Caisse nationale d'allocations familiales*) provides child and family benefits and manages minimum income programmes (90 % of benefits are paid by CNAF).
- - The CCMSA (Agriculture Mutual Benefit Fund) covers family benefits, housing allowances and old-age pensions and minimum old-age pension or Solidarity allowance for elderly.
- - The old age branch of *Caisse nationale d'assurance vieillesse*, CNAV (National Old Age Pension Fund) covers old-age pensions and minimum old-age pension for former employees.

4. Ex. 2 : Linkage to social security data for SILC

- 96 % of those who declared they received family allowances from CNAF were found in the social file.
- The linkage with MSA is lower (50%), as the information on the address of the beneficiary is poor in the MSA files.
- For minimum old-age pension, linkage done directly by the National Old Age Pension Fund. Among those aged 60+, 96 % are linked but only 1,4 % receive a minimum old-age pension, about half the expected proportion. Hence minimum old-age pension is often still imputed.

4. Ex. 2 : Linkage for SILC, Lessons

- A linkage based on identifying information such as name and address hinges on the quality of information. Default of linkage can be dealt with standard techniques of imputation. Necessary to ask questions on type of income.
- Even when the linkage works at the household level, some individuals such as young people cannot be linked. For children it has no impact, and for young adult, the problem can be solved through interviews.

4. Ex. 2 : Linkage for SILC, a methodological test

- Conducted to compare 2004 income distributions across sources.
- Wages: the vast majority of respondents correctly report whether they received wages during the year.

✓Mean wage very close between the two sources. Wages under-evaluated in the first quartile and over-evaluated in the last one.

✓For 80 % of employees, the difference between the two sources was small (less than 100 €/month). Most of those with a difference of less than 10 % had used their income tax return to answer. The difference were higher in case of a proxy interview.

4. Ex. 2 : Linkage for SILC, a methodological test

- Impact of the data source more important on distribution of retirement income.

✓ The amounts of pensions observed in the survey were higher than those of the tax files

✓ Some pensions are not taxable (additional pension for those who have raised at least three children, veteran pension...). Necessary to go on collecting by interview the non taxable components of income.

- Income from self-employment: the definitions are not the same between the two sources: priority given to survey .

4. Ex. 2 : Linkage for SILC

- In 2008, real estate income gathered from administrative data $\rightarrow$ total real estate income multiplied by two!

(1) under-estimation in the survey of the holding of real estate income;

(2) under-evaluation of the amounts : between 2007 and 2008, the average amount increased by 26 %.

- Less effect for other asset income: better-known by the households. Amounts were only slightly under-estimated.

4. Ex. 2 : Linkage for SILC, summary

- Better quality of some income data in the administrative file (wages). Linkage prevents under-evaluation, errors in collecting and reporting data in a survey. The impact of using a proxy is reduced.
- Still necessary to collect income for some categories of population (young people, self-employed) and to go on collecting the type of income received to impute income in case of error in the linkage.
- Combining the two sources is the best practice.
- Longitudinal interpretation modified by new linkage.

5. Social Security data : files on pension rights

- Many SHARE countries plan linkage to "social security" data, that is pension rights data.
- In France only a *sample* of future and current pensioners (*échantillons inter-régime de cotisants, EIC, et de retraités EIR*) is readily available.
- Not usable for linkage.
- Go back to data of various pension regimes (around 50 *caisses de retraite*).
- Time consuming. A compromise would be to concentrate of the data from the Caisse nationale d'Assurance Vieillesse (CNAV).

5. Social Security data : files on pension rights

- Most people contribute some day or the other to the CNAV, and CNAV has information (less precise though) on periods of contributions to other regimes for those persons.
- A linkage through the NIR : no technical difficulties, but needs a good legal base.
- If the NIR cannot be retrieved a linkage through demographic information could be tried.

This has to be tested (collaboration with CNAV needed…)

6. Conclusion: a triple issue

- **Scientific** issue

✓ If linkage with data already collected by interview, risk of "embarrassment of riches" , more can be less, time discontinuity, difference with the other countries. And little **financial or ethical** gain, if questionnaire can't be reduced.

✓ For SHARE in France an income linkage would be a means to know the after tax income. No *"à la source"* withdrawal of income tax.

✓ If linkage with data that have not been or cannot be got in the field (say pension rights), no such risks. More scientifically interesting if more SHARE countries are doing it too.

6. Conclusion: a triple issue

- **Financial (and practical)** issue

✓ Draw a refresher sample in the administrative source (for SS data)?

✓ Think ahead about the link identifier: the NIR or a reconstructed NIR (via birth name, first name, sex, day, month, year of birth.

✓ Remember that levels of information can vary in each sources : individual (pension), household (in some survey), fiscal unit (income tax returns), ayant-droit (health care insurance)…

6. Conclusion: a triple issue

- … “**ethical**”

✓ What about getting the respondent’s consent: implicit, explicit? Involves risk for the panel.

✓ In SHARE, the argument of reduced burden will be difficult to use if the CAPI remains the same whatever the link a country manages to do.

✓ In some countries, like France, some linkages might not be easily done outside the public service of statistics.