Microdata access for researchers in INSEE-France

Recent developments, open questions

Jean-François Royer –CREST-INSEE

Outline

- History
- Legal developments in 2008
- Current experiment of remote access
- Pending problems : some examples
- Conclusion

History

- Till 2008, under the French statistical law (1951)

- communication of individual data (issuing from a statistical process) when they are “related to personal or family life” is forbidden without exceptions (other than requisition by a judge)
- communication of such data when they are “of economic or financial order” is forbidden in general, but possible on demand (except for uses of administrative or fiscal control) ; these communications are under supervision of a special committee, “comité du secret”.

History - 2

- Accordingly, researchers in France have had :

- very difficult access to confidential microdata of household surveys, censuses, etc. ; the only possible ways were to be accepted inside INSEE temporarily (very rare), or to obtain treatments made by INSEE (tables only - costly)
- possibility to get files of business surveys, after approval by “comité du secret” ; many French research teams have taken advantage of that opportunity, which is generally not available abroad.

History - 3

- In the recent years, more and more anonymous microdata have been released : public use files, available on Internet increasingly since 2003 (22 files on www.insee.fr now) ; licensed files for researchers, called “fichiers de production et de recherche (FPR)” available through a partner of the statistical system “réseau Quetelet”

History - 4

- Restrictions on the dissemination of public use microdata have been enacted to keep them anonymous, and also, in some cases (censuses, for example), to meet the recommendations of the “National commission for data processing and the Liberties (CNIL)”, which acts according to a 1978 law, and wants to prevent small groups of population to be described and possibly stigmatized or harmed.

Legal developments in 2008

- Two modifications of the statistical law (1951) :
- (Creation of a Statistical Authority)
- Communication of individual data (issuing from a statistical process) when they are “related to personal or family life” is still forbidden, but a possible exception has been introduced : for *public statistical purposes* or *scientific or historical research purposes.*

Legal developments in 2008 - 2

- Communication of individual and household data will therefore be possible now, after hearing the “comité du secret”, whose scope is extended to these data (before, his scope was restricted to economic data).
- This applies also to statistical data issuing from administrative processes (for example, tax data, or social security data…) ; in such cases, the agreement of the original administration will be needed.

Legal developments in 2008 - 3

- Creating a research data center in France giving researchers access to confidential data about individuals or households is now legally possible, but not yet done.
- In 2008-2009, we are experimenting such a center, using the previous legal system. Researchers who take part in the experiment have been formally “integrated in INSEE” for that purpose (without pay ! Same system as in Canada “statisticiens réputés”).

Current experiment in remote access

- Organised by CREST “Centre de recherche en économie et statistique” and by ENSAE “Ecole nationale de la statistique et de l’administration économique”, which both belong to INSEE.
- Three research laboratories outside INSEE are involved : PSE (“Paris Sciences économiques”), INED (“Institut national d’études démographiques”), TSE (“Toulouse sciences économiques”).
- 11 research projects, 16 researchers in the experiment.

Current experiment in remote access - 2

- Microdata files in the experiment :

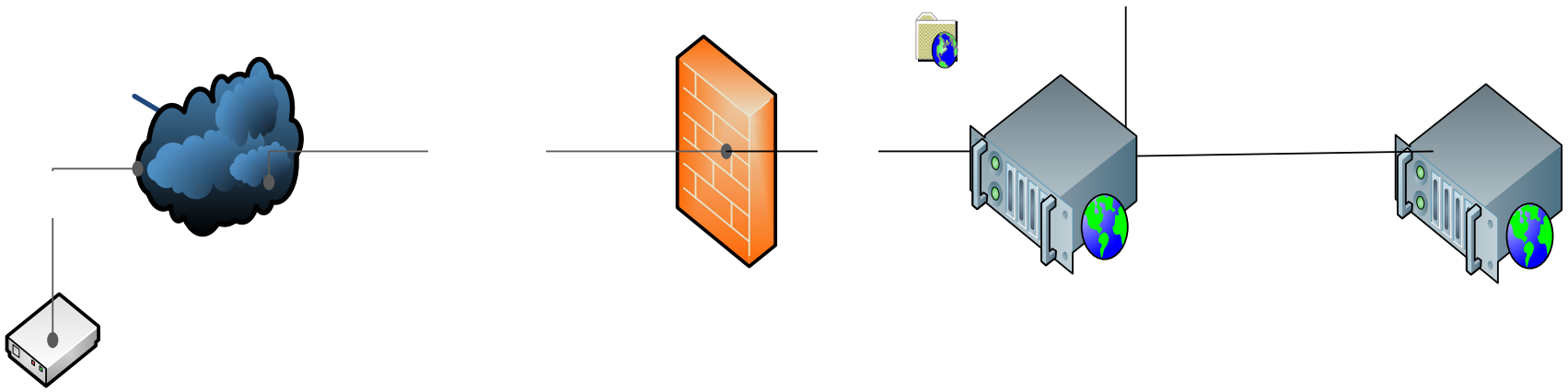
- household surveys : labor force, household budgets
- censuses : 1990, 1999
- yearly social data on employees from employers (“déclarations annuelles de données sociales DADS”)
- longitudinal sample of those data DADS
- births data
- permanent demographic sample (panel of census and register forms)

- No names, no addresses, no administrative personal identification numbers ; but the files include all interesting variables, and are therefore indirectly nominative.

Current experiment in remote access - 3

- Technical device : a central server in INSEE contains the confidential data files ; several terminals called “S-BOXes” in the laboratories (provided by INSEE) are used by the researchers.
- Connection between terminals and server : through Internet “securitized tunnels” .
- The researchers open sessions on the server, where they work with SAS, STATA,…

Current experiment in remote access - 4



Current experiment in remote access - 5

- Security devices :

- each researcher has a personal chip card, necessary to use the S-BOX, with a password
- each researcher can only see the data that he/she has been allowed to see for his/her project ; no possible communication between different projects (but there may be several researchers on one project)
- printing, exporting files, cut/copy are impossible for the researchers

Current experiment in remote access - 6

- Output control : we check every output , and keep a copy, before sending it to the researcher. Not a lot of work till now, but later…?

Current experiment in remote access - 7

- Results expected of the experiment : proof that “it can work” ; list of the crucial difficulties, either technical or relational ; estimation of the costs for a larger implementation (investment costs, functioning costs).
- Report to the direction of INSEE next month.

Pending problems

- There are many ! I will insist on three :
- 1° Who will manage the RDC ?

➢ INSEE ?

➢ The whole statistical service ? (INSEE and “services statistiques des ministères”)

➢ INSEE and representatives of the research “community” ? (for example “Réseau Quetelet”)

- Behind that question, who pays what ?
- Anyway, the “comité du secret” will play an important role.

Pending problems - 2

- 2° Security of the data : who is responsible ?

- The researchers who will be allowed to use the confidential data will be legally responsible for their uses.
- Nevertheless, the statisticians have to take dispositions that prevent accidental disclosure, and that keep everybody in sufficient care for this problem.
- Is there a risk that systematic output control shifts all responsibility towards the INS, and produces a lack of vigilance from researchers ?
- Symmetrically, wouldn't a posteriori sampled controls be interpreted as a lack of care from statisticians ?

Pending problems -3

- 3° Matching individual data

➢ Potentially huge demand from researchers

➢ If the files to be matched are both produced by INSEE, the problem could be solved outside the RDC : if the project is approved by “comité du secret” and CNIL, INSEE does the matching, removes the identifying variables, and gives access to the data through RDC

➢ If one of the files is provided by the researcher himself, the problem seems tougher, since we don’t want to include direct identifiers in RDC, neither to process unknown files ourselves.

Conclusion

- France is late on this topic : the main reason is the characteristics of its statistical law, till 2008.
- We have no experience of the techniques that modify the data for security purposes : there is no current or foreseen investment in this field.
- In the future, giving access to researchers should rely on the three channels : public use files ; if insufficient, licensed files (FPR) ; if insufficient, RDC with remote access.
- Now we have a better legal framework, and, I hope, a real will to make rapid progress.
- But some important decisions have not been taken yet.