



La gestion de la confidentialité à l'Insee

Réunion du groupe CAPPRIS

Maxime Bergeat
Département des méthodes statistiques

Insee
Mesurer pour comprendre

8 octobre 2014

En quelques mots

- Pourquoi ?
- Aujourd'hui, comment ?
- Quels projets pour faire mieux demain ?
- Qui ?

En quelques mots

- Pourquoi ?
- Aujourd'hui, comment ?
- Quels projets pour faire mieux demain ?
- Qui ?

Le secret, pourquoi ?

But : Protéger les données individuelles

- Empêcher les reconstructions des données individuelles
- Par des intrus potentiels
- Qui possèdent de l'information auxiliaire

Enjeux :

- Conserver la confiance des répondants et garantir les taux de réponse élevés
- Un cadre légal à prendre en compte
- Tout en cherchant à diffuser l'information la plus complète possible
- ➔ Un arbitrage à réaliser

Un cadre législatif à respecter

Loi du 7 juin 1951 (Article 6) :

- Définit le secret en matière de statistique
- Création du comité du secret
- Pas de règles précises

Code de bonnes pratiques de la statistique européenne :

- Principe 5 : Le respect de la vie privée ou du secret des affaires des fournisseurs de données (ménages, entreprises, administrations et autres répondants), la confidentialité des informations qu'ils communiquent et l'utilisation de celles-ci à des fins strictement statistiques doivent être absolument garantis.

- 6 indicateurs : loi, engagement du personnel, sanctions, instructions fournies lors de la production, dispositions matérielles et techniques, protocoles stricts d'accès aux microdonnées (pour les chercheurs aujourd'hui).

En quelques mots

- Pourquoi ?
- Aujourd'hui, comment ?
  - Données tabulées
  - Données individuelles
- Quels projets pour faire mieux demain ?
- Qui ?

Pour les données tabulées (1) – Un cadre juridique à prendre en compte

Des décisions de l'Insee qui font jurisprudence sur les données de statistique d'entreprises

- Depuis 1980
- Pour les tableaux de données agrégées
- Deux règles à respecter pour la diffusion (secret primaire)
  - Règle des trois unités
  - Règle de dominance (1, 85)
  - Communication effectuée sur ces règles (guide du secret statistique)
- Prise en compte du mécanisme d'échantillonnage
  - Hypothèse sous-jacente : pas de diffusion par ailleurs des poids de sondage

Pour les données tabulées (2) – Un cadre méthodologique et logiciel clairement défini

Une méthodologie fondée sur la suppression de cellules

- Les cases non diffusables sont « supprimées » (secret primaire)
- Des cases complémentaires sont supprimées, étant donné la diffusion des marges (secret secondaire)
- Des algorithmes de minimisation de l'information perdue sous contrainte de ne pas pouvoir estimer trop précisément les cases en secret primaire
  - Minimisation de l'information perdue (chaque case a un coût de suppression)
  - Sous contrainte de respecter des « intervalles de protection » pour les cases jugées sensibles

Outillage logiciel

- Tau-Argus, développé à l'origine par l'institut néerlandais CBS et financé par Eurostat
  - Portage en Open-Source en cours pour faciliter les futurs développements, qui pourront être effectués à l'échelle européenne
  - Logiciel le plus utilisé en production en Europe par les Instituts Nationaux de Statistique (17 pays sur 22 répondants à une enquête menée en 2013), souvent en complément d'autres procédures moins automatisées (gestion du secret secondaire « à la main », procédures SAS)...
- D'autres outils existent
  - Le package R sdcTable, développé en partie par Statistics Austria mais peu utilisé en Europe
  - L'ensemble de macros SAS CONFID2, utilisé à Statistique Canada

Pour les données tabulées (3) – Sensibiliser et former

Actions de formation

- Sur le cadre théorique de la gestion de la confidentialité pour les tableaux de données, incluant une initiation à Tau-Argus
- Public visé : producteurs de données dans la statistique d'entreprises ou dans les services statistiques ministériels, personnes élaborant des produits sur mesure dans les Directions régionales de l'Insee
- Deux formations par an d'un jour et demi
- D'autres actions de formation (plus générales et plus axées sur la législation) menées par le service d' « Action régionale »

Séminaires de sensibilisation aux questions liées à la confidentialité

- Séminaire de méthodologie statistique de l'Insee
- Coopération internationale (avec pays d'Afrique en particulier)

Élaboration et diffusion de « bonnes pratiques » méthodologiques

- Quel est le meilleur paramétrage du logiciel pour assurer une « bonne » protection ?
  - La notion des intervalles de protection
  - Minimisation de la perte d'information engendrée par le secret secondaire
  - Comparaison de différents algorithmes pour la suppression secondaire
- Soutien méthodologique et technique en continu

Pour les données tabulées (4) – Des difficultés à prendre en compte

Une difficulté majeure : prendre en compte les liens...

- Entre différents tableaux où la même variable est agrégée
- Entre différentes variables agrégées dans les tableaux et liées (par exemple par des égalités comptables)
- Entre différents demandeurs et producteurs de données, construisant des tableaux à partir des mêmes sources, et pas forcément en même temps

Des limitations informatiques

- Algorithmes de suppression secondaire « gourmands » en ressource
- Des « bridages » du logiciel Tau-Argus, qui empêchent de construire des tableaux avec de multiples variables de ventilation

Pour les données individuelles (1) – Trois niveaux de diffusion

Fichiers grand public diffusés sur insee.fr (*public use files*)

- Uniquement des données administratives pour les entreprises (dénombrement des entreprises avec localisation géographique, secteur d'activité et une variable de tranche d'effectif)
- Fichiers « ménage » issus du recensement de la population et de l'enquête Emploi

Fichiers pour les chercheurs diffusés via le réseau Quetelet

- Accréditation du chercheur nécessaire, qui est contractualisé et doit appartenir à un organisme reconnu
- Ensuite, accès à de nombreuses bases de données **sur les ménages**, où le niveau de détail est réduit pour limiter le risque de ré-identification

Fichiers diffusés au sein du Centre d'Accès Sécurisé Distant (CASD)

- Avant, il fallait venir travailler à l'Insee après la signature d'une convention
- Seulement suppression des identifiants directs (numéro de sécurité sociale, nom complet, numéro SIREN...)
- Accès à des données d'enquêtes auprès des ménages, à des données fiscales...
- Passage devant le Comité du secret, qui doit valider le projet de recherche, avant de pouvoir accéder aux données demandées. Le Comité se réunit quatre fois par an.
- Accès à distance aux données au sein d'un environnement sécurisé (identification digitale en particulier)
- En cas de volonté de sortir des données (pour insertion dans une publication par exemple), de l'*output checking* est réalisé

Pour les données individuelles (2) – Des protections statistiques simples

Pour les fichiers diffusés *via* le réseau Quetelet

- Définition d'une note d'objectifs de sécurité
- Qui donne la structure de diffusion de ce type de fichiers
- Méthodes d'anonymisation mises en œuvre : suppression de variables très identifiantes ou sensibles, réduction du niveau de détail (*global recoding*) pour les variables indirectement identifiantes
  - Dans les fichiers pour les chercheurs, la localisation géographique est rarement plus précise que le département.
  - Les poids d'échantillonnage sont généralement diffusés.

Pas de quantification systématique du risque de divulgation ou de l'utilité des fichiers produits

- Les décisions pour la diffusion sont prises au sein des Comités de pilotage de chaque enquête.
- Pas encore de méthodologie commune définie concernant l'anonymisation de tels fichiers.

Des méthodes pas encore utilisées

- Méthodes perturbatrices
- Génération de données synthétiques

Pour les données individuelles (3) – Des verrous informatiques pour les fichiers les plus détaillés

Pour les échanges internes (ou entre partenaires du Système Statistique Public) de données non anonymisées

- Des « tuyaux » bien sécurisés et des procédures d'échange optimisées
- Toute la partie légale (contractualisation et validité des échanges) est gérée en amont, par le service juridique de l'Insee

Pour les chercheurs accrédités qui veulent accéder à des données non anonymisées

- La solution du CASD (Centre d'Accès Sécurisé Distant)
- Une solution payante (environ 100€ HT / mois pour 2014, tarif en augmentation jusqu'en 2019)
- Les mêmes données que celles utilisées à l'Insee, les identifiants directs en moins

En quelques mots

- Pourquoi ?
- Aujourd'hui, comment ?
- Quels projets pour faire mieux demain ?
  - Au niveau de l'Insee
  - Au niveau européen
  - Et ailleurs ?
- Qui ?

À l'Insee, pour les données tabulées

Pour le travail de production courant

- Harmonisation des pratiques, qui peuvent différer entre les différents services de l'Insee
  - Procédures de transmission des données
  - Méthodologie utilisée dans la recherche du secret secondaire (méthodes algorithmiques, coût employé pour les suppressions secondaires, définition des intervalles de protection pour les cases jugées sensibles...)
  - Harmonisation logicielle

Des défis en cours

- De la possibilité d'opérer les traitements liés à la confidentialité directement avec SAS, par exemple avec un « encapsulage » de Tau-Argus, pour faciliter l'intégration à une chaîne de production courante
  - Des outils existent déjà dans d'autres instituts (Suède, Allemagne)
- La question de la gestion d'hypercubes de données (dont les utilisateurs pourraient extraire ce qu'ils veulent *via* un requêteur, par exemple)
  - D'un point de vue logiciel : les tests réalisés avec Tau-Argus sont peu probants
  - D'un point de vue méthodologique : à mener
- Le portage du logiciel Tau-Argus en Open-Source (première version Bêta disponible, version courante à venir d'ici fin 2014) qui ouvre de nouvelles possibilités de réflexion (l'implémentation de nouvelles méthodes sera facilitée)

À l'Insee, pour les données individuelles

Revue de littérature
- Les différentes méthodes
- L'outillage logiciel à disposition
- Le cadre législatif à respecter

Des tests opérés sur données réelles
- Un travail exploratoire débuté durant l'été 2014
- Dans le cadre de l'enquête Victimation menée par l'Insee
- Un objectif de réduction du risque de divulgation : la k-anonymisation
- Différentes méthodes testées : agrégation de l'information, suppression locale de modalités des variables, microagrégation
- Utilité des fichiers produits mesurée en regardant les analyses standard réalisées par les chargés d'études de l'Insee
  - Tabulations complexes
  - Statistiques descriptives bivariées
  - Modèles de régression logistique
- Des travaux à poursuivre

De grandes questions
- Une méthodologie unifiée pour toutes les enquêtes à diffuser (mesure du risque de divulgation, type de méthodes d'anonymisation, mesure de l'information perdue...)
- Quel(s) logiciel(s) utiliser pour mener l'anonymisation ?
  - µ-Argus
  - sdcMicro (package R)

Collaboration européenne – Eurostat (1)

Des réunions régulières sur la confidentialité

- Un groupe de travail où sont représentés des méthodologues et des experts juridiques
- Un groupe d'experts qui réunit les méthodologues de différents instituts européens

Des groupes consultés pour...

- Définir et diffuser des bonnes pratiques
  - Protection des tableaux de données
  - Protection des données individuelles
  - *Output checking*
- Effectuer des propositions d'anonymisations pour des fichiers de données individuelles
  - Raisonnement fichier par fichier
  - Décisions prises *in fine* dans les *Task forces* correspondant à l'enquête en question, par les instituts nationaux de statistique
- Discuter des évolutions logicielles en cours et à venir

Collaboration européenne – Eurostat (2)

Eurostat mis à contribution

- Élaboration de fichiers (pour les chercheurs pour le moment) selon les critères définis
- Un centre de recherche sécurisé (centre physique situé à Luxembourg)
- *Output checking* pour ce que veulent sortir les chercheurs du centre d'accès sécurisé

Collaboration européenne – Un nouveau projet prometteur

Framework Partnership Agreement on Statistical Disclosure Control

- Un projet impliquant 7 pays dont la France
- Prévu pour une durée de 4 ans, lancement officiel début 2015
- Un financement par Eurostat
- **L'objectif est d'aboutir à une meilleure coordination en ce qui concerne la confidentialité et l'accès aux fichiers de données individuelles**
  - Élaboration de « bonnes pratiques », en ce qui concerne les méthodes à privilégier mais également les meilleurs outils logiciels pour réaliser l'anonymisation
  - Effort de formation (à l'échelle européenne) pour diffuser les bonnes pratiques

En quelques mots

- Pourquoi ?
- Aujourd'hui, comment ?
- Quels projets pour faire mieux demain ?
- Qui ?

Qui ?

Des producteurs de données sensibilisées à ces problématiques

- Engagements pour le respect de la confidentialité
- Actions de formation
- Diffusion de « bonnes pratiques »
- Méthodologues prenant en charge certaines questions liées à la confidentialité répartis dans les services producteurs (au recensement par exemple)

Deux méthodologues spécialistes de la confidentialité

- Une personne chargée de la confection des masques de secret (pour les tableaux de données sur les entreprises) et de travaux méthodologiques sur les données tabulées
- Un expert référent sur les aspects méthodologiques et investi dans divers projets de coopération

Une unité juridique spécialiste des questions légales

- Dont une personne impliquée dans le groupe de travail européen sur la confidentialité

Des personnes qui travaillent sur ces problématiques au sein du CASD (entité du Genes : Groupe des Écoles Nationales d'Économie et Statistique)

- Sécurisation informatique
- *Output checking*

Conclusions

Pour les tableaux de données

- Une procédure bien rodée
- Un cadre méthodologique et logiciel
- Un effort de formation
- Des efforts pour l'harmonisation

Pour les fichiers de données individuelles

- De gros investissements en cours
  - Implication dans plusieurs projets visant à l'harmonisation des (et la définition de bonnes) pratiques, notamment en Europe
  - Doublement des experts « confidentialité » dans le service de méthodologie
- Et il reste à « choisir » un/des outil(s) informatique(s) pour pouvoir réaliser l'anonymisation des fichiers

Références bibliographiques

- M. Bergeat. *Un panorama de la protection des fichiers de données individuelles*, Séminaire de Méthodologie Statistique de l'Insee, juin 2014, disponible en ligne.

- M. Bergeat *& al*. *A French anonymization Experiment with Health Data*, "Privacy in Statistical Databases" conference, septembre 2014.

- Eurostat. *Results on the questionnaire on SDC tools*, 5th meeting of the Expert Group on Statistical Disclosure Control, octobre 2013.

- A. Hundepool *& al*. *Statistical disclosure control*, Wiley Series in Survey Methodology, 2012.

- Insee. *Guide du secret statistique*, octobre 2010, disponible en ligne.

La gestion de la confidentialité à l'Insee

Merci pour votre attention ☺

Contact :
Maxime Bergeat
01 41 17 64 86
maxime.bergeat@insee.fr

Insee
18 bd Adolphe-Pinard
75675 Paris Cedex 14
www.insee.fr
Informations statistiques :
www.insee.fr / Contacter l'Insee
09 72 72 4000
(coût d'un appel local)
du lundi au vendredi de 9h00 à 17h00

Insee