Protection des données à caractère personnel et qualité des enquêtes statistiques

Thomas Soubiran [1]

[1] Ceraps - UMR 8026 du CNRS

Journée d'études APPEL

« Le cadre juridique applicable aux traitements de données à caractère personnel »

Lille, le 28 avril 2017

Introduction :

L'utilisation des données à caractère personnel dans les enquêtes par questionnaires :

- Dans le cas des enquêtes par sondages aléatoires, la collecte repose sur la disposition de **données auxiliaires**
- Les données auxiliaires recouvrent les données connues *a priori* sur la population enquêtées pour en sélectionner des individus
- Ces données sont nécessaires pour assurer la qualité des données recueillies
- Ces données auxiliaires peuvent être **nominatives** ou, *a minima*, identifiantes

Introduction

- Dans les enquêtes par questionnaires (et pas seulement), **collecte et analyse** des données sont des phases **liées mais distinctes** :
  - Ces données ne sont mobilisées que dans la phase de collecte
  - L'exploitation des réponses aux questionnaires est souvent **anonyme**
- Cependant, la loi du 6 janvier 1978 modifiée (LIL) pas plus que le nouveau règlement européen sur la protection des données n'opèrent cette distinction
- Le traitement doit être considéré dans sa globalité et inclut donc la collecte
- Des analyses anonymes en pratique se retrouvent donc soumis aux obligations incombant au traitement de données à caractère personnel

Note : La suite ne traitera que des situations où les données analysées sont anonymes une fois la collecte terminée

Plan

- L'utilisation de données auxiliaires à caractère personnel dans la collecte des réponses à des enquêtes par questionnaires et leur importance pour la qualité des données
- L'impact de ce mode de collecte sur l'analyse juridique
- Ébauche d'une proposition visant à trouver un équilibre entre protection des individus et qualité des données dans le cas particulier du recours à des données auxiliaires

Plan

La théorie des sondages

- La sélection des individus est cruciale pour la qualité des données et donc des analyses
- Cet aspect a donc fait l'objet du développement d'une branche spécifique de la statistique à partir de la fin du XIX^{e} siècle : **la théorie des sondages**
- Cette branche a permis d'adapter la théorie statistique inférentielle au populations **finies** (66,9 millions d'habitants en France)
- La présentation s'appuiera sur certains résultats de la théorie des sondages en insistant sur leurs implications pratiques et l'utilisation des DCP

Note : Il existe plusieurs approches dans la théorie. Ce qui suit reprend les grandes lignes de l'approche par *plan de sondage*.

Motivation

- Il s'agit donc de trouver, à partir d'un échantillon, des valeurs proches de celles de la population dont l'échantillon est issu
- Il ne sera pas ici question de **« représentativité »** pas plus que de « photographies » ou de « miniatures »,. . .
- La question sera plutôt celle du **biais** et la **précision** des estimateurs et leur **optimisation** (minimisation|maximisation) comme critère de qualité des enquêtes.
- On recherche un estimateur $\hat{\theta}_{\mathcal{S}}$ des paramètres proche « en moyenne » du paramètre θ calculé sur l'intégralité de la population (minimisation du biais) :

$$\begin{aligned} B[\hat{\theta}_{\mathcal{S}}] &= E[\hat{\theta}_{\mathcal{S}} - \theta] \\ &= \sum_{s \subset U} \mathbb{P}(\mathcal{S} = s)[\hat{\theta}_s - \theta] \end{aligned}$$

  et ce, de façon à ce que cette espérance ne soit pas trop dispersée (maximisation de la précision) :

$$\begin{aligned} V[\hat{\theta}_{\mathcal{S}}] &= E[\hat{\theta}_{\mathcal{S}} - E(\hat{\theta}_{\mathcal{S}})^2] \\ &= \sum_{s \subset U} \mathbb{P}(\mathcal{S} = s)[\hat{\theta}_s - E(\hat{\theta}_{\mathcal{S}})] \end{aligned}$$

- Mais, pour cela, on a besoin de maîtriser le processus qui a généré les données (-ie : la collecte) et donc de disposer d'information sur la population enquêtée

Le sondage aléatoire simple

- La méthode fondamentale de la théorie des sondages est le sondage aléatoire simple (SAS) avec ou sans remise
- Il ne nécessite qu'une liste recensant les individus (unités) de la population visée
- On sélectionne les individus en procédant à un tirage dans la liste en utilisant un algorithme reposant sur un générateur de nombres (pseudo-)aléatoires
- Cela revient à générer une variable indicatrice $I_k = \mathbb{1}\{k \in \mathcal{S}\}$ marquant l'appartenance d'un individu k à un échantillon $\mathcal{S}$
- Comme il s'agit d'une variable aléatoire, on peut :
  - calculer la probabilité π_k qu'un individu k soit dans un échantillon (ici, la probabilité est la même pour tous $\pi = n/N$)
  - en dériver un estimateur
  - ainsi que la variance de cet estimateur (ou une approximation dans certains cas)

Exemples d'estimateurs

Pour un SAS :

- Estimateur d'un total : $\hat{t}_{y\pi} = \pi^{-1} \sum_{k\in\mathcal{S}} y_k$

  La moyenne de la population est donc ici égale à la moyenne de l'échantillon : $\hat{\bar{y}} = \frac{1}{N} \sum_s y_k \frac{N}{n} = \frac{1}{n} \sum_s y_k$
- Estimateur de sa variance :

$$\hat{V}(\hat{t}_{y\pi}) = N^2 \left(\mathbf{1} - \frac{\boldsymbol{n}}{\boldsymbol{N}}\right) \frac{1}{n} s_y^2, \quad s_y^2 = \frac{1}{n-1} \sum_{k\in\mathcal{S}} (y_k - \hat{\bar{y}})^2$$

- Le terme s_y^2 désigne la définition habituelle de la variance corrigée
- Le terme $(1 - n/N)$ (la correction de population finie) indique que plus le taux de sondage sera élevé, plus la variance de l'estimateur sera faible et, ce faisant, plus cet estimateur sera précis
- Cependant, la précision dépend de la taille de la population et de la dispersion de la variable d'intérêt

Sondages aléatoires et données à caractère personnel

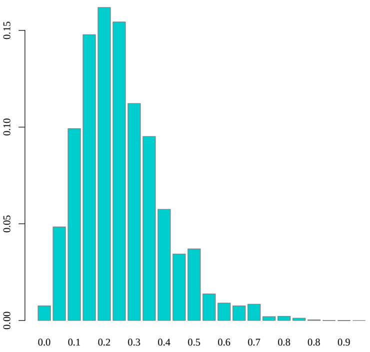
- La possibilité de calculer les probabilités d'inclusion π_k dans $\mathcal{S}$ conditionne la dérivation des estimateur
- Ce calcul est lui-même conditionné sur la disposition d'une liste d'individus **identifiés de façon univoque** auxquels **on est sûr d'avoir accès**
- La mise en œuvre du SAS nécessite donc de disposer de données auxiliaires à jour qui ont souvent un caractère personnel
- Il peut s'agir d'informations de contact : noms, prénoms, adresses postales, adresses électroniques,...

Les limites du SAS

- Ce sondage ne nécessite donc qu'un nombre limité de renseignements sur les individus de la population
- Néanmoins, le SAS est **rarement utilisé en pratique** :
  - Comme on l'a vu, la précision du sondage dépend notamment de la dispersion de la variable dans la population
  - En pratique, le SAS peut nécessiter un taux de sondage important pour donner un résultat satisfaisant en terme de qualité ce qui peut conduire à accroître significativement le coût de l'enquête.
- Si on dispose d'autres données auxiliaires (p. ex. des renseignements sur les personnes comme le sexe, l'age,...), il est possible de faire mieux pour un coût moindre
- La mise en œuvre d'un plan de sondage peut donc nécessiter de **disposer de DCP supplémentaires**

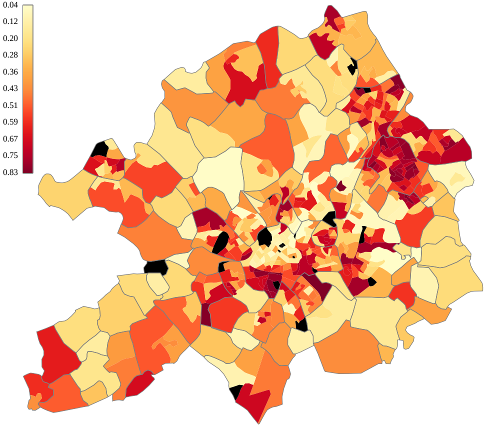
Hétérogénéité des populations

Pourcentage de chômage à l'Iris



Hétérogénéité spatiale des populations

Pourcentage de chômage à l'Iris dans l'agglomération lilloise



Source : IGN base Contours IRIS ® , INSEE enquête annuelle de recensement 2010

Les plans de sondages aléatoires complexes

- Le SAS peut conduire à sous|sur-représenter certaines caractéristiques en fonction de l'aléas des tirages
- Mobiliser des informations sur les individus et les contextes dans lesquels ils évoluent peut permettre de réduire le biais, l'erreur et le coût
- De nombreuses techniques ont été développées pour inclure des données auxiliaires dans la sélection des individus (sondages par stratification, par grappes, proportionnels à la taille, à plusieurs degrés,...)

Les données auxiliaires dans la conception des plans de sondages

Lorsqu'on dispose des données auxiliaires nécessaires à cet effet, la théorie des sondages propose de nombreuses techniques pour concevoir des plans de sondages optimaux :

- Les données auxiliaires peuvent aussi permettre de comparer *a priori* les mérites de différents plans de sondage :

$$\mathsf{deff} = \frac{V_{p(\mathcal{S})_1}(\hat{t}_\pi)}{V_{p(\mathcal{S})_2}(\hat{t}_\pi)}$$

- Disposer de données auxilaires permet, par exemple, de déterminer la taille suffisante de l'échantillon
- Elles peuvent aussi permettre d'allouer de façon optimale les ressources pour un sondage stratifié en minimisant :

$$\underset{V(\bar{y}_\mathcal{S})}{\arg\min} V(\bar{y}_\mathcal{S}) = \sum_{h=1}^{H} \frac{N_h^2}{N^2} \frac{S_{y\,h}^2}{n_h} \quad \text{avec la contrainte} \sum_{h=1}^{H} n_h = n$$

  Note : les variances $S_{y\,h}$ pour chacues des strates doivent être connues
- ...

Le traitement de la non-réponse

Pour compliquer un peu les choses (lors de l'analyse des données mais aussi du point de vue de l'analyse juridique) :

- Les données auxiliaires ne sont pas seulement utiles lors de la sélection des individus, elles sont aussi utiles **tout au long de la collecte**
- Les enquêtes sont en effet le plus souvent affligées par la **non-réponse** aux enquêtes qui peut conduire à de sévères biais
- Pouvoir observer la non-réponse au fil de la collecte peut permettre, par exemple, de concentrer les efforts sur certains individus pour limiter le biais (relances téléphoniques, papier,...)
- Une fois la collecte terminée, la pondération peut être corrigée en fonction des caractéristiques des non-répondants
- L'attention aux processus de non-réponses est d'autant plus importante que celle-ci ne cesse de croître en France et ailleurs

Plan

Les deux temps du traitement

- Les données auxiliaires (à caractère personnel ou non) revêtent donc un caractère crucial pour la qualité des données
- Elles peuvent se révéler utiles tant dans la phase de conception de l'enquête qu'au long de la collecte
- Collecte et analyse des données sont néanmoins **deux temps distincts**
- Les données auxiliaires n'ont généralement qu'un caractère instrumental : une fois la collecte terminée et les variables de pondération calculées, **elles ne sont plus nécessaires**

  Note : Certain éléments peuvent être conservés dans dans certains cas comme le tirage stratifié ou par grappe où les variables de stratification ou d'agrégation sont requis par les calculs. Toutefois, seule une variable indicatrice est nécessaire.

- Dans un nombre non négligeable de cas, lorsque l'analyse commence, les données ne contiennent plus d'informations (in)directement identifiantes

Les deux temps du traitement

- Même si les analyses ne nécessitent pas de manipuler de données à caractère personnel, le fait que la collecte des données en ait utilisé fait tomber le traitement sous le coup de la législation en vigueur en la matière :

  > « Constitue un traitement de données à caractère personnel toute opération ou tout ensemble d'opérations portant sur de telles données, quel que soit le procédé utilisé, et notamment *la collecte, l'enregistrement, l'organisation, la conservation*, l'adaptation ou la modification, l'extraction, la consultation, l'utilisation, la communication par transmission, diffusion ou toute autre forme de mise à disposition, le rapprochement ou l'interconnexion, ainsi que le verrouillage, l'effacement ou la destruction. »(`art. 2` de la LIL)

- En l'absence de distinction entre collecte et analyse, des analyses au final anonymes sont soumises à la loi
- Ceci complique fortement l'analyse juridique des traitements et les contraintes qui en découlent

Données sensibles

- Ces difficultés sont particulièrement manifestes lorsque des données **sensibles** sont collectées

  Exemple : la collecte d'informations sur les croyances et pratiques religieuses n'a rien d'exceptionnel même si l'on ne travaille pas sur le fait religieux (« fait social total » - M. Mauss)

- Plus généralement, la collecte de DCP doit répondre au principe de **proportionnalité** qui se traduit par la **minimisation** des données :
  - « Seules les données strictement nécessaires à la réalisation de la finalité peuvent être collectées »
  - La minimisation peut fortement interférer avec la problématique de l'enquête
  - Comment justifier de la pertinence d'une hypothèse avant de l'avoir testée ?
  - Cas limites : l'enquête exploratoire et, *a fortiori*, l'enquête prosopographique sont des oxymores (la collecte est la finalité)

Fazit

- Dans le cas des enquêtes par questionnaire, les obligations incombant aux responsables de traitement apparaissent parfois quelque peu disproportionnées par rapport aux **risques effectifs** encourus
- Lorsque les données à caractère personnel détenues sur les individus et leurs réponses ne sont jamais croisées, le risque se trouve plutôt du côté de **la sécurité et la confidentialité des données** et non au niveau des informations recueillies
- De plus, les procédures statistiques décrites sont suffisamment **standardisées** pour que les formalités qui s'y rattachent le soient aussi
- Quelles pourraient alors être les possibilités d'aménagements ?

Les méthodologies de référence

- Les méthodologies de référence désignent des procédures homologuées par la Cnil correspondant aux « catégories les plus usuelles de traitements automatisés de données de santé à caractère personnel à des fins de recherche, d'étude ou d'évaluation dans le domaine de la santé » (`art. 54` de la Lil)
- Elles ne concernent donc que la recherche en santé :
  - **MR-001** : *Recherches dans le domaine de la santé avec recueil du consentement*
  - **MR-002** : *Études non interventionnelles de performances concernant les dispositifs médicaux de diagnostic* in vitro
  - **MR-003** : *Recherches dans le domaine de la santé sans recueil du consentement*
- Chaque MR détaille le type de traitements auxquels elles s'applique (et ceux auxquels elle ne s'applique pas), la nature et l'origine des données, leurs destinataires,... et ce qu'elle autorise dans ce cadre précis

Une « méthodologies de référence » ?

- Même si elle sont limitées aux recherche en santé, les MR sont intéressantes à la fois du point de vue de la démarche mais aussi dans leur contenu (notamment la MR-003)
- Dans le cas qui nous intéresse ici, il s'agirait de clairement différencier collecte et analyse pour **distinguer les traitements** pour lesquels les DCP ne sont nécessaires que pour la collecte et non pour l'analyse
- Ce qui pourrait notamment impliquer que :
  - les DCP mobilisables pour la collecte soient précisées (origine et nature)
  - les nombre des destinataires des données auxiliaires à caractère personnel soit restreint aux personnes en charge de la réalisation de la collecte
  - le niveau de sécurité requis soit explicité
  - les personnes soient informées et puissent exercer un droit de retrait mais que certaines des données auxiliaires les concernant puissent malgré tout être conservées, hors données directement identifiantes (cf. traitement de la non-réponse)
  - la durée de conservation soit strictement limitée à la phase de collecte
  - les données analysées ne comportent aucune référence aux données auxiliaires à caractère personnel utilisées et que le lien soit impossible (p. ex. en détruisant les clefs d'appariement entre fichiers)

Une « méthodologies de référence » ?

- La question des DCP se cantonnerait donc à la collecte sans interférer sur les informations recueillies
- L'élaboration de ce type de procédure semble conforme à l'esprit du nouveau règlement européen :
  - Le règlement insiste en effet plus sur la mise en conformité ainsi que le contrôle *a posteriori* et moins sur les formalités préalables qui seront pour partie amenées à disparaître
  - Les responsables de traitement ne sont pas pour autant exempts de toutes obligations (au contraire, certaines sont mêmes renforcées)
  - Plutôt, le concept règlement introduit la notion de « *privacy by design* » (protection de la vie privée dès la conception) dans le droit européen qui implique d'adopter les mesures appropriées aux finalités et à la nature des données traitées dès la conception
  - Soit autant d'évolutions qui nécessitent de pouvoir disposer de lignes directrices précises à l'instar des MR

Collecter des données auxiliaires

- La rédaction d'une « méthodologies de référence » pose notamment la question de **l'origine des données**
- Pour disposer de ces données, il faut au préalable convertir des renseignements sur les individus composant la population visée en base de sondage
- Or, ces renseignements ont souvent été collectés à **d'autres fins**
- Sa réalisation peut aussi nécessiter le croisement de données collectées à des fins différentes
- Question d'autant plus importante que la **diffusion du numérique** présente des possibilités inédites en matière d'accès aux populations alors que les méthodes développées par le passé sont de moins en moins efficaces

Plan

Conclusion

- Le cas des données auxiliaires n'est qu'une des nombreuses illustrations des difficultés rencontrées lors de l'application du cadre juridique applicable aux DCP en SHS
- Il illustre de plus un processus courant d'anonymisation (ou de pseudonymisation) progressive lors de l'analyse des données après leur collecte qui n'est pas propre aux enquêtes statistiques
- Dans les enquêtes par entretiens, les choses sont moins tranchées que pour les enquêtes par questionnaires. Néanmoins, les interprétations s'éloignent le plus souvent des cas particuliers pour monter en généralité (passage du cas concret au cas pratique). Par exemple, les citations d'entretiens sont anonymisées lorsqu'elles sont publiées.
- Il ne s'agit pas à proprement parler d'une anonymisation au sens de la Cnil. Il s'agit plutôt du traitement lui-même qui devient anonyme.

Conclusion

- Les traitements en SHS ne visent souvent des personnes physiques que pour mieux s'en abstraire et produire au final des discours de portée générale non contingentés à un échantillon ou un autre
- Ceci pose la question de la spécificité de la finalité des traitements en SHS
- L'activité de recherche en SHS diffère de l'activité des entreprises, administrations, associations,...
- Nous n'avons pas à faire à des administrés, des assurés sociaux, des usagers, des employés, des clients,... mais bien à des *enquêtés*
- Il n'est pas question ici de prestations de services, de transactions commerciales, de relations de subordination, de surveillance, de monétarisation des DCP,... L'objet premier est de produire des *connaissances*.
- Ceci plaide pour une reconnaissance des spécificités des finalités de recherche en SHS qui, à l'instar de la recherche médicale, permettrait notamment de faciliter les démarches incombant aux traitement de DCP sans pour autant être dommageable pour la protection des personnes

Merci pour votre attention