Diffuser une base anonymisée : utopie ou réalitée ?

J. Djiriguian[1] N. Missègue[2] L. Ricroch[3]

[1]Unité Innovation et sécurité des données Drees

[2]Cheffe de projet statistique Drees

[3]Bureau Handicap dépendance Drees

JMS 2022 / 30 mars / Session 11

Plan de la présentation

- Bénéficiaires de l'APA vivant à domicile / environ 618 000 individus (96 départements)
- **L'APA ??** Allocation personnalisée d'autonomie : prise en charge de dépenses d'aide pour réaliser activités de la vie quotidienne
- Pas de variable directement identifiante (nom...), pas de NIR

Var. indirectement identifiantes et var. sensibles

Les diffuser ? En l'état ou tranches/regroupements ? Les "flouter" ?

Protéger notre fichier de 3 risques principaux :

1. divulgation d'identité : *Mme M. retrouvée comme bénéficiaire*
2. divulgation d'attribut : *divulguer le montant des ressources de Mme M.*
3. révélation inférentielle : *tous les bénéficiaires de tel département, sexe, etc. sont très dépendants*

Focalisation : **risque de ré-identification (autres risques traités)**

1. **Var. à ne pas diffuser** : risques de ré-identifier (ex. CP lieu domicile habituel)
2. **4 var. quasi-identifiantes** : département, sexe, situation familiale (vie en couple/non), âge

Clé d'identification

$clé_i$, $i \in \{1, 2, 3, 4\}$ = combinaison des modalités des variables quasi-identifiantes

3. **Var. sensibles** : GIR, axes du GIR, montant ressources, montant APA

- **Raisonnement** : on a des valeurs précises, on veut produire une approximation voisine et cohérente avec autres variables

Log-régression par intervalles (méthode résidus simulés)

Y, variable dépendante, appartenant à l'intervalle $[\log(\text{lower}_b) ; \log(\text{upper}_b)]$

L'équation s'écrit : $[\log(\text{lower}_b); \log(\text{upper}_b)] \sim X\gamma + \varepsilon$

- **Estimation par maximum de vraisemblance** : valeur prédite déterministe, $\widehat{Y}$, et écart-type des résidus noté σ

- $\widehat{Y}$: pas nécessairement dans l'intervalle, sans contrainte sur la variable estimée
- On veut : montant prédit dans intervalle $[\log(a)\ ;\ \log(b)]$
- **Ajout d'un résidu à cette prédiction** : simulé à partir de la loi normale tronquée issue de celle des résidus obtenus à l'étape précédente ($u \sim N_{[\alpha;\beta]}(0, \sigma)$, avec respectivement $\alpha = X_i\gamma - a$ et $\beta = X_i\gamma - b$). On note ce résidu simulé $residu_s$.
- **Variable floutée** : $e^{(\widehat{Y} + residu_s)}$

Variables : montant APA, ressources, âge (années révolues), date ouverture droits à APA

- **Choix des intervalles** : quantiles ou pré-déterminé (ex. montant APA : plans "saturés")
- **Choix des X_i** : sélection automatique (modifiable) de var. quanti et quali

  Préalable : calcul corrélations entre variables à flouter directement et chaque X_i potentielle

Attention

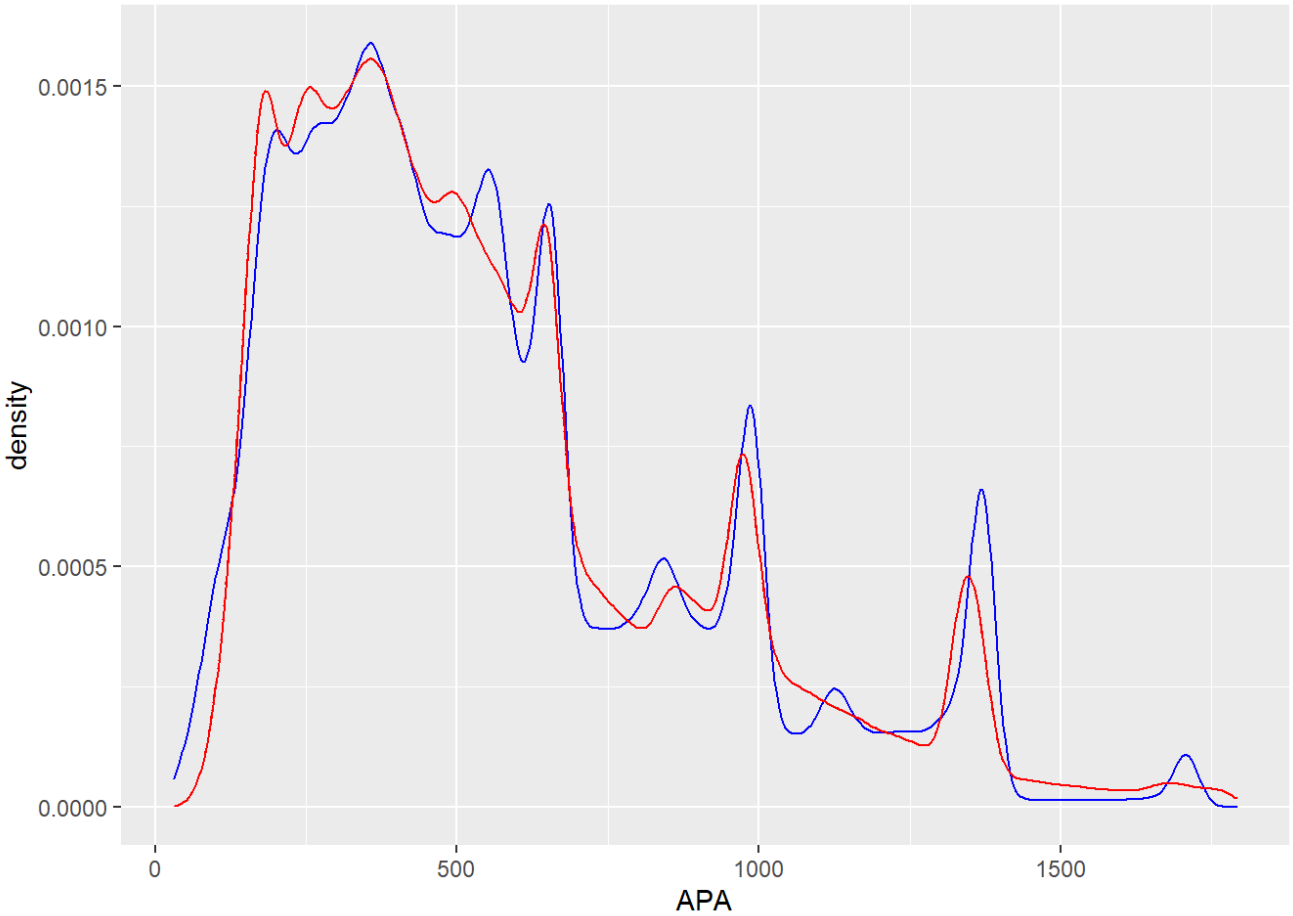
éviter colinéarité entre X_i : pas le GIR + axes du GIR

- **Autres variables liées** au montant **APA** : on conserve le lien observé
- **7 variables "floutées par règle"** : ex. montant de la participation financière du bénéficiaire : $NOTPB_F = (NOTPB/APA) \times APA_F$, avec APA_F floutée *via* log-régression par intervalles
- **Où** $APA_F \sim$ GIR, âge, etc.

4. Vérification de la cohérence

Comparaison graphique variables floutées-variables observées

Ex. : montant de l'APA ; bleu : observé / rouge : simulé



5. Vérification du niveau de confidentialité

Traitement préalable, k-anonymat/all-m

Validation de notre démarche : contrôler le niveau de confidentialité garanti par notre processus

- **Préalable** : département $<$ 2 000 : code non-signifiant affecté aléatoirement (7 départements)
- **clé d'identification** : $clé_i$, $i \in \{1, 2, 3, 4\}$ = combinaison des modalités des 4 quasi-identifiants (âge : environ 12 classes d'âge)
- **k-anonymat** : $k = 3$ pour la clé complète

5. Vérification du niveau de confidentialité

k-anonymat, k = 3

- 5 000 croisements clé complète d'identification
- 250 seulement ne respectent pas le 3-anonymat (5 % des croisements)
- 220 croisements : soit département "flou", soit situation familiale "non connue"
- environ **30** croisements problématiques : **méthode de suppression locale** (cf. Bergeat M.), c.a.d situation familiale à "non connue"

5. Vérification du niveau de confidentialité

l-diversité

- Prise en compte du GIR : variable sensible
- Protection supplémentaire en imposant que les croisements de la clé complète d'identification soient associés à au moins 2 GIR
- Donc une l-diversité $= 2$
- Sous-échantillon déterministe : individus ayant la même clé complète d'identification et possédant le même GIR exclus de la base anonymisée diffusée (seulement 140 individus)

6. Risque de ré-identification

- Risque global de ré-identification *(Scores avec sdcMICRO)*
- Nombre probable de ré-identifications exactes de bénéficiaires estimé à environ **5 000 individus** ; risque global du fichier de **0,8 %**

7. Échantillonnage et calage

- Au départ 618 000 individus : on tire un sous-échantillon aléatoire stratifié au 1/2
- **Pourquoi ?** : réduire encore plus le risque de ré-identification et *a fortiori* le risque de révélation d'attribut ou de divulgation inférentielle
- **Impératif ?** : ne pas divulguer la méthode de tirage

8. Validation par les résultats : comparaisons à la base CASD

1. **Statistiques descriptives** : comparables. Écarts faibles ($<$ 10 %)
   Écarts peu élevés en montants (10 € sur les petits montants)

2. **Régressions : aide humaine selon caractéristiques bénéficiaires**
   - Tous GIR : messages comparables
   - GIR 1 (les + dépendants) : idem, sauf pour des bénéficiaires parmi les moins aisés
   - autres GIR : parfois ampleur un peu différente pour les plus aisés

3. **Simulations Autonomix** *(modèle microsimulation)* : écarts par GIR faibles, écarts tranches ressources, mais montants en jeu comparables

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

Drees

Conclusions et prespectives

- **Des écarts** : populations fines / tranches ressources.
- **Autres méthodes de partitionnement** : clustering flou pour les intervalles ?
- **Autres méthodes perturbatrices** : éviter oscillations, mieux gérer écarts "flouté"-"réel"