Les méthodes perturbatives d'anonymisation de données individuelles

avantages et inconvénients, développements récents et exemples de mise en œuvre

Maël Buron 24/06/2019

Insee
Mesurer pour comprendre

Introduction

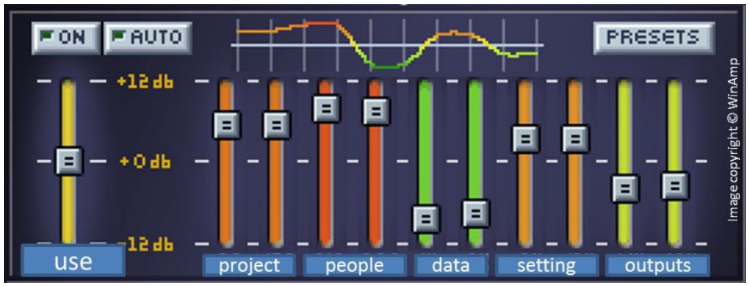
- S’assurer de la confidentialité des données diffusées, c’est nécessaire pour :
  - respecter les **obligations légales**
  - mais surtout conserver la **confiance des répondants** donc des **taux de réponses** élevés et une information fiable.
- Comment diffuser l’information la plus complète possible ?
  - compromis entre risque de ré-identification et utilité
- Peu d’expérience pour l’anonymisation de données individuelles contrairement aux données tabulées
- Différents types de fichiers pour l’accès aux données individuelles :
  - Quelques fichiers-détails « grand public » sur insee.fr, données très agrégées
  - Les Fichier de Production de Recherche (FPR), seulement pseudonymisés
  - Le Centre d’Accès Sécurisé aux Données (CASD), fichiers les plus complets

Introduction

- Mission inspection générale « Accès aux données individuelles de l'Insee » en 2013. Philippe Domergue et François Elissalt
  - Constatent que « le dispositif de l'Insee constitue un ensemble **riche et diversifié**, assurant un bon « continuum » susceptible de **répondre en principe aux besoins de publics variés**, dans le respect du cadre légal en vigueur »
  - Remarquent qu'au sein des fichiers-détails « **malgré les précautions** prises (limiter la précision de l'information, au détriment parfois assez dirimant de son intérêt), on ne saurait exclure que demeurent des **possibilités d'accéder à des données** que les enquêtés considèrent comme « **sensibles** » »
  - Recommandent « de **poursuivre la réflexion** du groupe de travail sur l'anonymisation des fichiers [...] afin de mieux définir à l'avenir la politique de l'Insee en la matière et d'être en mesure de l'expliquer »

Introduction

- « Five safes » Desai, Ritchie, Welpton, (2016)
  - Projets sûrs
  - Personnes sûres
  - Données sûres
  - Environnement sûr
  - Sorties sûres



- Pour les FPR et plus encore au CASD ces cinq leviers sont utilisés pour limiter le risque de rupture du secret lors de l'accès à des fichiers individuels non anonymisés.

- Diffuser sur internet ne laisse plus que la dimension « Données sûres »
  - et impose de limiter davantage le détail des fichiers,
    - Recodage de variable, suppressions locales, sous-échantillonnage
  - voire de les perturber en introduisant de l'incertitude.

01 Panorama des méthodes perturbatives

Anonymisation de données individuelles

- A partir d'un fichier confidentiel, on mesure le risque de ré-identification
  - Critères les plus souvent utilisés : k anonymat, l diversité
    - K anonymat : au moins k individus pour chaque combinaison de variables quasi-identifiantes (ex : variables socio démographiques)
    - L diversité : au moins l modalités différentes des variables d'intérêts pour chaque combinaison de variables quasi-identifiantes

- Si acceptable, diffusé tel quel.
- Si non, on anonymise jusqu'à ce que le risque résiduel soit acceptable.

- On mesure la perte d'information en comparant le fichier anonyme et le fichier original (erreurs quadratique, absolue, ou relative moyennes ; tables de contingences pour les variables qualitatives ; entropie)

Panorama des méthodes perturbatives

- Détails dans le document de travail 2016 de Maxime Bergeat « La gestion de la confidentialité pour les données individuelles » et le « Handbook on Statistical Disclosure Control » (Hundepool et al.)
- Anonymisation des données initiales X (matrice n,p) en Z(n,p)
  - en multipliant à gauche par A(n,n) de perturbation des individus
  - et à droite par B(p,p) de perturbations des variables,
  - et en ajoutant un bruit C(n,p)

$$Z = AXB + C$$

- Ajout de bruit (variables quantitatives)
- Post-Randomization Method (PRAM) (variables qualitatives)
- Microagrégation
- Matching k-anonyme
- Swapping
- Données synthétiques

Avantages et inconvénients

- Inconvénients :
  - Comment ça s'articule avec les **obligations légales ?** Les conventions donnent des règles claires sur le nombre d'individus minimal dans une case lorsqu'on tabule. C'est flou pour les fichiers individuels. Et quid des données perturbées ?
  - Comment **communiquer ?** Certaines méthodes perturbatives donnent l'impression que rien n'a été fait pour protéger le fichier.
- Avantages :
  - **Diminue le risque** de ré-identification : un intrus n'est plus sûr d'observer un enregistrement réel.
  - Permet de **conserver plus de détails** que les méthodes non perturbatives pour un même niveau de risque car on peut cibler les perturbations sur les individus à risque et garder le détail pour les autres.
  - Est-ce vraiment un problème d'ajouter du bruit à des données d'enquête ? Il y a déjà des incertitudes dans les aléas d'échantillonnage et du bruit même dans les registres avec par exemple les erreurs de saisies ou les redressements.

Ajout de bruit (variables quantitatives)

- Exemple bruit additif $\mathbf{Z} = \mathbf{X} + \epsilon,\ \mathbf{X} \sim \mathcal{N}(\mu, \sum),\ \epsilon \sim \mathcal{N}(\mathbf{0}, \sum_\epsilon)$

- Avec des bruits indépendants $\sum_\epsilon = \alpha \times diag(\sigma_1, ..., \sigma_p),\ \alpha > 0$
  - Pour préserver espérances et covariances

- Ou avec des bruits corrélés $\sum_\epsilon = \alpha \times \sum,\ \alpha > 0$
  - Pour préserver espérances et coefficients de corrélations linéaires

Post-Randomization Method (PRAM)

- Pour les variables qualitatives

- P matrice de perturbation PRAM. Matrice stochastique qui donne les probabilités de transition entre la modalité k pour la variable originale X et la modalité l pour la variable perturbée Z

$$\mathbf{P} = (p_{k,l})_{k,l \in [|1,K|]} = (\mathbb{P}(Z = l | X = k))_{k,l \in [|1,K|]}$$

- Attention au choix de P car risque de créer des combinaisons impossibles entre deux variables (ex : 5 ans & marié).
  - Possibilité d'échanges des combinaisons de variables
- Possibilité de construire des matrices PRAM invariantes pour conserver les distributions
- Technique utilisée pour les fichiers publics européens de l'enquête emploi

Microagrégation

- Objectif : fichier k-anonyme, c'est-à-dire que pour chaque combinaison de variables quasi-identifiantes (ex : variables sociodémographiques), on a au moins k individus.
- Formation de groupes d'individus de taille au moins k où l'on remplace la valeur des variables par la moyenne
- Minimisation de la variance intra-groupes pour limiter la déformation, sous la contrainte de la taille minimale k des groupes
- Possible sur une variable :
  - Indépendamment pour différentes variables — risque important
  - Ou sur un critère synthétique — utilité faible
- Ou multivarié
  - Mais temps de calcul important

Matching k-anonyme

- Imaginé et testé à l'Insee

- Première étape : suppressions locales
  - On supprime des valeurs des individus rares pour obtenir un fichier k-anonyme
- Deuxième étape : matching
  - On apparie avec un score de propension et on impute les valeurs supprimées
- Troisième étape : calage sur marge
  - On restaure certaines distributions en modifiant les poids

Swapping

- Echanges de variables entre deux individus
  - Introduit de l'incertitude dans le fichier
  - Possibilité de sélectionner aléatoirement les individus à échanger, ou de les cibler pour limiter la perturbation

- Exemple d'utilisation
  - données pour Eurostat du recensement 2011
    - Eurostat diffuse des tableaux avec beaucoup de dimensions
    - Certaines cases sont vides ou ont des valeurs très faibles → risque
  - variables nationalité et pays de naissance
  - avec une probabilité de l'ordre de $1/n^2$ où n est le nombre d'individus pondérés pour un couple (nationalité, pays de naissance) donné

Données synthétiques

- Les données anonymisées sont construites à partir d'un **modèle de simulation** construit sur les données originales.

- Complètement synthétique : toutes les variables sont simulées, pour tous les individus
- Partiellement synthétique : la simulation ne concerne que certains individus ou certaines variables.
- Hybrides : les variables diffusées sont construites comme une moyenne entre les données originales et les données simulées.

- Technique utilisée pour les fichiers publics européens de l'enquête ressources et conditions de vie

02 Fichiers publics européens

Center of Excellence on Statistical Disclosure Control

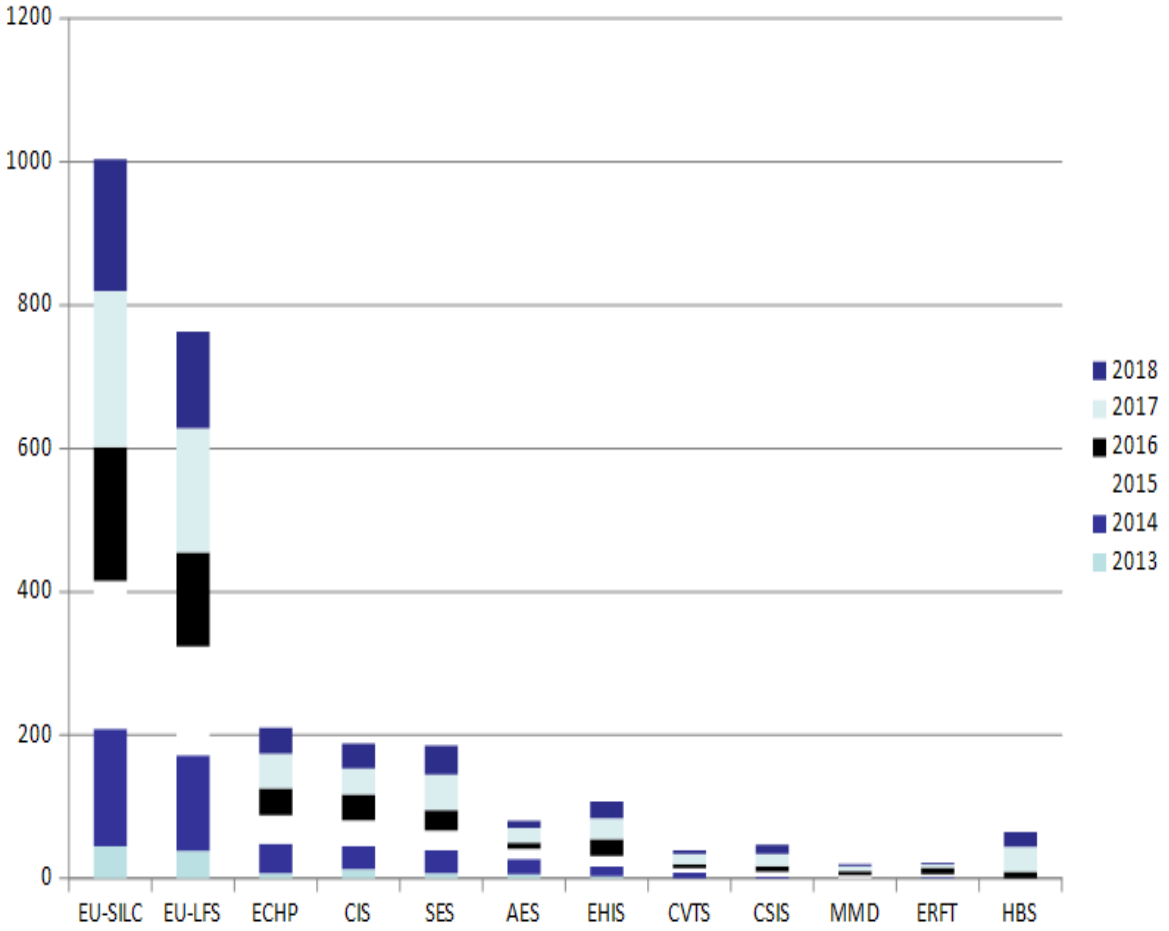
- Créé en octobre 2014 pour 4 ans. 7 pays participants :
  - Pays-Bas, Autriche, Allemagne, Finlande, France, Hongrie et Slovénie

- Objectifs : méthodes, outils, recommandations, support, formation

- 4 projets :
  - Public use files for Eurostat microdata (2015)
  - User support for and maintenance of SDC tools (04/2016 → 12/2018)
  - Harmonised protection of Census data in the ESS (09/2016 → 08/2017)
  - Open source tools for perturbative confidentiality methods (03/2018 → 09/2019)

- https://ec.europa.eu/eurostat/cros/content/centre-excellence-statistical-disclosure-control-0_en

Nombre de demandes par années des FPR européens



EU-SILC : European Union Statistics on Income and Living Conditions

EU-LFS : European Union Labour Force Survey

ECHP : European Community Household Panel

CIS : Community Innovation Survey

SES : Structure of Earnings Survey

AES : Adult Education Survey

EHIS : European Health Interview Survey

CVTS : Continuing Vocational Training Survey

CSIS : Community Statistics on Information Society

MMD : Micro-Moments Dataset

ERFT : European Road Freight Transport Survey

HBS : Household Budget Survey

Fichiers publics européens

- Objectif :
  - définition de bonnes pratiques
  - Fichiers pour l'enquête emploi (LFS : Labour Force Survey) et l'enquête sur les revenus et les conditions de vie (SILC : Statistics on Income and Living Conditions)

- Permettre aux chercheurs d'accéder à des fichiers qui ressemblent aux FPR avant de faire la demande qui prend du temps

- Permettre aux étudiants d'avoir des fichiers de données individuelles issus de la statistique publique

Anonymisation de l'enquête Emploi

- Objectifs de réduction de risque basés sur le k-anonymat
  - 13 variables jugées quasi-identifiantes
  - Objectif A : fichier 5-anonyme pour 7 (sur 13) variables quasi-identifiantes
  - Objectif B : au moins 10 individus pour tous les tableaux croisant 4 variables quasi-identifiantes sur 13
- Méthode choisie
  - Objectif A par agrégation et suppression pour ces 7 principaux quasi-identifiants
  - Puis perturbation PRAM pour les 6 autres
- Logiciel mu-Argus
- Utilité : comparaison pour des indicateurs standards (ex : taux de chômage, nombre de personnes en emploi) avant et après anonymisation
  - Résultat globalement satisfaisants
  - avec des écarts plus marqués pour les variables ayant été perturbées

Anonymisation de l'enquête ressources et conditions de vie

- Génération de données complètement synthétiques (simulation de toutes les variables pour tous les individus)
  - À partir d'un modèle de simulation estimé sur les données réelles utilisant les poids de sondage
  - Réplication bootstrap de la structure ménage/individu par âge et sexe des occupants du ménage
  - Pour les variables principales : simulation des variables en fonction des probabilités prédites par des modèles de régression logistique
  - Les variables continues sont discrétisées avant simulation (pour utiliser les modèles de régression logistique) puis rendues continues à nouveau par le mécanisme inverse
  - Pour les autres variables : simulation selon des méthodes basiques utilisant le quantile de revenu prédit

Anonymisation de l'enquête ressources et conditions de vie

- Mesure de l'utilité en comparant pour certains indicateurs classiques les résultats obtenus à partir des données originales et à partir des données synthétiques
  - Résultats discutables, particulièrement pour les variables continues comme les variables de revenu
  - Beaucoup plus de pauvres dans les données simulées

- Difficile de prouver que le risque de divulgation est nul dans le fichier synthétique, mais on peut s'y attendre si le modèle de simulation ne colle pas trop aux données

03 Recensement européen 2021

Harmonised protection of Census data in the ESS

- Pourquoi ce projet
  - 2011 hypercubes → méthodes différentes → comparaisons difficiles
  - 2021 ajout des données carroyées → différenciation ?

- Census Hub https://ec.europa.eu/CensusHub2 - Données de 32 pays

- Comment
  - Questionnaire
  - Tests et recommandations

Questionnaire

- Deux tiers des pays considèrent qu'il y a des variables sensibles
  - Et certains que toutes les variables sont sensibles
  - Alors qu'une poignée considère que rien n'est sensible
  - Le plus souvent : pays de naissance et nationalité

- La majorité des pays ont appliqué des méthodes spécifiques pour la confidentialité de leur Census 2011. Souvent : suppression

- Beaucoup se déclarent intéressés pour tester de nouveaux outils pour 2021
  - Suppressions avec Tau-Argus le plus cité
  - Peu de pays testeront autre chose si on ne leur propose pas d'autres outils
  - La moitié pourrait changer leur méthode suivant les recommandations

Identification des meilleures pratiques

- Aspects importants :
  - 1 - conserver la structure des hypercubes
  - 2 - prendre en charge le risque de divulgation d'attributs
  - 3 - minimiser la perte d'informations
  - 4 - être applicable par de nombreux États membres
- Les méthodes perturbatives semblent être les plus adaptées :
  - Recodage global - impossible
  - Suppression de cellules - difficile de manière cohérente entre les pays en raison des différences entre les règles
  - Et comment gérer le risque de différenciation entre hypercubes et données carroyées ?
- La méthode harmonisée proposée doit être :
  - flexible pour répondre à des besoins spécifiques
  - et adaptable simplement en changeant les paramètres

Tests et recommandations

- Le Royaume-Uni a fourni les codes SAS de son recensement de 2011
  - permutations ciblées de ménages *Targeted record swapping*
  - et bruit aléatoire C*ell-key method* basé sur les travaux de Australie

- On a adapté et testé les programmes SAS
  - Et conduit une analyse d'utilité / risque avec des mesures spécifiques de perte d'information.

- Ensuite, d'autres États membres ont été invités à tester les méthodes sur leurs données.

- Recommandation : combinaison de ces deux méthodes perturbatives

Targeted Record Swapping

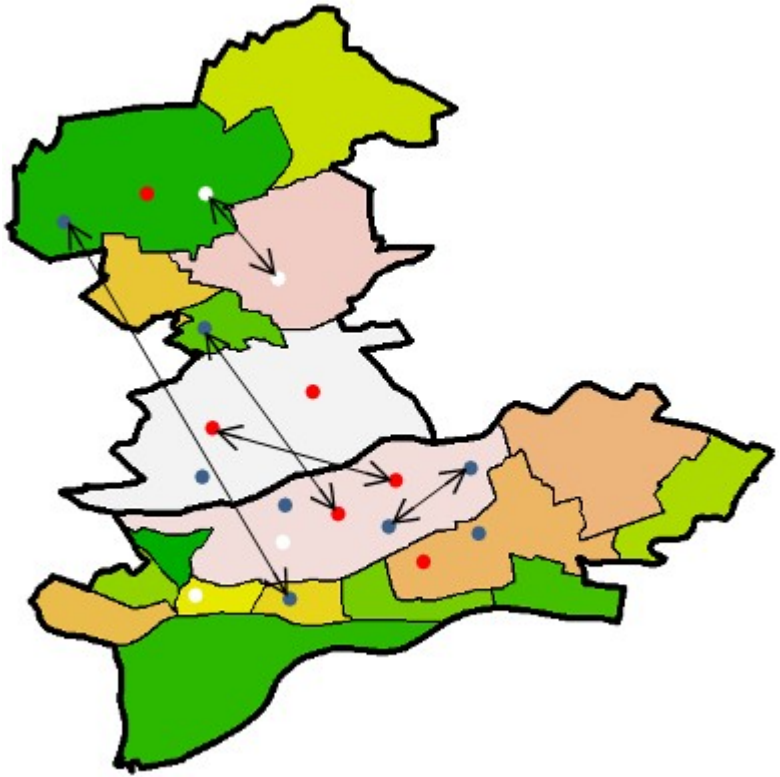
- Méthode avant tabulation

- Paramètres :
  - Choix des variables qui définissent le risque
  - Choix de variables géographiques
  - Calcul du risque pour les ménages à chaque niveau géographique
  - Choix des variables pour identifier les ménages qui se ressemblent
  - Choix d'un taux de swapping minimum

- 4 étapes : Targeting , sampling , matching , swapping.

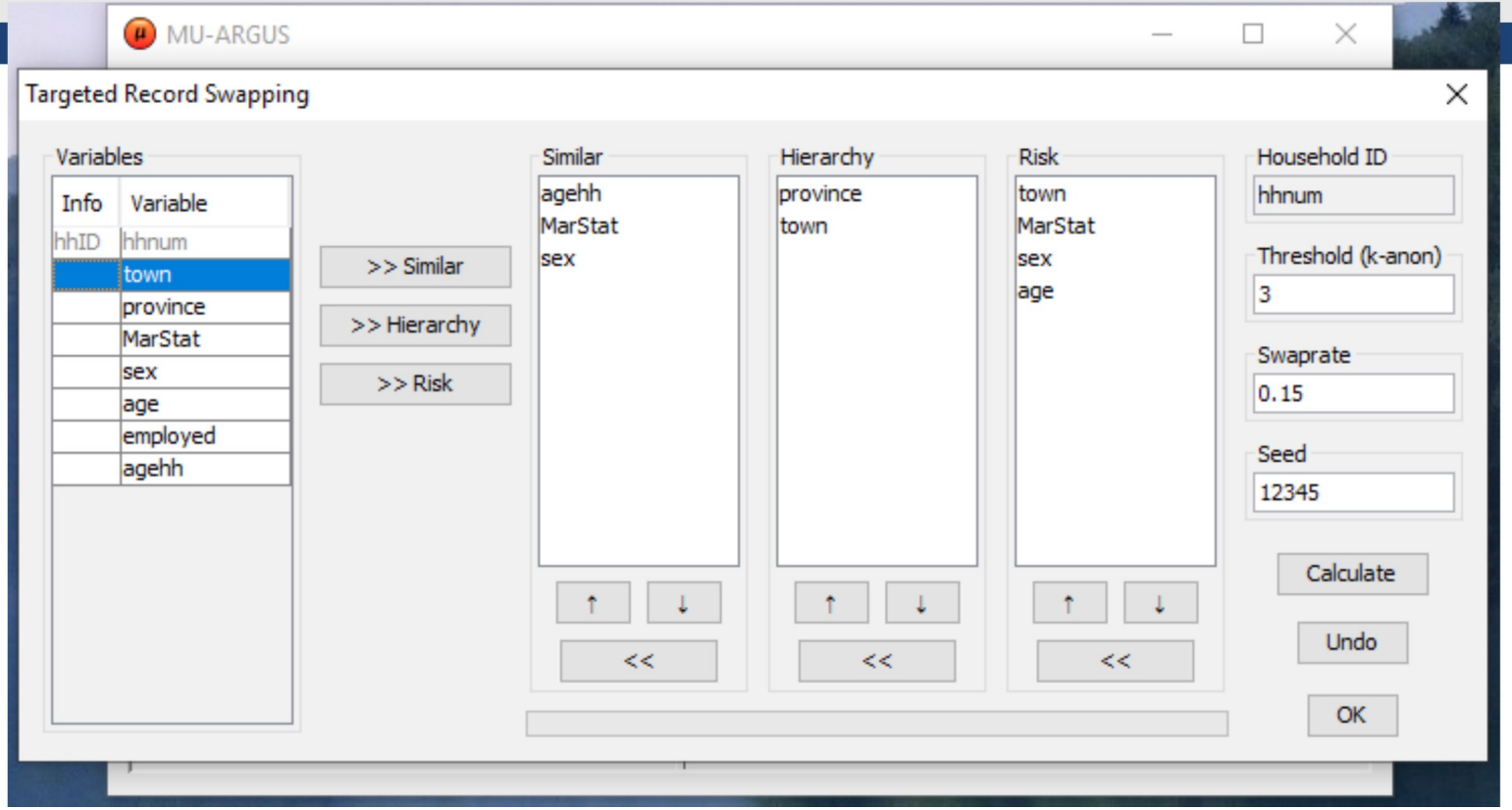
Targeted Record Swapping

- Par itération, du niveau géographique le plus agrégé au niveau le plus fin.
- On échange des ménages à risque avec d'autres qui leur ressemblent mais qui sont dans une zone différente.
- Puis on continu en passant à un niveau géographique plus fin.
- Exemple avec deux niveau géographiques :



Implémentation *Targeted Record Swapping*

- Code en C++
- Appel via µ-Argus :



- ou via R package `recordSwapping`

```
swapdata <- recordSwap(data,similar,hierarchy,risk,hid,th,swaprate,seed)
```

- Disponible sur https://github.com/sdcTools/protoTestCensus/

Références

- Bergeat, M. (2016). La gestion de la confidentialité pour les données individuelles. http://www.epsilon.insee.fr/jspui/bitstream/1/59179/1/m1607.pdf
- Desai, T., Ritchie, F., & Welpton, R. (2016). Five Safes: Designing data access for research. http://eprints.uwe.ac.uk/28124/1/1601.pdf
- Domergue, P., Elissalt, F. (2013). Accès aux données individuelles de l'Insee https://www.agora.insee.fr/files/live/sites/dg-ig/files/shared/audits/rapports/2013/2013-1.7.34_Acces%20aux%20donnees%20individuelles.pdf
- de Wolf, P. P. (2015). Public use files of EU-SILC and EU-LFS data. Joint UNECE-Eurostat work session on statistical data confidentiality, Helsinki, Finland. http://www.ksh.hu/statszemle_archive/2015/2015_11-12/2015_11-12_1140_2.pdf
- Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Longhurst, J., ... & Wolf, P. (2010). Handbook on statistical disclosure control. ESSnet on Statistical Disclosure Control. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.135.3606&rep=rep1