L'IMPACT REVU DES VALEURS MANQUANTES SUR LA MESURE DU RISQUE D'IDENTIFICATION BASÉE SUR LA RÈGLE DU *K*-ANONYMAT : PRÉSENTATION D'OPTA ET DE PESA

Ali Saoud

Haut-commissariat au Plan[1], Direction de la Statistique,
Rue Mohamed Belhassan El Ouazzani, Haut Agdal, 10001 Rabat, Maroc
a.saoud@hcp.ma

Résumé. Les microdonnées statistiques offrent un potentiel analytique considérable tout en regorgeant d'informations individuelles relevant de la vie privée des gens. Leur diffusion tant souhaitée et briguée nécessiterait l'occultation prudente de leur identité et de leurs attributs distinguables. Tel est le principe d'anonymisation des microdonnées qui vise à considérablement minimiser le risque d'identification.
Autant que nous sachions, la règle du *k*-anonymat, sur laquelle se base l'une des mesures du risque d'identification, s'applique actuellement en intégrant les valeurs manquantes selon une approche que nous avons jugée trop peu restrictive à notre goût.
La présente contribution introduit deux nouvelles alternatives pour l'estimation des valeurs manquantes, OPTA et PESA, toutes deux plus restrictives que l'approche en vigueur et pouvant convenir aux instituts statistiques publics, chacun selon l'ampleur de son besoin d'anonymisation des microdonnées qu'il produit.

Mots-clés. Approche optimiste (OPTA), approche pessimiste (PESA), confidentialité, contrôle de la divulgation statistique, anonymisation, microdonnées, données individuelles, risque d'identification, *k*-anonymat, valeurs manquantes.

Abstract. Statistical microdata has substantial potential for analysis while containing an overwhelming amount of individual-level information regarding people's privacy. The highly desired and pursued dissemination of microdata would require cautious concealment of individuals' identity and distinguishable attributes. This is the principle of statistical disclosure control for microdata which aims to greatly minimize disclosure risk.
As far as we know, the *k*-anonymity criterion, on which one of disclosure risk measures is based, is currently resorted to according to an approach which includes missing values in a non-restrictive enough way in our opinion.
This paper introduces two new alternatives for estimating missing values, OPTA and PESA, which are both more restrictive than the current approach and may be appropriate for National Statistical Institutes (NSI), depending on the extent of each one's need for statistical disclosure control for microdata it produces.

Keywords. Optimistic approach (OPTA), pessimistic approach (PESA), confidentiality, statistical disclosure control, anonymization, microdata, disclosure risk, *k*-anonymity, missing values.

[1] Les points de vue exprimés dans le présent article relèvent de la seule responsabilité de l'auteur et ne reflètent pas l'opinion ni la position du haut-commissariat au Plan du Maroc.

1. Introduction

En pleine ère digitale, l'humanité est en train d'assister à une révolution profonde de l'information et de la connaissance. L'essor vertigineux et continu des techniques numériques entraîne la circulation d'un flux exponentiellement grandissant de données sur l'internet. Celles-ci sont pour la plupart des informations à caractère individuel et peuvent éventuellement porter préjudice à la vie privée et à la liberté de tout un chacun.

Toujours est-il que le partage de données pour des desseins et des causes nobles et bénéfiques pour l'humanité doit être soutenu et encouragé. Les instituts statistiques publics disposent de mines d'informations mais ils ne peuvent accomplir le travail titanesque consistant à les exploiter et à les analyser dans l'intégralité des dimensions concevables et imaginables, même s'ils sont les entités les plus à même de faire cela, tant à l'échelle nationale qu'internationale.

C'est sur cette base que des initiatives de démocratisation des données ont vu le jour aux quatre coins du globe. Cependant, les producteurs et détenteurs de données sont incessamment tiraillés entre l'obligation de protéger la vie privée des individus et la volonté de dynamiser la recherche scientifique en mettant à disposition des utilisateurs potentiels des informations de plus en plus détaillées et pertinentes.

Face à ce dilemme, l'intérêt et le besoin d'anonymisation des microdonnées ont émergés afin que les instituts statistiques publics soient en mesure de diffuser des données le moins attentatoire possible à l'intimité des personnes tout en étant utiles en termes d'analyse autant que cela le permet.

Le processus d'anonymisation des microdonnées doit ainsi être assujetti à des règles d'éthique correspondant à un contexte propre aux personnes concernées. Il est ainsi crucial de religieusement définir quelles données les concernant revêtent un caractère sensible, en d'autres termes, quelles données un individu ne voudrait pas qu'autrui sache à propos de lui ou de son ménage. Cela peut être élargi pour également incorporer les données privées des entreprises.

Une personne peut être reconnue à partir des données de diverses manières, dépendant du type d'informations que détient l'utilisateur mal intentionné désireux d'opérer des identifications d'autrui, que nous appelons communément des scénarios de divulgation statistique. Généralement, les deux scénarios considérés sont la divulgation par appariement –se produisant en combinant plusieurs sources de données– et la divulgation par reconnaissance spontanée –survenant lorsque l'utilisateur mal intentionné connaît l'individu qu'il recherche (voisin, collègue, ami, parent, personnage public, etc.) et a des informations assez détaillées le concernant.

Afin de s'adonner adéquatement à l'anonymisation des microdonnées, il est nécessaire de modéliser l'éventualité qu'un quelconque utilisateur puisse retrouver tel ou tel individu, que nous appelons le risque d'identification. Parmi les différentes mesures dudit risque, nous allons nous intéresser à celle fondée sur la règle du *k*-anonymat.

Après avoir mené de fines recherches sur cette mesure du risque, l'approche que nous avons trouvée sur l'évaluation des valeurs manquantes –usuellement présentes dans les microdonnées statistiques– ne nous a pas fortement convaincus. C'est alors que nous proposons, à travers le présent article, une révision de l'influence qu'ont les valeurs manquantes sur l'estimation du risque d'identification basée sur la règle du *k*-anonymat. Mais d'abord, nous allons nous livrer à l'énonciation de définitions essentielles propres à l'anonymisation des microdonnées, pour ensuite présenter OPTA et PESA, les deux nouvelles approches que nous suggérons pour mieux appréhender l'intégration des valeurs manquantes au calcul du risque d'identification selon la règle du *k*-anonymat, puis enfin conclure en exposant brièvement les étapes à suivre pour mener à terme un processus d'anonymisation des microdonnées.

2. Définitions préliminaires

Dans le domaine de la statistique, nous entendons par un fichier de microdonnées un recueil de variables fournissant des informations individuelles –à l'opposé des données agrégées– sur les unités statistiques d'une population donnée (personnes, ménages, ou entreprises) ayant attrait à divers thèmes tels que la démographie, la santé, l'éducation, les conditions d'habitat, l'activité économique, etc.

Parmi les variables contenues dans un fichier de microdonnées, certaines permettent d'identifier un ensemble d'unités statistiques, voire leur totalité. Celles-ci peuvent être classées en deux groupes :

- **Les identifiants directs :** les variables qui permettent une identification claire et aisée, telles que le nom complet, l'adresse exacte, le numéro de téléphone, le numéro d'identification national, le code de l'entreprise, etc. La suppression de ces variables est intuitive et certes nécessaire mais pas suffisante. Nous ignorerons les identifiants directs dans le présent document, en les considérant systématiquement supprimés.
- **Les identifiants indirects :** les variables qui permettent une identification en combinant les informations qu'elles contiennent, telles que le sexe, la nationalité, la date de naissance, le revenu, etc. Ils peuvent correspondre à des variables catégorielles comme à des variables continues. Nous supposerons dans ce qui suit que tous les identifiants indirects sont des variables catégorielles –de nature ou par restructuration– pour les besoins de la mesure du risque d'identification objet d'étude.

Nous appellerons **clé d'identification** $c(i)$ d'une unité statistique u_i la combinaison de ses valeurs quant au *n*-uplet d'identifiants indirects $(V_p)_{p\in[\![1,\dots,n]\!]}$ présents dans un fichier de microdonnées de taille N. Ainsi, pour toute unité statistique u_i, la clé d'identification $c(i)$ est définie comme tel :

$$\forall i \in [\![1,\dots,N]\!], c(i) = (V_1(i),\dots,V_n(i)) \quad (1)$$

L'objectif d'un projet d'anonymisation d'un fichier de microdonnées est de faire en sorte qu'aucune unité statistique ne soit reconnaissable à travers sa clé d'identification, de façon plus ou moins certaine. Pour ce faire, il est nécessaire de définir des mesures permettant de quantifier la possibilité qu'un utilisateur mal intentionné réussisse à retrouver une ou plusieurs unités statistiques, qu'il faudra minimiser de manière drastique pour pouvoir prétendre qu'un fichier de microdonnées a été anonymisé.

À cet effet, plusieurs mesures du risque d'identification ont été élaborées. Comme l'affirment Samarati et Sweeney (1998a, [10], cité dans Ciglic *et al.*, 2014, [3]), l'un des concepts les plus répandus est la règle du *k*-anonymat. C'est à cette mesure particulière du risque que portera notre intérêt.

Selon la définition de Samarati et Sweeney (1998b, [11]), un fichier de microdonnées est conforme à la **règle du *k*-anonymat** si toute combinaison de valeurs des identifiants indirects peut indistinctement correspondre à au moins k unités statistiques.

En d'autres termes, le *k*-anonymat est vérifié pour un fichier de microdonnées si la fréquence $f(c(i))$ de chaque clé d'identification $c(i)$ est au moins égale à k, soit :

$$\forall i \in [\![1,\dots,N]\!], f(c(i)) \geq k \quad (2)$$

Or, cette définition ne mentionne pas les valeurs manquantes que contiennent éventuellement les identifiants indirects et qui peuvent perturber l'observation des fréquences $f(c(i))$ et aussi étendre l'ensemble des clés d'identification potentielles.

3. Valeurs manquantes et *k*-anonymat

Les fichiers de microdonnées comportent généralement des valeurs manquantes au niveau d'une ou de plusieurs variables. Celles-ci peuvent être avoir diverses causes selon lesquelles nous pouvons distinguer deux catégories de valeurs manquantes que nous appellerons :

- **« Sans objet » (N/A) :** valeur que prend une variable lorsque la personne en question n'est pas concernée par la question (p. ex. la profession manquante d'une personne au foyer).
- **« Non déterminée » (ND) :** valeur que prend une variable lorsque l'information afférente est manquante pour cause de non-réponse ou de codification erronée.

Pour les besoins du calcul du risque selon la règle du *k*-anonymat, les valeurs manquantes de la première catégorie peuvent être codifiées par une(des) nouvelle(s) modalité(s) (p. ex. codifier la profession des personnes au foyer par la modalité « Personne au foyer ») pour ainsi les différencier des valeurs manquantes de la seconde catégorie. Le présent document s'intéressera alors au traitement des seules valeurs « non déterminées ».

3.1 Deux nouvelles approches ajustées pour les valeurs « non déterminées »

Benschop et ses collaborateurs (2017, [1]) préconisent une approche pour intégrer les valeurs « non déterminées » au calcul des fréquences des clés d'identification –que nous appellerons dans ce qui suit l'approche orthodoxe. Celle-ci est implémentée dans le package *sdcMicro*[2] sous *R* et se base sur le fait qu'une valeur « non déterminée » peut prendre n'importe quelle valeur de la variable en question. Ainsi, selon cette approche, deux clés d'identification $c(i) = (V_1(i),...,V_n(i))$ et $c(j) = (V_1(j),...,V_n(j))$ correspondant à deux unités statistiques distinctes u_i et u_j $(i \neq j)$ sont concordantes si et seulement si :

$$\begin{cases} \forall (i,p) \in [\![1,...,N]\!] \times [\![1,...,n]\!], V_p(i) \in M_p \cup \{ND\} \\ \forall (i,j,p) \in [\![1,...,N]\!]^2 \times [\![1,...,n]\!], V_p(i) = V_p(j) \end{cases} \quad (3)$$

Où M_p est l'ensemble des valeurs valides de l'identifiant indirect V_p pour tout $p \in [\![1,...,n]\!]$.

Pour expliciter ce concept, considérons un fichier de microdonnées FM_1 comportant $N = 7$ personnes et un ensemble de variables dont $n = 2$ composent la clé d'identification :

- le « Sexe » : ayant pour modalités « Féminin » et « Masculin » et comprenant 3 valeurs « non déterminées » ;
- la « Nationalité » : ayant pour modalités « Marocaine » et « Étrangère » et comprenant 2 valeurs « non déterminées ».

Le calcul des fréquences des clés d'identification $f(c(i))$ pour chaque personne u_i figurant dans FM_1, selon l'approche orthodoxe, est illustré dans le tableau ci-après.

Indice i de u_i	**Sexe**	**Nationalité**	$f(c(i))$
1	Masculin	Marocaine	4
2	Féminin	Marocaine	4
3	Masculin	ND	5
4	Féminin	ND	5
5	ND	Étrangère	3
6	ND	Marocaine	6
7	ND	Marocaine	6

Tableau 1 : Le premier exemple de fichier de microdonnées FM_1 et les fréquences des clés d'identification des personnes le composant selon l'approche orthodoxe

[2] Un outil d'anonymisation développé en 2007 par Templ (2008, [12]) et continuellement amélioré. C'est l'un des outils les plus utilisés par les instituts statistiques européens notamment, avec *μ-Argus*, comme l'affirme Bergeat (2016, [2]).

Vu que la valeur minimale des fréquences de clés d'identification $f(c(i))$ est égale à 3, le fichier de microdonnées FM_1 peut être considéré 3-anonyme selon la formule (2). Cependant, si un utilisateur mal intentionné veut retrouver une femme étrangère, il n'y a que 2 personnes qui le sont potentiellement (u_4 et u_5), et de façon analogue s'il veut identifier un homme étranger (u_3 et u_5), soit :

$$f(Féminin; Étrangère) = f(Masculin; Étrangère) = 2 \quad (4)$$

Ainsi, en remplaçant à chaque fois le « Sexe » de u_5 par « Féminin » (Cas n^o 1) puis par « Masculin » (Cas n^o 2), nous obtenons les fréquences $f(c(i))$ indiquées dans le tableau qui suit.

	Cas n^o 1			**Cas n^o 2**		
Indice i de u_i	**Sexe**	**Nationalité**	$f(c(i))$	**Sexe**	**Nationalité**	$f(c(i))$
1	Masculin	Marocaine	4	Masculin	Marocaine	4
2	Féminin	Marocaine	4	Féminin	Marocaine	4
3	Masculin	ND	4	Masculin	ND	5
4	Féminin	ND	5	Féminin	ND	4
5	**Féminin**	Étrangère	**2**	**Masculin**	Étrangère	**2**
6	ND	Marocaine	6	ND	Marocaine	6
7	ND	Marocaine	6	ND	Marocaine	6

Tableau 2 : La simulation de FM_1 selon deux cas consistant à remplacer la valeur manquante du « Sexe » de u_5 par les valeurs possibles

Nous constatons que quel que soit le « Sexe » de u_5, sa fréquence $f(c(5))$ est égale à 2. Il est donc plus adéquat d'affirmer que :

$$f(ND; Étrangère) = f(c(5)) = 2 \quad (5)$$

Par conséquent, le fichier de microdonnées FM_1 ne peut plus être considéré 3-anonyme, mais plutôt 2-anonyme seulement.

Afin de remédier à cette sous-estimation du risque d'identification, il est préférable d'observer les fréquences des clés d'identification avec davantage de vigilance. À cet effet, nous définissons deux approches qui reposent sur la simulation du fichier de microdonnées en remplaçant les valeurs « non déterminées » par toutes les valeurs qu'elles peuvent éventuellement prendre :

- **Approche optimiste (OPTA) :** basée sur le fait qu'une clé contenant une(des) valeur(s) « non déterminée(s) » correspond à la plus fréquente des clés compatibles.
- **Approche pessimiste (PESA) :** basée sur le fait qu'une clé contenant une(des) valeur(s) « non déterminée(s) » correspond à la plus rare des clés compatibles.

En pratique, cela consiste à définir l'ensemble C de toutes les clés d'identification possibles c_q, à lister les personnes qui leur correspondent puis à calculer la fréquence potentielle de chaque clé c_q que nous noterons $f(c_q)$. Le tableau suivant met en exergue ce processus appliqué au fichier de microdonnées FM_1.

Indice q de c_q	**Sexe**	**Nationalité**	**Indices i des u_i correspondants**	$f(c_q)$
1	Féminin	Marocaine	2 ; 4 ; 6 ; 7	4
2	Masculin	Marocaine	1 ; 3 ; 6 ; 7	4
3	Féminin	Étrangère	3 ; 5	2
4	Masculin	Étrangère	3 ; 5	2

Tableau 3 : Les clés d'identification possibles au niveau de FM_1

Ensuite, il faut définir chaque ensemble C_i des clés d'identification qui correspondent potentiellement à la personne u_i comme tel :

$$\forall i \in [\![1,...,N]\!], C_i = \left\{ c_q = \left(V_{1q},...,V_{nq}\right) \mid \forall p \in [\![1,...,n]\!], V_p(i) \in \left\{V_{pq}; ND\right\} \right\} \quad (6)$$

Puis il faut calculer les fréquences des clés d'identification selon OPTA et PESA, que nous noterons respectivement $f_{OPTA}(c(i))$ et $f_{PESA}(c(i))$, selon les formules suivantes :

$$f_{OPTA}(c(i)) = \max_{c_q \in C_i} f(c_q) \quad (7)$$

$$f_{PESA}(c(i)) = \min_{c_q \in C_i} f(c_q) \quad (8)$$

Le tableau ci-après présente le résultat de ces calculs pour le fichier de microdonnées FM_1.

Indice i de u_i	**Sexe**	**Nationalité**	C_i	$f_{OPTA}(c(i))$	$f_{PESA}(c(i))$	$f(c(i))$
1	Masculin	Marocaine	$\{c_2\}$	4	4	4
2	Féminin	Marocaine	$\{c_1\}$	4	4	4
3	Masculin	ND	$\{c_2, c_4\}$	4	2	5
4	Féminin	ND	$\{c_1, c_3\}$	4	2	5
5	ND	Étrangère	$\{c_3, c_4\}$	2	2	3
6	ND	Marocaine	$\{c_1, c_2\}$	4	4	6
7	ND	Marocaine	$\{c_1, c_2\}$	4	4	6

Tableau 4 : Les fréquences des clés d'identification des personnes figurant dans FM_1 selon OPTA, PESA et l'approche orthodoxe

Nous constatons qu'OPTA et PESA permettent de mieux appréhender le problème soulevé au niveau du tableau 2 par rapport à l'approche orthodoxe. Le fichier de microdonnées FM_1 est donc 2-anonyme seulement, selon les deux approches proposées, alors qu'il était considéré 3-anonyme selon l'approche orthodoxe.

Par ailleurs, il est important de garder à l'esprit que les clés d'identification possibles ne sont pas nécessairement toutes explicitées dans un fichier de microdonnées. En faire abstraction porterait préjudice à l'assurance d'anonymat des unités statistiques.

3.2 Intégration des clés d'identification non explicites

Rappelons qu'une valeur « non déterminée » peut prendre n'importe quelle valeur valide, même celles qui ne sont pas représentées dans un fichier de microdonnées –brut ou prétendument anonymisé. La prise en compte des modalités non explicitées d'identifiants indirects pourrait chambouler le calcul des fréquences des clés d'identification. Elles requièrent ainsi une attention

particulière, d'autant plus qu'elles correspondent quasi-systématiquement à des sous-populations rares, et « sous-populations rares » rime avec « clés d'identification rares » et donc à risque.

Par conséquent, l'énumération de toutes les modalités des identifiants indirects –qu'elles soient représentées ou pas– revêt une importance capitale pour une évaluation plus réaliste du risque d'identification. Seule PESA en tient compte ; l'approches orthodoxe et OPTA en font fi.

Le nombre de clés d'identifications dépend directement du nombre de modalités de chaque identifiant indirect. Ainsi, en reprenant les notations utilisées précédemment et en supposant que toutes les combinaisons de modalités permettent de construire des clés d'identification cohérentes, nous avons que :

$$card(C) = \prod_{p=1}^{n} card(M_p) \tag{9}$$

Pour mettre cela en évidence, reprenons l'exemple du fichier de microdonnées FM_1 et considérons que la « Nationalité » est désormais une variable à trois modalités (Marocaine, Étrangère, Apatride). Le nombre d'éléments de l'ensemble C des clés d'identification possibles c_q devient donc $card(C) = 2 \times 3 = 6$, tel que le montre le tableau qui suit.

Indice q de c_q	**Sexe**	**Nationalité**	**Indices i des u_i correspondants**	$f(c_q)$
1	Féminin	Marocaine	2 ; 4 ; 6 ; 7	4
2	Masculin	Marocaine	1 ; 3 ; 6 ; 7	4
3	Féminin	Étrangère	3 ; 5	2
4	Masculin	Étrangère	3 ; 5	2
5	Féminin	Apatride	4	1
6	Masculin	Apatride	3	1

Tableau 5 : Les clés d'identification possibles au niveau de FM_1 suite à l'élargissement de l'ensemble des modalités de la « Nationalité »

Les personnes correspondant à c_5 et à c_6 en sont alors impactées puisque la fréquence de leur clé d'identification selon PESA $f_{PESA}(c(i))$ doit être revue à la baisse comme suit.

Indice i de u_i	**Sexe**	**Nationalité**	C_i	$f_{OPTA}(c(i))$	$f_{PESA}(c(i))$	$f(c(i))$
1	Masculin	Marocaine	$\{c_2\}$	4	4	4
2	Féminin	Marocaine	$\{c_1\}$	4	4	4
3	Masculin	ND	$\{c_2, c_4, c_6\}$	4	**1**	5
4	Féminin	ND	$\{c_1, c_3, c_5\}$	4	**1**	5
5	ND	Étrangère	$\{c_3, c_4\}$	2	2	3
6	ND	Marocaine	$\{c_1, c_2\}$	4	4	6
7	ND	Marocaine	$\{c_1, c_2\}$	4	4	6

Tableau 6 : Les fréquences des clés d'identification des unités de FM_1 selon OPTA, PESA et l'approche orthodoxe suite à l'élargissement de l'ensemble des modalités de la « Nationalité »

PESA prouve ainsi tout son potentiel en ne considérant pas le fichier de microdonnées FM_1 comme 2-anonyme. En effet, si un utilisateur mal intentionné essaie d'identifier un homme apatride, il peut être certain qu'il s'agit de u_3 et ainsi être en mesure de découvrir une information contenue dans une

variable à caractère sensible –que nous avons volontairement omis dans un souci de simplification de l'exemple étudié.

De ce fait, force est de rigoureusement définir l'ensemble C de toutes les clés d'identification possibles c_q afin de mettre en œuvre PESA dans les règles de l'art. Comme nous l'avons mentionné lors de la définition de la formule (9), il faut uniquement tenir compte des clés d'identification cohérentes et ainsi faire fi de celles qui sont incohérentes pour ne pas surestimer le risque d'identification.

Une généralisation de la formule (9) qui prend en considération ce souci de cohérence et de compatibilité des modalités des identifiants indirects serait :

$$card(C) = \left(\prod_{p=1}^{n} card(M_p) \right) - card(C_{incoh}) \qquad (10)$$

Où C_{incoh} est l'ensemble des clés d'identification qui comportent des valeurs d'identifiants indirects incompatibles entre elles, soit :

$$C_{incoh} = \left\{ c_q = (V_{1q}, ..., V_{nq}) \mid \exists t \in [\![2, ..., n]\!], \exists (p_1, ..., p_t) \in [\![1, ..., n]\!]^t, \mathrm{P}\left(\bigcap_{s=1}^{t} V_{p_s} = V_{p_s q} \right) = 0 \right\} \qquad (11)$$

Afin d'éclairer ce concept, considérons un fichier de microdonnées FM_2, explicité au niveau du tableau 7, comportant $N = 7$ personnes et un ensemble de variables dont $n = 2$ composent la clé d'identification :

- le « Type d'activité » : ayant pour modalités « Actif » et « Inactif » et comprenant une valeur « non déterminée » ;
- le « Statut d'emploi » : ayant pour modalités « Salarié », « Employeur », « Indépendant », « Aide familial », « Inactif » et comprenant deux valeurs « non déterminées ».

Indice i de u_i	**Type d'activité**	**Statut d'emploi**
1	Inactif	Inactif
2	Actif	Salarié
3	Actif	ND
4	Inactif	Inactif
5	Actif	Employeur
6	ND	ND
7	Actif	Indépendant

Tableau 7 : Le second exemple de fichier de microdonnées FM_2

Abstraction faite de la cohérence des modalités des identifiants indirects, le nombre d'éléments de l'ensemble C des clés d'identification possibles c_q serait $card(C) = 2 \times 5 = 10$, comme indiqué au niveau du tableau suivant.

Indice q de c_q	**Type d'activité**	**Statut d'emploi**	**Indices i des u_i correspondants**	$f(c_q)$
1	Actif	Salarié	2 ; 3 ; 6	3
2	Actif	Employeur	3 ; 5 ; 6	3
3	Actif	Indépendant	3 ; 6 ; 7	3
4	Actif	Aide familial	3 ; 6	2
5	Actif	Inactif	3 ; 6	2
6	Inactif	Salarié	6	1
7	Inactif	Employeur	6	1
8	Inactif	Indépendant	6	1
9	Inactif	Aide familial	6	1
10	Inactif	Inactif	1 ; 4 ; 6	3

Tableau 8 : Les clés d'identification possibles au niveau de FM_2 en ne tenant pas compte des incohérences existant entre les modalités des identifiants indirects

Nous pourrions naïvement calculer les fréquences des clés d'identification $f_{PESA}(c(i))$ selon PESA en utilisant la formule (8) pour obtenir le résultat exposé dans le tableau suivant.

Indice i de u_i	**Type d'activité**	**Statut d'emploi**	C_i	$f_{PESA}(c(i))$
1	Inactif	Inactif	$\{c_{10}\}$	3
2	Actif	Salarié	$\{c_1\}$	3
3	Actif	ND	$\{c_1, c_2, c_3, c_4, c_5\}$	2
4	Inactif	Inactif	$\{c_{10}\}$	3
5	Actif	Employeur	$\{c_2\}$	3
6	ND	ND	$\{c_1, c_2, c_3, c_4, c_5, c_6, c_7, c_8, c_9, c_{10}\}$	1
7	Actif	Indépendant	$\{c_3\}$	3

Tableau 9 : Les fréquences des clés d'identification des personnes figurant dans FM_2 selon PESA, abstraction faite de la cohérence des modalités des identifiants indirects

Le fichier de microdonnées FM_2 ne serait alors pas considéré 2-anonyme. Néanmoins, les combinaisons (Actif ; Inactif), (Inactif ; Salarié), (Inactif ; Employeur), (Inactif ; Indépendant), (Inactif ; Aide familial) ne relèvent pas du réel. De ce fait, l'ensemble $C_{incoh} = \{c_5, c_6, c_7, c_8, c_9\}$ et le nombre de clés d'identifications possibles est reconsidéré de la sorte : $card(C) = 2 \times 5 - 5 = 5$. Par conséquent, le calcul rigoureux des fréquences des clés d'identification $f_{PESA}(c(i))$ donne le résultat présenté au tableau qui suit, et fait de FM_2 un fichier de microdonnées bel et bien 2-anonyme. Retenons également que l'approche orthodoxe et OPTA l'auraient considéré comme étant 3-anonyme.

Indice i de u_i	**Type d'activité**	**Statut d'emploi**	C_i	$f_{PESA}(c(i))$
1	Inactif	Inactif	$\{c_{10}\}$	3
2	Actif	Salarié	$\{c_1\}$	3
3	Actif	ND	$\{c_1, c_2, c_3, c_4\}$	2
4	Inactif	Inactif	$\{c_{10}\}$	3
5	Actif	Employeur	$\{c_2\}$	3
6	ND	ND	$\{c_1, c_2, c_3, c_4, c_{10}\}$	**2**
7	Actif	Indépendant	$\{c_3\}$	3

Tableau 10 : Les fréquences revues des clés d'identification des personnes figurant dans FM_2 selon PESA, en veillant à la cohérence des clés d'identification prises en considération

4. Conclusion

La bonne évaluation du risque d'identification est une étape clé du processus d'anonymisation des microdonnées à laquelle il faut s'atteler minutieusement et prudemment. Par la suite, il se doit d'employer et de combiner plusieurs méthodes d'anonymisation –aux identifiants indirects et aux variables à caractère jugé sensible– dont nous citons non exhaustivement la méthode de recodage, la méthode des suppressions locales (Van Gelderen, 1995, [13], dans de Waal et Willenborg, 1998, [5]), la méthode PRAM (Gouweleeuw *et al.*, 1998, [7]), la méthode de microagrégation (Defays et Nanopoulos, 1993, [6], dans Mateo-Sanz et Domingo-Ferrer, 1998, [9]), la méthode de permutation par rang (Dalenius et Reiss, 1982, [4]). S'intéresser au concept de *l*-diversité introduit par Machanavajjhala et ses collègues (2006, [8]) est aussi essentiel. Ensuite, il est important de jauger la perte d'information indubitablement engendrée par l'application de méthodes d'anonymisation et d'œuvrer à la minimiser autant que faire se peut avant de pouvoir affirmer avoir anonymisé un fichier de microdonnées en bonne et due forme.

Suite à notre démonstration, nous pouvons affirmer notre penchant pour PESA qui permet une meilleure appréhension –à notre sens– du risque d'identification basé sur la règle du *k*-anonymat qu'OPTA, que nous jugeons déjà plus réaliste que l'approche orthodoxe.

L'appellation « approche pessimiste » que nous avons choisie a certes une connotation péjorative, mais dans le domaine de la protection des données individuelles, elle revêt un caractère prudent et soucieux de la préservation de l'intimité et des secrets des unités statistiques enquêtées.

Il va sans dire qu'une unité statistique protégée avec une méticulosité opiniâtre est une unité statistique qui fera à l'avenir davantage confiance aux instituts statistiques publics, et qui développe ainsi une prédisposition à devenir plus conciliante et coopérative lors d'enquêtes statistiques futures et à fournir des informations plus précises et véridiques la concernant.

De plus, une meilleure protection de la vie privée de la population enquêtée peut plaider pour une diffusion plus accrue des microdonnées dans une optique d'open data menant à un accroissement de la transparence des données et à une amélioration de la qualité des données à travers les commentaires constructifs des utilisateurs.

Bibliographie

[1] Benschop, T., Machingauta, C., and Welch, M. (2017). *Statistical Disclosure Control for Microdata: A Practice Guide (version 1.2).* Technical document, The World Bank.

[2] Bergeat, M. (2016). La gestion de la confidentialité pour les données individuelles. *Documents de travail de l'Insee*, M2016/07.

[3] Ciglic, M., Eder, J., and Koncilia C. (2014). *k*-Anonymity of Microdata with NULL Values, *Transactions on Large-Scale Data- and Knowledge-Centered Systems XXIV: Special Issue on Database- and Expert-Systems Applications*, 193–220.

[4] Dalenius, T. and Reiss, S. (1982). Data-swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference*, 6, 73–85.

[5] de Waal, A.G. and Willenborg, L.C.R.J. (1998). Optimal Local Suppression in Microdata. *Journal of Official Statistics*, Vol. 14, N^{o} 4, 421–435.

[6] Defays, D. and Nanopoulos, P. (1993). Panels of enterprises and confidentiality: the small aggregates method. *Proceedings of the 1992 Symposium on Design and Analysis of Longitudinal Surveys*, Statistics Canada, 195–204.

[7] Gouweleeuw, J.M., Kooiman, P., Willenborg, L.C.R.J., and de Wolf, P.-P. (1998). Post Randomisation for Statistical Disclosure Control: Theory and Implementation. *Journal of Official Statistics*, Vol. 14, N^{o} 4, 463–478.

[8] Machanavajjhala, A., Gehrke, J., Kifer, D., and Venkitasubramaniam, M. (2006). *l*-Diversity: Privacy Beyond *k*-anonymity. *ACM Trans. Knowl. Discovery From Data (TKDD)*, 1, 24–35.

[9] Mateo-Sanz, J. M. and Domingo-Ferrer, J., (1998). A comparative study of microaggregation methods, *Qiiestiió*, 22, 511–526.

[10] Samarati, P. and Sweeney, L. (1998a). Generalizing Data to Provide Anonymity when Disclosing Information. *Proc. of the 17th ACM Symp. on Principles of database systems*, PODS '98, ACM.

[11] Samarati, P. and Sweeney, L. (1998b). *Protecting Privacy when Disclosing Information: k-Anonymity and Its Enforcement through Generalization and Suppression.* Technical report.

[12] Templ, M. (2008). Statistical Disclosure Control for Microdata Using the R-Package sdcMicro. *Transactions on Data Privacy*, 1(2), 67–85.

[13] Van Gelderen, R.P. (1995). *ARGUS Statistical Disclosure Control of Survey Data.* Report, Statistics Netherlands.