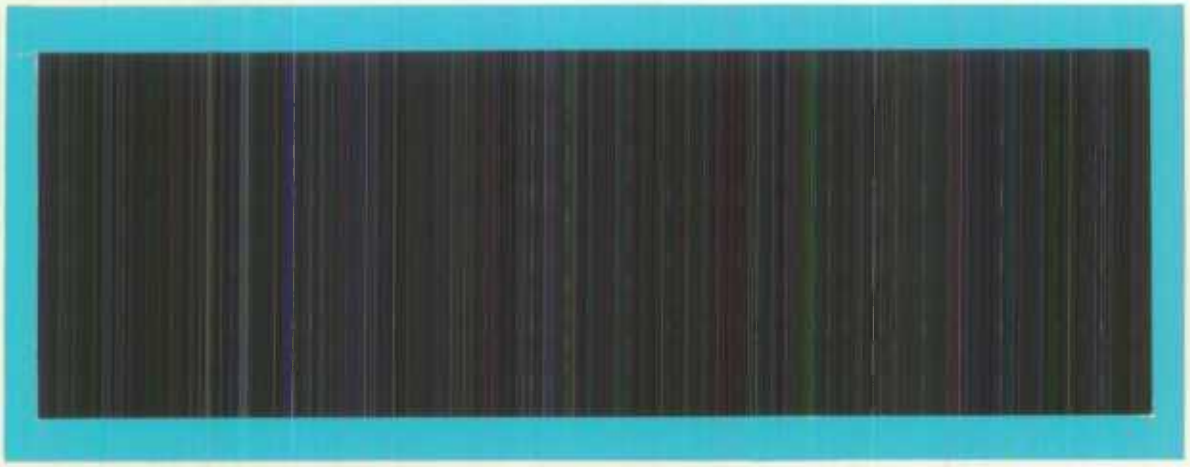
Statistics Canada Statistique Canada



Methodology Branch

Social Survey
Methods Division

Direction de la méthodologie

Division des méthodes
d'enquêtes sociales

Canada

CAHIER DE TRAVAIL NO. SSMD-91-017F

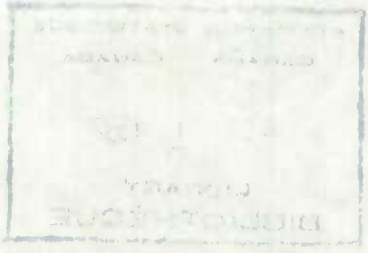
MÉTHODOLOGIE

ARRONDISSEMENT ALÉATOIRE OU CONVENTIONNEL: UN SECOND REGARD

D. ROBERGE, A. DEMNATI, R. KAUSHAL
DIVISION DES MÉTHODES D'ENQUÊTES SOCIALES

SSMD-91-017F

Novembre 1991



Arrondissement aléatoire ou conventionnel: un second regard

(D. Roberge, A. Demnati, R. Kaushal)

RÉSUMÉ

Les organismes statistiques recueillent presque toujours leurs données sous le sceau du secret. Afin de préserver le caractère confidentiel de ces données, ils utilisent diverses méthodes pour supprimer ou limiter les risques de divulgation. Dans cet article nous discutons de la protection des données lorsque les statistiques publiées sont des tableaux de fréquences. Plus particulièrement, nous évaluons les deux méthodes de protection les plus fréquemment utilisées, c.a.d. l'arrondissement aléatoire et conventionnel. Nous remettons en cause l'énoncé selon lequel les tableaux de fréquences arrondis aléatoirement ne peuvent être retrouvés. Nous proposons une méthodologie pour évaluer et comparer le risque de divulgation statistique après arrondissement.

ABSTRACT

Data collected by Statistical Agencies are almost always collected under the seal of confidentiality. To ensure confidentiality, Statistical Agencies use various methods to avoid or limit the risk of statistical disclosure. In this paper, we consider the methods use to protect the collected data when the published statistics are frequency tables. More specifically, we evaluate the two methods the most frequently used, i.e. random rounding and conventional rounding. We call into question the believe that: frequency tables random rounded cannot be undone. We also propose a methodology to evaluate and compare the risk of statistical divulgation after rounding.

1. INTRODUCTION

Les Agences Statistiques ont privilégiées depuis toujours l'utilisation des tableaux de fréquence pour présenter les données qualitatives (ou catégoriques) à un point tel qu'il n'est pas rare de voir des données quantitatives être transformées en données qualitatives pour être présentées dans des tableaux de fréquences. Par exemple, pour la variable quantitative "revenu", la moyenne et le total sont souvent accompagnés par un tableau de fréquences qui présente sa distribution en utilisant des classes de revenu arbitraires.

L'arrondissement aléatoire ne cesse de gagner en popularité comme méthode de protection pour les tableaux de fréquences. L'arrondissement aléatoire est utilisée depuis 1971 pour les données du recensement du Canada.

Dans la perspective de la protection de la confidentialité, l'avantage prétendu de l'arrondissement aléatoire sur l'arrondissement conventionnel est que les valeurs des cellules des tableaux de fréquences arrondies de façon aléatoire ne peuvent être retrouvées ; ce qui n'est pas le cas pour l'arrondissement conventionnel [Nargundkar et Saveland, 1972; Fellegi et Phillips, 1974]. Dans cet article nous montrons que cette affirmation est inexacte ; c'est-à-dire qu'il est possible de retrouver les valeurs initiales des cellules de tableaux de fréquences que ceux-ci soient arrondis aléatoirement ou conventionnellement.

Y-a-t-il cependant divulgation statistique parce qu'il est possible de retrouver les valeurs initiales des tableaux de fréquences arrondis ? Pour répondre à cette question, nous définissons dans la section 2 ce qu'est la divulgation statistique pour les tableaux de fréquences. Dans la section 3, nous démontrons comment un espion statistique s'y prendrait pour retrouver les valeurs initiales de tableaux de fréquences arrondis et ainsi menacer la confidentialité des unités enquêtées. Finalement, dans la section 4, nous concluons en soulignant les implications des résultats présentés sur le choix de la méthode d'arrondissement pour protéger les tableaux de fréquences.

2. LA DIVULGATION STATISTIQUE DANS LES TABLEAUX DE FRÉQUENCES

Lorsqu'il est possible de déduire à partir des données d'un tableau de fréquences de l'information précédemment inconnue au sujet d'une unité enquêtée et d'identifier cette unité, il y a divulgation statistique. Pour identifier une unité enquêtée, l'espion statistique utilisera un sous-ensemble des caractéristiques utilisées pour définir la population représentée et le tableau de fréquences. Si l'espion statistique doit utiliser l'ensemble des caractéristiques utilisées pour définir la population représentée et le tableau de fréquences, il n'a pas divulgation statistique car aucune nouvelle information n'est déduite. Lorsque la valeur d'une caractéristique peut être déduite, on dit qu'il y a divulgation statistique positive. La deuxième ligne du tableau 1 illustre un cas de divulgation positive. Dans ce cas, on peut conclure que toutes les personnes enquêtées âgées entre 35 et 49 ans appartiennent à la classe de revenu $ 50,000 et plus. Lorsque la valeur d'une caractéristique ne peut être déduite, mais qu'il est possible d'éliminer une ou plusieurs

valeurs, on dit qu'il y a divulgation statistique négative. La première ligne du tableau 1 illustre un cas de divulgation négative. Dans ce cas, on ne peut déduire la classe de revenu des personnes enquêtées âgées entre 20 et 34 ans, mais on peut déduire qu'aucune personne enquêtée de cette classe d'âge a un revenu supérieur à $ 49,999.

Tableau 1 - Illustrations de cas de divulgation statistique

		Classes de revenu			Total
		0 - 24,999	25,000 - 49,999	50,000 et plus	
Classes d'âge	20 - 34	2	5	0	7
	35 - 49	0	0	7	7
	50 - 64	3	3	1	7
Total		5	8	8	21

Le lecteur remarquera que les divulgations statistiques positives et négatives sont causées dans l'exemple ci-haut par la publication de cellules vides (zéro). Les cellules vides sont les seules cellules qui causent la divulgation statistique si nous supposons que l'espion statistique ne possède aucune connaissance à priori sur la population représentée. Dans le tableau 1, si l'espion statistique est la personne enquêtée âgée entre 50 et 64 ans avec un revenu de $50,000 et plus, alors il peut déduire qu'aucune autre personne enquêtée de cette classe d'âge a un revenu supérieur à $49,999 (divulgation statistique négative). Il est aussi possible que l'espion statistique connaisse à priori de l'information au sujet de plus d'une personne enquêtée. Si, par exemple l'espion statistique connaît les deux personnes enquêtées âgées entre 20 et 34 ans, il peut alors déduire que le revenue de toutes les autres personnes enquêtées de cette classe d'âge se situe entre $25,000 et 49,999 (divulgation statistique positive).

Il n'est pas possible de cerner avec précision les connaissances à priori dont disposent les espions statistiques. Il nous faut donc poser des hypothèses. On peut réduire le risque de divulgation en supposant que les connaissances à priori des espions statistiques sont considérables et protéger les tableaux de fréquence en conséquence. Cependant, toute réduction du risque de divulgation se traduit par une réduction de la quantité d'information publiée. En fait, la seule méthode absolue de protection consiste à supprimer entièrement toutes les données publiées. Étant donné que la raison d'être des organismes de statistiques est de publier des données dans

l'intérêt de la société, cette solution n'est pas acceptable. Il nous faut donc établir un équilibre entre le risque de divulgation statistique et la quantité d'information statistique publiée. Pour ce faire, nous nous devons de ne considérer que les menaces "raisonnables".

3. ARRONDISSEMENT

L'arrondissement prévient la divulgation statistique en altérant les cellules d'un tableau de fréquences. Nous allons brièvement décrire l'arrondissement aléatoire et conventionnel et la protection qu'elles offrent. Par la suite, nous présenterons la méthodologie utilisée pour retrouver les valeurs initiales des tableaux de fréquences arrondis.

L'arrondissement conventionnel consiste à arrondir la valeur de chaque cellule au plus près multiple d'un nombre entier donné appeler base d'arrondissement. Par exemple, si la valeur de la cellule est 2 est la base d'arrondissement est 5, la valeur arrondie sera 0. L'arrondissement aléatoire consiste à arrondir la valeur d'une cellule vers le multiple inférieur ou supérieur de la base d'arrondissement à partir d'un processus aléatoire. Les probabilités que la valeur soit arrondie vers le multiple supérieur et inférieur sont déterminées pour que l'espérance mathématique de la valeur arrondie soit égale à la valeur initiale. Par exemple, si la valeur de la cellule est 2 et que la base d'arrondissement est 5, la valeur arrondie sera 0 ou 5 avec les probabilités respectives de 3/5 et 2/5.

Ces deux méthodes d'arrondissement protègent un tableau de fréquences en créant une région d'incertitude autour de la valeur de chaque cellule de la table. Cette région d'incertitude prend la forme d'un intervalle qui inclut toutes les valeurs initiales possibles (avant arrondissement). La connaissance de la méthode d'arrondissement et de la base d'arrondissement est suffisante pour déduire ces intervalles. Le tableau suivant présente les intervalles pour une base d'arrondissement de 5 pour l'arrondissement aléatoire et conventionnel.

Tableau 2. Intervalles des valeurs initiales possibles

Valeur arrondie	Arrondissement conventionnel	Arrondissement aléatoire
0	0 - 2	0 - 4
5	3 - 7	1 - 9
10	8 - 12	6 - 14
...	...	...

Il est possible de réduire la grandeur de un ou plusieurs intervalles en combinant les intervalles des cellules avec celui du total. Prenons, par exemple, le tableau suivant arrondi conventionnellement avec une base d'arrondissement de 5.

I	II	III	IV	Total
5	5	10	5	35

Remplaçons les valeurs arrondies par les intervalles des valeurs initiales possibles.

I	II	III	IV	Total
3 - 7	3 - 7	8 - 12	3 - 7	33 - 37

Les sommes des bornes inférieures et supérieures des cellules sont respectivement 17 (3+3+7+3) et 33 (7+7+12+7). À partir de ces sommes, on obtient l'intervalle {17-33} pour le total. En combinant cette information avec l'intervalle des valeurs possibles pour le total obtenu directement {33-37}, on constate qu'il n'y a qu'une solution. Le total est 33, et cela correspond au tableau solutionné suivant:

I	II	III	IV	Total
7	7	12	7	33

Quoique ce tableau soit solutionné, il ne présente qu'un très petit risque de divulgation statistique. Pour qu'il y est divulgation statistique, il faudrait que l'espion statistique connaisse initialement 7 personnes enquêtées appartenant à la même cellule.

De façon similaire, un tableau de fréquences arrondi aléatoirement peut-être solutionné. Prenons, par exemple, le tableau suivant arrondis aléatoirement avec une base d'arrondissement de 5.

I	II	III	IV	Total
5	5	10	5	5

Remplaçons les valeurs arrondies par les intervalles des valeurs initiales possibles.

I	II	III	IV	Total
1 - 9	1 - 9	6 - 14	1 - 9	1 - 9

Les sommes des bornes inférieures et supérieures des cellules sont respectivement 9 (1+1+6+1) et 41 (9+9+14+9). À partir de ces sommes, on obtient l'intervalle {9-41} pour le total. En combinant cette information avec l'intervalle des valeurs possibles pour le total obtenu directement {1-9}, on constate qu'il n'y a qu'une solution. Le total est 9, et cela correspond au tableau solutionné suivant:

I	II	III	IV	Total
1	1	6	1	9

Ce tableau solutionné présente un risque de divulgation statistique élevé. Un espion statistique ne doit connaître initialement qu'une personne enquêtée pour qu'il y est divulgation statistique négative.

Nous devons en fait tenir compte des connaissances initiales de l'espion statistique lorsque nous voulons déterminer si un tableau peut-être solutionné ou non. Prenons, le tableau suivant arondi aléatoirement avec une base d'arrondissement de 5.

I	II	III	IV	Total
0	10	10	10	15

Remplaçons les valeurs arrondies, par les valeurs initiales possibles.

I	II	III	IV	Total
0 - 4	6 - 14	6 - 14	6 - 14	11 - 19

En utilisant la méthode décrite précédemment, le tableau ne peut-être solutionné mais les intervalles de valeurs possibles peuvent être réduits.

I	II	III	IV	Total
0 - 1	6 - 7	6 - 7	6 - 7	18 - 19

Si nous supposons maintenant que l'espion statistique connaît une personne enquêtée et que cette personne à la caractéristique I, alors l'intervalle des valeurs possibles pour cette cellule est réduit à une seule valeur {1}.

I	II	III	IV	Total
1	6 - 7	6 - 7	6 - 7	18 - 19

Ce tableau peut maintenant être complètement solutionné. Les sommes des bornes inférieures et supérieures des cellules sont respectivement 19 (1+6+6+6) et 25 (9+9+14+9). À partir de ces sommes, on obtient l'intervalle {19-25} pour le total. En combinant cette information avec l'intervalle des valeurs possibles pour le total obtenu plus tôt {18-19}, on constate qu'il n'y a qu'une solution. Le total est 19, est cela correspond au tableau solutionné suivant:

I	II	III	IV	Total
1	6	6	6	19

Des tableaux de deux dimensions et plus peuvent être solutionnés en utilisant le même principe. L'approche consiste alors à appliquer successivement la méthode pour chaque ligne et chaque colonne du tableau arrondis jusqu'à ce que celui-ci soit complètement solutionné ou qu'aucun intervalle ne puisse être réduit davantage.

Il est impossible de comparer d'une façon absolue le risque de divulgation statistique pour l'arrondissement aléatoire et l'arrondissement conventionnel car il dépend des tableaux initiaux à arrondir. Cependant, la comparaison des intervalles des valeurs possibles initiales, peut nous fournir des indications quant aux performances respectives des deux méthodes, puisque l'étendue de ces intervalles qui déterminent la protection. Par exemple, le tableau 3 présente les valeurs possibles initiales pour l'arrondissement conventionnel en base 5 et l'arrondissement aléatoire en base 3. Dans les deux cas pour la valeur arrondie 0, l'intervalle des valeurs possibles initiales est le même. Pour les autres valeurs arrondies, le nombre de valeurs possibles initiales est aussi le même, soit 5. Ceci indique que la performance dans ces deux cas pourrait bien être très similaire.

Tableau 3. Comparaison des intervalles des valeurs initiales possibles

Arrondissement Aléatoire Base 3		Arrondissement Conventionnel Base 5	
Valeur Arrondie	Valeurs Initiales Possibles	Valeur Arrondie	Valeurs Initiales Possibles
0	0 - 2	0	0 - 2
3	1 - 5	5	3 - 7
6	4 - 8	10	4 - 8

Nous devons souligner que nous avons jusqu'à maintenant uniquement considéré l'arrondissement d'un tableau de fréquence à une occasion. Alors que l'arrondissement conventionnel répété d'une valeur produit toujours le même résultat, l'arrondissement aléatoire

par sa nature peut générer des résultats différents. L'obtention de résultats différents en plus d'être une nuisance pour les utilisateurs réduit la protection de confidentialité offerte. L'observation de deux valeurs arrondies différentes réduit de plus de la moitié l'étendue de l'intervalle des valeurs possibles initiales. Par exemple, pour une base d'arrondissement de 5, l'étendue de l'intervalle des valeurs initiales possibles est de 9 pour toutes valeurs arrondies supérieure à 5. Cependant si deux valeurs arrondies différentes sont générés alors l'intervalle des valeurs possibles initiales devient l'intersection des intervalles et son étendue est réduit à 4. Le tableau suivant présente le cas où les deux valeurs arrondies observées sont 5 et 10.

Tableau 4. Exemple de réduction de l'intervalle des valeurs initiales possibles

	Valeur arrondie	Valeurs initiales possibles	Étendue des valeurs initiales possibles
Première Occasion	5	1 - 9	9
Seconde Occasion	10	6 - 14	9
Intersection	5 et 10	6 - 9	4

CONCLUSION

Nous avons démontré que l'arrondissement aléatoire ne possède pas sur l'arrondissement conventionnel l'avantage certain d'empêcher la divulgation statistique. Puisqu'il n'est pas possible de faire un choix sur cette base, qu'en est-il des autres caractéristiques de ces deux méthodes? Nous examinerons l'utilité des données après arrondissement et la facilité d'implantation des deux méthodes.

Toutes les méthodes de protection de la confidentialité produisent, sous une forme ou une autre, une perte d'information pour les utilisateurs. L'arrondissement n'y échappe pas. Il n'existe cependant pas de mesures absolues de la perte d'information, celle-ci dépendra des données et de leur utilisation. Quoiqu'il en soit, l'étendue des valeurs possibles initiales nous fournit une indication de la perte d'information. Ainsi, de la même façon dont nous avons conclu que la protection offerte par l'arrondissement aléatoire avec une base d'arrondissement de 3 est similaire à celle offerte par l'arrondissement conventionnel avec une base d'arrondissement de 5, nous pouvons conclure que la perte d'information est aussi similaire. Pour juger de l'utilité

de l'information, nous devons ajouter à cette perte d'information, la compréhension que les utilisateurs ont de la méthode et de ces impacts sur la qualité des données. Il ne fait aucun doute dans notre esprit que les utilisateurs ont une meilleure compréhension de l'arrondissement conventionnel que de l'arrondissement aléatoire.

Implanter l'arrondissement conventionnel est une tâche des plus simples. L'implantation de l'arrondissement aléatoire quoique légèrement plus complexe demeure une tâche relativement simple. Cependant, l'implantation de l'arrondissement aléatoire se complique rapidement si nous devons nous assurer que l'arrondissement répété d'une même statistique ne donnera pas lieu à des résultats différents.

Parce que, pour une même perte d'information, l'arrondissement conventionnel offre une protection similaire à celle offerte par l'arrondissement aléatoire, parce que son implantation est plus simple et parce que les utilisateurs en ont une meilleure compréhension, nous croyons que dans bien des situations cette méthode est plus appropriée que l'arrondissement aléatoire pour protéger les tableaux de fréquences.

RÉFÉRENCES

Fellegi I.P. and J.L. Phillips, Statistical Confidentiality: Some theory and applications to data dissemination, Annals of economic and social measurements, 3/2, 1974

Nragundkar, M.S. and Saveland, W., Random rounding: A means of preventing disclosure of information about individual respondents in aggregate data, American Statistical Association Proceedings, Social Statistical Section, 1972.