Recueil du Symposium 2001 de Statistique Canada
La qualité des données d'un organisme statistique : une perspective méthodologique

PROTECTION CONTRE LA DIVULGATION POUR LES PRODUITS STATISTIQUES NON CONVENTIONNELS : LES ESTIMATEURS PAR NOYAU DE LA DENSITÉ[1]

David R. Merrell[2] et Arnold P. Reznek[3]

RÉSUMÉ

Dans la présente communication, nous traitons d'un élément précis d'un programme de recherche axé sur la protection contre la divulgation dans le cas des produits statistiques « non conventionnels ». Nous soutenons que ces produits présentent des risques de divulgation différents de ceux qui existent habituellement et qu'il faut désormais en tenir compte. Plus précisément, nous soutenons que les estimateurs par noyau de la densité, s'ils constituent des descriptions puissantes (de grande qualité) d'échantillons représentatifs, présentent cependant des risques de divulgation qui dépendent essentiellement du choix d'une largeur de bande. Nous illustrons ces risques à l'aide d'un ensemble unique de données non confidentielles sur l'univers statistique des mines de charbon et nous proposons des solutions possibles. Enfin, nous décrivons les pratiques en usage au Center for Economic Studies du U.S. Census Bureau pour effectuer l'analyse de divulgation statistique portant sur les estimateurs par noyau de la densité.

MOTS-CLÉS : Estimateur par noyau de la densité; Protection contre la divulgation; Produits statistiques non conventionnels

1. INTRODUCTION

Les organismes statistiques gouvernementaux sont tenus par la loi de protéger la confidentialité des microdonnées fournies par les répondants lors des enquêtes et des recensements. Leurs produits conventionnels sont des agrégats des microdonnées—habituellement des tableaux de fréquences ou de totaux, ou encore des fichiers de microdonnées à grande diffusion. À l'égard de ces produits, les organismes emploient diverses méthodes de protection de la confidentialité (restriction de la divulgation statistique), dont la restriction des résultats diffusés (par exemple, suppression de cellules et réagrégation) et la perturbation (par exemple, arrondissement des résultats ou perturbation aléatoire des microdonnées sous-jacentes ou des résultats).

Les chercheurs du Center for Economic Studies (CES) et des Research Data Centers (RDC) du U.S. Census Bureau produisent couramment des produits statistiques qui sont (du moins pour le Census Bureau) « non

[1] La présente étude expose les résultats de recherches et d'analyses entreprises par le personnel du U.S. Census Bureau. Le Bureau en a fait une révision plus limitée que celle qu'il accorde aux publications officielles. La diffusion du présent document vise à informer les intéressés de la recherche continue et à susciter la discussion au sujet des travaux en cours. Nous tenons à remercier les participants au Symposium 2001 de Statistique Canada à Ottawa (Canada) ainsi que les participants au séminaire du Center for Economic Studies à Washington, DC pour leurs observations et leurs suggestions très utiles. Nous tenons également à remercier Timothy Dunne pour ses propos judicieux et ses suggestions avisées.
[2] David R. Merrell, Center for Economic Studies, U.S. Census Bureau et Department of Economics, UCLA, California Census Research Data Center, 4250 Public Policy Building, Box 951484, University of California Los Angeles, Los Angeles, CA 90095-1484 (dmerrell@ccrdc.ucla.edu)
[3] Arnold P. Reznek, Center for Economic Studies, U.S. Census Bureau, Washington Plaza II, Washington, DC 20233 (rezne001@ces.census.gov)

conventionnels »[4]. Ces produits statistiques comprennent habituellement les estimations de paramétres et les statistiques connexes tirées de modèles de régression linéaire et non linéaire et, parfois, des produits statistiques tirés de modèles d'estimation semi-paramétriques et non paramétriques ou de modèles de simulation[5]. Nous croyons que les méthodes actuelles de restriction de la divulgation ne sont pas toujours applicables à ces produits. Toutefois, il n'existe guère de lignes directrices à cet égard, bien qu'il s'agisse d'un sujet de recherche valable selon un rapport du U.S. Federal Committee on Statistical Methodology (FCSM, 1994, p. 87). Dans ces conditions, il est difficile de concilier la nécessité de publier des résultats de recherche significatifs et celle de protéger la confidentialité des renseignements fournis par les répondants. On a besoin de nouvelles lignes directrices, car de nombreux organismes, aux États-Unis et à l'étranger, ont établi ou envisagent d'établir des centres de données de recherche ou des entités semblables à accès limité.

La présente communication porte sur un aspect de travaux de plus grande envergure qui sont en cours. Dans le cadre de ce programme de recherche, nous tentons de décrire les risques de divulgation liés aux régressions linéaires et non linéaires et d'établir si les méthodes existantes de protection de la confidentialité sont applicables dans le contexte de la régression. Sur les méthodes existantes de protection de la confidentialité, voir Federal Committee on Statistical Methodology (1994); sur l'orientation de notre recherche, voir Merrell et Reznek (2001).

Nous examinons également les estimateurs non paramétriques des densités de probabilité, c.-à-d. les estimateurs par noyau de la densité, en nous intéressant particulièrement à l'incidence du choix d'une largeur de bande, du choix d'un support de densité et de la présence d'observations aberrantes sur le risque de divulgation. Enfin, nous évaluons l'efficacité des règles de divulgation existantes à l'égard du maintien de la confidentialité lorsqu'elles sont appliquées à des estimations non paramétriques des fonctions de densité. Nous ferons état de ces travaux de recherche dans un prochain article.

Dans la présente communication, nous nous concentrons sur le cas des estimateurs par noyau de la densité. Si nous présentons uniquement des exemples liés à des données sur des établissements, nos observations pourraient aussi porter sur des ménages, des personnes ou même des entreprises. Nous croyons que notre réflexion se rapporte à des types de données économiques ou démographiques. Dans un cas comme dans l'autre, le problème qui se pose est courant : nous cherchons à produire et à diffuser des estimations statistiques de grande qualité (comme en fournissent les estimations par noyau de la densité) tout en empêchant la divulgation des renseignements donnés par les répondants. Nous abordons les risques de divulgation que présentent ces formes d'estimateurs et proposons quelques moyens pour éliminer ces risques[6].

Le plan de la présente communication est le suivant : dans la section 2, nous définissons et décrivons les estimateurs par noyau de la densité. Dans la section 3, nous examinons les données, nous illustrons deux risques de divulgation connexes et des moyens permettant de les éliminer et nous résumons les rapports

[4] Pour en savoir plus sur le CES et les RDC, on visitera le site Web du CES à l'adresse http://www.ces.census.gov. Voir également Cooper, Merrell, Nucci et Reznek (1998) et Reznek, Cooper et Jensen (1997).

[5] En général, les types de produits statistiques en question sont évidemment très courants dans les domaines de la statistique et de la recherche en sciences sociales. Toutefois, ils sont « non conventionnels » dans le contexte de la restriction de la divulgation statistique; la mesure des risques de divulgation et les méthodes de restriction de la divulgation visent surtout les tableaux et les fichiers à grande diffusion.

[6] Il est certain que la plupart des utilisations des estimateurs par noyau de la densité consistent à examiner des écarts entre des échantillons représentatifs, c.-à-d. des variations de l'estimation par noyau de la densité. Par exemple, on pourrait s'intéresser aux variations de distribution de la taille des établissements entre deux cohortes d'entrée d'usines de fabrication ou de mines de charbon. Pour les examiner, il faudrait représenter graphiquement plus d'une estimation par noyau de la densité. Une fonction unique comme celles qui sont présentées plus loin serait normalement atypique du résultat d'une recherche. Nous utilisons des estimateurs uniques pour présenter les risques de divulgation liés à n'importe quelle estimation par noyau de la densité et nous soutenons implicitement que les risques, les solutions possibles et les politiques actuelles s'appliquent sans égard au mode de présentation des résultats. Néanmoins, nous sommes conscients que la plupart des résultats de recherche compareraient deux estimations de densité (ou plus).

entre ces deux préoccupations. Dans la section 4, nous expliquons la politique en vigueur au Center for Economic Studies et aux Research Data Centers du U.S. Census Bureau lorsqu'il s'agit d'effectuer une analyse de divulgation statistique à l'égard des estimations par noyau de la densité. Enfin, dans la section 5, nous présentons nos conclusions.

2. ESTIMATEURS PAR NOYAU DE LA DENSITÉ

2.1 Définition

Les estimateurs par noyau de la densité permettent aux chercheurs de décrire des distributions transversales de manière très simple et très claire. Ces estimateurs lissent les distributions selon une largeur de bande donnée et une fonction noyau. La fonction noyau attribue à chaque observation un poids qui rend cette observation proportionnelle à sa distance par rapport au centre d'une bande donnée. De manière formelle, l'estimateur par noyau de la densité est représenté par l'équation suivante :

$$\hat{f}_K = \frac{1}{nw} \sum_{i=1}^{n} K\left[\frac{x - X_i}{w} \right]$$

où n est la taille de l'échantillon, w est la largeur de bande, K est la fonction noyau, x est une observation et X_i est le point milieu d'une bande donnée. Il convient de noter que la fonction noyau, K, peut se présenter sous un certain nombre de types, notamment Epanechnikov, rectangulaire ou gaussien. Bref, l'estimation par noyau de la densité est la somme des écarts pondérés d'une observation par rapport au point milieu d'une bande où les poids sont attribués par la fonction noyau.

D'un point de vue mécanique, la fonction noyau a peu d'importance pour la plupart des chercheurs[7], et la plupart des logiciels choisissent une fonction implicite. Toutefois, le choix de la largeur de bande, w, présente un intérêt considérable. Comme la largeur de bande détermine le nombre d'observations comprises dans chaque bande, elle exerce une forte influence sur la forme de la densité de noyau calculée. S'il existe un certain nombre de méthodes de calcul de la largeur de bande, on peut aussi calculer une largeur de bande « optimale »; pour plus de détails, voir Tapia et Thompson (1978) ou Fox (1990).

2.2 Risques de divulgation

Les estimateurs par noyau de la densité constituent une façon simple et utile d'illustrer des échantillons représentatifs de données de manière non paramétrique. Ils s'avèrent ainsi très avantageux pour les utilisateurs de microdonnées confidentielles. Par contre, comme il s'agit d'estimateurs statistiques « non conventionnels », ils présentent différentes sortes de risques lorsqu'il s'agit de protéger la confidentialité des renseignements fournis par les répondants. Dans la présente section, nous décrivons brièvement ces risques et dans la suivante, nous les illustrons à l'aide de données non confidentielles ayant les mêmes propriétés et la même structure que certaines des données stockées au Center for Economic Studies et aux Research Data Centers du U.S. Census Bureau.

En général, nous trouvons que les estimateurs par noyau de la densité présentent des risques de divulgation de deux manières étroitement liées. Premièrement, ils présentent un risque lorsque les ensembles de données comportent un petit nombre de valeurs extrêmes. Comme nous allons le montrer, le support calculé de l'estimateur par noyau de la densité est parfois très proche du support réel de la distribution empirique à l'une des extrémités de la distribution, sinon aux deux. Deuxièmement, et d'une manière qui n'est pas étrangère à la première préoccupation, le choix d'une petite largeur de bande (choisie de manière

[7] D'ailleurs, la fonction noyau n'a pas plus d'importance lorsqu'il s'agit d'effectuer l'analyse de divulgation statistique. Toutefois, nous mentionnons plus loin un cas où la fonction noyau constitue un élément important du processus de protection contre la divulgation. Par contre, d'un point de vue général, elle a peu d'importance à l'égard de l'analyse de divulgation statistique.

optimale ou non) a tendance à rendre les observations individuelles directement observables (notamment celles qui se situent aux queues de la distribution empirique) et à faire en sorte que le support calculé de l'estimateur par noyau de la densité se rapproche du support réel de la distribution empirique.

3. ILLUSTRATION DES RISQUES DE DIVULGATION QUE PRÉSENTENT LES ESTIMATEURS PAR NOYAU DE LA DENSITÉ

3.1 Les données

Avant d'illustrer les deux risques de divulgation que présentent les estimateurs par noyau de la densité, nous décrivons les données utilisées à cette fin. Nous utilisons un ensemble de microdonnées unique qui contient l'univers statistique des mines de charbon répertoriées aux États-Unis en 1996. Ces données sont recueillies conformément à la réglementation et sous la surveillance de la U.S. Mine Safety and Health Administration et contiennent les réponses des mines concernant notamment l'emploi trimestriel, le nombre d'heures travaillées et le nombre de tonnes courtes de charbon produit. Il est certain que si ces données sont recueillies au niveau de l'établissement, elles ne sont pourtant pas confidentielles. Pour une description détaillée des données, voir Merrell (1999, 2000) et Dunne et Merrell (2001).

Pour nos besoins, ces données présentent un avantage important, car la structure de leurs fichiers et leurs propriétés statistiques sont semblables à celles des données recueillies par le U.S. Census Bureau, qui sont confidentielles. Par exemple, le recensement des manufactures américaines (ensemble de données qui est confidentiel en vertu des lois des États-Unis) contient des renseignements recueillis au niveau des établissements sur à peu près les mêmes variables (et plus encore), mais les distributions – par exemple, celle de la taille (mesurée en fonction de l'emploi) – sont très asymétriques. Ce phénomène empêche la création d'un ensemble de données à grande diffusion sur la fabrication et nous empêcherait également d'illustrer les problèmes liés aux estimateurs non conventionnels qui utilisent ce genre de données. Par contre, les données sur les mines de charbon présentent le même genre d'asymétrie dans les distributions de la taille mais, comme ces données ne sont pas confidentielles en vertu des lois des États-Unis, nous pouvons les utiliser en remplacement des données sur la fabrication et illustrer nos préoccupations et nos solutions à l'égard des produits statistiques non conventionnels, comme dans le cas des estimateurs par noyau de la densité.

Pour illustrer les risques de divulgation que présentent les estimateurs par noyau de la densité, nous calculons la distribution de l'effectif des mines en 1996. L'effectif d'une mine en 1996 correspond au nombre moyen de travailleurs par année (moyenne des trimestres). En 1996, il y avait 1 328 mines déclarées « actives »; encore une fois, pour plus de détails, voir Merrell (1999). La plus grande mine comptait 840,75 travailleurs et la plus petite, 0,5[8].

La figure 1 montre l'estimation par noyau de la densité à l'égard de ces mines. Certes, la fonction noyau est de type Epanechnikov, et la largeur de bande est calculée comme suit :

$$m = [\sqrt{\text{var}(x)}, \frac{IQR(x)}{1.349}]$$

$$w = \frac{0.9m}{n^{1/5}}$$

où x est une observation, IQR(x) est l'intervalle inter-quartile de x et n est la taille de l'échantillon. Il s'agit de la méthode implicite de calcul de la largeur de bande au moyen du logiciel économétrique Stata. Voir « Stata 7, référence H-P ».

[8] Plus loin, ces valeurs sont appelées valeurs maximale et minimale « réelles ».

La figure 1 présente cette estimation par noyau de la densité. Certains points sont dignes de mention. Premièrement, de façon très générale, l'estimation par noyau de la densité constitue une description très claire (donc de grande qualité) de la distribution de l'effectif des mines de charbon en 1996. Deuxièmement, de façon plus spécifique, il convient de noter que la masse de la distribution se situe à l'extrémité inférieure de la densité. Troisièmement, on trouve à l'extrémité supérieure de l'estimation par noyau de la densité une très longue queue due à la présence d'une seule très grande mine en 1996. Enfin, à l'intérieur de cette longue queue, il y a quelques mines de grande taille, mais moins extrême; elles sont représentées par les petites « bosses » qui marquent la longue queue de l'estimation de la densité. Ces mines, à l'extrémité supérieure de l'estimation par noyau de la densité, présentent un risque de divulgation lié à ce type de produit statistique non conventionnel.

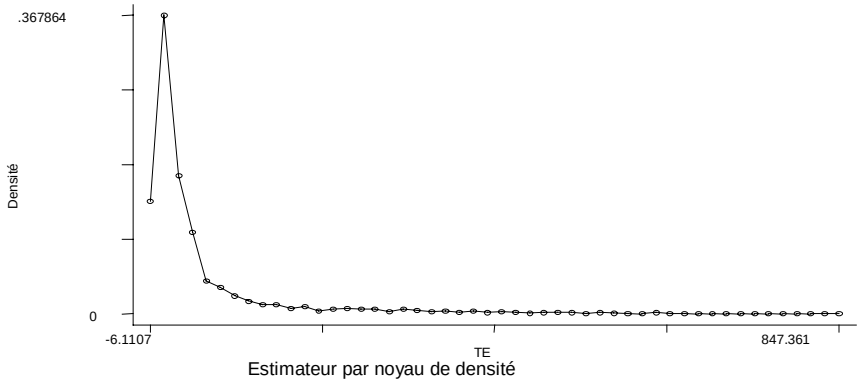

Figure 1. Estimation par noyau de la densité de la taille de l'effectif en 1996 (largeur de bande implicite)

3.2 Problèmes liés aux valeurs extrêmes et solutions possibles

S'il est vrai que l'estimateur par noyau de la densité fournit une estimation très claire, simple et concise de la distribution de l'effectif, il présente cependant le risque de divulguer les renseignements fournis par les répondants—encore une fois, de privilégier la nécessité de fournir des estimations statistiques de grande qualité aux dépens de celle de protéger la confidentialité des renseignements. Reportons-nous à la figure 1 et, plus précisément, aux valeurs extrêmes qui se situent dans la queue supérieure. On y trouve une seule mine qui est très grande par rapport à la plupart des autres, et l'estimation par noyau de la densité pour l'effectif de cette mine est de 847,36 travailleurs. Il convient de noter que la valeur maximale réelle (soit l'effectif réel de la plus grande mine) est de 840,75 travailleurs. Ces valeurs sont extrêmement proches; en effet, la valeur calculée se situe à moins de un pour cent de la valeur réelle.

Dans la plupart des hypothèses, on estimerait qu'il s'agit là d'une dérogation au principe de la non-divulgation, et ce, pour deux raisons. Premièrement, déclarer la valeur maximale réelle équivaut à identifier un établissement—au lieu de fournir une estimation de son effectif. Deuxièmement, comme l'estimation par noyau de la densité est pratiquement égale à la valeur maximale réelle, les protocoles de divulgation habituels en empêcheraient la publication[9].

[9] S'il est vrai qu'il existe de petites « pointes » observables entre la masse de la distribution et ses valeurs extrêmes, il est peu probable qu'on identifie ces mines aussi facilement que l'établissement représenté par la valeur extrême (si tant est qu'on arrive à les identifier); tout dépend du contexte. Il est donc probable qu'il y aura moins lieu de s'en préoccuper par rapport à l'observation maximale ou supérieure—du moins dans le contexte de l'analyse de divulgation statistique.

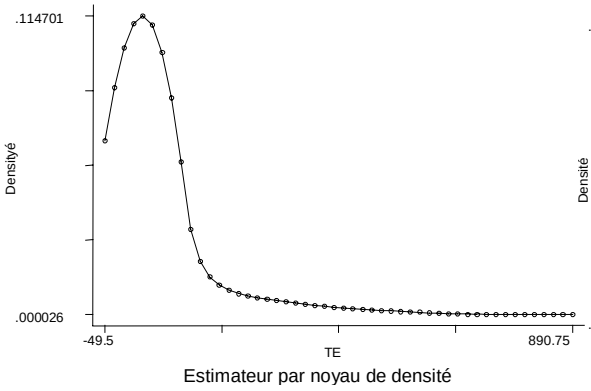

Figure 2a. Estimation par noyau de la densité, w=50

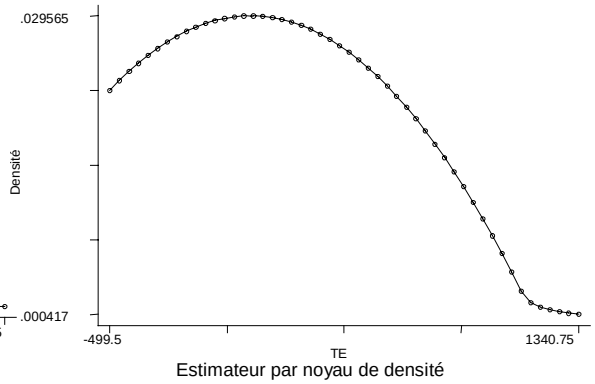

Figure 2b. Estimation par noyau de la densité, w=500

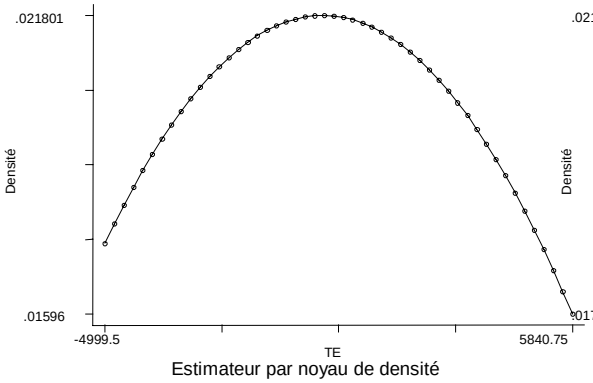

Figure 2c. Estimation par noyau de la densité, w=5 000

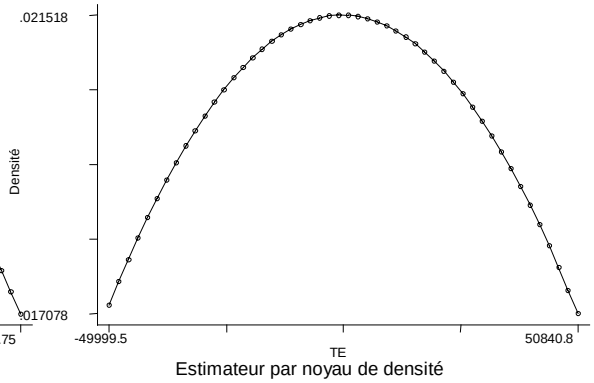

Figure 2d. Estimation par noyau de la densité, w=50 000

Comment donc concilier la nécessité d'obtenir des estimations statistiques de grande qualité à l'aide des estimateurs par noyau de la densité et celle de maintenir la confidentialité des renseignements fournis par les répondants? Les propriétés de l'estimation par noyau de la densité offrent une solution. En effet, nous savons que le choix de la valeur de la largeur de bande détermine essentiellement la courbe plus ou moins lisse de l'estimation de la densité. Or, le choix d'une plus grande largeur de bande devrait moins attirer l'attention sur une observation individuelle—c.-à-d. que la fonction noyau attribue un poids moindre à cette observation extrême, car elle se retrouve encore plus loin du centre d'une bande maintenant plus large. Ce phénomène a pour effet de masquer la valeur réelle de l'observation maximale en augmentant l'intervalle du support calculé de l'estimation de la densité.

Prenons les figures 2a, 2b, 2c et 2d, dans lesquelles on rajuste (manuellement) la largeur des bandes utilisées pour estimer la densité de noyau. Dans la figure 2a, on fixe la largeur de bande à 50; dans chacune des trois autres, la largeur de bande augmente d'un ordre de grandeur par rapport à la précédente. Les effets sont manifestes. Premièrement, la distribution devient de plus en plus lisse. Deuxièmement, et ce qui est plus important pour nos besoins, le support calculé à l'égard de la distribution passe de 847,36 travailleurs

dans la figure 1 à 890,75 travailleurs dans la figure 2a, puis à 50 840,8 travailleurs dans la figure 2d[10]. Il en résulte que plus la largeur de bande est grande, plus la valeur maximale réelle (soit 840,75 travailleurs) diffère de la valeur calculée. Lorsque la largeur de bande est fixée à 50, la valeur réelle se situe à moins de 6 % de la valeur calculée (contre 1 % dans le cas de la largeur de bande implicite, dans la figure 1). Plus la largeur de bande est grande, plus la valeur calculée à l'extrémité supérieure s'écarte de la valeur réelle[11].

D'un point de vue purement technique, il semble donc que la solution au problème des valeurs extrêmes consisterait à fixer une plus grande largeur de bande afin d'entraîner une divergence entre les valeurs réelles des observations extrêmes et les valeurs calculées obtenues d'après l'estimateur par noyau de la densité. Par contre, cette solution entraîne une perte de niveau de détail (et, donc, de qualité) de l'estimation. Si l'on compare les figures 1 et 2a à 2d, il est clair que dans la première, la courbe montre un niveau raisonnable de détail dans la distribution de l'effectif des mines de charbon et que dans la dernière, elle ne montre aucun niveau de détail—surtout lorsque la largeur de bande dépasse la taille de l'échantillon total. Plus la largeur de bande est grande, plus elle élimine le pouvoir de l'estimateur par noyau de la densité de décrire les distributions transversales de l'effectif[12].

3.3 Problèmes liés aux petites largeurs de bande et solutions possibles

La deuxième préoccupation générale concernant les estimateurs par noyau de la densité tient à une largeur de bande fixée « trop petite »—ce qui peut arriver même lorsqu'on choisit la largeur de bande « optimale ». Si une grande largeur de bande a tendance à lisser la distribution, voire à la « surlisser », l'inverse est également vrai : une petite largeur de bande a tendance à montrer un plus grand niveau de détail ou, du point de vue de la protection contre la divulgation, à « sous-lisser » la distribution, ce qui a pour effet de rendre plus d'observations individuellement observables aux extrémités inférieure et supérieure de la distribution. De plus, le support calculé de l'estimation par noyau de la densité a tendance à se rapprocher des valeurs réelles de la distribution empirique aux extrémités.

Pour illustrer ce point, prenons les figures 3a, 3b, 3c et 3d, dans lesquelles on réduit (manuellement) la largeur de bande en la faisant passer de la valeur implicite (figure 1) vers zéro. Les effets sont manifestes. Premièrement, si l'on compare d'une figure à l'autre, on remarque que le niveau de détail est de plus en plus élevé dans la queue supérieure de l'estimation par noyau de la densité. Ainsi, il y a plus de « pointes » dans la queue supérieure, ce qui signifie—comme on pouvait s'y attendre—que plus la largeur de bande diminue, plus on identifie de grandes mines de charbon.

La deuxième caractéristique qui se dégage est celle des valeurs calculées de l'estimation par noyau de la densité aux grandes et aux petites extrémités. À l'extrémité supérieure, la valeur maximale calculée en fonction de la valeur implicite est de 847,36 travailleurs dans la figure 1; elle est de 845,75 travailleurs

[10] Manifestement, les figures 2c et 2d ont peu de signification puisque la largeur de bande est plus grande que le nombre total d'observations disponibles. Ainsi, on utilise toutes les observations pour calculer l'estimation par noyau de la densité à chaque point. Toutes les observations se situent donc à l'intérieur du support calculé dans ces deux cas.

[11] On observe un effet semblable, quoique moins important à ce point, à l'extrémité inférieure de la distribution. Ainsi, la valeur calculée de la valeur extrême inférieure devient rapidement très différente de la valeur réelle.

[12] Un autre risque de divulgation (dont nous reparlerons à la section 4) tient aux échelles de l'axe horizontal. Il convient de noter que si l'on soustrait la largeur de bande (en supposant, pour les besoins de la discussion, qu'elle est connue) de la valeur maximale étiquetée sur le graphique, on se rapproche beaucoup de la valeur réelle et, dans certains cas, on pourrait obtenir la valeur exacte de la valeur maximale réelle. (Nous remercions Randy Becker et Kristen McCue d'avoir attiré notre attention sur ce point.) Comme nous le mentionnons plus loin dans la présente communication, nous préconisons, à l'égard de ce risque de divulgation, de supprimer l'étiquetage sur l'axe horizontal de toute représentation graphique de l'estimation par noyau de la densité.

lorsque w=5, de 841,75 travailleurs lorsque w=1 et de 841,25 travailleurs lorsque w=0,5. Manifestement, plus la largeur de bande diminue, plus la valeur calculée se rapproche de la valeur réelle[13].

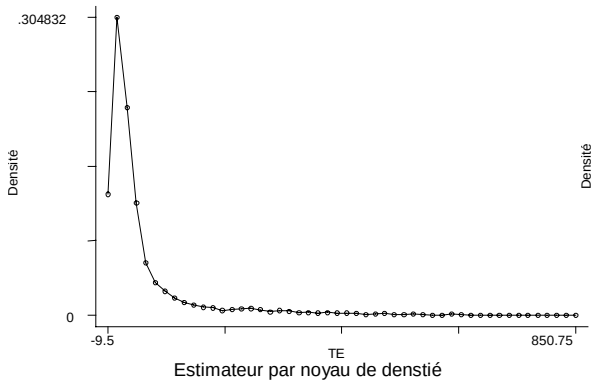

Figure 3a. Estimation par noyau de la densité, w=10

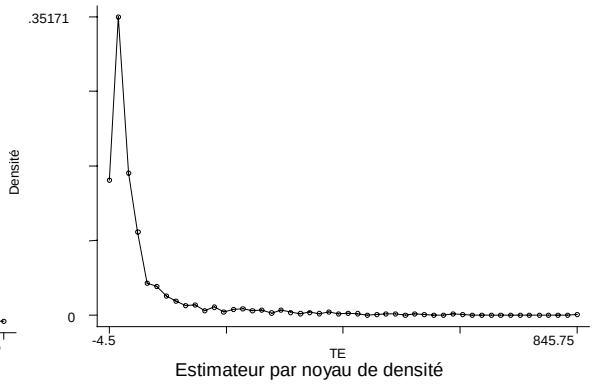

Figure 3b. Estimation par noyau de la densité, w=5

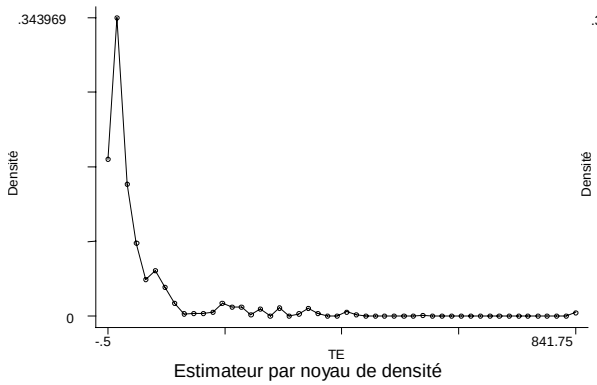

Figure 3c. Estimation par noyau de la densité, w=1

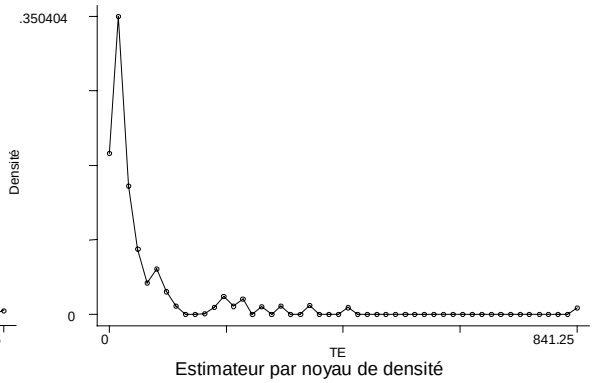

Figure 3d. Estimation par noyau de la densité, w=0,5

À l'extrémité inférieure de l'estimation par noyau de la densité, on observe un résultat semblable : la valeur minimale calculée se rapproche de la valeur minimale réelle. Il convient de noter que dans la figure 1, la valeur minimale calculée est de –6,11 travailleurs, ce qui diffère considérablement de la valeur réelle de 0,5 travailleur. Toutefois, plus la valeur fixée de la largeur de bande est petite (voir les figures 3b à 3d), plus la valeur minimale calculée se rapproche de la valeur minimale réelle. Donc, dans le cas où la largeur de bande est « trop petite », non seulement le niveau de détail est plus élevé au sujet d'un plus grand nombre d'établissements à l'extrémité supérieure de la distribution, mais les valeurs extrêmes minimale et maximale calculées se rapprochent davantage des valeurs maximale et minimale réelles. Encore une fois, le niveau de détail plus élevé au sujet des établissements compris dans la queue supérieure et le rapprochement accru du support calculé par rapport au support empirique réel risquent d'être considérés comme des dérogations au principe de la non-divulgation.

Donc, s'il est vrai qu'une grande largeur de bande a tendance à « surlisser » la densité de noyau, une petite largeur de bande a tendance à présenter un risque de divulgation non seulement en présentant plus de renseignements au sujet de valeurs élevées (puisqu'elle dénote un niveau de détail plus élevé dans la queue

[13] Il convient de noter que lorsque la largeur de bande est fixée à dix (figure 3a), les extrémités supérieure et inférieure de l'estimation par noyau de la densité deviennent plus grandes que dans la figure 1. On peut en déduire que la largeur de bande « optimale » calculée plus haut est inférieure à dix—dans cet exemple.

supérieure et que la valeur maximale calculée se rapproche de la valeur maximale réelle), mais aussi en présentant plus de renseignements au sujet de valeurs faibles (puisque la valeur minimale calculée se rapproche de la valeur minimale réelle). Du point de vue de la simplicité, il semble donc que la solution au problème d'une petite largeur de bande consisterait à en choisir une plus grande, soit la même solution que celle qui est présentée dans la section 3.2. Encore une fois, cette solution entraîne une réduction de la puissance de description et, donc, de la qualité de l'estimateur par noyau de la densité.

3.4 Synthèse

Dans les sections 3.2 et 3.3, nous soutenons que le choix de la largeur de bande constitue un facteur essentiel pour empêcher la divulgation des renseignements fournis par les répondants lorsqu'on utilise les estimateurs par noyau de la densité. Dans le cas des valeurs extrêmes, on peut les masquer en choisissant une plus grande largeur de bande et dans le cas de petites largeurs de bande, encore une fois, on peut masquer les valeurs extrêmes en choisissant une plus grande largeur de bande. Dans les deux cas, la solution entraîne une réduction de la qualité de l'estimation puisque le support calculé de la densité de noyau augmente considérablement selon la largeur de bande choisie. De plus, il est impossible de fixer objectivement une largeur de bande qui préserve la qualité de l'estimation tout en atténuant le risque d'identifier les renseignements fournis par un répondant donné, que cette largeur soit extrêmement grande ou extrêmement petite.

Une dernière solution, peut-être, consisterait à calculer des estimations de distributions de taille fondées sur des renseignements catégoriques. Autrement dit, on pourrait diviser le nombre d'observations en catégories d'effectif et en représenter graphiquement la fréquence pour produire un histogramme simple (un type spécial d'estimateur par noyau de la densité). Manifestement, cette solution est la plus restrictive de toutes les possibilités puisqu'elle proscrirait l'utilisation d'une méthode très simple, et pourtant puissante (c.-à-d. de grande qualité) pour décrire des échantillons représentatifs—d'autant plus qu'il existe des méthodes (si subjectives soient-elles) pour réduire les risques de divulgation.

4. PRATIQUES EN USAGE AU CENTER FOR ECONOMIC STUDIES DU U.S. CENSUS BUREAU

Tout en reconnaissant que les estimateurs par noyau de la densité présentent des risques de divulgation, nous estimons que leur capacité de présenter des estimations de grande qualité devrait inciter les analystes de la divulgation à mettre au point des méthodes d'atténuation de ces risques. Dans la présente section, nous abordons les pratiques en usage au Center for Economic Studies et aux Research Data Centers du U.S. Census Bureau en ce qui concerne l'analyse de divulgation statistique liée aux estimateurs par noyau de la densité. Il nous semble que ces pratiques, si subjectives soient-elles, concilient la nécessité de produire des estimations statistiques de grande qualité et la nécessité concomitante de maintenir la confidentialité des renseignements fournis par les répondants.

La première étape de l'analyse de divulgation statistique à l'égard des estimateurs par noyau de la densité consiste à tester un échantillon d'établissements ou de personnes en utilisant des protocoles établis de non-divulgation statistique. Ainsi, pour les établissements, nous examinons (de façon générale) le nombre d'entreprises et la prédominance des grandes entreprises (d'après les ventes réelles, l'emploi ou toute autre mesure dont on dispose et qu'on juge pertinente). Pour les données sur les personnes, nous examinons le nombre de répondants compris dans chaque cellule de données. Dans un cas comme dans l'autre (c.-à-d. pour les établissements ou pour les personnes), nous comparons les caractéristiques calculées des cellules à une valeur minimale spécifiée; pour plus de détails sur ces règles, voir Merrell et Reznek (2001). Pour que l'estimation par noyau de la densité permette d'effectuer une analyse de divulgation statistique claire, il est essentiel de bien respecter ces règles.

Après avoir déterminé que l'échantillon global qui sous-tend une estimation par noyau de la densité donnée respecte les protocoles établis de non-divulgation statistique, nous appliquons habituellement une autre condition qui, selon nous, atténue les préoccupations énoncées à la section 3. Ainsi, nous préconisons

comme pratique générale que les chercheurs s'abstiennent de préciser l'échelle dans toute représentation graphique de l'estimation par noyau de la densité. On arrive ainsi à faire abstraction du problème des largeurs de bande assez petites pour aggraver le risque d'identifier des observations extrêmement grandes ou petites, comme nous l'avons démontré plus haut et comme il en est question dans la note 12[14].

De plus, nous adoptons comme principe de publier uniquement des représentations graphiques d'estimateurs par noyau de la densité, et jamais des estimations ponctuelles. Supposons que les estimations ponctuelles (c.-à-d. les valeurs calculées) soient publiées. Dans ce cas, il suffirait de connaître la fonction noyau et la largeur de bande pour trouver les valeurs réelles des renseignements fournis par les répondants individuels. À partir de ces deux renseignements, on pourrait reconstruire l'ensemble réel de données, ce qui constituerait une dérogation sans équivoque au principe de la non-divulgation. En outre, il serait facile d'y arriver; il suffirait d'utiliser un tableur électronique standard pour récupérer les renseignements réels fournis par les répondants[15], comme le montre un simple examen de la formule de l'estimateur par noyau de la densité; voir l'équation formulée plus haut.

Enfin, nous adoptons une approche plutôt prudente à l'égard des produits statistiques non conventionnels. À ce titre, nous nous efforçons de demander conseil à un grand nombre d'analystes de la divulgation statistique. Nous espérons ainsi en arriver à un consensus en ce qui concerne ces méthodes statistiques innovatrices et de grande qualité, mais aussi déployer tous les efforts possibles pour éliminer les risques de divulgation primaire et secondaire. En cas de doute ou d'incertitude, nous consultons le Disclosure Review Board du U.S. Census Bureau pour connaître son avis collectif.

5. CONCLUSION

Les organismes statistiques qui autorisent l'accès à des microdonnées confidentielles font régulièrement face à un dilemme : la nécessité de produire des estimations statistiques de grande qualité (et d'en promouvoir la production) tout en maintenant des normes rigoureuses de confidentialité et de non-divulgation statistique. À mesure qu'évoluent les méthodes de production de ces estimations, notre conception de l'analyse de divulgation statistique et notre politique à cet égard doivent aussi évoluer. La présente communication propose un type d'estimateur (il y en a sans doute bien d'autres) qui produit des estimations statistiques de grande qualité ainsi que de nouveaux risques de divulgation des renseignements fournis par les répondants.

Nous présentons un exemple simple à l'aide d'un ensemble unique de données possédant des propriétés distributives très semblables à celles de données sur des établissements non publics, comme celles du Recensement des manufactures. Nous montrons que le risque de divulguer les renseignements fournis par les répondants dépend essentiellement du choix de la largeur de bande lorsqu'on produit des estimations par noyau de la densité. Le choix d'une largeur de bande « trop petite » (qu'on la choisisse selon une condition « optimale » ou manuellement) entraîne le risque d'identifier un plus grand nombre d'observations individuelles à l'extrémité supérieure de la densité de noyau, ainsi qu'un risque très élevé d'identifier les valeurs maximale et minimale réelles, car le support calculé de la fonction de noyau a tendance à se rapprocher du support réel de la distribution empirique. Par contre, le choix de largeurs de bande de plus en plus grandes pour masquer les identités de petites ou de grandes observations extrêmes sacrifie la clarté de la méthode statistique employée pour produire des descriptions simples mais significatives d'échantillons représentatifs.

Malheureusement, nous n'avons pu trouver que des solutions subjectives à ce problème. Nous appliquons des protocoles de non-divulgation classiques à des échantillons utilisés pour calculer les estimations par noyau de la densité et, une fois ces critères respectés, nous préconisons alors de supprimer les échelles dans les représentations graphiques de l'estimation par noyau de la densité. Nous arrivons ainsi à empêcher

[14] Cette exigence se heurte souvent à une vive résistance de la part des utilisateurs de microdonnées confidentielles.

[15] Nous remercions Timothy Dunne de nous avoir suggéré d'envisager ce point.

l'estimateur d'identifier les renseignements fournis par les répondants, mais au prix d'une réduction de la puissance explicative de l'estimation. En outre, nous adoptons comme principe de publier uniquement des représentations graphiques de l'estimateur par noyau de la densité, puisque les estimations ponctuelles, lorsqu'on connaît le type de la fonction noyau et la largeur de bande, peuvent servir à trouver les valeurs réelles correspondant aux répondants.

Bref, comme les organismes statistiques sont de plus en plus en mesure de produire des estimations innovatrices et de grande qualité, notre conception de l'analyse de divulgation statistique doit également évoluer. Nous devons arriver à concilier la nécessité de produire des estimations significatives et de grande qualité avec celle, tout aussi fondamentale, de protéger les renseignements fournis par les répondants. Les estimations par noyau de la densité constituent un cas intéressant, puisqu'il s'agit de descriptions simples mais puissantes d'échantillons représentatifs, mais elles présentent de nouveaux risques de divulgation qui échapperont peut-être aux méthodes utilisées traditionnellement pour limiter la divulgation. Nous prenons cet exemple pour illustrer le dilemme auquel font face de plus en plus souvent les organismes statistiques qui accordent un accès restreint à des microdonnées confidentielles, ainsi que des solutions possibles à ce dilemme.

BIBLIOGRAPHIE

Cooper, J.M.R, D.R. Merrell, A.R. Nucci, et A.P. Reznek (1998), "Protecting Confidential Data at Remote Access Sites: Lessons Learned from Research Data Centers," *American Statistical Association Proceedings of the Section on Government Statistics and Section on Social Statistics*, 1998, pp. 91-96.

Dunne, T. et D.R. Merrell (2001), "Gross Employment Flows in U.S. Coal Mining," *Economics Letters*, 2, pp. 217-224.

Federal Committee on Statistical Methodology (1994), "Report on Statistical Disclosure Limitation Methodology," Statistical Policy Working Paper 22, Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget.

Fox, J. (1990), "Describing Univariate Distributions," dans J. Fox et J.S. Long (éds) *Modern Methods of Data Analysis*, Newbury Park, CA: Sage Publications, pp. 58-125.

Merrell, D.R. (1999), "Acquisitions and Productivity in U.S. Coal Mining," Center for Economic Studies Working Paper No. 99-17, Washington, DC: U.S. Census Bureau.

Merrell, D.R. (2000), "Implications of Resource Exhaustion on Exit Patterns in U.S. Coal Mining," mimeo, Los Angeles, CA: UCLA.

Merrell, D.R. et A.P. Reznek (2001), "On Disclosure Protection for Non-Traditional Statistical Outputs," mimeo, Washington, DC: Center for Economic Studies, U.S. Bureau of the Census.

Reznek, A.P., J.M.R. Cooper, et J.B. Jensen (1997), "Increasing Access to Longitudinal Survey Microdata: the Census Bureau's Research Data Center Program," *American Statistical Association Proceedings of the Section on Government Statistics and Section on Social Statistics, 1997*, pp. 243-248.

StataCorp (2001), "Stata 7 Reference H-P," *Stata Statistical Software: Release 7.0*, College Station, TX: Stata Corporation.

Tapia, R.A. et J.R. Thompson (1978), *Nonparametric Probability Density Estimation*, Baltimore, MD: Johns Hopkins University Press.