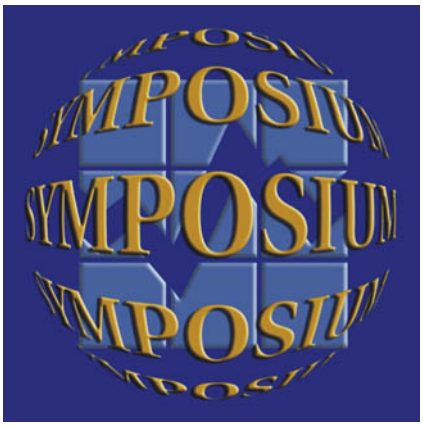
N° 11-522-XIF au catalogue

La série des symposiums internationaux de Statistique Canada - Recueil

Symposium 2003 : Défis reliés à la réalisation d'enquêtes pour la prochaine décennie

2003



Statistique Canada Statistics Canada

Canada

Recueil du Symposium 2003 de Statistique Canada
Défis reliés à la réalisation d'enquêtes pour la prochaine décennie

CONTRÔLE DE DIVULGATION STATISTIQUE POUR LES TABLEAUX : CHOIX D'UNE MÉTHODE

Paul B. Massell[1]

RÉSUMÉ

L'élaboration d'un certain nombre de nouvelles méthodes de contrôle de divulgation statistique (CDS) ces dernières années procure aux organismes statistiques des moyens supplémentaires de sauvegarde de la confidentialité de leurs données, mais cela veut aussi dire que ces organismes doivent choisir la méthode à appliquer à un ou plusieurs tableaux (ces tableaux sont peut-être liés les uns aux autres). Dans le présent document, nous analysons plusieurs méthodes CDS importantes en considérant un ensemble de facteurs d'intérêt. Le facteur dominant dans cette prise de décisions devrait être la nature de l'engagement pris par un organisme en matière de confidentialité. Un autre facteur d'importance est le temps d'élaboration et d'application du logiciel lié à chacune des méthodes en question. Autre facteur auquel on ne s'est guère attaché, la façon d'utiliser les tableaux doit être examinée. C'est, bien sûr, un facteur difficile à analyser, puisque l'usage varie selon les utilisateurs : certains chercheront seulement quelques valeurs de cellules et d'autres s'intéresseront aux subtilités de la modélisation statistique. Dans le présent document, nous proposons un certain nombre de méthodes statistiques d'analyse de méthodes CDS par rapport à ces facteurs. Nous espérons que notre analyse aidera les organismes à choisir la meilleure méthode CDS à appliquer à un jeu de tableaux.

MOTS CLÉS : Confidentialité, contrôle de divulgation statistique, perturbation de cellules, suppression de cellules, tableaux statistiques.

1. INTRODUCTION[2]

Un engagement en matière de confidentialité préside à la collecte de la plupart des données d'enquête et de recensement. Dans le présent document, nous supposons qu'un organisme statistique aura concrétisé cet engagement par des règles de confidentialité assez précises permettant de déterminer si un ou plusieurs tableaux renferment des données confidentielles. Supposons donc que l'organisme aura jugé, peut-être à l'aide d'un logiciel, qu'un jeu de tableaux pourrait contenir des renseignements confidentiels et devrait, par conséquent, faire l'objet d'un certain traitement de divulgation. Dans ce cas, il y a plusieurs facteurs à prendre en considération au moment de choisir la méthode de contrôle de divulgation statistique (CDS) à appliquer à de tels tableaux avant de les diffuser. Le choix d'une méthode comporte plusieurs étapes. D'abord, l'organisme doit savoir ce qui existe comme éventail de méthodes de contrôle de divulgation statistique de tableaux. Un grand nombre de ces méthodes ont été décrites dans des revues statistiques, des rapports techniques ou des livres (Willenborg et coll., 1996, 2001). Souvent, ces sources décrivent les propriétés mathématiques et statistiques d'une méthode, ainsi que le cheminement de ses concepteurs et des concepteurs du logiciel d'application avec les résultats des simulations qui se sont faites. Ces simulations sont importantes, car elles révèlent les limites de la méthode et indiquent les temps de traitement à prévoir pour des tableaux de taille diverse. Et les propriétés statistiques d'une méthode et les propriétés de son logiciel d'application détermineront si une technique CDS convient à la protection d'un jeu de tableaux.

À la section 2, nous définirons l'« application avant » d'une méthode de contrôle de divulgation statistique à des tableaux et, à la section 3, son « application arrière ». À la section 4, nous nous attacherons aux types de données des tableaux et aux diverses méthodes CDS qui y sont applicables. À la section 5, nous comparerons les types d'incertitude susceptibles

[1] Statistical Research Division, US Census Bureau, 4700, chemin Silverhill, pièce 3209-4, Washington, D.C. 20233, États-Unis, paul.b.massell@census.gov.

[2] Par ce rapport, nous voulons renseigner les intéressés sur les recherches qui se font et favoriser la discussion sur les travaux en cours. Les vues exprimées sont celles de l'auteur et ne correspondent pas nécessairement à celles du US Census Bureau.

d'être introduits à des fins de prévention de la divulgation. À la section 6, il sera question de l'utilité des tableaux après un traitement de divulgation. À la section 7 enfin, nous proposerons un cadre de prise de décisions pour toutes les questions que nous aurons exposées.

2. DÉFINITION DE L'APPLICATION AVANT D'UNE MÉTHODE DE CONTRÔLE DE DIVULGATION STATISTIQUE DE TABLEAUX

Nous jugeons bon de définir ce qu'il faut entendre par applications avant et arrière d'une méthode CDS. L'application **avant** est ce qui est normalement considéré comme la démarche clé de traitement de divulgation par laquelle un organisme statistique passe un programme de suppression ou de perturbation sur des tableaux, le but étant de produire un ensemble de tableaux protégés contre la divulgation. Il s'agit en général de modifier les valeurs d'un certain nombre de cellules, ce qui se fait en plusieurs étapes.

A. À l'aide d'une règle de sensibilité comme la règle p % ou (n,k) (Willenborg, 2001), on identifie quelles cellules sont « sensibles », ce qu'on appellera les cellules S. Dans le cas de tableaux liés les uns aux autres, la détermination de la sensibilité de cellules peut se révéler un exercice complexe. (À noter que, si les cellules S sont vides, aucune méthode CDS n'est à appliquer.)

B. Pour chaque cellule « sensible », on définit le minimum d'incertitude (MinU) que la méthode CDS doit introduire. On peut voir ce MinU comme une fonction productrice d'incertitude qui est appliquée à des cellules S dans les limites d'une certaine plage de valeurs. Une façon courante d'exprimer l'incertitude est de fixer un intervalle, auquel cas MinU(cellule i) sera un intervalle d'incertitude pour la cellule i. Idéalement, il serait bon de définir le maximum d'incertitude que doit introduire la méthode CDS dans chaque cellule (cellules S et autres). MaxU peut être considéré comme une fonction des cellules sur une certaine plage.

C. On applique la méthode CDS aux tableaux contenant des cellules S en respectant les contraintes imposées par MinU(cellules S) et MaxU(cellules). Le traitement de divulgation d'un tableau donnera les résultats escomptés s'il se crée un nouveau tableau conforme à ces contraintes d'incertitude. Pour juger si le traitement est réussi, on aura peut-être besoin d'une application arrière (voir plus loin).

D. Si la méthode employée a un caractère stochastique, la procédure pourrait prendre la forme d'un ensemble de fonctions de probabilité (fp) ou de densité où une de ces fonctions vise chacune des cellules. Il s'agit de la forme Pr(v-post = y | v-pre = x), où v-pre est la valeur de la cellule avant traitement de divulgation et v-post, sa valeur après traitement. Nous appellerons ces fonctions de probabilité ou de densité des probabilités **avant**. Les exemples les plus simples en sont le traitement stochastique en arrondis (Willenborg, 2001, p. 224) ou en perturbation de chiffres de tableaux. Dans ce cas, MinU(cellules S) et MaxU(cellules) sont des fp plutôt que des intervalles.

3. DÉFINITION DE L'APPLICATION ARRIÈRE D'UNE MÉTHODE DE CONTRÔLE DE DIVULGATION STATISTIQUE DE TABLEAUX

Il y a traitement arrière d'un tableau à la suite de l'application avant d'une méthode CDS. Un exemple en est l'exécution d'un programme de vérification de divulgation après suppression de cellules. S'il y a traitement arrière par un organisme statistique, l'exercice se fait généralement préalablement à la diffusion des tableaux, le but étant de juger si l'application avant a donné les résultats escomptés et peut-être de relever certains détails du traitement comme les niveaux précis de l'incertitude introduite dans certaines cellules (et peut-être dans toutes). Si le traitement avant a réussi, les tableaux peuvent être diffusés. Une fois qu'ils le sont, tout utilisateur peut en faire un traitement arrière (s'il dispose de ressources informatiques). L'objectif pourrait être de reconstituer le plus fidèlement possible le tableau d'origine (avant modification), peut-être pour préparer un exercice de modélisation. Ainsi, si un tableau est diffusé après suppression d'un certain nombre de cellules, un modélisateur pourrait vouloir imputer les valeurs des cellules en suppression avant de modéliser (voir plus loin). Bien sûr, il pourrait se contenter d'établir la meilleure estimation possible des valeurs de quelques cellules selon ce que permet le tableau. Dans un calcul de meilleure estimation, on établit l'intervalle

d'incertitude ou la fp (ou autre descripteur d'incertitude) d'une valeur de cellule, puis tire une valeur unique – qui est le mode de la densité, par exemple – de cette information.

Observations sur le traitement arrière
A. À propos de certaines méthodes de contrôle de divulgation statistique de tableaux, on a des garanties de succès, c'est-à-dire que ces techniques respectent toutes les contraintes d'incertitude, et ce, d'après des résultats mathématiques. Ainsi, des programmes de suppression de cellules qui font intervenir un algorithme d'analyse de réseau appliqué à de simples tableaux bidimensionnels assureront (au moins) le degré de protection qu'ils ont pour fonction de ménager. Dans ce cas, l'organisme n'a pas à recourir à un traitement arrière pour une vérification Min-U(cellules S), bien qu'une évaluation Max-U(cellules) ait peut-être encore son utilité. En d'autres termes, il s'agit de vérifier si on a introduit trop d'incertitude. Pour le cas évoqué de suppression de cellules, la théorie garantit que la méthode ne créera pas de sous-suppression, mais il pourrait y avoir sursuppression.

B. Si l'application avant crée un intervalle d'incertitude pour chaque cellule, on peut vérifier par le traitement arrière si l'intervalle est bien situé et assez étendu pour introduire l'incertitude recherchée.

C. Si le traitement avant est stochastique et que les probabilités antérieures sont clairement spécifiées, on peut faire un calcul bayésien de probabilités postérieures. Ce sont des probabilités postérieures qui peuvent permettre d'évaluer le degré d'incertitude introduit par une méthode CDS.

Ces probabilités postérieures sont de la forme Pr(v-pre = x | v-post = y) où v-pre est la valeur d'une cellule avant traitement de divulgation et v-post, sa valeur après traitement. Ces probabilités devraient comporter une incertitude qui satisfait MinU(cellules S).

4. TYPES DE DONNÉES ET MÉTHODES POSSIBLES DE CONTRÔLE DE DIVULGATION STATISTIQUE DE TABLEAUX

Dans le choix d'une méthode de contrôle de divulgation statistique de tableaux, on doit identifier en première étape la nature des données d'un ou de plusieurs tableaux.

A. Quel type de données le tableau renferme-t-il? S'agit-il de données d'ordre de grandeur ou de fréquence?
B. Si les données viennent d'une enquête par sondage, peut-on estimer l'erreur d'échantillonnage de chaque cellule?
C. Combien de tableaux doit-on élaborer? Quelle en est la taille? Sont-ils hiérarchisés?
D. Y a-t-il des cellules qui figurent dans deux tableaux et plus? Si tel est le cas, il s'agit de tableaux liés les uns aux autres. Sont-ils liés d'une façon moins évidente, par une relation additive, par exemple?

Nous verrons maintenant comment la réponse à ces questions influera sur le choix d'une méthode CDS.

A1 : Données de dénombrement (de fréquence)

Si les tableaux contiennent des données de dénombrement, le principal problème de divulgation réside dans l'existence de cellules aux valeurs de dénombrement très petites. Dans un certain nombre d'organismes, les cellules dont la valeur de dénombrement est de moins de 3 posent un problème que l'organisme doit résoudre par une méthode CDS qui élimine ces faibles valeurs. On reforme les tableaux en redéfinissant les catégories des variables en ligne et/ou en colonne. Le regroupement de catégories selon une dimension quelconque est un simple exemple de remodelage où on fond des catégories. Les données de dénombrement peuvent aussi faire l'objet d'un traitement en arrondis qui modifie presque toutes les cellules d'un tableau. C'est un traitement qui, d'une certaine façon, demeure excessif. Ajoutons qu'un traitement en arrondis présente un certain nombre de difficultés d'exécution : le traitement « classique » ne préserve pas l'additivité en général, d'autres le font en permettant l'arrondissement à un multiple qui n'est pas le plus proche dans le cas des valeurs marginales (dans ce cas, une valeur marginale ne correspond pas nécessairement à un des multiples les plus proches de la base). Ce qu'on appelle le traitement contrôlé en arrondis préserve l'additivité, mais son application prend du temps et rien ne garantit qu'elle pourra se faire dans un cadre tridimensionnel ou supérieur. Il y a aussi l'option de la perturbation aléatoire, dont l'application est rapide, mais aussi excessive. On peut en outre recourir à des méthodes

de perturbation plus récentes comme la méthode de correction contrôlée de tableaux (Cox, Dandekar, 2002). Dans une application de cette dernière méthode (Russell et coll., 2003), l'utilisateur peut juger « sensible » toute cellule dont la valeur de dénombrement est inférieure à n. Les valeurs de cellules qui se situent dans l'intervalle [1,n-1] passent à 0 ou à n. Souvent, on attribue la plus proche de ces deux valeurs. L'additivité est préservée en général, et on peut demander au programme d'essayer de la préserver sans changer les valeurs marginales. Enfin, on peut procéder par suppression (Willenborg et coll., 2001, p. 34), mais Sande (2003) dit de cette méthode qu'elle n'est pas normalement à recommander pour des tableaux à valeurs de dénombrement, car on ne s'en tient pas au type de protection qu'exigent de tels tableaux.

A2 : Données d'ordre de grandeur

Si des tableaux contiennent des données d'ordre de grandeur, chaque cellule représente la sommation d'une variable sur l'ensemble des répondants pour les valeurs des variables définissant cette cellule (il s'agit de variables en ligne et en colonne). Ainsi, avec des données d'ordre de grandeur économique, la variable peut être le chiffre d'affaires (en dollars) et les répondants, les établissements (entreprises) qui écoulent un certain produit (ligne) et se situent dans une certaine ville (colonne). Dans le cas de données économiques, la valeur de dénombrement des établissements dans une cellule est considérée comme généralement disponible, et aucune protection n'est jugée nécessaire. C'est la valeur (ordre de grandeur) de la variable de réponse d'une cellule qui peut être « sensible ». Le but est de rendre difficile son estimation précise. Ainsi, l'utilisateur et, en particulier, un répondant dont les données entrent dans la valeur d'une cellule ne pourront estimer avec précision la valeur (chiffre d'affaires, par exemple) à attribuer à un autre répondant pour cette cellule.

On peut procéder par suppression de cellules pour introduire au moins un certain degré d'incertitude, lequel peut prendre deux formes, à savoir (1) un intervalle de la forme [t - a, t + b] où t est la valeur réelle et a, b, les incertitudes limites de l'intervalle à introduire ou (2) un intervalle mobile d'étendue a + b qui contient la valeur réelle t. D'habitude, les organismes ne font pas connaître l'intervalle d'incertitude effectivement introduit dans chaque cellule en suppression, mais tout utilisateur disposant des ressources voulues pour élaborer un programme linéaire de modélisation d'additivité sera en mesure de calculer ces intervalles d'incertitude (aussi appelés intervalles de faisabilité ou de vérification). Dans certains types de perturbations, on crée une région d'incertitude pouvant comporter deux intervalles [t-a-c, t-a] et [t+b, t+b+d] ou seulement deux points t-a et t+b. Le traitement classique en arrondis convient moins à des données d'ordre de grandeur qu'à des données de dénombrement, les premières étant hautement asymétriques dans bien des cas. Si les valeurs de cellules ont des ordres de grandeur différents, il introduira sans doute trop d'incertitude dans certaines et trop peu dans d'autres. Il y a enfin l'option du regroupement de catégories, mais dans le cas des données économiques, on peut vouloir assurer la cohérence des définitions en ligne et en colonne d'une année à l'autre. Récemment, Sande (2003) a décrit un nouveau type de traitement en arrondis (qu'il appelle traitement de base élevée ou de variable) qui est conçu pour les données d'ordre de grandeur. Au début, l'utilisateur traite uniquement les cellules jugées « sensibles ». La procédure est en soi quelque peu complexe, visant à garantir l'introduction d'un degré spécifié d'incertitude et à empêcher qu'un calcul de point milieu (de l'intervalle d'arrondissement) ne livre une valeur proche de la valeur réelle.

B. Effet de données échantillonnées

Lorsqu'on calcule les valeurs de cellules de tableaux (valeurs de dénombrement ou d'ordre de grandeur) à partir de données échantillonnées, le risque de divulgation est ordinairement moins grand que si le calcul se fait à partir de données de population (données de recensement). Ainsi, dans l'incertitude introduite par une méthode CDS, on devrait retrouver l'erreur d'échantillonnage (et idéalement l'erreur non due à l'échantillonnage) toutes les fois que celle-ci peut être estimée. Pour la plupart, les méthodes CDS ne font rien en ce sens. Pour compenser l'absence de prise en compte directe de l'effet d'échantillonnage, les utilisateurs CDS fixent parfois pour leur programme un niveau de protection inférieur à ce qu'ils recherchent pour l'incertitude réelle. Cette stratégie fait intervenir l'idée que la protection réelle est supérieure à la protection nominale (spécifiée) à cause d'erreurs diverses (d'échantillonnage entre autres) dont sont entachées les données. Dans Willenborg (2001), il est question de l'effet sur l'incertitude dans le cas de données de dénombrement (p. 149) et d'ordre de grandeur (p. 144). Le traitement en arrondis de Sande (2003) dit à « base de variable » permet de faire intervenir tout type d'erreur – due ou non à l'échantillonnage – qui peut faire l'objet d'une estimation.

C. Nombre de tableaux avec leur taille et leur substructure

Il importe de savoir combien de tableaux doivent subir un traitement de divulgation et quelles en sont la taille et la substructure (hiérarchie), ces aspects nous indiquant l'importance de problèmes particuliers de traitement de divulgation. Pour des problèmes CDS d'envergure, on doit souvent écarter les méthodes CDS qui prennent beaucoup de temps, car leur temps d'exécution dépasse le temps alloué au contrôle de divulgation dans un passage de production.

D. Tableaux liés les uns aux autres

Dans (Willenborg, 2001, p. 150), on initie le lecteur à l'épineux problème des tableaux liés. Idéalement, un organisme devrait traiter simultanément les tableaux liés de manière à tenir compte de toutes leurs relations, mais il est souvent impossible d'agir ainsi, surtout dans le cas de gros tableaux, par limitation de puissance informatique. Quelquefois, les méthodes CDS permettent un traitement séquentiel de tableaux liés. Il faudra sans doute une méthode qui permette des compléments de traitement une fois effectué le traitement initial, ce qu'on appelle le retour arrière. Il y a retour arrière dans la plupart des programmes de suppression de cellules du US Census Bureau (Jewett, 1993). Cette reprise de traitement a lieu dans un passage de production. Idéalement, un organisme devrait veiller à la cohérence du traitement des tableaux liés entre les divers passages de production.

5. COMPARAISON DE DEGRÉ D'INCERTITUDE ET JUGEMENT DE SUFFISANCE SUR LE DEGRÉ D'INCERTITUDE INTRODUIT

Dans certains organismes fédéraux américains, on a l'obligation devant la loi de sauvegarder la confidentialité des données de tous les répondants aux enquêtes et aux recensements. Les prescriptions générales de la loi ne sont pas assez précises pour qu'on puisse en tirer des règles quantitatives d'application. Les organismes doivent se doter d'une politique en fonction d'expériences d'évaluation quantitative, de réflexion et/ou de simulation. Dans la politique adoptée, on doit en définitive répondre aux questions suivantes : (1) dans le cas de données de dénombrement, quel degré d'incertitude suffit à garantir, du point de vue de la capacité d'identification de répondants ou de divulgation par inférence, la sauvegarde de la confidentialité des données de tous les répondants? (2) dans le cas de données d'ordre de grandeur, quel degré d'incertitude doit-on introduire dans un tableau de sorte que la meilleure estimation qui en soit tirée ne permette pas à l'utilisateur de porter atteinte à la confidentialité des données des répondants? Ainsi, une politique de confidentialité transforme un énoncé qualitatif en un ensemble d'énoncés quantitatifs pouvant servir à évaluer le contrôle de divulgation statistique dans sa spécificité. Pour décrire cette notion, on pourrait employer le terme « exigences quantitatives de confidentialité des tableaux ».

5.1 Traitement avant : introduction d'un degré d'incertitude

5.1.1 Degré d'incertitude par intervalle : intervalles de faisabilité calculés par un programme de contrôle de divulgation

Dans une suppression de cellules, nous obtenons un intervalle d'incertitude pour chaque cellule mise en suppression. Comment calculer l'intervalle d'incertitude? Dans un tel traitement, on suppose fréquemment que le tableau visé est indépendant des autres, mais s'il est lié aux autres, le programme de contrôle devrait idéalement tenir compte de cette relation. Une question difficile à se poser dans une politique est la suivante : quel degré d'incertitude devrait-on introduire de manière à répondre aux exigences de confidentialité : 5 % suffit-il (intervalle d'incertitude (0,95*v, 1,05*v) pour une valeur de cellule)? Faut-il plutôt prendre 10 % ou 20 %? Devrait-on avoir des limites supérieures et inférieures de protection qui soient égales ou encore très différentes de manière à rendre plus difficile la stratégie d'identification par détermination de point milieu? Bien sûr, même avec des points zu et zl égaux (pour le degré requis d'incertitude), l'incertitude effectivement créée par un programme peut être des plus asymétriques autour de la valeur réelle. Une autre manière de procéder est la protection par intervalle mobile. Dans ce cas, il n'y a pas de stratégie d'identification par point milieu qui vaille et, comme les règles de protection sont moins contraignantes, on met souvent moins de cellules en suppression (Kelly et coll., 1992).

5.1.2 Incertitude par intervalle : intervalles de traitement en arrondis

Récemment, Sande (2003) a étudié une façon de former des intervalles de protection pour des cellules « sensibles » selon un mode assimilable au traitement en arrondis (méthode dite de base de variable). La création de tels intervalles présente des avantages par rapport à la suppression. Plus précisément, l'utilisateur est immédiatement en mesure d'associer un intervalle d'incertitude (protection) à chaque cellule « sensible ». On peut aussi attribuer un intervalle d'incertitude à d'autres cellules comportant des erreurs (due ou non à l'échantillonnage) susceptibles d'être estimées. Ainsi, l'utilisateur n'a plus à calculer d'intervalles d'incertitude, il peut simplement les relever avec les valeurs des cellules. Ce mode d'expression de l'incertitude est proche de la façon classique d'exprimer une estimation sous la forme x ± erreur. Avec cette méthode, la subtilité réside dans la façon de calculer les intervalles asymétriques autour de la valeur réelle. Il faut procéder ainsi pour qu'une stratégie de détermination de point milieu ne puisse que rarement livrer des valeurs proches des valeurs réelles.

5.1.3 Ensembles finis d'incertitude

Avec certains types de perturbation ou de traitement en arrondis, l'utilisateur avancé peut effectuer une rétroanalyse (voir plus haut le traitement dit de retour arrière) et calculer une fonction de probabilité sur le jeu de valeurs possibles (ensembles d'incertitude). Ces ensembles sont souvent finis. La question qui se pose sur le plan des politiques est la suivante : que doit être la forme de cette fonction pour que la protection de confidentialité soit suffisante? Ainsi, supposons que, dans une perturbation de données d'ordre de grandeur, l'utilisateur soit capable de déterminer que la valeur réelle est t1 ou t2 avec une probabilité de 1/2 dans chaque cas. Ce degré d'incertitude suffit-il? Qu'en serait-il d'une fonction probabiliste plus asymétrique, Pr (t1) = 0.9, Pr (t2) = 0,1, par exemple? La notion d'entropie comme elle est employée dans les ouvrages qui traitent de la théorie de l'information pourrait constituer une mesure utile de l'incertitude. [Rappelons-nous qu'une telle entropie se calcule par sommation sur l'ensemble des probabilités i de (p (i) * (-log (p (i))), où le logarithme est en base 2. La quantité n'est pas négative et a pour borne supérieure log (n), n étant le nombre de probabilités non nulles associées aux valeurs des cellules.]

5.1.4 Mise en forme de cellules à valeurs modifiées

En dernière étape dans le traitement avant, on modifie les valeurs des cellules selon le degré d'incertitude à introduire. En suppression de cellules, une seule lettre, D par exemple, peut être portée dans chacune des cellules. Dans le traitement classique en arrondis, on présente uniquement la valeur de la cellule après arrondissement à la base la plus proche. Dans des types plus complexes de traitement en arrondis, il peut suffire de mentionner une valeur unique dans une cellule, mais on devra probablement ajouter en clair des indications sur la façon de constituer l'intervalle d'incertitude à partir de cette valeur.

5.2 Traitement arrière : estimation de la valeur de la cellule et de l'incertitude

Supposons que l'utilisateur se voie présenter un intervalle explicite comme dans le traitement en arrondis ou implicite comme dans la suppression. On peut poser une certaine densité de probabilité f (v) pour cet intervalle, puis établir toute estimation de la valeur de la cellule à partir de f (v). Comme on le fait couramment dans une analyse bayésienne, on peut supposer, s'il n'y a pas d'indications antérieures au sujet de f (v), qu'il y a densité de probabilité uniforme sur l'intervalle d'incertitude [a, b]. Il faut se rappeler que, dans ce cas, on a f (v) = 1/(b-a) et que la moyenne et la variance sont respectivement (a+b)/2 et $(b-a)^2/12$.

5.3 Exemple d'analyse avant et arrière d'un algorithme : cas de la correction contrôlée de tableaux

5.3.1 Introduction d'un degré d'incertitude

Étape 1 : Pour chaque cellule « sensible » (valeur v), on choisit de modifier à la hausse ou à la baisse avec une probabilité de 1/2, ce qui donne soit v1 = (1+p/100)*v, soit v2=(1-p/100)*v.
Étape 2 : On emploie un modèle d'optimisation semblable à celui qui sert à la suppression, semblable en ce sens qu'on résout pour les modifications de cellules dans un tableau additif et qu'il existe des valeurs limites pour les perturbations.

Contrairement à ce qui se fait dans un grand nombre de programmes répandus de suppression où les cellules « sensibles » font l'objet d'une protection séquentielle, toutes ces cellules sont fixées simultanément à leurs valeurs de protection dans la procédure de correction contrôlée de tableaux. Ainsi, il n'y a qu'un passage à calculer (c'est-à-dire une seule exécution du solutionneur en programmation linéaire ou en programmation par nombres entiers). Des valeurs non nulles passées par les cellules non sensibles sont ajoutées aux valeurs réelles. On fixe habituellement les limites à un faible pourcentage des valeurs des cellules (les perturbations seront donc légères).

5.3.2 Modification des valeurs des cellules

On change la valeur de chaque cellule sensible à v1 ou à v2 avec une probabilité de 1/2. Chaque cellule non sensible passe à une valeur de son intervalle borné. En général, les valeurs limites de perturbation doivent être supérieures aux modifications à apporter aux cellules sensibles. Le choix d'un multiplicateur de limite dépend de l'ampleur d'une perturbation qui enlèvera toute utilité aux valeurs des cellules. Ajoutons que, en fixant une limite supérieure pour les cellules non sensibles, moins de ces cellules devront être perturbées pour la protection des cellules sensibles.

5.3.3 Estimation de l'incertitude

Si l'utilisateur connaît la valeur de p et sait qu'une cellule est sensible, il doit deviner si celle-ci a été en perturbation positive ou négative. Ainsi, si vn=v1 (perturbation en hausse), va = vn/(1+p/100) ou, si vn=v2 (perturbation en baisse), va = vn/(1-p/100), où va est l'estimation par vn et la détermination au jugé d'une perturbation positive ou négative. Le degré d'incertitude est en réalité plus élevé, car (1) l'utilisateur pourrait ne pas connaître p ou (2) ignorer si la cellule est sensible ou a fait du tout l'objet d'une modification.

6. DÉTERMINATION DE L'UTILITÉ DES TABLEAUX APRÈS LE TRAITEMENT DE DIVULGATION

Comme un organise se propose, en diffusant des tableaux, de les mettre à la disposition d'un certain nombre ou peut-être d'un grand nombre d'utilisateurs à diverses fins, on doit examiner l'effet du traitement de divulgation sur différents ensembles de tableaux. On pourrait ainsi procéder à diverses expériences statistiques.

Dans l'analyse qui suit, nous employons le terme « modèle » en son sens général. Il est autant question d'utilisation simple de tableaux (lecture des valeurs de cellules désignées pour des comparaisons ou des calculs simples) que d'application de modèles statistiques classiques (modèles loglinéaires), par exemple.

6.1 Expériences statistiques de détermination de l'effet d'une méthode CDS sur un modèle fondé sur un tableau

Expérience statistique 1 : détermination de l'effet d'une méthode CDS sur un type de modèle

Supposons qu'un organisme emploie actuellement une seule méthode CDS – que nous appellerons D1 – pour des tableaux à valeurs d'ordre de grandeur, disons. Supposons aussi que l'utilisateur calcule un certain modèle – que nous appellerons M – fondé sur un tableau diffusé qui a subi un traitement de divulgation. Nous appellerons ce modèle

M(post-D1). Supposons enfin que l'utilisateur a acquis à titre spécial le droit de se rendre dans l'organisme statistique pour appliquer le même type de modèle au tableau comme celui-ci existait avant le traitement. En se reportant à ce tableau avant traitement, il produit un modèle M (pré-D1). Nous appellerons « dist » une façon logique de mesurer la distance (ou la différence) entre ces deux modèles de type M. Ainsi, dist pourrait mesurer la variation des valeurs des paramètres. Nous désignerons ainsi la distance ou la différence entre les deux modèles que nous venons de décrire :

dist(M(pré-D1), M(post-D1)).

Si la différence est significative, nous pouvons en conclure que la méthode D1 a une grande incidence sur la modélisation de type M qui s'applique à ce tableau. Le modélisateur pourrait juger qu'il lui faut utiliser le tableau pré-D1 ou un tableau ayant fait l'objet d'un traitement de divulgation par une autre méthode.

Expérience statistique 2 : comparaison des effets de deux méthodes CDS sur un type de modèle

Supposons que, pour un tableau, le modélisateur désire élaborer un modèle de type M. Supposons également qu'il a accès à deux versions du tableau. Le tableau avant traitement est le même pour les deux versions. Le tableau commun a fait l'objet d'un traitement de divulgation par les méthodes D1 et D2. L'utilisateur voudrait utiliser la méthode Di qui minimise dist(M(pré-Di), M(post-Di)) sur i = 1,2. (Comme ci-dessus, nous posons que l'utilisateur a acquis à titre spécial un droit d'accès au tableau initial.)

6.2 Utilisation de tableaux et méthodes CDS

6.2.1 Utilisation simple : lecture de quelques valeurs de cellules pour des comparaisons ou des calculs simples

6.2.1.1 Tableau en suppression

L'utilisateur se retrouve devant deux cas extrêmes :
(1) Meilleure estimation de l'organisme statistique
CDS-Unc(v) = 0, où CDS-Unc(v) est l'intervalle d'incertitude de la valeur v.
(2) Aucune valeur directe; symbole de suppression seulement
CDS-Unc(v) = Unc(intervalle de suppression de v), où on définit le côté droit comme l'incertitude liée à l'intervalle de suppression pour v.

6.2.1.2 Tableau en perturbation par correction contrôlée

Nous avons $0 \leq$ SDC-Unc(v) $\leq$ Est(Max-Pert), où Est(Max-Pert) est l'estimation par l'utilisateur de la plus grande perturbation d'une cellule. D'ordinaire, Est(Max-Pert) « Unc(intervalle de suppression de v), mais l'utilisateur se retrouve devant une incertitude non nulle introduite dans chaque cellule, même dans celles qui n'ont pas été perturbées.

6.2.2 Modélisation

6.2.2.1 Tableau en suppression

Il y a bien des modèles qui ne peuvent être élaborés avec des tableaux comportant des cellules vierges (données manquantes). Nombre de méthodes exigent des tableaux complets. Il est possible d'imputer de diverses manières les valeurs manquantes de cellules en suppression au moyen d'un algorithme de maximum de vraisemblance (ajustement proportionnel itératif) ou d'entropie maximale. Il faudra cependant pousser nettement la recherche pour juger de l'incidence des cellules vierges sur le modèle final.

6.2.2.2 Tableau en perturbation par correction contrôlée

Dans ce cas, le tableau est complet, mais il existe une légère incertitude de degré inconnu dans chaque cellule. Il s'agit de déterminer la grandeur de l'incidence des perturbations sur le modèle final. Il faut des expériences statistiques comme celles que nous avons décrites.

7. CADRE PROPOSÉ DE PRISE DE DÉCISIONS

1re décision :

Décrire les tableaux devant faire l'objet d'un traitement de divulgation.
Quelles données essayez-vous de protéger?
Quel type d'incertitude essayez-vous d'introduire?

(dans le cas d'un tableau à valeurs de dénombrement, le degré d'incertitude doit suffire à prévenir toute divulgation par inférence)

(dans le cas d'un tableau à valeurs d'ordre de grandeur, on doit veiller à ce que les concurrents d'une entreprise ne puissent estimer le chiffre d'affaires de celle-ci en toute précision; il faut un degré d'incertitude d'au moins 10 %, par exemple)

Voulez-vous protéger les données au niveau de l'entreprise ou seulement au niveau de l'établissement?
Quel degré d'incertitude désirez-vous introduire?

(dans le cas de données de dénombrement, quelle devrait être la valeur seuil?)

(dans le cas de données d'ordre de grandeur, quel degré d'incertitude devrait-on choisir?)

Les tableaux sont-ils liés les uns aux autres et, si oui, de quelle manière?
Les spécialistes du domaine ont-ils des préférences quant à l'absence de modification de certaines cellules par la méthode CDS si la chose est possible?

(les cellules marginales devraient-elles être fixes?)

2e décision :

Quel type de méthode de contrôle de divulgation statistique de tableaux pourrait-on employer compte tenu de la nature des données?
(1) remodelage du tableau (regroupement de catégories);
(2) traitement en arrondis (divers types);
(3) suppression;
(4) perturbation (divers types).

3e décision :

Comment va-t-on utiliser le tableau?

(y aura-t-il une utilisation simple comme la lecture de valeurs des cellules ou des modèles statistiques fondés sur le tableau? Dans ce dernier cas, de quels types de modèles se servira-t-on?)

4e décision :

Étant donné les décisions qui précèdent, quelles méthodes vous paraissent les meilleures possible?

On aura peut-être à lire des ouvrages ou des articles qui donnent un aperçu et/ou une analyse des diverses méthodes CDS du point de vue de l'utilisateur. Des documents qui comparent les méthodes seront particulièrement utiles (Salazar-Gonzalez, J.J., 2002, Russell et coll., 2002, Sande, 2003).

5e décision :

Quelle devrait être l'application de la méthode retenue? La réponse à cette question peut dépendre de la taille du tableau et de la nature des données.

(1) dans le cas d'une programmation par nombres entiers à l'aide d'un programme général appelable, on obtient souvent la meilleure réponse, mais pas assez vite pour respecter les contraintes de temps;
(2) dans une programmation par nombres entiers qui est adaptée au problème à traiter, on peut procéder bien plus rapidement qu'en programmation générale par nombres entiers (c'est la « procédure accélérée » par les techniques de séparation-évaluation);
(3) programmation linéaire comme procédure heuristique; il y a « relaxation » du problème de programmation par nombres entiers et, souvent, on obtient une bonne approximation;
(4) métaheuristique (on peut procéder par « recuit simulé », recherche avec liste de tabous, etc.).

6^{e} décision

Quel logiciel utiliser?
Doit-on concevoir son propre logiciel? Peut-on employer un logiciel existant?
Certains organismes statistiques ont des logiciels qu'ils mettent gratuitement à la disposition des utilisateurs. On peut les télécharger d'un site Web ou se les procurer sous forme de disque compact.
Exemple : voir **http://neon.vb.cbs.nl/casc/** pour l'examen d'un progiciel Tau-Argus téléchargeable (version Argus pour les tableaux).
Exemple : voir **http://www.fcsm.gov/committees/cdac/cdac.html** pour les liens avec le logiciel d'un organisme américain (dans un proche avenir).
Enfin, il y a dans le secteur privé des experts-conseils qui vendent des logiciels ou des services.

8. CONCLUSION

Quelle recherche faciliterait une meilleure comparaison des méthodes CDS?

Nous mentionnons trois facteurs qui ont un net caractère appliqué. Il s'agit du mode d'utilisation des tableaux, des erreurs de données et des tableaux liés les uns aux autres. La question la plus générale que nous avons abordée est celle de la détermination de l'incidence sur l'utilisateur de la décision prise par un organisme statistique de recourir à telle ou telle méthode CDS pour protéger un ensemble de tableaux. Comme nous l'avons dit, la difficulté est qu'il existe sans doute une grande diversité d'utilisations, depuis la simple lecture de quelques valeurs de cellules jusqu'à l'élaboration de modèles statistiques complexes. En première étape d'une telle analyse, il faut donc s'enquérir auprès des utilisateurs (ou des gens qui entrent en interaction avec eux) de leur façon d'utiliser les tableaux. On peut penser que bien des problèmes statistiques épineux se dégageraient de l'analyse et que certains seraient des plus intéressants. Ainsi, on pourrait s'attacher à l'effet d'une perturbation par correction contrôlée sur des tableaux à valeurs de dénombrement lorsque l'utilisateur élabore une certaine catégorie de modèles loglinéaires. Il serait possible de lier l'incertitude qu'introduit une méthode CDS dans les valeurs des cellules à l'incertitude propre aux coefficients d'un modèle loglinéaire déterminé. Cette tâche serait réalisable s'il existait un petit nombre de tels modèles et d'autres utilisations des tableaux à examiner.

Un autre facteur d'intérêt est le rôle que joue l'analyse des erreurs de données dans l'application d'une méthode CDS. Dans un tel exercice qui dépend étroitement des méthodes d'enquête classiques, l'organisme statistique pourrait calculer l'effet des erreurs d'échantillonnage et celles non dues à l'échantillonnage sur la fixation des niveaux de protection. Pour certaines méthodes, on pourrait aisément adapter ces niveaux aux erreurs des données existantes. Bien sûr, on suppose que l'organisme est capable d'établir au moins des estimations grossières de ces erreurs. Il faut préciser que cette estimation est parfois difficile.

On doit également étudier la façon de protéger des tableaux liés les uns aux autres pour chaque nouvelle méthode CDS. Il y a des méthodes dont l'application peut facilement s'étendre d'un tableau unique à des tableaux liés, du moins si tous les tableaux en question font l'objet d'un traitement de protection simultané. Avec d'autres méthodes cependant, il est impossible de traiter des tableaux liés ou, du moins, cet exercice n'a rien de facile. Même là où la description théorique de l'algorithme indique que la méthode est applicable à des tableaux liés, le logiciel d'exécution de la méthode peut ne pas avoir cette capacité. Dans le cas des méthodes de traitement simultané d'un jeu de tableaux liés, il reste à savoir comment procéder avec des tableaux liés subissant des passages de traitement non simultanés. On pourrait en venir à une

conclusion générale au sujet du traitement possible de tableaux liés de sorte que chaque nouvelle méthode CDS puisse rapidement s'évaluer en fonction d'une telle capacité.

Dans notre exposé, nous avons seulement survolé les facteurs d'algorithmique des méthodes CDS et de leurs logiciels d'application. On constate heureusement que ces thèmes sont amplement examinés lorsqu'on introduit une méthode nouvelle ou qu'on découvre de nouvelles façons d'appliquer une méthode existante.

REMERCIEMENTS

J'aimerais remercier mon collègue Jim Fagan, du Census Bureau, Jose Dula, de l'Université du Mississippi (qui a séjourné un an au Census Bureau), et Steve Roehrig, de l'Université Carnegie Mellon, pour leur examen d'un grand nombre de sujets abordés dans le présent document.

RÉFÉRENCES

Cox, L.H., et R.A. Dandekar (2002), "A Disclosure Limitation Method For Tabular Data That Preserves Data Accuracy and Ease-of-use", présenté au Federal Comm. Stat. Methodology Conf., Nov. 2002.

Cox, L.H. et J. Kelly (2003), "Balancing Data Quality and Confidentiality for Tabular Data", Recueil de la session conjointe de travail 2003 Eurostat/Commission économique pour l'Europe sur la confidentialité des données, Luxembourg.

Dandekar, R.A. (2003), "Cost Effective Implementation of Synthetic Tabulation (a.k.a. Controlled Tabular Adjustments) in Legacy and New Statistical Data Publication Systems", Mars 2003, rapport non publié.

Dandekar, R.A. et L.H. Cox (2002), "Controlled tabular adjustment: an alternative to complementary cell suppression for disclosure limitation of tabular data", rapport non publié.

Dula, J., Massell, P., Fagan, J., (2003), "Statistical Disclosure Control for Tabular Data", rapport non publié.

Evans, T., Zayatz, L., Slanta, J. (1998), "Using Noise for Disclosure Limitation Establishment Tabular Data", Journal of Official Statistics, Décembre, 1998.

Fienberg, S. (2003), "Statistical Disclosure Limitation: Releasing Useful Data for Statistical Analysis", présenté au Bureau of Transportation Statistics, 28 avril, 2003, http://www.bts.gov/confidentiality_seminar_series/2003_04/

Fischetti, M., J.J. Salazar Gonzalez (2000), "Models and Algorithms for Optimizing Cell Suppression in Tabular Data with Linear Constraints", J. Amer. Stat. Assn., v.95, no.451, pp. 916-928; Sept 2000.

Glover, F. (1990), "Tabu Search: A Tutorial", Interfaces 20:4 Juillet-Août 1990 (pp. 74-94)

Jewett, R.(1993), "Disclosure Analysis for the 1992 Economic Census", rapport non publié du Recensement 1993.

Kelly, J.P., Golden, B.L., Assad, A.A. (1992), "Cell Suppression: Disclosure Protection for Sensitive Tabular Data", Networks, Vol. 22 (1992), p.397-417.

Massell, P. (2001), "Cell Suppression and Audit Programs used for Economic Magnitude Data," rapport de la division de recherche statistique, U.S. Census Bureau, http://www.census.gov/srd/papers/pdf/rr2001-01.pdf

Russell, J. N. et J. P. Kelly (2003), "Bureau of Transportation Statistics' Prototype Disclosure Limitation Software for Complex Tabular Data", Recueil de la session conjointe de travail 2003 Eurostat/Commission économique pour l'Europe sur la confidentialité des données, Luxembourg.

Russell, J. N., J. P. Kelly, et F. Glover (2002), “The Bureau of Transportation Statistics’ Statistical Disclosure Limitation Method for Tabular Data: A Review”, 2002 Proc. of Amer. Stat. Assn., Government Stat. Sect. [CD-ROM], Alexandria, VA.

Sande, G. (2003), “A Less Intrusive Variant on Cell Suppression to Protect the Confidentiality of Business Statistics”, rapport non publié.

Salazar-Gonzalez, J.J. (2002), “A Unified Mathematical Programming Framework for different Statistical Disclosure Limitation Methods for Tabular Data”, atelier de CASC, http://webpages.ull.es/users/casc/#Working %20papers %20and %20articles

Willenborg, L., et T. de Waal (1996), *Statistical Disclosure Control in Practice*, Lecture Notes in Statistics, v. 111, Springer, 1996.

Willenborg, L., et T. de Waal (2001), *Elements of Statistical Disclosure Control*, Lecture Notes in Statistics, v. 155, Springer, 2001.