No 11-522-XIF au catalogue

La série des symposiums internationaux de Statistique Canada - Recueil

Symposium 2005 : Défis méthodologiques reliés aux besoins futurs d'information

2005

Statistique Canada Statistics Canada

Canada

Recueil du Symposium 2005 de Statistique Canada
Défis méthodologiques reliés aux besoins futurs d'information

AJUSTEMENT TABULAIRE CONTRÔLÉ PRÉSERVANT LA QUALITÉ : UNE MÉTHODE DE RÉSOLUTION DES PROBLÈMES DE PROTECTION DES RENSEIGNEMENTS CONFIDENTIELS ET DE QUALITÉ DES DONNÉES POUR LES DONNÉES TABULAIRES

Lawrence H. Cox, Ph.D.[1]

RÉSUMÉ

Parmi les méthodes traditionnelles, dont l'objectif est de limiter la divulgation statistique dans les données tabulaires, on trouve la suppression de cellules, l'arrondissement de données et la perturbation de données. Toutes ces méthodes s'appliquent aux données de dénombrement, tandis que seule la suppression de cellules est efficace dans le cas d'observations quantitatives. La suppression de cellules élimine des données utiles par ailleurs, contrecarrant toute analyse directe de données. Puisque le mécanisme de suppression n'est pas descriptible en termes probabilistes, les tableaux supprimés ne peuvent être traités au moyen de méthodes statistiques telles que l'imputation. C'est pourquoi les caractéristiques de la qualité des données des tableaux supprimés sont faibles. On propose donc une solution de remplacement à la suppression de cellules, soit la méthode de l'ajustement tabulaire contrôlé (ATC). Cette méthode remplace les cellules tabulaires à l'origine de la divulgation par des valeurs sans risque de divulgation. Elle utilise la programmation linéaire ATC pour ajuster les valeurs restantes et rééquilibrer les équations tabulaires, et fait le tout de manière optimale en prenant en compte toute une gamme de mesures globales et locales liées à la« proximité » des données. Nos recherches récentes étendent l'utilisation de l'ATC pour en faire un *ATC préservant la qualité*, c'est-à-dire que son utilisation peut non seulement protéger la confidentialité, mais aussi préserver d'importants paramètres et statistiques de distribution dans un contexte à une variable et à plusieurs variables. Plus précisément, l'ATC préservant la qualité à une variable permet de garantir que les données ajustées préservent d'une manière approximative les moyennes et variances des données originales et que les données originales et ajustées affichent une corrélation positive élevée. Tout en préservant les propriétés univariées, l'ATC multivarié préserve les corrélations et les régressions entre deux variables issues de données originales. Toutes ces méthodes reposent sur des formes de programmation linéaire qui sont faciles à mettre en œuvre ou à améliorer.

MOTS CLÉS: Ajustement tabulaire contrôlé, programmation linéaire, covariance.

1. INTRODUCTION

Les données tabulaires sont très répandues et s'avèrent des éléments importants de la statistique officielle. La confidentialité des données a été étudiée pour la première fois dans le cas de données tabulaires (Fellegi 1972; Cox 1980). Les données tabulaires sont additives et sont donc naturellement reliées aux systèmes spécialisés d'équations linéaires : **TX = 0**, où **X** représente les *cellules tabulaires* et **T**, les *équations tabulaires*, les entrées de **T** font partie de l'ensemble {-1, 0, +1}, et chaque rangée de **T** contient précisément un -1.

Une méthodologie récente de *restriction de la divulgation statistique* (U.S. Department of Commerce, 1994) relativement aux données tabulaires est connue sous le nom d'*ajustement tabulaire contrôlé* (ATC). Son élaboration a été motivé par la complexité algorithmique, les obstacles analytiques et l'insatisfaction généralisée des utilisateurs face à la méthodologie dominante, soit la *suppression complémentaire de cellules* (Cox, 1980, 1995). La suppression complémentaire retire des publications toutes les *cellules confidentielles* – celles qui ne peuvent être publiées en raison des préoccupations de confidentialité – et des cellules non confidentielles afin de garantir que les valeurs des cellules confidentielles ne peuvent être reconstruites ou estimées de près par la manipulation des relations linéaires tabulaires. Pour les besoins de l'analyse statistique, la suppression de cellules comporte deux grands inconvénients : le retrait d'information utile par ailleurs et la difficulté à analyser les systèmes tabulaires en raison de valeurs manquantes non aléatoire dans les cellules. L'ATC remplace les valeurs confidentielles des cellules par des *valeurs sans risque de divulgation*, c'est-à-dire par des valeurs suffisamment distantes de la vraie valeur. Parce que les ajustements abîment presque certainement l'addivité du système tabulaire, l'ATC ajuste toutes les cellules non confidentielles, ou certaines d'entre elles, par de petites quantités afin de rétablir l'additivité. L'ATC est mis en

[1] Lawrence H. Cox, Ph.D. U.S. National Center for Health Statistics, LCOX@CDC.GOV

application au moyen des méthodes de programmation mathématique proposées par les logiciels commerciaux. Cox (2000) fournit le premier modèle mathématique d'ATC dans la littérature scientifique.

La principale question entourant l'ATC concerne le degré auquel il fausse les résultats analytiques. Autrement dit, les analyses fondées sur des données ajustées sont-elles dans un certain sens équivalentes à celles reposant sur des données originales? Nous démontrons dans ces pages comment il est possible d'utiliser l'ATC et la programmation linéaire afin de préserver les moyennes, les variances, les covariances, les corrélations et le coefficient de régression des données à une et à plusieurs variables. On peut obtenir davantage de détails sur les cas univariés dans Cox et Kelly (2004) et sur les cas multivariés dans Cox, Kelly et Patil (2004). Nos modèles s'appuient sur la programmation linéaire et sont donc faciles à concevoir et à utiliser, tout en s'appliquant à une gamme étendue de problèmes.

La section 2 présente un résumé de la méthodologie ATC originale (Cox, 2000; Dandekar et Cox, 2002). La section 3 fournit les méthodes linéaires pour les cas univariés qui permettent de préserver les moyennes et les variances des données originales et d'assurer une corrélation élevée entre les données originales et ajustées (Cox et Kelly, 2004). La section 4 élargit le paradigme au cas multivariés , fournissant des formes de programmation linéaire qui garantissent la préservation, dans les données ajustées, des covariances et des corrélations provenant des données originales (Cox, Kelly et Patil, 2004). Dans la section 5, nous présentons les résultats informatiques. La section 6 est consacrée au mot de la fin.

2. LA MÉTHODOLOGIE DE L'AJUSTEMENT TABULAIRE CONTRÔLÉ

L'ATC s'applique aux données tabulaires, quelles que soient leurs formes. Mais pour des raisons pratiques, nous nous concentrons sur les données quantitatives, où l'on constate les avantages les plus importants. Un paradigme simple de divulgation statistique de données quantitatives s'élabore ainsi : soit une cellule de tabulation, appelée *i*, comprend *k* répondants (par ex., des magasins de vêtements dans un comté) et leurs données (p. ex., données sur les ventes au détail et l'emploi). L'organisme statistique national (OSN) suppose que tout répondant connaît l'identité des autres répondants. La *valeur de la cellule* est la valeur totale de la statistique étudiée (par ex., les ventes au détail totales), additionnée selon les *contributions* (non négatives) à cette statistique de chaque répondant dans la cellule. Notons la valeur de la cellule par $v^{(i)}$ et les contributions des répondants par $v_j^{(i)}$, *qui sont placé*es par ordre décroissant. Il est possible pour n'importe quel répondant *J* de calculer $v^{(i)} - v_J^{(i)}$ qui fournit l'estimation supérieure de la contribution de tout autre répondant. Cette estimation est la plus rapprochée, en termes de pourcentage, lorsque J = 2 et j = 1. Une règle de divulgation standard, la *règle du p-pour cent*, stipule que la valeur de la cellule représente une *divulgation* si l'estimation se rapproche à moins de p pour cent de la contribution la plus grande. Les cellules confidentielles sont justement celles qui ne remplissent pas cette condition.

L'OSN peut aussi supposer que tout répondant peut avoir recours à la *notoriété publique* pour estimer la contribution de tout autre répondant à q pour cent près (q > p, par ex., q = 50 %). Cette information supplémentaires permet au deuxième en importance d'estimer $v^{(i)} - v_1^{(i)} - v_2^{(i)}$, soit la somme de toutes les contributions excluant la sienne et la plus importante, à q pour cent près. Cette estimation supérieure fournit au deuxième en importance une estimation inférieure de $v_1^{(i)}$. Les *limites inférieure* et *supérieure de protection* pour les valeurs des cellules sont respectivement égales au montant minimal qui doit être soustrait de la valeur de la cellule (ou lui être ajouté) de manière à ce que ces estimations inférieures (et supérieures) soient éloignées d'au moins p pour cent de la vraie valeur $v_1^{(i)}$. Les valeurs numériques sous la borne de protection inférieure et au-dessus de la borne supérieure sont des *valeurs sans risque de divulgation* pour la cellule. Une pratique commune de l'OSN consiste à supposer que toutes ces bornes de protection sont égales, à p_i. La suppression de cellules complémentaires supprime toutes les cellules confidentielles de la publication, remplaçant les valeurs confidentielles par des variables du système tabulaire **TX = 0**. Puisqu'il est presque certain qu'on peut estimer une ou plusieurs valeurs des cellules confidentielles supprimées à p pour cent de sa vraie valeur, il est nécessaire de supprimer certaines cellules non confidentielles jusqu'à ce qu'aucune estimation de cellules confidentielles ne soit à l'intérieur de p pour cent.

L'ajustement tabulaire contrôlé remplace chaque valeur confidentielle par une valeur sans risque de divulgation. Cela représente une amélioration par rapport à la suppression de cellule complémentaire, étant donné qu'on remplace un symbole de suppression par une valeur réelle. Cependant, les valeurs sans risque de divulgation ne sont pas nécessairement des estimations non biaisées des vraies valeurs. Afin de minimiser le biais, Dandekar et Cox (2002)

remplacent la vraie valeur par une des bornes de protection, $v^{(i)} - p_i$ ou $v^{(i)} + p_i$, c'est-à-dire par une des deux valeurs sans risque de divulgation étant les plus rapprochées de la valeur originale. Parce que ces attributions abîment presque certainement l'additivité du système tabulaire, l'ATC ajuste les valeurs non confidentielles afin de rétablir l'additivité. Du fait que les choix consistant à ajuster chaque valeur confidentielle à la baisse ou à la hausse sont binaires, ces étapes une fois combinées forment un système de programmation linéaire entier mixte [MILP] (Cox, 2000). En utilisant des heuristiques pour effectuer les choix binaires, l*e système de programmation linéaire avec relaxation* est facilement résolu.

Un système de programmation linéaire (entier mixte) ne garantit pas en soi que les propriétés analytiques des données originales et ajustées sont comparables. Cox et Dandekar (2003) abordent ces enjeux de trois manières. Premièrement, les valeurs confidentielles sont remplacées par des valeurs sans risque de divulgation aussi rapprochées que possible. Deuxièmement, les bornes (*capacités*) inférieure et supérieure sont imposées à la variation des valeurs non confidentielles afin de s'assurer que les ajustements apportés à chaque donnée individuelle sont suffisamment petits. Les capacités statistiques seraient, par exemple, basées sur une erreur de mesure estimée pour chaque cellule e_i. Troisièmement, le programme linéaire est optimisé en prenant en considération une mesure globale de distortion des données comme la somme minimale des ajustements absolus ou des ajustements absolus en pourcentage, comme suit.

Supposons qu'il y a *n* cellules dans un tableau parmi lesquelles les *s* premières sont confidentielles, que les données originales sont représentées par le vecteur nxl **a** et les données ajustées, par $a + y^+ - y^-$ et que $y = y^+ - y^-$. Le MILP de Cox (2000) qui minimise la somme des ajustements absolus est :

$$\min \sum_{i=1}^{n} (y_i^- + y_i^+) \qquad \text{compte tenu de :} \quad I_i \text{ binaire, i = 1, ..., s}$$

$$T(y) = 0 \tag{1}$$

$$y_i^- = p_i(1 - I_i), \quad y_i^+ = p_i I_i \qquad \text{i = 1, ... , s}$$

$$0 \le y_i^-, \; y_i^+ \le e_i \qquad \text{i = s+1, ..., n}$$

Les contraintes de Cox et Dandekar (2003) sont utiles. Malheureusement, les choix de mesure d'optimisation sont limités aux fonctions linéaires. Dans les deux prochaines sections, nous élargissons ce paradigme dans deux directions distinctes, en nous concentrant sur des démarches permettant de préserver la moyenne, la variance, la corrélation et la régression entre les données originales et ajustées.

La formulation (1) est un programme linéaire entier mixte, dont la partie entière peut être résolue au moyen de méthodes exactes dans les problèmes de petite ou moyenne envergure, ou au moyen d'heuristiques qui fixent les variables entières pour ensuite procéder à la programmation linéaire avec relaxation (Dandekar et Cox, 2002; Cox et Kelly, 2004). La suite du présent article se concentre sur le problème de la préservation de la qualité des données lors d'un ATC, et ne se préoccupe pas des modalités de résolution de la portion entière.

3. L'UTILISATION DE L'ATC POUR PRÉSERVER LES RELATIONS À UNE VARIABLE

Nous présentons les formulations de programmation linéaire permettant de préserver exactement ou approximativement la moyenne, la corrélation et la pente de régression entre les données originales et ajustées.

La préservation des valeurs moyennes est simple. Toute valeur de cellule $\mathbf{a_i}$ peut être maintenue fixe en forçant les variables d'ajustement correspondantes y_i^+, y_i^- à zéro, c'est-à-dire en forçant la capacité supérieure de chaque variable à zéro. Les moyennes sont calculées à partir de sommes. Par exemple, pour établir la moyenne générale, il suffit de calculer le total général. Afin d'établir la moyenne de n'importe quel ensemble de variables pour lequel une variable correspondante n'a pas encore été définie, il faut incorporer une nouvelle contrainte dans le système linéaire : $\sum(y_i^+ - y_i^-) = 0$, où la somme est calculée à partir de l'ensemble des variables d'intérêt. Le MILP correspondant est :

$$\min c(\mathrm{y}) \text{ compte tenu de :} \qquad I_i \text{ binaire, i = 1, ..., s}$$

$$\mathrm{T}(\mathrm{y}) = 0 \qquad (2)$$

$$\sum_{i=1}^{s} (y_i^+ - y_i^-) = 0$$

$$p_i(1 - I_i) \le y_i^- \le q_i(1 - I_i), \quad p_i I_i \le y_i^+ \le q_i I_i \qquad \text{i = 1, ... , s}$$

$$0 \le y_i^-, \; y_i^+ \le e_i \qquad \text{i = s+1, ..., n}$$

$c(\mathrm{y})$ sert à maintenir les ajustements près de leur borne inférieure, par exemple, $c(\mathrm{y}) = \sum \mathrm{y}^+ + \mathrm{y}^-$.

Pour la variance, tout sous-ensemble de cellules de dimension t avec $\bar{y} = 0$,

$$Var(\mathrm{a} + \mathrm{y}) = (1/\mathrm{t})(\Sigma((\mathrm{a_i} + \mathrm{y_i} - (\bar{\mathrm{a}} + \bar{\mathrm{y}}))^2)$$

$$= \mathrm{Var(a)} + (2/\mathrm{t}) \Sigma(\mathrm{a_i} - \bar{\mathrm{a}})\, \mathrm{y_i} + \mathrm{Var(y)}$$

Définissez $L(\mathrm{y}) = \mathrm{Cov(a, y)/Var(a)}$. Étant donné que $\bar{y} = 0$, alors $L(\mathrm{y}) = (1/(\mathrm{tVar(a)})) \sum_{\mathrm{i=1}}^{\mathrm{t}} (\mathrm{a_i} - \bar{\mathrm{a}})\, \mathrm{y_i}$, donc

$$Var(\mathrm{a} + \mathrm{y})/\mathrm{Var(a)} = 2\mathrm{L(y)} + (1 + \mathrm{Var(y)/Var(a)}) \text{ et}$$

$$|Var(\mathrm{a} + \mathrm{y})/\mathrm{Var(a)} - 1| = |2\mathrm{L(y)} + (\mathrm{Var(y)/Var(a)})|$$

Ainsi, il suffit de minimiser $|L(\mathrm{y})|$; de la manière suivante :

a) Incorporer deux nouvelles contraintes linéaires dans le système (2) :
$$w \ge L(\mathrm{y}), \; \mathrm{w} \ge -\mathrm{L(y)} \qquad (3)$$
b) minimiser *w*

En ce qui concerne la corrélation, l'objectif consiste à obtenir une corrélation positive élevée entre les valeurs originales et ajustées. Nous cherchons Corr (**a, a + y**) = 1, de manière exacte ou approximative. Lorsque $\bar{y} = 0$,

$$Corr(a, a+y) = Cov(a, a+y) / \sqrt{Var(a) Var(a+y)}$$

$$= (1 + L(y)) / \sqrt{\mathrm{Var(a + y)/ Var(a)}}$$

Puisque Var(**y**)/Var(**a**) est typiquement petit, le dénominateur devrait être près de un et le $\min |L(\mathrm{y})|$ sujet à (2) devrait permettre de bons résultats en termes de préservation de la corrélation. Il convient de noter qu'un dénominateur égal à un équivaut à préserver la variance, ce qui, comme nous l'avons vu, peut être accompli par $\min |L(\mathrm{y})|$.

Enfin, nous cherchons à préserver la régression des moindres carrés $Y = \beta_1 X + \beta_0$ des données ajustées Y = **a** + **y** par rapport aux données originales X = **a**, c'est-à-dire que nous voulons que β_1 soit près de un et que β_0 soit près de zéro.

$$\beta_1 = Cov(\mathrm{a} + \mathrm{y}, \mathrm{a}) / \mathrm{Var(a)} = 1 + \mathrm{L(y)}, \; \beta_0 = (\bar{\mathrm{a}} + \bar{\mathrm{y}}) - \beta_1 \bar{\mathrm{a}}$$

Sachant que $\bar{y} = 0$, alors $\beta_0 = 0$, $\beta_1 = 1$ chaque fois qu'il est possible que L(y) = 0. Encore une fois, cela revient à min $|L(\mathrm{y})|$ compte tenu des contraintes de (2), c'est-à-dire à (3).

4. L'UTILISATION DE L'ATC DANS LE BUT DE PRÉSERVER LES RELATIONS À PLUSIEURS VARIABLES

À la place d'un simple ensemble de données de données **a** organisées sous forme tabulaire, c'est-à-dire **Ta = 0**, auquel des ajustements **y** doivent être apportés pour les besoins de la confidentialité, nous allons désormais substituer des ensembles de données multiples qui sont tous organisés selon la même structure tabulaire commune **T**. Pour favoriser une démarche concrète, nous nous concentrons sur le cas à deux variables. Les données originales sont désignées par **a** et **b** et les ajustements correspondants portés aux valeurs originales sont désignés par les variables **y** et **z**. Dans une situation à plusieurs variables, la préservation de la covariance et de la variance est cruciale. Autrement dit, si nous pouvons préserver les valeurs moyennes et la matrice de variances-covariances des données originales, nous aurons donc préservé les propriétés essentielles des données originales, particulièrement dans le cas de modèles statistiques linéaires. Nous aimerions aussi préserver la régression linéaire simple des données originales **b** sur les données originales **a** dans les données ajustées. Voilà les objectifs de cette section.

4.1 La préservation de la matrice de variances-covariances

Les copies séparées du modèle (3) de la section précédente préservent les variances univariées Var (**a**) et Var (**b**). Afin de préserver Cov (**a,b**), il faut que :

$$\begin{aligned} \text{Cov}(\mathbf{a}, \mathbf{b}) &= \text{Cov}(\mathbf{a} + \mathbf{y}, \mathbf{b} + \mathbf{z}) \\ &= \text{Cov}(\mathbf{a}, \mathbf{b}) + \text{Cov}(\mathbf{a}, \mathbf{z}) + \text{Cov}(\mathbf{b}, \mathbf{y}) + \text{Cov}(\mathbf{y}, \mathbf{z}) \end{aligned}$$

Conséquemment, nous cherchons une solution précise ou approximative à :

$$\min |\text{Cov}(\mathbf{a}, \mathbf{z}) + \text{Cov}(\mathbf{b}, \mathbf{y}) + \text{Cov}(\mathbf{y}, \mathbf{z})| \text{ , compte tenu de (3)} \qquad (4)$$

Une démarche linéaire de résolution de (4) consiste à exécuter des optimisations linéaires en alternance, c'est-à-dire à résoudre (2) pour $\mathbf{y} = \mathbf{y_0}$, à substituer $\mathbf{y_0}$ dans (4) et résoudre pour $\mathbf{z} = \mathbf{z_0}$ et à continuer ainsi jusqu'à ce qu'une solution acceptable soit obtenue.

4.2 La préservation du coefficient de régression linéaire simple

Notre objectif est de préserver le coefficient estimé de régression lors de la régression linéaire simple de **b** sur **a**. Nous n'abordons pas ici les enjeux connexes de la préservation de l'écart-type de l'estimation et de la qualité de l'ajustement. Nous cherchons exactement ou approximativement :

$$\text{Cov}(\mathbf{a}, \mathbf{b}) / \text{Var}(\mathbf{a}) = \text{Cov}(\mathbf{a} + \mathbf{y}, \mathbf{b} + \mathbf{z}) / \text{Var}(\mathbf{a} + \mathbf{y})$$

$$\text{Var}(a + y) / \text{Var}(a) = \text{Cov}(a + y, b + z) / \text{Cov}(a, b)$$
$$= 1 + \text{Cov}(a, z) / \text{Cov}(a, b) + \text{Cov}(b, y) / \text{Cov}(a, b) + \text{Cov}(y, z) / \text{Cov}(a, b)$$

Un rappel : $\text{Var}(a + y)/\text{Var}(a) = 2L(y) + 1 + \text{Var}(y)/\text{Var}(a)$

$$2L(y) + \text{Var}(y)/\text{Var}(a) = \text{Cov}(a, z) / \text{Cov}(a, b) + \text{Cov}(b, y) / \text{Cov}(a, b) + \text{Cov}(y, z) / \text{Cov}(a, b)$$

afin de préserver les propriétés à une variable L(y) = 0 et la covariance à deux variables Cov (y, z) = 0 de manière exacte ou approximative. Donc, si nous cherchons en plus à préserver le coefficient de régression, nous devons satisfaire le programme linéaire :

$$\min |(\text{Cov}(\mathbf{a}, \mathbf{z}) + \text{Cov}(\mathbf{b}, \mathbf{y})) / \text{Cov}(\mathbf{a}, \mathbf{b})|, \text{ compte tenu de (4)} \qquad (5)$$

Lors de la mise en œuvre, l'objectif est représenté par une contrainte stipulant que la valeur absolue doit être près de zéro.

4.3 La préservation des corrélations

L'objectif ici est de s'assurer que les corrélations entre les variables calculées à partir de données ajustées se

rapprochent, en valeur, des corrélations basées sur les données originale, c'est-à-dire que Corr (**a, b**) = Corr (**a + y, b + z**) de manière exacte ou approximative. Après quelques calculs algébriques, la préservation de la corrélation équivaut à satisfaire, de manière exacte ou approximative :

$$\sqrt{\frac{Var(a+y)}{Var(a)}}\sqrt{\frac{Var(b+z)}{Var(b)}} = \frac{Cov(a+y,b+z)}{Cov(a,b)}$$

Dans bien des cas, la corrélation est préservée grâce aux méthodes déjà incluses en (5) et conçues pour préserver à la fois les variances et la covariance à une variable. Autrement, il peut s'avérer utile d'avoir recours à l'itération qui vise à contrôler le produit du deuxième membre.

5. RÉSULTATS DES SIMULATIONS INFORMATIQUES

Nous avons testé notre méthode sur trois tableaux à deux dimensions afin d'analyser la performance des formes linéaires proposées sur des mesures statistiques à une comme à deux variables. Trois tableaux ont été tirés d'un tableau à trois dimensions 4 x 9 x 9. Ce tableau contient des données quantitatives réelles, et la divulgation a été définie par la règle de dominance (1 contributeur, 70 %), c'est-à-dire qu'une cellule est confidentielle si la plus grande contribution excède 70 % de la valeur de la cellule. Il en résulte les niveaux de protection $p_i = (v_1^i)/0.7 - v^i$. Les tableaux A, B et C contiennent respectivement 6, 5 et 4 cellules confidentielles. Les limites supérieures (capacités) pour ajuster les cellules confidentielles et non confidentielles ont été fixées à 20 % de la valeur de la cellule.

Premièrement, nous avons utilisé la forme MILP pour calculer les solutions exactes pour les trois cas AB, AC et BC. La forme MILP a fait appel aux contraintes pour la préservation de la moyenne et de la variance et à un objectif qui minimise la variation de la covariance, afin de préserver les mesures univariées et multivariées.

Le tableau 1 présente les résultats pour les mesures de la covariance, de la corrélation, du coefficient de régression, de la variance des données A et de la variance des données B. Les valeurs du tableau sont des variations en pourcentage par rapport à la mesure statistique originale. Les moyennes ont été préservées dans les trois cas.

Cas	Variation de la covariance	Variation de la corrélation	Variation du coeff. régress.	Variation de la variance i	Variation de la variance j	Coeff. corr. original
AB	3,15	1,09	5,94	-3,22	6,2	0,77
AC	1,13	2,63	1,14	-2,43	0,1	0,40
BC	3,6	6,12	6,7	-3,6	-1,89	0,49
Moyenne	2,62	3,28	4,59	-3,08	1,47	0,55

Tableau 1: Résultats des formes linéaires sur les mesures statistiques (variation en pourcentage)

L'*heuristique de classement* (Dandekar et Cox, 2002) classe tout d'abord les cellules confidentielles par ordre décroissant et attribue l'orientation des cellules confidentielles de manière alternative. Cette méthode se veut un moyen de trouver de bonnes solutions en matière d'écart absolu des cellules d'une manière efficace sur le plan informatique. Nous avons étudié sa capacité de préserver des mesures statistiques en comparant les résultats de cette méthode avec ceux de la méthode MILP exacte et avec une solution optimale (non linéaire). Le tableau 2 présente cette comparaison. Étant donné la variation en ce qui concerne la qualité des solutions pour les mesures de la covariance et de la variance, la performance de l'heuristique de classement était bonne. En fait, l'heuristique de classement a amélioré la performance en matière d'écart absolu des cellules. Ce résultat est cohérent avec les publications, ce qui démontre la supériorité de la performance de l'heuristique de classement sur l'écart des cellules.

Méthode de résolution	Variation de la covariance	Variation de la corrélation	Variation de la variance A	Variation de la variance B	Écart absolu des cellules
Exacte	3,15	1,09	5,94	-3,22	8,81e+7
Heuristique de	5,34	2,19	4,49	-4,56	8,78e+7

classement					
Performance par rapport à la solution optimale	69	100	-24	41	-0,34

Tableau 2 : Résultats de l'heuristique de classement (en pourcentage)

6. LE MOT DE LA FIN

Les progrès des deux dernières décennies ont débouché sur plusieurs méthodes de limitation de la divulgation dans les données tabulaires. Parmi celles-ci, l'ajustement tabulaire contrôlé génère le produit le plus utilisable, multipliant ainsi les occasions d'analyser les données publiées. La question consiste maintenant à savoir dans quelle mesure des données ajustées préservent les résultats analytiques des données originales. Cox et Kelly (2003) se sont penchés sur cet enjeu dans le cas des variables uniques. Cet article a présenté des équations linéaires efficaces qui permettent de préserver les statistiques essentielles dans les cas des variables uniques et multiples, comme en témoigne également Cox (2004).

RÉFÉRENCES

Cox, L.H. (1980), "Suppression methodology and statistical disclosure control", *Journal of the American Statistical Association* 75, 377-385.

Cox, L.H. (1995), "Network models for complementary cell suppression", *Journal of the American Statistical Association* 90, 1153-1162.

Cox, L.H. (2000), Discussion. *ICES II: The Second International Conference on Establishment Surveys: Survey Methods for Businesses, Farms and Institutions*, Alexandria, VA: American Statistical Association, 905-907.

Cox, L.H. (2004), "Resolving Confidentiality and Quality Issues for Tabular Data", *European Conference on Quality and Methodology in Official Statistics (Q 2004): Conference Papers*, Weisbaden: Statistisches Bundesamt, 2004, Paper 28.01, CD ROM.

Cox, L.H. et Dandekar, R.A. (2003), "A new disclosure limitation method for tabular data that preserves data accuracy and ease of use", *Proceedings of the 2002 FCSM Statistical Policy Seminar*, Washington, DC: U.S. Office of Management and Budget (en cours d'impression).

Cox, L.H. et Kelly, J.P. (2004), "Balancing data quality and confidentiality for tabular data", *Proceedings of the UNECE/Eurostat Work Session on Statistical Data Confidentiality, Luxembourg, 7-9 April 2003, Monographs of Official Statistics*, Luxembourg: Eurostat (en cours d'impression).

Cox, L.H., Kelly, J.P. et Patil, R. (2004), "Balancing quality and confidentiality for multivariate tabular data", *Lecture Notes in Computer Science 3050*, New York: Springer Verlag (en cours d'impression).

Dandekar, R.A. et Cox, L.H. (2002), "Synthetic tabular data: an alternative to complementary cell suppression" (manuscrit).

Fellegi, I.P. (1972), "On the question of statistical confidentiality", *Journal of the American Statistical Association* 67, 7-18.

Fischetti, M. et Salazar-Gonzalez, J.J. (2000), "Models and algorithms for optimizing cell suppression in tabular data with linear constraints", *Journal of the American Statistical Association* 95, 916-928.

U.S. Department of Commerce (1994), "Statistical Disclosure and Disclosure Limitation Methods", Statistical Policy Working Paper 22, Washington, DC: Federal Committee on Statistical Methodology.