TRAITEMENT DE LA CONFIDENTIALITÉ STATISTIQUE DANS LES TABLEAUX : EXPÉRIENCE DE LA DIRECTION DES STATISTIQUES D'ENTREPRISES

Julien NICOLAS (*)

(*) Insee, Unité Méthodologie Statistique d'Entreprises

Introduction

Le problème du traitement de la confidentialité statistique dans les tableaux est désormais une question que les chargés d'enquêtes prennent en compte dans leur cahier des charges, en amont de l'enquête de plus en plus souvent.
La création d'un poste exclusivement chargé des problèmes de confidentialité des données, ayant vu le jour à la naissance de l'UMS-E en septembre 2008, y est pour beaucoup. Après, une année d'investissement méthodologique et logiciel, l'UMS-E a été capable de proposer son appui au sein de la DSE en matière de gestion du secret statistique dans les tableaux.

Plus largement qu'à la DSE, les instituts ont l'obligation légale et morale de garantir la confidentialité des informations qui leur ont été confiées par les répondants aux enquêtes (encadré juridique). Cette confidentialité est vitale pour obtenir une bonne coopération des répondants, auprès de qui il s'agit de collecter un maximum de données de la meilleure qualité possible.
Les instituts ont notamment l'obligation de contrôler la divulgation statistique dans les informations qu'ils mettent à disposition, en minimisant le risque que des informations sensibles sur des individus ou des entreprises puissent être divulguées à partir des données diffusées.

Cet article traite de la gestion du secret uniquement dans le cas de tableaux diffusant des statistiques sur les entreprises ; les méthodes statistiques mises en œuvre sont potentiellement les mêmes que dans le cas des tableaux statistiques sur les ménages, mais la jurisprudence diffère légèrement d'un cas à l'autre, modifiant quelque peu les règles en vigueur.

Encadré juridique : Les garanties données aux répondants

Le secret statistique est un concept du droit pénal. C'est l'application aux agents de la statistique publique de la notion de secret professionnel, définie par l'article L226-13 du code pénal. La loi du 7 juin 1951 définit le « secret en matière de statistiques ». Le 3ème alinéa de l'article 6 de cette loi astreint les agents de la statistique publique au secret professionnel dans l'exercice de leur profession. La loi de 1951 ne fait que renforcer cette obligation des fonctionnaires de la statistique publique, tout en l'étendant aux non titulaires, aux agents des organismes professionnels et des sous-traitants pouvant intervenir dans les activités de la statistique publique. L'alinéa 2 de l'article 7bis de la loi de 1951 s'applique aussi bien aux données collectées par la statistique publique sur des questionnaires statistiques qu'aux données collectées initialement par des formulaires administratifs et mobilisées par elle en vue d'une finalité statistique.
D'autre part, le principe 5 du code de bonnes pratiques de la statistique européenne garantit la confidentialité des répondants. Les indicateurs permettant l'évaluation de ce principe sont en France majoritairement totalement satisfaits. L'évaluation par les pairs propose néanmoins qu'un engagement ou serment à valeur juridique soit signé par le personnel, et que des instructions et lignes directrices soient fournies concernant la protection du secret statistique lors des processus de production et de diffusion.

1. Prolégomènes

Cette partie a d'abord pour but de fournir quelques définitions et vocabulaires facilitant la compréhension de l'article. Certaines de ces définitions seront parfois rappelées et/ou affinées au fur et à mesure de l'article. Ensuite, une première esquisse assez générale de la gestion du secret dans les tableaux de statistiques d'entreprises sera expliquée afin d'exposer le problème auquel il est possible de se confronter lors de la publication de tableaux, et la démarche grâce à laquelle ce problème peut être résolu.

1.1. Premières définitions

Les tableaux diffusant des statistiques sur les entreprises présentent en fait des sommes d'observations d'une variable quantitative, chaque somme faisant référence à un groupe d'observations défini par des variables catégorielles observées sur un ensemble de répondants.
Ces variables catégorielles sont les « variables de ventilation » du tableau, et la variable quantitative est la « variable de réponse ».

Les répondants sont ici des entreprises, et les variables de ventilation fournissent généralement des informations géographique, d'activité ou de taille sur les répondants. Les « cellules » d'un tableau sont définies par le croisement des variables de ventilations du tableau.
Ainsi chaque cellule d'un tableau présente une somme d'une variable quantitative telles que le chiffre d'affaires ou le nombre d'employés. Ces sommes sont les « valeurs des cellules » (parfois appelées « agrégats » ou totaux des cellules ») du tableau. Les observations individuelles de la variable de réponse, c'est-à-dire de chaque répondant, sont appelées les « contributions » aux valeurs des cellules.

La « dimension » d'un tableau est donnée par le nombre de variables de ventilation utilisées pour construire le tableau. Nous dirons que le tableau contient des « marges », s'il existe des totaux pour les variables de ventilation.

A première vue, il peut apparaître difficile de comprendre comment de l'information agrégée publiée dans des tableaux peut présenter un risque de divulgation. Cependant, il arrive fréquemment que des cellules fassent référence à peu de répondants, voire seulement un ou deux. Ce type de cellules est d'autant plus fréquent que le nombre de variables de ventilation est important dans un tableau, et que le niveau de détail de ces variables est fin. Si une cellule d'un tableau fait référence à un petit groupe de répondants, alors la publication de cette cellule peut engendrer un risque de divulgation. C'est le cas lorsque des répondants peuvent être identifiés par un intrus qui utiliserait l'information exposée dans ce tableau.

Exemple 1 :
Supposons qu'un tableau présente le chiffre d'affaires d'entreprises (variable de réponse) en fonction du secteur d'activité et de la région (variables de ventilation). Imaginons maintenant que dans l'activité de Cokéfaction et raffinage, il n'y a qu'une seule entreprise A dans une région particulière. Ce fait peut être connu par certains intrus (notamment par une entreprise B d'une région voisine qui aurait la même activité). Ainsi, si ce tableau est publié en l'état, l'entreprise B pourra connaître le chiffre d'affaires de l'entreprise A.

Afin d'établir si un risque de divulgation existe lors de la publication de valeur de cellules, et afin de se prémunir contre ce risque, les diffuseurs de données, et notamment les instituts statistiques nationaux, doivent mettre en œuvre des méthodes de protection des données. Comme nous le rappelions au début de ce document cette obligation légale et morale est nécessaire pour maintenir la confiance des répondants. En effet si dans l'exemple ci-dessus, l'entreprise A réalise que l'entreprise B peut en lisant le tableau publié connaître son chiffre d'affaires, et qu'elle considère cette information comme confidentielle, elle pourrait refuser de répondre aux prochaines enquêtes, répondre de manière incorrecte ou imprécise, voire même de porter plainte.

1.2. Premier pas dans la gestion du secret

Dans le cas des statistiques d'entreprises, la méthode de protection des tableaux la plus populaire est la « suppression des cellules ». Dans les tableaux protégés par cette méthode, toutes les cellules qui présentent un risque de divulgation sont supprimées de la publication. Il existe des méthodes alternatives telles que les méthodes d'arrondi ou de restructuration, mais nous nous focaliserons sur la suppression des cellules puisque c'est la méthode retenue à la DSE.

La deuxième partie de l'article introduit les concepts méthodologiques de la protection des tableaux. La première sous-partie intitulée « Les cellules sensibles dans les tableaux » présente les règles habituellement utilisées pour l'évaluation du risque de divulgation de chaque cellule d'un tableau. Ces règles sont appelées les « méthodes de contrôle de la divulgation primaire » (ou encore « règles de sécurité »). Les cellules pour lesquelles un risque est établi seront dites « sensibles », « confidentielles » ou « non sécurisées ». Ces cellules seront alors supprimées de la publication, ce seront les « suppressions primaires ». L'ensemble de ces suppressions s'appelle le « secret primaire ».

Cependant, dans bien des cas, supprimer les cellules sensibles ne suffit pas à les protéger car des marges (totaux de lignes ou de colonnes dans un tableau à deux dimensions) ou des marges intermédiaires (par exemple des niveaux agrégés d'activité ; il est courant de diffuser simultanément plusieurs niveaux de la Nace). À cause de cela, nous devons avoir recours à des suppressions supplémentaires, les « suppressions secondaires », pour garantir la protection des cellules sensibles. L'ensemble de ces suppressions s'appelle le « secret secondaire ».

2. Les concepts du contrôle de la divulgation dans les tableaux

La première sous-partie traite de l'évaluation du risque de divulgation de chaque cellule des tableaux qui présentent des agrégats d'une variable quantitative, et introduira les méthodes courantes d'évaluation de ce risque.
Afin de préserver les marges des tableaux alors que l'on souhaite en même temps protéger les cellules sensibles, des méthodes de contrôle de divulgation ont été développées : nous nous focaliserons sur la plus courante, la suppression des cellules.

2.1. Les cellules sensibles dans les tableaux

Nous commencerons cette partie en décrivant des scenarios typiques d'intrusion[1] qu'envisagent les instituts statistiques lors du contrôle de la divulgation des tableaux. Considérant ce type de scénarios d'intrusion, les instituts statistiques ont développé des « règles de sécurité » comme moyen de mesurer les risques de divulgation. Cette partie introduira les règles les plus courantes en s'appuyant sur des exemples.

2.1.1. Les scenarios d'intrusion

Lorsqu'une cellule d'un tableau se rapporte à un groupe restreint (voire un seul) de répondants, la publication de la valeur de la cellule comporte un risque de divulgation. C'est le cas lorsque les répondants peuvent être identifiés par un intrus utilisant l'information contenue dans le tableau diffusé. Dans l'exemple 1, il est suffisant pour l'intrus (l'entreprise B) de savoir que la valeur de la cellule se rapporte au chiffre d'affaires d'une entreprise de Cokéfaction et raffinage d'une région X. Dans cet

[1] L'intrusion est définie comme étant la capacité à divulguer de l'information individuelle.

exemple, l'entreprise B a supposé être capable d'identifier l'entreprise A comme étant la seule entreprise ayant cette activité dans cette région. Ainsi, la publication de la valeur de la cellule comporte un risque de divulgation puisque si l'entreprise B regarde la publication, elle connaîtra le chiffre d'affaires de l'entreprise A.
Le problème reste entier si une cellule se rapporte à deux répondants.

Exemple 2 :
Supposons maintenant que deux entreprises sont localisées dans une région, et qu'elles sont les seules entreprises de Cokéfaction et raffinage de cette région. Supposons de plus que les deux ont connaissance de cette situation. Alors de nouveau la publication de la valeur de la cellule comporte un risque de divulgation (cette fois pour les deux entreprises) : si l'une ou l'autre des entreprises regarde la publication et soustrait son chiffre d'affaires à la valeur de la cellule, elle connaîtra alors le chiffre d'affaires de l'autre entreprise.

L'exemple ci-dessus est basé sur un scenario typique de données d'entreprises : il est généralement supposé que les « intrus »[2] sont bien informés sur le secteur économique auquel se rapporte la cellule. Ce type de scenario a du sens car les tableaux de données que font paraître les instituts sont accessibles par tous, et notamment par les parties les mieux informées sur le sujet dont relève le tableau.
Dans ce scenario, il y a un risque pour que de l'information soit divulguée de manière exacte. Mais que fait-on de la divulgation statistique approchée?

Exemple 3 :
Supposons que dans une région X, il y a 4 entreprises qui travaillent dans le secteur de la Cokéfaction et du raffinage, c'est-à-dire qu'il y a l'entreprise A et 3 autres petites entreprises. Supposons que la contribution de l'entreprise A représente 90% du chiffre d'affaires de ce secteur dans la région. Dans ce scénario, la valeur de la cellule est une estimation très proche du chiffre d'affaires de l'entreprise A. Même si l'intrus potentiel (par exemple l'entreprise B d'une autre région) n'est pas capable d'identifier les 4 entreprises, elle peut tout de même savoir qu'il y a une très grande entreprise dans la région X, et quelle est cette entreprise.

2.1.2. La sensibilité des variables

La présomption de sensibilité d'une variable de réponse détermine le choix de la méthode de protection. Par exemple, spécialement dans le cas de données d'entreprises, les instituts décident de se prémunir contre le risque de divulgation approchée, comme présenté dans le dernier exemple, car la sensibilité de la variable est trop grande.

Ayant en tête les différents scenarios présentés précédemment, les instituts ont développé des « règles de sécurité » (aussi appelées « règles de sensibilité » ou « mesures de sensibilités »), qui permettent d'évaluer le risque de divulgation.

Nous allons maintenant évoquer les règles les plus courantes, et utilisées à la DSE, en présentant une vue d'ensemble de ces règles dans le tableau qui va suivre. Nous discuterons ensuite de ces règles en détail en expliquant dans quelles situations, l'utilisation de ces règles a un sens, avec des exemples simples comme illustration lorsque ce sera nécessaire.

Nous notons $x_1 \geq x_2 \geq \ldots \geq x_n \geq \ldots \geq x_{N(C)}$ les contributions des répondants participant à la valeur d'une cellule C, rangées par ordre décroissant. La valeur de la cellule C sera notée $T_C = \sum_{i=1}^{N(C)} x_i$.

[2] Ici, c'est une entreprise ou un individu qui cherche à divulguer de l'information sur une entreprise.

Règle	Une cellule est considérée comme non sécurisée, lorsque…
Règle de Fréquence minimale	La fréquence[3] de la cellule est inférieure à un seuil s prédéterminé (à la DSE s=3)
Règle de dominance du (n,k)	La somme des n plus grandes contributions dépasse k% le total de la cellule : $\sum_{i=1}^{n} x_i > \frac{k}{100} T_C$. (1)
Règle du p%	Le total de la cellule diminué des deux plus fortes contributions est inférieur à p% de la plus forte contribution : $(T_C - x_2) - x_1 = \hat{x}_1 - x_1 < {}^{p}\!/_{100}\, x_1$. (2)

Tableau 1: Les règles de sensibilité

On peut remarquer que la règle de dominance du (n,k) n'est pas définie pour k=100, et que celle du p% non plus pour p=0. En fait, ces deux règles sont asymptotiquement équivalentes à la règle de fréquence minimale.

Les règles de dominance et du p% sont ce que l'on appelle des « règles de concentration ». Elles pourront être désignées comme telle par la suite.

Lorsque la méthode des suppressions est utilisée pour gérer la confidentialité dans un tableau, les valeurs de cellules considérées comme non sécurisées (sensibles) au regard des règles de sensibilité employées, seront supprimées de la diffusion, et deviendront ainsi des « suppressions primaires ».

Le choix d'une règle particulière de sensibilité est généralement basé sur des scenarios d'intrusion particuliers prenant en compte l'information additionnelle pouvant être mobilisée, car disponible publiquement par exemple.

2.1.3. La règle de fréquence minimale

Les instituts statistiques qui diffusent de l'information fixent une fréquence minimum s. Les cellules d'un tableau qui ont un nombre de répondants supérieur ou égal à ce seuil sont considérées comme sécurisées. L'exemple 1 de l'introduction (la valeur de la cellule fait référence à une seule entreprise dans le secteur de la cokéfaction et du raffinage dans une région donnée) illustre le risque de divulgation pour les cellules de fréquence égale à 1. L'exemple 2 (sur deux entreprises du secteur de la cokéfaction et du raffinage) montre qu'il existe un problème similaire pour les cellules de fréquence égale à 2.

Normalement, le seuil s de fréquence minimale est fixé à 3. C'est ce qui est couramment pratiqué à la DSE.

Il y a une exception à l'utilisation de ce seuil lorsque l'institut estime qu'il est réaliste qu'une coalition de répondants peut mettre en commun leurs données afin de divulguer la contribution d'un autre répondant. C'est l'occasion d'insister sur le fait que de prendre un seuil supérieur à 3 dans le but de se prémunir de ce scenario d'intrusion n'a pas beaucoup de sens. Il n'y pas vraiment de moyen (peut-être mis à part les questions du profilage, et de la connaissance de la diffusion de données par les entreprises elles-mêmes) pour l'institut de savoir quelle hypothèse faire sur la taille d'une possible coalition. Si l'on suppose qu'une cellule fait référence à 100 répondants, et que 99 d'entre elles ont une obligation de diffusion publique de données car elles sont par exemple cotées en bourse, il devient alors possible qu'un intrus additionne l'ensemble des informations disponibles pour divulguer des informations sur la 100ème entreprise. L'institut devrait-il alors considérer que les cellules qui ont une fréquence inférieure à 101 sont considérées comme non sécurisées ?

[3] La fréquence d'une cellule correspond au nombre d'entreprises dans la population étudiée partageant les critères de ventilation de la cellule.

2.1.4. Les règles de concentration

Les valeurs des cellules que l'on publie sont naturellement des bornes supérieures de chacune des contributions à la cellule. Cette borne est évidemment plus proche d'une contribution en particulier, la plus grande. Autrement dit, la contribution la mieux estimée par le total de la cellule est la plus forte contribution à la cellule. Cette constatation simple est le fondement des règles de concentration. Ces règles de concentration n'ont de sens que si l'on considère que les intrus sont capable d'identifier les entreprises qui contribuent le plus, ce qui est le plus souvent le cas.

Lorsque l'on considère qu'une variable est sensible et doit rester confidentielle, se prémunir uniquement contre la divulgation exacte est jugé insuffisant. Dans ce cas, les règles de concentration doivent être utilisées.

On peut noter que normalement (c'est ce que préconise la théorie), il faudrait garder secret les paramètres de ces règles. C'est-à-dire que le seuil s de la règle de fréquence minimale, les paramètres n, k et p des règles de concentration ne devraient pas être communiqués.

Traditionnellement, les instituts statistiques - et c'est le cas à la DSE, sauf pour certaines enquêtes - se contentent d'utiliser la combinaison d'une règle de fréquence minimale et d'une règle de dominance (1,k). Cette approche peut s'avérer incomplète, car on oublie alors que le second plus grand contributeur peut correctement estimer le plus grand contributeur dans certains cas, simplement en soustrayant sa contribution au total de la cellule. L'exemple 4 qui suit illustre le problème.

Exemple 4 : Application de la règle de dominance (1,90).
Posons que le total d'une cellule vaut $T_C = 92000$, que les deux plus fortes contributions valent $x_1 = 81000$ et $x_2 = 8000$. On a que $x_1 < {}^{90}\!/_{100} \cdot 92000 = 82800$. La cellule est sécurisée du point de vue de la règle de dominance (1,90).
Par contre le second contributeur peut faire une estimation de la plus forte contribution en mobilisant seulement la valeur de la cellule et sa contribution : $(T_C - x_2) = \hat{x}_1 = 84000$.
Soit une estimation à environ 3.7% de la vraie valeur de x_1 ($\left(\hat{x}_1 - x_1\right) \Big/ x_1 \approx 0.037$). C'est donc plutôt une bonne estimation.

A l'inverse de la règle de dominance du (1,k), les règles de dominance du (2,k) et du p% prennent correctement en compte l'information que détient le second plus grand contributeur. De ces deux règles, la règle du p% est préférée car la règle du (2,k) a tendance à surprotéger, comme nous le verrons par la suite.

2.1.5. Règles du p% et dominance

Nous allons montrer que, selon les règles de concentration, un agrégat (valeur d'une cellule) est considéré comme sensible s'il fournit une estimation (par valeur supérieure) suffisamment proche d'une des contributions.

Supposons qu'il n'y a pas de coalition de répondants, c'est-à-dire qu'aucun intrus ne connaît plus d'une contribution. Alors l'estimation la plus précise qui puisse être réalisée, est celle que le second contributeur peut faire en soustrayant sa contribution au total de la cellule, pour estimer la plus forte contribution. On a $(T_C - x_2) = \hat{x}_1$ comme montré dans l'exemple 4.
La question que l'on se pose maintenant est comment déterminer qu'une telle estimation est trop proche de la vraie contribution.

L'application de la règle du p% assure que l'estimation $\hat{x}_1$ surestimera la vraie valeur x_1 d'au moins p% pour les cellules non sensibles, c'est-à-dire $\hat{x}_1 - x_1 \geq p/100 \cdot x_1$.

Lorsque l'on adapte la relation (1) du tableau 1 (voir la définition de la règle du (n,k)), au cas n=2, et que l'on soustrait cette inégalité à la valeur de la cellule, on obtient :

$(T_C - x_2) - x_1 = \hat{x}_1 - x_1 < \frac{100-k}{100} \cdot T_C$ (3).

Avec cette formulation la règle de dominance du (2,k) ressemble de près à la formulation de la règle du p%.

Les deux règles définissent un agrégat comme sensible lorsque l'estimation $\hat{x}_1 = (T_C - x_2)$ ne surestime pas « suffisamment » la vraie valeur x_1. La différence entre ces deux règles réside dans le « suffisamment ». Dans la règle du p%, cela est exprimé comme un pourcentage (p %) de la vraie valeur x_1 du plus grand contributeur, alors que dans la règle du (2,k), cela est exprimé comme un pourcentage (100-k %) du total T_C de la cellule. C'est en cela que la règle du (2,k) a une tendance à la surprotection. Il est en effet plus difficile que les cellules remplissent la condition de la règle du (2,k) que celle du p%.

2.1.6. Les poids de sondage

Beaucoup de tableaux sont fabriqués à partir de données d'enquêtes collectées sur un échantillon. Il y a deux raisons pour lesquelles l'évaluation du risque de divulgation peut être nécessaire pour les enquêtes échantillonnées. Tout d'abord, les grandes entreprises se retrouvent souvent dans la partie exhaustive de l'échantillon. En absence de non-réponse, les données de l'enquête sont alors les données de la population complète pour ces strates de tirage, et il se peut que des domaines de diffusion soient identiques à ces strates. Ensuite, même si l'on ne se trouve pas dans la partie exhaustive de l'échantillon, et si un agrégat fait référence à une population très asymétrique[4], l'estimateur de l'agrégat peut être une estimation assez précise de la plus forte contribution à cet agrégat. Par conséquent, il y a un sens à évaluer le risque de divulgation sur les agrégats issus d'enquêtes échantillonnées. La question est de savoir comment de tels agrégats peuvent être considérés comme sécurisés ou non.

La procédure décrite en suivant est celle utilisée dans τ-Argus, et c'est donc celle utilisée par la DSE.
Dans les données d'enquêtes échantillonnées, chaque enregistrement (ligne du fichier de microdonnées correspondant à une entreprise) se voit attribuer un poids. Ces poids sont construits de telle façon que les tableaux construits à partir de ces données sont représentatifs de la population., comme si finalement l'enquête avait été exhaustive.
Pour construire les tableaux, nous devons donc prendre en compte ces poids. C'est ce qui est fait dans τ-Argus lorsqu'on lui spécifie la présence de poids. Par ailleurs, les règles de sensibilité doivent être adaptées à la situation où l'on ne connaît pas les entreprises qui ont les plus fortes contributions. Une solution est l'approximation suivante.
Supposons qu'une cellule d'un tableau soit construite à partir de deux enregistrements ; l'un valant 100 avec un poids de 4, et l'autre valant 10 avec un poids de 7. Alors la valeur de la cellule est $T_C = (4 \times 100) + (7 \times 10) = 470$. Sans prendre en compte les poids, il n'y a que deux contributeurs au total de la cellule. Cependant, en tenant compte des poids, les contributions individuelles peuvent être approximées par la distribution 100, 100, 100, 100, 10, 10, 10, 10, 10, 10, 10. Les deux plus grandes contributions sont maintenant 100 et 100, et seront utilisées dans l'application des règles de sécurité.

[4] On entend par population asymétrique, un ensemble d'entreprises dont les contributions à l'agrégat sont très différentes. Par exemple, une entreprise avec une forte contribution et d'autres nettement plus petites. L'exemple 4 peut illustrer cette situation.

2.2. Les outils de la protection secondaire des données

Nous avons dit en amont que des marges ou marges intermédiaires étaient régulièrement diffusées dans les tableaux. Ainsi, il y a des relations d'additivité entre les cellules des tableaux.

A la DSE, les cellules qui présentent un risque de divulgation sont supprimées de la diffusion, ce sont les « suppressions primaires ». Cependant, à cause des relations d'additivité existantes entre les cellules, cela ne suffit pas, et il faut procéder à de nouvelles suppressions pour protéger les premières. Ces nouvelles suppressions s'appellent les « suppressions secondaires ».

2.2.1. L'intervalle des possibles

Dans l'exemple 5, si l'on considère que la valeur de la cellule x_{21} était sensible (suppression primaire), les trois autres valeurs des cellules x_{23}, x_{31} et x_{33} ont été supprimées (suppressions secondaires) pour empêcher la reconstruction, par différence au moyen des marges, de la cellule sensible.

Lorsqu'un tableau est protégé par la suppression de cellules, en utilisant les relations d'additivité existantes entre les cellules publiées et celles supprimées, il est possible de déduire pour chaque cellule supprimée une borne inférieure et une borne supérieure de sa vraie valeur. Ces deux bornes définissent l' « intervalle des possibles » de la vraie valeur de la cellule supprimée.
Cela tient aussi bien pour les tableaux ventilant une variable de réponse positive, que pour les tableaux ventilant une variable de réponse pouvant être négative, à condition qu'à la place de 0, une autre borne inférieure valable également pour chaque cellule soit disponible aux utilisateurs des tableaux. L'exemple 5 illustre le calcul de l'intervalle des possibles dans le cas d'un tableau simple bidimensionnel, où la variable de réponse ventilée est positive.

Exemple 5 :

Exemple	1	2	3	Total
1	20	50	10	80
2	x_{21}	19	x_{23}	49
3	x_{31}	32	x_{33}	61
Total	45	101	44	190

Tableau 2: Exemple de tableau diffusé protégé par suppressions

Les relations d'additivité des lignes et des colonnes du tableau nous donnent :

$$x_{21} + x_{23} = 49 - 19 = 30$$
$$x_{21} + x_{31} = 45 - 20 = 25$$
$$x_{31} + x_{33} = 61 - 32 = 29$$
$$x_{23} + x_{33} = 44 - 10 = 34$$
$$x_{21}, x_{23}, x_{31}, x_{33} \geq 0$$

En utilisant les relations précédentes pour trouver une borne supérieure et une borne inférieure à chacune des 4 cellules supprimées, on montre que le tableau 3 est équivalent au tableau 4.

Exemple	1	2	3	Total
1	20	50	10	80
2	$[0\ ,\ 25]$	19	$[5\ ,\ 30]$	49
3	$[0\ ,\ 25]$	32	$[4\ ,\ 29]$	61
Total	45	101	44	190

Tableau 3: Tableau avec intervalles des possibles

Il faut donc faire attention au fait que deux suppressions par ligne ou colonne ne suffisent pas nécessairement. Le risque est de pouvoir obtenir dans certains cas des intervalles des possibles suffisamment étroits pour que cela donne une estimation assez fine de la valeur initialement supprimée.
L'intervalle des possibles des suppressions primaires doit donc être suffisamment large. Il faut exiger une longueur d'intervalle des possibles minimum, c'est-à-dire un niveau de sécurité minimum pour l'ensemble des suppressions primaires.

2.2.2. Le niveau de sécurité

En principe, une structure de suppressions (masque de suppressions ou masques de secret) peut être considérée comme valable, si les bornes des intervalles des possibles des cellules sensibles (suppressions primaires) ne peuvent être utilisées pour déduire un encadrement trop précis, au sens des règles de sensibilité utilisées, de la contribution individuelle d'un répondant à la cellule.

Pour formaliser mathématiquement cette condition, nous déterminons des bornes de sécurité pour les suppressions primaires. Nous appelons l'écart entre ces bornes de sécurité et la vraie valeur de la cellule, les niveaux inférieur et supérieur de protection. Dans le tableau 4 qui suit sont répertoriées les formules de calcul des niveaux supérieurs de protection des règles de concentration. En dehors de toute considération de symétrie, les niveaux inférieurs de protection sont souvent identiques au niveau supérieur.

Règles de sensibilité	**Niveau supérieur de protection**
Règle de dominance du (1,k)	$100/k \cdot x_1 - T_C$
Règle de dominance du (n,k)	$100/k \cdot (x_1 + x_2 + ... + x_n) - T_C$
Règle du p%	$p/100 \cdot x_1 - (T_C - x_1 - x_2)$

Tableau 4: Niveaux supérieurs de protection

Si la distance entre la borne supérieure de l'intervalle des possibles et la vraie valeur de la cellule sensible est inférieure au niveau supérieur de protection calculé par l'une des formules du tableau 5, alors la borne supérieure de l'intervalle des possibles peut être utilisée pour estimer la contribution individuelle d'un répondant d'une cellule sensible (Cox 1981).

Exemple 6 : Preuve de la formule du tableau 5 pour la règle du p%

Supposons que $T_C + U$ soit la borne supérieure de l'intervalle des possibles pour une cellule dont la valeur est T_C. Le second plus grand contributeur peut alors déduire une borne supérieure de x_1,

notée x_1^U, en soustrayant sa propre contribution à la borne supérieure de l'intervalle des possibles : $x_1^U = T_C + U - x_2$. Une cellule est sécurisée selon la règle du p%, si l'on a $x_1^U \geq (1 + p/100) \cdot x_1$. Ainsi si l'intervalle des possibles fournit la protection voulue, on a alors $U \geq (1 + p/100) \cdot x_1 - T_C + x_2 = p/100 \cdot x_1 - (T_C - x_1 - x_2)$.

La distance entre la borne supérieure de l'intervalle des possibles et la vraie valeur de la cellule sensible doit majorer le niveau supérieur de protection. Dans le cas contraire, la cellule sensible n'est pas protégée correctement. Ce critère de sécurité est un critère nécessaire, mais pas suffisant pour obtenir une protection correcte.
En effet, même si les cellules supprimées fournissent l'intervalle des possibles exigé, la combinaison de cellules supprimées (sur la même ligne ou colonne que la cellule sensible que l'on souhaite protéger), ne satisfait pas nécessairement le secret primaire suivant cette règle.
Cela tient au fait que l'on ne contrôle pas la structure virtuelle issue de la combinaison des cellules supprimées : le second contributeur de la cellule combinée peut être le plus grand contributeur de la cellule ayant servie à protéger la cellule sensible au départ.

Exemple 7 : Posons pour cet exemple p=25%.

Exemple	1	2	3	Total
1	$x_{11} = 51$	42	$x_{13} = 55$	148
2	$x_{21} = 9$	10	$x_{23} = 31$	50
3	28	43	29	100
Total	88	95	115	298

Supposons qua la cellule x_{11} soit distribuée de la manière suivante : $x_{11} = 51 = 44 + 4 + 1 + 1 + 1$
On a alors : $51 - 44 - 4 = 3 < 25/100 \cdot 44 = 11$. La cellule est donc sensible selon cette règle.
Supposons également qua la cellule x_{21}, servant à protéger la précédente, soit distribuée de la manière suivante : $x_{21} = 9 = 6 + 1 + 1 + 1$
On a alors : $9 - 6 - 1 = 2 > 25/100 \cdot 6 = 1.25$. La cellule n'est donc pas sensible selon cette règle.

Regardons maintenant, la cellule virtuelle qu'est la combinaison de x_{11} et x_{21} :
$x_{11} + x_{21} = 60 = 44 + 6 + 4 + 1 + 1 + 1 + 1 + 1 + 1$.
On a alors $60 - 44 - 6 = 10 < 25/100 \cdot 44 = 11$. La cellule virtuelle combinaison des deux cellules supprimées n'est pas sécurisée.
Le second contributeur de la cellule sensible n'est plus le second contributeur de la cellule combinée.

Pour que le critère de sécurité soit nécessaire et suffisant pour l'obtention d'une protection correcte, il faut que le second contributeur de la cellule combinée soit celui de la cellule sensible.

3. La suppression de cellules dans τ-ARGUS : contexte méthodologique, conseils et recommandations

Le progiciel τ-ARGUS fournit des outils permettant le contrôle de la divulgation statistique dans les tableaux de données agrégées. La protection contre la divulgation de données sensibles peut être réalisée en modifiant les tableaux : soit en diffusant moins d'information, soit de l'information moins détaillée.
τ-ARGUS permet plusieurs types de modification des tableaux. Un tableau peut être remodelé, en fusionnant des lignes et/ou des colonnes ; les cellules sensibles peuvent être supprimées (suppressions primaires), et des suppressions additionnelles (suppressions secondaires) peuvent être trouvées pour protéger les premières. Il existe plusieurs algorithmes pour la sélection des suppressions secondaires : Hypercube, Modular, Optimal et Network.

3.1. Le problème de la protection des tableaux en pratique

Nous allons expliquer le concept basique des tableaux lors de la protection de données tabulaires, et également les structures plus complexes lors de la protection de tableaux hiérarchisés, et de tableaux liés. Afin de trouver un équilibre entre la protection des données individuelles et la mise à disposition d'information, il est nécessaire d'évaluer la perte d'information due à la suppression d'une cellule. Nous expliquerons donc le concept de coût des cellules pour évaluer la perte d'information.

3.1.1. La spécification des tableaux

Pour préparer la formalisation mathématique des problèmes de protection des données lors de la mise à disposition de tableaux sur un ensemble de données, toutes les relations linéaires entre les cellules de ces tableaux doivent être considérées.

Les tableaux dont il est question ici affichent des sommes d'observations d'une variable quantitative, appelée variable de réponse. Ces sommes sont calculées sur toutes les observations et/ou sont regroupées par groupe d'observations. Ces groupes peuvent être constitués en agrégeant les réponses de répondants selon des critères qui peuvent être l'activité économique, la région, des classes de chiffres d'affaires, des classes d'effectif ou encore la forme juridique. Dans ce cas, les groupes d'observations sont définis par des variables catégorielles observées sur chacune des unités répondantes, ce sont les variables explicatives.

La dimension d'un tableau est donnée par le nombre de variables de ventilation utilisées pour spécifier le tableau. Les marges, ou encore cellules marginales d'un tableau sont les cellules qui sont spécifiées par un plus petit nombre de variables que la valeur de la dimension. Plus ce nombre est petit et plus le niveau d'agrégation de ces marges est élevé. Par exemple, dans un tableau bidimensionnel de données d'entreprises, il est possible de ventiler une variable de réponse selon l'activité économique et la tranche de taille des entreprises. Dès lors, il est également possible de ventiler cette variable de réponse selon uniquement l'activité économique, et également uniquement selon les tranches de taille. Ces deux derniers types de ventilation sont les marges, ou cellules marginales du tableau. Si l'on diffuse également, la somme de la variable de réponse sur l'ensemble des observations, on appellera cette somme le total, ou grand total du tableau.

3.1.2. Tableaux liés et tableaux hiérarchisés

Les données collectées par le système statistique peuvent rencontrer des attentes différentes de la part des utilisateurs. Certains souhaitent des données publiées à des niveaux agrégés, alors que d'autres ont besoin de données plus détaillées. En tant que statisticiens, nous devons essayer de satisfaire au mieux les attentes des utilisateurs en publiant nos données à différents niveaux de détail. Nous combinons généralement plusieurs variables de ventilation de différentes façons lorsque nous

créons les tableaux à publier. Si deux tableaux qui présentent des données sur la même variable de réponse partagent tout ou partie des catégories d'au moins une variable de ventilation, il y aura alors de cellules communes dans les deux tableaux : ces tableaux sont donc des « tableaux liés » de part les cellules qu'ils ont en commun. Le cas le plus fréquent est le partage d'une variable de ventilation dans son intégralité.

Bien entendu, lorsque l'on utilise les méthodes de suppression de cellules pour protéger un ensemble de tableaux liés, il n'est pas suffisant de trouver de manière indépendante les cellules dites de suppressions secondaires. Sinon, il se pourrait qu'une même cellule soit supprimée dans un tableau mais pas dans un autre. Alors un utilisateur qui comparerait les deux tableaux serait capable de divulguer des cellules confidentielles du premier tableau.
Une approche classique est de protéger les tableaux individuellement, de reporter sur chacun des tableaux les suppressions secondaires des autres tableaux et de répéter la procédure de protection individuelle des tableaux.

Afin d'offrir des niveaux de détails, nous utilisons des schémas de classification élaborés pour catégoriser les répondants. Ainsi, un répondant se retrouvera le plus souvent dans plusieurs catégories du schéma de classification. Par exemple, une entreprise se situe dans une ville localisée dans un département appartenant à une région. Il en est de même pour les activités économiques des entreprises, puisque la nomenclature d'activités est une variable possédant une structure hiérarchisée. La figure 1 donne un exemple de hiérarchie d'activité, extrait de la nomenclature d'activités.

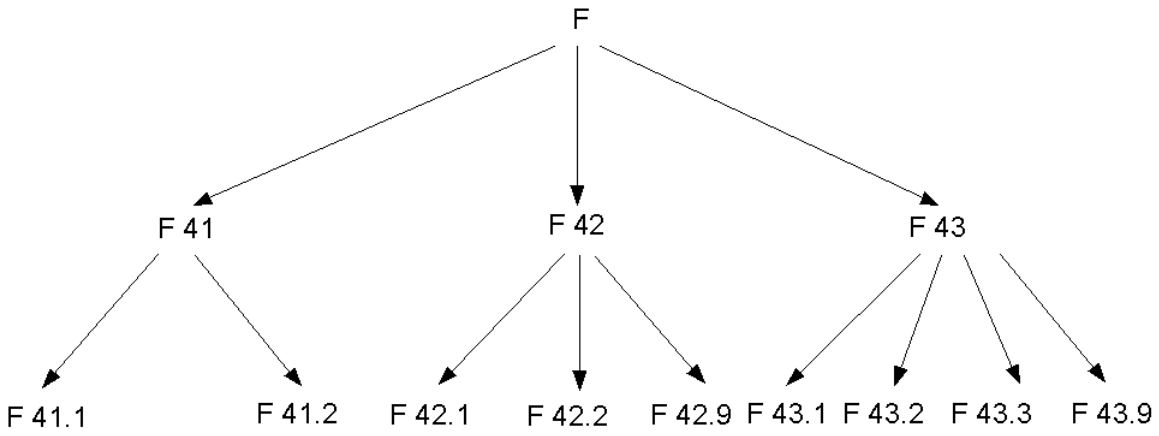

Figure 1: Exemple de hiérarchie

Pour gérer l'exemple de la figure 1, τ-ARGUS découpera l'arbre en quatre tableaux qui seront alors liés par les cellules qu'ils ont en commun.

F		F41		F42		F43
F41		F41.1		F42.1		F43.1
F42		F41.2		F42.2		F43.2
F43				F42.9		F43.3
						F43.9

Figure 2: Découpage de l'arbre en sous-tableaux

Le travail de découpage de l'arbre et de repérage des liens entre les sous-tableaux réalisé, le progiciel τ-ARGUS utilise la procédure de suppressions décrite pour les tableaux liés pour sécuriser le tableau faisant apparaître une structure hiérarchisée.

3.1.3. Coût des cellules

Le challenge de la protection des données tabulaires est de réussir à publier un maximum d'information, tout en générant suffisamment d'incertitudes sur les vraies valeurs des cellules sensibles. Il apparaît donc nécessaire d'estimer d'une manière ou d'une autre le contenu d'information des données diffusées, afin de pouvoir choisir un masque de secret « optimal »[5].
La perte d'information est exprimée comme somme des coûts associés aux suppressions secondaires.

Pour la suppression des cellules, l'idée de réaliser la plus petite perte d'information possible en supprimant un minimum de cellules est certainement le concept le plus naturel. C'est par exemple le cas lorsque l'on donne un coût identique à chaque cellule. Hors, cette technique aboutit parfois à supprimer des cellules dont la valeur est grande, ce qui est le plus souvent indésirable. Dans la pratique, d'autres fonctions de coût basées sur la valeur de la cellule ou d'une transformation de celle-ci donnent de meilleurs résultats. On peut noter que d'autres critères que celui de la valeur numérique de la cellule peuvent importer aux utilisateurs. Par exemple, la place d'une cellule dans un tableau (marges et sous-totaux sont généralement considérés comme très important). τ-ARGUS permet de choisir sa fonction de coût et d'effectuer une transformation de celle-ci, comme nous le montrons dans la suite.

3.1.4. La fonction de coût et λ

La fonction de coût indique l'importance relative donnée à une cellule. Dans τ-ARGUS, les fonctions de coût les plus classiques sont disponibles :

- L'unité, c'est-à-dire que chaque cellule possède le même poids (cela permet de minimiser le nombre de suppressions).
- Le nombre de contributeurs (cela permet de minimiser le nombre de contributeurs participant au calcul des cellules supprimées).
- La valeur de la cellule (cela permet de préserver autant que possible les cellules les plus grandes).
- La valeur de la cellule pour une autre variable de réponse (cette variable donnera l'importance des cellules).

Dans les deux derniers cas, la préférence est donnée aux cellules qui possèdent une grande valeur. Pour autant, l'importance donnée à une cellule est-elle directement proportionnelle à la valeur de celle-ci ? Autrement dit est-ce qu'une cellule, dont la valeur est 10 fois plus grande que celle d'une autre, est 10 fois plus importante ?
Il est parfois nécessaire d'être un peu plus modéré dans l'importance que l'on donne aux cellules. Ainsi, une transformation de la fonction de coût peut s'avérer nécessaire. Les transformations classiques sont la racine carrée, et le logarithme. Box et Cox (1964) avaient proposé la transformation suivante :

$$y = \frac{x^{\lambda} - 1}{\lambda}$$

Dans cette formule, λ=0 revient à une transformation logarithmique, et le -1 permet de prolonger la continuité en λ.

La formule implémentée dans τ-ARGUS est plus simple :

$$\begin{cases} y = x^{\lambda} \ si \ \lambda \neq 0, \\ y = \log(x) \ si \ \lambda = 0. \end{cases}$$

[5] L'optimalité est au sens de la minimisation du coût des suppressions secondaires.

Ainsi, choisir λ=1/2, revient à prendre la racine carrée de la fonction de coût initiale. Particulièrement dans les grands tableaux avec une structure hiérarchisée détaillée, on peut observer que le choix du lambda peut affecter significativement la structure de suppression.

3.2. Procéder efficacement à la protection des tableaux

Dans cette partie, nous donnons quelques lignes directrices pour faciliter l'analyse de la divulgation de données, et l'utilisation des méthodes disponibles dans τ-ARGUS.

3.2.1. Restructuration de tableaux

Lorsqu'un nombre important de cellules sensibles sont présentes dans un tableau, cela peut être une indication sur le fait que des variables de ventilation peuvent être trop détaillées. Dans ce cas, lorsque le plan de publication n'est pas encore fixé (ou obligatoire car répondant à un règlement par exemple), il est possible de fusionner des lignes et des colonnes du tableau. Dans le cas contraire, le nombre de suppressions secondaires pourraient être très important. Cette idée vient du fait que la fusion de cellules apporte davantage de sécurité aux contributions individuelles. Ainsi par cette opération, le nombre de cellules sensibles devrait être réduit. Par contre, à trop réaliser de regroupements, on perd alors l'utilité du tableau. S'il persiste des cellules sensibles, il faut alors avoir recours à d'autres méthodes telle que la méthode des suppressions.

En pratique, il est souvent impossible de réaliser l'analyse de la divulgation sur l'ensemble des tableaux que l'on souhaite publier. Une simplification bien connue, et utilisée à la DSE, est de définir des variables de réponses, dites clés. Les tableaux clés qui en résultent seront alors protégés avec τ-ARGUS, et les masques de secret trouvés seront appliqués tels quels à d'autres tableaux.
Cette stratégie possède l'avantage d'être plus efficace en temps, et de prévenir du recalcul grâce aux relations entre les variables de réponses de différents tableaux. Si l'on publie des variables de réponse ventilées selon les mêmes critères et qu'il existe une relation d'additivité entre ces variables de réponse, le fait de n'utiliser qu'un seul masque de secret pour toutes ces variables empêche d'utiliser ces relations d'additivité pour recalculer les suppressions effectuées.

Un compromis entre la copie intégrale du masque de secret, et la protection individuelle des tableaux, est la coordination des suppressions primaires uniquement. Cette stratégie est gérée dans τ-ARGUS grâce au concept de « shadow variable » (variable ombre). Cette variable est utilisée pour le tableau clé, et pour définir les cellules sensibles dans les autres tableaux. Ensuite, la recherche des suppressions secondaires se fait normalement, c'est-à-dire dans chaque tableau.

3.2.2. Le concept de « shadow variable »

La théorie des cellules non sécurisées de manière primaire, cellules sensibles, est basée sur le fait que les contributions les plus grandes peuvent être à risque et doivent être protégées. Par conséquent, les plus fortes contributions jouent un rôle important dans les règles de sensibilité.

Très souvent, la valeur de la cellule elle-même est un bon indicateur de la taille des contributeurs. Mais parfois, la valeur de la cellule n'est pas le meilleur indicateur de la taille des entreprises qui la composent. En effet, si par exemple la valeur de la cellule correspond à l'investissement réalisé dans des produits particuliers, il se peut que ce ne soit pas la plus grande entreprise (au sens de l'effectif ou du chiffre d'affaires par exemple) qui investit le plus dans ces produits. Il se pourrait même qu'au regard des règles de sensibilité, cette cellule soit mise en secret du fait d'une petite entreprise. Si l'on assume que cette entreprise n'est pas très connue, ni visible, pas vraiment mise en concurrence, il n'y pas de raison de protéger cette cellule.

Si l'on souhaite protéger la réelle entreprise dominante du secteur, car ces données sont très sensibles, une meilleure idée est d'appliquer les règles de sensibilité sur un autre tableau (construit à partir de la variable ombre) utilisant les mêmes variables de ventilation. Il suffit ensuite d'utiliser la structure des suppressions primaires, uniquement, de la copier dans le tableau que l'on souhaite sécuriser, puis de rechercher classiquement les suppressions secondaires dans ce tableau.

Dans τ-ARGUS, nous appelons le tableau qui sert à la recherche des cellules sensibles, le « shadow table », le tableau ombre.

Une autre raison d'utiliser une variable ombre peut être la coordination des structures de suppressions entre différents tableaux utilisant les mêmes variables de ventilation et une variable de réponse publiée périodiquement. On pourra se rapporter au paragraphe sur les publications périodiques.

3.2.3. Les interventions manuelles dans la procédure

Lors de procédure de contrôle de la divulgation statistique, il est parfois nécessaire d'avoir recours à des interventions manuelles. Pour un certain nombre de raisons, des cellules doivent être absolument supprimées de la diffusion (même si elle n'enfreigne pas les règles de sensibilité), ou au contraire exclues des structures de suppressions. Dans le premier cas, on peut imaginer que des données militaires seraient considérées comme extrêmement sensibles alors même que les cellules correspondantes à ces données remplissent tout à fait les règles de sécurité fixées. Dans le second cas, on peut imaginer qu'une entreprise dominante donne l'autorisation de diffusion de ces chiffres, souvent après demande de l'institut. Dans ce cas, nous devons éviter que les cellules correspondantes se retrouvent dans les structures de suppressions au regard de l'intérêt de diffusion de ces chiffres. On peut également imaginer que des cellules qui pourraient être utilisées comme suppressions secondaires soient connues par ailleurs, à travers d'autres publications par exemple.

Ainsi les préférences que l'on peut avoir lors de la sélection des suppressions secondaires peuvent être réalisées manuellement dans τ-ARGUS. Cependant la spécification de ces préférences peut être assez fastidieuse et laisse place à des erreurs possibles.
τ-ARGUS propose également l'utilisation d'un « fichier de préférences » (« a-priori file ») qui permet de spécifier des instructions pour chaque cellule, en donnant les modalités correspondantes des variables de ventilation :

- La cellule ne peut être utilisée comme suppression secondaire.
- La cellule doit être absolument supprimée.
- La cellule ne doit pas être supprimée.
- Spécifier une nouvelle valeur du coût de la cellule pour influencer son choix en tant que suppression secondaire.

Ces interventions peuvent trouver également leur intérêt dans les publications faites régulièrement.

3.2.4. Les publications périodiques

Lorsque des tableaux sont publiés régulièrement (mensuellement, trimestriellement, annuellement), des problèmes de gestion de la confidentialité peuvent apparaître. En effet, si l'on réalise l'analyse de la divulgation indépendamment d'une fois sur l'autre, les structures de suppressions peuvent évoluer.
Nous pouvons donc nous retrouver dans la situation où une cellule est publiée pour une période donnée alors qu'elle sera utilisée comme suppression secondaire lors de la période suivante. Il y a également le risque lorsqu'une cellule est publiée lors d'une période et devient une suppression primaire (cellule sensible) lors de la seconde période. Il est en effet possible que la cellule publiée lors de la première période soit un très bon proxy (estimation très proche de la vraie valeur) de la valeur de la cellule de la seconde période. Spécialement dans le cas où les observations tendraient à être relativement constante dans le temps, la valeur de la première période sera une estimation très proche de la valeur de la cellule supprimée lors de la période courante, qui pourra être utilisée pour divulguer des cellules sensibles de la période en cours.

Une solution simple est de copier la structure de suppression de la période précédente et d'ajouter si nécessaires les suppressions primaires et secondaires supplémentaires de la période en cours. En fait, il faut contrôler qu'en plus de la structure de suppression de la période précédente, il n'y aient pas des cellules sensibles supplémentaires non couvertes par la structure de suppression déjà mise en place, et qui engendreraient d'autres suppressions secondaires.

Une autre solution serait de donner un coût faible aux cellules supprimées lors de la période précédente lors de la recherche des suppressions secondaires de la période en cours pour inciter leur choix dans la structure des suppressions secondaires.

Cependant ces deux stratégies causeront une perte d'information d'autant plus grande que le changement des cellules sensible sera important.

Bibliographie

[1] Willenborg L., de Waal T., « Elements of Statistical Disclosure Control », *Lecture Notes in Statistics*, vol 155, Springer-Verlag, 2000.
[2] Nicolas J., « La gestion du secret dans les tableaux diffusant des statistiques d'entreprises », *La Lettre du SSE*, n°65, septembre 2010.
[3] Box G. E. P., Cox D. R., « A analysis of transformations », *Journal of Royal Statistical Society*, Series B, vol. 26, 1964.
[4] Cox L., « Linear Sensitivity Measures in Statistical Disclosure Control », *Journal of Planning and Inference,* 5, 1981.