Direction de la méthodologie et de la coordination statistique et internationale
Département des méthodes statistiques
Division Recueil et traitement de l'information

NOTE
À l'attention
de la directrice de la DSE,
des chefs de départements de la DSE,
des chefs de division de la DSE,
de la directrice de la DDAR,
du chef d'IIS
et de la cheffe de la division grands comptes

Dossier suivi par :
Maxime Bergeat
Tél. : 01 41 17 64 86
Mél : Maxime Bergeat

Paris, le 05 Février 2014
N° 81/ DG75-L120/MB

Objet : La méthodologie du traitement de la confidentialité pour les données tabulées - Foire Aux Questions

Cette note vise à recenser et à répondre aux questions que peut être amené à se poser un producteur de données lorsqu'il cherche à diffuser des tableaux de données. Les décisions concernant la protection des tableaux de données diffusés par l'Insee sont prises au final par le producteur des données, avec l'appui méthodologique éventuel du Département des méthodes statistiques (DMS). C'est en particulier le producteur qui choisit[1] les règles à appliquer pour détecter les cellules sensibles à cacher (secret primaire) ainsi que les paramétrages à effectuer lors de la recherche des cases sous secret secondaire. Les exemples concrets s'appuient sur l'utilisation du logiciel Tau-Argus.

- Quelles sont les règles à utiliser pour repérer le secret primaire ?
  Concernant les statistiques d'entreprises, une décision entérinée par le Directeur Général de l'Insee fait jurisprudence depuis 1980 et fixe deux règles à utiliser :
  - La règle des trois unités : toute case diffusée doit concerner au moins trois unités légales (entreprises par exemple)
  - La règle de dominance ou règle des 85 % : le premier contributeur d'une case diffusée doit contribuer pour moins de 85 % au total de la case.

  Les cases ne respectant pas une des deux règles décrites ci-dessus forment le secret primaire. On ne peut pas les diffuser. Il faut ensuite supprimer des cases supplémentaires afin de s'assurer qu'on ne peut retrouver une case en secret primaire en utilisant les marges diffusées au sein du tableau ou par ailleurs.

  Les pondérations liées à un éventuel échantillonnage sont prises en compte pour repérer le secret primaire : soit une case $T_C = \sum_{i=1}^{n} w_i x_i$. On considère que cette cellule est sensible si :

  - $\sum_{i=1}^{n} w_i < 3$ (règle des trois unités non respectée)

[1] Généralement, selon les sources, des règles existent déjà et le producteur les connaît, ce qui n'est pas forcément le cas de l'expert confidentialité : par exemple règle des 3 unités pour les données d'entreprise, des 11 ménages fiscaux pour les données fiscales, etc.

- ou $\max_{i=1}^{n} x_i > 0.85 \times T_C$ (atteinte à la règle de dominance)

Certains pays utilisent également en complément ou en remplacement de ces deux règles la règle du *p* %. Une cellule $T_C = \sum_{i=1}^{n} w_i x_i$ est jugée sensible si et seulement si, en notant $x_{(1)}$ et $x_{(2)}$ les deux premiers contributeurs de la case :

$\left(T_C - x_{(2)}\right) - x_{(1)} < \frac{p}{100} x_{(1)}$.

Prendre en compte cette règle permet de protéger les cases pour lesquelles le deuxième contributeur peut obtenir simplement (en considérant l'estimateur basique $\hat{x}_{(1)} = \left(T_C - x_{(2)}\right)$) une estimation précise (avec une imprécision inférieure à *p* %) de $x_{(1)}$.

Dans les traitements réalisés à l'heure actuelle par le DMS, on ne prend pas en compte la règle du *p* %. Toutefois, sur demande du producteur de données, il est possible de la considérer en complément des règles de fréquence (trois unités) et de dominance.

- Comment choisir les variables à utiliser pour effectuer les traitements liés à la confidentialité ?

  Parfois, le nombre de variables à ventiler dans les tableaux est très important. Dans de tels cas, une stratégie globale de protection des tableaux doit être adoptée, pour minimiser le temps requis pour traiter la confidentialité et maximiser la protection apportée. Le choix des variables de réponse utilisées pour faire les masques de secret doit se faire en fonction de deux critères principaux :

  - Tout d'abord, il faut se concentrer sur les variables « identifiantes », autrement dit celles qu'un intrus peut facilement estimer. Quand on applique une règle de dominance, on considère implicitement qu'un curieux est capable de classer dans l'ordre croissant (au sens croissance de la variable de réponse) les unités d'une case. Ceci peut être considéré comme vrai pour le chiffre d'affaires d'une entreprise, mais la pertinence de cette hypothèse peut être remise en cause pour un excédent brut d'exploitation par exemple.

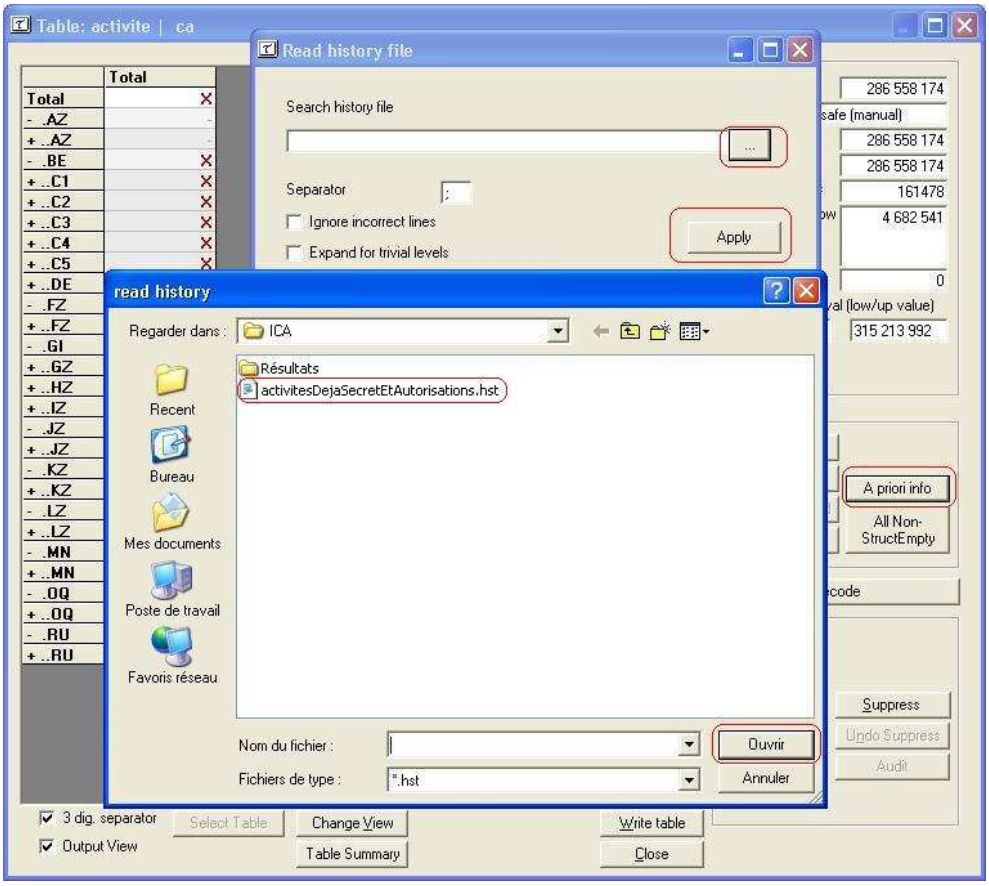
  - Il faut faire attention aux éventuels liens entre les variables de réponse ventilées dans les tableaux. Par exemple, s'il y a une égalité comptable entre ces variables, le secret peut être déjoué en utilisant cette relation si jamais les traitements sont réalisés indépendamment pour chaque variable.

  Par conséquent, il peut parfois s'avérer judicieux de choisir un petit nombre de variables « proxy » utilisées pour créer les masques de secret, et de reporter ces différents masques (mêmes structures de suppressions) pour d'autres variables de réponse liées à la variable proxy.

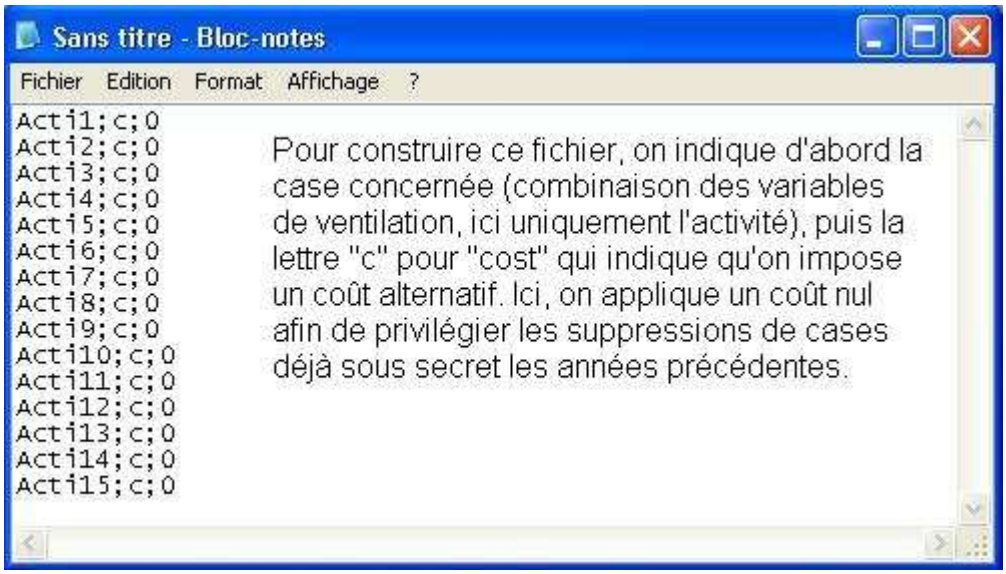
- Comment opérer des traitements répétés régulièrement, par exemple des données d'enquête actualisées chaque année ?

Plusieurs solutions sont envisageables. De façon générale, les traitements pour la confidentialité sont effectués indépendamment chaque année. Toutefois, il est possible d'effectuer les suppressions en fonction de ce qui a été supprimé les années précédentes (concernant le choix des cellules pour le secret secondaire, la détection du secret primaire étant uniquement due aux règles adoptées pour mesurer la sensibilité des cases du tableau). Ce choix est judicieux si on diffuse des séries temporelles, par exemple.

Pour appliquer cette méthode, il convient de considérer une fonction de coût alternative. En général, on cherche à supprimer les cases les plus petites possibles, et la fonction de coût est égale à la variable ventilée dans le tableau. Si on veut prendre en considération les suppressions des années précédentes pour les privilégier, on peut utiliser un fichier « *a priori* » qui permet de spécifier un coût alternatif pour les cases cachées les années précédentes.



Le fichier *a priori* est un fichier texte qui permet de spécifier des coûts alternatifs entre autres.

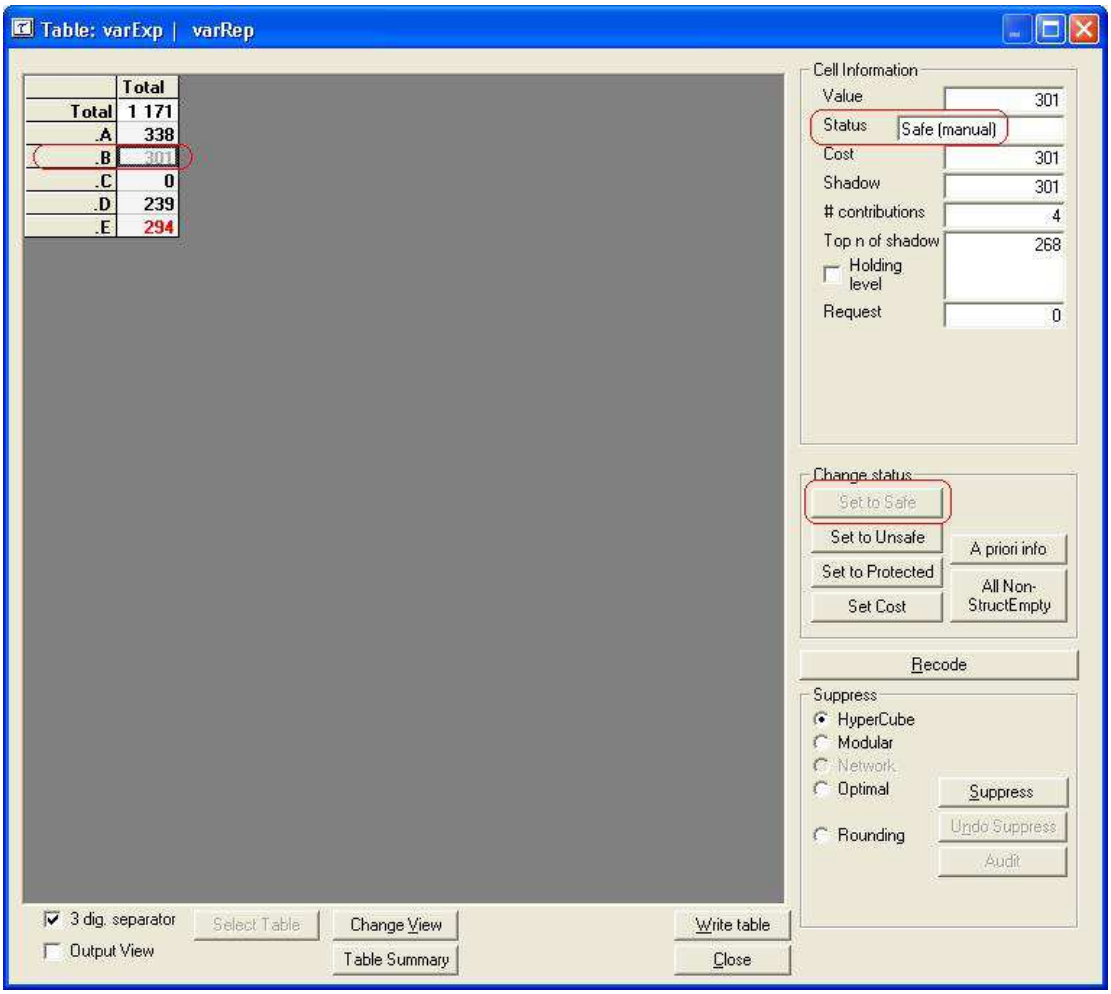


Ici, les secteurs portant les codes « Acti1 » à « Acti15 » seront privilégiés pour les suppressions.

- Comment procède t'on si une entreprise dominante dans un secteur a donné son accord pour la diffusion des données la concernant ?

  Dans certains secteurs, une entreprise fortement dominante publie également des indicateurs économiques la concernant sur son site Internet. C'est en particulier le cas de La Poste et de la SNCF, qui dominent respectivement les secteurs postaux et ferroviaires. Dans ces cas, on peut vouloir lever le secret primaire sur ces secteurs. Une autorisation spéciale de diffusion a été demandée à certaines entreprises (dont La Poste et la SNCF) qui permet de publier certains agrégats sectoriels malgré la dominance d'une entreprise.

  Pour pouvoir rendre diffusable une case normalement sensible, on utilise l'option « Set to safe » dans Tau-Argus.



On pourrait également utiliser un fichier *a priori* comme dans le point précédent, afin d'automatiser cette procédure si un nombre important de cases est concerné. Cela n'est toutefois pas recommandé car la protection d'une telle case sensible ne peut pas se faire si la case concernée contient moins de trois unités. En effet, dans ce cas-là, le deuxième contributeur de la case est divulgué, car le total est connu ainsi que le premier contributeur (donnée disponible *via* un autre canal).

La « déprotection » forcée d'une case pour laquelle le premier contributeur a donné son accord pour la divulgation ne peut donc se faire qu'à une des deux conditions suivantes :

- Il y a au moins trois unités contributrices dans cette case.
- L'unité dominante est l'unique contributrice du total.

- Quelle est la méthode à utiliser pour choisir les suppressions secondaires ?

Quatre méthodes sont implémentées pour rechercher les suppressions secondaires dans Tau-Argus : Hypercube, Network, Modular et Optimal. L'utilisation de la méthode Network est déconseillée car elle ne peut être utilisée que dans un nombre limité de cas (tableaux avec deux variables de ventilation dont une hiérarchique). De plus, elle ne permet pas de prendre en compte la non-divulgation des singletons. D'ailleurs, cette méthode ne sera plus disponible lors de la sortie fin 2014 de la prochaine importante mise à jour de Tau-Argus.

Pour choisir entre les trois autres méthodes, le premier critère à prendre en compte est celui de leur disponibilité. En effet, les méthodes Modular et Optimal font appel au solveur de calcul payant Xpress. Par conséquent, si on ne dispose pas de la licence pour cet outil, on ne peut utiliser que la méthode Hypercube. Bien que conduisant parfois à des situations où la perte d'information liée à la suppression de cases du tableau est importante, cette méthode est préférable à la protection « manuelle » des tableaux et permet en outre de s'assurer de la protection par intervalles. Si on a accès au solveur Xpress, il est conseillé d'utiliser la méthode Optimal sous réserve d'obtenir un schéma de suppressions secondaires dans un temps raisonnable[2].

- Comment protéger efficacement un fichier de données et éviter la divulgation par intervalles ?

Pour construire des tableaux sécurisés, il faut s'assurer qu'on est bien protégé contre tous les risques de divulgation possibles.

- Protection contre la divulgation exacte

Lors de la protection d'une table, après avoir détecté les problèmes de secret primaire, il faut cacher des cases supplémentaires pour qu'on ne puisse pas retrouver les données sensibles grâce aux marges diffusées, par simple différence. Là réside tout l'intérêt des traitements pour la confidentialité et de l'utilisation d'un logiciel permettant de les industrialiser comme Tau-Argus.

- Protection contre la divulgation par intervalles

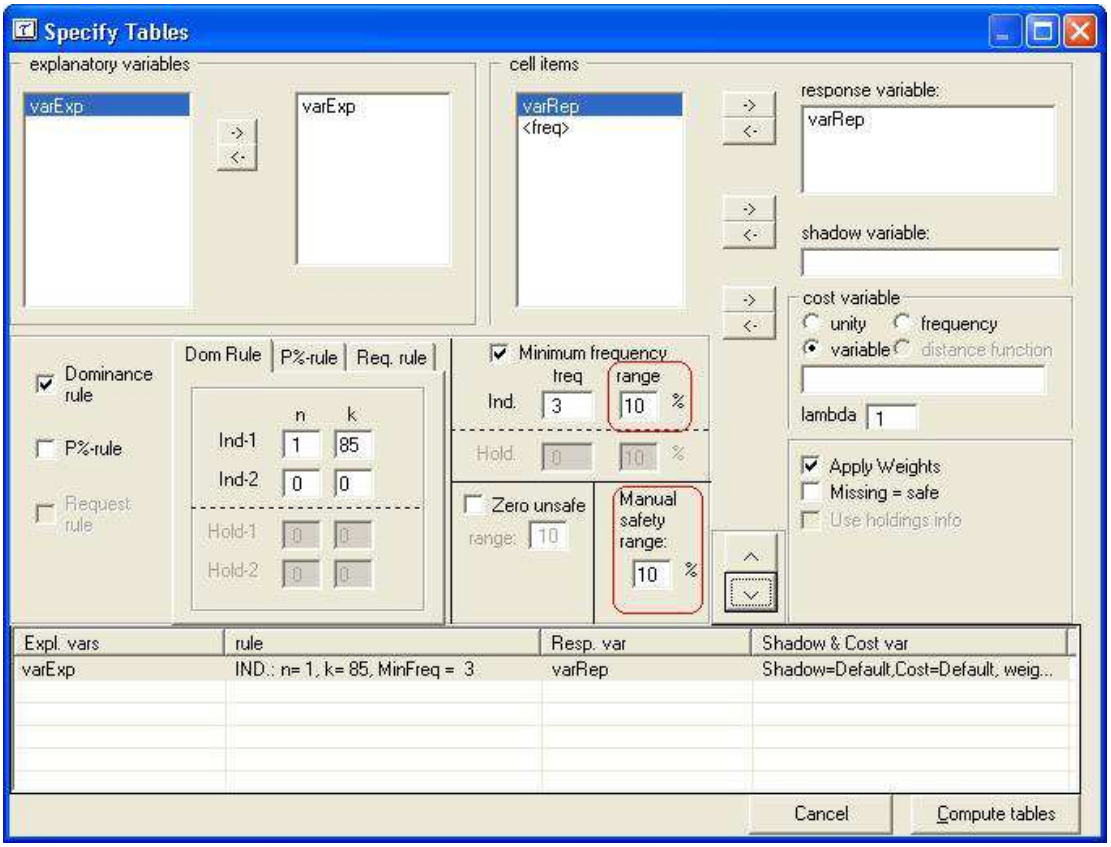
Ensuite, il convient, dans le choix des suppressions secondaires à effectuer, de ne pas supprimer des cases trop petites. En effet, lorsqu'une case est cachée, on peut en fait l'approcher par un intervalle en utilisant les relations d'additivité de la table et la positivité des cellules. Cet intervalle des possibles ne doit pas être trop fin. Sinon, la case sensible peut être approchée de façon très précise par un éventuel intrus.

Pour remédier à ce problème, on construit pour chaque cellule jugée sensible (secret primaire) un intervalle de protection, et les suppressions secondaires réalisées par la suite se font de façon à ce que les intervalles de protection soient inclus dans les intervalles des possibles.

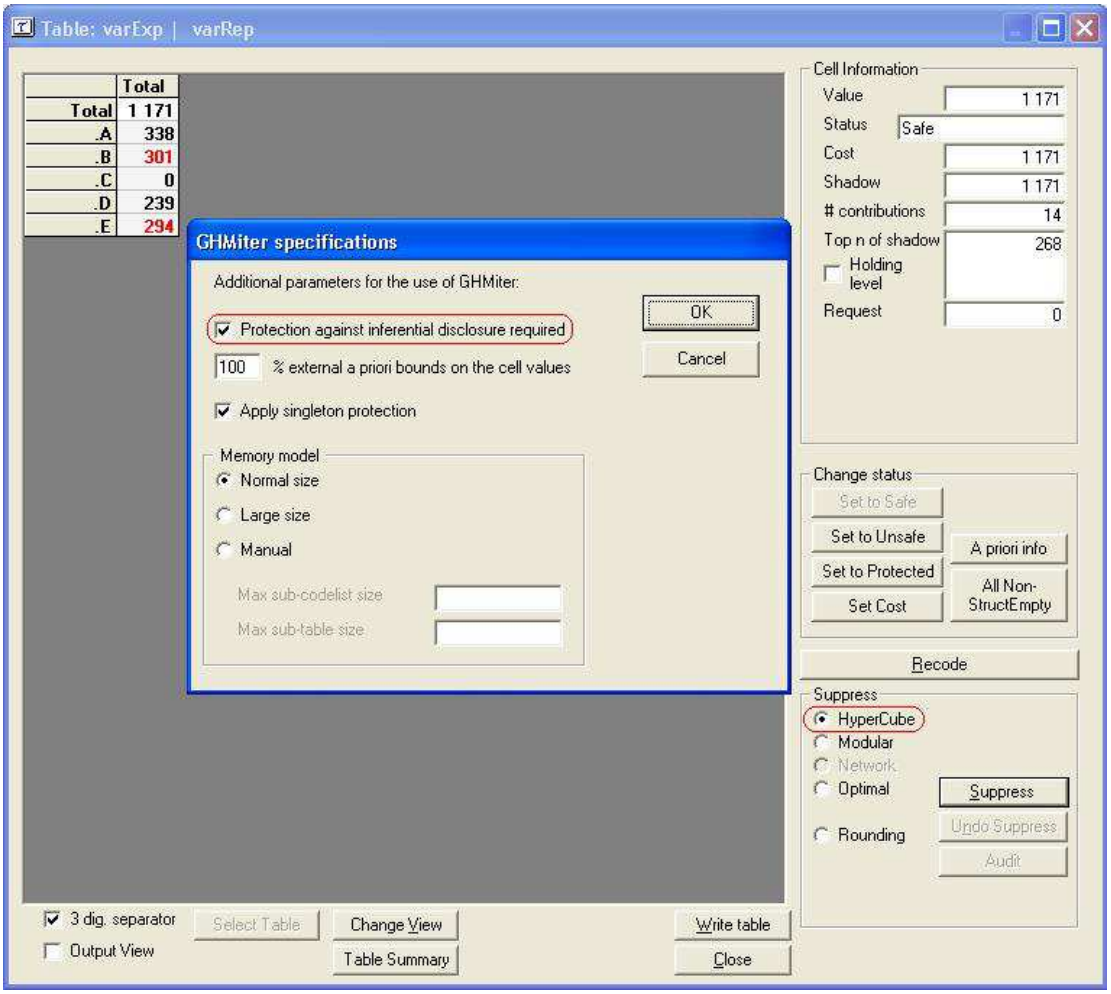
Le calcul de l'intervalle de protection est réalisé automatiquement lorsqu'il y a un problème de dominance (ou une atteinte à la règle du *p* %). Pour la règle de fréquence, il faut indiquer la largeur de l'intervalle de protection (relativement au total de la case) lors de l'étape de spécification des tables. Il en est de même si on désire forcer la protection d'une case (« manual safety range »). **Il est conseillé de choisir une largeur de 10 %**[3]. Les intervalles de protection sont ensuite construits de façon symétrique autour de la valeur de la case sensible.

[2] Pour plus de détails sur la comparaison entre les méthodes Hypercube, Modular et Optimal dans le cas où on disposerait de ces trois méthodes, on pourra par exemple se référer à l'étude suivante (en anglais) :
http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2013/Topic_4_France.pdf

[3] Il n'existe pas à l'heure actuelle de recommandations européennes en la matière, mais ce seuil de 10% est un standard préconisé par le groupe Eurostat d'experts sur la confidentialité des données.

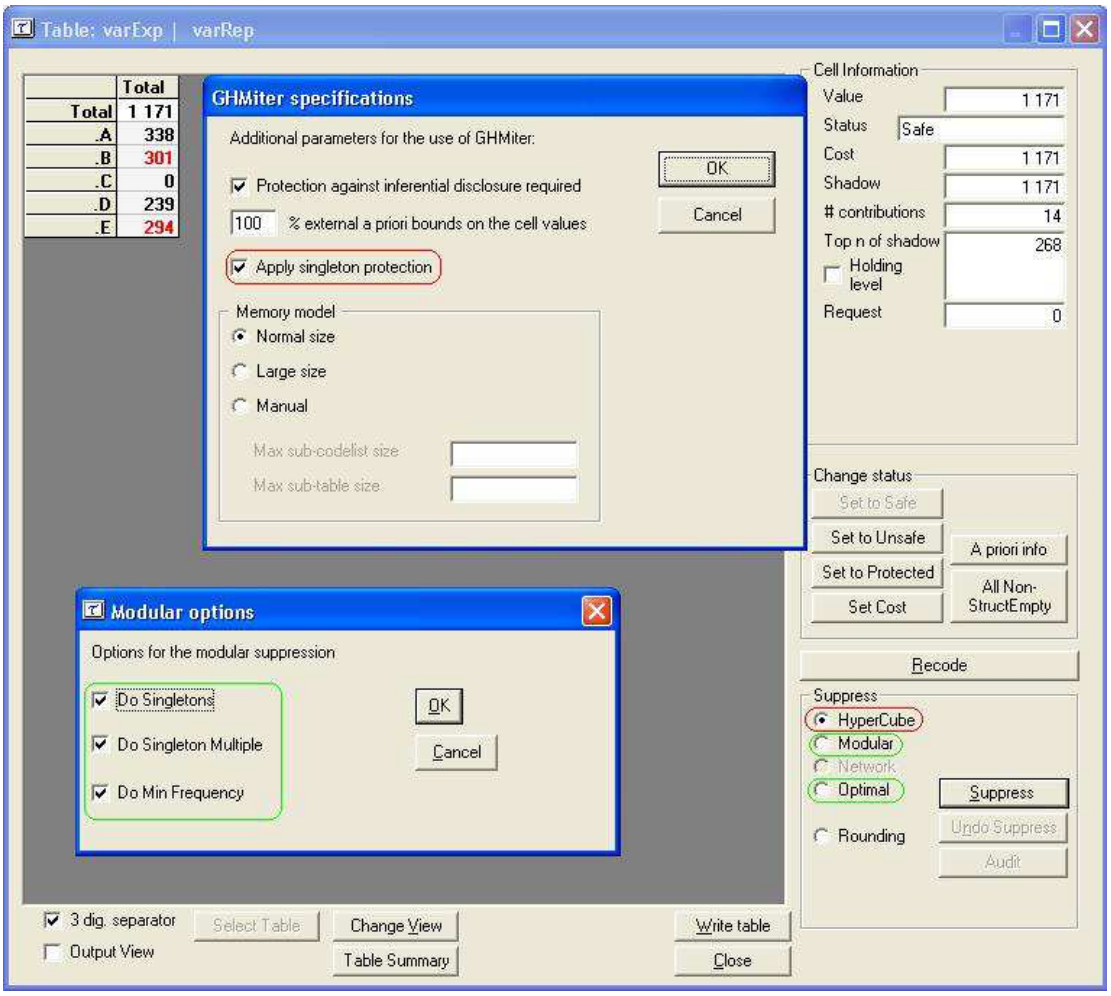


Pour chercher les suppressions secondaires, quand on utilise les méthodes Modular ou Optimal du logiciel Tau-Argus, la protection par intervalles est automatiquement réalisée. En revanche, en faisant appel à la solution Hypercube, on peut paramétrer le logiciel de façon à ne protéger que contre la divulgation exacte. L'utilisation de cette solution est à proscrire. Il convient de protéger contre la divulgation inférentielle en paramétrant l'utilisation de la méthode comme suit. Les hypothèses concernant les *a priori* sur les bornes des cellules sensibles (« external *a priori* bounds on the cell values ») sont prises égales à 100 %. Cela signifie qu'un intrus n'a pas d'autre indication sur le tableau que l'additivité et la positivité des cellules.



- Protection contre la divulgation des singletons

Enfin, il convient de protéger la table contre la divulgation des singletons, ou cellules avec un unique contributeur. Par exemple, si dans la même ligne ou la même colonne, on a deux singletons (donc en secret primaire pour respecter la règle des trois unités), une couche de secret secondaire supplémentaire doit être appliquée afin de protéger ces deux unités légales. En effet, même si, en apparence, la table est suffisamment protégée (on ne peut retrouver la valeur d'un des deux singletons en utilisant le total), en fait les deux contributeurs uniques des singletons peuvent déjouer le secret et retrouver la valeur de l'autre singleton.
Dans de tels cas, la protection secondaire ne suffit pas et il faut cacher des cases supplémentaires pour assurer la protection contre la divulgation des singletons. Les méthodes Hypercube, Modular et Optimal permettent d'assurer la protection des singletons.

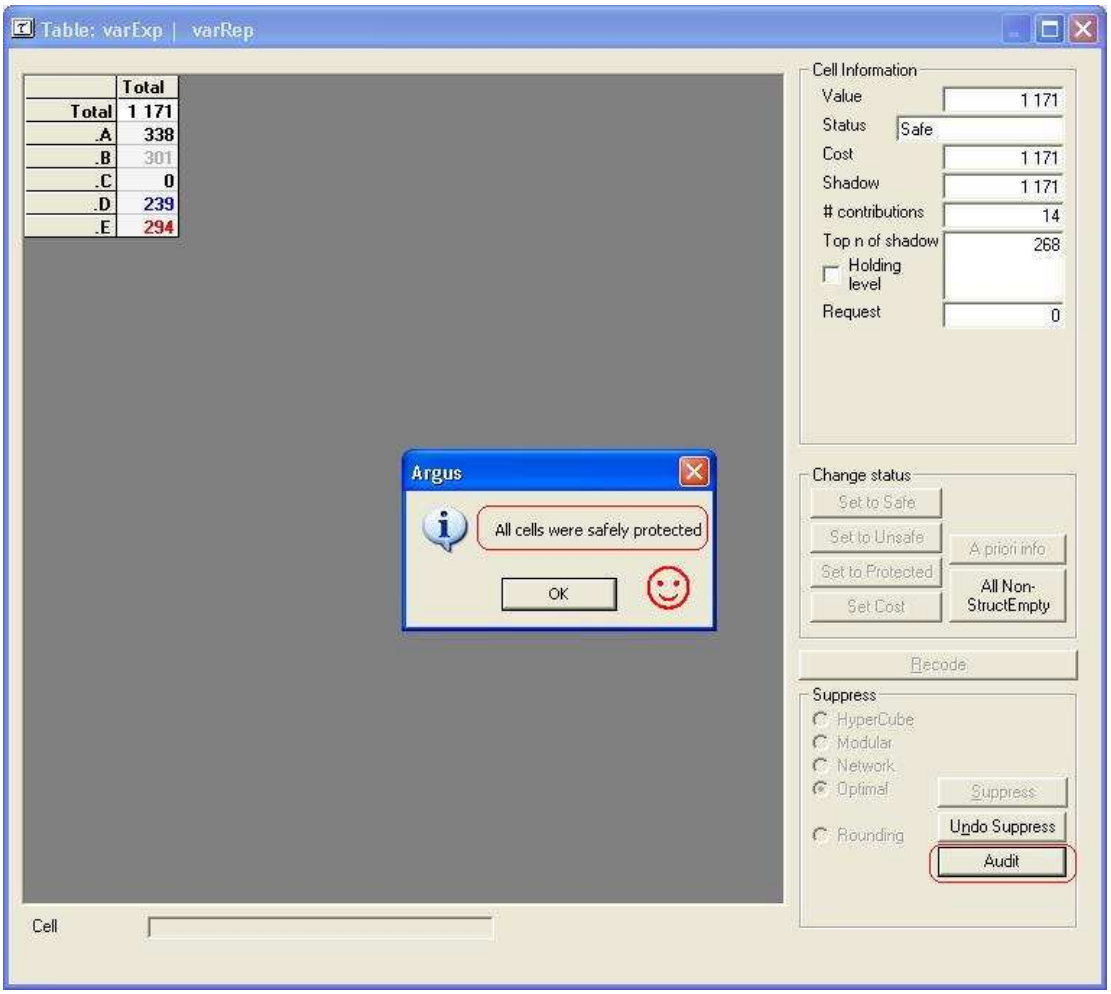


Pour les méthodes Modular et Optimal, trois sous-options sont proposées. Elles prennent en compte trois problèmes liés aux singletons (deux singletons dans la même ligne, un singleton et une autre cellule en secret primaire, deux cases en secret primaire à cause de la règle de fréquence dont la somme des fréquences est inférieure à la limite fixée pour la règle de fréquence minimale). Il est recommandé d'effectuer la protection pour ces trois types de problèmes.

Concernant la méthode Hypercube, la protection contre la divulgation des singletons est réalisée pour ces trois types de problèmes sans distinction dès lors que l'option « Apply singleton protection » est activée.

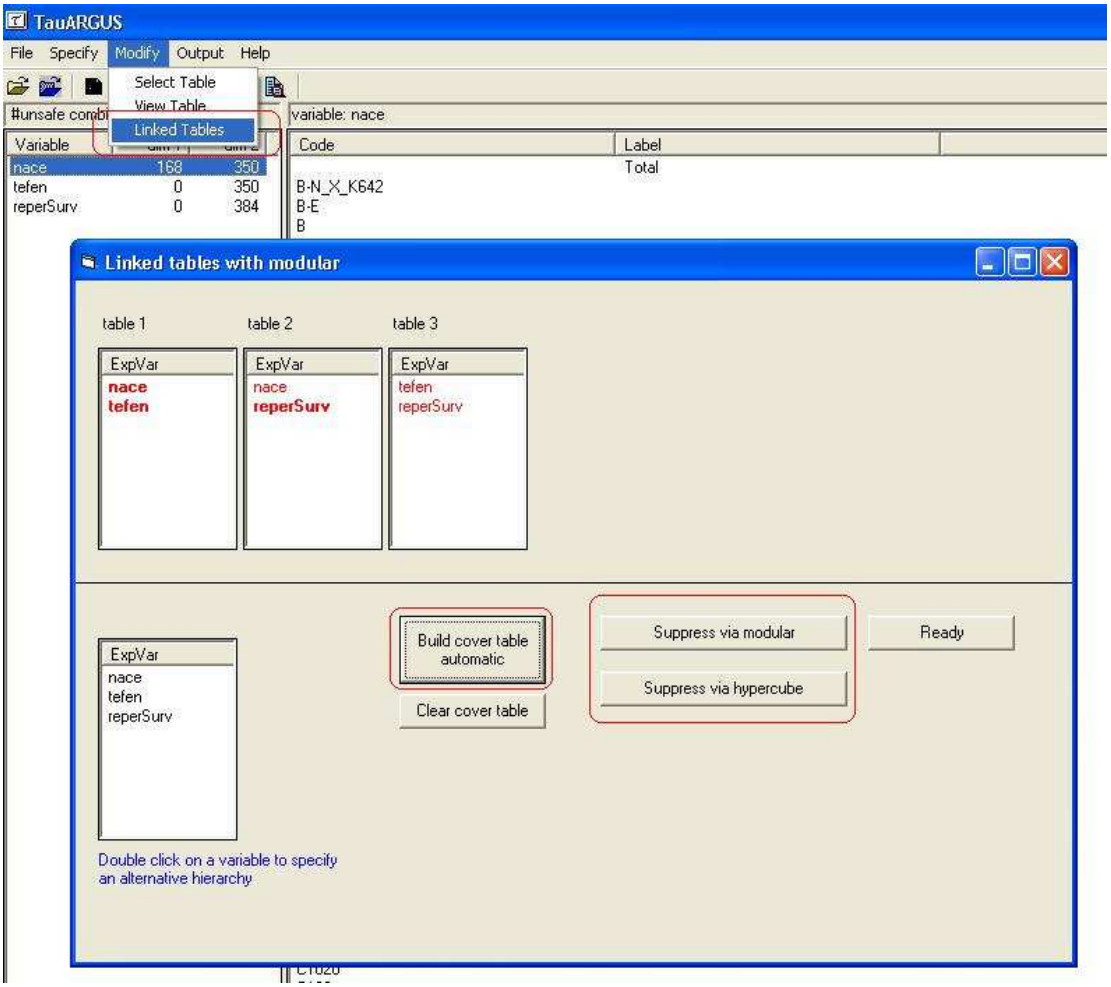
Après la recherche des suppressions secondaires, il est possible, sous réserve d'avoir accès aux fonctionnalités payantes de Tau-Argus[4], d'effectuer un « audit » de la table pour vérifier que les intervalles des possibles qu'on peut retrouver sont cohérents avec les intervalles de protection requis pour assurer la non-divulgation par intervalles. Il est conseillé d'effectuer cette vérification si cela est possible, qui permet de vérifier que la protection contre la divulgation par intervalles est bien réalisée.

[4] En effet, certaines fonctionnalités du logiciel nécessitent de recourir à un solveur de calcul payant (Xpress) qui permet de résoudre des programmes d'optimisation complexes. L'utilisation des méthodes Modular et Optimal pour la recherche du secret secondaire, ainsi que la procédure d'audit d'une table protégée, nécessitent en particulier le recours à ce solveur.

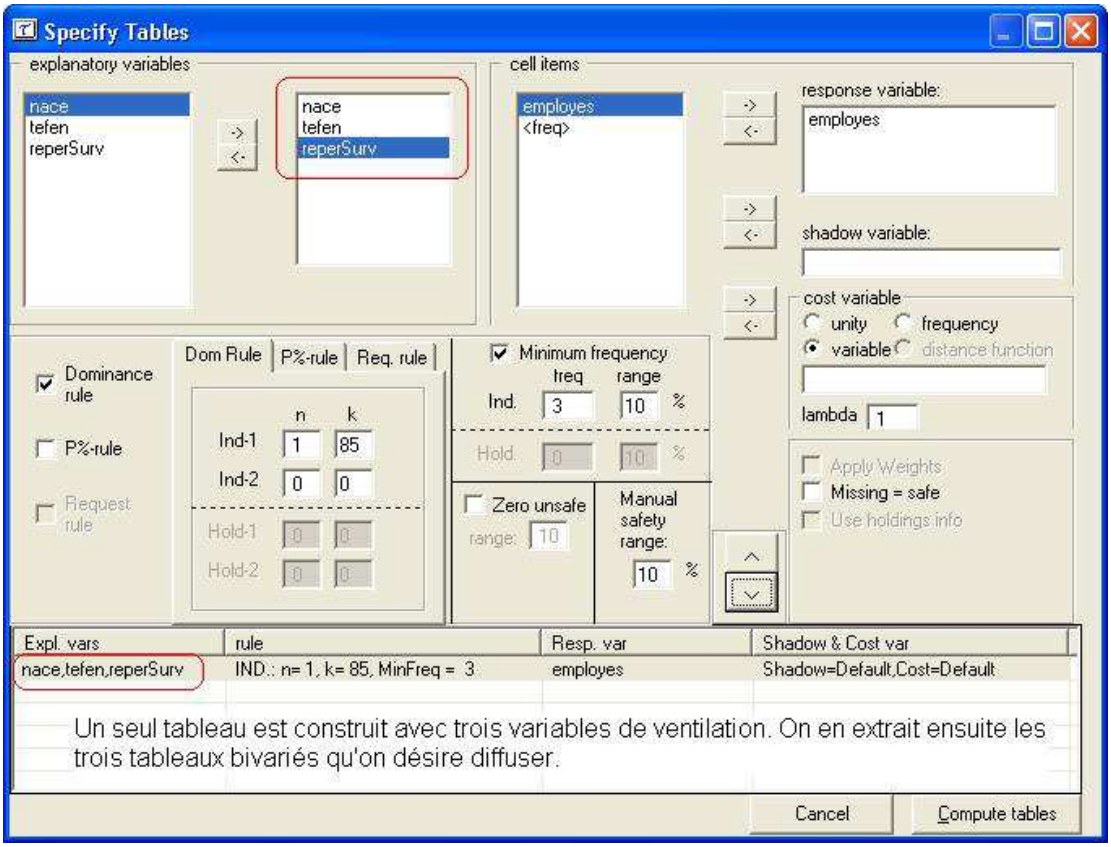


- Comment traiter simultanément une série de tableaux avec des variables de ventilation en commun ?

  Ce genre de problématique se rencontre souvent, par exemple pour les demandes Eurostat. Deux possibilités sont à envisager. Tout d'abord, le module « linked tables » permet de gérer simultanément le secret pour plusieurs tables. Les méthodes de protection Modular et Hypercube sont disponibles pour gérer la protection de tableaux liés.



Ensuite, il peut être intéressant de construire le tableau le plus fin (avec toutes les variables de ventilation comprises dans les tableaux liés), puis ensuite d'extraire les tables demandées. Cette solution est d'autant plus intéressante si on veut proposer à l'usager de construire son propre tableau *via* un requêteur.



- Comment effectuer la protection d'une table où les données diffusées sont des ratios, par exemple des moyennes ?

  La méthodologie utilisée dans Tau-Argus suppose que les données diffusées sont des agrégats : les marges diffusées correspondent à la somme des composantes correspondantes. Dans certains cas, la variable ventilée dans les tableaux est un ratio (une moyenne, une proportion, par exemple) :

$$R = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i y_i}$$

  Deux stratégies peuvent être utilisées pour protéger ce type de données en se ramenant à un problème classique.

  - Si on considère que la partie dénominateur est connue de l'utilisateur (ça peut être le cas si on s'intéresse à une proportion, par exemple), on peut effectuer la protection pour l'agrégat $\sum_{i=1}^{n} w_i x_i$.

- On peut également réaliser la protection pour les agrégats $\sum_{i=1}^{n} w_i x_i$ et $\sum_{i=1}^{n} w_i y_i$ et, dès que le numérateur **ou** le dénominateur doit être masqué, ne pas diffuser le ratio correspondant.

- Que diffuser concernant la méthodologie employée pour protéger les tableaux ?

  Concernant la diffusion aux utilisateurs des méthodes employées pour protéger les données, il ne faut pas donner trop d'indications qui permettraient de déjouer la protection apportée par Tau-Argus. En particulier, il est important dans les fichiers diffusés aux utilisateurs de ne pas distinguer les cases sous secret primaire et celles sous secret secondaire. En effet, dans certains cas, il est possible d'utiliser cette information auxiliaire pour déjouer la protection[5].

  De plus, connaître la fonction de coût utilisée pour choisir les suppressions secondaires peut également servir d'information auxiliaire pour déconstruire le processus de protection. Il peut par conséquent s'avérer judicieux de ne pas indiquer comment ont été sélectionnées les suppressions secondaires.

Le Chef de la division recueil et traitement de l'information

Signé : Gaël de Peretti

Pour information : Philippe Cuneo, Olivier Sautory, Benoît Rouppert, Odile Rascol

[5] Pour plus d'information sur les techniques utilisées pour déjouer la protection, on peut se rapporter au papier suivant (en anglais) : http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2013/Topic_4_Cox.pdf