Gérer la confidentialité des séries statistiques avec le logiciel Tau-Argus

L'exemple de l'Indice de Chiffre d'Affaires





Julien NICOLAS
Méthodologue, UMS-E

INSEE
INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES

20/12/2011

Introduction

But de la gestion de la confidentialité:

Protéger les données individuelles.

En empêchant la reconstruction de ces données à partir des diffusions.

Enjeux de la gestion de la confidentialité

Conserver la confiance des répondants aux enquêtes

afin de…

Récolter un maximum d'informations, de la meilleure qualité possible.

Garanties données aux répondants

Loi de 1951

- Définit le secret en matière de statistiques.
- Sanctions: 1 an d'emprisonnement, 15.000€ d'amende.

Code de bonnes pratiques de la statistique européenne

- Principe 5: Secret statistique
  - Le secret de la vie privée ou du secret des affaires des fournisseurs de données (ménages, entreprises, administrations et autres répondants), la confidentialité des informations qu'ils communiquent et l'utilisation de celles-ci à des fins strictement statistiques doivent être absolument garantis.
- 6 indicateurs:
  - Loi, engagement du personnel, sanctions, instructions fournies lors de la diffusion, dispositions matérielles et techniques, protocoles stricts d'accès aux extérieurs.

Jurisprudence française

Règle des 3 unités (fréquence minimale)

- Une cellule d'un tableau ne doit pas être construite à partir de moins de 3 unités.

Règle des 85% (dominance)

- Aucune des unités contributrices à une cellule ne doit contribuer à plus de 85% du total de celle-ci.

Règle de fréquence minimale

Une cellule est non sécurisée si le nombre de contributeurs au total de la cellule est strictement inférieur à un entier n.

Par exemple: n=3 (jurisprudence française).

Pour des données d'enquêtes, il faut considérer les poids.

Une cellule peut être construite à partir de seulement 2 répondants mais en représenter davantage.

Par contre, l'échantillon et la pondération associée doivent rester confidentiels.

Règle de dominance ou du (n,k)

Une cellule C est non sécurisée si la somme des n plus grandes contributions au total de la cellule dépasse k% du total de celle-ci:

$$x_1 \geq x_2 \geq \ldots \geq x_n \geq \ldots \geq x_{N(C)}$$

$$\sum_{i=1}^{n} x_i > \frac{k}{100} \sum_{i=1}^{N(C)} x_i$$

Interprétation:

On ne souhaite pas que les n plus gros contributeurs dominent trop le total de la cellule.

En France, on ne diffuse pas les agrégats pour lesquels une entreprise représente plus de 85% du total de l'agrégat.

Règle des 85% ou du (1,85)

Supposons que le total d'une cellule soit distribué comme suit:

Entreprise X: 81 000 000 €
Entreprise Y: 5 000 000 €
3 autres: 2 000 000 € chacune
Total: 92 000 000 €

Au regard de la règle (1,85), la cellule est non sécurisée car:

$$\frac{x_1}{\sum_{i=1}^{N(C)} x_i} = \frac{81 \cdot 10^6}{92 \cdot 10^6} \approx 0.88 > 0.85$$

Définition

Lors d'une diffusion, on vérifie que toutes les cellules d'un tableau respectent les règles fixées (fréquence minimale et règle de dominance).

Les cellules qui ne respectent pas l'une de ces règles, seront mises sous secret: ce sont les secrets primaires.

Que souhaite-t-on protéger ?

Indice de CA théorique:

$$I(n,m,sc) = \frac{CA(n,m,sc)}{\frac{1}{12}\sum_{m=1}^{12} CA(n_0,m,sc)} \times 100$$

sc: la sous-classe d'activité de diffusion (NACE)
m: mois
n: année
n_0: année de base (2005)

Sur quoi est alors posé le secret ?

On pose en fait le secret annuellement.

Le secret est posé sur la moyenne des CA mensuels de l'année n-1.

$$CA_{moyen,n-1}(sc) = \frac{1}{12}\sum_{m=1}^{12} CA(n-1,m,sc)$$

Le secret calculé est alors utilisé pour tous les mois de l'année n.

Le secret primaire

On repère dans un premier temps, tous les secteurs qui ne respectent pas les règles de secret fixées.

On vérifie qu'aucun secteur d'activité ne se rapporte à moins de trois unités (règle de fréquence minimale).

On vérifie qu'aucune unité ne domine à plus de 85% l'agrégat (ici, le CA) d'un secteur (règle de dominance).

Les secteurs qui ne respectent pas les règles ne sont pas diffusés.

Le secret primaire

Nace	CA moyen en n-1	Secret
…	…	…
10.4	$CA_{moyen,n-1}(10.4)$	V
10.41	$CA_{moyen,n-1}(10.41)$	V
10.41A	$CA_{moyen,n-1}(10.41A)$	V
10.41B	$CA_{moyen,n-1}(10.41B)$	V
10.42	$CA_{moyen,n-1}(10.42)$	A
10.42Z	$CA_{moyen,n-1}(10.42Z)$	A
10.5	$CA_{moyen,n-1}(10.5)$	V
…	…	…

Nous utilisons des codes **en interne** pour différencier les secteurs diffusables des secteurs non diffusables:

V: diffusable
A: non diff. (fréquence)
B: non diff. (dominance)

NB:
Lors de la mise à disposition des données au public, tous les agrégats non diffusables sont remplacés par un symbole.

Problème !

Il faut prendre en compte la structure de la variable de ventilation, ici la variable d'activité: Nace.

C'est une structure hiérarchisée: présence de totaux intermédiaires.

Des secteurs d'activités sont la simple agrégation de secteurs d'activités plus fins.

Le secret secondaire

Nace	CA moyen en n-1	Secret
…	…	…
10.4	**$CA_{moyen,n-1}$(10.4)**	**V**
10.41	$CA_{moyen,n-1}$(10.41)	V
10.41A	$CA_{moyen,n-1}$(10.41A)	V
10.41B	$CA_{moyen,n-1}$(10.41B)	V
10.42	$CA_{moyen,n-1}$(10.42)	A
10.42Z	$CA_{moyen,n-1}$(10.42Z)	A
10.5	**$CA_{moyen,n-1}$(10.5)**	**V**
…	…	…

C'est une structure hiérarchisée:

10.4 = 10.41 + 10.42

10.41 = 10.41A + 10.41B

10.42 = 10.42Z

On a même:

10=10.1+10.2+10.3+10.4+10.5+ 10.6+10.7

Le secret secondaire

Nace	CA moyen en n-1	Secret
…	…	…
10.4	**$CA_{moyen,n-1}$(10.4)**	**V**
10.41	$CA_{moyen,n-1}$(10.41)	D
10.41A	$CA_{moyen,n-1}$(10.41A)	V
10.41B	$CA_{moyen,n-1}$(10.41B)	D
10.42	$CA_{moyen,n-1}$(10.42)	A
10.42Z	$CA_{moyen,n-1}$(10.42Z)	A
10.5	**$CA_{moyen,n-1}$(10.5)**	**V**
…	…	…

La présence de totaux intermédiaires permet parfois de reconstruire les secteurs mis sous secret primaire.

Il faut procéder à de nouvelles mises sous secret: c'est ce que l'on appelle la pose du secret secondaire.

V: diffusable
A: non diff. (fréquence)
B: non diff. (dominance)
D: secret secondaire

La diffusion

Nace	Indice en (n,m)
…	…
10.4	$I_{n,m}$(10.4)
10.41	S
10.41A	$I_{n,m}$(10.41A)
10.41B	S
10.42	S
10.42Z	S
10.5	$I_{n,m}$(10.5)
…	…

Lors de la diffusion, nous appliquons le secret calculé sur le CA moyen en n-1.

On remplace tous les codes de secret par un même symbôle.

Comment réalise-t-on ce travail ?

La recherche des secteurs sous secret primaire, puis la recherche du secret secondaire sont effectuées au moyen d'un logiciel: Tau-Argus.

Démonstration du logiciel Tau-Argus.

Eurostat, Statistics Netherlands

Démonstration de la macro SAS %TauArgus.

D. Ladiray (code), J. Nicolas (spécif.)

Une alternative ?

La pose du secret secondaire amène à mettre en secret de nouveaux indices.

Nous nous sommes demandés si une alternative à cela était envisageable afin de permettre la diffusion d'un maximum d'indice.

Une piste ? (1)

Le schéma d'agrégation de la variable CA est simple:

L'agrégat d'un secteur d'activité est la somme des agrégats des sous-secteurs associés.

On a par exemple:

$$CA_{moyen,n-1}(10.41) = CA_{moyen,n-1}(10.41A) + CA_{moyen,n-1}(10.41B)$$
$$CA_{moyen,n-1}(10.4) = CA_{moyen,n-1}(10.41) + CA_{moyen,n-1}(10.42)$$

Alors que…

Une piste ? (2)

Nace	Indice en (n,m)	Poids
…	…	…
10.4	$I_{n,m}(10.4)$	$P_{10.4}$
10.41	$I_{n,m}(10.41)$	$P_{10.41}$
10.41A	$I_{n,m}(10.41A)$	$P_{10.41A}$
10.41B	$I_{n,m}(10.41B)$	$P_{10.41B}$
10.42	$I_{n,m}(10.42)$	$P_{10.42}$
10.42Z	$I_{n,m}(10.42Z)$	$P_{10.42Z}$
10.5	$I_{n,m}(10.5)$	$P_{10.5}$
…	…	…

Le schéma d'agrégation des indices utilise une pondération:

$$I_{n,m}(10.4) = \frac{P_{10.41} \cdot I_{n,m}(10.41) + P_{10.42} \cdot I_{n,m}(10.42)}{P_{10.4}}$$

avec

$$P_{10.4} = P_{10.41} + P_{10.42}$$

Ainsi, il n'est pas possible de reconstruire l'indice du secteur 10.42.

Une piste ? (3)

L'idée est de travailler simultanément sur la variable de pose du secret (ici le CA), et sur les poids associés aux indices.

1) On repère les indices sous secret primaire grâce au CA moyen en n-1.

2) On reporte ces secrets à l'identique sur les poids associés à ces indices.

3) On recherche les poids à mettre sous secret secondaire car des relations existent entre les poids.

$$P_{10.4} = P_{10.41} + P_{10.42}$$

Une piste ? (4)

Nace	Indice en (n,m)	Poids
…	…	…
10.4	$I_{n,m}(10.4)$	$P_{10.4}$
10.41	$I_{n,m}(10.41)$	D
10.41A	$I_{n,m}(10.41A)$	$P_{10.41A}$
10.41B	$I_{n,m}(10.41B)$	D
10.42	A	A
10.42Z	A	A
10.5	$I_{n,m}(10.5)$	$P_{10.5}$
…	…	…

$$I_{n,m}(10.4) = \frac{\overbrace{P_{10.41}}^{V} \cdot \overbrace{I_{n,m}(10.41)}^{D} + \overbrace{P_{10.42}}^{V} \cdot \overbrace{I_{n,m}(10.42)}^{A}}{P_{10.4}}$$

$$I_{n,m}(10.4) = \frac{\overbrace{P_{10.41}}^{D} \cdot \overbrace{I_{n,m}(10.41)}^{V} + \overbrace{P_{10.42}}^{D} \cdot \overbrace{I_{n,m}(10.42)}^{A}}{P_{10.4}}$$

Une impossibilité ! (1)

Les pondérations utilisées sont fixes pour une durée de 5 ans….

Dans cet exemple: $$I_{n,m}(10.4)=\frac{\overbrace{P_{10.41}}^{D}\cdot\overbrace{I_{n,m}(10.41)}^{V}+\overbrace{P_{10.42}}^{D}\cdot\overbrace{I_{n,m}(10.42)}^{A}}{P_{10.4}}$$

Nous avons uniquement que 2 inconnues, car:

$$P_{10.4}=P_{10.41}+P_{10.42}$$

D'où: $$I_{n,m}(10.4)=\frac{\overbrace{P_{10.41}}^{D}\cdot\overbrace{I_{n,m}(10.41)}^{V}+\overbrace{(P_{10.4}-P_{10.41})}^{D}\cdot\overbrace{I_{n,m}(10.42)}^{A}}{P_{10.4}}$$

Impossibilité ! (2)

Il va donc être possible de retrouver $I_{n,m}(10.42)$.

Car: $$I_{n,m}(10.4)=\frac{\overbrace{P_{10.41}}^{D}\cdot\overbrace{I_{n,m}(10.41)}^{V}+\overbrace{(P_{10.4}-P_{10.41})}^{D}\cdot\overbrace{I_{n,m}(10.42)}^{A}}{P_{10.4}}$$

$$et$$

$$n=1,...,12$$

Au bout de 2 mois de publication, on pourra…

- Recalculer les pondérations cachées, et donc…
- Recalculer l'indice sous secret.

Gérer la confidentialité des séries statistiques avec le logiciel Tau-Argus

Merci de votre attention !

Contact
M. NICOLAS Julien
Tél. : +33 (0) 1 41 17 6486
Courriel : julien.nicolas@insee.fr

Insee
18 bd Adolphe-Pinard
75675 Paris Cedex 14
www.insee.fr
Informations statistiques :
www.insee.fr / Contacter l'Insee
09 72 72 4000
(coût d'un appel local)
du lundi au vendredi de 9h00 à 17h00