Nº 11-522-X au catalogue
ISSN 1709-822X

Recueil du Symposium de 2022 de Statistique Canada : Désagrégation des données : dresser un portrait de données plus représentatif de la société

Outils de vérification des données de sortie du DataLab de l'ABS

par Chien-Hung Chien

Date de diffusion : le 25 mars 2024

Statistique Canada Statistics Canada

Canada

Recueil du Symposium de 2022 de Statistique Canada
Désagrégation des données : Dresser un portrait de données plus représentatif de la société

Outils de vérification des données de sortie du DataLab de l'ABS

Chien-Hung Chien[1]

Résumé

L'ABS (Bureau de la statistique de l'Australie) s'est engagé à offrir un meilleur accès à un plus grand nombre de microdonnées, tout en assurant la protection des renseignements personnels et la confidentialité, grâce à son DataLab (laboratoire de données) virtuel qui aide à entreprendre des recherches complexes plus efficacement. À l'heure actuelle, pour être autorisées, les données de sortie de recherche de DataLab doivent respecter des règles strictes afin de minimiser les risques de divulgation. Or le processus d'examen manuel n'est pas rentable et peut entraîner des erreurs. Le nombre croissant de résultats statistiques provenant de différents projets peut introduire des risques de divulgation résiduelle même si ces résultats de différents projets ont respecté des règles strictes en matière de sorties. L'ABS étudie la possibilité de fournir une vérification automatique des sorties au moyen de la méthodologie de la clé de cellule de l'ABS pour s'assurer que toutes les sorties des différents projets sont protégées de façon uniforme afin de minimiser les risques de divulgation résiduelle et de réduire les coûts associés à la vérification des sorties.

Mots clés : clé de cellule de l'ABS, vérification des sorties, cinq éléments de la sécurité

1. Introduction

Le DataLab de l'ABS, Bureau de la statistique de l'Australie, procure aux chercheurs un environnement sûr pour accéder à une variété de microdonnées sur les ménages et les entreprises[2]. Le DataLab de l'ABS est une solution d'analyse de données pour les utilisateurs de données professionnels qui souhaitent extraire toute la valeur des microdonnées de l'ABS (Parker, 2017). Il permet d'entreprendre des recherches complexes. Le nombre de sessions de DataLab a considérablement augmenté depuis 2019-2020, avec un accès à 15 520 séances en 2020-2021 et 24 037 séances en 2021-2022. À l'heure actuelle, l'autorisation des résultats de sortie de l'ABS est réalisée par un processus de vérification manuelle qui n'est ni évolutif, ni rentable, ni exempt d'erreurs humaines. Il existe également un risque que le nombre croissant de sorties de différents projets présente des risques de divulgation résiduelle, même si elles satisfont individuellement à des critères de sortie stricts.

Cette recherche s'appuie sur des recherches antérieures de l'ABS[3] et applique le concept des cinq éléments de la sécurité proposé par Desai et coll. (2016) pour fournir des outils de vérification des sorties aux chercheurs approuvés du DataLab de l'ABS. Ces chercheurs approuvés ont signé un engagement sur le mode d'utilisation des microdonnées et ont suivi une formation pour comprendre leurs responsabilités en matière d'utilisation efficace des microdonnées (Parker, 2017). Dans le contexte de l'accès sécurisé aux microdonnées confidentielles du secteur public dans le DataLab de l'ABS, ces chercheurs font partie d'une communauté d'utilisateurs sûrs qui travaillent à accroître la valeur des microdonnées du secteur public. Green et coll. (2017) ont constaté que les chercheurs sont plus susceptibles d'agir de façon appropriée lorsqu'ils sont considérés comme dignes de confiance et que le fait de faire partie d'une communauté entraîne des interventions positives. L'ABS coopérera avec des chercheurs de confiance pour rendre le processus de vérification des sorties évolutif en fournissant ces outils de vérification des sorties afin de faciliter et automatiser le processus.

---

[1] Chien-Hung Chien, Australian Bureau of Statistics, 45 Benjamin Way, Belconnen, Australie, ACT 2617 (joseph.chien@abs.gov.au)

[2] D'autres organismes de statistique officiels proposent aussi un service semblable par accès à distance ou sur place. Citons par exemple les Centres de données de recherche de Statistique Canada, le Stats New Zealand Data Lab, l'Office for National Statistics Secure Research Service et les Federal Statistical Research Data Centers des États-Unis.

[3] Voir Thompson et coll., 2013; Chipperfield and O'Keefe, 2014; et O'Keefe et coll., 2017.

Le présent article est organisé comme suit : la section 2 décrit le DataLab de l'ABS; la section 3 résume le fonctionnement des outils prototypes; la section 4 traite des considérations relatives à l'élaboration d'outils prototypes; et la section 5 propose une conclusion et des orientations dans la perspective de futures recherches.

2. DataLab de l'ABS et vérification des sorties

L'ABS utilise le cadre des cinq éléments de la sécurité proposé par Desai et coll. (2016) pour permettre aux chercheurs d'accéder à son DataLab. Le DataLab de l'ABS permet aux chercheurs d'analyser virtuellement des données d'enregistrement d'unité allant de fichiers de microdonnées de base à des fichiers de microdonnées détaillés et intégrés dans l'environnement sécurisé de l'ABS. L'environnement possède des versions récentes de logiciels d'analyse, notamment R, SAS, Stata et Python. Toutes les sorties analytiques que les chercheurs veulent utiliser hors de DataLab sont vérifiées par l'ABS avant leur diffusion (voir la description détaillée du processus dans Parker, 2017). Le contrôle des sorties joue un rôle important dans la minimisation des risques de divulgation statistiques. La vérification des sorties est utilisée couramment par d'autres organismes de statistique officiels (voir Stats NZ, 2020 et ONS, 2020). Afin d'assurer le respect des règles de confidentialité et de protection des renseignements personnels, les résultats du DataLab de l'ABS suivent des procédures strictes pour minimiser les risques de divulgation avant la diffusion des résultats.

3. Outils prototypes de vérification des données de sortie de l'ABS

L'ABS a mis au point des prototypes d'outils de protection des sorties tabulaires pour les données de dénombrement, les données continues et les résultats de régression, qui appliquent de façon uniforme la méthodologie de la clé de cellule de l'ABS aux sorties agrégées, minimisant ainsi les risques de divulgation résiduelle et réduisant les coûts de la vérification des sorties. Les outils prototypes actuels sont construits dans R[4]. Nous avons sélectionné quelques variables à partir de microdonnées accessibles au public[5] de la bibliothèque ***synthpop*** (voir Nowok et coll., 2016) et d'un calepin Jupyter pour présenter ces outils R. La figure 2-1 montre un exemple d'utilisation d'une fonction de tableau (***Conf_Table***) par les chercheurs pour produire un tableau de dénombrement et de revenu selon le niveau de scolarité atteint. Les chercheurs peuvent choisir d'afficher des tableaux agrégés protégés et originaux.

Figure 2-1
Tableaux dont la confidentialité est assurée

```
In [4]: Conf_Table(df=example,variable="income",factors=c("edu"),check=TRUE)
```

A data.table: 5 × 5

edu	pertcounts	counts	pertsum	sums
<fct>	<int>	<int>	<dbl>	<dbl>
VOCATION	913	914	1381340.0	1380168
PRIMARY	713	711	811787.1	805377
SECONDARY	957	956	1681874.4	1671247
HIGHER	666	665	1706695.6	1716742
sum	3247	3246	5559553.2	5573534

Le panneau de gauche de la figure 2-2 montre comment produire des résultats de régression protégés et originaux[6]. Les chercheurs peuvent également produire un graphique hexbin[7] pour montrer la relation entre les valeurs prédites

[4] Il est possible de construire une version similaire dans Python. Le Protari de l'ABS est construit dans Python (voir Protari).
[5] Nous utilisons le revenu, le niveau de scolarité, l'âge et le statut social du Diagnostic social 2011 – Qualité de vie objective et subjective en Pologne.
[6] O'Keefe et coll. (2017) analysent en détail la méthode robuste linéaire mise en œuvre pour les outils de vérification des résultats de la régression.
[7] Un graphique hexbin divise le domaine dans un graphique en hexagones mosaïque, puis nuance chaque hexagone en fonction du nombre d'observations qui se produisent dans cet hexagone.

protégées et les résidus à des fins de diagnostic du modèle. Un graphique hexbin permet aux utilisateurs de voir plus facilement les tendances tout en offrant un niveau élevé de protection contre la divulgation.

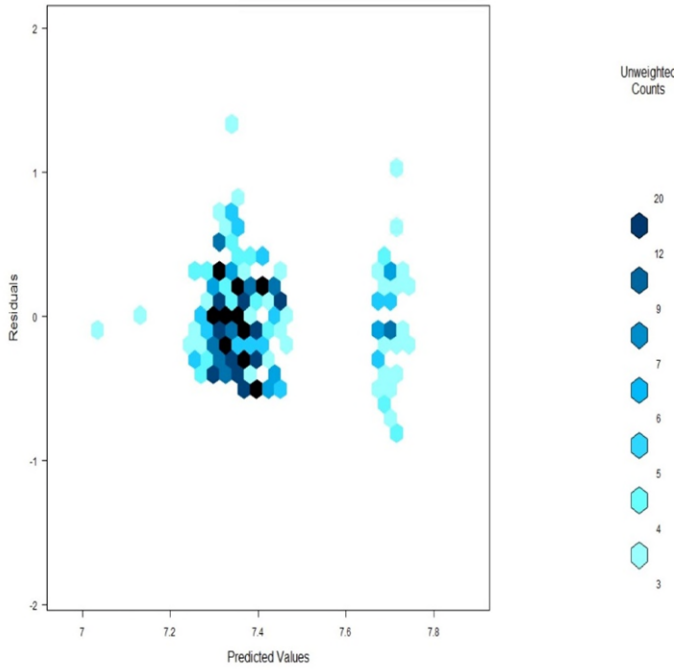
Figure 2-2
Outils prototypes de vérification des données de sortie pour les résultats de régression

fonction de régression (*glm_f*)

```
display_html(paste(capture.output(print(glm_f(
    type ="html"))), collapse="", sep=""))
```

	Dependent variable:	
	income_log	
	Linear	Conf. linear (Robust)
	(1)	(2)
	glm	glm
age	0.002***	0.002**
	(0.001)	(0.001)
socprofPUBLIC	0.154***	0.179***
	(0.028)	(0.025)
socprofSELF	0.372***	0.381***
	(0.048)	(0.044)
socprofFARMER	-0.466***	-0.414***
	(0.041)	(0.037)
socprofDISABLED	-0.630***	-0.616***
	(0.036)	(0.033)
socprofRETIRED	-0.379***	-0.350***
	(0.034)	(0.031)
Constant	7.341***	7.346***
	(0.038)	(0.034)
Note:		*p<0.1; **p<0.05; ***p<0.01

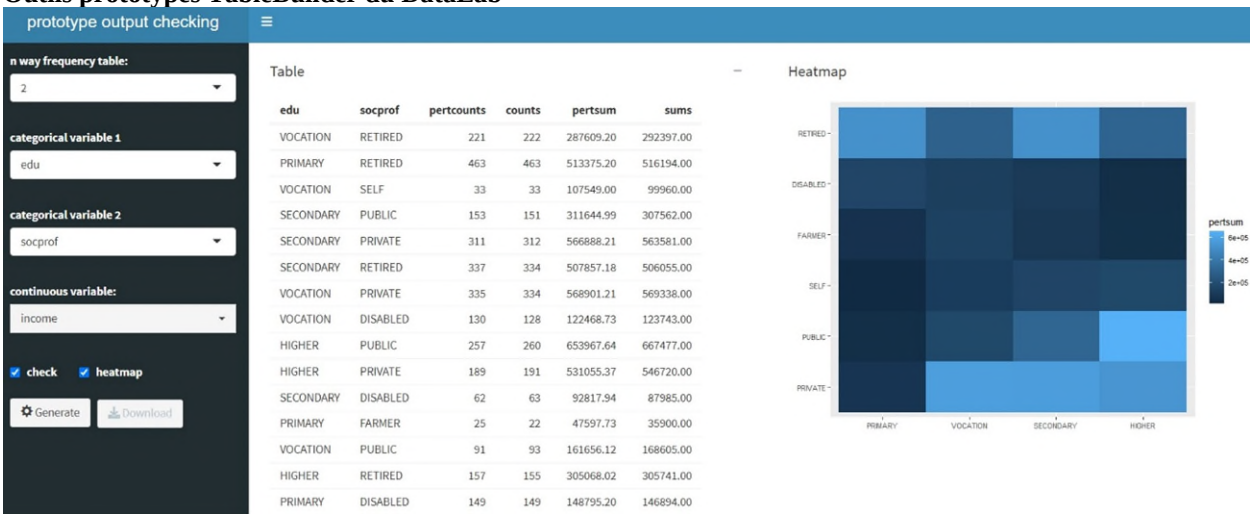
Graphique hexbin



Nous sommes conscients du fait que les chercheurs du DataLab de l'ABS sont plus ou moins à l'aise quand il s'agit d'utiliser un langage de programmation. C'est pourquoi nous avons développé une interface utilisateur de base qui interagit avec les fonctions de R. (Voir la figure 2-3). Les utilisateurs peuvent simplement transférer leur base de données d'analyse aux outils et produire des sorties sécurisées pour faciliter le processus de vérification des sorties. Les chercheurs peuvent produire une carte thermique servant leur analyse. Les outils de vérification des sorties

comportent des processus intégrés qui assurent la conformité. Par exemple, les chercheurs ne peuvent pas télécharger les fichiers CSV contenant les données originales du tableau.

Figure 2-3
Outils prototypes TableBuilder du DataLab



À l'heure actuelle, on a sélectionné des chercheurs du DataLab pour mettre à l'essai une version du prototype pour les données de dénombrement. Nous recueillons leurs commentaires afin d'améliorer la convivialité du prototype et d'adopter ses méthodes pour d'autres utilisations.

4. Considérations relatives au développement

Il reste à examiner une liste de questions en suspens pour le développement futur des outils de vérification des sorties, notamment :

- Les outils de vérification des sorties répondent-ils aux besoins des chercheurs?
- Les outils peuvent-ils répondre aux besoins d'utilisateurs aux capacités différentes?
- Comment pouvons-nous minimiser les erreurs dans l'utilisation des outils?
- Devrions-nous acheter des solutions dans le commerce ou développer des outils à l'interne?
- Comment pouvons-nous créer des outils conviviaux dont la maintenance est facile?

Il y a eu des développements importants dans la communauté du code de source ouverte[8] et les organismes nationaux de statistique dans la visualisation et les outils statistiques pour les sorties. Il est possible de créer des ensembles d'outils communs et durables en collaboration, qui assureront la sécurité des données de sortie puisque, de plus en plus, la diffusion se fait par l'intermédiaire des centres de données de recherche.

5. Conclusion et orientations en vue de futures recherches

Le DataLab de l'ABS est une solution d'analyse conçue pour les chercheuses et chercheurs qui veulent entreprendre une analyse complexe en temps réel de microdonnées détaillées. Le DataLab de l'ABS est devenu un outil de diffusion important permettant d'utiliser davantage les microdonnées du secteur public à des fins de recherche. On constate une croissance importante de la demande d'accès au DataLab de l'ABS. Le processus actuel de vérification manuelle des sorties de tous les résultats du DataLab n'est ni évolutif ni viable.

[8] Des progrès récents ont été réalisés dans le domaine des outils de visualisation avancés dans la communauté du code de source ouverte (voir Martoglio, 2018 et Kruchten, 2022).

L'automatisation du processus de vérification manuelle des sorties présente plusieurs avantages importants : elle permet notamment un processus plus ciblé et axé sur des processus complexes non automatisables. L'ABS et d'autres organismes de statistique souhaiteraient mettre au point des outils pour rendre le processus plus efficace et ainsi minimiser les erreurs.

Des discussions sont en cours sur l'affectation des ressources et les processus pour développer davantage les outils et les rendre plus largement disponibles dans le DataLab de l'ABS. Il existe des perspectives de collaboration entre organismes officiels de statistique et communautés du code de source ouverte pour l'élaboration de développer des ressources durables.

Bibliographie

Chipperfield, J. O. et C. M. O'keefe (2014), « Disclosure-protected Inference Using Generalised Linear Models », *International Statistical Review*, 82**,** p. 371-391.

Desai, T., Ritchie, F. et R. Welpton (2016), Five Safes: Designing data access for research.

Green, E., Ritchie, F., Newman, J. et T. Parker (2017), « Lessons learned in training 'safe users' of confidential data », Atelier sur la confidentialité des données statistiques.

Kruchten, N. (2022), *PivotTable.js - an open-source Javascript Pivot Table* [En ligne]. Disponible à l'adresse : https://pivottable.js.org/examples/

Martoglio, E. (2018), *rpivotTable: Build Powerful Pivot Tables and Dynamically Slice & Dice your Data* [En ligne]. Disponible à l'adresse : https://CRAN.R-project.org/package=rpivotTable

Nowok, B., Raab, G. M. et C. Dibben (2016), « synthpop: Bespoke creation of synthetic data in R », *Journal of statistical software*, 74**,** p. 1-26.

ONS (2020), SRS Researcher Output Clearance Guidance, août 2020, par Stats NZ Tatauranga Aotearoa, Wellington, Nouvelle-Zélande.

Parker, T. (2017), « The DataLab of the Australian bureau of statistics », *Australian Economic Review*, 50**,** p. 478-483.

Stats NZ (2020), Microdata output guide, août 2020, par Stats NZ Tatauranga Aotearoa, Wellington, Nouvelle-Zélande.

Thompson, G., Broadfoot, S. et D. Elazar (2013), « Methodology for the automatic confidentialisation of statistical outputs from remote servers at the Australian Bureau of Statistics », Atelier conjoint UNECE/Eurostat sur la confidentialité des données statistiques, Ottawa, Canada.