Catalogue no. 11-522-X
ISSN 1709-8211

Proceedings of Statistics Canada Symposium 2022:
Data Disaggregation: building a more representative data portrait of society

ABS DataLab output checking tools

by Chien-Hung Chien

Release date: March 25, 2024

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2022
Data Disaggregation: Building a more representative data portrait of society

ABS DataLab output checking tools

Chien-Hung Chien[1]

Abstract

The Australian Bureau of Statistics (ABS) is committed to improving access to more microdata, while ensuring privacy and confidentiality is maintained, through its virtual DataLab which supports researchers to undertake complex research more efficiently. Currently, the DataLab research outputs need to follow strict rules to minimise disclosure risks for clearance. However, the clerical-review process is not cost effective and has potential to introduce errors. The increasing number of statistical outputs from different projects can potentially introduce differencing risks even though these outputs from different projects have met the strict output rules. The ABS has been exploring the possibility of providing automatic output checking using the ABS cellkey methodology to ensure that all outputs across different projects are protected consistently to minimise differencing risks and reduce costs associated with output checking.

Key Words: ABS cellkey, output checking, five safes

1. Introduction

The Australian Bureau of Statistics (ABS) DataLab provides a safe environment for researchers to access a variety of household and business microdata[2]. The ABS DataLab is a data analysis solution for high-end data users who want to extract full value from ABS microdata (Parker, 2017). The DataLab supports researchers to undertake complex research. The number of DataLab sessions has increased substantially since 2019–20, with 15,520 sessions accessed in 2020-21 and 24,037 sessions accessed in 2021-22. Currently, the ABS output clearance is a manual checking process which is not scalable, cost effective or free from human error. There is also a risk that the increasing number of outputs from different projects could potentially introduce differencing risks even though these outputs have individually met the strict output criteria.

This research builds on previous ABS research[3] and applies the concept of five safes framework proposed by Desai et al. (2016) to provide output checking tools for the approved ABS DataLab researchers. These approved researchers have signed an undertaking on how microdata will be used and completed a training course to understand their responsibilities to use microdata effectively (Parker, 2017). Under the context of the trusted access to confidential public sector microdata in the ABS DataLab, these researchers belong to a researcher community of safe users who work to enable greater value from public sector microdata . Green et al. (2017) found that researchers are more likely to act appropriately when they are regarded as trustworthy, and they respond positively to being part of a community. The ABS will work with trusted researchers to make the output checking process scalable by providing these output checking tools to facilitate and automate the process.

The paper is organised as followed: Section 2 describes the ABS DataLab, Section 3 provides a summary of the prototype tools, Section 4 discusses considerations for developing prototype tools and Section 5 provides a conclusion and proposes future research directions.

---

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

[2] Other official statistical offices also provide similar service via remote or on-site access. For example, Statistics Canada Research Data Centres, Stats New Zealand Data Lab, Office for National Statistics Secure Research Service and US Federal Statistical Research Data Centers.

[3] See Thompson et al. (2013), Chipperfield and O'Keefe (2014) and O'Keefe et al. (2017)

2. ABS DataLab and output checking

The ABS uses the five safes framework proposed by Desai et al. (2016) while providing researchers access to the ABS DataLab. The ABS DataLab allows researchers to virtually analyse unit record information ranging from basic to detailed and integrated microdata files in the secure ABS environment. The environment has recent versions of analytical software, including R, SAS, Stata and Python. All analytical outputs that researchers want to use outside DataLab are checked by the ABS before release (see Parker (2017) for a detailed description of the process). Output checking plays an important role in minimising statistical disclosure risks. Output checking is commonly used by other official statistical offices (see Stats NZ (2020) and ONS (2020)). To ensure privacy and confidentiality rules are followed, ABS DataLab outputs follow strict procedures to minimise disclosure risks before releasing outputs.

3. ABS prototype output checking tools

The ABS has prototyped tabular output protection tools for count data, continuous data and regression outputs that consistently applies the ABS cellkey methodology to aggregate outputs, thereby minimising differencing risks and reducing the costs associated with output checking. The current prototype tools are built in R[4]. We selected a few variables from publicly available microdata[5] from the ***synthpop*** library (see Nowok et al. (2016)) and Jupyter notebook to showcase these R tools. Figure 2-1 shows an example of how researchers can use a table function (***Conf_Table***) to produce a table of counts and income by education attainment. Researchers can choose to display protected and original aggregate tables.

Figure 2-1
Confidentialised tables

```
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

The left panel in Figure 2-2 shows how researchers can produce protected and original regression outputs[6]. Researchers can also produce hexbin plot[7] to show the relationship between protected predicted values and residuals for model diagnostics. A hexbin plot allows users to see patterns more easily while still providing a high level of protection against disclosure.

Figure 2-2
Prototype output checking tools for regression outputs

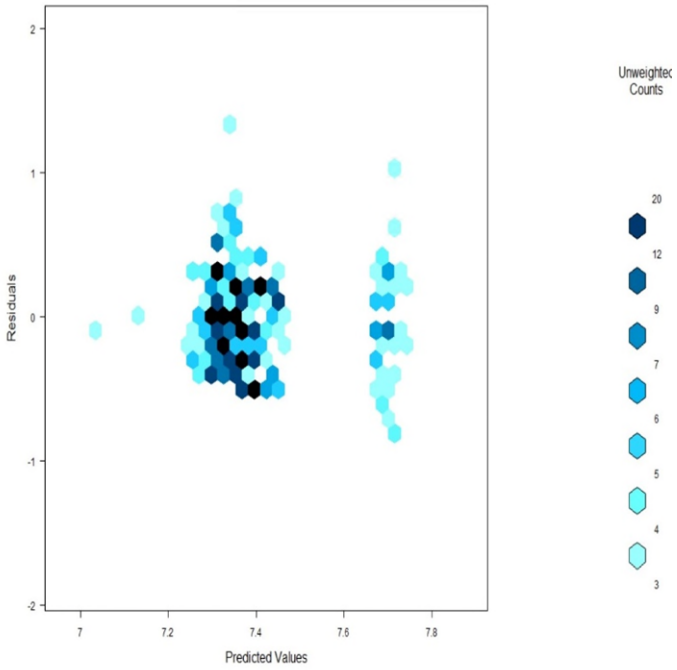
[4] There is a scope to build a similar version in Python. The ABS Protari is built on Python (see Protari).
[5] We use income, education attainment, age and social status from the Social Diagnosis 2011 - Objective and Subjective Quality of Life in Poland.
[6] O'Keefe et al. (2017) provide a detailed discussion on the linear robust method implemented for the regression output checking tools.
[7] A hexbin plot divides the area in a graph into tessellating hexagons, then shades each hexagon depending on the number of observations that occur in that hexagon.

regression function (*glm_f*)

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

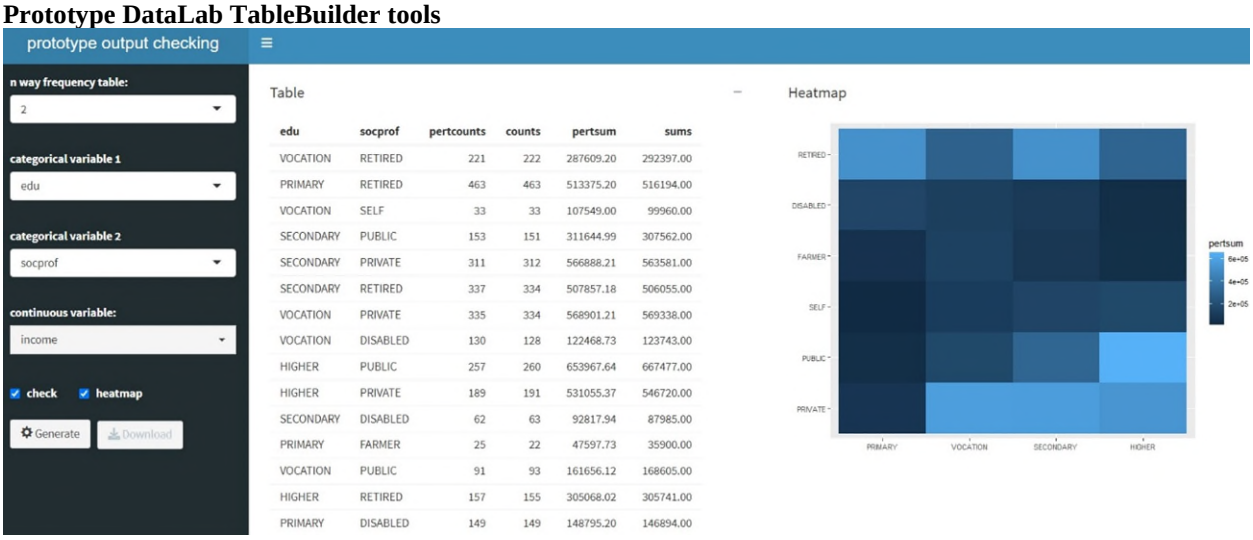
	Dependent variable:	
	income_log	
	Linear	Conf. linear (Robust)
	(1)	(2)
	glm	glm
age	0.002***	0.002**
	(0.001)	(0.001)
socprofPUBLIC	0.154***	0.179***
	(0.028)	(0.025)
socprofSELF	0.372***	0.381***
	(0.048)	(0.044)
socprofFARMER	-0.466***	-0.414***
	(0.041)	(0.037)
socprofDISABLED	-0.630***	-0.616***
	(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

Hexbin plot



We recognise researchers in the ABS DataLab have different comfort levels in using programming language. Therefore, we developed a basic user interface which interacts with the R functions. (See Figure 2-3). Users can simply pass their analysis data frame to the tools and produce safe outputs to facilitate the output checking process. Researchers can produce a heatmap for their analysis. The output checking tools have in-built processes to ensure compliance. For example, researchers cannot download csv files containing original table outputs.

Figure 2-3
Prototype DataLab TableBuilder tools



Currently, selected researchers in DataLab are trialling a version of the prototype for counts data, and we are gathering feedback to improve its usability and expand on its methods.

4. Development considerations

There is a list of outstanding questions to consider for further development of the output checking tools, including:

- Do output tools meet researchers' needs?
- Can the tools cater for different user capabilities?
- How can we minimise errors in using the tools?
- Should we buy off-the-shelf or develop in-house?
- How can we build tools that are user friendly and easily maintained?

There have been significant developments in the open-source community[8] and national statistics offices in the visualisation and statistical tools for outputs. There are collaborative opportunities to build common and enduring tool sets to ensure safe outputs as research data centres become a more popular dissemination tool.

5. Conclusion and future research directions

The ABS DataLab is the analysis solution for researchers who want to undertake real time complex analysis of detailed microdata. The ABS DataLab has become an important dissemination tool to make greater use of public sector microdata for research. There is significant growth in the demand for the ABS DataLab. The current manual output checking process for all DataLab outputs is not scalable or sustainable.

There are significant opportunities to be gained from automating the manual output checking process so that it can be more targeted and focused on complex processes that can not be automated. Both the ABS and other statistical offices are interested in developing tools to make the process more efficient to minimise any errors.

There are on-going discussions on resourcing and processes to develop the tools further and make them more widely available in the ABS DataLab. There are opportunities to seek collaborative effort from both official statistics and open-source communities to develop enduring assets.

[8] There are recent developments for advanced visualisation tools in the open-source community (see Martoglio (2018) and Kruchten (2022)).

References

CHIPPERFIELD, J. O., and C. M. O'KEEFE (2014), "Disclosure-protected Inference Using Generalised Linear Models", *International Statistical Review*, 82**,** pp. 371-391.

DESAI, T., RITCHIE, F., and R. WELPTON (2016). "*Five Safes: Designing data access for research*".

GREEN, E., RITCHIE, F., NEWMAN, J., and T. PARKER (2017), "Lessons learned in training 'safe users' of confidential data", Worksession on Statistical Data Confidentiality.

KRUCHTEN, N. (2022), "PivotTable.js - an open-source Javascript Pivot Table", Online Available: https://pivottable.js.org/examples/ [Accessed].

MARTOGLIO, E. (2018), "rpivotTable: Build Powerful Pivot Tables and Dynamically Slice & Dice your Data", Online Available: https://CRAN.R-project.org/package=rpivotTable [Accessed].

NOWOK, B., RAAB, G. M. and C. DIBBEN (2016), "synthpop: Bespoke creation of synthetic data in R", *Journal of statistical software*, 74**,** pp. 1-26.

ONS (2020), "SRS Researcher Output Clearance Guidance" in August 2020 by Stats NZ Tatauranga Aotearoa Wellington, New Zealand.

PARKER, T. (2017), "The DataLab of the Australian bureau of statistics", *Australian Economic Review*, 50**,** pp. 478-483.

STATS NZ (2020), "Microdata output guide", in August 2020 by Stats NZ Tatauranga Aotearoa Wellington, New Zealand.

THOMPSON, G., BROADFOOT, S. & ELAZAR, D. (2013), "Methodology for the automatic confidentialisation of statistical outputs from remote servers at the Australian Bureau of Statistics", Joint UNECE/Eurostat work session on statistical data confidentiality, Ottawa, Canada.