The Cell Key Method in τ-ARGUS

Keywords: *Tabular data, Statistical Disclosure Control, Cell Key Method, software, τ-ARGUS*

1 Introduction

The upcoming 2021 census has triggered the discussion again on which disclosure control technique would be favourable when publishing tables based upon the census data. It is a tradition that each member state of the EU in principle decides on the technique to protect their own census tables. From a European perspective however, a more harmonised approach would be favourable. That way it would be much easier to combine and correctly compare census tables from different member states.
Starting September 2016, a one year project partly funded by Eurostat[1] took place with the aim to identify methods for a 'Harmonised protection of census data in the ESS'. This project suggested to use one or both of two methods: Targeted Record Swapping (TRS) and noise addition using the Cell Key Method (CKM). In order to facilitate easy application of these methods a follow-up project was started (also partly funded by Eurostat[2]) to implement TRS and CKM into Open Source tools. As part of this project, not only a CKM method for census tables but a more general CKM method was implemented as well.
The implementation is available as an R-package (the cellKey package) as well as part of τ-ARGUS, a general purpose SDC tool for tabular data protection. Both tools can be found at the github environment sdcTools (https://sdctools.github.io/UserSupport). The sources can be found on https://github.com/sdcTools.
In this abstract we will briefly describe the method and give some information on the way it is implemented in τ-ARGUS. During a presentation at the NTTS and in a possible full paper we will further describe the way to use τ-ARGUS for application of the Cell Key Method using the idea of a 'walk through'.

2 Noise addition using the Cell Key Method

The Cell Key Method as implemented in τ-ARGUS is based on the ideas described in Fraser and Wooton [1] and elaborated on in Thompson et al. [2]. For the application of CKM to magnitude tables, we took up suggestions from Ma et al. [3]. More detailed information on the τ-ARGUS implementation can be found in e.g., Meindl et al. [4] and de Wolf [5].
Even though CKM is considered a post-tabular perturbative disclosure control method, it starts with the underlying microdata. CKM is designed to assure that whenever some table cell appears in another table based on the same microdata, it will get the exact same noise assigned. I.e., CKM assures consistency between tables. This is achieved by assigning some random value to each record in the microdata and using these values to define the stochastic component that is applied to a table cell. Whenever the same microdata with the same realisation of the random variable is used, a

[1] The SGA *Harmonised protection of census data in the ESS*, contract № 11111.2016.005-2016.367 under FPA № 11112.2014.005-2014.533

[2] The SGA *Open Source tools for perturbative confidentiality methods*, contract № 2018.0108 under FPA № 11112.2014.005-2014.533

table cell that consists of the same contributions will get the same amount of noise assigned.
We will call the random values for each record the recordkey. When calculating a table cell these record keys will be combined into a cellkey. The main difference in the way we implemented CKM into τ-ARGUS compared to the implementation in the Australian TableBuilder system is in the way we represent the record keys. In our approach the record keys are realisations from a Uniform$(0, 1)$ distribution.

3 Preparation

For both frequency count tables and magnitude tables one needs to assign record keys to the underlying microdata. This needs to be done *outside* τ-ARGUS. The reason being that it is essential for CKM to work that the same record keys are used whenever a table is constructed that is based on the same microdata.
To be able to apply CKM one also needs a so called perturbation table (p-table for short) that represents the distribution of the stochastic term that is applied to a cell. Such a p-table can be obtained using the R-package ptable[3] or could be constructed by yourself. That R package computes perturbation tables using a maximum entropy approach as suggested in Marley and Leaver [6] and illustrated in Giessing [7]. The resulting p-tables can be saved to file in a format that can be used by τ-ARGUS. For a description of that format, see de Wolf [5].
For the construction of a p-table for frequency count tables, one can specify the maximum possible perturbation D (specifying that the added noise is between $-$D and $+$D), the noise variance V (specifying the spread of the possible noise values) and a threshold js (specifying that the probability to add noise such that the cell value is less than or equal to js is zero).
A p-table for magnitude tables differs slightly from one for frequency count tables. In case of magnitude tables, values of the p-table will be interpolated to get the correct perturbation value for any possible value of the continuous response variable. This means that, additional to the parameters that need to be specified for a p-table for frequency count tables, a vector icat of integers between 1 and D needs to be specified (specifying the base cell values between which will be interpolated) along with a step width step (specifying the step width of the noise values) and a type (specifying whether the p-table is to be used for all cells or for cells with odd or even number of contributors).
Note that a p-table for frequency count tables with D $= 4$, V $= 3$ and js $= 0$ could also be obtained by specifying a p-table for magnitude tables with D $= 4$, V $= 3$, icat $= c(1, 2, 3, 4)$, step $= 1$ and type $=$ "all".

4 Frequency count tables

In case of frequency count tables, the cell key is obtained by adding the record keys of all contributors to that cell modulo 1. E.g., 4 records with record keys 0.68194, 0.81020, 0.01729 and 0.49102 being the only contributors a particular cell would give rise to a cell key with value 0.00045. Using the cell key together with the cell value, the p-table is used to determine the amount of noise added to that cell. For a detailed description of the way this is performed, see e.g., Appendix 1 in [8].
If you want to apply CKM to a frequency count table with τ-ARGUS, you can only work with microdata as input. Additionally to the 'standard' τ-ARGUS information, the .rda metadata file associated with the microdata needs to specify which variable

[3] Also available at https://github.com/sdcTools

is the record key and where the file with the p-table is located. At a later stage it is possible to change the p-table to a different one, if needed. To calculate a frequency count table for application of CKM, you need to select <freq> when specifying the table in τ-ARGUS. Then, when the main window of τ-ARGUS displays the table, you can select the Cell Key Method in the Suppress section of the window. Doing so, a button will appear allowing you to select a different p-table file. Finally, click the Cell Key button to apply CKM. Figure 1 shows the output τ-ARGUS will display after CKM has been applied. The darker the shade of a cell, the larger the noise that is applied. Additional information loss measures can be obtained by pressing the Table summary button.

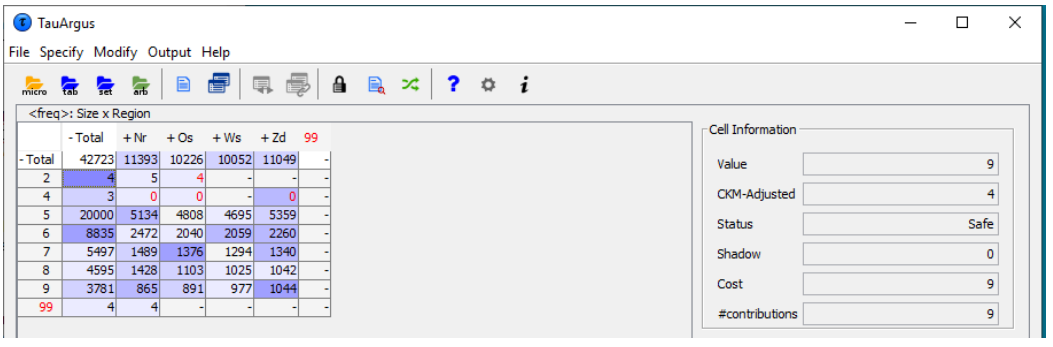

Figure 1: Example output of τ-ARGUS after applying CKM to a frequency count table.

5 Magnitude tables

The calculation of the cell key in case of magnitude tables is largely the same as for frequency count tables. The only difference being that there is an option to in- or exclude the record keys of records that contribute the *value* zero to a cell. See Giessing et al. [8] for more details on this option.
Again, the .rda metadata file should contain information that can be used when applying CKM to a magnitude table. This information needs to be specified for each numeric value in the microdata set to which you want to apply CKM. Note that later, in the GUI of τ-ARGUS, you also have the opportunity to change the metadata including the CKM-settings. See Figure 2 for an example to set CKM parameters for a variable. See Giessing et al. [8] for more details on the meaning of the different settings.
Several features for application of CKM to magnitude tables are available through τ-ARGUS, e.g., you can specify whether

- zero contributions should be included or excluded when calculating the cell keys;
- the noise should be scaled by using the largest K contributions, the cell mean, the distance between the maximum and the minimum contribution to a cell or the cell value itself;
- cells with an odd or even number of contributions should be perturbed differently;
- a so called 'flex-function' should be used for the scaling of the noise;
- small cells should be treated like counts.

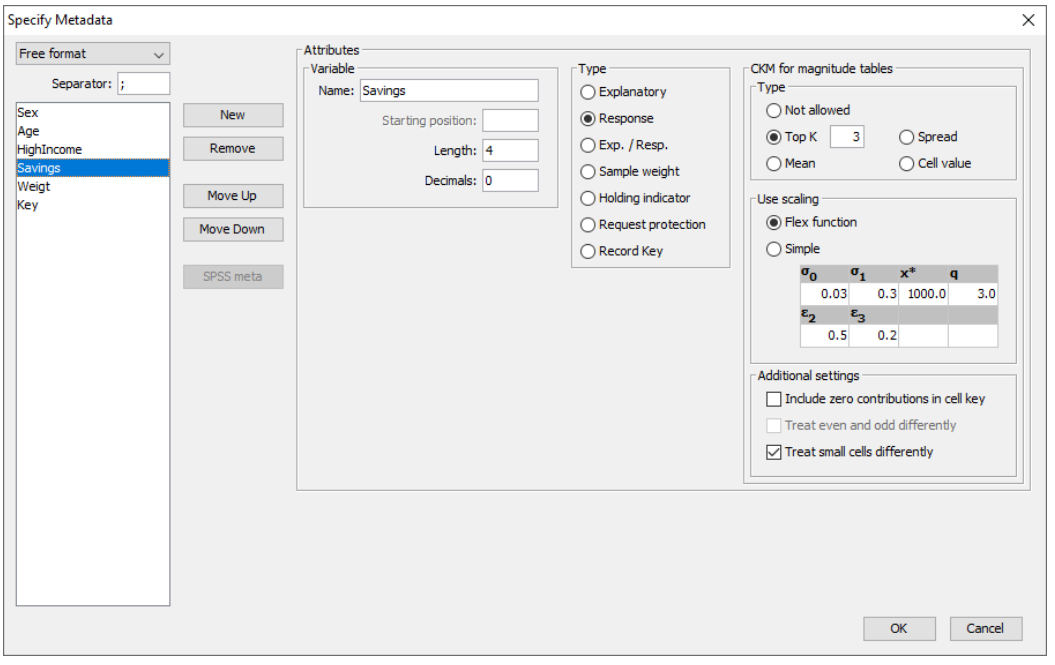

Figure 2: Example of the GUI to change CKM settings for a variable.

Moreover, we implemented that the cell key when used in the look-up phase of the procedure is changed in a deterministic way[4] when using the largest $K \geq 2$ contributions. We also allow for the specification of an additional amount of noise to be added to cells that are unsafe according to a pre-specified sensitivity rule (like the p%-rule). For magnitude tables there is currently no visual feedback of the amount of added noise using the shading of cells like in the case of frequency count tables. Also no additional information loss measures are available for magnitude tables.

6 Future work

Recently a new four year grant has started (STACE[5]) including a Centre of Excellence for SDC. Under this grant we will further elaborate on the maintenance and development of the SDC tools, including a further development of the CKM method for magnitude tables in τ-ARGUS. One of the first things that comes into mind is adding information loss measures for CKM protected magnitude tables.

References

[1] B. Fraser and J. Wooton. A proposed method for confidentialising tabular output to protect against differencing. Presented at Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality (Geneva, 9–11 November 2005), 2005. URL http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2005/wp.35.e.pdf.

[2] G. Thompson, S. Broadfoot, and D. Elazar. Methodology for the automatic confidentialisation of statistical outputs from remote servers at the australian bureau of statistics. Presented at Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality (Ottawa, 28–30 October 2013), 2013. URL http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2013/Topic_1_ABS.pdf.

[4] The first digit of a cell key is removed and added at the end of the number. E.g., a cell key 0.71926 would be changed into 0.19267.

[5] Partly funded by Eurostat, grant agreement № 899218, 2019-BG-Methodology

[3] Y. Ma, Y.X. Lin, J. Chipperfield, J. Newman, and V. Leaver. A new algorithm for protecting aggregate business microdata via a remote system. In J. Domingo-Ferrer and M. Pejić-Bach, editors, *Privacy in Statistical Databases, PSD 2016*, pages 210–221. Springer, 2016. DOI: 10.1007/978-3-319-45381-1_16. Lecture Notes in Computer Science, vol 9867.

[4] B. Meindl, A. Kowaric, and P.P. de Wolf. Prototype implementation of the cell key method, including test results and a description on the use for census data. In: Deliverable D3.1 of Work Package 3 'Prototypical implementation of the cell key/seed method' within the Specific Grant Agreement 'Open Source tools for perturbative confidentiality methods', 2018.

[5] P.P. de Wolf. Quick reference for CKM in τ-ARGUS 4.2.0, 2020. URL https://github.com/sdcTools/tauargus/releases/download/4.2.0b5/QuickReferenceCKM4.2.0.pdf.

[6] J.K. Marley and V.L. Leaver. A method for confidentialising user-defined tables: Statistical proper-ties and a risk-utility analysis. In *Proceedings of 58th World Statistical Congress, 1072–1081*, 2011.

[7] S. Giessing. Computational issues in the design of transition probabilities and disclosure risk estimation for additive noise. In J. Domingo-Ferrer and M. Pejić-Bach, editors, *Privacy in Statistical Databases, PSD 2016*, pages 237–251. Springer, 2016. DOI: 10.1007/978-3-319-45381-1_18. Lecture Notes in Computer Science, vol 9867.

[8] S. Giessing, R. van de Laar, T. Enderle, and R. Tent. Methodological report on options of generalising the cell key method and eventual restrictions. In: Deliverable D4.1 of Work Package 4 'More general implementation of the cell key/seed method' within the Specific Grant Agreement 'Open Source tools for perturbative confidentiality methods', 2019.