WP.29
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Manchester, United Kingdom, 17-19 December 2007)

Topic (iii): Applications (including practical implementation of SDC methods, actual issues within NSIs and software)

APPLYING TAU-ARGUS TO SUPERCROSS TABLES: A PRACTICAL EXAMPLE USING THE UK BUSINESS REGISTER UNIT DATA

Invited Paper

Prepared by Andrea Toniolo Staggemeier and Philip Lowthian (Office for National Statistics, United Kingdom) and Grant Lee (Space-Time Research Pty Ltd., Australia)

Applying Tau-Argus to SuperCROSS tables: A practical example using the UK Business Register Unit data

Andrea Toniolo Staggemeier[1], Philip Lowthian[2], and Grant Lee[3]

[1] Information Management (Strategies), Office for National Statistics, Newport, United Kingdom, andrea.staggemeier@ons.gsi.gov.uk

[2] Methodology Directorate (Statistical Disclosure Control), Office for National Statistics, London, United Kingdom, philip.lowthian@ons.gsi.gov.uk

[3] Space-Time Research Pty Ltd, Melbourne, Australia, grant.lee@str.com.au

Abstract. The Business Register Unit (BRU) at the Office for National Statistics (ONS) in the UK produces a wide range of tabular outputs, many of which are published on the ONS website. SuperCROSS is a tabulation tool used by the Office for National Statistics (ONS) and the idea of linking this program with the disclosure control package Tau-Argus has been developed over the last 3 years, culminating with the implementation of the SuperCROSS/Tau-Argus link live in production of BRU outputs from August 2007. A table created in SuperCROSS can be confidentialised by the user selecting the required rules from a drop down menu. This action opens Tau-Argus and uses a batch file to either round or suppress cells in the table. The safe table is then returned to SuperCROSS without the need for the user to interact with Tau-Argus.

This paper describes the interface between the two programs, typical rules that can be applied, how those rules are set, and the different output formats available. Also discussed in this paper are the reasons behind why a link between Tau-Argus and SuperWEB (thin client tool from SuperSTAR product suite) is not recommended as an alternative to SuperCROSS (thick client, i.e. desktop installation tool) within ONS. Finally benefits of this approach to both individual business areas and National Statistics Institutes (NSIs) in general are given.

Keywords. SuperCROSS, Tau-Argus, Statistical Disclosure Control, Controlled Rounding

1 Introduction

The idea to develop a link between a tabulation tool and a statistical disclosure tool was initiated by the Office for National Statistics (ONS) in 2004. One business area within ONS, the Business Register Unit (BRU), who have been long term users of SuperCROSS[1] for tabulation, had a business need for a more robust and efficient means of Statistical Disclosure Control (SDC) in order to protect their outputs. It was highly desirable to retain SuperCROSS for tabulation, and the tool of choice for SDC within ONS was Tau-Argus[2]. An interface has since been developed in SuperCROSS to allow tables to be passed seamlessly to Tau-Argus, and back to SuperCROSS once disclosure has been applied; all without any manual interaction from the user.

The process that BRU followed prior to this integration of tools was described as time consuming, labour intensive, and with some level of risk to publishing disclosive tables. As a consequence of this risk, less than optimal information was published in each table in order to minimise any event of disclosure.

1.1 Business Problem

This paper examines outputs produced by the BRU from the Inter-Departmental Business Register (IDBR[3]). This is the comprehensive list of UK businesses used for statistical purposes. IDBR also provides a sampling frame for surveys of businesses carried out by the ONS and by other government departments. It is therefore a key data source for analyses of business activity in the UK.

The IDBR covers businesses in all parts of the economy, other than some very small businesses (self-employed and those without employees and low turnover) and some non-profit organisations. With 2.1 million businesses listed, it provides nearly 99% coverage of UK economic activity. It holds a wide range of information on business units including: name; address; standard industrial classification; employment and employees; and turnover.

The amount of outputs that BRU produce require a well thought out process, including a procedure to apply the required disclosure control rules using Excel spread sheets. Historically, both rounding (for frequency tables) and suppression (for magnitude tables) have been carried out. Since this was not automated, the process was very time consuming and there was a potential risk that unsafe outputs might be produced.

[1] SuperCROSS is the client tabulation tool, and part of the SuperSTAR suite from Space-Time Research (http://www.spacetimeresearch.com)
[2] Tau-Argus (http://neon.vb.cbs.nl/casc)
[3] IDBR website (http://www.statistics.gov.uk/CCI/nugget.asp?ID=195)

The tabulation tool, SuperCROSS, already in use by BRU, was enhanced to provide an interface to support the disclosure control process in Tau-Argus. The following sections will briefly discuss Tau-Argus and SuperCROSS, the approach to integrating the two technologies, and a more detailed description on how the link and the application of safety rules were implemented.

1.2 Tau-Argus

Tau-Argus is a software tool which enables statistical disclosure control to be carried out to protect tabular output. It can be run in either interactive or batch mode and can import tables or microdata, allowing the user to create tables. Tau-Argus supports either frequency or magnitude data types and once imported along with a metadata file, the user can apply a number of confidentiality rules.

Typically for magnitude tables, safety rules such as threshold and dominance rules are set by the user, and cells failing these rules are highlighted, allowing the user to select them for suppression. In order to avoid disclosure by differencing, secondary suppression can be applied using a variety of techniques. For frequency tables, controlled rounding is commonly applied. This method rounds cell values to the nearest multiple of a user specified base, whilst maintaining the table additivity.

Tau-Argus was initiated as result of the CASC (Computational Aspects of Statistical Confidentiality)[4] project, which was European Union (EU) funded with additional support from many statistical organisations, including ONS, and EU Universities. Tau-Argus is currently being used by BRU and by a number of other Government Departments and Agencies who supply data for the ONS Neighbourhood Statistics website

1.3 SuperCROSS

SuperCROSS is part of the SuperSTAR Suite of products, developed by Space-Time Research. As a desktop tabulation tool, it allows a user to create tables via a drag and drop interface. It has features which include the ability for the user to derive new variables, add statistical calculations to tables, transform data dynamically, and to drill down on selected cells in a table to view the contributing unit records. These features give the user great flexibility for creating tables without the need for specialist skills such as SQL. SuperCROSS has been used in the ONS for many years, and is the current tabulation tool for BRU, Labour Force Survey, and Census.

[4] http://neon.vb.cbs.nl/casc

1.4 Proposed Solution

The proposed solution was based on the use of Tau-Argus in batch mode, integrated with the familiar SuperCROSS tabulation environment. SuperCROSS users can choose a pre-defined rule to be applied to a table, and Tau-Argus is invoked automatically. Once processed, the results would be passed back to the SuperCROSS environment. The key to this approach is that only a small number of SuperCROSS users would need to have some level of understanding of SDC. The SDC rules to be applied are stored in template files, which are specified and quality assured by a statistical disclosure control specialist.

Figure 1 shows the high level workflow of the Tau-Argus and SuperCROSS integration.

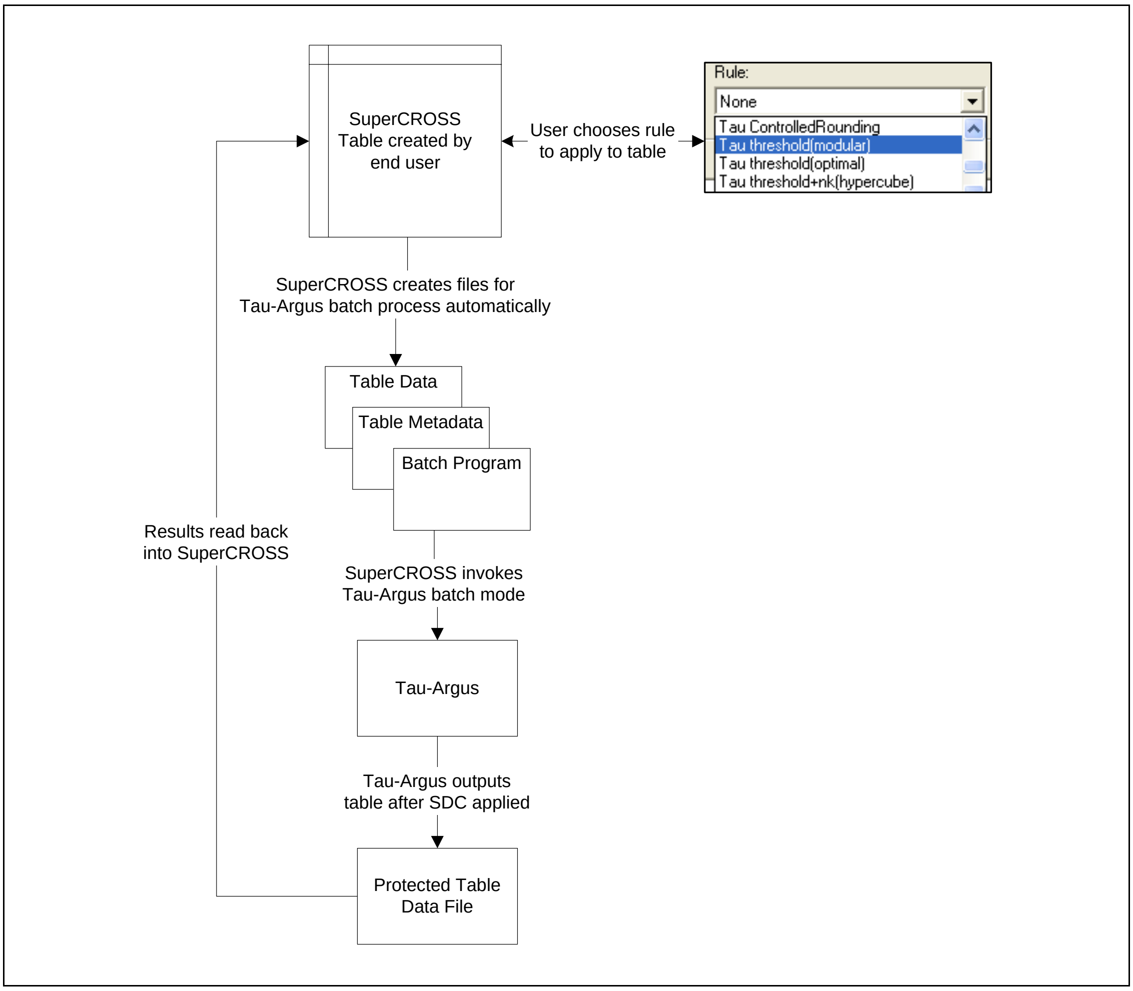

SuperCROSS Table created by end user
User chooses rule to apply to table
Rule:
None
Tau ControlledRounding
Tau threshold(modular)
Tau threshold(optimal)
Tau threshold+nk(hypercube)
SuperCROSS creates files for Tau-Argus batch process automatically
Table Data
Table Metadata
Batch Program
Results read back into SuperCROSS
SuperCROSS invokes Tau-Argus batch mode
Tau-Argus
Tau-Argus outputs table after SDC applied
Protected Table Data File

Figure 1. Tau-Argus and SuperCROSS workflow

2 Detail of Solution

2.1 How does Tau-Argus work in Batch Mode

Tau-Argus can be used through its own Graphical User Interface (GUI), or through a batch mode. Batches are composed of a set of instructions which allows Tau-Argus to identify the data source, metadata, disclosure rules to apply, and any transformation to the data that is required.

Through the batch file it is also possible to set parameters to create output files after disclosure rules are applied, as well as an HTML report file with summary information relating to the impact of the SDC method on the outputs produced. Figure 2 shows an example of a batch file created for Tau-Argus. It specifies the location of the table data and metadata, the specification of the table, the disclosure rules to apply, and the location of the output.

```
<OPENTABLEDATA>  "D:\Tau-Argus\temp_tauinput.tab"
<OPENMETADATA>   "D:\Tau-Argus\temp_taumetadata.rda"
<SPECIFYTABLE>   "var1""var2""var3"|"resp_var1"|"resp_var1"|"resp_var1"
<SAFETYRULE>     FREQ(3,30)|
<READTABLE>
<SUPPRESS>       RND(1,5,0,0,20,0,0,2)
<WRITETABLE>     (1,1,1,"D:\Tau-Argus\temp.csv")
```

Figure 2. Example of Tau-Argus batch file

For further information on each of the elements in the batch file, refer to the Tau-Argus manual available on the CASC website.

2.2 Integration of Tau-Argus with SuperCROSS

Figure 1 shows the workflow of how a user interacts with SuperCROSS.

a. Firstly, a rule template is defined by a SDC specialist. The user need not understand the rules in detail, nor how they are specified.

b. Next, the user creates a table in SuperCROSS, and chooses which rule they want to apply from a drop down menu.

c. SuperCROSS tabulates the table, and then generates the batch file, and table data and metadata files required for Tau-Argus. SuperCROSS then invokes Tau-Argus in batch mode, and the table is processed.

d. Once the Tau-Argus batch is finished, the results are read back into SuperCROSS and displayed to the user.

Figure 3 shows how a rule template file is defined for SuperCROSS. It contains the detail of the rule to apply when chosen by the user.

```
<SAFETYRULE>     FREQ(3,15)|
<READTABLE> 1
<SUPPRESS>       RND(1,10,0,0,20,0,0,3)
<WRITETABLE> (1,3,1,"C:\TEMP\report_tauoutput.csv")
```

Figure 3. Example Rule Template file for Controlled Rounding

This file is then referred to in a SuperCROSS configuration file (confid.ini), along with the location of the Tau-Argus executable. An example of the SuperCROSS configuration file is shown in Figure 4.

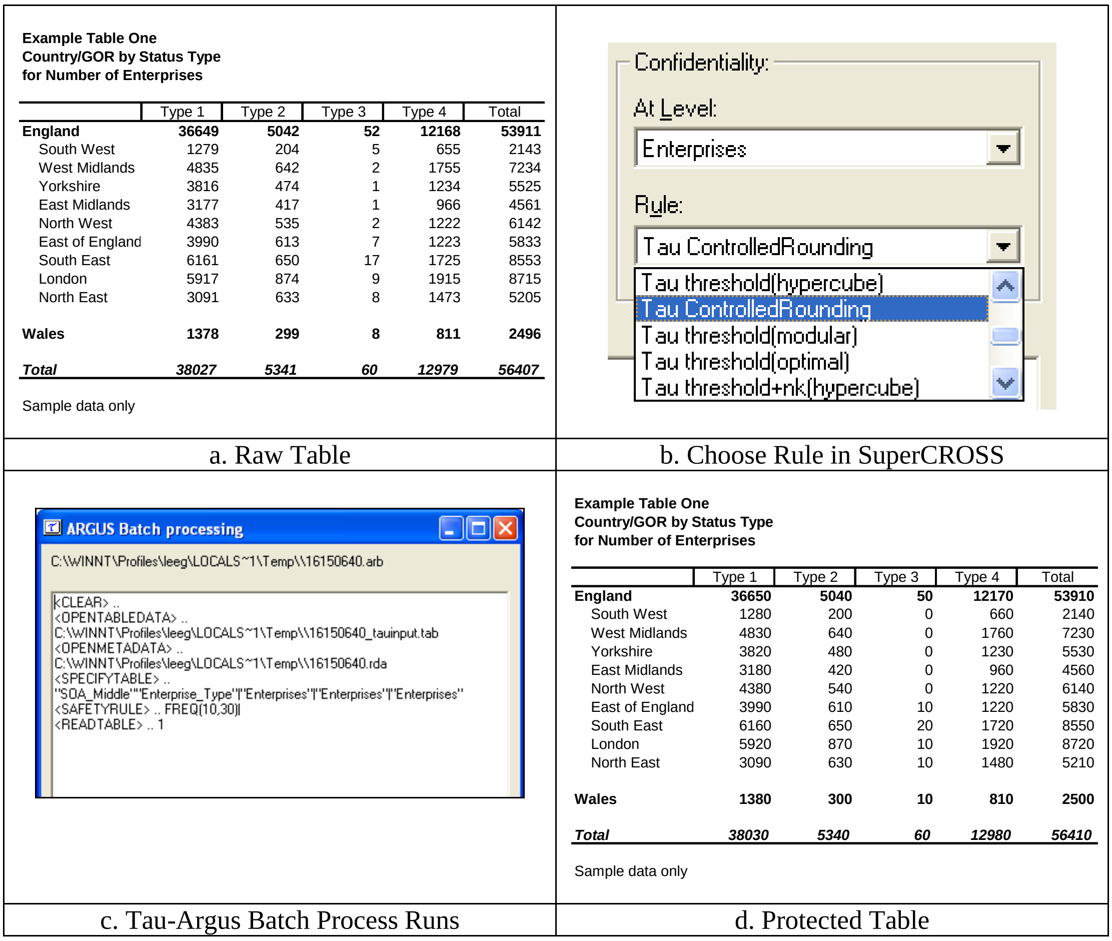
```
[Tau-Argus]
Tau-Argus Location=\\TauServer\TauArgus
Tau-Argus Exe=TauArgus.exe
Tau-Argus Log=C:\Program Files\TauArgus\sxtaulog.txt

[Rules]
Tau ControlledRounding=Tau:ControlledRounding

[ControlledRounding]
ARBFile=\\TauServer\RuleTemplates\TauControlledRounding.arb
TopN=0
```

Figure 4. SuperCROSS Configuration File

When the user tabulates a table in SuperCROSS with a chosen rule, the table data and metadata is saved to appropriately formatted text files, and the rule template file is copied and the location of the saved data is inserted. All of these operations are hidden from the SuperCROSS user and occur automatically.

2.3 Controlled Rounding Example

This section gives an example, not using standard ONS rules, of applying controlled rounding to a table within SuperCROSS. Figure 5 shows the steps required to apply the rule to a table.



Example Table One
Country/GOR by Status Type
for Number of Enterprises

	Type 1	Type 2	Type 3	Type 4	Total
England	**36649**	**5042**	**52**	**12168**	**53911**
South West	1279	204	5	655	2143
West Midlands	4835	642	2	1755	7234
Yorkshire	3816	474	1	1234	5525
East Midlands	3177	417	1	966	4561
North West	4383	535	2	1222	6142
East of England	3990	613	7	1223	5833
South East	6161	650	17	1725	8553
London	5917	874	9	1915	8715
North East	3091	633	8	1473	5205
Wales	**1378**	**299**	**8**	**811**	**2496**
Total	***38027***	***5341***	***60***	***12979***	***56407***

Sample data only

a. Raw Table

Confidentiality:

At Level: Enterprises

Rule: Tau ControlledRounding

b. Choose Rule in SuperCROSS

ARGUS Batch processing

```
C:\WINNT\Profiles\leeg\LOCALS~1\Temp\\16150640.arb

<CLEAR> ..
<OPENTABLEDATA> ..
C:\WINNT\Profiles\leeg\LOCALS~1\Temp\\16150640_tauinput.tab
<OPENMETADATA> ..
C:\WINNT\Profiles\leeg\LOCALS~1\Temp\\16150640.rda
<SPECIFYTABLE> ..
"SOA_Middle""Enterprise_Type"|"Enterprises"|"Enterprises"|"Enterprises"
<SAFETYRULE> .. FREQ(10,30)|
<READTABLE> .. 1
```

c. Tau-Argus Batch Process Runs

Example Table One
Country/GOR by Status Type
for Number of Enterprises

	Type 1	Type 2	Type 3	Type 4	Total
England	**36650**	**5040**	**50**	**12170**	**53910**
South West	1280	200	0	660	2140
West Midlands	4830	640	0	1760	7230
Yorkshire	3820	480	0	1230	5530
East Midlands	3180	420	0	960	4560
North West	4380	540	0	1220	6140
East of England	3990	610	10	1220	5830
South East	6160	650	20	1720	8550
London	5920	870	10	1920	8720
North East	3090	630	10	1480	5210
Wales	**1380**	**300**	**10**	**810**	**2500**
Total	***38030***	***5340***	***60***	***12980***	***56410***

Sample data only

d. Protected Table

Figure 5. Example of Controlled Rounding

a. The table is created by the user within SuperCROSS.

b. To apply disclosure control, the user can choose a pre-defined rule from a drop down menu in SuperCROSS. In this case, the rule is called “Tau ControlledRounding”. The user does not need to know, or understand, the specifics of how the actual rule works, although basic knowledge of the method is required. For reference, the rule used in this example was as follows:

```
<SAFETYRULE>     FREQ(10,30)|
<READTABLE>      1
<SUPPRESS>       RND(1,10,0,0,20,0,0,3)
```

In this example, a threshold rule is initially applied to the table. All cells with fewer than 10 contributors are defined as disclosive. After the table is checked for additivity and any undefined hierarchical levels added, controlled rounding is applied to base 10.

c. After the SuperCROSS tabulation process finishes, the table data is passed to Tau-Argus along with the batch file and other required information. Tau-Argus is invoked and the disclosure control rule selected is then applied.

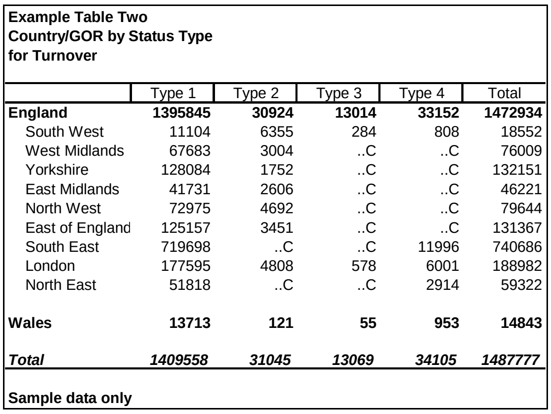
d. The protected table is now displayed in the SuperCROSS interface. The table can then be saved to the desired format for dissemination.

2.4 Example of Suppression

The original table in Figure 5 shows the count of the number of enterprises. If you change the definition of the table to sum turnover instead, it may be more appropriate to apply a different rule to the table. In this example, a suppression rule, not the ONS standard, is applied, which hides table cells rather than rounding them. In this case, the rule applied is as follows:

```
<SAFETYRULE>     P(15,100,1)| FREQ(3,30)
<READTABLE>
<SUPPRESS>       MOD(1)
```

This is a multi stage process. First, a threshold rule of 3 at a safety range of 30% is applied. In addition to this, the p% rule states that if the sum of all the contributors to a cell, excluding the top 2, is greater than 15% of the total value of the cell then the cell is not disclosive. This is applied to all cells in the table, with those failing either rule being suppressed. Finally, secondary suppression is applied using the modular method. Figure 6 shows the table results as the user would see in SuperCROSS, after the disclosure rules have been applied.



Example Table Two
Country/GOR by Status Type
for Turnover

	Type 1	Type 2	Type 3	Type 4	Total
England	**1395845**	**30924**	**13014**	**33152**	**1472934**
South West	11104	6355	284	808	18552
West Midlands	67683	3004	..C	..C	76009
Yorkshire	128084	1752	..C	..C	132151
East Midlands	41731	2606	..C	..C	46221
North West	72975	4692	..C	..C	79644
East of England	125157	3451	..C	..C	131367
South East	719698	..C	..C	11996	740686
London	177595	4808	578	6001	188982
North East	51818	..C	..C	2914	59322
Wales	**13713**	**121**	**55**	**953**	**14843**
Total	***1409558***	***31045***	***13069***	***34105***	***1487777***

Sample data only

Figure 6. Example of suppression rule

3 Business Benefits

The integration of the two tools, SuperCROSS and Tau-Argus, has advantages for both the producer of tabular statistics, and the user of the outputs. Tau-Argus has been shown to be a powerful disclosure control tool, enabling both suppression and controlled rounding to be carried out. The majority of producers of tables in a business area will not require any knowledge of Statistical Disclosure Control principles beyond the basics, as the rules required are pre-configured and quality assured by an SDC specialist. This means that a powerful disclosure control tool can be made available to users with minimum disruption to the office.

In general terms, any process which joins together important operations in the production of tables should be beneficial. A smoother output delivery process can be put in place and the users do not have to switch between different tools.

As the disclosure control procedure is now fully automated, this will result in savings in terms of time, particularly as users do not have to manually process data via Excel. There should also be quality improvements in the outputs (the tables have greater utility), with less likelihood of errors. This will assist in maintaining a positive reputation of the data supplier. Moreover, the business area could process more ad-hoc requests in the same amount of time.

The current implementation could easily be adapted for other business areas, with only minimal changes required to the rule templates for SuperCROSS, depending on the nature of the data.

4 Issues

The benefits to the business in terms of robustness and reduction in risk are obvious. However, the solution is not without its limitations, and these are predominately caused by the underlying methodology. There are also some disadvantages that can be found when comparing perceived and actual complexity of the tools.

The problem of the user creating the best design of a table that is statistically meaningful to data consumers, and to both of the tools involved, represents a major challenge to be overcome. SuperCROSS users have the flexibility to create tables in whichever fashion desired. All of these tables, prior to disclosure control, are valid. However, the tables do not necessarily have meaning when it comes to statistical disclosure control.

SuperCROSS users are able to define tables which span across multiple statistical unities (for example households and persons). This is an example of linking two or more tables from the same microdata, and there is currently no rule in Tau-Argus which solves this problem. Linked tables at the macrodata level are also not currently supported; tables which share one or more common variables cannot be protected simultaneously.

It is possible in SuperCROSS to create tables based on hierarchical variables, for example geography or industrial classification. SuperCROSS allows users to include any combination of values from any of the levels in the hierarchy, and it does not have to be complete. Although in certain instances Tau-Argus can calculate the missing hierarchical values, this must be used with great caution and it is recommended to supply a complete hierarchical structure.

The size of a table created in SuperCROSS can sometimes pose a problem for Tau-Argus in terms of the time required to solve the problem. It is often possible to overcome this by fine-tuning the rule applied asking Tau-Argus to find a feasible rather than an optimal solution, and then using table partitioning rules where available.

Many of the other features in SuperCROSS, such as grouping items within recodes, which are very useful in general for tabulation and table design purposes, also cause issues for the current methodology of Tau-Argus. Workarounds have been provided to BRU which do result in the desired table output format. Future development could overcome some of these limitations.

SuperWEB, also from Space-Time Research, is a browser based interface for self-service tabulation, accessed by the data consumer. This differs from SuperCROSS, which is typically used by the business area to generate pre-defined tables for publication. Given the nature of the SDC rules in Tau-Argus, it would not be recommended to apply this solution to a web-based, dynamic interface, either for data suppliers, or users of that data. Users of SuperWEB could not be expected to understand, nor appreciate, the limitations involved with attempting to apply SDC on demand through the web.

Aside from the methodological issues, any implementation of the SuperCROSS and Tau-Argus solution should not overlook the resources and effort required for installation and configuration of the tools. Tau-Argus requires mathematical solver software to be installed for some of the SDC rules, which adds a further layer of administration.

5 Conclusion

The integration of SuperCROSS and Tau-Argus will be beneficial in many ways to the ONS. The main benefit is that a business area can publish more, with better quality (i.e. data utility), and faster response times.

It is strongly advised that with any new implementation of this solution, the organisation invests in understanding the technology of the tools involved, and the administration that is required for all components. This should not be underestimated as Tau-Argus, SuperCROSS, and the mathematical solver of choice are from three different vendors, and all have their own method of dealing with issues.

Good user knowledge of SuperCROSS is required, but with little extra to learn. Once both products are installed the statistical disclosure business processes should flow more smoothly than previously experienced.

6 Acknowledgements

This paper was written with the cooperation and support of the Business Register Unit at the ONS.