WP.35
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)

IMPROVING OUR KNOWLEDGE OF METAHEURISTIC APPROACHES FOR CELL SUPPRESSION PROBLEM

Supporting Paper

Prepared by Andrea Toniolo Staggemeier[1], Alistair R. Clark[2], James Smith[3], and Jonathan Thompson [4]

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

[2] Principal Lecture in Operational Research at Computing, Engineering and Mathematical Sciences School, University of the West of England – Bristol, United Kingdom

[3] Reader at Computing, Engineering and Mathematical Sciences School, University of the West of England – Bristol, United Kingdom

[4] Principal Lecture in Operational Research at Maths Institute, University of Cardiff – Wales, United Kingdom

Improving our knowledge of metaheuristic approaches for cell suppression problem

Andrea Toniolo Staggemeier[1], Alistair R. Clark[2], James Smith[3], and Jonathan Thompson [4]

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

[2] Principal Lecture in Operational Research at Computing, Engineering and Mathematical Sciences School, University of the West of England – Bristol, United Kingdom

[3] Reader at Computing, Engineering and Mathematical Sciences School, University of the West of England – Bristol, United Kingdom

[4] Principal Lecture in Operational Research at Maths Institute, University of Cardiff – Wales, United Kingdom

Abstract. This work will discuss further tests carried out using a pool of operational research and artificial intelligence techniques to solve the cell suppression problem. Existing solutions to the problem available through Tau-Argus software are mathematically demanding and only enable a solution to the problem for small table sizes. The approaches investigated here are pseudo-optimum in quality but enable handling of large size tables with complex structure. The test bed for this work used artificially created data which represent real-world scenarios found at ONS, and a sample of real data created from IDBR[1] sources. These data type are summarised as magnitude data, in both hierarchical and non-hierarchical formats, including 2 sets of sensitivity (2% and 10% sensitivity) and 2 sets of sparsity measures (5% and 25% of the table's cells contain zero values).

Among the approaches discussed in this paper are: a hybrid Evolutionary Algorithm,;Ant Colony Optimization; and Greedy Randomising Adaptive Search Procedure (GRASP). The table safety criterion was met using the Attacker Model by Salazar and Fischetti (2001). A relaxed feasibility criterion was also used on the Ant Colony and GRASP approaches in order to try to accelerate the evaluation process. Initial results showed that all approaches are able to handle larger data sets than existing mathematical programming routines. However a trade-off analysis between time taken to solve and data size indicated that we still have to improve the total time, perhaps not by using a single cell pass for table safety evaluation but multiple cells at a time.

Keywords: Tau-Argus, Statistical Disclosure Control, Cell Suppression, Mathematical Programming, Evolutionary Algorithms, Ant Colony Optimization, Greedy Randomising Adaptive Search Procedure (GRASP)

[1] Inter-Departmental Business Register

1 Introduction

Cell Suppression is just one of many Statistical Disclosure Control methodologies for protecting individual respondents when publishing tabular statistical outputs. This particular methodology lends itself better to magnitude data, but because of the perception some statisticians have when their outputs are modified in any form, e.g. health data experts, there was a desire to apply the methodology to frequency data also . The problem is to find an ideally optimal pattern of suppressed cells for which key sensitive cells remain non-disclosive. This is done such that the objective for good solutions is measured in terms of the total information loss caused by the required additional suppressed cells.

This paper will discuss some of the problems of the existing suppression implementations available in Tau-Argus, and describe some initial results found when using alternative solutions using traditional mathematical programming techniques.

1.1 Business Problem

There are currently four suppression methodologies available within Tau-Argus: Network flow; Hypercube; Modular; and Optimal. For further details on all the relevant methodologies available in Tau-Argus, refer to the manual[2]. The ONS has a variety of data that needs to be confidentialised, and since the existing methodologies were not suitable to all magnitude and frequency data types, the Neighbourhood Statistics programme, in collaboration with Methodology Directory and Information Management Group, started an investigation to understand the potential of alternative approaches from different areas of Operational Research (OR) and Artificial Intelligence (AI).

This paper will present a short description of each alternative solution implemented, and highlight the findings of initial investigations. It will also discuss some of the issues that we consider outstanding from a modelling and algorithmic perspective, and a discussion on how some of these problems might be addressed.

1.2 Comparison of Existing Methods in Tau-Argus

A direct comparison between the existing Tau-Argus cell suppression approaches is not possible, due the inconsistent way each of the solutions incorporates protection levels and objective function definitions. This is certainly an excellent point of improvement for Tau-Argus contributor's community. However, this would only be of interest if the aim of the tool was to provide a framework for researching different approaches for the same problem in an easy way. At present, the plug-and-play architecture of Tau-Argus is reflected in the way each of the Tau-Argus contributors have their own preference on programming language and style, mathematical

[2] http://neon.vb.cbs.nl/casc/Software/TauManualV3.2.pdf

programming technology, and levels of documentation available. This in turn results in difficulties determining when is best to use one approach over another.

Other work from ONS, proposing quality measures for confidentialised tables, also did not take into account the fact that different solutions have slight different way in approaching the definition of cell safety, and the quality of the suppression pattern used. For example, some methodologies report on the number of suppressed cells, and others on the total of weighted sum for a suppression pattern. The ONS work , however, assesses the risk of disclosiveness of a table and trade-off against statistical quality of the outputs.

In other words, although the end result is the same, they are achieved by different formulations of the problem, and unless it is clearly defined how one protection interval is compared to another, we are not able to compare the quality of the outputs produced.

During the evaluation of alternative solutions, researchers from the University of the West of England and Cardiff University were given the task to follow the same rules of protection as per the Optimal Cell suppression routine (Salazar, 2001), and develop alternative methods that were capable of dealing with larger instances of the data, and also subject to the protection levels as defined in the Optimal model.

A first phase of this work concluded that there were some issues in the way the current model was expressed which meant not all the experiments were completed in terms of understanding the capability of the proposed methods. For highlights on the issues found please refer to the Issues section of this paper.

A second phase was then commissioned to work within the restrictions of the model found in phase one, but focussing on the capability of the approaches rather than definitions of protection. The remit of each investigation was to compare results when using different parameters, different operators, the table safety check, i.e. reliability of the approaches, and the implications for each approach when running on hierarchical data sets. This paper focuses on the results obtained from phase two.

1.3 Proposed Solutions

Three concepts of the optimization algorithms in use need to be considered before the brief explanation of the approaches we developed. they are the notion of:

An optimal solution is described when the upper and lower bounds for a problem are meeting the same results, i.e. can’t be further improved and very often referred to as a global optimum. Traditionally this technique uses mathematical modelling implementations such as Linear Programming and Mixed Integer Programming, which are time consuming in computational terms as the problem sizes increase. Optimal Suppression is an example of this type of solution approach.

A Heuristic solution, on the other hand, is a technique which does not involve mathematical proof for finding optimal results. It can be of two types; Constructive or Improvement; and works by searching through the solution space for a representation of the problem which satisfies its constraints. Heuristic techniques are mostly described as quick search algorithms. An example of this type of approach is a greedy[3] algorithm to select candidate secondary suppressed cells.

A Metaheuristic technique, however, tries to overcome a potential local optimum solution from a starting solutions by adding learning inputs to the process. They also don't have a mathematical proof for optimality but often can deal with large datasets in reasonable computational times. Examples of metaheuristics applications in SDC are Tabu Search, and Simulating Annealing for the Controlled Rounding problem, by James P. Kelly.

Using Salazar definitions, the cell suppression problem involves ensuring that a table of cells is protected, so that certain cells, denoted primary cells, cannot have their values deduced from the published values in the table. Each cell has an associated **weight** and the objective is to find the set of cells to suppress to ensure that confidentiality is maintained, but minimising the sum of the **weights** of the suppressed cells. This is different to the **values** of the cells.

The algorithms that were evaluated in the first phase of investigation were:

(a) four variations of Greedy heuristic;
(b) local search algorithm (Descent method);
(c) Greedy Randomised Adaptive Search Procedure (GRASP);
(d) Ant Colony Optimization (direct and indirect models); and
(e) Evolutionary Algorithm.

Due to some of the imposed requirements for a solution to be available in a "reasonable" amount of time, an alternative feasibility check was created to speed-up some of the computational problems we encountered during phase one of the project in 2006. This is referred to as **relaxed feasibility check** as opposed to **strict feasibility check,** when following Salazar's incremental attacker model (2001), and were applied to options developed by ONS and Cardiff University.

The aims of phase two of the project were to provide more insights on how the approaches behaved when exposed to other settings of the algorithm, and to try to tune them for best results. For that a subset of the problems investigated in phase one were used. Table 1 describes the factors we were interested in analysing when changing the parameters, operators, verifying the approach for safety and for the hierarchical variable cases. In phase two only magnitude data was used because in phase one we revealed the methodological problem of the definition of protection when using frequency data (see section 4 in this paper for a summary).

Label	Number	Rows	Columns	Sensitive	% zero cells	Av. No Primary
A	5	200	5	10%	25%	60
B	5	200	5	2%	5%	60
C	5	200	50	10%	25%	553

[3] Greedy algorithm works by choosing the cheapest cell in a row/column which minimise the information loss whilst still guaranteeing protection of the primary cells.

D	5	200	50	2%	5%	526
E	5	4000	10	10%	25%	2387
F	5	4000	10	25%	5%	2201
G	2	654	14	19%	16%	1913
H1	1	14	1433	16%	14%	3680
H2	1	14	1433	16%	14%	3641
H3	1	712	10	6%	49%	407
H4	1	712	10	8%	49%	495
H5	1	712	19	11%	35%	1432
H6	1	712	19	13%	35%	1616

Figure 1: Table of artificial and a sample of real data created for this work when variables were non-hierarchical (A-G) and hierarchical (H1-H6).

2 Experiments Design

2.1 Parameter Optimisation

The methods used in phase one of the project depended on a number of parameters; for example, GRASP requires a candidate list size and number of cycles, and Ant Colony Optimisation requires an evaporation value, weights on the visibility, trails, and possibly a candidate size. The first set of experiments will focus on taking the best performing heuristics from the previous research, seeking to produce a more thorough parameter optimisation. In this way, we can ensure that the proposed methods are as efficient as possible.

From an Evolutionary Algorithm (EA) perspective, the first set of experiments was designed to determine whether there was any benefit to the use of a population-based approach as opposed to a simple local search method. A second goal was to determine the effect of changing the way in which solutions are perturbed by mutation (in the EA) or in the Local Search (LS) routine.

2.2 Operator Approach

For the GRASP approach, the algorithm was extended to consider the protection levels of the primary cells, meaning that cells in rows and columns that have been chosen to become secondary suppressed are no longer the ones that possessed the lowest cost, but also have to assure the protection limits are ensured.

For the EA approach, three different neighbourhood generation operators were used for the Local Search/Mutation steps, namely:

- Insertion: pick two random values in the permutation, and move the second to just behind the first, moving the intermediate elements along to accommodate the change;
- Swap: pick two random elements in the permutation and swap their positions; and

- Inversion: pick two random elements and invert the entire sub-permutation between them.

2.3 Table Safety

There were many difficulties in phase one of the project, particularly ensuring that a table was completely protected. Eventually, a working incremental attacker heuristic model implemented in a mathematical solver did enforce feasibility. This work will look again at the difference between a safe solution to the relaxed cell suppression problem, and a safe solution to the tight cell suppression problem. The intention would be to attempt to identify where the solutions to the relaxed problem are not feasible, and to see if the definition of the relaxed variant can be improved so that solutions to the relaxed problem are more likely to be truly protected. This required analysis of datasets, looking at the differences between the relaxed solution and tight solution, and working out means of reducing the gap between the two feasibility definitions.

One aspect that could have a dramatic effect on this work would be a simpler feasibility check.

2.4 Hierarchical Tables

The methods (a) to (e) considered in phase one of this project were not designed to work for hierarchical tables. For phase two, the current methods were adjusted to ensure they were sufficiently robust to deal with hierarchical and non-hierarchical datasets. An EA approach was implemented in a way that is transparent to the algorithm whether or not the data was hierarchical. However, better understanding of how the approach works under these circumstances will be the focus of attention.

3 Issues

Many issues with the Cell Suppression model have arisen from phase one, and others were further identified during phase two. This section highlights some of the findings and suggests points of further research we intend to pursue.

3.1 External bounds in attacker model and tight intervals set in Tau-Argus

Note that an attacker is assumed to know the values lb_i and ub_i of the lower and upper “external bounds”. This may not be a realistic assumption. The values of lb_i and ub_i supplied in the Tau-Argus JJ-format file are currently specified as 0.5 and 1.5 times a cell’s nominal value respectively. This is an issue that should be further considered.

3.2 Upper, Lower, and Sliding Protection levels set in Tau-Argus

Protection levels are only a feature of primary cells. However, if the levels are allowed to be defined as in the Fischetti & Salazar (2001) model, it may be possible to identify a contributor to a primary cell due to the secondary chosen not pursuing insufficient boundary gap.

In other words, the “less/more than or equal to” inequalities in expression (3) from Fischetti and Salazar (2001) paper need to be replaced by “strictly less/more than” inequalities as in:

$$f_{i_k}^{k} < a_{i_k} - LPL_k \quad \textit{and} \quad g_{i_k}^{k} > a_{i_k} + UPL_k \quad \textit{and} \quad g_{i_k}^{k} - f_{i_k}^{k} > SPL_k \qquad (3^*)$$

thus consistent with Tau-Argus’ Optimal Suppression protection limits. This is not a trivial distinction given that many table data and protection limit values tend to be small integers. The result is usually a distinctly larger set of secondarily suppressed cells when the table has many integer values, i.e., frequency tables and certain magnitude tables. The discovery of these problems in Fischetti & Salazar (2001) obliged us to modify our method accordingly and rerun experimental tests.

Knowing the external bounds lb_i and ub_i for all cells i = 1,…, n and which cells have been suppressed in the published table, an attacker will try to discover the minimum and maximum possible values, $f_{i_k}^{k}$ and $g_{i_k}^{k}$, of each sensitive cell i_k. The attacker can do this “*by solving a linear program in which the values* y_{ij} *for ... [specific] missing cells (i, j) are treated as unknowns*” (Fischetti & Salazar, 2001, page 1009).

For a given sensitive cell i_k, the minimum possible value $f_{i_k}^{k}$ can be found by solving the following linear programme (LP):

$$\begin{array}{lll} \text{minimise} & y_{i_k} & \qquad (4) \\ \text{such that} & \boldsymbol{My = b} & \\ & lb_i \le y_i \le ub_i & \textit{for all } i \in SUP \\ & y_i = a_i & \textit{for all } i \notin SUP \end{array}$$

Similarly, for a given sensitive cell i_k, the maximum possible value $g_{i_k}^{k}$ can be found by solving the same LP, but maximising y_{i_k}, *i.e.*, replacing objective function (4) by:

$$\text{maximise} \quad y_{i_k} \qquad (5)$$

Fischetti & Salazar (2001) state that the sensitive cell i_k is sufficiently protected if the solutions to (4) and (5) satisfy:

$$\min(y_{i_k}) \le LPL_k \qquad \text{and} \qquad UPL_k \le \max(y_{i_k}) \qquad (6)$$

However, to conform to the TauArgus Optimal Supression protection definition, the solutions to (4) and (5) should be strictly outside the interval [LPL_k, UPL_k]. In other words, rather than (6), we should require

$$\min(y_{i_k}) < LPL_k \qquad \text{and} \qquad UPL_k < \max(y_{i_k}) \qquad (6^*)$$

This is not a trivial distinction, given that table data values and protection limit values tend to be integers and often small.

Fischetti & Salazar (2001) state that if this condition is satisfied for all sensitive cells i_k then the whole table is feasible, i.e., sufficiently protected. However, given that the attacker will not know which of the suppressed cells are the sensitive ones, this condition should really be satisfied not just for each sensitive cell i_k, but also for each secondarily suppressed cell within the set SUP. If not, then the values of certain secondarily suppressed cells might be guessed, subverting the protection of the sensitive cell. This issue merits further investigation and research than was possible within the resources and time frame of the current project.

The Sliding Protection Level SPL_k was zero for all cells in all the JJ-format files supplied for testing purposes.

3.3 The Incremental Attacker Heuristic

Fischetti & Salazar (2001) state that their branch-and-cut (BC) approach finds an optimal set of secondarily suppressed cells that guarantees protection for all sensitive cells in a table. The approach is sophisticated, time-consuming, and identifies optimal solutions only for moderately sized tables. However, the authors do make use of a fast heuristic to find incumbent solutions at each node of the BC tree, based on a heuristic procedure from Kelly et al. (1992) and Robertson (1995). The heuristic starts by taking as input:

a given sequence of all the sensitive cells $\{i_1, \ldots, i_p\}$ to be protected; this sequence is heuristically determined according to decreasing weight in Fischetti & Salazar (2001), but in our method it is the key decision, as it defines the solution space in our Evolutionary Algorithm. A set SUP of suppressed cells that is initially equal to the set sensitive cells $\{i_1, \ldots, i_p\}$

The set SUP of suppressed cells is then augmented by solving a series of Linear Programmes (LPs), two per sensitive cell i_k in the order of the given sequence. The LPs use the cell weights, consistency equations, upper & lower bounds, and upper & lower protection limits provided by the JJ files output by Tau-Argus. Note that this does not necessarily minimise the number of secondarily suppressed cells in SUP, but rather their total weight.

The first LP, known as the UPL *incremental attacker problem*, identifies which cells need to be added to the set SUP in order to guarantee that the sensitive cell i_k is protected with respect to its upper protection limit UPL_k. For a given sensitive cell i_k, the LP is:

$$\text{minimise} \quad \sum_{i=1}^{n} c_i (y_i^+ + y_i^-) \tag{7}$$

$$\text{such that} \quad \boldsymbol{M(y^+ - y^-) = b} \tag{8}$$

$$0 \le y_i^+ \le UB_i \qquad \text{for all } i = 1,\ldots, n \tag{9}$$

$$0 \le y_i^- \le LB_i \qquad \text{for all } i = 1,\ldots, n \tag{10}$$

$$y_{i_k}^- = 0 \quad \text{and} \quad y_{i_k}^+ = UPL_k \tag{11}$$

where $y_i = a_i + y_i^+ - y_i^-$ is the attacker's estimate of the value of sensitive cell $i \in \{1,\ldots, n\}$ so that the non-negative decision variables y_i^+ and y_i^- are respectively the deviations above and below of y_i from the cell value a_i. $UB_i = ub_i - a_i \ge 0$ is the relative external upper bound on y_i^+. $LB_i = a_i - lb_i \ge 0$ is the relative external lower bound on y_i^-. The objective function coefficient $c_i = 0$ for all $i \in$SUP and $c_i =$ cell weight w_i for all $i \notin$ SUP.

After solving LP (7)-(11), the set SUP is augmented with all cells $i \notin$ SUP for which $y_i^+ + y_i^- > 0$ in the optimal solution.

Setting $c_i = 0$ for the set SUP's newly added cells i resulting from the solution of (7)-(11), the second LP similarly identifies which cells need to be added to SUP so that sensitive cell i_k is protected with respect to its lower protection limit LPL_k. This LP constitutes expressions (7)-(10), but with (11) replaced by:

$$y_{i_k}^+ = 0 \quad \text{and} \quad y_{i_k}^- = LPL_k \tag{12}$$

Fischetti & Salazar (2001) state that: "this guarantees the fulfilment of the upper/lower protection level requirement for i_k with respect to the new set SUP of suppressions." However, our experimental tests found exceptions to this statement. It was agreed that, from ONS's perspective, the upper and lower protection limits, UPL_k and LPL_k, have to be strictly obeyed, i.e., < and > rather than <= and >=, contrary to Fischetti and Salazar (2001) and as discussed in section 3.2.1 above. This meant that expressions (11) and (12) were respectively replaced by

$$y_{i_k}^- = 0 \quad \text{and} \quad y_{i_k}^+ = UPL_k + 1 \tag{13}$$

$$y_{i_k}^+ = 0 \quad \text{and} \quad y_{i_k}^- = LPL_k + 1 \tag{14}$$

For tables with integer cell values this generally resulted in:

1. a substantial increase in the number of secondarily suppressed cells.
2. sufficient protection for the primarily sensitive cells, as defined by expression (6*)
3. occasionally insufficient protection for the secondarily suppressed cells, as defined by expression (6*)

With respect to this last observation (#3), it is possible that some or all of the insufficiently protected secondarily suppressed cells are redundant, (i.e., not needed for primary protection). This merits further investigation beyond the scope of the current project, and is an issue on which we would like to continue to collaborate.

Arising from phase one issues, a major concern towards the end of phase two was the behaviour of the system in that it was not clear whether tables were being adequately protected. Typically it was noted that for the best solutions found, the min and/or max attacker problems would be reported as "infeasible" for several of the primary cells while the suppression set was being incrementally built up. Also, when the problems were re-solved to check for protection using the complete suppression set, most, or indeed all, of these problems would have disappeared.

During detailed discussions it became apparent that exactly the same behaviour was being observed with the Dash Xpress-MP version used by Cardiff University, which was implemented completely independently by ONS and Cardiff University. This may be to do with the way in which the models were specified in the original Fischetti and Salazar paper, although the reasons would seem to be rather subtle.

Given the success of the joint ONS-UWE bid for an EPSRC three year CASE studentship to study this issue, it was decided that it would be more valuable to spend the remaining allocated time considering further improvements to the way in which the each of the (a) to (e) approaches worked.

At the end of the first project it was suggested that it might be worth amending the constructive heuristic, so that instead of incrementally the primary cells one-by-one and generating new suppression sets, it might be possible to treat the primary cells in groupings of some form.

It was also noted during the analysis of the initial results that on these problems the EA evaluated both the row-order heuristic, and the weight-ordered heuristic, and that the latter never gave the best results found. This suggests that there may be some merit to treating together groups of cells belonging to a common marginal total.

This idea has been discussed in some detail and it was pointed out that although this might potentially greatly reduce the number of max/min attacker Linear Programming problems to be solved, the complexity of each one would increase which might make the overall run-time little different. It was agreed that UWE would investigate the feasibility of this approach, but that it would be considered supplementary to the original specification since it might involve considerable modifications to the way that the problems were specified.

4 Conclusion

This paper stresses the importance of keeping a close link between ISIs, NSIs and universities so that creative thinking is applied to the challenges of large tables with multiple hierarchies and varying densities of zeros and sensitive cells. All this work is being developed in close partnership with two UK universities, namely the University of the West of England (UWE, Bristol-UK) and Cardiff University. Dr. Alistair R. Clark and Dr. James Smith (both from UWE-Bristol) lead the work on

Evolutionary Algorithms (Clark and Smith, 2006 and 2007) and Dr. Jonathan Thompson (Cardiff) leads the work on Ant Colony Optimization and GRASP algorithms (Thompson 2006 and 2007).

The experiments have shown that GRASP is the preferred solution method, as Ant Colony Optimisation requires too much time for learning to take place. Even GRASP had to be simplified for run times to be reasonable for the larger datasets.

Various parameters have been considered and it was shown that performing additional cycles was unlikely to significantly improve solution quality. The results from the GRASP method were then assessed by the Incremental Attacker model (Salazar 2001) and here, there is little that can be done to improve the run times. On the smaller datasets, the run times are well within the desired times, and indeed, it has been shown that several solutions can be assessed by the incremental attacker model in a relatively short time. However the larger datasets are different, as just assessing one solution required up to 9 hours of run time. It is difficult to assess solution quality without knowing the optimal results but they appear to be encouraging.

This work has also produced a solution method for hierarchical datasets, similar to the GRASP method for non-hierarchical datasets. This again worked well but required considerable run times.

There are considerable gaps in some instances between the solutions to the relaxed problems and to the real problem, however in many cases the solutions generated by GRASP were already feasible and the incremental attacker model did not add any further suppressions.

On many tables, the EAs find solutions with between 75-90% of the cost of the heuristic solutions. In some cases the cost is only 28% of the heuristic cost. On most types of tables one of the EA-based approaches gives the lowest mean and minimum cost. On most types of tables the Inversion mutation operator gives the lowest mean and minimum results.

Analysis shows that on tables such as 14x1433 and 712x12 where Local Search algorithms are more effective, the EAs are stopping because of the inbuilt convergence threshold. Given that the LS algorithms often find better solutions after a large number of unsuccessful attempts, this suggests that this parameter has been set to terminate the EA too quickly.

Statistical analysis by ANOVA suggests that the Local Search is marginally preferable to a population size of ten, and that the Swap operator is best. However the overwhelming factor is the observed difference in results comes from the choice of “seed” used to create the tables. Thus, for example, in some cases (200x5, 200x50, 4000x10) there is considerable difference between the minimum costs tables for different instances (seeds), and the number of runs for each method may be different, so comparisons based on variance and absolute costs must be treated with a certain amount of caution.

For the 4000 x 10 tables with more sensitive cells, the algorithms do not have time to evaluate sufficient solutions to find major improvements, but even so cost reductions of between 8% and 24% are observed with Local Search.

This project was intended to assess the viability of the approaches to cell suppression developed in the first phase of the project. To that end we have conducted an extensive and highly computationally intensive set of experiments, the results of which have been described above.

In terms of the quality of the results obtained we have demonstrated that both the Local Search and Evolutionary Algorithm approaches are able to systematically improve on the quality of the solutions provided by the initial heuristics used. In some cases the improvements the improvement is dramatic – for example cost savings of up to 72% have been reported.

We have further analysed the differences between the two approaches tested, and reported on the combination of settings that gives the best results across the fairly broad set of problems used, namely a “steady state” Genetic Algorithm with population size 10, inversion mutation, and termination of runs if the population has remained converged for 5000 iterations.

However, these results have clearly demonstrated that for the larger tables with many sensitive cells, using a constructive heuristic to build a suppression set by treating each primary cell in turn is not a time-effective approach. While improvements of up to 24% were still obtained with the Local Search, each solution took on average 20 minutes to evaluate, which is not promising as a scalable approach unless significant computational resources are available.

While not in the original scope of this project, we have developed an alternative “grouping” approach which considers a whole row or column at once. This has been implemented and initial results show that the scalability issue seems to be largely solved. The benefit of this approach is that it requires absolutely no modification to the way that the Evolutionary Algorithm functions, and should not affect the validity of the findings contained in this report concerning parameter settings.

There remains one aspect that it was originally intended to consider, and which was not possible. This was an analysis of the degree of protection afforded by the evolved solutions. As discussed in Section 4, results obtained by both UWE and Cardiff University showed that there appear to be further problems with the formulation of the min/max attacker problems within the constructive heuristic. Consideration of these issues would have taken considerably more time than was budgeted for, with no guarantee of successful resolution. Therefore in conjunction with ONS it was decided to leave this issue for further work. We are of course pleased to report that we have obtained funding from the Engineering and Physical Sciences Research Council for a three year project to focus specifically on this issue.

5 Acknowledgements

The authors would like to thank the Statistical Disclosure Control Branch in Methodology Directorate and Neighbourhood Statistics Programme at the ONS, as well as the Scottish Executive for their collaboration and input throughout the project, specially to Neighbourhood Statistics and the Scottish Executive for funding phase one of the project and the Information Management Group at ONS for funding phase two of this project.

6 References

Clark, A. R and Smith, J. EA and the Cell Suppression Problem – ONS Internal report – phase 1, (2006)

Clark, A. R and Smith, J. Further Experiments to investigate the Cell Suppression Problem – ONS Internal report – phase 2, (2007)

Fischetti, M. and Salazar, J.J. The cell suppression problem on tabular data with linear constraints, Management Science 47, 7, (2001).

Shlomo, N. and Young, C. Quality Measures for Statistical Disclosure Controlled Data, Proceedings of Q2006 - European Conference on Quality in Survey Statistics, 2006. See http://www.statistics.gov.uk/events/q2006/downloads/W21_Shlomo.doc

Thompson, J, Metaheuristics for Cell Suppression Problem – ONS Internal report – phase 1, (2006)

Thompson, J, Further experiments to investigate the Cell Suppression Problem – ONS Internal report – phase 2, (2007)