Distr.
GENERAL

ECE/CES/SEM.54/4 (Summary)
3 April 2006

Original: ENGLISH
ENGLISH AND RUSSIAN ONLY

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

EUROPEAN COMMISSION STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD) STATISTICS DIRECTORATE

Joint UNECE/Eurostat/OECD Seminar on the Management of Statistical Information Systems (MSIS)
Sofia, Bulgaria, 21-23 June 2006

Topic (i): Changes in statistical processes

APPLYING EVOLUTIONARY ALGORITHMS TO DISCLOSURE CONTROL PROBLEM

Invited Paper prepared by Andrea Toniolo Staggemeier and Vic Duoba,
Office for National Statistics, United Kingdom

Summary

1. ISIs and NSIs face a difficult task in protecting their sources of information from disclosure of individuals whilst increasing the amount of publicly available and informative data. One of the widely used methodologies to address this problem is called Cell Suppression. This consists of trying to find an ideally optimal pattern of secondary suppressed cells that minimizes the total information loss subject to preserving the table additivity and the requirements for protection levels. Another methodology used in this area is called Controlled Rounding which rounds the nominal values of a table up or down, to their nearest integer of a base number, also preserving the protection level requirements for each cell and the additivity of the table. Both of these methods are present in the Tau-Argus software tool, which has been funded through the EU CASC project. Statistics Netherlands (Anco Hundepool), and University of La Laguna, Tenerife (Professor Salazar), have been major contributors to the dissemination of Mathematical Programming in the EU context, with additional software contributions from Germany.

2. Although the current implementations of these two available methodologies have their own merit, there remain practical problems due to the large size of some table that ONS and its external data providers need to deal with. The computational resource demands of mathematical programming algorithms that seek ideal solutions mean that often confidentializing of large tables either exceeds the maximum allocated time or exceeds resources in some other way.

GE.06-22071

Hence there is a requirement to invest in research to try to overcome the numerical and computational difficulties experienced by the current available methods in Tau-Argus.

3. Hybrid approaches, that use a less formal mathematical definition to a problem, are generally called heuristic methods. These methods have the greatest ability to generate fast near-optimal solutions to large problems, but on the other hand, they also may get trapped on local solution space and not be capable of improving the quality of their results through longer runtimes. This is a classical problem for most optimisation algorithms.

4. To meet the challenge of finding adequate solutions to large table confidentializing using heuristic methods, a "meta" phase is added to the heuristic process. This additional phase enables solutions of large problems within a reasonable limited amount of computational effort. Often this approach to solving complex systems is linked with mathematical programming, i.e. linear programming formulations, which undertake the evaluation process (feasibility check) of each solution.

5. Metaheuristic approaches are well known in the Operational Research field and have been successfully adapted to a variety of optimization, and combinatorial problems. However, it is a relatively new approach to Statistical Disclosure Control (SDC). Cox et al., 2006, were pioneers in the introduction of the combined approach of exact methods, i.e. Linear Programming, and heuristic/metaheuristic. For example, Tabu Search and Scatter Search procedures were applied to disclosure control, with a different methodology (to that in the Tau Argus "Optimal" methods) called Controlled Tabular Adjustment (for the US Bureau of the Census).

6. This research has served to motivate the idea of using another metaphoric approach in ONS called the Evolutionary Algorithm (EA). The EA mimics the natural evolution process by representing a problem solution as genetic material, and through processes of selection, reproduction, and mutation evolves a population of solutions tending towards best results. The learning process is implemented by offspring inheriting good bits of each parent, and later a degree of *meme* can be added to each individual solution. In the context of disclosure control we are talking about evolving a population of suppressed patterns of a table such that at the end of the generating process we have selected the best individual solution to represent the algorithmic solution to the problem.

7. Other metaphoric approaches are also under investigation at ONS, such as the use of Ant Colony principles and Greedy Randomised Adaptive Procedure to guide the search for the best suppression pattern in a table.

8. 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 and Dr. Jonathan Thompson (Cardiff) leads the work on Ant Colony Optimization and GRASP algorithms. For the purpose of this paper we will focus on only one metaheuristic approach namely the Evolutionary

Algorithms. For more details on Ant Colony Optimization or GRASP please do not hesitate to contact the authors of this document.

9. ONS is supporting and participating in pioneering work in the field of the application of Evolutionary Algorithms to the SDC, bringing the state-of-art research and algorithmic developments in this area closer to real world applications. For convenience, we will focus the description of our approach on the Cell Suppression Problem and a description on how the heuristic methodology can be applied. We also believe that after the initial phase of research has been completed it can be extended to the Controlled Rounding Problem.

10. A trade-off between optimal solutions obtained by the current optimal (i.e. optimal wrt a specific mathematical model) suppression methodologies in Tau-argus (Fischetti and Salazar, 2001) and the solutions obtained by the new technology developed have recently been established that show that there is no detriment to safe suppressed patterns (as defined by the "optimal" method). A choice of a standard Incremental Attacker Model (Fischetti and Salazar, 2001), which verifies the solution for safety, is used to check solution feasibility and also serves as our evaluation criteria on how good/bad the solution generated is in relation to the total information loss. Initial results have shown that for small problems this computational effort is relatively insignificant and very effective to guide the search for optimal solutions, whist for large tables this can take for example 30 seconds to find a solution for a table with 10,000 cells and 700 primary disclosive cells. Further testing on a wide range of data sources will be undertaken in order to verify the initial promising findings.

11. If further testing yields results compatible with the early results, ONS will have determined a more scalable approach to confidentializing (although we emphasize that we acknowledge the need for further confirmatory experimentation and stress-testing). The new research approach relaxes the requirement of full optimality in favour of achieving protection with a suppression pattern, which is likely to be close to optimal. This can be achieved using heuristic, and/or metaheuristic methods. It is ONS's intention to make available any successful developments in the application of heuristics to protecting large tables through Tau Argus. Improvements in the way we handled the required number of feasibility checks and the process of guiding the Evolutionary Algorithm, through reproduction and mutation operators, are discussed on the paper. Initial results and further directions to be taken in this research are also presented at the end of the document.

12. This paper highlights the importance of keeping Statistical Offices closely linked to universities and other NSIs in order to encourage creative and innovative thinking for the solution of difficult problems that can be addressed by advanced applied mathematics (OR) and high-powered computing platforms.

References

Cox, L.H., Glover, F., Kelly, J.P. & Patil, R.J. (2006), "Confidentiality Protection By Controlled Tabular Adjustment: An Analytical and Empirical Investigation of Exact, Heuristic and Metaheuristic Methods," Decision Sciences Institute, to appear.

M. Fischetti, J.J. Salazar (2001). "Solving the Cell Suppression Problem on Tabular Data with Linear Constraints", Management Science, vol. 47, no. 7, pp 1008-1026.