CASC PROJECT

Computational Aspects of Statistical Confidentiality

September 2003

Changes to Plan in CASC: Replacing Testing of Masking Methods by Development of Record Linkage Software

Note concerning discontinuation of work on tasks TM4 and TM5 of workpackage 5, and replacement of Deliverable 6-D3

Sarah GIESSING,
Federal Statistical Office of Germany
65180 Wiesbaden
E-mail: sarah.giessing@statistik-bund.de

12.09.2002

Changes to Plan in CASC: Replacing Testing of Masking Methods by Development of Record Linkage Software

Note concerning discontinuation of work on tasks TM4 and TM5 of workpackage 5, and replacement of Deliverable 6-D3

Sarah GIESSING,
Federal Statistical Office of Germany
65180 Wiesbaden
E-mail: sarah.giessing@statistik-bund.de

1. *Introduction*

According to the original project plan, it was foreseen to include at least three alternative methods that seemed to be suitable for production of safe micro-data files from business statistics. In the course of the project, it became evident that one of these alternative approaches, Sullivan's algorithm, would probably turn out to be of less practical relevance. On the other hand, it became obvious that it would be a great benefit for µ-ARGUS to offer some kind of record linkage software for disclosure risk assessment which, however, was not foreseen in the project plan. Therefore, CASC partners decided to stop further work on Sullivan's algorithm. They agreed it would be better instead to use the part of the project budget originally meant to be spent on further research on applicability of Sullivan's method to pay work on implementation of a record linkage tool. During the 2nd project review the steering committee informed on this 'change to plan'. The proposal did not receive any negative comment from the reviewers, and was thus assumed to be accepted.
The following section will give some details about work relating to implementation and research on Sullivan's method foreseen in the project plan, the overall progress made to date, milestones and deliverables that were cancelled, and the amount of project resources which could be re-allocated because work on this part of the project was stopped. The last section will explain the need for implementation of the record linkage tool, and give an outline of this task.

2. *Tasks related to research and implementation of Sullivan's algorithm*

As explained in the description of work for the CASC project (annex 1 to contract no. IST-2000-25069), a part of the project resources were supposed to be spent on implementation and testing of a microdata masking method (Sullivan, 1989), e.g. task 2 of workpackage 1-1, tasks TM3 and TM4 of workpackage 5, and task TS3.2 of workpackage 6. The partner responsible for these tasks is the German Federal Statistical Office (DESTATIS). Sullivan's method which seemed to be especially well suited for application to business microdata was supposed to be integrated in µ-ARGUS.
Apart from this method, other alternative approaches to protect business microdata files were supposed to be included in the ARGUS framework, such as micro-aggregation (task T3 of workpackage 1-1), and model based methodology (task T1 of workpackage 1-1). In the course of the project, even one more method to protect business microdata (e.g. record swapping) which could be implemented without much effort was included in the package additionally.

Task 2 of workpackage 1-1 was completed duly. The conclusion of the report supplied as final deliverable of this task (deliverable 1.1-D1, (Brand, 2002)) was rather discouraging, emphasising that "... an algorithm as complex as the one proposed by Sullivan can only be applied by experts. Every application is very time-consuming and requires expert knowledge on the data and the algorithm.". The report insinuated that it would be hard for Statistical Institutes to practically apply the method properly. When the report was composed, first results on task TM4 of workpackage 5 (comparison of Sullivan's method vs. micro-aggregation techniques) were already available. According to these first results, there was some evidence that the

advantage of Sullivan's method over micro-aggregation in terms of data quality could be expected to be at most moderate, while the effort for using Sullivan's method is much higher. This conclusion gave the impression that for practical applications it was likely that the method would be of minor relevance.
After the decision to stop further research on Sullivan's method during CASC, task TM4 remained incomplete. Task TM3 (application of the method to various datasets in order to test the applicability) was dropped entirely, together with the deliverable concerned, e.g. deliverable 5-D4. Project resources saved this way accumulated to 11 PM's. Work on Task TS 3.2 of workpackage 6 (software testing) was redefined, now addressing a careful test of the implementation of the record linkage software in μ-ARGUS.

3. *Implementation of record linkage software*

Particularly during the discussions at the first CASC project meeting (Plymouth, April 2002) it became clear that it would be important to supply along with μ-ARGUS tools to help users choose between different methods, or parameters for methods. While it is relatively straightforward to provide some indicators for the information loss induced by application of SDC methods, it was emphasised that it would be irresponsible to provide users with measures for information loss, but without any indicators for disclosure risk remaining after application of that SDC methods. This might tempt users to choose parameters as to minimise information loss, without making them aware that the resulting data might not be much safer from re-identification as the original data set.
(Domingo-Ferrer, Torra, 2001) suggest to use empirical methods such as record linkage for disclosure risk assessment, when the goal is to compare disclosure risk in data sets resulting from application of different SDC measures. It therefore was considered necessary to implement such a software and integrate it into ARGUS. Such a task, however, had not been foreseen in the original project plan, so no project capacity had been included in the budget.
Thus it seemed to be an ideal solution to re-allocate project resources related to implementation and test of Sullivan's method to the new task of implementing software for record linkage. This could be seen as an additional task T3 (development of record linkage methodology and software) of workpackage 1-2 (Microdata: new disclosure risk assessment methodology) with an additional deliverable 1.2-D6 "C++ implementation of record linkage for disclosure risk assessment" to be delivered in month 33. In order to have the new software be tested properly, as mentioned above, task TS 3.2 of workpackage 6 (software package) will focus now on testing the record linkage tool, instead of the implementation of Sullivan's method. Again, the partner responsible for these tasks is DESTATIS. Meanwhile, the development of suitable record linkage methodology and software is nearly completed. A first prototype version has been delivered, which will be tested and refined during the rest of the project.

4. *Summary*

In the course of the CASC project it turned out that it would be highly desirable to supply along with μ-ARGUS record linkage software for disclosure risk assessment of protected micro-data sets. Therefore some of the tasks concerning research on Sullivan's masking algorithm were dropped and the project resources re-allocated to development and implementation of record linkage software. This affected in particular deliverable 5 D-4, which was replaced by a new deliverable 1.2-D6 "C++ implementation of record linkage for disclosure risk assessment." Sullivan's masking algorithm had, in earlier phases of the project, found to be difficult to use and therefore seemed to be of minor relevance for practical applications

References

Brand, R. (2002), 'Tests of the Applicability of Sullivan's Algorithm to Synthetic Data and Real Business Data in Official Statistics, deliverable 1.1-D1 of the CASC-project, unpublished report.
Domingo-Ferrer, J., Torra, V. (2001), 'A Quantitative Comparison of Disclosure Control Methods for Microdata', In: 'Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies' Doyle, Lane, Theeuwes, Zayatz (Eds), North-Holland
Sullivan, G.R. (1989): The Use of Added Error to Avoid Disclosure in Microdata Releases, unpublished PhD-Thesis, Iowa State University