Project N°: 262608

ACRONYM: Data without Boundaries

DELIVERABLE D11.4

(Software Code RCTA and ECTA Implementation)

WORK PACKAGE 11

(Improved Methodologies for Managing Risks of Access to Detailed OS Data)

REPORTING PERIOD:	From: Month 19	To: Month 24
PROJECT START DATE:	1st May 2011	DURATION: 48 Months
DATE OF ISSUE OF DELIVERABLE:	July 2013	
DOCUMENT PREPARED BY:	P10, P20	UPC, ULL

Combination of CP & CSA project funded by the European Community
Under the programme "FP7 - SP4 Capacities"
Priority 1.1.3: European Social Science Data Archives and remote access to Official Statistics

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 262608 (DwB - Data without Boundaries).

Preface

This document is the descriptive part associated with the source code that is delivered as deliverable D11.4. Different implementations of different approaches to the Controlled Tabular Adjustment (CTA) are produced by UPC (Technical University of Catalonia, Spain) and ULL (University of La Laguna, Spain).

The code provided by UPC mainly concerns the RCTA approach. The ULL code mainly concerns the ECTA approach. However, since all approaches are special versions of CTA, in the descriptions below, the overall term CTA is used as well.

The first part of this descriptive report concerns the code provided by UPC. The second part concerns the code provided by ULL.

TABLE OF CONTENTS

PART I - SOFTWARE CODE FOR RCTA IMPLEMENTATION

INTRODUCTION: PURPOSE OF SOFTWARE

The package implements the controlled tabular adjustment (CTA) method for the protection of statistical tabular data. It implements both the CTA version with binary decisions (a Mixed Integer Linear Programming – MILP – problem) and a new version that only needs the solution of four Linear Programming (LP) problems. This new version is based on a lexmin (lexicographic minimization) multi-objective approach.

It is a package mainly to be used by European National Statistical Institutes for the protection of tabular data.

DESCRIPTION OF CONTENT

Design and Usage

The package can be used in three different ways:

- As a standalone application through command line.
- As a standalone application through a Graphical User Interface (GUI), especially useful for non-expert users. Figures 1-3 show three screenshots for some particular states of the GUI. Figure 1 corresponds to the screen for the solution of a MILP problem; Figure 2 shows the screen for a LP; and Figure 3 shows a solver log file, and the output file with the input and protected cell table values.
- As a callable library to create an own program or used in other applications developing ad-hoc main programs.

The current version of the CTA package is linked with six state-of-the-art solvers: CPLEX, Xpress, GLPK, CBC, CLP and SYMPHONY. CLP is only valid for LPs; the other solvers can deal with both LP and MILP problems. CBC uses CLP as the LP solver. This multi-solver platform was developed using Osi (Open Solver Interface), which provides an abstract interface to communicate with solvers. CPLEX and Xpress are commercial solvers and a license is needed, but GLPK, CBC, CLP, and SYMPHOYNY are license free solvers.

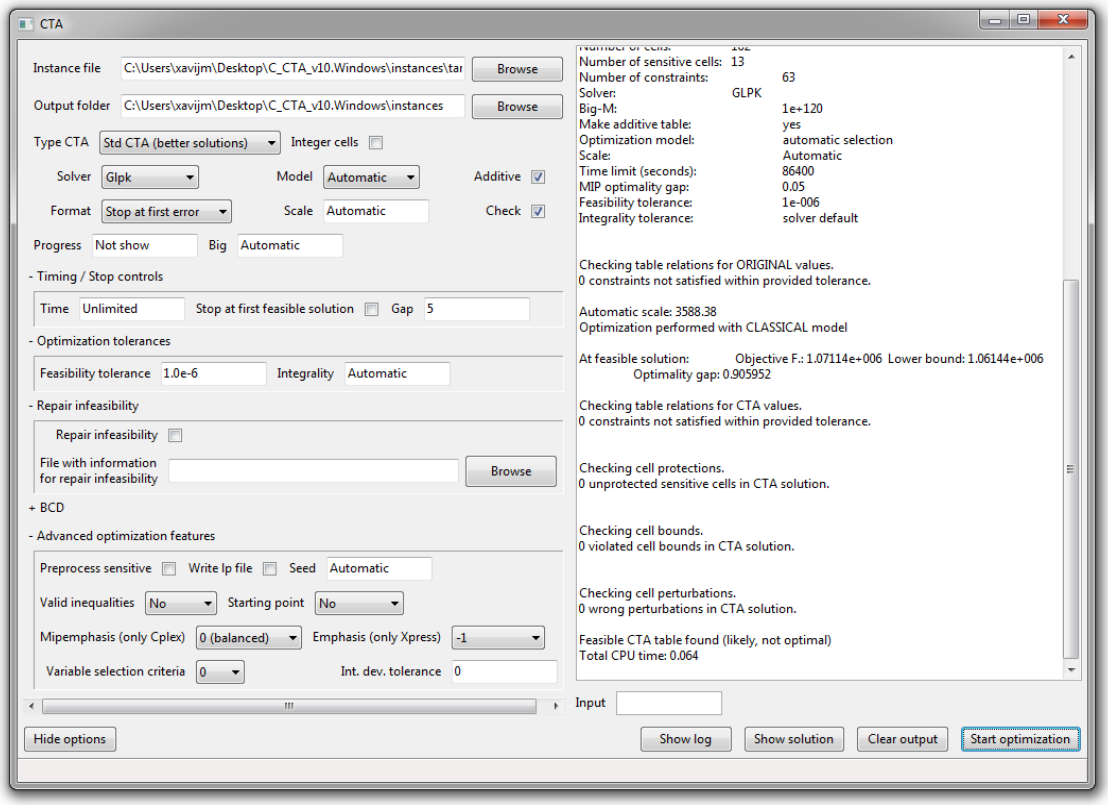

Figure 1. Screenshot of GUI for the solution of a MILP

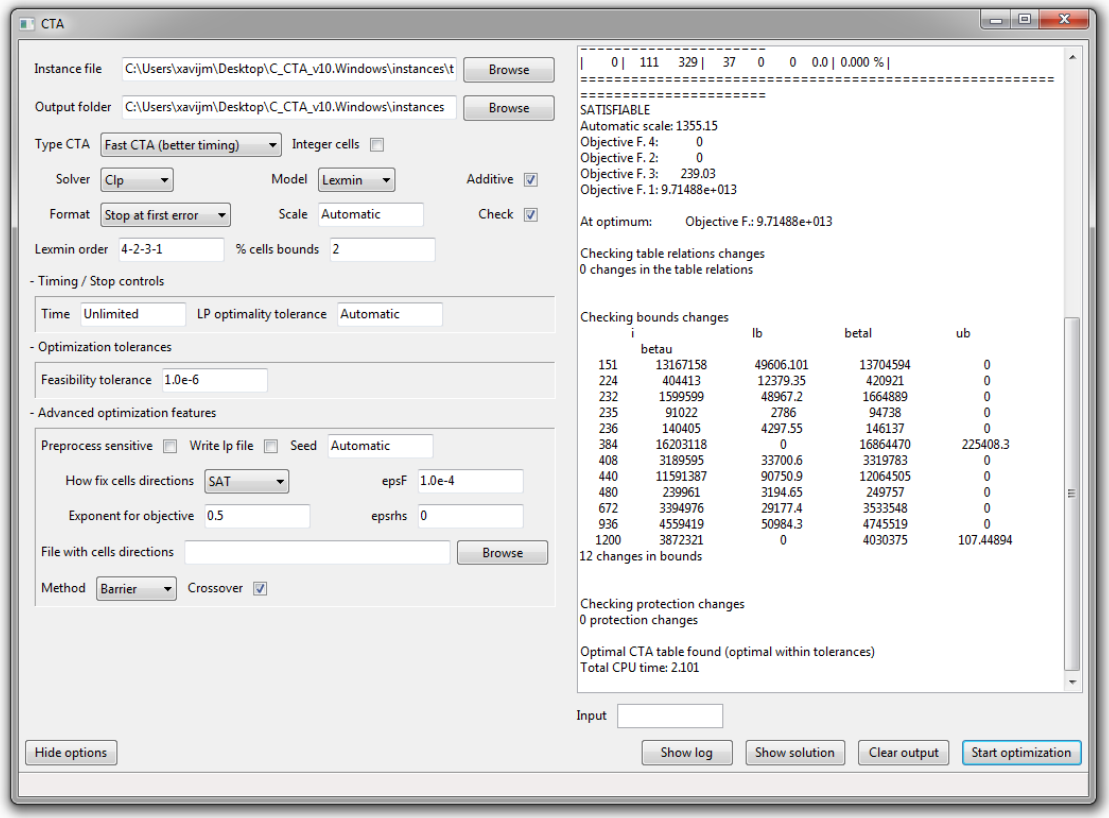

Figure 2. Screenshot of GUI for the solution of a LP

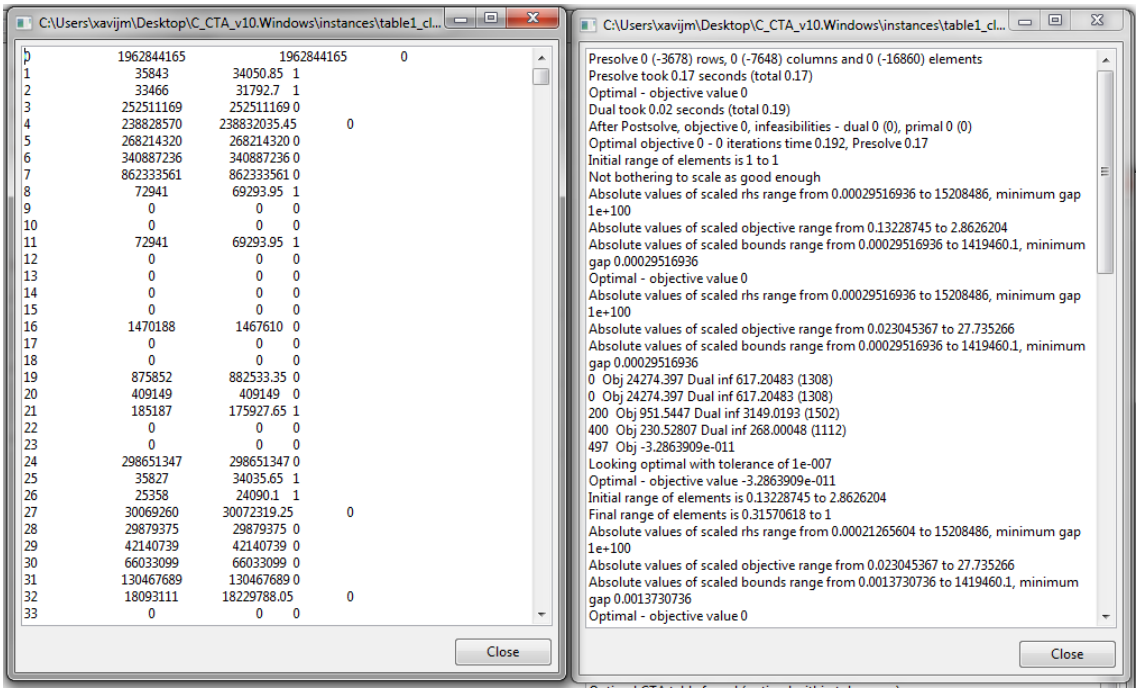

Figure 3. Screenshot of the solver log and output file with original and protected values

Outline of Package Features

The CTA package was started some years before DwB. Some new features have been added to the existing CTA package within DwB, as planed in the original working plan. The main features of the current version are:

- It has about 15000 lines of C/C++ code. Approximately 9000 of them have been developed within DwB.
- Addition of a Block Coordinate Descent (BCD) heuristic. This heuristic suboptimally solves the MILP CTA problem, by decomposing it into simpler subproblems.
- New fast CTA version, where binary decisions are pre-fixed. In this case CTA becomes a continuous LP. Since it may be infeasible, with this variant we allow for changes in the bounds, protection levels or the table additivity constraints. Together with the original objective this amounts to four objective functions, resulting in a multiobjective optimization problem. The whole problem is solved by assigning priorities to the four objectives, and applying a lexicographic minimization (lexmin) approach. This new version requires the solution of four LPs, and it may solve large tables in few seconds, though the quality of the resulting table may be lower than with the MILP CTA version.

- Graphical user interface.
- Both Linux and Windows versions.
- Extension for integrality in cell values. In general integrality is guaranteed, but in some tables it may be lost; this new option allows to force integrality. The resulting model is harder and more time consuming, and thus it is only recommended for not too large tables.
- Autoscaling of input data. This avoids several numerical problems related to the optimization solvers.
- Usage of Osi for communication with several LP and MILP solvers. The package is linked with two commercial (Cplex and Xpress) and four free solvers (CBC, Symphony, GLPK, CLP), whereas only one was required at the beginning of the project.

Many of the above features (multi-free solvers, graphical interface, autoscaling, integrality of cell values…) were not initially planned in the project. For this reason the number of person-months allocated to UPC will be higher than initially planned at the end of the project.

CONCLUSIVE REMARKS

The final version of the package is to be provided in D11.7 (month 48). However, there will be no significant differences from the current version (mainly tuning of the different features), since most of UPC person-months have already been consumed. It is worth noting that the software is being provided as of mid July 2013 (month 27), as in month 24 (initial delivery date) the lexmin approach was still under development.

The reference manual, and user's manual will be provided in deliverables D11.6 and D11.8 (respectively months 39 and 48).

The package is available in binary format as this is considered to be sufficient for most National Statistical Institute and/or researcher. It can be downloaded at: http://www-eio.upc.es/~jcastro/tmp/DwB/D11.4/CTAWindowsBinaries.zip

Those justifying the need for the source code may obtain it upon request to UPC. Source code cannot be redistributed to third parties.

PART II - SOFTWARE CODE FOR ECTA IMPLEMENTATION

INTRODUCTION: PURPOSE OF SOFTWARE

During the first two years of the DwB project we have being working on the implementation of computer programs as described in the WP11 tasks assigned to our team at ULL.

We have produced the following codes:

[CTA] A computer program to use Controlled Tabular Adjustment on tables

Controlled Tabular Adjustment (CTA) is a new methodology to protect private information when publishing statistical tables. It is based on perturbating the cell values of the table. It was a technique introduced by Dandekar and Cox (2002), and has attracted several practitioners in the last years, including EUROSTAT. The problem of finding the perturbation on each cell is also an optimization problem, easier to solve than the cell suppression one.

We have detected and study some disadvantages of CTA, and have proposed an alternative called Enhanced Controlled Tabular Adjustment (ECTA). During our research we have implemented a program for CTA that we have released as deliverable in WP11.

[ACTA] A computer program to audit a CTA solution

One of the disadvantages of CTA when compared with cell suppression is that a CTA solution may not be protected. For that reason, it is fundamental to have a procedure to check the protection of a CTA solution. Due to the potential large number of sensitive cells in a table, checking all the protection level requirements may consume too much time, and for that reason we have investigated and implemented a sophisticated auditing program based on a non-commercial code. This is the program that we have delivered in WP11.

[ECTA] A computer program to apply Enhanced Controlled Tabular Adjustment

Based on our analysis of CTA on real-world tables, we have proposed a new algorithm to achieve the same aim of CTA, but avoiding the major drawbacks of this technique. The new algorithm is called "Enhanced Controlled Tabular Adjustment" (ECTA), and is based on solving a

sequence of models. Each model is definitively much simple to solve than the CTA model, and hence also easier than the cell suppression model. Indeed, on magnitude tables, the ECTA models are polynomial-time solvable, while the CTA model remains NP-complete. In addition, the ECTA algorithm also considers the utility aspect of the table from the user point of view, not considered in CTA.

We have implemented and analyzed an algorithm to apply ECTA. Our implementation was done using non-commercial tools, and using the same libraries as the other methodologies to a fair comparison. The new code has proved to have much better performances on our benchmark tables.

CONTEXT & DISSEMINATION

We have presented our research results in different forums, including:

- Oral presentation at "Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality", Tarragona, Spain, 26-28 October 2011.
  Title: "Enhanced Controlled Tabular Adjustment"
  Web:
  http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2011/01Agenda.pdf

- A seminar at "Statistics and Operational Research Doctoral Training Centre", Lancaster University, 29 June 2012.
  Title: "Statistical Disclosure Control: Techniques to protect confidential information "
  Web: http://www.stor-i.lancs.ac.uk/event-info/stori-seminar-gonzalez

- A seminar at Westat Inc (Washington), 26 November 2012.
  Title: "Statistical Confidentiality: Modern Techniques to Protect Sensitive Cells when Publishing Tables"
  Web: http://washstat.org/sem2012.html

- A Seminar at U.S. Census (Washington), 27 November 2012.
  "Enhanced Controlled Tabular Adjustment: a new approach to protect sensitive cells when publishing a table".

- A seminar at C.B.S. (Netherlands), 29 January 2013.
  "Enhanced Controlled Tabular Adjustment: a new approach to protect sensitive cells when publishing a table"

HOW IT WORKS

A compressed file containing all source codes and documentation is available in https://www.webs.ull.es/users/jjsalaza/public/ECTA&CSP.zip
This file was given to the DwB WP11 Leader on January 29, during a visit to CBS. Some comments and remarks have been incorporated after and the compressed file updated.

The compressed file contains several folders, one for each code as described above.

Each folder contains a README file describing all the details of the code, including how to compile it, what it does, examples, etc. It is fully documented. It also contains two subfolders: one with the source code and another with the executable. The executable contains an instance to test that it runs properly. The source code allows a computer programmer to change the program and produce a new executable or a DLL library to be called from another software tool. All files have been created using non-commercial tools, so no commercial software is necessary to read, edit or compile them.

AVAILABILITY TO WIDER PUBLIC

All the material is publicly available at
https://www.webs.ull.es/users/jjsalaza/public/ECTA&CSP.zip