CASC PROJECT

Computational Aspects of Statistical Confidentiality

REFERENCE DATA SETS TO TEST AND COMPARE SDC METHODS FOR PROTECTION OF NUMERICAL MICRODATA

(Unscheduled Deliverable)

Ruth Brand (Destatis)
Josep Domingo-Ferrer (URV)
Josep M. Mateo-Sanz (URV)

April 2002

The CASC project is partly subsidised by the 5th framework program of the EU. Grant no IST-2000-25069.

0. Introduction

During the CASC Research Meeting held in April 2002 in Plymouth, the need was detected for reference data sets to test and compare microdata SDC methods. This document is a non-scheduled joint CASC deliverable by Destatis and URV and contains the description of three data sets which are proposed as reference data for numerical microdata protection.

The three data sets are described below and are called “CENSUS”, “EIA” and “Tarragona”. The “CENSUS” and “Tarragona” data sets have been used by the URV team in their publications. The “EIA” data set is proposed by Ruth Brand (Destatis) as an additional and somewhat larger data set. All three data sets seem to represent reasonably well real data sets used in business statistics.

1. “CENSUS” Data Set

The test data set was obtained on July 27, 2000 using the Data Extraction System of the U. S. Bureau of the Census (http://www.census.gov/DES/www/welcome.html).

1.1 Extraction procedure

The extraction procedure followed through DES was as follows:

- [Level 1] From the available data sources, the Current Population Survey was chosen.
- [Level 2] Year 1995 was chosen.
- [Level 3] The available filegroup chosen was ``March Questionnaire Supplement - Person Data Files".
- [Level 4] We used the ``File content documentation" option to find out which variables were numerical. We then requested MAX records and all 54 numerical variables (in alphabetical order the first and last selected variables were AERNLWT and WSWAL, respectively). Default values were taken for the rest of extraction parameters.

The above procedure yielded 149642 ASCII records which were fed to an MS-Excel spreadsheet.

1.2 Variable selection

Out of the 54 numerical variables arising from the extraction procedure, 38 variables took a 0 value for more than 120000 out of the 149642 extracted records; these variables were discarded. Out of the remaining 16, three variables were excluded because they took values over a very restricted set and could actually be regarded as categorical; these were HRSWK, MARGTAX and WKSWORK. Eventually, 13 variables remained, which were AFNLWGT, AGI, EMCONTRB, ERNVAL, FEDTAX, FICA, INTVAL, PEARNVAL, POTHVAL, PTOTVAL, STATETAX, TAXINC, WSALVAL. In addition the record sequence number formed by FSEQ, HPOS, HSEQ and PHFSEQ was also preserved in the resulting data.

1.3 Record selection

The 149642 records that were retrieved had a substantial number of zero values for the 13 variables that were selected. In a continuous variable, one would not expect zero values to appear very often. On the other hand, missing values (99999) appeared now and then. Therefore, records with zero or missing value for at least one of the 13 variables were discarded, which left 12062 records.

To further reduce the data set (so that there are less repeated values and many record linkage experiments can be carried out in a short time), the following was done:

1) For each different value of FEDTAX, only one record was preserved, so that the reduced data set had no repeated values for FEDTAX. The record preserved for each value of FEDTAX was the one with the lowest record sequence number.
2) The previous step was repeated for variables AFNLWGT, AGI, EMCONTRB, PTOTVAL, TAXINC, STATETAX (in this order). After this step, none of the above variables had repeated values.
3) To go below 1100 records, eliminating some repetitions of another variable was required. The remaining records were sorted by POTHVAL and repeated values of this variable were eliminated until there were 1080 records left.

In this way, we got a microdata set with the following properties:

- The number of records is less than 1200 and thus manageable by the Census probabilistic record linkage software.

The CASC project is partly subsidised by the 5th framework program of the EU. Grant no IST-2000-25069.

- There are seven variables with no repeated values: FEDTAX, AFNLWGT, AGI, EMCONTRB, PTOTVAL, TAXINC, STATETAX. The lack of repeated values is a feature of really continuous variables.
- 1080 (the number of records) is the largest integer less than
- 1200 which is a multiple of 5, 8 and 9. Thus, the data set can be microaggregated with minimal group sizes *k*=3,4,5,6,8,9 and 10 so that all groups have size *k*.

1.4 Description of variables

Variables in this data set are as follows:

AFNLWGT	Final weight (2 implied decimal places)
AGI	Adjusted gross income
EMCONTRB	Emplyr contribution for hlth insurance
ERNVAL	Business or Farm net earnings in 19..
FEDTAX	Federal income tax liability
FICA	Soc. sec. retirmnt payroll deduction
INTVAL	Amt of interest income
PEARNVAL	Total person earnings
POTHVAL	Total other persons income
PTOTVAL	Total person income
STATETAX	State income tax liability
TAXINC	Taxable income amount
WSALVAL	Amount: Total Wage & salary

2. “EIA” Data Set

Data set obtained from the U.S. Energy Information Authority. It consists of 4092 records and can be found the web address:
http://www.eia.doe.gov/cneaf/electricity/page/eia826.html

2.1 Extraction procedure

The procedure to obtain this data set was as follows:

1) Go to the web page above.
2) Choose “Download, Year 1996” which gets you an executable DOS file.
3) Run the DOS file. This results decompresses two files: one of them contains data in '.dbf' format (which can be opened from Excel); the second is a metadata file.

2.2 Description of variables

Variables (as described in the metadata file) are as follows:

Field	Field Name	Field Description	Type	Width	Dec
1	UTILITYID	UNIQUE UTILITY IDENTIFICATION NUMBER	Numeric	10	0
2	UTILNAME	UTILITY NAME	Character	30	NA
3	STATE	STATE FOR WHICH THE UTILITY IS REPORTING	Character	2	NA
4	YEAR	REPORTING YEAR FOR THE DATA	Numeric	2	0
5	MONTH	REPORTING MONTH FOR THE DATA	Numeric	2	0
6	RESREVENUE	REVENUE FROM SALES TO RESIDENTIAL CONSUMERS	Numeric	10	0
7	RESSALES	SALES TO RESIDENTIAL CONSUMERS	Numeric	10	0
8	COMREVENUE	REVENUE FROM SALES TO COMMERCIAL CONSUMERS	Numeric	10	0
9	COMSALES	SALES TO COMMERCIAL CONSUMERS	Numeric	10	0
10	INDREVENUE	REVENUE FROM SALES TO INDUSTRIAL CONSUMERS	Numeric	10	0
11	INDSALES	SALES TO INDUSTRIAL CONSUMERS	Numeric	10	0
12	OTHREVENUE	REVENUE FROM SALES TO OTHER CONSUMERS	Numeric	10	0
13	OTHRSALES	SALES TO OTHER CONSUMERS	Numeric	10	0
14	TOTREVENUE	REVENUE FROM SALES TO ALL CONSUMERS	Numeric	10	0
15	TOTSALES	SALES TO ALL CONSUMERS	Numeric	10	0

3. “TARRAGONA” Data Set

A real data set comprising figures of 834 companies in the Tarragona area. Data correspond to year 1995.

3.1 Description of variables

For each company, 13 quantitative variables are given:
1) Fixed assets
2) Current assets
3) Treasury
4) Uncommitted funds
5) Paid-up capital
6) Short-term debt
7) Sales
8) Labor costs
9) Depreciation
10) Operating profit
11) Financial outcome
12) Gross profit
13) Net profit.

The CASC project is partly subsidised by the 5th framework program of the EU. Grant no IST-2000-25069.