CASC PROJECT

Computational Aspects of Statistical Confidentiality

November 2003

Test report on network solution for large unstructured 2 dimensional tables in τ-Argus 2.2.2[1]

Marta Mas
Basque Statistics Office (Eustat)

Enric Ripoll
Julià Urrutia
Statistical Office of Catalonia (Idescat)

Deliverable No: 6D-8

[1] This paper has been carried out as a result of a collaboration agreement between Idescat and Eustat in the frame of the CASC project

CASC Project
Deliverable 6-D8

Test report on network solution for large unstructured 2 dimensional tables in τ-Argus 2.2.2[1]

Marta Mas
Basque Statistics Office (Eustat)

Enric Ripoll, Julià Urrutia
Statistical Office of Catalonia (Idescat)

1. Introduction

The objective of the deliverable is to publish the results of a test on the network solution for large unstructured 2 dimensional tables implemented in τ-Argus 2.2.2. The network flows protection method provides secondary suppression patterns to protect tabular data in the special cases of 2 dimensional tables with no hierarchical structure.

Recently, the dissemination in EUSTAT of a wide data bank with detailed information from large tables has lead to confidentiality problems. Some cells had to be aggregated under a safety criterion and therefore some accuracy in the information was lost. The aim of this work is to find a suitable suppression pattern for some of these tables by means of the network methodology and compare this solution with the aggregation of categories that is applied at this moment.

Particularly, data derived from the Industry and Construction Survey in the Basque Country will be used in this analysis. Two different heuristics described in Castro, J. [2] will be applied, as well as several combinations of parameters required by the network package.

2. Data description

The Industry and Construction Survey in the Basque Country 2000 is a sampling survey. Only companies with less than 20 employees are sampled and all the others are forced to enter the sample. The final sample file contains a total of 2935 records, each one representing an establishment which is an economic unit that carries out an economic activity (i.e. one company can have one or more establishments).

[1] This paper has been carried out as a result of a collaboration agreement between Idescat and Eustat in the frame of the CASC project
[2] Castro, J. (2003), *User's and programmer's manual of the network flows heuristics package for cell suppression in 2D tables*. Technical Report DR 2003-07

However, and due to the peculiarities of the industrial and construction sector in the Basque Country, the weighting procedure is made at activity level, particularly using the National Classification of Economic Activities (1993). Therefore, the file used for tabulation and analysis is already weighted and the record unit is the activity.

Finally we have a file with 842 records (activities x region) and 10 variables described here down:

A84 – Sectorization of the National Classification of Economic Activities. (48 categories)
CNAE93 - National Classification of Economic Activities. (330 categories)
ID_THL – Historic Territory -region- (3 categories).
ID_ESTB– Number of establishments.
GAS_COMPNET – Net purchases.
ING_VENTNET – Net sales.
PBSF.- Gross production.
VABCF.- Gross added value.
CP.- Personnel cost.
INVR.- Investment.

The tables considered for the analysis will be the following:

- Macro-figures for Industry and Construction by activity (A84) and historic territory.
- Profit and Loss accounts for the Industry and Construction by activity (A84) and historic territory.

The metadata file required by Argus is specified in a .rda file as follows:

```
<SEPARATOR>  ","

a84 2
  <RECODEABLE>

cnae93 5
  <RECODEABLE>

id_thl 1
  <RECODEABLE>

id_estb 8
  <NUMERIC>

gas_compnet 8
  <NUMERIC>

ing_ventnet 8
```

<NUMERIC>

pbsf 8
<NUMERIC>

vabcf 8
<NUMERIC>

cp 8
<NUMERIC>

invr 8
<NUMERIC>

3. Sensitive rules

Once we have read the file in τ-Argus, it is necessary to specify the tables to be protected and the sensitive rules that will determine the unsafe cells. The safety criterion imposed by EUSTAT for the tables published in the data bank, is based on the number of establishments which contribute to each cell. If there are only 3 or fewer establishments in a cell, this value cannot be published. However, our contributors are not establishments or companies but "activities". The only way to detect these sensitive cells is by representing the number of establishment as the cell value.

For the same reason, there are not specific rules for quantitative variables in terms of the contribution to the cell value. In fact, to establish a dominance rule in terms of the activity (record unit) is not a trivial issue. Only the manual safety range is given as sensitive parameter in τ-Argus and it is set as the given by default (30%).

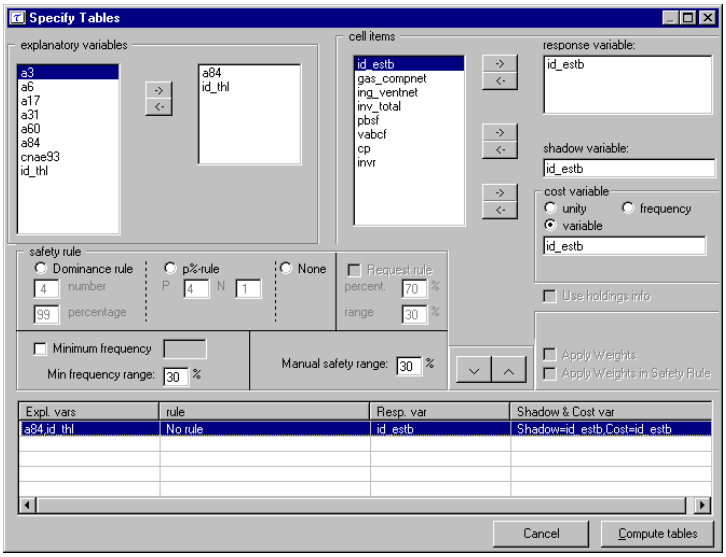

Figure 1: *Specify Table option in τ-Argus*

The pattern of primary suppressions will be set manually in τ-Argus or by means of an hist file containing the status of each cell [1]. As a consequence, this pattern will be the same for any quantitative variable that we represent in the cell value.

See the distribution of the number of establishments by activity and historic territory in Table 1:

	Historic Territory			
Activity (A84)	**Araba**	**Bizkaia**	**Gipuzkoa**	**Total**
8	-	3	-	3
9	11	29	28	68
10	31	62	51	144
11	28	20	40	88
12	4	34	52	90
13	74	289	220	583
14	29	42	51	122
15	427	93	101	621
16	-	-	1	1
17	25	80	67	172
18	56	329	178	563
19	11	24	25	60
20	104	521	413	1038
21	17	49	92	158
22	143	663	483	1289
23	-	2	-	2
24	11	27	17	55
25	27	46	30	103
26	11	32	16	59
27	13	108	51	172
28	51	163	143	357
29	24	39	30	93
30	1	2	3	6
31	63	167	119	349
32	36	36	39	111
33	-	21	9	30
34	30	79	55	164
35	160	543	455	1158
36	35	193	96	324
37	269	790	941	2000
38	103	336	483	922
39	18	39	113	170
40	10	27	38	75
41	137	407	488	1032
42	3	17	11	31
43	51	177	150	378
44	11	33	48	92
45	51	240	158	449

46	27	85	42	154
47	2	70	57	129
48	18	21	19	58
49	149	579	419	1147
50	27	159	109	295
51	4	16	1	21
52	10	25	48	83
53	4	11	7	22
54	7	8	20	35
55	2846	9741	8576	21163
Total	5169	16477	14593	36239

Table 1: *Number of establishments by activity (A84) and historic territory*.

Sensitive cells are shaded in grey. Until now, the protection problem was solved by aggregating the sensitive categories with others (related or not) in this way:

08/09-Petroleum and gas extraction/ Metal and non-metal minerals

14/16-Other food industry/ Tobacco

30/31-Concrete, lime and plaster/ Other non-metal industry

32/33-Steel industry/Non-ferrous metallurgy[3]

42/43-Office machinery and computer equipment/ Electric material

47/48-Ship construction/Other transport material

50/51-Other manufacturing/Recycling

The exact value of 56 cells (marginals included) is lost and an aggregated value is given instead.

4. The network flows solution

As it has been mention at the beginning of this document, the objective of this testing is to check the network flows solution for the secondary suppression problem. This protection technique has been implemented recently in τ-Argus and considers 2-dim tables with no hierarchical structure.

Detailed documentation about this package can be found in [2] and [4]. Only few remarks about the method and the parameters needed are given below.

[3] Although at this period (2000) these activities are not sensitive, the aggregation covers sensitive cells in other periods of time.

4.1 The heuristics

Two types of heuristics are implemented in the package:

1. *"0-1-flows" heuristic*: only flows 0 or 1 are sent through the network. Shortest-path subproblems are formulated and efficiently solved by Dijkstra's algorithm. See [4] and [5] for detail.
2. *The "n-flows" heuristics*: the network can transport any positive flow. The subproblems formulated are minimum-cost network problems, and are solved through PPRN. See [3] and [6] for detail.

Solutions provided by the n-flows heuristic are always equal or better than those computed by the 0-1 flows one. On the other hand, for large problems, the n-flows heuristic may be much slower than the 0-1 flows and Dijkstra combination. Choosing one or another option means a trade-off between quality of solution and efficiency. However, both options will be tested in this work.

4.2 Cell Weights

Cell weights are used in the objective function to be minimised by the heuristics. The default type of weights is the cell value.

4.3 Lower bounding

The heuristics provide an approximate solution to the cell suppression problem. To know how far that solution is from the optimal one, we should get some lower bound to the optimal objective function (i.e., minimum value or minimum weight suppressed). Computing the lower bound means solving a linear programming problem, and a CPLEX7.5 license is needed for that.

As we did not buy such a license, for our example this option will be not active.

4.4 Merit order for primary cells

The heuristics are iterative processes that sequentially protect each primary cell. The order primary cells are selected (named merit order in Network package) may modify the final solution. The user can choose between three merit orders: NORMAL, ASCENDENT and DESCENDENT. If the ASCENDENT order is selected, cells will be protected according to theirs cell values sorted in ascendant order (i.e., the first cell protected will be that with the lowest cell value, and so on).

4.5 Type of costs for objective function

The costs of arcs in the "0-1 flows" heuristic network are dynamically created for each primary cell by the heuristic. The purpose of these costs is to guide the protection procedure, making unsupervised cells with low weights better candidates for suppression than those with larger weights.

Computing these costs can be fairly expensive, and the package offers two alternatives. The first one, called FASTER WORSER, computes a set of costs efficiently; however these costs are not the best ones, and, usually, provide worse solutions than the second set of costs. This second set is the SLOWER BETTER. As its name shows, the heuristic is slower if these costs are computed, although the solution provided can be better.

4.6 Auditing

An auditing phase computes, after the protection process, the lower and upper bounds that an external attacker could derive from the primary cells after the publication of the table.

If the protection process is well done, the primary cells are protected between the required safety range (in our case the "manual" safety range). If x is the cell value of primary i, and l and u are the lower and upper safety bounds the attacker only knows that the real value of this primary is in the range $[x_i - l; x_i + u]$.

As a result, we have a protected and auditing table.

5. Analysis phase

At this point, we are ready to protect Table 1 using the Network flows solution. We are going to distinguish two different phases:

- *Combining parameters*. As we have seen in section 4, a set of parameters can be chosen to perform the method. Some of them are not compatible (i.e. it is not possible to run "n-flows" heuristic with Dijkstra's solver,…) but other are interchangeable and could provide different results.

- *Comparing results*. Outputs and results derived from the execution of the method in several "versions" (different heuristics, combination of parameters,…) will be compared and commented in this section.

5.1 Combining parameters

In the first running of the method the parameters were set to default. The values and options taken by default by the package are the following:

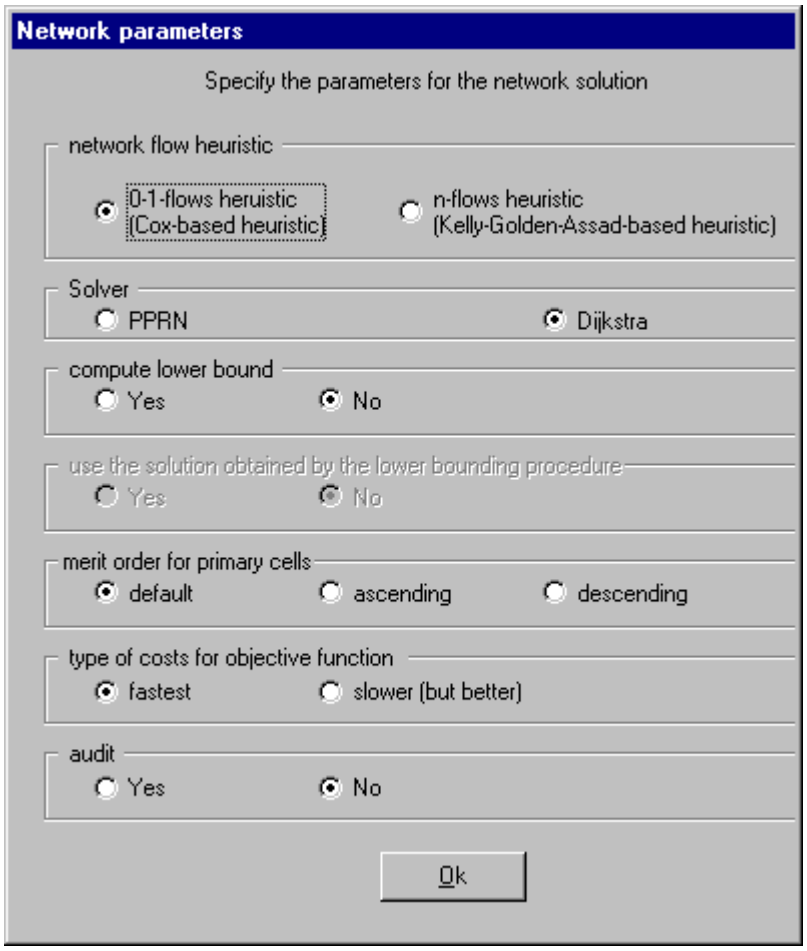

Figure 2 *Default Parameters for the Network method.*

The method runs and immediately this informative window appears. Four secondary suppressions were needed to protect the table under the conditions specified in Figure 2.

Figure 3

Execution time has not been taken into account for this study as the table is not very large and times are, in any case, very small. However, times given by the method (5 seconds, in Figure 3) are not representative of the real execution time of the method, as it seems to include the time spent choosing the parameters.

The table with the secondary suppressions looks as follows (**X**- Primary, **S**- Secondary):

	Historic Territory			
Activity (A84)	**Araba**	**Bizkaia**	**Gipuzkoa**	**Total**
8	-	**X**	-	**X**
9	11	29	28	68
10	31	62	51	144
11	28	20	40	88
12	4	34	52	90
13	74	289	220	583
14	29	42	51	122
15	427	93	101	621
16	-	-	**X**	**X**
17	25	80	67	172
18	56	329	178	563
19	11	24	25	60
20	104	521	413	1038
21	17	49	92	158
22	143	663	483	1289
23	-	**X**	-	**X**
24	11	27	17	55
25	27	46	30	103
26	11	32	16	59
27	13	108	51	172
28	51	163	143	357
29	24	39	30	93
30	**X**	**X**	**X**	**S**
31	63	167	119	349
32	36	36	39	111
33	-	21	9	30
34	30	79	55	164
35	160	543	455	1158
36	35	193	96	324
37	269	790	941	2000
38	103	336	483	922
39	18	39	113	170
40	10	27	38	75
41	137	407	488	1032
42	**X**	17	**S**	31
43	51	177	150	378
44	11	33	48	92
45	51	240	158	449
46	27	85	42	154
47	**X**	70	**S**	129
48	18	21	19	58
49	149	579	419	1147
50	27	159	109	295
51	**S**	16	**X**	21

52	10	25	48	83
53	4	11	7	22
54	7	8	20	35
55	2846	9741	8576	21163
Total	5169	16477	14593	36239

Table 2. *Table protected by Network method with secondary suppressions*

The same process is run for different combinations of parameters. Next table summarises values and results obtained for those combinations:

Heuristic	**Solver**	**Merit Order**	**Secondary Suppressions**
"0-1" flows	Dijkstra	Default	4
"0-1" flows	Dijkstra	Ascending	4
"0-1" flows	Dijkstra	Descending	3
"0-1" flows	PPRN	Default	3
"0-1" flows	PPRN	Ascending	3
"0-1" flows	PPRN	Descending	3
"n" flows	PPRN	Default	3
"n" flows	PPRN	Ascending	3
"n" flows	PPRN	Descending	3

Table 3. *Values of parameters and results for the Network method*

The lower bound option was not applied in any of the cases because a CPLEX license is needed, and it was not available at the moment of doing this work. The type of costs for the objective function in "0-1" flows heuristic, has been, in all the cases, the SLOWER (but better) option as the differences in execution time with the FASTER one were not noticeable, thus, we chose the best option. The option "audit" was active in all the executions in order to check and validate the method.

5.2 Comparing results

As the efficiency is not a problem here, it is clear that we will prefer the "n-flows" method, which gives a better solution in terms of number of suppressions. Nevertheless, we have tested that "0-1" flows heuristic gives also an optimal solution using PPRN solver and in one case (merit order: descending) with Dijkstra.

Finally, the suppression pattern chosen includes the following cells:

Primary			
Activity (A84)	**Historic Territory**		
08	Bizkaia		
08	Total		
16	Gipuzkoa		
16	Total		
23	Bizkaia		
23	Total		
30	Araba		
30	Bizkaia		
30	Gipuzkoa	**Secondary**	
42	Araba	Gipuzkoa	
47	Araba	Gipuzkoa	
51	Gipuzkoa	Araba	**Total**
Suppressions	12	3	15

Table 4. *Suppression pattern provided by the Network flows method*

A total of 15 suppressions are needed to protect Table 1. Now the decision consists on either publishing the aggregated values as explained in section 3, or not to release the exact values of 15 cells but to gain 41 real values suitable for publication.

The final decision lays on the responsible of the survey at EUSTAT, and , at this moment, a solution is being discussed.

6. Conclusions

The network flows package has provided a balanced solution suitable for publication. Although the table used is not very large (250 cells), the method seems to be very efficient.

Some exploratory analysis was made through bigger tables. See the table below, only to compare with other implemented solutions:

Number of cells	Primary suppressions	Method	Execution time (seconds)	Secondary suppressions
250	12	**Network flows**	1	3
250	12	GHMiter (Singleton)	1	8
250	12	GHMiter	4	4

1660	351	**Network flows**	7	62
1660	351	GHMiter (Singleton)	12	133
1660	351	GHMiter	146	100

Table 5. *Execution times and number of suppressions by method.*

Looking at the results, it would be interesting to test the same solution in case of hierarchical structures and/or more than 2 dimensions.

7. Report on problems

- Sometimes it is difficult to distinguish between parameters in τ-Argus and parameters of the Network flows method. "Cell weights" in Networks flows are similar (or the same) as "cost variable" in the table specification window in τ-Argus. Which one is used for the cost function?
  In order to avoid problems the cell value was specified as "cost variable" in τ-Argus option, as it is set by default in Network solution.

- It is not possible to control the cell status values from the Network parameter window. We assume that the status provided by Argus (Safe, unsafe and protected) are those taken by default by the heuristics in the Network flows package.

- As mentioned before in this report, the execution times given by the method at the end of the process are not very representative of the real time that the method takes to find the solution to the suppression problem. It seems to depend on the time spent choosing the parameters of the method.

- The computing of the lower bounding procedure for the optimal solution is not possible unless a commercial solver (CPLEX7.5) is bought.

- No information on the auditing phase is provided (neither by the method, nor by Argus).

- If n-flows heuristic is not compatible with Dijkstra algorithm, this option should not be available depending on the heuristic chosen. Now it is possible to choose n-flows and Dijkstra at the same time and, of course, the program fails (as it is expected).

References

[1] Anco Hundepool et al. *τ-Argus Version 2.2. User's Manual*. Document 4.2-D1. CASC project. April 2003.

[2] Castro, J.(2003), *User's and programmer's manual of the network flows heuristics package for cell suppression in 2D tables.* Technical Report DR 2003-07. Dept. of Statistics and Operations. Research, Universitat Politécnica de Catalunya, Barcelona, Spain

[3] Castro, J., *PPRN 1.0, User's Guide*, Technical report DR 94/06 Dept. of Statistics and Operations. Research, Universitat Politécnica de Catalunya, Barcelona, Spain, 1994.

[4] Castro, J., *Network flows heuristics for complementary cell suppression: an empirical evaluation and extensions*, in {\sl LNCS 2316, Inference Control in Statistical Databases}, J. Domingo-Ferrer (Ed), (2002) 59--73.

[5] Cox, L.H., *Network models for complementary cell suppression*. J. Am. Stat. Assoc. 90, (1995) 1453–1462.

[6] Kelly, J.P., Golden, B.L, Assad, A.A., *Cell Suppression: disclosure protection for sensitive tabular data*, Networks 22, (1992) 28–55.