



Nakładanie Tajemnicy Statystycznej w badaniu DG-1 z wykorzystaniem programu Tau-Argus

Marcin Milwicz, Magdalena Homenko

Óśrodek Statystyki Krótkookresowej

Dział Metodologii i Programowania

Plan prezentacji

1. Cel prezentacji
2. Definicja ogólna tajemnicy statystycznej
3. Krótkie wprowadzenie do badania DG-1
4. Omówienie procesu nakładania tajemnicy statystycznej w badaniu DG-1
5. Przygotowanie plików do programu τ -Argus
6. Etapy postępowania przy nakładaniu tajemnicy pierwotnej i wtórnej w programie τ -Argus
7. Wnioski



Cel prezentacji

- Praktyczne wykorzystanie programu τ -Argus na danych zagregowanych
- Zasady nakładania tajemnicy statystycznej w badaniu DG-1
- Znaczenie tajemnicy statystycznej w ochronie danych

SDC – Statistical Disclosure Control

- ***Statistical disclosure control*** techniques can be defined as set of methods to reduce the risk of disclosing information on individuals, businesses or other organisations. SDC methods minimise the risk of disclosure to an acceptable level while releasing as much information as possible.”
 - ***Handbook on Statistical Disclosure Control***
- **SDC** – zbiór metod redukujących ryzyko odkrycia informacji jednostkowych, są kompromisem pomiędzy publikowaniem danych jak najbardziej szczegółowych, a zapewnieniem jak największej ochrony danych jednostkowych

SDC – Statistical Disclosure Control

- Dla tabel z danymi agregowanymi stosuje się następujące metody SDC:
 - Reguły koncentracji (ukrycie pierwotne):
 - Reguła $p\%$ - gdzie p oznacza dokładność, z jaką możemy ustalić udział danych jednostkowych w agregacie
 - Reguła (n,k) – n jednostek stanowi więcej niż k procent udziału w agregacie
 - Łączenie wierszy / kolumn z danymi
 - Zaokrąglanie / zniekształcanie danych
 - Odpowiednia ochrona agregatów z małą liczbą jednostek (w szczególności singletonów)

Informacje ogólne o badaniu DG-1

- Badanie „Ocena bieżącej działalności gospodarczej przedsiębiorstw” (realizowane na formularzu DG-1 „Meldunek o działalności gospodarczej”) - źródło danych o kondycji podmiotów gospodarczych w Polsce
- Pozyskane dane dotyczą podstawowych mierników, które obrazują sytuację gospodarczą i społeczną w kraju oraz w każdym z województw
- Badanie DG-1 jest badaniem miesięcznym
- Obejmuje osoby prawne i jednostki organizacyjne niemające osobowości prawnej oraz osoby fizyczne (w których liczba pracujących wynosi 10 i więcej osób), których działalność została zaklasyfikowana wg PKD2007 do działalności w zakresie wybranych sekcji (oraz wg PKD2025)

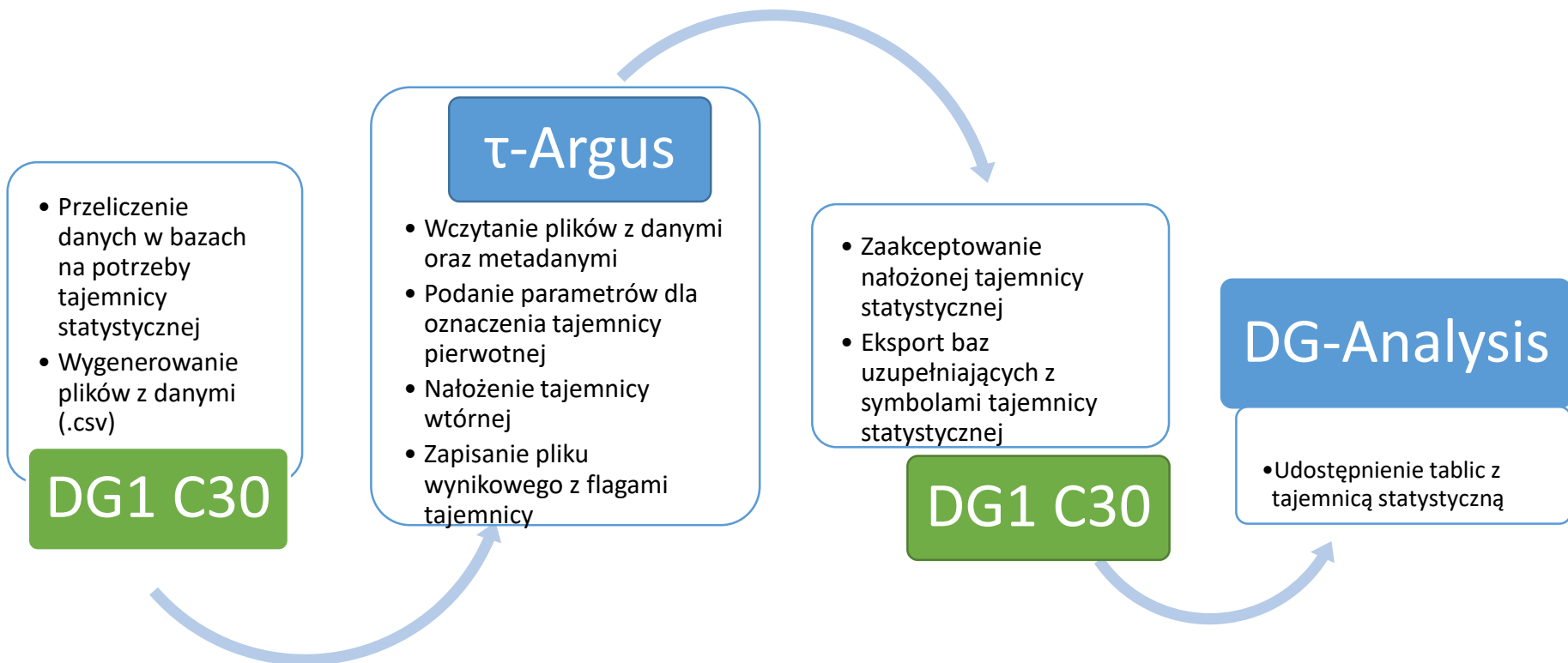
SDC w badaniu DG-1

- W badaniu DG1 oznacza się pierwotną i wtórną tajemnicę statystyczną (brak łączenia agregatów, zniekształcania danych)
- Za tajny lub jawny uznaje się cały agregat, a nie poszczególne zmienne
- Tajemnica nakładana jest na dane uogólnione i agregatowe
- Agregat jest tajny pierwotnie, gdy:
 - liczba jednostek tworzących go jest mniejsza niż 3 lub
 - udział jednej jednostki w agregacie przekracza 75% wartości tego agregatu – występuje dominacja reguła $(n,k) \rightarrow (1, 75)$
- Dominację wyznacza się na podstawie zmiennej obrót (ON_2B)
- Tajemnica wtórna jest wyznaczana metodą hipersześcianów (HypeCube)

Przetwarzanie danych wojewódzkich – ścieżka postępowania

- Standardowe przetwarzanie danych wojewódzkich
- Akceptacja danych wojewódzkich przez jednostki autorskie badania
- Nałożenie tajemnicy statystycznej na dane wojewódzkie i ogólnopolskie
- Akceptacja tajemnicy statystycznej
- Poinformowanie o zakończeniu nakładania tajemnicy statystycznej
- Technolodzy:
 - Eksport tablic dla urzędów statystycznych
 - Eksport baz uzupełniających z symbolami tajemnicy statystycznej
- Przeglądanie tablic z tajemnicą statystyczną (aplikacja DG-1 C30 oraz DG-Analysis)

Schemat procesu nakładania tajemnicy statystycznej



Nałożenie pierwotnej i wtórnej tajemnicy statystycznej w programie τ -Argus - przygotowanie

W programie τ -Argus potrzebne są następujące pliki:

- Plik z danymi
- Plik opisujący hierarchie danych (.hrc)
- Plik sterujący (.rda)

Wynikowo otrzymujemy [plik](#) z oznaczoną tajemnicą w formacie tekstowym (poniżej) oraz [raport](#) (plik html).

```
"Total","Total",1710473118.0,1
"Total","AA",8253189.1,1
"Total","A",8253189.1,1
"Total","A02",8213325.6,11
"Total","A03",39863.5,3
"Total","A9",867409759.6,1
"Total","B",31206397.3,1
"Total","B05",15575024.8,1
"Total","B051",14956988.4,1
"Total","B052",618036.4,1
"Total","B06",266018.2,1
"Total","B061",190574.1,3
"Total","B062",75444.1,11
"Total","B07",10455676.6,3
"Total","B071",- ,14
"Total","B072",10455676.6,3
```

Tworzenie pliku z danymi

Plik z danymi (np. .csv, .asc)

- kolumny z danymi bez nagłówków
- separator danych – średnik (może być inny)
- dla danych liczbowych – części dziesiętne / setne – po kropce

```
TOTAL;TOTAL;97271;242936012.8;5295040.7;  
TOTAL;AA;928;1223062.9;37141.3;  
TOTAL;A;928;1223062.9;37141.3;  
TOTAL;A02;911;1215536.9;37141.3;  
TOTAL;A03;17;7526.0;4907.3;  
TOTAL;A9;31550;114977759.4;5295040.7;  
TOTAL;B;460;3329996.7;886098.9;  
TOTAL;B05;28;1938464.3;787096.0;  
TOTAL;B06;6;30356.8;24824.5;  
TOTAL;B07;3;886500.1;886098.9;  
TOTAL;B08;342;318419.7;19720.9;  
TOTAL;B09;81;156255.8;35731.0;  
TOTAL;C;28837;89872713.9;5295040.7;  
TOTAL;C10;5508;16709414.4;351716.3;  
TOTAL;C11;170;1933169.1;277164.3;
```



Pliki opisujące hierarchie danych – np. teryt / kodAgregatu

Plik tekstowy z rozszerzeniem .hrc

- poziom ogółem – np. RAZEM, SUMA, TOTAL (ale tego poziomu nie uwzględniamy w pliku)
- znak określający poziom (zagłębienie) w hierarchii

```
AA
BA
  AA02
  AA03
  A9
  BA
    BB05
    BBB051
    BBB052
    BB06
    BBB061
    BBB062
    BB07
    BBB071
    BBB072
    BB08
    BBB081
    BBB089
    BB09
    BBB091
    BBB099
  C
    CC10
    CCC101
    CCC102
```

5
6

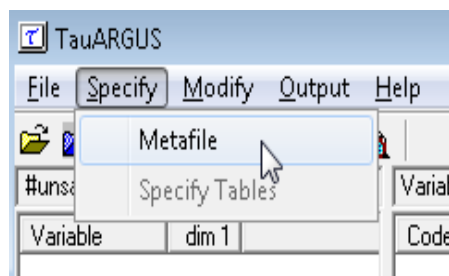
Plik sterujący

Plik tekstowy z rozszerzeniem .rda

- jakie informacje zawierają poszczególne kolumny w pliku z danymi
- ścieżka do pliku/ów z hierarchiami

```
<SEPARATOR> ";"
<SAFE> "S"
<UNSAFE> "U"
<PROTECT> "P"
WON
  <RECODEABLE>
  <TOTCODE> TOTAL
  <HIERARCHICAL>
  <HIERCODELIST> "D:\_tArgus\bazy\won.hrc"
  <HIERLEADSTRING> "@"
PKD
  <RECODEABLE>
  <TOTCODE> TOTAL
  <HIERARCHICAL>
  <HIERCODELIST> "D:\_tArgus\bazy\PKD_B3_A9.hrc"
  <HIERLEADSTRING> "@"
Ljed
  <NUMERIC>
  <FREQUENCY>
Obr
  <NUMERIC>
  <DECIMALS> 1
MaxOJ
  <DECIMALS> 1
  <MAXSCORE>
```

Tworzenie pliku sterującego sterującego



- explanatory variables
- frequency
- response variable (wg tej zmiennej wyznaczana jest dominacja)
- topN variable

Specify metafile

Free format

Separator ;

WON
PKD
Ljed
Qbr
MaxQJ

Attributes

name: MaxQJ

length: 0

decimals: 1

☐ explanatory variable

☐ response variable

☒ topN variable

☐ frequency

☐ shadow

☐ status indicator

☐ cost var

☐ lower prot. level

☐ upper prot. level

☐ distance for suppression weight

Codelist

☒ automatic

Code for Total

☐ codelist filename

Missings:

1: 2:

☐ hierarchical

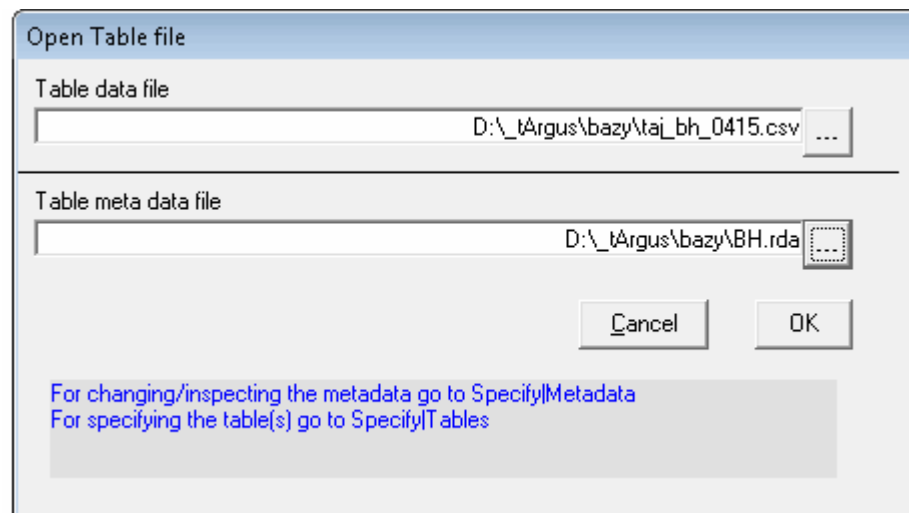
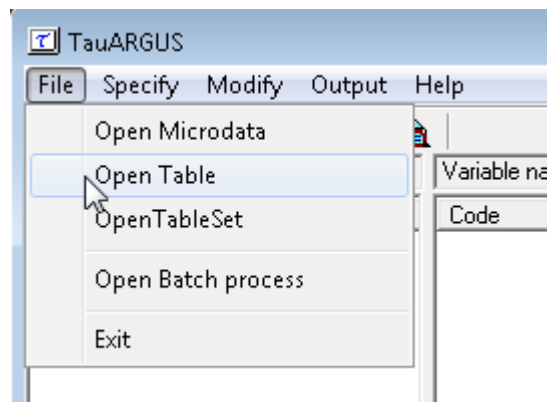
☐ Levels from microdata

☐ Levels from file

Leading string

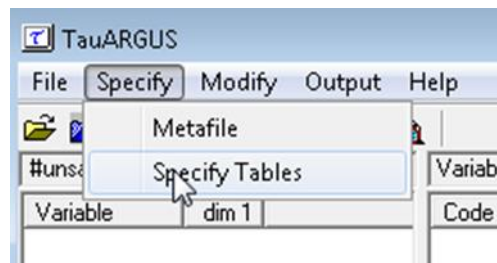
Cancel OK

Proces nakładania tajemnicy statystycznej – pojedyncze pliki z danymi



Proces nakładania tajemnicy statystycznej – pojedyncze pliki z danymi

- Tajemnica pierwotna



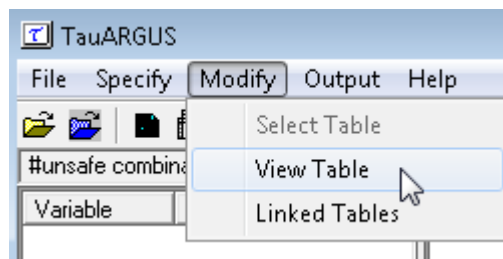
The 'Specify table' dialog box is shown with the following settings:

- Variables:**
 - Explanatory: WON, RD, RJ
 - Number: 3
- Status:**
 - Frequency
 - TopN: 1
- Cost:**
 - CostFunction for second. suppression: ☒ ResponsVar, ☐ Frequency, ☐ Unity, ☐ Distance
 - Lambda: 1
- Do not allow non-additivity** (selected) or **Allow non-additivity** (unselected)
- Compute incorrect totals** (unselected)
- safety rule:**
 - ☐ Use given status
 - ☒ Use safety rules
 - Manual safety range: 20 %
- Dominance rule** (checked) or **P% rule** (unchecked)
 - number: 1, P: 10, N: 1
 - percentage: 75
- Minimum frequency** (checked) with value 3, **Min frequency range:** 10 %
- Missing = safe** (unchecked)
- Zero unsafe** (unchecked)
- Zero margin:** 10

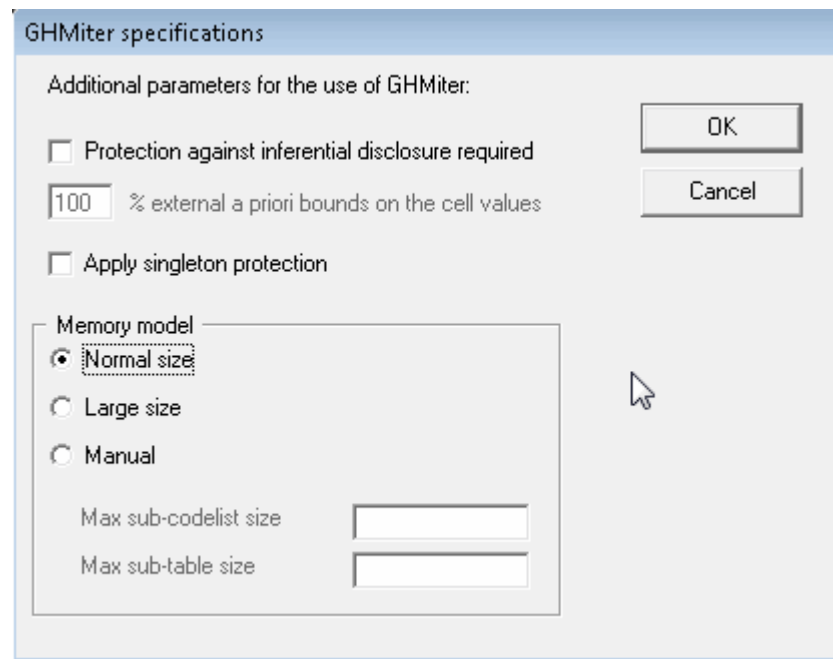
Ok

Proces nakładania tajemnicy statystycznej – pojedyncze pliki z danymi

- Tajemnica wtórna



następnie: Suppress ->
Hypercube



Proces nakładania tajemnicy statystycznej – pojedyncze pliki z danymi

- Tajemnica pierwotna i wtórna

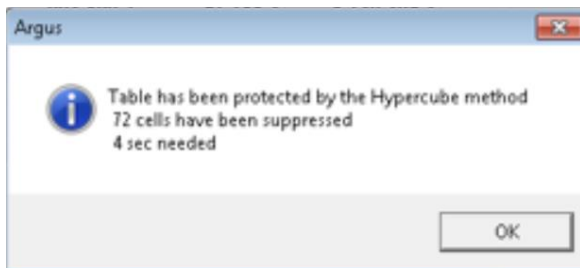


Table: WON x RO x RJ | Obr

RJ

TOTAL

	B2	B3	B4	B5
TOTAL	49 211 523,7	10 470 443,9	9 050 540,5	72 642 790,1
B2	660 793,1	2 009 497,0	93 644,0	1 000 570,1
B4	616 076,5	663 623,7	100 030,5	3 517 912,1
B6	492 640,0	010 043,1	57 052,1	2 659 240,1
B8	635 104,0	174 005,4	44 014,0	795 736,1
B10	730 740,0	297 600,5	59 940,4	4 001 329,1
B12	4 352 399,7	2 539 777,0	70 701,5	5 994 114,1
B14	21 400 402,2	4 003 120,0	721 037,9	21 005 053,1
B16	705 244,3	113 905,1	14 900,6	901 056,1
B18	3 015 001,0	506 904,0	100 230,5	1 136 231,1
B20	601 425,1	307 634,1	41 539,0	3 176 095,1
B22	10 463 992,2	204 374,3	141 790,0	3 090 296,1
B24	2 651 034,0	996 059,6	113 564,4	5 130 725,1
B26	152 525,2	252 529,4	10 020,2	1 005 652,1
B28	293 154,5	190 159,0	15 062,0	1 523 503,1
B30	1 133 505,4	5 105 604,5	0 231 911,0	12 325 358,1
B32	265 996,9	199 200,7	25 451,2	2 329 094,1

Cell Information

Value

663 427 164,7

Status

Safe

Cost

663 427 164,7

Shadow

663 427 164,7

B contributions

11,2017

Top n of shadow

10,000 007,1

Holding level

Request

0,0

Change status

Set to Safe

Set to Unsafe

Set to Protected

Set Cost

A good info

All Non-Struct/Empty

Suppress

HiddenCube

Module

Network

Optimal

Rounding

Suppress

Update Suppress

3-dig separator

Output View

Select Table

Change View

Table Summary

Write table

Done

Proces nakładania tajemnicy statystycznej – pojedyncze pliki z danymi

Zapis (write table)

Save Table

Format

☐ CSV-format

☐ CSV for pivot table

☒ Code-value

☐ SBS-format ☐ Add hierarchical levels

☐ Intermediate format ☐ Status only
☐ AddAuditResults
☐ UseHoldingsInfo

☐ JJ-format ☐ Remove trivial levels

General options:

☒ Add Status

☐ Suppress empty cells

☐ Var names on first line

☒ Embed spanning variables in quotes

Save Cancel

Podczas zapisu wynikowego pliku generowany jest także raport (plik html), który zawiera:

- informacje o nazwie pliku z danymi oraz metadanymi,
- strukturę zbioru danych,
- dokładny czas zapisania wyników,
- ustawienia parametrów przy jakich była nakładana tajemnica,
- podsumowanie agregatów w podziale na agregaty jawne, tajne, brakujące dane,
- kody dla każdego z wymiarów.

Proces nakładania tajemnicy statystycznej – przykład raportu

Var	Function	# codes
Response var:	Obr	
Explanatory var 1:	WON	17
Explanatory var 2:	RD	12
Explanatory var 3:	RJ	5
Shadow variable :	Obr	
Cost variable	Obr	

Safety Rule:

Dominance rule (Indiv. level) with $n = 1$ and $k = 75\%$
Minimum individual cell frequency: 3; safety margin: 10 %
Manual safety margin: 20%
Missing codes have been considered unsafe

GHMITER solution

GHMITER range ratio used: 0,000

GHMITER normal model used

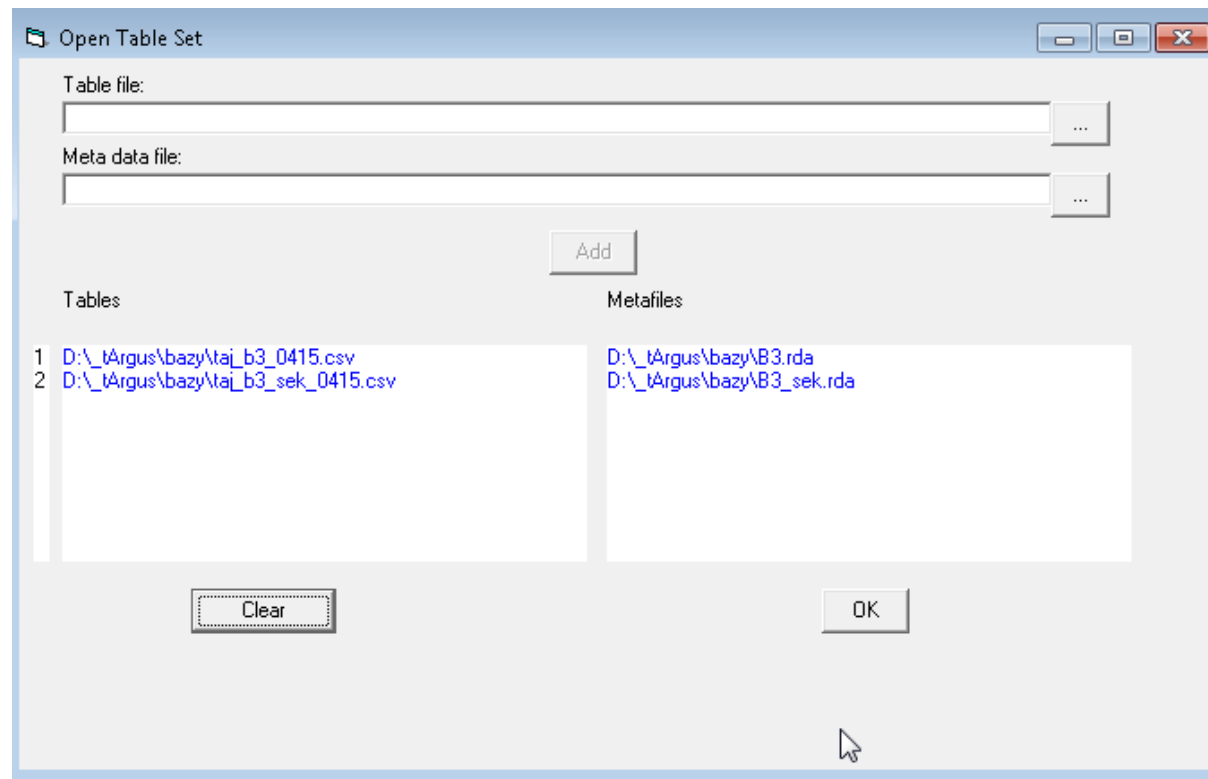
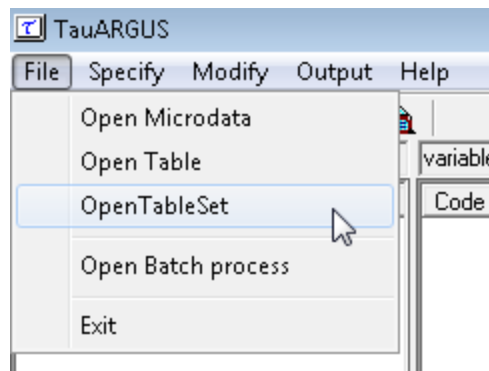
Singleton protection in GHMiter has not been applied

Time used to protect the table: 3 sec

Summary of the table

	Status	Number of cells	Number of respondents	Response value	Cost value
1	Safe	620	884404	5160150065,7	5160150065,7
2	Safe (manual)	0	0	0,0	0,0
3	Unsafe	37	2900	78465805,9	78465805,9

Proces nakładania tajemnicy statystycznej – tabele powiązane



Proces nakładania tajemnicy statystycznej – tabele powiązane

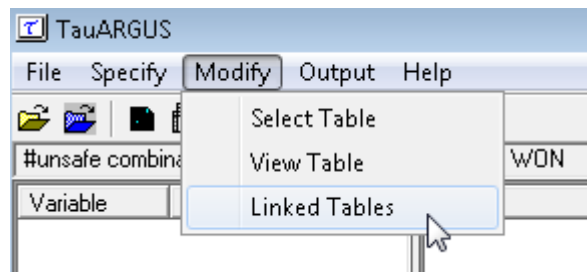
- Tajemnica pierwotna
(parametry dla obu
plików)

The 'Specify table' dialog box contains the following settings:

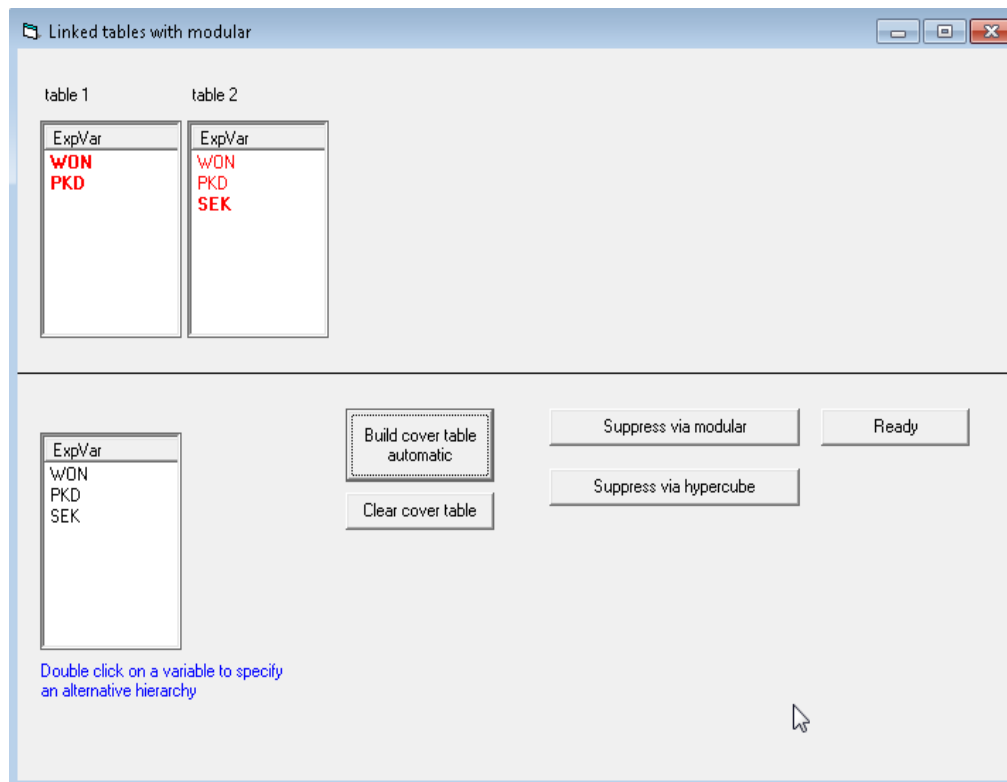
- Variables:**
 - Explanatory: WQON, PKD
 - Number: 2
- Status:**
 - Frequency
 - TopN: 1
- Cost:**
 - CostFunction for second. suppression:
 - ☒ ResponsVar
 - ☐ Frequency
 - ☐ Unity
 - ☐ Distance
 - Lambda: 1
- Do not allow non-additivity** (selected) ☐ Allow non-additivity
- ☐ Compute incorrect totals
- safety rule:**
 - ☐ Use given status
 - ☒ Use safety rules
- Dominance rule** (checked) ☐ P% rule
 - 1 number P 10 N 1
 - 75 percentage
- ☒ Minimum frequency 3 Min frequency range: 10 %
- ☐ Missing = safe
- ☐ Zero unsafe
- Zero margin 10

Ok

Proces nakładania tajemnicy statystycznej – tabele powiązane

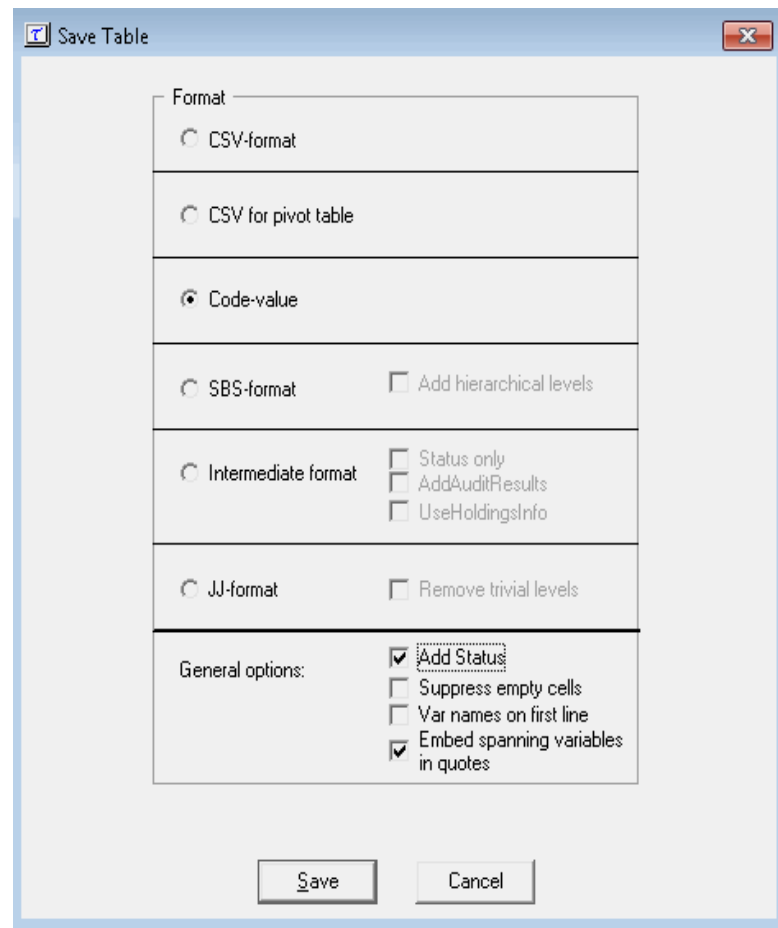
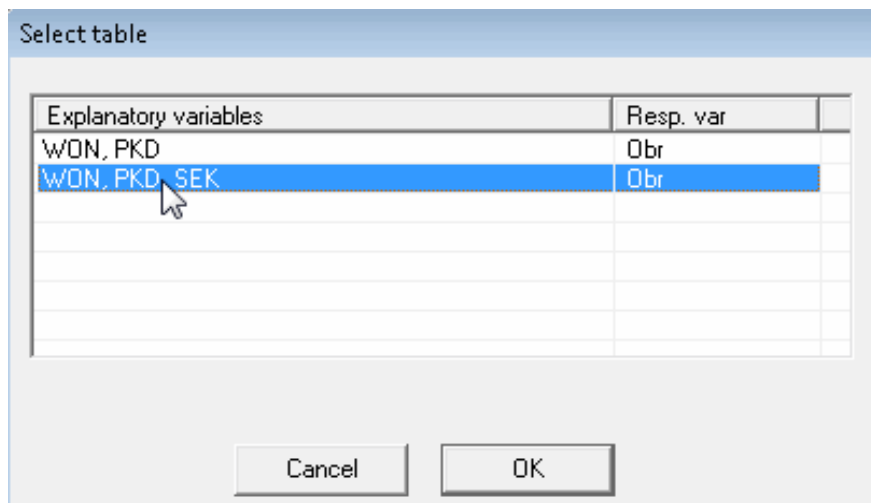
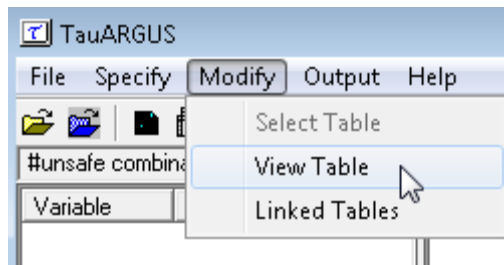


- build cover automatic
- suppress via hypercube



Proces nakładania tajemnicy statystycznej – tabele powiązane

- select table & write



Stosowane w programie statusy (flagi tajemnicy)

- 1 - Dane jawne
- 3 - Tajemnica pierwotna (dominacja)
- 5 - Tajemnica pierwotna (liczba jednostek < 3)
- 11 - Tajemnica wtórna
- 14 - Brak danych (agregat zawiera 0 jednostek)

Wnioski

- Nałożenie tajemnicy wtórnej jest dużo bardziej skomplikowane od tajemnicy pierwotnej – wykorzystanie programu τ -Argus upraszcza proces
- Program τ -Argus wykorzystujemy również do analiz tajemnicy statystycznej w innych przekrojach oraz badaniach
- Program oraz dokumentacja są dostępne: <https://research.cbs.nl/casc/tau.htm>

Literatura:

1. *Rozporządzenie Rady Ministrów z dnia 21 lipca 2015 r. w sprawie programu badań statystycznych statystyki publicznej na rok 2016 (Dz.U. z 4.IX.2015r. poz.1304).*
2. „τ-Argus 4.1 User's Manual”, Peter-Paul de Wolf (Modular), Anco Hundepool, Sarah Giessing (GHMITER, audit), Juan-José Salazar (Optimisation methods), Jordi Castro (Network solutions, CTA) Statistics Netherland, November 2014



Dziękuję za uwagę!