Perturbative Methods for Census 2021 tables

Presentation by Eric Schulte Nordholt
Geneva, September 2022

Statistics
Netherlands

Outline

- Census 2011 and confidentiality
- The project ‘Harmonised protection of census data in the ESS’
- Definition of test scenarios
- Tools for TRS and CKM
- Questions

Census 2011 and confidentiality (1)

European Census 2011 data represent an essential source of vital statistical information ranging from the lowest small-area geographical divisions to national and international levels

Harmonised census tables of 32 European countries are available via the Census Hub (https://ec.europa.eu/CensusHub2/)

Census data are detailed and confidential; protecting the census data is the responsibility of the member states

Census 2011 and confidentiality (2)

In spite of the output harmonisation international comparisons of census data are hampered by different statistical disclosure control approaches

Two Specific Grant Agreements (SGAs) to define and test best practices for the Census 2021:

- SGA on Harmonised protection of census data in the ESS (2016-2017)
- SGA on Perturbative confidentiality methods (2018-2019)

The project ‘Harmonised protection of census data in the ESS’ (1)

- Start: 1 September 2016
- End: 31 August 2017

- Four WPs:
  - WP 1 Management (7 deliverables)
  - WP 2 Questionnaire (2 deliverables)
  - WP 3 Development and testing of the recommendations; identification of best practices (4 deliverables)
  - WP 4 Dissemination (5 deliverables)

The project ‘Harmonised protection of census data in the ESS’ (2)

– Six countries involved:
  - CBS (Eric Schulte Nordholt, Peter-Paul de Wolf),
  - INSEE (Maël-Luc Buron),
  - Destatis (Sarah Gießing, Tobias Enderle),
  - HCSO (László Antal, Beata Nagy),
  - Statistics Finland (Annu Cabrera) and
  - SURS (Andreja Smukavec, Junoš Lukan)

Insee
Mesurer pour comprendre

DESTATIS
wissen. nutzen.

KSH

Statistics Finland

REPUBLIC OF SLOVENIA
STATISTICAL OFFICE RS

The project ‘Harmonised protection of census data in the ESS’ (3)

- Reviewed the country specific data protection regulations and methods
- Provided a harmonised approach to the protection of the Census 2021 (taking the national constraints into account)
- Recommended to Member States appropriate Statistical Disclosure Control methods for hypercubes
- Recommended how to handle efficiently confidential cells in grid squares and regional breakdowns (risk of disclosure due to differencing)

Definition of test scenarios (1)

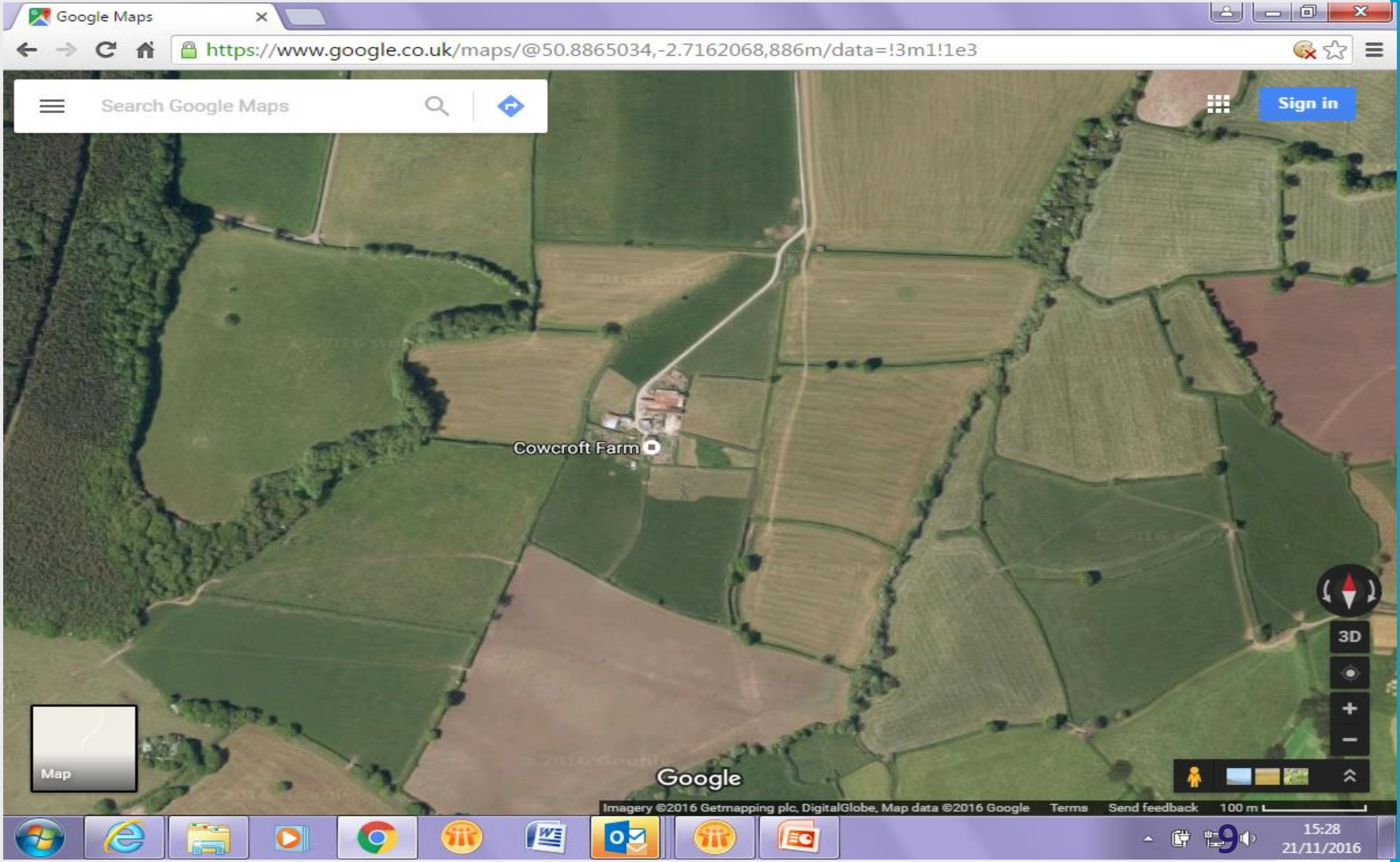
Restrictions:

- No table redesign or global recodes (lay-out of hypercubes fixed in implementing European census regulation)
- No cell suppressions (very difficult for linked high dimensional hypercubes and otherwise no European total can be calculated)

Complications:

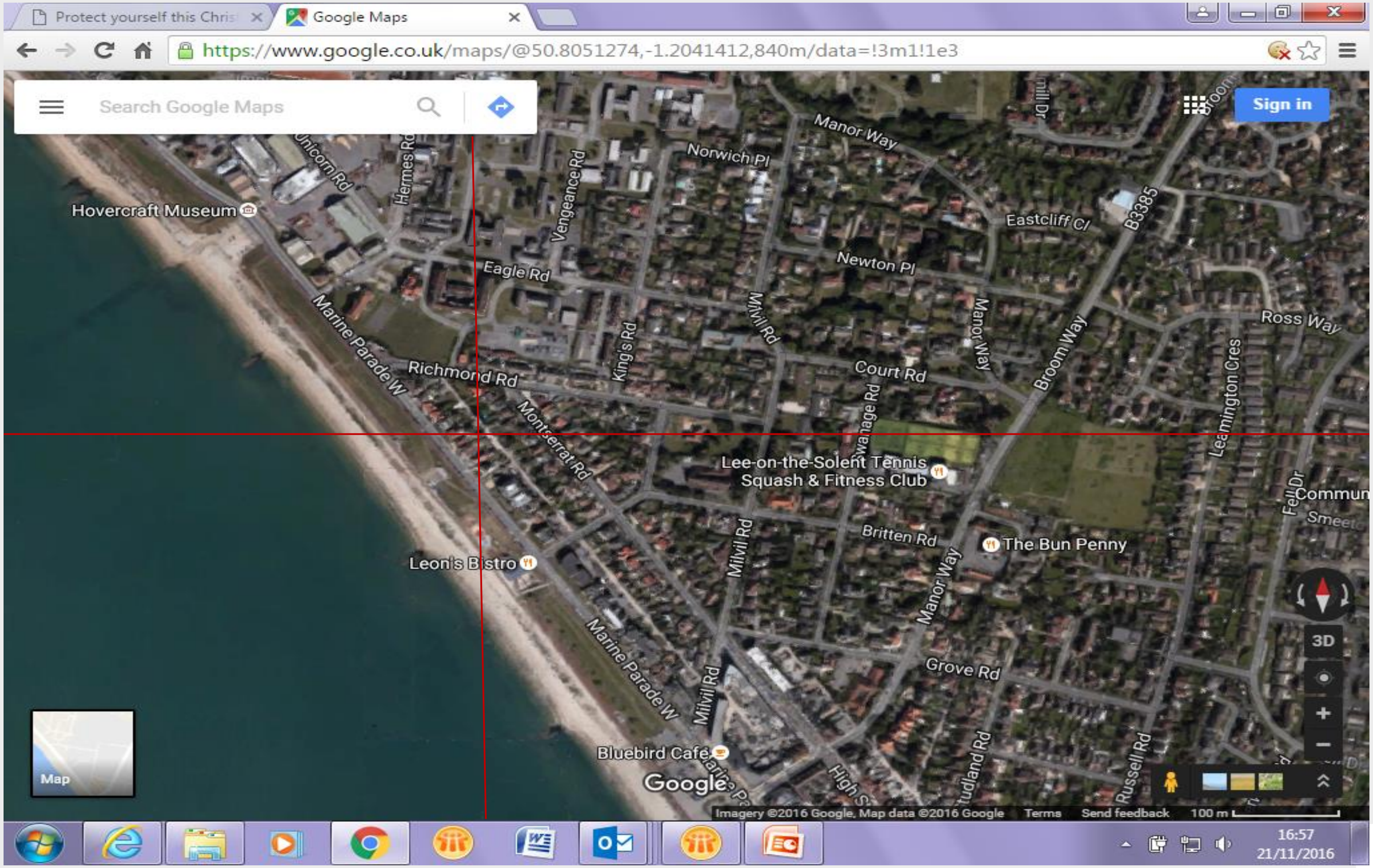
- 1 km² grid cells lead to many small cell values
- 1 km² grid cells ⇔ administrative regions (risk of disclosure due to differencing)

Definition of test scenarios (2)

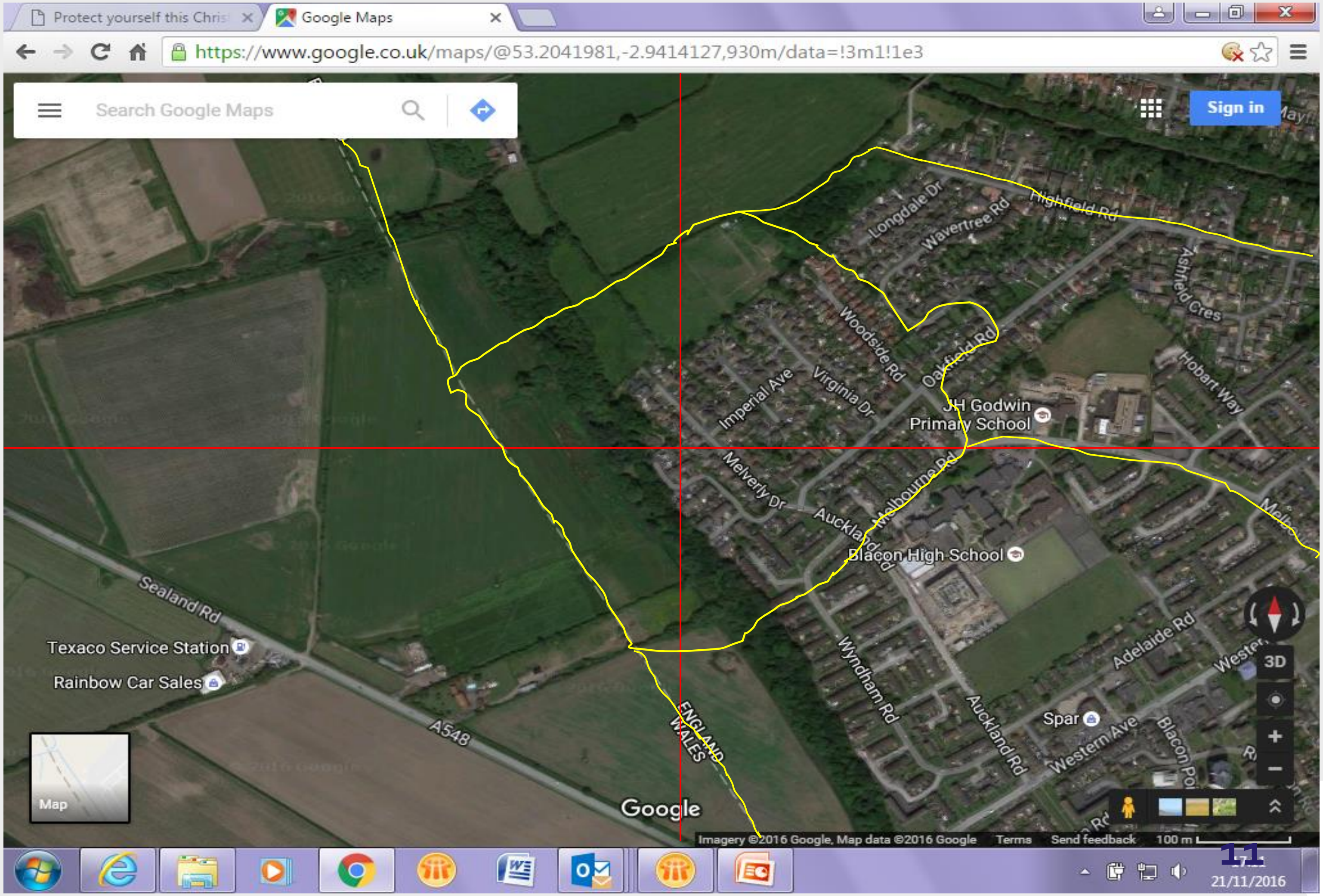


Google Maps
https://www.google.co.uk/maps/@50.8865034,-2.7162068,886m/data=!3m1!1e3
Search Google Maps
Sign in
Cowcroft Farm
Map
Google
3D
Imagery ©2016 Getmapping plc, DigitalGlobe, Map data ©2016 Google
Terms
Send feedback
100 m
15:28
21/11/2016

Definition of test scenarios (3)



Definition of test scenarios (4)



Definition of test scenarios (5)

The Statistical Disclosure Control solution should not alter the spatial distribution of the grid data too much:

- Zero frequencies grids should not too often be changed to positive frequencies
- Rare non-zero frequencies in an area should not be changed much

Usual disclosure risks:

- Small counts (may lead to direct identification)
- Attribute disclosure (a positive frequency may lead to disclosing information from a hypercube)

Definition of test scenarios (6)

Flexible method that can be adapted to national needs by the member states:

- Pre-tabular method of record swapping
- Post-tabular method of cell key method

Advantage: (estimates for) all cells become available

Disadvantages: relative error for small cell values may be large and loss of additivity in tables

Definition of test scenarios (7)

Record swapping and cell key method:

- Enhanced variant of cell key method developed by the Australian Bureau of Statistics (ABS)
- Provided by the Office for National Statistics (ONS) and adapted in this project

Tools for TRS and CKM (1)

- New Specific Grant Agreement (SGA) started in spring 2018: SGA on ‘Open source tools for perturbative confidentiality methods’ and run for 15 months (till autumn 2019)
- Aim of this new SGA: integrate the codes produced into user-friendly open source software packages
- Seven countries involved in this SGA: Austria, Finland, France, Germany, Hungary, Netherlands and Slovenia
- Software to test on census data (Targeted Record Swapping and the Cell Key Method) is available on github (https://github.com/sdcTools/CensusProtection)

Tools for TRS and CKM (2)

Targeted Record Swapping is a pre-tabular method (changes in microdata)

Preparation:

- **Specify variables that define risk (*k*-anonymity)**
- **Specify variables that define regional hierarchy**
- **Calculate risk for all households at each regional level**
- **Specify variables that define “similar” households**
- **Specify minimum swap rate**

Tools for TRS and CKM (3)

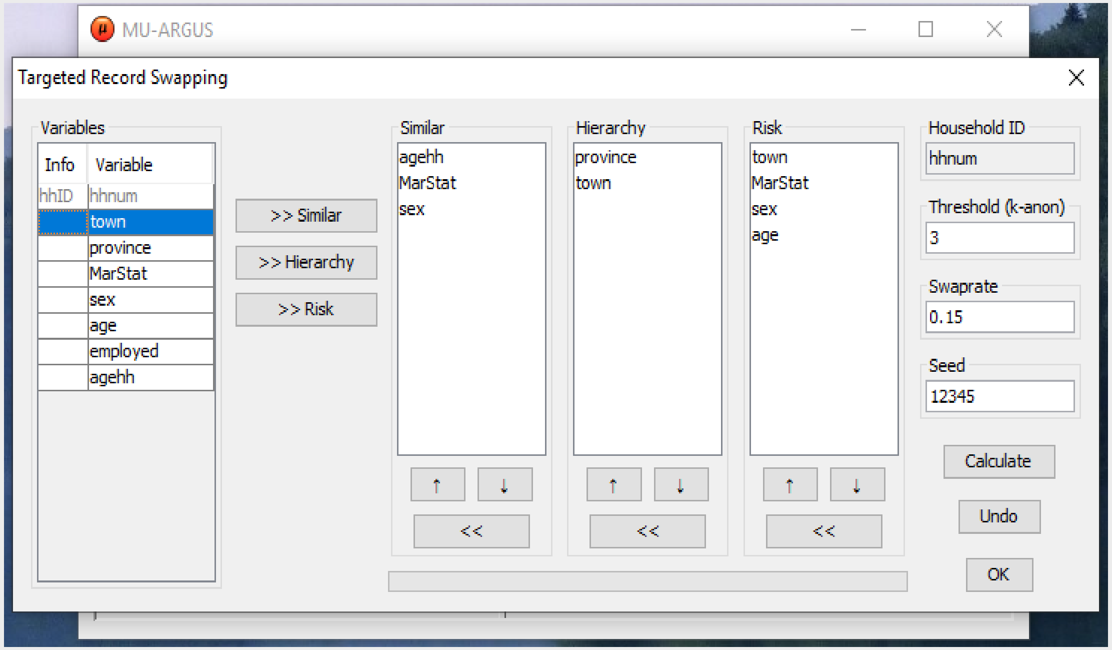
For implementation of TRS go from highest to lowest regional level :

- **Make donor-set of households**
  - "Similar" households as the high risk households
- **Draw a donor household for a high risk household**
  - Same regional *level*, different *region*
  - Swap all regional variables
- **If minimum swap rate is not reached, swap additional households at lowest regional level**

Tools for TRS and CKM (4)

C++ code

Callable from μ-Argus



Targeted Record Swapping

Variables

Info	Variable
hhID	hhnum
	town
	province
	MarStat
	sex
	age
	employed
	agehh

Similar: agehh, MarStat, sex

Hierarchy: province, town

Risk: town, MarStat, sex, age

Household ID: hhnum

Threshold (k-anon): 3

Swaprate: 0.15

Seed: 12345

Tools for TRS and CKM (5)

Cell key method is a post tabular method (noise added to table cells)

1. **Determine *p*-table**
2. **Draw $\mathcal{U}(0,1)$ value for each record = *record key***
3. **Sum record keys of records in each table cell and assign fractional part of that sum as *cell key* to each table cell**
4. **Use *cell value* AND *cell key* AND *p*-table to get amount of noise to add to that cell**

1. Determine *p*-table

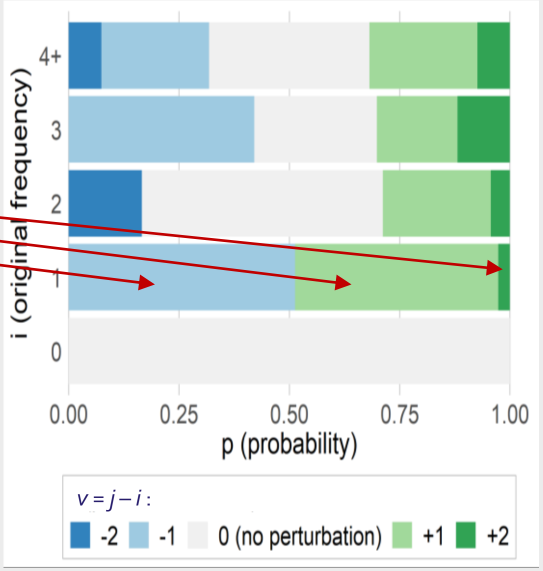
p-table: transition probabilities (R-package `ptable`)

p_{ij} = P(cell value *i* is changed into value *j*)

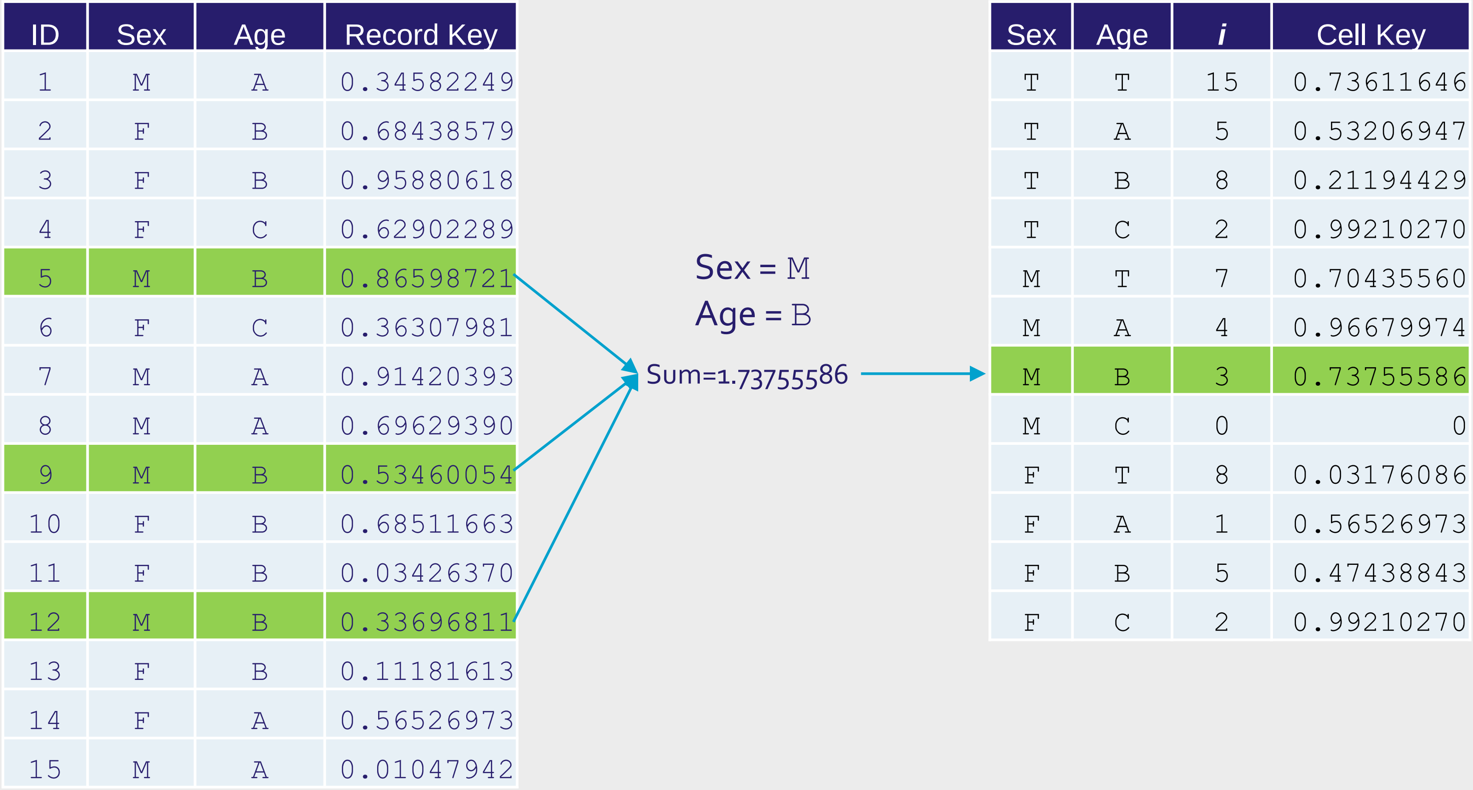
E.g., probabilities

i \ *j*	0	1	2	3	4	5	6
0	1	0	0	0	0	0	0
1	0.5133	0	0.4600	0.0267	0	0	0
2	0.1656	0	0.5463	0.2449	0.0432	0	0
3	0	0	0.4208	0.2776	0.1824	0.1192	0
4	0	0	0.0739	0.2442	0.3637	0.2442	0.0739

Cumulative



2.–3. Draw record keys and make cell keys



ID	Sex	Age	Record Key
1	M	A	0.34582249
2	F	B	0.68438579
3	F	B	0.95880618
4	F	C	0.62902289
5	M	B	0.86598721
6	F	C	0.36307981
7	M	A	0.91420393
8	M	A	0.69629390
9	M	B	0.53460054
10	F	B	0.68511663
11	F	B	0.03426370
12	M	B	0.33696811
13	F	B	0.11181613
14	F	A	0.56526973
15	M	A	0.01047942

Sex = M

Age = B

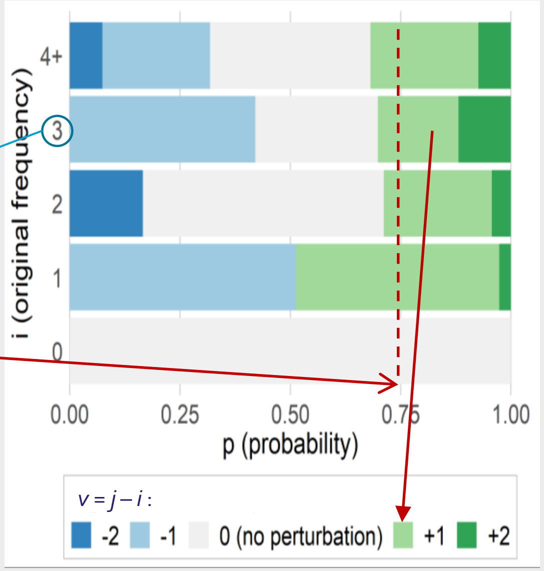
Sum=1.73755586

Sex	Age	*i*	Cell Key
T	T	15	0.73611646
T	A	5	0.53206947
T	B	8	0.21194429
T	C	2	0.99210270
M	T	7	0.70435560
M	A	4	0.96679974
M	B	3	0.73755586
M	C	0	0
F	T	8	0.03176086
F	A	1	0.56526973
F	B	5	0.47438843
F	C	2	0.99210270

4. Determine amount of noise to add

Sex	Age	*i*	Cell Key
T	T	15	0.73611646
T	A	5	0.53206947
T	B	8	0.21194429
T	C	2	0.99210270
M	T	7	0.70435560
M	A	4	0.96679974
M	B	3	0.73755586
M	C	0	0
F	T	8	0.03176086
F	A	1	0.56526973
F	B	5	0.47438843
F	C	2	0.99210270

Cumulative

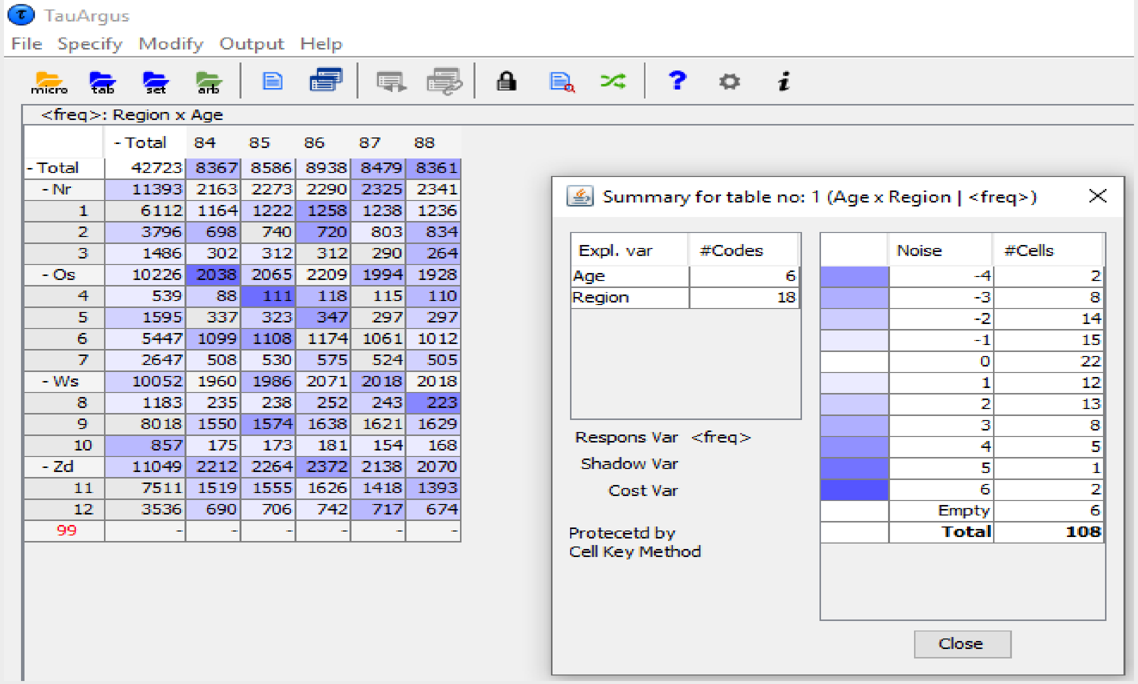


(M, B): *i* = 3

(M, B): *j* = *i* + 1 = 4

Tools for TRS and CKM (6)

Adding noise via CKM is part of τ-Argus and also callable from R



<freq>: Region x Age

	- Total	84	85	86	87	88
- Total	42723	8367	8586	8938	8479	8361
- Nr	11393	2163	2273	2290	2325	2341
1	6112	1164	1222	1258	1238	1236
2	3796	698	740	720	803	834
3	1486	302	312	312	290	264
- Os	10226	2038	2065	2209	1994	1928
4	539	88	111	118	115	110
5	1595	337	323	347	297	297
6	5447	1099	1108	1174	1061	1012
7	2647	508	530	575	524	505
- Ws	10052	1960	1986	2071	2018	2018
8	1183	235	238	252	243	223
9	8018	1550	1574	1638	1621	1629
10	857	175	173	181	154	168
- Zd	11049	2212	2264	2372	2138	2070
11	7511	1519	1555	1626	1418	1393
12	3536	690	706	742	717	674
99	-	-	-	-	-	-

Summary for table no: 1 (Age x Region | <freq>)

Expl. var	#Codes
Age	6
Region	18

Respons Var <freq>
Shadow Var
Cost Var

Protecetd by
Cell Key Method

	Noise	#Cells
▨	-4	2
▨	-3	8
▨	-2	14
▨	-1	15
	0	22
▨	1	12
▨	2	13
▨	3	8
▨	4	5
▨	5	1
▨	6	2
	Empty	6
	Total	**108**

Tools for TRS and CKM (7)

Testing has been done on datasets for TRS and large Census 2011 hypercubes for CKM with acceptable runtimes

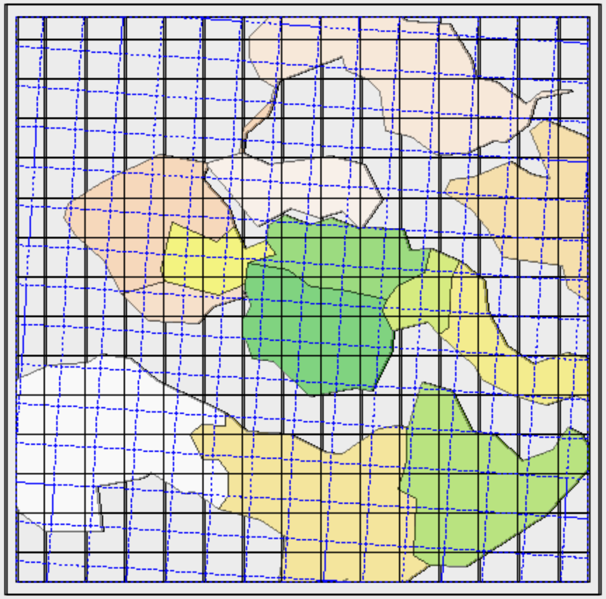
Help is available via github:
https://github.com/sdcTools/UserSupport/wiki

Tools for TRS and CKM (8)

Further activities:

- Risk and utility measures have been proposed
- Research for appropriate parameter values is going on in different countries
- Methods are known, software is available, tests are going on
- Communication of results will be important!
- Next ESTP training course on 25-27 January 2023 in Luxembourg

Illustration of different grids



Questions

Do you have any questions or remarks?