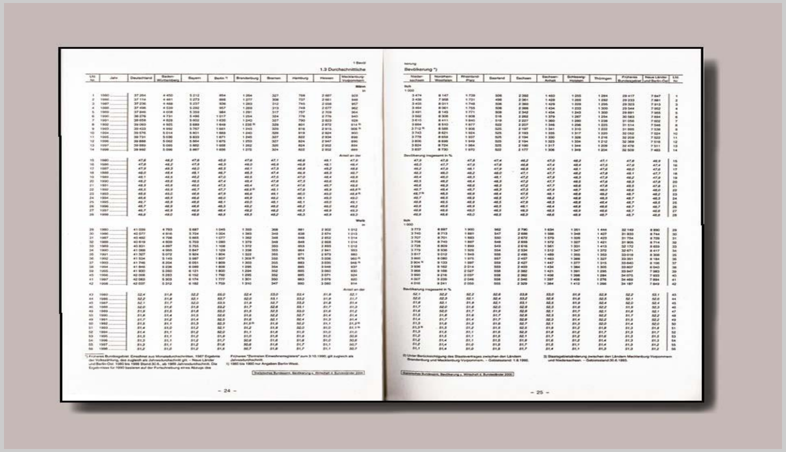
Testing CTA as Additivity Module



for Perturbed Census 2021 EU Hypercube Data

Tobias ENDERLE, Sarah GIESSING
Federal Statistical Office of Germany
Division Mathematical Statistical Methods

Background

EU-Project „Harmonized Protection of Census Data in the ESS“

- **recommended to test / use as SDC method for EU Census hypercubes and Grid Data**
  - **Targeted Record Swapping and/or**
  - **Random noise based on the Cell key method (Fraser and Wooton, ABS, 2005)**

Background

Properties of Cell key method noise:

- Statistical properties determined by noise design
  - Noise Variance V (constant, identical for all counts)
  - Noise Mean = 0 for each count („unbiased“)
  - Fixed Maximum Perturbation D
  - No small counts (1‘s, 2‘s) in noisy data
- Noise exactly consistent between hypercubes / grids
- Noise „drawn“ independently, based on Cell Keys

Background

Hypercube Data after CKM noise not additive

- **Viz.: Males + Females ≠ Total**

Naive approach

- **(hierarchical) aggregation of „interior“ cells**

Problem

- **Naive approach damages design properties of the noise:**
  - **Increase of maximum perturbation (like: from 3 to > 300)**
  - **Increase of noise variance (like: from 0,5 to 130)**
- **Large (absolute) inconsistencies between hypercubes**

„Calibrating“ noise might do better?

- **Iterative Methods (f.i. Iterative Proportional Fitting), or**
- **Optimization based method**
  - **Controlled Tabular Adjustment (CTA)**

Controlled Tabular Adjustment (CTA, Castro et al.)

Find vector of adjusted counts x_i subject to sets of constraints:

- (1) „Adjusted hypercube is additive“
- (2) Box constraints on deviations between „original“ (y_i) and adjusted (x_i) entries of the vector of counts
- (3) Protection levels for „sensitive“ cells ($i \in P$)
  - Upper protecion $x_i \geq y_i + upl_i$ **or**
  - Lower protection $x_i \leq y_i - lpl_i$

CTA Optimization problem in term of deviations $z_i := x_i - y_i$:

$$\min_z \sum_{i=1}^{n} w_i \, | z_i |$$

$$s.t. \quad (1)\ Az = 0$$

$$(2)\ l_i \leq z_i \leq u_i \quad , i = 1, \ldots, n$$

$$(3)\ z_i \leq lpl_i \text{ or } z_i \geq upl_i, \ i \in P$$

Scenario I: Test CTA to restore additivity to EU-Census hypercube 9.2 after perturbation by CKM noise

EU-Census hypercube 9.2 (one of ca. 150 EU-Census hypercubes)

- 4-dimensional, hierarchical hypercube presenting frequency counts by
  - geo_m 42 categories („synthetic" part of NW Germany)
  - sex 2 categories
  - age_m 21 categories
  - yae_h 24 categories
- Number of cells: 133,560
- Number of „additivity constraints": 129,822

Test hypercube 9.2: GEO.M x AGE.M x YAE.H x SEX

GEO.M

Format Ansicht ?

```
@01
@@010
@@@01051
@@@01053
@@@01054
@@@01055
@@@01056
@@@01057
@@@01058
@@@01059
@@@01060
@@@01061
@@@01062
@02
@@020
@@@02000
@03
@@031
@@@03151
@@@03152
@@@03153
@@@03154
@@@03155
@@@03156
@@@03157
@@@03158
@@032
@@@03251
@@@03252
@@@03254
@@@03255
@@@03256
@@@03257
@@033
@@@03351
@@@03352
@@@03353
@@@03354
@@@03355
@@@03356
@@@03357
@@@03358
@@@03359
@@@03360
@@@03361
@@034
@@@03451
@@@03452
@@@03453
@@@03454
@@@03455
@@@03456
```

AGE.M

Format Ansicht ?

```
@1.
@@1.1.
@@1.2.
@@1.3.
@2.
@@2.1.
@@2.2.
@@2.3.
@3.
@@3.1.
@@3.2.
@@3.3.
@@3.4.
@4.
@@4.1.
@@4.2.
@@4.3.
@5.
@@5.1.
@@5.2.
@@5.3.
@@5.4.
@6.
@@6.1.
@@6.2.
@@6.3.
@@6.4.
```

YAE.H

Format Ansicht ?

```
@1.
@@1.1.
@@@1.1.1.
@@@1.1.2.
@@1.2.
@@@1.2.1.
@@@1.2.2.
@@@1.2.3.
@@@1.2.4.
@@@1.2.5.
@@1.3.
@@@1.3.1.
@@@1.3.2.
@@@1.3.3.
@@@1.3.4.
@@@1.3.5.
@@1.4.
@@@1.4.1.
@@@1.4.2.
@@@1.4.3.
@@@1.4.4.
@@@1.4.5.
@@1.5.
@@1.6.
@@1.7.
@@1.8.
@@1.9.
@2.
@3.
```

SEX

Format Ansicht ?

```
@1
@2
```

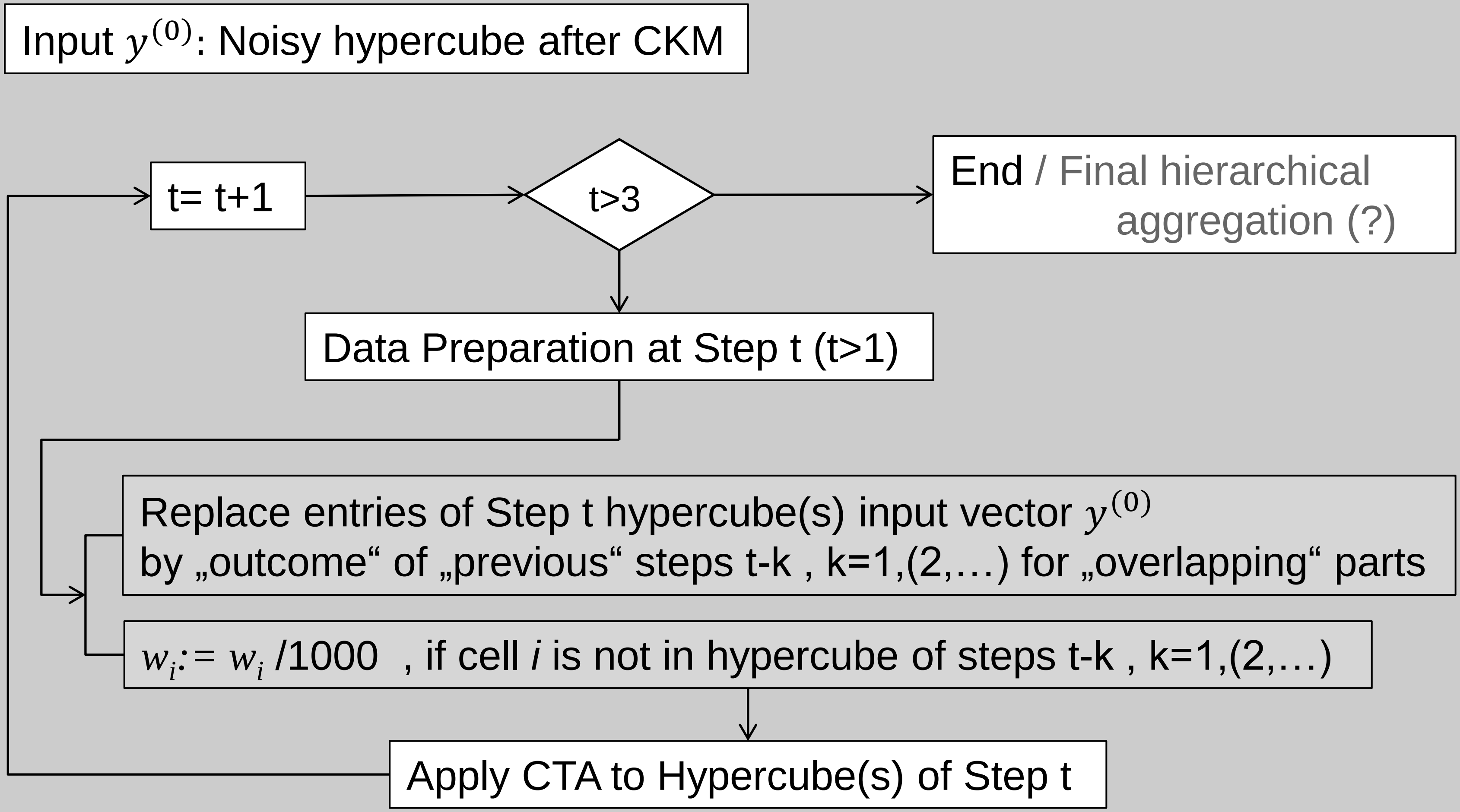
Settings of scenario I

- Cell key method noise
  - D=3 and V=1

- No sensitive cells ($P = \emptyset$)
- Weights: $w_i = (y_i)^{-\frac{1}{2}}$
- Box constraints: $l_i = -10$ und $u_i = +10$
- Commercial LP/MILP Solver

- Two approaches:
  - Solve in a single step
    - Not successfull (?)
    - Execution stopped (manually) after 1 week
  - Stepwise procedure
    - Blocking Strategy

Stepwise blocking strategy – 3 steps



Blocking strategy – Step 1

GEO.M

```
@01
@@010
@@@01051
@@@01053
@@@01054
@@@01055
@@@01056
@@@01057
@@@01058
@@@01059
@@@01060
@@@01061
@@@01062
@02
@@020
@@@02000
@03
@@031
@@@03151
@@@03152
@@@03153
@@@03154
@@@03155
@@@03156
@@@03157
@@@03158
@@032
@@@03251
@@@03252
@@@03254
@@@03255
@@@03256
@@@03257
@@033
@@@03351
@@@03352
@@@03353
@@@03354
@@@03355
@@@03356
@@@03357
@@@03358
@@@03359
@@@03360
@@@03361
@@034
@@@03451
@@@03452
@@@03453
@@@03454
@@@03455
@@@03456
```

AGE.M

```
@1.
@@1.1.
@@1.2.
@@1.3.
@2.
@@2.1.
@@2.2.
@@2.3.
@3.
@@3.1.
@@3.2.
@@3.3.
@@3.4.
@4.
@@4.1.
@@4.2.
@@4.3.
@5.
@@5.1.
@@5.2.
@@5.3.
@@5.4.
@6.
@@6.1.
@@6.2.
@@6.3.
@@6.4.
```

YAE.H

```
@1.
@@1.1.
@@@1.1.1.
@@@1.1.2.
@@1.2.
@@@1.2.1.
@@@1.2.2.
@@@1.2.3.
@@@1.2.4.
@@@1.2.5.
@@1.3.
@@@1.3.1.
@@@1.3.2.
@@@1.3.3.
@@@1.3.4.
@@@1.3.5.
@@1.4.
@@@1.4.1.
@@@1.4.2.
@@@1.4.3.
@@@1.4.4.
@@@1.4.5.
@@1.5.
@@1.6.
@@1.7.
@@1.8.
@@1.9.
@2.
@3.
```

SEX

```
@1
@2
```

Step 1
„CTA for hypercube of 3-way margins“

Blocking strategy – Step 1

GEO.M

Format Ansicht ?

```
@01
@@010
@@@01051
@@@01053
@@@01054
@@@01055
@@@01056
@@@01057
@@@01058
@@@01059
@@@01060
@@@01061
@@@01062
@02
@@020
@@@02000
@03
@@031
@@@03151
@@@03152
@@@03153
@@@03154
@@@03155
@@@03156
@@@03157
@@@03158
@@032
@@@03251
@@@03252
@@@03254
@@@03255
@@@03256
@@@03257
@@033
@@@03351
@@@03352
@@@03353
@@@03354
@@@03355
@@@03356
@@@03357
@@@03358
@@@03359
@@@03360
@@@03361
@@034
@@@03451
@@@03452
@@@03453
@@@03454
@@@03455
@@@03456
```

AGE.M

Format Ansicht ?

```
@1.
@@1.1.
@@1.2.
@@1.3.
@2.
@@2.1.
@@2.2.
@@2.3.
@3.
@@3.1.
@@3.2.
@@3.3.
@@3.4.
@4.
@@4.1.
@@4.2.
@@4.3.
@5.
@@5.1.
@@5.2.
@@5.3.
@@5.4.
@6.
@@6.1.
@@6.2.
@@6.3.
@@6.4.
```

YA H

Format Ansicht ?

```
@1.
@@1.1.
@@@1.1.1.
@@@1.1.2.
@@1.2.
@@@1.2.1.
@@@1.2.2.
@@@1.2.3.
@@@1.2.4.
@@@1.2.5.
@@1.3.
@@@1.3.1.
@@@1.3.2.
@@@1.3.3.
@@@1.3.4.
@@@1.3.5.
@@1.4.
@@@1.4.1.
@@@1.4.2.
@@@1.4.3.
@@@1.4.4.
@@@1.4.5.
@@1.5.
@@1.6.
@@1.7.
@@1.8.
@@1.9.
@2.
@3.
```

SEX

Format Ansicht ?

```
@1
@2
```

Step 1
„CTA for hypercube of 3-way margins“

X Skip variable

Blocking strategy – Step 2

GEO.M

Format Ansicht ?

```
@01
@@010
@@@01051
@@@01053
@@@01054
@@@01055
@@@01056
@@@01057
@@@01058
@@@01059
@@@01060
@@@01061
@@@01062
@02
@@020
@@@02000
@03
@@031
@@@03151
@@@03152
@@@03153
@@@03154
@@@03155
@@@03156
@@@03157
@@@03158
@@032
@@@03251
@@@03252
@@@03254
@@@03255
@@@03256
@@@03257
@@033
@@@03351
@@@03352
@@@03353
@@@03354
@@@03355
@@@03356
@@@03357
@@@03358
@@@03359
@@@03360
@@@03361
@@034
@@@03451
@@@03452
@@@03453
@@@03454
@@@03455
@@@03456
```

AGE.M

Format Ansicht ?

```
@1.
@@1.1.
@@1.2.
@@1.3.
@2.
@@2.1.
@@2.2.
@@2.3.
@3.
@@3.1.
@@3.2.
@@3.3.
@@3.4.
@4.
@@4.1.
@@4.2.
@@4.3.
@5.
@@5.1.
@@5.2.
@@5.3.
@@5.4.
@6.
@@6.1.
@@6.2.
@@6.3.
@@6.4.
```

YAE.H

Format Ansicht ?

```
@1.
@@1.1.
@@@1.1.1.
@@@1.1.2.
@@1.2.
@@@1.2.1.
@@@1.2.2.
@@@1.2.3.
@@@1.2.4.
@@@1.2.5.
@@1.3.
@@@1.3.1.
@@@1.3.2.
@@@1.3.3.
@@@1.3.4.
@@@1.3.5.
@@1.4.
@@@1.4.1.
@@@1.4.2.
@@@1.4.3.
@@@1.4.4.
@@@1.4.5.
@@1.5.
@@1.6.
@@1.7.
@@1.8.
@@1.9.
@2.
@3.
```

SEX

Format Ansicht ?

```
@1
@2
```

Step 2
„CTA for hypercube of 4-way GEO.L margins“

Blocking strategy – Step 2

GEO.M

```
@01
@@010
@@@01051
@@@01053
@@@01054
@@@01055
@@@01056
@@@01057
@@@01058
@@@01059
@@@01060
@@@01061
@@@01062
@02
@@020
@@@02000
@03
@@031
@@@03151
@@@03152
@@@03153
@@@03154
@@@03155
@@@03156
@@@03157
@@@03158
@@032
@@@03251
@@@03252
@@@03254
@@@03255
@@@03256
@@@03257
@@033
@@@03351
@@@03352
@@@03353
@@@03354
@@@03355
@@@03356
@@@03357
@@@03358
@@@03359
@@@03360
@@@03361
@@034
@@@03451
@@@03452
@@@03453
@@@03454
@@@03455
@@@03456
```

AGE.M

```
@1.
@@1.1.
@@1.2.
@@1.3.
@2.
@@2.1.
@@2.2.
@@2.3.
@3.
@@3.1.
@@3.2.
@@3.3.
@@3.4.
@4.
@@4.1.
@@4.2.
@@4.3.
@5.
@@5.1.
@@5.2.
@@5.3.
@@5.4.
@6.
@@6.1.
@@6.2.
@@6.3.
@@6.4.
```

YAE.H

```
@1.
@@1.1.
@@@1.1.1.
@@@1.1.2.
@@1.2.
@@@1.2.1.
@@@1.2.2.
@@@1.2.3.
@@@1.2.4.
@@@1.2.5.
@@1.3.
@@@1.3.1.
@@@1.3.2.
@@@1.3.3.
@@@1.3.4.
@@@1.3.5.
@@1.4.
@@@1.4.1.
@@@1.4.2.
@@@1.4.3.
@@@1.4.4.
@@@1.4.5.
@@1.5.
@@1.6.
@@1.7.
@@1.8.
@@1.9.
@2.
@3.
```

SEX

```
@1
@2
```

Step 2
„CTA for hypercube of 4-way GEO.L margins“

- GEO.L margins

Blocking strategy – Step 3

GEO.M

Format Ansicht ?

```
@01
@@010
@@@01051
@@@01053
@@@01054
@@@01055
@@@01056
@@@01057
@@@01058
@@@01059
@@@01060
@@@01061
@@@01062
@02
@@020
@@@02000
@03
@@031
@@@03151
@@@03152
@@@03153
@@@03154
@@@03155
@@@03156
@@@03157
@@@03158
@@032
@@@03251
@@@03252
@@@03254
@@@03255
@@@03256
@@@03257
@@033
@@@03351
@@@03352
@@@03353
@@@03354
@@@03355
@@@03356
@@@03357
@@@03358
@@@03359
@@@03360
@@@03361
@@034
@@@03451
@@@03452
@@@03453
@@@03454
@@@03455
@@@03456
```

AGE.M

Format Ansicht ?

```
@1.
@@1.1.
@@1.2.
@@1.3.
@2.
@@2.1.
@@2.2.
@@2.3.
@3.
@@3.1.
@@3.2.
@@3.3.
@@3.4.
@4.
@@4.1.
@@4.2.
@@4.3.
@5.
@@5.1.
@@5.2.
@@5.3.
@@5.4.
@6.
@@6.1.
@@6.2.
@@6.3.
@@6.4.
```

YAE.H

Format Ansicht ?

```
@1.
@@1.1.
@@@1.1.1.
@@@1.1.2.
@@1.2.
@@@1.2.1.
@@@1.2.2.
@@@1.2.3.
@@@1.2.4.
@@@1.2.5.
@@1.3.
@@@1.3.1.
@@@1.3.2.
@@@1.3.3.
@@@1.3.4.
@@@1.3.5.
@@1.4.
@@@1.4.1.
@@@1.4.2.
@@@1.4.3.
@@@1.4.4.
@@@1.4.5.
@@1.5.
@@1.6.
@@1.7.
@@1.8.
@@1.9.
@2.
@3.
```

SEX

Format Ansicht ?

```
@1
@2
```

Step 3
„CTA for six 4-way hypercubes“

Blocking strategy – Step 3

GEO.M

Format Ansicht ?

```
@01
@@010
@@@01051
@@@01053
@@@01054
@@@01055
@@@01056
@@@01057
@@@01058
@@@01059
@@@01060
@@@01061
@@@01062
@02
@@020
@@@02000
@03
@@031
@@@03151
@@@03152
@@@03153
@@@03154
@@@03155
@@@03156
@@@03157
@@@03158
@@032
@@@03251
@@@03252
@@@03254
@@@03255
@@@03256
@@@03257
@@033
@@@03351
@@@03352
@@@03353
@@@03354
@@@03355
@@@03356
@@@03357
@@@03358
@@@03359
@@@03360
@@@03361
@@034
@@@03451
@@@03452
@@@03453
@@@03454
@@@03455
@@@03456
```

AGE.M

Format Ansicht ?

```
@1.
@@1.1.
@@1.2.
@@1.3.
@2.
@@2.1.
@@2.2.
@@2.3.
@3.
@@3.1.
@@3.2.
@@3.3.
@@3.4.
@4.
@@4.1.
@@4.2.
@@4.3.
@5.
@@5.1.
@@5.2.
@@5.3.
@@5.4.
@6.
@@6.1.
@@6.2.
@@6.3.
@@6.4.
```

YAE.H

Format Ansicht ?

```
@1.
@@1.1.
@@@1.1.1.
@@@1.1.2.
@@1.2.
@@@1.2.1.
@@@1.2.2.
@@@1.2.3.
@@@1.2.4.
@@@1.2.5.
@@1.3.
@@@1.3.1.
@@@1.3.2.
@@@1.3.3.
@@@1.3.4.
@@@1.3.5.
@@1.4.
@@@1.4.1.
@@@1.4.2.
@@@1.4.3.
@@@1.4.4.
@@@1.4.5.
@@1.5.
@@1.6.
@@1.7.
@@1.8.
@@1.9.
@2.
@3.
```

SEX

Format Ansicht ?

```
@1
@2
```

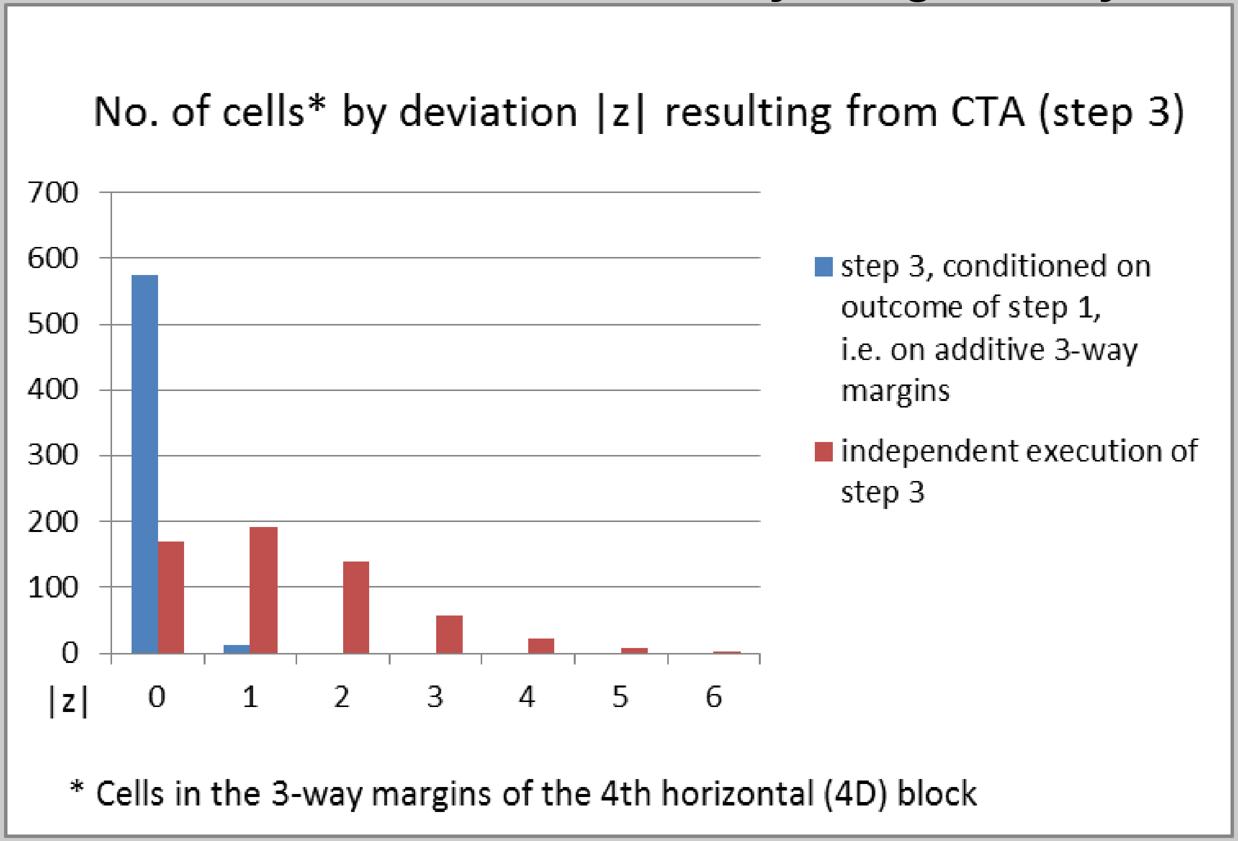
Step 3
„CTA for six
4-way hypercubes“

6 separate GEO blocks

Results for blocking approch in test scenario I

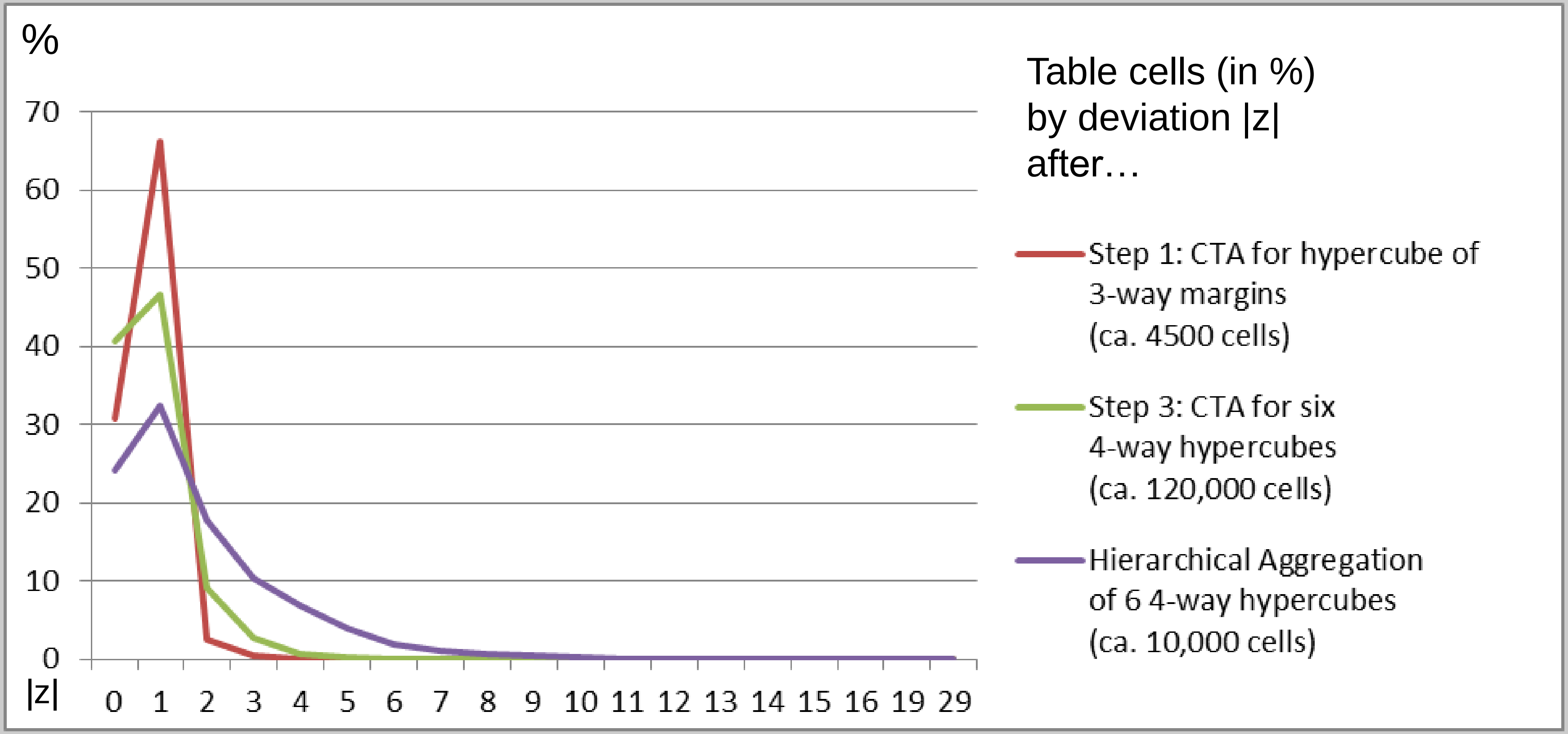
Step 1 (CTA on 3-way margin hypercube) preserves between-hypercube consistency very well

- **Example: CTA deviations after step 3, for block 4**
  - **Results for 3-way margins only**



Results for blocking approch in test scenario I

- „Smaller hypercube“ (Step 1) => „Smaller perturbation“
- Changes after CTA (Step 3) much smaller than after (final) hierarchical aggregation
  - Including Step 2 (CTA of Geo.L 4-way hypercube, not yet tried) might help



Scenario II - complex case: regard ones and twoes as sensitive cells

Find vector of adjusted counts x_i subject to sets of constraints:

- (1) „Adjusted hypercube is additive“
- (2) Box constraints on deviations between „original“ (y_i) and adjusted (x_i) entries of the vector of counts
- **(3) Protection levels for „sensitive“ cells ($i \in P$)**
  - **Upper protecion** $x_i \geq y_i + upl_i$ **or**
  - **Lower protection** $x_i \leq y_i - lpl_i$

CTA Optimization problem in term of deviations $z_i := x_i - y_i$:

$$\min_z \sum_{i=1}^{n} w_i \, | z_i |$$

$$s.t. \quad Az = 0$$

$$l_i \leq z_i \leq u_i \quad , i = 1, \ldots, n$$

$$z_i \leq -lpl_i \text{ or } z_i \geq upl_i, \quad i \in P$$

Scenario II: Restore additivity and don‘t allow for sensitive cells

EU-Census sub-hypercube 9.2

- 4-dimensional, hierarchical hypercube presenting frequency counts by
  - geo_m 42 categories („synthetic“ part of NW Germany)
  - sex 2 categories
  - age_m 21 categories
  - yae **3** categories
- Number of cells: **16,800**
- Number of „additivity constraints“: **16,352**
- Number of bottom level cells: 5,292

Iterative strategy for the complex case

Feasible solution: entries need to be zero or above a given threshold N (i.e. N=2)

Iterative strategy consists of three phases:

- Initial Phase
  - Starting with a suitable non-additive random noise solution
  - Run CTA instance as applied in the „simple case“ (scenario I)
- Second Phase (iteratively)
  - Define sensitive cells $P = \{i_1, i_2, \dots, i_p\}$ and protection levels
  - Run CTA instance using protection levels, modified boundaries (especially for non-sensitive bottom level cells) and weights
- Final Phase
  - Stop second phase if …
    - … feasible solution is available OR
    - … no further reduction of sensistive cells is possible
  - In case of remaining sensitive cells:
    deterministic rounding and/or aggegation of bottom level cells

Results of the iterative strategy

Cells	Before CTA	Initial Phase	Second Phase – Iteration …						Final Phase*
			1	2	3	4	5	5*	
0	1,474	1,488	1,552	1,564	1,570	1,570	1,570	1,075	1,076
1	0	61	16	10	4	5	4	1	0
2	0	106	31	14	6	2	2	1	0
> 2	15,326	15,145	15,201	15,212	15,220	15,223	15,224	4,215	4,216
All	16,800	16,800	16,800	16,800	16,800	16,800	16,800	5,292	5,292
CPU time		56 min	7 min	6 s	6 s	3 s	3 s		

* bottom level cells

Initial/Second Phase:

- Modified lower boundaries for non-sensitive cells: $\min\left(10; \max(y_i - 3; 0)\right)$
  - in order to prevent these cells to be changed to sensitive cells
  - ”Optimal CTA table found (optimal within tolerances)”

Results of the iterative strategy

Cells	Before CTA	Initial Phase	Second Phase – Iteration …						Final Phase*
			1	2	3	4	5	5*	
0	1,474	1,488	1,552	1,564	1,570	1,570	1,570	1,075	1,076
1	0	61	16	10	4	5	4	1	0
2	0	106	31	14	6	2	2	1	0
> 2	15,326	15,145	15,201	15,212	15,220	15,223	15,224	4,215	4,216
All	16,800	16,800	16,800	16,800	16,800	16,800	16,800	5,292	5,292
CPU time		56 min	7 min	6 s	6 s	3 s	3 s		

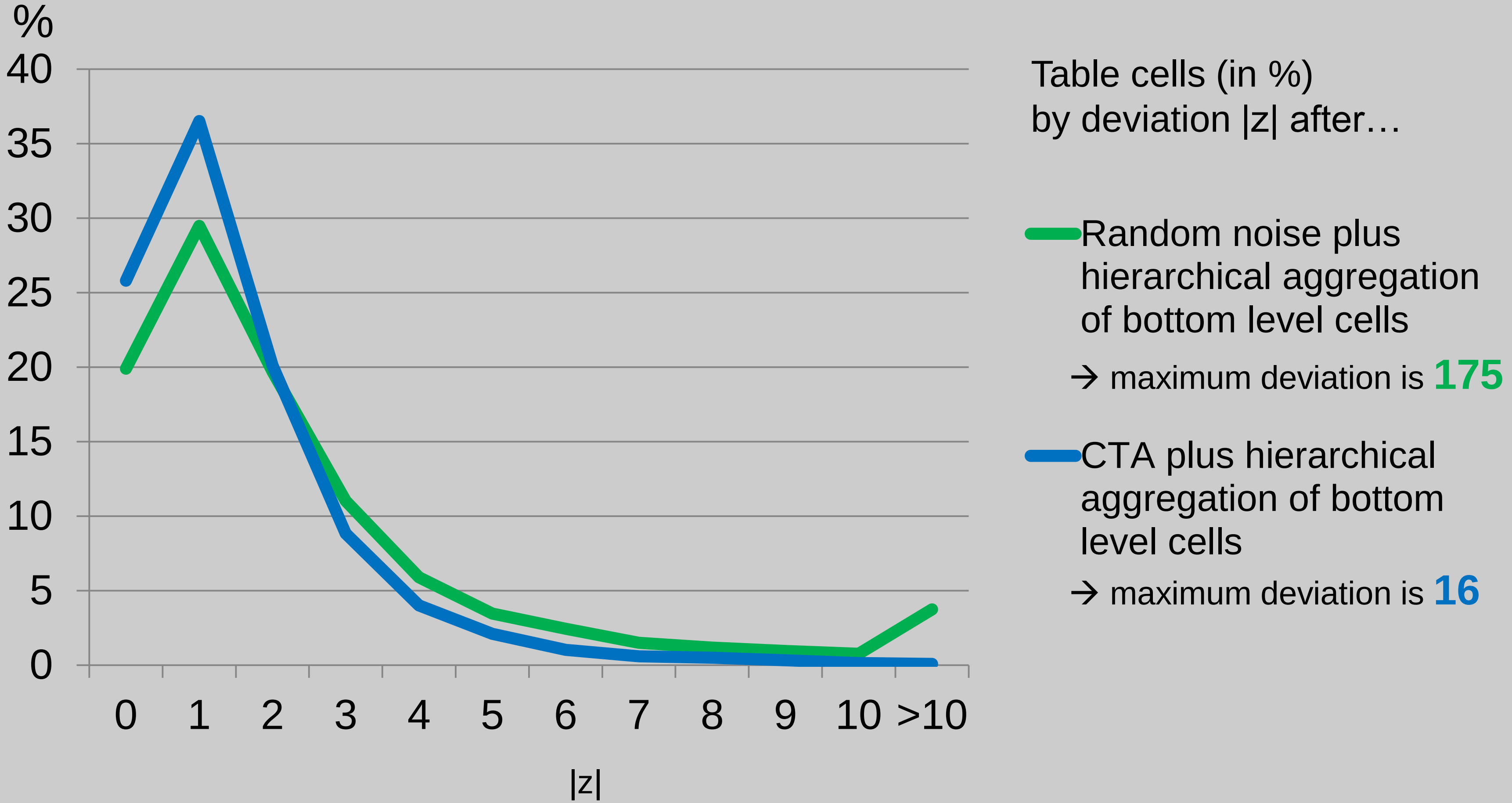
reduction of sensitive cells

* bottom level cells

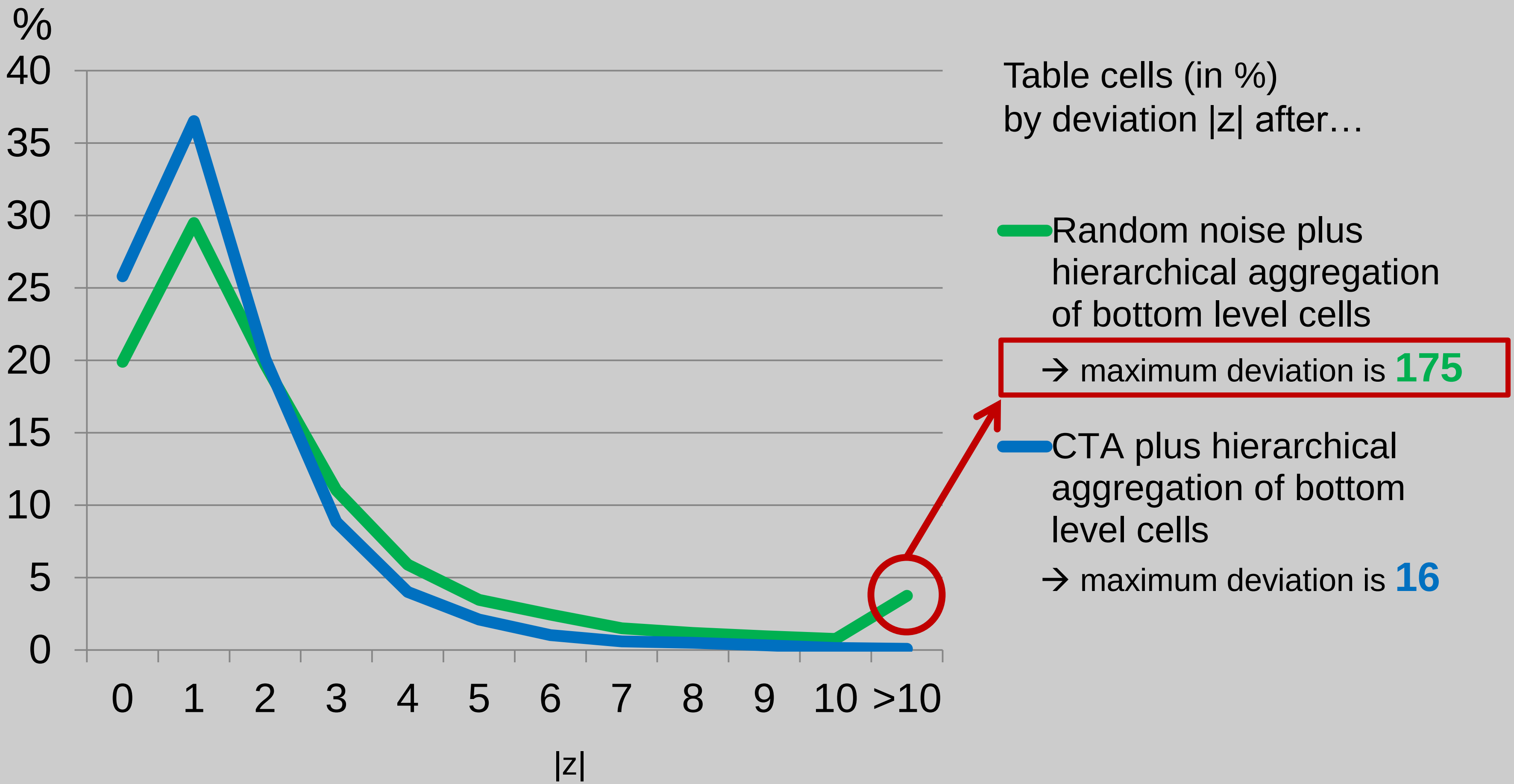
Initial/Second Phase:

- Modified lower boundaries for non-sensitive cells: $\min\left(10; \max(y_i - 3; 0)\right)$
  - in order to prevent these cells to be changed to sensitive cells
  - ”Optimal CTA table found (optimal within tolerances)”

Comparison: iterative CTA strategy vs. simple aggregation



Comparison: iterative CTA strategy vs. simple aggregation



Summary and Conclusion

Summary:

- Starting point: Non-additive but consistent hypercubes protected by random noise with Cell Key Method design properties
- Evaluations using synthetic EU hypercube data
- Restoring additivity at once too complex, therefore …
  - Blocking strategies and iterative approaches make CTA optimization problems feasible
  - CTA techniques lead to much less perturbation in hypercube margins than simple summation of the noisy lowest level counts

Conclusions:

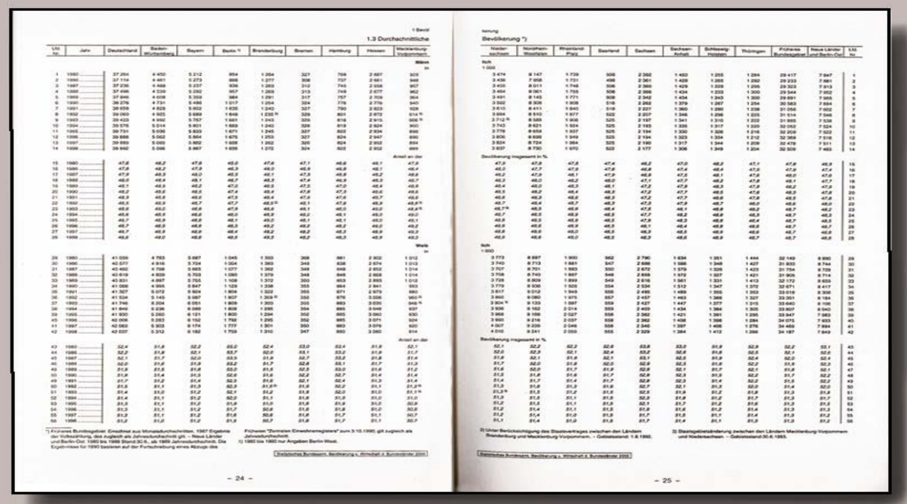
- Recommendation: to preserve consistency and keep perturbation “low” the “goals” of additivity “corrections” should be revised:
  - either not restore additivity at all, or
  - restore additivity to sub-hypercubes rather than to full hypercubes (i.e. devise heuristical strategies)

Acknowledgements

The research leading to these results has partially received funding from the EU project “Harmonized Protection of Census Data in the ESS” (Specific grant agreement n° 11111.2016.005-2016.367 under the Framework partnership agreement n° 11112.2014.005-2014.533).

References

- **Castro, J., Gonzalez, J.A., Baena, D. and Jimenez, X. (2013):** User's and programmer's manual of the RCTA package (v.2), Technical Report DR 2013-06, available from http://www-eio.upc.es/~jcastro
- **Enderle, T. and Gießing, S. (working paper, 2017):** Testing CTA as Additivity Module for Perturbed Census 2021 EU Hypercube Data; Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Skopje 2017.
- **Fraser, B., Wooton, J. (2006):** A proposed method for confidentialising tabular output to protect against differencing, in Monographs of Official Statistics. Work session on Statistical Data Confidentiality, Eurostat-Office for Official Publications of the European Com-munities, Luxembourg, 2006, S. 299-302.
- **Gießing, S. (2014):** Pre-tabular Perturbation with Controlled Tabular Adjustment: Some Considerations; Privacy in Statistical Databases, Balearen 2014.
- **Gießing, S. (2016):** Computational issues in the design of transition probabilities and disclosure risk estimation for additive noise; Privacy in Statistical Databases, Dubrovnik 2016.


1.3 Durchschnittliche
Bevölkerung *)
– 24 –
– 25 –

Thanks for your attention