Sensitivity analysis of disclosure control measures

EFGS conference, Sofia, 2013

- David Martin, Klaus Steinnocher, Ekkehard Petri

- Acknowledgements: Mario Köstl, Tanja Tötzer

AIT AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

eurostat

UNIVERSITY OF Southampton

Outline

- Previous presentation introduced „GEOSTAT 2011 – A population grid for Europe“
- This presentation describes related work investigating the potential disclosiveness of including social characteristics in the GEOSTAT grid
- Overview of datasets and potential risks
- Development of simulated disaggregated grid dataset
  - Reference data
  - Adjustment methodology
- Results: evaluation of alternative disclosure thresholds
- Conclusions

Overview of datasets and potential risks

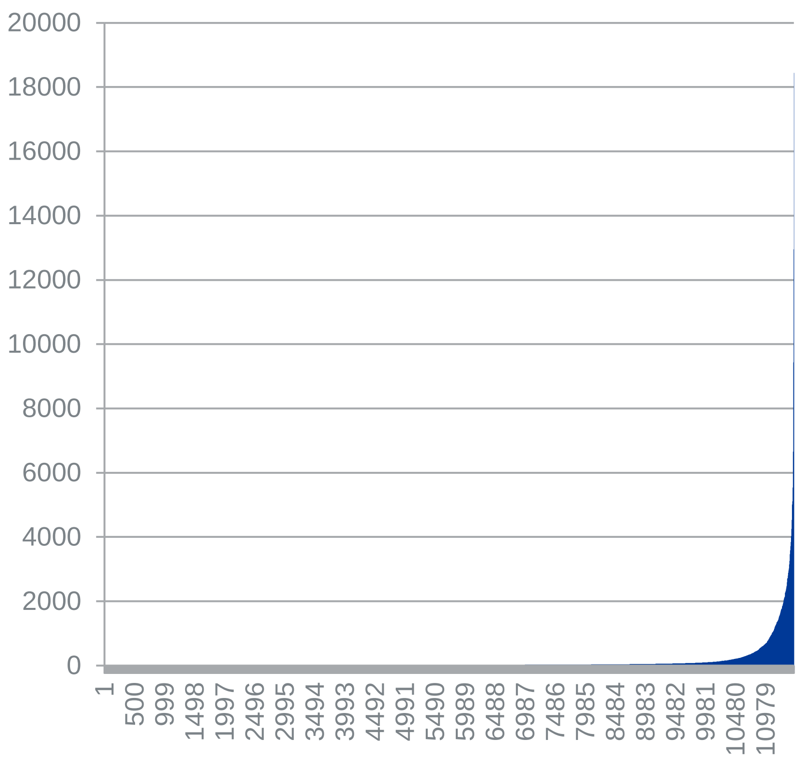
- Small population counts in grid cells present the risk of inadvertent disclosure of data about identifiable individuals
- The more unusual the socioeconomic characteristics, the greater the risk of disclosure and the more that might be learned by an “intruder”
- European NSIs adopt different confidentiality thresholds to reduce risk
- If a grid of socioeconomic characteristics were to be produced, what would be the impact of different thresholds on the utility of the data?
- Variables selected for sensitivity analysis:
  - pop > 65
  - male * pop > 65
  - women * employed
  - women * employed * in area

Development of a simulated disaggregated grid dataset

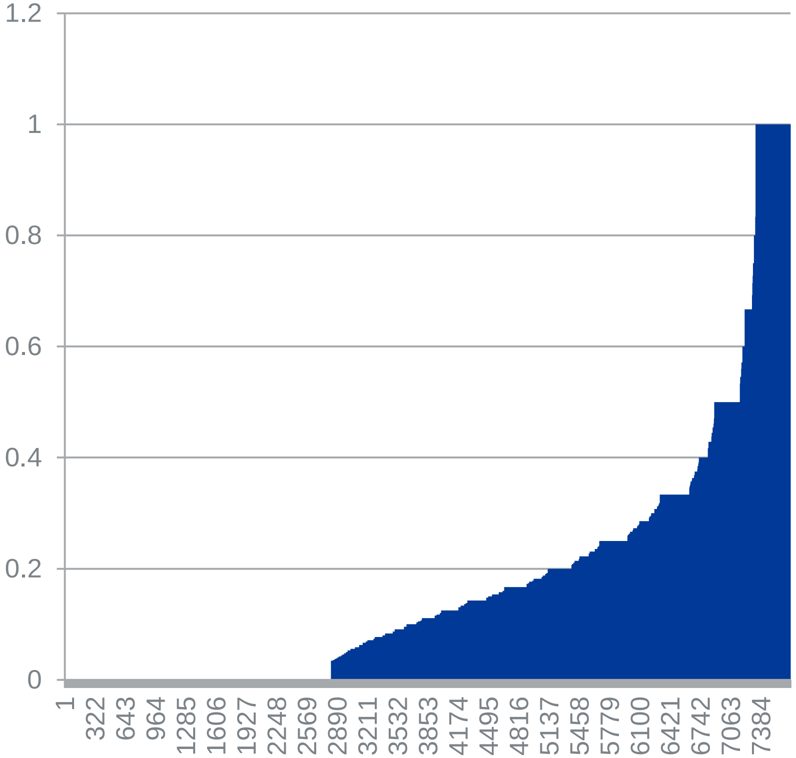
- Have total population per cell from GEOSTAT2006 grid
- Have social characteristics for whole grid area at LAU2 level
- Do not have social characteristics at cell level
- Linear disaggregation would simply assign LAU2 mean values of each variable to each cell
- Need a method to adjust these initial cell values to generate a more plausible statistical (and spatial) distribution
- Use reference distributions for appropriate variables from countries where cell or small area data are available
- Reference areas: Two urban and rural NUTS2 areas in each of Norway (NO01 Oslo-Akershus, NO02 Hedmark-Oppland) and Austria (AT13 Wien, AT31 Oberösterreich); All Output Areas in England

Example: Norway reference data (11455 cells with non-zero population)

Distribution of total population



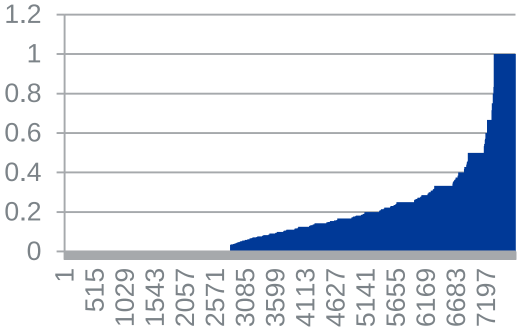
Distribution of proportion who are males aged over 68 in cells with population below 30



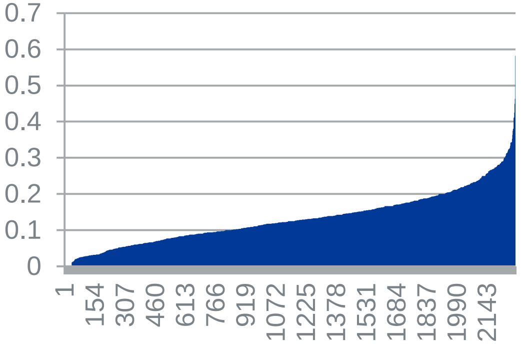
Population over 68/65

- Norway 68+

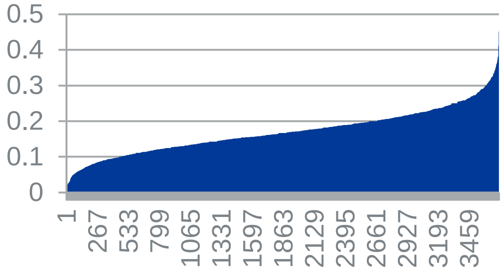
Cell Pop < 30



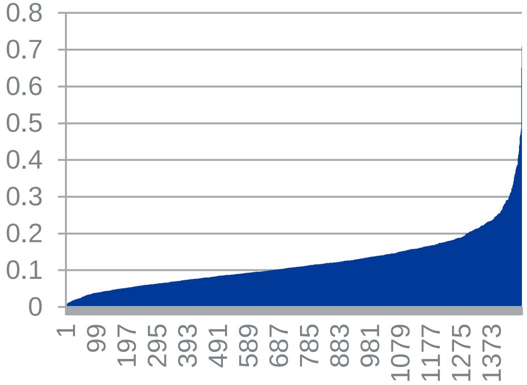
Cell Pop 30-100

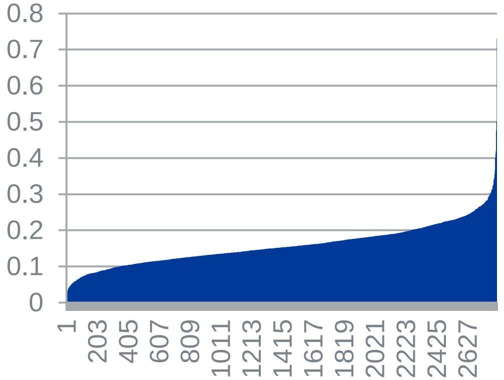


- Austria 65+

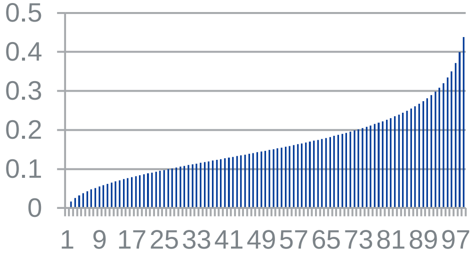


Cell/OA Pop 100+





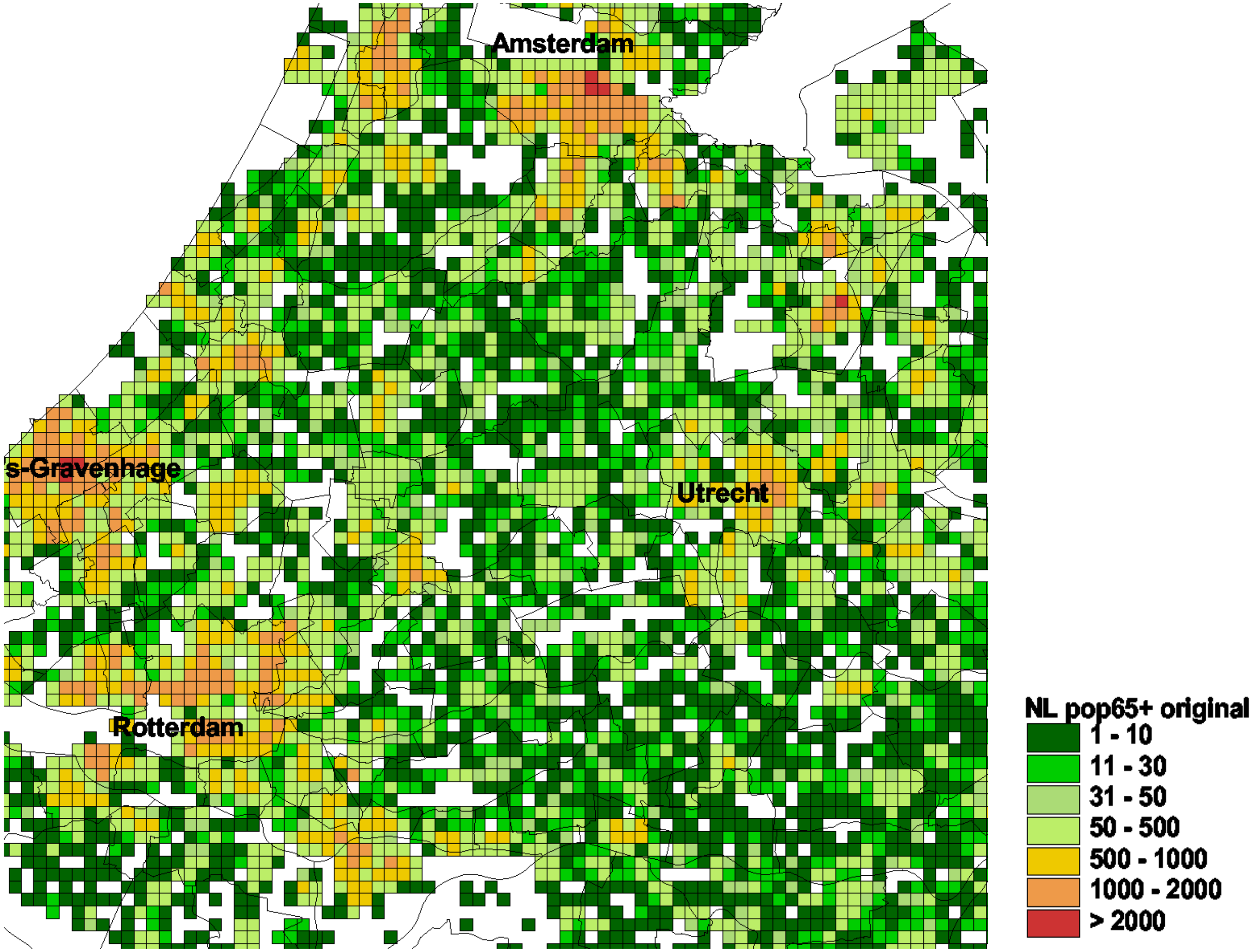
- England 65+ (centiles)



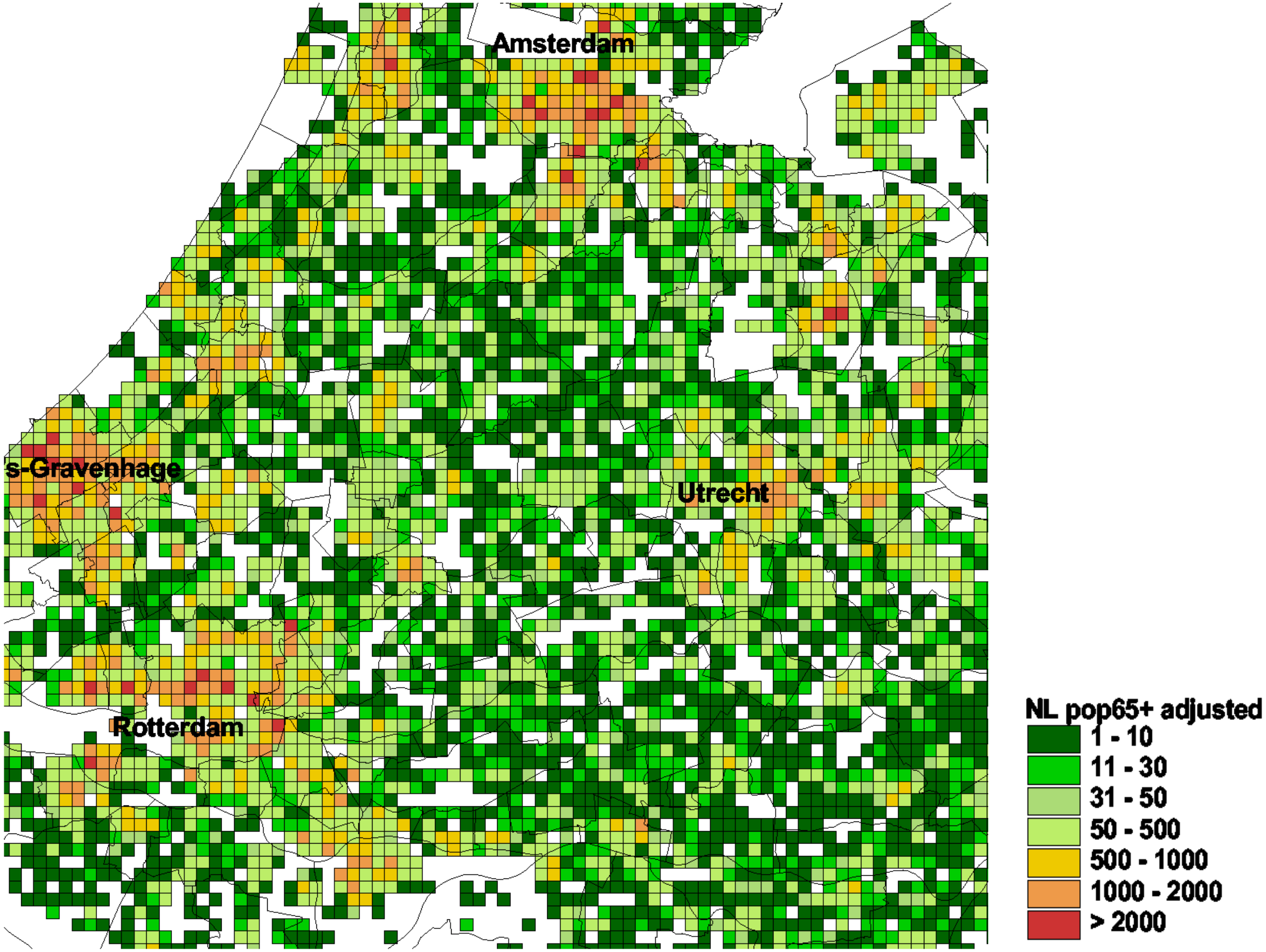
Adjustment methodology

- Identify best-matching reference distribution for each variable in each population size range: <30, 30-100, 100+
- Initially assign LAU2 mean proportion to every cell in that LAU2 for each social characteristic
- Draw random offsets from the best-matching reference distribution and adjust the initial values *in terms of offsets from the mean value*
  - (In the long run, the adjusted distribution would reflect the shape of the reference distribution, but preserving local mean)
- Skip very small LAU2s and very small populations (no balanced adjustment possible)
- Re-scale adjusted values to preserve total counts within each LAU2
- Use this dataset to assess the effect of different confidentiality thresholds
- (Methodology implemented in VBA)

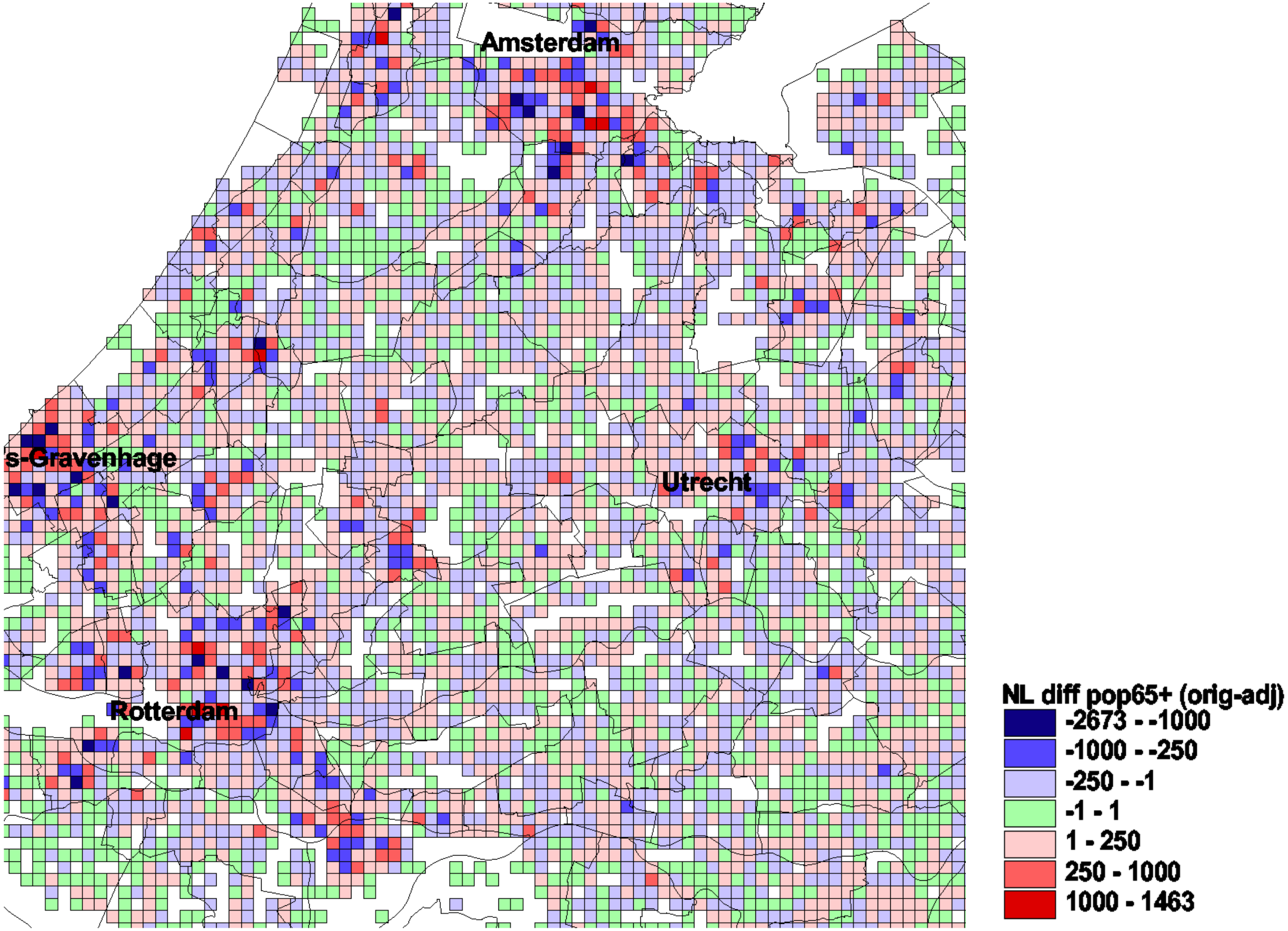
Adjustment example: NL Population 65+ LAU2 values



Adjustment example: NL Population 65+ adjusted values



Adjustment example: NL Population 65+ adjustment sizes



Reference distributions used to adjust each population characteristics grid

	Over 65	Male over 65	EmpEcAc Fem	EmpEcAc Fem InArea
Pop under 30	NW over 68	NW over 68	*NW over 68*	*NW over 68*
Pop 30-99	AT over 65	AT male over 65	*AT male over 65*	*AT male over 65*
Pop 100 and over	EN over 65	EN male over 65	EN EmpEcAc Fem	*EN EmpEcAc Fem*

Black = good match; *Red = approximate match*

Results: evaluation of alternative disclosure thresholds

- This adjustment methodology has been applied across the entire GEOSTAT 2006 grid for the selected social characteristics
- Four thresholds (3, 10, 30, 50) have been applied to the original and adjusted variables
  - These can be compared to the thresholds in the reference data (0 for Norway, 30 for Austria and 100 for England) – there is wide variation in current European threshold values
- We can assess the differences in the suppression of each variable before and after adjustment, according to cells and populations

- Percentage of cells suppressed (Netherlands), Pop 65+

Thresholds	3	10	30	50
Original distribution	22,4%	51,0%	71,6%	76,7%
Modelled distribution	27,7%	53,5%	72,2%	77,5%

- Percentage of cells suppressed (Finland), Pop 65+

Thresholds	3	10	30	50
Original distribution	63,1%	86,5%	94,7%	96,6%
Modelled distribution	60,7%	84,2%	94,1%	96,3%

- Percentage of cells suppressed (Netherlands)

Thresholds	3	10	30	50
Total population	3,8%	11,3%	29,1%	41,4%
65+	27,7%	53,5%	72,2%	77,5%
Male 65+	44,8%	68,6%	80,8%	85,2%
Female employed	21,8%	42,9%	66,5%	73,8%
Female employed in area	35,8%	60,0%	76,7%	81,4%

- Percentage of cells suppressed (Norway)

Thresholds	3	10	30	50
Total population	27,5%	50,0%	67,1%	73,4%
65+	60,9%	75,9%	85,9%	89,4%
Male 65+	71,9%	83,9%	91,6%	94,7%
Female employed	54,8%	71,0%	81,5%	86,1%
Female employed in area	61,0%	74,2%	85,1%	88,9%

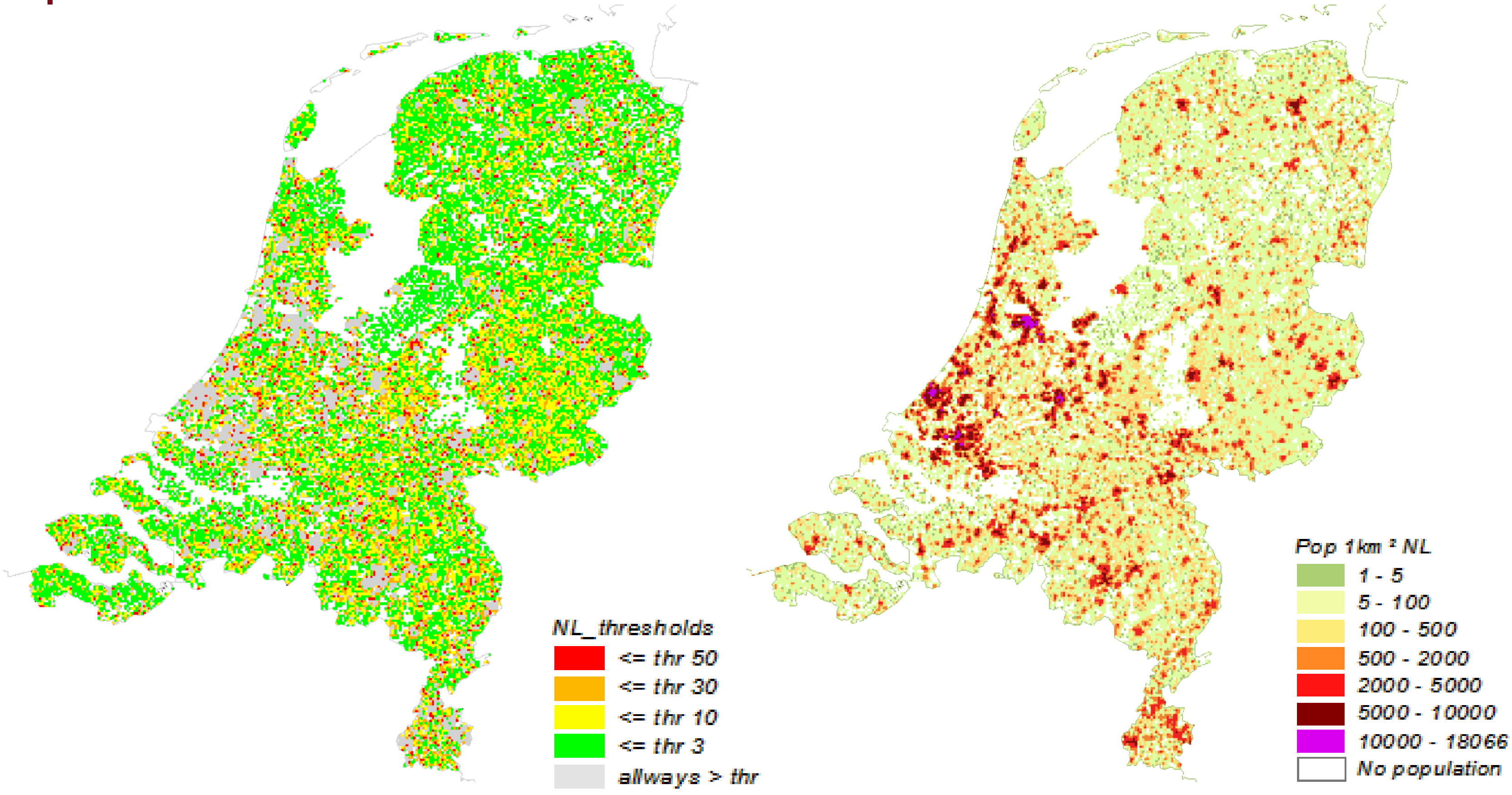
- Percentage of population suppressed (Netherlands)

Thresholds	3	10	30	50
total population	0,0%	0,1%	0,7%	1,6%
65+	0,5%	2,5%	6,7%	9,4%
Male 65+	1,5%	5,7%	12,4%	17,8%
Female employed	0,3%	1,4%	5,3%	7,8%
Female employed in area	0,7%	3,1%	8,0%	11,1%

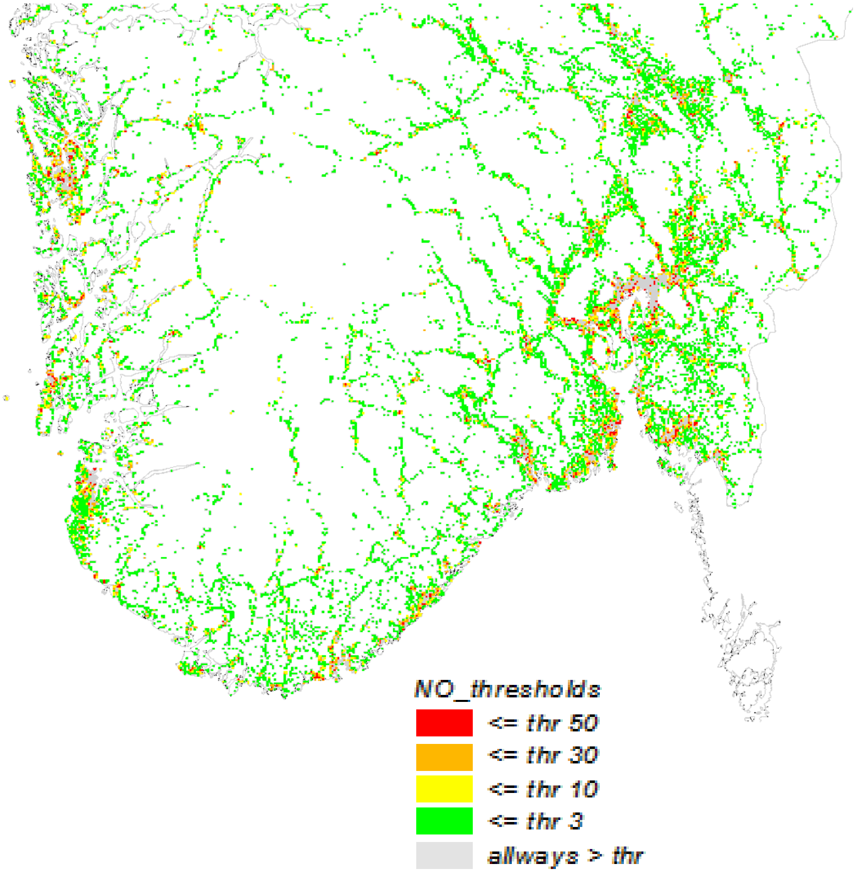
- Percentage of population suppressed (Norway)

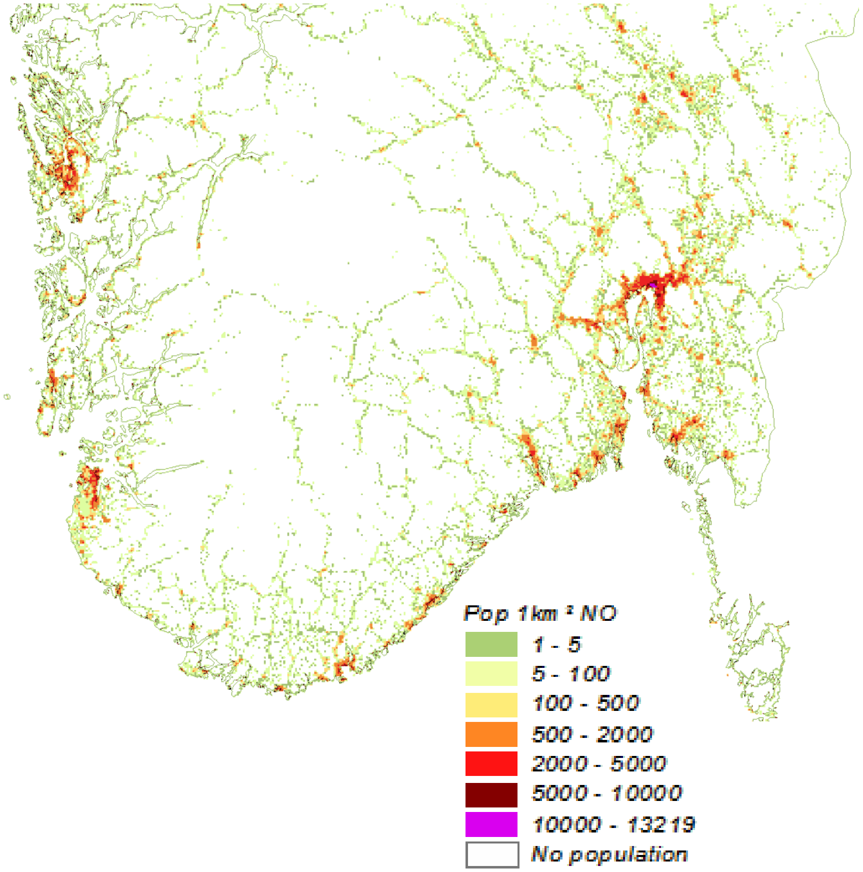
Thresholds	3	10	30	50
total population	0,3%	1,3%	3,3%	5,0%
65+	2,4%	6,1%	13,5%	19,5%
Male 65+	4,5%	11,6%	25,8%	38,3%
Female employed	1,5%	3,8%	9,0%	13,9%
Female employed in area	2,1%	4,8%	12,2%	17,7%

Netherlands – effect of thresholds on male 65+, beside total population



Norway – effect of thresholds on male 65+, beside total population





Conclusions

- The adjustment methodology permits the evaluation of thresholding impacts on more realistic distributions of social characteristics in cells
    - But it will not fully reflect spatial autocorrelation in the grid
- Extremely small cell values present in the grid present great challenges for disclosure control by thresholding
- Problems are most severe for unusual social variables and very small population sizes – especially in rural areas and sparse countries
- Impact on population is less severe than for cells, but there will still be large distortions in the maps
- If the thresholds used in the most conservative countries were to be applied across the grid, most of the data would be suppressed in some countries
- Potential value of exploring alternative perturbation or modelling methods that preserve totals but would not require such high levels of suppression