Johannes Gussenbauer
Alexander Kowarik
Klaudius Kalcher (mostlyAI)
Michael Platzer (mostlyAI)
Statistik Austria

AI-Based Privacy Preserving Census(like) Data Publication

Work Session on Statistical Data Confidentiality 2021

Wien
December 2021

Overview

- Project description
- Use case for Statistics Austria
- Utility and Risk

Project Description

- FFG-Project: "AI-Based Privacy-Preserving Big Data Sharing for Market Research" [1]
- Generate synthetic data from sequential or longitudinal micro data using generative deep learning models
- Partners: WU Vienna, MostlyAI, George Labs, Statistics Austria

[1] This research is supported by the "ICT of the Future" funding programme of the Austrian Federal Ministry for Climate Action, Environment, Energy, Mobility, Innovation and Technology.

Statistics Austria Use Case

- Generate synthetic micro data of the Austrian population

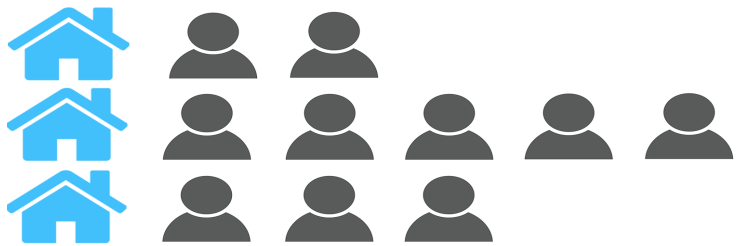
- Input is the "Richframe"

pseudonymized micro data set containing every person registered in the Austrian housing and living register as main residence within private (non-institutional) households

- Used "Richframe" from Q2 2021
  - Number of households 4015907
  - Number of persons 8845691

Statistics Austria Use Case

- How is population micro data sequential?



- Total of 25 variables
  - Household variables: NUTS region, urbanity, tenancy, type of housing, ...
  - Personal variables: age, sex, education, working status, citizenship (ISO 3), country of birth (ISO 3), yearly income, ...

Methodology

- MOSTLY AI synthetic data platform
  - installed locally at Statistics Austria

- Network is build automatically based on column types (numeric and categorical for this use case)
  - general model size
  - complexity of the link between the two tables
  - complexity of the sequential structure of the linked table

- After network is trained on input $\rightarrow$ generate synth. data from random seed

Utility

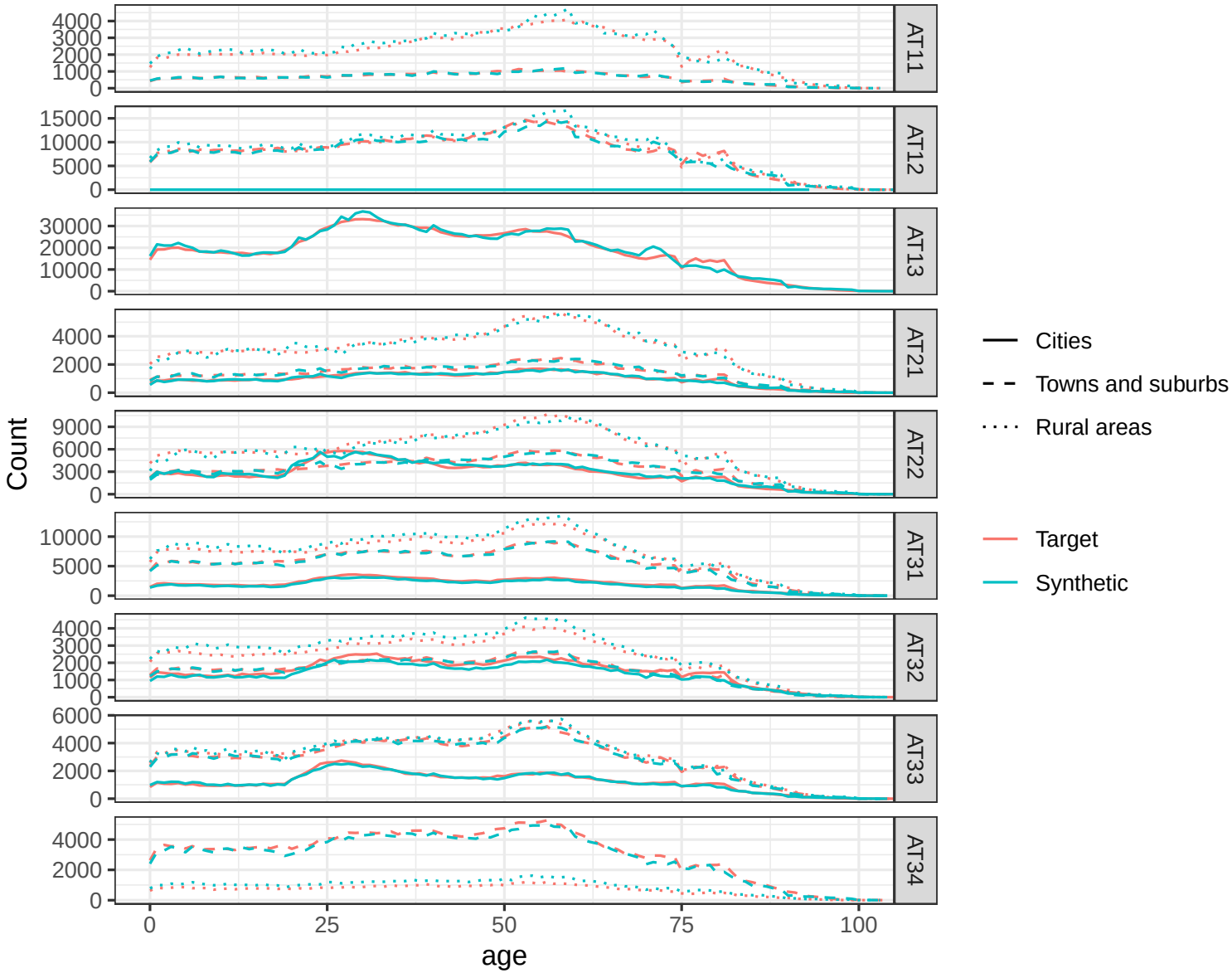
Utility

	Target	Synthetic	Difference (%)
Households	4015907	4015907	0.0000
Persons	8845691	8942421	1.0935

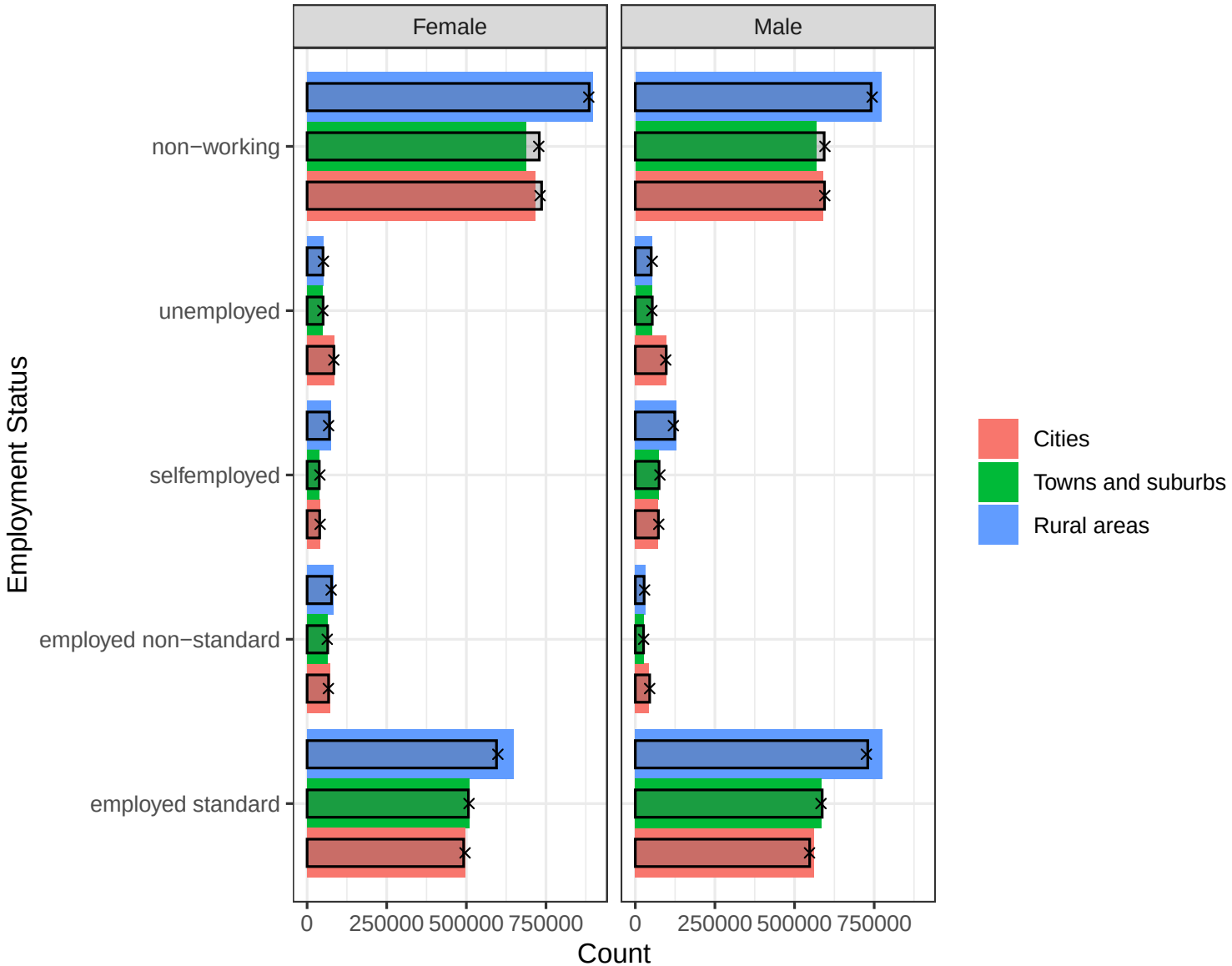
- Minor inconsistencies
  - Households conaining only minors: 535 cases
  - Households where county ≠ municipality: 352 cases
  - Persons with faulty citizenship or country of birth (low/mid/high): 2074/245 cases

CIT_high	CIT_mid	CIT_high	Cases
AT	AT	276	3

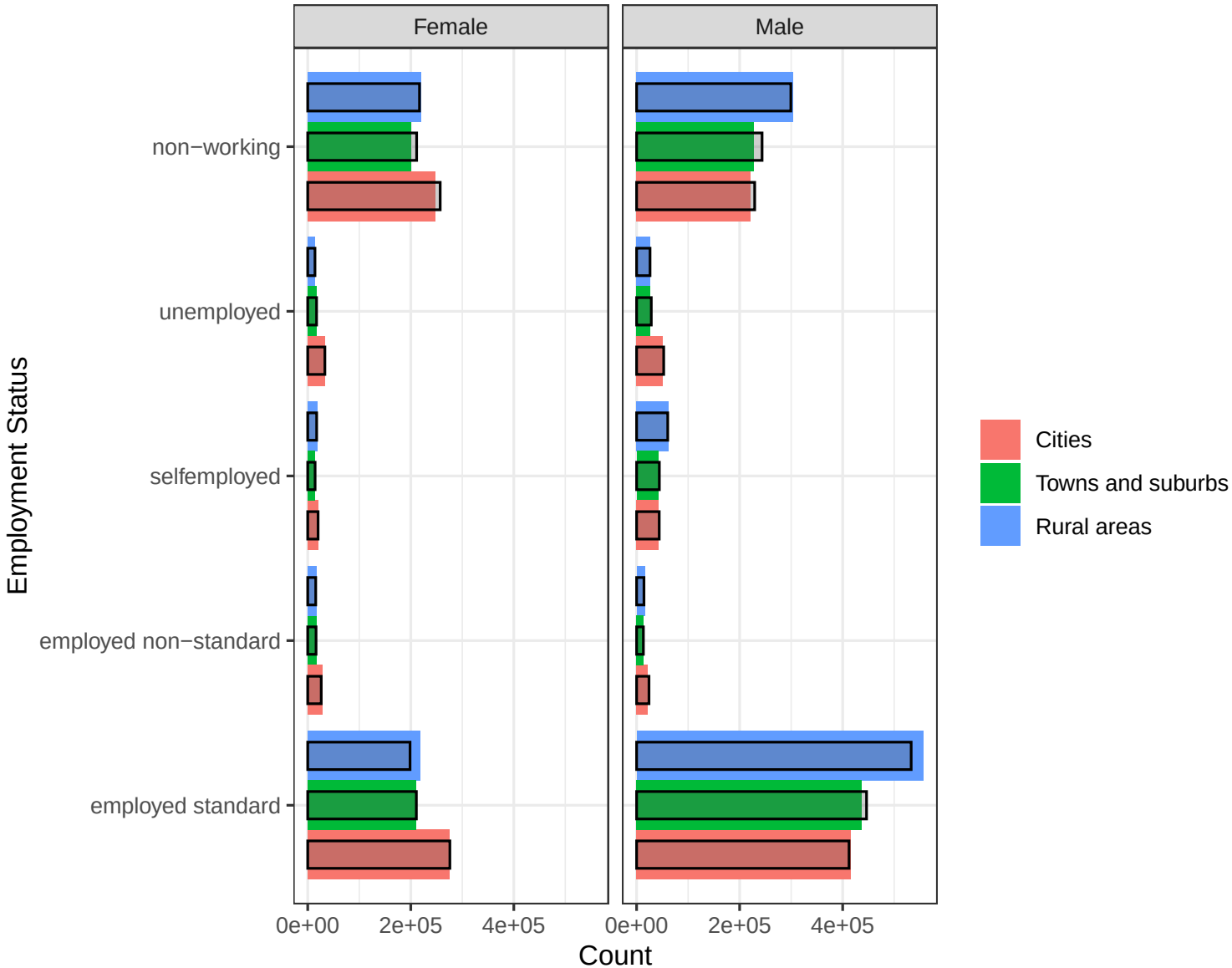
Distribution of age x nuts2 x urbanisation



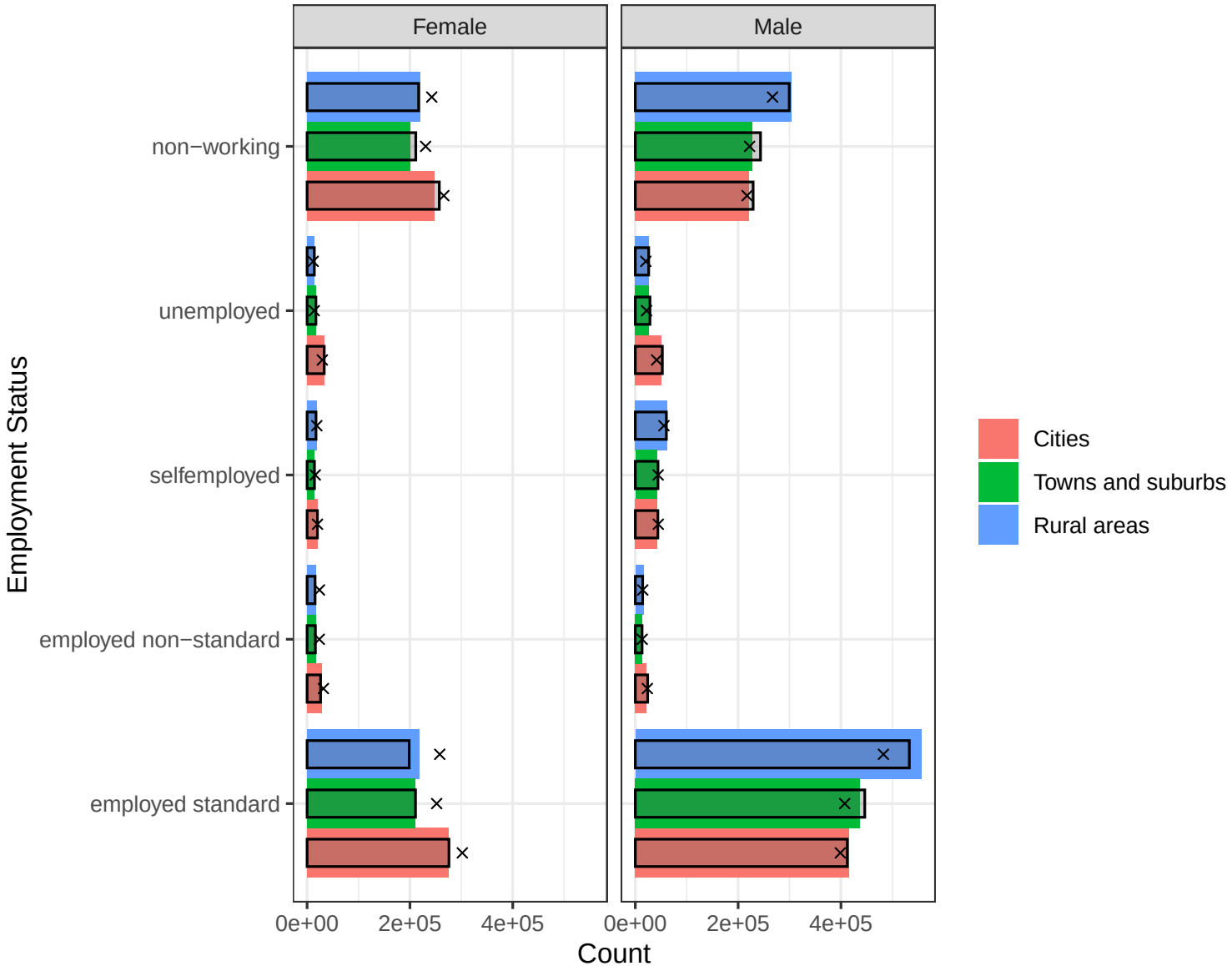
Distribution of income status x sex x urbanisation



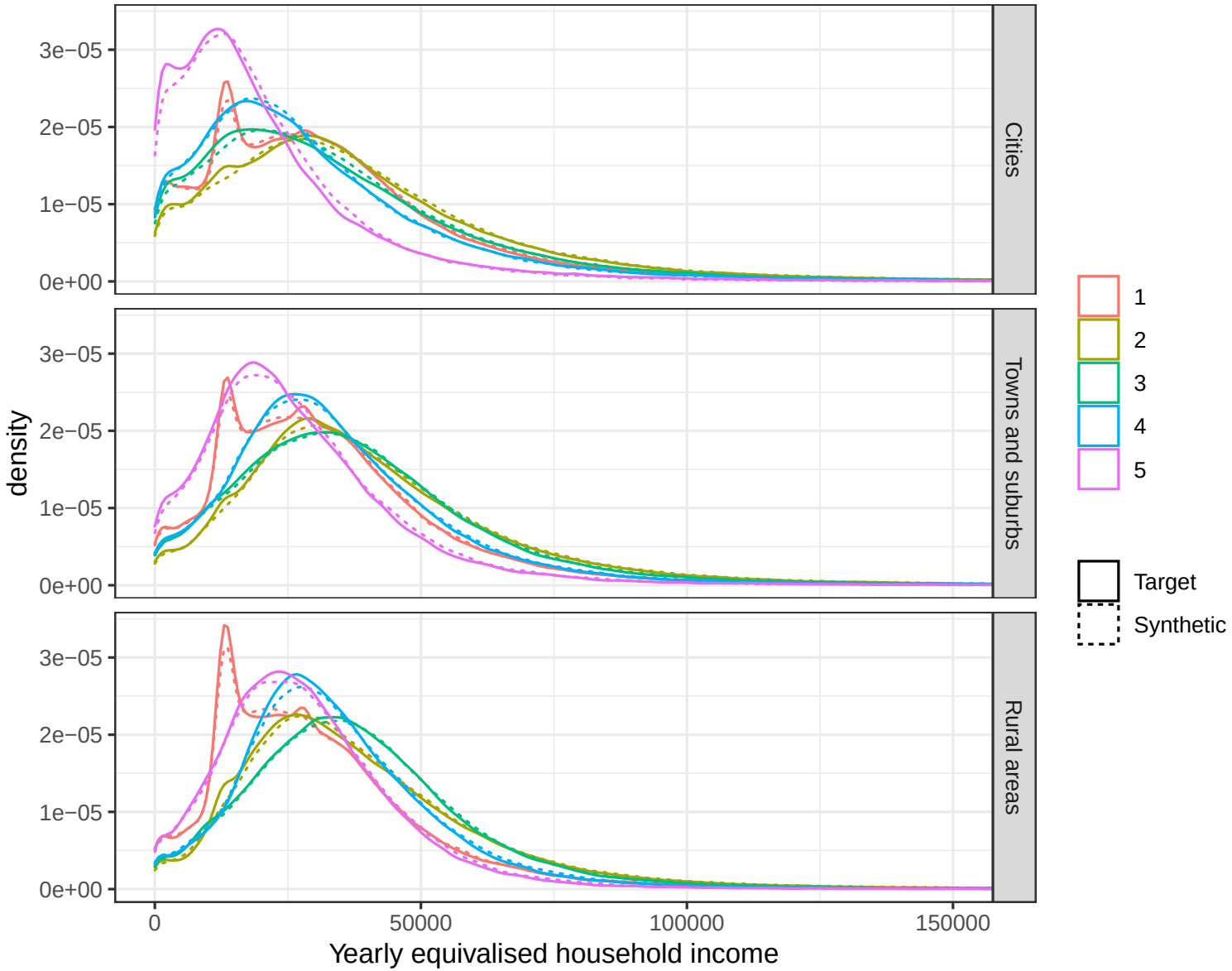
Distribution of income status x sex x urbanisation - highest earner in household



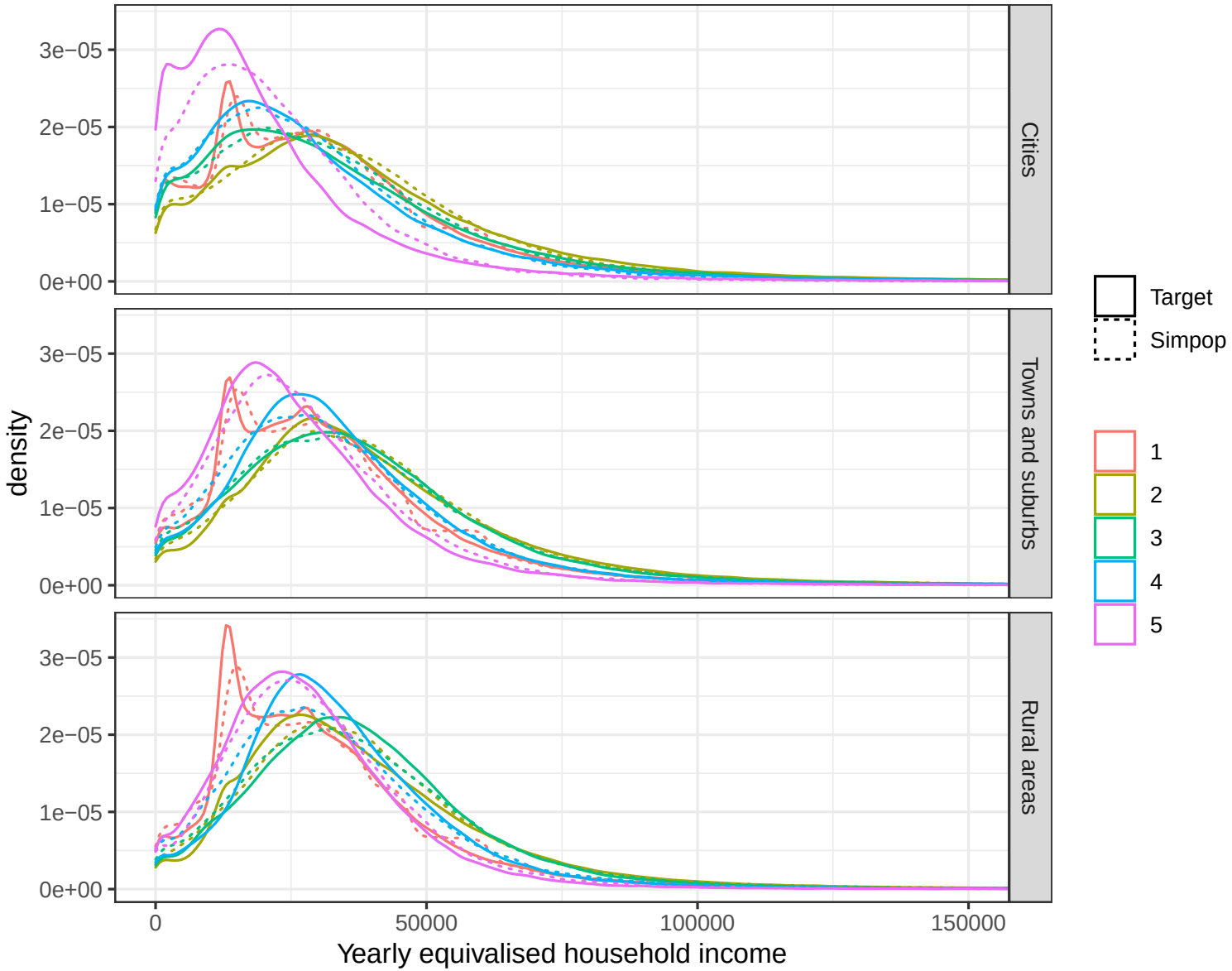
Distribution of income status x sex x urbanisation - highest earner in household



Distribution of income equiv. income x urbanisation x household size



Distribution of income equiv. income x urbanisation x household size



Risk

Risk

- Risk measures used by the mostlyAI-Software
  - Identical Match Share (IMS): identical records between synthetic data and input data
  - Distance to Closest Record (DCR): distance to the closest record in the input data
  - Nearest Neighbour Distance Ratio (NNDR): ratio between the distances to the nearest and second-nearest neighbors in the input data
- Select holdout data set before training model and generating synthetic data
- Compare IMS, DCR and NNDR between synthetic and input as well as holdout and input data

Table: Privacy measures of synthetic and holdout data set.

	IMS	DCR 5th percentile	NNDR 5th percentile
Holdout Data	0.01%	0.02	0.39
Synthetic Data	0.01%	0.02	0.41

- Still more risk analysis necessary

Conclusion

- Generating synthetic data using neural networks definitely feasible
- Difficult to tune neural network if generated data is not satisfactory
- Runtime quite considerable ~ 4 days to generate 4 mio households
- Generated data does need some attention to fix minor inconsistencies
- Komplex distributions seem to be quite well preserved
  - This would need special attention to modelling using other methods
- Need to analyse synthetic data more for more final conclusion

STATISTIK AUSTRIA
Die Informationsmanager

Rückfragen bitte an:
Johannes GussenbauerAlexander KowarikKlaudius Kalcher (mostlyAI)Michael Platzer (mostlyAI)

Kontakt:
Guglgasse 13, 1110 Wien
Tel: +43 (1) 71128–7327
Johannes.Gussenbauer@statistik.gv.at

AI-Based Privacy Preserving Census(like) Data Publication

Work Session on Statistical Data Confidentiality 2021