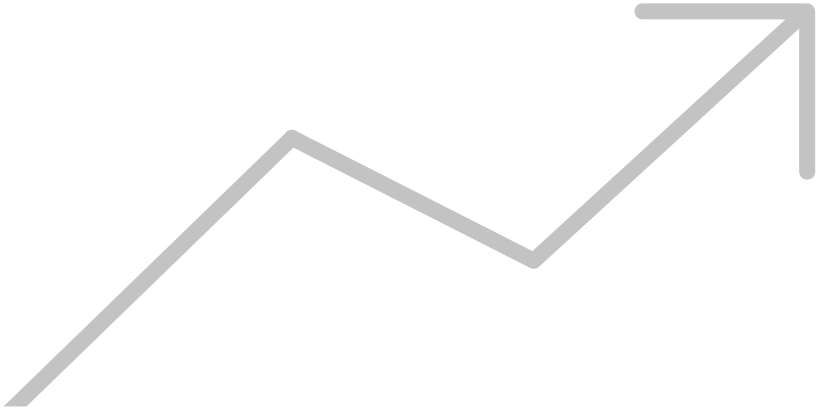
DESTATIS
Statistisches Bundesamt

Anonymization for Integrated and Georeferenced Data (AnigeD)

Yannik Garcia Ritz & Jannek Mühlhan

UNECE Expert meeting on Statistical Data Confidentiality 2023



AnigeD

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung
Finanziert von der Europäischen Union
NextGenerationEU

DESTATIS
Statistisches Bundesamt

Agenda

(1) Competency cluster AnigeD

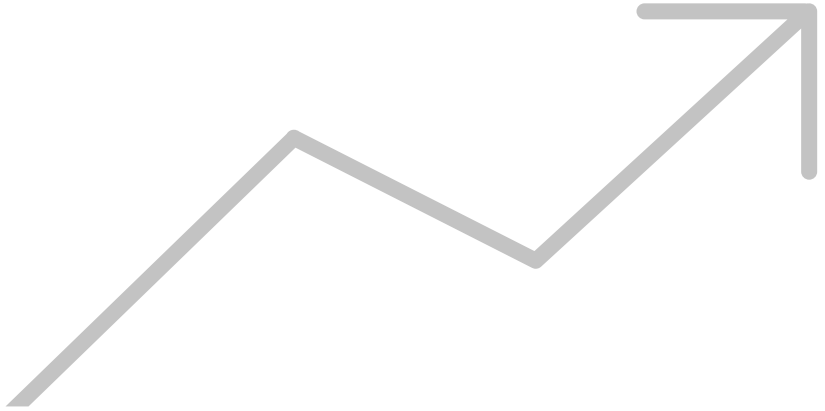
(2) Background of Data Synthesis

(3) Synthesis Approach

(4) Evaluation Approach

(5) Evaluation Results

(6) Discussion



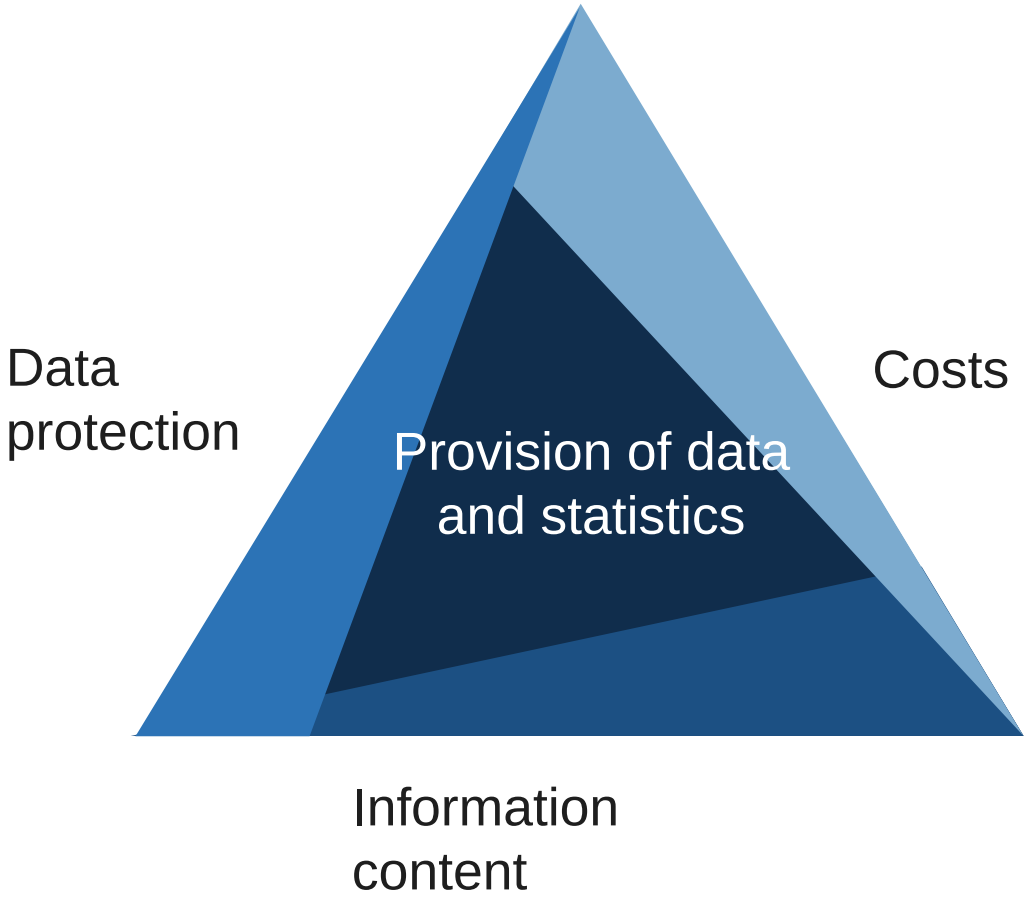
AnigeD

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Finanziert von der
Europäischen Union
NextGenerationEU

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung
Finanziert von der Europäischen Union NextGenerationEU

Initial situation

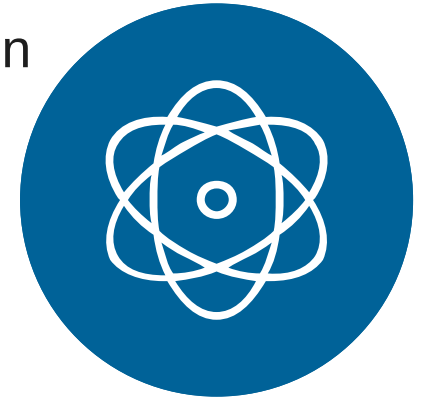


- Federal Ministry of Education and Research (BMBF) initiates nationwide “Anonymization for Secure Data Use” research network
  - Individual research projects and collaborative projects (competency clusters) are funded for three years
  - financed by the European Union – NextGenerationEU

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung
Finanziert von der Europäischen Union NextGenerationEU

Competency cluster AnigeD - Anonymization for Integrated and Georeferenced Data

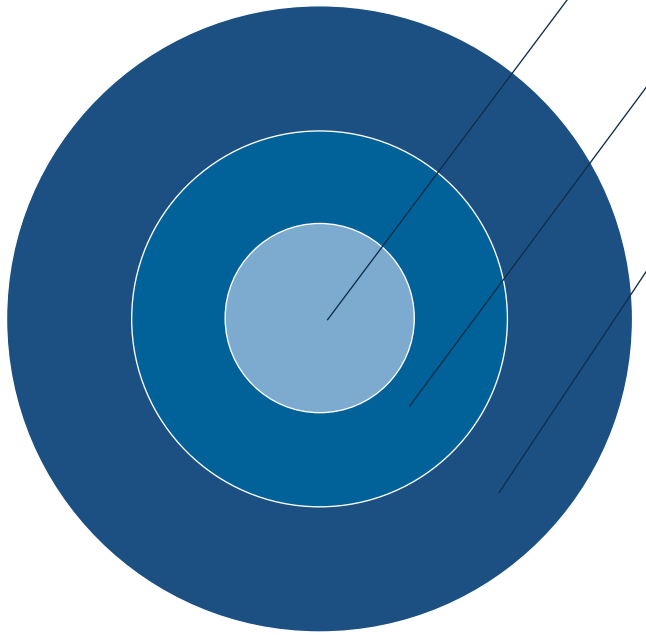
- AnigeD: Anonymization for Integrated and Georeferenced Data
  - Objective: securing and extending access to complex data while observing protection requirements
  - Total funding amount: EUR 4.37 million
  - Funding period: 11/2022 - 11/2025
  - Cluster coordination: Federal Statistical Office (Destatis), Wiesbaden



GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung
Finanziert von der Europäischen Union NextGenerationEU

Research partners

AnigeD

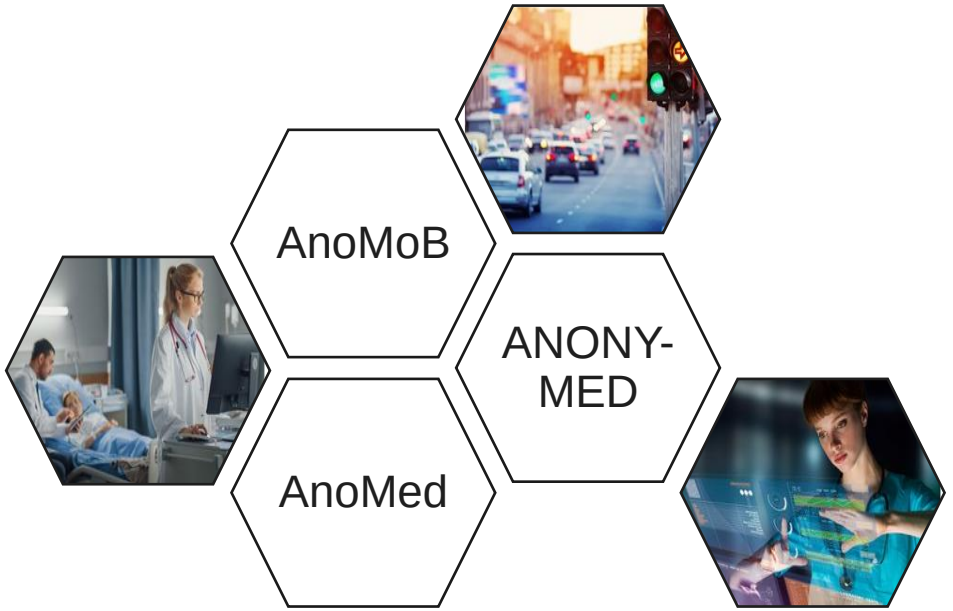


Project partners:
Destatis, FU Berlin, IAB, TH Köln, Speyer University

Associated partners:
DIW, EuroDaT, MPI-SWS, SMA Development GmbH, Telekom, Duisburg-Essen University,

Other research projects:
AnGer, DARIA, GANGES

Associated clusters



GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung
Finanziert von der Europäischen Union NextGenerationEU

Research priorities

- Evaluation of anonymization methods using legal criteria
- Potential of anonymization by synthetic data
- Anonymization of georeferenced data
- Testing of software tools for the efficient analysis and provision of anonymized, georeferenced data

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung
Finanziert von der Europäischen Union NextGenerationEU

Work package 3

- Comparison of procedures from statistics and informatics
- Systematic evaluation of criteria
- Analysis and methodological refinement of synthesis procedures
- Evaluation of synthetic data acceptance by the scientific community

Anonymization by means of synthetic data

Anonymization by means of synthetic data to provide microdata for the scientific community

Machine learning on anonymization by means of synthetic data

Basis for assessing synthetic data generation approaches for statistical confidentiality

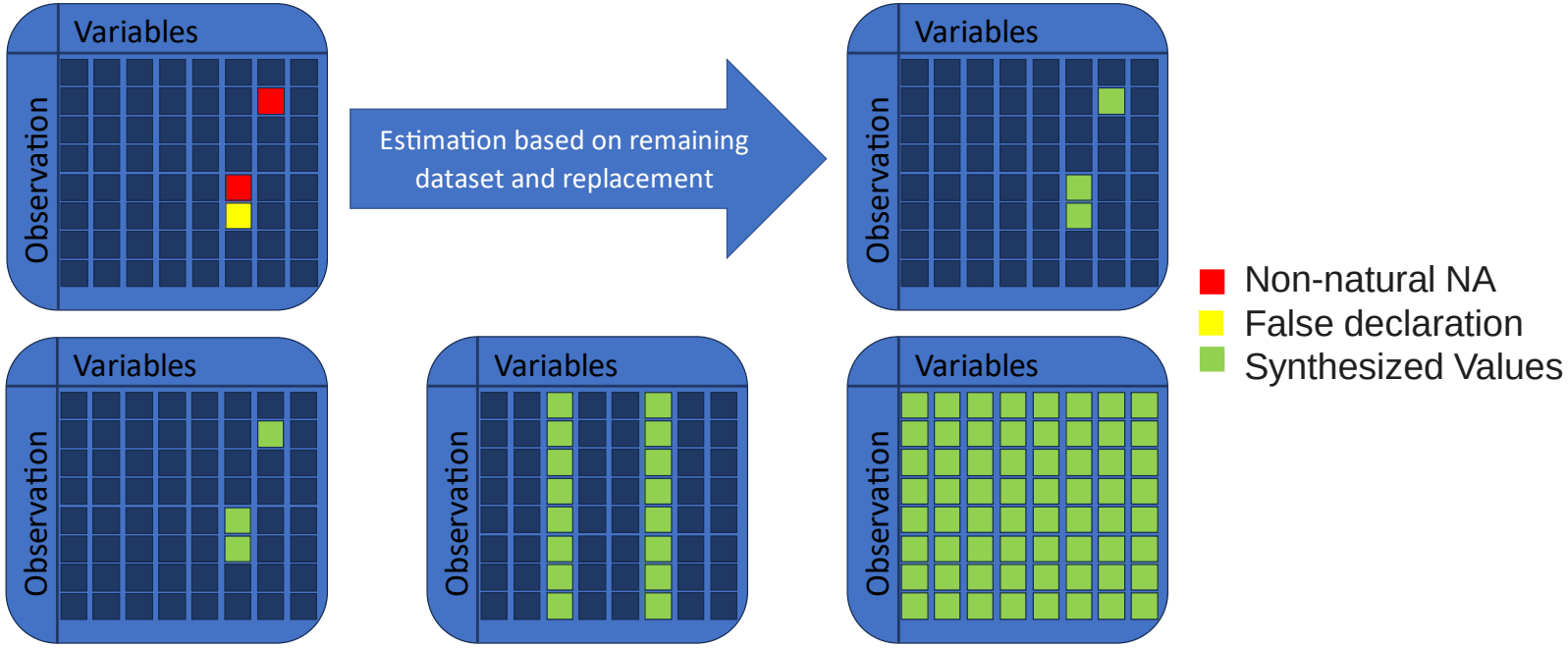
Synthetic data in statistics and informatics - Systematic comparison and methodological refinement (SynDeStatIk)

Context of Research on Anonymization Potentials of Data Synthesis

- Aim: reducing disclosure risks in provided datasets with less restrictions than established measures (suppression, top coding etc.)
  - Provide less aggregated microdata to enable better analysis for research at less complex ways of data access
- Builds on previous approaches on partial (e.g., Little, 1993) and full synthesis/imputation (Rubin, 1993)

Basic Idea of Data Synthesis

- Aim: generating data which mimics original data regarding distributions, relations etc. but with lower risks of reidentification
- Idea: apply approach of imputation to critical variables / all variables

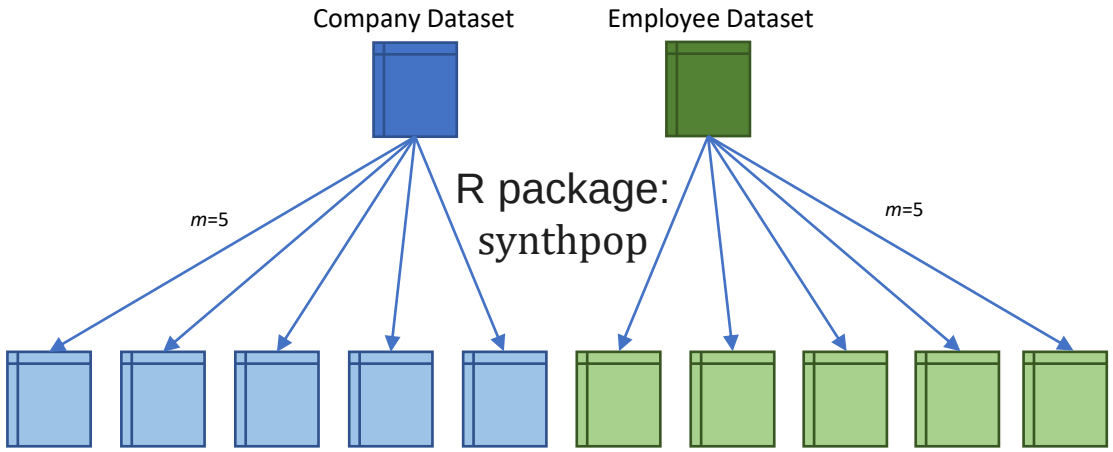


GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung
Finanziert von der Europäischen Union NextGenerationEU

Synthesis Approach

Increasing confidentiality

- Mostly CART-based synthesis
- Increasing the default minbucket parameter leads to tree pruning
- Smoothing for heavily skewed metric variables
  - Spline smoothing
  - Kernel density smoothing
- Minimal # of synthetic datasets to be provided: m (=5) ≥ *r (=2);* *Drechsler (2009); Reiter (2008)*



Nowok , Raab & Dibben (2016)

Evaluation Approach

Disclosure Risk Evaluation

- Risk Ratios (*k*-anonymity, high-risk observations)

$$\frac{Original\ On{-}Site\ Data}{Off{-}Site\ Data} \text{ vs. } \frac{Synthetic\ On{-}Site\ Data}{Off{-}Site\ Data}$$ *(Templ, Kowarik & Meindl, 2015)*

- *Drechsler & Reiter* (2008): Expected Match Risk & True Match Rate

Utility Evaluation

- Ensuring logical constraints, comparing descriptive key measures, examining distributions of analytic key variables and pMSE (global utility)
- Examining confidence interval overlaps of exemplary regression models (model-specific utility)

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung
Finanziert von der Europäischen Union NextGenerationEU

Evaluation Approach

Disclosure Risk Evaluation

Drechsler & Reiter (2008): Expected Match Risk & True Match Rate

- Expected Match Risk: data user randomly selects correct observation

$$\sum_{j \in T} \left(1/c_j\right) * I_j$$

- True Match Rate: share of truly matched targets w/ matches > 1 in D(m_1-m_5)

$$\sum_{j \in T} {}^{K_j}/_{\sum_{j \in T}(c_j=1)}$$

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung
Finanziert von der Europäischen Union NextGenerationEU

Evaluation Results

Disclosure Risk Evaluation – Risk Ratios

- Against first assumption increase in ratios
- Potential interpretation:
  - More unique (synthetic) observations
    -> larger pool of risky observation
    => Lower risk of true matches

	Original	Synthesized, spline smoothing	Synthesized, kernel density smoothing
	Company Dataset		
Ratio *k*=2 violating Obs. (on-/off-site)	5,977.50	7,356.9	6,459.9
Ratio *k*=3 violating Obs. (on-/off-site)	1,397.00	4,884.65	4,532.25
Ratio High-Risk Obs. (on-/off-site)	89.32	120.60	91.76721
	Employee Dataset		
Ratio *k*=2 violating Obs.	1.51	2.39	2.43
Ratio *k*=3 violating Obs.	1.07	3.81	3.82
Ratio High-Risk Obs. (On-/Off-site)	4.06	1.60	1.53

Source: SES 2018. RDCs of the Statistical Offices of the Federation and the Federal States.

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung
Finanziert von der Europäischen Union NextGenerationEU

Evaluation Results

Disclosure Risk Evaluation – Expected Match Risk & True Match Rate

	Synthesized (Spline smoothing)	Synthesized (Kernel density smoothing)
	Company Dataset	
Expect. Match Risk	0.0520 %	0.0502 %
True Match Rate	0.0000 %	0.0000 %
False Match Rate	100.0000 %	100.0000 %
	Employee Dataset	
Expect. Match Risk	0.0000 %	0.0000 %
True Match Rate	0.0000 %	0.0000 %
False Match Rate	100.0000 %	100.0000 %

Source: SES 2018. RDCs of the Statistical Offices of the Federation and the Federal States.

- Minor expected risk for a random match for the company data
- No expected match risk for a random match for the company data
- No true matches for both data materials

Evaluatio

Utility Evaluatic

- Ensuring logic

	Original data without Bavaria	Synthetic data without Bavaria [m_1]	Synthetic data without Bavaria [m_3]	Synthetic data without Bavaria [m_5]
		Sex		
Avg	1.4697	1.4696	1.4703	1.4697
median	1	1	1	1
SD	0.4991	0.4991	0. 4991	0.4991
		Year of Birth		
Avg	1973.6222	1973.6273	1973.6244	1973.6249
median	1972	1972	1972	1972
SD	13.0492	13.0436	13.0453	13.04500
		Year of entry into the company		
Avg	2005.0247	2005.2054	2005.2030	2005.2035
median	2010	2010	2010	2010
SD	12.6139	12.4221	12.4232	12.4226
		Gross monthly income		
Avg	2989.3089	2988.5670	2987.0762	2985.8921
median	2686	2685	2685	2688
SD	2490.2024	2518.3297	2453.5373	2371.0687
		Total earnings for overtime hours		
Avg	19.9182	19.4604	19.4583	19.5073
median	0	0	0	0
SD	126.1532	122.3120	122.0371	122.5444
		Shift- and night shift credits, weekend and holiday extra charges,		
Avg	29.5375	29.8927	29.7160	30.0267
median	0	0	0	0
SD	128.1537	127.4987	126.1038	127.4425
		Statutory deductions due to income tax and solidarity surcharge		
Avg	532.9553	518.4816	518.3125	517.7757
median	329	327	328	327
SD	875.9017	745.5711	766.2905	709.1242
		Statutory deductions due to social insurance		
avg	477.2533	484.0925	483.5502	483.7454
median	446	443	443	443
SD	309.9909	391.1970	352.7239	357.0693
		Gross yearly income		
avg	37903.7149	35862.8038	35844.9144	35830.7048
median	33367	32220	32220	32256
SD	36898.1805	30219.9569	29442.44793	28452.8244
		Net monthly income		
avg	1985.79994	1985.9929	1985.2135	1984.3709
median	1821	1812	1813	1814
SD	1494.4055	1545.2950	1511.5743	1467.8617

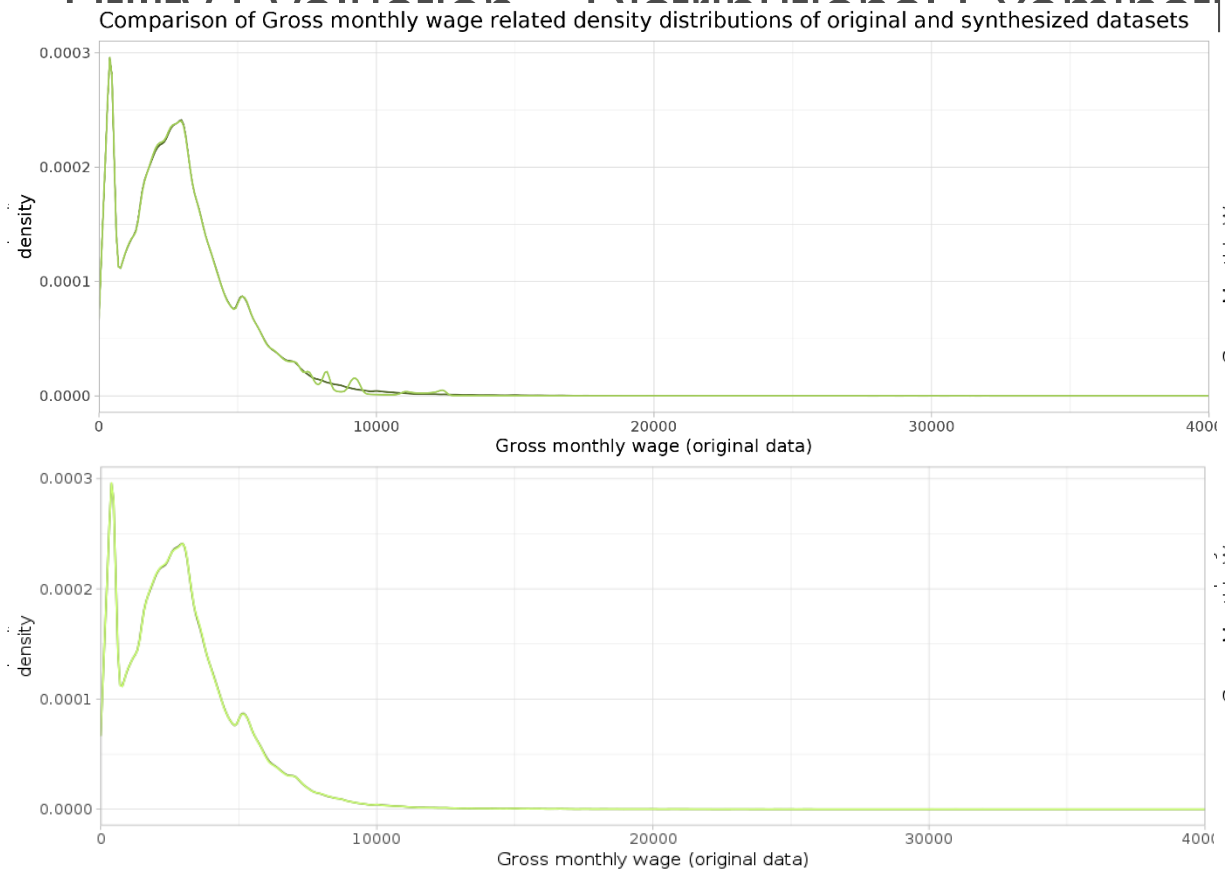
Source: SES 2018. RDCs of the Statistical Offices of the Federation and the Federal States.

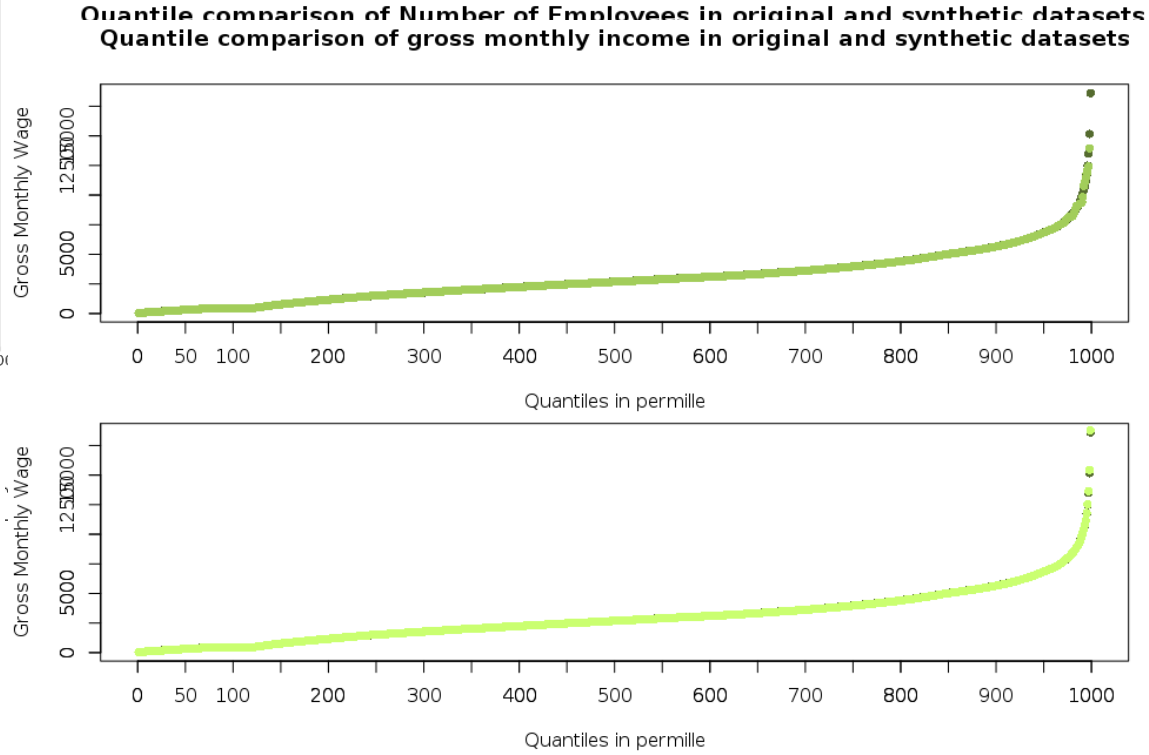
Evaluation Results

Firmendaten

Angestelltendaten

Utility Evaluation – Distributional Examination of Key Variables





GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung
Finanziert von der Europäischen Union NextGenerationEU

Evaluation Results

Utility Evaluation – Global Utility

- Ensuring logical constraints & comparing descriptive key measures
- Distributions of analytic key variables
- Propensity Mean-Squared Error (pMSE)

	Company Dataset		Employee Dataset	
	Spline Smoothing	Kernel Density Smoothing	Spline Smoothing	Kernel Density Smoothing
pMSE	0.1142	0.1142	0.1102	0.1110

Source: SES 2018. RDCs of the Statistical Offices of the Federation and the Federal States.

$$pMSE = \frac{1}{m}\sum_j^m \left(\frac{1}{N}\sum_i^N (\hat{p}_i - c)^2\right)$$

- pMSE interval per definition [0; 0.25]

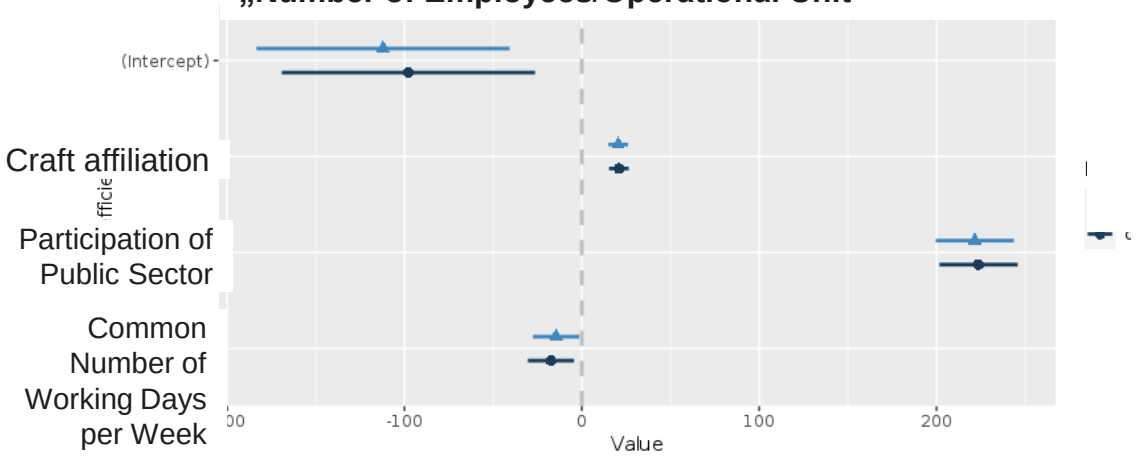
Evaluation Results

Firmendaten

Utility Evaluation – Model-Specific Utility (Company Material)

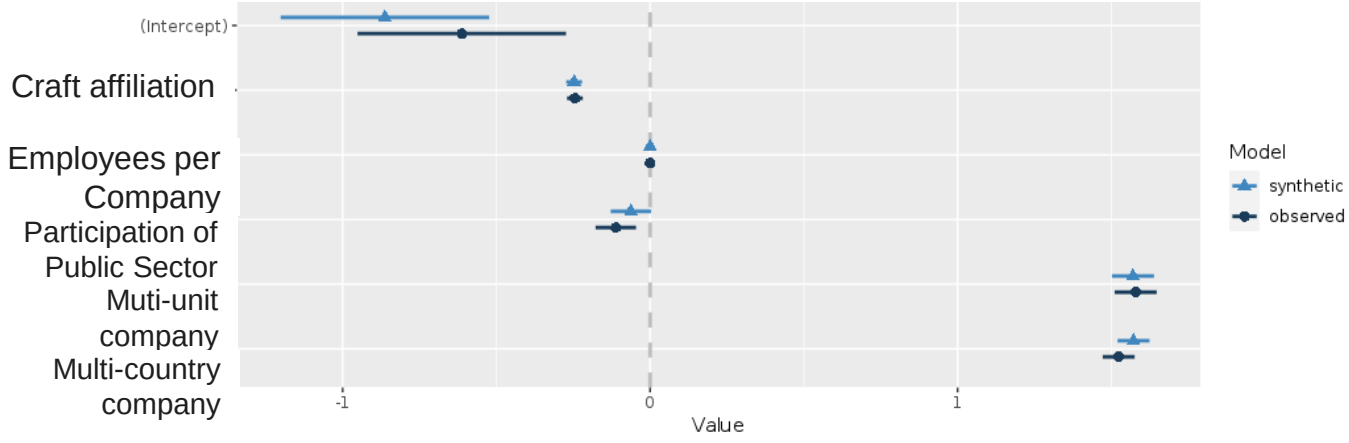
- Examining confidence interval overlaps of exemplary regression models

Coefficents for fit to „Number of Employees/Operational Unit“



mean Confidence Interval Overlap: 0.85

Coefficients for fit to „Company Loans are Negiotated Based on Collective Bargaining“



mean Confidence Interval Overlap: 0.74

Evaluation Results

Angestelltendaten

Utility Evaluation – Model-Specific Utility (Employee Material)

- Examining confidence interval overlaps of exemplary regression models

	Original on-site material (employee dataset) Coefficient (Std. error)	Synthesized on-site material (employee dataset) Coefficient (Std. error)	CI Overlap
Intercept	589.1931*** (2.569)	606.0611*** (2.56875)	-0.6752
Education	1.550*** (0.0197)	1.54555*** (0.01968)	0.9394
Sex	-3.3940*** (0.0249)	-3.18269*** (0.02488)	-1.1664
Year of Birth	-0.0304*** (0.0012)	-0.0682*** (0.0012)	-6.9942
Year of Entry	-0.2567*** (0.0015)	-0.2279*** (0.0015)	-3.7869
Restriction of term of contract	-1.4092*** (0.0101)	1.4784*** (0.01008)	-0.7522
Private sector	0.0716 (0.0542)	0.2492*** (0.0542)	0.1643
Company size	-0.0000*** (0.0000)	-0.0000*** (0.0000)	0.23399
Vocational education	3.5243*** (0.0124)	3.3684*** (0.01235)	-0.2990

Source: SES 2018. RDCs of the Statistical Offices of the Federation and the Federal States.
* p < 0.10, ** p < 0.05, *** p < 0.01

	Original on-site material (employee dataset) Coefficient (Std. error)	Synthesized on-site material (employee dataset) Coefficient (Std. error)	CI Overlap
Intercept	99.76548*** (0.9690)	84.8586*** (0.9690)	-2.9244
Sex	-0.24396*** (0.01248)	-0.2399*** (0.01248)	0.9172
Year of Birth	-0.04444*** (0.00049)	-0.0379*** (0.00049)	-2.3949
Education	-0.0408*** (0.0075)	-0.1054*** (0.0075)	-1.1864
Vocational Education	0.4637*** (0.00736)	0.5034*** (0.00736)	-0.3762
Restriction of term of contract	-1.93796*** (0.0090)	-1.6336*** (0.0090)	-7.6049
Weekly working hours	-0.16288*** (0.00086)	-0.1277*** (0.00086)	-9.3753
Private sector	0.2412*** (0.02997)	0.2035*** (0.02997)	0.6791
Company size	-0.0000*** (0.0000)	-0.0000*** (0.0000)	0.5961

Source: SES 2018. RDCs of the Statistical Offices of the Federation and the Federal States.
* p < 0.10, ** p < 0.05, *** p < 0.01

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung
Finanziert von der Europäischen Union NextGenerationEU

Discussion

Limitations

- Only partial synthesis is tested
- Results are not generizable for other surveys
- Assessment of Risk-Utility-Ratio yet not satisfying
- Research results cannot serve as legal report

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung
Finanziert von der Europäischen Union NextGenerationEU

Discussion

Research Implications

- Replicate studies with a full synthesis approach
- Evaluate utility and risk for longitudinal data (of other surveys)
- Examine further surveys regarding potentials of synthetic data
- Applying hyperparameter tuning to optimize cost-utility-ratio

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung
Finanziert von der Europäischen Union NextGenerationEU

Discussion

Practical Implications

- If future research is able to improve generated synthetic data regarding the Risk-Utility-Ratio:
  - Lawyers need to evaluate possibilities to provide synthetic on-site material via off-site access
- Synthetic data need to be provided as separate product at first without opportunities fo project-specific processing until insights are gained

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung
Finanziert von der Europäischen Union NextGenerationEU

Discussion

Conclusion

- Attempt to synthetize and evaluate official German on-site material
- New insights on synthesis of further official surveys
- Encouraging results regarding global utility and disclosure risks
- Improvable results concerning utility

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung
Finanziert von der Europäischen Union NextGenerationEU

Discussion

Conclusion

- Two Options:

1. release partially synthetic data tailored to specific research questions of the data users
2. release fully synthetic datasets if follow-up research is able to provide evidence for an improved model-specific

Contact

Statistisches Bundesamt
Postal address
65180 Wiesbaden

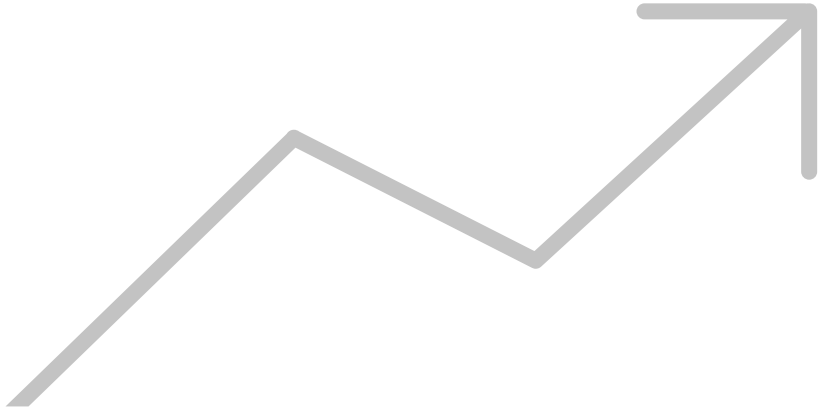
www.destatis.de

www.destatis.de/contact

Yannik.GarciaRitz@destatis.de

Jannek.Muehlhan@destatis.de

Functional mailbox
AnigeD@destatis.de



GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung
Finanziert von der Europäischen Union
NextGenerationEU

Sources

Drechsler, J. (2009). Generating multiply imputed synthetic datasets: theory and implementation. (Doctoral dissertation, Otto-Friedrich-Universität Bamberg, Fakultät Sozial-und Wirtschaftswissenschaften). Bamberg.

Drechsler, J., & Reiter, J. P. (2008). Accounting for Intruder Uncertainty Due to Sampling When Estimating Identification Disclosure Risks in Partially Synthetic Data. In J. Domingo-Ferrer, & Y. Saygin (Ed.), *Privacy in Statistical Databases. 5262*, pp. 227-238. Berlin: Springer. doi:10.1007/978-3-540-87471-3_19

Hafner, H.-P., & Lenz, R. (2011). Some aspects concerning analytical validity and disclosure risk of CART generated synthetic data. Joint UNECE/Eurostat work session on statistical data confidentiality, (pp. 1-10). Tarragona, Spain.

Loske, J., & Wolfanger, T. (2019). Entwicklung Synthetischer Datenstrukturfiles. Statistische Woche, (p. 113). Trier.

Karr, A. F., Kohnen, C. N., Oganian, A., Reiter, J. P., & Sanil, A. P. (2006). A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality. The American Statistician, 60(3), pp. 224-232. doi:10.1198/000313006X124640

Sources

Karr, A. F., Kohnen, C. N., Oganian, A., Reiter, J. P., & Sanil, A. P. (2006). A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality. The American Statistician, 60(3), pp. 224-232. doi:10.1198/000313006X124640

Little, R. J. (1993). Statistical analysis of masked data. *Journal of Official Statistics, 9*(2), pp. 407–426.

Nowok, B., Raab, G. M., & Dibben, C. (2016, October). synthpop: Bespoke Creation of Synthetic Data in R. Journal of Statistical Software, 74(11), pp. 1-26. doi:10.18637/jss.v074.i11

Order of the First Senate of 15, 1 BvR 209/83 -, paras. 1-214 (BVerfG December 1983).

Reiter, J. P. (2008). Selecting the number of imputed datasets when using multiple imputation for missing data and disclosure limitation. Statistics & Probability Letters, 78, pp. 15-20.

Rothe, D. (2015). Statistische Geheimhaltung - der Schutz vertraulicher Daten in der amtlichen Statistik - Teil 1: Rechtliche und methodische Grundlagen. Bayern in Zahlen, pp. 294-303.

Rubin, D. B. (1993). Discussion: Statistical disclosure limitation. Journal of Official Statistics, 9(2), pp. 462-468.

Sources

Rubin, D. B. (1993). Discussion: Statistical disclosure limitation. Journal of Official Statistics, 9(2), pp. 462-468.

Templ, M. (2017). Statistical Disclosure Control for Microdata - Methods and Applications in R (1. ed.). Basel: Springer Cham. doi:10.1007/978-3-319-50272-4

Templ, M., Kowarik, A., & Meindl, B. (2015, October). Statistical Disclosure Control for Micro-Data Using the R Package sdcMicro. Journal of Statistical Software, 67(4), pp. 1-36. doi:10.18637/jss.v067.i04

Woo, M.-J., Reiter, J. P., Oganian, A., & Karr, A. F. (2009). Global measures of data utility for microdata. Journal of Privacy and Confidentiality, 1(1), pp. 111-124.

Zühlke, S., Zwick, M., Scharnhorst, S., & Wende, T. (2005). The research data centres of the Federal Statistical Office and the statistical offices of the Länder. FDZ-Arbeitspapiere, 3, pp. 1-11.