INSTITUTE FOR EMPLOYMENT RESEARCH
The Research Institute of the Federal Employment Agency

AN OVERVIEW OF DATA PROTECTION STRATEGIES FOR INDIVIDUAL-LEVEL GEOCODED DATA

UNECE Expert meeting on Statistical Data Confidentiality

Wiesbaden, 26-28 September 2023

Maike Steffen
Konstantin Körner
Jörg Drechsler

BACKGROUND

- More and more geo-referenced data are being collected
  - Important for various research areas (e.g., to assess neighborhood effects, mobility patterns)
  - Highly identifying, availability for research is limited
  - IAB project on geo-referenced data → how to anonymize these data?
- Three main strategies for confidentiality protection
  - Aggregation
  - Geographic Masking
  - Synthetic data

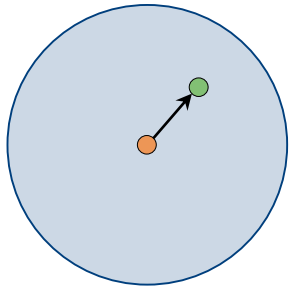
AGGREGATION

- Aggregation within pre-defined areas
  - Administrative areas
  - (Standardized) Grid cells
    - ⊕ External data can easily be linked
    - ⊖ Loss of spatial information
      Choice of aggregation level can bias results

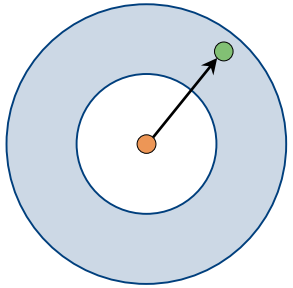
- Flexible aggregation
  - Population-adjusted grid cells
  - Microaggregation
    - ⊕ More efficient trade-off between confidentiality protection and utility
    - ⊖ Cannot easily be linkend to external data
      Harder to interprete

GEOGRAPHIC MASKING

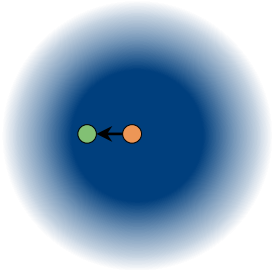
- Deterministic masking approaches
- Random perturbation
  - Original locations are randomly displaced
  - Different methods to draw maximum or minimum displacement distance
  - Possibility to adapt for population density



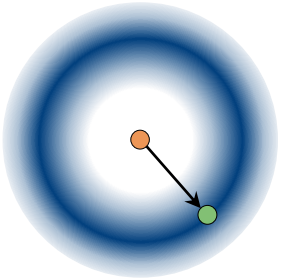
Displacement within a circle



Donut masking



Gaussian masking



Bimodal gaussian masking

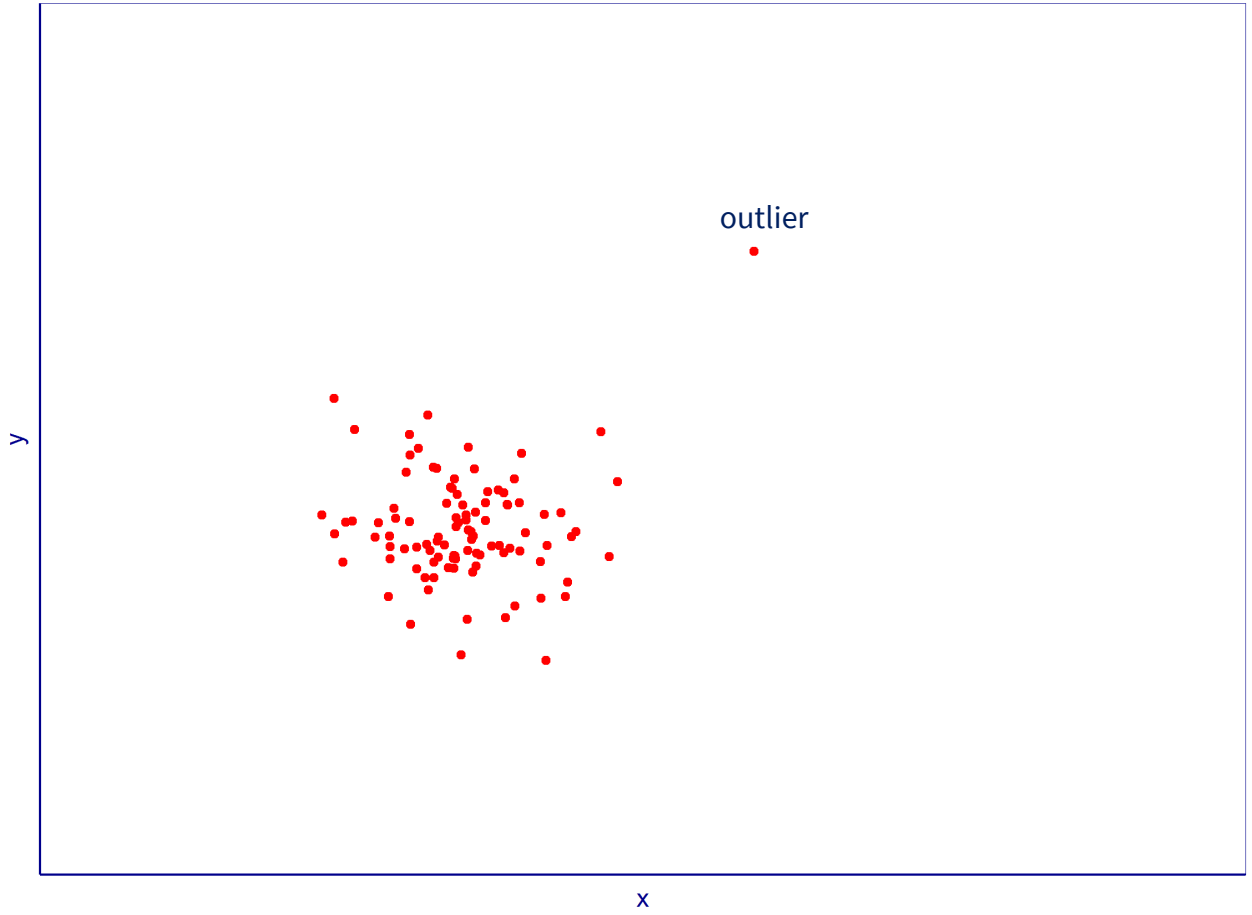
GEOGRAPHIC MASKING

- Deterministic masking approaches
- Random perturbation
  - Original locations are randomly displaced
  - Different methods to draw maximum or minimum displacement distance
  - Possibility to adapt for population density

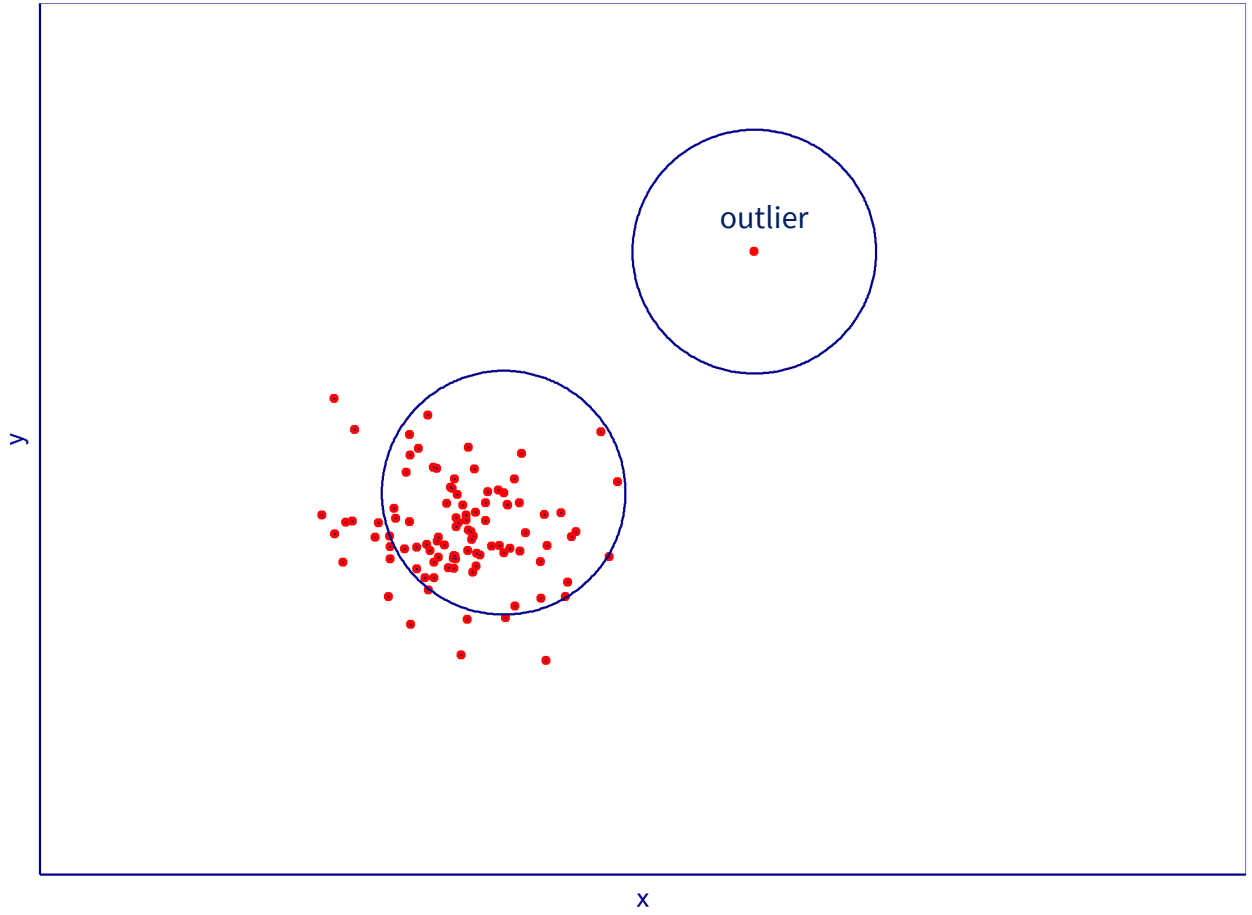
⊕ Widely used, straightforward method
Point-locations as output

⊖ No guaranteed level of privacy protection, especially in rural areas or areas with heterogenous population density

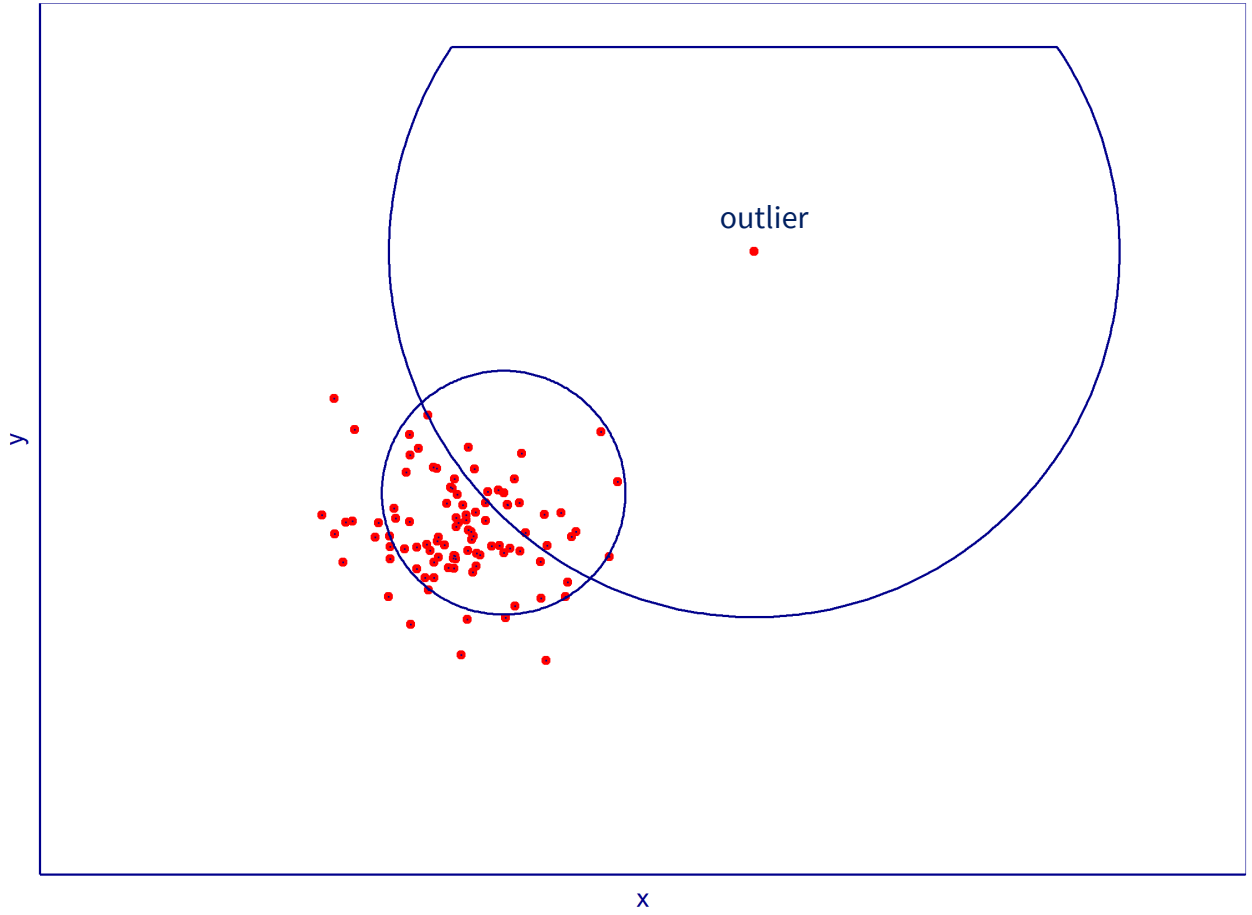
GEOGRAPHIC MASKING SOMETIMES OFFERS LITTLE PROTECTION



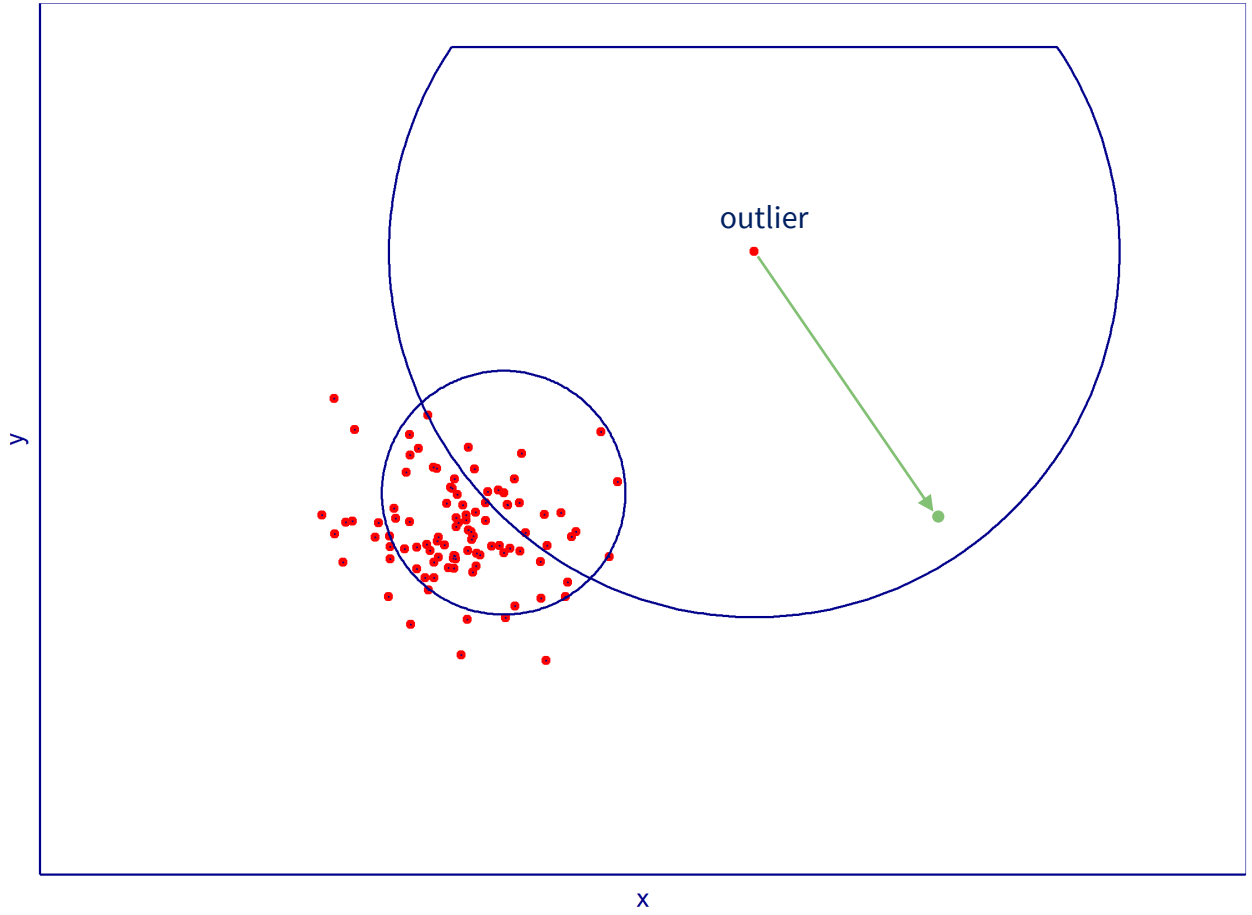
GEOGRAPHIC MASKING SOMETIMES OFFERS LITTLE PROTECTION



GEOGRAPHIC MASKING SOMETIMES OFFERS LITTLE PROTECTION



GEOGRAPHIC MASKING SOMETIMES OFFERS LITTLE PROTECTION



GEOGRAPHIC MASKING

- Location swapping (Zhang et al., 2017)
  - Original location is swapped with another location within a circle or donut
- Adaptive Areal Masking (Kounadi & Leitner, 2016)
  - random perturbation within pre-defined areas with at least k location points
  - Guarantees a certain level of anonymity
  - High alteration of locations

SYNTHETIC DATA

Synthesizing of non-geographic variables

- Account for spatial structure to synthesize non-geographic variables
- Data release
  - Detail level of geographic information
  - Separate release of 2 data sets (Koebe et al. 2023)

Synthesizing of geographic information

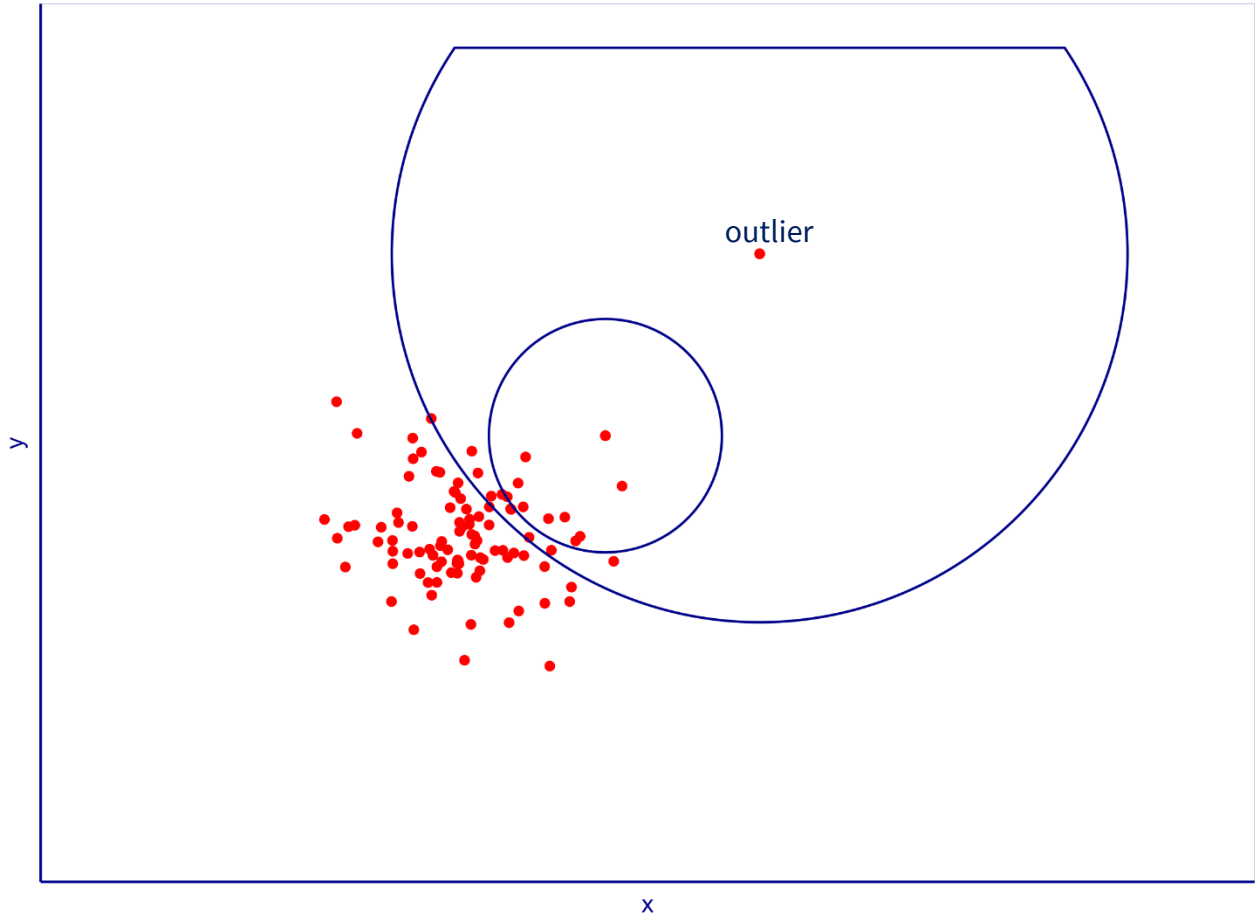
- Aggregated data (Quick, 2021; 2022; Paiva et al., 2014)
- Exact geographic coordninates (Wang & Reiter, 2012; Drechsler and Hu, 2021)
- Fully synthetic data (e.g., Quick et al., 2015)

RISK AND UTILITY ASSESSMENT

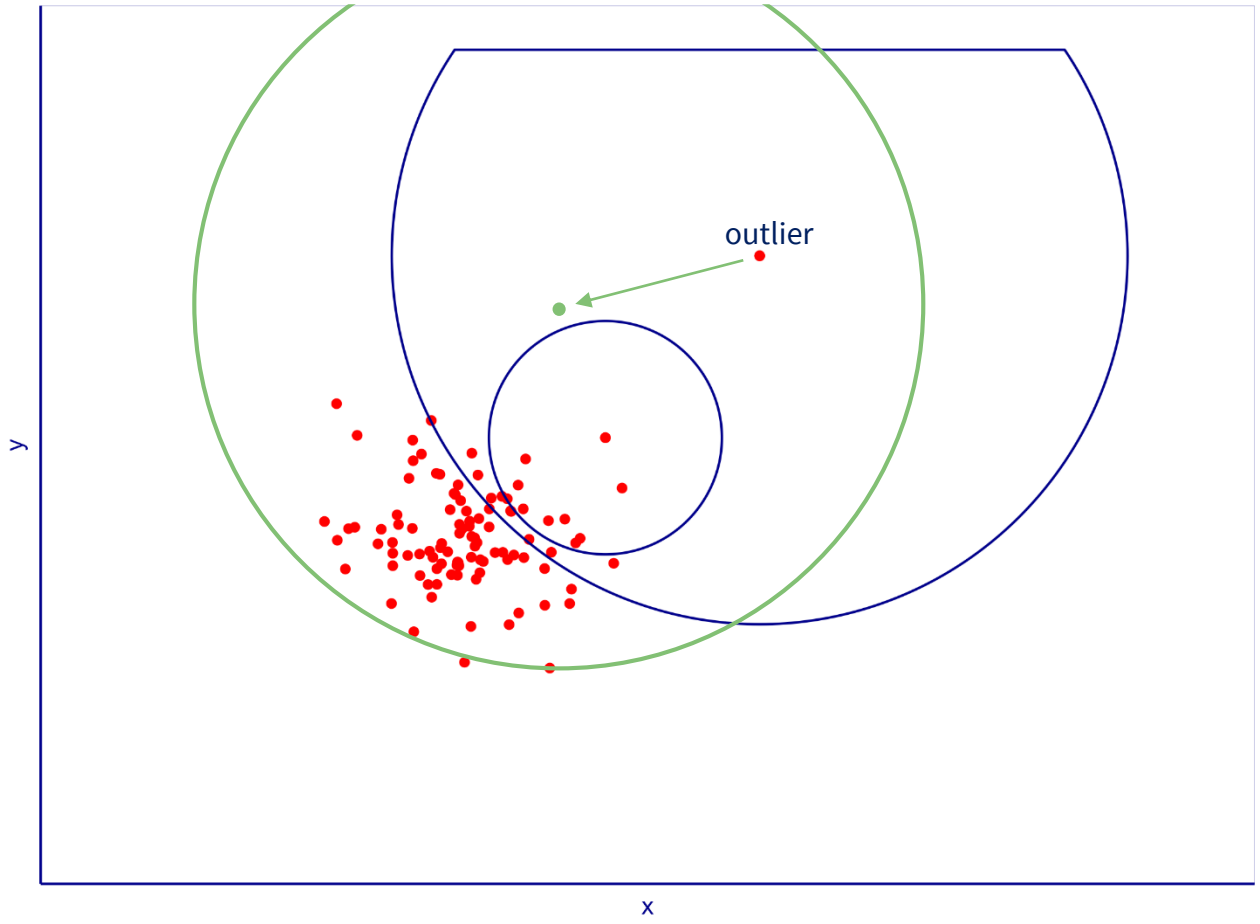
RISK ASSESSMENT

- K-anonymity
  - Definition: a record must be indistinguishable from at least $k - 1$ other records
  - Spatial k-anonymity for masking methods: measure the number of locations within a radius equal to the displacement distance
    (1) number of locations around the **original** point
    (2) number of locations around the **masked** location
  - Problems with this measurement
- Alternatives
  - Record linkage attacks (Drechsler and Hu, 2021; Quick et al. 2015)
  - Assessment of overfitting regarding spatial outliers (Quick et al. 2018)

SPATIAL K-ANONYMITY EXAMPLE



SPATIAL K-ANONYMITY EXAMPLE



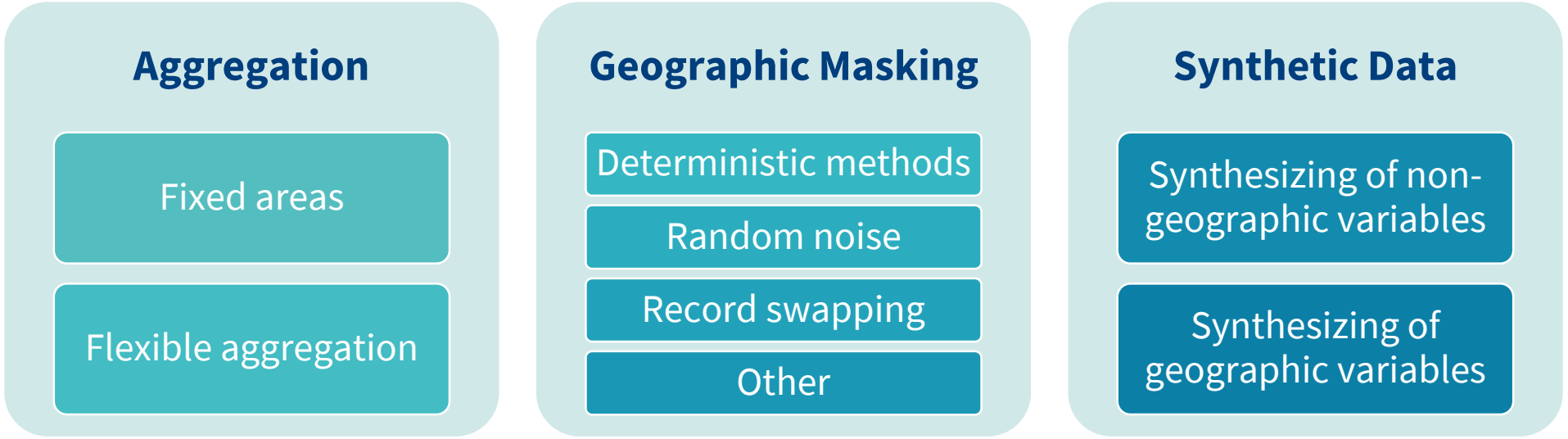
UTILITY ASSESSMENT

Comparison of original and anonymized data

1. Point locations and density measures
   - Distances between original and masked locations
   - Heatmaps using Kernel density estimation
2. Clustering
3. Spatial autocorrelation
4. Applied results

CONCLUSION

- Three main strands of confidentiality protecting strategies



- Some common masking techniques do not provide adequate confidentiality protection
- Common risk measures should be carefully evaluated

KEY REFERENCES

Drechsler, J. and J. Hu (2021). Synthesizing Geocodes to Facilitate Access to Detailed Geographical Information in Large-Scale Administrative Data. *Journal of Survey Statistics and Methodology 9*(3), 523–548.

Koebe, T., A. Arias-Salazar, and T. Schmid (2023). Releasing survey microdata with exact cluster locations and additional privacy safeguards. *Humanities and Social Sciences Communications 10*(1), 1–13.

Kounadi, O. and M. Leitner (2016). Adaptive areal elimination (aae): A transparent way of disclosing protected spatial datasets. *Computers, Environment and Urban Systems 57*, 59–67.

Paiva, T., A. Chakraborty, J. Reiter, and A. Gelfand (2014). Imputation of confidential data sets with spatial locations using disease mapping models. *Statistics in medicine 33*(11), 1928–1945.

Quick, H. (2021). Generating poisson-distributed differentially private synthetic data. *Journal of the Royal Statistical Society Series A: Statistics in Society 184*(3), 1093–1108.

Quick, H. (2022). Improving the utility of poisson-distributed, differentially private synthetic data via prior predictive truncation with an application to cdc wonder. *Journal of Survey Statistics and Methodology 10*(3), 596–617.

Quick, H., S. H. Holan, C. K. Wikle, and J. P. Reiter (2015). Bayesian marked point process modeling for generating fully synthetic public use data with point-referenced geography. *Spatial Statistics 14*, 439–451.

Quick, H., S. H. Holan, and C. K. Wikle (2018). Generating partially synthetic geocoded public use data with decreased disclosure risk by using differential smoothing. *Journal of the Royal Statistical Society Series A: Statistics in Society 181*(3), 649–661.

Quick, H. and L. A. Waller (2018). Using spatiotemporal models to generate synthetic data for public use. *Spatial and Spatio-Temporal Epidemiology 27*, 37–45.

Sakshaug, J. W. and T. E. Raghunathan (2010). Synthetic data for small area estimation. In J. Domingo-Ferrer and E. Magkos (Eds.), *Privacy in Statistical Databases*, Berlin, Heidelberg, pp. 162–173. Springer Berlin Heidelberg.

Sakshaug, J. W. and T. E. Raghunathan (2014). Generating synthetic data to produce public-use microdata for small geographic areas based on complex sample survey data with application to the national health interview survey. *Journal of Applied Statistics 41*(10), 2103–2122.

Wang, H. and J. P. Reiter (2012). Multiple imputation for sharing precise geographies in public use data. *The annals of applied statistics 6*(1), 229.

Zhang, S., Freundschuh, S. M., Lenzer, K., & Zandbergen, P. A. (2017). The location swapping method for geomasking. *Cartography and Geographic Information Science*, *44*(1), 22-34.

Zhou, Y., F. Dominici, and T. A. Louis (2010). A smoothing approach for masking spatial data. *The Annals of Applied Statistics 4*(3), 1451–1475. DOI: 10.1214/09-AOAS325.

CONTACT

Maike Steffen

maike.steffen@iab.de