

Anonymization for integrated and georeferenced Data (AnigeD)

Maurice Brandt¹, Markus Zwick¹

¹Federal Statistical Office Germany (Destatis)

Abstract

Data-based information is necessary for decision making in politics, business, science and public life. With digitization and the exponential growth of new data sources, stored data, as well as new analytical methods such as machine learning, the possibilities for evidence-based decision making have expanded and evolved significantly.

A key challenge in integrating disparate data sets from different data custodians is the protection of personal privacy and trade secrets within organizations. This currently hinders both the wider use of data as a product and the use of integrated data in policy advice and scientific research. Methods for anonymization and statistical confidentiality face the challenge of finding a compromise. On the one hand, they need to protect the information of the data subjects, while on the other hand, the chosen methods should still offer sufficient analysis and information potential for the anonymized data. Anonymization and confidentiality of individual data leads to information reduction. For new data sources like from Mobile Network Operators (MNO-data) and satellite-data the detailed regional information is usually the object of interest. The regional information is also one of the most critical one for reidentification in official statistics.

The challenge is to maintain the detailed regional level of georeferenced data and providing useful datasets for analysis. This can be done with anonymization and pseudonymization of data or geoinformation, which limits the risk of detection to an acceptable level while preserving sufficient analytical potential for wider use and value creation.

The AnigeD competence cluster is part of the "Research Network Anonymization for Secure Data Use" of the German Federal Ministry of Education and Research (BMBF) within the framework of the Federal Government's IT security research program "Digital. Secure. Sovereign". The thematic focus, which is supported by various research strands, is the further and new development of strategies for the protection of personal and company-related data when using complex integrated data sets. Not only the integration of different data via direct identifiers or probabilities is relevant, but also the integration and linking of data via regional information in the form of georeferencing.

The talk will introduce the challenges in the anonymization process and its possible impact on the quality of statistical products and publications. It will summarize the research results of the cluster in regard to the main research within the AnigeD project. It will present the evaluation of anonymized data according to formal criteria, anonymization of georeferenced data, and an example for MNO-data to apply the anonymization directly within the secured environment of a big European MNO.

Keywords: anonymisation, georeferenced data, integrated data