Anonymization for integrated and georeferenced Data (AnigeD)

Jannek Muehlhan[1], Markus Zwick[1]

[1]Federal Statistical Office Germany (Destatis)

Abstract

Data-based information plays a central role in politics, business, science and public life. With digitization and the exponential growth of 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.

In the past it has been shown that common anonymization strategies for individual data in economic statistics led to de facto or absolutely anonymized data sets, which were severely limited for scientific analyses due to the reduced or even distorted information potential. Anonymization and pseudonymization of data, which limits the risk of detection to an acceptable level while preserving sufficient analytical potential, is therefore essential 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 further describe the main research strands within the AnigeD project: 1) Evaluation of anonymized data according to formal criteria, 2) Anonymization through synthetic data, 3) Anonymization of georeferenced data, and 4) Open software tools for anonymization.

Keywords: Anonymisation, synthetic data, georeferenced data, integrated data

1. Introduction

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

The COVID-19 crisis highlighted that many valuable data sets exist in principle, but are often held in a decentralised manner in different silos by different actors, whether in companies or public institutions. At the same time, advances in big data, also referred to as non-traditional data, have shown that the greatest value comes especially when different non-traditional data sources are combined with traditional data, such as surveys and administrative data. Individual data sets are often only pieces of a puzzle, unable to paint a complete picture.

A key challenge in integrating disparate data sets from different data custodians is the protection of personal privacy and trade secrets within organisations. 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 anonymisation 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 anonymised data. Anonymisation and confidentiality of individual data go hand in hand with information reduction. Either information is suppressed to such an extent that attribution is no longer possible, or only with considerable effort, or this protection is achieved by modifying the data to such an extent that the usefulness of the data is partially lost. In either case, there is a loss of information. The application of anonymisation methods therefore always faces a dual optimisation problem. On the one hand, the analysis potential of the data should be maximised for a given protection requirement; on the other hand, the anonymisation concept should be adapted to a given necessary information content. In the past it has been shown that common anonymisation strategies for individual data in economic statistics led to de facto or absolutely anonymised data sets, which were severely limited for scientific analyses due to the reduced or even distorted information potential.

Anonymisation and pseudonymisation of data, which limits the risk of detection to an acceptable level while preserving sufficient analytical potential, is therefore essential for wider use and value creation.

The cluster is part of the "Research Network Anonymisation 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 programme "Digital. Secure. Sovereign". It is funded by the European Union - NextGenerationEU.

The AnigeD competence cluster is divided into the following research areas:

- Evaluation of anonymised data according to formal criteria.
- Anonymisation through synthetic data
- Anonymisation of georeferenced data
- Open software tools for anonymisation.

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.

2. Motivation:

Starting point: Anonymising data in such a way that the remaining information does not allow any conclusions to be drawn about individual data subjects (be they for example persons, households or companies), but still contains sufficient information potential, is a core concern of every data producer, whether private or public. In addition to various legal regulations (EU-DSGVO, BDSG, BStatG), the quality of the data products is of particular importance.

Methods for anonymisation or statistical confidentiality have to resolve a conflict of objectives. On the one hand, the information provided by the data subjects must be protected; on the other hand, the procedures must be chosen in such a way that the anonymised data still have sufficient potential for analysis or information. Anonymising and guaranteeing the confidentiality of individual data generally involves a reduction of information and thus a loss of information. The Federal Statistical Office already has extensive experience in anonymising large amounts of data.

Anonymisation through the use of synthetic data: The demand for greater availability and transparency of data, while maintaining confidentiality and data protection, can only be met by innovative methods of data processing, preparation and delivery. The use of classical anonymisation methods reaches its limits with increasing complexity and number of data usage requests. Synthetic data are a promising alternative to the publication of anonymised individual data in many areas. At the international level, there are already first applications by the national statistical authorities of New Zealand, Canada, Scotland and the United States of America, among others. The use of synthetic data to anonymize personal data has found wider application so far, as documented in the literature mentioned above.

The cluster builds on previous research that has addressed, among other things, the de facto anonymity of economic statistics. In the case of economic statistics data, these methods are sometimes limited by oligopolitical market structures, and there have been few applications for georeferenced data. Georeferenced data offer new possibilities for merging heterogeneous data. According to § 10 para. 3 BStatG, individual statistical data with regional information can be integrated on a hectare level. 3 BStatG, which allows for detailed regional information, but here too only a few anonymisation approaches have been developed for such integrated data. In this respect, AnigeD is expected to provide new insights that will be of great interest, especially for the commercial use of the data.

Anonymisation of georeferenced data: Georeferenced data has extraordinary potential for research and teaching as well as for public administration and the economy. Key questions about the future and sustainability of our society can only be answered with high-quality and accessible georeferenced data. On the one hand, georeferenced data is available when existing databases are expanded to include geocoded information. This area is growing rapidly. With the increasing amount of geocoded information available, more and more complex data sets can be designed by integrating geoinformation. Adequate methods for securing individual markers while maintaining the necessary amount of analysis potential are not yet available for either simple or highly complex georeferenced data.

The data protection problems with georeferenced data are currently usually circumvented by restricting the data to very high aggregation levels, which is associated with a considerable loss of information. Often, in-depth analyses cannot be carried out at all with the accessible data. In order to create improved data access options here, the development of privacy metrics for geodata, for example, is indispensable. In addition, privacy record linkage procedures have been developed in the literature, with which geodata can be stored as Bloom filters in individual records and used for linkage or distance calculation. The security of this and other methods of encrypting geocoordinates has so far been little investigated. The project is intended to make a contribution to this, especially with regard to the problems of statistical confidentiality caused by enriched data records. Within the scope of this subproject, new data sets with considerable potential for regional analyses, such as mobile phone data, will be considered. New procedures are to be developed here, but existing procedures and methods for anonymising data will also be further developed.

Anonymisation of mobile phone data: Concrete preliminary work has been done in the area of mobile phone signal data in recent years. Since 2017, the Federal Statistical Office has been researching possible applications of mobile phone signal data in official statistics. Within this framework, several studies have been carried out on different application purposes and quality

aspects. This has resulted in several modular software packages for geolocation, deduplication and aggregation of activities (see 'Mobile network data' of the ESSnet Big Data I and II project). In addition, the European Statistical System is working on the concrete technical implementation of privacy-compliant processing of mobile network data and on process models for cooperation between private data providers and official statistics. The implementation of such a process offers official statistics, and thus also research, society and politics, the possibility of making long-term statements on longitudinal changes in population distribution and mobility (e.g. long-term intra-German migration patterns, analysis of the effects of new forms of work and the development of sustainable means of transport).

3. Methods:

Within the framework of the research project "Anonymisation of official statistics through synthetic data", three lines of action are highlighted. The first line of action focuses on exploring the possible uses of synthetic data for the research data centres of the Federal Government and the Länder. In this context, methods for the (partly) automated creation of synthetic datasets will be developed and tested. These synthetic datasets will be used in various applications, such as data exploration, writing and testing of analysis programs, teaching, and anonymisation of particularly sensitive features and geocoordinates. It will also explore whether synthetic data can expand the range of data recipients, such as data journalists.

The second storyline looks at the potential of high-quality synthetic datasets for the way public and private data producers work to produce and publish aggregated results. Here we explore whether synthetic or partially synthetic data can be used directly in the production of results to resolve trade-offs between protecting confidentiality and making statistical results widely and flexibly available.

The third strand will systematically compare different approaches to synthetic data production. In particular, the extent to which the methods developed are also suitable for statistical analyses such as regression analyses will be investigated. It will be investigated how statistical approaches can be used in the context of machine learning and vice versa. In addition, existing approaches will be methodologically refined to address possible weaknesses, e.g. in the use of deep learning methods from computer science.

Even if the extensive use of synthetic data for the direct production of results is not always possible for quality reasons, there are scenarios where the use of synthetic data offers advantages. For all storylines, the standardisability of synthetic data generation and the effort involved is crucial.

The project will also develop privacy record linkage methods that allow geocoord data to be stored as Bloom filters in individual records and used for linkage or distance calculations. The security of these methods for encoding geocoordinates will be investigated, especially with regard to the problems of statistical secrecy caused by enriched datasets.
The plan is to align the environment term with typical applications or analysis models for the target data and then balance the two (usually conflicting) goals: Maximising the analysis potential and minimising the risk of reidentification.
The Chair of Statistics at the Department of Economics of the FU Berlin has developed advanced methods for the analysis of anonymised georeferenced data in cooperation with the company INWT Statistics. The focus of anonymisation is to reduce the accuracy of the georeferenced data in order to make it difficult or impossible to identify individual units in a dataset. Nevertheless, the dataset should remain usable for content related evaluations. This subproject deals with the use of anonymised georeferenced data and the limitations of anonymisation. Statistical methods will be developed that both consider the anonymisation process and enable typical evaluations of georeferenced data. These procedures will be demonstrated for different application areas. At the same time, user-friendly open source software will be developed for these applications.
Statistical procedures will be developed that allow for smooth map representations that are not bound to a specific area system, but are still compatible with the anonymised area values. The aim is to adapt the statistical evaluation of georeferenced data to the anonymisation procedure and to make the use of anonymised georeferenced data sets more efficient. To this end, adapted statistical estimation procedures will be developed and supported by open source software to facilitate their use by a wide range of users.
In order to make sound predictions about the capabilities of a potential attacker, a consistent formalisation of the material criteria specified by the legal system is required. To accomplish this legally and technically challenging task, the University Speyer adopts a research approach that measures the extent to which the provision of a data set increases the likelihood that an attacker will obtain new information about the data subject. This approach is based on the recognition that any natural person is already exposed to some basic risk from data that is generally accessible or available to a potential attacker, and that this risk remains even if the entity holding the data refrains from publishing or sharing it.
Another question that the University Speyer addresses is how the publication or dissemination of the dataset affects the pre-existing baseline risk. The University Speyer's approach is to examine existing proposals for measuring risk shift, taking into account their compatibility with the legal system and practice. In particular, two approaches will be considered: Differential

Privacy (DP) and GDA Score. However, it is not enough to merely measure the shift of the basic risk. In a third step, the University Speyer therefore plans to investigate in more detail the maximum extent to which the basic risk can be shifted so that the data-holder can legitimately assume that it is only passing on anonymised data.
The software system Diffix will be used as a demonstrator for the processing, evaluation and analysis of the data within the framework of the research tasks. It is used for the technical implementation of the anonymisation methods developed in the cluster. The aim is to make the best use of Diffix as a stand-alone application and as part of other programming languages such as Python and R, or anonymisation packages, to enable feasible solutions.

4. Aims:

AnigeD aims to advance current anonymisation methods and to identify and implement new solutions for new problems. This should not only secure but also extend the current state of data access for science. The methods developed and researched in the cluster will be made available not only to the project partners involved, but also to data-holding companies. In this way, the developed and new methods can generate added value for the companies on the one hand, and expand access to company data for science and official statistics on the other.
The main objective of AnigeD is to secure and expand access to complex data while protecting individual characteristics, and to create greater legal certainty for practitioners. Given the exponential growth of data volumes and the increasing complexity of data, especially in the context of georeferencing, current strategies for protecting individual identifiers are reaching their limits. Therefore, a sub goal of AnigeD is to secure and expand the supply of (complex) data for science in the research data network of research data centres.
In addition, existing methods will be further developed in cooperation with companies from the data industry and made available for commercial purposes. In this way, insights and applications developed for science through public funding of data access will also be opened up for data-driven business models. At the same time, data from the companies will be made available for use in science and society, with appropriate protection of feature carriers and trade secrets.
The talk at the European Conference on Quality in Official Statistics 2024 will give an overview over the research cluster “Anonymisation of integrated and georeferenced data” (AnigeD) and introduce different anonymisation techniques of georeferenced data and their potential in the use for NSIs.