UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS
Expert Meeting on Statistical Data Confidentiality
26-28 September 2023, Wiesbaden

Overview of the AnigeD Project and Potentials of Data Synthesis for Official Statistics and Research

Yannik Garcia Ritz, Safiyye Aydin, Jannek Mühlhan, Markus Zwick, (Federal Statistical Office, Germany)
Yannik.GarciaRitz@destatis.de
Safiyye.Aydin@destatis.de
Jannek. Mühlhan@destatis.de
Markus.Zwick@destatis.de

Abstract

Statistical Disclosure Control for integrated and georeferenced data is a new challenge for statistical institutes. New digital data in combination with traditional data offer many new analysis possibilities. Moreover, these complex datasets are usually georeferenced in a very fine-grained way. Traditional confidentiality procedures reach their limits here. Destatis, the German Federal Statistical Office, is working together with various universities on the further development of existing procedures in order to ensure the protection of individuals even for complex data. The lecture will present the first results of the project "Anonymization for integrated and georeferenced Data" (AnigeD) funded by the German Ministry of Research.

As part of AnigeD, Destatis further deals with dataset synthetization of population as well as economic statistics to examine probable potentials for data provision to the research community as answer to the growing scientific interest in official data. Confidentiality issues lead to several measures to ensure confidentiality of respondent units. Consequently, there is a diametral relationship between level of anonymity and analytic potential of provided datasets which may not completely satisfy the needs of the scientific community. Research on data synthetization is currently suspecting synthetic datasets to be a probable solution to this problem due to their artificial nature. With the following research an official statistics dataset will be synthesized and evaluated regarding analytical utility as well as the level of confidentiality. Furthermore, an evaluation regarding the ease of use of certain provision methods of synthetic datasets will be presented.

1 Introduction

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.

The COVID-19 crisis highlighted that many valuable data sets exist in principle, but are often held in a decentralized 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 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 go hand in hand with 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". It is funded by the European Union – NextGenerationEU. 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 AnigeD competence cluster is divided into the following research areas

- Formalization of substantive criteria for the success of anonymization provided by the legal system.
- Anonymization through synthetic data
- Anonymization of georeferenced data
- Evaluation of anonymized data according to formal criteria.
- Open software tools for anonymization

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 present paper presents insight from the research area of anonymization through data synthesis. Therefore, by synthesizing parts of the Structure of Earnings Survey 2018, a non-georeferenced database is chosen to gain further insights into the potentials of anonymization and provision of the data synthesis of official databases.

Previous research of *Loske & Wolfanger* (2019), *Hafner & Lenz* (2011) dealt with the synthesis of official data structural files, while *Templ* (2017) synthesized simulated data of the SES 2014. In contrast to the mentioned prior research approaches, the present work deals with the partial data synthesis of the on-site material of the company and employee datasets of the Structure of Earnings Survey 2018.

In addition to increasing the anonymity of respondent units through data synthesis, the potentially provided synthetic data needs to comply with high-quality requirements placed on official data (Zwick, 2016). Thus, the generated partially synthetic data material is evaluated concerning the attained global and analytic utility. Finally, this paper will provide a weighted evaluation of the data synthesis with respect to the anonymization and analysis potentials.

2 Background

Official data products and statistics face a steady increase in demand by the scientific community and the public, in general (Allin, 2021). Anonymising data in such a way that the remaining information does not allow any conclusions to be drawn about individual data subjects (be they 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 anonymization 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 anonymized data still have sufficient potential for analysis or information. Anonymizing 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 anonymizing large amounts of data (i.e., Ronning et al. (2005), Hundepool et al. (2012), Templ (2017)). In general, provided data is anonymized to a greater or lesser extent. In case of application of less anonymization measures the way of data access is made more difficult (Rothe, 2015).

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 anonymization methods reaches its limits with increasing complexity and number of data usage requests. Synthetic data offer opportunities to optimize the aspects of anonymization because respective measures can be integrated into the synthesis process (Drechsler & Haensch, 2023).

For on-site access, the full data material, except for direct identifiers, is provided to the scientific community. However, scientific data users must use the data either physically (safe center usage) or virtually (remote execution) at the providing institution. In contrast, off-site data can be used in the scientific institution of the contractor (Rothe, 2015).

The methodology of synthetic data generation is based on the principles of multiple imputation (Rubin, 1993). However, instead of only estimating values to replace missing values, false declarations etc., estimations are used to replace some or all variables of the original dataset (Little, 1993). Thus, there is a methodological distinction between the concepts of full (Rubin, 1993) and partial synthesis (Little, 1993). Only sensitive variables or variables which increase the reidentification risks are synthesized as part of partial synthesis. *Little* (1993) argues that focusing only on sensitive or reidentification risk increasing variables should prevent the analysis quality from being reduced too much by the estimation character of the synthesis approach. The possible, integrable anonymization measures can make an important contribution to balancing the protection of the respondents and ensuring of the analysis potential (Drechsler & Haensch, 2023).

The synthetic data should reflect the structure and relationships of the original data as closely as possible. Simultaneously, the level of anonymity should be increased in comparison to the original on-site material (Reiter, 2023). The Statistical Offices of the Federation and the Federal States are obliged to protect the respondent units according to Section 16 (1), Federal Statistics Act. At the same time, the must comply with the scientific privilege derived from Section 16 (6), Federal Statistics Act. Thus, the Federal Statistical Offices of the Federation and the Federal States founded the Research Data Centers (RDCs) in 2001 to enable scientific access to official data (Zühlke, Zwick, & Scharnhorst, 2003).

As part of the AnigeD project and competence cluster, one working package deals with the assessment of the supply potential of synthetic on-site material. 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.[1] The use of synthetic data to anonymize personal data has found wider application so far, as documented in the literature mentioned above (Burnett-Isaacs et al., 2021).

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 section 3 BStatG, individual statistical data with regional information can be integrated on a hectare level. § 10 section 3 BStatG, which allows for detailed regional information, but here too only a few anonymization 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.

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

[1] Burnett-Isaacs et al. (2021)

statistics (Hadam, Schmid, & Simm, 2020). 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.

Within the framework of the research project "Anonymization 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 RDCs of the Statistical Office of the Federation and the Federal States. 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 anonymization 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 semi-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.

So far, the RDCs do not provide synthetic data. Previous research regarding synthesis of official data dealt with data structural files (Loske & Wolfanger, 2019; Hafner & Lenz, 2011) or with simulations of official data (Templ, 2017). 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 standardizability of synthetic data generation and the effort involved is crucial.

The project will also develop privacy record linkage methods that allow geocoordinated 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: Maximizing the analysis potential and minimizing the risk of re-identification.

The Chair of Statistics at the Department of Economics of the FU Berlin has developed advanced methods for the analysis of anonymized georeferenced data in cooperation with the company INWT Statistics. The focus of anonymization 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 anonymized georeferenced data and the limitations of anonymization. Statistical methods will be developed that both take into account the anonymization 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 anonymized area values. The aim is to adapt the statistical evaluation of georeferenced data to the anonymization procedure and to make the use of anonymized 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 formalization of the material criteria specified by the legal system is required. To accomplish this legally and technically challenging task, the DUV (German University of Administrative Sciences Speyer, german: Deutsche Universität für Verwaltungswissenschaften 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 DUV addresses is how the publication or dissemination of the dataset affects the pre-existing baseline risk. The DUV'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 DUV 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 anonymized 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 anonymization 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 anonymization packages, to enable feasible solutions.

Aims:

AnigeD aims to advance current anonymization 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 RDCs.

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 mod-els. 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.

This paper reports first results from the research of the AnigeD project on the evaluation of the potentials of synthetic data for the scientific community as well as for the providing RDCs of the Statistical Offices of the Federation and the Federal States. Therefore, the company and the employee file of the Structure of Earnings Survey (SES) 2018 serve as base for several synthesis approaches and the respective evaluations regarding disclosure risks and utility of the generated synthetic data. The following section elaborates on the conceptualization of the synthesis approach and the subsequent assessment of the disclosure risks and utility of the synthetic data generated.

3 Conceptualization

Following the argumentation of *Little* (1993), the concept of partial synthesis is used to synthesize the on-site material of the SES 2018. The SES 2018 comprises a company and an employee dataset which are both partially synthesized as part of the present work. Various statistical techniques and machine learning approaches can be used to conduct data synthesis (Drechsler & Haensch). Research findings of *Grinsztajn, Oyallon, & Varoquaux* (2022) indicate that Classification And Regression Trees (CARTs) outperform conventional statistical techniques and other machine learning approaches in many occasions. Thus, CARTS are predominantly used to synthesize the two on-site datasets of the SES 2018.

Furthermore, data synthesis enables data providers to make use of different smoothing approaches as anonymization measure for variables obtaining highly skewed distributions. The resulting reduction in estimation accuracy leads to an increase in the level of anonymity (Nowok, Raab, Dibben, Snoke & van Lissa, 2022; Drechsler & Reiter, 2008). *Reiter* (2005) identifies several reasons for providing multiple synthetic datasets per original dataset. *Drechsler* (2009) suggests to provide at least as many synthetic datasets per original dataset as

the number of original datasets. In the present work, five synthetic datasets are generated based on the original company and original employee dataset, each. Hence, the minimal criterion of m ≥ r 8 (Drechsler, 2009) is complied with.

Following the partial data synthesis carried out, the generated partially synthetic datasets are checked concerning their disclosure risks. Here *k*-anonymity (Sweeney, 2002) is used as one measure to quantify the number of observations violating *k*=2 or *k*=3 anonymity (Templ, 2017) and the number of high-risk observations (Templ 2017). The mentioned key measures are calculated as ratios to the key measures of the respective off-site material as denominator. For baseline evaluations and to enable comparisons, the same is done for the original on-site material of the company and employee datasets of the SES 2018.

$$\hat{r}_k = \frac{\hat{p}_k}{1-\hat{p}_k} log\left(\frac{1}{\hat{p}_k}\right) \mid f_k = 1 \quad (1)$$

$$\hat{r}_k = \frac{\hat{p}_k}{1-\hat{p}_k} - \left(\frac{\hat{p}_k}{1-\hat{p}_k}\right)^2 log\left(\frac{1}{\hat{p}_k}\right) \mid f_k = 2 \quad (2)$$

$$\hat{r}_k = \frac{\hat{p}_k}{f_k - (1-\hat{p}_k)} \quad (3)$$

Observations are classified as high-risk observations if their estimated individual risk $\hat{r}_k$ is higher than 10 % and larger than the median individual risk $\hat{r}_k$ + factor δ times the median absolute deviation of $\hat{r}_k$ (δ ≥ 2; Templ, 2017).

Moreover, the generated partially synthetic data is evaluated regarding the number of expected random matches as well as the absolute number of true and false matches to the original data (Drechsler & Reiter, 2008).

- **Expected Match Risk** for a selection based on a random guess:

$$\sum_{j \in T} \left(\frac{1}{c_j}\right) * I_j \quad (4)$$

- **True Match Rate** for true matches of targets K_j among all matches identified within c_j units exemplarily examined:

$$\Sigma_{j \in T} K_j \Big/ \Sigma_{j \in T}(c_j = 1) \quad (5)$$

- **False Match Rate** for the share of incorrectly assumed matches within c_j units exemplarily examined:

$$1 - \left(\Sigma_{j \in T} K_j \Big/ \Sigma_{j \in T}(c_j = 1)\right) \quad (6)$$

Considering the high-quality standards for official data, it is important to further evaluate the analytic potential of the generated partially synthetic SES 2018 data, in addition to the disclosure risk assessment. Consequently, the generated partially synthetic company and employee datasets are examined concerning their global and model specific utility. Variable transformations according to *Raghunathan, Lepkowski, Van Hoewyk & Solenberger* (2001) are used to ensure compliance with basic logical constraints on variable relationships. Furthermore, descriptive statistics and distributions of the generated partially synthetic and the original data of both files of the

SES 2018 are compared to assess global utility. Finally, the propensity Mean-Squared Error (pMSE) is used as a final measure for global utility to rate the similarity of the generated partially synthetic datasets and the original database.

$$pMSE = \frac{1}{m}\sum_{j}^{m}\left(\frac{1}{N}\sum_{i}^{N}(\hat{p}_i - c)^2\right) \quad (7)$$

A model specific utility evaluation provides deeper insights to the usefulness of the partially synthetic company and employee datasets of the SES 2018. Considering that many research questions in the scientific community are worked with several models, underlines the importance of a model specific utility evaluation even further. The confidence interval overlap is a measure to assess the model specific utility and serves as an indicator for the accuracy of estimates obtained from models which are estimated on synthetic data (Karr, Kohnen, Oganian, Reiter, & Sanil 2006). Hence, the present work estimates exemplary linear and logistic regression models for partially synthetic company and employee data material, each. These exemplary regression models are used to estimate the average confidence interval overlap over all coefficients just as the separate confidence interval overlap.

$$J_k = \frac{1}{2} * \left[\frac{U_{over,k} - L_{over,k}}{U_{orig,k} - L_{orig,k}} + \frac{U_{over,k} - L_{over,k}}{U_{synth,k} - L_{synth,k}}\right] \quad (8)$$

4 Results

4.1 Disclosure Risk Evaluation

Spline smoothing has proven to be the best approach for the data synthesis of the *company data* of the SES 2018 regarding the cost-benefit ratio of disclosure risks and global/model specific utility. The evaluation of disclosure risks is executed by comparing k-anonymity key measures as well as the number of high-risk observations building up on *Templ* (2017), as already described in section 3. Contrary to first expectations increases in the mentioned key figures for the generated synthetic data material are recorded. Nevertheless, it needs to be underlined that the increase results through the data synthesis, so there is actually no increase in real high-risk observations which implies that the pool of partially synthetic high-risk observations is larger compared to the respective numbers in the original data material. It is believed that this is more likely to indicate increased security in terms of confidentiality, since the risk of finding a truly high-risk observation should decrease.

The examination of the key measures of *Drechsler & Reiter* (2008) slightly support this assumption because they reveal that a random disclosure only arises with a probability of less than 0.1 %. Furthermore, it turns out that, there is no true match to be observed in the generated partially synthetic company data.

In contrast to the company data, the best cost-benefit-ratio regarding disclosure risks and global/model specific data utility is achieved for the partially synthetic *employee data* if kernel density smoothing is applied to synthesis highly skewed variables (e.g., income-related variables). The data synthesis model which uses spline smoothing

leads to a noteworthy underestimation of outliers for the variable gross monthly income. Analogous to the disclosure risk evaluation of the company data, an adapted form of the approach of *Templ* (2017) is used in the first step. Thereby, increases in the ratios of the respective key measures are observed as well. However, these increases are less high compared to the increases observable for the partially synthetic company data.

In the second step, disclosure risks are again further evaluated by examining the expected match risk and the true match rate (Drechsler & Reiter, 2008). In contrast to the synthetic company data, there is no risk expected for a random match. Moreover, this is also observed for the true match rate indicating that all observations considered to be matching to original observations are actually false matches.

4.2 Utility Evaluation

As described in section 3 the utility of the generated partially synthetic data material based on the company and employee file of the SES 2018 is assessed both globally as well as specifically for exemplary regression models. Variable transformations as described by *Raghunathan, Lepkowski, Van Hoewyk & Solenberger* (2001) ensure that basic boundary values constraints are met. Thus, enabling to directly start with comparisons of the original and respective synthetic data material of the SES 2018 for global utility assessment. It can be observed that the basic descriptive statistical key measures (mean, median and standard deviation) are reflected well in the generated synthetic company and employee material.

Additionally, data utility is assessed by examining the mean pMSE for the partially synthetic company and employee material of the SES 2018. The mean pMSE of 0.1142 lies in the middle of the possible interval which indicates a still existing potential for utility improvement.

An exemplary linear regression model is estimated alongside an exemplary logistic regression model to evaluate the model specific utility of the generated partially synthetic company data. The exemplary linear regression model estimates potential effects of the craft affiliation of a company, participation of the public sector in company's capital as well as the number of common working days per week on the company's number of employees (see Table 1). In the next step, the average confidence interval of the exemplary synthetic data-based estimates is computed in relation to their counterparts of the original data over all $m = 5$ partially synthetic datasets. The average confidence interval overlap equals 85 % over all estimates of the exemplary linear regression. Observing a confidence interval overlap around 62 % for the explanatory variable "participation of the public sector in the company's capital" reveals that the overall mean confidence interval overlap of the exemplary linear regression model is negatively impacted by the CI of the estimate.

Table 1: Comparison of coefficients and confidence intervals of an exemplary linear regression on variable "Number of Employees" with both original and synthesized on-site company dataset of the SES 2018.

	Original on-site material (employee dataset) Coefficient (Std. error)	Synthesized on-site material (employee dataset) Coefficient (Std. error)	CI Overlap
Intercept	-97.7958*** (36.4197)	-92.9597** (36.4198)	0.9661
Craft affiliation	20.8993*** (2.8277)	19.6029*** (2.8277)	0.8830
Participation of the public sector in company's capital	223.6840*** (11.2451)	207.0839*** (11.2451)	0.6234
Working days per week	-17.4124*** (6.6388)	-15.5643 ** (6.6388)	0.9290

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

Table 2: Comparison of coefficients and confidence intervals of an exemplary logit regression on variable "collective bargaining" with both original and synthesized on-site company dataset of the SES 2018.

	Original on-site material (employee dataset) Coefficient (Std. error)	Synthesized on-site material (employee dataset) Coefficient (Std. error)	CI Overlap
Intercept	-0.6124*** (0.17298)	-0.8626*** (0.1730)	0.6311
Craft affiliation	-0.2445*** (0.01304)	-0.2471*** (0.0130)	0.9481
Number of Employees	0.00003*** (0.0000)	0.00003*** (0.0000)	0.7657
Working Days per Week	-0.1116*** (0.0336)	-0.0624* (0.0336)	0.6266
Type of corporate entity = Operation of a multi-business enterprise	1.5793*** (0.0349)	1.5705*** (0.0349)	0.9353
Type of corporate entity = Operation of a multi-country enterprise	1.5240*** (0.0265)	1.5723*** (0.0265)	0.5344

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

An exemplary logistic model estimates effects of several explanatory effects on a previously created binary variable that indicates whether wages are determined primarily on the basis of collective bargaining agreements (see Table 2). The highest confidence interval overlap is estimated for the coefficients of the explanatory variables "Craft affiliation" (94.81 %) and "Type of corporate entity = Operation of a multi-business enterprise" (93.53 %). However, both mentioned explanatory variables are the only variables exceeding 90 % and approximating the target value of 95 % confidence interval overlap.

Consequently, the overall mean confidence interval overlap for the exemplary logistic model, which is used to further evaluate the model specific utility of the generated synthetic company data, equals 74 %. Therefore, there is an even higher deviation for the confidence interval of explanatory variables in the exemplary logistic regression model in comparison to the explanatory variables in the exemplary linear regression model.

All in all, confidence interval overlaps of 85 % and 74 % suggests a good similarity between the exemplary linear and logistic regression based on the original and the generated partially synthetic data. Nevertheless, a further increase through extended tuning of the synthesis models for the company data is expected to achieve confidence interval overlaps close to 95 %.

In contrast to the partially synthetic company data, the partially synthetic *employee data* is estimated by making use of kernel density smoothing. The thereby generated partially synthetic employee data is assessed by comparing the descriptive statistics with the respective counterparts of the original data. This examination reveals that the descriptive key measures as well as the distribution of the monthly gross income is similarly well met as the key figures for the company data set. The same is true for the pMSE (0.1110) which is equally close to the pMSE of the company dataset. Consequently, this suggests that further tuning of the data synthesis model could also lead to a further increase in data utility in this case.

Table 3: Comparison of coefficients and confidence intervals of an exemplary linear regression on variable gross hourly income with both original and synthesized on-site employee dataset of the SES 2018.

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

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

Table 4: Comparison of coefficients and confidence intervals of an exemplary logit regression on variable "Gross Hourly Wages Above Minimal Wage" with both original and synthesized on-site employee dataset of the SES 2018.

	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

Moreover, an exemplary linear regression model is estimated on the generated variable gross hourly income (see Table 3). Additionally, an exemplary logistic regression model is estimated on a binary variable indicating whether the gross hourly income of an employee exceeds the minimal wage (see Table 4). Assessing the model-specific utility reveals that there is no mean confidence overlap to be observed, in average, for both exemplary regression models estimated for the partially synthetic and the original employee data.

5 Discussion

Limitations

The present research on synthesis potential of the SES 2018 does not check and ensure cress-file references between the company and employee file. It is likely that respective logical constraints need to be considered in future synthesis models. A similar train of thought is followed for the use case of panel data, since the present work only examines only a single survey year. It cannot be ruled out completely that relations in longitudinal context are not reflected accurately.

In the present research, it is dealt with a partial synthesis of the employee dataset and the company of the SES 2018. Consequently, the results are only valid for the examined datasets. It cannot be ruled out completely that the findings for the partial data synthesis of the SES 2018 files cannot be generalized for other official surveys such as microcensus, DRG, for example.

The examination of the partially synthesized SES 2018 datasets does only lead to suggestive conclusion that the partial synthesis led to a decrease in risks of deanonymization looking at *k*-anonymity and number of high-risks observations. It is theorized that an increase in the respective key measures after partial data synthesis reflects an increase of the pool of high-risk observations containing values which do not necessarily match the original data. Assessing the low expected match risks and true match rates of both partially synthesized company and employee files provides, further support for this hypothesis. Nevertheless, it needs to be acknowledged that there is yet no actual linkage between the estimated key measures yet.

Looking at the results it needs to be acknowledged that for both the employee and company file of the SES 2018, the disclosure risks and model-specific utility-related key measures do not meet the expectations. Thus, the current results do not provide evidence that the present generated partially synthetic data is ready for provision to the scientific community.

In principal, the assessment of all key measures provided in this paper is only offering a personal appraisal of partial synthetic data provision by official statistical offices. The present thesis is not to be considered as a legal report on how to deal with the provision of synthesized data material but is only offering a personal appraisal of potentials.

Research and Practical Implications

It is believed that existing cross-file references may not be accurately reflected in the partially synthesized SES 2018 data. This is to be investigated as part of further work on the third research area of the AnigeD project. If there are limitations concerning cross-file references, respective constraints need to be integrated into the partial synthesis models.

Additionally, future research should illuminate the utility and disclosure risk evaluations about panel data which has not yet been covered by the presented work. Since the scientific community is often interested in longitudinal research questions, it needs to be made sure that respective relations are reflected correctly, as well.

Furthermore, future research should check on the hypothesis that the increase in high-risk observations after partial data synthesis actually reflects a larger pool of untruthful high-risk observations, indicating lower disclosure risks. As part of this examination, it should also be exposed whether the key measures of Templ (2017) in combination with the key measures of Drechsler & Reiter (2008) could be harmonized.

Since the present work deals only with the evaluation of potentials of data synthesis of the on-site material of SES 2018, the generalizability of the presented findings for other official surveys should be investigated to provide lawyers with the knowledge needed for their legal assessment on simplified data access of less anonymized synthesized data to the scientific community.

In addition, future research should tie up to the present work. Work should be done to further improve the key figures, which have not been satisfactory in some places to date so that publication of synthetic data can be examined by the legal authorities in the future and implemented if necessary.

Current research results indicate potentials to increase confidentiality and keep the structure of original official survey data by making use of partial data synthesis of key and target variables of the SES 2018. However, the more precise examination of utility specific key measures (pMSE and confidence interval overlap) suggests that the partial data synthesis models need to be tuned before a release of partially synthetic SES 2018 data can be considered. Hyperparameter optimization seems to be a beneficial approach fur future data synthesis, enabling a structured search for hyperparameters which are able to maximize the desired result of utility metrics (Bergstra & Bengio, 2012).

The presented work is no legal report on the legal possibility of providing easier data access to less conservatively anonymized official data. Even a positive evaluation for the use case of partially synthesized data of the SES 2018 does not allow to make use of this approach for other survey years of the SES or other statistics. Consequently, a continuous legal monitoring needs to be implemented as soon as new research insights on the potential of synthesized official data are available.

Conclusion

All in all, the, so far, the generated partially synthetic data does not allow be made publicly available because they do not meet the expectations of the scientific community concerning the utility. Future research should focus on examining how to further increase the utility of the present partially synthetic on-site material of SES 2018. Only after that, a legal evaluation regarding the provision possibilities on simplified ways of access for the generated partially synthetic data on the SES 2018 is possible. Future research should also deal with further official statistics and panel data to further increase the knowledge on synthesis potentials for official on-site data.

6 Bibliography

Abowd, J., Stinson, M., & Benedetto, G. (2006). *Final Report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project.* Technical Report, US Census Bureau.

Allin, P. (2021, November). Opportunities and challenges for official statistics in a digital society. *Contemporary Social Science, 16*(2), pp. 156-169. doi:10.1080/21582041.2019.1687931

Brandt, M., Crößmann, A., & Gürke, C. (2011). Harmonisation of statistical confidentiality in the Federal Republic of Germany. *FDZ_Arbeitspapiere, 34*, pp. 1-12.

Caiola, G., & Reiter, J. P. (2010). Random Forests for Generating Partially Synthetic, Categorical Data. *Transactions on Data Privacy, 3*, pp. 27-42.

Drechsler, J. (2009). SYNTHETIC DATASETS FOR THE GERMAN IAB ESTABLISHMENT PANEL. *Joint UNECE/Eurostat work session on statistical data confidentiality*, (pp. 1-12). Bilbao, Spain.

Drechsler, J. (2011). Multiple imputation in practice—a case study using a complex German establishment survey. *AStA Advances in Statistical Analysis, 95*, pp. 1-26. doi:DOI 10.1007/s10182-010-0136-z

Drechsler, J., & Haensch, A.-C. (2023). 30 years of synthetic data. *arXiv:2304.02107*, pp. 1-42. doi:10.48550/arXiv:2304.02107

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

Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder. (2020a). *Metadatenreport. Teil I: Allgemeine und methodische Informationen zur Verdienststrukturerhebung 2018.* Metadatenreport, Düsseldorf.

Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder. (2020b). *Metadatenreport. Teil II: Produktspezifische Informationen zur Nutzung der Verdienststrukturerhebung 2018 per On-Site-Nutzung.* Metadatenreport, Wiesbaden. doi: 10.21242/62111.2018.00.00.1.1.0

Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder. (2022). Regelungen zur Auswertung in den Forschungsdatenzentren der Statistischen Ämter. (F. d. Länder, Ed.) pp. 1-25.

Hadam, S., Schmid, T., & Simm, J. (2020). Kleinräumige Prädiktion von Bevölkerungszahlen basierend auf Mobilfunkdaten aus Deutschland. In B. Klumpe, J. Schröder, & M. Zwick (Eds.), *Qualität bei zusammengeführten Daten* (pp. 31-48). Wiesbaden: Springer VS. doi:https://doi.org/10.1007/978-3-658-31009-7_3

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.

Hu, J., & Hoshino, N. (2018). The quasi-multinomial synthesizer for categorical data. *International Conference on Privacy in Statistical Databases* (pp. 75-91). Springer.

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

Kitchin, R. (2015). The opportunities, challenges and risks of big data for official statistics. *Statistical Journal of the IAOS, 31*, pp. 471-487. doi:DOI 10.3233/SJI-150906

Kursa, M., & Rudnicki, W. (2010). Feature Selection with the Boruta Package. *Journal of Statistical, 36*(11). Retrieved from https://doi.org/10.18637/jss.v036.i11

Manrique-Vallier, D., & Hu, J. (2018). Bayesian non-parametric generation of fully synthetic multivariate categorical data in the presence of structural zeros. *Journal of the Royal Statistical Society, 181*(3), pp. 635-647.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of Input. *Statistical Science, 9*(4), pp. 538-573.

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).

Pierson, S. (2015). Official statistics principles compared. *Statistical Journal of the IAOS, 31*, pp. 21-23. doi:10.3233/SJI-150886

Pistner, M., Slavkovic, A., & Vilhuber, L. (2018). Synthetic data via quantile regression for heavy-tailed and heteroskedastic data. In J. Domingo-Ferrer, & F. Montes (Ed.), *International Conference on Privacy in Statistical Databases* (pp. 92-108). Springer. doi:https://doi.org/10.1007/978-3-319-99771-1_7

Raab, G. M., Nowok, B., & Dibben, C. (2016). Practical data synthesis for large samples. *Journal of Privacy and Confidentiality, 7*(3), pp. 67-97. doi:https://doi.org/10.29012/jpc.v7i3.407

Reiter, J. P. (2005). Releasing multiply imputed, synthetic public use microdata: an illustration and empirical study. *Journal of the Royal Statistical Society, 168*(1), pp. 185-205.

Reiter, J. P. (2023). Synthetic Data: A Look Back and A Look Forward. *Transactions On Data Privacy, 16*, pp. 15-24.

Rothe, D. (2015, 05). 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.

Schäfer, A., & Gottschall, K. (2015). From wage regulation to wage gap: how. *Cambridge Journal of Economics, 39*, pp. 467-496. doi:doi:10.1093/cje/bev005

Statistisches Bundesamt. (2020a). Verdienststrukturerhebung - Niveau, Verteilung und Zusammensetzung der Verdienste und der Arbeitszeiten abhängiger Beschäftigungsverhältnisse - Ergebnisse für Deutschland - . *Fachserie, 16*(1), pp. 1-525.

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

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software, 45*(3), pp. 1-67.

van der Voort, H. G., Klievink, A. J., Arnaboldi, M., & Meijer, A. J. (2019, January). Rationality and politics of algorithms. Will the promise of big data survive the dynamics of public decision making? *Government Information Quarterly, 36*(1), pp. 27-38. doi:10.1016/j.giq.2018.10.011

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. (2001). Die Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder. (S. Bundesamt, Ed.) *Wirtschaft und Statistik, 10*, pp. 906-911.

Zühlke, S., Zwick, M., & Scharnhorst, S. (2003). Die Forschungsdatentren der Statistischen Ämter des Bundes und der Länder. (S. Bundesamt, Ed.) *Wirtschaft und Statistik 10*, pp. 906-911.

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.

Zwick, M. (2016). Big Data und amtliche Statistik. In B. Keller, H. Klein, & S. Tuschl, *Marktforschung der Zukunft - Mensch oder Maschine?* (pp. 157-172). Wiesbaden: Springer Gabler. doi:https://doi.org/10.1007/978-3-658-14539-2_10