Brenzel, Hanna; Ritz, Yannik Garcia

Article

Partielle Synthetisierung zur Anonymisierung der verknüpften Verdienststrukturerhebung 2018. Teil 1: Methodik und Evaluation der Datennützlichkeit

WISTA - Wirtschaft und Statistik

Provided in Cooperation with:
Statistisches Bundesamt (Destatis), Wiesbaden

Suggested Citation: Brenzel, Hanna; Ritz, Yannik Garcia (2025) : Partielle Synthetisierung zur Anonymisierung der verknüpften Verdienststrukturerhebung 2018. Teil 1: Methodik und Evaluation der Datennützlichkeit, WISTA - Wirtschaft und Statistik, ISSN 1619-2907, Statistisches Bundesamt (Destatis), Wiesbaden, Vol. 77, Iss. 4, pp. 83-94

PARTIELLE SYNTHETISIERUNG ZUR ANONYMISIERUNG DER VERKNÜPFTEN VERDIENSTSTRUKTUR-ERHEBUNG 2018

Teil 1: Methodik und Evaluation der Datennützlichkeit

Hanna Brenzel, Yannik Garcia Ritz

↘ **Schlüsselwörter:** partiell synthetische Daten – verknüpfte synthetische Daten – Automatisierung – verknüpfte Arbeitgeber-Arbeitnehmer-Daten – Nützlichkeitsbewertung

ZUSAMMENFASSUNG

Der Einsatz von synthetischen Daten wird als Alternative gehandelt, um das Spannungsfeld zwischen Erhalt des Analysepotenzials bei gleichzeitiger Gewährleistung des Schutzes der Daten optimal aufzulösen. Der Beitrag geht dieser Frage nach mit der Verdienststrukturerhebung 2018 als Datenbasis. Sie setzt sich aus einem Betriebsdatensatz und einem Angestelltendatensatz zusammen und bildet damit einen Linked-Employer-Employee-Datensatz. Bei verknüpften Datensätzen besteht die Schwierigkeit für die unabhängige Synthetisierung darin, die gewohnte Konsistenz zwischen den Datensätzen zu gewährleisten. Mithilfe eines schrittweisen partiellen multiplen Synthetisierungsansatzes konnte diese Konsistenz sichergestellt werden. Die Ergebnisse bezüglich der Nützlichkeit der generierten Daten sind vielversprechend.

↘ ***Keywords:*** *partially synthetic data – linked synthetic data – automation – linked employer-employee data – utility evaluation*

ABSTRACT

The use of synthetic data is considered a viable alternative in order to resolve the tension between the competing demands of maintaining the analytical potential of data while guaranteeing data protection. This article examines this issue, using the 2018 Structure of Earnings Survey (SES) as a data basis. The SES comprises an enterprise dataset and an employee dataset, which together form a linked employer-employee dataset. With linked datasets, the challenge for independent data synthesis is to ensure the customary consistency between the datasets. This was possible with the help of a partial multiple data synthesis approach, which was implemented on a step-by-step basis. The results regarding the utility of the generated data are promising.

Dr. Hanna Brenzel

studierte internationale Volkswirtschaft mit Ausrichtung auf Mittel- und Osteuropa an der Universität Regensburg. Sie leitet das Referat „Forschungsdatenzentrum, Methoden der Datenanalyse“ des Statistischen Bundesamtes.

Yannik Garcia Ritz

studierte nach seiner Ausbildung zum Fachangestellten für Markt- und Sozialforschung Wirtschaftswissenschaften an der Johann Wolfgang Goethe-Universität Frankfurt am Main (Bachelor of Science) sowie der Johannes Gutenberg-Universität Mainz (Master of Science). Er arbeitet aktuell als Wissenschaftlicher Mitarbeiter im Referat „Forschungsdatenzentrum, Methoden der Datenanalyse“ des Statistischen Bundesamtes.

Hinweis zur Förderung

Diese Forschung wurde durch das Cluster „Anonymisierung bei integrierten und georeferenzierten Daten (AnigeD)“ und seinen Träger, das Bundesministerium für Forschung, Technologie und Raumfahrt (BMFTR), gefördert.

1

Einleitung

In einer digitalisierten Welt steigt das Angebot an Daten. Dies fordert zunehmend den Bedarf an evidenzbasierter Politikberatung und damit die Nachfrage nach einem einfachen Datenzugang: sei es für Wissenschaft, Politik, Journalismus, Lehre oder die Gesellschaft im Allgemeinen. Gleichzeitig erhöht sich die Gefahr der Re-Identifikation durch die Fülle an öffentlich zugänglichen Daten, sodass der Schutz geheim zu haltender Daten komplexer und immer aufwendiger und schwieriger wird.

Häufig kommen traditionelle Anonymisierungsmethoden wie die Vergröberung oder das Entfernen von Merkmalen sowie die Ziehung von Stichproben an ihre Grenzen. Eine zu starke Anonymisierung verringert das Analysepotenzial der Daten, sodass nur eingeschränkte Fragestellungen mit den Datensätzen zu beantworten sind. Eine zu geringe Anonymisierung der Daten führt hingegen zu einem erhöhten Aufdeckungsrisiko. In diesem Spannungsfeld befinden sich Datenproduzenten und datenhaltende Stellen, die ihre Daten im Zwecke des Allgemeinwohls zugänglich machen wollen. Dies sowie größere Rechenleistungen haben insbesondere die Erstellung von synthetischen Daten als alternative Anonymisierungsmethodik in den letzten Jahren gefördert (Drechsler/Haensch, 2023).

Das durch das Bundesministerium für Forschung, Technologie und Raumfahrt sowie durch die Europäische Union (EU) im Rahmen des Konjunkturpakets „Next GenerationEU“ geförderte Forschungsnetzwerk „Anonymisierung für eine sichere Datennutzung“ und damit einhergehend das Forschungscluster „Anonymisierung bei integrierten und georeferenzierten Daten“ (AnigeD) greift dieses Spannungsfeld auf. Ziel des Forschungsclusters ist „gegenwärtig verwendete Anonymisierungsmethoden für georeferenzierte und verknüpfte Daten weiterzuentwickeln. Dies soll den gegenwärtigen Stand des Zugangs der Wissenschaft zu Einzeldaten nicht nur sichern, sondern auch ausbauen“[1].

Beteiligt im Forschungscluster AnigeD sind neben dem Statistischen Bundesamt die Technische Hochschule Köln, die Universität der Bundeswehr München, die Freie Universität Berlin sowie das Institut für Arbeitsmarkt- und Berufsforschung.

Der Beitrag ist dem dritten Arbeitspaket des Forschungsclusters zugeordnet, das sich mit der Evaluation und Weiterentwicklung von Methoden der Synthetisierung sowie mit deren Automatisierung beschäftigt.

Das Forschungsdatenzentrum des Statistischen Bundesamtes stellt gemeinsam mit dem Forschungsdatenzentrum der Statistischen Ämter der Länder[2] seit mehr als 20 Jahren Hochschulen oder sonstigen Einrichtungen mit der Aufgabe unabhängiger wissenschaftlicher Forschung (§ 16 Absatz 6 Bundesstatistikgesetz) Mikrodaten bereit (Brenzel/Zwick, 2022).

Die aktuelle Rechtsgrundlage nach § 16 Absatz 6 Nr. 1 Bundesstatistikgesetz erlaubt den Forschungsdatenzentren, faktisch anonymes Datenmaterial an die geprüften Einrichtungen zu übermitteln. Innerhalb speziell abgesicherter Bereiche darf der Zugang zu formal anonymen Daten gewährt werden (§ 16 Absatz 6 Nr. 2 Bundesstatistikgesetz). Letzteres ist mit einem relativ hohen Aufwand sowohl der Forschungsdatenzentren als auch der Datennutzenden verbunden. Bei einem Zugang über den sogenannten Gastwissenschaftsarbeitsplatz an einem der deutschlandweit derzeit 22 Standorte der Forschungsdatenzentren müssen die Datennutzenden Fahrtkosten sowie gegebenenfalls Übernachtungskosten aufbringen, die Forschungsdatenzentren müssen die entsprechenden Räumlichkeiten vorhalten. Bei der sogenannten kontrollierten Datenfernverarbeitung als Alternative zum Gastwissenschaftsarbeitsplatz hingegen liegt der Aufwand insbesondere auf Seiten der Mitarbeitenden der Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder, da diese die Auswertungsprogramme für die Wissenschaftlerinnen und Wissenschaftler manuell ausführen müssen. Im Anschluss durchlaufen alle Ergebnisse vor Übermittlung an die Datennutzenden eine Geheimhaltungsprüfung auf Einhaltung der absoluten Anonymität. ↘ Grafik 1

Ein deutlich geringerer kontinuierlicher Betreuungsaufwand entsteht bei den sogenannten faktisch anonymen Scientific-Use-Files, die an die zugangsberechtigte Einrichtung übermittelt werden dürfen. Hier liegt die

1 Mehr zum Thema unter: www.destatis.de

2 Nachfolgend kann „Forschungsdatenzentrum“ sowohl das Forschungsdatenzentrum des Statistischen Bundesamtes als auch das Forschungsdatenzentrum der Statistischen Ämter der Länder bezeichnen.

Grafik 1

Eigenschaften, Anonymisierungsgrad und Analysepotenzial der vom Forschungsdatenzentrum bereitgestellten Nutzungswege

Anonymisierung

	On-Site-Nutzung		Neu	Off-Site-Nutzung	
Zugangsweg	Kontrollierte Daten-fernverarbeitung	Gastwissenschafts-arbeitsplätze	Remote Access	Scientific-Use-Files	Public-Use-Files/ Campus-Files
Anonymisierungsgrad der Daten	formal anonym	formal anonym	faktisch anonym	faktisch anonym	absolut anonym
Nutzungsberechtigt	wissenschaftliche Einrichtung	wissenschaftliche Einrichtung	wissenschaftliche Einrichtung	wissenschaftliche Einrichtung	alle
Datenhaltung während der Nutzung	statistische Ämter	statistische Ämter	statistische Ämter	wissenschaftliche Einrichtung	beliebig
Aufenthaltsort der Nutzenden während der Nutzung	beliebig	statistische Ämter	wissenschaftliche Einrichtung	wissenschaftliche Einrichtung	beliebig

Analysepotenzial

Komplexität insbesondere beim Erstellen des Datenmaterials. Einerseits muss gewährleistet sein, dass die Einzelangaben in einem Datensatz nur mit einem unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft den Betroffenen zugeordnet werden können (§ 16 Absatz 6 Nr. 1 Bundesstatistikgesetz). Andererseits sinkt durch die Anonymisierung das Analysepotenzial der Daten, wodurch diese nur für eine eingeschränkte Zahl an Forschungsfragen verwertbar sind. Nimmt das Angebot öffentlich zugänglicher Informationen zu, steht mehr zusätzliches Wissen zur Verfügung, was eine potenzielle Re-Identifikation vereinfachen könnte. Im Umkehrschluss könnte künftig ein stärkerer Eingriff im Sinne einer Informationsreduktion in die Daten erforderlich sein, um einen faktisch anonymen Datensatz zu erhalten.

Wie erwähnt wird der Einsatz von synthetischen Daten als eine vielversprechende Alternative gehandelt, um das Spannungsfeld zwischen Erhalt des Analysepotenzials bei gleichzeitiger Gewährleistung des Schutzes der Daten optimal aufzulösen. Im Vergleich zu traditionellen Anonymisierungsmethoden birgt das Synthetisieren die Chance, die Analysemöglichkeiten weniger stark einzuschränken, ohne ein erheblich größeres Risiko einer Offenlegung von Identitäten und Verhältnissen der Auskunftgebenden zu erzeugen. Die entscheidende Frage ist daher: Lässt sich ein synthetischer Datensatz erstellen, der bei gesteigertem Analysepotenzial im Vergleich mit traditionellen Anonymisierungsmethoden die Anforderung an faktische Anonymität erfüllt und gleichzeitig näherungsweise Analyseergebnisse liefert, die mit den Originaldaten erzielt worden wären?

Der Artikel geht dieser Frage anhand eines konkreten und bestehenden Datensatzes der Forschungsdatenzentren nach. Kapitel 2 beschreibt das zugrunde liegende Datenmaterial. Auf die Synthetisierungsmethodik geht Kapitel 3 ausführlich ein, bevor in Kapitel 4 der Fokus darauf liegt, wie die Nützlichkeit des synthetisierten Datensatzes bewertet werden kann. Limitationen und Implikationen beleuchtet das abschließende fünfte Kapitel.

2

Datenbasis

Als Basisdatensatz wurde das formal anonyme Datenmaterial der Verdienststrukturerhebung 2018[3] herangezogen. Die Verdienststrukturerhebung setzt sich aus einem Betriebsdatensatz und einem Angestelltendatensatz zusammen und bildet damit einen Linked-Employer-Employee-Datensatz. Die Daten werden überwiegend dezentral mittels einer Stichprobe von den Statistischen Ämtern der Länder erhoben und durch das Statistische Bundesamt zu einem deutschlandweiten Datensatz aufbereitet. Der Original-Betriebsdatensatz umfasst 71 000 Betriebe, der Original-Angestelltendatensatz mehr als 1 Million Beschäftigte (Forschungsdatenzentren, 2020). Die beiden Teildatensätze lassen sich über die Betriebsnummern aus dem Unternehmensregistersystem (URS) zusammenfassen.[4] Eine Anspielung von detaillierten Unternehmensinformationen aus den Betriebsdaten an die Angestelltendaten ermöglicht es beispielsweise, Lohnunterschiede aufgrund von Unternehmens- und Branchencharakteristika zu untersuchen oder die entsprechenden Informationen als Kontrollvariablen in Regressionsmodelle einzubeziehen.

3

Synthetisierungsmethodik

Die Idee der Synthetisierung von Daten als Anonymisierungsmethode beruht auf den Prinzipien der Imputation. Im Gegensatz zur klassischen Imputation, bei der in der Regel Schätzergebnisse nur einzelne Beobachtungspunkte (fehlende Werte, implausible Angaben oder Ähnliches) ersetzen, verfolgt die Synthetisierung den Ansatz, ganze Variablen oder Beobachtungen durch Schätzergebnisse zu ersetzen. Werden alle Variablen und alle Beobachtungen innerhalb eines Datensatzes imputiert, spricht man von der vollen Synthese nach Rubin (1993). Wird nur ein Teil der Variablen durch Schätzergebnisse ersetzt, spricht man von einer partiellen Synthese. Dieser von Little (1993) vorgebrachte Ansatz ermöglicht es beispielsweise, sich bei der Synthetisierung ausschließlich auf die Anonymisierung sensibler Informationen beziehungsweise Variablen zu beschränken. Dafür kommen insbesondere Schlüssel- und Zielvariablen infrage. Als Schlüsselvariablen gelten Variable, die in Kombination eine hohe Wahrscheinlichkeit besitzen, bekannt zu sein und zum Zwecke einer Re-Identifikation von Berichtseinheiten genutzt zu werden (beispielsweise Geschlecht, Alter, Bildungsabschluss). Zielvariablen sind dagegen insbesondere sensible Informationen, welche nach der Re-Identifikation von Berichtseinheiten offengelegt werden könnten (beispielsweise einkommensbezogene Verhältnisse). Eine Beschränkung auf die genannten Variablengruppen bei der partiellen Synthetisierung bietet folglich die Chance, so viele Informationen wie möglich im Original zu belassen und für die Wissenschaft bereitzustellen. Vor diesem Hintergrund wurde die Verdienststrukturerhebung 2018 lediglich partiell synthetisiert, um die Interessen der Wissenschaft sowie der auskunftgebenden Berichtseinheiten bestmöglich zu gewichten und damit nur die Daten soweit zu verfremden, wie der Schutz der Daten dies erfordert.

Eine Analyse von Little und anderen (2021) zeigt, dass die Synthetisierung mittels Classification And Regression Trees (CART) eine Optimierung der Datennützlichkeit bei gleichzeitig adäquater Risikoreduzierung bietet. Die Autoren verwenden dabei für ihre Umsetzung das R-Paket synthpop von Raab und anderen (2016). Dieses ermöglicht eine anwenderfreundliche Synthetisierung unter Einsatz unterschiedlichster Methoden, wobei CART als Standardeinstellung im R-Paket verwendet wird. Die Schätzgenauigkeit der synthetisierten Informationen lässt sich unter anderem durch die Glättung metrischer Variablen, aber auch durch die Baumtiefe beeinflussen. Zudem bietet die Generierung von mehreren (partiell) synthetischen Datensätzen die Möglichkeit, die auf diesen Daten basierenden Schätzergebnisse zu stabilisieren und die Nützlichkeit damit zu erhöhen (Reiter, 2008; Drechsler, 2009).

Der Linked-Employer-Employee-Datensatz der Verdienststrukturerhebung 2018 enthält in beiden Datensätzen übergreifende Informationen. Dazu zählen beispielsweise die Information über das Bundesland, in welchem der jeweilige Betrieb ansässig ist, sowie Bezüge zwischen

3 DOI: 10.21242/62111.2018.00.00.1.1.0

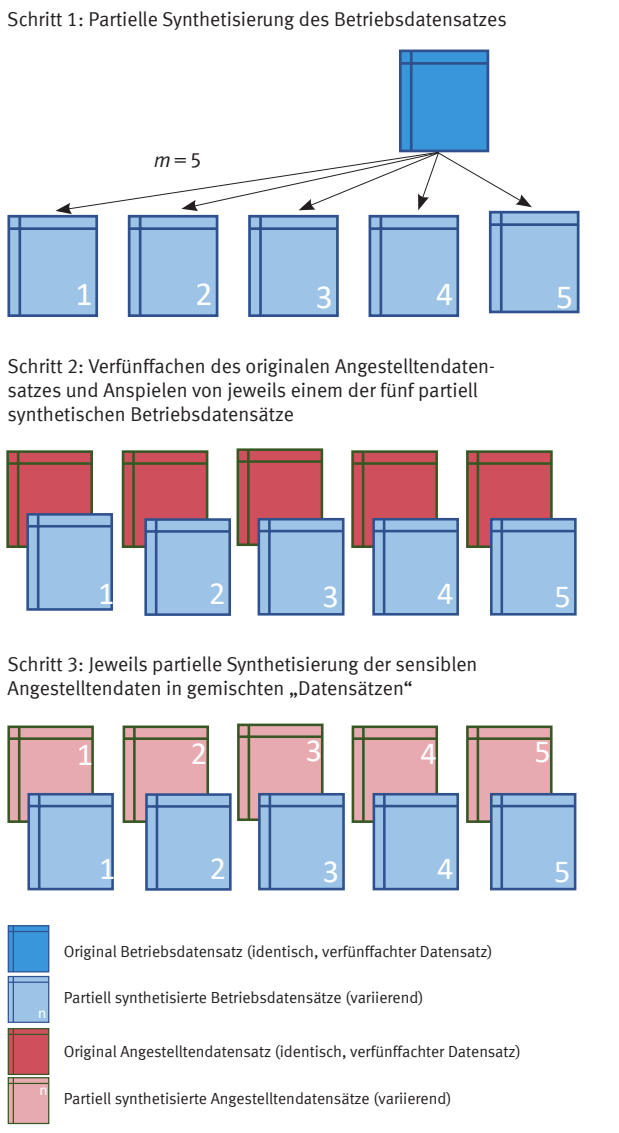
4 Zu beachten ist, dass bei der Zusammenfassung der fehlenden URS-Verknüpfungen einige Berichtseinheiten auf Angestelltenebene verloren gehen. Dies gilt sowohl für die Originaldaten als auch später für die partiell synthetischen Daten.

der Anzahl der Angestellten im Unternehmensdatensatz sowie der absoluten Anzahl im Betriebsdatensatz. Im Falle einer separaten Synthetisierung dieser übergreifenden Informationen sind logische Inkonsistenzen aufgrund der unterschiedlichen einfließenden erklärenden Variablen nicht nur nicht auszuschließen, sondern sehr wahrscheinlich. Um diese Problematik zu vermeiden, wurde ein schrittweiser Synthetisierungsansatz entwickelt: Zunächst werden fünf partiell synthetische Datensätze für den Betriebsdatensatz generiert (Schritt 1). Diese werden jeweils an einen der verfünffachten Original-Angestelltendatensätze angespielt (Schritt 2). In diesen fünf Datensätzen, die sich aus dem partiell synthetisierten Betriebsdatensatz sowie dem Original-Angestelltendatensatz zusammensetzen, werden nun jeweils sensible Schlüssel- und Zielvariablen des Original-Angestelltendatensatzes synthetisiert, sodass final fünf partiell synthetische Gesamtdatensätze zur Verdienststrukturerhebung 2018 entstehen (Schritt 3). Hierbei werden alle einkommensbezogenen Informationen (Brutto-/Nettoeinkommen, steuerliche und Sozialabgaben und so weiter) als Zielvariablen betrachtet und folglich synthetisiert. ↘ Grafik 2

Schlüsselvariablen sind unter anderem Geschlecht, Alter, Dauer der Unternehmenszugehörigkeit sowie Bildungsabschlüsse; diese werden ebenfalls synthetisiert.

Im Zuge der Synthetisierung können verschiedene Hyperparameter gesetzt werden. Für eine Hyperparameteroptimierung wurden mithilfe einer Rastersuche (grid search) im Laufe des Forschungsprozesses folgende Hyperparameter hinsichtlich ihres Nützlichkeits- und Risikoverhältnisses für Angestellten- und Betriebsdatensatz (jeweils auf Basis einer Stichprobe) unabhängig voneinander untersucht: die Anzahl der zu generierenden partiell synthetischen Datensätze, die Anzahl der minimal vorzuhaltenden Beobachtungen für das letzte Blatt im CART sowie die Anwendung von Glättungsverfahren. Als Kennzahlen für die Bewertung der Datennützlichkeit dienten der Propensity Score Mean-Squared-Error und der Confidence Interval Overlap für die Koeffizienten einer beispielhaften Regressionsanalyse, für die Risikobewertung dagegen Expected Match Risk und True Match Rate (Reiter/Mitra, 2009). Auf Basis der Ergebnisse wurde sich dafür entschieden, $m = 5$ partiell synthetische Datensätze zu generieren und auf Glättungsverfahren zu verzichten sowie mindestens drei Beobachtungen für die letzten Blätter der CART vorzuhalten.

Grafik 2
Schrittweises Vorgehen zur Synthetisierung der verknüpften Daten der Verdienststrukturerhebung 2018



4

Nützlichkeitsbewertung

Um die Nützlichkeit generierter synthetischer Datensätze zu bewerten, wird üblicherweise zwischen globaler Nützlichkeit und modellspezifischer Nützlichkeit unterschieden.

Die Bewertung der globalen Nützlichkeit erfolgt typischerweise anhand der Replikation allgemeiner statistischer Basiskennzahlen (Lage- und Streuungsmaße) sowie der Verteilungen (Drechsler/Haensch, 2023). Die Betrachtung der Lagemaße zeigt, dass insbesondere für die synthetisierten Variablen aus dem ursprünglichen Angestelltendatensatz sowohl arithmetisches Mittel als auch Median sowie die Quartilswerte sehr nah an den Originalwerten liegen. Auch die Standardabweichung wird, gemittelt über alle $m = 5$ partiell synthetischen Datensätze, sehr gut für diese Variablen wiedergegeben. Diese Kennzahlen werden für die Variable Unternehmensgröße (Anzahl der Mitarbeitenden), die ursprünglich aus dem Betriebsdatensatz stammt, auf Beschäftigtenebene unterschätzt. ↘ Tabelle 1 Dieses Phänomen tritt unabhängig davon auf, ob eine Glättung vorgenommen wird oder nicht und ebenfalls unabhängig davon, ob man den verknüpften Datensatz oder singulär den Betriebsdatensatz betrachtet. Hier sollte künftig geprüft werden, ob statt CART eine alternative Synthetisierungsmethode für diese Variablen bessere Ergebnisse liefert.

Tabelle 1
Vergleiche deskriptiver Statistiken für exemplarische Variablen aus Originaldatensatz und partiell synthetischen Datensätzen

	Originaldaten	Partiell synthetische Daten
	Bruttomonatseinkommen	
Mittelwert	2 709,00	2 707,72
Median	2 365,00	2 365,00
Standardabweichung	2 590,98	2 585,50
P25 | P75	1 206 | 3 562	1 206 | 3 561,40
	Nettomonatsverdienst	
Mittelwert	1 811,43	1 811,84
Median	1 629,00	1 629,60
Standardabweichung	1 554,90	1 565,58
P25 | P75	908 | 2 336	910,40 | 2 341,60
	Geschlecht	
Mittelwert	1,45	1,45
Median	1,00	1,00
Standardabweichung	0,4973	0,4973
P25 | P75	1 | 2	1 | 2
	Geburtsjahr	
Mittelwert	1 974,65	1 974,65
Median	1 974,00	1 974,00
Standardabweichung	13,28	13,28
P25 | P75	1 964 | 1 986	1 964 | 1 986
	Jahr des Unternehmenseintritts	
Mittelwert	2 009,21	2 009,29
Median	2 013,00	2 013,00
Standardabweichung	9,72	9,53
P25 | P75	2 005 | 2 016	2 005 | 2 016
	Unternehmensgröße	
Mittelwert	4 339,354	3 708,452
Median	121,00	76,40
Standardabweichung	21 712,43	19 080,55
P25 | P75	27 | 581	12 | 459,20

Berechnungen auf Basis der Verdienststrukturerhebung 2018.

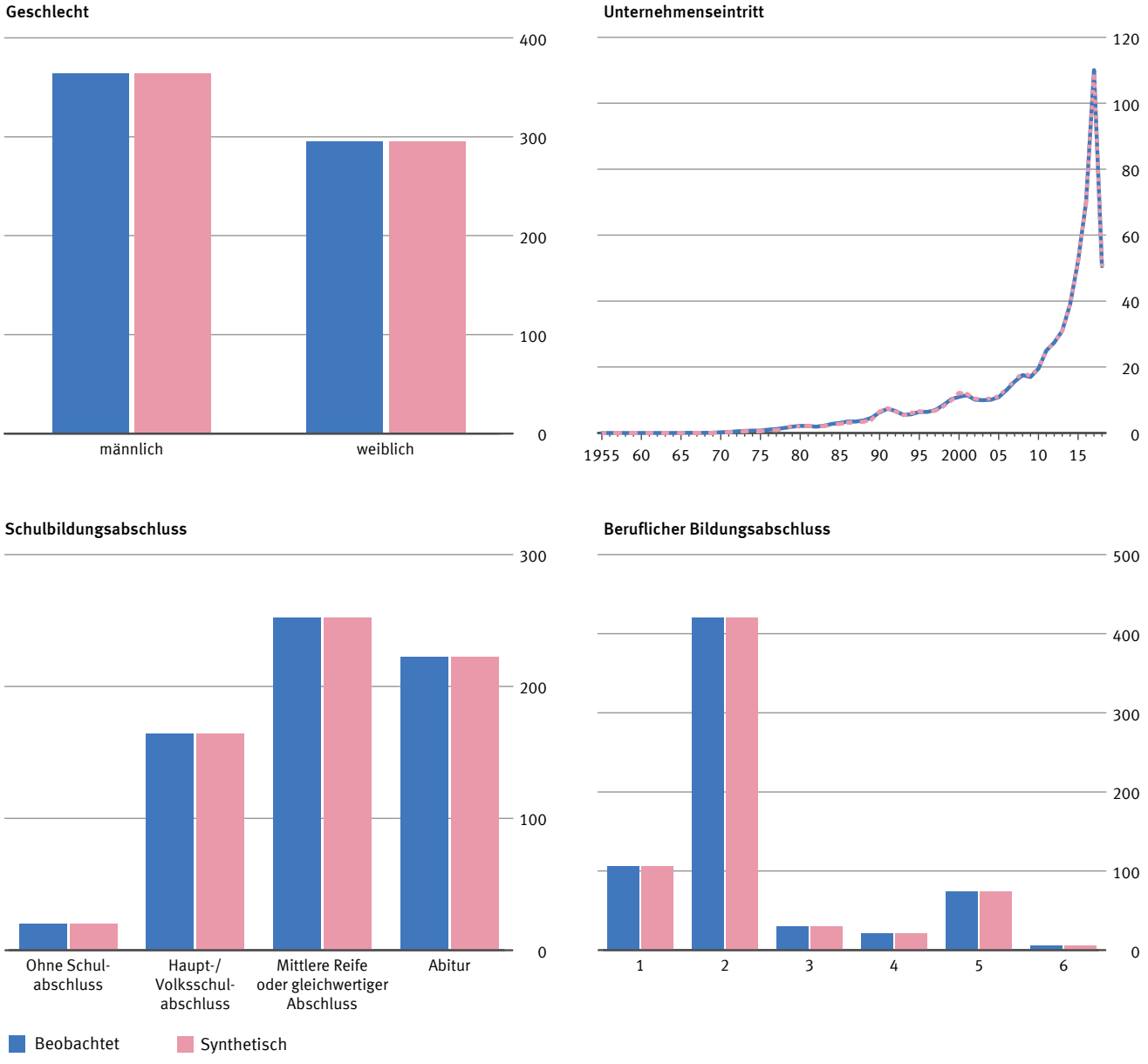
Sowohl die grafische Betrachtung der Verteilungen für einige beispielhafte nicht metrische Variablen in ↘ Grafik 3 als auch die Auswertung der Hellingerdistanz für die metrischen Variablen zeigen, dass die generierten, multiplen, partiell synthetischen Daten die Gegebenheiten der Originaldaten gut replizieren. Die Hellingerdistanz dient dazu, den Abstand und damit die Unterscheidbarkeit von zwei Verteilungen zu berechnen (Le Cam/Yang, 2000). Folglich eignet sie sich, um die Ähnlichkeit von anonymisierten oder synthetisierten Variablen zu Originalvariablen bezüglich ihrer Verteilungen zu bewerten (Gomatam und andere, 2005). Per Definition liegt die Hellingerdistanz D_H zwischen 0 und 1, wobei die Verteilungen sich umso ähnlicher sind, je näher der Wert bei null liegt.

$$D_H^2(V, V_{syn}) = \frac{1}{2\sqrt{2}} \int \left(\sqrt{f(V_i)} - \sqrt{g(V_{syn,i})}\right)^2 dV \quad (1)$$

Auch die Ergebnisse der Analysen zur Hellingerdistanz bestätigen den Eindruck aus den optischen Vergleichen ausgewählter Verteilungen. Für die beiden Zeitinformationen (Geburtsjahr und Jahr des Eintritts in das Unternehmen) sind die Verteilungen aus den synthetischen Datensätzen nahezu identisch zur Verteilung der jeweiligen Variablen in den Originaldaten. Ebenfalls sehr gering ist der Abstand zwischen der synthetischen Ver-

Grafik 3

Univariate Verteilungen der beispielhaften diskreten Merkmale Geschlecht, Unternehmenseintritt, schulischer und beruflicher Bildungsabschluss
in 1 000



Berechnungen auf Basis der Verdienststrukturerhebung 2018.

teilung und der Verteilung in den Originalvariablen für den „Gesamtverdienst für Überstunden“ ($D_H^2 < 0{,}05$) sowie für die Variable „Gesetzliche Abzüge durch die Sozialversicherung (insgesamt)“. Für letztere liegt die Hellingerdistanz knapp über 0,05. Für „Brutto-, Nettomonats- und Bruttojahresverdienst“ ist die Hellingerdistanz zwar doppelt so hoch im Vergleich zur Variablen „Gesamtverdienst für Überstunden“, mit $D_H^2 < 0{,}10$ aber immer noch sehr gering. Lediglich die Variable „Unternehmensgröße“ beziehungsweise „Anzahl der Mitarbei-

ter des Unternehmens“, welche ursprünglich aus dem Betriebsdatensatz stammt, ist mit $D_H^2 = 0{,}6431$ sehr hoch. ↘ Tabelle 2 Diese Beobachtung geht einher mit den Ergebnissen beziehungsweise Abweichungen der deskriptiven Kennzahlen (siehe Tabelle 1). Folglich ist vor einer finalen Bereitstellung der Daten zu prüfen, ob im Synthetisierungsschritt 1 durch eine Anpassung der Modellparameter weitere Verbesserungen möglich sind (siehe Grafik 2).

Tabelle 2

Hellingerdistanzen zwischen den Verteilungen der originalen und partiell synthetischen (metrischen) Einkommensvariablen und Jahresangaben

Bruttomonatsverdienst	0,0798
Gesamtverdienst für Überstunden	0,0404
Gesetzliche Abzüge durch die Sozialversicherung (insgesamt)	0,0602
Bruttojahresverdienst	0,0826
Nettomonatsverdienst	0,0776
Geburtsjahr	0,0024
Jahr des Unternehmenseintritts	0,0007
Unternehmensgröße	0,6431

Berechnungen auf Basis der Verdienststrukturerhebung 2018.

Darüber hinaus kann die globale Nützlichkeit der synthetisierten Daten durch den sogenannten Propensity Score Mean-Squared-Error (pMSE) bewertet werden. Dieser beschreibt ebenfalls die Ähnlichkeit der Originaldaten und der generierten synthetischen Daten. Dafür werden der Originaldatensatz und jeweils ein synthetisierter Datensatz „gestapelt“ (Summe der Beobachtungen: $n_{orig} + n_{syn} = N$). Anschließend werden propensity scores berechnet in Bezug auf eine Indikatorvariable *I*, welche angibt, ob Beobachtungen aus den Originaldaten oder synthetischen Daten stammen. Final wird dann der durchschnittliche Fehler dieser propensity scores berechnet und das Ergebnis für alle $m = 5$ partiell synthetischen Datensätze gemittelt. Da in der vorliegenden Arbeit der Originaldatensatz bezüglich der Anzahl an Berichtseinheiten 1:1 repliziert wurde, liegt der pMSE per Definition zwischen 0 und 0,25. Dabei ist die Nützlichkeit der Daten umso besser, desto näher der pMSE bei null liegt (Woo und andere, 2009). Zu beachten ist, dass auch die Beobachtungszahl den pMSE beeinflusst (Snoke und andere, 2018).

$$(2) \qquad pMSE = \frac{1}{N}\sum_{i=1}^{N}(p_i - \frac{n_{syn}}{N})^2$$

Die propensity scores für den pMSE wurden mithilfe von CART-Modellen geschätzt. Als Prädiktorvariablen sind dabei die Variablen Alter und Geschlecht sowie alle einkommensbezogenen Variablen zur Schätzung herangezogen worden. Der pMSE liegt mit 0,0042 sehr nahe an null und damit an der maximal möglichen Nützlichkeit. Das heißt das Modell kann nicht perfekt zwischen den Original- und den synthetischen Daten unterscheiden.

Hierbei ist jedoch zu berücksichtigen, dass eine Bewertung lediglich einzelfallbezogen stattfinden kann und keine globalen Aussagen abgeleitet werden können. Um die modellspezifische Nützlichkeit zu bewerten, werden beispielhafte Analysen sowohl auf den partiell synthetisierten Daten als auch auf den Originaldaten durchgeführt und die jeweiligen Ergebnisse miteinander verglichen. Neben der bloßen Betrachtung der Schätzer und ihrer Streuungen und Signifikanzniveaus ist die Betrachtung der Überlappung der Konfidenzintervalle (Karr und andere, 2006) eine etablierte Methode. Für die Bewertung der modellspezifischen Nützlichkeit wurde eine beispielhafte lineare Regressionsanalyse auf die abhängige Variable Bruttostundenlohn berechnet. Als mögliche erklärende Variablen für Unterschiede im Bruttostundenlohn wurden dabei Informationen zur beruflichen und schulischen Bildung, zum Geschlecht, Geburtsjahr und Jahr des Unternehmenseintritts sowie zu einer möglichen Befristung und zur Unternehmensgröße in das Modell aufgenommen. Das Modell zeigt sowohl auf Basis der Originaldaten als auch der multiplen, partiell synthetischen Daten einen negativen Zusammenhang zwischen den unabhängigen und den abhängigen Variablen. Ausnahmen sind die Bildungsinformationen und die Unternehmensgröße. Bis auf die Regressionskonstante, das Geschlecht und die mögliche Beteiligung der öffentlichen Hand am Unternehmenskapital sind auch nur vernachlässigbare Unterschiede bei der Größe der Koeffizienten zu beobachten. Bei den drei genannten unabhängigen Variablen weichen die Koeffizienten teils mehr und teils weniger deutlich von ihren auf den Originaldaten basierenden Gegenspielern ab. Unabhängig von der Datenbasis wird für die Regressionskonstante sowie alle unabhängigen Variablen ein statistisch höchst signifikanter Einfluss auf die abhängige Variable Bruttostundenlohn beobachtet. Die wichtigsten Ergebnisse der beispielhaften linearen Regression werden also auch mit den multiplen, partiell synthetischen Datensätzen der Verdienststrukturerhebung 2018 wiedergegeben. Die Überlappung der

Tabelle 3
Vergleich einer beispielhaften Regressionsanalyse (abhängige Variable: Bruttostundenlohn) basierend auf Original- und partiell synthetischen Daten

	Partiell synthetische Daten	Originaldaten	Überlappung des Konfidenzintervalls
Konstante	657,9256	643,0927***	0,0704
Schulbildungsabschluss	1,6492	1,7106***	0,2916
Beruflicher Bildungsabschluss	3,6022	3,6179***	0,7476
Geschlecht	– 3,3084	– 3,6585***	0,0000
Geburtsjahr	– 0,0366	– 0,0409***	0,2123
Unternehmenseintritt	– 0,2850	– 0,2733***	0,5591
Befristung	– -1,3867	– 1,3712***	0,6593
Privatwirtschaft (dummy)	– 0,2421	– 0,4013***	0,3259
Unternehmensgröße	0,0000	0,0000***	0,0000
Durchschnittliche Überlappung des Konfidenzintervalls			0,3185

Berechnungen auf Basis der Verdienststrukturerhebung 2018.
Anmerkung: ***p<0.001

Konfidenzintervalle der Koeffizienten der unabhängigen Variablen lässt jedoch ein weiteres Optimierungspotenzial vermuten. Nur für ein Drittel der Koeffizienten ist eine Überlappung von mehr als 50 % zu beobachten (beruflicher Bildungsabschluss, Jahr des Unternehmenseintritts, Befristung des Arbeitsverhältnisses). Da unter diesen Variablen eine metrische, eine nominale und eine ordinale Variable enthalten sind, ist zunächst einmal nicht von einem Einfluss des Variablentyps auf die Überlappung des Konfidenzintervalls auszugehen. ↘ Tabelle 3

5 Diskussion

Der vorliegende Beitrag hat einen Ansatz zur multiplen, partiellen Synthetisierung eines verknüpften Datensatzes vorgestellt. Dafür wurde die multiple, partielle Synthetisierung des Gesamtdatenmaterials der Verdienststrukturerhebung 2018 in drei Schritte unterteilt und zunächst nur sensible Daten der Betriebsdaten fünffach synthetisiert. Im Anschluss erfolgte die Zusammenführung mit den zu diesem Zeitpunkt unveränderten Angestelltendaten, bevor schließlich die sensiblen Variablen aus den Angestelltendaten in den fünf zusammengeführten Gesamtdatensätzen synthetisiert wurden. Dieses Vorgehen ermöglicht es, die logische Konsistenz von übergreifend in beiden Teildatensätzen vorhandenen Informationen sicherzustellen.

Die Ergebnisse zur Nützlichkeitsbewertung der erstellten multiplen, partiell synthetischen Daten zur Verdienststrukturerhebung 2018 deuten auf eine gute Nützlichkeit des Datenprodukts für die Wissenschaft hin. Die Ergebnisse der globalen Nützlichkeit zeigen, dass grundlegende deskriptive Statistiken (Mittelwert, Median, Standardabweichung und Quartile) zur Datensatzstruktur durch die multiplen, partiell synthetischen Daten wiedergegeben werden. Auch die Analysen zum Abgleich der Verteilungen der synthetischen Variablen mit ihren Gegenstücken in den Originaldaten mittels pMSE und Hellingerdistanz zeigen, dass die synthetisierten Datensätze die Verteilungen aus den Originaldaten gut abbilden. Lediglich für die Variable zur Anzahl der Angestellten im Unternehmen zeigt bezüglich der Verteilung größere Abweichungen. Vor einer finalen Bereitstellung als partiell synthetisches Datenprodukt sollte an dieser Stelle nochmals geprüft werden, ob es möglich ist, die Verteilungsgüte dieser Variable durch eine Anpassung der Modellparameter zu optimieren.

Grundsätzliches Potenzial für die wissenschaftliche Nutzung eines solchen Datenprodukts in den Forschungsdatenzentren zeigen auch die Ergebnisse einer beispielhaft durchgeführten Regressionsanalyse auf die abhängige Variable Bruttostundenlohn. Die Effektrichtung und das Signifikanzniveau der berechneten Regressionskoeffizienten der in das exemplarische Modell aufgenommenen unabhängigen Variablen werden zu jeder Zeit korrekt wiedergegeben. Bei der Effektstärke sind teilweise kleinere Abweichungen zu beobachten und die Überlappung der Konfidenzintervalle ist nur bedingt zufriedenstellend.

An den aufgezeigten Ergebnissen sind zum aktuellen Forschungsstand und unter den aktuellen technischen Möglichkeiten keine signifikanten Verbesserungen zu erwarten. Bei einem qualitativen Befragungsprozess wird im weiteren Verlauf des Projekts untersucht, ob die zugrunde liegenden Charakteristika der Wissenschaft zur Nutzung als separates Datenprodukt einen Mehrwert bieten könnten. Unabhängig von den Ergebnissen dieses qualitativen Prozesses könnten die vorgestellten Daten der Wissenschaft zumindest für erste explorative Einblicke und für eine bessere Unterstützung bei der Entwicklung des Programmcodes im Rahmen der kontrollierten Datenfernverarbeitung behilflich sein.

Inwieweit die Daten der breiten Öffentlichkeit zum Download oder nur der Wissenschaft als Scientific-Use-File oder per Remote Access zur Verfügung gestellt werden können, hängt von den Ergebnissen der Risikoanalyse ab. Diese wird ein weiterer Artikel in dieser Zeitschrift näher beleuchten. Zudem steht eine finale juristische Bewertung noch aus. Abschließend lässt sich festhalten, dass ein partiell synthetisierter Datensatz mit einer zufriedenstellenden Nützlichkeit generiert werden konnte. Ob sich der Aufwand zur Erstellung solch synthetisierter Datensätze allerdings lohnt, hängt maßgeblich davon ab, wie die Risikoanalyse ausfällt und – damit einhergehend – welcher Zugangsweg gewählt werden kann. Nicht zuletzt ist die Akzeptanz durch die Wissenschaft ein entscheidendes Kriterium.

LITERATURVERZEICHNIS

Brenzel, Hanna/Zwick, Markus. *Eine informationelle Infrastruktur in Deutschland ist erwachsen – das Forschungsdatenzentrum des Statistischen Bundesamtes*. In: WISTA Wirtschaft und Statistik. Ausgabe 6/2022, Seite 54 ff.

Drechsler, Jörg. *Generating Multiply Imputed Synthetic Datasets: Theory and Implementation*. Bamberg 2010. [Zugriff am 26. Juni 2025]. Verfügbar unter: fis.uni-bamberg.de

Drechsler, Jörg/Haensch, Anna-Carolina. *30 years of synthetic data*. arXiv preprint arXiv:2304.02107. 2023. DOI: 10.48550/arXiv.2304.02107

Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder. *Metadatenreport. Teil II: Produktspezifische Informationen zur Nutzung der Verdienststrukturerhebung 2018 per On-Site-Nutzung*. DOI: 10.21242/62111.2018.00.00.1.1.0

Gomatam, Shanti/Karr, Alan F./Sanil, Ashish P. *Data swapping as a decision problem*. In: Journal of Official Statistics. Statistics Sweden. Jahrgang 21. Ausgabe 4/2005, Seite 635 ff.

Karr, Allan F./Kohnen, Christine N./Oganian, Anna/Reiter, Jerome P./Sanil, Ashish P. *A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality*. In: The American Statistician. Jahrgang 60. Ausgabe 3/2006, Seite 224 ff. DOI: 10.1198/000313006X124640

Le Cam, Lucien/Yang, Grace Lo. *Contiguity – Hellinger Transforms*. In: Asymptotics in Statistics. New York 2000. DOI: 10.1007/978-1-4612-1166-2_3

Little, Roderick J. A. *Statistical Analysis of Masked Data*. In: Journal of Official Statistics. Statistics Sweden. Jahrgang 9. Ausgabe 2/1993, Seite 407 ff.

Little, Claire/Elliot, Mark/Allmendinger, Richard/Samani, Sahel Shariati. *Generative Adversarial Networks for Synthetic Data Generation: A Comparative Study*. arXiv preprint arXiv:2112.01925. 2021.

Raab, Gillian. M./Nowok, Beata/Dibben, Chris. *Practical data synthesis for large samples*. In: Journal of Privacy and Confidentiality. Jahrgang 7. Ausgabe 3/2016, Seite 67 ff. DOI: 10.29012/jpc.v7i3.407

Reiter, Jerome P. *Selecting the number of imputed datasets when using multiple imputation for missing data and disclosure limitation*. In: Statistics & Probability Letters. Jahrgang 78. Ausgabe 1/2008, Seite 15 ff.

Reiter, Jerome P./Mitra, Robin. *Estimating risks of identification disclosure in partially synthetic data*. In: Journal of Privacy and Confidentiality. Jahrgang 1. Ausgabe 1/2009, Seite 99 ff.

Rubin, Donald B. *Discussion: Statistical disclosure limitation*. In: Journal of Official Statistics. Statistics Sweden. Jahrgang 9. Ausgabe 2/1993, Seite 462 ff.

LITERATURVERZEICHNIS

Snoke, Joshua/Raab, Gillian M./Nowok, Beata/Dibben, Chris/Slavkovic, Aleksandra. *General and specific utility measures for synthetic data*. In: Journal of the Royal Statistical Society. Series A: Statistics in Society. Jahrgang 181. Ausgabe 3/2018, Seite 663 ff. DOI: 10.1111/rssa.12358

Woo, Mi-Ja/Reiter, Jerome P./Oganian, Anna/Karr, Alan F. *Global measures of data utility for microdata*. In: Journal of Privacy and Confidentiality. Jahrgang 1. Ausgabe 1/2009, Seite 111 ff. DOI: 10.29012/jpc.v1i1.568

RECHTSGRUNDLAGEN

Gesetz über die Statistik für Bundeszwecke (Bundesstatistikgesetz – BStatG) in der Fassung der Bekanntmachung vom 20. Oktober 2016 (BGBl. I Seite 2394), das zuletzt durch Artikel 14 des Gesetzes vom 8. Mai 2024 (BGBl. I Nr. 152) geändert worden ist.

Herausgeber
Statistisches Bundesamt (Destatis), Wiesbaden

Schriftleitung
Dr. Daniel Vorgrimler
Redaktion: Ellen Römer

Ihr Kontakt zu uns
www.destatis.de/kontakt

Erscheinungsfolge
zweimonatlich, erschienen im August 2025
Ältere Ausgaben finden Sie unter www.destatis.de sowie in der Statistischen Bibliothek.

Artikelnummer: 1010200-25004-4, ISSN 1619-2907