

WISTA

Wirtschaft und Statistik

Sonja Leischner | Angela Kolbe

Zum Einfluss des Grundrechts auf informationelle Selbstbestimmung auf die Bundesstatistik

Stefanie Setzer | Johannes Rohde |
Volker Güttgemanns | Patrick Rothe

**Die Cell-Key-Methode in den Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder –
Teil 1: Vorstellung des neuen Geheimhaltungsverfahrens**

Patrick Rothe | Volker Güttgemanns |
Johannes Rohde | Stefanie Setzer

**Die Cell-Key-Methode in den Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder –
Teil 2: Auswirkungen des neuen Geheimhaltungsverfahrens**

Igor Franjić | Thomas Kolvenbach |
Mingyong Tong

Der europäische Mikrodatabankenaustausch – neue Datenquelle für die Außenhandelsstatistik

Julius Weißmann | Tim Herbst

Maschinelles Lernen im Basisregister für Unternehmen

Stefan Linz | Luis Federico Flores |
Peter Mehlhorn

Umstellung der Umsatz-, Auftragseingangs- und Auftragsbestandsindizes im Verarbeitenden Gewerbe auf das Basisjahr 2021

Anke Rink | Ines Seiwert |
Raimund Rödel

Regionale Ergebnisse der Unternehmensdemografie

3 | 2024

ABKÜRZUNGEN

D	Durchschnitt (bei nicht addierfähigen Größen)
Vj	Vierteljahr
Hj	Halbjahr
a. n. g.	anderweitig nicht genannt
o. a. S.	ohne ausgeprägten Schwerpunkt
Mill.	Million
Mrd.	Milliarde

ZEICHENERKLÄRUNG

–	nichts vorhanden
0	weniger als die Hälfte von 1 in der letzten besetzten Stelle, jedoch mehr als nichts
.	Zahlenwert unbekannt oder geheim zu halten
. . .	Angabe fällt später an
X	Tabellenfach gesperrt, weil Aussage nicht sinnvoll
I oder —	grundsätzliche Änderung innerhalb einer Reihe, die den zeitlichen Vergleich beeinträchtigt
/	keine Angaben, da Zahlenwert nicht sicher genug
()	Aussagewert eingeschränkt, da der Zahlenwert statistisch relativ unsicher ist
	Abweichungen in den Summen ergeben sich durch Runden der Zahlen.
	Tiefer gehende Internet-Verlinkungen sind hinterlegt.

INHALT

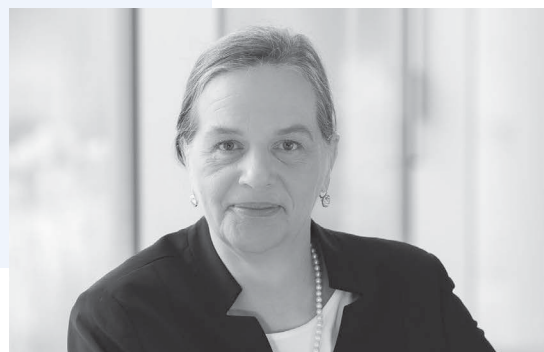
3	Editorial
4	Kennzahlen
8	Aktuelle Informationsangebote
10	Kurznachrichten
17	<p>Sonja Leischner, Angela Kolbe</p> <p>Zum Einfluss des Grundrechts auf informationelle Selbstbestimmung auf die Bundesstatistik</p> <p><i>Influence of the right to informational self-determination on federal statistics</i></p>
31	<p>Stefanie Setzer, Johannes Rohde, Volker Güttgemanns, Patrick Rothe</p> <p>Die Cell-Key-Methode in den Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder – Teil 1: Vorstellung des neuen Geheimhaltungsverfahrens</p> <p><i>The cell key method in the Research Data Centres of the statistical offices of the Federation and the Länder – Part 1: presenting the new disclosure control method</i></p>
45	<p>Patrick Rothe, Volker Güttgemanns, Johannes Rohde, Stefanie Setzer</p> <p>Die Cell-Key-Methode in den Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder – Teil 2: Auswirkungen des neuen Geheimhaltungsverfahrens</p> <p><i>The cell key method in the Research Data Centres of the statistical offices of the Federation and the Länder – Part 2: effects of the new disclosure control method</i></p>

INHALT

55	Igor Franjić, Thomas Kolvenbach, Mingyong Tong Der europäische Mikrodatenaustausch – neue Datenquelle für die Außenhandelsstatistik <i>The European microdata exchange – A new data source for foreign trade statistics</i>
67	Julius Weißmann, Tim Herbst Maschinelles Lernen im Basisregister für Unternehmen <i>Machine learning in the basic register of enterprises</i>
80	Stefan Linz, Luis Federico Flores, Peter Mehlhorn Umstellung der Umsatz-, Auftragseingangs- und Auftragsbestandsindizes im Verarbeitenden Gewerbe auf das Basisjahr 2021 <i>Rebasing the indices of turnover, new orders and the stock of orders in manufacturing to the year 2021</i>
93	Anke Rink, Ines Seiwert, Raimund Rödel Regionale Ergebnisse der Unternehmensdemografie <i>Regional results of business demography</i>

EDITORIAL

Dr. Ruth Brand



LIEBE LESERIN, LIEBER LESER,

Informationen zu Unternehmen, zum Beispiel zu Gründungen und Schließungen, zu schnell wachsenden Unternehmen und zu Überlebensraten neu gegründeter Unternehmen, sind wichtige Kennzahlen, um die Dynamik einer Volkswirtschaft zu beurteilen. Sie fließen beispielsweise in die Strukturindikatoren ein, die die Fortschritte bei der Verwirklichung der Strategie für Wachstum und Beschäftigung [Europa 2020](#) überwachen. Bislang lagen Analysen zur Unternehmensdemografie für Deutschland insgesamt vor, ab dem Berichtsjahr 2021 ist eine kleinräumigere Darstellung möglich. In Zusammenarbeit mit dem Bayerischen Landesamt für Statistik ist ein Beitrag entstanden, der in dieser WISTA-Ausgabe die Methodik und erste Ergebnisse der Unternehmensdemografie auf regionaler Ebene vorstellt. Neben der Zuordnung nach Kreisen wird eine auf Koordinaten basierende Kartendarstellung gezeigt, ebenso eine Auswertung nach Raumkategorien.

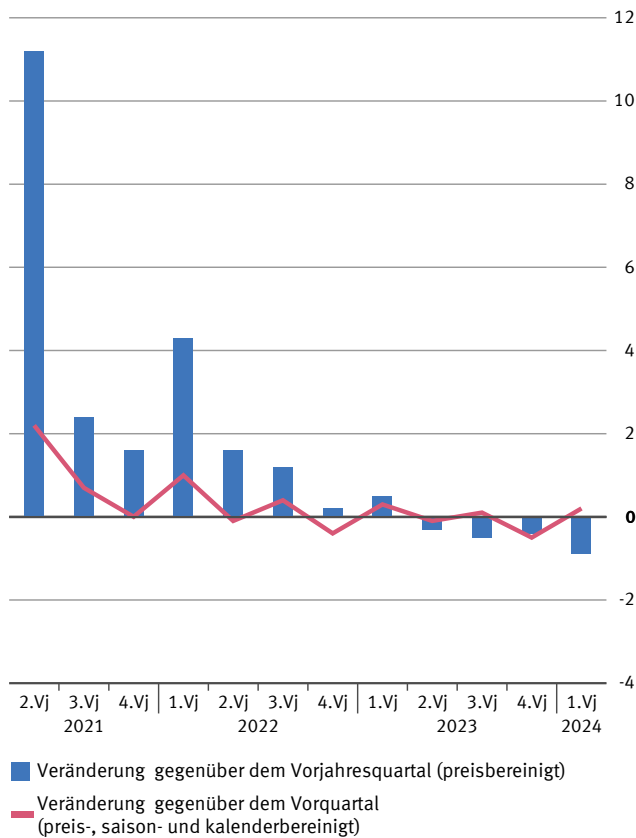
Die Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder stellen der Wissenschaft für tiefgehende Analysen Mikrodaten zur Verfügung. Um unserer gesetzlichen Pflicht zur Geheimhaltung nachzukommen, wird derzeit zumeist das Verfahren der Zellspernung angewandt, welches aufwandsstark ist und bestimmte Werte den Ergebnistabellen entnimmt. Für ausgewählte Statistiken führen die Forschungsdatenzentren mit der Cell-Key-Methode ein neues Geheimhaltungsverfahren ein. Darüber berichtet diese Ausgabe in zwei Artikeln: Der erste Aufsatz stellt die Funktionsweise der Cell-Key-Methode vor. Sie erzeugt einen Schutz vor der Reidentifikation von Befragten, indem eine Überlagerung von Fallzahlen eine Unsicherheit über die Anzahl der tatsächlich zum Ergebnis beitragenden Fälle schafft. Die Auswirkungen des neuen Verfahrens auf die Ergebnisse beschreibt der zweite Aufsatz.

Informieren Sie sich darüber hinaus zu Themen wie 40 Jahre Grundrecht auf informationelle Selbstbestimmung, dem erstmaligen Mikrodatenaustausch zu innereuropäischen Warenexporten und dem Nutzungspotenzial maschinellen Lernens im Basisregister für Unternehmen.

Ruth Brand

Präsidentin des Statistischen Bundesamtes

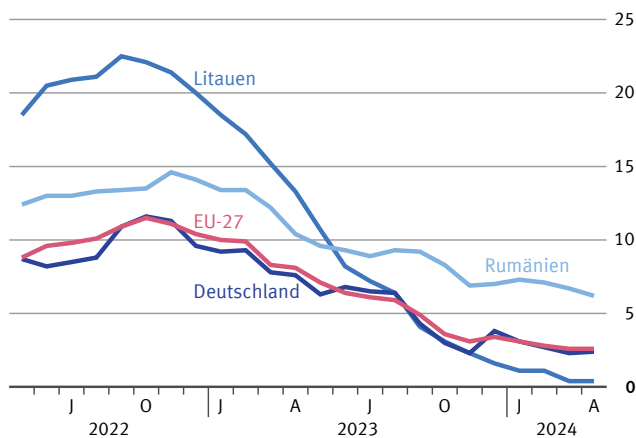
Bruttoinlandsprodukt
in %



Verbraucherpreisindex
2020 = 100

2023		2024	
Januar	114,3	Januar	117,6
Februar	115,2	Februar	118,1
März	116,1	März	118,6
April	116,6	April	119,2
Mai	116,5	Mai	119,3
Juni	116,8		2,4 %
Juli	117,1		Veränderung zum Vorjahresmonat
August	117,5		
September	117,8		
Oktober	117,8		
November	117,3		
Dezember	117,4		

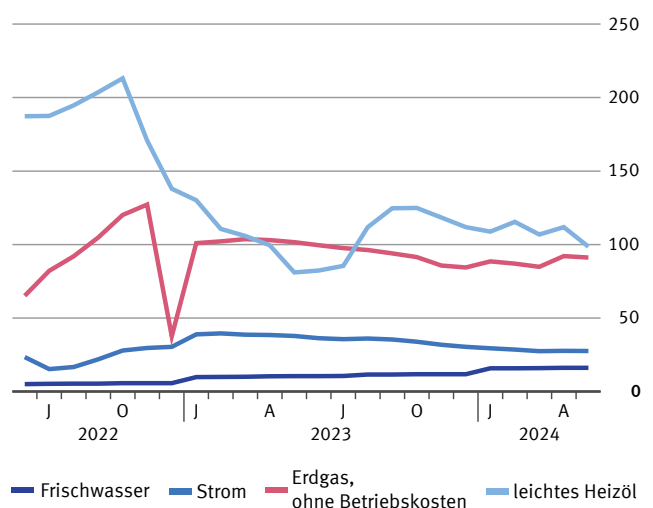
Harmonisierter Verbraucherpreisindex insgesamt
Veränderung gegenüber dem Vorjahresmonat in %



Dargestellt sind neben Deutschland und der Europäischen Union insgesamt (EU-27) die Länder mit der höchsten und der niedrigsten Veränderungsrate innerhalb der EU.

Stand: 13.06.2024

Entwicklung der Verbraucherpreise für Energie und Wasser
Preisabstand in % gegenüber dem Jahr 2020



The chart displays the monthly number of deaths in Germany and abroad. The red line, labeled 'Ausland', shows a significant peak in late 2022, reaching nearly 10 deaths, followed by a sharp decline and then a rise to around 7 deaths by January 2024. The blue line, labeled 'Deutschland', remains consistently low, mostly below 1 death per month, with a slight increase in late 2023 and early 2024.

Month	Ausland (Deaths)	Deutschland (Deaths)
Apr 2022	6.5	0.5
May 2022	7.5	0.6
Jun 2022	8.5	0.5
Jul 2022	8.0	0.4
Aug 2022	7.5	0.5
Sep 2022	7.5	0.6
Oct 2022	5.5	0.5
Nov 2022	5.0	0.4
Dec 2022	4.5	0.5
Jan 2023	4.5	0.6
Feb 2023	5.5	0.7
Mar 2023	6.5	0.8
Apr 2023	7.5	0.9
May 2023	8.0	0.8
Jun 2023	8.5	0.7
Jul 2023	9.5	0.8
Aug 2023	9.0	0.7
Sep 2023	8.5	0.8
Oct 2023	8.5	0.9
Nov 2023	5.5	0.8
Dec 2023	6.5	0.7
Jan 2024	5.0	0.6
Feb 2024	5.5	0.7
Mar 2024	7.0	0.8

Kalender- und saisonbereinigte Werte nach dem Verfahren X13 JDemetra+. – Vorläufiges Ergebnis.

The chart displays monthly precipitation (mm) as blue bars and monthly temperature (°C) as a red line. The x-axis represents time from January 2022 to April 2024, with labels for J, O, J, A, O, and A. The y-axis for precipitation ranges from 0 to 2.0 mm, and the y-axis for temperature ranges from -1.0 to 2.0 °C. Precipitation is highest in winter (around 1.7 mm in Jan 2022) and lowest in summer (around 0.1 mm in Jul 2023). Temperature is highest in summer (around 0.4 °C in Aug 2023) and lowest in winter (around -0.5 °C in Jan 2024).

Month	Precipitation (mm)	Temperature (°C)
Jan 2022	1.7	0.3
Feb 2022	1.6	0.1
Mar 2022	1.3	0.0
Apr 2022	1.3	0.2
May 2022	1.3	0.7
Jun 2022	1.3	0.8
Jul 2022	1.3	0.4
Aug 2022	1.3	0.2
Sep 2022	1.2	0.1
Oct 2022	1.2	-0.1
Nov 2022	1.1	-0.3
Dec 2022	1.1	-0.5
Jan 2023	1.1	-0.3
Feb 2023	1.1	0.3
Mar 2023	1.1	0.3
Apr 2023	1.0	0.3
May 2023	0.9	0.2
Jun 2023	0.9	0.1
Jul 2023	0.8	0.1
Aug 2023	0.7	0.4
Sep 2023	0.6	0.3
Oct 2023	0.6	0.2
Nov 2023	0.6	0.1
Dec 2023	0.6	-0.1
Jan 2024	0.5	-0.5
Feb 2024	0.4	0.2
Mar 2024	0.3	0.2
Apr 2024	0.3	0.3

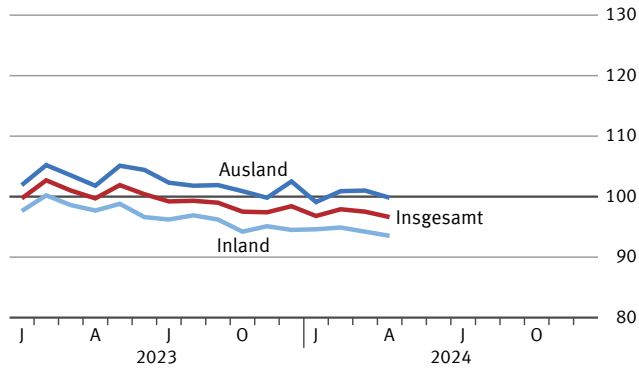
Stand: 13.06.2024

The chart displays three data series over time from August 2022 to June 2024. The y-axis represents a value ranging from 0 to 50. The x-axis shows months, with labels for August (A), January (J), and October (O) for the years 2022, 2023, and 2024. The 'Originalwerte' (red line) shows significant seasonal fluctuations, with peaks around January and October. The 'Trend-Konjunktur-Komponente (Berliner Verfahren)' (dark blue line) and 'saison- und kalenderbereinigte Werte' (light blue line) show a smoother, generally downward trend, with the latter exhibiting some seasonal variation.

Month	Originalwerte	Trend-Konjunktur-Komponente (Berliner Verfahren)	saison- und kalenderbereinigte Werte
Aug 2022	35	35	35
Sep 2022	34	34	34
Oct 2022	33	33	33
Nov 2022	32	32	32
Dec 2022	31	31	31
Jan 2023	33	30	30
Feb 2023	30	29	29
Mar 2023	28	28	28
Apr 2023	30	27	27
May 2023	28	26	26
Jun 2023	25	25	25
Jul 2023	23	24	24
Aug 2023	25	24	24
Sep 2023	28	24	24
Oct 2023	25	24	24
Nov 2023	23	24	24
Dec 2023	22	24	24
Jan 2024	25	24	24
Feb 2024	23	24	24
Mar 2024	22	24	24
Apr 2024	23	24	24
May 2024	22	24	24
Jun 2024	21	24	24

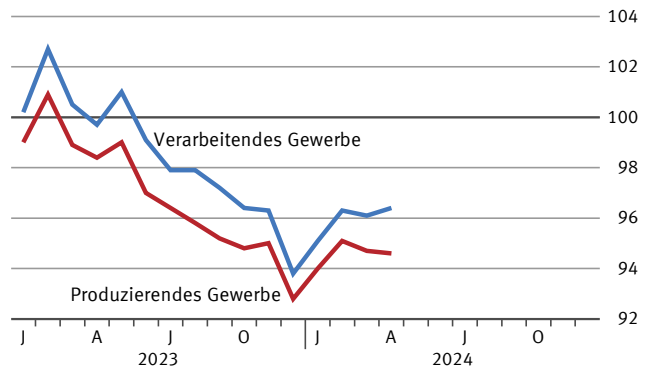
Kennzahlen und Indikatoren

Auftragseingang im Verarbeitenden Gewerbe
Volumenindex 2021 = 100



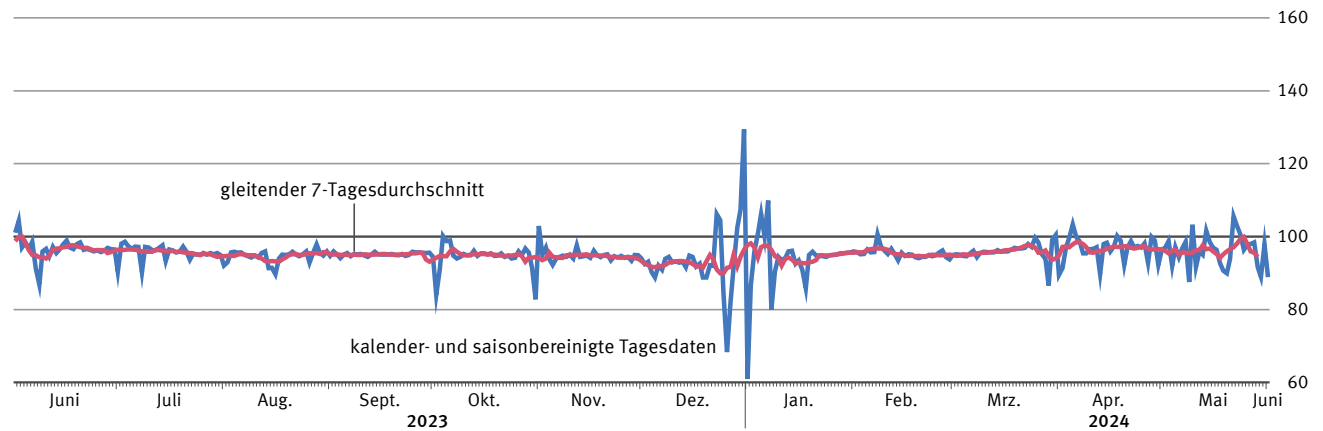
Kalender- und saisonbereinigter Wert nach dem Verfahren X13 JDemetra+. – Vorläufiges Ergebnis.

Produktion im Produzierenden und Verarbeitenden Gewerbe
Index 2021 = 100



Kalender- und saisonbereinigte Werte nach dem Verfahren X13 JDemetra+. – Vorläufiges Ergebnis.

Lkw-Maut-Fahrleistungsindex
2021 = 100



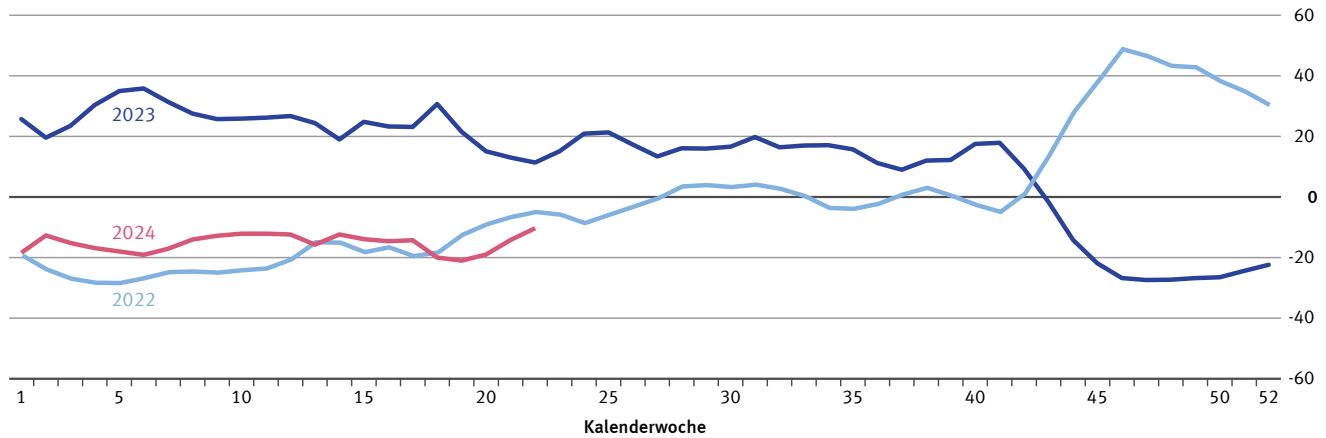
Quellen: Bundesamt für Logistik und Mobilität, Deutsche Bundesbank, Statistisches Bundesamt

Stand: 13.06.2024

Kennzahlen und Indikatoren

Neue Kreditverträge nach Kalenderwochen

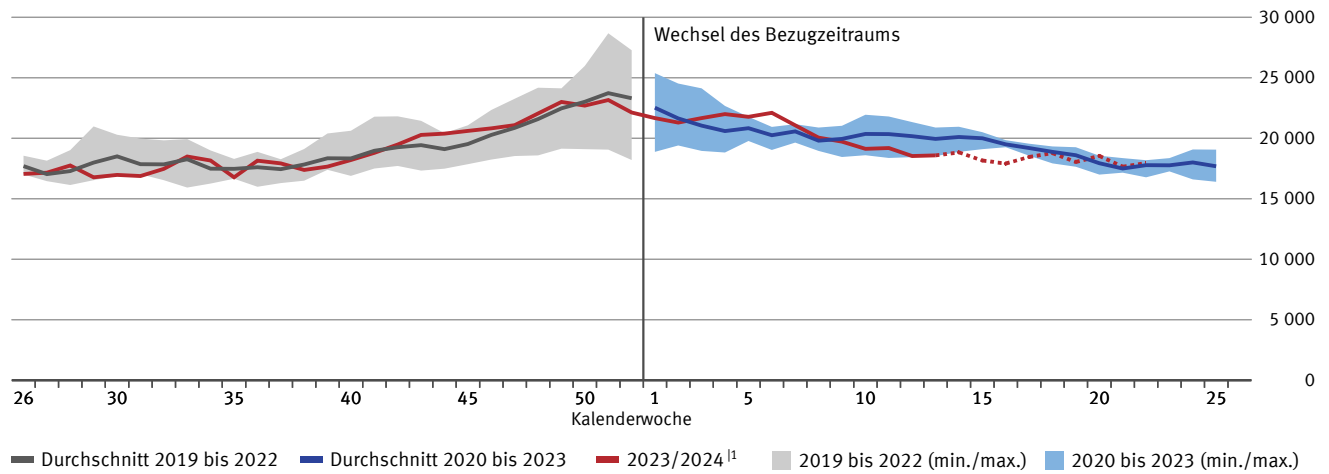
Veränderung gegenüber der entsprechenden Vorjahreswoche in %



Anfang 2022 zeigte sich mit Veränderungsraten von 100 % und mehr ein starker Anstieg im Vergleich zum Vorjahr; dabei handelt es sich um Sondereffekte, die seitens des Datenlieferanten nicht bereinigt werden konnten.

Quelle: SCHUFA Holding AG; Berechnung: Statistisches Bundesamt

Wöchentliche Sterbefallzahlen in Deutschland



Gestrichelte Werte enthalten Schätzanteil.

¹ Sonderauswertung der vorläufigen Sterbefallzahlen.

Stand: 13.06.2024



Ukraine

Der Angriff Russlands auf die Ukraine und die damit verbundenen Sanktionen haben starke Auswirkungen auf Wirtschaft und Bevölkerung sowie den Energie-sektor. Auf einer Sonderseite zum Thema stellt das Statistische Bundesamt relevante Daten zur Verfügung. Über die Seite gelangt man auch zum [zentralen Hilfs-portal](#) der Bundesregierung für Geflüchtete aus der Ukraine.

➤ www.destatis.de/Im-Fokus/Ukraine



Dashboard Deutschland

Das vom Statistischen Bundesamt entwickelte Datenportal bietet hochaktuelle und hochfrequente Zahlen, Daten und Fakten zu den Themen Arbeitsmarkt, Bauen und Wohnen, Energie, Finanzen, Konjunktur und Wirtschaft sowie Ukraine. Es trägt damit zu einem faktenbasierten demokratischen Diskurs der Öffentlichkeit und zur evidenzbasierten Entscheidungsfindung durch Politik und Verwaltung bei. Der integrierte Pulsmesser Wirtschaft bietet Einblicke in das aktuelle wirtschaftliche Geschehen, intuitives und einfaches Vergleichen von Daten sowie das Erkennen von konjunkturellen Entwicklungen und Zusammenhängen mithilfe täglicher, wöchentlicher, monatlicher und vierteljährlicher Indikatoren.

➤ www.dashboard-deutschland.de



EXSTAT – Experimentelle Statistiken

In der Rubrik „EXSTAT – Experimentelle Statistiken“ veröffentlicht das Statistische Bundesamt regelmäßig neue, innovative Projektergebnisse. Sie entstehen auf der Grundlage neuer Datenquellen und Methoden. Im Reifegrad und in der Qualität unterscheiden sie sich von amtlichen Statistiken, insbesondere in Bezug auf Harmonisierung, Erfassungsbereich und Methodik. Dennoch sind es Ergebnisse der Statistischen Ämter des Bundes und der Länder, die interessante, neue Perspektiven auf verschiedene Themenfelder der Statistik bieten.

➤ www.destatis.de/exstat

im Fokus

Inflation – das statistische Angebot rund ums Thema

Die Inflationsraten stehen stets im Fokus der Öffentlichkeit. Aktuelle Zahlen und Fakten sowie weiterführende Informationen stellt das Statistische Bundesamt auf der [Themenseite Verbraucherpreisindex und Inflationsrate](#) zur Verfügung. Das Video „[Verbraucherpreisindex und Inflation kurz erklärt](#)“ bietet einen kurzen, kompakten Einstieg ins Thema. Und mithilfe des persönlichen [Inflationsrechners](#) kann ermittelt werden, wie sehr die persönliche von der amtlichen Teuerungsrate abweicht.



Klima

Der Klimawandel ist eine der größten Herausforderungen der heutigen Zeit, alle Bereiche der Gesellschaft sind betroffen. Wie beeinflusst unsere Lebens- und Wirtschaftsweise das Klima? Wie wirkt sich die Umstellung hin zu mehr Klimaschutz gesamtgesellschaftlich aus? Was bedeutet sie für unseren Alltag – vom Weg zur Arbeit bis zum aktuellen Strompreis? Wo zeigen sich die Folgen des Klimawandels? Daten und Fakten zum Thema Klima, Klimawandel und Klimaschutz sind gebündelt unter

➤ www.destatis.de/klima



Fachkräfte

Fachkräftemangel und Arbeitskräftebedarf sind zunehmend wichtige Faktoren für die wirtschaftliche Entwicklung in Deutschland. Daten und Fakten dazu bündelt das Statistische Bundesamt auf einer eigenen Sonderseite. Das Angebot umfasst die Bereiche Demografie, Erwerbstätigkeit, Bildung und Zuwanderung – und wird sukzessive erweitert.

➤ www.destatis.de/fachkraefte

KURZNACHRICHTEN

IN EIGENER SACHE

75 Jahre Grundgesetz – Daten zu Grundrechten

Das Grundgesetz für die Bundesrepublik Deutschland ist 75 Jahre alt geworden. Am 8. Mai 1949 wurde es vom Parlamentarischen Rat beschlossen, von den Alliierten genehmigt und am 23. Mai 1949 in einer feierlichen Sitzung in Bonn ausgefertigt und verkündet. Mit seinen Verfassungsprinzipien wie Demokratie, Sozialstaat und Rechtsstaat sowie seinen in den Artikeln 1 bis 19 niedergelegten Grundrechten steht es über allen anderen deutschen Rechtsnormen.

Das Grundgesetz garantiert unter anderem die Würde des Menschen, die Gleichberechtigung von Frauen und Männern und den Schutz von Familie, Ehe oder Eigentum. Die Grundrechte spiegeln sich in einer Vielzahl von amtlichen Statistiken wider. Diese geben zum Beispiel Aufschluss darüber, wie es um menschenwürdige Lebensbedingungen und eine möglichst gleichberechtigte Teilhabe an Gesellschaft und Wirtschaft in Deutschland bestellt ist.

Anlässlich des Jubiläums präsentiert das Statistische Bundesamt auf einer Themenseite Indikatoren zu zehn Aspekten der in Artikel 1 bis 19 des Grundgesetzes festgeschriebenen Grundrechte. Diese Indikatoren verdeutlichen, dass Demokratie Daten braucht, um die Einhaltung der Grundrechte gewährleisten zu können: So zeigt der Gender Pay Gap, wie es um die Gleichberechtigung auf dem Arbeitsmarkt bestellt ist, Statistiken zur Kinder- und Jugendhilfe geben Aufschluss über staatliche Unterstützung für Familien, die Sozialhilfestatistik liefert Daten zu Asylbewerberinnen und -bewerbern.

➤ www.destatis.de

Indikatoren zur Generationengerechtigkeit

Bei Generationengerechtigkeit geht es um die gerechte Verteilung gesellschaftlicher Belastungen beziehungsweise des gesellschaftlichen Wohls auf die verschiedenen Generationen. Im Jahr 2021 hat das Bundesverfassungsgericht entschieden, dass bestimmte Teile des Klimaschutzgesetzes von 2019 verfassungswidrig seien, da sie die Reduktion von Treibhausgasen nach 2030 unzureichend regeln und somit die Grundrechte der nach 2030 lebenden Menschen gefährden würden.

An diese Logik angelehnt, geben die Indikatoren der Generationengerechtigkeit des Statistischen Bundesamtes auf einer neuen Themenseite einen Überblick über andere Messgrößen, die durch das Handeln heutiger Generationen beeinflusst sind. Abhängig vom heutigen Handeln, werden die unterschiedlichen Lebensgrundlagen der Menschen in Deutschland auf eine irreversible oder langfristige Art beeinträchtigt. Dadurch könnten die im Grundgesetz verankerten Freiheiten, wie das Recht auf Leben und körperliche Unversehrtheit, für künftige Generationen verletzt oder gefährdet werden. Neben natürlichen Lebensgrundlagen wie Klima, Boden oder Wasser werden dabei auch soziale und wirtschaftliche Lebensgrundlagen betrachtet, deren Schutz für den Erhalt unserer Freiheiten ebenfalls von Bedeutung ist.

➤ www.destatis.de

AUS EUROPA

Wie steht es um Europa und die EU?

Wie geht es dem deutschen Arbeitsmarkt im Vergleich zu anderen EU-Staaten? Wofür benötigen private Haushalte in der EU am meisten Energie? Welches Land gibt am meisten Geld für sein Bildungssystem aus? Die Webseite „Europa in Zahlen“ des Statistischen Bundesamtes beleuchtet wichtige Kennzahlen aus Wirtschaft, Umwelt und Gesellschaft für Deutschland und die anderen EU-Staaten. Die Palette der Daten reicht vom Mobilitätsverhalten von Männern und Frauen und Kraftstoffpreisen über die Anzahl der Ökobauern bis zum Thema Ernährung und Übergewicht. Der Brexit Monitor verfolgt die Entwicklung im Vereinigten Königreich anhand ausgewählter sozioökonomischer Indikatoren.

🔗 www.destatis.de

56. Sitzung des AESS

Der Ausschuss für das Europäische Statistische System (AESS) tagte am 22. Mai 2024 in Luxemburg. Er hat folgende Durchführungsverordnungen der Kommission verabschiedet:

- › Durchführungsbestimmungen zur Verordnung (EU) 2022/2379 des Europäischen Parlaments und des Rates in Bezug auf Statistiken zu Nährstoffen,
- › Entwurf über die indikative Übersicht der Umweltgüter und -dienstleistungen gemäß der Verordnung (EU) Nr. 691/2011 des Europäischen Parlaments und des Rates über europäische umweltökonomische Gesamtrechnungen,
- › Gewährung von Ausnahmeregelungen für bestimmte Mitgliedstaaten in Bezug auf die Übermittlung von Statistiken über landwirtschaftliche Betriebsmittel und landwirtschaftliche Erzeugung gemäß den Durchführungsverordnungen (EU) 2023/1538 und (EU) 2023/1579,
- › Entwurf über einheitliche Bedingungen für die Übermittlung von Zeitreihen für die neue regionale Gliederung nach der Verordnung (EG) Nr. 1059/2003 des Europäischen Parlaments und des Rates,

- › Änderung der Verordnung (EU) 2020/1148 in Bezug auf die Erstellung harmonisierter Verbraucherpreisindizes (HVPI),
- › Änderung der Verordnung (EU) Nr. 1445/2007 des Europäischen Parlaments und des Rates bezüglich der Liste der Einzelpositionen für Kaufkraftparitäten.

Einer der wichtigen Tagesordnungspunkte der 56. AESS-Sitzung war der ergänzende Workshop zur Klärung der methodischen Bedenken der Mitgliedstaaten des Europäischen Statistischen Systems (ESS) bei der Berücksichtigung von selbst genutztem Wohneigentum im Harmonisierten Verbraucherpreisindex. Während einige ESS-Mitgliedstaaten eine jeweils national zugeschnittene Lösung präferieren, betont Deutschland zusammen mit weiteren Ländern die Notwendigkeit einer einheitlichen Methodik. Die Workshops, die sich unter anderem auch mit Maßnahmen zur Erhöhung der Aktualität beschäftigen, sind für Herbst 2024 und das Jahr 2025 vorgesehen.

Das vom Statistischen Amt der Europäischen Union (Eurostat) für das kommende Jahr vorgestellte Arbeitsprogramm enthält in seinem jetzigen Entwurf unter anderem folgende Themen:

- › Update der Standards für das System of National Accounts (SNA – System Volkswirtschaftlicher Gesamtrechnungen),
- › Verbesserung des Harmonisierten Verbraucherpreisindex und der Immobilienpreisstatistiken,
- › Mikrodaten-Linking im Bereich der Unternehmensstatistiken und andere innovative Datenquellen,
- › statistische Daten zur Digitalisierung,
- › Modernisierung und methodische Verbesserungen der Umweltgesamtrechnungen (UGR), der Ökosystemgesamtrechnungen, der Abfallstatistiken und der Statistiken zum Europäischen Green Deal,
- › Erhebungen 2026 zur Verwendung von Pflanzenschutzmitteln (SAIO),
- › Modernisierung der Sozialstatistiken, einschließlich Machbarkeits- und Pilotstudien, Überprüfung der Arbeitsmarktstatistiken für Unternehmen und weitere Implementierung der IESS-Verordnung.

Der AESS diskutierte die weitere Ausrichtung des EMOS-Labels an Hochschulen, da das Programm europaweit in seiner aktuellen Form nicht den ursprünglichen Erwartungen entspricht. Der „European Master in Official Statistics“ (EMOS) bezeichnet ein Zertifikat, das besondere Kenntnisse der amtlichen Statistik bescheinigt. Der AESS begrüßte mehrheitlich die in Aussicht gestellten Reformvorschläge, um die Attraktivität für die beteiligten Hochschulen und für die Teilnehmenden zu erhöhen. Der AESS sieht darüber hinaus weiteren Handlungsbedarf. Eurostat sagte zu, die im schriftlichen Verfahren geäußerten Anmerkungen zu sichten und dem AESS eine überarbeitete Unterlage vorzulegen.

Die europäische Datenstrategie betrifft insbesondere auch die amtliche Statistik in Europa und wird deshalb seit einiger Zeit in verschiedenen Gremien und Foren des ESS diskutiert. Eurostat legte in der aktuellen Sitzung ein Positionspapier vor und empfiehlt folgendes Vorgehen:

- › ein besseres Verständnis von Common European Data Spaces erlangen,
- › Datenräume untersuchen und priorisieren (vor allem Landwirtschaft, Energie, Mobilität, Gesundheit, Tourismus, Green Deal und Kompetenzen),
- › Chancen und Risiken der Datenräume erfassen sowie Informationen zu Datenmanagement und Interoperabilität der Daten beziehungsweise Datenräume einholen.

Um die bisher geleisteten Arbeiten zu intensivieren, empfiehlt der AESS mehrheitlich, eine Task Force einzusetzen. Eurostat wird diesen Vorschlag prüfen. Zudem wird die Positionierung der europäischen amtlichen Statistik auf der diesjährigen europäischen Statistikkonferenz (DGINS) in Tallinn, Estland, im Herbst 2024 weiter behandelt.

Eurostat stellte eine leicht modifizierte Geschäftsordnung für den AESS und die dem AESS vorgelagerte strategische Partnerschaftsgruppe vor. Im Kern geht es um die Bereiche Format und Dauer, Häufigkeit und Art der Sitzungen sowie um Verfügbarkeit der Sitzungsunterlagen. Für einige ESS-Mitgliedstaaten greifen die Vorschläge zu kurz, da strategische Aspekte in Bezug auf die Zusammenstellung der Tagesordnungspunkte unerwähnt bleiben. So wurde beispielsweise vorgeschlagen, dass Eurostat den AESS frühzeitig über geplante EU-Ver-

ordnungen mit Bezug zur amtlichen Statistik informiert. Die Diskussionen zur Geschäftsordnung werden in der nächsten Sitzung des AESS im Oktober 2024 fortgeführt.

AUS DEM INLAND

Indexrevision der Erzeugerpreisindizes für Dienstleistungen

Die Erzeugerpreisindizes für Dienstleistungen werden planmäßig in fünfjährigem Abstand einer Überarbeitung (Revision) unterzogen. Aufgrund der europäischen Rechtsgrundlagen ist das Jahr 2021 das aktuelle Basisjahr. Mit der Veröffentlichung der Ergebnisse für das erste Berichtsquartal 2024 am 19. Juni 2024 ist die Umstellung auf das Basisjahr 2021 = 100 erfolgt.

Bei der Überarbeitung wurden die Wägungsschemata auf der Grundlage der Umsatzstrukturen des Jahres 2021 neu berechnet und methodische Änderungen eingearbeitet. Damit verbunden ist auch eine Neuberechnung der Erzeugerpreisindizes ab dem ersten Quartal 2021. Die Veröffentlichungspositionen, die bis vor 2021 zurückreichen, werden für die Zeiträume vor 2021 nicht neu berechnet, sondern lediglich verkettet, das heißt formal auf das neue Basisjahr umgerechnet.

Die Umsetzung der neuen europäischen Rechtsgrundlagen machte es nötig, den Erfassungsbereich der Erzeugerpreisindizes für Dienstleistungen auszuweiten. Es war nicht für alle Branchen des Dienstleistungssektors möglich, die erforderlichen neuen Erhebungen aufzubauen. Die Erzeugerpreisindizes für Wirtschaftszweigabteilungen des Dienstleistungssektors werden deshalb teilweise um geeignete Schätzungen für die Preisentwicklung einzelner Branchen aus anderen Preisstatistiken ergänzt.

Auf Ebene der Wirtschaftsabschnitte und Wirtschaftsabteilungen wurde das Veröffentlichungsprogramm für Wirtschaftszweigindizes für Erzeugerpreise an die Veröffentlichungen des europäischen Statistikamts Eurostat und damit an die europäischen Rechtsgrundlagen angepasst. Über das europäische Veröffentlichungsprogramm hinaus stehen Erzeugerpreisindizes für die 3-stelligen Gruppen der Wirtschaftszweigsystematik und für Dienstleistungsarten in der Datenbank [GENESIS-Online](#) des Statistischen Bundesamtes zur Verfügung.

Eine Themenseite des Statistischen Bundesamtes hält über aktuelle Entwicklungen bezüglich der Erzeugerpreisindizes für Dienstleistungen auf dem Laufenden.

➤ www.destatis.de

VERANSTALTUNGEN

StatistikTage Bamberg|Fürth 2024

Das Bayerische Landesamt für Statistik und die Otto-Friedrich-Universität Bamberg organisieren im Rahmen des Statistik Netzwerks Bayern vom 11. bis 12. Juli 2024 die zwölften StatistikTage Bamberg|Fürth. Die Veranstaltung findet als reine Präsenzveranstaltung in der Aula der Otto-Friedrich-Universität Bamberg statt.

Das Thema der diesjährigen StatistikTage lautet „Zensus – jetzt und in Zukunft“. Im Jahr der Veröffentlichung der Zensusergebnisse erhält dieses Großprojekt der amtlichen Statistik besonders große Aufmerksamkeit. Die Ergebnisse des Zensus 2022 geben Aufschluss darüber, wie viele Menschen in Deutschland wohnen, wie sie leben und arbeiten. Sie sind maßgebend für zahlreiche finanz- und gesellschaftspolitische Entscheidungen.

Der erste Tag der StatistikTage beleuchtet die Themen Zensus international und in Deutschland, Modell und Durchführung des Zensus 2022 sowie Möglichkeiten zur Anwendung der Zensusdaten. Den Schwerpunkt des zweiten Tages bilden die Zukunftsperspektiven des Zensus. Erstmals wird die Veranstaltung mit einer hochrangig besetzten Podiumsdiskussion zum Thema „Zukunft des Zensus in Deutschland“ geschlossen.

➤ www.statistiknetzwerk.bayern.de

11. Konferenz „Forschen mit dem Mikrozensus“

Die 11. Konferenz „Forschen mit dem Mikrozensus“ findet am 14. und 15. November 2024 in Mannheim statt. Sie bietet eine Plattform zur Diskussion von inhaltlichen und methodischen Beiträgen aus allen Arbeitsgebieten der Sozialwissenschaften und angrenzenden Disziplinen. Darüber hinaus dient die Konferenz dem

Erfahrungsaustausch zwischen Datennutzenden und Vertreterinnen und Vertretern der amtlichen Statistik. Sie wendet sich sowohl an Wissenschaftlerinnen und Wissenschaftler, die bereits mit dem Mikrozensus arbeiten, als auch an jene, die dies planen.

Die Konferenz wird gemeinsam vom German Microdata Lab (GML) der GESIS, dem Statistischen Bundesamt und den Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder durchgeführt. Die Frist für den Call for Papers läuft noch bis zum 15. Juli 2024, die Anmeldefrist zur Teilnahme an der Konferenz bis zum 31. August 2024.

➤ www.gesis.org

Call for Abstracts für die NTS 2025

Die nächste Konferenz New Techniques and Technologies for Statistics (NTS) findet vom 11. bis 13. März 2025 in Brüssel statt.

Die internationale wissenschaftliche Konferenzreihe wird alle zwei Jahre von Eurostat organisiert und befasst sich mit neuen Techniken und Methoden für die amtliche Statistik sowie mit den Auswirkungen neuer Technologien auf die Systeme zur Erhebung, Erstellung und Verbreitung von Statistiken.

Ziel der Konferenz ist es, Ergebnisse aus laufenden Forschungs- und Innovationsprojekten in der amtlichen Statistik zu präsentieren und neue innovative Projekte anzuregen und zu erleichtern, um die Qualität und den Nutzen der amtlichen Statistik zu verbessern.

➤ cros.ec.europa.eu

Internationale Machine-Learning-Tagung in Wiesbaden

Vom 3. bis 5. April 2024 lud das Statistische Bundesamt zur internationalen Konferenz „Foundations and Advances of Machine Learning in Official Statistics“ nach Wiesbaden ein. Die Fachtagung zog Vortragende sowie Teilnehmende aus statistischen Ämtern und Zentralbanken im In- und Ausland, aus weiteren deutschen Behörden sowie von Universitäten an. Die Tagung

bot eine Reihe spannender Vorträge und ein breites Themenspektrum, das von konkreten Anwendungen maschineller Lernverfahren in der amtlichen Statistik über Querschnittsthemen wie Recht, Qualität und Prozessrationalisierung bis hin zu Fragestellungen der statistischen Methodik und der erforderlichen Technologie reichte.

Zur Tagungsdokumentation wurden alle Abstracts und die Foliensätze zu den Vorträgen auf der Webseite des Statistischen Bundesamtes veröffentlicht.

➤ www.destatis.de

AnigeD auf dem Kongress „Anonymisierung für eine sichere Datennutzung“

Am 16. und 17. April 2024 diskutierten in Lübeck mehr als 200 Teilnehmende aus Wissenschaft, Wirtschaft, Politik und Gesellschaft auf dem Kongress „Anonymisierung für eine sichere Datennutzung“ (AnoSiDat) über Themen wie Anonymisierungstechnologien, Datenschutz sowie sichere und innovative Datennutzung. Den Kongress ausgerichtet hat das vom Bundesministerium für Bildung und Forschung geförderte Forschungsnetzwerk Anonymisierung.

Das vom Statistischen Bundesamt geleitete Kompetenzcluster „Anonymität bei integrierten und georeferenzierten Daten“ (AnigeD) war auf dem Kongress durch Vorträge und Posterpräsentationen vertreten. Zusätzlich stellten die dem Cluster assoziierten Forschungsprojekte „Datenschutzkonforme Informationsfusion und Risikobewertung zur Prävention von Identitätsbetrug und Minderung von Ausfallrisiko“ (DARIA), „Anonymisierung von Gerichtsentscheidungen für E-Justice und Legal Tech“ (AnGer) und „Gewährleistung von Anonymitäts-Garantien in Enterprise-Streaminganwendungen“ (GANGES) ihre Arbeiten vor, was die Vielfalt und Breite der Forschungsaktivitäten des Clusters betonte. Die Projekte brachten ihre eigenen innovativen Ansätze und Ergebnisse in die Diskussion ein und bereicherten den Diskurs des Kongresses.

Weitere Informationen zu den Forschungsarbeiten im Kompetenzcluster AnigeD sind auf der Webseite des Projektes veröffentlicht.

➤ www.destatis.de

NEUERSCHEINUNGEN

Die Gesundheitsökonomischen Gesamtrechnungen der Länder

Zur statistischen Darstellung der Gesundheitswirtschaft hat die amtliche Statistik in Deutschland mit den Gesundheitsökonomischen Gesamtrechnungen (GGR) ein eigenes Rechenwerk entwickelt. Dieses enthält Daten zu Ausgaben, Beschäftigten und der Wertschöpfung ab 2008 auf Ebene der Bundesländer. Informationen zur Methodik und Ergebnisse für Niedersachsen enthält ein Beitrag in „Statistisch gesehen“, dem Online-Magazin des Landesamtes für Statistik Niedersachsen.

➤ magazin.statistik.niedersachsen.de

Neue Veröffentlichungen der OECD

Infrastruktur für eine klimaresiliente Zukunft

Globale Rekordtemperaturen von rund 1,4 Grad Celsius über dem vorindustriellen Durchschnitt haben 2023 zu mehr Hitzewellen und Überschwemmungen sowie Dürren und länger andauernden Waldbränden geführt. Solche klimatischen Ereignisse üben zunehmend Druck auf die Infrastruktur aller Sektoren aus – von der Stromversorgung über Kommunikations- und Verkehrsnetze bis hin zur Wasser- und Abfallwirtschaft.

„Infrastructure for a Climate-Resilient Future“ gibt einen Überblick über die Auswirkungen des Klimawandels auf die Infrastruktur und die wichtigsten Politikbereiche, die berücksichtigt werden müssen, um die Infrastruktur widerstandsfähiger zu machen. Der Bericht diskutiert Fortschritte und Lücken bei der Planung und Entwicklung von Infrastrukturen über ihren gesamten Lebenszyklus hinweg und informiert über standortbezogene Ansätze.

➤ manage.oecd-berlin.de

OECD-Wirtschaftsausblick – Frühjahrsausgabe 2024

Die Weltwirtschaft wächst weiterhin mit moderatem Tempo. Der Wirtschaftsausblick prognostiziert ein stabiles globales Wachstum des Bruttoinlandsprodukts von 3,1 % im Jahr 2024, nach 3,1 % im Jahr 2023, gefolgt von einem leichten Anstieg auf 3,2 % im Jahr 2025. Die Auswirkungen der restriktiven monetären Bedingungen sind nach wie vor spürbar, insbesondere auf den Immobilien- und Kreditmärkten, aber die globale Wirtschaftstätigkeit erweist sich als relativ widerstandsfähig, die Inflation geht weiter zurück und das Vertrauen des privaten Sektors verbessert sich.

Der „OECD Economic Outlook, Volume 2024 Issue 1“ liefert eine allgemeine Bewertung der makroökonomischen Lage und Projektionen zur gesamtwirtschaftlichen Produktion, Beschäftigung, Preisentwicklung, Haushaltssalden und Leistungsbilanzen. Der Bericht formuliert eine Reihe von Empfehlungen.

➤ manage.oecd-berlin.de

Taxing Wages 2024

Die hohen Inflationsraten der letzten zwei Jahre haben die Steuern und Abgaben auf Arbeit 2023 in den OECD-Ländern nach oben getrieben. Die effektiven Steuersätze auf Arbeitseinkommen sind in den meisten OECD-Ländern gestiegen. Gemessen am Nettoeinkommen Alleinstehender ging dadurch der Durchschnittsverdienst in 21 von 38 OECD-Ländern zurück.

„Taxing Wages 2024: Tax and Gender through the Lens of the Second Earner“ vergleicht Steuern und Sozialabgaben auf Arbeitseinkommen in allen 38 OECD-Ländern. Die Studie veranschaulicht, wie diese berechnet werden und sich auf die Einkommen auswirken und ermöglicht länderübergreifende Vergleiche für acht verschiedene Haushalts- und Einkommensstypen (Alleinstehende, Alleinerziehende, Ein- oder Zweiverdienerhaushalte, mit Kindern oder ohne Kind) hinsichtlich der Arbeitskosten sowie der gesamten Steuer- und Sozialleistungen. Die diesjährige Ausgabe von Taxing Wages enthält ein Sonderkapitel, das untersucht, wie sich die Steuerlast von Erstverdienenden und Zweitverdienenden unterscheidet.

➤ manage.oecd-berlin.de

OECD-Ausblick auf die digitale Wirtschaft 2024

Der Kommunikationstechnologiesektor (IKT) wuchs zwischen 2013 und 2023 um durchschnittlich 6,3 % und damit etwa dreimal so schnell wie die Gesamtwirtschaft in den 27 untersuchten OECD-Ländern. Auch im Jahr 2023 wird der Sektor diese starke Leistung mit einer durchschnittlichen Wachstumsrate von 7,6 % beibehalten. In vielen OECD-Ländern war 2023 ein Rekordjahr, wobei fünf OECD-Länder (das Vereinigte Königreich, Belgien, Deutschland, Österreich und die Niederlande) im Jahr 2023 Wachstumsraten von über 10 % erzielten.

Diese erste Ausgabe des „OECD Digital Economy Outlook“ bietet neue Einblicke in die Schlüsseltechnologien, die dem digitalen Technologieökosystem zugrunde liegen. Unter Verwendung von Big Data und maschinellen Lernverfahren liefert Band 1 neue Schätzungen der Wachstumsrate des IKT-Sektors. Zudem bietet der Bericht einen Ausblick auf die Zukunft der künstlichen Intelligenz (KI) und erläutert, wie sie sich zu einer positiven Kraft entwickeln kann.

➤ manage.oecd-berlin.de

ZUM EINFLUSS DES GRUNDRECHTS AUF INFORMATIONELLE SELBST- BESTIMMUNG AUF DIE BUNDES- STATISTIK

Sonja Leischner, Angela Kolbe

📌 **Schlüsselwörter:** informationelle Selbstbestimmung – Datenschutz – statistische Geheimhaltung – Bundesstatistikgesetz – Volkszählungsurteil

ZUSAMMENFASSUNG

Mit dem sogenannten Volkszählungsurteil von 1983 wurde in Deutschland das Grundrecht auf informationelle Selbstbestimmung eingeführt. Dieses hat das Datenschutzrecht in Deutschland auf neue Grundlagen gestellt sowie den Rechtsrahmen für die Bundesstatistik und deren Selbstverständnis maßgeblich beeinflusst. Der Beitrag erläutert, wie sich das Recht auf informationelle Selbstbestimmung auf die Entwicklung der Bundesstatistik seitdem ausgewirkt hat und welche rechtlichen Hürden und Chancen es bei deren Fortentwicklung gibt.

📌 **Keywords:** *informational self-determination – data protection – statistical confidentiality – Federal Statistics Act – population census judgment*

ABSTRACT

In Germany, the “population census judgment” of 1983 introduced the fundamental right of the individual to determine the use of his or her data, referred to as informational self-determination. It placed data protection legislation in Germany on a new footing, profoundly influenced the legal framework governing federal statistics and played a key role in shaping the philosophy that underpins federal statistics. This paper explains how the right to informational self-determination has influenced the development of federal statistics ever since and describes the legal obstacles and opportunities presented in connection with the further development of these statistics.



Dr. Sonja Leischner

ist seit 2017 im Statistischen Bundesamt tätig; sie war zuvor im Statistischen Landesamt Rheinland-Pfalz Referatsleiterin für Rechts- und Europaangelegenheiten sowie behördliche Datenschutzbeauftragte. Seit 2022 leitet sie die Gruppe „Recht, Compliance“ im Statistischen Bundesamt sowie den Bund-Länder-Arbeitskreis „Rechtsfragen der Statistik“.



Dr. Angela Kolbe

ist Volljuristin und seit 2011 im Statistischen Bundesamt tätig. Sie leitet das Referat „Statistikrecht“ und war zuvor behördliche Datenschutzbeauftragte und Leiterin des Datenschutzreferats.

1

Einleitung

Vor 40 Jahren hat das Bundesverfassungsgericht im sogenannten Volkszählungsurteil vom 15. Dezember 1983 das Grundrecht auf informationelle Selbstbestimmung formuliert (BVerfGE 65, 1). Dieses Grundrecht hat nicht nur das Datenschutzrecht in Deutschland auf neue Grundlagen gestellt, sondern auch den Rechtsrahmen für die Bundesstatistik und deren Selbstverständnis maßgeblich beeinflusst. Für bundesstatistische Erhebungen hat sich so über die Jahre eine Kultur etabliert, die neben den Charakteristika der Neutralität, Objektivität und fachlichen Unabhängigkeit durch die Fundamente der statistischen Geheimhaltung und Zweckbindung geprägt wird. Die Bundesstatistik muss aber zugleich dem inzwischen rasant fortgeschrittenen digitalen und gesellschaftlichen Wandel gerecht werden und sich kontinuierlich weiterentwickeln. Daher muss sie ausloten, welche Chancen das Recht auf informationelle Selbstbestimmung ihr in der Zukunft hierzu gewährt.

Wie hat sich das Recht auf informationelle Selbstbestimmung auf die Entwicklung der Bundesstatistik in den vergangenen 40 Jahren ausgewirkt? Welche rechtlichen Hürden und Chancen gibt es bei der künftigen Fortentwicklung der Bundesstatistik? Nach einer Einordnung des Volkszählungsurteils 1983 in Kapitel 2 und einem Rückblick auf das Zensusurteil 2011 in Kapitel 3 gibt Kapitel 4 einen Überblick darüber, welchen Herausforderungen sich die Bundesstatistik der Zukunft stellen muss. Bisherige und künftige Weiterentwicklungen des Statistikrechts werden in Kapitel 5 ausführlich diskutiert. Der Artikel schließt mit einem kurzen Ausblick.

2

Das Volkszählungsurteil 1983

Volkszählungen sind Erhebungen von Daten über die gesamte Bevölkerung eines Landes zu einem Stichtag. Sie sind in der Geschichte nicht neu, sondern bereits aus dem Altertum bekannt, beispielsweise aus dem Neuen Testament der Bibel. Nach dem Zweiten Weltkrieg und vor der deutschen Vereinigung wurden in beiden Teilen Deutschlands mehrere Volkszählungen durchgeführt.¹ Als mit der 1983 geplanten Volkszählung „der Staat seine Schäflein zählen wollte“, formierte sich Widerstand in der Bevölkerung (Widmann, 2023). Rechtsgrundlage des damals vorgesehenen Zensus war das sogenannte Volkszählungsgesetz 1983 vom 25. März 1982. Dass die bei den Befragten erhobenen Daten mit denen der Melderegister abgeglichen und erstmals mithilfe von Computern ausgewertet und gespeichert werden sollten, weckte in der Bevölkerung die Angst, zum „gläsernen Bürger“ zu werden. Dies führte zu Verfassungsbeschwerden verschiedener Bürgerinnen und Bürger, sodass sich das Bundesverfassungsgericht mit der Rechtmäßigkeit der geplanten Maßnahme auseinandersetzen musste.

2.1 „Geburtsstunde“ des Grundrechts auf informationelle Selbstbestimmung

Das Bundesverfassungsgericht stellte in seinem Urteil vom 15. Dezember 1983 (BVerfGE 65, 1) fest, dass zahlreiche Vorschriften des Volkszählungsgesetzes 1983 in Grundrechte des Einzelnen eingriffen. Diese Vorschriften erklärte es für nichtig und das gesamte Bundesgesetz für verfassungswidrig, da es die Beschwerdeführer und die Beschwerdeführerinnen in ihrem Recht auf informationelle Selbstbestimmung verletzte. Das Bundesverfassungsgericht leitete dieses, bisher nicht gegebene Recht aus Artikel 2 Absatz 1 Grundgesetz, dem Recht auf freie Entfaltung der Persönlichkeit, und aus Artikel 1 Absatz 1 Grundgesetz, der Unantastbarkeit der Menschenwürde, ab. Es schuf damit ein neues Grundrecht (Wieduwilt, 2024) und hob den Datenschutz auf Verfassungsrang.

1 In den Jahren 1950, 1956 (Gebäude- und Wohnungszählung), 1961, 1970 und 1987 im früheren Bundesgebiet sowie in der ehemaligen DDR in den Jahren 1950, 1964, 1971 und 1981.

Unter dem Recht auf informationelle Selbstbestimmung ist die Befugnis der einzelnen Person zu verstehen, grundsätzlich selbst über die Preisgabe und Verwendung ihrer personenbezogenen Daten zu bestimmen.¹² Dieses Recht ist jedoch nicht schrankenlos gewährleistet. Die einzelne Person hat kein Recht im Sinne einer absoluten, uneinschränkbaren Herrschaft über „ihre“ Daten; sie ist vielmehr eine sich innerhalb der sozialen Gemeinschaft entfaltende, auf Kommunikation angewiesene Persönlichkeit. Grundsätzlich muss die einzelne Person daher Einschränkungen ihres Rechts auf informationelle Selbstbestimmung im überwiegenden Allgemeininteresse hinnehmen, die allerdings auf gesetzliche Grundlagen gestützt werden müssen.¹³

Auch wenn das Bundesverfassungsgericht die Vorgehensweise der Bundesstatistik grundsätzlich nicht infrage gestellt hat, hat es den Gesetzgeber hinsichtlich mancher Aspekte der geplanten Volkszählung zur Nachbesserung aufgefordert.

Die im Volkszählungsurteil 1983 getroffenen Vorgaben und die dort aufgestellten Grundsätze haben dem Statistikrecht seine immer noch geltende Verfasstheit gegeben (siehe hierzu Statistisches Bundesamt, 1985).

2.2 Auswirkungen auf das Statistikrecht

Das Bundesverfassungsgericht erkannte an, dass für die Verarbeitung von Daten für statistische Zwecke bestimmte Besonderheiten gelten, die auch das Verfassungsrecht berücksichtigen müsse. So könne bei einer Datenerhebung für statistische Zwecke eine enge und konkrete Zweckbestimmung der Daten nicht verlangt werden, da es zum Wesen der Statistik gehöre, dass Daten für die verschiedensten, nicht von vornherein bestimmbar Aufgaben verwendet werden sollen. Gerade mit einer Volkszählung solle eine Datenbasis geschaffen werden, die für weitere statistische Untersuchungen sowie für politische Planungsprozesse zur Verfügung steht.¹⁴

Weiterhin gab das Bundesverfassungsgericht vor, dass es zur Sicherung des Rechts auf informationelle Selbstbestimmung besonderer Vorkehrungen hinsichtlich der

Vorbereitung und Durchführung statistischer Erhebungen bedürfe, da die erhobenen Daten während des Aufbereitungsprozesses noch eine Zeitlang identifizierbar seien. Zudem müssten bezüglich dieser identifizierenden Merkmale (sogenannte Hilfsmerkmale) Löschungsregelungen vorgesehen werden. Und schließlich seien wirksame Vorkehrungen zur Abschottung nach außen unabdingbar. Nur unter diesen Voraussetzungen eröffne sich der Zugang der staatlichen Organe zu den für die Planungsaufgaben erforderlichen Informationen und könne und dürfe von den Bürgerinnen und Bürgern erwartet werden, die von ihnen zwangsweise verlangten Auskünfte zu erteilen. Dürften personenbezogene Daten, die zu statistischen Zwecken erhoben wurden, gegen den Willen oder ohne Kenntnis der Betroffenen weitergeleitet werden, so würde das nicht nur das verfassungsrechtlich gesicherte Recht auf informationelle Selbstbestimmung unzulässig einschränken, sondern auch die vom Grundgesetz selbst in Artikel 73 Nr. 11 vorgesehene und damit schutzwürdige amtliche Statistik gefährden. Denn für die Funktionsfähigkeit der amtlichen Statistik ist ein möglichst hoher Grad an Genauigkeit und Wahrheitsgehalt der erhobenen Daten notwendig. Dieses Ziel könne nur erreicht werden, wenn bei den Auskunftgebenden das notwendige Vertrauen in die Abschottung ihrer für statistische Zwecke erhobenen Daten geschaffen werde.¹⁵

Diesbezüglich bestehen für den Gesetzgeber Informationspflichten; insbesondere bei Massenerhebungen sind Befragte über ihre Rechte im Vorhinein aufzuklären.¹⁶

So wurde insbesondere die Verbindung der Volkszählung für statistische Zwecke mit dem Melderegisterabgleich der Kommunen nach § 9 Absatz 1 Volkszählungsgesetz 1983 als verfassungswidrig angesehen, denn hier waren nach Ansicht des Bundesverfassungsgerichts unvereinbare Zwecke miteinander kombiniert worden.¹⁷ Die beiden Zwecke (Erstellung der Statistik und Durchführung des Melderegisterabgleichs) schlossen sich gegenseitig aus, da die Beachtung des Statistikgeheimnisses mit den weitreichenden Übermittlungsregelungen des Melderechtes unvereinbar sei.¹⁸

2 BVerfGE 65, 1 (43).

3 BVerfGE 65, 1 (34).

4 BVerfGE 65, 1 (36).

5 BVerfGE 65, 1 (37 f.).

6 BVerfGE 65, 1 (44).

7 BVerfGE 65, 1 (46).

8 BVerfGE 65, 1 (47).

Weiterhin sah das Bundesverfassungsgericht die Übermittlung statistischer Einzelangaben an oberste Bundes- oder Landesbehörden sowie Gemeinden zur Verwendung für deren Aufgabenerfüllung – also für nicht statistische Zwecke – aus den gleichen Gründen als mit der Verfassung unvereinbar an.⁹ Zudem sei nicht hinreichend erkennbar gewesen, zu welchem konkreten Zweck die Daten weitergegeben werden, insbesondere ob nur zu statistischen oder auch zu Verwaltungsvollzugszwecken.¹⁰ Nach Auffassung des Bundesverfassungsgerichts fehlten im gemeindlichen Bereich überdies organisatorische Vorkehrungen, welche die Zweckbindung der Daten garantierten; erforderlich sei die Trennung der Kommunalstatistik von anderen Aufgabenbereichen der Gemeinde.¹¹

Demgegenüber wurde eine Übermittlung von Daten für wissenschaftliche Zwecke an Amtsträger und für den öffentlichen Dienst besonders Verpflichtete als verfassungsgemäß angesehen, soweit sich die Übermittlung in den Grenzen des für wissenschaftliche Zwecke Erforderlichen hält; Name und Anschrift dürfen nicht weitergegeben werden.¹²

Die Vorgaben des Volkszählungsurteils 1983 wurden sodann im Volkszählungsgesetz von 1987 berücksichtigt und die Volkszählung im selben Jahr durchgeführt. Auch in den derzeitigen Regelungen des Bundesstatistikgesetzes finden sich die Vorgaben des Bundesverfassungsgerichts wieder (zu den Auswirkungen des Volkszählungsurteils auf das Bundesstatistikgesetz siehe Statistisches Bundesamt, 1988).

9 BVerfGE 65, 1 (48).

10 BVerfGE 65, 1 (48 ff.).

11 BVerfGE 65, 1 (49). Dieser Vorgabe trägt mittlerweile die Regelung des § 16 Absatz 5 Bundesstatistikgesetz Rechnung.

12 BVerfGE 65, 1 (50).

3

Das Zensusurteil 2011

Bei der folgenden Volkszählung – dem Zensus 2011 – wurde nicht nur eine neue Methode zur Ermittlung der Bevölkerungszahlen angewendet (erstmalig wurden hier Register als weitere Datenquellen herangezogen), auch öffentliche Wahrnehmung und Kritik fielen deutlich geringer aus. Dennoch fand sich auch der Zensus 2011 vor dem Bundesverfassungsgericht wieder. Kritikpunkte der klagenden Stadtstaaten Hamburg und Berlin waren insbesondere die angewandte Methodik der Stichprobe sowie die sich daraus mittelbar ergebenden finanziellen Folgen für die Bundesländer.

Bereits im Volkszählungsurteil von 1983 hatte das Bundesverfassungsgericht gefordert, dass wegen der erforderlichen Vielfalt der Verwendungs- und Verknüpfungsmöglichkeiten der statistischen Informationsverarbeitung zum Ausgleich entsprechende Schranken gegenüberstehen müssen. Es sind also immer Vorkehrungen zu treffen, die einer Verletzung des Rechts auf informationelle Selbstbestimmung entgegenwirken (Bierschenk/Leischner, 2019, hier: Seite 15 ff.; Kienle, 2018; Leischner/Weigelt, 2019, hier: Seite 1731). Diese Schranken hat das Bundesverfassungsgericht im Zensusurteil 2011 ausdrücklich bestätigt.¹³

Mit Blick auf die besonderen Gefährdungen, die sich durch die Nutzung automatisierter Datenverarbeitung ergeben, gelte weiterhin Folgendes: Soweit das Grundrecht auf informationelle Selbstbestimmung durch ein Gesetz beschränkt werde, sind durch Gesetz organisatorische und verfahrensrechtliche Vorkehrungen zu treffen, die einer Verletzung dieses Grundrechts entgegenwirken.¹⁴ Hierzu zählen Transparenz, aufsichtliche Kontrolle und ein effektiver Rechtsschutz bei der Speicherung und Nutzung personenbezogener Daten, organisatorischer und verfahrensrechtlicher Schutz gegen Zweckentfremdung durch Weitergabe- und Verwertungsverbote sowie Aufklärungs-, Auskunfts- und Löschpflichten.¹⁵ Neben der Bestätigung der im Volkszählungsurteil 1983 getroffenen Vorgaben zu Geheimhaltung, Rückspielverbot und Abschottung stellte das Bundes-

13 BVerfGE 150, 1 ff.

14 BVerfG, Urteil vom 19. September 2018, Rz. 221.

15 BVerfG, Urteil vom 19. September 2018, Rz. 223.

verfassungsgericht nun auch fest, dass der Grundsatz der Verhältnismäßigkeit die Prüfung beinhalte, ob aufgrund der Fortentwicklung der statistischen Wissenschaft Möglichkeiten einer grundrechtsschonenderen Datenerhebung bestehen.¹⁶

Der registerbasierte Zensus 2011 hielt nach Wertung des Bundesverfassungsgerichts diesen Maßstäben stand. Auch künftige bundesstatistische Erhebungen haben sich an den Anforderungen des Rechts auf informationelle Selbstbestimmung zu orientieren und sie zu wahren.

4

Die Bundesstatistik der Zukunft

4.1 Der Anspruch: zukunftssicher, funktionsfähig und realitätsgerecht

Die Statistik für Bundeszwecke hat nach § 1 Bundesstatistikgesetz die Aufgabe, unter Verwendung wissenschaftlicher Erkenntnisse sowie unter Einsatz der jeweils sachgerechten Methoden und Informationstechniken Daten zu gewinnen. Durch ihre Ergebnisse werden gesellschaftliche, wirtschaftliche und ökologische Zusammenhänge für Bund, Länder, Gemeinden und Gemeindeverbände, Gesellschaft, Wirtschaft, Wissenschaft und Forschung aufgeschlüsselt. Nur wenn es der Bundesstatistik gelingt, eine empirische Grundlage für fundierte Entscheidungen zu schaffen, ist eine am Sozialstaatsprinzip ausgerichtete Politik gesichert, wie sie vom Gesetzgeber und der Verfassung¹⁷ ausdrücklich gefordert wird. Ebenso wie die Funktionsfähigkeit der Statistik verfassungsrechtlich vorgegeben ist, trifft dies auch für die Notwendigkeit einer realitätsgerechten Statistik zu. Die Bundesstatistik muss mithin leistungsfähig sein (Kühling, 2023, hier: Einleitung, Rnr. 71).

Die Corona-Pandemie hat gezeigt, wie wichtig das flexible Vorhalten umfassender statistischer Daten für politische Entscheidungsprozesse ist (Kühling, 2023, hier: Rnr. 88) und dass die amtliche Statistik hier an ihre Grenzen stoßen kann. Qualitativ hochwertige Statistiken

sind gerade in Krisenzeiten eine wichtige Grundlage für faktenbasierte Entscheidungen (Schliffka/Polus, 2020).

Zur Deckung kurzfristiger Datenbedarfe ist dabei von besonderer Relevanz, aus neuen administrativen und privaten Datenquellen [experimentelle Statistiken](#) bereitzustellen (Schliffka/Polus, 2020). Mit der digitalen Revolution hat sich die Lebenswirklichkeit nachhaltig gewandelt (Wiengarten/Zwick, 2017, hier: Seite 20); Nutzerinnen und Nutzer erwarten maßgeschneiderte statistische Analysen quasi auf Knopfdruck, die ihnen genau dann angeboten werden, wenn sie sie brauchen (Riede und andere, 2018, hier: Seite 103). In der Big-Data-Welt haben neue digitale Daten das Potenzial, amtliche Statistiken erheblich zu optimieren und dabei die Auskunftsgebenden spürbar zu entlasten (Wiengarten/Zwick, 2017, hier: Seite 21).

Das umfangreiche Produktportfolio und das Veröffentlichungsprogramm des Statistischen Amtes der Europäischen Union (Eurostat) verdeutlichen, wie umfassend die Anforderungen der Europäischen Union (EU) an die Produktion von Bundesstatistiken sind (O'Donnell, 2006; Klumpen/Köhler, 2003). Die Rahmenverordnung für Unternehmensstatistiken, der Verhaltenskodex für europäische Statistiken (sogenannter Code of Practice; Eurostat, 2017), die Europäische Statistikverordnung und andere europäische Rahmenverordnungen begründen diese Anforderungen.

Auch die Wissenschaft, die eigenen Verfassungsrang genießt (Artikel 5 Absatz 3 Grundgesetz), steigert kontinuierlich ihren Bedarf an Daten. Der Rechtsrahmen über den Zugang der Wissenschaft zu statistischen Daten ist aus diesem Grund forschungsfreundlich auszugestalten (Kühling/Sauerborn, in Kühling, 2023, hier: § 16, Rnr. 53).

Das grundlegende Bedürfnis, statistische Daten über die bisherigen Zusammenführungsbeschränkungen des § 13a Bundesstatistikgesetz hinaus zu verknüpfen, ist inzwischen auch in der Finanzverwaltung angekommen. Mit dem Ziel einer verbesserten evidenzbasierten Steuergesetzgebung sucht das Netzwerk für empirische Steuerforschung (NeSt)¹⁸ nach Möglichkeiten, auch

¹⁶ BVerfG, Urteil vom 19. September 2028, Rz. 226.

¹⁷ Sozialstaatsprinzip des Artikels 20 Absatz 1 Grundgesetz.

¹⁸ Das [Netzwerk empirische Steuerforschung \(NeSt\)](#) ist eine unter dem Dach des Bundesministeriums der Finanzen errichtete und von ihm betriebene offene und interdisziplinär ausgerichtete Plattform. Sie dient insbesondere der Vernetzung von empirisch Forschenden auf dem Gebiet der Besteuerung mit der amtlichen Statistik und der Finanzverwaltung.

durch die Verknüpfung von Daten mit steuerstatistischen Daten eine verbesserte Datenbasis zu gewinnen.

Die Kommission Zukunft Statistik¹⁹ legt in ihrem Bericht vom 15. Januar 2024 mehrere Handlungsempfehlungen vor, um unter anderem die Datenverfügbarkeit zu verbessern. Einhergehen soll die verbesserte Datenverfügbarkeit mit einer Reform des Bundesstatistikgesetzes – weg von der Regelung der Datenerhebung (Inputseite) hin zur Regelung der Ergebnisse der Statistik (Outputseite).

4.2 Neue Technologien

Das Volkszählungsurteil und die Begründung des Rechts auf informationelle Selbstbestimmung entstammen einer Zeit, in der die automatisierte Datenverarbeitung im Vergleich zu heute zwar noch in den Kinderschuhen steckte, aber dabei war, an Fahrt aufzunehmen.

Entwicklungsschritte im privaten Bereich gingen über den Heimcomputer und erste Mobiltelefone mit geringer Speicherkapazität über Smartphones mit der Möglichkeit, mobile Anwendungen (Apps) zu nutzen, sowie die weitreichende Nutzung der Social-Media-Plattformen. Im wirtschaftlichen Bereich sind die Einführung von Großrechnersystemen, das Teilen von Computerressourcen in Form von Servern sowie Cloud-Lösungen hervorzuheben. Parallel dazu hat auch die amtliche Statistik vom technischen Fortschritt profitiert: Auskunftgebende können online – teilweise mit Unterstützung von Apps – melden, die Eingangsprüfung der Daten erfolgt immer stärker auch mithilfe von Verfahren der Künstlichen Intelligenz (KI) und nicht mehr manuell. Durch eine gut ausgebaute IT-Infrastruktur lassen sich die Daten schnell und sicher übermitteln und die statistischen Ergebnisse so mit einem hohen Anspruch an Validität und Aktualität veröffentlichen. Die Methoden der Datenverarbeitung vor der Formulierung des Grundrechts auf informationelle Selbstbestimmung, als beispielsweise Lochkarten zum Einsatz kamen, lassen sich mit denen im Jahr 2024 somit kaum vergleichen. Die Frage, ob die in den 1980er-Jahren aufgestellten Grundsätze auch heute noch gültig sind, erscheint daher legitim.

Im Hinblick auf die technische Entwicklung ist hiervon vor allem das Abschottungs- und Trennungsgebot relevant. Es besagt, dass die Verarbeitung von statistischen Einzeldatensätzen in einem von den übrigen Verwaltungsstrukturen getrennten Bereich erfolgen muss. Dem entsprechend waren die Daten des Statistischen Bundesamtes zunächst auf eigener Hardware, zum Beispiel einem Großrechner, gespeichert, danach – mit zunehmender Zentralisierung beziehungsweise Konsolidierung der Informationstechnik in der Bundesverwaltung – erst durch die Bundesstelle für Informationstechnik (BIT) und mittlerweile durch das Informationstechnikzentrum Bund (ITZBund) auf deren Servern.

Die Bedeutung und die rasante Entwicklung technischer Möglichkeiten hat auch der Verfassungsgeber erkannt und die Regelungen des Artikels 91c Absätze 1 und 3 in das Grundgesetz eingefügt. Danach können Bund und Länder bei Planung, Errichtung und Betrieb der für ihre Aufgabenerfüllung benötigten informationstechnischen Systeme zusammenwirken. Das verdeutlicht, dass eine leistungsfähige IT-Infrastruktur der öffentlichen Verwaltung ein Gut von Verfassungsrang ist (Kühling, 2023, hier: Rnr. 48).

Die IT-gestützte Datenverarbeitung ist mittlerweile von besonderer Bedeutung für die Funktionsfähigkeit der Bundesstatistik. Um schnelle Ergebnisse erzielen, flexibel auf neue Datenbedarfe reagieren und weiterhin relevant bleiben zu können benötigt die Statistik State-of-the-art-Methoden, wie große Rechnerkapazitäten, Einsatz von Machine Learning und KI. Nur mit einer zeitgemäßen IT-Infrastruktur ist es möglich, die komplexen gesellschaftlichen Strukturen statistisch abzubilden.

Die Vorteile dieser Technologien müssen allerdings deren Nachteile überwiegen, um etwaige Eingriffe in das Grundrecht auf informationelle Selbstbestimmung zu rechtfertigen (Kühling, 2023, hier: Rnr. 47 ff.). Dann steht das Abschottungs- und Trennungsgebot einer Verarbeitung statistischer Daten auch auf virtuellen Servern nicht entgegen.

Es ist somit möglich und geboten, die Grundsätze des 40 Jahre alten Volkszählungsurteils auf heutige technische Grundlagen und Methoden der Statistikproduktion anzuwenden.

19 Die [Kommission Zukunft Statistik \(KomZS\)](#) wurde vom Statistischen Bundesamt eingerichtet und mit der Erarbeitung von Empfehlungen für eine vorausschauende Programmplanung und eines Zielbilds der amtlichen Statistik für das Jahr 2030 beauftragt. Der Abschlussbericht der Kommission wurde am 16. Januar 2024 der Amtsleitung des Statistischen Bundesamtes überreicht.

4.3 Der Registerzensus als fachliches Zukunftsprojekt

Die nächste EU-weite Zensusrunde findet im Jahr 2031 statt. Der vorliegende europäische Verordnungsentwurf sieht vor, dass die Mitgliedstaaten jährlich Bevölkerungszahlen bereitstellen. Da das bisherige Verfahren die zu erwartende Verkürzung der Periodizität nicht erfüllen kann, wird die Methode der Volkszählung erneut weiterentwickelt. Von der Totalerhebung durch Erhebungsbeauftragte über den registergestützten Zensus der Jahre 2011 und 2022 wandelt sich die Methodik der Ermittlung der Einwohnerzahlen zu einem Zensus, der perspektivisch möglichst rein registerbasiert erfolgen soll. Demnach sollen Daten, die bereits in der Verwaltung oder Statistik vorliegen, für den Zensus genutzt werden, ohne diese Daten spezifisch zu Zensuszwecken nochmals bei den Bürgerinnen und Bürgern erheben zu müssen (Söllner/Körner, 2022).

Den Wandel der Methodik hat letztlich auch national das Bundesverfassungsgericht vorgegeben: Der Gesetzgeber ist aufgerufen, aufgrund der stetigen wissenschaftlichen Entwicklungen in der Statistik stets die „Möglichkeiten einer grundrechtsschonenderen Datenerhebung“ zu prüfen und anzuwenden.¹²⁰

Das Bundesverfassungsgericht stellte im Zensusurteil 2011 fest, dass die Nutzung von bereits vorhandenen Daten aus Verwaltungsregistern die Belastung von Auskunftgebenden reduziere. Zudem führe sie im Vergleich zur Vollerhebung zu einer geringeren Eingriffstiefe in das Grundrecht auf informationelle Selbstbestimmung. Grund dafür sei, dass die Datenübermittlungen lediglich solche Daten betreffen würden, die in den genutzten Verwaltungsregistern ohnehin vorliegen und bereits erhoben wurden (Bierschenk/Leischner, 2019, hier: Seite 12).¹²¹ Dieses grundrechtsschonendere Verfahren wird beim Registerzensus mit der Nutzung der Register mit dem Once-Only-Prinzip umgesetzt (Thiel/Puth, 2023, hier: Seite 307 f.; Gößl, 2019, hier: Seite 49).

Weder verstößt die vorgesehene Verknüpfung von Daten aus unterschiedlichen Quellen gegen das informationelle Selbstbestimmungsrecht, noch birgt sie die damit grundsätzlich einhergehende Gefahr einer Profilbildung.

Denn die Datenverarbeitung zu statistischen Zwecken erfolgt niemals, um Profile oder Prognosen einzelner Personen zu erstellen. Vielmehr hat die Bundesstatistik die Aufgabe, Massenerscheinungen abzubilden, unter anderem als Grundlagen für politische Entscheidungen. Im Registerzensus erfolgende Zusammenführungen werden stets im von der Verwaltung abgeschotteten Bereich der statistischen Ämter stattfinden, sodass keine Daten an die Verwaltung übermittelt werden. Aktive gesetzliche, organisatorische und technische Schutzmaßnahmen sorgen für eine effektive Trennung von jeglichen Exekutivzwecken anderer Behörden. Beispiele dafür sind die gesonderte Sicherung und getrennte Speicherung von für die Zuordnung der Person erforderlichen Merkmalen sowie Rechte- und Rollenkonzepte, die Zugriffsrechte auf die Daten auf absolute Notwendigkeiten beschränken (Thiel/Puth, 2023, hier: Seite 308).

Im Juni 2021 ist das Registerzensuserprobungsgesetz in Kraft getreten. Mit dieser Rechtsgrundlage kann ein registerbasiertes Verfahren der Datenerhebung umfassend erprobt werden. Zudem werden die zuverlässige Verknüpfung der Registerdaten und ein Verfahren zur Klärung von Unstimmigkeiten in Bezug zum Wohnsitz einer Person, die sogenannte Wohnsitzanalyse, untersucht.

5

Weiterentwicklung des Statistikrechts

Die Rechtsprechung des Bundesverfassungsgerichts belässt Legislative und Exekutive einen weiten Spielraum, alle Voraussetzungen für eine funktionsfähige Statistik zu schaffen (Kühling, 2023, hier: Rnr. 89). Dies kann zum einen durch eine funktionsgerechte Interpretation der Anforderungen des Bundesverfassungsgerichts an das Recht auf informationelle Selbstbestimmung erreicht werden und zum anderen durch eine Anpassung des Statistikrechts.

5.1 Bisherige Fortentwicklung der Gesetzgebung

In der Vergangenheit hat der Gesetzgeber diesbezüglich unterschiedliche Gesetzgebungsverfahren auf den Weg gebracht.

20 BVerfGE 65, 1, S. 41; BVerfGE 89, 1 (124 f.).

21 BVerfGE 89, 1104.

Ein wichtiger Schritt für die Bundesstatistik erfolgte bereits im Jahr 2005 mit Einführung des § 3a Bundesstatistikgesetz. Die Regelung stellt als besondere Form der Amtshilfe die Möglichkeit der arbeitsteiligen Zusammenarbeit von Bund und Ländern nach dem Prinzip „Einer oder einige für alle“ klar. Sie war Ausgangspunkt für eine Intensivierung der Zusammenarbeit im Statistischen Verbund²² mit dem Ziel, die Durchführung der Bundesstatistik effizienter und wirtschaftlicher zu gestalten.

Dem Fortschritt der Digitalisierung Rechnung tragend hat die 2013 eingeführte Regelung des § 11a Bundesstatistikgesetz die Rechtsgrundlage dafür geschaffen, dass von Unternehmen und Betrieben unter bestimmten Voraussetzungen zur Erfüllung der statistischen Auskunftspflicht elektronische Verfahren zur Datenübermittlung zu nutzen sind.

Mit der Änderung des § 10 Absatz 2 Bundesstatistikgesetz wurde ein „Meilenstein hin zu einer flexibleren räumlichen Auswertung bundesstatistischer Daten gelegt“ (Isfort/Dommermuth, 2023), indem bundesstatistische Angaben mit Bezug auf quadratische Gitterzellen mit einer Fläche von mindestens einem Hektar gespeichert werden dürfen.

Im Jahr 2016 erfolgte eine gebündelte Novellierung des Bundesstatistikgesetzes (Engelter/Sommer, 2016). Der neu eingeführte § 5a Bundesstatistikgesetz enthält die Prüfpflicht des Statistischen Bundesamtes, ob bereits vorhandene Daten öffentlicher Stellen zur Erstellung einer Bundesstatistik qualitativ geeignet sind; er trägt damit zu einer Verbesserung des Once-Only-Prinzips bei. Eine Anpassung des § 16 Absatz 6 Bundesstatistikgesetz ermöglicht der Wissenschaft grundsätzlich Zugriff auf formal anonymisierte Einzelangaben innerhalb speziell abgesicherter Bereiche statt wie bisher lediglich auf faktisch anonymisierte Daten.

Im Zuge der Umsetzung der europäischen Verordnung zu Unternehmensstatistiken im Jahr 2020 wurde § 5a Bundesstatistikgesetz erneut angepasst: Einerseits wurde er hinsichtlich der Möglichkeiten der Eignungsprüfung fortentwickelt, andererseits enthält er nun die rechtlichen Rahmenbedingungen für die Einrichtung einer Verwaltungsdaten-Informationsplattform.

Eine weitere Entlastung der Auskunftgebenden und größere Flexibilisierung der amtlichen Statistik (Isfort/Dommermuth, 2023) wurde dadurch erreicht, dass auf direkte Befragungen verzichtet werden kann, wenn Angaben vorangegangener Erhebungen eine entsprechende Erhebung ersetzen können. Gleiches gilt für die Verwendung von Daten aus Wirtschafts- und Umweltstatistiken untereinander sowie für die Nutzung von Informationen aus allgemein zugänglichen Quellen.

5.2 Statistikrecht in einer „Verrechtlichungsfalle“?

Den Handlungsspielraum der amtlichen Statistik bestimmen zunehmend verschiedene Ebenen der Verrechtlichung (Hoffmann-Riem, 1998): Die Zahl rechtlicher Regelungen, um das informationsbezogene Handeln zu legitimieren, nimmt zu: Im Bundesdatenschutzgesetz, in der europäischen Datenschutz-Grundverordnung und in bereichsspezifischen Normen sind Regelungen verstreut, ebenso besteht ein Spannungsverhältnis zwischen dem Recht auf informationelle Selbstbestimmung und der Normenklarheit (Franzius, 2015, hier: Seite 264). Einerseits fordert die Verfassung eine funktionsfähige und realitätsgerechte Statistik und letztendlich leistungsfähige Verwaltung (Kühling, 2023, hier: Einleitung, Rnr. 71). Diese muss aber andererseits den verfassungsrechtlichen Restriktionen genügen.

Schon nach dem Volkszählungsurteil 1983 kam die Frage auf, ob Datenschutz und Statistik Gegensätze oder „im Prinzip natürliche Verbündete“ seien (Poppenhäger, 1995, hier: Seite 20 mit Hinweis auf Hölder, 1985, hier: Seite 56). Im Zentrum der Kritik standen die „Überdehnung“ des Rechts auf informationelle Selbstbestimmung und die Forderungen nach „Abrüstung“ verfassungsrechtlicher Vorgaben (Poppenhäger, 1995, hier: Seite 20; Bull, 2011, hier: Seite 36).

Zur Lösung dieses Spannungsverhältnisses darf das Grundrecht auf informationelle Selbstbestimmung allerdings nicht neu konzipiert werden (Franzius, 2015), sondern es ist zu prüfen, ob die bundesstatistischen Regelungen – insbesondere das Bundesstatistikgesetz – nicht selbst die Anforderungen des Rechts auf informationelle Selbstbestimmung überspannen oder „überinterpretieren“. Denn das Bundesverfassungsgericht hat bereits im Volkszählungsurteil 1983 festgehalten, dass

22 Den Statistischen Verbund bilden die Statistischen Ämter des Bundes und der Länder.

das Recht auf informationelle Selbstbestimmung besonders schützenswert sei, allerdings unter den heutigen und künftigen Bedingungen der automatisierten Datenverarbeitung.¹²³ Mit Blick auf den möglichen Fortschritt der Digitalisierung sind die Aussagen des Bundesverfassungsgerichts daher vor dem Hintergrund des jeweils aktuellen Standes der Technik „funktional“ zu interpretieren (Kühling, 2023, hier: Rnr. 47).

5.3 Mögliche künftige Fortentwicklung der Gesetzgebung

Das Bundesstatistikgesetz als gesetzlicher Rahmen der Bundesstatistik bleibt an vielen Stellen hinter den Potenzialen zurück, die eine optimierte Wirksamkeit bundesstatistischer Erhebungen ermöglichen würden.

Flexibilisierung bundesstatistischer Erhebungen

Die geforderte Flexibilisierung des bundesstatistischen Rechtsrahmens hin zu einer outputorientierten Gesetzgebung (Kühling/Schmid in Kühling, 2023, hier: § 5a, Rnr. 57) wäre durch verschiedene Anpassungen des Bundesstatistikgesetzes aussichtsreich, ohne dass sie mit verfassungsrechtlichen Vorgaben kollidiert.

Haben Verwaltungsdaten das qualitative Prüfverfahren des Statistischen Bundesamtes nach § 5 Absatz 4 Bundesstatistikgesetz erfolgreich durchlaufen, könnte eine erweiterte Nutzung für statistische Zwecke realisiert werden, wenn sie ohne weitergehende gesetzliche Regelung für konkrete bundesstatistische Zwecke übermittelt werden müssen. Wären jedoch personenbezogene Daten von der Übermittlung betroffen, besteht mit Blick auf das informationelle Selbstbestimmungsrecht ein Gesetzesvorbehalt (Kühling/Schmid in Kühling, 2023, hier: § 5a, Rnr. 57).

Eine weitere Möglichkeit zur Flexibilisierung besteht darin, dass die Bundesstatistiken künftig auf einem vom Statistischen Bundesamt erarbeiteten Programm basieren, das gesetzlich zu regeln wäre¹²⁴ (Kühling und andere in Kühling, 2023, hier: § 9, Rnr. 22; Kühling/Schmid in Kühling, 2023, hier: § 5, Rnr. 59).

Oft ist bei Schaffung einer Rechtsgrundlage für eine statistische Erhebung vorhersehbar, dass sich die Merkmalsausprägungen ändern können. Insbesondere bei Unternehmens- und Betriebsstatistiken bietet es sich in diesen Fällen an, anstelle von konkreten Merkmalen Merkmalskomplexe in der die Statistik anordnenden Rechtsvorschrift aufzunehmen oder Verweise auf Merkmalsvorgaben in europäischen Regelungen (Kühling, 2023, hier: § 10, Rnr. 35 f.).

Um in nicht vorhersehbaren Situationen, beispielsweise bei einer Pandemie, auf Datenbedarfe reagieren zu können, sollte im Bundesstatistikgesetz eine Regelung aufgenommen werden,¹²⁵ nach der eine oberste Bundesbehörde über die Möglichkeiten des § 7 Bundesstatistikgesetz hinaus eine Bundesstatistik anordnen kann, um aufgrund von nationalen Krisensituationen auftretende Datenbedarfe abzudecken.

Erweiterung der Verknüpfungsmöglichkeiten statistischer Daten

Die Zusammenführungsregelungen des § 13a Bundesstatistikgesetz sind im Hinblick auf die gestiegenen Datenbedarfe in ihrer aktuellen Form sehr restriktiv gehalten, obwohl gerade bei statistischen Erhebungen eine enge und konkrete Zweckbindung nicht verlangt wird.¹²⁶ Gefahren einer Verletzung des informationellen Selbstbestimmungsrechts könne durch organisatorische und verfahrensrechtliche Vorkehrungen entgegengewirkt werden.¹²⁷ Entsprechende Möglichkeiten haben sich durch den technischen Fortschritt deutlich weiterentwickelt. Im Hinblick auf die verfassungsrechtliche Forderung, dass stets zu prüfen ist, ob aufgrund der Fortentwicklung der statistischen Wissenschaft Möglichkeiten bestehen, Daten grundrechtsschonender zu erheben,¹²⁸ ist eine Erweiterung der verknüpfbaren Datenquellen angezeigt. Aktuell wurde durch eine Änderung des § 13a Bundesstatistikgesetz, die am 16. Mai 2024 in Kraft getreten ist, bereits eine Erweiterung erreicht, wonach nunmehr Daten oberster Bundesbehörden, die diese zur Erfüllung statistischer Berichtspflichten nach dem Recht der Europäischen Union erhoben haben oder die zu

23 BVerfGE 65, 1 (42).

24 Beispielsweise in § 5 Bundesstatistikgesetz oder § 9 Bundesstatistikgesetz.

25 Beispielsweise durch entsprechende Ergänzung des § 7 Bundesstatistikgesetz.

26 BVerfGE 150, 1 (1).

27 BVerfGE 150, 1 (1).

28 BVerfGE 150, 1 (1).

diesem Zweck in deren Auftrag erhoben wurden, ebenfalls unter anderem mit den Wirtschaftsstatistiken und dem Statistikregister zusammengeführt werden dürfen. Gerade mit dem Argument der Entlastung der Betroffenen und gestiegenen Informationsbedarfen ist auch die Zulässigkeit der Verknüpfung personenbezogener Daten durchaus begründbar (Kühling, 2023, hier: § 13a, Rnr. 25) und sind weitergehende Reformbestrebungen angezeigt.

Statistische Geheimhaltung

Die in § 16 Bundesstatistikgesetz definierte statistische Geheimhaltung ist zweifellos ein „Herzstück“ der gesetzlichen Rahmenbedingungen für bundesstatistische Erhebungen. Es stellt sich aber die Frage, ob der Gesetzgeber seinerzeit die vom Volkszählungsurteil 1983 erlassenen Prämissen unter dem Druck der gesellschaftspolitischen Diskussionen etwas zu dogmatisch umgesetzt hat. Der Begriff der zu schützenden Einzelangaben aus § 16 Bundesstatistikgesetz wird in der rechtlichen Diskussion bislang so verstanden, dass – anders als im allgemeinen Datenschutzrecht – hierunter nicht nur die Einzelangaben natürlicher Personen zu verstehen sind, sondern auch diejenigen juristischer Personen (Dorer und andere, hier: § 16, Rnr. 14). Das Volkszählungsurteil 1983 selbst enthält keine solch weit gefasste Aussage, sondern die Formulierung „personenbezogene Daten“.

Zwar ergibt sich ein verfassungsrechtlicher Schutz von Unternehmens- und Betriebsdaten durch einen drohenden Eingriff in die verfassungsrechtlich geschützte Berufsausübungsfreiheit, allerdings nur, soweit dadurch Betriebs- und Geschäftsgeheimnisse offenbart würden.¹²⁹ Unter solchen Betriebs- und Geschäftsgeheimnissen versteht das Bundesverfassungsgericht alle auf ein Unternehmen bezogene Tatsachen, Umstände und Vorgänge, die nicht offenkundig, sondern nur einem begrenzten Personenbereich zugänglich sind und an deren Nichtverbreitung der Rechtsträger ein berechtigtes Interesse hat.¹³⁰

Hinsichtlich der Informationen zu verstorbenen Personen hat sich die Rechtsauffassung durchgesetzt, dass diese ebenfalls der statistischen Geheimhaltung unterliegen,

gleichwohl entsprechende Informationen überhaupt nicht vom Grundrecht auf informationelle Selbstbestimmung geschützt werden. Das Bundesverfassungsgericht hat in seiner Mephisto-Entscheidung¹³¹ begründet, dass es kein postmortales Persönlichkeitsrecht gibt. Korrespondierend mit der Strafbarkeit der Verunglimpfung des Andenkens Verstorbener nach § 189 Strafgesetzbuch wird lediglich der Anspruch der Angehörigen auf Schutz vor besonders schwerwiegenden Entstellungen des Andenkens der verstorbenen Person oder groben Ehrverletzungen gewährt. Bundesstatistiken, für die Daten Verstorbener erhoben werden,¹³² stellen mit ihrem Veröffentlichungsinhalt einen derartigen einschneidenden Eingriff in den postmortalen Persönlichkeitsschutz regelmäßig nicht dar.

Eine Veränderung des Umfangs der statistischen Geheimhaltung scheint daher auch unter verfassungsrechtlichen Gesichtspunkten grundsätzlich zulässig und sollte weiterverfolgt werden.

Privat gehaltene Daten

Privat gehaltene Daten gewinnen für statistische Erhebungen zunehmend an Bedeutung. Daher hält auch der Statistische Beirat¹³³, das nach § 4 Bundesstatistikgesetz berufene Gremium der Nutzerinnen und Nutzer der Bundesstatistik, die Einräumung eines gesetzlichen Zugangs für bundesstatistische Zwecke für dringend erforderlich (Statistischer Beirat, 2023). Auf europäischer Ebene gibt es mit einem Vorschlag zur Novellierung der Europäischen Statistikverordnung 223/2009 einen Vorstoß, entsprechende Daten für europäische Statistiken zugänglich zu machen. Um einen solchen Zugang möglichst grundrechtsschonend zu gestalten, wäre es denkbar, eine dem § 5a Bundesstatistikgesetz vergleichbare Regelung einzuführen.¹³⁴ Die angesprochene Regelung lässt auf einer ersten Stufe für Eigentumsprüfungen die Anforderung von Metadaten privater Einheiten zu und auf zweiter Stufe die Anforderung formal anonymisierter Einzelangaben zur Verarbeitung

29 BVerfGE 15, 205 ff.

30 BVerfGE 15, 205 ff.

31 Grundsatzentscheidung zur Kunstfreiheit und zum allgemeinen Persönlichkeitsrecht: BVerfG, Beschluss vom 24. Februar 1971, 1 BvR 435/68.

32 Beispielsweise für die Sterbefall- oder die Todesursachenstatistik.

33 Der Statistische Beirat berät das Statistische Bundesamt in Grundsatzfragen und vertritt die Belange der Nutzerinnen und Nutzer, sowie Befragten und Produzenten der Bundesstatistik.

34 Zum Beispiel als § 5b Bundesstatistikgesetz neu.

im Testbetrieb (Kühling/Schmid in Kühling, 2023, hier: § 5, Rnr. 59).

Unter bestimmten gesetzlichen Rahmenbedingungen und Schranken, ähnlich wie sie mit § 11a Bundesstatistikgesetz bereits für öffentliche Stellen, Betriebe und Unternehmen gelten, sollten künftig auch Privatpersonen verpflichtet werden, ihre Daten für statistische Erhebungen elektronisch zu übermitteln. Mit dieser Online-First-Strategie wäre es möglich, die Auskunftspflichtigen weiter zu entlasten und methodische Zielsetzungen zu erreichen.

Kompetenz des Statistischen Bundesamtes zur Datenverwaltung


Der Entwurf der novellierten Europäischen Statistikverordnung unterstreicht in seiner Begründung das hohe technische Fachwissen der statistischen Ämter in den Bereichen Metadatenverwaltung, Datenqualität und Datenschutz. Er ermutigt die Mitgliedstaaten, den nationalen Statistikämtern eine wichtige Rolle in der nationalen Datenverwaltung zuzuweisen. Diesem Anliegen sollte eine entsprechende Ergänzung des Aufgabenkatalogs des § 3 Absatz 1 Bundesstatistikgesetz Rechnung tragen.

Verbesserter Datenzugang für Forschung und Wissenschaft

Am 7. März 2024 hat das Bundesministerium für Bildung und Forschung ein Eckpunktepapier für ein geplantes Forschungsdatengesetz, welches den Zugang zu Daten für die öffentliche und private Forschung verbessern und vereinfachen soll, veröffentlicht (Bundesministerium für Bildung und Forschung, 2024). Darin enthalten ist auch das Vorhaben, einen Online-Zugang (Remote Access) zu formal anonymisierten statistischen Daten zu schaffen. Dieser geht einher mit einem gesetzlichen Forschungsauftrag für das Statistische Bundesamt, der durch entsprechende Anpassungen des Bundesstatistikgesetzes abzusichern sein wird.

6

Ausblick

Das Recht auf informationelle Selbstbestimmung wird auch in Zukunft die Ausgestaltung der Bundesstatistik prägen. Dies bedeutet aber nicht, dass ihre gestalterische Weiterentwicklung und dafür etwa erforderliche Anpassungen des Rechtsrahmens ausgeschlossen sind. Das vom Bundesverfassungsgericht formulierte Grundrecht stellt zwar klare Grundanforderungen an die Ausgestaltung der Bundesstatistik, lässt allerdings Gestaltungsspielräume, von denen der Gesetzgeber zum Teil noch keinen Gebrauch gemacht hat. Damit die Bundesstatistik vor dem Hintergrund des gesellschaftlichen und digitalen Wandels ihren gesetzlichen Auftrag der Informationsbereitstellung sachgerecht erfüllen kann, ist eine Anpassung des Bundesstatistikgesetzes entlang dieser Gestaltungsspielräume geboten. 

LITERATURVERZEICHNIS

Bierschenk, Michaela/Leischner, Sonja. [Zur Verfassungsmäßigkeit der Vorschriften über den Zensus 2011](#). In: WISTA Wirtschaft und Statistik. Ausgabe 1/2019, Seite 11 ff.

Bull, Hans Peter. *Informationelle Selbstbestimmung – Vision oder Illusion?* 2. Auflage. Tübingen 2011.

Bundesministerium der Finanzen. *Netzwerk empirische Steuerforschung*. [Zugriff am 18. April 2024]. Verfügbar unter: www.bundesfinanzministerium.de

Bundesministerium für Bildung und Forschung. *Eckpunkte BMBF Forschungsdatengesetz*. 2024. [Zugriff am 19. April 2024]. Verfügbar unter: www.bmbf.de

Dorer, Peter/Mainusch, Helmut/Tubies, Helga. *Bundesstatistikgesetz. Kommentar*. München 1988.

Engelter, Marion/Sommer, Kay. [Die Novellierung des Bundesstatistikgesetzes 2016](#). In: WISTA Wirtschaft und Statistik. Ausgabe 6/2016, Seite 11 ff.

Eurostat (Statistisches Amt der Europäischen Union). *Verhaltenskodex für europäische Statistiken*. 2017. [Zugriff am 18. April 2024]. Verfügbar unter: www.destatis.de

Franzius, Claudio. *Das Recht auf informationelle Selbstbestimmung*. In: Zeitschrift für das Juristische Studium (ZfS). Ausgabe 3/2015, Seite 259 ff.

Göbl, Thomas. *Der Zensus vor dem Bundesverfassungsgericht*. In: Statistisches Monatsheft Baden-Württemberg. Ausgabe 1/2019, Seite 41 ff.

Hoffmann-Riem, Wolfgang. *Informationelle Selbstbestimmung in der Informationsgesellschaft – Auf einem Weg zu einem neuen Konzept des Datenschutzes*. In: Archiv des öffentlichen Rechts (AöR). Ausgabe 123/1998, Seite 513 ff.

Hölder, Egon. *Durchblick ohne Einblick. Die amtliche Statistik zwischen Datennot und Datenschutz*. 2. Auflage. Zürich 1985.

Isfort, Claudia/Dommermuth, Silke. [Der neue Kommentar zum Bundesstatistikgesetz: zur Weiterentwicklung des Statistikrechts seit 1988](#). In: WISTA Wirtschaft und Statistik. Ausgabe 2/2023, Seite 19 ff.

Kienle, Thomas. *Anmerkung zu BVerfG, 19.09.2018 – 2 BvF 1/15: BVerfG: Vorschriften über den Zensus 2011 verfassungsgemäß*. In: Zeitschrift für Datenschutz (ZD). Ausgabe 12/2018, Seite 578 ff.

Klumpen, Dorothea/Köhler, Sabine. [Aktuelle Anforderungen an die amtliche Statistik in Europa](#). In: Wirtschaft und Statistik. Ausgabe 11/2003, Seite 981 ff.

Kommission Zukunft Statistik. *Bericht*. 2024. [Zugriff am 6. Mai 2024]. Verfügbar unter: www.destatis.de

Kühling, Jürgen. *Bundesstatistikgesetz: BStatG. Kommentar*. München 2023.

LITERATURVERZEICHNIS

Kühling, Jürgen. *Neues Bundesdatenschutzgesetz – Anpassungsbedarf bei Unternehmen*. In: Neue Juristische Wochenschrift (NJW). 2017. Seite 1985 ff.

Leischner, Sonja/Weigelt, Sabine. *Vorschriften über Volkszählung 2011 verfassungsgemäß, Anmerkung zu BVerfG, Urteil vom 15.12.1983 – 1 BvR 209/83 u.a.* In: NVwZ Neue Zeitschrift für Verwaltungsrecht. Jahrgang 37. Heft 22/2018, Seite 1703, 1731 ff.

O'Donnell, Daniel. *Nutzerleitfaden zur EU-Statistik*. In: Wirtschaft und Statistik. Ausgabe 5/2006, Seite 443 ff.

Poppenhäger, Holger. *Die Übermittlung und Veröffentlichung statistischer Daten im Lichte des Rechts auf informationelle Selbstbestimmung*. Schriften zum Recht des Informationsverkehrs und der Informationstechnik (RIVT). Band 12. Berlin 1995.

Riede, Thomas/Tümmler, Thorsten/Wondrak, Stefan. *Die digitale Agenda des Statistischen Bundesamtes*. In: WISTA Wirtschaft und Statistik. Ausgabe 1/2018, Seite 102 ff.

Schliffka, Christina/Polus, Dominique. *Das Europäische Statistische System als Datenmanager – verlässliche Daten für Europa*. In: WISTA Wirtschaft und Statistik. Ausgabe 4/2020, Seite 30 ff.

Söllner, René/Körner, Thomas. *Der Registerzensus: Ziele, Anforderungen und Umsetzungsansätze*. In: WISTA Wirtschaft und Statistik. Ausgabe 4/2022, Seite 13 ff.

Statistischer Beirat. *Bericht über die 70. Tagung*. 2023. [Zugriff am 19. April 2024]. Verfügbar unter: www.statistischebibliothek.de

Statistisches Bundesamt. *Datennotstand und Datenschutz. Die amtliche Statistik nach dem Volkszählungsurteil. Ergebnisse des 1. Wiesbadener Gesprächs am 30./31. Oktober 1984*. Band 3 der Schriftenreihe Forum der Bundesstatistik. Wiesbaden 1985. Verfügbar unter: www.statistischebibliothek.de

Statistisches Bundesamt. *Zum Gesetz über die Statistik für Bundeszwecke*. Band 9 der Schriftenreihe Forum der Bundesstatistik. Wiesbaden 1988. Verfügbar unter: www.statistischebibliothek.de

Thiel, Georg/Puth, Marie-Christin. *Der Zensus der Zukunft: Registerzensus*. In: Neue Zeitschrift für Verwaltung (NVwZ). 2023, Seite 305 ff.

Widmann, Arno. *Volkszählung 1983: Als der Staat seine Schäflein zählen wollte*. In: Frankfurter Rundschau. 12. April 2023. [Zugriff am 17. April 2024]. Verfügbar unter: www.fr.de

Wieduwilt, Hendrik. *Ein Grundrecht aus Versehen: 40 Jahre Volkszählungsurteil*. 2024. [Zugriff am 17. April 2024]. Verfügbar unter: anwaltsblatt.anwaltverein.de

Wiengarten, Lara/Zwick, Markus. *Neue digitale Daten in der amtlichen Statistik*. In: WISTA Wirtschaft und Statistik. Ausgabe 5/2017, Seite 19 ff.

RECHTSGRUNDLAGEN

Bundesdatenschutzgesetz (BDSG) vom 30. Juni 2017 (BGBl. I Seite 2097), das zuletzt durch Artikel 10 des Gesetzes vom 22. Dezember 2023 (BGBl. I Nr. 414) geändert worden ist.

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.

Gesetz über eine Volks-, Berufs-, Wohnungs- und Arbeitsstättenzählung (Volkszählungsgesetz 1983) in der Fassung der Bekanntmachung vom 25. März 1982 (BGBl. I Seite 369), außer Kraft getreten am 15. November 1985.

Gesetz über eine Volks-, Berufs-, Gebäude-, Wohnungs- und Arbeitsstättenzählung (Volkszählungsgesetz 1987 – VoZählG 1987) vom 8. November 1985 (BGBl. I Seite 2078), aufgehoben durch Artikel 5 des Gesetzes vom 9. Juni 2021 (BGBl. I Seite 1649).

Gesetz zur Erprobung von Verfahren eines Registerzensus (Registerzensuserprobungsgesetz – RegZensErpG) vom 9. Juni 2021 (BGBl. I Seite 1649).

Grundgesetz für die Bundesrepublik Deutschland in der im Bundesgesetzblatt Teil III Gliederungsnummer 100-1 veröffentlichten bereinigten Form, das zuletzt durch Artikel 1 des Gesetzes vom 19. Dezember 2022 (BGBl. I Seite 2478) geändert worden ist.

Strafgesetzbuch in der Fassung der Bekanntmachung vom 13. November 1998 (BGBl. I Seite 3322), das zuletzt durch Artikel 12 des Gesetzes vom 27. März 2024 (BGBl. I Nr. 109) geändert worden ist.

Verordnung (EG) Nr. 223/2009 des Europäischen Parlaments und des Rates vom 11. März 2009 über europäische Statistiken und zur Aufhebung der Verordnung (EG, Euratom) Nr. 1101/2008 des Europäischen Parlaments und des Rates über die Übermittlung von unter die Geheimhaltungspflicht fallenden Informationen an das Statistische Amt der Europäischen Gemeinschaften, der Verordnung (EG) Nr. 322/97 des Rates über die Gemeinschaftsstatistiken und des Beschlusses 89/382/EWG, Euratom des Rates zur Einsetzung eines Ausschusses für das Statistische Programm der Europäischen Gemeinschaften (Amtsblatt der EU Nr. L 87, Seite 164).

Verordnung (EU) 2016/679 des Europäischen Parlaments und des Rates vom 27. April 2016 zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten, zum freien Datenverkehr und zur Aufhebung der Richtlinie 95/46/EG (Datenschutz-Grundverordnung) (Amtsblatt der EU Nr. L 119, Seite 1).

Verordnung (EU) 2019/2152 des Europäischen Parlaments und des Rates vom 27. November 2019 über europäische Unternehmensstatistiken, zur Aufhebung von zehn Rechtsakten im Bereich Unternehmensstatistiken (Amtsblatt der EU Nr. L 327, Seite 1).

Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates zur Änderung der Verordnung (EG) Nr. 223/2009 über europäische Statistiken, COM(2023) 402 final 2023/0237(COD).

DIE CELL-KEY-METHODE IN DEN FORSCHUNGSDATENZENTREN DER STATISTISCHEN ÄMTER DES BUNDES UND DER LÄNDER

Teil 1: Vorstellung des neuen Geheimhaltungsverfahrens

Stefanie Setzer, Johannes Rohde, Volker Güttgemanns, Patrick Rothe

➤ **Schlüsselwörter:** Datenschutz – Geheimhaltung – stochastische Überlagerung – Aufdeckungsschutz – post-tabular

ZUSAMMENFASSUNG

Die Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder führen für ausgewählte Statistiken die Cell-Key-Methode als neues Verfahren zur Ergebnisgeheimhaltung ein. Dieses Verfahren schützt die Befragten vor der Reidentifikation, indem es durch die Überlagerung der Fallzahlen eine Unsicherheit über die Anzahl der tatsächlich zum Ergebnis beitragenden Fälle schafft. Der Artikel stellt die Funktionsweise der Cell-Key-Methode vor und bietet dabei sowohl eine einfach zu verstehende Einführung in die Thematik als auch detaillierte methodische Informationen.

➤ **Keywords:** data protection – confidentiality – stochastic perturbation – disclosure protection – post-tabular

ABSTRACT

The Research Data Centres of the statistical offices of the Federation and the Länder are introducing the cell key method for selected statistics as a new method for keeping results confidential. This method protects respondents against re-identification by applying a level of perturbation to cells in a table to create uncertainty about the number of cases actually contributing to a result. This article explains how the cell key method works and provides both an easy-to-understand introduction to the topic and detailed methodological information.

Stefanie Setzer

ist Diplom-Soziologin und Referentin im Referat „Forschungsdatenzentrum, Methoden der Datenanalyse“ des Statistischen Bundesamtes. Schwerpunkt ihrer Arbeit ist die fachliche und methodische Weiterentwicklung des Arbeitsbereichs.

Dr. Johannes Rohde

hat Wirtschaftswissenschaften an der Leibniz Universität Hannover studiert und dort 2015 seine Promotion im Bereich Statistik abgeschlossen. Bei IT.NRW leitet er den Service „Mathematisch-statistische Methoden und experimentelle Statistik“.

Volker Güttgemanns

hat einen Master of Science in Wirtschaftswissenschaften und war von 2017 bis 2023 stellvertretende Leitung der Geschäftsstelle des Forschungsdatenzentrums der Statistischen Ämter der Länder.

Patrick Rothe

hat Sozialwissenschaften an der Universität Mannheim studiert und ist seit 2011 im Bayerischen Landesamt für Statistik tätig. Seit 2018 leitet er dort das Sachgebiet „Grundsatzfragen der amtlichen Statistik, Digitalisierung, Forschungsdatenzentrum, Kompetenzzentrum Analyse“. Inhaltlich beschäftigt er sich schwerpunktmäßig unter anderem mit der statistischen Geheimhaltung.

1

Einleitung

Die Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder (FDZ) stellen Mikrodaten für die wissenschaftliche Nutzung bereit. Um den Datenschutz hierbei zu gewährleisten, gibt es zwei Möglichkeiten: entweder die Anonymisierung, bei der die Daten vor der Bereitstellung an die Wissenschaft so verändert werden, dass bei der Auswertung keine Geheimhaltungsrisiken entstehen können, oder die Geheimhaltung, bei welcher der Schutz der Daten durch eine Veränderung der Ergebnisse erzeugt wird. Die Wahl der Vorgehensweise hängt dabei vom gewählten Zugangsweg ab: Bei den sogenannten Off-Site-Zugangswegen, bei denen die Daten an die Nutzenden übermittelt werden, erfolgt eine Anonymisierung der Daten, was jedoch immer mit einem Informationsverlust einhergeht. Bei Nutzung der On-Site-Zugangswegen, also eines Gastwissenschaftsarbeitsplatzes oder der kontrollierten Datenfernverarbeitung, verbleiben die Daten in den geschützten Räumen der amtlichen Statistik. Dort kann in der Regel das volle Informationspotenzial der Daten erhalten bleiben, die erzeugten Ergebnisse werden dafür aber einer Geheimhaltungsprüfung unterzogen.

Die Auswirkungen dieser Geheimhaltungsprüfung kennt jede Person, die schon einmal einen der On-Site-Zugangswegen der Forschungsdatenzentren genutzt hat: Nach Bereitstellung der Ergebnisse springen häufig drei große X ins Auge. Dieses Sperrmuster verwenden die Forschungsdatenzentren üblicherweise, wenn die Veröffentlichung eines Ergebnisses ein Geheimhaltungsrisiko darstellt. Doch warum nehmen die Forschungsdatenzentren die Geheimhaltung überhaupt so ernst? Gäbe es Alternativen zu den drei großen X?

Kapitel 2 beantwortet zunächst die erste Frage, warum die Geheimhaltung den Forschungsdatenzentren so wichtig ist. Wie die Cell-Key-Methode als alternatives Verfahren zur Sperrung mit den drei großen X funktioniert, erläutert Kapitel 3. Danach erläutert Kapitel 4 die Methodik der Cell-Key-Methode formell. Ein kurzes Fazit mit dem Hinweis auf den zweiten Aufsatzteil beschließt den Beitrag.

2

Geheimhaltung in den Forschungsdatenzentren

2.1 Warum nehmen die Forschungsdatenzentren Geheimhaltung so ernst?

Diese Frage lässt sich einfach beantworten: weil sie gesetzlich dazu verpflichtet sind. Die Pflicht zur Geheimhaltung ist in §16 Bundesstatistikgesetz (BStatG) geregelt. Danach sind „Einzelangaben über persönliche und sachliche Verhältnisse, die für eine Bundesstatistik gemacht werden, [...] geheim zu halten“ (§16 Absatz 1 BStatG). Dieser Absatz regelt aber auch Ausnahmen, die es den statistischen Ämtern und der Wissenschaft ermöglichen, Daten und Ergebnisse unter bestimmten Voraussetzungen zu veröffentlichen. Eine Veröffentlichung ist beispielsweise möglich, wenn die Einzelangaben mit den Ergebnissen anderer Befragter zusammengefasst wurden oder wenn die Einzelangaben den Betroffenen nicht zuzuordnen sind. Diese beiden Ausnahmen ermöglichen die Bereitstellung von Daten und Ergebnissen und begründen gleichzeitig die Pflicht zur Geheimhaltung. Ergebnisveröffentlichungen sind erlaubt, solange aus den Ergebnissen keine Rückschlüsse auf Einzelne gezogen werden können.

Der Grund für diese gesetzliche Regelung und den dadurch festgelegten hohen Stellenwert der Geheimhaltung ist gut nachvollziehbar: Für die Erhebungen der amtlichen Statistik besteht oft Auskunftspflicht. Die Befragten – seien es Personen, Unternehmen, Betriebe oder Sonstige – können demnach häufig nicht selbst entscheiden, welche Informationen sie von sich preisgeben wollen. Um diesen Eingriff in die informationelle Selbstbestimmung auszugleichen, garantiert der Gesetzgeber den Befragten, dass ihnen ihre Angaben nicht zugeordnet werden können. Gleiches gilt für Erhebungen mit freiwilliger Teilnahme.

Das Vertrauen der Befragten in die Nicht-Zuordenbarkeit ihrer Angaben ist die Grundlage dafür, dass Fragen ohne Sorge vor Enthüllung wahrheitsgemäß beantwortet werden, und trägt somit maßgeblich zur hohen Qualität der Daten bei. Der Schutz der anvertrauten Daten hat daher für die amtliche Statistik – und damit auch für die Forschungsdatenzentren – stets die oberste Priorität.

2.2 Der bisherige Standard: die Zellspernung

Bisher stellen die statistischen Ämter die Geheimhaltung in der Regel mithilfe der Zellspernung sicher¹. Bei dieser Form der Geheimhaltung werden alle Angaben, die ein Geheimhaltungsrisiko darstellen, durch ein Sperrmuster („XXX“) ersetzt. Dieses Verfahren hat sich in der amtlichen Statistik bewährt, weist jedoch einige gravierende Nachteile auf:

- › **Informationsverlust:** Bei der Zellspernung werden nicht nur die kritischen Angaben selbst gesperrt (Primärspernung). Um eine Rückrechnung dieser Werte zu verhindern, müssen sie mit an sich unkritischen Angaben gegengespart werden (Sekundärspernung). Wenn gesperrte Angaben über andere Tabellen rückrechenbar sind, müssen auch hier Sperrungen umgesetzt werden (tabellenübergreifende Sperrung). So kann ein einzelner zu sperrender Wert schnell eine Vielzahl weiterer Sperrungen an sich unkritischer Werte nach sich ziehen.
- › **Hoher Aufwand:** Da die Geheimhaltungsprüfung bei der Zellspernung in der Regel nicht vollständig automatisiert erfolgen kann, ist die Geheimhaltung sehr zeit- und ressourcenintensiv und Nutzende müssen teils lange auf ihre Ergebnisse warten.
- › **Unzufriedenheit:** Durch dieses Verfahren können nicht alle interessierenden Werte veröffentlicht werden, teilweise müssen sogar ganze Tabellen gesperrt werden. Daher führt die Zellspernung häufig zu unzufriedenen Datennutzenden.

2.3 Das neue Geheimhaltungsverfahren: die Cell-Key-Methode

Aufgrund der beschriebenen Nachteile der Zellspernung haben die statistischen Ämter für erste Statistiken die Einführung der Cell-Key-Methode (CKM) beschlossen.¹² Diese Entscheidung wirkt sich unmittelbar auf die Datenbereitstellung in den Forschungsdatenzentren aus, da die Sicherstellung der Geheimhaltung stets in Einklang mit den fachseitig festgelegten Geheimhaltungsregeln erfolgt. Bei der Cell-Key-Methode handelt es sich um ein datenveränderndes Verfahren für die Geheimhaltung von Fallzahltabellen. Die Vorteile des Verfahrens sind:

- › Mit der Cell-Key-Methode gibt es keine Sperrungen.
- › Die Ergebnisse weisen eine hohe Datenqualität auf und sind tabellenübergreifend konsistent.
- › Der Aufwand für die Geheimhaltungsprüfung von Tabellen ist deutlich geringer.
- › Aufdeckungsrisiken durch Fehler bei der tabellenübergreifenden Geheimhaltung können ausgeschlossen werden.

Diesen Vorteilen stehen aber auch Nachteile gegenüber:

- › Tabellen sind nach der Anwendung der Cell-Key-Methode nicht mehr additiv.
- › Gerade bei kleinen Fallzahlen kann die Veränderung der Werte relativ stark ausfallen.
- › Da es sich bei der Cell-Key-Methode originär um ein Verfahren für die Geheimhaltung von Fallzahltabellen handelt, können zusätzliche Aufwände bei der Prüfung der Ergebnisse multivariater Analysemethoden entstehen.
- › Die Cell-Key-Methode ist weniger eingängig als andere Geheimhaltungsverfahren und bedarf daher umfangreicher Erläuterung.

Das Verfahren der Cell-Key-Methode wird im Folgenden für die Grundform der Geheimhaltung von Fallzahltabellen vorgestellt.

1 Beim Zensus 2011 wurde außerdem das Verfahren SAFE – Sichere Anonymisierung Für Einzelangaben genutzt (Höhne, 2015).

2 Die Cell-Key-Methode wurde ursprünglich vom australischen Statistikamt (Australian Bureau of Statistics) entwickelt.

3

Funktionsweise der Cell-Key-Methode

Bei der Cell-Key-Methode handelt es sich um ein post-tabulares datenveränderndes Geheimhaltungsverfahren. Das bedeutet, dass das Verfahren erst bei der Ergebniserstellung ansetzt und dass die Geheimhaltung durch eine Veränderung von Fallzahlen erfolgt. Die Schutzwirkung wird dadurch erzielt, dass Unsicherheit bezüglich der Originalfallzahl geschaffen wird, indem Tabellenwerte mit einem Fehlerterm überlagert werden. Das Verfahren stellt dabei sicher, dass die Überlagerung konsistent ist, dass also logisch identische Fallzahlen über alle Tabellen hinweg identisch bleiben. So nehmen Randsummen einer Kreuztabelle, beispielsweise „Bundesland x Alter“, immer die gleichen Werte an, die auch in den entsprechenden Fallzahltabellen der beiden Merkmale ausgegeben werden. Das gilt unabhängig davon, ob der Wert als Innen- oder als Randfeld einer Tabelle auftritt.

Die folgenden Abschnitte erläutern die Funktionsweise der Cell-Key-Methode Schritt für Schritt. Hierbei ist zu beachten, dass ein großer Vorteil der Cell-Key-Methode gerade darin besteht, dass die eigentliche Geheimhaltung von Fallzahltabellen weitgehend automatisiert erfolgt.

3.1 Record Keys


Für die Anwendung der Cell-Key-Methode wird an den Ausgangsdatensatz zunächst ein zusätzliches Merkmal angespielt, das den sogenannten Record Key enthält. Dieser besteht aus einer Zufallszahl zwischen 0 und 1, die jeder Beobachtungseinheit fest zugeordnet wird.

Zur Veranschaulichung zeigt [Übersicht 1](#) Daten für eine fiktive Gemeinde, in der das Alter und das Einkommen aller erwachsenen Bewohner klassiert erfasst ist. Aus Gründen der Übersichtlichkeit wird ein Record Key mit zwei Nachkommastellen vergeben, in echten Anwendungsfällen ist die Anzahl der Nachkommastellen in der Regel höher.


Übersicht 1

Anfügen der Record Keys an den Ausgangsdatensatz

ID	Alter	Einkommen	Record Key
1	jung	mittel	0,54
2	jung	hoch	0,68
3	alt	niedrig	0,14
4	alt	mittel	0,93
5	jung	mittel	0,51
6	alt	mittel	0,37
7	alt	niedrig	0,84
8	alt	hoch	0,19
9	alt	mittel	0,26
10	jung	hoch	0,43
11	alt	mittel	0,99
12	jung	mittel	0,74
13	jung	mittel	0,65
14	alt	niedrig	0,79
15	alt	mittel	0,25



Ausgangsdatensatz



Angespielter Record Key

3.2 Übergangsmatrix

In der Übergangsmatrix wird festgelegt, mit welchem Wert eine Fallzahl überlagert wird (Kleber/Gießing, 2018). Dafür wird für jede Originalfallzahl beschlossen, mit welcher Wahrscheinlichkeit diese zu einem bestimmten anderen Wert verändert wird. In [Tabelle 1](#) bliebe die Originalfallzahl 10 zum Beispiel mit einer Wahrscheinlichkeit von 0,7 eine 10, würde mit einer Wahrscheinlichkeit von je 0,1 zu einer 9 oder 11 und mit einer Wahrscheinlichkeit von je 0,05 zu einer 8 oder 12. So lassen sich verschiedene Rahmenbedingungen festlegen, beispielsweise die maximale Abweichung, die Wahrscheinlichkeit für den Erhalt einer Originalfallzahl, der Ausschluss von 1 und 2 in der überlagerten Tabelle oder eine höhere Bleibewahrscheinlichkeit für höhere Fallzahlen.

Dabei ist zu beachten, dass es sich hierbei lediglich um ein fiktives Beispiel einer Übergangsmatrix handelt, um deren mögliche Ausgestaltung vereinfacht darzustellen. Denkbar wäre beispielsweise auch die Festlegung, dass alle ausgewiesenen Nullen echte Nullen sind, dass kein Wert unverändert bleibt oder dass starke Überlagerungen wahrscheinlicher sind als geringe. Da die Übergangsmatrix somit steuert, wie (un-)ähnlich sich die

Die Cell-Key-Methode in den Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder – Teil 1: Vorstellung des neuen Geheimhaltungsverfahrens

Tabelle 1

Fiktives Beispiel einer Übergangsmatrix

Original- häufigkeit	Zielhäufigkeit																	
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0,7	0	0	0,3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0,3	0	0	0,7	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0,5	0,35	0,15	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0,25	0,5	0,2	0,05	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0,05	0,2	0,5	0,2	0,05	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0,05	0,2	0,5	0,2	0,05	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0,05	0,2	0,5	0,2	0,05	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0,05	0,2	0,5	0,2	0,05	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0,05	0,2	0,5	0,2	0,05	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0,05	0,1	0,7	0,1	0,05	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0,05	0,1	0,7	0,1	0,05	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0,05	0,1	0,7	0,1	0,05	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0,05	0,1	0,7	0,1	0,05	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0,05	0,1	0,7	0,1	0,05	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0,05	0,1	0,7	0,1	0,05

■ Nullen bleiben unverändert
■ Keine 1 oder 2 in der überlagerten Tabelle
■ Maximalabweichung +/-2
■ Höhere Bleibewahrscheinlichkeit bei höheren Fallzahlen

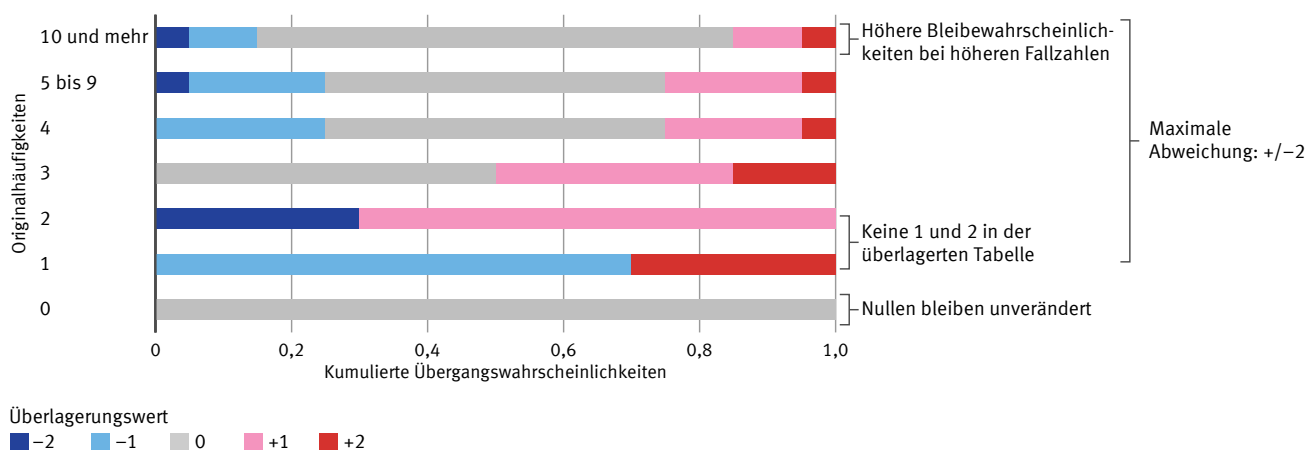
originale und die überlagerte Tabelle sind, stellt sie das Kernstück der Cell-Key-Methode dar, das mit viel Aufwand ausgestaltet wird. Diese tatsächlich verwendeten Übergangsmatrizen unterliegen ebenso wie die festgelegten Rahmenbedingungen der strengen Geheimhaltung und werden an die jeweiligen Bedarfe der Statistiken spezifisch angepasst.

3.3 Überlagerungstableau

Auf Basis der Übergangsmatrix wird ein Überlagerungstableau erstellt (Kleber/Gießing, 2018). Zur Veranschaulichung stellt [Grafik 1](#) das Überlagerungstableau auf Basis der Übergangsmatrix aus Tabelle 1 als Raster dar. Hierfür werden die Wahrscheinlichkeiten der einzelnen Zeilen der Übergangsmatrix kumuliert. Daraus lässt sich

Grafik 1

Fiktives Beispiel eines Überlagerungstableaus



im sogenannten Lookup-Schritt ablesen, welcher Überlagerungswert für die jeweilige Originalfallzahl aus der zugehörigen kumulierten Übergangswahrscheinlichkeit resultiert.

Zum besseren Verständnis stellt [Übersicht 2](#) die Veränderungen der Originalfallzahl „4“ exemplarisch dar.

Übersicht 2

Kumulierte Übergangswahrscheinlichkeiten für die Originalfallzahl „4“

Veränderung zu	Entspricht Überlagerung mit	Wahrscheinlichkeit	Kumulierte Wahrscheinlichkeit
3	-1	0,25	0,25
4	0	0,5	0,75
5	1	0,2	0,95
6	2	0,05	1

3.4 Tabellenerstellung

Bei der Erstellung der überlagerten Tabellen kommen schließlich die Originalfallzahl, die kumulierte Übergangswahrscheinlichkeit für diese Originalfallzahl und die eingangs erzeugten Record Keys zusammen, um den Überlagerungswert zu bestimmen:

Übersicht 3

Erstellen der Cell Keys für zwei Beispiele durch die Addition der zugehörigen Record Keys (RK) und das Auf-null-Setzen der Zahl vor dem Komma

ID	Alter	Einkommen	Record Key
1	jung	mittel	0,54
2	jung	hoch	0,68
3	alt	niedrig	0,14
4	alt	mittel	0,93
5	jung	mittel	0,51
6	alt	mittel	0,37
7	alt	niedrig	0,84
8	alt	hoch	0,19
9	alt	mittel	0,26
10	jung	hoch	0,43
11	alt	mittel	0,99
12	jung	mittel	0,74
13	jung	mittel	0,65
14	alt	niedrig	0,79
15	alt	mittel	0,25

Im Zuge der Erstellung von Fallzahltabellen mit der Cell-Key-Methode werden nicht nur die Fallzahlen ermittelt. Für jede Tabellenzeile werden darüber hinaus die Record Keys aller Beobachtungseinheiten addiert, die zur entsprechenden Tabellenzeile beitragen. Relevant sind von der Summe der Record Keys allerdings nur die Nachkommastellen, der Wert vor dem Komma wird daher auf 0 gesetzt. So entsteht ein Wert zwischen 0 und kleiner 1, der sogenannte Cell Key.

[Übersicht 3](#) veranschaulicht diesen Mechanismus für die Merkmalskombination „junge Befragte mit mittlerem Einkommen“ sowie für die Summe der Personen mit niedrigem Einkommen.

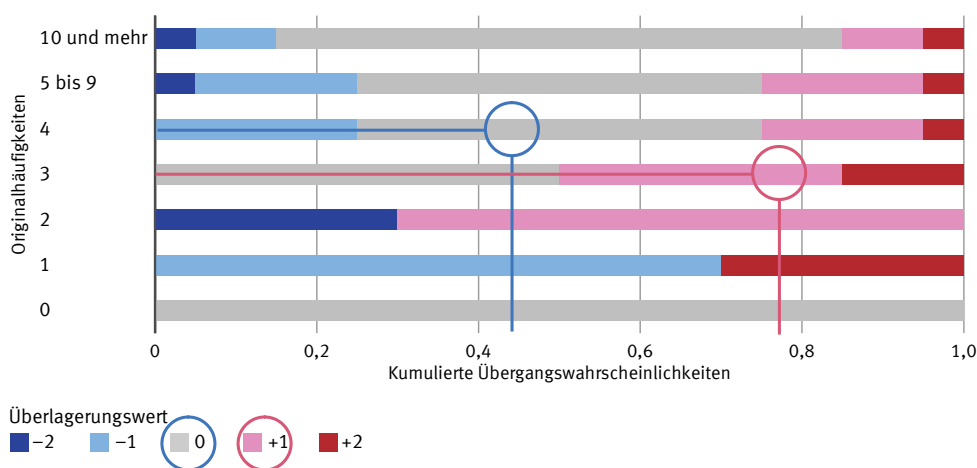
Die für alle Originalfallzahlen berechneten Cell Keys werden jetzt an das Übergangstableau zurückgespielt. Hier erfolgt der Abgleich, in welchem Raster die kumulierte Wahrscheinlichkeit dem ermittelten Cell Key entspricht. Aus dieser Spalte wird dann der Überlagerungswert für die jeweilige Originalfallzahl abgelesen. Die farbige Markierung des betreffenden Rasters verdeutlicht, mit welchem Wert die Originalfallzahl überlagert wird.

Im Beispiel ergibt sich für die vier jungen Personen mit mittlerem Einkommen aus dem ermittelten Cell Key von 0,44 ein Überlagerungswert von 0. Diese Fallzahl bleibt also unverändert. Für die drei Personen mit niedrigem

		Alter		
		jung	alt	Summe
Einkommen	niedrig	0	3	3 $\Sigma RK = 0,14 + 0,84 + 0,79 = 1,77$ → Cell Key = 0,77
	mittel	4 $\Sigma RK = 0,54 + 0,51 + 0,74 + 0,65 = 2,44$ → Cell Key = 0,44	5	9
	hoch	2	1	3
	Summe	6	9	15

Grafik 2

Ablesen ("Lookup") der Überlagerungswerte aus dem Überlagerungstableau für die 4 jungen Personen mit mittlerem Einkommen und einem Cell Key von 0,44 (blau) und den 3 Personen mit niedrigem Einkommen und einem Cell Key von 0,77 (rot)



Einkommen ergibt sich ein Cell Key von 0,77, was laut Überlagerungstableau einer Überlagerung von +1 entspricht. Diese Fallzahl wird also von 3 auf 4 verändert.

➔ Grafik 2

Wird das Verfahren auf alle Felder der Tabelle angewandt, verändert sich die originale Fallzahltablette anhand des in [Übersicht 4](#) dargestellten Mechanismus.

Der überlagerten Tabelle sieht man zunächst nicht an, dass sie nicht die Originalfallzahlen enthält. Auf den zweiten Blick wird aber schnell deutlich, dass sich die Innenfelder in der Regel nicht zu den Randsummen summieren. Diese und andere Auswirkungen der Cell-Key-Methode stellt ein weiterer Beitrag vor, der ebenfalls in derselben Ausgabe dieser Zeitschrift erschienen ist (Rothe und andere, 2024).

4

Formelle Erläuterung der Cell-Key-Methode

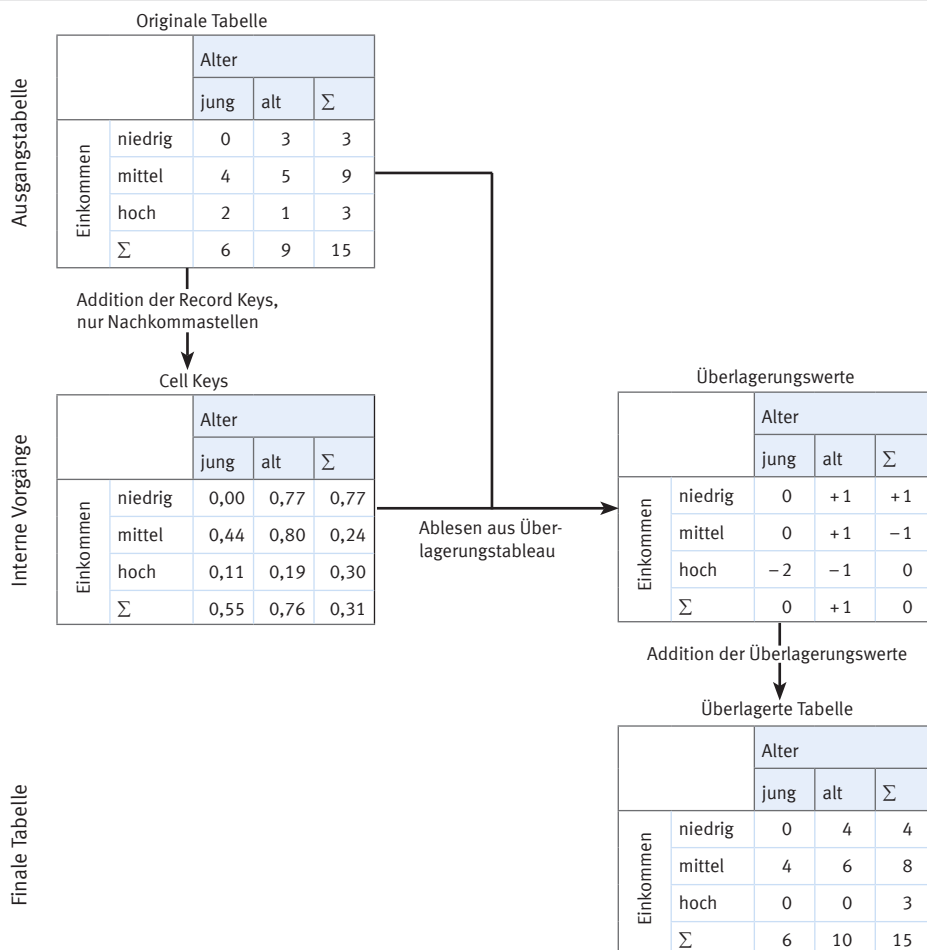
Die Cell-Key-Methode eignet sich in ihrer Grundform für die Geheimhaltung von Fallzahltablets. Mittlerweile steht auch eine Erweiterung der Cell-Key-Methode auf Wertetabellen zur Verfügung (Gießing/Tent, 2019), die in der deutschen amtlichen Statistik bislang allerdings lediglich für einige Wertmerkmale des Zensus 2022 angewendet wird. Im Folgenden beschränkt sich die formelle Erläuterung der Methodik der Cell-Key-Methode daher auf deren Grundform für Fallzahltablets in Anlehnung an die Originalveröffentlichung von Fraser/Wooton (2005).

Die Grundidee der Cell-Key-Methode besteht darin, bei den Nutzenden Unsicherheit dahingehend zu erzeugen, ob eine veröffentlichte Fallzahl der Originalfallzahl entspricht oder diese leicht verändert wurde. Dazu erfolgt eine Perturbation aller Originalfallzahlen mittels eines Überlagerungswertes. Bezeichnet man mit $d_i \in \mathbb{Z}$ den dem i -ten Tabellenfeld mit Originalfallzahl $j_i \in \mathbb{N}_0$ zugeordneten Überlagerungswert, so ergibt sich die veränderte Fallzahl k_i gemäß

$$k_i = j_i + d_i.$$

Übersicht 4

Darstellung der Arbeitsschritte der Cell-Key-Methode



Wie bei allen datenverändernden Verfahren gehört zu den wesentlichen Eigenschaften der Cell-Key-Methode, dass alle Originalfallzahlen unabhängig ihrer Kritikalität beziehungsweise ihres Aufdeckungsrisikos verändert werden können.

4.1 Anforderungen an die Überlagerungswerte

Für die Überlagerungswerte d_i gelten für alle i folgende Eigenschaften:

- a) **Unverzerrtheit:** $E[d_i] = 0$, das heißt aus der Überlagerung der Tabellenfelder ergibt sich keine systematische Verzerrung des Gesamtergebnisses.

- b) **Nicht-Negativität:** $d_i \geq -j_i$, das heißt für jedes Tabellenfeld wird sichergestellt, dass sich durch die Überlagerung keine negative Fallzahl ergibt ($\forall i: k_i \geq 0$).

- c) **Ganzzahligkeit:** $d_i \in \mathbb{Z}$, das heißt die Überlagerungswerte müssen ganzzahlig sein, damit auch für die veränderten Fallzahlen Ganzzahligkeit gewährleistet ist. Aus den Bedingungen (b) und (c) ergibt sich somit $k_i \in \mathbb{N}_0$.

4.2 Parameter der Cell-Key-Methode

Die Anwendung der Cell-Key-Methode setzt die Festlegung von Parametern und Bedingungen voraus, durch welche das für die Zuweisung der Überlagerungswerte maßgebende stochastische Modell determiniert wird. Dabei wird zwischen zwingend festzulegenden Parametern

tern sowie zusätzlichen optional zu setzenden Restriktionen unterschieden. Die konkrete Ausgestaltung der Parametrisierung erfolgt dabei stets durch die für die betreffende Statistik zuständige Fachseite.

Konkret ist die Festlegung zweier Parameter obligatorisch:

- › **Maximale Abweichung** $D \in \mathbb{N}$ zwischen originalen und veränderten Fallzahlen, sodass $\forall i: |d_i| \leq D$.
- › **Varianz** V der Abweichungen von den Originalfallzahlen: Die Varianz entspricht

$$V := \text{Var}[d_i] = E[d_i^2] = \sum_{i=-D}^D (p_i \cdot d_i^2).$$

Durch die Festlegung von D ergibt sich als Wertebereich für V implizit $(0; D^2]$.

Die Festlegung der maximalen Abweichung zwischen originalen und veränderten Fallzahlen dient insbesondere der Steuerung des durch die Datenveränderung induzierten Informationsverlustes und damit der Qualität der geheim gehaltenen Ergebnisse. Mit der Varianz der Abweichungen von den Originalfallzahlen wird die Unsicherheit kalibriert, die bei den Nutzenden durch die Datenveränderung erzeugt werden soll. Eine geringe Varianz impliziert dabei, dass ein hoher Anteil der Originalfallzahlen nur geringfügig oder gar nicht (also $d_i = 0$) verändert wird. Je größer V gewählt wird, umso größer ist die Wahrscheinlichkeit, dass (unter Berücksichtigung der festgelegten Maximalabweichung D) hohe Abweichungen zwischen originalen und veränderten Fallzahlen auftreten.

4.3 Übergangsmatrix

Die maximale Abweichung von der Originalfallzahl sowie die Varianz der Überlagerungswerte bestimmen maßgeblich die Wahrscheinlichkeiten, mit der ein bestimmter Überlagerungswert einer Originalfallzahl zugeordnet wird. Im Folgenden bezeichnet $p_{jk} \in [0; 1]$ die Wahrscheinlichkeit, dass die Originalfallzahl j zur Fallzahl k verändert wird (beziehungsweise den Überlagerungswert $d = |j - k|$ erhält). Da diese Wahrscheinlichkeit nur von der Höhe der Originalfallzahl abhängt, jedoch nicht von einem konkreten Tabellenfeld i , kann hier auf den Index i verzichtet werden. Die (bedingten) Übergangswahrscheinlichkeiten werden für alle Kombinationen

möglicher Originalfallzahlen und veränderter Fallzahlen in der sogenannten Übergangsmatrix \mathbf{T} gesammelt:

$$\mathbf{T} = \begin{pmatrix} p_{00} & p_{01} & p_{02} & \dots & p_{0k} & \dots \\ p_{10} & p_{11} & p_{12} & \dots & p_{1k} & \dots \\ p_{20} & p_{21} & p_{22} & \dots & p_{2k} & \dots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots \\ p_{j0} & p_{j1} & p_{j2} & & p_{jk} & \\ \vdots & \vdots & \vdots & & \vdots & \ddots \end{pmatrix}$$

\mathbf{T} ist eine quadratische Matrix, die folgende Eigenschaften aufweist:

- › \mathbf{T} enthält auf der Hauptdiagonalen die sogenannten Bleibewahrscheinlichkeiten p_{jj} mit $j = k$, das heißt die Wahrscheinlichkeiten, dass eine Originalfallzahl j unverändert bleibt.
- › Zeilenweise enthält \mathbf{T} die bedingten Wahrscheinlichkeitenverteilungen für die Überlagerungen der Originalfallzahlen j , das heißt $\forall j: \sum_k p_{jk} = 1$.
- › Durch die zwingende Festlegung einer maximalen Abweichung D von den jeweiligen Originalfallzahlen ist $p_{jk} = 0$ für alle $k \notin \{j - D, \dots, j + D\}$.
- › Für kleine Fallzahlen $j < D$ kann das eigentlich durch D festgelegte Intervall für die möglichen Fallzahlen nach Datenveränderung $[j - D, \dots, j + D]$ nicht vollständig ausgeschöpft werden, um der Bedingung der Nicht-Negativität der veränderten Fallzahlen zu genügen. Das tatsächliche Intervall für die möglichen Fallzahlen nach Veränderung der Originalfallzahl j lautet somit $\Pi = [\max\{j - D; 0\}, \dots, j + D]$. Negative Überlagerungswerte sind bei kleinen Originalfallzahlen (bei gleichzeitiger Gewährleistung der Unverzerrtheit der veränderten Fallzahlen) somit nur eingeschränkt möglich (Höhne/Höninger, 2018).

4.4 Weitere optionale Restriktionen zur Kalibrierung der Übergangsmatrix

Neben den oben genannten zwingend festzulegenden Parametern D und V können optional weitere Nebenbedingungen formuliert werden, welche die Gestalt der Übergangsmatrix beeinflussen. Mögliche Restriktionen lauten:

- › **Festlegung einer Bleibewahrscheinlichkeit:** Es kann festgelegt werden, dass ein bestimmter Anteil P_V aller Originalfallzahlen unverändert bleiben soll. In diesem Fall entsprechen $P_V \cdot 100\%$ aller Fallzahlen in den geheim gehaltenen Tabellen ihrem jeweiligen Originalwert.
- › **Original-Nullen sollen unverändert bleiben:** $p_{00} = 1$ beziehungsweise $\forall k > 0: p_{0k} = 0$. In der Realität nicht existierende Ausprägungen von Merkmalskombinationen bleiben auch nach Anwendung der Cell-Key-Methode ausgeschlossen.
- › **Ausschluss kleiner Fallzahlen nach Datenveränderung:** Die Ausgabe sehr kleiner veränderter Fallzahlen kann optional bis einschließlich eines Schwellenwerts $m > 0$ ausgeschlossen werden. Für die entsprechenden veränderten Fallzahlen $k = 1, \dots, m$ gilt dann $\forall j: p_{jk} = 0$. Die entsprechenden Spaltenvektoren für die ausgeschlossenen veränderten Fallzahlen in \mathbf{T} entsprechen dann dem Nullvektor. Häufig wird facheitig die 1 als veränderte Fallzahl ausgeschlossen ($\forall j: p_{j1} = 0$).
- › **Gewährleistung einer symmetrischen Verteilung:** $\forall d \in \{-D, \dots, D\}: p_{j(j-d)} = p_{j(j+d)}$, das heißt Veränderungen einer Originalfallzahl um die Überlagerungswerte $-d$ und d besitzen die identische Wahrscheinlichkeit. Hieraus ergibt sich eine symmetrische Verteilung der Übergangswahrscheinlichkeiten für jede Fallzahl j um die entsprechende Bleibewahrscheinlichkeit p_{jj} . Dabei ist zu beachten, dass die Symmetrieeigenschaft der Überlagerungsverteilung nur für solche Originalfallzahlen j gewährleistet werden kann, für welche $j \geq D$ gilt. Werden zusätzlich kleine veränderte Fallzahlen bis einschließlich des Schwellenwerts m ausgeschlossen, erweitert sich diese Bedingung zu $j > D + m$.

4.5 Berechnung der Übergangsmatrix

Auf Basis der obligatorisch festzulegenden Parameter sowie unter Berücksichtigung der weiteren optionalen Restriktionen an die Ausgestaltung der Übergangsmatrix ist für jede Originalfallzahl die konkrete (bedingte) Verteilung der Überlagerungswerte zu berechnen. Marley/Leaver (2011) schlagen hierzu die Maximierung der Entropie der (bedingten) Überlagerungsverteilungen vor. Die Entropie stellt ein Streuungsmaß einer Wahrscheinlichkeitsverteilung dar, dessen Maximierung in diesem Zusammenhang als Minimierung des Aufdeckungsrisikos interpretiert werden kann (Marley/Leaver, 2011). Bezeichnet Π die Menge der für die betreffende Originalfallzahl möglichen Fallzahlen nach Datenveränderung, so lautet das Maximierungsproblem zur Berechnung der Überlagerungsverteilung für Originalfallzahl j

$$\max_{p_{jk}} \left\{ - \sum_{k \in \Pi} (p_{jk} \cdot \log_2 p_{jk}) \right\}.$$

Einschließlich aller formulierten Nebenbedingungen ergibt sich somit ein nicht-lineares Gleichungssystem. Gießing (2016) stellt einen Ansatz zur Lösung des Optimierungsproblems mittels eines Lagrange-Ansatzes vor, welcher im R-Paket *ptable* (Enderle, 2023) zur Erstellung von CKM-Übergangsmatrizen angewendet wird. Weitere Ausführungen zur Maximierung der Entropie sind Enderle/Vollmar (2019) zu entnehmen.

Für alle Originalfallzahlen $j > m + D$ ist die Überlagerungsverteilung strukturell identisch, das heißt $\forall l \geq 0: p_{jk} = p_{(j+l)(k+l)}$. Die resultierende Überlagerungsverteilung für $j = m + D + 1$ gilt somit – jeweils um $a \in \mathbb{N}$ Spalten in \mathbf{T} nach rechts verschoben – auch für alle größeren Originalfallzahlen $j + a$.

4.6 Record Keys und Cell Keys

Nach der erfolgten Spezifizierung der Übergangsmatrix \mathbf{T} werden den Originalfallzahlen konkrete Überlagerungswerte zugeordnet. Dies erfolgt anhand des vorliegenden Datenmaterials.

Dazu wird zunächst jedem Merkmalsträger $s = 1, \dots, S$ im Mikrodatensatz eine feste Zufallszahl $r_s \sim U[0; 1]$ zugewiesen. Die Zufallszahlen werden aus einer stetigen Gleichverteilung gezogen und als **Record Keys** bezeichnet. Der einem Merkmalsträger zugewiesene Record Key bleibt für alle Auswertungen, die für die betroffene Statistik erstellt werden, mindestens für die laufende Berichtsperiode identisch.

Auf Ebene der Tabellenfelder wird anhand der Record Keys eine „Kennziffer“ für jedes einzelne Tabellenfeld gebildet, welche zur Zuweisung des Überlagerungswertes für das betreffende Tabellenfeld verwendet wird. Dieser sogenannte **Cell Key** c_i wird für Tabellenfeld i gemäß

$$c_i = \sum_{s \in I} r_s - \left\lfloor \sum_{s \in I} r_s \right\rfloor \sim U[0; 1]$$

berechnet, wobei die Menge I alle Merkmalsträger s enthält, die zu Tabellenfeld i beitragen. Zur Berechnung des Cell Keys für Tabellenfeld i wird die Summe der Record Keys aller zu i beitragenden Merkmalsträger um deren nächst kleineren ganzzahligen Betrag reduziert (hinterer Term mit unterer Gauß-Klammer). Diese Rechenoperation ist notwendig, da gleichverteilte Zufallsvariablen ihre Verteilungseigenschaft bei Summierung (vorderer Term) verlieren und durch die Korrektur die Eigenschaft einer stetigen Gleichverteilung für die Cell Keys wiederhergestellt wird. Im Ergebnis weist nach diesem Schritt jedes zu überlagernde Tabellenfeld einen spezifischen Cell Key mit einem Wert $c_i \in [0; 1)$ auf, der von den konkreten Merkmalsträgern beziehungsweise deren Record Keys abhängt, die zum Tabellenfeld beitragen. Durch die hier dargestellte Vorgehensweise wird tabellenübergreifende Konsistenz der veränderten Fallzahlen sichergestellt, da logisch identische Tabellenfelder (das heißt mit identischen beitragenden Merkmalsträgern) stets den gleichen Cell Key erhalten.

4.7 Zuweisung der Überlagerungswerte

In einem letzten Schritt werden die generierten Cell Keys genutzt, um anhand der spezifizierten Übergangsmatrix zu entscheiden, welcher konkrete Überlagerungswert einem Tabellenfeld zugewiesen wird.

Die Übergangsmatrix wird dazu in ein sogenanntes Überlagerungstableau überführt, indem die Überlagerungsverteilung für jede Originalfallzahl j schrittweise aggregiert wird. Dazu sei die Verteilungsfunktion F_j gemäß

$$F_j(d) := \sum_{b \leq d; d \in \Lambda} p_{b|j}$$

definiert, wobei die Menge $\Lambda_j = \{d_1, d_2, \dots, d_{n-1}, d_n\}$ alle für j infrage kommenden Überlagerungswerte enthalte und $p_{d|j}$ die Wahrscheinlichkeit einer Überlagerung der Originalfallzahl j mit Überlagerungswert d bezeichnet. Die Höhe der einzelnen Übergangswahrscheinlichkeiten wird dabei über die Breite der Intervalle $[0; F_j(d_1)]$, $(F_j(d_1); F_j(d_2)]$, ..., $(F_j(d_{n-1}); F_j(d_n)]$ abgebildet, wobei $F_j(d_n) = 1$ ist. Die Vereinigungsmenge aller Intervalle deckt das Intervall $[0; 1]$ somit vollständig und überlappungsfrei ab.

Da die Cell Keys auf dem Intervall $[0; 1)$ gleichverteilt sind, kann die Zuweisung des Überlagerungswerts auf Basis eines einfachen Abgleichs zwischen dem Cell Key eines Tabellenfeldes und der Verteilungsfunktion der Überlagerungswerte vorgenommen werden. Dabei wird der Überlagerungswert d_i zum Tabellenfeld i mit Cell Key c_i anhand des folgenden Mechanismus zugeordnet:

$$d_i = d_r | d_r \in \Lambda: c_i \in (F_j(d_{r-1}); F_j(d_r)]$$

Als Überlagerungswert für das Tabellenfeld i wird somit der Überlagerungswert d_r ausgewählt, falls der Cell Key des Tabellenfeldes in das Teilintervall fällt, welches durch die Verteilungsfunktionswerte von d_{r-1} und d_r aufgespannt wird (und dessen Breite der Wahrscheinlichkeit einer Überlagerung mit d_r entspricht). Dieser Zuweisungsmechanismus wird auf alle zu überlagernden Tabellenfelder angewendet. Durch die Gleichverteilungseigenschaft der Cell Keys ist gewährleistet, dass über alle Tabellen hinweg der Anteil an mit einem bestimmten Überlagerungswert überlagerten Tabellenfeldern der jeweiligen Übergangswahrscheinlichkeit $p_{d|j} \cdot 100\%$ entspricht.

5

Fazit

Mit der Cell-Key-Methode hält ein neues Geheimhaltungsverfahren Einzug in die amtliche Statistik und damit auch in die Forschungsdatenzentren. Die Cell-Key-Methode ist ein Verfahren, das auf einer post-tabularen stochastischen Überlagerung basiert. Die feste Zuordnung eines Record Keys zu jeder Beobachtungseinheit stellt sicher, dass Veränderungen der originalen Fallzahlen konsistent und über verschiedene Ergebnisläufe hinweg replizierbar erfolgen. Dies geht jedoch zulasten der Additivität der Ergebnistabellen. Welche Auswirkungen die Anwendung der Cell-Key-Methode auf Tabellenergebnisse und darauf basierende Kennzahlen darüber hinaus hat, beschreibt im Detail der Artikel „Die Cell-Key-Methode in den Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder – Teil 2: Auswirkungen des neuen Geheimhaltungsverfahrens“ in derselben Ausgabe dieser Zeitschrift (Rothe und andere, 2024). [u](#)

LITERATURVERZEICHNIS

Enderle, Tobias/Vollmar, Meike. *Geheimhaltung in der Hochschulstatistik*. In: WISTA Wirtschaft und Statistik. Ausgabe 6/2019, Seite 87 ff.

Enderle, Tobias. *ptable: Generation of Perturbation Tables for the Cell-Key Method. R package version 1.0.0*. 2023. [Zugriff am 30. April 2024]. Verfügbar unter: CRAN.R-project.org

Fraser, Bruce/Wooton, Janice. *A proposed method for confidentialising tabular output to protect against differencing*. Work session on statistical data confidentiality. Supporting paper. Genf 2005. [Zugriff am 30. April 2024]. Verfügbar unter: unece.org

Giessing, Sarah/Tent, Reinhard. *Concepts for generalising tools implementing the cell key method to the case of continuous variables*. In: Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality. Den Haag 2019. [Zugriff am 30. April 2024]. Verfügbar unter: unece.org

Giessing, Sarah. *Computational Issues in the Design of Transition Probabilities and Disclosure Risk Estimation for Additive Noise*. In: Domingo-Ferrer, Josep/Peji-Bach, Mirjana (Herausgeber). *Privacy in Statistical Databases*. LNCS (Lecture Notes in Computer Science). 2016. Ausgabe 9867, Seite 237 ff. DOI: [10.1007/978-3-319-45381-1_18](https://doi.org/10.1007/978-3-319-45381-1_18)

Höhne, Jörg. *Das Geheimhaltungsverfahren SAFE*. In: Zeitschrift für amtliche Statistik Berlin Brandenburg. Ausgabe 2/2015, Seite 16 ff. [Zugriff am 30. April 2024]. Verfügbar unter: www.statistischebibliothek.de

Höhne, Jörg/Höninger, Julia. *Die Cell-Key-Methode – ein Geheimhaltungsverfahren*. In: Zeitschrift für amtliche Statistik Berlin Brandenburg. Ausgabe 3+4/2018, Seite 14 ff. [Zugriff am 30. April 2024]. Verfügbar unter: www.statistischebibliothek.de

Kleber, Birgit/Gießing, Sarah. *Geheimhaltung beim Zensus 2021*. In: Methoden – Verfahren – Entwicklungen. Nachrichten aus dem Statistischen Bundesamt. Ausgabe 2/2018, Seite 3 ff. [Zugriff am 30. April 2024]. Verfügbar unter: www.destatis.de

Marley, Jennifer K./Leaver, Victoria L. *A Method for Confidentialising User-Defined Tables: Statistical Properties and a Risk-Utility Analysis*. In: Proceedings of 58th World Statistical Congress. 2011. [Zugriff am 30. April 2024]. Verfügbar unter: 2011.isiproceedings.org

Rohde, Johannes/Seifert, Christiane/Gießing, Sarah/Setzer, Stefanie (unter Mitarbeit von Breitenfeld, Jörg/Brings, Stefan/Höhne, Jörg/Höninger, Julia/Rothe, Patrick/Schedding-Kleis, Ulrike). *Entscheidungskriterien für die Auswahl eines Geheimhaltungsverfahrens. Version 1.1 vom 23.04.2021*. Internes Dokument des Statistischen Verbunds (Statistische Ämter des Bundes und der Länder).

Rothe, Patrick/Güttgemanns, Volker/Rohde, Johannes/Setzer, Stefanie. *Die Cell-Key-Methode in den Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder. – Teil 2: Auswirkungen des neuen Geheimhaltungsverfahrens*. In: WISTA Wirtschaft und Statistik. Ausgabe 3/2024, Seite 45 ff.

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.

DIE CELL-KEY-METHODE IN DEN FORSCHUNGSDATENZENTREN DER STATISTISCHEN ÄMTER DES BUNDES UND DER LÄNDER

Teil 2: Auswirkungen des neuen Geheimhaltungsverfahrens

Patrick Rothe, Volker Güttgemanns, Johannes Rohde, Stefanie Setzer

➤ **Schlüsselwörter:** Geheimhaltung – stochastische Überlagerung – post-tabular – Verhältniszahlen – Zeitreihen

ZUSAMMENFASSUNG

Die Cell-Key-Methode ist ein hauptsächlich für die Geheimhaltung von Fallzahltabellen entwickeltes Geheimhaltungsverfahren. In den Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder werden häufig umfangreiche Analysen vorgenommen, auf deren Ergebnisdarstellung sich das neue Verfahren ebenfalls auswirken kann. Der Artikel beschreibt die Auswirkungen, die das Verfahren auf die Ergebnisqualität, Fallzahltabellen, Verhältniszahlen und Zeitreihen hat.

➤ **Keywords:** confidentiality – stochastic perturbation – post-tabular – ratios – time series

ABSTRACT

The cell key method is a disclosure control method that was primarily developed to ensure the confidentiality of frequency tables. Extensive analyses are frequently carried out in the Research Data Centres of the statistical offices of the Federation and the Länder, and the new method can affect the presentation of the results. This article describes the effects that the method has on the quality of results, frequency tables, ratios and time series.

Patrick Rothe

hat Sozialwissenschaften an der Universität Mannheim studiert und ist seit 2011 im Bayerischen Landesamt für Statistik tätig. Seit 2018 leitet er dort das Sachgebiet „Grundsatzfragen der amtlichen Statistik, Digitalisierung, Forschungsdatenzentrum, Kompetenzzentrum Analyse“. Inhaltlich beschäftigt er sich schwerpunktmäßig unter anderem mit der statistischen Geheimhaltung.

Volker Güttgemanns

hat einen Master of Science in Wirtschaftswissenschaften und war von 2017 bis 2023 stellvertretende Leitung der Geschäftsstelle des Forschungsdatenzentrums der Statistischen Ämter der Länder.

Dr. Johannes Rohde

hat Wirtschaftswissenschaften an der Leibniz Universität Hannover studiert und dort 2015 seine Promotion im Bereich Statistik abgeschlossen. Bei IT.NRW leitet er den Service „Mathematisch-statistische Methoden und experimentelle Statistik“.

Stefanie Setzer

ist Diplom-Soziologin und Referentin im Referat „Forschungsdatenzentrum, Methoden der Datenanalyse“ des Statistischen Bundesamtes. Schwerpunkt ihrer Arbeit ist die fachliche und methodische Weiterentwicklung des Arbeitsbereichs.

1

Einleitung

Die Statistischen Ämter des Bundes und der Länder führen für ausgewählte Statistiken ein neues Geheimhaltungsverfahren ein: die Cell-Key-Methode (CKM)¹. Um die Geheimhaltung der Ergebnisse über alle Veröffentlichungen hinweg sicherzustellen, wenden die Forschungsdatenzentren der amtlichen Statistik dieses Verfahren entsprechend an. Der Artikel „Die Cell-Key-Methode in den Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder – Teil 1: Vorstellung des neuen Geheimhaltungsverfahrens“ (Setzer und andere, 2024) beschreibt das Verfahren ausführlich.

Die Schutzwirkung der Cell-Key-Methode entsteht durch eine post-tabulare Überlagerung aller Fallzahlen. In der überlagerten Tabelle lässt sich hierdurch letztendlich nicht mehr erkennen, welche Werte überlagert wurden und welche weiterhin dem Originalwert entsprechen. Diese Vorgehensweise stellt einen großen Unterschied zur bisher überwiegend angewandten Zellsperre dar, in der ausgewiesene Werte immer dem Originalwert entsprechen, insbesondere kleine Fallzahlen aber gesperrt werden. Die Auswirkungen des neuen Verfahrens auf die Ergebnisqualität, Fallzahltabellen, Verhältniszahlen und Zeitreihen werden in diesem Beitrag vorgestellt. Dazu beschreibt Kapitel 2 detailliert die Auswirkungen auf Fallzahltabellen, Kapitel 3 befasst sich mit den Effekten, die die Cell-Key-Methode auf Verhältniszahlen haben kann. Auch auf Zeitreihen kann sich die Anwendung der Cell-Key-Methode auswirken, wie in Kapitel 4 dargestellt wird. Ein kurzes Fazit zur Nutzung des neuen Geheimhaltungsverfahrens in den Forschungsdatenzentren beschließt den Artikel.

¹ Die Cell-Key-Methode wurde vom australischen Statistikamt (Australian Bureau of Statistics) entwickelt (Fraser/Wooton, 2005).

2

Auswirkungen auf Fallzahltabellen

2.1 Kurzüberblick

Bei Fallzahltabellen, die mit der Cell-Key-Methode überlagert wurden, sind zwei wesentliche Auswirkungen besonders hervorzuheben:

- › Mit der Cell-Key-Methode geheim gehaltene Tabellen sind immer **konsistent**. Kommen logisch identische Tabellenfelder also in verschiedenen Tabellen vor (zum Beispiel in einer univariaten Fallzahltable und als Randwert einer Kreuztable), wird immer die gleiche überlagerte Fallzahl ausgegeben. Dies gewährleisten eine vorher festgelegte Übergangsmatrix und der Record Key, der den einzelnen Erhebungseinheiten fest zugeordnet ist.
- › Mit der Cell-Key-Methode geheim gehaltene Tabellen sind **nicht additiv**. Da Tabelleninnen- und -randfelder unabhängig voneinander überlagert werden, addieren sich die überlagerten Innenfelder nicht (oder nur zufällig) zu den überlagerten Randsummen. Eine Wiederherstellung der Additivität wäre zwar theoretisch möglich, würde aber die Konsistenz und Qualität der Ergebnisse beeinträchtigen und zudem die benötigte Rechenzeit erhöhen, sodass darauf verzichtet wird.

2.2 Auswirkungen im Detail

Additivität und Konsistenz der geheim gehaltenen Tabellenergebnisse stellen zwei wesentliche Anforderungen an statistische Verfahren zur Geheimhaltung von Tabellen dar. Gerade im Kontext der Verwendung der Cell-Key-Methode sind diese beiden Eigenschaften von besonderer Bedeutung.

Additivität einer Tabelle ist dann gegeben, wenn sich die Innenfelder der Tabelle zur ausgewiesenen Summe im entsprechenden Randfeld aufaddieren lassen – was bei einer unbearbeiteten Tabelle der Normalfall ist und in aller Regel von den Nutzenden auch so erwartet wird. Bestimmte Verfahren zur statistischen Geheimhaltung, die aus der Familie der datenverändernden Methoden stammen, führen im Ergebnis jedoch dazu, dass diese

Eigenschaft verletzt wird. Somit entspricht die Summe der aufaddierten Innenfelder nicht zwangsläufig der in der Tabelle ausgewiesenen Randsumme. Dieser geschilderte Effekt („Nicht-Additivität“) tritt auch bei der Cell-Key-Methode auf und entsteht, indem jedes Feld einer Tabelle separat – das heißt unabhängig von allen anderen Tabellenfeldern – dem datenverändernden Algorithmus unterzogen wird. Innenfelder werden somit genauso wie die in der Tabelle enthaltenen Zwischen- oder Gesamtsummen behandelt. Dieses separate Vorgehen bewirkt in der Regel einen Verlust der Additivitätseigenschaft.

Auf den ersten Blick kann dieser Effekt für die Nutzenden irritierend erscheinen, letztlich führt dieses Vorgehen unter Gesichtspunkten der Datenqualität und Informationserhaltung jedoch zu einem besseren Ergebnis: Die separate Überlagerung jedes Tabellenfeldes verhindert, dass sich Abweichungen über eine Tabellenzeile oder -spalte hinweg aufaddieren und die nach Geheimhaltung ausgewiesene Randsumme gegenüber dem Originalwert stark abweicht. Eine Veränderung um die Höhe der Maximalabweichung multipliziert mit der Anzahl der beteiligten Tabellenfelder wäre dabei im Extremfall für Randsummen nicht ausgeschlossen. Die unabhängige Überlagerung aller Tabellenfelder führt jedoch auch bei Tabellenfeldern mit Zwischen- oder Gesamtsummen dazu, dass der veränderte Wert vom Originalwert niemals weiter abweichen kann als es die vorher festgelegte Maximalvorgabe zulässt.¹²

Dieser Vorteil hinsichtlich der Datenqualität wird jedoch durch die Diskrepanz zwischen den aufaddierten Summen der einzelnen Tabellenfelder und den in der Tabelle ausgewiesenen Zwischen- und Gesamtsummen erkauft. Dieser Unterschied kann unter Umständen deutlich ausfallen. Daher sollten Nutzende darauf verzichten, selbstständig Rechenoperationen mit den Angaben aus den per Cell-Key-Methode geheim gehaltenen Ergebnistabellen vorzunehmen, sondern direkt auf die in der Tabelle ausgewiesenen Summenfelder zurückgreifen.

Eine Nicht-Additivität fällt immer dann besonders auf, wenn nur sehr wenige Tabellenfelder zu einer Randsumme beitragen und diese durch einfaches Kopfrechnen sehr schnell ermittelt werden kann. Ein Beispiel hierfür ist die Aufgliederung des Merkmals „Geschlecht“

nach den drei Merkmalsausprägungen „weiblich“, „männlich“ und „divers“ in einer fiktiven Tabelle: Hierbei ist die Summation der Fallzahlen in den drei Innenfeldern sehr einfach möglich, wobei es sehr wahrscheinlich ist, dass die selbstständig berechnete Summe von der in der geheim gehaltenen Tabelle ausgewiesenen Gesamtsumme abweicht. In größeren Tabellen mit einer größeren Anzahl an beitragenden Spalten oder Zeilen fällt dieses Problem jedoch nicht direkt ins Auge.

Konsistenz hingegen bezeichnet die Eigenschaft eines spezifischen Tabellenfeldes, immer den identischen inhaltlichen Wert auszuweisen – unabhängig von der konkreten Tabelle, in der diese Merkmalskombination ausgewiesen wird. Nicht alle Geheimhaltungsverfahren können dies gewährleisten, da hierfür zwingend gesichert sein muss, dass ein- und dasselbe inhaltliche Tabellenfeld in allen Fällen durch die gewählte Methode identisch behandelt wird. Die Cell-Key-Methode ist aufgrund ihrer Ausgestaltung in der Lage, diese Eigenschaft zu erfüllen. Gewährleistet wird dies durch die Aggregation der zu einem individuellen Tabellenfeld beitragenden Record Keys zum für das Verfahren namensgebenden Cell Key. Da dieser bei Vorhandensein derselben Kombination von Merkmalsträgern jeweils identisch ausfällt, ist auch die Datenveränderung stets identisch. Das gilt unabhängig davon, in welcher Tabelle und auf welchem Veröffentlichungsweg das entsprechende Tabellenfeld publiziert wird.

Neben einer höheren Nutzendenfreundlichkeit – die Ergebnisse hängen beispielsweise nicht vom Abrufzeitpunkt ab – geht diese Eigenschaft auch mit einem gesteigerten Schutz der Daten und der für deren Überlagerung genutzten Parameter einher. Bei nicht konsistentem Verhalten könnte im Gegensatz hierzu bei einer größeren Anzahl an Ergebnisabrufen – zumindest näherungsweise – auf die dahinterliegende Originalangabe geschlossen werden, da sich die positiven und negativen Abweichungen vom sich ergebenden Mittelwert bei einem erwartungstreuen Verfahren im Mittel ausgleichen.

¹² Weitere Betrachtungen zur Nicht-Additivität der Cell-Key-Methode finden sich in Höhne/Höninger (2018).

Übersicht 1

Auswirkungen der Cell-Key-Methode auf die Additivität einer Fallzahltablelle

Originaltabelle					Überlagerte Tabelle				
Einkommen	Alter				Einkommen	Alter			
	jung		alt	Σ		jung		alt	Σ
	niedrig	0	3	3		niedrig	0	4	4
	mittel	4	5	9		mittel	4	6	8
Σ	hoch	2	1	3	Σ	hoch	0	0	3
	Σ	6	9	15		Σ	6	10	15

Cell-Key-Methode

Farblegende			
Ausgangswert unverändert	Randsumme ist Summe der Innenfelder		
	ja		nein
	ja		
nein	ja		
	nein		

➤ Übersicht 1 basiert auf dem Beispiel des Artikels „Die Cell-Key-Methode in den Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder – Teil 1: Vorstellung des neuen Geheimhaltungsverfahrens“ (Setzer und andere, 2024) und zeigt die Veränderungen, die durch die Cell-Key-Methode bezüglich der Randsummen ausgelöst werden können.

2.3 Weitere Qualitätskriterien

Ein Geheimhaltungsverfahren sollte stets so konzipiert sein, dass das Gleichgewicht zwischen dem Schutz der Angaben der Befragten und dem Erhalt des Informationspotenzials einer Statistik sichergestellt ist. Höhne/Höninger (2018) benennen dafür neben den bereits erläuterten Aspekten der Konsistenz und Additivität drei weitere grundsätzliche Ziele von Geheimhaltungsverfahren, wobei unterschiedliche Verfahren die einzelnen Ziele unterschiedlich priorisieren. ➤ Übersicht 2 stellt diese Ziele vor und geht auf deren Erfüllung durch die Cell-Key-Methode und die aktuell in den Forschungsdatenzentren überwiegend genutzte Zellsperren ein.

Zur Frage der Akzeptanz der Ergebnisse (Punkt 3): Insgesamt haben interne Testrechnungen der statistischen Ämter ergeben, dass der durch Anwendung der Cell-Key-Methode verursachte Informationsverlust bei der Wahl

geeigneter Parameter gering ist (Rohde und andere, 2021). Dabei ist zu beachten, dass sich bei kleinen (Original-)Fallzahlen auch Überlagerungen mit einem absolut gesehen geringen Wert stark auswirken können: Wird beispielsweise die Originalfallzahl 3 mit dem Wert +6 überlagert (+200%), ist die relative Abweichung deutlich höher als bei der entsprechenden Überlagerung einer höheren Fallzahl (zum Beispiel +0,6% bei einer Originalfallzahl von 1000). Die Generierung von Tabellen mit vielen kleinen Fallzahlen ist daher nach Möglichkeit zu vermeiden, indem Auswertungen nicht auf einer unnötig tiefen regionalen oder fachlichen Gliederungsebene vorgenommen werden.

Übersicht 2

Vergleich von Cell-Key-Methode und Zellspernung bei Erreichen der Geheimhaltungsziele

Ziel	Cell-Key-Methode	Zellspernung
1. Keine unplausiblen Werte	Durch die Ausgestaltung der Übergangsmatrix kann sichergestellt werden, dass in Fallzahltabellen keine unplausiblen Ergebnisse generiert werden (Originalfallzahl 0 wird nicht verändert). Bei der Berechnung von Kennzahlen kann es jedoch vor allem bei kleinen Fallzahlen zu Unplausibilitäten kommen, wenn beispielsweise Zähler und Nenner von Verhältniszahlen gegenläufig verändert werden oder sich bei Trendanalysen das Vorzeichen ändert. Diese ungewünschten Effekte lassen sich jedoch durch geeignete Maßnahmen vermeiden.	Fallzahlen werden nicht verändert, sondern bei zu geringer Besetzung gesperrt. Daher können keine unplausiblen Fallzahlen entstehen. Dies gilt auch für die Berechnung von Kennzahlen und Zeitreihen.
2. Schutz der Einzelangaben	Bei der Ausgestaltung der Übergangsmatrix wird sichergestellt, dass die Veränderungen groß genug sind, um den Schutz der Einzelangaben sicherzustellen.	Fallzahlen unterhalb einer festgelegten Mindestfallzahl werden gesperrt, durch Gegensperrungen wird eine Rückrechnung verhindert. Der Schutz der Einzelangaben ist also bei korrekter Anwendung des Verfahrens sichergestellt.
3. Akzeptanz der Ergebnisse	Bei der Ausgestaltung der Übergangsmatrix wird berücksichtigt, dass die Veränderungen (unter Berücksichtigung von Punkt 2) klein genug sind, um bei einem Großteil der Nutzenden auf Akzeptanz zu stoßen.	Je nach Größe der betrachteten (Sub-)Population und der Detailtiefe der betrachteten Merkmale werden die Sperrungen von einigen Nutzenden als zu restriktiv betrachtet.
4. Ergebnisse sind konsistent	Logisch identische Tabellenfelder werden immer mit dem gleichen Wert überlagert.	Sperrmuster werden auch tabellenübergreifend umgesetzt, so dass Veränderungen bei korrekter Umsetzung des Zellsperverfahrens konsistent sind. Bei häufig genutzten Statistiken und vielen Ergebnistabellen können übersehene Zusammenhänge zwischen Tabellenfeldern allerdings zu Fehlern führen.
5. Ergebnisse sind additiv	Durch die unabhängige Überlagerung der einzelnen Fallzahlen sind geheim gehaltene Tabellen nicht additiv. ¹	Additivität bleibt in Tabellen bestehen, sofern keine Sperrungen erfolgen.

Ziel voll erreicht

Ziel teilweise erreicht

Ziel nicht erreicht

¹ Die Additivität könnte durch einen Algorithmus wiederhergestellt werden, der die überlagerten Werte so verändert, dass sich die Tabelleninnenfelder wieder zu den Randfeldern addieren. Dieses Vorgehen hätte aber zwei entscheidende Nachteile: Zum einen erfordert die Herstellung der Additivität zusätzliche Rechenleistung und verlängert die für die Tabellenerstellung benötigte Laufzeit, zum anderen wären die Ergebnisse danach nicht mehr konsistent, was die Qualität der Daten entscheidend beeinträchtigen würde.

3

Verhältniszahlen

Verhältniszahlen sind mathematische Beziehungen zwischen statistischen Größen, die einen sinnvollen Zusammenhang darstellen. Generell können drei Kategorien von Verhältniszahlen unterschieden werden (Bourier, 2011):

- Gliederungszahlen:** Diese setzen eine Teilgröße in Beziehung zur Gesamtgröße und werden oft als Anteilswerte bezeichnet, wie bei Geschlechterquoten.
- Beziehungszahlen:** Hier werden unterschiedliche Größen miteinander in Beziehung gesetzt, die in einem sachlichen Zusammenhang stehen, wie das Bruttoinlandsprodukt je Einwohner/-in.
- Messzahlen:** Diese setzen inhaltlich ähnliche Größen in Beziehung, jedoch zu unterschiedlichen Zeitpunkten, Zeiträumen oder Regionen, wie die Veränderung des Bruttoinlandsprodukts im Vergleich zum Vorjahr.

In der wissenschaftlichen Forschung ist die Berechnung von Verhältniszahlen eine gängige Methode. Dabei ist es wichtig, die geltenden Geheimhaltungsvorschriften und -regelungen zu beachten, insbesondere auch im Rahmen von Forschungsprojekten in den Forschungsdatenzentren. Eine grundlegende Anforderung besteht vor diesem Hintergrund darin, die Geheimhaltung gemäß den fachspezifischen Konzepten sicherzustellen, um ein konsistentes Vorgehen über verschiedene Analysen und Veröffentlichungen hinweg zu gewährleisten. Dies kann auch spezielle statistische Regelungen für den Umgang mit Verhältniszahlen erforderlich machen, die von den Forschungsdatenzentren wie von ihren Nutzenden zu berücksichtigen sind.

Für Nutzende der Forschungsdatenzentren ist es entscheidend, die Methodik der Geheimhaltung und deren Auswirkungen zu verstehen, um die Qualität der berechneten Verhältniszahlen selbst einschätzen zu können. Der folgende Abschnitt erläutert die Auswirkungen der Cell-Key-Methode auf Verhältniszahlen sowie auf deren Aussagekraft.

3.1 Auswirkungen auf Verhältniszahlen

Basis für die Geheimhaltung von Verhältniszahlen mit der Cell-Key-Methode ist die Veränderung der Zähler und Nenner entsprechend ihrer Cell Keys. Dadurch kann es vor allem bei kleinen Fallzahlen passieren, dass sich das errechnete Verhältnis deutlich vom Wert der nicht überlagerten Verhältniszahl unterscheidet. Diese Ungenauigkeiten verringern sich mit größeren Fallzahlen. Aus diesem Grund wird von einer Berechnung von Verhältniszahlen abgeraten, wenn diese auf nur wenigen Fällen beruhen.

3.2 Umgang mit Verhältniszahlen

Häufig werden Verhältniszahlen aus den überlagerten Fallzahlen von Zähler und Nenner berechnet, was hier als „A-posteriori-Verhältniszahl“ bezeichnet wird.¹³ Daher kann es bei datenverändernden Geheimhaltungsverfahren wie der Cell-Key-Methode zu unerwünschten Effekten kommen:

- › **Ungenauigkeit:** Bei der Anwendung der Cell-Key-Methode kann die Veränderungsrichtung der Fallzahlen zufällig stark gegenläufig sein, das heißt Zähler und Nenner werden um einen hohen positiven beziehungsweise negativen Wert (oder umgekehrt) verändert. Das kann insbesondere bei kleinen Fallzahlen zu erheblichen Abweichungen zwischen der ursprünglichen und der A-posteriori-Verhältniszahl führen. Dieser Effekt verringert sich jedoch mit steigenden Fallzahlen.
- › **Unplausibilität:** Ein weiterer unerwünschter Effekt kann auftreten, wenn zum Beispiel der ursprünglich kleinere Zähler nach der Veränderung größer ist als der veränderte Nenner. In diesem Fall würden sich A-posteriori-Anteilswerte über 100 % ergeben.
- › **Veränderung der Aussage:** Besonders bei der Analyse von Zeitreihen kann die Anwendung der Cell-Key-Methode zu Trendverzerrungen oder sogar zu einer Trendumkehr führen (siehe Kapitel 4).

Während die Überlagerung von Fallzahlen eine feste Varianz aufweist, stellen Enderle und andere (2018)

fest, dass dies nicht auf Verhältniszahlen zutrifft, die auf Basis zweier überlagerter Werte berechnet wurden. Für das Ausmaß der Abweichung des tatsächlichen Verhältniszahls $R := X/Y$ zum überlagerten Verhältniszahl $\hat{R} := \hat{X}/\hat{Y}$ spielen die Originalfallzahlen, aus denen das Verhältnis berechnet wird, eine große Rolle: je größer die Fallzahlen X und Y , desto geringer die Abweichung des Verhältniszahls. Zur Veranschaulichung der Problematik kleiner Fallzahlen nennen Enderle und andere (2018) folgendes Beispiel: Die relativ kleinen Originalfallzahlen $x = 4$ und $y = 4$ werden zueinander ins Verhältnis gesetzt. Für den tatsächlichen Verhältniszahl gilt damit: $R = 4/4$, also 100 %. Bei einer Maximalabweichung von $D = 3$ könnten die überlagerten Fallzahlen die Werte $\hat{x} = x + D = 7$ und $\hat{y} = y - D = 1$ annehmen. Der überlagerte Verhältniszahl wäre dann $\hat{R} = 7/1$, also 700 %.

3.3 Beurteilung der Qualität eines Anteilswertes

Die Qualität eines Anteilswertes muss von Wissenschaftlerinnen und Wissenschaftlern bewertet werden können. Dafür sind Informationen über die Verteilung der zufälligen Abweichungen von der Originalgröße erforderlich, wobei jedoch die Details dieser Abweichungen nicht offengelegt werden dürfen. Um die Qualität von Anteilswerten zu beurteilen kann die relative Standardabweichung als ein Maß für die Streuung um die Originalfallzahl verwendet werden.¹⁴

Die relative Standardabweichung für Anteilswerte oder für alle Verhältniszahlen, die als Quotienten aus veränderten Fallzahlen im Zähler und Nenner gebildet werden, wird wie folgt berechnet:

Angenommen, q_{xy} repräsentiert die Wahrscheinlichkeit, dass einem veränderten Anteilswert $\tilde{v} := \left(\frac{\tilde{x}}{\tilde{y}}\right)$ der Originalzähler $x \in \mathcal{D}_x$ und der Originalnenner $y \in \mathcal{D}_y$ zugrunde liegen, und $d(v)_{xy}$ repräsentiert die Abweichungen zwischen dem veränderten Anteilswert \tilde{v} und den möglichen originalen Anteilswerten v , dann kann die (absolute) Standardabweichung für den veränderten Anteilswert \tilde{v} wie folgt approximiert werden:

¹³ Weiterführende Vorschläge für Techniken zur Anwendung stochastischer Überlagerung bei Verhältniszahlen, die als Quotient aus zwei Wertsummen gebildet werden, finden sich in Gießing (2013).

¹⁴ Anstelle der absoluten Standardabweichung wird die relative Standardabweichung als Gütemaß herangezogen, da diese die zufällige Streuung um die Originalfallzahl ins Verhältnis zur Größe des Originalwertes setzt.

$$\sigma(\tilde{v}) = \sqrt{\sum_{\substack{x \in \mathcal{D}_x \\ y \in \mathcal{D}_y}} q_{xy} \cdot d(v)_{xy}^2}$$

Der entsprechende Relative Root Mean Square Error (RRMSE) oder die approximierte relative Standardabweichung (auch Variationskoeffizient genannt) für den veränderten Anteilswert \tilde{v} ergibt sich dann gemäß

$$k(\tilde{v}) = \frac{\sigma(\tilde{v})}{\tilde{v}}.$$

Für Mittelwerte oder Verhältniswerte, bei denen die Wertsumme im Zähler und/oder Nenner (gegebenenfalls verändert) in die Berechnung einfließt, erfordert das beschriebene Verfahren einige Anpassungen. Eine alternative Vorgehensweise besteht darin, die relative Standardabweichung von v mithilfe der Übergangswahrscheinlichkeiten zu bestimmen.

Für die Hochschulstatistik wurden die Auswirkungen unterschiedlicher Parametrisierungen auf die Qualität von Verhältniszahlen untersucht. Das Ergebnis zeigt, dass die Wahl von Bleibewahrscheinlichkeiten und maximaler Abweichung von der Originalfallzahl nicht die entscheidenden Faktoren für die Qualität von Verhältniszahlen darstellen. Vielmehr hängt die Qualität vor allem von der Größe der zugrunde liegenden Fallzahlen ab, die in die Berechnung einfließen. Demnach beeinträchtigen insbesondere kleine Fallzahlen die Qualität und Aussagekraft der Verhältniszahlen. Erst ab einer bestimmten Größe der Basiszahlen im Zähler und Nenner kann eine ausreichende statistische Aussagekraft gewährleistet werden (Enderle/Vollmar, 2019).

3.4 Empfehlungen bei der Berechnung von Verhältniszahlen

Ohne Kenntnis der unveränderten Verhältniszahlen ist es für Nutzende nicht möglich, die Qualität einer Verhältniszahl zu beurteilen. Die Höhe der relativen Standardabweichung oder des RRMSE hängt in erster Linie von der Größe der einbezogenen Fallzahlen ab. Daher wird allgemein empfohlen, Verhältniszahlen nur auf Basis ausreichend hoher Fallzahlen im Zähler und Nenner zu berechnen, denn die Varianz geht in diesen Fällen ohnehin gegen Null.

Sollte die Berechnung einzelner Verhältniszahlen auf Basis geringer Fallzahlen für eine Forschungsfrage unerlässlich sein, können Nutzende der Forschungsdatenzentren ihren betreuenden FDZ-Standort um Unterstützung bitten.

4

Zeitreihen

Zeitreihen sind wertvolle Instrumente, um zeitliche Verläufe und Entwicklungen darzustellen. Hierbei werden Kennzahlen auf der Basis von Zeitpunkten oder Zeiträumen berechnet. Dies ermöglicht die Analyse von absoluten oder relativen Veränderungen über die Zeit hinweg.

Wie bei der Berechnung von Verhältniszahlen sind auch bei der Betrachtung von Zeitreihen die geltenden Geheimhaltungsvorschriften und -regelungen der Fachstatistik zu beachten.

4.1 Auswirkungen auf Zeitreihen

Die Anwendung der Cell-Key-Methode kann negative Auswirkungen auf die Analyse von Zeitreihen haben. Das gilt sowohl beim Vergleich von zwei mittels Cell-Key-Methode geheim gehaltenen Erhebungswellen als auch bei der Betrachtung von zwei Erhebungswellen mit unterschiedlicher Geheimhaltung, also beim Vergleich der Erhebungsjahre vor und nach Einführung der Cell-Key-Methode. Besonders bei sehr geringen Unterschieden in den unveränderten Fallzahlen der beiden verglichenen Beobachtungszeiträume kann es bei gegenläufiger Veränderung der betrachteten Werte vorkommen, dass Unterschiede zwischen den Erhebungswellen vergrößert oder verkleinert werden. Im Extremfall ist sogar eine Trendumkehr möglich. Sehr schwache Veränderungen im Zeitverlauf sind bei Anwendung der Cell-Key-Methode daher mit Vorsicht zu interpretieren.

4.2 Umgang mit Zeitreihen

Zeitreihen basieren auf Daten aus verschiedenen Zeiträumen und werden verwendet, um die zeitliche Entwicklung von Sachverhalten zu analysieren. Die einzelnen Datenpunkte zu den verschiedenen Berichtszeiträumen bilden die Basis für entsprechende Betrachtungen.

Die Bildung von Differenzen aus veränderten Datenpunkten kann zu weniger sicheren Ergebnissen führen. Das gilt insbesondere dann, wenn eine Differenz anhand kleiner Fallzahlen berechnet wird.

Bei der Berechnung von Differenzen sind bei der Geheimhaltung mit der Cell-Key-Methode spezifische Herausforderungen zu berücksichtigen:

I. Umgang mit Differenzen, wenn beide Datenpunkte aus CKM-Datenbeständen stammen

Bei der Differenz zwischen zwei zufällig veränderten Datenpunkten können ähnliche Herausforderungen auftreten wie bei Saldierungen (siehe Abschnitt 4.3). Dies trifft insbesondere bei kleinen Fallzahlen zu und wenn die einzelnen Punkte starke, gegenläufige Veränderungen aufweisen. In der Konsequenz kann die Anwendung der Cell-Key-Methode theoretisch Veränderungen in der Trendstärke und sogar eine Änderung des Trendvorzeichens verursachen.

Für diese Herausforderungen gibt es (statistikspezifische) Lösungsansätze. Im Zensus 2022 wird eine Methode gewählt, die einen Vorzeichenwechsel vermeidet. In der Bevölkerungsstatistik und der Hochschulstatistik hingegen wird die Differenz aus den beiden überlappenden Datenpunkten gebildet. Aus methodischer Sicht führt dies zu einer Verdopplung der in den CKM-Parametern vorgegebenen Varianz. Wenn X_1 und X_2 zwei veränderte Datenpunkte mit identischer, als CKM-Parameter festgelegter Varianz V sind, gilt für ihre Differenz:

$$\text{Var}(X_1 - X_2) = \text{Var}(X_1) + \text{Var}(X_2) = 2V$$

Dies erhöht im Vergleich zu den einzelnen absoluten Fallzahlen das Risiko einer größeren Abweichung von der Originaldifferenz. Basieren die Zeitreihen (oder die einzelnen Datenpunkte) jedoch auf ausreichend großen Fallzahlen, wird der Effekt einer Verzerrung der Trendstärke vernachlässigbar.

Um das Risiko einer wesentlichen Veränderung der Trendstärke zu minimieren, empfehlen die Forschungsdatenzentren, in wissenschaftlichen Veröffentlichungen Analysen auf Basis von Zeitreihen nur auf Basis einer ausreichend hohen Fallzahl durchzuführen. Wie hoch diese Fallzahlgrenze ist, hängt von der genutzten Statistik ab sowie von den betrachteten Merkmalen. Der für die jeweilige Statistik fachlich zuständige FDZ-Standort kann hierzu beratend unterstützen.

II. Umgang mit Differenzen, wenn die Datenpunkte aus unterschiedlich geheim gehaltenen Datenbeständen stammen

Ändert sich das Geheimhaltungsverfahren zwischen zwei Berichtszeitpunkten, beispielsweise durch den Wechsel von der Zellsperrung zur Cell-Key-Methode, führt dies zu einem methodischen Bruch in der Zeitreihe. Die Forschungsdatenzentren haben für diesen Fall geregelt, dass an dem Punkt des Bruchs, also dort, wo der Wechsel des Geheimhaltungsverfahrens erfolgt, keine Differenzen berechnet werden dürfen. Dies hat zur Konsequenz, dass die Zeitreihe am Bruchpunkt unterbrochen ist. Vergleiche zwischen den Berichtszeitpunkten der Umstellung sind daher nicht möglich. Einige Fachstatistiken, wie der Zensus, haben jedoch entschieden, frühere Wellen ihrer Statistik nachträglich ebenfalls mit der Cell-Key-Methode geheim zu halten. Entsprechende Vergleiche anhand der beiden mit der Cell-Key-Methode geheim gehaltenen Datenbestände sind dann wieder möglich. Die Information, ob das für eine bestimmte Statistik durchgeführt wurde, findet sich in den entsprechenden von den Forschungsdatenzentren bereitgestellten Metadatenreports.

4.3 Umgang mit Saldierungen


Der Umgang mit Saldierungen ist in Bezug auf die grundlegende Problematik sehr ähnlich dem Umgang mit Zeitreihen. Wird ein Saldo, also die Differenz zweier veränderter Fallzahlen, berechnet (beispielsweise der Wanderungssaldo in der Bevölkerungsstatistik), so führt die Verknüpfung zweier stochastischer Größen zu einer höheren Unsicherheit des Ergebnisses. Dies geschieht, weil die Varianz der Differenz im Vergleich zur vorhandenen Varianz bei der Veränderung einzelner Fallzahlen verdoppelt wird (siehe Abschnitt 4.2). Dadurch kann die Aussagekraft des Saldos bei kleinen Fallzahlen verzerrt werden. Zudem besteht die theoretische Möglichkeit, dass bei Anwendung der Cell-Key-Methode das Vorzeichen des Saldos wechselt, wenn beide Originalfallzahlen gegenläufig verändert werden und die Fallzahlen sowohl klein sind als auch nahe beieinanderliegen. Der relative Effekt solcher Verzerrungen aufgrund möglicher großer Abweichungen von der ursprünglichen Differenz nimmt jedoch ab, je größer die zugrunde liegenden Fallzahlen sind – ähnlich wie bei Verhältniszahlen und Zeitreihendifferenzen.

Die Forschungsdatenzentren empfehlen daher auch bei Saldierungen, entsprechende Auswertungen nur auf Basis ausreichend hoher Fallzahlen vorzunehmen. Bei vielen kleinen Fallzahlen sollte die Auswertung nach Möglichkeit auf einer höheren regionalen oder fachlichen Aggregationsebene (zum Beispiel Landkreise statt Gemeindeebene oder Wirtschaftszweig-4-Steller statt -5-Steller) erfolgen, ebenso sollten schwach besetzte Kategorien gruppiert werden.

5

Fazit

Im Vergleich mit den traditionellen Geheimhaltungsverfahren, allen voran der weit verbreiteten Zellsperre, weist die Cell-Key-Methode mit Blick auf die Abwägung zwischen Informationsverlust und Aufdeckungsrisiko einige Vorteile auf. Allerdings hat dieser Beitrag ausführlich dargestellt, dass die Datennutzenden sowie die Forschungsdatenzentren auch neue Besonderheiten bei der Auswertung und Interpretation der mit der Cell-Key-Methode geheim gehaltenen Ergebnisse berücksichtigen müssen.

Da die Cell-Key-Methode für die Geheimhaltung von Fallzahltabellen ausgelegt ist, kann es bei darüber hinausgehenden Analysen auf Basis von Verhältniszahlen, Zeitreihen und Saldierungen zu Problemen kommen. Diese sind vermeidbar, sofern Auswertungen immer auf Basis ausreichend großer Fallzahlen vorgenommen werden. Sollten bei der Nutzung in den Forschungsdatenzentren Probleme oder Unsicherheiten entstehen, können Nutzende sich jederzeit an ihren betreuenden FDZ-Standort wenden. 

LITERATURVERZEICHNIS

Bourier, Günther. *Beschreibende Statistik. Praxisorientierte Einführung – Mit Aufgaben und Lösungen*. Wiesbaden 2011, Seite 19 ff. DOI: [10.1007/978-3-8349-6556-1](https://doi.org/10.1007/978-3-8349-6556-1)

Enderle, Tobias/Giessing, Sarah/Tent, Reinhard. *Designing Confidentiality on the Fly Methodology – Three Aspects*. In: Domingo-Ferrer, Josep/Montes, Francisco (Herausgeber). *Privacy in Statistical Databases*. LNCS (Lecture Notes in Computer Science). 2018. Ausgabe 11126, Seite 28 ff. DOI: [10.1007/978-3-319-99771-1_3](https://doi.org/10.1007/978-3-319-99771-1_3)

Enderle, Tobias/Vollmar, Meike. *Geheimhaltung in der Hochschulstatistik*. In: WISTA Wirtschaft und Statistik. Ausgabe 6/2019, Seite 87 ff.

Fraser, Bruce/Wooton, Janice. *A proposed method for confidentialising tabular output to protect against differencing*. Work session on statistical data confidentiality. Supporting paper. Genf 2005. [Zugriff am 30. April 2024]. Verfügbar unter: unece.org

Giessing, Sarah. *What shall we do with the ratios?* Work session on statistical data confidentiality. Supporting paper. Ottawa 2013. [Zugriff am 7. Mai 2024]. Verfügbar unter: unece.org

Höhne, Jörg/Höninger, Julia. *Die Cell-Key-Methode – ein Geheimhaltungsverfahren*. In: Zeitschrift für amtliche Statistik Berlin Brandenburg. Ausgabe 3+4/2018, Seite 14 ff. [Zugriff am 30. April 2024]. Verfügbar unter: www.statistischebibliothek.de

Marley, Jennifer K./Leaver, Victoria L. *A Method for Confidentialising User-Defined Tables: Statistical Properties and a Risk-Utility Analysis*. In: Proceedings of 58th World Statistical Congress. 2011. [Zugriff am 30. April 2024]. Verfügbar unter: 2011.isiproceedings.org

Rohde, Johannes/Seifert, Christiane/Gießing, Sarah/Setzer, Stefanie (unter Mitarbeit von Breitenfeld, Jörg/Brings, Stefan/Höhne, Jörg/Höninger, Julia/Rothe, Patrick/Schedding-Kleis, Ulrike). *Entscheidungskriterien für die Auswahl eines Geheimhaltungsverfahrens. Version 1.1 vom 23.04.2021*. Internes Dokument des Statistischen Verbunds (Statistische Ämter des Bundes und der Länder).

Setzer, Stefanie/Rohde, Johannes/Güttgemanns, Volker/Rothe, Patrick. *Die Cell-Key-Methode in den Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder - Teil 1: Vorstellung des neuen Geheimhaltungsverfahrens*. In: WISTA Wirtschaft und Statistik. Ausgabe 3/2024, Seite 31 ff.

DER EUROPÄISCHE MIKRODATEN-AUSTAUSCH – NEUE DATENQUELLE FÜR DIE AUßENHANDELSSTATISTIK

Igor Franjić, Thomas Kolvenbach, Mingyong Tong

➤ **Schlüsselwörter:** Außenhandel – Intrahandel – Intrastat – EBS-Verordnung – Geheimhaltung

ZUSAMMENFASSUNG

Seit dem Berichtsmonat Januar 2022 tauschen die nationalen Statistikämter der EU-Mitgliedstaaten untereinander Mikrodaten zu den innereuropäischen Warenexporten aus. Dabei wurde erstmals in der Geschichte des Europäischen Statistischen Systems ein Mikrodatabaustausch zwischen den EU-Mitgliedstaaten rechtlich verankert und über eine zentrale Infrastruktur abgewickelt. Der Aufsatz beschreibt Inhalt und Umfang der ausgetauschten Daten und vergleicht sie mit den spiegelbildlichen Daten zu innereuropäischen Warenimporten, die das Statistische Bundesamt selbst erhebt. Dabei werden die Qualität und Nutzbarkeit der Mikrodaten in der deutschen Außenhandelsstatistik sowie die mit dem Austausch und der Nutzung der Daten verbundenen Herausforderungen diskutiert.

➤ **Keywords:** foreign trade – intra-EU trade – Intrastat – EBS regulation – confidentiality

ABSTRACT

The national statistical institutes of the EU Member States have been exchanging microdata on intra-EU exports of goods since the reference month of January 2022. This marked the first time in the history of the European Statistical System that microdata exchange between the Member States was legally established and processed through a centralised infrastructure. This article describes the scope and content of the exchanged data and compares them to the mirror data on intra-EU imports of goods collected by the Federal Statistical Office. The quality of the microdata and their usability for German foreign trade statistics is examined and the challenges involved in exchanging and using the data are discussed.

Igor Franjić

hat Betriebswirtschaftslehre (M.Sc.) an der Heinrich-Heine-Universität Düsseldorf studiert und arbeitet seit 2022 als Wissenschaftlicher Mitarbeiter im Referat „Grundsatzfragen, Qualitätssicherung, Verbreitung“ der Gruppe „Außenhandel“ des Statistischen Bundesamtes. Im Projekt „Mikrodatabaustausch“ analysiert er Qualität und Verwendungsmöglichkeiten der empfangenen Mikrodaten in der Außenhandelsstatistik.

Thomas Kolvenbach

hat Economics (M.Sc.) an der Universität zu Köln studiert. Seit 2022 ist er im Statistischen Bundesamt im Referat „Grundsatzfragen, Qualitätssicherung, Verbreitung“ der Gruppe „Außenhandel“ tätig. Er bereitet Steuerdaten zur Nutzung für die Außenhandelsstatistik auf und erstellt Analysen zur künftigen Nutzung von Mikrodaten.

Mingyong Tong

hat Politikwissenschaft (M.A.) an der Rheinischen Friedrich-Wilhelms-Universität Bonn studiert. Seit 2022 ist er im Statistischen Bundesamt tätig, zunächst im Referat „Grundsatzfragen, Qualitätssicherung, Verbreitung“ der Gruppe „Außenhandel“. Derzeit erstellt er im Referat „Mikrozensus – Auswertung und Analyse“ Sonderauswertungen zum Mikrozensus und Standardtabellen.

1

Einleitung

Seit 1993 der Europäische Binnenmarkt geschaffen wurde und dadurch die Zollanmeldungen weggefallen sind, werden die Daten zum Warenhandel innerhalb der Europäischen Union (EU) durch die Intrahandelsstatistik erhoben: Monatlich melden Unternehmen ihre innergemeinschaftlichen Importe und Exporte an die nationalen statistischen Ämter, in Deutschland an das Statistische Bundesamt. Obgleich etwa 92 % der außenhandelsaktiven Unternehmen durch die geltenden Anmeldeschwellen¹ von der Meldepflicht befreit sind, stellen die sogenannten Intrastat-Meldungen einen erheblichen Aufwand für die übrigen, meldepflichtigen Unternehmen dar (Schüßler und andere, 2024). Schon bald nach der Einführung von Intrastat gab es deswegen Bestrebungen, auf unterschiedlichen Wegen den Beantwortungsaufwand zu reduzieren.

Ausgangspunkt vieler Überlegungen war, dass die meisten innergemeinschaftlichen Warenbewegungen zweimal erhoben werden, nämlich durch eine Versendungsmeldung im Exportmitgliedstaat und eine Eingangsmeldung im Importmitgliedstaat. Diese Tatsache legt den Schluss nahe, dass Versendungs- und Eingangsmeldung prinzipiell redundant sein sollten und es genügt, nur eine der beiden Meldungen einzufordern und die Informationen anschließend zwischen den statistischen Ämtern auszutauschen. Der Austausch wurde im Projekt SIMSTAT (Single Market Statistics) vor einigen Jahren erfolgreich erprobt (Steinfelder, 2016).

Die Ergebnisse dieses Projektes fanden daraufhin Niederschlag in der European-Business-Statistics(EBS)-Verordnung, welche auch die Außenhandelsstatistik auf eine neue rechtliche Grundlage stellte (Herzog, 2020). Erstmals wurde auf europäischer Ebene rechtlich verankert, dass die nationalen statistischen Ämter ab Berichtsmonat 2022 statistische Mikrodaten verpflichtend austauschen. Gleichzeitig wurde die bisherige Verpflichtung, mindestens 93 % der innergemeinschaftlichen Importe selbst zu erheben (Artikel 10 Absatz 3 der Verordnung [EG] Nr. 638/2004), aufgehoben. Die

nationalen statistischen Ämter wurden ermächtigt, für die Importstatistik unterschiedliche Datenquellen unter Einhaltung definierter Qualitätsstandards flexibel zu nutzen. Damit verbindet sich die bereits in der [Vision 2020](#) des Europäischen Statistischen Systems (ESS) formulierte Hoffnung, Unternehmen von Meldepflichten zu entlasten, gleichzeitig aber die Qualität der Außenhandelsstatistik zu steigern (Erwägungsgrund 16 EBS-Verordnung).

Da die EU-Kommission anstrebt, den Austausch von Mikrodaten zwischen den nationalen statistischen Ämtern und dem Statistischen Amt der Europäischen Union (Eurostat) sowie untereinander auszuweiten (Absatz 7 Entwurf der Neufassung der Verordnung [EG] Nr. 223/2019), nimmt die Außenhandelsstatistik mit dem Mikrodatenaustausch eine Vorreiterrolle ein. Einige Punkte sprechen dafür, dass die ausgetauschten Daten in der Außenhandelsstatistik besonders einfach zu verwenden sein sollten:

- › Die Methodik der Außenhandelsstatistik ist weitgehend europäisch harmonisiert.
- › Der Merkmalskranz ist in allen Mitgliedstaaten im Wesentlichen der gleiche.
- › Der Erhebungsgegenstand ist in Form innergemeinschaftlicher Warenbewegungen vergleichsweise klar und einfach definiert.

In der Realität zeigt sich jedoch, dass es sowohl technisch als auch organisatorisch ein äußerst komplexes Unterfangen darstellt, einen solch umfassenden Austausch von Daten zu planen, zu implementieren und durchzuführen. Insbesondere die Nutzung der Daten ist mit zahlreichen Schwierigkeiten behaftet.

Ein Ziel dieses Artikels ist, am Beispiel der erhaltenen Außenhandelsdaten zu verdeutlichen, welchen Herausforderungen die Arbeit mit europäischen Mikrodaten und deren Auswertung gegenübersteht. Die hierbei gesammelten Erfahrungen und identifizierten Schwierigkeiten können auch bei künftigen Projekten in anderen Statistikbereichen helfen, konzeptionelle und praktische Probleme beim Datenaustausch und bei der Datennutzung bereits im Vorfeld zu identifizieren und zu beheben. Die folgenden Kapitel 2 und 3 geben zunächst einen Überblick über das System des Mikrodatenaustauschs und die übermittelten Daten sowie darüber, nach welchen Kriterien die Daten analysiert und bewertet wurden.

¹ Für das Berichtsjahr 2024 liegt die Anmeldeschwelle zur Feststellung der Auskunftspflicht für die Warenversendung bei 500 000 Euro und für den Wareneingang bei 800 000 Euro.

Anschließend beschreibt Kapitel 4 die wesentlichen Charakteristika der empfangenen Daten und vergleicht sie mit den spiegelbildlich vom Statistischen Bundesamt erhobenen Daten. Zuletzt gibt Kapitel 5 einen Ausblick auf noch offene Problemstellungen und mögliche weitere Verwendungsmöglichkeiten der ausgetauschten Daten.

2

Der Mikrodatabaustausch

Der Mikrodatabaustausch (micro data exchange – MDE) zwischen den nationalen statistischen Ämtern erfolgt über eine zentralisierte Infrastruktur. Jeder Mitgliedstaat schickt spätestens 30 Tage nach Ende des Berichtsmonats eine Datei, die sämtliche seit der vorherigen Übermittlung erhobenen Mikrodaten¹² über Warenexporte in die anderen EU-Mitgliedstaaten enthält, an einen von Eurostat betriebenen Knotenpunkt, den sogenannten MDE-Hub. Konkret enthält diese Datei die vorliegenden Mikrodaten für den aktuellen Berichtsmonat sowie Korrekturen, Nachmeldungen und Löschungen für vorangegangene Berichtszeiträume.

Spätestens 35 Tage nach Ende des Berichtsmonats verschickt jeder Mitgliedstaat dann zusätzlich eine Datei mit den dazugehörigen Metadaten (coverage metadata – MDC, deutsch etwa „Metadaten zum Abdeckungsgrad“) an den MDE-Hub. Dabei handelt es sich um die Aggregate der zuvor übermittelten MDE-Daten nach Bestimmungsland, Zweistellern des Warenverzeichnisses für die Außenhandelsstatistik und Berichtsmonat sowie die zugehörigen Zuschätzungen für Antwortausfälle und Warenexporte von Unternehmen unter der Anmeldeschwelle.¹³

Der Hub überprüft, ob die übermittelten Dateien technisch und formal korrekt sind und spaltet sie dann nach dem angegebenen Bestimmungsland in sogenannte Split Files auf, die an die jeweiligen nationalen statistischen Ämter versendet werden. So wird im Sinne des Datenschutzes und der Datenökonomie sichergestellt,

dass einzelne Mitgliedstaaten nur diejenigen Mikrodaten erhalten, die sie auch betreffen.¹⁴

Das Statistische Bundesamt erhält somit jeden Monat von allen EU-Mitgliedstaaten und Nordirland¹⁵ jeweils die Mikrodaten über die Warenexporte mit Bestimmungsland Deutschland (MDE-Daten) und aggregierte Angaben über die Mikrodaten mit den zugehörigen

- 4 Auf demselben Weg werden jeden Monat Zollnoten über bestimmte Warenverkehre mit Drittländern zwischen den EU-Mitgliedstaaten (CDE-Daten) ausgetauscht; diese sind nicht Gegenstand der folgenden Untersuchungen und Vergleiche.
- 5 Trotz des Ausscheidens des Vereinigten Königreiches aus der Europäischen Union 2020 ist Nordirland unter dem Nordirland-Protokoll weiterhin Teil des Europäischen Binnenmarktes für Güter und nimmt deshalb am Mikrodatabaustausch teil. Ist im Folgenden von (Mitglied-) Staaten die Rede, sind damit die EU-Mitgliedstaaten und Nordirland gemeint.

Übersicht 1

Die wichtigsten in den ausgetauschten MDE- und MDC-Daten enthaltenen Merkmale

Meldemerkmal	Erläuterung	MDE	MDC
URI	Eindeutige Meldungs-Identifikationsnummer (ID) des Mikrodatabasatzes	x	
TIME_PROD	Zeitstempel	x	
RECORD_ACTION	Meldungsart	x	
DATA_SOURCE	Datenquelle	x	x
PARTNER_MS	Bestimmungsland	x	x
REF_PERIOD	Berichtsmonat	x	x
FLOW	Verkehrsrichtung	x	x
PARTNER_ID	Umsatzsteuer-Identifikationsnummer (USt-ID) des Handelspartners (in der Regel des Importeurs) im Bestimmungsland	x	
COMMODITY	Achtstellige Warennummer	x	x
COUNTRY_ORIGIN	Ursprungsland	x	x
NATURE_TRANS	Art des Geschäfts (Kaufgeschäft, Lohnveredelung, Leasing und so weiter)	x	x
MODE_TRANSPORT	Verkehrszweig	x	x
DELIVERY_TERMS	Lieferbedingungen	x	x
QTY_NET_MASS	Eigenmasse in Kilogramm	x	x
QTY_SU	Besondere Maßeinheit (beispielsweise Stückzahl)	x	x
TAX_AMOUNT	Rechnungswert	x	x
STAT_VAL	Statistischer Wert	x	x
RECORD_CONF_STATUS	Geheimhaltungskennzeichnung	x	

MDE: Mikrodatabaustausch; MDC: Mikrodaten mit Zuschätzungen.

12 Produkt X wird aus Land A nach Land B importiert und geht dort an Unternehmen Z.

13 Eine detailliertere Aggregation ist zulässig, aber nicht verpflichtend, da die Zuschätzungen in den einzelnen Mitgliedstaaten eine unterschiedliche Detailtiefe aufweisen.

Zuschätzungen (MDC-Daten). Damit ein sinnvoller Vergleich der MDE-Daten mit den deutschen Eingangsdaten möglich ist, sind ihnen die in den MDC-Daten enthaltenen Zuschätzungen hinzuzufügen; diese Kombination aus MDE-Daten mit den Zuschätzungen aus den MDC-Daten wird im Folgenden kurz als EU-Daten bezeichnet. Die wichtigsten in den Dateien enthaltenen Merkmale zeigt [Übersicht 1](#).

3

Qualitäts- und Nutzbarkeitskriterien

Um als zusätzliche Datenquelle für die Außenhandelsstatistik dienen zu können, müssen die von den EU-Mitgliedstaaten übermittelten Daten eine Reihe formaler wie inhaltlicher Anforderungen erfüllen:

- › **Pünktlichkeit:** Zunächst ist es wichtig, dass die MDE- und MDC-Daten pünktlich zu den gesetzlich festgeschriebenen Lieferterminen vorliegen, da das Statistische Bundesamt die Ergebnisse des deutschen Außenhandels im jeweiligen Berichtsmonat bereits wenige Tage nach dem Liefertermin veröffentlicht.
- › **Vollständigkeit:** Darüber hinaus ist bedeutend, wie vollständig die Daten zu diesem Zeitpunkt sind, also in welchem Umfang folgende MDE-Lieferungen Daten des aktuellen Berichtsmonats nachmelden oder korrigieren.
- › **Konsistenz:** Außerdem ist es notwendig, dass die MDE- und MDC-Daten inhaltlich konsistent sind, da sonst keine korrekte Zuordnung der Zuschätzungen zu den Mikrodaten erfolgen kann. Vor allem müssen die gelieferten MDE-Daten mit den spiegelbildlich in Deutschland erhobenen Intrastat-Daten im Wesentlichen übereinstimmen, um bei Verwendung der MDE-Daten in der deutschen Importstatistik die Vergleichbarkeit mit früheren Berichtsperioden zu gewährleisten. Die Übereinstimmung ist auf drei verschiedenen Ebenen nötig:
 - formal: die MDE- und MDC-Daten enthalten die gleichen Erhebungsmerkmale,
 - methodisch: sie bilden methodisch den gleichen Sachverhalt ab wie die spiegelbildlichen Intrastat-Eingangsdaten,

- inhaltlich: die Ergebnisse passen sowohl bezüglich des Wertes (insbesondere des statistischen Wertes) als auch bezüglich der erfassten Partnerländer und Warenkategorien zusammen.

- › **Geheimhaltung:** Die Mikrodaten dürfen vom versendenden Mitgliedstaat nicht in solchem Ausmaß als geheimzuhalten gekennzeichnet sein, dass der Importmitgliedstaat sie in seiner Veröffentlichung großflächig sperren muss.

Die empirischen Erkenntnisse zu diesen Kriterien, die durch den Vergleich der empfangenen MDE- und MDC-Daten mit den deutschen Intrastat-Eingangsdaten gewonnen wurden, werden im Folgenden vorgestellt.

4

Vergleich mit den Intrastat-Eingangsdaten

4.1 Pünktlichkeit und Vollständigkeit

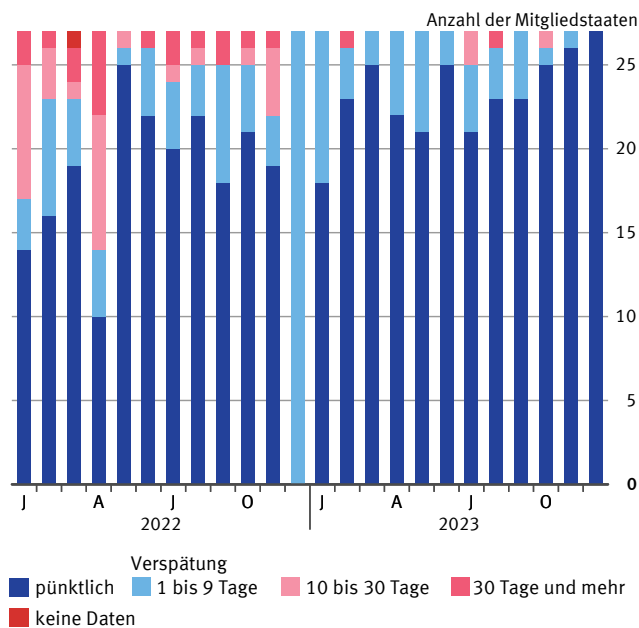
Besonders zu Beginn des Mikrodatenaustauschs im Jahr 2022 waren die Datenlieferungen vieler Mitgliedstaaten zunächst sehr unpünktlich. Für den Berichtsmonat April 2022 beispielsweise schickten nur zehn Staaten pünktlich MDE-Daten, während aus fünf Staaten die MDE-Daten mit mehr als 30 Tagen Verzögerung eingingen. Ähnliche Probleme gab es auch bei der Pünktlichkeit der MDC-Datenlieferungen. Während des Jahres 2023 verbesserte sich die Pünktlichkeit deutlich, sodass für Berichtsmonat Dezember 2023 erstmals MDE-Daten aus allen Mitgliedstaaten innerhalb der Frist von 30 Tagen nach Ende des Berichtsmonats vorlagen. Dennoch gibt es immer wieder Verzögerungen, die teilweise auch technischen Problemen geschuldet sind. So wurden für den Berichtsmonat Dezember 2022 die MDE-Daten aller 27 Mitgliedstaaten wegen eines Fehlers im MDE-Hub erst nach der Frist an die Partnerländer übermittelt.

➤ [Grafik 1](#)

Dabei sind das Revisionsverhalten und damit die Vollständigkeit der MDE-Daten zu einem frühen Zeitpunkt von Land zu Land unterschiedlich. So liegen aus Frankreich und Polen schon 30 Tage nach Ende des Berichtsmonats stets über 90% des Gesamtwertes aller MDE-

Grafik 1

Pünktlichkeit der ersten MDE-Datenlieferung für einen Berichtsmonat, gemessen am Zeitpunkt des Dateneingangs im Statistischen Bundesamt



Daten vor, die überhaupt von dem entsprechenden Land für den jeweiligen Berichtsmonat übermittelt werden. Mit der ersten Revision folgen die restlichen Meldungen, während danach kaum noch korrigiert und nachgemeldet wird. Österreich hingegen stellt zur ersten Datenlieferung typischerweise erst 30% des Gesamtwertes bereit; der Anteil an Zuschätzungen in den MDC-Daten ist zu diesem Zeitpunkt entsprechend hoch. [Tabelle 1](#)

4.2 Formale und methodische Übereinstimmung

Trotz der weitgehenden methodischen Harmonisierung der Außenhandelsstatistik auf europäischer Ebene sind die Erhebungsmerkmale weiterhin nicht vollständig harmonisiert. So handelt es sich beim Verkehrszweig an der Grenze und den Lieferbedingungen im Mikrodatabaustausch um fakultative Merkmale, die nur von etwa der Hälfte der Mitgliedstaaten erfragt werden.¹⁶ Zudem wird der statistische Wert der Waren nicht in allen Ländern erhoben, sondern zum Teil aufgrund fester Quoten aus dem Rechnungswert errechnet. Auch das Aggregationsniveau der MDC-Daten, insbesondere der darin enthaltenen Zuschätzungen, ist unterschiedlich: So schätzen einige Mitgliedstaaten nur auf Ebene der Warenkapitel, andere auf Warennummer-Ebene und wiederum andere auf Ursprungsland-Kapitelebene zu.

Problematisch in Bezug auf die deutsche Außenhandelsstatistik ist zudem, dass das Bestimmungsbundesland in den anderen Mitgliedstaaten nicht erhoben wird und somit, anders als in den deutschen Spiegeldaten, in den MDE-Daten auch nicht zur Verfügung steht. Hier zeigt sich die prinzipielle Schwierigkeit, nationale Datenbedarfe aus einer europäischen Datenquelle zu decken. Ersatzweise lässt sich als Indiz das Sitz-Bundesland des in der MDE-Meldung angegebenen Handelspartners heranziehen. Wie später im Text deutlich wird, handelt es sich dabei aber nicht um eine gleichwertige Information.

Eine wichtige Grundannahme hinter der Idee des Mikrodatabaustausches war, dass die Versendungsmeldung im Exportmitgliedstaat und die spiegelbildliche Eingangsmeldung im Importmitgliedstaat zu einer Waren-

¹⁶ In der deutschen Intrahandelsstatistik ist der Verkehrszweig an der Grenze ein Pflichtmerkmal in beiden Verkehrsrichtungen, während die Lieferbedingungen nicht erhoben werden.

Tabelle 1

Zum jeweiligen Zeitpunkt vorliegende Mikrodaten eines Berichtsmonats für ausgewählte Versendungsländer

	Österreich			Frankreich			Polen		
	September 2023	Oktober 2023	November 2023	September 2023	Oktober 2023	November 2023	September 2023	Oktober 2023	November 2023
	am statistischen Wert gemessener Anteil in %								
nach 30 Tagen	0	29	27	96	93	96	92	92	100
nach 60 Tagen	95	97	100	100	99	100	99	100	100
nach 90 Tagen	100	100	100	100	100	100	100	100	100

bewegung konzeptionell den gleichen Sachverhalt erfassen und entsprechend austauschbar sein sollten. Die beiden Meldungen sollten also nicht nur formal, sondern auch methodisch übereinstimmen. In dieser Denkweise sind Asymmetrien zwischen den spiegelbildlichen Statistiken immer falsch. Bei genauer Betrachtung ist dies aus mehreren Gründen jedoch nicht zwangsläufig der Fall:

- › Erstens ist es möglich, dass die Wertstellung der Ware für die Exportstatistik des einen Mitgliedstaats richtigerweise von der Wertstellung der Ware für die Importstatistik des anderen Mitgliedstaats abweicht. Nach den internationalen Richtlinien ist jeweils der Wert der Ware an der Grenze des betreffenden Landes zu erheben. Findet beispielsweise eine innergemeinschaftliche Warenbewegung im Rahmen eines Dreiecksgeschäfts statt – das heißt mit einem Zwischenhändler in einem dritten Mitgliedstaat –, so ist der Wert der Ware an der Grenze des Importmitgliedstaats mindestens um die Gewinnspanne des Zwischenhändlers höher als an der Grenze des Exportmitgliedstaats. Gerade Dreiecksgeschäfte sind innerhalb der EU indes weit verbreitet.
- › Zweitens ist die Außenhandelsstatistik als Monatsstatistik vergleichsweise hoch frequent und so kann es vorkommen, dass die Warenbewegung richtigerweise in verschiedenen Berichtsperioden ausgewiesen wird. Wird eine Ware beispielsweise Ende März von Portugal nach Deutschland verschifft, kommt sie dort aller Wahrscheinlichkeit nach erst Anfang April an.
- › Drittens können Asymmetrien, zumindest auf detaillierter Warenebene, auch durch die Meldeschwellen im Intrahandel entstehen, ohne dass in irgendeiner Form ein Anmeldefehler vorliegt. Der Exporteur einer Ware kann im Exportmitgliedstaat meldepflichtig sein, da seine innergemeinschaftlichen Umsätze den Schwellenwert überschreiten. Gleichzeitig kann der Importeur im Importmitgliedstaat nicht meldepflichtig sein, da er die entsprechende Meldeschwelle unterschreitet. Umgekehrt ist es möglich, dass der Importeur im Importmitgliedstaat meldepflichtig ist, während dies zurzeit nicht für alle Exporteure, die ihn beliefern, in den entsprechenden Exportmitgliedstaaten gilt. Dieses Problem verstärken die unterschiedlichen Detailtiefen der Zuschätzungen zwischen den Mitgliedstaaten, ebenso die Tatsache, dass die Anmeldeschwellen europaweit nicht einheitlich sind. Sie schwanken viel

mehr durch die methodischen Vorgaben je nach Mitgliedstaat zwischen wenigen Tausend Euro und mehreren Millionen Euro.

Es zeigt sich also, dass Asymmetrien entweder methodisch bedingt sind oder im Einzelfall schwer aufzudecken und konzeptionell nicht aufzulösen sind.

4.3 Inhaltliche Übereinstimmung

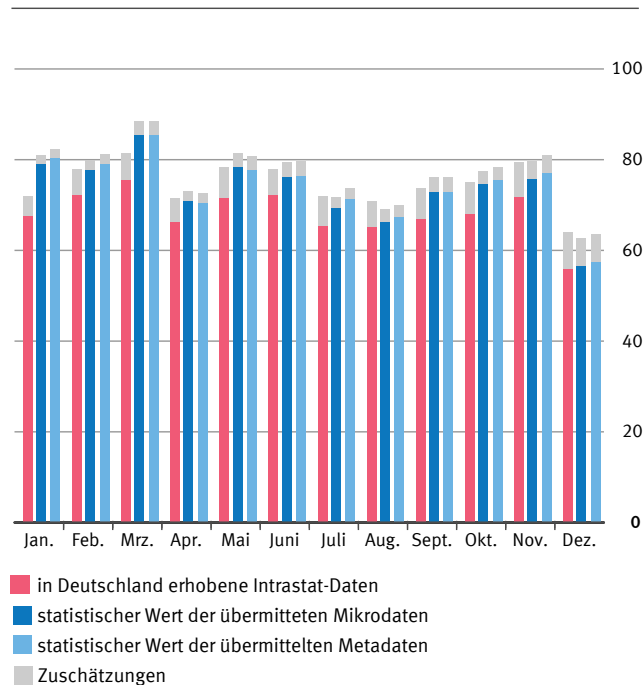
Auf hoch aggregierter Ebene weichen die EU-Daten nur moderat von den deutschen Eingangsdaten ab. Im Monatsdurchschnitt des Jahres 2023 lag der Gesamtwert der EU-Daten um 2,9 % höher als die erhobenen Eingangsdaten. Die prozentuale Differenz bewegt sich zwischen – 3 % und + 13 %.¹⁷ Es fällt allerdings schon auf dieser Ebene auf, dass die Gesamtsumme der MDE-Daten laut MDC-Datei zum Teil von der Gesamtsumme der MDE-Daten laut MDE-Datei abweicht. Diese Inkonsistenz, für die vermutlich die abweichenden Lieferfristen (t+30 für die MDE-Datei, t+35 für die MDC-Datei) verantwortlich sind, stellt ein großes Problem bei der korrekten Zuordnung von Zuschätzungen zu Mikrodaten dar und ist bis heute nicht von allen Mitgliedstaaten behoben worden. Solange dieses Problem nicht gelöst ist, lässt sich die Gesamtsumme der innergemeinschaftlichen Importe Deutschlands aus den EU-Daten nicht berechnen. Somit sind die Daten nur stark eingeschränkt verwendbar. ➤ Grafik 2

Verlässt man die Ebene des Gesamtwerts, zeigen sich auf Ursprungslandebene im Jahresüberblick 2023 bereits deutliche Unterschiede zwischen EU-Daten und deutschen Eingangsdaten. Das Ursprungsland ist die wesentliche Partnerlandangabe auf der Importseite und gibt das Land an, in dem die importierte Ware hergestellt oder – bei grenzüberschreitenden Verflechtungen in der Warenproduktion – zum letzten Mal vor dem Import wesentlich verändert worden ist. Es zeigt sich, dass in den EU-Daten auf viele europäische Ursprungs-

¹⁷ Im folgenden Text ist die absolute Abweichung definiert als die Summe des statistischen Wertes der EU-Daten minus der Summe des statistischen Wertes der deutschen Eingangsdaten. Die prozentuale Abweichung ist die absolute Abweichung geteilt durch die Summe des statistischen Wertes der deutschen Eingangsdaten multipliziert mit Hundert. Somit sind die absolute und die prozentuale Abweichung **positiv**, wenn der statistische Wert in den EU-Daten **höher** als in den deutschen Eingangsdaten ist, und **negativ**, wenn der statistische Wert in den EU-Daten **niedriger** als in den deutschen Eingangsdaten ist.

Grafik 2

Statistischer Wert nach Berichtsmonat und Datenquelle für das Berichtsjahr 2023
Mrd. EUR



länder ein niedrigerer statistischer Wert als in den deutschen Eingangsdaten entfällt. Dagegen wird umgekehrt für manche außereuropäischen Ursprungsländer wie China oder Indien ein höherer statistischer Wert als in den deutschen Eingangsdaten ausgewiesen. Möglicherweise geben die deutschen Importeure mangels genauer Kenntnis des Ursprungslands den Mitgliedstaat an, aus dem die Ware nach Deutschland kommt, während die dortigen Exporteure unmittelbar Kontakt mit dem Lieferanten im Drittland haben und deshalb das Ursprungsland besser kennen. [➤ Grafik 3](#)

Problematisch ist in diesem Zusammenhang, dass in den MDE-Daten das Ursprungsland in relevantem Umfang als unbekannt gekennzeichnet ist: Insgesamt ist im Berichtsjahr 2023 das Ursprungsland für 3 % des in den MDE-Daten gemeldeten Wertes unbekannt. Dabei sticht mit 7 % des Importwerts das Kapitel 27 (Brennstoffe) heraus, in dem bei 5 % der Pipeline-Erdgas-Importe und sogar bei 54 % des importierten Stroms das Ursprungsland fehlt.

In den Detailergebnissen zeigen sich die Asymmetrien noch deutlicher. [➤ Tabelle 2](#) vergleicht die EU-Daten

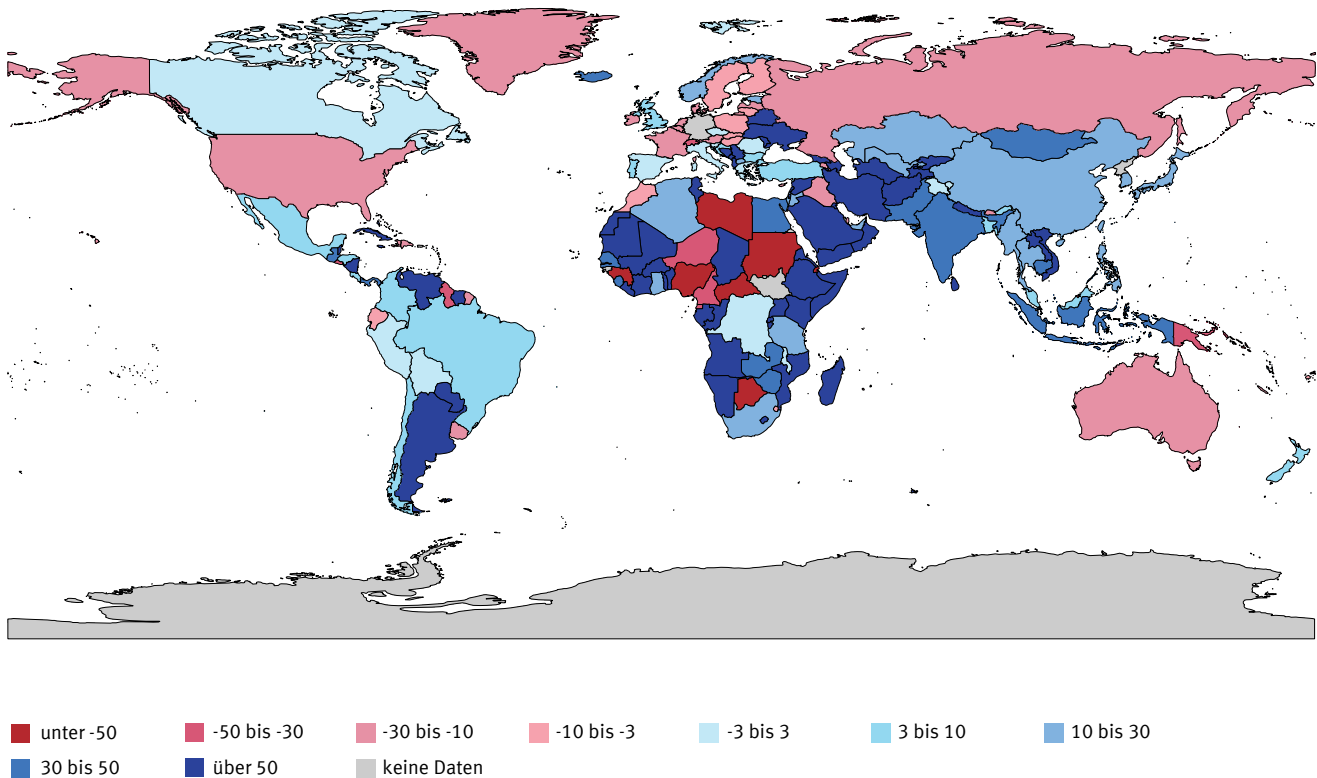
und die deutschen Eingangsdaten auf der Ebene der bedeutendsten Versandungsländer und der wertmäßig größten Kapitel des Warenverzeichnisses für die Außenhandelsstatistik in den Berichtsjahren 2022 und 2023. Aus dem Versandungsland sind die Waren mit Bestimmung Deutschland versandt worden, es entspricht somit dem Mitgliedstaat, der die jeweiligen MDE-Daten an das Statistische Bundesamt übermittelt. Dementsprechend handelt es sich gerade hierbei um spiegelbildliche Daten zum selben Sachverhalt, von denen anzunehmen wäre, dass sie sich weitgehend entsprechen. Jedoch weisen sie zum Teil massive und über die Zeit volatile Abweichungen in beide Richtungen auf.

Vor allem die Asymmetrien in Kapitel 27 (Brennstoffe) fallen auf: Laut EU-Daten waren die Importe aus Belgien und Italien im Berichtsjahr 2022 deutlich höher als nach den deutschen Eingangsdaten. Die Importe aus den restlichen Versandungsländern fielen gemäß EU-Daten nur halb so groß aus wie aus den deutschen Eingangsdaten zu erwarten wäre. Für das Berichtsjahr 2023 sind diese Abweichungen zwar geringer, aber weiterhin vorhanden. Dagegen sind den EU-Daten zufolge in beiden Jahren aus fast allen wichtigen Partnerländern weniger Arzneimittel (Kapitel 30) importiert worden als die deutschen Eingangsdaten belegen. Ein Großteil der Gesamtabweichung in diesem Kapitel geht auf die Niederlande zurück. Die Gründe für diese beträchtlichen Abweichungen lassen sich nur zum Teil aufklären:

So enthält das Kapitel 27 den komplexen Handel von Strom und Pipeline-Gas über feste Übertragungsnetze. Die korrekte und konsistente Erfassung dieser Warenbewegungen ist besonders kompliziert. Auch in Kapitel 87 (Beförderungsmittel) treten erhebliche Asymmetrien auf, obwohl die Waren dieses Kapitels konzeptionell vergleichsweise einfach zu erfassen sind. Dabei wurden persistente Asymmetrien bei den Beförderungsmitteln auch von den Statistikämtern anderer EU-Mitgliedstaaten festgestellt, die sich trotz intensiver Nachforschungen bei den größten Meldeeinheiten nicht vollständig auflösen ließen: Einerseits meldeten die Unternehmen in Deutschland andere als die in den versendenden Mitgliedstaaten angemeldeten Warenmengen. Andererseits wichen sowohl die verwendeten Warennummern als auch der Warenwert voneinander ab, obwohl es sich rein konzeptionell um dieselben Warenbewegungen handelte. Bestätigen die Meldeeinheiten jedoch auf beiden Seiten die Richtigkeit ihrer Angaben, sind der

Grafik 3

Abweichungen zwischen EU-Daten und deutschen Eingangsdaten nach Ursprungsland 2023
in %



© EuroGeographics bezüglich der Verwaltungsgrenzen

tatsächlich richtige Wert oder die tatsächlich richtige Warennummer in der Regel nicht verifizierbar.

Auch im Kapitel 30 waren solche Phänomene zu beobachten. So kann als Importeur in den MDE-Daten sowohl eine Konzernmutter als auch eine Tochtergesellschaft angegeben sein, und gleichzeitig kann es sowohl auf deutscher Seite als auch auf Seiten des Exportmitgliedstaats zu Meldeausfällen gekommen sein. Daher sind die entstehenden Differenzen auch auf Mikrodatenebene äußerst schwierig nachzuvollziehen und lassen sich – wenn überhaupt – nur im aufwendigen Kontakt mit den beteiligten Unternehmen und den Daten versendenden Mitgliedstaaten klären.

Hinzu kommt, wie oben bereits beschrieben, dass Asymmetrien konzeptionell durchaus richtig sein können. Mithin zeigt der Vergleich der EU-Daten mit den

deutschen Eingangsdaten auf detaillierter Warennummer-Land-Ebene, dass spiegelbildliche Ergebnisse sehr unterschiedlich ausfallen können und die Gründe hierfür oft nur schwer zu ermitteln sind. Dies schränkt die Nutzbarkeit der EU-Daten für die deutsche Importstatistik wesentlich ein (siehe Tabelle 2).

Ein weiteres, jedoch anders gelagertes Problem stellt die Importstatistik nach Bundesländern dar. Wie oben bereits beschrieben, enthalten die EU-Daten keine Informationen über das Bestimmungsbundesland der nach Deutschland gelieferten Ware. Zugleich ist die Zuordnung der MDE-Daten zu Bestimmungsbundesländern auf Grundlage des Sitzbundeslandes des in den Daten angegebenen Importeurs methodisch mit großen Einschränkungen verbunden. Der Vergleich der MDE-Daten mit den deutschen Eingangsdaten zeigt, dass dadurch

Tabelle 2

Abweichungen zwischen EU-Daten und deutschen Eingangsdaten für die fünf wertmäßig größten Warenkapitel und Versandungsländer

	Kapitelnummer des Warenverzeichnisses für die Außenhandelsstatistik					Sonstige Warenkapitel	Abwei- chungen insgesamt
	27 ¹	30 ²	84 ³	85 ⁴	87 ⁵		
	%						
Berichtsjahr 2022							
Belgien	+ 264,66	– 16,89	– 16,39	+ 0,79	– 16,50	– 13,08	+ 9,78
Frankreich	+ 63,81	– 4,50	+ 39,55	+ 18,33	+ 12,23	– 0,27	+ 8,12
Italien	+ 268,48	– 4,23	+ 5,84	– 6,17	+ 2,00	+ 3,40	+ 3,61
Niederlande	+ 29,92	– 25,14	+ 8,09	+ 7,67	+ 8,68	+ 10,49	+ 12,19
Polen	– 10,42	– 0,87	– 10,70	+ 2,43	– 3,50	+ 1,28	– 0,73
Sonstige Versen- dungsländer	– 46,39	– 8,23	– 8,65	– 2,67	+ 23,51	– 6,95	– 3,69
Insgesamt	+ 39,51	– 13,33	– 0,76	+ 1,75	+ 13,58	– 1,39	+ 3,24
Berichtsjahr 2023							
Belgien	+ 91,54	– 3,99	– 12,07	+ 13,12	– 13,20	– 6,32	+ 2,49
Frankreich	+ 58,39	– 6,63	+ 33,38	+ 15,84	– 0,71	– 1,15	+ 5,52
Italien	+ 73,98	+ 0,69	+ 5,52	– 4,52	– 5,53	+ 2,21	+ 1,29
Niederlande	+ 18,00	– 24,17	+ 3,77	+ 0,91	– 2,24	+ 7,64	+ 5,87
Polen	– 10,63	– 1,14	– 6,64	+ 11,70	+ 4,39	+ 4,28	+ 3,66
Sonstige Versen- dungsländer	– 20,28	– 3,88	– 0,05	+ 1,84	+ 9,27	– 0,27	+ 0,98
Insgesamt	+ 22,85	– 8,66	+ 2,61	+ 3,59	+ 3,54	+ 1,39	+ 2,91

1 Mineralische Brennstoffe; Mineralöle und Erzeugnisse ihrer Destillation; bituminöse Stoffe; Mineralwächse.

2 Pharmazeutische Erzeugnisse.

3 Kernreaktoren, Kessel, Maschinen, Apparate und mechanische Geräte; Teile davon.

4 Elektrische Maschinen, Apparate, Geräte und andere elektrotechnische Waren, Teile davon; Tonaufnahme- oder Tonwiedergabegeräte, Bild- und Tonaufzeichnungs- oder -wiedergabegeräte, für das Fernsehen, Teile und Zubehör für diese Geräte.

5 Zugmaschinen, Kraftwagen, Krafträder, Fahrräder und andere nicht schienengebundene Landfahrzeuge, Teile davon und Zubehör.

weit höhere Anteile der deutschen Importe – gemessen an den MDE-Daten – den Stadtstaaten Berlin und Hamburg sowie Hessen zugeordnet werden, als nach den deutschen Eingangsdaten zu erwarten wäre. Grund dafür ist vermutlich, dass sich die Unternehmenssitze auf einige wenige Städte konzentrieren, ohne dass dies mit den tatsächlichen innerdeutschen Warenströmen korrespondiert. Außerdem lassen sich einerseits die MDE-Daten, bei denen der Importeur als unbekannt gekennzeichnet ist, und andererseits die Zuschätzungen aus den MDC-Daten, welche grundsätzlich nicht auf Unternehmensebene vorliegen, auf diese Weise prinzipiell nicht bestimmten Bundesländern zuordnen. Sie müssen pauschal verteilt werden. [Grafik 4](#)

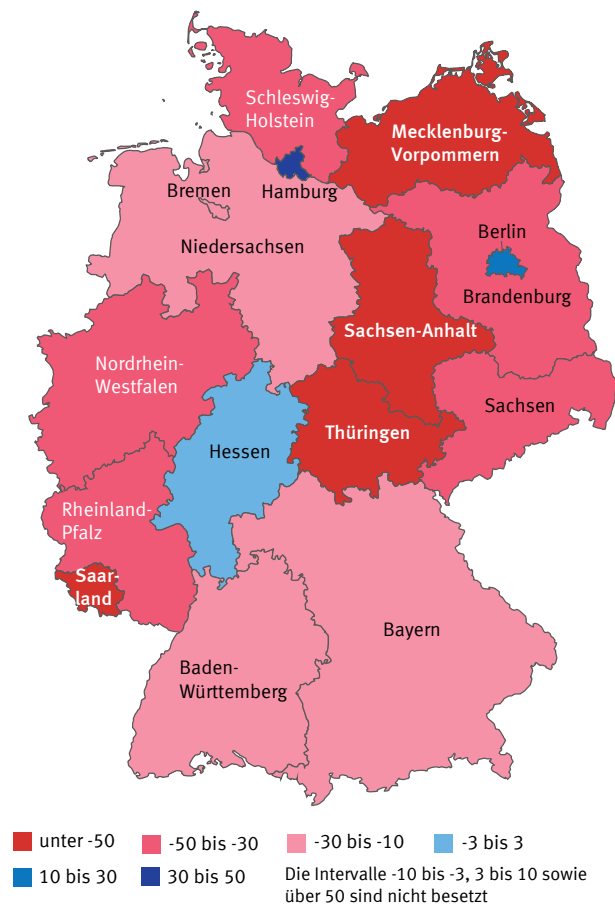
Jenseits der Bundesländerproblematik sind die in den MDE-Daten enthaltenen Umsatzsteuer-Identifikationsnummern des deutschen Handelspartners, das heißt in der Regel des Importeurs, das einzige Mittel, die MDE-Daten auf Unternehmensebene den deutschen Ein-

gangsdaten zuzuordnen. Dabei ist die Verfügbarkeit dieses Merkmals zum Teil noch unbefriedigend: Für jeden Berichtsmonat des Jahres 2023 entfallen etwa 5 bis 7 % des statistischen Wertes in den MDE-Daten auf Meldungen, bei denen der Handelspartner als unbekannt gekennzeichnet ist. Für einige Partnerländer liegt dieser Wert sogar über 10 %. In einem bedeutenden Anteil dieser Fälle kann die fehlende Angabe gar nicht erhoben werden, weil dem Versender der tatsächliche Warenempfänger nicht bekannt ist (beispielsweise bei Dreiecksgeschäften) oder weil dieser keine Umsatzsteuer-Identifikationsnummer besitzt (zum Beispiel im Falle von Privatpersonen).

Über das gesamte Berichtsjahr 2023 hinweg finden sich 96 % der deutschen Unternehmen, die der Intrahandelsstatistik Eingänge melden, als Handelspartner in den MDE-Daten. Auf diese Unternehmen entfallen 98 % des im Intrastat-Eingang gemeldeten statistischen Wertes. Die zugehörigen MDE-Daten machen aber um-

Grafik 4

Abweichung des Gesamtwerts zwischen den EU-Daten und den deutschen Eingangsdaten (jeweils ohne Zuschätzungen) auf Bundesländerebene im Berichtsjahr 2023 in %



© GeoBasis-DE / BKG 2022

gekehrt nur 85 % des im Mikrodatabaustausch insgesamt gemeldeten Wertes aus. Gleichzeitig handelt es sich dabei nur um 5 % aller im Mikrodatabaustausch übermittelten Handelspartner. Die restlichen 95 % der MDE-Handelspartner entsprechen im Wesentlichen Unternehmen in Deutschland, die sich unter der Anmeldeschwelle bewegen und deshalb nicht erfasst werden. Die MDE-Daten zu diesen Unternehmen können zwar keinen Intrastat-Eingangsdaten zugeordnet werden, bilden aber eine wertvolle zusätzliche Informationsquelle, die für detailliertere Zuschätzungen für Warenverkehre unter der Intrastat-Meldeschwelle nutzbar gemacht werden könnte. Die MDE-Daten enthalten

auch die Warenexporte an – in der Intrahandelsstatistik grundsätzlich nicht meldepflichtige – Privatpersonen und erlauben somit eine ungefähre Abschätzung des von Privatpersonen aus dem EU-Ausland bezogenen Waren. Daneben kann eine Zuordnung daran scheitern, dass dem exportierenden Unternehmen, das die Intrastat-Versendungsmeldung im Exportmitgliedstaat abgibt, der tatsächliche Warenempfänger unbekannt ist und es stattdessen die Umsatzsteuer-Identifikationsnummer des Rechnungsempfängers oder einen zugelassenen Platzhalter-Code angibt. Auch bei Importgeschäften innerhalb verbundener Unternehmen können Zuordnungsprobleme auftreten, beispielsweise wenn der Exporteur die Ware an einen deutschen Konzern verkauft und die Umsatzsteuer-Identifikationsnummer der Muttergesellschaft als Handelspartner angibt, die Ware aber an eine Tochtergesellschaft verschickt, welche die Intrastat-Meldung in Deutschland abgibt.

Trotz der genannten Probleme ist eine verlässliche Zuordnung der MDE-Daten auf Unternehmensebene äußerst wichtig. In einem Szenario, in dem die EU-Daten die vom Statistischen Bundesamt erhobenen Daten nicht komplett ersetzen, sondern ergänzen sollen, können Daten zu einer erfassten Warenbewegung in einer oder beiden Datenquellen erscheinen. Wenn es möglich ist, MDE-Daten deutschen Unternehmen zuverlässig zuzuordnen, könnten sie zur Plausibilisierung von Daten, die dieses Unternehmen selbst gemeldet hat, herangezogen werden. Außerdem wäre es so realisierbar, Warenverkehre unter der Anmeldeschwelle auf der Ebene einzelner Unternehmen, die keine Meldungen abgeben, zuzuschätzen.

4.4 Geheimhaltung

Die statistische Geheimhaltung stellt im Kontext des Mikrodatabaustausches ein ebenso bedeutendes wie komplexes Thema dar. Besonders tief greift es in einem sogenannten Einstrom-Szenario, in dem die Mitgliedstaaten die Erhebung der innergemeinschaftlichen Importe vollständig einstellen und entsprechend die Importstatistik im Intrahandel allein auf Grundlage der erhaltenen MDE-Daten erstellen. In einem solchen Fall entspricht die Importstatistik nach Versendungsländern im Importmitgliedstaat tatsächlich spiegelbildlich der Exportstatistik des Exportmitgliedstaats. Hält

beispielsweise der Exportmitgliedstaat A die Exporte in Warennummer x geheim, muss der Importmitgliedstaat B die Importe aus A in Warennummer x ebenfalls geheim halten. Ansonsten wäre der geheim zu haltende Wert unmittelbar in der Statistik des Importmitgliedstaates B ersichtlich. Gleichzeitig muss B die Importe aus A in einer anderen Warennummer y ebenfalls geheim halten, da B auch die Gesamtsumme der Importe aus A veröffentlicht. Um wiederum diese Geheimhaltung abzusichern, muss auch A die Exporte in Warennummer y nach B geheim halten und gleichzeitig die Exporte in Warennummer y nach einem dritten Mitgliedstaat C, um eine Rückrechnung über die innereuropäische Gesamtsumme der Exporte zu verhindern. Diese Kaskade lässt sich unendlich fortsetzen und verdeutlicht, dass die Geheimhaltung in einem Einstromverfahren ein ungemein komplexes, internationales Koordinationsproblem darstellt.


Tatsächlich zeigt sich, dass bei Umsetzung eines Einstromverfahrens in Deutschland schon allein durch die primären Geheimhaltungswünsche der Exportmitgliedstaaten – das heißt die erste Runde der beschriebenen Kaskade – viel mehr Kombinationen aus Warennummern und Partnerländern von Geheimhaltung betroffen wären, als dies im Status quo der Fall ist. Hätten die MDE-Daten die deutschen Eingangsdaten in den Berichtsjahren 2022 und 2023 vollständig ersetzt, hätte sich die Anzahl an gesperrten Kombinationen allein dadurch schon fast vervierfacht. Dabei wären auch wichtige Warengruppen wie Brennstoffe oder Arzneimittel von Sperrungen betroffen gewesen: 2022 hätten beispielsweise die Daten der Mineralölimporte komplett gesperrt werden müssen. Hinzu kommt, dass durch die Volatilität und hohe Anzahl der im laufenden Jahr auftretenden, zusätzlich geheim zu haltenden Kombinationen eine Geheimhaltung, wie sie derzeit in der deutschen Außenhandelsstatistik praktiziert wird und die auf größtmögliche Verfügbarkeit und Vollständigkeit der veröffentlichten Ergebnisse abzielt, nicht mehr durchführbar wäre.

5

Fazit

Die Auswertung der Berichtsjahre 2022 und 2023 zeigt, dass die Daten, die das Statistische Bundesamt im Rahmen des Mikrodatabaustauschs der Außenhandelsstatistik von anderen Mitgliedstaaten erhalten hat, nicht ohne Weiteres verwendbar sind. Sie erfüllen die an ihre Qualität und Nutzbarkeit für die deutsche Importstatistik gesetzten Kriterien – zumindest als hauptsächliche Datenquelle – derzeit nicht. Dazu sind die genannten Probleme mit der Verfügbarkeit, Übereinstimmung und Geheimhaltung der Daten zu schwerwiegend und die Zuverlässigkeit und Vollständigkeit der Datenlieferungen unzureichend. Insbesondere ist die Umsetzung eines Einstromverfahrens verbunden mit einem vollständigen Wegfall der Intrastat-Importerhebung derzeit nicht realistisch.

Dennoch bietet der Mikrodatabaustausch Ansatzpunkte, um die durch die EBS-Verordnung vorgegebenen Ziele zu erreichen: Unter Berücksichtigung der methodischen Asymmetrien ergeben sich neue Möglichkeiten, die bisher erhobenen Eingangsdaten zu plausibilisieren und zu ergänzen. Weiterhin können Warenverkehre, die bislang nicht von der Intrahandelsstatistik erfasst werden konnten, einbezogen werden. Insbesondere liegen erstmals tief gegliederte Einzeldaten auch zu jenen Unternehmen vor, die in Deutschland von der Meldepflicht befreit sind. Dies eröffnet neue Schätzmöglichkeiten für die innergemeinschaftlichen Warenimporte von Unternehmen unterhalb der Anmeldeschwelle auf detaillierter Warenebene, die bisher nur auf Ebene der Kapitel des Warenverzeichnisses für die Außenhandelsstatistik vorliegen. So kann künftig die Anmeldeschwelle der Importerhebung deutlich angehoben und damit ein wesentlicher Teil der derzeit meldepflichtigen Unternehmen entlastet werden, ohne die Qualität der Außenhandelsstatistik zu beeinträchtigen. Die entsprechenden Verfahren werden derzeit erarbeitet und werden Gegenstand eines weiteren Beitrags in dieser Zeitschrift sein.

Schon jetzt ist indes festzuhalten, dass die Daten aus dem Mikrodatabaustausch insgesamt eine wertvolle Ergänzung des bisherigen Datenmaterials zu den deutschen Importen aus anderen EU-Mitgliedstaaten sind, ohne es vollständig ersetzen zu können. 

LITERATURVERZEICHNIS

European Statistical System Committee. [*ESS Vision 2020*](#). In: WISTA Wirtschaft und Statistik. Ausgabe 3/2016, Seite 11 ff.

Herzog, Natascha. [*Auswirkungen der neuen europäischen Verordnung für Unternehmensstatistiken auf das nationale statistische System*](#). In: WISTA Wirtschaft und Statistik. Ausgabe 5/2020, Seite 47 ff.

Schüßler, Simone/Herold, Lucie/Roller, Jonas. [*Datenaktualisierung des Belastungsbarometers: aktuelle Zahlen zu Bürokratiekosten durch amtliche Statistiken*](#). In: WISTA Wirtschaft und Statistik. Ausgabe 1/2024, Seite 109 ff.

Steinfelder, Joseph. [*SIMSTAT als „business case“ für einen statistischen Datenaustausch in der Europäischen Union*](#). In: WISTA Wirtschaft und Statistik. Ausgabe 4/2016, Seite 25 ff.

RECHTSGRUNDLAGEN

Verordnung (EG) Nr. 638/2004 des Europäischen Parlaments und des Rates vom 31. März 2004 über die Gemeinschaftsstatistiken des Warenverkehrs zwischen Mitgliedstaaten und zur Aufhebung der Verordnung (EWG) Nr. 3330/91 des Rates (Amtsblatt der EG Nr. L 102, Seite 1), die zuletzt durch die Verordnung (EU) Nr. 659/2014 des Europäischen Parlaments und des Rates (Amtsblatt der EG Nr. L 189, Seite 128) geändert wurde.

Verordnung (EU) 2019/2152 des Europäischen Parlaments und des Rates vom 27. November 2019 über europäische Unternehmensstatistiken, zur Aufhebung von zehn Rechtsakten im Bereich Unternehmensstatistiken (Amtsblatt der EU Nr. L 327, Seite 1).

Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates vom 10. Juli 2023 zur Änderung der Verordnung (EG) Nr. 223/2009 über europäische Statistiken.

MASCHINELLES LERNEN IM BASIS-REGISTER FÜR UNTERNEHMEN

Vorstudie zum Potenzial automatischer Konsolidierung von Unternehmensstammdaten

Julius Weißmann, Tim Herbst

➤ **Schlüsselwörter:** Datensatzverknüpfung – Registerdaten – Textähnlichkeit – überwachtes Lernen – Datenanalyse

ZUSAMMENFASSUNG

Das Statistische Bundesamt hat den gesetzlichen Auftrag, ein Register über Unternehmensbasisdaten für Deutschland aufzubauen. Die Inhalte und Ziele des Basisregisters für Unternehmen werden in dem zugrunde liegenden Gesetz definiert. Diese liegen vor allem in der Konsolidierung der Basisdaten von Unternehmen aus verschiedenen Quellregistern an einer zentralen Stelle sowie die Gewährleistung von deren Aktualität und Richtigkeit. Die korrekte und effiziente Verknüpfung der Basisdaten von Unternehmen aus den verschiedenen Quellregistern hat hierbei eine herausragende Bedeutung. Dafür hat das Statistische Bundesamt intern eine explorative Studie durchgeführt, die den Nutzen von maschinellem Lernen zur Verknüpfung der Quelldaten evaluiert hat. Darüber hinaus informiert der Beitrag über einen teilautomatisierten Ansatz, welcher Verknüpfungen nur unter der Voraussetzung von hinreichend sicheren Vorhersagen vornimmt.

➤ **Keywords:** data record linkage – register data – text similarity – supervised learning – data analysis

ABSTRACT

The Federal Statistical Office has a legal mandate to establish a register of basic enterprise data for Germany. The content and objectives of the basic register of enterprises are defined in the underlying legal act. Its main aims are to consolidate basic enterprise data from various source registers in a central location and to ensure the timeliness and accuracy of the data. Correct and efficient linkage of the basic enterprise data from various source registers is of paramount importance in this context. For this purpose, the Federal Statistical Office conducted an internal exploratory study to evaluate the benefits of machine learning for linking source data.



Julius Weißmann

ist Data Scientist und wissenschaftlicher Mitarbeiter im Referat „Künstliche Intelligenz, Big Data“ des Statistischen Bundesamtes. Er befasst sich mit dem Einsatz von maschinellem Lernen in der amtlichen Statistik sowie der Automatisierung von Statistik- und Nichtstatistikprozessen.



Tim Herbst

ist Data Analyst und im Referat „Basisregister für Unternehmen – Fachverfahren“ des Statistischen Bundesamtes tätig. Er befasst sich mit der fachlichen Integration und Zusammenführung der Quellregisterdaten für das Basisregister.

1

Einleitung

Der Aufbau einer modernen Registerlandschaft durch übergreifende Nutzbarmachung von in Registern gespeicherten Daten ist nicht erst seit dem aktuellen Koalitionsvertrag (SPD, Bündnis 90/Die Grünen und FDP, 2021) politischer Wille in Deutschland. Ziel ist, mit einer effizienten, digitalen Verwaltung sowohl Mehrwert zu generieren als auch die digitale Handlungsfähigkeit des Staates sicherzustellen. Dies stellt eine große Herausforderung, aber auch eine große Chance für die öffentliche Verwaltung dar. Das gilt gleichermaßen für Aufbau und Inbetriebnahme des Basisregisters für Unternehmen (im Folgenden Basisregister) mit dem Statistischen Bundesamt als registerführender Behörde (§ 1 Absatz 1 bis 3 Unternehmensbasisdatenregistergesetz).

➤ Nach § 3 Absatz 2 Unternehmensbasisdatenregistergesetz im Basisregister als Unternehmen geführte und in den Quellregistern gespeicherte Einheiten:

1. Kaufleute im Sinne des Handelsgesetzbuchs;
2. Genossenschaften im Sinne des Genossenschaftsgesetzes;
3. Partnerschaften im Sinne des Partnerschaftsgesellschaftsgesetzes;
4. Vereine im Sinne des Bürgerlichen Gesetzbuchs;
5. wirtschaftlich Tätige im Sinne der Abgabenordnung:
 - a) natürliche Personen, die wirtschaftlich tätig sind,
 - b) juristische Personen und
 - c) Personenvereinigungen; sowie
6. weitere Unternehmen im Sinne des Siebten Buches Sozialgesetzbuch.

Als bundeseinheitliche Wirtschaftsnummer dient dabei die Wirtschafts-Identifikationsnummer nach § 139 c der Abgabenordnung (§ 2 Absatz 1 Unternehmensbasisdatenregistergesetz), welche das Bundeszentralamt für Steuern vergibt. Für die Umsetzung haben sich moderne technologische Ansätze wie maschinelles Lernen (ML) als mögliche Katalysatoren im Bereich der Informationsverknüpfung erwiesen (Schnell, 2021). Aus diesem Grund stellte auch in der im Statistischen Bundesamt durchgeführten Vorstudie zum Basisregister für Unternehmen maschinelles Lernen einen zentralen

Baustein in den Untersuchungen dar. Sie hatte zum Ziel, Erkenntnisse darüber zu erlangen, wie für die Zusammenführung der bestehenden Einzelregister automatisierte Verfahren gewinnbringend eingesetzt werden können.

Das folgende Kapitel 2 stellt das Basisregister für Unternehmen allgemein vor. Kapitel 3 beschreibt die Vorstudie zum Basisregister, informiert zu deren Datenbasis, zur Effizienz der Datensatzverknüpfung und zeigt Ähnlichkeiten in den Quellregistern auf. Wie maschinelles Lernen in der Vorstudie eingesetzt wurde, erläutert Kapitel 4 mit Erläuterungen zum Datensatz, zur maschinellen Lernstrategie, zur Datensatzverknüpfung und Deep-Learning-Ansätzen. Abschließend bewertet Kapitel 5 die Vorstudie zum Basisregister und die mit ihr gewonnenen Erkenntnisse.

2

Das Basisregister für Unternehmen

Die Einführung des Basisregisters soll erstmals den zentralen und registerübergreifenden Zugriff auf Unternehmensbasisdaten unter den gesetzlich definierten Rahmenbedingungen ermöglichen. Es birgt somit erhebliches Potenzial, um Verwaltungsprozesse effizienter zu gestalten und die Bürokratiekosten der Unternehmen zu senken. Die deutsche Registerlandschaft umfasst rund 120 einzelne Register mit Unternehmensbezug, die alle zweckgebunden und weitgehend unabhängig voneinander agieren (BMWK, 2021). Dies bedeutet Pflegeaufwand für jedes einzelne Register und führt vor allem dazu, dass Daten zwischen den Registern inkonsistent sind. Das Basisregister soll definierte Registerinformationen mit dem Ziel der „Single Source of Truth“ als konsistente Datenbasis zusammenführen. Mehrfachmeldungen der Unternehmen zur Datenaktualisierung sollen vermieden und gleichzeitig ein effizienter Datenaustausch zwischen den Registern ermöglicht werden. Das Basisregister umfasst künftig für diesen Zweck die qualitätsgesicherten Stammdaten aller Unternehmen und führt in diesem Zusammenhang die eindeutige bundeseinheitliche Wirtschaftsnummer als Identifikator ein. Dies verspricht mit Inbetriebnahme nicht nur einen effizienten Datenaustausch in der Verwaltung, sondern entlastet auch die Unternehmen selbst, da Mehrfachmeldungen bei Veränderungen entfallen.

➤ Nach § 3 Absatz 3 Unternehmensbasisdatenregistergesetz im Basisregister gespeicherte Stammdaten:

1. Für den Rechtsverkehr verbindliche Angabe der Firma oder des Namens entsprechend der Eintragung im Handelsregister, Partnerschaftsregister, Genossenschaftsregister oder Vereinsregister,
2. für Verwaltungszwecke aktuelle Angabe der Firma oder des Namens entsprechend der Führung im Datenbestand der öffentlichen Stelle nach § 4 Absatz 1,
3. Verwaltungsanschrift unter Angabe von Straße, Hausnummer, Postfach, Postleitzahl, Ort und Länderkennzeichen,
4. Sitz (Ort),
5. inländische Geschäftsanschrift entsprechend der Eintragung im Handelsregister, Partnerschaftsregister, Genossenschaftsregister oder Vereinsregister unter Angabe von Straße, Hausnummer, Postleitzahl, Ort und Länderkennzeichen, soweit die Pflicht zur Eintragung besteht,
6. Rechtsform und
7. Haupttätigkeit nach Klassifikation der Wirtschaftszweige.

Verwaltungsstellen können nach § 5 Unternehmensbasisdatenregistergesetz nach der Inbetriebnahme des Basisregisters auf dieses zugreifen und aktuelle Stammdaten zu Unternehmen abrufen. Angebundene Register verfügen so über aktuelle Daten hoher Qualität, wodurch Kosten durch veraltete Informationen vermieden werden können. Das Statistische Bundesamt spielt dabei als erfahrene registerführende Stelle eine entscheidende Rolle und trägt dazu bei, die Register- und Verwaltungslandschaft in Deutschland zu modernisieren.

3

Vorstudie zum Basisregister

Die Daten für das Basisregister werden für den Betrieb aus mehreren Quellen (sogenannte Quellregister) zur Verfügung gestellt. Die Quellregister – namentlich das Bundeszentralamt für Steuern, das zentrale Unternehmerverzeichnis der Deutschen Gesetzlichen Unfallversicherung und das gemeinsame Registerportal der Länder (Justiz) – führen ihre Datenbanken nach unter-

schiedlichen Kriterien und verfügen jeweils über einen voneinander abweichenden Umfang an Datensätzen beziehungsweise Unternehmen. Das Basisregister wird künftig all diese Daten zusammenbringen und vereinheitlichen. Um aus den bestehenden Datenbeständen erste Schlüsse auf die zu erwartenden Dateninhalte, die Datenqualität und die Herausforderungen bei der Verknüpfung ziehen zu können, hat das Statistische Bundesamt eine Vorstudie zum Basisregister durchgeführt. Dafür haben die oben genannten Quellen für das Bundesland Hamburg Daten bereitgestellt, die für die geplante Zusammenführung untersucht wurden. Um zu aussagekräftigen Ergebnissen zu gelangen war es wichtig, bei den Untersuchungen auf Echtdaten zurückzugreifen und die Konsolidierung in einem definierten Rahmen möglichst realitätsnah zu simulieren. Hierzu wurden ausschließlich Daten juristischer Personen herangezogen. Ziel war, Einblicke in die Datenqualität und den Datenumfang der einzelnen Quellen zu erhalten und Potenziale für die automatisierte Zusammenführung der Datenbestände zu identifizieren.

3.1 Datenbasis

Eine vorbereitende Datenanalyse hat zu Beginn erste Einblicke in die Daten ermöglicht, Muster identifiziert und Hypothesen generiert. Die Datenlieferungen zur Vorstudie zeigen, dass jede Quelle über eine unterschiedliche Anzahl an Unternehmensdatensätzen in ihrer Datenbank verfügt. Gleichzeitig unterscheiden sich die zur Verfügung gestellten Datensätze quellspezifisch auch in der Definition, der Verfügbarkeit und der Formatierung der Attribute, welche für die Datensatzverknüpfung vorgesehen sind. Mit der bundeseinheitlichen Wirtschaftsnummer soll künftig ein quellenübergreifender Identifikator zur Konsolidierung geschaffen werden. Dieser existierte für die erste Zusammenführung der Daten jedoch nicht, sodass eine andere Verknüpfungsstrategie erforderlich war. Die Daten mussten anhand

Tabelle 1
Umfang der Vorstudien Daten

	Zur Verfügung stehende Datensätze
Bundeszentralamt für Steuern	347 493
Zentrales Unternehmerverzeichnis der Deutschen Gesetzlichen Unfallversicherung	151 884
Gemeinsames Registerportal der Länder (Justiz)	137 464
Insgesamt	636 841

Übersicht 1

Bewertung der Attribute für die Zuordnungsfindung hinsichtlich Güte und Verfügbarkeit

Attributgruppe	Verfügbarkeit der Attribute in Prozent der gesamten Datensätze	Einschätzung zur Verfügbarkeit und Nutzbarkeit der Attribute
Unternehmensbezeichnung	100	Immer verfügbar, je Amtsgericht und je Quellregister eindeutig, anfällig für Schreibfehler, bedingt normierbar
Adressinformationen	99	Hohe Verfügbarkeit, normierbar, für Clusterbildung geeignet
Registerstring ¹	30	Nur für juristische Personen verfügbar, dort jedoch per Definition eindeutig
Rechtsform	100	Immer verfügbar, kein einheitliches Schema, als Attribut für Clusterbildung geeignet

¹ Der für die Vorstudie definierte Registerstring setzt sich aus dem Registergericht, an dem die Eintragung stattgefunden hat, der Registerart und der Registernummer zusammen. Der angegebene Wert bezieht sich auf alle Einheiten, jedoch sind viele Einheiten nicht eintragungspflichtig, weshalb fehlende Werte hier toleriert werden müssen. Dies ist gleichzeitig die größte Herausforderung bei der Zusammenführung, da der eindeutige Identifikator bisher fehlt und erst mit Vergabe der bundeseinheitlichen Wirtschaftsnummer (beWiNr.) eingeführt wird.

der übermittelten Attribute identifiziert und so miteinander verknüpft werden, dass ein konsolidierter Datensatz entstand, der korrekte vorhandene Informationen aus bis zu drei Quellen umfasst. Für die explorative Vorstudie wurden insgesamt 636 841 Datensätze ausgewertet, die sich auf die einzelnen Quellregister aufteilen. Dabei wurden mehrere Attribute zur Verknüpfung der Daten in Betracht gezogen und auf ihre Eignung hinsichtlich der Nutzbarkeit für ein Matching der Daten eingeschätzt.

↗ Tabelle 1

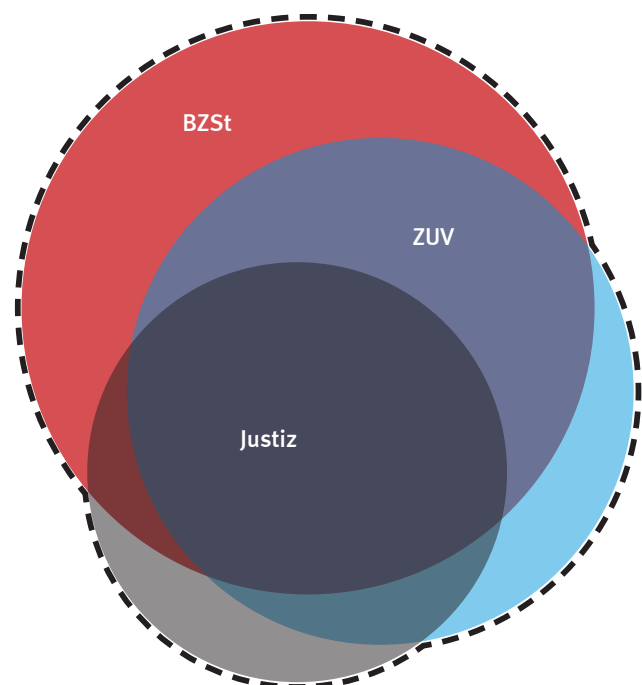
↗ Übersicht 1 zeigt, dass sich die verfügbaren Attribute unterschiedlich gut zur Verknüpfung der Datensätze zwischen den Quellregistern eignen. Hierbei sind neben der Vollständigkeit auch andere Faktoren wie Verfügbarkeit oder Eindeutigkeit entscheidend. Deshalb muss der Worst Case, das heißt die geringste Verfügbarkeit eines Attributes, angenommen werden. Außerdem sind die eingehenden Informationen nicht immer auf demselben Stand und können so zu Widersprüchen bei der Zuordnung führen. Um dieser Problematik zu begegnen, sollte die Vorstudie testen, inwiefern Algorithmen aus dem maschinellen Lernen in diesem Kontext geeignet sind, Zuordnungen hoher Güte mithilfe der verfügbaren Attribute vorzunehmen. Ziel ist, die automatische Zuordnungsgüte weiter zu erhöhen und – falls manuelle Entscheidungen zu treffen sind – die Sachbearbeitung durch passende Vorschläge zu unterstützen.

Aus der verfügbaren Datenbasis wird deutlich, dass beim Zusammenführen der übermittelten Informationen Herausforderungen für die Konsolidierung entstehen können. Kein Quellregister geht vollständig in der Menge eines anderen Quellregisters auf. Dabei kann

einerseits noch nicht exakt bestimmt werden, wie groß der Abdeckungsgrad der Quellregister gegeneinander ist; dies ist schematisch in ↗ Grafik 1 dargestellt. Andererseits sind die Attribute nicht fehlerfrei und teilweise nicht übereinstimmend, sodass eine gewisse Fehlertoleranz und ein eindeutiges Regelwerk für die Verknüpfung übereinstimmender Datensätze erforderlich

Grafik 1

Schematische Darstellung des geplanten Registerumfangs als Venn-Diagramm



BZSt: Bundeszentralamt für Steuern; ZUV: Zentrales Unternehmensverzeichnis der Deutschen Gesetzlichen Unfallversicherung; Justiz: Gemeinsames Registerportal der Länder (Justiz).

sind. Sind Attribute nicht übereinstimmend, ist es Aufgabe des Basisregisters, den richtigen Eintrag je Attribut zu identifizieren und in das konsolidierte Unternehmen zu übernehmen. Hierbei soll der Automatisierungsgrad durch maschinelle Entscheidungen so groß wie möglich werden, ohne Qualitätseinbußen in Kauf zu nehmen. Das Basisregister wird nach der Aufbauphase Daten aus allen Quellregistern enthalten, sodass die Gesamtzahl an Unternehmen größer ist als in jedem Quellregister. Der äußere Rand des gesamten Venn-Diagramms stellt hierbei den Umfang des Basisregisters dar.

3.2 Effiziente Datensatzverknüpfung

Ein grundlegendes Problem bei der Datensatzverknüpfung ist die Komplexität der Daten, da eine komplette Überprüfung zur Folge hätte, dass alle Einträge unter den Quellregistern miteinander verglichen werden müssten (Christen, 2012). Gerade bei anspruchsvollen Ähnlichkeitsberechnungen lässt sich die vollständige Überprüfung nicht mehr gewährleisten. Daher wird in der Praxis häufig über das Blocking eine Vorauswahl an Einträgen getroffen, welche überhaupt für einen Vergleich infrage kommen. Zwar lässt sich über das Blocking die Menge der Vergleiche reduzieren, allerdings führt das im Umkehrschluss zu einem Ausschluss von potenziell zusammengehörenden Einträgen. Im Kompromiss aus Genauigkeit und Geschwindigkeit wurden unter der vorhandenen IT-Infrastruktur die Postleitzahl und die Hausnummer für das Blocking verwendet. Diese liegen ausreichend fehlerfrei vor und können somit als notwendige Voraussetzung für eine effiziente Verlinkung gesehen werden. Durch das Blocking ließ sich die Anzahl der notwendigen Vergleiche von 720 385 600 auf 463 690 (0,06 %) reduzieren.

3.3 Ähnlichkeiten in den Quellregistern

Ziel der Untersuchung war, die automatisierte Verknüpfung der unterschiedlichen Quellregister zu bewerten. Generell erfolgt die Verknüpfung mehrerer Datenbestände idealerweise für jeden Eintrag über einen eindeutigen Identifikator. In den Daten aus Quellregistern gibt es jedoch keinen solchen Identifikator, weshalb gängige Ähnlichkeitsmaße herangezogen wurden (de Bruin, 2022).

Neben schnell berechenbaren Ähnlichkeitsmaßen¹ wurden für die Verlinkung weitere Ähnlichkeiten, basierend auf dem Tf-idf-Maß² erzeugt. In der Verarbeitung natürlicher Sprache ist das ein übliches Vorgehen, so werden Texte häufig in N-Grams³ repräsentiert und durch das Tf-idf-Maß bewertet.

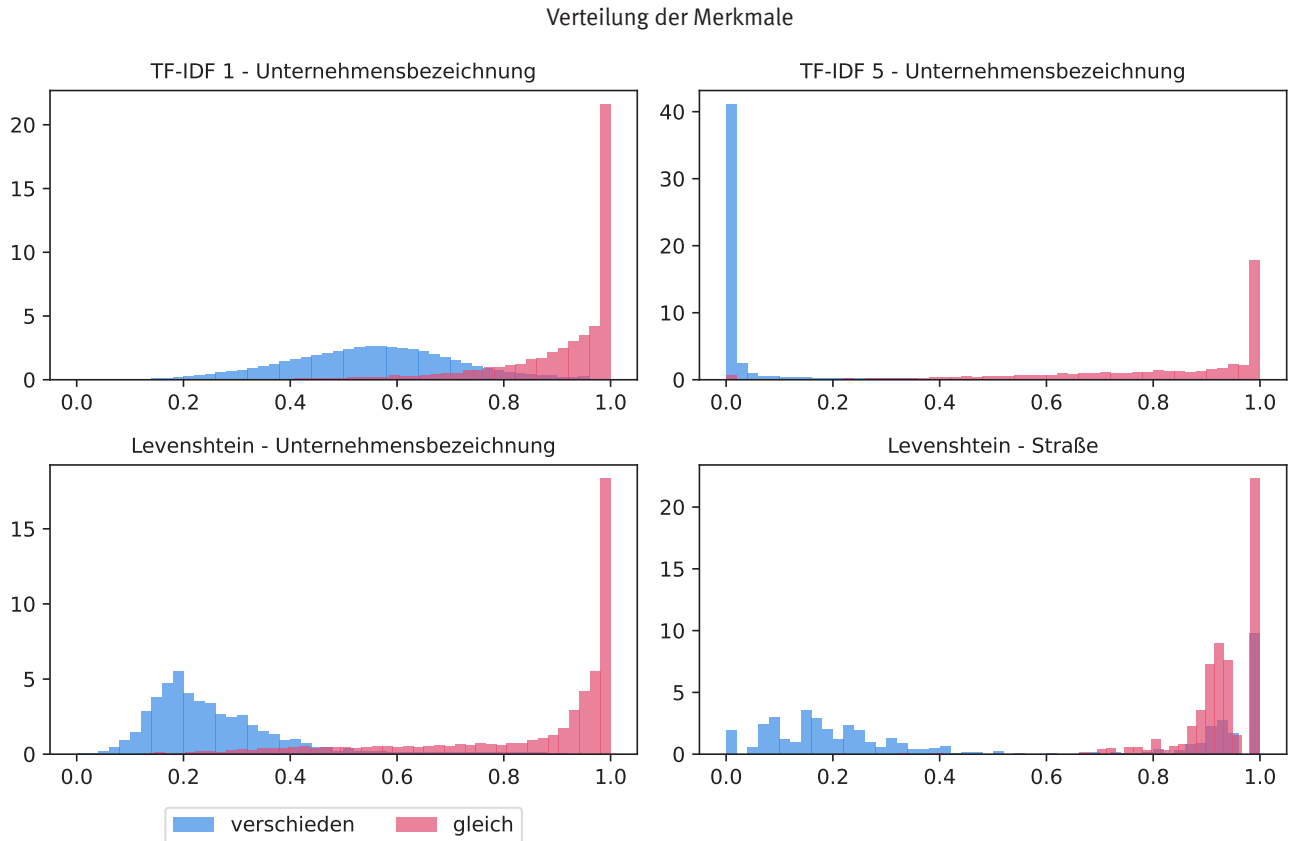
Bei der Betrachtung der Ähnlichkeitsmaße fällt auf, dass sich beispielsweise die Levenshtein-Ähnlichkeit (Yujian/Bo, 2007) zwar schnell berechnen lässt, dafür jedoch in der Schwellenwert-basierten Einordnung fehleranfällig ist. Besonders bei den nicht normierten Straßennamen fallen die Fehlzuordnungen ins Gewicht. Das lässt erahnen, dass eine geeignete Normierung von zentraler Bedeutung sein wird. [↩ Grafik 2 auf Seite 72](#)

Ebenfalls lässt sich in Grafik 2 erkennen, dass ein größeres N-Grams zu deutlich aussagekräftigeren Ähnlichkeitsmaßen führt, was eine verlässlichere Verlinkung der Quellregister ermöglicht. Zwar scheinen besonders längere Vektoren – wie sie durch größere N-Grams entstehen – erfolgversprechend für die Verlinkung anhand der Unternehmensnamen im Basisregister zu sein, allerdings benötigen diese besonders viel Rechenspeicher. Um diesem Effekt entgegenzuwirken wurde (wie in Abschnitt 4.4 beschrieben) versucht, durch einen Autoencoder⁴ eine speichereffizientere Repräsentation der Vektoren zu gewährleisten.

- 1 Eine umfassende Übersicht aller Ähnlichkeitsmaße enthält Grafik 4.
- 2 Tf-idf steht für Term Frequency – Inverse Document Frequency und lässt sich mit „Vorkommenshäufigkeit – inverse Dokumenthäufigkeit“ übersetzen. Dieses statistische Maß wird eingesetzt, um die Relevanz der N-Grams in Bezug auf den gesamten Datensatz zu bewerten.
- 3 N-Grams, auch Q-Grams genannt, sind die Fragmente (in diesem Fall Buchstaben), in welche ein Text zerlegt wurde. Bei diesem Verfahren werden N aufeinanderfolgende Buchstaben zusammengefasst. Durch die Überführung eines Wortes in N-Grams lassen sich Muster in den Zeichenketten erkennen und ähnliche zusammengesetzte Wörter identifizieren.
- 4 Autoencoder gehören den künstlichen neuronalen Netzen an und können wie in diesem Fall genutzt werden, um durch nicht lineare Operationen effiziente Kodierungen zu erzeugen (Bank und andere, 2020).

Grafik 2

Gegenüberstellung der Verteilung von vier Ähnlichkeitsmaßen



Tf-idf: Term Frequency – Inverse Document Frequency (Vorkommenshäufigkeit – inverse Dokumenthäufigkeit). Idealerweise sollten die zusammengehörigen Einträge (rot) eine Säule nahe bei 100 bilden und äquivalent die ungleichen Einträge eine blaue Säule nahe bei null.

4

Einsatz von maschinellem Lernen in der Vorstudie zum Basisregister für Unternehmen

4.1 Datensatz zum maschinellen Lernen

Unter den untersuchten Ähnlichkeitsmaßen eignete sich keines einzeln, um eine verlässliche Verknüpfung der Quellregister zu gewährleisten. Aus diesem Grund wurden Experimente mit Algorithmen aus dem maschinellen Lernen durchgeführt. Beim maschinellen Lernen lassen sich aus mehreren Merkmalen, welche in diesem Fall die generierten Ähnlichkeitsmaße aus den Attributen Unter-

nehmensbezeichnung und Straßenname sind, nicht lineare Zusammenhänge ableiten. Das trainierte Modell soll damit in der Lage sein, automatisiert verlässlichere Verknüpfungen zwischen den Quellregistern vorzunehmen, als es mit einem einfachen Schwellenwert der Fall wäre. Modelle aus dem überwachten maschinellen Lernen benötigen für das Training jedoch Labels, welche in diesem Fall dem nicht vorhandenen eindeutigen Identifikator entsprechen. Einige Einheiten enthalten durch ihre Eintragungspflicht Informationen über Registernummer, Registerart und Amtsgericht, anhand deren Kombination sich eine eindeutige Verknüpfung durchführen lässt. Unter der Annahme, dass die Attributinhalt der Einheiten mit vorhandenem Registerstring auch denen ohne Registerstring und damit der Grundgesamtheit entsprechen (siehe Übersicht 1), wurden die Informationen über die Zusammengehörigkeit durch Übereinstimmung

des Registerstrings gewährleistet und entsprechend aus den Trainingsdaten entfernt. Dieser Schritt ist für das überwachte maschinelle Lernen in diesem Fall die bestmögliche Annäherung an die in den Realdaten tatsächlich fehlenden Identifikatoren. Das Modell wird somit gezwungen, anhand der übrigen Merkmale eine Verknüpfung zu erstellen, wenngleich die Labels, also die Zusammengehörigkeit, zur Evaluierung des Modells bekannt sind. Im tatsächlichen Einsatz würden die Attribute des Registerstrings, wenn vorhanden, zusätzlich für die Datensatzverknüpfung verwendet werden, da sie eine definitorisch eindeutige Zuordnung zulassen und sehr effizient sind. Bedingt durch die Kapazität der nutzbaren IT-Infrastruktur hat sich die Vorstudie für die Untersuchungen im Bereich des maschinellen Lernens auf die Datenbestände des Bundeszentralamts für Steuern und des zentralen Unternehmerverzeichnisses der Deutschen Gesetzlichen Unfallversicherung beschränkt.

Mit diesem aufbereiteten Datenbestand soll anhand der verbleibenden Attribute, bestehend aus Unternehmensbezeichnung, Postleitzahl, Hausnummer und Straßenname, die Verknüpfung der Quellregister umgesetzt werden. Insgesamt lassen sich aus den Datenbeständen des Bundeszentralamts für Steuern und des zentralen Unternehmerverzeichnisses der Deutschen Gesetzlichen Unfallversicherung durch dieses Vorgehen jeweils 26 840 gelabelte Proben erstellen. Davon besteht für die Hälfte eine Verknüpfung durch den erzeugten Identifikator zum anderen Datenbestand.

4.2 Maschinelle Lernstrategie

Beim überwachten maschinellen Lernen werden unterschiedliche Modelle mit gelabelten Daten trainiert und anschließend auf für das Modell unbekannten Daten getestet, um eine Aussage über die Transferleistung des Modells zu ermöglichen. Für eine zuverlässige Bewertung der ML-Algorithmen wurde die Datensatzverknüpfung mit einer fünffachen Kreuzvalidierung durchgeführt und das Hyperparametertuning⁵ in einer erneuten fünffachen Kreuzvalidierung verschachtelt. Für

⁵ Das Hyperparametertuning ist entscheidend für den Erfolg eines Modells. Dabei wird das Modell mit unterschiedlichen Voreinstellungen (das heißt Hyperparametern) auf denselben Daten trainiert und die beste Hyperparameter-Konstellation für die weitere Anwendung übernommen. Da es viele mögliche Konstellationen für jedes Modell gibt, wurde hier mit einer zufälligen Suche (englisch: Random Search) gearbeitet.

die Analysen wurden insgesamt acht unterschiedliche ML-Algorithmen zur Klassifikation herangezogen (siehe Grafik 3). Für die Hyperparameteroptimierung wurden etwa 60 Konstellationen je Klassifikator verwendet, wobei die rechenaufwendige Support Vector Maschine aus Performancegründen nur in vier Konstellationen verglichen werden konnte. Darüber hinaus wurde mit einem Oversampling⁶ gearbeitet, um Verzerrungen im Modell zu reduzieren (Géron, 2019; Hasti und andere, 2009). Es wurde in einer Python-Umgebung gearbeitet und größtenteils scikit-learn als ML-Bibliothek (Pedregosa und andere, 2011) und das Python Record Linkage Toolkit (de Bruin, 2024) für die Datensatzverknüpfung verwendet.

4.3 Datensatzverknüpfung durch maschinelles Lernen

➤ Grafik 3 zeigt einen Vergleich der Ergebnisse der trainierten ML-Modelle aus der fünffachen Kreuzvalidierung. Da bei der Datensatzverknüpfung vor allem falsche Verknüpfungen vermieden werden sollen, wird der positive prädiktive Wert (PPV)⁷ als entscheidendes Maß für die Evaluation herangezogen. Das Random-Forest-Modell erweist sich mit einem PPV von 0,978 am leistungsstärksten. Im Einklang mit dem hohen PPV fällt der Standardfehler beim Random-Forest-Modell mit 0,003 am geringsten aus und unterstreicht damit die Aussagekraft des Modells. Neben dem PPV wird auch der F1-Wert⁸ dargestellt, um einen allgemeinen Eindruck über die Performance des Modells zu gewährleisten. Hier liegt der Random-Forest-Klassifikator mit 0,988 nur knapp hinter XGBOOST mit einem Wert von 0,989. Basierend auf PPV und dem F1-Wert kann das Random-Forest-Modell für die Datensatz-Verlinkung als geeignetes Modell betrachtet werden.

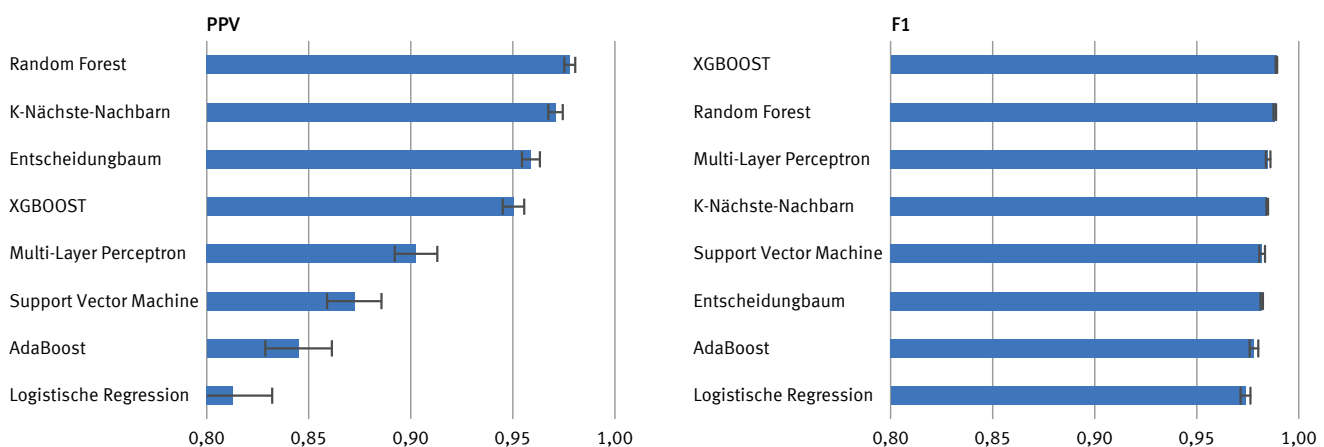
⁶ Das Oversampling (auch Upsampling genannt) lässt sich mit überrepräsentierter Stichprobe übersetzen und wird verwendet, um Klassenungleichgewichte zu adressieren. Es hilft die Fähigkeit des Modells zur genauen Vorhersage der Minderheitsklasse zu verbessern, indem es mehr Beispiele zum Lernen zur Verfügung stellt.

⁷ Der positive prädiktive Wert (auch positiver Vorhersagewert; englisch: Precision oder Positive Predictive Value) ist die Wahrscheinlichkeit, dass eine Verlinkung mit positivem Testergebnis tatsächlich einer Verlinkung entspricht. Der Wert ergibt sich aus folgender Formel: $PPV = \text{richtig-positiv} / (\text{richtig-positiv} + \text{falsch-positiv})$

⁸ Der F1-Wert ist eine Kennzahl, die sich zur Bewertung von Klassifizierungsmodellen eignet. Er berechnet sich aus dem harmonischen Mittel aus dem positiven Vorhersagewert und der Sensitivität.

Grafik 3

Modellvergleich aus fünffacher Kreuzvalidierung



Anmerkungen: Modellvergleich durch PPV und F1-Wert mit ± 2 Standardfehler in gemittelter fünffacher Kreuzvalidierung. – Zu beachten ist, dass die Skala nicht bei 0 beginnt. – Der PPV ist das entscheidende Maß für die Modellwahl, da er die Wahrscheinlichkeit schätzt, dass eine vom Modell vermutete Verlinkung tatsächlich korrekt ist. Die Modelle mit höherem PPV weisen darüber hinaus auch geringere Standardabweichungen über die Kreuzvalidierung auf. Unter Betrachtung des F1-Wertes verändert sich insbesondere die Einordnung des Random-Forest-Modells nur geringfügig.

Anhand der Permutation Importance⁹ wird die Bedeutung der Merkmale für das erfolgreichste Modell (Random-Forest-Klassifikator) veranschaulicht. Es fällt auf, dass besonders die Tf-idf-Maße eine entscheidende Rolle für den Erfolg des Modells einnehmen, wobei das 5-Gram besonders wichtig für den Erfolg ist. Der Straßenname scheint hingegen in nicht normierter Form für den Erfolg des Modells nur von vernachlässigbar geringer Bedeutung zu sein. [➤ Grafik 4](#)

[➤ Grafik 5](#) auf Seite 76 verdeutlicht den Mehrwert des Random-Forest-Modells für die automatisierte Verknüpfung zwischen den Quellregistern. Eine Alternative zur Verwendung des ML-Modells, welches sich mehrerer Merkmale bedient, stellt eine rein Schwellenwert-basierte Zuordnung auf Basis der Kosinus-Ähnlichkeit aus den 5-Grams dar. Es ist deutlich zu erkennen, dass die Kurve des PPV beim ML-Modell deutlich schneller ansteigt und dass diese mit einer höheren Sensitivität einhergeht, wodurch eine höhere Abdeckung der relevanten Verknüpfungen erreicht wird. Das ML-Modell verknüpft bei einem PPV von 0,95 im direkten Vergleich 95,1 % der zusammengehörigen Daten, wohingegen beim Tf-idf-Maß nur 77,1 % der Daten zusammengeführt

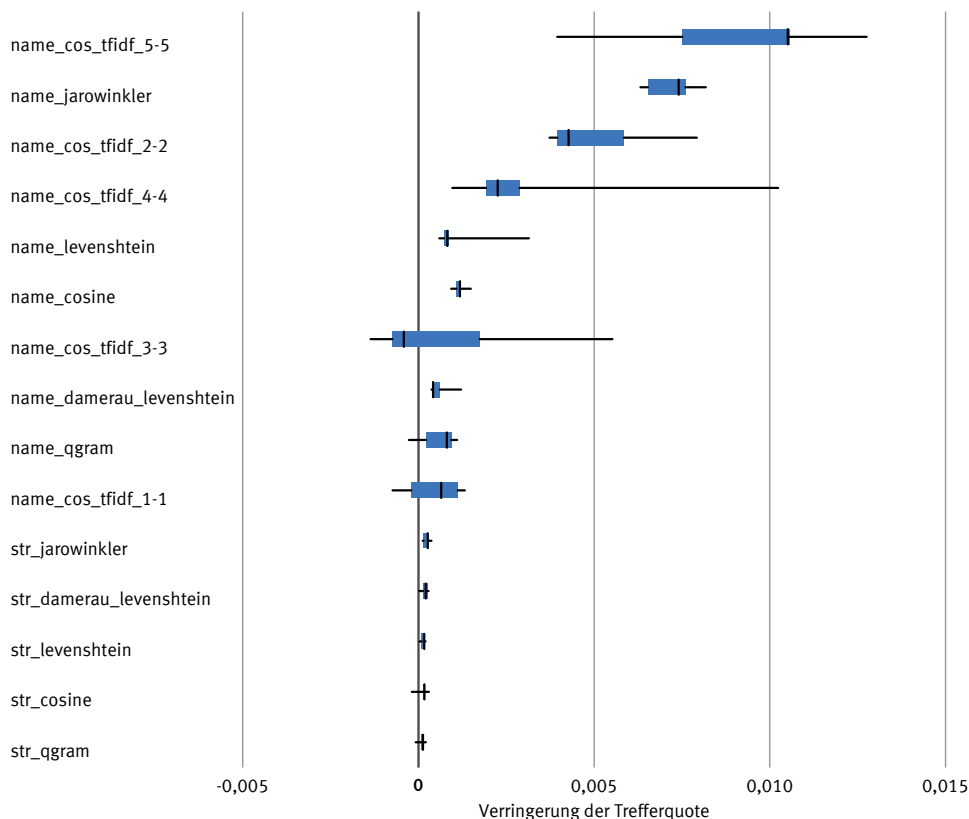
werden. Ähnlich verhält es sich bei einem PPV von 0,99, welcher im Gegensatz zum rein Tf-idf-basierten Verfahren nicht mehr bei einer Sensitivität von 0,571, sondern bei einem Wert von 0,883 liegt. Grafik 5 verdeutlicht daher, dass sich durch maschinelles Lernen anhand der verschiedenen Merkmale mehr Daten mit einer höheren Präzision automatisiert verarbeiten lassen. Dies ist insbesondere vor dem Hintergrund bemerkenswert, dass im Basisregister Fehlzuordnungen zwingend vermieden werden sollen und somit eine hohe Präzision erforderlich ist. Anhand der PPV-Werte könnten perspektivisch verschiedene Sicherheitsniveaus innerhalb der Verarbeitung der Daten abgeleitet werden.

⁹ Die Permutation Importance ist eine Technik zur Bewertung der Bedeutung von Merkmalen in einem maschinellen Lernmodell. Es wird gemessen, wie stark die Trefferquote des Modells abnimmt, wenn die Werte eines bestimmten Merkmals zufällig geändert werden. Dies hilft zu verstehen, welche Merkmale am einflussreichsten für die Vorhersagen des Modells sind.

Grafik 4

Permutation Importances an den Testdaten

Fünffache Kreuzvalidierung



Anmerkungen: Die Permutation Importances (Technik zur Bewertung der Bedeutung von Merkmalen in einem maschinellen Lernmodell) wurden für die Merkmale anhand der Testdaten in fünffacher Kreuzvalidierung berechnet. Für die Unternehmensbezeichnung (name) und die Straßennamen (str) wurden verschiedene Ähnlichkeiten berechnet: qgram: q-gram, jarowinkler: Jaro-Winkler, levenshtein: Levenshtein, damerau_levenshtein: Damerau-Levenshtein, cosine: Kosinus. Für die Unternehmensbezeichnung wurden außerdem die Kosinus-Ähnlichkeiten aus dem Tf-idf-Maß für N-Grams von 1 (cos_tfidf_1-1) bis 5 (cos_tfidf_5-5) berechnet.

4.4 Deep-Learning-Ansätze

Als besonders erfolgversprechend erwiesen sich für das Random-Forest-Modell die Ähnlichkeitsmaße durch den Tf-idf-Algorithmus. Diese Art der Kodierung hat bereits lange Tradition in der Verarbeitung natürlicher Sprache. In den letzten zehn Jahren werden Zeichenketten jedoch zunehmend mit künstlichen neuronalen Netzen verarbeitet, da in diesem Forschungsbereich zahlreiche bahnbrechende Erfolge erzielt wurden (Li und andere, 2020; Mueller/Thyagarajan, 2024; Santos und andere, 2024; Shuangli und andere, 2019; Vaswani und andere, 2017; Xu und andere, 2022). Aufgrund der gegebenen Hardware konnten die Analysen in diesem Themenfeld nicht vollumfänglich durchgeführt werden, was insbe-

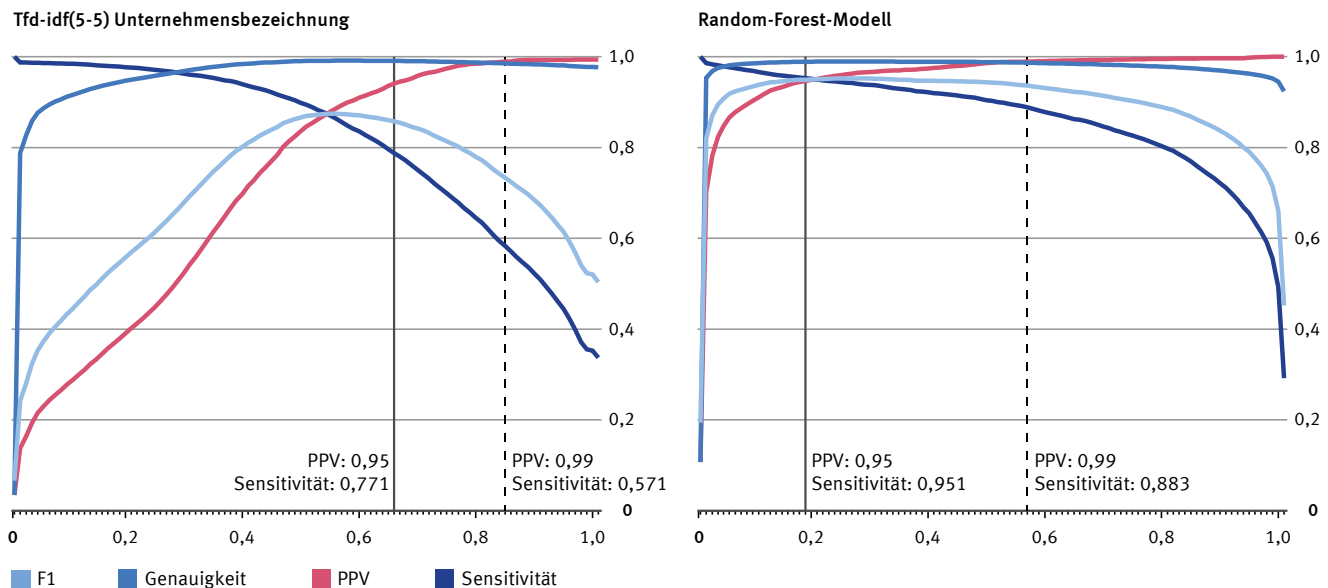
sondere an fehlenden GPUs¹⁰ liegt. Sie mussten daher auf einen späteren Zeitpunkt verschoben werden, wenn geeignete IT-Kapazitäten bereitstehen.

Auf Basis von künstlichen neuronalen Netzen wurden zwei Ansätze erarbeitet, anhand derer weitere Ähnlichkeitsmaße aus den Zeichenketten extrahiert werden. Im ersten Ansatz sollte die Dimensionierung der erfolgreichen N-Grams reduziert werden, damit diese in einem umfangreicheren Ausmaß mit geringerem Speicher- und Rechenaufwand verarbeitet werden können. Eine potenzielle Lösung hierfür könnten Autoencoder sein, welche den Ähnlichkeitsvektor unter Berücksichtigung

¹⁰ GPU (Graphics Processing Unit) bedeutet zu Deutsch Grafikprozessor. Durch GPUs lassen sich künstliche neuronale Netze besonders effizient verarbeiten.

Grafik 5

Schwellenwert-basierter Performancevergleich zweier Klassifikatoren



Anmerkungen: In der Gegenüberstellung wird die rein Schwellenwert-basierte Klassifikation des aussagekräftigsten Merkmals (Tfd-idf(5,5) Unternehmensbezeichnung) mit dem erfolgreichsten ML-Modell (Random-Forest-Modell) verglichen. In der Gegenüberstellung ist der PPV die entscheidende Metrik.

der wesentlichen Informationen reduziert abbilden. Es wurde zwar ein kleines Netzwerk mit effizienten Parametern gewählt, allerdings ließen sich die Berechnungen in keinem zurzeit zeitlich vertretbaren Rahmen umsetzen.

Um neue Merkmale zu generieren, wurde One-Shot-Learning¹¹ durch einen Transformer-basierten¹² Encoder¹³ in einem Metric-Learning-Ansatz¹⁴ erprobt. Anhand dieser Netzwerkarchitektur wird das Modell optimiert, Zei-

chenketten über die Kosinus-Ähnlichkeit¹⁵ richtig zuzuordnen. Aufgrund der eingeschränkten Hardware wurde hierbei erneut bewusst eine Architektur mit möglichst wenigen Parametern genutzt. In ersten Analysen an Teilstichproben ließ sich zwar eine Transferleistung des trainierten Modells erkennen, um diese Beobachtung allerdings zu bestätigen, müssten weitere Untersuchungen durchgeführt werden.

11 One-Shot-Learning ist eine Methode aus dem maschinellen Lernen. Üblicherweise wird ein ML-Modell mit möglichst vielen Beispielen aus einer Klasse trainiert. Stehen beispielsweise nur wenige oder nur ein Beispiel für eine Klasse bereit, kann One-Shot-Learning eine geeignete Lösung darstellen, um die Übereinstimmung zweier Proben zu verifizieren. Im vorliegenden Fall steht nur eine Unternehmensbezeichnung für jede Unternehmenseinheit zur Verfügung, weshalb das One-Shot-Learning als Methode gewählt wurde.

12 Transformer sind Modelle, die zu den künstlichen neuronalen Netzen gehören (Vaswani und andere, 2017). Basierend auf dieser Architektur wurden im Bereich des Neuro-Linguistischen Programmierens (NLP) populäre Modelle wie GPT, Bert, Claude, Gemini oder Mistral entwickelt (Zhao und andere, 2024).

13 Bei dem Encoder handelt es sich um einen essenziellen Block aus der Transformer-Architektur. In diesem Fall wurde der Encoder so gebaut, dass er Zeichenketten in semantisch aussagekräftige Vektoren überführt, anhand welcher Aussagen über die Ähnlichkeit der Zeichenketten getätigt werden können.

14 Das Metric Learning wird in diesem Fall verwendet, um das passende Abstandsmaß zwischen Datenpunkten zu erlernen.

15 Mit der Kosinus-Ähnlichkeit lassen sich zwei Vektoren vergleichen.

5

Fazit


Die Experimente veranschaulichen, dass trotz der teils herausfordernden Datenlage weitere Arbeiten an einer automatisierten Verknüpfung der Quellregister gewinnbringend sein können und Verfahren des maschinellen Lernens hierbei eine wichtige Rolle zukommen wird.

Um den Unsicherheiten in den Daten entgegenzuwirken, wurden die Attribute schrittweise verarbeitet. Dabei erwies es sich als geeignetes Vorgehen, zu Beginn die robusteren Attribute (Postleitzahl und Hausnummer) für das Blocking heranzuziehen und dann mit anspruchsvolleren Ähnlichkeitsmaßen in der fehleranfälligen Unternehmensbezeichnung die Zuordnung zu verfeinern. Die Straßennamen erwiesen sich hingegen in den Experimenten in nicht normierter Form als ungeeignet. Künftig könnten daher geeignete Normierungsregeln und eine entsprechende Qualitätssicherung potenziell vorteilhaft für die Verknüpfung der Straßennamen sein. Das Blocking durch die Postleitzahl und die Hausnummer führte zu einem Verlust an Verknüpfungen, dieser wurde in der Vorstudie allerdings aus Ressourcengründen hingenommen. Daher wäre es denkbar, dass sich künftig das Blocking optimieren lässt, indem die Daten vor dem Blocking bereinigt werden, und dass durch eine umfangreichere IT-Infrastruktur Verknüpfungen durch weniger Blocking durchgeführt werden können. Wie auch in Grafik 2 deutlich wurde, wäre in weiteren Untersuchungen eine Vorverarbeitung in Form einer Adressnormierung insbesondere für die Straßennamen gewinnbringend. Darüber hinaus könnten künftig Labels aus dem Echtbetrieb die Datenqualität und den Umfang der Trainingsdaten bereichern.

Die Experimente zu den unterschiedlichen Ähnlichkeiten verdeutlichen, wie gewinnbringend es ist, wenn mehrere ähnlichkeitsbasierte Merkmale und insbesondere auch 5-Grams in Verbindung mit dem Tf-idf-Algorithmus in der Entscheidungsfindung hinzugezogen werden. Anhand dieser Ähnlichkeitsmerkmale erwies sich das Random-Forest-Modell unter Berücksichtigung des PPV mit einem Wert von 0,978 als verlässlichstes ML-Modell mit gleichzeitig niedrigem Standardfehler.

Besonders interessant dürfte für den laufenden Betrieb die Einbindung von Schwellenwerten bei der Aussage des

ML-Modells sein. Wie Grafik 5 zeigt, lassen sich dadurch unter der Berücksichtigung individueller Genauigkeiten entsprechend viele Verknüpfungen automatisiert verarbeiten. Mithilfe des Random-Forest-Modells lassen sich somit 95,1 % der Verknüpfungen mit einem positiven prädiktiven Wert von 0,95 vollautomatisiert auffinden.

Bei der Erzeugung der Merkmale wie auch beim Training der Modelle wurde im Hinblick auf den Bedarf einer tagesaktuellen Verarbeitung der Daten deutlich, dass die IT-Infrastruktur entscheidend für Durchsatz, Methodik und Genauigkeit ist. Unter diesen Gesichtspunkten lässt sich der Einsatz von maschinellem Lernen in der Vorstudie zum Basisregister als gewinnbringende Möglichkeit für die automatisierte Verknüpfung der Quellregister betrachten. Gleichwohl gilt es zu beachten, dass die Daten nicht registrierter Unternehmen (das sind unter anderem natürliche Personen) in der Vorstudie nur simuliert werden konnten, sodass hier weiterer Forschungsbedarf besteht. 

LITERATURVERZEICHNIS

Aghamohamadi, Zhina/Rezaei Ghahroodi, Zahra. *Record Linkage with Machine Learning Methods*. In: Journal of Statistical Sciences. Jahrgang 16. Ausgabe 1/2022, Seite 1 ff. [Zugriff am 8. Mai 2024]. Verfügbar unter: jss.irstat.ir

Bank, Dor/Koenigstein, Noam/Giryes, Raja. *Autoencoders*. In: CoRR, abs/2003.05991. 2020. [Zugriff am 8. Mai 2024]. Verfügbar unter: [arxiv.org](https://arxiv.org/abs/2003.05991)

BMWK (Bundesministerium für Wirtschaft und Klimaschutz). *Gesetzentwurf zur Umsetzung des Basisregisters für Unternehmensstammdaten mit bundeseinheitlicher Wirtschaftsnummer*. 2021. [Zugriff am 16. Januar 2024]. Verfügbar unter: www.bmwk.de

Christen, Peter. *The Data Matching Process*. In: Christen, Peter. Data Matching. Data-Centric Systems and Applications. Berlin, Heidelberg 2012. DOI: [10.1007/978-3-642-31164-2_2](https://doi.org/10.1007/978-3-642-31164-2_2)

de Bruin, Jonathan. *Record Linkage Toolkit Documentation*. 2022. [Zugriff am 8. Mai 2024]. Verfügbar unter: buildmedia.readthedocs.org

Géron, Aurélien. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. Zweite Auflage. 2019. [Zugriff am 8. Mai 2024]. Verfügbar unter: www.oreilly.com

Hasti, Trevor/Tibshirani, Robert/Friedman, Jerome. *The Elements of Statistical Learning*. Zweite Auflage. 2009.

Yujian, Li/Bo, Liu. *A normalized Levenshtein distance metric*. In: IEEE transactions on pattern analysis and machine intelligence. Jahrgang 29. Ausgabe 6/2007, Seite 1091 ff. DOI: [10.1109/TPAMI.2007.1078](https://doi.org/10.1109/TPAMI.2007.1078)

Li, Pengpeng/Luo, An/Liu, Jiping/Wang, Yong/Zhu, Jun/Deng, Yue/Zhang, Junjie. *Bidirectional Gated Recurrent Unit Neural Network for Chinese Address Element Segmentation*. In: isprs International Journal of Geo-Information. Ausgabe 9/2020, Seite 635 ff. DOI: [10.3390/ijgi9110635](https://doi.org/10.3390/ijgi9110635)

Mueller, Jonas/Thyagarajan, Aditya. *Siamese Recurrent Architectures for Learning Sentence Similarity*. In: Proceedings of the AAAI Conference on Artificial Intelligence. Jahrgang 30. Ausgabe 1/2016. doi.org

Pedregosa, Fabian/Varoquaux, Gaël/Gramfort, Alexandre/Michel, Vincent/Thirion, Bertrand/Grisel, Olivier/Blondel, Mathieu/Prettenhofer, Peter/Weiss, Ron/Dubourg, Vincent/Vanderplas, Jake/Passos, Alexandre/Cournapeau, David/Brucher, Matthieu/Perrot, Matthieu/Duchesnay, Édouard. *Scikit-learn: Machine Learning in Python*. In: Journal of Machine Learning Research. Ausgabe 12/2011, Seite 2825 ff. [Zugriff am 13. Mai 2024]. Verfügbar unter: scikit-learn.org

LITERATURVERZEICHNIS

Santos, Rui/Murrieta-Flores, Patricia/Calado, Pável/Martins, Bruno. *Toponym matching through deep neural networks*. In: International Journal of Geographical Information Science. Jahrgang 32. Ausgabe 2/2018, Seite 324 ff.

DOI: [10.1080/13658816.2017.1390119](https://doi.org/10.1080/13658816.2017.1390119)

Schnell, Rainer. *Maschinelles Lernen für Record Linkage*. Technischer Bericht für das Statistische Bundesamt. 2021.

Shan, Shuangli/Li, Zhixu/Quiang, Yang/Liu, An/Xu, Jiajie/Chen, Zhigang. *DeepAM: Deep Semantic Address Representation for Address Matching*. In: Shao, Jie und andere. Web and Big Data. 2019. Lecture Notes in Computer Science.

Ausgabe 11641. doi.org

SPD; Bündnis 90/Die Grünen und FDP. *Mehr Fortschritt wagen – Bündnis für Freiheit, Gerechtigkeit und Nachhaltigkeit*. Koalitionsvertrag 2021-2025. [Zugriff am 16. Mai 2024]. Verfügbar unter: www.bundesregierung.de

Vaswani, Ashish/Shazeer, Noam/Parmar, Niki/Uszkoreit, Jakob/Jones, Llion/Gomez, Aidan N./Kaiser, Lukasz/Polosukhin, Illia. *Attention Is All You Need*. In: CoRR, abs/1706.03762. 2017. [Zugriff am 14. Mai 2024]. Verfügbar unter: arxiv.org

Xu, Liuchang/Mao, Ruichen/Zhang, Chengkun/Wang, Yuanyuan/Zheng, Xinyu/Xue, Xingyu/Xia, Fang. *Deep Transfer Learning Model for Semantic Address Matching*. In: Applied Sciences. Ausgabe 12.19/2022. doi.org

Zhao, Wayne Xin/Zhou, Kun/Li, Junyi/Tang, Tianyi/Wang, Xiaolei/Hou, Yupeng/Min, Yingqian/Zhang, Beichen/Zhang, Junjie/Dong, Zican/Du, Yifan/Yang, Chen/Chen, Yushuo/Chen, Zhipeng/Jiang, Jinhao/Ren, Ruiyang/Li, Yifan/Tang, Xinyu/Liu, Zikang/Liu, Peiyu/Nie, Jian-Yun/Wen, Ji-Rong. *A Survey of Large Language Models*. In: CoRR, abs/2303.18223. 2023. [Zugriff am 14. Mai 2024]. Verfügbar unter: arxiv.org

RECHTSGRUNDLAGEN

Abgabenordnung in der Fassung der Bekanntmachung vom 1. Oktober 2002 (BGBl. I Seite 3866; 2003 I Seite 61), die zuletzt durch Artikel 14 des Gesetzes vom 27. März 2024 (BGBl. I Nr. 108) geändert worden ist.

Gesetz zur Errichtung und Führung eines Registers über Unternehmensbasisdaten und zur Einführung einer bundeseinheitlichen Wirtschaftsnummer für Unternehmen (Unternehmensbasisdatenregistergesetz – UBRG) vom 9. Juli 2021 (BGBl. I Seite 2506), das zuletzt durch Artikel 1 des Gesetzes vom 22. Dezember 2023 (BGBl. I Nr. 404) geändert worden ist.

Dr. Stefan Linz

leitet das Referat „Konjunkturindizes, Saisonbereinigung“ des Statistischen Bundesamtes.

Luis Federico Flores

ist als Referent im Referat „Konjunkturindizes, Saisonbereinigung“ des Statistischen Bundesamtes für die Methodik der Indexberechnung und Saisonbereinigung zuständig.

Peter Mehlhorn

ist seit 1986 in verschiedenen Bereichen der Unternehmensstatistik des Statistischen Bundesamtes tätig und seit 1999 für die Berechnung des Auftragseingangs- und Umsatzindex in der Industrie zuständig.

UMSTELLUNG DER UMSATZ-, AUFTRAGSEINGANGS- UND AUFTRAGSBESTANDSINDIZES IM VERARBEITENDEN GEWERBE AUF DAS BASISJAHR 2021

Stefan Linz, Luis Federico Flores, Peter Mehlhorn

➤ **Schlüsselwörter:** Umbasierung – Konjunkturindizes – Produzierendes Gewerbe – Indexgewichte – Gewichtungsstruktur – Unternehmensstatistiken

ZUSAMMENFASSUNG

Im Berichtsmonat Januar 2024 hat beim Umsatzindex für den Bergbau und das Verarbeitende Gewerbe sowie bei den Auftragseingangs- und Auftragsbestandsindizes im Verarbeitenden Gewerbe das neue Basisjahr 2021 turnusmäßig das bisher geltende Basisjahr 2015 abgelöst. Neben der Anpassung des Basisjahres als Bezugsgröße der Indizes wurden die Indexgewichte aktualisiert. Dieser Aufsatz beschreibt die neuen Gewichtungsstrukturen.

➤ **Keywords:** *rebasing – short-term economic indices – industry – weights – weighting structure – business statistics*

ABSTRACT

As of the reference month of January 2024, the index of turnover in mining and manufacturing as well as the indices of new orders and the stock of orders in manufacturing have been rebased from the previous base year of 2015 to 2021, as is regular practice. In addition to the change of the base year as the reference period of the indices, the index weights were updated. This article describes the new weighting structures.

1

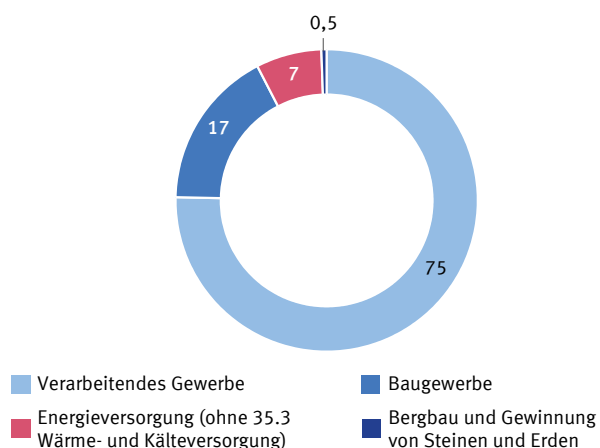
Einleitung

Die amtliche Statistik stellt mit den Statistiken zum Produzierenden Gewerbe monatlich Konjunkturindizes zur wirtschaftlichen Leistung der Industriebetriebe in Deutschland bereit, welche die Entwicklung von Produktion, Umsätzen, Auftragseingängen und Auftragsbeständen beschreiben. Mit Veröffentlichung der Ergebnisse für den Berichtsmonat Januar 2024 wurden die Konjunkturindizes auf das Basisjahr 2021 umgestellt. In diesem Aufsatz geht es um den Umsatzindex für den Bergbau und das Verarbeitende Gewerbe sowie die Auftragseingangs- und Auftragsbestandsindizes im Verarbeitenden Gewerbe. Auch beim Produktionsindex für das Produzierende Gewerbe wurden die Gewichtungsstrukturen aktualisiert und im Bereich des Baugewerbes wurde die zugrunde liegende Branchenklassifikation umgestellt (Linz und andere, 2024).

Das Verarbeitende Gewerbe sowie der Bergbau und die Gewinnung von Steinen und Erden erwirtschaften zusammen rund 76 % der Wertschöpfung des Produzierenden Gewerbes, wobei der Bergbau einen nur sehr geringen Anteil einnimmt. [↗ Grafik 1](#)

Grafik 1

Wertschöpfungsanteile in den Abschnitten¹ des Produzierenden Gewerbes in %



1 Abschnitte der Klassifikation der Wirtschaftszweige, Ausgabe 2008 (WZ 2008).

Die Zuordnung der wirtschaftlichen Tätigkeiten zu Wirtschaftszweigen erfolgt nach der [Klassifikation der Wirtschaftszweige, Ausgabe 2008 \(WZ 2008\)](#), die wiederum auf der [statistischen Systematik der Wirtschaftszweige in der Europäischen Gemeinschaft](#) (NACE Rev. 2) aufbaut.

Nach einer kurzen Erläuterung zur Konstruktion der Indizes und zur Basisjahrumstellung in Kapitel 2 skizziert Kapitel 3 das konjunkturelle Umfeld in den Jahren 2021 und 2015. Die folgenden Kapitel 4 bis 6 beschreiben für den Umsatzindex, den Auftragseingangsindex sowie den Auftragsbestandsindex im Einzelnen die Änderungen der Gewichtungsstrukturen, die sich mit der Umstellung auf das neue Basisjahr 2021 ergeben haben.

2

Indexkonstruktion und Basisumstellung

Die Indizes zum Umsatz im Bergbau und Verarbeitenden Gewerbe sowie zum Auftragseingang und Auftragsbestand im Verarbeitenden Gewerbe geben jeweils das Verhältnis der aktuellen Werte zu den entsprechenden Werten im Basisjahr an. Ein Indexwert von 110 bedeutet zum Beispiel beim Umsatzindex, dass das Umsatzvolumen im aktuellen Monat um 10 % höher liegt als im Durchschnitt des Basisjahres. In Deutschland werden die Konjunkturindizes als Festbasisindizes berechnet, die in der Regel alle fünf Jahre auf ein neues Basisjahr umgestellt werden. Die Basisumstellung umfasst bei diesen Indizes die folgenden zwei Aspekte:¹

(1) Aus praktischen Gründen wird die Bezugsgröße der Indizes auf das neue Basisjahr umgestellt, damit die Zahlenwerte einfach zu interpretieren sind und nicht zu groß werden. Ab Berichtsmonat Januar 2024 gibt etwa der Umsatzindex nicht mehr das Verhältnis des Umsatzvolumens zum Durchschnitt des Jahres 2015, sondern zu dem des Jahres 2021 an. Die in Tabellen und Grafiken verwendete Kurzbezeichnung des Basisjahres wird entsprechend von „2015 = 100“ auf „2021 = 100“ geändert.

1 Methodische Umstellungen, die häufig im Zusammenhang mit der Basisumstellung eingeführt werden, gab es bei diesen Indizes nicht. Eine ausführliche Darstellung der Konstruktionsprinzipien, Geltungsbereiche und Funktionen der Konjunkturindizes im Bereich des Produzierenden Gewerbes findet sich in Linz und andere (2008).

(2) Mit der Gewichtungsstruktur wird festgelegt, mit welchem Gewicht die Indexergebnisse für einzelne Wirtschaftszweige in den jeweiligen Gesamtindex eingehen. Die Gewichtungsstruktur bezieht sich auf die wirtschaftlichen Verhältnisse im Basisjahr. Sie wird mit der Basisjahrumstellung aktualisiert und zwischen den Basisjahren konstant gehalten. Die aktuellen Gewichte der Indizes beziehen sich nun auf das Jahr 2021 anstelle von 2015, die Gewichte für frühere Zeitpunkte bleiben unverändert.

Für die Berechnung des Umsatzindex ist seit Januar 2024 die Verordnung über europäische Unternehmensstatistiken (EBS-Verordnung) maßgeblich.¹² Die genauen Anforderungen sind in einer Durchführungsverordnung geregelt. Dort ist festgelegt: „Das erste Basisjahr ist 2015, das zweite Basisjahr ist 2021 und das dritte Basisjahr ist 2025. Danach basieren die Mitgliedstaaten die Indizes alle fünf Jahre um, wobei sie die mit 0 oder 5 endenden Jahre als Basisjahre verwenden. Sämtliche Indizes sind innerhalb von drei Jahren nach Ablauf des neuen Basisjahrs auf dieses neue Jahr umzubasieren.“¹³ Die Wahl des Basisjahres 2021 anstelle von 2020 ist verwaltungstechnisch begründet und beruht darauf, dass die EBS-Verordnung erst ab Januar 2021 wirksam wurde. Für die mit der Verordnung neu eingeführten [Dienstleistungsproduktionsindizes](#) bestand somit erst ab 2021 eine gesetzliche Grundlage für die Datenerhebung. Aus Gründen der Vergleichbarkeit wurde für alle Konjunkturindizes der EBS-Verordnung das Jahr 2021 als Basisjahr festgelegt.

Für die Auftragseingangs- und Auftragsbestandsindizes im Verarbeitenden Gewerbe gibt es keine europäische Verordnung, sie werden aufgrund des Gesetzes über die Statistik im Produzierenden Gewerbe berechnet. Auch für sie wurde wegen der Vergleichbarkeit als Basisjahr 2021 festgelegt.

12 Verordnung (EU) 2019/2152 des Europäischen Parlaments und des Rates vom 27. November 2019 über europäische Unternehmensstatistiken, zur Aufhebung von zehn Rechtsakten im Bereich Unternehmensstatistiken. Zuvor war die sogenannte Konjunkturstatistik-Verordnung relevant: Verordnung (EG) Nr. 1165/98 des Rates vom 19. Mai 1998 über Konjunkturstatistiken.

13 Durchführungsverordnung (EU) 2020/1197 der Kommission vom 30. Juli 2020 zur Festlegung technischer Spezifikationen und Einzelheiten nach der Verordnung (EU) 2019/2152 des Europäischen Parlaments und des Rates über europäische Unternehmensstatistiken, zur Aufhebung von zehn Rechtsakten im Bereich Unternehmensstatistiken, hier: Anhang VII Absatz 2.

3

Konjunkturelles Umfeld in den Jahren 2015 und 2021

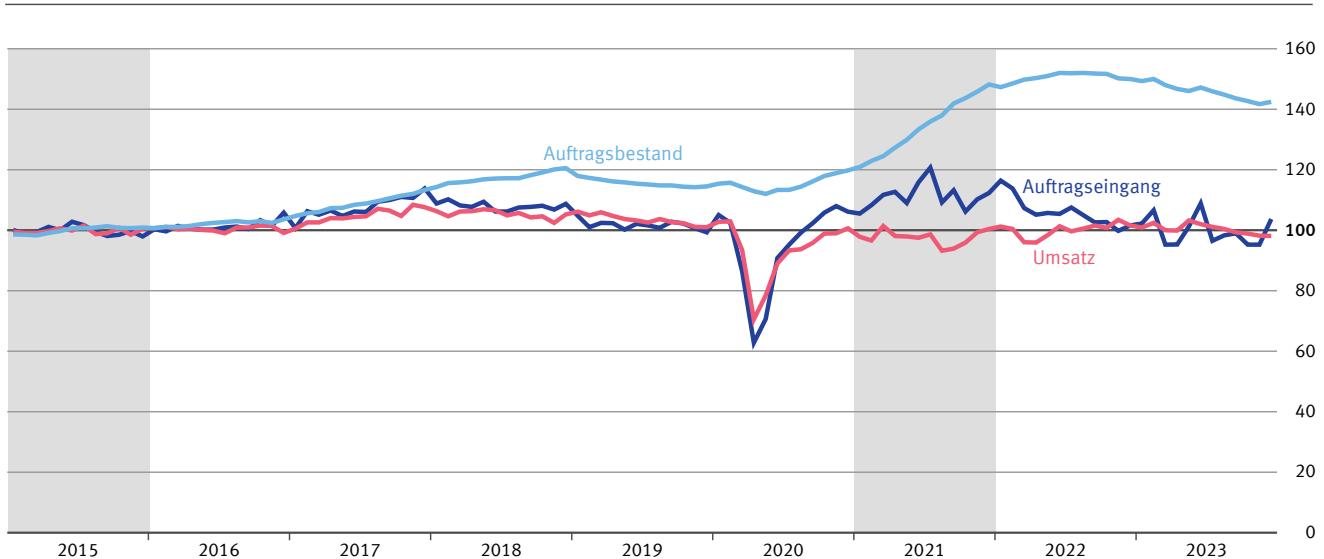
➤ Grafik 2 zeigt die Entwicklung der Umsatz-, Auftragseingangs- und Auftragsbestandsindizes. Dabei ist zu beachten, dass der Umsatzindex alle Wirtschaftszweige des Verarbeitenden Gewerbes und Bergbaus umfasst, während sich die Auftragseingangs- und Auftragsbestandsindizes nur auf ausgewählte Wirtschaftszweige beziehen. Um die Veränderungen von 2015 bis 2021 zu betrachten, werden die noch nicht umbasierten Indizes dargestellt (2015 = 100).

In der unterschiedlichen Entwicklung der Indizes im Jahr 2021, dem neuen Basisjahr der Indizes, spiegeln sich die wirtschaftlichen Auswirkungen der Coronakrise wider. In der ersten Phase der Pandemie wurden mit den Lockdowns über viele Branchen hinweg die Produktions- und Transportkapazitäten heruntergefahren, in Deutschland ebenso wie in vielen anderen Regionen der Welt. Zugleich war das öffentliche Leben eingeschränkt, was eine Verschiebung der privaten Konsumstruktur von Dienstleistungen, zum Beispiel Reisedienstleistungen oder Gastronomie, zu industriell gefertigten Konsumgütern wie Elektronikgeräten oder Möbeln bewirkt hat. Der mit der Lockerung der Lockdowns einsetzende, international synchrone Konjunkturaufschwung erzeugte eine starke Nachfrage nach Industrieprodukten. Durch die pandemiebedingten Einschränkungen waren jedoch in vielen Regionen noch die Produktionskapazitäten und die internationalen Transportkapazitäten beeinträchtigt. Die plötzliche und teilweise unerwartete Nachfrage nach Konsumgütern im Anschluss an die erste Pandemiephase konnte durch die Industrie nicht vollständig bedient werden, insbesondere fehlten Rohstoffe und Elektronikbauteile (Linz und andere, 2022). Der Auftragseingangsindex, als Indikator für die Nachfrageentwicklung, erreichte im Jahr 2021 einen Höchststand, während Materialmangel dazu führte, dass die Produktion und damit die Umsätze der Industriebetriebe stagnierten. Die Schere zwischen Nachfrage und Angebot öffnete sich und in den Betrieben stauten sich die Aufträge auf. In der Folge stieg der Auftragsbestand, also der Bestand an nicht abgearbeiteten Aufträgen, im Verlauf des Jahres 2021 in erheblichem Umfang an. Dass der Materialmangel in den verschiedenen Wirtschafts-

Umstellung der Umsatz-, Auftragseingangs- und Auftragsbestandsindizes im Verarbeitenden Gewerbe auf das Basisjahr 2021

Grafik 2

Preis- und saisonbereinigte Konjunkturindizes im Verarbeitenden Gewerbe
2015 = 100



zweigen unterschiedlich stark ausgeprägt war, führte zu Verschiebungen in der relativen wirtschaftlichen Bedeutung der verschiedenen Wirtschaftszweige. Die Verschiebung der relativen Bedeutung spiegelt sich in Veränderungen der Gewichtung der hier beschriebenen Wirtschaftszweigindizes wider, auf die später im Text eingegangen wird.

Der Wettbewerb der Industriebetriebe um Rohstoffe, Vorprodukte und Transportdienstleistungen ließ auch die Einfuhrpreise für diese Güter stark ansteigen. Hinzu kamen erhebliche Preissteigerungen bei Energie, denn die plötzliche Erholung der Weltwirtschaft nach der Coronakrise führte zu einem unerwartet hohen Energiebedarf. Weitere Ursachen für hohe Energiepreise im Jahr 2021 (dem Jahr vor dem Angriff Russlands auf die Ukraine) waren neben der steigenden Nachfrage auch gedrosselte Gasexportmengen aus Russland, wenig gefüllte Gasspeicher und zusätzliche Kosten für CO₂-Emissionszertifikate. In einigen Wirtschaftszweigen konnten die Industriebetriebe die Preissteigerungen beim Einkauf von Energie und Rohstoffen auf ihre Fertigprodukte aufschlagen. So gaben in einer Umfrage des ifo Instituts im Sommer 2021 die befragten Unternehmen in den Bereichen Chemie und Metallherstellung an, dass sie Steigerungen ihrer Einkaufspreise vollständig an die Abnehmer ihrer Produkte weitergeben konnten (Wohlrabe, 2021). In anderen Bereichen, etwa bei der

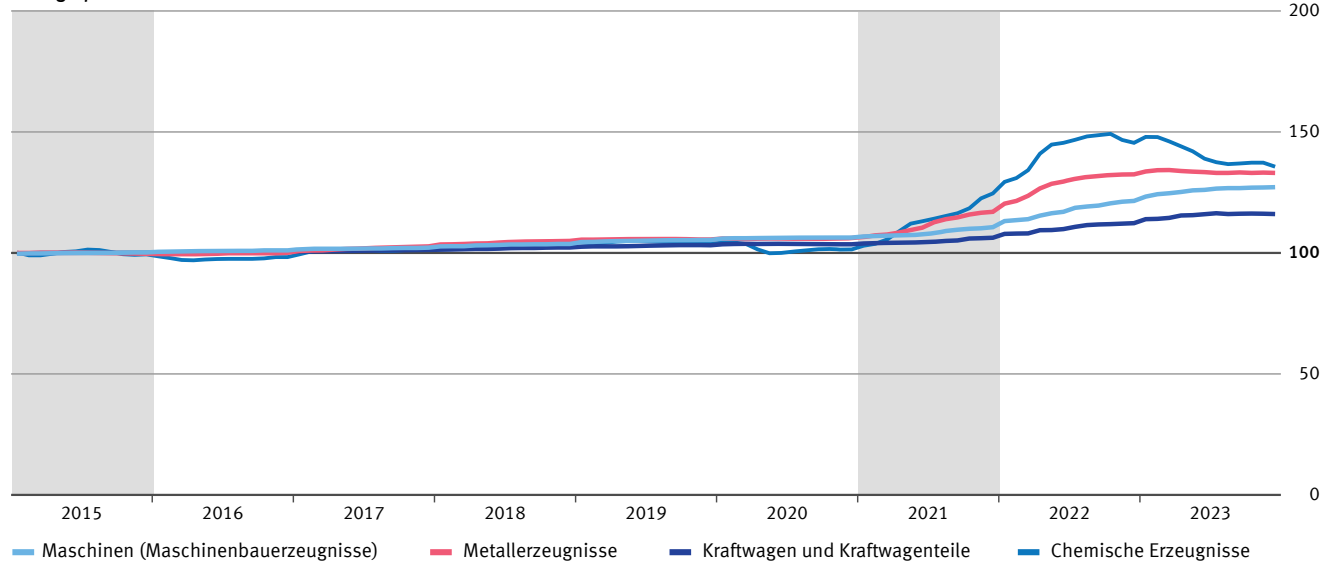
Herstellung von Industrieprodukten wie Maschinen oder Kraftwagen, sind die Absatzpreise in geringerem Maße oder erst mit Verzögerung gestiegen. Die Preisentwicklung seit dem Jahr 2015 für ausgewählte Wirtschaftsbereiche und Energieträger zeigt [Grafik 3](#) auf Seite 84.

Auch die unterschiedlichen Preisentwicklungen in den verschiedenen Wirtschaftszweigen haben zu den Verschiebungen in den relativen Prozentgewichten der Konjunkturindizes beigetragen, wie unten näher erläutert wird.

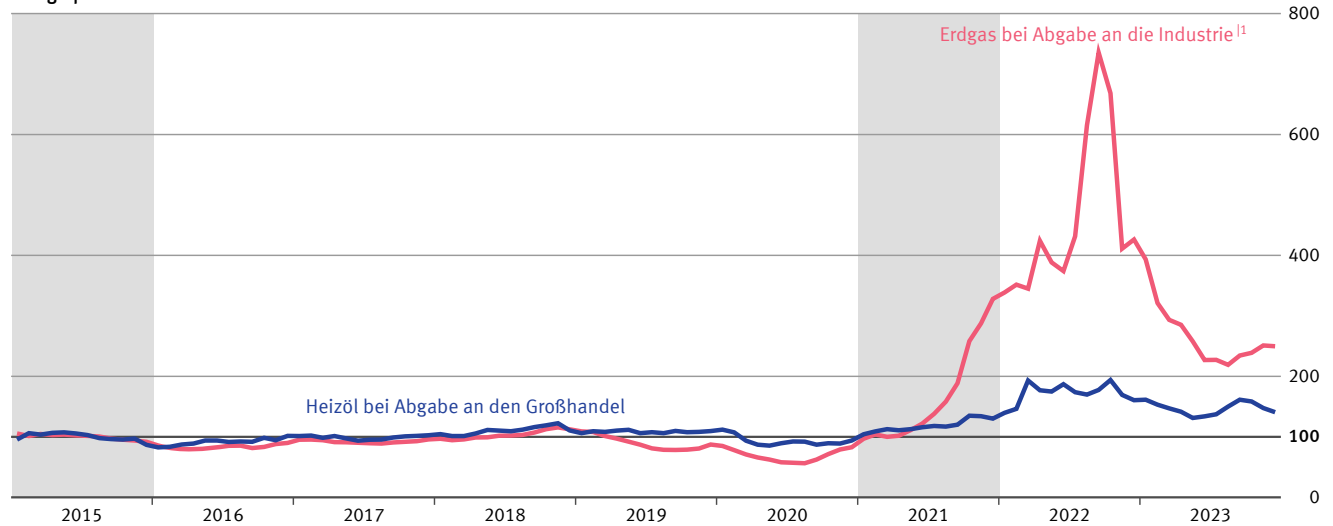
Grafik 3

Entwicklung von Erzeuger- und Energiepreisen in Deutschland
2015 = 100

Erzeugerpreise



Energiepreise



1 Jahresabgabe über 500 000 MWh.

4

Umsatzindex

Der Umsatzindex misst die monatliche Entwicklung der preisbereinigten Umsätze in den industriellen Wirtschaftszweigen. Der Umsatz umfasst die Summe der Rechnungsendbeträge (ohne Umsatzsteuer) aus Lieferungen und Leistungen an andere Betriebe oder Unternehmen. Lieferungen und Leistungen zwischen Betrieben desselben Unternehmens werden bei der Ermittlung des Umsatzes nicht berücksichtigt. Die Daten resultieren aus einer monatlichen Erhebung bei Industriebetrieben mit 50 und mehr Beschäftigten, das waren im Jahr 2023 knapp 23 000 Betriebe. Bei der Datenerhebung wird unterschieden zwischen Umsätzen, die von den deutschen Industriebetrieben durch Verkäufe an Unternehmen im Inland, aus der Eurozone oder dem restlichen Ausland erzielt wurden.

Der Umsatzindex bezieht sich auf die 245 Klassen der Abschnitte B und C der WZ 2008. In den Wirtschaftszweigen wird jeweils ein Index für den Umsatz im Inland, mit der Eurozone und mit dem restlichen Ausland ausgewiesen. Der Gesamtindex für das Verarbeitende Gewerbe wird als gewichteter Mittelwert der untergeordneten Wirtschaftszweigindizes berechnet. Für die Gewichtung werden dieselben Umsatzdaten herangezogen, die für die Messgröße bei der Berechnung der Wirtschaftszweigindizes verwendet werden. Die Gewichte ergeben sich aus der nominalen (nicht preisbereinigten) Umsatzsumme, berechnet als Durchschnitt über alle Monate des Basisjahres in den betreffenden Wirtschaftszweigen. Die Gesamtheit der zugrunde liegenden Gewichte wird als Wägungsschema bezeichnet, welches sich nun auf das Jahr 2021 bezieht. In [Tabelle 1](#) ist das neue Wägungsschema für den Umsatzindex für den Bergbau und das Verarbeitende Gewerbe zusammenfassend dargestellt.⁴ Ebenfalls angegeben ist die Gewichtungsstruktur für das vorherige Basisjahr 2015.

Bei den Veränderungen gegenüber dem Jahr 2015 fällt insbesondere der Rückgang des Wägungsanteils im Bereich 29 Herstellung von Kraftwagen und Kraftwagen-

teilen auf. Zum Teil kam der Rückgang des relativen Gewichtsanteils der Automobilindustrie dadurch zustande, dass die Umsätze in der Automobilbranche im Jahr 2015 überdurchschnittlich hoch gelegen haben, da sie besonders in den Jahren 2014 und 2015 stark gestiegen sind. Dieser Wachstumstrend schwächte sich ab dem Jahr 2018 deutlich ab. Hinzu kam, dass sowohl der Produktionsrückgang in der ersten Phase der Coronakrise als auch die anschließenden Produktionsbehinderungen durch Materialknappheit in der Automobilindustrie besonders stark ausgeprägt waren. Mikrochips gehörten zu den in Umfragen häufig genannten Engpassbauteilen; deren Verknappung behinderte die Pkw-Produktion im Jahr 2021 deutlich (Wohlrabe, 2021, hier: Seite 61). In der Folge lag der Umsatz aus der Herstellung von Kraftwagen und Kraftwagen teilen im neuen Basisjahr 2021 deutlich niedriger als im Basisjahr 2015. Der Rückgang des Gewichts der Automobilbranche im Umsatzindex spiegelt also einerseits einen längerfristigen Rückgang des Wachstumstrends und andererseits einen Sondereffekt infolge der Materialknappheit im Jahr 2021 wider.

Eine ähnliche Entwicklung zeigte sich im Wirtschaftszweig 28 Maschinenbau. Auch hier waren eine Abschwächung des längerfristigen Entwicklungstrends ab dem Jahr 2018 und zusätzlich im Jahr 2021 starke Produktionsbehinderungen durch Materialmangel zu beobachten. Im Gegensatz zur Automobilindustrie befand sich der Maschinenbau im früheren Basisjahr 2015 jedoch nicht in einer ausgeprägten konjunkturellen Hochphase und der rückläufige Entwicklungstrend ab dem Jahr 2018 fiel weitaus schwächer aus als in der Kraftfahrzeug-Industrie. Entsprechend ist der Umsatzrückgang im Vergleich zum Jahr 2015 weniger stark als bei der Herstellung von Kraftfahrzeugen, der relative Wägungsanteil ist im Maschinenbau nur leicht gesunken.

Bei der Herstellung von chemischen Erzeugnissen verlief die Entwicklung nach der Coronakrise etwas anders. Hier war das Problem des Materialmangels weniger verbreitet und aufgrund der hohen Nachfrage gerade nach Vorleistungsgütern, zu denen die chemischen Erzeugnisse gehören, konnten die Betriebe ihre Produktion im Jahr 2021 ausweiten. In diesem Umfeld war auch die Weitergabe von Preissteigerungen bei den Rohstoffen und Vorprodukten an die Kunden der Chemieindustrie möglich. Neben der guten Auftragslage sowie den relativ geringen Produktionsbehinderungen durch Materialmangel

⁴ Es handelt sich um eine zusammenfassende Darstellung, weil nur die Wägungsanteile der übergeordneten Wirtschaftszweig-Abteilungen der WZ 2008 angegeben sind. Die Berechnung des Umsatzindex erfolgt auf der detaillierteren Ebene von Wirtschaftszweig-Klassen.

Tabelle 1

Wägungsschema zum Umsatzindex für den Bergbau und das Verarbeitende Gewerbe

	Insgesamt		Inland		Eurozone		Übriges Ausland	
	2015	2021	2015	2021	2015	2021	2015	2021
	%							
B + C Bergbau und Gewinnung von Steinen und Erden sowie Verarbeitendes Gewerbe	100	100	50,77	50,87	20,18	20,29	29,05	28,85
B Bergbau und Gewinnung von Steinen und Erden	0,40	0,30	0,13	0,25	0,01	0,02	0,00	0,02
05 Kohlenbergbau	0,14	0,08	0,13	0,08	0,01	0,00	0,00	0,00
06 Gewinnung von Erdöl und Erdgas	0,13	0,07	0,01	0,07	0,01	0,00	0,01	0,00
08 Gewinnung von Steinen und Erden, sonstiger Bergbau	0,11	0,13	0,01	0,10	0,01	0,02	0,01	0,02
09 Erbringung von Dienstleistungen für den Bergbau und für die Gewinnung von Steinen und Erden	0,02	0,01	0,00	0,01	0,00	0,00	0,00	0,00
C Verarbeitendes Gewerbe	99,60	99,70	50,40	50,61	20,14	20,26	29,03	28,83
10 Herstellung von Nahrungs- und Futtermitteln	8,44	8,64	6,48	6,45	1,39	1,49	0,57	0,70
11 Getränkeherstellung	1,18	1,14	1,03	0,98	0,08	0,07	0,06	0,09
12 Tabakverarbeitung	0,44	0,32	0,34	0,26	0,05	0,04	0,05	0,03
13 Herstellung von Textilien	0,67	0,61	0,33	0,29	0,18	0,17	0,15	0,14
14 Herstellung von Bekleidung	0,45	0,29	0,30	0,18	0,09	0,06	0,07	0,05
15 Herstellung von Leder, Lederwaren und Schuhen	0,18	0,15	0,12	0,10	0,03	0,03	0,03	0,02
16 Herstellung von Holz-, Flecht-, Korb- und Korkwaren (ohne Möbel)	1,05	1,37	0,78	0,98	0,16	0,20	0,11	0,19
17 Herstellung von Papier, Pappe und Waren daraus	2,39	2,39	1,44	1,41	0,56	0,61	0,39	0,37
18 Herstellung von Druckerzeugnissen; Vervielfältigung von bespielten Ton-, Bild- und Datenträgern	0,88	0,64	0,74	0,53	0,08	0,06	0,07	0,06
19 Kokerei und Mineralölverarbeitung	2,58	4,93	2,26	4,44	0,21	0,30	0,09	0,19
20 Herstellung von chemischen Erzeugnissen	7,82	8,40	3,21	3,29	2,01	2,36	2,59	2,76
21 Herstellung von pharmazeutischen Erzeugnissen	2,33	2,52	0,80	0,92	0,65	0,55	0,87	1,05
22 Herstellung von Gummi- und Kunststoffwaren	4,36	4,40	2,58	2,50	1,00	1,05	0,78	0,85
23 Herstellung von Glas und Glaswaren, Keramik, Verarbeitung von Steinen und Erden	1,79	1,95	1,23	1,35	0,28	0,30	0,27	0,30
24 Metallerzeugung und -bearbeitung	5,92	6,19	3,50	3,60	1,36	1,49	1,05	1,10
25 Herstellung von Metallerzeugnissen	6,13	6,22	4,06	4,04	1,11	1,14	0,98	1,05
26 Herstellung von Datenverarbeitungsgeräten, elektronischen und optischen Erzeugnissen	4,63	4,97	1,77	1,62	0,92	0,98	1,95	2,37
27 Herstellung von elektrischen Ausrüstungen	5,28	5,58	2,55	2,78	1,12	1,22	1,60	1,57
28 Maschinenbau	13,66	13,10	5,29	4,69	2,83	2,97	5,57	5,43
29 Herstellung von Kraftwagen und Kraftwagenteilen	21,25	17,72	7,48	6,01	4,04	3,24	9,71	8,46
30 Sonstiger Fahrzeugbau	2,96	2,83	1,02	1,08	1,12	1,02	0,81	0,73
31 Herstellung von Möbeln	1,15	1,08	0,79	0,72	0,22	0,23	0,14	0,12
32 Herstellung von sonstigen Waren	1,41	1,60	0,60	0,65	0,27	0,34	0,53	0,62
33 Reparatur und Installation von Maschinen und Ausrüstungen	2,65	2,65	1,70	1,73	0,36	0,34	0,60	0,59

Abschnitte und Abteilungen der Klassifikation der Wirtschaftszweige, Ausgabe 2008 (WZ 2008). – Differenzen zwischen Abschnitten und Abteilungssummen sind rundungsbedingt.

hat auch die Preisentwicklung zum Anstieg des relativen Wägungsanteils der Chemieindustrie im Umsatzindex beigetragen. Eine ähnliche Situation war im Bereich der Metallerzeugung und -bearbeitung zu beobachten, gleichwohl der Gewichtsanteil hier etwas weniger stark zugenommen hat.

Im Wirtschaftszweig 19 „Kokerei und Mineralölverarbeitung“ sind Preissteigerungen der wesentliche Grund für

den starken Anstieg des Gewichtsanteils im Umsatzindex. Bei der Mineralölverarbeitung wird der Rohölbedarf der deutschen Raffinerien überwiegend durch Importe gedeckt. Wie bei vielen anderen Rohstoffen und Vorprodukten waren auch bei Erdöl die Einfuhrpreise nach der Coronakrise stark angestiegen. Das hat dazu geführt, dass der Verkaufswert der Mineralölerzeugnisse in Deutschland und damit der nominale Umsatz der mineralölverarbeitenden Betriebe im Jahr 2021 stark ange-

Umstellung der Umsatz-, Auftragseingangs- und Auftragsbestandsindizes im Verarbeitenden Gewerbe auf das Basisjahr 2021

wachsen ist. Das (preisbereinigte) Umsatzvolumen im Bereich der „Kokerei und Mineralölverarbeitung“ war im Jahr 2021 hingegen etwas niedriger als 2015.

Inlands- und Auslandsumsätze

In Tabelle 1 sind auch die relativen Wägungsanteile der Absatzmärkte (Inland, Eurozone, übriges Ausland) angegeben. Bei der Aufteilung der Gewichtungsanteile auf Inlands- und Auslandsumsätze sind nur moderate Änderungen eingetreten. Insgesamt lagen die Umsätze der deutschen Industriebetriebe im Jahr 2021 rund 9 % höher als im Jahr 2015, eine ähnliche Umsatzsteigerung war sowohl für den Inlandsumsatz als auch für den in der Eurozone⁵ erzielten Umsatz deutscher Betriebe zu beobachten. Der Umsatz mit dem übrigen Ausland war im gleichen Zeitraum jedoch nur etwa 7 % höher. Der Wägungsanteil des im nicht zur Eurozone gehörenden Ausland erzielten Umsatzes ist damit leicht gesunken, während der im Inland und der mit der Eurozone erzielte Umsatz leicht angestiegen ist.

5 Die Eurozone bilden die Mitgliedstaaten der Europäischen Union mit dem Euro als offizieller Währung.

5

Auftragseingangsindex

Der Auftragseingangsindex gibt Auskunft über die monatliche Entwicklung des (preisbereinigten) Volumens der in den Unternehmen ausgewählter Wirtschaftszweige des Verarbeitenden Gewerbes jeweils neu eingegangenen Aufträge. Als Auftragseingänge gelten die im Berichtsmonat von den Betrieben fest akzeptierten Aufträge auf Lieferung selbst hergestellter oder in Lohnarbeit gefertigter Erzeugnisse. Auch hier wird unterschieden zwischen Aufträgen, die den deutschen Betrieben von Unternehmen im Inland, aus der Eurozone oder dem übrigen Ausland erteilt wurden.

In den Auftragseingangsindex werden nur die in [Tabelle 2](#) aufgeführten Abteilungen der WZ 2008 und deren untergeordnete Wirtschaftszweig-Klassen einbezogen.⁶ Die 128 Wirtschaftszweigindizes der einbezogenen Klassen werden jeweils für Aufträge aus dem Inland, der Eurozone und dem übrigen Ausland berechnet.

6 Die Auswahl der Wirtschaftszweige für den Auftragseingangsindex orientiert sich an einer früheren Auftragseingangserhebung im Europäischen Statistiksistem, die auf europäischer Ebene zwischenzeitlich abgeschafft wurde, in Deutschland jedoch wegen nationaler Nutzerbedarfe fortgeführt wird.

Tabelle 2

Zusammengefasstes Wägungsschema für den Auftragseingangsindex im Verarbeitenden Gewerbe

	Insgesamt		Inland		Eurozone		Übriges Ausland	
	2015	2021	2015	2021	2015	2021	2015	2021
	%							
Einbezogene Teile des Verarbeitenden Gewerbes	100	100	43,12	41,73	21,66	22,19	35,22	36,08
13 Herstellung von Textilien	0,91	0,78	0,45	0,36	0,25	0,23	0,21	0,19
14 Herstellung von Bekleidung	0,65	0,44	0,42	0,26	0,14	0,09	0,11	0,08
17 Herstellung von Papier, Pappe und Waren daraus	3,29	3,12	1,97	1,82	0,76	0,80	0,55	0,50
20 Herstellung von chemischen Erzeugnissen	10,27	10,69	4,27	4,14	2,67	2,97	3,34	3,58
21 Herstellung von pharmazeutischen Erzeugnissen	3,23	3,13	1,09	1,12	0,92	0,78	1,22	1,22
24 Metallerzeugung und -bearbeitung	7,67	8,25	4,58	4,66	1,82	2,03	1,28	1,57
25 Herstellung von Metallerzeugnissen	8,41	8,41	5,54	5,44	1,54	1,55	1,34	1,41
26 Herstellung von Datenverarbeitungsgeräten, elektronischen und optischen Erzeugnissen	6,40	7,32	2,44	2,51	1,30	1,50	2,64	3,31
27 Herstellung von elektrischen Ausrüstungen	7,43	7,78	3,57	3,89	1,58	1,58	2,29	2,31
28 Maschinenbau	19,15	19,93	7,27	6,82	4,06	4,58	7,80	8,53
29 Herstellung von Kraftwagen und Kraftwagenteilen	28,33	25,43	10,16	8,52	5,01	4,69	13,16	12,22
30 Sonstiger Fahrzeugbau	4,26	4,73	1,36	2,18	1,60	1,38	1,30	1,17

Abteilungen der Klassifikation der Wirtschaftszweige, Ausgabe 2008 (WZ 2008). – Differenzen in den Summen sind rundungsbedingt.

Die Datengewinnung, Indexberechnung, Preisbereinigung und die Berechnung des Wägungsschemas erfolgt analog zur Vorgehensweise beim Umsatzindex; für die Gewichtung wird hier der Mittelwert der Auftragseingänge über die Monate des Basisjahres herangezogen. Die aktuellen Gewichtungsstrukturen sowie die des Basisjahres 2015 sind in Tabelle 2 zusammenfassend angegeben.

Bei den Veränderungen der Wägungsanteile gegenüber dem Jahr 2015 ist beim AuftragseingangsindeX ein starker Rückgang im Bereich der Herstellung von Kraftwagen und Kraftwagenteilen zu beobachten. Die seit 2018 tendenziell rückläufige Konjunkturentwicklung in der Automobilindustrie in Deutschland ist auch bei den Auftragseingängen zu erkennen. Dass der Rückgang des Wägungsanteils im AuftragseingangsindeX etwas schwächer ausfällt als im Umsatzindex, dürfte sowohl auf den in der Kraftfahrzeug-Industrie besonders starken Produktionsrückgang während der Corona-Lockdowns im Jahr 2020 als auch auf den anschließenden Materialmangel im Jahr 2021 zurückzuführen sein – beides wirkt sich auf den Umsatz stärker aus als auf den Auftragseingang.

Der Bereich der Herstellung von Datenverarbeitungsgeräten, elektronischen und optischen Erzeugnissen befand sich im Jahr 2021 in einer deutlichen Aufschwungphase. Die Auftragseingänge wuchsen, während die Produktionsausweitung nicht Schritt halten konnte, denn auch hier waren die Betriebe stark mit fehlenden Inputmaterialien konfrontiert. Im AuftragseingangsindeX hat daher der Wägungsanteil stärker zugenommen als im Umsatzindex.

Im Maschinenbau ist der relative Wägungsanteil leicht gestiegen, was vor allem auf die Materialknappheit im Jahr 2021 in diesem Wirtschaftszweig zurückgeht. Die damit einhergehende Auftragshäufung hat den seit 2018 leicht rückläufigen längerfristigen Entwicklungstrend im Maschinenbau kompensiert.

Schließlich gibt es im Bereich der Chemieindustrie ebenso wie in der Metallerzeugung und -bearbeitung einen Anstieg des Wägungsanteils im AuftragseingangsindeX, der sich durch die deutlichen Preissteigerungen im Jahr 2021 erklärt.

Beim Vergleich der Wägungsanteile zwischen Auftrags- eingangs- und Umsatzindex ist auch zu berücksichtigen,

dass die Bezugsgrößen der Prozentgewichte unterschiedlich abgegrenzt sind. Wie oben erwähnt, wird im AuftragseingangsindeX die Auftragseingangssumme der ausgewählten Wirtschaftszweige als Bezugsgröße verwendet, während beim Umsatzindex alle industriellen Zweige einbezogen sind.

6

Auftragsbestandsindex

Für den Auftragsbestandsindex werden die gleichen Datenquellen, Methoden und Gliederungen verwendet wie beim AuftragseingangsindeX. Auch die Auswahl der einbezogenen Wirtschaftszweig-Klassen entspricht derjenigen beim AuftragseingangsindeX. Allerdings wird bei der Erhebung des Auftragsbestands nur zwischen Aufträgen aus dem Inland und dem Ausland unterschieden, es gibt keine weitere Unterscheidung zwischen Eurozone und dem übrigen Ausland. Für den Auftragsbestandsindex im Verarbeitenden Gewerbe werden wie beim AuftragseingangsindeX 126 Wirtschaftszweig-indizes berechnet, hier jeweils nur für Aufträge aus dem Inland und aus dem gesamten Ausland. Die Gewichte für den Auftragsbestandsindex ergeben sich aus der durchschnittlichen Auftragsbestandssumme in den Monaten des Basisjahres 2015. Eine zusammenfassende Gegenüberstellung der Gewichtungsstrukturen für die Basisjahre 2015 und 2021 enthält [Tabelle 3](#).

Beim Auftragsbestandsindex ist die Konzentration der Werte auf wenige große Wirtschaftszweige stark ausgeprägt, besonders bei den Aufträgen aus dem Ausland. Hier entfallen fast 70 % der gesamten Auftragsbestandssumme auf die drei größten Wirtschaftszweige Maschinenbau, Herstellung von Kraftwagen und Kraftwagenteilen und sonstiger Fahrzeugbau (Flugzeuge, Schiffe, Züge und Spezialfahrzeuge). In anderen Branchen, zum Beispiel bei der Herstellung von Bekleidung, chemischen Produkten oder Pharmazeutika, sind die Produkte weniger „maßgeschneidert“. Sie können somit eher auf Lager produziert und direkt verkauft werden, sodass sich kein hoher Auftragsbestand aufbaut. Auch dürfte in einigen dieser Wirtschaftszweige die Just-in-time-Produktion eine wichtige Rolle spielen, bei der Produkte auf Abruf produziert und geliefert werden, also ebenfalls zeitnah zum Auftragseingang, sodass kein großer Auftragsbestand entsteht.

Umstellung der Umsatz-, Auftragseingangs- und Auftragsbestandsindizes im Verarbeitenden Gewerbe auf das Basisjahr 2021

Tabelle 3

Zusammengefasstes Wägungsschema für den Auftragsbestandsindex im Verarbeitenden Gewerbe

	Insgesamt		Inland		Ausland	
	2015	2021	2015	2021	2015	2021
	%					
Einbezogene Teile des Verarbeitenden Gewerbes	100	100	32,71	34,16	67,29	65,84
13 Herstellung von Textilien	0,32	0,36	0,16	0,16	0,17	0,20
14 Herstellung von Bekleidung	0,38	0,44	0,22	0,26	0,17	0,18
17 Herstellung von Papier, Pappe und Waren daraus	0,71	0,81	0,38	0,42	0,34	0,40
20 Herstellung von chemischen Erzeugnissen	1,88	2,95	0,75	1,00	1,13	1,94
21 Herstellung von pharmazeutischen Erzeugnissen	0,85	1,11	0,15	0,36	0,70	0,75
24 Metallerzeugung und -bearbeitung	4,51	4,40	2,40	2,27	2,12	2,12
25 Herstellung von Metallerzeugnissen	6,70	7,02	4,00	4,29	2,68	2,73
26 Herstellung von Datenverarbeitungsgeräten, elektronischen und optischen Erzeugnissen	5,02	6,40	1,89	2,32	3,12	4,07
27 Herstellung von elektrischen Ausrüstungen	6,11	6,25	2,88	3,18	3,23	3,07
28 Maschinenbau	29,64	26,39	8,71	7,53	20,95	18,87
29 Herstellung von Kraftwagen und Kraftwagenteilen	14,96	18,34	5,58	5,77	9,38	12,57
30 Sonstiger Fahrzeugbau	28,92	25,53	5,59	6,60	23,29	18,93

Abteilungen der Klassifikation der Wirtschaftszweige, Ausgabe 2008 (WZ 2008). – Differenzen in den Summen sind rundungsbedingt.

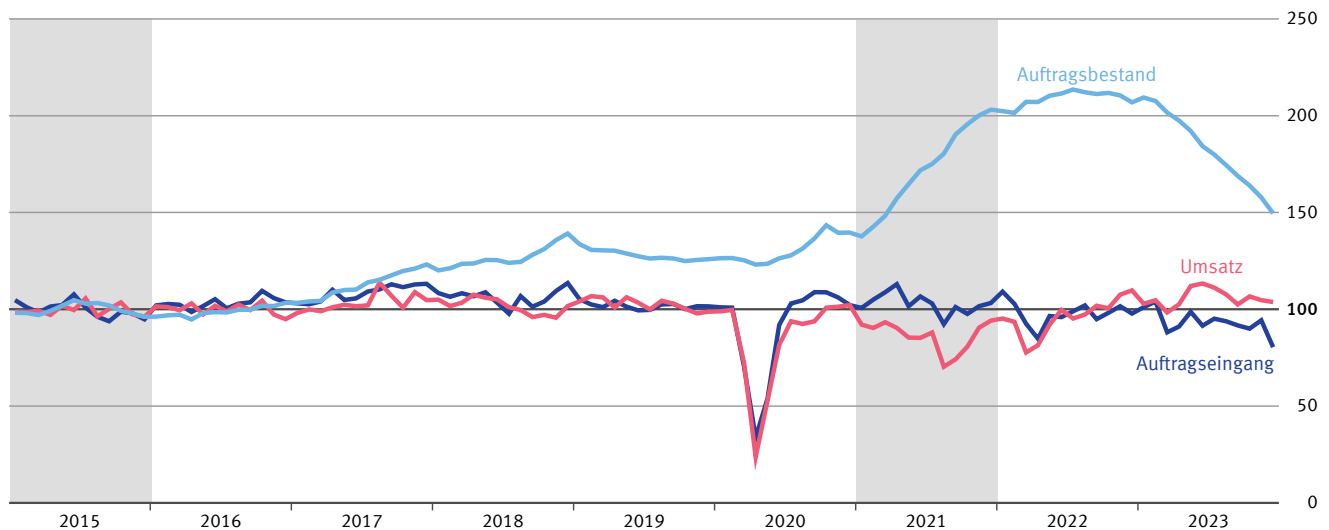
Bei der Veränderung der Wägungsstruktur fällt zunächst auf, dass der Auftragsbestand im Jahr 2021 im Bereich der Herstellung von Kraftwagen und Kraftwagenteilen stark zugenommen hat. Das ist vor allem auf den in der

Automobilindustrie im Jahr 2021 besonders stark verbreiteten Materialmangel zurückzuführen, der zu einem Auftragsstau in außergewöhnlicher Höhe geführt hat.

➤ Grafik 4

Grafik 4

Preis- und saisonbereinigte Konjunkturindizes der Abteilung¹ "Herstellung von Kraftwagen und Kraftwagenteilen" 2015 = 100



¹ Abteilung der Klassifikation der Wirtschaftszweige, Ausgabe 2008 (WZ 2008).

Ähnliches gilt für die Herstellung von Datenverarbeitungsgeräten, elektronischen und optischen Erzeugnissen: Auch in diesem Wirtschaftszweig waren die Hersteller im Jahr 2021 mit Materialmangel konfrontiert und der Auftragsbestand ist stark angewachsen. Schließlich war auch in der Chemieindustrie das Problem des Materialmangels relevant, sodass der Auftragsbestand hier ebenfalls außergewöhnlich stark anstieg.


In der Chemieindustrie und im Automobilbau wurden die Auftragsbestände in den Folgejahren relativ schnell wieder abgebaut, weil das (preisbereinigte) Volumen der Neuaufträge tendenziell rückläufig war. In der Chemiebranche begann der Abbau bereits im Jahr 2022, im Automobilbereich erst ab 2023. Damit ist bei diesen beiden Wirtschaftszweigen das höhere Gewicht im Auftragsbestandsindex des Basisjahres 2021 eher auf Sondereffekte infolge des Materialmangels zurückzuführen als auf konjunkturelle Entwicklungen. Der Bereich der Herstellung von Datenverarbeitungsgeräten, elektronischen und optischen Erzeugnissen befindet sich hingegen seit vielen Jahren in einer Aufschwungphase, hier steigt der Auftragsbestand weiterhin an.

Im Maschinenbau und im Sonstigen Fahrzeugbau ist der Auftragsbestand gesunken. Der Rückgang des Wägungsanteils im Auftragsbestandsindex kommt vor allem dadurch zustande, dass die Auftragsbestände in diesen beiden Wirtschaftszweigen im vorigen Basisjahr 2015 relativ hoch waren. Absolut betrachtet ist in diesen beiden Wirtschaftszweigen der Auftragsbestand angestiegen.

7

Fazit

Mit Berichtsmonat Januar 2024 wurde beim Umsatzindex für den Bergbau und das Verarbeitende Gewerbe sowie bei den Auftragseingangs- und Auftragsbestandsindizes im Verarbeitenden Gewerbe das bisherige Basisjahr 2015 turnusmäßig durch das neue Basisjahr 2021 abgelöst. Neben der Anpassung des Basisjahres als Bezugsgröße der Indizes wurden die Indexgewichte aktualisiert.

Die speziellen wirtschaftlichen Entwicklungen im Jahr 2021 – Coronakrise, Materialknappheit, fehlende Transportkapazitäten, hohe Energiepreise – erschweren die Interpretation der Umsatz-, Auftragseingangs- und Auftragsbestandsindizes an einigen Stellen. Wegen dieser Besonderheiten wurde im Europäischen Statistiksistem eine Verschiebung der Basisjahrumstellung oder die Berechnung von Durchschnittsn mehrerer Jahre für die Gewichtung erwogen. Alle in Betracht gezogenen Alternativen waren jedoch ihrerseits mit gravierenden Nachteilen verbunden, sodass letztlich auch aus administrativen Gründen am Basisjahr 2021 festgehalten wurde. 

LITERATURVERZEICHNIS

Linz, Stefan/Flores, Luis Federico/Bolz, Maria/Schächer, Jennifer/Eid, Nicole. [Umstellung des Produktionsindex im Produzierenden Gewerbe auf das Basisjahr 2021](#). In: WISTA Wirtschaft und Statistik. Ausgabe 2/2024, Seite 55 ff.

Linz, Stefan/Neumann, Malte David/Abdalla, Salima/Gladis-Dörr, Gerda. [Auswirkungen der Corona-Pandemie: Lieferengpässe bremsen Industrie und treiben Preise](#). In: WISTA Wirtschaft und Statistik. Ausgabe 1/2022, Seite 71 ff.

Linz, Stefan/Möller, Hans-Rüdiger/Mehlhorn, Peter. [Umstellung der Konjunkturindizes im Produzierenden Gewerbe auf das Basisjahr 2021](#). In: WISTA Wirtschaft und Statistik. Ausgabe 2/2018, Seite 49 ff.

Statistisches Bundesamt. *Klassifikation der Wirtschaftszweige, Ausgabe 2008 (WZ 2008)*. Wiesbaden 2008. [Zugriff am 6. Juni 2024]. Verfügbar unter: www.destatis.de

Statistisches Bundesamt. *Monatliche Konjunkturstatistik im Dienstleistungsbereich*. [Zugriff am 3. Mai 2024]. Verfügbar unter: www.destatis.de

Statistisches Amt der Europäischen Union (Eurostat). *Statistische Systematik der Wirtschaftszweige in der Europäischen Gemeinschaft – NACE Rev. 2*. 2008. [Zugriff am 11. Juni 2024]. Verfügbar unter: ec.europa.eu

Wohlrabe, Klaus. *Materialengpässe in der Industrie: Wer ist betroffen, und wie reagieren die Unternehmen?* In: Ifo Schnelldienst, München. Jahrgang 74. Ausgabe 9/2021, Seite 60 ff. [Zugriff am 14. Mai 2024]. Verfügbar unter: www.ifo.de

RECHTSGRUNDLAGEN

Durchführungsverordnung (EU) 2020/1197 der Kommission vom 30. Juli 2020 zur Festlegung technischer Spezifikationen und Einzelheiten nach der Verordnung (EU) 2019/2152 des Europäischen Parlaments und des Rates über europäische Unternehmensstatistiken, zur Aufhebung von zehn Rechtsakten im Bereich Unternehmensstatistiken (Amtsblatt der EU Nr. L 271, Seite 1).

Gesetz über die Statistik im Produzierenden Gewerbe (ProdGewStatG) in der Fassung der Bekanntmachung vom 21. März 2002 (BGBl. I Seite 1181), das zuletzt durch Artikel 7 des Gesetzes vom 22. Februar 2021 (BGBl. I Seite 266) geändert worden ist.

Verordnung (EG) Nr. 1893/2006 des Europäischen Parlaments und des Rates vom 20. Dezember 2006 zur Aufstellung der statistischen Systematik der Wirtschaftszweige NACE Revision 2 und zur Änderung der Verordnung (EWG) Nr. 3037/90 des Rates sowie einiger Verordnungen der EG über bestimmte Bereiche der Statistik (Amtsblatt der EU Nr. L 393, Seite 1).

Verordnung (EU) 2019/2152 des Europäischen Parlaments und des Rates vom 27. November 2019 über europäische Unternehmensstatistiken, zur Aufhebung von zehn Rechtsakten im Bereich Unternehmensstatistiken (Amtsblatt der EU Nr. L 327, Seite 1).

Verordnung (EG) Nr. 1165/98 des Rates vom 19. Mai 1998 über Konjunkturstatistiken (Amtsblatt der EG Nr. L 162, Seite 1).

REGIONALE ERGEBNISSE DER UNTERNEHMENSDEMOGRAFIE

Anke Rink, Ines Seiwert, Raimund Rödel

➤ **Schlüsselwörter:** Unternehmen – Arbeitgeberdemografie – Unternehmensregister – EBS-Verordnung – Regionalität – NUTS-Gründungskarten

ZUSAMMENFASSUNG

Die Unternehmensdemografie betrachtet das Gründungs- und Schließungsgeschehen von Unternehmen sowie deren Bestand am Markt. Diese Informationen stehen nun auch auf regionaler Ebene europaweit zur Verfügung. Der Artikel präsentiert erste Ergebnisse auf regionaler Ebene für Deutschland und stellt Möglichkeiten weiterer regional tief gegliederter Auswertungen dar.

➤ **Keywords:** enterprises – employer business demography – business register – EBS Regulation – regionalität – NUTS maps of enterprise births

ABSTRACT

Business demography looks at the birth and death of enterprises and their survival in the market. This information is now also available at regional level across Europe. The article presents first results at regional level for Germany and outlines the possibilities for further in-depth regional analyses.

Anke Rink

ist Diplom-Geografin und als Referentin im Referat „Unternehmensregister, -demografie, Verwaltungsdatenspeicher, Handwerk“ des Statistischen Bundesamtes für das Fachthema Unternehmensdemografie zuständig. Sie vertritt das Statistische Bundesamt in diesem Bereich in den europäischen Arbeitsgruppen, erstellt methodische Konzepte und verantwortet neben den Datenlieferungen an das Statistische Amt der Europäischen Union die Georeferenzierung des Unternehmensregisters.

Ines Seiwert

ist Diplom-Verwaltungswirtin (FH) und im Referat „Unternehmensregister, -demografie, Verwaltungsdatenspeicher, Handwerk“ des Statistischen Bundesamtes für das Fachthema Unternehmensdemografie zuständig. Die Erstellung von methodischen Konzepten, Datenlieferungen an das Statistische Amt der Europäischen Union und Veröffentlichung der Daten zur Unternehmensdemografie gehören zu ihren Aufgaben.

Dr. Raimund Rödel

ist Geoinformatiker und Geograph und für das statistische Unternehmensregister im Bayerischen Landesamt für Statistik tätig. Ein besonderes Augenmerk richtet er auf die Chancen, die in der Arbeit mit Geodaten im Unternehmensregister und in den Wirtschaftsstatistiken liegen.

1

Einleitung

Im Mittelpunkt der Unternehmensdemografie stehen sowohl Unternehmensgründungen und -schließungen der gesamten Unternehmenspopulation als auch Unternehmensgründungen und -schließungen von Arbeitgeberunternehmen sowie schnell wachsende Unternehmen. Bei Unternehmensgründungen und -schließungen von Arbeitgeberunternehmen handelt es sich um Unternehmen, die ihre erste beschäftigte Person einstellen beziehungsweise die letzte beschäftigte Person entlassen. Für neu gegründete Unternehmen werden Überlebensraten ermittelt. Diese Informationen sind wichtige Indikatoren für die Dynamik einer Volkswirtschaft. Neben dem Wettbewerb und dem Strukturwandel beeinflussen aber auch unvorhersehbare Ereignisse, beispielsweise die Corona-Pandemie, wirtschaftliche Entwicklungen und verändern die Unternehmenspopulation.

Bis einschließlich des Berichtsjahres 2020 liegen Analysen zur Unternehmensdemografie nur für Deutschland insgesamt vor. Um die Dynamik verschiedener regionaler Märkte zu analysieren, ist jedoch eine kleinräumigere Darstellung der Daten der Unternehmensdemografie, der Arbeitgeberdemografie¹ und der schnell wachsenden Unternehmen erforderlich.

Ab dem Berichtsjahr 2021 bieten die Daten der Unternehmensdemografie auf Kreisebene diese wichtigen Informationen für politische Entscheidungen und für die Indikatoren der [Strategie Europa 2020](#) an. Weiterhin sind unternehmensdemografische Daten ein wichtiger Bestandteil des [Entrepreneurship Indicators Programme](#) der Organisation für wirtschaftliche Zusammenarbeit und Entwicklung (OECD). Die Gründungs-, Schließungs- und Überlebensraten fließen in die Strukturindikatoren ein, die die Fortschritte bei der Verwirklichung der Strategie Europa 2020 überwachen.

Die rechtliche Grundlage für die Darstellung von kleinräumigeren Daten ist die Verordnung der Europäischen Union (EU) über europäische Unternehmensstatistiken sowie der dazugehörigen Durchführungsrechtsakte.

Die Statistiken der Unternehmensdemografie, einschließlich der regionalen Aufteilung, basieren auf der statistischen Einheit [Unternehmen](#) nach der EU-Unternehmensdefinition. Demnach entspricht das Unternehmen „(...) der kleinsten Kombination Rechtlicher Einheiten, die eine organisatorische Einheit zur Erzeugung von Waren und Dienstleistungen bildet und insbesondere in Bezug auf die Verwendung der ihr zufließenden laufenden Mittel über eine gewisse Entscheidungsfreiheit verfügt. Ein Unternehmen übt eine Tätigkeit oder mehrere Tätigkeiten an einem Standort oder an mehreren Standorten aus. Ein Unternehmen kann einer einzigen Rechtlichen Einheit (einfaches Unternehmen) entsprechen.“

Die nationalen Veröffentlichungen erfolgen in [GENESIS-Online](#), der Datenbank des Statistischen Bundesamtes. Daneben veröffentlicht das Statistische Amt der Europäischen Union (Eurostat) alle Kennzahlen zur regionalen Unternehmensdemografie im [EC Data Browser](#).

Das anschließende Kapitel 2 stellt die Methodik der regionalen Unternehmensdemografie dar, die Präsentation der ersten Ergebnisse folgt in Kapitel 3. Der Beitrag endet mit einem kurzen Fazit in Kapitel 4.

2

Methodik

Das Ziel der Unternehmensdemografie ist es, echte Gründungen und echte Schließungen zu erkennen und darzustellen. Echte (originäre) Gründungen beziehungsweise Schließungen nach der Definition der Unternehmensdemografie umfassen die Schaffung oder Auflösung von Produktionsfaktoren, ohne dass andere Unternehmen an diesem Vorgang beteiligt sind.

Diese unternehmensdemografischen Daten werden belastungsarm aus vorhandenen Datenquellen generiert (European Communities, 2007a). Die Hauptquelle für die Unternehmensdemografie ist dabei das statistische Unternehmensregister (URS) (Möding/Philipp, 2007). Dieses enthält neben den Unternehmen nach EU-Unternehmensdefinition (nachfolgend Unternehmen genannt) auch die zugehörigen Rechtlichen Einheiten und Niederlassungen.

1 Diese umfasst nur Unternehmen mit Beschäftigten.

Für die Auswertung relevant sind dabei alle Unternehmen, deren Rechtliche Einheiten im Berichtsjahr mehr als 22 000 Euro Umsatz erzielen, oder deren Niederlassungen im Berichtsjahr im Jahresdurchschnitt sozialversicherungspflichtig Beschäftigte haben, oder deren Niederlassungen im Berichtsjahr im Jahresdurchschnitt mindestens eine Person geringfügig entlohnt beschäftigen.

Ein Unternehmen zählt als **Gründung**, wenn alle ihm zugehörigen Rechtlichen Einheiten Gründungen sind oder lediglich Hilfstätigkeiten für das Unternehmen ausführen. Ein Unternehmen zählt als **Schließung**, wenn alle ihm zugehörigen Rechtlichen Einheiten Schließungen sind oder lediglich Hilfstätigkeiten ausführen. Dieses Vorgehen berücksichtigt auch die Restrukturierungen von Rechtlichen Einheiten in komplexen Unternehmen oder Unternehmensgruppen. Neben den Informationen zum Gründungs- und Schließungsgeschehen liefert die Unternehmensdemografie Ergebnisse zum Fortbestand von neu gegründeten Unternehmen und deren Beschäftigtenentwicklung. Ein Unternehmen überlebt, wenn mindestens eine der zugehörigen Rechtlichen Einheiten zu einem beliebigen Zeitpunkt des Folgejahres über Beschäftigte verfügt oder Umsätze erzielt. Das Überleben von Neugründungen wird über einen Zeitraum von fünf Jahren verfolgt. Die Methodik zur Berechnung demografischer Ereignisse, zur Ermittlung der echten Gründungen und Schließungen und zur Ableitung der Ergebnisse von Unternehmen stellen vorhergehende Aufsätze zur Unternehmensdemografie in dieser Zeitschrift ausführlich dar (Rink und andere, 2013; Rink/Seiwert, 2021).

Die Unternehmensdemografie bildet alle Unternehmen der Gesamtwirtschaft² nach der [Klassifikation der Wirtschaftszweige, Ausgabe 2008](#) (WZ 2008) ab, die am Markt tätig sind.

Die Bestimmung der Markttätigkeit eines Unternehmens erfolgt nach dem [Europäischen System Volkswirtschaftlicher Gesamtrechnungen](#) (ESVG 2010).

Marktproduktion ist dabei die Herstellung von Gütern, die auf dem Markt verkauft werden oder verkauft werden sollen.

Für die regionale Aufgliederung der Ergebnisse wird die NUTS-Klassifikation³ (Nomenclature des Unités territoriales statistiques) verwendet. Sie lehnt sich eng an die Verwaltungsgliederung der einzelnen Länder der Europäischen Union an. Dabei untergliedert die NUTS die Mitgliedstaaten in vier Hierarchieebenen. Die den NUTS-Leveln entsprechenden regionalen Untergliederungen in Deutschland zeigt [Übersicht 1](#).

Übersicht 1

NUTS-Level und ihre Entsprechung in Deutschland

NUTS-Level	Entsprechung in Deutschland	Ausprägungen
NUTS 0	Nationalstaat	1
NUTS 1	Bundesländer	16
NUTS 2	Regierungsbezirke	38
NUTS 3	Landkreise/Kreise und kreisfreie Städte beziehungsweise Stadtkreise (in Baden-Württemberg)	401

Die NUTS-Schlüssel (European Communities, 2007b) werden aus dem Amtlichen Gemeindeschlüssel abgeleitet, der im statistischen Unternehmensregister für jede Niederlassung vorliegt. Bei mehreren Niederlassungen wird der Amtliche Gemeindeschlüssel der Sitzniederlassung der Rechtlichen Einheit verwendet.

Ähnlich wird dieser auch für das Unternehmen ermittelt. Hier wird der Amtliche Gemeindeschlüssel der bestimmenden Rechtlichen Einheit auf das gesamte Unternehmen übertragen. Für die Unternehmensdemografie wird der Amtliche Gemeindeschlüssel verwendet, der am jeweiligen Ende des Berichtsjahres bei der Einheit gespeichert ist.

Die schwerpunktmäßige regionale Zuordnung kann zu Unschärfen führen, was jedoch aus den folgenden Gründen zu vernachlässigen ist: Im Berichtsjahr 2021 waren 99,7 % der Gründungen einfache Unternehmen, in denen 93,7 % aller abhängig Beschäftigten in Neugründungen beschäftigt waren. Von diesen neu gegründeten einfachen Unternehmen wiederum bestanden 99,8 % aus Rechtlichen Einheiten mit nur einer Niederlassung. Bei den Schließungen waren 99,9 % einfache Unternehmen und bei 99,1 % der Rechtlichen Einheiten bestanden diese nur aus einer Niederlassung.

2 Produzierendes Gewerbe und Dienstleistungsbereich; Abschnitte B bis N und P bis S ohne Interessenvertretungen sowie kirchliche und sonstige religiöse Vereinigungen (S 94) der WZ 2008.

3 Verordnung (EG) Nr. 1059/2003 des Europäischen Parlaments und des Rates vom 26. Mai 2003 über die Schaffung einer gemeinsamen Klassifikation der Gebietseinheiten für die Statistik (NUTS).

Auch die regionale Aufgliederung der einzelnen Bundesländer weist Unterschiede auf: Nur bei der Hälfte der Bundesländer wird die NUTS-2-Ebene tiefer aufgliedert. Hierbei variiert die Anzahl der Kreise je Regierungsbezirk (bei den Bundesländern mit mehr als einer NUTS-2-Ausprägung) von 3 Kreisen je Regierungsbezirk in Sachsen bis hin zu 23 Kreisen je Regierungsbezirk in Bayern stark. Das Land mit der größten Anzahl an regionalen Untergliederungen auf NUTS-2-Ebene ist Bayern. Diese Unterschiede wirken sich auf kartografische Darstellungen aus, da kleinere gründungsstarke kreisfreie Städte nicht so prägnant wahrgenommen werden wie größere Kreise, die eine Stadt mit einschließen.

Da Unternehmen umziehen, kann ein und dasselbe Unternehmen in verschiedenen Populationen der Unternehmensdemografie in unterschiedlichen Regionen nachgewiesen werden. Ein Beispiel: Ein Unternehmen wird im Main-Taunus-Kreis gegründet. Es wächst schnell und zieht nach zwei Jahren nach Wiesbaden, wo es ein weiteres Jahr lang schnell wächst. Nach weiteren sechs Jahren zieht es nach Mainz, wo es nach weiteren zehn Jahren geschlossen wird.

Die Regionen, in denen das Unternehmen in der Unternehmensdemografie nachgewiesen wird, sind:

- › als aktives Unternehmen:
 - Main-Taunus-Kreis (in den ersten zwei Jahren);
 - Wiesbaden (in den nächsten sechs Jahren);
 - Mainz (in den letzten zehn Jahren)
- › Gründung: Main-Taunus-Kreis
- › Arbeitgebergründung: Main-Taunus-Kreis
- › Dreijahresüberleben: Main-Taunus-Kreis (obwohl das Unternehmen mittlerweile in Wiesbaden aktiv ist)
- › Schnell wachsendes Unternehmen: Wiesbaden (letztes Jahr mit hohem Wachstum)
- › Schließung: Mainz

3

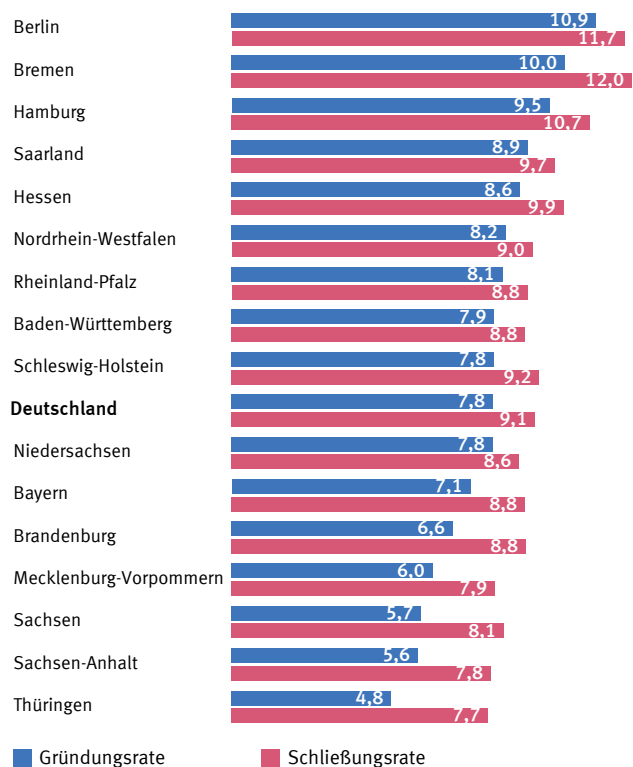
Erste Ergebnisse für die regionale Unternehmensdemografie

Bislang lagen die Daten der Unternehmensdemografie nur für Deutschland und die anderen europäischen Länder vor. Das Gründungs- und Schließungsgeschehen ist jedoch in den einzelnen Regionen Deutschlands und Europas sehr unterschiedlich. Nachfolgend werden die Ergebnisse der Unternehmensdemografie nach Kreisen und auch in der Unterscheidung städtisch, ländlich und intermediär (Vororte, kleinstädtische Bereiche) analysiert.

➤ **Grafik 1** vergleicht die Gründungs- und Schließungsraten der Gesamtwirtschaft über die Bundesländer hinweg. Hier fallen besonders die Stadtstaaten mit hohen


Grafik 1

Gründungs- und Schließungsraten im Produzierenden Gewerbe und Dienstleistungsbereich¹ im Berichtsjahr 2021 je 100 aktive Unternehmen



¹ Abschnitte B bis N und P bis S (ohne S 94) der Klassifikation der Wirtschaftszweige, Ausgabe 2008 (WZ 2008).

Raten auf. Dies betrifft sowohl die Gründungs- als auch die Schließungsraten, sodass die Unternehmenslandschaft dort eine höhere Dynamik aufweist.

Die regionale Aufgliederung der Ergebnisse ermöglicht auch weitergehende und übergreifende Analysen der Unternehmensdemografie. Für eine kartografische Darstellung wird die Zahl der Unternehmensgründungen mit der Zahl der aktiven Unternehmen je Kreis standardisiert. Eine Standardisierung wird angewendet, wenn die unterschiedliche Flächengröße der regionalen Einheiten die Darstellung absoluter Werte verzerren würde.  **Grafik 2** stellt die Unternehmensgründungen je aktiven Unternehmen nach Kreisen dar. Die Einteilung der Klassen erfolgt über die Quantile.


Die kleinste Klasse wird durch jene 67 Kreise mit den Gründungsraten zwischen 34 und 57 Gründungen je 1 000 aktiver Unternehmen repräsentiert. Die größte Klasse repräsentiert jenes Sechstel von 67 Kreisen in Deutschland mit Gründungsraten zwischen 84 und 117 Gründungen je 1 000 aktiver Unternehmen.

Inhaltlich ist dennoch folgender Aspekt zu diskutieren: Aus einem wirtschaftsgeografischen Blickwinkel kann die Aussagekraft von Gründungsraten, die durch Standardisierung an der Zahl der aktiven Unternehmen gebildet wurden, hinterfragt werden. Unternehmensgründungen können als Indikator für die unternehmerische Risikofreudigkeit und das Innovationspotenzial in der Bevölkerung einer Region wahrgenommen werden. Eine Region mit einem hohen Unternehmensbestand – diese darf als wirtschaftlich stark gelten – wird durch die hohe Zahl der Unternehmen im Nenner eine niedrigere Gründungsrate aufweisen als eine Region mit gleicher Zahl an Gründungen, aber einem niedrigeren Unternehmensbestand. Die Region allein mit dem geringeren Unternehmensbestand gilt somit merkwürdigerweise als gründungsstärker. Der Unternehmensbestand ist als Maßstab für die Risikofreudigkeit und das Innovationspotenzial einer Region demnach weniger geeignet.

Die Gründungsfreudigkeit drückt sich unter Berücksichtigung der wirtschaftlichen Rahmenbedingungen eher im Anteil der Unternehmensgründungen bei vergleichbarer erwerbsfähiger Bevölkerung aus. Gründungen entstehen weniger aus einem Reservoir an vorhandenen Unternehmen, diese wirken im Zweifelsfall eher als Konkurrenten. Um die Ergebnisse der regionalen Unternehmensdemografie und hier besonders die Gründungsraten sinnvoll

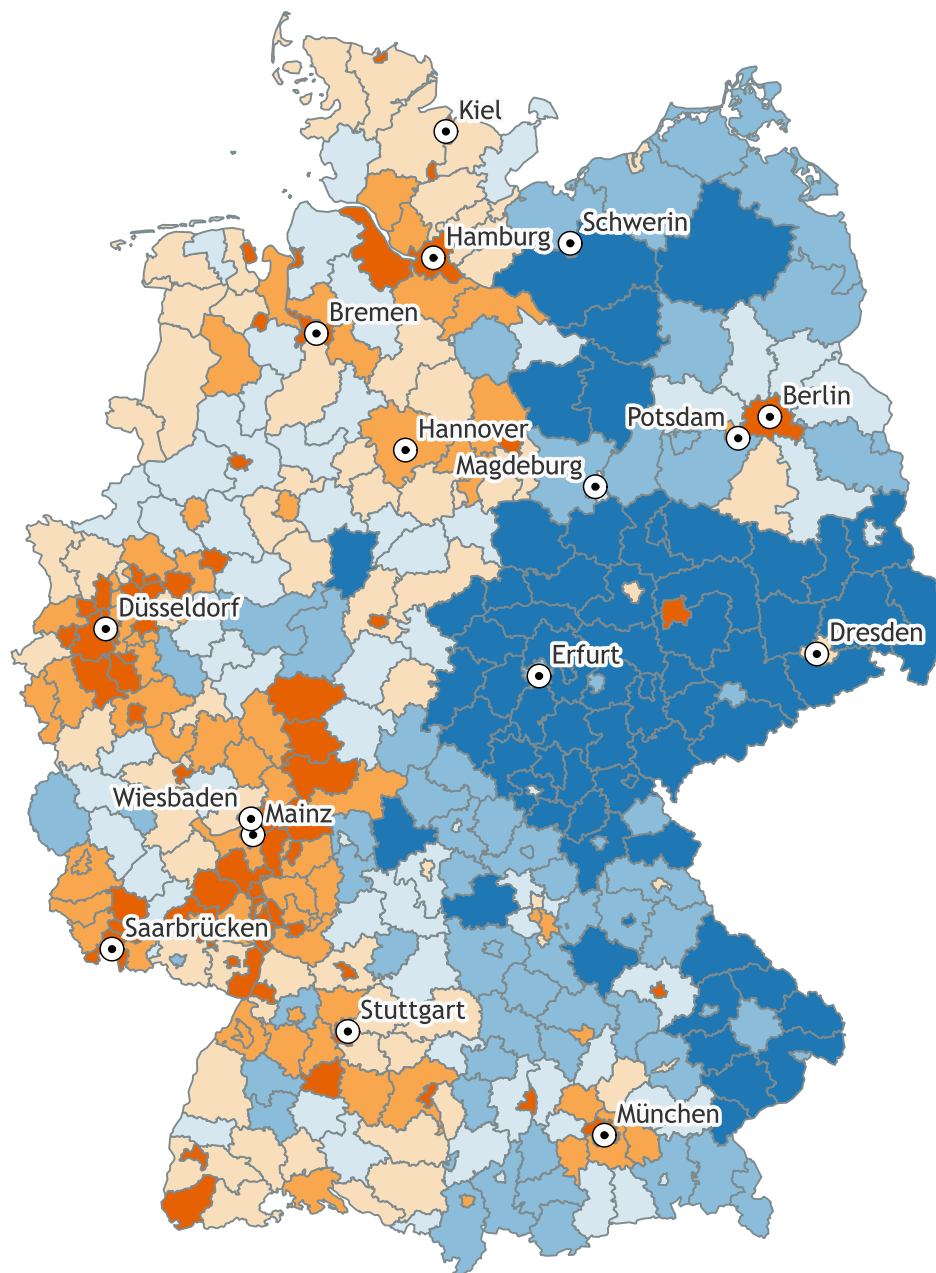
abzubilden, werden die Gründungsraten daher alternativ als Anteil der Unternehmensgründungen je erwerbsfähiger Bevölkerung im Alter zwischen 15 und 65 Jahren dargestellt. Dieses Vorgehen ermöglicht einen Vergleich zwischen Gewerbeanzeigenstatistik und Unternehmensdemografie.

Untersuchungen zum Zusammenhang zwischen den Gewerbeanmeldungen und der Unternehmensdemografie haben gezeigt, dass überschlägig etwa knapp die Hälfte der zur Anzeige gebrachten Gewerbeanmeldungen später in einem neugegründeten Unternehmen in der Unternehmensdemografie resultieren (Rödel/Stephan, 2022).

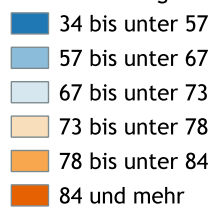
 **Grafik 3** stellt die Gründungsraten für das Berichtsjahr 2021 in den Kreisen Deutschlands als Zahl der Unternehmensgründungen je 10 000 Erwerbsfähigen dar. Das Kartenbild wird wesentlich durch die Umrisse der administrativen Kreise geprägt. Hier fällt auf, dass gerade im südlichen Deutschland kreisfreie Städte mit einer geringen Fläche mit ihren teilweise hohen Gründungsraten im Kartenbild optisch weniger in Erscheinung treten als größere Flächen. Sind Städte und ihr Umland dagegen wie in Teilen des nördlichen Deutschlands in der Kreisgliederung zusammengefasst, nivelliert dies vermutlich auch die Gründungsraten von Stadt und Umland. Die Geometrie der Verwaltungsgliederung bestimmt also das Kartenbild. Dadurch ist die Ausdehnung von Schwerpunkten oder Clustern eines intensiveren Gründungsgeschehens deutlich schwieriger zu erkennen als in den nachfolgenden Karten mit einer Darstellung in flächengleichen Hexagonen.

Grafik 2

Gründungsraten im Berichtsjahr 2021
Unternehmensgründungen je 1 000 aktive Unternehmen



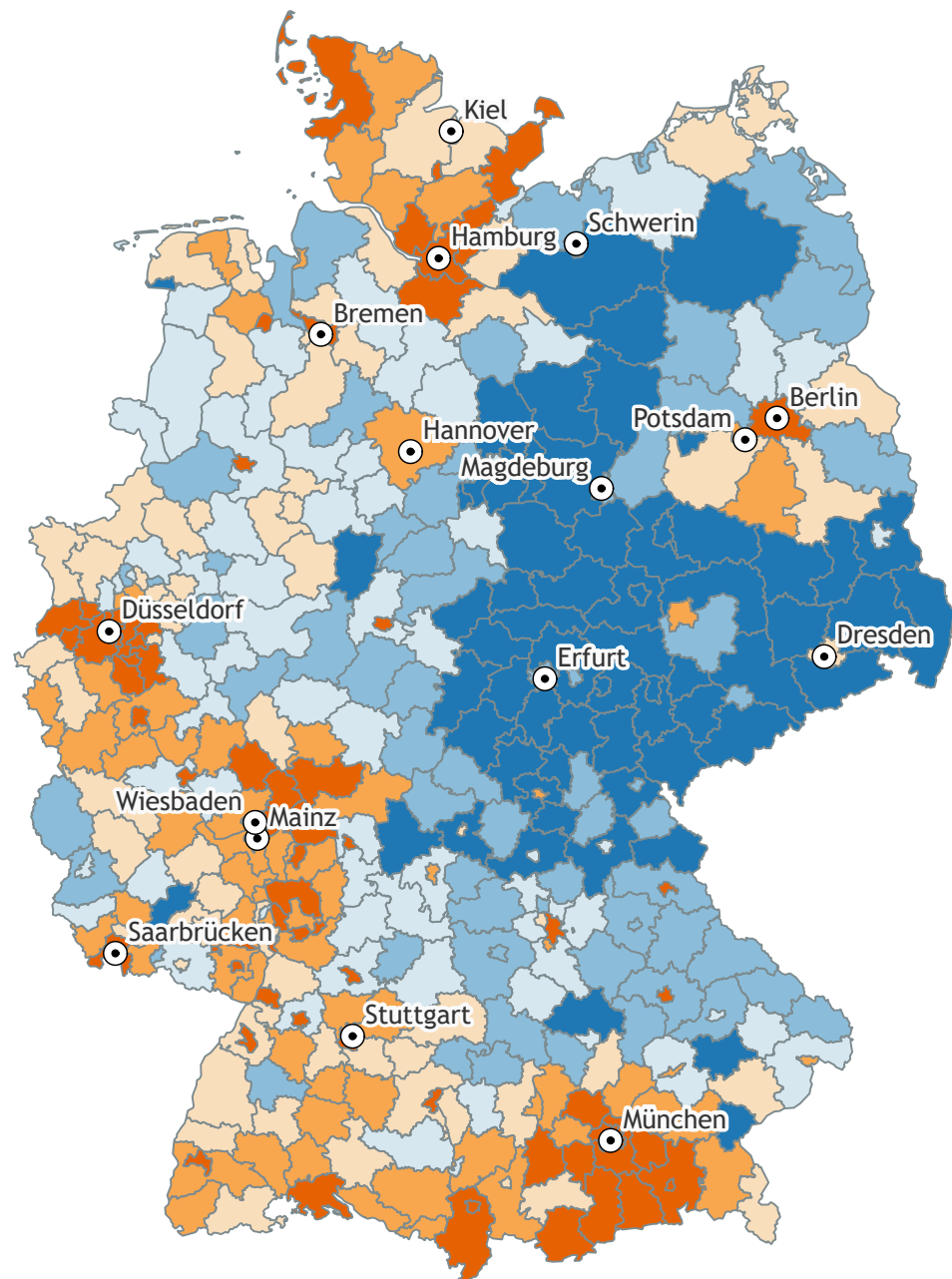
Unternehmensgründungen je Tsd. Unternehmen



Grafik 3

Gründungsraten im Berichtsjahr 2021

Unternehmensgründungen je 10 000 Erwerbsfähige im Alter von 15 bis unter 65 Jahren

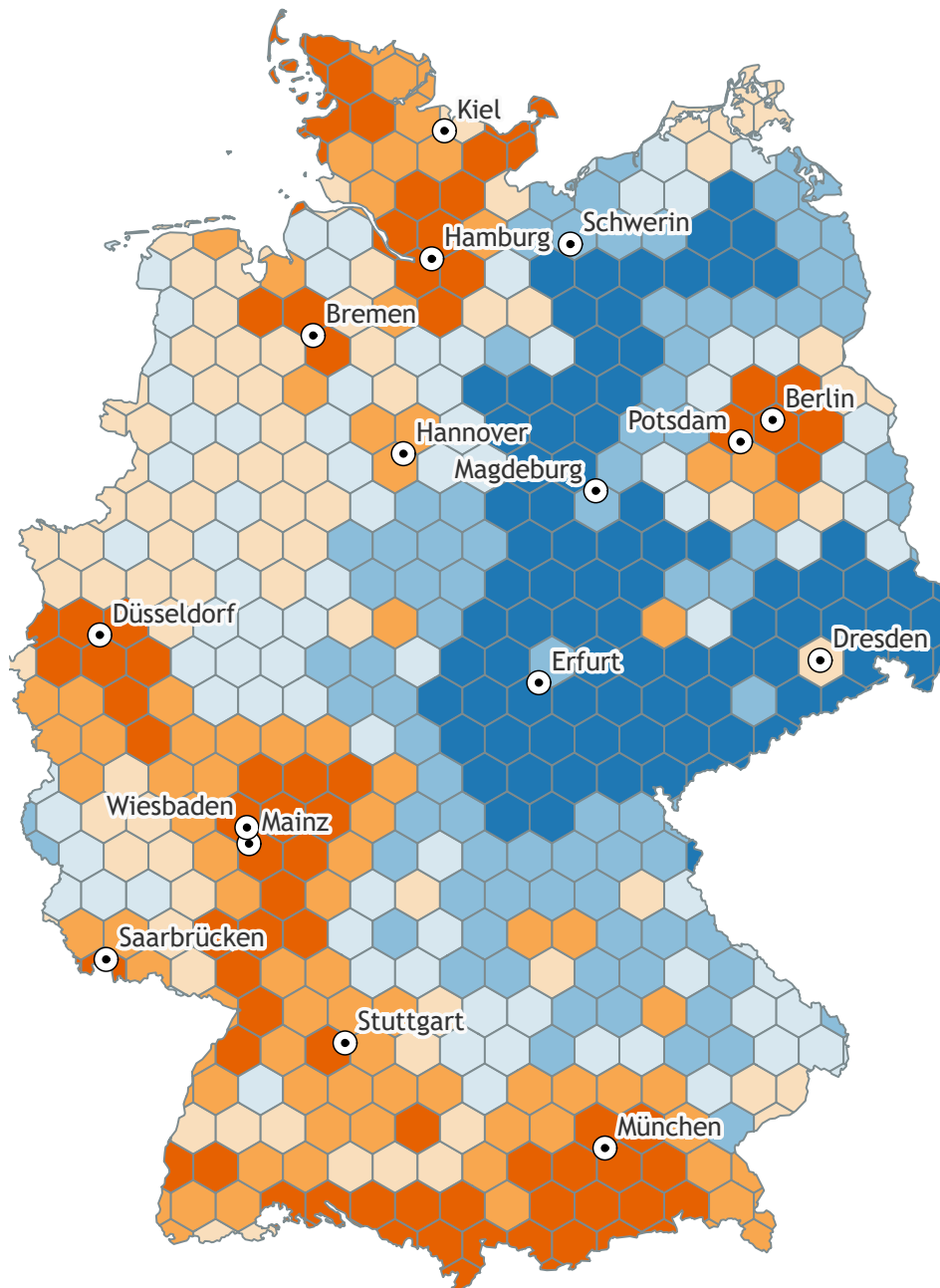


Unternehmensgründungen je 10 Tsd. Erwerbsfähige

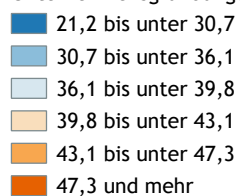
- 16,1 bis unter 31,5
- 31,5 bis unter 37,1
- 37,1 bis unter 41
- 41 bis unter 44,6
- 44,6 bis unter 49,7
- 49,7 und mehr

Grafik 4

Gründungsraten in Deutschland in flächengleichen Hexagonen im Berichtsjahr 2021
Unternehmensgründungen je 10 000 Erwerbsfähige im Alter von 15 bis unter 65 Jahren



Unternehmensgründungen je 10 Tsd. Erwerbsfähige



Um den Einfluss der administrativen Gliederungen aufzulösen, wurde die Karte in [Grafik 4](#) so modifiziert, dass flächengleiche Hexagone die Gründungsraten wiedergeben. Für diese Kartendarstellung wurde das R Package cartography mit der Methode getGridLayer verwendet. Hierbei wurde aus der Geometrie der Unternehmensgründungen und der Geometrie zur Zahl der Erwerbsfähigen die Geometrie der Gründungsrate berechnet. Die Größe der Hexagone entspricht dem Median des Flächeninhalts aller Kreise in Deutschland. Allerdings handelt es sich hierbei nicht um eine echte Wiedergabe der Gründungsrate für jedes Hexagon, sondern um interpolierte Werte anhand der Werte aus der administrativen Regionalgliederung in Grafik 3. Trotzdem sind Cluster oder zusammenhängende Bereiche erhöhter Gründungsaktivität besser erkennbar.

Im Gegensatz zur Darstellung basierend auf der Zuordnung zu Kreisen (administrative Gliederung) zeigt [Grafik 5](#) eine Karte von Bayern, die auf geografisch

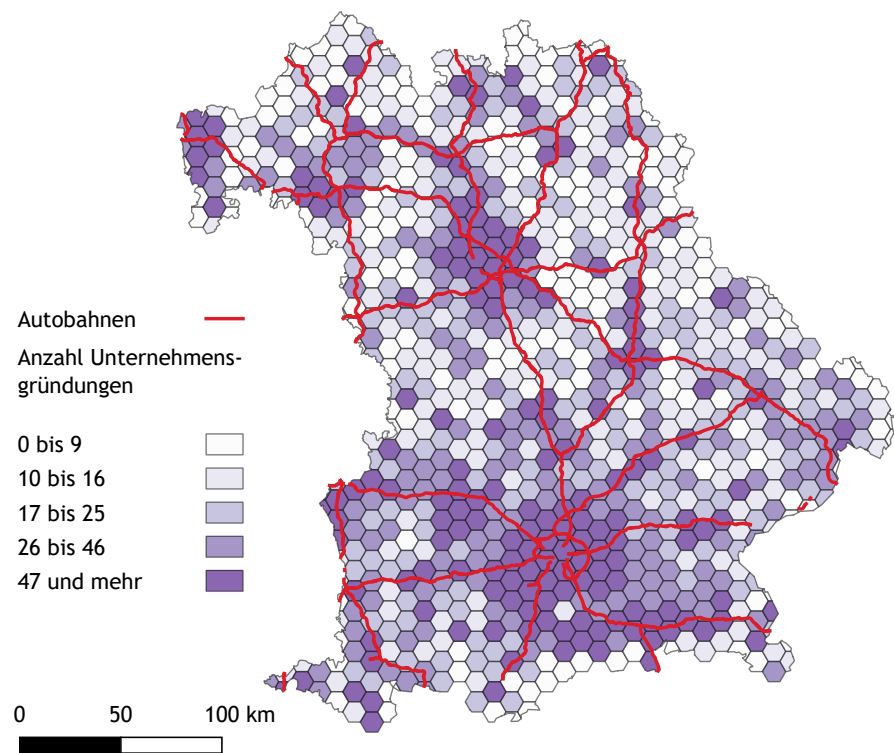
verorteten Angaben (Koordinaten) zu Unternehmensgründungen basiert.

Da die für Unternehmensgründungen betreffenden Einheiten als georeferenzierter Datenbestand vorliegen, wurden diese einem aufgespannten Feld von Hexagonen mit 10 km Innendurchmesser zugeordnet. Die Hexagone haben einen jeweils gleichen Flächeninhalt, deshalb zeigt Grafik 5 anders als die vorangegangenen Karten administrativer Regionen nun eine echte flächentreue Darstellung. Damit ist es möglich, die Gründungsaktivität ohne den Umweg über eine Interpolation anhand der Zahl der Unternehmensgründungen zu visualisieren. Die Farbgebung unterteilt die Zahl der Unternehmensgründungen in fünf gleich große, durch Quantile gebildete Klassen.

In Grafik 5 werden nicht nur Cluster oder zusammenhängende Bereiche erhöhter Gründungsaktivität erkennbar, auch das räumliche Muster verstärkter Gründungsaktivität kann verfolgt werden. Sehr eindrücklich sind die

Grafik 5

Unternehmensgründungen in Bayern im Berichtsjahr 2021
Hexagone mit einem Innendurchmesser von 10 km



Kartengrundlage für Straßen: Bundesamt für Kartographie und Geodäsie: Digitales Landschaftsmodell 1:1 000 000, DLM1000, Datenlizenz: dl-de/by-2-0.

Schwerpunkte von Unternehmensgründungen in und um die Metropolregionen München und Nürnberg zu erkennen. Ebenso werden aber auch Achsen entlang von Verkehrsadern wie der Autobahnen von München nach Salzburg, nördlich von Regensburg und entlang der Linie Nürnberg-Bamberg-Schweinfurt deutlich.

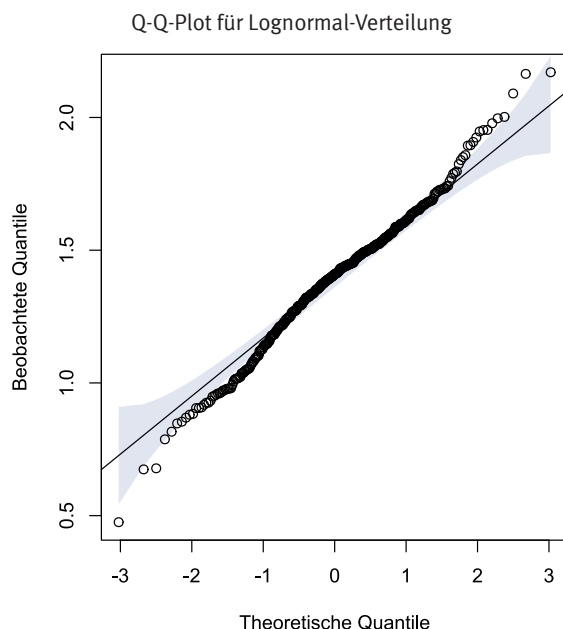
➤ Methodische Erwägungen zu Unternehmensgründungen je erwerbsfähiger Bevölkerung (15 bis unter 65 Jahre)

Die in diesem Beitrag bevorzugte Darstellung von Gründungsdaten als Unternehmensgründungen je erwerbsfähiger Bevölkerung wird durch ein randliches methodisches Detail gestützt.

➤ Grafik 6 zeigt deutlich, dass Unternehmensgründungen eine regionale Häufung oder Clusterung aufweisen. Sie sind also nicht gleichmäßig und auch nicht zufällig im Raum verteilt. Eher sind sie als überdispers zu charakterisieren und folgen einer gruppierten (contagious) Verteilung. Wird der Raum in gleichmäßige Flächen (wie in Grafik 5 in Hexagone) aufgeteilt, gibt es zahlreiche gering besetzte Flächen und einige Flächen mit einer hohen Zahl an Unternehmensgründungen. Eine solche Verteilung ist nicht symmetrisch. Diese Art der Verteilung

Grafik 6

Q-Q-Plot für die logarithmierte Zahl der Unternehmensgründungen je Erwerbsfähige in den Kreisen Deutschlands



wird typischerweise durch eine verbundene Verteilung (mixture distribution) aus einer Normalverteilung und einer endlastigen Verteilung (heavy-tailed) beschrieben. Für eine inhaltlich zutreffende regionale Darstellung, die die Grundzüge einer solchen gruppierten Verteilung zutreffend widerspiegelt, sollte der gewählte Indikator der Gründungsrate damit im Bereich hoher Gründungsdaten einer rechtsschiefen statistischen Verteilung folgen. Eine derartige rechtsschiefe, endlastige Verteilung stellt die Logarithmische Normal-Verteilung (kurz: Log-Normalverteilung) dar.

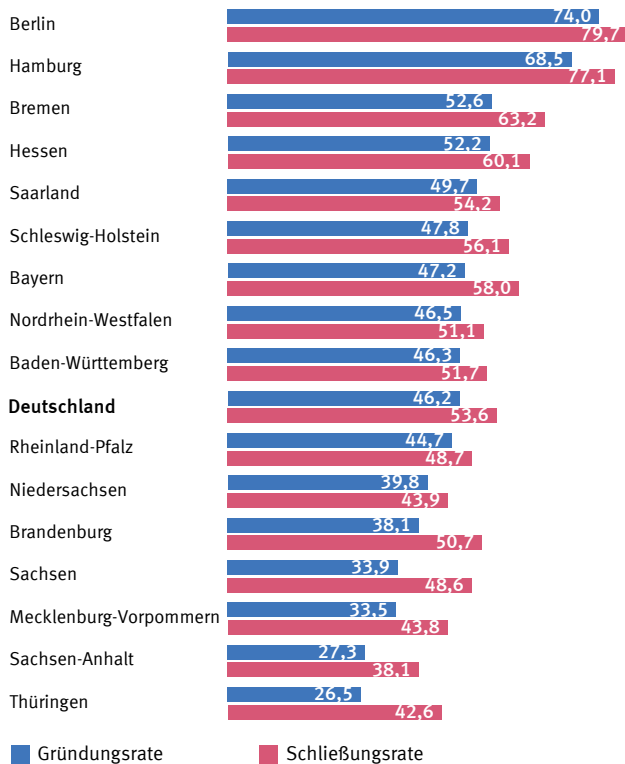
In Grafik 6 ist der Quantil-Quantil-Plot für die logarithmierte Gründungsrate anhand der Erwerbsfähigen dargestellt. Diese Gründungsrate folgt augenscheinlich einer Log-Normalverteilung und zeigt damit die einer gruppierten Verteilung innewohnende Endlastigkeit. Tatsächlich liegen die beobachteten Häufigkeiten in den Größenklassen hoher Gründungsdaten sogar noch geringfügig über den theoretischen Häufigkeiten der bereits rechtsschiefen Log-Normalverteilung. Anders verhält es sich mit der Gründungsrate, die wie in Grafik 2 anhand einer Standardisierung an der Zahl aktiver Unternehmen erhalten wurde. Diese hat die Eigenschaft einer Log-Normalverteilung verloren und nähert sich eher einer symmetrischen Normalverteilung, was mit dem Testergebnis des Shapiro-Wilk Normality-Test zusätzlich nahegelegt wird. Die charakteristisch gruppierten Verteilungseigenschaften von Unternehmensgründungen werden also besser durch die Standardisierung der Unternehmensgründungen an der erwerbsfähigen Bevölkerung dargestellt.

Letztlich sind Gründungsdaten ein Weg, um die unterschiedliche Gründungsintensität zwischen verschiedenen Regionaleinheiten vergleichbar zu gestalten. Die vorangegangene Diskussion zeigt, dass die Wahl des Standardisierungsverfahrens einen erheblichen Einfluss auf die Aussagekraft der Gründungsdaten hat. Diese Vergleichbarkeit wird somit für Kreise wie auf den oben gezeigten Karten hergestellt. Sie ist aber auch geeignet, um die Gründungsintensität für einzelne Bundesländer darzustellen.

Für eine vollständige Darstellung der Ergebnisse zur regionalen Unternehmensdemografie zeigen die folgenden beiden Grafiken die Gründungs- und Schließungsdaten sowie die Überlebensraten von Unternehmen (zum Zeitpunkt $t-3$) für Deutschland unterteilt nach Bundesländern. Die in den dargestellten Karten erkennbare Struktur hoher Gründungsdaten in Agglomerations-

Grafik 7

Gründungs- und Schließungsraten im Berichtsjahr 2021 je 10 000 Erwerbsfähige im Alter von 15 bis unter 65 Jahren



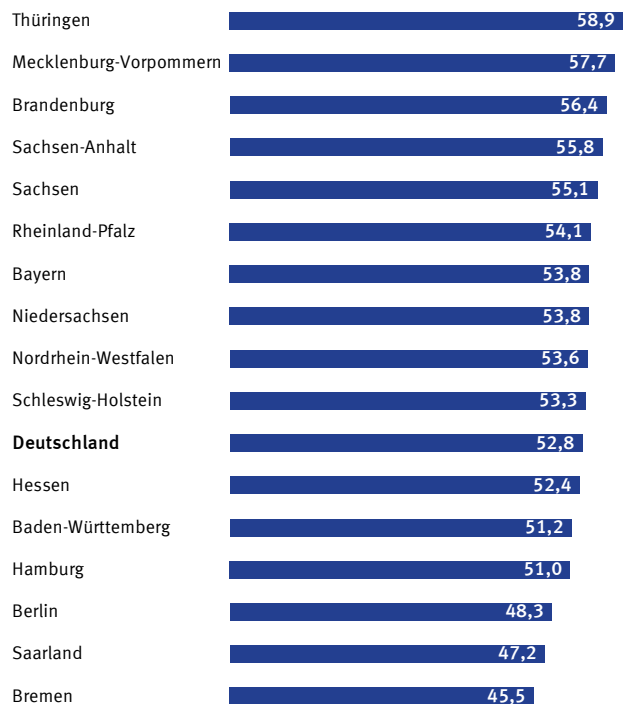
räumen spiegelt sich auch hier wider. [↗ Grafik 7](#) zeigt, dass Berlin und Hamburg sowohl mit hohen Gründungs- als auch Schließungsraten führen. Das darf als Hinweis auf eine höhere Dynamik in einer städtisch geprägten Unternehmenslandschaft gelten.

Für die Normierung von Überlebensraten ist es methodisch sinnvoll, auf eine Standardisierung der Überlebensrate der neu gegründeten Unternehmen anhand des zum Zeitpunkt vor n Jahren gegründeten Unternehmensbestandes zurückzugehen. Das ist angemessen, da hier die Aussage im Vordergrund steht, wie sich diese Unternehmensgründungen am Markt behaupten. [↗ Grafik 8](#) stellt daher den Vergleich der Bundesländer anhand der überlebenden Unternehmen dividiert durch die Zahl der Unternehmensgründungen in Prozent dar.

Bei einem Blick auf die Überlebensraten der Unternehmen nach Bundesländern schlagen sich niedrige Gründungs- und Schließungsraten und damit die den Grün-

Grafik 8

Überlebensraten nach drei Jahren von neu gegründeten Unternehmen im Produzierenden Gewerbe und Dienstleistungsbereich¹ im Jahr 2021 in %

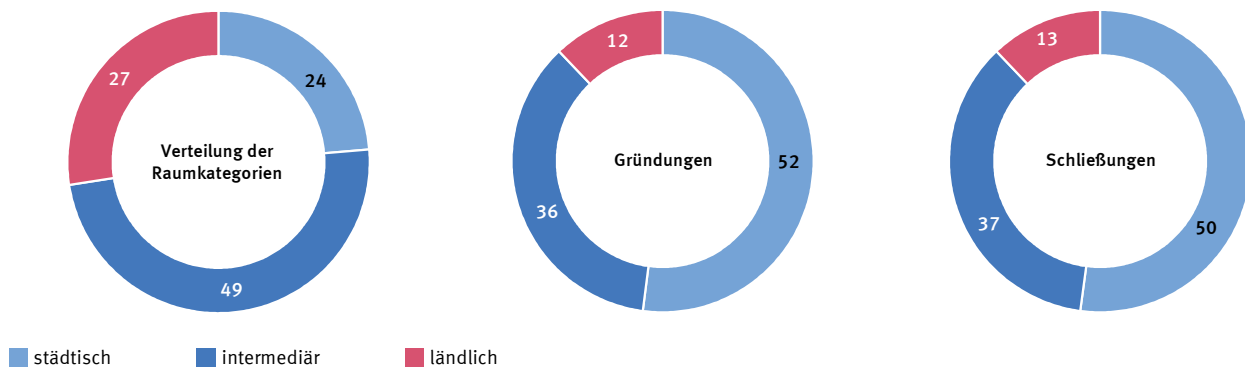


¹ Abschnitte B bis N und P bis S (ohne S 94) der Klassifikation der Wirtschaftszweige, Ausgabe 2008 (WZ 2008).

dungen und Schließungen innewohnende geringere Dynamik in tendenziell höheren Überlebensraten nieder. Bundesländer mit höheren Gründungs- und Schließungsraten weisen eine höhere Dynamik des Gründungsgeschehens auf und haben fallweise geringere Überlebensraten. Das betrifft augenscheinlich vor allem einige ostdeutsche Bundesländer, deren Gründungs- und Schließungsraten niedriger, deren Überlebensraten jedoch leicht höher als im gesamtdeutschen Vergleich sind.

Grafik 9

Gründungen und Schließungen in städtischen, intermediären und ländlichen Regionen im Berichtsjahr 2021
in %



Mehr Unternehmensgründungen in den städtischen Regionen

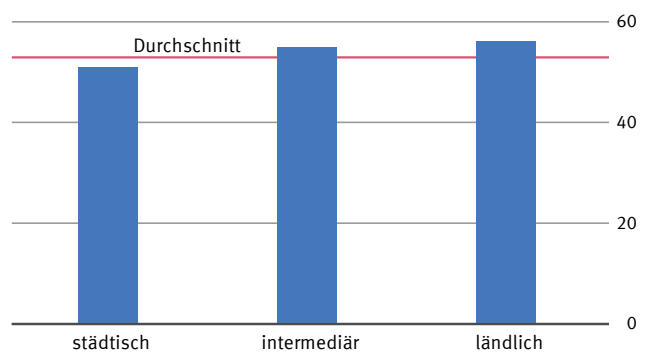
Neben der klassischen regionalen Gliederung nach Bundesländern und Kreisen ist auch eine strukturelle Gliederung der Regionen möglich. Hierbei wird jeder Kreis als städtische, ländliche oder intermediäre (Vororte, kleinstädtische Bereiche) Region⁴ klassifiziert.

Bei einem Vergleich der Wirtschaftsdynamik in [Grafik 9](#) unterscheiden sich der städtische Bereich, die intermediären Bereiche und die ländlichen Regionen deutlich. Betrachtet man hingegen die Verteilung der Regionen in Deutschland, dominieren intermediäre Kreise. Der Anteil der Gründungen im städtischen Bereich ist doppelt so hoch wie deren Anteil an allen Regionen.

[Grafik 10](#) zeigt dies auch im Vergleich der Überlebensraten der Verstädterungstypen. In der Stadt wird deutlich mehr gegründet, jedoch sind diese Gründungen weniger dauerhaft als auf dem Land. In der Stadt überleben nur 51 % der Neugründungen die ersten drei Jahre. Auf dem Land überleben von neu gegründeten Unternehmen nach drei Jahren 56 % und damit rund fünf Prozentpunkte mehr als in der Stadt. Die Unternehmensgründungen im ländlichen Raum sind somit nachhaltiger für die Wirtschaft.

Grafik 10

Überlebensraten von vor drei Jahren gegründeten Unternehmen im Berichtsjahr 2021 nach Raumkategorien
in %

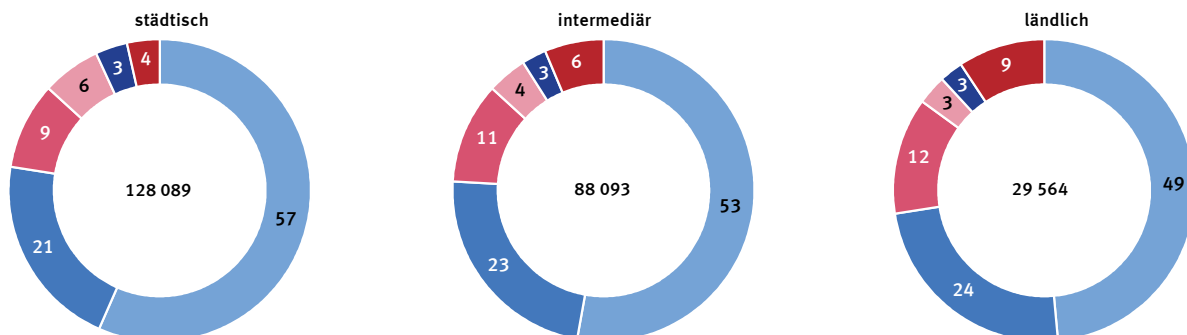


⁴ Zur Klassifikation siehe ec.europa.eu

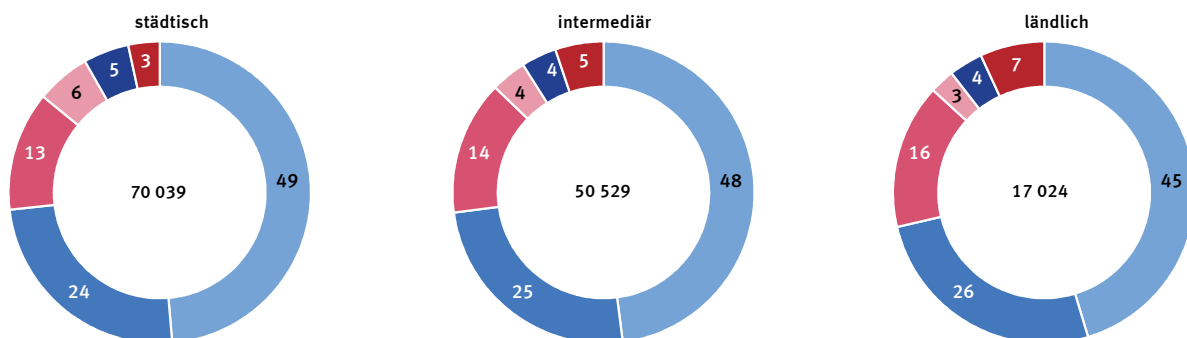
Grafik 11

Gründungen in städtischen und ländlichen Regionen im Berichtsjahr 2021 nach Wirtschaftsbereichen in %

Gründungen insgesamt



Arbeitgebergründungen



■ Dienstleistungen (K-N, P-S)
 ■ Handel und Gastgewerbe (G+I)
 ■ Baugewerbe (F)
 ■ Information/Kommunikation (J)
 ■ Verkehr (H)
 ■ Produzierendes Gewerbe (ohne Baugewerbe) (B-E)

Abschnitte der Klassifikation der Wirtschaftszweige, Ausgabe 2008 (WZ 2008).

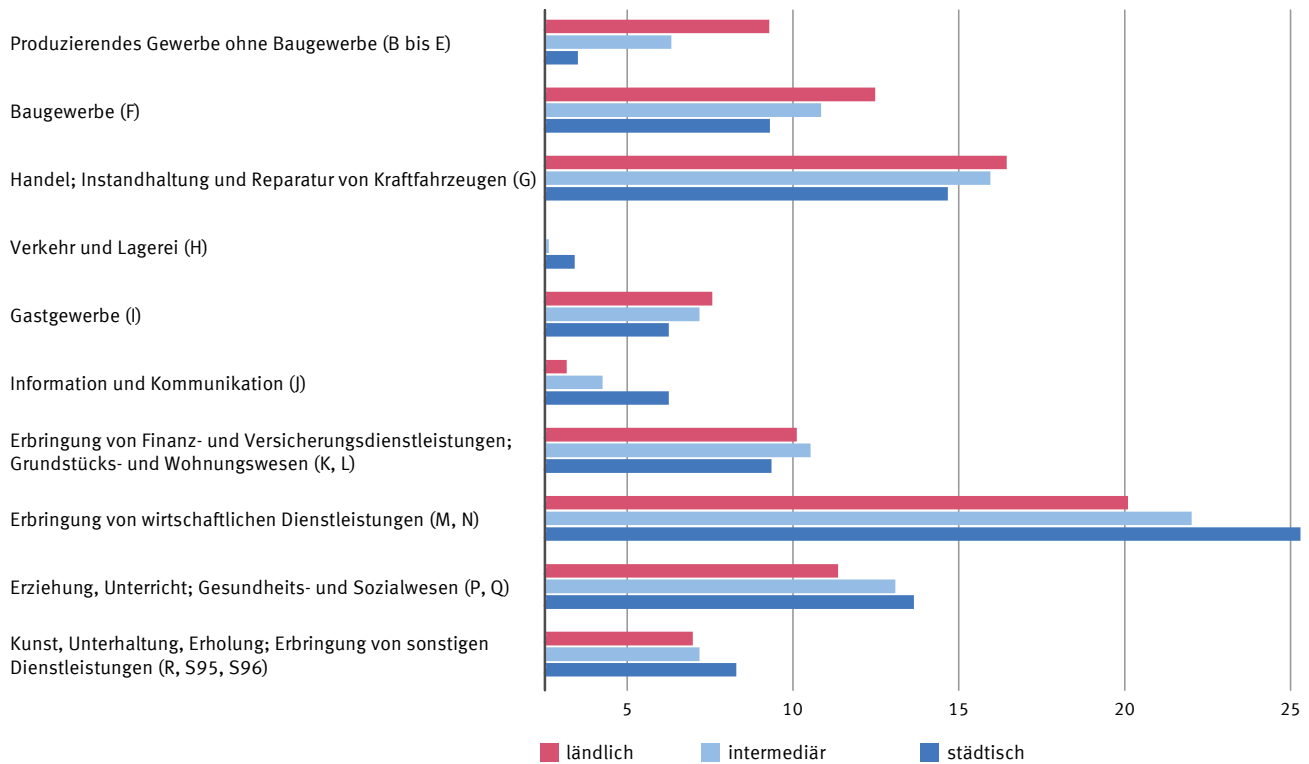
Nach Wirtschaftsbereichen untergliedert, entfallen 57 % aller Gründungen in der Stadt auf die Dienstleistungsbereiche (Abteilungen P bis R sowie Abschnitte S95 und S96 der WZ 2008), gefolgt von 21 % Handel und Gastgewerbe. Der Anteil der Arbeitgebergründungen in den Dienstleistungsbereichen liegt in der Stadt bei 49 % und ist somit deutlich geringer als im Durchschnitt.

➤ Grafik 11

In den ländlichen Regionen liegt der Anteil der Gründungen im Dienstleistungsbereich bei 49 %. Hier weisen Handel und Gastgewerbe (24 %), Baugewerbe (12 %) und Produzierendes Gewerbe (9 %) höhere Anteile als in der Stadt auf. Die Verteilung zwischen Gründungen und Arbeitgebergründungen ist fast identisch.

Grafik 12

Gründungsraten je Raumkategorie gemessen an allen Gründungen je Raumkategorie im Berichtsjahr 2021
in %



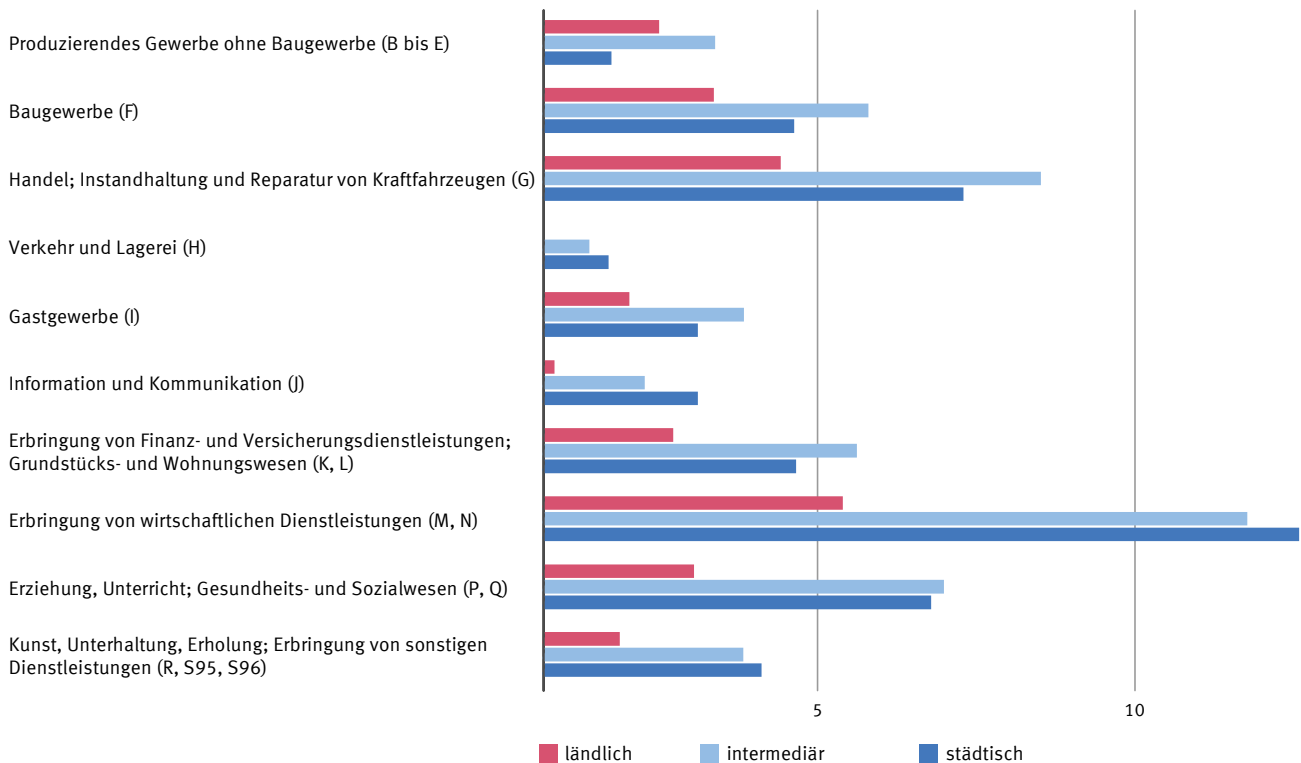
Abteilungen der Klassifikation der Wirtschaftszweige, Ausgabe 2008 (WZ 2008).

➤ **Grafik 12** unterscheidet das Gründungsgeschehen nach Raumkategorie und Wirtschaftsabschnitt. Im Bereich Erbringung von wirtschaftlichen Dienstleistungen (Abteilungen M und N der WZ 2008) liegt der Anteil der Neugründungen vor allem im städtischen Bereich höher. Im Gegensatz dazu liegt der Anteil an neugegründeten Unternehmen in den Wirtschaftsbereichen Produzierendes Gewerbe, Energieversorgung, Bau und Handel im ländlichen Bereich höher. Diese Verteilung hängt vermutlich mit dem größeren Flächenbedarf für die Schaffung von Produktionsfaktoren in den Wirtschaftsbereichen Verarbeitendes Gewerbe, Energieversorgung, Bau und Handel zusammen. Ein Bauunternehmen, welches schweres Gerät und Baumaterial benötigt, braucht mehr Raum als ein Start-up im IT-Bereich, bei dem ein großer Teil der Beschäftigten im Homeoffice arbeitet.

➤ **Grafik 13** normiert die Gründungsraten auf die erwerbstätige Bevölkerung. Besonders in den Dienstleistungsbereichen und bei Information und Kommunikation liegen die höchsten Gründungsraten in den städtischen Regionen. In den übrigen Wirtschaftsbereichen dominiert meist die intermediäre Raumkategorie.

Grafik 13

Gründungsraten je Raumkategorie im Berichtsjahr 2021 nach zusammengefassten Wirtschaftsabschnitten¹
je 10 000 Erwerbsfähige im Alter von 15 bis unter 65 Jahren



¹ Klassifikation der Wirtschaftszweige, Ausgabe 2008 (WZ 2008).


4

Fazit

Die regionale Darstellung der Unternehmensdemografie liefert wichtige Informationen für Wirtschaft und Politik. Bei der regionalen Analyse ist dabei die Auswahl der entsprechenden Vergleichsbasis entscheidend.

Bei den Gründungs- und Schließungsraten bietet die Standardisierung auf die erwerbsfähige Bevölkerung Vorteile gegenüber der Normierung auf den Bestand der aktiven Unternehmen, beispielsweise können die Daten leichter mit den Gewerbeanmeldungen verglichen werden. Die Überlebensraten der neugegründeten Unternehmen sind hingegen nur mit Bezug zum Bestand der aktiven Unternehmen sinnvoll interpretierbar.

Neben der Darstellung in administrativen Gliederungen ist es aufschlussreich, auf regionaler Ebene die administrativen Einflüsse zu reduzieren und eine rein geobasierte Darstellung zu wählen. Hierbei wurden verschiedene Ansätze skizziert, nämlich die hexagonale Darstellung der Gründungen bundesweit und die punktgenaue Verortung der Unternehmensgründungen als Basis einer Kartendarstellung am Beispiel Bayerns.

Im regionalen Bereich bietet die Unternehmensdemografie noch viele Möglichkeiten zur weiteren Analyse. 

LITERATURVERZEICHNIS

European Communities. *Eurostat – OECD Manual on Business Demography Statistics*. 2007a. [Zugriff am 8. Mai 2024]. Verfügbar unter: ec.europa.eu

European Communities. *Regions in the European Union – Nomenclature of territorial units for statistics*. 2007b. [Zugriff am 8. Mai 2024]. Verfügbar unter: ec.europa.eu

Europäische Union. *Europäisches System Volkswirtschaftlicher Gesamtrechnungen – ESVG 2010*. 2014. [Zugriff am 8. Mai 2024]. Verfügbar unter: ec.europa.eu

Möding, Patrizia/Philipp, Katja. *Erweiterte Auswertungen mit dem Unternehmensregister*. In: *Wirtschaft und Statistik*. Ausgabe 4/2007, Seite 342 ff.

Rink, Anke/Seiwert, Ines/Opfermann, Rainer. *Unternehmensdemografie: methodischer Ansatz und Ergebnisse 2005 bis 2010*. In: *Wirtschaft und Statistik*. Ausgabe 6/2013. Seite 422 ff.

Rink, Anke/Seiwert, Ines. *Aktuelle Entwicklungen in der Unternehmensdemografie*. In: *WISTA Wirtschaft und Statistik*. Ausgabe 2/2021, Seite 41 ff.

Rödel, Raimund/Stephan, Frank. *Von den „Neugründungen“ in der Gewerbeanzeigenstatistik bis zur „Gründung“ in der Unternehmensdemografie – Eine Analyse auf der Basis des Statistischen Unternehmensregisters in Bayern im Berichtsjahr 2019*. In: *Bayern in Zahlen*. Ausgabe 02/2022, Seite 33 ff. [Zugriff am 8. Mai 2024]. Verfügbar unter: www.statistik.bayern.de

RECHTSGRUNDLAGEN

Durchführungsverordnung (EU) 2020/1197 der Kommission vom 30. Juli 2020 zur Festlegung technischer Spezifikationen und Einzelheiten nach der Verordnung (EU) 2019/2152 des Europäischen Parlaments und des Rates über europäische Unternehmensstatistiken, zur Aufhebung von zehn Rechtsakten im Bereich Unternehmensstatistiken (Amtsblatt der EU Nr. L 271, Seite 1).

Verordnung (EG) Nr. 1059/2003 des Europäischen Parlaments und des Rates vom 26. Mai 2003 über die Schaffung einer gemeinsamen Klassifikation der Gebiets-einheiten für die Statistik (NUTS) (Amtsblatt der EU Nr. L 154, Seite 1).

Verordnung (EG) Nr. 295/2008 des Europäischen Parlaments und des Rates vom 11. März 2008 über die strukturelle Unternehmensstatistik. (Amtsblatt der EU Nr. L 97, Seite 13).

Verordnung (EG) Nr. 250/2009 der Kommission vom 11. März 2009 zur Durchführung der Verordnung (EG) Nr. 295/2008 des Europäischen Parlaments und des Rates im Hinblick auf die Definitionen der Merkmale, das technische Format für die Datenübermittlung, die erforderlichen Doppelmeldungen gemäß NACE Rev. 1.1 und NACE Rev. 2 und die zuzulassenden Abweichungen bei der strukturellen Unternehmensstatistik. (Amtsblatt der EU Nr. L 86, Seite 1).

Verordnung (EG) Nr. 251/2009 der Kommission vom 11. März 2009 zur Durchführung und Änderung der Verordnung (EG) Nr. 295/2008 des Europäischen Parlaments und des Rates im Hinblick auf die zu erstellenden Datenreihen für die strukturelle Unternehmensstatistik bzw. die nach der Überarbeitung der statistischen Güterklassifikation in Verbindung mit den Wirtschaftszweigen (CPA) erforderlichen Anpassungen. (Amtsblatt der EU Nr. L 86, Seite 170).

Verordnung (EU) 2019/2152 des Europäischen Parlaments und des Rates vom 27. November 2019 über europäische Unternehmensstatistiken, zur Aufhebung von zehn Rechtsakten im Bereich Unternehmensstatistiken (Amtsblatt der EU Nr. L 327, Seite 1).

Herausgeber
Statistisches Bundesamt (Destatis), Wiesbaden

Schriftleitung
Dr. Daniel Vorgrimler
Redaktion: Ellen Römer

Ihr Kontakt zu uns
www.destatis.de/kontakt

Erscheinungsfolge
zweimonatlich, erschienen im Juni 2024
Ältere Ausgaben finden Sie unter www.destatis.de sowie in der [Statistischen Bibliothek](#).

Artikelnummer: 1010200-24003-4, ISSN 1619-2907

© Statistisches Bundesamt (Destatis), 2024
Vervielfältigung und Verbreitung, auch auszugsweise, mit Quellenangabe gestattet.