Scientific-Use-Files zu den Lebenslagenbefragungen 2023 – Anonymisierungskonzept –

1. Einführung

Das vorliegende Konzept beschreibt die Anonymisierung der Lebenslagenbefragungen der Bürgerinnen und Bürger und Unternehmen 2023 zur Bereitstellung von den beiden faktisch anonymisierten Scientific Use Files (SUFs). Auf Antrag können diese Daten von den Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder für wissenschaftliche Projekte zur Verfügung gestellt werden.

Das Bundesstatistikgesetz § 16 Abs. 6 ermöglicht die Nutzung von Individualdaten der Statistischen Ämter des Bundes und der Länder durch die Wissenschaft, „wenn die Einzelangaben nur mit einem unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft zugeordnet werden können“. Dies wird als faktische Anonymisierung bezeichnet. Forschungsarbeiten beschäftigten sich in der Vergangenheit genauer damit, wie sich diese Anforderungen interpretieren und praktisch umsetzen lassen (Höhne, 2010; Ronning und andere, 2005; Ronning und andere, 2003; Müller und andere, 1991). Hierbei ist eine Kompromisslösung nötig. Auf der einen Seite muss die Anonymität der Befragten ausreichend gesichert sein. Auf der anderen Seite soll der Wissenschaft ein Höchstmaß an Analysepotenzial geboten werden.

Zur Bereitstellung des anonymisierten Datensatzes werden mögliche Analysen bereits antizipiert, das Risiko der Identifikation dabei eingeschätzt und die Daten gegebenenfalls entsprechend abgeändert. Der faktischen Anonymität wird darüber hinaus aber auch dadurch entsprochen, dass Datennutzende vertraglich zusichern müssen, sich an das Deanonymisierungsverbot zu halten und dass sie bei Fehlverhalten eine Vertragsstrafe zu zahlen haben. Zudem sind Analysen nur im Rahmen des angegebenen wissenschaftlichen Vorhabens gestattet und die Nutzenden werden nach § 16 Abs. 7 BStatG auf die Wahrung der statistischen Geheimhaltung besonders verpflichtet.

Der folgende Abschnitt beschreibt die Lebenslagen-Daten der Bürgerinnen und Bürger und Unternehmen 2023 und ihren Schutzbedarf. Der Hauptteil des Konzepts fasst wissenschaftliche Empfehlungen zur faktischen Anonymisierung zusammen und erläutert deren Umsetzung zur Bereitstellung eines entsprechenden SUFs für die Lebenslagenbefragungen 2023. Das Konzept schließt mit einem Fazit.

2. Datenmaterial

Die Lebenslagenbefragungen des Statistischen Bundesamtes liefern Informationen zum Kontakt der Bürgerinnen und Bürger sowie der Unternehmen mit der der öffentlichen

Verwaltung und zu ihrer Zufriedenheit mit der Dienstleistungserbringung durch die zuständigen Behörden.

Seit 2015 fanden alle zwei Jahre Lebenslagenbefragungen statt. Die Daten von 2019 konnten erstmals der Wissenschaft über faktische Anonymisierungsmaßnahmen zur Verfügung gestellt werden, da die von den Befragten abzugebende Datenschutzerklärung zwischenzeitlich angepasst wurde. Die fünfte Welle an Personen und Unternehmen wurde 2023 befragt. Auch die aktuellen Daten werden im Forschungsdatenzentrum angeboten.

Die Grundgesamtheit bei der Befragung setzt sich bei den Bürgerinnen und Bürgern aus der Wohnbevölkerung Deutschlands ab 16 Jahren und bei den Unternehmen aus den Firmen mit mindestens einem Behördenkontakt innerhalb der letzten zwei Jahre innerhalb einer von 23 bei den Bürgerinnen und Bürgern und 11 bei den Unternehmen vordefinierten Lebenslagen zusammen. Mit dem Begriff Lebenslagen werden dabei typische Situationen im Leben der Betroffenen bzw. im Lebenszyklus des Unternehmens bezeichnet, in denen sie Kontakt zur öffentlichen Verwaltung gehabt haben können. Die Listen der Lebenslagen wurden erstmals anhand einer eigenständigen Erhebungen ermittelt und anschließend von Welle zu Welle aktualisiert (siehe Übersichten 1 und 2).

Übersicht 1: 23 bei den Bürgerinnen und Bürgern abgefragte Lebenslagen 2023

- Berufsausbildung
- Studium
- Beantragung von Ausweisdokumenten
- Führerschein/Fahrzeugregistrierung
- Arbeitslosigkeit
- Finanzielle Probleme
- Steuererklärung
- Heirat/Lebenspartnerschaft
- Scheidung/Aufhebung Lebenspartnerschaft
- Geburt eines Kindes
- Kinderbetreuung
- Umzug
- Immobilienerwerb
- Eintritt in Ruhestand
- Altersarmut
- Gesundheitliche Willensbekundung
- Längerfristige Krankheit
- Behinderung
- Pflegebedürftigkeit
- Tod einer nahestehenden Person
- Ehrenamtliche Tätigkeit im Verein
- Kontakt mit dem Gesundheitsamt während der Corona-Pandemie
- Beteiligung an einem Gerichtsverfahren

Übersicht 2: 11 bei den Unternehmen abgefragte Lebenslagen 2023

- Gründung eines Unternehmens
- Steuern und Finanzen
- Einstellen von Beschäftigten
- Ausbildung
- Arbeitssicherheit und Gesundheitsschutz
- Bau einer Betriebsstätte
- Forschung & Entwicklung, Patent- und Markenschutz
- Teilnahme an Ausschreibungsverfahren
- Import / Export
- Geschäftsaufgabe und –übergabe
- Beantragung von Corona-Hilfen

Lebenslagen mit mehreren sich systematisch unterscheidenden Verwaltungsverfahren wurden darüber hinaus für die Befragungen in homogene Untergruppen unterteilt und die relevanten Kriterien zusätzlich erfragt. Bei der Kinderbetreuung wurde beispielsweise zwischen vorschulischer und schulischer Kinderbetreuung unterschieden und bei der Geschäftsaufgabe und -übergabe wurde betrachtet, ob eine Insolvenz beantragt wurde oder nicht. Für jede der so unterschiedenen insgesamt 58 Lebenslagenschichten war eine zuvor bestimmte Anzahl an Interviews zu realisieren. Diese bestimmte sich nach verschiedenen Kriterien wie z. B. der Anzahl möglicher Behördenkontakte in der Schicht.

Für die Stichprobenziehung und die anschließenden telefonischen und online-Interviews beauftragte das Statistische Bundesamt über eine öffentliche Ausschreibung 2023 *Ipsos*. Die Stichprobenziehung für die größere Telefonstrichprobe erfolgte bei den Bürgerinnen und Bürgern anhand zufällig generierter Telefonnummern nach dem ADM-Dual-Frame-Ansatz für Festnetz- und Mobilfunknummern (Glemser und andere, 2014). Die Ziehung der Firmen für die Unternehmensbefragung erfolgte zufällig aus der externen Adressdatenbank w:inform (Himmelsbach und andere, 2016).

Daneben wurde eine kleinere Zusatzstichprobe online befragt, um die Möglichkeit einer Umstellung von CATI auf Online zu testen. Die Stichprobe der Bürgerinnen und Bürger wurde dabei aus dem Ipsos Online Access Panel gezogen. Zur Kontaktierung der Unternehmen wurden zufällig Unternehmensadressen mit zugehörigen E-Mail-Adressen aus dem umfassenden Datenbestand des Anbieters Dun & Bradstreet gezogen. Bei allen Stichproben wurden verschiedene Quotenpläne verwendet, um möglichst repräsentative Nettostichproben z.B. hinsichtlich Verteilung der Bundesländer, Alter, Geschlecht, verwendetem Endgerät, Unternehmensgrößenklasse und Branche zu gewährleisten.

Die beiden Befragungen gliederten sich jeweils analog in mehrere Teile. Zunächst wurden die Bürgerinnen und Bürger bzw. Geschäftsführenden von Unternehmen dazu interviewt, welche der vordefinierten Lebenslagenschichten auf sie selbst bzw. ihr Unternehmen innerhalb der vergangenen zwei Jahre zugetroffen hatten. Diesen Screening-Teil beantworteten insgesamt etwa 24 750 Personen und 6 100 Unternehmen. Hier wurde auch abgeglichen, welches Alter auf die Befragten bzw. welche Rechtsform

auf das Unternehmen befragten zutraf, um beurteilen zu können, ob diese wirklich zur Zielgruppe gehören. Anhand der Antworten in diesem Abschnitt beurteilte die Software, ob die Befragten für Interviews zu Lebenslagenschichten mit noch zu realisierenden Interviews in Frage kamen.

Traf dies zu, so wurde eine der Schichten gezogen und die Person innerhalb dieser näher dazu befragt, mit welchen Behörden sie Kontakt hatte und um welche konkreten Anliegen es dabei ging. Hierzu identifizierte das Statistische Bundesamt vorab für jede Lebenslage die dazugehörigen Ämter und die von ihnen angebotenen Dienstleistungen und ließ deren Relevanz im Prozess von betroffenen Personen, Unternehmen sowie Expertinnen und Experten regelmäßig validieren.

Sofern die Personen von einem befragungsrelevanten Behördenkontakt berichteten, so wählte der Computer zufällig eine der angegebenen Dienstleistungen aus und startete das detaillierte Interview. Zunächst konnten die Befragten dabei angeben, wie zufrieden sie insgesamt mit der Bearbeitung ihres jeweiligen Anliegens durch die Behörde waren. Etwaige Probleme durften sie in einer offenen Frage erläutern. Anschließend beantworteten die Betroffenen detaillierte Fragen zu den von ihnen genutzten und präferierten Kommunikationsmedien mit den Behörden. Falls keine elektronischen Medien für den Austausch genutzt wurden, wurden zudem die Gründe hierfür erfragt.

Anschließend beurteilten die Interviewten ihre Zufriedenheit mit der Erbringung der jeweiligen Dienstleistung durch die Behörde genauer anhand von 17 Zufriedenheitsfaktoren (siehe Übersicht 3).

Übersicht 3: 18 Zufriedenheitsfaktoren

- Informationen zu Verfahrensschritten und zum weiteren Ablauf
- Verständlichkeit der Formulare und Anträge
- Zugang zu notwendigen Formularen und Anträgen
- Online-Angebote
- Zugang zur richtigen Stelle
- Räumliche Erreichbarkeit
- Öffnungszeiten
- Wartezeit
- Hilfsbereitschaft
- Fachkompetenz
- Gesamte Verfahrensdauer
- Vertrauen in Behörde
- Diskriminierungsfreiheit
- Unbestechlichkeit
- Verständlichkeit des Rechts
- Verständlichkeit offizieller Schreiben
- Digitale Transaktionsmöglichkeit

Weiterhin wurden sie gefragt, ob sie bei dem Anliegen ihr Ziel erreicht hatten, inwiefern das Ergebnis ihren Erwartungen entsprach, welche Zeit sie aufgewendet hatten, um die jeweilige Dienstleistung zu beantragen, wie lange die Bearbeitungszeit bei der Behörde war, als wie kompliziert sie das jeweilige Verfahren einstufen würden und falls zutreffend warum sie es kompliziert fanden. Sodann gaben die Befragten die von ihnen genutzten Informationsquellen sowie die Zufriedenheit mit ausgewählten Quellen an. Schließlich durften die Personen noch etwaige Verbesserungsvorschläge zur Erbringung der jeweiligen Dienstleistung durch die Behörde benennen.

Sofern die Befragten weitere relevante Behördenkontakte angegeben hatten, wurden sie anschließend um ihre Einschätzung zu weiteren zusätzlichen zufällig ausgewählten Dienstleistungen gebeten. Auf freiwilliger Basis konnte eine Person Fragen zu bis zu zwölf Dienstleistungen im Rahmen von maximal drei verschiedenen Lebenslagen beantworten.

Im letzten Fragebogenteil wurden den Bürgerinnen und Bürgern noch allgemeine Fragen zu ihrem sozioökonomischen Hintergrund sowie zu den für die Designgewichtung relevanten Informationen wie der Haushaltsgröße und der Anzahl an Telefonnummern unter denen sie erreichbar waren gestellt. Bei den Unternehmen wurden analog die Merkmale Unternehmensalter und Umsatzgrößenklasse sowie die für die Gewichtung noch relevanten Informationen über die Branche und Beschäftigtengrößenklasse erhoben.

Ein Interview zählt zum Nettodatensatz, wenn zur jeweiligen Dienstleistung die detaillierten Antworten vorliegen und die erforderlichen Informationen für die Designgewichtung bekannt sind. Insgesamt wurden ca. 7 600 Personen und 3 100 Unternehmen erfolgreich interviewt. Es liegen von ihnen insgesamt Angaben zu rund 9 500 individuellen bzw. 3 900 unternehmensspezifischen Behördenkontakten vor.

Als Ergebnis veröffentlicht wurde die Gesamtzufriedenheit als einfaches arithmetisches Mittel aus allen abgefragten Faktorzufriedenheiten (Statistisches Bundesamt, 2019a-2019d). Hintergrund ist, dass dieser Wert die Zufriedenheit anhand von objektivierten Kriterien widerspiegelt und nicht so stark situativ schwankt wie die direkt abgefragte Gesamtzufriedenheit der Befragten. Bei allen Analysen wurde das Gesamtdesigngewicht zur Korrektur der unterschiedlichen Auswahlwahrscheinlichkeiten der Personen innerhalb der Haushalte bzw. der Unternehmen nach Branche und Größenklasse, sowie der Lebenslagen, Schichten, Behörden und Dienstleistungen für ein Interview angewendet. Die Nutzung dieses Gewichtungsfaktors wird auch für unabhängige Analysen mit den Daten empfohlen, um die Ergebnisse von der Stichprobe auf die Grundgesamtheit zu projizieren.

Die Datensatzbeschreibungen und die Fragebogen als leserfreundliche Version und als Programmiervorlage inklusive Interviewhinweise und Filterführung werden mit den SUFs bereitgestellt. Weitere Details zur Methodik der Lebenslagenbefragungen beschreiben Himmelsbach und andere (2016), Walprecht und andere (2020), Schmidt und andere (2015) und Kühnhenrich/Michalik (2019).

Die Daten der Lebenslagenbefragungen sind zum Teil sensibel, etwa wenn es darum geht, welche Person bzw. welches Unternehmen welche (Hilfs-)leistungen beantragt hat. Daneben sind einige sozioökonomische Merkmale wie der Behinderungsgrad, die Staatsbürgerschaft, ein etwaiger Migrationshintergrund und das Haushaltseinkommen der Befragten oder auch die Umsatzgrößenklasse eines Unternehmens, sofern letztere nicht öffentlich einsehbar ist, als private Informationen und damit als schutzbedürftig einzustufen.

3. Anonymisierung

3.1 Vorüberlegungen

Wie eingangs dargestellt bestehen die beiden in Einklang zu bringenden Ziele der faktischen Anonymisierung darin, die Anonymität der Befragten zu sichern und zugleich möglichst umfassende und valide Analysen zu ermöglichen. Diese Kriterien sind jedoch gegensätzlich: Je stärker Individualdaten anonymisiert werden, desto höher ist zwar der Schutz, desto geringer fallen jedoch auch Bandbreite und Qualität möglicher Analysen aus (Höhne, 2010). Deshalb muss eine Kompromisslösung gefunden werden.

Beim Löschen und Vergröbern von Variablen reduziert sich der Informationsgehalt der Daten. Bei datenverändernden Methoden wie dem Löschen, Tauschen oder der Imputation von Beobachtungen können darüber hinaus die Fallzahlen für mögliche Analysen sinken und es werden Verteilungsmomente der Daten wie Mittelwert, Varianz, Verteilung und Kovarianzen beeinflusst (Ronning/Gnoss, 2003). Im schlimmsten Fall sind die Daten und somit auch Analyseergebnisse nach der Anonymisierung nicht mehr repräsentativ. In welchem Umfang und nach welcher Methode anonymisiert wird, ist deshalb von entscheidender Bedeutung.

Im Regelfall ist der den Befragten zugesicherte Schutz ihrer Einzeldaten als gewichtiger einzuschätzen als die Möglichkeit zusätzlicher wissenschaftlicher Analysen. Die Maßgabe ist dabei laut Gesetz, dass es nur mit unverhältnismäßig großem Aufwand möglich sein soll, Einzeldaten zuzuordnen. In der Vergangenheit wurde dies in der Praxis zum Teil eher großzügig gehandhabt.

Es sind inzwischen allerdings diverse erfolgreiche Identifizierungsversuche dokumentiert, bei denen schutzbedürftige Daten über bekannte Identifikatoren mit anderen Informationsquellen zusammengespielt und so Einzelsubjekte identifiziert wurden. Dabei waren teilweise keine eindeutigen individuellen Schlüsselmerkmale wie der Name oder die Sozialversicherungsnummer der Person oder bekannte niedrigbesetzte und damit relativ eindeutige Variablenkombinationen wie die aus Postleitzahl und Geburtsdatum zum Abgleich nötig. In einigen Fällen wurden auch mehrere auf den ersten Blick harmlos erscheinende Variablen, welche in mehreren Datensätzen vorkommen (beispielsweise zu Filmpräferenzen oder mehrere sozioökonomische Merkmale) kombiniert, um eindeutige Fälle herauszufiltern (Rocher und andere 2019; Lubarsky, 2017).

Bei einem systematischen Literaturreview aus dem Jahr 2011 stellte sich das Risiko der Re-Identifikation durch Forscherinnen und Forscher als niedriger heraus, wenn gängige Standard-Empfehlungen eingehalten wurden (El Emam und andere, 2011). Die deutlich gestiegene Anzahl erfolgreicher Re-Identifikationsversuche in den letzten Jahren zeigt dennoch, dass die Unsicherheit seitdem deutlich gestiegen ist (Rocher und andere, 2019). Dies hängt sicherlich mit der kontinuierlich steigenden Anzahl verfügbarer Daten und den fortschreitenden technischen Möglichkeiten zusammen.

Technisch gesehen ist eine Kombination mehrerer Datensätze mit wenig Vorwissen und Standard-Tabellenkalkulationssoftware möglich. Selbst über den gedanklichen Abgleich seltener Merkmalskombinationen mit der Grundgesamtheit ist eine Identifikation denkbar, insbesondere wenn die angreifende Person die Information hat, dass eine bestimmte Person an einer Befragung teilgenommen hat. Beide Identifikationswege können ohne datenverändernde Maßnahmen nur vollständig ausgeschlossen werden, wenn keinerlei Merkmale mit eindeutigen Kombinationen im Datensatz enthalten sind, welche aus anderen Quellen bekannt sein könnten.

Empfehlungen aus der Literatur und das Vorgehen bei ähnlichen Projekten unter Berücksichtigung der genannten Probleme werden bei der im folgenden beschriebenen Anonymisierung der Lebenslagen-Daten berücksichtigt. Die auf die Lebenslagenbefragungen 2023 konkret angewandten Maßnahmen werden im Folgenden jeweils kurz erläutert und deren Anwendung auf die betroffenen Variablen im Datensatz beschrieben.

3.2 Ziehen einer Teilstichprobe

Durch Vorwissen über die Teilnahme bestimmter Personen oder Firmen an der Befragung können diese wie beschrieben identifiziert werden, falls die zugehörige Kombination von Schlüsselmerkmalen einmalig im Datensatz ist. Beispielsweise könnte jemandem aus Gesprächen bekannt sein, dass eine bestimmte Person oder ein Unternehmen Fragen zu einer seltenen Kombination behördlicher Dienstleistungen beantwortet hat. Das Risiko einer falschen Zuordnung steigt jedoch, wenn nur eine Zufallsstichprobe der Originaldaten vorliegt. Somit könnte es theoretisch sein, dass noch eine Person oder ein Unternehmen mit der seltenen Merkmalskombination an der Befragung teilnahm. Dann ist die Zuordnung mit Unsicherheit behaftet.

Die Fallzahl sinkt durch das Ziehen einer Substichprobe jedoch. Damit werden mögliche Analysen eingeschränkt. Aufgrund der in der Tiefe teilweise geringen Fallzahlen kann es zudem auch bei einer zufälligen Ziehung zu einem gewissen Einfluss auf detailliertere Analyseergebnisse kommen. In Abwägung beider Überlegungen wird im vorliegenden Fall jeweils zufällig eine 95 %- Stichprobe der Originaldaten (bezogen auf die Personen, respektive Unternehmen) gezogen.

Hierbei wird jeder Beobachtung im Datensatz durch den Computer eine systemfrei generierte Pseudozufallszahl zugewiesen. Durch die Vergabe eines festen Startwertes

(auch random seed) für den Zufallszahlengenerator ist dieses Prozedere replizierbar. Es folgt bei der Prozedur Proc Survey Select in SAS der Methode von Floyd (Bentley and Floyd, 1987). Die zugewiesenen Zahlen folgen einer stetigen Gleichverteilung zwischen 0 und 1, d.h. jede Zahl in diesem Intervall wird mit gleich großer Wahrscheinlichkeit ausgewählt. Die 5 Prozent der Beobachtungen mit Zufallswert über 0,95 werden schließlich aus der Stichprobe gelöscht.

3.3 Systemfreie Anordnung

Die Fragebogen-Nummerierung erfolgt systematisch nach der Reihenfolge der Teilnahme an der Befragung und birgt mit Zusatzwissen über den Zeitraum der Feldphase somit ein gewisses Identifikationsrisiko. Für Analysen ist eine Identifikationsnummer aber teilweise nötig, um Verläufe nachvollziehen zu können. Daher wird die Identifikationsnummer für die teilnehmenden Personen und Unternehmen zufällig neu generiert.

3.4 Entfernen von Variablen

Folgende Variablen bergen ein hohes Identifikationsrisiko oder werden für die Auswertung nicht unbedingt benötigt und werden daher vollständig gelöscht:

Beide Befragungen

- Interviewdatum
- Interviewdauer
- Teilstichprobe (Festnetz, mobil oder online)

Befragung der Bürgerinnen und Bürger

- Postleitzahl
- Wohnort
- Anzahl der Mobilfunknummern, unter denen die Befragungsperson erreichbar ist
- Anzahl der Festnetznummern unter der der Haushalt der befragten Person erreichbar ist
- Anzahl Kinder im Alter von 16 und 17 Jahren im Haushalt

Das Interviewdatum war für Plausibilitätsprüfungen am Datensatz bedeutsam und ist ansonsten nicht weiter relevant für Auswertungen. Wenn jemand den Tag eines Interviews kennen würde, dann würde zudem hierüber ein Wiedererkennungsrisiko bestehen. Gleiches gilt für die Interviewdauer.

Postleitzahl und Wohnort der Bürgerinnen und Bürger weisen in Kombination mit zusätzlichen soziodemographischen Merkmalen ein hohes Identifikationsrisiko auf, da spezifische Kombinationen in der Grundgesamtheit sehr selten sein können und somit eine eindeutige Zuordnung nicht ausgeschlossen werden könnte.

Anzahl der Telefonnummern, Teilstichprobe und Anzahl Kinder zwischen 16 und 17 Jahren im Haushalt sind lediglich für die Berechnung der Designgewichtung bei der Befragung der Bürgerinnen und Bürger relevant, welche für die unterschiedlichen Auswahlwahrscheinlichkeiten der Personen innerhalb der Haushalte korrigiert. Für Auswertungen der Daten ist die Gesamtgewichtungsvariable im Datensatz ausreichend.

3.5 Zusammenfassen von Variablenausprägungen

Werden Variablenausprägungen verringert, dann wird damit einer zu geringen Besetzung einzelner Ausprägungen vorgebeugt und es wird somit unwahrscheinlicher, einzelne Subjekte durch einmalige Kombinationen zu identifizieren. Grundsätzliche Zusammenhänge zu anderen Variablen bleiben bei dieser Methodik erhalten, was einen Vorteil gegenüber der Methode des Ersetzens von Werten darstellt, durch welche sich multivariate Zusammenhänge ändern und die Varianz verringern können. Je stärker eine Variable vergröbert wird, desto weniger Informationen liefert sie allerdings noch.

Wenn alle soziodemographischen bzw. Unternehmensmerkmale im jeweiligen Lebenslagen-Datensatz kombiniert werden, dann ergeben sich aufgrund der Vielzahl von Variablen selbst unter Aggregation noch eindeutige Kombinationen in der Stichprobe. Allerdings kommen diese Konstellationen in der der Grundgesamtheit (der deutschen Bevölkerung im Alter ab 16 Jahren bzw. der Unternehmen) mehrfach vor, wenn die übriggebliebenen Merkmale ausreichend vergröbert sind.

Durch das Ziehen der 95%-Stichprobe kann man dann selbst mit dem Wissen darüber, dass eine Person oder ein Unternehmen mit einer bestimmten Merkmalskombination an der Befragung teilgenommen hat, diese(s) nicht eindeutig identifizieren. Vor der Identifikation schützt auch, dass die sozioökonomischen und Unternehmensangaben auf freiwilliger Basis erteilt und nicht überprüft worden sind und daher zum Teil fehlende oder falsche Werte aufweisen. Aus den genannten Gründen werden die im Folgenden beschriebenen Maßnahmen als ausreichend eingestuft.

a) Befragung der Bürgerinnen und Bürger

bula Regionalinformation

Die fünfstellige Gemeindekennziffer kann nicht herausgegeben werden, da zu diversen Kreisen bzw. kreisfreien Städten nur sehr wenige Interviews vorliegen. Die gröbere Information über das Bundesland bietet demgegenüber ausreichend Schutz für die Teilnehmerinnen und Teilnehmer und ermöglicht gleichzeitig eine zumindest grobe regionale Zuordnung der Befragten. Als tiefste regionale Ebene wird daher die Angabe zum Bundesland bereitgestellt.

F1 Alter

Die numerische Variable für das Alter in Jahren wird in eine kategoriale Variable mit den Ausprägungen „unter 20 Jahre“, „20 bis 29 Jahre“, „30 bis 39 Jahre“, „40 bis 49 Jahre“, „50 bis 59 Jahre“ und „60 Jahre und älter“ überführt, da einzelne Altersjahre und vor allem die Ränder der Altersverteilung nur spärlich besetzt sind.

F60 Haushaltsgröße

Befragte mit 6 und mehr Personen im Haushalt werden gemeinsam in einer Kategorie ausgewiesen, anstatt die genaue Anzahl Haushaltsmitglieder anzugeben, da Haushalte mit mehr Mitgliedern relativ selten vorkommen.

F61 Kinder unter 18 Jahre im Haushalt

Befragte mit 4 und mehr Kindern unter 18 Jahre im Haushalt werden gemeinsam in einer Kategorie ausgewiesen, anstatt die genaue Anzahl an Kindern im Haushalt anzugeben, da Haushalte mit mehr Kindern eher selten sind.

F64 Familienstand

Folgende Kategorien werden aufgrund niedriger Fallzahlen bei den gleichgeschlechtlichen Ausprägungen jeweils zusammengefasst:

- „verheiratet“ und „in einer eingetragenen Lebenspartnerschaft“
- „geschieden“ und „in einer eingetragenen Lebenspartnerschaft, die aufgehoben wurde
- „verwitwet“ und „in einer eingetragenen Lebenspartnerschaft, bei der der Partner verstorben ist“

F57 Staatsbürgerschaft

Die Kategorien bei der Staatsbürgerschaft „Die des Vereinigten Königreichs (UK), der Schweiz, von Norwegen, Liechtenstein oder Island“ und „die eines anderen Landes“ werden zu einer Ausprägung „die eines anderen Landes außerhalb der EU“ zusammengefasst.

F55 Höchster Bildungsabschluss

Die beiden Kategorien „noch in Schulausbildung“ und „von der Schule abgegangen ohne Schulabschluss“ weisen jeweils nur geringe Rückläufe auf und werden daher zusammen ausgewiesen. Alle anderen Abschlussvarianten („Haupt- oder Volksschulabschluss“, „Mittlere Reife oder Abschluss der polytechnischen Oberschule“, „Abitur, Fachhochschulreife Gymnasium oder erweiterte Oberschule EOS“ und „Hochschulabschluss“) können wie erfragt separat ausgewiesen werden.

b) Befragung der Unternehmen

ostwest Regionalinformation

Die fünfstellige Gemeindekennziffer kann nicht herausgegeben werden, da zu diversen Kreisen bzw. kreisfreien Städten nur sehr wenige Interviews vorliegen. Gleiches gilt auch für die nächstgröbere Information über das Bundesland. Erst die Region (neue oder alte Bundesländer) bietet demgegenüber ausreichend Schutz für die Teilnehmerinnen und Teilnehmer und ermöglicht gleichzeitig eine zumindest grobe regionale Zuordnung der Befragten. Als tiefste regionale Ebene wird daher die Angabe zum Bundesgebiet bereitgestellt.

F1 Rechtsform

Folgende Rechtsformen werden aufgrund weniger Nennungen zu einer Kategorie „Mischformen, andere Rechtsformen“ zusammengefasst:

- Mischformen wie z. B. GmbH & Co. KG
- Genossenschaft, Versicherungsverein auf Gegenseitigkeit
- Sonstiges

Die anderen Kategorien „Personengesellschaften“, „Kapitalgesellschaften“ und „Einzelunternehmen, Freie Berufe“ bleiben unverändert.

F2 Beschäftigtengrößenklasse

Auch bei diesem Merkmal werden Kategorien zusammengefasst, um die Identifikation einzelner Unternehmen durch eine erhöhte Fallzahl je Kategorie zu erschweren:

Ursprüngliche Kategorisierung	Kategorisierung im SUF
0 Beschäftigte	0 Beschäftigte
1 bis 9 Beschäftigte	1 bis 9 Beschäftigte
10 bis 19 Beschäftigte	10 bis 19 Beschäftigte
20 bis 49 Beschäftigte	20 bis 49 Beschäftigte
50 bis 249 Beschäftigte	50 oder mehr Beschäftigte
250 oder mehr Beschäftigte	

F55 Umsatzgrößenklasse

Analog zur Beschäftigtengrößenklasse wird bei der Umsatzklassifizierung vorgegangen, um die Rückverfolgung auch hier zu erschweren:

Ursprüngliche Kategorisierung	Kategorisierung im SUF
Unter 17 500 Euro	bis unter 500 000 Euro
von 17 500 Euro bis unter 500 000 Euro	
von 500 000 bis unter 10 Mio. Euro	von 500 000 bis unter 10 Mio. Euro
10 Mio. bis unter 40 Mio. Euro	10 Mio. Euro und höher
40 Mio. Euro und höher	

F56 & F57: aggregierter Sektor

Einige Branchen sind in der Stichprobe selten. Selbst unter Aggregation auf Sektorenebene bleibt das Problem bestehen, da der Sektor „Land- oder Forstwirtschaft, Fischerei“ in der Stichprobe nur aus einer Branche besteht und selten vorkommt. Daher werden die Sektoren „Land- oder Forstwirtschaft, Fischerei“ und „Produzierendes Gewerbe“ gemeinsam ausgewiesen.

Ursprüngliche Kategorisierung	Kategorisierung im SUF
Baugewerbe	Produzierendes Gewerbe, oder Land-/ Forstwirtschaft, Fischerei
Land- oder Forstwirtschaft, Fischerei	
Verarbeitendes Gewerbe	
Energieversorgung	
Wasserversorgung, Abwasser- und Abfallentsorgung	
Bergbau	
Handel, Kfz-Reparatur	Dienstleistungen
Freiberufliche, wissenschaftliche oder technische Dienstleistungen	
Grundstücks- und Wohnungswesen	
Gastgewerbe	
Gesundheits- und Sozialwesen	
Information und Kommunikation	
Verkehr und Lagerei	
Kunst, Unterhaltung, Erholung	
Erziehung, Unterricht	
Finanzen, Versicherungen	
Sonstige wirtschaftliche Dienstleistungen	
Sonstige Dienstleistungen	

Durch die kombinierten Maßnahmen verkleinert sich die Anzahl an Kombinationsmöglichkeiten der Variablenausprägungen sehr deutlich und alle sind ausreichend mit Fallzahlen hinterlegt.

3.6 Unverändert bereitgestellte Variablen

Geschlossene Fragen zu Kontakten mit der öffentlichen Verwaltung

Die Häufigkeit der Behördenkontakte, der spezifischen Kontaktwege mit den Behörden, der genutzten Informationsquellen sowie die Zufriedenheit der Befragten mit verschiedenen Aspekten der Erbringung der behördlichen Dienstleistungen sind zentrale Inhalte der Erhebung. Zudem verteilen sich die Antworten auf diese Fragen relativ breit. Eine Antwort auf die Detailfragen liegt auch nicht bloß einmal pro Befragtem vor, sondern für jeden spezifischen Behördenkontakt, sodass die Fallzahlen deutlich höher sind. Da diese Informationen nur mit einem sehr hohen Informationsverlust zu anonymisieren wären und aus der Bereitstellung der unveränderten Informationen gleichzeitig kein erhöhtes Aufdeckungsrisiko resultiert, werden diese Variablen in unveränderter Form bereitgestellt. Gleiches gilt für die Informationen zum Teilnahmestatus und zu den Screen-Out-Gründen und die für Analysen nötige Designgewichtung.

Offene Fragen

Es gibt jeweils vier Freitext-Fragen in den beiden Datensätzen:

- etwaige *andere Anliegen* bei der Behörde als die geschlossen abgefragten konnten von den Befragten benannt werden
- Personen die alles in allem nicht mindestens zufrieden mit ihrem Behördenkontakt waren, wurden nach ihren konkreten *Problemen* gefragt
- Personen die ihren Behördenkontakt kompliziert faden, wurden um eine Begründung gebeten
- Falls Personen keine ihrer benötigten Formulare und Nachweise online ausgefüllt haben, wurden nach den Gründen hierfür befragt
- zu jeder abgefragten behördlichen Dienstleistung konnten die Befragten *Verbesserungsvorschläge* abgeben

Diese Freitextangaben wurden bereits vom für die Erhebung verantwortlichen Institut derart anonymisiert, dass Namen und Orte mit „XXX“ ersetzt und damit unkenntlich gemacht wurden.

Eine Person oder ein Unternehmen lässt sich anhand der Freitextangaben alleine oder auch in Kombination mit anderen Merkmalen kaum identifizieren, da die Angaben kurzgehalten und eher allgemein sind.

Mit den offenen Fragen sind gleichzeitig aufschlussreiche inhaltliche Auswertungen möglich. Daher sollen sie Forscherinnen und Forschern für Analysen zur Verfügung gestellt werden. Veröffentlicht werden dürfen die Ergebnisse letztlich nur in aggregierter Form, z. B. nach Anwendung und Auszählung von Codier-Systematiken oder über die Formulierung des Tenors in eigenen Worten.

Weitere sozioökonomische Variablen

a) beide Befragungen

Die Gemeindegrößenklasse kann unverändert bereitgestellt werden.

a) Bürgerinnen und Bürger

Das diverse Geschlecht kommt tendenziell jeweils nur selten in der Erhebung vor. Allerdings ist dieses in der Grundgesamtheit häufiger anzutreffen oder dort nicht immer nachvollziehbar, sodass eine eindeutige Identifikation nicht möglich erscheint und bezüglich dieser Ausprägung keine weiteren Anonymisierungsmaßnahmen ergriffen werden.

Ebenfalls unverändert bereit gestellt werden können die sozioökonomischen Merkmale Erwerbstatus, Nettohaushaltseinkommen (kategorial), Behinderung (kategorial), Partner im Haushalt (kategorial), Migrationshintergrund (kategorial) und primär genutztes Gerät zur Internetnutzung (kategorial). Hier liegen für jede Kategorie jeweils genügend Rückläufe vor.

b) Unternehmen

Das Unternehmensalter (kategorial) wird unverändert bereitgestellt, da in der aktuellen Welle für jede Ausprägung der Variablen hinreichend Rückläufe vorliegen.

4. Fazit

Durch die beschriebenen Anonymisierungsmaßnahmen wird die faktische Anonymität der Daten sichergestellt. Daher können die Daten in dieser Form für Analysen zu Forschungszwecken bereitgestellt werden.

Falls darüberhinausgehende Forschungsfragen bestehen, welche sich nur mit den Originaldaten beantworten lassen, so kann die Erstellung entsprechender Sonderauswertungen durch die Gruppe I2 unter der funktionalen E-Mailadresse erfuellungsaufwand@destatis.de angefragt werden.

Literatur

Bentley, Jon/Floyd, Bob, *A Sample of Brilliance*. In: Communications of the Association for Computing Machinery, Ausgabe 30/1987, S. 754ff. Zugriff am 09. April 2021]. Verfügbar unter: https://dl.acm.org/doi/abs/10.1145/30401.315746?download=true.

El Emam, Khaled/Jonker, Elizabeth/Arbuckle, Luk/Malin, Bradley. *A Systematic Review of Re-Identification Attacks on Health Data*. In: PLoS ONE, Ausgabe 6/2012. [Zugriff am 06. August 2020]. Verfügbar unter: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0028071.

Glemser, Axel/Meier, Gerd/Heckel, Christiane. Dual-Frame: *Stichprobendesign für CATI-Befragungen im mobilen Zeitalter*. In: ADM Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute e. V. (Herausgeber). Stichproben-Verfahren in der Umfrageforschung: Eine Darstellung für die Praxis. 2. Auflage. Wiesbaden 2014, Seite 167 ff.

Himmelsbach, Elke/Hornbach, Carsten/Michalik, Susanne/Kühnhenrich, Daniel. *Methodische Grundlagen der Zufriedenheitsbefragungen zu behördlichen Dienstleistungen*. In: Wirtschaft und Statistik, Ausgabe 4/2016. Seite 54 ff.

Höhne, Jörg. *Verfahren zur Anonymisierung von Einzeldaten*. Statistik und Wissenschaft, Bd. 16. Wiesbaden 2010.

Kühnhenrich, Daniel/Michalik, Susanne. *Verwaltungssprache, schwere Sprache? – Ergebnisse zur Verständlichkeit von behördlichen Formularen und Schreiben aus der Lebenslagenbefragung 2019*. In: Fisch, Rudolf (Herausgeber). Verständliche Verwaltungskommunikation in Zeiten der Digitalisierung. Konzepte – Lösungen – Fallbeispiele. Baden-Baden 2020, Seite 47ff.

Lubarsky, Boris. *Re-Identification of „Anonymized Data“*. In: Geo. L. Tech. Rev., Ausgabe 202/2017. [Zugriff am 06. August 2020]. Verfügbar unter:

https://georgetownlawtechreview.org/wp-content/uploads/2017/04/Lubarsky-1-GEO.-L.-TECH.-REV.-202.pdf.

Müller, Walter/Bien, Uwe/Knoche,Peter/Wirth, Heike u.a.. *Die faktische Anonymität von Mikrodaten*. Schriftenreihe Forum Bundesstatistik, Bd. 19. Wiesbaden 1991.

Rocher, Luc/Henrickx, Julien M./de Montjoye, Yves-Alexandre. *Estimating the success of re-identifications in incomplete datasets using generative models*. In: Nature communications, 10/2019. [Zugriff am 06. August 2020]. Verfügbar unter: https://www.nature.com/articles/s41467-019-10933-3.

Ronning, Gerd/Gnoss, Roland. *Anonymisierung wirtschaftsstatistischer Einzeldaten*. Beiträge zum Workshop am 20./21. März 2003 in Tübingen. Schriftenreihe Forum Bundesstatistik, Bd. 42. Wiesbaden 2003.

Ronning, Gerd/Sturm, Roland/Höhne, Jörg/Lenz, Rainer/Rosemann, Martin/Scheffler, Michael/Vorgrimler, Daniel u.a.. *Handbuch zur Anonymisierung wirtschaftsstatistischer Mikrodaten*. Statistik und Wissenschaft, Bd. 4. Wiesbaden 2005.

Schmidt, Bernd/Kuehnhenrich, Daniel/Zipse, Christian/Vorgrimler, Daniel. *Entlastungen spürbarer machen – Wie wird der Kontakt zur Verwaltung wahrgenommen?* In: Wirtschaft und Statistik, Ausgabe 2/2015.

Statistisches Bundesamt. *Zufriedenheit der Bürgerinnen und Bürger mit behördlichen Dienstleistungen: Ausgewählte Ergebnisse der Lebenslagenbefragung 2019*. 2019a. [Zugriff am 31. Juli 2020]. Verfügbar unter: https://www.amtlich-einfach.de/SharedDocs/Downloads/Ergebnisse_Buerger_2019.pdf?__blob=publicationFile&v=2.

Statistisches Bundesamt. *Zufriedenheit der Unternehmen mit behördlichen Dienstleistungen: Ausgewählte Ergebnisse der Lebenslagenbefragung 2019*. 2019b. [Zugriff am 31. Juli 2020]. Verfügbar unter: https://www.amtlich-einfach.de/SharedDocs/Downloads/Ergebnisse_Wirtschaft_2019.pdf?__blob=publicationFile&v=4.

Statistisches Bundesamt. *Datentabellen der Lebenslagenbefragung 2019. Zufriedenheit der Bürgerinnen und Bürger mit behördlichen Dienstleistungen*. 2019c. [Zugriff am 31. Juli 2020]. Verfügbar unter: https://www.amtlich-einfach.de/SharedDocs/Downloads/Datentabellen_Buerger_2019.xlsx?__blob=publicationFile&v=1.

Statistisches Bundesamt. *Datentabellen der Lebenslagenbefragung 2019. Zufriedenheit der Bürgerinnen und Bürger mit behördlichen Dienstleistungen*. 2019d. [Zugriff am 31. Juli 2020]. Verfügbar unter: https://www.amtlich-einfach.de/SharedDocs/Downloads/Datentabellen_Wirtschaft_2019.xlsx?__blob=publicationFile&v=2

Walprecht, Sylvana/ Schulze, Claudia/ Kühnhenrich, Daniel. *Nutzerorientierte Weiterentwicklung der Lebenslagenbefragungen von 2015 bis 2019*. In: Wirtschaft und Statistik, Ausgabe 5/2020.