Anonymisierung von Einzeldaten aus dem Datenbestand der Lohnsteuer-Statistik 2020

Bernhard Meindl

14. Dezember 2021

Inhaltsverzeichnis

1 Einleitung

Ein Ziel der Bundesanstalt STATISTIK AUSTRIA ist es, ausgewählte Mikrodatensätze der amtlichen Statistik für die wissenschaftlichen Forschung und Lehre aufzubereiten und in Form von Standardisierten Datensätzen (SDS) über die Webseite der Statistik Austria bereitzustellen. Als SDS werden Einzeldatensätze bezeichnet, die in Bezug auf die Wahrung des Datenschutzes der Respondenten durch Anwendung statistischer Anonymisierungsverfahren angepasst wurden. Durch gezielte Reduktion des Informationsgehalts des Datensatzes wird erreicht, dass das Risiko einer erfolgreichen Identifikation einer konkreten statistischen Einheit sehr gering ausfällt. Weiters müssen Datennutzer Nutzungsbestimmungen akzeptieren, in denen etwa festgehalten wird, dass keine De-Anonymisierungsversuche der Daten durchgeführt werden dürfen.

In dieser Arbeit wird die Erstellung eines anonymisierten Datensatzes aus der Lohnsteuerstatistik (LST-Datensatz) 2020 beschrieben. Der anonymisierte Datensatz enthält insgesamt 70376 Beobachtungen und besteht aus 17 Variablen. Die Anonymisierung wurde in mehreren Schritten durchgeführt wird und lehnt sich methodisch stark an die von der Statistik Austria bereits durchgeführte Erstellung von standardisierten Datensätzen der Lohnsteuerstatistik 2005 bis 2019 an.

Der anonymisierte Datensatz kann sowohl als reine Text-Datei (csv-File zum einfachen Import etwa in Microsoft Excel) als auch als R-Datensatz bezogen werden.

2 Besonderheiten des Datensatzes

2.1 Originaldaten

Der SDS der Lohnsteuerstatistik 2020 basiert auf dem authentischen Datenbestand der Lohnsteuerstatistik 2020. Dieser Datenbestand enthält Information über Alter, Geschlecht oder die soziale Stellung sowie über Bezüge und Steuerabgaben von etwas mehr als 6 Millionen steuerpflichtiger Arbeitnehmer und Arbeitnehmerinnen sowie Pensionsbezieher und Pensionsbezieherinnen. Aus den umfangreichen Informationen des Datenbestandes wurden schlussendlich jene Variablen ausgewählt, die - zusammen mit den durchgeführten Umkodierungen - in Anhang (A) beschrieben sind.

2.2 Modifikation der Originaldaten

Es wird nun kurz beschrieben, wie bestehende Variablen aus dem authentischen Datenbestand bei der Erstellung des SDS modifiziert wurden, um den bestehenden Geheimhaltungsbestimmungen Rechnung zu tragen.

Die Variable *gebjahr* enthält das Geburtsjahr einer Person. Die Angabe des Geburtsjahres könnte unter Umständen - zusammen in Verbindung mit zusätzlicher Information - die Reidentifikation einer Person ermöglichen. Deshalb wurde die Variable in eine neue Variable *alter* - bestehend aus 8 Altersgruppen - umkodiert. Aus Anhang (A) ist die Definition der Altersgruppen ersichtlich.

Die im authentischen Datenbestand vorhandene Gliederung der Wirtschaftsklassifikation ÖNACE auf 5-Steller Ebene ist für den anonymisierten Datensatz zu detailliert. Angreifer könnten durch Kombination dieser Information mit anderen (kategoriellen) Variablen unter Umständen einzelne Personen korrekt identifizieren. Aus diesem Grund ist es notwendig, die ursprüngliche Gliederung der Wirtschaftsklassifikation zu vergröbern und in eine neue Variable *oenace08_1steller* umzukodieren. In dieser Variable wurden verschiedene ÖNACE 2-Steller zu 1-Stellern aggregiert. Die genaue Gliederung der neu erstellten Variable geht aus Anhang (A) hervor. Zu erwähnen ist, dass im anonymisierten Datensatz für 2020 die ÖNACE im Unterschied zu den standardisierten Datensätzen für die Jahre 2005-2007 nicht in der Version von 2003 sondern in der aktuellen Version von 2008 enthalten ist.

3 Geheimhaltung

Im folgenden wird die Anonymisierungsprozedur zur Erzeugung eines SDS aus dem authentischen Datenbestand der Lohnsteuerstatistik 2020 beschrieben.

3.1 Software

Nach dem Vorbereiten des Rohdatensatzes mit SAS wurde die weitergehende Anonymisierungsprozedur mit der freien Statistiksoftware *R* (R Development Core Team, 2006) sowie dem von Statistik Austria entwickelten und frei verfügbaren `R`-Package *sdcMicro* (Templ, 2007) (**s**tatistical **d**isclosure **c**ontrol for **micro**data) durchgeführt. Das Package kann von den Servern des R Comprehensive Archive Network (CRAN) heruntergeladen werden.

3.2 Direkte Identifikationsvariablen

Direkte Identifikationsvariablen ermöglichen es einem Datenangreifer, bestimmte Personen in einem Datensatz eindeutig zu identifizieren. Solche Variablen müssen daher aus einem Standardisierten Datensatz entfernt werden um den Geheimhaltungsanforderungen gerecht werden zu können. Ein Beispiel für eine direkte Identifikationsvariable könnte etwa die Sozialversicherungsnummer sein, die von einem Angreifer dazu genutzt werden könnte, eine Person im Standardisierten Datensatz eindeutig zu identifizieren.

In den Datenbeständen der Lohn- und Einkommensteuerstatistiken ist als eindeutige Kennung für eine Person (= Personenschlüssel) nur das "bereichsspezifische Personenkennzeichen - Amtliche Statistik" (*bPK AS*) enthalten, das aber insofern keine "richtige" Identifikationsvariable darstellt, als daraus ein Rückschluss auf die zugehörige Person nicht möglich ist. Das *bPK AS* wurde nicht in den SDS aufgenommen.

3.3 Indirekte Identifikationsvariablen

Kann durch Kombination mehrerer (meist kategorieller) Variablen eine Person eindeutig im Datensatz identifiziert werden, so werden die diese Variablen als indirekte Identifikationsvariablen bezeichnet. In diesem Zusammenhang ist es jedoch wichtig festzustellen, dass keine der indirekten Identifikationsvariablen für sich selbst zur eindeutigen Identifizierung einer Person im Datensatz ausreicht.

Indirekte Identifikationsvariablen in den Lohnsteuerdaten sind etwa das Geburtsdatum (*vgebdat*), die soziale Stellung der Person (*sozst*), das Bundesland (*bl*) oder Information über die Anzahl der Lohnzettel (*zlz*) einer Person. Kategorielle Variablen können vergröbert oder umkodiert werden um das Risiko einer Reidentifikation einer Person gering zu halten. Letztlich kann es sein, dass in den indirekten Identifikationsvariablen wenige Werte unterdrückt bzw. gelöscht werden müssen um weitestgehende Anonymität gewährleisten zu können. Die an den im anonymisierten Datensatz vorhandenen Variablen durchgeführten Umkodierungen und Vergröberungen sind in Anhang (A) aufgelistet.

3.4 Die Stichprobe

Der erste Schritt bei der Erstellung eines Standardisierten Datensatzes für die Lohnsteuerstatistik 2020 besteht - wie schon bei der Vorgehensweise für den SDS von 2005-2019 - darin, eine Stichprobe aus dem vollständigen, authentischen Datensatz des Jahres 2020 zu ziehen. Hinsichtlich der Geheimhaltung bietet eine Stichprobe den Vorteil, dass nicht alle Objekte der Grundgesamtheit im veröffentlichten Mikrodatensatz enthalten sind. Daher kann sich ein Angreifer selbst bei einer vermeintlichen Identifikation einer Person nicht sicher sein, dass die identifizierte Person die ist, an der er interessiert ist. Durch die Auswahl einer Teilmenge für den Standardisierten Datensatz kann ein Angreifer nicht wissen, ob eine Zielperson in der Stichprobe enthalten ist.

Analog zum Stichprobendesign der standardisierten Datensätze von 2005 bis 2019 wurde eine geschichtete Zufallsstichprobe mit einheitlichem Auswahlsatz von 1% innerhalb der Schichten gezogen. Als Schichtungsvariablen wurden folgende Variablen ausgewählt:

- **bl**: 10 Ausprägungen
- **geschl**: 2 Ausprägungen
- **alter**: 8 Ausprägungen

Die gezogene Stichprobe besteht aus insgesamt 70376 Beobachtungen und enthält zu den 16 ausgewählten Variablen auch noch das resultierende Hochrechnungsgewicht.

3.5 Schlüsselvariablen für die Geheimhaltung

Indirekte Identifikationsvariablen, deren Ausprägungskombinationen ein Angreifer verwenden könnte, um eine eindeutige Identifikation einer Person im Datensatz vorzunehmen, werden als Schlüsselvariable bezeichnet. Für die LST-Daten wurden - wiederum analog zu den bereits vorhandenen SDS - folgende Schlüsselvariablen definiert.

- *bl* (10 Ausprägungen)
- *sozst* (9 Ausprägungen)
- *oenace08_1steller* (11 Ausprägungen)
- *geschl* (2 Ausprägungen)
- *alter* (8 Ausprägungen)

Zu bemerken ist, dass die Variable *sozst* im SDS von 2020 analog zur Variable in den SDS ab 2007, aber unterschiedlich zu den standardisierten Datensätzen aus dem Bereich der Lohnsteuerstatistik von 2005 und 2006 kodiert ist.

Im Zuge der Anonymisierungsprozedur werden einzelne Schlüselvariablen modifiziert indem sie entweder vergröbert oder umkodiert werden. Eine weitere Möglichkeit besteht darin, einzelne Werte in den Schlüsselvariablen zu löschen um schließlich einen sicheren SDS mit hohem Analysepotential zu erhalten.

3.6 Lokale Unterdrückung

Für jede Merkmalskombination der Schlüsselvariablen wird - nach dem Modell von Benedetti and Franconi (1998) - das individuelle Reidentifikationsrisiko berechnet. Dabei ist neben der Anzahl an Personen, die eine spezifische Ausprägungskombination der Schlüsselvariablen aufweist auch das Hochrechnungsgewicht wesentlich. Der Einfluss des Hochrechnungsgewichtes ergibt sich aus der Tatsache, dass Personen mit einem hohen Hochrechnungsgewicht grundsätzlich ein höheres Reidentifikationsrisiko aufweisen. Diese Personen müssen besonders geschützt werden.

Basierend auf der gezogenen Stichprobe, den ausgewählten (und gegebenenfalls modifizierten) Schlüsselvariablen weisen 591 Beobachtungen eine einzigartige (unique) Ausprägungskombination in den Schlüsselvariablen auf. Außerdem gibt es weitere 628 Personen, deren Ausprägungskombination der Schlüsselvariablen genau zweimal vorkommen. Im Zuge der Anonymisierungsprozedur soll durch gezielte Sperrungen in den Schlüsselvariablen erreicht werden, dass jeder Ausprägungskombination zumindest 3 Personen zugeordnet werden kann. Dies wird auch als *3-Anonymity* bezeichnet.

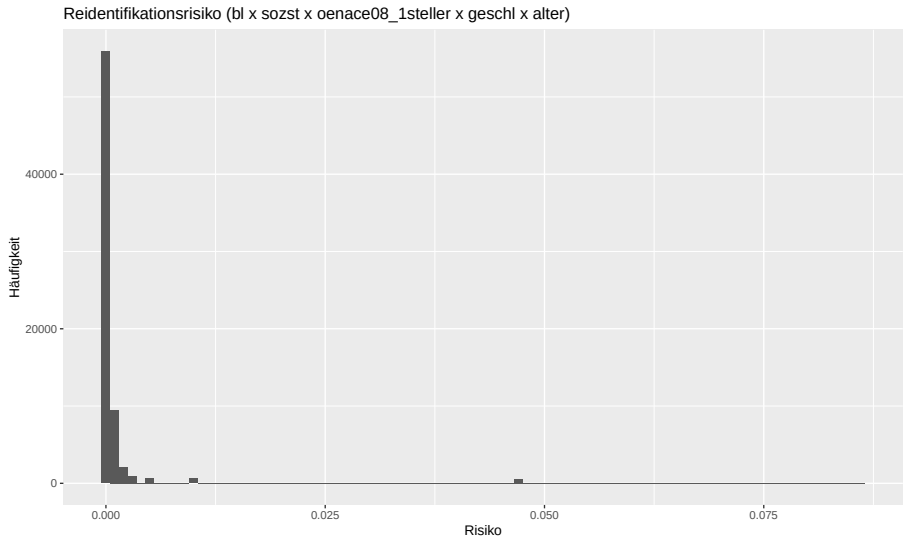

Abbildung 1: Individuelles Identifikationsrisiko in den Originaldaten.

Abbildung (1) zeigt das individuelle Reidentifikationsrisiko vor der Unterdrückung von Werten in den Schlüsselvariablen. Man erkennt, dass das Reidentifikationsrisiko für die allermeisten Beobachtungen sehr gering ist. Personen mit hohem Reidentifikationsrisiko müssen zusätzlich geschützt werden. Dies geschieht indem schrittweise Sperrungen in Schlüsselvariablen durchgeführt werden. Es ergeben sich folgende Sperrungen:

- Variable *alter*:
  1072 Beobachtungen ($\approx 1.52\%$) wurden unterdrückt.

- Variable *bl*:
  134 Beobachtungen ($\approx 0.19\%$) wurden unterdrückt.
- Variable *sozst*:
  27 Beobachtungen ($\approx 0.04\%$) wurden unterdrückt.

Durch das Sperren dieser Werte im Datensatz kann *3-Anonymity* gewährleistet werden. Das bedeutet, dass jede Ausprägungskombination der Schlüsselvariablen im Datensatz zumindest dreimal existiert.

Abbildung (2) zeigt das individuelle Reidentifikationsrisiko im Datensatz nach Durchführen der lokalen Unterdrückung in den Schlüsselvariablen. Man erkennt, dass für die im SDS vorhandenen Personen ein sehr geringes Reidentifikationsrisiko besteht. Insbesondere sei auf die unterschiedliche Skalierung der x-Achsen in Abbildung (1) und (2) hingewiesen.

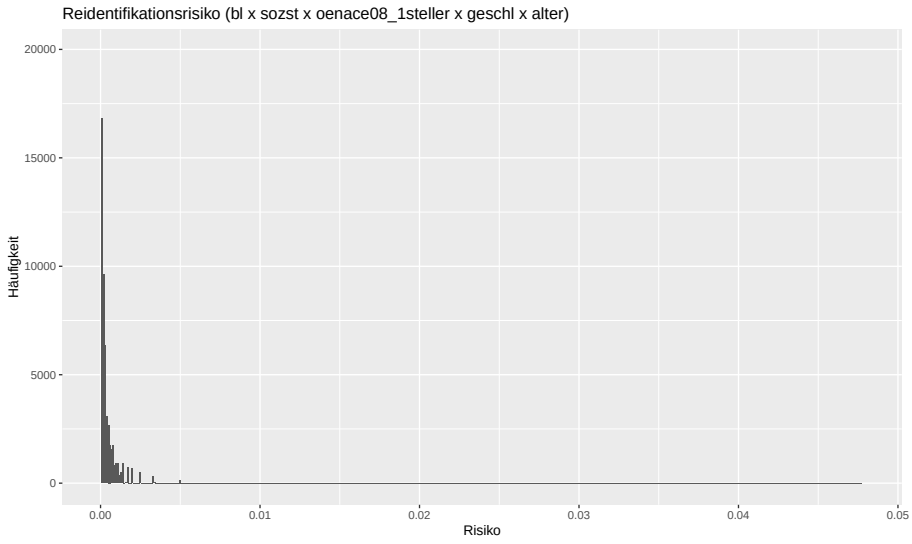

Abbildung 2: Individuelles Reidentifikationsrisiko im anonymisierten Datensatz.

3.7 Mikroaggregation

Unter Umständen kann auch bekannte Informationen über (Prozent)Werte numerischer Variablen verwendet werden, um Einheiten erfolgreich zu identifizieren. Insbesondere "Ausreißer" in numerischen Variablen können in Verbindung mit Informationen über andere Schlüsselvariablen dazu verwendet werden, eine positive Identifizierung zu erreichen. Die *Mikroaggregation* numerischer Variablen bietet zusätzlichen Schutz gegen Identifizierungsversuche. Die Idee dabei ist, dass möglichst "ähnliche" Objekte in einem ersten Schritt zu gruppieren und anschließend in einem zweiten Schritt die Ausprägungen numerischen Variablen der gewählten Personen durch eine Statistik zu ersetzen. Dadurch wird sichergestellt, dass jede einzelne Ausprägung mehrfach im Datensatz auftritt.

Aus Anhang (A) geht hervor, welche Variablen des Standardisierten Datensatzes der Lohnsteuerstatistik 2020 mikroaggregiert wurden. Bei diesen Variablen wird sichergestellt,

dass jeder Wert zumindest 4-fach pro Variable auftritt. Als Mikroaggregationsmethode wurde ein Verfahren verwendet bei dem die Gruppierung der Beobachtungen mittels multivariat berechneter Distanzen durchgeführt wurde.

4 Zusammenfassung

Die Aufbereitung und Bereitstellung sensibler Mikrodaten - wie etwa Steuerdaten - für wissenschaftliche Forschung und Lehre ist ein komplexer Prozess. Insbesondere muss Hauptaugenmerk auf die Anonymisierung der Daten gelegt werden um die gegebenen rechtlichen Anforderungen zu erfüllen.

Da nur eine 1% Stichprobe des vollständigen authentischen Datenbestands für die Veröffentlichung als Standardisierter Datensatz aufbereitet wurde, kann sich ein Datenangreifer auch bei vermeintlicher positiver Identifizierung einer Person aufgrund mehrerer Variablen nicht sicher sein, ob die identifizierte Person überhaupt für die Stichprobe ausgewählt wurde. Durch die weiters angewandten Anonymisierungsverfahren wie die Aggregation beziehungsweise das Umkodieren kategorieller Variablen, dem Ersetzen kritischer Werte in den Schlüsselvariablen durch *missings* und durch Mikroaggregation numerischer Variablen wurde erreicht, dass das Reidentifikationsrisiko aller im SDS verbleibenden Daten sehr gering ist. Allerdings ist anzumerken, dass es 100%-igen Schutz vor Reidentifizierung nicht geben kann. Ein Restrisiko bleibt immer bestehen.

Die gewählte Methodik zur Erstellung des Standardisierten Datensatzes für die Lohnsteuerstatistik 2020 ist praktisch identisch mit der Erstellung der standardisierten Datensätze aus dem Datenbestand der Lohnsteuerstatistik für 2005 bis 2019. Der Anonymisierungsprozess wurde mit der hausintern entwickelten Software *sdcMicro* durchgeführt. Beim Erstellen des SDS zur Lohnsteuerstatistik 2020 wurde immer darauf geachtet, trotz der notwendigen Anonymisierung der Daten das hohe Analysepotential der Daten zu erhalten. Der vorliegende standardisierte Datensatz wird diesem Anspruch gerecht.

Literatur

Benedetti, R. and Franconi, L. (1998). Statistical and technological solutions for controlled data dissemination. In Pre-proceedings of New Techniques and Technologies for Statistics, volume 1, pages 225–232.

R Development Core Team (2006). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Templ, M. (2007). *scdMicro*: A package for statistical disclosure control in R. In ISI 2007, Lissabon.

A Datenbeschreibung

In Tabelle (A) werden die Variablen beschrieben, die im SDS für 2020 enthalten sind. In Spalte 4 ist überblicksmäßig die für diese Variable angewandte Geheimhaltungaktion beschrieben. Als Unterstützung wurden die Variablennamen verschiedenfärbig markiert, wobei Variablen, die in schwarzer Schrift aufscheinen, nicht verändert wurden. Variablen, die mit blau gekennzeichnet sind wurden verändert oder neu erzeugt und rot bedeutet, dass diese Variable mikroaggregiert wurde.

Variablen-name	**Skalierung / Identifier**	**Beschreibung**	**Aktion**	**Kodierung/ Spezifizierung**
SOZST	kat. / indirekt	Soziale Stellung		1=Personen mit sonst. Aktivbezügen 2=Lehrling 3=Arbeiter(in) 4=Angestellte(r) 5=Beamte(r) 6=Vertragsbedienstete(r) 7=Beamte(r) i.R. 8=Pensionisten ohne Beamte i.R. 9=Personen mit nur Pflegegeldbezug NA=fehlend
GESCHL	kat. / indirekt	Geschlecht		1=männlich 2=weiblich
BL	kat. / indirekt	Bundesland		0=Ausland/unbekannt 1=Burgenland 2=Kärnten 3=Niederösterreich 4=Oberösterreich 5=Salzburg 6=Steiermark 7=Tirol 8=Vorarlberg 9=Wien NA=fehlend
OENACE	kat. / indirekt	ÖNACE-Klassifikation	vergröbert	1-Steller in Variable OENACE08_1steller
OENACE08_1steller	kat. / indirekt	ÖNACE08 1-Steller	erstellt aus Variable OENACE	0=unbekannt 1=Abschnitte A,B 2=Abschnitt C 3=Abschnitte D,E 4=Abschnitt F 5=Abschnitt G 6=Abschnitt I 7=Abschnitte H,J 8=Abschnitt K 9=Abschnitte L,M,N 10=Abschnitte O,P,Q 11=Abschnitte R,S,T,U NA=fehlend
VTBESCH	kat. / nein	Beschäftigung: Vollzeit / Teilzeit		1=(überwiegend) Vollzeit 2=(überwiegend) Teilzeit 3=nicht zutreffend 4=(überwiegend) unbekannt
GEBJAHR	kat. / indirekt	Geburtsjahr	vergröbert	Altersklassen in Variable ALTER
ALTER	kat. / indirekt	Altersklassen	erstellt aus Variable GEBJAHR	1=15 Jahre und jünger 2=16-25 Jahre 3=26-35 Jahre 4=36-45 Jahre 5=46-55 Jahre 6=56-60 Jahre 7=61-65 Jahre 8=66 Jahre und älter NA=fehlend

ZLZ	num. / nein	Anzahl der Lohnzettel		1-8=tatsächliche Anzahl 9=9 und mehr
BEZD	num. / nein	Bezugsdauer in Tagen	vergröbert	Bezugsdauer in Wochen in Variable BEZW
BEZW	kat. / nein	Bezugsdauer in Wochen	erstellt aus Variable BEZD	1=1-7 Tage 2=8-14 Tage ... 52=358-365 Tage
KZ210	num. / nein	Bruttobezüge <210>	mikroaggregiert	
KZ220	num. / nein	Sonstige Bezüge gem. §67 (1,2)	mikroaggregiert	
KZ230	num. / nein	Sozialvers., Kammeruml., Wohnbauförderung	mikroaggregiert	
FESTSAT	num. / nein	Steuerfreie bzw. mit festen Sätzen versteuerte Bezüge gem §67, Abs. 3-8	mikroaggregiert	
KZ245	num. / nein	Steuerpflichtige Bezüge	mikroaggregiert	
EINBLST	num. / nein	Insgesamt eingehaltene Lohnsteuer	mikroaggregiert	
LFESTSA	num. / nein	Lohnsteuer mit festen Sätzen	mikroaggregiert	
NTSONST	num. / nein	Nach Tarif versteuerte sonst. Bezüge §67 (2,6,10)	mikroaggregiert	
SAMPLINGWEIGHT	num. / nein	Stichprobengewicht		

Tabelle 1: Beschreibung der Variablen aus dem Standardisierten Datensatz der Lohnsteuerstatistik 2020