Anonymisierung von Einzeldaten aus dem Datenbestand der integrierten Lohn- und Einkommenssteuerstatistik 2010

Bernhard Meindl

14. August 2013

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 Standardisierte Datensätze werden in diesem Rahmen Einzeldatensätze bezeichnet, die insbesondere in Bezug auf die Wahrung des Datenschutzes angepasst wurden. Die Wahrung der Interessen der Respondenten wird dabei durch die Anwendung statistischer Anonymisierungsverfahren gewährleistet. Durch gezielte Reduktion des Informationsgehalts des entsprechenden 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 durchgeführt werden dürfen.

In dieser Arbeit wird die Erstellung eines anonymisierten Datensatzes aus der integrierten Lohn- und Einkommenssteuerstatistik 2010 beschrieben. Der anonymisierte Datensatz enthält insgesamt 66754 Beobachtungen und besteht aus 14 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 2006-2009 sowie der Erstellung eines SDS aus dem Bestand der Einkommenssteuer für 2005 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 integrierten Lohn- und Einkommenssteuer 2010 basiert auf dem zugrundeliegenden authentischen Datenbestand. Aus den umfangreichen Informationen des authentischen Datenbestandes wurden schließlich insgesamt 14 Variablen ausgewählt, die - zusammen mit den durchgeführten Umkodierungen - in Anhang (A) beschrieben sind.

2.2 Modifikation der Originaldaten

Es wird nun kurz beschrieben, auf welche Art und Weise bestehende Variablen aus dem authentischen Datenbestand für den SDS modifiziert wurden, um der Geheimhaltung Rechnung zu tragen.

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

Die im authentischen Datenbestand existierende tiefe Gliederung der Wirtschaftsklassifikation ÖNACE (2003) ist für den anonymisierten Datensatz zu detailliert. Angreifer könnten durch Kombination dieser Information mit anderen (kategoriellen) Variablen unter Umständen einzelne Einheiten korrekt identifizieren. Aus diesem Grund war es notwendig, die ursprüngliche Gliederung der Wirtschaftsklassifikation zu vergröbern und in eine neue Variable *OENACE* umzukodieren. In dieser Variable wurden verschiedene ÖNACE 2-Steller zu 1-Stellern aggregiert. Die exakte Gliederung der neu erstellten Variable *OENACE* geht aus Anhang (A) hervor.

3 Geheimhaltung

Es wird nun die Anonymisierungsprozedur beschrieben, die durchgeführt wurde um aus dem authentischen Datenbestand der integrierten Lohn- und Einkommenssteuerstatistik 2010 einen SDS-File zu erzeugen.

3.1 Software

Nach dem Vorbereiten des Rohdatensatzes mit SAS wurde die weitergehende Anonymisierungsprozedur mit der freien Statistiksoftware *R* (**?**) sowie dem von Statistik Austria entwickelten und frei verfügbaren `R`-Package *sdcMicro* (**?**) (**s**tatistical **d**isclosure **c**ontrol for **micro**data) durchgeführt. Das Package kann von den Servern des R Comprehensive Archive Network (CRAN) heruntergeladen werden. *sdcMicro* weist wesentliche Vorteile gegenüber der für Geheimhaltung von Mikrodaten empfohlenen "Standardsoftware" μ-Argus auf. Außerdem wird *sdcMicro* ständig aktualisiert, verbessert und weiterentwickelt.

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. Als Beispiel für eine direkte Identifikationsvariable könnte etwa die Sozialversicherungsnummer genannt werden, die von einem Angreifer dazu genutzt werden könnte, eine Person im Standardisierten Datensatz eindeutig zu identifizieren.

Im Datenbestand der Lohnsteuerstatistik existiert für jede Person eine weitere mögliche direkte Identifikationsvariable, die Subjektidentifikationsnummer (*SID*). Da diese Variable eventuell beim Verknüpfen der Mikrodaten mit weiteren Quellen als direkte Identifikationsvariable verwendet werden könnte, wurde diese Variable gelöscht.

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 ausreichen muss.

Indirekte Identifikationsvariablen in den Variablen der integrierten Lohn- und Einkommenssteuer, die für die Erstellung des SDS vorgesehen sind, sind etwa das Geburtsdatum (*VGEBDAT*), das Bundesland (*BLD*) oder Information über den Schwerpunkt der Tätigkeit (*SP8*) 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 integrierte Lohn- und Einkommenssteuerstatisik 2010 besteht darin, eine Stichprobe aus dem vollständigen, authentischen Datensatz 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.

Für die Erstellung eines SDS aus der integrierten Lohn- und Einkommenssteuerstatistik des Jahres 2010 wurde eine geschichtete Zufallsstichprobe mit einheitlichem Auswahlsatz von 1% innerhalb der Schichten gezogen. Als Schichtungsvariablen wurden folgende Variablen ausgewählt:

- **BLD**: 10 Ausprägungen
- **GESCHL**: 2 Ausprägungen
- **ALTER**: 9 Ausprägungen

Die gezogene Stichprobe besteht aus insgesamt 66754 Beobachtungen und enthält zu den 14 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 Schlüsselvariablen bezeichnet. Für die diesen Datenbestand wurden folgende Schlüsselvariablen definiert.

- **BLD** (10 Ausprägungen)
- **ALTER** (9 Ausprägungen)
- **OENACE** (13 Ausprägungen)
- **GESCHL** (2 Ausprägungen)
- **SP8** (7 Ausprägungen)

Zu bemerken ist, dass die Variable *SP8* für den Vorgang der Anonymisierung mit nur 3 Ausprägungen verwendet wurde. Im finalen SDS ist diese Variable jedoch mit 7 Ausprägungen (siehe Anhang A) enthalten.

Im Zuge der Anonymisierungsprozedur werden einzelne Schlüsselvariablen 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.

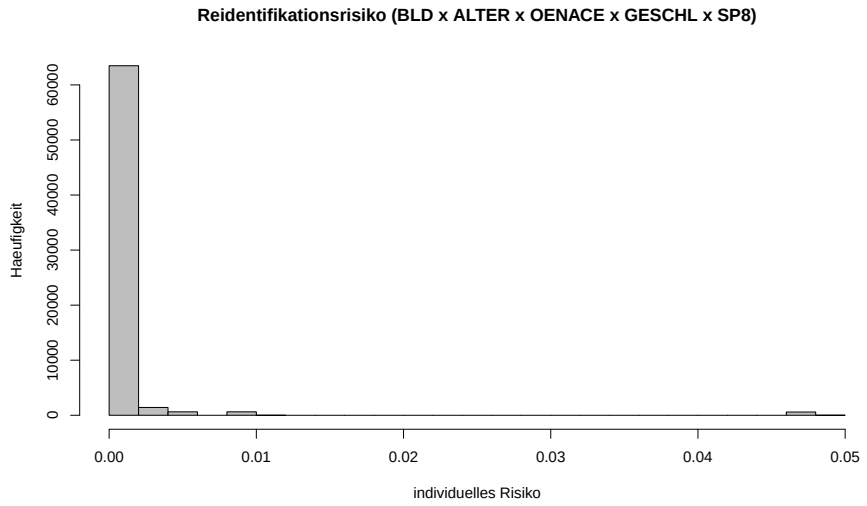

Abbildung 1: Individuelles Identifikationsrisiko in den Originaldaten.

3.6 Lokale Unterdrückung

Für jede Merkmalskombination der Schlüsselvariablen wird - nach dem Modell von **?**) - 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 und den ausgewählten (und gegebenenfalls modifizierten) Schlüsselvariablen weisen 598 Beobachtungen eine einzigartige (unique) Ausprägungskombination in den Schlüsselvariablen auf. Außerdem gibt es weitere 630 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) 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 **SP8**:
  598 Beobachtungen ($\approx 0.9\%$) wurden bei einem Grenzwert für das individuelle Risiko von 4% unterdrückt.

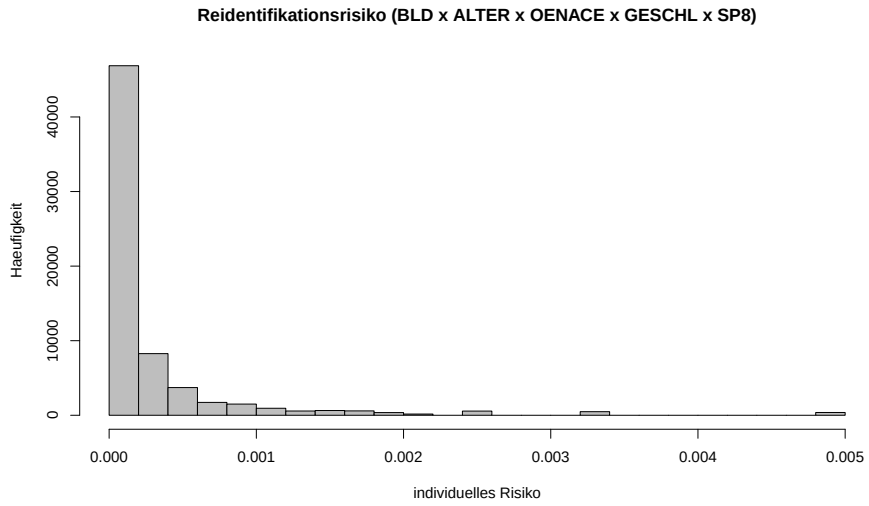

Abbildung 2: Individuelles Reidentifikationsrisiko im anonymisierten Datensatz.

- Variable **OENACE**:
  175 Beobachtungen ($\approx 0.26\%$) wurden bei einem Grenzwert für das individuelle Risiko von 1% unterdrückt.
- Variable **ALTER**:
  306 Beobachtungen ($\approx 0.46\%$) wurden bei einem Grenzwert für das individuelle Risiko von 0.5% 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.

3.7 Mikroaggregation

Unter Umständen besteht die Möglichkeit dass ein Datenangreifer ihm bekannte Informationen über einen (Prozent)Wert einer numerischen Variable heranzieht, um eine Person im Datensatz erfolgreich zu identifizieren. Insbesondere "Ausreißer" in numerischen Variablen können in Verbindung mit Informationen über andere Schlüsselvariablen dazu verwendet werden, eine positive Identifizierung zu erreichen.

Mikroaggregation numerischer Variablen bietet zusätzlichen Schutz gegen Reidentifizierungsversuche. Mikroaggregation bedeutet grundsätzlich, dass möglichst "ähnliche" Objekte in einem ersten Schritt gruppiert werden. In einem zweiten Schritt werden schließlich

die Ausprägungen einer numerischen Variablen der gewählten Personen durch eine Statistik ersetzt. Bei der verwendeten Statistik handelt es sich oftmals um den Mittelwert. Durch die Mikroaggregation numerischer Variable wird sichergestellt, dass jede einzelne Ausprägung mehrfach im Datensatz auftritt.

Aus Anhang (A) geht hervor, welche Variablen des Standardisierten Datensatzes der integrierten Lohn- und Einkommenssteuerstatistik 2010 mikroaggregiert wurden. Bei diesen Variablen wird sichergestellt, dass jeder Wert einer mikroaggregierten Variablen zumindest 4-fach in dieser Variable auftritt. Als Mikroaggregationsmethode wurde *individual ranking* ausgewählt, da dieses Verfahren auch für Variablen, die fehlende Werte enthalten, verwendet werden kann. Alle Variablen, die mikroaggregiert werden, werden zuerst unabhängig voneinander sortiert. In einem zweiten Schritt findet die Mikroaggregation selbst statt. Den Abschluss des Verfahrens bildet das Zurücksortieren der Ausprägungen nach der ursprünglichen Ordnung.

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 sehr geringes Restrisiko bleibt bestehen.

Die gewählte Methodik zur Erstellung des Standardisierten Datensatzes für die integrierte Lohn- und Einkommenssteuerstatistik 2010 gleicht in vielen Bereichen der Erstellung standardisierter Datensätze aus dem Datenbestand der Lohnsteuerstatistik bzw. der Einkommenssteuerstatistik. Der Anonymisierungsprozess wurde mit der hausintern entwickelten Software *sdcMicro* durchgeführt. Beim Erstellen des SDS wurde immer darauf geachtet, trotz der notwendigen Anonymisierung der Daten das hohe Analysepotential der Daten bestmöglich 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 2010 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**
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 1-Steller	erstellt aus Oenace 2-Stellern	0=unbekannt 1=2-Steller < 15 2=2-Steller >=15 und <40 3=2-Steller >=40 und <45 4=2-Steller >=45 und <50 5=2-Steller >=50 und <55 6=2-Steller >=55 und <60 7=2-Steller >=60 und <65 8=2-Steller >=65 und <70 9=2-Steller >=70 und <75 10=2-Steller >=75 und <90 11=2-Steller >=90 NA=fehlend
SP8	kat. / indirekt	Schwerpunkt der Beschäftigung		1=Arbeitnehmer ausschließlich 2=Arbeitnehmer schwerpunktmäßig. 3=Arbeitnehmer nicht schwerpunktmäßig 4=Pensionisten ausschließlich 5=Pensionisten schwerpunktmäßig. 6=Pensionisten nicht schwerpunktmäßig 7=Bezieher von übrigen Einkünften
ALTER	kat. / indirekt	Altersklassen	erstellt aus dem Geburtsdatum	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
GESEIN1	num. / nein	Gesamteinkommen mit Transfereinkünften	mikroaggregiert	
STEUGES	num. / nein	Gesamtsteuer	mikroaggregiert	
NETTO	num. / nein	Nettoeinkommen	mikroaggregiert	
KZ0210A	num. / nein	Lohneinkünfte (inkl. Pensionseinkünfte)	mikroaggregiert	

EINK	num. / nein	übrige Einkünfte	mikroaggregiert	
TRANSGES	num. / nein	Transfereinkünfte insgesamt	mikroaggregiert	
STBEMGR	num. / nein	Steuerbemessungsgrundlage	mikroaggregiert	
SAMPWEIGHT	num. / nein	Stichprobengewicht		

Tabelle 1: Beschreibung der Variablen aus dem Standardisierten Datensatz der integrierten Lohn- und Einkommenssteuerstatistik 2010