Anonymisierung von Einzeldaten aus dem Datenbestand der Lohnsteuer-Statistik 2005

Bernhard Meindl und Matthias Templ

27. August 2007

Inhaltsverzeichnis

1 Einleitung

Die Bundesanstalt STATISTIK AUSTRIA stellt ausgewählte Mikrodaten der amtlichen Statistik zur Nutzung in der wissenschaftlichen Forschung und Lehre bereit. Die vorliegende Arbeit beschreibt die Herstellung eines anonymisierten Datensatzes aus der Lohnsteuerstatistik (LST-Datensatz). Ziel war es, einen Standardisierten Datensatz (SDS) zu generieren. Standardisierte Datensätze sind Subsamples von Mikrodatensätzen[1], die Forscherinnen und Forscher nach einer Registrierung von der Website der Statistik Austria herunterladen können. Durch gezielte Reduktion des Informationsgehalts der Datensätze wird eine Identifikation eines konkreten Einzelfalls unmöglich gemacht.

[1] Mikrodaten sind die zu den jeweiligen Beobachtungseinheiten gehörenden Daten.

Bei der Veröffentlichung von anonymisisierten Mikrodaten ist besonderes Augenmerk auf die Wahrung des Datenschutzes zu legen. Das wird durch die Anwendung von Anonymisierungsverfahren und das Auferlegen von Nutzungsbestimmungen gewährleistet. Die Anonymisierung erfolgt mit Blick auf die Erhaltung eines möglichst hohen Analysepotentials der Daten. Im Falle des SDS-Datensatzes aus den österreichischen Lohnsteuerdaten konnte insbesondere die multivariate Struktur der Daten erhalten werden.

Der anonymisierte Datensatz der LST-Daten enthält insgesamt 59279 Beobachtungen und besteht aus 17 Variablen. Die Anonymisierung wurde in mehreren Schritten durchgeführt wird und lehnt sich an die von der Statistik Austria bereits durchgeführte Anonymisierung von Daten zur beruflichen Weiterbildung in Unternehmen (CVTS2) an. Auf die einzelnen Anonymisierungsschritte wird in den folgenden Kapiteln noch genau eingegangen. Der anonymisierte Datensatz wird sowohl als Text-Datei (csv-File zum einfachen Import in Microsoft Excel) als auch als R-Datensatz veröffentlicht, da das Statistikprogrammpaket R unter Studierenden, Lehrenden und Forschenden an Universitäten sehr gebräuchlich ist.

2 Besonderheiten des Datensatzes

2.1 Originaldaten

Grundlage für das Erstellen eines Standardisierten Datensatzes für die LST-Daten war der Datenbestand zur Lohnsteuerstatistik 2005. Dieser Datensatz enthält - bereits von Lohnzettelebene auf Personenebene aggregiert - Informationen wie Alter, Geschlecht oder die soziale Stellung sowie Informationen über Bezüge und Steuerabgaben über die Person selbst. Aus den 74 Variablen, die für insgesamt knapp 6 Millionen Personen vorhandenen waren, wurden schließlich für die Erstellung des SDS 17 Variablen ausgewählt. Diese Variablen sind im Anhang (A) beschrieben.

2.2 Modifikation der Originaldaten

Im nun folgenden Abschnitt wird kurz darauf eingegangen, auf welche Art und Weise bereits bestehende Variablen modifiziert wurden, um der Geheimhaltung Rechnung zu tragen.

Basierend auf der Variable *nuts3*, die die Nuts3-Gliederung Österreichs enthält, wurde eine Variable *bld* berechnet, die die 9 Bundesländer Österreichs sowie eine zusätzliche Kategorie enthält. Die zusätzliche Kategorie enthält alle Personen, die aus dem Ausland stammen oder deren Herkunft unbekannt ist.

Die Variable *gebjahr* enthält das Geburtsjahr einer Person. Die Angabe des Geburtsjahres könnte - in Verbindung mit zusätzlichem Wissen über eine Person - die Reidentifikation einer Person ermöglichen. Deshalb wurde diese Variable vergröbert und eine neue Variable *alter* berechnet, die aus insgesamt 8 Altergruppen besteht. Die genaue Kodierung der Variable Alter ist ebenfalls aus Anhang (A) ersichtlich.

Auch die in den Originaldaten vorhandene Gliederung der Wirtschaftsklassen ist für einen anonymisierten Datensatz zu genau. Angreifer könnten durch Kombination des vorhandenen ÖNACE 6-Stellers mit anderen (kategoriellen) Variablen möglicherweise einzelne Individuen korrekt reidentifizieren. Daher wurde die Variable *oenace* vergröbert und in die neue Variable *oenace_1steller* umkodiert. Diese neue Variable enthält Information über den ÖNACE 1-Steller. Welche 2-Steller zusammengefasst wurden ist wieder in Anhang (A) aufgelistet.

3 Geheimhaltung

In den nun folgenden Kapiteln werden die einzelnen Schritte beschrieben, die unternommen wurden, um aus dem Datenbestand der Lohnsteuerstatistik einen SDS-File zu erzeugen. Zuerst werden wir jedoch kurz die Software beschreiben, mit deren Hilfe die Anonymisierung eines Datensatzes flexibel und einfach durchgeführt werden kann.

3.1 Software

Nach dem Vorbereiten des Rohdatensatzes mit SAS wurde die weitere Bearbeitung mit der Statistiksoftware *R* (R Development Core Team, 2006) durchgeführt. Zur Anonymisierung der Lohnsteuerdaten selbst wurde das von Statistik Austria entwickelte und frei verfügbare R-Package *sdcMicro* (Templ, 2007) (**s**tatistical **d**isclosure **c**ontrol for **micro**data) verwendet. Das Package kann von den Servern des R Comprehensive Archive Network (CRAN)[2] heruntergeladen werden. *sdcMicro* weist zahlreise Verbesserungen gegenüber der für Geheimhaltung empfohlenen "Standardsoftware" μ-Argus auf. So ist es beispielsweise bei Verwendung von *sdcMicro* möglich, das Reidentifikationsrisiko einer Beobachtung auch für mehr als 4 sogenannte Schlüsselvariablen zu berechnen, während μ-Argus bei dieser Aufgabe scheitert.

Die Vorgehensweise der Anonymisierung ist weitestgehend explorativ. Durch die vorhandene Flexibilität von *sdcMicro* in Verbindung mit der mächtigen Syntax von *R* ist das Durchführen der Geheimhaltung fast spielerisch und einfach möglich.

3.2 Direkte Identifier

Durch direkte Identifier lässt sich eine eindeutige Identifizierung von einzelnen Personen in einem Datensatz vornehmen. Direkte Identifizierungsvariablen dürfen daher aus Gründen des Datenschutzes auf keinen Fall im anonymisierten Datensatz enthalten sein. Ein klassisches Beispiel für eine direkte Identifizierungsvariable ist die Sozialversicherungsnummer, die - falls sie im anonymisierten SDS-File vorhanden wäre - im Falle eines Angriffs dazu genutzt werden könnte, die Daten einer Person im Datensatz zu identifizieren.

Im Fall der Lohnsteuerdaten gibt es für jede Arbeitnehmerin und für jeden Arbeitnehmer noch eine weitere mögliche direkte Identifikationsvariable, die Subjektidentifikationsnummer (*sid*). Diese Variable könnte eventuell von Angreifern durch Verknüpfen des

[2] etwa unter http://cran.at.r-project.org

Datenbestandes mit anderen Quellen als direkte Identifikationsvariable verwendet werden. Aus diesem Grund wurde diese Variable gelöscht.

3.3 Indirekte Identifier

Ist durch Kombination mehrerer, meist kategorieller Variablen eine eindeutige Identifikation einer Person möglich, so bezeichnet man diese Variablen als indirekte Identifier. Es ist jedoch zu betonen, dass jede dieser Variablen für sich alleine nicht zur eindeutigen Identifizierung einer Person im Datensatz ausreichend ist.

Indirekte Identifikationsvariablen in den Lohnsteuerdaten sind etwa das Geburtsdatum (*vgebdat*), die soziale Stellung der Person (*sozst*), das Bundesland (*bl*) oder Information darüber, ob eine Person im Jahr 2005 Anspruch auf Bundespflegegeld (*bundpfg*) besaß. Um das Risko einer Reidentifkation gering zu halten werden kategorielle Variablen vergröbert, mehrere mögliche Ausprägungen zusammengefasst und letztlich wenige Werte (lokal) unterdrückt. Welche Umkodierungen an den im anonymisierten Datensatz vorhandenen Variablen durchgeführt wurden, ist im Anhang (A) beschrieben.

3.4 Die Stichprobe

Ein erster Schritt hinsichtlich Geheimhaltung des Lohnsteuerdatensatzes besteht darin, eine Stichprobe aus den vorhandenen Lohnsteuerdaten für das Jahr 2005 zu ziehen. Hinsichtlich der Geheimhaltung bietet sie den Vorteil, dass im Falle eines Deanonymisierungsversuchs die angreifende Person nicht sicher sein kann, ob die von ihm möglicherweise identifizierte Person wirklich die interessierende Person ist oder nur zufällig die gleichen Ausprägungen in manchen Schlüsselvariablen besitzt. Durch die Stichprobenziehung kann ein Angreifer nicht wissen, ob eine Zielperson in die Stichprobe gezogen wurde oder nicht.

Um das Analysepotential für weitere Auswertungen zu bewahren, wurde eine geschichtete Zufallsstichprobe mit einheitlichem Auswahlsatz von 1% innerhalb der Schichten gezogen. Als Schichtungsvariablen wurden gewählt:

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

Als Grundlage für die weiteren Anonymisierungsschritte ergab sich ein Datensatz mit insgesamt 59279 Beobachtungen, den im Anhang ersichtlichen Variablen sowie einer Variable für das Stichprobengewicht. Das Stichprobengewicht ist im konkreten Fall jedoch kaum relevant, da die Stichprobe mit einem einheitlichen Auswahlsatz von 1% gezogen wurde und das Hochrechnungsgewicht praktisch immer dem Reziprokwert des Auswahlsatzes entspricht.

3.5 Schlüsselvariablen für die Geheimhaltung

Als Schlüsselvariable werden jene - meist kategoriellen - indirekten Identifizierungsvariablen bezeichnet, die ein Angreifer durch Kombination ihrer verschiedenen Ausprägungsmöglichkeiten verwenden könnte, um eine positive Identifikation einer Person vorzunehmen. Für die LST-Daten wurden folgende Schlüsselvariablen definiert.

- *bld* (10 Ausprägungen)
- *sozst* (7 Ausprägungen)
- *oenace_1steller* (12 Ausprägungen)
- *geschl* (2 Ausprägungen)
- *alter* (8 Ausprägungen)

In einer explorativen Art und Weise, unterstützt durch das R-Package *sdcMicro*, werden einzelne Schlüsselvariable vergröbert und/oder umkodiert, um schließlich am Ende dieses Prozesses einen SDS mit ausreichender Geheimhaltung zu erhalten.

3.6 Lokale Unterdrückung

Dem Modell von Benedetti und Franconi (1998) folgend, wird für jede Ausprägungskombination der Schlüsselvariablen das Reidentifkationsrisiko berechnet. Dabei ist jedoch nicht nur die Anzahl der Personen, die eine spezifische Ausprägungskombination der Schlüsselvariablen aufweist, wesentlich. Eine wichtige Rolle spielt auch das Stichprobengewicht. Denn Personen mit einem niedrigen Stichprobengewicht besitzen ein höheres Reidentifikationsrisiko. Der Grund hierfür ist, dass das Hochrechnungsgewicht ja angibt, wieviele Personen der Grundgesamtheit durch dieses Stichprobenobjekt repräsentiert werden. Ist das Gewicht also klein, so handelt es sich um eine in der Grundgesamtheit selten auftretende Merkmalskombination. Diese Personen müssen besonders geschützt werden.

Im Originaldatensatz sind insgesamt 623 Beobachtungen vorhanden, die eine einzigartige (unique) Ausprägungskombination der Schlüsselvariablen aufweisen. Weiters ergibt sich bei der getroffenen Wahl der Ausprägungen für die Schlüsselvariablen, dass 604 Personen in ihrer Ausprägungskombination der Schlüsselvariablen genau zweimal vorkommen. Das bedeutet, dass in manchen Ausprägungskombination der Schlüsselvariablen nur zwei Personen vorhanden sind.

Aus Abbildung (1) geht hervor, dass das Risiko zur Reidentifikation für die allermeisten Beobachtungen sehr gering ist. Für Personen, die ein relativ hohes Reidentifikationsrisiko aufweisen, werden schließlich schrittweise Werte in den Schlüsselvariablen auf *missing* gesetzt. Dies ist notwendig, um ausreichende Anonymität gewährleisten zu können. Diese Vorgehensweise ergab die folgenden Sperrungen.

- Variable *oenace_1steller*:
  631 Beobachtungen ($\approx 1.06\%$) wurden bei einem Grenzwert für das individuelle Risiko von 1% unterdrückt.

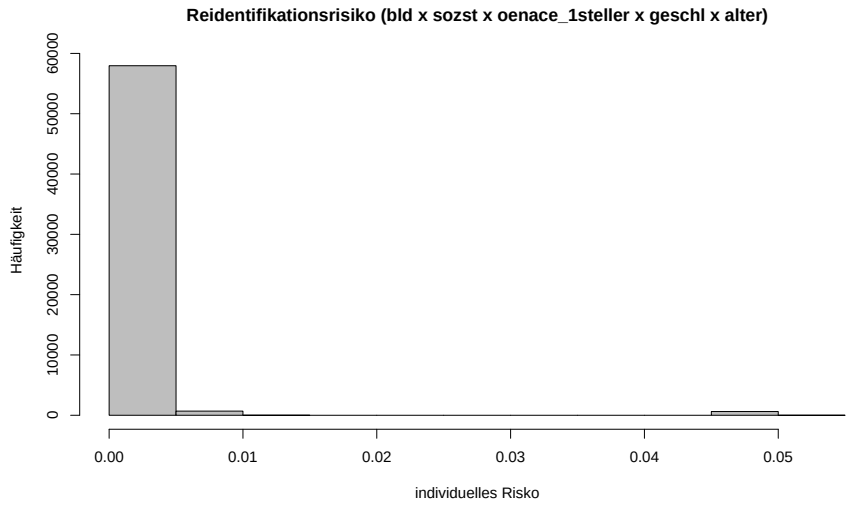

Abbildung 1: Individuelles Risiko für die Reidentifikation in den Originaldaten.

- Variable *bld*:
  154 Beobachtungen ($\approx 0.26\%$) wurden bei einem Grenzwert für das individuelle Risiko von 1% unterdrückt.
- Variable *sozst*:
  93 Beobachtungen ($\approx 0.16\%$) wurden bei einem Grenzwert für das individuelle Risiko von 0.9% unterdrückt.

Diese Ersetzungen führen dazu, dass im Datensatz keine Beobachtung mehr unique ist, dass also jede Ausprägungskombination in den Schlüsselvariablen zumindest zweimal vorkommt. Um weitestgehende Anonymisierung zu gewährleisten, sollte weiters sichergestellt werden, dass jede Ausprägungskombination der Schlüsselvariablen im Datensatz mindestens dreimal existiert. Um diese Vorgabe zu erreichen war es jedoch nicht notwendig, zusätzliche explizite Sperrungen in anderen Variablen durchzuführen.

Schlußendlich ergibt sich ein Datensatz, in dem jede Ausprägungskombination der Schlüsselvariablen zumindest dreimal vorhanden ist. Weiters geht aus Abbildung (2) hervor, dass für die im Datensatz vorhandenen Personen ein sehr geringes Reidentifikationsrisiko besteht.[3]

3.7 Mikroaggregation

Ein Angreifer kann nicht nur Kenntnisse über kategorielle Variablen einer Person ausnutzen, um eine Person zu identifizieren. Auch Informationen über einen (Prozent-)Wert einer numerischen Variable kann unter Umständen dazu ausgenutzt werden, ein Individuum eindeutig zu identifizieren. So kann etwa im Bereich der Lohnsteuerstatistik das Wissen um sehr hohe sonstige Bezüge einer Person (Variable *ntsonst*) in Verbindung mit

[3]Man beachte die unterschiedliche Skalierung der x-Achse in Abbildung (1) und (2).

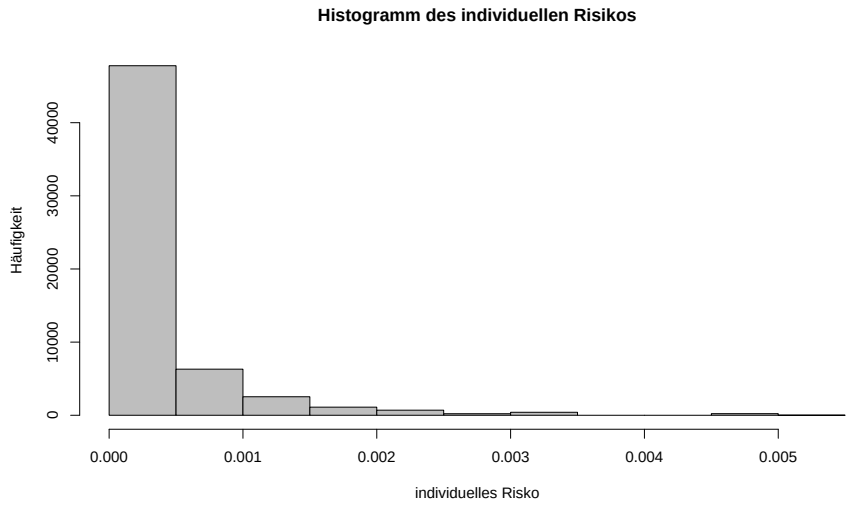

Abbildung 2: Individuelles Risiko für die Reidentifikation im anonymisierten Datensatz.

der Kenntnis über andere indirekte Schlüsselvariable dazu verwendet werden, um eine Person zu identifizieren.

In diesem Fall ist *Mikroaggregation* ein geeignetes Mittel, um zusätzlichen Schutz für den Datensatz zu gewährleisten. Die Idee ist recht einfach. Möglichst "ähnliche" Objekte werden gruppiert und die Ausprägungen einer numerischen Variablen durch den Mittelwert der Ausprägungen der gewählten Objekte ersetzt. Damit wird sichergestellt, dass bei numerischen Variablen jede Ausprägung mehrfach im Datensatz vorhanden ist. Dabei ist zu bemerken, dass die "Verschmutzung" der Daten durch die Mikroaggregation gering ist und die Ergebnisse univariater bzw. multivariater Analyseverfahren gegenüber den Originaldaten approximativ gleich sind.

Welche Variablen im konkreten Anwendungsfall der Erstellung des SDS-Files für Lohnsteuerdaten mikroaggregiert wurden, ist aus Anhang (A) ersichtlich. Es ist bei der Mikroaggregation sichergestellt, dass für jede numerische Variable zumindest 4 Werte gleich sind. Weiters ist zu bemerken, dass die Methode des *individual rankings* verwendet wurde. Diese Methode ist insoferne brauchbar, da sie auch auf Daten mit fehlenden Werten angewendet werden kann. Die Variablen werden unabhängig voneinander in einem ersten Schritt sortiert, anschließend mikroaggregiert und schließlich wieder zurücksortiert.

4 Zusammenfassung und Ausblick

Steuerdaten sind von Natur aus sehr sensibel. Daher lag das Hauptaugenmerk bei der Erstellung des standardisierten Datensastzes für Lohnsteuerdaten naturgemäß auf den Methoden zur Geheimhaltung. Es sollte einem potentiellen Angreifer möglichst schwer gemacht werden, Personen aus dem anonymisierten Datensatz zu identifizieren. Sollte es dennoch gelingen eine Person zu identifizieren, so kann der Angreifer dennoch nicht die

persönlichen Steuerdaten derjenigen Person erfahren. Dies wurde dadurch sichergestellt, dass alle numerischen Variablen mikroaggregiert wurden. Durch die durchgeführten Geheimhaltungsverfahren wie dem Ziehen einer 1% Stichprobe aus dem gesamten Datenbestand, gezieltes Umkodieren sowie Ersetzen kritischer Werte in den Schlüsselvariablen durch *missings* sowie durch Mikroaggregation numerischer Variablen, wurde ein sehr hoher Grad an Anonymität erreicht. Es muss jedoch erwähnt werden, dass es 100%-ige Anonymität nicht geben kann und immer ein, wenn auch sehr kleines, Restrisiko für eine positive Reidentifkation einer Person gegeben ist.

Bei der Erstellung des SDS-Files für Lohnsteuerdaten profitierten wir von Vorarbeiten, die für die Anonymisierung des CVTS2-Datensatzes bereits geleistet wurden. Mit der von Statistik Austria entwickelten Software *sdcMicro* war es in einem explorativen Vorgehen sehr einfach, die unterschiedlichen Geheimhaltungsmethoden anzuwenden. Die Flexibilität, Schnelligkeit und die weitreichende Möglichkeit in der Methodenwahl machen *sdcMicro* zu einem geradezu optimalen Instrument zur Erstellung von anonymisierten Datensätzen.

Ein weiteres wesentliches Ziel war es, trotz aller notwendigen Anonymisierungsmaßnahmen der Wissenschaft einen Datensatz zur Verfügung stellen zu können, der ein hohes Analysepotential bietet. Der vorliegende standardisierte Datensatz wird diesem Anspruch gerecht.

Literatur

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

R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2006. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

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

A Datenbeschreibung

Es werden nun die Variablen beschrieben, die im SDS enthalten sind. Überblicksmässig wird die „Aktion“ der Geheimhaltung in Spalte 4 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=Lehrling 2=Arbeiter(in) 3=Angestellte(r) 4=Beamte(r) 5=Vertragsbedienstete(r) 6=Beamte(r) i.R. 7=Pensionisten ohne Beamte i.R. NA=fehlend
GESCHL	kat. / indirekt	Geschlecht		1=männlich 2=weiblich
NUTS3	kat. / indirekt	NUTS-Gliederung, 3-Steller	vergröbert	Bundesland in Variable BLD
BLD	kat. / indirekt	Bundesland		11=Burgenland 12=Niederösterreich 13=Wien 21=Kärnten 22=Steiermark 31=Oberösterreich 32=Salzburg 33=Tirol 34=Vorarlberg 99=Ausland/unbekannt NA=fehlend
OENACE	kat. / indirekt	ÖNACE-Klassifikation	vergröbert	1-Steller in Variable OENACE_1steller
OENACE_1steller	kat. / indirekt	ÖNACE 1-Steller	erstellt aus Variable OENACE	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
VTBESCH	kat. / nein	Beschäftigung: Vollzeit / Teil-zeit		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 GEB-JAHR	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		