Standardisierter Mikrodatensatz der Reisegewohnheiten für das Jahr 2015

Bernhard Meindl, Magdalena Six

Letztes Update der Dokumentation: Juli 2016

Inhaltsverzeichnis

1 Einleitung **1**

2 Besonderheiten des Datensatzes **2**

2.1 Originaldaten 2

2.2 Modifikation der Originaldaten 2

3 Geheimhaltung **3**

3.1 Software 3

3.2 Direkte Identifikationsvariablen 3

3.3 Indirekte Identifikationsvariablen 3

3.4 Schlüsselvariablen für die Geheimhaltung 3

3.5 Lokale Unterdrückung 4

3.6 Mikroaggregation 4

3.7 Postrandomisierung 5

4 Zusammenfassung **5**

5 Anhang: Datenbeschreibung **5**

5.1 Infos zur Datenbeschreibung 5

1 Einleitung

Ein strategisches Ziel der Bundesanstalt STATISTIK AUSTRIA ist es, für den Zweck der wissenschaftlichen Forschung und Lehre ausgewählte Mikrodatensätze der amtlichen Statistik bereitzustellen. Diese werden in Form von Standardisierten Datensätzen (SDS) über die Webseite der Statistik Austria zugänglich gemacht. Als SDS werden grundsätzlich Einzeldatensätze bezeichnet, die vor der Veröffentlichung so aufbereitet wurden, dass die gesetzlichen Regelungen hinsichtlich Datenschutz erfüllt sind. Durch die Anwendung statistischer Anonymisierungsverfahren wird das Risiko, dass auf Informationen über einzelne statistische Einheiten rückgeschlossen werden kann, minimiert. Des Weiteren müssen potentielle Datennutzer Nutzungsbestimmungen akzeptieren, bevor der Zugriff auf einen SDS ermöglicht wird.

In diesem Dokument wird die Erstellung eines anonymisierten Datensatzes aus dem Bereich der Tourismusstatistik beschrieben. Der SDS liefert für das Jahr 2015 Informationen zu den Reisegewohnheiten der Österreicher und Österreicherinnen. Die Anonymisierungsschritte wurden so gewählt, dass der anonymisierte Mikrodatensatz höchstmöglichen Informationsgehalt bei gleichzeitig möglichst geringem Identifikationsrisiko aufweist.

In der vorliegenden Struktur sind Mikrodatensätze durchgängig seit 2012 verfügbar.

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 über die Reisegewohnheiten der Österreicher und Österreicherinnen basiert auf vierteljährlichen Stichprobenerhebungen, die laufend mit dem Ziel durchgeführt werden, das nationale Reiseverhalten der im Inland wohnhaften Personen ab 15 Jahren abzubilden. Je Quartal werden im Rahmen dieser Erhebung rund 3500 ausgewählte, im Inland wohnhafte Personen ab 15 Jahren, verteilt über Österreich, telefonisch befragt (Grundgesamtheit von rd. 7.2 Millionen.). Die Teilnahme an der Erhebung ist freiwillig. Der Auswahlrahmen für die proportional geschichtete Stichprobe ist das Zentrale Melderegister (ZMR). Die Telefonnummern werden dem öffentlichen Telefonbuch entnommen. Zusätzlich wurden Informationen über das Alter, das Wohnsitzbundesland sowie das Geschlecht der Reisenden und Nicht-Reisenden aus dem Zentralen Melderegister hinzugefügt. Aus den umfangreichen Informationen des Datenbestandes wurden schlussendlich insgesamt 30 Variablen ausgewählt, die - zusammen mit den durchgeführten Umkodierungen - im Anhang (Kapitel 5) 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, umkodiert bzw. neu erstellt wurden, um der Geheimhaltung Rechnung zu tragen. Nachfolgend werden alle Variablen, die verändert bzw. neu erstellt worden sind, aufgelistet.

- *persID*: fortlaufende Identifizierungsnummer einer Person wurde neu erstellt
- *Alter*: das Alter in Jahren wurde zu 10-Jahres Altersgruppen vergröbert
- *Reiseland* einzelne Zielländer wurden zusammengefasst
- *Gesamtausgaben*: Die Gesamtausgaben (in Euro) der jeweiligen Reise wurden berechnet
- *p_Ausgaben_Transport*: Anteil der Ausgaben für Transport an den Gesamtausgaben (neu)
- *p_Ausgaben_Unterkunft* Anteil der Ausgaben für Unterkünfte an den Gesamtausgaben (neu)
- *p_Ausgaben_Andere* Anteil der Ausgaben für Sonstiges an den Gesamtausgaeen (neu)
- *p_Ausgaben_Wertgueter* Anteil der Ausgaben für Wertgüter an den Gesamtausgaben (neu)

Im Anhang (Kapitel 5) ist die tatsächliche Kodierung aller im SDS enthaltenen Variablen ersichtlich.

3 Geheimhaltung

Es wird nun die Anonymisierungsprozedur beschrieben, die durchgeführt wurde, um aus dem authentischen Datenbestand für das Jahr 2015 einen SDS-File zu erzeugen.

3.1 Software

Die Anonymisierungsprozedur mit der freien Statistiksoftware *R* sowie dem von Statistik Austria entwickelten und frei verfügbaren **R**-Paket *sdcMicro* (**s**tatistical **d**isclosure **c**ontrol for **micro**data) durchgeführt. Das Paket 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 diesem SDS zugrundeliegt, wurde die *ZMR-Nummer* der Reisenden durch eine fortlaufende ID (Variable *persID*) ersetzt.

3.3 Indirekte Identifikationsvariablen

Kann durch Kombination mehrerer (meist kategorieller) Variablen eine Person eindeutig im Datensatz identifiziert werden, so werden diese Variablen als indirekte Identifikationsvariablen bezeichnet. Wichtig zu bemerken ist, dass keine indirekte Identifikationsvariable für sich selbst zur eindeutigen Identifizierung einer Person im Datensatz ausreicht.

Als indirekte Identifikationsvariablen in den Daten über die Reisegewohnheiten der Österreicher und Österreicherinnen können beispielsweise die höchste abgeschlossene Schulbildung (Variable *Schulbildung*), Information über die berufliche Stellung (Variable *Berufstaetigkeit*), das Alter (Variable *Alter*), Geschlecht (Variable *Geschlecht*) oder Information über den Wohnsitz eines Befragten (Variable *Wohnsitzbundesland*) herangezogen werden. 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 höchstmögliche Anonymität gewährleisten zu können.

3.4 Schlüsselvariablen für die Geheimhaltung

Indirekte Identifikationsvariablen, deren Ausprägungskombinationen ein Angreifer verwenden könnte, um eine eindeutige Identifikation einer Person im Datensatz vorzunehmen, werden allgemein als

Schlüsselvariablen oder *Key-Variablen* bezeichnet. Für den vorliegenden Datenbestand wurden folgende Variablen als *Schlüsselvariablen* definiert.

- *Wohnsitzbundesland*: 10 Ausprägungen
- *Geschlecht*: 2 Ausprägungen
- *Schulbildung*: 4 Ausprägungen
- *Berufstaetigkeit*: 5 Ausprägungen
- *Alter*: 6 Ausprägungen

Eine Möglichkeit, einen sicheren SDS mit hohem Analysepotential zu erhalten, besteht darin, einzelne Werte in den Schlüsselvariablen bewußt zu löschen um schließlich k-Anonymität gewährleisten zu können.

3.5 Lokale Unterdrückung

Im Zuge der Anonymisierungsprozedur wurde durch gezielte Sperrungen von einzelnen Ausprägungen in den Schlüsselvariablen erreicht, dass jeder Ausprägungskombination in den definierten Schlüsselvariablen zumindest 3 Personen zugeordnet werden können. Dies wird auch als *3-Anonymity* bezeichnet. Die nachfolgende Tabelle zeigt die Anzahl der notwendigen Sperrungen in den Schlüsselvariablen.

Tabelle 1: Sperrungen in Schlüsselvariablen zur Erreichung von 3-Anonymität

Wohnsitzbundesland	**Geschlecht**	**Schulbildung**	**Berufstaetigkeit**	**Alter**
5	0	26	177	0

3.6 Mikroaggregation

Unter Umständen besteht die Möglichkeit, dass ein Datenangreifer ihm bekannte Informationen über einen 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 - üblicherweise den Mittelwert ersetzt. Durch die Mikroaggregation numerischer Variablen wird sichergestellt, dass jede einzelne Ausprägung mehrfach im Datensatz auftritt. Die *Verschmutzung* der Daten selbst durch das Mikroaggregationsverfahren ist gering. Dies ergibt sich aus dem Vergleich univariater bzw. multivariater Statistiken der Originaldaten mit den mikroaggregierten Daten.

Aus der Datenbeschreibung (Kapitel 5) geht hervor, dass die Variable *Gesamtausgaben* in diesem Datensatz mikroaggregiert wurde. Es wurde sichergestellt, dass jeder Wert der mikroaggregierten Variablen zumindest 3 Mal in dieser Variable auftritt.

3.7 Postrandomisierung

Um Variablen, die für die Datenanalyse absolut notwendig sind und aus diesem Grunde nicht aus dem Datensatz entfernt oder vergröbert werden können zu schützen, besteht die Möglichkeit, ein Postrandomisierungsverfahren anzuwenden. Die grundlegende Idee bei PRAM ist es, dass grundsätzlich jede Ausprägung in einer Variable entweder unverändert bleibt oder in eine andere Kategorie wechseln kann. Hinter dem Verfahren stehen Übergangsmatrizen, die den Zufallsprozess steuern. Die tatsächlich gewählten Wahrscheinlichkeiten für das Wechseln oder Nicht-Wechseln können nicht publiziert werden. Es ist jedoch zu erwähnen, dass die Wahrscheinlichkeit für einen Verbleib in der jeweiligen Kategorie aus den Originaldaten hoch ist um die innere Struktur des Datensatzes möglichst wenig zu verändern.

Die Variablen des vorliegenden Datensatzes, auf die Postrandomisierungstechniken angewendet wurden, ist in Kapitel 5 ersichtlich.

4 Zusammenfassung

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

Durch die angewandten Anonymisierungsverfahren wie die Aggregation beziehungsweise das Umkodieren kategorieller Variablen, dem Ersetzen kritischer Werte in den Schlüsselvariablen durch *missings*, durch Mikroaggregation numerischer Variablen sowie Postrandomisierung wurde erreicht, dass das Reidentifikationsrisiko aller im SDS verbleibenden Daten sehr gering ist. Allerdings ist anzumerken, dass es 100%-igen Schutz vor Aufdeckung sensibler Information nicht geben kann. Ein (geringes) Restrisiko bleibt also bestehen.

Beim Erstellen des Standardisierten Datensatzes über die Reisegewohnheiten der Österreicher und Österreicherinnen wurde darauf geachtet, trotz der notwendigen Anonymisierung der Daten das hohe Analysepotential der Daten zu erhalten. Der vorliegende standardisierte Datensatz wird diesem Anspruch gerecht. Des Weiteren ist es möglich, den SDS in Zukunft mit aktualisierten und neuen Daten zu erweitern.

5 Anhang: Datenbeschreibung

5.1 Infos zur Datenbeschreibung

Im folgenden Teil dieses Dokuments werden die Variablen beschrieben, die im SDS enthalten sind. Wurde bei einer Variable eine Geheimhaltungsaktion angewandt, so ist diese in Klammer neben dem Variablennamen angebeben. 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 postrandomisiert und rot bedeutet, dass diese Variable mikroaggregiert wurde.

5.1.1 persID

eindeutiger Identifier einer Person. Haben mehrere Datenzeilen dieselbe persID, so sind dies die verschiedenen Reisen einer Person.

Codes:

- fortlaufend

5.1.2 Geschlecht (Modifikation: Schlüsselvariable)

enthält das Geschlecht der Person.

Codes:

- 1: männlich
- 2: weiblich

5.1.3 Alter (Modifikation: Schlüsselvariable | vergröbert)

Das Alter der Reisenden wurde in 10-Jahres Gruppen umgeschlüsselt.

Codes:

- 15-24: 15 bis einschließlich 24 Jahre
- 25-34: 25 bis einschließlich 34 Jahre
- 35-44: 35 bis einschließlich 44 Jahre
- 45-54: 45 bis einschließlich 54 Jahre
- 55-64: 55 bis einschließlich 64 Jahre
- 65+: 65 Jahre oder älter

5.1.4 Wohnsitzbundesland (Modifikation: Schlüsselvariable)

enthält das Wohnbundesland des Reisenden.

Codes:

- AT11: Burgenland
- AT12: Niederösterreich
- AT13: Wien
- AT21: Kärnten

- AT22: Steiermark
- AT31: Oberösterreich
- AT32: Salzburg
- AT33: Tirol
- AT34: Vorarlberg

5.1.5 Schulbildung (Modifikation: Schlüsselvariable)

höchste abgeschlossene Ausbildung der/des Reisenden nach der International Standard Classification of Education (ISCED).

Codes:

- 1: ISCED 2011 Level 0-2 (Primärschulbildung und Pflichtschulabschluss)
- 2: ISCED 2011 Level 3 and 4 (mittlere und höhere Schulbildung)
- 3: ISCED 2011 Level 5-8 (Hochschulbildung)

5.1.6 Berufstaetigkeit (Modifikation: Schlüsselvariable)

berufliche Stellung der/des Reisenden,

Codes:

- 1: Erwerbstätig (unselbstständig oder selbstständig)
- 2: Arbeitslos
- 3: Student/Schüler
- 4: Andere Nicht-Erwerbspersonen

5.1.7 Anzahl_der_Personen_im_Haushalt (Modifikation: Top-Coding)

Anzahl der im Haushalt lebenden Personen, Haushalte mit mehr als 5 Personen werden zu Haushalten mit 6 Personen zusammengefasst.

Codes:

- 1-5: numerisch
- 6: 6 Personen oder mehr

5.1.8 Anzahl_der_Personen_bei_der_Reise (Modifikation: Top-Coding)

Anzahl der bei der Reise teilnehmenden Personen, die mit dem bzw. der Befragten im Haushalt leben, inklusive dem bzw. der Befragten. Reisen mit mehr als 4 teilnehmenden Personen wurden zu Reisen mit 5 Teilnehmern zusammengefasst.

Codes:

- 1-4: numerisch
- 5: 5 Personen oder mehr

5.1.9 Abreisemonat (Modifikation: PRAM)

Kalendermonat der Abreise.

Codes:

- 1: Jänner
- ...
- 12: Dezember

5.1.10 Abreisejahr

Kalenderjahr der Abreise: 2015. Standardisierte Mikrodatensätze gibt es derzeit für die Jahre 2012, 2013, 2014, 2015. Das Kalendermonat ist aus der Variable *Abreisemonat* ersichtlich.

5.1.11 Anzahl_Naechtigungen

Gesamtanzahl der Nächtigungen.

Codes:

- numerisch

5.1.12 Anzahl_Naechtigungen_Oest

Im Fall von Auslandsreisen die Anzahl der Nächtigungen bei der Anreise bzw. Abreise, die in Österreich verbracht wurden.

Codes:

- numerisch für Auslandsreisen
- NA für Inlandsreisen

5.1.13 Reisezweck

Hauptsächlicher Reisezweck

Codes:

- 1: Strand- und Badeaufenthalt
- 2: Aktivurlaub
- 3: Erholungsurlaub
- 4: Wellness-/Schönheitsurlaub
- 5: Gesundheitsurlaub
- 6: Verwandten/Bekanntenbesuche
- 7: (nicht-berufliche) Ausbildung
- 8: Kultur und Besichtigung
- 9: Shopping
- 10: Besuch einer Veranstaltung, eines Events oder Festivals
- 11: Sonstige Urlaubsreisezwecke
- 12: Geschäftlich: Kongresse, Messen, berufliche Weiter- bzw. Fortbildung
- 13: Sonstige geschäftliche Zwecke

5.1.14 Stadt

Art des Urlaubsreiseziels - Stadt, Mehrfachnennungen in der Art des Urlaubsreiseziels sind möglich

Codes:

- 1: Stadt: Ja
- 2: Stadt: Nein
- 9: Nicht anwendbar da Geschäftsreise

5.1.15 Meer

Art des Urlaubsreiseziels - Meer, Mehrfachnennungen in der Art des Urlaubsreiseziels sind möglich

Codes:

- 1: Reiseziel Meer: Ja
- 2: Reiseziel Meer: Nein
- 9: Nicht anwendbar da Geschäftsreise

5.1.16 Land

Art des Urlaubsreiseziels - Ort in ländlichem Gebiet, Mehrfachnennungen in der Art des Urlaubsreiseziels sind möglich

Codes:

- 1: Reiseziel ländliches Gebiet: Ja
- 2: Reiseziel ländliches Gebiet: Nein
- 9: Nicht anwendbar da Geschäftsreise

5.1.17 Kreuzfahrtsschiff

Art des Urlaubsreiseziels - Kreuzfahrtsschiff, Mehrfachnennungen in der Art des Urlaubsreiseziels sind möglich

Codes:

- 1: Reiseziel Kreuzfahrtsschiff: Ja
- 2: Reiseziel Kreuzfahrtsschiff: Nein
- 9: Nicht anwendbar da Geschäftsreise

5.1.18 Berg

Art des Urlaubsreiseziels - Gebirge, Mehrfachnennungen in der Art des Urlaubsreiseziels sind möglich

Codes:

- 1: Reiseziel Gebirge: Ja
- 2: Reiseziel Gebirge: Nein
- 9: Nicht anwendbar da Geschäftsreise

5.1.19 Anderes

Art des Urlaubsreiseziels - Sonstiges, Mehrfachnennungen in der Art des Urlaubsreiseziels sind möglich

Codes:

- 1: Reiseziel Sonstiges: Ja
- 2: Reiseziel Sonstiges: Nein
- 9: Nicht anwendbar da Geschäftsreise

5.1.20 Kinder

Mitreisende Kinder

Codes:

- 1: Mindestens ein mitreisendes Kind
- 2: Keine mitreisenden Kinder
- 9: Nicht anwendbar da Geschäftsreise

5.1.21 Transportmittel

hauptsächlich benutztes Transportmittel

Codes:

- 1: Flugzeug
- 2: Schiff
- 3: Zug
- 4: Bus/Reisebus
- 5: PKW (eigener/gemietet)
- 6: Sonstige

5.1.22 Unterbringung (Modifikation: vergröbert)

vorwiegende Unterbringung

Codes:

- 1: Hotels und ähnliche Betriebe
- 2: Private Unterkünfte (Bezahlung)
- 3: Private Unterkünfte (Gratis)
- 4: Sonstige Beherbergungsbetriebe und spezielle Unterkünfte

5.1.23 Pauschalreise

Pauschalreise

Codes:

- 1: Ja
- 2: Nein

5.1.24 Reiseland (Modifikation: vergröbert)

Ausländischer Staat bzw. österr.Bundesland, in welchem die Reise hauptsächlich verbracht wurde.

Codes:

- AT11: Burgenland
- AT12: Niederösterreich
- AT13: Wien
- AT21: Kärnten
- AT22: Steiermark
- AT31: Oberösterreich
- AT32: Salzburg
- AT33: Tirol
- AT34: Vorarlberg
- AT: Rundreise in Österreich
- 1: Belgien
- 2: Dänemark
- 3: Deutschland
- 4: Finnland
- 5: Frankreich
- 6: Griechenland
- 7: Vereinigtes Königreich
- 8: Irland
- 9: Italien
- 10: Luxembourg
- 11: Niederlande
- 12: Portugal
- 13: Schweden
- 14: Spanien
- 15: Island
- 16: Norwegen

- 17: Schweiz
- 18: Baltikum (Estland, Lettland, Litauen)
- 19: Kroatien
- 22: Malta
- 23: Polen
- 24: Rumänien
- 25: Slowakei
- 26: Slowenien
- 27: Türkei
- 28: Tschechische Republik
- 29: Ungarn
- 30: Zypern
- 31: Bosnien Herzegowina
- 32: Serbien
- 33: Bulgarien
- 34: Russland
- 35: restl.Europa inkl. Bosnien Herzigowina, Serbien, Bulgarien
- 36: Ägypten
- 37: In den Jahren 2012-2014 wurde Tunesien extra ausgegeben, aufgrund einer zu geringen Anzahl an Reisen im Jahr 2015 wird Tunesien 2015 zur Kategorie 38 (restl. Afrika) gezählt
- 38: restl. Afrika inklusive Tunesien
- 39: USA
- 40: Kanada
- 41: Mittel- und Südamerika
- 42: China
- 43: restliches Asien
- 44: Australien, Neuseeland und Inseln nordöstlich davon im Indischen Ozean

5.1.25 Gesamtausgaben (Modifikation: Mikroaggregiert)

Gesamtausgaben (in Euro) der Reise, Summe aus Ausgaben für Transport, Unterkunft und sonstigen Ausgaben - numerisch codiert. Die Gesamtausgaben beziehen sich auf alle im gemeinsamen Haushalt lebenden Mitglieder der Reisegruppe. Kinder wurden bei der Berechnung berücksichtigt.

5.1.26 p_Ausgaben_Transport (Modifikation: neu erstellt)

Anteil der Ausgaben für Transport an den Gesamtausgaben - numerisch kodiert.

5.1.27 p_Ausgaben_Unterkunft (Modifikation: neu erstellt)

Anteil der Ausgaben für Unterkünfte an den Gesamtausgaben - numerisch kodiert.

5.1.28 p_Ausgaben_Andere (Modifikation: neu erstellt)

Anteil der sonstigen Ausgaben an den Gesamtausgaben, Ausgaben für Wertgüter sind ein Teil der sonstigen Ausgaben - numerisch kodiert

5.1.29 p_Ausgaben_Wertgueter (Modifikation: neu erstellt)

Anteil der Ausgaben für Wertgüter an den Gesamtausgaben - numerisch kodiert.

5.1.30 Gewicht

Stichprobengewicht zur Hochrechnung auf die Grundgesamtheit.