Nutzungskonzept Taxpayer-Panel 2001-2012

Ulrike Gerber, Stefanie Uhrich

1. Vorbemerkungen

Paneldaten sind Daten, die für dieselben Beobachtungseinheiten Angaben für mehrere Zeitpunkte enthalten. Sie ermöglichen die Betrachtung von Phänomenen im Zeitverlauf und erfreuen sich daher in der wirtschafts- und sozialwissenschaftlichen Forschung wachsender Beliebtheit. Das Statistische Bundesamt stellt seit 2007 Paneldaten aus dem Bereich der Einkommensteuer zu Analysezwecken bereit, das sogenannte Taxpayer-Panel (TPP). Die ersten Erfahrungen mit Auswertungen zum TPP 2001-2004 zeigten, dass Berechnungen auf Grundlage des Gesamtmaterials sehr zeitaufwändig sind und erhebliche Speicherkapazitäten erfordern. Im Hinblick auf die Ausweitung des Nutzerkreises und die sukzessive Ergänzung des TPP um weitere Wellen wurde deshalb das vorliegende Nutzungskonzept erarbeitet.

Mit dem TPP 2001-2012 wurden einige grundlegende Neuerungen in der Methodik umgesetzt.

2. Datenmaterial

Die Wellen 2001 bis 2011 des TPP wurden auf Basis der jährlichen Einkommensteuerstatistik (Geschäftsstatistik) erstellt. Diese Querschnittsdaten enthalten die Angaben aus den Einkommensteuererklärungen der rund 27 Millionen veranlagten deutschen Steuerpflichtigen[1] und wurden über die Steuernummern sowie indirekte Indikatoren zu einem Panel verknüpft.

Ab dem Veranlagungsjahr 2012 entfällt die Geschäftsstatistik. Die bis dahin dreijährliche Bundesstatistik zur Lohn- und Einkommensteuer wird fortan jährlich erhoben. Das TPP wird deshalb ab Welle 2012 mit den Daten der Bundesstatistik fortgeführt. Die Bundesstatistik umfasst neben den veranlagten Steuerpflichtigen auch rund 12 Millionen nicht veranlagte Steuerpflichtige, die sogenannten Lohnsteuerfälle. Aufgrund von Verzögerungen bei der Lieferung dieser Fälle, werden sie aber erst ab Veranlagungsjahr 2013 im TPP enthalten sein.

Des Weiteren steht seit dem Veranlagungsjahr 2010 die Steueridentifikationsnummer als Verknüpfungsmerkmal zur Verfügung, so dass auch das Konzept für die Zusammenführung der Datensätze umgestellt wurde. Die Verknüpfungen werden ab Welle 2012 ausschließlich über eindeutige Indikatoren (Steueridentifikations- und Steuernummern) vorgenommen. Dadurch ist es möglich, Daten auch über mehr als zwei unbesetzte Jahre hinweg einem bestimmten Steuerpflichtigen zuzuordnen.[2]

Das Panel 2001-2012 weist insgesamt 40.724.427 Datensätze auf, zu denen Angaben für mindestens zwei Jahreswellen vorliegen. Für Auswertungen sind grundsätzlich alle Variablen der jährlichen Einkommensteuerstatistiken verfügbar.[3] Es ist zu beachten, dass es durch den Wechsel der Datengrundlage (Bundes- anstelle Geschäftsstatistik) teilweise zu **Verschiebungen bei der Belegung der Kennzahlen ab Welle 2012** kommt. So werden in der Bundesstatistik alle Informationen zu Frauen in den B-Kennziffern ausgegeben, in den bisherigen Geschäftsstatistiken stehen die Werte alleinveranlagter Frauen in den A-Kennziffern (analog zu alleinveranlagten Männern). Die unterschiedliche Belegung dieser Kennzahlen muss insbesondere bei ge-

[1] Im Jahr 2008 wurden die hessischen Steuerpflichtigen nicht vollständig erfasst.
[2] Bisher wurden Verknüpfungen nur für drei zurückliegende Jahre durchgeführt, da die Finanzverwaltung nach Ablauf zweier Jahre alte Steuernummern neu vergeben kann.
[3] Schlüsselnummern (z.B. Steuer- oder Vertragsnummern) sowie Gebietszuordnungen unterhalb der Kreisebene wurden entfernt.

schlechtsspezifischen Auswertungen beachtet werden. Eine Aufstellung der Variablen - differenziert nach Veranlagungsjahren - ist in der Anlage enthalten.

In der Einkommensteuerstatistik (und somit im TPP) sind fehlende Werte aus Speicherplatzgründen durch „0“ ersetzt. Da die Finanzverwaltungen nicht einheitlich zwischen errechneten „0“-Werten und fehlenden Werten unterscheiden, entsteht dadurch kein Informationsverlust.

3. Nutzungskonzept

Den Anforderungen der Nutzer aus der Wissenschaft und den IT-technischen Gegebenheiten entsprechend werden vier unterschiedliche Zugangswege für die Analyse des TPP angeboten. Untenstehende Tabelle fasst die vier Produkte zum Panel zusammen.

Produkt	Anzahl der Datensätze	Maximale Variablen-anzahl	Anonymisierung	Form der Nutzung	Analyse-software
Struktur-datenfile	703	Keine Beschrän-kung	Vollständige Anonymisierung	Off-Site	SAS, SPSS, Stata
0,5%-Stichprobe	140.867	Keine Beschrän-kung	Teilweise Verallgemeinerung von Bundesland und amtlichem Gemeindeschlüssel, Verallgemeinerung von Geburts-datum und Religion, Löschung der Höchstverdiener, Anonymisierung der Ergebnisse	Gastwissen-schaftler-arbeitsplatz im FDZ (Bund und Länder)	SAS, SPSS, Stata[4]
5%-Stichprobe	1.405.387	Keine Beschrän-kung	Teilweise Verallgemeinerung von Bundesland und amtl. Gemein-deschlüssel, Anonymisierung der Ergebnisse	Kontrollierte Datenfernver-arbeitung im FDZ (Bund)	SAS, Stata
Gesamt-material (mind. zwei Jahre besetzt)	40.724.427	25 je Welle	Keine Anonymisierung der Daten, Anonymisierung der Ergebnisse	Kontrollierte Datenfernver-arbeitung in der Fachabtei-lung (Bund)	SAS

Das **Strukturdatenfile** enthält 703 absolut anonymisierte Datensätze im Umfang der vom Nutzer beantragten Variablen. Die absolute Anonymisierung wird hier durch zufälliges Vertauschen von Merkmalsausprägungen erreicht. Dabei bleibt die Charakteristik der Variablen erhalten, der Datensatz kann aber in sich inkonsistent sein. Das Strukturdatenfile wird direkt an den Nutzer verschickt und ist ausschließlich zur Überprüfung der erstellten Syntaxen gedacht.

Für die Nutzung am Gastwissenschaftlerarbeitsplatz in den Forschungsdatenzentren (FDZ) der Statistischen Ämter des Bundes und der Länder ist eine **0,5%-Stichprobe** vorgesehen. Dabei handelt es sich um eine faktisch anonymisierte Substichprobe der geschichteten 5%-Stichprobe (siehe unten). Die Anonymisierung wird erreicht, indem neben der Zusammenfassung von Bundesländern nach West/Ost in den zwei höchsten Einkunftsklassen zusätzlich

- alle Geburtsdaten auf den 31. Dezember des jeweiligen Jahres gesetzt werden,
- jeweils die 10 Datensätze mit den höchsten Einkommen für West- und Ostdeutschland gelöscht werden und

[4] Für den Fall, dass die bereits am Gastwissenschaftlerarbeitsplatz entwickelten Syntaxen anschließend noch über die 5% Stichprobe für die kontrollierte Datenfernverarbeitung laufen gelassen werden sollen, kann dies nur mit der Analysesoftware SAS oder Stata geschehen.

- die Variablen ef13 (Religion -A-) und ef14 (Religion -B-) auf folgende Ausprägungen umkodiert werden[5]:
  10 – evangelisch
  14 – katholisch
  16 – sonstige
  20 – keine.

Dieses Material hat den Vorteil, dass der Forscher seine Untersuchung unmittelbar an den Originaldaten entwickeln kann. Die 0,5%-Stichprobe kann mit den Analysesoftwares SAS, SPSS oder STATA bearbeitet werden.

Tiefer gehende Analysen sind mit der gewichteten **5%-Stichprobe** möglich, die über das FDZ des Statistischen Bundesamtes im Rahmen der kontrollierten Datenfernverarbeitung ausgewertet werden kann. Die Stichprobe wurde aus allen Steuerpflichtigen mit Angaben in mindestens fünf Jahren gezogen (28.725.488 Datensätze)[6] und ist nach bestimmten Variablen geschichtet:

(1) Bundesland,
(2) Grund-/Splittingtabelle,
(3) Überwiegender Einkunftsart,
(4) Median des Gesamtbetrages der Einkünfte über die besetzten Wellen und
(5) Relativer Variation des GdE zwischen den Jahren.

In der höchsten Einkunftsklasse wurden die Bundesländer nach West/Ost zusammengefasst:

Bundesland (bl): 98 – Alte Bundesländer
99 – Neue Bundesländer und Berlin.
Amtlicher Gemeindeschlüssel (ef7):
98888 – Alte Bundesländer
99999 – Neue Bundesländer und Berlin.

Eine ausführliche Beschreibung der Stichprobenziehung findet sich in der Anlage. Die Anzahl der Variablen ist bei dieser Zugangsform nicht beschränkt. Für dieses Produkt ist grundsätzlich die vorherige Arbeit mit der 0,5%-Stichprobe am Gastwissenschaftlerarbeitsplatz vorgesehen[7].

Das **Gesamtmaterial** zum Taxpayer-Panel (mind. zwei Jahre besetzt) wird nur für einzelne Auswertungen zu ausgewählten Variablen (max. 25 je Veranlagungsjahr) bereitgestellt. Interessierte haben die Möglichkeit, SAS-Codes an das Statistische Bundesamt zu schicken (Referat F308). Vor Herausgabe werden die Ergebnisse zur Sicherstellung der statistischen Geheimhaltung überprüft. Einzeldaten wie z.B. minimale und maximale Ausprägungen werden grundsätzlich gelöscht. Bei der Nutzung des Gesamtmaterials wird vorausgesetzt, dass die Auswertungen bereits anhand der 0,5%-Stichprobe durchgeführt wurden.

Wiesbaden, November 2016

Anlagen

DSB_Panel_2016.xls (Variablenauswahl)
Zusammenfassung der Religionsschlüssel.doc
Stichprobenziehung Taxpayerpanel 01-12

[5] Eine Aufstellung, wie die Religionsschlüssel zusammengefasst wurden, findet sich im Anhang.
[6] Bisher bildete ein balanciertes Panel (alle Wellen besetzt) die Grundlage der Stichprobenziehung, weshalb Untersuchungen zu Aufnahme und Aufgabe bestimmter Tätigkeiten anhand der Stichproben nicht möglich waren.
[7] Hierfür entstehen keine zusätzlichen Kosten. Die Ergebnisfreigabe erfolgt dabei ausschließlich über die kontrollierte Datenfernverarbeitung. Dies gilt entsprechend für die Nutzung des Gesamtmaterials.