Nutzungskonzept Taxpayer-Panel 2001-2014

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, die sich auch auf das aktuelle Panel auswirken.

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[2] auch rund 12 Millionen nicht veranlagte Steuerpflichtige, die sogenannten Lohnsteuerfälle.

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 Identifikatoren (Steueridentifikations- und Steuernummern) vorgenommen. Dadurch ist es möglich, Daten auch über mehr als zwei unbesetzte Jahre hinweg einem bestimmten Steuerpflichtigen zuzuordnen.[3]

Das Panel 2001-2014 weist insgesamt 54.614.370 Datensätze auf, zu denen Angaben für mindestens zwei Jahreswellen vorliegen. Für Auswertungen sind grundsätzlich alle Variablen der jährlichen Einkommensteuerstatistiken verfügbar. Schlüsselnummern (z.B. Steuer- oder Vertragsnummern) wurden entfernt. 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 geschlechtsspezifischen

[1] Im Jahr 2008 wurden die hessischen Steuerpflichtigen nicht vollständig erfasst.
[2] Zu den veranlagten Steuerpflichtigen zählen ab Veranlagungsjahr 2012 zusätzlich etwa 12.000 Verlustfeststellungs- und 15.000 Nur-Sparzulagenfälle, für die nur sehr wenige Angaben vorliegen.
[3] Bisher wurden Verknüpfungen nur für drei zurückliegende Jahre durchgeführt, da die Finanzverwaltung nach Ablauf zweier Jahre alte Steuernummern neu vergeben kann.

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 Datenprodukte 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
Datenstruktur-file	790	Keine Beschrän-kung	Vollständige Anonymisie-rung der Daten, keine inhaltlichen Aussagen möglich	Off-Site	SAS, SPSS, Stata
0,5%-Stichprobe	158.008	Keine Beschrän-kung	Formale Anonymisierung der Daten[4], absolute Anonymisierung der Ergebnisse	Gastwissenschafts-arbeitsplatz in den FDZ (derzeit nur Bund)	SAS, SPSS, Stata[5]
5%-Stichprobe	1.579.443	Keine Beschrän-kung	Formale Anonymisierung der Daten, absolute Anonymisierung der Ergebnisse	Kontrollierte Datenfernverarbei-tung im FDZ (Bund)	SAS, Stata
Gesamt-material (mind. zwei Jahre besetzt)	54.614.370	25 je Welle	Keine Anonymisierung der Daten, absolute Anonymisierung der Ergebnisse	Kontrollierte Datenfernverarbei-tung in der Fachab-teilung (Bund)	SAS

Der **Datenstrukturfile** enthält 790 absolut anonymisierte Datensätze. 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 und lässt keine inhaltlichen Aussagen zu. Der Datenstrukturfile wird direkt an den Nutzer verschickt und ist ausschließlich zur Überprüfung der im Rahmen der kontrollierten Datenfernverarbeitung erstellten Syntaxen gedacht.

Für die Nutzung an einem Gastwissenschaftsarbeitsplatz in den Forschungsdatenzentren (FDZ) der Statistischen Ämter des Bundes und der Länder ist eine **0,5%-Stichprobe** vorgesehen. Dabei handelt es sich um eine Zufallsstichprobe aus der geschichteten und formal anonymisierten 5%-Stichprobe. Da der Amtliche Gemeindeschlüssel für bayerische Gemeinden am Gastwissenschaftsarbeitsplatz lediglich in anonymisierter Form bereitgestellt werden kann, sind bei diesen Datensätzen die letzten drei Stellen des Amtlichen Gemeindeschlüssels pseudoanonymisiert.

[4] Für bayerische Gemeinden kann der Amtliche Gemeindeschlüssel lediglich in anonymisierter Form am Gastwissenschaftsarbeitsplatz bereitgestellt werden.
[5] 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.

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 (31.588.881 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.

Eine ausführliche Beschreibung der Stichprobenziehung findet sich in der Anlage. Die Anzahl der Variablen ist bei dieser Zugangsform nicht beschränkt, Schlüsselnummern wie z.B. Steuer- oder Vertragsnummern wurden entfernt. Für dieses Produkt wird die vorherige Arbeit mit der 0,5%-Stichprobe am Gastwissenschaftsarbeitsplatz empfohlen[7].

Das **Gesamtmaterial** zum Taxpayer-Panel (mind. zwei Jahre besetzt) wird nur für einzelne Auswertungen bereitgestellt. Dafür wird in der Fachabteilung des Statistischen Bundesamtes (Referat F306) ein nutzerspezifisches Panel mit maximal 25 Variablen je Veranlagungsjahr erstellt und der vom Nutzer vorab getestete SAS-Code angewendet. Vor Herausgabe werden die Ergebnisse zur Sicherstellung der statistischen Geheimhaltung überprüft. Einzeldaten wie z.B. minimale und maximale Ausprägungen werden grundsätzlich gelöscht.[8] Bei der Nutzung des Gesamtmaterials wird vorausgesetzt, dass die Auswertungen bereits anhand der 0,5%-Stichprobe durchgeführt wurden.

Wiesbaden, Dezember 2018

Anlagen

DSB_TPP2014.xlsx (Variablenauswahl)
Stichprobenziehung Taxpayerpanel 01-14

[6] Bis einschließlich TPP2010 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] Bei gleichzeitiger Nutzung der kontrollierten Datenfernverarbeitung entstehen hierfür keine zusätzlichen Kosten. Die Ergebnisfreigabe erfolgt dabei ausschließlich über die kontrollierte Datenfernverarbeitung. Dies gilt entsprechend für die Nutzung des Gesamtmaterials.
[8] Hinweise zu den Geheimhaltungsverfahren bieten die „Regeln zur Auswertung der Geheimhaltung“ mit der zugehörigen „Übersicht Geheimhaltungsregeln“:
http://www.forschungsdatenzentrum.de/de/geheimhaltung