

FDZ-Arbeitspapier
Nr. 14

Daniel Vorgrimler
Christopher Gräß
Susan Kriete-Dodds



STATISTISCHE ÄMTER
DES BUNDES UND DER LÄNDER
FORSCHUNGSDATENZENTREN

Zur Konzeption eines
Taxpayer-Panels für
Deutschland

2006

FDZ-Arbeitspapier
Nr. 14

Daniel Vorgrimler
Christopher Gräß
Susan Kriete-Dodds

 **STATISTISCHE ÄMTER**
DES BUNDES UND DER LÄNDER
FORSCHUNGSDATENZENTREN

Zur Konzeption eines
Taxpayer-Panels für
Deutschland

2006

Herausgeber: Statistische Ämter des Bundes und der Länder
Herstellung: Statistisches Bundesamt

Fachliche Informationen
zu dieser Veröffentlichung:

Statistisches Bundesamt
Gruppe VI D „Steuern“
Tel.: 06 11 / 75 23 85
Fax: 06 11 / 72 40 00
steuern@destatis.de

Erscheinungsfolge: unregelmäßig
Erschienen im November 2006

Informationen zum Datenangebot:

Statistisches Bundesamt
Forschungsdatenzentrum
Tel.: 06 11 / 75 42 20
Fax: 06 11 / 72 39 15
forschungsdatenzentrum@destatis.de

Forschungsdatenzentrum der
Statistischen Landesämter
– Geschäftsstelle –
Tel.: 0211 / 9449 41 47
Fax: 0211 / 9449 40 77
forschungsdatenzentrum@lds.nrw.de

Diese Publikation wird kostenlos als **PDF-Datei** zum Download unter www.forschungsdatenzentrum.de angeboten.

© Statistisches Bundesamt, Wiesbaden 2005
(im Auftrag der Herausbergemeinschaft)

Für nichtgewerbliche Zwecke sind Vervielfältigung und unentgeltliche Verbreitung, auch auszugsweise, mit Quellenangabe gestattet. Die Verbreitung, auch auszugsweise, über elektronische Systeme/Datenträger bedarf der vorherigen Zustimmung. Alle übrigen Rechte bleiben vorbehalten.

Zur Konzeption eines Taxpayer-Panels für Deutschland

Dr. Daniel Vorgrimler / Christopher Gräß / Susan Kriete-Dodds¹

Zusammenfassung

Paneldaten erfreuen sich in der wirtschafts- und sozialwissenschaftlichen Forschung wachsender Beliebtheit. Auch die finanzwissenschaftliche Forschung erhält durch sie neue Impulse. Hier hat sich in der angelsächsischen Literatur mit der „new tax responsiveness literature“ (NTR-literature) sogar eine eigene Forschungsrichtung etabliert, die steuerliche Fragestellungen auf Basis von Paneldaten analysiert. In Deutschland existieren aufgrund mangelnder Daten nur vereinzelt Beiträge zu diesem Forschungszweig. Die nur beschränkt zur Verfügung stehenden Paneldaten sind ursächlich für die geringe Bedeutung, die solche Daten bisher in der wissenschaftlichen Diskussion und in der steuerpolitischen Beratung haben.

Mit den Daten der jährlichen Einkommensteuerstatistik liegen dem Statistischen Bundesamt Querschnittsdaten vor, die über Identifikationsmerkmale (Steuernummer) prinzipiell miteinander verknüpft werden können. Da die Daten im Haus bereits vorliegen wird durch ein Panel kein Auskunftspflichtiger zusätzlich belastet. Angesichts des Mehrwertes an Information, der durch ein Panel entsteht, berücksichtigt dies die Belange der Auskunftspflichtigen im besonderen Maße.

Der Beitrag stellt die bisherigen Arbeiten des Statistischen Bundesamts für ein Panel aus den Daten der jährlichen Einkommensteuerstatistik vor. Im ersten Teil wird auf die Vorteile und Notwendigkeit von Paneldaten im Allgemeinen und einem Panel aus der Einkommensteuerstatistik im Besonderen eingegangen. Im zweiten Teil wird kurz auf die Problematik der Organisation von großen Datenmengen eingegangen. Im Gegensatz zu herkömmlichen Befragungspanels, bei denen ex ante bestimmt wird welche Merkmalsträger wiederholt befragt werden, muss für ein Panel aus der Einkommensteuerstatistik die separat vorliegenden Jahre (Wellen) ex post zu einem Panel zusammengeführt werden. Dieser Fragestellung wird im dritten Teil am Beispiel der Zusammenführung der Jahre 2001 und 2002 nachgegangen. Abschließend werden einige einfache Anwendungsbeispiele von steuerstatistischen Paneldaten aufgezeigt. Ein Ausblick rundet den Beitrag ab.

• **Stichwörter: Einkommensteuerstatistik, Paneldaten, Panelmortalität, Panelanalysen**

¹ Dr. Daniel Vorgrimler, Christopher Gräß, Susan Kriete-Dodds, Statistisches Bundesamt Wiesbaden

1 Warum Paneldaten?

Als Gründe für die wachsende Beliebtheit von Paneldaten finden sich in der Literatur die zusätzlichen Analysemöglichkeiten, die mit der zusätzlichen Dimension Zeit verbunden sind, die neuen Entwicklungen bei den Analysetechniken und nicht zuletzt das verbesserte Angebot an Paneldaten² (so z. B. das SOEP des DIW).

In der angelsächsischen wissenschaftlichen Literatur hat sich mit der „new tax responsiveness literature“ (NTR-literature) eine eigene Forschungsrichtung etabliert, die steuerliche Fragestellungen auf Basis von Paneldaten analysiert.³ Dies zeigt, dass auch für finanzwissenschaftliche Fragestellungen von Paneldaten wichtige Impulse ausgehen können. Ein erster deutscher Beitrag ist hierbei der von Gottfried/Schellhorn im Jahr 2003 erschienene Aufsatz „Die Elastizität des zu versteuernden Einkommens“. Diese Arbeit basiert auf einem Panel für Steuerpflichtige aus Baden-Württemberg mit rund 50.000 Steuerfällen aus den Jahren 1988-1991. Die nur beschränkt zur Verfügung stehenden Paneldaten sind ursächlich für die geringe Bedeutung, die solche Daten bisher in der wissenschaftlichen Diskussion und in der steuerpolitischen Beratung haben.

Aktuelle und repräsentative Paneldaten sind notwendig, wenn deutsche Wissenschaftler in dem Forschungszweig der NTR-literature international konkurrenzfähig sein wollen. Hier kann das Statistische Bundesamt mit der Entwicklung eines Taxpayer-Panels für Deutschland die führende Rolle als Datenanbieter für die deutsche Wissenschaft übernehmen. Ein Panel aus der Einkommensteuerstatistik soll aber nicht nur ein Angebot an die Wissenschaft sein, vielmehr ergeben sich für das Statistische Bundesamt selbst Möglichkeiten die eigenen statistischen Produkte in ihrer Aussagefähigkeit zu verbessern.

Mit den Daten der jährlichen Einkommensteuerstatistik liegen ab dem Veranlagungsjahr 2001 dem Statistischen Bundesamt Querschnittsdaten vor, die über Identifikationsmerkmale (Steuernummer) prinzipiell miteinander verknüpft werden können (siehe Kapitel 3).⁴ Da die Daten dem Statistischen Bundesamt bereits vorliegen, wird durch ein Panel kein Auskunftspflichtiger zusätzlich belastet. Angesichts des Mehrwertes an Information, der durch ein Panel entsteht, berücksichtigt dies die Belange der Auskunftspflichtigen im besonderen Maße.

² Vgl. Hsiao, Cheng, Why Panel Data?, Institute of Economic Policy Research, Working Paper 05.33, www.usc.edu/iepr, 2005.

³ So widmen sich zahlreiche Arbeiten der Verhaltensanalyse der Steuerpflichtigen auf Veränderungen von Steuersätzen und versuchen so die so genannte Laffer-Kurve empirisch herzuleiten, siehe z. B. Goolsbee, A.: What happens when you tax the rich?, 2000.

⁴ Zur jährlichen Einkommensteuerstatistik vgl. Lietmeyer et al.: Jährliche Einkommensteuerstatistik auf Basis der bisherigen Geschäftsstatistik der Finanzverwaltung, WiSta, Heft 7, 2005, S. 671 ff.

2 Die jährliche Einkommensteuerstatistik und Überlegungen zur Datenhaltung

Dem Statistischen Bundesamt wurde vom Bundesministerium der Finanzen die statistische Aufbereitung der Geschäftsstatistik zur Einkommensteuer ab dem Veranlagungsjahr 2001 übertragen (jährliche Einkommensteuerstatistik). Dieses Datenmaterial umfasst jährlich ca. 28 Millionen Einkommensteuerveranlagungen und ab 2002 die Anträge zur Förderung der zusätzlichen Altersvorsorge (Riester Rente). Durch die Aufbereitung der Daten durch das Statistische Bundesamt wird es zukünftig möglich jährliche Auswertungen zur Struktur und Wirkungsweise der Einkommensteuer, zu deren wirtschaftlicher und sozialer Bedeutung und insbesondere zur Einkommens- und Steuerlastverteilung durchzuführen. Das umfangreiche Material wird sowohl die Politikberatung als auch die wissenschaftliche Forschung in vielen Bereichen erheblich verbessern. Die jährliche Einkommensteuerstatistik wird in der nächsten Zeit kontinuierlich allen Nutzergruppen zugänglich gemacht. Die jährliche Einkommensteuerstatistik ist nicht zu verwechseln mit der dezentralen dreijährlichen Bundesstatistik zur Lohn- und Einkommensteuer, die zuletzt für das Veranlagungsjahr 2001 aufbereitet wurde.⁵

Zur Erstellung des Testpanels lagen dem Statistischen Bundesamt im Rahmen der jährlichen Einkommensteuerstatistik ein komplettes Jahr 2001 und ein fast vollständiges Jahr 2002 vor (ca. 95 % der Daten waren vorhanden). Diese beiden Jahre stehen in diesem Beitrag daher im Mittelpunkt. Für das Jahr 2003 standen nur erste Daten zur Verfügung, sodass diese für bestimmte Fragestellungen mit in die Betrachtung eingeschlossen werden konnten. Seit Erstellung des hier vorgestellten Testpanels sind die restlichen Daten für 2002 sowie über 95% der Daten für 2003 eingegangen. Des Weiteren stehen erste Daten für 2004 zur Verfügung, so dass die in diesem Beitrag beschriebenen Ergebnisse aktualisiert und die Betrachtung auf mindestens drei Jahre ausgedehnt werden können⁶.

Wie erwähnt besteht die jährliche Einkommensteuerstatistik aus knapp 28 Mill. Steuerpflichtigen mit zurzeit rund 1.000 Merkmalen pro Jahr. Ein Jahr beansprucht knapp 70 GB Speicherplatz. Die Aufbereitung der Daten erfolgt unter Einsatz der SAS-Software. Die maschinelle Infrastruktur ist mit dem SAS-Serversystem gegeben. Trotz der gut infrastrukturellen Ausstattung stellt sich aufgrund der Datenmenge die Frage nach der effizienten Haltung der Paneldaten. Hierbei sind drei mögliche Arten der Datenhaltung zu diskutieren:

- *Vollversion*: Verknüpfung aller Merkmalsträger mit jeweils allen Merkmalen für jedes Paneljahr,
- *Stichprobenversion*: Verknüpfung eines bestimmten Prozentsatzes der Merkmalsträger mit allen Merkmalen für jedes Paneljahr,

⁵ zur Abgrenzung der jährlichen zur dreijährlichen Statistik vgl. Lietmeyer et al.: Jährliche Einkommensteuerstatistik auf Basis der bisherigen Geschäftsstatistik der Finanzverwaltung, WiSta, Heft 7, 2005, S. 671 ff.

⁶ Die endgültigen Ergebnisse für das Panel 2001/2002 werden in der Reihe Wirtschaft und Statistik veröffentlicht.

- *Masterdateiversion*: Verknüpfung aller Merkmalsträger mit minimaler Merkmalsanzahl in einer Masterdatei für jedes Paneljahr mit Zugriff auf die restlichen Merkmale in den bereits vorhandenen Grunddateien (Querschnittsdaten).

Bei einer *Vollversion* liegt ein Komplettmaterial ständig zur Verfügung. Nachteil ist allerdings, dass es zu einer doppelten Datenhaltung mit extremen Platzanforderungen kommt und dass die Laufzeiten aufgrund der entstehenden Breite des Datensatzes (Anzahl der Merkmale) sehr hoch werden. Des Weiteren wird das Material mit jedem Jahr schwieriger zu verwalten. So müssten nach fünf verknüpften Jahren ungefähr 4.500 Merkmale verarbeitet werden.

Der Vorteil beim *Stichprobenmodell* liegt in den kürzeren Laufzeiten. Dem stehen aber die Nachteile eines möglichen Stichprobenfehlers und eines zusätzlichen Mortalitätsproblems entgegen. Die Stichprobe muss über die Jahre hinweg repräsentativ bleiben und Personen, die aus der Stichprobe fallen, müssen durch gleichartige ersetzt werden. Darüber hinaus ist das Problem mit der Breite des entstehenden Panels auch im Rahmen dieses Modells nicht gelöst.

Beim Modell der *Masterdatei* werden die Probleme mit der Handhabbarkeit und der Kapazität am besten gelöst. In der Masterdatei stehen nur die über die Zeit konstanten oder relativ konstanten Merkmale wie Steuernummer⁷ und amtlicher Gemeindeschlüssel. Hierdurch bleibt die Größe der Datei konstant. Mit dieser Datei greift man fallbezogen auf die einzelnen Jahre und die für die Auswertung notwendigen Merkmale in den Grunddateien zu. Diese Grunddateien entsprechen den Dateien, die auch im normalen Aufbereitungsprozess vorhanden sind. Doppelte Datenhaltung wird dadurch fast komplett verhindert. Der Nachteil besteht jedoch darin, dass bei jeder Auswertung das Panel im Prinzip neu zusammengestellt werden muss, was z. T. sehr laufzeitintensiv sein wird. Am Ende muss, um unnötige Datenhaltung zu verhindern, die analysierte Datei wieder gelöscht werden, was die Reproduzierbarkeit der Auswertungen erschwert. Abgeschwächt kann dieses Problem dadurch werden, dass einzelne besonders wichtige und damit oft zu analysierende Merkmale – wie zum Beispiel die festgesetzte Einkommensteuer eines Jahres – mit in die Masterdatei aufgenommen werden. Auswertungen mit einem solchen Panel sind auch mit allen 28 Mill. Steuerpflichtigen mit geringen Laufzeiten möglich, solange in der Masterdatei nicht zu viele Merkmale enthalten sind.

Trotz der genannten Nachteile scheint mit den bestehenden technischen Ressourcen das dritte Modell am Besten geeignet zu sein. Sinnvoll dürfte es sein, wichtige Merkmale wie die Einkunftsarten, das zu versteuernde Einkommen und die festgesetzte Einkommensteuer für jedes Jahr in der Masterdatei zu integrieren. Die Masterdatei wächst dadurch im Zeitablauf nur beschränkt, aber sie ist für einen Großteil der in Frage kommenden Auswertungen bereits ausreichend. Ein weiterer Vorteil dieses Modells ist, dass die

⁷ Die Steuernummer liegt dem Statistischen Bundesamt für Plausibilitätsprüfungen vor und wird nicht veröffentlicht.

anderen beiden Modelle jederzeit aus dieser Variante produzierbar sind. So kann z. B. eine Stichprobe für die Weitergabe der Daten an die Wissenschaft erstellt werden.

3 Zur Verknüpfung von Merkmalsträgern über die Zeit

3.1 Verknüpfung mit Hilfe der Steuernummer

Die Steuernummer ist auf Ebene der Bundesländer ein eindeutiger Identifikator für die Steuerpflichtigen, der sich nur zu bestimmten Anlässen ändert. Steuerpflichtige können daher prinzipiell über die Zeit verfolgt werden. Problematisch ist ein Steuernummerwechsel eines Steuerpflichtigen. Die Finanzverwaltung sieht hierfür vor, die alte Steuernummer separat mitzuliefern. Aus verschiedenen Gründen ist dies aber oftmals nicht möglich (so z. B. wenn ein Steuerpflichtiger über die Bundeslandgrenzen hinweg umzieht und das neu zuständige Finanzamt eine Steuernummer vergibt ohne die alte zu kennen) und wird von den Finanzverwaltungen der Länder auch unterschiedlich umgesetzt. Daher wird das Problem des Steuernummerwechsels von der Finanzverwaltung nur unzulänglich gelöst. Eine erste Zusammenführung der Veranlagungsjahre 2001 mit 2002 ergab ein Testpanel von 23,790 Mill. Steuerpflichtigen. Hiervon konnten lediglich 45.000 Steuerpflichtige mit Hilfe der im Veranlagungsjahr 2002 als „alt“ geführten Steuernummer zusammengeführt werden. Tabelle 1 enthält die Anzahl an Steuerpflichtigen bei den verknüpften Daten und die Anzahl der jeweiligen verbliebenen Merkmalsträger. Dabei ist zu beachten, dass für das Veranlagungsjahr 2002 und insbesondere für das Jahr 2003 noch weitere Daten geliefert werden, sodass erst danach ein korrekter Überblick über die nicht verknüpften Daten möglich sein wird.

Tabelle 1: Anzahl verknüpfter bzw. nicht verknüpfter Merkmalsträger im Testpanel 2001/2002

Datei	Anzahl Merkmalsträger	Nicht verknüpft in %
Panel 01/02	23.790.324	
Panel 01/02/03	13.444.950	
Panel 01/03	182.912	
Rest 01	3.919.057	14,1
Rest 02	2.713.979	10,2
Rest 03	3.191.671	19,1

Das Zweijahrespanel 2001/02 enthält über 85 % der Steuerpflichtigen der jeweiligen Jahre. Der überwiegende Teil der Steuerpflichtigen kann demnach ohne Schwierigkeiten zu einem Panel verbunden werden. Allerdings bedeutet eine Ausfallrate von jährlich 15 %, dass bei einem Panel mit vier Wellen bei nur noch etwas mehr als 61 % der Steuerpflichtigen die Verknüpfung über den gesamten Zeitraum gelingt. Neben diesem Problem einer schleichenden Erosion der Datenbasis, ist der Ausfall von Merkmalsträgern bzw. die Panelmortalität dann problematisch, wenn hierdurch Verzerrungen in den statistischen Ergebnissen entstehen. In Tabelle 2 sind verschiedene deskriptive Maße für den Gesamtbetrag der Einkünfte für die Jahre 2001 und 2002 enthalten, berechnet aus dem Gesamtmaterial der jeweiligen Jahre und aus dem Zweijah-

respanel. Der durchschnittliche Gesamtbetrag der Einkünfte liegt für beide Jahre aus dem Panel errechnet signifikant höher als aus dem Gesamtmaterial der einzelnen Jahre. Es ist daher nicht nur zum Erhalt der Datengrundlage notwendig sich mit den nicht verknüpften Steuerpflichtigen näher zu beschäftigen, sondern auch um die Repräsentativität des Panels zu wahren.

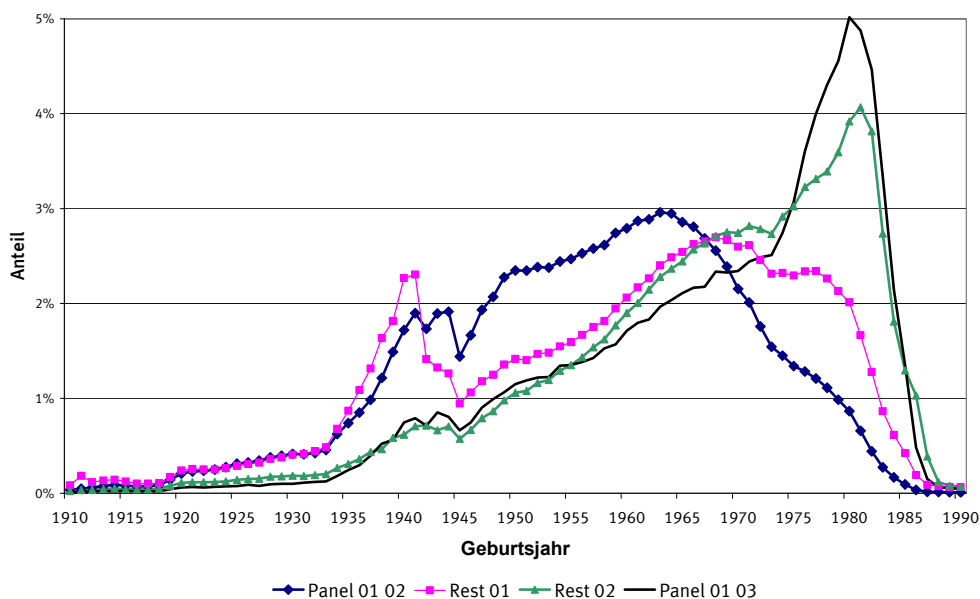
Tabelle 2: Statistiken des Testpanels im Vergleich mit der Grundgesamtheit (Werte für den Gesamtbetrag der Einkünfte)

Jahr	Variationskoeffizient	Konfidenzintervall		Mittelwert	N	Median	Schiefe
		untere 99 % Grenze	obere 99 % Grenze				
Panel 2001	301	36.232	36.347	36.289	23.790.324	28.026	300
2001	336	34.535	34.649	34.592	27.709.160	26.194	387
Panel 2002	295	35.811	35.923	35.867	23.790.324	28.194	454
2002	313	34.230	34.338	34.284	26.504.230	26.792	553

3.2 Mögliche Gründe für Panelmortalität und Verzerrungen in den Panelergebnissen

Ein Steuerpflichtiger kann entweder aufgrund eines fehlerhaften Identifikators oder eines tatsächlichen Ausscheidens aus der bzw. Eintretens in die Population nicht über die Zeit zu verknüpfen sein. Während es gilt die erste Art der Mortalität zu minimieren, ist die zweite Art nicht nur nicht zu verhindern, sondern es kann für bestimmte Fragestellungen sogar erwünscht sein, diese Mortalität richtig zu identifizieren. Da beide Fälle gleichzeitig auftreten, gilt es im ersten Schritt die beiden Mortalitätsarten bestmöglich voneinander abzugrenzen. Die Abbildung 1 zeigt die Verteilung der Geburtsjahre bei verknüpften Merkmalsträgern (Panel 01/02) im Vergleich zu den Merkmalsträgern, die nicht verknüpft werden konnten. Merkmalsträger aus dem Veranlagungsjahr 2001 können u. a. dann nicht verknüpft werden, wenn sie 2002 keine Einkommensteuererklärung abgegeben haben. Gründe hierfür wären das Erreichen des Rentenalters oder der Tod. Dagegen können die restlichen Merkmalsträger aus dem Veranlagungsjahr 2002 erstmalig eine Einkommensteuererklärung abgegeben haben und daher nicht mit 2001 zu verknüpfen sein. Während der erste Fall hauptsächlich auf ältere Steuerpflichtige zutrifft, trifft der zweite tendenziell bei jüngeren Steuerpflichtigen zu. Daher ist bei der Verteilung der Geburtsjahre für die restlichen Merkmalsträger des Jahres 2001 eine Verschiebung zu den älteren Geburtsjahrgängen zu erwarten und für die des Jahres 2002 zu den jüngeren. Die Abbildung bestätigt diese Vermutung. Für die restlichen Merkmalsträger aus 2001 ergibt sich eine starke Ausprägung bei den Jahrgängen die typischerweise in Rente gegangen sind. Darüber hinaus sind die Jahrgänge vor 1935 überproportional vertreten. Dies kommt aufgrund des Maßstabes der Abbildung nicht zur Geltung. Bei den restlichen Merkmalsträgern aus 2002 sind dagegen die jüngeren Jahrgänge überproportional vertreten.

Abbildung 1: Verteilung der Geburtsjahre bei den verknüpften Merkmalsträgern und den nicht verknüpften Merkmalsträgern im Vergleich



Steuerpflichtige, die z. B. aufgrund unregelmäßiger Erwerbsverläufe nicht für jedes Jahr eine Einkommensteuererklärung abgeben, können nur für diese Jahre verknüpft werden, in denen sie zur Einkommensteuer veranlagt werden. Dieses Problem kann zumindest teilweise für diejenigen Jahre aufgefangen werden, die innerhalb der zeitlichen Grenzen des Panels liegen. So z. B. für das Jahr 2002, wenn ein Panel bestehend aus den Jahren 2001 bis 2003 betrachtet wird (siehe Kapitel 3.3).

Steuerpflichtige können aufgrund einer Heirat als eigenständiger Merkmalsträger aus der Population ausscheiden. Bei einer Heirat erhält i. d. R. die Frau die Steuernummer des Mannes. (Eine Ausnahme ist hier Baden-Württemberg, das nach Datenlage bei einer Heirat i. d. R. eine neue Steuernummer vergibt). Eine Heirat im Jahr t ist im Panel dadurch zu erkennen, dass der Steuerpflichtige im Jahr $(t-1)$ nach der Grundtabelle und im Jahre t nach der Splittingtabelle veranlagt wird. Darüber hinaus besteht im Jahr $(t-1)$ der Steuerpflichtige noch aus einem Steuerfall und im Jahr t aus dem alten Steuerfall und einem neu hinzugekommenen. Im Jahr 2002 konnten im Panel rund 260.000 Steuerpflichtige identifiziert werden, die diesem Muster entsprechen und von denen auszugehen ist, dass sie geheiratet haben (das Statistische Jahrbuch weist 390.000 Eheschließungen für das Jahr 2002 aus). Eheschließungen sind demnach der Grund für die Panelmortalität bei über 300.000 Merkmalsträgern aus dem Jahr 2001, da bei Zusammenveranlagung die „Frau“ nicht als eingeständiger Steuerpflichtiger weitergeführt wird.

Genau umgekehrt verhält es sich bei einer Ehescheidung. Hier findet ein Wechsel in der Veranlagungsart von der Splittingtabelle in $(t-1)$ zur Grundtabelle in t . Existiert in $(t-1)$ ein Steuerpflichtiger mit zwei Steuerfällen, so werden daraus in t zwei Steuerpflichtige mit jeweils einem Steuerfall. Der Mann behält i. d. R.

seine Steuernummer und die Frau kommt in t als neuer Steuerpflichtiger hinzu. Die Frau kann als neuer Grundtabellenfall in t nicht mit einem Datensatz aus (t-1) verknüpft werden. Im Panel 2001/02 konnten rund 160.000 Ehescheidungen identifiziert werden. Insgesamt gab es 2002 rund 200.000 Ehescheidungen.⁸ Ehescheidungen sind nach diesen Ergebnissen der Grund für die Panelmortalität bei über 160.000 Merkmalsträgern aus dem Jahr 2002.

Die ersten Überlegungen zur Panelmortalität zeigen bereits, dass hiervon spezielle Gruppen überdurchschnittlich betroffen sind. So fallen ältere Steuerpflichtige aufgrund von Tod oder Verrentung aus der Einkommensteuerveranlagung raus und kommen jüngere Steuerpflichtige aufgrund des Eintritts in das Erwerbsleben in die Einkommensteuerstatistik hinzu. Da es in beiden Fällen zu keiner Aufnahme in das Panel kommt und es sich bei diesen Gruppen um Steuerpflichtige mit tendenziell geringeren Einkünften handelt, führt dies zu höheren Einkommensdaten der Steuerpflichtigen die im Panel vertreten sind gegenüber allen Steuerpflichtigen.

3.3. Alternative Verknüpfungsstrategien

Im Folgenden werden alternative Verknüpfungsstrategien aufgezeigt, mit denen die Panelmortalität begrenzt werden soll. Wie Tabelle 1 bereits gezeigt hat, verblieben aus dem Veranlagungsjahr 2001 3,9 Mill. Steuerpflichtige und aus dem Veranlagungsjahr 2002 2,7 Mill. Steuerpflichtige, die nicht in das Panel aufgenommen werden konnten. Ziel der folgenden Ausführungen ist es, diese Steuerpflichtigen entweder auf einen anderen Weg über die Zeit zu verknüpfen oder solche zu identifizieren, die entweder tatsächlich aus der Besteuerung und damit aus dem Panel ausscheiden oder die Besteuerung neu aufgenommen haben und entsprechend neu in das Panel zu integrieren sind.

- **Zeitweiliges Ausscheiden aus dem Panel**

Unregelmäßige Erwerbsverläufe drücken sich durch eine unregelmäßige Abgabe der Einkommensteuererklärung aus. So können Merkmalsträger aus dem Jahre 2001 deswegen nicht mit einem 2002 Datensatz verknüpft worden sein, weil in diesem Jahr aufgrund des zu geringen Einkommens keine Erklärung abgegeben wurde. Um dies zu testen, wurden die verbliebenen Merkmalsträger des Jahres 2001 mit den bereits vorhandenen Merkmalsträgern aus dem Jahr 2003 zusammengeführt. Hierdurch konnten über 180.000 Steuerpflichtige identifiziert werden, die im Jahr 2002 keine, aber in den Jahren 2001 und 2003 eine Einkommensteuererklärung abgegeben haben. Die Steuerpflichtigen haben demnach ein Jahr mit der Veranlagung ausgesetzt. Die Struktur der Geburtsjahre zeigt eine starke Konzentration bei den jungen Jahrgängen (vgl. Abbildung 1). Unstetige Erwerbsverläufe dürften überproportional bei jüngeren Steuerpflichtigen in der Ausbildung bzw. im Studium auftreten, was eine Erklärung für die konzentrierte Verteilung wäre. Eine weitere Analyse der Daten zeigt, dass es sich zu 75 % um Grundtabellenfälle handelt (im

⁸ Vgl. Emmerling, D.: Ehescheidungen 2003, in WiSta, Heft 2, 2005, Seite 97-108.

Gesamtmaterial liegt dieser Anteil bei knapp 50 %), deren Median beim Gesamtbetrag der Einkünfte in beiden Jahren unter 10.000 € liegt. Auch dies bestätigt die Vermutung, dass es sich zu einem Großteil um Studenten und Auszubildende handelt. Die geringen Einkünfte und deren Ausfall aus dem Panel führen schließlich zu nach oben verzerrten Ergebnissen im Panel.

- **Eheschließungen**

Durch Heirat bekommt die Frau die Steuernummer des Mannes und wird mit ihm gemeinsam als Splittingfall veranlagt.⁹ Aus den restlichen Daten von 2001 werden alle weiblichen Fälle der Grundtabelle separiert (1,3 Mill. Steuerpflichtige). Aus dem Panel werden alle Fälle gezogen, die in 2002 nach der Splitting- und in 2001 nach der Grundtabelle veranlagt wurden (260.000 Steuerpflichtige). Die beiden auf diese Weise gewonnenen Datenbestände werden mit Hilfe indirekter Identifikatoren zusammengeführt. Diese sind:

- Geburtsdatum der Frau,
- Religion der Frau,
- Wohnort.

Konkret sollen mit dieser Vorgehensweise diejenigen identifiziert werden, die 2001 als eigenständige Steuerpflichtige existierten und 2002 „nur“ noch als Steuerfall in den Daten enthalten sind. Hiermit wird ein weiterer Teil der Panelmortalität erklärt.

Aufgrund der Bedingungen werden aus den beiden Datenbeständen diejenigen Sätze als identisch zusammengeführt, die innerhalb des gleichen Ortes wohnen, am gleichen Tag geboren sind und der gleichen Religion angehören. Implizit setzt dies voraus, dass nach oder aufgrund einer Heirat kein Umzug in einen anderen Ort stattgefunden hat und es zu keinem Religionswechsel gekommen ist. Diese Bedingungen werden in der Regel erfüllt sein, aber sicher nicht in allen Fällen. Treten bei der Zusammenführung Doppelfälle auf, so werden diese nicht betrachtet.¹⁰

Auf die beschriebene Weise konnten 130.000 Steuerpflichtige identifiziert werden, die im Jahr 2002 „nur“ noch als Steuerfall in den Daten enthalten sind und 2001 als eigenständige Steuerpflichtige geführt wurden. Mit anderen Worten kann für diese Anzahl an Steuerpflichtigen die Panelmortalität durch Heirat erklärt werden. Doppelfälle waren lediglich knapp 10.000 zu verzeichnen. Da es sich nicht um Zusammenführungen über eindeutige Identifikatoren, sondern über indirekte Identifikatoren handelt, stellt sich die Frage nach der Qualität der Zusammenführungen. Von einer hohen Qualität in der Summe kann ausgegangen werden, wenn bestimmte statistische Maße den Erwartungen entsprechen. So ist die Korrelation

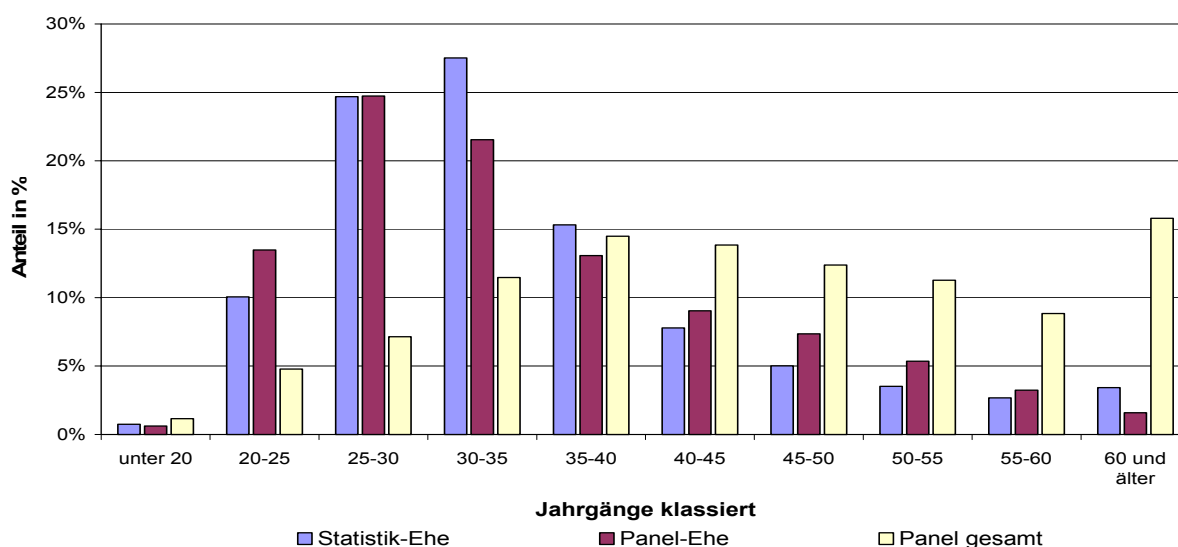
⁹ Der umgekehrte Fall, dass der Mann die Steuernummer der Frau übernimmt kommt nur sehr selten vor, führt nur zu wenigen Zusammenführungen und wird daher nicht explizit betrachtet.

¹⁰ Derzeit werden doppelte Zuordnungen als fehlerhafte Zuordnungen behandelt und entsprechend gelöscht. Weitere Untersuchungen müssen zeigen, inwieweit die doppelten Zuordnungen noch in richtige und falsche unterschieden werden können.

zwischen den Einkünften aus nichtselbständiger Arbeit aus verschiedenen Jahren erfahrungsgemäß sehr hoch (gemessen nach dem Spearman-Koeffizienten liegt diese bei 0,9). Bei den 130.000 frisch verheirateten Steuerpflichtigen wird eine Korrelation von 0,71 nach Spearman erreicht, was ebenfalls sehr hoch ist und als Indikator für die Qualität der Zusammenführung angesehen werden kann. Ein weiterer Indikator ist die Altersstruktur. Diese sollte zu den jüngeren Jahrgängen hin verschoben sein, da das durchschnittliche Heiratsalter geringer ist als das Durchschnittsalter.

Abbildung 2 enthält die Verteilung der klassierten Jahrgänge, die sich aus dem Statistischen Jahrbuch für die heiratenden Personen insgesamt ergibt und die Verteilung der Jahrgänge, wie sie sich aus dem Panel ergibt. Diese werden mit der Altersverteilung für alle Steuerpflichtige verglichen. Die Abbildung lässt erkennen, dass die Altersstruktur derjenigen die Heiraten sich tatsächlich deutlich von der Gesamtheit unterscheidet. Dies gilt sowohl für die Verteilung aus dem Statistischen Jahrbuch als auch für die Verteilung aus dem Panel. Die Panelverteilung gibt dagegen gut die Verteilung des Jahrbuchs wieder, was wiederum auf eine hohe Qualität der Zuordnungen schließen lässt.

Abbildung 2: Verteilung der klassierten Jahrgänge bei Heirat und insgesamt



Die knapp 130.000 aufgrund von Heirat ausgefallenen Steuerpflichtigen zeichnen sich dadurch aus, dass sie zum einen jünger sind und zum anderen, dass es sich i. d. R. um Grundtabellenfälle handelt. Von daher ist es wenig verwunderlich, dass das durchschnittliche Einkommen mit 23.000 € deutlich unter dem Gesamtdurchschnitt mit 36.000 € liegt. Auch diese Gruppe ist demnach mit verantwortlich dafür, dass die durchschnittlichen Einkünfte im Panel nach oben verzerrt sind.

- **Ehescheidungen**

Die Vorgehensweise bei der Identifikation von Merkmalsträgern, die aufgrund von Ehescheidungen im Jahr 2002 erstmalig auftauchen, entspricht dem spiegelbildlichen Vorgehen bei der Eheschließung. Durch eine Scheidung erhält die Frau eine neue Steuernummer und wird nach der Grundtabelle veranlagt.¹¹ Nun werden aus den restlichen Daten von 2002 alle weiblichen Fälle der Grundtabelle separiert (ca. 1 Mill. Steuerpflichtige). Aus dem Panel werden alle Fälle gezogen, die in 2001 nach der Splitting- und in 2002 nach der Grundtabelle veranlagt wurden (ca. 160.000 Steuerpflichtige). Diese beiden Datenbestände werden mit Hilfe indirekter Identifikatoren zusammengeführt. Sie entsprechen den bei den Eheschließungen verwendeten Identifikatoren:

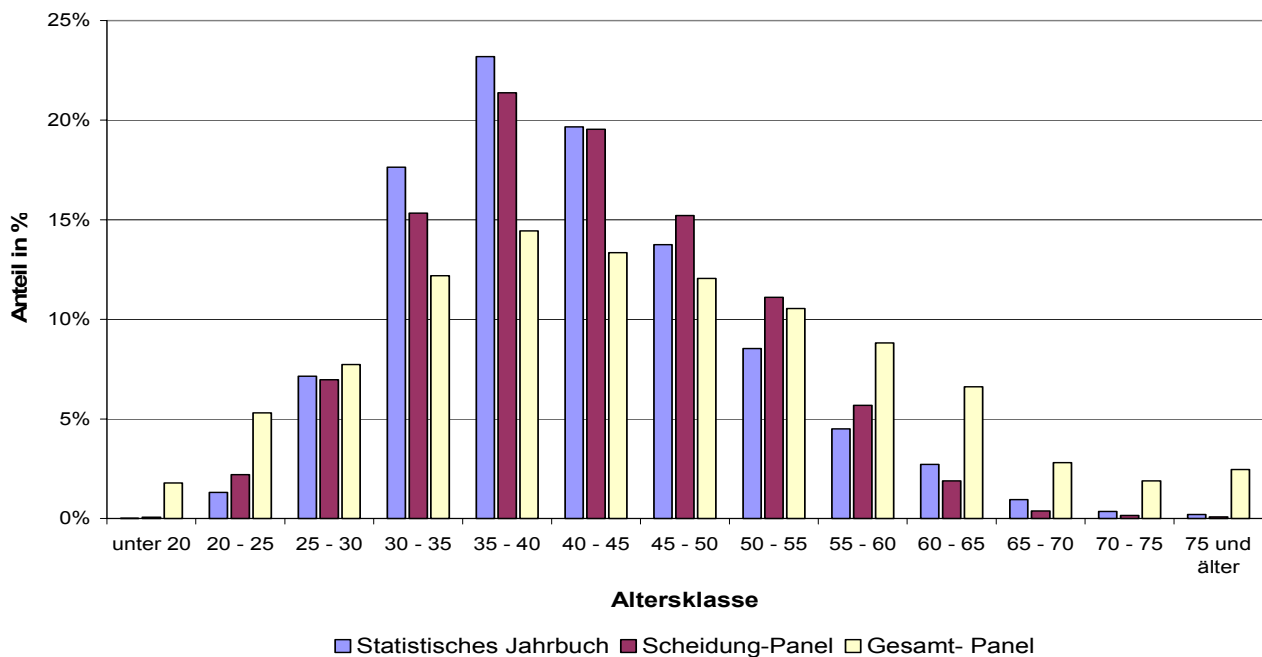
- Geburtsdatum der Frau,
- Religion der Frau,
- Wohnort.

Es gelten daher bezüglich des Wohnorts und der Religion die gleichen Bedingungen wie im Heiratsfall. Diese Bedingungen dürften in diesem Fall seltener gelten, da ein Umzug nach einer Scheidung wahrscheinlicher als nach einer Heirat ist.

Auf die beschriebene Weise konnten von den Steuerpflichtigen aus dem Jahr 2002 rund 42.000 identifiziert werden, die im Jahr 2001 „nur“ als Steuerfall in den Daten enthalten sind und 2002 als eigenständiger Steuerpflichtiger neu hinzukommen. Bei diesen ist demnach eine Scheidung der Grund, warum sie nicht mit einem Datensatz aus 2001 verknüpft werden konnten. Auch in diesem Fall spielten doppelte Zuordnungen mit lediglich knapp 1.700 Fällen nur eine untergeordnete Rolle. Die Frage nach der Qualität der Zusammenführungen stellt sich auch bei dieser Art der Zusammenführung. Der Korrelationskoeffizient nach Spearman zwischen den Einkünften aus nichtselbständiger Arbeit aus den beiden Jahren ist mit 0,81 sehr hoch. Die Altersstruktur wurde wiederum mit der Grundgesamtheit und mit der Altersstruktur verglichen, die sich aus den Angaben des Statistischen Jahrbuchs für 2002 ergeben und ist in Abbildung 3 abgetragen. Auch in diesem Fall spiegelt die Altersstruktur im Panel gut die tatsächliche Struktur wider.

¹¹ Auch hier gilt, dass der umgekehrte Fall nur sehr selten vorkommt und nur zu wenigen Zusammenführungen führt und daher hier nicht explizit betrachtet wird.

Abbildung 3: Verteilung der klassierten Jahrgänge bei Scheidung und insgesamt



Unter bestimmten Bedingungen können Unterhaltsleistungen an den geschiedenen Ehepartner als Sonderausgaben abgezogen werden. Hierzu muss u. a. der Empfänger der Unterhaltsleistung diese als sonstige Einkünfte bei der Steuererklärung angeben. Diese Regelung müsste sich bei den zusammengeführten Daten dadurch wieder finden, dass die Sonderausgaben mit den sonstigen Einkünften beim ehemaligen Ehepartner übereinstimmen. Allerdings nutzen nur sehr wenige diese Möglichkeit, so dass die Kontrollgruppe mit rund 2.200 Steuerpflichtigen auch sehr klein ausfällt. Von diesen Steuerpflichtigen haben über 1.900 Übereinstimmungen in den beiden Merkmalen, was wiederum für eine hohe Qualität der Zusammenführung spricht.

Was die Einkommensstrukturen angeht gilt hier ähnliches wie beim Heiratsfall. Die neu geschiedenen Steuerpflichtigen sind zwar älter wie die frisch vermählten, aber immer noch jünger als der durchschnittliche Steuerpflichtige. Darüber hinaus handelt es sich wiederum um Grundtabellenfälle, die durch die Scheidung im Panel verloren gehen. Die Einkünfte dieser Gruppe sind entsprechend mit durchschnittlich 26.700 € geringer als der Gesamtdurchschnitt, wodurch ebenfalls die Ergebnisse des Panels nach oben verzerrt werden.

- **Zusammenführung der Splittingtabellenfälle mit Hilfe indirekter Identifikatoren**

Die Zusammenführung der Splittingtabellenfälle stellt geringere Ansprüche als die der Grundtabellenfälle. Daher werden beide Fälle getrennt behandelt. Wie bereits erwähnt, können Steuerpflichtige dann nicht ohne weiteres über die Zeit beobachtet werden, wenn sich ihre Steuernummer ändert. Dies ist besonders

bei einem Wohnortwechsel der Fall. Daher ist zu erwarten, dass viele der verbliebenen Datensätze der beiden Jahre daher nicht zusammengeführt wurden, weil ein Umzug der Steuerpflichtigen vorliegt. Der Wohnort als indirekter Identifikator scheidet demnach aus.¹² Stattdessen werden bei den Splittingfällen die Geburtstage und die Religion der Ehepartner als indirekte Identifikatoren verwendet. Die Wahrscheinlichkeit, dass bei zwei Ehepaare die auf anderem Wege bisher noch nicht im Panel enthalten sind, die Geburtstage der jeweiligen Männer und Frauen übereinstimmen ist bereits so gering, dass bei einer Übereinstimmung dieser Merkmalsausprägungen in den beiden restlichen Datenbeständen davon ausgegangen werden kann, dass es sich bei dem Steuerpflichtigen der beiden Jahre um denselben handelt.¹³ Das verwendete Merkmal der Religionszugehörigkeit dient als zusätzliche Sicherheit, welches letztlich dafür sorgt, dass die Kombination der indirekten Identifikatoren fast wie ein eindeutiger Identifikator wirkt. Die Qualität der Zuordnung kann wiederum über den Korrelationsvergleich – diesmal für die Einkünfte des Mannes und der Frau – getestet werden. Zusätzlich können die Geburtstage der Kinder abgeglichen werden.

Es konnten über 130.000 Zuordnungen auf diesem Weg gefunden werden. Die Korrelationen bei den Einkünften aus nichtselbständiger Arbeit sind mit 0,78 (Mann) bzw. 0,82 (Frau) hoch. Bei lediglich 7.500 Kindern mussten unterschiedliche Geburtstage festgestellt werden. Auch dies zeigt die hohe Qualität der Zuordnung zumal es nicht bedeutet, dass es sich bei den 7.500 Fällen automatisch um falsche Zuordnungen handelt. Oftmals handelt es sich lediglich um Tippfehler oder es haben sich Änderungen bei den zu berücksichtigenden Kindern ergeben. Die Einkommen der zusammengeführten Steuerpflichtigen sind nicht signifikant verschieden von den Einkünften aller nach der Splittingtabelle veranlagten Steuerpflichtigen. Rund 50.000 der 130.000 Steuerpflichtigen haben ihren Wohnort gewechselt, was bei diesen der Grund für die neue Steuernummer sein dürfte.

- **Zusammenführung der Grundtabellenfälle mit Hilfe indirekter Identifikatoren**

Wesentlich schwieriger gestaltet es sich, die nach der Grundtabelle versteuerten Steuerpflichtigen der beiden Jahre zusammen zu führen. Grundtabellenfälle sind aus verschiedenen Gründen stärker von der Panelmortalität betroffen als Steuerpflichtige, die nach der Splittingtabelle besteuert werden.¹⁴ Daher ist die Anzahl der über indirekte Identifikatoren zu verknüpfenden Steuerpflichtigen in diesem Fall höher (in den restlichen Dateien der Jahre 2001/02 sind über 2 Mill. Grundtabellenfälle und nur ca. 1 Mill. Splitting-

¹² Von denjenigen Steuerpflichtigen die mit Hilfe der alten Steuernummer zusammengeführt wurden, die also eine neue Steuernummer erhielten und dies verfolgt werden kann, haben gut 10 % ihren Wohnsitz gewechselt. Wird daher der Wohnort als indirekter Identifikator verwendet, würde man entsprechend viele eigentlich identische Steuerpflichtige nicht als solche erkennen.

¹³ Wird für das Alter der Ehepaare eine Zeitspanne von 30 Jahren angenommen, dann gibt es bereits knapp 120 Mill. unterschiedliche Kombinationsmöglichkeiten. In den beiden Datenbeständen sind dagegen nur ca. 3 Mill. Steuerpflichtige und davon ca. 1 Mill. Steuerpflichtige enthalten, die nach der Splittingtabelle veranlagt werden und daher in Frage kommen.

¹⁴ So scheiden aufgrund eines Todesfalls überproportional Grundtabellenfälle aus, da beim Tod eines Ehepartners der andere als Grundtabellenfall veranlagt wird und entsprechend weiterhin im Panel vertreten sein kann. Einen ähnlichen Zusammenhang ergibt sich bei der Verrentung. Auch ist – wie bereits gesehen – eine unregelmäßige Abgabe der Einkommensteuererklärung aufgrund eines unregelmäßigen Erwerbsverlaufs eher bei Grundtabellenfällen als bei Splittingfällen zu beobachten.

fälle enthalten). Dies ist ein Grund für die größeren Schwierigkeiten, aber bei weitem nicht der wichtigste. Am wichtigsten ist, dass nur noch ein Geburtsdatum je Datensatz vorhanden ist. Die Wahrscheinlichkeit, dass mindestens zwei einzelne Personen am gleichen Tag Geburtstag haben ist um ein vielfaches größer, als dass zwei Ehepaare exakt die gleichen Geburtsdaten aufweisen. So sind von den ca. 36.000 unterschiedlichen Geburtstagen der 2,5 Mill. Steuerpflichtigen aus 2001 die bisher noch nicht zusammengeführt wurden lediglich 1.300 einmalig. Aus diesem Grund müssen weitere indirekte Identifikatoren zur Zusammenführung verwendet werden.

Ebenso wie bei den zusammen veranlagten Steuerpflichtigen (Splittingfälle) liegt ein Grund für eine neue Steuernummer im Wohnortswechsel. Daher sollte auf den Wohnort als indirekter Identifikator ebenfalls verzichtet werden. Leider konnte auf diese Weise trotz des Einbeziehens mehrerer zusätzlicher indirekter Identifikatoren kein zufrieden stellendes Ergebnis erzielt werden. Aufgrund dieser Problematik wurde ein zweistufiges Verfahren gewählt. Auf der ersten Stufe wurde kein Wohnortswechsel zugelassen, während in der zweiten Stufe ein Wechsel des Wohnsitzes innerhalb eines Bundeslandes zugelassen wurde. Diese Vorgehensweise weist einige Mängel auf,¹⁵ ist jedoch nach dem derzeitigen Stand der Arbeiten (leider) ohne Alternative.

Inklusive Wohnort wurden auf der ersten Stufe elf indirekte Identifikatoren eingesetzt. Im Einzelnen sind dies:

- Wohnort,
- Religion,
- Geschlecht,
- Gewerbekeznzahl,
- Soziale Gliederung: Nicht rentenversicherungspflichtige Arbeitnehmer (gekürzte Vorsorgepauschale),
 Rentenversicherungspflichtige Arbeitnehmer (ungekürzte Vorsorgepauschale),
 Kein Bruttolohn vorhanden,
- Geburtsdatum der ersten drei Kinder,
- Einkünfte aus Land und Forstwirtschaft (ja/nein),
- Einkünfte aus Gewerbebetrieb (ja/nein),

¹⁵ So wären z.B. beim Splittingmatch von den 130.000 Zuordnungen 50.000 nicht gefunden worden, weil bei diesen der AGS nicht übereinstimmt. Solche Fälle können bei dem gewählten Vorgehen im Grundtabellenfall nicht gefunden werden.

- Einkünfte aus selbständiger Arbeit (ja/nein),
- Einkünfte aus Vermietung und Verpachtung (ja/nein).

Auf diese Weise kam es zu ca. 135.000 eindeutigen Zusammenführungen. Der Anteil an doppelten Zuordnungen war mit 15.000 höher als in den anderen Fällen. Die Korrelation bei den Einkünften aus nichtselbständiger Arbeit lag mit 0,56 zwar deutlich im positiven Bereich, der Wert liegt aber einiges unter den Werten, die ansonsten erreicht wurden und deutet damit auf eine geringere Qualität der Zuordnung hin.

Die hohe Anzahl an indirekten Identifikatoren erhöht das Risiko, dass aufgrund von fehlerhaften Abweichungen in den Ausprägungen zwischen den Jahren identische Steuerpflichtige nicht als solche erkannt werden. Gründe für solche Abweichungen können sowohl veränderte Lebensumstände der Steuerpflichtigen, als auch Fehler in der Datenerfassung sein. Darüber hinaus sind die verwendeten indirekten Identifikatoren teilweise mit dem Problem behaftet, dass die Merkmalsausprägungen sehr stark konzentriert sind. So haben nur die wenigsten ledigen Steuerpflichtigen Kinder oder sind Gewerbetreibende. Dies führt dazu, dass durch diese Merkmale die Steuerpflichtigen in einen sehr kleinen und einen sehr großen Block unterteilt werden. Die Chancen für eine richtige Zusammenführung der Steuerpflichtigen der kleinen Blöcke sind zwar sehr hoch, derjenigen im großen Block aber sehr klein. Dieser Effekt lässt sich gut an der Gruppe der allein erziehenden Eltern zeigen. 15.700 der zusammengeführten Steuerpflichtigen haben mindestens ein Kind. Die Korrelation der Einkünfte dieser Gruppe liegt bei 0,88 und bestätigt eine hohe Qualität der Zuordnung bei dieser Subpopulation. Ebenfalls hohe Korrelationswerte werden erreicht, sobald eine der Gewinneinkunftsarten bzw. die Einkünfte aus Vermietung und Verpachtung besetzt sind.

Auf der zweiten Stufe der Zusammenführung wurden zunächst die bisher noch nicht zusammengeführten Merkmalsträger der Jahre 2001/02 um diejenigen bereinigt, die mit Hilfe der ersten Stufe zusammengeführt wurden. Dadurch reduzieren sich in beiden Dateien die gesuchten Merkmalsträger um jeweils rund 130.000. Dies erhöht die Chancen einer erfolgreichen Zusammenführung, allerdings ist die verbleibende Menge von rund 2 Mill. Grundtabellenfällen in beiden Datenbeständen weiterhin hoch. Entsprechend brachten Versuche, analog zur Zusammenführung der Splittingfälle komplett auf ein regionales Kriterium zu verzichten, keine zufrieden stellenden Resultate. Aus diesem Grund wurden Umzüge auf Bundeslandebene beschränkt und das Bundesland als indirekter Identifikator eingeführt. Als weitere indirekte Identifikatoren wurden mit den bereits auf der ersten Stufe verwendeten gearbeitet.

Die mit 427.000 Steuerpflichtigen vergleichsweise hohe Anzahl an Zuordnungen, die auf beschriebenen Weg erfolgten – bei gleichzeitig verhältnismäßig geringen Korrelationswerten (0,3) – deuten auf einen relativ hohen Anteil an Falschzuordnungen hin. Es gilt auch hier, dass eine hohe Qualität, gemessen am Korrelationskoeffizient, erzielt wird sobald entweder eine der Gewinneinkunftsarten besetzt ist oder Kinder vorhanden sind (0,62). Dies ist bei 15.000 Zuordnungen der Fall. Die 13.500 zusammengeführten

Merkmalsträger mit einer gekürzten Vorsorgepauschale – was u. a. bei Beamten der Fall ist –, erzielen ebenfalls erhöhte Korrelationswerte (0,56).

Wird keine erfolgreichere Art der Zusammenführung gefunden, ergeben sich verschiedene Möglichkeiten mit den Zuordnungen umzugehen. Sie können einerseits mangels Qualität grundsätzlich ignoriert werden. Andererseits könnte die Qualität als ausreichend erachtet und die Paneldatei entsprechend um diese Steuerpflichtigen ergänzt werden. Das Kriterium hierfür ist, inwieweit die zusätzlichen Steuerpflichtigen zu einem Qualitätsgewinn des Panels beitragen. Positiv für die Qualität sind die zusätzlich (richtig zusammengeführten) Merkmalsträger. Nachteilig sind die Falschzuordnungen. Was letztlich mehr zählt, muss abgeschätzt werden. Alternativ zu diesen grundsätzlichen Möglichkeiten könnte man auch nur die Gruppen in das Panel übernehmen, von denen die hohe Qualität der Zuordnungen bekannt ist. Dies wären alle Alleinerziehenden, die Bezieher von Gewinneinkünften und die Steuerpflichtigen mit gekürzter Vorsorgepauschale (z. B. Beamte). Dies hätte aber den Nachteil, dass es zu einer weiteren Verzerrung des Panels kommt. Allerdings dürfte diese bei der geringen Fallzahl nicht sonderlich ins Gewicht fallen.

3.4 Das erweiterte Testpanel 2001/02

Das erweiterte Testpanel besteht aus den Merkmalsträgern, die aufgrund der direkten (Steuernummer neu/alt) und aufgrund der indirekten Identifikatoren zusammengeführt wurden. Darüber hinaus sind diejenigen identifizierten Steuerpflichtigen enthalten, die 2001 zur Einkommensteuer veranlagt wurden aber nicht 2002. Zum einen sind dies die Fälle die 2002 aufgrund einer Heirat ausgefallen sind und zum anderen, diejenigen, die 2002 mit der Veranlagung ausgesetzt haben. Diese Merkmalsträger weisen Ausprägungen für 2001 auf. Die sich auf 2002 beziehenden Merkmale sind lediglich als Missing-Values enthalten. Darüber hinaus sind diejenigen Steuerpflichtigen enthalten, die 2002 aufgrund einer Scheidung erstmalig wieder in den Daten enthalten sind. Jeder im Panel enthaltene Datensatz erhält einen Merker der darüber informiert, auf welchem Wege der Datensatz Eingang in das Panel gefunden hat. Tabelle 3 enthält die Verteilung der Merker im Panel.

Tabelle 3: Verteilung der Merkmalsträger im Testpanel nach Zusammenführungsart

Merker	Zusammenführungsart	Anzahl		Kumulative Häufigkeit
		absolut	in %	
1	Steuernummer	23.745.287	95,60	23.745.287
2	Alte Steuernummer	45.037	0,18	23.790.324
3	Indirekte Identifikatoren	694.845	2,80	24.485.169
4	2002 ohne Erklärung	182.912	0,74	24.668.081
5	Ausfall durch Heirat	128.941	0,52	24.797.022
6	Neuaufnahme durch Scheidung	42.248	0,15	24.839.270

Aufgrund der zusätzlichen Zusammenführungen ging die Ausfallquote gemessen am Veranlagungsjahr 2001 von 14,1 auf 10,4 % zurück. Da einige der nicht verknüpften Merkmalsträger deswegen nicht verknüpft sind, weil der dazugehörige Datensatz in den 2002 Daten noch fehlte, wird dieser Wert aufgrund der nächsten Datenlieferung weiter sinken. Repräsentativer erscheinen daher die Zahlen für 2002, hier ging die Ausfallquote von 10,2 auf 6,4 % zurück. Hochgerechnet auf ein Panel mit vier Wellen, wären in diesem ca. 75 % aller Steuerpflichtigen enthalten. In Tabelle 4 sind analog zur Tabelle 2 die beschreibenden Statistiken des – in diesem Fall erweiterten – Panels im Vergleich zur Grundgesamtheit der jeweiligen Jahre enthalten.

Tabelle 4: Statistiken des erweiterten Testpanels 2001/2002 im Vergleich mit der Grundgesamtheit (Werte für den Gesamtbetrag der Einkünfte)

Jahr	Variationskoeffizient	Konfidenzintervall		Mittelwert	N	Median	Schiefe
		untere 99 % Grenze	obere 99 % Grenze				
Panel 2001	318	35.869	35.928	35.987	24.616.700	27.624	424
2001	336	34.535	34.592	34.649	27.709.160	26.194	387
Panel 2002	311	35.711	35.769	35.827	24.614.647	27.704	540
2002	313	34.230	34.284	34.338	26.504.230	26.792	553

Die Mittelwerte haben sich im Vergleich zur Tabelle 2 den Werten aus der Grundgesamtheit angenähert. Sie sind aber weiterhin signifikant erhöht. Steuerpflichtige, die nicht in das Panel integriert werden konnten, haben demnach geringere Einkünfte. Typische Fälle sind dabei die Berufsanfänger, die Todesfälle, oder diejenigen, die nach dem Eintritt in den Ruhestand keine Steuererklärung mehr abgeben müssen.

4 Beispiele von neuen Analysemöglichkeiten

In diesem Abschnitt sollen einige typische Analysen aufgezeigt werden, die mit Hilfe von Paneldaten entweder überhaupt erst möglich werden oder aber aufgrund dieser Daten mehr an Informationen aus den Analysen gewonnen werden können. Als Datenbasis dient das Testpanel 2001/02, wie es sich aufgrund der in Abschnitt 3 beschriebenen Zusammenführung ergibt.

- **Analysen von Bruttoeffekten**

Brutto- und Nettoeffekte zu unterscheiden und zu analysieren ist ein typisches Untersuchungsgebiet der Paneldatenanalyse.¹⁶ Hierbei wird der Nettoeffekt als die Veränderung einer Merkmalssumme zwischen zwei Zeitpunkten verstanden. So z. B. der Anstieg der Zahl von Arbeitslosen innerhalb eines Jahres. Diese Veränderung kann durch Wanderungsbewegungen in beide Richtungen ausgelöst sein, d. h. jemand der arbeitslos war hat eine Arbeit gefunden und jemand der beschäftigt war wurde arbeitslos. Der Nettoeffekt

¹⁶ Vgl. Hsiao, Cheng, Why Panel Data? a. a. O. Seite 3-4.

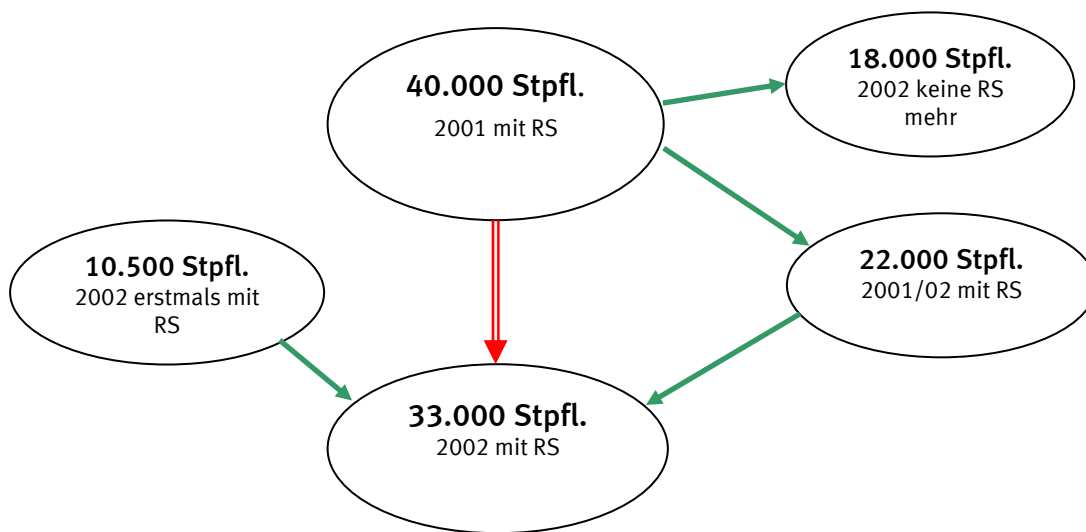
ist dann lediglich die Saldierung der Wanderungsbewegungen. Der Bruttoeffekt zeigt hingegen die Wanderungsbewegungen und damit die wahre Dynamik der Veränderung über die Zeit auf.

Im Zuge der Koalitionsverhandlungen 2005 wurde die Frage diskutiert, wie viele Steuerpflichtige von der geplanten Reichensteuer betroffen sind. Im beschriebenen Testpanel der Jahre 2001/02 sind rund 41.000 Steuerpflichtige enthalten, die aufgrund ihres zu versteuernden Einkommens des Veranlagungsjahres 2001 den um drei Prozentpunkte erhöhten Steuertarif bezahlen müssen¹⁷. Für das Jahr 2002 konnten rund 33.000 Steuerpflichtige ermittelt werden. Inwieweit es tatsächlich zu einem Rückgang kommt, kann nicht abschließend beurteilt werden, da noch eine Datenlieferung für das Jahr 2002 aussteht und die derzeit vorhandenen Daten gerade im höheren Einkommensbereich Lücken aufweisen. Die Frage nach dem Nettoeffekt, der hier einen Rückgang um 9.000 Steuerpflichtigen signalisieren würde, ist hier aber auch nicht von Interesse. Es soll nur der Unterschied zwischen Netto- und Bruttoeffekt aufgezeigt werden. Die wahre Dynamik wird bei einer Betrachtung der Wanderungsbewegungen deutlich. Sie ist im folgenden Schaubild skizziert.

Im konkreten Fall zeigen die Daten, dass fast 18.000 Steuerpflichtige, die im Jahr 2001 die Grenzen zur Reichensteuer überschritten, im Jahr 2002 nicht mehr in diese Regelung fallen. Umgekehrt kommen im Jahr 2002 10.500 Steuerpflichtige neu hinzu, die aufgrund ihres zu versteuernden Einkommens 2001 in diesem Jahr keine Reichensteuer zu zahlen hatten. Bei der politischen Beratung ist daher die alleinige Aussage „wie viele Steuerpflichtige von der Reichensteuer jährlich betroffen sind“ zu kurz gegriffen. Konkret müsste diese Aussage durch den Hinweis ergänzt werden, dass über die Zeit gesehen deutlich mehr von der Reichensteuer betroffen sind, oder positiv formuliert, dass die aufgrund der Reichensteuer entstehende Steuerlast mehrperiodisch betrachtet auf mehr Steuerpflichtige verteilt wird als es bei der Betrachtung der Querschnittsdaten zum Ausdruck kommt. Im konkreten Fall wären über zwei Jahre gesehen rund 51.000 Steuerpflichtige mindestens einmal von der Neuregelung betroffen, ca. 22.600 von ihnen müssten in beiden Jahren eine Reichensteuer bezahlen.

¹⁷ Als Grundlage für die Berechnungen gilt die ursprünglich geplante Reichensteuer, von der Steuerpflichtige mit einem zu versteuernden Einkommen (zvE) von über 250.000 € (Grundtabellenfälle) bzw. 500.000 € (Splittingfälle) betroffen waren.

Schaubild: Auswertung des Testpanel - Bruttoströme und Nettoströme im Vergleich



- RS = festgesetzter „Reichensteuer“ (zvE \geq 250.000 bzw. 500.00)
- Stpfl. = Steuerpflichtige
- ==> = Nettostrom
- > = Bruttoströme

• **Analysen von Progressionseffekten**

Bei der Einkommensbesteuerung wird der Grundsatz der horizontalen Gerechtigkeit postuliert, die sich dadurch ausdrückt, dass diejenigen mit gleichem Einkommen die gleiche Steuerlast tragen sollen. Ob dieser Grundsatz innerhalb eines Steuerjahres erfüllt wird, ist umstritten. Unstrittig ist, dass aufgrund der Ausgestaltung und der jährlichen Periodizität der Steuerfestsetzung der Grundsatz der horizontalen Gerechtigkeit über mehrere Jahre nicht erfüllt werden kann. Wie im Folgenden gezeigt wird, minimiert derjenige Steuerpflichtige intertemporal seine Steuerlast, der sein Einkommen auf die unterschiedlichen Perioden (Jahre) gleichmäßig verteilen kann. Je ungleicher die Einkommenserzielung, desto höher die Steuerlast. Der Grund hierfür liegt in der überproportionalen Besteuerung hoher Einkommen.

Diese Aussage ist rein theoretisch ableitbar und bedarf keiner empirischen Überprüfung anhand einer Panelanalyse. Die tatsächliche Bedeutung dieses intertemporalen Progressionseffekts zu ermitteln, ist allerdings nur mit Hilfe von Paneldaten möglich. So zeigen die Daten des Panels, dass ein Steuerpflichtiger, der nach der Splittingtabelle besteuert wird und innerhalb von zwei Jahren 100.000 € verdient, mindestens 19.000 € Steuern für diese beide Jahre zahlen muss. Maximal wurde dagegen ein Steuerbetrag von 46.000 € ermittelt. Der Progressionseffekt beträgt in diesem Fall 27.000 €. Um den gesamten Effekt

zu messen, wurde in einer Simulationsrechnung für jeden Steuerpflichtigen des Panels, der in den Jahren 2001/02 zur Einkommensteuer veranlagt wurde, die Summe der tariflichen Einkommensteuer der beiden Jahre berechnet. Dies erfolgte auf zwei Wegen: Zum einen, wie sie sich aufgrund der tatsächlich realisierten Einkommens ergibt und zum anderen wie sie sich ergeben würde, wenn der Steuerpflichtige das Einkommen gleichmäßig auf beide Jahre verteilt hätte. Die Differenz ergibt den Progressionseffekt, d. h. die Steuer, die der Steuerpflichtige nur aufgrund seiner ungleichen Einkommenserzielungen in den beiden Jahren bezahlen muss. Beschränkt wurde die Analyse auf Steuerpflichtige, welche die Besteuerungsart (Grund/Splitting) nicht gewechselt haben, da die dadurch ausgehenden Effekte ausgeschlossen werden sollten. Die Ergebnisse der Simulationsrechnungen sind in Tabelle 5 enthalten.

Tabelle 5: Steuerliche Mehrbelastung aufgrund ungleichmäßiger Einkommen der im Testpanel vertretenden Steuerpflichtigen in den Jahren 2001/02

Progressionsverlust							
Tarif	Maximum	Arithmetisches Mittel	Steuerpflichtige	Median	Perzentile		Summe (Mill. €)
					90	99	
	€	€	Anzahl		€	€	
2001	1.077.144	180	23.927.539	4	401	2.192	4.310
2005	932.135	166	23.927.539	4	375	2.124	3.967

Die Tabelle verdeutlicht, dass es zu 4.310 Mill. € an zusätzlicher Steuerbelastung aufgrund unregelmäßigem Erwerbsverlauf innerhalb der Jahre 2001/02 gekommen ist. Führt man die Berechnung nicht mit dem damals gültigen Tarif 2001, sondern mit dem aktuell gültigen Tarif 2005 durch, dann geht der Progressionseffekt auf 3.967 Mill. € zurück. Die Zahlen signalisieren hiermit einen Rückgang der Progressionswirkung aufgrund der Umstellung des Steuertarifs. Die Tabelle zeigt darüber hinaus, dass die Lasten sehr schief verteilt sind. Der Median beträgt lediglich 4 €, das arithmetische Mittel dagegen bereits 180 € (2001) und die 1 % Steuerpflichtigen, welche die höchsten Lasten tragen müssen, haben eine zusätzliche Steuerbelastung von mindestens 2.192 €. Diese Verteilung, die 2005 etwas an Schiefe verliert, zeigt, dass der „normale“ Steuerpflichtige nicht von dem Effekt der intertemporären Progression betroffen ist.

- **Analysen von Einkommensentwicklungen**

Die Analyse von Einkommensentwicklungen auf Basis von reinen Querschnittsdaten kann unvollständig sein. Wie im folgenden Beispiel verdeutlicht werden soll, können durch die Einbeziehung von Paneldaten zusätzliche Informationen gewonnen werden.

Tabelle 4 zeigt nicht nur, dass die durchschnittlichen Einkommen der im Panel enthaltenen Steuerpflichtigen signifikant höher liegen als es der Grundgesamtheit entspricht, sondern es wird auch ein Rückgang des arithmetischen Mittel der Einkommen von 2001 auf 2002 signalisiert (der Median steigt hingegen

gen leicht an).¹⁸ Mit Paneldaten wird es möglich, die Entwicklung jedes Steuerpflichtigen einzeln zu betrachten und so Aussagen über die individuelle Einkommensentwicklung zu treffen. Z. B. kann ermittelt werden, wie hoch der Anteil derjenigen Steuerpflichtigen ist, deren Einkommen ansteigen bzw. sinken. Darüber hinaus kann die durchschnittliche Differenz der Einkommen beider Jahre bezogen auf einen Steuerpflichtigen bestimmt werden. Der Median dieser Differenz stellt dann die Einkommenssteigerung bzw. den -rückgang eines „normalen“ Steuerpflichtigen dar.

Die Ergebnisse zeigen, dass knapp 60 % der Steuerpflichtigen Einkommenssteigerungen erzielen konnten, während die restlichen gut 40 % Einkommensverluste, gemessen am Gesamtbetrag der Einkünfte, hinnehmen mussten. Das arithmetische Mittel der Gewinne betrug 6.560 € (Median: 2.472 €), das der Verluste 10.800 € (Median: 3.766 €). Der obere Teil der Tabelle 6 zeigt für alle Steuerpflichtigen und nach Veranlagungsart aufgeteilt einige beschreibende Kennzahlen für die Einkommensänderungen. Wie der Median zeigt, stieg das Einkommen eines „normalen“ Steuerpflichtigen um 563 € an.

Tabelle 6: Auswertung Veränderung der Einkommen der im Testpanel vertretenen Steuerpflichtigen gemessen am Gesamtbetrag der Einkünfte 2001/02

Veranlagungsart	Anzahl Beobachtungen	arithmetisches Mittel in €	Standardabweichung	Median in €
Alle Steuerpflichtigen				
Grund	11.470.267	282	81.533	580
Splitting	12.575.939	-1.143	109.438	515
alle	24.046.206	-457	97.830	563
Steuerpflichtige mit Wohnortwechsel				
Grund	701.405	453	38.718	787
Splitting	166.513	-1.027	220.060	628
alle	869.041	176	103.134	813

Inwieweit mit einem Wohnortwechsel eine Erhöhung des Einkommens verbunden ist, kann mit den Paneldaten ebenfalls analysiert werden.¹⁹ Erstaunlicherweise konnten nur 56 % derjenigen Steuerpflichtigen, die den Wohnort gewechselt haben, dies mit Einkommenssteigerungen verbinden. Dies sind weniger als bei der Betrachtung aller Steuerpflichtigen. Trotzdem kann – wie der untere Teil der Tabelle 6 zeigt – diese Gruppe der Steuerpflichtigen höhere Einkunftssteigerungen erzielen als alle Steuerpflichtigen. Bei den nach der Grundtabelle besteuerten und bei den Steuerpflichtigen insgesamt sind die Unterschiede signifikant (nicht aber bei den nach der Splittingtabelle besteuerten). Der „normale“ Steuerpflichtige erzielt bei einem Umzug eine Einkommenssteigerung von 813 €.

¹⁸ Wie schon in dem Beispiel der Spitzenverdiener muss auch in diesem darauf hingewiesen werden, dass für das Jahr 2002 noch eine Datenlieferung fehlt, da der Bearbeitungszeitraum noch nicht abgeschlossen ist. Wie aber in Lietmeyer et al. gezeigt, steigt das durchschnittliche Einkommen während des Bearbeitungszeitraums an. Inwieweit die durchschnittlichen Einkommen tatsächlich sinken oder steigen kann daher noch nicht abschließend beurteilt werden. Für die zu zeigenden Effekte spielt dies aber keine Rolle. Vgl. Lietmeyer et al.: Jährliche Einkommensteuerstatistik auf Basis der bisherigen Geschäftsstatistik der Finanzverwaltung, WiSta, Heft 7, 2005, Seite 679.

¹⁹ Allerdings ist wie mehrfach beschrieben der Wohnortwechsel auch ein Grund für Panelmortalität. Von daher dürfte das Ausmaß des Wohnortwechsels unterschätzt sein.

- **Weitere Anwendungsbereiche**

Neben den aufgezeigten Anwendungsmöglichkeiten gibt es weitere Bereiche, die im Rahmen dieses Beitrags nicht oder noch nicht aufgezeigt werden können. Zu denken ist hier an Analysen im Bereich der Panelökonometrie²⁰ oder an Analysen zur Diffusion neuer „Steuersparmodelle“. Steuersparmodelle – wie der Sonderausgabenabzug nach § 10a EStG (Riester-Rente) – können als Angebote des Staates an seine Kunden (den Steuerpflichtigen) aufgefasst werden. Vor diesem Hintergrund wird es mit Paneldaten möglich, wie in der Marktforschung Untersuchungen zu deren Akzeptanz, Reichweite und ähnlichem durchzuführen. Mit den derzeitigen Testdaten sind diese Untersuchungen aufgrund der noch zu geringen Anzahl an verfügbaren Wellen nicht möglich. Umfasst der Zeitraum des Panels eine Steuerreform, so kann diese als „natürliches Experiment“ zur Beobachtung der Reaktion der Steuerpflichtigen auf die Reform aufgefasst werden.²¹ Dies fördert das Verständnis der Wirkungen von Steuerreformen.

5 Fazit

Die Entwicklung eines Taxpayer-Panels steht zwar noch am Anfang, dennoch zeigen bereits die ersten Arbeiten das Potenzial, welches mit solchen Daten gegeben ist. Die ersten Ergebnisse haben darüber hinaus gezeigt, dass ein Panel für die jährliche Einkommensteuerstatistik zu realisieren ist. Fragen wie die der Datenhaltung, von Verknüpfungsstrategien und dem Umgang mit der Panelmortalität bedürfen allerdings noch weiterer Untersuchungen. Nicht zuletzt muss noch eine ausreichende Datenbasis aufgebaut werden, um überhaupt ein aussagefähiges Panel aufbauen zu können. Dies wird in den nächsten Jahren schrittweise geschehen, so dass die Möglichkeit ein Panel über mehrere Jahre zu erstellen gegeben sein werden. Ein weiterer Schritt in diese Richtung wird noch in diesem Jahr folgen.

Literatur

EMMERLING, D.: Ehescheidungen 2003, in: WiSta, Heft 2, 2005, Seite 97-108.

GOTTFRIED, P.; SCHELLHORN, H.: Empirical Evidence on the effects of Marginal Tax Rates on Income – The German Case, IAW-Diskussionspapier, Nr. 15, 2004.

GOTTFRIED, P.; SCHELLHORN, H.: Die Elastizität des zu versteuernden Einkommens, IAW-Diskussionspapier Nr. 14, 2003

GOOLSBEE, A.: What happens when you tax the rich?, Journal of political economy, University of Chicago Press, vol. 108(2), pages 352-378, 2000.

²⁰ Eine Übersicht über die Methoden der Panelökonometrie findet sich in Lechner, M.: Eine Übersicht über gängige Modelle der Panelökonometrie und ihre kausale Interpretation, Universität St. Gallen, Department of Economics, Discussion paper No. 2001-18, 2001.

HSIAO, C.: Why Panel Data? Institute of Economic Policy Research, Working Paper 05.33, www.usc.edu/iepr, 2005.

LECHNER, M.: Eine Übersicht über gängige Modelle der Panelökonometrie und ihre kausale Interpretation, Universität St. Gallen, Department of Economics, Discussion paper No. 2001-18, 2001.

LIETMEYER, V.; KORDSMEYER, V.; GRÄB, C.; VORGRIMLER, D.: Jährliche Einkommensteuerstatistik auf Basis der bisherigen Geschäftsstatistik der Finanzverwaltung, in: WiSta, Heft 7, 2005, Seite 671-681.

²¹ So verwenden Gottfried und Schellhorn in ihren Arbeiten die Steuerreform von 1990 als ein „quasi natürliches Experiment“ um die Einflüsse der Änderungen der Steuersätze auf das zu versteuernde Einkommen zu messen. Vgl. Gottfried, P.; Schellhorn, H.: Empirical Evidence on the effects of Marginal Tax Rates on Income – The German Case, IAW-Diskussionspapier, Nr. 15, 2004.