BGPE Discussion Paper

No. 122

Entwicklung des integrierten Mikrosimulationsmodells EITDsim

Georg Struch

August 2012

ISSN 1863-5733

Editor: Prof. Regina T. Riphahn, Ph.D.
Friedrich-Alexander-University Erlangen-Nuremberg

B G P E Discussion Paper Series

Entwicklung des integrierten Mikrosimulationsmodells EITDsim

Georg Struch
Universität Potsdam

13. August 2012

Zusammenfassung

Dieser Beitrag dokumentiert den Aufbau des Mikrosimulationsmodells EITDsim. EITDsim kann zur empirischen (ex ante) Evaluation von steuer- und finanzpolitischen Fragestellungen, insbesondere von (potentiellen) Steuerreformen, verwendet werden. Dafür kombiniert EITDsim auf Basis des erweiterten Mikrodatensatzes FAST 2004 ein statisches Steuermikrosimulationsmodell zur Schätzung kurzfristiger Erst-Rundeneffekte mit einem diskreten Arbeitsangebotsmodell zur Schätzung mittelfristiger Zweit-Rundeneffekte. Aufgrund unvollständiger Informationen in der FAST 2004 wird als Datengrundlage für das Arbeitsangebotsmodell in EITDsim das Sozioökonomische Panel (SOEP) des DIW verwendet. Auf dem Wege einer Datensatzfusion werden sodann die Informationen aus beiden Mikrodatensätzen kombiniert.

JEL-Klassifikation: C51, C53, H24, J22
Schlagworte: Mikrosimulation, Arbeitsangebot, Datensatzfusion

Georg Struch, Lehrstuhl für Finanzwissenschaft, Wirtschafts- und Sozialwissenschaftliche Fakultät, Universität Potsdam, August-Bebel-Straße 89 , 14482 Potsdam. Tel. +49 (0) 331 - 977 32 59, E-Mail: Georg.Struch@Uni-Potsdam.de

Vorbemerkungen zum Forschungsprojekt EITD

Bevor in den folgenden Abschnitten der Aufbau und die Funktionsweise des Mikrosimulationsmodells EITDsim erläutert wird, soll an dieser Stelle kurz auf die Anfänge des Forschungsprojektes eingegangen werden. Das Akronym EITD steht für *Extended Income Tax Dataset* und ist das Ergebnis eines ersten Forschungsansatzes, eine geeignete Datenbasis für die Simulation steuer- und sozialpolitischer Reformvorschläge und -alternativen zu konstruieren.[1] EITD erweiterte im Zuge eines Fusions- bzw. Erweiterungsalgorithmus den Informationsgehalt der amtlichen Steuerstatistik (FAST 2001) zum einen mit zusätzlichen sozioökonomischen Variablen sowie zusätzlichen Einzeldatensätzen durch das SOEP 2001. Zum anderen wurden, der demographischen und monetären Entwicklung folgend, sowohl die Fallgewichte der Einzeldatensätze als auch die Einkommens- und Ausgabenbestandteile bis an den (damals) aktuellen Rand fortgeschrieben. EITD stellte daher für die Evaluierung von Steuerrechtsänderungen eine umfassendere Datengrundlage dar als die Ausgangsdaten der amtlichen Steuerstatistik.

Die Konstruktion des EITD erfolgte zunächst durch die Erweiterung der Datengrundlage FAST 2001 durch das SOEP. Hierbei wurden nach struktureller Angleichung beider Datensätze mit der Methode des Propensity-Score-Matchings zunächst diejenigen Fälle identifiziert, welche sich in beiden Datensätzen nur geringfügig unterscheiden, und anschließend wurden die Informationen aus beiden Datensätzen derart fusioniert, dass die Informationsbreite der amtlichen Steuerstatistik stieg. Um zu erreichen, dass der neue integrierte Mikrodatensatz tatsächlich alle steuerlich veranlagten Fälle in Deutschland abbildet, wurden, dem Ansatz von Bach et al. (2009) folgend, diejenigen Einzeldatensätze, welche nach dem Matching "übrig" waren, dem neuen Datensatz angehangen. Mit diesem Vorgehen wurden die amtlichen Steuerdaten (hochgerechnet) um über 18 Mio. Fälle erweitert. Diese erweiterte Datengrundlage wurde dann mit einem static-aging-Ansatz an die demographische Struktur des (damals) aktuellen Randes angepasst. Hierbei wurden die Altersstruktur, Familienstände, Erwerbsbeteiligung und Erwerbstätigkeit berücksichtigt. Neben der veränderten demographischen Struktur wurden schließlich die Einkommens- und Ausgabengrößen nominal fortgeschrieben. Hierbei wurden verschiedene Einkommens- und Ausgabenarten, soweit möglich, getrennt mit den entsprechenden Fortschreibungsgrößen angepasst.

Nach der Veröffentlichung der FAST 2004 im Jahr 2010 wurde der Ansatz des For-

[1] Struch & Jenderny (2010).

schungsprojektes EITD vollständig überarbeitet. Das Ziel war nun nicht mehr (nur) die Konstruktion einer informationell erweiterten Datengrundlage für Mikrosimulationsmodelle, sondern die Konstruktion eines integrierten Steuer-Transfer-Mikrosimulationsmodells zur Evaluation von kurz- und mittelfristigen Effekten von (potentiellen) Steuerreformvorschlägen.

1 Einleitung

Gerade in wirtschaftlich schwierigen Zeiten für die öffentlichen Haushalte braucht eine nachhaltig gestaltende Finanz- und Steuerpolitik ein hohes Maß an Information über die Wirkungszusammenhänge von verschiedenen Politikoptionen. Im Vorfeld einer Steuerreform sind die konkreten Auswirkungen nur schwer abzuschätzen. Aufgrund der hohen Komplexität von Steuer- und Abgabensystemen werden zur ex-ante Evaluation von Reformvorschlägen deshalb vermehrt Simulationsmodelle eingesetzt. Dabei wird versucht mit dem Modell eine möglichst genaue Abbildung des realen Wirtschafts- und Steuersystems zu erreichen, um davon ausgehend die Wirkungen einzelner finanzpolitischer Instrumente identifizieren und quantifizieren zu können. In den vergangenen Jahren sind zum Zwecke der wissenschaftsorientierten Politikberatung eine Reihe von mikrodatenbasierten Simulationsmodellen entwickelt worden, u.a. von Steiner et al. (2005), Fuest et al. (2005), Wagenhals & Buck (2006) Steiner & Wakolbinger (2009), Flory & Stöwhase (2010) oder Peichl et al. (2010), die sich aber sowohl hinsichtlich der zugrundeliegenden Datenbasis als auch hinsichtlich der methodischen Konzeption zum Teil erheblich unterscheiden. Einig sind sich all diese Modelle aber in der Zielsetzung, dem politischen Entscheidungsträger Informationen über die fiskalischen, allokativen und distributiven Effekte einzelner Reformoptionen bereitzustellen und zwar noch bevor diese in das reale Steuer- und Transfersystem implementiert werden.

In diesem Kontext steht auch das Mikrosimulationsmodell EITDsim, welches zur empirischen (ex ante) Analyse von Reformen des bestehenden Steuer- und Transfersystems auf das verfügbare Einkommen und das Arbeitsangebot von privaten Haushalten in Deutschland entwickelt wurde. EITDsim kombiniert ein Steuer-Transfermodell, das die Effekte von Sozialabgaben, Einkommensteuer & Solidaritätszuschlag sowie staatlicher Transferzahlungen auf das Nettohaushaltseinkommen simuliert, mit einem diskreten Arbeitsangebotsmodell, welches zur ökonometrischen Schätzung von Arbeitsangebotsreaktionen der Haushalte verwendet wird. Mit EITDsim ist es daher möglich, Aussagen über die, kurz- und mittel-

fristigen, fiskalischen, allokativen und distributiven Effekte einer (hypothetischen) Reform des bestehenden Steuer-Transfersystems zu treffen.[2]

Voraussetzung für die detaillierte Abbildung der komplexen institutionellen Zusammenhänge des Steuer- und Abgabensystems und die damit simulierbaren Politikoptionen ist jedoch eine aktuell repräsentative und qualitativ hochwertige Datenbasis. Angesichts der hohen Informationsanforderungen kann es durchaus sein, dass eine neue (artifizielle) Datenbasis aus verschiedenen Mikrodatensätzen konstruiert werden muss, um über alle notwendigen Variablen für die Evaluation der zu untersuchenden Politikmaßnahme zu verfügen. Mit der vorliegenden Arbeit wird die Vorgehensweise bei der Konzeption des Mikrosimulationsmodells EITDsim sowie die Fusion zweier repräsentativer Mikrodatensätze, der amtlichen Steuerstatistik und des Sozioökonomischen Panels, dokumentiert.

Die vorliegende Arbeit ist wie folgt gegliedert: In Kapitel 2 werden die theoretischen Grundlagen der Mikrosimulation aufgezeigt und der Modul-Aufbau des integrierten Mikrosimulationsmodell EITDsim vorgestellt. Die einzelnen Module werden anschließend in den Kapiteln 3, 4 und 5 erörtert. Abschließende Bemerkungen und einen Ausblick auf weiterführende Forschungsprojekte beinhaltet Kapitel 6.

2 Steuer-Transfer-Mikrosimulation mit EITDsim

2.1 Grundlagen

Seit Orcutt (1957) vor über 50 Jahren den (gedanklichen) Grundstein für einen neuen sozio-ökonomischen Modelltyp legte, hat sich die Verwendung von Mikrosimulationsmodellen (MSM) in den Sozial- und Wirtschaftswissenschaften aufgrund der hohen Flexibilität immens ausgebreitet.[3] MSM zählen zu den partial-analytischen Modellen und ermöglichen eine detaillierte Abbildung und (empirische) Analyse eines sozio-ökonomischen Systems sowie der darin befindlichen Mikroeinheiten. Insbesondere für die Finanzwissenschaft stellt die Mikrosimulation ein ideales Instrument für (ex ante) empirische Untersuchungen der Wirkungen von steuerpolitischen Reformmaßnahmen dar. Als grundlegende Voraussetzungen für die Methode der Mikrosimulation gelten dabei die Verfügbarkeit von repräsentativen Mikrodaten und die Vollständigkeit relevanter Variablen.

[2] Allerdings betrachtet EITDsim nur die Angebotsseite des Arbeitsmarktes, ist also kein allgemeines Gleichgewichtsmodell (CGE-Modell), weshalb keine Aussagen über Drittrundeneffekte möglich sind.

[3] Baroni & Richiardi (2007) geben einen Überblick über die Entwicklung der MSM in verschiedenen Wissenschaftsbereichen in den letzten 50 Jahren.

MSM können grundsätzlich in statische und dynamische Modellklassen unterschieden werden. Mit dynamischen Mikrosimulations- bzw. Arbeitsangebotsmodellen können die Langzeitwirkungen einer betrachteten Maßnahme auf den gesamten Lebenszyklus und z.B. den Aufbau von Humankapital oder die Ersparnis eines Untersuchungsobjektes analysiert werden. Aufgrund verschiedener Probleme (z.B. fehlende Variablen, Item-Nonresponse, Unit-Nonresponse) in Panel-Datensätzen spielen dynamische MSM in der empirischen Finanzwissenschaft immer noch eine untergeordnete Rolle. Einen Überblick über verschiedene Modelltypen und Anwendungsgebiete zeigen Blundell & MaCurdy (1999) in Kapitel 8 .

Mit statischen MSM wird dagegen anhand einer komparativ-statischen Analyse der kurzfristige Effekt (Erstrundeneffekt) von z.B. einer Steuerreform auf das Steueraufkommen oder die Verteilung der verfügbaren Einkommen gemessen. Die zentrale Annahme dieser Modellklasse lautet, dass die Untersuchungssubjekte vor und nach der simulierten Reformmaßnahme die gleichen Verhaltensmuster aufweisen. Diese eher restriktive Annahme einer vollkommen unelastischen Anpassungsreaktion kann für die meisten steuerlichen Maßnahmen in der kurzen Frist aber durchaus gerechtfertigt werden. Mithin ist es grundsätzlich plausibel, dass die Reaktion der Steuersubjekte auf z.B. die Abschaffung eines steuerlichen Ausnahmetatbestandes erst mit einer gewissen Verzögerung einsetzt, weil u.U. die konkreten Auswirkungen auf das verfügbare Einkommen unsicher sind und ein schnelles Handeln der Steuersubjekte unwahrscheinlich macht. Ebendieser mittelfristige Effekt (Zweitrundeneffekt) kann aber mit statischen Arbeitsangebotsmodellen qualitativ und quantitativ geschätzt werden. Zu diesem Zweck sind in den letzten 50 Jahren verschiedene Modelltypen konzipiert worden.[4] Das wesentliche Unterscheidungsmerkmal dieser statischen Arbeitsangebotsmodelle ist die Konstruktion der Budgetrestriktion. Demnach kann zwischen kontinuierlichen und diskreten Arbeitsangebotsmodellen differenziert werden.

Kontinuierlichen Arbeitsangebotsmodellen liegt die Annahme zugrunde, dass die Mikroeinheiten ihre Arbeitszeit frei wählen können. Für die, zur Schätzung der nutzenmaximalen Arbeitszeit notwendige, Budgetrestriktion existieren wiederum verschiedene Konstruktionsansätze: Burtless & Hausman (1978) bilden das Steuersystem möglichst exakt anhand einer stückweise linearisierten Funktion nach. Im Gegensatz dazu schlägt Hall (1973) eine vereinfachte Darstellung des Steuersystems anhand einer linearen Funktion vor, während Flood & MaCurdy (1992) das Steuersystem durch eine kontinuierliche und differenzierbare Funktion approximieren. Trotz dieser unterschiedlichen Ansätze sind die Schätzergebnissen von kontinuierlichen Arbeitsangebotsmodellen laut Creedy & Duncan (2002) ungenau

[4] Einen umfassenden Überblick bieten hierzu Blundell & MaCurdy (1999) oder Creedy & Duncan (2002).

oder verzerrt. Darüber hinaus wird der extensive Arbeitsangebotseffekt laut Brenneisen & Peichl (2007) nur unzureichend abgebildet und Van Soest et al. (1990) zufolge wird auch der Anteil der Teilzeitbeschäftigten tendenziell überschätzt.

Diskrete Arbeitsangebotsmodelle fußen auf der Annahme, dass die Mikroeinheiten ihre Arbeitszeit nicht frei, sondern aus einem endlichen Kontingent diskreter Arbeitszeiten wählen können. Diese grundlegende Annahme vereinfacht zum Einen den Rechenaufwand bei der Ermittlung der Budgetrestriktion und zum Anderen die Komplexität der Fragestellung, da der Nutzen eines betrachteten Haushaltes nur für einzelne Punkte aber nicht die gesamte Budgetrestriktion geschätzt werden muss. Auch bei der Spezifikation der Nutzenfunktion bieten diskrete Arbeitsangebotsmodelle mehr Flexibilität, da erst ex post überprüft werden muss, ob die grundsätzlichen Anforderungen an die Nutzenfunktion erfüllt sind. Darüber hinaus ermöglichen diskrete Arbeitsangebotsmodelle auch die Analyse von extensiven Arbeitsangebotseffekten. Ein Nachteil diskreter Arbeitsangebotsmodelle ist sicherlich der Verzicht auf Informationen über die tatsächlich gearbeiteten Stunden zugunsten der Modellierung von diskreten Arbeitszeitkategorien. Der damit verbundene Rundungsfehler wiegt die Vorteile der diskreten gegenüber der kontinuierlichen Modellierung des Arbeitsangebots nach herrschender Meinung aber nicht auf.[5]

2.2 Datengrundlage von EITDsim

Elementarer Bestandteil eines MSM ist die zugrundeliegende Datenbasis. Die Entscheidung für oder gegen eine bestimmte (verfügbare) Datengrundlage ist zuvorderst durch die mit dem MSM zu beantwortende Forschungsfrage bestimmt. Darüber hinaus kann wiederum, wenn z.B. relevante Informationen nicht verfügbar sind, die Datenbasis selbst die potentiellen Möglichkeiten eines MSM limitieren.

Aufgrund der Konzeption von EITDsim als Instrument zur Beantwortung bzw. Evaluation von insbesondere finanz- und sozialpolitisch motivierter Forschungsfragen wird als Primärdatenbasis in EITDsim die Faktisch Anonymisierte Lohn- und Einkommensteuerstatistik (FAST) 2004 verwendet.[6] Dieser Datensatz basiert auf einer geschichteten 10%-Zufallsstichprobe der Lohn- und Einkommensteuerstatistik aus dem Veranlagungsjahr 2004

[5] Zum Beispiel Van Soest (1995, S.83-84), Van Soest & Das (2000, S.1-2,26-28), Fuest et al. (2005, S.24) oder Peichl et al. (2010, S.14-15).

[6] Weil es bei der Veranlagung zur Einkommensteuer zu gewissen zeitlichen Verschiebungen kommen kann, erscheint die amtliche Statistik mit einer etwa fünf- bis sechsjährigen Verzögerung. Aufgrund dessen sind bis zum jetzigen Zeitpunkt keine aktuelleren Steuerdaten verfügbar.

und wird der Wissenschaft vom Statistischen Bundesamt als Scientific Use File zur Verfügung gestellt. FAST weist ca. 3,5 Millionen Einzeldatensätze auf und stellt eine repräsentative Stichprobe der rund 35 Millionen Einkommensteuerpflichtigen dar. Ehepaare, die eine gemeinsame Veranlagung gewählt haben, werden jedoch als ein Fall ausgewiesen. Unter Berücksichtigung dieses Aspekts repräsentiert die FAST 2004 insgesamt rund 52,78 Millionen Personen.

FAST 2004 liefert detaillierte Informationen über steuerlich relevante Einkommensquellen, Sonderausgaben, außergewöhnliche Belastungen sowie einige (steuerlich relevante) soziodemographische Merkmale, wie Alter, Geschlecht, Religion, Kinderzahl und Familienstand. Diese informationelle Tiefe macht FAST besonders für MSM zur Beantwortung finanz- und steuerpolitischer Forschungsfragen interessant. Ein nicht unerheblicher Nachteil der FAST ist jedoch, dass alle Einkommensbestandteile, die nicht steuerpflichtig sind sowie Informationen über das soziale Umfeld und insbesondere das Arbeitsverhalten der Steuerpflichtigen nicht erfasst werden.[7]

Als Sekundärdatenbasis wird in EITDsim das Sozioökonomische Panel (SOEP) verwendet. Das SOEP ist eine repräsentative Wiederholungsbefragung privater Haushalte in Deutschland, die als anonymisierter Mikrodatensatz vom DIW Berlin für die wissenschaftliche Forschung bereitgestellt wird. Im, für EITDsim relevanten, Erhebungsjahr 2004 wurden 11.796 Haushalte mit 22.019 Menschen befragt.[8] Diese Stichprobe ist repräsentativ für rund 40 Mio. Menschen in Deutschland, wobei jedoch besonders die unteren und mittleren Einkommensbereiche abgebildet werden. Es werden sowohl kontinuierlich Informationen über u.a. Persönlichkeitsmerkmale, Erwerbs- und Einkommensverläufe, Arbeitsverhalten, Haushaltszusammensetzungen als auch Informationen zu Schwerpunktthemen, wie z.B. Weiterbildung und Qualifikation, erhoben.[9] Das SOEP eignet sich aufgrund der Fülle an soziodemographischen bzw. -ökonomischen Informationen besonders für die Beantwortung von sozialpolitisch motivierter Forschungsfragen und wird in EITDsim insbesondere für die ökonometrische Schätzung von Arbeitsangebotsreaktionen in Folge einer Politikmaßnahme verwendet.

Aufgrund des Panelcharakters und der Erhebung von Informationen zu Haushalten

[7] Vgl. Forschungsdatenzentrum (2008).

[8] Die Befragung wird mittels Interviews durchgeführt und erfordert teilweise eine Selbsteinschätzung der Teilnehmer. Insbesondere bei Fragen zur Einkommenshöhe bzw. steuerlich relevanten Themen können hier jedoch Verzerrungen entstehen.

[9] vgl. Haisken-DeNew & Frick (2011).

sowie Einzelpersonen unterscheidet sich das Datendesign des SOEP deutlich vom Aufbau des FAST. Informationen über gemeinsam veranlagte Ehepaare werden im SOEP, anders als im FAST, nicht übergreifend als ein (Steuer-)Fall dargestellt. Es ist jedoch grundsätzlich möglich, den Ehepartner anhand einer eindeutigen Zuordnungsnummer zu identifizieren und die Informationen beider Einzeldatensätze zu kombinieren.

2.3 Modellaufbau

Eine schematische Darstellung der Grundstruktur von EITDsim zeigt die Abbildung 1. In Anbetracht der komplexen Modellierungsanforderungen, wird der Aufbau des gesamten Modells in drei Module unterteilt. Nach einer Aufbereitung der Daten sowie der Datensatzstruktur werden zunächst sowohl in der FAST als auch im SOEP die Einzeldatensätze mit Hilfe von Dummy-Variablen dahingehend kategorisiert bzw. selektiert, ob dem Haushalt ein (teilweise) flexibles oder unflexibles Arbeitsangebot unterstellt werden kann.[10] Dies geschieht vor dem Hintergrund, dass (1) die ökonometrische Schätzung des Arbeitsangebotsverhaltens nur für Einzeldatensätze des SOEP mit (teilweise) flexiblen Arbeitsangebot durchgeführt wird und (2) nur diese ausgewählten SOEP-Einzeldatensätze später auf dem Wege der Datensatzfusion mit (möglichst) ähnlichen Einzeldatensätzen der FAST verknüpft werden sollen. Von der Datensatzfusion werden jedoch diejenigen SOEP-Einzeldatensätze ausgenommen, welche im Status Quo erwerbslos sind.[11] Diese SOEP-Einzeldatensätze werden der FAST angehangen und erweitern so die Informationsbreite der Datenbasis.

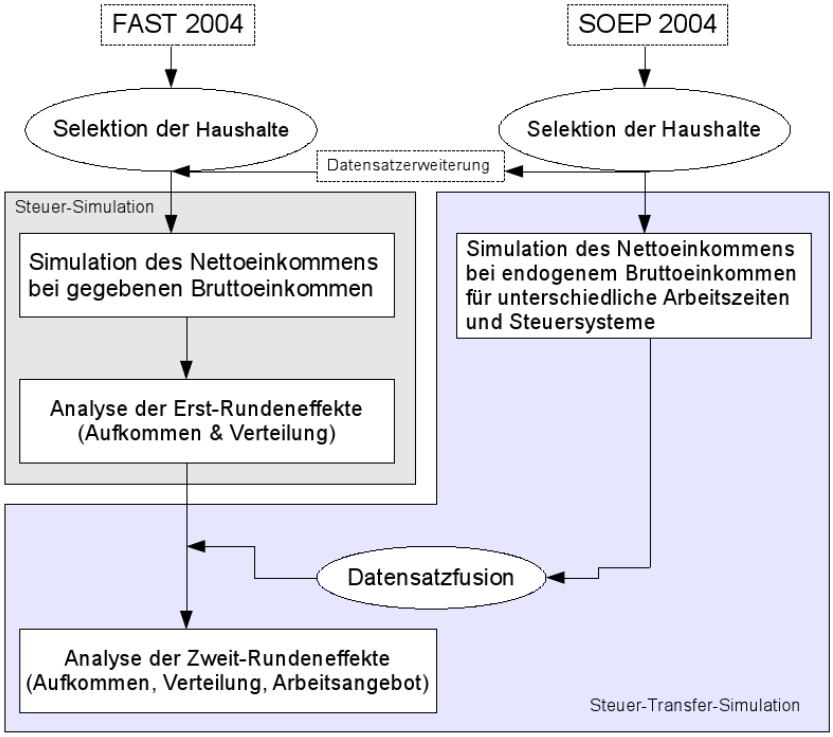
Das erste Modul enthält ein statisches Steuer-Simulationsmodell auf Basis der (erweiterten) FAST 2004 und dient zur Ermittlung der Aufkommens- und Verteilungseffekte einer potentiellen Steuerreform in der kurzen Frist. Die Simulation des aktuellen Steuersystems dient dabei als Referenzmodell für potentielle Steuerreformvorschläge. Dabei wird implizit angenommen, dass das Arbeitsangebot vollkommen unelastisch ist (Erst-Rundeneffekte). Das zweite Modul enthält ein Steuer-Transfer-Simulationsmodell sowie ein ökonometrisch geschätztes Arbeitsangebotsmodell auf Basis der (selektierten) SOEP-Daten. Das dritte Modul stellt eine Art "interaktive" Schnittstelle, mit der eine Verknüpfung der Module 1 und 2 hergestellt wird. Genauer gesagt werden über einen Fusionsalgorithmus die SOEP-

[10] Näheres hierzu folgt in Abschnitt 4.2.

[11] Da der aktuelle Beschäftigungsstatus in der FAST nicht beobachtet werden kann, wird angenommen, dass alle in der FAST enthaltenen Mikroeinheiten mit flexiblem Arbeitsangebot auch tatsächlich erwerbstätig sind.

Einzeldatensätze, für die eine Arbeitsangebotsreaktion ökonometrisch (valide) geschätzt werden konnte, mit möglichst ähnlichen FAST-Einzeldatensätzen verbunden. Mit diesem (artifiziellen) Mikrodatensatz ist es dann möglich, die Aufkommens-, Verteilungs- und Arbeitsangebotseffekte einer potentiellen Steuerreform in der mittleren Frist zu ermitteln (Zweit-Rundeneffekte). Es wird in EITDsim außerdem implizit angenommen, dass ein zusätzliches Arbeitsangebot auch tatsächlich nachgefragt wird.[12]

Abbildung 1: Modellaufbau EITDsim



Quelle: Eigene Darstellung.

3 Das Steuer-Simulationsmodul in EITDsim

Mit dem Steuer-Simulationsmodul von EITDsim ist es möglich die kurzfristigen Effekte einer Veränderung der steuerlichen Rahmenbedingungen auf das Steueraufkommen oder

[12] EITDsim ist also kein allgemeines Gleichgewichtsmodell (CGE-Modell), da hierfür zusätzlich noch eine endogene Arbeitsnachfrage modelliert werden müsste.

die empirische Einkommensverteilung zu simulieren. Das Modul kann jedoch keine Verhaltensanpassung der Steuerpflichtigen endogenisieren. D.h. es wird implizit unterstellt, dass die Steuerpflichtigen kurzfristig z.B. einer höheren Steuerbelastung nicht ausweichen bzw. das Arbeitsangebot vollkommen unelastisch ist. Diese starke Vereinfachung beschränkt die Ergebnisse der Analyse einer (potentiellen) Tarifreform dahingehend, dass nur Aussagen über Einkommenseffekte, sogenannte 'Erst-Rundeneffekte', möglich sind. Der für EITDsim verwendete Ansatz der Steuer-Simulation basiert größtenteils auf Struch (2012).

3.1 Datenbasis

Die Grundlage für die komparativ-statische Analyse einer (potentiellen) Steuerreform in EITDsim stellt der erweiterte Mikrodatensatz FAST 2004 dar. Diese erweiterte FAST setzt sich aus den Einzeldatensätzen der FAST 2004 und den als erwerbslos identifizierten SOEP-Einzeldatensätzen des Jahres 2004 zusammen.

3.2 Modellierung des aktuellen Einkommensteuerechts

Der Aufbau des Steuersimulationsmodells orientiert sich am Berechnungsschema der tariflichen Einkommensteuer in § 2 EStG. Nach der Ermittlung des zu versteuernden Einkommens (zvE) erfolgt die Berechnung der Einkommensteuer nach § 32a EStG. Nach geltender Rechtslage 2012 wird die Einkommensteuer durch Anwendung des Formeltarifs in § 32a Abs. 1 EStG berechnet. Bis zu einem zvE von 8.004 Euro (Grundfreibetrag) muss keine Einkommensteuer gezahlt werden. Bei einem zvE von 8.005 Euro beträgt der Eingangssteuersatz 14%. Innerhalb dieser ersten Progressionszone steigt die Grenzbelastung bis auf knapp 24% bei einem zvE von 13.469 Euro stark an. In der anschließenden zweiten Progressionszone steigt die Grenzbelastung langsamer, aber stetig bis auf knapp 42% bei einem zvE von 52.881 Euro an. Innerhalb der folgenden ersten Proportionalzone (ab einem zvE von von 52.882 Euro) und bis zu einem zvE von 250.730 Euro beträgt die Grenzbelastung konstant 42% und springt danach auf den Spitzensteuersatz von 45%. Die Durchschnittsbelastung steigt mit dem zvE, wobei die Steigung der Funktion gegen Null strebt, und nähert sich dem Spitzensteuersatz asymptotisch an. Ehepaare können auf Antrag eine gemeinsame Veranlagung wählen. Die Einkommensteuer wird in diesem Fall nach § 32a Abs. 5 EStG (Splittingverfahren) ermittelt: Die Einkommensteuer beträgt das Doppelte des Steuerbetrags, der sich für die Hälfte des gemeinsamen zvE nach Anwendung des Formeltarifs in § 32a Abs. 1 EStG ergibt. Für eine korrekte Erfassung der Einkommensteuer werden

außerdem folgende Vorschriften berücksichtigt:

- § 32 EStG Freibeträge für Kinder (Günstigerprüfung)
- § 32b EStG Progressionsvorbehalt
- § 32d EStG Gesonderter Tarif für Einkünfte aus Kapitalvermögen (Abgeltungsteuer)

Bei Inanspruchnahme des Kinderfreibetrages erfolgt eine Hinzurechnung des Kindergeldes zur ermittelten Einkommensteuer. Ferner wird im Modell angenommen, dass jeder Steuerpflichtige die Veranlagungsoption nach § 32d Abs. 6 EStG immer dann wählt, wenn die Abgeltungsteuer höher als die Steuer ist, die sich ergeben würde, wenn die Einkünfte aus Kapitalvermögen gemäß § 32a EStG besteuert werden würden. Das (Brutto-)Markteinkommen wird als Aggregat aus dem zu versteuernden Einkommen (zuzüglich etwaige Frei- und Entlastungsbeträge), den Bruttoeinkünften aus Kapitalvermögen und den Einkünften, die dem Progressionsvorbehalt des § 32b EStG unterliegen, definiert. Das verfügbare Einkommen ergibt sich nach Abzug der tariflichen Einkommensteuer, der Abgeltungsteuer und des Solidaritätszuschlags vom (Brutto-)Markteinkommen.[13]

4 Das Arbeitsangebotsmodul von EITDsim

Mit dem Arbeitsangebotsmodul von EITDsim ist es möglich Verhaltensanpassungen der Haushalte auf eine Veränderung der steuerlichen Rahmenbedingungen zu simulieren. Der für EITDsim verwendete Ansatz basiert auf dem neoklassischen Lehrbuchmodell, wonach ein Individuum bei der Wahl seines Arbeitsangebots vor einem Trade-off zwischen Freizeit und Einkommen bzw. Konsum steht. Das Individuum maximiert seinen Nutzen, welcher positiv vom Einkommen bzw. Konsum und negativ von der Arbeitszeit abhängt, unter den Nebenbedingungen, dass das gesamte Einkommen verkonsumiert wird und außerdem nur eine begrenzte Freizeitausstattung möglich ist. Die Lösung des Problems ergibt für das Individuum ein optimales Bündel aus Freizeit und Einkommen sowie das dafür notwendige Arbeitsangebot.

Mit EITDsim wird zunächst das nutzenmaximierende Arbeitsangebot unter Maßgabe des aktuellen Steuersystems mit einem statisch diskreten Haushaltsarbeitsangebotsmodell

[13] Bei dem hier verwendeten steuerrechtlichen Nettoeinkommenskonzept werden also weder fehlende Einkunftsdaten imputiert noch werden Transferzahlungen, wie z.B. Kindergeld, Leistungen nach dem SGB II, etc., berücksichtigt. D.h. aus der FAST wird nicht, wie z.B. Bönke et al. (2007) vorschlagen, ein "ökonomisches Einkommen" zurückberechnet. Insbesondere für die Bezieher niedriger Einkommen werden damit die Konsummöglichkeiten eher unterschätzt.

nach Van Soest (1995) ökonometrisch geschätzt. Die Identifizierung von Arbeitsangebotsreaktionen der Haushalte auf eine Veränderung des Steuerregimes erfolgt dann über einen Vergleich der prognostizierten Arbeitszeitkategorien vor und nach der Reform.

4.1 Modell

Als theoretische Grundlage dient das diskrete Haushaltsarbeitsangebotsmodell von Van Soest (1995). Van Soest nimmt an, dass die Entscheidung der Haushaltsmitglieder über ihre Arbeitszeit nicht (völlig) beliebig sei. Aus diesem Grund könne man die Wahl der Arbeitszeit statt als kontinuierliche Variable auch als kategoriale Variable modellieren, d.h. die Arbeitsangebotsentscheidung könne als Wahl zwischen einer bestimmten Anzahl von Arbeitszeitalternativen betrachtet werden. Mit diskreten Auswahlmodellen wird die bedingte Wahrscheinlichkeit geschätzt, dass ein Haushalt i aus K Alternativen immer dann die Alternative k wählt, wenn der damit einhergehende Nutzen U_{ik} größer als der Nutzen aller anderen Alternativen l ist:

$$P_{ik} = P(U_{ik} > U_{il}) \; \forall \; l = 1, \ldots, K, \, l \neq k \tag{1}$$

Dem Prinzip der stochastischen Nutzenmaximierung von McFadden (1973) folgend, kann der Nutzen U_{ik} in einen deterministischen (direkt beobachtbaren) Teil V_{ik} und einen zufälligen (unbeobachtbaren) Teil ε_{ik} aufgespalten werden.

$$U_{ik} = V_{ik} + \varepsilon_{ik} \tag{2}$$

Der direkte Nutzen V_{ik} ist eine Funktion von Nettoeinkommen und Freizeit sowie weiteren Merkmalen des Haushaltes. Der stochastische Teil ε_{ik} enthält alle unbeobachteten Faktoren, die den Nutzen des Haushaltes beeinflussen. Durch Einsetzen von (2) in (1) und Umformen kann die bedingte Wahrscheinlichkeit, dass Haushalt i die Alternative k allen anderen Alternativen vorzieht wie folgt dargestellt werden:

$$P(U_{ik} > U_{il}) = P(V_{ik} - V_{il} > \varepsilon_{il} - \varepsilon_{ik}). \tag{3}$$

Unter der Annahme, dass die Fehlerterme ε_{ik} über alle Haushalte i und Arbeitszeitkategorien K unabhängig und identisch verteilt sind und einer Extremwertverteilung vom Typ I (Gumbel) folgen, kann die bedingte Auswahlwahrscheinlichkeit dann für jede mögliche Arbeitszeitkategorie über ein Conditional-Logit-Modell nach McFadden (1973) berechnet werden:

$$P(U_{ik} > U_{il}) = \frac{\exp(V_{ik})}{\sum_{l=1}^{K} \exp(V_{il})}. \tag{4}$$

Um das nutzenmaximierende Verhalten der Haushalte abbilden zu können, werden die Parameter der (direkten) Nutzenfunktion V_{ik} empirisch geschätzt. Dabei wird angenommen, dass Ehepartner eine gemeinsame Nutzenfunktion maximieren.[14]

In der Literatur finde sich verschiedene funktionale Spezifikationen für den zu schätzenden (direkten) Haushaltsnutzen V_{ik}.[15] Notwendige Bedingung für eine sinnvolle ökonomische Spezifikation ist jedoch, dass der Grenznutzen von Einkommen und Freizeit positiv ist.[16] In EITDsim wird der (direkte) Nutzen eines Haushaltes i bei Wahl einer bestimmten Arbeitszeitkategorie k mit einer Translog-Funktion modelliert:

$$V_{ik}(x_{ik}) = x'_{ik} A x_{ik} + \beta' x_{ik}, \tag{5}$$

wobei $x = (\ln ekn,\ \ln lm,\ \ln lf)'$.[17] Die (beobachtbaren) Elemente von x sind das Nettohaushaltseinkommen (ekn), die Freizeit des Mannes (lm) und die Freizeit der Frau (lf). In die Nutzenfunktion gehen die Elemente von x sowohl in linearer und quadratischer Form als auch in Form von Kreuztermen ein. Die Matrix A mit den Elementen α_{ij}, $ij = (1,2,3)$, beinhaltet die (zu schätzenden) Koeffizienten der quadratischen Terme und der Kreuzterme. Der Vektor β_j, $j = (1,2,3)$ die (zu schätzenden) Koeffizienten der linearen Terme.

Statt in Matrix-Schreibweise kann (5) umgeformt auch wie folgt dargestellt werden:

$$\begin{aligned} V_{ik}(x_{ik}) &= \beta_1 \ln ekn_{ik} + \beta_2 \ln lm_{ik} + \beta_3 \ln lf_{ik} + \alpha_{11}(\ln ekn_{ik})^2 + \alpha_{22}(\ln lm_{ik})^2 \\ &+ \alpha_{33}(\ln lf_{ik})^2 + 2\alpha_{12} \ln ekn_{ik} \ln lm_{ik} + 2\alpha_{13} \ln ekn_{ik} \ln lf_{ik} + 2\alpha_{23} \ln lm_{ik} \ln lf_{ik}. \end{aligned} \tag{6}$$

Ferner wird die (beobachtbare) Heterogenität der Haushalte wie bei Van Soest (1995) durch folgende Spezifizierung der Parameter β_m, α_{mn} berücksichtigt:

$$\beta_m = \beta_{m0} + \sum_{p=1}^{P} \beta_{mp} z_p \tag{7}$$

$$\alpha_{mn} = \alpha_{mn0} + \sum_{p=1}^{P} \alpha_{mnp} z_p, \tag{8}$$

[14] Sog. Household Utility oder Unitary Model.

[15] Stern (1986) gibt einen Überblick über verschiedene Spezifikationen von direkten Nutzenfunktionen für Arbeitsangebotsfunktionen.

[16] Diese Bedingung muss allerdings erst ex post überprüft werden.

[17] Der entscheidende Vorteil dieser Spezifikation liegt in ihrer Flexibilität. D.h. durch die Modellierung von Kreuztermen ist es möglich, die Abhängigkeit von z.B. dem Freizeitnutzen eines Haushaltsmitglieds auf den Freizeitnutzen seines Partners zu schätzen.

wobei $m, n = 1, 2, 3$ und z_p $(p = 1, \ldots, P)$ die Fall-spezifischen Kontrollvariablen darstellen. Dazu zählen z.B. Alter, Nationalität, Schulabschluss, Bruttostundenlohn, Anzahl und Alter der Kinder im Haushalt sowie die Wohnregion (Ost/West).

Nach der Schätzung muss geprüft werden, ob das geschätzte Haushaltsnutzen-Modell auch mit der zugrunde liegenden ökonomischen Theorie übereinstimmt. Wie oben ausgeführt, muss der Grenznutzen der Freizeit sowie des Einkommens positiv, aber abnehmend sein.[18] D.h. unter der Voraussetzung, dass Freizeit für die Haushaltsmitglieder ein normales Gut darstellt, sollte die erste Ableitung der Nutzenfunktion nach der Freizeit eines Haushaltsmitglieder positiv und die die zweite Ableitung negativ sein. Darüber hinaus ist das Vorzeichen des Substitutionseffektes der Freizeit der Haushaltsmitglieder untereinander bzw. von Freizeit und Einkommen der Haushaltsmitglieder jedoch theoretisch unbestimmt.

$$\frac{\partial\, V}{\partial\, ekn} = \frac{\beta_1 + 2\alpha_{11} \ln ekn + 2\alpha_{12} \ln lm + 2\alpha_{13} \ln lf}{ekn} \tag{9}$$

$$\frac{\partial\, V}{\partial\, lm} = \frac{\beta_2 + 2\alpha_{22} \ln lm + 2\alpha_{12} \ln ekn + 2\alpha_{23} \ln lf}{lm} \tag{10}$$

$$\frac{\partial\, V}{\partial\, lf} = \frac{\beta_3 + 2\alpha_{33} \ln lf + 2\alpha_{13} \ln ekn + 2\alpha_{23} \ln lm}{lf} \tag{11}$$

[18] Vgl. hierzu Steiner & Wrohlich (2004) und Brenneisen & Peichl (2007).

4.2 Datensatz, Selektion und Variable

Als Datenbasis für die ökonometrische Schätzung und Simulation des Arbeitsangebotsverhaltens dient in EITDsim das Sozioökonomische Panel (SOEP) des DIW Berlin.[19] Im, für die Simulation des Arbeitsangebotsverhaltens relevanten, Erhebungsjahr 2004 wurden 11.796 Haushalte mit 22.019 Menschen befragt.

Wie in der Literatur üblich, wird in einem ersten Schritt die Schätzung des Arbeitsangebotes auf jene Personen beschränkt, denen ein flexibles Arbeitsangebot unterstellt werden kann.[20] Dazu gehören sowohl alle abhängig Beschäftigten als auch alle Arbeitslosen. Ein unflexiblen Arbeitsangebot wird Personen unterstellt, die eines der folgenden Merkmale aufweisen:

- Personen, die jünger als 16 oder älter als 65 Jahre sind
- Beamte
- Bezieher von Altersrente, Altersübergangs- oder Vorruhestandsgeld
- Auszubildende und Studenten
- Frauen im Mutterschutz
- Wehr- oder Zivildienstleistende
- Selbstständige

Der zweite Schritt stellt die Zuordnung der Personen in eine bestimmte Haushaltskategorie dar. Es werden dafür in Abhängigkeit der Arbeitsangebotsflexibiltät drei Haushaltstypen definiert.[21] Der erste Haushaltstyp wird als "flexibler Haushalt" bezeichnet, d.h. sowohl dem Haushaltsvorstand als auch seinem Partner kann ein flexibles Arbeitsangebot unterstellt werden. Der zweite Haushaltstyp wird als "gemischter Haushalt" charakterisiert, wenn entweder dem Haushaltsvorstand oder seinem Partner ein flexibles Arbeitsangebot unterstellt werden kann. Den dritten Haushaltstyp, der sogenannte "unflexible Haushalt", kennzeichnet, dass neben dem Haushaltsvorstand auch sein Partner ein unflexibles Arbeitsangebot im obigen Sinne aufweist. Haushalte, die der Kategorie des "unflexiblen Haushaltes" zugeordnet werden können, werden aus der Schätzstichprobe entfernt. Darüber hinaus wird bei der Modellierung zwischen Single- und Paarhaushalten unterschieden. Zwei Personen zählen nur dann als Paarhaushalt, wenn sie zusammen veranlagt werden. "Gemischte

[19] Ausführliche Informationen zum SOEP geben Haisken-DeNew & Frick (2011).

[20] Vgl. Van Soest & Das (2000), Steiner & Wrohlich (2004), Fuest et al. (2005) oder Franz et al. (2007).

[21] Ähnlich gehen Fuest et al. (2005) oder Steiner & Wakolbinger (2009) vor.

Haushalte" verbleiben in der Schätzstichprobe, jedoch wird der Partner mit dem flexiblen Arbeitsangebot wie ein "technischer" Single-Haushalt behandelt.

Die Schätzstichprobe umfasst (ohne Berücksichtigung der Hochrechnungsfaktoren) insgesamt 2.806 Paar-Haushalte und 5.470 (technische) Single-Haushalte, welche wiederum in 2.386 Single-Männer und 3.084 Single-Frauen unterschieden werden können.

4.2.1 Budgetrestriktion des Haushaltes

Um den geschlechtsspezifischen Unterschieden bei der tatsächlichen Wochenarbeitszeit Rechnung zu tragen, wird für Männer und Frauen eine unterschiedliche Anzahl möglicher Arbeitszeitkategorien (k) modelliert. Dies bedeutet, dass Männer zwischen drei Alternativen wählen können. Die Modellierung der Arbeitszeitkategorien für Männer ist in Tabelle 1 dargestellt.

Tabelle 1: Arbeitszeitkategorien - Männer

k	Wochenarbeitszeit (in Stunden)	
0	0	Arbeitslos
1	1 - 40	Voll erwerbstätig
2	> 40	Überstunden

Bei Frauen wird angenommen, dass sie sich zwischen fünf Alternativen entscheiden können. Die Modellierung der Arbeitszeitkategorien für Frauen ist in Tabelle 2 dargestellt. Um die Arbeitsangebotsentscheidung von Paar-Haushalten modellieren zu können, wird

Tabelle 2: Arbeitszeitkategorien - Frauen

k	Wochenarbeitszeit (in Stunden)	
0	0	Arbeitslos
1	1 - 15	Teilzeit erwerbstätig
2	16 - 34	Teilzeit erwerbstätig
3	35 - 40	Vollzeit erwerbstätig
4	> 40	Überstunden

angenommen, dass diese Haushalte eine gemeinsame Nutzenfunktion maximieren. Zu diesem Zweck werden die geschlechtsspezifischen Arbeitszeitkategorien für Männer und Frauen in Arbeitszeitkombinationen für Paare überführt. Dieses Vorgehen führt zu insgesamt

15 verschiedenen Arbeitszeitkombinationen, aus denen ein Paar-Haushalt das gemeinsame Arbeitsangebot wählen kann.

Van Soest (1995) folgend, wird die Freizeitausstattung eines Haushalts i für jede korrespondierende Arbeitszeitkategorie k gemäß folgender Regel bestimmt:

$$lm_{ik} = TE - hm_{ik} \text{ bzw. } lf_{ik} = TE - hf_{ik},$$

wobei lm (lf) für die Freizeitausstattung des Mannes (der Frau), TE für das Zeitbudget pro Woche und hm (hf) für die Arbeitszeit des Mannes (der Frau) stehen. Es wird angenommen, dass die gesamte Zeitausstattung TE eines Haushaltsmitgliedes pro Woche 80 Stunden beträgt.[22] Die Arbeitszeiten hm bzw. hf werden definiert als Produkt aus der Arbeitszeitkategorie k und einem bestimmten Intervall IL_g, $g \in \{m, f\}$:

$$hm_{ik} = k \cdot IL_m \text{ bzw. } hf_{ik} = k \cdot IL_f.$$

Aufgrund der unterschiedlichen Anzahl an Arbeitszeitkategorien für Männer und Frauen, wird für Männer eine Intervalllänge von $IL_m = 24$ und für Frauen eine Intervalllänge von $IL_f = 12$ definiert.

Für die Simulation des Nettohaushaltseinkommens wird in EITDsim ein Steuer-Transfer-Modell, das alle relevanten Komponenten des EStG beinhaltet, verwendet. Der vorgenommenen Selektion des Datensatzes entsprechend, stellen die Einkünfte aus abhängiger Beschäftigung für die Mehrzahl der betrachteten Haushalte die Haupteinkommenskomponente dar. Die (beobachtbare) Information über das Bruttomonatseinkommen wird, zusammen mit der (beobachtbaren) Information über die Wochenarbeitszeit des Haushaltsmitglieds, benutzt, um einen Bruttostundenlohn zu ermitteln. Bei der Simulation des (hypothetischen) Bruttomonatseinkommens eines Haushaltsmitgliedes für jede Arbeitszeitkategorie k wird der ermittelte Bruttostundenlohn mit der durchschnittlichen Wochenarbeitszeit jeder Arbeitszeitkategorie k multipliziert. Bei Ehepaaren wird das für jeden Ehepartner ermittelte (hypothetische) Bruttomonatseinkommen einer Arbeitszeitkategorie zu einem gemeinsamen Bruttohaushaltseinkommen aufsummiert. Wie in der Literatur üblich wird dabei implizit angenommen, dass der Bruttostundenlohn unabhängig von der tatsächlichen Arbeitszeit ist.[23] Für Haushaltsmitglieder, die entweder erwerbslos sind oder bei denen keine Informationen über den Bruttostundenlohn vorliegen, wird mit einem zweistufigen Selektionsmodell ein (hypothetischer) Bruttostundenlohn geschätzt. Zu diesem Zweck wird

[22] Van Soest (1995), Van Soest & Euwals (1999) und Steiner & Wrohlich (2004) zeigen, dass alternative Zeitbudgetwerte, z.B. $TE = 60$, kaum Einfluss auf die Schätzergebnisse haben.

[23] Vgl. Van Soest (1995), Van Soest & Das (2000) oder Fuest et al. (2005).

eine Lohn- bzw. Partizipationsgleichung durch eine Regression empirisch geschätzt.[24] Die Schätzergebnisse sind in der Tabelle 13 im Appendix abgetragen.

Für die Simulation des Nettoeinkommens werden ausgehend vom (hypothetischen) Bruttomonatseinkommen eines Haushaltes in der Kategorie k die Sozialversicherungsbeiträge des Arbeitnehmers, die simulierte Einkommensteuer sowie der Solidaritätszuschlag abgezogen und (potentielle) staatliche Transferzahlungen hinzu addiert. Die Simulation der Einkommensteuer orientiert sich am Berechnungsschema der tariflichen Einkommensteuer in § 2 EStG. Nach Ermittlung des zu versteuernden Einkommens (zvE) erfolgt die Berechnung der Einkommensteuer entsprechend § 32a EStG. Für eine korrekte Erfassung der Einkommensteuer werden außerdem folgende Vorschriften berücksichtigt:

- § 32 EStG Freibeträge für Kinder (Günstigerprüfung)
- § 32b EStG Progressionsvorbehalt
- § 32d EStG Gesonderter Tarif für Einkünfte aus Kapitalvermögen (Abgeltungsteuer)

Bei Inanspruchnahme des Kinderfreibetrages erfolgt eine Hinzurechnung des Kindergeldes zur ermittelten Einkommensteuer. Ferner wird im Modell angenommen, dass jeder Steuerpflichtige die Veranlagungsoption nach § 32d Abs. 6 EStG immer dann wählt, wenn die Abgeltungsteuer höher als die Steuer ist, die sich ergeben würde, wenn die Einkünfte aus Kapitalvermögen gemäß § 32a EStG besteuert würden.

Die Modellierung des staatlichen Transferanspruchs bei Erwerbslosigkeit umfasst in EITDsim neben dem o.g. Kindergeld auch das Arbeitslosengeld I, das Arbeitslosengeld II, das Kindersozialgeld sowie die Zulagen für Unterkunft und Heizung. Der Transferanspruch verringert sich durch einen Hinzuverdienst wie folgt: Nach aktueller Rechtslage bleiben die ersten einhundert Euro anrechnungsfrei. Zwischen 100 und 1.000 Euro werden 80 Prozent und zwischen 1.000 und 1.200 Euro 90 Prozent vom Nettoeinkommen auf den Transferanspruch angerechnet.

4.2.2 Deskriptive Analysen der Variablen

Die Erwerbsquote in der Grundgesamtheit (GG) beträgt 56,78%.[25] Tabelle 3 zeigt die Erwerbsquoten für die verschiedenen Gruppen der Schätzstichprobe bzw. der Grundgesamtheit.

[24] Vgl. hierzu Heckman (1976) und Heckman (1979).

[25] Die Erwerbsquote gibt das Verhältnis der erwerbstätigen Personen zur Gesamtbevölkerung im erwerbsfähigen Alter (16-65 Jahre) an.

Tabelle 3: Erwerbsquoten

	Paare		Techn. Singles		
	Männer	Frauen	Männer	Frauen	GG
ZV	81,68%	57,05%	73,83%	50,10%	56,51%
EV	-	-	75,43%	75,20%	57,09%
GG	74,81%	50,70%	58,30%	55,35%	56,78%

Quelle: Eigene Berechnungen auf Basis der SOEP-Welle 2004.

Die Erwerbsquote für Männer (Frauen) innerhalb der Schätzstichprobe liegt bei 78,41% (62,80%) und damit nur marginal unter den Werten, die das Statistische Bundesamt für 2004 ermittelte.[26] Hinsichtlich der Veranlagungsart können in der GG keine gravierenden Unterschiede in der Erwerbsquote festgestellt werden. In der Schätzstichprobe unterscheidet sich der Anteil der erwerbstätigen Single-Männer nur geringfügig vom Anteil der erwerbstätigen verheirateten Männer, wohingegen in der GG eine deutlich höhere Erwerbsquote für verheiratete Männer festgestellt werden kann. Auffällig ist weiterhin, dass vor allem in der Schätzstichprobe aber auch in der GG der Anteil der Erwerbstätigen bei den Single-Frauen höher ist als bei verheirateten Frauen. Eine naheliegende Erklärung hierfür wäre, dass der potentielle Arbeitslohn von verheirateten Frauen durch das Splitting-Verfahren (bei Erwerbstätigkeit des Mannes) mit einem hohen Grenzsteuersatz belastet wird und diese Frauen dann eher von einer Arbeitsaufnahme absehen.

Tabelle 4 gibt einen Überblick über die Verteilung der Arbeitszeitkategorien der (technischen) Single-Haushalte in der Schätzstichprobe. Fast ein Viertel aller Single-Männer ist nicht erwerbstätig. Der überwiegende Teil (69%) der Single-Männer arbeitet zwischen 1 und 40 Stunden pro Woche, wobei die durchschnittliche Wochenarbeitszeit rund 38 Stunden beträgt. Nur 6% haben eine Wochenarbeitszeit von mehr als 40 Stunden. Im Mittel arbeiten diese Single-Männer rund 49 Stunden pro Woche.

Fast jede dritte Single-Frau in der Schätzstichprobe ist nicht erwerbstätig. Ungefähr 22% der Single-Frauen arbeiten in Teilzeit, d.h. nicht mehr als 34 Stunden pro Woche. Für rund 43% aller Single-Frauen beträgt die Wochenarbeitszeit wenigsten 35 und höchsten 40

[26] Die vom Statistischen Bundesamt (2006) ermittelte Erwerbsquote bezogen auf die männlichen (weiblichen) Erwerbspersonen beträgt 79,3% (65,2%).

Tabelle 4: Durchschnittliche Wochenarbeitszeit und relative Verteilung der Single Haushalte auf die Arbeitszeitkategorien

(a) Männer

Arbeitszeitkategorie k	0	1	2
Wochenarbeitszeit	0	1≤40	>40
∅ Wochenarbeitszeit	0	37,65	48,92
Relativer Anteil	24,80%	69,16%	6,03%

(b) Frauen

Arbeitszeitkategorie k	0	1	2	3	4
Wochenarbeitszeit	0	1≤15	16≤34	35≤40	>40
∅ Wochenarbeitszeit	0	10,19	25,03	38,41	48,46
Relativer Anteil	32,73%	4,15%	18,33%	42,85%	1,94%

Quelle: Eigene Berechnungen auf Basis der SOEP-Welle 2004.

Stunden. Nur knapp 2% arbeiten aber mehr als 40 Stunden pro Woche.

Tabelle 5 zeigt die Verteilung der Arbeitszeitkategorien für verheiratete Paare in der Schätzstichprobe. In fast 9% aller Paar-Haushalte sind beide Partner nicht erwerbstätig. Wie schon in Tabelle 3 gezeigt wurde, gibt es in Paar-Haushalten die Tendenz zum Allein-Verdienermodell, wobei der Anteil der erwerbstätigen Ehemänner im Vergleich deutlich größer ist als der Anteil der erwerbstätigen Ehefrauen. Im Vergleich zu den Single-Frauen arbeiten auch relativ mehr Ehefrauen in Teilzeit, d.h. zwischen 1 und 15 bzw. 16 und 34 Stunden pro Woche.

Tabelle 5: Relative Verteilung der Paar-Haushalte auf die Arbeitszeitkategorien

			Frauen					
	Arbeitszeitkategorie k		0	1	2	3	4	
		Wochenarbeitszeit	0	1≤15	16≤34	35≤40	>40	Summe
Männer	0	0	8,77%	2,10%	3,32%	4,14%	0,24%	18,58%
	1	1≤40	30,68%	9,60%	17,88%	14,56%	0,53%	73,25%
	2	>40	3,49%	0,97%	2,04%	1,56%	0,10%	8,16%
		Summe	42,95%	12,67%	23,24%	20,26%	0,87%	100%

Quelle: Eigene Berechnungen auf Basis der SOEP-Welle 2004.

Das durchschnittliche Bruttoeinkommen (incl. Transfers) eines Single-Mannes beträgt im Falle der Erwerbslosigkeit 10.235 Euro und bei Erwerbstätigkeit rund 25.640 Euro im Jahr. Demgegenüber beläuft sich das durchschnittliche Bruttoeinkommen einer Single-Frau bei Erwerbslosigkeit auf 9.588 Euro und bei Erwerbstätigkeit auf rund 18.687 Euro pro Jahr. Ehepaare verfügen bei gemeinsamer Erwerbslosigkeit über ein Bruttoeinkommen von durschnittlich 19.122 Euro im Jahr. Ist nur die Ehefrau erwerbstätig, dann steigt das gemeinsame Bruttoeinkommen marginal auf 19.273 Euro im Jahr an. Arbeitet dagegen nur der Mann, beträgt das gemeinsame Bruttojahreseinkommen im Durchschnitt 32.058 Euro. Wenn beide Ehepartner erwerbstätig sind, erwirtschaften sie gemeinsam durchschnittlich 48.811 Euro jährlich.

4.3 Spezifizierung des Schätzmodells

Zur ökonometrischen Schätzung der Parameter der (direkten) Nutzenfunktion des Haushaltes wird in Stata ein Conditional-Logit-Modell nach McFadden (1973) spezifiziert. Auf das Vorliegen von beobachtbarer Heterogenität wird mittels folgender Fall-spezifischer (Dummy-)Variablen kontrolliert: Alter, Gesundheitsstatus, Nationalität, Schulabschluss, Anzahl und Alter der Kinder im Haushalt und Wohnregion (Ost/West).

Die Spezifikation des Schätzmodells wird für jede Gruppe (Paar-Haushalte, Single-Männer/-Frauen) getrennt vorgenommen. Dabei wird implizit angenommen, dass alle Beobachtungseinheiten innerhalb einer Schätzgruppe ein vergleichbares Einkommens- bzw. Freizeitnutzenkalkül aufweisen. Der Nutzen eines Haushaltes in jeder möglichen Arbeitszeitkategorie wird anschließend durch Einsetzen der geschätzten Parameter und der Variablenwerte in (6) bestimmt. Für jeden Haushalt können dann die bedingten Auswahlwahrscheinlichkeiten der jeweiligen Arbeitszeitkategorien gemäß (4) ermittelt werden.

4.3.1 Schätzergebnisse

Die oben spezifizierte Nutzenfunktion wird in allen drei Schätzgruppen mit der Maximum-Likelihood Methode geschätzt. Die Tabellen 6 und 7 zeigen die geschätzten Parameter der direkten Nutzenfunktion.

Aufgrund der Logarithmierung und der Modellierung von Kreuz- sowie Interaktionstermen von Einkommen und Freizeit können die geschätzten Koeffizienten jedoch nicht als marginale Effekte interpretiert werden. Die Interaktion von *ekn* mit den sozioökonomi-

Tabelle 6: Parameter der Arbeitsangebotsschätzung - Singles

Variable	Koeffizient	Standard-fehler	Koeffizient	Standard-fehler
	Männer		Frauen	
Einkommen				
ln ekn	-6.285	4.933	-79.962***	5.599
$(\ln \text{ekn})^2$	-0.702**	0.339	3.885***	0.318
ln ekn × Ost	4.705***	1.731	1.583	2.289
$(\ln \text{ekn})^2$ × Ost	-0.518***	0.158	-0.026	0.161
ln ekn × Abitur	-5.896***	1.782	-0.924	2.044
$(\ln \text{ekn})^2$ × Abitur	0.601***	0.162	-0.011	0.144
ln ekn × Deutsch	-4.440	3.532	7.952**	3.895
$(\ln \text{ekn})^2$ × Deutsch	0.411	0.320	-0.523*	0.274
Freizeit				
ln l	93.936***	12.916	-40.839***	8.002
$(\ln l)^2$	-13.313***	1.568	-0.170	0.846
ln l × Ost	-0.731	6.183	1.099	3.963
$(\ln l)^2$ × Ost	0.035	0.810	-0.021	0.502
ln l × Deutsch	-10.934	12.417	9.805	6.712
$(\ln l)^2$ × Deutsch	1.370	1.601	-1.231	0.849
ln l × Alter	-0.462***	0.044	-0.296***	0.037
$(\ln l)^2$ × Alter^2	0.000***	0.000	0.000***	0.000
ln l × Abitur	-5.758	6.326	3.468	4.157
$(\ln l)^2$ × Abitur	0.755	0.838	-0.631	0.533
ln l × Erwerbsbehinderung	1.483***	0.296	-0.027	0.283
ln l × Kinder 0-6 Jahre			8.753***	0.983
ln l × Kinder 7-16 Jahre			6.100***	0.517
ln l × Kinder ab 17 Jahre			4.458***	0.320
Interaktion von Einkommen und Freizeit				
ln ekn × ln l	2.905***	0.500	6.662***	0.432
ln ekn × ln l × Anzahl Kinder	-0.216	0.166	0.103**	0.709
ln ekn × ln l × Kinder 0-6 Jahre	0.882	0.489	-0.999***	0.234
ln ekn × ln l × Kinder 7-16 Jahre	0.119	0.375	-0.938***	0.131
ln ekn × ln l × Kinder ab 17 Jahre	0.345	0.285	-0.904***	0.096
Loglikelihood-Wert	-1491		-4220	
Pseudo-R^2	0.4309		0.1494	
Beochbachtungszahl N	7158		15415	

Anmerkungen: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Quelle: Eigene Berechnungen auf Basis von EITDsim.

schen Merkmalen z ist in den beiden Single-Schätzgruppen nur in drei Fällen signifikant: Für alleinstehende Männer mit Abitur hat das Nettoeinkommen bei der Arbeitszeitwahl ceteris paribus einen niedrigeren Stellenwert als bei denen ohne Abitur. Ferner hat für alleinstehende Männer in den neuen Bundesländern das Nettoeinkommen ceteris paribus ein höheres Gewicht bei der Arbeitszeitentscheidung als bei denen in den alten Bundesländern. Darüber hinaus spielt das Nettoeinkommen für deutsche Single Frauen eine stärkere Rolle bei der Arbeitszeitentscheidung als für nicht-deutsche Single Frauen. Die Interaktionsterme von Freizeit und den sozioökonomischen Merkmalen z sind in der Gruppe der Single Männer nur in den Fällen Alter und Erwerbsbehinderung signifikant: Mit zunehmenden

Alter sinkt der Freizeitnutzen, wohingegen er bei einer gesundheitlichen Beeinträchtigung ceteris paribus steigt. Bei Single Frauen spielt das Alter bei der Freizeitabwägung eine ähnliche Rolle wie bei Single Männern. Ferner erhöht das Vorhandensein von Kindern im Haushalt den Wert der Freizeit für Single Frauen. Je jünger die Kinder, desto stärker ist dieser Effekt.

Tabelle 7 zeigt, dass in der Schätzgruppe der Paarhaushalte nur die Interaktion von Einkommen und Wohnregion signifikant ist. D.h. für Ehepaare in den neuen Bundesländern spielt bei der Arbeitsangebotsentscheidung das gemeinsame Nettoeinkommen im Vergleich zu den Ehepaaren in den alten Bundesländern ceteris paribus eine untergeordnete Rolle. Auffällig ist außerdem, dass deutsche Ehefrauen der Freizeit einen höheren Nutzen beimessen als nicht-deutsche Ehefrauen. Die Alterseffekte auf den Freizeitnutzen sind bei Ehepaaren, wie bei den Single Haushalten, negativ. Auch die Präsenz und das Alter der Kinder hat einen ähnlichen Einfluss auf den Freizeitwert von Ehefrauen wie bei Single Frauen. Darüber hinaus sind die Interaktionsterme von Einkommen und Freizeit in allen drei Schätzgruppen positiv und hochsignifikant.

In allen drei Schätzgruppen erfüllen die geschätzten Parameter die notwendige Bedingung für eine sinnvolle ökonomische Nutzenspezifikation, d.h. der marginale Nutzen von Einkommen und Freizeit ist in allen Beobachtungspunkten positiv. Wäre diese Bedingung z.B. in einem Beobachtungspunkten verletzt, würde sich die Auswahlwahrscheinlichkeit dieser Arbeitszeitkategorie bei einer Reformsimulation erhöhen, auch wenn reformbedingt das Einkommen sinkt.[27]

[27] Vgl. Jacobebbinghaus (2006, S. 109).

Tabelle 7: Parameter der Arbeitsangebotsschätzung - Ehepaare

Variable	Koeffizient	Standard-fehler
Einkommen		
ln ekn	-23.259***	6.220
$(\ln \text{ekn})^2$	0.409	0.370
ln ekn × Ost	-6.784**	2.972
$(\ln \text{ekn})^2 \times$ Ost	0.582**	0.232
ln ekn × Abitur_m	-1.373	2.854
$(\ln \text{ekn})^2 \times \text{Abitur}_m$	0.094	0.221
ln ekn × Abitur_f	-0.510	3.208
$(\ln \text{ekn})^2 \times \text{Abitur}_f$	0.054	0.247
ln ekn × Deutsch_m	-1.696	5.607
$(\ln \text{ekn})^2 \times \text{Deutsch}_m$	0.106	0.433
ln ekn × Deutsch_f	-4.008	5.266
$(\ln \text{ekn})^2 \times \text{Deutsch}_f$	0.333	0.409
Freizeit		
ln lm	55.342***	9.325
$(\ln \text{lm})^2$	-9.627***	0.951
ln lf	2.382	7.591
$(\ln \text{lf})^2$	-1.300*	0.752
ln lm × Ost	-7.666	5.484
$(\ln \text{lm})^2 \times$ Ost	0.559	0.636
ln lf × Ost	4.317	4.711
$(\ln \text{lf})^2 \times$ Ost	-1.210**	0.545
ln lm × Deutsch_m	-1.216	7.773
$(\ln \text{lm})^2 \times \text{Deutsch}_m$	0.030	0.973
ln lf × Deutsch_f	26.334***	6.134
$(\ln \text{lf})^2 \times \text{Deutsch}_f$	-3.401***	0.768
ln lm × Alter_m	-0.450***	0.068
$(\ln \text{lm})^2 \times \text{Alter}_m^2$	0.000***	0.000
ln lf × Alter_f	-0.417***	0.069
$(\ln \text{lf})^2 \times \text{Alter}_f^2$	0.000***	0.000
ln lm × Abitur_m	-8.514*	4.863
$(\ln \text{lm})^2 \times \text{Abitur}_m$	0.963	0.620
ln lf × Abitur_f	-0.256	4.591
$(\ln \text{lf})^2 \times \text{Abitur}_f$	-0.055	0.579
ln lm × $\text{Erwerbsbehinderung}_m$	0.944***	0.232
ln lf × $\text{Erwerbsbehinderung}_f$	0.542*	0.312
ln lf × Kinder 0-6 Jahre	6.811***	0.716
ln lf × Kinder 7-16 Jahre	3.382***	0.442
ln lf × Kinder ab 17 Jahre	1.855***	0.282
ln lm × ln lf	1.987***	0.413
ln lm × ln lf × Ost	1.112**	0.547
ln lm × ln lf × Deutsch_m	0.014	0.103
Interaktion von Konsum und Freizeit		
ln ekn × ln lm	3.390***	0.369
ln ekn × ln lf	1.135***	0.284
ln ekn × ln lm × Anzahl Kinder	-0.063*	0.035
ln ekn × ln lm × Kinder 0-6 Jahre	-0.011	0.107
ln ekn × ln lm × Kinder 7-16 Jahre	-0.011	0.093
ln ekn × ln lm × Kinder ab 17 Jahre	0.095	0.073
ln ekn × ln lf × Anzahl Kinder	0.121***	0.024
ln ekn × ln lf × Kinder 0-6 Jahre	-0.101	0.091
ln ekn × ln lf × Kinder 7-16 Jahre	-0.122*	0.071
ln ekn × ln lf × Kinder ab 17 Jahre	-0.146***	0.053
Loglikelihood-Wert	-5747	
Pseudo-R^2	0.2428	
Beochbachtungszahl N	42045	

Anmerkungen: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Quelle: Eigene Berechnungen auf Basis von EITDsim.

4.3.2 Analyse der Prognosegüte des Modells

Für die Evaluation von Arbeitsangebotsreaktionen werden nur jene Haushalte herangezogen, deren geschätzte nutzenmaximale Arbeitszeitkategorie auch der tatsächlich gewählten entspricht. Die Prognosegüte des Modells wird durch einen Vergleich der geschätzten nutzenmaximalen Arbeitszeitkategorie mit der tatsächlich gewählten Arbeitszeitkategorie eines Haushaltes ermittelt. Tabelle 8 zeigt, dass die Voraussagekraft des geschätzten Modells für die drei Schätzgruppen recht unterschiedlich ist. Bei fast 72% der Single-Männer entspricht die geschätzte nutzenmaximale Arbeitszeitkategorie auch der tatsächlich gewählten. Die Prognosegüte des Modells für die Gruppe der Single-Frauen ist im Vergleich zu den Single-Männer zwar deutlich geringer, aber mit mehr als 56% noch durchaus zufriedenstellend. Im Gegensatz dazu kann für Paar-Haushalte nur in knapp 31% aller Fälle durch das geschätzte Modell die tatsächlich gewählte Arbeitszeitkategorie vorausgesagt werden.

Tabelle 8: Prognosegüte

		Techn. Singles	
	Paare	Männer	Frauen
Prognosegüte	30,43%	71,58%	56,57%

Quelle: Eigene Berechnungen auf Basis von EITDsim.

Darüber hinaus muss konstatiert werden, dass bestimmte Arbeitszeitkategorien überhaupt nicht valide prognostiziert werden konnten. Zu diesen Arbeitszeitkategorien gehören $k = 2$ bei Single-Männern, $k = 1$ und $k = 4$ bei Single-Frauen. Für Paar-Haushalte konnten insgesamt 5 von 15 Arbeitszeitkategorien nicht valide vorausgesagt werden.

5 Das Datensatzfusionsmodul in EITDsim

Die Verbindung des Arbeitsangebotsmoduls mit dem Steuer-Simulationsmodul wird in EITDsim mit der Methode des *Statistical Matchings* vollzogen. Das Ziel dieser Datensatzfusion ist es, aus den zwei zugrundeliegenden Datenbeständen diejenigen Beobachtungseinheiten zu identifizieren, welche sich in bestimmten sozioökonomischen Merkmalen möglichst ähnlich sind. Wird ein sogenannter statistischer Zwilling gefunden, können die Informationen kombiniert werden. Dieser neue (artifizielle) Mikrodatensatz dient in EITD-

sim dann zur Evaluation von mittelfristigen Effekten (Zweit-Rundeneffekten) potentieller Politikmaßnahmen.

5.1 Grundlagen

Den Auftakt für die theoretische und praktische Forschung zur Verknüpfung zweier (oder mehrerer) unabhängiger Datensätze setzte Okner (1972), indem er aus dem 1966er US Tax File und dem 1967er Survey of Economic Opportunities eine neue artifizielle Mikrodatenbasis konstruierte. Seitdem hat sich ein außerordentlich breites Literaturspektrum zur Problematik der Datensatzfusion entwickelt.[28] Bei einem Teil der Wissenschaft stößt jedoch die Idee, zwei (oder mehrere) unabhängige Datenbestände zu verknüpfen, auf Skepsis oder Ablehnung.[29] Andererseits hat in den letzten Jahren die theoretische Forschung (im Zusammenspiel mit der immens gestiegenen Rechenkapazität aktueller PCs) zu den Möglichkeiten (und Grenzen) der Datensatzfusion erhebliche Fortschritte gemacht.[30] Darüber hinaus werden auch in der Praxis Datensatzverknüpfungen sehr häufig angewendet. Zum Beispiel entwickelte die GfK (Gesellschaft für Konsum-, Markt- und Absatzforschung) vier Analyseinstrumente auf dem Weg der Datensatzfusion. Ferner wurden integrierte Mikrodatensätze als Basis für Steuersimulationsmodelle in den letzten Jahren u.a. von Bork (2000), Steiner et al. (2005), Fuest et al. (2005), Wagenhals & Buck (2006), Steiner & Wakolbinger (2009) und Peichl et al. (2010) entwickelt.

Die in diesem Beitrag durchgeführte Datensatzfusion ist zuallererst von einer sogenannten exakten Datenzusammenführung abzugrenzen. Von einer exakten Zusammenführung spricht man, wenn Informationen von identischen Beobachtungseinheiten aus getrennten Datensätzen in einem gemeinsamen Datenbestand verschmolzen werden. Die Voraussetzung für eine exakte Zusammenführung ist grundsätzlich das Vorliegen eines eindeutigen Primärschlüssels, z.B. einer Personen-ID. Liegt kein derartiger Schlüssel vor, dann ist die Zusammenführung von Informationen in der Regel nur zwischen (möglichst) ähnlichen Merkmalsträgern möglich.[31] Ähnlichkeit heißt hier, dass auf Basis eines gemeinsamen Variablenvektors z, diejenigen Mikroeinheiten in beiden Datenbeständen identifiziert werden

[28] Einen guten Überblick bieten z.B. Rodgers (1984), Rässler (2002) oder D'Orazio et al. (2006).

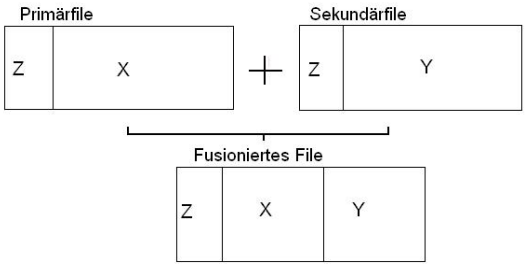
[29] Zum Beispiel bei Sims (1972a), Sims (1972b) oder Rodgers (1984). Hauptkritikpunkt ist die implizite Annahme der bedingten Unabhängigkeit zwischen den niemals gemeinsam beobachtbaren Variablen in den zu fusionierenden Datensätzen für gegebene gemeinsame Variable.

[30] Vgl. Gu & Rosenbaum (1993).

[31] Vgl. Rosenbaum & Rubin (1983).

können, welche sich nur marginal unterscheiden. Das Prinzip ist in Abbildung 2 dargestellt.

Abbildung 2: Prinzip der Datensatzfusion



Quelle: Eigene Darstellung nach Buck (2006).

Für die Durchführung einer Datensatzfusion stehen in der Literatur eine Reihe von Fusionsalgorithmen zur Auswahl, die spezifische Vor- und Nachteile haben.[32] Einige Fusionsverfahren, sog. Covariate-Matching-Methods (CVM), basieren auf Distanzfunktionen zur Messung der Ähnlichkeit zwischen einer Beobachtung i aus dem Primär- und einer Beobachtung j aus dem Sekundärdatensatz. Anhand der Kovariaten z werden dabei zunächst Abstandsmaße berechnet. Die Ermittlung des Fusionspartners erfolgt dann durch einen Vergleich der (skalaren) Maße: Mit, beispielsweise, einem Nearest-Neighbor-Verfahren können die Beobachtungseinheiten aus Primär- und Sekundärdatensatz fusioniert werden, welche einen minimalen Abstand aufweisen. Gängige Distanzfunktionen zur Bestimmung des Fusionspartners stellen z.B. die absolute Distanz (auch City-Block-Metric genannt), die euklidische Distanz, die quadratische Distanz oder die Mahalanobis-Distanz dar. Die Verwendung von Distanzfunktionen erfordert teilweise eine Normierung und eine Gewichtung der Kovariaten z, weil es bei der Berechnung der Abstandsmaße ansonsten zu einer Übergewichtung der Variablen mit dem größten Wertebereich kommen kann.[33] Ein weiterer Nachteil dieser Fusionsalgorithmen ist, dass bei sehr großen Datenbeständen und mit zunehmender Dimension des Kovariatenvektors die Rechenintensität stark zunimmt, weil für jede Beobachtungseinheit des Primärdatensatzes der Abstand zu allen Beobachtungen im Sekundärdatensatz berechnet werden muss. Darüber hinaus sinkt die Wahrscheinlichkeit statistische Zwillinge zu identifizieren mit zunehmender Anzahl der Kovariaten in z.[34]

[32] Eine ausführliche Diskussion verschiedener Matching-Methoden liefert z.B. Zhao (2004).

[33] Dies gilt jedoch nicht für die Mahalanobis-Distanz.

[34] Vgl. Rosenbaum & Rubin (1985), Gu & Rosenbaum (1993).

Eine Alternative zur Verwendung von Distanzfunktionen stellt das Propensity Score-Matching (PSM) dar.[35] Dieser Ansatz wurde zuerst vorn Rosenbaum & Rubin (1983) vorgeschlagen. Zur Bestimmung des Fusionspartners wird im PSM zunächst mittels einer Probit- oder Logitschätzung auf der Basis der Kovariaten Z ein Ähnlichkeitsindex bzw. Propensity Score ermittelt. Anhand dieses (skalaren) Maßes kann dann der Fusionspartner, z.B. mit einem Nearest-Neighbor-Verfahren, bestimmt werden. Bei großen Datensätzen weist dieser Ansatz einige Vorteile gegenüber der Verwendung von Distanzfunktionen auf. Zwar sind auch Propensity Score-Verfahren rechenintensiv, da aber die eigentliche Fusion wiederum nur auf Basis des Propensity Scores vorgenommen wird, relativiert sich bei großen Datensätzen dieser Nachteil. Ein weiterer Vorteil des PSM besteht ferner darin, dass mit dem Schätzverfahren eine implizite Gewichtung der Kovariaten einhergeht.

Mithin ist es aber auch möglich, die zwei oben beschriebenen Matching-Ansätze zu kombinieren. So schlagen beispielsweise Rosenbaum & Rubin (1985) ein Verfahren vor, bei dem zunächst mit einem Logit-Modell auf Basis aller (vorhandenen) unabhängigen Variablen für jede Beobachtungseinheit ein Propensity Score geschätzt wird. Anschließend wird mit einer Teilmenge der Kovariaten ein Mahalanobis-Metric-Matching mit Caliper, auf Basis des geschätzten Propensity Scores, durchgeführt. Rosenbaum & Rubin (1985), Rubin & Thomas (2000) und Baser (2006) zeigen, dass dieser Fusionsansatz den (vergleichsweise) besten Ausgleich zwischen Primär- und Sekundärdatensatz bezüglich der Kovariaten z herstellt, d.h. die Selektionsverzerrung minimiert.

5.2 Umsetzung im Modell

Als (erklärende) Matchingvariablen z werden folgende Kovariate verwendet: Religion (Mann: ef13/Frau: ef14), Bundesland (ef62), Ost-West-Dummy (ef63), Alter (Mann: ef64/Frau: ef67), Alterskategorie (Mann: ef65/Frau: ef68), Kinderanzahl (ef70), Einkommen aus nichtselbstständiger Beschäftigung (Mann: inc1a/ Frau: inc1b/ Summe: inc1) und sonstiges Einkommen (Mann: inc2a/ Frau: inc2b/ Summe: inc2). Vor der Anwendung der Fusionsalgorithmen wird der Datensatz ferner wieder in drei Subfiles geteilt, d.h. es wird eine Selektion entsprechend Kapitel 4.2 vorgenommen.[36]

[35] Eine anwenderorientierte Einführung in das PSM geben Calliendo & Kopeinig (2005).

[36] Nach Buck (2006) ist dieses Vorgehen dann sinnvoll, wenn a priori Informationen vorliegen, dass hinreichende Unterschiede in den Datensätzen existieren, die eine Fusion dieser heterogenen Datensätze nicht ratsam erscheinen lassen. (Vgl. Buck (2006), Kapitel 2.5.) Durch dieses Vorgehen wird somit der potentielle Fehler ausgeschlossen, dass es bei der Datensatzfusion z.B. zu einem Matching von Single-Frauen mit Single-Männern kommt.

Tabelle 9: Mittelwertvergleich, Unmatched, Einzelveranlagte Männer (Subsample I)

Variable	Mean Treated	Mean Control	%bias*	t-test t	t-test p>|t|
ef13	2.6929	2.7215	-2.2	-0.81	0.416
ef62	8.4336	8.2237	5.2	1.86	0.063
ef63	1.3149	1.2741	8.9	3.24	0.001
ef64	32.266	33.6	-12.2	-4.53	0.000
ef65	5.0511	5.3034	-11.4	-4.24	0.000
ef70	0.2132	0.0154	48.2	13.00	0.000
inc1	14946	25780	-67.8	-30.29	0.000
inc2	96.06	244.59	-6.0	-6.90	0.000
$\sum$inc	15042	26025	-68.6	-30.59	0.000

*Durchschnittliche Differenz der Kovariaten von Primär- und Sekundärdatensatz als prozentualer Anteil an der Quadratwurzel der durchschnittlichen Standardabweichung: $100 * (Z_i - Z_j) / [0.5 * (S_{Zi} + S_{Zj})]^{1/2}$, wobei Z_i bzw. Z_j die Mittelwerte und S_{Zi} bzw. S_{Zj} die Standardabweichungen der einzelnen Kovariaten in Primär- bzw. Sekundärdatensatz darstellen.

Quelle: Eigene Berechnungen.

Tabelle 9 zeigt, dass im Subfile der Single-Männer (Subsample I) vor der Datensatzfusion, mit Ausnahme von ef13, deutliche Unterschiede zwischen Merkmalsträgern aus der FAST (Treatment-Gruppe) und dem SOEP (Control-Gruppe) vorliegen.

Tabelle 10: Mittelwertvergleich, Unmatched, Einzelveranlagte Frauen (Subsample II)

Variable	Mean Treated	Mean Control	%bias*	t-test t	t-test p>|t|
ef14	2.558	2.6063	-3.7	-1.16	0.245
ef62	8.2111	8.2177	-0.2	-0.05	0.960
ef63	1.2954	1.2543	9.2	2.82	0.005
ef67	33.453	32.948	4.5	1.41	0.160
ef68	5.2914	5.1801	4.9	1.54	0.125
ef70	0.3395	0.0926	46.4	11.55	0.000
inc1	14231	22222	-64.4	-21.85	0.000
inc2	70.181	20.753	9.5	2.43	0.015
$\sum$inc	14301	22243	-63.9	-21.65	0.000

*Durchschnittliche Differenz der Kovariaten von Primär- und Sekundärdatensatz als prozentualer Anteil an der Quadratwurzel der durchschnittlichen Standardabweichung: $100 * (Z_i - Z_j) / [0.5 * (S_{Zi} + S_{Zj})]^{1/2}$, wobei Z_i bzw. Z_j die Mittelwerte und S_{Zi} bzw. S_{Zj} die Standardabweichungen der einzelnen Kovariaten in Primär- bzw. Sekundärdatensatz darstellen.

Quelle: Eigene Berechnungen.

Tabelle 10 zeigt, dass auch im Subfile der Single-Frauen (Subsample II) signifikante Differenzen in den Mittelwerten der Kovariaten hervortreten. Lediglich in ef62 ähneln sich Treatment- und Control-Gruppe. Im Subfile der zusammenveranlagten Paare (Subsample III) weisen sogar alle Kovariate der Treatment- und Control-Gruppe in den Mittelwerten signifikante Differenzen auf (Tabelle 11).

Nach Durchführung der Datensatzfusion sollten diese Differenzen möglichst nicht mehr vorliegen.

Tabelle 11: Mittelwertvergleich, Unmatched, Zusammenveranlagte Paare (Subsample III)

Variable	Mean Treated	Control	%bias*	t-test t	p>|t|
ef13	2.7835	2.5218	20.1	9.66	0.000
ef14	2.6625	2.3946	20.8	9.82	0.000
ef62	8.6101	7.8491	18.7	8.46	0.000
ef63	1.3478	1.2307	26.0	11.81	0.000
ef64	46.594	48.243	-15.1	-7.29	0.000
ef65	7.9186	8.2445	-14.8	-7.16	0.000
ef67	43.632	45.419	-17.2	-8.28	0.000
ef68	7.3281	7.6792	-16.8	-8.04	0.000
ef70	1.0123	0.7857	22.3	10.41	0.000
inc1	22797	37392	-60.0	-41.78	0.000
inc2	1872.6	3969.9	-17.3	-14.79	0.000
$\sum$inc	24669	41362	-65.6	-47.38	0.000

*Durchschnittliche Differenz der Kovariaten von Primär- und Sekundärdatensatz als prozentualer Anteil an der Quadratwurzel der durchschnittlichen Standardabweichung: $100 * (Z_i - Z_j) / [0.5 * (S_{Zi} + S_{Zj})]^{1/2}$, wobei Z_i bzw. Z_j die Mittelwerte und S_{Zi} bzw. S_{Zj} die Standardabweichungen der einzelnen Kovariaten in Primär- bzw. Sekundärdatensatz darstellen.

Quelle: Eigene Berechnungen.

Da ex ante nicht bestimmbar ist, welches Verfahren optimal für die Fusion der vorliegenden Datensätze wäre, werden zunächst drei verschiedene Fusionsalgorithmen durchgeführt und die Fusionsergebnisse in einer Qualitätsanalyse miteinander verglichen.[37] Folgende Fusionsverfahren werden nachfolgend evaluiert:

- (M1) - PSM: Nearest-Neighbor-Matching auf Propensity-Score-Basis
- (M2) - CVM: Mahalanobis-Metric-Matching mit Caliper
- (M3) - PSM/CVM: Mahalanobis-Metric-Matching (inkl. Propensity-Score) mit Caliper auf Propensity-Score-Basis

Letztendlich wird die Zusammenführung von FAST und SOEP in EITDsim dann mit dem Verfahren durchgeführt, welches die (vergleichsweise) geringste Selektionsverzerrung aufweist.

Die Datenfusion wird mit der Statistiksoftware STATA und dem Modul PSMATCH2 durchgeführt.[38] Dafür ist gemäß Gleichung (12) zunächst in den Datensatz eine künstliche Treatment-Variable S einzuführen. Diese nimmt den Wert 1 an, wenn die Beobachtungseinheit im Primärdatensatz FAST enthalten ist (Treament-Gruppe) und den Wert 0, falls der Merkmalsträger Teil des Sekundärdatensatzes SOEP (Control-Gruppe) ist. Anhand dieser

[37] In der Literatur finden sich keine Anhaltspunkte für die Existenz eines überlegenen Fusionsalgorithmus. Vgl. S. 380 in Baser (2006).

[38] Für eine ausführliche Dokumentation des Moduls PSMATCH2 siehe Leuven & Sianesi (2003).

Treatment-Variablen S kann PSMATCH2 unterscheiden, zu welchem Ursprungsdatensatz der Merkmalsträger gehört.

$$S_i = \begin{cases} 1, \text{ falls die Beobachtung Teil der FAST ist und} \\ 0, \text{ falls die Beobachtung Teil des SOEP ist.} \end{cases} \tag{12}$$

In der Literatur finden sich verschiedene Ansichten zur Frage, ob ein vergleichsweise großer Datensatz wie die FAST eher als Primär- oder Sekundärdatensatz im Matchingprozess verwendet werden sollte. So ist z.B. Rässler (2002) der Ansicht, ein großer Datensatz sollte eher als Primärdatensatz verwendet werden, damit keine Informationen außen vor bleiben.[39] Demgegenüber meinen D'Orazio et al. (2006), dass ein vergleichsweise großer Datensatz eher als Sekundärdatensatz verwendet werden sollte, weil damit die Wahrscheinlichkeit für gute Fusionsergebnisse steigen würde.

Da die FAST sich v.a. aufgrund ihrer informationellen Tiefe hinsichtlich verschiedener Einkommensquellen, Sonderausgaben, außergewöhnliche Belastungen etc. als Datenbasis für steuerorientierte MSM empfiehlt, wird in EITDsim FAST als Primär- und das SOEP als Sekundärdatenbasis verwendet.

(M1) - PSM: Nearest-Neighbor-Matching auf Propensity-Score-Basis Auf Basis der Kovariaten Z wird für jede der drei Subfiles eine Schätzung der Propensity Scores durchgeführt. Der Propensity Score $e(z_i)$ für Beobachtungseinheit i gibt dabei die bedingte Wahrscheinlichkeit dafür an, dass i, gegeben die Kovariate z_i, dem Primärdatensatz $(S = 1)$ angehört.

$$e(z_i) = P(S_i = 1 | Z_i = z_i) \tag{13}$$

Für die Schätzung der Propensity Scores wird ein Logit-Modell verwendet. Die geschätzten Koeffizienten und deren Signifikanzniveaus finden sich in den Tabellen 16, 14 und 15 im Appendix. Zur Bestimmung des Fusionspartners werden dann die geschätzten Propensity Scores $\hat{e}(z)$ verglichen. Dabei wird für jede Beobachtungseinheit des Primärdatensatzes ein Partner aus dem Sekundärdatensatz gesucht, dessen Propensity Score sich nur minimal unterscheidet (Nearest-Neighbor-Verfahren).

$$d(i, j) = |\hat{e}(z_j) - \hat{e}(z_i)| \tag{14}$$

Die Beobachtungseinheit j des Sekundärdatensatzes, welcher eine minimale Distanz $d(i, j)$ aufweist, wird als Fusionspartner für die Beobachtungseinheit i der des Primärdatensatzes

[39] Vgl. S. 18 in Rässler (2002).

ausgewählt und i wird aus dem Matching-Pool entfernt. Dieser Prozess wird solange wiederholt, bis für jeden Merkmalsträger i des Primärdatensatzes ein hinreichend ähnlicher Fusionspartner j aus dem Sekundärdatensatz gefunden wurde. Durch die Berücksichtigung der Common Support Bedingung wird ferner sichergestellt, dass nur jene Beobachtungseinheiten des Primärdatensatzes FAST bei der Fusion berücksichtigt werden, für die auch ein passender Partner im Sekundärdatensatz gefunden werden kann. D.h. es bleiben alle Beobachtungseinheiten des Primärdatensatzes FAST bei der Fusion außen vor, bei denen der geschätzte Propensity Score höher als das Maximum bzw. niedriger als das Minimum der geschätzten Propensity Scores der SOEP-Beobachtungseinheiten ist.

(M2) - CVM: Mahalanobis-Metric-Matching Für das Mahalnobis-Metric-Matching werden zunächst alle Beobachtungseinheiten in den Subfiles in eine zufällige Ordnung gebracht und anschließend wird die Distanz von der ersten Beobachtungseinheit der Treatment-Gruppe (FAST) zu allen Beobachtungseinheiten des Control-Gruppe (SOEP) bestimmt. Die Mahalanobis-Distanz einer Beobachtungseinheit i der Treatment-Gruppe zu einer Beobachtungseinheit j der Control-Gruppe entspricht dabei

$$d(i,j) = (u-v)^T C^{-1}(u-v), \tag{15}$$

wobei u und v die Werte von der verwendeten Matchingvariablen enthält und C die Kovarianz-Matrix der Matchingvariablen aller Beobachtungseinheiten des Sekundärdatensatzes darstellt. Die Beobachtungseinheit j des Sekundärdatensatzes SOEP, welche eine minimale Distanz $d(i,j)$ aufweist, wird als Fusionspartner für die Beobachtungseinheit i der des Primärdatensatzes FAST ausgewählt und i wird aus dem Matching-Pool entfernt. Dieser Prozess wird solange wiederholt, bis für jede Beobachtungseinheit i des Primärdatensatzes ein hinreichend ähnlicher Fusionspartner j aus dem Sekundärdatensatz gefunden wurde. Das Mahalanobis-Metric-Matching wird für jede der drei Subfiles auf Basis der Kovariaten Bundesland (ef62), Alter (ef64, ef67), Anzahl der Kinder (ef70) und Einkommen (inc). Darüber hinaus wird zur Verbesserung der Fusionsergebnisse ein Caliper i.H.v. 0,1 verwendet. Der Caliper definiert für jede Beobachtungseinheit einen Common Support Bereich, d.h. potentielle Fusionspartner mit einer Distanz außerhalb dieses Bereichs werden als Fusionspartner ausgeschlossen.[40]

(M3) - PSM/CVM: Mahalanobis-Metric-Matching (inkl. Propensity-Score) mit Caliper auf Propensity-Score-Basis Bei diesem Verfahren werden die beiden oben be-

[40] Vgl. S. 10 in Calliendo & Kopeinig (2005).

schriebenen Algorithmen kombiniert. D.h. es wird Mahalanobis-Metric-Matching auf Basis der Kovariaten $z^m = \{$ef62, ef64, ef67, ef70, inc1, inc2$\}$ sowie dem Logarithmus des zuvor geschätzten Propensity Score,

$$\hat{q}(z) = \log\left(\frac{1-\hat{e}(z)}{\hat{e}(z)}\right), \tag{16}$$

durchgeführt. Dem Ansatz von Rosenbaum & Rubin (1985) folgend, wird zur Verbesserung der Fusionsergebnisse wiederum ein Caliper, diesmal auf Basis des geschätzten Propensity Scores, von 0,1 verwenddet. D.h. liegt der geschätzte Propensity Score von SOEP-Beobachtungseinheit j außerhalb dieses Bereichs, wird j als möglicher Fusionspartner für FAST-Beobachtungseinheit i ausgeschlossen.

5.3 Qualitätsanalyse der Matchingergebnisse

Um zu entscheiden, welches Fusionsverfahren für die Zusammenführung von FAST und SOEP letztlich am geeignetsten ist, werden, dem Ansatz von Rosenbaum & Rubin (1985) bzw. Baser (2006) folgend, nach dem Matching vier Kriterien geprüft:

(1) *Mittelwertvergleich*: Mit zweiseitigen t-Tests wird überprüft, ob sich die Mittelwerte der einzelnen Kovariaten in Primär- und Sekundärdatensatz unterscheiden.

(2) *Standardisierte mittlere Differenz*: Berechnung der durchschnittlichen Differenz der Kovariaten von Primär- und Sekundärdatensatz als prozentualer Anteil an der Quadratwurzel der durchschnittlichen Standardabweichung:

$$\frac{100 * (Z_i - Z_j)}{[0.5 * (S_{Zi} + S_{Zj})]^{1/2}}, \tag{17}$$

wobei Z_i bzw. Z_j die Mittelwerte und S_{Zi} bzw. S_{Zj} die Standardabweichungen der einzelnen Kovariaten in Primär- bzw. Sekundärdatensatz darstellen.[41]

(3) *Prozentualer Mittelwertangleich*: Ermittlung der prozentualen Angleichung (oder Verzerrungsreduzierung) in den Mittelwerten der Kovariaten vor und nach der Fusion:

$$\frac{(Z_{Ni} - Z_{Nj}) - (Z_{Vi} - Z_{Vj})}{Z_{Vi} - Z_{Vj}} * 100, \tag{18}$$

wobei Z_{Vi} bzw. Z_{Vj} die Mittelwerte der Kovariaten in Primär- respektive Sekundärdatensatz vor der Fusion und Z_{Ni} bzw. Z_{Nj} die Mittelwerte der Kovariaten in Primär- bzw. Sekundärdatensatz nach der Fusion sind.

[41] Vgl. S. 34 in Rosenbaum & Rubin (1985).

(4) *Dichteschätzung der Kovariaten*: Mit einem Kalmogorov-Smirnov-Test bei kontinuierlichen Variablen bzw. einem χ^2-Homogenitätstest bei kategorialen Variablen wird überprüft, ob die Kovariaten in Primär- bzw. Sekundärdatensatz die gleiche Dichteverteilung aufweisen.

Diese vier Kriterien können einen Aufschluss darüber geben, welche qualitative Güte die verwendeten Fusionsalgorithmen haben. Tabelle 12 zeigt die Ergebnisse der Qualitätsanalyse für alle drei Subsamples (Subsample I-III) und alle drei Fusionsalgorithmen (M1-M3).

Der Fusionsalgorithmus M1, also das Nearest-Neighbor-Matching auf Basis eines empirisch geschätzten (skalaren) Propensity Scores, ist in allen drei Subsamples die qualitativ schlechteste Matchingmethode. Nach Kriterium (1) weisen alle Kovariate von Treatment- und Control-Gruppen nach dem Matching signifikante Unterschiede auf. Lediglich in Subsample II zeigt ein Dichtevergleich (Kriterium (4)) für ef14, ef67, ef68 und inc2 insignifikante Unterschiede. Die Ergebnisse legen die Vermutung nahe, dass mit M1 auch weniger exakte Matches vollzogen werden. D.h. es werden auf Basis des geschätzten Propensity Scores mithin auch Merkmalsträger aus Treatment- und Control-Gruppe als statistische Zwillinge identifiziert, die sich hinsichtlich ihrer sozioökonomischen Eigenschaften deutlich unterscheiden.

Die Fusionsergebnisse der Methoden M2 und M3 unterscheiden sich qualitativ in den drei Subsamples nur wenig. Dem Fusionsalgorithmus M3 kann jedoch eine, v.a. im Hinblick auf Kriterium (1), leicht bessere Performance attestiert werden. Mithin weisen deutlich mehr Kovariate der Treatment- und Control-Gruppen nach dem Matching mit M3 insignifikante Unterschiede hinsichtlich der Mittelwerte auf als nach dem Matching mit M2. Die standardisierten mittleren Differenzen der Kovariaten (Kriterium (2)) sind für M2 und M3 nahezu identisch, nur in Subsample III (Zusammenveranlagte Paare) fallen die standardisierten mittleren Differenzen der Kovariate nach dem Matching mit M3 etwas geringer aus als nach M2. Ähnliches gilt für die Bewertung der Fusionsergebnisse nach Kriterium (3). Sowohl M2 als auch M3 reduzieren die ex ante vorliegenden mittleren Unterschiede in den Kovariaten von Treatment- und Control-Gruppen nahezu gleich stark, wobei aber M3 den (marginal) stärkeren Effekt hat. Darüber hinaus zeigt auch ein Vergleich der Ergebnisse der Dichteschätzungen (Kriterium (4)) nur geringfügige Unterschiede zwischen M2 und M3 in den Subsamples II und III. Nur im Subsample I (Ledige Männer) werden nach dem Matching mit M3 (im Gegensatz zu M2) auch für die Verteilungen von ef62, ef63, ef64 und ef65 insignifikante Unterschiede ermittelt.

Tabelle 12: Qualitätsanalyse der Matchingergebnisse

	Subsample I			Subsample II			Subsample III		
Variable	M1	M2	M3	M1	M2	M3	M1	M2	M3
(1) p-Werte der Mittelwerte									
ef13	0.000	0.000	0.000	-	-	-	0.000	0.000	0.000
ef14	-	-	-	0.000	0.000	0.030	0.000	0.000	0.004
ef62	0.000	0.866	0.861	0.000	0.677	0.962	0.000	0.783	0.848
ef63	0.000	0.599	1.000	0.000	0.560	0.466	0.000	0.872	1.000
ef64	0.000	0.299	0.536	-	-	-	0.000	0.119	0.943
ef65	0.000	0.000	0.103	-	-	-	0.000	1.000	1.000
ef67	-	-	-	0.000	0.257	0.521	0.000	0.281	0.816
ef68	-	-	-	0.000	0.012	1.000	0.000	1.000	1.000
ef70	0.000	1.000	1.000	0.000	1.000	1.000	0.000	1.000	1.000
inc1	0.000	0.830	0.790	0.000	0.902	0.622	0.000	0.011	0.728
inc2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.063	0.000
$\sum$inc	0.000	0.812	0.802	0.000	0.893	0.618	0.000	0.009	0.681
(2) Standardisierte mittlere Differenz in %									
ef13	5.6	-2.2	2.4	-	-	-	7.5	18.0	6.3
ef14	-	-	-	9.4	-7.7	-0.8	7.6	11.2	3.2
ef62	1.3	0.0	0.1	15.1	-0.1	0.0	9.0	0.4	0.2
ef63	-3.7	-0.1	0.0	11.0	-0.2	-0.3	11.7	-0.2	0.0
ef64	6.5	0.2	0.2	-	-	-	4.8	-1.9	-0.1
ef65	6.6	0.9	0.5	-	-	-	5.3	0.0	0.0
ef67	-	-	-	14.2	0.3	0.2	6.4	-1.3	0.3
ef68	-	-	-	15.1	0.7	0.0	6.5	0.0	0.0
ef70	0.8	0.0	0.0	15.4	0.0	0.0	3.8	0.0	0.0
inc1	4.6	0.0	-0.1	9.6	0.0	0.2	2.5	1.8	0.3
inc2	1.7	0.0	0.0	-4.7	0.1	0.0	2.3	0.1	0.1
$\sum$inc	4.9	0.1	-0.1	9.3	0.0	0.2	3.4	1.7	0.3
(3) Prozentualer Mittelwertangleich									
ef13	-156.8	0.4	-10.2	-	-	-	62.8	10.7	68.8
ef14	-	-	-	-152.5	-106.3	78.3	63.6	45.9	84.4
ef62	75.3	99.1	98.8	-8960.2	17.7	89.7	51.9	98.0	98.9
ef63	58.6	98.4	100.0	-18.9	98.0	97.2	55.0	99.1	100.0
ef64	46.8	98.1	98.6	-	-	-	68.0	87.4	99.4
ef65	41.8	92.2	96.0	-	-	-	64.4	100.0	100.0
ef67	-	-	-	-214.5	93.3	95.8	62.8	92.3	98.4
ef68	-	-	-	-206.2	86.2	100.0	61.2	100.0	100.0
ef70	98.4	100.0	100.0	66.8	100.0	100.0	82.8	100.0	100.0
inc1	93.2	99.9	99.9	85.2	99.9	99.8	95.9	97.0	99.5
inc2	71.1	99.5	99.5	50.8	99.2	99.6	87.0	99.5	99.4
$\sum$inc	92.9	99.9	99.9	85.4	99.9	99.8	94.8	97.3	99.5
(4) p-Werte der Dichteschätzungen									
ef13	0.392	0.000	0.000	-	-	-	0.000	0.000	0.000
ef14	-	-	-	0.484	0.000	0.000	0.000	0.000	0.000
ef62	0.000	0.000	0.000	0.000	0.012	0.000	0.000	0.000	0.000
ef63	0.005	0.039	0.039	0.008	0.000	0.000	0.000	0.000	0.304
ef64	0.000	0.000	0.000	-	-	-	0.000	0.000	0.000
ef65	0.000	0.000	0.000	-	-	-	0.000	0.000	0.000
ef67	-	-	-	0.469	0.000	0.000	0.000	0.000	0.000
ef68	-	-	-	0.481	0.000	0.000	0.000	0.000	0.000
ef70	0.000	0.999	0.996	0.000	0.035	0.018	0.000	0.000	0.000
inc1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
inc2	0.080	1.000	1.000	0.891	1.000	1.000	0.000	0.876	0.505
$\sum$inc	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Quelle: Eigene Berechnungen auf Basis von EITDsim.

Die Ergebnisse der Qualitätsanalyse in Tabelle 12 zeigen, dass sich das Verfahren M3, eine Kombination aus Propensity Score Matching und Mahalanobis-Metric-Matching, deutlich vom Verfahren M1 und leicht vom Verfahren M2 abhebt. D.h. die Wahrscheinlichkeit einen statistischen Zwilling aus der FAST und dem SOEP zu identifizieren ist mit dem

Fusionsalgorithmus M3 größer als bei Anwendung eines der anderen beiden Verfahren.

6 Zusammenfassung und Ausblick

Mit diesem Beitrag wurde der Aufbau und die Funktionsweise des integrierten Mikrosimulationsmodells EITDsim dokumentiert. EITDsim stellt ein flexibel einsetzbares Instrument zur empirischen ex ante Evaluation und Analyse von (potentiellen) finanz- und sozialpolitischen Politikmaßnahmen dar. Dafür kombiniert EITDsim auf Basis des erweiterten Mikrodatensatzes FAST ein statisches Steuermikrosimulationsmodell zur Schätzung kurzfristiger Erst-Rundeneffekte mit einem diskreten Arbeitsangebotsmodell zur Schätzung mittelfristiger Zweit-Rundeneffekte. Aufgrund unvollständiger Informationen in der FAST wird als Datengrundlage für das Arbeitsangebotsmodell in EITDsim das Sozioökonomische Panel (SOEP) des DIW verwendet. Auf dem Wege einer Datensatzfusion werden sodann die Informationen aus beiden Mikrodatensätzen kombiniert. Hierfür werden zunächst drei verschiedene Fusionsalgorithmen hinsichtlich ihrer zu erwartenden Gütequalität evaluiert. Die Analyse der Fusionsergebnisse zeigt, dass ein kombiniertes Verfahren aus Propensity Score Matching und Mahalanobis-Metric Matching am ehesten dazu geeignet ist, aus den Beobachtungseinheiten der FAST und des SOEP statistische Zwillinge zu identifizieren.

Eine zukünftig denkbare und auch sinnvolle Modul-Erweiterung für EITDsim würde zum Einen die Modellierung einer endogenen Arbeitsnachfrage und zum Anderen die Konzeption eines Wohlfahrtsmoduls zur Evaluation von Wohlfahrtswirkungen (potentieller) Politikmaßnahmen beinhalten. Darüber hinaus stellt sich grundsätzlich die Frage, ob, in Bezug auf die Bevölkerungsentwicklung und die Einkommen, eine Anpassung der Datengrundlagen bis an den aktuellen Rand möglich und zielführend wäre.

Auf Basis des hier dargelegten Entwicklungsstandes kann EITDsim jedoch bereits jetzt als Instrument für die wissenschaftsorientierte Politikberatung eingesetzt werden.

Literatur

Bach, S., Corneo, G., & Steiner, V. (2009). *From Bottom to Top: The Entire Income Distribution in Germany, 1992-2003.* Review of Income and Wealth. Nr. 55, S. 303-330.

Baroni, E. & Richiardi, M. (2007). *Orcutt's Vision, 50 years on.* LABORatorio R. Revelli. Working Paper, Nr. 65.

Baser, O. (2006). *Too much ado about propensity score models? Comparing methods of propensity score matching.* Value in health the journal of the International Society for Pharmacoeconomics and Outcomes Research. Vol. 9 , Nr. 6, S. 377-385.

Blundell, R. & MaCurdy, T. (1999). Labor Supply: A Review of Alternative Approaches. In: Ashenfelter, O. & Card, D., (Hrsgb.), *Handbook of Labor Economics.* Elsevier. Vol. 3A, S. 1559-1695.

Bönke, T., Neher, F., & Schröder, C. (2007). *Bestimmung ökonomischer Einkommen und effektiver Einkommensteuerbelastungen mit der Faktisch Anonymisierten Lohn- und Einkommensteuerstatistik.* Schmollers Jahrbuch. 127 (4), S. 585-623.

Bork, C. (2000). *Steuern, Transfers und private Haushalte: Eine mikroanalytische Simulationsstudie der Aufkommens- und Verteilungswirkungen.* Peter Lang Verlag, Frankfurt/Main.

Brenneisen, F. & Peichl, A. (2007). *Dokumentation des Wohlfahrtsmoduls von FiFoSiM.* Universität zu Köln.

Buck, J. (2006). *Datenfusion und Steuersimulation - Theorie und Empirie im Rahmen des Mikrosimulationsmodells GMOD.* Shaker Verlag, Aachen.

Burtless, G. & Hausman, J. (1978). *The Effect of Taxation on Labor Supply: Evaluating the Gary Negative Income Tax Experiment.* Journal of Political Economy. Nr. 86(6), S. 1103-1130.

Calliendo, M. & Kopeinig, S. (2005). *Some Practical Guidance for the Implementation of Propensity Score Matching.* IZA Bonn. Discussion Paper Series, Nr. 1588.

Creedy, J. & Duncan, A. (2002). *Behavioural Microsimulation with Labour Supply Responses.* Journal of Economic Surveys. Vol. 16, Nr.1, S. 1-39.

D'Orazio, M., Di Zio, M., & Scanu, M. (2006). *Statistical Matching: Theory and Practice (Wiley Series in Survey Methodology).* John Wiley & Sons.

Flood, L. & MaCurdy, T. (1992). *Work disincentive effects of taxes: An empirical analysis of Swedish men.* Carnegie-Rochester Conference Series on Public Policy. Nr. 37, S. 239-277.

Flory, J. & Stöwhase, S. (2010). *MIKMOD-ESt- A Static Microsimulation Model for the Evaluation of Personal Income Taxation in Germany.* Preliminary Version.

Forschungsdatenzentrum (2008). *Beschreibung Lohn- und Einkommensteuerstatistik.* Forschungsdatenzentrum-Website, `http://www.forschungsdatenzentrum.de/bestand/lest/index.asp`, zuletzt besucht am 30.07.2008.

Franz, W., Guertzgen, N., Schubert, S., & Clauss, M. (2007). *Reformen im Niedriglohnsektor: Eine integrierte CGE-Mikrosimulationsstudie der Arbeitsangebots- und Beschäftigungseffekte.* ZEW Discussion Papers. Nr. 07-085.

Fuest, C., Peichl, A., & Schaefer, T. (2005). *Dokumentation FiFoSiM: Integriertes Steuer-Transfer-Mikrosimulations- und CGE-Modell.* Universität zu Köln. Finanzwissenschaftliche Diskussionsbeiträge Nr. 05-3.

Gu, X. S. & Rosenbaum, P. R. (1993). *Comparison of Multivariate Matching Methods: Structures, Distances, and Algorithms.* Journal of Computational and Graphical Statistics. Vol. 2, Nr. 4, S. 405-420.

Haisken-DeNew, J. & Frick, J. (2011). *DTC - Desktop Companion to the German Socio-Economic Panel (SOEP).* DIW-Website: `http://www.diw.de/documents/dokumentenarchiv/17/diw_01.c.38951.de/dtc.409713.pdf`, Zuletzt aufgerufen am 30.05.2012.

Hall, R. E. (1973). Wages, Incomes and Hours of Work in the US Labor Force. In: Cain, C. & Watts, H., (Hrsgb.), *Income Maintence and Labour Supply.* Institute for Research on Poverty Monography Series, Academic Press, London, UK. S. 102-162.

Heckman, J. (1976). *The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models.* Annals of Economic and Social Measurement. Nr. 5, S. 475-492.

Heckman, J. (1979). *Sample Selection Bias as a Specification Error.* Econometrica. Nr. 47, S. 153-161.

Jacobebbinghaus, P. (2006). *Steuer-Transfer-Mikrosimulation als Instrument zur Bestimmung des Einflusses von Steuern und Transfers auf Einkommen und Arbeitsangebot einzelner Haushalte.* Dissertation, Universität Bielefeld.

Leuven, E. & Sianesi, B. (2003). Psmatch2: Stata module to perform full mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing. Statistical Software Components, Boston College Department of Economics.

McFadden, D. (1973). *Conditional Logit Analysis of Qualitative Choice Behaviour.* Frontiers in Econometrics. Nr. 1(2), S. 105-142.

Okner, B. (1972). Constructing a new data base from existing microdata sets: The 1966 merge file. In: *Annals of Economic and Social Measurement, Volume 1, number 3*, NBER Chapters. National Bureau of Economic Research, Inc. S. 325-341.

Orcutt, G. (1957). *A new type of socio-economic system.* Review of Economics and Statistics. Nr. 39(2), S. 116-123.

Peichl, A., Schneider, H., & Siegloch, S. (2010). *Documentation IZAΨMOD: The IZA Policy SImulation MODel.* IZA Bonn. IZA Discussion Paper Nr. 4865.

Rodgers, W. L. (1984). *An Evaluation of Statistical Matching.* Journal of Business & Economic Statistics. Vol. 2, Nr. 1, S. 91-102.

Rosenbaum, P. R. & Rubin, D. B. (1983). *The Central Role of the Propensity Score in Observational Studies for Causal Effects.* Biometrika. Vol. 70, S. 41-55.

Rosenbaum, P. R. & Rubin, D. B. (1985). *Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score.* The American Statistician. Vol. 39, S. 33-38.

Rässler, S. (2002). *Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches.* Lecture Notes in Statistics. Springer.

Rubin, D. B. & Thomas, N. (2000). *Combining Propensity Score Matching With Additional Adjustments for Prognostic Covariates.* Journal of the American Statistical Association. Vol. 95, Nr. 450, S. 573-585.

Sims, C. A. (1972a). Comments. In: *Annals of Economic and Social Measurement, Volume 1, number 3*, NBER Chapters. National Bureau of Economic Research, Inc. S. 343-345.

Sims, C. A. (1972b). Rejoinder. In: *Annals of Economic and Social Measurement, Volume 1, number 3*, NBER Chapters. National Bureau of Economic Research, Inc. S. 355-357.

Statistisches Bundesamt (2006). *Datenreport 2006 - Zahlen und Fakten über die Bundesrepublik Deutschland.* Statistisches Bundesamt, Bonn.

Steiner, V., Haan, P., & Wrohlich, K. (2005). *Dokumentation des Steuer-Transfer-Mikrosimulationsmodells STSM 1992-2002.* DIW Berlin. Data Documentation 9.

Steiner, V. & Wakolbinger, F. (2009). *The Austrian Tax Transfer Modell ATTM, Version 1.1.* ATTM Research.

Steiner, V. & Wrohlich, K. (2004). *Household Taxation, Income Splitting and Labor Supply Incentives - A Microsimulation Study for Germany.* DIW Berlin. Discussion Papers Nr. 421.

Stern, N. (1986). On the specification of labour supply functions. In: Blundell, R. & Walker, I., (Hrsgb.), *Unemployment, search and labour supply.* Cambridge University Press, Cambridge, UK. S. 143-189.

Struch, G. (2012). *Eine verteilungspolitische Beurteilung aktueller Reformvorschläge zur deutschen Einkommensbesteuerung.* Journal of Economics and Statistics. i.E.

Struch, G. & Jenderny, K. (2010). *Dokumentation des Extended Income Tax Dataset EITD.* Universität Passau. Working Paper.

Van Soest, A. (1995). *Structural Models of Family Labor Supply - A Discret Choice Approach.* The Journal of Human Ressources. Nr. 30, S. 63-88.

Van Soest, A. & Das, M. (2000). *Family Labor Supply and Proposed Tax Reforms in the Netherlands.* Tilburg University, Center for Economic Research. Nr. 2000-20.

Van Soest, A. & Euwals, R. (1999). *Desired and Actual Labour Supply of Unmarried Men and Women in the Netherlands.* Labour Economics. Nr. 6, S. 95-118.

Van Soest, A., Woittiez, I., & Kapteyn, A. (1990). *Labor Supply, Income Taxes and Hours Restrictions in The Netherlands.* The Journal of Human Ressources. Nr. 25, S. 517-558.

Wagenhals, G. & Buck, J. (2006). GMOD+: An Innovative Tax-Benefit Microsimulation Modeling Tool. In: Borutzky, W., Orsoni, A., & Zobel, R., (Hrsgb.), *Proceedings 20th European Conference on Modelling and Simulation.*, ECMS. S. 354-359.

Zhao, Z. (2004). *Using Matching to Estimate Treatment Effects: Data Requirements, Matching Metrics, and Monte Carlo Evidence.* The Review of Economics and Statistics. Vol. 86, Nr. 1, S. 91-107.

Appendix

Tabelle 13: Lohnregression mit Selektionskorrektur

	Singles		Ehepaare	
	Männer	Frauen	Männer	Frauen
Lohngleichung				
20≤Alter<25	-26.047***	-5.217***		
25≤Alter<30	-19.076***	-4.124***	-14.357***	-1.771
30≤Alter<35	-16.165***	-4.144***	-16.093***	-0.385
35≤Alter<40	-14.097***	-4.071***	-13.231***	-0.016
40≤Alter<45	-10.798**	1.709**	-10.459***	0.780
45≤Alter<50	-7.769**	1.445*	-7.782***	1.353
50≤Alter<55	-5.539	2.085***	-5.885***	1.932**
55≤Alter<60	-5.156*	1.943**	-3.994***	1.659*
VZ-Beschäftigung	-0.071	0.111***	-0.216***	0.125***
ohne Beschäftigung	-1.365***	-0.595***	-1.250***	-0.546***
Abitur	1.890	5.609***	9.924***	8.625***
Fachabitur	0.659***	4.278	7.518***	4.286***
Realschule	-2.038	1.443**	1.701	2.521***
Hauptschule	-1.132	-0.308	-1.136	-0.425
BL: SH	-0.550	1.894	4.967***	2.796*
BL: HH	-1.940	3.470**	8.200***	5.163**
BL: N	2.192	2.159**	4.726***	2.358*
BL: HB	-2.759	2.965	5.585**	4.016
BL: NRW	0.666	1.684**	6.399***	3.486***
BL: H	2.329	3.368***	7.137***	4.250***
BL: RP, S	1.050	2.058**	6.326***	2.441*
BL: BW	2.283	3.127***	7.837***	4.065***
BL: BY	2.942	2.680***	5.963***	2.819**
BL: B	3.354	0.826	3.693**	6.004***
BL: BB	1.356	-0.807	0.528	1.694
BL: MV	1.582	-0.456	-0.423	2.441
BL: SN	0.142	-0.866	-0.759	1.144
BL: ST	-0.242	-1.162	-0.369	1.038
Konstante	36.938***	8.737***	28.621***	5.618***
Partizipationsgleichung				
20≤Alter<25	0.829***	0.885***		
25≤Alter<30	0.297***	1.066***	0.179	0.435*
30≤Alter<35	0.362***	1.145***	0.607***	0.777***
35≤Alter<40	0.474***	1.117***	0.772***	0.843***
40≤Alter<45	0.422***	1.131***	0.785***	0.923***
45≤Alter<50	0.298**	1.058***	0.741***	0.948***
50≤Alter<55	0.399***	0.764***	0.668***	0.734***
55≤Alter<60	0.235*	0.666***	0.508***	0.531***
Ost	-0.420***	-0.021	-0.294***	0.267***
Abitur	0.653***	0.479***	0.458***	0.332***
Fachabitur	0.482***	0.396***	0.603***	0.410***
Realschule	0.422***	0.406***	0.575***	0.327***
Hauptschule	0.104	0.146*	0.380***	0.071
Erwerbsbehinderung	-0.255**	-0.006	-0.419***	-0.245**
Kinder 0-6 Jahre		-1.545***		-1.760***
Kinder 7-16 Jahre		-0.972***		-1.057***
Kinder ab 17 Jahre		-0.497***		-0.590***
Konstante	0.292**	-0.387***	0.125	-0.236**
Number of obs	2385	3082	2803	2803
Censored obs	517	986	500	1100
Uncensored obs	1868	2096	2303	1703
Wald χ^2	198	618	801	342

Anmerkungen: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Quelle: Eigene Berechnungen auf Basis von EITDsim.

Tabelle 14: Propensity Score Schätzung für Single Männer

Logistic regression	Number of obs	=	439131
	LR chi2(14)	=	1388
	Prob > chi2	=	0.0000
Log pseudolikelihood = -6692	Pseudo R2	=	0.0940

S	Coef.	Std. Err.	z	P>|z|
ef13d1	-.1106853	.0787616	-1.41	0.160
ef13d2	-.0055481	.0820486	-0.07	0.946
ost	-.1680993	.0790589	-2.13	0.033
ef65d1	-.246171	.1801946	-1.37	0.172
ef65d2	.3594968	.1525363	2.36	0.018
ef65d3	.3995235	.146058	2.74	0.006
ef65d4	.1656025	.1402531	1.18	0.238
ef65d5	.21382	.1426471	1.50	0.134
ef65d6	.1737337	.1493075	1.16	0.245
ef65d7	.1818119	.1602164	1.13	0.256
ef65d8	.4486051	.1852519	2.42	0.015
ef70	2.598535	.2708752	9.59	0.000
inc1a	-.0000571	1.83e-06	-31.21	0.000
inc2a	.0006431	.0001909	3.37	0.001
Konstante	6.901472	.1445892	47.73	0.000

Quelle: Eigene Berechnungen.

Tabelle 15: Propensity Score Schätzung für Single Frauen

Logistic regression	Number of obs	=	357840
	LR chi2(12)	=	200
	Prob > chi2	=	0.0000
Log pseudolikelihood = -1332	Pseudo R2	=	0.0699

S	Coef.	Std. Err.	z	P>|z|
ef14d1	-.2421101	.1998115	-1.21	0.226
ef14d2	-.3407477	.2039485	-1.67	0.095
ost	.0085922	.2103296	0.04	0.967
ef68d2	.452149	.3751764	1.21	0.228
ef68d3	-.177554	.3364703	-0.53	0.598
ef68d4	.3512903	.3777184	0.93	0.352
ef68d5	-.1346164	.3508124	-0.38	0.701
ef68d6	.1621796	.3808164	0.43	0.670
ef68d7	-.154939	.3703103	-0.42	0.676
ef68d8	-.2063471	.3779888	-0.55	0.585
ef70	1.131361	.2557227	4.42	0.000
inc1b	-.000055	4.74e-06	-11.60	0.000
Konstante	8.704929	.3519164	24.74	0.000

Quelle: Eigene Berechnungen.

Tabelle 16: Propensity Score Schätzung für zusammenveranlagte Paare

Logistic regression	Number of obs	=	419149
	LR chi2(24)	=	1571
	Prob > chi2	=	0.0000
Log pseudolikelihood = -5272	Pseudo R2	=	0.1297

S	Coef.	Std. Err.	z	P>|z|
ef13d1	-.2942798	.1098817	-2.68	0.007
ef13d2	-.0470112	.1205213	-0.39	0.696
ef14d1	.1123965	.1134494	0.99	0.322
ef14d2	.0594697	.1224935	0.49	0.627
ost	.5731324	.1036187	5.53	0.000
ef65d2	-.1534547	.6777943	-0.23	0.821
ef65d3	.2888939	.3664884	0.79	0.431
ef65d4	.0167116	.2431407	0.07	0.945
ef65d5	-.1825969	.2021501	-0.90	0.366
ef65d6	.2087837	.1877607	1.11	0.266
ef65d7	.055867	.1579591	0.35	0.724
ef65d8	.3742914	.1372218	2.73	0.006
ef68d2	1.217362	.5185519	2.35	0.019
ef68d3	.6705692	.2965562	2.26	0.024
ef68d4	.1709825	.2324874	0.74	0.462
ef68d5	.1411816	.2096322	0.67	0.501
ef68d6	.283409	.1898162	1.49	0.135
ef68d7	.0056311	.1572991	0.04	0.971
ef68d8	.183664	.128148	1.43	0.152
ef70	.3873502	.0468647	8.27	0.000
inc1a	-.0000541	1.62e-06	-33.31	0.000
inc1b	-.0000245	3.43e-06	-7.15	0.000
inc2	-.0000808	4.90e-06	-16.50	0.000
inc2a	.0000257	5.58e-06	4.61	0.000
Konstante	7.115027	.1172488	60.68	0.000

Quelle: Eigene Berechnungen.