Himmelreiche, Ralf K.; Vom Berge, Philipp; Fitzenberger, Bernd; Günther, Roland; Müller, Dana

Working Paper

Überlegungen zur Verknüpfung von Daten der Integrierten Erwerbsbiographien (IEB) und der Verdienststrukturerhebung (VSE)

RatSWD Working Paper, No. 262

Provided in Cooperation with:
German Data Forum (RatSWD)

Suggested Citation: Himmelreiche, Ralf K.; Vom Berge, Philipp; Fitzenberger, Bernd; Günther, Roland; Müller, Dana (2017) : Überlegungen zur Verknüpfung von Daten der Integrierten Erwerbsbiographien (IEB) und der Verdienststrukturerhebung (VSE), RatSWD Working Paper, No. 262, Rat für Sozial- und Wirtschaftsdaten (RatSWD), Berlin, https://doi.org/10.17620/02671.14

This Version is available at:
https://hdl.handle.net/10419/162146

RatSWD Working Paper Series

www.ratswd.de

RatSWD.
Rat für Sozial- und WirtschaftsDaten

262

Überlegungen zur Verknüpfung von Daten der Integrierten Erwerbsbiographien (IEB) und der Verdienststrukturerhebung (VSE)

Ralf Himmelreicher, Philipp vom Berge, Bernd Fitzenberger, Roland Günther, Dana Müller

Mai 2017

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Working Paper Series des Rates für Sozial- und Wirtschaftsdaten (RatSWD)

Die *RatSWD Working Papers* Reihe startete Ende 2007. Seit 2009 werden in dieser Publikationsreihe nur noch konzeptionelle und historische Arbeiten, die sich mit der Gestaltung der statistischen Infrastruktur und der Forschungsinfrastruktur in den Sozial-, Verhaltens- und Wirtschaftswissenschaften beschäftigen, publiziert. Dies sind insbesondere Papiere zur Gestaltung der Amtlichen Statistik, der Ressortforschung und der akademisch getragenen Forschungsinfrastruktur sowie Beiträge, die Arbeit des RatSWD selbst betreffend. Auch Papiere, die sich auf die oben genannten Bereiche außerhalb Deutschlands und auf supranationale Aspekte beziehen, sind besonders willkommen.

RatSWD Working Papers sind nicht-exklusiv, d. h. einer Veröffentlichung an anderen Orten steht nichts im Wege. Alle Arbeiten können und sollen auch in fachlich, institutionell und örtlich spezialisierten Reihen erscheinen. Die *RatSWD Working Papers* können nicht über den Buchhandel, sondern nur online über den RatSWD bezogen werden.

Um nicht deutsch sprechenden Nutzer/innen die Arbeit mit der Reihe zu erleichtern, sind auf den englischen Internetseiten der *RatSWD Working Papers* nur die englischsprachigen Papers zu finden, auf den deutschen Seiten werden alle Nummern der Reihe chronologisch geordnet aufgelistet.

Einige ursprünglich in der *RatSWD Working Papers* Reihe erschienenen empirischen Forschungsarbeiten sind ab 2009 in der RatSWD Research Notes Reihe zu finden.

Die Inhalte der *RatSWD Working Papers* stellen ausdrücklich die Meinung der jeweiligen Autor/innen dar und nicht die des RatSWD. Das Bundesministerium für Bildung und Forschung hat die Publikationen nicht beeinflusst.

Überlegungen zur Verknüpfung von Daten der Integrierten Erwerbsbiographien (IEB) und der Verdienststrukturerhebung (VSE)

Ralf Himmelreicher
(Geschäfts- und Informationsstelle für den Mindestlohn, FU Berlin)

Philipp vom Berge
(IAB)

Bernd Fitzenberger
(HU Berlin)

Roland Günther
(Statistisches Bundesamt)

Dana Müller
(IAB)

Inhalt

1 Einleitung 2
2 Verknüpfung von Daten 4
3 Die Ausgangsdatensätze VSE und IEB 7
4 Direkte Verknüpfung von IEB und VSE 14
5 Indirekte Verknüpfung von IEB und VSE 16
6 Fazit und nächste Schritte 21
7 Literatur 22
8 Anhang 24

1. Einleitung

Für quantitative Analysen zu den Wirkungen des Mindestlohns stehen in Deutschland eine Reihe von Datensätzen aus wissenschaftlichen Umfragen und administrativen Prozessen zur Verfügung (Mindestlohnkommission 2016; vgl. ausführlich vom Berge et al. 2014). Im Vorfeld der Einführung des gesetzlichen Mindestlohns wurde die Datenbasis für eine wissenschaftliche Evaluation erheblich verbessert. Hierzu zählen beispielsweise die Erweiterung der Verdienststrukturerhebung (VSE) um Betriebe mit weniger als zehn Beschäftigten, die Durchführung einer Verdiensterhebung 2015 (VE2015), um die Situation direkt nach Einführung des Mindestlohns erfassen zu können, sowie die Berücksichtigung mindestlohnspezifischer Fragen im Sozio-oekonomischen Panel (SOEP) und im IAB-Betriebspanel. Dennoch kann die aktuelle Datenlandschaft in Bezug auf Forschungsvorhaben im Bereich des Mindestlohns noch weiter verbessert werden. Insbesondere mit Blick auf die Validität von Arbeitszeitangaben sind erhebliche Verbesserungen anzustreben. So sind Informationen zur wöchentlichen Arbeitszeit in den Verwaltungsdaten der Bundesagentur für Arbeit (BA) lediglich sehr grob als Voll- oder Teilzeitbeschäftigung ausgewiesen. Hinreichend exakte Angaben zur Arbeitszeit sind allerdings zentral, wenn es darum geht, aus Bruttomonatslöhnen und wöchentlichen Arbeitszeiten Bruttostundenlöhne zu berechnen (Dütsch et al 2017).

Die Mindestlohnkommission hat sich daher in ihrem ersten Bericht für eine Verbesserung der Datenqualität und Datenverfügbarkeit für die Evaluation der Auswirkungen des gesetzlichen Mindestlohns ausgesprochen (Mindestlohnkommission 2016). Dies bezog sich insbesondere auf eine Verknüpfung von Daten der Integrierten Erwerbsbiographien (IEB) des Instituts für Arbeitsmarkt- und Berufsforschung (IAB) mit der Verdienststrukturerhebung (VSE) des Statistischen Bundesamtes. Ziel war und ist es vor allem, die Informationen zu den Arbeitszeiten aus der VSE mit der IEB zu verbinden. Bereits vor der Einführung des gesetzlichen Mindestlohns in Deutschland wurde die Verknüpfung von IEB und VSE vorgeschlagen (vom Berge et al. 2014). Jedoch konnte das Vorhaben nicht umgesetzt werden. Vor diesem Hintergrund und die Empfehlungen der Mindestlohnkommission aufgreifend hat die Geschäfts- und Informationsstelle für den Mindestlohn im November 2016 einen Workshop organisiert, auf dem die Potentiale, Möglichkeiten, aber auch Grenzen einer Verknüpfung dieser beiden für die Mindestlohnevaluation zentralen Datensätze diskutiert wurden.[1]

Im Workshop bestand von Seiten der Wissenschaft, des IAB sowie des Statistischen Bundesamts Einigkeit, dass das erheblich vergrößerte Analysepotenzial eines aus IEB und VSE verknüpften Datensatzes eine wesentliche Verbesserung der Evaluationsmöglichkeiten des Mindestlohns in Deutschland darstellen würde. Allerdings wurde seitens des Statistischen Bundesamts darauf hingewiesen, dass ohne rechtliche Grundlage eine eindeutige Verknüpfung der Datensätze nicht vorstellbar ist.

[1] An der Veranstaltung nahmen Vertreterinnen und Vertreter des Statistischen Bundesamtes, des Forschungsdatenzentrums der Bundesagentur für Arbeit im Institut für Arbeitsmarkt- und Berufsforschung (IAB), des Bundesministeriums für Arbeit und Soziales, des Bundesministeriums für Wirtschaft und Energie sowie aus der Wissenschaft teil. Zudem waren der Vorsitzende der Mindestlohnkommission, Mitarbeiterinnen und Mitarbeiter aus den Arbeitsstäben der Mitglieder der Mindestlohnkommis-sion sowie der Geschäfts- und Informationsstelle für den Mindestlohn anwesend.

Eine direkte Verknüpfung von IEB und VSE würde nach aktuellem Stand der Mindestlohnforschung das Analysepotential enorm erhöhen. So könnte auf Basis eines eindeutig verknüpften Arbeitgeber-Arbeitnehmer-Datensatzes Verlaufsanalysen für Betriebe und Beschäftigte unter Kontrolle von Interaktionseffekten durchgeführt werden. Im Detail würden valide Informationen zu den vertraglichen und bezahlten Arbeitsstunden sowie zu Zeiten der Nichtbeschäftigung, Informationen zu Bruttolöhnen und Sonderzahlungen sowie Angaben zur ökonomischen Situation der Betriebe, die Tarifbindung, die industriellen Beziehungen und die Arbeitsorganisation in den Betrieben ein dem Forschungsgegenstand angemessenes Analysepotential bieten. Auf der Grundlage solcher Informationen können Reaktionen der Betriebe wie der Beschäftigten nach Einführung oder Anpassung des Mindestlohns im Hinblick auf mögliche Veränderungen der Arbeitszeitgestaltung, der Arbeitsentgelte wie auch der Arbeitsorganisation untersucht werden. Um eine direkte Verknüpfung von Datensätzen einfacher und aus datenschutzrechtlicher Sicht verlässlicher zu machen empfiehlt der RatSWD (2017a) die Einführung einer eineindeutigen anonymisierten Kennung.

Als Ausgangspunkt für die weitere rechtliche wie politische Diskussion bedarf es zunächst eines klaren fachlichen Konzepts. Darin sollen die wesentlichen Fragen mit Blick auf eine Verknüpfung der beiden Datenquellen diskutiert werden: Welche Merkmale aus den jeweiligen Datensätzen sind relevant? Wie kann eine Verknüpfung konkret gestaltet werden? Welche Auswertungsmöglichkeiten sollen für den neuen Datensatz zugelassen werden?

Das vorliegende Papier kann auch die Grundlage für eine juristische Bewertung sein, um eine rechtliche Grundlage zur künftigen Verknüpfung von IEB und VSE zu schaffen. Hinsichtlich der bereits durchgeführten Erhebung VSE2014 sollen Vorschläge formuliert werden, die eine Verknüpfung der Daten im Rahmen der vorhandenen rechtlichen Möglichkeiten (insbesondere nach statistischen Kriterien) ermöglichen.

Im Hinblick auf die Verdienststrukturerhebungen des Statistischen Bundesamtes fokussiert dieses Papier auf die VSE2014 und nachfolgende Erhebungen, wie etwa die VSE2018. Keine Berücksichtigung finden die Verdiensterhebungen VE2015, VE2016 und VE2017. Der Grund für diese Vorgehensweise ist, dass sich VSE und VE in mehreren Punkten deutlich unterscheiden, die die Suche nach rechtlichen und technischen Möglichkeiten zur Verknüpfung erheblich erschweren würden. Erstens unterscheidet sich die rechtliche Grundlage. Die rechtliche Grundlage zur Erhebung der VSE ist das Verdienststatistikgesetz (VerdStatG), während die Verdiensterhebungen auf dem Bundesstatistikgesetz (BStatG) basieren. Zweitens gibt es unterschiedliche zeitliche Perspektiven: Nach aktuellem Stand werden lediglich drei Verdiensterhebungen ins Feld gehen, die VE2015, VE2016 und letztmalig die VE2017. Für eine Erhebung einer auf die VSE2018 folgende VE2019 gibt es derzeit keine rechtliche Grundlage. Drittens unterscheiden sich VSE und VE hinsichtlich möglicher Identifikatoren zur Verknüpfung mit der IEB. In den VSE werden die Rentenversicherungsnummern erhoben, während die VE aus rechtlichen Gründen auf Personalnummern abstellt. Vor diesem Hintergrund wird der Schwerpunkt der Ausführungen auf die Verknüpfung von IEB und VSE gelegt.

Das vorliegende Papier gliedert sich wie folgt: Im zweiten Abschnitt werden verschiedene Möglichkeiten zur Verknüpfung unterschiedlicher Datenquellen vorgestellt. Der dritte Abschnitt geht auf die Spezifika der zu verknüpfenden Datenquellen ein. Daran schließt sich im vierten Abschnitt eine Darstellung möglicher direkter Identifikatoren an, auf deren Grundlage eine Verknüpfung der beiden Datensätze vorgenommen werden könnte. Der fünfte Abschnitt diskutiert Aspekte indirekter Verknüpfungsmöglichkeiten. Der sechste und letzte Abschnitt fasst zentrale Befunde zusammen und schlägt einige Maßnahmen vor, die eine erfolgreiche Verknüpfung von IEB und VSE begünstigen würden.

2. Verknüpfung von Daten

Weltweit ist ein zunehmender Trend in den Wirtschafts- und Sozialwissenschaften zur Verknüpfung verschiedener Datenquellen zu verzeichnen (RfII 2017). Durch die Verknüpfung von Daten wird das Analysepotenzial vorhandener Datenquellen vergrößert. Zudem lassen sich solche Merkmale, die in beiden Datensätzen vorhanden sind, validieren. Ferner kann durch die Verknüpfung von Daten der Aufwand für die befragten Arbeitnehmerinnen und Arbeitnehmer sowie für Unternehmen verringert werden. Hierdurch werden sowohl Bürokratiekosten als auch Kosten, die für zusätzliche Erhebungen anfallen würden, eingespart.[2]

Grundsätzlich können zwei verschiedene Methoden der Datenverknüpfung unterschieden werden, dies sind die direkte und die indirekte Verknüpfung. Bei der direkten Verknüpfung werden die Datensätze an Hand eindeutiger Identifikatoren (Versicherungs- oder Betriebsnummer), die in beiden Datensätzen vorhanden sind, verbunden (deterministisches Record Linkage). Ziel ist eine exakte Übereinstimmung der Fälle in beiden Datensätzen. Bei der indirekten Verknüpfung wird die Forderung nach exakter Übereinstimmung aufgegeben. Anhand von Wahrscheinlichkeiten oder Ähnlichkeiten wird nach statistischen Zwillingen gesucht (distanzbasiertes oder probalistisches Record Linkage). In Abhängigkeit von den zu verknüpfenden Datensätzen ist entweder eine Einverständniserklärung der Befragten, die im Rahmen der Befragung oder im Nachhinein eingeholt werden kann, oder eine rechtliche Regelung, die eine Verknüpfung vorsieht, erforderlich.

Die Verknüpfung von Umfragedaten mit prozessproduzierten Daten etwa der Träger der sozialen Sicherung wird auch deshalb zunehmend prominenter, weil die Ausschöpfungsquoten von Erhebungen oftmals zurückgehen. Als Ursachen von Nicht-Teilnahme oder vollständigem Antwortausfall (‚Unit-Nonresponse') werden schwere Erreichbarkeit, Teilnahmeverweigerung und Nicht-Befragbarkeit (z. B. durch Krankheit, Sprachprobleme oder auch Todesfälle in Panelstichproben) angeführt (Schnell et al. 2013). Insofern gewinnen prozessproduzierte Daten wie die IEB sowie verpflichtende Erhebungen wie die VSE zunehmend an Bedeutung. So verzeichnet die verpflichtende VSE2014 eine Rücklaufquote von nahezu 100 % (Statistisches Bundesamt 2015), während bei der VE2015 mit freiwilliger Teilnahme die

[2] Zu grundlegenden Informationen zum Thema Verknüpfung von verschiedenen Daten siehe das Internetportal des ‚German Record Linkage Center (GermanRLC)'. Das GermanRLC wurde im Jahr 2011 gegründet, um die Forschung bei der Verknüpfung von Daten (engl. record linkage) in technischer und methodischer Hinsicht wie auch grundsätzlich bei Verknüpfungsvorhaben zu unterstützen (Schnell und Bender 2010). Weitere Informationen sind abgelegt unter http://www.record-linkage.de

Teilnahmequote bei 12,8 % liegt (Statistisches Bundesamt 2017). In Bezug auf verdienststatistische Erhebungen mit freiwilliger Teilnahme konstatiert das Statistische Bundesamt daher: „Belastbare detaillierte Statistiken lassen sich auf diesem Weg jedoch nicht gewinnen. Freiwillige Erhebungen stellen somit keine Alternative für dauerhafte, amtliche Verdienststatistiken mit Auskunftspflicht dar." (Statistisches Bundesamt 2017: 66)

Nach aktuellem Stand der Mindestlohnforschung würde ein verknüpfter Arbeitgeber-Arbeitnehmer-Datensatz („Linked-Employer-Employee-Datensatz"), der Verlaufsanalysen für Betriebe und Beschäftigte zulässt, Angaben zu den vertraglichen und bezahlten Arbeitsstunden sowie zu Zeiten der Nichtbeschäftigung enthält, Informationen zu Bruttolöhnen und Sonderzahlungen aufweist sowie die ökonomische Situation der Betriebe, die Tarifbindung, die industriellen Beziehungen und die Arbeitsorganisation in den Betrieben erfasst, ein dem Forschungsgegenstand angemessenes Analysepotential bieten. Die VSE und die IEB ermöglichen gehaltvolle Analysen und haben ihre spezifischen Stärken. Beide Datenquellen sind aus der Perspektive der Wissenschaft jedoch mit Einschränkungen behaftet,[3] so dass eine Verknüpfung der beiden Datensätze eine deutliche Verbesserung der Analysemöglichkeiten für die Mindestlohnforschung ermöglichen würde. Ferner könnten für die Mindestlohnforschung höchst relevante Messfehler evaluiert und gegebenenfalls korrigiert werden. Zudem könnten Synergien aus den spezifischen Vorzügen beider Datensätze genutzt werden. Solche Synergien könnten mit gewissen Einschränkungen und unter speziellen Annahmen selbst dann genutzt werden, wenn keine direkte Verknüpfung der beiden Datenquellen VSE und IEB erfolgen würde, sondern stattdessen mit Hilfe statistischer Coupling- oder Matchingverfahren bestimmte Merkmale aus den Datensätzen durch indirekte Verknüpfung analysiert werden könnten (siehe Abschnitt 5).

In den letzten Jahren wurden in Deutschland einige vielversprechende Projekte mit verknüpften Datensätzen realisiert. So wurden unter anderem im Rahmen des SHARE-RV Projektes Daten des „Survey of Health, Ageing and Retirement in Europe (SHARE)" und der Deutschen Rentenversicherung über die Versicherungsnummer direkt verknüpft (Czaplicki und Korbmacher 2010). Ein weiteres Beispiel ist das IAB-SOEP Migration Sample, bei dem die Befragungsdaten des SOEP mit denen der IEB verknüpft wurden. Ähnlich wie bei SHARE-RV wurde den Interviewten im Rahmen der Datenerhebung folgende Frage zur direkten Verknüpfung gestellt: „Wir haben aber noch eine Bitte an Sie: Für wissenschaftliche Untersuchungen ist es zunehmend wichtig, mehr über die Erwerbsbiografie von Menschen zu erfahren und bei statistischen Analysen zu berücksichtigen. Zu diesem Zweck möchten wir gerne die Befragungsdaten aus den Interviews mit den Sozialversicherungsdaten des Instituts für Arbeitsmarkt- und Berufsforschung verknüpfen. Dies ist aus datenschutzrechtlichen Gründen nur mit Ihrer Einwilligung möglich. Selbstverständlich ist Ihre Einwilligung freiwillig." (Eisnecker und Kroh 2017: Appendix 1, 2)

[3] Als eine wesentliche Einschränkung wird der fehlende Haushaltskontext angesehen, der insbesondere im Fall von Pflege und Kinderbetreuung das Arbeitsangebotsverhalten der erwerbsfähigen Haushaltsmitglieder beeinflusst. Zudem ist die Vermögensausstattung der Personen und Haushalte nicht bekannt.

Die beiden Projekte zum direkten Record Linkage zeigen, dass die Verknüpfung von Umfragedaten mit administrativen Daten der sozialen Sicherung in Deutschland ein erprobtes sowie datenschutzrechtlich geklärtes Verfahren darstellt. Datenschutzrechtliche Voraussetzung für die Verknüpfung mit Sozialdaten ist nach SGB X die schriftliche Einverständniserklärung der Befragten zum direkten Record Linkage.[4] Zur Sicherung der informationellen Selbstbestimmung der befragten Personen wird aus der Perspektive des Datenschutzes die informierte Einwilligung (informed consent) als hinreichende Grundlage betrachtet. Das Einholen solcher Einverständniserklärungen ist außerhalb des Kontexts von laufenden Befragungen eher schwierig, weil oftmals kein vorheriger Kontakt zur Zielperson bestanden hat oder eine Kontaktaufnahme grundsätzlich nicht möglich ist. Für den Fall, dass eine nachträgliche Kontaktaufnahme möglich sein sollte wird von niedrigen Teilnahmequoten ausgegangen (Sakshaug et al. 2016). Zudem sind schriftliche Anschreiben, Rückantworten mir frankierten Briefumschlägen für die schriftliche Einverständniserklärung und deren nachträgliche Zuordnung zu den konkreten Untersuchungseinheiten kostenintensiv und vergleichsweise fehleranfällig.

Neben diesen beispielhaft angeführten Verfahren zur direkten Verknüpfung von Umfrage- mit Sozialdaten auf der Basis einer prospektiven schriftlichen Einverständniserklärung gibt es in der Forschungspraxis weitere Konstellationen zu bedenken. Hierbei spielt vor allem die Verknüpfung von bereits erhobenen Daten eine Rolle. In diesem Fall müsste bei den teilnehmenden Personen oder Betrieben im Nachhinein angefragt werden, ob sie einer rückwirkenden Verknüpfung zustimmen. Das Einholen einer solchen rückwirkenden Einverständniserklärung wird als besonders schwierig angesehen, weil sich Befragungspersonen nicht mehr an die Befragungen erinnern können oder wollen oder eventuell verzogen und damit nicht mehr erreichbar sind. Insgesamt wird in Bezug auf das Einholen nachträglicher Einverständniserklärungen von geringen Rücklaufquoten und eher hohen Kosten ausgegangen.

Die Möglichkeit der Einholung einer schriftlichen Einverständniserklärung zur Datenverknüpfung besteht für die Daten der Statistischen Ämter grundsätzlich auch. Sie scheidet für die VSE2014 jedoch aus Kostengründen bzw. praktischen Erwägungen aus (siehe Abschnitt 3.4)

Vor diesem Hintergrund ist zunächst eine rechtliche Grundlage zur direkten Verknüpfung von zukünftigen Erhebungen zu schaffen, auf deren Grundlage die VSE2018 mit der entsprechenden IEB verknüpft werden kann. Solange diese nicht vorliegt, können die Daten nicht eindeutig, etwa über die Versicherungsnummer, zusammengefügt werden. Als mögliche Alternative zum eindeutigen Record Linkage gibt es eine ganze Reihe von statistischen Verfahren. Für den Einsatz solcher Verfahren ist ein tieferes Verständnis der Datensätze und ihrer Merkmale erforderlich, weshalb zunächst die beiden Datensätze und ihre Merkmale genauer dargestellt werden.

[4] Grundsätzlich können Einverständniserklärungen aktiv (opting-in) oder passiv (opting-out) eingeholt werden. Die aktive und ausdrückliche Einverständniserklärung setzt aktives Handeln (etwa Namenseingabe und Unterschrift) der Befragten voraus, damit ein Einverständnis erklärt ist. Bei einer passiven Einverständniserklärung entsteht eine Einverständniserklärung immer dann, wenn nicht aktiv widersprochen wird. Zu den jeweiligen Effekten siehe Sakshaug et al. (2016).

3. Die Ausgangsdatensätze VSE und IEB

3.1 Die Verdienststrukturerhebung als Datenquelle für eine verknüpfte IEB-VSE

Die VSE ist die umfassendste Erhebung zu Verdiensten und bezahlten Arbeitszeiten in Deutschland und wird alle vier Jahre durchgeführt (Statistisches Bundesamt 2013; Günther 2013). Für die angeschriebenen Betriebe besteht eine gesetzliche Auskunftspflicht. Die letzte Erhebung aus dem Jahr 2014, also unmittelbar vor Einführung des gesetzlichen Mindestlohns, umfasste rund 60.000 Betriebe mit mindestens einem Beschäftigten aus allen Wirtschaftszweigen außer privaten Haushalten und exterritorialen Organisationen. Es wurden Daten zu rund einer Million Beschäftigungsverhältnisse gesammelt (Statistisches Bundesamt 2016). Die Daten wurden aus der betrieblichen Personalverwaltung, v.a. der Entgeltabrechnung übernommen und weisen dementsprechend eine hohe Zuverlässigkeit auf. Das gilt insbesondere im Hinblick auf die Entgeltangaben. „Lediglich bei der Arbeitszeit könnte es im Einzelfall Probleme geben, jedoch könnten die wegen der Mindestlohngesetzgebung neuen strengeren Aufzeichnungspflichten die Meldung hierzu erleichtern." (Zimmer 2015)

Im Vergleich zu anderen Datenquellen liegen die Stärken der VSE für die Mindestlohnevaluation vor allem darin, dass neben monatlichen Bruttoverdiensten auch monatliche bezahlte Arbeitsstunden erfasst werden. Das erlaubt zum einen die Berechnung von Stundenlöhnen, auf die die Mindestlohngesetzgebung abstellt. Zum anderen können so aber auch Veränderungen in den Arbeitszeiten als Reaktion auf den Mindestlohn untersucht werden.

Zudem erlaubt die Datenbasis der VSE eine vergleichsweise starke Differenzierung nach Wirtschaftszweigen und Regionen. Für die VSE2014 wurde Wirtschaftszweigen mit hohem Niedriglohnanteil im Vergleich zu anderen Wirtschaftszweigen ein größerer Stichprobenumfang zugewiesen, damit die Ergebnisse des Niedriglohnbereichs belastbar und differenziert auswertbar werden. Methodische Änderungen der VSE2014 im Vergleich zur VSE2010 werden in FDZ der Statistischen Ämter der Länder (2016a) erläutert.

Merkmale

Die VSE ist ein Linked-Employer-Employee-Datensatz. Ihre Merkmale lassen sich nach betrieblichen Angaben und nach Angaben über die Beschäftigten unterscheiden (Statistisches Bundesamt 2016). Die entsprechenden Fragebögen befinden sich in den Anhangtabellen A2 und A3.

Zentrale Merkmale für die ML-Forschung im Rahmen eines verknüpften Datensatzes sind auf Ebene der Beschäftigten:

- das Merkmalsset zur Arbeitszeit (regelmäßige wöchentliche Arbeitszeit, bezahlte Stunden ohne Überstunden sowie bezahlte Überstunden pro Monat) und
- das Merkmalsset zur Entlohnung (gesamter Bruttomonatsverdienst, darunter der Gesamtverdienst für Überstunden und für Zuschläge).

sowie auf Ebene des Betriebes:

- der Tarifvertrag, kodiert nach Tarifvertragsnummer des Statistischen Bundesamtes bzw. die Tarifbindung (Tarifbindung vs. keine Tarifbindung) des Betriebes und
- die betriebsübliche Wochenarbeitszeit von Vollzeitbeschäftigten.

Zusätzlich hilfreiche Merkmale im Rahmen eines verknüpften Datensatzes sind auf Ebene der Beschäftigten:

- die Vertragsform Vollzeit/Teilzeit/Befristet/Unbefristet (Stelle 9 des Tätigkeitsschlüssels 2010)

und auf Ebene des Betriebes:

- der Wirtschaftszweig, kodiert nach WZ2008.

Die beiden Merkmale Vertragsform und Wirtschaftszweig liegen auch in den IEB vor. Hier wurden die Angaben jedoch ohne weitere Überprüfung vom Betrieb übernommen, während die Angaben in der VSE von den Statistischen Landesämtern explizit überprüft und im Kontakt mit dem Betrieb korrigiert wurden. Die Angaben der VSE dürften damit zuverlässiger sein. Die Frage der Genauigkeit von Angaben ist nicht unbedeutend, wie die Analysen zur Lohnungleichheit von Möller (2016) zeigen. Eine umfangreiche Dokumentation der Merkmale der VSE2014 findet sich in den Metadaten zum entsprechenden Scientific Use File (FDZ der Statistischen Ämter der Länder 2016b).

Verknüpfbarer Teil des Datenbestands

Der Datenbestand der VSE2014 setzt sich aus drei Teilen zusammen, da die Daten aus drei verschiedenen Quellen und mit verschiedener Methodik gewonnen wurden. Lediglich einer dieser drei Teile ist mit der IEB direkt verknüpfbar. Dies ist die originär erhobene Stichprobe der VSE2014, da lediglich dort die Versicherungsnummer erhoben wird. Tabelle 1 gibt einen Überblick über die drei Datenquellen, die der VSE2014 zugrunde liegen.

Tabelle 1: Die Datenquellen der VSE2014

	Stichproben-erhebung	Personalstand-statistik	Betriebe ohne sozi-alversicherungs-pflichtige Beschäf-tigte (SVB)
Grundgesamtheit			
Abgrenzung nach Wirtschaftsabschnit-ten*	Beschäftigte in Betrieben mit SVB der Wirt-schaftsabschnitte A bis S ohne O und ohne den überwiegenden Teil des Abschnitts P	Beschäftigte der Wirt-schaftsabschnitte O und P der Personalstandsta-tistik	Beschäftigte in Betrieben ohne SVB der Wirt-schaftsabschnitte A bis S ohne O
Betriebe	2,053 Mill.	nicht bekannt	0,425 Mill.
Jobs	32,1 Mill.	4,2 Mill.	0,9 Mill.
Stichprobe			
Betriebe	55.882	4.847 (künstliche)	10.000
Jobs	763.000	248.000	22.000
Verfahren der Datengewinnung	Erhebung der Merkmale beim Arbeitgeber (Be-trieb) per Online-Formular oder elektroni-scher Datenübermittlung	Übernahme bzw. Ablei-tung der Merkmale aus der Personalstandstatis-tik. Die Personalstandsta-tistik erhob die Daten per elektronischer Daten-übermittlung beim Ar-beitgeber (zentrale Per-sonalabrechnungsstellen)	Hot-deck-Imputation ganzer Jobdatensätze aus der Stichprobenerhebung
Identifikatoren für deterministische Verknüpfung			
Betrieb	Betriebsnummer BA	keine	keine
Job	Rentenversicherungs-nummer nach Verfrem-dung per Hash-Verfahren	keine	keine

Anmerkungen: *Zur Klassifikation der Wirtschaftsabschnitte siehe Anlage A1.
SVB – sozialversicherungspflichtig Beschäftigte.
Quelle: eigene Abbildung, zusammengestellt aus Statistisches Bundesamt (2016).

Beschreibung der Stichprobenziehung des verknüpfbarer Teils

Der verknüpfbare Teil ist eine Stichprobe der abgebildeten Grundgesamtheit von Betrieben und Jobs. Die Stichprobe wurde nach statistischen Kriterien so gestaltet, dass sie sowohl zuverlässige repräsentative Gesamtergebnisse sicherstellt, als auch Differenzierungen nach Wirtschaftsabschnitten und Bundesländern ermöglicht. Die Zuverlässigkeit der differenzierten Befunde ist dabei umso größer, je mehr Beschäftigte die jeweilige Grundgesamtheit umfasst.

Das Verfahren zur Auswahl von Beschäftigungsverhältnissen war zweistufig: Zunächst wurden die Betriebe und dann die Beschäftigungsverhältnisse ausgewählt. Für die Auswahl der Betriebe diente die Bundeskopie des statistischen Unternehmensregisters mit Stand Mai 2014 als Auswahlgrundlage. Alle Betriebe mit sozialversicherungspflichtigen Beschäftigten wurden in die Auswahlgrundlage einbezogen. Es erfolgte eine Schichtung nach Bundesland,

Wirtschaftszweig (84 Abteilungen) und der Zahl der Beschäftigten des Betriebs (7 Größenklassen). Der nominale Stichprobenumfang betrug 60.000 Betriebe. Die Auswahlgrundlage für die Beschäftigungsverhältnisse umfasste alle Beschäftigte eines in der ersten Stufe ausgewählten Betriebs. Vom Statistischen Bundesamt wurde ein fester Auswahlsatz (bzw. Auswahlabstand) für jede Beschäftigtengrößenklasse vorgegeben. Die Betriebe hatten die Möglichkeit, die Auswahl entweder selbst vorzunehmen oder dies den Statistischen Landesämtern zu überlassen. Im ersten Fall hatten die Betriebe zunächst alle Beschäftigtensätze in eine beliebige Reihenfolge zu bringen, etwa durch eine Sortierung nach der Personalnummer oder nach dem Namen. Danach war aus der Reihe jener Satz als erster auszuwählen, der an der Position stand, die als zufällige Startzahl vom statistischen Amt vorgegeben worden war. Schließlich war im vorgegebenen Abstand jeder weitere Satz der Reihung auszuwählen. Im zweiten Fall mussten die Betriebe die Angaben aller Beschäftigten übermitteln und die Ämter übernahmen die zufallsgesteuerte Auswahl sowie die Löschung der überzähligen Datensätze.[5]

Tabelle 2: Wahrscheinlichkeit der Auswahl nach Beschäftigtengrößenklasse des Betriebs

Beschäftigtengrößenklasse des Betriebs in SVB	Auswahlabstand im Betrieb	1. Stufe: Durchschnittlicher Auswahlsatz	2. Stufe: Auswahlsatz	Durchschnittlicher Auswahlsatz beider Stufen zusammen
1 – 9	1	1,66%	100%	1,66%
10 – 49	2	5,39%	50%	2,70%
50 – 99	3	10,1%	33,3%	3,37%
100 – 249	6	20,8%	16,7%	3,47%
250 – 499	10	32,7%	10%	3,27%
500 – 999	20	46,2%	5%	2,31%
1000 und mehr	40	100%	2,5%	2,50%
Insgesamt	X	X	X	2,61%

Anmerkungen: SVB – sozialversicherungspflichtig Beschäftigte.
Quelle: eigene Abbildung, zusammengestellt aus Statistisches Bundesamt (2016).

Nach Beschäftigtengrößenklasse des Betriebs ergaben sich unterschiedliche Auswahlwahrscheinlichkeiten, die in Tabelle 2 dargestellt sind. Durch die Kombination von beiden Stufen ergaben sich gleichmäßigere Auswahlsätze, die die hochgerechneten Ergebnisse stabilisierten.[6]

[5] 78% der Betriebe mit zehn und mehr SVB entschieden sich für Fall 1 und übermittelten eine Auswahl der Beschäftigten. 619 000 Datensätze wurden dabei ausgewählt und übermittelt, 5,3 Millionen Datensätze nicht. 22% der Betriebe mit zehn und mehr SVB entschieden sich für Fall 2 und übermittelten Angaben zu allen Beschäftigten. Die Statistischen Ämter der Länder nahmen die Auswahl von 144 000 Datensätzen vor und löschten rund 890 000 überzählige Datensätze.
[6] Ein detaillierter Methodenbericht zur Stichprobenziehung der VSE2014 wird im Laufe des Jahres 2017 in der vom Statistischen Bundesamt herausgegebenen Zeitschrift Wirtschaft und Statistik erscheinen.

3.2 Die Integrierten Erwerbsbiographien als Datenquelle für eine verknüpfte IEB-VSE

Die Daten der Bundesagentur für Arbeit (BA) sowie des Instituts für Arbeitsmarkt- und Berufsforschung (IAB) stellen für die Mindestlohnevaluation eine wichtige Informationsquelle dar. Auf individueller Ebene sind die Integrierten Erwerbsbiographien (IEB) ein wesentlicher Datensatz, der detaillierte Informationen zu den Erwerbsverläufen von Personen enthält. Die IEB werden aus verschiedenen administrativen Daten zusammengeführt und enthalten Informationen zu Beschäftigungszeiten, des Bezugs von Leistungen, Zeiten der Arbeitslosigkeit, Arbeitsuche sowie zur Teilnahme an Maßnahmen (vom Berge et al. 2014). Aufgrund sozialdatenschutzrechtlicher Bestimmungen unterliegt die Nutzung der IEB gewissen Zugangsbeschränkungen. In der Regel wird daher die Stichprobe der Integrierten Arbeitsmarktbiographien (SIAB) genutzt, die eine 2-Prozent-Stichprobe der IEB darstellt (Antoni et al. 2016). Die IEB sind aufgrund der hohen Differenzierbarkeit und Datenqualität sowie der Erfassung tagesgenauer Erwerbsverläufe für die Evaluation von Beschäftigungs- und Lohneffekten grundsätzlich gut geeignet (vom Berge et al. 2014). Auch Längsschnittuntersuchungen auf Personen- oder Betriebsebene sind möglich. Eine problematische Einschränkung ist allerdings, dass Arbeitszeiten nicht oder nur grob nach Voll- und Teilzeitbeschäftigung erfasst werden und somit keine genaue Berechnung von Stundenlöhnen möglich ist. Unter bestimmten Annahmen kann die Arbeitszeit der Beschäftigten aus anderen Datensätzen, wie zum Beispiel dem Mikrozensus imputiert werden (IAB/RWI/ISG 2011). Solche Maßnahmen sind als Hilfskonstruktionen zu bewerten, weil unklar ist, ob die tatsächliche individuelle Arbeitszeit der imputierten entspricht. Ein solches Defizit könnte vor allem im Rahmen einer direkten Verknüpfung ausgeräumt werden. In Ermangelung direkter Verknüpfungsmöglichkeiten stellen solche Imputationsverfahren allerdings eine innovative Annäherung an die Realität dar, weshalb sie im Rahmen der Evaluation der Branchenmindestlöhne eingesetzt wurden (IAB/RWI/ISG 2011).

Erhebungsart und Inhalte

Die IEB sind als Vollerhebung konzipiert und somit für den erhobenen Personenkreis repräsentativ. Sie enthalten Informationen zu allen Personen, die im Beobachtungszeitraum mindestens einmal einen der folgenden Zustände aufweisen (Antoni et al. 2016):

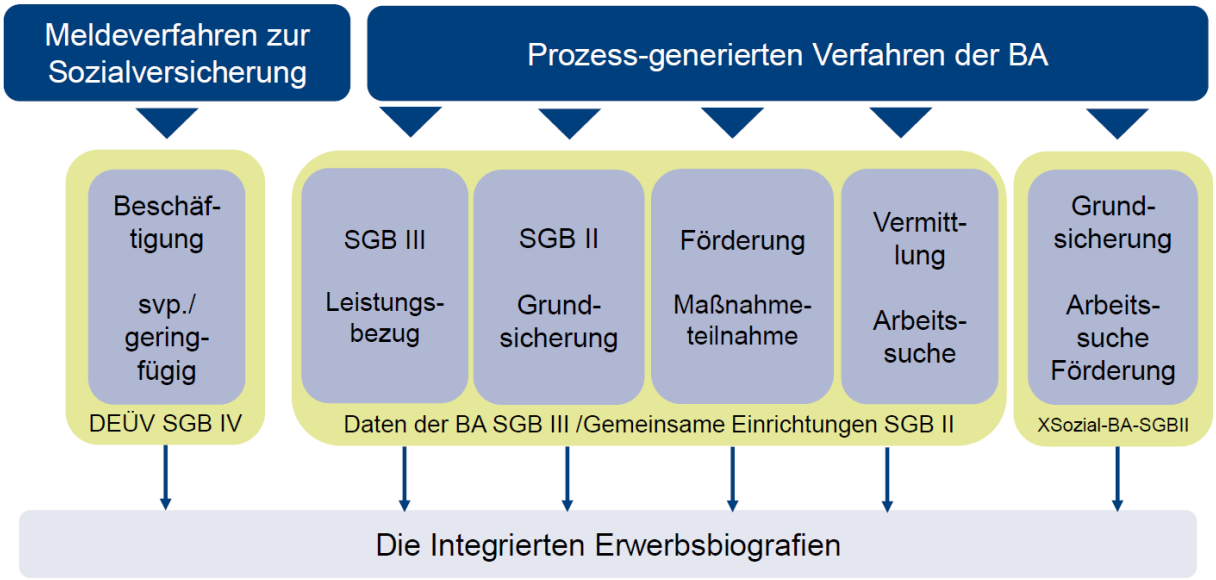
- sozialversicherungspflichtige Beschäftigung in Deutschland (erfasst ab 1975)
- geringfügige Beschäftigung in Deutschland (erfasst ab 1999)
- Bezug von Leistungen nach dem Rechtskreis SGB III (erfasst ab 1975) oder SGB II (erfasst ab 2005)
- bei der Bundesagentur für Arbeit (BA) bzw. den Grundsicherungsträgern als arbeitsuchend gemeldet (erfasst ab 2000)
- Teilnahme an arbeitsmarktpolitischer Maßnahme der BA (erfasst ab 2000).

Hierzu werden Daten aus dem Meldeverfahren zur gesetzlichen Kranken-, Pflege-, Renten- und Arbeitslosenversicherung mit Meldungen aus mehreren Fachverfahren der BA zusammengeführt. Dies geschieht über einen quellenübergreifenden Personen-Identifikator, der sogenannten Einheitlichen Statistischen Person (ESP), die in

Zusammenarbeit mit dem IAB von der BA-Statistik entwickelt wurde. Obwohl die IEB für viele der enthaltenen Personen das gesamte Erwerbsleben weitestgehend tagesgenau abbilden können, bestehen dennoch einige Ausnahmefälle, in denen die IEB keine entsprechenden Informationen enthalten. Hierzu zählen unter anderem Zeiten, in denen eine Person selbstständig beschäftigt oder verbeamtet ist.[7]

Im Hinblick auf die Möglichkeiten der statistischen Datenanalyse ist es von Vorteil, dass die Episoden in der IEB, für die sowohl ein Beginn- als auch ein Enddatum angegeben ist, aufgeteilt (‚gesplittet') werden können. Dabei werden solche Spells, die sich zeitlich überschneiden, gesplittet. Innerhalb eines Personenkontos gibt es dann nur noch überschneidungsfreie oder vollständig parallele Spells.

Abbildung 1: Die Integrierte Erwerbsbiografie (IEB) innerhalb der Datenlandschaft von BA und IAB und FDZ-Daten



Quelle: Forschungsdatenzentrum der BA im IAB (o. J.).

Vor- und Nachteile

Die im Rahmen eines Verwaltungsprozesses erhobenen Daten haben einige besondere Vor-, aber auch einige Nachteile (Kröger et al. 2011). Besondere Vorteile sind, dass es keinen selektiven Antwortausfall, keine soziale Erwünschtheit bei der Beantwortung von Fragen und auch keine Panelmortalität gibt. Zudem liegen oftmals hohe Fallzahlen (Vollerhebung) über lange Zeiträume vor. Für die untersuchten Personen entsteht kein zusätzlicher zeitlicher und finanzieller Aufwand, da auf vorhandenes Datenmaterial zurückgegriffen wird, weshalb die Kosten der Datenbeschaffung in der Regel deutlich unter denen von Primärerhebungen liegen. Nachteilig ist, dass die Daten für Verwaltungszwecke und nicht zur Beantwortung von wissenschaftlichen Fragestellungen erhoben wurden. Insofern liegen oftmals wichtige erklärende Merkmale nicht vor. Zudem kann sich die Erhebungsart und der Erhebungsumfang, etwa durch das in Kraft treten neuer gesetzlichen Regelungen ändern. Im Rahmen der Sozialdaten

[7] Eine genauere Auflistung relevanter Lücken findet sich in Antoni et al. (2016), Tabelle 6.

liegt der Fokus auf Arbeitnehmerinnen und Arbeitnehmern. Insofern liegen Informationen für Selbstständige und Beamte nicht vor. Zudem handelt es sich in der Regel um Individualdaten, die den jeweiligen Haushaltskontext nicht erfassen, außer etwa bei Bedarfsgemeinschaften im SGB II-Bezug (siehe Abbildung 1).

Die Vorteile der IEB für die Mindestlohnforschung sind insbesondere die Informationen zu Personen und Betrieben im Längsschnitt. Die Nachteile sind vor allem in fehlenden aussagekräftigen Informationen zu den wöchentlichen Arbeitsstunden zu sehen. Deshalb werden auf Datenbasis der IEB oftmals Analysen ausschließlich für Vollzeitbeschäftigte durchgeführt (Möller 2016).[8] Bekannt ist allerdings, dass niedrige Löhne eher bei Teilzeit- oder befristet Beschäftigten verbreitet sind, jedoch auch bei Vollzeitbeschäftigten zugenommen haben (Mindestlohnkommission 2016).

Merkmale in der IEB

Mögliche Identifikatoren in der IEB zur Verknüpfung mit der VSE sind die ESP-ID und die systemfreie Betriebsnummer; über diese Nummern können zur Verknüpfung mit weiteren Datenquellen folgende Identifikatoren hinzugefügt werden. Dies sind die Versicherungsnummer, die BA-Kundennummer, die A2LL-Kundennummer, die XSozial-Kundennummer sowie die Betriebsnummer.

Weitere Merkmale sind u. a. Beginn- und Enddatum des Spells, Quelldatensatz, Erwerbsstatus, Geburtsdatum, Geschlecht, Erwerbsstatus vor Arbeitssuche, Abmeldegrund/Abgabegrund/Beendigungsgrund, Status nach Arbeitssuche, Beginn Arbeitslosigkeit, Dauer Arbeitslosigkeit, Tagesentgelt/Leistungssatz, Restanspruch/geplante Dauer, Wohnort und Arbeitsort (Geschäftsstelle, Gemeinde), SGB-II-Trägerart, SGB-II-Träger-Geschäftsstelle, Staatsangehörigkeit, Schulabschluss, Ausbildung, Beruf, Stellung im Beruf, Wirtschafts(unter)klasse (73/93/03/08), Gleitzone sowie Maßnahme-ID.

Zentrale Merkmale für die Mindestlohnforschung im Rahmen eines verknüpften Datensatzes sind insbesondere die tagesgenauen Angaben zu Beschäftigungs- und Arbeitslosigkeitszeiten, das Bruttotagesentgelt (bis zur Beitragsbemessungsgrenze erfasst), Beruf, Ausbildung, Arbeitsort und Wirtschaftszweig.

Verknüpfung

Aus der Perspektive der BA ist eine Verknüpfung von IEB und VSE technisch unproblematisch. Ein Verknüpfung könnte eindeutig über die Versicherungsnummer vorgenommen werden oder im Rahmen anderer personen-identifizierender Merkmale, wie es bereits häufiger erfolgreich im German Record Linkage Center praktiziert wurde.

§75 SGB X bietet die rechtliche Grundlage, nach der eine Verknüpfung von Sozialdaten für wissenschaftliche Forschungsvorhaben im Bereich der sozialen Sicherung oder der wissenschaftlichen Arbeitsmarkt- und Berufsforschung oder der Planung im Sozialleistungsbereich

[8] Zur Imputation fehlender Arbeitszeitinformationen in den Registerdaten der BA siehe Ludsteck und Thomsen (2016).

durch eine öffentliche Stelle im Rahmen ihrer Aufgaben durchgeführt werden kann. Dabei dürfen die schutzwürdigen Interessen der Betroffenen nicht beeinträchtigt werden, sofern nicht das öffentliche Interesse an der Forschung oder Planung das Geheimhaltungsinteresse erheblich überwiegt. Zudem sind die Forschungsvorhaben zeitlich befristet anzulegen und erfordern eine Genehmigung durch das BMAS.

4. Direkte Verknüpfung von IEB und VSE

Sowohl für eine direkte als auch für eine indirekte Verknüpfung von IEB und VSE bedarf es einer Klärung zahlreicher fachlicher Fragen. Die nachfolgenden Unterabschnitte beschreiben die verschiedenen Optionen. Im Hinblick auf die Verknüpfung von verschiedenen Datenquellen ist grundsätzlich zu unterscheiden, ob es sich um die Verknüpfung von Datenquellen handelt, die bereits erhoben wurden, oder ob es sich um in der Zukunft liegende Datenfusionen handelt.

4.1 Identifikatoren

Direkte Identifikatoren (deterministic record linkage): Auf Basis der Versicherungsnummer könnten Personen in der IEB mit Beschäftigungsverhältnissen in der VSE auf der Individualebene direkt verknüpft werden. Da Personen mehrere Beschäftigungsverhältnisse haben können, ist die Beschäftigungsform zu kontrollieren, und u. U. die Sozialversicherungsnummer bzw. deren Pseudonym zu erweitern, um diese kontrollieren zu können. Grundsätzlich sollte man darauf vorbereitet sein, dass bei der Verknüpfung über die Versicherungsnummer im Fall von Haupt- und Nebenbeschäftigung eine beschäftigte Person mehrere Beschäftigungsverhältnisse mit identischer Versicherungsnummer hat.

Über Betriebsnummern könnten die Betriebe aus beiden Datenquellen direkt miteinander verknüpft werden. Basierend auf Versicherungs- und Betriebsnummern könnten Personen in Betrieben dargestellt werden.

Die Verknüpfung an Hand der genannten direkten Identifikatoren bedarf dieser nicht unmittelbar im datenschutzrechtlich sensiblen Originalformat. Es genügt, sie in verfremdeter Form etwa als Pseudonym einzusetzen, solange die Verfremdung eindeutig und auf beiden Seiten (VSE und IEB) gleich ist. Hierzu können so genannte Hash-Verfahren eingesetzt werden, die die Verfremdung zwar eindeutig, aber nicht umkehrbar vornehmen. D. h., aus den verfremdeten Identifikatoren kann nicht mehr auf die ursprünglichen Identifikatoren geschlossen werden, was eine erhebliche Anonymisierung darstellt. Bei den Statistischen Landesämtern wird bislang die Versicherungsnummer getrennt von den Erhebungsdaten gespeichert vorgehalten, bis das anzuwendende Hash-Verfahren festgelegt wurde. Geplant ist, die Versicherungsnummer zu löschen, sobald per Hash-Verfahren die Verfremdung durchgeführt wurde.

4.2 Datennutzung und Datenzugang

In diesem Abschnitt soll es darum gehen aufzuzeigen, inwieweit die verschiedenen Formen der Datennutzung die Aufbereitung von Daten und deren datenschutzrechtliche Zulässigkeit beeinflussen.

Für die datenschutzrechtliche Prüfung ist die Art und Weise der Nutzung der Daten nach den jeweiligen Zugangswegen in den beteiligten Forschungsdatenzentren zu beachten. Mögliche Zugangswege zu den Daten können Mikrodatenfiles in Form von absolut anonymisierten Public Use Files (PUF) oder faktisch anonymisierte Scientific Use Files (SUF), Gastwissenschaftleraufenthalte oder die kontrollierte Datenfernverarbeitung (KDFV) sein (Bender et al. 2010). Wichtig ist hervorzuheben, dass beim KDFV wie auch beim Remote Access keine Weitergabe der Daten an die Wissenschaftlerinnen und Wissenschaftler stattfindet. Das bedeutet, dass Forschende die Daten selbst nicht sehen. Sie erhalten lediglich datenschutzrechtlich geprüfte Outputs aus statistischen Verfahren, wie etwa Regressionskoeffizienten (RatSWD 2017b). Die Wahrscheinlichkeit einer Re-Identifikation oder das so genannte Risiko der „Inferenz-Enthüllung“ (Ronning et al. 2011), anhand welcher man etwa von einem Regressionskoeffizient auf eine bestimmte Person oder einen bestimmten Betrieb schließen kann, ist – nach obligatorischer „Outputkontrolle“ in den Forschungsdatenzentren – grundsätzlich sehr gering. Diese Outputkontrolle kann jedoch vor allem bei größeren Forschungsprojekten aufwändig und damit kostenintensiv sein (RatSWD 2017c; Schiller et al. 2017).

4.3 Rückwirkende direkte Verknüpfung von IEB und VSE

Für die VSE2014 liegt keine rechtliche Grundlage für eine deterministische Datenverknüpfung vor. Um dennoch eine direkte Verknüpfung realisieren zu können, wäre die Einholung einer nachträglichen Zustimmung der Befragten zur Verknüpfung der Daten notwendig. Wie bereits in Abschnitt 2 angemerkt wurde, ist das nachträgliche Einholen einer schriftlichen Einverständniserklärung zur Verknüpfung von Daten oftmals mit geringen Rücklaufquoten und hohen Kosten verbunden.

Für die VSE2014 ist die rückwirkende Einholung der Zustimmung zur Datenverknüpfung faktisch unmöglich, da die in der 2. Stufe der Stichprobenziehung von den Betrieben zufällig ausgewählten Arbeitnehmerinnen und Arbeitnehmer den statistischen Ämtern gar nicht bekannt sind. Die Einholung könnte allenfalls über die Arbeitgeber geschehen, die jedoch nicht zu einer Kooperation verpflichtet sind. Die Kooperationsbereitschaft dürfte sehr gering sein, denn den Arbeitgebern könnte nicht mitgeteilt werden, welche Arbeitnehmerinnen und Arbeitnehmer betroffen sind. Denn das einzige Identifikationsmerkmal - die erhobene Versicherungsnummer – darf für diesen Zweck nicht verwendet werden. Die Arbeitgeber könnten allenfalls alle ihre Beschäftigten kontaktieren. Das würde die Kosten jedoch drastisch erhöhen und möglicherweise Unmut in einer Belegschaft erzeugen, die von der Übermittlung der Daten bislang keine Kenntnis hatte.

Insofern ist eine rückwirkende Verknüpfung grundsätzlich eher als Kompromiss für solche Konstellationen anzusehen, in denen keine Alternative zu dieser Möglichkeit besteht. Alternativ kann eine indirekte Verknüpfungsmöglichkeit (probabilistic record linkage) genutzt werden. Letztere wäre im Übrigen auch für die VSE2018 eine Option, soweit eine rechtliche Grundlage für eine direkte Verknüpfung nicht rechtzeitig vorliegen sollte.

5. Indirekte Verknüpfung von IEB und VSE

Aus methodischer Sicht ist die direkte Verknüpfung der Einzeldaten der IEB und der VSE stets die beste Wahl, denn auf dieser Basis entstehen keinerlei Informationsverluste: Die durch die Kombination beider Datenquellen neu verfügbaren Informationen stehen in vollem Umfang für die Analyse zur Verfügung. Die neuen Informationen können dabei als der Zusammenhang der in beiden Datensätzen getrennt beobachteten Merkmale verstanden werden. Aus statistischer Perspektive entspricht dies der gemeinsamen empirischen Verteilung von getrennt beobachteten Merkmalen. Soweit eine direkte Verknüpfung nicht in Betracht kommt, sind verschiedenen Varianten einer indirekten Verknüpfung denkbar. Wie bei der direkten spielt auch bei einer indirekten Verknüpfung die Art und Weise der Datennutzung, etwa ob der Datenzugang über SUF oder über kontrolliertes Fernrechnen erfolgen soll, eine wesentliche Rolle für die datenschutzrechtliche Bewertung der Datenverknüpfung (vgl. 4.2). Zu den vorbereitenden Maßnahmen für eine indirekte Verknüpfung eines SUF der Einkommens- und Verbrauchsstichprobe (EVS) mit der Versicherungskontenstichprobe (VSKT) siehe Beckers et al. (2012).

5.1 Identifikatoren

Indirekte Identifikatoren (probabilistic record linkage): Die Verknüpfung über indirekte Identifikatoren bietet einen erheblich größeren Grad der Anonymisierung des verknüpften Datensatzes. Bei einer indirekten Verknüpfung besteht keine Gewissheit, dass die Datensätze ein und derselben Person miteinander verknüpft wurden, hierfür besteht lediglich eine statistische Wahrscheinlichkeit. Diese Wahrscheinlichkeit hängt unter anderem von der Trennschärfe der verwendeten Identifikatoren ab. Als Faustregel gilt, dass die Trennschäfte eines Identifikators mit der Zahl seiner Ausprägungen wächst. Tabelle 3 enthält potentielle Identifikatoren, die sowohl in der VSE als auch in der IEB vorhanden sind, und eine Einschätzung über deren Trennschärfe.

Tabelle 3: Potenzielle indirekte Identifikatoren in VSE und IEB

Merkmal der IEB	Merkmal der VSE	Trennschäfte
Betriebsnummer BA	Betriebsnummer BA*	Hoch
Arbeitsort	Regionalschlüssel des Arbeitsorts (31.12.2014)	Hoch
WZ08 5-Steller	Wirtschaftszweig (WZ2008)	Hoch
Berufsuntergruppe (KldB 2010)	Ausgeübter Beruf (KldB 2010) aus TS2010	Hoch
Beginndatum des Beschäftigungsverhältnisses (Tag/Monat/Jahr)	Datum des Beschäftigungsbeginns (Monat/Jahr)	Hoch (trotz nicht erhobener Tagesangaben in der VSE)
Tagesentgelt	Bruttomonatsverdienst (April 2014)	Mittel (denn die Verdienstdefinitionen sind verschieden)
Erwerbsstatus	Personengruppe	Mittel
Geburtsjahr	Geburtsjahr	Mittel
Ausbildung	Höchster beruflicher Ausbildungsabschluss aus TS2010	Gering
Schulausbildung	Höchster allgemeinbildender Schulabschlussabschluss aus TS2010	Gering
Anzahl Beschäftigte gesamt	Arbeitnehmer des Betriebs	Gering
Geschlecht	Geschlecht	Gering

* Das Merkmal liegt bisher nicht im Datensatz vor, kann jedoch hinzugefügt werden.
Quelle: eigene Abbildung.

Das Ziel von statistischen Verknüpfungsverfahren ist, die Wahrscheinlichkeit der Treffer, also der Verknüpfung der Daten ein und derselben Person, zu maximieren. Anders gesprochen geht es darum, statistische Zwillinge etwa anhand von Nearest-Neighbour-Verfahren zu finden. Gleichzeitig besteht die Möglichkeit, die Zahl der Treffer nicht zu groß werden zu lassen, um Aspekten des Datenschutzes Rechnung zu tragen. Dies ließe sich über die Auswahl der verwendeten indirekten Identifikatoren sowie deren Kodierung steuern. So könnten z. B. Identifikatoren mit hoher Trennschärfe, wie etwa die Betriebsnummer, bewusst nicht verwendet werden. Hinsichtlich der Kodierung der Identifikatoren könnte zudem eine gezielte Reduzierung der Zahl der Ausprägungen vorgenommen werden. Im Unterschied zur direkten Verknüpfung von IEB und VSE bliebe bei einer indirekten Verknüpfung die empirische gemeinsame Verteilung unbekannt und könnte mit Hilfe von statistischen Verfahren geschätzt werden. Eine indirekte Verknüpfung erlaubt es die interessierenden Verteilungsparameter ohne systematische Verzerrungen zu schätzen, sofern ein hinreichendes Verständnis der statistische Daten erzeugenden Prozesse vorliegt.

Es gibt verschiedene methodische Hilfskonstruktionen für die indirekte Verknüpfung der beiden Datensätze. Grundsätzlich ist die indirekte Verknüpfung von Datensätzen ein regelmäßig

angewendetes Verfahren. Als Beispiel hierfür kann stellvertretend ein Projekt herangezogen werden, in dem Informationen aus Datensätzen des Forschungsdatenzentrums der Rentenversicherung (Versicherungskontenstichprobe, VSKT) und des Sozio-oekonomisches Panels (SOEP) per Datenfusion indirekt verknüpft wurden (Rasner et al. 2007). Auf den positiven Erfahrungen dieses Projekts setzte kurz darauf das von der Volkswagen Stiftung geförderten Projekt „Lebensläufe und Alterssicherung im Wandel“ (LAW) auf. Zur Beantwortung von Fragen im Hinblick auf die Höhe und Zusammensetzung der Altersvorsorge verschiedener Geburtskohorten wurden Informationen aus drei verschiedenen Datenquellen indirekt verknüpft: Es wurden die Versicherungskontenstichprobe (VSKT), der Deutsche Alterssurvey (DEAS)[9] und das Sozio-oekonomisches Panel (SOEP) per Datenfusion kombiniert (https://www.dza.de/forschung/abgeschlossene-projekte/projekt-law.html).

Verfahren zur indirekten Verknüpfung haben im Rahmen der oben genannten Projekte zu belastbaren Befunden geführt. Ferner gilt die zunehmende Verknüpfung verschiedener und zum Teil verteilter Datenquellen kombiniert mit neuen Verknüpfungsmethoden als eine von drei entscheidende Neuerung des Digitalzeitalters (RfII 2017). Entscheidend für den Erfolg der indirekten Verknüpfung wird sein, die von Rubin und Little (1987) gelegten Grundlagen mit neueren Ansätzen, etwa von Schnell (2014), zu kombinieren. Eine solche Vorgehensweise sollte eine Verwendung von State-of-the-Art-Methoden gewährleisten. Trotz methodischer Innovationen sollte berücksichtigt werden, dass bei indirekt verknüpften Datenquellen Verzerrungen vorliegen können, die Effekte auf die interne und externe Validität der Befunde haben können (Sakshaug und Antoni 2017).

Das zu lösende Problem bei einer indirekten Verknüpfung kann vereinfachend wie folgt beschrieben werden: In den VSE-Angaben fehlt das Merkmal Y für alle Datensätze, ist jedoch in den IEB für alle dortigen Datensätze verfügbar.[10] Merkmal Y kann dabei ein stetiges Merkmal sein, wie z. B. der Tagesverdienst am 30. Juni 2015, oder ein kategoriales Merkmal, wie etwa der Indikator, ob die Person am 30. Juni 2015 sozialversicherungspflichtig beschäftigt war (Y=1) oder nicht sozialversicherungspflichtig beschäftigt war (Y=0).[11] Es gilt, das Merkmal Y in der VSE zu befüllen. In beiden Datensätzen, IEB und VSE, liegen zudem gleiche Hilfsmerkmale X vor, die bei der Lösung des Problems herangezogen werden. Als X kommen vor allem die Merkmale der Tabelle 3, also die potenziellen indirekten Identifikatoren in VSE und IEB in Frage. In der Wissenschaft wird diese Problemstellung meist als Datenfusion (data fusion, file matching) bezeichnet (Rässler 2004). Hier sollen drei nahe liegende Verfahren beispielhaft angesprochen werden, um vor allem damit in Zusammenhang stehende datenschutzrechtliche Implikationen zu verdeutlichen. Wie vorne angedeutet, werden dem jeweiligen Stand der Forschung entsprechende State-of-the-Art-Methoden zur indirekten Verknüpfung verwendet.

[9] Zur Verknüpfung von VSKT und DEAS siehe Simonson et al. (2012).

[10] Grundsätzlich könnte der umgekehrte Fall genauso gut betrachtet werden: Y würde dann in den IEB Angaben enthalten und wäre in der VSE zu befüllen. Das erscheint hier jedoch nicht praktikabel. Die IEB sind der weitaus größere und für die Analysen wichtigere Datensatz. Insbesondere geht es vor allem darum, einige wenige Informationen aus der VSE in die IEB zu übertragen.

[11] Aus den IEB lässt sich zu einem Stichtag bestimmen, ob jemand sozialversicherungspflichtig beschäftigt ist und welchen Tagesverdienst er oder sie in dem betreffenden Beschäftigungsverhältnis aufweist.

5.2 Imputation

Das Problem fehlender Informationen wird in diesem Fall als Missing-Data-Problem betrachtet, wobei der Ausfallmechanismus vom Datenaufbau erzwungen ist (Missing by Design). Den hier üblicherweise angewandten Imputations-Verfahren ist gemein, dass sie sich den Hilfsmerkmalen X bedienen, die in einem möglichst engen Zusammenhang mit Y stehen. Das entspricht der Behandlung eines zufälligen, jedoch von X abhängigen Ausfalls (MAR, missing at random). Im vorliegenden Fall wird im ersten Schritt der Zusammenhang zwischen Y und X auf Basis der IEB geschätzt. Im zweiten Schritt zieht man den geschätzten Zusammenhang auf die Ausprägungen von X in der VSE heran und prognostiziert für jeden Datensatz in der VSE das Merkmal Y.

Alle Imputationsverfahren, die eine Übertragung des geschätzten Zusammenhangs erlauben, ohne dass dabei Einzeldaten der IEB übertragen werden, sind auf Basis der existierenden Rechtsgrundlagen zulässig und können unmittelbar eingesetzt werden. Diese Voraussetzung erfüllen z. B. flexible und semiparametrische Regressionsverfahren, nicht jedoch Nearest-Neighbour-Verfahren, die bekanntlich Einzeldaten als Spenderdatensätze benötigen. Die Wirksamkeit des Verfahrens hängt von der Fähigkeit der X ab, Y erklären zu können. Die X sollten deshalb mit maximalem Informationsgehalt benutzt werden. Das bedeutet, dass Variablen wie Wirtschaftszweig, Beruf und Regionalschlüssel bevorzugt in tiefster Gliederung eingesetzt werden sollten. Faktisch anonymisierte Datensätze der VSE und IEB, die zwar leicht zugänglich sind, aber genau dies verhindern, sollten nicht ohne Prüfung des vermuteten Informationsverlustes verwendet werden. Das Verfahren und die Erklärungskraft der X sollte deshalb auf den in den FDZ der Statistischen Ämter und des IAB vorliegenden Original-Datensätzen zumindest untersucht werden. Nur dann, wenn kein Informationsverlust beim Übergang zu faktisch anonymisierten Datensätzen festgestellt werden kann, können auch letztere für das Verfahren herangezogen werden.

5.3 Support Vector Machines

Support Vector Machines (SVM) stammen aus dem maschinellen Lernen und sind mathematische Verfahren der Mustererkennung (Hamel 2009). Sie werden u. a. als Klassifikator eingesetzt, wenn es darum geht, Objekte vorgegebenen Klassen zuzuweisen. Hierbei ist in der Regel unbekannt, in welche Klasse ein Objekt tatsächlich gehört. Die SVM unterteilt die Objekte in Klassen, indem es in dem Raum der Hilfsmerkmale X der Objekte Klassengrenzen einzieht. Um die Klassengrenzen herum soll dabei ein möglichst breiter Bereich frei von Objekten bleiben, sodass die Zuweisung möglichst trennscharf ist.

Ausgangsbasis für eine SVM ist stets eine Menge von Trainingsobjekten X1, für die jeweils bekannt ist, welcher Klasse Y sie zugehören. Anhand der Trainingsobjekte sucht die Methode die trennschärfsten Klassengrenzen. Mathematisch werden diese durch mehrdimensionale Vektoren beschrieben, was dem Verfahren den Namen gab. Die Klassengrenzen werden anschließend benutzt, neue Objekte X2 einer Klasse Y zuzuordnen.

Unter der hier zu lösenden Problemstellung wären die Trainingsobjekte X1 die Daten der IEB, die neuen Objekte X2 die Daten der VSE. X sind die bekannten Hilfsmerkmale in IEB und VSE. Für Y kommt jedes Merkmal infrage, wie z.B. (wie oben schon erwähnt) der Tagesverdienst am 30. Juni 2015, oder ein kategoriales Merkmal, wie etwa der Indikator, ob die Person am 30. Juni 2015 sozialversicherungspflichtig beschäftigt war (Y=1) oder nicht sozialversicherungspflichtig beschäftigt war (Y=0).[12]

Die durch die SVM festgelegten Klassengrenzen lassen sich von der IEB auf die VSE übertragen. Dabei werden jedoch gewisse Einzeldaten X1 mit übertragen, die Bestandteil der Klassengrenzen sind. Auf jetziger Rechtsgrundlage ist jedoch eine Übermittlung von Originaldaten der VSE (oder der IEB) nicht zulässig. Zulässig könnte allein die Übermittlung von faktisch anonymisierten Einzeldaten sein. Die damit verbundene Vergröberung von Variablen wie Wirtschaftszweig, Beruf und Regionalschlüssel könnte jedoch die Trennschärfe des Verfahrens erheblich schwächen.

5.4 Random (Decision) Forests

Auch Random Forests stammen aus dem maschinellen Lernen und werden u. a. als Klassifikator eingesetzt (James et al. 2015). Bei Random Forests werden die Merkmale X jedes Objekts einer Reihe von Fallprüfungen (Entscheidungen) unterworfen, um am Ende der Entscheidungskette das Objekt eindeutig einer Klasse Y zuzuweisen. Die Reihe der Entscheidungen nennt man Entscheidungsbäume (decision tree). Weil die Wahl der Entscheidungen und die Ausprägungen der X das Ergebnis beeinflussen, werden mehrere verschiedene Entscheidungsbäume eingesetzt, die im Lernprozess zufällig (random) erzeugt werden. Einem Objekt wird jene Klasse Y zugeordnet, die für dessen gegebene X am Ende der meisten Entscheidungsbäume steht. Wie bei SVMs lernt das Verfahren anhand eines Trainingsdatensatzes, der für die Trainingsobjekte die wahre Klasse Y kennt. Das Gelernte – die als optimal ermittelten Zuordnungsvorschriften – wird anschließend auf die zu klassifizierenden Objekte angewandt.

Im Rahmen der hier zu lösenden Problemstellung wären die Trainingsobjekte X1 die Daten der IEB, die zu klassifizierenden Objekte X2 die Daten der VSE. X sind die bekannten Hilfsmerkmale in IEB und VSE. Für Y kommt jedes stetige oder kategoriale Merkmal infrage (siehe oben).[13]

Von IEB zu VSE sind die ermittelten Zuordnungsvorschriften zu übertragen. Diese enthalten keine Einzeldaten. Wie einige Imputationsverfahren sind Random Forests damit auf Basis der existierenden Rechtsgrundlagen zulässig und könnten unmittelbar eingesetzt werden. Auch hier sollte Lernen und Anwenden aus den genannten Gründen zumindest zunächst auf den Original-Datensätzen stattfinden, nicht auf faktisch anonymisierten Daten.

Das Statistische Bundesamt testet Random Forests auf Basis der VSE 2014. Die Ergebnisse werden im Laufe des Jahres 2017 vorliegen.

[12] Bei stetigen Merkmalen führen die SVMs Regressionen durch.

[13] Bei stetigen Merkmalen führen die Random Forests Regressionen durch.

6. Fazit und nächste Schritte

Belastbare Information zu Arbeitsentgelten und Arbeitszeiten sind zentral für die Qualität der Evaluation eines stundenbezogenen Mindestlohns. Insofern sind valide Informationen zu Arbeitsentgelten und Arbeitszeiten für die Mindestlohnforschung von großer Bedeutung. Eine direkte Verknüpfung der IEB- und VSE-Datensätze bietet dabei die beste Option, um eine belastbare Datenbasis für die Mindestlohnforschung zu schaffen. Aus wissenschaftlicher Sicht ist sie damit eindeutig die erste Wahl. Eine indirekte Verknüpfung würde, je nach Ausgestaltung, demgegenüber einen deutlichen Qualitätsverlust in der Kombination der beiden Datenquellen bedeuten. Gleichzeitig würde auch eine lediglich indirekt verknüpfte IEB/VSE eine Verbesserung des Analysepotentials im Vergleich zum Status quo, in dem beide Datenquellen nebeneinander stehen, mit sich bringen. Insbesondere mit Blick auf die VSE2014, für die keine rechtliche Grundlage zur direkten Verknüpfung mit der IEB existiert, ist eine indirekte Verknüpfung letztlich die einzige Möglichkeit für eine Verknüpfung beider Datenquellen.

Aktuell besteht für die Evaluation des Mindestlohns seitens des Statistischen Bundesamtes keine rechtliche Regelung, die eine direkte Verknüpfung von IEB und VSE ermöglicht. Die nachträgliche Einholung einer Zustimmung zur Datenverknüpfung scheidet aus praktischen Gründen aus. Somit ist eine direkte Verknüpfung der VSE2014 mit der IEB nicht möglich. Für eine zukünftige direkte Verknüpfung von VSE2018 mit der entsprechenden IEB und den darauf folgenden Erhebungen sollte daher eine gesetzliche Grundlage geschaffen werden. Statt der Durchführung neuer, kostenintensiver Befragungen, bei denen alle relevanten Informationen gemeinsam erhoben werden könnten, würden die relevanten Daten auf Basis vorhandener Datenquellen zusammengespielt, ohne zusätzlichen Aufwand für Betriebe oder Beschäftigte zu erzeugen. Gleichzeitig würde die Verknüpfung eine enorme Erhöhung des Analysepotenzials insbesondere für längerfristige Wirkungsanalysen mit Blick auf die Bezieher niedriger Stundenlöhne sowie die Differenzierungsmöglichkeiten nach Betrieben, Branchen und Regionen bedeuten. Für eine datenschutzrechtlich konforme Verknüpfung der VSE2014 und der entsprechenden IEB kommen nach aktueller Rechtslage lediglich indirekte Verknüpfungsverfahren in Betracht, wie etwa das statistische Matching oder die Übermittlung von geschätzten Werten.

In einem nächsten Schritt sollte auf Basis des vorliegenden Papiers daher erstens geprüft werden, welche (mit Blick auf künftige Erhebungen) rechtlichen Voraussetzungen für eine direkte Verknüpfung geschaffen werden müssten. Zweitens sollte für die abgeschlossene Erhebung ein Verfahren für ein statistisches Matching entwickelt werden. Dieses ist grundsätzlich auch ohne die Schaffung einer neuen Rechtsgrundlage zulässig.

7. Literatur

Antoni, Manfred, Andreas Ganzer und Philipp vom Berge (2016), Stichprobe der Integrierten Arbeitsmarktbiografien (SIAB) 1975-2014, FDZ-Datenreport 04/2016, Forschungsdatenzentrum der Bundesagentur für Arbeit im Institut für Arbeitsmarkt- und Berufsforschung, Nürnberg.

Beckers, Benjamin, Ralf Himmelreicher und Carsten Schröder (2012), The evolution of tangibles, financial and social security wealth over the lifecycle: estimates for Germany, *Historical Social Research/Historische Sozialforschung*, 37(2), 165-184.

Bender, Stefan, Ralf Himmelreicher, Sylvia Zühlke und Markus Zwick (2010), Access to microdata from official statistics, in: German Data Forum & Rat für Sozial- und Wirtschaftsdaten (Hrsg.), Building on progress. Expanding the research infrastructure for the social, economic, and behavioral sciences, Opladen: Budrich UniPress, 215-230.

Czaplicki, Christin und Julie Korbmacher (2010), SHARE-RV: Verknüpfung von Befragungsdaten des Survey of Health, Ageing and Retirement in Europe mit administrativen Daten der Rentenversicherung, Gesundheit, Migration und Einkommensungleichheit, DRV-Schriften Band 55/2010, 28-37.

Dütsch, Matthias, Ralf Himmelreicher und Clemens Ohlert (2017), Zur Berechnung von Bruttostundenlöhnen - Verdienst(struktur)erhebung und Sozio-oekonomisches Panel im Vergleich, SOEPpaper 911 des DIW, Berlin.

Eisnecker, Philipp Simon und Martin Kroh (2017), The informed consent to record linkage in panel studies. Optimal starting wave, consent refusals, and subsequent panel attrition, *Public Opinion Quarterly*, 81(1), 131-143.

FDZ der Statistischen Ämter der Länder [Forschungsdatenzentrum der Statistischen Ämter der Länder] (2016a), Metadaten zur Verdienststrukturerhebung 2014 EVAS 62111. Teil I - Erhebung, FDZ-Standort Hessen.

FDZ der Statistischen Ämter der Länder [Forschungsdatenzentrum der Statistischen Ämter der Länder] (2016b), Metadaten zur Verdienststrukturerhebung 2014 EVAS 62111. Teil II - Scientific-Use-File, FDZ-Standort Hessen.

Forschungsdatenzentrum der BA im IAB (o. J.), Übersicht zum Datenangebot. Abrufbar unter: http://fdz.iab.de/de/FDZ_Overview_of_Data.aspx [Abfragedatum: 09.12.2016].

Hamel, Lutz (2009), Knowledge discovery with support vector machines, Hoboken John Wiley & Sons, Inc.

IAB/RWI/ISG [Institut für Arbeitsmarkt- und Berufsforschung, Rheinisch-Westfälisches Institut für Wirtschaftsforschung, Institut für Sozialforschung und Gesellschaftspolitik] (2011), Evaluation bestehender gesetzlicher Mindestlohnregelungen - Branche: Bauhauptgewerbe, Nürnberg u. a.

James, Gareth, Daniela Witten, Trevor Hastie und Robert Tibshirani (2015), An introduction to statistical learning with applications in R, New York: Springer.

Kröger, Katharina, Uwe Fachinger und Ralf Himmelreicher (2011), Empirische Forschungsvorhaben zur Alterssicherung. Eine kritische Anmerkung zur aktuellen Datenlage, RatSWD Working Paper 170, Rat für Sozial- und Wirtschaftsdaten, Berlin.

Ludsteck, Johann und Ulrich Thomsen (2016), Imputation of the working time information for the employment register data, FDZ-Methodenreport 01/2016 EN, Forschungsdatenzentrum der Bundesagentur für Arbeit im Institut für Arbeitsmarkt- und Berufsforschung, Nürnberg.

Mindestlohnkommission (2016), Erster Bericht zu den Auswirkungen des gesetzlichen Mindestlohns. Bericht der Mindestlohnkommission an die Bundesregierung nach § 9 Abs. 4 Mindestlohngesetz, Berlin.

Möller, Joachim (2016), Lohnungleichheit – Gibt es eine Trendwende?, *Wirtschaftsdienst*, 96(13), 38-44.

Rasner, Anika, Ralf K. Himmelreicher, Markus G. Grabka und Joachim R. Frick (2007), Best of both worlds - Preparatory steps in matching survey data with administrative pension records. The case of the German Socio-Economic Panel and the scientific use file completed insurance biographies 2004, SOEPpapers on Multidisciplinary Panel Data Research 70, Berlin.

Rässler, Susanne (2004), Data fusion: identification problems, validity, and multiple imputation, *Austrian Journal of Statistics*, 33(1&2), 153-171.

RatSWD [Rat für Sozial- und WirtschaftsDaten] (2017a), Die sozial-, verhaltens- und wirtschaftswissenschaftliche Survey-Landschaft in Deutschland. Empfehlungen des RatSWD, Output Series 6.

RatSWD [Rat für Sozial- und WirtschaftDaten] (2017b), Handreichung Datenschutz, Output Series 5.

RatSWD [Rat für Sozial- und WirtschaftsDaten] (2017c), Tätigkeitsbericht der akkreditierten Forschungsdatenzentren des Rates für Sozial- und Wirtschaftsdaten (RatSWD) für das Jahr 2015, Output Series 7.

RfII [Rat für Informationsinfrastrukturen] (2017), Datenschutz und Forschungsdaten. Aktuelle Empfehlungen, Göttingen.

Ronning, Gerd, Philipp Bleninger, Jörg Dreschler und Christopher Gürke (2011), Remote Access. Eine Welt ohne Mikrodaten??, FDZ-Arbeitspapier 33, Statistische Ämter des Bundes und der Länder, Forschungsdatenzentren.

Rubin, Donald B. und Roderick J. A. Little (1987), Statistical analysis with missing data, Hoboken: John Wiley & Sons.

Sakshaug, Joseph W. und Manfred Antoni (2017), Errors in Linking Survey and Administrative Data, in: Paul P. Biemer, Edith Leeuw, Stephanie Eckman, Brad Edwards, Frauke Kreuter, Lars E. Lyberg, N. Clyde Tucker und Brady T. West (Hrsg.), Total Survey Error in Practice, Hoboken: John Wiley & Sons, 557-573.

Sakshaug, Joseph W., Alexandra Schmucker, Frauke Kreuter, Mick P. Couper und Eleanor Singer (2016), Evaluating active (opt-in) and passive (opt-out) consent bias in the transfer of federal contact data to a third-party survey agency, *Journal of Survey Statistics and Methodology*, 4(3), 382-416.

Schiller, David H., Johanna Eberle, Daniel Fuß, Jan Goebel, Jörg Heining, Tatjana Mika, Dana Müller, Frank Röder, Michael Stegmann und Karsten Stephan (2017), Standards des sicheren Datenzugangs in den Sozial- und Wirtschaftswissenschaften: Überblick über verschiedene Remote-Access-Verfahren, RatSWD Working Paper Series 261.

Schnell, Rainer (2014), An efficient privacy-preserving record linkage technique for administrative data and censuses, *Statistical Journal of the IAOS*, 30(3), 263–270.

Schnell, Rainer und Stefan Bender (2010), Summary of the grant proposal "Einrichtung eines Zentrums für Record-Linkage", Nürnberg.

Schnell, Rainer, Paul B. Hill und Elke Esser (2013), Methoden der empirischen Sozialforschung, München: Oldenbourg Verlag.

Simonson, Julia, Laura Romeu Gordo und Nadiya Kelle (2012), Statistical matching of the german aging survey and the sample of active pension accounts as a source for analyzing life courses and old age incomes, *Historical Social Research / Historische Sozialforschung*, 37(2), 185-210.

Statistisches Bundesamt (2008), Klassifikation der Wirtschaftszweige, Wiesbaden.

Statistisches Bundesamt (2015), Vorschlag zur Verdiensterhebung 2015. Vorschlag zur Erhebung personenbezogener Daten über Bruttostundenverdienste und verdiensterklärende Merkmale bei einer repräsentativen Stichprobe von Arbeitgebern zum Berichtsmonat April 2015, Wiesbaden.

Statistisches Bundesamt (2016), Verdienststrukturerhebung. Erhebung der Struktur der Arbeitsverdienste nach § 4 Verdienststatistikgesetz, VSE 2014 Qualitätsbericht, Wiesbaden.

Statistisches Bundesamt (2017), Verdiensterhebung 2015. Abschlussbericht einer Erhebung über die Wirkung des gesetzlichen Mindestlohns auf die Verdienste und Arbeitszeiten der abhängig Beschäftigten, Wiesbaden.

vom Berge, Philipp, Hans Verbeek, Matthias Umkehrer, Michael Fertig und Stefan Bender (2014), Vorbereitende Forschung für die zweite Evaluationsrunde Mindestlöhne - Erschließung neuer Datenquellen, FDZ-Methodenreport 3/2014, Forschungsdatenzentrum der Bundesagentur für Arbeit im Institut für Arbeitsmarkt- und Berufsforschung, Nürnberg.

Zimmer, Elke (2015), Verdiensterhebung 2015. Eine Erhebung nach § 7 Bundesstatistikgesetz, *Zeitschrift für amtliche Statistik Berlin Brandenburg*, 9(4), 14-17.

8. Anhang

Anhangverzeichnis:

A1 Tabelle: Klassifikation der Wirtschaftsabschnitte
A2 VSE2014-Betriebsbogen
A3 VSE2014-Beschäftigtenbogen A4 Tabelle: Merkmale und Datenquelle

Anhang A1:

Tabelle: ***Klassifikation der Wirtschaftsabschnitte***

Abschnitt	**Wirtschaftsabschnitte***
A	Land- und Forstwirtschaft, Fischerei
B	Bergbau und Gewinnung von Steinen und Erden
C	Verarbeitendes Gewerbe
D	Energieversorgung
E	Wasserversorgung; Abwasser- und Abfallentsorgung und Beseitigung von Umweltverschmutzungen
F	Baugewerbe
G	Handel; Instandhaltung und Reparatur von Kraftfahrzeugen
H	Verkehr und Lagerei
I	Gastgewerbe
J	Information und Kommunikation
K	Erbringung von Finanz- und Versicherungsdienstleistungen
L	Grundstücks- und Wohnungswesen
M	Erbringung von freiberuflichen, wissenschaftlichen und technischen Dienstleistungen
N	Erbringung von sonstigen wirtschaftlichen Dienstleistungen
O	Öffentliche Verwaltung, Verteidigung; Sozialversicherung
P	Erziehung und Unterricht
Q	Gesundheits- und Sozialwesen
R	Kunst, Unterhaltung und Erholung
S	Erbringung von sonstigen Dienstleistungen
T	Private Haushalte mit Hauspersonal; Herstellung von Waren und Erbringung von Dienstleistungen durch private Haushalte für den Eigenbedarf ohne ausgeprägten Schwerpunkt
U	Exterritoriale Organisationen und Körperschaften

Anmerkung: Klassifikation der Wirtschaftszweige, Ausgabe 2008 (WZ 2008)

Quelle: (Statistisches Bundesamt 2008)

Anhang A2:

VSE 2014 - Betriebsbogen

Verdienststrukturerhebung 2014 **VS0** Rücksendung bitte bis 31. März 2015

Betriebsbogen

Ansprechpartner/-in für Rückfragen (freiwillige Angabe)

Name:

Telefon oder E-Mail:

Vielen Dank für Ihre Mitarbeit.

Rechtsgrundlagen und weitere rechtliche Hinweise entnehmen Sie der beigefügten Unterlage, die Bestandteil dieses Fragebogens ist.
Bitte beachten Sie bei der Beantwortung der Fragen die Erläuterungen zu 1 bis 24 in der separaten Unterlage.

Falls Anschrift oder Firmierung nicht mehr zutreffen, bitte auf Seite 2 korrigieren.

Identnummer (bei Rückfragen bitte angeben)

A Angaben über das Unternehmen

Wirtschaftszweig | 0 Bogenart | Identnummer

1 Einfluss der öffentlichen Hand auf die Unternehmensführung
Bitte zutreffende Ziffer in das nebenstehende Feld eintragen. 09

1 = Kein oder eingeschränkter Einfluss der öffentlichen Hand auf die Unternehmensführung durch Kapitalbeteiligung (50 % oder weniger), Satzung oder sonstige Bestimmungen.

2 = Beherrschender Einfluss der öffentlichen Hand auf die Unternehmensführung durch Kapitalbeteiligung (mehr als 50 %), Satzung oder sonstige Bestimmungen.

2 Anzahl aller Arbeitnehmerinnen/Arbeitnehmer **des Unternehmens am 30. April 2014** 1 10

B Angaben über den Betrieb

1 **Wirtschaftliche Tätigkeit**
Falls die wirtschaftliche Tätigkeit von der bereits vorgedruckten abweicht, korrigieren Sie diese bitte. Bei der Ausführung verschiedener Tätigkeiten geben Sie diejenige an, mit der die überwiegende Anzahl der Arbeitnehmerinnen/Arbeitnehmer beschäftigt ist.

2 Anzahl aller Arbeitnehmerinnen/Arbeitnehmer **im Betrieb** mit Vergütung für den gesamten Monat April 2014. 1 2

Männer 11

Frauen 12

3 In Betrieben ab einer bestimmten Größe muss nicht für alle unter B2 erfassten Beschäftigten der Arbeitnehmerbogen ausgefüllt werden. Sofern diese Möglichkeit für Sie besteht, sind hier Auswahlvorgaben eingetragen. Erfassen Sie die Beschäftigten Ihrer Verdienstliste ab der Startzahl fortlaufend nach dem Auswahlabstand. 3

Startzahl

Auswahlabstand

Alternativ können Sie alle unter B2 erfassten Beschäftigten im Arbeitnehmerbogen eintragen, die Auswahl übernimmt das statistische Amt.

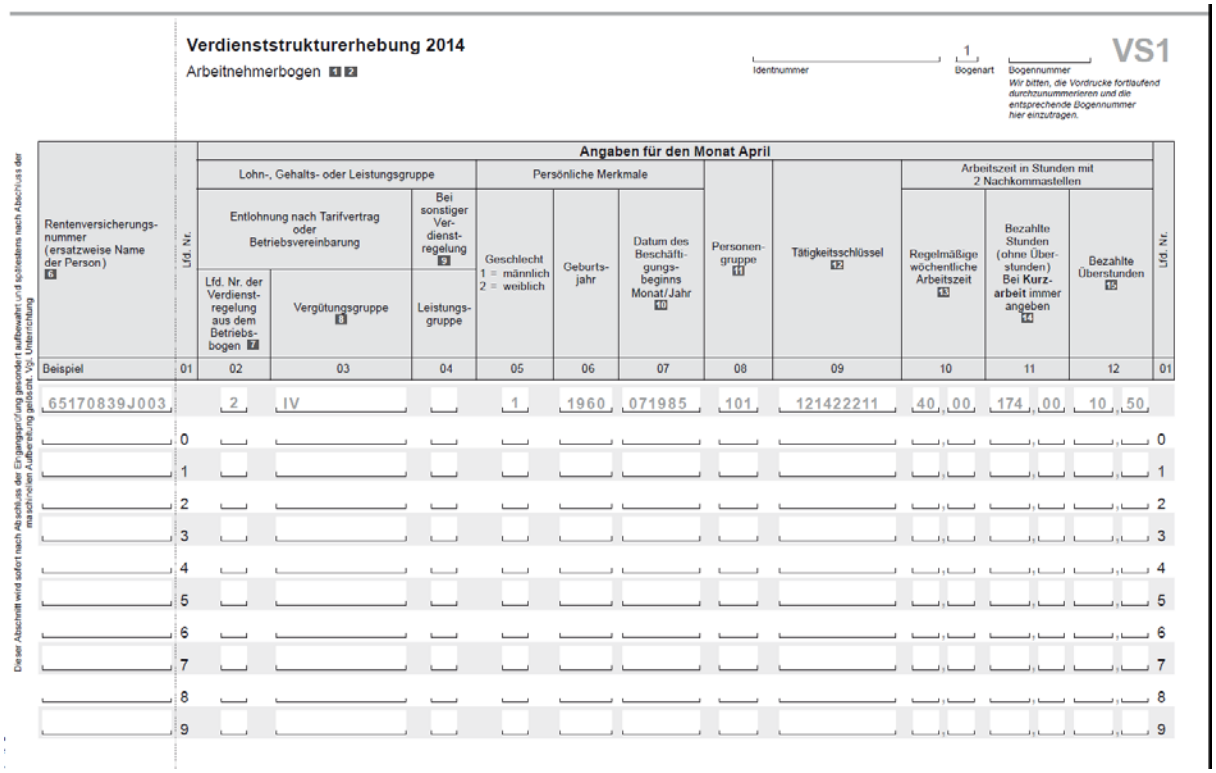
Anzahl der von Ihnen insgesamt beigefügten, ausgefüllten Arbeitnehmerbogen .....

4 Anzahl der Wochentage, die der Berechnung des Urlaubsanspruchs eines Vollzeitbeschäftigten zugrunde liegt. 14

5 Betriebsübliche Wochenarbeitszeit eines Vollzeitbeschäftigten in Stunden. ..... 15 ,

Anhang A3:

VSE 2014 – Arbeitnehmerbogen



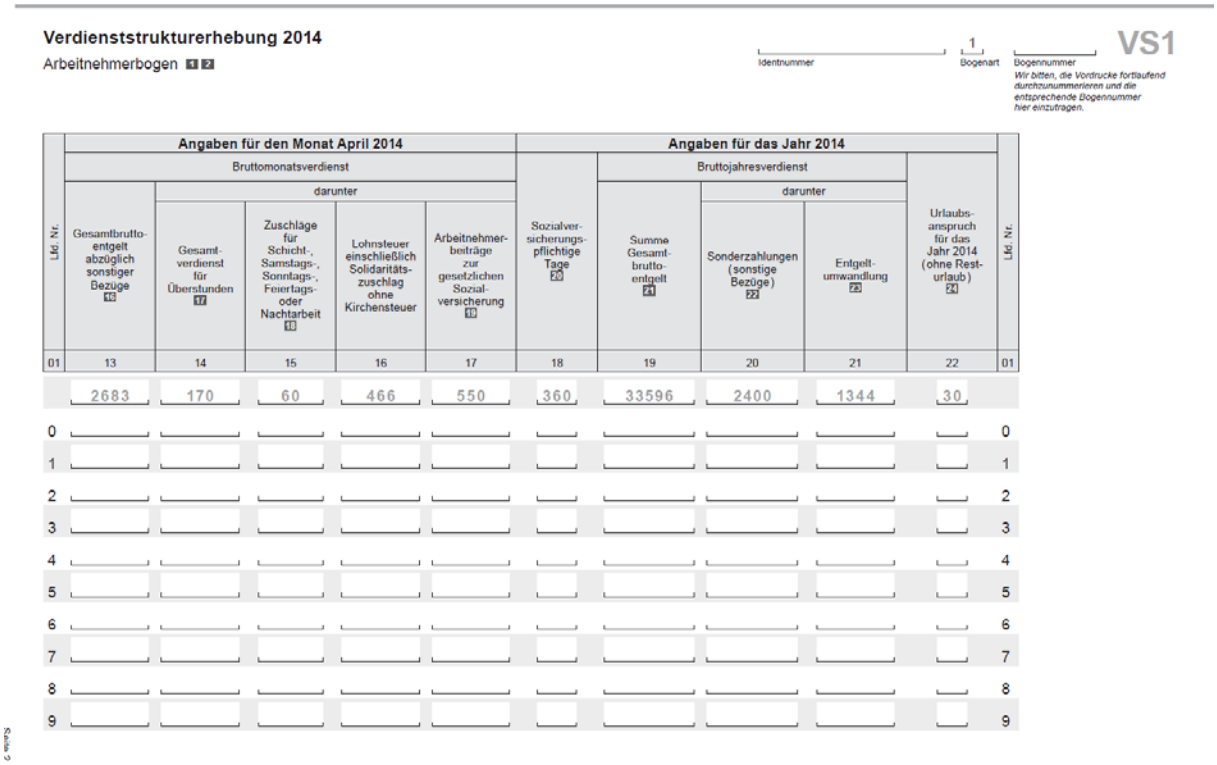
Verdienststrukturerhebung 2014
Arbeitnehmerbogen 1 2

Identnummer | 1 Bogenart | Bogennummer | VS1

Wir bitten, die Vordrucke fortlaufend durchzunummerieren und die entsprechende Bogennummer hier einzutragen.

Dieser Abschnitt wird sofort nach Abschluss der Eingangsprüfung gesondert aufbewahrt und schnellstmöglich nach Abschluss der maschinellen Aufbereitung gelöscht. Vgl. Unterrichtung

Rentenversicherungsnummer (ersatzweise Name der Person) 6	Lfd. Nr.	Angaben für den Monat April: Lohn-, Gehalts- oder Leistungsgruppe: Entlohnung nach Tarifvertrag oder Betriebsvereinbarung: Lfd. Nr. der Verdienstregelung aus dem Betriebsbogen 7	Entlohnung nach Tarifvertrag oder Betriebsvereinbarung: Vergütungsgruppe 8	Bei sonstiger Verdienstregelung 9: Leistungsgruppe	Persönliche Merkmale: Geschlecht 1 = männlich 2 = weiblich	Geburtsjahr	Datum des Beschäftigungsbeginns Monat/Jahr 10	Personengruppe 11	Tätigkeitsschlüssel 12	Arbeitszeit in Stunden mit 2 Nachkommastellen: Regelmäßige wöchentliche Arbeitszeit 13	Bezahlte Stunden (ohne Überstunden) Bei Kurzarbeit immer angeben 14	Bezahlte Überstunden 15	Lfd. Nr.
Beispiel	01	02	03	04	05	06	07	08	09	10	11	12	01
65170839J003		2	IV		1	1960	071985	101	121422211	40,00	174,00	10,50	
	0												0
	1												1
	2												2
	3												3
	4												4
	5												5
	6												6
	7												7
	8												8
	9												9



Verdienststrukturerhebung 2014
Arbeitnehmerbogen 1 2

Identnummer | 1 Bogenart | Bogennummer | VS1

Wir bitten, die Vordrucke fortlaufend durchzunummerieren und die entsprechende Bogennummer hier einzutragen.

Lfd. Nr.	Angaben für den Monat April 2014: Bruttomonatsverdienst: Gesamtbruttoentgelt abzüglich sonstiger Bezüge 16	darunter: Gesamtverdienst für Überstunden 17	darunter: Zuschläge für Schicht-, Samstags-, Sonntags-, Feiertags- oder Nachtarbeit 18	darunter: Lohnsteuer einschließlich Solidaritätszuschlag ohne Kirchensteuer	darunter: Arbeitnehmerbeiträge zur gesetzlichen Sozialversicherung 19	Angaben für das Jahr 2014: Sozialversicherungspflichtige Tage 20	Bruttojahresverdienst: Summe Gesamtbruttoentgelt 21	darunter: Sonderzahlungen (sonstige Bezüge) 22	darunter: Entgeltumwandlung 23	Urlaubsanspruch für das Jahr 2014 (ohne Resturlaub) 24	Lfd. Nr.
01	13	14	15	16	17	18	19	20	21	22	01
	2683	170	60	466	550	360	33596	2400	1344	30	
0											0
1											1
2											2
3											3
4											4
5											5
6											6
7											7
8											8
9											9