

Erweiterung der Schätzmodellierung zum Ausgleich operativ untererfasster Eckwerte auf Regionalebene in der Leistungsstatistik SGB II

Projektion der vorläufigen, untererfassten Eckwerte
(erhoben ohne Wartezeit) auf das Wartezeitniveau

Nürnberg, im November 2011



Impressum

Titel:	Methodenbericht der Statistik der BA Ausgleich operativ untererfasster Eckwerte auf Regionalebene in der Leistungsstatistik SGB II
Herausgeber:	Bundesagentur für Arbeit Statistik Nürnberg
Erstellungsdatum:	02.11.2011
Autor(en):	Christiane Tentscher Matthias Wolff

Weiterführende statistische Informationen:

Internet	http://statistik.arbeitsagentur.de
Hotline	01801 / 78 722 10 (Hotline) *
Fax	01801 / 78 722 11 * *) 3,9 Cent je Minute aus dem Festnetz der Deutschen Telekom. Bei Anrufen aus Mobilfunknetzen höchstens 42 ct / min.
E-Mail	service-haus.datenzentrum@arbeitsagentur.de

© Statistik der Bundesagentur für Arbeit, Nürnberg, 2011

Für nichtgewerbliche Zwecke sind Vervielfältigung und unentgeltliche Verbreitung, auch auszugsweise, mit genauer Quellenangabe gestattet.

Die Verbreitung, auch auszugsweise, über elektronische Systeme/Datenträger bedarf der vorherigen Zustimmung.

Alle übrigen Rechte vorbehalten.

Zusammenfassung

Gesicherte statistische Aussagen über Niveau und Entwicklungen lassen sich im Bereich der Grundsicherungsstatistik nach dem SGB II aufgrund der operativen Untererfassungen am aktuellen Rand nur über Zeiträume treffen, die mindestens drei Monate zurückliegen (Wartezeit). Die Nutzer der amtlichen Statistik nach dem SGB II vor allem aus Politik, Verwaltung und Medien möchten aber möglichst zeitnah schon vor Ablauf der dreimonatigen Wartezeit Entwicklungen erkennen können. Damit für die Beurteilung der aktuellen Entwicklungen zeitnahe Informationen bereitgestellt werden können, wurden seit Januar 2011 für Kreise, Länder und Deutschland insgesamt die Eckwerte für Bedarfsgemeinschaften (BG), erwerbsfähige (eLb) und nicht erwerbsfähige Leistungsberechtigte (nEf) auf den erwarteten 3-Monatswert hochgerechnet. Das Hochrechnungsverfahren wurde in einem Methodenbericht vorgestellt, in dem auch ein Review der Modellierung angekündigt wurde. Dieses Review der Modellierung umfasste die folgenden Punkte.

- Erweiterung der Berichtsfähigkeit für die SGBII und SGBIII- Trägerstrukturen
- Einbezug von Nettoarbeitstagen für zugelassene kommunale Träger
- Verwendung der integrierten Datenbasis der Grundsicherungsstatistik
- Einbezug sukzessiv vervollständigter Wartezeitinformationen

Einige dieser Punkte wurden schon zu früheren Zeitpunkten umgesetzt und in der Schätzmodellierung und der statistischen Berichterstattung berücksichtigt werden. Andere Zielstellungen bedurften umfänglicher Vorbereitungen und kommen erst mit der Veröffentlichung der Stichtagsergebnisse vom Oktober 2011 erstmals als Teil der Schätzmodellierung zum Einsatz.

In diesem Methodenbericht soll es nun darum gehen, zu erläutern wie die einzelnen Review-Bestandteile umgesetzt wurden und wie sie im Kontext der Gesamtkonzeption der Schätzmodellierung wirken.

Dabei werden zunächst die datenseitigen Rahmenbedingungen dargelegt um anhand dessen die Konzeption der Schätzmodellierung zu erläutern. In diesem Kontext werden die einzelnen Regressoren und Ihr jeweiliger Erklärungsgehalt für die Schätzung insgesamt beschrieben. Abschließend wird ein zusammenfassender Überblick über die Ergebnislage gegeben, wobei die neuen Ergebnisse mit denen der alten Konzeption verglichen werden.

Inhaltsverzeichnis

Zusammenfassung	3
Inhaltsverzeichnis	4
Abbildungsverzeichnis	5
Tabellenverzeichnis	5
Formelverzeichnis	5
1 Einleitung	6
2 Rahmenbedingungen und Ausgangssituation	8
3 Konzeption des Schätzverfahrens	13
3.1 Modellierung – Multivariate lineare Regression	13
3.1.1 Erklärende Variablen: Regressoren	14
3.1.1.1 Aktueller Bestand ohne bzw. mit geringer Wartezeit	16
3.1.1.2 Trend des Nacherfassungsverhaltens	17
3.1.1.3 Nettoarbeitstage – der Einfluss des statistischen Stichtags	18
3.1.1.4 Arbeitslosigkeit als Frühindikator für Hilfebedürftigkeit	20
3.1.1.5 Interaktionsvariablen im Modell	22
3.1.2 Ermittlung der Schätzwerte	23
3.1.2.1 Ermittlung des Schätzwertes zum Bestand Personen	24
3.1.2.2 Ermittlung der Länderergebnisse	24
3.1.2.3 Ermittlung der Ergebnisse für SGBII und SGBIII-Trägergebiete	24
3.1.2.4 Rundung	25
3.1.3 Standardfehler der Schätzung	25
4 Ergebnislage	27
5 Datenqualität und monatliche Qualitätssicherung	30
6 Angebot Datenmaterial	31
7 Ansprechpartner	32
Anhang	33
A) Begriffsdefinition – Stichtag	33
B) Begriffsdefinition – Nacherfassungsquote	33
C) Begriffsdefinition – Stock-Flow-Modell	33
D) Anwendung und Interpretation von Boxplots	34
E) Anwendung und Interpretation von Histogrammen und Dichtefunktionen	35
F) Wartezeitkonzept in der Grundsicherungsstatistik	37
Quellenverzeichnis	39

Abbildungsverzeichnis

<u>Abbildung 1:</u> Streuung der Nacherfassungsquoten für unterschiedliche Wartezeitstände.....	8
<u>Abbildung 2:</u> Zeitreihe zum Nacherfassungsverhalten am Bestand an BGn	11
<u>Abbildung 3:</u> Wirkung des Einbezugs aktuellerer Wartezeitstände in der Schätzung.....	16
<u>Abbildung 4:</u> Trend der Nacherfassungsquoten nach Wartezeit	17
<u>Abbildung 5:</u> Einfluss der Nettoarbeitstage auf die Nacherfassungsquoten	18
<u>Abbildung 6:</u> trägerformbezogene Ergebnisbetrachtung	20
<u>Abbildung 7:</u> Einfluss der Regressoren der Arbeitslosenstatistik	21
<u>Abbildung 8:</u> kreisbezogene Abweichungen der Schätzergebnisse vom Datenstand WZ3 ...	27
<u>Abbildung 9:</u> Entwicklung der Schätzergebnisse auf Bundesebene – absolut	28
<u>Abbildung 10:</u> Lesehilfe Boxplot	34
<u>Abbildung 11:</u> Boxplot für Normalverteilung.....	35
<u>Abbildung 12:</u> Histogramm und Dichtekurve zu Nacherfassungsquoten (Bestand BGn).....	37
<u>Abbildung 13:</u> Wartezeitkonzept - Sicht aktueller Erhebungszeitpunkt.....	38
<u>Abbildung 14:</u> Wartezeitkonzept - Sicht Wartezeitstände für ersten Erhebungszeitpunkt.....	38

Tabellenverzeichnis

<u>Tabelle 1:</u> Nacherfassungsquoten in Prozent; Daten ohne Wartezeit	10
<u>Tabelle 2:</u> Nacherfassungsquoten in Prozent; Daten nach einmonatiger Wartezeit.....	10
<u>Tabelle 3:</u> Nacherfassungsquoten in Prozent; Daten nach zweimonatiger Wartezeit.....	10
<u>Tabelle 4:</u> Terminübersicht zur Berechnung der Nettoarbeitstage	19
<u>Tabelle 5:</u> Ergebnisübersicht der Schätzung Oktober 2011 für einen Beispielkreis.....	25
<u>Tabelle 6:</u> Ergebnislage alte Konzeption	29
<u>Tabelle 7:</u> Ergebnislage neue Konzeption	29

Formelverzeichnis

<u>Formel 1:</u> Berechnung Regressionskoeffizienten, allgemein.....	15
<u>Formel 2:</u> Berechnung der Regressionskoeffizienten für Daten ohne Wartezeit	15
<u>Formel 3:</u> Berechnung der Regressionskoeffizienten für Daten mit einmonatiger Wartezeit .	15
<u>Formel 4:</u> Berechnung der Regressionskoeffizienten für Daten mit zweimonatiger Wartezeit .	15
<u>Formel 5:</u> Schätzgleichung für Daten ohne Wartezeit.....	23
<u>Formel 6:</u> Schätzgleichung für Daten mit einmonatiger Wartezeit.....	23
<u>Formel 7:</u> Schätzgleichung für Daten mit zweimonatiger Wartezeit	23
<u>Formel 8:</u> Berechnung des Standardfehlers der Schätzung für Schätzlauf Oktober 2011.....	25
<u>Formel 9:</u> Berechnung der Nacherfassungsquoten.....	33

1 Einleitung

Im Rahmen des zweiten Sozialgesetzbuches (SGB II) übernehmen die Träger der Grundsicherung für Arbeitsuchende die Aufgabe der umfassenden Betreuung der hilfebedürftigen Personen. Hierzu gehört unter anderem auch die Erfassung der leistungsrelevanten Informationen zum Leistungsberechtigten und seiner Bedarfsgemeinschaft in das jeweilige Fachverfahren im Rahmen der Leistungsgewährung.

Ein wichtiger Punkt im Rahmen dieser Zuständigkeit ist, dass der Prozess der Leistungsgewährung bzw. Feststellung der Hilfebedürftigkeit einige Zeit in Anspruch nimmt, da alle leistungsrelevanten Informationen (Bedarfe, Einkommen, Wohninformationen etc.) für die einzelnen Mitglieder der Bedarfsgemeinschaft festgestellt und erfasst werden müssen. Zusätzlich können Konstellationen eintreten, in denen Leistungsberechtigte nicht oder erst verspätet Änderungen in Ihren Beschäftigungs-, Einkommens- oder Vermögensverhältnissen angeben, was zu höheren Untererfassungen am aktuellen Rand führt.

Dies beeinflusst letztlich auch die statistische Berichterstattung, da zum statistischen Stichtag¹ noch nicht alle berichtsrelevanten Daten erfasst sind. In der statistischen Berichterstattung über die Lage am aktuellen Stichtag entsteht damit ein Wissensdefizit. Wie in allen leistungsgewährenden Verfahren bedeutet dies auch im SGB II, dass zum Stichtag noch nicht alle Bedarfe, Ansprüche und Leistungen bekannt sind, die für diesen Monat realisiert werden. Diesem Wissensdefizit ist die Notwendigkeit geschuldet, die Berichterstattung auf eine gesicherte Datenbasis zu stellen, welche nach einer definierten Wartezeit verfügbar ist. Mit der Wartezeit erhöht sich der Wissenstand über den Berichtszeitraum, es lässt sich erst dann ein endgültiges Bild der tatsächlichen Anspruchslage zeichnen.

Die Statistik der Bundesagentur für Arbeit hat bereits einen Monat nach Inkrafttreten des SGB II, am 2. Februar 2005, die Berichterstattung über Personen, Bedarfsgemeinschaften und deren Leistungen im SGB II aufgenommen. Zu diesem Zeitpunkt konnten die Informationen nur aus den aktuellen Daten, ohne Wartezeit, gewonnen werden. Mitte 2005 wurde im Bereich der Leistungsstatistik SGB II zusätzlich die Berichterstattung auf Basis vervollständigter Daten nach einer Wartezeit etabliert. Aus den oben genannten Gründen lassen allein Auswertungen nach einer Wartezeit verlässliche und belastbare Aussagen über die Entwicklung der Grundsicherung zu. Die Erfahrung hat gezeigt, dass für die Berichterstattung aus der Grundsicherungsstatistik eine Wartezeit von drei Monaten zur Vervollständigung der Datenbasis ausreicht. Die statistische Berichterstattung auf Basis von Daten ohne Wartezeit wurde sukzessive reduziert.

Das Wartezeitkonzept² hat den Nachteil, dass gesicherte Aussagen über Lage und Entwicklung in der Grundsicherungsstatistik nur mit einem Zeitverzug von drei Monaten und nicht monatsaktuell wie in der Arbeitslosenstatistik gemacht werden können. Am aktuellen Rand stehen nur die wenig belastbaren vorläufigen (und unvollständigen) Daten zur Verfügung. In einem ersten Schritt wurde deshalb Anfang 2008 ein Schätzmodell eingeführt, welches auf der Ebene des Bundes sowie der Länder die Untererfassung durch ein Hochrechnungsverfahren ausgleicht. Basierend auf den Informationen in der Vergangenheit sowie der Bearbei-

¹ Begriffsdefinition im Anhang → A) Begriffsdefinition – Stichtag

² Erläuterung zum Wartezeitkonzept der Grundsicherungsstatistik im Anhang → F) Wartezeitkonzept in der Grundsicherungsstatistik

tungszeit bis zum statistischen Stichtag wurden für alle Bundesländer Schätzfaktoren ermittelt, die das Wissensdefizit vorläufig ausgleichen und die statistische Berichterstattung bis zum aktuellen Rand ermöglichen konnten.

Mit Veröffentlichung der Ergebnisse im Januar 2011 gab es dann erstmals auch auf Kreis- und Trägerebene hochgerechnete Ergebnisse. Die hierfür eingesetzte Modellierung konnte als Erweiterung des bestehenden Modells beschrieben werden und hatte sich im Vergleich zu alternativen Schätzmodellierungen als besonders treffsicher herausgestellt.

Einige Defizite konnten aber schon im Rahmen der erstmaligen Veröffentlichung identifiziert werden, sodass mit der Erstveröffentlichung der Ergebnisse auch eine Entwicklungsperspektive benannt wurde. Die darin benannten und im Folgenden aufgeführten Punkte wurden bis Oktober 2011 umgesetzt.

- Erweiterung der Berichtsfähigkeit für die SGBII und SGBIII- Trägerstrukturen
- Einbezug von Nettoarbeitstagen für zugelassene kommunale Träger
- Verwendung der integrierten Datenbasis der Grundsicherungsstatistik
- Einbezug sukzessiv vervollständigter Wartezeitinformationen

Einige dieser Punkte wurden schon zu früheren Zeitpunkten umgesetzt und in der Schätzmodellierung und der statistischen Berichterstattung berücksichtigt werden. Andere Zielstellungen bedurften umfänglicher Vorbereitungen und kommen erst mit der Veröffentlichung der Stichtagsergebnisse vom Oktober 2011 erstmals als Teil der Schätzmodellierung zum Einsatz.

2 Rahmenbedingungen und Ausgangssituation

a) Nacherfassungsquoten im Vergleich der Grundsicherungsträger

Bezogen auf die kreisübergreifende Betrachtung der Bestandseckwerte am Stichtag – Anzahl Bedarfsgemeinschaften und Personen (darunter: erwerbsfähige und nicht erwerbsfähige Leistungsberechtigte) liegt der Mittelwert der Nacherfassungsquoten³ von Daten ohne Wartezeit gegenüber Daten mit dreimonatiger Wartezeit bei etwa 4 bis 5 Prozent. Die Unterfassung wird nach Ablauf der definierten Wartezeit schrittweise aufgelöst. So sinkt der Mittelwert der Nacherfassungsquoten nach Ablauf eines Monats Wartezeit um etwa 3,5 Prozentpunkte, nach Ablauf des zweiten Monats Wartezeit nochmals um rund 0,5 Prozentpunkte. Abbildung 1 visualisiert den Effekt anhand der Bestandsdaten zu Bedarfsgemeinschaften. Die obere Abbildung zeigt den direkten Vergleich der Verteilungen der Nacherfassungsquoten anhand sogenannter Boxplots⁴ für die kreisbezogenen Beobachtungen von Januar 2007 bis Juni 2011 (insgesamt 54 Monate x 412 Kreise – 22.248 Beobachtungen). Zu erkennen ist

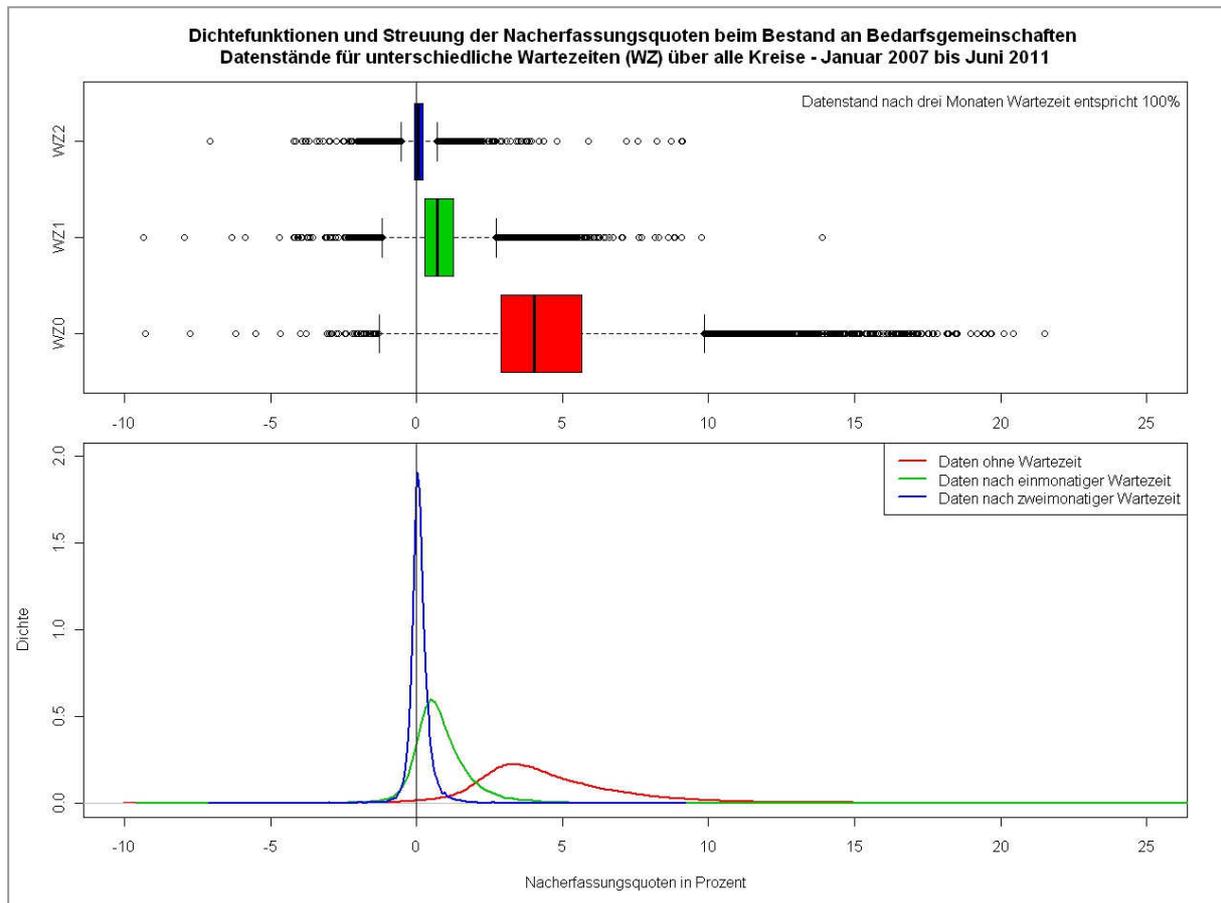


Abbildung 1: Streuung der Nacherfassungsquoten für unterschiedliche Wartezeitstände

einerseits, dass die Boxen mit zunehmendem Wartezeitstand deutlich schmaler werden, was darauf hinweist, dass die mittleren 50% der Daten näher beieinander liegen, der Interquartilsabstand mit Ablauf der Wartezeit also deutlich kleiner wird. Des Weiteren wird auch eine

³ Die Nacherfassungsquote ist ein Maß für die Unterfassung von Daten ohne Wartezeit (WZ0) bzw. mit geringer Wartezeit (WZ1 und WZ2) im Vergleich zu der nach dreimonatiger Wartezeit vervollständigten Datenbasis (WZ3) über einen Berichtsmonat. Siehe hierzu auch Anhang → B) Begriffsdefinition – Nacherfassungsquote

⁴ Lesehilfe zu Boxplots im Anhang → D) Anwendung und Interpretation von Boxplots

Verschiebung der Boxen hin zur 0%-Linie der Nacherfassungsquoten deutlich. Zu erkennen ist darüber hinaus, dass der größte Zugewinn an Informationen nach Wartezeit von einem Monat zu verzeichnen ist.

Die untere Graphik zeigt die korrespondierenden Dichtefunktionen⁵. Man erkennt zunächst dass die Verteilungen im Rahmen der Wartezeit schmäler werden, was den Effekt verdeutlicht, dass die kreisbezogenen Nacherfassungsquoten näher beieinander liegen, die Streuung der Daten also abnimmt. Auch hier ist die Verschiebung der Kurven hin zur 0%-Grenze deutlich. Die Erwartungswerte der Dichtefunktionen, abzulesen am Maximum der jeweiligen Kurve, nähern sich deutlich der Nacherfassungsquote von 0% an. Dichtefunktionen liest und interpretiert man grundsätzlich über die Betrachtung der Fläche unter der Kurve. Die Fläche unter der Kurve links und rechts des Erwartungswertes der Verteilungen gibt an, wie groß der prozentuale Anteil der monatsweisen Kreisbeobachtungen ist, die eine Nacherfassungsquote um den Erwartungswert aufweisen. Die Wahrscheinlichkeit größere oder kleinere Nacherfassungsquoten zu beobachten sinkt mit zunehmender Wartezeit deutlich.

Die Verteilungen gleichen zudem Normalverteilungen. Sie sind insbesondere für Zufallsgrößen typisch, die von sehr vielen externen und zusammenwirkenden Einflussgrößen bestimmt sind. Typisch für diese Kurvendarstellung ist, dass sie links und rechts vom Maximum aus glockenförmig gegen Null abfällt.

Was an beiden Darstellungsformen deutlich wird ist überdies, dass die Ausgangsbedingungen von regional starken Disparitäten geprägt sind. Beide Darstellungen basieren auf kreisbezogenen Nacherfassungsquoten im Zeitraum Januar 2007 bis Juni 2011, weshalb neben dem Gros der erwartungswertnahen Beobachten eben auch Ausreißer zu beobachten sind. Die Streuung dieser verringert sich zwar mit Ablauf der Wartezeit. In Relation des jeweiligen Interquartilsabstandes sind diese Ausreißer aber nach wie vor als wichtiger Teil der Ausgangssituation einer Schätzmodellierung (auf Regionalebene) zu berücksichtigen. Tabelle 1 bis 3 zeigen die Ausgangssituation anhand ausgewählter Verteilungsmaße. Für den jeweiligen Bestandseckwert kann hierin die Verteilung der Daten insgesamt abgelesen werden. Auf die Darstellung der minimalen und maximalen Nacherfassungsquoten wurde aufgrund der Ausreißerproblematik verzichtet. Stattdessen wurden die Grenzen der oberen und unteren 2,5%-Marke datenmäßig ausgewiesen (wie beim „seven-number-summary“). Unter Auslassung der oberen und unteren 500 Fälle (bezogen auf 22.248 Beobachtungen insgesamt) markieren sie die um Ausreißer bereinigten minimalen und maximalen Ausprägungen der Verteilung der Nacherfassungsquoten auf Regional- und Monatsebene. Die in der Tabelle ausgewiesenen negativen Nacherfassungsquoten beschreiben eine Ausnahmesituation beim Nacherfassungsverhalten der Träger der Grundsicherung. Neben dem typischen Fall der nachträglichen Erfassung und Bewilligung von Leistungsfällen kann es auch zu rückwirkenden Aufhebungen von Fällen kommen. Bei einzelnen Grundsicherungsträgern kann dies auch negative Nacherfassungsquoten nach sich ziehen.

⁵ Lesehilfe zur empirischen Verteilungsfunktion im Anhang → E) Anwendung und Interpretation von Histogrammen und Dichtefunktionen

Nacherfassungsquoten über alle Kreise und Monate (Januar 2007- Juni 2011)	BG	eLb	nEf	Pers
Untere 2,5%	0,7	-0,6	-0,5	0,2
Untere 10% (unteres Dezil)	2,0	-0,1	-0,2	1,8
Untere 25% (unteres Quantil)	2,9	0,3	-0,1	2,9
Median	4,0	0,7	0,1	4,1
Obere 25% (oberes Quantil)	5,7	1,3	0,2	5,9
Obere 10% (oberes Dezil)	7,6	2,0	0,5	8,2
Obere 2,5%	10,5	3,2	0,9	11,8

Tabelle 1: Nacherfassungsquoten in Prozent; Daten ohne Wartezeit

Nacherfassungsquoten über alle Kreise und Monate (Januar 2007- Juni 2010)	BG	eLb	nEf	Pers
Untere 2,5%	-0,6	-0,5	0,2	-0,7
Untere 10% (unteres Dezil)	-0,1	-0,2	1,8	-0,1
Untere 25% (unteres Quantil)	0,3	-0,1	2,9	0,3
Median	0,7	0,1	4,1	0,7
Obere 25% (oberes Quantil)	1,3	0,2	5,9	1,3
Obere 10% (oberes Dezil)	2,0	0,5	8,2	2,2
Obere 2,5%	3,2	0,9	11,8	3,6

Tabelle 2: Nacherfassungsquoten in Prozent; Daten nach einmonatiger Wartezeit

Nacherfassungsquoten über alle Kreise und Monate (Januar 2007- Juni 2011)	BG	eLb	nEf	Pers
Untere 2,5%	-0,5	0,2	-0,7	-0,5
Untere 10% (unteres Dezil)	-0,2	1,8	-0,1	-0,2
Untere 25% (unteres Quantil)	-0,1	2,9	0,3	-0,1
Median	0,1	4,1	0,7	0,1
Obere 25% (oberes Quantil)	0,2	5,9	1,3	0,3
Obere 10% (oberes Dezil)	0,5	8,2	2,2	0,5
Obere 2,5%	0,9	11,8	3,6	1,0

Tabelle 3: Nacherfassungsquoten in Prozent; Daten nach zweimonatiger Wartezeit

b) Nacherfassung im Zeitverlauf

Das Nacherfassungsverhalten unterliegt im zeitlichen Verlauf Schwankungen. Was bei einzelnen zugelassenen kommunalen Trägern zeitweilig auch durch Umstellungsprobleme in der Software bedingt ist. In einigen Berichtsmonaten liegen teilweise stark variierende Nacherfassungsquoten vor, so dass Zeitreihenbetrachtungen auf Basis von Daten ohne Wartezeit zu Fehlinterpretationen verleiten können.

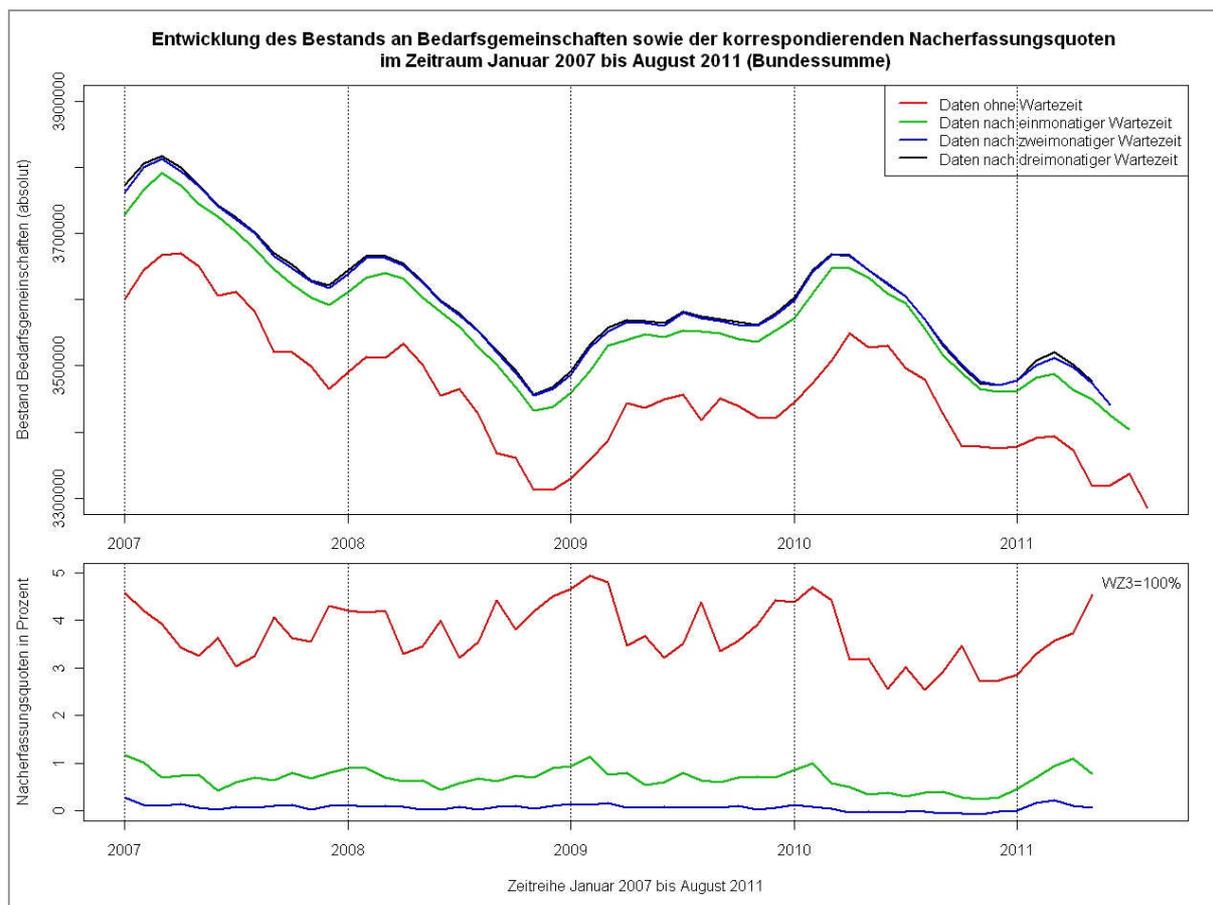


Abbildung 2: Zeitreihe zum Nacherfassungsverhalten am Bestand an BGn

Abbildung 2 verdeutlicht diese Problematik anhand der Bundesergebnisse zur Bestandsentwicklung der Bedarfsgemeinschaften im Zeitraum Januar 2007 bis August 2011. Es lassen sich vor allem bei den Nacherfassungsquoten für Daten ohne Wartezeit starke Schwankungen, die auch von einem auf den anderen Monat abrupt eintreten können. Der Entwicklungsverlauf der Bestandszahlen am aktuellen Rand kann in diesem Kontext als deutlich „eckiger“ beschrieben werden, vergleicht man ihn mit den Verläufen nach Wartezeit.

Dieser Effekt ist in Regionalbetrachtungen teilweise deutlicher ausgeprägt, da hier regionale Spezifika eine größere Rolle spielen, die sich bei einer Betrachtung auf Bundesebene wieder ausgleichen. Exemplarisch können regional unterschiedliche Urlaubszeiten oder Feiertagsregelungen, personelle Umstrukturierungen, softwareseitige Neuerungen und/ oder Korrekturen, die nur einen Teil der Träger der Grundsicherung betreffen, benannt werden.

Des Weiteren haben aber auch regionale „Rahmenbedingungen“ einen gewichtigen Einfluss auf das Nacherfassungsverhalten. Vor allem die Größe des zu betreuenden Kundenstamms spielt eine große Rolle. Diese Problematik kann an den Grundsicherungsträgern in Bayern deutlich gemacht werden. Viele dieser Träger weisen regelmäßig eine überproportionale Untererfassung auf. Die Analyse zeigt, dass sich hier überwiegend sehr kleine Träger mit kleinem Kundenstamm, kürzerer Leistungsbezugsdauer, höherem relativen Umschlag und einer relativ kleinen Personalausstattung finden. Es ist zu vermuten, dass hier Verschiebungen und besondere Ereignisse auf dem regionalen Arbeitsmarkt (z.B. Schließung eines Betriebs) sehr viel stärker zur Geltung kommen als bei großen Trägern mit großem Kunden-

stamm (z.B. JC Hamburg). Das Problem lässt sich allgemein damit umschreiben, dass kleine „Bestände“ variabler reagieren als große.

Darüber hinaus können auch überregionale Neuerungen zu kurzfristig „nicht modellierbaren“ Umbrüchen führen; zu denken ist hier vor allem an Änderungen der gesetzlichen Grundlage und an Veränderungen in den Vorgaben zur Erfassung in den operativen Verfahren.

Die Herausforderung bestand im ersten Schritt 2010 darin, unter Beachtung der regional divergenten Rahmenbedingungen, ein allgemein gültiges Set an erklärenden Größen zu bestimmen, welches für die einzelne Regionaleinheit eine hohe „Vorhersagekraft“ aufweist und zugleich robust auf Umbrüche und Ausreißer reagiert. Im Rahmen der Umsetzung des Methodenreviews zum Oktober 2011 ging es nun vor allem darum die erkannten Schwächen der Ausgangsmodellierung auszumerzen, um damit die Treffsicherheit der Schätzung zu erhöhen.

3 Konzeption des Schätzverfahrens

In diesem Kapitel zur Konzeption des Schätzverfahrens soll es darum gehen, zu erläutern wie die einzelnen Review- Bestandteile gemäß Tz. 1 (Seite 6) umgesetzt wurden und wie sie im Kontext der Gesamtkonzeption der Schätzmodellierung wirken. In diesem Zusammenhang wurde das ursprüngliche Kapitel des Methodenberichtes vom Februar 2011 hier überarbeitet und neu ausgerichtet, einige Passagen wurde übernommen, andere zugunsten neuer Inhalte ausgelassen. Ein umfassender Überblick zur initialen Entwicklung der Modellierung kann dem Methodenbericht vom Februar 2011 entnommen werden.

3.1 Modellierung – Multivariate lineare Regression

Hochrechnungen können allgemein formuliert als Projektion auf Basis eines Teilergebnisses beschrieben werden. Sie kommen in der Regel dann zur Anwendung, wenn noch nicht alle Informationen für das Gesamtergebnis vorliegen oder die Erhebung der Gesamtinformation grundsätzlich nicht zu bewältigen ist. Je umfassender dabei das verwendete Teilergebnis ausfällt (oder auch je repräsentativer das Sample) desto sicherer kann auch die Hochrechnung das tatsächliche Ergebnis bestimmen. Das für den Ausgleich untererfasster Eckwerte auf Regionalebene angewandte Modell kann in diesem Zusammenhang als Hochrechnung bezeichnet werden, die sich unter Einbezug weiterer erklärender Größen der Methodik einer linearen Regression⁶ bedient.

In die Modellierung fließt ein Set von zwölf Regressoren⁷ ein, was aber nicht bedeutet, dass alle zwölf erklärenden Größen auch stets für jeden einzelnen Kreis Anwendung für die Schätzung finden. Im Rahmen der Regression wird der „Erklärungsgehalt“ jedes Regressors für die Datenlage des einzelnen Kreises quantifiziert und es fließen nur diejenigen Regressoren ein, die unter Betrachtung der Datenlage des einzelnen Kreises, auch Erklärungsgehalt liefern. Kriterium für den „Ausschluss“ ist die Teilkorrelation⁸ zwischen dem Regressor und der abhängigen Variablen (Daten mit dreimonatiger Wartezeit). Wenn der Wert der Teilkorrelation zu klein⁹ ist, wird der entsprechende Regressor aus der Schätzung für den Kreis entfernt.

Datenbasis bildet ein Betrachtungszeitraum, welcher im Januar 2007 beginnt und jeweils bis an den aktuellen Rand heranreicht, im Oktober 2011 also 58 Monate umfasst. Kleinste Dateneinheit ist die Information zu einem bestimmten Kreis. Die Schätzmodellierung setzt auf Kreisebene an und wird entsprechend für jeden Kreis einzeln durchgeführt. Schätzergebnisse für Bundesländer, den Bund gesamt sowie die Unterscheidung in neue und alte Bundesländer basieren auf den entsprechenden Aggregationen.

⁶ Lineare Regressionen sind eine Form der Regressionsanalyse, bei der angenommen wird, dass zwischen der abhängigen und den unabhängigen Variablen ein linearer Zusammenhang besteht. Die abhängige Variable wird als Linearkombination der mit den Regressionskoeffizienten gewichteten Regressoren definiert.

⁷ Unabhängige Variablen sind mit Blick auf die Modellierung Größen, die das Modell „von außen“ beeinflussen. Diese so genannten Regressoren sind unabhängig von der zu erklärenden Variable und liefern in diesem Kontext Erklärungsgehalt für ihre Entwicklung.

⁸ Der Begriff der Korrelation beschreibt in der Statistik das Vorliegen einer wechselseitigen Beziehung zwischen zwei oder mehr Variablen. Dabei muss aber kein Kausalzusammenhang zwischen den Variablen bestehen.

⁹ Es wurde die Standardeinstellung im verwendeten Statistik-Softwareprogramm SPSS genutzt - hier ist ein Toleranzwert von 0,0001 vorgegeben.

Neu ist in diesem Kontext, dass für Hochrechnungen jetzt auch die statistischen Messergebnisse mit einem Monat bzw. zwei Monaten Wartezeit als Grundlage verwendet werden. Bisher wurden bei der Aktualisierung der vorläufig hochgerechneten Eckwerte jeweils nur die Messergebnisse ohne Wartezeit verwendet und es wurden nur aktuellere Rahmendaten und die etwas näher an den aktuellen Rand gerückten Ergebnisse zu den tatsächlichen Nachbarwilligungsfaktoren in der Vergangenheit berücksichtigt. Vorteil des jetzigen Ansatzes ist, dass mit fortgeschrittenem Wartezeitstand die statistische Ausgangslage wesentlich besser ist und eine Hochrechnung auf Basis dieser sukzessiv vervollständigten Teilergebnisse nur noch eine geringere Untererfassung auszugleichen hat; der Ansatz kann damit wesentlich treffsicherer die endgültige Ergebnislage projizieren.

3.1.1 Erklärende Variablen: Regressoren

Das Schätzmodell ermittelt den Bestand des jeweiligen Eckwerts (Bedarfsgemeinschaften, Personen und darunter erwerbsfähige und nicht erwerbsfähige Personen), der nach Ablauf der definierten Wartezeit von drei Monaten erwartet werden kann bzw. höchstwahrscheinlich eintritt, und nutzt dazu folgende Informationen.

- aktueller Bestand des Eckwerts ohne Wartezeit (BG_t^{WZ0}) bzw. mit einmonatiger (BG_t^{WZ1}) oder zweimonatiger Wartezeit (BG_t^{WZ2})
- linearer Trend der Nacherfassungsquoten basierend auf den jeweiligen Wartezeitständen, im Sinne eines langfristigen Lerneffektes im Nacherfassungsverhalten ($NQ_BG_t^{PRED}$)
- Einfluss des statistischen Stichtags im Rahmen des Einbezugs der Nettoarbeitstage als erklärende Größe (NAT_t)
- Abbildung von nicht-linearen Effekten durch den Einbezug der Quadrate aus den Nettoarbeitstagen sowie dem linearen Trend der Nacherfassungsquoten ($(NAT_t)^2$, $(NQ_BG_t^{PRED})^2$)
- Wechselwirkungen von verschiedenen Einflüssen im Kontext des Einbezugs von Interaktionsvariablen aus dem Datenstand ohne Wartezeit (bzw. mit einmonatiger oder zweimonatiger Wartezeit), den im jeweiligen Berichtsmonat maßgeblichen Nettoarbeitstagen sowie dem linearen Trend der Nacherfassungsquoten ($(NAT_t \times NQ_BG_t^{PRED})$, $(BG_t^{WZ0/1/2} \times NAT_t)$, $(BG_t^{WZ0/1/2} \times NQ_BG_t^{PRED})$, $(BG_t^{WZ0/1/2} \times NQ_BG_t^{PRED} \times NAT_t)$)
- Informationen aus der Arbeitslosenstatistik: aktueller Bestand an Arbeitslosen sowie Zugänge in und Abgänge aus Arbeitslosigkeit im Rechtskreis SGB II (ALO_t , Zug_ALO_t , Abg_ALO_t)

$$\beta_i = \frac{\text{cov}(x_i, y)}{\text{var}(x_i)}$$

$$\alpha = \bar{y} - \sum_{i=1}^n \beta_i \bar{x}_i$$

Formel 1: Berechnung Regressionskoeffizienten, allgemein

Die Berechnung der Regressionskoeffizienten (α , β_i) erfolgt im Rahmen einer einfachen Regressionsanalyse¹⁰, dessen Formulierung links abgebildet ist. Unten wurden die konkreten Formeln für die Schätzung des Wartezeitbestandes an Bedarfsgemeinschaften exemplarisch anhand der Regressionskoeffizienten β_1 und α dargelegt.

$$\beta_1 = \frac{\text{cov}(BG_t^{WZ0}, BG_t^{WZ3})}{\text{var}(BG_t^{WZ0})}$$

$$\alpha^{WZ0} = \overline{BG^{WZ3}} - \beta_1 \overline{BG^{WZ0}} - \beta_2 \overline{NQ^{WZ0_BG^{PRED}}} - \beta_3 \overline{NAT^{WZ0}} - \beta_4 \overline{ALO} - \beta_5 \overline{Zug_ALO} - \beta_6 \overline{Abg_ALO} - \beta_7 \overline{(NAT^{WZ0})^2} - \beta_8 \overline{(NQ^{WZ0_BG^{PRED}})^2} - \beta_9 \overline{(NAT^{WZ0} * NQ^{WZ0_BG^{PRED}})} - \beta_{10} \overline{(BG^{WZ0} * NAT^{WZ0})} - \beta_{11} \overline{(BG^{WZ0} * NQ^{WZ0_BG^{PRED}})} - \beta_{12} \overline{(BG^{WZ0} * NQ^{WZ0_BG^{PRED}} * NAT^{WZ0})}$$

Formel 2: Berechnung der Regressionskoeffizienten für Daten ohne Wartezeit

$$\beta_1 = \frac{\text{cov}(BG_t^{WZ1}, BG_t^{WZ3})}{\text{var}(BG_t^{WZ1})}$$

$$\alpha^{WZ1} = \overline{BG^{WZ3}} - \beta_1 \overline{BG^{WZ1}} - \beta_2 \overline{NQ^{WZ1_BG^{PRED}}} - \beta_3 \overline{NAT^{WZ1}} - \beta_4 \overline{(NAT^{WZ1})^2} - \beta_5 \overline{(NQ^{WZ1_BG^{PRED}})^2} - \beta_6 \overline{(NAT^{WZ1} * NQ^{WZ1_BG^{PRED}})} - \beta_7 \overline{(BG^{WZ1} * NAT^{WZ1})} - \beta_8 \overline{(BG^{WZ1} * NQ^{WZ1_BG^{PRED}})} - \beta_9 \overline{(BG^{WZ1} * NQ^{WZ1_BG^{PRED}} * NAT^{WZ1})}$$

Formel 3: Berechnung der Regressionskoeffizienten für Daten mit einmonatiger Wartezeit

$$\beta_1 = \frac{\text{cov}(BG_t^{WZ2}, BG_t^{WZ3})}{\text{var}(BG_t^{WZ2})}$$

$$\alpha^{WZ2} = \overline{BG^{WZ3}} - \beta_1 \overline{BG^{WZ2}} - \beta_2 \overline{NQ^{WZ2_BG^{PRED}}} - \beta_3 \overline{NAT^{WZ2}} - \beta_4 \overline{(NAT^{WZ2})^2} - \beta_5 \overline{(NQ^{WZ2_BG^{PRED}})^2} - \beta_6 \overline{(NAT^{WZ2} * NQ^{WZ2_BG^{PRED}})} - \beta_7 \overline{(BG^{WZ2} * NAT^{WZ2})} - \beta_8 \overline{(BG^{WZ2} * NQ^{WZ2_BG^{PRED}})} - \beta_9 \overline{(BG^{WZ2} * NQ^{WZ2_BG^{PRED}} * NAT^{WZ2})}$$

Formel 4: Berechnung der Regressionskoeffizienten für Daten mit zweimonatiger Wartezeit

¹⁰ Die Varianz beschreibt den Charakter der Verteilung einer Variablen. Varianz misst in diesem Zusammenhang die Streuung der Werte relativ zu einem Erwartungswert (in der Regel geht man vom Mittelwert der Variablen aus). Hierbei werden die Quadrate der Abweichungen entsprechend Ihrer Häufigkeit bzw. Wahrscheinlichkeit gewichtet. Kovarianz ist in diesem Kontext als Maßzahl für den Zusammenhang zweier Variablen zu verstehen. Sie gibt die Richtung des Zusammenhangs an, nicht aber die Stärke. Ist die Kovarianz positiv, so kann von einem gleichläufigen linearen Zusammenhang ausgegangen werden. Negative Ausprägungen der Kovarianz lassen auf gegenläufige lineare Zusammenhänge schließen. Ist die Kovarianz Null, so besteht kein linearer Zusammenhang zwischen beiden Variablen (aber: kein Ausschluss von nichtlinearen Zusammenhängen).

3.1.1.1 Aktueller Bestand ohne bzw. mit geringer Wartezeit

Der aktuelle Bestand des jeweiligen Eckwerts ohne bzw. mit einmonatiger und zweimonatiger Wartezeit ist als unabhängige Größe im Modell explizit aufgegriffen worden und kann dabei als bereits bekanntes Teilergebnis im Rahmen der Hochrechnung beschrieben werden. Im alten Verfahren wurden nur die Daten ohne Wartezeit und die Daten mit dreimonatiger Wartezeit verwendet. Die Aktualisierung der Schätzergebnisse basierte allein auf dem monatsweisen Anwachsen der Zeitreihe und nicht auf zumindest teilweise vervollständigten Wartezeitinformationen. Das hatte zur Folge, dass die Schätzung zum Wartezeitstand zwei Monate teilweise einen größeren Fehler aufwies, als die reale Datenlage ohne Hochrechnung. Abbildung 3 verdeutlicht diesen Effekt graphisch. Die oberen drei Boxplots zeigen die Verteilung der Nacherfassungsquoten für Daten ohne (WZ0), mit einmonatiger (WZ1) und zweimonatiger Wartezeit (WZ2). Die Boxen markieren dabei jeweils den Bereich in dem die mittleren 50% der Daten liegen. Der jeweilige Median (mittlerer Wert) ist als vertikaler Strich in der Box abgetragen. Die „Antennen“ markieren das jeweils das 1,5-fache des Interquartilsabstandes (der durch die Breite der Verteilung der mittleren 50% der Daten definiert ist).

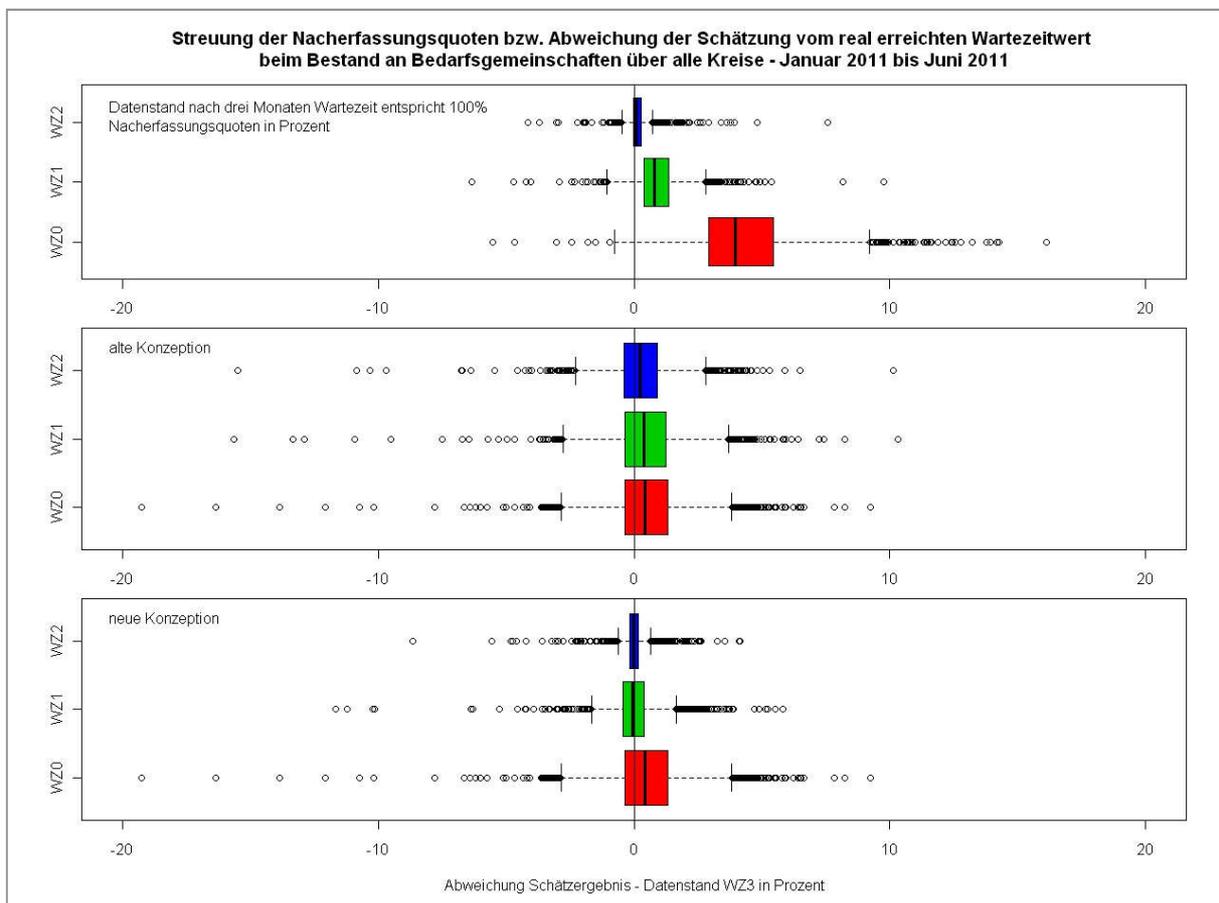


Abbildung 3: Wirkung des Einbezugs aktuellerer Wartezeitstände in der Schätzung

Die abgetragenen Punkte außerhalb dieser Begrenzungen zeigen die Ausreißer der Verteilung. Der Zugewinn an Treffsicherheit bezogen auf Daten mit Wartezeitstand mit einem und zwei Monaten wird daran deutlich, dass die Verteilungen der Nacherfassungsquoten mit Wartezeit schmaler werden und in Richtung Null-Prozent-Marke tendieren. Die unteren beiden Abbildungen zeigen die Abweichungen der Schätzergebnisse erzielt mit der alten (mittlere Graphik) und neuen Schätzkonzeption (untere Graphik) von den Wartezeitergebnissen

nach dreimonatiger Wartezeit. Zu erkennen ist deutlich, dass die Schätzung mittels alter Modellierung die Wartezeitentwicklung der Daten nur bedingt nachzeichnen konnte, während die neue Modellierung nicht nur die Untererfassung am aktuellen Rand gut ausgleicht sondern auch die Verschmälerung der Verteilungen nachzeichnet. Jeder Schätzlauf bringt einen echten Informationszugewinn.

Durch den Einbezug der Datenstände nach ein- bzw. zweimonatiger Wartezeit wird der verbleibende „zu erklärende“ Anteil der Wartezeitentwicklung geringer, so dass auch eine Hochrechnung auf Basis dieser sukzessiv vervollständigten Teilergebnisse weniger Untererfassung auszugleichen hat und damit treffsicherer die endgültige Ergebnislage projizieren kann.

3.1.1.2 Trend des Nacherfassungsverhaltens

Der lineare Trend der Nacherfassungsfaktoren bildet die Entwicklung des Nacherfassungsverhaltens im jeweiligen Kreis ab. Er wird zunächst auf den Ist-Daten ermittelt, die bis Januar 2007 zurückreichen und dann genutzt um die Entwicklung des Nacherfassungsverhaltens am aktuellen Rand abzuschätzen.

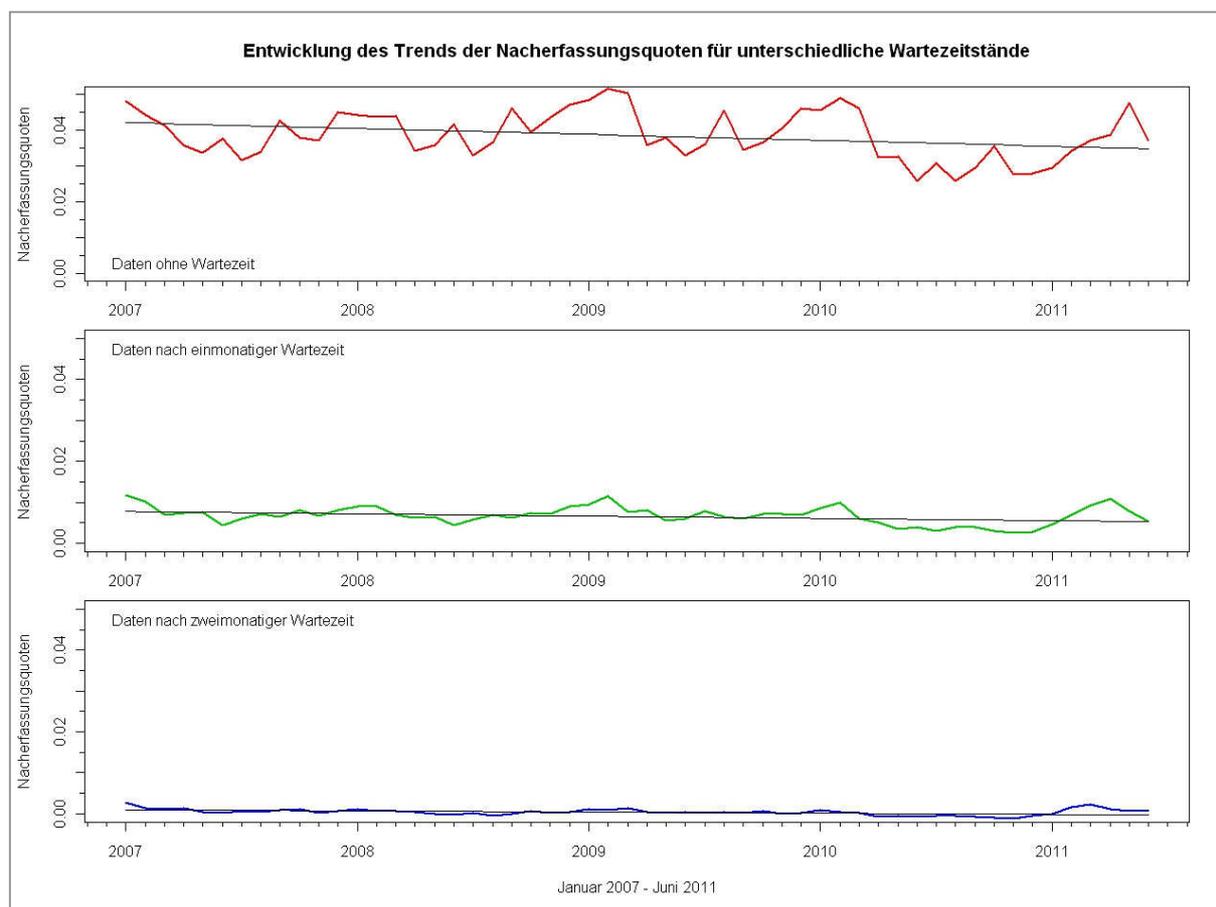


Abbildung 4: Trend der Nacherfassungsquoten nach Wartezeit

Der Trend der Nachbewilligungsfaktoren kann als langfristiger Lerneffekt im Hinblick auf das Nacherfassungsverhalten des einzelnen Kreises bzw. des zuständigen Trägers der Grundversicherung im Rahmen regionaler Rahmenbedingungen interpretiert werden.

Überlegungen den Trend über einen kürzeren Zeitraum zu schätzen um damit kurzfristig möglicherweise treffsicherere Verläufe der Trendgeraden zu erreichen müssen abgelehnt werden, denn kurzfristige, extern bedingte Schwankungen im Nacherfassungsverhalten

könnten das Schätzergebnis in die eine oder andere Richtung verzerren. Grundsätzlich ist die Unsicherheit einer Prognose bezogen auf die implementierten Einflussgrößen davon abhängig wie lang die zur Verfügung stehende Zeitreihe ist; eine längere Zeitreihe wirkt stabilisierend, soweit keine Änderungen bei den Einflussgrößen selbst stattfinden.

In der überarbeiteten Schätzmodellierung wird für jeden Schätzlauf je Wartezeitstand jeweils der entsprechende Trend der Nacherfassungsquoten angesetzt. Zum Beispiel wird für den Schätzlauf nach zweimonatiger Wartezeit der lineare Trend der Nacherfassungsquoten für Daten nach zweimonatiger Wartezeit eingesetzt. Abbildung 4 visualisiert die datenseitigen Grundlagen.

3.1.1.3 Nettoarbeitstage – der Einfluss des statistischen Stichtags

Als Nettoarbeitstage (NAT) bezeichnet man die Zahl an Tagen zwischen dem Monatsersten und dem Tag des Datenabzugs abzüglich der Wochenenden und der bundeseinheitlichen Feiertage.

Die Zahl der Nettoarbeitstage quantifiziert die Zeit, die den Sachbearbeitern verbleibt, um erst- oder wiederbewilligte Leistungsfälle für den (aktuellen) Berichtsmonat zu bearbeiten. Je nach Monat kann diese Zeitspanne kürzer oder länger ausfallen, sie schwankt für Daten ohne Wartezeit in der Regel zwischen fünf und zehn Arbeitstagen.

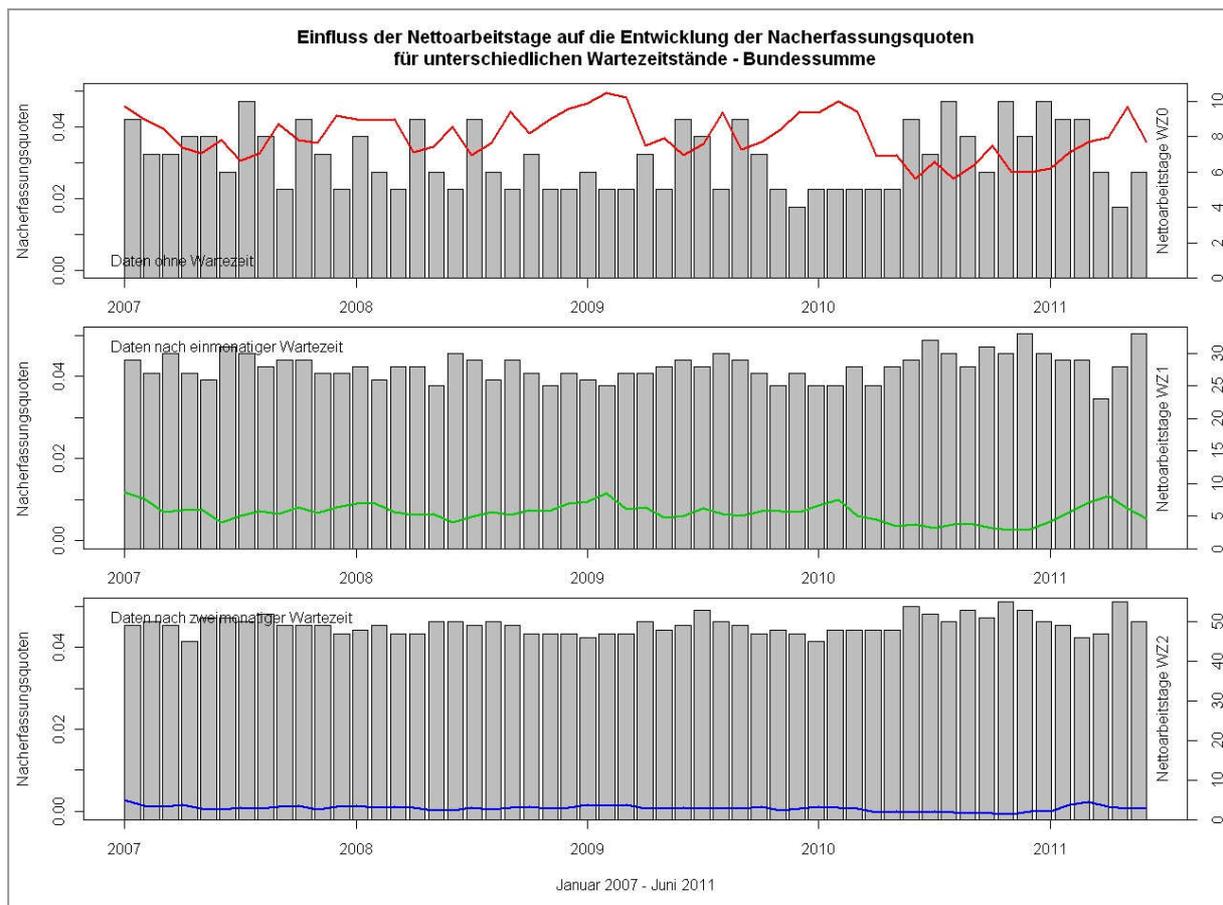


Abbildung 5: Einfluss der Nettoarbeitstage auf die Nacherfassungsquoten

Die Grundannahme ist, dass die Zahl der Nettoarbeitstage ein Maß dafür ist wie stark die Daten ohne Wartezeit untererfasst sind. Allgemein sollte gelten, je weniger Nettoarbeitstage, desto größer ist die Untererfassung bzw. je mehr Nettoarbeitstage, desto kleiner die Unterer-

fassung. Abbildung 5 visualisiert den beschriebenen Effekt anhand der Nacherfassungsquoten für den Bestand an Bedarfsgemeinschaften (Bundessumme). Man erkennt vor allem für den Datenstand ohne Wartezeit, dass Monate mit einer geringen Anzahl an Nettoarbeitstagen hohe Nacherfassungsquoten nach sich ziehen, während in Monaten mit vielen Nettoarbeitstagen eine entsprechende Verringerung der Nacherfassungsquoten zu beobachten ist. Dieser Einfluss lässt mit steigender Wartezeit nach. Beim Datenstand nach einmonatiger Wartezeit ist er noch latent abzulesen, während er bei Daten mit zweimonatiger Wartezeit kaum noch eine Rolle spielt.

Für die Umsetzung der Hochrechnung von Wartezeitständen wurde der Einsatz der Nettoarbeitstage erweitert. So fließt für die Schätzung von Daten ohne Wartezeit der originäre Nettoarbeitstageansatz ein, bei dem man vom Monatsanfang bis Datenabzug A2LL bzw. Stichtag die um Wochenenden und bundeseinheitliche Feiertage bereinigte Zahl an Arbeitstagen zählt. Für die Schätzung von Daten nach einer Wartezeit von einem Monat bezieht man die Nettoarbeitstage des gesamten ersten Monats bis hin zum Termin des Datenabzugs A2LL bzw. Stichtags des Folgemonats ein, denn das quantifiziert den Zeitraum, in dem die Träger der Grundsicherung vor Ort Informationen für den entsprechenden Berichtsmonat erfassen können. Für die Schätzung von Daten nach zweimonatiger Wartezeit gilt dieses Prinzip analog. Man „zählt“ die Nettoarbeitstage über den ersten und zweiten Monat insgesamt bis hin zum dritten Monat, in dem man den Zeitraum bis Datenabzug A2LL bzw. Stichtag zählt.

Seit Juli 2011 wird die Zahl der Nettoarbeitstage verfahrensorientiert bestimmt. Dabei orientiert sich die Zahl der Nettoarbeitstage für zugelassene kommunale Träger am statistischen Stichtag, während für Jobcenter (JC) in gemeinsamer Einrichtungen und JC in getrennter Aufgabenwahrnehmung die Zahl der Nettoarbeitstage am Datenabzug A2LL festgemacht wird. Tabelle 4 zeigt für Januar bis Juni 2011 die relevanten Termine.

Jahr	Monat	Stichtag	Datenstand A2LL	Feiertage1	Feiertage2	NAT A2LL	NAT zkT
2011	Januar	17.01.2011	14.01.2011	01.01.2011		10	11
2011	Februar	15.02.2011	11.02.2011			9	11
2011	März	15.03.2011	11.03.2011			9	11
2011	April	12.04.2011	08.04.2011	22.04.2011	25.04.2011	6	8
2011	Mai	12.05.2011	05.05.2011	01.05.2011		5	9
2011	Juni	15.06.2011	09.06.2011	02.06.2011	13.06.2011	7	9

Tabelle 4: Terminübersicht zur Berechnung der Nettoarbeitstage

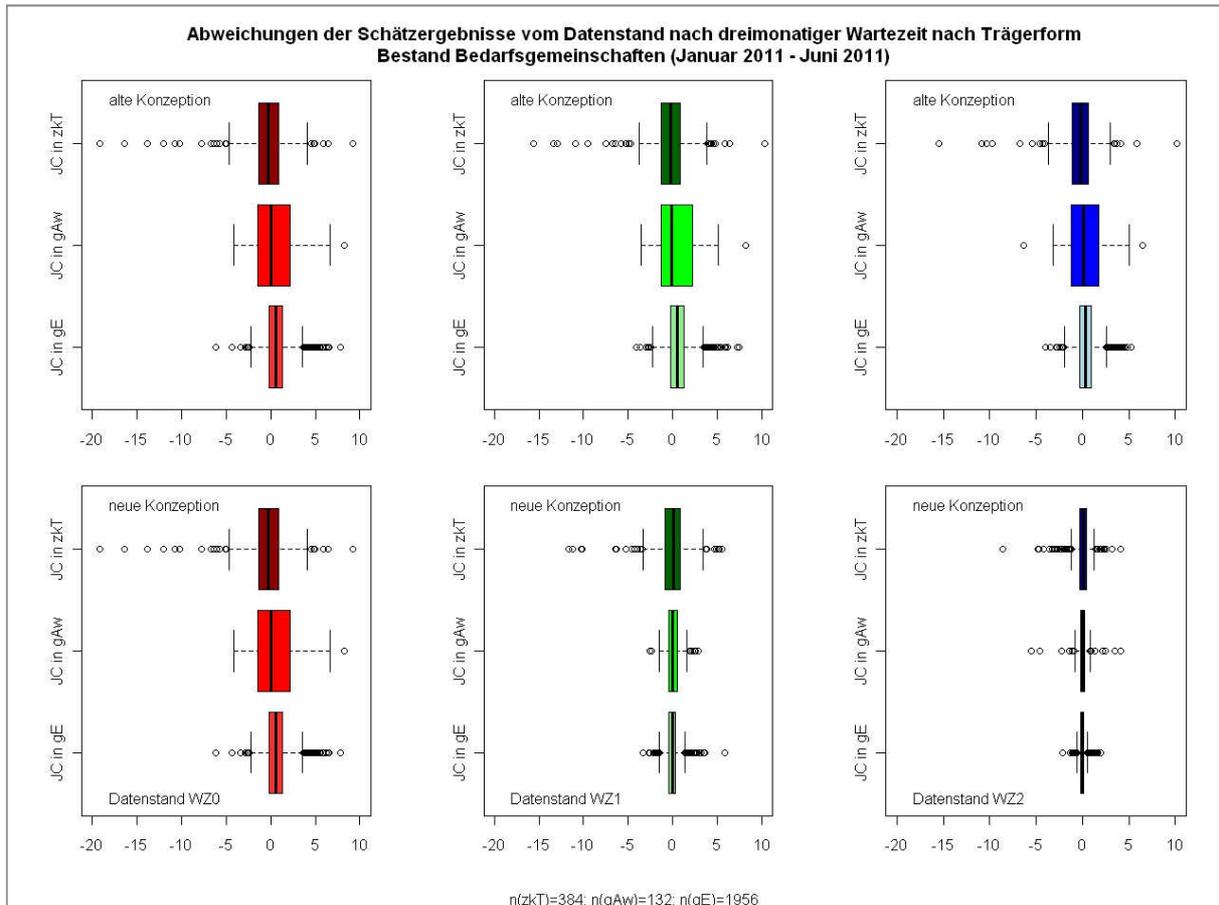


Abbildung 6: trägerformbezogene Ergebnisbetrachtung

Abbildung 6 zeigt den Effekt des trägerformbezogenen Einsatzes der Nettoarbeitstage auf das Ergebnis der Schätzung über alle Kreise und die ersten sechs Monate in 2011. Im oberen Teil der Graphik sind jeweils die Verteilungen der Abweichungen zwischen dem Wartezeitenergebnis und der alten Schätzkonzeption dargestellt, während im unteren Teil der Graphik das Äquivalent für die überarbeitete Konzeption abgebildet ist. Man erkennt zunächst den Zugewinn an Datenqualität vor allem in den Graphiken zur Schätzung nach einmonatiger und zweimonatiger Wartezeit. Für die Schätzung am aktuellen Rand ergibt sich rein optisch keine Veränderung der trägerformbezogenen Verteilungen. Die Tatsache, dass die Neukonzeption der Nettoarbeitstage zumindest graphisch keinen größeren Effekt hat, ist vor allem darauf zurück zu führen, dass im Rahmen der inhaltlich-methodischen Überarbeitung der Hochrechnung keine isolierte Betrachtung der Einzeleinflüsse möglich ist und somit eine Effekte-Überlagerung im Ergebnis deutlich wird. In diesem Kontext kann man davon ausgehen, dass die Einführung von trägerformspezifischen Nettoarbeitstagen einen eher kleinen Teil der Ergebnisverbesserung bewirkt, da auch nur ein verhältnismäßig kleiner Teil der Kreise hiervon betroffen ist. Nichtsdestotrotz macht die Graphik deutlich, dass der Informationszugewinn durch die Neukonzeption der Schätzung trägerformübergreifend deutlich ausfällt.

3.1.1.4 Arbeitslosigkeit als Frühindikator für Hilfebedürftigkeit

Die Entwicklung der Arbeitslosigkeit im Berichtsmonat wird für die Schätzmodellierung über den aktuellen Bestand, den Zugang in und Abgang aus Arbeitslosigkeit im Rechtskreis SGBII quantifiziert. Grundannahme ist, dass die Entwicklung der Arbeitslosigkeit prädiktiven Erklä-

rungsgehalt für die Entwicklung der Bestandszahlen ohne Wartezeit liefert, also als Frühindikator zu Rate gezogen werden kann. Analysen haben aber gezeigt, dass für die Schätzungen auf Basis von Daten nach einmonatiger und zweimonatiger Wartezeit die Arbeitslosigkeit keinen erwähnenswerten Beitrag mehr leistet und deshalb dort nicht mehr als Prädiktoren verwendet wird.

Grundsätzlich ist zu erwarten, dass von einem konjunkturellen Abschwung in Kombination mit steigender Arbeitslosigkeit zunächst vor allem solche Beschäftigungsverhältnisse betroffen sind, die als instabil im Sinne der Dauer und Kontinuität des Beschäftigungsverhältnisses beschrieben werden können. Ein direkter Effekt auf den Personenbestand im SGB II lässt sich vor allem aus den Annahmen folgern, dass solche Beschäftigungsverhältnisse zum einen häufig die Anwartschaftszeiten für einen Bezug von Arbeitslosengeld nicht erfüllen und zum anderen die Höhe der Bezüge sowie der daraus resultierende Entgeltersatz per Arbeitslosengeld mit Grundsicherungsleistungen aufgestockt werden muss.

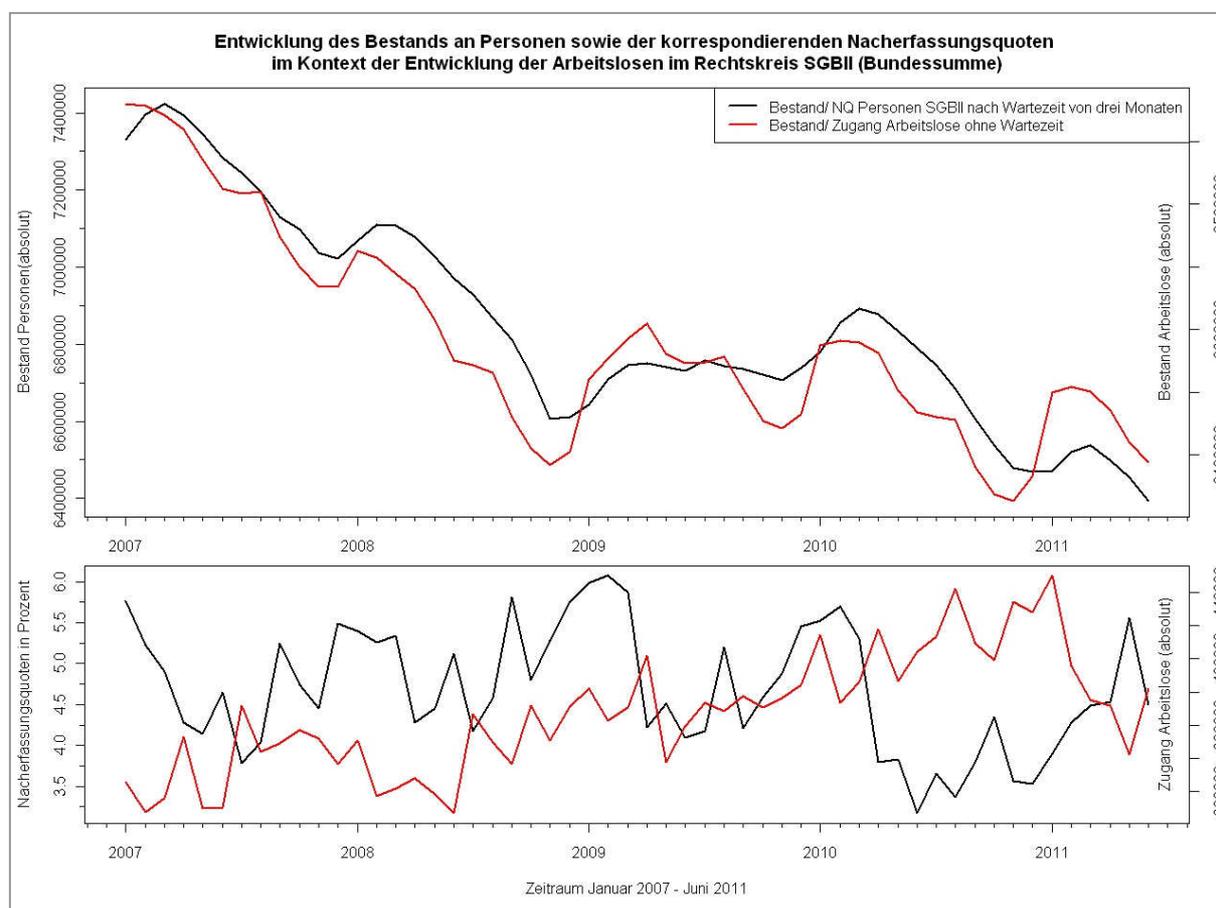


Abbildung 7: Einfluss der Regressoren der Arbeitslosenstatistik

Ein Konjunkturabschwung mit negativem Einfluss auf die Zahl der Beschäftigtenverhältnisse würde also zunächst einen Anstieg des Bestands an Personen im SGB II zur Folge haben.

Darüber hinaus ist anzunehmen, dass sich die allgemeine Situation am Arbeitsmarkt verschärft und dass Neuabschlüsse von Beschäftigungsverhältnissen unwahrscheinlicher werden. Dies bewirkt wiederum eine Verlängerung von individuellen Arbeitslosigkeitsdauern und führt zeitversetzt zu einem verstärkten Übergang von Personen aus dem Leistungsbezug SGB III in den Leistungsbezug der Grundsicherung. Mit Blick auf die Schätzmodellierung

kann man feststellen, dass der fluktuierende Teil des Personenbestands im SGB II direkt sowie indirekt von den allgemeinen Entwicklungen am Arbeitsmarkt abhängig ist und somit auch Veränderungen im Bestand an Bedarfsgemeinschaften und Personen in der Grundsicherung durch Veränderungen am Arbeitsmarkt erklärt werden können. Abbildung 7 verdeutlicht die Zusammenhänge an zwei Schaubildern. Im oberen Teil der Graphik ist für den Zeitraum Januar 2007 bis Juni 2011 die Entwicklung des Bestands an Arbeitslosen im Rechtskreis SGB II gegen die Entwicklung des Bestands an Personen im SGB II nach einer Wartezeit von drei Monaten geplottet. Die Entwicklung beider Graphen kann als im Verlauf recht ähnlich beschrieben werden (unter Berücksichtigung der unterschiedlichen Skalenniveaus). Vorteil des Einsatzes der Arbeitslosenzahlen im Kontext der Schätzmodellierung ist nun der zeitliche Vorlauf dieser Informationen der vor allem daraus resultiert, dass der Meldeprozess zur Arbeitslosigkeit dem Leistungsgewährungsprozess im SGBII vorgelagert ist.

In Zusammenschau der Informationen zum Bestand sowie der Zugänge in und Abgänge aus Arbeitslosigkeit im Rechtskreis SGB II kann ein Stock-Flow-Zusammenhang¹¹ beschrieben werden, welcher wiederum Einfluss auf das Nacherfassungsverhalten im jeweiligen Kreis hat. Die untere Graphik zeigt diesen Zusammenhang, indem die Zugangszahlen in Arbeitslosigkeit gegen die Entwicklung der Nacherfassungsquoten (für Daten ohne Wartezeit) für den Personenbestand abgebildet sind. Es lässt sich zeigen, dass Zugang von Arbeitslosen und Nacherfassungsquoten miteinander korrespondieren, wobei die Zugänge in Arbeitslosigkeit einen leicht vorlaufenden Charakter haben.

3.1.1.5 Interaktionsvariablen im Modell

Mit der Implementierung von Interaktionsvariablen in die Schätzmodellierung wird versucht, Wechselwirkungen von verschiedenen erklärenden/externen Variablen abzubilden. Interaktionseffekte sind nicht additive Effekte zweier oder mehrerer unabhängiger Variablen in einer Schätzmodellierung. Dabei wird angenommen, dass der Erklärungsgehalt einer Variablen von den Ausprägungen einer anderen Variablen abhängig ist. Interaktionseffekte werden über das Produkt zweier oder mehrerer Variablen im Modell abgebildet.

Im Einzelnen wurden folgende Interaktionsterme in der Schätzmodellierung berücksichtigt.

- Nettoarbeitstage und Trend der Nachbewilligungsfaktoren für Daten ohne, mit einmonatiger und zweimonatiger Wartezeit
- aktueller Bestand und Nettoarbeitstage für Daten ohne, mit einmonatiger und zweimonatiger Wartezeit
- aktueller Bestand und Trend der Nachbewilligungsfaktoren für Daten ohne, mit einmonatiger und zweimonatiger Wartezeit
- aktueller Bestand und Trend der Nachbewilligungsfaktoren und Nettoarbeitstage für Daten ohne, mit einmonatiger und zweimonatiger Wartezeit

Quadrate aus erklärenden Variablen werden vor allem deshalb in die Schätzmodellierung einbezogen, um nicht-lineare Effekte abbilden zu können. Lineare Zusammenhänge können die Realität nicht immer treffsicher abbilden. Grundannahme eines linearen Zusammenhangs ist, dass sich die erklärende Größe stets gleichartig auf die Entwicklung der zu erklärenden Größe auswirkt. Fügt man neben der erklärenden Variable noch das Quadrat dieser in eine

¹¹ Definition des Begriffs „Stock-Flow-Modell“ im Anhang → C) Begriffsdefinition – Stock-Flow-Modell

Schätzmodellierung ein, so kann man auch variierende Einflüsse der jeweiligen Regressoren auf die abhängige Variable im Modell abbilden und erreicht flexiblere Anpassungen der Schätzgleichung an die abzubildende Realität.

Konkret ist in die Schätzmodellierung neben dem Quadrat der Nettoarbeitstage für Daten ohne, mit einmonatiger und zweimonatiger Wartezeit auch das Quadrat der Entwicklung der Nachbewilligungsfaktoren für Daten ohne, mit einmonatiger und zweimonatiger Wartezeit eingeflossen.

3.1.2 Ermittlung der Schätzwerte

Im Rahmen des Ausgleichs von operativ untererfassten Eckwerten in der Grundsicherungsstatistik wird mittels des dargelegten Schätzverfahrens jeweils ein Monat entsprechend des zur Verfügung stehenden Wartezeitstandes geschätzt. Somit ergibt sich ein dreigeteiltes Verfahren, bei dem in Abhängigkeit vom aktuellst vorliegenden Wartezeitstand das erwartete Ergebnis nach dreimonatiger Wartezeit abgeschätzt wird. Dieser Logik folgend werden also für den aktuellsten Berichtsmonat ohne Wartezeit, für die Daten des Vormonats nach einmonatiger und des Vormonats nach zweimonatiger Wartezeit Hochrechnungen vorgenommen. Es ergeben sich exemplarisch für den Bestand an Bedarfsgemeinschaften die folgenden Schätzgleichungen, die Größe u_t beschreibt dabei die nicht zu modellierende Störgröße (Schätzgleichungen für erwerbsfähige und nicht erwerbsfähige Leistungsbezieher sind völlig analog).

$$\begin{aligned}
 BG_t^{PRED} = & \alpha + \beta_1 BG_t^{WZ0} + \beta_2 NQ_{BG_t^{PRED}}^{WZ0} + \beta_3 NAT_t^{WZ0} + \beta_4 ALO_t + \beta_5 Zug_{ALO_t} + \beta_6 Abg_{ALO_t} \\
 & + \beta_7 (NAT_t^{WZ0})^2 + \beta_8 (NQ_{BG_t^{PRED}}^{WZ0} \cdot BG_t^{PRED})^2 + \beta_9 (NAT_t^{WZ0} * NQ_{BG_t^{PRED}}^{WZ0} \cdot BG_t^{PRED}) \\
 & + \beta_{10} (BG_t^{WZ0} * NAT_t^{WZ0}) + \beta_{11} (BG_t^{WZ0} * NQ_{BG_t^{PRED}}^{WZ0}) + \beta_{12} (BG_t^{WZ0} \\
 & * NQ_{BG_t^{PRED}}^{WZ0} * NAT_t^{WZ0}) + u_t
 \end{aligned}$$

Formel 5: Schätzgleichung für Daten ohne Wartezeit

$$\begin{aligned}
 BG_t^{PRED} = & \alpha + \beta_1 BG_t^{WZ1} + \beta_2 NQ_{BG_t^{PRED}}^{WZ1} + \beta_3 NAT_t^{WZ1} + \beta_4 (NAT_t^{WZ1})^2 \\
 & + \beta_5 (NQ_{BG_t^{PRED}}^{WZ1} \cdot BG_t^{PRED})^2 + \beta_6 (NAT_t^{WZ1} * NQ_{BG_t^{PRED}}^{WZ1} \cdot BG_t^{PRED}) + \beta_7 (BG_t^{WZ1} \\
 & * NAT_t^{WZ1}) + \beta_8 (BG_t^{WZ1} * NQ_{BG_t^{PRED}}^{WZ1}) + \beta_9 (BG_t^{WZ1} * NQ_{BG_t^{PRED}}^{WZ1} \\
 & * NAT_t^{WZ1}) + u_t
 \end{aligned}$$

Formel 6: Schätzgleichung für Daten mit einmonatiger Wartezeit

$$\begin{aligned}
 BG_t^{PRED} = & \alpha + \beta_1 BG_t^{WZ2} + \beta_2 NQ_{BG_t^{PRED}}^{WZ2} + \beta_3 NAT_t^{WZ2} + \beta_4 (NAT_t^{WZ2})^2 \\
 & + \beta_5 (NQ_{BG_t^{PRED}}^{WZ2} \cdot BG_t^{PRED})^2 + \beta_6 (NAT_t^{WZ2} * NQ_{BG_t^{PRED}}^{WZ2} \cdot BG_t^{PRED}) + \beta_7 (BG_t^{WZ2} \\
 & * NAT_t^{WZ2}) + \beta_8 (BG_t^{WZ2} * NQ_{BG_t^{PRED}}^{WZ2}) + \beta_9 (BG_t^{WZ2} * NQ_{BG_t^{PRED}}^{WZ2} \\
 & * NAT_t^{WZ2}) + u_t
 \end{aligned}$$

Formel 7: Schätzgleichung für Daten mit zweimonatiger Wartezeit

Die Schätzung wird jeden Monat jeweils nach Validierung und Aufbereitung der Stichtagsdaten vorgenommen und ist so konzipiert, dass sich die der Schätzung zugrunde liegende Datenbasis mit jedem Monat vergrößert und zwar um die der aktuellen Stichtagsverarbeitung entstammenden neuen Informationen zur Grundsicherungsstatistik.

Mit Einbezug der jeweils aktuellsten Daten ohne, mit einmonatiger, mit zweimonatiger und mit dreimonatiger Wartezeit erhöht sich also der dem Schätzverfahren zugrundeliegende Informationsstand. Hieraus ergibt sich mit jedem Schätzlauf eine Revision der Schätzergebnisse, die auf einer sukzessiv vervollständigten Wartezeitinformation beruht.

3.1.2.1 Ermittlung des Schätzwertes zum Bestand Personen

Der Eckwert Bestand an Personen im SGB II (insgesamt) wird als Summe aus den beiden geschätzten Eckwerten erwerbsfähige (eLb) und nicht erwerbsfähige Leistungsberechtigte (nEf) gebildet.

Für die Summenbildung werden die nicht gerundeten Schätzwerte verwendet. Dies kann bei Verwendung von gerundeten veröffentlichten Werten bei dem einzelnen Kreis zu einer Abweichung von maximal einer Person in der Summe führen. Für die Aufsummierung auf Bundesländer bzw. Bundesebene hat das keinen gravierenden Einfluss, weil es bei dieser Fehlertoleranz von einer Person pro Berichtsebene bleibt. Die Wahrscheinlichkeit, dass sich der Rundungsfehler sogar verkleinert steigt mit jeder Aggregationsstufe.

3.1.2.2 Ermittlung der Länderergebnisse

Die Ermittlung der Länderergebnisse, der Aggregationsstufen „Ost“ und „West“ sowie der Bundesergebnisse basiert auf additiver Aggregation. Grundsätzlich wäre es auch hier möglich, die Schätzmodellierung auf die jeweilige Aggregationsstufe anzuwenden. Schätzungen auf Basis der Aggregationsstufen lassen zwar grundsätzlich eine höhere Schätzgüte im Rahmen der angewandten Modellierung erwarten, weil die Schätzung auf stärker geglätteten Daten beruht und extreme Abweichungen damit weniger ins Gewicht fallen. Dies konnte aber anhand der vorliegenden Datenbasis nicht bestätigt werden, weshalb entschieden wurde, die Aggregationsstufen mittels Addition der nicht gerundeten Kreisergebnisse zu ermitteln. Neben einer etwas höheren Schätzgüte hat dieses Vorgehen den großen Vorteil, dass auf Ausgleichsrechnungen – zur Herstellung der Kongruenz der Schätzergebnisse – nach der Schätzung verzichtet werden kann.

3.1.2.3 Ermittlung der Ergebnisse für SGBII und SGBIII-Trägergebiete

Die SGBII- und die SGBIII-Gebietsstruktur sind nicht identisch mit der politischen Gebietsstruktur. Ein Beispiel hierfür kann man sich Berlin als Kreis vorstellen, hier agieren mehrere Jobcenter für die im Zweifelsfall nur ein Gesamtergebnis berichtet werden könnte. Daher müssen die Hochrechnungsergebnisse für Träger- und Agenturebene auf Kreisebene umgerechnet werden.

Die träger- und agenturbezogene Bereitstellung der Hochrechnungsergebnisse beruht auf einem Anteilswertverfahren bei dem die letzt-verfügbaren Wartezeitergebnisse (nach dreimonatiger Wartezeit) für Trägergebiet und Kreise gegeneinander gestellt werden. Dabei wird der jeweilige Anteilswert des Trägergebietes am Kreiswert ermittelt. Dieser trägerbezogene Anteilswert wird in einem zweiten Schritt mit dem Kreisergebnis aus der aktuellen Hochrechnung multipliziert und man erhält so das trägergebietsspezifische Ergebnis.

Vorteil dieses Vorgehens ist, dass keine separaten Hochrechnungen für die jeweiligen Trägerstrukturen angewandt werden müssen und sich so auch keine Differenzen in den Gesamtsummen ergeben. Darüber hinaus ist dieses Verfahren sehr schnell und effektiv einsetzbar und aufgrund der gesicherten Ergebnisqualität auf Kreisebene kann von einer ebenso hohen Schätzgüte für die Trägergebiete ausgegangen werden.

3.1.2.4 Rundung

Im Kontext der Schätzung wird auf den Einsatz gerundeter Zwischenergebnisse verzichtet. Für die Bereitstellung der Prognose-Ergebnisse ist allerdings eine Darstellung ohne Dezimalstellen vorgesehen. Bei der Weiterverarbeitung der Ergebnisse kann es zu rundungsbedingten Abweichungen kommen, dies muss entsprechend beachtet werden.

3.1.3 Standardfehler der Schätzung

Der Standardfehler einer Schätzung kann als die durchschnittliche Abweichung des geschätzten Wertes von dem real erreichten oder allgemein „wahren“ Wert¹² beschrieben werden. Je kleiner dieser Standardfehler ist, desto genauer kann der unbekannte Wert mit Hilfe der eingesetzten Schätzmodellierung prognostiziert werden. Er liefert damit ein Maß für die Güte der Schätzung.

Die untere Formel zeigt die Berechnung des Standardfehlers der Schätzung für den Schätzlauf Oktober 2011 allgemein formuliert auf.

$$\text{Standardfehler der Schätzung}_t = \sqrt{\frac{\sum_{t=\text{Januar 2007}}^{\text{Oktober 2011}} (BG_t^{PRED} - BG_t^{WZ3})^2}{\text{Anzahl Monate} - \text{Anzahl Regressoren}}}$$

Formel 8: Berechnung des Standardfehlers der Schätzung für Schätzlauf Oktober 2011

Zu beachten ist, dass der Standardfehler der Schätzung von zwei Größen abhängt. Das ist zum einen der Umfang des der Regression zugrunde liegenden Datensets – je größer dieses Datenset ist, desto kleiner wird tendenziell der Standardfehler ausfallen. Zum anderen hängt er auch von der Varianz des Datensatzes ab – je kleiner die Varianz umso kleiner wird in der Regel der Standardfehler der Schätzung ausfallen.

	BG_t^{WZ0}	$BG_t^{WZ0_PRED}$	BG_t^{WZ1}	$BG_t^{WZ1_PRED}$	BG_t^{WZ2}	$BG_t^{WZ2_PRED}$	BG_t^{WZ3}
Juni 2011	6276	6471	6457	6425	6455	6444	6455
Juli 2011	6379	6456	6502	6486	6510	6498	6498
August 2011	6261	6392	6417	6408	6438	6427	
September 2011	6179	6352	6373	6371			
Oktober 2011	6119	6380					
Standardfehler der Schätzung	27,2726		13,5520		7,4965		
Prozentuale Abweichung	0,45%		0,21%		0,12%		

Tabelle 5: Ergebnisübersicht der Schätzung Oktober 2011 für einen Beispielkreis

¹² In der Grundsicherungsstatistik wird der Datenstand nach einer Wartezeit von drei Monaten als „wahrer“ Wert angesehen.

In der Berichterstattung zu den aktuellen Ergebnissen der Hochrechnung wird weiterhin nur ein Standardfehler veröffentlicht und zwar der zu den Ergebnissen der Hochrechnung der Daten am aktuellen Rand. Dieser Standardfehler stellt im Kanon der dreiteiligen Berechnung von Schätzergebnissen den größten Fehler dar und kann ebenso im Rahmen einer Abschätzung der Schätzgüte für Daten mit einmonatiger und zweimonatiger Wartezeit herangezogen werden, auch wenn wie im Beispiel in Tabelle 5 eine deutliche Verringerung für die Schätzung anhand der Ergebnisse nach einmonatiger und zweimonatiger Wartezeit beobachtet werden kann.

4 Ergebnislage

Einige Ergebnisse wurden schon im Rahmen der Erläuterungen der vorherigen Kapitel insbesondere zur Konzeption der Schätzmodellierung vorgestellt. Um einen zusammenfassenden und damit abgerundeten Blick auf die Ergebnislage im Vergleich zur alten Konzeption anzubieten wird hier nochmals die Ergebnisentwicklung auf Kreisebene sowie auf Bundesebene dargestellt.

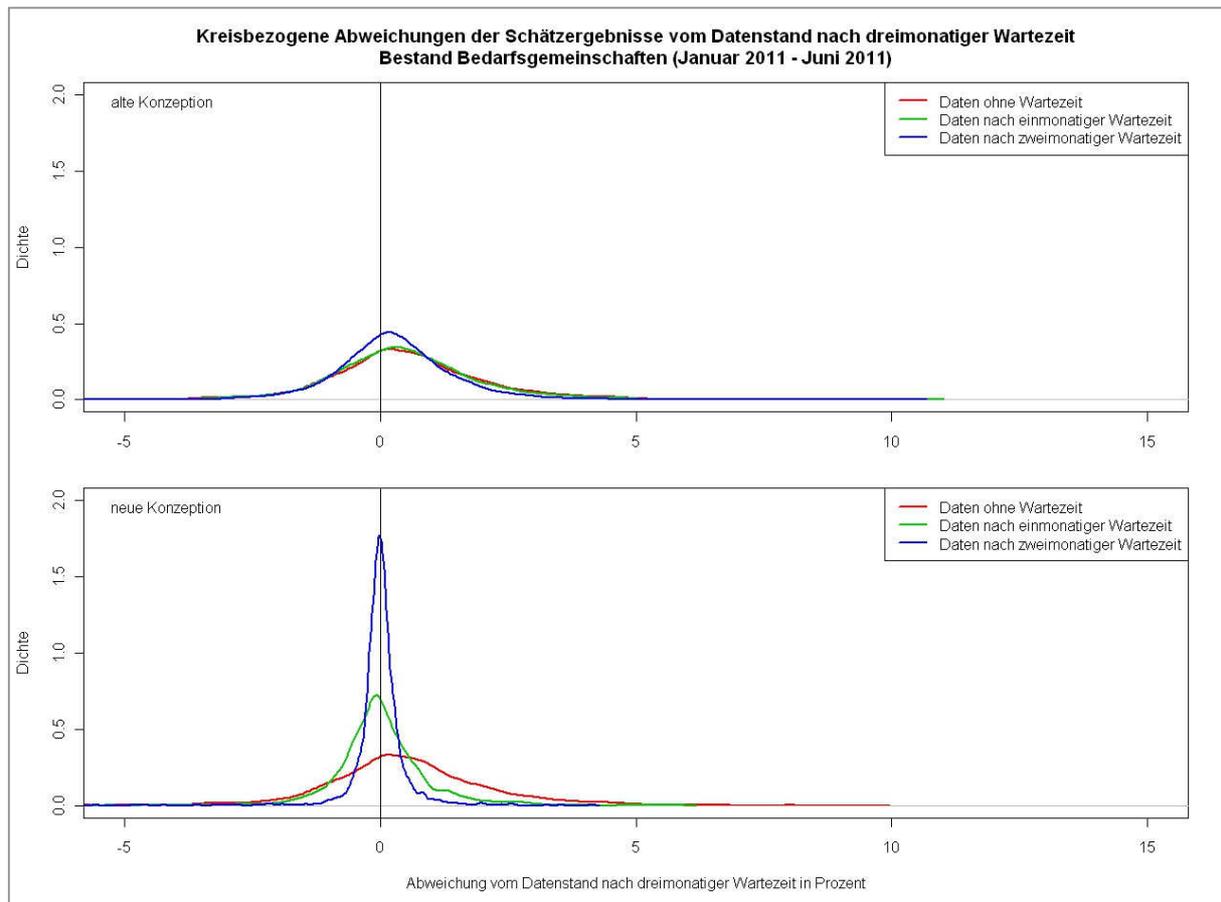


Abbildung 8: kreisbezogene Abweichungen der Schätzergebnisse vom Datenstand WZ3

Abbildung 8 zeigt für den Bestand an Bedarfsgemeinschaften die Abweichungen der Schätzergebnisse vom real erreichten Wartezeitergebnis nach drei Monaten. Dabei sind mehrere Effekte abzulesen. Zunächst erkennt man, dass die alte Konzeption im Ablauf der Wartezeit keinen echten Zugewinn an Information und damit Treffsicherheit verbuchen konnte. Alle drei Kurven sind in etwa gleich „hoch“ und gleich breit gestreut. Hier wird nochmals der Effekt unterstrichen, dass der Zugewinn an Information in der alten Konzeption allein dadurch begründet war, dass sich der zugrunde liegende Datensatz verlängerte. Ein echter Informationszugewinn – begründet durch eine sukzessiv vervollständigte Datenbasis – zeigt sich an den Schätzergebnissen auf Basis der neuen Konzeption. Man erkennt dass die Treffsicherheit unter Einbezug aktuellerer Wartezeitstände erhöht. Die Abweichungskurven für die Schätzung von Wartezeitergebnissen werden schmäler und steiler, was im Kontext der gewählten Dichtekurvendarstellung bedeutet, dass mehr Beobachtungen rund um den Erwartungswert (das Maximum der Kurven) von 0% Abweichung zwischen Schätzung und echtem Wert verteilt sind.

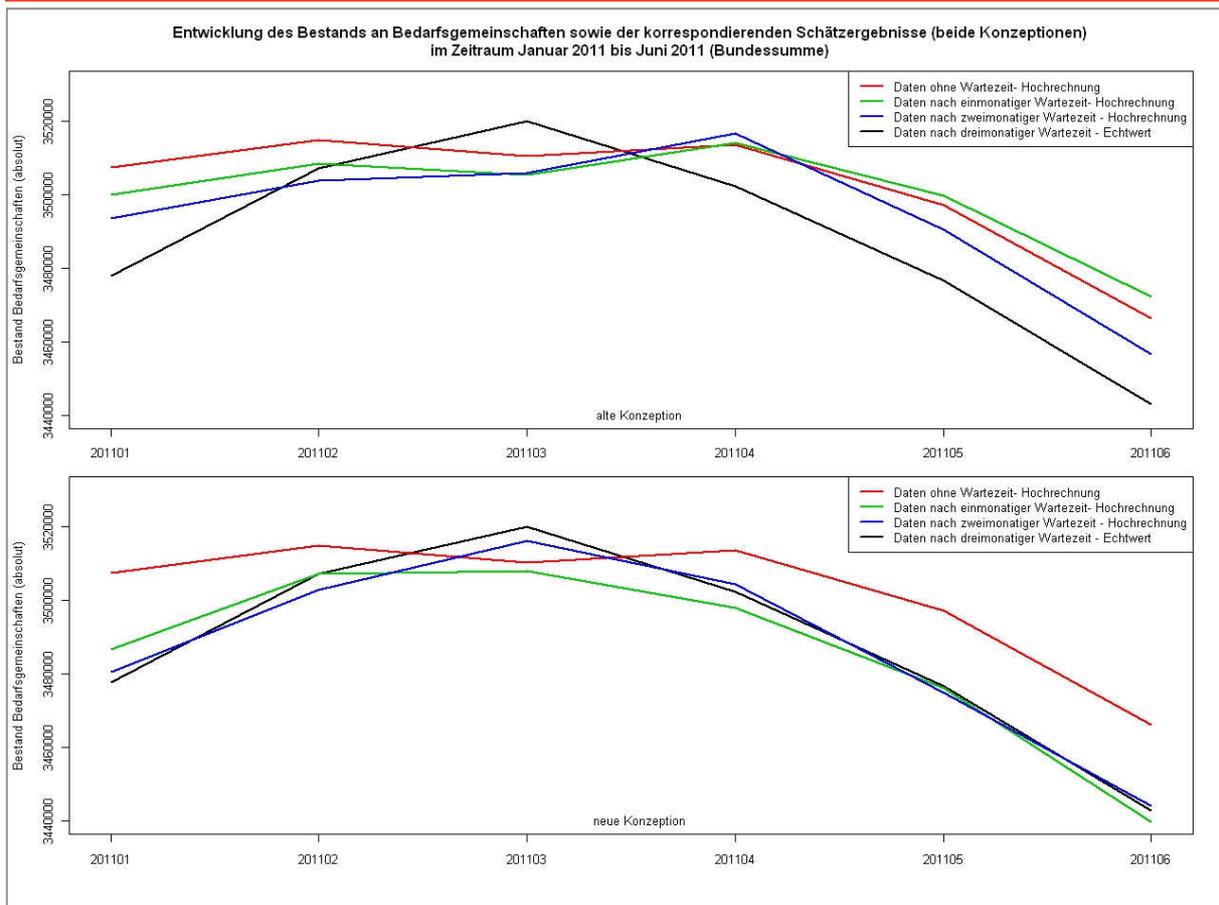


Abbildung 9: Entwicklung der Schätzergebnisse auf Bundesebene – absolut

Abbildung 9 zeigt die Ergebnisse der Schätzung auf Bundesebene in absoluten Werten. Auch hier wurden alte und neue Konzeption vergleichend untereinander dargestellt. Man erkennt für die neue Konzeption ein schrittweises Angleichen der Schätzkurven an den Verlauf der Echtwerte erzielt nach dreimonatiger Wartezeit, der im Kontext der „alten“ Konzeption deutlich schwächer ist. Auch wird die Verbesserung der Treffsicherheit der Schätzkonzeption deutlich. Es sollte aber beachtet werden, dass auf Bundesebene auch für die Schätzung am aktuellen Rand das Auseinanderlaufen von Schätzung und „Echtwert“ noch unter einem Prozent liegt und dass somit auch das „sichtbare“ Auseinanderdriften der Ergebnisse in der Datenlage eher gering ausfällt.

Tabelle 6 und Tabelle 7 zeigen die Ergebnislage kreis- und monatsübergreifend in Zahlen ausgedrückt, wobei auch hier die alte Konzeption mit der neuen direkt vergleichbar dargestellt wurde.

Ausgleich untererfasster Eckwerte auf Regionalebene

Abweichung Hochrechnung- WZ3 in Prozent Alte Konzeption / Januar 2011- Juni 2011 (Bestand BGn, alle Kreise)	Datenstand WZ0	Datenstand WZ1	Datenstand WZ2
Untere 2,5%	-2,58	-2,35	-2,10
Untere 10% (unteres Dezil)	-1,17	-1,13	-1,03
Untere 25% (unteres Quantil)	-0,36	-0,37	-0,39
Median	0,43	0,39	0,23
Obere 25% (oberes Quantil)	1,32	1,25	0,89
Obere 10% (oberes Dezil)	2,43	2,26	1,68
Obere 2,5%	3,99	3,79	2,82

Tabelle 6: Ergebnislage alte Konzeption

Abweichung Hochrechnung- WZ3 in Prozent Neue Konzeption / Januar 2011- Juni 2011 (Bestand BGn, alle Kreise)	Datenstand WZ0	Datenstand WZ1	Datenstand WZ2
Untere 2,5%	-2,58	-1,71	-0,81
Untere 10% (unteres Dezil)	-1,17	-0,89	-0,36
Untere 25% (unteres Quantil)	-0,36	-0,43	-0,16
Median	0,43	-0,05	-0,01
Obere 25% (oberes Quantil)	1,32	0,40	0,16
Obere 10% (oberes Dezil)	2,43	1,01	0,39
Obere 2,5%	3,99	2,20	0,93

Tabelle 7: Ergebnislage neue Konzeption

5 Datenqualität und monatliche Qualitätssicherung

In der Grundsicherungsstatistik werden die Daten aktuell auf Trägerebene mittels eines Fortschreibungsmodells auf Plausibilität geprüft. Das Verfahren stützt sich auf eine Typisierung der Träger der Grundsicherung nach geographischen Merkmalen und Bedingungen regionaler Arbeitsmärkte durch das Institut für Arbeitsmarkt- und Berufsforschung¹³. Auf Basis aller Träger einer Vergleichsgruppe wird für jeden Träger monatlich ein gewichteter Fortschreibungsfaktor ermittelt. Der Fortschreibungswert kann in diesem Zusammenhang als eine Prognose basierend auf der mittleren Entwicklung der Trägertypengruppe bezeichnet werden. Der Vergleich mit dem real beobachteten Monatswert bietet die Möglichkeit anhand von Konfidenzintervallen Probleme in Bezug auf die Validität des Datenmaterials zu identifizieren.

In der Grundsicherungsstatistik auf Basis des BA-Fachverfahrens A2LL wird neben dem Vergleich des per Fortschreibung erreichten Wertes und dem real beobachteten Ergebnis auch die Konsistenz des Stock-Flow-Zusammenhangs geprüft.

Analog wird mit der Datenbasis aus dem Datenübermittlungsstandard XSozial verfahren. Grundlage des Einsatzes des Fortschreibungsmodells ist die validierte Datenbasis der Jobcenter in gemeinsamer Einrichtung und getrennten Aufgabenwahrnehmung, welche als Referenzgröße im Rahmen der Betrachtung der Trägertypengruppen eingesetzt wird. Werden die Daten eines Trägers als unplausibel eingestuft bzw. führen technische oder organisatorische Probleme vor Ort zu unvollständigen Datenlieferungen oder einem kompletten Ausfall der Datenlieferung¹⁴, so werden für die Eckwerte (Bestand der Bedarfsgemeinschaften, der erwerbsfähigen Hilfebedürftigen, der nicht erwerbsfähigen Hilfebedürftigen sowie der Hilfebedürftigen insgesamt) Fortschreibungswerte auf Träger- und Kreisebene bestimmt. Mit diesem Verfahren können auch im Fall von Datenausfällen lückenfreie und konsistente Zeitreihen erreicht werden.

Im Zeitraum von Oktober 2009 bis einschließlich Oktober 2011 wurden für die Daten aller zugelassenen kommunalen Träger ohne und mit einmonatiger Wartezeit jeweils 21-mal der Fortschreibungswert benutzt, für Daten nach zweimonatiger Wartezeit kam der Fortschreibungswert in 18 Fällen zum Einsatz und für „endgültige“ Daten nach einer Wartezeit von drei Monaten 10-mal. Wobei das Gros der Ausfälle durch Umstrukturierungen der Gebietseinheiten und Trägerformen begründet war. Somit konnten im angegebenen Zeitraum etwa 98,8% der gemeldeten Daten für das Hochrechnungsmodell verwendet werden.

¹³ Rüb; Werner (2007): Typisierung von SGB II-Trägern; IAB Forschungsbericht Nr. 1/2007; Nürnberg

¹⁴ Es werden zurzeit nur ausgefallene Daten ohne Wartezeit durch Fortschreibungswerte ersetzt. Bei Ausfällen von Daten mit einer dreimonatigen Wartezeit wird auf die Daten mit zweimonatiger Wartezeit zurückgegriffen. Die Daten mit zweimonatiger Wartezeit sind in der Regel nur marginal von Untererfassungen betroffen und können daher als plausibel eingestuft werden.

6 Angebot Datenmaterial

Im Rahmen der Veröffentlichung der Grundsicherungsstatistik werden monatlich kreisbezogen für die Bestandseckwerte zu Bedarfsgemeinschaften, erwerbsfähigen Hilfebedürftigen, nicht erwerbsfähigen Hilfebedürftigen sowie den Personen insgesamt drei vorläufig hochgerechnete Wartezeitergebnisse bereitgestellt. Dies entspricht im Rahmen der Wartezeitsystematik der Grundsicherungsstatistik dem Dreimonatszeitraum am aktuellen Rand, über den vor Ablauf der definierten Wartezeit noch keine gesicherten Aussagen zur Bestandsentwicklung getroffen werden können.

Die Schätzung wird nur für die Bestandseckwerte vorgenommen. Eine Schätzung auf Basis von Strukturwerten (beispielsweise untergliedert nach Alter, Staatsangehörigkeit und Familienstand) ist bislang nicht möglich, da Strukturwerte grundsätzlich stärkeren Schwankungen unterliegen als Bestandseckwerte. Dadurch fallen die Ergebnisse je nach Gliederung in einen kleinteiligen Wertebereich, indem eine Wartezeitabschätzung nicht mehr verlässlich funktioniert.

Die Schätzergebnisse werden auf Kreisebene berechnet und bereit gestellt. Die entsprechenden Aggregatebenen (Landes- und Bundesergebnisse) basieren auf diesen Ergebnissen. Im Kontext der Schätzung wird aufgrund dessen auf den Einsatz gerundeter Zwischenergebnisse verzichtet. Für die Bereitstellung der Hochrechnungsergebnisse ist allerdings eine Darstellung ohne Dezimalstellen, also ganzzahliger Ergebnisse, vorgesehen. In diesem Kontext kann es bei der Weiterverarbeitung der Ergebnisse zu rundungsbedingten Abweichungen kommen.

Die Schätzergebnisse unterliegen einer monatlichen Aktualisierung, was darauf zurückzuführen ist, dass der den Regressionsanalysen zugrundeliegende Datensatz monatlich eine Erweiterung um die aktuellsten Stichtagsergebnisse erfährt und damit aktualisierte Ergebnisse zur Folge hat.

In Ergänzung zu den Schätzergebnissen findet eine Veröffentlichung des Schätzfehlers für die Schätzung der Daten ohne Wartezeit statt. Der Standardfehler einer Schätzung kann als die durchschnittliche Abweichung des geschätzten Wertes von dem real erreichten oder allgemein „wahren“ Wert¹⁵ beschrieben werden. Die Veröffentlichung dieses Ergebnisses soll vor allem der Quantifizierung der mittleren Güte der veröffentlichten Ergebnisse dienen. Je kleiner dieser Standardfehler ausfällt, desto genauer konnte der unbekanntes Wert mit Hilfe der eingesetzten Schätzmodellierung im Mittel prognostiziert werden.

Die Schätzergebnisse werden monatlich zum statistischen Veröffentlichungstermin bereit gestellt.

¹⁵ In der Grundsicherungsstatistik wird der Datenstand nach einer Wartezeit von drei Monaten als „wahrer“ Wert angesehen.

7 Ansprechpartner

Folgende Ansprechpartner stehen Ihnen mit Rat und Tat zur Seite sollten Sie Fragen zur Schätzmodellierung zum Ausgleich operativ untererfasster Eckwerte auf Regionalebene in der Leistungsstatistik SGB II haben.

Name	Telefon	E-Mail
Hofmann, Dr. Bernd	0911/ 179-1175	Bernd.Hofmann5@arbeitsagentur.de
Lorenz, Sebastian	0911/ 179-8728	Sebastian.Lorenz@arbeitsagentur.de
Tentscher, Christiane	0911/ 179-6476	Christiane.Tentscher@arbeitsagentur.de
Wolff, Matthias	0911/ 179-6519	Matthias.Wolff@arbeitsagentur.de

Anhang

A) Begriffsdefinition – Stichtag

Der Begriff Stichtag kann allgemein definiert als ein Datum beschrieben werden, mit dem der Eintritt eines bestimmten Ereignisses direkt verknüpft ist. In der Statistik ist der Begriff des Stichtages häufig synonym zum Begriff des „Zähltag“ in Gebrauch, was vor allem darauf zurückzuführen ist, dass am statistischen Stichtag die Quantifizierung von Beständen vorgenommen wird. Darüber hinaus ermöglicht die Terminierung von Stichtagen aber auch ereignisbezogene Betrachtungen zwischen zwei Stichtagen, also definierter Zeitintervalle.

Die für die Arbeitsmarkt- und Grundsicherungsstatistik relevanten statistischen Stichtage werden von der Statistik der Bundesagentur für Arbeit jeweils für ca. zwei Jahre im Voraus festgelegt und können unter folgendem Link eingesehen werden.

<http://statistik.arbeitsagentur.de/Navigation/Statistik/Service/Veroeffentlichungskalender/Veroeffentlichungskalender-Nav.html>

B) Begriffsdefinition – Nacherfassungsquote

Der Begriff der Nacherfassungsquote (NQ) ist ein Maß für die Unterfassung von Daten ohne Wartezeit (WZ0) bzw. mit geringer Wartezeit (WZ1 und WZ2) im Vergleich zu der nach dreimonatiger Wartezeit vervollständigten Datenbasis (WZ3) über einen Berichtsmonat. Die Nacherfassungsquote ist über folgende Formel – exemplarisch anhand des Bestands an Bedarfsgemeinschaften im Berichtsmonat dargestellt – definiert.

$$NQ_{-BG_t^{WZ3}} = \left(\frac{BG_t^{WZ3} - BG_t^{WZ0}}{BG_t^{WZ3}} \right) \times 100$$

Formel 9: Berechnung der Nacherfassungsquoten

C) Begriffsdefinition – Stock-Flow-Modell

Das „Stock-Flow-Modell“ beschreibt allgemein den Zusammenhang zwischen Stromgrößen und Bestand. Von Bestandseinheiten spricht man wenn der Zustand betrachtet werden soll. Von Ereigniseinheiten wenn ein Zustands- oder Statuswechsel gemeint ist. Eine Bestandsgröße wird dadurch gewonnen, dass man alle Einheiten, die zu einem bestimmten Zeitpunkt in einem bestimmten Zustand (z.B. Arbeitslosigkeit) existieren, zusammenfasst. Ereigniseinheiten erhält man, wenn man alle Ereignisse (z.B. Zugang in oder Abgang aus Arbeitslosigkeit) in einem bestimmten Zeitraum aufsummiert. Bei Ereignisgesamtheiten werden Zugangsgesamtheiten (Zugänge) und Abgangsgesamtheiten (Abgänge) unterschieden.

In Kapitel 3.1.1.4 wird der Stock-Flow-Zusammenhang mittels Bestand an Arbeitslosen sowie dem Zugang in und dem Abgang aus Arbeitslosigkeit abgebildet. Der Bestand an Arbeitslosen wird im Berichtsmonat jeweils am Stichtag abgezogen, während die Bewegungen als Ereignisgesamtheit im Zeitraum zwischen zwei Stichtagen gemessen werden. In diesem Kontext gilt, dass sich der Bestand eines Monats aus dem Bestand des Vormonats plus Zugänge im aktuellen Berichtsmonat minus Abgänge im aktuellen Berichtsmonat ergibt.

D) Anwendung und Interpretation von Boxplots

In einem Boxplot werden fünf verschiedene Streuungs- und Lagemaße in einer Graphik zusammengefasst, so kann man die Verteilung statistischer Daten kompakt visualisieren.

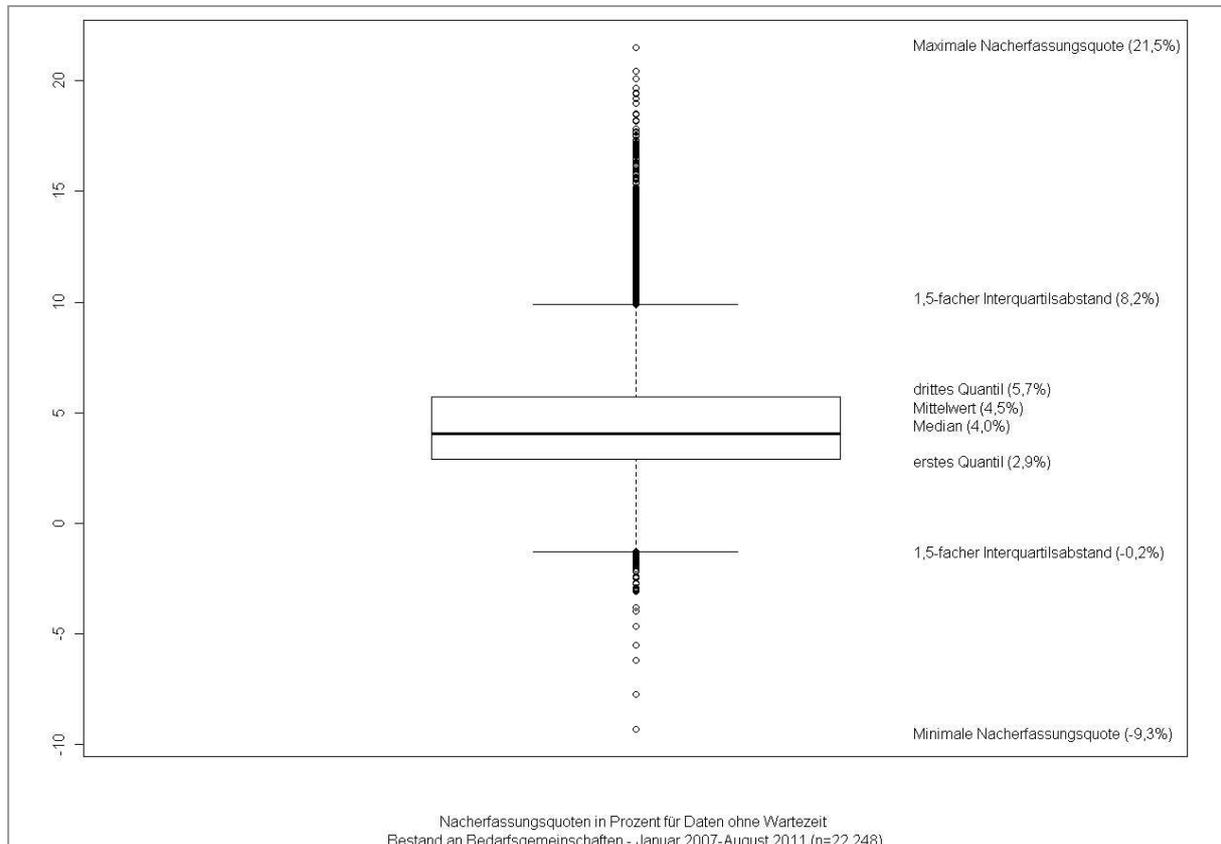


Abbildung 10: Lesehilfe Boxplot

Ein Boxplot stellt neben dem Median, den zwei Quartilen auch die Extremwerte der Verteilung dar. Die Box markiert den Bereich, in dem die mittleren 50% der Daten liegen, sie wird durch das obere und untere Quartil begrenzt. Die Länge der Box entspricht in diesem Kontext dem Interquartilsabstand und ist ein Maß für die Streuung der Daten. Der Median wird als durchgehender Strich in der Box abgetragen. Der Median teilt den Datensatz in zwei Hälften, in denen jeweils 50% der Daten liegen. Seine Lage in der Box vermittelt einen Eindruck über die Schiefe der Verteilung der Daten. Liegt der Median genau in der Mitte des Boxplot, so kann man von einer symmetrischen Verteilung der Daten des Datensatzes ausgehen. Die „Antennen“ am Boxplot stellen die außerhalb der Box liegenden Werte dar, sie zeigen das Minimum und Maximum einer Verteilung, sofern diese nicht um mehr als das 1,5-fache des Interquartilsabstand vom oberen und unteren Quartil abweichen. Datenpunkte, die darüber hinaus gehen gelten als Ausreißer und werden als einzelne Datenpunkte im Diagramm separat abgetragen.

In [Abbildung 10](#) wurde für die Nacherfassungsquoten des Bestands an Bedarfsgemeinschaften (Daten ohne Wartezeit) eine entsprechende Boxplot-Darstellung gewählt. Die mittleren 50% der Daten liegen hier im Intervall zwischen 2,9% und 5,7% - die äußeren Begrenzungen der Box markieren diesen Bereich. Der Interquartilsabstand beträgt damit im Beispiel 2,8 Prozentpunkte. Minimum und Maximum der beobachteten Nacherfassungsquoten werden in der Darstellung als Ausreißer gewertet und dargestellt, da sie um mehr als das 1,5-fache des

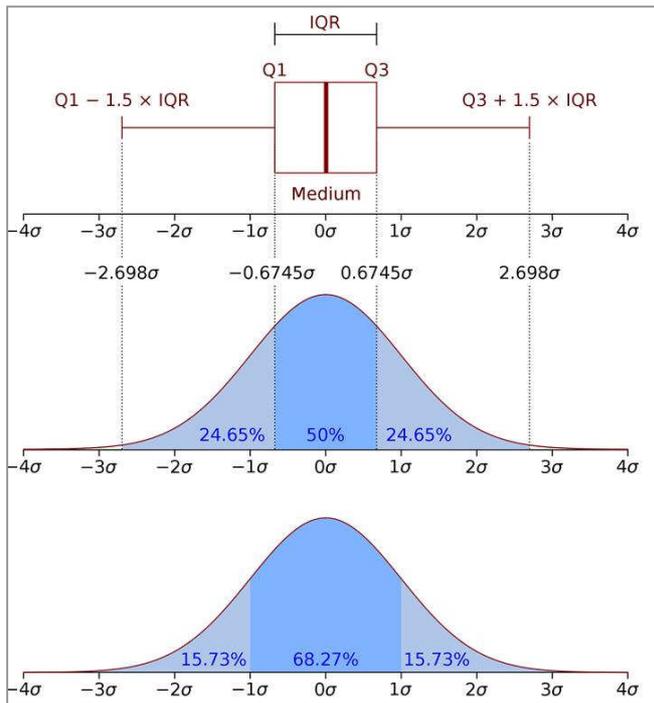


Abbildung 11: Boxplot für Normalverteilung

Interquartilsanstandes (IQA) vom Median abweichen. Sie können als äußerste Punkte in der Graphik bei -9,3% und 21,5% abgelesen werden. Das erste oder untere Quartil wird durch eine Nacherfassungsquote von 2,9% markiert - 25% der Beobachtungen haben eine Nacherfassungsquote bei Daten ohne Wartezeit die kleiner oder gleich 2,9% ist. Umgekehrt wird das dritte oder obere Quartil durch eine Nacherfassungsquote von 5,7% begrenzt, was bedeutet, dass 75% aller Beobachtungen im Datensatz unterhalb dieser Nacherfassungsquote liegen aber eben auch noch ein Viertel der Beobachtungen größere Nacherfassungsquoten aufweisen.

Der Median teilt den Datensatz in zwei Hälften, im Beispiel liegt er bei 4,0% und kann anhand des „Strichs“ in der Box des Boxplots abgelesen werden.

Die Antennen tragen das jeweils 1,5-fache des Interquartilsabstands vom oberen und unteren Quartil aus gesehen ab. Bei einer Normalverteilung liegen etwa 99% der Beobachtungen zwischen diesen Antennengrenzen. Beobachtungen die über diese Grenzen hinaus ragen werden deshalb als Ausreißer klassifiziert. Abbildung 11¹⁶ zeigt für den Fall normalverteilter Messergebnisse den Zusammenhang zwischen den Elementen des Boxplot und den Verteilungsmassen unter der Glockenkurve der Normalverteilung. Dargestellt ist der Zusammenhang zwischen Interquartilsabstand und der Standardabweichung. Die Abbildung legt den Hintergrund für die Festlegung der Kriterien für Ausreißer offen: Innerhalb der unteren und oberen Grenzwerte liegen bei einer Normalverteilung 99,3% aller Fälle, außerhalb sind nur 0,7% zu erwarten, also z.B. bei 427 Jobcentern nur ca. 3 Fälle. Im Falle „schiefer“ Verteilungen liegt die Anzahl als potentielle Ausreißer markierter Fälle häufig etwas höher.

E) Anwendung und Interpretation von Histogrammen und Dichtefunktionen

Ein Histogramm ist eine graphische Darstellung der Häufigkeitsverteilung metrisch skalierten Merkmale. Ein Histogramm erfordert die Einteilung der Daten in Klassen. Im Histogramm werden über den Klassen direkt aneinander angrenzende Rechtecke errichtet, deren Flächeninhalt die Klassenhäufigkeit darstellt.

Mit Histogrammen können sowohl absolute - (Frequenzen) als auch relative Häufigkeiten (Dichten) dargestellt werden. In Abbildung 12 wurde zur Abgrenzung der Visualisierungs-

¹⁶ Quelle: http://en.wikipedia.org/wiki/Box_plot; IQR = IQA, Medium = Median

möglichkeiten die Darstellung als relative Häufigkeit ausgewählt. Hier wurden Histogramm und Dichtekurve für die Nacherfassungsquoten des Bestands an Bedarfsgemeinschaften (Daten ohne Wartezeit) übereinander gelegt um den Einsatz beider Darstellungen vergleichend betrachten zu können.

„Histogramme entsprechen auch optisch nicht der gedanklichen Vorstellung, daß die Verteilung eines stetigen Merkmals besser durch eine glatte Kurve als durch eine – oft sehr sprunghafte – Treppenfunktion repräsentiert wird. Der Vorstellung einer stetigen Verteilung kommt bei umfangreichen Datensätzen die empirische Verteilungsfunktion deutlich näher (...).“¹⁷

Besonderheiten von Histogramm und Verteilungs- oder auch Dichtefunktion sind, dass zum einen die von der Treppenfunktion/Kurve überdeckte Gesamtfläche gleich eins ist und dass zum anderen keine negativen Funktionswerte eintreten können.

Zu lesen ist die Dichtefunktion als eine Vorstellung der Fläche unter der Kurve.

„Die Fläche, die von der Kurve über einem bestimmten Intervall $[a, b]$ von x -Werten begrenzt wird, ist dann als prozentualer Anteil der x -Werte zu interpretieren, die in dieses Intervall fallen. Diese Fläche wird beim Histogramm durch das (...) Rechteck über $[a, b]$ approximiert.“¹⁸

In Abbildung 12 wurde ebendiese Vorstellung graphisch dargestellt. Ein gedachter Bereich unter der Kurve beispielsweise zwischen zwei und fünf Prozent Nacherfassungsquote beschreibt den prozentualen Anteil der monatsweisen Kreisbeobachtungen, die eine Nacherfassungsquote größer oder gleich zwei und kleiner oder gleich fünf Prozent Nacherfassungsquote aufweisen. Der entsprechende Bereich ist rot schattiert dargestellt. Das darunter liegende Histogramm approximiert diesen Bereich als entsprechende Treppenschritte.

¹⁷ Fahrmeier et.al (2004): S.87

¹⁸ Fahrmeier et.al (2004): S.88

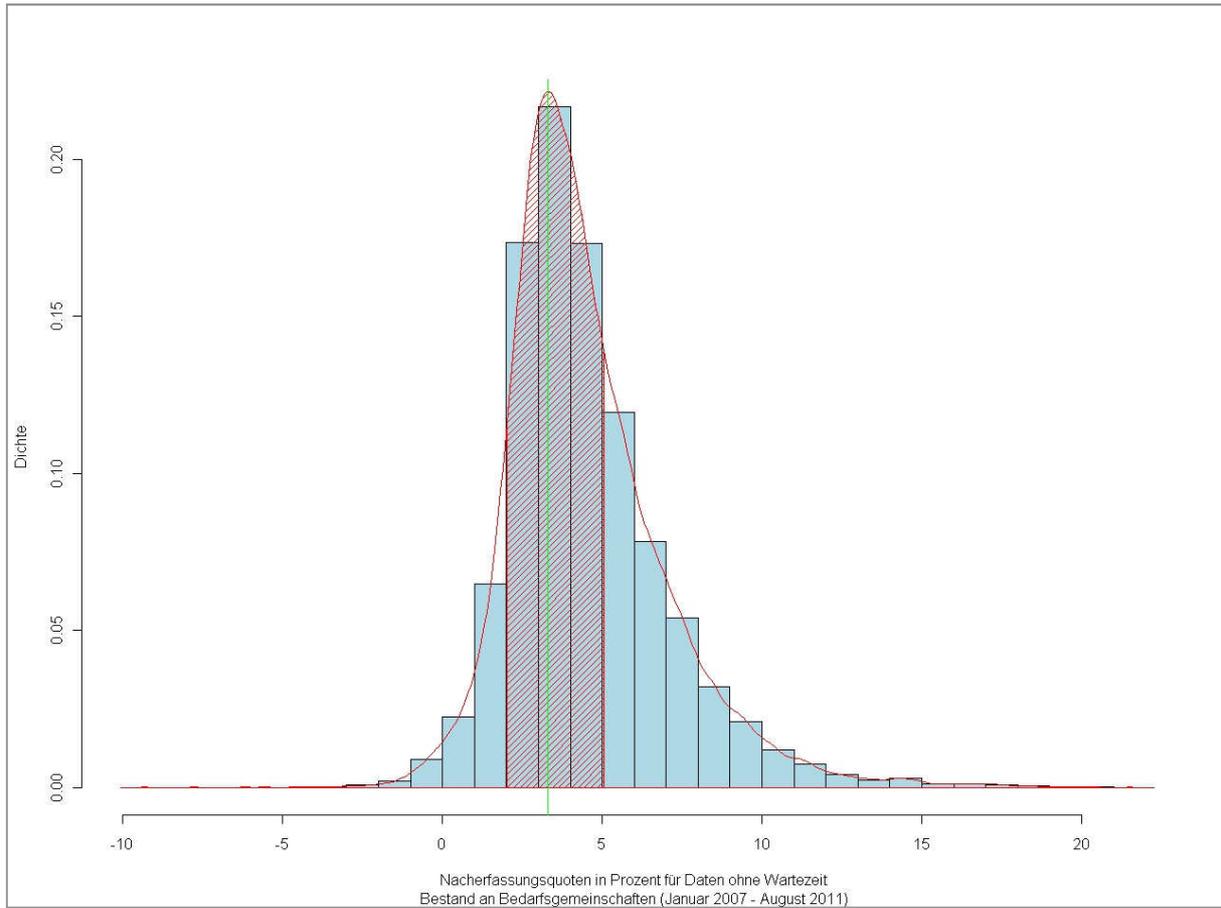


Abbildung 12: Histogramm und Dichtekurve zu Nacherfassungsquoten (Bestand BGN)

Man erkennt, dass der Erwartungswert der Verteilung bei einer Nacherfassungsquote von etwa 3,3% (grüne Linie) liegt. Der Erwartungswert wird durch das Maximum der Dichtekurve beschrieben. In diesem Bereich liegen über den Betrachtungszeitraum Januar 2007 bis August 2011 die meisten kreisbezogenen Beobachtungen. Dem folgend ist die Wahrscheinlichkeit kleinere (oder sogar negative) und größere Nacherfassungsquoten zu beobachten, verhältnismäßig kleiner.

Die hier vorliegende Verteilung gleicht nahezu einer Normalverteilung. Sie ist insbesondere für Zufallsgrößen typisch, die von sehr vielen externen und zusammenwirkenden Einflussgrößen bestimmt sind. Typisch für diese Kurvendarstellung ist, dass sie links und rechts vom Maximum aus glockenförmig gegen Null abfällt. Je steiler dieser Abfall ausfällt und je enger damit die Kurve an sich ist, umso weniger streuen die x-Werte um den Erwartungswert und umso kleiner ist die Standardabweichung.

F) Wartezeitkonzept in der Grundsicherungsstatistik

Vom aktuellen Erhebungszeitpunkt (t_E) aus betrachtet besteht die Möglichkeit auf die vergangenen Stichtage mit unterschiedlichen Wartezeiten zu blicken. Für Stichtag t_{S+3} liegen soweit nur wenig belastbare vorläufige (und unvollständige) Daten ohne Wartezeit vor, während für Stichtag t_{S+2} schon Daten nach einer Wartezeit von einem Monat, für Stichtag t_{S+1} Daten nach einer Wartezeit von zwei Monaten und für Stichtag t_S endgültige, belastbare Daten nach einer Wartezeit von drei Monaten zur Verfügung stehen.

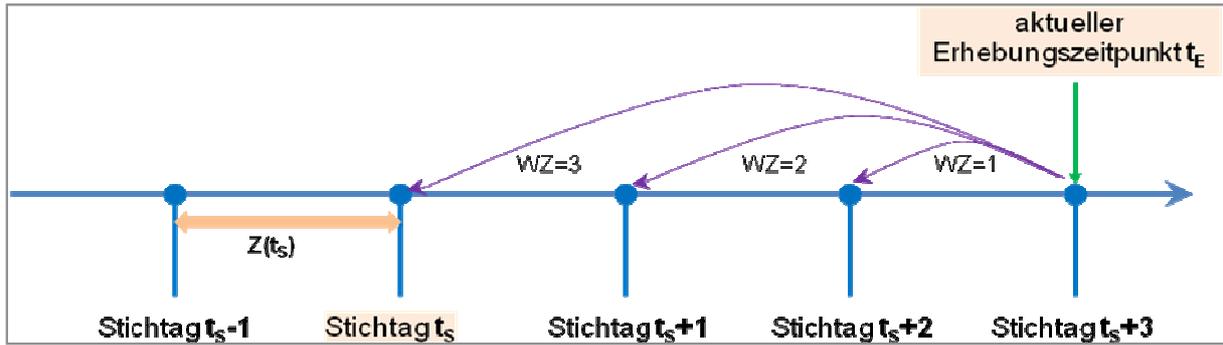


Abbildung 13: Wartezeitkonzept - Sicht aktueller Erhebungszeitpunkt

Gleichzeitig gibt es über den Stichtag t_s vier sukzessiv vervollständigte Sichten, die sich aus dem Ablauf der Wartezeit ergeben. Die Betrachtung des ersten Erhebungszeitpunktes liefert am Stichtag t_s einen Datenstand ohne Wartezeit. Nach Ablauf eines weiteren Monats kann am Stichtag t_{s+1} der Blick auf t_s zurück gerichtet werden und man kann eine teilweise vervollständigte Datenbasis analysieren, am Stichtag t_{s+2} liegen dann weiter vervollständigte Daten nach einer Wartezeit von zwei Monaten und zum Stichtag t_{s+3} endgültige und damit belastbare Daten nach einer Wartezeit von drei Monaten vor.

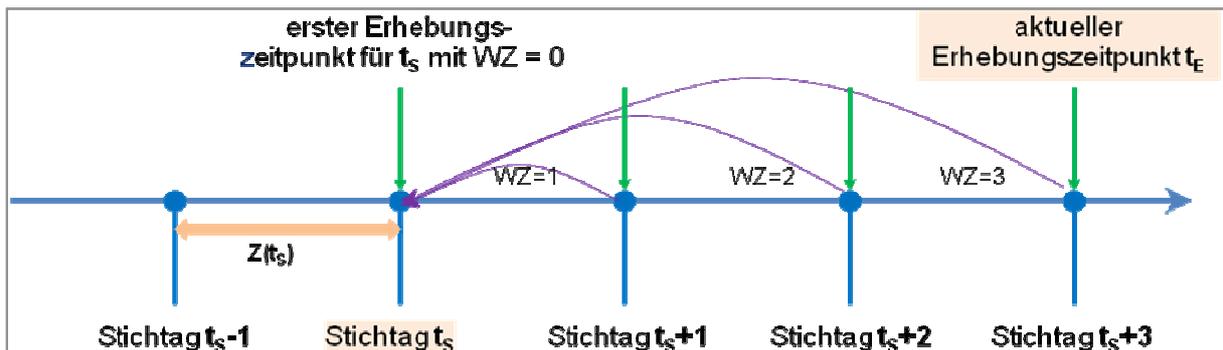


Abbildung 14: Wartezeitkonzept - Sicht Wartezeitstände für ersten Erhebungszeitpunkt

Quellenverzeichnis

Fahrmeier, Ludwig et al. (2004): Statistik - Der Weg zur Datenanalyse; 5. Auflage; Springer Verlag, München

Rüb; Werner (2007): Typisierung von SGB II-Trägern; IAB Forschungsbericht Nr. 1/2007; Nürnberg

Tentscher, Wolff, Seidel (2011): Ausgleich operativ untererfasster Eckwerte auf Regionalebene in der Leistungsstatistik SGB II; Methodenbericht der Statistik der BA; Nürnberg

Wikipedia, the free encyclopedia (2011): Box plot; http://en.wikipedia.org/wiki/Box_plot (abgefragt am 27.10.2011)

Statistik-Infoseite

Im **Internet** finden Sie weiterführende Informationen der Statistik der Bundesagentur für Arbeit unter:

<http://statistik.arbeitsagentur.de>

Statistische Daten erhalten Sie unter "Statistik nach Themen":

[Statistik nach Themen](#)

Es werden folgende Themenbereiche angeboten:

[Arbeitsmarkt im Überblick](#)
[Arbeitslose und gemeldetes Stellenangebot](#)
[Arbeitsmarktpolitische Maßnahmen](#)
[Ausbildungsstellenmarkt](#)
[Beschäftigung](#)
[Grundsicherung für Arbeitsuchende \(SGB II\)](#)
[Leistungen SGB III](#)
[Statistik nach Berufen](#)
[Statistik nach Wirtschaftszweigen](#)
[Zeitreihen](#)
[Eingliederungsbilanzen](#)
[Einnahmen/Ausgaben der BA](#)

Daten bis 12/2004 finden Sie unter dem Menüpunkt "Archiv bis 2004"

Glossare zu den verschiedenen Fachstatistiken finden Sie hier:

[Glossare](#)

Es werden folgende Themenbereiche angeboten:

[Arbeitsmarkt](#)
[Ausbildungsstellenmarkt](#)
[Beschäftigung](#)
[Förderstatistik/Eingliederungsbilanzen](#)
[Grundsicherung für Arbeitsuchende \(SGB\)](#)
[Leistungen SGB III](#)

Hintergründe zur Statistik nach dem SGB II und III und zur Datenübermittlung nach § 51b SGB II finden Sie unter dem Auswahlpunkt "Grundlagen":

[Grundlagen der Statistik](#)

Für weitere Datenwünsche, Sonderauswertungen und Auskünfte:

Bundesagentur für Arbeit
Statistik Datenzentrum
Hotline: 01801 / 78 722 10 *)
Fax: 01801 / 78 722 11 *)
E-Mail: statistik-datenzentrum@arbeitsagentur.de
Post: Regensburger Straße 104, 90478 Nürnberg

^{*)} 3,9 Cent je Minute aus dem Festnetz der Deutschen Telekom. Bei Anrufen aus Mobilfunknetzen höchstens 42 ct / min.