DIW BERLIN

217

Politikberatung kompakt

Deutsches Institut für Wirtschaftsforschung 2026

Prognosemodelle für das Medianäquivalenzeinkommen

Geraldine Dany-Knedlik, Laura Pagenhardt und Jana Wittich

IMPRESSUM

DIW Berlin, 2026

DIW Berlin
Deutsches Institut für Wirtschaftsforschung
Anton-Wilhelm-Amo-Straße 58
10117 Berlin
Tel. +49 (30) 897 89-0
Fax +49 (30) 897 89-200
www.diw.de

ISBN 978-3-911978-08-8
ISSN 1614-6921

Alle Rechte vorbehalten.
Abdruck oder vergleichbare
Verwendung von Arbeiten
des DIW Berlin ist auch in
Auszügen nur mit vorheriger
schriftlicher Genehmigung
gestattet.

DIW Berlin: Politikberatung kompakt 217

Dr. Geraldine Dany-Knedlik *
Dr. Laura Pagenhardt **
Jana Wittich ***

Prognosemodelle für das Medianäquivalenzeinkommen

Kurzexpertise im Auftrag des Bundesministeriums der Finanzen

Berlin, 6. Oktober 2026

* DIW Berlin, Abteilung Makroökonomie, GDanyKnedlik@diw.de
** DIW Berlin, Abteilung Makroökonomie, LPagenhardt@diw.de
*** DIW Berlin, Abteilung Makroökonomie, JWittich@diw.de

Wir danken Nadja Kockrow für ihre Unterstützung bei der Recherche und Datenaufbereitung.

Inhaltsverzeichnis

1 Einleitung und Hintergrund 1

2 Rahmenbedingungen der Prognoseevaluation 3

2.1 Datengrundlage der Schätzung 3

2.2 Evaluationsdesign (*Pseudo-Echtzeit*) 4

2.3 Prognosehorizonte 5

2.4 Evaluationsmaßstab und Evaluationszeitraum 6

3 Modelle 6

3.1 Bezugsmodelle 7

3.2 Trendfortschreibungen 7

3.3 Univariate Zeitreihenmodelle 8

3.4 Makroökonomische Brückenmodelle 9

3.5 Fehlermaße und Verlustfunktionen 10

4 Ergebnisse der Modellevaluation 11

4.1 Prognosegüte im Überblick 11

4.2 Fehler im Einkommensniveau 16

4.3 Bewertung unter asymmetrischem Verlust 20

5 Schlussfolgerungen und Handlungsempfehlungen 22

A Anhang 27

A.1 Zusätzliche Informationen zu den Schätzbedingungen 27

A.2 Gewählte ARIMA-Ordnungen 29

A.3 Ausführliche Schätzergebnisse: Prognosegüte am EU-SILC-Maßstab 29

A.4 Zusätzliche Schätzergebnisse: Herbst 32

Tabellenverzeichnis

1 Verfügbarer Datenstand zum Prognosezeitpunkt im Jahr t 6

A.1 Augmented-Dickey-Fuller-Tests: Niveaus und Veränderungsraten 28

A.2 ARIMA(p, d, q)-Ordnungen pro Modell und Prognosezeitpunkt 29

Abbildungsverzeichnis

1 Prognosegüte am EU-SILC-Maßstab (ohne Bruchjahr 2020) – Frühjahr, Horizonte $t-2$ und $t-1$ . . . 12
2 Prognosegüte am EU-SILC-Maßstab (ohne Bruchjahr 2020) – Frühjahr, Horizonte $t+0$ und $t+1$ . . . 13
3 Prognosegüte am EU-SILC-Maßstab (ohne Bruchjahr 2020) – Herbst, Horizonte $t-1$ und $t+0$ . . . 13
4 Prognosegüte am EU-SILC-Maßstab (ohne Bruchjahr 2020) – Herbst, Horizonte $t+1$ und $t+2$ . . . 14
5 Verkettete Prognosepfade ausgewählter Modelle und realisierte EU-SILC-Reihe – Frühjahr . . . 16
6 Fehler der verketteten Niveauprognosen – Frühjahr I/II . . . 18
7 Fehler der verketteten Niveauprognosen – Frühjahr II/II . . . 19
8 Rangverschiebung zwischen symmetrischer und asymmetrischer Bewertung – Frühjahr I/II . . . 21
9 Rangverschiebung zwischen symmetrischer und asymmetrischer Bewertung – Frühjahr II/II . . . 22
A.1 Medianäquivalenzeinkommen verschiedener SOEP-Wellen . . . 27
A.2 Prognosegüte am EU-SILC-Maßstab (ohne Bruchjahr 2020), alle Modelle – Frühjahr . . . 30
A.3 Prognosegüte am EU-SILC-Maßstab (ohne Bruchjahr 2020), alle Modelle – Herbst 31
A.4 Verkettete Prognosepfade ausgewählter Modelle und realisierte EU-SILC-Reihe – Herbst . . . 32
A.5 Fehler der verketteten Niveauprognosen – Herbst I/II . . . 33
A.6 Fehler der verketteten Niveauprognosen – Herbst II/II . . . 33
A.7 Rangverschiebung zwischen symmetrischer und asymmetrischer Bewertung – Herbst I/II . . . 34
A.8 Rangverschiebung zwischen symmetrischer und asymmetrischer Bewertung – Herbst II/II . . . 35

1 Einleitung und Hintergrund

In den vergangenen Jahren hat sich das Medianäquivalenzeinkommen zu einer Schlüsselgröße sozial- und verteilungspolitischer Analysen entwickelt. Zusätzliches Gewicht erhält diese Kennziffer durch den Beschluss des Bundesverfassungsgerichts vom 17. September 2025 (2 BvL 5/18 u.a.). Danach hat sich die Mindestbesoldung von Beamtinnen und Beamten an einem Zielwert von 80 Prozent des Medianäquivalenzeinkommens zu orientieren, der sogenannten Prekaritätsschwelle. Da diese Rechtsprechung voraussichtlich auch auf weitere anhängige Verfahren Anwendung finden wird, rücken Entwicklung und Prognostizierbarkeit dieser Größe unmittelbar in den Mittelpunkt der Besoldungsgesetzgebung von Bund und Ländern.

So ist durch die verbindliche Anknüpfung der Mindestbesoldung an das Medianäquivalenzeinkommen eine möglichst präzise Prognose dieser Kennzahl wichtig für die mittelfristige Haushaltsplanung. Dabei sind die beiden Richtungen eines Prognosefehlers jedoch unterschiedlich zu bewerten: Eine Unterschätzung der künftigen Entwicklung kann zu Nachzahlungen führen, während eine Überschätzung vermeidbare Haushaltsbelastungen nach sich zieht. Hinzu kommt, dass sich das Medianäquivalenzeinkommen aus Haushaltsnettoeinkommen und Äquivalenzgewichtung zusammensetzt und daher eigene analytische und prognostische Herausforderungen aufwirft, die es von vielen anderen gesamtwirtschaftlichen Aggregaten mit etablierten Prognoseinstrumenten unterscheidet.

Eine erste Kurzexpertise des DIW Berlin hat für die Analyse der Prognostizierbarkeit des Medianäquivalenzeinkommens die methodische Grundlage gelegt (Dany-Knedlik, Hüttl und Pagenhardt, 2026). Darin wurde gezeigt, dass die Daten aus der EU-Statistik über Einkommen und Lebensbedingungen (*European Union Statistics on Income and Living Conditions*, EU-SILC),[1] auf denen das offizielle Medianäquivalenzeinkommen für Deutschland basiert, statistische Umstellungen und damit Strukturbrüche aufweisen, die bei der Anwendung gängiger Analyse- und Prognosemodelle berücksichtigt werden müssen. Zugleich zeigt die Analyse, dass eine aus dem Sozio-oekonomischen Panel (SOEP) gewonnene Zeitreihe das Medianäquivalenzeinkommen valide nachbildet und sich damit als Basis für die Evaluation von Prognosemodellen eignet.

[1] Verfügbar in Eurostat als *Durchschnittliches und Median-Einkommen nach Alter und Geschlecht* (`ilc_di03`); Link zur Eurostat-Datenbank.

Erste deskriptive Befunde deuten zudem darauf hin, dass makroökonomische Prognosen der verfügbaren Einkommen ein geeigneter Indikator sein könnten, um die mittlere Entwicklung der Kennziffer nachzuzeichnen. Daraus ergab sich die Frage, wie verschiedene gängige Prognoseverfahren hinsichtlich ihrer Vorhersagegüte und Robustheit im systematischen Vergleich abschneiden.

Hier setzt die vorliegende Kurzexpertise an. Ihr Ziel ist die Entwicklung eines operativ einsetzbaren Prognoseinstruments für Horizonte von einem bis zwei Jahren, das zuverlässige Schätzungen liefert und zugleich möglichst einfach, nachvollziehbar und reproduzierbar ist. Dazu werden mehrere Klassen von Prognoseverfahren mit zunehmender Komplexität – von naiven Richtwerten über einfache Trendfortschreibungen und univariate Zeitreihenmodelle bis hin zu makroökonomischen Brückenmodellen – in einem einheitlichen Rahmen gegeneinander getestet und ihre Prognosegüte in einer Pseudo-Echtzeit-Schätzung (*Pseudo-Out-of-Sample-Evaluation*) anhand des offiziellen Medianäquivalenzeinkommens auf Basis der EU-SILC-Daten bewertet.[2] Für jeden historischen Prognoseursprung wird der damalige Informationsstand nachgebaut, und die Modelle werden ausschließlich mit den Daten geschätzt, die zum jeweiligen Zeitpunkt verfügbar gewesen wären.

Konkret beantwortet die Kurzexpertise die Fragen: Welches der untersuchten Prognoseverfahren liefert über die Horizonte von einem bis zwei Jahren die höchste und robusteste Vorhersagegüte für das Medianäquivalenzeinkommen in Deutschland? Wie schneiden naive Bezugsmodelle, einfache Trendfortschreibungen, univariate Zeitreihenmodelle und makroökonomische Brückenmodelle im Vergleich zueinander ab, und welches Verfahren bietet den besten Kompromiss zwischen Prognosegüte, Einfachheit und praktischer Anwendbarkeit?

Der weitere Bericht ist wie folgt aufgebaut. Abschnitt 2 legt die einheitlichen Rahmenbedingungen der Evaluation dar. Abschnitt 3 stellt die vier Modellklassen und die verwendeten Fehlermaße vor. Abschnitt 4 berichtet über die Ergebnisse der Modellevaluation. Abschnitt 5 fasst die Befunde zusammen und leitet Handlungsempfehlungen ab.

[2] Die vergleichende Out-of-Sample- bzw. Pseudo-Out-of-Sample-Evaluation konkurrierender Prognosemodelle ist dabei ein etabliertes Verfahren der empirischen Prognoseliteratur; vgl. etwa Diebold und Lopez, 1996 und West u. a., 2006.

2 Rahmenbedingungen der Prognoseevaluation

Bevor die einzelnen Modellklassen vorgestellt und evaluiert werden, legt dieser Abschnitt den gemeinsamen Rahmen fest, innerhalb dessen alle Verfahren geschätzt, selektiert und bewertet werden. Die dargestellten Rahmenbedingungen betreffen die Datengrundlage der Schätzung, den Nachbau des historischen Informationsstands (Pseudo-Echtzeit), die Prognosehorizonte, die verwendeten makroökonomischen Indikatoren sowie den Evaluationsmaßstab. Sie gelten einheitlich für alle Modellklassen und stellen sicher, dass die Ergebnisse vergleichbar sind.

2.1 Datengrundlage der Schätzung

Die Schätzung und Selektion der Prognosemodelle erfolgt auf Grundlage der aus dem SOEP gewonnenen Zeitreihe des Medianäquivalenzeinkommens. Ausschlaggebend für diese Wahl sind die im ersten Bericht dokumentierten Befunde: Das SOEP liefert eine lange, bruchfreie Reihe ab 1991, die die offizielle EU-SILC-Kennzahl valide nachbildet, während die EU-SILC-Reihe selbst zu kurz und durch den Strukturbruch im Erhebungsjahr 2020 zu stark beeinträchtigt ist, um darauf eine belastbare Modellschätzung aufzubauen (Dany-Knedlik, Hüttl und Pagenhardt, 2026).

Am aktuellen Rand wird die SOEP-Reihe um jeweils ein EU-SILC-Jahr ergänzt. Hintergrund ist die unterschiedliche Publikationsgeschwindigkeit der beiden Erhebungen: EU-SILC-Ergebnisse liegen rund ein Jahr früher vor als die entsprechende SOEP-Welle. Somit wird die beobachtete Reihe um ein Jahr verlängert und die Modelle können auf dem aktuellsten verfügbaren Informationsstand aufsetzen. Die hohe Übereinstimmung der Veränderungsraten beider Quellen, die im ersten Bericht dokumentiert wurde, rechtfertigt dieses Vorgehen.

Als Prognoseindikatoren für die Brückenmodelle dienen die Prognosen der Gemeinschaftsdiagnose (GD), die ab 1991 vorliegen.[3] Berücksichtigt werden die verfügbaren Einkommen der privaten Haushalte, die Nettolöhne und -gehälter sowie die Bruttolöhne und -gehälter je Be-

[3] Die GD ist die seit 1950 halbjährlich gemeinschaftlich erstellte Konjunkturprognose der führenden deutschen Wirtschaftsforschungsinstitute (derzeit DIW Berlin, ifo, IfW Kiel, IWH und RWI) im Auftrag des Bundeswirtschaftsministeriums. Sie liefert eine vollständig ausformulierte Prognose der deutschen Volkswirtschaftlichen Gesamtrechnungen (VGR) und dient der Bundesregierung als Referenz für ihre wenig später veröffentlichte gesamtwirtschaftliche Projektion, die die Projektgruppe als unabhängige Institution nach dem Vorausschätzungsgesetz zugleich auf Plausibilität prüft.

schäftigten (im Folgenden BLG pro Kopf). Letztere gehen trotz ihres Bruttokonzepts in den Modellvergleich ein, da sie die einzige Pro-Kopf-Größe der Einkommensentwicklung sind, die über den für die Prognoseevaluation erforderlichen Zeitraum konsistent in den Prognosen der GD explizit ausgewiesen wird.[4]

Alle Modelle werden auf die jährlichen Veränderungsraten geschätzt. Dafür sprechen zwei Gründe. Erstens ist die Veränderungsrate das operative Zielformat: Der Prognosewert ergibt sich als Verkettung prognostizierter Raten ausgehend vom jeweils letzten amtlichen EU-SILC-Wert, sodass das Niveau der SOEP-Reihe selbst nicht in die Prognose eingeht. Zweitens erfüllt erst die Veränderungsrate die für die Zeitreihenmodellierung zentrale Stationaritätsanforderung (Anhang A.1, Tabelle A.1).

2.2 Evaluationsdesign (*Pseudo-Echtzeit*)

Die Prognosegüte der Modelle wird in einem Pseudo-Echtzeit-Rahmen ermittelt: Für jeden historischen Prognoseursprung wird der damalige Informationsstand nachgebaut, und die Modelle werden ausschließlich mit den Daten geschätzt, die zum jeweiligen Zeitpunkt verfügbar gewesen wären. Für die Prognosen der GD wird dies strikt umgesetzt, indem ausschließlich die seinerzeit veröffentlichten Werte aus den Original-Gutachten verwendet werden (Echtzeit-Daten), nicht die heute vorliegenden revidierten Reihen.

Für die Einkommensreihen aus SOEP und EU-SILC wird demgegenüber auf den aktuellen Datenstand zurückgegriffen. Streng genommen handelt es sich damit nicht um eine richtige Echtzeit-Evaluation, da jede neue Umfragewelle rückwirkend Gewichte und generierte Variablen revidiert. Diese Vereinfachung ist jedoch unschädlich, da die Revisionen der SOEP-Reihe über die Wellen hinweg vernachlässigbar klein sind (Anhang A.1, Abbildung A.1). Für EU-SILC stellt sich die Revisionsfrage ohnehin kaum, da einmal veröffentlichte Erhebungsjahre nicht in Form von

[4] In älteren GD-Veröffentlichungen werden teilweise auch die Nettolöhne und -gehälter je Beschäftigten angegeben. Diese Reihe steht jedoch nicht über den gesamten Zeitraum in vergleichbarer Form zur Verfügung. Nettolöhne und -gehälter sowie verfügbare Einkommen werden daher nur als gesamtwirtschaftliche Aggregate verwendet. Entsprechende Pro-Kopf-Größen ließen sich zwar grundsätzlich aus den Aggregaten und den prognostizierten Beschäftigtenzahlen ableiten. Der vorliegende Modellvergleich stützt sich jedoch bewusst auf Größen, die konsistent in der GD ausgewiesen werden, um Nachvollziehbarkeit und Reproduzierbarkeit ohne zusätzliche Umrechnungsschritte zu gewährleisten. Die aggregierten BLG wurden ebenfalls geprüft, siehe Abbildungen A.2 und A.3 in Anhang A.3.

Vintages fortgeschrieben werden. Relevant ist hier vielmehr der methodische Bruch der deutschen EU-SILC-Erhebung im Jahr 2020, als die Erhebung in den Mikrozensus integriert wurde, verbunden mit dem Wechsel von der Panel- zur Flächenzufallsstichprobe, einer Verdopplung des Stichprobenumfangs, einer teilweisen Auskunftspflicht sowie geänderter Gewichtung und Imputation. Ergebnisse ab 2020 sind daher nach Angaben des Statistischen Bundesamtes nicht mit denen der vorherigen Erhebungsjahre vergleichbar. Die mit dem finalen Datenstand ermittelte Prognosegüte kann daher als gute Näherung der operativ erreichbaren Güte gelten.

2.3 Prognosehorizonte

Die relevanten Prognosehorizonte ergeben sich aus dem Zusammenspiel von Verwendungszweck und Publikationsverzögerung der Einkommensdaten. Für die Besoldungsgesetzgebung wird üblicherweise das Medianäquivalenzeinkommen des laufenden und des folgenden Jahres, unter Umständen auch bis zu zwei Jahre im Voraus, benötigt. Zugleich liegen die amtlichen Einkommensdaten zum Entscheidungszeitpunkt stets mit erheblicher Verzögerung von teilweise mehreren Jahren vor. Ein Teil des Prognosepfads betrifft daher Einkommensjahre, die kalendarisch bereits abgeschlossen, aber noch nicht beobachtet sind.

Der verfügbare Informationsstand unterscheidet sich zudem danach, wann die Prognose im Jahr t erstellt wird (Tabelle 1). Im vorliegenden Bericht werden zwei Zeitpunkte herangezogen, die sich an den Veröffentlichungen der GD orientieren – Frühjahrs- und Herbstprognosen werden in der Evaluation getrennt abgebildet, da sie unterschiedliche effektive Prognosehorizonte implizieren.[5]

Zu diesem Zeitpunkt sind die aktuellen Umfrageergebnisse in der Regel noch nicht veröffentlicht. Bei einer Vorausschätzung im Frühjahr (Herbst) reicht die SOEP-Reihe bis zum Einkommensjahr $t-4$ ($t-3$), und die EU-SILC-Reihe bis zum Jahr $t-3$ ($t-2$). Schließlich wird die Reichweite der zukunftsgerichteten Prognosen durch die Verfügbarkeit der makroökonomischen Indikatoren begrenzt: Die Herbstgutachten der GD reichen seit 2016 bis zum Jahr $t+2$, davor lediglich bis zum Jahr $t+1$. Ein weiteres Prognosejahr im Voraus mit Horizont $t+3$ ließe

[5] Bei den GD-Prognosen ist darüber hinaus zu beachten, dass sich deren Reichweite im Zeitverlauf verändert hat. Die Frühjahrsgutachten enthalten seit 1997 eine Prognose für das Folgejahr $t+1$, zuvor beschränkten sie sich auf das laufende Jahr t. Die historisch kürzere Reichweite der GD-Prognosen begrenzt zugleich, ab wann die längeren Horizonte der Brückenmodelle evaluiert werden können.

Tabelle 1: Verfügbarer Datenstand zum Prognosezeitpunkt im Jahr t

Prognosezeitpunkt	Zeitraum	GD-Prognosen	SOEP	EU-SILC
Frühjahr (April)	bis 1996	$t+0$	$t-4$	$t-3$
	ab 1997	$t+1$		
Herbst (Sept./Okt.)	bis 2015	$t+1$	$t-3$	$t-2$
	ab 2016	$t+2$		

Anmerkung: Die Angaben beziehen sich auf das jeweils zuletzt verfügbare Medianäquivalenzeinkommen aus den Erhebungen von SOEP und EU-SILC bzw. auf die zum jeweiligen Prognosezeitpunkt maximal verfügbaren Vorausschätzungen der entsprechenden Prädiktoren aus den GD-Prognosen.

sich daher nur mit univariaten Modellen abdecken. Entsprechend wurden die Prognosegüten der verschiedenen Modelle von $t-2$ bis $t+2$ verglichen.

2.4 Evaluationsmaßstab und Evaluationszeitraum

Die Prognosegüte wird an der amtlichen EU-SILC-Reihe bewertet. Evaluiert werden die Jahre 2010 bis 2024; für Modellvarianten, deren Eingangsdaten erst später verfügbar sind, wird der Zeitraum entsprechend verkürzt. Zu berücksichtigen ist schließlich der Strukturbruch der EU-SILC-Reihe (vgl. Dany-Knedlik, Hüttl und Pagenhardt, 2026, Abschnitt 2.2): Mit dem Erhebungsjahr 2020 (Einkommensjahr 2019) wurde die Befragung in Deutschland in den Mikrozensus integriert, wodurch die zugehörige Veränderungsrate nicht als tatsächliche Einkommensveränderung interpretierbar ist. Diese Beobachtung wird daher sowohl aus der Schätzung als auch aus der Evaluierung ausgeschlossen.

3 Modelle

Die untersuchten Prognoseverfahren lassen sich in vier Klassen einteilen, die sich in aufsteigender Reihenfolge durch zunehmende Komplexität und einen wachsenden Informationseinsatz auszeichnen: naive Richtwerte, deterministische Trendfortschreibungen, univariate Zeitreihenmodelle sowie makroökonomische Brückenmodelle, die externe Prognoseinformation der GD einbeziehen. Für die Prognoseevaluation der Modelle gelten die in Abschnitt 2 beschriebenen Rahmenbedingungen. Prognoseobjekt ist die Veränderungsrate des Medianäquivalenzeinkommens und geschätzt wird auf der SOEP-Reihe, die ergänzt wird um den aktuellsten Wert aus der

EU-SILC-Erhebung.[6] Die Prognosen basieren auf Modellen, die anhand des jeweils verfügbaren Informationsstandes geschätzt wurden, sodass das Schätzfenster mit der Zeit expandiert. Der prognostizierte Einkommenswert ergibt sich anschließend stets durch Verkettung der Ratenprognosen ausgehend vom letzten amtlichen EU-SILC-Wert (vgl. Abschnitt 2.3).

3.1 Bezugsmodelle

Als Referenzmaßstab dienen zwei bewusst einfach gehaltene Verfahren. Der *Random Walk* setzt die Prognose für alle Horizonte gleich der zuletzt beobachteten Veränderungsrate. Er ist der geeignete Vergleichsmaßstab, wenn die Veränderungsrate hochpersistent ist und das aktuelle Umfeld die beste Auskunft über die nahe Zukunft gibt. Das *historische Mittel* verwendet stattdessen den Durchschnitt aller seit 1991 beobachteten Veränderungsraten als Prognose für alle Horizonte. Es ist der geeignete Maßstab, wenn die Rate um einen stabilen Mittelwert schwankt und kurzfristige Ausschläge keine prognostische Information tragen.
Die Funktion der beiden Referenzmodelle ist disziplinierend. Sie definieren die Untergrenze dessen, was ohne jeden Modellierungsaufwand erreichbar ist. Ein komplexeres Verfahren rechtfertigt seinen zusätzlichen Aufwand – und die mit ihm verbundenen Ermessensspielräume – nur, wenn es diese Untergrenze systematisch unterschreitet.

3.2 Trendfortschreibungen

Trendfortschreibungen nutzen den ausgeprägten langfristigen Aufwärtstrend des nominalen Medianäquivalenzeinkommens. Auf das logarithmierte Einkommensniveau wird ein linearer Trend geschätzt:

$$\ln y_t \;=\; a + b\,t + u_t, \tag{1}$$

wobei y_t das nominale Medianäquivalenzeinkommen im Jahr t, b die Trendsteigung und u_t die Abweichung vom Trend bezeichnet. Die Steigung b des log-linearen Trends impliziert eine konstante jährliche Veränderungsrate $\hat{g} = e^b - 1 \approx b$, die als Prognose für alle Horizonte dient. Die zentrale Spezifikationsentscheidung ist der Beginn des Schätzfensters. Ein langes Fenster

[6] Siehe dazu Dany-Knedlik, Hüttl und Pagenhardt, 2026.

liefert stabile, wenig revisionsanfällige Schätzungen, unterstellt aber, dass das Wachstumstempo der frühen 1990er Jahre für die Gegenwart noch repräsentativ ist. Ein kurzes Fenster reagiert schneller auf Veränderungen des Trendwachstums, ist dafür aber anfälliger für zufällige Schwankungen weniger Beobachtungen. Um dieses Spannungsfeld abzubilden, werden vier Varianten mit Fensterbeginn 1991, 2000, 2005 und 2010 geführt.[7] Da die Trends an jedem Prognoseursprung neu geschätzt werden, expandieren die Fenster über die Evaluation hinweg.

3.3 Univariate Zeitreihenmodelle

Trendfortschreibungen ignorieren die kurzfristige Dynamik der Reihe. Univariate Zeitreihenmodelle der ARIMA-Klasse ergänzen diese Perspektive, indem sie die Veränderungsrate aus ihrer eigenen Vergangenheit erklären: Autoregressive Terme (AR) nutzen zurückliegende Werte der Rate, Moving-Average-Terme (MA) vergangene Prognosefehler. Für die stationäre Veränderungsrate lautet die Modellgleichung:

$$g_t \;=\; c + \sum_{i=1}^{p} \phi_i \, g_{t-i} + \sum_{j=1}^{q} \theta_j \, \varepsilon_{t-j} + \varepsilon_t \qquad \varepsilon_t \sim \mathsf{WN}(0, \sigma^2), \tag{2}$$

mit der Konstanten c, den AR-Koeffizienten $\phi_1, \ldots, \phi_p$, den MA-Koeffizienten $\theta_1, \ldots, \theta_q$ und dem Fehlerterm ε_t. Die Ordnungen p und q bestimmen, wie viele Terme der jeweiligen Art eingehen.

Die konkrete Ordnung wird an jedem Prognoseursprung ausgewählt. Dafür wird zunächst eine Maximalordnung ($p, q \;=\; 3$) festgelegt und durch alle möglichen Kombinationen iteriert. Die optimale Ordnung wird mittels Akaike-Informationskriterium (AIC) ermittelt (vgl. Tabelle A.2 in Anhang A.2). Dieses Vorgehen bildet ab, wie das Verfahren auch operativ eingesetzt würde, und liefert zugleich einen eigenständigen Befund: Eine über die Ursprünge hinweg instabile Selektion wäre ein gewichtiges Argument gegen komplexe Spezifikationen, deren Ergebnisse sich von Jahr zu Jahr schwer nachvollziehbar ändern würden.

Es werden zwei Varianten geprüft, die sich im Integrationsgrad unterscheiden. Die Hauptvariante modelliert direkt die Veränderungsrate. Der Differenzenparameter wird dabei auf $d = 0$

[7] Auf einen späteren Fensterbeginn wird aufgrund der geringen Anzahl an Evaluierungspunkten verzichtet.

fixiert, da die Veränderungsrate, wie in Abschnitt 2.1 gezeigt, stationär ist. Als Alternative wird eine Spezifikation auf dem (logarithmierten) Einkommensniveau mit fixiertem $d = 1$ geschätzt. Sie modelliert damit ebenfalls die erste Differenz der Reihe, lässt der Selektion aber einen anderen Spielraum bei Trend- und Konstantenbehandlung.

3.4 Makroökonomische Brückenmodelle

Alle bisherigen Verfahren sind univariate Modelle, nutzen also ausschließlich die Vergangenheit der Zielreihe selbst. Brückenmodelle beziehen externe Informationen ein, nämlich die in Abschnitt 2.1 beschriebenen Prognosen der GD für die verfügbaren Einkommen, die Nettolöhne und -gehälter sowie die BLG pro Kopf.[8] Die Grundidee ist, die makroökonomische Prognoseexpertise über eine „Brücke“ auf das Medianäquivalenzeinkommen zu übertragen.
Die Hauptspezifikation ist eine horizontspezifische Regression der Veränderungsrate des Medianäquivalenzeinkommens auf die GD-Rate des jeweiligen Zieljahres:

$$g_{t+h} \;=\; \alpha_h + \beta_h \, x^{(h)}_{t+h} + \varepsilon_{t+h}, \tag{3}$$

wobei $x^{(h)}_{t+h}$ diejenige GD-Prognose bezeichnet, die zum Entscheidungszeitpunkt die jüngste verfügbare Information über das Zieljahr $t + h$ darstellt. Achsenabschnitt α_h und Steigung β_h werden je Horizont frei geschätzt, sodass systematische Niveau- und Sensitivitätsunterschiede zwischen den Konzepten der GD und des Medianäquivalenzeinkommens von den Koeffizienten aufgefangen werden. Für den Prognoseursprung t ergibt sich die Punktprognose als $\hat{g}_{t+h} = \hat{\alpha}_h + \hat{\beta}_h x^{(h)}_{t+h}$.
Ergänzend werden zwei Varianten geprüft, die die Brückengleichung in entgegengesetzte Richtungen abwandeln. Zum einen wird Gleichung (3) um die verzögerte abhängige Variable erweitert (+γg_{t+h-1}), sodass neben der GD-Information auch die Persistenz der Rate selbst eingeht. Dabei ist g_t die am Prognoseursprung zuletzt beobachtete Veränderungsrate. Zum anderen wird die GD-Rate ohne Schätzung direkt als Prognose übernommen ($\alpha_h = 0$, $\beta_h = 1$),

[8] Für die verfügbaren Einkommen und die Nettolöhne und -gehälter wird die Entwicklung in absoluten Größen verwendet (vgl. Fußnote 4).

sodass $\hat{g}_{t+h} = x^{(h)}_{t+h}$. Diese „blanke“ Anwendung entspricht der naheliegenden Praxis, die makroökonomische Prognose unverändert auf das Medianäquivalenzeinkommen zu übertragen.

3.5 Fehlermaße und Verlustfunktionen

Alle Fehlermaße werden für jedes Modell und jeden Horizont identisch berechnet. Ausgangspunkt ist der Prognosefehler der Veränderungsrate:

$$e_{t+h} \;=\; \hat{g}_{t+h} - g_{t+h} \tag{4}$$

in Prozentpunkten, wobei $\hat{g}_{t+h}$ die am Ursprung t erstellte Prognose für den Horizont h und g_{t+h} die tatsächlich realisierte Veränderungsrate desselben Zieljahres ist. Ein negativer Fehler bedeutet eine Unterschätzung des Einkommenswachstums (Risiko einer zu niedrig bemessenen Besoldung mit Nachzahlungsbedarf), ein positiver eine Überschätzung (vermeidbare Haushaltsbelastung). Die Fehler werden getrennt je Horizont berechnet und bilden die Grundlage der Modellvergleiche. Da der operative Prognosewert durch Verkettung der Raten am EU-SILC-Anker entsteht, kumulieren sich die Ratenfehler zudem zu einem Niveaufehler in Euro bzw. Prozent.

Symmetrische Gütemaße Über die n Evaluationspunkte eines Horizonts hinweg werden drei symmetrische Maße berechnet, wobei e_i die n Einzelfehler des betrachteten Horizonts indiziert:

$$\mathrm{ME} = \frac{1}{n}\sum_{i=1}^{n} e_i, \qquad \mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |e_i|, \qquad \mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2}. \tag{5}$$

Der mittlere Fehler (ME) misst die systematische Verzerrung. Ein Wert nahe null zeigt eine unverzerrte Prognose an, ein deutlich negativer oder positiver Wert eine systematische Unter- oder Überschätzung. Der mittlere absolute Fehler (MAE) beziffert die typische Fehlergröße unabhängig vom Vorzeichen. Die Quadratwurzel des mittleren quadratischen Fehlers (RMSE) bestraft große Abweichungen überproportional.

Asymmetrischer Verlust Symmetrische Fehlermaße behandeln Unter- und Überschätzung als gleichermaßen kostspielig. Sollen Über- und Unterschätzung unterschiedlich stark „bestraft“ werden, erhalten negative und positive Fehler im Verlustmaß unterschiedliche Gewichte (so genannter Lin-Lin-Verlust, vgl. z.B. Elliott und Timmermann, 2008):

$$L_{a:b}(e) \;=\; \begin{cases} a \cdot |e| & \text{falls } e < 0 \quad \text{(Unterschätzung)}, \\ b \cdot |e| & \text{falls } e \geq 0 \quad \text{(Überschätzung)}. \end{cases} \tag{6}$$

Im vorliegenden Bericht wird beispielhaft eine $2{:}1$-Gewichtung angewendet, d.h. eine Unterschätzung geht mit doppelter Stärke in das Fehlermaß ein wie eine Überschätzung.[9]

4 Ergebnisse der Modellevaluation

Die folgenden Ergebnisse beziehen sich auf die Prognosefehler gegenüber der amtlichen EU-SILC-Reihe unter Ausschluss des Bruchjahres 2020. Berichtet werden insgesamt fünf relevante Horizonte: für den Prognosezeitpunkt im Frühjahr die Echtzeitschätzungen für $t-2$ bis $t+0$ und die Vorausschätzung für $t+1$ sowie im Herbst die Echtzeitschätzungen für $t-1$ und $t+0$ und die Vorausschätzungen für $t+1$ und $t+2$.

4.1 Prognosegüte im Überblick

Die Modelle werden zunächst anhand der in Abschnitt 3.5 beschriebenen symmetrischen Fehlermaße bewertet. Neben mittlerem Fehler (ME), mittlerem absoluten Fehler (MAE) und der Quadratwurzel des mittleren quadrierten Fehlers (RMSE) wird der RMSE zusätzlich relativ zum Bezugsmodell des historischen Mittels ausgewiesen. Werte unter eins zeigen an, dass ein Modell den naiven Richtwert unterbietet. Abbildungen 1 und 2 zeigen die Ergebnisse für den Prognosezeitpunkt im Frühjahr (Horizonte $t-2$ bis $t+1$), Abbildungen 3 und 4 jene für den Herbst

[9] Die Festlegung des maßgeblichen Verhältnisses ist eine normative Entscheidung der Anwenderin; so können auch alternative Verhältnisse festgelegt werden, beispielsweise $1,5{:}1$ mit geringerer Gewichtung der Unterschätzung als im Beispiel oder $3{:}1$ mit stärkerer Gewichtung der Unterschätzung. Vor extremer Asymmetrie ist allerdings zu warnen, da systematisch überhöhte Planungswerte im Besoldungsgefüge nicht zurückgenommen werden.

(Horizonte $t-1$ bis $t+2$).[10] Zur Übersichtlichkeit wird nur eine Auswahl an Modellen gezeigt, der vollständige Modellvergleich findet sich in Anhang A.3, Abbildungen A.2 und A.3.

Abbildung 1: Prognosegüte am EU-SILC-Maßstab (ohne Bruchjahr 2020) – Frühjahr, Horizonte $t-2$ und $t-1$

Horizont t-2

Modell	n	ME	MAE	RMSE	rel.
Rate BLG pro Kopf (GD)	11	+0.45	1.47	1.87	0.71
Rate Verf. Eink. (GD)	11	+0.75	1.34	1.92	0.72
Brücke BLG pro Kopf (GD)	11	+0.06	1.67	2.19	0.83
Brücke Verf. Eink. (GD)	11	-0.08	1.69	2.21	0.83
Brücke Nettolöhne (GD)	11	+0.28	1.60	2.27	0.86
Rate Nettolöhne (GD)	11	+1.58	1.81	2.38	0.90
Historischer Mittelwert	11	-0.06	2.06	2.65	1.00
Log-Trend ab 2010	11	-0.49	2.15	2.70	1.02
ARIMA	11	+0.01	1.64	2.73	1.03
Log-Trend ab 1991	11	-0.67	2.37	2.75	1.04
Random Walk (Rate)	11	-0.35	2.38	3.55	1.34

Horizont t-1

n	ME	MAE	RMSE	rel.
10	+0.48	1.37	1.89	0.68
10	+0.63	1.30	1.90	0.68
10	-0.09	1.73	2.23	0.80
10	+0.02	1.57	2.15	0.77
10	+0.10	1.64	2.28	0.82
10	+1.60	1.85	2.33	0.84
10	-0.29	2.17	2.78	1.00
10	-0.79	2.61	3.06	1.10
10	-0.15	2.29	3.07	1.10
10	-0.88	2.51	2.88	1.04
10	-0.83	3.74	4.77	1.72

Anmerkungen: Fehlermaße in Prozentpunkten der Veränderungsrate. ME = Mittlerer Fehler. MAE = Mittlerer absoluter Fehler. RMSE = Quadratwurzel des mittleren quadrierten Fehlers. rel. = RMSE relativ zum Bezugsmodell (historischer Mittelwert), < 1: besser als Bezugsmodell. n = Zahl der Evaluationspunkte, definiert durch das Modell mit der geringsten Zahl an Evaluationspunkten.

Zwei Befunde bestätigen sich über alle Horizonte und beide Prognosezeitpunkte hinweg. Erstens ist das historische Mittel als Prognoseansatz bemerkenswert schwer zu schlagen. Sein RMSE bewegt sich über die Horizonte zwischen rund 2,7 und 3,4 Prozentpunkten, und sein mittlerer Fehler bleibt durchgehend nahe null, sodass er kaum systematisch verzerrt. Die langen Trendfortschreibungen und das ARIMA-Modell erreichen ähnliche Fehlerniveaus, unterschreiten den Richtwert über weite Strecken aber nicht. Zugleich fallen die Verfahren, die allein auf die kurzfristige Dynamik der Reihe setzen, klar zurück. Der Random Walk weist mit relativen RMSE-Werten zwischen 1,3 und 1,7 durchgehend die größten Fehler auf, und das ARIMA-Modell bleibt mit Werten um 1,0 bis 1,1 nahe an der Güte des Bezugsmodells, erreicht sie aber nicht ganz.

[10] Da die Frühjahrsgutachten der GD keine Angaben für das vorvergangene Jahr enthalten, wird für den Horizont $t-2$ der im vorangegangenen Herbstgutachten veröffentlichte Wert herangezogen. Zwischenzeitliche Revisionen der VGR bleiben dabei vereinfachend unberücksichtigt. Operativ kann für diesen Horizont alternativ auf den jeweils aktuellen Datenstand der VGR zurückgegriffen werden.

Abbildung 2: Prognosegüte am EU-SILC-Maßstab (ohne Bruchjahr 2020) – Frühjahr, Horizonte $t+0$ und $t+1$

Horizont t+0

Modell	n	ME	MAE	RMSE	rel.
Rate Verf. Eink. (GD)	9	+0.30	1.11	1.57	0.55
Rate BLG pro Kopf (GD)	9	+0.52	1.36	1.93	0.68
Brücke Verf. Eink. (GD)	9	-0.25	1.80	2.12	0.74
Rate Nettolöhne (GD)	9	+1.57	1.59	2.20	0.77
Brücke BLG pro Kopf (GD)	9	-0.09	1.89	2.37	0.83
Brücke Nettolöhne (GD)	9	+0.00	1.92	2.50	0.88
Historischer Mittelwert	9	-0.10	2.19	2.85	1.00
Log-Trend ab 1991	9	-0.69	2.50	2.91	1.02
ARIMA	9	+0.06	2.20	2.97	1.04
Log-Trend ab 2010	9	-0.68	2.67	3.02	1.06
Random Walk (Rate)	9	-0.93	3.69	4.66	1.64

Horizont t+1

n	ME	MAE	RMSE	rel.
8	+1.04	1.94	3.11	1.03
8	+0.58	2.48	3.32	1.10
8	-0.39	2.59	3.00	1.00
8	+1.62	1.96	3.34	1.11
8	-0.16	2.54	3.08	1.02
8	-0.48	2.59	3.00	1.00
8	-0.08	2.41	3.01	1.00
8	-0.67	2.69	3.07	1.02
8	+0.17	2.34	3.05	1.01
8	-0.72	2.88	3.17	1.05
8	-0.65	3.66	4.51	1.50

Anmerkungen: Fehlermaße in Prozentpunkten der Veränderungsrate. ME = Mittlerer Fehler. MAE = Mittlerer absoluter Fehler. RMSE = Quadratwurzel des mittleren quadrierten Fehlers. rel. = RMSE relativ zum Bezugsmodell (historischer Mittelwert), < 1: besser als Bezugsmodell. n = Zahl der Evaluationspunkte, definiert durch das Modell mit der geringsten Zahl an Evaluationspunkten.

Abbildung 3: Prognosegüte am EU-SILC-Maßstab (ohne Bruchjahr 2020) – Herbst, Horizonte $t-1$ und $t+0$

Horizont t-1

Modell	n	ME	MAE	RMSE	rel.
Rate BLG pro Kopf (GD)	11	+0.45	1.47	1.87	0.71
Rate Verf. Eink. (GD)	11	+0.75	1.34	1.92	0.72
Brücke BLG pro Kopf (GD)	11	+0.06	1.67	2.19	0.83
Brücke Verf. Eink. (GD)	11	-0.08	1.69	2.21	0.83
Brücke Nettolöhne (GD)	11	+0.28	1.60	2.27	0.86
Rate Nettolöhne (GD)	11	+1.58	1.81	2.38	0.90
Historischer Mittelwert	11	-0.06	2.06	2.65	1.00
Log-Trend ab 2010	11	-0.49	2.15	2.70	1.02
ARIMA	11	+0.01	1.64	2.73	1.03
Log-Trend ab 1991	11	-0.67	2.37	2.75	1.04
Random Walk (Rate)	11	-0.35	2.38	3.55	1.34

Horizont t+0

n	ME	MAE	RMSE	rel.
10	+0.15	1.29	1.66	0.60
10	+0.37	1.18	1.67	0.60
10	-0.34	1.84	2.20	0.79
10	-0.24	1.76	2.24	0.81
10	-0.08	1.77	2.30	0.83
10	+1.24	1.57	2.13	0.77
10	-0.29	2.17	2.78	1.00
10	-0.79	2.61	3.06	1.10
10	-0.15	2.29	3.07	1.10
10	-0.88	2.51	2.88	1.04
10	-0.83	3.74	4.77	1.72

Anmerkungen: Fehlermaße in Prozentpunkten der Veränderungsrate. ME = Mittlerer Fehler. MAE = Mittlerer absoluter Fehler. RMSE = Quadratwurzel des mittleren quadrierten Fehlers. rel. = RMSE relativ zum Bezugsmodell (historischer Mittelwert), < 1: besser als Bezugsmodell. n = Zahl der Evaluationspunkte, definiert durch das Modell mit der geringsten Zahl an Evaluationspunkten.

Abbildung 4: Prognosegüte am EU-SILC-Maßstab (ohne Bruchjahr 2020) – Herbst, Horizonte $t+1$ und $t+2$

Horizont t+1

Modell	n	ME	MAE	RMSE	rel.
Rate Verf. Eink. (GD)	9	+0.62	1.46	2.58	0.91
Brücke BLG pro Kopf (GD)	9	+0.10	1.95	2.60	0.91
Brücke Verf. Eink. (GD)	9	-0.32	2.13	2.65	0.93
Rate BLG pro Kopf (GD)	9	+0.90	1.80	2.67	0.94
Brücke Nettolöhne (GD)	9	+0.24	1.96	2.68	0.94
Historischer Mittelwert	9	-0.10	2.19	2.85	1.00
Log-Trend ab 1991	9	-0.69	2.50	2.91	1.02
Rate Nettolöhne (GD)	9	+1.59	1.89	2.97	1.04
ARIMA	9	+0.06	2.20	2.97	1.04
Log-Trend ab 2010	9	-0.68	2.67	3.02	1.06
Random Walk (Rate)	9	-0.93	3.69	4.66	1.64

Horizont t+2

n	ME	MAE	RMSE	rel.
6	+1.15	2.77	3.56	1.03
6	+0.26	2.69	3.31	0.96
6	+0.14	2.82	3.35	0.97
6	+0.94	2.78	3.46	1.01
6	+0.53	2.71	3.39	0.99
6	+0.17	2.95	3.44	1.00
6	-0.44	3.13	3.45	1.00
6	+1.82	2.87	3.84	1.12
6	+0.47	2.88	3.50	1.02
6	-0.11	2.99	3.35	0.97
6	+0.12	3.89	4.92	1.43

Anmerkungen: Fehlermaße in Prozentpunkten der Veränderungsrate. ME = Mittlerer Fehler. MAE = Mittlerer absoluter Fehler. RMSE = Quadratwurzel des mittleren quadrierten Fehlers. rel. = RMSE relativ zum Bezugsmodell (historischer Mittelwert), < 1: besser als Bezugsmodell. n = Zahl der Evaluationspunkte, definiert durch das Modell mit der geringsten Zahl an Evaluationspunkten.

Zweitens, der entscheidende Zugewinn entsteht durch die Einbindung der GD-Information. Über die kurzen Horizonte liefern die blanken Übertragungen der GD-Raten der verfügbaren Einkommen und der BLG pro Kopf die besten Prognosen. Bei $t-2$ und $t-1$ im Frühjahr sowie bei $t-1$ und $t+0$ im Herbst liegen beide im RMSE praktisch gleichauf (relative RMSE-Werte von 0,60 bis 0,72). Am aktuellen Rand im Frühjahr setzt sich die Rate der verfügbaren Einkommen dagegen deutlich ab: Bei $t+0$ sinkt ihr relativer RMSE auf 0,55, gegenüber 0,68 für die BLG pro Kopf. Das kommt einer nahezu halbierten Fehlergröße gegenüber dem historischen Mittel gleich, und der mittlere absolute Fehler geht von rund 2,2 auf etwa 1,1 Prozentpunkte zurück. Zudem erzielt die Rate der verfügbaren Einkommen an sämtlichen Horizonten (bis auf $t+2$ für den Herbst) den niedrigsten mittleren absoluten Fehler aller untersuchten Modelle.

Der spätere Prognosezeitpunkt verschafft den GD-Modellen zusätzliche Reichweite. Während im Frühjahr der Vorsprung bei $t+1$ bereits aufgezehrt ist und die GD-Modelle den Richtwert nur noch erreichen oder geringfügig überschreiten (relative RMSE-Werte zwischen 1,0 und 1,1), unterbieten sie ihn im Herbst auch bei $t+1$ noch, etwa mit 0,91 für die blanke Rate der verfügbaren Einkommen, 0,91 bis 0,94 für die geschätzten Brücken und 0,94 für die Rate der BLG pro

Kopf. Der Grund liegt in der aktuelleren GD-Information, die im Herbst ein Jahr weiter reicht, sodass sich der effektive Prognosehorizont verkürzt. Die auf der eigenen Historie beruhenden Verfahren ändern sich zwischen beiden Zeitpunkten dagegen kaum, sodass der Unterschied im Wesentlichen auf die GD-Modelle zurückgeht.

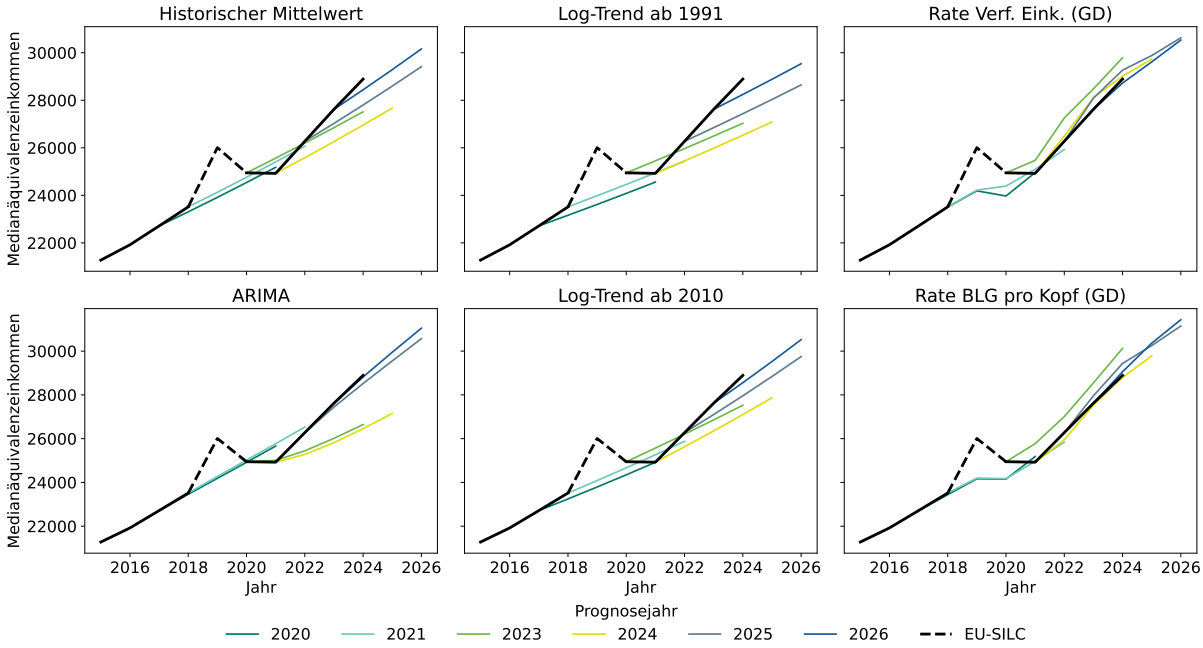
Zwischen den GD-Varianten zeigt sich ein systematisches Muster. An den kurzen Horizonten bis $t + 0$ bleiben die geschätzten Brücken der verfügbaren Einkommen und der BLG pro Kopf hinter den jeweiligen unadjustierten Raten zurück, bei $t + 0$ im Frühjahr etwa mit relativen RMSE-Werten von 0,74 gegenüber 0,55 für die verfügbaren Einkommen bzw. 0,83 gegenüber 0,68 für die BLG pro Kopf. Die GD-Prognosen sind dort bereits so eng mit der Dynamik des Medianäquivalenzeinkommens verbunden, dass eine zwischengeschaltete Regression auf kurzer Reihe eher Schätzrauschen hinzufügt als Information. Am längsten Horizont $t + 2$ kehrt sich das Bild jedoch um. Dort unterbieten die geschätzten Brücken (relative RMSE-Werte von 0,96 bis 0,99, angeführt von der Brücke der BLG pro Kopf) sowie der kurze Log-Trend das Bezugsmodell knapp, während die blanken Raten der verfügbaren Einkommen (1,03) und der BLG pro Kopf (1,01) zurückfallen. Dieser Befund beruht bei $t + 2$ allerdings auf lediglich sechs Evaluationspunkten und ist entsprechend vorsichtig zu interpretieren. Die direkte Übertragung der Nettolohnrate erreicht über alle Horizonte günstige absolute Fehler, überschätzt das Einkommenswachstum aber systematisch, was sich in einem durchgehend positiven mittleren Fehler zwischen rund 1,2 und 1,8 Prozentpunkten niederschlägt.

Ein Detail verdient mit Blick auf eine Verlustbetrachtung Beachtung. Die blanke Rate der verfügbaren Einkommen weist über alle Horizonte einen leicht positiven mittleren Fehler auf und überschätzt das Wachstum des Medianäquivalenzeinkommens im Durchschnitt also geringfügig. Zugleich bleibt ihr mittlerer absoluter Fehler auch dort am niedrigsten, wo ihr RMSE den Richtwert nicht mehr unterbietet, etwa bei $t+1$ im Frühjahr mit 1,94 Prozentpunkten bei einem relativen RMSE von 1,03. Einzelne große Abweichungen fallen bei diesem Verfahren also stärker ins Gewicht als die typische Fehlergröße vermuten lässt. Die Frage, wie sich die leichte Überschätzung auswirkt, wenn Unterschätzungen stärker gewichtet werden, wird in Abschnitt 4.3 wieder aufgegriffen.

4.2 Fehler im Einkommensniveau

Für die praktische Anwendung ist nicht allein die Prognosegüte der jährlichen Veränderungsraten entscheidend. Maßgeblich ist vielmehr, wie stark sich Prognosefehler im fortgeschriebenen Einkommensniveau niederschlagen. Hierzu werden die prognostizierten Raten auf den jeweils letzten verfügbaren Wert aufgerechnet. Die so verketteten Prognosepfade lassen sich der realisierten EU-SILC-Reihe gegenüberstellen (Abbildung 5). Gezeigt werden sechs ausgewählte Modelle, deren Pfade die Erkenntnisse aus dem Ratenvergleich nun im Niveau veranschaulichen.

Abbildung 5: Verkettete Prognosepfade ausgewählter Modelle und realisierte EU-SILC-Reihe – Frühjahr



Anmerkungen: Realisiertes Niveau (schwarze Linie) und prognostizierter Verlauf des Medianäquivalenzeinkommens für verschiedene Prognosezeitpunkte (farbige Linien) für ausgewählte Modelle. Der prognostizierte Verlauf ergibt sich jeweils als (letzter beobachteter Wert $\times$ prognostizierte Raten).

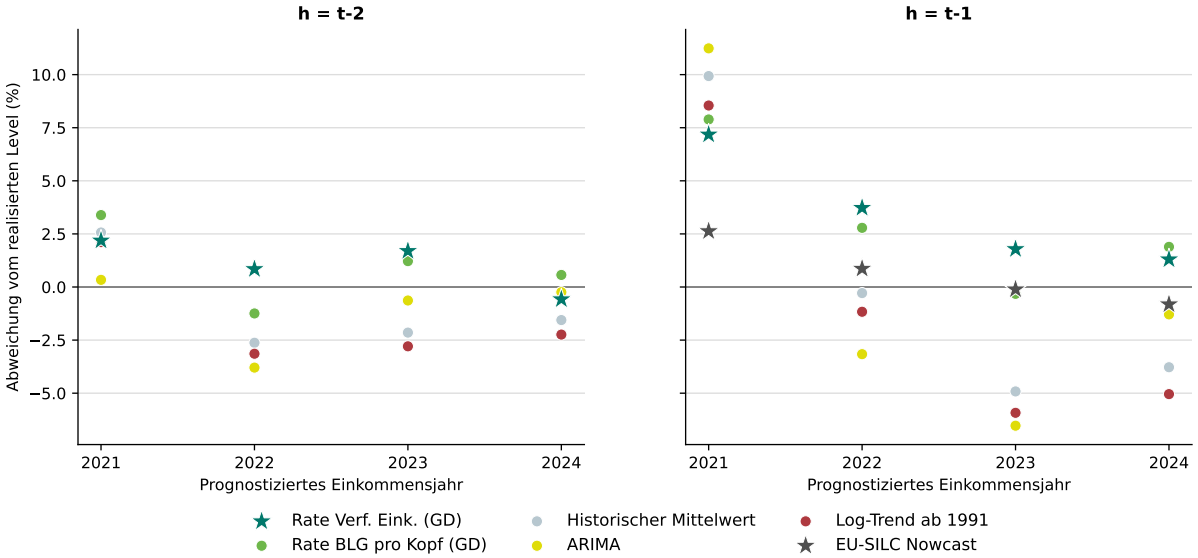
Die verketteten Prognosepfade bestätigen zunächst die Ergebnisse der Bewertung auf Basis der Veränderungsraten. Die Verfahren unterscheiden sich weniger durch einzelne große Prognosefehler als vielmehr durch ihre Fähigkeit, den mittelfristigen Wachstumspfad des Medianäquivalenzeinkommens nachzuzeichnen. Besonders deutlich wird dies in den Jahren 2022 bis 2024, in denen das nominale Medianäquivalenzeinkommen infolge der außergewöhnlich hohen Inflation erheblich stärker zunahm als in den vorangegangenen Jahren. Während das historische

Mittel und langfristige Trendmodelle diesen Anstieg nur verzögert nachvollziehen, folgen die auf den GD-Prognosen basierenden Verfahren dem tatsächlichen Verlauf deutlich enger.[11]
Auch hinsichtlich der Verzerrung ergeben sich Unterschiede zwischen den Modellklassen. Die direkten GD-Prognosen der verfügbaren Einkommen und der BLG pro Kopf überschätzen die tatsächliche Einkommensentwicklung zwar im Mittel leicht, bleiben jedoch über nahezu alle Horizonte hinweg näher an der realisierten Entwicklung als die übrigen Verfahren. Demgegenüber unterschätzen historische Mittelwerte und langfristige Trendfortschreibungen insbesondere in den Jahren des außergewöhnlich starken Einkommenswachstums den tatsächlichen Verlauf deutlich. Diese Unterschätzung ist weniger auf einzelne Ausreißer als vielmehr auf die geringe Anpassungsgeschwindigkeit langfristiger Verfahren an veränderte makroökonomische Rahmenbedingungen zurückzuführen. Die Pfade des ARIMA-Modells fächern dagegen deutlich stärker auf, weil sie den jeweils zuletzt beobachteten Ausschlag der Rate in die Zukunft verlängern und dadurch von Prognoseursprung zu Prognoseursprung stärker schwanken.
Für den Prognosezeitpunkt im Herbst ergibt sich ein ähnliches Bild (Anhang A.4, Abbildung A.4). Auch dort begleiten die glatten Verfahren die realisierte Reihe im Trendtempo und bleiben ab 2023 unter ihr, während die GD-gestützten Modelle den jüngsten Anstieg am engsten nachzeichnen. Die Unterschiede zum Frühjahr gehen auf die neueren Informationen im Herbst zurück, die nun ein Jahr weiter reichen, sodass die GD-Pfade dem realisierten Verlauf bis in das kommende Jahr hinein noch enger folgen.
Insgesamt bestätigt die Betrachtung des Einkommensniveaus somit die Schlussfolgerungen aus der Evaluation der Veränderungsraten. Für kurzfristige Prognosen liefern die makroökonomischen Prognosen der verfügbaren Einkommen und der BLG pro Kopf nicht nur die geringsten Fehler in den Jahresraten, sondern auch die realistischsten Fortschreibungen des gesamten Einkommenspfades. Gleichzeitig zeigt sich, dass die Unterschiede zwischen den Modellklassen im Niveau anschaulicher hervortreten als bei der isolierten Betrachtung einzelner Veränderungsraten und deshalb für die praktische Bewertung der Prognoseverfahren besonders aufschlussreich sind.

[11] Die Niveaufehler verdeutlichen zudem, dass sich selbst vergleichsweise kleine Unterschiede in den jährlichen Veränderungsraten durch die Verkettung über mehrere Jahre merklich aufsummieren können. Besonders bei längeren Prognosehorizonten entstehen dadurch systematische Abweichungen im Einkommensniveau, obwohl sich die Fehler einzelner Jahresraten jeweils in einem moderaten Bereich bewegen.

Zusätzlich lässt sich die Abweichung der verketteten Niveaus vom realisierten Wert in Prozent für die einzelnen Modelle und Prognosejahre auftragen (Abbildungen 6 und 7). Ein positiver Wert steht für eine Überschätzung des Einkommensniveaus, ein negativer für eine Unterschätzung.

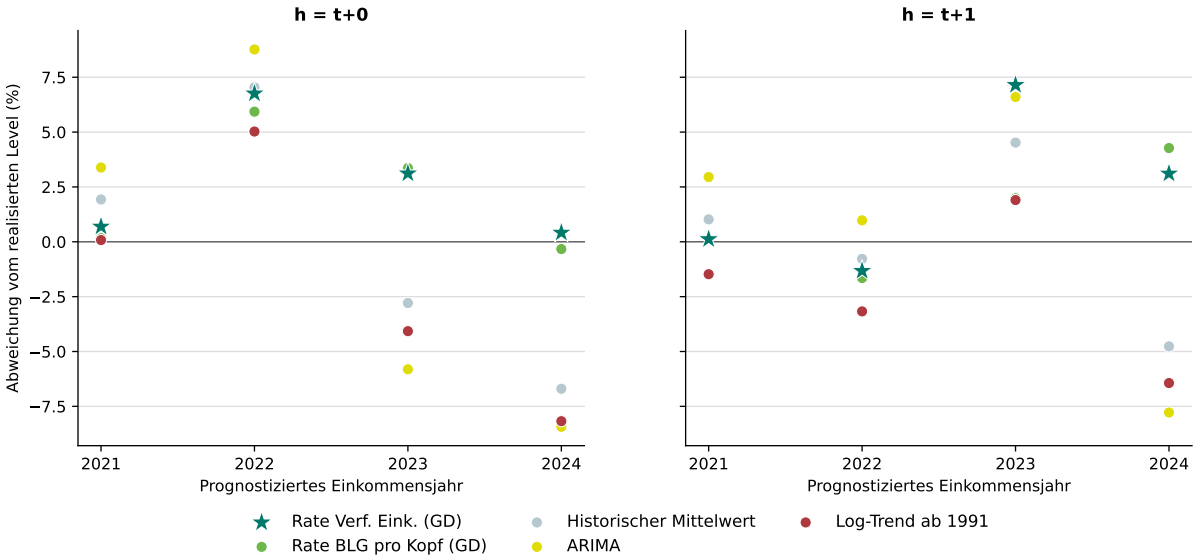
Abbildung 6: Fehler der verketteten Niveauprognosen – Frühjahr I/II



Anmerkungen: Abweichung der verketteten Niveauprognosen (Anker-Level × prognostizierte Raten) vom realisierten Wert für verschiedene Modelle, Prognosehorizonte und Jahre. J bezeichnet das prognostizierte Einkommensjahr, Anker ist jeweils der letzte verfügbare EU-SILC-Wert. Links ($h = t - 2$): Prognosezeitpunkt im Frühjahr des Jahres $J + 2$, Anker = Einkommensjahr $J - 1$. Rechts ($h = t - 1$): Prognosezeitpunkt im Frühjahr des Jahres $J + 1$, Anker = Einkommensjahr $J - 2$. Fehler > 0 = Überschätzung.

Am Rückschätzhorizont $h = t - 2$ bleiben die Fehler klein und die Modelle liegen nah beieinander, meist innerhalb von etwa drei Prozent. Die Prognosejahre 2021 (für $h = t - 1$) und 2022 (für $h = t + 0$) stechen zudem heraus: Hier überzeichnen sämtliche Modelle das realisierte Niveau um jeweils fünf bis elf Prozent. Diese gleichgerichtete Verfehlung geht weniger auf die Modelle als auf den Maßstab zurück – die betroffenen Pfade verketten über das bruchverzerrte Ankerniveau hinweg und erben dessen Überzeichnung. In den Jahren des kräftigen Einkommenswachstums 2023 und 2024 kehrt sich das Bild um. Die auf der eigenen Historie beruhenden Verfahren unterschätzen das Niveau nun um rund zweieinhalb bis acht Prozent, weil das nominale Einkommenswachstum dieser Jahre deutlich über dem langjährigen Trendtempo lag. Einzig die direkten Übertragungen der GD-Raten der verfügbaren Einkommen und der BLG pro Kopf bleiben für $h = t - 1$ und $h = t + 0$ in beiden Jahren nahe am realisierten Wert und

Abbildung 7: Fehler der verketteten Niveauprognosen – Frühjahr II/II



Anmerkungen: Abweichung der verketteten Niveauprognosen (Anker-Level $\times$ prognostizierte Raten) vom realisierten Wert für verschiedene Modelle, Prognosehorizonte und Jahre. J bezeichnet das prognostizierte Einkommensjahr, Anker ist jeweils der letzte verfügbare EU-SILC-Wert. Links ($h = t + 0$): Prognosezeitpunkt im Frühjahr des Jahres J, Anker = Einkommensjahr $J - 3$. Rechts ($h = t + 1$): Prognosezeitpunkt im Frühjahr des Jahres $J - 1$, Anker = Einkommensjahr $J - 4$. Fehler > 0 = Überschätzung.

im Wesentlichen innerhalb von drei Prozent; beim Prognosehorizont $h = t + 1$ überschätzen jedoch beide das Einkommensniveau für 2023 und 2024 deutlich.

Für den Herbst ergibt sich ein ähnliches Bild; insbesondere die GD-gestützten Verfahren liegen dort aufgrund der aktuelleren Informationen am jeweiligen Horizont meist näher am realisierten Niveau (Anhang A.4, Abbildungen A.5 und A.6).

Für den Horizont $h = t - 1$ steht mit dem EU-SILC-Nowcast zusätzlich eine externe Schätzung des noch unveröffentlichten Einkommensjahres zur Verfügung (Eurostat, 2025; Dany-Knedlik, Hüttl und Pagenhardt, 2026, Abschnitt 3.2.1). Er trifft das Medianäquivalenzeinkommen des zuletzt abgeschlossenen Einkommensjahres besonders gut und stellt die genaueste verfügbare Schätzung des Ausgangsniveaus dar, sodass es naheliegt, ihn als Anker der Verkettung heranzuziehen.

4.3 Bewertung unter asymmetrischem Verlust

Die bisherigen Modellvergleiche beruhen auf symmetrischen Fehlermaßen, bei denen Unter- und Überschätzungen gleichermaßen in die Bewertung eingehen. Im Kontext der Besoldungsgesetzgebung erscheint diese Annahme jedoch nicht zwingend angemessen. Während eine Unterschätzung des Medianäquivalenzeinkommens zu einer zu niedrigen Bemessung der Mindestbesoldung und damit zum fiskalischen Risiko späterer Nachzahlungen führen kann, verursacht eine Überschätzung zunächst höhere Haushaltsansätze. Vor diesem Hintergrund werden die Modellränge unter dem symmetrischen MAE denen unter einem Lin-Lin-Verlust im Szenario $2{:}1$ gegenübergestellt (Abbildungen 8 und 9).

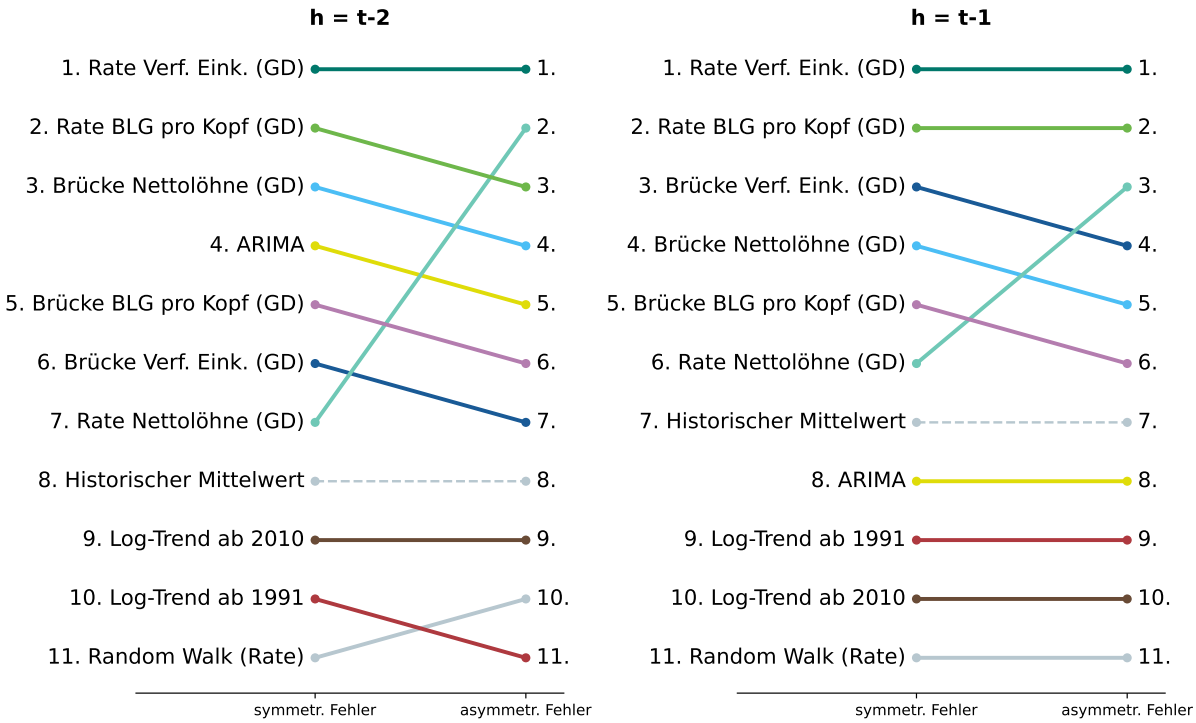
Ein Wechsel der Verlustfunktion verschiebt die Rangfolge nicht zufällig, sondern nach einem klaren Prinzip. Da eine Unterschätzung in diesem Fall doppelt gewichtet wird, verbessern sich Modelle, die das Einkommenswachstum eher zu hoch ansetzen, während Verfahren mit systematischer Unterschätzung zurückfallen.

Am deutlichsten zeigt sich dieses Muster bei der blanken Übertragung der GD-Nettolohnrate: Sie weist über alle Horizonte einen ausgeprägt positiven mittleren Fehler von rund 1,2 bis 1,8 Prozentpunkten auf und verbessert sich unter asymmetrischer Bewertung spürbar, bis hin zum ersten Rang beim Horizont $h = t+1$. Auf der anderen Seite verlieren die glatten Verfahren, die das jüngste Wachstum untertreiben. Die Log-Trends und das historische Mittel rücken unter der asymmetrischen Gewichtung nach teilweise hinten, weil ihre negativen Fehler nun doppelt zählen. Der Random Walk bleibt an allen Horizonten, bis auf $h = t - 2$, das schwächste Verfahren.

Das primär empfohlene Modell, die blanke Übertragung der GD-Rate der verfügbaren Einkommen, behauptet dagegen an den kurzen und mittleren Horizonten ($h = t$–2 bis $h = t + 0$) den ersten Rang unter beiden Verlustfunktionen. Da dieses Modell selbst einen leicht positiven mittleren Fehler aufweist, wird es durch die asymmetrische Gewichtung nicht bestraft, sondern in seiner führenden Position bestätigt. Die direkte Übertragung der Prognose der verfügbaren Einkommen bleibt somit sowohl unter symmetrischer als auch unter stärker planungsorientierten asymmetrischen Bewertungskriterien das insgesamt leistungsstärkste Verfahren.[12]

[12] Gleichzeitig verdeutlicht die Analyse einen grundsätzlichen Unterschied zwischen Prognose und Planung. So ist unter Lin-Lin-Verlust nicht der Median der Fehlerverteilung die optimale Punktprognose, sondern das $a/(a+b)$-

Abbildung 8: Rangverschiebung zwischen symmetrischer und asymmetrischer Bewertung – Frühjahr I/II

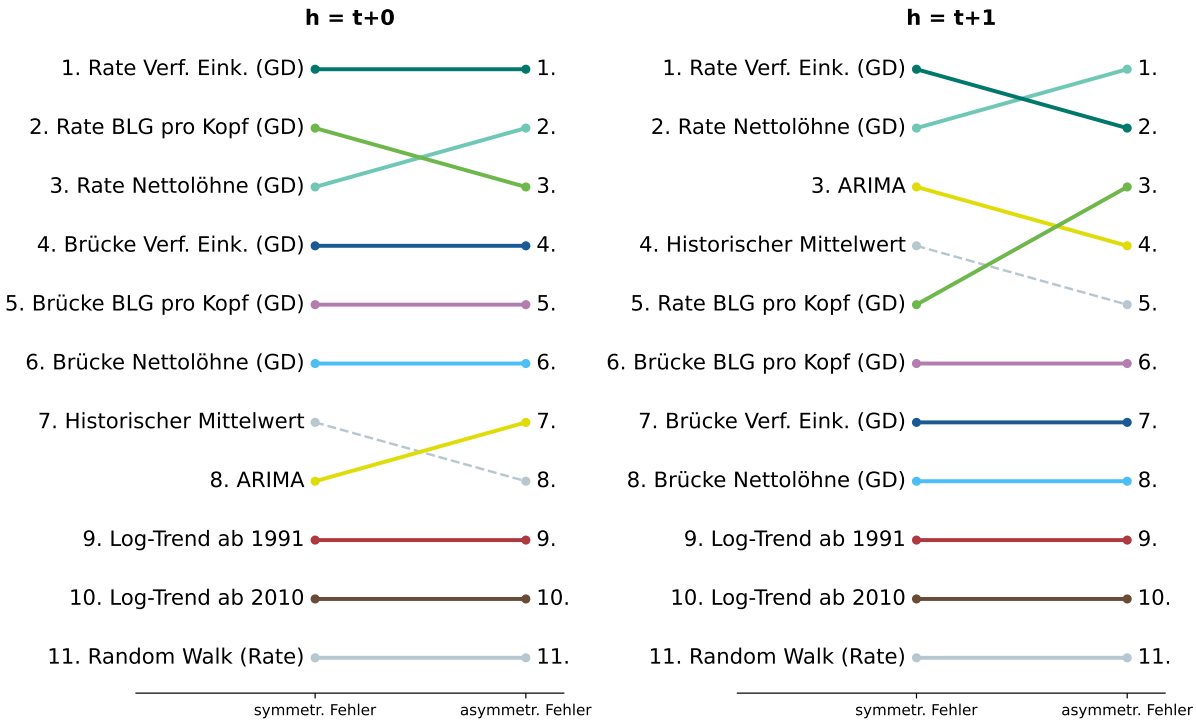


Anmerkungen: Rang jedes Modells nach mittlerem absolutem Fehler (MAE) unter symmetrischer Bewertung (links) gegenüber dem Rang unter Lin-Lin-Verlust im Szenario 2:1 (rechts). Rang 1 = geringster (gewichteter) absoluter Fehler.

Für den Prognosezeitpunkt im Herbst zeigt sich ein ähnliches Muster (Anhang A.4, Abbildungen A.7 und A.8). Während bei $h = t + 0$ und $h = t + 1$ die Rangfolge weitgehend stabil bleibt, ordnet sie sich beim längsten Horizont $h = t + 2$ spürbar um, weil die Fehler größer werden und das Vorzeichen des mittleren Fehlers stärker durchschlägt. Dabei profitiert insbesondere die blanke Übertragung der Nettolohnrate von einer asymmetrischen Verlustfunktion, während die geschätzten Brückenmodelle zurückfallen.

Quantil der Verteilung der Zielgröße, bei $2{:}1$ also das 67-Prozent-Quantil, sodass Zentralprognose und Planungswert voneinander abweichen; der Planungswert ergäbe sich als Aufschlag aus dem gewählten Quantil der Fehlerverteilung.

Abbildung 9: Rangverschiebung zwischen symmetrischer und asymmetrischer Bewertung – Frühjahr II/II



Anmerkungen: Rang jedes Modells nach mittlerem absolutem Fehler (MAE) unter symmetrischer Bewertung (links) gegenüber dem Rang unter Lin-Lin-Verlust im Szenario 2:1 (rechts). Rang 1 = geringster (gewichteter) absoluter Fehler.

5 Schlussfolgerungen und Handlungsempfehlungen

Die vorliegende Kurzexpertise liefert eine systematische empirische Bewertung verschiedener Prognoseverfahren für das Medianäquivalenzeinkommen. Sämtliche Modelle wurden unter einheitlichen Pseudo-Echtzeit-Bedingungen geschätzt und anhand ihrer Fähigkeit bewertet, die später veröffentlichten amtlichen EU-SILC-Daten möglichst genau vorherzusagen. Aus den Ergebnissen lassen sich fünf Schlussfolgerungen sowie darauf aufbauende Handlungsempfehlungen für den operativen Einsatz ableiten.

Schlussfolgerungen

1. Für kurze Prognosehorizonte liefert die direkte Übertragung der GD-Prognosen der verfügbaren Einkommen die höchste Prognosegüte. Die blanke Übertragung der Veränderungsra-

te der Prognosen der verfügbaren Einkommen erzielt an den kurzen Prognosehorizonten – im Frühjahr bis $h = t + 0$, im Herbst sogar bis $h = t + 1$ – insgesamt die beste Prognosegüte. Einzig die ebenfalls unadjustiert übertragene Rate der BLG pro Kopf erreicht an den kürzesten Horizonten vergleichbare Prognosefehler; am aktuellen Rand im Frühjahr fällt sie jedoch deutlich zurück. Mit einer Reduktion des RMSE gegenüber dem historischen Mittel um bis zu rund 45 Prozent schlägt die Rate des verfügbaren Einkommens alle geschätzten Alternativen, einschließlich der aufwendigeren Brückenregressionen. Der Zusammenhang zwischen der prognostizierten Größe und dem Medianäquivalenzeinkommen ist bereits so eng, dass die zusätzliche Schätzung von Achsenabschnitt und Steigung mehr Rauschen als zusätzlichen Informationsgewinn liefert. Zusätzliche Modellkomplexität führt somit nicht automatisch zu besseren Prognosen.

2. Einfache historienbasierte Verfahren und Brückenmodelle erweisen sich vor allem an langen Prognosehorizonten als leistungsfähig. Dabei bildet insbesondere das historische Mittel einen bemerkenswert treffsicheren und zugleich einfachen Referenzmaßstab. Weder langfristige Trendfortschreibungen noch ARIMA-Modelle erzielen über die meisten Horizonte eine systematisch höhere Prognosegüte. Am längsten Schätzhorizont $h = t + 2$ sind auch die aufwendigeren Brückenmodelle konkurrenzfähig und unterbieten den RMSE des historischen Mittels leicht. Gleichzeitig zeigt sich, dass Verfahren, die ausschließlich auf kurzfristiger Ratendynamik beruhen, insbesondere der Random Walk, kaum verwertbares Signal übertragen, sondern primär Erhebungsrauschen hinzufügen.

3. Die Ergebnisse bleiben auch unter außergewöhnlichen konjunkturellen Rahmenbedingungen weitgehend stabil. Der starke nominale Einkommensanstieg in den Jahren 2023 und 2024 infolge der Hochinflationsphase stellt einen anspruchsvollen Test für die untersuchten Verfahren dar. Während historienbasierte Modelle diesen nur verzögert nachvollziehen und deshalb deutliche Unterschätzungen aufweisen, bilden die GD-Prognosen der verfügbaren Einkommen und der BLG pro Kopf den Anstieg wesentlich genauer ab. Die außergewöhnlichen Abweichungen der Jahre 2021 und 2022 sind dagegen überwiegend auf den beschriebenen Strukturbruch der EU-SILC-Reihe zurückzuführen und nicht als Versagen einzelner Prognose-

verfahren zu interpretieren. Insgesamt bestätigt sich damit die Robustheit der Hauptergebnisse auch unter außergewöhnlichen makroökonomischen Rahmenbedingungen.

4. Die grundsätzliche Rangfolge der Modelle bleibt unter asymmetrischer Verlustfunktion im Szenario 2:1 weitgehend erhalten. Werden Unterschätzungen des Medianäquivalenzeinkommens stärker gewichtet als Überschätzungen, verschieben sich einzelne Modellränge zugunsten leicht optimistischer Verfahren. Sowohl im Frühjahr als auch im Herbst bleibt die blanke Übertragung der Rate der verfügbaren Einkommen an den kurzen Horizonten (im Frühjahr bis $h = t + 0$, im Herbst sogar bis $h = t + 1$) das stärkste Verfahren.

5. Für längere Prognosehorizonte und für Übertragungen auf Länderebene besteht weiterer Forschungsbedarf. Die Aussagekraft der Evaluation nimmt mit zunehmendem Prognosehorizont naturgemäß ab. Da für ein drittes Prognosejahr keine Prognosen der GD zur Verfügung stehen, können dafür ausschließlich univariate Verfahren eingesetzt werden. Hier erscheinen insbesondere das historische Mittel und langfristige Trendfortschreibungen als geeignete Ausgangspunkte, die empirische Evidenz beruht jedoch auf einer vergleichsweise kleinen Zahl von Evaluationspunkten. Auch eine Übertragung der hier entwickelten Verfahren auf Länderebene erscheint grundsätzlich möglich, setzt jedoch längere und ausreichend stabile regionale Zeitreihen voraus. Da die makroökonomischen Prognosen der GD ausschließlich auf nationaler Ebene vorliegen und die verfügbaren Stichprobenumfänge des SOEP in einzelnen Ländern begrenzt sind, ist die Übertragung auf regionale Anwendungen sorgfältig abzuwägen.

Handlungsempfehlungen

Aus den Ergebnissen ergibt sich eine klare Handlungsempfehlung. Für kurzfristige Prognosen des Medianäquivalenzeinkommens sollte die direkte Übertragung der Prognosen der verfügbaren Einkommen aus der GD als Primärmodell verwendet werden. Dieses Verfahren liefert an den kurzen Prognosehorizonten (im Frühjahr bis $h = t + 0$, im Herbst bis $h = t + 1$) die höchste Prognosegüte und erfüllt gleichzeitig die eingangs formulierte Zielsetzung eines einfachen, nachvollziehbaren und reproduzierbaren Prognoseinstruments. Die direkte Übertragung der BLG-pro-Kopf-Rate sollte zudem als makroökonomisches Vergleichsmodell stets mitgeführt

werden. Sie erzielt insbesondere an den kürzesten Horizonten eine nahezu vergleichbare Prognosegüte und ermöglicht damit eine zusätzliche Plausibilisierung der auf den verfügbaren Einkommen beruhenden Prognose. Das historische Mittel dient zudem als Plausibilitätsanker und bietet sich insbesondere an langen Prognosehorizonten, die mitunter über die Reichweite der GD hinausgehen, als transparente und robuste Alternative an. Zusätzlich zu den Schätzungen der GD könnten auch die Projektionen der Bundesregierung, die einem strengen, rechtlich vorgegeben Befürwortungsverfahren durch die GD unterliegen und als „statistisch unverzerrt" gelten dürften, als Fortschreibungsvariable zugrunde gelegt werden.

Für den operativen Einsatz ergibt sich daraus eine einfache Prognosestrategie. Mittels des horizontabhängigen Verfahrens sollten zunächst die Veränderungsraten des Medianäquivalenzeinkommens prognostiziert werden. Für die Übertragung auf das Einkommensniveau werden die Raten anschließend durch Verkettung auf den letzten verfügbaren amtlichen EU-SILC-Wert aufgerechnet. Als Anker für die Verkettung für das zuletzt abgeschlossene Einkommensjahr bietet sich insbesondere der EU-SILC-Nowcast von Eurostat an, da er die genaueste verfügbare Schätzung des Ausgangsniveaus darstellt. Das festgelegte Primärmodell wird regelmäßig gegen alternative Modelle evaluiert. Ein Modellwechsel erfolgt erst, wenn ein alternatives Verfahren über mehrere Aktualisierungen hinweg einen substanziellen und robusten Prognosevorteil zeigt.

Literatur

Dany-Knedlik, Geraldine, Pia Hüttl und Laura Pagenhardt (2026). *Bestimmungsfaktoren und Prognostizierbarkeit des Medianäquivalenzeinkommens*. Politikberatung kompakt 215.

Diebold, Francis X und Jose A Lopez (1996). „Forecast evaluation and combination“. *Handbook of statistics* 14, S. 241–268.

Elliott, Graham und Allan Timmermann (2008). „Economic Forecasting“. *Journal of Economic Literature* 46, S. 3–56.

Eurostat (2025). *Flash estimates of income inequalities and poverty indicators for 2024 (FE 2024): Methodological note*.

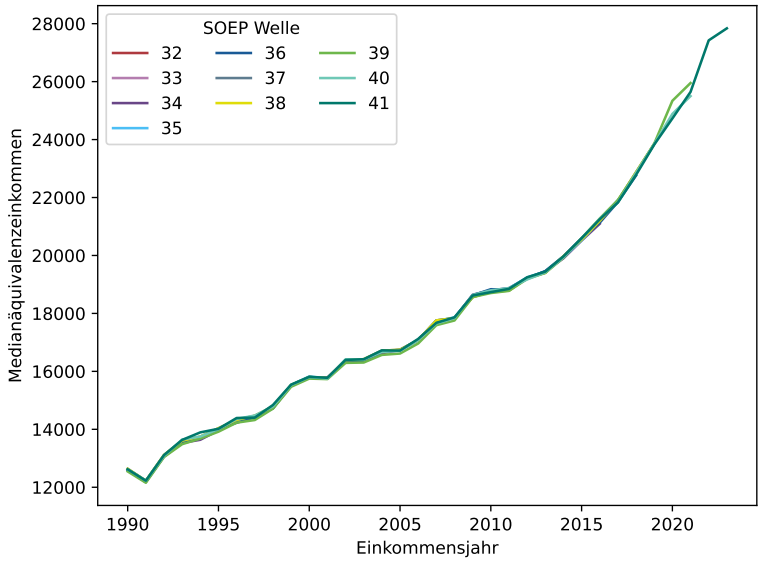
West, KD u. a. (2006). „Handbook of economic forecasting“. *Forecast Evaluation*, S. 101–133.

A Anhang

A.1 Zusätzliche Informationen zu den Schätzbedingungen

Revisionen. Im Pseudo-Echtzeit-Design (Abschnitt 2.2) wird auf die Einkommensreihe aus dem SOEP basierend auf dem aktuellen Datenstand (v41) zurückgegriffen. Streng genommen wäre hier eine Echtzeit-Betrachtung angezeigt, denn jede neue SOEP-Welle liefert die gesamte Einkommenshistorie neu aus und revidiert dabei rückwirkend Hochrechnungsfaktoren und generierte Einkommensvariablen. Für das Medianäquivalenzeinkommen sind diese Revisionen jedoch praktisch bedeutungslos: Die aus den Wellen 32 bis 41 berechneten Medianäquivalenzeinkommen sind über den gesamten Zeitraum ab 1991 nahezu deckungsgleich und im Schaubild kaum voneinander zu unterscheiden (Abbildung A.1); nennenswerte Revisionen treten weder am aktuellen Rand noch in der Historie auf. Eine neue Welle verlängert die Reihe somit im Wesentlichen nur um ein weiteres Einkommensjahr, ohne die bereits vorliegenden Werte substanziell zu verändern. Gleiches gilt für die EU-SILC-Reihe, deren veröffentlichte Werte nachträglich nur im Ausnahmefall – etwa bei methodischen Anpassungen oder Korrekturen – revidiert werden.

Abbildung A.1: Medianäquivalenzeinkommen verschiedener SOEP-Wellen



Anmerkung: Nominales Medianäquivalenzeinkommen in Euro nach Einkommensjahr, jeweils berechnet mit dem Datenstand der SOEP-Wellen 32 bis 41 (Umfragejahre 2015 bis 2024).

Stationarität. Laut Augmented-Dickey-Fuller-Tests (ADF, Tabelle A.1) kann die Nullhypothese einer Einheitswurzel für die Einkommensniveaus beider Quellen nicht verworfen werden – die Reihen sind trendbehaftet und damit nicht stationär. Für die SOEP-Veränderungsrate wird die Einheitswurzel hingegen klar verworfen; sie erfüllt damit die Voraussetzung für die Zeitreihenmodellierung.

Bemerkenswert ist, dass dies für die EU-SILC-Veränderungsrate nicht gilt: Hier kann die Einheitswurzel nicht verworfen werden. Dieses Ergebnis ist angesichts von lediglich zehn verwertbaren Beobachtungen und des Strukturbruchs mit Vorsicht zu interpretieren – die Teststärke ist in derart kurzen Reihen gering. Gerade darin liegt aber der Kern des Befundes: Die EU-SILC-Reihe ist zu kurz, um ihre statistischen Eigenschaften überhaupt verlässlich zu bestimmen, und liefert damit keine tragfähige Basis für die Modellschätzung. Die Stationaritätsergebnisse stützen die Entscheidung, die Modelle anhand der SOEP-Reihe zu schätzen.

Tabelle A.1: Augmented-Dickey-Fuller-Tests: Niveaus und Veränderungsraten

	ADF-Stat.	p-Wert	verz.	Beob.	1 %-krit. Wert	5 %-krit. Wert	stationär
SOEP (€)	2,33	0,99	1	38	$-3{,}62$	$-2{,}94$	nein
EU-SILC (€)	2,09	0,99	7	11	$-4{,}22$	$-3{,}19$	nein
SOEP (%)	$-3{,}50$	0,01	2	36	$-3{,}63$	$-2{,}95$	ja
EU-SILC (%)	$-1{,}17$	0,69	7	10	$-4{,}33$	$-3{,}23$	nein

Anmerkungen: (€) = Niveaus in Euro, (%) = Veränderungsraten. Wahl der Verzögerungen (verz.) mittels Akaike-Informations-Kriterium. EU-SILC ohne Bruchjahr (Umfragejahr 2020).

A.2 Gewählte ARIMA-Ordnungen

Abschnitt 3.3 beschreibt die Verwendung von ARIMA-Modellen als univariate Modellvarianten, die die kurzfristige Dynamik der Einkommensentwicklung abbilden können. Dabei wird die Ordnung der ARIMA-Terme für jeden Prognosezeitpunkt neu bestimmt. Tabelle A.2 gibt die jeweils selektierten Modell-Ordnungen.

Tabelle A.2: ARIMA(p, d, q)-Ordnungen pro Modell und Prognosezeitpunkt

Prognosezeitpunkt	2011–2023	2024	2025	2026
Anker	2009–2021	2022	2023	2024
Level	(0,1,0)	(0,1,3)	(1,1,0)	(0,1,3)
Rate	(1,0,0)	(1,0,0)	(1,0,0)	(0,0,3)

Anmerkungen: (p, d, q)-Ordnungen für ARIMA-Schätzungen, gewählt anhand Akaike-Informations-Kriterium (AIC). Die Integrationsordnung d ist festgesetzt auf $d = 1$ für die Levelschätzung und $d = 0$ für die Schätzung in Raten.

A.3 Ausführliche Schätzergebnisse: Prognosegüte am EU-SILC-Maßstab

Aus Gründen der Übersichtlichkeit beschränken sich Abbildungen 1 bis 4 in Abschnitt 4 auf eine Auswahl der untersuchten Modelle. Darüber hinaus wurden weitere Modellvarianten geschätzt (vgl. Abschnitt 3): zusätzliche Log-Trends mit Beginn im Jahr 2000 bzw. 2005, ein ARIMA-Modell auf dem logarithmierten Einkommensniveau, um die verzögerte abhängige Variable erweiterte Brückenmodelle (Brücke-AR) sowie Varianten auf Basis der aggregierten Bruttolöhne und -gehälter. Die Abbildungen A.2 und A.3 weisen die Ergebnisvergleiche inklusive dieser Modelle für die Prognosezeitpunkte im Frühjahr und im Herbst aus.

Abbildung A.2: Prognosegüte am EU-SILC-Maßstab (ohne Bruchjahr 2020), alle Modelle – Frühjahr

	Horizont t-2					Horizont t-1					Horizont t+0					Horizont t+1				
Modell	n	ME	MAE	RMSE	rel.	n	ME	MAE	RMSE	rel.	n	ME	MAE	RMSE	rel.	n	ME	MAE	RMSE	rel.
Rate BLG pro Kopf (GD)	11	+0.45	1.47	1.87	0.71	10	+0.48	1.37	1.89	0.68	9	+0.52	1.36	1.93	0.68	8	+0.58	2.48	3.32	1.10
Rate Verf. Eink. (GD)	11	+0.75	1.34	1.92	0.72	10	+0.63	1.30	1.90	0.68	9	+0.30	1.11	1.57	0.55	8	+1.04	1.94	3.11	1.03
Rate BLG (GD)	11	+1.46	1.62	1.95	0.74	10	+1.43	1.62	1.92	0.69	9	+1.51	1.52	1.98	0.69	8	+1.79	1.95	3.27	1.09
Brücke BLG (GD)	11	+0.26	1.50	2.09	0.79	10	+0.11	1.52	2.13	0.77	9	+0.04	1.83	2.41	0.85	8	-0.05	2.40	2.91	0.97
Brücke BLG pro Kopf (G	11	+0.06	1.67	2.19	0.83	10	-0.09	1.73	2.23	0.80	9	-0.09	1.89	2.37	0.83	8	-0.16	2.54	3.08	1.02
Brücke Verf. Eink. (GD)	11	-0.08	1.69	2.21	0.83	10	+0.02	1.57	2.15	0.77	9	-0.25	1.80	2.12	0.74	8	-0.39	2.59	3.00	1.00
Brücke-AR BLG	11	+0.07	1.58	2.27	0.86	10	-0.09	1.74	2.33	0.84	9	-0.15	2.02	2.55	0.89	8	-0.22	2.52	2.95	0.98
Brücke Nettolöhne (GD)	11	+0.28	1.60	2.27	0.86	10	+0.10	1.64	2.28	0.82	9	+0.00	1.92	2.50	0.88	8	-0.48	2.59	3.00	1.00
Brücke-AR Nettolöhne	11	+0.06	1.60	2.38	0.90	10	-0.12	1.87	2.48	0.89	9	-0.16	2.08	2.61	0.92	8	-0.71	2.76	3.08	1.02
Rate Nettolöhne (GD)	11	+1.58	1.81	2.38	0.90	10	+1.60	1.85	2.33	0.84	9	+1.57	1.59	2.20	0.77	8	+1.62	1.96	3.34	1.11
Brücke-AR BLG pro Kopf	11	+0.07	1.86	2.47	0.93	10	-0.15	1.80	2.28	0.82	9	-0.13	1.94	2.43	0.85	8	-0.28	2.69	3.16	1.05
Brücke-AR Verf.Eink.	11	-0.28	1.83	2.52	0.95	10	-0.19	1.88	2.45	0.88	9	-0.31	1.99	2.36	0.83	8	-0.51	2.72	3.07	1.02
Historischer Mittelwert	11	-0.06	2.06	2.65	1.00	10	-0.29	2.17	2.78	1.00	9	-0.10	2.19	2.85	1.00	8	-0.08	2.41	3.01	1.00
Log-Trend ab 2005	11	-0.50	2.26	2.67	1.01	10	-0.75	2.45	2.84	1.02	9	-0.59	2.46	2.87	1.01	8	-0.59	2.66	3.03	1.01
Log-Trend ab 2010	11	-0.49	2.15	2.70	1.02	10	-0.79	2.61	3.06	1.10	9	-0.68	2.67	3.02	1.06	8	-0.72	2.88	3.17	1.05
ARIMA in Levels	11	-0.80	2.32	2.72	1.03	10	-1.13	2.63	2.96	1.06	9	-1.06	2.72	3.04	1.07	8	-1.04	2.89	3.19	1.06
ARIMA	11	+0.01	1.64	2.73	1.03	10	-0.15	2.29	3.07	1.10	9	+0.06	2.20	2.97	1.04	8	+0.17	2.34	3.05	1.01
Log-Trend ab 2000	11	-0.72	2.37	2.74	1.03	10	-0.96	2.57	2.91	1.05	9	-0.79	2.57	2.92	1.02	8	-0.79	2.76	3.08	1.02
Log-Trend ab 1991	11	-0.67	2.37	2.75	1.04	10	-0.88	2.51	2.88	1.04	9	-0.69	2.50	2.91	1.02	8	-0.67	2.69	3.07	1.02
Random Walk (Rate)	11	-0.35	2.38	3.55	1.34	10	-0.83	3.74	4.77	1.72	9	-0.93	3.69	4.66	1.64	8	-0.65	3.66	4.51	1.50

Anmerkungen: Fehlermaße in Prozentpunkten der Veränderungsrate. ME = Mittlerer Fehler. MAE = Mittlerer absoluter Fehler. RMSE = Quadratwurzel des mittleren quadrierten Fehlers. rel. = RMSE relativ zum Bezugsmodell (historischer Mittelwert), < 1: besser als Bezugsmodell. n = Zahl der Evaluationspunkte, definiert durch das Modell mit der geringsten Zahl an Evaluationspunkten.

Abbildung A.3: Prognosegüte am EU-SILC-Maßstab (ohne Bruchjahr 2020), alle Modelle – Herbst

Horizont t-1

Modell	n	ME	MAE	RMSE	rel.
Rate BLG pro Kopf (GD)	11	+0.45	1.47	1.87	0.71
Rate Verf. Eink. (GD)	11	+0.75	1.34	1.92	0.72
Rate Bruttolöhne (GD)	11	+1.46	1.62	1.95	0.74
Brücke Bruttolöhne (GD	11	+0.26	1.50	2.09	0.79
Brücke BLG pro Kopf (G	11	+0.06	1.67	2.19	0.83
Brücke Verf. Eink. (GD)	11	-0.08	1.69	2.21	0.83
Brücke-AR Bruttolöhne	11	+0.07	1.58	2.27	0.86
Brücke Nettolöhne (GD)	11	+0.28	1.60	2.27	0.86
Brücke-AR Nettolöhne	11	+0.06	1.60	2.38	0.90
Rate Nettolöhne (GD)	11	+1.58	1.81	2.38	0.90
Brücke-AR Verf.Eink.	11	-0.28	1.83	2.52	0.95
Historischer Mittelwert	11	-0.06	2.06	2.65	1.00
Log-Trend ab 2005	11	-0.50	2.26	2.67	1.01
Log-Trend ab 2010	11	-0.49	2.15	2.70	1.02
ARIMA in Levels	11	-0.80	2.32	2.72	1.03
ARIMA	11	+0.01	1.64	2.73	1.03
Log-Trend ab 2000	11	-0.72	2.37	2.74	1.03
Log-Trend ab 1991	11	-0.67	2.37	2.75	1.04
Random Walk (Rate)	11	-0.35	2.38	3.55	1.34

Horizont t+0

n	ME	MAE	RMSE	rel.
10	+0.15	1.29	1.66	0.60
10	+0.37	1.18	1.67	0.60
10	+1.08	1.33	1.58	0.57
10	-0.06	1.58	2.08	0.75
10	-0.34	1.84	2.20	0.79
10	-0.24	1.76	2.24	0.81
10	-0.22	1.79	2.30	0.83
10	-0.08	1.77	2.30	0.83
10	-0.22	1.93	2.48	0.89
10	+1.24	1.57	2.13	0.77
10	-0.42	2.14	2.67	0.96
10	-0.29	2.17	2.78	1.00
10	-0.75	2.45	2.84	1.02
10	-0.79	2.61	3.06	1.10
10	-1.13	2.63	2.96	1.06
10	-0.15	2.29	3.07	1.10
10	-0.96	2.57	2.91	1.05
10	-0.88	2.51	2.88	1.04
10	-0.83	3.74	4.77	1.72

Horizont t+1

n	ME	MAE	RMSE	rel.
9	+0.90	1.80	2.67	0.94
9	+0.62	1.46	2.58	0.91
9	+1.57	1.71	2.81	0.99
9	+0.52	1.65	2.54	0.89
9	+0.10	1.95	2.60	0.91
9	-0.32	2.13	2.65	0.93
9	+0.40	1.72	2.54	0.89
9	+0.24	1.96	2.68	0.94
9	+0.07	2.05	2.69	0.94
9	+1.59	1.89	2.97	1.04
9	-0.39	2.20	2.69	0.94
9	-0.10	2.19	2.85	1.00
9	-0.59	2.46	2.87	1.01
9	-0.68	2.67	3.02	1.06
9	-1.06	2.72	3.04	1.07
9	+0.06	2.20	2.97	1.04
9	-0.79	2.57	2.92	1.02
9	-0.69	2.50	2.91	1.02
9	-0.93	3.69	4.66	1.64

Horizont t+2

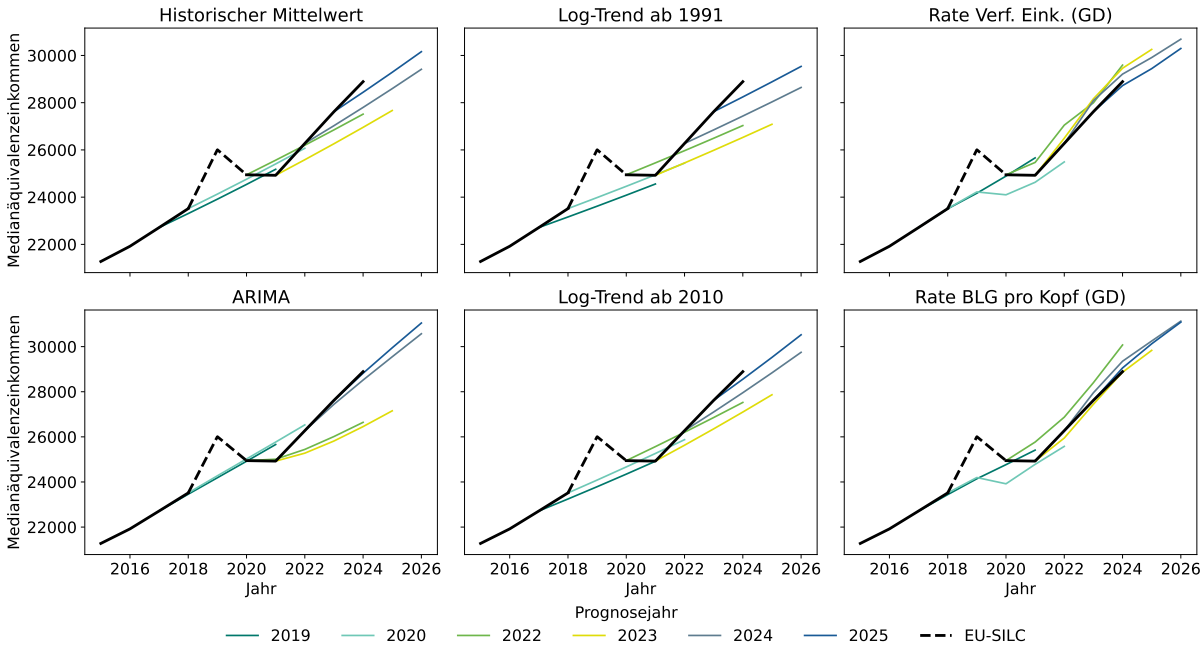
n	ME	MAE	RMSE	rel.
6	+0.94	2.78	3.46	1.01
6	+1.15	2.77	3.56	1.03
6	+1.72	2.63	3.69	1.07
6	+0.74	2.49	3.36	0.98
6	+0.26	2.69	3.31	0.96
6	+0.14	2.82	3.35	0.97
6	+0.72	2.48	3.34	0.97
6	+0.53	2.71	3.39	0.99
6	+0.48	2.71	3.36	0.98
6	+1.82	2.87	3.84	1.12
6	+0.10	2.83	3.34	0.97
6	+0.17	2.95	3.44	1.00
6	-0.31	3.06	3.39	0.99
6	-0.11	2.99	3.35	0.97
6	-0.81	3.28	3.54	1.03
6	+0.47	2.88	3.50	1.02
6	-0.51	3.14	3.43	1.00
6	-0.44	3.13	3.45	1.00
6	+0.12	3.89	4.92	1.43

Anmerkungen: Fehlermaße in Prozentpunkten der Veränderungsrate. ME = Mittlerer Fehler. MAE = Mittlerer absoluter Fehler. RMSE = Quadratwurzel des mittleren quadrierten Fehlers. rel. = RMSE relativ zum Bezugsmodell (historischer Mittelwert), < 1: besser als Bezugsmodell. n = Zahl der Evaluationspunkte, definiert durch das Modell mit der geringsten Zahl an Evaluationspunkten.

A.4 Zusätzliche Schätzergebnisse: Herbst

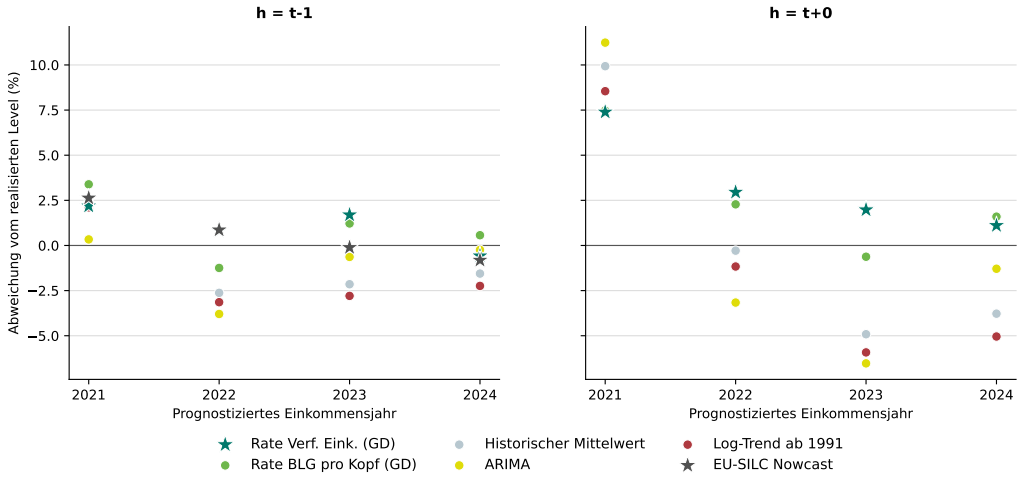
Abschnitt 4 stellt die Ergebnisse für beide Prognosezeitpunkte dar, für den Herbst jedoch nur in Auswahl. Analog zu den Abbildungen 5 bis 9 dokumentiert dieser Anhang die Ergebnisse für den Prognosezeitpunkt im Herbst, zu dem die Einkommensdaten bereits für ein weiteres Jahr vorliegen und die GD-Prognosen (seit 2016) für zwei Jahre im Voraus erstellt.

Abbildung A.4: Verkettete Prognosepfade ausgewählter Modelle und realisierte EU-SILC-Reihe – Herbst



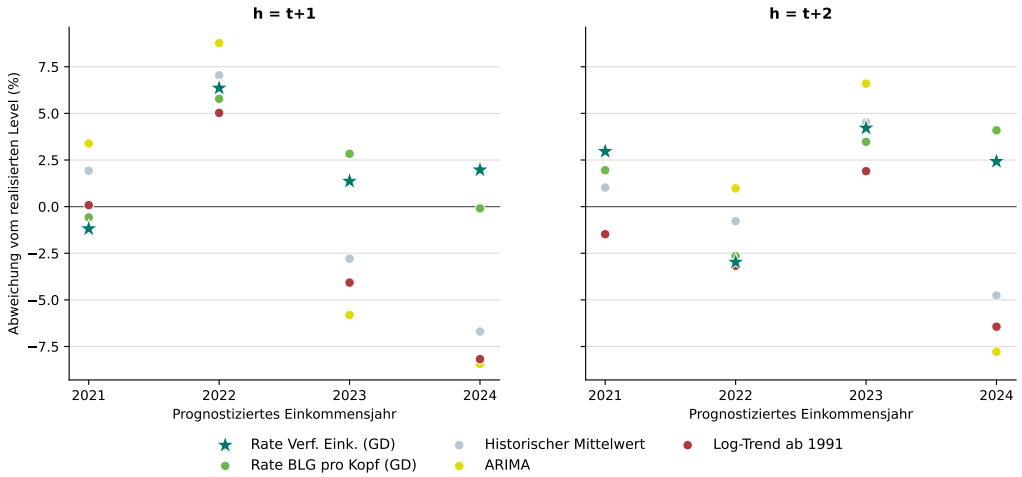
Anmerkungen: Realisiertes Niveau (schwarze Linie) und prognostizierter Verlauf des Medianäquivalenzeinkommens für verschiedene Prognosezeitpunkte (farbige Linien) für ausgewählte Modelle. Der prognostizierte Verlauf ergibt sich jeweils als (letzter beobachteter Wert $\times$ prognostizierte Raten).

Abbildung A.5: Fehler der verketteten Niveauprognosen – Herbst I/II



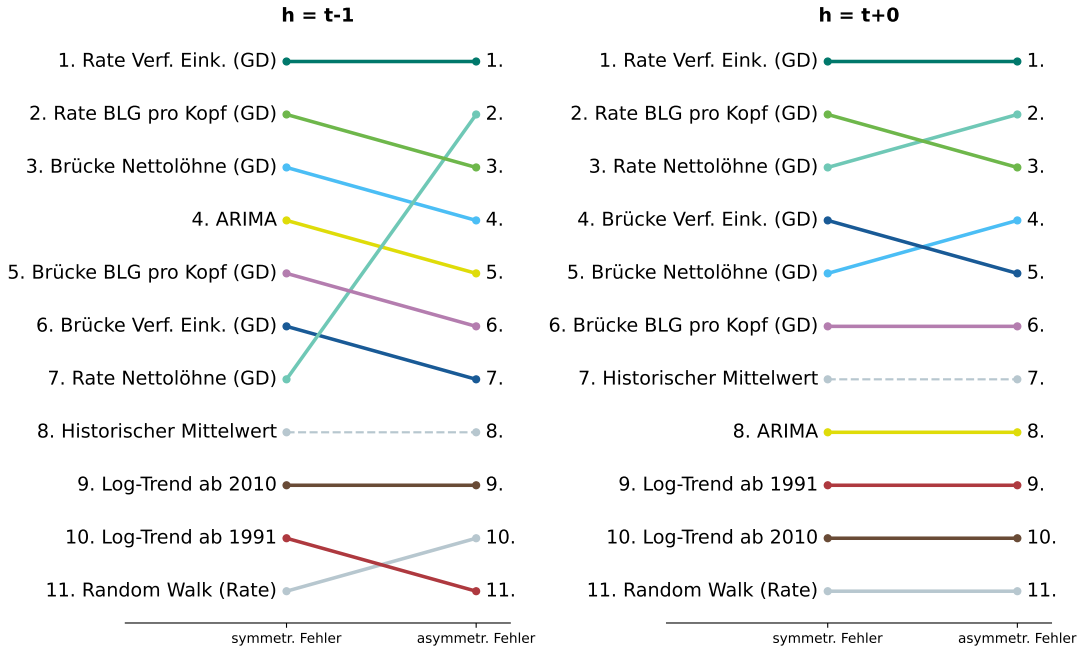
Anmerkungen: Abweichung der verketteten Niveauprognosen (Anker-Level × prognostizierte Raten) vom realisierten Wert für verschiedene Modelle, Prognosehorizonte und Jahre. J bezeichnet das prognostizierte Einkommensjahr, Anker ist jeweils der letzte verfügbare EU-SILC-Wert. Links ($h = t - 1$): Prognosezeitpunkt im Herbst des Jahres $J + 1$, Anker = Einkommensjahr $J - 1$. Rechts ($h = t + 0$): Prognosezeitpunkt im Herbst des Jahres J, Anker = Einkommensjahr $J - 2$. Fehler > 0 = Überschätzung.

Abbildung A.6: Fehler der verketteten Niveauprognosen – Herbst II/II



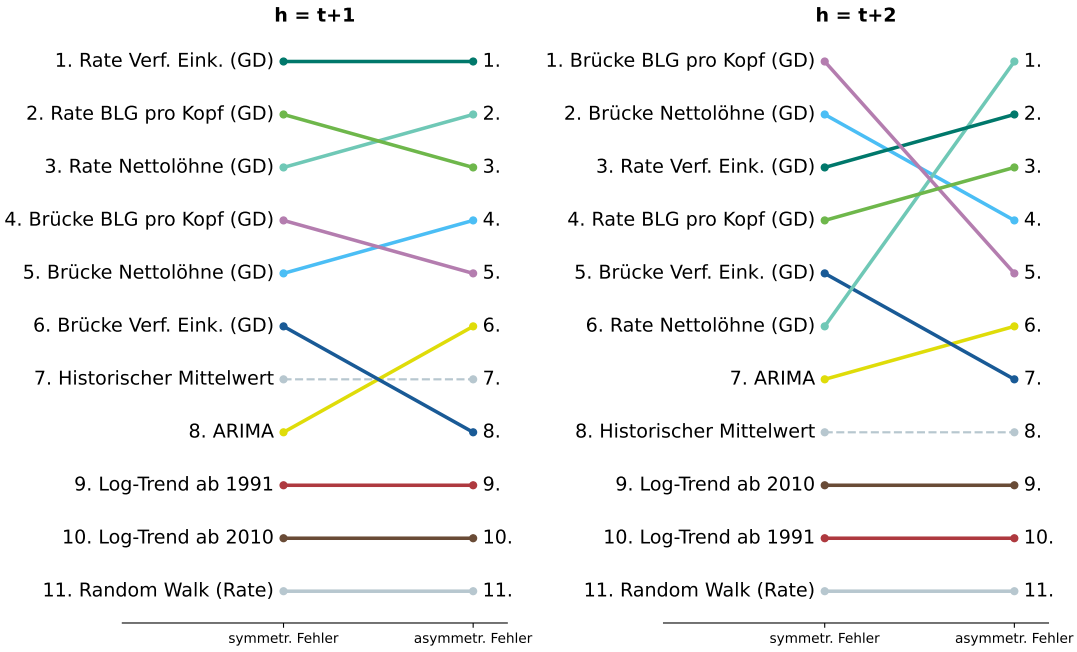
Anmerkungen: Abweichung der verketteten Niveauprognosen (Anker-Level × prognostizierte Raten) vom realisierten Wert für verschiedene Modelle, Prognosehorizonte und Jahre. J bezeichnet das prognostizierte Einkommensjahr, Anker ist jeweils der letzte verfügbare EU-SILC-Wert. Links ($h = t + 1$): Prognosezeitpunkt im Herbst des Jahres $J - 1$, Anker = Einkommensjahr $J - 3$. Rechts ($h = t + 2$): Prognosezeitpunkt im Herbst des Jahres $J - 2$, Anker = Einkommensjahr $J - 4$. Fehler > 0 = Überschätzung.

Abbildung A.7: Rangverschiebung zwischen symmetrischer und asymmetrischer Bewertung – Herbst I/II



Anmerkungen: Rang jedes Modells nach mittlerem absolutem Fehler (MAE) unter symmetrischer Bewertung (links) gegenüber dem Rang unter Lin-Lin-Verlust im Szenario 2:1 (rechts). Rang 1 = geringster (gewichteter) absoluter Fehler.

Abbildung A.8: Rangverschiebung zwischen symmetrischer und asymmetrischer Bewertung – Herbst II/II



Anmerkungen: Rang jedes Modells nach mittlerem absolutem Fehler (MAE) unter symmetrischer Bewertung (links) gegenüber dem Rang unter Lin-Lin-Verlust im Szenario 2:1 (rechts). Rang 1 = geringster (gewichteter) absoluter Fehler.