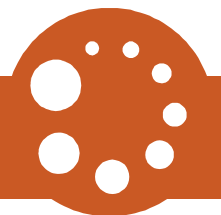




Modellgestützte Schätzung der Bruttowertschöpfung und verwandter Indikatoren in der Statistik der Unternehmensgruppen (VAB-STAGRE)

EXPERIMENTAL STATISTICS



Neuchâtel, 2024

Herausgeber: Bundesamt für Statistik (BFS)
Auskunft: stagre@bfs.admin.ch, Tel. +41 58 463 66 13
Redaktion: Marius Ley, BFS
Themenbereich: 06 Industrie, Dienstleistungen
Originaltext: Deutsch

Layoutkonzept: Sektion PUB
Download: www.statistik.ch
Copyright: BFS, Neuchâtel 2024
Wiedergabe unter Angabe der Quelle
für nichtkommerzielle Nutzung gestattet

Modellgestützte Schätzung der Bruttowertschöpfung und verwandter Indikatoren in der Statistik der Unternehmensgruppen (VAB-STAGRE)

Marius Ley, BFS*

23. Mai 2024

1 Einleitung

1.1 Kontext und Ziele

Seit dem Publikationsjahr 2020 werden in der Statistik der Unternehmensgruppen (STAGRE) Resultate zum Umsatz veröffentlicht, die auf einem Modellierungsansatz beruhen. Dies bedeutet, dass als massgebliche Komponenten für die Berechnung von aggregierten Umsätzen in der STAGRE verwendet werden:

- a) der gemäss **Wertschöpfungsstatistik (WS)** erhobene “Umsatz betriebswirtschaftlich” für jene Einheiten, für welche dieser vorhanden ist (also für die Einheiten der Nettostichprobe der WS);
- b) ein **Schätzwert** dieses Umsatzes gemäss WS für jene Einheiten des Rahmens der WS, die nicht in der Nettostichprobe enthalten sind, modelliert auf der Basis verfügbarer Hilfsvariablen.

Diese Vorgehensweise basiert somit auf für “**Small Area Estimates**” (SAE) typischen Grundideen. Einen detaillierten Beschrieb dazu liefert die folgende Publikation: *Schätzung des Umsatzes in der Statistik der Unternehmensgruppe (STAGRE): Methodische Grundlagen*, Bundesamt für Statistik, Neuchâtel 2023 (im Folgenden jeweils mit “Methodische Grundlagen Umsatz” abgekürzt).¹

Das vorliegende Projekt hat zum Ziel, diesen für den Umsatz verwendeten Modellierungsansatz so zu erweitern, dass er sich auch zur Produktion von Resultaten der Variablen **Bruttoproduktionswert (BPW)**, **Vorleistungen (VL)** und **Bruttowertschöpfung (BWS)** in der STAGRE einsetzen lässt. Diese drei Variablen werden ebenfalls in der WS erhoben und sind grundlegend für die Berechnung des Produktionskontos innerhalb des Systems der Volkswirtschaftlichen Gesamtrechnung (VGR).² Angestrebt wird, wie bereits beim Umsatz, für

die Schnittmenge aus der STAGRE und dem Rahmen der WS Totalschätzungen dieser Variablen bereitzustellen.³ Dabei sollen diese Ergebnisse auch nach diversen Kategorien (darunter insbesondere: Art der Unternehmensgruppe, Grössenklassen, Sitzland der ausländisch kontrollierten Gruppen, Branchen gemäss BFS-50) aufgeschlüsselt werden können.

1.2 Spezifische Herausforderungen

In Bezug auf die bisher (also für die Zielvariable “Umsatz”) erfolgten Arbeiten zur Modellierung präsentieren sich bei den Variablen BPW, VL und BWS zwei grundlegende neue Herausforderungen:

1. Die Variablen des Produktionskontos weisen gegenüber dem Umsatz gewisse **konzeptionelle und statistische Besonderheiten** auf (präziser erläutert in den Abschnitten 4 bis 6), die demzufolge im Rahmen der bisher erfolgten Arbeiten zur Modellierung des Umsatzes noch nicht behandelt wurden.
2. Der **Erklärungsgehalt der verfügbaren Hilfsvariablen** mit diesen drei neuen Zielvariablen dürfte geringer ausfallen, als dies beim Umsatz gemäss WS der Fall war. Eine ziemlich hohe Korrelation hatte sich insbesondere zwischen der Hilfsvariablen MWST-Umsatz und dem WS-Umsatz gezeigt, was nicht überrascht, da eher geringe konzeptionelle und erhebungstechnische Differenzen zwischen diesen zwei Variablen bestehen. Es ist deshalb noch abzuklären und aufzuzeigen, ob der Modellierungsansatz auch für die drei neuen Zielvariablen ausreichend zuverlässige Schätzungen liefert.

*Dank gebührt insbesondere Anne Massiani und Daniel Assoulin von der Sektion Statistische Methoden (METH) des BFS für die sorgfältige Relektüre des Dokuments.

¹In dieser Publikation wurde der Begriff der “Small Area Estimates” allerdings nicht explizit verwendet.

²Eine Bereitstellung dieser Variablen in der STAGRE ist deshalb anzustreben, weil in makroökonomischen und struktu-

rellen Analysen die Variablen BPW oder BWS für die Messung des wirtschaftlichen Outputs einer Einheit aussagekräftiger sind als der Umsatz. Zudem ist für die Mitgliedsländer der Europäischen Union die Bereitstellung dieser Variablen obligatorisch, gemäss der “European Business Statistics”-Richtlinie im Bereich der “Foreign Affiliates Statistics” (FATS; übertragen auf den Schweizerischen Kontext ist die FATS-Population eine Teilmenge der STAGRE). Siehe https://ec.europa.eu/eurostat/statistics-explained/index.php?title=European_business_statistics_manual

³Somit wird in Kauf genommen, dass für einen bedeutenden Teil der STAGRE keine Ergebnisse ausgewiesen werden können; siehe Unterabschnitt 2.1.

Der **ersten Herausforderung** sind die Abschnitte 4 bis 6 gewidmet. Dort werden die konzeptionellen und statistischen Eigenheiten der Variablen des Produktionskontos erläutert und jeweils Lösungsansätze dazu vorgeschlagen und evaluiert. Diese Evaluation erfolgt, indem jeweils unterschiedliche Modellvarianten gerechnet und in Bezug auf die Präzision deren Vorhersagen miteinander verglichen werden. Dies geschieht grundsätzlich mittels des Determinationskoeffizienten R^2 (Definition siehe Unterabschnitt 2.4) als Qualitätsmass. R^2 stellt ein Mass für den linearen Zusammenhang der Vorhersagen mit den erhobenen Werten dar; somit sind hier möglichst hohe Werte anzustreben (ein Wert von eins entspricht einer perfekten Vorhersage). Gegenüber dem bis anhin (siehe Methodische Grundlagen Umsatz) für die Evaluation verwendeten RMSE hat R^2 den Vorteil, dass dessen Werte nicht nur innerhalb einer gegebenen Zielvariablen, sondern auch zwischen den verschiedenen Zielvariablen eine gewisse Vergleichbarkeit ermöglichen. Somit erlaubt die Verwendung des R^2 , eine quantitative Beurteilung der oben als **zweite Herausforderung** formulierten Besorgnis, dass unter dem mutmasslich geringeren Erklärungsgehalt der Hilfsvariablen mit den Zielvariablen des Produktionskontos die Präzision der Vorhersagen geringer ausfallen könnte als beim Umsatz.

2 Grundsätzliches zum Vorgehen

2.1 Relevante Populationen

Wie bereits bei der Modellierung des Umsatzes erfolgt die **Schätzung** der Modelle unter Verwendung aller Einheiten der Nettostichprobe der WS; also auch derjenigen Einheiten, die nicht als Unternehmensgruppen organisiert sind und somit ausserhalb der STAGRE-Population liegen. Dieses Vorgehen erlaubt, auf eine grössere Zahl von Beobachtungen zurückzugreifen und somit präzisere Schätzergebnisse zu erhalten, als wenn die Schätzpopulation auf diejenigen Einheiten der WS eingeschränkt würde, die sich gleichzeitig im STAGRE-Perimeter befinden.⁴ Dies erfordert allerdings, dass in den beiden Subpopulationen — Unternehmen des STAGRE-Perimeters; und Unternehmen, die keiner Gruppe angeschlossen sind — die relevanten Variablen der gleichen Verteilung folgen, da ansonsten verzerrte Modellvorhersagen resultieren würden. Um dem Risiko einer solchen Verzerrungen zu begegnen, enthält die gewählte Modellspezifikation (siehe Unterabschnitt 2.2) eine binäre Variable für die Zugehörigkeit einer Einheit zu einer multinationalen Unternehmensgruppe.⁵

⁴In den verfügbaren Referenzjahren sind in der Nettostichprobe der WS (jeweils zwischen 13'995 und 14'831 Beobachtungen) Einheiten in Unternehmensgruppen anzahlmässig zu weniger als der Hälfte vertreten (jeweils zwischen 5'365 und 6'045 Beobachtungen).

⁵Zudem könnten geeignete Residuenplots erlauben, allfällige vom Modell nicht in ausreichender Form berücksichtigte Unterschiede in der Verteilung zwischen Einheiten innerhalb vs. ausserhalb der STAGRE sichtbar zu machen; dieser Punkt ist für wei-

Die **Vorhersage** von Schätzwerten der Modelle geschieht – ebenfalls wie beim Umsatz — für alle Einheiten, die der Schnittmenge aus dem Rahmen der WS und der STAGRE angehören. Eine Bereitstellung von Schätzwerten für Einheiten der STAGRE, die ausserhalb des Rahmens der WS liegen, ist bis auf weiteres nicht vorgesehen. Bei diesen ausgeschlossenen Einheiten handelt es sich um

- Einheiten in Kombinationen aus Branchen und Rechtsformen, die im Rahmen der WS per Definition nicht berücksichtigt wurden. Im Referenzjahr 2020 entfielen in der STAGRE auf solche Einheiten 16.1% der Beschäftigung und 13.5% des Umsatzes gemäss MWST-Daten.⁶
- Einheiten mit weniger als drei Beschäftigten. Im Referenzjahr 2020 entfielen in der STAGRE auf solche Einheiten 0.4% der Beschäftigung und 4.3% des Umsatzes gemäss MWST-Daten.

Bei der Veröffentlichung von auf diesen Schätzungen basierenden Resultaten ist folglich in geeigneter Art und Weise auf diese Einschränkungen hinzuweisen.

2.2 Modellspezifikation

An der grundlegenden Modellspezifikation, die bereits zur Vorhersage des Umsatzes verwendet wurde, wird auch im Rahmen des vorliegenden Projektes festgehalten. Formell ausgedrückt wurden Modelle konzipiert, die auf Ebene der Einheiten i des WS-Rahmens eine Vorhersage zu einer Zielvariablen Y_i liefern,⁷ basierend auf den Hilfsvariablen “Umsatz gemäss MWST-Daten” T_i und “Arbeitseinkommen gemäss AHV” L_i . Dabei wurde der folgende multiplikative Zusammenhang unterstellt:

$$Y_i = A_{g(i)} T_i^{\beta_{g(i)}} L_i^{\gamma_{g(i)}} e^{\delta' x_i} e^{\varepsilon_i} \quad (1)$$

Der Term $g(i)$ verweist auf die Branchenzugehörigkeit des Unternehmens i . Der Einfluss, den die beiden verwendeten Hilfsvariablen T_i und L_i auf die Zielvariable Y_i haben, variiert somit in Abhängigkeit der wirtschaftlichen Tätigkeit des Unternehmens. Der Vektor x_i enthält die folgenden erklärenden binären Variablen:

- ob i Teil einer multinationalen Unternehmensgruppe ist;

terführende Analysen (siehe Abschnitt 8) angedacht.

⁶Bei den in der STAGRE bedeutsamen Branchen, die im WS-Rahmen ganz oder teilweise ausgeschlossen wurden, handelt es sich insbesondere um die NOGA-Abteilungen 64–66 (Finanz- und Versicherungsdienstleistungen) sowie 86 (Gesundheitswesen).

⁷Der Einfachheit halber wird die Zielvariable durchwegs mit Y_i bezeichnet, ohne in der Notation explizit zwischen den unterschiedlichen Zielvariablen (BPW, VL und BWS) zu unterscheiden. Dies deshalb, weil angestrebt wird, für die Schätzung der drei Variablen jeweils ein möglichst ähnlich spezifiziertes Modell zu verwenden.

- ob i im Verlauf des jeweiligen Referenzjahres Waren exportiert hat;
- ob i im Verlauf des jeweiligen Referenzjahres Waren exportiert hat mit einem Wert, der den Wert des MWST-Umsatzes T_i übersteigt.

Um auf lineare Schätzverfahren zurückgreifen zu können, wird Gleichung (1) in logarithmierter Form geschätzt:

$$y_i = \alpha_{g(i)} + \beta_{g(i)} t_i + \gamma_{g(i)} l_i + \delta' x_i + \varepsilon_i \quad (2)$$

Wobei hier zwecks einfacherer Darstellung die Logarithmen der Variablen mittels derer Kleinbuchstaben ausgedrückt werden, also beispielsweise $\log Y_i = y_i$, und die Logarithmen von A_i mittels α_i . Gleichung (2) wird als *Linear Mixed Model* (LMM) geschätzt, da dieser Ansatz erlaubt, den Parameter δ als Fix-Effekt, die Parameter $\alpha_{g(i)}$, $\beta_{g(i)}$ und $\gamma_{g(i)}$ hingegen als “Random”-Effekte zu spezifizieren.⁸ Diese Modellwahl beruht auf zwei Verteilungsannahmen: Einerseits einer Annahme zur Verteilung der Zielvariablen Y_i , andererseits einer Annahme zur Verteilung der “Random”-Effekte (die im Folgenden zwecks besserer Lesbarkeit als $b_{g(i)}$ zusammengefasst werden).

Für die **Verteilung der Zielvariablen** wird angenommen, dass y_i (also der Logarithmus von Y_i) einer bedingten Normalverteilung folgt:

$$(y_i | x_i, z_i, b_{g(i)}) \sim \mathcal{N}(\delta' x_i + b'_{g(i)} z_i, \sigma^2)$$

Der Vektor $b_{g(i)}$ der “Random”-Effekte beinhaltet die bereits beschriebenen Koeffizienten:

$$b_{g(i)} \equiv (\alpha_{g(i)}, \beta_{g(i)}, \gamma_{g(i)})$$

Der Vektor der dazugehörigen Kovariate z_i beinhaltet folglich eine Konstante sowie die Logarithmen der beiden Hilfsvariablen (MWST-Umsatz und Arbeitseinkommen):

$$z_i \equiv (1, t_i, l_i)$$

Bezüglich der **Verteilung der “Random”-Effekte** werden zwei unterschiedliche Varianten von Annahmen postuliert (und evaluiert). In der detaillierteren der beiden Varianten (mit *genesteten* “Random”-Effekten) werden die “Random”-Effekte als Summe einer fixen Komponente μ_b (die dem Erwartungswert entspricht) und zweier zentrierter normalverteilter Komponenten spezifiziert, wovon die erste auf Stufe der **noga2r**-Branche⁹

$G[g(i)]$ der Beobachtung i und die zweite auf Stufe deren NOGA-Art (also die Branchenzugehörigkeit gemäss der detailliertesten verfügbaren, sechsstelligen Gliederungsstufe) $g(i)$ bestimmt wird:

$$\begin{aligned} b_{g(i)} &\equiv \mu_b + e_{G[g(i)]} + f_{g(i)} \\ \text{mit } \mu_b &\equiv (\alpha, \beta, \gamma), \\ e_{G[g(i)]} &\sim \mathcal{N}(\mathbf{0}, \Sigma_G) \quad \text{und} \quad f_{g(i)} \sim \mathcal{N}(\mathbf{0}, \Sigma_g) \end{aligned}$$

Wobei es sich bei Σ_G und Σ_g um die Varianz-Kovarianzmatrizen der “Random”-Effekte für die beiden Gliederungsstufen handelt. Diese Matrizen sind per Definition symmetrisch und positiv semidefinit, können ansonsten aber beliebige Werte annehmen. Dies erlaubt für Korrelation unter den “Random”-Effekten jeweils innerhalb der beiden Gliederungsstufen (nicht jedoch zwischen den beiden Gliederungsstufen; das heisst, $e_{G[g(i)]}$ und $f_{g(i)}$ werden als miteinander nicht korreliert betrachtet).¹⁰ Der Begriff *genestet* bezieht sich auf die Tatsache, dass es sich bei der **noga2r**-Branche $G[g(i)]$ um eine Aggregationsstufe auf Basis der NOGA-Art $g(i)$ handelt. In der einfacheren (nicht *genesteten*) der hier postulierten Varianten wird keine Variabilität der “Random”-Effekte auf Stufe NOGA-Art unterstellt, das heisst $f_{g(i)} = 0$; es werden also lediglich “Random”-Effekte auf Stufe **noga2r**, also $e_{G[g(i)]}$, geschätzt.

LMM-Spezifikationen zeigten bereits für die Schätzung des Umsatzes die besten Resultate und ermöglichen insbesondere, für die NOGA-Zugehörigkeit der Beobachtung i eine detailliertere Gliederungsstufe zu verwenden, als dies mit reinen Fix-Effekten möglich wäre.

2.3 Schätzungen und Vorhersage verschiedener Modellvarianten

Die Bereitstellung der Modellvorhersagen — ausgehend von der grundlegenden Schätzgleichung (2) — beinhaltet mehrere Schritte, um der Komplexität der Variablen, Modelle und Datenlage gerecht zu werden. Um eine bessere Übersicht über diese Schritte zu ermöglichen, werden diese hier kurz erläutert und motiviert, wobei zwecks detaillierterer Ausführungen jeweils auf nachfolgende Abschnitte verwiesen wird.

1. **Schätzen von Teilmodellen**, jeweils in verschiedenen Varianten für jede Zielvariable. Das Schätzen verschiedener Varianten von Teilmodellen für dieselbe Zielvariable erlaubt insbesondere, die folgenden Eigenheiten der Daten zu berücksichtigen:

⁸Zur Implementierung der LMM-Schätzungen wurde auf das R-Package **lme4** zurückgegriffen.

⁹Die Branchenklassifizierung **noga2r** stützt sich auf die NOGA-Abteilungen (erste zwei Stellen der NOGA-Klassifizierung), wobei vier Abteilungsgruppen jeweils zusammengefasst werden (05–09, 11–12, 19–20, 38–39).

¹⁰Die verwendete Prozedur (siehe Fussnote 8) bietet die Option, “Random”-Effekte auch innerhalb einer Gliederungsstufe als untereinander nichtkorreliert zu spezifizieren. Der Einfachheit halber, und da die festgestellten Korrelationen als plausibel erschienen, wurde — wie schon bei der Schätzung des Umsatzes — darauf verzichtet, Varianten mit unterstellter Nichtkorrelation zu schätzen und zu evaluieren.

- (a) Die Hilfsvariable MWST-Umsatz ist nicht für alle Beobachtungen verfügbar.¹¹
- (b) Bei gewissen Modellspezifikationen können in einzelnen Subgruppen $g(i)$ unplausible Schätzwerte für die “Random”-Effekte resultieren (siehe Abschnitt 4).¹²
- (c) Für die Zielvariable BWS muss auch die Möglichkeit negativer Werte berücksichtigt werden (siehe Abschnitt 5).

Darüber hinaus bildet das Schätzen verschiedener Varianten von Teilmodellen die Basis für die Evaluation (weiter unten, Schritt 4) unterschiedlicher Modellvarianten, was wiederum eine empirisch gestützte Wahl der besten Modellspezifikation ermöglicht. Implementiert werden sowohl “Linear Mixed Models” (LMM) für Gleichung (2) für die drei Zielvariablen, als auch binomiale “Generalized Linear Mixed Model” mit Logit-Funktion (für die Wahrscheinlichkeit des Auftretens negativer Werte in der Variablen BWS; siehe Abschnitt 5).

2. **Vorhersagen der Teilmodelle** berechnen. Da das logarithmierte Modell eine Vorhersage für $E(\log Y_i | x_i, z_i, b_{g(i)})$ liefert, und da die gewünschte Vorhersage $E(Y_i | x_i, z_i, b_{g(i)})$ sich nicht einfach aus $\exp[E(\log Y_i | x_i, z_i, b_{g(i)})]$ ergibt (der Logarithmus ist keine lineare Transformation), ist eine Skalierung erforderlich. Konkret berechnet sich die Vorhersage der Zielvariablen in absoluten Werten $\hat{Y}_i$ aus der logarithmierten Vorhersage $\hat{y}_i$ wie folgt:

$$\hat{Y}_i \equiv \exp \left(\hat{y}_i + \frac{1}{2} \hat{\sigma}_{G[g(i)]}^2 \right) \quad (3)$$

Dieser Formulierung liegt die Überlegung zugrunde, dass für eine logarithmisch normalverteilte Zufallsvariable X , mit $\log X \sim \mathcal{N}(\mu, \sigma)$, für den Erwartungswert $E(X) = \exp(\mu + \sigma^2/2)$ gilt. In der hier implementierten Methode bezeichnet $\hat{\sigma}_{G[g(i)]}^2$ den um Extremwerte bereinigten, nach **noga2r**-Branchen $G[g(i)]$ differenzierten Mittelwert der quadrierten Schätzfehler $\hat{\varepsilon}_i^2$. Die Bereinigung um Extremwerte erfolgt, indem von der Berechnung dieser Mittelwerte jene Beobachtungen i ausgeschlossen werden, deren Absolutwert von $\hat{\varepsilon}_i$ grösser als drei mal die geschätzte residuale Standardabweichung des Teilmodells ist. Diese Variante der Bereinigung,

¹¹Um eine Vorhersage der Zielvariablen auch für Einheiten mit fehlendem MWST-Umsatz zu ermöglichen, wird im Rahmen des *hybrid-gepoolten Ansatzes* (erläutert im Technischen Anhang der Methodischen Grundlagen zum Umsatz) jeweils immer auch eine Variante der Modellgleichung unter Weglassung der erklärenden Variablen T_i geschätzt.

¹²Bei der Schätzung des Umsatzes zeigte sich, dass Subgruppen $g(i)$ mit negativen und somit mutmasslich unplausiblen Schätzwerten für $\beta_{g(i)}$ oder $\gamma_{g(i)}$ auftreten können. Deshalb werden für alle Zielvariablen zusätzlich auch Modellvarianten gerechnet und evaluiert, die solche Schätzwerte unterdrücken, indem in diesen Fällen automatisch Vorhersagen aus “Ausweich”-Teilmodellen zum Zuge kommen, in denen auf den Einbezug der betroffenen Hilfsvariablen verzichtet wird.

die den unerwünschten Einfluss allfälliger Extremwerte auf die Vorhersagen eindämmt, kommt bereits für die Produktion der Variablen Umsatz zum Einsatz und hat bei der damaligen Evaluation die besten Ergebnisse erzielt (siehe Methodische Grundlagen Umsatz).

3. Varianten von (Haupt-) **Modellvorhersagen** implementieren, wobei ein (Haupt-) Modell nach bestimmten Regeln auf die Vorhersagen eines oder mehrerer Teilmodelle zurückgreift. Damit wird die Berücksichtigung der Eigenheiten der Daten, die weiter oben in Schritt 1 aufgelistet sind, umgesetzt.
4. Gewährleistung von **untereinander konsistenten Vorhersagen** (siehe Abschnitt 6), das heisst der folgenden definitorischen Gleichheit bzw. Ungleichheit zwischen den drei Variablen des Produktionskontos:
 - (a) $BWS = BPW - VL$ (Unterabschnitt 6.1)
 - (b) $BWS \leq BPW$ (Unterabschnitt 6.2)
5. **Evaluation**, in Bezug auf vordefinierte Qualitätsmasse, der in den vorherigen Schritten implementierten Modellvorhersagen, und Bestimmung des **favorisierten Modells**.

Schritte 1 bis 4 (Schätzung und Vorhersage) werden isoliert für jedes der Referenzjahre implementiert, während für den letzten Schritt 5 (Evaluation) Qualitätsmasse sowohl zusammengefasst über alle Referenzjahre als auch nach einzelnen Referenzjahren ausgewiesen und analysiert werden. Im Rahmen des vorliegenden Projekts werden Daten der Referenzjahre 2014 bis 2021 verwendet und ausgewertet.

2.4 Evaluation der Modellvarianten

Das grundlegende hier verwendete Qualitätsmass — der bereits erwähnte Determinationskoeffizient R^2 — ist wie folgt definiert:

$$R^2 = 1 - \frac{\sum_{t=1}^T \sum_{i=1}^{N_t} (Y_{i,t} - \hat{Y}_{i,t})^2 w_{i,t}}{\sum_{t=1}^T \sum_{i=1}^{N_t} (Y_{i,t} - \bar{Y}_t)^2 w_{i,t}}$$

mit

$$\bar{Y}_t = \frac{1}{\sum_{i=1}^{N_t} w_{i,t}} \sum_{i=1}^{N_t} Y_{i,t} w_{i,t} \quad \forall t \in \{2014, \dots, 2021\}$$

wobei $\hat{Y}_{i,t}$ die vom Modell vorhergesagten und $Y_{i,t}$ die effektiv erhobenen Werte sind, und $w_{i,t}$ ein geeignetes Gewicht; jeweils definiert für N_t Einheiten der Nettostichprobe in den Referenzjahren $t \in \{2014, \dots, 2021\}$.¹³

¹³Im Rahmen der Evaluation der verschiedenen Modellvarianten für den Umsatz kam der wie folgt definierte RMSE zum Einsatz:

$$RMSE_w = \sqrt{\frac{1}{\sum_{t=1}^T \sum_{i=1}^{N_t} w_{i,t}} \sum_{t=1}^T \sum_{i=1}^{N_t} (\hat{Y}_{i,t} - Y_{i,t})^2 w_{i,t}}$$

Tabelle 1: Qualitätsmasse verschiedener Modellvarianten für die Variablen BPW, BWS, VL und Umsatz

Variable	“Genestete” Random-Effekte	Gewich- tung	Unterdrückung negativer $\hat{\beta}_{g(i)}$, $\hat{\gamma}_{g(i)}$	Berücksicht. Werte ≤ 0	Qualitätsmasse, gewichtet R^2 R_{abs}	
BPW	nein	–	–	–	0.6875	0.7365
BPW	nein	ja	–	–	0.6257	0.7170
BPW	ja	–	–	–	0.7474	0.7749
BPW*	ja	–	ja	–	0.7474	0.7751
BPW		(Berechnung mittels $VL^* + BWS^*$)			0.7129	0.7654
BPW		(Maximalwert aus BPW^* und BWS^*)			0.7474	0.7751
VL	nein	–	–	–	0.6850	0.6553
VL*	ja	–	–	–	0.6982	0.7023
VL	ja	–	ja	–	0.5829	0.6930
VL		(Berechnung mittels $BPW^* - BWS^*$)			0.7519	0.7108
VL		(Berechnung mittels (Maximalwert aus BPW^* und BWS^*) $- BWS^*$)			0.7519	0.7108
BWS	nein	–	–	–	0.5372	0.7180
BWS	ja	–	–	–	0.5638	0.7357
BWS	ja	–	ja	–	0.5639	0.7358
BWS*	ja	–	ja	ja	0.5611	0.7372
BWS		(Berechnung mittels $BPW^* - VL^*$)			0.0333	0.6847
BWS		(Minimalwert aus BPW^* und BWS^*)			0.5611	0.7372
Umsatz	ja	–	ja	–	0.8683	0.8343

Mit einem Stern (*) bezeichnet sind diejenigen Modellvarianten, die für die Gewährleistung der definitorischen Gleichheit (siehe Abschnitt 6) herangezogen wurden.

Mit fetter Schrift (in der ersten Spalte) sind die schlussendlich favorisierten Modellvarianten gekennzeichnet.

Als Gewicht für R^2 und R_{abs} dient das Hochrechnungsgewicht der WS.

Ein vom R^2 abgeleitetes, alternatives Qualitätsmass — hier als R_{abs} bezeichnet — erhält man, wenn statt der quadrierten Abweichungen die Absolutwerte der Abweichungen verwendet werden:

$$R_{abs} = 1 - \frac{\sum_{t=1}^T \sum_{i=1}^{N_t} |Y_{i,t} - \hat{Y}_{i,t}| w_{i,t}}{\sum_{t=1}^T \sum_{i=1}^{N_t} |Y_{i,t} - \bar{Y}_t| w_{i,t}}$$

R_{abs} hat die Eigenschaft, weniger sensibel auf einzelne Beobachtungen mit sehr grossen Vorhersagefehlern (ausgedrückt durch den Absolutwert von $Y_{i,t} - \hat{Y}_{i,t}$) zu reagieren als R^2 , da die Vorhersagefehler nicht in ihrem Quadrat, sondern nur als Absolutwert Eingang in die Berechnung finden. Dies kann deshalb einen Vorteil darstellen, weil die hier untersuchten Modelle die Tendenz aufweisen, dass einige wenige Beobachtungen sehr grosse Vorhersagefehler aufweisen.

Diese Problematik lässt sich für die Zielvariable BPW und für das mit BPW* (gemäss Tabelle 1) bezeichnete Modell illustrieren, indem die Aufsummierung der $(Y_{i,t} - \hat{Y}_{i,t})^2 w_{i,t}$ analysiert wird (unter Verwendung des Hochrechnungsgewichtes der Wertschöpfungsstatistik als $w_{i,t}$). Wird, innerhalb der einzelnen Referenzjahre t , alleine für die fünf Beobachtungen mit den jeweils höchsten Werten dieser gewichteten quadrierten Vorhersagefehler die Summe aus diesen Werten gebildet, macht diese durchwegs mehr als die Hälfte der Gesamtsumme (über

alle Beobachtungen des Referenzjahres) der gewichteten quadrierten Vorhersagefehler aus. Eine analoge Analyse für die gewichteten *Absolutwerte* (statt der quadrierten Werte) der Vorhersagefehler ergibt, dass hier die fünf Beobachtungen mit den jeweils höchsten Werten je nach Referenzjahr einen deutlich geringeren Anteil der Gesamtsumme ausmachen, nämlich zwischen 12% und 15%.

Tabelle 1 fasst für die evaluierten Modelle, die in den folgenden Abschnitten detailliert erläutert werden, die Resultate für R^2 und R_{abs} zusammen. Zuunterst in dieser Tabelle sind auch Resultate für R^2 und R_{abs} für die Variable Umsatz enthalten, die sich aus einem Modell ergeben, dessen Spezifikation derjenigen entspricht, die für die bereits erfolgte Veröffentlichung dieser Variablen im Jahr 2022 verwendet wurde. Auf den auffallend tiefen Wert (0.0333) für R^2 der Modellvariante zur Variablen BWS *Berechnung mittels $BPW^* - VL^*$* wird in Unterabschnitt 6.1 eingegangen.

Es lässt sich leicht zeigen (da die im RMSE enthaltene Summierung jener im Zähler des von eins abgezogenen Terms von R^2 entspricht, und da sämtliche anderen Terme in den beiden Massen für eine gegebene Zielvariable invariant sind), dass RMSE und R^2 unter Verwendung des gleichen Gewichts $w_{i,t}$ zu gleichlautenden Einschätzungen der relativen Modellpräzisionen für eine gegebene Zielvariable führen.

3 Grundlegende Spezifikation: “Random”-Effekte, Gewichtung

Bei der Schätzung des Umsatzes erzielten Modelle mit “genesteter” Spezifikation der **“Random”-Effekte**, und zwar auf den zwei Stufen **noga2r** (siehe Fussnote 9) und NOGA-Art, die besten Ergebnisse. Wie im Unterabschnitt 2.2 erläutert, wurden im vorliegenden Projekt für sämtliche Zielvariablen auch Modellvarianten mit vereinfachter Spezifikation der “Random”-Effekte (lediglich auf Stufe **noga2r**, also nicht “genestet”) evaluiert. Die Resultate in Tabelle 1 deuten darauf hin, dass die “genestete” Spezifikation für alle drei Zielvariablen überlegen ist, wobei für die Variable BPW ein etwas grösserer Präzisionsgewinn resultiert als für die beiden anderen Variablen. Somit wird für die in den folgenden Abschnitten betrachteten Modellvarianten jeweils an dieser Spezifikation festgehalten.

Die Analysen zur Schätzung des Umsatzes ergaben auch, dass die Berücksichtigung eines **Gewichts zur Schätzung** der logarithmierten Version von Gleichung (1) zu keiner Verbesserung der verwendeten Qualitätsmasse führte. Getestet wurde dort insbesondere das Hochrechnungsgewicht der WS, das in dieser Erhebung zum Zweck der Ermittlung der direkten Schätzungen bereitgestellt wird (und das im vorliegenden Projekt für die Berechnung der gewichteten Qualitätsmasse herangezogen wird). In der vorliegenden Analyse wird der Einbezug dieses Gewichts in die Schätzung ebenfalls evaluiert, und zwar in einer Modellvariante für die Variable BPW mit vereinfachten “Random”-Effekten. Die Resultate in Tabelle 1 deuten jedoch darauf hin, dass dies eine deutliche Verringerung in den betrachteten Qualitätsmassen zur Folge hätte. Somit wird im weiteren Verlauf und für die weiteren Variablen auf die Verwendung eines Gewichts zur Schätzung der Modellgleichung verzichtet. Weiterführende Analysen und theoretische Überlegungen erscheinen jedoch angebracht, um die Frage des Einbezugs von Gewichten in den Modellschätzungen zu klären.

4 Unterdrückung negativer Schätzwerte für $\beta_{g(i)}$ und $\gamma_{g(i)}$

Die Resultate in Tabelle 1 deuten darauf hin, dass dieses Vorgehen (das bereits für die Variable Umsatz evaluiert und angewendet wurde) bei den Variablen BPW und BWS kaum messbare Veränderungen in den Qualitätsmassen mit sich bringt (keine oder nur geringfügige Verbesserungen). Bei der Variablen VL verschlechtert die Unterdrückung negativer Schätzwerte für $\beta_{g(i)}$ und $\gamma_{g(i)}$ die Vorhersage hingegen deutlich. Dies liegt mutmasslich daran, dass für diese Variable in vielen Fällen negative Werte für $\hat{\gamma}_{g(i)}$ (also den Koeffizienten für die Arbeitseinkommen L_i) erzielt wurden, die in der Regel einen ökonomisch

plausiblen Zusammenhang widerspiegeln: Die Vorleistungen (VL) stellen, wie auch die mittels L_i vergütete Arbeit, einen Input im Produktionsprozess der Unternehmen dar, wobei diese beiden Inputs zu einem gewissen Grad substituierbar sind. Das heisst, innerhalb einer gegebenen Subgruppe $g(i)$ können verschiedene Unternehmen mit einem ähnlichen MWST-Umsatzvolumen T_i auftreten, die sich durch voneinander abweichende Inputkombinationen unterscheiden: Einheiten mit hohem Personalaufwand und tiefen Vorleistungen können ebenso gut existieren wie Einheiten mit tiefem Personalaufwand und hohen Vorleistungen. Dies äussert sich statistisch in einer negativen bedingten Korrelation zwischen den beiden Inputs Vorleistungen und Personalaufwand und in der Folge in negativen Werten für $\hat{\gamma}_{g(i)}$.

Auf Grund dieses empirischen Befunds und dieser Überlegungen werden für die beiden Variablen BPW und BWS die Modellvarianten mit Unterdrückung negativer Schätzwerte für $\beta_{g(i)}$ und $\gamma_{g(i)}$ beibehalten, während für die Variable VL auf eine solche Unterdrückung verzichtet wird.

5 Berücksichtigung nichtpositiver erhobener Werte

5.1 Kontext und Lösungsansatz

Im Kontext der Schätzung des Umsatzes konnte davon ausgegangen werden, dass in der Zielvariablen keine negativen Werte auftreten (da solche ökonomisch kaum Sinn ergeben, und da der Fragebogen der WS die Möglichkeit negativer Umsätze gar nicht zulässt). Etwas anders gestaltet sich die Situation bei den drei Variablen BPW, VL und BWS, da diese in der WS durch Aufsummierung und Saldierung mehrerer Positionen des Fragebogens erzielt werden. Dadurch können negative Werte auftreten, die in den meisten Fällen aus ökonomischer Sicht auch plausibel sind. Tabelle 2 fasst die Häufigkeit des Auftretens negativer Werte zusammen:

Tabelle 2: Anzahl der Beobachtungen mit negativen Werten für die relevanten Zielvariablen und Anzahl Beobachtungen total in der WS-Nettostichprobe, nach Referenzjahr

Ref.- Jahr	Anzahl Beob. mit Wert < 0 für...	Anzahl			
	Umsatz	BPW	VL	BWS	Beob. Total
2014	0	3	0	330	14310
2015	0	13	0	352	14659
2016	0	9	0	342	14136
2017	0	6	0	372	13995
2018	0	4	0	383	14831
2019	0	5	0	370	14439
2020	0	7	1	488	14443
2021	0	15	1	541	14629
Total	0	62	2	3178	115442

Insbesondere die Variable BWS ist von einer nicht vernachlässigbaren Zahl von negativen erhobenen Werten betroffen. Eine weitergehende Analyse zeigt zudem, dass für diese Variable in drei NOGA-Abschnitten (M, P und R) jeweils mittlere Volumina (über alle Referenzjahre hinweg) von negativen Werten von 5% oder mehr im Verhältnis zum Volumen positiver BWS-Werte auftreten. Solche negative Werte sind deshalb problematisch, weil eine Schätzung von (2) lediglich die Eventualität positiver Werte von Y_i abzudecken vermag, da der Logarithmus einer Variablen nur für positive Werte definiert ist. Formell ausgedrückt: (2) erlaubt — unter Einbezug aller Beobachtungen i mit positiven Werten für Y_i — eine Schätzung des bedingten Erwartungswertes $E(Y_i|Y_i > 0, \dots)$.¹⁴ Anzustreben ist jedoch eine Schätzung des bedingten Erwartungswertes $E(Y_i|\dots)$, also ohne Vernachlässigung negativer Werte der Zielvariablen. Formell existiert zwischen $E(Y_i|\dots)$ und $E(Y_i|Y_i > 0, \dots)$ der folgende Zusammenhang:

$$\begin{aligned} E(Y_i|\dots) &= P(Y_i > 0|\dots)E(Y_i|Y_i > 0, \dots) \\ &\quad + P(Y_i = 0|\dots)E(Y_i|Y_i = 0, \dots) \\ &\quad + P(Y_i < 0|\dots)E(Y_i|Y_i < 0, \dots) \\ &= P(Y_i > 0|\dots)E(Y_i|Y_i > 0, \dots) \\ &\quad + P(Y_i < 0|\dots)E(Y_i|Y_i < 0, \dots) \end{aligned}$$

Die Vereinfachung ergibt sich dadurch, dass $E(Y_i|Y_i = 0, \dots)$ immer 0 ergibt.

Das heisst, für eine konsistente Vorhersage von $E(Y_i|\dots)$ ist es erforderlich, nebst für $E(Y_i|Y_i > 0, \dots)$ auch jeweils für $P(Y_i > 0|\dots)$, $P(Y_i < 0|\dots)$ und $E(Y_i|Y_i < 0, \dots)$ über Schätzwerte zu verfügen. In der Praxis kann die Eventualität $Y_i = 0$ als vernachlässigbar eingestuft werden,¹⁵ womit die folgende Vereinfachung akzeptabel erscheint:

$$E(Y_i|\dots) \approx [1 - P(Y_i < 0|\dots)] E(Y_i|Y_i > 0, \dots) + P(Y_i < 0|\dots)E(Y_i|Y_i < 0, \dots) \quad (4)$$

Deshalb wird der folgende Lösungsansatz für eine konsistente Schätzung der Variablen BWS vorgeschlagen:

- $E(Y_i|Y_i > 0, \dots)$ lässt sich, wie bereits erläutert, unter Verwendung von (2) und mit Einbezug aller Beobachtungen i mit positiven Werten für Y_i schätzen.
- $E(Y_i|Y_i < 0, \dots)$ lässt sich schätzen, indem für die Beobachtungen i mit *negativen* Werten für Y_i eine angepasste Variante von Gleichung (2), mit $\log(-Y_i) = \dots$, spezifiziert wird. Da die Anzahl der

zur Verfügung stehenden Beobachtungen (je nach Referenzjahr zwischen 330 und 541) im Vergleich mit der Schätzung von $E(Y_i|Y_i > 0, \dots)$ um ein Vielfaches kleiner ist, musste dazu eine Vereinfachungen der Spezifikation vorgenommen werden (Random-Effekte lediglich auf Stufe **noga2r** statt “genestet” auf Stufe **noga2r** und NOGA-Arten).

- $P(Y_i < 0|\dots)$ lässt sich im Rahmen eines *Generalized Linear Mixed-Effects Model* der **binomial**-Familie mit **logit** als Link-Funktion schätzen. Hier wird ebenfalls auf die erklärenden Variablen gemäss Gleichung (2) zurückgegriffen, wobei eine Vereinfachung erforderlich war (in Form der Einschränkung $\gamma_{g(i)} = -\beta_{g(i)}$; d.h. anstelle von $\dots + \beta_{g(i)}t_i + \gamma_{g(i)}l_i + \dots$ wird $\dots + \beta_{g(i)}(t_i - l_i) + \dots$ geschätzt), da das Modell sonst nicht konvergierte.
- Die schliesslich verwendete Vorhersage für die Variable BWS wird dann gemäss (4) aus diesen drei Vorhersagen ermittelt.

Für die Variablen BPW und VL wurde die Anzahl der Beobachtungen mit negativen Werten (je nach Referenzjahr um 0.1% oder weniger der Fälle) wie auch deren Volumina als zu geringfügig erachtet. Die Berechnung von Modellen für $E(Y_i|Y_i < 0, \dots)$ und $P(Y_i < 0|\dots)$ wäre in der Folge unzuverlässig und das Ausmass der statistischen Inkonsistenzen, die sich damit vermeiden liessen, vernachlässigbar. Deshalb wird für die Vorhersage dieser Variablen lediglich auf den gemäss (2) ermittelten Wert — also $E(Y_i|Y_i > 0, \dots)$ — zurückgegriffen.

5.2 Resultate

Zunächst erfolgt in Tabelle 3 eine grobe Einschätzung der Vorhersagekraft des Modells für $P(Y_i < 0|\dots)$. Dabei werden sämtliche Beobachtungen des Rahmens und insbesondere der Nettostichprobe gemäss ihren vorhergesagten Werten der Wahrscheinlichkeit einer negativen BWS in sieben Kategorien eingeteilt. Es zeigt sich, dass sich in den vier obersten Kategorien der vorhergesagten Wahrscheinlichkeit (jene mit $\hat{P} \geq 1\%$) keine Hinweise auf verzerrte Vorhersagen finden lassen: Der effektiv beobachtete Anteil der Beobachtungen mit negativer BWS liegt jeweils komfortabel in der Spanne der vorhergesagten Wahrscheinlichkeiten, die diesen Kategorien zu Grunde liegen. Anders sieht es in den folgenden zwei Kategorien aus: Für vorhergesagte Wahrscheinlichkeiten zwischen 0.3% und 1% sowie zwischen 0.1% und 0.3% kommen die effektiv beobachteten Anteile mit 0.16% und 0.01% jeweils deutlich tiefer als die untere Schwelle zu liegen. Für eine vorhergesagte Wahrscheinlichkeit von unter 0.1% wiederum ist effektiv keine negative BWS beobachtet worden, was mit der Vorhersage kompatibel ist. Dieser Befund lässt vermuten, dass beim Wahrscheinlichkeitsmodell für negative BWS noch Optimierungspotenzial besteht. Jedoch erscheint das Risiko grösserer Verzerrungen als Folge der Implementierung dieses Lösungs-

¹⁴Der besseren Lesbarkeit halber werden im vorliegenden Abschnitt Ausdrücke der Art $E(Y_i|Y_i > 0, x_i, z_i, b_{g(i)})$ jeweils mit $E(Y_i|Y_i > 0, \dots)$ (oder analog) abgekürzt.

¹⁵Nullwerte machen, in jedem Referenzjahr und für jede der hier untersuchten drei Variablen, weniger als 0.1% der Fälle aus.

ansatzes eher gering, da sich der hier quantifizierbare Vorhersagefehler im Bereich von einigen Zehntelprozentpunkten bewegt.

Tabelle 3: Anzahl der Beobachtungen (im Rahmen und erhoben¹⁶), Anzahl und Anteil der erhobenen Beobachtungen mit negativer BWS, nach Intervallen der vorhergesagten Wahrscheinlichkeit einer negativen BWS, summiert über alle Referenzjahre

Vorhergesagte Wahrsch. $BWS < 0$	Anzahl Beob.:		Erh. $BWS < 0$	
	Rahmen	Erh.	Anzahl	in %
$\hat{P} \geq 30\%$	7612	1850	988	53.41
$30\% > \hat{P} \geq 10\%$	29922	4233	871	20.58
$10\% > \hat{P} \geq 3\%$	98278	9986	616	6.17
$3\% > \hat{P} \geq 1\%$	190238	18395	299	1.63
$1\% > \hat{P} \geq 0.3\%$	386384	34627	57	0.16
$0.3\% > \hat{P} \geq 0.1\%$	514213	35859	5	0.01
$\hat{P} < 0.1\%$	48708	5849	0	0.00
Total	1275355	110799	2836	2.56

Unbefriedigend ist hingegen, dass sich R^2 als Folge des Einbezugs der Eventualität negativer BWS-Werte verschlechtert (von 0.5639 auf 0.5611). Dieses Kriterium stützt somit den hier vorgeschlagenen Lösungsansatz nicht. Eine gegenläufige Tendenz (leichte Verbesserung von 0.7358 auf 0.7372) zeigt sich im Wert von R_{abs} . In der Summe ist der Nutzen des hier vorgeschlagenen Lösungsansatzes zur Berücksichtigung negativer Werte der BWS somit unschlüssig. Da die theoretische Grundlage dafür gegeben ist, wird an diesem Lösungsansatz vorläufig festgehalten.

6 Gewährleistung von untereinander konsistenten Vorhersagen

Mit den Erkenntnissen aus den vorangegangenen drei Abschnitten lässt sich für jede der drei Zielvariablen eine favorisierte Modellvariante bestimmen. Diese bis hierhin favorisierten Varianten sind in Tabelle 1 in der ersten Spalte jeweils mit einem Stern gekennzeichnet. Es ist jedoch nicht gewährleistet, dass die Vorhersagen gemäss diesen Varianten für die drei Zielvariablen untereinander konsistent sind im Sinne der in Punkt 4 in Unterabschnitt 2.3 erläuterten Erfordernisse. Im vorliegenden Abschnitt soll deshalb untersucht werden, wie sich Vorhersagen bereitstellen lassen, die diese Erfordernisse erfüllen.

6.1 Definitiorische Gleichheit

Eine Besonderheit im Kontext der Schätzung der drei Variablen BPW, VL und BWS besteht darin, dass sich, für jede Einheit i , der Wert der letztgenannten Grösse per Definition als Saldo der ersten zwei ergibt:

$$BWS_i = BPW_i - VL_i \quad (5)$$

Werden diese drei Variablen mit dem bisher beschriebenen Modellierungsansatz isoliert voneinander geschätzt, ist nicht gewährleistet, dass die resultierenden Vorhersagen ($\widehat{BPW}_i$, $\widehat{VL}_i$, $\widehat{BWS}_i$) diesen buchhalterischen Zusammenhang erfüllen. Es ist also damit zu rechnen, dass $\widehat{BWS}_i \neq \widehat{BPW}_i - \widehat{VL}_i$. Eine naheliegende Lösung für dieses Problem würde darin liegen, für die BWS nicht auf die Vorhersage aus dem Modell für diese Variable zurückzugreifen, sondern — inspiriert aus der Definition in (5) — stattdessen einfach die Differenz der Vorhersage der beiden anderen Variablen, also $\widehat{BPW}_i - \widehat{VL}_i$ zu verwenden. Es ist jedoch nicht a priori garantiert, dass diese Lösung vorteilhafter ist als die weiteren zwei existierenden Optionen, um für eine der drei Variablen die Vorhersage aus den Vorhersagen der jeweils zwei anderen Variablen gemäss (5) abzuleiten. Deshalb werden in Tabelle 1 für alle der drei Variablen Ergebnisse der Qualitätsmasse ausgewiesen, die resultieren, wenn gemäss (5) die Vorhersage der jeweiligen Variable aus den Vorhersagen der jeweils anderen zwei Variablen bestimmt wird.

In der Tat zeigt sich, dass die Berechnung der Vorhersage für die BWS als Saldo $\widehat{BPW}_i - \widehat{VL}_i$ die unvorteilhafteste der drei Varianten ist, da dies bei dieser Zielvariablen einen Rückgang in den betrachteten Qualitätsmassen (sehr markant im R^2 , nämlich von 0.5611 auf 0.0333) zur Folge hätte. Dieser erstaunlich tiefe Wert für R^2 scheint in erster Linie durch Beobachtungen mit einer negativen vorhergesagten BWS (sprich mit höheren vorhergesagten Werten für die VL im Vergleich zum BPW) getrieben zu sein, obschon solche Vorhersagen in der Nettostichprobe mit weniger als 1.5% nicht sehr zahlreich sind (wird anstelle dieser negativen Vorhersagen ein Wert von null eingesetzt, würde ein R^2 von 0.5155 resultieren). Würde hingegen für die Vorhersage der BPW die Summe $\widehat{VL}_i + \widehat{BWS}_i$ verwendet, wäre der resultierende Rückgang in den Qualitätsmassen vergleichsweise gering (R^2 : von 0.7474 auf 0.7129, weniger ausgeprägt im R_{abs}). Schliesslich würde die Verwendung des Saldos $\widehat{BPW}_i - \widehat{BWS}_i$ als Vorhersage für die VL dazu führen, dass sich die Wert der Qualitätsmasse bei dieser Variablen deutlich verbessern würde, und zwar von 0.6982 auf 0.7519 im R^2 und wiederum etwas weniger ausgeprägt im R_{abs} . Vor diesem Hintergrund wurde die letztgenannte Variante bevorzugt und die Modellvorhersagen für die Variable VL verworfen; die zu verwendenden Vorhersagen für diese Variable werden stattdessen unmittelbar als Saldo aus den Vorhersagen des BPW und der BWS ermittelt.

¹⁶Im Kontext dieser Tabelle wurden jene Beobachtungen, welche dem “ex post” konstruierten Rahmen nicht mehr angehören (z.B. weil deren Beschäftigtenzahl zwischenzeitlich unter drei gefallen ist), nicht mitgezählt. Somit fällt das Total der Anzahl erhobener Beobachtungen hier leicht geringer aus als das Total in Tabelle 2.

6.2 Einheiten mit $\widehat{BWS}_i > \widehat{BPW}_i$

Bei der bisher vorgeschlagenen separaten Modellierung der Variablen können sich Situationen ergeben mit Vorhersagen mit einem tieferen Wert für die Variable BPW als für die Variable BWS. Unter Verwendung der favorisierten Spezifikation des Modells machen solche Situationen, je nach Referenzjahr, zwischen 0.5% und 0.8% der Fälle aus. Dies würde, gemäss der in (5) gegebenen buchhalterischen Gleichheit, eine negative Vorhersage für die Variable VL implizieren. Wie in Abschnitt 5.1 gezeigt wurde, sind negative Werte für die Variable VL in den erhobenen Daten jedoch äusserst selten zu beobachten. Es scheint deshalb vernünftig, solche Situationen nicht zuzulassen, und in diesen Fällen stattdessen eine Vorhersage von null für die VL und somit einen identischen Wert für die Vorhersagen von BPW und BWS zu erzwingen.

Um zu entscheiden, welcher Wert für die Vorhersagen für BPW und BWS in diesen Fällen verwendet werden soll, wurde sowohl die Variante mit Einsetzung von $\widehat{BWS}$ nach $\widehat{BPW}$ (was, für jede Beobachtung i , der Verwendung des Maximalwerts aus $\widehat{BPW}_i$ und $\widehat{BWS}_i$ zur Vorhersage des BPW entspricht), als auch die umgekehrte Variante (Verwendung des Minimalwerts aus $\widehat{BPW}_i$ und $\widehat{BWS}_i$ zur Vorhersage der BWS) evaluiert. Es zeigt sich, dass sich die resultierenden Variationen der Qualitätsmasse im Bereich einiger Hunderttausendstel bewegen. Während mit der ersten Variante eine geringfügige Verbesserung der Qualitätsmasse für die Vorhersage des BPW resultiert, führt die zweite Variante zu einer ebenso geringfügigen Verschlechterung für die Vorhersage der BWS. Somit wird die erste Variante (Verwendung des Maximalwerts aus $\widehat{BPW}_i$ und $\widehat{BWS}_i$ zur Vorhersage des BPW) bevorzugt.

7 Zusammenfassung und abschliessende Bereinigungen

Die in den Abschnitten 3 bis 6 erzielten Erkenntnisse erlauben, für die beiden Variablen BPW und BWS die jeweils **favorisierten Modellvarianten** zu identifizieren (in der ersten Spalte der Tabelle 1 mit fetter Schrift gekennzeichnet). Insbesondere hat sich herausgestellt, dass es — zur Gewährleistung der buchhalterischen Gleichung (5) und auch motiviert durch das verwendete Qualitätsmass — optimal ist, als Vorhersage für die Variable VL den Saldo der favorisierten Vorhersagen der beiden erstgenannten Variablen zu verwenden.¹⁷

¹⁷Die für die Variable VL zu verwendende Vorhersage berechnet sich folglich als Saldo zwischen der gemäss Unterabschnitt 6.2 favorisierten Vorhersage für den BPW (sprich dem Maximalwert aus $\widehat{BPW}^*$ und $\widehat{BWS}^*$) und der favorisierten Vorhersage für die BWS (sprich $\widehat{BWS}^*$), und nicht (wie zum Zweck der Evaluation in Unterabschnitt 6.1) direkt als Saldo zwischen $\widehat{BPW}^*$ und $\widehat{BWS}^*$. Wie in Tabelle 1 ersichtlich ist, ändert sich das resultierende Qualitätsmass für die Variable VL dadurch praktisch nicht.

Der Erklärungsgehalt der für die Modellierung herangezogenen Hilfsvariablen (MWST-Umsatz und AHV-Arbeitseinkommen) fällt, im Vergleich zur Zielvariablen Umsatz gemäss WS, für die drei Zielvariablen BPW, VL und BWS tatsächlich geringer aus. Dies zeigt sich in den Werten für R^2 , die für den WS-Umsatz über 0.86 liegen, in den bevorzugten Varianten für die Variablen BPW und VL hingegen nur bei 0.75 und für die BWS gar nur bei 0.56. Insgesamt erfreulicher gestalten sich die Resultate für R_{abs} : Mit Werten von 0.78 wird der BPW im Vergleich zum WS-Umsatz (0.83) in dieser Betrachtung ziemlich gut erklärt, und auch die Vorhersage der VL und der BWS erscheint mit Werten von 0.71 bzw. 0.74 unter diesem Licht akzeptabel.

Wie schon bei der Produktion des Umsatzes wurden für die favorisierten Vorhersagen der beiden Variablen BPW und BWS **Einzelfälle von Beobachtungen vertieft untersucht**, die nicht in der Nettostichprobe der WS enthalten sind und deren Vorhersagen einen hohen Einfluss auf die Totalwerte ihrer **noga2r**-Branchen ausüben. Konkret wurden jene Einheiten ausserhalb der Nettostichprobe analysiert, für die in einem oder mehreren Referenzjahren die vorhergesagten Werte des BPW oder der BWS (a) mindestens CHF 10 Millionen betragen und (b) 10% oder mehr der entsprechenden Totalwerte (hochgerechnet gemäss WS) innerhalb der **noga2r**-Branche ausmachen.¹⁸ Diese Analyse ermöglichte, unter den 16 die Filterkriterien erfüllenden Einheiten deren zehn zu identifizieren, für die in einem oder mehreren Referenzjahren die Vorhersagen für BPW oder BWS als unplausibel zu erachten waren und stattdessen auf die Vorhersage anderer Modellvarianten zurückgegriffen wurde:

- Für neun Einheiten wurden in einem oder mehreren Referenzjahren jeweils für beide Zielvariablen die Vorhersagen des entsprechenden Teilmodells ohne Berücksichtigung der MWST-Umsätze eingesetzt;
- für eine Einheit wurde in einem Referenzjahr für den BPW die Vorhersage des Teilmodells mit ausschliesslicher Berücksichtigung des MWST-Umsatzes und für die BWS ein Wert von null eingesetzt.

Die mutmasslichen Ursachen für die angezweifelte Eignung der favorisierten Modellvarianten in diesen Fällen waren vielfältig (z.B. Erhebungsfehler beim MWST-Umsatz, zu hohe verteilte MWST-Umsätze in Gruppen mit untypischer Branchenzusammensetzung, problematische Änderungen der NOGA-Einteilung der Einheit im Zeitverlauf) und lieferten in der Summe keine Hinweise auf systematische Schwächen der verwendeten Modelle.

Wie schon bei der schlussendlich für den Umsatz verwendeten Variante wurde darauf verzichtet, Bereinigungen (oder anderwertige spezifische Behandlungen) von

¹⁸Bedingung (a) ist deshalb sinnvoll, weil in einzelnen **noga2r**-Branchen und Referenzjahren die hochgerechnete BWS gemäss WS nahe bei null zu liegen kommt, wodurch die alleinige Anwendung von Bedingung (b) in einer zu hohen Zahl von zu untersuchenden Fällen führen würde.

Extremwerten zum Zweck der Modellschätzungen und -evaluationen vorzunehmen.

Um den Anforderungen des Datenschutzes gerecht zu werden, sind für die Publikation von Resultaten die folgenden Kriterien massgebend: Es werden in den Tabellen keine Werte ausgewiesen für Zellen, in denen weniger als vier Einheiten von null abweichende Werte aufweisen, oder in denen der Absolutwert einer einzigen Beobachtung 60% oder mehr des Totals der Absolutwerte aller Beobachtungen beträgt. Diese Einschränkungen kommen bereits für die Publikation des Umsatzes zur Anwendung und haben sich nicht als übermässig restriktiv herausgestellt: Für die grosse Mehrheit der Zellen ist es möglich, für die drei Zielvariablen in der STAGRE Resultate nach den eingangs in Unterabschnitt 1.1 erwähnten Gliederungskriterien zur Verfügung zu stellen.

8 Ausblick auf weiterführende Arbeiten und Analysen

Die bis hierhin umgesetzten und dokumentierten Arbeiten zeigen auf, dass die modellgestützte Schätzung der Variablen BPW, VL und BWS ein vielversprechender Ansatz darstellt, um im Rahmen der Statistik der Unternehmensgruppen Totalschätzungen dieser Variablen bereitzustellen, die zudem nach verschiedenen weiteren Untersuchungsbereichen aufgeschlüsselt werden können. Mehrere Aspekte dieses Ansatzes erfordern jedoch noch weiterführende Analysen, um sicherzustellen, dass diese Ergebnisse den geforderten Qualitätskriterien entsprechen und gleichzeitig auch bezüglich ihrer Aussagekraft den potentiellen Nutzerinnen und Nutzern dieser Statistik einen Mehrwert erbringen. Die folgende Auflistung fasst diese für weiterführende Analysen vorgesehenen Aspekte zusammen:

- Eine nach verschiedenen Untersuchungsbereichen differenzierte Betrachtung der Qualitätsmasse. Dimensionen, gemäss derer diese Betrachtung erfolgen soll, sind insbesondere:
 - Branchen,
 - Zugehörigkeit zur STAGRE-Population oder nicht,
 - Referenzjahre,
 - Grössenklassen.

Der Zweck einer solchen differenzierten Betrachtung besteht insbesondere darin, auszuschliessen, dass einzelne der genannten Untersuchungsbereiche von Verzerrungen oder übermässig hoher Varianz der Modellvorhersagen betroffen sind, was durch die bis hierhin erfolgte gesamthafte Betrachtung der Qualitätsmasse nicht gewährleistet werden kann. Ergänzend könnten geeignete Residuenplots Hinweise darauf liefern, ob zwischen verschiedenen Untersuchungsbereichen (insbesondere: Zugehörigkeit

zur STAGRE-Population oder nicht) problematische Abweichungen in der Verteilung bestehen.

- Überprüfung der Kohärenz mit gewissen direkten Schätzungen der WS (z.B. Totale auf Gliederungsstufe BFS-50 der WS-Population), wobei insbesondere der Zeitraum nach Revision der WS von Interesse ist.¹⁹
- Im Kontext der Schätzung der Teilmodelle und der Berechnung deren Vorhersagen (die in Unterabschnitt 2.3 beschriebenen Schritte 1 und 2):
 - Weiterführende Analysen und Überlegungen theoretischer Art zu einem allfälligen Einbezug der Hochrechnungsgewichte;
 - Analysen zum Auftreten von Extremwerten und der dadurch potentiell aufgeworfenen Probleme;
 - Analysen zur Verteilung der Schätzfehler (z.B. mittels grafischer Mittel wie QQ-Plots) und, falls erforderlich, Entwicklung von Strategien zum Umgang mit allfälligen Verletzungen der postulierten Normalitätsannahmen;
 - Untersuchung möglicher Alternativen zur verwendeten Rücktransformation der logarithmierten Vorhersagen in absolute Werte (Gleichung 3);
 - Untersuchung möglicher Alternativen zum verwendeten Wahrscheinlichkeitsmodell für negative BWS (Abschnitt 5).
- Entwicklung von Methoden, die eine verlässliche Einschätzung der Varianz (beispielsweise mittels Bootstrap) erlauben.
- Überlegungen zur Möglichkeit, die Modellvorhersagen so zu kalibrieren, dass sie besser in Übereinstimmung mit den im Produktionskonto der Volkswirtschaftlichen Gesamtrechnung (VGR) publizierten Ergebnisse stehen.²⁰
- Differenzierung zwischen den Einkommen der Angestellten und der Selbstständigen in der Hilfsvariable Arbeitseinkommen. Die AHV-Ausgleichskassen, die

¹⁹Seit der Revision der Wertschöpfungsstatistik (umgesetzt ab dem Referenzjahr 2021) werden, nebst der vollzeitäquivalenten Beschäftigung, neu auch die MWST-Umsätze als Kalibrierungsvariable verwendet. Somit beruhen die Ergebnisse der WS ab 2021 auf einer Datengrundlage, die dem hier verwendeten Modellierungsansatz sehr ähnlich ist. Infolgedessen erscheint es sinnvoll, zu untersuchen, ob statistisch signifikante Abweichungen der hier erzielten Ergebnisse mit jenen der WS ab 2021 bestehen oder nicht.

²⁰Als Synthesestatistik wird die VGR nicht ausschliesslich auf der Basis einer einzelnen Quelle wie der WS erstellt. Vielmehr werden die von der WS hochgerechneten Resultate unter Berücksichtigung einer Vielzahl weiterer Quellen ergänzt und plausibilisiert; das heisst, die Resultate der WS erfahren bei Bedarf Korrekturen, bevor sie Eingang in das Produktionskonto finden. Deshalb kann die Vergleichbarkeit der im Rahmen des vorliegenden Projekts erhaltenen Ergebnisse mit jenen der VGR nicht ohne weiteres gewährleistet werden.

als Quelle für diese Hilfsvariable dienen, unterscheiden zwischen diesen zwei Kategorien von Arbeitseinkommen.²¹ Theoretische Überlegungen sowie bereits durch die Sektion VGR erzielte Erfahrungen legen nahe, dass der Zusammenhang zwischen den Zielvariablen und den Arbeitseinkommen nicht einheitlich ist für die beiden Kategorien. Eine geeignete Berücksichtigung eines solchen nicht einheitlichen Zusammenhangs könnte die Präzision der Vorhersagen erhöhen, da im Rahmen der WS der Anteil von Einheiten, die Selbstständigeneinkommen entrichten, immerhin ungefähr 20% beträgt.²²

- Analysen zur Möglichkeit, verlässliche Vorhersagen für Einheiten mit weniger als drei Beschäftigten zu produzieren.

²¹Siehe <https://www.bsv.admin.ch/bsv/de/home/sozialversicherungen/ueberblick/beitraege.html> für mehr Informationen.

²²Allerdings ist in der WS-Nettostichprobe der Anteil solcher Einheiten mit unter 10% weniger als halb so hoch als im Rahmen. Grund dafür ist, dass die Einheiten der kleinsten Grössenklasse mit einer geringen Ziehungsrates in die Stichprobe aufgenommen werden; Selbstständigeneinkommen werden jedoch im Wesentlichen nur in kleinen und sehr kleinen Unternehmen entrichtet. Dieser relativ geringe Anteil solcher Einheiten in der Nettostichprobe stellt eine bedeutende Herausforderung für eine differenzierte Berücksichtigung der Selbstständigeneinkommen in der Schätzung dar.