OJA-Index

Ein mehrstufiges Verfahren zur Berechnung von Konjunkturindizes aus Online-Stellenanzeigen



DESTATIS
Statistisches Bundesamt

BertelsmannStiftung

© Bertelsmann Stiftung, Gütersloh
November 2026

Herausgeber
Bertelsmann Stiftung
Carl-Bertelsmann-Straße 256
33311 Gütersloh
www.bertelsmann-stiftung.de

Verantwortlich
Gunvald Herdin, Senior Project Manager, Bertelsmann Stiftung

Autor:innen
Dr. Martina Rengers, Statistisches Bundesamt, Destatis
Johannes Müller, &effect data solutions GmbH

Datenarbeit
Dr. Martina Rengers, Statistisches Bundesamt, Destatis
Johannes Müller, &effect data solutions GmbH

Datenquelle
Textkernel BV

Lektorat
Jan Haas

Zitationshinweis
Rengers, M. und Müller, J. (2026). OJA-Index – Ein mehrstufiges Verfahren zur Berechnung von Konjunkturindizes aus Online-Stellenanzeigen. Bertelsmann Stiftung

Layout und Datenvisualisierung
Linda Wedi

Titelgestaltung
Linda Wedi

DOI 10.11586/2026134

ID_3217

OJA-Index

Ein mehrstufiges Verfahren zur Berechnung von Konjunkturindizes aus Online-Stellenanzeigen

BertelsmannStiftung

Inhaltsverzeichnis

Zusammenfassung ... 6

Abstract ... 7

1. Das Ziel der Veröffentlichung von Online-Stellenindizes ... 8

2. Daten und Datenqualität von Online-Stellenanzeigen ... 11
   2. 1 Coverage-Probleme ... 11
   2.2 Dubletten, Stock versus Flow und technische Aspekte ... 11
   2.3 Fake-Stellenanzeigen, Personalvermittlung und Zeitarbeit ... 12
   2.4 Benchmarks ... 13

3. Die Textkernel-Daten ... 15

4. Ein mehrstufiges Verfahren zur Berechnung von OJA-Indizes ... 19
   4.1 Stock ... 19
   4.2 Ensemble-Ansätze und Ausreißerbereinigung ... 19
      4.2.1 Die Idee von Ensemble-Ansätzen ... 19
      4.2.2 Ausreißerbereinigung der absoluten Zahlen ... 20
      4.2.3 Winsorisierung der Änderungsrate ... 20

5. Bewertung der Ergebnisse ... 22
   5.1 Vorgehensweise ... 22
   5.2 Korrelationskoeffizienten und Wendepunktanalyse ... 22
   5.3 Volatilitätsmaße und grafische Vergleichsanalyse ... 25

6. Fazit und Ausblick ... 31

Literaturverzeichnis ... 33

Abbildungsverzeichnis

ABBILDUNG 1 **Textkernel-Daten: neue Online-Stellenanzeigen pro Woche vor und nach Deduplizierung 17**

ABBILDUNG 2 **Textkernel-Daten: pro Woche erfasste Online-Stellenanzeigen eines großen deutschen Portals ... 18**

ABBILDUNG 3 **Textkernel-Daten: Outlier-Bereinigung am Beispiel eines großen Portals 21**

ABBILDUNG 4 **Textkernel-Daten: OJA-Indizes unter verschiedenen Winsor-Raten 23**

ABBILDUNG 5 **Wendepunktanalyse bei winsorisierten Textkernel-Daten im Vergleich zu BA- und IAB-Daten 24**

ABBILDUNG 6 **OJA-Indizes: Textkernel-Daten, Outlier und Winsor15 im Vergleich zu BA-X und IAB 28**

ABBILDUNG 7 **OJA-Indizes: Textkernel-Daten mit Winsor10, 15 und 20 im Vergleich zu BA-X und IAB 29**

ABBILDUNG 8 **OJA-Indizes: Textkernel-Daten, BA-X, Index.de und Indeed im Vergleich zu IAB 30**

Tabellenverzeichnis

TABELLE 1 **Textkernel-Daten: Art der Quellen und Anteil der erfassten Stellenanzeigen (einschließlich Dubletten) 15**

TABELLE 2 **Textkernel-Daten: Anzahl der Stellenanzeigen und Dubletten 16**

TABELLE 3 **Korrelationskoeffizienten der Veränderungsraten der Indizes im Vergleich zu BA- und IAB-Daten 23**

TABELLE 4 **Volatilitätsmaße für verschiedene OJA-Indizes aus Textkernel-Daten 26**

TABELLE 5 **Volatilitätsmaße für OJA-Indizes aus Textkernel-Daten im Vergleich zu BA-X, Index.de und Indeed 27**

TABELLE 6 **Ausmaß des Corona-Bruchs bei verschiedenen Indizes (in %) 30**

Zusammenfassung

Die Covid-19-Krise hat den bereits bestehenden Druck hinsichtlich der Bereitstellung hochfrequenter und zeitnah verfügbarer Daten deutlich verstärkt. In diesem Kontext eröffnet die internetbasierte Erhebung von Daten zu Online-Stellenanzeigen (Online Job Advertisements – OJA) das Potenzial, die Lage des Arbeitsmarktes nahezu in „Echtzeit“ abzubilden. Durch die Konstruktion eines monatlichen Index zur (konjunkturellen) Entwicklung von Online-Stellenanzeigen können diese Echtzeitinformationen systematisch analysiert und aufbereitet werden. Auf diese Weise lassen sich zusätzliche Erkenntnisse generieren, die über die etablierten Statistiken zur Arbeitskräftenachfrage hinausgehen.

Gleichzeitig zeigen zahlreiche methodische Studien der jüngeren Zeit, dass die Erstellung monatlicher (oder sogar wöchentlicher) Indizes auf Basis von OJA-Daten zur Abbildung der kurzfristigen Arbeitskräftenachfrage mit erheblichen Herausforderungen verbunden ist und bislang häufig nur eingeschränkt gelingt. Diese Schwierigkeiten erwachsen in erster Linie aus spezifischen Limitationen von OJA-Daten, insbesondere hinsichtlich ihrer Repräsentativität, technisch bedingter Artefakte sowie der Problematik der Deduplizierung.

Der vorliegende Beitrag begegnet diesen Herausforderungen, indem er ein mehrstufiges Verfahren zur Berechnung von Konjunkturindizes aus Online-Stellenanzeigen vorstellt. Dieses Verfahren trägt zentralen Anforderungen an Konjunkturindikatoren Rechnung, die gewährleisten sollen, dass beobachtete Volatilitäten nicht durch methodisch-technische Effekte verzerrt und zugleich konjunkturelle Wendepunkte sowie Trends in der Entwicklung der Arbeitskräftenachfrage bestmöglich identifiziert werden. Die methodische Grundlage bildet eine australische Studie, in der ein Algorithmus entwickelt wurde, der die Nutzbarkeit von Online-Stellenanzeigen zur Abbildung kurzfristiger und hochaktueller Entwicklungen des Arbeitsmarktes verbessert.

Das Verfahren basiert auf einem sogenannten Ensemble-Ansatz, bei dem Stellenanzeigen aus unterschiedlichen Quellen zunächst separat verarbeitet und erst anschließend aggregiert werden. Dieses Vorgehen unterscheidet sich grundlegend von sonst üblichen Methoden, bei denen Daten zunächst quellenübergreifend zusammengeführt und dann anschließend gemeinsam verarbeitet werden, beispielsweise durch eine einheitliche Glättung aller Stellenanzeigen über sämtliche Quellen hinweg.

Insgesamt zeigen die Ergebnisse, dass auf OJA-Daten basierende Indizes nicht nur als komplementäres Instrument, sondern in spezifischen Anwendungskontexten auch als eigenständiges Analysewerkzeug für den Arbeitsmarkt eingesetzt werden können. Insbesondere die Kombination von OJA-Daten mit traditionellen Erhebungsverfahren bietet ein erhebliches Potenzial zur Ermöglichung einer umfassenderen, dynamischeren und zeitnäheren Analyse des Arbeitsmarktes. Zukünftige Forschungsarbeiten sollten sich daher auf die Entwicklung hybrider Modelle sowie auf die weitere methodische Optimierung der OJA-Datenanalyse konzentrieren, um deren volles Potenzial für die Arbeitsmarktforschung auszuschöpfen. Zugleich erfordern die mit OJA-Daten verbundenen Herausforderungen eine kontinuierliche Weiterentwicklung der methodischen Ansätze, um die Aussagekraft und Qualität der Ergebnisse langfristig sicherzustellen und weiter zu verbessern.

Zusatzinformation

Die Bertelsmann Stiftung arbeitet seit 2021 mit dem Anbieter Textkernel für die Analysen auf jobmonitor.de. Ausschlaggebend dafür war unter anderem die gute Datenverfügbarkeit für Deutschland.

Abstract

The COVID-19 crisis has significantly intensified the existing pressure to provide high-frequency and timely data. In this context, the internet-based collection of data on online job advertisements (OJA) offers the potential to map the state of the labour market almost in 'real time'. By constructing a monthly index of the (cyclical) trends in online job advertisements, this real-time information can be systematically analysed and processed. In this way, additional insights can be generated that go beyond the established statistics on labour demand.

At the same time, numerous recent methodological studies show that the creation of monthly (or even weekly) indices based on OJA data to reflect short-term labour demand is associated with considerable challenges and has so far often been only partially successful. These difficulties stem primarily from specific limitations of OJA data, particularly regarding its representativeness, technical artefacts and the issue of deduplication.

This paper addresses these challenges by applying a multi-stage procedure for calculating economic indices from online job advertisements. This procedure takes into account key requirements for economic indicators, which are designed to ensure that observed volatility is not distorted by methodological and technical effects, whilst at the same time identifying economic turning points and trends in the development of labour demand as effectively as possible. The methodological basis is an Australian study in which an algorithm was developed to improve the usability of online job advertisements for mapping short-term and highly topical developments in the labour market.

The method is based on a so-called ensemble approach, in which job advertisements from different sources are first processed separately and only subsequently aggregated. This approach differs fundamentally from conventional methods, in which data is first consolidated across sources and then processed collectively, for example by applying a uniform smoothing process to all job advertisements across all sources.

Overall, the results show that indices based on OJA data can be used not only as a complementary instrument but also, in specific application contexts, as a standalone analytical tool for the labour market. In particular, combining OJA data with traditional survey methods offers significant potential for enabling a more comprehensive, dynamic and timely analysis of the labour market. Future research should therefore focus on the development of hybrid models and on further methodological optimisation of OJA data analysis in order to realise its full potential for labour market research. At the same time, the challenges associated with OJA data require continuous refinement of methodological approaches to ensure and further improve the validity and quality of the results in the long term.

Additional Information

The Bertelsmann Foundation has been working with Textkernel since 2021 to conduct analyses for jobmonitor.de. One of the key factors in this decision was the good availability of data for Germany.

1. Das Ziel der Veröffentlichung von Online-Stellenindizes

Die Covid-19-Krise hat den vorhandenen Druck hinsichtlich der Bereitstellung von hochfrequenten und hochaktuellen Daten erhöht. Die internetbasierte Erfassung von Daten zu Online-Stellenanzeigen (Online Job Advertisements – OJA) bietet dabei das Potenzial einer schnellstmöglichen Abbildung der Arbeitsmarktlage in „Echtzeit". Mit einem monatlichen Index zur (konjunkturellen) Entwicklung von Online-Stellenanzeigen können diese echtzeitbezogenen Informationen zum Online-Stellenmarkt zeitnah analysiert und aufbereitet werden. Darüber hinausgehende regionale Differenzierungen sowie berufsgruppen- oder wirtschaftszweigspezifische Auswertungen des Online-Stellenmarktes liefern neue kurzfristige Indikatoren, die aktuelle und politikrelevante Entwicklungen in Bezug auf Veränderungen der Arbeitskräftenachfrage darstellen. Denkbar sind beispielsweise Frühwarnungen in Bezug auf drohende Engpässe, Veränderungen in den Qualifikationsanforderungen oder regionale Konzentrationen. Damit können Informationen bereitgestellt werden, die über die bisherigen etablierten Statistiken zur Arbeitskräftenachfrage hinausgehen.

Dies erkannten auch Unternehmen, die entweder als Jobportalbetreiber selbst oder als kommerzielle Drittanbieter solche Daten erzeugen. So veröffentlichten erste Unternehmen während der Covid-19-Pandemie – unabhängig von ihrem originären Geschäftsfeld – frei zugängliche, sehr zeitnahe Informationen zur kurzfristigen Entwicklung des Online-Stellenmarktes. Der Trend zu solchen Veröffentlichungen ist bis heute ungebrochen und wird sogar um zusätzliche private Datenanbieter erweitert.

Online-Stellenanzeigen sind damit in den letzten Jahren zu einer wertvollen Quelle für die Analyse von Arbeitsmarktdynamiken geworden. Sie bieten Einblicke in aktuelle Trends und Veränderungen im Einstellungsverhalten von Unternehmen. Ein zentraler Aspekt hierbei ist die Anzahl der offenen Stellen, die häufig als Vorlaufindikator für die Entwicklung von Neueinstellungen verstanden wird und für die Online-Stellenanzeigen als hochaktueller Proxy verwendet werden könnten. In Phasen wirtschaftlichen Wachstums steigt in der Regel die Anzahl der Neueinstellungen, während sie in wirtschaftlichen Abschwungphasen tendenziell sinkt. Die Annahme, dass es einen starken Zusammenhang zwischen offenen Stellen und Neueinstellungen gibt, wird von unterschiedlichen Studien gestützt (Bleninger u. a. 2012; Kettner/Klinger 2012; Napierala 2023).

Eine Analyse der Auswirkungen der Covid-19-Pandemie und insbesondere des ersten Shutdowns zeigt, dass die Anpassungen der Arbeitskräftenachfrage weniger über Entlassungen als vielmehr „in erster Linie über Veränderungen des Einstellungsverhaltens und damit der Ausschreibung offener Stellen zu beobachten sind" (Bossler u. a. 2020b, S. 11). In diesem Zusammenhang können Online-Stellenanzeigen dabei helfen, auch berufs- und regionalspezifische Meldequoten zu schätzen und diese kurzfristiger als bisher verfügbar zu machen (Burstedde u. a. 2020, S. 44).

Die Erfassung von OJA ermöglicht es, zeitnah und detailliert Informationen über die Nachfrage nach Arbeitskräften zu sammeln. Diverse Studien konnten zeigen, dass Daten zu Online-Stellenanzeigen hierfür genutzt werden können (USA: Azar u. a. 2020; Slowakei: Štefánik/Lyócsa/Bilka 2022; Australien: Evans u. a. 2023).

Bereits vor der Covid-19-Pandemie war das Potenzial von OJA-Daten zur schnellstmöglichen Abbildung der wirtschaftlichen Lage bekannt. Rengers (2018b) liefert eine (unvollständige) Bestandsaufnahme von weltweiten Beispielen öffentlich-rechtlicher Institutionen und Firmen, die teilweise schon lange vor der Corona-Krise Online-Stellenanzeigen in der Arbeitsmarkt- und Konjunkturberichterstattung verwendeten und daraus veröffentlichungsfähige Indikatoren entwickelten.[1]

1 Ein Update dieser Liste findet man bei Rengers (2020).

Exemplarisch kann hier Australien genannt werden, wo die australische Regierung einen Index der monatlichen Online-Stellenanzeigen basierend auf den drei größten Jobportalen Australiens (SEEK, CareerOne und Workforce Australia online job boards) veröffentlicht. In einem monatlichen Report[2] wird die zeitliche Entwicklung ab dem Jahr 2006 dargestellt; es handelt sich damit um die längste kontinuierliche Zeitreihe für Stellenanzeigen in Australien. Dabei werden nicht nur ein Internet Vacancy Index (IVI), sondern auch monatliche Absolutzahlen zu den neu eingegangenen Stellenanzeigen innerhalb eines Monats ausgewiesen. Zusätzliche Angaben sind die prozentuale Veränderung gegenüber dem Vormonat und Vorjahr. Eine Differenzierung erfolgt darüber hinaus nach Berufen, Qualifikationsanforderungen und 37 Regionen.

Noch während der Corona-Krise und auch danach sind weltweit weitere OJA-datenbasierte Statistikberichterstattungen hinzugekommen. Dazu zählen Unternehmen wie Lightcast (ehemals Burning Glass) und Textkernel, aber auch die Jobportalanbieter Indeed und LinkedIn (de Lazzer/Rengers 2021). Seit September 2023 stellt auch die index-Gruppe einen hochaktuellen Stellenmarktindikator zur Verfügung.

Im Bereich der amtlichen Statistik stellte beispielweise das Statistikamt des Vereinigten Königreichs (ONS) experimentelle Statistiken unter dem Label „Real time indicators of ONS“ zur Entwicklung von Online-Stellenanzeigen von 2021 bis 2024 zunächst in einer wöchentlichen Periodizität dar, differenziert nach Branchenkategorien und Regionen mithilfe OJA-Daten von Adzuna UK (siehe ONS 2021).[3] Seit 2025 werden – weniger prominent – OJA-Daten der Firma Textkernel als „early experimental data“ beziehungsweise „official statistics in development“ zur Verfügung gestellt (siehe ONS 2024).

Das Interesse an der Entwicklung monatlicher (oder sogar wöchentlicher) Indizes aus OJA-Daten ist insbesondere durch die Corona-Krise deutlich gestiegen. Gleichzeitig weisen aber viele methodische Untersuchungen zu OJA-Daten der jüngeren Vergangenheit darauf hin, dass eine solche Indexbildung zur konjunkturellen Abbildung mit großen Herausforderungen verbunden ist und bislang tatsächlich häufig scheitert:

- „Vergleiche der Ergebnisse verschiedener OJA-Daten zeigten deutliche Unterschiede. So wurde aus den CEDEFOP-Daten ein monatlicher Online-Stellenindex den vergleichbar gemachten Ergebnissen von Indeed und der Firma Textkernel gegenübergestellt. Hier zeigten die CEDEFOP-Daten eine höchst volatile Zeitreihe, die im Gegensatz zu den anderen OJA-Daten konjunkturelle Einbrüche wie die in der Pandemie zum einen schwer erkennbar machte, diese jedoch zum anderen im Ausmaß deutlich überzeichneten.“ (Rengers/Hornik 2024, S. 52)

- „Es zeigt sich, dass bei der Berechnung von Indizes zum Online-Stellenmarkt die Wahl der Quelle(n) und die Art der Datenerhebung starke Auswirkungen auf den Verlauf haben. Eine große Herausforderung besteht also darin, Indexausschläge durch ökonomische Veränderungen von Ausschlägen durch Veränderungen im Datenerhebungs- und Aufbereitungsprozess zu unterscheiden.“ (de Lazzer/Rengers 2021, S.85)

- „Auch für [die amtliche Statistik] Deutschland[s] bleibt die Frage, ob ein Rückgriff auf die Datenzusammenstellungen kommerzieller Anbieter eine Option für die Gewinnung von Online-Stellenanzeigen sein kann. Zwar würde man damit rechtliche und technische Schwierigkeiten des Web Scrapings umgehen, die Beurteilung der Datenqualität ist jedoch bedeutend schwieriger. Dazu wäre zum einen die Aneignung eines tiefergehenden Wissens über Big-Data-Technologien generell erforderlich, das die Funktionsweisen von Algorithmen und Technologien wie Web Scraping, Machine Learning oder Text Mining umfasst. Zum anderen wäre erforderlich, dass kommerzielle Firmen der amtlichen Statistik gegenüber die „Black Box“ öffnen und unter anderem folgende Angaben offenlegen: die Auswahl der Jobportale, die eingesetzten Algorithmen zur Dublettenbereinigung, das genaue Vorgehen beim Einsatz von Techniken des Maschinellen Lernens oder Text Minings zur Erzeugung generierter Variablen.“ (Rengers 2018b, S.31)

- „Limitations of Adzuna data:
The number of job adverts being posted is not a direct measure of labour force demand; the number could respond to other changes such as how po-

2 Siehe https://www.jobsandskills.gov.au/data/internet-vacancy-index.

3 Adzuna UK veröffentlich einen eigenen „The Adzuna Job Market Report“ unter https://www.adzuna.co.uk/adzuna-intelligence/job-market-report, der nach namentlicher Anmeldung zum Download bereitsteht.

sitions are recruited for (that is, decreased activity from recruitment agencies could lead to decreased duplication of multiple adverts for one post).
Job adverts may not be removed from online job vacancy boards immediately when the position is filled so the indices may not fully reflect companies who have halted active recruitment. Note that Adzuna performs data cleaning to remove adverts that have not been observed as live for 30 days.
The data are compiled from multiple job vacancy boards and adverts are considered 'live' if the posting is still live on any board, even when it has already been removed from an alternative source.
The scope of online job adverts does not fully capture the scope of UK economic activity because of differing advertising methods, for example, casual work may be advertised by word-of-mouth or in shop windows as opposed to online.
There are points in the time series where we know there are increased levels of duplication in the dataset that are resulting in a potentially inflated value of job adverts; where this is the case, footnotes were added to the data tables to identify these anomalies." (ONS 2021, S. 6 [im Original als Auflistung dargestellt])

- ONS 2024 unterteilt das Kapitel zu Stärken und Schwächen ihrer OJA-Daten von Textkernel mit folgenden Überschriften: (1) Revisions, (2) Coverage (3) Deduplication (4) Future improvements in coverage (5) Data consistency (6) Standard Occupation Classification SOC allocation: matching, (7) Standard Occupation Classification SOC allocation: method performance (8) Similarity metric for understanding accuracy (9) Future improvements of occupation assignment (10) Geography allocation. (ONS 2024, S. 6–13)

Der vorliegende Beitrag stellt ein mehrstufiges Verfahren zur Berechnung von Konjunkturindizes aus Online-Stellenanzeigen vor. Dieses trägt den allgemeinen Anforderungen an Konjunkturindikatoren in dem Sinne Rechnung, dass diese in ihrer Volatilität nicht durch technisch-methodische Effekte überzeichnet sind und (konjunkturelle) Wendepunkte und Trends in der Arbeitskräftenachfrageentwicklung identifizieren können. Basis dieses Verfahrens ist die Studie von Evans u. a. (2023) von der staatlichen Behörde Australiens für wissenschaftliche und industrielle Forschung CSIRO[4], in der ein Algorithmus entwickelt wird, der die Nutzbarkeit von Online-Stellenanzeigen zur Abbildung zeitnaher, hochaktueller Entwicklungen des Stellenmarktes verbessert.

Zunächst werden in Kapitel 2 die wichtigsten Aspekte benannt, die einen Einfluss auf die Qualität von OJA-Daten haben, und erläutert, welche Herausforderungen damit einhergehen. Gleichzeitig werden Referenzgrößen vorgestellt, die als Benchmark herangezogen werden können, um die Datenqualität und Geeignetheit von Online-Stellenanzeigen zur Abbildung kurzfristiger Arbeitsmarktsentwicklungenzu beurteilen. Im nachfolgenden Kapitel 3 werden sodann die in dieser Studie verwendeten OJA-Daten der Firma Textkernel betrachtet und einige deskriptive datensatzbeschreibende Charakteristiken dargestellt. Kapitel 4 und 5 evaluieren, ob und wie Daten aus Online-Stellenausschreibungen generelle Trends widerspiegeln, mit welchen Methoden die Anzahl offener Stellen geschätzt wird und ob ein so gebildeter Index bestehende Trendindikatoren ergänzen kann. Für die Schätzung der offenen Stellen werden zwei Ensemble-Modelle getestet und validiert. Im letzten Kapitel werden die gewonnenen Erkenntnisse diskutiert und darauf basierend ein Fazit formuliert sowie ein Ausblick gegeben.

4 CSIRO: Commonwealth Scientific and Industrial Research Organisation.

2. Daten und Datenqualität von Online-Stellenanzeigen

2. 1 Coverage-Probleme

Die Nutzung von OJA-Daten zur Schätzung offener Stellen bringt mehrere methodische Herausforderungen mit sich. In einem ILO[5] Working Paper beschreiben Fabo/Kureková (2022) im Rahmen einer ausführlichen internationalen Review von Studien, die OJA-Daten nutzen, die folgenden Verzerrungen: OJA-Daten unterrepräsentieren oft bestimmte Sektoren (beispielsweise Bau oder Landwirtschaft) und kleinere Unternehmen, die weniger häufig online inserieren. Gleichzeitig werden Jobs, die höhere Bildungsabschlüsse erfordern, häufiger online ausgeschrieben, was zu einer Überrepräsentation von Stellenangeboten für besser ausgebildete Arbeitskräfte führt.[6]

In einer einem Cedefop Working Paper zur Repräsentativität von OJA-Daten in Europa finden Napierala u. a. (2022) jedoch eine relativ gute Abdeckung über unterschiedliche Wirtschaftsbereiche (gemessen über Berufe der ISCO-Klassifikation).

Auf den deutschen Arbeitsmarkt bezogen stellen Rengers (2018b) und de Lazzer/Rengers (2021) zwei Coverage-Probleme dar. Ein erstes Coverage-Problem entsteht dabei durch die unvollständige Abdeckung der Gesamtheit der offenen Stellen durch Online-Stellenanzeigen. So hatten frühere Untersuchungen ergeben, dass zwischen 41 % und 70 % der Unternehmen ihre offenen Stellen auf Online-Jobportalen bekanntgeben, Online-Stellenanzeigen demnach nicht das gesamte Stellenangebot des Marktes repräsentieren.

Die Vielzahl von Jobportalen macht darüber hinaus eine Auswahl der OJA-Quellen erforderlich, was zu einem zweiten Coverage-Problem führen kann, wenn dabei die Repräsentativität des Online-Stellenmarktes verloren geht. Das Treffen einer Auswahl von Jobportalen ist auch aufgrund der in Deutschland zu beobachtenden Schnelllebigkeit in der Branche der Jobportalanbieter nicht zu vermeiden, sodass der Online-Stellenmarkt nicht vollständig abgebildet werden kann.[7]

2.2 Dubletten, Stock versus Flow und technische Aspekte

Bei der internetgestützten[8] Erfassung von Online-Stellenanzeigen *verschiedener Jobportale* kann es zur mehrfachen Erfassung identischer Stellenanzeigen und damit zu Dubletten kommen. Deren Bereinigung ist komplex und erfordert den Einsatz algorithmischer Verfahren oder Methoden des maschinellen Lernens. Dubletten entstehen allerdings nicht nur durch die parallele Erfassung von OJA-Daten verschiedener Portale, sondern auch *bei einem Jobportal zu unterschiedlichen Zeitpunkten*. Die wichtigsten Kriterien bei der Dublettenprüfung sind deshalb inhaltliche Nähe und zeitnahe Veröffentlichung von Online-Stellenanzeigen. Eine grundlegende Heuristik zur Reduktion zeitlich

5 Internationale Arbeitsorganisation (International Labour Organization – ILO).

6 Weitere Verzerrungen (nach Fabo/Kureková 2022) können auftreten durch informelle Beschäftigung, einen ungleichen Zugang zum Internet, einen hohen Anteil selbstständiger Arbeit und stark variierende Muster über Regionen. Diese Verzerrungen beziehen sich aber auf den weltweiten Vergleich und sind von geringerer Bedeutung für den deutschen Arbeitsmarkt.

7 „In Deutschland gibt es mehr als 1000 Jobportale. Die inhaltlich vollständige Erfassung der Online-Stellenanzeigen all dieser Portale wäre – trotz aller Automatisierungen – nicht nur mit einem hohen (Rechner-)Zeitaufwand verbunden, sondern würde auch immense Speicherkapazitäten (Daten-Storage-Problem) benötigen. Daher muss eine Auswahl getroffen werden.“ (de Lazzer/Rengers 2021).

8 Die Datenerfassung kann über API, Scraping oder Crawling durchgeführt werden. Bei der Datenerhebung mittels Application *Programming Interfaces (APIs)* werden vom Portalbetreiber bereitgestellte Schnittstellen zur strukturierten Abfrage von Inhalten verwendet. Beim *Scraping* erfolgt eine gezielte Extraktion eindeutig definierter Inhalte mithilfe speziell auf einzelne Webseiten zugeschnittener Software. *Crawling* hingegen ist die automatisierte und weitgehend unspezifische Erfassung von Webinhalten durch Programme, die selbstständig nach relevanten Informationen suchen.

bedingter Dubletten ist die Unterscheidung zwischen Bestands- und Stromgrößen. Wird ein zu einem Zeitpunkt dublettenbereinigter Bestand ausschließlich um eindeutige Stromgrößen ergänzt, lässt sich dieses Problem weitgehend vermeiden (siehe dazu auch Rengers 2018b, S. 13f.).

Theoretisch könnten OJA-Daten – wie bei einem Schnappschuss (Screenshot) – alle zu einem bestimmten Zeitpunkt t im Internet vorhandenen Stellenanzeigen als Bestandsgröße (Stock) abbilden. In der Realität sind damit allerdings mehrere Probleme verbunden. Zu nennen sind hier zunächst rein technische Aspekte. So benötigt der Datenerfassungsvorgang selbst eine gewisse Zeitspanne, innerhalb derer sich bereits Bestandsveränderungen ergeben (können).

Aktive Maßnahmen zur Unterbindung von Scraping, zeitweise Nichterreichbarkeit von Webseiten oder Änderungen im Quellcode können außerdem dazu führen, dass Daten teilweise gar nicht erhoben werden. Neben dem Wegfall einzelner Portale können auch strukturell unvollständige Stellenanzeigen zu Verzerrungen im resultierenden Datensatz beitragen und vor allen Dingen die erforderliche Dedublizierung *zwischen den verschiedenen zu einem Zeitpunkt* t *erfassten OJA-Quellen* erschweren (siehe auch Rengers/Hornik 2024, S. 56).

Zur Abbildung des Bestands innerhalb eines Zeitraumes wie beispielsweise eines Monats muss darüber hinaus eine Dedublizierung *im Zeitverlauf* durchgeführt werden. Das heißt, jede einzelne in $t + x$ erfasste Online-Stellenanzeige muss daraufhin geprüft werden, ob es sich dabei um eine neue Stellenanzeige handelt, die als Flussgröße (Flow) innerhalb des Zeitraumes interpretiert werden kann.

Wie das genannte Stock-Flow-Problem mithilfe von Annahmen zum Beispiel über die durchschnittliche Vakanzdauer einer offenen Stelle gelöst werden kann, wird in Kapitel 4.1 thematisiert.

2.3 Fake-Stellenanzeigen, Personalvermittlung und Zeitarbeit

Das Problem von gefälschten Online-Stellenanzeigen tauchte bereits im Jahr 2009 im Zusammenhang mit Vorwürfen auf, das Jobportal der Bundesagentur für Arbeit werde von Datenhändlern missbraucht und würde seine Datenschutz- und Datensicherheitsverpflichtungen nicht hinreichend erfüllen. Dies geht aus einer Antwort der Bundesregierung auf eine Anfrage zum „Sozialdatenschutz und zu möglichen fingierten Stellenanzeigen im Stellenportal ‚Jobbörse' der Bundesagentur für Arbeit" (Deutscher Bundestag 2018) hervor. Die Fragesteller hatten unter anderem mit Bezug auf einen Artikel der *Süddeutschen Zeitung*[9] aus dem Jahr 2011 darauf hingewiesen, dass gefälschte Online-Stellenanzeigen beim Jobportal der Bundesagentur für Arbeit (BA) bereits länger bekannt seien. Konkret hatten Datensammler gefälschte Online-Stellenanzeigen geschaltet, um an sensible Bewerberdaten von Jobsuchenden zu kommen. Die damalige Bundesregierung erläuterte in ihrer Antwort ausführlich die verschiedenen Prüfverfahren der Bundesagentur für Arbeit zur Identifizierung möglicher Fakes bei betreuten und unbetreuten Stellenanzeigen. Allerdings wurde auch darauf hingewiesen, dass die Identifikation von gefälschten oder fingierten Stellenanzeigen eine „permanente Herausforderung für die Betreiber privater und öffentlicher Stellenbörsen" sei.

Im Mai 2019 deckten SWR-Reporter erneut Tausende Fake-Stellenanzeigen beim Jobportal der Bundesagentur für Arbeit auf. Geschaltet wurden sie von Datensammlern, die sich als private Arbeitsvermittler ausgegeben hatten. Als Konsequenz änderte die BA-Jobbörse im August 2019 ihre Standardsucheinstellungen zulasten von Online-Stellenanzeigen von privaten Arbeitsvermittlern.[10] Aktuell werden beim BA-Jobportal (www.arbeitsagentur.de/jobsuche/) weiterhin standardmäßig keine Online-Stellenanzeigen der Personalvermittlung angezeigt, wohl aber Stellenanzeigen von Zeitarbeitsfirmen.

Personalvermittlung und Zeitarbeit stellen zwei voneinander abzugrenzende Formen von Personaldienstleistungen dar. Während bei der Personalvermittlung

9 Siehe https://www.sueddeutsche.de/karriere/datenmissbrauch-bei-der-jobsuche-bewerber-und-betrueger-1.126185 [Zugriff am 19.11.2025].

10 Siehe Meldung von O-Ton-Arbeitsmarkt.de vom 23.08.2019: https://www.o-ton-arbeitsmarkt.de/o-ton-news/jobboerse-der-arbeitsagentur-nach-datenklau-ueberarbeitet [Zugriff am 19.11.2025].

geeignete Bewerber durch einen Dienstleister direkt in ein unbefristetes oder befristetes Beschäftigungsverhältnis beim Kundenunternehmen integriert werden, beschreibt Zeitarbeit (Arbeitnehmerüberlassung) die zeitlich begrenzte Überlassung von Arbeitskräften an ein Einsatzunternehmen, wobei das Beschäftigungsverhältnis mit dem Zeitarbeitsunternehmen besteht.

Stellenausschreibungen von Personalvermittlern können verstärkt zu Dubletten führen, da Arbeitgeber häufig nicht nur ausschließlich eine Personalvermittlung beauftragen, sondern – parallel zur Stellenanzeige des Personalvermittlers – auch in Eigenregie mit einer eigenen Stellenanzeige suchen. Wie bereits oben erwähnt, ist dieses Phänomen nicht auf die Bundesagentur für Arbeit beschränkt, sondern gilt generell für alle Jobportale.

2.4 Benchmarks

Zur Beurteilung der Datenqualität von OJA-Daten und ihrer Geeignetheit zur Abbildung der kurzfristigen (konjunkturellen) Arbeitskräftenachfrage dienen Referenzgrößen, die als Benchmark herangezogen werden können. So gibt es bereits andere Indizes, die die Entwicklung von offenen Stellen in Deutschland beschreiben. Grundsätzlich lassen sich drei Arten von Erhebungsmethoden unterscheiden: 1) stichprobenbasierte Umfragen, 2) Erhebungen von offiziell gemeldeten Stellen, 3) computergestützte Verfahren, die auf Daten von Online-Plattformen aufbauen. Die Datenbasis und das Ziel der Indizes sind dabei jeweils unterschiedlich, was bei der Interpretation einer vergleichenden Analyse zu berücksichtigen ist.

Die im Folgenden beschriebenen Benchmarks dieser Studie sind insbesondere die beiden etablierten Arbeitsmarktstatistiken, die Stellenerhebung (Job Vacancy Survey – JVS) des Instituts für Arbeitsmarkt und Berufsforschung (IAB) sowie die Statistik über die Zahl der bei der Bundesagentur für Arbeit (BA) gemeldeten Stellen. Darüber hinaus kommen zwei weitere Datenquellen aus dem Bereich der Online-Stellenanzeigen zur Anwendung: zum einen der Indeed-Stellenindikator und zum anderen der index-Stellenmarktindikator der index-Gruppe, die beide als Konjunkturindikatoren geeignet scheinen.

Sowohl national als auch international von herausragender Bedeutung ist der so genannte *Job Vacancy Survey (JVS)*, der in großen Teilen der Welt schon seit Jahrzehnten Auskunft über die Zahl der offenen Stellen in der Gesamtwirtschaft gibt, jedoch häufig nur auf Quartalsebene zur Verfügung steht. Für Deutschland wird dafür die Stellenerhebung des Instituts für Arbeitsmarkt- und Berufsforschung genutzt. Die *IAB-Stellenerhebung* (eine ausführliche Methodenbeschreibung findet man bei Brenzel u. a. 2016) wird quartalsweise durchgeführt und erfragt auf freiwilliger Basis Informationen zu Anzahl und Struktur offener Stellen. In dieser repräsentativen Umfrage werden jedes Quartal ca. 10.000 bis 15.000 Betriebe und Verwaltungen befragt. Sie existiert seit 1989 in West- und seit 1992 in Ostdeutschland. Ursprünglich als jährliche schriftliche Erhebung im vierten Quartal konzipiert, wurde sie ab 2006 durch telefonische Kurzinterviews in den ersten drei Quartalen ergänzt. Das vierte Quartal weist dabei den größten Stichprobenumfang auf.[11]

Die Daten der IAB-Stellenerhebung zeigen eine sehr hohe Datenqualität und können auf unterschiedlichen Ebenen segmentiert werden (Bossler u. a. 2020a). Trotz der hohen Datenqualität gibt es auch einige Herausforderungen (Napierala/Kvetan/Branka 2022): Da nicht alle Unternehmen der Aufforderung zur vierteljährlichen Meldung offener Stellen nachkommen, müssen fehlende Daten teilweise durch Imputation korrigiert werden, um Non-Response-Bias zu vermeiden. Neu gegründete Unternehmen werden möglicherweise nicht erfasst, da das Unternehmensregister nicht regelmäßig aktualisiert wird, was zu Verzögerungen zwischen der Stichprobenauswahl und der Datenerhebung führt. Darüber hinaus können Meldefehler auftreten, da die Daten auf Selbstauskünften basieren.

Eine Vergleichbarkeit der Ergebnisse von OJA-Daten mit denen der IAB-Stellenerhebung ist nur eingeschränkt gegeben (siehe dazu ausführlich Rengers 2018a, 2018b, S. 14ff.). Ein Grund ist, dass auf der einen Seite der Online-Stellenmarkt und auf der anderen der Gesamtstellenmarkt betrachtet wird. Auch besteht das Problem, dass die OJA-Daten in der Regel keine Angaben zur Anzahl der Stellen je Stellenanzeige enthalten. Darüber hinaus können aus der IAB-Stellenerhebung nur Quartalsindizes erstellt werden, während dieser Beitrag die Entwicklung und Prüfung von monatlichen Indizes aus OJA-Daten zum Ziel hat.

11 Ergebnisse der IAB-Stellenerhebung können unter https://iab.de/das-iab/befragungen/iab-stellenerhebung/ abgerufen werden.

Die *Bundesagentur für Arbeit* gibt monatlich einen Index *BA-X* von gemeldeten Stellen heraus. Der BA-X wurde zum ersten Mal im Jahr 2007 veröffentlicht und in der Methodik zuletzt im Januar 2020 revidiert. Seitdem dient das Jahr 2015 als Referenzjahr mit einen Basiswert von 100. Für diese Studie wurde er zur Vergleichbarkeit umbasiert.

Mit dem BA-X werden alle Stellen erfasst, die offiziell gemeldet wurden (siehe Bundesagentur für Arbeit 2020, 2025). Die gemeldeten Stellen werden vor der Indexberechnung saison- und kalenderbereinigt. In Deutschland hat der BA-X unter den Stellenindizes die höchste Aktualität. In Bezug auf die Schätzung der offenen Stellen ergibt sich allerdings eine Verzerrung, da keine Meldepflicht für zu besetzende Stellen besteht. „Die Selektivität, mit der Arbeitgeber die BA bei der Stellenbesetzung einschalten oder nicht, bestimmt den stichprobenbedingten Fehler in Bezug auf die gesamtwirtschaftliche [!] Arbeitskräftenachfrage" (Bundesagentur für Arbeit 2018, S. 17).[12] Die Meldequote lag in den vergangenen zehn Jahren zwischen 38 % und 58 % (IAB-Stellenerhebung).

Das Online-Stellenportal *Indeed* hat im Zuge der Coronakrise Indikatoren zum Verlauf von Online-Stellenanzeigen entwickelt und veröffentlicht. Es gibt Indizes heraus, die sowohl die Anzahl der aktiven Stellenanzeigen auf Indeed.com als auch die täglich neu eingestellten Stellenanzeigen abbilden (Glynn 2022). Außerdem wird ein Index berechnet, der die prozentuale Veränderung der saisonbereinigten Stellenanzeigen seit dem 1. Februar 2020 unter Verwendung eines gleitenden Sieben-Tage-Durchschnitts aufzeigt (der 1. Februar 2020 dient als vorpandemische Basislinie). Weiterhin werden auch Daten nach Wirtschaftssektoren und Bundesländern bereitgestellt. Der Index basiert allerdings ausschließlich auf Stellen, die bei indeed.com eingetragen sind, wodurch sich ebenfalls Verzerrungen ergeben.[13]

Seit September 2023 stellt auch die index-Gruppe einen hochaktuellen *index-Stellenmarktindikator* der Öffentlichkeit zur Verfügung. Für den Basiswert 100 wird auch hier – wie beim BA-X – das Jahr 2015 als Referenzjahr verwendet, weswegen nachfolgend umbasierte Werte für Vergleichszwecke verwendet werden. Im Gegensatz zu Indeed verwendet die index-Gruppe nicht nur Stellenanzeigen von Jobportalen. Ausgewertet werden nach eigenen Angaben für Deutschland 137 Printmedien und 42 Online-Stellenbörsen, wobei Quellen mit ausschließlich bezahlten Stellenzeigen überrepräsentiert seien.[14]

12 Daten zum BA-X werden unter https://statistik.arbeitsagentur.de/DE/Navigation/Statistiken/Statistiken-aktuell/BAX/BAX-Nav.html veröffentlicht.

13 Der Indeed-Stellenindikator heißt „Indeed Job Postings Index". Daten können unter https://data.indeed.com/#/postings abgerufen werden. Es gibt kein eigenes Methodenpapier. Neben der o. g. Quelle sind einige Hinweise bei den FAQ (https://www.hiringlab.org/indeed-data-faq-2/) zusammengestellt. Weitere methodische Hinweise sind in einzelnen Pressemitteilungen integriert (siehe https://www.hiringlab.org/de/).

14 Weitere Infos zum index-Stellenmarktindikator in monatlichen Pressemitteilungen, beispielsweise https://www.index.de/pressemitteilungen/stellenangebot-blieb-im-februar-niedrig/ (darunter auch ein Methodenpapier und Daten-Downloadmöglichkeiten).

3. Die Textkernel-Daten

Die oben beschriebenen Datenqualitätsprobleme sind je nach Zusammenstellung verschiedener OJA-Quellen, eingesetztem Algorithmus zur Entdeckung von Dubletten sowie technischer Probleme (wie Abbruch des Scrapings) in Ausmaß und Umfang unterschiedlich stark ausgeprägt. Letztlich ist die genaue Art der Datenerfassung häufig (Betriebs-)Geheimnis verschiedener Anbieter von OJA-Daten. Die nachfolgenden Ausführungen zeigen, welche Qualitätsprobleme die OJA-Daten von Textkernel haben.

Grundlage für die Analysen ist die Jobmonitor-Datenbank der Bertelsmann Stiftung mit über 400 Millionen Online-Stellenanzeigen und regelmäßig über 50.000 Quellen. Die Daten stammen vom niederländischen Anbieter Textkernel, der seit dem Jahr 2014 Daten in Deutschland erhebt und aufbereitet. Diese werden dedupliziert und von Textkernel mit Metadaten angereichert (beispielsweise zum Ort der Tätigkeit oder zum ausschreibenden Unternehmen). Für die vorliegende Studie werden Daten von Januar 2019 bis einschließlich September 2025 genutzt.

Tabelle 1 stellt die unterschiedlichen Quellen dar. Die Gesamtzahl der von Textkernel verwendeten Quellen lag im Jahr 2019 bei über 64.100, stieg im Jahr 2022 bis auf gut 86.500 an und sank ab dem Jahr 2024 deutlich. In der Anzahl machten Unternehmenswebseiten mit regelmäßig über 90 % in allen Jahren den Großteil der Quellen aus, gefolgt von Zeitungen/Sonstige und Jobportalen. Beispielsweise setzen sich im Jahr 2019 die gut 64.100 Quellen aus knapp 61.600 Unternehmenswebseiten, gut 900 Jobportalen und rund 1.600 Zeitungen/Sonstige zusammen. Bezüglich der jeweils erfassten Stellenanzeigen sehen die Verhältnisse dagegen anders aus: Ein Großteil der erfassten Stellenanzeigen kommt von den Jobportalen, wo die Anteile im beobachteten Zeitraum zwischen 75,6 % (2020) und 86,6 % (Jan.–Sept. 2025) schwanken.

In der Tabelle 1 ebenfalls angegeben ist die Zahl von Vermittler-Sites, die unabhängig von der sonstigen Klassifizierung der Quellen von Textkernel als Information zur Verfügung gestellt wurde. Im Jahr 2019 waren gut 1.800 solcher Quellen enthalten, ab dem Jahr 2024 dagegen weniger als 1.000. Von diesen Vermittler-Sites

TABELLE 1 **Textkernel-Daten: Art der Quellen und Anteil der erfassten Stellenanzeigen (einschließlich Dubletten)**

Jahr	insgesamt	Unternehmens-webseiten	Jobportale	Zeitungen, Sonstige	darunter Vermittler-Sites
	Anzahl Quellen				
2019	64.138	61.594	933	1.611	1.862
2020	73.230	70.649	949	1.632	1.808
2021	66.591	64.064	893	1.634	1.648
2022	86.563	84.105	897	1.561	1.562
2023	78.017	75.917	798	1.302	1.288
2024	25.826	24.270	630	926	917
2025/01-09	19.227	17.911	529	787	796
	% der insgesamt erfassten Stellenanzeigen				
2019	100	7,6	79,1	13,3	10,6
2020	100	8,9	75,6	15,5	10,0
2021	100	8,6	77,6	13,8	9,3
2022	100	8,5	76,0	15,5	11,6
2023	100	8,5	79,0	12,5	10,8
2024	100	5,6	83,7	10,7	9,6
2025/01-09	100	5,4	86,6	8,0	7,5

Quelle: Eigene Berechnungen.

TABELLE 2 **Textkernel-Daten: Anzahl der Stellenanzeigen und Dubletten**

Jahr	Stellenanzeigen einschließlich Dubletten	Stellenanzeigen ohne Dubletten	Dubletten	
			pro erfasster Stellenanzeige	Anteil an allen Stellenanzeigen
	Anzahl			%
2019	51.506.462	11.967.401	4,3	76,8
2020	51.712.793	11.098.489	4,7	78,5
2021	61.991.767	13.314.849	4,7	78,5
2022	79.624.330	15.141.419	5,3	81,0
2023	73.512.917	15.193.480	4,8	79,3
2024	58.945.503	12.490.678	4,7	78,8
2025/01-09	37.967.228	8.983.230	4,2	76,3

Quelle: Eigene Berechnungen.

wurden zwischen 11,6 % (2022) und 7,5 % (Jan.–Sept. 2025) der insgesamt erfassten Stellenanzeigen gewonnen.

Tabelle 2 zeigt die Anzahl der erfassten Stellenanzeigen vor und nach Bereinigung von Dubletten. Zu erwähnen ist, dass hier und im Folgenden generell Stellenanzeigen von Personalvermittlern, die insbesondere aus Gründen des in Kapitel 2.3 beschriebenen Problems der verstärkten Dublettenerzeugung kritisch zu betrachten sind, im Datensatz enthalten bleiben. Dies zum einen, weil eine Abgrenzung von Personalvermittlung und Zeitarbeit häufig nicht trennscharf erfolgen kann; zum anderen, weil auch unklar ist, in welchem Ausmaß Personalvermittlungen die Identifizierung von Dubletten erschweren oder sogar verhindern. Beide Aspekte bedürften einer intensiven, mit einem großen Zeitaufwand verbundenen Analyse, sind aber für den Hauptzweck dieser Studie von untergeordneter Relevanz – auch deshalb, weil die bisher identifizierten Stellenanzeigen von Personalvermittlern von der Größenordnung her noch keinen Anlass zur Sorge gegeben haben.

Inklusive Dubletten enthält der Textkernel-Datensatz pro Jahr zwischen 79,6 Mio. (2022) und 51,5 Mio. (2019) beziehungsweise 37,9 Mio. (Jan.–Sept. 2025) Stellenanzeigen. Pro erfasster Stellenanzeige werden zwischen 5,3 (2022) und 4,3 (2019) beziehungsweise 4,2 Dubletten (Jan.–Sept. 2025) identifiziert. Somit haben Dubletten einen Anteil zwischen 76,3 % und 81 % an allen erfassten Stellenanzeigen.

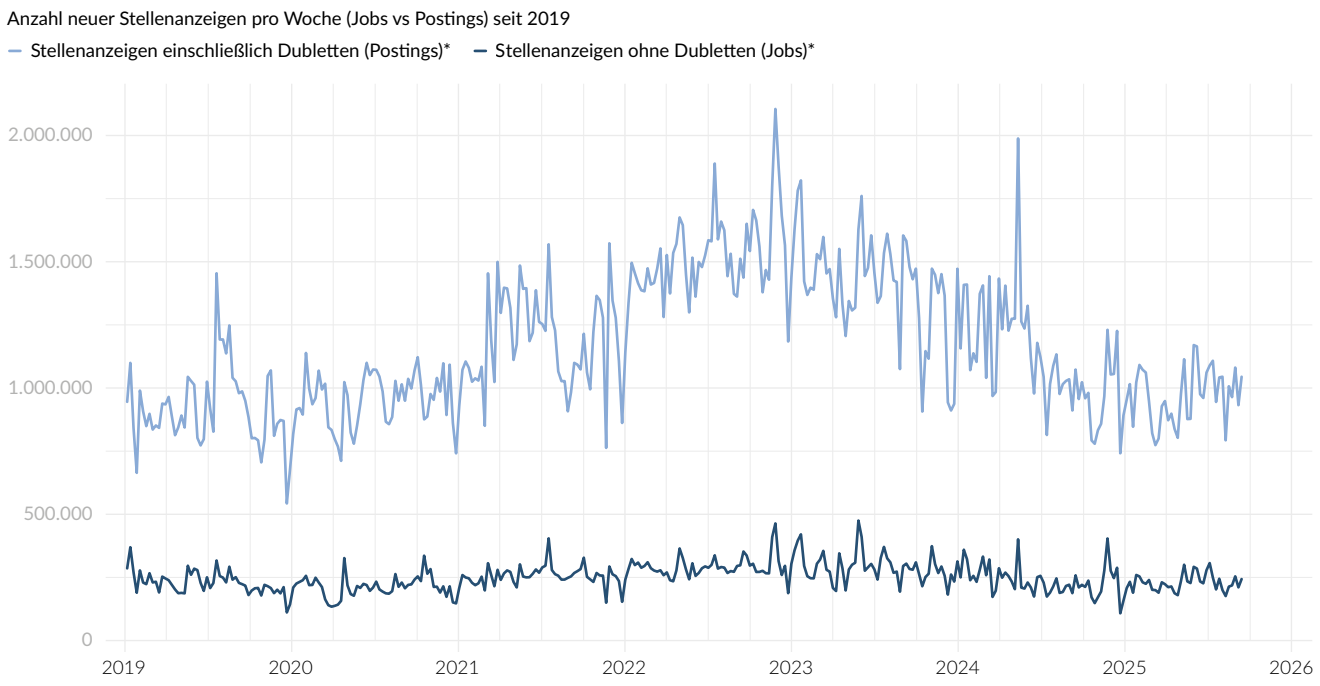
Tabelle 2 zeigt nicht, wie der unterjährige Verlauf der Anzahl deduplizierter Stellenanzeigen aussieht und ob diese gegebenenfalls bereits die Anforderungen an einen Konjunkturindikator erfüllen. Vorweg: Es ist nicht so, dass sich die pro Jahr erfassten und deduplizierten Stellenanzeigen gleichmäßig über die Monate verteilen. Wie die nachfolgenden Ausführungen zeigen, ist diese unterjährige Volatilität allerdings wohl eher ein überwiegend technisch bedingtes Datenerfassungsphänomen als ein saisonal bedingtes Arbeitskräftenachfragephänomen.

Wie bereits in Kapitel 2.2 beschrieben, wird die Datenqualität auch bei den Textkernel-Daten durch technische Aspekte beeinflusst. Abbildung 1 zeigt den Verlauf von neu eingestellten Online-Stellenanzeigen pro Woche (Flow) vor und nach Dublettenbereinigung. Auch wenn sich die sehr hohe Volatilität mit der Dublettenbereinigung verringert, lässt der genauere Blick in die Daten erkennen, dass technische Aspekte zu Artefakten führen.

Ein Problem dabei ist, dass nicht alle Portale und Webseiten vollständig und gleichmäßig gescraped werden können. Das führt zu mehreren Herausforderungen, die sich als die im Folgenden erläuterten Artefakte beschreiben lassen. Teilweise sind diese Artefakte bei der Betrachtung einzelner Zeitreihen zu sehen.

Drift: Die Anzahl der Portale nimmt kontinuierlich zu. Trotz Deduplizierung und einer gewissen Sättigung führt das insgesamt zu einem Anstieg der Postings, der nicht konjunkturell bedingt sein muss, sondern schlicht auf eine Zunahme an gesammelten Daten zurückzuführen ist. Der Effekt dieses Drifts lässt sich nicht zuverlässig bestimmen.

ABBILDUNG 1 **Textkernel-Daten: Neue Online-Stellenanzeigen pro Woche vor und nach Deduplizierung**



** Die Begriffe „postings" und „jobs" werden insbesondere in der englischsprachigen Literatur verwendet, sollen in dieser Studie aber möglichst vermieden werden. Der Grund hierfür ist, dass der Begriff „job" im Allgemeinverständnis mit **einer** Stelle gleichgesetzt wird. Hinter einer deduplizierten Stellenanzeige können aber auch **mehrere** Stellen stehen. Die Anzahl der Stellen pro Stellenanzeige ist bei OJA-Daten allerdings häufig unbekannt oder wird nicht erfasst.*

Quelle: Eigene Darstellung.

JOBMONITOR | BertelsmannStiftung

Step Changes: Für manche Portale bleibt zunächst die durchschnittliche Rate an gesammelten Postings über einen bestimmten Zeitraum hinweg relativ konstant – bis sich das Volumen von Monat zu Monat plötzlich ändert und schließlich auf einem anderen Niveau verharrt.

Implosions: Für einen gewissen Zeitraum werden regelmäßig Postings gesammelt – bis das Volumen einbricht und deutlich weniger bis gar keine Postings mehr gesammelt werden.

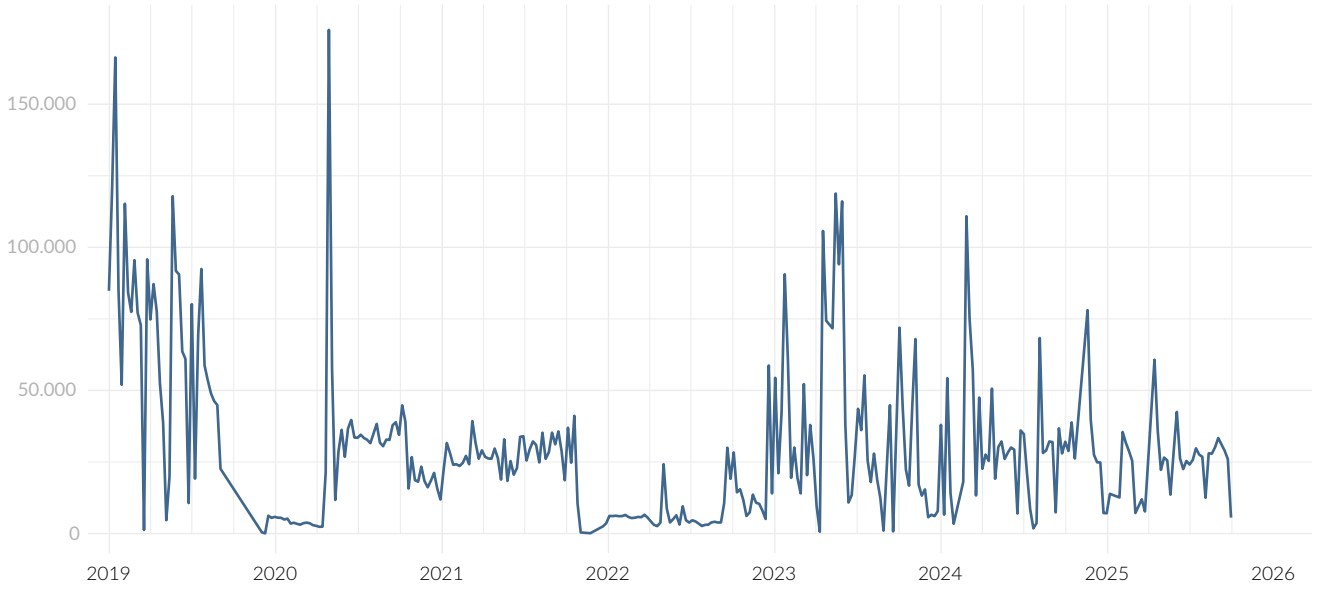
Explosions: Für einen kurzen Zeitraum werden plötzlich sehr viele Postings gesammelt. Dies hängt auch mit den Implosions zusammen. Wenn über einen bestimmten Zeitraum gar keine Postings gesammelt werden, kann der Scraper beim nächsten Besuch der Webseite ein deutlich höheres Volumen an neuen Postings scrapen.

Die beschriebenen Artefakte führen dazu, dass man die Zeitreihen für einzelne Portale als *intermittent-time-series mit step-changes* beschreiben kann (dies gilt teilweise auch für die aggregierte Zeitreihe). Das wiederum hat weitreichende Implikationen für die Modellierung mithilfe von statistischen Methoden.

Abbildung 2 zeigt diese technisch bedingten Herausforderungen noch einmal beispielhaft für ein großes deutsches Portal. Ab Mitte 2019 werden fast gar keine Postings mehr gesammelt (Implosion), bis Anfang 2020 innerhalb einer Woche über 175.000 Postings gesammelt werden (Explosion).

ABBILDUNG 2 **Textkernel-Daten: Pro Woche erfasste Online-Stellenanzeigen eines großen deutschen Portals**

Datenqualitätsprobleme am Beispiel eines großen Portals; Anzahl neuer Postings seit 2019



Quelle: Eigene Darstellung.

JOBMONITOR | BertelsmannStiftung

4. Ein mehrstufiges Verfahren zur Berechnung von OJA-Indizes

4.1 Stock

Das in Kapitel 2.2 beschriebene Stock-Flow-Problem kann mithilfe von Annahmen über die durchschnittliche Vakanzdauer einer offenen Stelle gelöst werden (siehe de Lazzer/Rengers 2021, Evans u. a. 2023). Auch wenn OJA-Daten generell nur eingeschränkt Auskunft darüber geben können, wie lange eine Stelle unbesetzt bleibt, wäre – rein gedanklich – auch eine indirekte Schätzung der Vakanzdauer aus den OJA-Daten selbst möglich. Dazu müsste allerdings jede gesammelte Stellenausschreibung täglich daraufhin überprüft werden, ob sie noch verfügbar ist, was einen großen technischen Aufwand bedeuten würde. Alternativ könnte die Differenz zwischen Erstveröffentlichungsdatum und Ablaufdatum einer Online-Stellenanzeige verwendet werden, die jedoch für einen Großteil der OJA-Daten nicht vorliegt.[15]

Umfragebasierte Werte geben die durchschnittliche Vakanzdauer bei sozialversicherungspflichtigen Neueinstellungen mit 93 Tagen an (Gürtzgen/Kubis/Popp 2023). Davon sind 62 Tage die durch Unternehmen geplante Vakanzdauer und 31 Tage ungeplant. Vakanzzeiten nehmen in Phasen des wirtschaftlichen Aufschwungs zu und in Rezessionsphasen ab (Bossler u. a. 2018).

Bei der Schätzung von Vakanzzeiten sollte allerdings ein wichtiger Unterschied erwähnt werden: Die IAB-Stellenerhebung fragt explizit nach offenen Stellen, während Online-Stellenausschreibungen bei längeren Vakanzzeiten auch erneut veröffentlicht werden können – abhängig vom Deduplizierungsverfahren können somit bei einem längeren Zeitraum zwischen zwei Stellenausschreibungen für dieselbe Position Duplikate nicht erkannt werden (siehe Evans u. a. 2023). Zhao/Chen/Mason 2021 diskutieren unterschiedliche Deduplizierungsansätze (unter anderem den Ansatz von Textkernel) und schlagen eine Dauer von 60 Tagen vor, bevor ein ähnliches Posting als neue Stelle behandelt wird. Das würde bedeuten, dass längere Vakanzzeiten auch zu einem höheren Aufkommen an neuen Postings führen würden.

Zur Bildung einer Stock-Variablen, die den durchschnittlichen Bestand an Online-Stellenanzeigen eines Monats abbildet, wurden vier, acht und zehn Wochen als Parameter für die Vakanzdauer geprüft. Die Wahl fiel schließlich auf acht Wochen (56 Tage), sodass die Flow-Variable „neue wöchentliche Stellen“ $\mathrm{p_t}$ durch nachfolgend gezeigte gleitende Summenbildung der neuen Postings über einen bestimmten Zeitraum $\mathrm{x_t}$ (Monat),

$$\mathrm{x_t} = \sum_{\mathrm{t}=1}^{8} \mathrm{p_t}$$

in eine Stock-Variable umgewandelt und durch anschließende Mittelwertbildung zu einer Schätzung des durchschnittlichen monatlichen Bestands an Online-Stellenanzeigen wird.

4.2 Ensemble-Ansätze und Ausreißerbereinigung

4.2.1 Die Idee von Ensemble-Ansätzen

Nachfolgend werden zwei unterschiedliche Ensemble-Ansätze getestet, um eine bereinigte Zeitreihe für die Entwicklung der offenen Stellen zu schätzen. Grundlage beider Verfahren ist die Studie von Evans u. a. (2023) der australischen staatlichen Forschungsorganisation CSIRO, in der ein Algorithmus vorgestellt wird, der die Nutzbarkeit von Online-Stellenanzeigen zur Abbildung zeitnaher, hochaktueller Entwicklungen des Stellenmarktes verbessert.

15 Eine Ausnahme sind hier OJA-Daten, die direkt von Stellenportalen (wie beispielsweise Indeed) angeboten werden, da diese einen direkten Zugriff auf den aktuellen Bestand ihrer angebotenen beziehungsweise aktiven Stellenanzeigen haben.

Die Studie von Evans u. a. (2023) basiert auf Daten, die von Adzuna Australia bereitgestellt wurden. Adzuna ist ein hybrides Jobportal, das OJA-Daten sowohl aus direkt von Arbeitgebern und Personalvermittlern veröffentlichten Online-Stellenanzeigen als auch aus einer Vielzahl externer Quellen erfasst, darunter auch Webseiten größerer Arbeitgeber. Obwohl OJA-Daten von Adzuna Australia bereits einem Dedublizierungsprozess unterzogen wurden, entfernten Evans und andere im Rahmen ihrer Analyse zusätzliche potenzielle Dubletten. Zur Beurteilung der Qualität ihrer Vorgehensweise zur Abbildung offener Stellen aus OJA-Daten verwendeten sie als Referenzgröße (Benchmark) die zeitliche Entwicklung der JVS-Daten, die vom Australian Bureau of Statistics veröffentlicht wird. Die OJA-Daten von Adzuna Australia sind nicht öffentlich zugänglich.[16]

Der hier angewandte Ensemble-Ansatz basiert darauf, Stellenanzeigen aus unterschiedlichen Quellen zunächst separat zu verarbeiten und erst im Anschluss zusammenzuführen. Dieses Vorgehen unterscheidet sich grundlegend von Methoden, bei denen die Daten zunächst quellenübergreifend aggregiert und dann gemeinsam verarbeitet werden, etwa durch eine einheitliche Glättung sämtlicher Stellenanzeigen über alle Quellen hinweg.

Es werden zwei Ansätze getestet. Im ersten Ansatz wird jede Zeitreihe der *absoluten Zahlen* der einzelnen OJA-Quellen um Ausreißer bereinigt, die Anzahl der Stellenanzeigen pro Monat dann aggregiert und schließlich indexiert. Im zweiten Ansatz, unmittelbar basierend auf Evans u. a. (2023), wird jede Zeitreihe der Änderungsraten der einzelnen OJA-Quellen um Ausreißer bereinigt und zu einem Index aggregiert.

4.2.2 Ausreißerbereinigung der absoluten Zahlen

Für die Identifizierung von Ausreißern in den Zeitreihen der absoluten Anzahl von Online-Stellenanzeigen jeder einzelnen OJA-Quelle wird zunächst ein Trend berechnet. Hierfür wird Friedman's Super Smoother (Luedicke 2015) verwendet, der für jede Woche ***t*** und jede Quelle ***j*** lokale Mittelwerte $\mathrm{x}^*_{\mathrm{jt}}$ bildet.

Für jede Woche wird nun die Distanz (Remainder) zwischen dem lokalen Mittelwert und dem tatsächlichen Wert berechnet: $\mathrm{d}_{\mathrm{jt}} = \mathrm{x}^*_{\mathrm{jt}} - \mathrm{x}_{\mathrm{jt}}$. Für die Verteilung der Remainder wird anschließend die Interquartilspanne (die Spanne zwischen dem 25%- und dem 75%-Quartil) berechnet. Beobachtungen werden als Ausreißer klassifiziert, wenn sie drei Interquartilspannen unter dem ersten Quartil oder über dem dritten Quartil liegen. Hyndman (2021) beschreibt, dass diese Definition von Tukey (1977, S. 44) in seinem ursprünglichen Boxplot-Vorschlag für „extrem abweichende" Werte verwendet wurde:

a. Wenn $\mathrm{x}_{\mathrm{jt}} < \mathrm{Q}_1 - 3 \cdot \mathrm{IQR}$, dann $\mathrm{x}_{\mathrm{jt}} = \mathrm{x}^*_{\mathrm{jt}}$

b. Wenn $\mathrm{x}_{\mathrm{jt}} > \mathrm{Q}_3 + 3 \cdot \mathrm{IQR}$, dann $\mathrm{x}_{\mathrm{jt}} = \mathrm{x}^*_{\mathrm{jt}}$

Abbildung 3 zeigt das Ergebnis der bereinigten neuen Postings am Beispiel der Online-Stellenanzeigen des ausgewählten großen Portals von Abbildung 2. Für eine OJA-Indexbildung müssen sodann die bereinigten neuen Postings aller einzelnen OJA-Quellen in die Posting-Stock-Variable (wie in Kapitel 4.1 beschrieben) umgewandelt werden. Der Index wird durch die Summe der Stellenanzeigen aller Portale gebildet und in Bezug zu einer Basisperiode gesetzt.

4.2.3 Winsorisierung der Änderungsrate

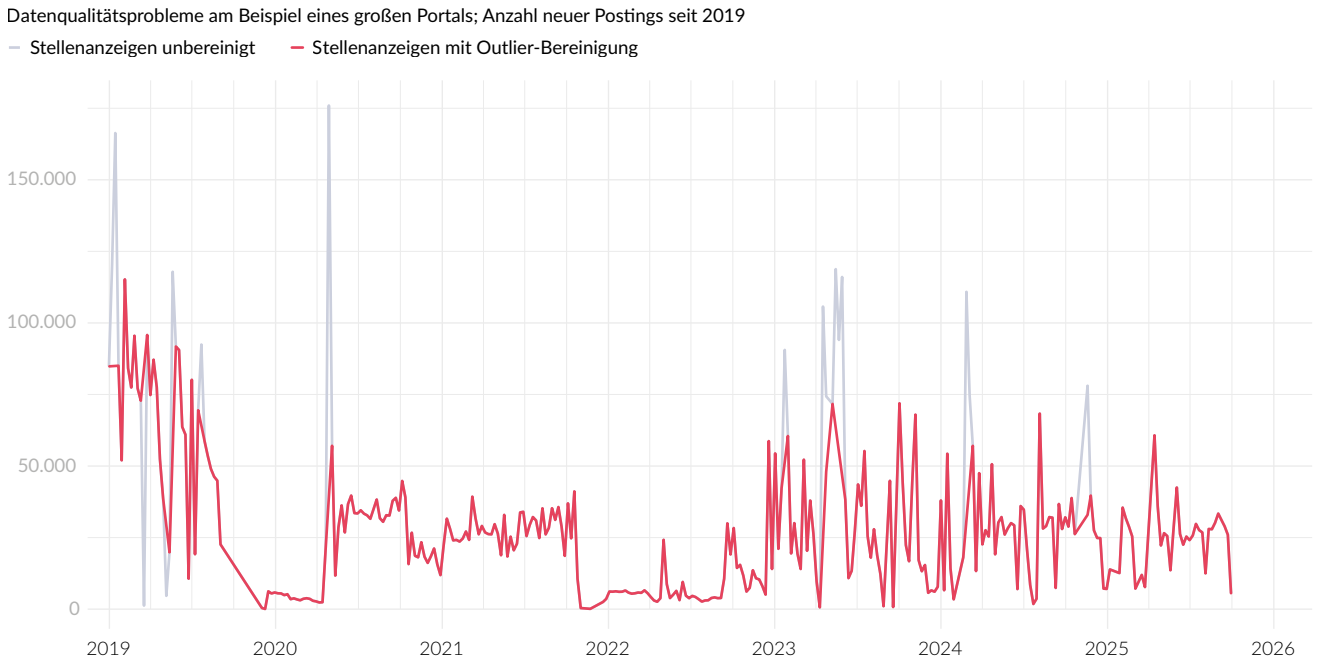
Bei der Winsorisierung – benannt nach dem Ingenieur und Biostatistiker Charles P. Winsor (1895–1951) – wird der Einfluss von Ausreißern auf einen Mittelwert begrenzt, indem Extremwerte auf bestimmte Perzentile begrenzt werden. Gleichzeitig bleibt aber die Gesamtstruktur der Daten erhalten. Evans u. a. (2023) beschreiben folgenden Winsor-basierten Algorithmus, um die Änderungsrate an offenen Stellen aus OJA-Daten zu schätzen:

1) Die neuen Postings werden in die Posting-Stock-Variable (wie in 4.1 beschrieben) umgewandelt.

2) Es werden alle OJA-Quellen ***j*** identifiziert, die sowohl in der Periode ***t*** als auch in der Periode ***t−1*** mehr als $\mathrm{x}_{\mathrm{min}}$ Postings aufweisen. Evans u. a. (2023) nutzen den Wert 100 als Parameter, den sie auf ihren Daten validieren konnten. Für die vorliegende Studie werden ebenfalls 100 Online-Stel-

16 Siehe Evans u. a. (2023) S. 3, 8.

ABBILDUNG 3 **Textkernel-Daten: Outlier-Bereinigung am Beispiel eines großen Portals**



Quelle: Eigene Darstellung.

JOBMONITOR | BertelsmannStiftung

lenanzeigen als Ausgangswert genutzt, aber auch andere Werte für x_{min} getestet.

3) Für jedes Portal wird die prozentuale Änderungsrate Δx_{jt} berechnet. Dies kann auf Monats- oder Quartals-Ebene stattfinden.

Änderungsrate: $\Delta x_{jt} = \frac{x_{jt} - x_{j,t-1}}{x_{j,t-1}}$

4) Die Änderungsrate wird für jede Periode um einen Winsorisierungsfaktor $k\%$ bereinigt. Dabei werden die Portale nach der Größe der Änderungsrate ihrer Stellenanzeigen geordnet. Anschließend werden die Quartilswerte von Δx_{jt}^{k} (als Schwelle für Ausreißer nach unten) und Δx_{jt}^{1-k} (als Schwelle für Ausreißer nach oben) berechnet. Alle Änderungsraten, die jeweils unter Δx_{jt}^{k} beziehungsweise über Δx_{jt}^{1-k} liegen, werden mit dem jeweiligen Quartilswert ersetzt.

a. Wenn $\Delta x_{jt} < \Delta x_{jt}^{k}$, dann $\Delta x_{jt}^{w} = \Delta x_{jt}^{k}$

b. Wenn $\Delta x_{jt} > \Delta x_{jt}^{1-k}$, dann $\Delta x_{jt}^{w} = \Delta x_{jt}^{1-k}$

5) Für jede Periode wird der Mittelwert der Änderungsraten gebildet.

Ein wesentliches Merkmal des Algorithmus ist, „that it treats each source's percentage change in job postings as an equally weighted signal of the change in job vacancies (subject to winsorization)." (Evans u. a. 2023, S. 4).

Die Berechnung von OJA-Indizes nach der Methode von Evans u. a. (2023) beruht auf (monatlichen) Änderungsraten beziehungsweise Wachstumsraten der OJA-Absolutwerte *jeder* einzelnen OJA-Quelle, nicht aber auf den kumulierten Absolutwerten aller OJA-Quellen. Es erfolgt damit sozusagen eine „Egalisierung" der verschiedenen OJA-Quellen, das heißt, die Entwicklung der Anzahl der Online-Stellenanzeigen auf kleineren Jobportalen erhält dasselbe Gewicht wie diejenige größerer Jobportale. Die Gewichtung von Portalen aufgrund ihrer Größe wird mit dieser Methode verhindert.

Bevor OJA-Indizes auf Basis dieser Änderungsraten berechnet werden, werden diese über die Winsorisierung von Ausreißern bereinigt. Die Indexberechnung erfolgt dann auf Basis der winsorisierten Wachstumsraten. Für die Entwicklung des Modells werden sowohl die Mindestanzahl an Postings x_{min} als auch die Winsorisierungsrate $k\%$ als Parameter mit unterschiedlichen Werten getestet.

5. Bewertung der Ergebnisse

5.1 Vorgehensweise

Im Folgenden werden die Ergebnisse für den Winsor-Ensemble-Index und den Outlier-Ensemble-Index dargestellt (siehe dazu die entsprechende Beschreibung in den Kapiteln 4.2.2 und 4.2.3). Die Parameter wurden initial gesetzt, mit einem Rolling Window von acht Wochen für die Schätzung der offenen Stellen. Beim Ensemble-Ansatz der Winsorisierung der Änderungsraten wurden außerdem Portale mit weniger als durchschnittlich 100 Stellenanzeigen in einem Monat oder dem Vormonat für die Berechnung im jeweiligen Monat ausgeschlossen. Die Winsorierungsrate wurde auf unterschiedlichen Levels getestet.

Für die Parameterbestimmung der Winsorisierung sowie für die generelle Beurteilung der mit den oben beschriebenen Ensemble-Ansätzen gebildeten OJA-Indizes lassen sich die gleichen Kriterien heranziehen. Es werden dabei zum einen Kennzahlen zur Korrelation der OJA-Indexreihen mit verschiedenen Benchmark-Zeitreihen und die Volatilitätsmaße verschiedener Zeitreihen verglichen, zum anderen erfolgen eine Wendepunktanalyse und eine grafische Vergleichsanalyse.

Primär soll die Anzahl der offenen Stellen in Deutschland abgebildet werden. Als Benchmark werden dabei vorrangig die IAB-Stellenerhebung und die BA-Statistik für den sekundären Vergleich verwendet. Darüber hinaus werden mit dem Indeed-Index und dem index-Stellenmarktindikator zwei weitere Indizes aus dem Bereich der Online-Stellenanzeigen in die Betrachtung miteinbezogen, die ebenfalls als potenziell geeignete Konjunkturindikatoren angesehen werden können und der Öffentlichkeit frei zur Verfügung stehen.

5.2 Korrelationskoeffizienten und Wendepunktanalyse

Abbildung 4 zeigt zunächst, dass aus Textkernel-Daten berechnete OJA-Indizes je nach Berechnungsverfahren sehr unterschiedlich ausfallen. Die Verläufe der einzelnen Zeitreihen weichen teilweise deutlich voneinander ab.

Für die Evaluation werden mehrere *Korrelationskoeffizienten*[17] ermittelt, die jeweils unterschiedliche Ebenen der Übereinstimmung zwischen den Indizes verschiedener Datenquellen bewerten. Tabelle 3 zeigt die Ergebnisse für den Outlier-Ensemble-Index und mehrere Winsor-Ensemble-Indizes mit verschiedenen Parametern *k* in Korrelation zu IAB- und BA-Daten. Diese Korrelationen werden nicht nur auf Basis der Textkernel-OJA-Daten, sondern auch für den nicht saisonbereinigten Indeed-OJA-Index und den nicht saisonbereinigten index-Stellenmarktindikator ermittelt.

Da die IAB-Stellenerhebung eine besondere Rolle als Benchmark spielt, aber nur quartalsweise durchgeführt wird, werden hier Korrelationen aus Quartalsveränderungen betrachtet. Die Korrelation der Veränderung zwischen Quartalen gibt an, wie ähnlich kurzfristige

17 Der Produkt-Moment-Korrelationskoeffizient (Bravais-Pearson-Korrelationskoeffizient) berechnet sich über

$$r = \frac{\mathrm{cov}(x, y)}{s_x \cdot s_y}$$

.

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

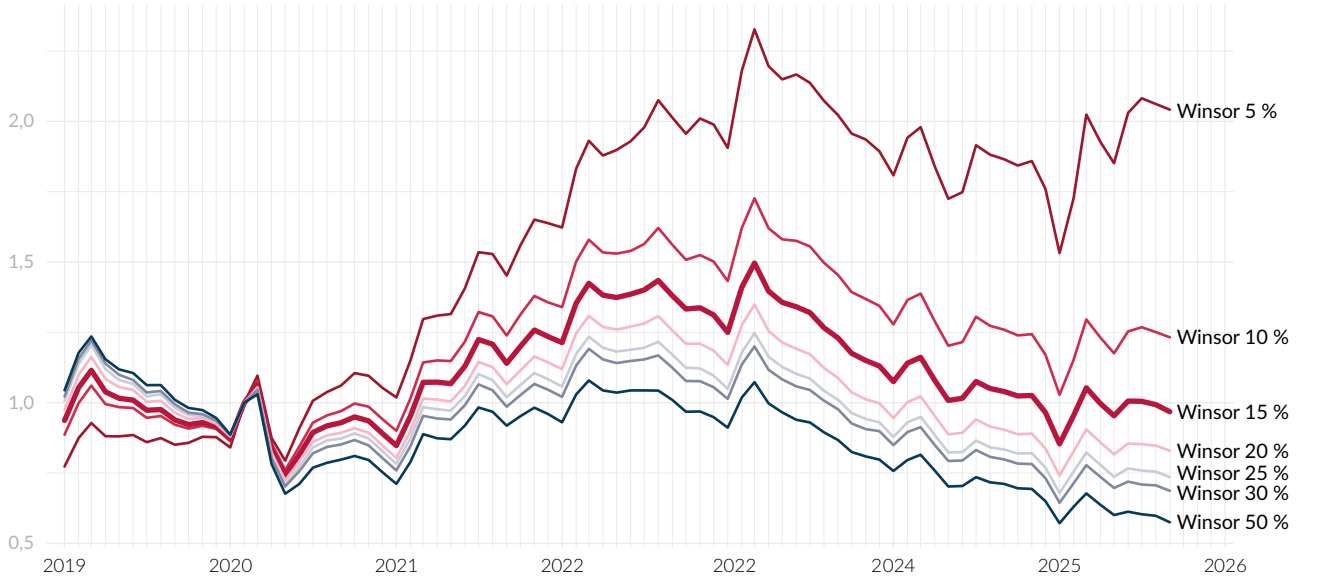
bzw.

$$r = \frac{n \cdot \sum_{i=1}^{n} x_i \cdot y_i - \left(\sum_{i=1}^{n} x_i\right) \cdot \left(\sum_{i=1}^{n} y_i\right)}{\sqrt{\left[n \cdot \sum_{i=1}^{n} {x_i}^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right] \cdot \left[n \cdot \sum_{i=1}^{n} {y_i}^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right]}}$$

Wertebereich $-1 \leq r \leq 1$
$r = 0$ keine Korrelation, unkorreliert, kein linearer Zusammenhang

ABBILDUNG 4 **Textkernel-Daten: OJA-Indizes unter verschiedenen Winsor-Raten**

Stock-Index unter verschiedenen Winsor-Raten



Quelle: Eigene Darstellung.

JOBMONITOR | BertelsmannStiftung

Schwankungen verschiedener Indizes ausfallen. Die Korrelation zwischen IAB-Daten und BA-Daten beträgt dabei nur 0,52 (siehe Tabelle 3). Die Korrelation zwischen dem Winsor-Index und dem IAB-Index bewegt sich ab einer Winsorisierungsrate von 5 % zwischen 0,38 und 0,51, während der Outlier-Index mit einem Wert von 0,22 deutlich darunter liegt. Bei Indeed und der index-Gruppe liegen die Korrelationen bei 0,57 beziehungsweise 0,44.

Die Korrelation mit dem BA-Index ist insgesamt höher; beispielweise liegen die Werte bei den Winsor-Indizes zwischen 0,52 und 0,68. Änderungen zwischen Quartalen im Outlier-Index sind mit 0,32 kaum mit dem BA-Index korreliert.

Quartalsänderungsraten zum vorletzten Quartal sind stärker korreliert: Die Winsor-Indizes korrelieren mit dem IAB-Index mit 0,62 bis 0,71 und mit dem BA-Index mit 0,62 bis 0,82. Die Änderungsraten zum vorletzten Quartal sind damit bei den Winsor-Indizes

TABELLE 3 **Korrelationskoeffizienten der Veränderungsraten der Indizes im Vergleich zu BA- und IAB-Daten**

Datenquelle	Veränderungsraten im Vergleich zum ...						Index	
	Vorquartal		vorletzten Quartal		Vorjahres-Quartal			
	IAB	BA	IAB	BA	IAB	BA	IAB	BA
IAB	1	0,52	1	0,64	1	0,92	1	0,84
BA	0,52	1	0,64	1	0,92	1	0,84	1
Outlier-Index	0,22	0,32	0,41	0,35	0,56	0,42	0,82	0,63
Winsor 5 %	0,38	0,52	0,62	0,62	0,85	0,75	0,64	0,29
Winsor 10 %	0,41	0,59	0,65	0,72	0,91	0,84	0,83	0,60
Winsor 15 %	0,43	0,61	0,67	0,75	0,92	0,87	0,90	0,79
Winsor 20 %	0,44	0,63	0,67	0,77	0,93	0,89	0,87	0,88
Winsor 25 %	0,46	0,63	0,69	0,78	0,93	0,90	0,77	0,88
Winsor 30 %	0,46	0,65	0,69	0,80	0,94	0,91	0,71	0,86
Winsor 50 %	0,51	0,68	0,71	0,82	0,94	0,93	0,52	0,77
Indeed	0,57	0,66	0,80	0,83	0,94	0,92	0,94	0,90
Index-Gruppe	0,44	0,61	0,71	0,82	0,90	0,94	0,90	0,84

Quelle: Eigene Berechnungen.

überwiegend stärker mit denjenigen des IAB-Index korreliert, als dies beim BA-Index der Fall ist (0,64).

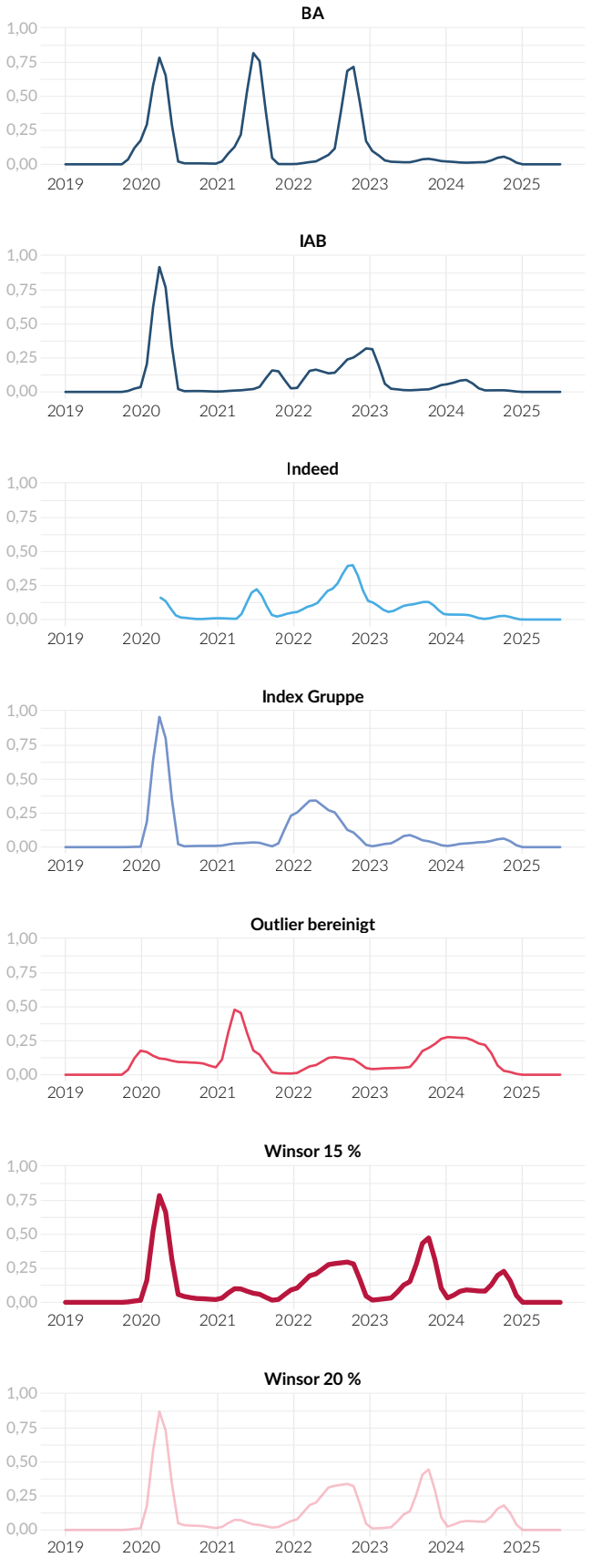
Die Korrelation der Änderungsrate steigt weiter mit einem Vergleich zum Vorjahres-Quartal. Der Winsor-Index (ab *k* = 5 %) korreliert mit dem IAB- beziehungsweise BA-Index mit 0,85 und 0,94 beziehungsweise mit 0,75 und 0,93. Der Outlier-Index weist eine deutlich schwächere Korrelation auf (IAB: 0,56; BA: 0,42).

Während bei den Änderungsraten eine höhere Winsorisierungsrate zu einer höheren Übereinstimmung mit den Vergleichsindizes führt, zeigt sich beim direkten Vergleich der Indizes, dass eine Winsorisierungsrate von 10 %, 15 % oder 20 % mit Werten von über 0,8 die größte Korrelation aufweist.

Für die Bestimmung der *Wendepunkte* in den Index-Zeitreihen wird ein Bayesian Change-Point Detection Algorithmus genutzt (BEAST). Dabei werden nicht nur abrupte Änderungen im Trend der Daten festgestellt, sondern auch berechnet, wie hoch die Wahrscheinlichkeit ist, dass es sich um einen wahren Wendepunkt handelt (in Vergleich zu einem zufälligen Artefakt der Daten). Wie in Abbildung 5 zu erkennen, zeigen die *winsorisierten Textkernel-Indizes* (Winsor15 und Winsor20) mehrere Wendepunkte: Der ausgeprägteste Wendepunkt ist der Beginn der Covid-19-Pandemie Anfang 2020 mit einer Wahrscheinlichkeit von 76 % beziehungsweise 88 %. Zudem gibt es Anhaltspunkte für weitere Wendepunkte. Anfang 2022 liegt die Wahrscheinlichkeit für einen Wendepunkt bei etwa 29 % beziehungsweise 31 % und Ende 2023 bei fast 50 % beziehungsweise 46 %. Der *IAB-Index* zeigt einen ausgeprägten Wendepunkt im Jahr 2020 (92 %). Anfang 2023 tritt ein weiterer Peak auf; die Wahrscheinlichkeit für einen Wendepunkt liegt dort bei etwa 30 %. Die Wendepunkte treten damit zu ähnlichen Zeitpunkten ein wie bei dem Stock-Index. Der *BA-Index* zeigt drei sehr klare Wendepunkte: Anfang 2020 (Covid-19-Pandemie, 77 %), Mitte 2021 (84 %) und Ende 2022 (74 %). Der *Oulier-Ensemble-Index* zeigt nur einen signifikanten Wendepunkt Anfang 2021 (51 %). Die Wendepunkte der anderen Indizes sind sehr schwach bis gar nicht ausgeprägt.

ABBILDUNG 5 **Wendepunktanalyse bei winsorisierten Textkernel-Daten im Vergleich zu BA- und IAB-Daten**

Wahrscheinlichkeit eines wahren Wendepunkts (0–1)



Quelle: Eigene Darstellung.

5.3 Volatilitätsmaße und grafische Vergleichsanalyse

Volatilitätsmaße bieten die Möglichkeit, unterschiedliche Ausmaße von Volatilität einer Zeitreihe in einer Kennzahl auszudrücken.[18] Tabelle 4 und Tabelle 5 zeigen die Ergebnisse für verschiedene Zeitreihen aus OJA-Daten von Textkernel sowie im Vergleich zu den Zeitreihen von Index.de, Indeed und BA-X. Alle Zeitreihen sind monatliche Indexreihen, die im Februar 2020 den Basiswert 100 haben. Dabei werden Minimum und Maximum der Vormonatsveränderungen in Prozentpunkten angegeben und Kennzahlen zu so genannten „Large M-M-changes", „Large Inversions" und „Double Large Inversions" berechnet. *Large M-M-changes* zeigen, wie häufig Vormonatsveränderungen in der absoluten Größenordnung von zehn beziehungsweise fünf Prozentpunkten innerhalb der letzten 36 Monate und in den einzelnen Berichtsjahren auftraten. Die Kennzahlen zu *Large Inversions* geben an, wie häufig der Index im Vergleich zum Vormonat zunächst um mindestens zehn beziehungsweise fünf Prozentpunkte fällt (steigt), aber im darauffolgenden Monat in genau dieser Größenordnung wieder steigt (fällt). Bei *Double Large Inversions* fällt (steigt) der Index im Vormonatsvergleich um mindestens zehn beziehungsweise fünf Prozentpunkte, steigt (fällt) im darauffolgenden Monat und fällt (steigt) im übernächsten Monat wieder jeweils in dieser Größenordnung.

Diese Kennzahlen sollen Hinweise darauf geben, in welchem Ausmaß die Zeitreihen Volatilitäten aufweisen, die methodisch-technisch bedingt sind und nicht die reale Entwicklung am Arbeitsmarkt widerspiegeln. Die zugrundeliegende Idee ist, dass (konjunkturelle) Wendepunkte in der Arbeitskräftenachfrageentwicklung eher einmalige Ereignisse sind, die in einer solchen Größenordnung nicht monatlich auftreten. Da die Corona-Krise genau in den hier betrachteten Zeitraum fällt (Beginn 2020), sollte sich dies auch in den Volatilitätskennzahlen widerspiegeln.

Tabelle 4 zeigt zunächst die verschiedenen OJA-Indizes aus Textkernel-Daten. Der mit „Textkernel" bezeichnete Index basiert auf dem Verfahren zur Behandlung des Stock-Flow-Problems (Kapitel 4.1) und enthält ansonsten keine weiteren Bereinigungen. Die Indizes „Textkernel Outlier" und „Textkernel Winsor-k%" wurden dagegen den in Kapitel 4.2.2 (Ausreißerbereinigung der absoluten Zahlen) beziehungsweise Kapitel 4.2.3 (Winsorisierung der Änderungsraten) beschriebenen Verfahren unterzogen.

Wie erwartet, zeigt der Textkernel-Index im Corona-Jahr 2020 mit -20,4 Prozentpunkten den größten Rückgang bei den Vormonatsveränderungen auf. Auch weisen die Large-|10|-M-M-changes dort mit 33,3 % ein Maximum auf, während alle Large Inversions und Double Large Inversions bei null liegen. Der Textkernel-Outlier-Index verändert die Eigenschaften und schwächt beispielsweise den coronabedingten Einbruch bei den Vormonatsveränderungen im Jahr 2020 auf ein Minimum von -19,6 Prozentpunkten ab; er zeigt gleichzeitig für das Jahr 2019 einen noch stärkeren Einbruch mit -21,8 Prozentpunkten. Darüber hinaus verschlechtern sich auch die Werte bei den Large-|10|-Inversions.
Bei allen Textkernel-Winsor-k%-Indizes wird dagegen der Einbruch im Jahr 2020 gegenüber dem einfachen Textkernel-Index stärker herausgehoben (Winsor10 mit -22,2 Prozentpunkten; Winsor15 mit -22,8 Prozentpunkten; Winsor20 mit -23,2 Prozentpunkten) und bleibt auch der stärkste Einbruch im Beobachtungszeitraum. Sowohl Textkernel-Winsor10 als auch Textkernel-Winsor15 verschlechtern sich hinsichtlich der Performance bei den Large-|10|-Inversions, wobei dies bei Winsor15 weniger deutlich ist. Alle winsorisierten Textkernel-Indizes reduzieren in den einzelnen Jahren die Large-|10|-M-M-changes gegenüber dem einfachen Textkernel-Index und stellen auch hier die Coronakrise klarer heraus. Am deutlichsten erfolgt dies bei Textkernel-Winsor20, gefolgt von Textkernel-Winsor15.

Die Volatilitätsmaße der Tabelle 4 zeigen insgesamt, dass Bereinigungsprozesse nicht unbedingt das liefern, was im Vorfeld von ihnen (theoretisch) erwartet wird. Dies liegt unter anderem an der außerordentlichen Komplexität des gesamten Erfassungs- und Aufbereitungsprozesses von OJA-Daten. Es stellt sich dabei jedoch die Frage, ob andere Indizes gegebenenfalls eine bessere Performance zeigen. In Tabelle 5 wird deshalb beispielhaft der Textkernel-Winsor15 ausgewählt und vergleichbaren monatlichen Indizes von BA-X, Index.de und Indeed gegenübergestellt (siehe dazu auch Kapitel 2.4). Alle Indizes werden in ihrer nicht saisonbereinigten Form verwendet, mit Ausnahme des BA-X-Index, der ausschließlich saisonbereinigt vorliegt.

18 Die Grundidee der nachfolgend dargestellten Volatilitätsmaße geht zurück auf eine EU-Task-Force zu Ergebnissen der monatlichen Erwerbslosenquoten, zu der die Autorin gehörte.

TABELLE 4 **Volatilitätsmaße für verschiedene OJA-Indizes aus Textkernel-Daten**

Scoreboard: Volatilität
Monatlicher OJA-Index mit Basis Feb 2020, NICHT saisonbereinigt

Index		Vormonats-veränderungen in %-Punkten		%-Anteil Vormonatsveränderungen in %-Punkten der Größenordnung L					
				Large: >= |10|			Large: >= |5|		
Daten	Zeitraum	MIN	MAX	Large M-M-changes	Large Inversions	Double Large Invers.	Large M-M-changes	Large Inversions	Double Large Inversions
Textkernel	letzte 36 Mon.	-12,2	22,2	11,1	0,0	0,0	52,8	8,3	0,0
	2019	-13,4	7,0	9,1	0,0	0,0	54,5	20,0	22,2
	2020	-20,4	11,6	33,3	0,0	0,0	50,0	8,3	0,0
	2021	-12,2	10,6	25,0	0,0	0,0	58,3	8,3	0,0
	2022	-10,1	15,5	16,7	0,0	0,0	33,3	8,3	0,0
	2023	-9,3	22,2	8,3	0,0	0,0	50,0	0,0	0,0
	2024	-9,2	14,0	8,3	0,0	0,0	41,7	0,0	0,0
	2025/01-09	-11,3	12,4	22,2	0,0	0,0	77,8	33,3	0,0
Textkernel	letzte 36 Mon.	-14,0	37,3	33,3	2,8	0,0	72,2	22,2	2,8
Outlier	2019	-21,8	18,7	18,2	0,0	0,0	54,5	40,0	44,4
	2020	-19,6	14,8	50,0	16,7	0,0	50,0	25,0	8,3
	2021	-9,9	13,3	16,7	0,0	0,0	50,0	8,3	0,0
	2022	-9,7	12,2	16,7	0,0	0,0	66,7	33,3	8,3
	2023	-14,0	37,3	41,7	8,3	0,0	83,3	33,3	8,3
	2024	-11,7	11,3	33,3	0,0	0,0	66,7	8,3	0,0
	2025/01-09	-13,4	16,0	22,2	0,0	0,0	66,7	22,2	0,0
Textkernel	letzte 36 Mon.	-14,2	18,8	16,7	5,6	0,0	52,8	16,7	0,0
Winsor10	2019	-6,5	11,2	9,1	0,0	0,0	27,3	10,0	0,0
	2020	-22,2	13,7	16,7	0,0	0,0	50,0	16,7	0,0
	2021	-6,8	12,8	25,0	0,0	0,0	58,3	8,3	0,0
	2022	-6,0	16,2	8,3	0,0	0,0	41,7	8,3	0,0
	2023	-10,6	18,8	25,0	8,3	0,0	50,0	16,7	0,0
	2024	-9,8	9,0	0,0	0,0	0,0	50,0	8,3	0,0
	2025/01-09	-14,2	14,3	33,3	11,1	0,0	66,7	33,3	0,0
Textkernel	letzte 36 Mon.	-10,0	16,0	11,1	2,8	0,0	47,2	13,9	0,0
Winsor15	2019	-7,6	11,7	9,1	0,0	0,0	27,3	20,0	22,2
	2020	-22,8	12,6	16,7	0,0	0,0	50,0	16,7	0,0
	2021	-6,8	12,0	16,7	0,0	0,0	58,3	8,3	0,0
	2022	-5,4	14,0	8,3	0,0	0,0	25,0	0,0	0,0
	2023	-10,0	16,0	16,7	0,0	0,0	50,0	16,7	0,0
	2024	-8,2	6,6	0,0	0,0	0,0	50,0	8,3	0,0
	2025/01-09	-11,1	10,1	22,2	11,1	0,0	55,6	22,2	0,0
Textkernel	letzte 36 Mon.	-9,6	14,1	2,8	0,0	0,0	36,1	11,1	0,0
Winsor20	2019	-7,9	12,4	9,1	0,0	0,0	27,3	20,0	22,2
	2020	-23,2	12,0	16,7	0,0	0,0	50,0	16,7	0,0
	2021	-6,1	11,5	8,3	0,0	0,0	50,0	0,0	0,0
	2022	-5,1	12,8	8,3	0,0	0,0	25,0	0,0	0,0
	2023	-9,4	14,1	8,3	0,0	0,0	41,7	16,7	0,0
	2024	-6,9	5,7	0,0	0,0	0,0	41,7	8,3	0,0
	2025/01-09	-9,6	8,6	0,0	0,0	0,0	33,3	11,1	0,0

Quelle: Eigene Darstellung.

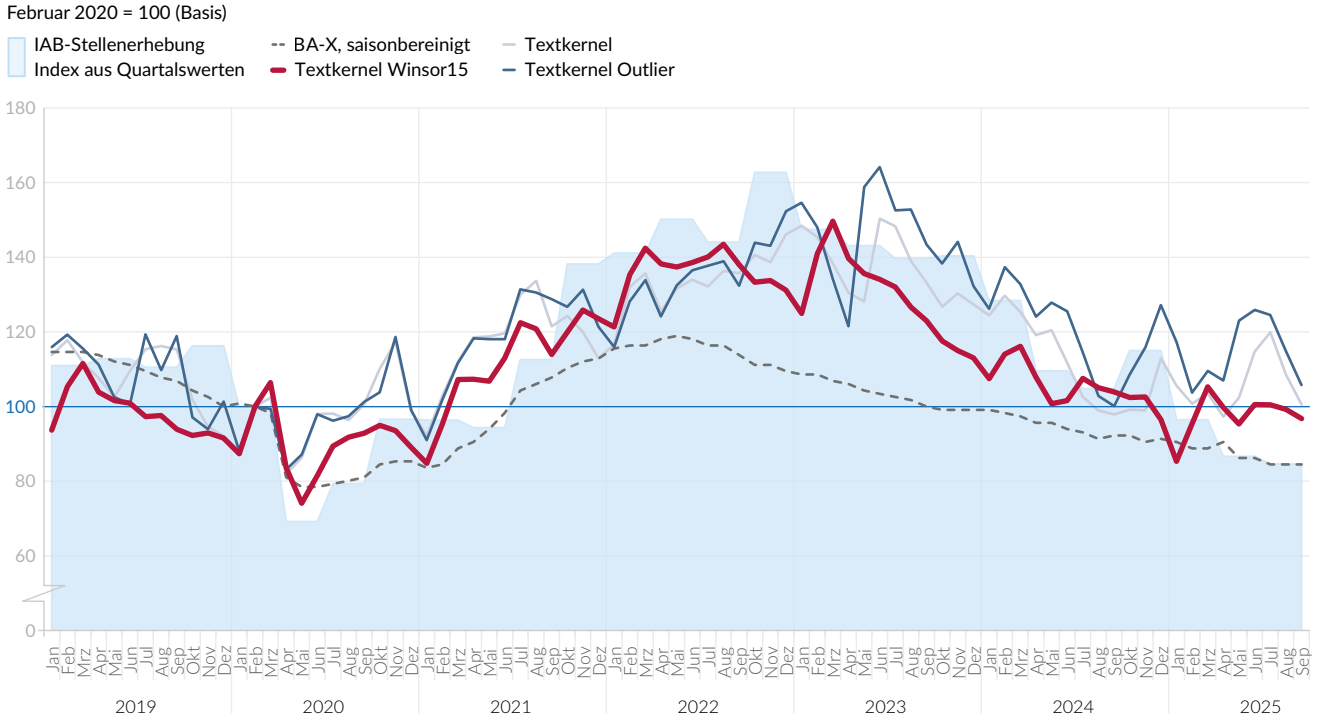
TABELLE 5 **Volatilitätsmaße für OJA-Indizes aus Textkernel-Daten im Vergleich zu BA-X, Index.de und Indeed**

Scoreboard: Volatilität
Monatlicher OJA-Index mit Basis Feb 2020, NICHT saisonbereinigt

Index		Vormonats-veränderungen in %-Punkten		%-Anteil Vormonatsveränderungen in %-Punkten der Größenordnung L					
				Large: >= |10|			Large: >= |5|		
Daten	Zeitraum	MIN	MAX	Large M-M-changes	Large Inversions	Double Large Invers.	Large M-M-changes	Large Inversions	Double Large Inversions
Textkernel	letzte 36 Mon.	-10,0	16,0	11,1	2,8	0,0	47,2	13,9	0,0
Winsor15	2019	-7,6	11,7	9,1	0,0	0,0	27,3	20,0	22,2
	2020	-22,8	12,6	16,7	0,0	0,0	50,0	16,7	0,0
	2021	-6,8	12,0	16,7	0,0	0,0	58,3	8,3	0,0
	2022	-5,4	14,0	8,3	0,0	0,0	25,0	0,0	0,0
	2023	-10,0	16,0	16,7	0,0	0,0	50,0	16,7	0,0
	2024	-8,2	6,6	0,0	0,0	0,0	50,0	8,3	0,0
	2025/01-09	-11,1	10,1	22,2	11,1	0,0	55,6	22,2	0,0
BA-X	letzte 36 Mon.	-4,3	1,7	0,0	0,0	0,0	2,8	0,0	0,0
saison-	2019	-2,6	0,0	0,0	0,0	0,0	0,0	0,0	0,0
bereinigt	2020	-17,2	3,4	8,3	0,0	0,0	8,3	0,0	0,0
	2021	-1,7	6,0	0,0	0,0	0,0	8,3	0,0	0,0
	2022	-2,6	2,6	0,0	0,0	0,0	0,0	0,0	0,0
	2023	-1,7	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	2024	-1,7	0,9	0,0	0,0	0,0	0,0	0,0	0,0
	2025/01-10	-4,3	1,7	0,0	0,0	0,0	0,0	0,0	0,0
Index.de	letzte 36 Mon.	-26,8	29,7	38,9	11,1	5,6	75,0	38,9	22,2
	2019	-15,3	13,1	45,5	10,0	0,0	63,6	50,0	44,4
	2020	-31,3	27,4	33,3	8,3	0,0	50,0	8,3	0,0
	2021	-5,6	20,5	33,3	0,0	0,0	66,7	16,7	8,3
	2022	-12,1	14,4	33,3	0,0	0,0	91,7	41,7	25,0
	2023	-26,8	21,7	41,7	25,0	16,7	66,7	41,7	33,3
	2024	-20,9	17,9	41,7	0,0	0,0	83,3	33,3	8,3
	2025/01-10	-20,9	29,7	20,0	10,0	0,0	70,0	50,0	30,0
Indeed	letzte 36 Mon.	-4,7	2,6	0,0	0,0	0,0	0,0	0,0	0,0
	2019	--	--	--	--	--	--	--	--
	2020	-14,5	4,5	10,0	0,0	0,0	10,0	0,0	0,0
	2021	-0,8	9,2	0,0	0,0	0,0	41,7	0,0	0,0
	2022	-2,1	8,5	0,0	0,0	0,0	16,7	0,0	0,0
	2023	-4,5	2,6	0,0	0,0	0,0	0,0	0,0	0,0
	2024	-4,7	-0,1	0,0	0,0	0,0	0,0	0,0	0,0
	2025/01-10	-4,0	1,4	0,0	0,0	0,0	0,0	0,0	0,0

Quelle: Eigene Darstellung.

ABBILDUNG 6 **OJA-Indizes: Textkernel-Daten, Outlier und Winsor15 im Vergleich zu BA-X und IAB**



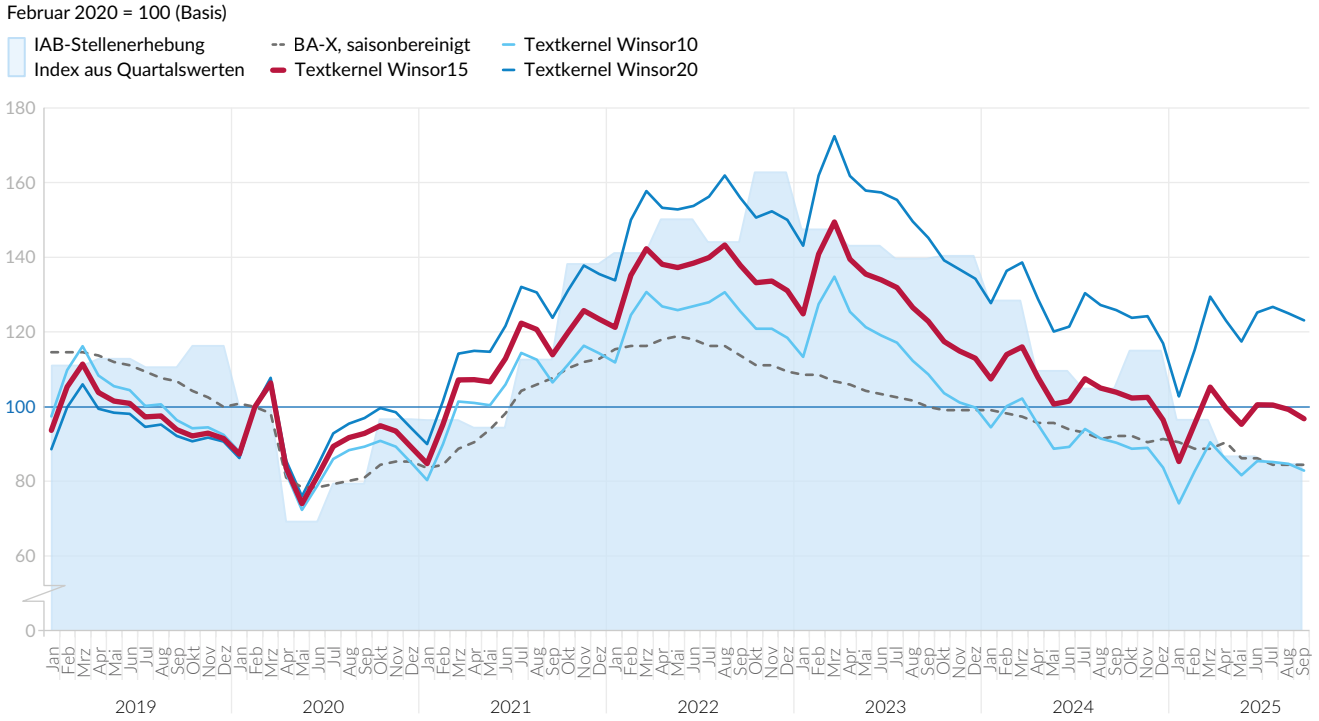
Quelle: Eigene Darstellung.

JOBMONITOR | BertelsmannStiftung

Auch die Indizes von BA-X, Index.de und Indeed weisen mit -17,2 Prozentpunkten, -31,3 Prozentpunkten und -14,2 Prozentpunkten jeweils die stärkste Vormonatsveränderung im Corona-Jahr 2020 auf. Sowohl der (saisonbereinigte) BA-X als auch der (nicht saisonbereinigte) Indeed-Index zeigen dabei nur im Krisenjahr 2020 Large-|10|-M-M-changes an, während diese in den anderen Jahren bei null liegen. Auch nehmen sowohl die Large-|10|-Inversions als auch die Double-Large-|10|-Inversions in allen anderen Jahren den Wert Null an. Beide Indizes bilden damit eindeutig die Corona-Krise ab und zeigen gleichzeitig die geringste Volatilität im Zeitverlauf. Dies stellt sich beim index-Stellenindikator etwas anders dar: Er ist deutlich volatiler und kann die Corona-Krise nicht anhand der Large-|10|-M-M-changes identifizieren, da auch in anderen Berichtsjahren vergleichbar hohe oder sogar höhere Werte auftauchen. Auch weist er mehrfach Large-|10|-Inversions und sogar Double-Large-|10|-Inversions auf, sodass der Index von Index.de anhand der Volatilitätsmaße insgesamt schlechter als der Textkernel-Winsor15-Index dasteht.

Viele der oben beschriebenen Ergebnisse zu den Volatilitätsmaßen zeigen sich auch in der grafischen Vergleichsanalyse. Der besseren Übersichtlichkeit halber wurde diese nachfolgend in die Abbildungen 6–8 aufgeteilt. Abbildung 6 vergleicht den einfachen Textkernel-Index, den Textkernel-Outlier-Index und den Textkernel-Winsor15-Index mit dem BA-X und einem Index aus den vierteljährlichen offenen Stellen der IAB-Stellenerhebung. Es zeigt sich, dass Textkernel-

ABBILDUNG 7 **OJA-Indizes: Textkernel-Daten mit Winsor10, 15 und 20 im Vergleich zu BA-X und IAB**



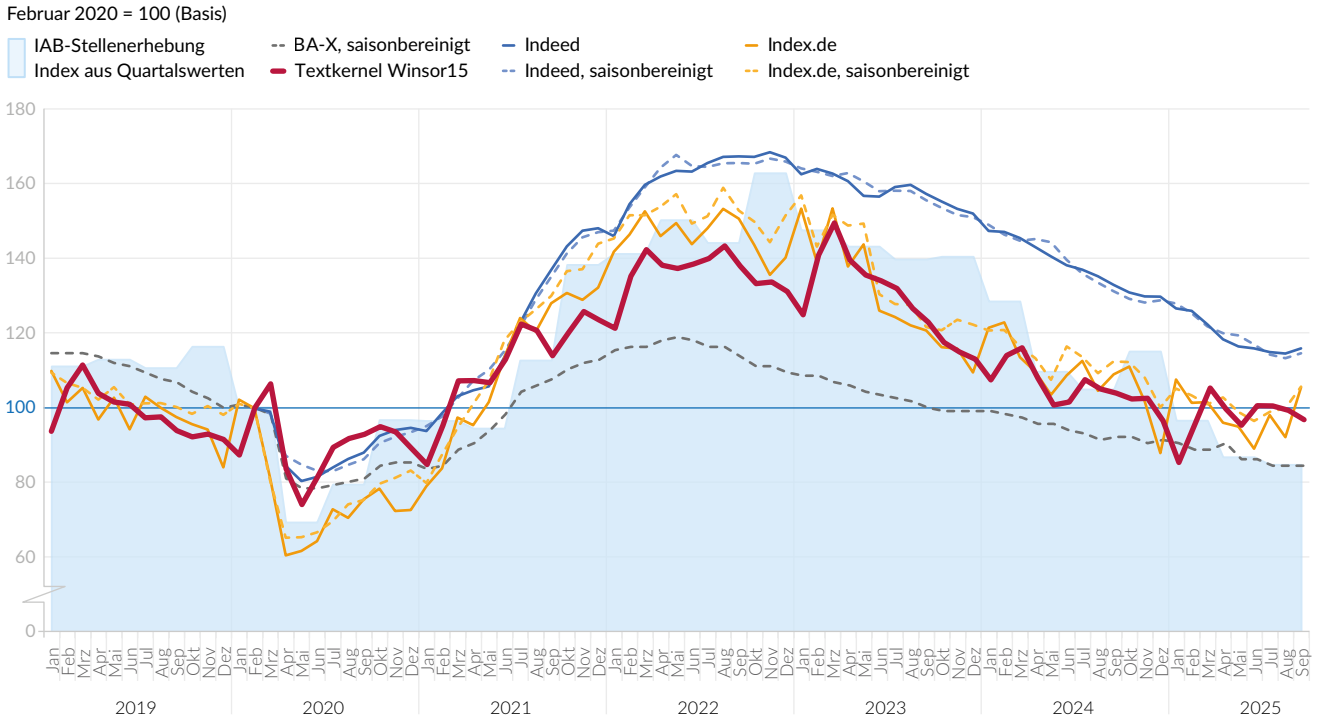
Quelle: Eigene Darstellung.

JOBMONITOR | BertelsmannStiftung

Index und Textkernel-Outlier-Index sich vom Verlaufsmuster her wenig verändern, während die Bereinigung nach dem Ensemble-Ansatz der Winsorisierung der Änderungsraten den Zeitreihenverlauf deutlich verändert (so zum Beispiel im Juni 2019, in dem der Outlier-Index steigt und der Winsor15-Index sinkt, und im März 2023, in dem der Winsor15-Index fällt und der Outlier-Index steigt). Wie Abbildung 7 zeigt, hat die Höhe der Winsorisierungsrate kaum Einfluss auf das Verlaufsmuster. Die unterschiedlichen Winsoriserungsraten verändern in erster Linie das Niveau des Index und das Ausmaß der Amplituden, was sich auch in den Volatilitätsmaßen widerspiegelt.

Abbildung 8 zeigt alle verschiedenen monatlichen Indizes im Vergleich zum Index aus den vierteljährlichen offenen Stellen der IAB-Stellenerhebung, wobei hier für die Textkernel-Daten exemplarisch der Winsor15-Index dargestellt wurde, obwohl der Winsor20-Index durchaus mindestens genauso gut geeignet wäre. Es ist deutlich zu erkennen, dass der BA-X-Index und die Indeed-Indizes die geringste Volatilität aufweisen, was an der Saisonbereinigung liegt. Allerdings gilt dies auch für den nicht saisonbereinigten Indeed-Index, der sich kaum vom saisonbereinigten Indeed-Index unterscheidet. Die hohe Volatilität des index-Stellenmarktindikators ist bei dem saisonbereinigten Indikator von Index.de nur mäßig abgeschwächt.

ABBILDUNG 8 **OJA-Indizes: Textkernel-Daten, BA-X, Index.de und Indeed im Vergleich zu IAB**



Quelle: Eigene Darstellung.

JOBMONITOR | BertelsmannStiftung

Alle Indizes zeigen den Corona-Bruch. Der coronabedingte Abschwung wird allerdings je nach Index in völlig unterschiedlicher Größenordnung abgebildet. Tabelle 6 vergleicht, welchen Rückgang die einzelnen Indizes von Februar bis Mai 2020 aufzeigen. So verzeichnet der aus den vierteljährlichen offenen Stellen der IAB-Stellenerhebung ermittelte Index einen Rückgang um 30,8 %, während der saisonbereinigte BA-X um 21,6 %, der Textkernel-Winsor15-Index um 27,9 %, der nicht saisonbereinigte index-Stellenindikator um 38,3 % und der nicht saisonbereinigte Indeed-Index um 15,1 % zurückgeht.

TABELLE 6 **Ausmaß des Corona-Bruchs bei verschiedenen Indizes (in %)**

Index	Rückgang von 2020-02 zu 2020-05
Index aus IAB-Stellenerhebung (bis 2023-09)	-30,8
Textkernel	-13,7
Textkernel Outlier	-12,8
Textkernel winsor10	-23,9
Textkernel winsor15	-25,9
Textkernel winsor20	-27,6
BA-X, saisonbereinigt	-21,6
Indexgruppe, nicht saisonbereinigt	-38,3
Indexgruppe, saisonbereinigt	-34,6
Indeed	-15,1
Indeed, saisonbereinigt	-19,5
BIP (preisbereinigt)*	-10,5

* Nur zu Informationszwecken. | Quelle: Eigene Berechnungen.

6. Fazit und Ausblick

Die Studie zeigt, dass Online-Stellenanzeigen ein wertvolles Instrument zur Analyse der Arbeitsmarktnachfrage darstellen und traditionelle Ansätze wie die BA-Statistik und die IAB-Stellenerhebung sinnvoll ergänzen können. Die Stärke des OJA-Ansatzes liegt in seiner zeitnahen Datenverfügbarkeit, die insbesondere in Phasen wirtschaftlicher Unsicherheit oder Krisen wertvolle Einblicke liefert. Gleichzeitig wurden Limitationen deutlich, insbesondere in Bezug auf die Repräsentativität der Daten, technisch bedingte Artefakte und die Notwendigkeit von Annahmen zur Transformation von Flow- in Stock-Variablen.

Ein zentraler Aspekt ist die Datenqualität, die maßgeblich von der Erhebung und Bereinigung der OJA-Daten abhängt. Online-Stellenanzeigen unterliegen Verzerrungen durch eine ungleiche Repräsentation verschiedener Branchen und Unternehmensgrößen. Beispielsweise werden Positionen im Baugewerbe, in der Landwirtschaft oder in kleinen Unternehmen seltener online ausgeschrieben, während höherqualifizierte Stellen überproportional stark vertreten sind. Darüber hinaus führen technische Artefakte wie Drift (zunehmende Anzahl erfasster Portale), Duplikate (mehrfache Ausschreibungen derselben Stelle) oder Step-Changes (plötzliche Veränderungen in der Datenmenge durch Scraping-Prozesse) zu zusätzlichen Herausforderungen. Zudem können OJA-Daten generell das gesamte Stellenangebot des Marktes nur unvollständig abdecken, weil sie nur den Online-Stellenmarkt abbilden (erstes Coverage-Problem). Auch die Repräsentativität des Online-Stellenmarktes kann aus verschiedenen Gründen beeinträchtigt sein (zweites Coverage-Problem).

Die Transformation der Flow-Variable (neue Stellenanzeigen pro Zeitraum) in eine Stock-Variable (Gesamtbestand an Stellenanzeigen) erfordert zudem Annahmen über die durchschnittliche Vakanzdauer, die zwischen Branchen und Regionen stark variieren kann. Diese Annahmen beeinflussen die Genauigkeit der Ergebnisse und können insbesondere in dynamischen wirtschaftlichen Phasen zu Abweichungen führen. Weiterhin ist die Datenverfügbarkeit von der Zuverlässigkeit der Scraping-Prozesse abhängig, die durch technische Änderungen auf den Plattformen oder Zugriffsbeschränkungen eingeschränkt werden können. Und schließlich hängt die Repräsentativität der OJA-Daten auch von den verwendeten Quellen und deren Abdeckung ab, was zu systematischen Lücken in den erfassten Stellenanzeigen führen kann.

Während das Interesse an der Entwicklung monatlicher (oder sogar wöchentlicher) Indizes auf Basis von OJA-Daten insbesondere infolge der Corona-Krise signifikant zugenommen hat, haben gleichzeitig zahlreiche methodische Studien zu OJA-Daten der jüngeren Vergangenheit die damit verbundenen großen Herausforderungen aufgezeigt. Im vorliegenden Beitrag wurde versucht, diesen Herausforderungen mit einem mehrstufigen Verfahren zur Berechnung von Konjunkturindizes aus Online-Stellenanzeigen zu begegnen. Ausgangsbasis war dabei die Studie von Evans u. a. (2023), die einen Algorithmus zur verbesserten Abbildung zeitnaher, hochaktueller Entwicklungen des Stellenmarktes durch OJA-Daten entwickelt hatten.

Dabei wurden zwei unterschiedliche Ensemble-Ansätze vorgestellt, die darauf beruhen, Stellenanzeigen aus unterschiedlichen Quellen zunächst unabhängig voneinander zu verarbeiten und sie erst im Anschluss zu einem Gesamtbild zusammenzuführen. Diese Vorgehensweise unterscheidet sich wesentlich von Methoden, bei denen die Daten zu Beginn quellenübergreifend aggregiert und anschließend gemeinsam verarbeitet werden, beispielsweise durch eine einheitliche Glättung sämtlicher Stellenanzeigen über alle Quellen hinweg.

Im ersten Ansatz wird jede Zeitreihe der absoluten Werte der einzelnen OJA-Quellen von Ausreißern bereinigt, die monatliche Anzahl der Stellenanzeigen dann aggregiert und anschließend indexiert (Textkernel-Outlier-Index). Im zweiten Ansatz, der unmittelbar auf der Arbeit von Evans und anderen basiert, erfolgt eine Ausreißerbereinigung der Zeitreihen über die Winsorisierung der Änderungsraten der einzelnen

OJA-Quellen, bevor die Daten zu einem Index aggregiert werden (Textkernel-Winsor-Index).

Die Validierung der entwickelten Indizes aus Textkernel-Daten zeigt, dass der Textkernel-Winsor-Index gegenüber dem Textkernel-Outlier-Index deutliche Vorteile bietet. So weist der Winsor-Index beispielweise eine höhere Korrelation mit den Indizes der etablierten IAB- und BA-Statistiken auf. Ein zentraler Grund für die bessere Performance des Winsor-Index ist die robuste Bereinigung um Ausreißer durch die Winsorisierung. Diese Methode stellt sicher, dass extreme Werte in den Änderungsraten einzelner Portale das Aggregat nicht verzerren, während gleichzeitig eine größere Datenbasis berücksichtigt wird. Der Outlier-Index hingegen eliminiert extreme Werte komplett, was zu einer Reduktion der Sensitivität für kurzfristige Schwankungen führt.

Insgesamt verdeutlichen die Ergebnisse, dass OJA-basierte Indizes nicht nur als Ergänzung, sondern in spezifischen Anwendungsbereichen als eigenständiges Instrument zur Arbeitsmarktanalyse genutzt werden können. Die Kombination von OJA-Daten mit traditionellen Erhebungsmethoden bietet großes Potenzial, um eine umfassendere, dynamischere und zeitnähere Analyse des Arbeitsmarktes zu ermöglichen. Künftige Forschungen könnten sich darauf konzentrieren, hybride Modelle zu entwickeln und die Methodik der OJA-Datenanalyse weiter zu verbessern, um ihr volles Potenzial für die Arbeitsmarktforschung zu erschließen.

Die mit OJA-Daten verbundenen großen Herausforderungen erfordern allerdings eine kontinuierliche Weiterentwicklung der Methodik, um die Aussagekraft und Qualität der Ergebnisse weiter zu optimieren. Ideen für verbesserte Modelle existieren bereits. Ansatzpunkt kann beispielsweise die durch den Ensemble-Ansatz der winsorisierten Änderungsraten bedingte Egalisierung der verschiedenen OJA-Quellen sein. Beim hier beschriebenen Verfahren wurden die winsorisierten Änderungsraten jeder einzelnen OJA-Quelle im letzten Schritt vor der Indexbildung durch den Mittelwert der Änderungsraten zusammengefasst. Dadurch bekommt die Entwicklung der Anzahl der Online-Stellenanzeigen auf kleineren Jobportalen dasselbe Gewicht wie diejenige größerer Jobportale.

Dieses Vorgehen kann sinnvoll, aber auch problematisch sein. In der Realität kann die Entwicklung des Online-Stellenmarktes durchaus stark von einer Branche dominiert werden. Die Entwicklung einer großen beschäftigungsintensiven Branche (auch arbeits- und lohnintensiv genannt) hat einen stärkeren Einfluss auf die Gesamtentwicklung der Arbeitskräftenachfrage als die Entwicklung einer großen kapitalintensiven Branche. Entwickeln sich einzelne OJA-Quellen gegenläufig zur allgemeinen Entwicklung, *weil* sie speziell auf bestimmte Branchen ausgerichtet sind, kann dies zu einem verzerrten Gesamtindikator führen. Beispielsweise gibt es einzelne Jobportale, die auf bestimmte Branchen wie das Hotelgewerbe oder Gastronomie spezialisiert sind.

Eine ungewichtete Mittelwertbildung der Änderungsraten aller einzelnen OJA-Quellen kann also bei gegenläufigen Entwicklungen der Arbeitskräftenachfrage zu einer verzerrten Abbildung der gesamtwirtschaftlichen Entwicklung führen. Dem könnte prinzipiell mit einem Branchengewicht wie etwa der branchenspezifischen Erwerbstätigenzahl entgegenwirkt werden, wenn die oben genannten Änderungsraten nicht für jeweils eine OJA-Quelle, sondern getrennt für jede Branche berechnet werden.

Auch Evans und andere schlagen ein Gewichtungsschema vor, allerdings haben sie eine andere Idee als die gerade skizzierte: Sie nennen als Beispiel die am JVS gemessene Vorhersagegenauigkeit einer Quelle für einen bestimmten Zeitraum (Evans u. a. 2023).

Unabhängig vom Einsatz einer Gewichtung der Änderungsraten gibt es schließlich auch die ganz einfache Möglichkeit, einen OJA-Index über einen Bottom-Up-Ansatz aus Subindizes zu bilden, die beispielweise branchen-, berufsgruppen- oder regional differenziert ermittelt wurden.

Literaturverzeichnis

Azar, J./Marinescu, I./Steinbaum, M./Taska, B. (2020). Concentration in US labor markets: Evidence from online vacancy data. Labour Economics, 66, 101886.

Bleninger, P./Kettner, A./Pausch, S./Szameitat, J. (2012). Können offene Stellen als Vorlaufindikator für Neueinstellungen dienen? Ergebnisse aus der IAB-Erhebung des gesamtwirtschaftlichen Stellenangebots. IAB-Forschungsbericht 201204. Institut für Arbeitsmarkt- und Berufsforschung (IAB), Nürnberg. https://doku.iab.de/forschungsbericht/2012/fb0412.pdf

Bossler, M./Gürtzgen, N./Kubis, A./Moczall, A. (2018). IAB-Stellenerhebung von 1992 bis 2017: So wenige Arbeitslose pro offene Stelle wie nie in den vergangenen 25 Jahren. IAB-Kurzbericht 23/2018. Institut für Arbeitsmarkt- und Berufsforschung (IAB), Nürnberg. https://doku.iab.de/kurzber/2018/kb2318.pdf

Bossler, M./Gürtzgen, N./Kubis, A./Küfner, B./Lochner, B. (2020a). The IAB Job Vacancy Survey: Design and research potential. Journal for Labour Market Research, 54(1), 1–12. https://doi.org/10.1186/s12651-020-00278-6

Bossler, M./Gürtzgen, N./Kubis, A./Küfner, B. (2020b). IAB-Stellenerhebung im ersten Quartal 2020: Mit dem Corona-Shutdown ging zuerst die Zahl der offenen Stellen zurück. IAB-Kurzbericht Nr. 12/2020. Institut für Arbeitsmarkt- und Berufsforschung (IAB), Nürnberg. https://www.econstor.eu/bitstream/10419/234200/1/kb2020-12.pdf

Brenzel, H./Czepek, J./Kiesl, H./Kriechel, B./Kubis, A./Moczall, A./Rebien, M./Röttger, C./Szameitat, J./Warning, A./Weber, E. (2016): Revision der IAB-Stellenerhebung: Hintergründe, Methode und Ergebnisse. IAB-Forschungsbericht Nr. 4/2016. Institut für Arbeitsmarkt- und Berufsforschung (IAB), Nürnberg. https://doku.iab.de/forschungsbericht/2016/fb0416.pdf

Bundesagentur für Arbeit (2018). Statistik/Arbeitsmarktberichterstattung. Grundlagen: Qualitätsbericht – Statistik der gemeldeten Arbeitsstellen, Nürnberg, August 2018. https://statistik.arbeitsagentur.de/DE/Statischer-Content/Grundlagen/Methodik-Qualitaet/Qualitaetsberichte/Generische-Publikationen/Qualitaetsbericht-Statistik-gemeldete-Arbeitsstellen.pdf

Bundesagentur für Arbeit (2020). Statistik/Arbeitsmarktberichterstattung. Grundlagen: Methodenbericht – Revision des BA-Stellenindex (BA-X), Nürnberg, Januar 2020. https://statistik.arbeitsagentur.de/DE/Statischer-Content/Grundlagen/Methodik-Qualitaet/Methodenberichte/gemeldete-Arbeitsstellen/Generische-Publikationen/Methodenbericht-Revision-BAX.pdf

Bundesagentur für Arbeit (2026). Kurzinfo BA-Stellenindex BA-X. September 2026. https://statistik.arbeitsagentur.de/Statistikdaten/Detail/202609/arbeitsmarktberichte/bax-ba-x/ba-x-d-0-202609-pdf.pdf?__blob=publicationFile&v=2

Burstedde, A./Flake, R./Jansen, A./Malin, L./Risius, P./Seyda, S./Schirner, S./Werner, D. (2020). Die Messung des Fachkräftemangels. Methodik und Ergebnisse aus der IW-Fachkräftedatenbank zur Bestimmung von Engpassberufen und zur Berechnung von Fachkräftelücken und anderen Indikatoren. IW-Report Nr. 59. Köln. https://www.iwkoeln.de/fileadmin/user_upload/Studien/Report/PDF/2020/IW-Report_2020_Messung_des_Fachkraeftemangels.pdf

De Lazzer, J./Rengers, M. (2021). Auswirkungen der Coronakrise auf den Arbeitsmarkt: experimentelle Statistiken aus Daten von Online-Jobportalen. WISTA – Wirtschaft und Statistik, 73(3), 71–88. https://www.destatis.de/DE/Methoden/WISTA-Wirtschaft-und-Statistik/2021/03/auswirkungen-coronakrise-arbeitsmarkt-032021.pdf

Deutscher Bundestag (2018). Antwort der Bundesregierung auf die Kleine Anfrage der Abgeordneten Jessica Tatti, Susanne Ferschl, Matthias W. Birkwald, weiterer Abgeordneter und der Fraktion DIE LINKE. Fragen zum Sozialdatenschutz und zu möglichen fingierten Stellenanzeigen im Stellenportal „Jobbörse“ der Bundesagentur für Arbeit v. 31.05.2018, BT-Drucksache 19/2417.

Evans, D./Mason, C./Chen, H./Reeson, A. (2023). An algorithm for predicting job vacancies using online job postings in Australia. Humanities and Social Sciences Communications, 10, 102. https://doi.org/10.1057/s41599-023-01562-9

Fabo, B./Kureková, L. M. (2022). Methodological issues related to the use of online labour market data. ILO Working Paper 68. International Labour Organization, Genf. https://webapps.ilo.org/static/english/intserv/working-papers/wp068/index.html

Glynn, C. (2022). Introducing the Indeed Job Postings Index. https://www.hiringlab.org/2022/12/15/introducing-the-indeed-job-postings-index/

Gürtzgen, N./Kubis, A./Popp, M. (2023). IAB-Monitor Arbeitskräftebedarf 1/2023: Verhältnis von Arbeitslosen zu offenen Stellen ist mit 1,5 nach wie vor niedrig. IAB-Forum, 1. Juni 2023. https://www.iab-forum.de/iab-monitor-arbeitskraeftebedarf-1-2023/

Hyndman (2021). Detecting time series outliers. https://robjhyndman.com/hyndsight/tsoutliers

Kettner, A./Klinger, S. (2012). Entwicklung der offenen Stellen: Überlagerung von Konjunktur und Struktur. Wirtschaftsdienst, 92(10), 712–714. https://www.econstor.eu/bitstream/10419/68419/1/728593467.pdf

Luedicke, J. (2015). SUPSMOOTH: Stata module to perform Friedman's super smoother. Statistical Software Components, Boston College Department of Economics. https://ideas.repec.org/c/boc/bocode/s458030.html

Napierala, J. (2023). The feasibility of using online job advertisements in analysing unmet EU demand. Cedefop Working Paper No. 18. Luxembourg: Publications Office. https://www.cedefop.europa.eu/files/6218_en.pdf

Napierala, J./Kvetan, V./Branka, J. (2022). Assessing the representativeness of online job advertisements. Cedefop Working Paper No. 17. Luxembourg: Publications Office. https://www.cedefop.europa.eu/files/6217_en.pdf

Office for National Statistics ONS (2021). Using Adzuna data to derive an indicator of weekly vacancies: Experimental Statistics. https://www.ons.gov.uk

Office for National Statistics ONS (2024). Measuring labour demand volumes across the UK using Textkernel data user guide. https://www.ons.gov.uk

Rengers, M. (2018a). Internetbasierte Erfassung offener Stellen im Statistischen Bundesamt. In: König, C. u. a. (Hrsg.). Big Data – Chancen, Risiken, Entwicklungstendenzen. Schriftenreihe der ASI – Arbeitsgemeinschaft Sozialwissenschaftlicher Institute. Wiesbaden 2018, S. 61ff.

Rengers, M. (2018b). Internetgestützte Erfassung offener Stellen – Machbarkeitsstudie im Rahmen eines ESSnet-Projekts zu Big Data. WISTA – Wirtschaft und Statistik, 5/2018, 11–33. https://www.destatis.de/DE/Methoden/WISTA-Wirtschaft-und-Statistik/2018/05/internetgestuetzte-erfassung-052018.pdf

Rengers, M (2020). Developing meaningful indicators based on OJA [online job vacancies] data. In: ESSnet Big Data, Work Package B: Implementation – Online Job Vacancies, August 2020. https://doi.org/10.13140/RG.2.2.30283.04646

Rengers, M./Hornik, A. (2024). Daten zu Online-Stellenanzeigen im ESSnet WIN. WISTA – Wirtschaft und Statistik, 5/2024, 49–61. https://www.destatis.de/DE/Methoden/WISTA-Wirtschaft-und-Statistik/2024/05/online-stellenanzeigen-essnet-win-052024.pdf

Štefánik, M./Lyócsa, Š./Bilka, M. (2022). Using online job postings to predict key labour market indicators. Social Science Computer Review, 41(5), 1630–1649. https://doi.org/10.1177/08944393221085705

Tukey, J. (1977). Exploratory data analysis. Reading: Addison-Wesley.

Zhao, Y./Chen, H./Mason, C. M. (2021). A framework for duplicate detection from online job postings. In IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, S. 249–256.

jobmonitor.de

Ansprechpartner
Gunvald Herdin
Senior Project Manager
Nachhaltige Soziale Marktwirtschaft
E-Mail: gunvald.herdin@bertelsmann-stiftung.de
Telefon +49 5241 81-81311

bertelsmann-stiftung.de

Bertelsmann Stiftung
Carl-Bertelsmann-Straße 256
33311 Gütersloh
Telefon +49 5241 81-0

DESTATIS
Statistisches Bundesamt

BertelsmannStiftung