FORTSCHRITTE BEIM EINSATZ VON WEB SCRAPING IN DER AMTLICHEN VERBRAUCHERPREISSTATISTIK – EIN WERKSTATTBERICHT

Christian Blaudow, Daniel Seeger

Schlüsselwörter: Verbraucherpreisindex – Web Scraping – Digitalisierung – dynamische Preissetzung

ZUSAMMENFASSUNG

Für den deutschen Verbraucherpreisindex und den Harmonisierten Verbraucherpreisindex werden derzeit monatlich für etwa 10 000 Produkte Preise im Internet erhoben. Im Regelfall erfolgt dies manuell durch Preiserheberinnen und Preiserheber, und zwar zu einem festgelegten Zeitpunkt. Im Statistischen Bundesamt wird intensiv daran gearbeitet, diese manuellen Preiserhebungen bis Ende 2020 mit der Hilfe von Web Scraping zu automatisieren. Auch werden neue Methoden entwickelt, um große Datenmengen in die Berechnung von Preisindizes zu integrieren. Der vorliegende Artikel stellt das in der Entwicklung befindliche generische Programm für die Nutzung von Web Scraping vor und gibt darüber hinaus einen Überblick über den Forschungsstand in der Methodenentwicklung.

Keywords: *consumer price index – web scraping – digitalisation – dynamic pricing*

ABSTRACT

At present, the prices of roughly 10,000 products are collected on the internet each month for the German Consumer Price Index and the Harmonised Index of Consumer Prices. Usually, this is done manually by price collectors, and at a given point of time. At the Federal Statistical Office, intensive work is underway to automate the manual price collection by means of web scraping by the end of 2020. Also, new methods are being developed to include large data volumes in the calculation of price indices. This article presents the generic program that is being developed for the use of web scraping and gives an overview of the state of research in the field of methods development.

Christian Blaudow

hat International Economics an der Universität Göttingen studiert und ist Referent im Referat „Zwischenörtliche Preisvergleiche, Parteienindex“ des Statistischen Bundesamtes. Schwerpunkte seiner Arbeit sind die Automatisierung von Preiserhebungen im Internet sowie die Entwicklung von Methoden für die Integration von und den Umgang mit großen Datenmengen.

Daniel Seeger

ist Diplom-Volkswirt und leitet das Referat „Verbraucherpreise“ im Statistischen Bundesamt. Er ist insbesondere zuständig für die Koordinierung der Zusammenarbeit mit den Statistischen Ämtern der Länder und arbeitet auf europäischer Ebene an der Weiterentwicklung der Methoden des Harmonisierten Verbraucherpreisindex mit.

1

Einleitung

Der Internethandel gewinnt stetig an Bedeutung. Dementsprechend erfasst das Statistische Bundesamt für die Verbraucherpreisstatistik zunehmend Preise der im Internet gehandelten Waren und Dienstleistungen. Derzeit werden für die Verbraucherpreisstatistik jeden Monat Preise für etwa 10 000 Produkte im Internet erhoben, überwiegend manuell. Der Anteil der Produkte, für die Preise im Internet erhoben werden, hat sich im Verbraucherpreisindex (VPI) beziehungsweise im harmonisierten Verbraucherpreisindex (HVPI) seit dem Basisjahr 2005 bis zum Basisjahr 2015 von 5 % auf mehr als 10 % erhöht. Diese Aussage gilt zumindest bei den Gütern des Warenkorbs, für die eine Einteilung der Konsumausgaben der privaten Haushalte in verschiedene Geschäftstypen[1] (Sandhop, 2012) möglich ist. Der steigende Anteil resultiert auch daraus, dass große Einzelhandelsunternehmen neben ihrem Verkauf in stationären Geschäften ihre Produkte verstärkt auch in Onlineshops anbieten. Zum Teil reduzieren sie ihr Sortiment im stationären Handel sogar zugunsten des Internethandels. Der steigende Anteil der Waren und Dienstleistungen, die im Internet gehandelt werden, hat zur Folge, dass die Stichprobe der Preiserhebung im Onlinehandel künftig weiter ausgeweitet werden muss.

Parallel wird in der Preisstatistik daran gearbeitet, die Frequenz der Preiserhebungen zu optimieren. Studien über das Preissetzungsverhalten von Onlinehändlern haben gezeigt, dass sich Preise im Onlinehandel häufiger ändern als im stationären Handel (Blaudow/Burg, 2018). Eine höhere Frequenz der Preiserhebungen ist nur durch einen verstärkten Einsatz automatisierter Verfahren wie Web Scraping zu erreichen. Zusätzlich werden neue Methoden entwickelt, um die zu erwartenden Datenmengen zu validieren und in die monatliche Indexberechnung des VPI und des HVPI zu integrieren.

Der vorliegende Beitrag gibt einen Überblick über die anstehenden Veränderungen in der Internetpreiserhebung, sowohl aus technischer als auch aus methodischer Sicht.

1 Hierbei muss festgehalten werden, dass sich die Geschäftstypengewichtung nur auf etwa ein Drittel aller Güter im Warenkorb bezieht.

2

Der Einsatz von Web Scraping in der Verbraucherpreisstatistik

Seit dem Jahr 2012 werden im Statistischen Bundesamt die Möglichkeiten automatisierter Preiserhebungen im Internet mithilfe von Web Scraping untersucht. Finanziert wurden und werden diese Arbeiten weitgehend durch das Statistische Amt der Europäischen Union (Eurostat). Dabei wurde von Beginn an ein imitierender Ansatz gewählt, bei dem durch Web Scraping einzelne, manuelle Preiserhebungen im Internet automatisiert werden. Das heißt, zu bestimmten Zeitpunkten werden automatisiert Preise für zuvor definierte Produkte extrahiert, die ansonsten manuell erhoben wurden. Die Internetseite, auf der das jeweilige Produkt angeboten wird, kann entweder mit der entsprechenden URL[2] direkt aufgerufen werden. Oder es wird anhand von Informationen, die in einer Eingangsdatenbank liegen oder im Programm errechnet werden, Schritt für Schritt zur Internetseite gesteuert. Das Programm erkennt anhand der festgelegten Definition das Produkt und extrahiert den gesuchten Preis. Somit wird die manuelle Arbeit der Preiserheberinnen und Preiserheber imitiert, ist jedoch in beliebiger Frequenz wiederholbar. Im Gegensatz dazu könnten auch die Preise aller verfügbaren Produkte beispielsweise eines Onlinehändlers ohne vorausgehende Produktauswahl[3] abgegriffen werden. Dieser Ansatz wird vom Statistischen Bundesamt bislang nicht verfolgt.

Derzeit wird Web Scraping in der Preisstatistik isoliert vom hausinternen Netzwerk des Statistischen Bundesamtes mit projekteigener IT-Ausstattung betrieben. Das eigentliche Programm, welches die Automatisierungsprozesse steuert, ist in der Programmiersprache Java geschrieben. Es verwendet SQL[4]-Datenbanken für Eingangs- und Ausgangsdaten sowie entweder das Scrapingtool iMacros oder Selenium. Die Programmierung kann ebenso wie die Programm- und Datenpflege nur von wenigen Personen mit speziellen IT-Kenntnissen durchgeführt werden.[5]

2 Uniform Resource Locator – einheitlicher Ressourcenzeiger. Die URL ist ein Standard, mit dem bestimmte Inhalte aufgerufen werden; häufig wird darunter die Adresse einer Website verstanden.

3 Sogenanntes bulk scraping.

4 Structured Query Language.

5 Für zusätzliche Erläuterungen siehe Blaudow (2018).

Aktuell werden bereits für mehrere Güterbereiche Preise für die Verbraucherpreisstatistik erfolgreich mittels Web Scraping erhoben, beispielsweise für Mietwagen, für zahlreiche Arten von Medikamenten oder auch für Fernbus- und Bahnreisen.

Über die Preiserhebungen für den VPI/HVPI hinaus wird Web Scraping auch in anderen Bereichen der Preisstatistik erfolgreich eingesetzt:

- Für die Kaufkraftparitäten werden Preise für Hotelübernachtungen und Flugreisen erhoben.
- Für den Preisindex des selbstgenutzten Wohneigentums werden Wohngebäudeversicherungstarife erhoben.
- Für den Erzeugerpreisindex werden Gas- und Strompreise von mehreren Energiebörsen automatisiert aus dem Internet gewonnen.

Je nach Statistik und Volatilität der Preise werden die automatisierten Preiserhebungen täglich, wöchentlich oder monatlich durchgeführt.

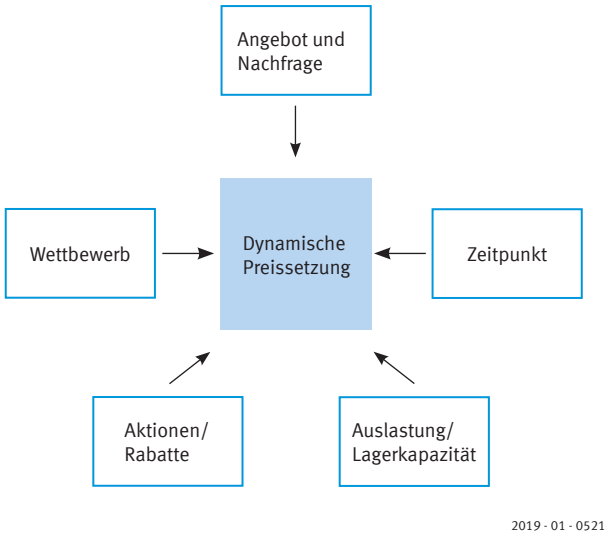
3

Die Herausforderung durch dynamische Preissetzung im Onlinehandel

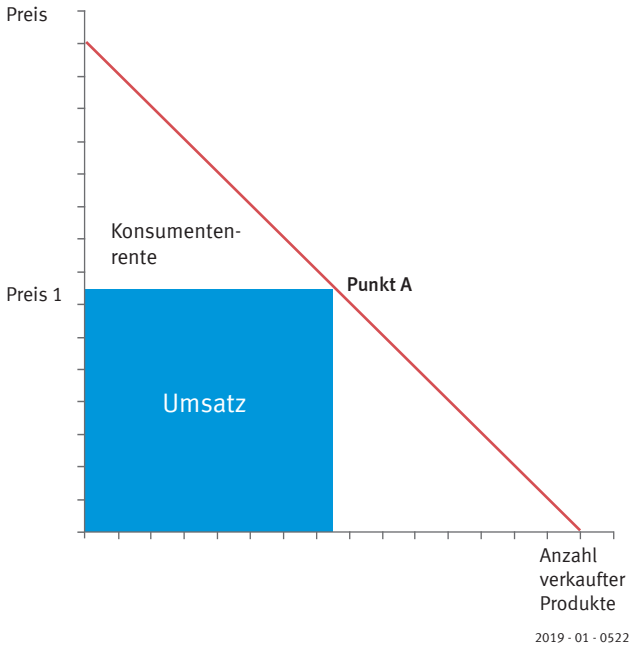
Im Internethandel ist es mit wenig Aufwand möglich, Preise anhand von Algorithmen in sehr kurzen Zeitabständen automatisiert und annähernd unbegrenzt zu ändern. Dieses Phänomen wird allgemein als dynamische Preissetzung bezeichnet. Die hierfür verwendeten Algorithmen werden durch Faktoren wie Angebot und Nachfrage, Zeitpunkt des Kaufs und Rabattaktionen beeinflusst.[6] ↘ Grafik 1

In einem Markt mit einer statischen Preisgestaltung (siehe ↘ Grafik 2) bieten die Produzenten ihre Produkte zu einem Preis an, der für alle Kundinnen und Kunden gleich ist und für einen absehbaren Zeitabschnitt (beispielsweise über einen Monat) gleich bleibt. Ein Produzent kennt nicht die individuelle Zahlungsbereitschaft der Kundschaft und wird einen Preis 1 setzen, der möglichst viele Kundinnen und Kunden zum Kauf seines Produkts veranlasst. Die Kundschaft kann auf diese Weise das Produkt und dessen Preis mit Produkten anderer

Grafik 1
Auswahl an Einflussfaktoren für dynamische Preissetzung



Grafik 2
Statische Preisgestaltung



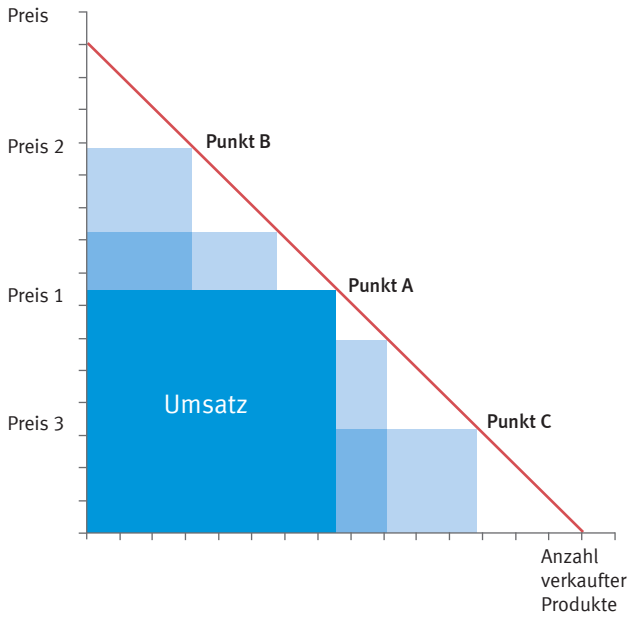
6 In der Studie von Blaudow/Burg (2018) wurde beispielweise festgestellt, dass ein Großteil der Preisänderungen bei größeren Onlineshops während der ersten Tageshälfte auftritt.

Produzenten vergleichen, um ihre Konsumentenrente[7] zu maximieren.

Liegt die individuelle Zahlungsbereitschaft einer Kundin oder eines Kunden oberhalb des gesetzten Preises, ist die Konsumentenrente beim Kauf des Produkts positiv. Liegt der Preis hingegen oberhalb der individuellen Zahlungsbereitschaft, würde sich die Kundin oder der Kunde gegen einen Kauf entscheiden und möglicherweise ein Substitut bei einem anderen Produzenten zu einem geringeren Preis erwerben.

Mit einer dynamischen Preissetzungsstrategie kann gegenüber einer statischen Preissetzung der Gewinn des Verkäufers ausgeweitet werden: Dazu müsste es dem Verkäufer gelingen, die Preise so zu setzen, dass sie idealerweise genau die Zahlungsbereitschaft des potenziellen Käufers treffen. Kundinnen und Kunden mit hoher Zahlungsbereitschaft würden einen höheren Preis zahlen (Preis 2 in ↘ Grafik 3) entsprechend ihrer Zahlungsbereitschaft und die Konsumentenrente tendiert gegen null. Kundinnen und Kunden mit einer Zahlungsbereitschaft unter dem statischen Preis würde ein entsprechend niedrigerer Preis (Preis 3) angeboten. Sie würden zum niedrigeren Preis kaufen und nicht auf ein Substitut ausweichen. Der Absatz würde steigen, der Gewinn des Verkäufers optimiert und die Konsumentenrente abgeschöpft.

Grafik 3
Dynamische Preisgestaltung

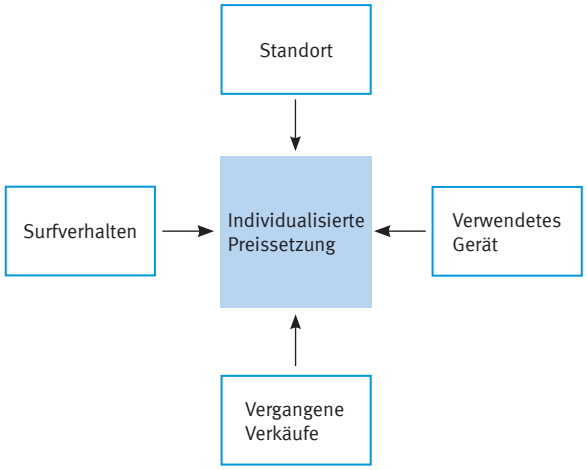


2019 - 01 - 0523

Im Onlinehandel ermöglichen die verfügbaren Algorithmen sehr schnelle Preisanpassungen. Ein Wechsel zwischen verschiedenen Produktpreisen (Preis 1, Preis 2 und Preis 3) ist ohne großen Aufwand durchzuführen und jederzeit wiederholbar, wodurch ein Produzent seine Absatzstrategie stets den Marktgegebenheiten anpassen kann.

Das Phänomen der dynamischen Preissetzung ist zu trennen vom Phänomen der individualisierten Preissetzung. Bei den Algorithmen für die individualisierte Preissetzung sind personenabhängige Eigenschaften die entscheidenden Kriterien, die die Preissetzung beeinflussen. ↘ Grafik 4

Grafik 4
Auswahl an Einflussfaktoren für individualisierte Preissetzung



2019 - 01 - 0524

Sowohl das Phänomen der dynamischen als auch das der individualisierten Preissetzung im Internet stellt eine Herausforderung für die amtliche Preisstatistik dar. Stark volatile Preise machen eine repräsentative Preismessung schwieriger. Während die individualisierte Preissetzung konzeptionell bei der Inflationsmessung noch

7 Als Konsumentenrente bezeichnet man in der Wohlfahrtstheorie die Differenz zwischen der individuellen Zahlungsbereitschaft einer Käuferin oder eines Käufers für ein Produkt und dem tatsächlich gezahlten Preis beim Kauf [siehe beispielsweise Varian (2010), hier: Seite 264].

nicht berücksichtigt werden soll, ist eine dynamische Preissetzung in der durchschnittlichen Preisentwicklung korrekt zu erfassen. Deshalb wird die Preisvolatilität bei den für die Verbraucherpreisstatistik relevanten Produkten kontinuierlich untersucht. Derzeit wird in einer Studie mit rund 9 000 Produkten das Preissetzungsverhalten von etwa 250 Onlinehändlern beobachtet und analysiert. Dazu werden stündlich Preise erhoben und ausgewertet. Ziel der Studie ist es, die Preisvolatilität der einzelnen Produkte festzustellen und bei Bedarf die Preiserhebung zu verbessern, um die Repräsentativität und die Qualität der monatlichen Preiserhebung des VPI/HVPI zu sichern oder sogar zu steigern.

4

Die Zukunft des Web Scrapings in der Preisstatistik: das generische Programm

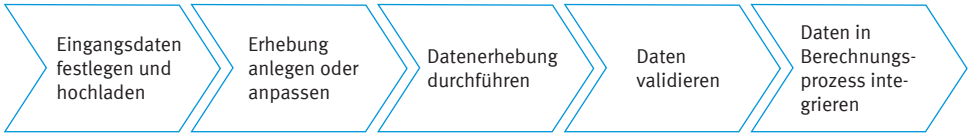
Es ist vorgesehen, mit Web Scraping sukzessive die bisherigen manuellen Preiserhebungen im Internet abzulösen. Damit können in kurzer Zeit größere Datenmengen erhoben werden und die Datenerhebungen werden flexibler. So können beispielsweise Erhebungen an Wochenenden oder Nacherhebungen ohne großen Aufwand durchgeführt werden. Allerdings kann die Wartungsintensität hoch ausfallen, wenn eine Internetseite häufig verändert wird (technische Wartung) oder die zu erhebenden Produkte häufig wechseln und ersetzt werden müssen (fachliche Wartung).

Die fachliche Wartung erfordert unter anderem regelmäßige Kontrollen der Input- und Outputdatenbanken, da das Produktsortiment im Onlinehandel durch häufige Wechsel geprägt ist. Hierzu ist geschultes Personal in den Fachreferaten erforderlich, welches die automatisiert erhobenen Preisdaten auf deren Plausibilität hin überprüft und bei auftretenden Schwierigkeiten die notwendigen Gegenmaßnahmen einleitet (zum Beispiel Produkte manuell nacherheben). Die qualitativen Anforderungen an die Mitarbeiterinnen und Mitarbeiter steigen dadurch. Die Preiserhebungen im Internet, die bisher überwiegend einmal monatlich für jedes Produkt durchgeführt werden, erfolgen künftig automatisiert und in vielen Fällen in einer sehr viel höheren Frequenz. Die bestehende Stichprobe der etwa 10 000 Produkte im Onlinehandel muss dabei nicht ausgeweitet werden, sondern entsprechend der Volatilität der Preise beziehungsweise der Preissetzungsstrategien der Onlinehändler werden mehr Preise je Produkt erfasst. Plausibilisierungsschritte, die bislang im Zuge der Preiserhebung erledigt wurden, müssen nachgelagert und sehr zeitnah im Erhebungsprozess für einen deutlich größeren Datensatz erfolgen.

Um die technische Wartung zu vereinfachen, wird seit Mitte des Jahres 2018 ein generisches Programm für Web Scraping in der Preisstatistik entwickelt. Mit dessen Hilfe sollen der Wartungsaufwand und die notwendigen IT-Kenntnisse für Web Scraping deutlich reduziert werden. Ziel ist, dass Web Scraping ohne tiefergehende IT-Kenntnisse von vielen Mitarbeiterinnen und Mitarbeitern eingesetzt und gesteuert werden kann. Ein technisches Grundverständnis über den Aufbau von Internetseiten oder den Umgang mit Datenbanken wird dennoch auch weiterhin unabdingbar sein. Spezielle Schulungen werden die Einführung des generischen Programms begleiten.

Das generische Programm wird als nutzerfreundliche Anwendung konzipiert: Die Benutzerinnen und Benutzer müssen keine Codes programmieren, sondern lediglich Abläufe anlegen, Variablen definieren und HTML-Positionen der Informationen auf Internetseiten in entsprechenden Masken einfügen. Eingangsdaten, die zur Identifizierung des zu erhebenden Produkts notwendig sind, können außerhalb der Anwendung gepflegt und aktualisiert werden und für die Erhebung in die Anwendung geladen werden. Die Anwendung wird darüber hinaus in der Lage sein, aus den definierten Variablen automatisch zusätzliche Variablen für die weitere Verwendung im Programm anzulegen. Die Ausgabedaten, also die Ergebnisse der automatischen Preiserhebungen, werden in einer Datenbank abgelegt und können über einen Browser evaluiert werden. Der gesamte Erhebungsprozess einschließlich der Navigation im Internet wird von einem herkömmlichen Internetbrowser übernommen, um Ladezeiten und Skriptfehler zu minimieren.

Grafik 5
Prozess der Datenerhebung mithilfe des generischen Programms



2019 - 01 - 0525

5

Prozess der Datenerhebung mithilfe des generischen Programms

Mit dem Einsatz des generischen Programms muss der Prozess der Datenerhebung im Vergleich zum bisherigen Vorgehen erweitert werden:

Zu Beginn sind die wesentlichen Informationen zum Produkt, wie Name, Artikelnummer und exakte Internetadresse (URL), in einer Tabelle zusammenzufassen. Hinzu kommen wichtige Informationen für die spätere Weiterverwendung, wie die (COICOP[8]-)Produktnummer, die Berichtsstellennummer und die Produktvariantennummer. Diese auch bisher schon erforderlichen Eingangsdaten werden in die Anwendung importiert und für die weitere Verwendung als Variablen angelegt. Daraufhin kann die Benutzerin oder der Benutzer den Ablauf der Preiserhebung definieren. So besteht die Möglichkeit, das Programm direkt zur Produktseite navigieren zu lassen oder schrittweise durch mehrere Seiten zu steuern (beispielsweise über die Auswahl mehrerer Filter). Schleifen und Bedingungen können ebenfalls angelegt werden. Mit Schleifen und Bedingungen kann die Datenerhebung präziser ablaufen und die Qualität der Ergebnisse gesteigert werden. In einem Zeitplan kann die Benutzerin oder der Benutzer festlegen, ob die Erhebung stündlich, täglich, wöchentlich oder jährlich wiederholt werden soll. Die Tage und genauen Uhrzeiten können ebenfalls eingestellt werden.

Unmittelbar nach einer Datenerhebung oder spätestens zum Ende eines gesamten Berichtszeitraums werden die erhobenen Daten von Preiserheberinnen und Preiserhebern validiert. Dieser Prozess ist in vier Schritte unterteilt, die im nachfolgenden Kapitel näher beschrieben werden. Im Anschluss an die Validierung werden die plausiblen Preise entweder direkt in das Berechnungsprogramm der Verbraucherpreisstatistik übernommen oder es müssen Preisindizes berechnet werden, welche als sogenannte Messzahlen in den weiteren Produktionsprozess gegeben werden. Damit wäre dann ein Erhebungsprozess abgeschlossen und kann wieder von vorne beginnen. ↘ Grafik 5

6

Neue Validierungsprozesse in der Verbraucherpreisstatistik

Die Verwendung großer Datenmengen in der monatlichen Statistikproduktion setzt einen umfassenden und effizienten Validierungsprozess voraus. Insbesondere bei der Preiserhebung im Internet können unplausible Ergebnisse das Gesamtergebnis verfälschen. Aus diesem Grund müssen dem aktuellen Validierungsprozess in der Verbraucherpreisstatistik, der auf eine manuelle Preiserhebung ausgerichtet ist, weitere Validierungsprozesse vorgeschaltet werden.

6.1 Validierung durch das Scrapingprogramm

Die Validierung beginnt schon während der automatisierten Preiserhebung. Das generische Programm wird in der Lage sein, bestimmte Charakteristika eines Produkts zu prüfen, bevor der dazugehörige Preis erhoben wird. Solche Charakteristika können Artikelnummern der zu erhebenden Produkte oder – zum Beispiel beim Personentransport – die Verbindungsmerkmale sein. Dieses Vorgehen erhöht die Präzision der automatisierten Preiserhebung beträchtlich und reduziert die ungeeigneten Daten in der Ausgangsdatenbank.

8 COICOP: Classification of Individual Consumption of Purpose – Klassifikation des individuellen Verbrauchs nach Verwendungszwecken.

6.2 Technische Validierung

Nach Abschluss der Preiserhebung erzeugt das generische Programm automatisch eine Datei mit einer Zusammenfassung des gerade abgeschlossenen Erhebungslaufs, in der Kennzahlen (zum Beispiel die Laufzeit der Erhebung, die Anzahl der erhobenen Preise, erkannte Fehler oder Warnungen) zusammengestellt werden. Darüber hinaus werden für zuvor definierte Fehlerarten Screenshots abgelegt. Diese Informationen ermöglichen es der oder dem für den Ablauf der Erhebung verantwortlichen Mitarbeiterin oder Mitarbeiter, die technische Wartung durchzuführen. Bei Bedarf kann ein vollständiger Ersatzlauf gestartet werden, um entstandene Lücken der Erhebung zu schließen.

6.3 Fachliche Datenvalidierung

Nach einem aus technischer Sicht erfolgreichen Erhebungslauf kann die Datenvalidierung durch Produktexpertinnen und Produktexperten beginnen. Vorübergehend nicht verfügbare Produkte sind im stationären und im Onlinehandel normal. Um im Nachgang der Erhebung beurteilen zu können, ob ein Produkt nur vorübergehend nicht verfügbar war oder dauerhaft aus dem Sortiment genommen wurde, werden Informationen zur Verfügbarkeit des Produkts in der Ausgangsdatenbank abgelegt. Sollte ein Produkt mehr als zwei Monate nicht verfügbar sein, wird das Produkt durch die zuständige Produktexpertin oder den zuständigen Produktexperten ersetzt. Auch bei fehlerhaft vorgegebenen Positionen der zu erhebenden Informationen auf den Produktseiten oder gänzlich fehlerhaften URLs sind die Datenbanken entsprechend zu pflegen und für die nächsten Erhebungen vorzubereiten. Die erhobenen Preisdaten werden anschließend auf Ausreißer und nicht plausible Werte hin überprüft. Nicht plausible Werte sind manuell aus dem Datensatz zu entfernen. Bei großem Datenverlust – hervorgerufen beispielsweise durch technische Schwierigkeiten – muss an dieser Stelle notfalls imputiert werden, sofern eine Wiederholung des Erhebungslaufs nicht möglich ist.

6.4 Weiterverwendung von Web-Scraping-Preisindizes

Die monatliche Datenaufbereitung und Datenvalidierung der Verbrauchpreisstatistik insgesamt wird einheitlich in einem gemeinsamen Berechnungsprogramm der Länder und des Bundes durchgeführt. Dieses sogenannte Verbundprogramm[9] ist auf eine traditionelle Preiserhebung ausgerichtet, daher akzeptiert es für die monatliche Indexberechnung nur einen einzelnen Preis oder eine Messzahl für das jeweilige Produkt. Messzahlen sind außerhalb des Erhebungsprogramms zu berechnen und anschließend im Verbundprogramm zu erfassen.

Aktuell werden in der Preisstatistik neue Methoden getestet, wie aus großen Datenmengen [die zum Beispiel mittels Web Scraping oder auch mithilfe anderer moderner Datenquellen wie Scannerdaten (Bieg, 2019) gewonnen werden] verwertbare Preisindizes auf Produktebene für die monatliche Indexproduktion berechnet werden können.

Im Folgenden werden einige Mittelwerte erläutert und erste Ergebnisse dieser Untersuchungen vorgestellt.

7

Methodenentwicklung

Die automatisierten Preiserhebungen ermöglichen es, fast unendlich viele Preise aus dem Internet zu erfassen. Wie bereits erwähnt, ist dies insbesondere in solchen Produktbereichen beziehungsweise bei solchen Onlineshops zwingend nötig, bei denen die Preise sehr volatil sind. Der Wandel von der traditionellen Indexberechnung mit monatlich einem Preis je Produkt hin zu mehreren Preisen je Produkt und Monat muss mit statistischen

Übersicht 1
Methoden zur automatisierten Preiserhebung

Methode	In Produktion	Experimentell
Modalwert	X	
Ungewichteter Durchschnitt	X	
Gewichteter Durchschnitt	X	
COICOP-Index		X

9 Für zusätzliche Erläuterungen siehe Burg/Seeger (2009).

Methoden begleitet werden. Drei einfache Methoden zur Durchschnittsberechnung werden bereits in der monatlichen Indexproduktion der Verbraucherpreisstatistik eingesetzt, eine Methode zur Aggregation großer Datenmengen wird auf ihre Anwendbarkeit geprüft. ↘ Übersicht 1

7.1 Modalwert

Als Modalwert bezeichnet man den Wert, der am häufigsten in der Erhebung vorkommt. Diese Methode wird derzeit für Preiserhebungen verwendet, die im wöchentlichen Turnus durchgeführt werden. Bei monatlichen Statistiken werden also maximal vier Werte herangezogen, bei Quartalsstatistiken maximal zwölf Werte. Durch die Verwendung des Modalwerts einer Preiserhebung fließt ein tatsächlicher Preis in die Indexberechnung ein, die anderen Preise eines Produkts dienen der Plausibilisierung. Die Verwendung dieser Methode setzt voraus, dass die Preise nicht extrem volatil sind.

7.2 Ungewichteter Durchschnitt

Ein ungewichteter Durchschnittswert $\bar{x}$ ist geeignet, wenn die Käufe eines beobachteten Produkts im Zeitverlauf gleich verteilt sind. Diese Methode wird verwendet, wenn weder der Wochentag noch die Tageszeit des Verkaufs gesondert gewichtet werden müssen. Es wird angenommen, dass sich die Preise nicht datumsabhängig verändern.

$$\bar{x} = \sum_{i=1}^{n} x_i \tag{1}$$

7.3 Gewichteter Durchschnitt

Die Verwendung eines gewichteten Durchschnitts $\bar{x}_w$ ist geeignet, wenn zu den erhobenen Preisen sehr unterschiedlich häufig gekauft, aber gleichmäßig erhoben wurde. Beispielsweise ist anzunehmen, dass Reisende schnelle Zugverbindungen bevorzugen, dass Tickets in der 2. Klasse häufiger gekauft werden als Tickets in der 1. Klasse (zumindest zu bestimmten Tagen oder Zeiten) oder dass bestimmte Verbindungen (zum Beispiel Frankfurt – Köln) gefragter sind als andere. In der Berechnung des Durchschnittspreises würde demnach eine Gewichtung w_i für mehr Repräsentativität sorgen.

$$\bar{x}_w = \sum_{i=1}^{n} w_i x_i \tag{2}$$

7.4 COICOP-Index

Neben der Anwendung dieser klassischen statistischen Methoden werden in der Preisstatistik weitere Ansätze zur Integration von großen Datenmengen untersucht. Ein Ansatz beispielsweise besteht darin, bei volatilem Preissetzungsverhalten dennoch einen repräsentativen Produktpreis für den Betrachtungszeitraum zu ermitteln, der möglichst alle (relevanten) Preisveränderungen erfasst. Mehrere statistische Ämter in der EU forschen an neuen Methoden zur Verwendung von großen Datenmengen im VPI/HVPI (Nygaard, 2015; van Loon/Roels, 2018; Griffioen/ten Bosch, 2016). Aktuell werden diese Methoden in der deutschen Preisstatistik evaluiert und weiterentwickelt beziehungsweise angepasst.

Das niederländische Statistikamt verwendet beispielsweise einen sogenannten lowest COICOP-Index (Griffioen/ten Bosch, 2016). Dort wurde festgestellt, dass sich dieser besonders für Produktgruppen mit vielen Produktvarianten und gleichzeitig häufigen Produktwechseln eignet (zum Beispiel bei Bekleidung).

Ein analoger Ansatz wird in der deutschen Preisstatistik getestet: Der COICOP-Index wird aus den arithmetischen Mittelwerten der online erhobenen Preise für eine Produktgruppe berechnet, gewichtet mit den Anteilen eines Onlinehändlers für eine Produktgruppe. Als Produktgruppe wird die niedrigste COICOP-Produktklassifikation (in Deutschland die COICOP-10-Steller-Ebene) angesehen, weil dafür eine Gewichtung vorliegt.

Der COICOP-Index für die Produktgruppe g des Onlinehändlers r im Monat m wird auf folgende Weise berechnet:

$$CI_{G,m}^{r} = \sum_{g \in G} w_{g,m}^{r} I_{g,m}^{r} \tag{3}$$

Die Gewichtung einer Berichtsstelle wird durch w^r ausgedrückt und ergibt sich aus dem Anteil der Produkte eines Onlineshops an der Stichprobe des VPI/HVPI.

Die Stichprobe der für den VPI/HVPI relevanten Onlinehändler wird bisher im Zuge der fünfjährlichen Revision anhand der Umsatzzahlen der Onlinehändler angepasst. Dazu werden auch Daten aus der Marktforschung verwendet. Die Gewichtung nach Onlinehändlern ist in solchen Produktbereichen sinnvoll, in denen bei vielen Onlinehändlern Preise erhoben werden und auch viele Produktvarianten erfasst werden (zum Beispiel Bekleidung). Außerdem ist sie sinnvoll bei Onlinehändlern, die zwar ein breites Angebot anbieten, aber nicht in allen Produktbereichen gleich führend sind (zum Beispiel Amazon).

Der Elementarindex $I^r_{g,m}$ stellt einen Elementarindex einer Produktgruppe *g* beim Onlinehändler *r* dar, beispielsweise Herrenjeans oder Damenfreizeitschuhe.[10] Dieser Elementarindex wird mit Preisdaten berechnet, die mit der Hilfe von Web Scraping erhoben wurden.

$$(4) \qquad I^r_{g,m} = \frac{P^r_{g,m}}{P^r_{g,m-1}}$$

Der Preis $P^r_{g,m}$ einer Produktgruppe *g* ergibt sich aus dem arithmetischen Mittel aller Preise, welche im Monat *m* beim Onlinehändler *r* erhoben wurden. Die Anzahl der verwendeten Preise ist besonders hoch, weil die Preise stündlich (*h*) erhoben werden. Es erfolgt keine Gewichtung nach Tageszeit oder Wochentag.[11]

$$(5) \qquad P^r_{g,m} = \frac{1}{n^r_{g,m}} \sum_{h \in m} \sum_{a \in g} p^r_{a,m,h}$$

Der Faktor *p* repräsentiert den Preis eines Produkts *a*, der auf der Internetseite eines Onlinehändlers *r* stündlich erhoben wurde und zur jeweiligen Produktgruppe *g* gehört. Der Buchstabe *n* gibt die Anzahl an Preisen an, die im Monat *m* auf der Internetseite gefunden wurden.

Die Verwendung des COICOP-Index hat den Vorteil, dass alle erhobenen Preisdaten verwendet werden können und mit entsprechenden Gewichten der Berichtsstellen nach Produktgruppe ein repräsentativer Markt abgebildet werden kann. Allerdings werden hier alle erhobenen Preise implizit gleich gewichtet, unabhängig davon, wie oft beziehungsweise ob überhaupt zum jeweiligen Preis verkauft wurde.

Die Elementarindizes $I^{web\ scraping}_{g,m}$ der jeweiligen Produktgruppe, welche aus den Web-Scraping-Daten errechnet werden, können mit den korrespondierenden Elementarindizes $I^{CPI}_{g,m}$ verglichen werden. Diese werden monatlich gemeinsam von den Statistischen Ämtern der Länder und dem Statistischen Bundesamt im bisherigen Verfahren im Onlinehandel ermittelt und fließen in die laufende Produktion des VPI/HVPI ein. Gegenübergestellt werden also Indizes berechnet aus mit Web Scraping gewonnenen Massendaten und Indizes aus der regulären Berechnung des VPI/HVPI.

$$(6) \qquad I^{web\ scraping}_{g,m} = \frac{P_{g,m}}{P_{g,m-1}}$$

$$(7) \qquad I^{CPI}_{g,m} = \sum_{s=1}^{16} w_s \frac{P_{g,m}}{P_{g,m-1}}$$

Der Elementarindex $I^{CPI}_{g,m}$ entspricht dem gewichteten arithmetischen Mittel der Elementarindizes der einzelnen Bundesländer *s* für die jeweilige Produktgruppe *g*. Diese Elementarindizes werden mit Preisdaten berechnet, die von den Statistischen Ämtern der Länder und vom Statistischen Bundesamt im Onlinehandel erhoben werden und in die laufende Produktion des VPI/HVPI einfließen. Die Elementarindizes einer Produktgruppe liegen für die einzelnen Länder vor und werden entsprechend der Größe des Bundeslands gewichtet – w_s – und zu einem Ergebnis für Deutschland aggregiert. Die regulären (gesamtdeutschen) Elementarindizes werden mit den Elementarindizes verglichen, die sich aus der stündlichen (automatisierten) Preiserhebung ergeben.

Die vorläufigen Ergebnisse der experimentellen Elementarindizes seit Erhebungsbeginn im Dezember 2018 zeigen teilweise einen sehr unterschiedlichen Verlauf gegenüber den regulären Daten auf. Aus den bisher betrachteten Produktgruppen lässt sich jedoch schließen, dass bei häufigeren Preiserhebungen von Bekleidung im Onlinehandel die Entwicklung der Elementarindizes volatiler verlaufen kann als bei den regulären Daten. Betrachtet man beispielhaft die Elementarindizes aus den Studiendaten der Produktgruppen Herrenstoffhose und Herrenjeans, weisen diese zwar einen ähn-

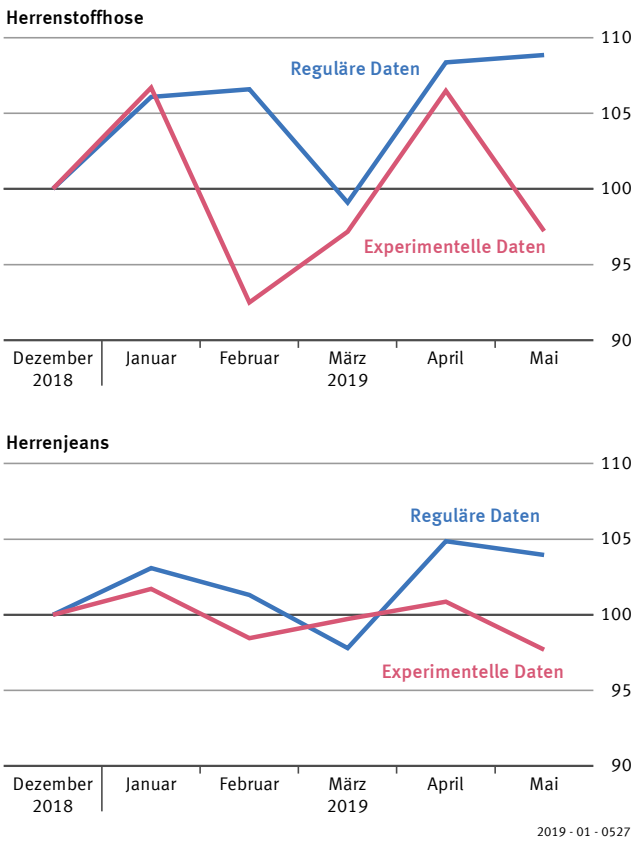
10 Ein Elementarindex ergibt sich für jede Produktgruppe als Veränderungsrate der Durchschnittspreise der darin erhobenen Produkte zwischen aktuellem Monat und Vormonat.

11 An diesem Ansatz wird allerdings ebenfalls intensiv geforscht. In früheren Studien wurde beispielsweise bereits festgestellt, dass die Anzahl der Preisänderungen auch abhängt vom Wochentag und der Tageszeit. Darüber hinaus ist anzunehmen, dass Konsumenten und Konsumentinnen ihre Einkäufe im Onlinehandel weniger häufig nachts tätigen.

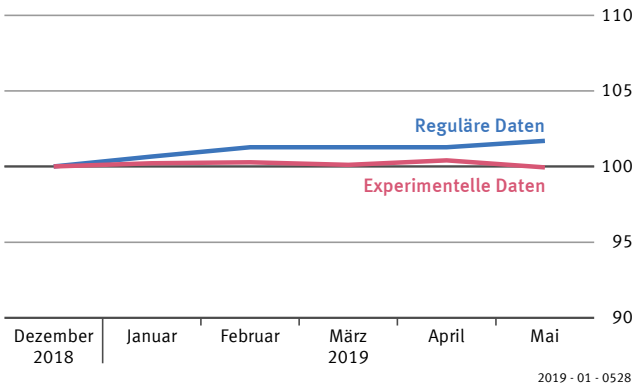
lichen Verlauf, aber stärkere Preisschwankungen auf. Insbesondere beim Vergleich der Elementarindizes der Herrenstoffhose fällt auf, dass der Elementarindex nach regulären Daten dem Elementarindex nach experimentellen Daten zeitverzögert nachfolgt. Eine Erklärung für den zeitverzögerten Verlauf könnte sein, dass die experimentellen Daten bereits Preisänderungen enthalten, die die reguläre Erhebung aufgrund ihrer Ausgestaltung aktuell erst im Folgemonat berücksichtigen kann. Besonders auffällig ist dies in den Monaten Februar und März 2019: Bei der Herrenstoffhose handelt es sich um ein saisonales Produkt, bei dem zumeist in den Monaten Februar und März der Übergang von der Winter- zur Sommerkollektion vollzogen wird. Die Preiserhebungen sind in diesen beiden Monaten durch Sonderangebote geprägt. Der massive Abverkauf der Winterware und die Einführung der Sommerartikel zum regulären Preis werden durch die häufigeren Preiserhebungen zeitlich nach vorne verlagert. Im weiteren Verlauf der aktuellen Studien muss geprüft werden, ob es sich hierbei um eine noch zu korrigierende Verzerrung oder tatsächlich um eine bessere zeitliche Abdeckung handelt. Auch bei den Herrenjeans ist diese Verschiebung zu erkennen, wenn auch weit weniger ausgeprägt. ↘ Grafik 6

Bei Produkten aus der Produktgruppe „Möbel, Leuchten, Geräte und anderes Haushaltszubehör“ (hier beispielhaft Pflegemittel) machen häufige Preiserhebungen offensichtlich kaum einen Unterschied für die Indexberechnung gegenüber der traditionellen Preiserhebung. ↘ Grafik 7

Grafik 6
Vorläufige Ergebnisse der experimentellen Elementarindizes für die Beispiele Herrenstoffhose und Herrenjeans
Dezember 2018 = 100



Grafik 7
Vorläufige Ergebnisse der experimentellen Elementarindizes für das Beispiel Pflegemittel
Dezember 2018 = 100



Erweiterungen des COICOP-Index hinsichtlich umfassenderer Preisindizes im Onlinehandel sind vorstellbar. Die Analyse experimenteller Indexformeln und die Prüfung, ob diese für die deutsche Verbraucherpreisstatistik anwendbar sind, sind Bestandteil der laufenden Studie und werden in den folgenden Jahren weiter vorangetrieben.

8

Fazit und Ausblick

Den Herausforderungen durch die gestiegene Bedeutung des Internethandels begegnet die amtliche Preisstatistik auf zwei Arten:

1. Manuelle Preiserhebungen im Internet für die laufende Indexproduktion werden sukzessive automatisiert. Somit können diese Preiserhebungen künftig mit einer optimierten Anzahl an Wiederholungen durchgeführt werden. Auch die Stichprobe im Onlinehandel kann durch Web Scraping in kurzer Zeit ausgeweitet werden.
2. Derzeit werden neue Validierungsprozesse entwickelt und Berechnungsmethoden getestet, mit dem Ziel, auch sehr volatile Preise mittels großer Datenmengen besser abzubilden und in die monatliche Indexproduktion zu integrieren.

Die Einführung des generischen Programms in die Produktion des VPI/HVPI ist sukzessive ab der zweiten Jahreshälfte 2019 geplant. Mit dessen Hilfe wird die manuelle Preiserhebung im Internet schrittweise automatisiert beziehungsweise bestehende Automatisierungsprozesse in den generischen Standard überführt. Parallel werden die Preiserheberinnen und Preiserheber in der Anwendung des Programms geschult, um ihre Preiserhebungen eigenständig zu automatisieren und anschließend zu pflegen. Das Ziel ist, bis Ende des Jahres 2020 die vormals manuellen Preiserhebungen im Internet für 10 000 Produkte vollständig zu automatisieren. Zunächst werden dabei die bisherigen manuellen Erhebungen imitiert und bei volatilen Preisen die Erhebungsfrequenz angepasst. Methodische Änderungen sind in diesem ersten Schritt nicht vorgesehen. Erforderlich sind jedoch zusätzliche Validierungsprozesse während des Erhebungsprozesses.

Darüber hinaus werden neue Methoden untersucht und entwickelt, um die Möglichkeiten des Web Scrapings möglichst weitgehend auszuschöpfen. Das Ziel ist, zur nächsten Revision der Verbraucherpreisstatistik die oben beschriebenen neuen Berechnungsmethoden, soweit sinnvoll, einzubauen. Langfristig kann das generische Programm auch für die Datenerhebung anderer Statistiken erweitert und verwendet werden.

Die flächendeckende Integration von Web Scraping in die laufende Produktion der Preisstatistik stellt einen wichtigen Schritt hin zur Digitalisierung im Statistischen Bundesamt dar. Das generische Programm wird die Onlinepreiserhebungen deutlich verbessern und damit auch die Qualität des VPI und des HVPI steigern.

LITERATURVERZEICHNIS

Blaudow, Christian. *Fortschritte und Herausforderungen beim Web Scraping – Automatisierung von Preiserhebungen im Internet*. In: Methoden – Verfahren – Entwicklungen. Ausgabe 1/2018, Seite 3.

Blaudow, Christian/Burg, Florian. *Dynamische Preissetzung als Herausforderung für die Verbraucherpreisstatistik*. In: WISTA Wirtschaft und Statistik. Ausgabe 2/2018, Seite 11 ff.

Bieg, Matthias. *Nutzung von Scannerdaten in der Preisstatistik – eine Untersuchung anhand von Marktforschungsdaten*. In: WISTA Wirtschaft und Statistik. Ausgabe 2/2019, Seite 25 ff.

Sandhop, Karsten. *Geschäftstypengewichtung im Verbraucherpreisindex*. In: Wirtschaft und Statistik. Ausgabe 3/2012, Seite 266 ff.

Burg, Florian/Seeger, Daniel. *Das neue Verbundprogramm der Verbraucherpreisstatistik*. In: Wirtschaft und Statistik. Ausgabe 2/2009, Seite 169 ff.

Griffioen, Robert/ten Bosch, Olav. *On the use of Internet data for the Dutch CPI*. 2016. [Zugriff am 12. Juni 2019]. Verfügbar unter: www.researchgate.net

Van Loon, Ken/Roels, Dorien. *Integrating big data in the Belgian CPI*. 2018. Meeting of the Group of Experts on Consumer Price Indices. Verfügbar unter: www.unece.org

Nygaard, Ragnhild. *The use of online prices in the Norwegian Consumer Price Index*. Tokio 2015.

Varian, Hal R. *Intermediate Microeconomics. A Modern Approach*. 8. Auflage. 2010.

Herausgeber
Statistisches Bundesamt (Destatis), Wiesbaden

Schriftleitung
Dr. Daniel Vorgrimler
Redaktionsleitung: Juliane Gude
Redaktion: Ellen Römer

Ihr Kontakt zu uns
www.destatis.de/kontakt

Erscheinungsfolge
zweimonatlich, erschienen im August 2019
Das Archiv älterer Ausgaben finden Sie unter www.destatis.de

Print
Einzelpreis: EUR 19,– (zzgl. Versand)
Jahresbezugspreis: EUR 114,– (zzgl. Versand)
Bestellnummer: 1010200-19004-1
ISSN 0043-6143
ISBN 978-3-8246-1083-9

Download (PDF)
Artikelnummer: 1010200-19004-4, ISSN 1619-2907

Vertriebspartner
IBRo Versandservice GmbH
Bereich Statistisches Bundesamt
Kastanienweg 1
D-18184 Roggentin
Telefon: +49 (0) 382 04 / 6 65 43
Telefax: +49 (0) 382 04 / 6 69 19
destatis@ibro.de

Papier: Design Offset, FSC-zertifiziert

© Statistisches Bundesamt (Destatis), 2019
Vervielfältigung und Verbreitung, auch auszugsweise, mit Quellenangabe gestattet.