Hauke, Oliver; Kristiansen, Annette

Article

Effiziente Auswertung des Business-Tax-Panels mithilfe der Statistiksoftware R

WISTA - Wirtschaft und Statistik

Provided in Cooperation with:
Statistisches Bundesamt (Destatis), Wiesbaden

Suggested Citation: Hauke, Oliver; Kristiansen, Annette (2026) : Effiziente Auswertung des Business-Tax-Panels mithilfe der Statistiksoftware R, WISTA - Wirtschaft und Statistik, ISSN 1619-2907, Statistisches Bundesamt (Destatis), Wiesbaden, Vol. 78, Iss. 1, pp. 127-139

This Version is available at:
https://hdl.handle.net/10419/337208

EFFIZIENTE AUSWERTUNG DES BUSINESS-TAX-PANELS MITHILFE DER STATISTIKSOFTWARE R

Oliver Hauke, Annette Kristiansen

↘ **Schlüsselwörter:** effiziente Programmierung – Datenverarbeitung – Apache Parquet – Steuerstatistiken – Forschungsdatenzentrum

ZUSAMMENFASSUNG

Das Statistische Bundesamt steht ständig vor der Herausforderung, deutlich größere Datenvolumen zu verarbeiten; zugleich steigen die Erwartungen der Öffentlichkeit an die Aktualität und Granularität der amtlichen Statistiken. Aktuelle technologische Weiterentwicklungen bieten das Potenzial, diese Herausforderungen zu bewältigen. So können geeignete Werkzeuge Auswertungen, die zuvor mehrere Stunden gedauert haben, in wenigen Sekunden ausführen. Der Artikel evaluiert am Beispiel des Business-Tax-Panels die Performanz und Handhabung entsprechender Werkzeuge in der Statistiksoftware R mit dem Ziel, Handlungsempfehlungen für die Praxis abzuleiten.

↘ ***Keywords:*** *efficient programming – data processing – Apache Parquet – tax statistics – research data centre*

ABSTRACT

The Federal Statistical Office of Germany faces the constant challenge of having to process significantly increasing data volumes, while public expectations regarding the granularity and timeliness of official statistics continue to increase. Recent technological developments hold the potential to meet this challenge. Analyses that used to take several hours can be conducted within seconds using suitable tools. Taking the business tax panel as an example, the article assesses the performance and user-friendliness of tools provided in the statistical software R, with the aim of deriving recommendations for their use in practice.

Oliver Hauke

ist Mathematiker und Referent im Referat „IT-Kompetenzzentrum Auswertung und Analyse“ des Statistischen Bundesamtes. Er verantwortet die Weiterentwicklung der technischen Infrastruktur des Forschungsdatenzentrums des Statistischen Bundesamtes und berät das Netzwerk empirische Steuerforschung in der effizienten Nutzung der Systeme.

Annette Kristiansen

ist Ökonomin und Referentin im Referat „Unternehmenssteuern“ des Statistischen Bundesamtes. Sie beschäftigt sich mit der Zusammenführung der Unternehmenssteuerstatistiken und betreut Arbeiten für das Netzwerk empirische Steuerforschung.

1

Einleitung

Die zunehmende Digitalisierung hat Auswirkungen auf viele Lebensbereiche und somit auf die Gesellschaft. Sie beeinflusst und verändert beispielsweise Kommunikation, Arbeitswelt und Bildung. Als ein Produkt der Digitalisierung entstehen große Datenmengen, die die amtliche Statistik vor die Herausforderung stellen, diese deutlich größeren Datenmengen zu verarbeiten. Zugleich nehmen die Erwartungen der Öffentlichkeit an die Aktualität, Genauigkeit und Gliederungstiefe von amtlichen Statistiken zu. Technologische Weiterentwicklungen bieten das Potenzial, diese Herausforderungen zu bewältigen. Geeignete Werkzeuge können Auswertungen, die bisher Stunden dauerten, in wenigen Sekunden ausführen. Der Artikel evaluiert die Leistung solcher Werkzeuge in der Statistiksoftware R sowie deren Handhabung durch Nutzende von amtlichen Mikrodaten.

Die Analysebedarfe der Wissenschaft in den Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder (FDZ)[1] verdeutlichen die Herausforderungen besonders. Die On-Site-Zugangswege bieten formal anonymisierte Mikrodaten der amtlichen Statistik in großem Umfang für wissenschaftliche Auswertungen an. Diese umfassen Gastwissenschaftsarbeitsplätze (GWAP), die Forschende direkt in den Liegenschaften der Forschungsdatenzentren nutzen können, oder kontrollierte Datenfernverarbeitung (KDFV), bei der entwickelte Auswertungsprogramme durch die FDZ-Mitarbeitenden ausgeführt werden. Die Infrastruktur der Forschungsdatenzentren muss in der Lage sein, umfangreiche Forschungsprojekte auf Basis großer Datenmengen effizient zu unterstützen.

Auch aus der Politik und der evidenzbasierten Politikberatung haben sich weitere Anforderungen ergeben. So fordert der Wissenschaftliche Beirat beim Bundesministerium der Finanzen (2020) eine Ausweitung des Datenbestands für die empirische Steuerforschung, während der Sachverständigenrat zur Begutachtung der gesamtwirtschaftlichen Entwicklung (2023) die Forschungsdateninfrastruktur zudem als rückständig bezeichnete.

Seit seiner Gründung im Jahr 2022 ist die Verbesserung der Datenbereitstellungsinfrastruktur für die empirische Steuerforschung ein Fokusthema des Netzwerks empirische Steuerforschung (NeSt). Insgesamt vernetzt das Netzwerk Kompetenzen aus der Finanzverwaltung, Steuerforschung und amtlichen Statistik mit dem Ziel, die evidenzbasierte Steuerforschung zu stärken. Es hat die Erweiterung des Datenbestands an Steuerstatistiken seither maßgeblich vorangetrieben. Seit 2024 steht das Business-Tax-Panel (BTP) mit fast 3 000 Merkmalen und 67 Millionen Beobachtungen über die Forschungsdatenzentren zur Verfügung (Kapitel 2) – ein enormes Analysepotenzial, das zugleich hohe Anforderungen an die Analysesysteme und deren Nutzende stellt.

Um diese Herausforderungen zu bewältigen, wurde ein interner R-Server des Statistischen Bundesamtes im Januar 2025 um erhebliche Rechenkapazitäten erweitert. Vorangetrieben durch das Netzwerk empirische Steuerforschung stehen auch der Wissenschaft seit November 2025 im FDZ Bund die On-Site-Server mit erhöhten Rechenkapazitäten für Analysen in R und Stata exklusiv für GWAP- und KDFV-Nutzungen zur Verfügung (Kapitel 3) – ein bedeutender Fortschritt für die Infrastruktur des FDZ Bund.[2]

Die freie Statistiksoftware R hat sich zu einem Standardwerkzeug für Statistik und Datenanalyse entwickelt, das zunehmend in der amtlichen Statistik sowie in der Wissenschaft genutzt wird. Eine weltweite Gemeinschaft erweitert die Open-Source-Software R laufend um neue Funktionalitäten in Form von R-Paketen – insbesondere für effiziente Datenverarbeitung (Kapitel 4). Dieser Artikel bewertet die Eignung von R-Paketen für die Verarbeitung großer Datenmengen im Statistischen Bundesamt und FDZ Bund (Kapitel 5) und demonstriert, wie mit R relevante Auswertungen des Business-Tax-Panels sekundenschnell durchgeführt werden können (Kapitel 6). Das Ergebnis zeigt, dass geeignete Technologien die Effizienz erheblich steigern, zugleich jedoch weiterer Handlungsbedarf besteht (Kapitel 7).

1 FDZ Bund steht im Folgenden für das Forschungsdatenzentrum des Statistischen Bundesamtes.

2 Demonstriert in einem Vortrag anlässlich eines NeSt-Webinars des Bundesministeriums der Finanzen am 28. März 2025: www.bundesfinanzministerium.de

2

Anwendungsfall Business-Tax-Panel

2.1 Datengrundlage

Das Business-Tax-Panel kombiniert die Angaben aus verschiedenen Ertrags- sowie Umsatzsteuerstatistiken im Quer- und Längsschnitt. Verfügbar sind Informationen aus den folgenden Statistiken: Gewerbesteuerstatistik, Körperschaftsteuerstatistik, Umsatzsteuer-Voranmeldung und Veranlagung, Statistik über die Personengesellschaften und Gemeinschaften, Einnahmenüberschussrechnung sowie einige Merkmale aus dem statistischen Unternehmensregister. Insgesamt deckt der Forschungsdatensatz eine Vielzahl an Analysemöglichkeiten ab, wie die Analyse von Anpassungsreaktionen oder Verteilungsanalysen, die für evidenzbasierte steuerpolitische Entscheidungen nötig sind (Kristiansen, 2023).

Für die vielfältigen Analysewünsche der Forschenden, die damit einhergehende Dokumentation des Datensatzes in den FDZ-Metadatenreports (FDZ Bund, 2024) und den Auswertungsbedarf des Fachbereichs im Statistischen Bundesamt ist der Datensatz möglichst wenig beschränkt worden und dadurch ressourcenintensiv. Anstelle einer Stichprobe wurden die zugrunde liegenden Vollerhebungen verwendet. Die erste Version des Business-Tax-Panels (2013 bis 2019) besteht aus knapp 67 Millionen Beobachtungen und liegt in komprimierter Form mit insgesamt 314 Gigabyte[3] vor. Mit der Aktualisierung um das Berichtsjahr 2020 wird sich die Zahl der Beobachtungen auf etwa 78 Millionen erhöhen. Ebenso steigt der Merkmalsumfang von fast 3 000 Merkmalen durch weitere Berichtsjahre. Bereits jetzt – und insbesondere durch die anstehende Aktualisierung – stehen der Fachbereich im Statistischen Bundesamt sowie das FDZ Bund vor der Herausforderung, den hohen Bedarf an Festplattenspeicher, Arbeitsspeicher (sogenanntes Random-Access-Memory, kurz RAM) und Prozessorleistung zu decken.

2.2 Auswertungsbedarfe

Neben den standardisierten Veröffentlichungsprodukten (wie Tabellen in der Datenbank GENESIS-Online, Statistische Berichte, Pressemitteilungen) der zugrunde liegenden Statistiken erstellt das Statistische Bundesamt zusätzlich Sonderauswertungen. Diesen liegen ebenso wie den Standardauswertungen meist spezifische Anforderungen zugrunde:

Ein häufiger Anwendungsfall ist die deskriptive Auswertung in der Statistikproduktion. Dafür sind verschiedene Bearbeitungsschritte in Form von Gruppierung der Beobachtungen, Auswahl von Merkmalen und Bearbeitung der Daten nötig. Auch für den FDZ-Metadatenreport werden deskriptive Auswertungen benötigt. Daher wird als Anwendungsfall die Tabelle zur Häufigkeitsverteilung der Beobachtungen je Jahr und Steuerstatistik (FDZ Bund, 2024) herangezogen. ↘ Tabelle 1 auf Seite 130

Ein weiterer Anwendungsfall simuliert eine exemplarische Datenanalyse, die den klassischen Bedarf der Wissenschaft repräsentiert. Die verschiedenen Datenbearbeitungsschritte und Analysen zeigen ein potenzielles Forschungsprojekt und sind entsprechend vereinfacht. Konkret wird das Vorliegen eines Verlustvortrages[4] im jeweiligen Berichtsjahr anhand verschiedener Einflussfaktoren untersucht. Die Körperschaftsteuerstatistik zeigt über die Jahre eine Zunahme der Zahl der Steuerpflichtigen mit Verlustvortrag im jeweiligen Berichtsjahr (GENESIS-Tabellen-Code 73211-0002). Anzunehmen ist, dass Personalkosten den Verlustvortrag erhöhen können, daher dienen die Zahl der sozialversicherungspflichtig Beschäftigten sowie die Summe der Löhne und Gehälter als erklärende Variable. Der jeweilige Gesamtbetrag der Einkünfte und die abgezogenen Spenden können ebenfalls den Verlustvortrag beeinflussen. Um den Internationalisierungsgrad des Unternehmens zu erfassen, wird ein Dummy aufgenommen. Die Ausgabe der Regressionsergebnisse ist hier nicht von Relevanz und wird daher nicht erzeugt.

Die beiden beschriebenen Anwendungsfälle (Fallzahl und Regression) werden für den Performanzvergleich genutzt (siehe Kapitel 5).

3 Der Datenumfang ist in komprimierter SAS-Datei gemessen. Dies variiert für andere Dateitypen (siehe Kapitel 5).

4 Verluste aus einem Veranlagungsjahr können mit Gewinnen in späteren Veranlagungsjahren verrechnet werden.

Tabelle 1
Häufigkeitsverteilung der Beobachtungen je Jahr und Steuerstatistik

	Gewerbesteuer	Körperschaft-steuer	Statistik der Personengesell-schaften und Gemeinschaften	Umsatzsteuer-Voranmeldung	Umsatzsteuer-Veranlagung	Statistisches Unternehmens-register	Einnahmen-überschuss-rechnung
2013	3 633 451	1 210 838	1 205 110	3 243 538	6 435 280	3 474 431	3 703 982
2014	3 714 176	1 236 129	1 220 521	3 240 221	6 446 620	3 550 775	3 720 318
2015	3 819 938	1 264 717	1 234 537	3 255 537	6 535 948	3 634 643	4 230 871
2016	3 887 556	1 293 431	1 247 699	3 266 429	6 550 960	3 674 354	4 581 982
2017	3 981 853	1 316 520	1 220 067	3 266 806	6 664 535	3 708 949	5 685 661
2018	4 050 674	1 350 837	1 284.121	3 279 136	6 831 036	3 736 550	6 310 028
2019	4 099 690	1 388 679	1 294 898	3.288 306	6 972 252	3 723 169	6 501 292

Anmerkung: Das statistische Unternehmensregister ist nicht komplett im Business-Tax-Panel enthalten. Nur Einheiten, die in den Steuerstatistiken enthalten sind, werden um Angaben des statistischen Unternehmensregisters ergänzt.

3

Ausbau der Analyseserver

Mit der Bereitstellung neuer Analyseserver im Jahr 2025 wurden entscheidende Fortschritte beim Ausbau der Infrastruktur des Statistischen Bundesamtes und des FDZ Bund erzielt. Die Einführung der neuen Systeme ist als Ergebnis langjähriger Planung und Umsetzung ein bedeutender Schritt.

Änderungen von Softwareversionen sind in einer abgeschotteten Infrastruktur wie im Statistischen Bundesamt organisatorisch aufwendig. Daher standen für die folgenden Darstellungen nicht immer die neuesten Softwareversionen zur Verfügung – dies spiegelt die realistischen Bedingungen der Statistikproduktion wider. Die neue Infrastruktur wurde daher von Beginn an zukunftsorientiert konzipiert und bietet moderne Software auf deutlich erweiterter Hardwarebasis.

3.1 Ausgangslage

Das Hauptwerkzeug der Statistikproduktion des Statistischen Bundesamtes ist die kommerzielle Analysesoftware SAS 9.4, die auf dedizierten Servern bereitgestellt wird.[5] SAS verarbeitet Daten überwiegend zeilenorientiert und belastet den Arbeitsspeicher nur gering – ein entscheidender Vorteil in der amtlichen Statistik, da große Datenmengen mit begrenzter Hardware verarbeitet werden können. Für viele Anwendungsfälle ist jedoch die spaltenbasierte Datenverarbeitung effizienter. In R sind Verfahren verfügbar, die moderne Prozessor- und Arbeitsspeicher gezielt ausnutzen und damit Laufzeiten gegenüber SAS erheblich verkürzen (siehe die Kapitel 4 und 5).

Im FDZ Bund ist die Statistiksoftware Stata das am häufigsten genutzte Werkzeug. Viele wissenschaftliche Fachrichtungen nutzen es, da weniger tiefgreifende Kenntnisse nötig sind und spezialisierte Verfahren, insbesondere für ökonometrische Analysen, bereitstehen. Bei großen Datenmengen stößt Stata jedoch an Grenzen, da Daten vollständig in den Arbeitsspeicher geladen werden müssen und in Stata nur wenige Operationen auf mehrere Prozessorkerne parallelisiert durchgeführt werden. Nur durch manuelle Variablenauswahl lassen sich größere Datensätze bearbeiten. Bisher waren die Gastwissenschaftsarbeitsplätze des FDZ Bund auf die Single-Core-Version von Stata mit begrenztem Arbeitsspeicher beschränkt, wodurch viele FDZ-Produkte ohne Datenzuschnitt[6] an den Gastwissenschaftsarbeitsplätzen nicht nutzbar waren. Leistungsfähigere Server standen vor der Bereitstellung der On-Site-Server lediglich in der kontrollierten Datenfernverarbeitung zur Verfügung.

5 Da SAS bereits langjährig eingesetzt wird, ist für die kommenden Performanzvergleiche zu beachten, dass die SAS-Server nicht die neueste Hardware nutzen. Im Jahr 2025 war SAS auf Version 9.4M8 verfügbar.

6 Ein Datenzuschnitt ist eine Teilmenge der Datengrundlage, die aus projektspezifischen Merkmalen und zufällig ausgewählten Beobachtungen besteht.

3.2 Anforderungen an die Analyseinfrastruktur

Damit eine moderne Infrastruktur den technischen Anforderungen des Statistischen Bundesamtes und des FDZ Bund gerecht wird, muss sie ein breites Spektrum an Datenverarbeitungswerkzeugen mit hoher Performanz leicht handhabbar bereitstellen.

Die eingesetzte Infrastruktur muss skalierbar die Verarbeitung wachsender Datenmengen zuverlässig mit den vorhandenen Rechenkapazitäten ermöglichen, damit steigende Datenverarbeitungsbedarfe gedeckt werden. Um eine hohe Reaktionsfähigkeit und eine reibungslose GWAP- beziehungsweise KDFV-Nutzung im FDZ Bund sicherzustellen, sollten die Systeme kurze Rechenzeiten ermöglichen – und idealerweise Auswertungen großer Datenmengen in Echtzeit (das heißt innerhalb weniger Sekunden) ausführen können.

Ein wesentlicher Erfolgsfaktor der Systeme ist eine intuitive Handhabung. Moderne Interfaces und nachvollziehbare Programmiersyntax müssen sicherstellen, dass die Infrastruktur für verschiedene Fachrichtungen leicht zugänglich ist. So kann die simple Nutzung den Gesamtaufwand der Datenauswertung reduzieren, da neben der Rechenzeit auch die Entwicklungszeit berücksichtigt werden muss (Gomolka und andere, 2021).

Ebenso entscheidend sind die Bereitstellungs- und Wartungsaufwände der Systeme innerhalb der abgeschotteten Systemarchitektur des Statistischen Bundesamtes für einen stabilen Einsatz in der Statistikproduktion und die Verfügbarkeit im FDZ Bund.

3.3 Erweiterter R- und Python-Server

Die Statistiksoftware R bietet hohe Anwenderorientierung und fokussiert sich auf Datenverarbeitung und -auswertung. Als Open-Source-Software erfreut sich R einer großen internationalen Gemeinschaft, die R um mehr als 20 000 R-Pakete erweitert.[7] Passend zu den Anforderungen aus Abschnitt 3.2 sind darunter viele R-Pakete, die die Performanz und/oder Handhabung von R verbessern. Auch die amtliche Statistik hat die Vorzüge von R erkannt und die Software erfreut sich nun steigender Beliebtheit in den Statistikämtern.[8] Diese Entwicklungen sind ebenfalls im Statistischen Bundesamt zu beobachten. So steht bereits seit 2020 ein Server für R und Python zur Verfügung, der für die Datenverarbeitung in der Statistikproduktion sowie für die kontrollierte Datenfernverarbeitung im FDZ Bund genutzt wurde.

Im Januar 2025 wurde der interne R- und Python-Server, der exklusiv für die Mitarbeiterinnen und Mitarbeiter des Statistischen Bundesamtes zur Verfügung steht, erheblich erweitert. Dabei wird die kommerziell unterstützte Posit Workbench verwendet, die kollaboratives Arbeiten und den Betrieb der Systeme vereinfacht. So werden die aktuellsten R-Versionen und R-Pakete in einer modernen Oberfläche bereitgestellt. Zusätzlich wurden die Systemkapazitäten deutlich ausgebaut, sodass Grafikkarten ganz neue Anwendungsfälle ermöglichen. ↘ Tabelle 2

Tabelle 2
Ausbau der R-Server des Statistischen Bundesamtes im Januar 2025

Komponente	Vorher	Nachher
Anzahl Server	1	10
Arbeitsspeicher	700 Gigabyte	je 2 Terabyte
Kernanzahl	72	je 96
Kerntaktfrequenz	je bis zu 4,0 Gigahertz	je bis zu 3,4 Gigahertz
Grafikkarte	keine	je zwei Nvidia A6000 (mit je 48 Gigabyte VRAM)

Mehr als 200 Mitarbeiterinnen und Mitarbeiter setzen den R- und Python-Server inzwischen regelmäßig ein.

3.4 On-Site-Server des FDZ Bund

Neben dem internen Server sind seit November 2025 die sogenannten On-Site-Server im FDZ Bund einsatzfähig. Sie bieten der Wissenschaft dedizierte Stata- und R-Server mit umfangreichen Systemka-

7 Am 9. Dezember 2025 waren 23 035 Pakete auf CRAN, dem umfassenden R-Archivnetzwerk, verfügbar.

8 Dies belegt beispielsweise die Entwicklung der Teilnahmezahlen der Konferenz „The Use of R in Official Statistics“: r-project.ro/conference2025.html

pazitäten für die exklusive Nutzung an den Gastwissenschaftsarbeitsplätzen sowie der kontrollierten Datenfernverarbeitung. Alle Analyseserver sind über schnelle Direktverbindungen an einen zentralen Netzwerkspeicher angebunden, sodass R und Stata kombiniert werden können. ↘ Tabelle 3

Tabelle 3
On-Site-Server des Forschungsdatenzentrums des Statistischen Bundesamtes: Spezifikationen

Komponente	Anzahl/Größe
Stata-Server	3
R-Server	6
Arbeitsspeicher	je 512 Gigabyte
Kernanzahl	je 16
Kerntaktfrequenz	je 3,1 Gigahertz

Stand: November 2025

Stata ist in der effizienten Version „MP“ einheitlich auf leistungsstarken Servern verfügbar. Zudem steht seit 2025 ein R-Server für GWAP-Nutzungen im FDZ Bund bereit. Der Systemaufbau orientiert sich am System aus Abschnitt 3.3. Beide R-Server unterstützen moderne, speicherschonende Werkzeuge, mit denen auch Datenmengen oberhalb des Arbeitsspeichers effizient verarbeitet werden können.

4 Datenverarbeitungswerkzeuge

Im Folgenden wird die Performanz von Datenverarbeitungswerkzeugen, das heißt Softwarelösungen, die Daten einlesen, modifizieren, aggregieren, auswerten oder analysieren,[9] betrachtet.

Ein Datenverarbeitungswerkzeug wird als hochperformant bezeichnet, wenn es Datenmengen über dem verfügbaren Arbeitsspeicher verarbeiten kann und dabei eine flexible und effiziente Nutzung der Ressourcen ermöglicht. Denn für die Skalierung auf größere Datenmengen ist ein Wechsel von der vollständigen Verarbeitung im Arbeitsspeicher zur flexiblen Auslagerung von Daten nötig. Im Arbeitsspeicher ist zwar eine schnelle Datenverarbeitung möglich, doch das Einlesen großer Datenmengen kann einen Großteil der Gesamtlaufzeit ausmachen, und der verfügbare Arbeitsspeicher ist meist begrenzt. Geeignete Software und moderne Hardware ermöglichen heute sehr schnelle Verarbeitungen von Datenmengen oberhalb des Arbeitsspeichers.

4.1 Klassische Werkzeuge in R

Die Kernfunktionen von R (sogenannte Base R) bieten grundlegende Werkzeuge, die für Einsteiger oder in stark abgeschotteten Infrastrukturen zweckdienlich sind. Base R lässt sich in Bezug auf Handhabung und Performanz durch eine Vielzahl von R-Paketen verbessern, wie die folgenden R-Pakete zeigen:

> data.table (Barrett und andere, 2024) gilt als „Goldstandard“ für die schnelle Datenverarbeitung im Arbeitsspeicher sowie das effiziente Einlesen von CSV-Dateien. Dank der Anlehnung an die Base-R-Syntax ist es sehr beliebt. Solange die Daten vollständig in den Arbeitsspeicher passen, ist data.table in den meisten Fällen die schnellste klassische Lösung. Dabei ist einschränkend zu beachten, dass eine CSV-Datei im Arbeitsspeicher ein Vielfaches (etwa das Drei- bis Vierfache) ihrer Größe beansprucht.

> tidyverse (Wickham und andere, 2016) ist ein Ökosystem von R-Paketen, das besonders auf einfache und nachvollziehbare Programmierung setzt. Die eigene dplyr-Syntax ist auch für neue Nutzende gut verständlich. ↘ Tabelle 4 Diese intuitive Syntax wird von vielen anderen R-Paketen adaptiert. Die exemplarische Syntax gibt, nach ergänzenden Transformationen, die in Tabelle 1 dargestellten Fallzahlen wider.

4.2 Hochperformante Werkzeuge in R

Für Datenmengen, die einen Großteil des Arbeitsspeichers belegen, ist die nachfolgend beschriebene Variablenselektion eine oft notwendige Best Practice, die direkt spürbare Performanzgewinne

9 Das Schreiben von Daten wird nicht vertieft behandelt, da die betrachteten Ergebnistabellen in der Regel klein sind. Die genannten Lösungen sind dennoch im Schreiben von Daten ebenfalls effizient. Aufgrund externer Abhängigkeiten werden Datenbanksysteme sowie verteilte Rechenlösungen nicht berücksichtigt. Der Fokus liegt auf den Standard-R-Nutzenden, die mit den üblichen R-Werkzeugen vertraut sind.

Tabelle 4
Exemplarische Auswertung des Business-Tax-Panels in dplyr-Syntax

Nr.	Befehl	Beschreibung
1	ergebnis <-	Speichert die Fallzahl je Jahr und Statistik des Business-Tax-Panels.
2	read_csv('daten/btp/daten.csv') |>	Einlesen des Business-Tax-Panels im CSV-Format.
3	group_by(jahr) |>	Gruppiert die Daten nach Jahren.
4	summarize(	Aggregiert die Daten auf eine Beobachtung je Gruppe.
5	g = sum(str_detect('g', verk)),	Zählt in jeder Gruppe die Fälle mit 'g' in verk.
6	k = sum(str_detect('k', verk)),	Zählt in jeder Gruppe die Fälle mit 'k' in verk.
7	p = sum(str_detect(‚p', verk)),	Zählt in jeder Gruppe die Fälle mit ‚p' in verk
8	u = sum(str_detect(‚u', verk)),	Zählt in jeder Gruppe die Fälle mit ‚u' in verk .
9	v = sum(str_detect(‚v', verk)),	Zählt in jeder Gruppe die Fälle mit ‚v' in verk.
10	r = sum(str_detect(‚r', verk)),	Zählt in jeder Gruppe die Fälle mit ‚r' in verk.
11	e = sum(str_detect(‚e', verk)),	Zählt in jeder Gruppe die Fälle mit ‚e' in verk .
12	)	

bringt. Besonders für klassische Werkzeuge kann das Risiko für Arbeitsspeicher-Abbrüche und Wartezeiten deutlich reduziert werden. Zusätzlich ist die Datenaufteilung kombiniert mit sogenannter Lazy-Computation für große Datenmengen oder komplexe Analysen vorteilhaft.

> Variablenselektion: Die Einlesefunktion sollte bei größeren Datenmengen stets auf notwendige Variablen begrenzt werden, um Ladezeit und Arbeitsspeicher-Bedarf massiv zu reduzieren. Das Einlesen zunächst aller Daten (gegebenenfalls mit späterer Variablenselektion) verursacht unnötige Wartezeiten, Ressourcenbelegung und gegebenenfalls Abbrüche. Auch Anwendungsfälle in Stata können von einer Variablenselektion profitieren – zum Beispiel mit einer Beschleunigung um 33,9 % in Anwendungen der Forschungsdatenzentren (Gomolka und andere, 2021).

> Datenaufteilung: Wenn der Arbeitsspeicher nicht ausreicht, um die benötigten Daten einzulesen, kann eine Datenaufteilung auf mehrere Arbeitsspeicher-gerechte Dateien helfen. Hochperformante Werkzeuge können auch direkt mehrere Dateien nutzen, ohne dass sie manuell zusammengeführt werden müssen.

> Lazy-Computation: Ohne Anpassungen in der Programmierung kann die Lazy-Computation geplante Datenverarbeitungsschritte speichern, ohne sie sofort auszuführen. Erst bei expliziter Abfrage eines Ergebnisses führt das System die Berechnung durch. Dadurch werden nur tatsächlich benötigte Daten verarbeitet und unnötige Schritte vermieden – eine manuelle Variablenauswahl ist nicht erforderlich.

Apache Parquet ist ein spaltenbasiertes, komprimiertes Open-Source-Datenformat, das speziell für effiziente Speicherung und Verarbeitung großer Datenmengen entwickelt wurde. Im Folgenden wird das Format vereinfacht als Parquet bezeichnet. Durch eine spaltenbasierte Struktur ermöglicht Parquet den schnellen Zugriff auf einzelne Merkmale und reduziert gleichzeitig den Speicherbedarf erheblich. Gomolka und andere (2021) zeigten im Kontext der Forschungsdatenzentren massive Geschwindigkeitsgewinne anhand Parquet, sodass im Forschungsdaten- und Servicezentrum der Deutschen Bundesbank bereits standardmäßig Parquet-Dateien angeboten werden.

Für die hochperformante Verarbeitung von Datenmengen oberhalb des Arbeitsspeichers werden die folgenden Werkzeuge betrachtet, die in Form von R-Paketen verfügbar sind:

> arrow (Richardson und andere, 2024) nutzt speziell für die Arbeit mit großen Datensätzen intelligente Datenaufteilung und Parallelisierung durch C++ im Hintergrund. Es verarbeitet Daten spaltenweise und verwendet Lazy-Computation, was

besonders bei aufwendigen analytischen Abfragen auf großen Datensätzen Vorteile bietet. arrow nutzt standardmäßig dplyr-Syntax (in Tabelle 4 muss lediglich „csv“ durch „parquet“ ersetzt werden). arrow ist oft die beste Wahl in R, um mehrere Parquet-Dateien zu nutzen, die insgesamt ein Vielfaches größer sind als der vorhandene Arbeitsspeicher.

- duckdb (Mühleisen/Raasveldt, 2024) ist ein eingebettetes Datenbanksystem, das ohne separaten Datenbankserver direkt in R ebenfalls effizient Daten spaltenweise verarbeitet. Nativ nutzt duckdb SQL-Syntax. Durch ergänzende R-Pakete kann auch dplyr-Syntax verwendet werden. Eine direkte Schnittstelle zu arrow ermöglicht die nahtlose Kombination von arrow und duckdb.
- polars (Shima und andere, 2024) ist ein jüngeres R-Paket, das anhand der Programmiersprache Rust für hohe Geschwindigkeit und sparsame Ressourcennutzung sorgt. Nativ nutzt es Python-orientierte Syntax und kann auch durch das Paket tidypolars minimal verlangsamt dplyr-Syntax verwenden.

4.3 Externe Benchmarks

Öffentlich verfügbare Benchmarks können zu abweichenden Ergebnissen führen.[10] Daher ist die Untersuchung in der eigenen Infrastruktur mit relevanten Anwendungsfällen (insbesondere Datenmengen) und den verfügbaren Software-Versionen wichtig. Die zum Zeitpunkt des Performanzvergleichs (Kapitel 5) betrachteten R-Paket-Versionen können im Statistischen Bundesamt aus organisatorischen Gründen nicht flexibel angepasst werden. Sie sind im Literaturverzeichnis explizit aufgeführt. Entscheidend ist bei der Versionswahl nicht das absolute Performanz-Optimum, sondern eine Version mit langfristig stabiler und verlässlich guter Performanz.

5

Performanzvergleich

Um zwischen den in Kapitel 4 vorgestellten Datenverarbeitungswerkzeugen abzuwägen, wurde deren Performanz in einem umfangreichen Vergleich auf dem SAS-Server des Statistischen Bundesamtes und den On-Site-R- und Stata-Servern des FDZ Bund exploriert. Dabei wurde schrittweise die Datenmenge erhöht und nur Werkzeuge weiterbetrachtet, die eine hohe Performanz aufzeigten. Weiterführende Performanzmessungen sowie der vollständig reproduzierbare Quellcode sind auf Opencode verfügbar.[11]

5.1 Versuchsaufbau

Für die Untersuchung wurden simulierte Einzeldaten in verschiedenen Größen (1 % sowie 10 % der insgesamt 67 Millionen Beobachtungen des Business-Tax-Panels) betrachtet. Sie wurden aus öffentlich verfügbaren Informationen generiert und bilden die wesentlichen Strukturen des Business-Tax-Panels für die Analyse künstlich nach. Abgrenzend zu den im FDZ Bund bereitgestellten Datenstrukturfiles wurden die Testdaten nicht aus echten Mikrodaten abgeleitet. Der simulierte Datensatz ist ausschließlich für technische Testzwecke der folgenden Anwendungsfälle (siehe Kapitel 2) bestimmt:

1. Fallzahlen (je Statistik und Jahr; siehe Tabelle 1),
2. Regression (Einflussfaktoren auf Verlustvorträge).

Um eine zuverlässige Performanz-Messung zu erhalten, wurde jedes Datenverarbeitungswerkzeug dreifach angewandt. Der Vergleich erfolgt anhand der mittleren Messung der Laufzeit (real verstrichene Zeit).

Daneben ist der beanspruchte Arbeitsspeicher der Auswertung entscheidend, da ein übermäßiger Arbeitsspeicherverbrauch zu Abbrüchen führen kann. Dennoch wurde der Arbeitsspeicher-Verbrauch

10 Benchmark von data.table (cran.r-project.org), duckdb (duckdblabs.github.io) und polars (ddotta.github.io).

11 gitlab.opencode.de

nur für ausgewählte Tests anhand eines Systemmonitors beobachtet.[12]

Das Ziel war, Werkzeuge mit einer Mindestperformanz in verschiedenen Szenarien zu finden und nicht in jedem Anwendungsfall ein Werkzeug für die schnellste Verarbeitung zu bestimmen. Denn der Einsatz von möglichst wenigen verschiedenen Werkzeugen reduziert Entwicklungs- und Wartungsaufwände.

Feine Abstufungen der Zielgrößen sind für die Aufgaben nicht gleichermaßen relevant. Sobald eine Mindestperformanz erreicht wurde, rückt die Reduktion des Entwicklungsaufwands in den Vordergrund. Entsprechend sind Aspekte wie einfache Anwendbarkeit und Nachvollziehbarkeit – auch für unerfahrene Nutzende – ausschlaggebend.

5.2 Vergleich von R zu SAS und Stata

Zunächst werden 1 % der simulierten BTP-Daten (etwa 700 000 Beobachtungen; 12 Gigabyte als CSV-Datei) geprüft. Um die nachfolgenden Ergebnisse einzuordnen, wurden die Performanzmessungen auch in SAS und Stata als Vergleichsmaßstab durchgeführt. ↘ Tabelle 5

SAS zeigt seine Stärke anhand der geringen Beanspruchung des Arbeitsspeichers. Mit etwa 55 Sekunden je Auswertung ist Stata aber etwa doppelt so schnell wie SAS.

Für die Datenmenge von 1 % der simulierten BTP-Daten beschleunigen alle betrachteten R-Datenverarbeitungswerkzeuge die Berechnungen um mindestens 76 % gegenüber SAS und Stata. ↘ Tabelle 6 Die schnellsten Ausführungen benötigen in R für Fallzahlen oder Regression etwa eine Sekunde.

Aus weiteren Untersuchungen für 1 % der simulierten Daten lassen sich die folgenden Erkenntnisse ableiten:

1. Das Parquet-Dateiformat ermöglicht die schnellsten Ergebnisse und spart gleichzeitig 50 % Festplattenspeicher ein.
2. Die Datenaufteilung verlangsamt für geringere Datenmengen die Berechnung der Fallzahlen. In der Berechnung der Regression profitiert arrow von einer Aufteilung.
3. Die Variablenauswahl zeigte immer eine Reduktion der Laufzeit und Arbeitsspeicher-Beanspruchung, schränkt aber die Flexibilität der Auswertung ein.
4. Unter den klassischen Werkzeugen war data.table deutlich am schnellsten[13], benötigte aber etwa die dreifache Größe der CSV-Datei im Arbeitsspeicher. data.table wird hier durch Variablenauswahl konkurrenzfähig zu den besten Methoden.
5. In der Berechnung der Fallzahl sind alle hochperformanten Werkzeuge – arrow, duckdb und polars – nahezu gleichwertig und benötigen höchstens 2 Sekunden für das Ergebnis.
6. In der Berechnung der Regression war duckdb mit nur etwa 1 Sekunde deutlich am schnellsten. Hier ist der klassische Ansatz (data.table mit Variablenselektion) sogar schneller als arrow.
7. Bei der Regression wurde polars nicht betrachtet, da es im R-Kontext weniger intuitiv zu nutzen war als die anderen Werkzeuge und die Implementierung aufgrund von Fehlern nicht durchgelaufen ist.

12 Im Rahmen des Experiments konnte der Arbeitsspeicher-Verbrauch nicht systematisch korrekt in R gemessen werden, da der Verbrauch von C++ oder Rust nicht mit R-Paketen gemessen werden kann. Dies zeigt eine Warnung des R-Pakets bench in dieser Benchmark: tidypolars.etiennebacher.com

13 Übersichtshalber wird data.table als repräsentativer Vergleichswert für klassische Werkzeuge betrachtet, da das R-Paket deutlich die beste Performanz zeigte. Der Vergleich zu anderen klassischen Werkzeugen ist zu finden auf gitlab.opencode.de

Tabelle 5
Baselines in SAS und Stata (anhand 1 % des simulierten Business-Tax-Panels)

Software	Anwendung	Arbeitsspeicher	Laufzeit	Festplattenspeicher
SAS	Fallzahlen	65 Megabyte	2,4 Minuten	33 Gigabyte
	Regression	7 Megabyte	2,8 Minuten	33 Gigabyte
Stata	Fallzahlen	33 Gigabyte	55,2 Sekunden	32 Gigabyte
	Regression	33 Gigabyte	55,8 Sekunden	32 Gigabyte

Tabelle 6
Benchmarks für die Berechnung der Fallzahlen und der Regression anhand 1 % des simulierten Business-Tax-Panels

R-Paket	Dateiformat	Anpassung	Laufzeit	Arbeitsspeicher	Festplatte
			Fallzahlen		
polars	Parquet		0,97 Sekunden	–	5 Gigabyte
duckdb	Parquet		1,68 Sekunden	–	5 Gigabyte
duckdb	Parquet	Datenaufteilung (Berichtsjahr)	1,94 Sekunden	–	4 Gigabyte
arrow	Parquet	Variablenauswahl	2,07 Sekunden	–	5 Gigabyte
data.table	CSV	Variablenauswahl	2,38 Sekunden	41 Megabyte	12 Gigabyte
arrow	Parquet		2,72 Sekunden	–	5 Gigabyte
arrow	Parquet	Datenaufteilung (gleichmäßig)	2,99 Sekunden	–	5 Gigabyte
arrow	Parquet	Datenaufteilung (Berichtsjahr)	3,42 Sekunden	–	4 Gigabyte
arrow	Parquet		4,18 Sekunden	–	5 Gigabyte
data.table	CSV		8,76 Sekunden	31 Gigabyte	12 Gigabyte
data.table	CSV	Datenaufteilung (Berichtsjahr)	13,32 Sekunden	47 Gigabyte	10 Gigabyte
			Regression		
duckdb	Parquet		1,13 Sekunden	–	5 Gigabyte
data.table	CSV	Variablenauswahl	2,90 Sekunden	552 Megabyte	12 Gigabyte
arrow	Parquet	Datenaufteilung (Berichtsjahr)	4,02 Sekunden	–	4 Gigabyte
data.table	CSV		7,51 Sekunden	32 Gigabyte	12 Gigabyte
arrow	Parquet		13,35 Sekunden	–	5 Gigabyte

Anhand der hier betrachteten Datenmenge von 700 000 Beobachtungen können die schnellsten Werkzeuge kaum differenziert werden. Daher wurde im nächsten Experiment für ausgewählte Werkzeuge die Datenmenge auf 7 000 000 Beobachtungen erhöht (etwa 10 % des Business-Tax-Panels).

5.3 Größere Datenmengen in R

Auch für 10 % der simulierten BTP-Daten können hochperformante Werkzeuge in R beide Anwendungsfälle in unter 10 Sekunden beantworten. ↘ Tabelle 7 Die Auswertungswerkzeuge mit der höchsten Performanz sind:

> arrow mit nur 4 Sekunden für die Berechnung der Fallzahl,

> duckdb mit nur 5,2 Sekunden für die Berechnung der Regression.

Für 10 % der simulierten Daten lassen sich weitere Erkenntnisse ableiten:

1. Das Parquet-Dateiformat ermöglicht die schnellsten Ergebnisse und reduziert die Datenmengen um 75 Gigabyte gegenüber den CSV-Dateien (etwa 62 %).
2. duckdb verbraucht weniger als 1 Gigabyte Arbeitsspeicher und liefert beide Ergebnisse in etwa 5 Sekunden. Eine Datenaufteilung hat das Ergebnis verschlechtert.
3. Mit der schnellsten Fallzahlberechnung zeigt arrow seine Kernkompetenz, mehrere Parquet-Dateien besonders effizient nutzen zu können. Die spezifische Aufteilung hat einen großen Einfluss auf Arbeitsspeicher-Bedarf und Laufzeit. Unerwarteterweise hat eine zufällige Aufteilung bessere Ergebnisse erzielt als die Aufteilung nach Berichtsjahren.
4. Klassische Werkzeuge kommen bei der Größe der CSV-Datei von 121 Gigabyte bereits an Kapazitätsgrenzen der Server. Somit ist bei der Anwendung klassischer Methoden deutlich mehr Erfahrung notwendig, um zwischen Variablenauswahl und Arbeitsspeicher-Verbrauch abzuwägen.
5. Die Laufzeit von 41 Sekunden von polars ist um den Faktor 10 langsamer als arrow und duckdb. Da polars das jüngste dieser R-Pakete ist, könnte

Tabelle 7
Benchmarks für die Berechnung der Fallzahlen und der Regression anhand 10 % des simulierten Business-Tax-Panels

R-Paket	Dateiformat	Anpassung	Laufzeit	Arbeitsspeicher	Festplatte
Fallzahlen					
arrow	Parquet	Datenaufteilung (gleichmäßig)	4,00 Sekunden	2,7 Gigabyte	45 Gigabyte
duckdb	Parquet		5,39 Sekunden	400 Megabyte	46 Gigabyte
arrow	Parquet	Variablenauswahl	7,45 Sekunden	–	46 Gigabyte
duckdb	Parquet	Datenaufteilung (Berichtsjahr)	7,57 Sekunden	–	45 Gigabyte
arrow	Parquet	Datenaufteilung (Berichtsjahr)	9,61 Sekunden	8,3 Gigabyte	45 Gigabyte
arrow	Parquet		14,85 Sekunden	163 Gigabyte	46 Gigabyte
data.table	CSV	Variablenauswahl	20,80 Sekunden	408 Megabyte	121 Gigabyte
polars	Parquet		41,02 Sekunden	186 Gigabyte	46 Gigabyte
data.table	CSV		53,04 Sekunden	313 Gigabyte	121 Gigabyte
data.table	CSV	Datenaufteilung (Berichtsjahr)	4 Minuten 41 Sekunden	472 Gigabyte	97 Gigabyte
Regression					
duckdb	Parquet		5,19 Sekunden	760 Megabyte	46 Gigabyte
arrow	Parquet	Datenaufteilung (Berichtsjahr)	7,21 Sekunden	8 Gigabyte	45 Gigabyte
data.table	CSV	Variablenauswahl	27,16 Sekunden	5 Gigabyte	121 Gigabyte
data.table	CSV		1 Minuten 12 Sekunden	317 Gigabyte	121 Gigabyte
arrow	Parquet		2 Minuten 59 Sekunden	336 Gigabyte	46 Gigabyte

ein Update der vorliegenden Version 1.0.1[14] dies beheben.

6

Ergebnis für das vollständige Business-Tax-Panel

Hochperformante R-Werkzeuge sind in der Lage, die betrachteten Anwendungsfälle in Echtzeit zu beantworten. Die Verarbeitung von Parquet-Dateien mit arrow und duckdb für die Auswertung großer Datenmengen wird empfohlen. Parquet-Dateien haben in allen Untersuchungen deutlich schnellere Auswertungen ermöglicht und gleichzeitig den Speicherbedarf reduziert. Arrow und duckdb können ohne Verzögerungen kombiniert werden und bieten die nachvollziehbare dplyr-Syntax. Für besonders große Datenmengen kann arrow aufgeteilte Parquet-Dateien besonders gut nutzen.

Das vollständige Business-Tax-Panel 2013 bis 2019 hat einen Umfang von 610 Gigabyte in Stata-Dateien, sodass selbst der Arbeitsspeicher der On-Site-Stata-Server mit 512 Gigabyte nicht ausreicht, um das Business-Tax-Panel vollständig in Stata einzulesen. Dadurch muss in Stata zwingend mit Variablenselektion gearbeitet werden. Für SAS liegt das Business-Tax-Panel in einzelnen komprimierten Jahresscheiben vor und belegt insgesamt 314 Gigabyte Festplattenspeicher. SAS kann die betrachteten Anwendungsfälle für diese Datenmengen weiterhin durchführen und erreicht die in ↘ Tabelle 8 dargestellte Performanz.

14 Das R-Package polars in der verwendeten Version 1.0.1 war nur der erste Minor Release nach einer kompletten Überarbeitung. Inzwischen sind fünf neuere Major Releases mit diversen Performanz-Verbesserungen verfügbar.

Tabelle 8
Business-Tax-Panel 2013 bis 2019 in SAS als komprimierte Jahresscheiben

Software	Anwendung	Arbeitsspeicher	Laufzeit	Festplattenspeicher
SAS	Fallzahlen	1 Gigabyte	93 Minuten	314 Gigabyte
SAS	Regression	141 Megabyte	92 Minuten	314 Gigabyte

Tabelle 9
Business-Tax-Panel 2013 bis 2019 in R mit arrow und Berichtsjahren in Parquet-Dateien

Software	Anwendung	Arbeitsspeicher	Laufzeit	Festplattenspeicher
R	Fallzahlen	etwa 7 Gigabyte	5,5 Sekunden	17,5 Gigabyte
R	Regression	etwa 10 Gigabyte	11,8 Sekunden	17,5 Gigabyte

Anhand einer Aufteilung von Parquet-Dateien je Berichtsjahr zeigt arrow für das gesamte Business-Tax-Panel erst das volle Potenzial von R-Werkzeugen. ↘ Tabelle 9.

Gegenüber SAS reduziert R die Laufzeit von über 90 Minuten auf etwa 12 Sekunden mit dennoch geringem Arbeitsspeicher-Bedarf (unter 10 Gigabyte). Die Parquet-Dateien beanspruchen nur noch 17,5 Gigabyte Speicher. Hier lässt sich beobachten, dass arrow und Parquet-Dateien für große Datenmengen höhere Effizienz bieten. Obwohl gegenüber Abschnitt 5.3 die Datenmenge um 90 % erhöht wurde, steigt die Laufzeit der Berechnung der Fallzahlen nur um eine Sekunde (38 %) und der Regression um 4,6 Sekunden (64 %).

7

Fazit

Eine Datenhaltung im Parquet-Format für Datenverarbeitungen beziehungsweise -auswertungen in der Statistikproduktion und in den Forschungsdatenzentren ist gewinnbringend. Es reduziert den notwendigen Festplattenspeicher stark und mehrere R-Pakete (arrow, duckdb und polars) können Daten großen Umfangs im Parquet-Format sehr schnell verarbeiten. Daher wird empfohlen, eine Datenhaltung im Parquet-Format und eine Datenverarbeitung in R für große Datensätze in der amtlichen Statistik und in den Forschungsdatenzentren zu erproben. Denn die Untersuchung demonstriert, dass dadurch relevante analytische Fragestellungen anhand großer Datenmengen in Echtzeit ressourcenschonend beantwortet werden können.

Das R-Paket arrow (kombiniert mit duckdb) zeigte die intuitivste Handhabung, sodass die Entwicklungsaufwände am geringsten waren. Das neuere R-Paket polars[15] bietet eine vielversprechende Alternative, die sich bei kleineren Datenmengen als sehr schnell erwiesen hat. In der vorliegenden Version zeigte es jedoch bei größeren Datenmengen keine durchgängig stabile Performanz, und die Umsetzung in R war vergleichsweise weniger intuitiv. Dadurch wird deutlich, dass es für einen optimalen Einsatz der Werkzeuge entscheidend ist, die Ergebnisse des Statistischen Bundesamtes mit der Open-Source-Gemeinschaft zu teilen, sodass die Anwendungen reproduziert und verbessert werden können.

In der Statistikproduktion und den Forschungsprojekten der Forschungsdatenzentren bieten die vorgestellten Werkzeuge das Potenzial, Effizienz erheblich zu steigern sowie die Nachvollziehbarkeit der Programmierung zu erhöhen. Dafür ist der Aufbau eines Datenbestands in Parquet notwendig und die schrittweise Ausweitung auf weitere Statistiken. Begleitend sind Schulungsangebote für die Nutzenden und Mitarbeitenden der Forschungsdatenzentren und der amtlichen Statistik wesentlich für den Erfolg der Umsetzung. Nutzende der effizienten Werkzeuge können dann sekundenschnelle Auswertungen großer Datensätze im Vollmaterial durchführen, wodurch sich Analysen in der empirischen Forschung mithilfe von R sowohl beschleunigen als auch ausweiten lassen.

Die Untersuchung hat gezeigt, dass R als zentrales Analysewerkzeug im FDZ Bund und im Statistischen Bundesamt eingesetzt werden kann. Da R vielseitig hohe Performanz und intuitive Handhabung bietet, kann ein Fokus auf R Hardwarekosten senken und die Softwarevielfalt verschlanken. Die Analyse großer Datensätze in der amtlichen Statistik und den Forschungsdatenzentren kann von einer massiven Performanz-Verbesserung profitieren, wenn die vorgestellten Werkzeuge von Infrastrukturbetreibern, amtlicher Statistik, Forschungsdatenzentren und Forschenden adaptiert werden.

15 Siehe Fußnote 14.

LITERATURVERZEICHNIS

Barrett, Tyson/Dowle, Matt/Srinivasan, Arun/Gorecki, Jan/Chirico, Michael/Hocking, Toby/Schwendinger, Benjamin/Krylov, Ivan. *data.table: Extension of 'data.frame'* (Version 1.17.8). 2024. [Zugriff am 16. Januar 2026]. Verfügbar unter: CRAN.R-project.org

Forschungsdatenzentrum der Statistischen Ämter des Bundes und der Länder. *Metadatenreport. Teil II. Produktspezifische Informationen zum Business-Tax-Panel 2013–2019 für die On-Site-Nutzung. Version I*. Wiesbaden 2024. [Zugriff am 16. Januar 2026]. Verfügbar unter: www.forschungsdatenzentrum.de

Gomolka, Matthias/Blaschke, Jannick/Hirsch, Christian. *Working with Large Data at the RDSC*. In: Deutsche Bundesbank (Herausgeber). Technical Report 2021-04. [Zugriff am 16. Januar 2026]. Verfügbar unter: www.bundesbank.de

Kristiansen, Annette. *Business-Tax-Panel – Zusammenführung von Unternehmenssteuerstatistiken*. In: WISTA Wirtschaft und Statistik. Ausgabe 4/2023, Seite 47 ff.

Mühleisen, Hannes/Raasveldt, Mark. *duckdb: DBI Package for the Duckdb Database Management System (Version 1.3.2)*. 2024. [Zugriff am 16. Januar 2026]. Verfügbar unter: CRAN.R-project.org

Richardson, Neal/Cook, Ian/Crane, Nic/Dunnington, Dewey/François, Romain/Keane, Jonathan/Mecum, Bryce und andere. *Arrow: Integration to 'Apache' 'Arrow' (Version 20.0.0.2)*. 2024. [Zugriff am 16. Januar 2026]. Verfügbar unter: CRAN.R-project.org

R Core Team. *R: A language and environment for statistical computing (Version 4.5.0)*. R Foundation for Statistical Computing, Vienna, Austria. 2021. [Zugriff am 16. Januar 2026]. Verfügbar unter: www.R-project.org

Sachverständigenrat zur Begutachtung der gesamtwirtschaftlichen Entwicklung. *Jahresgutachten 2023/24: Wachstumsschwäche überwinden – In die Zukunft investieren*. 2023. [Zugriff am 16. Januar 2026]. Verfügbar unter: www.sachverstaendigenrat-wirtschaft.de

Shima, Tatsuya/Bacher, Etienne und andere. *polars: R Bindings for the 'polars' Rust Library (1.0.1)*. 2024. [Zugriff am 16. Januar 2026]. Verfügbar unter: pola-rs.github.io

Wickham, Hadley/Averick, Mara/Bryan, Jennifer/Chang, Winston/ D'Agostino McGowan, Lucy/François, Romain/Grolemund, Garrett und andere. *Welcome to the tidyverse (Version 2.0.0)*. In: Journal of Open Source Software. Ausgabe 4 (43): 1686. 2019. DOI: 10.21105/joss.01686.

Wissenschaftlicher Beirat beim Bundesministerium der Finanzen. *Notwendigkeit, Potenzial und Ansatzpunkte einer Verbesserung der Dateninfrastruktur für die Steuerpolitik*. 2020. Gutachten. [Zugriff am 16. Januar 2026]. Verfügbar unter: www.bundesfinanzministerium.de

Herausgeber
Statistisches Bundesamt (Destatis), Wiesbaden

Schriftleitung
Dr. Daniel Vorgrimler
Redaktion: Ellen Römer

Ihr Kontakt zu uns
www.destatis.de/kontakt

Erscheinungsfolge
zweimonatlich, erschienen im Februar 2026
Ältere Ausgaben finden Sie unter www.destatis.de sowie in der Statistischen Bibliothek.

Artikelnummer: 1010200-26004-4, ISSN 1619-2907