Corpus der Entscheidungen des Bundesfinanzhofs

(CE-BFH)

Codebook

Version 2025-01-14

DOI: 10.5281/zenodo.14622341

Titel	Corpus der Entscheidungen des Bundesfinanzhofs
Abkürzung	CE-BFH
Autor	Seán Fobbe
Version	2025-01-14
Download	https://doi.org/10.5281/zenodo.14622341
Lizenz	CC0 1.0 Universal

Zitiervorschlag

Seán Fobbe (2025). Corpus der Entscheidungen des Bundesfinanzhofs (CE-BFH). Version 2025-01-14. Zenodo. DOI: 10.5281/zenodo.14622341.

Digital Object Identifier (DOI): Concept DOI und Version DOI

Soweit nicht anders angegeben ist die DOI immer eine »Version DOI« und bezieht sich nur auf eine bestimmte Version des Datensatzes. Sie verweist daher nur auf Version 2025-01-14. Für das Gesamtkonzept dieses Datensatzes steht eine »Concept DOI« zur Verfügung, die auf der Zenodo-Seite jeder Version unter »Cite all versions?« zu finden ist. Sie lautet 10.5281/zenodo.7691840. Die »Concept DOI« verlinkt immer die aktuellste Version.

Urheberrecht

Disclaimer

Dieser Datensatz ist eine private wissenschaftliche Initiative und steht in keiner Verbindung zu Behörden, Gerichten oder anderen amtlichen Stellen der Bundesrepublik Deutschland.

Inhaltsverzeichnis

1 Einführung

Der **Bundesfinanzhof (BFH)** ist einer der fünf obersten Gerichtshöfe des Bundes und steht an der Spitze der Finanzgerichtsbarkeit der Bundesrepublik Deutschland (Art. 95 Abs. 1 GG, §§ 2, 10 f. FGO). Der BFH ist die höchste Instanz in Steuer- und Zollsachen und entscheidet über Revisionen gegen Urteile der Finanzgerichte, gegen urteilsgleiche Entscheidungen und Beschwerden gegen andere Entscheidungen der Finanzgerichte (§ 36 FGO).[1] Er wurde mit dem »Gesetz über den Bundesfinanzhof« vom 29. Juni 1950 errichtet und hat seinen Sitz in München (§ 2 FGO).

Im Jahr 2022 sind am Bundesfinanzhof 11 Senate eingerichtet, bestehend aus ca. 60 Richter:innen.[2] Zudem besteht ein Großer Senat, dem Richter:innen aller Senate angehören. Präsident:in und Vizepräsident:in vertreten das Gericht nach Außen. Entscheidungen der Senate ergehen in einer Besetzung von 5 Richter:innen, Beschlüsse außerhalb der mündlichen Verhandlung können von 3 Richter:innen getroffen werden (§ 10 Abs. 3 FGO). Der Große Senat besteht aus Gerichtspräsident:in und einer Richter:in jedes Senates, in dem der Vorsitz nicht von der Gerichtspräsident:in geführt wird (§ 11 Abs. 5 FGO), d.h. aktuell 11 Mitgliedern. Zu Beginn jeden Jahres bestimmt das Präsidium des Gerichts die thematische Geschäftsverteilung zwischen den Senaten.

Wieso dieser Datensatz? Die quantitative Analyse von juristischen Texten, insbesondere denen des BFH, ist in den deutschen Rechtswissenschaften ein noch junges und kaum bearbeitetes Feld.[3] Zu einem nicht unerheblichen Teil liegt dies auch daran, dass die Anzahl an frei nutzbaren Datensätzen außerordentlich gering ist.

Die meisten hochwertigen Datensätze lagern (fast) unerreichbar in kommerziellen Datenbanken und sind wissenschaftlich gar nicht oder nur gegen Entgelt zu nutzen. Frei verfügbare Datenbanken wie *Opinio Iuris*[4] und *openJur*[5] verbieten ausdrücklich das maschinelle Auslesen der Rohdaten. Wissenschaftliche Initiativen wie der Juristische Referenzkorpus (JuReKo) sind nach jahrelanger Arbeit hinter verschlossenen Türen verschwunden.

In einem funktionierenden Rechtsstaat muss die Rechtsprechung öffentlich, transparent und nachvollziehbar sein. Im 21. Jahrhundert bedeutet dies auch, dass sie systematischer Überprüfung mittels quantitativen Analysen zugänglich sein muss. Der Erstellung und Aufbereitung des Datensatzes liegen daher die Prinzipien der allgemeinen Verfügbarkeit durch Urheberrechtsfreiheit, strenge Transparenz und vollständige wissenschaftliche Reproduzierbarkeit zugrunde. Die FAIR-Prinzipien (Findable, Accessible, Interoperable and Reusable) für freie wissenschaftliche Daten inspirieren sowohl die Konstruktion, als auch die Art der Publikation.[6]

[1] Der Finanzrechtsweg ist vergleichsweise kurz: in erster Instanz entscheiden die Finanzgerichte (§ 35 FGO) und es steht im Anschluss nur noch die Anrufung des BFH offen (§ 36 FGO).

[2] https://www.bundesfinanzhof.de/de/gericht/organisation/.

[3] Besonders positive Ausnahmen finden sich unter: https://www.quantitative-rechtswissenschaft.de/

[4] https://opinioiuris.de/

[5] https://openjur.de/

[6] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for Scientific Data Management and Stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

2 Nutzung

Die Daten sind in offenen, interoperablen und weit verbreiteten Formaten (CSV, TXT, PDF) veröffentlicht. Sie lassen sich grundsätzlich mit allen modernen Programmiersprachen (z.B. Python oder R), sowie mit grafischen Programmen nutzen.

Wichtig: Nicht vorhandene Werte sind sowohl in den Dateinamen als auch in der CSV-Datei mit "NA" codiert.

2.1 CSV-Dateien

Am einfachsten ist es die **CSV-Dateien** einzulesen. CSV[7] ist ein einfaches und maschinell gut lesbares Tabellen-Format. In diesem Datensatz sind die Werte komma-separiert. Jede Spalte entspricht einer Variable, jede Zeile einer Entscheidung. Die Variablen sind unter Punkt 5 genauer erläutert.

Zum Einlesen empfehle ich für **R** das package **data.table** (via CRAN verfügbar). Dessen Funktion **fread()** ist etwa zehnmal so schnell wie die normale **read.csv()**-Funktion in Base-R. Sie erkennt auch den Datentyp von Variablen sicherer. Ein Beispiel:

```
library(data.table)
dt <- fread("filename.csv")
```

2.2 TXT-Dateien

Die **TXT-Dateien** inklusive Metadaten können zum Beispiel mit **R** und dem package **readtext** (via CRAN verfügbar) eingelesen werden. Ein Vorschlag:

```
library(readtext)
df <- readtext("./*.txt",
               docvarsfrom = "filenames",
               docvarnames = c("gericht",
                               "bfhe",
                               "datum",
                               "spruchkoerper_az",
                               "registerzeichen",
                               "eingangsnummer",
                               "eingangsjahr_az",
                               "zusatz_az",
                               "bfh_id",
                               "kollision"),
               dvsep = "_",
               encoding = "UTF-8")
```

[7] Das CSV-Format ist in RFC 4180 definiert, siehe https://tools.ietf.org/html/rfc4180

3 Konstruktion

3.1 Beschreibung des Datensatzes

Dieser Datensatz ist eine digitale Zusammenstellung von möglichst allen begründeten Entscheidungen, die auf der amtlichen Internetpräsenz des Bundesfinanzhofs (BFH) am jeweiligen Stichtag veröffentlicht waren. Die Stichtage für jede Version entsprechen exakt der Versionsnummer.

Zusätzlich zu den aufbereiteten maschinenlesbaren Formaten (HTML und CSV) sind die PDF-Daten enthalten, damit Analyst:innen gegebenenfalls eine unabhängige Konvertierung vornehmen können. Die PDF-Rohdaten wurden inhaltlich nicht verändert und nur die Dateinamen angepasst, um die Lesbarkeit für Mensch und Maschine zu verbessern.

Speziell an Praktiker:innen richten sich die PDF-Sammlungen aller in der amtlichen Sammlung abgedruckten Entscheidungen (V-Entscheidungen).

3.2 Datenquellen

Datenquelle	Fundstelle
Primäre Datenquelle	https://www.bundesfinanzhof.de
Source Code	https://doi.org/10.5281/zenodo.14622342
Registerzeichen	https://doi.org/10.5281/zenodo.4569564

Die Tabelle der Registerzeichen und der ihnen zugeordneten Verfahrensarten stammt aus dem folgenden Datensatz: “Seán Fobbe (2021). Aktenzeichen der Bundesrepublik Deutschland (AZ-BRD). Version 1.0.1. Zenodo. DOI: 10.5281/zenodo.4569564.”

3.3 Sammlung der Daten

Die Daten wurden unter Beachtung des Robot Exclusion Standard (RES) gesammelt. Der Abruf geschieht ausschließlich über TLS-verschlüsselte Verbindungen. Die Entscheidungen sind laut dem Gericht anonymisiert, aber ungekürzt.

3.4 Source Code und Compilation Report

Der gesamte Source Code — sowohl für die Erstellung des Datensatzes, als auch für dieses Codebook — ist öffentlich einsehbar und dauerhaft erreichbar im wissenschaftlichen Archiv des CERN unter dieser Addresse hinterlegt: https://doi.org/10.5281/zenodo.14622342

Mit jeder Kompilierung des vollständigen Datensatzes wird auch ein umfangreicher **Compilation Report** in einem attraktiv designten PDF-Format erstellt (ähnlich diesem Codebook). Der Compilation Report enthält den kommentierten Source Code für die Daten-Pipeline, dokumentiert relevante Rechenergebnisse, gibt sekundengenaue Zeitstempel an und ist mit einem klickbaren Inhaltsverzeichnis versehen. Er ist zusammen mit dem Source Code hinterlegt. Wenn Sie sich für Details der Herstellung interessieren, lesen Sie diesen bitte zuerst.

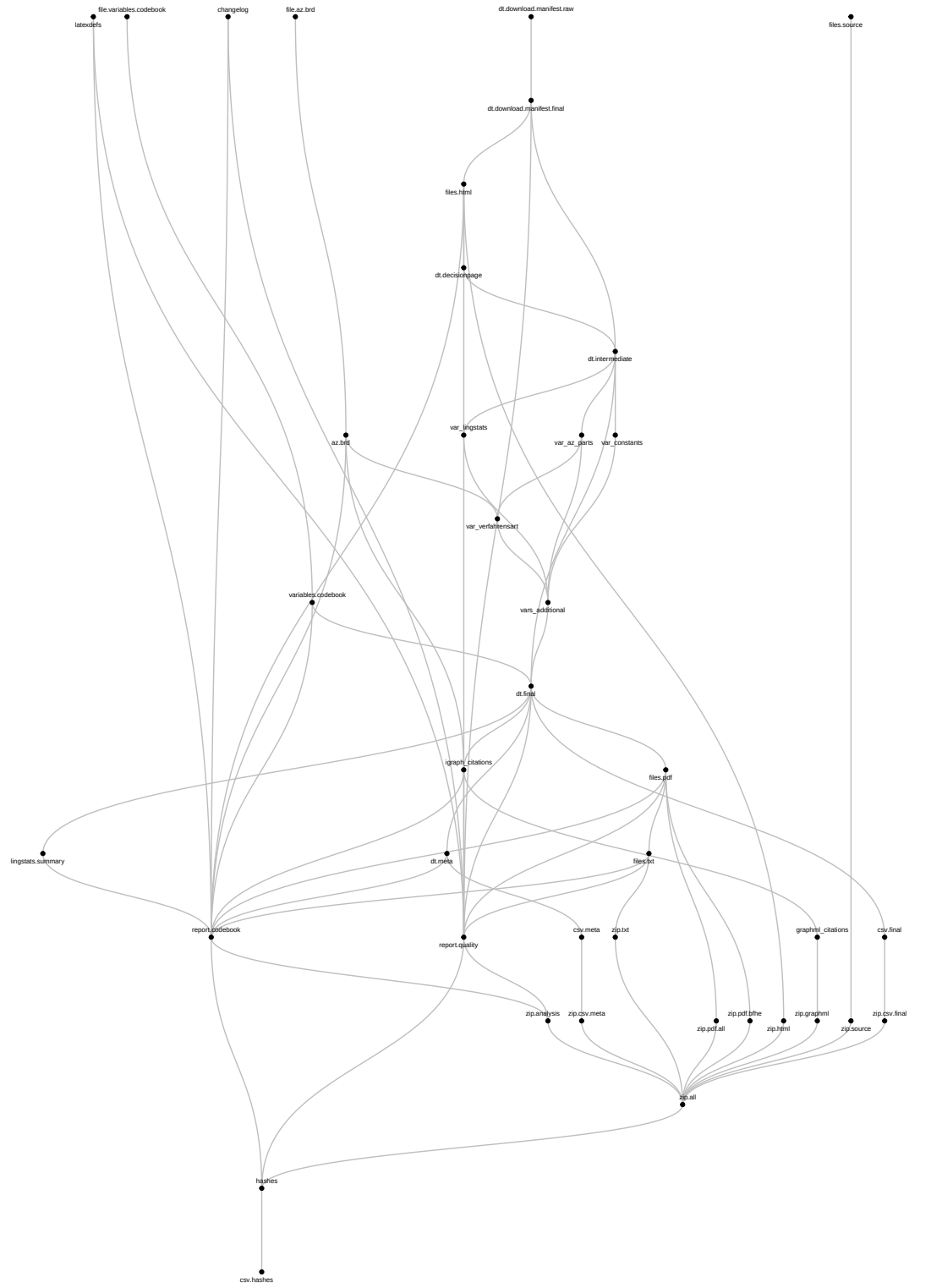


Abbildung 1: Der vollständige Prozess der Datensatz-Kompilierung.

3.5 Grenzen des Datensatzes

Nutzer:innen sollten folgende wichtige Grenzen beachten:

- Der Datensatz enthält nur das, was das Gericht auch tatsächlich veröffentlicht, nämlich begründete Entscheidungen (*publication bias*).
- Es kann aufgrund technischer Grenzen bzw. Fehler sein, dass manche — im Grunde verfügbare — Entscheidungen nicht oder nicht korrekt abgerufen werden (*automation bias*).
- Es werden HTML- und PDF-Dateien abgerufen (*file type bias*). Der Text der Entscheidungen in der CSV-Datei stammt aus den HTML-Dateien. Die PDF-Dateien sind beigefügt, falls bei der Extraktion fehler auftreten sollten, sind in der Regel aber für den Gebrauch in der traditionellen Rechtswissenschaft und -praxis gedacht.
- Erst ab dem 1. Januar 2010 sind begründete Entscheidungen des Bundesfinanzhofs einigermaßen vollständig veröffentlicht (*temporal bias*). Die Frequenztabellen geben hierzu genauer Auskunft.

3.6 Urheberrechtsfreiheit von Rohdaten und Datensatz

An den Entscheidungstexten und amtlichen Leitsätzen besteht gem. § 5 Abs. 1 UrhG kein Urheberrecht, da sie amtliche Werke sind. § 5 UrhG ist auf amtliche Datenbanken analog anzuwenden (BGH, Beschluss vom 28.09.2006, I ZR 261/03, »Sächsischer Ausschreibungsdienst«).

Alle eigenen Beiträge (z.B. durch Zusammenstellung und Anpassung der Metadaten) und damit den gesamten Datensatz stelle ich gemäß einer *CC0 1.0 Universal Public Domain Lizenz* vollständig urheberrechtsfrei.

3.7 Metadaten

3.7.1 Allgemein

Die Metadaten in den Dateinamen sind größtenteils unverändert von den jeweiligen Datenbankeinträgen aus der amtlichen Datenbank des Bundesfinanzhofs entnommen. Berechnet und hinzugefügt wurden durch den Autor des Datensatzes eine Reihe weitere Variablen, sowie in den Dateinamen der PDF/TXT-Dateien Unter- und Trennstriche, um die Maschinenlesbarkeit zu erleichtern. Der volle Satz an Metadaten ist nur in den CSV-Dateien enthalten. Alle hinzugefügten Metadaten sind vollständig maschinenlesbar dokumentiert. Sie sind entweder im Source Code enthalten, mit dem Source Code zusammen dokumentiert oder über dauerhaft stabile Identifikatoren (z.B. DOI) zitiert.

Die Dateinamen der PDF- und TXT-Dateien enthalten Gerichtsname, die Bezeichnung als V- oder NV-entscheidung, Datum, das offizielle Aktenzeichen, einen Zusatz zum Aktenzeichen und die vom BFH in der Datenbank genutzte einzigartige ID.

3.7.2 Schema für die Dateinamen

```
[gericht]_[bfhe]_[datum]_[spruchkoerper_az]_
[registerzeichen]_[eingangsnummer]_[eingangsjahr_az]_[bfh_id]
```

3.7.3 Beispiel eines Dateinamens

```
BFH_V_2023-07-11_X_R_17_22_STRE202310190.pdf
```

3.8 Qualitätsprüfung

Die Inhalte der Variablen wurden strikt validiert. Die möglichen Werte der jeweiligen Variablen wurden zudem durch Frequenztabellen und Visualisierungen auf ihre Plausibilität geprüft. Insgesamt werden zusammen mit jeder Kompilierung über 30 automatisierte Tests zur Qualitätsprüfung durchgeführt. Alle Ergebnisse der Qualitätsprüfungen sind aggregiert im Robustness Checks Report, im Compilation Report und einzeln im Archiv »ANALYSE« zusammen mit dem Datensatz veröffentlicht.

3.9 Grafische Darstellung

Die Robenfarbe der Richter:innen des Bundesfinanzhofs ist »karmesinrot«. Der Hex-Wert hierfür ist vermutlich #7e0731. Das ist besonders bei der Erstellung thematisch passender Diagrammen hilfreich. Alle im Compilation Report und diesem Codebook präsentierten Diagramme sind in diesem karmesinrot gehalten.

4 Varianten und Zielgruppen

Dieser Datensatz ist in verschiedenen Varianten verfügbar, die sich an unterschiedliche Zielgruppen richten. Zielgruppe sind nicht nur quantitativ forschende Rechtswissenschaftler:innen, sondern auch traditionell arbeitende Jurist:innen. Idealerweise müssen quantitative Methoden ohnehin immer durch qualitative Interpretation, Theoriebildung und kritische Auseinandersetzung verstärkt werden (*mixed methods approach*).

Lehrende werden von den vorbereiteten Tabellen und Diagrammen besonders profitieren, die bei der Erläuterung der Charakteristika der Daten hilfreich sein können und Zeit im universitären Alltag sparen. Alle Tabellen und Diagramme liegen auch als separate Dateien vor, um sie einfach z.B. in Präsentations-Folien oder Handreichungen zu integrieren.

4.1 CSV_Datensatz

Diese CSV-Datei ist die für statistische Analysen empfohlene Variante des Datensatzes. Sie enthält den Volltext aller Entscheidungen, sowie alle in diesem Codebook beschriebenen Metadaten. Jede Spalte entspricht einer Variable, jede Zeile einer Entscheidung.

Die Texte der Entscheidungen wurden aus dem HTML-Quelltext extrahiert, nicht aus den PDF-Dateien.

Empfohlen für Legal Tech und quantitative Forschung

4.2 CSV_Metadaten

Wie die primäre CSV-Variante, nur ohne die Entscheidungstexte. Sinnvoll für Analyst:innen, die sich nur für die Metadaten interessieren und Speicherplatz sparen wollen. Jede Spalte entspricht einer Variable, jede Zeile einer Entscheidung.

4.3 HTML

Diese Variante enthält die ursprünglichen HTML-Dateien, wie sie auf der Webseite des BFH präsentiert werden. Die HTML-Dateien sind die Grundlage für die CSV-Variante. Hilfreich, falls Probleme bei der Nutzung der CSV-Dateien auftreten.

4.4 PDF

Die PDF-Dokumente wie sie vom BFH auf der amtlichen Webseite bereitgestellt werden, jedoch verbessert durch semantisch hochwertige Dateinamen, die der leichteren Auffindbarkeit von Entscheidungen dienen. Die Dateinamen sind so konzipiert, dass sie auch für die traditionelle qualitative juristische Arbeit einen erheblichen Mehrwert bieten.

Im Vergleich zu den CSV-Dateien enthalten die Dateinamen nur einen reduzierten Umfang an Metadaten, um Kompatibilitätsprobleme zu vermeiden und die Lesbarkeit zu verbessern.

Neben dem vollen Datensatz ist für Praktiker:innen auch eine Variante aufbereitet, die nur *V-Entscheidungen* der amtlichen Sammlung enthalten.

Empfohlen für traditionelle juristische Forschung

4.5 TXT

Diese Variante enthält die vollständigen, aus den PDF-Dateien extrahierten Entscheidungstexte, aber nur einen reduzierten Umfang an Metadaten, der dem der PDF-Dateien entspricht. Die TXT-Dateien sind optisch an das Layout der PDF-Dateien angelehnt.

Geeignet für qualitativ arbeitende Forscher:innen, die nur wenig Speicherplatz oder eine langsame Internetverbindung zur Verfügung haben oder für quantitativ arbeitende Forscher:innen, die beim Einlesen der CSV-Dateien Probleme haben.

4.6 ANALYSE

Dieses Archiv enthält alle während dem Kompilierungs- und Prüfprozess erstellten Tabellen (CSV) und Diagramme (PDF, PNG) im Original. Sie sind inhaltsgleich mit den in diesem Codebook verwendeten Tabellen und Diagrammen.

Das PDF-Format eignet sich besonders für die Verwendung in gedruckten Publikationen, das PNG-Format besonders für die Darstellung im Internet. Analyst:innen mit fortgeschrittenen Kenntnissen in R können auch auf den Source Code zurückgreifen. Empfohlen für Nutzer:innen die einzelne Inhalte aus dem Codebook für andere Zwecke (z.B. Präsentationen, eigene Publikationen) weiterverwenden möchten.

Hilfreich bei der Vorbereitung von Lehre und Forschung

5 Variablen

5.1 Datenstruktur

```
## Classes 'data.table' and 'data.frame':   10885 obs. of  33 variables:
##  $ aktenzeichen          : chr  "VII S 34/09" "IV R 43/07" "VII B 118/09" "VII
    B 165/09" ...
##  $ bfh_id                : chr  "STRE201050294" "STRE201050190" "STRE201050120
    " "STRE201050238" ...
##  $ doc_id                : chr  "BFH_NV_2010-01-04_VII_S_34_9_STRE201050294" "
    BFH_NV_2010-01-05_IV_R_43_7_STRE201050190" "BFH_NV_2010-01-07_VII_B_118_9_
    STRE201050120" "BFH_NV_2010-01-07_VII_B_165_9_STRE201050238" ...
##  $ ecli                  : chr  NA NA NA NA ...
##  $ gericht               : chr  "BFH" "BFH" "BFH" "BFH" ...
##  $ text_leitsatz         : chr  "1. NV: Ob bei Verpachtung eines Betriebes der
    Pächter oder der Verpächter Milcherzeuger ist, bedarf einer umfas"| __
    truncated__ "NV: Der Gesellschafter einer zweigliedrigen GbR ist nach § 48
    Abs. 1 Nr. 3 FGO klagebefugt und damit notwendig "| __truncated__ "NV:
    Restschuldbefreiung erlangt der Insolvenzschuldner nicht mit dem Ablauf der
    sog. Wohlverhaltensphase, sonde"| __truncated__ "1. NV: Die für die
    Erstattung nach Art. 901 Abs. 2 ZKDVO erforderlichen Nachweise können nicht
    allein mit den i"| __truncated__ ...
##  $ url_html              : chr  "https://www.bundesfinanzhof.de/de/
    entscheidung/entscheidungen-online/detail/STRE201050294/" "https://www.
    bundesfinanzhof.de/de/entscheidung/entscheidungen-online/detail/STRE201050190
    /" "https://www.bundesfinanzhof.de/de/entscheidung/entscheidungen-online/
    detail/STRE201050120/" "https://www.bundesfinanzhof.de/de/entscheidung/
    entscheidungen-online/detail/STRE201050238/" ...
##  $ url_pdf               : chr  "https://www.bundesfinanzhof.de/de/
    entscheidung/entscheidungen-online/detail/pdf/STRE201050294?type=1646225765"
    "https://www.bundesfinanzhof.de/de/entscheidung/entscheidungen-online/detail/
    pdf/STRE201050190?type=1646225765" "https://www.bundesfinanzhof.de/de/
    entscheidung/entscheidungen-online/detail/pdf/STRE201050120?type=1646225765"
    "https://www.bundesfinanzhof.de/de/entscheidung/entscheidungen-online/detail/
    pdf/STRE201050238?type=1646225765" ...
##  $ zeichen               : num  14086 8666 4027 6056 11204 ...
##  $ tokens                : num  2140 1494 629 1037 1909 ...
##  $ typen                 : num  648 494 290 395 583 764 540 968 302 370 ...
##  $ saetze                : num  47 99 29 70 87 75 78 164 25 49 ...
##  $ adv                   : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ bfhe                  : chr  "NV" "NV" "NV" "NV" ...
##  $ normen                : chr  "FGO § 133a|FGO § 118 Abs 2|FGO § 126 Abs 3 S
    1 Nr 2|EGV 1788/2003 Art 5 Buchst c|GG Art 103 Abs 1|FGO § 96 Abs 1 S 1" "FGO
     § 48 Abs 1 Nr 3|FGO § 60 Abs 3" "InsO § 294 Abs 3|InsO § 300|InsO § 301" "ZK
     Art 238|ZK Art 239|ZKDV Art 901 Abs 2|ZKDV Art 902 Abs 1|ZKDV Art 904 Buchst
     a|EWGV 2454/93 Art 901 Abs 2|EW"| __truncated__ ...
##  $ pkh                   : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ registerzeichen       : chr  "S" "R" "B" "B" ...
##  $ spruchkoerper_az      : chr  "VII" "IV" "VII" "VII" ...
##  $ spruchkoerper_db      : chr  "VII. Senat" "IV. Senat" "VII. Senat" "VII.
    Senat" ...
##  $ titel                 : chr  "Tatsachenfeststellung und rechtliche Wü
    rdigung bei Ermittlung des Milcherzeugers - Keine Bindung an rechtliche "| __
    truncated__ "Notwendige Beiladung des aus einer zweigliedrigen GbR
```

```
    ausgeschiedenen Gesellschafters bei Streit über Einkunftsart der GbR" "Kein
    insolvenzrechtliches Aufrechnungsverbot zwischen Aufhebung des
    Insolvenzverfahrens und Erteilung der Restschuldbefreiung" "Zoll: Erstattung
    der Einfuhrabgaben bei Wiederausfuhr der Ware" ...
##  $ verfahrensart         : chr  "Sonstige Verfahren" "Revision" "Beschwerden"
    "Beschwerden" ...
##  $ vorinstanz            : chr  "vorgehend BFH , 25. Mai 2009, Az: VII R
    28/08" "vorgehend Finanzgericht Berlin-Brandenburg , 13. Juni 2007, Az: 15 K
    3202/04 B" "vorgehend Finanzgericht des Landes Sachsen-Anhalt , 17. März
    2009, Az: 2 K 1682/08" "vorgehend FG Hamburg, 26. Mai 2009, Az: 4 K 58/07"
    ...
##  $ datum                 : Date, format: "2010-01-04" "2010-01-05" ...
##  $ entscheidungsjahr     : int  2010 2010 2010 2010 2010 2010 2010 2010 2010
    2010 ...
##  $ eingangsjahr_az       : int  9 7 9 9 9 7 8 7 9 9 ...
##  $ eingangsjahr_iso      : num  2009 2007 2009 2009 2009 ...
##  $ eingangsnummer        : int  34 43 118 165 99 34 159 24 113 110 ...
##  $ veroeffentlichung     : Date, format: "2010-05-19" "2010-04-14" ...
##  $ veroeffentlichungsjahr: int  2010 2010 2010 2010 2010 2010 2010 2010 2010
    2010 ...
##  $ doi_concept           : chr  "10.5281/zenodo.7691840" "10.5281/zenodo
    .7691840" "10.5281/zenodo.7691840" "10.5281/zenodo.7691840" ...
##  $ doi_version           : chr  "10.5281/zenodo.14622341" "10.5281/zenodo
    .14622341" "10.5281/zenodo.14622341" "10.5281/zenodo.14622341" ...
##  $ lizenz                : chr  "Creative Commons Zero 1.0 Universal" "
    Creative Commons Zero 1.0 Universal" "Creative Commons Zero 1.0 Universal" "
    Creative Commons Zero 1.0 Universal" ...
##  $ version               : chr  "2025-01-14" "2025-01-14" "2025-01-14"
    "2025-01-14" ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

5.2 Allgemeine Hinweise

- **Doppelte Codierung der Spruchkörper** — Für viele Urteile sind die Spruchkörper doppelt enthalten, einmal aus der Datenbank (Variable »spruchkoerper_db«), einmal durch das Aktenzeichen (Variable »spruchkoerper_az«).
- **Fehlende Werte** sind immer mit »NA« codiert.
- **Strings** können grundsätzlich alle in UTF-8 definierten Zeichen (insbesondere Buchstaben, Zahlen und Sonderzeichen) enthalten.
- Die **Reihenfolge** der Variablen entspricht der im CSV-Datensatz. Der Datensatz wird automatisiert darauf getestet, ob alle Variablen im Datensatz auch in diesem Codebook dokumentiert sind.

5.3 ID-Variablen

ID-Variablen stellen verschiedene Identifikatoren für die Entscheidung zur Verfügung, beispielsweise Aktenzeichen oder ECLI.

Variable	Type	Description
aktenzeichen	String	Das amtliche Aktenzeichen im Format [senatsnummer] [registerzeichen] [eingangsnummer] / [eingangsjahr] [ggf. zusatz_az]. Quelle: BFH-Datenbank.
bfh_id	String	Die vom BFH vergebene ID der Entscheidung in der amtlichen Datenbank. Die ID ist einzigartig. Quelle: BFH-Datenbank.
doc_id	String	(Nur CSV) Dateiname der PDF- und TXT-Dateien, ohne Dateiendung. Quelle: Kompositum verschiedener Variablen des Datensatzes.
ecli	String	(Nur CSV-Datei) Der European Case Law Identifier (ECLI) der Entscheidung. Die ECLI ist einzigartig. Die ECLI ist vor allem dann hilfreich, wenn dieser Datensatz mit anderen Datensätzen zusammengeführt und Duplikate vermieden werden sollen. Nicht für alle Entscheidungen vorhanden. Quelle: BFH-Datenbank, aus den HTML-Dateien extrahiert
gericht	String	Name des Gerichts. Es ist nur der Wert »BFH« vergeben. Dies ist der ECLI-Code für »Bundesfinanzhof«. Diese Variable dient vor allem zur einfachen und transparenten Verbindung der Daten mit anderen Datensätzen. Quelle: Autor des Datensatzes.

5.4 Text-Variablen

Text-Variablen enthalten den Volltext der Entscheidung, Teilstücke davon (z.B. Leitsätze), den Umfang des Volltextes (Zeichen, Tokens, Typen, Sätze) und dessen Quelle (URLs zu Volltexten).

Variable	Type	Description
text	String	(Nur CSV) Volltext der Entscheidung. Achtung: wenige Entscheidungen (ca. 20) haben keinen Text, weil es Parallelentscheidungen sind. Quelle: BFH-Datenbank, aus den HTML-Dateien extrahiert

(continued)

Variable	Type	Description
text_leitsatz	String	(Nur CSV) Text der Leitsätze der Entscheidung. Quelle: BFH-Datenbank, aus den HTML-Dateien extrahiert.
url_html	String	(Nur CSV) Link zum Volltext der Entscheidung als HTML in der amtlichen Datenbank des Gerichts. Quelle: BFH-Datenbank.
url_pdf	String	(Nur CSV) Link zum Volltext der Entscheidung als PDF in der amtlichen Datenbank des Gerichts. Quelle: BFH-Datenbank.
zeichen	Integer	(Nur CSV) Die Anzahl Zeichen eines Dokumentes. Quelle: Mit R berechnet.
tokens	Integer	(Nur CSV) Die Anzahl Tokens (beliebige Zeichenfolge getrennt durch whitespace) eines Dokumentes. Diese Zahl kann je nach Tokenizer und verwendeten Einstellungen erheblich schwanken. Für diese Berechnung wurde eine reine Tokenisierung ohne Entfernung von Inhalten durchgeführt. Benutzen Sie diesen Wert eher als Anhaltspunkt für die Größenordnung denn als exakte Aussage und führen sie ggf. mit ihrer eigenen Software eine Kontroll-Rechnung durch. Quelle: Mit R berechnet
typen	Integer	(Nur CSV) Die Anzahl *einzigartiger* Tokens (beliebige Zeichenfolge getrennt durch whitespace) eines Dokumentes. Diese Zahl kann je nach Tokenizer und verwendeten Einstellungen erheblich schwanken. Für diese Berechnung wurde eine reine Tokenisierung und Typenzählung ohne Entfernung von Inhalten durchgeführt. Benutzen Sie diesen Wert eher als Anhaltspunkt für die Größenordnung denn als exakte Aussage und führen sie ggf. mit ihrer eigenen Software eine Kontroll-Rechnung durch. Quelle: mit R berechnet.

(continued)

Variable	Type	Description
saetze	Integer	(Nur CSV) Die Anzahl Sätze. Die Definition entspricht in etwa dem üblichen Verständnis eines Satzes. Die Regeln für die Bestimmung von Satzanfang und Satzende sind im Detail allerdings sehr komplex und in »Unicode Standard: Annex No 29« beschrieben. Diese Zahl kann je nach Software und verwendeten Einstellungen erheblich schwanken. Für diese Berechnung wurde eine reine Zählung ohne Entfernung von Inhalten durchgeführt. Benutzen Sie diesen Wert eher als Anhaltspunkt für die Größenordnung denn als exakte Aussage und führen sie ggf. mit ihrer eigenen Software eine Kontroll-Rechnung durch. Quelle: Mit R berechnet

5.5 Thematische Variablen

Thematische Variablen geben Auskunft über eine grobe thematische Zuordnung der Entscheidung, beispielsweise zu Registerzeichen, Verfahrensart, Normen, Vorinstanz.

Variable	Type	Description
adv	Logical	Ob es sich um eine Entscheidung zur Aufhebung der Vollziehung (AdV) handelt. Entweder TRUE oder FALSE. Quelle: REGEX-Suche nach “AdV” im Aktenzeichen.
bfhe	String	(Nur CSV) Ob die Entscheidung in der amtlichen Sammlung veröffentlicht wird (»V«) oder nicht (»NV«). Der BFH spricht hier auch von V-Entscheidungen und NV-Entscheidungen. Quelle: BFH-Datenbank.
normen	String	(Nur CSV) Die rechtlichen Normen, die von der Entscheidung betroffen sind. Normen beginnen jeweils mit dem Gesetzesnamen, gefolgt von der genauen Fundstelle. Mehrere Normen sind durch einen vertikale Balken (»|«) getrennt. Quelle: BFH-Datenbank.
pkh	Logical	(Nur CSV) Ob es sich um eine Entscheidung zur Prozesskostenhilfe (PKH) handelt. Entweder TRUE oder FALSE. Quelle: REGEX-Suche nach “PKH” im Aktenzeichen.

(continued)

Variable	Type	Description
registerzeichen	String	Das amtliche Registerzeichen. Eine Erläuterung der Abkürzungen findet sich im Abschnitt 6. Quelle: Mit REGEX aus Variable “aktenzeichen” extrahiert.
spruchkoerper_az	String	Der im Aktenzeichen angegebene Spruchkörper. Die Senate sind mit römischen Ziffern nummeriert. Der Große Senat ist mit »GrS« gekennzeichnet. Quelle: Mit REGEX aus Aktenzeichen extrahiert.
spruchkoerper_db	String	(Nur CSV) Der Spruchkörper, wie er in der amtlichen Datenbank des Gerichts eingetragen ist. Quelle: BFH-Datenbank
titel	String	(Nur CSV) Der Titel der Entscheidung. Enthält eine kurze thematische und rechtliche Einordnung. Quelle: BFH-Datenbank
verfahrensart	String	(Nur CSV) Die Verfahrensart, auf die das Registerzeichen hinweist. Siehe auch Abschnitt 6. Quelle: Abgleich von Registerzeichen mit AZ-BRD-Datensatz.
vorinstanz	String	(Nur CSV) Die Vorinstanz des Verfahrens. Quelle: BFH-Datenbank.

5.6 Temporale Variablen

Temporale Variablen bieten Informationen zu wichtigen Zeitpunkten, wie Verkündung der Entscheidung, Veröffentlichung der Entscheidung oder Eingang des Verfahrens.

Variable	Type	Description
datum	Date	Datum der Entscheidung im Format YYYY-MM-DD (ISO-8601). Quelle: BFH-Datenbank.
entscheidungsjahr	Integer	Jahr der Entscheidung im Format YYYY (ISO-8601). Quelle: Berechnet aus Variable “datum”.
eingangsjahr_az	Integer	Eingangsjahr laut Aktenzeichen. Das Jahr in dem das Verfahren beim Gericht anhängig wurde. Das Format ist eine zweistellige Jahreszahl (YY). Quelle: Mit REGEX aus Variable “aktenzeichen” extrahiert.

(continued)

Variable	Type	Description
eingangsjahr_iso	Integer	(Nur CSV) Eingangsjahr im Format YYYY-MM-DD (ISO-8601). Quelle: Aus Variable “eingangsjahr_az” berechnet.
eingangsnummer	Integer	Eingangsnummer. Verfahren des gleichen Eingangsjahres erhalten vom Gericht eine fortlaufene Nummer (Ordinalzahl) in der Reihenfolge ihres Eingangs. Quelle: Mit REGEX aus Variable “aktenzeichen” extrahiert.
veroeffentlichung	Date	(Nur CSV) Das Datum der Veröffentlichung der Entscheidung im Format YYYY-MM-DD (ISO-8601). Quelle: BFH-Datenbank.
veroeffentlichungsjahr	Integer	(Nur CSV) Das Jahr der Veröffentlichung der Entscheidung im Format YYYY (ISO-8601). Quelle: Aus Variable “veroeffentlichung” berechnet.

5.7 Meta-Variablen

Meta-Variablen beziehen sich auf den Datensatz selbst. Sie dokumentieren Versionsnummer, verschiedene DOIs und die Lizenz des Datensatzes. Streng genommen sind sie innerhalb des Datensatzes Konstanten (weil der Inhalt immer gleich ist) und nur im Vergleich zwischen Datensätzen echte Variablen.

Variable	Type	Description
doi_concept	String	(Nur CSV) Der Digital Object Identifier (DOI) des Gesamtkonzeptes des Datensatzes. Dieser ist langzeit-stabil (persistent). Über diese DOI kann via www.doi.org immer die **aktuellste Version** des Datensatzes abgerufen werden. Prinzip F1 der FAIR-Data Prinzipien (»data are assigned globally unique and persistent identifiers«) empfiehlt die Dokumentation jeder Messung mit einem persistenten Identifikator. Selbst wenn die CSV-Dateien ohne Kontext weitergegeben werden kann ihre Herkunft so immer zweifelsfrei und maschinenlesbar bestimmt werden. Quelle: Vom Autor hinzugefügt.

(continued)

Variable	Type	Description
doi_version	String	(Nur CSV) Der Digital Object Identifier (DOI) der **konkreten Version** des Datensatzes. Dieser ist langzeit-stabil (persistent). Über diese DOI kann via www.doi.org immer diese konkrete Version des Datensatzes abgerufen werden. Prinzip F1 der FAIR-Data Prinzipien (»data are assigned globally unique and persistent identifiers«) empfiehlt die Dokumentation jeder Messung mit einem persistenten Identifikator. Selbst wenn die CSV-Dateien ohne Kontext weitergegeben werden kann ihre Herkunft so immer zweifelsfrei und maschinenlesbar bestimmt werden. Quelle: Vom Autor hinzugefügt.
lizenz	String	Die Lizenz für den Gesamtdatensatz. In diesem Datensatz immer »Creative Commons Zero 1.0 Universal«. Quelle: Vom Autor hinzugefügt.
version	Date	(Nur CSV) Die Versionsnummer des Datensatzes im Format YYYY-MM-DD (Langform nach ISO-8601). Die Versionsnummer entspricht immer dem Datum an dem der Datensatz erstellt und die Daten von der Webseite des Gerichts abgerufen wurden. Quelle: Vom Autor hinzugefügt.

6 Registerzeichen

Die Tabelle der Registerzeichen und der ihnen zugeordneten Verfahrensarten stammt aus dem folgenden Datensatz: »Seán Fobbe (2021). Aktenzeichen der Bundesrepublik Deutschland (AZ-BRD). Version 1.0.1. Zenodo. DOI: 10.5281/zenodo.4569564.«

Die im Datensatz enthaltenen Registerzeichen wurden jeweils um die runden Klammern bereinigt, um Probleme bei der Nutzung unter Windows zu vermeiden.

Die Bedeutung des Registerzeichens »ER-S« ist mir nicht klar, aber möglicherweise ist es eine Kombination aus den Registerzeichen »E«, »R« und »S«. Ich arbeite an der Aufklärung.

Registerzeichen	Verfahrensart
AR	Allgemeines Register: Vorverfahren oder sonstige Verfahrensarten
B	Beschwerden
E	Erinnerung zu Kosten und Streitwert, Erinnerung in Kostenfestsetzungsverfahren
GrS	Großer Senat
K	Entschädigungsklagen wegen überlanger Verfahrensdauer
PKH	Prozesskostenhilfe
R	Revision
S	Sonstige Verfahren
ER-S	Unklar

7 Linguistische Kennzahlen

7.1 Erläuterung der Kennzahlen und Diagramme

Zur besseren Einschätzung des inhaltlichen Umfangs des Korpus dokumentiere ich an dieser Stelle die Verteilung der Werte für einige klassische linguistische Kennzahlen:

Kennzahl	Definition
Zeichen	Zeichen entsprechen grob den *Graphemen*, den kleinsten funktionalen Einheiten in einem Schriftsystem. Beispiel: das Wort »Richterin« besteht aus 9 Zeichen.
Tokens	Eine beliebige Zeichenfolge, getrennt durch whitespace-Zeichen, d.h. ein Token entspricht in der Regel einem »Wort«, kann aber auch Zahlen, Sonderzeichen oder sinnlose Zeichenfolgen enthalten, weil es rein syntaktisch berechnet wird.
Typen	Einzigartige Tokens. Beispiel: wenn das Token »Finanzrecht« zehnmal in einer Entscheidung vorhanden ist, wird es als ein Typ gezählt.
Sätze	Entsprechen in etwa dem üblichen Verständnis eines Satzes. Die Regeln für die Bestimmung von Satzanfang und Satzende sind im Detail aber sehr komplex und in »Unicode Standard: Annex No 29« beschrieben.

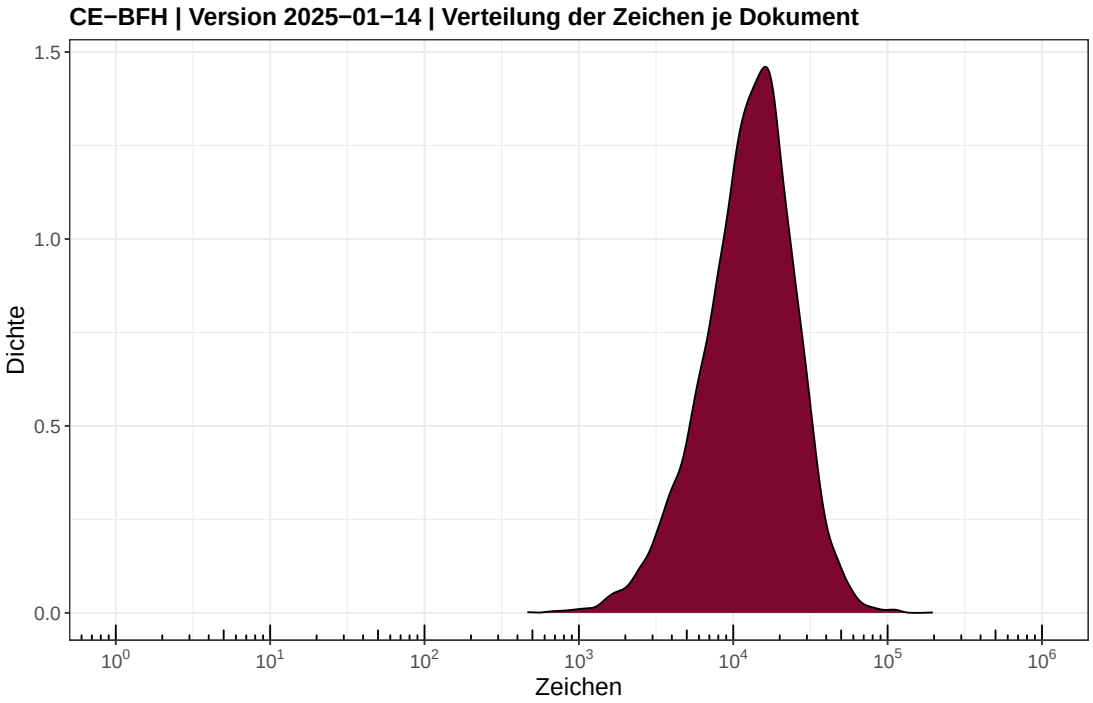
Es handelt sich bei den Diagrammen jeweils um »Density Charts«, die sich besonders dafür eignen die Schwerpunkte von Variablen mit stark schwankenden numerischen Werten zu visualisieren. Die Interpretation ist denkbar einfach: je höher die Kurve, desto dichter sind in diesem Bereich die Werte der Variable. Der Wert der y-Achse kann außer Acht gelassen werden, wichtig sind nur die relativen Flächenverhältnisse und die x-Achse.

Vorsicht bei der Interpretation: Die x-Achse ist logarithmisch skaliert, d.h. in 10er-Potenzen und damit nicht-linear. Die kleinen Achsen-Markierungen zwischen den Schritten der Exponenten sind eine visuelle Hilfestellung um diese nicht-Linearität zu verstehen.

7.2 Werte der Kennzahlen

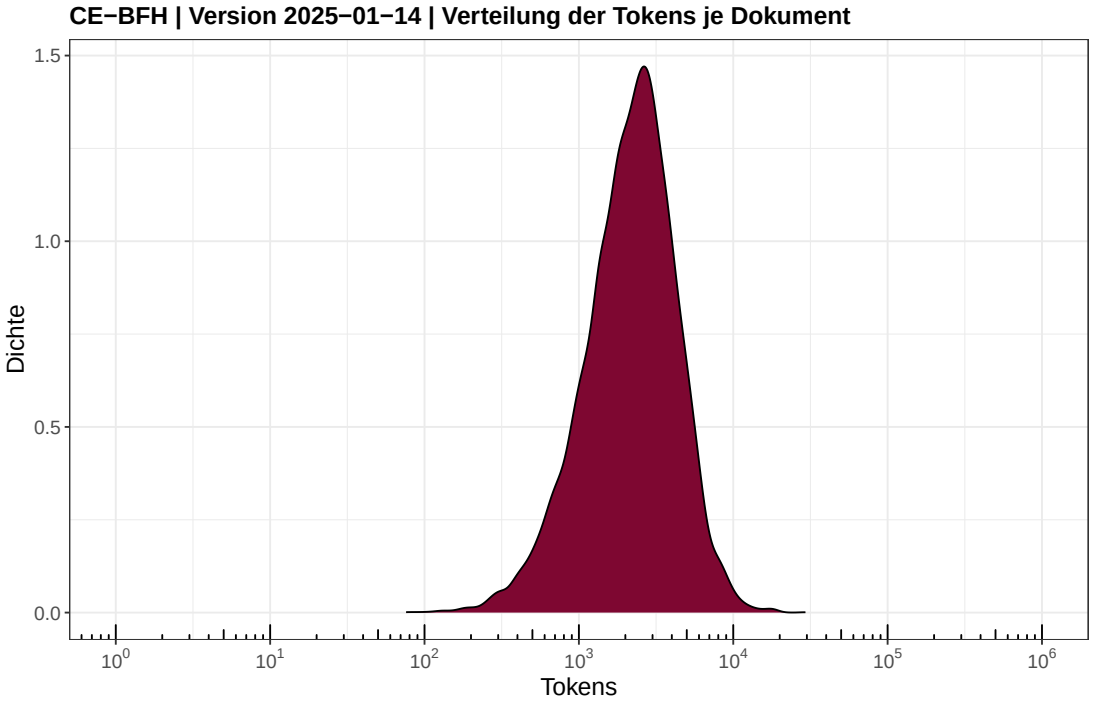
Variable	Summe	Min	Quart1	Median	Mittel	Quart3	Max
zeichen	167,148,907	0	8,169	13,203	15,355.89	19,701	196,294
tokens	28,573,459	0	1,394	2,264	2,625.03	3,383	29,344
typen	277,562	0	490	691	736.19	917	3,610
saetze	1,204,320	0	61	96	110.64	142	947

7.3 Verteilung Zeichen



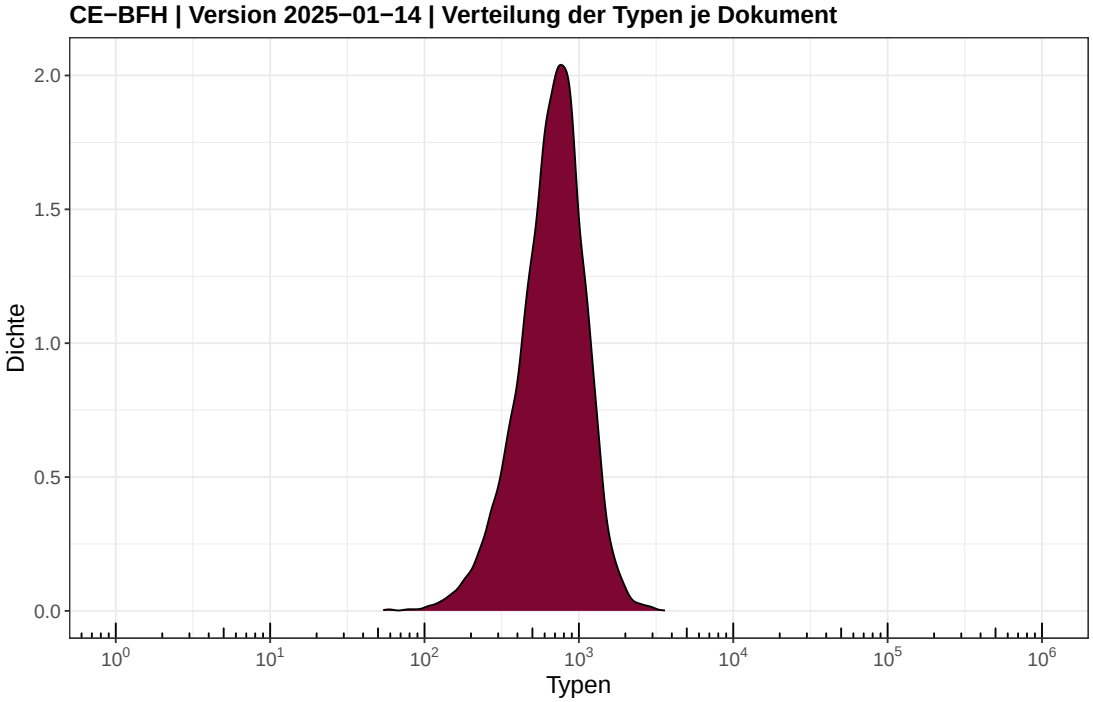
Fobbe | DOI: 10.5281/zenodo.14622341

7.4 Verteilung Tokens



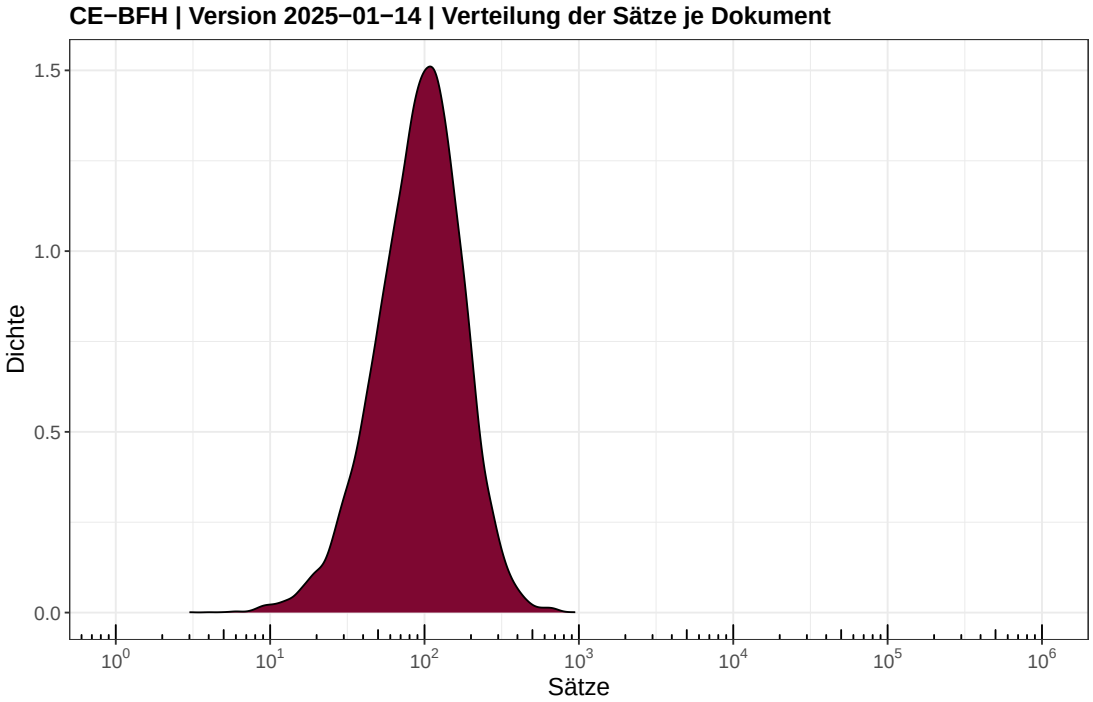
Fobbe | DOI: 10.5281/zenodo.14622341

7.5 Verteilung Typen



Fobbe | DOI: 10.5281/zenodo.14622341

7.6 Verteilung Sätze



Fobbe | DOI: 10.5281/zenodo.14622341

8 Zitationsnetzwerk des Bundesfinanzhofs (Beta)

8.1 Überblick

Der Datensatz enthält zusätzlich eine spezialisierte Variante, die Zitate des BFH aus den Entscheidungstexten zu seiner eigenen Rechtsprechung extrahiert und in strukturierter Form aufbereitet.

Diese Variante ist noch in der Beta-Testphase. Folgende Rechtsprechungszitate sind enthalten:

- Zitate von Aktenzeichen zu Aktenzeichen
- Zitate von Aktenzeichen zu BFHE

Achtung: Zitierende Entscheidungen können nur solche sein, für die im Korpus der Volltext dokumentiert ist (d.h. normalerweise ab 2010). Zitierte Entscheidungen können aus der gesamten Rechtsprechung stammen (d.h. aus allen Jahren).

Zitate unter Angabe des Aktenzeichens sind weniger genau als Zitate zu konkreten Entscheidungen (bei denen Datum und ggf. Kollisionsziffer nötig sind). Sie stellen aber eine gute Näherung dar. Die Auflösung der Zitate auf Entscheidungsebene ist geplant und wird in Zukunft mitveröffentlicht.

BFHE-Zitate können exakt einer zitierten Entscheidung zugeordnet werden. Die Quell-Dokumente sind jedoch nur mit dem Aktenzeichen hinterlegt, um eine Konkordanz mit dem Rest des Datensatzes herzustellen. Die Auflösung von Quell-Dokumenten nach BFHE ist mir leider derzeit nicht möglich, da ich aktuell keine Entsprechungstabelle zwischen amtlichen Sammlungen und Aktenzeichen/Datum-Zitaten habe.

8.2 Technische Hinweise

Das Zitationsnetzwerk wird als GraphML-Datei angeboten und kann z.B. einfach in graphische Software wie Gephi[8] importiert und ohne Programmierkenntnisse genutzt werden. Formal handelt sich um einen gewichteten, gerichteten Graphen (Digraph). Die Anzahl der Knoten gibt die Anzahl der BFHE-Entscheidungen und Aktenzeichen mit eingehenden und/oder ausgehenden Zitaten an. Die Anzahl der Kanten gibt die Anzahl der Knoten-Paare mit mindestens einem Zitat an. Die Gewichte der Kanten geben die Anzahl der Zitate zwischen Knoten an. Die Ausgangsstärke gibt die Summe aller einfachen Zitate an.

Beachten Sie auch bitte folgende Punkte:

- Das gesamte Netzwerk ist sehr groß und die Analyse ist daher ohne weitere Einschränkungen rechenintensiv und anspruchsvoll. In der Regel sollten Sie das Netzwerk auf die für Sie interessanten Teile reduzieren.
- Zur Reduktion des Netzwerks auf eine handliche Größe stelle ich v.a. zwei Variablen bereit: den Senat und das Registerzeichen.
- Die Extraktion mit *regular expressions* ist nicht perfekt. Es kann daher sein, dass Zitate fehlen, wenn sie nicht als solche erkannt wurden, wegen Tippfehlern, ungewöhnlichem Textumfeld etc. Es ist aktuell unklar wieviele Zitate fehlen könnten, weil es keinen Goldstandard zum Abgleich gibt. Wenn Ihnen größere Fehlbestände auffallen, melden Sie sich bitte.

[8] https://gephi.github.io/

- Sie können über die Variable »bfh-alternative« alle BFHE-Zitate auswählen (TRUE) oder nur Aktenzeichen-Zitate betrachten (FALSE)

Metric	Value	Metric	Value
Number of Nodes	39,875.00	Mean In-Degree	3.59
Number of Edges	143,281.00	Max In-Degree	120.00
Strength (Out)	178,806.00	Min In-Degree	0.00
Mean Degree	7.19	Mean Out-Degree	3.59
Max Degree	148.00	Max Out-Degree	139.00
Min Degree	0.00	Min Out-Degree	0.00

8.3 Metadaten

Die Knoten des Netzwerks sind mit Metadaten aus dem Hauptdatensatz angereichert. Deshalb sind grundsätzlich nur im Hauptdatensatz vorhandene Aktenzeichen (d.h. solche die in der BFH-Datenbank veröffentlicht sind) mit allen Metadaten verbunden.

Für alle anderen Zitate konnte ich nur solche Metadaten hinterlegen, die aus dem Aktenzeichen (Registerzeichen, Senatsnummer) oder dem BFH-Zitat (BFH ja/nein, Nummer des Bandes) mit REGEX zu extrahieren waren.

Hinweis: die Variable »bfhe« gibt an, ob die das Aktenzeichen in der BFH-Datenbank als V- oder NV-Entscheidung markiert ist. Die Variable »bfhe-alternative« dagegen ist TRUE/FALSE und gibt an, ob das Zitat (d.h. der Name des Knotens) die Zeichenkette »BFHE« enthält.

Folgende Metadaten-Variablen sind enthalten:

```
##  [1] "adv"                 "band"                "bfh_id"
##  [4] "bfhe"                "bfhe-alternative"    "datum"
##  [7] "doc_id"              "doi_concept"         "doi_version"
## [10] "ecli"                "eingangsjahr_az"     "eingangsjahr_iso"
## [13] "eingangsnummer"      "entscheidungsjahr"   "gericht"
## [16] "lizenz"              "name"                "normen"
## [19] "pkh"                 "registerzeichen"     "saetze"
## [22] "spruchkoerper_az"    "spruchkoerper_db"    "text_leitsatz"
## [25] "titel"               "tokens"              "typen"
## [28] "url_html"            "url_pdf"             "verfahrensart"
## [31] "veroeffentlichung"   "veroeffentlichungsjahr" "version"
## [34] "vorinstanz"          "zeichen"
```

8.4 Methodik Aktenzeichen

Dieser Datensatz enthält sowohl zitierte Aktenzeichen (Aktenzeichen-zu-Aktenzeichen-Zitate), als auch Zitate von Aktenzeichen zu BFHE-Zitate (Aktenzeichen-zu-Sammlung-Zitate).

Aktenzeichen sind verhältnismäßig einfach zu erfassen. Die Funktion *f.citation_network.R* erstellt erstellt eine komplexe REGEX, die jeweils die relevanten Registerzeichen in die

Suche aufnimmt. Der Source Code ist zu komplex um ihn hier im Detail zu besprechen, sehen Sie sich bei Interesse bitte die Funktion genauer an.

Um konkrete Entscheidungen zu zitieren müsste zusätzlich zum Aktenzeichen noch das Datum berücksichtigt werden. Weil dies die REGEX deutlich komplizierter macht, ist dieser Schritt noch in Arbeit. Im Hauptdatensatz sind allerdings 98.73 % aller ausgehenden Aktenzeichen einzigartig (unabhängig vom Datum), sodass das Aktenzeichen eine gute Näherung darstellt.

8.5 Methodik BFHE

Die Zitate zu der amtlichen Sammlung BFHE werden aus dem Volltext in einem Zwei-Stufen-Verfahren extrahiert, ähnlich wie in Coupette, *Juristische Netzwerkforschung* (Mohr Siebeck 2019), S. 241–244.

8.5.1 Erste Stufe

In der **ersten Stufe** werden die Zitierblöcke lokalisiert und aus dem Volltext gesammelt. Es wird die starke Annahme getroffen, dass Zitierblöcke mit »BFHE« (ignoriert Groß- und Kleinschreibung) eingeleitet werden und nur Whitespace, Zahlen, gewisse Sonderzeichen und gewisse Buchstaben enthalten.

Zitierblöcke enden in der Regel mit einer runden Klammer, die in der REGEX nicht enthalten ist, um sie als Grenzzeichen zu nutzen. Auch Gleichheitszeichen (=) sind nicht enthalten, damit die REGEX vor einem Hinweis auf einen alternative Abdruck abbricht.

Die konkreten regular expressions (REGEX) sind die folgenden:

```
"BFHE[\\s\\d\\[\\];,\\.<>Rnfu-]+"
```

8.5.2 Zweite Stufe

In der **zweiten Stufe** werden aus allen Zitierblöcken die einzelnen Zitate extrahiert, standardisiert und mit der Ausgangsentscheidung verbunden. Die Extraktion trifft die starke Annahme, dass eine Entscheidung der amtlichen Sammlungen entweder mit »BFHE« oder bei einem Mehrfachzitat in einem Zitierblock mit einem Semikolon eingeleitet wird. Folgende REGEX kommen dabei zum Einsatz:

```
regex.cite <- paste0("(BFHE|;)\\s*", # hooks
                     "\\d{1,3},\\s*", # Volume
                     "\\d{1,3}") # Page

print(regex.cite)
```

```
## [1] "(BFHE|;)\\s*\\d{1,3},\\s*\\d{1,3}"
```

Damit findet man zwei Varianten von Einzelzitaten:

- »BFHE 248, 287«

- »; 248, 287«

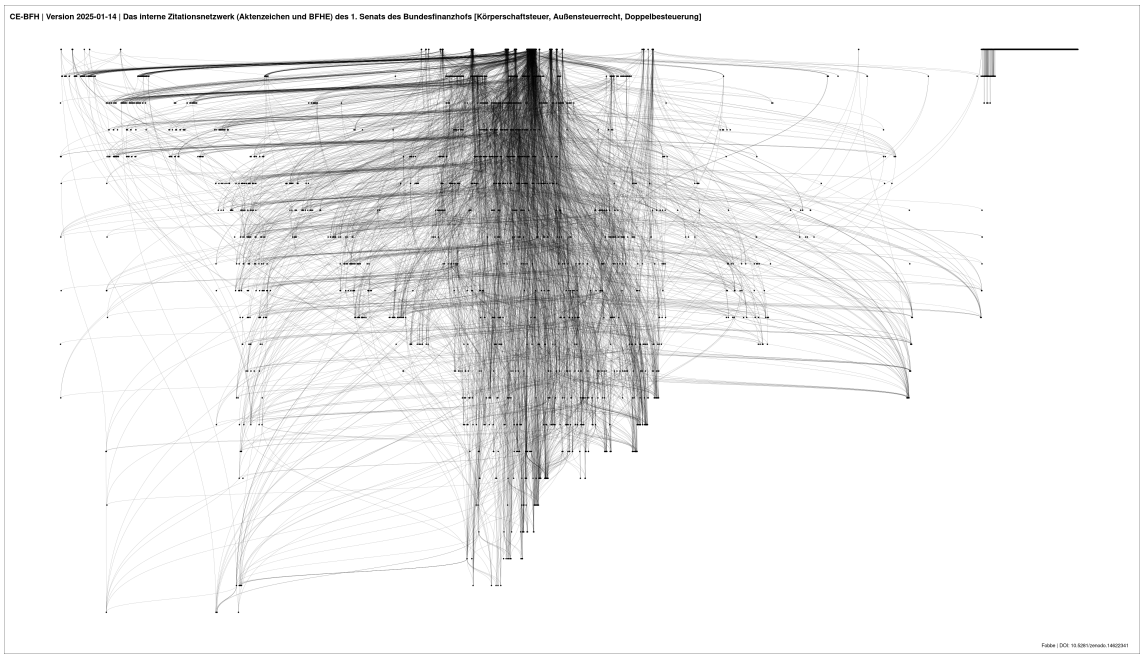
Die Einzelzitate werden anschließend bereinigt und standardisiert. Zum Ende hin werden Selbstzitate entfernt und Metadaten hinzugefügt.

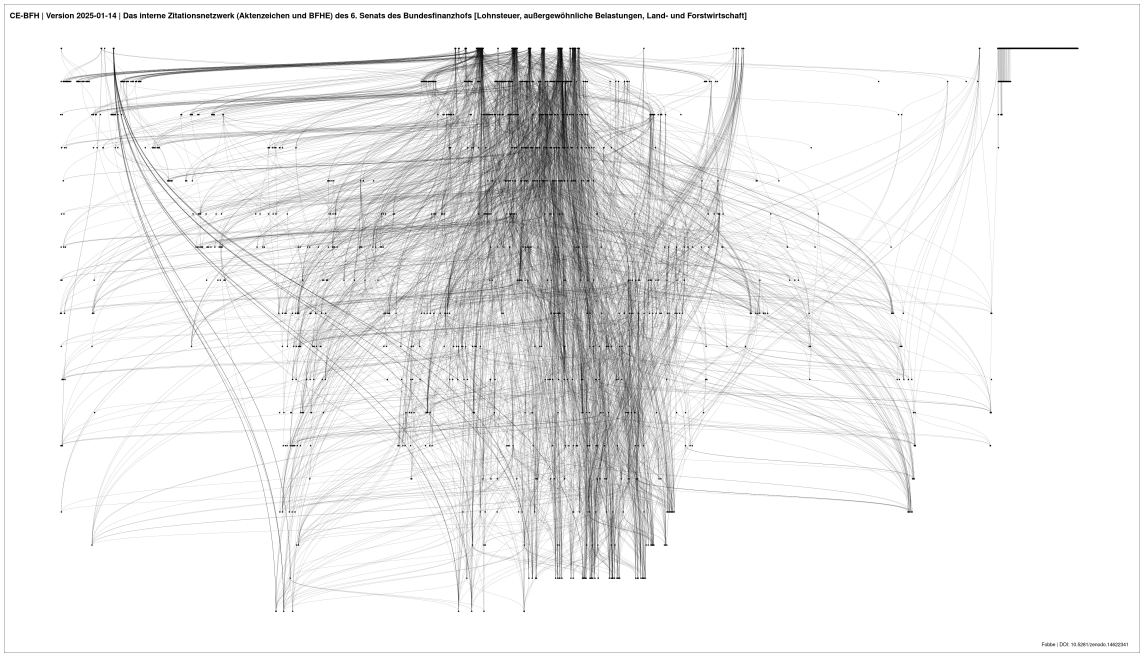
8.5.3 Grenzen

Die Extraktion mit regulären Ausdrücken hat Grenzen. Insbesondere folgende Probleme führen zur Nichterkennung von Zitaten:

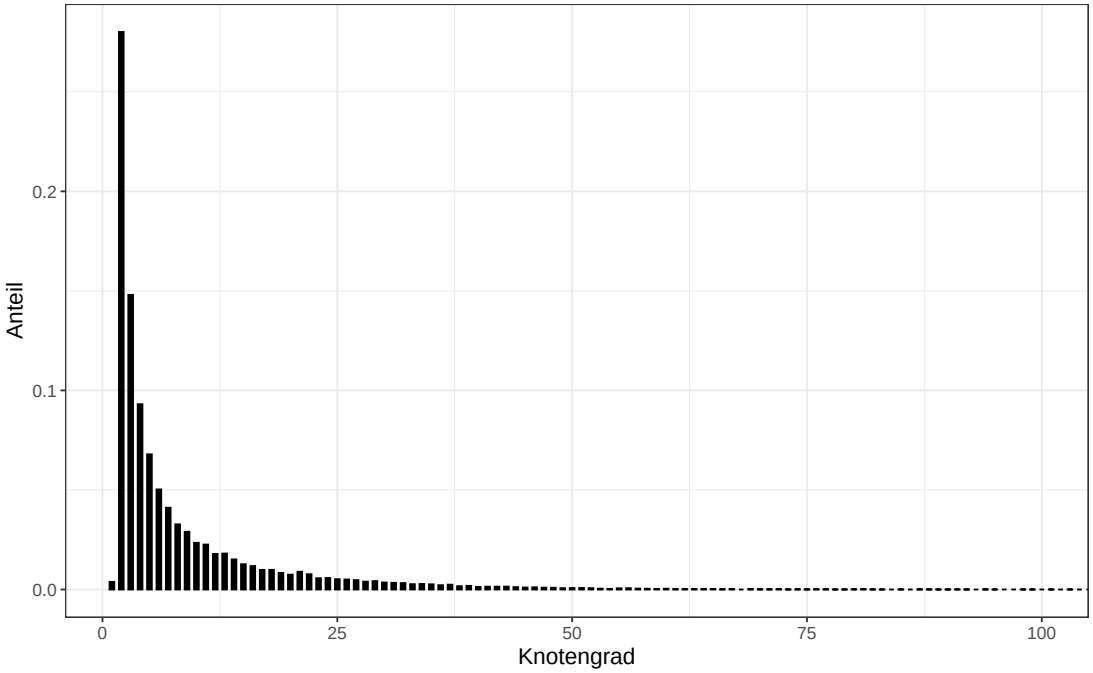
- Tippfehler (außer Groß- und Kleinschreibung)
- Unregelmäßge Zitierweise
- Verkürzte Schreibweisen, beispielsweise BVerfGE 60, 162: »BVerfGE 3, 19 (27), 383 (394); 4, 375 (381 f.);« — das Beispiel stammt aus Coupette (2019: 246)
- Einfügung von Entscheidungsnamen wie in BVerfGE 42, 143: »BVerfGE 7, 198 (205ff) - Lüth -; 18, 85 (92f); 30, 173 (187f, 196f) - Mephisto -; 32, 311 (316)« — das Beispiel stammt aus Coupette (2019: 246)

8.6 Visualisierungen des Zitationsnetzwerks



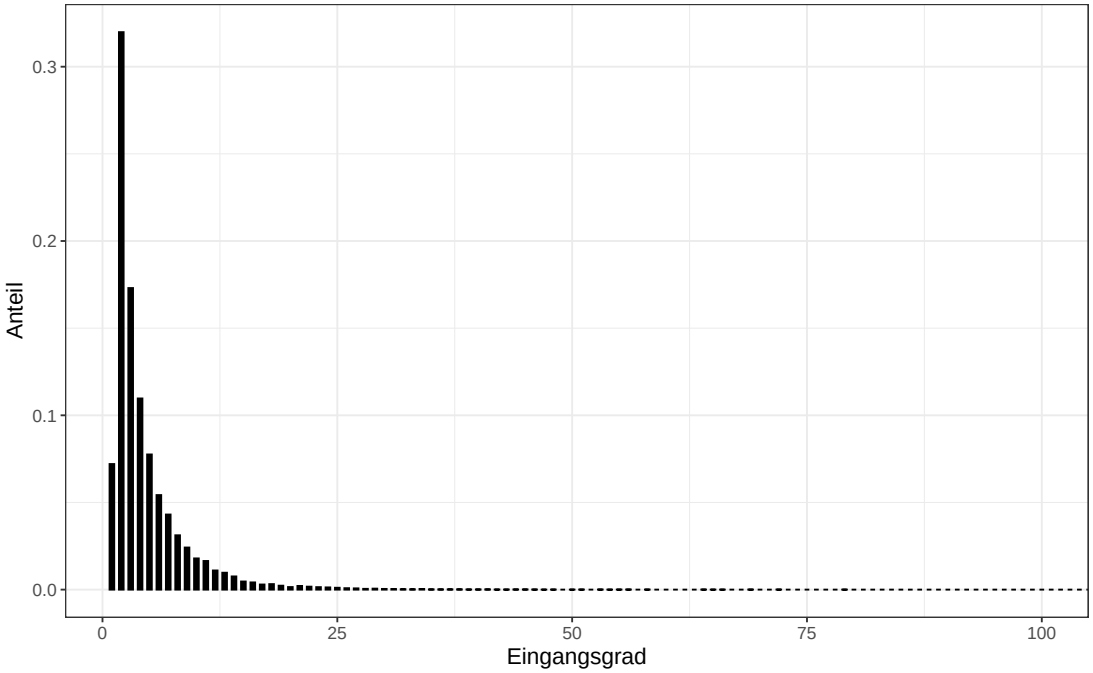


CE−BFH | Version 2025−01−14 | Verteilung der Knotengrade bis 100 im Zitationsnetzwerk



Fobbe | DOI: 10.5281/zenodo.14622341

CE−BFH | Version 2025−01−14 | Verteilung der Eingangsgrade bis 100 im Zitationsnetzwerk



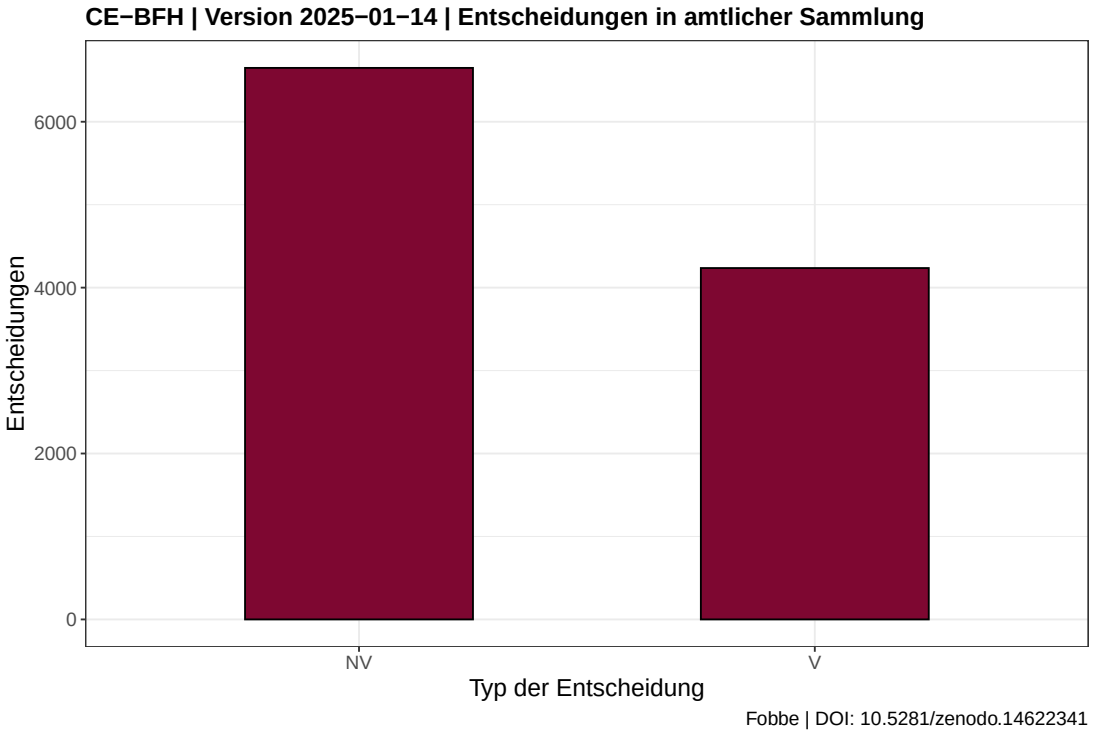
Fobbe | DOI: 10.5281/zenodo.14622341

9 Inhalt des Korpus

9.1 Zusammenfassung

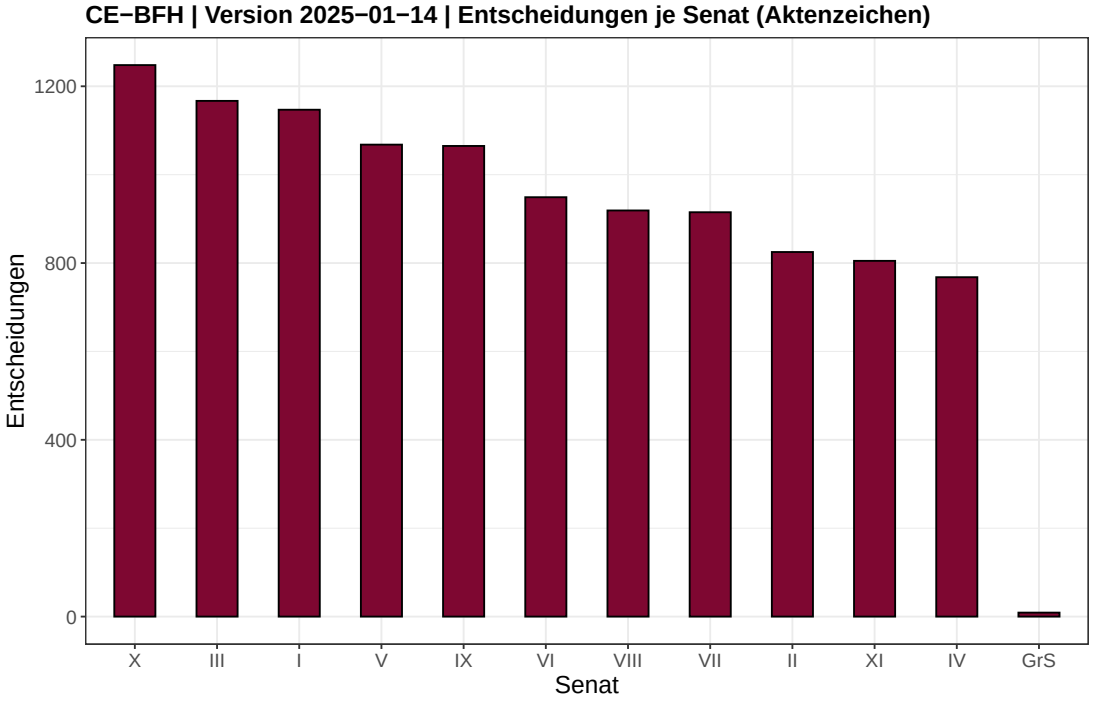
Variable	Anzahl	Min	Quart1	Median	Mean	Quart3	Max
entscheidungsjahr	15	2010	2012	2014	2015.31	2019	2024
eingangsjahr_iso	23	2001	2011	2013	2013.88	2017	2024
eingangsnummer	266	1	17	35	49.62	64	281

9.2 Nach Typ der Entscheidung



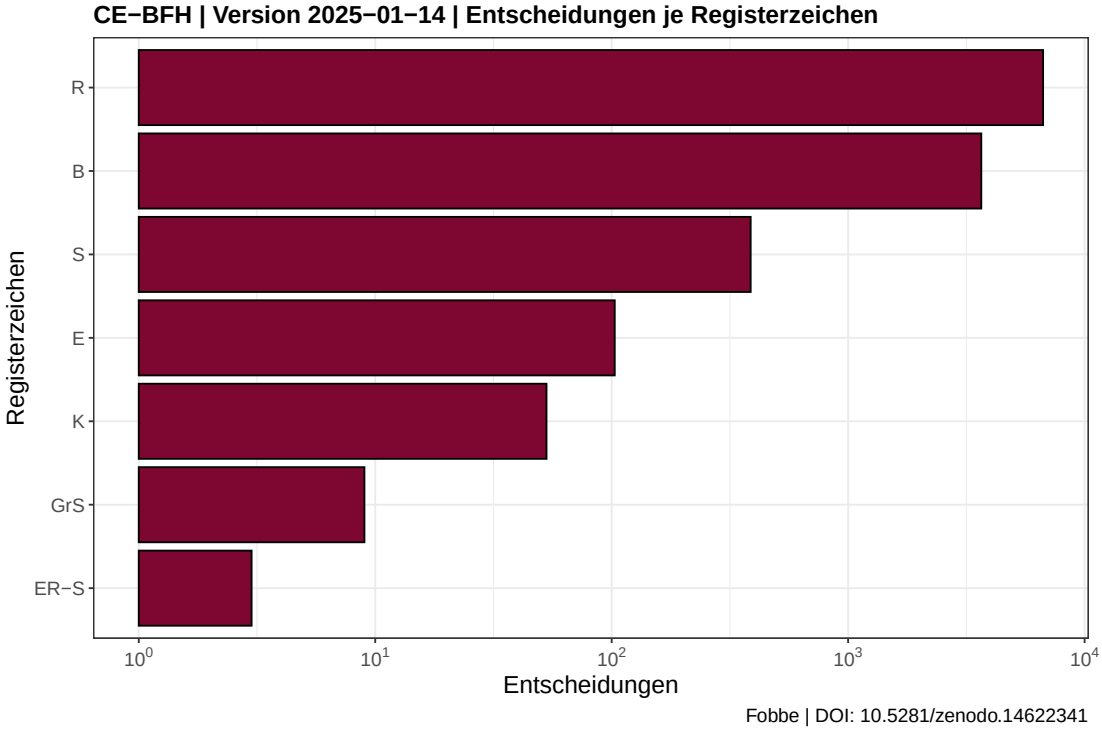
Typ	Entscheidungen	% Gesamt	% Kumulativ
NV	6649	61.08	61.08
V	4236	38.92	100.00
Total	10885	100.00	100.00

9.3 Nach Spruchkörper (Aktenzeichen)



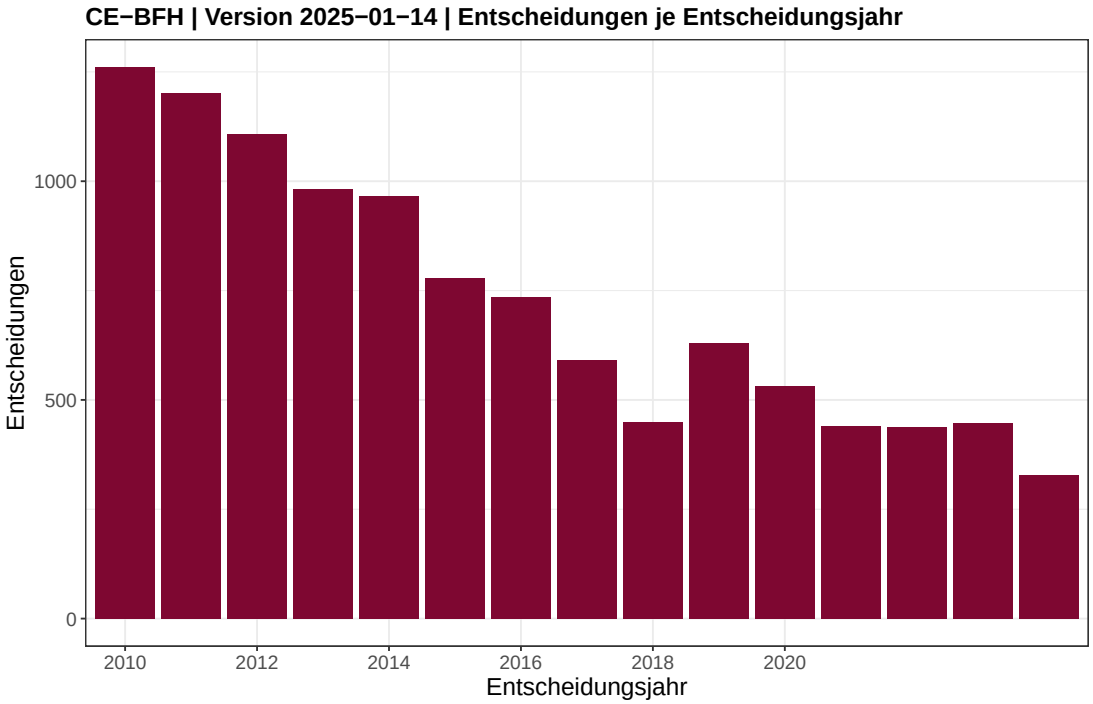
Senat	Entscheidungen	% Gesamt	% Kumulativ
GrS	9	0.08	0.08
I	1147	10.54	10.62
II	825	7.58	18.20
III	1167	10.72	28.92
IV	768	7.06	35.98
IX	1065	9.78	45.76
V	1068	9.81	55.57
VI	949	8.72	64.29
VII	915	8.41	72.70
VIII	919	8.44	81.14
X	1248	11.47	92.60
XI	805	7.40	100.00
Total	10885	100.00	100.00

9.4 Nach Registerzeichen



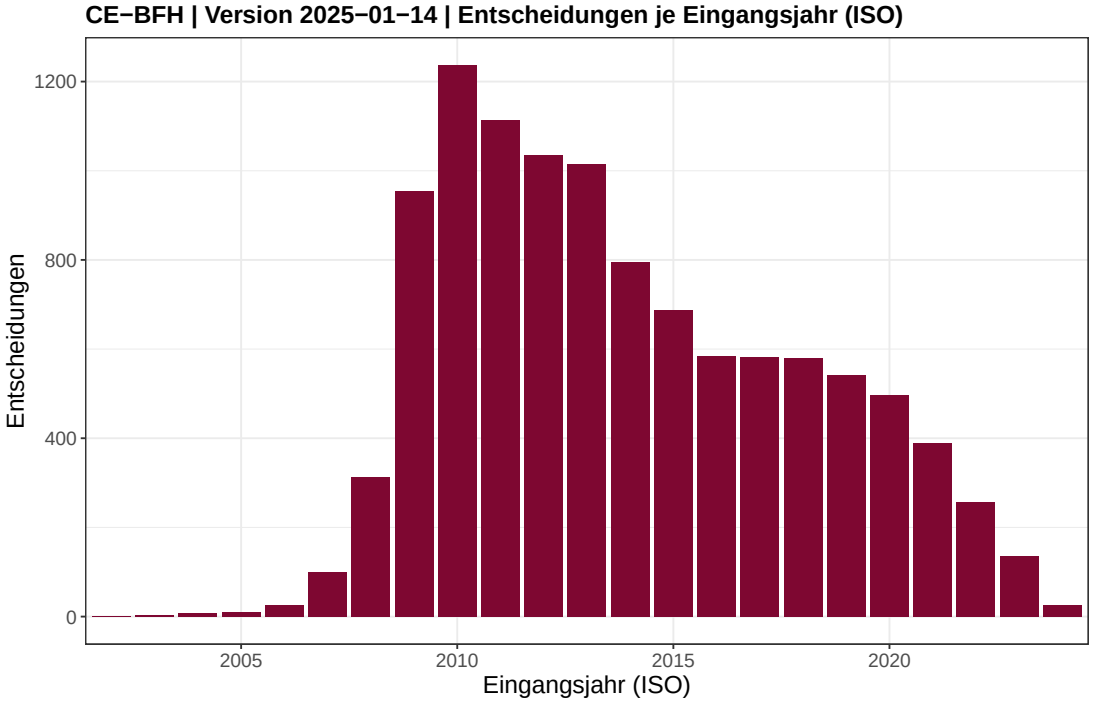
Registerzeichen	Entscheidungen	% Gesamt	% Kumulativ
B	3656	33.59	33.59
E	103	0.95	34.53
ER-S	3	0.03	34.56
GrS	9	0.08	34.64
K	53	0.49	35.13
R	6674	61.31	96.44
S	387	3.56	100.00
Total	10885	100.00	100.00

9.5 Nach Entscheidungsjahr



Entscheidungsjahr	Entscheidungen	% Gesamt	% Kumulativ
2010	1261	11.58	11.58
2011	1201	11.03	22.62
2012	1107	10.17	32.79
2013	981	9.01	41.80
2014	967	8.88	50.68
2015	778	7.15	57.83
2016	735	6.75	64.58
2017	591	5.43	70.01
2018	449	4.12	74.14
2019	629	5.78	79.92
2020	532	4.89	84.80
2021	439	4.03	88.84
2022	438	4.02	92.86
2023	448	4.12	96.98
2024	329	3.02	100.00
Total	10885	100.00	100.00

9.6 Nach Eingangsjahr (ISO)

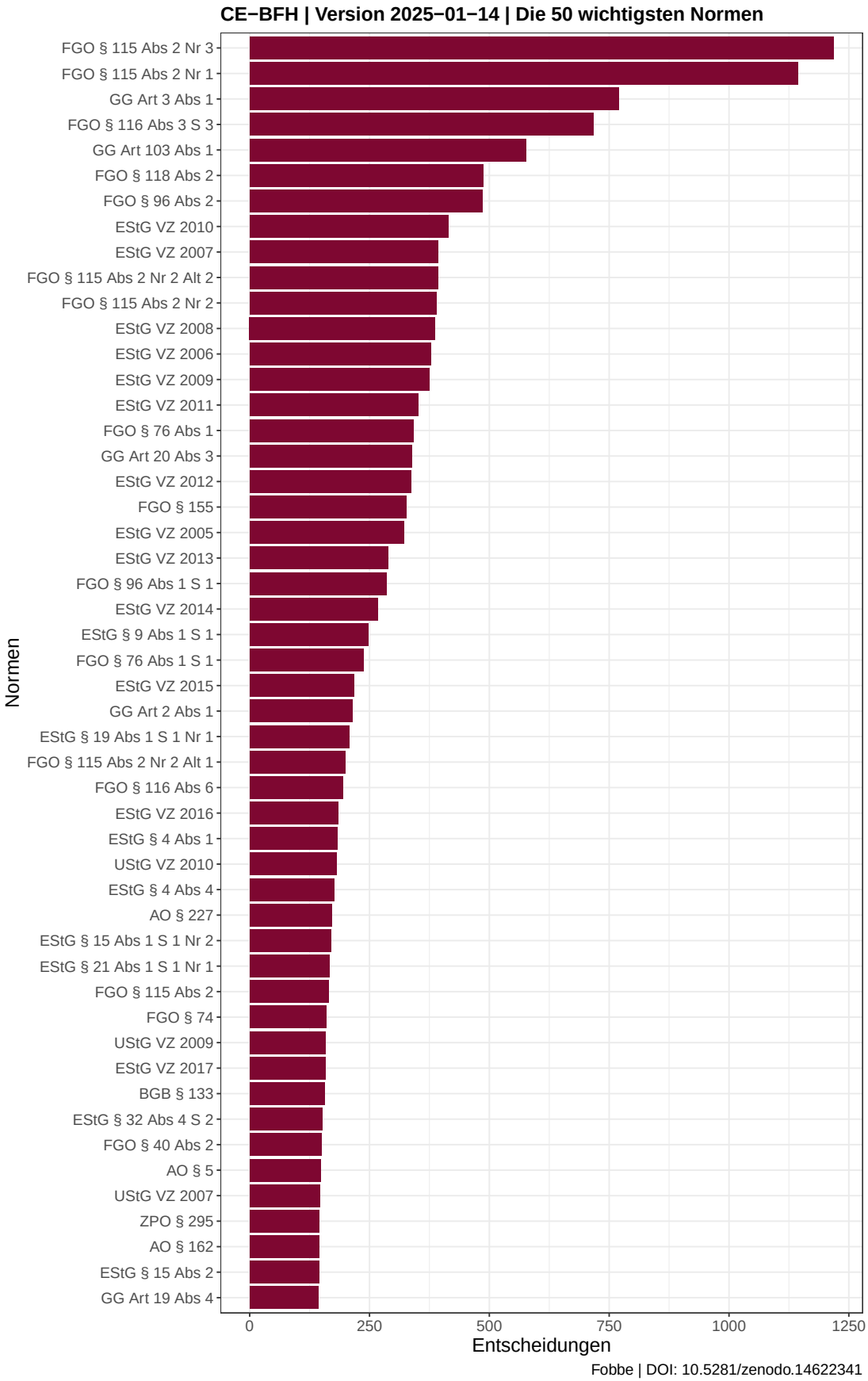


Eingangsjahr	Entscheidungen	% Gesamt	% Kumulativ
2001	1	0.01	0.01
2003	3	0.03	0.04
2004	7	0.06	0.10
2005	10	0.09	0.19
2006	26	0.24	0.43
2007	100	0.92	1.35
2008	312	2.87	4.22
2009	954	8.76	12.98
2010	1237	11.36	24.35
2011	1114	10.23	34.58
2012	1035	9.51	44.09
2013	1016	9.33	53.42
2014	794	7.29	60.72
2015	687	6.31	67.03
2016	584	5.37	72.39
2017	581	5.34	77.73
2018	579	5.32	83.05
2019	541	4.97	88.02
2020	496	4.56	92.58

(continued)

Eingangsjahr	Entscheidungen	% Gesamt	% Kumulativ
2021	389	3.57	96.15
2022	257	2.36	98.51
2023	136	1.25	99.76
2024	26	0.24	100.00
Total	10885	100.00	100.00

9.7 Nach Normen



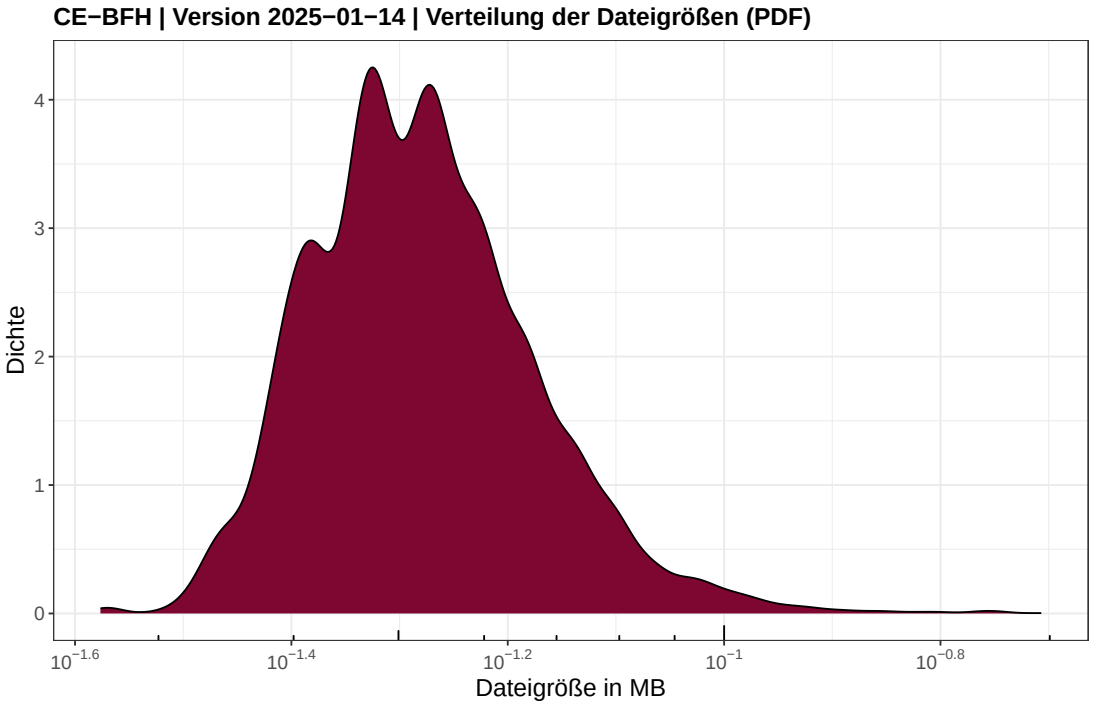
Fobbe | DOI: 10.5281/zenodo.14622341

Normen	Entscheidungen	% Gesamt
FGO § 115 Abs 2 Nr 3	1218	1.64
FGO § 115 Abs 2 Nr 1	1144	1.54
GG Art 3 Abs 1	770	1.04
FGO § 116 Abs 3 S 3	717	0.96
GG Art 103 Abs 1	577	0.78
FGO § 118 Abs 2	487	0.65
FGO § 96 Abs 2	485	0.65
EStG VZ 2010	415	0.56
EStG VZ 2007	393	0.53
FGO § 115 Abs 2 Nr 2 Alt 2	392	0.53
FGO § 115 Abs 2 Nr 2	390	0.52
EStG VZ 2008	387	0.52
EStG VZ 2006	378	0.51
EStG VZ 2009	374	0.50
EStG VZ 2011	351	0.47
FGO § 76 Abs 1	341	0.46
GG Art 20 Abs 3	338	0.45
EStG VZ 2012	336	0.45
FGO § 155	326	0.44
EStG VZ 2005	321	0.43
EStG VZ 2013	288	0.39
FGO § 96 Abs 1 S 1	285	0.38
EStG VZ 2014	267	0.36
EStG § 9 Abs 1 S 1	248	0.33
FGO § 76 Abs 1 S 1	238	0.32
EStG VZ 2015	217	0.29
GG Art 2 Abs 1	214	0.29
EStG § 19 Abs 1 S 1 Nr 1	208	0.28
FGO § 115 Abs 2 Nr 2 Alt 1	200	0.27
FGO § 116 Abs 6	195	0.26
EStG VZ 2016	185	0.25
EStG § 4 Abs 1	182	0.24
UStG VZ 2010	181	0.24
EStG § 4 Abs 4	176	0.24
AO § 227	171	0.23
EStG § 15 Abs 1 S 1 Nr 2	170	0.23

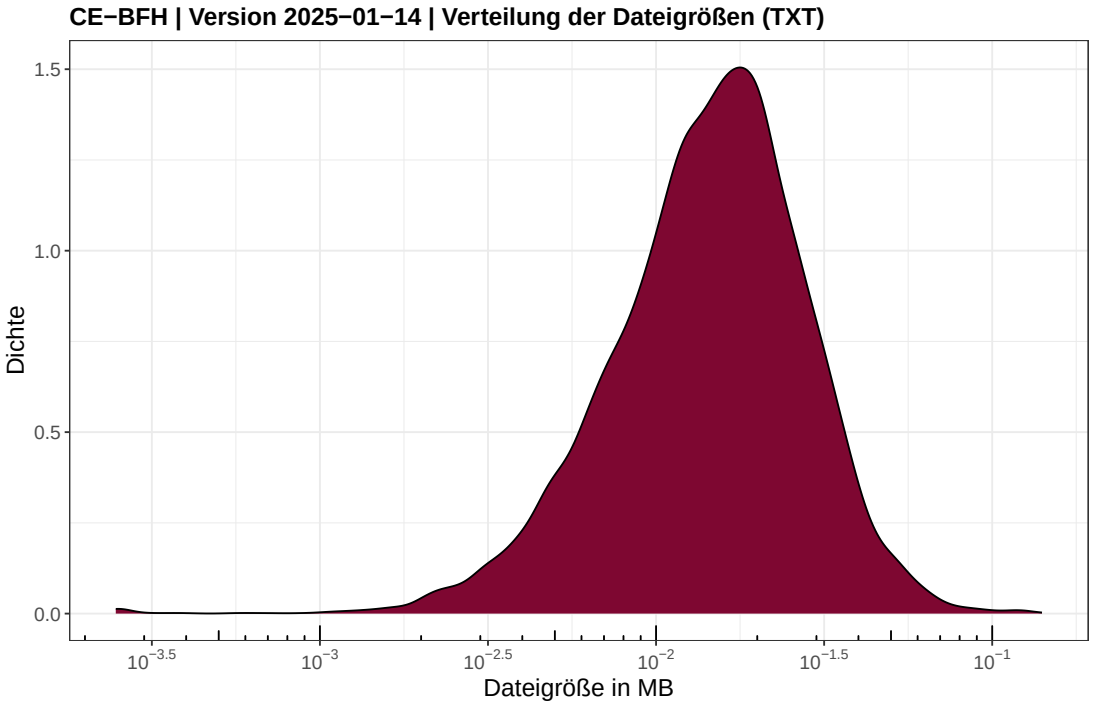
(continued)

Normen	Entscheidungen	% Gesamt
EStG § 21 Abs 1 S 1 Nr 1	166	0.22
FGO § 115 Abs 2	164	0.22
FGO § 74	159	0.21
EStG VZ 2017	158	0.21
UStG VZ 2009	158	0.21
BGB § 133	157	0.21
EStG § 32 Abs 4 S 2	152	0.20
FGO § 40 Abs 2	150	0.20
AO § 5	148	0.20
UStG VZ 2007	146	0.20
AO § 162	145	0.19
ZPO § 295	145	0.19
EStG § 15 Abs 2	144	0.19
GG Art 19 Abs 4	143	0.19

10 Dateigrößen

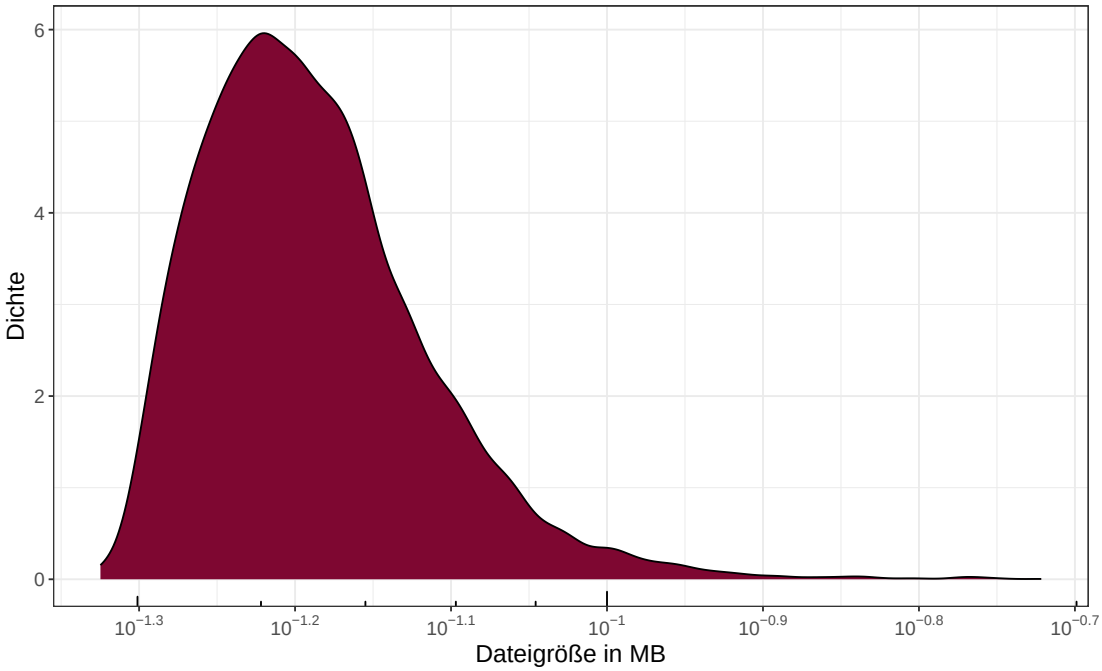


Fobbe | DOI: 10.5281/zenodo.14622341



Fobbe | DOI: 10.5281/zenodo.14622341

CE−BFH | Version 2025−01−14 | Verteilung der Dateigrößen (HTML)



Fobbe | DOI: 10.5281/zenodo.14622341

11 Kryptographische Signaturen

11.1 Zwei-Phasen-Signatur

Die Integrität und Echtheit der einzelnen Archive des Datensatzes sind durch eine Zwei-Phasen-Signatur sichergestellt.

In **Phase I** werden während der Kompilierung für jedes ZIP-Archiv, das Codebook und die Robustness Checks Hash-Werte in zwei verschiedenen Verfahren (SHA2-256 und SHA3-512) berechnet und in einer CSV-Datei dokumentiert.

In **Phase II** werden diese CSV-Datei und der Compilation Report mit meinem persönlichen geheimen GPG-Schlüssel signiert. Dieses Verfahren stellt sicher, dass die Kompilierung von jedermann durchgeführt werden kann, insbesondere im Rahmen von Replikationen, die persönliche Gewähr für Ergebnisse aber dennoch vorhanden bleibt.

11.2 Persönliche GPG-Signatur

Die während der Kompilierung des Datensatzes erstellte CSV-Datei mit den Hash-Prüfsummen und der Compilation Report sind mit meiner persönlichen GPG-Signatur versehen. Der mit dieser Version korrespondierende Public Key ist sowohl mit dem Datensatz als auch mit dem Source Code hinterlegt. Er hat folgende Kenndaten:

Name: Sean Fobbe (fobbe-data@posteo.de)

Fingerabdruck: FE6F B888 F0E5 656C 1D25 3B9A 50C4 1384 F44A 4E42

12 Changelog

12.1 Version 2025-01-14

- Vollständige Aktualisierung der Daten
- LIZENZÄNDERUNG: Source Code jetzt unter GNU General Public License Version 3 (GPLv3) oder später lizenziert
- NEU: Zitationsnetzwerk des BFH von Aktenzeichen-zu-Aktenzeichen und Aktenzeichen-zu-BFHE als GraphML mit vielen Metadaten
- NEU: Option für Clean Runs in Konfiguration eingefügt (löscht alle Daten vor dem eigentlichen Run)
- NEU: Test auf geringen oder fehlenden Text-Inhalt
- NEU: Automatische Archivierung der Zwischenergebnisse in der Pipeline als ZIP-Archiv
- Docker Image auf R 4.4.0 aktualisiert (wegen CVE-2024-27322)
- Expliziter R Package Version Lock für 2024-06-13 (CRAN Date)
- Überarbeitung des Dockerfiles
- Überarbeitung der Dokumentation zu den Varianten des Datensatzes
- Vereinheitlichung der Komponenten für PDF-Extraktion, linguistische Statistiken und Berechnung kryptographischer Hashes
- Vereinfachung der Run-Skripte und stärkere Integration mit Docker Compose
- Erweiterung des Lösch-Skriptes
- Docker Zeitzone auf Berlin eingestellt
- Entfernung der Nummerierung von Diagrammen
- Entfernung der Tesseract Dependencies
- Entfernung der Python Toolchain
- Aktualisierung des Public GPG Keys im Repository

12.2 Version 2023-10-15

- Erstveröffentlichung

13 Parameter für strenge Replikationen

```
## [1] "OpenSSL 3.0.2 15 Mar 2022 (Library: OpenSSL 3.0.2 15 Mar 2022)"
```

```
## R version 4.4.0 (2024-04-24)
## Platform: x86_64-pc-linux-gnu
## Running under: Ubuntu 22.04.4 LTS
##
## Matrix products: default
## BLAS:   /usr/lib/x86_64-linux-gnu/openblas-pthread/libblas.so.3
## LAPACK: /usr/lib/x86_64-linux-gnu/openblas-pthread/libopenblasp-r0.3.20.so;
    LAPACK version 3.10.0
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## time zone: Europe/Berlin
## tzcode source: system (glibc)
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] future.apply_1.11.2 future_1.33.2       quanteda_4.0.2
##  [4] readtext_0.91       data.table_1.15.4   scales_1.3.0
##  [7] ggraph_2.2.1        igraph_2.0.3        ggplot2_3.5.1
## [10] pdftools_3.4.0      kableExtra_1.4.0    knitr_1.47
## [13] rvest_1.0.4         httr_1.4.7          zip_2.3.1
## [16] fs_1.6.4            testthat_3.2.1.1    RcppTOML_0.2.2
## [19] tarchetypes_0.9.0   targets_1.7.0
##
## loaded via a namespace (and not attached):
##  [1] tidyselect_1.2.1    viridisLite_0.4.2   dplyr_1.1.4
##  [4] farver_2.1.2        viridis_0.6.5       fastmap_1.2.0
##  [7] tweenr_2.0.3        stringfish_0.16.0   digest_0.6.35
## [10] base64url_1.4       lifecycle_1.0.4     secretbase_0.5.0
## [13] qpdf_1.3.3          processx_3.8.4      magrittr_2.0.3
## [16] compiler_4.4.0      rlang_1.1.4         tools_4.4.0
## [19] utf8_1.2.4          yaml_2.3.8          labeling_0.4.3
## [22] askpass_1.2.0       stopwords_2.3       graphlayouts_1.1.1
## [25] xml2_1.3.6          withr_3.0.0         purrr_1.0.2
## [28] grid_4.4.0          polyclip_1.10-6     fansi_1.0.6
## [31] colorspace_2.1-0    globals_0.16.3      MASS_7.3-60.2
## [34] cli_3.6.2           rmarkdown_2.27      generics_0.1.3
## [37] RcppParallel_5.1.7  rstudioapi_0.16.0   RApiSerialize_0.1.3
## [40] cachem_1.1.0        ggforce_0.4.2       stringr_1.5.1
## [43] parallel_4.4.0      vctrs_0.6.5         Matrix_1.7-0
## [46] callr_3.7.6         ggrepel_0.9.5       listenv_0.9.1
```

```
## [49] systemfonts_1.1.0   tidyr_1.3.1          glue_1.7.0
## [52] parallelly_1.37.1   codetools_0.2-20     ps_1.7.6
## [55] stringi_1.8.4       gtable_0.3.5         munsell_0.5.1
## [58] tibble_3.2.1        pillar_1.9.0         htmltools_0.5.8.1
## [61] brio_1.1.5          R6_2.5.1             tidygraph_1.3.1
## [64] evaluate_0.24.0     lattice_0.22-6       qs_0.26.3
## [67] backports_1.5.0     memoise_2.0.1        Rcpp_1.0.12
## [70] fastmatch_1.1-4     svglite_2.1.3        gridExtra_2.3
## [73] xfun_0.44           pkgconfig_2.0.3
```

Literaturverzeichnis

Allaire, JJ, Yihui Xie, Christophe Dervieux, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, et al. 2024. *Rmarkdown: Dynamic Documents for R*. https://github.com/rstudio/rmarkdown.

Barrett, Tyson, Matt Dowle, Arun Srinivasan, Jan Gorecki, Michael Chirico, and Toby Hocking. 2024. *Data.table: Extension of 'Data.frame'*. https://r-datatable.com.

Bengtsson, Henrik. 2021. "A Unifying Framework for Parallel and Distributed Processing in R Using Futures." *The R Journal* 13 (2): 208–27. https://doi.org/10.32614/RJ-2021-048.

———. 2024a. *Future.apply: Apply Function to Elements in Parallel Using Futures*. https://future.apply.futureverse.org.

———. 2024b. *Future: Unified Parallel and Distributed Processing in R for Everyone*. https://future.futureverse.org.

Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo. 2018. "Quanteda: An R Package for the Quantitative Analysis of Textual Data." *Journal of Open Source Software* 3 (30): 774. https://doi.org/10.21105/joss.00774.

Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, Akitaka Matsuo, and William Lowe. 2024. *Quanteda: Quantitative Analysis of Textual Data*. https://quanteda.io.

Csárdi, Gábor. 2024. *Zip: Cross-Platform Zip Compression*. https://github.com/r-lib/zip.

Csardi, Gabor, and Tamas Nepusz. 2006. "The Igraph Software Package for Complex Network Research." *InterJournal* Complex Systems: 1695. https://igraph.org.

Csárdi, Gábor, Tamás Nepusz, Vincent Traag, Szabolcs Horvát, Fabio Zanini, Daniel Noom, and Kirill Müller. 2024. *Igraph: Network Analysis and Visualization*. https://r.igraph.org/.

Eddelbuettel, Dirk. 2023. *RcppTOML: Rcpp Bindings to Parser for "Tom's Obvious Markup Language"*. http://dirk.eddelbuettel.com/code/rcpp.toml.html.

Gagolewski, Marek. 2022. "stringi: Fast and Portable Character String Processing in R." *Journal of Statistical Software* 103 (2): 1–59. https://doi.org/10.18637/jss.v103.i02.

Gagolewski, Marek, Bartek Tartanus, others; Unicode, Inc., and others. 2024. *Stringi: Fast and Portable Character String Processing Facilities*. https://stringi.gagolewski.com/.

Landau, William Michael. 2021a. *Tarchetypes: Archetypes for Targets*.

———. 2021b. "The Targets R Package: A Dynamic Make-Like Function-Oriented Pipeline Toolkit for Reproducibility and High-Performance Computing." *Journal of Open Source Software* 6 (57): 2959. https://doi.org/10.21105/joss.02959.

———. 2024a. *Tarchetypes: Archetypes for Targets*. https://docs.ropensci.org/tarchetypes/.

———. 2024b. *Targets: Dynamic Function-Oriented Make-Like Declarative Pipelines*. https://docs.ropensci.org/targets/.

Ooms, Jeroen. 2023. *Pdftools: Text Extraction, Rendering and Converting of Pdf Documents*. https://docs.ropensci.org/pdftools/.

Pedersen, Thomas Lin. 2024. *Ggraph: An Implementation of Grammar of Graphics for Graphs and Networks*. https://ggraph.data-imaginist.com.

Ushey, Kevin, and Hadley Wickham. 2024. *Renv: Project Environments*. https://rstudio.github.io/renv/.

Wickham, Hadley. 2024. *Rvest: Easily Harvest (Scrape) Web Pages*. https://rvest.tidyverse.org/.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC.

———. 2015. *Dynamic Documents with R and Knitr*. 2nd ed. Boca Raton, Florida: Chapman; Hall/CRC. https://yihui.org/knitr/.

———. 2024. *Knitr: A General-Purpose Package for Dynamic Report Generation in R*. https://yihui.org/knitr/.

Xie, Yihui, J. J. Allaire, and Garrett Grolemund. 2018. *R Markdown: The Definitive Guide*. Boca Raton, Florida: Chapman; Hall/CRC. https://bookdown.org/yihui/rmarkdown.

Xie, Yihui, Christophe Dervieux, and Emily Riederer. 2020. *R Markdown Cookbook*. Boca Raton, Florida: Chapman; Hall/CRC. https://bookdown.org/yihui/rmarkdown-cookbook.

Zhu, Hao. 2024. *KableExtra: Construct Complex Table with Kable and Pipe Syntax*. http://haozhu233.github.io/kableExtra/.