Regelungen für die Prüfung von Forschungsergebnissen auf Geheimhaltung und Datenvertraulichkeit („Outputprüfung“)

Version 2.2

Stand: Oktober 2018

Verfasser: Forschungsdaten- und Servicezentrum

Inhaltsverzeichnis

Vorwort

Liebe Datennutzerinnen und Datennutzer,

die Deutsche Bundesbank ermöglicht Forschenden einen transparenten und kostenfreien Zugang zu ausgewählten Mikrodatenbeständen. Zum einen eröffnet der Zugang zu diesen Mikrodaten den Forschenden eine erhebliche Steigerung des Analysepotentials. Zum anderen stellt die Arbeit mit diesen sensiblen Daten gegenüber der Arbeit mit aggregierten Daten deutlich höhere Anforderungen hinsichtlich der Wahrung des Datenschutzes und der Datenvertraulichkeit an die Forschenden.

Die Forschenden sind dafür verantwortlich, dass keine vertraulichen Daten an die Öffentlichkeit gelangen. Ihre veröffentlichten Forschungsergebnisse müssen absolut anonym sein. Sie dürfen keine Rückschlüsse auf einzelne Wirtschaftssubjekte wie Banken, Unternehmen, Personen oder Haushalte ermöglichen.

Die folgenden Kriterien und Regelungen sollen sicherstellen, dass diese Anforderungen an Forschungsergebnisse immer erfüllt sind. Darüber hinaus tragen sie dazu bei, das Vertrauen der Datengeber, wie z.B. Banken und Unternehmen, und der Produzenten der Mikrodatensätze, wie den Fachbereichen der Bundesbank, im sicheren Umgang mit ihren Daten zu stärken. Dieses Vertrauen in die Einhaltung des Datenschutzes und der Datenvertraulichkeit durch die Bundesbank und ihrer Datennutzerinnen und Datennutzer ist zwingend erforderlich für die zukünftige Bereitstellung dieses öffentlichen Gutes für die Forschung.

1 Einleitung

Forschende, welche mit Mikrodaten der Deutschen Bundesbank arbeiten, sind dafür verantwortlich, dass ihre Berechnungsergebnisse nicht die Datenvertraulichkeit verletzen. Sie haben dafür Sorge zu tragen, dass auf Basis ihrer Berechnungsergebnisse keine Rückschlüsse auf einzelne Merkmalsträger oder statistische Einheiten wie beispielsweise Banken, (nichtfinanzielle) Unternehmen, Personen oder Haushalte möglich sind. Das sichere Umfeld der Deutschen Bundesbank dürfen Ergebnisse nur absolut anonymisiert verlassen.

Aus diesem Grund muss von Mitarbeitern des Forschungsdaten- und Servicezentrums (FDSZ) oder eines anderen im Einzelfall für die Prüfung zuständigen Bereichs der Bundesbank zuvor geprüft werden, ob bei der Erstellung dieser Ergebnisse die Anforderungen an die Datenvertraulichkeit gewahrt wurden. Sollte dies nicht der Fall sein, dann können die Berechnungsergebnisse nicht freigegeben werden.

In Abschnitt 2 wird dargelegt, in welcher Form Berechnungsergebnisse dem FDSZ oder einem anderen im Einzelfall für die Prüfung zuständigen Bereich der Bundesbank zur hier beschriebenen Prüfung (Outputprüfung) vorgelegt werden müssen.

Des Weiteren bietet das Dokument einen Überblick, welche Kriterien bei der Erstellung der Berechnungsergebnisse von Forschenden zu beachten sind, damit die Datenvertraulichkeit gewährleistet ist. Hierbei wird zwischen allgemeinen und datensatzspezifischen Kriterien unterschieden. Allgemeine Kriterien müssen grundsätzlich beachtet werden, unabhängig davon, mit welchen Daten gearbeitet wird. Die allgemeinen Kriterien werden in Abschnitt 3 aufgeführt. Für bestimmte Einzeldaten gelten besondere Geheimhaltungsvorschriften. Aus diesem Grund sind bei der Arbeit mit solchen Daten zusätzliche Kriterien zu beachten. Diese werden im Abschnitt 4 aufgezeigt.

Bei den in diesem Dokument aufgeführten Kriterien handelt es sich nicht um ein abgeschlossenes Regelwerk. Die Kriterien und Regelungen können im Laufe der Zeit angepasst, ergänzt und erweitert werden. Darüber hinaus kann das FDSZ oder ein anderer im Einzelfall für die Prüfung zuständiger Bereich auch jetzt schon Berechnungsergebnisse hinsichtlich weiterer Kriterien prüfen, falls dies erforderlich ist (vgl. Abschnitt 5).

Schließlich wird im letzten Abschnitt auf eine vom FDSZ erstellte Prozedur verwiesen, die bei Nutzung der gängigen Analysesoftware Stata die für die Prüfung der Berechnungsergebnisse erforderlichen Fallzahlen ermittelt.

Sollten die in Abschnitt 2 beschriebenen Anforderungen oder die in den Abschnitten 3 und 4 genannten Kriterien nicht erfüllt sein, ist das FDSZ oder ein anderer im Einzelfall für die Prüfung zuständiger Bereich der Bundesbank dazu berechtigt, die Outputprüfung abzulehnen, die Freigabe der Berechnungsergebnisse zu verweigern und eine Nachbesserung einzufordern.

Falls Forschende unsicher sind, ob ihre Berechnungsergebnisse diesen Anforderungen und Kriterien entsprechen, wird empfohlen mit den Mitarbeiterinnen und Mitarbeitern des FDSZ mögliche kritische Punkte gemeinsam zu besprechen. Bei Gastforscherinnen und Gastforschern sollte dies bereits während des Forschungsaufenthalts erfolgen. Dieses Gespräch ersetzt jedoch nicht die Outputprüfung durch das FDSZ. Ebenso ist es weiterhin die Aufgabe und Verantwortlichkeit der Forschenden, bei der Erstellung der Berechnungsergebnisse sicherzustellen, dass die Datenvertraulichkeit nicht verletzt wird.

2 Anforderungen an Berechnungsergebnisse

2.1 Mindestanforderungen des FDSZ an Berechnungsergebnisse

Zu den Mindestanforderungen an Berechnungsergebnisse, die zum Zwecke der Outputprüfung vorgelegt werden, gehören:

- Ergebnisdateien (z.B. log-files in Stata, Nutzung von rdsc_log in R)
- hinreichend kommentierte Programmcodes, so dass diese nachvollziehbar sind
- alle für die Outputprüfung relevanten Informationen und Fallzahlen
- zugrunde liegende Häufigkeiten als Tabellen (insbesondere bei Grafiken)
- Ablage der Ergebnisse sowie Programmcodes im dafür vorgesehenen Projektverzeichnis (Unterordner „transfer“ im Projektordner, siehe hierzu auch Vertrag über den Zugang zu Einzeldaten im Rahmen eines Forschungsprojekts im Forschungsdaten- und Servicezentrum der Deutschen Bundesbank, Anlage 2, Punkt 9: Ordnerstruktur am Gastforscherarbeitsplatz)
- Ergebnisse und Programmcodes werden nicht zur Outputprüfung akzeptiert, wenn sie in softwarespezifischen Formaten vorgelegt werden, wie z.B. .Rdata. Stattdessen müssen „plain text files“ genutzt werden (beispielsweise .txt, .csv, .tsv, .R, .do, .log oder LaTeX-Dateien). Bitte beachten Sie, dass Dateien im .doc oder .docx-Format keine „plain text files“ sind.

2.2 Protokollierung von Berechnungsergebnissen

- Bis auf die Master-Datei soll jede Auswertungsdatei (für Stata beispielsweise: „do-file“) ein log-file erzeugen, in der die generierten Berechnungsergebnisse vollständig protokolliert werden.
- Neben log-files sind für deskriptive Tabellen und Regressionstabellen außerdem weitere „plain text files“ zulässig (z.B .csv oder LaTeX-Dateien). Aus dem log-file muss jedoch ersichtlich sein, an welcher Stelle diese Tabelle erzeugt und exportiert wurde.
- Die log-files müssen editierbar sein, damit bei der Prüfung ggf. kritische Ergebnisse hinsichtlich der Datenvertraulichkeit gesperrt werden können (beispielsweise sind .smcl-Dateien aus Stata unzulässig).
- Für die Prüfung hinsichtlich der Geheimhaltung von einzelnen Merkmalsträgern (vgl. hierzu Abschnitt 2.5) ist es erforderlich, dass für jegliche Form der Ergebnisdarstellung (z.B. Tabellen, Grafiken, Regressionen, deskriptive Statistiken) die zugrundeliegenden Fallzahlen immer mit angegeben werden.[1] Ist dies nicht der Fall, dann werden die Ergebnisse nicht freigegeben. Sie müssen beim nächsten Aufenthalt überarbeitet und einschließlich der Fallzahlen neu erzeugt werden.

2.3 Nutzung von Microsoft Excel

An allen Gastforscherarbeitsplätzen ist die Software Microsoft Excel installiert. Bitte beachten Sie, dass mit Microsoft Excel erzeugte Berechnungsergebnisse nicht zur Outputprüfung akzeptiert und nicht freigegeben werden. Ebenso werden in ein Excel-Dokument (.xls- oder .xlsx-Datei) exportierte Berechnungsergebnisse nicht zur Outputprüfung akzeptiert und nicht freigegeben.

[1] Für die Anforderungen an die Fallzahlen vgl. Abschnitt 3.1.

2.4 Eindeutigkeit von Variablennamen und Wertelabels

Variablennamen und Labels der Ausprägungen innerhalb einer Variablen sind eindeutig zu vergeben. Dies gilt besonders für die Variablennamen. Wird eine Variable neu erzeugt oder eine bestehende Variable verändert, dann sind auch die zugehörigen Variablenbezeichnungen ausnahmslos neu zu vergeben und in der Variablenliste im Programmkopf der Syntax aufzuführen. Dabei ist auf „sprechende Namen“ zu achten. Werden (kategoriale) Variablen neu erzeugt oder verändert, dann müssen für die Ausprägungen entsprechende Wertelabels vergeben werden.

2.5 Prüfung der Berechnungsergebnisse

Berechnungsergebnisse sind dem FDSZ oder einem anderen im Einzelfall für die Prüfung zuständigen Bereich der Bundesbank zur Outputprüfung vorzulegen. Berechnungsergebnisse dürfen die Räume des FDSZ oder vergleichbar ausgestattete Räume der Bundesbank erst verlassen, wenn das FDSZ oder ein anderer im Einzelfall für die Prüfung zuständiger Bereich der Bundesbank dem schriftlich zugestimmt hat.

Falls nach einem Forschungsaufenthalt Berechnungsergebnisse durch das FDSZ geprüft und freigegeben werden sollen, so sind die Berechnungsergebnisse und alle für die Prüfung erforderlichen Dateien im dafür vorgesehenen Projektverzeichnis abzulegen (vgl. Abschnitt 2.1). Das FDSZ kontrolliert nicht von sich aus, ob nach einem Gastaufenthalt im Projektverzeichnis Berechnungsergebnisse zur Prüfung vorliegen. Es ist deshalb per E-Mail darüber zu informieren, dass Berechnungsergebnisse geprüft werden sollen und in welchem Ordner diese zu finden sind. Bitte senden Sie aus Geheimhaltungsgründen keine Berechnungsergebnisse per E-Mail an das FDSZ!

Neben den Berechnungsergebnissen sind auch die entsprechenden Programmcodes in diesem Ordner abzulegen. Dies ist erforderlich, damit kritische Berechnungsergebnisse rekapituliert werden können. Ist dies anhand der abgelegten Programmcodes nicht möglich, können die Berechnungsergebnisse nicht freigegeben werden. Falls diese Programmcodes ebenfalls außerhalb der Bundesbank benötigt werden, ist dies dem FDSZ mitzuteilen. Ansonsten werden nur die Berechnungsergebnisse versandt.

Forschende haben dafür zu sorgen, dass die unter Abschnitt 2.1 beschriebenen Mindestanforderungen erfüllt werden und der Prüfaufwand seitens des FDSZ oder eines anderen im Einzelfall für die Prüfung zuständigen Bereichs der Bundesbank soweit wie möglich minimiert wird.

Es ist nicht vorgesehen, dass Forschende zunächst sämtliche Berechnungsergebnisse zur Prüfung vorlegen und unter diesen erst im Nachhinein diejenigen auswählen, die für eine Veröffentlichung (z.B. Vortrag, Arbeitspapier, Zeitschriftenartikel) geeignet sind. Grundsätzlich sollen sämtliche Arbeiten innerhalb des Forschungsprojekts, die einen Bezug zu den Mikrodaten der Bundesbank haben, während des Forschungsaufenthalts in den Räumen des FDSZ oder vergleichbar ausgestatteten Räumen der Bundesbank durchgeführt werden. Dies umfasst neben der Datenaufbereitung sowie der Analyse der Daten auch die Auswahl der für eine Veröffentlichung benötigten Berechnungsergebnisse.

Der Umfang der zur Prüfung vorgelegten Berechnungsergebnisse soll eine sinnvolle und nachvollziehbare Größe nicht überschreiten. Richtlinie sollte hierbei die Menge an Berechnungsergebnissen sein, die für die Erstellung einer Veröffentlichung benötigt werden und dort Platz finden. Die Auswahl der zur Prüfung vorgelegten Berechnungsergebnisse ist auf diejenigen zu beschränken, welche zur finalen Publikation geeignet erscheinen. Somit sind von Forscherinnen und Forschern umfangreiche Berechnungsergebnisse, wie sie insbesondere im Rahmen der explorativen Analysen oder der Prüfung verschiedener Modellspezifikationen bei Regressionen anfallen, be-

reits während des Forschungsaufenthalts durchzusehen und sorgfältig die publikationswürdigen zur Prüfung auszuwählen.

Berechnungsergebnisse sollten durch das FDSZ lediglich einmalig geprüft und freigegeben werden müssen. Dies dient dazu, den Prüfungsaufwand zu reduzieren. Aus diesem Grund dürfen identische Berechnungsergebnisse nicht wiederholt zur Outputprüfung vorgelegt werden. Falls Berechnungsergebnisse in Ausnahmefällen doch erneut erstellt und zur Outputprüfung vorgelegt werden, dann nur mit Begründung und Verweis auf die entsprechenden früheren Auswertungen.

3 Allgemeine Kriterien zur Prüfung der Berechnungsergebnisse

Bei Analysen und Berechnungen mit Mikrodaten der Deutschen Bundesbank sind die in diesem Abschnitt aufgeführten Kriterien immer zu beachten. Diese Kriterien gelten unabhängig davon, welche Datensätze in einem Forschungsprojekt genutzt werden.

3.1 Fallzahlen

Sämtliche Berechnungsergebnisse, die zur Prüfung vorgelegt werden, müssen auf mindestens drei unterschiedlichen Wirtschaftssubjekten (Berichtspflichtige oder andere juristische bzw. natürliche Personen, Rechtssubjekte oder Niederlassungen, wie Banken, (nichtfinanzielle) Unternehmen oder Haushalte) basieren. Forschende müssen bereits bei der Erstellung ihrer Berechnungsergebnisse darauf achten, dass diese Bedingung erfüllt ist.

Daher müssen grundsätzlich für **alle** Ergebnisse die zugrundeliegenden Fallzahlen (d.h. Anzahl der in das Ergebnis eingehenden verschiedenen Merkmalsträger) kontrolliert und ausgewiesen werden. Bei gewichteten Berechnungsergebnissen sind die ungewichteten Fallzahlen ebenso einzubeziehen. In der Regel werden alle Ergebnisse sowie die dazugehörigen Fallzahlen gesperrt, wenn die Ergebnisse auf weniger als drei Wirtschaftssubjekten basieren.

Bei der Ermittlung der relevanten Fallzahlen ist zu berücksichtigen, dass die Anzahl der unterschiedlichen Merkmalsträger entscheidend ist, die den Ergebnissen zugrunde liegen, und nicht die Anzahl der Beobachtungen. Die Merkmalsträger sind die „zu schützenden“ Einheiten. In der Regel ist die Zahl der zugrundeliegenden Merkmalsträger kleiner als die Zahl der Beobachtungen in einem Datensatz. Es ist daher darauf zu achten, dass Merkmalsträger nicht mehrfach gezählt werden, wenn die Fallzahlen ermittelt werden.

Wiederholte Beobachtungen für einzelne Merkmalsträger sind ein wesentlicher Bestandteil der meisten Mikrodaten der Bundesbank. Dadurch ergeben sich viele Beobachtungen für einen Merkmalsträger. So werden beispielsweise für ein und dieselbe Bank oder ein und dasselbe Unternehmen verschiedene Transaktionen mit unterschiedlichen Partnern zu mehreren Zeitpunkten ausgewiesen. Im vorangegangenen Beispiel dürfte der Merkmalsträger die Bank oder das nichtfinanzielle Unternehmen sein. Bei der Ermittlung der Fallzahlen darf jeder Merkmalsträger nur einmal gezählt werden. Es muss also in diesem Fall die Anzahl unterschiedlicher Banken bzw. die Anzahl unterschiedlicher Unternehmen überprüft und ausgewiesen werden, die den Ergebnissen zugrunde liegen. Die Gesamtzahl der Beobachtungen im Datensatz reicht nicht aus.

In einigen Datensätzen werden fehlende Werte grundsätzlich mit Null belegt. Deshalb sind in solchen Datensätzen Nullen zunächst als nicht valide Beobachtungen anzusehen. Es muss dann im Einzelfall überlegt werden, wie

die Nullen zu interpretieren sind. Das heißt, wenn einem Wert beispielsweise Zahlen von fünf Merkmalsträgern zugrunde liegen, könnte ein Problem vorliegen, wenn drei der Werte Null betragen, obwohl es sich dabei eigentlich um fehlende Werte handelt und somit nur Werte von zwei Merkmalsträgern in das Ergebnis eingehen.

3.2 Berechnung der einem Aggregat zugrundeliegenden Fallzahlen

Für jede mittels Aggregation erzeugte Variable muss die Anzahl der dem Aggregat zugrundeliegenden Merkmalsträger, für die gültige Einzelwerte vorliegen, ausgewiesen werden.

Bei der Bildung von Quoten oder Anteilen ist sicherzustellen, dass für jede in die Berechnung eingehende (aggregierte) Variable die Zahl der zugrundeliegenden Merkmalsträger (für die gültige Einzelwerte vorliegen) ausreichend hoch ist. Das bedeutet, dass die Fallzahlen nicht nur für den Nenner, sondern auch für den Zähler geprüft werden müssen.

Einige Datensätze liefern Informationen zu nach- und übergeordneten (wirtschaftlichen) Einheiten (Beispiele: Aktien sowie Derivate mit übergeordneter Einheit Aktiengesellschaften; Tochterunternehmen mit übergeordneter Einheit Mutterunternehmen). Werden im Rahmen der Analyse solcher Datensätze Werte zu Einheiten ausgegeben, die sich zum Teil den gleichen übergeordneten Einheiten zuordnen lassen, so ist hier immer die Anzahl der Merkmalsträger auf der jeweils obersten Ebene entscheidend. So sind beispielsweise nicht nur einzelne Unternehmen als zu schützende Einheit zu betrachten, sondern auch der Konzern, dem diese Unternehmen angehören. Beziehen sich Werte auf eine Reihe von Unternehmen, die alle einem Konzern angehören, dann würden diese Werte möglicherweise einer Einzelangabe dieses Konzerns entsprechen. Daher ist bei der Erstellung der Berechnungsergebnisse stets auf die entsprechend zu schützende Einheit zu achten.

Bitte beachten Sie, dass auch das Fallenlassen von Datenpunkten mit bestimmten Werten (etwa mit fehlenden Werten in Stata) zur Ausgabe einer Aggregatinformation (über die Gruppe der Fälle mit diesem Wert) führen kann.

3.3 Grafiken

Bei jeder Grafik (etwa einer relativen Häufigkeitsfunktion) müssen für jeden Datenpunkt in der Grafik die Fallzahlen kontrolliert und ausgewiesen werden, z.B. indem man eine entsprechende Tabelle (etwa eine Tabelle mit absoluten Häufigkeiten) generiert, anhand der man die Fallzahlen einzelner Klassen oder Untergruppen unmittelbar erkennt. Das bedeutet beispielsweise im Fall einer Grafik mit der Summe der Bankaktiva für die Monate Januar 2012 bis Dezember 2015, dass für jeden einzelnen Monat die Zahl der Banken ausgewiesen werden muss, auf denen die Berechnung der Summe der Bankaktiva basiert. Wenn diese Tabellen nicht vorliegen, dann können die Grafiken nicht freigegeben werden.

Grafiken sind in einem nicht weiter zu verarbeitenden Format abzuspeichern. Dies soll verhindern, dass dahinterliegende Werte- oder Fallzahltabellen verschickt werden. Aus diesem Grund dürfen Grafiken nicht als Vektorgrafiken zur Prüfung vorgelegt werden. Empfohlenes Datei-Format für Stata-Grafiken ist .png.

3.4 Fallzahlenausgabe bei Ausgabe von Quantilen

Werden Quantile (q) einer Grundgesamtheit oder Untergruppe berechnet, so muss zugleich die Fallzahl (n) der entsprechenden Gruppe berechnet werden und die auf dieser Basis errechnete Anzahl der Fälle pro Quantilsbe-

reich (beispielsweise beträgt für den Fall eines 99 %-Quantils (q=99) und 1.000 Fällen die Zahl der Fälle pro Quantilsbereich (1000 x (1-0,99) = 10)). In der Praxis ergibt sich ein Quantilswert aufgrund von Nichtteilbarkeiten häufig als gewichtetes Mittel zweier benachbarter Werte. Mathematische und praktische Erwägungen haben gezeigt, dass ein Wert von 2,3 als geeignete Grenze für die Anzahl der Fälle pro Quantilsbereich angesehen werden kann. Deshalb wären beispielsweise Angaben zum Median für die Publikation zu sperren, wenn für:

$$q' = \begin{cases} q, & q \leq 50 \\ 100 - q, & q > 50 \end{cases}$$

$$(n+1)\frac{q'}{100} \leq 2{,}3$$

beziehungsweise

$$(n+1)q' \leq 230$$

gilt.

3.5 Tabellenübergreifende Geheimhaltung

Sofern Berechnungsergebnisse zunächst als Gesamtergebnisse und die Einzeldaten anschließend nach bestimmten Kriterien unterteilt werden, muss die tabellenübergreifende Geheimhaltung beachtet werden. Wird beispielsweise eine „alle Unternehmen“-Tabelle und eine Tabelle „Maschinenbauunternehmen“ erzeugt, ist die Differenz „Nicht-Maschinenbauunternehmen“. Forschende müssen in einem solchen Fall die Tabelle „Nicht-Maschinenbauunternehmen“ immer mit erzeugen, da diese ebenfalls bei der Geheimhaltungsprüfung berücksichtigt werden; über Differenzbildung könnten sonst Ausprägungen einzelner Merkmalsträger identifiziert werden.

3.6 p%-/Dominanzregel

Merkmalsträger unterscheiden sich hinsichtlich ihrer Größe. Auch wenn ein Wert basierend auf drei oder mehr Merkmalsträgern ausgewiesen wird, dann besteht die Gefahr einer indirekten Identifizierung großer Merkmalsträger, falls diese einen ausreichend hohen Anteil zu diesem Wert beisteuern (Dominanzkriterium). Die Möglichkeit einer solchen indirekten Identifizierung muss ausgeschlossen sein. Wenn unter den Merkmalsträgern, die der Berechnung eines Werts zugrunde liegen, ein oder zwei Merkmalsträger zu finden sind, die im Vergleich zu den übrigen Merkmalsträgern deutlich größer sind, dann muss geprüft werden, ob diese einen dominanten Beitrag zu diesem Wert leisten. Neben der Anzahl der zugrundeliegenden Merkmalsträger ist in einem solchen Fall stets der größte und zweitgrößte Wert auszuweisen. Als Faustregel gilt: Der Anteil des größten und zweitgrößten Werts am Gesamtwert sollte zusammen 85 % nicht übersteigen.

3.7 Minima und Maxima

Minima und Maxima sind grundsätzlich Einzeldaten und somit geheim zu halten. Ausnahme bilden dichotome Merkmale, sofern beide Werte besetzt sind.

Es ist aber möglich, alternativ zu Minima und Maxima einer Variablen den Durchschnittswert der drei Merkmalsträger mit dem niedrigsten und der drei Merkmalsträger mit dem höchsten Wert auszuweisen. Bei der Berechnung ist darauf zu achten, dass diese insgesamt sechs Merkmalsträger unterschiedlich sind.

3.8 Dummy-Variablen

Wird der Mittelwert einer mit 0 und 1 kodierten Dummy-Variablen ausgegeben, dann ist zu berücksichtigen, dass dieser Wert einer Quote entspricht. Sie gibt den Anteil der Beobachtungen mit der Ausprägung 1 an allen Beobachtungen an. Ein Mittelwert von 0,7 bei einer solchen Dummy-Variable bedeutet, dass in 70 % der Fälle die Variable die Ausprägung 1 hat und in 30 % der Fälle die Ausprägung 0. Daher gelten bei der Ausgabe von Mittelwerten für eine Dummy-Variable die gleichen Regeln wie bei der Berechnung von Quoten (vgl. Abschnitt 3.2). Es müssen mindestens drei unterschiedliche Merkmalsträger die Ausprägung 0 haben und mindestens drei unterschiedliche Merkmalsträger die Ausprägung 1.

Dummy-Variablen können in Regressionen aufgenommen werden, auch wenn weniger als drei Beobachtungseinheiten die Ausprägung 0 bzw. 1 haben. In diesem Fall werden aber die Regressionskoeffizienten für diese Dummy-Variablen nicht freigegeben und gesperrt. Die übrigen Regressionskoeffizienten können hingegen freigegeben werden. Daher muss für jede Dummy-Variable in einer Regression geprüft werden, ob die Variable für eine ausreichende Anzahl von Merkmalsträgern die Ausprägung 0 bzw. die Ausprägung 1 hat. Überprüft werden kann dies in Stata unmittelbar im Anschluss an die Regression mit dem Befehl „`nobsreg`“ (vgl. Abschnitt 6). Diese Information ist nur für Dummy-Variablen erforderlich, die sich direkt auf die Merkmalsträger beziehen. Ansonsten werden diese Informationen nicht benötigt.

3.9 Identifikatoren

Weder in den do-files noch in den Berechnungsergebnissen, die zur Prüfung vorgelegt werden, dürfen Identifikatoren enthalten sein (z.B. BAID, Unternehmensnummern etc.).[2]

3.10 Vorherige Ergebnisse

Neben den vorgelegten Berechnungsergebnissen werden bei der Geheimhaltungsprüfung auch Berechnungsergebnisse herangezogen, die Forschende bereits zuvor erhalten haben. Da die Zusammenschau den Informationswert der zur Prüfung vorgelegten Berechnungsergebnisse erhöhen kann, kann sich hieraus die Notwendigkeit zusätzlicher Sperrungen ergeben.

3.11 Analysen auf regionaler Ebene

Bei der Arbeit mit regionalen Informationen oder einer Analyse auf regionaler Ebene ist zu beachten, dass regionale Informationen das Reidentifikationsrisiko erhöhen können. Dadurch ergeben sich besondere Anforderungen an die Datenvertraulichkeit und die Outputprüfung. Auswertungen und insbesondere deskriptive Analysen für

[2] Für den Umgang mit ISINs vgl. Abschnitt 4.2.

einzelne Regionen werden grundsätzlich nicht zur Outputprüfung akzeptiert, wenn die regionalen Einheiten kleinräumiger als Bundesländer (NUTS1) sind. Dies gilt unabhängig davon, ob für die regionale Einheit drei oder mehr Merkmalsträger vorliegen. Berechnungsergebnisse können nur freigegeben werden, wenn sich kein Bezug zu einer bestimmten Region herstellen lässt bzw. wenn aus den Ergebnissen nicht erkennbar ist, welche Region betrachtet wird. Auch in diesem Fall gilt weiterhin, dass die Zahl der zugrundeliegenden Merkmalsträger ausreichend hoch sein muss und das Dominanzkriterium nicht verletzt sein darf.

4 Zusätzliche Kriterien zur Prüfung der Berechnungsergebnisse für bestimmte Datensätze

Für die Arbeit mit den meisten Datensätzen ist es ausreichend, wenn die allgemeinen Kriterien beachtet werden. Jedoch gelten für manche Datensätze besondere Geheimhaltungsrichtlinien. In solch einem Fall müssen zusätzlich zu den in Abschnitt 3 beschriebenen allgemeinen Kriterien die für diesen Datensatz spezifischen Kriterien beachtet werden. Diese werden im Folgenden für die einzelnen Datensätze dargestellt.

4.1 Auslandsstatus der Banken (MFIs) (AUSTA)

- Eine Bank (Mutter) kann mehrere rechtlich unselbständige Zweigniederlassungen (Auslandsfilialen) und/oder rechtlich selbständige Tochterbanken (Auslandstöchter) unterhalten. Daher muss bei der Analyse auf der Ebene der Auslandsniederlassungen immer die Anzahl der zugrundeliegenden Mütter geprüft werden (vgl. auch Abschnitt 3.2).
- In vielen Ländern sind nur wenige Mütter und Töchter vertreten. Werden für diese Länder aggregierte Werte ausgewiesen, dann kann der Anteil einer Bank sehr hoch ausfallen. Bei begründetem Verdacht muss immer das Dominanzkriterium überprüft werden, wie in Abschnitt 3.6 beschrieben.

4.2 Statistik über Wertpapierinvestments (SHS-Base plus)

- Die schützenswerte Einheit ist das meldepflichtige Institut und nicht das Wertpapier. D.h. Auswertungen nach Wertpapieren sind grundsätzlich unproblematisch. Auch ist es unproblematisch, wenn ISINs in den Berechnungsergebnissen ausgewiesen werden.
- Kritisch sind Berechnungsergebnisse, die eine Kombination von Deponentensektor 1221 / 1222 / 1223, ISIN und Bank enthalten. Hierbei handelt es sich um Eigenbestände von meldepflichtigen Instituten an eigenen Wertpapieren. Es besteht die Möglichkeit anhand der ISIN herauszufinden, wer das Wertpapier emittiert hat. Dadurch ist direkt ersichtlich, um welche Bank es sich hierbei handelt.
- Es besteht die Möglichkeit, über die ISIN Informationen zu den Wertpapieren aus weiteren bzw. externen Quellen zuzuspielen. Dafür kann Gastforscherinnen und Gastforschern eine Liste der im Datensatz enthaltenen ISINs zur Verfügung gestellt werden. Aufgrund der unter Punkt 2 beschriebenen Problematik ist ein Zuspiel von Wertpapiernamen nicht erlaubt.

4.3 Daten der Millionenkreditevidenz

- In den Daten der Millionenkreditevidenz werden Kreditbeziehungen zwischen Kreditgeber und Kreditnehmer ausgewiesen. Jeder ausgewiesene Wert muss sich daher auf mindestens drei unterschiedliche Kreditgeber als auch auf drei unterschiedliche Kreditnehmer beziehen.
- Wenn aus den Einzeldaten aggregierte Werte berechnet werden, dann muss ebenfalls die Zahl der zugrundeliegenden Kreditgeber und Kreditnehmer ausgewiesen werden.

- Die Millionenkreditevidenz enthält auch Informationen zu übergeordneten Kreditgebern (Kreditgeberkonzern) und Kreditnehmern (Kreditnehmereinheit). Diese sind ebenfalls als zu schützende Einheit zu betrachten.

4.4 Mikrodatenbank Direktinvestitionen (MiDi)

- Die MiDi ist ein Datensatz, der Investitionsbeziehungen abbildet. D.h. jede Zeile in der MiDi entspricht einer in einem bestimmten Jahr gemeldeten Investitionsbeziehung. Daher reicht es in keinem Fall, nur die Anzahl der Zeilen zu zählen, auch nicht wenn eine Investitionsbeziehung nur in einem Jahr existiert. Denn in der Regel gehören mehrere Zeilen zu einem Unternehmen (sämtliche Investitionsbeziehungen des Unternehmens oberhalb der Meldegrenzen). Bei der Ermittlung der Fallzahlen ist darauf zu achten, dass jedes Unternehmen nur einmal gezählt wird, auch wenn es mehrere Investitionen tätigt, und es nicht zu einer Mehrfachzählung kommt.
- Implizit enthält die MiDi auch Informationen zu Konzernstrukturen und damit Informationen zu übergeordneten Unternehmen. Die übergeordneten Unternehmen sind ebenfalls als zu schützende Einheit zu betrachten. Dies betrifft etwa im Falle eines Outward-Investments den deutschen Melder (die „num“), sowie ggf. die deutsche Mutter des Melders (die „nui“). Das bedeutet, es reicht nicht, wenn ein berechnetes Aggregat auf mehr als drei ausländischen Tochterunternehmen beruht. Stattdessen muss hier in einem ersten Schritt die Anzahl der deutschen Melder (num), welche hinter dem Aggregat stehen, ausgezählt und als Fallzahl ausgegeben werden. Falls für deutsche Melder ein Mutterunternehmen (nui) vorhanden ist, dann muss für diese Fälle „nui“ anstelle von „num“ in die Berechnung der Fallzahl einbezogen werden.[3] Bei einem Inward-Investment sind entsprechend zu schützen: num (deutscher Melder), nu4 (ausländische Mutter), noa (ggf. Konzernmutter der ausländischen Mutter).
- In vielen Ländern sind nur wenige Mütter und Töchter vertreten. Werden für diese Länder aggregierte Werte ausgewiesen, dann kann der Anteil einer zu schützenden Unternehmenseinheit (beispielsweise im Fall eines Outward-Investments: einer deutschen Mutter) sehr hoch ausfallen. Ein Anteil von mehr als 85 % einer einzelnen Einheit am Aggregat kommt deshalb durchaus vor und muss bei begründetem Verdacht untersucht werden. Dies macht es erforderlich, das Reidentifikationsrisiko hinsichtlich des Dominanzkriteriums zu prüfen. Daher müssen neben dem aggregierten Wert auch die beiden höchsten Werte ausgewiesen werden, die in den aggregierten Wert eingehen.

[3] Praktisch lässt sich dies z.B. so durchführen, dass zuerst eine Hilfsvariable gebildet wird, welche den Wert der „num“ bekommt. Falls eine „nui“ vorhanden ist, wird dann diese Hilfsvariable mit dem Wert der „nui“ überschrieben. Anschließend wird diese Hilfsvariable für die Fallauszählungen anstatt der „num“ verwendet. Bei einem Paneldatensatz muss zusätzlich beachtet werden, dass ein Melder im Laufe der Zeit unterschiedliche Mütter haben kann. Ein deutlicher Hinweis hierfür ist, wenn im Datensatz die Anzahl der unterschiedlichen Mütter höher ausfällt als die Anzahl der unterschiedlichen Melder. In diesem Fall liefert die Hilfsvariable irreführend Ergebnisse. Dem kann z.B. dadurch Rechnung getragen werden, indem die Auszählung der Fallzahlen mit Hilfe der Hilfsvariablen für jede einzelne Zeiteinheit durchgeführt wird.

4.5 Statistik zum internationalen Dienstleistungshandel (SITS)

- Der Datensatz SITS bezieht sich auf Dienstleistungstransaktionen. D.h. jede Zeile bildet eine für ein bestimmtes Datum (Jahr und Monat) gemeldete Transaktion ab. Da ein Unternehmen in der Regel mehrere Transaktionen tätigt, können dadurch mehrere Zeilen Beobachtungen zu ein und demselben Unternehmen beinhalten. Die Anzahl der Zeilen mit ein und demselben Monat entspricht daher nicht der Zahl der Unternehmen. Bei der Ermittlung der Fallzahlen ist darauf zu achten, dass es nicht zu Mehrfachzählungen kommt, sondern jedes Unternehmen nur einmal gezählt wird, auch wenn es mehrere Transaktionen tätigt.
- In vielen Ländern tätigen nur wenige Unternehmen Dienstleistungstransaktionen. Werden für diese Länder aggregierte Werte ausgewiesen, dann kann der Anteil eines Unternehmens sehr hoch ausfallen. Ein Anteil von 85 % oder darüber einer einzelnen Einheit am Aggregat kommt deshalb durchaus vor und muss bei begründetem Verdacht untersucht werden. Dies macht es erforderlich, das Reidentifikationsrisiko hinsichtlich des Dominanzkriteriums zu prüfen. Daher müssen neben dem aggregierten Wert auch die beiden höchsten Werte ausgewiesen werden, die in den aggregierten Wert eingehen.

4.6 Unternehmensbilanzen aus dem Refinanzierungsgeschäft (Ustan)

- Manche Unternehmen melden mehrere Abschlüsse in einem Jahr. Bei der Ermittlung der Fallzahlen ist daher darauf zu achten, dass solche Unternehmen nur einmal gezählt werden und es nicht zu Mehrfachzählungen kommt.

5 Weitere Kriterien

Das FDSZ oder ein anderer prüfender Bereich der Bundesbank kann die Berechnungsergebnisse auch hinsichtlich weiterer Kriterien prüfen, soweit dies erforderlich erscheint.

6 Ermittlung der relevanten Fallzahlen in Stata

Fallzahlen lassen sich mit Stata auf unterschiedliche Weise ermitteln. Das FDSZ hat hierfür zwei Stata-Befehle `nobsreg` und `nobsdes` erstellt. `nobsreg.ado` und `nobsdes.ado` sind mit den zugehörigen .hlp-files in der ado_library und auf der Webseite des FDSZ verfügbar.[4] Falls Sie `nobsreg` oder `nobsdes` verwenden möchten, schauen Sie sich bitte die .hlp-files mit ausführlichen Beispielen an; hier werden die Prozeduren nur skizziert.

`nobsreg` ohne weitere Angabe direkt im Anschluss an einen Regressionsbefehl berichtet die Anzahl der unterschiedlichen Identifikatoren (IDs), die der Regression zugrunde liegen. Zudem kontrolliert es, ob eine Variable nur für ein oder zwei Identifikatoren Werte ungleich Null aufweist. Falls ja, wird eine Warnmeldung ausgegeben. `nobsreg` versucht also, sparsam im Output zu sein.

Im deskriptiven Fall, dazu gehören auch Schaubilder, ermittelt `nobsdes ID Variable [if] [, options]` für eine Variable die Anzahl unterschiedlicher IDs mit validen Beobachtungen und kontrolliert, ob das Dominanzkriterium verletzt wird. Benötigt wird der Name der ID und der interessierenden Variablen, in dieser Reihenfolge. Außerdem gibt es mehrere Optionen. Werden Kategorien mit `by()` benannt, wird als Standard eine

[4] https://www.bundesbank.de/resource/blob/604786/c9c14eade63b7fb3f85f38233bbacaa8/mL/number-of-observations-data.zip

Tabelle mit der Anzahl der unterschiedlichen IDs mit validen Beobachtungen angezeigt, ohne dass eine zusätzliche Variable erzeugt wird. Es kann für den weiteren Gebrauch aber auch eine Variable mit Nullen und Einsen ausgegeben werden, deren Summe die Anzahl unterschiedlicher IDs mit validen Beobachtungen angibt.