Forschungsdatenzentrum
der Bundesagentur für Arbeit
im Institut für Arbeitsmarkt-
und Berufsforschung

Vorgaben des FDZ der BA im IAB zur Nutzung von Datenfernverarbeitung und Gastaufenthalten

Zur effizienten Bearbeitung der Anfragen ist eine Standardisierung der Auswertungen erforderlich. Hierdurch wird ebenfalls sichergestellt, dass am FDZ jede Anfrage zügig bearbeitet werden kann. Für Sie kann dies u.U. ein Abweichen vom gewohnten Programmierstil bedeuten.

1 Allgemeines

Senden Sie Dateien für Gastaufenthalte und Aufträge zur Datenfernverarbeitung immer an das FDZ-Postfach iab.fdz@iab.de. Standardmäßig bietet das FDZ Analysen mit den Softwareprogrammen Stata und SPSS an. Wenn Sie andere Statistikprogramme für Ihre Auswertungen benötigen, setzten Sie sich bitte mit dem FDZ in Verbindung.

Kenngrößen auf aggregiertem Niveau (z.B. Arbeitslosenquoten nach Kreisen) dürfen den Daten zugespielt werden. Die entsprechenden Kenngrößen müssen dem FDZ mit einer Beschreibung der Merkmale, der Aggregationsebene und der Quellenangabe übermittelt werden.

Ado-Dateien für Stata müssen ebenfalls dem FDZ übermittelt werden. Wie die Ado-Dateien in die Programme eingebunden werden müssen, ist im Beispiel auf Seite fünf ersichtlich.

Es kommt vor, dass E-Mails von Firewall oder SPAM-Filter abgefangen werden, ohne dass wir Notiz davon erhalten. Daher bestätigen wir den Eingang jedes Auftrags per E-Mail. Falls Sie innerhalb von zwei Werktagen keine Eingangsbestätigung erhalten haben, geben Sie uns bitte Bescheid.

Vor der Übersendung der Ergebnisse nach Gastaufenthalt oder Datenfernverarbeitung erfolgt eine Datenschutzprüfung. Oberstes Gebot ist dabei die Vermeidung von Reidentifikationsrisiken für Personen und Betriebe. Zunächst wird geprüft, ob unsere Dokumentationsvorgaben eingehalten wurden (siehe Abschnitt 5). Anschießend werden die Inhalte der Ergebnisdateien betrachtet und gegebenenfalls einzelne Ergebnisse gelöscht. Der Nutzer bzw. die Nutzerin verpflichtet sich, bei wiederholten Auswertungen keine Rückrechnungen zu vorherigen Auswertungen vorzunehmen, um gelöschte Werte zu ermitteln und Personen bzw. Betriebe zu reidentifizieren. Veröffentlichte Werte müssen in den zugesandten Ergebnisdateien enthalten sein. Bei Verstoß wird der Nutzer bzw. die Nutzerin für die Dauer von zwei Jahren von jeder weiteren Inanspruchnahme der

Ferndatenrechnung, von Datenlieferungen und Gastaufenthalten am FDZ der Bundesagentur für Arbeit ausgeschlossen. Informationen über Verstöße gegen die o.g. Verpflichtungen werden an andere Forschungsdatenzentren weitergegeben.

Wir bemühen uns, Ihre Ergebnisse möglichst bald, spätestens nach 5 Arbeitstagen zurückzuschicken. Sollte es einmal länger als 5 Arbeitstage dauern, werden wir Sie informieren.

2 Gastaufenthalte

Wir empfehlen, dass Sie Ihre Auswertungsprogramme anhand unserer Testdaten vor dem Gastaufenthalt vorbereiten. Wenn keine Testdaten vorliegen und der von Ihnen verwendete Datensatz auch als Scientific Use File vorliegt, empfehlen wir vor dem Gastaufenthalt mit dem SUF zu arbeiten. Die für den Gastaufenthalt vorbereiteten Auswertungsprogramme (und ggf. weitere im FDZ benötigte Dateien) sollten spätestens drei Arbeitstage vor dem ersten Aufenthaltstag an uns geschickt werden. Bei später übermittelten oder erst zum Gastaufenthalt mitgebrachten Dateien können wir nicht gewährleisten, dass diese bei Aufenthaltsbeginn auf dem Rechenserver zur Verfügung stehen.

Ihr Gastaufenthalt dient der Bearbeitung von Auswertungsprogrammen anhand der Originaldaten. Diese Programme werden am Ende des Gastaufenthaltes von uns gestartet. Nach Ende des Programmdurchlaufs erhalten Sie von uns ausschließlich die in diesem Programmdurchlauf erzeugten Ergebnisdateien. Während des Gastaufenthalts erzeugte Ergebnisdateien werden nicht weitergegeben.

Bitte beachten Sie bei der Programmierung die Vorgaben in Abschnitt 5, insbesondere die Anlage der Master-Datei samt Dokumentation des Programmablaufs.

3 Datenfernverarbeitung

Wir verändern die von Ihnen eingeschickten Auswertungsprogramme nicht. Sie müssen also vollständig lauffähig übermittelt werden. Wir erwarten, dass die Programme soweit möglich anhand unserer Testdaten überprüft wurden. Neu eingeschickte Dateien überschreiben bereits vorhandene, wenn sie den gleichen Dateinamen haben.

Zur Bearbeitung Ihrer Aufträge benötigen wir in jeder E-Mail die folgenden Angaben:

Projektnummer – Speicherbedarf – erwartete Laufzeit – Datum Vertragsende – Sonstiges

Auf Basis der Information zu Speicherbedarf und erwarteter Laufzeit können wir die Auslastung der Server verbessern. Ziel der Angabe des Vertragsendes ist, dass Sie und wir rechtzeitig an die Verlängerung auslaufender Verträge erinnert werden. Unter Sonstiges bitte

schon in der Betreffzeile auf Dinge hinweisen, die bei der Bearbeitung der Programme oder der Datenschutzprüfung zu beachten sind (z.B. wenn mit MATLAB oder TSP gerechnet werden soll). Damit Ihre Angaben maschinenlesbar sind, müssen sie in standardisierter Form erfolgen, also exakt nach folgendem Muster eingegeben werden.

fdz### – # {mb, gb} – # {m, h, d, w} – yyyy-mm-dd – [Sonstiges]

```
# steht für Ziffern
mb = Megabyte
gb = Gigabyte
m = Minuten
h = Stunden
d = Tage
w = Wochen
yyyy-mm-dd = Datum im Format Jahr-Monat-Tag
```

Beispiel Stata: **fdz041 – 3 gb – 14 h – 2011-01-31**

Die Angabe des Speicherbedarfs ist nur bei Verwendung von Stata erforderlich. Als Nutzer anderer Software setzen Sie bitte die Ziffer 0 ein.

Beispiel Sonstige: **fdz041 – 0 gb – 14 h – 2011-01-31**

Wir bitten um Verständnis dafür, dass wir Aufträge unbearbeitet zurücksenden, wenn relevante Informationen in der Betreffzeile fehlen oder die Formvorgaben nicht erfüllt sind.

4 Struktur der Projektverzeichnisse

Jedes Projekt erhält eigene Projektordner, die über die gesamte Laufzeit des Projektes bestehen bleiben. Werden in einem Projekt mehrere Datensätze benötigt, z.B. BAP und IABS, dann wird für jeden Datensatz ein separater Projektordner angelegt.

Jeder Projektordner enthält folgende Unterverzeichnisse:

orig: In diesem Verzeichnis werden alle Originaldaten bereitgestellt. Das sind neben den beantragten Daten gegebenenfalls auch mitgebrachte oder von uns bereitgestellte Aggregatdaten. Für dieses Verzeichnis erhalten Sie nur Leserechte.

data: Hier sollten alle generierten Datensätze abgelegt werden, Gehen Sie sparsam mit diesem Speicherplatz um, belegen sie bitte hier nicht mehr als 20 GB.

prog: In diesem Ordner befinden sich alle Auswertungsprogramme (do-Dateien, ado-Dateien oder sps-Dateien).

log: Dieser Ordner enthält alle Ergebnisdateien, die bei der Analyse der Daten entstehen, inklusive der Grafiken. Stata-Ergebnisdateien müssen die Endung „.log“ tragen (das scml-Format ist zur Vereinfachung der Datenschutzprüfung nicht zulässig).

doc: Dieser Ordner ist für alle weiteren Dateien, z.B. Word und Excel.

Das Anlegen von weiteren Verzeichnissen oder Unterverzeichnissen ist nicht erlaubt. Die Projektverzeichnisse sind nach den BA-Standards gesichert.

Damit bei Änderungen der Verzeichnisstruktur nicht alle Pfade geändert werden müssen, sind ausschließlich relative Pfade zu verwenden. Die relativen Pfade sind so zu setzen, dass beim Doppelklick auf eine Master-Datei (siehe Abschnitt 5.1) alles Gewünschte durchläuft und die neu erzeugten Dateien in die richtigen Ordner geschrieben werden. Wir empfehlen daher, dass Sie in der Master-Datei zunächst aus dem *prog*-Ordner eine Ebene höher wechseln (`cd ..`). Anschließend können Sie die Unterordner *prog, orig, data* und *log* wie im Beispiel in Abschnitt 5.2 ansprechen. Die Master-Datei steht im *prog*-Ordner. Wenn Ihnen die Verwendung relativer Pfade nicht geläufig ist, sprechen Sie uns an, wir helfen Ihnen gerne.

Der Speicherplatz ist auf 20 GB pro Projekt begrenzt. Dateien im Umfang von bis zu 30 GB können archiviert werden. Die Dateien werden bis zu 10 Jahre nach Projektende gespeichert. Auf archivierte Dateien kann für weitere Berechnungen zurückgegriffen werden, allerdings benötigen wir eine Vorlaufzeit von bis zu zwei Wochen zum Zurückspielen auf den Rechenserver.

5 Organisation der Auswertungsprogramme

Die folgenden Regeln dienen der Minimierung des Bearbeitungsaufwandes und der Nachvollziehbarkeit der Analysen.

5.1 Dokumentation des Programmablaufs per Master-Datei

Alle im Verlauf des Projektes verwendeten Programme sind von einem übergeordneten Programm aufzurufen, welches wir als „Master-Datei“ bezeichnen. Ziel der Master-Datei ist zum einen, dass wir pro Datenfernverarbeitungsauftrag nur eine Datei starten müssen, zum anderen, dass wir den Ablauf der Analysen auch dann noch nachvollziehen können, wenn sich die einzelnen Programmschritte auf viele Dateien verteilen. Dazu ist wichtig, dass in der Master-Datei hinter jedem Programmaufruf eine kurze Beschreibung der Inhalte des jeweiligen Unterprogramms erfolgt.

Die Master-Datei soll alle im Projektverlauf relevanten Auswertungsprogramme enthalten, so dass es jederzeit möglich ist, alle Programmschritte vom Einlesen der Originaldaten bis zur Erzeugung der Endergebnisse mit einem Aufruf auszuführen. Bereits ausgeführte Programmschritte sind im Projektverlauf auszukommentieren. Die Master-Datei sollte in allen Projekten exakt „master.do“ heißen, um das Starten der Programme am FDZ zu erleichtern. Am Ende von Gastaufenthalten werden anhand der Master-Datei die während des

Aufenthalts erarbeiteten Programme gestartet. Nur die dabei erzeugten Ergebnisdateien werden verschickt.

Bei der Nutzung von **SPSS** ist es nicht sinnvoll eine entsprechende Master-Datei anzulegen, da die erzeugten Outputs nur sehr schwer zu kontrollieren sind. Zum einen würden alle Ergebnisse in eine Datei geschrieben, zum anderen ist die Zuordnung der Analyseergebnisse zu den entsprechenden Befehlen nur schwer möglich. Damit wir die einzelnen Analyseschritte dennoch nachvollziehen können, legen Sie bitte eine Textdatei an, die analog zur Master-Datei die Reihenfolge und Inhalte der einzelnen Analysedateien auflistet. Die SPSS-Syntaxen starten wir in der angegebenen Reihenfolge und die einzelnen Ausgaben erhalten den gleichen Namen wie die dazugehörige Syntaxdatei. Bitte schicken sie uns diese Datei mit der Liste der Analyseprogramme immer mit, wenn Sie Veränderungen daran vornehmen.

Die so gestaltete Master-Datei gewährleistet die vollständige Reproduzierbarkeit der Ergebnisse. Damit wird zum einen eine zentrale Anforderung der guten wissenschaftlichen Praxis erfüllt, zum anderen sind nach Projektende nur Originaldaten und Programm-Dateien zu archivieren.

Beispiel für den Inhalt der Datei master.do für Stata:

```
* Master-Datei des Projektes "Projektname"

clear
set more off          // damit das Programm durchläuft
set linesize 255      // damit Tabellen nicht unnötig umgebrochen werden
cd ..                 // Wechsel aus dem prog-Ordner
adopath ++ prog       // Angabe, wo die eigenen ado-Dateien liegen
which command         // so können Sie sicherstellen, dass Sie die
                      // gewünschte Version der ado-Datei ausführen
*do prog/1_PersDat.do      // Aufbereitung der Personendaten
*do prog/2_BetrDat.do      // Aufbereitung der Betriebsdaten
do prog/3_DatXY.do         // Zusammenspielen … Aggregation auf Branchen
do prog/4a_AnalyseA.do     // Regression …
do prog/4b_AnalyseB.do     // Kommentar

capture log close          // Ausgabe einer Dateiliste in eine log-Datei
log using log/Dateiliste.log, replace
dir prog\*
dir log\*
dir data\*
dir orig\*
dir doc\*
log close
```

5.2 Ablauf der Auswertungsschritte

Zur besseren Nachvollziehbarkeit sollten die Analyseschritte wie folgt auf Programmdateien aufgeteilt werden:

- Datenaufbereitung
- Deskriptionen
- Analysen

Zu jedem Thema kann es mehrere Dateien geben. In den Teilen zur Datenaufbereitung und Analyse soll die Zahl der Tabellen begrenzt sein, die wesentlichen sollen sich im Teil Deskriptionen befinden.

5.3 Dateinamen

Die Master-Datei muss in allen Projekten, für die Stata genutzt wird, exakt „master.do“ heißen. Die Auswertungsprogramme sind so zu nummerieren, dass deren Abfolge auch durch Sortieren nach dem Dateinamen im Dateimanager sichtbar wird (siehe Namensgebung in der Beispiel-Master-Datei). Ergebnisdateien und erzeugte Daten-Dateien oder Grafiken sollen den gleichen Dateinamen tragen wie das Programm, mit dem diese Dateien generiert werden. Zur Illustration: in dem Auswertungsprogramm „1_PersDat.do“ werden eine Datendatei, eine Grafik und eine Ergebnisdatei generiert. Zur besseren Zuordnung sollen diese als „1_PersDat.dta“ und als „1_PersDat.log“ bezeichnet werden. Die Grafik erhält analog hierzu den Namen „1_PersDat.gph“ (das Suffix der Dateinamen bezieht sich hier auf Stata). Erzeugt ein Programm mehrere Daten-,Ergebnis-Dateien oder Grafiken, sollen sich deren Dateinamen durch unterschiedliche Endungen unterscheiden. Bei Analysen mit SPSS werden die Outputdateien vom FDZ unter dem entsprechenden Dateinamen gespeichert.

Beispiel für den Inhalt eines Stata-Programms zur Datenaufbereitung:

```
* Programm-Datei des Projektes "Projektname" - 1_PersDat.do
* Inhalt: Aufbereitung der Personendaten

* Aggregationsniveau: Branchen nach Bundesland
*                     aggregiert wird in 1_PersDat.do
*                     mind. 30 Betriebe pro Datenzeile
*                     die Personenzahl pro Zeile steht in Faelle

clear
set mem 600m
capture log close
log using log/1_PersDat.log, replace

use orig/originalname, clear

* Aufbereitung der Variablen ...
bysort id: egen varname2 = modus(varname1), max

* Aggregation auf die Ebene Branchen nach Bundesland
collapse (count) Faelle = varname2 (mean) varname2, by(branche bula)

* Deskriptionen
summarize varlist

save data/1_PersDat.dta, replace
log close
```

5.4 Dokumentation innerhalb der Auswertungsprogramme

Neben der Dokumentation des Programmablaufs in der Master-Datei sollen auch die einzelnen Programme so dokumentiert sein, dass man die einzelnen Schritte nachvollziehen kann. Im Einzelnen erwarten wir:

- Zwischenüberschriften, wie "Aufbereitung der Variablen X" oder "Aggregation auf Betriebsebene",
- die Beschreibung der Inhalte/Funktion von Schleifen,
- die Bildung von Variablenlabels für alle wichtigen Variablen, die neu gebildet wurden,
- die Verwendung möglichst "sprechender" Variablennamen.

Werden die ursprünglichen Mikrodaten für weitere Berechnungen oder Analysen aggregiert (z.B. auf die Ebene von Betrieben, Regionen, Wirtschaftzweigen, Berufen, etc.), muss dies im Analyseprogramm entsprechend deutlich gemacht werden. Bei Aggregationen bitten wir auch darum, dass eine Variable gebildet wird, welche die Anzahl der Personen oder ggfs. Betriebe pro Datenzeile (Aggregat) enthält (z.B. bysort branche: gen N = _N). Dies ist

erforderlich, damit wir die Datenschutzprüfungen vornehmen können. Alle im Anschluss an eine Aggregation erstellten Ergebnisdateien sollen daher am Anfang folgende Informationen enthalten:

- auf welche Ebene aggregiert wurde,
- in welchem Programmschritt aggregiert wurde (dieser Programmschritt sollte auch in der Master-Datei entsprechend kommentiert sein),
- wie viele Personen und ggfs. Betriebe mindestens pro Datenzeile eingegangen sind,
- Name der Variablen, welche die Personen- oder Betriebsanzahl pro Datenzeile enthält.

Fehlt der Hinweis auf die Aggregation am Beginn der Ergebnisdatei, kann es dazu kommen, dass wir bei der Datenschutzprüfung mehr Werte als erforderlich löschen.

5.5 Dokumentation der Fallzahlen

Eine Maßnahme zur Vermeidung von Reidentifikationsrisiken für Personen und Betriebe ist, dass wir im Rahmen unserer Datenschutzprüfung vor dem Übersenden der Ergebnisse sämtliche Werte, die auf weniger als 20 Beobachtungen (Personen oder Betriebe) basieren, durch einen "/" ersetzen. Damit wir die zu löschenden Werte erkennen, sind bei allen statistischen Kennzahlen oder multivariaten Analysen die zugrunde liegenden Fallzahlen auszuweisen. Fehlt die Fallzahl, löschen wir die betreffenden Ergebnisse. Da bei Variablen mit nur zwei Ausprägungen (z.B. Dummys) anhand des Mittelwerts die Anzahl der Fälle in den beiden Ausprägungen berechnet werden kann, müssen bei Mittelwerten neben der Fallzahl immer auch Minimum, Maximum und Standardabweichung ausgegeben werden. Anhand dieser Angaben können wir Variablen mit zwei Ausprägungen von anderen Variablen unterscheiden.

Ebenfalls zur Erkennung der Fallzahlen ist bei Verwendung von Hochrechnungsfaktoren parallel ohne Hochrechnung zu rechnen. Die hochgerechneten und die dazugehörigen nicht-hochgerechneten Ergebnistabellen müssen dabei immer direkt untereinander stehen. Bei umfangreichen Berechnungen mit Hochrechnungsfaktoren empfiehlt es sich, zwei parallele Analyseprogramme, jeweils mit und ohne Hochrechnung, zu verwenden.

5.6 Zulässige Ausgabeformate

Wenn Sie Stata nutzen, geben Sie die Ergebnisse ausschließlich in ASCII-Text-Dateien aus. Bei Formaten wie xls, doc oder rtf ist der Aufwand für die Durchsicht für uns zu hoch. Ergebnisse, die in außerhalb der durch Stata ausgegeben Log-Files in gesonderte Dateien abgespeichert werden, müssen wieder in das Log-File eingebunden werden, um sie zu

erhalten. Bei Berechnungen mit SPSS exportieren wir die Ergebnisse zur Durchführung der Datenschutzprüfung in der Regel nach Excel. Das heißt, Sie erhalten Ihre Ergebnisse als Excel-Dateien.

Aus Stata Grafiken lassen sich teilweise die Ursprungsdaten erzeugen. Daher versenden wir nur noch Stata Grafiken, die mit der Option "asis" erzeugt wurden:

`graph ... , ... saving(`*`filename`*`, replace asis)` oder
`graph save` *`filename`*`, replace asis`

Die Formate wmf, eps und png sind ebenfalls zulässig. Bitte generieren Sie nur dann Grafiken, wenn es nicht möglich ist, die Grafiken durch die Ergebnisdateien z.B. in Excel selbst zu erstellen. Da der Aufwand der Datenschutzprüfungen von Grafiken sehr hoch ist, sollte die Anzahl der erstellten Grafiken so gering wie möglich gehalten werden. Bei jeder Grafik ist nachzuweisen, auf wie vielen Beobachtungen die einzelnen dargestellten Werte basieren. Dies kann über die Angabe der Fallzahlen entweder direkt in den Grafiken oder in Tabellen, die sich unmittelbar vor oder nach der Erstellung der Grafiken befinden, erfolgen.

Falls Sie aggregierte Datensätze als Ergebnisdatei erhalten möchten, sprechen Sie die Vorgehensweise bitte vorher mit uns ab. Grundsätzlich gilt, dass für jede aggregierte Variable im Datensatz ein zusätzliches Merkmalenthalten sein muss, das die Anzahl der gültigen Beobachtungen, auf denen die Aggregation besteht, enthält. Die vergebenen Variablennamen sollen sprechend sein. Für jedes Projekt kann nur einmal ein solcher Datensatz weitergegeben werden. Wir empfehlen in diesen Fällen einen Gastaufenthalt.

6 Publikation der Ergebnisse

In Publikationen, die Auswertungen von FDZ-Daten enthalten, sind Daten und Dokumentationen entsprechend den Vorgaben auf der Homepage des FDZ (http://doku.iab.de/fdz/access/Zitierweisen.pdf) zu zitieren. Zudem ist jeweils ein Belegexemplar an das FDZ zu senden.