1. Nicht aller Anfang ist schwer

1.1. Was ist Datenmanagement? Braucht man das?

Datenmanagement ist die Grundlage jeder Datenverarbeitung. Es wird so selbstverständlich praktiziert, dass es als Management von Daten kaum bewusst wahrgenommen wird. Beispiele sind das Hinzufügen neuer Variablen oder Werte, Korrigieren, Sortieren oder das Ergänzen von Daten. Viele statistische Verfahren führen zuerst Datenmanagementoperationen durch, bevor sie die eigentliche Analyse vornehmen. Nichtparametrische Tests bilden vor dem eigentlichen Test z.B. Ränge.

Datenmanagement bedeutet jedoch noch viel mehr, z.B. Transponieren, Bilden von Subsets über Filter, Gruppenvariablen oder Zufallsfunktionen, Bilden neuer Variablen oder Werte über Umkodierungen oder arithmetische Operationen, Zusammenfügen von Datensätzen, uvam.

Datenmanagement ist essentiell. Professionelles Datenmanagement ist die Voraussetzung für einen korrekten Datensatz und ein Datensatz ist die Basis Ihrer wissenschaftlichen Arbeit. Jede Analyse bzw. jede Fragestellung *setzt voraus*, dass die Transformation bzw. das Ergebnis (Zustand) von Datensätzen (Dateien), Werten (Zeichen) oder Variablen (Datenfelder) korrekt ist. Diese Überprüfung übernimmt kein Statistikprogramm der Welt; diese Prüfungen müssen und können nur die Anwender selbst vornehmen. Nicht überprüftes Datenmanagement bedeutet nicht automatisch korrektes Datenmanagement. Erst wenn Sie *100% genau wissen,* dass Ihre Datengrundlage *völlig in Ordnung* ist (und nicht mehr *vermuten*, dass sie *wahrscheinlich* in Ordnung ist), können und dürfen Sie mit Auswertungen fortfahren.

Beherzigen Sie für Ihr weiteres wissenschaftliches Arbeiten u.a. folgende Ratschläge (vgl. dazu auch Kap. 8).

- Trauen Sie Ihren Daten nicht, trauen Sie schon gar nicht den Daten von Dritten und erst recht nicht einem Analyseprogramm, nicht einmal SPSS.
- Prüfen Sie *jede* Eingabe doppelt. Bei der menschlichen Dateneingabe per Hand treten ca. 5% Fehler auf.
- Erstellen Sie Sicherheitskopien von *jeder* Datensatzversion.
- Protokollieren Sie sorgfältig *jede* Änderung an Ihrem Datensatz und zwar per Syntax. Stellen Sie sich nur mal vor, Sie nehmen stundenlang von Hand aufwendige, wenn nicht sogar komplizierte Korrekturen an ihrem Datensatz vor und dann stellt sich heraus, dass aufgrund eines Hard- oder Softwarefehlers keine der Änderungen umgesetzt wurde. Dieser Aufwand war umsonst. Dies wäre nicht passiert, wenn die Modifikationen per Syntax vorgenommen worden wären.
- Nicht alle Voreinstellungen von SPSS (z.B. die für Missings) werden Ihrer Analyse immer automatisch angemessen sein. Darüber hinaus funktioniert nicht einmal ein Statistikprogramm 100% ordnungsgemäß. Deshalb: Vertrauen ist gut, Kontrolle jedoch viel besser.
- Prüfen Sie *jede* unprotokollierte Änderung an Ihrem Datensatz daraufhin, ob diese gezielt und legitim oder zufällig und unberechtigt ist, u.a. am Umfang des Datensatzes (Variablen, MB/KB-Größe) oder auch am Speicherdatum.
- Prüfen Sie immer *mehrmals* die Funktionalität Ihrer Syntax. Kommentieren Sie in Ihrer Syntax die Operationen, die diese ausführen soll.

Planung ist alles...

Viele Projekte bleiben „hängen", weil die Verantwortlichen nicht wissen, wie die Daten für die Analyse mit SPSS vorbereitet werden müssen.

Datenmanagement kommt immer vor der eigentlichen Analyse. Erstens zeitlich: Datenmanagement kommt immer zeitlich vor der eigentlichen Analyse bzw. Entscheidung; auch unterlassenes (und daher nicht geprüftes) Datenmanagement *ist* Datenmanagement und womöglich sogar *fehlerhaftes* Datenmanagement. Zweitens kausal: Geplantes und explizites Datenmanagement machen Analysen und darauf aufbauende Entscheidungen oft erst möglich.

Vielen ist nicht klar, wie aufwendig oder kompliziert die Datenvor- und -nachbereitung sein kann. Hinweise auf die Dimensionen können Veröffentlichungen zu Data Warehouses entnommen werden. Nach Cabena et al. (1998, 43) entfallen darauf z.B. 90% des Zeitaufwands. Der Verfasser betreute unter anderem ein Projekt, bei dem das Verhältnis der Aufwendungen für Datenmanagement und Datenanalyse in einem Verhältnis von 20:1 standen. Solche Ausmaße sind durchaus realistisch und werden ausschließlich und alleine von den Eigenschaften der betroffenen Datensätze (Dateien), Variablen (Datenfelder) und Werte (Zeichen) bestimmt. Man sollte erst dann behaupten, dass Datensätze, Variablen und Werte in Ordnung sind, wenn man sich davon überzeugt hat. Die dazu erforderlichen Ressourcen und Fähigkeiten sollten keinesfalls unterschätzt werden.

Wenn der Aufwand und die Komplexität des Datenmanagements nicht angemessen eingeschätzt werden, gerät man bei einem Projekt leicht in Konflikt mit Ressourcen und Deadlines. Bei einem Projekt (Analyse, Fragestellung) vermeidet man die skizzierten Probleme (Verzögerungen, Ressourcenvergeudung, Qualitätseinbußen, nicht eingehaltene Dokumentations- bzw. Nachweispflicht) während der Analysephase, wenn von vorneherein eindeutig transparent und geplant ist, wie die Daten für die weiteren Projektphasen aufbereitet werden sollen und welche Maßnahmen und Fähigkeiten dafür erforderlich sind.

Datenmanagement: Ein Thema mit Variationen...

Datenmanagement kann verschiedene Ebenen formatierter Informationen betreffen, z.B. Datensätze (Dateien), Variablen (Datenfelder) und Werte (Zeichen). Je nachdem, welche Ebene betroffen ist, kann Datenmanagement in unterschiedlichster Weise ablaufen.

Umgangssprachlich laufen scheinbar dieselben Prozesse ab, z.B. das „Zusammenfügen" von Datensätzen, Variablen und Daten. Da die dahinter ablaufenden technischen Prozesse jedoch völlig andere sind, werden die Abläufe im Detail mit unterschiedlichen, aus der Informatik stammenden Begriffen auseinandergehalten, die sich z.T. wiederum in den einzelnen Programmierbefehlen bzw. -funktionen wiederfinden können. Das „Zusammenfügen" von Datensätzen wird z.B. je nach Art und Weise als „Verketten" bzw. „Joinen" bezeichnet. Das „Zusammenfügen" von Variablen kann als „Neuberechnung" oder „Aggregation" beschrieben werden. Das „Zusammenfügen" von Daten bzw. Werten (z.B. innerhalb einer numerischen Variablen) kann als „Concatenate" bzw. „Verbinden" bezeichnet werden.

Die Art der Analyse bestimmt die Weise des vorangehenden Umgangs mit den Daten. Je nach gewünschter Analyse werden also jeweils andere, vorbereitende Schritte des Datenmanagements erforderlich sein; manchmal sogar mehrere kombiniert. SPSS bietet standardmäßig verschiedene Ansätze zum Aufbereiten von Datensätzen, Variablen und Werten an.

- auf der Ebene von Datensätzen (Dateien): Einlesen, Zusammenfügen bzw. Aufteilen, Umstrukturieren, Transponieren, Fälle bzw. Untergruppen (Subsets) auswählen über Filter, Gruppenvariablen oder Zufallsfunktionen, Anlegen von Zufallsdatensätzen, Speichern.
- auf der Ebene von Variablen (Datenfeldern): Vereinheitlichen, Formatieren und Bilden neuer Variablen (z.B. über Bedingungen, arithmetische Funktionen oder sonstige Operationen), Definieren bzw. Überprüfen von Missings, Gewichtungen.
- auf der Ebene von Werten (Zeichen): Suchen, Zählen, Formatieren, Umkodieren (automatisch, gezielt) uvam. Für den Umgang mit Datums- und Zeitangaben gibt es z.B. spezielle Funktionen für das Zusammenfassen, Konvertieren und Extrahieren.

Die eigentliche Power von SPSS geht allerdings über diese Standardoperationen weit hinaus.

... und noch viel mehr

Die Power von SPSS beginnt an der Stelle, an der diese Basisfunktionen anwendungs- bzw. praxisorientiert umgesetzt werden. Datenmanagement kann somit im Prinzip die Daten für jede denkbare Anwendung, Fragestellung bzw. Analyse vorbereiten. Praxisorientierte Anwendungen von Datenmanagement sind zum Beispiel

- die Analyse von Mehrfachantworten.
- die Analyse von (halb)offenen Textangaben.
- Makroprogrammierung.

In einer separaten Veröffentlichung werden vom selben Autor weitere Möglichkeiten des Datenmanagements mit dem Schwerpunkt der Sicherung von Datenqualität mit SPSS vorgestellt.

- Überprüfung von Datensätzen, Variablen, Werten bzw. Missings auf Vollständigkeit.

- Vereinheitlichung von Kodierungen, Werten bzw. Strings, Datumsvariablen und -werten, Währungen und Messeinheiten, Symbolen oder Sonderzeichen oder auch über Zählen von Schablonen.
- Identifizieren, Evaluieren und Entfernen doppelter Datenzeilen.
- Überprüfen von Missings (Anzeigen, Löschen, Rekonstruktion bzw. Imputation).
- Überprüfen von Ausreißern (uni-, multivariat).
- Überprüfung von Plausibilität (Datenqualität).
- Überprüfen mehrerer Variablen und Datensätze gleichzeitig.

Ein weiteres interessantes Thema wäre die Arbeit mit Datenbanken. Dies kann aber in diesem noch eher grundsätzlichen Rahmen nicht vorgestellt werden. Die Makroprogrammierung wird im Kontext von Datenmanagement deshalb vorgestellt, weil SPSS-Makros ein ausgezeichnetes Instrument sind, um durch Automatisierung wiederholtes bzw. multivariates Management von Daten oder Analysen extrem zu beschleunigen.

Absicht dieses Buches

Eine der vielen Ursachen für die fälschliche Annahme, dass sich Daten *automatisch* in einem auswertungsfähigen Zustand befinden, ist, dass derzeit zu anwendungsorientiertem Datenmanagement mit SPSS keine Literatur zur Verfügung steht. Dieses Buch versucht diese Lücke in mehrerer Hinsicht zu schließen. Es möchte die fundamentale Rolle von explizitem und kontrolliertem Datenmanagement für die Analyse von Daten hervorheben, vor allem, indem es den Zugang zur eigentlichen Power von SPSS zu Einsatzmöglichkeiten öffnet, die weit über die Optionen für Mauslenker hinausgehen.

Die Didaktik des Buches ist anwendungs- und erfolgsorientiert. Die Einsatzmöglichkeiten des vielleicht anfangs noch ungewohnten Datenmanagements mit SPSS-Syntax sind dabei nach praxisorientierten Fragestellungen geordnet, z.B. Analyse von (halb)offenen Textfragen, Umstrukturieren von Datensätzen oder Makroprogrammierung und darin wieder von den unkomplizierten Zugängen, die bis zu recht anspruchsvollen und mächtigen Ansätzen gesteigert werden.

Der Sinn dieser sukzessiven Steigerung anhand von SPSS-Syntax ist dabei nicht nur zu zeigen, wie nachhaltig effektiv und effizient Datenmanagement sein kann, sondern auch, dass Datenmanagement sorgfältig geplant, explizit und transparent sein muss. Letztlich soll dies eine Sensibilisierung dafür sein, bei Analysen nicht bloß ergebnisorientiert mit dem

SPSS-Ausgabefenster zu arbeiten, sondern vor allem und zunächst *ablauforientiert* mit dem ausgegebenen Syntaxprotokoll und dem selbst angelegten Syntaxprogramm. Die Ergebnisse sind erst dann einzusehen, wenn Optionen und Art ihrer Anforderung überprüft und für in Ordnung befunden sind.

Die Herabsetzung von Datenmanagement zu „Fußarbeit" wäre eine völlige Verkennung des Stellenwerts dieser Tätigkeit und ihrer Komplexität. Ein Anliegen dieses Buches ist es daher auch, diese zu oft vernachlässigte Tätigkeit und ihre Bedeutung in den Vordergrund zu rücken. Intelligentes Datenmanagement *ist* unabdingbare Grundlage und Voraussetzung für informationsbasierte Entscheidungen in allen Anwendungsbereichen und -ebenen.

1.2. Wieso Syntax?

Dies ist die am häufigsten gestellte Frage und soll daher auch gleich beantwortet werden. Wieso Syntax? Nicht nur, weil Sie per Syntaxsteuerung letztlich schneller, sorgfältiger und flexibler arbeiten können. Jemand, der bereits einigermaßen gut mit Syntax programmieren kann, schlägt Maustenker um Längen.

Der Hauptgrund ist: Ein Syntaxprotokoll bzw. ein Syntaxprogramm ist die einzige Möglichkeit, die getätigten Mausklicks zu kontrollieren; es gibt dazu keine Alternative. Die Abfolge von Mausklicks wird sonst in keiner anderen Form protokolliert. Anzunehmen, dass am Ergebnis, einer SPSS-Ausgabe, die Art und Abfolge von Mausklicks kontrolliert werden kann, ist ein grundlegender Irrtum.

Eine Ausgabe gibt nur das deskriptive, grafische oder inferenzstatistische Ergebnis wieder, protokolliert aber nicht alle SPSS-Voreinstellungen, z.B. den Umgang mit Missings. Auch ist es eine Fehleinschätzung davon auszugehen, sich immer die getätigten Mausklicks merken zu können; das geht schon gar nicht in der Situation, wenn man sich verklickt hat. Die Maussteuerung ist dafür typischerweise sehr anfällig.

Ablauforientiertes Arbeiten vor ergebnisorientiertem Arbeiten!

Letztlich soll dies eine Schulung dafür sein, bei Analysen ablauforientiert mit dem ausgegebenen Syntaxprotokoll und dem selbst angelegten Syntaxprogramm zu arbeiten. Die Ergebnisse sind erst dann einzusehen, wenn die SPSS-Optionen und die Art ihrer Anforderung überprüft und für in Ordnung befunden wurden.

Syntaxprogrammierung hat nur Vorteile

- Validierung: Syntaxsteuerung birgt in sich den konstruktiv zu sehenden Zwang zur inhaltlichen Validierung einer Analyse; damit ist gemeint, dass Programmieren eher dazu zwingt nachzudenken, warum und wieso etwas von SPSS ausgeführt werden soll als Maussteuerungen, die durchaus auch mal gedankenlos erfolgen können. Die mechanische Anwendung von Menüs, Buttons und Optionen ist generell nicht zu empfehlen.
- SPSS als Syntaxgenerator: SPSS kann so eingestellt werden, dass es zu den Mausklicks und Eingaben die im Hintergrund generierte Befehlssyntax ausgibt, die Sie dann für eigene Zwecke abspeichern, kopieren, umschreiben, erweitern (und vieles andere mehr) können.
- Automatisierbarkeit und Wiederverwendbarkeit: Einmal geschrieben oder gespeichert, können Sie ein Syntaxprogramm immer wieder verwenden.
- Geschwindigkeit: Die Abarbeitung eines Syntaxprogramms ist um ein Vielfaches schneller als das (wiederholte) Anklicken von Menüs.
- Offenheit: Sie können ein Programm immer wieder durch direktes Hineinkopieren von Codezeilen oder auch von Hand erweitern bzw. überarbeiten.
- Erweiterbarkeit: Sie können den regulären Leistungsumfang von SPSS erweitern, indem Sie über Programme oder Skripte zusätzliche Funktionen in SPSS integrieren.
- Effizienz: Sie können Programmcodes zu Makros umschreiben, die die Automatisierbarkeit und Effizienz von Prozessabläufen noch weiter erhöhen. Mit zunehmender Professionalisierung sind Sie mit Syntax in der Lage, (u.a. über Makros) Programme zu schreiben, die z.B. mit nur einem Bruchteil an Codezeilen denselben Leistungsumfang erreichen.
- Flexibilität: Syntaxsteuerung ist flexibler und bietet mehr Möglichkeiten des Datenmanagements als eine Menüsteuerung; es gibt in SPSS einige Funktionen, die Sie nicht über die Maus-, sondern nur über die Syntaxsteuerung ansprechen können (z.B. MANOVA, Ridge Regression).
- Übersichtlichkeit und Systematisierung: Syntaxsteuerung bietet Übersichtlichkeit bei der Auswertung auch von mehreren hundert Variablen. Syntax ist für die Analyse großer Datensätze weitaus geeigneter als Maussteuerung.

- Einheitlichkeit: Syntax ist eine einheitliche und technisch klar definierte Sprache und erklärt sich im Prinzip stets selbst. Eine Anleitung für die Syntaxsteuerung ist insofern auch eine Anleitung für die Maussteuerung.
- Kommunikation: Der Austausch von prinzipiell selbsterklärender Syntax zwischen oder innerhalb international arbeitender Forschungsprojekte erleichtert die Kommunikation, Evaluation und Interaktion und trägt zu ihrer Präzisierung bei. In den syntaxorientierten Kapiteln kann z.B. über die Erläuterung der dazugehörigen Syntax der Leistungsumfang der diversen Prozeduren differenzierter beschrieben werden als über Abbildungen.
- Individualisierung: Syntaxsteuerung erlaubt, anhand der Syntax jede eingestellte Option zu überprüfen; das bedeutet, Sie entdecken dadurch auch (zwar gutgemeinte) Voreinstellungen seitens SPSS, die aber gerade für Ihre individuelle Datensituation definitiv dysfunktional sein können. Vertrauen ist gut, Kontrolle besser.
- Austausch: Syntaxprogramme können als Textdokumente in alle Welt verschickt werden; falls Sie z.B. Fragen zur Angemessenheit einer Analyse haben, kopieren Sie einfach die Syntax in eine E-Mail und schicken diese ab. Mausklicks können Sie nicht versenden.
- Protokollierung und Dokumentation: Dieser Aspekt ist nicht unwichtig und kann helfen, v.a. bei Peer-Reviews und Evaluationen peinliche Situationen zu verhindern. Falls Sie z.B. eine langwierige Analyse per Maussteuerung vorgenommen haben und jemand möchte sehen, *wie* Sie die Analyse gerechnet haben und Sie haben dann kein Syntaxprogramm parat, dann müssen Sie mindestens mit kritischen Fragen, Mehrarbeit (die Sie sich hätten ersparen können) oder sogar (evtl. unnötige) Ablehnung von Veröffentlichungen oder Ressourcen rechnen. Manche Institutionen fordern standardmäßig die Analysesyntax mit ein, um Arbeiten oder Projekte begutachten zu können.
- Permanenz: Jahrelanges Mausklicken können Sie nicht speichern. Aber jahrelanges Syntaxprogrammieren. Wenn Sie einmal ein Programm geschrieben haben, können Sie es auch Jahre später unverändert wieder verwenden. Einmal seitens SPSS angebotene Syntax wird auch nicht „weggeworfen". Wird über die Menüsteuerung eine Syntax (z.B. LOGLINEAR, gibt keine Abweichungsresiduen aus) durch eine andere ersetzt (z.B. GENLOG, gibt Abweichungsresiduen aus), kann dennoch auf die Vorteile der ersetzten Verfahren zurückgegriffen werden. LOGLINEAR ermöglicht z.B. im Gegensatz zu GENLOG die Katego-

rien eines Faktors über Kontraste zu reparametrisieren. Was man also mit GENLOG nicht über den Maus- und Syntaxzugriff berechnen kann, schafft LOGLINEAR über die Syntaxsteuerung.

- Unabhängigkeit: SPSS-Syntax ist aufwärtskompatibel und weitestgehend plattformunabhängig. Ist einmal ein SPSS-Programm geschrieben, läuft es auf jeder höheren SPSS-Version (was auch bedeutet, dass einmal geschriebene Syntax automatisch auch ggf. optimierte Algorithmen anspricht). Schließt das SPSS-Programm keine hardwarebezogenen Spezifika mit ein, ist der SPSS-Code darüber hinaus plattformunabhängig (vgl. die Anmerkungen zu SPSS for Macintosh am Ende des Buches). Wenn Sie andere Betriebssysteme anschaffen, können Sie SPSS-Programme weiterverwenden.
- Fehlerresistenz: Syntax ist generell weniger fehleranfällig und funktioniert auch dann, wenn Buttons oder Menüs bei der Maussteuerung versagen (siehe z.B. bei MAPS). Bei mausgesteuerter Analyse können nicht selten Fehler in der Programmierung der Buttons dazwischenfunken.

Was spricht dann eigentlich noch gegen die Syntaxprogrammierung? Nicht einmal mehr die Standardantwort, Syntaxprogrammieren sei schwer. Das einzige, was jetzt zu tun ist, ist zu zeigen, dass Syntaxprogrammieren ziemlich leicht ist und meiner Erfahrung nach mehr Spaß macht als Maussteuerung. Syntax und Produktionsmodus sind in der SPSS-Studentenversion nicht verfügbar.