Datenqualität mit SPSS

von
Dipl.-Psych. Christian FG Schendera

Oldenbourg Verlag München Wien

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.d-nb.de> abrufbar.

Rosenheimer Straße 145, D-81671 München
Telefon: (089) 45051-0
oldenbourg.de

Lektorat: Wirtschafts- und Sozialwissenschaften, wiso@oldenbourg.de
Herstellung: Anna Grosser
Coverentwurf: Kochan & Partner, München
Gedruckt auf säure- und chlorfreiem Papier
Druck: Grafik + Druck, München
Bindung: Thomas Buchbinderei GmbH, Augsburg

ISBN 978-3-486-58214-7

Vorwort

Nicht nur bei der Arbeit mit SPSS gilt:
Datenqualität ist nicht alles, aber ohne Qualität der Daten ist alles nichts.

Seit knapp 20 Jahren bin ich u.a. mit der professionellen Arbeit mit Daten konfrontiert. „Konfrontation" deshalb, weil die Analyse von Daten tatsächlich eine Herausforderung darstellt, genau betrachtet sogar auf drei Ebenen der Arbeit mit Daten: Datenmanagement – Datenqualität – Datenanalyse (Überschneidungen sind möglich). Datenmanagement ist dabei das eher allgemeine Aufbereiten von Daten, Datenqualität das eher kriteriengeleitete Bereitstellen von Daten, Datenanalyse die (vorrangig) inferenzstatistische Analyse von Daten. Bereits während meines Studiums an der Universität Heidelberg hatte ich zahlreiche Gelegenheiten, meine Analysefertigkeiten (damals noch an Großrechnern) anzuwenden und zu erweitern. Etwas fiel mir dabei immer wieder auf: Lehrbücher (z.B. zu SPSS oder zur Statistik) bezogen sich immer nur auf *ideale* Datensituationen. Wenn ich eine Erfahrung während all dieser Jahre, Hunderten von Projekten und unzähligen Analysen, gemacht habe, dann ist es die, dass Daten *üblicherweise nicht ideal* sind: Daten sind fehlerhaft, haben Ausreißer, Lücken, Doppelte und weisen alle denkbare und undenkbaren Arten von Fehlern auf. *Daten sind „schmutzig"*. Saubere Daten sind die Ausnahme, verunreinigte Daten die Regel.

Was ich mir damals oft gewünscht hatte, war eine Zusammenstellung von Regeln, sowie ein übergeordnetes Konzept, das mir gesagt hätte: „So, wenn Du mit Daten arbeitest, können u.a. folgende Probleme auftreten. Wenn Du dieses Problem mit Deinen Daten hast: Prüf' dieses. Wenn Du jenes Problem hast, prüf' jenes." Erstens wäre das jedoch schon mal die falsche Herangehensweise gewesen: Datenqualitätsproblemen begegnet man nicht erst *reaktiv* bei erkennbaren Fehlern, sondern *proaktiv* wegen der Möglichkeit nicht per Augenschein erkennbarer Probleme. Zweitens, leider, gab es ein solches Buch bzw. Konzept nicht. Im Gegenteil könnte dieses Buch, das Sie gerade in Händen halten, vielleicht sogar das erste Buch zu SPSS überhaupt sein, das sich ausschließlich mit ausgewählten Problemen im Zusammenhang mit Datenqualität und ihrer praktischen Lösung beschäftigt. Ich wünsche mir selbst aufrichtig, dieses Buch hätte es schon vor zwanzig Jahren gegeben. Was glauben Sie, wieviel Arbeit, Ärger und Zeit auch mir erspart worden wäre? Und was glauben Sie, welche Datenaspekte überhaupt Probleme und Fallstricke bereiten könnten? Der nächste Abschnitt verrät es Ihnen.

Dieses Buch führt zunächst ein in Kriterien und Maßnahmen zur Gewährleistung einer definierbaren und nachweisbaren optimalen Datenqualität mit SPSS. Die Kriterien wurden aus anerkannten Standards, wie auch Vorschlägen und Empfehlungen von Freunden und Kollegen aus dem DWH- und Methodenbereich zusammengestellt. Trotz des Umfangs dieser

Veröffentlichung muss betont werden, dass es sich nur um eine *Auswahl* von Kriterien und Empfehlungen handeln kann.

Das Buch ist so allgemein gehalten, dass die zusammengestellten Kriterien auf SPSS Datensätze, aber im Prinzip auch auf Data Warehouses (z.B. mittels Clementine, vgl. Kapitel 16) angewandt werden können. Mögliche Unterschiede in der Definition von Daten (z.B. Geschäftsdaten vs. Daten der Grundlagenforschung) und im Ablauf ihrer Prüfung werden in Kapitel 2 angesprochen. Aspekte wie z.B. die Analyse von komplexen (Geschäfts-, Daten-, usw.)Prozessen oder auch die konkrete Planung von Datenqualitätsprojekten können nur angedeutet werden. Auf Data Warehousing oder Data Mining im Zusammenhang mit speziellen Anforderungsbereichen (z.B. Basel II) kann aus Platzgründen leider nicht eingegangen werden.

Kapitel 1 führt in die am häufigsten auftretenden Problembereiche ein, z.B. Vollständigkeit, Einheitlichkeit, Doppelte, Missings, Ausreißer und Plausibilität. Ein erstes, schematisches Konzept verdeutlicht die Zusammenhänge der Kriterien untereinander und die grundsätzliche Bedeutung der Qualität von Daten. Weitere Kriterien für die Qualität von Daten, sowie ihre Kommunikation werden in den Kapiteln 13 und 19 vorgestellt.
Kapitel 2 skizziert grundsätzliche Rahmenbedingungen zur Herstellung von Datenqualität, u.a. Ressourcen, die Priorisierung von Zielen (Relevanz) und Kontrolle durch Protokolle (SPSS Syntax).
Kapitel 3 beschreibt erste Kontrollmöglichkeiten der Vollständigkeit von Datensätzen, Fällen (Zeilen), Variablen (Spalten) und Werten bzw. Missings.
Kapitel 4 stellt zahlreiche Möglichkeiten vor, Uneinheitlichkeit zu identifizieren bzw. in numerischen Werten, Zeiteinheiten und Strings zu vereinheitlichen.
Kapitel 5 führt in das Problemfeld des Erkennens, Interpretierens und (ggf.) Ausfilterns von mehrfachen Werten bzw. Datenzeilen ein. Dieser Abschnitt ist um Hinweise zur Identifikation von Doppelten beim Einlesen von gruppierten sowie genesteten Daten ergänzt.
Kapitel 6 stellt das Umgehen mit fehlenden Daten (Missings) vor. Nach der Beurteilung von Missings im Hinblick auf Ursachen (Muster), Folgen, Ausmaß und Mechanismen werden zahlreiche Methoden der Rekonstruktion und des Ersetzen von Missings beschrieben: U.a. über die Cold deck-Imputation, zufallsbasierte bzw. logische Ansätze, univariate Schätzung, multivariate Ähnlichkeit (Hot deck-Imputation) oder auch multivariate Schätzung (Ansatz der internen Konsistenz, Missing Value Analysis, MVA).
Kapitel 7 erläutert das Erkennen, Interpretieren und Umgehen von bzw. mit Ausreißern. Im Zusammenhang mit den Merkmalen von Ausreißern wird zunächst die besonderen Rolle der Erwartungshaltung („Frames") diskutiert. Im Anschluss an die Identifikation von univariaten bzw. multivariaten Ausreißern über Maße, Regeln, Tests und Diagramme werden Möglichkeiten des Umgehens mit Ausreißern vorgestellt.
Kapitel 8 beschreibt qualitative und quantitative Ansätze zur Überprüfung der Plausibilität. Die Überprüfung der Qualität von Daten in der Praxis wird zunächst an einer einzelnen Variablen erläutert (mit Beispielen für eine kategoriale Variable, eine Stringvariable und eine metrische Variable). Im Anschluss daran wird die Überprüfung der multivariaten Qualität von Daten anhand eines qualitativen, sowie eines genuin quantitativen Ansatzes (Anomalie-Ansatz) vorgestellt.

Kapitel 9 stellt das effiziente Überprüfen mehrerer Variablen und Kriterien mittels Prüfregeln vor. Dieses Kapitel stellt den seit SPSS 14 verfügbaren (sofern lizenziert), mächtigen Menüpunkt „Validierung" bzw. die SPSS Prozedur VALIDATEDATA vor. Das Kapitel führt zunächst in die Ansteuerung per Maus ein; abschließend wird auf die Erweiterung durch selbstgeschriebene Prüfprogramme in SPSS Syntax übergegangen.
Kapitel 10 umfasst zahlreiche weitere Beispiele für das Überprüfen mehrerer Werte, Zeilen und Spalten *in einem Datensatz* auf einmal. Von besonderem Interesse dürften u.a. die zahlreichen Varianten der vorgestellten Zählvariablen (Counter) sein, sowie weitere spezielle Anwendungen, z.B. das Umbenennen zahlreicher Variablennamen (Präfixe, Suffixe).
Kapitel 11 umfasst zahlreiche weitere Beispiele für die Arbeit *mit mehreren (separaten) Datensätzen* auf einmal. Von besonderem Interesse dürften u.a. die diversen Makros zum Screenen, Aufteilen oder auch Zusammenfügen mehrerer Datensätze sein. Auch werden diverse Anwendungsmöglichkeiten der Optionen des SPSS Befehls DATASET vorgestellt.
Kapitel 12 befasst sich mit zeit- bzw. datumsbezogenen Problemen, und ihrem Erkennen und Lösen. Von besonderem Interesse dürfte u.a. der Abschnitt zum Zeitstempel sein.
Kapitel 13 stellt weitere Kriterien für die Qualität von Daten vor, u.a. Menge, Eindeutigkeit, Relevanz, Genauigkeit oder auch Verständlichkeit. Die Einhaltung dieser Kriterien kann üblicherweise eher vom Anwender bzw. Auftraggeber anhand genau formulierter (semantischer) Zielvorstellungen beurteilt werden als seitens SPSS mittels formeller Prüfregeln.
Die *Kapitel 14 bis 18* enthalten eine kleine Übungsaufgabe (Kapitel 14), ein Programmbeispiel für die Umsetzung einer ersten Strategie (Kapitel 15), Hinweise für SPSS Syntax und Datenqualität in Clementine (Kapitel 16) und Macintosh User (Kapitel 17), sowie eine Checkliste (Prüfdokumentation). Diese weniger mit Kriterien, sondern eher mit der konkreten Arbeitspraxis mit SPSS befassten Kapitel bilden zusammen auch eine Abgrenzung zum sich anschließenden Kapitel zur Planungs-, Analyse- und Ergebnisqualität. Kapitel 16 demonstriert insbesondere, wie Clementine über das Einbinden von SPSS Syntax und Prozeduren allgemein und insbesondere hinsichtlich der Gewährleistung von Datenqualität erweitert werden kann. Kapitel 18 enthält eine kopierfähige Liste ausgewählter Kriterien, anhand der Anwender die Art und Weise der Umsetzung von Qualitätskriterien protokollieren können.
Kapitel 19 stellt kommentierte Kriterien für die *Kommunikation* der Qualität von Daten, Erhebungen und Analysen zusammen. Wenn man sich während der Analysephase um die Qualität von Daten und Ergebnissen bemüht, ist zu vermeiden, diesen positiven Impetus durch eine suboptimale Kommunikation zunichte zu machen. Ein separates Kapitel ist den „Todsünden" professionellen Arbeitens vorbehalten. Der Sinn dieses Kapitels ist, bereits im Ansatz zu vermeiden, dass bestimmte Praktiken oft gar nicht erst als unseriös oder auch als unprofessionell erkannt werden, bis es zu spät ist. Mit diesem Kapitel soll auch der zunehmenden Qualitätsabnahme in wissenschaftlichen Veröffentlichungen entgegengetreten werden, die vom professionell arbeitenden Teil der Wissenschaft mit wachsender Sorge beobachtet wird. Es gilt zu vermeiden, dass wissenschaftliche Arbeit von der Öffentlichkeit und auch von Teilen der scientific community selbst nicht mehr ernst genommen wird. Ich erlaube mir bereits an dieser Stelle alle Leser aufzufordern, durch Rückmeldungen und Vorschläge zur Ergänzung oder Verbesserung dieses Buches gerade für dieses Kapitel (aber nicht nur) in zukünftigen Auflagen beizutragen (vgl. Kapitel 21).

Die *Kapitel 20 bis 24* enthalten die Literatur, die Aufforderung zur Rückmeldung Ihrer Meinung zu diesem Buch, eine Kurzvorstellung des Autors, sowie abschließend die Verzeichnisse zur SPSS Syntax, sowie zu den Stichworten.
Am Ende dieses Buches sollten Sie zahlreiche Kriterien für Datenqualität, und die wichtigsten Funktionen von SPSS für ihre Gewährleistung kennen, anhand Ihres Optimalitätsstandards definieren und über Maus- oder Syntaxsteuerung auf Ihre Daten anwenden können. Anwender können dabei bereits vorliegende Daten ex post daraufhin prüfen, ob sie definierten Kriterien entsprechen, wie auch über die Entwicklung eines Standards (Regeln) vor der Dateneingabe bzw. –migration auch antizipierend dafür Sorge tragen, dass neue Daten erst gar nicht fehlerhaft ins Analysesystem gelangen können, z.B. bei der Integration von Filtern bereits beim Datenzugriff oder bei der Dateneingabe.

Die Maßnahmen zur Gewährleistung von Datenqualität werden in der Literatur unter den verschiedensten Oberbegriffen zusammengefasst. Die Begriffe variieren vom Beschreiben des *Prozesses* (u.a. Datenbereinigung, Datensäuberung, Datenstandardisierung, Datenprüfung, Data Cleaning/Cleansing, Deduplikation, Datenveredelung, Datenhygiene, Preprocessing, Plausibilitätsanalyse, Plausibilitätsprüfung, Data Scrubbing, Data Stewardship) bis hin zu dem des *Zieles* (u.a. Datenplausibilität, Datenqualität, Datenintegrität, Qualitätssicherung, Qualitätsmanagement, Quality Assurance). Diese Begriffe stammen v.a. aus dem Kontext des Data Warehousing, das ein weites Feld (oder eher: eine Spielwiese?) für viele weitere technizistisch anmutende -ing und -ung-Anglizismen (v.a.) zu sein scheint (vgl. u.a. Matching, Parsing, Householding, Konsolidierung, Standardisierung uam.). Die Maßnahmen unterstützen Sie nicht nur beim Gewährleisten der Qualität quantitativer Einträge, sondern auch qualitativer Einträge, z.B. Buchstaben oder Texten. Als eine allererste Anwendung wäre z.B. die Überprüfung auf bzw. Gewährleistung der Einheitlichkeit von Produktnamen oder anderer qualitativer Informationen oder Kodierungen zu nennen.
Allen Maßnahmen zur Gewährleistung von Datenqualität ist jedoch gemeinsam, dass sie allererste Priorität vor der eigentlichen Analyse haben und anstrengend, aufwendig, aber auch gefährlich sein können. „Gefährlich" im wahrsten Sinne des Wortes: Gerade Langeweile oder Stress verursachen aufgrund nachlassender Konzentration oder Motivation schnell Fehler.
Sie werden daher sehr gefordert sein: Das Anspruchsvolle der Maßnahmen zur Gewährleistung von Datenqualität zielt im Gegensatz zum allgemein recht unkomplizierten Hypothesentest nicht darauf ab, bekannte Faktoren oder Variablen auf erwartete Zusammenhänge oder Unterschiede zu untersuchen, sondern unbekannte Faktoren oder Variablen auf auch unerwartete Effekte zu untersuchen. Denken Sie bei all der Detektivarbeit und Ursachenforschung an eine Maxime von Sherlock Holmes: „Wenn man das Unmögliche ausschließt, dann ist das, was übrig bleibt – und sei es noch so unwahrscheinlich – unabdingbar die Wahrheit."

Bei der Datenanalyse ist es dabei nicht viel anders wie im realen Leben. „Schmutz", „Nebel" oder „Rauschen" (welches Bild auch immer Sie für Datenverunreinigungen wählen mögen) lassen die eigentliche „Natur der Dinge" erst dann erkennen, wenn sie nicht mehr da sind. Ich verbinde daher mit diesem Buch die Hoffnung, dass Sie später feststellen können: Dieses Handbuch *kann* eine Art Filter, eine Art Türöffner sein, mit dem Sie über Fehler- oder auch Plausibilitätsanalysen die „schmutzige Schicht" von den Daten nehmen und an die wahren

Phänomene gelangen können. Ich würde mich freuen, wenn dieses Buch Anwender ebenfalls dabei unterstützen und anregen könnte, auch die (in diesem Buch notwendigerweise nur gestreiften) Möglichkeiten komplexer multivariater Statistik auszuschöpfen, zu interpretieren und wirkliche Phänomene zu entdecken. Die Sicherstellung und Gewährleistung von Datenqualität gleicht einer Kunst und ist der „eigentlichen" Analyse von Daten im Hinblick auf Aufwand und Komplexität mindestens ebenbürtig, wenn nicht sogar oft anspruchsvoller (vgl. Schendera, 2005). Nicht nur Wilcox (1999) fragt sich, wieviele Phänomene *nicht* entdeckt wurden, nur weil Anwender nicht in der Lage waren und sind, die Möglichkeiten komplexer multivariater Statistik im Hinblick auf Datenanalyse und Datenqualität anzuwenden und auszuschöpfen. Der Leser wird auch dringend auf frühzeitig einzuplanende und nicht zu unterschätzende Ressourcen für die Datenqualität hingewiesen (Kapitel 2).

Die Qualität von Daten ist kein Selbstzweck. Daten am Ende eines Qualitätsprozesses sind somit immer auch Informationen. Wenn Information die Grundlage von Wissen ist, Wissen wiederum Macht bedeutet, dann sollte klar sein, was fehlerhafte Daten bedeuten. Datenqualität kommt vor Analysequalität und ist die allererste Grundlage der Wissenskonstruktion, also der Objektivität, Zuverlässigkeit und Gültigkeit von Wissen u.a. auch mittels Forschungsmethoden und Statistik. Mit der Qualität von Daten steht und fällt demnach nicht „nur" eine Doktor- oder Diplomarbeit. Die Qualität von Daten entscheidet *auch* über den Ruf ganzer Disziplinen, Unternehmen oder auch Forschungsbereiche, über wissenschaftliche Professionalität und Glaubwürdigkeit per se (vgl. dazu auch Kapitel 19).

Der hier vertretene Begriff der „Qualität" (zur Elaboration dieses Begriffes vgl. die weiteren Kapitel) wird mithin als unabhängig von der Strategie wissenschaftlicher Forschungsausrichtung verstanden, also ob es sich z.B. um „quantitative" Ansätze oder „qualitative" Forschungssätze handelt. Diese begriffliche Dichotomie „quantitativ vs. qualitativ" ist in Bezug auf die eigentlich methodisch differenzierenden Merkmale nur artifiziell trennscharf, realiter jedoch irreführend (z.B. Kromrey, 2005). Die Methodik beider Disziplinen ist z.T. durch grundlegende Gemeinsamkeiten ausgezeichnet: Verschiedene Fragestellungen erfordern notwendigerweise unterschiedliche methodische Ansätze, was diese Disziplinen und ihre Verfahren gleichberechtigt macht und sie einander i.S. eines Methodenpluralismus konstruktiv ergänzt. Verschiedene Fragestellungen erfordern zwar eine differentielle, aber jeweils immer auch eine professionelle (also regel- und kriteriengeleitete) Vorgehensweise. Einige qualitative Ansätze skizzieren in ihrem Vorgehen sogar eine explizite Quantifizierbarkeit (z.B. Mayring, 1990), während umgekehrt quantitative Ansätze durchaus in der Lage sind, auch mit qualitativen „Daten", in unterschiedlichem Maße quantifiziert, zu arbeiten (siehe Beispiele dazu unten). Beide Forschungsrichtungen setzen z.B. selbstverständlich jeweils auch optimal hochqualitative „Daten" (Messungen, Texte, Interviews, Bilder usw.) i.S.v. Abbildungen der empirischen Realität in der Folge einer Gegenstands(vorverständnis)-Methoden-Interaktion voraus. Beiden Disziplinen sind z.B. auch Problembereiche gemeinsam, die die Qualität des jeweiligen Ansatzes einzuschränken vermögen, z.B. Bias, Subjektivität oder Selektivität (v.a. Wilson, 1981, 57–61). Qualitativ- wie auch quantitativbasierte Veröffentlichungen können bei kritischer Betrachtung z.T. massive Fehler zutage treten lassen. In wissenschaftlichen Einrichtungen ist das kritische Lesen solcher Veröffentlichungen z.T. Bestandteil der Methodenausbildung. Dem Verfasser liegen sogar Veröffentlichun-

gen vor, die z.T. mit Preisen ausgezeichnet wurden, die aber noch nicht einmal in der Lage waren, einen Pearson Korrelationskoeffizienten korrekt zu interpretieren.

Die vorgestellten Qualitätskriterien (z.B. Vollständigkeit, Einheitlichkeit, oder auch die abschließend später vorgestellten Kriterien, wie z.B. Aktualität, Menge oder Relevanz) können uneingeschränkt auch auf die Arbeit mit qualitativen Daten angewandt werden. Ein weiterer Grund ist, wie oben bereits angedeutet, dass die Semantik qualitativer „Daten" (z.B. Texte, Interviewtranskriptionen) oft auch mittels Software weiterverarbeitet werden können, sowohl „qualitativ", wie auch „quantitativ":

- z.B. in SPSS als standalone-Tool (vgl. z.B. Schendera, 2005, Kapitel 7)
- z.B. in speziellen Lösungen der SPSS Predictive Text Analytics-Familie: Je nach Fragestellung und Anwendungsbereich können Anwender z.B. die Lösungen einsetzen und kombinieren: „Text Mining" für Clementine (Aufschlüsselung auf Freitextdaten verschiedenster Art mittels „Natural Language Processing"-Techniken), „LexiQuest Mine" (Elaborieren zentraler Themen und Konzepte in großen Textmengen) „LexiQuest Categorize" (Kategorisieren von Texten) oder auch „Text Mining Builder" (für das Fein-Tuning linguistischer Analysen).
- z.B. im Zusammenspiel mit Tools von Drittanbietern, z.B. als Anwendung auf die Inhaltsanalyse, z.B. MAXqda (früher: WinMax), atlas.ti oder NUD*IST.

Für beide methodischen Zugänge (Disziplinen) gilt daher gleichermaßen: Datenqualität kommt vor Analysequalität. Die vorgestellten Kriterien sind daher auch für die Arbeit mit qualitativen „Daten" (Texten, Strings) geeignet und (je nach Kriterium, Maß und Ansatz) auch für große bis sehr große Datenmengen, wie sie z.B. im Data Mining mit Clementine anfallen können.
Das Buch wurde überwiegend auf der Grundlage von und für SPSS Syntax (derzeit Version 15) entwickelt. Einsteiger in SPSS für Windows sollten dabei wissen, dass SPSS Syntax über Mausklicks automatisch angefordert werden kann bzw. selbst einfach zu programmieren ist, vgl. dazu „Datenmanagement mit SPSS" (Schendera, 2005). Dieses Buch und „Datenqualität mit SPSS" (Schendera, 2007) wurden übrigens von Anfang an als separate, aber einander ergänzende Handbücher entwickelt. „Datenmanagement mit SPSS" bietet dabei eine tragfähige Einführung in die Arbeit mit SPSS Syntax einschließlich einem ersten Programmieren von Makros; „Datenqualität mit SPSS" umfasst darauf aufbauend eine vielschichtige und praxisorientierte Anwendung der SPSS Syntax auf die Überprüfung von Daten und ggf. der Optimierung ihrer Qualität. Erfahrene SPSS Anwender kennen vermutlich bereits die zahlreichen Vorteile der Arbeit mit SPSS (vgl. Kapitel 2). Für sie soll schon an dieser Stelle darauf hingewiesen werden, dass sie mittels SPSS Syntax die Möglichkeit haben, auch in Clementine zu arbeiten. Clementine Anwender könnten es evtl. von Vorteil finden, dass sie über die Knoten „SPSS Transform" und „SPSS Output" über SPSS Syntax und Prozeduren in der Lage sind, die Performanz von Clementine im Hinblick auf die kriteriengeleitete Überprüfung und Gewährleistung von Datenqualität deutlich zu erweitern, und zwar sowohl durch das Einbinden von einfachen SPSS Funktionen (wie z.B. REPLACE), komplexen, von Anwendern entwickelten Syntaxprogrammen, als auch über das Einbinden der Performanz spezieller SPSS Prozeduren, wie z.B. VALIDATEDATA. Da dadurch Daten, Ergebnisse und Modellierungen geliefert werden können, die auf einem systematischen Kriterienkanon für

die Qualität der zugrundeliegenden Daten aufbauen (und dies jederzeit nachweisbar ist), erhöht dies nachhaltig die Transparenz, Glaubwürdigkeit und damit auch der Professionalität des Data Mining mit Clementine. Das Kapitel zu Clementine wurde auf der Grundlage der Version 11.1 (englisch) verfasst.

In den vorangegangenen Abschnitten sind auch die Begriffe von „Wahrheit" und „Professionalität" gefallen. Die zentrale Absicht dieses Buch ist es, klare und überprüfbare Kriterien, eindeutige einzuhaltende Standards und zu unterlassende Verhaltensweisen zusammenzustellen. Diese Anforderungen sind selbstverständlich kein Selbstzweck, sondern Ausdruck einer professionellen Ethik, der Verantwortung gegenüber professioneller wissenschaftlicher Arbeit. Ich verbinde mit den zusammengestellten Kriterien, Standards und (zu unterlassenden) Verhaltensweisen auch den Wunsch, dass eindeutig und nachhaltig kommuniziert werden konnte, welche grundlegenden Erwartungen an die Qualität von Daten, Analysen und Forschung im Allgemeinen gestellt wurden und auch zukünftig gestellt werden.
In einem Punkt bitte ich um Nachsicht und möchte deutlich hervorheben: Trotz der Fülle des zusammengestellten Materials kann es sich *nicht* um eine erschöpfende Darstellung handeln. Im Gegenteil, es handelt sich nur um eine, durchaus von einer gewissen Subjektivität geleiteten *Auswahl* der wichtigsten Aspekte. Über die Auswahl der Aspekte, wie auch die ihre jeweilige Gewichtung ließe sich sicherlich diskutieren, wozu ich alle Anwender ausdrücklich auffordern möchte.
Ich möchte bereits an dieser Stelle nochmals alle Leser einladen, durch Rückmeldungen und Vorschläge zur Ergänzung oder Verbesserung dieses Buches in zukünftigen Auflagen beizutragen (vgl. Kapitel 21).
Auf keinen Kompromiss würde ich mich jedoch einlassen wollen, würde ein Leser annehmen, dass er bzw. sie nach dem Lesen dieses Buch *alles* über die Qualität von Daten, Analysen oder auch Forschung wüsste. Ich muss dieser Annahme deutlich entgegentreten: Im Gegenteil, dies ist nur ein Anfang, ein erster Versuch einer explorativen Systematisierung, ein erster Ausgangspunkt für interessierte Leser und professionell arbeitende Wissenschaftler.

执子之手,与子偕老---献给我珍爱的中国公主 .

Ebenfalls gewidmet meinen Großeltern Lore & Albert Schmid und Klara Schendera

Zu Dank verpflichtet bin ich für fachlichen Rat und/oder auch einen Beitrag in Form von Syntax, Daten und/oder auch Dokumentation unter anderem: Prof. Gerd Antos (Martin-Luther-Universität Halle-Wittenberg), Prof. Johann Bacher (Johannes-Kepler-Universität Linz, Österreich), Prof. Vijay Chatterjee (Mount Sinai Medical School, New York University, USA), Prof. Mark Galliker (Universität Bern, Schweiz), Werner E. Helm (FH Darmstadt), Prof. Jürgen Janssen (Universität Hamburg), Raynald Levesque (Boucherville QC, Canada), Prof. Roderick J.A. Little (University of Michigan USA), Prof. Daniel McFadden (University of Berkeley USA), Dr. James W. McNally (University of Michigan USA), Prof. Theo Van der Weegen (Radboud Universität Nijmegen, Niederlande), Prof. Rainer Schlittgen (Universität Hamburg), Dr. Jonathan D. Shanklin (Head of Meteorology & Ozone Monitoring Unit,

British Antarctic Survey, Madingley Road, Cambridge, England, United Kingdom), Prof. Stephen G. West (Arizona State University USA), Matthew M. Zack (Centers for Disease Control, Atlanta, Georgia (USA).
Mein Dank gilt auch Herrn Alexander Bohnenstengel, sowie Frau Sabine Wolfrum und Frau Ingrid Abold von der Firma SPSS GmbH Software (München) für die großzügige Bereitstellung der Software und der technischen Dokumentation. Herrn Dr. Schechler vom Oldenbourg Verlag danke ich für das Vertrauen, auch dieses Buch zu veröffentlichen, sowie die immer großzügige Unterstützung. Volker Stehle (Eppingen) gestaltete die Druckformatvorlage. Stephan Lindow (Hamburg) entwarf die Grafiken. Falls in diesem Buch noch irgendwas unklar oder fehlerhaft sein sollte, so liegt die Verantwortung alleine beim Autor.
Forschen zu dürfen ist ein *Privileg*. Forschen zu können ein *Wert*.
Verleihen Sie beidem Nachhaltigkeit.

Heidelberg, Juli 2007
CFG Schendera

Inhalt

1 Einführung: Sechs erste Kriterien

Datenqualität ist essentiell und allgegenwärtig

- z.B. in Regierungen, Ämtern und Behörden (Office of Management and Budget, 2007, Chapter III, 2006; US Census Bureau, 2006; United Nations, 2003, 1995, 1983; OECD, 2003; Eurostat, 2004, 2003, 2002, 1999/1998; Statistische Ämter des Bundes und der Länder, 2003; Körner & Schmidt, 2006; Blanc et al., 2001 uam.)
- z.B. in Banken und Unternehmen (Lee et al., 2006; Konno, 2006; Willeke et al., 2006; Bettschen, 2005; Goerk, 2005, 2004; Infeld & Sebastian-Coleman, 2004; McKeon, 2003; Wan et al., 2002; Ofori-Kyei et al., 2002; English, 1999; Internationaler Währungsfond (IMF, z.B. Carson & Liuksila, 2001; Carson, 2000) uam.)
- z.B. in Forschungsverbänden (z.B. DFG, 2005, 1998; DeGEval, 2004, Bundesärztekammer, 2003; Wilkinson & APA Task Force on Statistical Inference, 1999 uam.).

Die Qualität von Daten ist jedoch nichts Selbstverständliches. Nur wenn die Qualitätskriterien bzw. die zu ihrer Überprüfung erforderlichen Maßnahmen bekannt sind, kann die Qualität von Daten hergestellt werden. Anders formuliert: Nur wer weiß, wonach er suchen soll, hat auch die Chance, die Fehler in seinen Daten zu finden und (hoffentlich) rechtzeitig zu korrigieren. Dieses Buch geht von der Annahme aus, dass die Qualität von Daten nicht etwas automatisch Gegebenes, sondern ein Produkt ist, das nur durch explizite Definitionen, Maßnahmen, sowie Kriterien gezielt herbeigeführt und nachgewiesen werden kann.
Datenqualität ist ein genuin interdisziplinäres Thema. Die im Buch vorgestellten Qualitätskriterien stammen u.a. aus der Umfrageforschung (primäres Ziel: Datenerhebung), aber auch dem Bereich der Data Warehouses (primäres Ziel: Datenhaltung). Die Statistik liefert jeweils weitere Kriterien, Annahmen und dazugehörigen Hintergrundkonzepte. Entsprechend heterogen ist die Literatur, vgl. z.B. Batini & Scannapieco, 2006, Kap. 2; Lee et al., 2006; Gackowski, 2004; Laliberté et al.; 2004; Fugini et al., 2002; Pernici & Scannapieco, 2002; English, 2002, 1999; Helfert et al., 2001; Carson & Liuksila, 2001; Carson, 2000; Berry & Linoff, 2000; Naumann & Rolker, 2000; Brackstone, 1999; Chapman et al., 1999; Garvin, 1998; Wang & Strong, 1996 uam.
Die folgende Anekdote mag ein banales Beispiel sein, ist aber aufschlussreich für ein notwendigerweise sinnvoll aufeinander abzustimmendes Zusammenspiel von Datenqualität, Data Warehouses und Expertise bereits in grundlegenden Konzepten der Statistik. In einem DHW eines großen Unternehmens war die Implementierung eines Prüfalgorithmus geplant, der auf dem Nullhypothesen-Test basiert. Die Überlegung des verantwortlichen Projektlei-

ters war dabei folgende: Wenn ein bestimmter Wert sich zu sehr von allen übrigen Werten im DWH unterscheidet, erreicht der Wert statistische Signifikanz und kann dadurch als Fehler identifiziert werden. Übersehen wurde dabei, dass das DWH Millionen an Einträgen enthielt und annähernd jeder geprüfte Wert alleine schon aufgrund der riesigen Datenmenge „Signifikanz" erzielte (vgl. zum Signifikanz-Konzept auch Kapitel 19.3.), weniger wegen des absoluten Abweichens von den übrigen Werten. Nach Rücksprache mit statistisch erfahrenen Data Analysten wurde ein Prüfalgorithmus implementiert, der fehlerhafte Werte unabhängig von der Menge der Daten erfolgreich identifizierte. Der Fairness halber muss betont werden, dass der verantwortliche Projektleiter top-level Managementinformationen „aufsaß", in denen die Funktionsweise des klassischen Signifikanztests nicht korrekt dargestellt war.

Je nach primärem Ziel gibt es derzeit noch anwendungsbedingt unterschiedliche Akzentuierungen inerhalb der einzelnen Systematisierungen und Hierarchisierungen in den jeweiligen Disziplinen. Es kann jedoch angenommen werden, dass sie sich im Laufe der Zeit annähern werden, sind sie doch letztlich Elemente ein und desselben Prozesses: Eine Datenerhebung erfordert immer eine Datenhaltung für die gewonnen Daten. Eine Datenhaltung setzt somit immer voraus, dass überhaupt eine Datenerhebung stattfand bzw. erhobene Daten vorliegen. Beide Seiten setzen jedoch voraus bzw. setzen sich idealiter proaktiv dafür ein, dass Datenqualität gegeben ist. Die meisten Methodologien, sei es als systematische Erfassung der subjektiven Einschätzung der Datenqualität durch die Anwender (z.B. als IQA Survey), sei es als objektive Prüfregel auf der konkreten Datenebene (z.B. als Codd Integrity Constraint) prüfen mehr oder weniger denselben Kriterienkanon (z.B. Lee et al., 2006, Kap. 3, 4, sowie Appendix 3; Batini & Scannapieco, 2006, Ka. 7.2.; Laliberté et al.; 2004; Long et al., 2004, 201–203; Gackowski, 2004; Statistics Canada, 2000; Carson, 2000; Davies & Smith, 1999c, d). Auch augenscheinlich theoretisch divergente Hintergrundrahmen, wie z.B. empirische, semiotische, ontologische oder auch TDQM (Total Data Quality Management) Ansätze weisen in Bezug auf die konkreten Kriterien weite Überschneidungen auf, die sich dann aber im Einzelnen dann z.T. doch wiederum völlig unterschiedliche Definitionen aufweisen (vgl. Gackowski, 2004, 127–130). Die Terminologie für die Kriterien im Einzelnen, sowie als Oberbegriff, ist derzeit noch alles andere als einheitlich. Diverse Veröffentlichungen verwenden z.B. den Begriff „Dimension" (u.a. Lee et al., 2006; Batini & Scannapieco, 2006, z.B. Kap. 2; Redman, 2004, 2001), manche dagegen *auch* „Attribut" (Gackowski, 2004; O'Brien & Marakas, 2003) usw. Diese Begriffe repräsentieren jedoch (statistische) Konzepte, die in Bezug auf Komplexität über einzelne, eher aus der Informatik stammende (techno)logische Begriffe, wie z.B. des datensatzzentrierten Terms „integrity constraint," z.T. deutlich hinausgehen.

Sechs erste Kriterien

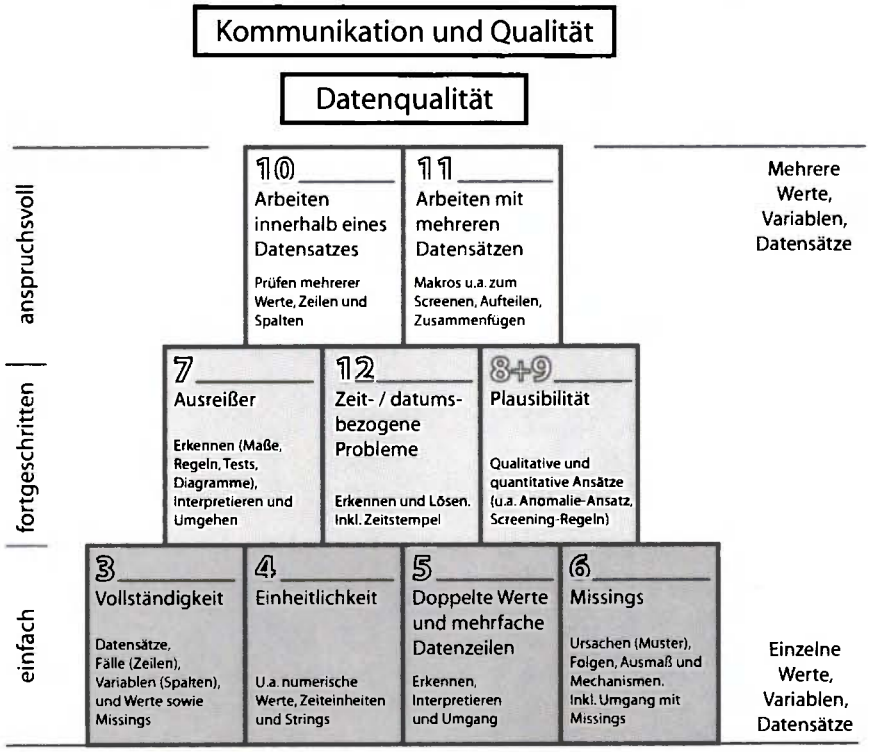
Von den zahlreichen Qualitätskriterien (vgl. zu weiteren Kriterien die Kapitel 13 und 19) werden zunächst Vollständigkeit, Einheitlichkeit, Missings, Doppelte, Ausreißer, sowie Plausibilität vorgestellt. Viele weitere Kriterien bauen auf diesen auf. Die Kriterien „Vollständigkeit", „Einheitlichkeit", „Doppelte", sowie „Missings" bilden z.B. eine Basis. Die Überprüfung dieser Kriterien setzt jedoch unter Umständen das Gegebensein weiterer Rahmenbedingungen voraus, z.B. dass u.a. ein Plan und genug Zeit für die Überprüfung mit SPSS vorliegen (vgl. Kapitel 2). Vor dem Einsatz von SPSS wird empfohlen, sich einen Überblick über die anderen Kriterien und ihr Erfülltsein zu verschaffen. Die in den folgenden Kapiteln vertieften sechs Kriterien für Datenqualität

- Vollständigkeit (Kapitel 3) bzw.
- kontrollierte Missings (Kapitel 6),
- Vermeidung von doppelten Daten (Kapitel 5),
- Einheitlichkeit (Kapitel 4),
- Beurteilung von Ausreißern und (Kapitel 7)
- Plausibilität (Interpretierbarkeit) (Kapitel 8)

sind eng miteinander verwoben. Diese Kriterien, sowie weitere, die in Kapitel 13 vorgestellt werden, bilden zusammen die Grundlage der Checkliste in Kapitel 18.

Vollständigkeit ist z.B. so definiert, dass die Anzahl der Daten in einem schlussendlich vorliegenden Analysedatensatz exakt der Summe der gültigen und fehlenden Angaben in einer strukturierten Umgebung entspricht, also z.B. in einer Erhebung (Fragebogen) oder auch den Teildatensätzen, aus denen er gebildet wurde. Fehlen z.B. in einem Fragebogen Angaben, so sollten diese kontrollierten Missings im Datensatz entsprechen. Liegen im Analysedatensatz mehr Daten bzw. Missings vor, als erhoben bzw. geliefert wurden, so wird von (zu vermeidenden) Doppelten gesprochen. Unkontrollierte Missings und Doppelte sind nach der Gewährleistung von Vollständigkeit sorgfältig zu prüfen.

Für viele Analyse-, Transformations- und Prüfmaßnahmen ist die Einheitlichkeit der Daten, v.a. wenn sie aus verschiedenen Quellen stammen, eine unabdingbare Voraussetzung. Erst wenn die Einheitlichkeit der Daten gegeben ist, z.B. von Währungen oder Zeiteinheiten, ist eine Überprüfung auf Ausreißer möglich.
Das folgende Diagramm soll das Konzept vorstellen, dass diesem Buch zugrunde liegt.



Die Kriterien „Vollständigkeit", „Einheitlichkeit", „Doppelte", sowie „Missings" bilden die Basis (vgl. die unterste Ebene im obigen Diagramm). Alle weiteren Kriterien bauen darauf auf (fur andere Hierarchien vgl. z.B. Gackowski, 2004, 139–140; Oliveira et al., 2005). Die Pyramide ist somit von unten nach oben zu lesen; die Basis selbst ist von links nach rechts zu lesen (wie auch die darüberliegende, nicht jedoch die dritte Schicht): Vollständigkeit ist Voraussetzung fur die Gewährleistung von Einheitlichkeit, den Ausschluss von doppelten Zeilen bzw. Werten, sowie den angemessenen Umgang mit möglichen Missings.

Wird die Pyramide von unten nach oben zu lesen, macht sie evident, dass solange z.B. keine vollständigen, einheitlichen usw. Daten vorliegen, es wenig Sinn macht, Daten auf weitere Kriterien zu überprüfen (z.B. Plausibilität), geschweige denn inferenzstatistische Analysen durchzufuhren oder etwaige „Ergebnisse" (voreilig) zu kommunizieren. Wird die Pyramide von unten nach oben gelesen, kann sie somit eine erste Orientierung sein, die eigenen Datenprüfungen in sinnvoll aufeinander aufbauende Schritte zu organisieren. Die Pyramide kann jedoch auch von oben nach unten gelesen werden: Angefangen von der „Krone" des professionellen wissenschaftlichen Arbeitens, einer Veröffentlichung oder eines Berichts z.B. an das Management oder Vorgesetzte, kann bereits die Publikation bzw. der Report selbst auf das Einhalten von Qualitätsstandards überprüft werden, bis hinunter zum konkreten professionellen Umgehen mit weiteren Kriterien, wie z.B. Ausreißern, Missings oder auch Vollständigkeit. Mit der „bottom-up"-Perspektive wird die Hoffnung verbunden, dass diese Pyramide bereits den *Prozess* des wissenschaftlichen Arbeitens und Kommunizierens in Bezug auf

ausgewählte Aspekte der Qualität von Daten optimiert bzw. eine gewisse Sicherheit verleiht. Die „top-down"-Perspektive kann u.a. helfen, ein *Produkt* wissenschaftlichen Arbeitens und Kommunizierens (ggf. vor einer Veröffentlichung bzw. Weiterleitung) zu evaluieren und ggf. rechtzeitig zu optimieren (zu einer differenzierteren Interpretation der Pyramide in Bezug auf das Kriterium „Plausibilität" wird auf Kapitel 8 verwiesen).

Dieses Diagramm versucht auch zu verdeutlichen, nach welchen Dimensionen dieses Buch aufgebaut ist. Das Buch ist nach *Qualitätskriterien* unterteilt: Jedes Kapitel stellt ein Qualitätskriterium vor. Kapitel 3 behandelt z.B. die Vollständigkeit, Kapitel 4 die Einheitlichkeit usw. Manche Kapitel behandeln dasselbe Kriterium, die Kapitel 8 und 9 befassen sich z.B. jeweils mit der Plausibilität. Kapitel 8 führt in das Grundprinzip und zunächst einfache Ansätze ein; das Kapitel 9 stellt den Einsatz deutlich anspruchsvollerer Screening-Programme vor. Manche (spätere) Kapitel behandeln die Arbeit mit mehreren Kriterien, bzw. stellen weitere Kriterien vor (z.B. die Kapitel 9, 13, 16 und 19). In Kapitel 13 werden Kriterien vorgestellt, für die der Einsatz von SPSS nicht notwendigerweise erforderlich ist, z.B. Menge, Eindeutigkeit, Relevanz, Genauigkeit oder auch Verständlichkeit; nichtsdestotrotz sind diese Kriterien für die professionelle Arbeit mit Daten relevant. In Kapitel 19 werden z.B. Kriterien für die *Kommunikation* der Qualität von Daten, Erhebungen und Analysen vorgestellt. Jedem Kapitel zu den Qualitätskriterien wird eine solche Pyramide vorangestellt, um die Vernetzung des vorgestellten Kriteriums (hervorgehoben durch einen dezenten 3d-Effekt) mit den anderen nochmals zu veranschaulichen. Eine weitere Dimension des Aufbaus des Buches ist die *Komplexität* der SPSS Programme; diese wurde, sofern möglich, immer von einfach bis anspruchsvoll geordnet. Mit Komplexität in Zusammenhang stehen auch die *Anwendungsmöglichkeiten* von SPSS; diese beginnen bei einzelnen Werten, Variablen und Datensätzen und laufen auf das durchaus anspruchsvolle Arbeiten mit bzw. Programmieren für mehrere Werte, Variablen bzw. Datensätze hinaus (z.B. die Kapitel 10 bzw. 11). Weil die Pyramide nach Qualitätskriterien aufgebaut ist, enthält das Diagramm nicht alle Kapitel dieses Buches, z.B. werden nicht die Kapitel zu Clementine (Kapitel 16), MacIntosh (Kapitel 17), sowie zur Kommunikation von Daten- und Analysequalität (Kapitel 19) wiedergegeben. Kapitel 12 ist das einzige Kapitel bzw. Kriterium (für *zeit- bzw. datumsbezogene* Probleme), dessen Abfolge in der Hierarchie der Pyramide nicht mit der Abfolge im Inhaltsverzeichnis des Buches übereinstimmt. Dieses Kapitel hat deshalb eine Sonderstellung, weil der Faktor „Zeit" nicht immer geprüft werden muss. Damit ist gemeint: Kapitel 12 ist nur dann erforderlich, wenn die zu prüfenden Daten tatsächlich auch Zeit- bzw. Datumsvariablen enthalten. In diesem Falle sind die Kriterien aus Kapitel 12 *vor* der Überprüfung der allgemeinen Plausibilität im Sinne der Kapitel 8 bzw. 9 anzuwenden (vgl. Diagramm). Da Datensätze jedoch nicht immer Zeit- bzw. Datumsvariablen enthalten, wurde das Kapitel 12 nach den Kapiteln 10 bzw. 11 platziert, weil es u.U. weniger häufig eingesetzt wird als die anderen Kapitel (vgl. Inhaltsverzeichnis).

Plausibilität

Alle Maßnahmen zur Gewährleistung von Datenqualität streben zusammen zunächst das Ziel der Plausibilität an. Plausibilität ist so definiert, dass der Inhalt der Daten in einem Datensatz nicht nur exakt dem Inhalt einer bereits vorliegenden Datenhaltung oder auch Angaben in einer strukturierten Erhebung entspricht, also z.B. in einem Datensatz oder einem Fragebo-

gen, sondern qualitativ darüber hinausgeht, indem z.B. auch versehentlich (oder auch absichtlich) gemachte Fehler identifiziert, korrigiert oder entfernt werden konnten. Daten mit höherer Qualität sind somit genauer und semantisch eindeutiger interpretierbar. Abgeleitete Informationen sind exakter und besser begründbar bzw. ohne Unschärfe und Unsicherheit. Daten am Ende eines Qualitätsprozesses *sind* Informationen.
Plausibilität ist am anspruchsvollsten zu prüfen und setzt (wie in der Pyramide zu sehen) u.a. Vollständigkeit, Einheitlichkeit und weitere formale Korrektheit der Daten voraus (tatsächlich können die diversen Kriterien im Prinzip konkreten „Graden" von „Plausibilität" zugewiesen werden, vgl. Kapitel 8). Die Plausibilität der zu prüfenden Daten setzt die Plausibilität *anderer* Daten zur Rekonstruktion von Missings oder zur Überprüfung von Ausreißern voraus. Jeder Ansatz zur Sicherung von Plausibilität basiert darauf, dass die Daten, die die eigentliche Entscheidungsgrundlage für die Plausibilität bilden, selbst plausibel sind. Insofern sind solche Prüf- und Korrekturmaßnahmen mit großer Sorgfalt vorzunehmen. Mit der Plausibilität von Daten, Analysen, sowie Ergebnissen ist jedoch noch nicht die Arbeit im Hinblick auf Qualität beendet. Qualität muss auch korrekt kommuniziert werden (vgl. Kapitel 19); es ist unbedingt zu vermeiden, dass suboptimale Kommunikation einen dilettantischen Eindruck vom eigentlich professionellen Arbeiten vermittelt.

Die Überprüfung der Datenqualität mit SPSS geht von der Annahme aus, dass Datenqualität objektiv definierbar, messbar und somit nachweisbar ist. Vollständigkeit ist z.B. objektiv definierbar, messbar und nachweisbar: Entweder liegt in einem Feld ein Eintrag vor (i.S.v. gültiger Wert) oder nicht (i.S.v. Missing). Dasselbe gilt für alle Werte einer Datenhaltung: Entweder ist eine Datenhaltung zu 100% vollständig oder nur zu 95%, 88% usw. Dieselbe Sichtweise lässt sich auf doppelte Einträge usw. übertragen: Entweder eine Datenhaltung enthält keine unzulässigerweise Doppelte oder falls doch, dann in einem objektiv und genau quantifizierbaren Ausmaß. Anders sieht es jedoch mit den Anforderungen an die Qualität von Daten aus. Anforderungen an Daten können durchaus subjektiv sein und von optimalen objektiven Größen abweichen. Je nach Anwendung kann z.B. eine Datenhaltung aus der Sicht eines Entscheiders nur zu 95% vollständig sein, ist nach Maßgabe eines Sachverständigen bei bestimmten Größen eine gewisse Streubreite zulässig usw. Die *subjektiven* Anforderungen an Datenqualität sollten jedoch nicht mit ihrer *faktisch objektiven Messbarkeit* verwechselt werden. Welche Gründe, Kriterien oder Ursachen auch immer von Anwendern vorgebracht werden, um die Abweichung von einer optimalen Datenqualität zu erklären, die Konsequenzen sind dieselben:

- Jede Abweichung von einer perfekten Datenqualität ist immer eine suboptimale Datenqualität.
- Eine suboptimale Datenqualität ist daher immer mittels der vorgebrachten subjektiven Gründe, Kriterien oder Ursachen seitens der Anwender zu erklären bzw. zu verantworten.
- Eine (sub)optimale Datenqualität ist kein Selbstzweck, sondern eine Eigenschaft mit Folgen.

Datenqualität: Definition
Als klassische Definition für Datenqualität gilt „data to be of high quality if they are fit for their intended uses in operations, decision making and planning" (Juran & Godfrey, 1999[5], 2;

vgl. dazu die Definition des US Census Bureau, 2006, 1: „The Census Bureau defines quality as 'fitness for use'."). So zutreffend diese Definition in ihrer Allgemeinheit ist, so ungenau ist sie gleichermaßen, da diese Definition nicht angibt, welches *Kriterium* die Qualität von Daten ausmacht (z.B. Genauigkeit, Einheitlichkeit usw.) und auch, in welchem Ausmaß.
Die Qualität von Daten ist ein mehrstelliges Relationsprädikat, das sich aus der Art und Anzahl der *erforderlichen* Kriterien („Kriterienkanon"), den Methoden ihrer Prüfung, den dabei eingesetzten Toleranzen/Grenzwerten der jew. Kriterien (z.B. 100%, 3Sigmas usw.), sowie den auch jeweils *ausgeschlossenen* Kriterien usw. ergibt. Der Anwender hat dabei nicht nur das Optimum zu definieren; dies ist selbstverständlich. Die eigentliche Herausforderung wird jedoch sein, jede (ggf. durchaus notwendige) Abweichung vom Optimum zu vermeiden, falls aus pragmatischen Gründen jedoch nicht unumgänglich, zu begründen und die damit verbundenen Konsequenzen zu rechtfertigen. Um hier einem Missverständnis gleich von vorneweg entgegenzutreten: Datenqualität ist nicht relativ. Jeder Kontext mag durchaus eine *gegenstands- bzw. datenangemessen notwendigerweise eigene* Definition von Datenqualität vertreten. Die jeweilige Definition und Gewährleistung von Datenqualität sind zentrale Aufgaben des Anwenders. Die einzufordernde Begründung dient dem Nachweis der Gegenstandsangemessenheit der Qualitätsdefinition und mithin auch dem Nachweis ihrer Objektivität, also in der Praxis auch dem Ausschluss willkürlicher Maßnahmen. Dieses Buch wird daher keine Definition der Qualität von Daten i.S. eines Globalkriteriums vorgeben, jedoch ein Rahmenkonzept bzw. einen Kriterienkanon, die es erlauben sollten, jegliche Definition von Datenqualität zu analysieren. Der Ausgangspunkt hierfür ist die Gesamtheit aller Kriterien, optimaler Methoden bzw. Toleranzen/Grenzwerten als Ziel. Jedes Ausklammern von Kriterien, jeder Einsatz (Verzicht) auf Prüfmethoden, wie auch jedes angewandte Toleranzkriterium ist jedoch explizit transparent zu machen und zu begründen.

Datenqualität vor Analysequalität I : Erfahrungen mit DWH
Datenqualität kommt vor Analysequalität. Dass schmutzige Daten verzerrte Ergebnisse nach sich ziehen, ist sachevident und kann, je nach Art der „Verschmutzung", z.B. für Vollständigkeit (Kapitel 3), Doppelte (Kapitel 5), sowie Ausreißer (Kapitel 6) an zahlreichen spezifischen Beispielen ausführlich demonstriert werden. Ohne konkrete Art, Ursache und Ausmaß der Datenfehler zu kennen, sind ihre Auswirkungen uneinheitlich und nicht vorhersagbar (z.B. Haughton et al., 2003, 69–75 anhand einer Simulation mittels Kreditdaten). Entscheidungen auf der Grundlage solcher Daten können somit entspechend völlig inkorrekt sein. Datenqualität schafft erst die Grundlage und Voraussetzung für sinnvolle Analysen. Die Sicherung und Gewährleistung von Datenqualität ist nicht selbstverständlich; Komplexität oder Aufwand können die der geplanten „eigentlichen" Analysen ohne weiteres übertreffen. In manchen Projekten sind die Phasen der Datenmanagement und Datenbereinigung um ein Vielfaches aufwendiger als die Analyse selbst. Berry & Linoff (2000, 177) bezeichnen daher „dirty data" als den „curse of data mining". In Data Warehouses entfallen ca. 90% des Zeitaufwandes auf die Datenvor- und -nachbearbeitung (Cabena et al. 1998, 43); circa 70% der Kosten werden z.B. durch Maßnahmen zur Sicherung der Datenqualität verursacht. Die dazu erforderlichen Ressourcen und Fähigkeiten sollten keinesfalls unterschätzt werden. Eckerson (2002) zitiert eine Studie des Data Warehouse Institute, wonach „managing data quality and consistency" und „reconciling customer records" zu den größten Herausforderungen in

CRM-Projekten gehören. Data Warehouses scheitern oft an einer nicht ausreichenden Datenqualität. Die FleetBoston Financial Corp. (USA) scheiterte z.B. 1996 trotz eines millionenschweren Budgets daran, Kundendaten zusammenzuführen, weil unterschätzt wurde, wie schwierig und zeitaufwendig es sein kann, Daten aus 66 Systemen zu vereinheitlichen und zusammenzuführen (Eckerson, 2002, 7). Neben Einschnitten in das eigentlich zur Verfügung stehende Budget ist die Gewährleistung von Datenqualität das Top 2 Problem (Peterson, 2003). Der Verfasser hat z.B. ein anfangs völlig verfahrenes Projekt übernommen, bei dem die erforderlichen Maßnahmen der Qualitätssicherung zum Aufwand der eigentlichen Analyse schlussendlich in einem Verhältnis von ca. 20:1 standen. Ursache waren v.a. Datensätze und -banken mit uneinheitlichen Formaten und Inhalten. Eine Herabsetzung zu „Fußarbeit" ist daher eine völlige Verkennung des Stellenwerts dieser Tätigkeit und ihrer Komplexität (vgl. auch Lee et al., 2006, 2). Ein Anliegen dieses Buches ist es daher auch, diese zu oft vernachlässigte Tätigkeit und ihre Bedeutung in den Vordergrund zu rücken.
Ein professionell durchgeführtes Datenqualitätsprojekt ist demgegenüber durchaus in der Lage, über einen Zeitraum von mehreren Jahren mehrere hundert Millionen an Euros einzusparen. Batini & Scannapieco (2006, 188–199) führen z.B. einen Fall an, bei dem ein Budget von ca. 6 Millionen € für Architektur und Wartung über einen Zeitraum von drei Jahren Einsparungen von mindestens von 600 Milionen € ermöglichen sollte, je nach Rechenweise sogar bis zu 1,2 Milliarden €.

Datenqualität vor Analysequalität II : Erfahrungen in der Forschung
Forscher betreiben oft größten Aufwand, um ihre Analysen möglichst anspruchsvoll auswerten und publizieren zu können. Im Gegensatz dazu steht ein oft erschreckend naiver und konzeptloser Umgang mit ihrem Fundament, der Qualität der Daten. Man könnte fast formulieren, als ob manche Anwender auf diesem Auge (der Datenqualität) blind sind und nur die Qualität der Analyse im (anderen) Auge haben. Der Verfasser könnte in diesem Zusammenhang zahlreiche, ziemlich drastische Beispiele aus der eigenen Erfahrung vorbringen.
Der Verfasser wurde z.B. vor nicht allzu langer Zeit beauftragt, die Sicherung und Gewährleistung der Datenqualität einer von Dritten betreuten multizentrischen klinischen Studie mit einer mehrjährigen Laufzeit zu prüfen. Der konkrete Anlass war, dass wichtige Veröffentlichungen anstanden, aber per Zufall Widersprüchlichkeiten in den Daten festgestellt wurden, die man weder zu prüfen, noch zu erklären verstand. Nach einer mehrere Tage dauernden Prüfung von Daten und Dokumentation stand fest, dass die Studienleiter und ihre Mitarbeiter fast alle Fehler begangen hatten, die sie hatten begehen können. Als massivste Fehler fielen sofort u.a. doppelte Datensätze und -zeilen auf, aber auch Datenlücken im Masterdatensatz trotz kompletter Einzeldatensätze. Unbemerkt wurden auch Kodierungen für Missings als reale Werte in die Analyse einbezogen. Beim Überprüfen der Programme für Datenmanagement und Datenanalyse (sog. Code-Review) im Detail fielen weitere Fehler logischer oder statistischer Art auf. Am Ende kam eine seitenlange Liste an Fehlern und Versäumnissen zusammen, die jedoch gleichzeitig die Voraussetzungen für einen konstruktiven Anfang schuf. Anhand dieser Liste wurden im Rahmen eines trouble-shooting nacheinander alle möglichen Fehlerquellen systematisch eliminiert. Analysen der nunmehr qualitativ optimierten Datenlage ergaben dann abschließend zahlreiche aufschlussreiche Effekte, deren Nuanciertheit und Differenziertheit in zahlreichen internationalen Veröffentlichungen Eingang

fanden. Während dieses Projekt noch mit einem „blauen Auge" davonkam, sind dem Verfasser allerdings aufgrund eigener Reanalysen weitere Projekte bekannt, deren nicht korrekte Ergebnisse z.T. bis in Veröffentlichungen auf Regierungsebene einfließen konnten. Manche Projekte konnte man nur noch als unrettbar gescheitert erklären. Über die Konsequenzen mag sich jeder Leser selbst so seine Gedanken machen. Die eingangs skizzierten Beispiele verdeutlichen hoffentlich, dass man nicht davon ausgehen kann, dass die Daten in Ordnung sind, nur weil man seine Daten oder auch ihre Transformation nicht überprüft. Mangelnde Datenqualität ist keine Fiktion, sondern kann jederzeit und in jeder Größenordnung passieren, von der studentischen Hausarbeit bis hin zum Data Warehouse.

Lohnt sich der Aufwand für Datenqualität? Kosten-Nutzen-Überlegungen

„Ob sich der Aufwand lohnt?", sicher ist die nächste, aber eindeutig absurde Frage. Man schnallt sich im Auto ja auch deshalb an, weil es sicherer ist, aber nicht deshalb, weil ein Unfall passieren wird (dieser Logik dürften sich auch Gurtgegner anschließen). Natürlich lohnt sich der Aufwand. Daten mit höherer Qualität sind genauer und interpretierbarer. Enthalten die Daten z.B. geldwerte Informationen, dann gewährleistet Datenqualität genauere Analysen (z.B. Kundengruppen, Kundenbeziehungsphasen), konkretere Vorhersagen (z.B. Produktakzeptanz, künftiges Kaufverhalten), begründbare Entscheidungen, letztlich einen Wettbewerbsvorteil. Jede Ungenauigkeit kostet bares Geld, z.B. wenn z.B. ein Unternehmer mit ungenauen Absatzdaten oder Einkaufsprofilen seiner Kunden arbeitet und diese Informationsunschärfe in Strategien einfließen sollte. Datenqualität führt zu mehr (weil korrekter) Information und somit auch zu mehr Umsatz und Gewinn. Datenqualität hilft, den Nutzen-Faktor zu optimieren und unterstützt eine präzisere Identifikation von potentiellen Kunden (Target-Marketing, Up-/Cross-Selling), erhöhte Trefferquoten beim Direkt-Marketing, und somit Senkung von Werbekosten. Datenqualität schützt somit auch vor Fehlverhalten und Fehlinvestitionen. Im CRM (Customer Relationship Management) verhindert Datenqualität z.B., dass mit Kunden falsch umgegangen wird. Im Marketing verhindert Datenqualität z.B., dass Aktionen auf der Basis doppelter, ungenauer oder fehlerhafter Informationen einen Imageschaden produzieren oder Interessenten abschrecken (z.B. Anschreiben mit falsch geschriebenen Nachnamen, Bewerbung unerwünschter Produkte oder Leistungen, oder auch Ausstellen fehlerhafter Rechnungen). Optimierte Prozesse, Produkte und Strategien unterstützen Nachfrage bzw. Angebotsakzeptanz, Kundenzufriedenheit und Kundenbindung. Fehlgeleitete Marketinginvestitionen (z.B. durch mehrfaches Mailing an dieselbe Adresse) können unmittelbar eingespart, Verluste durch falsch berechnete oder übermittelte Preise direkt verhindert werden (vgl. für zahlreiche Beispiele u.a. Redman, 2004, 2001; Eckerson, 2002 passim). Mangelhafte Daten verursachen auch fehlerhafte Analysen, die wiederum zu falschen Entscheidungen und letztlich einer Verweigerungshaltung der Entscheider gegenüber den Datenhaltungen führen (Helfert, 2000, 65).

Das Arbeiten mit qualitativ hochwertigen Daten erhöht nicht nur die Transparenz innerhalb des Unternehmens, sondern führt auch zu einem Qualitätsbewusstsein, das sich wiederum positiv auf das Arbeits- und Geschäftsklima bis hin zum Interessenten auswirken wird. Besonders eine nach außen getragene Dokumentation von Qualitätssicherungsmaßnahmen und ihre Erfüllung unterstützt das Vertrauen von Interessenten in die Datenqualität. Der Return of Invest (ROI) zeigt sich also in zwei Seiten: Vermeiden von Kosten (Fehlinvestitionen, Fehl-

entscheidungen) und Erhöhung des Nutzen (Gewinn, Wettbewerbsvorteil, Konkurrenzfähigkeit). Der Aufwand für die Gewährleistung von Datenqualität lohnt sich also solange, wie diese Maßnahmen Fehlinvestitionen und Fehlentscheidungen verhindern und Gewinne und Wettbewerbsvorteile maximieren helfen. Der ROI für die Sicherung von Datenqualität dürfte sich nicht nur bei Unternehmensdaten schnell amortisiert haben. An Datenqualität sind letztlich eindeutige materielle Konsequenzen gekoppelt, vor allem auch die Glaubwürdigkeit zur Befähigung des eigenen Tuns. Um abschließend einmal Fakten zu nennen: Redman (2004) entwickelte z.B. im Laufe der Jahre den COPDQ-Faktor (COPDQ, Cost of poor data quality) und geht davon aus, dass der COPDQ-Faktor die Kosten verzehnfacht. Wenn z.B. Maßnahmen mit Daten, die in Ordnung sind, 1€ kosten, dann würden diesselben Maßnahmen mit Daten, die nicht in Ordnung sind, 10€ kosten. Redman schätzt jedoch, dass die wahren Kosten deutlich höher liegen und dass suboptimale Datenqualität die Industrie mindestens 10%, wenn nicht sogar zwischen 20% an Einkünften kostet. Das Data Warehousing Institute schätzt, dass z.B. amerikanischen Firmen alleine 2001 durch „schmutzige" Daten über 610 Milliarden Dollar an Einnahmen entgangen sind; die wahren Kosten seien jedoch wahrscheinlich viel höher anzusetzen (Eckerson, 2002, 5; Peterson, 2003).

Ursachen suboptimaler Datenqualität

Die Ursachen suboptimaler Datenqualität sind vielschichtig und v.a. bei zunehmend großen heterogenen Datenhaltungen auch technisch gesehen ausgesprochen komplex. Bei Data Warehouses sind u.a. folgende Ursachen möglich (vgl. auch Lee et al., 2006, 80–108; Eckerson, 2002):

- suboptimale (zu restriktiv/zu tolerant), nur temporär arbeitende oder gar keine Prüfalgorithmen oder –tools.
- subjektive Faktoren bei der Datenlieferung: z.B. un-/absichtliche Verzerrungen.
- schwammige Definitionen („fuzzy definitions"): In Datenhaltungen muss mit präzisen Datendefinitionen gearbeitet werden. „Alt" kann jedoch z.B. relativ sein. Im Lebensmittelhandel werden z.B. Waren möglichst schnell umgesetzt, weil sie sonst nicht mehr verzehrt werden können. Eine fehlerhafte Definition von „Alter" in der Datenbank eines großen Lebensmittelhändlers führte dann jedoch dazu, dass auch Weine noch schnell zu einem günstigen Preis verkauft wurden. Hochklassige „alte" Weine gelangten wegen diesem Fehler zu einem Bruchteil ihres eigentlichen Werts in den Verkauf.
- hohe Volatilität des Inhalts der Datenhaltung: Bestimmte Datenarten sind invariant, z.B. Geburtsdaten, Veröffentlichungs- oder Gründungsdaten), während andere dagegen hochvolatil sein können (z.B. Aktienkurse, Fahrpläne, Umsatzdaten). Personendaten unterliegen z.B. einer permanenten Dynamik (z.B. Heirat, Umzug, Adressänderung usw.).
- uneinheitliche Repräsentation von Daten: z.B. bei Namen in der Form Vorname-Mittelname-Nachname im Gegensatz zu Nachname-Vorname-Mittelname.
- komplexe Repräsentationen von Daten: z.B. in Form von Grafiken oder ausführlichen Texten anstelle von komprimierten (numerischen) Kodierungen.
- gestiegene oder inhaltlich geänderte Anforderungen an die Daten: Lässt die Datenhaltung keine flexible Änderung von Daten zu, besteht das Risiko, dass Daten „provisorisch" an das DWH übergeben werden, mit der Folge, dass u.a. Inkonsistenzen auftreten können.

- verschiedene Speicherorte bzw. –formate für Daten: z.B. Ablage von Daten u.a. in dBASE, Microsoft Access, ORACLE und Microsoft Excel (z.T. verursacht durch einen Umstieg auf eine andere Software, durch geänderte Organisations- oder Abteilungsstrukturen usw.).
- unzureichende Hard- und Software: z.B. unzuverlässige Rechner (Server, Systeme), Datenübertragungen, ungenaue Statistikprogramme oder Algorithmen (z.B. Keeling & Pavur, 2007; Knusel, 2005; McCullough & Wilson, 2005, 2002, 1999; McCullough, 1999, 1998; Davies & Smith, 1999b).
- Konflikte innerhalb und zwischen Hardware- und Software: Die Funktionalität von Zugriffen (z.B. SQL-Queries) kann z.B. von Hardware, Betriebssystem oder auch dem Format der Daten abhängen, auf die zugegriffen werden soll. Die Daten können somit selbst in Ordnung sein, jedoch kann ihre (technisch gesehen: komplexere) Umgebung die Probleme verursachen, v.a. wenn es sich um (geographisch) verteilte Systeme handelt.
- riesige Datenmengen: Je mehr Daten sich z.B. ansammeln, desto weniger Zeit bleibt oft für gründliche Prüfung und Implementierung langfristiger Lösungen für die Gewährleistung von Datenqualität.
- unzureichendes Verständnis von Datenqualität (u.a. von Eingebern, Anwendern oder Entscheidern): z.B. die Identifikation, Interpretation und Behebung möglicher Fehler (z.B. Laliberté et al., 2004, 16), sowie in Bezug auf realistische, aber durchaus komplexe Kosten-Nutzen-Beurteilungen (vgl. auch Lee et al. 2006, 13–26; Batini & Scannapieco, 2006, 88–95; Eppler & Helfert, 2004).

Reichweite von Wirkungen: Über Signifikanzen hinaus (Fact vor Fiction)
Über Ursachen, Dunkelziffern und Folgen unentdeckter Fehler zu spekulieren ist müßig. In jedem Fall wurde die Richtigkeit des eigenen Tuns entweder (vom jeweiligen Anwender oder von anderen) nicht hinterfragt oder Korrekturprozesse wurden (vom jeweiligen Anwender oder von anderen) nicht (effektiv) daraufhin überprüft, ob sie denn richtig oder erfolgreich waren.

In den Grundlagenwissenschaften dürfte z.B. interessieren, wie die Anwender und andere den erzielten Ergebnissen vertrauen können. In einem auf „publish or perish" aufbauenden Wissenschaftsbetrieb ist nichts so sehr gefürchtet, als wenn nach der Veröffentlichung festgestellt wird, dass die Ergebnisse auf Daten- oder einfachsten Rechenfehlern beruhen (vgl. auch Schendera, 2006; Höding et al., 2005; Friedmann et al., 2004; McKinsey & Company, 2004). Enthalten die Daten z.B. klinische Informationen, dann riskieren Patienten, Ärzte und Pharmaunternehmen bei fehlerhaften Daten nicht selten darüber hinaus Ruf, Vertrauen und Gesundheit, was bei einer verantwortungsvoll gesicherten Datenqualität von vorneherein ausgeschlossen ist und oft genug der Nutzen die Kosten überwiegt (vgl. Richardson & Chen, 2001; McFadden, 1998; Gassman et al., 1995).

Effekte können also nicht bloß in Richtung Alpha- oder in Richtung Beta-Fehler gehen, also Unterschiede vorgaukeln, wo keine sind, oder Effekte verdecken, die höchst interessant wären (vgl. auch Haughton et al., 2003, 72). Bei Gewährleistung der Datenqualität hat man auch nicht bloß die Gewissheit, dass die vorliegenden Ergebnisse (unabhängig davon, ob sie in Richtung Signifikanz oder Nichtsignifikanz gehen) nicht durch eine mangelhafte Daten-

qualität beeinträchtigt sind. Andere haben nicht nur die Sicherheit, ihre Entscheidungen auf der Grundlage zuverlässiger Informationen treffen zu können. Manche szientistisch-beschönigenden Formulierungen gaukeln vor, dass evtl. nicht gegebene Datenqualität nur über feine Nuancen in der Korrektheit von Überlegungen oder Hypothesentests entscheidet, die auch durchaus hätten anders, aber zumindest hätten ähnlich ausfallen können. Um es in aller Deutlichkeit zu sagen: „Suboptimale" Datenqualität kann auch bedeuten, dass die Daten und alles, das darauf aufbaut, völliger Schrott sein kann. Unrettbar verloren, nicht recyclingfähig, eine völlige Verschwendung an Ressourcen mit Konsequenzen bis in den Regress oder Rechtsstreit.
Mangelhafte Datenqualität kann mitunter weit verheerendere Auswirkungen haben: Falsch konvertierte Daten verursachten z.B. den Absturz der ersten Ariane 5 Rakete am 4. Juni 1996. Falsche Adressdaten verursachten den versehentlichen Angriff der chinesischen Botschaft im Kosovokrieg am 8. Mai 1999 durch die NATO. Aufgrund eines falschen Signals verlor die US-Weltraumbehörde Nasa im November 2006 den Kontakt zur Mars-Sonde „Mars Global Surveyor". Die Bundesagentur für Arbeit meldete z.B. in den Monaten Dezember 2006 bis April 2007 falsche Arbeitslosenzahlen (SPIEGEL ONLINE, 2007). Falsche Daten gefährden Jahresberichte von Unternehmen und schrecken u.U. Investoren ab, gefährden das Vertrauen in demokratische Wahlen (z.B. Auszählung von Wählerstimmen), den Kampf gegen Armut, Terrorismus oder Kriminalität usw. (vgl. auch Atherton, 2007; Beikler, 2005; HEISE online news, 2005; Klotz, 2005; IMF Survey, 2004; Redman, 2004, 2001; Bange & Schinzer, 2001; Eurostat, 1999/1998).

Plausibilitätsanalysen machen nicht nur Sinn: Plausibilitätsanalysen schaffen Sinn. Plausibilitätsanalysen schaffen erst die Glaubwürdigkeit von Daten. Ohne Datenplausibilität keine Analyse. Plausibilitätsanalysen machen Sinn für Datenmengen jeder Größenordnung bis zum Data Warehouse: Bei sehr kleinen Datensätzen schließen Plausibilitätsanalysen aus, dass sich ein einzelner Datenfehler massiv nachteilig auf die Ergebnislage durchsetzen kann. Bei sehr großen Datensätzen schließen Plausibilitätsanalysen aus, dass sich scheinbar vereinzelte Fehler zu unüberschaubaren Fehlinvestitionen aufsummieren. Nicht nur bei sehr großen Datenhaltungen hat Datenqualität Priorität. Die Gewährleistung und Bereitstellung von Datenqualität ist ein Element von Phase 2 („Datenverständnis") der CRISP-DM Methodologie (vgl. SPSS, 2000; Chapman et al., 1999). Erfahrungsgemäß hat man es (auch) im Data Mining mit Daten ungeprüfter Qualität zu tun, bei denen sich die diversen Fehler und Probleme oft erst im Moment der Anlayse bzw. Modellierung zeigen (Berry & Linoff, 2000, 177–181). „Data Cleaning" gehört zu den Best Practises, das Unterlassen zum eindeutigen „Mistake" (Rud, 2001). Plausibilitätsanalysen bewahren das professionelle Data Mining vor einem Data Mülling und schaffen die Grundlage für glaubwürdige Ergebnisse, Überlegungen und Entscheidungen. Und das Beste ist: Datenqualität ist keine Glaubenssache, sondern eine überprüfbare Tatsache. Kriterien der Datenqualität (z.B. Vollständigkeit) können objektiv und messbar sein (entweder liegt z.B. ein Wert vor oder nicht). Wie, zeigen die folgenden Maßnahmen zur nachweisbaren Gewährleistung einer optimalen Datenqualität in den am häufigsten auftretenden Problembereichen Vollständigkeit, Einheitlichkeit, Doppelte, Missings, Ausreißer und Plausibilität. Daten, die erfolgreich diesen Maßnahmen unterzogen wurden, können begründet und nachweisbar als von hoher Qualität bezeichnet werden.

Optimale Datenqualität spielt vor allem bei Studien mit extremen Konsequenzen (Letalität), Aufwand (Ressourcen) und Phänomenen in Grenzbereichen eine fundamentale Rolle. Möglichen Konsequenzen von suboptimaler Datenqualität lassen sich am Beispiel eines Falles zeigen, das sicher auf optimaler Datenqualität basiert. Im Jahr 2004 wurde z.B. im Skandal um das Schmerzmedikament Vioxx und dessen Nebenwirkungen (u.a. Verdoppelung von Herzattacken und Schlaganfällen bei einer Medikamenteneinnahme von mehr als 18 Monaten) ein großer Pharmakonzern von Konsumenten und Angehörigen tatsächlicher und angeblicher Todesopfer bei der US-Börsenaufsicht SEC wegen „falschen und fehlleitenden Angaben bezüglich des Sicherheitsprofils" angezeigt. Nur wenige Minuten nach der Meldung rauschte der Aktienkurs des Unternehmens in den Keller. Einige Wochen später schaltete sich sogar das US-Justizministerium ein. Laut der US-Gesundheitsbehörde Food and Drug Administration (FDA) gab es zwar bei der Zulassung des Medikaments ein sogenanntes „Signal" auf eine potenzielle Nebenwirkung, aber in einem statistisch nicht signifikanten Ausmaß. Einige Monate später schätzt die FDA, dass alleine in den USA zwischen 88.000 und 140.000 Menschen durch die Einnahme von Vioxx seit seiner Zulassung im Jahre 1999 schwere Herz-Kreislauf-Krankheiten davongetragen haben könnten (vgl. auch Graham et al., 2005). Ende September 2004 wurde das Medikament vom Markt genommen. Laut eigenen Angaben sieht sich der Pharmakonzern mit mehreren tausend Klagen konfrontiert. Im August 2006 wurde der Arzneimittelhersteller von einem US-Gericht nach einer Klage zu einer Millionenentschädigung verurteilt, weil er Medizinern wissentlich falsche Angaben gemacht und fahrlässig gehandelt habe, weil er nicht ausreichend über die Risiken bei der Einnahme von Vioxx aufgeklärt habe. In einem weiteren Fall hob das Oberste Gericht des US-Bundesstaates New Jersey ein zugunsten des Pharmakonzerns ausgefallenes Urteil auf und ordnete eine neue Verhandlung an. Das Gericht begründete sein Urteil damit, dass seit dem ersten Urteil neue Beweise dahingehend aufgetaucht seien, dass der Arzneimittelhersteller gewusst habe, dass Patienten auch dann Herzinfarkte erleiden konnten, wenn sie Vioxx weniger als 18 Monate eingenommen hatten.

Die Literatur zu Datenqualität ist heterogen, meist theorielastig und nicht selten auch widersprüchlich. Analog zum Impetus der sog. Good Clinical Practise, Qualitätsprobleme zu identifizieren und zu beheben (z.B. in Gestalt der ICH Guidelines), gibt es immer wieder Bestrebungen, im Zusammenhang mit Qualitätssicherung und -management auch Normen für Datenqualität (inter)national festzulegen, z.B. in Gestalt von DIN- bzw. ISO-Standards (z.B. ISO 9000, ISO 8402). Bisher hat sich jedoch kein Standard durchsetzen können, weil es anscheinend schwierig ist, sich auf eine Auswahl und Hierarchisierung von Qualitätskriterien zu einigen und diese dann auch als verbindlich durchzusetzen. Das u.a. von der Bundesärztekammer (2003) herausgegebene „Curriculum Qualitätssicherung / Ärztliches Qualitätsmanagement" betont z.B. einerseits, dass „klinische Studien ein Höchstmaß an Datenqualität erfordern" (vgl. p. 78), definiert jedoch andererseits an anderer Stelle Datenqualität ausgesprochen nichtssagend als die „Eigenschaft eines Datums in bezug auf die Gütekriterien Objektivität, Validität und Reliabilität" (p. 72). Wie später auch zu sehen sein wird, ist diese Definition eindeutig nicht korrekt. Die ISO 9000 ist z.B. eine Normenreihe (ISO 9001 – ISO 9004, vgl. auch: ISO 10011–ISO 10013) mit Empfehlungen und Standards zum Qualitätsmanagement. Konkrete Vorgaben zur Qualität eines Produkts oder einer Dienstleistung und den vorzunehmenden Maßnahmen gehören nicht zum Inhalt dieser Normen. Die Definition

des Begriffs der Qualität sowie die zur Zielerreichung erforderlichen Kriterien und Prozesse legt die zu zertifizierende Einrichtung fest. Die ISO 9000 besagt also nur, dass eine zertifizierte Einrichtung lediglich in der Lage ist, gewisse Abläufe und Dokumentationsformen anzubieten, jedoch frei in der eigenen Ausgestaltung ist und auch nicht dazu verpflichtet ist, diese einzuhalten.

Das nächste Kapitel wird nun erste Hinweise für die Planung eines Datenqualitätsprojektes einschl. der Priorisierung erster Teilschritte geben.

2 Zielsetzung, Konzept und Grundlagen

Grundsätzlich: Stellen Sie sicher, dass Ihnen für Ihre Arbeit die notwendige Zeit, Ruhe und Ressourcen zur Verfügung stehen (z.B. Batini & Scannapieco, 2006, 7.5.; Lee et al., 2006, Kap. 2, 7 und 10; Redman, 2004, 2001; OECD, 2003; Berry & Linoff, 2000, Part Three; Kimball & Merz, 2000, Kap. 15; Calvert & Ma, 1996, Part II).

Zielsetzung und Konzept
Bevor Sie sich an die Arbeit machen, stellen Sie sicher, dass Zielsetzung und Konzept vorliegen und akzeptiert sind. Als übergeordnetes Ziel kann z.B. die Verbesserung eines Produktes, einer Dienstleistung oder eines Prozesses (Behandlung, Marketing, Produktion) gesetzt sein. Als konkretes Ziel ist dann gemeint, dass die zur Verfügung stehenden Daten zu Produkt, Dienstleistung oder Prozess auf ihre Qualität hin geprüft und optimiert werden. Dieser Vorgang der Prüfung der Datenqualität kann somit auch nur von jemand mit Sachnähe durchgeführt werden. Dieses Expertenwissen ist dabei einerseits notwendig für die Beurteilung der inhaltlichen Plausibilität von Daten über ein Produkt, eine Dienstleistung oder eines technischen bzw. Geschäftsprozesses (Abgleich mit inhaltlichen Erwartungen). Andererseits ist dieses Expertenwissen notwendig für das Verstehen des Zustandekommens (Erfassung, Transformation, Migration usw.) der Daten selbst und ihrer Variations- und Fehlerquellen. Datenqualität als Wert ist das Ziel und somit gleichzeitig auch Anreiz bzw. Motivation.

Überblick, Priorisierung und Konfliktlösung
Verschaffen Sie sich einen Überblick über die vorzunehmenden Prüfungen und Datentransformationen. Schaffen Sie dabei Prioritäten. Jeder Anwendungsbereich kann dabei andere Prioritäten haben. Generell haben *relevante* Daten Vorrang vor *interessanten* Daten. In der klinischen Forschung haben z.B. sog. Primärvariablen Vorrang vor Sekundär-, sowie weiteren Variablen. In Data Warehouses haben u.a. Daten mit hohem ROI Vorrang vor Daten mit niedrigem ROI usw. Versuchen Sie mit allen Parteien eine allgemein akzeptierte Prioritätenliste abzusprechen. Da unterschiedliche Wünsche (schnelle Maßnahmenumsetzung vs. gründliche Maßnahmen) i.S.v. unterschiedlichen Kriterien (timeliness vs. consistency) oft konfligieren können bzw. ein und derselben Anforderung oft unterschiedliche Prioritäten zugewiesen werden (z.B. 100% vs. 75% Elimination von Inkonsistenzen), ist dies in der Praxis nicht immer ganz einfach (vgl. weiter unten auch die Anmerkungen unerfüllbaren

Maimalforderungen und ihrer Auflösung). Ein damit zusammenhängendes Problem ist, dass Kriterien der Datenqualität zwar objektiv und kontextunabhängig sein können, die Anforderungen seitens der Beteiligten an diese Kriterien jedoch oft subjektiv und kontextabhängig. Die Festlegung von Kriterien gehört zu den höchsten Anforderungen, der erfahrungsgemäß oft gleich mehrere Schwierigkeiten entgegenstehen können:

- Entscheider und manchmal auch Analysten können oder wollen nicht immer den Status Quo vor einer Datenqualitätsmaßnahme beurteilen. Ursachen sind oft mangelnde Zeit, fehlende Geduld und/oder nicht genügend technisches Verständnis für die Komplexität der erforderlichen Maßnahmen (Lee et al., 2006).
- Entscheider und manchmal auch Analysten können oder wollen nicht immer die erforderliche Komplexität der Datenqualitätsmaßnahmen kommunizieren. Besonders schwierig ist diese Situation, wenn diese Maßnahmen externen Dritten kommuniziert werden sollten, z.B. Finanzierern oder Kunden, die gar kein technisches Wissen oder oft (viel schlimmer) ein Pseudo-Verständnis mitbringen (Kromrey, 1999).
- Eine Folge ist letztlich, dass erforderliche Datenqualitätsmaßnahmen oft nicht genügend priorisiert werden können. Manche Entscheider wollen dann entweder oft alles auf einmal, und zwar möglichst sofort, was aber de facto nicht funktionieren kann. Manche Entscheider begnügen sich u.U. mit kurzsichtigen, weil scheinbar kostengünstigen Kriterien, langfristig jedoch unzureichenden und ggf. noch teureren Folgen, was nur als strategischer bzw. selbstberuhigender Pseudoaktivismus anstelle der eigentlich erforderlichen Maßnahmen verstanden werden kann. Dies kann auf Dauer ebenfalls nicht funktionieren.

Setzen Sie Prioritäten (z.B. Ersetzen von Missings) in konkrete Einzelmaßnahmen um (z.B. Ersetzen von Missings in Namen, Datumsangaben und Werten). Wenn keine Einigung auf die jeweiligen Maximalanforderungen möglich ist, dann sollten wenigstens Mindesterfüllungsanforderungen formuliert sein. Prüfen Sie Ihre Maßnahmen auf Erfolg und Akzeptanz, z.B. von Maßnahmen, Kodierungen oder Labels; z.B. ist es für international ausgerichtete Unternehmen eine Überlegung wert, ob z.B. Datumsformate zunächst auf europäisch oder amerikanisch standardisiert werden sollen.

ex post-Prävention

Die Überprüfung auf Datenqualität ist ein Prozess, der oft erst dann einsetzt, wenn die Daten bereits vorliegen, sozusagen ein ex post-Prozess. Das bei dieser Gelegenheit gewonnene Wissen kann jedoch in ein Präventionskonzept umgesetzt werden, das zukünftige Qualitätssicherungsmaßnahmen auf ein Minimum zu reduzieren erlaubt. Wenn z.B. bei der Qualitätskontrolle festgestellt wird, dass bestimmte Prozessphasen fehleranfällig sind, dann können z.B. diese Phasen gezielt optimiert werden und somit die Fehlerhäufigkeit drastisch reduzieren. Wenn z.B. festgestellt wurde, dass einer der am aufwendigsten zu korrigierende Fehler die uneinheitliche Schreibweise von z.B. Produktbezeichnungen war, dann können präventive Maßnahmen (z.B. standardisierende Kodierungen, Hinweise, Schulungen) das Fehlerausmaß ebenfalls drastisch reduzieren. Je früher Fehlerquellen identifiziert und behoben sind, desto besser.

Prozessorientiertes vs. zustandsorientiertes Vorgehen
Bei den Maßnahmen kann zwischen einem eher *prozessorientierten* oder einem eher *zustandsorientierten* Vorgehen unterschieden werden. Beim eher *prozessorientierten* Vorgehen gelangen fehlerhafte Daten über geeignete Maßnahmen (z.B. Filter) gar nicht erst in die Analyse. Beim eher *zustandsorientierten* Vorgehen werden die fehlerhaften Daten bereits in der Datenhaltung korrigiert. Der Unterschied ist, dass das *prozessorientierte* Vorgehen bei jedem Zugriff/Prozess auf die Daten abläuft (und entsprechend mehrfach Ressourcen benötigt), das *zustandsorientierte* Vorgehen dagegen nur einmal notwendig ist. Das prozessorientierte Vorgehen ist für einen einzelnen Prozesses effizienter (weil keine Filterungen usw. vorgenommen werden müssen), jedoch nicht für iterative bzw. verschiedene Abläufe. Das *zustandsorientierte* Vorgehen ist für einen singulären Prozess oft zu aufwendig, v.a. wenn die Korrektur der kompletten Datenhaltung recht umfangreich ist, jedoch nicht für iterative bzw. wechselnde Abläufe.

Relevanz
Trennen Sie ggf. in einem ersten Schritt relevante bzw. aktuelle und somit zu prüfende Daten von Daten, die weder aktuell, noch relevant sind, noch jemals in die Analyse(n) eingehen werden. Denken Sie dabei langfristig. Wenn Sie nicht sicher sind, ob bestimmte Daten nicht doch irgendwann einmal relevant werden könnten, dann ist es ökonomischer, diese bei dieser Gelegenheit gleich mit zu prüfen. Vergegenwärtigen Sie sich: Während Sie die vorliegenden Daten überprüfen, machen Sie eigentlich nichts anderes, als ein Konzept für die Sicherung der Datenqualität zu entwerfen. Der Prüfprozess, der nun durchaus Tage oder Wochen in Anspruch nehmen kann, bis die Daten einsatzbereit sind, wird beim nächsten Durchlauf nur noch wenige Stunden in Anspruch nehmen.

Zwei mögliche Vorgehensweisen
Beim Überprüfen auf Datenqualität kann oft zunächst nach folgendem (eher *datengeleiteten*) Schema vorgegangen werden. Das Grundprinzip gilt dabei für kleinere SPSS Datensätze, wie auch komplette Data Warehouses (z.B. Batini & Scannapieco, 2006, 170–188, Lee et al., 2006, 65ff.; Eckerson, 2002):

- Definition der Daten
- Erstellen eines Prüfplans (inkl. Ressourcen, Datenprozessen usw.)
- Überprüfen der Daten (anhand festgelegter Kriterien, Maßnahmen und Toleranzen)
- Diagnose der Ursachen (bei Abweichungen von Definitionen)
- Entwicklung einer Lösung für Daten und Ursachen (Planung, Kalkulation)
- Korrektur der Daten
- Überprüfen der Daten (nach Korrektur)
- Monitoring der Daten und ihrer Qualität (idealerweise inkl. Präventionsmaßnahmen und Auditing der Projektaktivitäten)

Ein (sehr einfaches) Beispiel für dieses Schema kann z.B. die Ersetzung von zweistelligen Jahreswerten durch vierstellige Jahreswerte sein (sog. Y2K-Problem). Dieses Schema kann jedoch nicht immer angewendet werden. Ein Beispiel ist z.B. die Grundlagenforschung.

Wenn hier z.B. unerwartete Daten auftreten, dann sind nicht notwendigerweise die Daten, sondern unter Umständen die dahinterstehende Theorie, die Definition der Daten, anzupassen. In diesem Falle könnte ein (eher *theoriegeleitetes*) Schema folgendermaßen aussehen:

- Definition der Daten
- Erstellen eines Prüfplans (inkl. Ressourcen, Datenprozessen usw.)
- Überprüfen der Daten (anhand festgelegter Kriterien, Maßnahmen und Toleranzen)
- Diagnose der Ursachen (bei Abweichungen von Definitionen)
- Entwicklung einer Lösung für Daten und Ursachen (Planung, Kalkulation)
- Re-Definition der Daten (z.B. anhand optimierter Theorien)
- Überprüfung der Theorie (Ursache)
- Monitoring der Daten und ihrer Qualität (Prozesskontrolle und Auditing der Projektaktivitäten)

Ein Beispiel für dieses Schema kann z.B. die geänderte Sichtweise auf die Entwicklung der Ozon-Konzentration über der Antarktis sein. Ausreißer sind demnach nicht mehr Fehler i.S.v. Abweichungen von einer erwarteten Entwicklung, sondern empirisch gültige Repräsentationen eines unerwarteten Verhaltens der (veränderten) Wirklichkeit (vgl. Kapitel 7 und 8).

Kontrolle und Protokolle

Protokollieren Sie jede Ihrer Maßnahmen so ausführlich, dass ggf. bestimmte Korrekturen auch für Dritte nachvollziehbar sind und ggf. auch wieder rückgängig gemacht werden können. Nehmen Sie an Daten keine Änderungen von Hand vor. Änderungen von Hand können Sie nicht oder nur aufwendig protokollieren und sich schon gar nicht merken. Arbeiten Sie bei der Sicherung von Datenqualität also nur mit Syntax (vgl. Schendera, 2005).

Der Hauptgrund ist: Ein Syntax-Protokoll bzw. ein Syntax-Programm 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 jedoch nicht alle SPSS-Voreinstellungen, z.B. den Umgang mit Missings. Auch ist es eine Fehleinschätzung davon ausgehen zu können, sich immer die getätigten Mausklicks merken zu können; das geht schon gar nicht in der Situation, wenn man sich verklickt hat. Die Maussteuerung ist typischerweise dafür sehr anfällig. Letztlich soll dies auch eine Schulung dafür sein, sich bei Analysen nicht ergebnisorientiert am Ausgabefenster, sondern ablauforientiert mit dem ausgegebenen Syntax-Protokoll und dem selbst entwickelten Syntax-Programm zu arbeiten. Die Ergebnisse sind erst dann einzusehen, wenn Optionen und Art ihrer Anforderung überprüft und für in Ordnung befunden sind.

- Validierung: Syntax-Steuerung 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 den im Hintergrund generierten Befehlssyntax ausgibt, den Sie dann für eigene Zwecke abspeichern, kopieren, umschreiben und was auch immer können.
- Automatisierbarkeit und Wiederverwendbarkeit: Einmal geschrieben oder gespeichert, können Sie ein Syntaxprogramm immer wieder verwenden. Darüber hinaus können Sie aus einem SPSS Programm heraus über INSERT (ab Version 13; stoppt bei Fehlern, kann aber mittels ERROR=CONTINUE auch so eingestellt werden, dass es weiterläuft) bzw. INCLUDE (alte Version, stoppt bei Fehlern) weitere SPSS Programme ausführen lassen. Über SCRIPT können Sie aus einem SPSS Programm heraus ebenfalls weitere Befehle ausführen lassen.
- 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.
- Effizienz: Sie können Programmcode 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: Syntax-Steuerung 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 im Prinzip stets selbst. Eine Anleitung für die Syntaxsteuerung ist insofern auch eine Anleitung für die Maussteuerung.
- Der Austausch von prinzipiell selbsterklärender Syntax zwischen oder innerhalb international arbeitender Forschungsprojekte erleichtert die Kommunikation, Evaluation und Interaktion, und trägt zu ihrer Präzisierung bei. Der Leistungsumfang der diversen SPSS-Prozeduren kann über die dazugehörige Syntax differenzierter erläutert werden als über Abbildungen (Screenshots).
- 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 ab dafür. Versuchen Sie das mal mit Mausklicks.....
- Protokollierung und Dokumentation: Dieser Aspekt ist nicht unwichtig, und kann v.a. bei Peer-Reviews und Evaluationen peinliche Situationen verhindern helfen. 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 sehen Sie in dieser Situation nicht unbedingt gut aus. 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 Kategorien 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 plattformabhängig. Wenn Sie andere Betriebssysteme anschaffen, können Sie SPSS-Programme weiterverwenden.
- Fehleranfälligkeit: Syntax ist generell weniger fehleranfällig, sondern 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".
- Lernerleichterung: Das erleichterte Lernen auch anderer Programmiersprachen (z.B. C, GPL, Python, .NET, SaxBasic, XML) ist ebenfalls eine zusätzlich qualifizierende Investition in die Zukunft. Kenntnisse in weiteren externen Programmiersprachen können auch für das Erweitern des Leistungsumfangs von SPSS oder auch Clementine nützlich sein. Python kann z.B. auch in Clementine zum Einsatz kommen.
- Erweiterbarkeit: Sie können den regulären Leistungsumfang von SPSS erweitern, indem Sie über vorgefertigte SPSS Makros, SPSS Programme oder Skripte zusätzliche Funktionen in SPSS integrieren (z.B. über Sax Basic).
- Customizing: SPSS ist z.B. auch über eigene Programmierungen in .NET oder Python individuell erweiterbar. Mit Python können z.B. eigene SPSS Prozeduren entwickelt werden, z.B. vergleichbar zu REGRESSION bzw. SPSS Daten zur weiteren Verarbeitung an externe Programmiersprachen übergeben werden (z.B. Levesque, 2007[4], Part II). Mittels BEGIN PROGRAM – END PROGRAM können Anweisungen und Daten an weitere Programmiersprachen übergeben werden, z.B. Python. Für Python werden dazu das SPSS-Python Integration Plug-In (ab SPSS Version 14.0.1 oder später), für andere externe Sprachen wie z.B. .NET u.a. die Installation des Microsoft .NET Frameworks benötigt.
- Clementine-Qualifikation: Von Programierkenntnissen in SPSS profitieren Sie in zweifacher Hinsicht, da Sie SPSS Programmierkenntnisse, sowie SPSS Programme auch in Clementine (vgl. dazu Kapitel 16) einbringen können. In Clementine kann SPSS Syntax über spezielle Knoten erstellt bzw. der Inhalt bereits erstellter SPSS Syntaxprogramme eingefügt und in den gewünschten Analyseablauf (Stream) integriert werden. Die Arbeit

mit SPSS Syntax in Clementine entspricht der Arbeit in der gewohnten SPSS Umgebung. In Clementine entwickelte Syntax kann wie in SPSS abspeichert und (automatisiert) wiederverwendet werden.

- Erweiterung von Clementine: Clementine berechnet als Qualitätsmaße (je nach Version) den Anteil der Missings (Vollständigkeit) in Prozent, die Anzahl der fehlenden Werte und Leerzeiche), sowie Ausreißer und Extremwerte (basierend auf der Standardabweichung bzw. dem Interquartilbereich). Über SPSS Syntax und Prozeduren (wie z.B. VALIDATEDATA) kann die Performanz von Clementine im Hinblick auf die Überprüfung von Datenqualität, sowie auch anderen Anwendungszwecke deutlich erweitert werden.

Kopien, Kopien und nochmals Kopien

Legen Sie grundsätzlich Sicherheitskopien der Originaldaten, sowie Ihrer Programme an. Arbeiten Sie grundsätzlich nur mit Sicherheitskopien. Entwerfen Sie „Worst Case", sowie „Data Recovery" Szenarios. Entwickeln Sie z.B. über einfache Backups hinausgehende Strategien. Legen Sie vor Transformationen etc. von Variablen oder Werten immer Kopien der Originalvariablen an; legen Sie nach Transformationen immer Sicherheitskopien des Outputs an, um z.B. bei Versionswechseln (Upgrades) undokumentierte Veränderungen der Softwarealgorithmen und damit Ihrer Ergebnisse entdecken zu können. Nehmen Sie Transformationen etc. systematisch und immer nur an den Kopien der Originalvariablen vor. Zur Orientierung können Sie diese Variablenkopien z.B. vor dem numerischen Suffix mit einem „T" für „Transformiert", einem „N" für „Neu", oder einem „C" für „Corrected" o.ä. versehen (vor dem numerischen Suffix deshalb, weil dieser u.U. noch für Vektoren (SPSS-Befehl VECTOR) benötigt wird). Überprüfen Sie jeden Ihrer Schritte mehrfach. Nehmen Sie von Zeit zu Zeit Vergleiche zwischen den transformierten Daten und den Originaldaten vor; bei komplexen Transformationen nach jedem Einzelschritt. Ihre Arbeit ist die Grundlage für jede darauf aufbauende Arbeit, Analyse und Entscheidung.

Vermeiden Sie ad hoc-Ansätze (Effizienz und Konsequenz)

Ad hoc-Ansätze sind Vorgehensweisen, die sich durch folgende Merkmale auszeichnen können:

- Das Vorgehen ist konzeptlos. Dem Vorgehen liegen weder Plan, logische Folge der Arbeitsschritte, noch Ziel zugrunde. Suboptimale Arbeitsbedingungen (z.B. Terminzwänge) werden z.B. auch nach der Installation völlig neuer Systeme suboptimale Datenqualität verursachen.
- Es werden nicht alle relevanten Kriterien überprüft, sondern nur eine willkürliche Auswahl. Es werden also weniger Kriterien überprüft als erforderlich wäre.
- Die zur Überprüfung der ausgewählten Kriterien eingesetzten Maßnahmen sind banal, z.B. nur ein univariates Rekonstruieren von Missings anstelle des erforderlichen multivariaten Ansatzes.
- Die eingestellten Kriterien sind zu tolerant. Anstelle von z.B. von 100% korrekten Daten werden die Daten nur zu 75% korrigiert, und dies auch ohne weitere Begründung.

- Die Ad hoc-Maßnahme(n) erfolgen über Mausführung und sind nirgendwo dokumentiert. Anstelle eines prozessorientierten Vorgehens per Syntax wird ein zustandsorientierter Ansatz per Maus gewählt.
- Die eigentlichen Ursachen der Datenprobleme werden nicht identifiziert und behoben. Die Datenprobleme treten weiterhin auf und müssen immer wieder von vorne behoben werden bzw. werden zunehmend „toleriert".

Unterlassen Sie von vorneherein ad hoc-Ansätze. Ad hoc-Ansätze sind nicht nur langfristig ineffektiv, sondern verursachen immer mehr und immer neue Kosten (Zeit, Geld). Unterstützende Ursachen können u.a. sein: eine Ignoranz gegenüber der Komplexität des Problems und einer Illusion des Geschütztseins vor seinen weitreichenden materiellen Konsequenzen, eine irrationale Fehlertoleranz gegenüber immer noch bzw. wieder auftretenden (ungeklärten) Datenproblemen, sowie auch eine gewisse Frustration anstelle der eigentlich erforderlichen Motivation.

Gehen Sie ein Datenqualitätsprojekt mit den erforderlichen Ressourcen an, als allererste zählt Professionalität. Konzipieren Sie Lösungen bzw. vermeiden Sie Probleme, indem Sie langfristig und vorausschauend planen. Berücksichtigen Sie dabei je nach Art Ihres Projekts u.a. folgende Faktoren (vgl. u.a. Lee et al., 2006; Totterdell, 2005; Dravis, 2004, u.a. Appendix A; OECD, 2003; United Nations, 2003):

- Management (Anzahl, Hierarchien, Kompetenzen, Verfügbarkeit, Unterstützung),
- Personal (Anzahl, Kompetenzen, Verfügbarkeit, Unterstützung),
- finanzielles Budget (Volumen, Verfügbarkeit, Flexibilität),
- Zeitrahmen (Wochen, Monate; Dringlichkeit, Flexibilität),
- Computer (Hardware, Programmversionen, Support),
- Daten (Volumen, Speicherort, Kontaktpersonen, Primär-/Sekundärvariablen, usw.),
- Projektphasen und –komplexität (z.B. Informationen über u.a. Projektabläufe, Qualitätskriterien, Lösungsansätze),
- bereits vorhandene Materialien (u.a. Dokumenationen oder Standards zu Datenqualität bzw. Datenproblemen), Programme (Module) oder auch SPSS Syntax),
- ggf. erforderliche Ressourcen für Monitoring und Auditing und ggf.
- zu beschaffendes bzw. einzukaufendes Know-How (Workshops, Schulungen, externer Support usw.).

Nehmen Sie im Bereich Datenqualität (zunächst) nichts für selbstverständlich an; stellen Sie grundsätzlich alles in Frage. Praktizieren Sie die Maxime: „Vertrauen ist gut, Kontrolle ist besser". Entwickeln Sie eine umfassende und detaillierte Dokumentation. Entwickeln Sie eine Ablaufstruktur (z.B. entweder nach Qualitätskriterien, nach Variablen oder Datenhaltungen). Berücksichtigen Sie im Projektablauf ggf. Alternativen. Betreiben Sie permanentes Monitoring, z.B. zwischen geplanten und tatsächlich notwendigen Projektphasen, z.B. einen Abgleich zwischen geplanter und tatsächlicher Zeit usw.

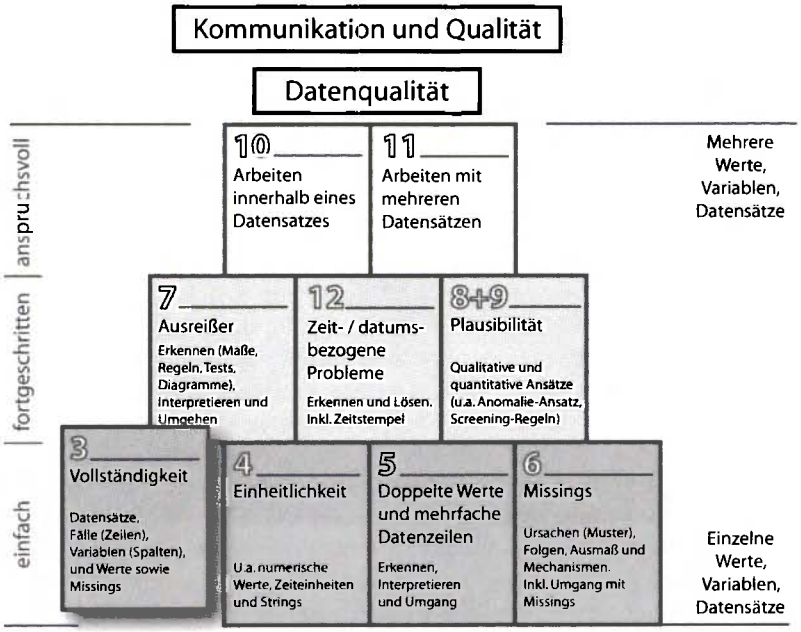
Warum Sie „Optimale Datenqualität in kürzester Zeit bei minimalem Aufwand" vermeiden sollten

Abschließend nochmals: Stellen Sie sicher, dass Ihnen für Ihre Arbeit die notwendige Zeit und Ruhe zur Verfügung steht. Stress schlägt sich immer negativ auf die Qualität von Daten durch. Vermeiden Sie ein Zerriebenwerden zwischen den drei „Mühlsteinen" eines möglichen Anspruchs in der Form (1) *optimale Datenqualität* (2) in *kürzester Zeit* (3) bei *minimalem Aufwand.* Dieser Anspruch ist leicht dahingesagt, alle drei Kriterien können jedoch unmöglich gleichzeitig erfüllt werden. Optimale Datenqualität erfordert nun mal den notwendigen (maximalen) zeitlichen, sowie materiellen Aufwand. Vordergründig können immer nur zwei beliebige Seiten dieses Anspruchs erfüllt werden, ist im Prinzip eine beliebige dritte als nicht erfüllt in Kauf zu nehmen. Da das Arbeitsziel jedoch *optimale Datenqualität* lautet, sind nur Auflösungen in Richtung (a) nur minimaler Kosten (bei maximaler Zeit), (b) nur maximaler Kosten (in minimaler Zeit) oder auch (c) maximal erforderlichen Kosten und Zeit zulässig. Wird diese Maximalforderung nicht nach einer Seite aufgelöst, endet man in maximalen Kosten (bis hin zum persönlichen Zusammenbruch), indiskutabler Datenqualität und/oder nicht eingehaltenen Deadlines. Stehen nicht die erforderlichen Ressourcen für den notwendigen Aufwand zur Verfügung, ist damit die Datenqualität jedoch noch nicht direkt gefährdet, sofern zumindest genug Zeit zur Verfügung steht. Positiv formuliert lauten die Auflösungen: Maximaler Aufwand (Kosten) und maximale Zeit bedingen maximale Datenqualität. Maximaler Aufwand (Kosten) und höchste Datenqualität bedingen nicht notwendigerweise eine lange Projektzeit. Höchste Datenqualität in maximaler Zeit verursacht nicht notwendigerweise maximalen Aufwand (Kosten). Kommunizieren Sie mit Vorgesetzten (Betreuern), Mitarbeitern oder Angestellten die drei Seiten der Maximalforderung und legen Sie eine verbindliche Auflösung fest. Ist eine verbindliche Auflösung der Maximalforderung nicht durchsetzbar, so sollten die Seiten zumindest priorisiert werden.

Haben Sie nun Ihr Projekt strukturiert und geplant, können Sie mit der Überprüfung der einzelnen Kriterien beginnen, naheliegenderweise mit dem Kriterium „Vollständigkeit" (vgl. das nächste Kapitel).

3 Vollständigkeit

Vollständigkeit ist so definiert, dass die Anzahl der Daten in einem schlussendlich vorliegenden Analysedatensatz exakt der *Summe der gültigen und fehlenden Angaben* in einer strukturierten Umgebung entspricht, also z.B. in einem Fragebogen, einer Kundendatei, Textsammlungen, Teildatensätzen usw. Diese erste Definition von externer Vollständigkeit bezieht sich auf die vollständige Abbildung *externer* Informationen durch eine Datenhaltung. Eine weitere, quantifizierende Definition von *interner* Vollständigkeit bezieht sich auf das *Verhältnis der gültigen Werte und (idealerweise: kontrolliert) fehlenden Angaben (Missings)* innerhalb der Datenhaltung. Diese Unterscheidung erlaubt, Vollständigkeit besser zu kontrollieren und zu beurteilen. Wenn z.B. alle verfügbaren Daten ohne Angaben sind, dann sind diese Daten insgesamt extern vollständig, aber intern unvollständig, also komplett leer.



„Vollständigkeit" gehört neben „Einheitlichkeit", „Doppelte", sowie „Missings" zu den grundlegenden Kriterien von Datenqualität, zu deren Prüfung SPSS eingesetzt werden kann. Die Prüfung aller weiteren Kriterien baut auf diesen auf. Vollständigkeit ist wiederum Voraussetzung für die Gewährleistung von Einheitlichkeit, den Ausschluss von doppelten Zeilen

bzw. Werten, sowie den angemessenen Umgang mit möglichen Missings. Vollständigkeit ist sozusagen die wichtigste Voraussetzung, die sine non qua-Bedingung (mit *einer* Präzisierung!): es muss sich um den *richtigen* Datensatz handeln. Den *falschen* Datensatz (u.a. auf Vollständigkeit) zu überprüfen ist einer der größten Fehler, die passieren können.

Vollständigkeit lässt sich auf mehreren Ebenen prüfen:

- Die Anzahl der einzelnen Datensätze (z.B. Updates), die in einen Masterdatensatz eingegangen sind.
- Die Anzahl der Fälle (Beobachtungen), die in Form von Datenzeilen in den Datensatz eingegangen sind.
- Die Anzahl der Variablen, die in Form von Datenspalten in den Datensatz eingegangen sind.
- Die vorhandenen (gültigen) Werte bzw. kontrollierten Missings für jeden Fall, z.B. Teilnehmer an einer Erhebung.

Anmerkung:
Im Data Mining werden oft auch Fälle als „Datensätze" bezeichnet. Dieses Buch reserviert den Begriff „Datensatz" ausschließlich für eine Datei (z.B. einen SPSS Datensatz). Für eine Datenzeile verwendet dieses Buch ausschließlich die Begriffe „Fall" bzw. „Zeile" (je nach Kontext) und nicht den Terminus „Datensatz". Der Begriff „Datensatz" ist ausschließlich für die Bezeichnung der Gesamtheit an Zellen vorbehalten (auch wenn diese Datei nur eine Zeile enthalten sollte).

3.1 Kontrollmöglichkeiten auf der Ebene der Anzahl der Datensätze

Mit Vollständigkeit auf der Ebene der Datensätze ist im Folgenden definiert, dass alle Datenhaltungen vorliegen, also z.B. alle SPSS Dateien. Die Frage, ob diese Dateien auch wie erwartet Daten enthalten, wird in einem nächsten Schritt überprüft. Zunächst wird aber überprüft, ob alle Datenhaltungen vollständig vorliegen. Die Frage, ob alle Dateien vorliegen, scheint ein triviales Problem zu sein, ist es aber nicht; wenn gleich komplette Datensätze fehlen, dann fehlen auch gleich eine Menge Daten. Die Konsequenzen können weitreichend sein. Die Bundesagentur für Arbeit meldete z.B. in den Monaten Dezember 2006 bis April 2007 falsche Arbeitslosenzahlen. Die Ursache war, dass bei der Datenübertragung ein kompletter Datensatz verloren gegangen war. Das für solche Fälle vom System erstellte Fehlerprotokoll wurde jedoch anscheinend ebenfalls übersehen (SPIEGEL ONLINE, 2007). Ein Computertechniker der Steuerbehörde von Alaska löschte im Juli 2006 versehentlich Daten im Wert von ca. 38 Milliarden Dollar; als eigentliches Problem kam hier hinzu, dass die nächtlichen Sicherungsbänder nicht lesbar waren (Sutton, 2007). Das Zentrum für das Nachrichtenwesen der Bundeswehr, gebaut für die sensibelsten Daten Deutschlands (u.a. für Geheimdienstberichte, Meldungen von CIA und anderen Nachrichtendiensten), verlor Ende 2004, nach dem gegenwärtigen Kenntnisstand im Zusammenhang mit einem defekten Daten-

sicherungsroboter, militärisch äußerst sensible Daten und löste damit einen politischen, sowie auch IT-bezogenen Skandal aus (REPORT MAINZ, 2007) (vgl. 6.1.2.).

Komplette Datensätze können schneller verloren gehen als man vermutet; mögliche Ursachen können z.B. sein: Abspeichern auf defekten Festplattensektoren, beim Versand über Email/Internet (z.B. Antivirenprogramme oder Firewalls) oder auch Konflikte mit anderer Software, z.B. Zipp-Programmen. Die verschiedenen Datensätze in einem Projekt sollten auf jeden Fall in mehr als einer Form protokolliert sein, z.B. als Syntax, Log, Projektdokumentation (Audit Trail) oder auch als regelmäßige Kopie der Festplatte oder des Servers.

Kollegen passierte in diesem Zusammenhang ein recht interessantes Problem mit einem Zipp-Programm. Ein Zipp-Programm dient im Allgemeinen dazu, die Größe von Dateien zu komprimieren. Wenn mehrere Dateien gleichzeitig gezippt werden, sind zusammengehörige Dateien auch gleich zusammen in einem Ordner, was praktisch und übersichtlich ist. Eine Abteilung eines Unternehmens packte nun mehrere Dateien in eine einzelne Zipp-Datei und schickte diese zur Überprüfung in eine andere Abteilung. Dummerweise kam die zweite Abteilung immer zu anderen Ergebnissen als die erste Abteilung. Die Aufregung war groß. Bis sich nach langwierigem Suchen und nur durch Zufall herausstellte, dass es sich zwar um die korrekten Kommunikationswege, die korrekten Analyseverfahren und auch die korrekte Zipp-Datei handelte, dass aber die Zipp-Datei selbst ein „Eigenleben" führte. Das Zipp-Programm "verschluckte" Daten. Nach dem Entzippen war, von ursprünglich drei gezippten Dateien, nur noch eine vorhanden. Wurden die Dateien dann einzeln gezippt und versandt, waren die Daten vollständig und die Analysergebnisse stimmten überein.

Die Vollständigkeit eines Datensatzes definiert sich als die Summe seiner Teile, also z.B. die Anzahl von Teildatensätzen, Updates, Variablen usw. Die Vollständigkeit von Datensätzen ist üblicherweise nicht ohne weiteres *innerhalb* einer Datenhaltung überpüfbar. Üblicherweise *sollte* auf externe Information zurückgegriffen werden können (vgl. die obigen Hinweise auf die Protokollierungen). Die Anzahl der Variablen kann je nach Anwendungsbereich auf unterschiedliche Weise überprüft werden, z.B. über den Abgleich mit einer Datendokumentation (z.B. Projektplan, Metadaten) oder den Abgleich mit anderen, bereits vorliegenden Datensätzen. Einzelne Datensätze sollten auf ein Teil-Ganzes-Verhältnis überprüft werden, z.B. wenn sich einzelnen Datensätze (sog. Masterdatensätze) aus anderen Teildatensätzen zusammensetzen. Diese Situation liegt oft dann vor, wenn z.B. innerhalb eines Unternehmens Informationen aus verschiedenen Abteilungen zusammengeführt werden, z.B. Marketing und Controlling, oder wenn z.B. Wissenschaftler mehrere Fragebögen einsetzen und diese dann bei der Analyse zusammen in einer Datei auswerten wollen.

- *Prüfmöglichkeit Augenschein:* Überprüfen Sie anhand der Variablen- und Datenansicht die Vollständigkeit aller Zeilen und Spalten per Augenschein. SPSS nummeriert am jeweils linken Rand der Variablen- und Datenansicht die vorhandenen Einträge durch. Für relativ geringe Datenmengen kann diese Methode ausreichend sein, für grosse bis sehr große Datenmengen sollten andere Ansätze gewählt werden.
- *Prüfmöglichkeit Metadaten:* Anhand von Projektdaten (z.B. Projektplan, Fragebogen, Kodierungsliste usw.) wird geprüft, wie viele Variablen der Datensatz mindestens enthal-

ten müsste. Ist z.B. die Anzahl der Variablen (Spalten) geringer, so ist der geprüfte Datensatz vermutlich nicht vollständig.

- *Prüfmöglichkeit Vergleichsdatensatz:* Der vorliegende Datensatz wird hinsichtlich seiner Vollständigkeit mit einem anderen Datensatz (z.B. aus dem Vorjahr) verglichen, z.B. hinsichtlich der Anzahl seiner Variablen (Spalten). Ist z.B. die Anzahl der Spalten geringer, so ist der geprüfte Datensatz vermutlich nicht vollständig. Wenn man es ganz genau wissen will, kann man mit SPSS auch zwei strukturell gleiche Datensätze auf absolut identische Inhalte überprüfen (vgl. 11.5.).
- *Prüfmöglichkeit Syntax:* Die (Teil)Datensätze in einem Masterdatensatz lassen sich z.B. an der ausgeführten Syntax oder auch an ID-Variablen kontrollieren. Unabhängig davon, ob UPDATE, ADD FILES oder MATCH FILES eingesetzt wurde: Am Programm selbst kann zwar abgelesen werden, welche Datensätze zusammengefügt wurden, nicht notwendigerweise, wie oft. Prüfen Sie auch, ob die Ausgabe (Log) irgendwelche Fehler rückmeldet.

Makroprogramme können ihre eigenen Tücken aufweisen, z.B. wenn zentrale Variablen für die Makrosteuerung Missings enthalten. Kapitel 11.7.1. zeigt ein Beispiel, in dem ein Datensatz z.B. deshalb nicht angelegt werden konnte, weil die erforderliche Variable einen Missing enthielt. Fügt nun ein späterer Schritt des Makros alle automatisch angelegten Dateien ebenfalls automatisch zusammen, dann durchaus, ohne dass dabei die gar nicht angelegte Datei entdeckt wird.
Selbst an einem korrekten Programm kann nicht abgelesen werden, wie oft die Datensätze zusammengefügt wurden. Ein mehrfaches Ausführen eines ADD FILES-Befehls hängt z.B. den neuen Datensatz mehrfach an den Masterdatensatz an; es besteht die Möglichkeit, dass Datensätze zu oft aufgenommen wurden. ADD FILES oder MATCH FILES erlauben nur das Handling von maximal 50 Datensätzen auf einmal; es besteht die Möglichkeit, dass die übrigen Datensätze nicht aufgenommen wurden. Um diese und andere Möglichkeiten auszuschließen, stehen weitere Optionen zur Verfügung:

- *Prüfmöglichkeit Zeilenzahl*: Enthalten die Updates immer dieselbe Zeilenzahl, kann die Zeilenzahl mit der Anzahl der Updates multipliziert werden, um zu der Gesamtzeilenzahl zu gelangen, die der Masterdatensatz aufweisen müsste. Enthalten die Update-Datensätze Kontrollvariablen (sog. „Counter" für die Nummer des Update, vgl. dazu 10.1), kann anhand der Analyse dieser Zählvariablen das korrekte einmalige (oder evtl. versehentlich mehrmalige) Zusammenfügen überprüft werden. Liegen keine Zählvariablen oder unregelmäßige Variablenstrukturen vor, können doppelte Datensätze über die Identifikation von Doppelten identifiziert und entfernt werden. Speichern Sie einen aktualisierten Masterdatensatz immer unter einem anderen Namen wie den alten Masterdatensatz ab. Es wäre verheerend, wenn der Masterdatensatz eventuell versehentlich mit den falschen Werten überschrieben werden würde, aber keine Kopie mehr zur Verfügung stünde.
- *Prüfmöglichkeit Variablen:* Die Vollständigkeit der einzelnen Datensätze kann anhand der Bezeichnung der enthaltenen Variablen kontrolliert werden. Übliche Bezeichnungen sind dabei systematisch Prä- oder Suffixe in den Datensatzbezeichnungen. Die Bezeichnungen der Datensätze aus dem Sozio-Ökonomischen Panel (SOEP) des Deutschen Instituts für Wirtschaftsforschung sind z.B. nach den Präfixen variiert, z.B. „AHBRUTTO",

„BHBRUTTO", „CHBRUTTO" usw. Die Präfixe „A", „B" und „C" repräsentieren das jeweilige Jahr der Erhebungswelle, in diesem Falle 1984, 1985 und 1986. Die restliche Zeichenfolge „BRUTTO" bezeichnet den wesentlichen Inhalt der jährlich erhobenen Datensätze, in diesem Falle der jährliche Personen-Bruttobestand. Alternativ wären Datensatzbezeichnungen denkbar, an denen die Präfixe konstant bleiben und nur die Suffixe variieren, z.B. HBRUTO84, HBRUTO85, HBRUTO86, usw. Für beide Varianten gilt: Für jeden Datensatz, der vorliegen sollte, muss auch die Bezeichnung in der Syntax vorkommen. Liegen die erwarteten Teildaten (z.B. aus dem Jahr 1985) nicht im Gesamtdatensatz vor, war entweder die Bezeichnung in der Syntax falsch oder die Syntax war richtig, aber der Datensatz versehentlich falsch bezeichnet.

- *Grenzen der Prüfbarkeit:* Eine Aussage über die Vollständigkeit einer Datenhaltung steht und fällt mit der Qualität (Vollständigkeit) ihrer Dokumentation (Metadaten). Wenn eine Datendokumentation keine Aussage über die SOLL-Vollständigkeit zulässt, dann ist eine Überprüfung nicht in der Lage, eine abweichende IST-Vollständigkeit bzw. Un- bzw. Nicht-Vollständigkeit abzuleiten.

Die praktische Arbeit mit mehreren Datensätzen gleichzeitig wird im Allgemeinen zu den anspruchsvolleren Anforderungen im Bereich Datenmanagement/Datenanalyse gezählt und daher in diesem Buch im Kapitel 11 behandelt.

3.2 Kontrollmöglichkeiten auf der Ebene der Anzahl der Fälle (Zeilen)

Dateien setzen sich u.a. aus Zeilen (Fälle, Beobachtungen, Fragebögen, usw.) zusammen, ggf. auch in Form zeilenweiser Updates. Die Kontrolle der Vollständigkeit von Spalten (Variablen) wird in 3.3. behandelt.

- *Prüfmöglichkeit Augenschein:* Überprüfen Sie anhand der Datenansicht die Vollständigkeit aller Zeilen per Augenschein. SPSS nummeriert am linken Rand der Datenansicht die vorhandenen Einträge durch. Für relativ geringe Datenmengen kann diese Methode ausreichend sein, für grosse bis sehr große Datenmengen sollten andere Ansätze gewählt werden.
- *Prüfmöglichkeit Metadaten:* Anhand von Projektdaten (z.B. Rücklaufprotokollierungen, Posteingangslisten, Strichlisten, elektronischen Protokollen usw.) wird geprüft, wieviele Fälle (Zeilen) der Datensatz mindestens enthalten müsste. Ist z.B. die Anzahl der Fälle (Zeilen) geringer, so ist der geprüfte Datensatz vermutlich nicht vollständig.
- *Prüfmöglichkeit Zählvariable:* Die Anzahl der Fälle (syn.: Zeilen, Beobachtungen, Fragebögen) kann über das Anlegen einer systematisch hochzählenden Zählvariablen (sog. Counter, vgl. 10.1) kontrolliert werden, die für jeden neu aufgenommenen Fragebogen um eine Einheit ansteigt. Entspricht die Anzahl der (vollständigen) Fragebögen dem Stand der Zählervariablen, so sind die Fragebögen vollständig eingegeben. Die Zählvariable kann ebenfalls kontrolliert werden. Sofern eine Zählvariable bei 1 beginnt und bis

zum Höchstwert lückenlos ist, braucht nur die Zeilennummer des letzten Datensatzes mit dem Counter verglichen werden; stimmen beide überein, sind die Daten lückenlos, also alle Fragebögen aufgenommen. Sofern eine Zählvariable nicht bei 1 beginnt, sind die Lücken sorgfältig zu prüfen.
Für zeilenweise Updates (s.o.) kann die Zeilenzahl mit der Anzahl der Updates multipliziert werden, um zu der Gesamtzeilenzahl zu gelangen, die der Masterdatensatz aufweisen müsste. Enthalten die Update-Datensätze Kontrollvariablen, kann anhand der Analyse dieser Zählvariablen die Vollständigkeit der zeilenweisen Updates überprüft werden (zum auch gruppenweisen Zählen von Zeilen vgl. 10.1).

- *Prüfmöglichkeit Syntax:* Prüfen Sie auch, ob die Ausgabe (Log) irgendwelche Fehler rückmeldet. SAVE bzw. XSAVE sollten z.B. nach einem Schreibvorgang die Anzahl der gespeicherten Fälle zurückmelden; fehlt diese Rückmeldung, besteht die Möglichkeit, dass der Datensatz nicht vollständig gespeichert wurde.
- *Grenzen der Prüfbarkeit*: Auch hier steht und fällt ein Vollständigkeitsurteil mit der Qualität (Vollständigkeit) der Datendokumentation (Metadaten). Wenn eine Datendokumentation keine SOLL-Aussage zur Vollständigkeit zulässt, dann ist eine Überprüfung nicht in der Lage, eine abweichende IST- bzw. Nicht-Vollständigkeit abzuleiten.
  Die Analyse von Daten aus Datenhaltungen und Erhebungen Dritter kann besondere Probleme in sich bergen. Aufgrund erfahrungsgemäß eingeschränkter Kontrolle ist es empfehlenswert, genau festgelegte, vollständige und eindeutige Absprachen zu treffen, auch über Kontrollmöglichkeiten hinsichtlich Vollständigkeit, Bias usw. Beauftragen Sie z.B. einen Drittanbieter, die Daten online zu erheben, besteht die Möglichkeit, dass er nur komplette Datenzeilen bereitstellt, aber z.B. nicht die der Abbrecher oder auch Daten, die außer der ID komplett fehlen. Wenn Sie nur die „vollständigen" Datenzeilen vorliegen haben, haben Sie nicht ohne weiteres die Möglichkeit, das quantitative und qualitative Ausmaß der Vollständigkeit abschätzen zu können. Sie wissen einfach nicht, wieviel fehlt (evtl. feststellbar an einer lückenhaften ID). Sie wissen auch nicht, was (aus welchen Gründen) fehlt. Mit der Vollständigkeit der Daten ist das Problem der Interpretierbarkeit der Daten eng verknüpft; es ist ohne einen Rückgriff auf die fehlenden Daten nicht sicher auszuschließen, dass die vorhandenen Daten einen Bias aufweisen, dessen Art und Ausmaß gerade deshalb nicht rekonstruiert werden kann, weil die Missings nicht zur Verfügung gestellt wurden. Auch die Vollständigkeit von Missings hat ihren Sinn, wie der Abschnitt zum Umgang mit Missings zeigen wird.

3.3 Kontrollmöglichkeiten auf der Ebene der Anzahl der Variablen (Spalten)

Die Vollständigkeit von Variablen (Spalten; im Data Mining auch: Felder) wird üblicherweise über die Anzahl von Variablen bestimmt. Die Vollständigkeit von Variablen ist ebenfalls üblicherweise nicht nicht ohne weiteres innerhalb einer Datenhaltung überpüfbar. Üblicherweise *sollte* auf externe Information zurückgegriffen werden können. Die Anzahl der Variablen kann je nach Anwendungsbereich auf unterschiedliche Weise überprüft werden (vgl. 3.1.):

- *Prüfmöglichkeit Augenschein:* Überprüfen Sie anhand der Variablenansicht die Vollständigkeit aller Spalten per Augenschein. SPSS nummeriert am linken Rand der Variablenansicht die vorhandenen Einträge durch. Für relativ geringe Datenmengen kann diese Methode ausreichend sein, für grosse bis sehr große Datenmengen sollten andere Ansätze gewählt werden.
- *Prüfmöglichkeit Metadaten:* Über den Abgleich mit einer Datendokumentation (z.B. Kodierungsliste, Projektplan, Metadaten usw.) wird geprüft, wieviele Variablen (Spalten) der Datensatz mindestens enthalten müsste. Ist z.B. die Anzahl der Spalten (Variablen) geringer, so ist der geprüfte Datensatz vermutlich nicht vollständig.
- *Prüfmöglichkeit Variablen:* Sind die enthaltenen Variablen nach einer bestimmten Systematik vergeben (z.B. ITEM_1, ITEM_2, usw.), dann sind Brüche in der Systematik ein Hinweis darauf, dass Variablen fehlen. Solche Brüche sind v.a. auch über das Prüfen von Syntaxprogrammen zu erkennen.
- *Prüfmöglichkeit Syntax:* Prüfen Sie auch, ob die Ausgabe (Log) irgendwelche Fehler rückmeldet.
- *Prüfmöglichkeit Vergleichsdatensatz:* Der vorliegende Datensatz wird hinsichtlich seiner Vollständigkeit (Anzahl seiner Variablen) mit anderen, bereits vorliegenden Datensätzen, oder auch über das Anlegen eines Testdatensatzes als Maske zur Eingabe für einen Fragebogen geprüft. Können anhand des Testdatensatzes nicht alle Angaben aus dem Fragebogen untergebracht werden, so ist der Datensatz nicht vollständig, sondern muss um die fehlenden Variablen ergänzt werden. Dieses Vorgehen empfiehlt sich v.a. bei Fragebögen, die leichte Variationen aufweisen (was in der Praxis allerdings von vorneherein vermieden werden sollte). Diese Vorgehensweise ist v.a. für das rechtzeitige Entdecken von Eingabeproblemen bei Mehrfachantworten zu empfehlen.
- *Grenzen der Prüfbarkeit:* Eine Aussage über die Vollständigkeit einer Datenhaltung steht und fällt der Qualität (Vollständigkeit) ihrer Dokumentation (Metadaten). Der Abgleich mit einem vollständigen Testdatensatz ist z.B. erfahrungsgemäß aufwendiger als der Abgleich mit einem Kodierungsplan, jedoch wesentlich sicherer, da der Abgleich mit einem Kodierungsplan von einer Vollständigkeit abhängt, die ja gerade überprüft werden soll. Der „Qualitätsknoten" in Clementine 9 berechnet z.B. als Qualitätsmaß u.a. die Vollständigkeit von Variablen (Feldern) in Prozent. Clementine 9 überprüft allerdings nicht, ob die *Anzahl der Variablen* vollständig ist, sondern die Lückenlosigkeit der Einträge, also im Prinzip die *Vollständigkeit der Werte.*

3.4 Kontrollmöglichkeiten auf der Ebene von Werten bzw. Missings

Die Vollständigkeit der vorhandenen Werte bzw. Missings kann über einfache Prüfprogramme ermittelt werden. Die Vollständigkeit von Werten ist in zweierlei Hinsicht zu überprüfen. Erstens, sind die Wertereihen lückenlos (also ohne Missings) und zweitens, sind auch die Daten vollständig (es fehlen also keine bestimmten Werte) (vgl. 10.2.).

Der „Qualitätsknoten" in Clementine 9 berechnet z.B. als Qualitätsmaß die Vollständigkeit von Variablen (Feldern) in Prozent, sowie die Anzahl der fehlenden Werte und Leerzeichen. Die Qualität einer Variablen wird somit über die Lückenlosigkeit der Einträge hergeleitet.

Die folgenden Prüfprogramme ermitteln z.B. die Summe der systemdefinierten Missings, das Programm darunter die Summe der system- **und** benutzerdefinierten Missings (vg. Schendera, 2005). Entspricht die Anzahl der vorhandenen Werte der Anzahl der Variablen, dann sind die Daten vollständig (vgl. 10.3.).
Ein weiterer Ansätz wäre, Quersummen zu bilden, z.B. über die Funktion SUM. Diese Quersummen können untereinander oder auch mit externen Referenzwerten (z.B. Summen anderer Datenhaltungen) verglichen werden. Übereinstimmende Summen weisen dann darauf hin, dass z.B. die gültigen Werte innerhalb der ausgewählten Variablen, die in die Summenbildung einbezogen werden, einander entsprechen und somit vollständig sind, sofern davon ausgegangen werden kann, dass die herangezogenen Referenzwerte selbst valide sind. Unterschiede zwischen Quersummen können auf fehlende Variablen bei der Berechnung und/oder fehlende Werte zurückgeführt werden. Die mit der Berechnung von Vollständigkeitsindizes zusammenhängende Problematik fehlender Werte und eines häufigen Fehlers bei ihrer Berechnung werden im Kapitel 6 zu den Missings vorgestellt.

```
count SYSMISUM=ITEM1 ITEM2 ITEM3 (SYSMIS).
exe.

count MISSUM =ITEM1 ITEM2 ITEM3 (MISSING).
exe.

compute SYSMISUM=SYSMIS(ITEM1) + SYSMIS(ITEM2) +
        SYSMIS(ITEM3).
exe.

compute MISSUM=MISSING(ITEM1) + MISSING(ITEM2) +
        MISSING(ITEM3).
exe.
```

Eine Voraussetzung für dieses Vorgehen ist, dass bereits bei der Dateneingabe fehlende Werte mit einem anwenderdefinierten Kode versehen wurden. Dieses Vorgehen gewährleistet, dass zum Zeitpunkt der Analyse zwischen Missings unterschieden werden kann, die als bekannt, also *kontrolliert*, fehlen, und solchen, die sich der Kontrolle bei der Dateneingabe entzogen hatten und nochmals überprüft werden müssten. Dieses Vorgehen könnte auf einen Index für ein Qualitätskriterium im Allgemeinen (für einen Vollständigkeitsindex vgl. 6.1.2.) bzw. die Plausibilität, also Datenqualität im Besonderen (vgl. Kapitel 8 und 9) erweitert werden. Hierbei werden die Inhalte verschiedener Variablen zunächst auf Plausibilität überprüft. Ist ein Wert korrekt, erhält eine Indexvariable einer ersten geprüften Variablen den Wert 1, andernfalls den Wert 0. Sind alle Variablen geprüft, werden alle Indexvariablen aufaddiert und durch ihre Anzahl dividiert. Der entstehende Quotient gibt das Ausmaß der korrekten Werte pro Fall (Zeile an). Andere Berechnungsvarianten sind möglich, z.B. 1 minus Anzahl der fehlerhaften Einträge dividiert durch ihre Anzahl usw. (vgl. auch Batini &