Zur Effizienz einiger Missing-Data-Techniken — Ergebnisse einer Computer-Simulation

1. Einleitung

Kaum ein Datensatz im Rahmen sozialwissenschaftlicher Forschung enthält alle Daten, die zu erheben beabsichtigt war. Das Problem unvollständiger Datensätze findet im Gegensatz zu seiner Verbreitung in der Forschungspraxis allerdings nur einen geringen Niederschlag in der Lehrbuchliteratur und in noch geringerem Ausmaß in veröffentlichten Forschungsberichten. Zwar gehört mittlerweile die zumindest formale Beachtung des damit verwandten Problems der Totalausfälle ("Unitnonresponse", Verweigerungen) zum Standard bei der Beurteilung empirischer Untersuchungen, für "partiell" fehlende Werte ("Itemnonresponse", "missing data", im folgenden "MD") gilt dies jedoch nicht. In der Forschungspraxis wird das Vorliegen eines ernsthaften "MD-Problems" häufig erst dann bemerkt, wenn ein Statistikprogrammpaket aufgrund seiner Voreinstellungen an einem Datensatz scheitert. Zumeist wird dann zu ad-hoc-Lösungen gegriffen, die weder auf statistischen Modellen beruhen noch durch eine soziologische "Theorie" gestützt werden.

Nahezu alle in der sozialwissenschaftlichen Forschung üblichen "Missing-Data-Techniken" basieren auf der Annahme, daß der Prozeß, der MD produziert, mit der "inhaltlichen" Fragestellung nicht zusammenhängt: Die Daten werden als zufällig fehlend betrachtet ("missing at random", im folgenden "MAR").[1)] Nun lassen sich aber für jeden Abschnitt der traditionellen Art der Datenerhebung und Datenerfassung (Interview, Codierung, Erfassung, Verarbeitung) plausible Vermutungen über systematische Zusammenhänge zwischen der Entstehung fehlender Werte und inhaltlichen Fragestellungen anstellen. Die Oberprüfung der MAR-Annahme mit "Tests auf MAR" (vgl. MONTMANN, BOLLINGER & HERRMANN, 1983) zeigt dann häufig, daß von "zufälligem" Fehlen in sozialwissenschaftlichen Datensätzen nicht ohne weiteres die Rede sein kann. Damit stellt sich die Frage, wie ein Datensatz, bei dem die MAR-Annahme verworfen werden muß, ausgewertet werden sollte. Das häufigste Vorgehen basiert auf einer zumeist implizit gelassenen Annahme über die Robustheit der MD-Techniken gegenüber dem Ausfallmechanismus: Bewährt sich eine MD-Technik unter der MAR-Annahme, so wird angenommen, daß diese Technik auch bei anderen Ausfallmechanismen vergleichsweise gute Resultate erzielt. Damit wird

unterstellt, daß die relative Leistungsfähigkeit der Verfahren unter anderen Ausfallmechanismen erhalten bleibt, also kein "Interaktionseffekt" zwischen Ausfallmechanismus und Leistungsfähigkeit der MD-Verfahren besteht. Um die Frage nach diesem Interaktionseffekt zu beantworten, wurde eine Simulationsstudie durchgeführt, in der eine Reihe von MD-Techniken (neben MAR) auch unter der Annahme von systematischen Ausfallmechanismen getestet wurden.

2. Missing-Data-Techniken

Aus der kaum noch überschaubaren Vielzahl von MD-Techniken (vgl. z.B. ANDERSON, BASILEVSKY & HUM, 1983) sollen hier nur einige Techniken herausgegriffen werden.

Prinzipiell kann zwischen Methoden zur Schätzung fehlender Werte und Methoden zur Parameterschätzung in Datensätzen mit fehlenden Werten unterschieden werden. Diese Unterscheidung basiert auf dem Kriterium, ob nach der Anwendung der Methode ein vervollständigter Datensatz, bei dem die fehlenden Werte durch Schätzungen ersetzt wurden (Imputation), existiert oder nicht.

Die in diesem Zusammenhang wichtigsten Parameterschätzverfahren stellen die Berechnungsmethoden von Korrelationsmatrizen bei unvollständigen Datensätzen dar. Die einfachste Methode besteht im Ausschluß aller Fälle mit fehlenden Werten aus den Berechnungen (im folgenden: "LISTWISE"). Neben LISTWISE können Korrelationsmatrizen u.a. mit den für jedes Variablenpaar getrennt berechneten bedingten Mittelwerten und Varianzen bei paarweiser Berechnung der Kovarianzen berechnet werden (im folgenden: "PAIRWISE"). Weiterhin können Korrelationsmatrizen aus den Daten berechnet werden, nachdem die fehlenden Werte durch die Mittelwerte der jeweiligen Variablen ersetzt wurden ("WILK's method", im folgenden: "MEAN").

In der neueren MD-Literatur besitzt insbesondere der sogenannte "EM-Algorithmus" als Parameterschätzmethode für unvollständige Datensätze eine erhebliche Bedeutung. Der "expectations-maximization" (EM)-Algorithmus stellt eine allgemeine Methode zur Berechnung von Maximum-Likelihood-Schätzungen dar, die sich prinzipiell auf jedes MD-Problem anwenden läßt (ORCHARD & WOODBURY, 1972; DEMPSTER, LAIRD & RUBIN 1977). Der EM-Algorithmus arbeitet iterativ; jede Iteration besteht aus einem M-Schritt, in dem die "averaged

log-likelihood" berechnet wird, und einem M-Schritt, in dem diejenigen Parameter gesucht werden, die die "averaged log-likelihood" maximieren (vgl. LITTLE, 1983:369-373). Im hier interessierenden Zusammenhang multivariater Normalverteilung kann der EM-Algorithmus mit wiederholten multiplen Regressionen folgendermaßen durchgeführt werden:

1. Berechnung einer Anfangsschätzung des Mittelwertvektors und der Kovarianzmatrix aus den Fällen ohne MD (LISTWISE);
2. Für jeden Fall mit fehlenden Werten: Partitionieren des Mittelwertvektors und der Kovarianzmatrix in einen vollständigen und einen unvollständigen Teil;
3. Schätzung der fehlenden Werte durch multiple Regression mit allen vorhandenen Variablen unter Benutzung der geschätzten Mittelwerte und Kovarianzmatrix;
4. Berechnung eines neuen geschätzten Mittelwertvektors und einer neuen geschätzten Kovarianzmatrix;
5. Korrektur der Kovarianzmatrix: Für jeden Fall mit fehlenden Werten wird die residuale Kovarianz der Variablen zu dem entsprechenden Element der Kovarianzmatrix addiert;
6. Berechnen eines Konvergenzkriteriums. Falls keine Konvergenz: Wiederholung der Schritte 2-6; sonst Abbruch.

Einige spezielle MD-Programme bieten iterative multiple Regressionsschätzungen ("iterative least squares", im folgenden "ILS") ohne die Korrektur der Kovarianzmatrix[2)]

Neben den bisher erörterten Parameterschätzverfahren existieren Verfahren, die fehlende Werte durch Schätzwerte ersetzen ("Imputation") und damit "vervollständigte" Datensätze, die keine fehlenden Werte enthalten, produzieren (da der EM-Algorithmus auch MD-Ersetzungen liefert, nimmt er hier eine Zwischenposition ein).

DEAR (1959) wird eine Ersatzmethode zugeschrieben, die auf der Berechnung der Hauptkomponenten basiert. Ausgehend von einer ursprünglichen Schätzung der Korrelationsmatrix wird die erste Hauptkomponente berechnet, und fehlende Daten werden durch die entsprechenden Werte des Falles auf die Hauptkomponenten ersetzt.

Zu den am häufigsten vorgeschlagenen Methoden der MD-Schätzung gehört die multiple Regressionsschätzung nach BUCK (1960). BUCK berechnet die Korrelationsmatrix für die vollständigen Fälle und dann für jeden Fall und für

jede unvollständige Variable eine multiple Regressionsschätzung mit allen verfügbaren Variablen des Falls. Da die vorhergesagten Werte alle auf den Regressionsgraden liegen, ist eine Verzerrung der dann aus den geschätzten Daten berechneten Korrelationsmatrix zu erwarten. BUCK entwickelte daher eine Korrektur der geschätzten Kovarianzmatrix, die allerdings in der Regel in der Literatur nicht angeführt wird. Wie BEALE & LITTLE (1975:134-137) zeigen, ist der EM-Algorithmus identisch mit der multiplen Regressionsersetzung nach BUCK, sofern diese iteriert angewandt wird.

Seit relativ kurzer Zeit finden sich in der Literatur Arbeiten, die eine Anwendung der aus der amtlichen Statistik stammenden "Hotdeck"-Techniken auf unvollständige "sozialwissenschaftliche" Datensätze versuchen. Es gibt keine feststehenden Definitionen der Hotdeck-Verfahren (FORD, 1983:185-196). Den Definitionsversuchen gemein ist die Zuschreibung des stochastischen Charakters des Verfahrens, die Verwendung von Daten des aktuellen Datensatzes zur Zuschreibung und die Betonung des Dopplungscharakters der Ersetzung. Zwei Grundtypen der Hotdeck-Verfahren können unterschieden werden: Sequentielle und simultane Verfahren. Die sequentiellen Verfahren ("traditional hot-deck", vgl. KALTON & KASPRZYK, 1982:23) beginnen mit der Festlegung der "Imputationsklassen" und einer ersten Besetzung der Zellen. Imputationsklassen stellen in Hinsicht auf die unvollständige Variable als homogen angenommene Gruppen dar, die nur "zufällig" in Hinsicht auf ihr Antwortverhalten ("Response-Nonresponse") unterschiedliche Reaktionen zeigen (das Konzept der Imputationsklassen hängt eng mit dem sogenannten MARC-Modell = "missing at random within classes" zusammen; fehlende Daten eines Nonrespondenten werden durch Daten eines "ähnlichen" Respondenten ersetzt). Für die Startwertefestlegung existieren mehrere Vorschläge. Einer der einfachsten besteht in der Zuweisung des Zellenmittelwertes aus früheren Untersuchungen. Der Datensatz wird dann sequentiell abgearbeitet. Für jeden Fall des Datensatzes wird die zugehörige Imputationsklasse mit den Klassifikationsvariablen bestimmt. Ist ein gültiger Wert für den Fall auf der zu vervollständigenden Variablen vorhanden, so wird der aktuelle Wert dieses Falles ("Donor") zum Inhalt der Imputationsklasse. Fehlt für den aktuellen Fall der Wert der Variablen, so wird dem Fall ("Rezipient" oder "Kandidat") der Wert der Imputationsklasse zugewiesen (der Name "Hotdeck" geht auf diese andauernde Veränderung des Zelleninhalts durch im aktuellen Datensatz vorhandene Werte zurück). Simultane Verfahren beruhen auf der Zuweisung von

Werten eines Falles, der die geringste Distanz in bezug auf die "matching variables" zum zu vervollständigenden Fall besitzt ("nearest neighbor rule").

3. Effizienz der MD-Techniken bei der Kovarianzschätzung

Die in der Literatur und vor allen in der Forschungspraxis vorherrschende Auffassung des MD-Problems als bloß technisches Problem führt zur Frage nach der Effizienz der MD-Verfahren. Diese Frage setzt entweder die MAR-Annahme voraus oder die Annahme, daß ein eventueller systematischer Ausfallmechanismus nicht so stark mit den interessierenden Konstrukten zusammenhängt, daß eine Änderung substantieller Beziehungen erfolgt. Daneben könnte die Annahme erfolgen, daß ein MD-Verfahren in der Lage sei, einen systematischen Ausfallmechanismus zu kompensieren. Nur unter einer dieser Prämissen stellt sich die Frage nach einem "optimalen" MD-Verfahren, die in der Literatur einen großen Raum einnimmt.

Selbst unter der MAR-Annahme sind analytische Berechnungen zur Leistungsfähigkeit der verschiedenen MD-Techniken mit extremen Schwierigkeiten behaftet (SANTOS, 1981:27, ANDERSON, BASILEVSKI & HUM, 1983:459). Es existieren demzufolge nur für sehr einfache Fälle exakte analytische Ergebnisse für Parameterschätzverfahren. Die damit verbundenen Probleme nehmen zu, wenn der Effekt auf multivariate Statistiken Gegenstand des Interesses wird, insbesondere bei den MD-Ersetzungsverfahren. Mit Ausnahme der Arbeit von SANTOS (1981) existieren für die meisten Ersetzungsverfahren keine analytischen Ergebnisse über ihren Effekt auf multivariate Statistiken, sondern lediglich für einfache univariate Statistiken (fast ausschließlich Mittelwert bzw. "totals"). Das Verhalten sowohl der Schätzverfahren als auch der Ersetzungsverfahren unter anderen Ausfallmodellen als MAR ist analytisch weitgehend unbekannt.

Das Verhalten der MD-Verfahren hängt von einer großen Zahl von Bedingungen ab, so z.B. dem Ausfallmechanismus, der Zahl der Fälle, der Zahl der Variablen usw. Um die Effekte der großen Zahl von Parametern ohne analytische Versuche dennoch zu studieren, wurden für die meisten MD-Verfahren Simulationsstudien durchgeführt. Simulationsstudien sind bei diesen Problemen die einzige Möglichkeit, zu verallgemeinerbaren Aussagen zu gelangen. Der Vergleich verschiedener Verfahren an einem Datensatz läßt die Spezifika des

jeweiligen Datensatzes unkontrolliert, Verallgemeinerungen sind nicht möglich. In noch stärkerem Ausmaß trifft dies auf Studien zu, die von vornherein mit unvollständigen Datensätzen arbeiten. In diesen Fällen ist kein Vergleich mit den "wahren" Werten möglich. Nur falls alle Bedingungen, die das Verhalten der MD-Verfahren beeinflussen, kontrolliert werden können, kann ein Vergleich verschiedener Verfahren zu prinzipiell verallgemeinerbaren Aussagen gelangen (HAITOVSKY, 1965:4; ANDERSON, BASILEVSKY & HUME, 1983:446). Der größte Teil der Literatur besteht daher aus Simulationsstudien, die das Verhalten einiger weniger Verfahren unter verschiedenen Bedingungen zeigen sollen. Mit wenigen Ausnahmen erfolgen die Studien explizit unter der MAR-Annahme; die Kritik der vorliegenden Arbeiten zeigt aber, daß selbst unter der MAR-Annahme kaum durch vorliegende Studien gestützte Aussagen über die Wahl eines "optimalen" MD-Verfahrens möglich sind (SCHNELL, 1985). Das Problem zeigt sich bei der Einbeziehung systematischer Ausfallmechanismen noch wesentlich deutlicher: Mit sehr wenigen Ausnahmen existiert hierzu keine Literatur; die Auswirkungen systematischer Ausfälle auf die meisten MD-Techniken sind unbekannt.

Es gibt bisher keine Arbeit, die einen Gesamtüberblick über die vorgenommenen Simulationen zu geben versucht, wenn auch Ansätze hierzu existieren: HAMILTON (1975:41-47); ANDERSON, BASILEVSKY & HUM (1983:459-472). Zu den einflußreichsten Arbeiten gehören HAITOVSKY (1969), TIMM (1970), GLEASON & STAELIN (1975), BEALE & LITTLE (1975), KIM & CURRY (1977) und LITTLE (1979). Die Arbeit von HAITOVSKY wird in der Regel herangezogen, um die Überlegenheit von LISTWISE gegenüber PAIRWISE zu belegen (z.B. LITTLE, 1979:76), die Arbeit von KIM & CURRY (1977), um das Gegenteil zu belegen (z.B. HOLM, 1979:152), BEALE & LITTLE (1975) und LITTLE (1979) werden zitiert, um die Überlegenheit des EM-Algorithmus zu demonstrieren. Zum Einfluß von Hotdeck-Verfahren auf die Kovarianzstruktur existieren nur zwei größere Simulationsstudien (VACEK & HIKAGA, 1980; KAISER, 1983).

Die in der Literatur zu findenden Empfehlungen (z.B. LÖSEL & WÜSTENDÖRFER, 1974; HOLM, 1979) beruhen auf der MAR-Annahme, wurden weitgehend ohne analytische Ergebnisse gegeben und basieren überwiegend auf Simulationen, die nur sehr schwer verallgemeinerbar sind. Mit Ausnahme sehr weniger Ansätze (vgl. VAN GUILDER & AZEN, 1981) existiert zur Frage der Robustheit der MD-Techniken gegenüber verschiedenen Ausfallmechanismen keine Literatur; ebenso sind Abschätzungen möglicher Verzerrungen inhaltlicher Ergebnisse durch die Existenz systematischer Ausfallmechanismen analytisch nahezu unmöglich. Um die Frage nach der Robustheit zu klären, wurde daher eine Simulationsstudie durchgeführt.

4. Design der Simulationsstudie

Die Fragestellung wurde auf zwei Aspekte eingeengt: Erstens auf eine mögliche Aussage über die Güte der Schätzung der fehlenden Werte und zweitens auf die Frage nach der jeweils erhaltenen "Korrelationsstruktur".

In der Simulation wurde entsprechend den noch darzustellenden Vorüberlegungen mit Hilfe eines Simulationsprogramms zunächst ein simulierter Datensatz erzeugt, der jeweils einem realen Datensatz, wie er in der Forschungspraxis vorkommen könnte, entsprechen sollte, dessen "Eigenschaften" aber vollständig bekannt waren, so z.B. die Populationskorrelationen. Für diesen Datensatz wurde dann die Stichprobenkorrelation berechnet und anschließend - entsprechend den zu simulierenden Ausfallmodellen - ein Teil der Daten als "fehlend" ausgewiesen. Auf diesen unvollständigen Datensatz wurden dann eine Reihe unterschiedlicher MD-Ersetzungsverfahren angewandt und jeweils für jede Ersetzungstechnik die Abweichungen der geschätzten Daten von den bekannten "wahren" Daten und die Abweichungen der aus den geschätzten Daten berechneten Korrelationsmatrizen von den bekannten "wahren" Stichprobenkorrelationsmatrizen berechnet.

Insgesamt wurden vier Faktoren, von denen ein Einfluß auf die Leistungsfähigkeit der MD-Ersetzungstechniken erwartet wurde, systematisch variiert:

- Umfang der Stichprobe,
- mittlere Interkorrelation in der Population,
- Anteil fehlender Werte pro Variable in der Stichprobe,
- der Ausfallmechanismus.

Jeder dieser Faktoren besaß drei Faktorstufen, so daß bei einem vollständigen faktoriellen Design (3x3x3x3=) 81 verschiedene Bedingungen entstanden. Für jede der Bedingungen wurden 10 voneinander unabhängige Datensätze erzeugt, insgesamt also 810 Datensätze. Die Datensätze umfaßten 100, 300 und 500 Fälle, die Zahl der Variablen wurde rein willkürlich auf 10 festgelegt. Die Variablen der Datensätze sind Realisationen von in der Population multivariatnormal verteilten Variablen, die mit einem Zufallszahlengenerator mit vorgegebener Kovarianzstruktur erzeugt wurden [3]. Für die Simulation wurden Variablen mit sieben Kategorien verwendet [4]. Zusätzlich zu den 10 "internen" Variablen wurden drei weitere Variablen erzeugt, auf denen keine

fehlenden Werte generiert wurden. Diese drei zusätzlichen Variablen werden im folgenden als "externe Variablen" (V1-V3) bezeichnet. Sie sollten übliche demographische Variablen wie "Geschlecht" und "Alter" simulieren. Für die externen Variablen wurde angenommen, daß auf ihnen keine fehlenden Werte existieren. Die Analysen der Abweichungen der geschätzten Korrelationen beziehen sich auf die (10x10)-Korrelationsmatrizen der internen Variablen mit ihren jeweils 45 Korrelationskoeffizienten. Die Interkorrelationsstruktur der internen Variablen sämtlicher Matrizen wurde als homogen vorgegeben, alle Koeffizienten nahmen damit in der Population den selben Wert an. Diese Struktur wurde wegen ihrer Einfachheit und der Nähe zu Interitemkorrelationsmatrizen gewählt. Die Interkorrelation (CM) nahm die Werte 0.2, 0.4 und 0.6 an und dürfte damit in der Regel in der Praxis vorkommenden Größenordnungen entsprechen.

Tabelle 1: Vorgegebene Korrelationsstruktur

	externe Variablen			interne Variablen				Selektionsvariable
	V1	V2	V3	V4	V5	...	V13	V14
V1		.2	.2	.15	.15	...	.15	.15
V2			.2	.15	.15	...	.15	.15
V3				.15	.15	...	.15	.15

Tabelle 1 zeigt die vorgegebene Struktur, die allen Matrizen zugrundelag. Die externen Variablen weisen eine niedrige homogene Interkorrelation von 0.2 auf und korrelieren mit den internen Varibalen jeweils mit 0.15 [5].

Drei verschiedene Ausfallmechanismen wurden simuliert. Neben MAR sollten zwei andere plausibel erscheinende Mechanismen zum Vergleich herangezogen werden; die Generierung der fehlenden Daten erfolgte für alle Mechanismen unabhängig voneinander. "Mischungen" der Ausfallmechanismen, wie sie in den meisten Datensätzen in der Praxis vorkommen dürften, wurden nicht vorgenommen. In allen drei Ausfallmodellen wurde der Prozentsatz fehlender Werte pro Variable dreistufig variiert: 1%, 5% und 10%. Die Generierung der MD erfolgte ohne Restriktionen, so daß auch Fälle mit auf allen Variablen fehlenden Werten möglich waren.

Neben einem reinen MAR-Modell wurden ein Hazard-Modell (MAT) und ein MARC-Modell simuliert. Im Hazard-Modell fehlen Daten dann, wenn eine nur indirekt gemessene Variable einen Schwellenwert überschreitet. Der Ausfall erfolgte, um den in sozialwissenschaftlichen Datensätzen unrealistischen Fall des "complete truncation" zu vermeiden, nicht deterministisch. Das realisierte Modell zog eine Zufallsstichprobe aus denjenigen Fällen des generierten Datensatzes, die einen in Abhängigkeit vom Anteil fehlender Werte bestimmten Schwellenwert auf der "latenten" Variablen überschritten; die Korrelationen mit der Selektionsvariablen (V14) wurden als in der Population gleich vorgegeben. Die Höhe dieser Korrelationen wurde jeweils mit einer leicht stärkeren Korrelation als der jeweiligen Interkorrelation (r=CM+0.1) angenommen (das Modell könnte z.B. Ausfallmechanismen bei Leistungstests, bei denen entweder oberhalb eines Leistungslevels durch wahrgenommene Trivialität des Tests oder unterhalb eines Leistungslevels durch Frustration Ausfälle auftreten, entsprechen).

Das MARC-Modell sollte Stichproben simulieren, deren Elemente aus zwei verschiedenen Populationen stammen. Die "Populationen" sollten sich sowohl in Hinsicht auf die Art und Höhe des Zusammenhangs der internen Variablen, als auch in ihren Ausfallwahrscheinlichkeiten unterscheiden. Innerhalb der Populationen wurde MAR angenommen. Im realisierten Modell wurden die Sampleanteile mit 0.8 und 0.2 vorgegeben. Das Verhältnis der Ausfallwahrscheinlichkeiten wurde ebenso willkürlich mit 1:10 festgelegt, um ein möglichst stark verzerrendes Modell zu erhalten. Die MD-Anteile wurden so bestimmt, daß sich im zusammengefaßten Sample 1%, 5% und 10% fehlende Daten pro Variable ergaben. Die Differenz der mittleren Interkorrelation wurde indirekt über die variierende Interkorrelation im Sampleteil 1 bei konstanter 0.1-Korrelation im Sampleteil 2 simuliert.

Zunächst wurden 12 MD-Verfahren ausgewählt. Die Auswahl beruhte auf zwei Kriterien: Das Verfahren sollte entweder weit verbreitet und/oder von besonderem theoretischen Interesse sein. Nach einer Reihe von Voruntersuchungen wurden zwei Versionen der Hotdeck-Verfahren ausgeschieden, deren Ergebnisse so weit unter denen der anderen Verfahren lagen, daß eine vollständige Prüfung unangebracht schien. Die folgenden Verfahren wurden damit in der Hauptuntersuchung verwendet:

- LISTWISE (vollständige Fälle)
- PAIRWISE (paarweise vorhandene Werte)
- MEAN (Mittelwertersetzung)
- CELLMEAN (Zellenmittelwertersetzung)
- HOT-DECK (Dopplungsverfahren)
- REGRESSION (Einfachregression)
- HAZARD (Zweifachregrssionmit MD-Indikator)
- DEAR (Hauptkomponentenverfahren)
- MULT (Multiple Regression, BUCK-ähnlich)
- ILS (iterierte multiple Regression, EM-ähnlich)

Einige Verfahren bedürfen kurzer Erläuterungen. MEAN verwendete alle vorhandenen Werte einer Variablen für die Mittelwertersetzung. CELLMEAN baute für jede unvollständige Variable eine Crossbreak-Tabelle (mit 2x5 Zellen) aus den vollständigen Variablen V1 und V2 auf. War die Zelle unbesetzt, wurde für die entsprechende Zelle der 10-Zellen-Tabelle die zweite Variable kollabiert. HOTDECK benutzte ebenfalls die Variablen V1 und V2 als Imputationsklassenvariablen; als Startwerte wurden die Variablenmittelwerte benutzt. Fehlende Daten wurden variablenweise ersetzt und nicht fallweise, d.h. ein unvollständiger Fall wurde eventuell durch gültige Werte mehrerer verschiedener Fälle komplettiert. Das Programmsegment REGRESSION schätzte fehlende Werte mit einer linearen Einfachregression mit der jeweils am stärksten korrelierenden Variablen, wobei die PAIRWISE-Korrelationsmatrix den Ausgangspunkt darstellt. HAZARD verwendete zusätzlich zu der von REGRESSION verwendeten Variablen die Anzahl fehlender Werte pro Fall zur Vorhersage. DEAR ersetzte fehlende Werte durch Werte der ersten Hauptkomponente, die aus der PAIRWISE-Matrix berechnet wurde. MULT verwendete die LISTWISE-Matrix als Ausgangspunkt der multiplen Regressionsschätzung; ILS verwendete die geschätzte Datenmatrix von MULT und die zugehörige vollständige Korrelationsmatrix als Ausgangspunkt. ILS konvergierte während der Simulation immer mit weniger als 100 Iterationen. Konvergenzkriterium war eine maximale Veränderung zwischen zwei aufeinanderfolgenden Korrelationsmatrizenschätzungen kleiner als 0.0001.

Als unabhängige Variablen wurden eine Reihe von "Ähnlichkeitsmaßen" der "wahren" Korrelationsmatrix und der geschätzten Matrix definiert. Als Ähnlichkeitsmaß der "Struktur" der Matrix wurde (wegen seiner unmittelbaren Anschaulichkeit) der Korrelationskoeffizient (r) der beiden als Datenvektoren aufgefaßten Korrelationsmatrizen berechnet. Da der Korrelationskoeffizient invariant gegenüber linearen Transformationen ist, bezieht sich dieses "Ähnlichkeitsmaß" nicht auf die Reproduktion einzelner Koeffizienten, sondern auf den Erhalt der Struktur, genauer: auf den Erhalt relativer Größenordnungen. Die Verteilung dieses Koeffizienten ist selbstverständlich extrem schief (in der Simulation ergab sich über alle Bedingungen eine

"skewness" von -2.083), daher wurde die Auswertung mit FISHER-z-transformierten Koeffizienten durchgeführt (im folgenden "ZMKS"). Daneben wurde die mittlere Abweichung der geschätzten Korrelationskoeffizienten berechnet.

Neben der Genauigkeit der Schätzung der Korrelationsmatrix sollte die Güte der Schätzung der fehlenden Werte untersucht werden. In Anlehnung an GLEASON & STAELIN (1975:243) und KALTON & SANTOS (1983:88) wurden u.a. die mittleren Abweichungen der geschätzten Daten von den "wahren" Werten berechnet. Da die Werte gruppiert wurden, kann auch die Frage nach dem Anteil "exakt" geschätzter Werte beantwortet werden. Hierzu wurde die Variable HIT berechnet, wobei HIT dem Quotienten entspricht:

$$\text{HIT} = \frac{\text{Zahl der Übereinstimmungen von wahren und geschätzten Werten}}{\text{Gsamtzahl fehlender Werte im Datensatz}}$$

Nach den dargestellten Überlegungen ergab sich folgendes fünffaktorielles Design:

- MTYPE: Ausfallmechanismus (MAR, MAT, MARC)
- N : FALLZAHL (100, 300, 500)
- CM : Interkorrelation (.2, .4, .6)
- MPER : Anteil fehlender Werte (1%, 5%, 10%)
- PROG : MD-Verfahren (LIST, PAIR, MEAN, CELL, DECK, REGR, HAZD, DEAR, MULT, ILS)

Das Simulationsprogramm erzeugte für jede Kombination der Faktoren 1-4 ("Zellen") einen Datensatz, der dann den 10 bzw. 6 "Behandlungen", den MD-Verfahren, unterworfen wurde. Für jede Zelle des Designs wurden insgesamt 10 voneinander vollständig unabhängige Datensätze erzeugt [6]. Das Design stellt damit ein fünffaktorielles (3x3x3x3x10) balanciertes Design mit wiederholter Messung nur des letzten Faktors dar. Mit den 81 Zellen und 10 Wiederholungen wurden insgesamt 810 vollständig voneinander unabhängige Datensätze erzeugt. Die die Korrelationsmatrizen betreffenden abhängigen Variablen wurden für alle 10 Verfahren berechnet, die die geschätzten Daten betreffenden abhängigen Variablen dagegen nur für die Verfahren 5-10, da weder PAIRWISE noch LISTWISE Ersetzungen vornehmen und der Einsatz von MEAN und CELLMEAN als MD-Ersetzungsverfahren zwar möglich, aber nicht sinnvoll schien. Für die die geschätzten Daten betreffenden Variablen reduzierte sich das Design auf ein (3x3x3x3x6)-Design.

5. Ergebnisse der Simulation

Aus der Fülle der Ergebnisse der Simulation sollen lediglich einige in Hinsicht auf praktische Konsequenzen besonders wichtige Ergebnisse berichtet werden.

Die zentrale Frage nach dem Interaktionseffekt zwischen Ausfallmechanismus und der Leistungsfähigkeit der Verfahren wurde mit mehreren Techniken untersucht. So zeigten z.B. Clusteranalysen den Wechsel der Clusterzugehörigkeit der Verfahren bei wechselnden Ausfallmechanismen. Dem Design entsprechen natürlich vor allem Varianzanalysen, insbesondere in Form von Profilanalysen. Die Varianzanalysen zeigten für alle Ähnlichkeitsmaße in bezug auf die Schätzung der Korrelationsmatrizen für alle Designvariablen hochsignifikante Interaktionseffekte ($p < 0.001$) mit den MD-Verfahren: Die Reaktionsprofile der MD-Verfahren verlaufen nicht parallel. In bezug auf die Schätzung fehlender Werte sind die Ergebnisse vergleichbar, wenn auch verständlicherweise die Effekte der Fallzahl und des Anteils fehlender Werte auf die Größe der Abweichung der geschätzten fehlenden Werte in der Regel nicht so stark waren (vgl. Tabellen 2 und 3).

Bei allen abhängigen Variablen stellen die Effekte der MD-Verfahren und die Ausfallmechanismen die stärksten Einflußfaktoren dar. Die getrennt nach Ausfallmechanismen und MD-Verfahren gerechneten Regressionen zeigen innerhalb der Zellen des Wiederholungsdesigns in Hinsicht auf die anderen Designvariablen in bezug auf die Schätzung der Korrelationsmatrizen die Linearität der resultierenden Verzerrungen: Die Schätzungen werden in der Regel umso schlechter, je niedriger die Zusammenhänge in der Population werden und je mehr die Daten fehlen. Tendenziell zeigt sich mit steigendem N unter sonst gleichen Bedingungen ein leichter Rückgang der resultierenden Verzerrungen.

Tabelle 2: MANOVA - Korrelationsabweichungsindizes (ZMKS)

ZWISCHEN

	F	SIGNIF.
MTYPE	40.342	< .001
N	3.655	< .050
MPER	1421.117	< .001
CM	94.191	< .001

BINNEN

	WILK'S LAMBDA	MULT.-F	DF	SIGNIF.
PROG	.10747	665.337	9/ 721	< .001
PROG x MTYPE	.27745	71.979	18/1442	< .001
PROG x N	.76850	11.273	18/1442	< .001
PROG x MPER	.24809	80.728	18/1442	< .001
PROG x CM	.39193	47.853	18/1442	< .001

Tabelle 3: MANOVA - "Exakte" MD-Schätzungen (HIT)

ZWISCHEN

	F	SIGNIF.
MTYPE	54.555	< .001
N	7.296	< .001
MPER	13.387	< .001
CM	44.608	< .001

BINNEN

	WILK'S LAMBDA	MULT.-F	DF	SIGNIF.
PROG	.25906	414.719	5/ 725	< .001
PROG x MTYPE	.69353	29.114	10/1450	< .001
PROG x N	.96579	2.546	10/1450	< .010
PROG x MPER	.96437	2.655	10/1450	< .010
PROG x CM	.79536	17.588	10/1450	< .001

5.1 Relative Effizienz der Verfahren

Um die Frage nach dem "besten" Verfahren unter den verschiedenen Bedingungen zu beantworten, wurden in Anlehnung an TIMM (1969:28; 1970:427) relative Effizienzindizes berechnet. Die relative Effizienz eines Verfahrens in bezug auf eine bestimmte Variable ist definiert als der Quotient zwischen dem Meßwert des jeweiligen Verfahrens und dem Maximum bzw. Minimum der Meßwerte aller Verfahren unter der jeweiligen Bedingung. Für die Variablen ZMKS und HIT wurde zunächst für jede experimentelle Bedingung das Maximum, für alle anderen Variablen das Minimum gesucht und der Quotient berechnet. Die Tabellen 4 und 5 geben für zwei Variablen (ZMKS und HIT) vereinfachte

Zusammenfassungen (bei denen auf eine Aufspaltung nach anderen experimentellen Faktoren verzichtet wurde) dieser Quotienten wieder.

Tabelle 4: Korrelationsabweichungsindex ZMKS: Mittlere relative Effizienz der Verfahren

Ausfall-mechanismus	Verfahren									
	LIST	PAIR	MEAN	CELL	DECK	REGR	HAZD	DEAR	MULT	ILS
MAR	.560	.923	.860	.854	.694	.899	.896	.984	.933	.968
MAT	.882	.926	.841	.838	.692	.923	.928	.957	.968	.978
MARC	.629	.951	.936	.932	.752	.921	.917	.978	.952	.960

Tabelle 5: Exakt geschätzte MD - Mittlere relative Effizienz der Verfahren

Ausfall-mechanismus	Verfahren					
	DECK	REGR	HAZD	DEAR	MULT	ILS
MAR	.579	.834	.836	.923	.881	.900
MAT	.513	.777	.871	.782	.900	.917
MARC	.657	.857	.851	.918	.887	.886

In bezug auf die Effizienz bei der Anzahl exakt geschätzter Werte zeigt HOTDECK über alle Bedingungen stets die niedrigsten Werte. Auffällig ist allerdings das Ansteigen der Effizienz im MARC-Modell. Während REGRESSION und HAZARD sich unter MAR und MARC nur wenig voneinander unterscheiden, wird die Differenz beider Verfahren unter MAT deutlich: HAZARD gehört unter diesem Modell zu den drei besten Verfahren. DEAR stellt unter MAR eines der drei besten Verfahren dar, wobei die Effizienz immer mit steigendem Zusammenhang in der Korrelationsmatrix wächst. Unter MAT ändert sich das Vorzeichen dieser Beziehung: Je höher CM, desto niedriger die Effizienz von DEAR. In diesem Modell gehört DEAR nicht mehr zu den besten Verfahren. Unter MARC verbessert sich DEAR wieder: Der Zusammenhang mit CM wird wieder positiv, unter den meisten Bedingungen liegt DEAR über allen anderen Verfahren. MULT und ILS ähneln sich unter allen Bedingungen sehr stark, eine leichte Verbesserung durch die Iteration ist in der Regel feststellbar. Beide Verfahren zeigen insgesamt die höchsten Effizienzen. Bemerkenswert ist vor allem die erstaunlich hohe Effizienz von ILS unter MAT bei hohem CM und 10% Ausfällen: Unter dieser Bedingung zeigt ILS stets die höchste Effizienz.

Die Effizienz der Verfahren in Hinsicht auf die Reproduktion der relativen Größenordnungen der Korrelationskoeffizienten (ZMKS) zeigt die Ähnlichkeit von PAIRWISE, DEAR, MULT und ILS unter MAR (0.923, 0.984, 0.933 und 0.968) ebenso deutlich wie die außergewöhnlich schlechten Ergebnisse von LISTWISE (0.560): Lediglich HOT-DECK zeigt vergleichbar schlechte Ergebnisse (0.694). Interessant auf diesem Hintergrund ist die relative Verbesserung der Effizienz von LISTWISE unter MAT (0.882), wobei das HOTDECK die schlechtesten Ergebnisse (0.692) zeigt und DEAR deutlich absinkt (0.957). Sowohl REGRESSION als auch HAZARD zeigen unter MAT deutlich bessere Effizienten: 0.923 bzw. 0.928. Unter MARC verschlechtert sich LISTWISE auf 0.629 und liegt damit im Mittel noch schlechter als HOTDECK mit 0.752. DEAR steigt im Mittel auf 0.978 und erzielt damit unter MARC fast durchgehend die besten Ergenisse.

5.2. Genauigkeit der Schätzungen

Betrachtet man die absoluten Ergebnisse der Verfahren in Hinsicht auf die Schätzung fehlender Daten, so ist der Anteil "exakt" geschätzter Daten überraschend hoch: Trotz der 7-stufigen Variablen liegt der Anteil "exakter" Schätzungen beim besten Verfahren niemals unter 28%, unter MAR liegt der Anteil im Durchschnitt für ILS bei 38%, unter MAT bei 36% und unter MARC bei 33%. Weitgehend unabhängig vom Anteil fehlender Werte erreicht HIT unter hohem CM und N bis zu 45%. Auffällig ist das Absinken von HIT unter MARC ebenso wie die große Streuung der Verfahren (wie sie auch durch die Betrachtung der relativen Effizienzen deutlich wird). Aufgrund der Ersetzungsmethode verständlich erscheint die Empfindlichkeit der DEAR gegenüber dem Ausfallmechanismus: Unter MAT entspricht der Anteil exakter Schätzungen bei DEAR ungefähr dem Anteil von REGRESSION und gehört damit zu den schlechteren Verfahren (DEAR: MAR 39%, MAT 30%, MARC 34%).

Die mittlere Abweichung der geschätzten Daten unter MAR schwankt im Durchschnitt über alle Bedingungen und Verfahren zwischen 0.017 (DEAR) und 0.025 (REGRESSIONS); damit findet tendenziell eine leichte Unterschätzung statt. Unter MAT steigt die mittlere Abweichung der geschätzten Daten um mindestens eine Zehnerpotenz an: Die mittleren Unterschätzungen liegen zwischen 0.295 bei ILS und 1.05 bei HOTDECK. Unter MARC gehen die mittleren Abweichungen auf -0.01 (HAZARD) bis 0.002 (DEAR) zurück, wobei häufige Oberschätzungen insbesondere bei kleineren Fallzahlen auffallen.

Auf der Ebene der Korrelationsmatrizen liegen unter MAR die mittleren Abweichungen zwischen -.012 (ILS) und 0.033 (HOTDECK), wobei durchgängig eine Unterschätzung stattfindet. Die Ausnahmen hiervon bilden DEAR, MULT und ILS, bei denen durchgängig Oberschätzungen stattfinden, wobei die Oberschätzung mit dem Anteil fehlender Werte ansteigt. Unter MAT steigen die mittleren Abweichungen auf -0.04 (ILS) bis 0.049 (HOTDECK), wobei alle Verfahren mit Ausnahme von ILS Unterschätzungen der Korrelationskoeffizienten zeigen. ILS zieht tendenziell sehr kleine Oberschätzungen nach sich, der Extremwert wird mit -0.016 bei N=300, CM=.6 und MPER=10% erreicht. Unter MARC zeigen alle Verfahren (mit Ausnahme von HOTDECK und MEAN) deutliche Oberschätzungen der Korrelationskoeffizienten: -0.038 (LISTWISE) bis -0.002 (CELLMEAN). HOTDECK und MEAN zeigen Unterschätzungen (0.0219 und 0.0002). Die größte mittlere Oberschätzung erreicht LISTWISE mit -.0082 bei N=500, CM=.6 und MPER=10%.

Betrachtet man die Ergebnisse in bezug auf ZMKS, so fällt die relative Konstanz der Ergebnisse von PAIRWISE, MULT und ILS auf: Die Koeffizienten liegen im MIttel zwischen 1.89 und 2.02, die Korrelationen zwischen geschätzten und "wahren" Korrelationsmatrizen liegen damit zwischen 0.955 und 0.97 und reproduzieren damit die "Struktur" der Matrizen weitgehend. Trotz dieser überwiegend guten Ergebnisse existieren für alle MD-Techniken Bedingungen, bei denen die Reproduktion nur ungenügend gelingt: Das Minimum erreicht ILS mit 1.20 (r=0.835) unter MARC bei N=300, CM=.6 und MPER=10%. Das schlechteste Ergebnis aller Verfahren unter allen Bedingungen erzielt LISTWISE mit .577 (R=.052) unter MAR, N=300, CM=.2 und MPER=10%.

5.3 Zusammenfassung

Die wichtigsten Ergebnisse der Simulation können in drei Punkten zusammengefaßt werden:

1. Die MD-Verfahren bilden durch ihre Konstruktion Gruppen mit in sich relativ homogenen Reaktionsmustern. Die Ergebnisse der Clusteranalysen wurden durch den Vergleich der Regressionsparameter der getrennt durchgeführten Regressionen ebenso bestätigt wie durch Plots.

2. Die MD-Verfahren reagieren auf die vermuteten Einflußfaktoren "mittlere Interkorrelation" und vor allem "Anteil fehlender Werte" erwartungsgemäß; der Effekt der Fallzahl kann nicht so deutlich demonstriert werden.

3. Ein unter allen Bedingungen effizientes Schätzverfahren existiert nicht. Die Ergebnisse der Simulation stützen die Vermutung der Existenz der postulierten Interaktionseffekte zwischen Ausfallmechanismus und der Leistungsfähigkeit der MD-Verfahren. Verallgemeinerungen von Ergebnissen über die Leistungsfähigkeit der MD-Verfahren ausgehend von den Resultaten unter MAR sind problematisch, die Wahl eines "optimalen" MD-Verfahrens bei unbekanntem Ausfallmechanismus unmöglich. Die zentrale Frage der Simulation, die Frage nach der "Robustheit" der Verfahren gegenüber dem Ausfallmechanismus, muß negativ beantwortet werden: Die Clusterzugehörigkeiten ändern sich ebenso wie die Regressionsparametervorzeichen; die Interaktionseffekte in den Varianzanalysen sind dementsprechend signifikant.

Neben diesen Hauptergebnissen lassen sich aus der Simulation weitere Resultate gewinnen, die für die sozialwissenschaftliche Forschungspraxis relevant sind.

So zeigt entgegen dem Großteil der Literatur die allgemein für MAR nahegelegte fallweise Löschung fehlender Werte (LISTWISE) gegenüber der PAIRWISE-Methode überwiegend enttäuschende Ergebnisse. In diesem Zusammenhang soll auf die Ergebnisse von BROWN (1983) und FINKBEINER (1979) hingewiesen werden. FINKBEINER (1979:419) stellt fest: "If the researcher finds the CPO (PAIRWISE, Anm.d.Verf.) method usable in his application, it can be expected to produce relatively accurate results", und: "If cost is a concern, then the MR (MEAN, Anm.d.Verf.) and CPO methods are to be recommended as being effective yet relatively inexpensive and easy to implement" (FINKBEINER, 1979:420). FINKBEINER bezieht sich auf die Ergebnisse seiner Simulationsstudie in Hinsicht auf Faktorenanalysen; seine "CPO method" liegt bei den Ergebnissen besser als die Hauptkomponentenmethode und besser als MEAN und LISTWISE. BROWN argumentiert ebenfalls in Hinsicht auf Faktorenanalysen, allerdings mit analytischer Zielsetzung. BROWN (1983:286) stellt fest: "... PP (PAIRWISE, Anm.d.Verf.) performs reasonably well relative to ML provided the measure of sparseness is small", und: "Qualitatively, PP estimates become more efficient compared to DEL (LISTWISE, Anm.d.Verf.) as the

signal-to-noise level in a variable decreases. Improvement of DEL over PP is realized only when one variable is measured almost without error (communality above 0.80) and there is five percent or less missing data" (BROWN, 1983:281). Möglicherweise liegt in diesem letzten Ergebnis von BROWN die Erklärung sowohl für die vergleichsweise guten Resultate für PAIRWISE im Rahmen dieser Simulation als auch für die starke Verbreitung der PAIRWISE-Methode in der sozialwissenschaftlichen Forschungspraxis, die sich damit im Gegensatz zu den Ergebnissen der theoretischen Statistiker befindet (vgl. HEIBERGER & LITTLE, 1981).

Unter Beachtung der theoretischen Vorbehalte gegenüber einer Datenanalyse ohne explizite (und notwendige) Modellierung des Ausfallprozesses lassen sich aus der Simulation in ihrer Verallgemeinerbarkeit stark begrenzte Empfehlungen ableiten, falls eine Analyse ohne die weiter unten noch zu erwähnenden "multiple Imputations" erfolgen muß: Die Ergebnisse der Simulation legen für homogene Korrelationsmatrizen die Verwendung von ILS bzw. DEAR (je nach Ausfallmechanismus) nahe. Ist dies technisch nicht möglich, stellt PAIRWISE eine akzeptable Alternative dar.

In bezug auf die "Leistungsfähigkeit" der Verfahren sollten die teilweise recht hohen mittleren Werte nicht darüber hinwegtäuschen, daß bei bestimmten Bedingungen alle Verfahren so niedrige Werte nach sich ziehen, daß die "Struktur" der Matrix als zerstört angesehen werden kann: Die relative Größe der Koeffizienten wird nur ungenügend reproduziert; daher müssen alle Verfahren, die auf der relativen Größe aufbauen, notwendigerweise weitgehend irreführende Ergebnisse erbringen. Auch die sozialwissenschaftlich realistischen Modelle zeigen unter einigen Bedingungen so kleine Koeffizienten, daß z.B. eine Faktorenanalyse, gleich mit welcher Methode die Korrelationsmatrix berechnet wurde, mit Sicherheit sehr stark verzerrte Faktorenstrukturen zeigen würde. Diese potentiellen Gefahren durch systematische Ausfallmechanismen sollen anhand eines Nebenergebnisses der Simulation noch einmal demonstriert werden.

Die Standardbehandlung fehlender Werte im Rahmen empirischer Sozialforschung erschöpft sich in der Regel in der Entscheidung zwischen PAIRWISE und LISTWISE als Berechnungsmethoden der Korrelationsmatrizen (vgl. KIM & CURRY, 1977:215). Ein damit verbundenes Problem zeigt die Tabelle 6. Kon-

ventionellerweise wird in der Pfadanalyse und in der konfirmatorischen Faktorenanalyse ein Modell verworfen, wenn die Abweichungen zwischen dem Modell und den beobachteten Koeffizienten größer als 0.1 sind (WEEDE & JAGODZINSKI, 1977:321). Nun ist nur in den seltensten Fällen die Stichprobe Gegenstand des Interesses, dieses richtet sich dagegen auf Aussagen über Populationen - auch wenn inferenzstatistische Probleme zumeist vernachlässigt werden. So ist es durchaus möglich, ein Modell entsprechend dem oben genannten Kriterium anzupassen, aber die Anpassung trifft eben nur auf die jeweilige Stichprobe, nicht hingegen auf die Population zu, falls ein systematischer Ausfallmechanismus vorliegt und somit die beobachteten Korrelationen selbst mehr als .1 von den "wahren" Werten abweichen. Die Tabelle 6 zeigt für die beiden häufigsten Berechnungen von "MD-Korrelationen" den in der Simulation ermittelten Anteil derjenigen geschätzten ("beobachteten") Korrelationsmatrizen, bei denen die mittlere Abweichung vom wahren Wert größer als .1 ist. Der Anteil derart stark verzerrter Korrelationen erreicht unter MAT und 10% fehlender Werte bei LISTWISE 18.9%. Die Akzeptanz oder Verwerfung eines Modells wird dementsprechend häufig fehlerhaft sein. Die Tabelle bezieht sich nicht auf einzelne Koeffizienten (wie das Kriterium), sondern auf die mittlere Abweichung aller Koeffizienten.

Tabelle 6: Korrelationsmatrizen mit mittleren Abweichungen < .10 (Angaben in Prozent)

Ausfall-typ	Methode	% MD pro Variable (1-10%)	(10%)
MAR	LIST	0.7	2.2
	PAIR	0.0	0.0
MAT	LIST	6.7	18.9
	PAIR	0.4	0.1
MARC	LIST	3.7	7.8
	PAIR	0.0	0.0

Diese hohe Zahl potentieller Fehlentscheidungen und die durch die Simulation demonstrierte Existenz des Interaktionseffektes zwischen der Leistungsfähigkeit der MD-Verfahren und den Ausfallmechanismen stellen weitere Argumente für die Notwendigkeit der theoretischen Begründung vermuteter Ausfallmechanismen dar. Eine "automatische" Kompensation systematischer Ausfallmechanismen kann von den MD-Techniken nach den Ergebnissen der Simulation nicht erwartet werden. Damit stützen die empirischen Ergebnisse die

theoretische Kritik an den Annahmen der Ersetzungsverfahren: Die auf den MAR- bzw. MARC-Annahmen basierten Techniken können bei Vorliegen anderer Ausfallmechanismen bzw. bei einer rein "mechanischen" Anwendung zu Fehlschlüssen führen.

6. Explizite Modellierung des Ausfallmechanismus als alternative Analysestrategie: Multiple Imputation (MI)

Die somit verbleibende Alternative bei der Analyse von Datensätzen, bei denen systematische Ausfallmechanismen nicht ausgeschlossen werden können, besteht in der explizit theoretisch geleiteten Ersetzung fehlender Werte. Da diese theoretischen Annahmen mit den verbundenen Daten unüberprüfbar sind, verbleibt lediglich die Möglichkeit, unterschiedliche Hypothesen über den Ausfallprozeß in ihren Konsequenzen auf Entscheidungen über Akzeptanz oder Verwerfung substantieller theoretischer Hypothesen zu untersuchen. Die einzige Möglichkeit hierzu stellt die Anwendung des Konzepts der "multiple imputations" dar.

D.B. RUBIN entwickelt seit 1977 in einer Reihe von Arbeiten (RUBIN, 1977, 1978, 1979, 1980, 1981) das Konzept der "multiple imputation". [7] Ausgehend von der Tatsache, daß jede Analysestrategie beim Vorliegen eines MD-Problems ein Modell des Ausfallprozesses verwenden muß, besteht das Vorgehen bei einer multiplen Imputationsstrategie aus drei Schritten.

Zunächst müssen explizite Modellierungen des Ausfallprozesses erfolgen. Entsprechend den Annahmen des Modells werden dann Ersetzungen der fehlenden Werte mit einem nichtdeterministischen MD-Ersetzungsverfahren vorgenommen (das Ergebnis ist ein vollständiger Datensatz, zu dessen Analyse Standardtechniken benutzt werden können). Dieser Schritt wird nun I-mal wiederholt; da das MD-Verfahren bei Wiederholungen nicht dieselben Ersetzungen vornimmt, entstehen I vollständige (unterschiedliche) Datensätze. Im letzten Schritt werden die aus den I-Datensätzen berechneten Statistiken miteinander verglichen. Der Vergleich erlaubt (falls ein Modell des Ausfallprozesses benutzt wurde) die Simulation der Verteilung der Statistiken bei gegebenen Daten und gegebenem Ausfallprozeß. Werden hingegen unterschiedliche Annahmen über den Ausfallprozeß gemacht, d.h. also: mehrere explizite Modellierungen vorgenommen, erlaubt der Vergleich die Beurteilung der Sensitivität der Ergebnisse gegenüber unterschiedlichen Annahmen über den Aus-

fallprozeß. Durch den Vergleich der Statistiken der unter verschiedenen Annahmen über den Ausfallmechanismus vervollständigten Datensätze lassen sich Aussagen über die Sensitivität der Ergebnisse gegenüber den Ausfallmechanismen gewinnen. MI liefert keine eindeutigen Resultate, sondern zeigt die Spannweite möglicher Ergebnisse auf.

Die Idee, mehrfache Ersetzungen für fehlende Daten vorzunehmen, um den Effekt verschiedener Ausfallmodelle zu testen, ist keineswegs neu. Explizit findet sich der Vorschlag z.B. schon bei FINNEY (1974:10). RUBIN hat dann aber die theoretischen Grundlagen einer vollständigen Bayes-Analyse für die MI formuliert, deren Durchführung in der Praxis jedoch fast unmöglich ist, selbst wenn die Beschränkung auf "ignorable mechanisms" erfolgt: Selbst mit einer einzigen Variablen ist der Arbeitsaufwand außerordentlich hoch (HERZOG & RUBIN, 1983:242).

Die Durchführung der MI mit (in Hinsicht auf die Bayes-Grundlagen) impliziten Techniken, wie z.B. einem Hotdeck ist hingegen vergleichsweise einfach. Wenn auch gegen die ursprünglichen Intentionen des MI-Ansatzes gerichtet, ist eine von den theoretischen Grundlagen der MI abgelöste Anwendung möglich. Beim derzeitigen Stand des Wissens zu systematischen Ausfällen scheint diese Anwendung der MI gerade für den Normal-Nutzer die einzige Möglichkeit zu sein, ein MD-Problem in seinen möglichen Auswirkungen abzuschätzen, wenn auch alle Schritte einer MI z.Z. noch Schwierigkeiten bieten. Wie schon erwähnt, existieren bis auf extrem vereinfachte Beispiele keine brauchbaren Standardmodelle, an denen zumindest eine Orientierung möglich wäre. Die Modelle müssen völlig neu entwickelt werden. Da auch nur vereinzelte und in der vorhandenen Form in diesem Zusammenhang nicht brauchbare sozialwissenschaftliche ("inhaltliche") Modelle existieren, bleibt als relativ einfache Möglichkeit nur die Entwicklung simpler Modelle, die den "ungünstigsten Fall" eines systematischen Ausfalls für die jeweilige Fragestellung erfassen sollen. Diese "worst reasonable models (WRM's)" (DEMPSTER, 1978:30) können nur die Extremwerte der interessierenden Statistiken zeigen, eine "realistische" Modellierung der Ausfallprozesse erfolgt nicht. Trotzdem stellen "WRM's" durch die durch sie möglichen Sensitivitätstest einen brauchbaren ersten Versuch dar.

Neben der Entwicklung von Sensitivitätsindizes bleibt vor allem die Frage nach der Änderung der Entscheidungen über Akzeptanz oder Verwerfung einer konkreten Hypothese bei Annahme verschiedener Ausfallprozesse.

Zusammenfassend muß betont werden, daß die Durchführung einer MI z.Z. mit einer so großen Zahl praktischer Probleme behaftet ist, daß die Anwendung im Rahmen einer Standardanalyse weitgehend ausscheidet. Da MI bisher jedoch die einzige Möglichkeit zur Abschätzung der Konsequenzen eines MD-Problems darstellen, muß die Entwicklung von empirisch gut abgesicherten Theorien zur Erklärung der Prozesse, die zum "Ausfall" bzw. zu "missing data" führen, als Voraussetzung für ein effizientes Imputationssystem als dringend notwendig betrachtet werden.

Dieser Beitrag wurde von Rainer Schnell verfaßt, der mit Hartmut Esser in Fragen der Erklärung und Analyse von Fehlern bei der Datenerhebung zusammenarbeitet.

Anmerkungen

1) Zu den wenigen MD-Techniken, die explizit nicht auf der MAR-Annahme beruhen, vgl. COHEN (1968) und COHEN & COHEN (1975) sowie BERK & RAY (1982) und BERK (1983).
2) Eine iterative Version der multiplen Regressionsschätzung ohne Korrektur wurde bereits von FEDERSPIEL, MONROE & GREENBERG (1959) verwendet. Ebenso ist diese Option in den Programmen von BERGER (1979) und O'GRADY (1982) vorhanden. SCHMEE & HAHN (1979) schlugen eine "ILS"-Technik explizit zur Lösung von "Censoring"-Problemen vor.
3) Die Variation der Verteilungsform wäre zwar wünschenswert gewesen, aber die Aufnahme eines weiteren experimentellen Faktors hätte mindestens eine Verdreifachung der Rechenzeit bedingt.
4) Die Ausgangsdaten wurden in Anlehnung an BOLLEN & BARB (1981) kategorisiert, die Abweichungen der geschätzten Daten basieren entsprechend auf kategorisierten Schätzungen; den Berechnungen der Korrelationsabweichungen liegen hingegen nicht-kategorisierte Schätzungen zugrunde.
5) Die Setzung dieser Korrelationen hat einen theoretisch interessanten Hintergrund. Unter der Voraussetzung, daß die Globalvariablen nur deshalb in empirischen Untersuchungen mit kognitiven Variablen Korrelationen aufweisen, weil in der Regel typische Kombinationen von Merkmalsausprägungen vermittelt über jeweils typische handlungsrelevante Mechanismen, z.B. spezielle, Motivationen vermittelnde Sozialisationspraktiken, jeweils ähnliche Strukturen hervorbringen (LANGENHEDER, 1975:61), wurden diese Korrelationen konstant niedrig angesetzt.
6) Zur Generierung der Pseudo-Zufallszahlen und zur Realisation des Simulationsprogramms wurden Subroutinen der NUMERICAL ALGORITHM GROUP (NAG) FORTRAN LIBRARY Mark 10 verwendet. Der zugrundeliegende "multiplicative congruential"-Generator G05CAF wurde durch G05CBF gestartet, die Erzeugung der multivariatnormal verteilten Daten erfolgte mit G05EAF und G05EZF. Das Programm wurde auf den IBM 4341 des HRZ Essen unter VM/SP und

SSI mit dem IBM-VS-FORTRAN Compiler Level 1.3.0 erstellt und gerechnet. Einzelheiten und das Programmlisting finden sich bei SCHNELL (1985).

7) Eng mit diesem Konzept verbunden ist das Konzept der "Repeated Replication Imputation Procedure" (RRIP) von KISH (1983) für Unitnonresponse.

Literatur

ANDERSON, A.B., BASILEVSKY, A. & HUM, D.J. Missing data: a review of the literature. In: P.H. ROSSI, J.D. WRIGHT & A.B. ANDERSON (Hrsg.), Handbook of survey research. New York, 1983, 415-493.

BEALE, E.M.L. & LITTLE, R.J.A. Missing values in multivariate analysis. Journal of the Royal Statistical Society, Series B, 37, 1975, 129-146.

BERGER, M.P.F. A FORTRAN IV program for the estimation of missing data. Behavior Research Methods and Instrumentation, 11, 1979, 395-396.

BERK, R.A. & RAY, S.C. Selection biases in sociological data. Social Science Research, 11, 1982, 352-398.

BERK, R.A. An introduction to sample selection bias in sociological data. American Sociological Review, 46, 1981, 386-398.

BOLLEN, K.A. & BARB, K.H. Pearson's r and coarsely categorized measures. American Sociological Review, 46, 1981, 232-239.

BROWN, C.H. Asymptotic comparison of missing data procedures for estimating factor loadings. Psychometrika, 48, 1983, 269-291.

BUCK, S.F. A method of estimation of missing values in multivariate data suitable for use with an electronic computer. Journal of the Royal Statistical Society, Series B, 22, 1960, 302-306.

COHEN, J. Multiple regression as a general data-analytic system. Psychological Bulletin, 70, 1968, 426-443.

COHEN, J. & COHEN, P. Applied multiple regression. Correlation analysis for the behavioral sciences. Hillsdale, 1975.

DEAR, R.E. A principal component missing data method for multiple regression models. Technical report SP-86, System Development Corporation, Santa Monica/Calif., 1959.

DEMPSTER, A.P., LAIRD, N.M. & RUBIN, D.B. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, 39, 1977, 1-22.

DEMPSTER, A.P. Discussion. ASA Proceedings of the Section on Survey Research Methods, 1978, 30-31.

ESSER, H. Fehler bei der Datenerhebung. Kurseinheit 3: Datenerhebung als sozialer Prozeß. Studienbrief, Fernuniversität Hagen, 1983.

ESSER, H. Determinanten des Interviewer- und Befragtenverhaltens: Probleme der theoretischen Erklärung und empirischen Untersuchung von Interviewereffekten. In: K.U. MAYER & P. SCHMIDT (Hrsg.), Allgemeine Bevölkerungsumfrage der Sozialwissenschaften. Frankfurt, 1984, 26-71.

FEDERSPIEL, C.F., MONROE, R.J. & GREENBERG, B.G. An investigation of some multiple regression methods for incomplete samples. University of North Carolina, Institute of Statistics, Memeo Series #236.

FINKBEINER, C. Estimation for the multiple factor model when data are missing. Psychometrika, 44, 1979, 409-420.

FINNEY, D.J. Problems, data, and inference. Journal of the Royal Statistical Society, Series A, 137, 1974, 1-19.

FORD, B.L. An overview of hot-deck procedures. In: W.G. MADOW, I. OLKIN & D.B. RUBIN (Hrsg.), Incomplete data in sample surveys, vol. 2. New York, 1983, 185-207.

GLEASON, T.C. & STAELIN, R. A proposal for handling missing data. Psychometrika, 40, 1975, 229-252.

HAITKOVSKY, Y. Estimation of multivariate statistics from grouped and missing data. Ph.D. Thesis, Harvard University, 1966.

HAITKOVSKY, Y. Missing data in regression analysis. Journal of the Royal Statistical Society, Series B, 63, 1968, 67-82.

HAMILTON, M.A. Regression analysis when there are missing observations - a survey and bibliography. Technical report 1-3-75. Statistical Laboratory, Montana State University, 1975.

HEIBERGER, R.M. Regression with pairwise-present covariance matrix: a dangerous practice. ASA Proceedings of the Statistical Computation Section, 1977, 38-47.

HEIBERGER, R.M. Sample indicators of perturbation for missing value regression techniques. In: International Association for Statistical Computation (Hrsg.), COMPSTAT 1978. Wien, 1978, 88-94.

HERZOG, T.N. & RUBIN, D.B. Using multiple imputations to handle nonresponse in sample surveys. In: W.G. MADOW, I.OLKIN & D.B. RUBIN (Hrsg.), Incomplete Data in sample survey, vol. 2. New York, 1983, 209-245.

KAISER, J. The effectiveness of hot-deck procedures in small samples. ASA Proceedings of the Section on Survey Research Methods, 1983, 523-528.

KALTON, G. Compensating for missing survey data. Survey Research Center, Institute for Social Research, University of Michigan, 1983.

KALTON, G. & KASPRZYK, D. Imputing for missing survey responses. ASA Proceedings of the Section on Survey Research Methods, 1982, 22-31.

KIM, J.O. & CURRY, J. The treatment of missing data in multivariate analysis. Sociological Methods and Research, 6, 1977, 215-239.

KISH, L. Repeated replication imputation procedure (RRIP). In: G. KALTON (Hrsg.), Compensating for missing survey data. Survey Research Center, Institute for Social Research, University of Michigan, 1983, 131-136.

LANGENHEDER, W. Theorie menschlicher Entscheidungshandlungen. Stuttgart, 1975.

LITTLE, R.J.A. Maximum likelihood inference for multiple regression with missing values: a simulation study. Journal of the Royal Statistical Society, Series B, 41, 1979, 76-87.

LITTLE, R.J.A. The ignorable case. In: W.G. MADOW, I. OLKIN & D.B. RUBIN (Hrsg.), Incomplete data in sample surveys, vol. 2. New York, 1983, 341-382.

LITTLE, R.J.A. The ignorable case. In: W.G. MADOW, I. OLKIN & D.,B. RUBIN (Hrsg.), Incomplete data in sample surveys, vol. 2. New York, 1983b, 383-413.

LITTLE, R.J.A. & RUBIN, D.B. Incomplete data. In: S. KOTZ & N.L. JOHNSON (Hrsg.), Encyclopedia of Statistical Sciences, vol. 4. New York, 1983, 46-53.

LITTLE, R.J.A. & RUBIN, D.B. Missing data in large data sets. In: WRIGHT, T. (Hrsg.), Statistical methods and the improvement of data quality. Orlando, 1983b, 215-243.

LÖSEL, F. & WÖSTENDÖRFER, W. Zum Problem unvollständiger Datenmatrizen in der empirischen Sozialforschung. Kölner Zeitschrift für Soziologie und Sozialpsychologie, 26, 1974, 342-357.

MÜNTMANN, V., BOLLINGER, G. & HERRMANN, A. Tests auf Zufälligkeit von "missing data". In: H. WILKE (Hrsg.), Statistik-Software in der Sozialforschung. Berlin, 1983, 87-101.

O'GRADY, K.E. Regression estimation of missing data. Behavior Research Methods and Instrumentation, 14, 1982, 359-360.

ORCHARD, T. & WOODBURY, M.A. A missing information principle: theory and practice. Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability, 1, 1970, 697-715.

RUBIN, D.B. Formalizing subjective notions about the effect of nonrespondents in sample surveys. Journal of the American Statistical Association, 72, 1977, 359, 538-543.

RUBIN, D.B. Multiple imputations in sample surveys - a phenomenological Bayesian approach to nonresponse. ASA Proceedings of the Section on Survey Research Methods, 1978, 20-28.

RUBIN, D.B. Illustrating the use of multiple imputations to handle nonresponse in sample surveys. Bulletin of the International Statistical Institute, 1979, 517-531.

RUBIN, D.B. Handling nonresponse in sample surveys by multiple imputation. Monography, US Bureau of the Census, 1980.

RUBIN, D.B. The Bayesian bootstrap. The Annals of Statistics, 9, 1981, 130-134.

SANTOS, R.L. Effects of imputation on regression coefficients. ASA Proceedings of the Section on Survey Research Methods, 1981, 140-145.

SANTOS, R.L. Effects of imputation on complex statistics. Survey Research Center, Institute for Social Research, University of Michigan, 1981b.

SCHMEE, J. & HAHN, G.J. A simple method for regression analysis with censored data. Technometrics, 21, 4, 1979, 417-432.

SCHNELL, R. Missing Data Probleme in der empirischen Sozialforschung. Unveröffentlichtes Manuskript, Fachbereich 1, Universität Essen, 1985.

TABACHNICK, B.G. & FIDELL, L.S. Using multivariate statistics. New York, 1983.

TIMM, N.H. Estimating variance-covariance and correlation matrices from incomplete data. Ph.D. Thesis, University of California, Berkley, 1969.

TIMM, N.H. The estimation of variance-covariance and correlation matrices from incomplete data. Psychometrika, 35, 417-437.

VACEK, P.M. & ASHIKAGA, T. An examination of the nearest neighbor rule for imputing missing values. ASA Proceedings of the Statistical Computing Section, 1980, 326-331.

VAN GUILDER, M. & AZEN, S. Conclusions regarding algorithms for handling incomplete data. ASA Proceedings of the Statistical Computing Section, 1981, 53-56.

WEEDE, E. & JAGODZINSKI, W. Einführung in die konfirmatorische Faktorenanalyse. Zeitschrift für Soziologie, 6, 3, 1977, 315-333.