MATCHING KLEINER STICHPROBEN

EIN VERGLEICH VERSCHIEDENER VERFAHREN

DISSERTATION

zur Erlangung des Grades
Doktor der Wirtschaftswissenschaft (Dr. rer. pol.)
der Juristischen und Wirtschaftswissenschaftlichen Fakultät
der Martin-Luther-Universität Halle-Wittenberg

eingereicht am 12.08.2008
verteidigt am 25.11.2008
von **Dipl.-Vw. Dipl.-Kfr. Eva Reinowski**

Gutachter
Prof. Dr. Claudia Becker
Prof. Dr. Heinz P. Galler

urn:nbn:de:gbv:3-000014674
[http://nbn-resolving.de/urn/resolver.pl?urn=nbn%3Ade%3Agbv%3A3-000014674]

Zusammenfassung

Die durchgeführte Studie liefert einen Beitrag zur Entwicklung von „Standards“ für den Einsatz von Matchingverfahren in empirischen Evaluationsstudien.
Innerhalb einer Simulation werden die Ergebnisse verschiedener in der Literatur diskutierter Distanzmaße und Zuordnungsprozesse miteinander verglichen. Die strenge Orientierung an realen Entscheidungssituationen stellt dabei eine Ergänzung zu den meisten bisher bekannten Studien dar. Sie erklärt zum einen die Fokussierung auf kleine Stichproben, zum anderen die explizite Berücksichtigung unterschiedlich skalierter Variablen, die im Matchingprozess berücksichtigt werden müssen. Um eine Annäherung an realistische Verteilungen dieser Variablen zu erreichen, wird der Mikrozensus des Statistischen Bundesamtes als Vorbild für die generierten Zufallsvariablen der Simulation verwendet.

Die Betrachtung der Distanzmaße umfasst die in der Literatur als vorteilhaft angesehenen Maße, die Mahalanobisdistanz und Balancing Scores sowie aus der Statistik bekannte – in Evaluationsstudien bisher allerdings nicht verwendete – aggregierte Distanzmaße. Die Auswahl der analysierten Zuordnungsprozesse orientiert sich ebenfalls an den Ergebnissen bisheriger Studien. In die Analyse werden Replacement Matching, Random Matching, Optimal Nearest Neighbor Matching, Ridge Matching und Optimal Full Matching einbezogen.

Der Vergleich der Matchingergebnisse der verschiedenen Distanzmaße anhand nichtparametrischer skalenspezifischer Tests der Übereinstimung der Merkmalsverteilungen zeigt, dass aggregierte Distanzmaße in kleinen Stichproben besser in der Lage sind, Ähnlichkeiten in unterschiedlich skalierten Merkmalen zusammenzufassen als die bisher gebräuchlichen Maße.
Hinsichtlich des mittleren quadratischen Fehlers und seiner Bestanteile ist Optimal Full Matching den anderen analysierten Zuordnungsprozessen vorzuziehen, bei Betrachtung der quadrierten Distanzsumme als Näherungswert für die Angleichung der Merkmalsverteilungen erreicht Replacement Matching die beste Zuordnung. Der erwartete Vorzug optimaler Zuordnungsprozesse gegenüber anderen Nearest Neighbor Matching Algorithmen wird durch die Simulationsergebnisse nicht bestätigt.

Stichwörter

Simulation, Matching, Distanzmaße, Zuordnungsalgorithmen, optimale Zuordnungsprozesse, mikroökonometrische Evaluation, Berufsausbildung

Abstract

The study contributes to the development of „standards“ for the application of matching algorithms in empirical evaluation studies.
Various distance measures and matching processes that are dicussed in the current literature are compared among each other in a simulation study. Supplementary to former studies, the simulation setup strongly orientates on real evaluation situations. This reality orientation requires to focus on small samples, and differently scaled variables must be considered explicitly in the matching process. In order to approximate realistic distributions, the random variables in the simulation are generated after the example of the German Microcensus.

In the simulation, the Mahalanobis distance and two Balancing Scores are considered because their use is recommended in evaluation literature. Additionally, statistical aggregated distance measures not yet used for empirical evaluation are included. The choice of matching algorithms is orientated on the results of former studies: Replacement Matching, Random Matching, Optimal Nearest Neighbor Matching, Ridge Matching and Optimal Full Matching are analyzed.

The matching outcomes of the analyzed distance measures are compared by nonparametrical scale-specific tests for identical distributions of the characteristics in the participant's and the control group. In small samples, aggregated distance measures are the better choice for summarizing similarities in diffently scaled variables compared to commonly used measures.
Regarding the Mean Square Error and its parts, bias and variance, using Optimal Full Matching results in favourable matching outcomes. In terms of the sum of the squared distances – as an approximation for the similarity of the variable's distributions –, Replacement Matching is able to identify the best control groups. The expected superiority of Optimal Nearest Neighbor Matching is not confirmed by the simulation results.

Keywords

simulation study, matching algorithms, optimal matching, distance measures, microeconometric evaluation, vocational training

Danksagung

Ich möchte mich bei allen Personen, die mich bei meiner Arbeit unterstützt und zu dieser Dissertation beigetragen haben, herzlich bedanken.
Mein besonderer Dank gilt meinen Betreuern, Frau Prof. Claudia Becker und Herrn Prof. Heinz P. Galler, für die umfassende Begleitung der Studie, Ihre hilfreichen Kommentare zu früheren Fassungen der einzelnen Kapitel und die zahlreichen fruchtbaren Diskussionen, die sehr zur Verbesserung und Vervollkommnung dieser Arbeit beigetragen haben.
Außerdem möchte ich Heiner Dettmann, Wilfried Ehrenfeld und Christian Schmeißer für Ihre Unterstützung bei der Konzeption und Durchführung der Simulation danken. Ohne ihre Ideen und Vorschläge wäre die Studie nicht das, was sie jetzt ist.
Den Kollegen des Zentrums für Sozialforschung Halle, insbesondere Dr. Christine Steiner, danke ich ebenfalls sehr herzlich für die Möglichkeit, das zsh-Jugendpanel zu nutzen, und für ihre schnelle und unkomplizierte Hilfe bei der Aufbereitung der Daten.

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

Kapitel 1

Einleitung

Matching gehört zu den am weitesten verbreiteten Verfahren in der empirischen Evaluationsforschung. Diese Popularität ist u.a. mit der leicht verständlichen Grundidee des Verfahrens – der Suche nach „statistischen Zwillingen“ für die beobachteten Personen – zu erklären und macht Matching besonders geeignet als Grundlage der wissenschaftlichen Politikberatung zum Einsatz bzw. der Verteilung öffentlicher Mittel.

Das wachsende Interesse an der Entwicklung von Evaluationsverfahren steht in engem Zusammenhang mit der Ausweitung der Förderprogramme auf dem Arbeitsmarkt – zunächst in den USA, seit Beginn der 1990er Jahre auch in Europa. Die steigenden Ausgaben für die aktive Arbeitsmarktpolitik machte die Überpüfung der Wirkung der eingeführten Förderprogramme – und damit die Rechtfertigung der öffentlichen Ausgaben auf diesem Gebiet – notwendig.[1] Die Bedeutung der Evaluation des Einsatzes öffentlicher Mittel in Deutschland lässt sich u.a. daran erkennen, dass für die Reform der Arbeitsmarktpolitk durch die sog. Hartz-Gesetze vom Bundestag eine umfassende Evaluation aller implementierten Arbeitsmarktprogramme beschlossen wurde.[2]

Bei der Beurteilung der Auswirkungen eines Förderprogramms werden zwei Ebenen der Betrachtung unterschieden. Die Effekte für ausgesuchte Personengruppen sind Gegenstand der mikroökonometrischen Evaluation. Mit Hilfe makroökonometrischer Verfahren ist die Berücksichtigung gesamtgesellschaftlicher Effekte und „unerwünschter Nebenwirkungen“ wie Mitnahme- oder Substitutionseffekte möglich.[3] Die in diser Arbeit im Mittelpunkt stehenden Matchingverfahren gehören zur Gruppe der mikroökonometrischen Evaluationsverfahren.

[1] Im europäischen Raum beziehen sich die meisten Evaluationsstudien auf arbeitsmarktpolitische Instrumente in Deutschland und Schweden. Vgl. u.a. Fitzenberger und Prey (2000); Fitzenberger und Speckesser (2002); Hujer, Caliendo und Thomsen (2003); Hujer, Maurer und Wellner (1997); Lechner (1998); Lechner, Miquel und Wunsch (2004) bzw. Larsson (2003); Richardson und van den Berg (2001); Sianesi (2004).

[2] Die zeitnahe Evaluierung der Umsetzung der Gesetze für moderne Dienstleistungen am Arbeitsmarkt (Hartz-Gesetze) wurde im November 2002 vom Bundestag beschlossen (Kaltenborn, 2002). Mit der – in diesen Gesetzen beschlossenen – Gründung des Rates für Sozial- und Wirtschaftsdaten und der Einrichtung von Forschungsdatenzentren war eine entscheidende Verbesserung der Datenlage verbunden, die als eine wichtige Voraussetzung für den Einsatz und die Weiterentwicklung geeigneter Evaluationsverfahren anzusehen ist.

[3] Diese Effekte bezeichnen zum einen die „Mitnahme“ staatlicher Fördergelder für Entscheidungen, die so auch ohne Förderung getroffen worden wären, zum anderen die Verbesserung der Situation einer Personengruppe zulasten einer anderen. Erläuterungen zur makroökonomischen Evaluation finden sich u.a. in Arntz, Jacobebbinghaus und Spermann (2004), Bernhard et al. (2008) oder Calmfors, Forslund und Hemström (2002).

Zur Feststellung der Wirkung einer Fördermaßnahme wurden zu Beginn der mikroökonometrischen Evaluationsforschung parametrische Verfahren, v.a. die sog. Heckman-Korrektur (Heckman, 1978; 1979; 1980) und der Instrumentalvariablenansatz, eingesetzt. Im Rahmen solcher Modelle wird der Maßnahmeeffekt durch die Integration eines Zusatzterms bzw. einer zusätzlichen Variable ermittelt.[4] In der Folgezeit wurde allerdings festgestellt, dass die Einhaltung der dabei getroffenen Modellannahmen in empirischen Untersuchungen nicht immer gewährleistet ist. So wird in der Literatur v.a. bezweifelt, dass eine Förderung auf alle Personen die gleiche Wirkung hat und das Verhalten der beobachteten Personen mit den impliziten Verhaltensannahmen der Modelle übereinstimmt (Heckman, 1997, S. 446ff.).

Hier liegt der Vorteil nichtparametrischer Verfahren. Für ihre Anwendung werden keine Verhaltensannahmen und Annahmen über funktionale Zusammenhänge benötigt, und sie liefern auch konsistente Ergebnisse, wenn der Maßnahmeeffekt in der Bevölkerung heterogen ist (Schmidt, 1999, S. 16).

Innerhalb der nichtparametrischen Verfahren wird der Effekt einer Förderung durch den Vergleich der beobachteten Situation mit einer möglichst gut vergleichbaren Situation – meist der einer anderen Personengruppe – ermittelt. Diese Verfahren können als alternative Entwicklung zu sozialen Experimenten innerhalb der statistischen Literatur angesehen werden.[5] Im Unterschied zur Durchführung sozialer Experimente wird dabei keine neue Datenbasis geschaffen, sondern ein bereits bestehender Datensatz genutzt.

Alternativ zu den im folgenden Kapitel näher zu erläuternden Matchingverfahren werden in der Literatur weitere nichtparametrische Ansätze diskutiert.[6]

[4] Erläuterungen zu beiden Verfahren finden sich in Blundell und Costa Dias (2000) oder Heckman und Navarro-Lozano (2004). Alternative parametrische Ansätze werden u.a. in Heckman und Hotz (1989), Heckman, LaLonde und Smith (1999) oder Heckman und Robb (1985) vorgestellt.

[5] Datensätze aus sozialen Experimenten sind bisher vor allem für die USA verfügbar. Eins der bekanntesten Beispiele ist der Datensatz der National Supported Work Demonstration (NSW). Die Daten enthalten Informationen über ein Förderprogramm für benachteiligte Personengruppen, das in verschiedenen Regionen der USA angeboten wurde. Die Besonderheit dieses Programms besteht in der zufälligen Einteilung der teilnahmeberechtigten Personen in Teilnehmer und Nichtteilnehmer. Nähere Informationen zu diesem Programm finden sich in LaLonde (1986).
In Rosenbaum (2002), Kap. 2 wird eine ausführliche Darstellung der auf solchen Daten basierenden Verfahren gegeben.

[6] Im Folgenden wird die Grundidee dieser Schätzer kurz erläutert. Ausführliche Beschreibungen der Verfahren, der nötigen Annahmen sowie ihrer Vor- und Nachteile finden sich u.a. in Heckman, LaLonde und Smith (1999), Reinowski (2006) oder Schmidt (1999).

Der intuitiv am einfachsten zugängliche Schätzer vergleicht eine Personengruppe mit sich selbst – zu unterschiedlichen Zeitpunkten. Zur Ermittlung des Maßnahmeeffekts werden beim **Vorher-Nachher-Vergleich** die Ausprägungen einer vorher festgelegten Zielgröße in der Teilnehmergruppe vor und nach der Maßnahme miteinander verglichen. Dabei wird unterstellt, dass sich diese Zielgröße im Zeitablauf nicht verändern würde, wenn die betrachtete Personengruppe nicht an einer Fördermaßnahme teilnehmen würde.[7]

Der **Kreuz-Vergleich** stellt zu einem Zeitpunkt nach dem Ende der Maßnahme die Teilnehmer einer geeigneten Gruppe von Nichtteilnehmern gegenüber. Dabei wird angenommen, dass die beobachtete Zielgröße in der Gruppe der Nichtteilnehmer mit der hypothetischen Zielgröße der Teilnehmer übereinstimmt.[8] Die Zielgrößen dürfen dabei weder durch beobachtbare, nicht bei der Gruppenbildung berücksichtigte, noch durch unbeobachtbare Unterschiede zwischen den Merkmalen der Personen beider Gruppen beeinflusst werden.

Das **Differenz-von-Differenzen-Verfahren** ist eine Kombination aus beiden Verfahren. Hier werden nicht die Zielgrößen selbst miteinander verglichen, sondern deren Veränderungen. Dabei wird angenommen, dass der durchschnittliche Unterschied in der Zielgröße der Nichtteilnehmer vor und nach der Maßnahme dem hypothetischen Unterschied dieser Größe unter den Teilnehmern entspricht. Eine unverzerrte Schätzung des Maßnahmeeffekts ist möglich, wenn sich die Zielgrößenveränderung in den verglichenen Personengruppen im Durchschnitt nur aufgrund der Maßnahmeteilnahme der einen und der Nichtteilnahme der anderen Gruppe unterscheidet. Dies ist der Fall, wenn unbeobachtete Merkmale in beiden Personengruppen zeitinvariant sind und sich die Zielgrößen in den verglichenen Personengruppen im Durchschnitt gleich entwickeln.

Im Gegensatz zu diesen Verfahren werden beim Matching nicht pauschal zwei Personengruppen miteinander verglichen, sondern für jeden Teilnehmer einzeln eine (oder mehrere) Vergleichsperson(en) gesucht. Diese einzeln zugeordneten Vergleichsperso-

[7] Diese Annahme lässt Veränderungen in der Zielgröße zwischen beiden Beobachtungszeitpunkten für einzelne Personen zu, im Durchschnitt müssen diese Abweichungen aber ausgeglichen sein (Schmidt, 1999, S. 29).

[8] Die Zielgrößen einzelner Personen in den Gruppen können sich unterscheiden, im Durchschnitt müssen sich die Unterschiede aber aufheben (Schmidt, 1999, S. 30).

nen werden in einer Kontrollgruppe zusammengefasst und zur Ermittlung des Maßnahmeeffekts mit der Teilnehmergruppe verglichen. Das ist aufwändig, hat aber den Vorteil, dass für jeden Teilnehmer ein Pendant in der Kontrollgruppe existiert.

In der Literatur finden sich sehr unterschiedliche Vorschläge für die Identifizierung und Zuordnung geeigneter Vergleichspersonen mit Hilfe von Matchingverfahren. Die Frage nach der Bestimmung des „richtigen“ Verfahrens gewinnt deshalb zunehmend an Bedeutung. Allerdings gibt es keinen Matchingalgorithmus, der in jeder Situation den anderen überlegen ist. Bei der Auswahl muss berücksichtigt werden, dass sowohl die Fragestellung als auch die zur Verfügung stehende Datenbasis sowie die beobachtete Personengruppe und die potenziellen Vergleichsgruppen die Eignung der verfügbaren Matchingverfahren für die konkrete Anwendung beeinflussen.
Die Diskussion über die Vorzüge und Nachteile der einzelnen Schätzer sowie die Bedingungen, unter denen ein Verfahren besser ist als andere, nimmt in der jüngeren Evaluationsforschung einen breiten Raum ein. Allerdings ist die Suche nach allgemein gültigen Regeln für die Wahl und den Einsatz eines Matchingverfahrens noch nicht abgeschlossen (Imbens, 2004, S. 26).

Die vorliegende Arbeit ist als Beitrag zu dieser Diskussion zu sehen und soll ein weiterer Schritt auf dem Weg der „Standardisierung“ des Einsatzes von Matchingverfahren sein. Die Orientierung an praktisch vorzufindenden Entscheidungssituationen stellt dabei eine Ergänzung zu den meisten bisher in der Literatur zu findenden Studien dar. Die Realitätsnähe wird durch die Reproduktion eines existierenden Datensatzes im Rahmen einer Simulation erreicht. Sie hat zum einen Auswirkungen auf die betrachteten Stichproben – in der empirischen Forschung stehen i.d.R. relativ kleine Stichproben zur Verfügung – zum anderen auf die einsetzbaren Verfahren.

Im folgenden Kapitel werden verschiedene in der Literatur diskutierte Matchingverfahren erläutert. An alle Verfahren werden zwei Anforderungen gestellt. Zum einen ist es notwendig, die relevanten Merkmale im Matchingprozess adäquat zu berücksichtigen. Dazu muss ein Distanz- oder Ähnlichkeitsmaß verwendet werden, das in der Lage ist, die Informationen in den vorliegenden Daten möglichst gut wiederzugeben. Zum anderen müssen auf Grundlage dieses Maßes die besten Partner für jeden Teilnehmer aus der Gruppe der Nichtteilnehmer identifiziert und zugeordnet werden.

In der Arbeit werden die in der Statistik gebräuchlichen Ähnlichkeits- und Distanzmaße zur Zusammenfassung von Merkmalen eines Skalenniveaus sowie in der Evaluationsliteratur zu findende Balancing Scores dargestellt. Zusätzlich werden Methoden der Zusammenfassung verschiedener Merkmale aus anderen Wissenschaftsbereichen erläutert, die insbesondere die Berücksichtigung unterschiedlich skalierter Variablen bei der Bildung eines aggregierten Distanzmaßes ermöglichen. Es wird erwartet, dass diese aggregierten Distanzmaße besser als die bisher eingesetzten Balancing Scores in der Lage sind, die für Matching nötigen Informationen zusammenzufassen und damit eine bessere Grundlage für die Zuordnung passender Partner bieten.

Die Algorithmen, die für die Zuordnung von Partnern eingesetzt werden können, sind ebenfalls Gegenstand des zweiten Kapitels. Sie lassen sich nach der Anzahl der jeweils zugeordneten Partner, deren Gewichtung sowie der Möglichkeit der mehrfachen Zuordnung einer Person unterscheiden. Besonderes Augenmerk wird bei der Darstellung auf optimale Zuordnungsprozesse gelegt, die in der empirischen Evaluationsliteratur zwar die bestmögliche Zuordnung von Personen ermöglichen, aber bisher selten zu finden sind.

Die Darstellung der für den Matchingprozess einsetzbaren Vefahren wird ergänzt durch eine kurze Vorstellung von Kombinations- und Erweiterungsmöglichkeiten, die in der Literatur diskutiert werden.

Der Stand der Forschung zur Ermittlung von „Standards“ bei der Wahl geeigneter Matchingverfahren ist Gegenstand des dritten Kapitels. Der Frage, welches Verfahren in welcher konkreten Situation das beste ist, wird in der Literatur auf verschiedene Weise nachgegangen. Einige Autoren geben allgemeine Handlungempfehlungen, andere untersuchen die asymptotischen Eigenschaften der einzelnen Verfahren. Den größten Teil dieses Zweigs der Evaluationsforschung machen allerdings Vergleichsstudien aus, die unterschieden werden können nach der Art der verwendeten Daten. Sensitivitätsanalysen nutzen bereits vorhandene Daten, um die Ergebnisse der analysierten Verfahren mit einem Benchmark zu vergleichen. In Simulationsstudien wird eine eigene Datenbasis geschaffen, die die Untersuchung verschiedener Eigenschaften der Verfahren erlaubt.

In der Arbeit wird ein Überblick über die jeweils wichtigsten Studien und ihre Ergebnisse gegeben.

Aufbauend auf diesen Untersuchungen wird eine eigene Simulationsstudie durchgeführt, die im vierten Kapitel erläutert wird. Im Mittelpunkt dieser Studie steht die Frage, welches Distanzmaß am besten in der Lage ist, unterschiedlich skalierte Merkmale, die im Matchingprozess berücksichtigt werden, zusammenzufassen, und welcher Zuordnungsprozess die Identifikation und Zuordnung der besten Partner für die betrachteten Personen ermöglicht. In die Untersuchung werden diejenigen Distanzmaße und Zuordnungsprozesse einbezogen, die sich in früheren Studien als vorteilhaft gegenüber anderen erwiesen haben, sowie zusätzlich zwei der vorgestellten aggregierten Distanzmaße und zwei optimale Zuordnungsprozesse.
In der Simulation werden jeweils zwei Personengruppen – eine Teilnehmer- und eine Nichtteilnehmergruppe – kombiniert, die sich in verschiedenen Aspekten unterscheiden. In den einzelnen Schritten der Analyse wird die Größe der Stichprobe insgesamt, das Verhältnis der Teilnehmer- zur Nichtteilnehmeranzahl sowie das Ausmaß der Übereinstimmung der betrachteten Matchingvariablen variiert. Der Fokus liegt dabei auf kleinen Stichproben. Die Grundlage der Analyse bildet ein in der Arbeitsmarktforschung häufig verwendeter Datensatz, der für die Studie „nachgebildet“ wird, um eine größtmögliche Nähe zu real vorzufindenden Entscheidungssituationen zu erreichen.
Zur Beurteilung der Ergebnisse der einzelnen Verfahren werden neben den in der Evaluationsliteratur gebräuchlichen Gütemaßen nichtparametrische skalenpezifische Tests zur Überprüfung der Übereinstimmung der Mittelwerte bzw. Häufigkeitsverteilungen der verwendeten Variablen eingesetzt.

Den Abschluss der Arbeit bildet ein Anwendungsbeipiel. Mit Hilfe eines der vorgestellten Matchingverfahren werden die Beschäftigungseffekte der Förderung der Berufsausbildung in Ostdeutschland ermittelt. Mit der empirischen Untersuchung soll die Frage beantwortet werden, ob die Absolventen betriebsnaher und außerbetrieblicher Berufsausbildungsgänge schlechtere Chancen auf einen ihrer Qualifikation entsprechenden Berufseinstieg haben als Absolventen ungeförderter Ausbildungen. Dazu wird zum einen der quantitative Effekt der Förderung ermittelt, zum anderen werden verschiedene qualitative Merkmale der aufgenommenen Erwerbstä-

tigkeit betrachtet. Als Datenbasis dient das Jugendpanel des Zentrums für Sozialforschung Halle, aus dem eine Stichprobe aller Jugendlichen, die eine Berufsausbildung erfolgreich abgeschlossen haben, gezogen wird.

Kapitel 2

Vorstellung verschiedener Matchingverfahren

Die Idee der Evaluation der Entscheidungen von Personen besteht im Ex-post-Vergleich der Situation, die aus einer Entscheidung resultiert, mit derjenigen, die sich aus einer alternativen Entscheidung ergeben hätte. Der Unterschied zwischen beiden Situationen gibt Aufschluss über die Wirkung der getroffenen Entscheidung. Im Folgenden wird zur Veranschaulichung unterstellt, dass Entscheidungen von Personen hinsichtlich der Teilnahme oder Nichtteilnahme an einer Fördermaßnahme auf dem Arbeitsmarkt evaluiert werden. Als Kriterium zur Beurteilung des Erfolgs dieser Maßnahme dient das Einkommen.

2.1 Theoretische Grundlage

Für die Darstellung des Maßnahmeeffekts wird i.d.R. das Modell der potenziellen Einkommen verwendet.[1] Das Modell basiert auf der Annahme, dass die Teilnahmeentscheidung von einem nutzenmaximierenden Individuum auf Grundlage seines individuellen erwarteten Nutzens aus der Maßnahme getroffen wird.[2] Dabei wird unterstellt, dass für die Akteure die persönliche Situation ohne Teilnahme repräsentativ für die persönliche Situation ohne Maßnahme ist – d.h. die Maßnahme hat keine Auswirkungen auf Personen, die nicht an ihr teilnehmen.[3]

Der individuelle Effekt einer Fördermaßnahme ME_{it} wird innerhalb des Modells als Unterschied zwischen zwei potenziellen Einkommen einer Person i zum Zeitpunkt t

[1] Dieses Modell findet sich in der Literatur unter verschiedenen Namen. Es wird auch als Switching Regression Model (Heckman, LaLonde und Smith, 1999, S. 1915), Roy-Modell der Einkommensverteilung (Heckman und Vytlacil, 1999, S. 4730) bzw. kurz als Roy-Rubin-Modell (Hujer und Caliendo, 2000, S. 9) bezeichnet.

[2] Der Nutzen wird definiert als der Gegenwartswert der erwarteten Einnahmen nach Maßnahmeteilnahme, abzüglich der Teilnahmekosten (die sich aus direkten Kosten und entgangenen Einnahmen aufgrund der Teilnahme zusammensetzen). Individuelle Erwartungen werden unter Verwendung der dem Entscheidungsträger zur Verfügung stehenden Informationen gebildet und können daher – je nach Informationsquelle – zwischen den Teilnehmern variieren. Für nähere Erläuterungen zur unterstellten Entscheidungsregel und den getroffenen Annahmen vgl. Heckman, LaLonde und Smith (1999) S. 1915f..

[3] Diese Annahme impliziert, dass nur direkte Effekte auf die teilnehmenden Personen betrachtet werden und indirekte Effekte (z.B. Steuerzahlungen zur Finanzierung der Maßnahme oder Verdrängungseffekte auf dem regionalen Arbeitsmarkt) unberücksichtigt bleiben (Heckman, LaLonde und Smith, 1999, S. 1880).

dargestellt – dem Einkommen Y_{it}^T nach Teilnahme an einer Fördermaßnahme und dem ohne Maßnahmeteilnahme Y_{it}^C:

$$ME_{it} = Y_{it}^T - Y_{it}^C. \tag{2.1}$$

Eins der beiden Einkommen ist allerdings für ein und dieselbe Person zu einem Zeitpunkt nicht beobachtbar – und der individuelle Effekt ME_{it} damit nicht ermittelbar.[4] Das Grundproblem der Evaluation ist also ein Problem fehlender Informationen. Der individuelle Effekt einer Fördermaßnahme ist nicht ermittelbar, da eine der beiden Situationen nicht beobachtbar ist und damit der Vergleichsmaßstab für das beobachtete Einkommen fehlt.

Anstelle des individuellen Effekts wird deshalb ein durchschnittlicher Maßnahmeeffekt bestimmt.[5] Grundlage für die Ermittlung dieses Effekts sind die beobachteten individuellen Einkommen Y_{it}:

$$Y_{it} = D_i \cdot Y_{it}^T + (1 - D_i)\, Y_{it}^C. \tag{2.2}$$

Die Dummy-Variable D_i gibt an, welches der beiden potenziellen Einkommen tatsächlich beobachtet wird: das Teilnahmeeinkommen Y_{it}^T, wenn die betrachtete Person an einer Fördermaßnahme teilgenommen hat ($D_i = 1$), oder das Nichtteilnahmeeinkommen Y_{it}^C, wenn sie nicht teilgenommen hat ($D_i = 0$).
Der durchschnittliche Effekt für die Teilnehmer wird ermittelt aus der Differenz der durchschnittlichen Teilnehmereinkommen nach der Maßnahme und dem durch-

[4] Für einen Teilnehmer lässt sich nicht feststellen, wie hoch sein Einkommen gewesen wäre, wenn er nicht an der Fördermaßnahme teilgenommen hätte. Und für eine nicht geförderte Person kann kein Einkommen nach Teilnahme an einer Maßnahme ermittelt werden.
Die unbeobachtbare der beiden Situationen wird in der Literatur als hypothetisch bzw. counterfactual (outcome) bezeichnet (Hujer und Caliendo, 2000, S. 10).

[5] In der Literatur werden unterschiedliche Schätzer des durchschnittlichen Maßnahmeeffekts diskutiert. Einen systematischen Überblick geben u.a. Hübler (2001) und Imbens (2004). Da diese Schätzer infolge der Heterogenität des Maßnahmeeffekts in der Bevölkerung unterschiedliche Ergebnisse liefern, ist eine gründliche Überlegung über die gewünschte Aussage nötig (Heckman, 1997, S. 443ff.).
Für die Evaluation von Fördermaßnahmen wird i.d.R. der durchschnittliche Effekt für die Teilnehmer an einer Maßnahme eingesetzt. Er ist interpretierbar als Antwort auf die Frage „Wie groß ist der durchschnittliche Vorteil einer Person, der sich durch eine Maßnahmeteilnahme ergibt, im Vergleich zu seiner Situation ohne Teilnahme?“. Dieser Effekt dient im Folgenden als Beispiel für den zu ermittelnden Maßnahmeeffekt.

schnittlichen hypothetischen Einkommen, das die Teilnehmer erzielt hätten, wenn sie nicht teilgenommen hätten:

$$E\left(ME_t|D=1\right) = E\left(Y_t^T|D=1\right) - E\left(Y_t^C|D=1\right). \tag{2.3}$$

Dabei wird der durchschnittliche Maßnahmeeffekt mit $E\left(ME_t|D=1\right)$, das Einkommen der Teilnehmer nach der Maßnahme mit $E\left(Y_t^T|D=1\right)$ und das hypothetische Einkommen mit $E\left(Y_t^C|D=1\right)$ bezeichnet.

Da man für die Teilnehmergruppe kein Nichtteilnahmeeinkommen beobachten kann, muss für den letzten Term der Gleichung ein geeigneter Ersatz gefunden werden, so dass gilt: $E\left(Y_t^C|D=1\right) - E\left(Y_t^C|D=0\right) = 0$. Die Indikatorvariable D gibt dabei an, ob die beobachteten Personen zum Zeitpunkt t Teilnehmer $(D=1)$ oder Nichtteilnehmer $(D=0)$ sind.

Beim Matching wird das hypothetische Einkommen der Teilnehmer durch das Durchschnittseinkommen geeigneter Nichtteilnehmer ersetzt. Dabei wird für jeden Teilnehmer einzeln nach passenden Personen gesucht. Die Einkommen dieser in einer Kontrollgruppe zusammengefassten Nichtteilnehmer bilden die Vergleichsgröße zur Ermittlung des Maßnahmeeffekts. Damit die o.g. Bedingung erfüllt ist, dürfen in der Kontrollgruppe nur Personen zu finden sein, die sich in keinem für das Einkommen und die Maßnahmeteilnahme relevanten Merkmal von den Teilnehmern unterscheiden.[6]

Wenn keine solche Kontrollgruppe gefunden werden kann, unterscheiden sich die durchschnittlichen Einkommen in beiden Gruppen auch ohne die Teilnahme der einen und die Nichtteilnahme der anderen Gruppe, was zu einer Über- oder Unterschätzung des Maßnahmeeffekts führt (Caliendo und Hujer, 2006, S. 202). Dieses Problem $\left(\left(Y_t^C|D=1\right) - E\left(Y_t^C|D=0\right) \neq 0\right)$ wird als Selektionsverzerrung bezeichnet.[7]

[6] Als relevant gelten dabei solche Merkmale, die die Zuordnung zu einer der beiden Gruppen und das potenzielle Einkommen beeinflussen (Eichler und Lechner, 2001, S. 230). Dies gilt sowohl für beobachtbare Merkmale (wie Qualifikation, Alter, Geschlecht) als auch für unbeobachtbare (z.B. Motivation, Selbstbewusstsein).

[7] In Konle-Seidl (2005) wird zwischen negativer und positiver Selektion unterschieden. Eine negative Selektion liegt vor, wenn die Teilnehmer an einer Maßnahme einer Problemgruppe am Arbeitsmarkt angehören. In diesem Fall würde der Maßnahmeeffekt unterschätzt, wenn das Ein-

Um den Maßnahmeeffekt als kausalen Zusammenhang zwischen Maßnahmeteilnahme und Einkommenshöhe interpretieren zu können, müssen zwei weitere Voraussetzungen erfüllt sein. Zum einen dürfen die berücksichtigten Merkmale nicht durch die Teilnahmeentscheidung beeinflusst werden. Die Exogenität der erklärenden Variablen ist insbesondere für zeitveränderliche Merkmale, die kurz vor Beginn oder nach Beendigung der Maßnahme beobachtet werden, zu beachten.[8] Zum anderen muss die Ermittlung eines Durchschnittseffekts unabhängig von Größe und Zusammensetzung der Teilnehmergruppe möglich sein. Anders ausgedrückt: Der Maßnahmeeffekt auf eine Person darf nicht beeinflusst werden von der Teilnahme anderer Personen an der gleichen Maßnahme.[9]

In der folgenden Darstellung der Annahmen des Matchingansatzes und der einzelnen Matchingverfahren werden drei Personengruppen unterschieden. Die Teilnehmergruppe – im folgenden gekennzeichnet mit dem Index $.^T$ – besteht aus allen Personen $i = 1, \ldots, I$, die an einer bestimmten Fördermaßnahme teilgenommen haben. Das erzielte individuelle Einkommen dieser Personen wird mit Y_i^T bezeichnet. Die Nichtteilnehmergruppe wird im Vorfeld der Evaluation definiert[10] und trägt im Folgenden den Index $.^{NT}$. Sie besteht aus allen potenziellen Partnern $j = 1, \ldots, J$ für die Teilnehmer. Aus dieser Gruppe wird die Kontrollgruppe gebildet, die mit dem Index $.^C$ bezeichnet wird. Sie setzt sich aus den Mitgliedern der einzelnen

kommen der Teilnehmer mit dem Durchschnittseinkommen aller anderen Personen verglichen würde. Die Nichtbeachtung einer Positivselektion (die Teilnehmer haben von vornherein eine bessere Position auf dem Arbeitsmarkt) würde zur Überschätzung der Wirkung führen.

[8] Ein Beispiel für die offensichtliche Verletzung dieser Annahme wäre die Berücksichtigung des Qualifikationsniveaus nach Beendigung einer Weiterbildungsmaßnahme. Zwar determiniert das Qualifikationsniveau das potenziell erzielbare Einkommen auf dem Arbeitsmarkt und wäre insofern wichtig für die Vergleichbarkeit zweier Personen. Da die Qualifikation eines Teilnehmers aber eine direkte Folge der Fördermaßnahme ist, ist dieses Merkmal nicht mehr exogen. Es müsste statt dessen die Qualifikation vor Beginn der Förderung berücksichtigt werden.

[9] In der Literatur wird diese Annahme als Stable Unit Treatment Value Assumption (SUTVA) bezeichnet (Rubin, 1986, S. 961). In der Praxis ist sie nicht selbstverständlich erfüllt. So können bspw. große Teilnehmerzahlen an gleichartigen Förderprogrammen in einer Region die Beschäftigungsaussichten eines einzelnen Teilnehmers nach Beendigung einer Maßnahme negativ beeinflussen.

[10] Die Nichtteilnehmergruppe umfasst i.d.R. alle Personen, die bestimmte Auswahlkriterien erfüllen, z.B. der gleichen Altersgruppe wie die Teilnehmer angehören oder – im Fall arbeitsmarktpolitischer Maßnahmen – ebenfalls arbeitslos sind (Reinowski, Schultz und Wiemers (2005); Sianesi (2004)). Es ist auch möglich, die Teilnehmer an einer anderen Maßnahme als Nichtteilnehmer zu verwenden (Lechner, Miquel und Wunsch, 2004).

Unter-Kontrollgruppen C_i für jeden der Teilnehmer zusammen. Die individuellen Einkommen der Personen in der Kontrollgruppe werden mit Y_j^C bezeichnet.

2.2 Annahmen

Für die Bildung der Kontrollgruppe gelten zwei Annahmen. Die Annahme der bedingten Unabhängigkeit (CIA) besagt, dass für alle Personen mit gleichen Merkmalen das potenzielle Einkommen bei Teilnahme und Nichtteilnahme übereinstimmt:[11]

$$Y_t^T, Y_t^C \perp D|X. \tag{2.4}$$

Dabei steht $\perp$ für die Unabhängigkeit der Einkommen Y_t^T und Y_t^C in der Teilnehmer- und der Kontrollgruppe. Anders ausgedrückt: Gegeben die relevanten Merkmale X, ist die Höhe des potenziell erzielten Einkommens in der Teilnehmer- und der Kontrollgruppe unabhängig davon, welche Personen diesen Gruppen zugeordnet werden. Das bedeutet, dass alle Merkmale, die den Selektionsprozess und das potenzielle Einkommen beeinflussen, beobachtbar sein müssen und bei der Schätzung des Maßnahmeeffekts berücksichtigt werden.

Unter Gültigkeit der Annahme der bedingten Unabhängigkeit ist die Identifikation des Durchschnittseffekts für Personen mit den beobachteten Merkmalen theoretisch möglich. Um den Effekt tatsächlich schätzen zu können, ist eine weitere Annahme nötig. Es müssen sowohl Teilnehmer als auch Nichtteilnehmer mit den relevanten Merkmalen X zu finden sein:

$$0 < \Pr(D = 1|X) < 1. \tag{2.5}$$

Diese Annahme wird als Overlap (Imbens, 2004, S. 7) oder Common Support Condition (CSC) (Lechner, 2001b, S. 5) bezeichnet. Sie ist eine notwendige Bedingung,

[11] Diese Annahme ist in der Literatur unter verschiedenen Namen zu finden: Conditional Independence Assumption (Lechner, 2001b, S. 44), Ignorable Treatment Assignment (Rosenbaum und Rubin, 1983, S. 43) oder Unconfoundedness (Imbens, 2004, S. 7). In der letztgenannten Quelle wird sie mit der Standardannahme der Exogenität der erklärenden Variablen in Regressionsmodellen verglichen.

weil nur innerhalb des Common-Support-Bereichs die Schätzung der beiden potenziellen Einkommen möglich ist. Außerhalb des Common-Support-Bereichs kann entweder nur Y_t^T oder nur Y_t^C geschätzt werden (Imbens, 2004, S. 8). Der durchschnittliche Maßnahmeeffekt kann also nur für Personen innerhalb des CS-Bereichs bestimmt werden.[12]

Für die Ermittlung des durchschnittlichen Maßnahmeeffekts für die Teilnehmer können diese Annahmen abgeschwächt werden. Da das durchschnittliche Einkommen der Teilnehmer direkt aus den beobachteten Daten ermittelt werden kann, muss die CIA nur für das Nichtteilnehmereinkommen gelten (Imbens, 2004, S. 8):

$$Y_t^C \perp D|X. \tag{2.6}$$

Weiterhin ist es ausreichend, dass für jeden Teilnehmer ein ähnlicher Nichtteilnehmer zu finden ist (Smith und Todd, 2005a, S. 313):

$$\Pr(D = 1|X) < 1. \tag{2.7}$$

2.3 Ermittlung des Maßnahmeeffekts

Wenn diese Annahmen erfüllt sind, lässt sich der durchschnittliche Maßnahmeeffekt für die Teilnehmer konsistent aus der durchschnittlichen Einkommensdifferenz zwischen Teilnehmern und Kontrollgruppe ermitteln. Der Unterschied zum Kreuz-Vergleich besteht darin, dass nicht mehr alle Nichtteilnehmer berücksichtigt werden und das gleiche Gewicht erhalten, sondern die Gewichtung jedes Nichtteilnehmers für verschiedene Teilnehmer variiert. Anstelle des Vergleichs zweier Durchschnitts-

[12] In der konkreten Anwendung ist die Einhaltung dieser Bedingung daher zu prüfen und die Schätzung des Maßnahmeffekts auf die Personen zu beschränken, die innerhalb des Common-Support-Bereichs liegen. In Caliendo und Kopeinig (2005) werden die in der empirischen Literatur zu findenden Verfahren zur Prüfung der Einhaltung der CSC zusammengefasst.
Wenn durch die Bereinigung ein großer Teil der Stichprobe nicht mehr berücksichtigt werden kann, ist der Schätzer nicht mehr aussagefähig für die Stichprobe. In diesem Fall sollten die nicht berücksichtigten Personen näher untersucht werden, um den Maßnahmeeffekt für die verbleibenden Personen richtig interpretieren zu können. Lechner (2001a) schlägt bspw. die Nutzung dieser Informationen zur Schätzung von Schranken für den Maßnahmeeffekt vor.

einkommen wird der Maßnahmeeffekt ME_{M_t} als gewichteter Durchschnitt individueller Einkommensdifferenzen ermittelt:

$$ME_{M_t} = \sum_{i=1}^{I} w\,(i) \left(Y_{it}^{T} - \bar{Y}_{it}^{C}\right). \tag{2.8}$$

Die Anzahl der berücksichtigten Einkommensdifferenzen wird dabei mit $i = 1, \ldots, I$ bezeichnet, die Gewichtung jeder einzelnen Differenz mit $w\,(i)$. Y_{it}^{T} gibt das Einkommen eines Teilnehmers i zum Zeitpunkt t, $\bar{Y}_{it}^{C}$ sein Vergleichseinkommen an, das aus dem gewichteten Durchschnitt aller Nichtteilnehmereinkommen in seiner Unter-Kontrollgruppe gebildet wird:

$$\bar{Y}_{it}^{C} = \sum_{j \in C_i} W\,(i,j)\, Y_{jt}^{C}. \tag{2.9}$$

Mit Y_{jt}^{C} wird das Einkommen eines Nichtteilnehmers j zum Zeitpunkt t bezeichnet, mit $W\,(i,j)$ die individuelle Gewichtung dieses Einkommens für Teilnehmer i. Für die Gewichtung der einzelnen Nichtteilnehmereinkommen gilt:

$$0 \leq W\,(i,j) \leq 1 \;\text{ und }\; \sum_{j \in C_i} W\,(i,j) = 1. \tag{2.10}$$

Die Anzahl der Nichtteilnehmer in der Unter-Kontrollgruppe C_i eines Teilnehmers wird durch J^{C_i} angegeben.

Wie groß die Unter-Kontrollgruppe jedes Teilnehmers ist und wie die einzelnen Einkommen gewichtet werden, hängt von der Wahl des Zuordnungsprozesses ab.

Vor der Auswahl eines Zuordnungsprozesses muss allerdings die Übereinstimmung zwischen Teilnehmern und Nichtteilnehmern in den relevanten Merkmalen ermittelt werden.
Eine Möglichkeit besteht in der Überprüfung jedes einzelnen Merkmals für jeden Teilnehmer und jeden Nichtteilnehmer. Dieses Verfahren wird als exaktes Matching bezeichnet (Schmidt, 1999, S. 26). Mit seiner Anwendung ist das sog. Dimensionsproblem verbunden: Je mehr Merkmale berücksichtigt werden, desto größer ist die Gefahr, dass für einige Personen keine möglichen Partner mit den gleichen Merkmalen gefunden werden können (Black und Smith, 2004, S. 109f.). Darüber hinaus

steigt der Rechenaufwand mit jedem berücksichtigten Merkmal exponentiell.[13] Besonders problematisch ist in diesem Zusammenhang die Berücksichtigung metrisch skalierter und nominaler Merkmale mit einer großen Anzahl möglicher Ausprägungen.
Eine Alternative bietet die Zusammenfassung der merkmalsspezifischen Informationen in einem Ähnlichkeits- oder Distanzmaß.

2.3.1 Ähnlichkeits- und Distanzmaße

Die Feststellung der Ähnlichkeit zwischen zwei Objekten anhand eines Maßes hat – neben ihrer Verwendung in der Evaluation – sehr verschiedene Einsatzgebiete, z.B. in der Clusteranalyse, der Mustererkennung oder der Biometrie. Entsprechend ihres Einsatzgebietes finden sich sehr unterschiedliche Ähnlichkeits- bzw. Distanzmaße in der Literatur. Welches konkrete Maß im Zusammenhang mit der Evaluation eingesetzt wird, hängt im Wesentlichen von der Art der beobachteten Merkmale ab. Für qualitative (nominal und ordinal skalierte) Merkmale werden i.d.R. Ähnlichkeitskoeffizienten ermittelt, für quantitative Merkmale Distanzmaße. Beide Maße lassen sich – nach vorheriger Normierung – ineinander überführen.

Zunächst werden Ähnlichkeits- bzw. Distanzmaße für einheitlich skalierte Merkmale erläutert. Für die empirische Anwendung ist allerdings von größerer Bedeutung, wie sich unterschiedlich skalierte Merkmale in einem gemeinsamen Distanzmaß zusammenfassen lassen. Diese Frage ist Gegenstand der darauffolgenden Abschnitte.

Nominal skalierte dichotome Merkmale

Nominal skalierte Variablen besitzen mehrere Ausprägungen, die sich nicht in eine Rangfolge bringen lassen. Eine Variable mit zwei möglichen Ausprägungen wird als dichotom bezeichnet.

[13] In Anlehnung an Hujer, Caliendo und Radić (2001) S. 10 lässt sich dieses Problem mit Hilfe eines Beispiels verdeutlichen. Die Ähnlichkeit von Personen soll anhand von $n = 2$ Merkmalen überprüft werden, für die jeweils zwei Ausprägungen zu beobachten sind (z.B. Beschäftigungsstatus: beschäftigt/nicht beschäftigt; abgeschlossene Berufsausbildung: ja/nein). Es ergeben sich $2^2 = 2^n$ zu überprüfende Matching-Möglichkeiten. Die Berücksichtigung eines zusätzlichen Merkmals (z.B. Alter: bis 45 Jahre/älter) erhöht diese Anzahl auf $2^3 = 2^{n+1}$.

Für die gemeinsame Beobachtung zweier Objekte hinsichtlich ihrer Ausprägungen in dichotomen Variablen lassen sich folgende Fälle unterscheiden: das gemeinsame Auftreten der Ausprägung Eins bei beiden Objekten, das Auftreten von Null bei einem und Eins bei dem anderen Objekt sowie ein gemeinsames Auftreten der Null bei beiden Objekten. Diese Fälle können mit einer 4-Felder-Tafel veranschaulicht werden (vgl. Abbildung 2.1).

		Objekt i	
		0	1
Objekt j	0	$z_{i0,j0}$	$z_{i1,j0}$
	1	$z_{i0,j1}$	$z_{i1,j1}$

Abbildung 2.1: *Kombination der Merkmalsausprägungen eines dichotomen Merkmals für zwei Objekte*

Quelle: Eigene Darstellung in Anlehnung an Sokal und Sneath (1963) S. 126.

Dabei bezeichnet $z_{i1,j1}$ die Anzahl der beobachteten Übereinstimmungen der Ausprägung Eins, $z_{i0,j0}$ die Anzahl der übereinstimmenden Ausprägungen Null, $z_{i1,j0}$ und $z_{i0,j1}$ die Anzahl der Merkmale mit jeweils unterschiedlicher Ausprägung: $z_{i1,j0}$ steht für den Fall Eins bei Objekt i und Null bei Objekt j, $z_{i0,j1}$ für den umgekehrten Fall.

Für die Feststellung der Ähnlichkeit von Objekten anhand von dichotomen Merkmalen existiert eine Reihe von Ähnlichkeitsmaßen, die sich darin unterscheiden, ob und in welcher Form übereinstimmende und nicht übereinstimmende Ausprägungen der Merkmale berücksichtigt werden.

Die meisten Maße lassen sich durch folgende Funktion zusammenfassen (Backhaus et al., 2000, S. 484):

$$s_{ij} = \frac{z_{i1,j1} + \theta\, z_{i0,j0}}{z_{i1,j1} + \theta\, z_{i0,j0} + \lambda\,(z_{i1,j0} + z_{i0,j1})}. \tag{2.11}$$

Die Ähnlichkeit s_{ij} zwischen zwei Objekten i und j wird bestimmt durch den Anteil ihrer übereinstimmenden Merkmale $z_{i1,j1} + z_{i0,j0}$ an allen beobachteten Merkmalen $N = z_{i1,j1} + z_{i1,j0} + z_{i0,j1} + z_{i0,j0}$. Mit den beiden Gewichtungsfaktoren θ und λ wird festgelegt, welche der Terme bei der Feststellung der Ähnlichkeit berücksichtigt und wie stark sie gewichtet werden.

Die Gewichtung der übereinstimmenden Ausprägungen wird mit dem Faktor θ beeinflsst. Sollen die Übereinstimmungen in beiden Ausprägungen gleich gewichtet werden, gilt $\theta = 1$. Im Fall $\theta = 0$ wird nur eine gemeinsame Ausprägung Eins berücksichtigt. Mit dem Faktor λ wird der Grad des Einflusses nichtübereinstimmender Merkmale auf die Ähnlichkeitsfeststellung zweier Objekte determiniert. $\lambda < 1$ bedeutet eine stärkere Betonung der Übereinstimmungen, $\lambda > 1$ relativiert das Gewicht der Übereinstimmungen im Ähnlichkeitsmaß durch eine stärkere Berücksichtigung der nicht übereinstimmenden Merkmale. Sollen die übereinstimmenden Merkmale im Verhältnis zu allen beobachteten Merkmalen betrachtet werden, muss gelten $\lambda = 1$.
Die Auswahl eines Ähnlichkeitsmaßes für die empirische Anwendung wird wesentlich dadurch beeinflusst, ob die gemeinsame Ausprägung Null in einem Merkmal einen eigenen Aussagegehalt besitzt.

Der Matchingkoeffizient Der Matchingkoeffizient (Kaufmann und Pape, 1996) wird eingesetzt, wenn Übereinstimmungen in den Ausprägungen Null und Eins den gleichen Aussagegehalt haben. Er gibt den Anteil der Merkmale mit übereinstimmenden Ausprägungen an der Gesamtzahl der untersuchten Merkmale an. Übereinstimmungen in beiden Ausprägungen der Merkmale werden gleich gewichtet (es gilt $\theta = 1$ und $\lambda = 1$):

$$\begin{aligned} MC_{ij} &= \frac{z_{i1,j1} + z_{i0,j0}}{N} \\ &= \frac{1}{N}\left(\sum_{n=1}^{N} Q_1(x_{ni}, x_{nj}) + \sum_{n=1}^{N} Q_0(x_{ni}, x_{nj})\right). \end{aligned} \tag{2.12}$$

Dabei bezeichnet MC_{ij} den Matchingkoeffizienten, $Q_1(x_{ni}, x_{nj})$ und $Q_0(x_{ni}, x_{nj})$ sind Indikatoren für die Übereinstimmung der Objekte i und j im Merkmal x_n. Sie werden wie folgt definiert:

$$Q_1(x_{ni}, x_{nj}) = \begin{cases} 1 & \text{wenn } x_{ni} = x_{nj} = 1 \\ 0 & \text{sonst} \end{cases} \quad \text{und}$$

$$Q_0(x_{ni}, x_{nj}) = \begin{cases} 1 & \text{wenn } x_{ni} = x_{nj} = 0 \\ 0 & \text{sonst.} \end{cases}$$

Der Jaccardkoeffizient Wenn die Übereinstimmung in der Ausprägung Null keinen Aussagegehalt besitzt, wird sie bei der Ähnlichkeitsermittlung nicht berücksichtigt. (Dann gilt $\theta = 0$ und $\lambda = 1$.) In diesem Fall kann bspw. der Jaccardkoeffizient angewendet werden (Backhaus et al., 2000):[14]

$$\begin{aligned} JC_{ij} &= \frac{z_{i1,j1}}{z_{i1,j1} + z_{i1,j0} + z_{i0,j1}} \\ &= \frac{1}{N - \sum_{n=1}^{N} Q_0(x_{ni}, x_{nj})} \sum_{n=1}^{N} Q_1(x_{ni}, x_{nj}). \end{aligned} \tag{2.13}$$

Der Jaccardkoeffizient JC_{ij} gibt den Anteil der übereinstimmenden Ausprägungen Eins an allen Merkmalen an, für die mindestens eins der Objekte die Ausprägung Eins besitzt.

Beide Ähnlichkeitsmaße können als „Grundformen“ aufgefasst werden – der Matchingkoeffizient für die Berücksichtigung aller Übereinstimmungen, der Jaccardkoeffizient für die Berücksichtigung der übereinstimmenden Ausprägung Eins. Die Anwendung der beiden Koeffizienten resultiert (mit Ausnahme des Spezialfalls $z_{i0,j0} = 0$) in unterschiedlichen Ähnlichkeits-Rangfolgen (Kaufmann und Pape, 1996, S. 445).
Daneben existieren zahlreiche weitere Maße, die sich durch ihre Gewichtung der Anzahlen $z_{i1,j1}$, $z_{i1,j0}$, $z_{i0,j1}$ und $z_{i0,j0}$ von den beiden vorgestellten Maßen unterscheiden.[15]

Nominal skalierte polytome Merkmale

Merkmale mit mehr als zwei Ausprägungen, die sich nicht in eine Rangfolge bringen lassen, werden als polytome Variablen bezeichnet. Für die Feststellung der Ähnlichkeit zweier Objekte anhand nominal skalierter polytomer Merkmale finden sich weniger zahlreiche Maße in der Literatur.[16]

[14] In der Literatur findet sich dieses Maß auch unter dem Namen Tanimoto- bzw. Similarity-Koeffizient (Steinhausen und Langer, 1977, Tab. 3.2.1.2).

[15] Eine Übersicht über weitere Ähnlichkeitsmaße findet sich bspw. in Cheetham und Hazel (1969) oder Sokal und Sneath (1963).
Bei der Anwendung einiger dieser Maße ergeben sich die gleichen Ähnlichkeits-Rangfolgen wie beim Matching- und dem Jaccardkoeffizienten (Steinhausen und Langer, 1977, S. 54).

[16] Anders als bei der Ähnlichkeitsfeststellung dichotomer Variablen wird in ihnen eine gemeinsame Ausprägung Null gleich gewichtet wie Übereinstimmungen in anderen Ausprägungen.

Der verallgemeinerte Matchingkoeffizient Eine einfache Möglichkeit der Ähnlichkeitsmessung bietet der verallgemeinerte Matchingkoeffizient (Kaufmann und Pape, 1996), bei dem – analog zum Matchingkoeffizienten für dichotome Variablen – die Summe der Merkmale mit übereinstimmenden Ausprägungen mit der Gesamtzahl der beobachteten Merkmale gewichtet wird :

$$gMC_{ij} = \frac{z_s}{N} = \frac{1}{N} \sum_{n=1}^{N} Q_s(x_{ni}, x_{nj}). \tag{2.14}$$

Dabei steht gMC_{ij} für den verallgemeinerten Matchingkoeffizienten, z_s für die Anzahl der Merkmale mit übereinstimmenden Ausprägungen, $Q_s(x_{ni}, x_{nj})$ ist ein Indikator für die Übereinstimmung von Objekt i und Objekt j im Merkmal x_n:

$$Q_s(x_{ni}, x_{nj}) = \begin{cases} 1 & \text{wenn } \; x_{ni} = x_{nj} \\ 0 & \text{sonst.} \end{cases}$$

Alle übereinstimmenden Merkmale werden dabei gleich gewichtet, unabhängig von der Anzahl ihrer möglichen Ausprägungen.

Das Ähnlichkeitsmaß von Hyvärinen Wenn die unterschiedliche Anzahl möglicher Ausprägungen im Ähnlichkeitsmaß berücksichtigt werden soll, kann das Ähnlichkeitsmaß von Hyvärinen (1962) angewendet werden. In diesem Maß erhalten Übereinstimmungen in Merkmalen mit vielen möglichen Ausprägungen ein höheres Gewicht als solche in Merkmalen mit wenigen:

$$HC_{ij} = \sum_{n=1}^{N} Q_v(x_{ni}, x_{nj}). \tag{2.15}$$

HC_{ij} steht für das Ähnlichkeitsmaß von Hyvärinen, $Q_v(x_{ni}, x_{nj})$ ist ein Indikator für die Übereinstimmung von Objekt i und Objekt j im Merkmal x_n, in dem die Anzahl der möglichen Ausprägungen einer Variable V_n berücksichtigt wird. Er wird wie folgt definiert:

$$Q_v(x_{ni}, x_{nj}) = \begin{cases} V_n & \text{wenn } \; x_{ni} = x_{nj} \\ 0 & \text{sonst.} \end{cases}$$

Der Smirnoffkoeffizient Im Smirnoffkoeffizienten (Sokal und Sneath, 1963) wird die übereinstimmende Ausprägung einer Variable mit der Wahrscheinlichkeit des gemeinsamen Auftretens dieser Ausprägung gewichtet. Im Unterschied zu den bisher vorgestellten Ähnlichkeitsmaßen wird hier die Übereinstimmung in jeder Ausprägung einer Variable überprüft und bspw. das gemeinsame Fehlen einer Ausprägung als Übereinstimmung gewertet:

$$SC_{ij} = \frac{1}{V} \sum_{n=1}^{N} \sum_{v_n=1}^{V_n} w_{v_n}. \tag{2.16}$$

Dabei bezeichnet SC_{ij} den Smirnoffkoeffoizienten, w_{v_n} die gewichtete Übereinstimmung in Ausprägung v_n des Merkmals x_n, V_n die Anzahl der möglichen Ausprägungen dieses Merkmals und V die Gesamtanzahl der möglichen Ausprägungen über alle Variablen: $V = \sum_{n=1}^{N} V_n$. Bei Nichtübereinstimmung wird $w_{v_n} = -1$ gesetzt. Die Größe des Gewichts bei Übereinstimmung der Objekte ist abhängig von der Häufigkeit, mit der diese Ausprägung in der Stichprobe beobachtet wird. Für alle Ausprägungen v_n einer Variable x_n ergeben sich damit unterschiedliche Gewichtungen. Übereinstimmungen in selten auftretenden Ausprägungen werden stärker gewichtet als solche in häufig auftretenden.

Von Sokal und Sneath (1963) wird kritisch angemerkt, dass die Berücksichtigung der Wahrscheinlichkeit des Auftretens von Gemeinsamkeiten bei der Ähnlichkeitsfeststellung dazu führt, dass die gleichen Objekte in unterschiedlichen Stichproben als unterschiedlich ähnlich angesehen werden.

Eine alternative Lösung zu skalenspezifischen Ähnlichkeitsmaßen stellt die Transformation der nominal skalierten polytomen Merkmale in dichotome Variablen dar. Die Anzahl der neu gebildeten Variablen muss der Anzahl der möglichen Ausprägungen des ursprünglichen Merkmals entsprechen. Das Vorhandensein einer bestimmten Ausprägung der Ursprungsvariable wird dann mit Eins, das Nichtvorhandensein mit Null bezeichnet. Nach dieser Umwandlung wird in der Literatur die Anwendung des oben erläuterten Jaccardkoeffizienten (bzw. jedes Ähnlichkeitskoeffizienten, der eine gemeinsame Null-Ausprägung nicht berücksichtigt) empfohlen (Backhaus et al., 2000, S. 490).

Für die empirische Anwendung scheint der verallgemeinerte Matching-Koeffizient am besten geeignet zu sein, wenn man davon ausgeht, dass jedes Merkmal den gleichen Beitrag zur Erklärung der Ähnlichkeit zwischen zwei Objekten liefert. Die Gewichtung von Merkmalen in Abhängigkeit von der Anzahl möglicher Ausprägungen oder der Häufigkeit ihres Auftretens würde die Bedeutung einzelner Merkmale im Vergleich zu anderen überzeichnen. Dies trifft ebenfalls auf die Bildung von dichotomen Hilfsvariablen zu. Hier resultiert die Verzerrung des Erklärungsanteils aus der „Vervielfältigung“ einer Information durch die Bildung mehrerer Variablen.
Zu berücksichtigen ist bei der Anwendung des Matchingkoeffizienten allerdings die gleiche Gewichtung von übereinstimmenden Ausprägungen Null und Eins.

Ordinal skalierte Merkmale

Ordinal skalierte Merkmale sind solche, deren Ausprägungen sich in eine Rangfolge bringen lassen. Zwei Objekte sind umso ähnlicher, je näher die Ausprägungen der betrachteten Variablen hinsichtlich ihrer Rangordnung beieinander liegen.

In der Literatur finden sich keine spezifischen Ähnlichkeitsmaße für ordinal skalierte Merkmale. Es besteht aber auch für diese Merkmalsgruppe die Möglichkeit der Umwandlung in dichotome Variablen. Die Anzahl der neu gebildeten Variablen entspricht der Anzahl der Ränge der Ursprungsvariablen minus Eins (Steinhausen und Langer, 1977, S. 56). Anders als bei nominalen Merkmalen wird das Vorhandensein jeder Ausprägung nicht separat betrachtet. Die neu gebildeten Variablen werden der Rangfolge der Ausprägungen entsprechend geordnet. Das Vorhandensein einer Ausprägung der Ursprungsvariable wird wieder mit Eins bezeichnet. Im Unterschied zu nominalen Variablen wird allen Variablen mit niedrigerem Rang ebenfalls der Wert Eins zugewiesen. Das Nichtvorhandensein einer ranghöheren Ausprägung wird mit Null bezeichnet. Zur Ähnlichkeitsfeststellung können dann die für dichotome Variablen gebräuchlichen Ähnlichkeitsmaße angewendet werden. Dabei sollte allerdings berücksichtigt werden, dass die Anzahl der gebildeten Hilfsvariablen von der Anzahl der Ausprägungen der Ursprungsvariablen bestimmt wird und sich dadurch eine unterschiedliche Gewichtung der Informationen ergibt.
Um das auszugleichen, könnte jede einzelne neu gebildete Variable mit der Anzahl

der möglichen Ausprägungen der Ursprungsvariable (und damit der Anzahl der aus einem Ursprungsmerkmal neu gebildeten Variablen) gewichtet werden. Diese Gewichtung ließe sich in das skalenspezifische Ähnlichkeitsmaß integrieren.

Metrisch skalierte Merkmale

Quantitative (metrisch skalierte) Merkmale geben über die Rangordnung von Ausprägungen hinaus Auskunft über das Ausmaß der Unterschiede zwischen Objekten (Opitz, 1980, S. 36). Diese Unterschiede werden üblicherweise anhand von Distanzmaßen gemessen. Sie geben i.d.R. die Summe der paarweise ermittelten Differenzen zwischen zwei Objekten in den jeweiligen Ausprägungen an.

City-Block-Metrik und Euklidische Distanz Zwei sehr häufig angewendete Distanzmaße lassen sich mit Hilfe der Minkowski-Metrik zusammenfassen (Backhaus et al., 2000, S. 491):

$$MM_{ij} = \left[\sum_{n=1}^{N} (x_{ni} - x_{nj})^a \right]^{\frac{1}{a}} . \tag{2.17}$$

Dabei bezeichnet MM_{ij} die Minkowski-Metrik, $(x_{ni} - x_{nj})$ die Differenz zwischen Objekt i und Objekt j im Merkmal x_n und a eine positive Konstante, deren Wert über das verwendete Distanzmaß entscheidet. Das zu $a = 1$ gehörige Maß ist die City-Block-Metrik, aus $a = 2$ resultiert die Euklidische Distanz. Mit der Quadrierung der Differenzen in der Euklidischen Distanz werden die großen Abweichungen stärker gewichtet als die kleinen.

Wenn die beobachteten Merkmale unterschiedliche Maßeinheiten aufweisen, ist für die Anwendung beider Distanzmaße die Normierung der Merkmale auf eine gemeinsame Maßeinheit nötig. Beide Distanzmaße sind nicht skaleninvariant, d.h. die Größe der jeweiligen Distanz hängt von der Maßeinheit des beobachteten Merkmals ab (Kaufmann und Pape, 1996, S. 448f.). Ein gängiges Standardisierungsverfahren ist die z-Transformation, bei der Variablen so umgewandelt werden, dass sie einen Mittelwert von Null und eine Standardabweichung von Eins haben (Brosius, 1998, S. 696f.). Dadurch wird allerdings auch die unterschiedliche Streuung der Variablen

beseitigt. Wenn diese Information erhalten bleiben soll, ist bspw. die Normierung der Ausgangsinformationen durch Division der betrachteten Ausprägung durch die jeweils größte beobachtete Ausprägung der Variable vorzuziehen (Cain und Harrison, 1958, S. 91). Das Ergebnis dieser Normierungen sind Distanzen im Bereich zwischen Null und Eins – ohne Maßeinheiten. Die unterschiedliche Varianz der Merkmale und Korrelationen zwischen Merkmalen werden mit beiden Distanzmaßen nicht berücksichtigt.

Die Mahalanobisdistanz Sind diese Informationen für die Feststellung der Ähnlichkeit von Bedeutung, muss die Mahalanobisdistanz verwendet werden. Hier wird nicht nur der Abstand zwischen zwei Objekten in einem Merkmal ermittelt, sondern auch die Varianz jedes Merkmals und evtl. Korrelationen mit anderen Merkmalen berücksichtigt (Opitz, 1980, S. 52). Je größer die Gesamtvarianz eines Merkmals ist, desto geringer wird die entsprechende Differenz $(x_{ni} - x_{nj})$ für die Gesamtdistanz zwischen den Objekten i und j gewichtet. Sind zwei Merkmale hoch korreliert (liefern also beide annähernd die gleiche Information), wird der gemeinsame Erklärungsbeitrag an der Distanz entsprechend niedrig gewichtet.
Das Distanzmaß wird von Mahalanobis (1936) zur Feststellung der Distanz zwischen zwei Stichproben mit normalverteilten Merkmalen hergeleitet. Die Distanz wird ermittelt aus der Summe der Differenzen der einzelnen Merkmale, gewichtet mit der inversen Varianz-Kovarianz-Matrix:

$$MD_{ij} = \left[(\mathbf{x_i}) - (\mathbf{x_j})\right]' \mathbf{Cov}^{-1} \left[(\mathbf{x_i}) - (\mathbf{x_j})\right]. \tag{2.18}$$

Dabei bezeichnen $\mathbf{x_i}$ und $\mathbf{x_j}$ die Merkmalsvektoren der Objekte i und j und $\mathbf{Cov}$ die Varianz-Kovarianz-Matrix: $\mathbf{Cov} = \frac{1}{I+J-1} \sum (\mathbf{x_i} - \bar{\mathbf{x}}_\mathbf{i}) (\mathbf{x_j} - \bar{\mathbf{x}}_\mathbf{j})$.

Die Biaskorrektur Aus der Mahalanobisdistanz leiten Abadie und Imbens (2002) ein bias-korrigiertes Distanzmaß ab. Anstelle der Varianz-Kovarianz-Matrix wird die Varianzmatrix $diag(\mathbf{Cov})$ in der Abadie-Imbens-Metrik AI_{ij} verwendet:

$$AI_{ij} = \left[(\mathbf{x_i}) - (\mathbf{x_j})\right]' diag(\mathbf{Cov})^{-1} \left[(\mathbf{x_i}) - (\mathbf{x_j})\right]. \tag{2.19}$$

Die einzelnen Merkmale werden nicht mehr unter Berücksichtigung ihrer Korrelation zu den anderen Merkmalen gewichtet, sondern nur in Abhängigkeit ihrer eigenen Streuung. Variablen mit einer geringen Streuung erhalten auch hier ein größeres Gewicht.

Unterschiedlich skalierte Merkmale

Die einfachste Möglichkeit der gemeinsamen Betrachtung verschieden skalierter Variablen besteht in der Transformation aller betrachteten Merkmale auf ein gemeinsames Skalenniveau. Dabei ist es ohne weitere Annahmen möglich, Variablen von einem höheren auf ein niedrigeres Skalenniveau zu transformieren. Mit einer solchen Niveauregression sind allerdings immer Informationsverluste verbunden (Kaufmann und Pape, 1996, S. 452). Eine umgekehrte Transformation (Niveauprogression) ist eher problematisch, da den „hochtransformierten“ Variablen mehr Aussagegehalt unterstellt wird als die ursprünglichen Merkmale besitzen.

Alternativ können unterschiedlich skalierte Merkmale in einem gemeinsamen Distanz- oder Ähnlichkeitsmaß zusammengefasst werden.

Anforderungen an aggregierte Ähnlichkeits- und Distanzmaße Für die Interpretierbarkeit eines gemeinsamen Koeffizienten ist die Transformation der Distanzmaße in Ähnlichkeitsmaße (oder umgekehrt) nötig.[17] Am häufigsten ist in der Literatur die folgende Transformation für normierte Maße zu finden: $s_{n,ij} = 1 - d_{n,ij}$. Dabei steht $d_{n,ij}$ für die Distanz und $s_{n,ij}$ für die Ähnlichkeit zwischen zwei Objekten i und j im Merkmal x_n.

Um einen adäquaten Beitrag jeder Einzeldistanz zur Gesamtdistanz zu gewährleisten, müssen darüber hinaus evtl. unterschiedliche Schwankungsbreiten und verschiedene Skalenniveaus der einzelnen Distanzen berücksichtigt werden. Um eine Über- oder Untergewichtung einer der Merkmalsgruppen zu verhindern, werden die Distanzen auf den Bereich zwischen Null und Eins normiert. Eine verbreitete Normierungsmethode ist die Division der Einzeldistanzen durch die Schwankungsbreite

[17] Eine Übersicht über mögliche Transformationen findet sich in Steinhausen und Langer (1977).

(maximale Distanz) der entsprechenden Variable (Diday und Simon, 1976, S. 52). Allerdings ist mit dieser Normierung der Einfluss von Ausreißern noch nicht berücksichtigt. Hierfür kann die Schwankungsbreite „getrimmt“ – d.h. die Ränder abgeschnitten – oder an ihrer Stelle die Standardabweichung einer Variable verwendet werden (Wilson und Martinez, 1997, S. 4).

Die gewichtete Mahalanobis-Matching-Distanz Das in Kaufmann und Pape (1996) vorgestellte aggregierte Distanzmaß ist eine Kombination skalenspezifischer Distanzmaße, die jeweils mit der Anzahl der Variablen des entsprechenden Skalenniveaus gewichtet werden.[18]

Für die gemeinsame Distanzmessung metrischer und nominaler Variablen können in diesem Rahmen die Mahalanobisdistanz und der verallgemeinerte Matchingkoeffizient kombiniert werden. Die erforderliche Normierung der metrischen Variablen erfolgt mit Hilfe der maximalen Distanz. Die Ähnlichkeiten der nominalen (dichotomen und polytomen) Variablen werden in Distanzen umgewandelt. Beide Distanzmaße werden mit der Anzahl der beobachteten Variablen mit dem entsprechenden Skalenniveau gewichtet. Die Matching-Mahalanobis-Distanz ergibt sich wie folgt:

$$MDMC_{ij} = \frac{1}{N}\left[me \cdot MD_{ij} + no \cdot (1 - gMC_{ij})\right]. \tag{2.20}$$

Mit $MDMC_{ij}$ und MD_{ij} bzw. gMC_{ij} werden die Mahalanobis-Matching-Distanz und die skalenspezifischen Distanzmaße bezeichnet, N gibt die Anzahl der Variablen an, die sich aus der Anzahl der metrischen Variablen, me, und der nominalen Variablen, no, zusammensetzt: $N = me + no$.

Die folgenden Beispiele alternativer Aggregationen zeigen, dass die Zusammenfassung von Variablen mit unterschiedlichen Skalenniveaus auch in anderen Wissenschaftsbereichen von Bedeutung ist.

Die verallgemeinerte Minkowski-Metrik Von Ichino und Yaguchi (1994) wird eine Zusammenfassung unterschiedlich skalierter Merkmale für die Anwendung in

[18] Diese Art der Zusammenfassung verschiedener Merkmale wird in Opitz (1980) S. 59 als linearhomogene Aggregation bezeichnet. Sie erfüllt die dort genannten Bedingungen für aggregierte Distanzindizes.

der statistischen Mustererkennung entwickelt. Die Grundannahme besteht darin, dass sich Objekte mit allen ihren Eigenschaften in einem Kartesischen Raummodell darstellen lassen. Die Objekte werden charakterisiert durch N beobachtete Merkmale. Objekt i wird bspw. beschrieben durch die Produktmenge: $i = i_1 \times i_2 \times ... \times i_N$, Objekt j durch $j = j_1 \times j_2 \times ... \times j_N$. Der Merkmalsraum wird entsprechend bezeichnet mit: $U^{(N)} = U_1 \times U_2 \times ... \times U_N$. Dabei ist – im Unterschied zu den bisher betrachteten Distanzmaßen – nicht nur die Betrachtung von Skalaren möglich, sondern auch von Merkmalsintervallen.[19] Die Definition des Merkmalsraumes für ein Merkmal wird bestimmt durch seine Skalierung. Für metrisch und ordinal skalierte Variablen bildet das geschlossene Intervall $U_n = [v_{min_n}, v_{max_n}]$ den Merkmalsraum, wobei v_{min_n} den kleinsten und v_{max_n} den größten möglichen Wert der Variable x_n bezeichnet. Für eine nominal skalierte Variable bildet die Menge aller möglichen Ausprägungen den Merkmalsraum: $U_n = v_{min_n}, v_{1_n}, v_{2_n}, ..., v_{max_n}$.

Für die Distanzmessung werden zwei Größen berücksichtigt: die Vereinigungsmenge zweier Objekte sowie deren Schnittmenge. Die Vereinigungsmenge der Objekte i und j in Bezug auf metrisch und ordinal skalierte Variablen bildet das geschlossene Intervall $v_{ni} \oplus v_{nj} = [\min(v_{min_n,i}, v_{min_n,j}), \max(v_{max_n,i}, v_{max_n,j})] \ \forall \ x_n = 1, 2, ..., N$. Dabei bezeichnet $\oplus$ den kartesischen Verbindungsoperator (Cartesian Joint Operator). Für nominale Variablen bildet die Menge der bei mindestens einem der beiden Objekte beobachteten Ausprägungen der betrachteten Variable den gemeinsamen Raum: $v_{n,i} \oplus v_{n,j} = v_{n,i} \cup v_{n,j}$. Die Schnittmenge (Cartesian Meet) beider Objekte für eine Variable x_n beinhaltet nur die gemeinsamen Ausprägungen dieser Variable: $v_{n,i} \otimes v_{n,j} = v_{n,i} \cap v_{n,j}$. Dabei bezeichnet $\otimes$ den kartesischen Schnittoperator (Cartesian Meet Operator).

Beide Größen werden wie folgt zusammengefasst, um die Distanz zwischen den Objekten i und j im Merkmal x_n zu ermitteln:

$$d_{n,ij} = [v_{n,i} \oplus v_{n,j}] - [v_{n,i} \otimes v_{n,j}] + \lambda \left(2 [v_{n,i} \otimes v_{n,j}] - [v_{n,i}] - [v_{n,j}]\right).$$

[19] Diese Erweiterung ist für die Betrachtung beobachteter Merkmale im Rahmen einer Evaluation nicht von Interesse. Daraus erklärt sich aber die Komplexität des beschriebenen Vorgehens.

Dabei bezeichnet [...] jeweils die Länge des Intervalls (für metrisch und ordinal skalierte Variablen) bzw. die Menge aller Ausprägungen (für nominal skalierte Variablen). Mit dem Faktor λ kann die Gewichtung der Übereinstimmungen gegenüber dem gemeinsamen Merkmalsraum beeinflusst werden. Je größer dieser Faktor ist, desto stärker wird die Schnittmenge beider Objekte gewichtet. Es gilt $0 \leq \lambda \leq 0,5$.

Diese Gleichung vereinfacht sich, wenn für ein Merkmal x_n für beide Objekte jeweils nur eine Ausprägung beobachtet wird. Dann gilt: $[v_{n,i}] = [v_{n,j}] = 0$. Bei Übereinstimmung der beobachteten Merkmalsausprägungen entfällt der erste Term der Gleichung: $[v_{n,i} \oplus v_{n,j}] = 0$, bei Nichtübereinstimmung der zweite: $[v_{n,i} \otimes v_{n,j}] = 0$. Bei Übereinstimmung der beobachteten Ausprägung ergibt sich folgende Distanz:

$$d_{n,ij} = -\left[v_{n,i} \otimes v_{n,j}\right] + 2\lambda\left[v_{n,i} \otimes v_{n,j}\right],$$

bei Nichtübereinstimmung gilt:

$$d_{n,ij} = \left[v_{n,i} \oplus v_{n,j}\right].$$

Die für jede einzelne betrachtete Variable gebildeten Distanzen lassen sich zusammenfassen. Um eine ungleichmäßige Gewichtung der einzelnen Variablen durch unterschiedliche Maßeinheiten zu vermeiden und die Dimension der beobachteten Größen zu vereinheitlichen, schlagen Ichino und Yaguchi (1994) folgende Normierung vor:

$$d_{n,ij}^{norm} = \frac{d_{n,ij}}{[U_n]},$$

wobei $[U_n]$ die Größe des Merkmalsraums der Variable x_n (das Intervall bzw. die Menge aller möglichen Werte) bezeichnet. Damit wird $d_{n,ij}^{norm}$ zu einer dimensionslosen Größe, für die gilt: $0 \leq d_{n,ij}^{norm} \leq 1$.
Die dimensionslosen Größen lassen sich in einer verallgemeinerten Minkowski-Metrik zusammenfassen:

$$gMM_{ij} = \left[\sum_{n=1}^{N} (d_{n,ij}^{norm})^a\right]^{\frac{1}{a}}. \tag{2.21}$$

Durch die Wahl des Exponenten a wird das angewendete Distanzmaß festgelegt.[20]

[20] Eine sehr ähnliche Art der Zusammenfassung verschieden skalierter Merkmale findet sich u.a. bei Gowda und Diday (1992).

Die heterogene Wertdifferenz In neuronalen Netzwerken zur Klassifikation werden Distanzfunktionen zum fallbasierten Lernen eingesetzt. In diesem Zusammenhang stellen Wilson und Martinez (1997) eine Distanzfunktion für die Verarbeitung von Informationen über metrische und nominale Variablen vor. Diese heterogene Wertdifferenzmetrik zweier Objekte wird wie folgt definiert:

$$HD_{ij} = \sqrt{\sum_{n=1}^{N} d_{n,ij}^2}. \tag{2.22}$$

Dabei bezeichnet HD_{ij} die heterogene Wertdifferenzmetrik, N die Anzahl der betrachteten Merkmale und $d_{n,ij}$ die skalenspezifische Distanz zwischen den betrachteten Objekten i und j.

Die Distanzmessung für nominale Variablen erfolgt in Wilson und Martinez (1997) mit Hilfe von Klassen, in die Objekte mit den beobachteten Ausprägungen eingeordnet werden. Diese Art der Distanzfeststellung ist im Zusammenhang mit der Evaluation nicht möglich. Es kann statt dessen eins der vorgestellten Ähnlichkeitsmaße für nominale Variablen, z.B. der verallgemeinerte Matchingkoeffizient, verwendet werden.[21] Die Distanz der metrischen Variablen wird mit Hilfe der normierten absoluten Differenz zwischen beiden Ausprägungen ermittelt. Zur Normierung wird die Standardabweichung verwendet.

Die einzelnen Wertdifferenzen ergeben sich dann wie folgt:[22]

$$d_{n,ij} = \begin{cases} \frac{|x_{ni} - x_{nj}|}{4\sigma_n} & \text{wenn } n \text{ metrisch} \\ 1 - gMC_{n,ij} & \text{wenn } n \text{ nominal.} \end{cases}$$

Dabei bezeichnet $|x_{ni} - x_{nj}|$ die absolute Differenz der Ausprägung des metrisch skalierten Merkmals x_n zwischen Objekt i und Objekt j, σ_n die Standardabweichung dieses Merkmals und $gMC_{n,ij}$ den verallgemeinerten Matchingkoeffizienten.

[21] Eine gemeinsame Ausprägung im Merkmal x_n wird mit Eins bewertet, eine Nichtübereinstimmung mit Null.

[22] Da für alle untersuchten Objekte die Ausprägungen der betrachteten Variablen bekannt sind, fällt der in Wilson und Martinez (1997) S. 8 genannte erste Fall weg.

Der aggregierte Ähnlichkeitskoeffizient von Gower Eine ähnliche Art der Aggregation verschieden skalierter Merkmale stellt die Bildung des gewichteten Durchschnitts aus skalenspezifisch ermittelten Ähnlichkeitsmaßen dar. Sie wird in Gower (1971) vorgestellt:

$$SG_{ij} = \frac{\sum_{n=1}^{N} w_n s_{n,ij}}{\sum_{n=1}^{N} pc_{n,ij}}. \tag{2.23}$$

Dabei bezeichnet SG_{ij} den Ähnlichkeitskoeffizienten von Gower, $s_{n,ij}$ die merkmalsspezifisch ermittelte Ähnlichkeit zwischen den Objekten i und j im Merkmal x_n, w_n einen merkmalsspezifischen Gewichtungsfaktor und $pc_{n,ij}$ die Anzahl der Variablen, für die für beide Objekte eine Ausprägung beobachtet wurde. Wenn der Vergleich aller Objekte in allen betrachteten Variablen möglich ist, gilt: $\sum_{n=1}^{N} pc_{n,ij} = N$.
Wie die Ähnlichkeit in einem Merkmal ermittelt wird, hängt von seiner Skalierung ab. Für die Ermittlung der „Einzelähnlichkeiten“ dichotomer Variablen werden – wie beim Jaccard-Koeffizienten – nur gemeinsame Eins-Ausprägungen berücksichtigt. Für die Feststellung der Ähnlichkeit in nominal skalierten Variablen wird – der Idee des verallgemeinerten Matchingkoeffizienten folgend – eine gemeinsame Ausprägung mit Eins bewertet, jede Nichtübereinstimmung mit Null. Für metrisch skalierte Variablen wird die absolute Differenz der beobachteten Ausprägungen mit der maximalen Differenz $diff_{max_n}$ normiert und in ein Ähnlichkeitsmaß transformiert: $s_{n,ij} = 1 - \frac{|x_{ni} - x_{nj}|}{diff_{max_n}}$.

Balancing Scores Eine Alternative zu den bisher vorgestellten statistischen Distanzmaßen stellt die Bildung sog. Balancing Scores dar. Unter diesem Begriff werden alle Funktionen zusammengefasst, für die gilt:

$$X \perp D | BS(X)\,. \tag{2.24}$$

Die bedingte Verteilung der Merkmale, gegeben die Funktion $BS(X)$, ist in der Teilnehmergruppe und der Kontrollgruppe gleich. In Rosenbaum und Rubin (1983) wird bewiesen, dass in großen Stichproben auch die Verwendung eines Balancing Scores die Zuordnung von Personen zur Teilnehmer- oder Kontrollgruppe unabhängig vom potenziell erzielten Einkommen erlaubt, wenn der Score aus Merkmalen gebildet

wird, die die Annahme bedingter Unabhängigkeit (2.4) erfüllen. Kurz gesagt: Auch Balancing Scores erfüllen diese Annahme asymptotisch:

$$Y_t^T, Y_t^C \perp D | BS(X). \tag{2.25}$$

Der gröbste – und in empirischen Studien am häufigsten verwendete – Balancing Score ist ein eindimensionales Maß, der Propensity Score. Er wird definiert als: $PS(X) = \Pr(D = 1|X)$, die Wahrscheinlichkeit der Teilnahme an einer Fördermaßnahme. Wenn die Teilnahmewahrscheinlichkeit nicht beobachtbar ist, muss der Propensity Score auf Grundlage der relevanten Merkmale geschätzt werden. Für die Schätzung wird i.d.R. ein Probitmodell verwendet. Dabei wird angenommen, dass die beobachtete Teilnahme oder Nichtteilnahme Ausdruck einer nicht beobachtbaren (latenten) Variable ist (Greene, 2003, S. 668f.). Diese Variable kann mit Hilfe der sog. Indexfunktion dargestellt werden:

$$IN_i = \beta X_i + \varepsilon_i \quad \text{mit} \quad \varepsilon_i \sim N(0, \sigma^2). \tag{2.26}$$

Dabei bezeichnen IN_i die latente Variable, X_i die individuellen Merkmale und β ihren Einfluss auf diese latente Variable. Der beobachtbare Modellteil, die Teilnahmeentscheidung, wird beschrieben durch:

$$D_i = \begin{cases} 1 & \text{wenn } IN_i > 0 \\ 0 & \text{sonst.} \end{cases}$$

Der Propensity Score lässt sich in diesem Modellrahmen unter Verwendung der Verteilungsfunktion der Standardnormalverteilung Φ schätzen:

$$\widehat{PS}(X_i) = \Phi\left(\hat{\beta} X_i\right). \tag{2.27}$$

Zur Spezifikation der Schätzgleichung für den Propensity Score wird von Dehejia und Wahba (1999) ein iteratives Verfahren eingesetzt. Für verschiedene Spezifikationen werden die beiden Teilstichproben in Untergruppen entsprechend der Größe des Propensity Scores unterteilt. Für jede Untergruppe wird anschließend auf signifikante Differenzen in den Mittelwerten und Verteilungen der einzelnen Variablen

zwischen Teilnehmern und Nichtteilnehmern getestet. Sind solche Unterschiede vorhanden, wird die Spezifikation des Modells verändert und Terme höherer Ordnung bzw. Interaktionsterme einbezogen.[23]

Aus der Propensity-Score-Schätzung wird ein anderes Ähnlichkeitsmaß, die Partizipationsneigung, abgeleitet (Christensen, 2001):

$$\begin{aligned} \widehat{PN}_i &= \hat{\beta} X_i \\ &= \widehat{IN}_i. \end{aligned} \tag{2.28}$$

Für den linearen Schätzer der Indexfunktion gelten die oben genannten Annahmen. Er hat den Vorteil, dass die Unterschiede zwischen Teilnehmer- und Kontrollgruppe an den Rändern der Verteilung deutlicher werden (Lechner, 1998, S. 115) und damit die Zuordnung von Personen, deren Propensity Scores nahe bei Null oder Eins liegen, exakter möglich ist.

In der empirischen Evaluationsliteratur werden beide Scores häufig angewendet. Allerdings ist zu berücksichtigen, dass der „wahre“ Propensity Score (also die tatsächliche Teilnahmewahrscheinlichkeit) in den wenigsten Fällen bekannt ist. Muss diese Größe geschätzt werden, gilt die Einhaltung der Grundannahme nur asymptotisch. Gerade in kleinen Stichproben kann der Informationsverlust, der mit der Verwendung einer geschätzten Größe anstelle der zugrunde liegenden Merkmale selbst verbunden ist, Auswirkungen auf die Qualität des Matchingergebnisses haben. So ist nicht auszuschließen, dass einige Merkmale eine andere Bedeutung für die Teilnahmeentscheidung als für die zukünftige Situation einer Person haben, sodass Personen mit identischem Propensity Score unterschiedliche Aussichten auf dem Arbeitsmarkt haben. Diesem Problem kann begegnet werden, indem – zusätzlich oder alternativ zum Propensity Score – wichtige Determinanten der zukünftigen Situation bei der Feststellung der Ähnlichkeit von Personen berücksichtigt werden.[24]

[23] In der jüngeren empirischen Literatur wird dieses Verfahren häufig angewendet, bspw. in Becker und Ichino (2002).

[24] Dieses Problem wird in Fröhlich (2004a) ausführlich erläutert. Der aus der zusätzlichen Berücksichtigung persönlicher Merkmale resultierende Balancing Score wird hier als Augmented Propensity Score bezeichnet.

In der Literatur finden sich verschiedene Kombinationen aus den oben vorgestellten statistischen Distanzmaßen und dem Propensity Score.
In Qian (2004) wird der Propensity Score anhand der beobachteten qualitativen Merkmalen ermittelt und als zusätzliche Variable in die Mahalanobisdistanz integriert. Einem Vorschlag von Rosenbaum und Rubin (1985) folgend, wird von Lechner (1998) ein zweistufiger Prozess angewendet, in dem der Propensity Score als Vorauswahlkriterium für mögliche Kontrollgruppenmitglieder dient, aus denen mit Hilfe der Mahalanobisdistanz dann der ähnlichste Partner ausgewählt wird. In Zhao (2004, 2006) dienen die Koeffizienten der Propensity-Score-Schätzung zur Gewichtung Differenzen der jeweiligen Merkmale.

2.3.2 Zuordnungsprozesse

Anhand der vorgestellten Distanzmaße lassen sich mögliche Partner für jeden betrachteten Teilnehmer identifizieren und in einer eigenen Unter-Kontrollgruppe zusammenfassen. Die Unter-Kontrollgruppe eines Teilnehmers C_i setzt sich aus denjenigen Nichtteilnehmern zusammen, deren Charakteristika X_j denen des Teilnehmers X_i ähnlich sind. Mit $C(X_i)$ wird dabei die Menge aller Merkmale bezeichnet, die denen des Teilnehmers i ähnlich sind:

$$C_i = \{j | X_j \in C\left(X_i\right)\} \quad \forall j \in \{D = 0\}\,. \tag{2.29}$$

Die Auswahlprozesse für die Personen in den einzelnen Unter-Kontrollgruppen lassen sich danach unterscheiden, ob genau ein Nichtteilnehmer ausgewählt wird, oder ob mehrere Nichtteilnehmer in einer Unter-Kontrollgruppe vertreten sind. Die Wahl eines Prozesses bewegt sich im Spannungsfeld zwischen Verzerrung und Varianz des Schätzers: Mit zunehmender Anzahl der berücksichtigten Nichtteilnehmer verringert sich die Varianz des Schätzers, aber die Gefahr der Verzerrung nimmt zu (weil auch solche Nichtteilnehmer berücksichtigt werden, die dem Teilnehmer weniger ähnlich sind).[25]

[25] In Caliendo und Kopeinig (2005) findet sich eine Übersicht über den Zusammenhang zwischen Zuordnungsprozessen und diesen beiden Größen.

Nearest Neighbor Matching

Auswahlprozesse, in denen genau ein Nichtteilnehmer die Unter-Kontrollgruppe bildet, werden als 1:1-Matching oder Nearest Neighbor Matching bezeichnet. Für das Nearest Neighbor Matching ergibt sich folgende Definition der Ähnlichkeit (2.29) zwischen den Merkmalen eines Teilnehmers und denen eines Nichtteilnehmers:

$$C\left(X_i\right) = \left\{ j | \underset{j \in NT}{Min} \ \langle X_i - X_j \rangle \right\}. \tag{2.30}$$

Dabei bezeichnet $\langle X_i - X_j \rangle$ die Distanz zwischen einem Teilnehmer und einem Nichtteilnehmer, die anhand der beobachteten Merkmale ermittelt wird. Es wird derjenige Nichtteilnehmer als ähnlichster bezeichnet, dessen Merkmale den geringsten Abstand zu den Teilnehmermerkmalen aufweisen.[26] Für die Gewichtung des Nichtteilnehmereinkommens (2.10) gilt dann:

$$W\left(i,j\right) = \begin{cases} 1 & \text{wenn } \ j \in C\left(X_i\right) \\ 0 & \text{sonst.} \end{cases} \tag{2.31}$$

Caliper Matching Um zu verhindern, dass ein Teilnehmer mit einem unähnlichen Nichtteilnehmer verglichen wird, wenn es keine ähnlichen Nichtteilnehmer gibt, kann ein zulässiger Maximalabstand zwischen den Merkmalen festgelegt werden. Diese Form wird als Caliper Matching bezeichnet (Cochran und Rubin, 1973, S. 420f.). Die Menge aller Nichtteilnehmer mit ähnlichen Merkmalen (2.29) wird dann wie folgt definiert:

$$C\left(X_i\right) = \begin{cases} \left\{ j | \underset{j \in NT}{Min} \ \langle X_i - X_j \rangle \right\} & \text{wenn } \ \langle X_i - X_j \rangle < \psi \\ \emptyset & \text{sonst.} \end{cases} \tag{2.32}$$

Es wird nur unter denjenigen Nichtteilnehmern der ähnlichste ausgewählt, deren Merkmale nur innerhalb der vorgegebenen Toleranz ψ von denen des Teilnehmers abweichen. Existiert kein solcher Nichtteilnehmer, kann für den betrachteten Teil-

[26] Wenn mehrere Nichtteilnehmer einem Teilnehmer gleich ähnlich sind, wird unter ihnen ein Nichtteilnehmer ausgewählt.

nehmer kein Vergleichseinkommen ermittelt werden. Er kann bei der Ermittlung des Maßnahmeeffekts nicht berücksichtigt werden. Allerdings ist a priori schwer einzuschätzen, welcher maximal zulässige Abstand sinnvoll ist (Caliendo und Kopeinig, 2005, S. 10).

Die Zuordnungsprozesse, mit denen den Teilnehmern ihre Partner zugewiesen werden, lassen sich einteilen in Zuordnungsprozesse mit Zurücklegen und solche nach dem Prinzip Ziehen ohne Zurücklegen.

Zuordnung mit Zurücklegen Bei der Zuordnung mit Zurücklegen wird jedem Teilnehmer der ähnlichste Nichtteilnehmer zugewiesen, unabhängig davon, für wie viele Teilnehmer ein Nichtteilnehmer als Partner verwendet wird. Dabei besteht die Gefahr, dass nur wenige Nichtteilnehmer zur Bildung der Kontrollgruppe benutzt werden, auch wenn noch andere – sehr ähnliche – Nichtteilnehmer zur Verfügung stehen.[27] Die Anwendung dieser Zuordnung empfiehlt sich für Matchingprozesse, bei denen Teilnehmer- und Nichteilnehmergruppe (annähernd) gleich groß sind, da damit am ehesten gewährleistet ist, dass jeder Teilnehmer zur Feststellung des Maßnahmeeffekts berücksichtigt wird, wenn er ein Pendant in der Nichtteilnehmergruppe hat.[28]

Zuordnung ohne Zurücklegen Wenn die Nichtteilnehmergruppe deutlich größer als die Teilnehmergruppe ist, werden Zuordnungsprozesse ohne Zurücklegen angewendet. Bei dieser auch als Pair Matching bezeichneten Art der Gruppenbildung wird jedem Teilnehmer genau ein Nichtteilnehmer zugeordnet, wobei ein Nichtteilnehmer nicht als Partner für mehrere Teilnehmer eingesetzt werden kann.

Ein häufig eingesetztes Verfahren, das als **Random Matching** (Dehejia und Wahba, 2002, S. 154) oder **Greedy Pair Matching** (Augurzky, 2000a, S. 4) bezeichnet wird, legt nach dem Zufallsprinzip eine Reihenfolge der Teilnehmer fest. In dieser Reihenfolge wird dann jedem Teilnehmer ein passender Nichtteilnehmer zugeordnet.

[27] Gerfin und Lechner (2002) weisen darauf hin, dass damit eine erhebliche Steigerung der Varianz der Schätzung verbunden sein kann.

[28] Gleich große Teilnehmer- und Nichtteilnehmergruppen treten vor allem bei der Evaluation mehrerer Maßnahmen auf, wo die Teilnehmer der einen Maßnahme gleichzeitig Nichtteilnehmer für eine andere Maßnahme sind.

Einmal zugeordnete Nichtteilnehmer werden nicht noch einmal verwendet. Bei diesem Verfahren kann nicht verhindert werden, dass den Teilnehmern am Ende des Zuordnungsprozesses möglicherweise sehr unähnliche Nichtteilnehmer zugeordnet werden. Es ist auch möglich, dass keine passenden Partner mehr gefunden werden, obwohl es ähnliche Nichtteilnehmer gibt.[29]

Diejenigen Teilnehmer, für die kein passender Partner gefunden wird, müssen aus der weiteren Analyse ausgeschlossen werden.

Aufbauend auf dem Random Matching kann eine **iterative Annäherung an die bestmögliche Zuordnung** durchgeführt werden.[30] In der Abbildung 2.2 wird der Zuordnungsprozess verdeutlicht.

In diesem Prozess wird für alle Pärchen der festgelegten Anfangszuordnung die individuelle Distanz ermittelt und in der Summe der quadrierten Distanzen zusammengefasst. Die Minimierung dieser Summe ist Ziel dieses Algorithmus', der nach folgenden Regeln abläuft: Es wird ein Teilnehmer ermittelt, der neben dem aktuellen noch weitere mögliche Partner hat. Unter diesen möglichen Partnern wird einer zufällig ausgewählt. Für den ausgewählten Nichtteilnehmer werden zwei Fälle unerschieden: Er ist entweder noch keinem anderen Teilnehmer zugewiesen. Dann wird er gegen den ursprünglich zugeordneten Nichtteilnehmer ausgetauscht, wenn sich dadurch die Summe der quadrierten Distanzen verringert. Oder der ausgewählte Nichtteilnehmer ist bereits einem anderen Teilnehmer zugeordnet. Dann wird zusätzlich überprüft, ob dieser Teilnehmer noch weitere mögliche Partner hat. Wenn weitere Partner vorhanden sind, wird nach dem oben beschriebenen Muster ausgetauscht. Dieser Prozess wird so oft wiederholt, bis mit einer vorher festgelegten Anzahl von Durchläufen keine Verringerung der Distanzsumme mehr erreicht werden kann.[31]

[29] Diese Problematik wird in Rosenbaum (1989) anhand von Beispielen erläutert.

[30] Als alternative Anfangszuordnung können auch Pärchen aus jeweils einem Teilnehmer und einem zufällig zugeordneten Nichtteilnehmer gebildet werden.

[31] Dieses Verfahren wird in Reinowski, Schultz und Wiemers (2003) zur Evaluation arbeitsmarktpolitischer Maßnahmen für Langzeitarbeitslose angewendet.

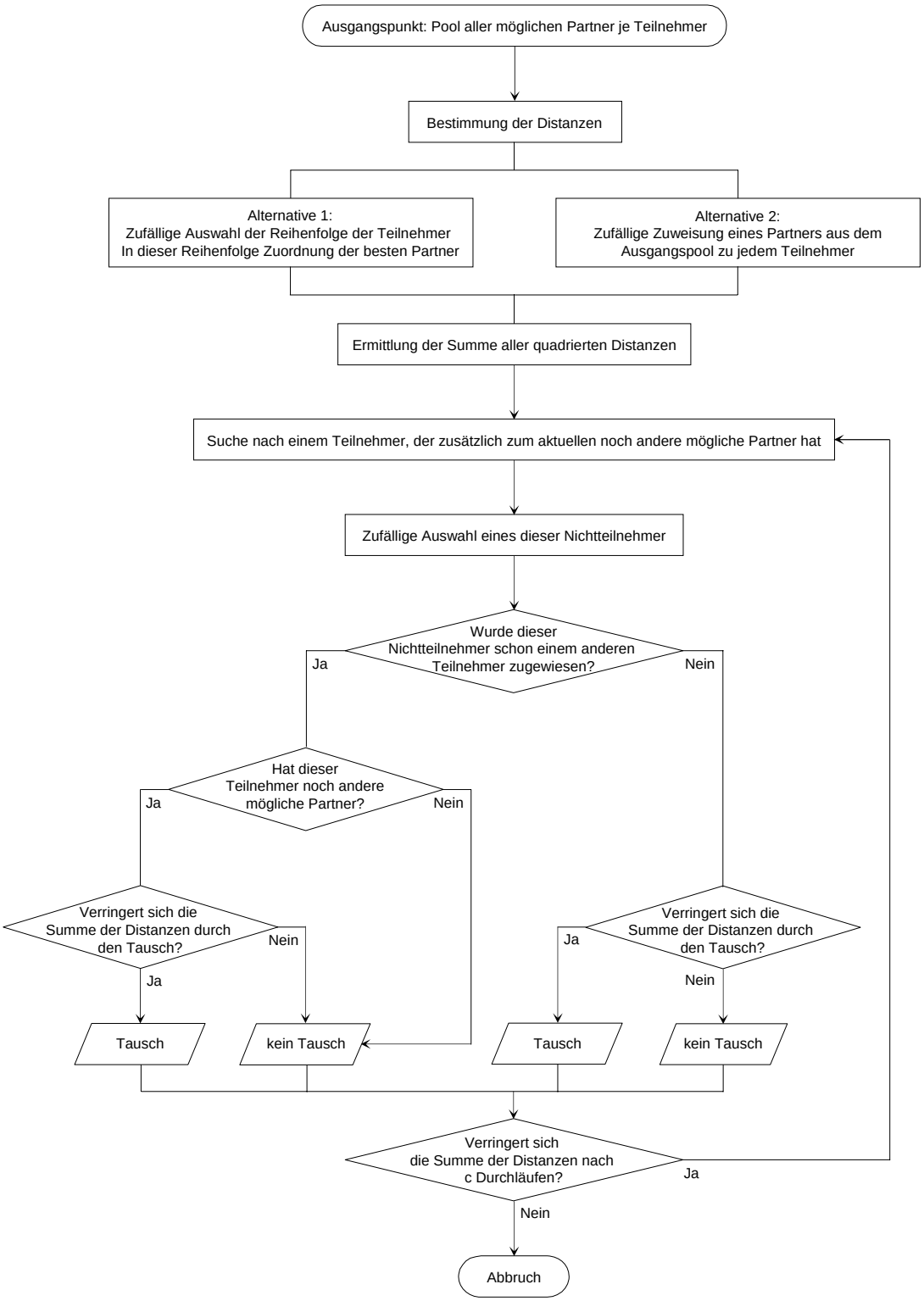

Abbildung 2.2: *Ablaufschema des iterativen Verfahrens zur Zuordnung möglichst ähnlicher Nichtteilnehmer zu den untersuchten Teilnehmern*

Quelle: Reinowski, Schultz und Wiemers (2003) S. 18.

Wenn verhindert werden soll, dass mehr Teilnehmer als nötig aus der Analyse ausgeschlossen werden und für alle Teilnehmer möglichst ähnliche Partner gefunden werden sollen, können Methoden aus der linearen Optimierung bzw. der Graphentheorie eingesetzt werden. Solche **optimalen Matchingalgorithmen** liefern eine vollständige Zuordnung von Nichtteilnehmern zu Teilnehmern (wenn diese möglich ist), bei der die Gesamtsumme der Abstände zwischen Teilnehmern und Nichtteilnehmern minimal ist. Der Vorteil optimaler Zuordnungen gegenüber Greedy Matching wird in Rosenbaum (1989) anhand eines Beispiels erläutert: optimale Algorithmen finden die – hinsichtlich des definierten Kriteriums – beste Lösung, während bei Greedy Matching nicht klar ist, ob ein besseres als das gefundende Ergebnis existiert.

Ein Beispiel für optimale Zuordnungsprozesse ist der **Ungarische Algorithmus** zur Lösung des klassischen Zuordnungsproblems.[32] Von Kuhn (1955) wird, basierend auf den Arbeiten von König (1916) und Egervary (1931), ein matrixbasiertes Lösungsverfahren vorgeschlagen.[33] Das Ziel dieses Lösungsverfahrens besteht in der kostenminimalen vollständigen Zuordnung von Personen zu Arbeitsplätzen. Diese Idee kann auf die Zuordnung möglichst ähnlicher Personen übertragen werden. Die Kosten werden dabei ersetzt durch die individuellen Distanzen zwischen den Teilnehmern und den Nichtteilnehmern; die Summe dieser Distanzen ist das Optimierungskriterium.

Den Ausgangspunkt des Lösungsalgorithmus von Kuhn bildet eine quadratische Matrix, in deren Zeilen die Teilnehmer, in den Spalten die Nichtteilnehmer stehen. Die Elemente dieser Matrix geben die jeweilige Distanz zwischen einem Teilnehmer und einem Nichtteilnehmer an.
Für jede Zeile dieser Matrix wird das Zeilenminimum ermittelt und von jedem Element in der entsprechenden Zeile subtrahiert. Die ursprünglichen Elemente der Distanzmatrix werden durch die jeweilige Differenz ersetzt. In der resultierenden Matrix werden die Spaltenminima ermittelt und von jedem Element der entspre-

[32] Das Zuordnungsproblem gilt als ein Spezialfall des Transportproblems der linearen Optimierung (Bazaraa, Jarvis und Sherali, 1990, S. 499).

[33] Die Grundidee dieses Lösungsverfahrens kann mit Hilfe der Matrixnotation oder innerhalb der Graphentheorie erläutert werden. Die relevanten Begriffe aus der Graphentheorie und die Zusammenhänge zwischen einem Matchingproblem und der Graphentheorie werden in Rosenbaum (1989) ausführlich erläutert. Eine kurze Darstellung des Algorithmus auf Grundlage der Graphentheorie findet sich in Reinowski, Schultz und Wiemers (2005).

chenden Spalte subtrahiert, die Elemente werden wiederum ersetzt. Aus der so modifizierten Matrix können mögliche Zuordnungen von Nichtteilnehmern zu Teilnehmern abgelesen werden: jedes Null-Element gibt eine mögliche Zuordnung mit minimaler Distanz an. Die Anfangszuordnung unter diesen möglichen Zuordnungen wird durch die Markierung von Zeilen und Spalten erreicht. Dabei werden mit einer möglichst geringen Anzahl markierter Linien (Zeilen oder Spalten) alle Null-Elemente der Matrix abgedeckt. Die markierten Zeilen geben die Teilnehmer an, denen ein Partner zugewiesen wurde, die markierten Spalten die zugeordneten Nichtteilnehmer. Wenn die minimale Anzahl markierter Linien der Anzahl der Zeilen der Matrix entspricht, ist kein weiterer Schritt nötig. Mit der Anfangszuordnung konnte dann bereits allen Teilnehmern ein Partner zugeordnet werden.

Durch die Markierung von Linien entsteht die sog. reduzierte Matrix. Wenn die Zahl der markierten Linien kleiner als die Zeilenanzahl ist, wird die Matrix weiter modifiziert. Unter den nicht markierten Elementen wird das Minimum ermittelt. Von allen nicht markierten Elementen der reduzierten Matrix wird dieses Minimum subtrahiert und die ursprünglichen Elemente durch die Differenz ersetzt. Zu allen doppelt markierten Elementen (Schnittpunkten von markierten Zeilen und markierten Spalten) wird das Minimum addiert und das ursprüngliche Element durch die Summe ersetzt. Alle einfach markierten Elemente bleiben unverändert.
In der so modifizierten Matrix geben die Null-Elemente erneut mögliche Zuordnungen an. Die aktuelle Zuordnung wird wieder durch die Markierung von Zeilen und Spalten unter der o.g. Bedingung vorgenommen, die ursprüngliche Markierung (und damit die Anfangszuordnung) wird aufgehoben. Wenn die minimale Anzahl markierter Linien der Anzahl der Zeilen der Matrix entspricht, ist die aktuelle Zuordnung vollständig und optimal hinsichtlich der Summe der individuellen Distanzen.

Ist die Anzahl markierter Linien kleiner als die Zeilenanzahl, wird die Matrix erneut nach dem o.g. Muster (Ermittlung des kleinsten nicht markierten Elements und Modifizierung der Elemente der Matrix) verändert. Dieser Schritt wird wiederholt bis die optimale Zuordnung erreicht ist.

In der Literatur werden darüber hinaus **Auktionsalgorithmen** für die Lösung des Zuordnungsproblems von Objekten zu Bietern diskutiert (Bertsekas, 1992a; Bikhchandani und Ostroy, 2006).[34]

Für den Auktionsprozess wird unterstellt, dass jedes Objekt jedem Bieter einen individuellen Nutzen stiftet. Dieser Nutzen wird bestimmt aus dem Unterschied zwischen der Wertschätzung des Produkts durch den Bieter und dem Objektpreis. Es wird angenommen, dass der Nutzen jedes Bieters dem Auktionator bekannt ist. Das Ziel des Auktionsprozesses besteht darin, jedem Bieter mindestens ein Objekt zuzuordnen – bei maximalem Gesamtnutzen (der Summe der individuellen Nutzen, der durch die Verteilung der Produkte an die Bieter erreicht wird).

Bei der Zuordnung von Nichtteilnehmern zu Teilnehmern mit Hilfe eines Auktionsalgorithmus können die Teilnehmer als Bieter, die Nichtteilnehmer als Objekte angesehen werden. Die Distanz- bzw. Ähnlichkeitsmatrix kann als Wertschätzung jedes Bieters für jedes Objekt interpretiert werden.

Die einzelnen Schritte der einfachsten Form des Auktionsalgorithmus werden im Folgenden kurz skizziert.[35] In jeder Iteration des Auktionsprozesses wird für einen Bieter, der aktuell kein Objekt besitzt, der individuelle Nutzen aus seinem Lieblingsobjekt und seinem zweitliebsten Objekt ermittelt. Wenn das Lieblingsobjekt noch keinen anderen Besitzer hat, erhält der Bieter dieses Objekt. Wenn das Objekt schon einen Besizer hat, werden zwei Fälle unterschieden: Ist dem Bieter sein zweitliebstes Objekt gleich viel wert, erhält er den Zuschlag für sein zweitliebstes Objekt. Ist das nicht der Fall, wird die ursprüngliche Zuordnung aufgelöst, der aktuelle Bieter erhält sein Lieblingsobjekt. Dem ehemaligen Besitzer wird ein anderes Objekt zugeordnet, wenn das möglich ist (d.h., wenn ein anderes Objekt, aus dem er einen positiven Nutzen ziehen kann, für eine Zuordnung frei ist). Ist dies nicht möglich, muss der ehemalige Besitzer erneut an der Auktion teilnehmen, d.h. in einem der nächsten Schritte wird ihm erneut ein Objekt zugeordnet. Die Auktion ist beendet, wenn jeder Bieter ein Objekt besitzt. Die Zuordnung ist optimal hinsichtlich des kollektiven Gesamtnutzens.

[34] Die Grundidee solcher Algorithmen ist vergleichbar mit dem Konzept der unsichtbaren Hand für die gesamtwirtschaftlich effiziente Güterallokation (Bertsekas, 2001, S. 2).

[35] Die Darstellung orientiert sich an Bertsekas (1981) S. 156ff.

Um zu verhindern, dass dieser Prozess unendlich lange läuft, werden zu Beginn der Auktion Objektpreise festgelegt und im Laufe des Auktionsprozesses verändert.[36] Mit jedem Schritt wird der Preis für das aktuell zugeordnete Objekt erhöht. Damit vermindert sich der Nutzen dieses Objekts für alle Bieter gleichermaßen – seine Attraktivität gegenüber den anderen Objekten nimmt ab. Bieter, deren Wertschätzung dem jetzt teureren Objekt gegenüber eher gering ist, weichen in den folgenden Iterationen auf andere Objekte aus.[37]

Im einfachsten – hier beschriebenen – Fall ist die Anzahl der Bieter gleich der Anzahl der verfügbaren Objekte. Eine Erweiterung des Auktionsalgorithmus erlaubt darüber hinaus eine 1:1-Zuordnung, wenn die Objektanzahl größer als die Anzahl der Bieter ist (Bertsekas, 1992b). Um auch für diese Erweiterungsmöglichkeit eine optimale Zuordnung zu erreichen, muss der beschriebene Auktionsalgorithmus mit einer sog. Rückwärtsauktion kombiniert werden (Bertsekas, Castanon und Tsaknakis, 1993, S. 13). Der beschriebene Auktionsprozess bildet dann die Ausgangszuordnung. Für alle nicht zugeordneten Objekte wird eine vergleichbare Auktion mit „vertauschten Rollen“ durchgeführt (Bertsekas, 1992a, S. 36ff.). In jedem Schritt werden für ein Objekt der beste und der zweitbeste Bieter ermittelt. Der beste Bieter wird dem Objekt zugeordnet, wenn sich dadurch der kollektive Gesamtnutzen erhöht. Die ursprüngliche Zuordnung wird dadurch aufgehoben.

Die Gefahr aller Zuordnungsprozesse ohne Zurücklegen besteht darin, dass der festgestellte Maßnahmeeffekt nicht repräsentativ für die gesamte Gruppe der Teilnehmer ist. Dies ist der Fall, wenn Teilnehmer mit ganz bestimmten Merkmalen systematisch aus der Gruppe der untersuchten Personen entfernt werden, weil kein passender Nichtteilnehmer gefunden wird (Augurzky, 2000a, S. 3). Diesem Problem kann – neben dem Einsatz von Zuordnungsverfahren mit Zurücklegen – auf verschiedene Weise begegnet werden.

[36] Der Anfangspreis der Objekte kann willkürlich festgelegt werden. Er liegt i.d.R. bei Null (Bertsekas, 1981, S. 160).

[37] Damit wird gewährleistet, dass ein Objekt nur einem Bieter zugeordnet wird, dessen Wertschätzung dafür relativ hoch ist.

Local Polynomial Regression

Eine Alternative zur Zuordnung einzelner Personen stellt die Berücksichtigung aller Nichtteilnehmer in der Umgebung eines Teilnehmers dar. Zur Definition dieser Umgebung werden die Merkmale des Teilnehmers verwendet. Dabei wird ausgenutzt, dass sich der Zusammenhang zwischen den beobachteten Merkmalen und dem Einkommen einer Person durch ein Polynom r-ter Ordnung approximieren lässt. Analog zur parametrischen gewichteten KQ-Schätzung kann die Schätzung der Parameter dieser nichtparametrischen Regression dann als Lösung des folgenden Minimierungsproblems für $l = 0, ..., r$ aufgefasst werden:

$$\underset{\kappa_0,...\kappa_r}{min} \sum_{j \in NT} \left(Y_j - \sum_{l=0}^{r} \kappa_l \left(X_i - X_j \right)^l \right)^2 K \left(\frac{X_i - X_j}{b} \right) \tag{2.33}$$

Der Parameter r gibt die polynomiale Ordnung des Minimierungsproblems an. X_i und X_j bezeichnen jeweils die Merkmale des Teilnehmers i und des Nichtteilnehmers j, Y_j das Einkommen des Nichtteilnehmers, $K(.)$ eine nichtnegative Gewichtungsfunktion und b die Bandbreite, die die Umgebung eines Teilnehmers definiert.

Die Vergleichsgröße (2.9) eines Teilnehmers ergibt sich aus dem gewichteten Durchschnitt der Einkommen derjenigen Nichtteilnehmer, die in der Umgebung dieses Teilnehmers liegen:

$$\hat{\bar{Y}}_{i,r}^{C} = \sum_{j \in NT} W \left(\frac{X_i - X_j}{b} \right) Y_j. \tag{2.34}$$

Die Gewichtung (2.10) jedes einzelnen Nichtteilnehmereinkommens hängt vom Abstand des Nichtteilnehmers zum Teilnehmer ab. Je größer dieser Abstand, desto geringer das Gewicht:

$$W(\frac{X_i - X_j}{b}) = \mathbf{e}'\mathbf{G}^{-1} \left\{ 1, \left(X_i - X_j \right), ..., \left(X_i - X_j \right)^r \right\}' K \left(\frac{X_i - X_j}{b} \right). \tag{2.35}$$

Die sog. Glättungsmatrix $\mathbf{G} = \begin{pmatrix} G_0(X_j) & G_1(X_j) & \dots & G_r(X_j) \\ G_1(X_j) & G_2(X_j) & \dots & G_{r+1}(X_j) \\ \vdots & \vdots & \ddots & \vdots \\ G_r(X_j) & G_{r+1}(X_j) & \dots & G_{2r}(X_j) \end{pmatrix}$ für $l = 0, ..., 2r$ besteht aus den Elementen $G_l(X_j) = \sum_{j \in NT} K(\frac{X_i - X_j}{b})(X_i - X_j)^l$. Die

Anzahl der Elemente des Einheitsvektors $\mathbf{e}' = (1, 0, 0, ...)$ wird durch r vorgegeben. Alle Nichtteilnehmer außerhalb der festgelegten Bandbreite werden mit Null gewichtet.

Durch die Wahl der polynomialen Ordnung für die Approximation des Vergleichseinkommens wird die angewendete Methode festgelegt. So lässt sich das Vergleichseinkommen mit Hilfe des Kern Matching als Lösung der Local Polynomial Regression für $r = 0$ darstellen. Die Wahl von $r = 1$ impliziert die Anwendung der Local Linear Regression. Polynome höherer Ordnung sind zur Lösung des Selektionsproblems nicht gebräuchlich.
Der Unterschied zwischen beiden im Folgenden näher erläuterten Methoden wird deutlich, wenn man berücksichtigt, dass r die Anzahl der Parameter einer Schätzung definiert: Kern Matching entspricht der Schätzung einer Konstanten, während die Local Linear Regression eine Konstante und einen linearen Term enthält (Heckman und Smith, 1995, S. 342). Die Berücksichtigung eines linearen Terms zusätzlich zur Konstante in der Schätzung ist immer dann vorteilhaft, wenn potenzielle Kontrollgruppenmitglieder asymmetrisch um die Teilnehmer verteilt sind (Smith und Todd, 2005a, S. 317).

Kern Matching Beim Kern Matching können alle Nichtteilnehmereinkommen zur Konstruktion der Vergleichsgröße eingesetzt werden:

$$C(X_i) = \{j \in \{D = 0\}\}. \tag{2.36}$$

Die Gewichtung (2.10) jedes einzelnen Kontrollgruppenmitglieds für den betrachteten Teilnehmer ist abhängig von der Ähnlichkeit seiner Merkmale zu denen des Teilnehmers. Sie ist nicht auf Null oder Eins beschränkt:

$$W(i, j) = \frac{K_{ij}}{\sum_{j \in NT} K_{ij}}. \tag{2.37}$$

Dabei wird mit $K_{ij} = K\left(\frac{X_i - X_j}{b}\right)$ die Kernfunktion bezeichnet, die zur Gewichtung der einzelnen Nichtteilnehmereinkommen für Teilnehmer i eingesetzt werden soll, und mit b die gewählte Bandbreite. Als Kernfunktion kann jede symmetrische, po-

sitive Funktion, deren Integral eins ist, verwendet werden.[38] Die gebräuchlichsten Funktionen sind die Normalverteilung (Gaußscher Kern) und der Epanechnikovkern.

Das Vergleichseinkommen (2.9) ergibt sich wie folgt:

$$\hat{\bar{Y}}_{i,0}^{C} = \sum_{j \in NT} \frac{K\left(\frac{X_i - X_j}{b}\right)}{\sum_{j \in NT} K\left(\frac{X_i - X_j}{b}\right)} Y_j. \tag{2.38}$$

Durch die Wahl der Bandbreite wird festgelegt, wie viele Nichtteilnehmereinkommen für die Schätzung der Vergleichsgröße für einen Teilnehmer genutzt werden. Dabei besteht ein Zielkonflikt zwischen Unverzerrtheit und Varianz der Schätzung. Mit zunehmender Größe der Bandbreite verringert sich die Varianz des Schätzers (weil mehr Nichtteilnehmereinkommen zur Schätzung verwendet werden), aber die Gefahr der Verzerrung nimmt zu (weil auch solche Nichtteilnehmereinkommen verwendet werden, die dem Teilnehmer weniger entsprechen). Die optimale Bandbreite (im Sinne einer Minimierung des mittleren quadratischen Fehlers) ist abhängig von verschiedenen Faktoren und kann mit Hilfe verschiedener Methoden näherungsweise bestimmt werden. Die am häufigsten eingesetzten Verfahren sind Silvermans Faustregel[39] und das Cross-Validation-Verfahren.

Local Linear Regression Auch bei der Local Linear Regression wird die individuelle Vergleichsgröße aus den Einkommen aller Nichtteilnehmer konstruiert:

$$C(X_i) = \{j \in \{D = 0\}\}. \tag{2.39}$$

Für die Gewichtung (2.10) jedes einzelnen Kontrollgruppenmitglieds für den betrachteten Teilnehmer gilt:

$$W(i,j) = \frac{K_{ij} \sum_{j \in NT} K_{ij} (X_i - X_j)^2 - [K_{ij} (X_i - X_j)] \left[\sum_{j \in NT} K_{ij} (X_i - X_j)\right]}{\sum_{j \in NT} K_{ij} \sum_{j \in NT} K_{ij} (X_i - X_j)^2 - \left[\sum_{j \in NT} K_{ij} (X_i - X_j)\right]^2}. \tag{2.40}$$

[38] Diese Anforderungen werden von jeder symmetrischen Verteilungsfunktion erfüllt.

[39] Die Anwendung dieser Regel erfordert die Verwendung der Gaußschen Kernfunktion. Eine Anwendung findet sich bspw. in Bergemann, Fitzenberger und Speckesser (2001).

Dabei bezeichnet $K_{ij} = K\left(\frac{X_i - X_j}{b}\right)$ die Kernfunktion und b die Bandbreite. Für die Wahl der Kernfunktion gelten die oben getroffenen Aussagen. Die Festlegung der Bandbreite erfolgt i.d.R. mit Hilfe des Cross-Validation-Verfahrens.
Als Vergleichsgröße (2.9) ergibt sich:

$$\hat{\bar{Y}}^C_{i,1} = \frac{G_2(X_j)GY_0(X_j) - G_1(X_j)GY_0(X_j) + G_0(X_j)GY_1(X_j) - G_1(X_j)GY_1(X_j)}{G_0(X_j)G_2(X_j) - G_1^2(X_j)}. \tag{2.41}$$

Die Terme dieser Gleichung werden nach dem oben erwähnten Muster gebildet:

$$G_l(X_j) = \sum_{j \in NT} K(\frac{X_i - X_j}{b})(X_i - X_j)^l \qquad \text{mit } \; l = 0, 1, 2 \quad \text{und}$$

$$GY_l(X_j) = \sum_{j \in NT} K(\frac{X_i - X_j}{b})(X_i - X_j)^l \, Y_j \qquad \text{mit } \; l = 0, 1.$$

Bei der Anwendung der Local Linear Regression auf kleine Stichproben kann in Datenbereichen, in denen wenige Beobachtungen zu finden sind, die Varianz der Schätzung sehr groß werden. Um die Varianz in diesen Bereichen zu verringern, wurden verschiedene Modifikationen entwickelt. Dazu zählen u.a. die Vergrößerung der Bandbreiten für diese Bereiche (Fan et al., 1996; Ruppert, 1997), die Verwendung extrapolierter Datenpunkte (Hall und Turlach, 1997) sowie Trimming und Ridging.[40] Beim Trimming werden nur Daten von Personen zur Schätzung verwendet, deren Merkmalsausprägungen hinreichend häufig auftreten (Heckman et al., 1998, S. 1078f.). Unter Ridging ist die Vergrößerung des Nenners in der Gewichtungsgleichung durch einen stichprobenabhängigen Zusatzterm zu verstehen (Fan, 1992, S. 1000).

[40] Die beiden letztgenannten Modifizierungsvarianten werden in Fröhlich (2004a) näher erläutert.

Eine Variante des Ridge Matching von Seifert und Gasser (1996, 2000) kombiniert Kern Matching und Local Linear Regression und verändert die Vergleichsgröße (2.9) wie folgt:

$$\begin{aligned}\hat{\bar{Y}}_{i,1}^{C} &= (1-\bar{R})\left(\frac{GY_0(X_j)}{G_0(X_j)}\right) + \bar{R}\left(\frac{GY_0(X_j)}{G_0(X_j)} + \frac{GY_1(X_j)(X_i-\bar{X}_j)}{G_2(X_j)}\right) \\ &= \frac{GY_0(X_j)}{G_0(X_j)} + \frac{GY_1(X_j)(X_i-\bar{X}_j)}{G_2(X_j+R)} \\ &= \frac{GY_0(X_j)}{G_0(X_j)} + \frac{GY_1(X_j)(X_i-\bar{X}_j)}{G_2(X_j)+|X_i-\bar{X}_j|ba.}\end{aligned} \tag{2.42}$$

Dabei bezeichnet $\bar{X}_j$ die durchschnittliche Merkmalsausprägung der Nichtteilnehmer in der Umgebung des Teilnehmers i, $\bar{R}$ wird definiert als $\bar{R} = \frac{G_2(X_j)}{G_2(X_j)+R}$. Der Rigde-Parameter R setzt sich wie folgt zusammen: $R = |X_i - \bar{X}_j|ba$, wobei a eine Konstante ist, deren Größe von der Wahl der Kernfunktion bestimmt wird. Die Terme $GY_0(X_j)$ und $GY_1(X_j)$ sowie $G_0(X_j)$ und $G_2(X_j)$ werden wie oben definiert.

1:k-Matching und Full Matching

Eine andere Erweiterungsmöglichkeit des Pair Matching besteht in der Auswahl mehrerer Nichtteilnehmer für jeden Teilnehmer. Dabei kann entweder eine feste Anzahl Partner je Teilnehmer vorgegeben werden (1:k-Matching), oder es werden alle Nichtteilnehmer unter den Teilnehmern aufgeteilt (Full Matching).[41]
Beim Full Matching wird eine Mindestanzahl von Personen, die jedem Teilnehmer zugeordnet werden müssen, festgelegt (i.d.R. einer). Auch die Festlegung von Höchstgrenzen ist möglich, um ausgewogene Gruppengrößen zu erreichen. Die Unter-Kontrollgruppen der Teilnehmer setzen sich aus unterschiedlich vielen Nichtteilnehmern zusammen. Für die Größe der Unter-Kontrollgruppe jedes Teilnehmers gilt: $N_{min}^{C_i} \leq N^{C_i} \leq N_{max}^{C_i}$. Dabei bezeichnet N^{C_i} die Anzahl der Nichtteilnehmer in der Kontrollgruppe, $N_{min}^{C_i}$ gibt die minimale Anzahl der Partner eines Teilnehmers, $N_{max}^{C_i}$ die maximal mögliche Anzahl an.

[41] Die Zuordnung einer flexiblen Anzahl von Kontrollgruppenmitgliedern für jeden Teilnehmer erhöht die Ähnlichkeit der Gruppen im Vergleich zu Algorithmen mit einer vorher festgelegten Anzahl Partner, wie in Ming und Rosenbaum (2000) festgestellt wird. In Hansen und Ohlsen-Klopfer (2006) wird dieser Vorteil anhand eines Beispiels erläutert.

Für die Zuordnung der Nichtteilnehmer sind verschiedene Adaptionen der für Nearest Neighbor Matching vorgestellten Verfahren denkbar.

Greedy Full Matching In Augurzky (2000a) wird ein Full Matching entwickelt, das ausgehend vom Greedy Pair Matching nach dem gleichen Prinzip weitere Nichtteilnehmer zu den gebildeten Unter-Kontrollgruppen zuordnet, bis kein Nichtteilnehmer mehr übrig ist. Anschließend werden diejenigen Teilnehmer, für die bisher kein Partner gefunden wurde, auf die kleinsten Kontrollgruppen aufgeteilt, indem in zufälliger Reihenfolge den Nichtteilnehmern die ähnlichsten verbliebenen Teilnehmer zugeordnet werden. Dieses Verfahren weist die gleichen Probleme auf wie Greedy Pair Matching.

Radius Matching Ebenfalls mehrere Nichteilnehmer werden beim Radius Matching verwendet, das – aufbauend auf dem Caliper Matching – alle Nichtteilnehmer, die innerhalb des Toleranzbereichs liegen, für die Konstruktion des Vergleichseinkommens verwendet (Dehejia und Wahba, 2002).

Auktionsalgorithmen Sowohl für die Zuordnung einer festen als auch einer variablen Anzahl Partner können Algorithmen aus der linearen Optimierung angewendet werden. In der Literatur werden dafür v.a. Auktionsalgorithmen empfohlen (Bertsekas, 1992a; Bikhchandani und Ostroy, 2006).

Der im Abschnitt 2.3.2 beschriebene Algorithmus zur optimalen 1:1-Zuordnung wird erweitert, sodass die Zuordnung von mehr als einem Objekt zu jedem Bieter möglich ist. Um auch für diese Erweiterungsmöglichkeit eine optimale Zuordnung zu gewährleisten, wird der einfache Auktionsalgorithmus mit einer Rückwärtsauktion kombiniert (Bertsekas, Castanon und Tsaknakis, 1993, S. 13), in dessen Verlauf eine dem beschriebenen Auktionsalgorithmus vergleichbare Auktion für alle nicht zugeordneten Objekte durchgeführt wird (Bertsekas, 1992b). In jedem Schritt dieser Rückwärtsauktion wird einem Objekt – zusätzlich zu den bereits zugeordneten Bietern – sein favorisierter Bieter zugeordnet, wenn sich dadurch der kollektive Gesamtnutzen erhöht.

Nach Abschluss der Zuordnung kann das Vergleichseinkommen (2.9) jedes Teilnehmers als Mittelwert der Nichteilnehmereinkommen seiner Untergruppe ermittelt werden. Die einzelnen Nichtteilnehmereinkommen erhalten dabei entweder alle das gleiche Gewicht:

$$W(i,j) = \frac{1}{N^{C_i}}, \tag{2.43}$$

oder werden – analog zum Kern Matching – in Abhängigkeit von der Distanz der Nichtteilnehmer zum Teilnehmer gewichtet.

Der Nachteil aller in den vorangegangenen Abschnitten vorgestellten Matchingverfahren besteht darin, dass mit ihrer Hilfe nur Heterogenitäten zwischen Teilnehmern und Nichtteilnehmern aufgrund beobachtbarer Merkmale beseitigt werden können. Es besteht also die Gefahr, dass das Selektionsproblem nicht vollständig gelöst wird, da die Heterogenitäten aufgrund unbeobachtbarer Eigenschaften unberücksichtigt bleiben. Um dieses Problem zu lösen, werden in der Literatur verschiedene Erweiterungen des Matchingansatzes diskutiert. Im Folgenden wird ein kurzer Überblick über die Verknüpfungsmöglichkeiten von Matching mit anderen Verfahren gegeben.

2.4 Kombination mit anderen Verfahren

2.4.1 Bedingtes Differenz-von-Differenzen-Verfahren

Eine Erweiterungsmöglichkeit besteht in der Kombination von Matching und dem Differenz-von-Differenzen-Verfahren. Dadurch können nach dem Matching noch verbleibende Heterogenitäten aufgrund unbeobachtbarer Faktoren beseitigt werden, falls sie sich im Zeitablauf nicht ändern.[42] Für die Schätzung des durchschnittlichen Maßnahmeeffekts gelten die oben erläuterten Bedingungen. Das Teilnehmereinkommen wird ersetzt durch die entsprechende Einkommensdifferenz zwischen zwei Zeitpunkten t und $t-q$: $\Delta Y_i^T = Y_{it}^T - Y_{i,t-q}^T$. Die Vergleichsgröße (2.9) bildet

[42] Erläuterungen dazu finden sich in Heckman, Ichimura und Todd (1997) sowie Heckman und Smith (1999). Angewendet wird diese Kombination bspw. in Bergemann, Fitzenberger und Speckesser (2001).

dementsprechend der gewichtete Durchschnitt der Einkommensdifferenzen zwischen zwei Zeitpunkten in der Unter-Kontrollgruppe:

$$\Delta\bar{Y}_i^C = \sum_{j=1}^{N^{C_i}} W(i,j) \left(Y_{jt}^C - Y_{j,t-q}^C\right). \tag{2.44}$$

Der Maßnahmeeffekt (2.8) wird dann ermittelt als gewichteter Durchschnitt der Differenzen aus den individuellen Einkommensdifferenzen der Teilnehmer:

$$ME_{cDiD_t} = \sum_{i=1}^{N^T} w(i) \left(\Delta Y_i^T - \Delta\bar{Y}_i^C\right). \tag{2.45}$$

Der Maßnahmeeffekt wird mit ME_{cDiD_t}, die Anzahl der Teilnehmer mit N^T, die Einkommensdifferenzen mit $\Delta Y_i^T - \Delta\bar{Y}_i^C$ und deren Gewichtung mit $w(i)$ bezeichnet.

Diese Kombination ist robuster als die isolierte Verwendung des Matchingansatzes, weil beide möglichen Quellen der Selektionsverzerrung berücksichtigt werden können, wie in Smith und Todd (2005a) bei einem Vergleich der Ergebnisse verschiedener Methoden festgestellt wird.

2.4.2 Regression-adjusted Matching

Alternativ dazu kann mit der Nutzung einer gebräuchlichen Modellannahme für die unterstellte Einkommensgleichung und die Einbeziehung von Ausschlussbedingungen der Einfluss unbeobachtbarer Heterogenitäten berücksichtigt werden. Das Einkommen der Teilnehmer und der Nichtteilnehmer lässt sich mit Hilfe eines herkömmlichen Modells darstellen, in dem beobachtbare und unbeobachtbare Faktoren additiv separierbar sind: $Y_t^T = g_t^T(X) + \epsilon_t^T$ bzw. $Y_t^C = g_t^C(X) + \epsilon_t^C$. Dabei bezeichnen g_t^T und g_t^C jeweils den strukturellen Zusammenhang zwischen dem Einkommen Y_t^T bzw. Y_t^C und den beobachtbaren Faktoren X, mit ϵ_t^T und ϵ_t^C wird der Einfluss der unbeobachtbaren Faktoren bezeichnet. Diese Separierung ist die Grundlage für die Isolierung des Selektionseffekts der Teilnahmeentscheidung, der annahmegemäß nur über die unbeobachtbaren Faktoren im Fehlerterm wirkt (Heckman, LaLonde und Smith, 1999, S. 1923). Unter Anwendung von Ausschlussbedingungen wird X

in verschiedene Komponenten zerlegt: $X = (X_O, X_P)$. In X_O werden alle Variablen zusammengefasst, die das Einkommen determinieren, in X_P diejenigen, die die Teilnahmewahrscheinlichkeit beeinflussen.

Die Berücksichtigung des Einflusses der in X_O enthaltenen Variablen auf den Maßnahmeeffekt ist bspw. mit Hilfe einer partiellen linearen Regression möglich. So verwenden Heckman, Ichimura und Todd (1997) anstelle der Einkommen Y_{it}^T und Y_{jt}^C die Residualgrößen $\tilde{Y}_{it}^T = Y_{it}^T - X_{O,i}\hat{\beta}_t$ und $\tilde{Y}_{jt}^C = Y_{jt}^C - X_{O,j}\hat{\beta}_t$. Mit $\hat{\beta}_t$ werden die Koeffizienten der Schätzung des Effekts der beobachteten Merkmale X_O auf das jeweilige Einkommen bezeichnet. Die individuelle Vergleichsgröße (2.9) ergibt sich als gewichteter Durchschnitt der entsprechenden Residualgrößen der Nichtteilnehmereinkommen:

$$\tilde{\tilde{Y}}_{it}^C = \sum_{j=1}^{N^{C_i}} W(i,j)\, \tilde{Y}_{jt}^C. \tag{2.46}$$

Der durchschnittliche Maßnahmeeffekt (2.8) wird aus der Summe der gewichteten Differenzen der Residualgrößen ermittelt:

$$ME_{RM_t} = \sum_{i=1}^{N^T} w(i) \left(\tilde{Y}_{it}^T - \tilde{\tilde{Y}}_{it}^C \right). \tag{2.47}$$

Dabei bezeichnet ME_{RM_t} den Maßnahmeeffekt zum Zeitpunkt t, $\tilde{Y}_{it}^T$ und $\tilde{\tilde{Y}}_{it}^C$ die Residualgrößen der Teilnehmereinkommen und der entsprechenden Vergleichseinkommen, $w(i)$ gibt die Gewichtung jeder einzelnen Differenz für den Gesamteffekt an.

2.4.3 Korrektur der Abweichungen

Eine Kombination zwischen Matching und linearer Regression wird ebenfalls in Abadie und Imbens (2002) vorgeschlagen. Diese Biaskorrektur soll evtl. Verzerrungen aufgrund von nicht-exaktem Matching berücksichtigen. Das Vergleichseinkommen wird um einen Korrekturterm ergänzt, der die geschätzte Abweichung der Einkommensgrößen zwischen dem Teilnehmer i und der entsprechenden Unter-Kontrollgruppe C_i, die auf Unterschiede in den Merkmalen zurückzuführen ist, enthält. Dieser Korrekturterm setzt sich zusammen aus konsistenten Schätzern für das

hypothetische Einkommen des Teilnehmers i bei Nichtteilnahme und das hypothetische Einkommen des Nichtteilnehmers j bei Teilnahme. Im einfachsten Fall können dafür – in Anlehnung an Regression-adjusted Matching – jeweils die geschätzten Einkommen $\hat{Y}_{it}^{T} = X_i\hat{\beta}_t$ und $\hat{Y}_{jt}^{C} = X_j\hat{\beta}_t$ verwendet werden. Die Vergleichsgröße (2.9) in der Schätzung des Maßnahmeeffekts verändert sich dadurch wie folgt:

$$\hat{\hat{Y}}_{it}^{C} = \sum_{j=1}^{N^{C_i}} W\,(i,j)\,[Y_{jt}^{C} + (X_i\hat{\beta}_t - X_j\hat{\beta}_t)]. \tag{2.48}$$

Der Maßnahmeeffekt (2.8) ergibt sich aus dem Durchschnitt der gewichteten Einkommensdifferenzen zwischen beobachtetem Teilnahmeeinkommen und entsprechendem Vergleichseinkommen:

$$ME_{BC_t} = \sum_{i=1}^{N^{T}} w(i)(Y_{it}^{T} - \hat{\hat{Y}}_{it}^{C}). \tag{2.49}$$

Der Maßnahmeeffekt wird mit ME_{BC_t}, die berücksichtigten individuellen Einkommensdifferenzen mit $Y_{it}^{T} - \hat{\hat{Y}}_{it}^{C}$ und deren Gewichtung wieder mit $w(i)$ bezeichnet.

In den vorangegangenen Abschnitten wird unterstellt, dass nur zwei Entscheidungsalternativen zu einem Zeitpunkt zur Verfügung stehen. Diese Annahme ist in der Realität allerdings selten erfüllt. Im folgenden Kapitel werden deshalb kurz Anpassungsmöglichkeiten für die vorgestellten Verfahren zur Berücksichtigung von mehr als zwei Handlungsalternativen (multiple treatments), der Möglichkeit des Maßnahmeeintritts zu verschiedenen Zeitpunkten (timing of events) sowie der Möglichkeit der mehrfachen Maßnahmeteilnahme (sequential treatment) erläutert.[43]

[43] Da sich die grundlegenden Entscheidungen – die Wahl eines den Daten angemessenen Distanzmaßes und eines geeigneten Zuordnungsprozesses – nicht verändern, werden diese Modellerweiterungen im Rahmen dieser Arbeit nicht ausführlich erläutert. Eine detaillierte Darstellung aller genannten Erweiterungsmöglichkeiten findet sich u.a. in Reinowski (2006).

2.5 Erweiterungen

2.5.1 Berücksichtigung mehrerer Handlungsalternativen

Die Berücksichtigung von mehreren Handlungsalternativen zum gleichen Zeitpunkt verändert den geschätzten Maßnahmeeffekt. Der Teilnahmeeffekt für die Teilnehmer wird nicht mehr als „absoluter Effekt“ bestimmt, sondern im Vergleich zu jeder Handlungsalternative einzeln. Bei jedem dieser paarweisen Vergleiche kann die Existenz der anderen Wahlmöglichkeiten vernachlässigt werden (Lechner, 2001b, S. 49). Bei der Auswahl eines geeigneten Matchingalgorithmus ist allerdigs zu beachten, dass hier die Teilnehmer an einer alternativen Maßnahme die Kontrollgruppe bilden und damit eine eingeschränkte Anzahl von Nichtteilnehmern zur Verfügung steht.[44]

Eine weitere Annäherung an real vorzufindende Entscheidungssituationen stellt die Berücksichtigung der zeitlichen Dimension der Entscheidungen dar. Dabei kann zum einen die Möglichkeit des Maßnahmeeintritts zu verschiedenen Zeitpunkten (timing of events) eine Rolle spielen. Zum anderen ist es in vielen Ländern möglich, mehr als einmal an einer Fördermaßnahme am Arbeitsmarkt teilzunehmen (sequential treatment).

2.5.2 Möglichkeit der Maßnahmeteilnahme zu verschiedenen Zeitpunkten

Wenn die beobachteten Personen zu verschiedenen Zeitpunkten ihrer Arbeitslosigkeit bzw. nach unterschiedlich langer Arbeitslosigkeitsdauer an inhaltlich ähnlichen Maßnahmen teilnehmen können, stellt sich nicht mehr die Frage „Teilnahme oder Nichtteilnahme“, sondern „Teilnahme jetzt oder jetzt (noch) nicht“ (Sianesi, 2004, S. 133). Zur Berücksichtigung dieser Veränderung der Entscheidungssituation sind in der empirischen Literatur im wesentlichen drei Ansätze zu finden.

[44] Aus diesem Grund wird in Lechner (2001b) die Ermittlung der Kontrollgruppe unter Mehrfachnutzung eines Nichtteilnehmers (Zuordnung mit Zurücklegen) vorgeschlagen.
Anwendungsbeispiele solcher paarweisen Vergleiche verschiedener Maßnahmen finden sich u.a. in Gerfin und Lechner (2002), Lechner, Miquel und Wunsch (2004) und Sianesi (2001).

Matching von Zeitdauern

Beim Duration Matching werden die zur Verfügung stehenden Informationen über den Zeitpunkt des Maßnahmeeintritts reduziert. Dazu wird ein fiktiver Stichtag (z.B. 6 Monate nach Beginn der Arbeitslosigkeit) festgelegt. Alle Personen, die bis zu diesem Zeitpunkt eine Maßnahme begonnen haben, gelten als Teilnehmer, alle anderen sind Nichtteilnehmer - unabhängig davon, ob sie später (also nach längerer Arbeitslosigkeit) noch an einer Maßnahme teilnehmen oder nicht. Dieses Vorgehen reduziert die Analyse auf eine statische Betrachtung von „Teilnahme-oder-Nichtteilnahme"-Entscheidungen. Für die Analyse können dann die vorgestellten Verfahren ohne Modifikationen genutzt werden.

Allerdings sollte bei der Festlegung der Nichtteilnehmergruppe eine Mindestbedingung eingeführt werden: Nur diejenigen Nichtteilnehmer, die mindestens genauso lange arbeitslos sind wie ein Teilnehmer bis zum Beginn seiner Maßnahme, sollten als potenzielle Kontrollgruppenmitglieder für diesen Teilnehmer berücksichtigt werden (Lechner, 1999, S. 79). Zur Ermittlung des Maßnahmeeffekts können dann nur Teilnehmer berücksichtigt werden, für die Nichtteilnehmer zu finden sind, die mindestens genauso lange arbeitslos sind und die gleichen relevanten Merkmale besitzen.

Die Anwendung dieses Verfahrens wird in der Literatur allerdings als problematisch angesehen, weil mit der Wahl des Stichtages evtl. wichtige Informationen über den Zeitpunkt des Maßnahmeeintritts verloren gehen. Darüber hinaus wird mit der Einteilung der Stichprobe in Teilnehmer und Nichtteilnehmer implizit eine Annahme über die Zukunft getroffen: Es wird unterstellt, dass die zum Stichtag als Nichtteilnehmer angesehenen Personen auch in der Zukunft nicht an einer Maßnahme teilnehmen werden (Fredriksson und Johansson, 2003, S. 10f.).

Zeitabhängiger Teilnahmeindikator

Die Einführung eines zeitabhängigen (bzw. arbeitslosigkeitsdauerabhängigen) Teilnahmeindikators ermöglicht die Berücksichtigung des Zeitpunkts einer Maßnahme. Der Vorteil dieser Erweiterung des Matchingansatzes liegt in der Möglichkeit der

exakten Zuordnung der betrachteten Personen in Teilnehmer- und Kontrollgruppe (für jeden Zeitpunkt) – ohne implizite Annahmen über die Zukunft.

Allerdings wird mit Hilfe dieses Indikators nicht der Maßnahmeeffekt für die Teilnehmer ermittelt, sondern ein partieller Effekt für Personen, die innerhalb einer bestimmten Arbeitslosigkeitsdauer eine Maßnahme begonnen haben, im Vergleich zur (Noch-) Nichtteilnahme (Fredriksson und Johansson, 2003, S. 14). Für die Definition des Teilnahmeindikators wird unterstellt, dass sich die Arbeitslosigkeitsdauer bis zur Teilnahme in diskrete Zeiteinheiten einteilen lässt (Sianesi, 2004, S. 136). Für jede Zeiteinheit (z.B. Monate) werden Teilnehmer- und Nichtteilnehmergruppen analog zu dem in den vorangegangen Abschnitten beschriebenen Vorgehen gebildet. Der ermittelte Effekt bezieht sich dann nur auf den Maßnahmeeintritt nach einer bestimmten Arbeitslosigkeitsdauer. Er gibt den durchschnittlichen Maßnahmeeffekt für diejenigen an, die nach der Dauer einer bestimmten Anzahl Zeiteinheiten in Arbeitslosigkeit an einer Maßnahme teilnehmen, im Vergleich zur (Noch-) Nichtteilnahme. Um einen „Gesamteffekt“ zu ermitteln, der dem Maßnahmeeffekt für die Teilnehmer vergleichbar ist, müsste ein durchschnittlicher Effekt über alle betrachteten Arbeitslosigkeitsdauern geschätzt werden (Fredriksson und Johansson, 2003, S. 14).

Alternative: Verweildaueranalyse

Eine weitere Möglichkeit der Berücksichtigung der Informationen über unterschiedliche Eintrittszeitpunkte besteht in der Anwendung von Modellen der Verlaufsanalyse. Zur Analyse der Effekte der Arbeitsmarktpolitik werden vor allem Mixed Proportional Hazard Models verwendet.[45] Der Effekt einer Maßnahmeteilnahme auf die Übergangsgeschwindigkeit in Beschäftigung wird in diesem Rahmen mit Hilfe eines rekursiven bivariaten Modells für den Übergang in Beschäftigung und den Übergang in eine Maßnahme ermittelt (Abbring und van den Berg, 2003, S. 1503).

[45] Solche Modelle finden sich in der jüngeren Evaluationsliteratur recht häufig (Caliendo und Hujer, 2006; Lalive, van Ours und Zweimüller, 2002; Zijl, van den Berg und Heyma, 2004), basieren aber auf einer anderen theoretischen Grundlage als Matchingverfahren. Aus diesem Grund werden sie hier nur kurz erwähnt.

2.5.3 Möglichkeit der mehrfachen Teilnahme im Zeitablauf

Im Falle einer mehrfachen Teilnahme an (gleichartigen oder unterschiedlichen) Maßnahmen ist es möglich, dass sich frühere Maßnahmen auf aktuelle Entscheidungen und den beobachteten Arbeitsmarkterfolg auswirken (Lechner und Miquel, 2001, S. 3). In diesem Fall ist es wichtig, den Untersuchungsgegenstand und die Vergleichsgröße genau zu definieren (Sianesi, 2004, S. 135). Eine Möglichkeit besteht darin, den Effekt einer spezifischen Maßnahme zu evaluieren, wofür i.d.R. nur die erste Maßnahme, die eine Person absolviert hat, in Frage kommt.[46] In diesem Fall kann die Evaluation innerhalb des oben beschriebenen statischen Rahmens erfolgen. Dabei gehen allerdings Informationen darüber verloren, wie sich diese erste Maßnahme auf spätere Entscheidungen auswirkt und wie evtl. später absolvierte Maßnahmen auf den Arbeitsmarkterfolg wirken.

Getrennte Schätzung der Effekte

Mit Hilfe eines in Bergemann, Fitzenberger und Speckesser (2004) verwendeten bedingten Differenz-von-Differenzen-Schätzers ist die Schätzung der Effekte der ersten und zweiten Maßnahme auf den Arbeitsmarkterfolg möglich. Durch die Gegenüberstellung der Einkommensdifferenz der Teilnehmer an der zweiten Maßnahme mit dem entsprechenden Vergleichseinkommen lassen sich zwei Effekte ermitteln. Die gemeinsame Wirkung der ersten und zweiten Maßnahme (combined effect) wird ermittelt, indem die Einkommensdifferenz nach der zweiten Maßnahme mit derjenigen vor Beginn der ersten Maßnahme verglichen wird. Welche Wirkung die zweite Maßnahme hat (incremental effect), wird durch den Vergleich der Situationen vor Beginn und nach Beendigung der zweiten Maßnahme festgestellt. Die Kontrollgruppe wird dabei aus allen Personen gebildet, die nicht an zwei gleichartigen Maßnahmen teilgenommen haben, also sowohl aus „wahren Nichtteilnehmern“ und Teilnehmern an nur einer Maßnahme als auch Teilnehmern an einer gleichartigen Maßnahme und anderen (nicht berücksichtigten) Maßnahmen.

[46] Ein solches Vorgehen findet sich häufig in der empirischen Literatur (Reinowski, Schultz und Wiemers, 2005; Sianesi, 2001).

Dieser Ansatz ermöglicht die Bestimmung der Wirkung verschiedener Maßnahmen auf den Arbeitsmarkterfolg, aber nicht die Berücksichtigung des Einflusses dieser Maßnahmen auf spätere Entscheidungen.

Dynamischer Modellrahmen

Unter Verwendung eines dynamischen Modellrahmens ist es möglich, auch solche Effekte zu berücksichtigen (Lechner und Miquel, 2001, S. 28). Der Unterschied zu den bisher vorgestellten Ansätzen besteht in der Zusammenfassung der in der Vergangenheit getroffenen Entscheidungen zu Entscheidungssequenzen, deren Ergebnisse miteinander vergleichbar sind.[47] Wie aufwendig die Suche nach geeigneten Kontrollgruppenmitgliedern für diesen Vergleich ist, hängt davon ab, in wie vielen Perioden sich die Entscheidungen der verglichenen Gruppen unterscheiden.

2.6 Zusammenfassung

Matchingverfahren sind die in der Literatur am weitesten verbreiteten nichtparametrischen Lösungsverfahren für das Selektionsproblem. Sie beruhen auf zwei grundlegenden Annahmen. Die Annahme bedingter Unabhängigkeit besagt, dass alle Personen mit identischen beobachtbaren (und beobachteten) Merkmalen die gleichen potenziellen Einkommen haben. Dies trifft sowohl auf den Teilnahme- als auch auf den Nichtteilnahmefall zu. Weiterhin wird angenommen, dass die beobachteten Merkmalsausprägungen in der Teilnehmer- und auch in der Nichtteilnehmergruppe auftreten. Nur innerhalb des Common-Support-Bereichs ist die Schätzung des Maßnahmeeffekts möglich.

Im Zusammenhang mit dem Einsatz von Matching in der empirischen Forschung müssen zwei Entscheidungen getroffen werden: Zum einen muss ein Distanzmaß gewählt werden, mit dem die zur Verfügung stehenden Informationen zusammengefasst werden können, zum anderen muss der eingesetzte Zuordnungsprozess den

[47] Die Grundidee der Modellierung solcher Entscheidungssequenzen wird in Lechner (2004) ausführlich erläutert.

Daten gerecht werden. Für beide Entscheidungen stehen verschiedene Alternativen zur Verfügung, die in diesem Kapitel erläutert werden. Für den Einsatz in der empirischen Evaluation sind diejenigen Distanzmaße vorzuziehen, mit deren Hilfe die Berücksichtigung verschieden skalierter Merkmale ohne Informationsverlust möglich ist. Dafür kommen aggregierte Distanzmaße und Balancing Scores in Frage. Der Propensity Score und der Index Score werden in empirischen Studien sehr häufig angewendet. Aus theoretischen Überlegungen heraus bietet sich v.a. für kleine Stichproben allerdings eher die Nutzung aggregierter Distanzmaße an. Von den hier vorgestellten Maßen werden die Mahalanobis-Matching-Distanz und das Ähnlichkeitsmaß von Gower – neben Propensity Score und Index Score – in einem späteren Kapitel im Rahmen einer Simulation miteinander verglichen.

Für die Zuordnung von Partnern auf Grundlage der ermittelten Distanzen bzw. Ähnlichkeiten stehen ebenfalls sehr unterschiedliche Verfahren zur Verfügung. Welche Verfahren in Frage kommen, richtet sich nach der Größe des Datensatzes, d.h. der Größe der Stichprobe der Nichtteilnehmer im Vergleich zur Teilnehmerstichprobe. Sind beide Stichproben etwa gleich groß, bietet sich die Nutzung von Zuordnungen mit Zurücklegen oder Local Polynomial Regression an. Stehen dagegen relativ viele potenzielle Partner für die Teilnehmer zur Verfügung, können Zuordnungsverfahren ohne Zurücklegen eingesetzt werden. Der Einsatz optimaler Zuordnungsprozesse erscheint sinnvoll, da mit ihnen die Realisierung der bestmöglichen Zuordnung gewährleistet ist. Aus der Gruppe der optimalen Zuordnungen werden deshalb zwei Verfahren in die Simulation einbezogen: der Ungarische Algorithmus für optimale 1:1-Zuordungen sowie ein Auktionsalgorithmus für Optimal Full Matching.

In der Literatur werden verschiedene Erweiterungen und Ergänzungen des „einfachen“ Matchingverfahrens – zur Lösung des Problems der unbeobachtbaren Heterogenität bzw. zur Annäherung an in der Realität vorzufindende Entscheidungssituationen – diskutiert, die in diesem Kapitel ebenfalls kurz erläutert werden. Diese Erweiterungsmöglichkeiten ändern aber nichts an den grundlegenden Entscheidungen im Zusammenhang mit dem Einsatz von Matchingverfahren für empirische Untersuchungen. In der Simulation wird deshalb keine dieser Erweiterungen berücksichtigt.

Kapitel 3

Stand der Forschung zur Entwicklung von Handlungsempfehlungen

Die Betrachtung der Eigenschaften der verschiedenen Matchingverfahren, der Vor- und Nachteile der einzelnen Distanzmaße und Zuordnungsprozesse und die Frage, welches Verfahren in welcher konkreten Situation das beste ist, gewinnen in der Literatur zunehmend an Bedeutung.
In der jüngeren Evaluationsliteratur lassen sich verschiedene Zweige der Analyse identifizieren. So findet man Studien zu den (asymptotischen) Eigenschaften der Verfahren, allgemeine Handlungsempfehlungen zur Auswahl eines Verfahrens sowie Vergleiche verschiedener Ansätze. Im Folgenden werden die jeweils wichtigsten Studien kurz vorgestellt.[1]

3.1 Untersuchung asymptotischer Eigenschaften

Eine vielzitierte Arbeit von Rosenbaum und Rubin (1983) beschäftigt sich mit den asymptotischen Eigenschaften von Balancing Scores und dem Nachweis, dass der Propensity Score (PS) der gröbste Balancing Score ist. Das sog. „Rosenbaum-Rubin-Theorem“ (Angrist und Hahn, 2004, S. 58) besagt, dass zur Erfüllung der Annahme der bedingten Unabhängigkeit asymptotisch die Verwendung eines Balancing Scores ausreicht, wenn der Score aus Merkmalen gebildet wird, die diese Annahme erfüllen.

In der Folgezeit wurde der Propensity Score zum beliebtesten Distanzmaß für empirische Anwendungen. Diese Popularität spiegelt sich auch in der analytischen Literatur wider. So zeigen Heckman, Ichimura und Todd (1998), dass sich PS-Matching und exaktes Matching weder im asymptotischen Bias noch in der asymptotischen Varianz voneinander unterscheiden.[2] Es wird betont, dass der Vorteil von PS-Matching in der Dimensionsreduktion und damit der leichteren Handhabung dieses Distanzmaßes liegt. In Hahn (1998) werden Effizienzgrenzen für die Varianz PS-basierter Schätzungen des Maßnahmeeffekts entwickelt. Es wird gezeigt, dass

[1] Die Darstellung beschränkt sich auf Studien, deren Gegenstand der durchschnittliche Effekt für die Teilnehmer ist.

[2] Als exaktes Matching werden Verfahren bezeichnet, in denen die Kontrollgruppe anhand der einzelnen Matchingvariablen gebildet wird, d.h. die Übereinstimmung der verglichenen Personen wird für jedes Merkmal einzeln überprüft (Schmidt, 1999, S. 26). Nicht-exaktes Matching bezeichnet Verfahren, in denen die Merkmale zu einem Distanz- oder Ähnlichkeitsmaß bzw. einem Score zusammengefasst werden.

die Kenntnis des „wahren“ Propensity Score die Varianz der Schätzung des Maßnahmeeffekts für die Teilnehmer verringert.

Die von Hahn (1998) entwickelten Effizienzgrenzen werden in einer Untersuchung der asymptotischen Eigenschaften von exakten Matchingverfahren aufgegriffen. Abadie und Imbens (2002) beschränken die Menge der untersuchten Schätzer auf Verfahren mit Zurücklegen mit einer festen Anzahl von Partnern. Die wichtigsten Ergebnisse der Analyse lassen sich wie folgt zusammenfassen: Beim exakten Matching sind die Schätzer konsistent und asymptotisch normalverteilt. Allerdings ist die Varianz der Schätzer größer als die von Hahn (1998) entwickelte Effizienzgrenze. Damit sind sie nicht effizient. Bei nicht-exaktem Matching beinhalten sie einen Biasterm, der mit zunehmender Größe der Stichprobe nicht gegen Null tendiert und dessen Größe von der Anzahl der berücksichtigten stetigen Variablen abhängt. Damit sind Standard-Konfidenzintervalle und Standard-Tests nicht anwendbar.

In Angrist und Hahn (2004) werden exaktes Matching und PS-Matching hinsichtlich ihrer asymptotischen Effizienz miteinander verglichen. Dabei wird festgestellt, dass PS-Matching besser im Sinne einer höheren Effizienz als exaktes Matching ist. Es ist insbesondere dann vorzuziehen, wenn der Erklärungswert der einzelnen Variablen für die Erfolgsgröße gering ist, die Teilnahmewahrscheinlichkeit nahe Null oder Eins liegt und nur wenige potenzielle Partner für jeden Teilnehmer verfügbar sind. Diese Ergebnisse bestätigen sich auch bei Verwendung einer alternativen Definition der Asymptotik.[3]

Von verschiedenen Autoren wird kritisch angemerkt, dass die theoretische Analyse der asymptotischen Eigenschaften für Matchingverfahren nur wenig Hilfe für die Anwendung – auf Stichproben begrenzter Größe – liefert (Angrist und Hahn, 2004). In der Literatur findet sich daher auch eine wachsende Anzahl von Studien, die allgemeine Hinweise für den Gebrauch von Matchingverfahren liefern. Die inhaltlichen Schwerpunkte dieser Studien bilden zum einen Plausibilitätsüberlegungen zur Ein-

[3]Diese Definition ist angelehnt an die Definition der Panelasymptotik. Hier ist die Zellgröße fix (d.h. die Anzahl möglicher Partner ist festgelegt) und die Anzahl der Zellen geht gegen unendlich (d.h. die Anzahl möglicher Ausprägungen der Kovariate nimmt zu). Im Gegensatz dazu steht die koventionelle Definition von Asymptotik, nach der die Anzahl der Zellen fix ist und ihre Größe gegen unendlich geht.

haltung der Annahmen, zum anderen Strategien zur Identifizierung der relevanten Merkmale.

3.2 Allgemeine Handlungsempfehlungen

Einen relativ breiten Raum nehmen Überlegungen zur Plausibilität der Einhaltung der nicht testbaren Grundannahmen des Matching ein.

Zur Überprüfung der Vereinbarkeit der Annahme bedingter Unabhängigkeit (CIA) mit den zugrunde liegenden Daten werden in der Literatur zwei Verfahren vorgeschlagen. Die Tests beruhen auf der Idee, dass bei Gültigkeit der CIA die Durchschnittseinkommen in zwei Teilstichproben übereinstimmen müssten, wenn keine der beiden Gruppen an der evaluierten Maßnahme teilnimmt. Das gilt sowohl im Vorfeld der Maßnahme als auch danach.

Im Pre-Program-Test (Heckman und Hotz, 1989) werden die Durchschnittseinkommen der Teilnehmer- und der Nichtteilnehmergruppe für einen Zeitpunkt vor Beginn der Maßnahme verglichen. Wenn hier ein Unterschied auftritt, wird vermutet, dass nicht alle Ursachen der Selektionsverzerrung beseitigt worden sind. Die Schlüsselannahme dafür – ähnlich wie für das Differenz-von-Differenzen-Verfahren – besteht darin, dass die Unterschiede zwischen Teilnehmern und Nichtteilnehmern zeitinvariant sind.

Von der Grundidee sehr ähnlich ist der Post-Program-Test (Rosenbaum, 1987). Dafür ist eine zusätzliche Personengruppe nötig, die als potenzielle Kontrollgruppe für die Teilnehmergruppe in Frage käme.[4] Aus der zusätzlichen potenziellen Kontrollgruppe wird eine Kontrollgruppe für die in der eigentlichen Untersuchung verwendete Kontrollgruppe konstruiert. Dafür werden dieselben Variablen und dasselbe Matchingverfahren verwendet wie zur Ermittlung dieser Kontrollgruppe. Anschließend wird für die in der Analyse verwendete Kontrollgruppe der „Maßnahmeeffekt“ geschätzt. Dieser Effekt sollte Null sein. Ein signifikanter Unterschied von Null lässt

[4] In Heckman, Ichimura und Todd (1997) wird dieser Test bspw. mit einer Gruppe nicht teilnahmeberechtigter Personen und einer Gruppe berechtigter, aber nicht an der untersuchten Maßnahme teilnehmender Personen, durchgeführt.

vermuten, dass die in der Analyse verwendete Kontrollgruppe noch systematische Unterschiede zur Teilnehmergruppe aufweist.

Imbens (2004) betont, dass die Nicht-Ablehnung der Nullhypothese in beiden Fällen kein statistisch gesicherter Beweis für die Erfüllung der CIA ist, aber als Indiz für deren Gültigkeit gewertet werden kann.

Für die Überprüfung der Common-Support-Bedingung (CSC) lassen sich in der Literatur ebenfalls verschiedene Strategien finden. In Hujer, Maurer und Wellner (1997) werden die Verteilungen des Propensity Scores in der Teilnehmer- und der Kontrollgruppe geplottet. Anhand dieser Plots wird der Common-Support-Bereich grafisch überprüft.[5] Eine Alternative zur grafischen Überprüfung bietet der Vergleich der minimalen und maximalen PS-Werte in beiden Teilstichproben (Caliendo und Kopeinig, 2005). Allerdings können mit dieser Methode Lücken in der PS-Verteilung nicht identifiziert werden. Dies ist möglich, wenn der Common-Support-Bereich definiert wird als diejenigen PS-Werte, deren Dichtefunktion in beiden Teilstichproben größer als Null ist (Caliendo und Kopeinig, 2005). Imbens (2004) empfiehlt die Überprüfung des schlechtesten Matches für jede betrachtete Variable. Wenn diese maximale Abweichung im Vergleich zur Standardabweichung der Variable in der Stichprobe groß ist, ist eine Verletzung der CS-Bedingung zu vermuten. Er gibt allerdings keinen Hinweis darauf, welche Abweichung als „groß“ anzusehen ist.

Hinsichtlich der Strategie zur Identifizierung der relevanten Merkmale und der Spezifikation des Modells zur Schätzung des Propensity Scores gibt es in der Literatur sehr unterschiedliche Empfehlungen. Eine weit verbreitete Methode ist der Rückgriff auf die ökonomische Theorie, die Ergebnisse bisheriger Studien und die Berücksichtigung des Selektionsprozesses (Christensen, 2001; Sianesi, 2002) für die Auswahl der einbezogenen Variablen. In Black und Smith (2004) werden statistische Gütemaße zur Auswahl des Variablensets genutzt. Allerdings wird in Heckman und Navarro-Lozano (2004) anhand verschiedener Szenarien verdeutlicht, dass die Nutzung statistischer Gütemaße keine Garantie für die vollständige Beseitigung der Selektionsverzerrung bietet.

[5] Eine grafische Überprüfung ist ebenfalls mit Hilfe von Histogrammen möglich (Smith und Todd, 2005a).

Basierend auf der Zerlegung des Schätzfehlers in seine drei Bestandteile (fehlender Common Support, unterschiedliche Verteilung der Variablen in beiden Teilstichproben und Selektionsverzerrung aufgrund unbeobachteter Heterogenitäten) stellen Heckman, Ichimura und Todd (1997, 1998) und Heckman et al. (1998) fest, dass Matching in der Lage ist, die ersten beiden Fehlerquellen zu beseitigen. Für eine unverzerrte Schätzung mit Hilfe von Matchingalgorithmen sind allerdings umfangreiche Informationen über den Selektionsprozess und alle einkommensrelevanten Variablen nötig, insbesondere Informationen über die Arbeitsmarkt-Vorgeschichte der beobachteten Personen. Darüber hinaus sollte die potenzielle Vergleichsgruppe aus demselben Datensatz wie die Teilnehmergruppe stammen, und für beide Teilstichproben sollten identische Informationen vorliegen. Eine weitere Bedingung für eine unverzerrte Schätzung besteht in der Berücksichtigung regionaler Besonderheiten bzw. der Beschränkung des Vergleichs auf Personen eines regionalen Arbeitsmarktes.

Neben der Überprüfung der Plausibilität der Annahmen und der Identifizierung relevanter Variablen ist für die empirische Forschung interessant, welches Verfahren in welcher Ausgangssituation die besten Matchingergebnisse verspricht. Der weitaus größte Teil der Literatur zur Entwicklung von Handlungsempfehlungen beschäftigt sich mit dieser Frage. In den Studien werden verschiedene Verfahren in unterschiedlichen Situationen miteinander verglichen.

3.3 Vergleich von Matchingschätzern

Die Studien zur Wahl eines geeigneten Matchingverfahrens in verschiedenen Situationen lassen sich in zwei Gruppen einteilen. Sensitivitätsanalysen nutzen bereits vorhandene Datensätze und vergleichen die Schätzergebnisse der analysierten Verfahren mit einem Benchmark, dem „wahren“ Maßnahmeeffekt. Dieser Benchmarkeffekt wird i.d.R. mit Hilfe von Daten aus sozialen Experimenten ermittelt. Die Größe der Abweichung gibt Auskunft über die Eignung des analysierten Verfahrens für die Abbildung und Nutzung der zugrunde liegenden Datenbasis. In Simulationsstudien werden künstliche Datensätze generiert, um bestimmte Eigenschaften mit Hilfe statistischer Kennzahlen untersuchen zu können.

3.3.1 Sensitivitätsanalysen

Eine der ersten Studien zum Vergleich verschiedener Lösungsverfahren für das Selektionsproblem ist die von LaLonde (1986), in der mehrere nichtparametrische Verfahren mit den Ergebnissen eines sozialen Experiments, der National Supported Work Demonstration (NSW), verglichen werden. Es wird festgestellt, dass die Verfahren sehr unterschiedliche Ergebnisse produzieren und teilweise erheblich vom Benchmark der experimentellen Daten abweichen. LaLonde schlussfolgert daraus, dass diese Verfahren nicht geeignet sind, die Selektionsverzerrung effektiv zu beseitigen.

Viele der darauf folgenden Studien beziehen sich auf diese Analyse. Ein sehr bekanntes Beispiel bildet die Debatte zwischen Dehejia/Wahba und Smith/Todd über die Möglichkeiten von Matchingverfahren, LaLondes Kritik zu begegnen.

In Dehejia und Wahba (1999) wird der gleiche Datensatz (NSW) zur Überprüfung der Performance von Propensity Score Matching verwendet. Es wird festgestellt, dass die Ergebnisse der experimentellen Daten besser repliziert werden können als in LaLonde (1986), wenn das Verfahren von Dehejia und Wahba (1999) zur Spezifikation der Schätzgleichung für den Propensity Score angewendet wird. Daraus wird die Schlussfolgerung gezogen, dass Propensity Score Matching geeignet ist, valide Schätzungen des Maßnahmeeffekts zu erzielen, wenn keine experimentellen Daten vorliegen.

Dehejia und Wahba (2002) vergleichen verschiedene Schätzer auf Basis des Propensity Scores. Dabei werden Zuordnungsverfahren ohne Zurücklegen und Zuordnungsverfahren mit Zurücklegen betrachtet. Als Qualitätskriterium wird wieder die Abweichung des geschätzten Maßnahmeeffekts vom Benchmark eingesetzt. Die wichtigsten Schlussfolgerungen lassen sich wie folgt zusammenfassen: Die Wahl eines Matchingverfahrens ist abhängig von der Größe des Common-Support-Bereichs. Bei relativ kleinem Common Support sind Verfahren mit Zurücklegen besser. Bei großem Common Support sind Verfahren ohne Zurücklegen die bessere Wahl. Darüber hinaus wird festgestellt, dass der Vorteil von Caliper Matching gegenüber dem Nearest Neighbor Matching sehr gering ist.

Ebenfalls mit den gleichen Daten untersuchen Smith und Todd (2005a) Matchingverfahren auf Grundlage des Propensity Scores mit den von LaLonde (1986) und

Dehejia und Wahba (1999) verwendeten Spezifikationen. Als Gütemaß wird die Abweichung der Mittelwerte der einzelnen Variablen zwischen beiden Teilstichproben verwendet. Aus der Analyse geht hervor, dass Propensity Score Matching ein sinnvolles Instrument, aber keine generelle Lösung des Selektionsproblems ist. Alle Schätzer sind sehr sensitiv gegenüber der Wahl der für die Schätzung des Propensity Scores verwendeten Variablen und der zugrunde liegenden Stichprobe. Die Kombination von Matching mit dem Differerenz-von-Differenzen-Verfahren wird favorisiert, da bei Anwendung dieser Verfahrenskombination die geringsten Aweichungen der Merkmalsmittelwerte auftreten.

Die Studie von Smith und Todd (2005a) wird ergänzt durch eine Analyse von Zhao (2006) auf Basis derselben Daten. Gegenstand dieser Studie ist der Vergleich von Propensity Score Matching mit Matchingverfahren auf Grundlage der Euklidischen Distanz und der Mahalanobisdistanz sowie Matching mit gewichteten Kovariaten in zwei Varianten und der bias-korrigierte Schätzer von Abadie und Imbens (2002).[6] Es wird kein Distanzmaß identifiziert, das den anderen überlegen ist.

Abadie und Imbens (2002) vergleichen die Performance des von ihnen entwickelten bias-korrigierten Schätzers mit der eines einfachen Matchings (auf Basis aller Variablen sowie auf Basis des Propensity Scores) und der Performance von Regressionen in verschiedenen Spezifikationen. Die Grundlage bilden ebenfalls die NSW-Daten sowie eine Simulation, in der diese Daten rekonstruiert werden. Als Benchmark dient der „wahre“ Maßnahmeeffekt, der mit Hilfe der experimentellen Daten ermittelt wurde. Die Simulation ermöglicht darüber hinaus die Schätzung von Bias, Standardabweichung sowie mittlerer quadratischer Abweichung (MSE). Als wichtigstes Ergebnis lässt sich festhalten: Matching, besonders der hier eingeführte bias-korrigierte Schätzer, ist hinsichtlich der untersuchten Gütemaße besser als die getesteten Alternativen. Unter den Matchingverfahren ist die Angleichung der Verteilungen der Variablen in Teilnehmer- und Nichtteilnehmergruppe bei beiden Formen des einfachen Schätzers abhängig von der Größe der individuellen Unter-Kontrollgruppen: je größer die Zahl der zugeordneten Partner, desto schlechter die Angleichung der

[6] Dieser Schätzer wird im Abschnitt 2.4 näher erläutert.

Verteilungen. Der bias-korrigierte Schätzer ist robuster gegenüber der Anzahl der einbezogenen Partner.

Auf Basis der Daten des National Job Training Partnership Act (JTPA) vergleichen Heckman, Ichimura und Todd (1997, 1998) und Heckman et al. (1998) ein einfaches Propensity Score Matching mit verschiedenen Matching-Erweiterungen.[7] Als Gütemaß dient der geschätzte Bias der einzelnen Verfahren. Hinsichtlich dieses Kriteriums sind das einfache Propensity Score Matching und Local Linear Matching annähernd gleich zu bewerten. Die Kombination der erweiterten Schätzer mit dem Differenz-von-Differenzen-Verfahren liefert den geringsten Bias.

In Augurzky (2000b) wird die Sensitivität des Einkommenseffekts eines Bachelorabschlusses gegenüber der Wahl verschiedener Distanzmaße und Zuordnungsalgorithmen für eine Stichprobe aus dem National Longitudinal Survey of Youth 1979 untersucht. Es werden drei verschiedene Distanzmaße (Propensity Score, Index Score sowie die Mahalanobisdistanz) und drei Zuordnungsalgorithmen (Optimal Full Matching, Greedy Pair Matching sowie ein von Augurzky entwickeltes Greedy Full Matching) miteinander verglichen.[8] Zur Beurteilung der Ergebnisse werden zwei Kriterien eingesetzt: die Ähnlichkeit der Verteilungen der Merkmale in der Teilnehmer- und der Kontrollgruppe sowie die Varianz des Ergebnisses. Die wichtigsten Resultate können in drei Punkten zusammengefasst werden: Greedy Full Matching verteilt die Nichtteilnehmer gleichmäßiger über alle individuellen Unter-Kontrollgruppen als Optimal Full Matching, die Verteilungen der Merkmale in Teilnehmer- und Kontrollgruppe sind für beide Verfahren etwa gleich ähnlich. Greedy Pair Matching errreicht – definitionsgemäß – den höchsten Grad der Uniformität der Größe der individuellen Unter-Kontrollgruppen und die ähnlichsten Merkmalsverteilungen. Allerdings ist ein systematischer Verlust von Beobachtungen festzustellen.[9] Dieser Verlust würde bei heterogenen Maßnahmeeffekten dazu führen, dass das Ergebnis nicht mehr repräsentativ für alle Teilnehmer ist. In der Studie wird die Angleichung der Verteilungen der

[7] Erläuterungen zu den hier verwendeten Verfahren (Regression-adjsuted Matching, Local Linear Matching sowie die Kombination beider Schätzer mit dem Differenz-von-Differenzen-Verfahren) finden sich in den Abschnitten 2.3.2 und 2.4.

[8] Vgl. die Abschnitte 2.3.1 und 2.3.2 für eine Erläuterung der Verfahren.

[9] Das Problem des Verlustes von Beobachtungen wird im Zusammenhang mit der Vorstellung des Greedy Pair Matching im Abschnitt 2.3.2 erläutert.

Merkmale in Teilnehmer- und Nichtteilnehmergruppe deshalb als Hauptkriterium für die Wahl eines Matchingverfahrens erachtet. In Stichproben begrenzter Größe mit starkem Selektionseffekt und heterogenem Maßnahmeeffekt wird allerdings ein Tradeoff zwischen der Varianz des Ergebnisses und der Angleichung der Verteilungen der Merkmale festgestellt.

Die gleichen Distanzmaße und Zuordnungsalgorithmen werden von Augurzky und Kluve (2004) noch einmal untersucht. Die Ergebnisse dieser Studie bestätigen Augurzkys Resultate. Ausgehend von den Untersuchungsergebnissen wird die Verwendung eines (im Sinne minimaler Distanzen) optimalen Zuordnungsalgorithmus' mit Beschränkung der maximal zulässigen Anzahl der Personen in einer Unter-Kontrollgruppe empfohlen.

3.3.2 Simulationsstudien

Augurzky (2000a) analysiert den Einfluss von Fehlern bei der Spezifikation des Modells zur Propensity-Score-Schätzung auf die Ähnlichkeit der gematchten Gruppen. Zur Berücksichtigung von Fehlspezifikationen hinsichtlich der einbezogenen Variablen werden zwei Modellvarianten betrachtet. Das partielle Modell enthält nur die Variablen, die sowohl die Teilnahmeentscheidung als auch das Einkommen beeinflussen. Im vollständigen Modell werden alle verfügbaren Variablen berücksichtigt. Eine Fehlspezifikation hinsichtlich der angenommenen Verteilung wird durch eine Abweichung von der Standardnormalverteilung und die Hinzunahme von Interaktions- und quadratischen Termen simuliert. Als wichtigstes Ergebnis lässt sich festhalten: Das partielle Probitmodell liefert – in Abhängigkeit der Wichtigkeit der unberücksichtigten Variablen in der Propensity-Score-Schätzung – teilweise stark verzerrte Matchingergebnisse. Das vollständige Probitmodell dagegen erreicht in den meisten Fällen eine unverzerrte Schätzung des Maßnahmeeffekts. Je stärker der Einfluss der Variablen auf den Teilnahmeindikator ist, desto größer ist der mittlere quadratische Fehler (MSE) des vollständigen Modells. Der MSE des partiellen Modells dagegen verändert sich nur geringfügig in den verschiedenen Designs. Das unterschiedliche Verhalten erklärt sich aus der unterschiedlichen Varianz der Ergebnisse beider Modelle. Während der Verlust von Beobachtungen im vollständigen Modell

immer größer wird, je wichtiger die Variablen werden, verändert sich die Anzahl der betrachteten Fälle im partiellen Modell nur geringfügig.
Die Analyse eines alternativen Modells, in dem die Verteilung der Variablen von der angenommenen Standardnormalverteilung abweicht und Interaktionsterme einbezogen werden, verändert die Ergebnisse des Basismodells insofern, als sich Bias und MSE beider Modelle vergrößern, wobei die Veränderung des partiellen Modells größer ist. Aufbauend auf den Ergebnissen dieser Simulation wird – entgegen der bis dahin gültigen Praxis, alle verfügbaren Variablen, die die Teilnahmeentscheidung beeinflussen, in der Schätzung zu berücksichtigen – argumentiert, dass die Beschränkung auf diejenigen Variablen, die hoch signifikant sind, zu besseren Matchingergebnissen im Sinne einer geringeren mittleren quadratischen Abweichung führt.

Mit Hilfe der Monte-Carlo-Simulation eines linearen Zusammenhangs zwischen Outcomegröße und normalverteilten Merkmalen und Störtermen werden von Zhao (2004) vier Distanzmaße miteinander verglichen: Propensity Score, Mahalanobisdistanz sowie Matching mit gewichteten Merkmalsdifferenzen in zwei Varianten.[10] In der Simulation werden die Stichprobengröße und die Anzahl der berücksichtigten Merkmale sowie die Stärke des Selektionseffekts und die relative Bedeutung der Merkmale für die Teilnahmeentscheidung variiert. Als Qualitätskriterien werden Bias, Standardfehler und MSE betrachtet. Hinsichtlich dieser Kriterien erreicht Propensity Score Matching relativ gute Ergebnisse, wenn die Korrelation zwischen dem Teilnahmeindikator und den berücksichtigten Merkmalen hoch ist. Dieser Effekt ist umso stärker, je größer die Anzahl der Kovariate ist. In kleinen Stichproben sind die Ergebnisse des Propensity Scorse im Vergleich zu den anderen Distanzmaßen relativ schlecht. Die Mahalanobisdistanz ist dagegen ein relativ robustes Distanzmaß unter verschiedenen Stichprobendesigns.

In ihrer Analyse vergleichen Gu und Rosenbaum (1993) verschiedene Distanzmaße (Propensity Score, Mahalanobisdistanz sowie eine zweistufige Distanzmessung) und Zuordnungsprozesse (Greedy Pair Matching und optimale Zuordnungen) miteinander. Dazu werden die Anzahl der Kovariate, die Gruppengröße der zur Verfügung

[10] Die Differenzen der Kovariate werden mit den Koeffizienten der Propensity-Score-Schätzung bzw. mit ihrem geschätzten Effekt auf das potenzielle Einkommen gewichtet.

stehenden Nichtteilnehmer im Vergleich zur Größe der Teilnehmergruppe und die Stärke der Heterogenität der Merkmale in beiden Gruppen variiert. Zur Beurteilung der Ergebnisse werden zwei Kriterien eingesetzt: die durchschnittliche Distanz zwischen einem Teilnehmer und den Nichtteilnehmern in seiner Unter-Kontrollgruppe sowie die Ähnlichkeit der Verteilungen der Merkmale in Teilnehmer- und Kontrollgruppe. Es wird festgestellt, dass für eine kleine relative Zahl der Nichtteilnehmer mit einer optimalen Zuordnung eine geringere durchschnittliche Distanz als mit Greedy Pair Matching erreicht wird. Bei einer großen relativen Nichtteilnehmeranzahl sind dagegen keine wesentlichen Unterschiede feststellbar. In Bezug auf die Ähnlichkeit der Merkmale in Teilnehmer- und Kontrollgruppe ist kein wesentlicher Vorteil des Optimal Matching zu erkennen. Darüber hinaus liefert Propensity Score Matching die beste Lösung bei einer großen Anzahl der Kovariate, das zweistufige Distanzmaß liegt häufig nahe am besten Ergebnis. Full Matching ist hinsichtlich der Gütekriterien besser als die Zuordnung einer festen Anzahl von Partnern, sowohl bei der Erreichung einer möglichst geringen durchschnittlichen Distanz als auch hinsichtlich der Ähnlichkeit der Verteilung der Kovariate. Full Propensity Score Matching liefert bessere Ergebnisse als Full Mahalanobis Matching, vor allem bei großer Anzahl der Kovariate.

Fröhlich (2004a) vergleicht in einer Studie die Performance verschiedener Matchingalgorithmen bei unterschiedlichem Common Support und unterschiedlicher Form des Vergleichs-Outcomes. Folgende Algorithmen werden betrachtet: Pair Matching, Kern Matching, Local Linear Matching mit und ohne Trimming sowie Ridge Matching (in der von Seifert und Gasser (1996; 2000) empfohlenen Form) – jeweils mit Epanechnikov- und Gaußscher Kernfunktion. Zusätzlich wird die Zuordnung einer festen Anzahl von Partnern in zwei Formen berücksichtigt: das Vergleichs-Outcome wird als einfacher Durchschnitt des Outcomes der berücksichtigten Nichtteilnehmer bzw. als gewichteter Durchschnitt ermittelt. Die Gewichtung jedes Nichtteilnehmers ist dabei abhängig vom Abstand zum Teilnehmer.

In der Simulation werden das Größenverhältnis von Teilnehmer- und Nichtteilnehmergruppe, die Größe des Common-Support-Bereichs (die Lage der Dichtefunktionen des Propensity Scores zueinander) sowie die Regressionsgleichung zur Schätzung des Vergleichs-Outcomes (insbesondere der Grad und die Lage von Nicht-

linearitäten) variiert. Der Propensity Score ist dabei eine lineare Funktion eines Merkmals. Zur Beurteilung der Matchingergebnisse wird der mittlere quadratische Fehler (MSE) des geschätzten Vergleichs-Outcomes genutzt. Als Benchmark wird Pair Matching verwendet, der MSE aller anderen Algorithmen wird in Relation zu dem des Pair Matching angegeben.

Ein wichtiges Ergebnis dieser Studie ist die Feststellung, dass die Form der Regressionsgleichung (und damit Lage und Ausmaß der Nichtlinearität) die Qualität des Matchingergebnisses nicht bzw. nur in geringem Maße beeinflusst.

Bei Annahme der optimalen Bandbreite für die o.g. Schätzer ergeben sich folgende Beobachtungen: Ridge Matching hat den kleinsten MSE in allen Designs. Damit kann die größte Verbesserung gegenüber Pair Matching erreicht werden. Die Performance von Local Linear Matching ist sehr sensitiv gegenüber Veränderungen des Designs. Kern Matching ist weniger designabhängig, aber immer schlechter als Ridge Matching. Die Art der Kernfunktion ist dabei von geringer Bedeutung. Der MSE der Zuordnung einer festen Anzahl von Partnern ist immer größer als der von Ridge Matching, dabei ist der gewichtete Durchschnitt schlechter als die einfache Durchschnittsermittlung.

Wird die Bandbreite mit Hilfe des Cross-Validation-Verfahrens ermittelt, ist die Performance der meisten Schätzer abhängig vom Design: Kern Matching, Local Linear Matching und die Zuordnung einer festen Anzahl von Partnern verschlechtern sich bei Einschränkung des Common-Support-Bereichs. Welches die beste Trimmingregel für Local Linear Matching ist, hängt ebenfalls vom Design ab. Es bleibt unklar, welche Regel für die Praxis empfehlenswert ist.
Ridge Matching ist dagegen relativ unempfindlich gegenüber dem Design.

Die Veränderung des Verhältnisses der Teilnehmeranzahl zur Nichtteilnehmeranzahl beeinflusst die Performance der Schätzer relativ zum Pair Matching. Wenn die Anzahl der Nichtteilnehmer viel größer ist, sind alle Schätzer besser; ist die Anzahl der Teilnehmer größer, liefert nur Ridge Matching bessere Ergebnisse.

Mit wachsender Größe der Stichprobe verbessert sich die Performance aller Schätzer.

3.3.3 Gütemaße zur Beurteilung der Ergebnisse

In den vorgestellten Vergleichsstudien werden verschiedene Gütemaße zur Beurteilung der Matchingverfahren eingesetzt. Diesen Maßen liegen unterschiedliche Kriterien der Qualitätsmessung zugrunde, die verschiedene Anforderungen an die Matchingergebnisse widerspiegeln. Die wichtigsten Gütemaße und die mit ihnen möglichen Aussagen werden in diesem Abschnitt kurz erläutert.

Der mittlere quadratische Fehler Häufig wird, wie in Fröhlich (2004a), die Güte des Schätzergebnisses anhand des mittleren quadratischen Fehlers (MSE) überprüft. Dieses Effizienzmaß findet Anwendung, wenn der „wahre“ Maßnahmeeffekt bekannt ist, bspw. aus experimentellen Datensätzen oder Simulationen. Der MSE setzt sich aus der Abweichung der einzelnen Schätzwerte von ihrem Mittelwert, also der Varianz, und der Abweichung des Mittelwertes vom „wahren“ Wert, dem Bias, zusammen (Bleymüller, Gehlert und Gülicher, 1996):

$$\begin{aligned} MSE(\widehat{ME}) &:= E[(\widehat{ME} - ME)^2] \\ &= E[\widehat{ME} - E(\widehat{ME})]^2 + [E(\widehat{ME}) - ME]^2 \\ &= Var(\widehat{ME}) + [B(\widehat{ME})]^2. \end{aligned} \tag{3.1}$$

Dabei bezeichnet ME den „wahren“ Maßnahmeeffekt, $\widehat{ME}$ eine Schätzung des Maßnahmeeffekts. $E(.)$ steht für den Erwartungswert, Var für die Varianz und B für den Bias des Maßnahmeeffekt-Schätzers.

Allgemein bedeutet ein geringer MSE, dass der Schätzer im Mittel in der Nähe des „wahren“ Maßnahmeeffekts liegt und die Schätzwerte relativ gering streuen.

Die prozentuale Reduzierung des Bias Ein Gütemaß, mit dem nicht das Matchingergebnis, sondern die Ähnlichkeit der Verteilungen der Variablen in Teilnehmer- und Nichtteilnehmergruppe überprüft wird, wird als prozentuale Reduzierung des Bias (Percent Bias Reduction) bezeichnet.[11] Die Bezeichnung Bias ist hier auf die

[11] Dieses Maß wird von Cochran (1968) in die Literatur eingeführt und in der Folgezeit häufig zur Begutachtung verschiedener Matchingalgorithmen verwendet, bspw. in Cochran und Rubin (1973), Rubin (1973). Die folgende Darstellung orientiert sich an Ming und Rosenbaum (2000).

Abweichung der Mittelwerte der Merkmale in beiden Gruppen bezogen. Die Reduzierung des Bias einer Variable wird anhand eines Vorher-Nachher-Vergleichs der Summe der Differenzen der Variablenmittelwerte festgestellt. Zur Ermittlung des Bias $B_{vor}(x_n)$ vor dem Matching wird die Differenz der Durchschnittswerte dieses Merkmals in der Teilnehmer- und der Nichtteilnehmergruppe betrachtet:

$$\begin{aligned} B_{vor}(x_n) &= \frac{1}{I}\sum_{i\in T} x_{ni} - \frac{1}{J}\sum_{j\in NT} x_{nj} \\ &= \bar{x}_n^T - \bar{x}_n^{NT}. \end{aligned} \tag{3.2}$$

Dabei bezeichnet I die Anzahl der Teilnehmer, J die Anzahl der Nichtteilnehmer, x_{ni} steht für die Ausprägung des Merkmals x_n bei Teilnehmer i, x_{nj} für die des Nichtteilnehmers j, $\bar{x}_n^T$ und $\bar{x}_n^{NT}$ geben die jeweiligen Durchschnittswerte in beiden Gruppen an.

Der Bias $B_{nach}(x_n)$ nach dem Matching ergibt sich aus der Summe der Differenzen, die jeweils zwischen der Ausprägung beim Teilnehmer und dem Durchschnittswert der ihm zugeordneten Nichtteilnehmer auftreten:

$$B_{nach}(x_n) = \frac{1}{I}\sum_{i\in T}\left(x_{ni} - \frac{1}{J^{C_i}}\sum_{j\in C_i} x_{nj}\right). \tag{3.3}$$

Dabei bezeichnet der Index C_i die individuelle Kontrollgruppe des Teilnehmers i, und J^{C_i} die entsprechende Anzahl der Kontrollgruppenmitglieder.

Die prozentuale Reduzierung des Bias B_{red} durch Matching erhält man durch die Verknüpfung der beiden Maße:

$$B_{red}(x_n) = 100 \cdot \left(1 - \left|\frac{B_{nach}}{B_{vor}}\right|\right). \tag{3.4}$$

Die standardisierte Differenz der Variablen Ein sehr ähnliches Maß für die Angleichung der Verteilungen ist die standardisierte Differenz der beim Matching verwendeten Variablen. Üblich ist auch hier ein Vergleich der Differenzen vor und nach dem Matching.[12] Die standardisierte Differenz $sd_{vor}(x_n)$ einer Variable vor

[12]Die Darstellung der standardisierten Differenzen orientiert sich an Rosenbaum und Rubin (1985). Für weitere Anwendungen vgl. bspw. Hujer und Thomsen (2006) oder Sianesi (2004).

Matching berechnet sich aus der Differenz der Stichprobendurchschnitte der gesamten Teilnehmergruppe $\bar{x}_n^T$ und der gesamten Nichtteilnehmergruppe $\bar{x}_n^{NT}$ im Verhältnis zur durchschnittlichen Standardabweichung beider Teilstichproben:

$$sd_{vor}(x_n) = 100 \cdot \frac{\bar{x}_n^T - \bar{x}_n^{NT}}{\sqrt{\frac{s^2(x_n^T)+s^2(x_n^{NT})}{2}}}. \tag{3.5}$$

Dabei bezeichnen $\bar{x}_n^T$ und $\bar{x}_n^{NT}$ die jeweiligen Stichprobenmittelwerte und s^2 die Stichprobenvarianz.

Zur Ermittlung der standardisierten Differenz nach Matching wird der Zähler der Gleichung verändert. Die Mittelwertdifferenz eines Merkmals wird bestimmt zwischen der Teilnehmergruppe $\bar{x}_n^T$ und der Kontrollgruppe $\bar{x}_n^C$. Der Nenner bleibt unverändert:

$$sd_{nach}(x_n) = 100 \cdot \frac{\bar{x}_n^T - \bar{x}_n^C}{\sqrt{\frac{s^2(x_n^T)+s^2(x_n^{NT})}{2}}}. \tag{3.6}$$

Von Smith und Todd (2005b) wird kritisiert, dass sich kein statistisch gesichertes Kriterium für die zulässige Größe der nach dem Matching verbleibenden Differenzen finden lässt.

Alternativ zu diesem Maß werden deshalb in verschiedenen Studien t-Tests zur Überprüfung der Übereinstimmung der Mittelwerte der einzelnen Matchingvariablen (Augurzky und Kluve, 2004) bzw. Hotelling-T^2-Tests zur gemeinsamen Prüfung aller Variablen durchgeführt (Smith und Todd, 2005b).

3.4 Zusammenfassung

Die Analyse der Vor- und Nachteile der einzelnen Distanzmaße und Zuordnungsverfahren nimmt in der jüngeren Literatur immer breiteren Raum ein. In diesem Kapitel werden die wichtigsten Studien vorgestellt, die auf verschiedenen Wegen versuchen, die Wahl eines für eine konkrete Ausgangssituation am besten geeigneten Matchingverfahrens zu erleichtern.

In den beiden letztgenannten Quellen wird jeweils der Median aller Differenzen vor und nach dem Matching ausgewiesen.

Zusammenfassend für die zitierten Studien lassen sich folgende allgemeine Erkenntnisse festhalten:

- Für die Anwendung von Matchingverfahren sind umfangreiche Daten über die zu vergleichenden Personen nötig. Sind solche Informationen verfügbar, sind Matchingverfahren besser als parametrische und andere nichtparametrische Verfahren in der Lage, das Selektionsproblem zu lösen.
- Allerdings gibt es keinen Matchingprozess, der in jeder Situation den anderen überlegen ist. Die Wahl eines geeigneten Algorithmus ist vielmehr abhängig von den zur Verfügung stehenden Daten.
- Eine Kombination von Matching mit dem Differenz-von-Differenzen-Verfahren erscheint sinnvoll, um evtl. nach dem Matching noch verbleibende unbeobachtbare Heterogenitäten zu beseitigen.

Für die Auswahl geeigneter Distanzmaße geben die zitierten Studien unterschiedliche Hinweise. So stellt Zhao (2004) fest, dass der MSE von Schätzern auf Basis des Propensity Scores in kleinen Stichproben größer ist als der von exaktem Matching. Die Mahalanobisdistanz wird hier als besseres, gegenüber verschiedenen Stichprobendesigns relativ robustes Maß angesehen. Demgegenüber stellen Gu und Rosenbaum (1993) fest, dass der Propensity Score bei einer großen Anzahl der Kovariate das beste Distanzmaß darstellt.

In Bezug auf Zuordnungsprozesse wird in Gu und Rosenbaum (1993) Full Matching der Zuordnung einer festen Anzahl von Partnern für die Konstruktion möglichst ähnlicher Gruppen (gemessen an der Verteilung der Merkmale in beiden Gruppen) vorgezogen. Fröhlich (2004a) stellt in seiner Studie fest, dass Ridge Matching hinsichtlich des MSE besser als alle überprüften Alternativen ist. In beiden Studien wird die Sensitivität der analysierten Algorithmen gegenüber der Relation zwischen Teilnehmer- und Nichtteilnehmeranzahl betont: Je mehr Nichtteilnehmer pro Teilnehmer zur Verfügung stehen, desto ähnlicher sind die Kontrollgruppen den Teilnehmern – bei jedem Zuordnungsalgorithmus.
Bei kleinem Common-Support-Bereich oder relativ kleiner Nichtteilnehmergruppe im Verhältnis zur Teilnehmeranzahl werden in Dehejia und Wahba (2002) Verfahren

mit Zurücklegen empfohlen, bei großem Common-Support-Bereich können dagegen Zuordnungsverfahren ohne Zurücklegen angewendet werden.

Aufbauend auf den Ergebnissen dieser Studien und den theoretischen Erkenntnissen des vorangegangenen Kapitels wird im Folgenden eine eigene Simulationsstudie durchgeführt. Dabei stehen kleine Stichproben im Mittelpunkt der Analyse. In der Literatur findet sich keine einheitliche Aussage darüber, aus wie vielen Personen eine kleine Stichprobe besteht. Betrachtet man die zitierten Studien, scheint eine Teilnehmerstichprobe von 100 Personen die Grenze zwischen kleinen und größeren Stichproben zu markieren.[13]

[13] In der Studie von Fröhlich (2004a) besteht die kleinste Teilnehmerstichprobe aus 100 Personen, ebenso wie in Zhao (2004). Eine strengere Definition von „klein“ findet sich lediglich in Gu und Rosenbaum (1993) mit einer Teilnehmerstichprobe der Größe 50.

Kapitel 4

Simulation zum Vergleich verschiedener Matchingverfahren

Die Analyse im folgenden Teil der Arbeit leistet einen Beitrag zur Entwicklung von Entscheidungshilfen bzw. allgemeinen Regeln für die Auswahl eines geeigneten Matchingverfahrens auf Grundlage des Vergleichs verschiedener Verfahren. Der Fokus der Studie liegt dabei auf kleinen Stichproben. Entsprechend der Zweistufigkeit des Entscheidungsprozesses ist die Untersuchung in zwei Schritte gegliedert: im ersten Schritt werden Distanzmaße untersucht, im zweiten Schritt erfolgt die Analyse verschiedener Zuordnungsprozesse. Dabei werden die Ergebnisse früherer Studien einbezogen, indem diejenigen Distanzmaße und Zuordnungsalgorithmen verglichen werden, die sich in diesen Studien als vorteilhaft erwiesen haben.

Der wichtigste Unterschied zu den bisher in der Literatur zu findenden Studien liegt in der expliziten Berücksichtigung des unterschiedlichen Skalenniveaus der verwendeten Variablen. Damit soll eine größere Nähe zu real vorzufindenden Entscheidungssituationen erreicht werden. In der Simulation wird ein realer Datensatz nachgebildet, um die Anforderungen, die an Matchingverfahren in der Praxis gestellt werden, möglichst gut abbilden zu können. Das hat zum einen Auswirkungen auf das gewählte Distanzmaß, zum anderen auf die einsetzbaren Tests zur Überprüfung der Matchingergebnisse.

In der Simulation werden der Propensity Score, der Indexscore und die Mahalanobisdistanz verglichen, da sich diese Maße in früheren Studien als vorteilhaft gegenüber anderen Distanzmaßen erwiesen haben. Zusätzlich zu diesen in anderen Studien bereits analysierten Maßen werden aggregierte Distanzmaße eingeführt, die die adäquate Berücksichtigung metrischer und nominaler Variablen ermöglichen. Die gewichtete Mahalanobis-Matching-Distanz und der Ähnlichkeitskoeffizient von Gower sollten besser als die bisher verwendeten Distanzmaße in der Lage sein, in der Gesamtdistanz den Beitrag jeder einzelnen Variable unter Berücksichtigung des spezifischen Skalenniveaus adäquat zu berücksichtigen.

Die Wahl der zu vergleichenden Zuordnungsprozesse ergibt sich ebenfalls aus den Ergebnissen früherer Studien. Ridge Matching (Local Linear Regression in der von Fröhlich (2004a) empfohlenen Spezifikation), Optimal Full Matching sowie die Zuordnung mit Zurücklegen werden in der Literatur empfohlen. Zusätzlich wird ein optimales Nearest Neighbor Matching betrachtet. Damit wird angestrebt, den Vorteil von Nearest Neighbor Matching (eine ähnlichere Verteilung der Merkmale in

Teilnehmer- und Nichtteilnehmergruppe) auszunutzen und gleichzeitig den Nachteil des bisher häufig verwendeten Random Matchings (der Verlust von Beobachtungen, wenn kein passender Partner mehr gefunden werden kann) auszugleichen.

Für den Vergleich der Distanzmaße und Zuordnungsalgorithmen wird ein Simulationsdesign gewählt, in dem die Struktur der Merkmale, die relative Anzahl der Teilnehmer und Nichtteilnehmer, der Common-Support-Bereich sowie die Stichprobengröße insgesamt variiert werden können.

4.1 Hypothesen

Eine der Herausforderungen bei der Anwendung von Matchingverfahren besteht in der adäquaten Berücksichtigung unterschiedlich skalierter Merkmale. In der Literatur finden sich dazu verschiedene Möglichkeiten, die im Abschnitt 2.3.1 vorgestellt werden.

Das am einfachsten zu handhabende (und daher sehr häufig angewandte) Distanzmaß ist der Propensity Score. Der daraus abgeleitete lineare Index Score ist – insbesondere für Personen mit einer Teilnahmeneigung nahe Null oder Eins – trennschärfer und daher besser in der Lage, auch in den Randbereichen der Dichteverteilung ähnliche Personen zu erkennen (Lechner, 1998, S. 115).[1]

Ein Nachteil beider Maße, der besonders in kleinen Stichproben zum Tragen kommt, ist die implizite Gewichtung der Merkmale nach ihrer Bedeutung für die Teilnahmeentscheidung (Zhao, 2004, S. 94). Aus dieser Gewichtung erwächst das von Fröhlich (2004b) beschriebene Problem der möglicherweise unvollständigen Angleichung der Verteilungen der Determinanten der Einkommenshöhe zwischen Teilnehmern und Nichtteilnehmern trotz individuell identischer Propensity Scores. Diese ungleiche Merkmalsverteilung bedeutet eine Verletzung der Annahme bedingter Unabhängigkeit und führt zu einer Verzerrung des geschätzten Maßnahmeeffekts.

Im Vergleich zum Propensity Score und dem Index Score sollte mit der Berücksichtigung jedes einzelnen relevanten Merkmals die Bedeutung der Variablen für das

[1] Dieser Unterschied soll allerdings in der Simulation nicht näher untersucht werden.

Einkommen besser abgebildet werden können. Zur Berücksichtigung aller Merkmale – unabhängig von ihrem Skalenniveau – wird u.a. in Zhao (2004) die Mahalanobisdistanz eingesetzt. Diese hat den Vorteil, dass durch die Berücksichtigung der Varianz-Kovarianz-Matrix „automatisch“ eine Normierung der einbezogenen Variablen erfolgt. Die Varianz-Kovarianz-Matrix dient gleichzeitig der Gewichtung der Merkmale. Ein Nachteil dieses Distanzmaßes ist die nicht adäquate Berücksichtigung nicht metrisch skalierter Merkmale.[2] Um diesem Problem zu begegnen, werden in der vorliegenden Arbeit zusätzlich zwei aggregierte Distanzmaße verwendet.

Aus den angestellten Überlegungen ergeben sich folgende Hypothesen hinsichtlich der Wahl des Distanzmaßes:

- Die Ähnlichkeit der Personen bezüglich der Determinanten einer Outcome-Größe kann mit der Mahalanobisdistanz und den aggregierten Distanzmaßen besser als mit Hilfe des Propensity Scores wiedergegeben werden.
- Die aggregierten Distanzmaße sind besser als die Mahalanobisdistanz in der Lage, nominale Variablen ihrer Bedeutung entsprechend zu gewichten.

Wie stark sich die Wahl eines Zuordnungsprozesses auf das Matchingergebnis auswirkt, ist abhängig von der Größe der Übereinstimmung der untersuchten Gruppen und der Anzahl der zur Verfügung stehenden Nichtteilnehmer pro Teilnehmer. Sind beide relativ groß, sollten alle Algorithmen annähernd gleiche Ergebnisse liefern. Je kleiner der Common-Support-Bereich und je geringer die relative Anzahl der Nichtteilnehmer jedoch wird, desto größer ist die Herausforderung, die an den Zuordnungsprozess gestellt wird.

In der Literatur werden drei unterschiedliche Ansätze, dieser Herausforderung zu begegnen, favorisiert. Die Zuordnung mit Zurücklegen wird die größte Ähnlichkeit der beiden Teilstichproben erreichen, da die mehrfache Zuordnung eines Nichtteilnehmers möglich ist, wenn dieser der ähnlichste Partner für verschiedene Teilnehmer ist. Allerdings besteht die Vergleichsgruppe evtl. nur aus sehr wenigen Personen, was

[2] Dies hat insbesondere Auswirkungen auf die Behandlung von Variablen, bei denen bestimmte Ausprägungen in der Stichprobe nur selten auftreten, und Variablen, für die in der Stichprobe „Ausreißer“ beobachtet werden (Gu und Rosenbaum, 1993, S. 419).

die Varianz des Schätzers erheblich vergrößern kann. Das „Gegenstück“ dazu bildet das Full Matching, bei dem alle verfügbaren Nichtteilnehmer unter den Teilnehmern aufgeteilt werden. Eine Mehrfachnutzung ist dabei nicht möglich. Dadurch entsteht eine relativ große Vergleichsgruppe, die aber im Vergleich zum Ziehen mit Zurücklegen unähnlicher sein wird, woraus ein größerer Bias des Ergebnisses resultieren kann.
Gemeinsamkeiten mit beiden Verfahren hat das Ridge Matching, bei dem das Vergleichseinkommen aus dem Einkommen mehrerer Nichtteilnehmer gebildet wird und jeder Nichtteilnehmer mehrfach zugeordnet werden kann. Diese Mehrfachnutzung von Nichtteilnehmern stellt v.a. in Stichproben mit annähernd gleich großer Teilnehmer- und Nichtteilnehmeranzahl einen Vorteil gegenüber Zuordnungsverfahren ohne Zurücklegen dar, weil damit der Verlust von Beobachtungen aufgrund fehlender Partner vermieden werden kann.

Zusätzlich zu den in der Literatur zu findenden Ansätzen werden zwei Nearest Neighbor Matchingverfahren ohne Zurücklegen in die Analyse einbezogen. Es wird vermutet, dass ein optimales Nearest Neighbor Matching auf Grundlage des Ungarischen Algorithmus besser als Random Matching in der Lage ist, den Teilnehmern aus den verfügbaren Nichtteilnehmern die passenden Partner zuzuordnen. Daraus sollte zum einen ein geringerer Verlust von Beobachtungen und zum anderen eine ähnlichere Kontrollgruppe resultieren. Die Vermutung über die Ähnlichkeit der Kontrollgruppe konnte in Gu und Rosenbaum (1993) allerdings nicht bestätigt werden. Als mögliche Begründung dafür wird in der zitierten Studie angeführt, dass Random Matching und Optimal Nearest Neighbor Matching tendenziell die gleichen Nichtteilnehmer auswählen.[3]

Aus diesen Überlegungen und den Ergebnissen der im Kapitel 3.3 vorgestellten Studien lassen sich die folgenden Hypothesen hinsichtlich der Wahl des Zuordnungsalgorithmus für die Analyse ableiten:

- Wenn Teilnehmer- und Nichtteilnehmerstichproben sich nur geringfügig unterscheiden, sind die Ergebnisse aller Zuordnungsverfahren sehr ähnlich.

[3]Die Nichtteilnehmer werden von beiden Verfahren unterschiedlichen Teilnehmern zugeordnet. Die entstehende Kontrollgruppe insgesamt ist aber in beiden Verfahren mehr oder weniger gleich (Gu und Rosenbaum, 1993, S. 413).

- Bei der Zuordnung mit Zurücklegen ist die Abweichung zwischen dem „wahren“ und dem geschätzten Maßnahmeeffekt geringer als bei anderen Zuordnungsverfahren, dafür ist die Streuung des Ergebnisses größer.

- Beim Optimal Full Matching ist die Streuung des Maßnahmeeffekts geringer als bei anderen Zuordnungsverfahren, allerdings ist die Abweichung zwischen dem „wahren“ und dem geschätzten Maßnahmeeffekt größer.

- Durch eine optimale Nearest Neighbor Zuordnung wird der Verlust von Teilnehmern aufgrund einer suboptimalen Zuordnung vermieden. Der Unterschied zum Random Matching wird umso deutlicher, je kleiner die verfügbare Nichtteilnehmergruppe wird und je unähnlicher die darin enthaltenen Nichtteilnehmer den Teilnehmern sind.

- Durch die Anwendung von Ridge Matching und Zuordnungen mit Zurücklegen auf Stichproben mit nahezu gleich großer Teilnehmer- und Nichtteilnehmeranzahl kann der Verlust von Teilnehmern aufgrund fehlender Partner vermieden werden.

4.2 Untersuchungsdesign

Die aufgestellten Hypopthesen werden mit Hilfe einer Simulation überprüft. Die Ausgestaltung des Simulationsdesigns orientiert sich an theoretischen Überlegungen und bisher in der Literatur zu findenden Studien.

Als wichtigste Anforderungen an ein Distanzmaß lassen sich die folgenden formulieren: Ein Distanzmaß sollte zum einen Übereinstimmungen und Unterschiede in den betrachteten Merkmalen so exakt wie möglich wiedergeben und zum anderen jedes Merkmal unabhängig vom Skalenniveau gleich gewichten, wenn keine andere Gewichtung, z.B. entsprechend der Bedeutung, angestrebt wird. Um zu überprüfen, wie die einzelnen Distanzmaße diesem Anspruch gerecht werden, wird die Variablenstruktur in der Simulation variiert. Alle anderen Größen (Verhältnis der Anzahl der Teilnehmer zu den Nichtteilnehmern sowie Größe der Stichproben) bleiben dabei unverändert.

Zur Prüfung der Hypothesen über die Zuordnungsalgorithmen spielen zwei Größen eine Rolle: die Relation der Anzahl der Teilnehmer und Nichtteilnehmer sowie die Größe des Common-Support-Bereichs. Beide Größen werden in der Simulation variiert. Mit der Verringerung der Anzahl Nichtteilnehmer je Teilnehmer sowie der Einschränkung des Common-Support-Bereichs wird der wachsende Anspruch an die Zuordnungsverfahren durch die Verringerung der zur Verfügung stehenden Wahlmöglichkeiten verdeutlicht.

Die Veränderung der Stichprobengröße insgesamt wird motiviert durch die in der Literatur zu findende Aussage, dass sich die Matchingergebnisse mit zunehmender Stichprobengröße verbessern.

In der Simulation werden verschiedene Stichproben generiert, die sich in der Anzahl der Personen, der Art der Übereinstimmung der Variablen sowie der Größe des Common-Support-Bereichs unterscheiden.

4.2.1 Stichprobendesign

Für jede Stichprobe werden Variablen mit unterschiedlichen Skalenniveaus konstruiert. Die Ausprägung der einzelnen Variablen orientiert sich an einem realen Datensatz, um einen möglichst engen Bezug zur praktischen Anwendung herzustellen.

Als Grundlage dient die amtliche Repräsentativstatistik über die Bevölkerung und den Arbeitsmarkt (Mikrozensus) des Statistischen Bundesamtes Deutschland. Er basiert auf der Befragung einer 1%-Stichprobe aller Haushalte in Deutschland und liefert Informationen zu allen in den befragten Haushalten lebenden Personen. Neben demografischen Angaben enthält er Informationen über das allgemeine Ausbildungsniveau, berufliche Qualifikationen, Weiterbildungsaktivitäten, die aktuelle Erwerbsbeteiligung und evtl. frühere Beschäftigungen der Befragten. Darüber hinaus werden Aussagen zum Familien- und Haushaltszusammenhang und zur wirtschaftliche Lage der einzelnen Personen und des Haushalts insgesamt erfasst. Die Mikrozensusergebnisse bieten statistische Informationen, auf deren Basis politische Entscheidungen in der Bundesrepublik und der EU getroffen werden und bilden

gleichzeitig die Grundlage für die laufende Arbeitsmarkt- und Berufsforschung.[4] Als Vorbild für die Stichproben der Simulation dient die Unterstichprobe der Personen im Alter zwischen 25 und 55 Jahren des Mikrozensus 2004.[5]

Die Angaben in diesem Datensatz treten in unterschiedlichen Skalenniveaus auf. Als metrische Variablen liegen bspw. Informationen über Alter, Ausbildungsdauer und Einkommenshöhe vor, dichotome nominale Variablen beinhalten bspw. Angaben über Geschlecht, Staatsangehörigkeit oder Erwerbstätigkeit. Eine dritte Gruppe von Informationen bilden ordinale und polytome nominale Variablen, z.B. Ausbildungsniveau, Familienstand oder Beschäftigungstyp. Diese Gruppe dient als Orientierung für die Bildung polytomer nominaler Variablen. Ordinale Variablen werden für die Simulation nicht generiert, da sie für die Ermittlung von Ähnlichkeiten zwischen Objekten in dichotome nominale Variablen umgewandelt werden.

Für die Stichproben werden jeweils 5 metrische, 5 dichotome und 5 polytome nominale Variablen generiert.[6] Für alle metrischen Variablen wird Normalverteilung unterstellt. Die Mittelwerte und Standardabweichungen, die der Zufallsziehung zugrunde gelegt werden, orientieren sich an folgenden Merkmalen im Mikrozensus: Alter, Kinderanzahl, Ausbildungsdauer, Dauer der Betriebszugehörigkeit sowie Nettoeinkommen. Dichotome Variablen werden auf Grundlage der Binomialverteilung generiert. Der jeweilige Mittelwert orientiert sich an den Angaben über Geschlecht, Familienstand „verheiratet“, deutsche Staatsbürgerschaft, Arbeit im öffentlichen Dienst sowie Wohnort in den Neuen Bundesländern. Für die Generierung polytomer nominaler Merkmale wird die Häufigkeit des Auftretens verschiedener Ausprägungen der jeweiligen Variablen vorgegeben. Als Vorbild dafür dient die beobachtete Häufigkeit jeweils ausgewählter nominaler oder ordinaler Merkmale: Schulbildungsniveau, Ausbildungsniveau, Beschäftigungstyp, Betriebsgröße (jeweils in 4 Kategorien) und

[4] Die Daten werden bspw. für Entscheidungen über die Mittelverteilung des EU-Regional- und Sozialfonds herangezogen und fließen in die Jahresgutachten des Sachverständigenrates zur Begutachtung der gesamtwirtschaftlichen Entwicklung ein.

[5] Diese Einschränkung der Altersstruktur entspricht dem üblichen Vorgehen bei Arbeitsmarktanalysen.

[6] Die Auswahl der Variablen orientiert sich an den in empirischen Studien zum Arbeitsmarkt häufig verwendeten Informationen.

Wirtschaftszweige (in 3 Kategorien), in denen die Befragten beschäftigt sind. Alle Variablen werden aus univariaten Verteilungen gezogen.[7]

Darüber hinaus werden verschiedene Linearkombinationen dieser Variablen definiert. Als Orientierung dabei dient das Vorgehen bei der Evaluation einer wirtschaftspolitischen Maßnahme. Denkbar ist bspw. die Beurteilung einer Weiterbildungsmaßnahme anhand ihres Einkommenseffektes für die Teilnehmer. Dafür sind zusätzliche Variablen nötig: der Maßnahmeeffekt, das Einkommen nach Teilnahme und das Nichtteilnahmeeinkommen. Die Spezifikation der Einkommen orientiert sich an der Einkommensschätzung des Sachverständigenrates zur Begutachtung der gesamtwirtschaftlichen Entwicklung im Jahresgutachten 2004/2005.[8] Der Unterschied zwischen Teilnahme- und Nichtteilnahmeeinkommen wird bestimmt durch einen individuellen Maßnahmeeffekt, der ebenfalls als Linearkombination verschiedener Faktoren definiert wird.[9]

Die für die Analyse verwendeten Stichproben setzen sich jeweils aus einer Teilnehmer- und einer Nichtteilnehmerstichprobe zusammen. Es werden Teilnehmerstichproben unterschiedlicher Größe erzeugt (50 Personen, 100 Personen und 300 Personen), deren Struktur sich in der gesamten Simulation nicht verändert. Alle Strukturveränderungen in den Gesamtstichproben werden durch die Zuordnung unterschiedlicher Nichtteilnehmerstichproben zu einer dieser Teilnehmerstichproben erreicht. Als Kriterium für die Veränderungen wird die Größe des Common-Support-Bereichs verwendet, die je nach gewünschter Übereinstimmung beider Teilstichproben angepasst wird. Bei den metrischen Variablen wird diese Anpassung durch die Abweichung der Merkmalsmittelwerte um eine merkmalsspezifische Konstante erreicht. In dieser

[7] Damit wird die Korrelationsstruktur der Variablen nicht berücksichtigt. Dies bedeutet eine geringfügige Einschränkung der Realitätsnähe der Simulation und könnte Auswirkungen auf das Verhalten und die Performance der analysierten Matchingverfahren haben. Allerdings würde die Abbildung der gemeinsamen Verteilung unterschiedlich skalierter Variablen den Rahmen dieser Arbeit sprengen.

[8] Als Bestimmungsfaktoren werden in der Studie genannt: Alter, Anzahl der Kinder, Ausbildungsdauer, Dauer der Betriebszugehörigkeit, Geschlecht, deutsche Staatsbürgerschaft, Wohnort Neue Bundesländer, Arbeit im öffentlichen Dienst sowie quadratische Terme aus den genannten Faktoren. Vgl. dazu Sachverständigenrat zur Begutachtung der gesamtwirtschaftlichen Entwicklung (2004), Kasten 30. Die Parameter dieser Schätzung werden zur Definition des Einkommens in der generierten Stichprobe verwendet.

[9] Eine Beschreibung der generierten Variablen und ihrer Verwendung zur Definition des Maßnahmeeffekts und der Einkommensgrößen sowie der für Matching eingesetzten Variablen findet sich in der Tabelle B.1 im Anhang.

Konstante werden der gewünschte Umfang der Abweichung und die Streuung der Variable berücksichtigt. Für die nominalen Variablen bestimmt die Konstante den Unterschied in der Häufigkeit des Auftretens der einzelnen Ausprägungen in beiden Teilstichproben.[10] Der gewünschte Umfang der Abweichung wird für sehr ähnliche Nichtteilnehmerstichproben auf 1% der Streuung eines Merkmals, für mittlere Übereinstimmungen auf 10% und für relativ unähnliche Nichtteilnehmerstichproben auf 25% der Streuung festgelegt. Für jede Stichprobe wird nach dem Zufallsprinzip entschieden, ob eine positive oder eine negative Konstante generiert wird.[11]

4.2.2 Simulationsaufbau

In der Simulation werden jeweils 100 Durchläufe mit derselben Konstellation von Teilnehmer- und Nichtteilnehmerstichproben durchgeführt. In jedem Simulationslauf wird dabei eine von 100 Teilnehmerstichproben einer bestimmten Größe mit jeweils einer von 100 Nichtteilnehmerstichproben mit entsprechender Merkmalsausprägung kombiniert.[12]

Die Simulationsstudie setzt sich aus zwei Teilen zusammen, die den beiden grundlegenden Entscheidungen bei der Wahl eines Matchingverfahrens entsprechen. Im ersten Teil der Analyse werden Distanzmaße hinsichtlich ihrer Fähigkeit der Berücksichtigung unterschiedlicher Skalenniveaus für die Ermittlung der Gesamtdistanz verglichen. Den zweiten Teil bildet die Untersuchung der Güte der Matchingergebnisse, die mit verschiedenen Zuordnungsprozessen erzielt werden.

An beide Teiluntersuchungen werden unterschiedliche Anforderungen gestellt, denen mit einem entsprechenden Simulationsdesign und der Auswahl geeigneter Gütemaße

[10] Die Größe der merkmalsspezifischen Konstante orientiert sich jeweils an der Prüfgröße eines statistischen Anpassungstests (der Teststatistik des χ^2-Homogenitätstests für die dichotomen und polytomen nominalen Variablen und derjenigen des t-Tests gleicher Mittelwerte im Falle der metrischen Variablen).

[11] Da für nominale Variablen kein statistisches Streuungsmaß existiert, wird die „Streuung“ dieser Variablen durch eine merkmalsspezifische Abweichung vom Median approximiert. Für metrische Variablen wird die Standardabweichung zur Definition der merkmalsspezifischen Abweichung genutzt. Detaillierte Informationen über die Definition der Unterschiede zwischen Teilnehmer- und Nichtteilnehmerstichprobe finden sich in der Tabelle B.1 im Anhang.

[12] Für die Durchführung der Simulation wird MATLAB 6.5 für Windows genutzt.

Rechnung getragen wird. Zur Analyse der Distanzmaße werden die Teilnehmerstichproben der Größe 100 Personen mit unterschiedlichen Nichtteilnehmerstichproben kombiniert. Das Größenverhältnis bleibt dabei konstant. Für die Untersuchung der Zuordnungsprozesse werden dagegen Teilnehmerstichproben unterschiedlicher Größe mit unterschiedlich großen und unterschiedlich ähnlichen Nichtteilnehmerstichproben kombiniert.

Der eigentlichen Untersuchung geht in beiden Analyseschritten eine Überprüfung der Common-Support-Bedingung für die jeweils verwendeten Stichproben voran. Die Einhaltung dieser Bedingung wird für jedes Merkmal einzeln überprüft. Sie gilt als eingehalten, wenn sich mindestens ein Nichtteilnehmer mit der gleichen Ausprägung des Merkmals wie ein Teilnehmer findet und umgekehrt. Für die metrischen Variablen gilt die Bedingung als eingehalten, wenn die Abweichung einen Toleranzbereich von 10% des merkmalsspezifischen Mittelwertes nicht überschreitet, für die nominalen Variablen gilt nur die exakte Übereinstimmung. Personen, für die diese Bedingung nicht erfüllt ist, können bei der Zuordnung passender Partner nicht berücksichtigt werden und bleiben bei der Ermittlung des Maßnahmeeffekts unberücksichtigt.

Analyse der Distanzmaße

In der Literatur werden drei Distanzmaße empfohlen: der Propensity Score, der Indexscore und die Mahalanobisdistanz. Diese Maße werden im Folgenden miteinander verglichen. Darüber hinaus werden zwei aggregierte Distanzmaße, der Ähnlichkeitskoeffizient von Gower und die gewichtete Mahalanobis-Matching-Distanz, betrachtet.

Alle genannten Distanzmaße werden unter Verwendung derselben Variablen ermittelt. Einbezogen werden dabei dichotome und polytome nominale Variablen sowie metrisch skalierte Merkmale. Interaktionsterme und quadratische Terme werden in der Spezifikation nicht berücksichtigt. Der Propensity Score (2.27) und der Indexscore (2.28) werden mit Hilfe eines Probitmodells geschätzt.[13] Die Mahalanobisdistanz (2.18) wird in der in Abschnitt 2.3.1 beschriebenen Form verwendet. Die

[13] Dazu wird ein Tool von Le Sage (1999) genutzt.

Spezifikation der gewichteten Mahalanobis-Matching-Distanz (2.20) entspricht der Beschreibung im gleichen Abschnitt.

Als weiteres aggregiertes Distanzmaß wird der – ebenfalls im Abschnitt 2.3.1 beschriebene – Ähnlichkeitskoeffizient von Gower (2.23) verwendet. Dazu werden folgende Modifikationen vorgenommen: Das Ähnlichkeitsmaß wird in ein Distanzmaß umgewandelt. Der Nenner der ursprünglichen Gleichung vereinfacht sich, da die Ausprägungen der betrachteten Variablen für alle Personen beobachtbar sind. Die Distanzen in den einzelnen Merkmalen werden mit Eins gewichtet. Damit hat das Distanzmaß nach Gower folgende Form:

$$DG_{ij} = \frac{1}{N} \sum_{n=1}^{N} d_{n,ij}. \tag{4.1}$$

Dabei bezeichnet DG_{ij} das Distanzmaß nach Gower, $d_{n,ij}$ die Distanz zwischen Teilnehmer i und Nichtteilnehmer j im Merkmal x_n und N die Anzahl der Merkmale.

Als skalenspezifische Maße werden die normierte absolute Distanz (für metrische Variablen) und der verallgemeinerte Matchingkoeffizient (für nominale Variablen) angewendet. Für die einzelnen Distanzen gilt:

$$d_{n,ij} = \begin{cases} \frac{|x_{ni} - x_{nj}|}{diff_{max_n}} & \text{wenn } x_n \text{ metrisch} \\ 1 - gMC_{n,ij} & \text{wenn } x_n \text{ nominal.} \end{cases}$$

Dabei gibt $|x_{ni} - x_{nj}|$ die absolute Differenz zwischen Teilnehmer i und Nichtteilnehmer j im betrachteten Merkmal x_n, x_{ni} und x_{nj} jeweils die Ausprägung des betrachteten Mekmals und $diff_{max_n}$ die maximale Differenz in diesem Merkmal an. Der verallgemeinerte Matchingkoeffizient wird mit $gMC_{n,ij}$ bezeichnet.[14]

Simulationsdesign für die Analyse Zur Analyse der Distanzmaße werden vier verschiedene Stichprobendesigns eingesetzt, die sich in der Stärke und der Art der

[14] Bei einzelner Betrachtung jeder nominalen Variable entspricht der verallgemeinerte Matchingkoeffizient einem Indikator für die Übereinstimmung der Ausprägungen:

$$gMC_{n,ij} = \begin{cases} 1 & \text{wenn } x_{ni} = x_{nj} \\ 0 & \text{sonst.} \end{cases}$$

.

Abweichung der Merkmalsausprägungen der Nichtteilnehmerstichproben von den Teilnehmerstichproben unterscheiden. Die jeweilige Veränderung der Übereinstimmung wird durch die Variation der Größe des Common-Support-Bereichs für die entsprechenden Variablen erreicht. Den Ausgangspunkt bilden Stichproben mit nahezu identischen Merkmalsverteilungen aller Skalenniveaus in beiden Teilstichproben. In diesem Design liegt die Abweichung der Merkmalsmittelwerte bzw. der Häufigkeit des Auftretens einer Merkmalsausprägung zwischen Teilnehmer- und Nichtteilnehmerstichprobe bei durchschnittlich 1% der merkmalsspezifischen Streuung. Diese Übereinstimmung wird zunächst nur für die metrisch skalierten Merkmale, danach nur für die nominalen (dichotomen und polytomen) Merkmale eingeschränkt, die Abweichung der Merkmalsmittelwerte bzw. Häufigkeitsverteilungen in der Nichtteilnehmerstichprobe der jeweiligen Variablen liegt dann bei durchschnittlich 25% der merkmalsspezifischen Streuung. Den letzten Schritt der Analyse bilden Teilnehmer- und Nichtteilnehmerstichproben, in denen die Merkmale aller Skalenniveaus relativ stark – durchschnittlich 25% der merkmalsspezifischen Streuung – voneinander abweichen.

Da in diesem Schritt nur die Distanzmaße von Interesse sind, wird für alle Distanzmaße der gleiche Zuordnungsalgorithmus (optimales Nearest Neighbor Matching mit dem Ungarischen Algorithmus) verwendet. Die Stichprobengröße und das Verhältnis zwischen Teilnehmer- und Nichtteilnehmeranzahl bleiben in allen Simulationsläufen konstant. Es werden jeweils Teilnehmerstichproben von 100 Personen und Nichtteilnehmerstichproben von 1000 Personen generiert.

Gütemaße zur Beurteilung der Distanzmaße Für die Verzerrung von Schätzergebnissen sind zwei mögliche Quellen zu berücksichtigen. Zum einen treten Schätzfehler aufgrund der ungenügenden Angleichung der Verteilungen der Merkmale in beiden Teilstichproben auf, zum anderen kann der Verlust von Beobachtungen zur Verzerrung des Ergebnisses führen. Während die erstgenannte Fehlerquelle – eine Verletzung der Annahme bedingter Unabhängigkeit – in jeder Situation zu Verzerrungen führen kann, ist die zweite nur bedeutsam beim Auftreten heterogener Maßnahmeeffekte, da dann der Verlust von Beobachtungen gleichbedeutend mit einem Informationsverlust ist. Beide Fehlerquellen müssen bei der Beurteilung der

Matchingergebnisse berücksichtigt werden. Bei der Beurteilung von Distanzmaßen kann allerdings nur die erste Fehlerquelle betrachtet werden.[15]

Distanzmaße haben die Aufgabe, Unterschiede bzw. Ähnlichkeiten zwischen Teilnehmern und Nichtteilnehmern deutlich zu machen. Beim Einsatz ein und desselben Zuordnungsalgorithmus' entspricht die Frage nach der Güte eines Distanzmaßes also der Frage nach der Angleichung der Merkmalsverteilungen. Im Abschnitt 3.3.3 werden die in der Literatur häufig angewandten Gütemaße zur Beantwortung dieser Frage dargestellt. Allerdings muss festgestellt werden, dass keins dieser Maße eine statistisch gesicherte Aussage über Eignung oder Nichteignung eines Matchingverfahrens zulässt. Für Bias Reduction und standardisierte Differenz ist keine Aussage über die zulässige Größe der nach dem Matching verbleibenden Mittelwertdifferenzen möglich. Für den t-Test und den Hotelling-T^2-Test ist zweifelhaft, ob die impliziten Annahmen (unabhängige Stichproben und normalverteilte Merkmale) in einer empirischen Analyse erfüllt sind.

Zur Beurteilung der Güte der Matchingergebnisse werden in dieser Analyse deshalb skalenspezifische parameterfreie Tests für verbundene Stichproben verwendet: für metrische Variablen der Wilcoxon-Vorzeichen-Rangtest und für dichotome nominale Variablen der McNemartest. Da für polytome nominale Variablen kein Test für verbundene Stichproben verfügbar ist, wird hier der χ^2-Test auf Homogenität eingesetzt.

Der Vorzeichen-Rangtest von Wilcoxon Der Wilcoxontest für verbundene Stichproben gehört zu den trennschärfsten parameterfreien Tests für Variablen mit kardinalem Messniveau (Clauss und Ebner, 1967, S. 225) und stellt damit eine sinnvolle Alternative zu den in der Evaluationsliteratur zu findenden Tests für metrisch skalierte Variablen dar. Seine Aussage ist vergleichbar mit der eines Mittelwerttests.

Mit dem Vorzeichen-Rangtest wird für jedes metrisch skalierte Merkmal überprüft, ob sich die Ausprägungen in den verbundenen Stichproben (der Teilnehmer- und der Kontrollgruppe) im Mittel unterscheiden oder nicht. Dabei wird angenommen, dass die Differenzen der einzelnen Beobachtungen unabhängige Stichprobenvariablen und

[15] Die zweite Fehlerquelle wird bei der Analyse der Zuordnungsprozesse berücksichtigt.

identisch verteilt sind. Darüber hinaus wird von der stetigen symmetrischen Verteilung der Differenzen um ihren Median ausgegangen (Büning und Trenkler, 1994, S. 171). Die Nullhypothese gleich verteilter Merkmalsausprägungen in beiden Teilstichproben lässt sich unter diesen Annahmen mit Hilfe des Medians der Differenzen formulieren:

$$H_0 : M_n = 0 \text{ vs. } H_1 : M_n \neq 0.$$

Dabei bezeichnet M_n den Median der Differenzen in der Ausprägung eines Merkmals x_n. Die einzelnen Differenzen geben jeweils den Unterschied in der Merkmalsausprägung zwischen einem Teilnehmer und seinem Partner an.

Die Prüfgröße des Wilcoxontests wird anhand der Rangfolge dieser Differenzen $\mathit{diff}_{ni} = x_{ni} - x_{nj}$ $(\forall i = 1, \ldots, I)$ ermittelt. Entsprechend ihres Absolutbetrages werden sie in eine Rangfolge gebracht, beginnend mit der kleinsten Differenz. Alle exakt gleichen Paare (mit $\mathit{diff}_{ni} = 0$) werden dabei nicht berücksichtigt. Treten gleiche Differenzen (sog. Bindungen) auf, werden diese Paare in einer Bindungsgruppe zusammmengefasst, für die ein mittlerer Rangplatz gebildet wird.[16] Die Summe der Rangplätze der positiven Differenzen bildet die Prüfgröße P_n:

$$P_n = \sum_{i=1}^{I} R_{ni}^{+} Q_{ni}. \tag{4.2}$$

Mit R_{ni}^{+} wird der Rang der positiven Differenz diff_{ni} bezeichnet, Q_{ni} dient als Indikator, für den gilt:

$$Q_{ni} = \begin{cases} 1 & \text{wenn } \mathit{diff}_{ni} > 0 \\ 0 & \text{sonst.} \end{cases}$$

Unter der Nullhypothese ist die Prüfgröße symmetrisch verteilt um den Mittelwert $E[P_n] = \frac{I(I+1)}{4}$ mit der Varianz $Var[P_n] = \sqrt{\frac{I(I+1)(2I+1)}{24}}$. Die Nullhypothese wird abgelehnt, wenn die Prüfgröße den kritischen Wert überschreitet. Die Verteilung der Prüfgröße entspricht asymptotisch einer Normalverteilung mit den o.g. Parametern. Für Stichproben mit einer Größe von mehr als 25 Teilnehmern kann deshalb als

[16] Für eine ausführliche Beschreibung dieser Durchschnittsrangbildung vgl. Büning und Trenkler (1994).

Prüfverteilung die Normalverteilung verwendet werden (Clauss und Ebner, 1967, S. 225f.). Das gewählte Signifikanzniveau zur Überprüfung der Matchingergebnisse liegt bei $\alpha = 5\%$.

Der χ^2-Homogenitätstest Der χ^2-Test auf Homogenität zweier unabhängiger Stichproben ist ein Verteilungstest für nominal skalierte Variablen. Anhand der beobachteten Häufigkeit des Auftretens der einzelnen Ausprägungen der Variablen wird deren Übereinstimmung mit der jeweils erwarteten Häufigkeit überprüft.[17]

Bei der Durchführung des Tests wird davon ausgegangen, dass die beobachteten Personen eine Zufallsstichprobe bilden und die Merkmalsausprägungen von Person zu Person unabhängig sind (Büning und Trenkler, 1994, S. 222ff.). Die Nullhypothese des Homogenitätstests besagt, dass die Wahrscheinlichkeitsverteilung der Ausprägungen des untersuchten Merkmals sich zwischen den beiden Stichproben nicht unterscheidet:

$$H_0 : p^T_{V_n} = p^C_{V_n} \text{ vs. } H_1 : p^T_{V_n} \neq p^C_{V_n}.$$

Dabei bezeichnet $p^T_{V_n}$ die Wahrscheinlichkeitsverteilung der Ausprägungen des Merkmals x_n in der Teilnehmergruppe und $p^C_{V_n}$ die entsprechende Wahrscheinlichkeitsverteilung in der Kontrollgruppe.

Die Prüfgröße des Homogenitätstests wird ermittelt aus der Summe der Anteile der quadrierten Abweichung der Auftrittshäufigkeit einer Ausprägung an der erwarteten Häufigkeit über alle untersuchten Stichproben und alle Ausprägungen (Siegel, 1997, S. 102). Für die Überprüfung der Matchingergebnisse vereinfacht sich die Prüfgröße, da nur zwei Stichproben (die Teilnehmer- und die Kontrollgruppe) verglichen werden und die erwarteten Häufigkeiten der einzelnen Ausprägungen den beobachteten Häufigkeiten in der Teilnehmergruppe entsprechen. Für die Prüfgröße P_n ergibt sich:

$$P_n = \sum_{v_n=1}^{V_n} \frac{\left(z^C_{v_n} - z^T_{v_n}\right)^2}{z^T_{v_n}}. \tag{4.3}$$

[17] Für die Anwendung des Tests zur Überprüfung der Matchingergebnisse entspricht die erwartete Häufigkeit der beobachteten Häufigkeit in der Teilnehmerstichprobe und die beobachtete Häufigkeit der in der Kontrollgruppe.

Dabei bezeichnet $z^T_{v_n}$ die beobachtete Häufigkeit des Auftretens einer Ausprägung v_n in der Teilnehmergruppe, $z^C_{v_n}$ die Häufigkeit in der Kontrollgruppe und V_n die Anzahl der Ausprägungen einer Variable.

Die Verteilung dieser Prüfgröße entspricht unter der Nullhypothese annähernd der χ^2-Verteilung. Die Nullhypothese gleicher Wahrscheinlichkeitsverteilungen wird abgelehnt, wenn gilt: $P_n > \chi^2_{(V_n-1),(1-\alpha)}$, d.h. wenn die Prüfgröße den kritischen Wert übersteigt. Als Signifikanzniveau wird ebenfalls $\alpha = 5\%$ gewählt.

Der McNemartest Der McNemartest ist ein Homogenitätstest für dichotome Merkmale in zwei verbundenen Stichproben. Das Testprinzip beruht ebenfalls auf dem Vergleich der beobachteten mit den erwarteten Häufigkeiten. Anders als beim χ^2-Homogenitätstest werden aber nur die Unterschiede betrachtet. Dies lässt sich mit Hilfe einer Vier-Felder-Tafel verdeutlichen (vgl. Abbildung 4.1). Dabei entspricht die erwartete Häufigkeit der beobachteten Auftrittshäufigkeit einer Ausprägung in der Teilnehmerstichprobe, die beobachtete Häufigkeit derjenigen in der Kontrollgruppe.

C \ T	0	1	$\sum_C$
0	$z^T_{0,C0}$	$z^T_{1,C0}$	z^C_0
1	$z^T_{0,C1}$	$z^T_{1,C1}$	z^C_1
$\sum_T$	z^T_0	z^T_1	$z^T = z^C$

Abbildung 4.1: *Kombination der Merkmalsausprägungen eines dichotomen Merkmals im Zwei-Stichproben-Fall*

Quelle: Eigene Darstellung in Anlehnung an Siegel (1997) S. 61.

In jedem der vier Felder finden sich die Häufigkeiten des Auftretens folgender Fälle: im oberen linken Feld die Häufigkeit des gemeinsamen Auftretens der Ausprägung Null in beiden Stichproben, im oberen rechten Feld die des Auftretens von Null in der Kontrollgruppe C und Eins in der Teilnehmerstichprobe T, im unteren linken Feld der umgekehrte Fall (Null in der Teilnehmer- und Eins in der Kontrollgruppe) sowie im unteren rechten Feld die Auftrittshäufigkeit einer gemeinsamen Eins. Im McNemartest werden nur das obere rechte und das untere linke Feld betrachtet, die

beiden anderen bleiben unberücksichtigt.[18] Dabei gelten die gleichen Annahmen wie für den χ^2-Homogenitätstest.

Die erwarteten Häufigkeiten in beiden betrachteten Feldern stimmen überein. Anders ausgedrückt: Unter der Nullhypothese ist die erwartete Auftrittshäufigkeit beider Unterschiede identisch:

$$H_0 : z^T_{1,C0} = z^T_{0,C1} = \frac{z^T_{1,C0} + z^T_{0,C1}}{2} \quad \text{vs.} \quad H_1 : z^T_{1,C0} \neq z^T_{0,C1}.$$

Dabei wird die Gesamtzahl der betrachteten Fälle durch $z^T_{1,C0}+z^T_{0,C1}$ angegeben. Unter Nutzung dieses Zusammenhanges und der Verwandtschaft zum χ^2-Homogenitätstest kann die Prüfgröße P_n abgeleitet werden (Siegel, 1997, S. 61f.):

$$P_n = \frac{\left(z^T_{1,C0} - z^T_{0,C1}\right)^2}{z^T_{1,C0} + z^T_{0,C1}}.$$

Dabei bezeichnet $z^T_{1,C0}$ die Häufigkeit des Auftretens von Null in der Kontrollgruppe und Eins in der Teilnehmerstichprobe und $z^T_{0,C1}$ den umgekehrten Fall (Null in der Teilnehmer- und Eins in der Kontrollgruppe). Die Kontinuitätskorrektur von Yates verändert die Prüfgröße (Siegel, 1997, S. 62):[19]

$$P_n = \frac{\left(|z^T_{1,C0} - z^T_{0,C1}| - 1\right)^2}{z^T_{1,C0} + z^T_{0,C1}}. \tag{4.4}$$

Die Nullhypothese gleicher Häufigkeiten wird abgelehnt, wenn die Prüfgröße den kritischen Wert übersteigt, d.h. wenn gilt: $P_n > \chi^2_{1,(1-\alpha)}$. Auch für diesen Test wird $\alpha = 5\%$ als Signifikanzniveau gewählt.[20]

Darüber hinaus soll betrachtet werden, wie viel der ursprünglichen Ungleichverteilung der Merkmale in Teilnehmer- und Nichtteilnehmergruppe durch das Matching beseitigt wurde. Dazu wird die prozentuale Reduzierung des Bias verwendet. Abwei-

[18] Der oben beschriebene χ^2-Homogenitätstest würde dagegen alle vier Felder in den Test einbeziehen.

[19] Durch diese Korrektur wird die Annäherung der Stichprobenverteilung an die χ^2-Verteilung verbessert (Siegel, 1997, S. 62).

[20] Wenn die erwarteten Häufigkeiten kleiner als 5 sind, wird anstelle des McNemartests ein Binomialtest durchgeführt (Siegel, 1997, S. 64).

chend von der in der Literatur üblichen Berechnung wird anstelle des Durchschnitts der Einzeldifferenzen nach Matching (vgl. Abschnitt 3.3.3) die Differenz der Mittelwerte betrachtet:[21]

$$B_{nach}(x_n) = \frac{1}{I}\sum_{i \in T} x_{ni} - \frac{1}{J^C}\sum_{j \in J^C} x_{nj}. \tag{4.5}$$

Dabei bezeichnet I die Anzahl der Teilnehmer, J^C die Anzahl der Kontrollgruppenmitglieder, x_{ni} steht für die Ausprägung des Merkmals x_n bei einem Teilnehmer i, x_{nj} für die des Nichtteilnehmers j. Diese Mittelwertdifferenz wird – wie das in Abschnitt 3.3.3 beschriebene Gütemaß – mit der Differenz vor Matching verglichen.

In beiden Maßen sollte sichtbar werden, welches der Distanzmaße am besten in der Lage ist, Unterschiede zu identifizieren und bei der Zuordnung zu berücksichtigen. Da für jedes Distanzmaß die gleiche Stichprobe und die gleiche Zuordnung gewählt wird, sind Unterschiede im Ergebnis allein auf die „Fähigkeiten“ der Distanzmaße zurückzuführen.

Analyse der Zuordnungsalgorithmen

Im zweiten Teil der Analyse werden verschiedene Zuordnungsprozesse miteinander verglichen. In der Literatur haben sich Ridge Matching, Optimal Full Matching sowie die Zuordnung mit Zurücklegen als vorteilhaft gegenüber anderen Matchingverfahren erwiesen. Die untersuchten Algorithmen werden ergänzt durch einen weiteren optimalen Zuordnungsprozess, ein Nearest Neighbor Matching auf Grundlage des Ungarischen Algorithmus. Darüber hinaus wird in die Analyse – zum Vergleich der empfohlenen Verfahren mit einem „Standardansatz“ – ein in der Anwendung weit verbreiteter Algorithmus, das Random Matching, einbezogen.

Die Zuordnung mit Zurücklegen und das Random Matching werden in der im Abschnitt 2.3.2 dargestellten Form analysiert. Für die Anwendung des Ungarischen Algorithmus, dessen Schrittfolge ebenfalls im Abschnitt 2.3.2 erläutert wird, wird ein Tool von Borlin (1999) genutzt. Da sich die Anzahl der Teilnehmer von der der

[21] Dieses Maß erscheint sinnvoller zur Beurteilung des Matchingergebnisses insgesamt als eine Betrachtung der Angleichung jeder einzelnen Unter-Kontrollgruppe an „seinen“ Teilnehmer, die bei der gängigen Biasdefinition im Mittelpunkt steht (Gu und Rosenbaum, 1993, S. 411).

Nichtteilnehmer unterscheidet, muss die Distanzmatrix vorher in eine quadratische Form gebracht werden. Dazu werden zusätzliche Zeilen oder Spalten eingefügt, deren Elemente aus großen Zahlen (9999) bestehen – und damit vom Algorithmus nicht berücksichtigt werden. Da in den meisten Fällen die Anzahl der Nichtteilnehmer die der Teilnehmer übersteigt, entspricht die Zeilenanzahl in der resultierenden Distanzmatrix der Anzahl der Nichtteilnehmer.

Die Grundlage für die optimale Zuordnung mit Hilfe des Full Matching bildet ein Tool von Kumar (2007) zur optimalen 1:1-Zuordnung mit Hilfe eines Auktionsalgorithmus. Dieses Tool wird für eine vollständige Zuordnung entsprechend erweitert. Wie im Abschnitt 2.3.2 beschrieben wird, handelt es sich bei Auktionsalgorithmen um Maximierungsprozesse. Die ermittelten Distanzen müssen also im Vorfeld in Ähnlichkeiten umgewandelt werden. Die Vergleichsgröße für jeden Teilnehmer wird nach Abschluss der Zuordnung als arithmetisches Mittel der Einkommen aller Nichtteilnehmer in seiner Unter-Kontrollgruppe ermittelt.

Die Spezifikation des Ridge Matchings (2.42) entspricht der in Abschnitt 2.3.2 vorgestellten Kombination aus Kern Matching und Local Linear Regression.[22] Im Unterschied zu den anderen analysierten Zuordnungsprozessen basiert das Ridge Matching auf Propensity Scores. Als Kernfunktion zur Gewichtung der Differenzen zwischen den Propensity Scores der Teilnehmer und Nichtteilnehmer d wird ein Epanechnikovkern $K(d) = 0.75\,(1 - d^2)\,1_{[-1,1]}\,(d)$ verwendet. Die optimale Bandbreite liegt im Bereich $0 < b^{opt} \leq 2$. Mit Hilfe eines Leave-One-Out-Prozesses wird eine datenabhängige Näherungslösung für die optimale Bandbreite ermittelt. Für die Schrittgröße gilt: $0.01\sqrt{1.2^{step-2}}$ mit $step = 1, 4, 7, \ldots, 52, 54, \ldots, 60$. Die Auswahl der „besten“ Bandbreite erfolgt nach folgender Regel:

$$b^* = \underset{b}{argmin} \sum_{j=1}^{J} [Y_j - Y\_j(PS_j)]^2. \tag{4.6}$$

Dabei bezeichnet b die generierten Bandbreiten, b^* die ausgewählte Bandbreite, Y_j das Einkommen eines Nichtteilnehmers j und Y_j die ohne Berücksichtigung des Nichtteilnehmers j geschätzte Vergleichsgröße für dieses Einkommen.

[22] Die Auswahl der Kernfunktion und die Ermittlung der Bandbreite erfolgt in Anlehnung an Fröhlich (2004a).

Simulationsdesign für die Analyse Die Güte des Ergebnisses jedes Zuordnungsprozesses wird von zwei Faktoren beeinflusst. Das Größenverhältnis der Teilnehmer- zur Nichtteilnehmerstichprobe gibt an, wie viele Nichtteilnehmer je Teilnehmer im Durchschnitt für eine Zuordnung zur Verfügung stehen. Die Größe des Common-Support-Bereichs gibt Auskunft darüber, wie viele dieser Personen mögliche Partner für die Teilnehmer sind. Je geringer das Verhältnis zwischen Anzahl der Teilnehmer und Anzahl der Nichtteilnehmer wird und je kleiner der Common-Support-Bereich, desto schwieriger ist die Suche nach geeigneten Partnern – und umso höher damit die Anforderung an den Zuordnungsprozess.

Beide Faktoren werden in der Analyse betrachtet. Eine gegebene Teilnehmerstichprobe wird dazu mit unterschiedlich großen und unterschiedlich ähnlichen Nichtteilnehmerstichproben kombiniert. Es werden drei verschiedene Größenverhältnisse zwischen Teilnehmer- und Nichtteilnehmeranzahl festgelegt: 1:1, 1:3 sowie 1:10. Innerhalb eines Größenverhältnisses wird dann die Übereinstimmung der Merkmalsausprägungen in beiden Teilstichproben variiert. Als ähnliche Merkmale in den Nichtteilnehmerstichproben gelten solche, deren Merkmalsmittelwert bzw. Häufigkeitsverteilung der Ausprägungen um 1% der merkmalsspezifischen Streuung abweicht.[23] Als eingeschränkt ähnlich werden Variablen mit einer Abweichung von 10% der merkmalsspezifischen Streuung, als unähnlich solche mit einer Abweichung von 25% bezeichnet. Um zusätzlich den Einfluss der Stichprobengröße auf das Matchingergebnis berücksichtigen zu können, wird dieser Vorgang für Teilnehmerstichproben in drei unterschiedlichen Größen (50, 100 und 300 Personen) durchgeführt. Die Stichprobe mit 50 Teilnehmern kann als „Untergrenze“ für die Personenanzahl einer analysierbaren Stichprobe angesehen werden.[24] Die Stichprobe mit 100 Teilnehmern entspricht der in der Literatur gebräuchlichen Definition kleiner Stichproben (Fröhlich, 2004a; Zhao, 2004). Die Stichprobe mit 300 Teilnehmern gilt als „groß“ in dieser Studie. Sie wird verwendet, um den Einfluss der Größe der Ausgangsstichproben auf die Qualität der Matchingergebnisse festzuhalten.[25]

[23] Die Abweichung in den Variablen wird auf die in Abschnitt 4.2.1 beschriebene Weise erzeugt. Informationen dazu finden sich außerdem in der Tabelle B.1 im Anhang.

[24] Eine Stichprobe entsprechender Größe wird in einer Studie von Gu und Rosenbaum (1993) verwendet.

[25] Die Ergebnisse von Fröhlich (2004a) legen nahe, dass zwischen der Qualität der erzielten Matchingergebnisse und der Größe der Stichprobe ein positiver Zusammenhang besteht.

Da für Stichproben mit einem Verhältnis der Personenanzahl von 1:1 der Einsatz von Zuordnungsverfahren ohne Mehrfachnutzung von Personen nicht sinnvoll ist, werden für diese Stichproben nur Ridge Matching und die Zuordnung mit Zurücklegen untersucht.[26] Zum Vergleich wird allerdings auch in diesen Stichproben Random Matching eingesetzt.

Es ergeben sich nur vergleichbare Aussagen, wenn die Zuordnungen auf dem gleichen Distanzmaß beruhen. Deshalb wird für alle Zuordnungsprozesse die gewichtete Mahalanobis-Matching-Distanz verwendet.[27] Eine Ausnahme bildet das Ridge Matching, das auf Grundlage des Propensity Scores durchgeführt wird.

Gütemaße zur Beurteilung der Zuordnungsprozesse Als Hauptkriterium zur Beurteilung der Distanzmaße wird die Angleichung der Mittelwerte der einzelnen Merkmale in Teilnehmer- und Nichtteilnehmerstichprobe eingesetzt. Dieses Kriterium ist für die Beurteilung der Zuordnungsprozesse nicht anwendbar, da nicht bei allen untersuchten Algorithmen bekannt ist, aus welchen Personen (mit welchen Merkmalen) die Kontrollgruppen für die einzelnen Teilnehmer bestehen.[28] Als Näherungswert für die Übereinstimmung der Teilstichproben in den Merkmalen wird deshalb die Übereinstimmung im verwendeten Distanzmaß – als Zusammenfassung aller Merkmale – betrachtet. Als Gütekriterium dient die Summe der quadrierten Distanzen zwischen Teilnehmern und Nichtteilnehmern.

Darüber hinaus stellt sich gerade in kleinen Stichproben die Frage nach dem Verlust von Teilnehmern durch Unzulänglichkeiten des Zuordnungsprozesses. Zur Beurteilung der Güte der Matchingergebnisse in diesem Punkt wird deshalb die Anzahl

[26] Die Anwendung von Zuordnungsverfahren ohne Zurücklegen ist bei gleich großen Ausgangsstichproben problematisch, da der Verlust von Teilnehmern aufgrund fehlender Nichtteilnehmer (wenn weniger Nichtteilnehmer die Common-Support-Bedingung erfüllen) bzw. die Zuordnung sehr unähnlicher Partner, weil keine anderen mehr verfügbar sind, in besonderem Maße auftritt. Aus diesem Grund werden in der Literatur v.a. Zuordnungsverfahren mit Zurücklegen angewendet (Lechner, Miquel und Wunsch, 2004; Sianesi, 2001).

[27] Die Mahalanobis-Matching-Distanz erweist sich im ersten Teil der Analyse als vorteilhaft gegenüber den anderen Distanzmaßen.

[28] Beim Ridge Matching dienen die Merkmale der Personen – bzw. der daraus ermittelte Propensity Score – zur Definition der Bandbreite, innerhalb derer alle Einkommen zur Konstruktion der Vergleichsgröße verwendet werden. Die Merkmale der einzelnen Personen werden dabei nicht näher betrachtet.

der entfernten (Teilnehmer-)Beobachtungen, absolut und in Relation zur Größe der Ausgangsstichprobe, eingesetzt.

Als weiteres Kriterium wird ein Effizienzmaß, der mittlere quadratische Fehler (MSE) bzw. seine Wurzel, der Root Mean Square Error (RMSE), betrachtet. Dieses Maß setzt sich – wie im Abschnitt 3.3.3 beschrieben – aus der Varianz und dem Bias des Maßnahmeeffektschätzers zusammen. Bias bedeutet in diesem Zusammenhang die Abweichung des durchschnittlichen geschätzten vom durchschnittlichen „wahren“ Nichtteilnahmeeinkommen (Dehejia und Wahba, 2002, S. 158) und wird ermittelt als Differenz der mittleren Einkommensgrößen über alle Stichproben:

$$B\left(\widehat{ME}\right) = \bar{Y}^T_{NT_{wahr}} - \hat{\bar{Y}}^C. \tag{4.7}$$

Dabei bezeichnet $B\left(\widehat{ME}\right)$ den Bias des Maßnahmeeffektschätzers, $\bar{Y}^T_{NT_{wahr}}$ das durchschnittliche „wahre“ Nichtteilnahmeeinkommen der Teilnehmer und $\hat{\bar{Y}}^C$ das arithmetische Mittel der Nichtteilnahmeeinkommen in den gebildeten Kontrollgruppen.

Der Bias gibt Auskunft über die Fähigkeit der untersuchten Matchingverfahren, den „wahren“ Wert des Vergleichseinkommens mit den gegebenen Daten abzubilden. Die Varianz des geschätzten Maßnahmeeffekts ist ein Maß für die Streubreite der einzelnen Schätzwerte um ihren Mittelwert. Beide Bestandteile des MSE werden – zusätzlich zum MSE – in die Beurteilung der Zuordnungsprozesse einbezogen. Als Zusatzinformation für den Vergleich der Zuordnungsalgorithmen wird die durchschnittliche Abweichung der geschätzten von der „wahren“ Vergleichsgröße und die Veränderung dieser Abweichung durch Matching (Bias Reduzierung) – als Mittelwert über alle Stichproben – ermittelt.[29]

Das Problem aller zur Beurteilung der Zuordnungsprozesse verwendeten Gütemaße besteht darin, dass sie – im Unterschied zu den für Distanzmaße eingesetzten Tests – keine statistisch gesicherte Aussage über die Güte der erzielten Ergebnisse ermöglichen. Die Simulationsergebnisse für die Zuordnungsprozesse können also nur

[29] Im Unterschied zur Biasdefinition des MSE wird dabei der Bias vor und nach dem Matching als Duchschnittswert über die Abweichungen in jeder Stichprobe gebildet („Durchschnitt der Differenzen“ anstelle der „Differenz der Durchschnitte“). Für beide Gütemaße gilt: $|B_{nach}| = \left|B\left(\widehat{ME}\right)\right|$.

untereinander verglichen und die Algorithmen hinsichtlich der genannten Gütemaße in eine Rangfolge gebracht werden.

4.3 Simulationsergebnisse

Zur Überprüfung der einzelnen Distanzmaße und Zuordnungsprozesse werden jeweils 100 Simulationsläufe durchgeführt. Die Durchschnitte der Simulationsergebnisse für diese 100 Durchläufe bilden die im Folgenden präsentierten Ergebnisse.

4.3.1 Analyse der Distanzmaße

In der Studie zur Analyse der Distanzmaße werden Stichproben mit 100 Teilnehmern mit verschiedenen Nichtteilnehmerstichproben der Größe 1000 Personen kombiniert. Zur Beurteilung der Güte der Matchingergebnisse werden zwei Kriterien angewendet: skalenspezifische Tests der Übereinstimmung der Mittelwerte bzw. Häufigkeitsverteilungen der Variablen sowie die Reduzierung der ursprünglichen Abweichung durch Matching. Auf Übereinstimmung der metrischen Variablen wird mit Hilfe des Vorzeichen-Rangtests von Wilcoxon getestet, für dichotome Merkmale wird dafür der McNemartest, für polytome der χ^2-Anpassungstest eingesetzt.

In den folgenden Tabellen werden die wichtigsten Ergebnisse für jedes der vier Untersuchungsdesigns – Stichproben mit ähnlichen Merkmalen, Stichproben mit unähnlichen metrischen Merkmalen bei ähnlichen nominalen Merkmalen, Stichproben mit unähnlichen nominalen Merkmalen bei ähnlichen metrischen Variablen sowie Stichproben mit unähnlichen Merkmalen – zusammengefasst.[30]

In der Tabelle 4.1 werden die Ergebnisse für das Ausgangsdesign der Simulation, ähnliche Stichproben, dargestellt. Dieses Design ist gekennzeichnet durch eine Abweichung der Merkmalsmittelwerte bzw. Häufigkeitsverteilungen der Merkmalsausprägungen von 1% der Streuung des jeweiligen Merkmals zwischen den Teilnehmer- und den Nichtteilnehmerstichproben.

[30] Im Anhang B.2 finden sich detaillierte Ergebnisse für jedes der untersuchten Distanzmaße.

Die erste Spalte enthält Informationen über die für die Bildung der Distanzmaße genutzten Merkmale: Variablen 1-5 sind metrisch skaliert, 6-9 sind dichotome und 10-12 polytome nominale Merkmale.[31] Für alle Distanzmaße werden die gleichen Variablen verwendet. In der zweiten Spalte werden die Merkmalsmittelwerte der jeweiligen Variablen aus den Teilnehmerstichproben zusammengefasst. Diese Mittelwerte sind innerhalb eines Stichprobendesigns für alle Distanzmaße gleich, da für alle Untersuchungen dieselben Teilnehmerstichproben verwendet werden. Es können graduelle Unterschiede zwischen den vier Stichprobendesigns auftreten, die darauf zurückzuführen sind, dass während der Prüfung der Common-Support-Bedingung Personen aus der Analyse entfernt werden, wenn sie keinen potenziellen Partner haben. Das gleiche gilt für die Mittelwerte in den Nichtteilnehmerstichproben, die in der dritten Spalte zu finden sind.

Unterschiedliche Merkmalsmittelwerte in den gebildeten Kontrollgruppen sind ausschließlich auf die unterschiedliche Gewichtung der Merkmale in den einzelnen Distanzmaßen zurückzuführen, die zur Zuordnung unterschiedlicher Partner zu den jeweiligen Teilnehmern und damit zu unterschiedlich gut angepassten Kontrollgruppen führt. Ein Einfluss des Zuordnungsprozesses ist ausgeschlossen, da für alle Maße der Ungarische Algorithmus angewendet wird.

Die Spalten 4-15 geben die Analyseergebnisse für die einzelnen Distanzmaße wieder. Es werden jeweils die Merkmalsmittelwerte in der gebildeten Kontrollgruppe, die Ergebnisse der skalenspezifischen Tests sowie die prozentuale Reduzierung der Abweichung durch Matching (Bias Reduzierung) präsentiert. Die Testergebnisse werden als durchschnittliche Ablehnungsrate der Nullhypothese gleicher Mittelwerte bzw. Häufigkeitsverteilungen über alle Simulationsläufe ausgewiesen, die Werte liegen zwischen Null und Eins. Je kleiner der ausgewiesene Wert ist, desto besser gelingt die Angleichung der Merkmalsverteilungen in Teilnehmer- und Kontrollguppe – umso besser ist das eingesetze Distanzmaß. Ein positiver Wert der Bias Reduzierung gibt an, dass die Abweichung der Merkmalsmittelwerte sich durch das Matching verringert hat, ein negativer Wert dagegen bedeutet eine Vergrößerung der Abweichung.

[31] Eine Übersicht über die Merkmale, die im Matchingprozess verwendet werden, liefert die Tabelle B.1 im Anhang.

Tabelle 4.1: **Ergebnisse für Stichproben mit ähnlichen metrischen und nominalen Variablen**

Ausgangsdaten			Propensity Score			Mahalanobisdistanz			Mahalanobis-Matching			Gowerdistanz		
X[a]	T[b]	NT[b]	C[b]	Test[c]	BiasR[d]	C[b]	Test[c]	BiasR[d]	C[b]	Test[c]	BiasR[d]	C[b]	Test[c]	BiasR[d]
1	40,05	39,80	37,15	0,67	-155,07	40,03	0,15	56,85	39,94	0,20	8,99	40,02	0,08	80,86
2	0,98	1,04	0,91	0,36	-46,76	1,00	0,08	55,29	1,00	0,15	26,83	0,98	0,12	74,74
3	12,02	11,99	11,12	0,65	-99,70	12,00	0,18	59,67	12,07	0,15	31,10	12,04	0,11	78,46
4	11,95	12,28	10,65	0,42	-59,02	11,98	0,16	50,29	12,30	0,19	9,56	11,92	0,12	75,85
5	1302,33	1323,95	1164,49	0,47	-83,74	1302,71	0,20	50,27	1332,45	0,19	-7,43	1298,01	0,06	79,15
6	0,50	0,51	0,51	0,46	-55,11	0,50	0,01	76,57	0,50	0,00	89,64	0,50	0,15	2,18
7	0,66	0,63	0,71	0,44	-49,00	0,67	0,06	77,05	0,66	0,02	87,39	0,64	0,14	8,18
8	0,90	0,89	0,90	0,26	-18,17	0,91	0,01	74,73	0,93	0,12	51,25	0,90	0,07	-1,87
9	0,20	0,19	0,34	0,62	-144,62	0,20	0,00	81,90	0,18	0,08	68,78	0,19	0,15	0,57
10	2,84	2,86	2,65	0,79	-539,66	2,88	0,43	-65,71	2,87	0,00	-3,10	2,86	0,59	-124,57
11	2,36	2,37	2,22	0,74	-576,42	2,34	0,50	-63,65	2,34	0,01	-48,47	2,37	0,62	-236,62
12	3,15	3,16	2,94	0,72	-565,61	3,23	0,46	-137,31	3,21	0,00	-34,12	3,15	0,64	-104,91

Anmerkungen:
Durchschnittsergebnisse für 100 Stichproben.
Abweichung der Mittelwerte bzw. Häufigkeitsverteilungen: 1% der merkmalsspezifischen Streuung.
[a] In die Analyse einbezogene Variablen; Skalenniveau: 1-5 metrisch, 6-9 dichotom, 10-12 polytom;
[b] Durchschnitttliche Ausprägung des Merkmals in den Teilnehmerstichproben (T), Nichtteilnehmerstichproben (NT) bzw. den Kontrollgruppen (C);
[c] Skalenspezifische Tests (metrische Variablen: Wilcoxontest, dichotome: McNemartest, polytome: χ^2-Test), durchschnittliche Ablehnungsrate der Nullhypothese gleicher Mittelwerte bei einem Signifikanzniveau von 5%;
[d] Prozentuale Reduzierung der Abweichung der Merkmalsmittelwerte durch Matching.

Die durchgeführte Untersuchung liefert nahezu identische Ergebnisse für den Propensity Score und den Index Score.[32] In den Tabellen werden deshalb nur die Analyseergebnisse für den Propensity Score (in den Spalten 4-6) ausgewiesen. Die Spalten 7-9 enthalten die Ergebnisse der Analyse der Mahalanobisdistanz, die Spalten 10-15 die Ergebnisse für die aggregierten Distanzmaße – die gewichtete Mahalanobis-Matching-Distanz (in den Spalten 10-12) und das Distanzmaß nach Gower (in den Spalten 13-15).

Die Gütekriterien zur Beurteilung der Analyseergebnisse zeichnen ein deutliches Bild. Die Mahalanobisdistanz und die beiden aggregierten Distanzmaße sind besser in der Lage als Propensity Score und Index Score, geeignete Partner für die Teilnehmer zu identifizieren. Die Unterschiede in der Güte der gebildeten Kontrollgruppen sind erheblich.

Der Vergleich der Mittelwerte der (durchschnittlichen) Merkmalsausprägungen in den mit Hilfe der Scores gebildeten Kontrollgruppen mit denen der Ausgangsstichproben (der Nichtteilnehmer) zeigt, dass anhand von Propensity Score und Index Score Nichtteilnehmer ausgewählt werden, deren Merkmalsausprägungen im Durchschnitt stärker von denen der Teilnehmer abweichen als vor dem Matching. Das zeigt sich auch in der negativen prozentualen Reduzierung des Bias. Diese Zuordnung „unpassender“ Nichtteilnehmer führt dazu, dass die Hypothese gleicher Merkmalsmittelwerte in den durchgeführten Tests für beide Scores in etwa der Hälfte aller Stichproben (im Durchschnitt über alle Merkmale) abgelehnt wird.

Die durchschnittlichen Merkmalsmittelwerte in den Kontrollgruppen, die anhand der anderen drei Distanzmaße gebildet werden, sind denen der Teilnehmerstichprobe insgesamt ähnlicher. Im Vergleich zu den Mittelwerten der Ausgangsstichproben werden auf Grundlage der Mahalanobisdistanz und des Distanzmaßes nach Gower Nichtteilnehmer zugeordnet, deren Merkmalsausprägungen im Durchschnitt weniger oder genauso abweichen. Ein weniger eindeutiges Bild ergibt der entsprechende

[32] Beide Distanzmaße unterscheiden sich geringfügig in der durchschnittlichen Ablehnungsrate der Nullhypothese gleicher Mittelwerte für einige Variablen; die durchschnittlichen Merkmalsausprägungen sowie die prozentuale Reduzierung des Bias sind für alle untersuchten Merkmale identisch. Vgl. dazu die Tabellen B.2 und B.3 im Anhang B.2.

Vergleich der auf Basis der gewichteten Mahalanobis-Matching-Distanz gebildeten Kontrollgruppe.

Die Betrachtung der durchschnittlichen Merkmalsmittelwerte bietet eine erste Orientierung zur Beurteilung der Distanzmaße. Allerdings sind aus diesen Werten die Abweichungen der Mittelwerte in den einzelnen Stichproben nicht ersichtlich. Auskunft über die Größe der einzelnen Abweichungen nach Matching – und die Verringerung ursprünglich vorhandener Unterschiede – geben die Ergebnisse der durchgeführten Tests und die prozentuale Reduzierung des Bias.

Sowohl mit der Mahalanobisdistanz als auch mit beiden aggregierten Distanzmaßen wird die Abweichung der Merkmalsmittelwerte zwischen den einzelnen Teilnehmer- und Nichteilnehmerstichproben verringert. Das Ausmaß dieser Reduzierung ist dabei sehr unterschiedlich. Eine Ausnahme bilden die polytomen Merkmale, deren Unterschied durch das Matching vergrößert wird.[33] Interessant für die empirische Forschung ist, mit welchem Distanzmaß die ähnlichsten Partner identifiziert und damit die ähnlichsten Kontrollgruppen gebildet werden können. Diese Information liefern die Ergebnisse der skalenspezifischen Tests.[34]

Die durchschnittliche Ablehnungsrate der Hypothese gleicher Mittelwerte bzw. Häufigkeitsverteilungen liegt bei Mahalanobisdistanz und dem Distanzmaß nach Gower bei ca. einem Fünftel der Stichproben, bei der gewichteten Mahalanobis-Matching-Distanz nur bei ca. 10%. Das Ziel des Matchings, die Angleichung der Merkmalsverteilungen in Teilnehmer- und Kontrollguppe – und damit die Nicht-Ablehnung der Nullhypothese in den durchgeführten Tests –, wird mit der gewichteten Mahalanobis-Matching-Distanz am häufigsten erreicht. Insgesamt scheint dieses Distanzmaß bei sehr ähnlichen Stichproben am besten in der Lage zu sein, die jeweils besten Partner für die Teilnehmer zu identifizieren.

Die Betrachtung der Testergebnisse in den unterschiedlichen Skalenniveaus der Variablen offenbart darüber hinaus interessante Unterschiede zwischen den Distanzmaßen. So zeigt sich bei den beiden erstgenannten Maßen eine „Schwäche“ bei der

[33] Bei Betrachtung dieses Gütekriteriums muss allerdings berücksichtigt werden, dass die Ausgangsstichproben der Nichtteilnehmer den Teilnehmern definitionsgemäß sehr ähnlich sind, was keinen großen Spielraum für die Verringerung der Abweichung lässt.

[34] Die Tests werden auf einem Signifikanzniveau von $\alpha = 5\%$ durchgeführt.

Angleichung der polytomen Merkmale – die durchschnittlichen Ablehnungsraten sind mit ca. 45% bzw. 60% sehr hoch – während die Angleichung dieser Variablen mit der gewichteten Mahalanobis-Matching-Distanz sehr gut gelingt. Die „Schwäche“ dieses Maßes scheint bei den metrisch skalierten Merkmalen zu liegen, bei denen die Ablehnungsrate knapp ein Fünftel beträgt – ebenso wie bei der Mahalanobisdistanz. Für metrisch skalierte Merkmale gelingt die Angleichung mit dem Distanzmaß nach Gower am besten. Die Angleichung dichotomer, nominal skalierter Merkmale gelingt mit allen drei Maßen relativ gut, wobei die Ablehnungsrate beim Distanzmaß nach Gower etwas höher liegt als bei beiden anderen Maßen.

Als Zwischenergebnis kann festgehalten werden, dass bei Vorliegen ähnlicher Stichproben der Propensity Score und der Index Score nicht geeignet sind, passende Partner für die Teilnehmer zu identifizieren. Das gelingt mit Mahalanobisdistanz und den aggregierten Distanzmaßen besser. Zwischen diesen drei Maßen lassen sich Unterschiede in der Anpassungsgüte der Merkmalsmittelwerte in Abhängigkeit vom Skalenniveau der Merkmale feststellen. Insgesamt können mit Hilfe der gewichteten Mahalanobis-Matching-Distanz die (im Vergleich zu den anderen analysierten Distanzmaßen) jeweils besten Partner für die Teilnehmer identifiziert und zugeordnet werden.

Die folgenden Tabellen fassen die Simulationsergebnisse für die Stichprobendesigns mit einer Abweichung in jeweils einer Variablenart zusammen. In der Tabelle 4.2 werden die Ergebnisse für unähnliche metrische Merkmale bei ähnlichen nominalen Merkmalen dargestellt, in Tabelle 4.3 der umgekehrte Fall unähnlicher nominaler (dichotomer und polytomer) Merkmale bei ähnlichen metrischen Merkmalen. Stichproben mit unähnlichen metrischen Merkmalen sind durch eine durchschnittliche Abweichung der Mittelwerte um 25% der merkmalsspezifischen Streuung gekennzeichnet. In den Stichproben mit unähnlichen nominalen Variablen weicht die Häufigkeitsverteilung der Ausprägungen der dichotomen und polytomen Merkmale um 25% der merkmalsspezifischen Abweichung vom Median ab.

Tabelle 4.2: **Ergebnisse für Stichproben mit unähnlichen metrischen Variablen (bei ähnlichen nominalen Variablen)**

Ausgangsdaten			Propensity Score			Mahalanobisdistanz			Mahalanobis-Matching			Gowerdistanz		
X[a]	T[b]	NT[b]	C[b]	Test[c]	BiasR[d]	C[b]	Test[c]	BiasR[d]	C[b]	Test[c]	BiasR[d]	C[b]	Test[c]	BiasR[d]
1	40,06	40,30	38,12	0,58	-3,59	40,20	0,65	55,55	40,34	0,71	14,20	40,08	0,31	82,41
2	0,98	1,08	0,94	0,31	32,52	1,01	0,62	52,52	1,03	0,53	31,45	0,99	0,31	80,62
3	12,03	11,80	11,73	0,41	38,31	11,94	0,62	60,74	11,94	0,65	35,23	11,97	0,47	78,68
4	11,97	12,99	10,89	0,50	23,06	12,33	0,62	58,56	12,77	0,72	14,83	12,08	0,37	80,70
5	1301,84	1383,92	1188,05	0,53	15,49	1330,01	0,64	58,22	1387,07	0,75	1,27	1308,83	0,41	82,69
6	0,50	0,49	0,54	0,25	6,43	0,50	0,04	71,86	0,50	0,00	89,45	0,49	0,16	-3,16
7	0,66	0,65	0,70	0,19	15,76	0,66	0,04	75,61	0,66	0,00	87,53	0,65	0,14	-4,60
8	0,90	0,90	0,91	0,14	8,23	0,91	0,02	68,78	0,93	0,09	48,01	0,90	0,06	3,80
9	0,20	0,21	0,25	0,25	-16,62	0,20	0,02	75,95	0,18	0,08	73,47	0,21	0,14	-0,04
10	2,84	2,85	2,77	0,76	-384,16	2,89	0,49	-64,45	2,87	0,00	14,02	2,85	0,71	-134,32
11	2,35	2,36	2,32	0,78	-449,91	2,34	0,50	-72,42	2,33	0,01	-32,41	2,36	0,63	-147,09
12	3,16	3,16	3,05	0,76	-473,92	3,23	0,49	-128,76	3,21	0,00	-42,57	3,17	0,60	-141,09

Anmerkungen:
Durchschnittsergebnisse für 100 Stichproben.
Abweichung der Mittelwerte bzw. Häufigkeitsverteilungen: ähnliche Merkmale 1% der merkmalsspezifischen Streuung, unähnliche 25%.
[a] In die Analyse einbezogene Variablen; Skalenniveau: 1-5 metrisch, 6-9 dichotom, 10-12 polytom;
[b] Durchschnitttliche Ausprägung des Merkmals in den Teilnehmerstichproben (T), Nichtteilnehmerstichproben (NT) bzw. den Kontrollgruppen (C);
[c] Skalenspezifische Tests (metrische Variablen: Wilcoxontest, dichotome: McNemartest, polytome: χ^2-Test), durchschnittliche Ablehnungsrate der Nullhypothese gleicher Mittelwerte bei einem Signifikanzniveau von 5%;
[d] Prozentuale Reduzierung der Abweichung der Merkmalsmittelwerte durch Matching.

Tabelle 4.3: **Ergebnisse für Stichproben mit unähnlichen nominalen Variablen (bei ähnlichen metrischen Variablen)**

Ausgangsdaten			Propensity Score			Mahalanobisdistanz			Mahalanobis-Matching			Gowerdistanz		
X[a]	T[b]	NT[b]	C[b]	Test[c]	BiasR[d]	C[b]	Test[c]	BiasR[d]	C[b]	Test[c]	BiasR[d]	C[b]	Test[c]	BiasR[d]
1	40,05	40,08	37,98	0,58	-87,46	40,07	0,16	54,37	40,12	0,16	15,80	40,06	0,05	82,08
2	0,98	1,05	0,92	0,24	-30,25	1,00	0,13	47,96	1,01	0,18	23,86	0,98	0,05	81,10
3	12,03	12,04	11,36	0,57	-69,51	12,01	0,22	56,91	12,07	0,19	21,39	12,01	0,15	78,91
4	11,96	12,42	10,78	0,44	-60,76	12,07	0,20	48,96	12,38	0,22	2,35	11,93	0,06	80,80
5	1302,72	1343,39	1182,57	0,46	-68,08	1308,43	0,20	54,86	1348,28	0,22	-11,18	1303,41	0,09	75,96
6	0,50	0,48	0,58	0,49	-57,32	0,50	0,02	77,72	0,49	0,00	89,35	0,48	0,15	2,16
7	0,66	0,65	0,69	0,44	-36,72	0,67	0,05	75,55	0,66	0,01	87,91	0,66	0,20	0,47
8	0,90	0,89	0,92	0,27	-3,41	0,91	0,00	70,98	0,93	0,14	51,59	0,89	0,12	-9,76
9	0,20	0,20	0,31	0,53	-107,78	0,20	0,00	79,79	0,18	0,05	71,97	0,20	0,12	4,12
10	2,84	2,92	2,60	0,88	-215,19	2,92	0,35	-0,42	2,88	0,00	49,07	2,93	0,68	-32,40
11	2,36	2,49	2,09	0,91	-95,59	2,37	0,52	64,88	2,36	0,00	81,69	2,50	0,85	-9,40
12	3,16	3,17	3,02	0,79	-386,00	3,22	0,48	-107,85	3,22	0,00	-62,65	3,18	0,61	-126,16

Anmerkungen:
Durchschnittsergebnisse für 100 Stichproben.
Abweichung der Mittelwerte bzw. Häufigkeitsverteilungen: ähnliche Merkmale 1% der merkmalsspezifischen Streuung, unähnliche 25%.
[a] In die Analyse einbezogene Variablen; Skalenniveau: 1-5 metrisch, 6-9 dichotom, 10-12 polytom;
[b] Durchschnitttliche Ausprägung des Merkmals in den Teilnehmerstichproben (T), Nichtteilnehmerstichproben (NT) bzw. den Kontrollgruppen (C);
[c] Skalenspezifische Tests (metrische Variablen: Wilcoxontest, dichotome: McNemartest, polytome: χ^2-Test), durchschnittliche Ablehnungsrate der Nullhypothese gleicher Mittelwerte bei einem Signifikanzniveau von 5%;
[d] Prozentuale Reduzierung der Abweichung der Merkmalsmittelwerte durch Matching.

Die generierten Abweichungen der Variablen sind in beiden Stichprobendesigns nicht an den Mittelwerten der durchschnittlichen Merkmalsausprägungen erkennbar. Dies erklärt sich daraus, dass es sich um Durchschnittswerte über 100 Stichproben handelt und bei der Generierung der Nichtteilnehmerstichproben zufällig festgelegt wird, ob eine positive oder ein negative Abweichung vom Mittelwert in der Teilnehmerstichprobe erzeugt wird. Beide Fälle sind damit gleich wahrscheinlich, d.h. im Mittel unterscheiden sich die Merkmalsmittelwerte nicht.

Sichtbar werden Veränderungen im Vergleich zum Ausgangsdesign dagegen im Bias vor Matching (vgl. dazu die Tabellen im Anhang B.2). Für die Stichproben mit unähnlichen metrischen Variablen ist ungefähr eine Verdopplung des Bias vor Matching zu beobachten. Unterschiede in den polytomen Variablen sind dagegen kaum zu erkennen.[35] Für die Bias Reduzierung durch Matching werden nur geringe Veränderungen über die verschiedenen Stichprobendesigns festgestellt.

Die Betrachtung der Ergebnisse der skalenspezifischen Tests liefert in beiden Stichprobendesigns eine sehr ähnliche Einschätzung der Güte der einzelnen Distanzmaße wie für das Ausgangsdesign. Für Propensity Score und Index Score sind deutlich höhere Ablehnungsraten der Nullhypothese gleicher Mittelwerte bzw. Häufigkeitsverteilungen festzustellen als für Mahalanobisdistanz, gewichtete Mahalanobis-Matching-Distanz und das Distanzmaß nach Gower (bei unähnlichen metrischen Variablen durchschnittlich 45% im Vergleich zu ca. 40%, 30% und 35%; bei unähnlichen nominalen Variablen ca. 55% im Vergleich zu 20%, 10% und 26%).
Auch die Berücksichtigung der einzelnen Skalenniveaus zeigt ein ähnliches Muster wie im Ausgangsdesign. Die Angleichung der polytomen Variablen ist problematisch mit der Mahalanobisdistanz und – in besonderem Maße – mit dem Distanzmaß nach Gower. Die durchschnittlichen Ablehnungsraten liegen bei Variablen dieses Skalenniveaus bei 65% bzw. 70%. Für das Distanzmaß nach Gower ist ebenfalls eine deutlich schlechtere Angleichung der dichotomen Variablen als bei Mahalanobisdistanz

[35] Vermutlich ist dies mit der Konstruktion des Bias als Differenz von Mittelwerten zu erklären, mit der Abweichungen in der Häufigkeitsverteilung nicht adäquat identifiziert werden können. Der Bias vor und nach Matching sowie die Bias Reduzierung werden deshalb nur als ergänzende Information zur Güte der Ergebnisse behandelt. Das Hauptaugenmerk für die Beurteilung der Ergebnisse liegt auf den Ergebnissen der skalenspezifischen Mittelwert- bzw. Häufigkeitsverteilungstests.

und gewichteter Mahalanobis-Matching-Distanz festzustellen. Die Angleichung der polytomen Variablen gelingt mit dem letztgenannten Distanzmaß dagegen sehr gut – in beiden Stichprobendesigns sind die Häufigkeitsverteilungen dieser Variablen in Teilnehmer- und Kontrollgruppe in allen 100 Stichproben gleich; die durchschnittliche Ablehnungsrate liegt bei 0%. Allerdings gelingt die Angleichung der Merkmalsmittelwerte der metrischen Variablen mit der Mahalanobis-Matching-Distanz relativ schlecht, wogegen das Distanzmaß nach Gower in diesem Skalenniveau die beste Angleichung erzielt. Die Ergebnisse für die Mahalanobisdistanz liegen für alle Skalenniveaus zwischen denen der beiden aggregierten Distanzmaße.

Der Vergleich der Ergebnisse beider Stichprobendesigns macht wichtige Unterschiede in den Fähigkeiten der einzelnen Distanzmaße deutlich. Für die Stichproben mit unähnlichen metrischen Merkmalen ist eine deutliche Verschlechterung der Ergebnisse im Vergleich zum Ausgangsdesign festzustellen, die für Mahalanobisdistanz und Mahalanobis-Matching-Distanz besonders stark ausfällt. Die durchschnittliche Ablehnungrate der Nullhypothese für metrische Merkmale liegt für beide Maße bei ca. 65% und ist damit sogar höher als die Ablehnungsraten für Propensity Score und Index Score. Dem Distanzmaß nach Gower gelingt dagegen die Angleichung der Merkmalsmittelwerte der metrischen Variablen auch bei relativ unterschiedlichen Ausgangsdaten – die durchschnittliche Ablehnungsrate liegt hier bei ca. einem Drittel. Unähnliche nominale Variablen verändern die beschriebenen Ergebnisse des Ausgangsdesigns dagegen nur in sehr geringem Maße.

Insgesamt – für beide Stichprobendesigns und über alle Skalenniveaus – verändert sich die Rangfolge der Distanzmaße nicht, allerdings verringert sich der „Vorsprung“ der Mahalanobis-Matching-Distanz gegenüber dem Distanzmaß nach Gower, besonders im Stichprobendesign unähnlicher metrischer Variablen. Für das letztgenannte Distanzmaß ist darüber hinaus eine gleichmäßigere Angleichung der Merkmalsmittelwerte über alle Skalenniveaus zu beobachten.

Die folgende Tabelle 4.4 enthält die Ergebnisse für Stichproben mit unähnlichen Merkmalen. Die Abweichung der Merkmalsmittelwerte bzw. der Häufigkeitsverteilungen der Merkmalsausprägungen beträgt 25% der merkmalsspezifischen Streuung.

Tabelle 4.4: **Ergebnisse für Stichproben mit unähnlichen metrischen und nominalen Variablen**

Ausgangsdaten			Propensity Score			Mahalanobisdistanz			Mahalanobis-Matching			Gowerdistanz		
X[a]	T[b]	NT[b]	C[b]	Test[c]	BiasR[d]	C[b]	Test[c]	BiasR[d]	C[b]	Test[c]	BiasR[d]	C[b]	Test[c]	BiasR[d]
1	40,06	40,22	38,93	0,38	23,77	40,14	0,65	56,77	40,30	0,76	13,35	40,09	0,21	83,94
2	0,98	1,06	0,97	0,24	34,57	1,01	0,53	51,41	1,03	0,57	28,79	0,98	0,25	81,19
3	12,03	12,15	11,65	0,42	38,49	12,07	0,69	59,53	12,13	0,73	30,18	12,03	0,48	79,83
4	11,94	12,45	11,38	0,39	24,45	12,07	0,51	57,56	12,49	0,64	13,86	11,98	0,25	81,01
5	1301,08	1370,79	1232,12	0,50	28,79	1324,59	0,64	57,30	1378,85	0,69	1,60	1309,36	0,37	81,14
6	0,50	0,50	0,53	0,14	16,15	0,50	0,07	72,11	0,50	0,00	89,16	0,50	0,25	-13,26
7	0,65	0,64	0,69	0,24	1,72	0,67	0,06	75,40	0,67	0,00	84,88	0,65	0,14	1,52
8	0,90	0,89	0,92	0,17	17,24	0,92	0,01	68,44	0,93	0,14	52,35	0,90	0,15	-2,27
9	0,20	0,20	0,25	0,25	-20,19	0,20	0,02	77,23	0,18	0,07	70,35	0,20	0,13	-4,35
10	2,84	2,91	2,77	0,83	-138,45	2,91	0,52	-16,72	2,88	0,00	32,83	2,91	0,69	-27,97
11	2,36	2,49	2,24	0,83	-11,42	2,40	0,61	54,95	2,36	0,00	78,21	2,49	0,79	-2,25
12	3,16	3,17	3,12	0,84	-350,86	3,23	0,42	-135,21	3,22	0,00	-64,77	3,18	0,58	-123,65

Anmerkungen:
Durchschnittsergebnisse für 100 Stichproben.
Abweichung der Mittelwerte bzw. Häufigkeitsverteilungen: 1% der merkmalsspezifischen Streuung.
[a] In die Analyse einbezogene Variablen; Skalenniveau: 1-5 metrisch, 6-9 dichotom, 10-12 polytom;
[b] Durchschnitttliche Ausprägung des Merkmals in den Teilnehmerstichproben (T), Nichtteilnehmerstichproben (NT) bzw. den Kontrollgruppen (C);
[c] Skalenspezifische Tests (metrische Variablen: Wilcoxontest, dichotome: McNemartest, polytome: χ^2-Test), durchschnittliche Ablehnungsrate der Nullhypothese gleicher Mittelwerte bei einem Signifikanzniveau von 5%;
[d] Prozentuale Reduzierung der Abweichung der Merkmalsmittelwerte durch Matching.

Die Ergebnisse für die so generierten Stichproben zeichnen ein ähnliches Bild der Fähigkeiten der einzelnen Distanzmaße wie die vorangegangenen Analyseschritte. Besonders auffällig sind auch hier die konträren Ergebnisse für die gewichtete Mahalanobis-Matching-Distanz und das Distanzmaß nach Gower. Den schlechten Testergebnissen des erstgenannten Maßes hinsichtlich der metrischen Variablen – mit einer durchschnittlichen Ablehnungsrate von ca. 65% – stehen sehr geringe Ablehnungsraten (von 0,5% bzw. 0%) bei den nominalen Variablen gegenüber. Das Gegenteil gilt für das Distanzmaß nach Gower, für das die durchschnittlichen Ablehnungsraten der metrischen Variablen im Mittel bei 30%, die der polytomen Variablen bei knapp 70% liegen. Die Mahalanobisdistanz liegt bei der skalenspezifischen Betrachtung wieder zwischen beiden aggregierten Maßen, schneidet im vorliegenden Stichprobendesign insgesamt aber schlechter ab. Interessant ist die Verbesserung der Ergebnisse von Propensity Score und Index Score in Relation zu den drei anderen Maßen. Die Nullhypothese gleicher Mittelwerte bzw. Häufigkeitsverteilungen wird durchschnittlich in 44% der Stichproben abgelehnt – im Vergleich zu 39%, 36% bzw. 30% im Falle der Mahalanobisdistanz, des Distanzmaßes nach Gower sowie der gewichteten Mahalanobis-Matching-Distanz.

Zusammenfassend für alle Stichprobendesigns kann ein wiederkehrendes Muster für die Güte der Distanzmaße hinsichtlich ihrer Fähigkeit zur Angleichung der Merkmalsverteilungen (Merkmalsmittelwerte und Häufigkeitsverteilungen der Ausprägungen) verschieden skalierter Variablen festgestellt werden. Der Propensity Score und der Index Score sind in allen untersuchten Designs deutlich schlechter als die drei anderen Distanzmaße in der Lage, Ähnlichkeiten bzw. Unterschiede in unterschiedlich skalierten Merkmalen adäquat abzubilden. Im Durchschnitt über alle vier Stichprobendesigns liegen die Ablehnungsraten der skalenspezifischen Tests bei beiden Scores bei ca. 50% gegenüber 30% bei Mahalanobisdistanz und Distanzmaß nach Gower und ca. 20% im Falle der gewichteten Mahalanobis-Matching-Distanz. Die Ergebnisse für Propensity Score und Index Score legen die Vermutung nahe, dass die Nutzung parametrischer Modelle zur Schätzung von Distanzmaßen und die implizite Gewichtung der einbezogenen Merkmale in Abhängigkeit ihres Einflusses auf die Teilnahmewahrscheinlichkeit in kleinen Stichproben problematisch ist. Sie

bestätigen die in Fröhlich (2004b) geäußerten Vorbehalte gegen die Nutzung von Propensity Scores in kleinen Stichproben.

Die Simulationsergebnisse stimmen mit der Aussage in Zhao (2004) hinsichtlich der – im Vergleich zum Propensity Score – besseren Eignung der Mahalanobisdistanz zur Wiedergabe der Unterschiede in den einzelnen Merkmalen in kleinen Stichproben überein.[36] Die festgestellten Unterschiede zwischen den Ergebnissen für Propensity Score und Index Score sind – anders als in Augurzky (2000a) – sehr gering; die bessere Unterscheidungsfähigkeit des Index Scores zwischen Personen an den „Rändern“ der Verteilung kann in den Simulationsergebnissen nicht beobachtet werden.

Die Berücksichtigung der Skalenniveaus der Variablen gibt zusätzlichen Einblick in die Vorzüge und Nachteile der untersuchten aggregierten Distanzmaße. Das Distanzmaß nach Gower ist nicht bzw. nur bedingt geeignet zur Angleichung der Verteilung polytomer Variablen. Auch die Ergebnisse für die dichotomen Variablen sind schlechter im Vergleich zu Mahalanobisdistanz und Mahalanobis-Matching-Distanz. Die „Stärke“ dieses Distanzmaßes liegt bei metrischen Variablen. Nahezu das Gegenteil gilt für die Mahalanobis-Matching-Distanz. Während die Angleichung nominaler (dichotomer und vor allem polytomer) Variablen gut gelingt, tritt die „Schwäche“ dieses Maßes besonders deutlich hervor, wenn die metrischen Merkmale unähnlich sind (vgl. Tabellen 4.2 und 4.4). Die Testergebnisse der Mahalanobisdistanz liegen zwischen denen der beiden genannten Maße.
Von den analysierten Distanzmaßen scheint die gewichtete Mahalanobis-Matching-Distanz am besten in der Lage zu sein, in verschiedenen Stichprobendesigns Ähnlichkeiten und Unterschiede in den einzelnen Variablen adäquat wiederzugeben. Berücksichtigt man allerdings die deutliche Verschlechterung der Ergebnisse, die bei einer Einschränkung der Übereinstimmung der metrischen Variablen eintritt, kann die Anwendung dieses Distanzmaßes nicht uneingeschränkt für die empfohlen werden.

[36] Eine ähnliche Aussage für eine geringe Anzahl von Merkmalen (5) findet sich auch in Gu und Rosenbaum (1993), allerdings wird in der gleichen Studie für eine große Anzahl Variablen (20) das Gegenteil festgestellt. Es ist möglich, dass die Anzahl der in der Simulationsstudie verwendeten Variablen (12) noch „klein genug“ ist, so dass die Ergebnisse auch keinen Widerspruch zu den Aussagen in der genannten Studie darstellen.

Die Analyseergebnisse lassen dagegen eine Kombination aus Mahalanobis-Matching-Distanz und dem Distanzmaß nach Gower interessant erscheinen. Betrachtet man die „Bestandteile“ dieser Distanzmaße, fällt auf, dass die Verwendung des verallgemeinerten Matchingkoeffizienten für alle nominalen Variablen gemeinsam – wie bei der Mahalanobis-Matching-Distanz – ein gutes Maß für die Abbildung von Ähnlichkeiten und Unterschieden in nominalen Variablen zu sein scheint.[37] Für die Feststellung von Übereinstimmungen und Unterschieden in metrischen Variablen scheint die normierte absolute Merkmalsdifferenz, die im Distanzmaß nach Gower verwendet wird, eine bessere Alternative zur Mahalanobisdistanz darzustellen. In einer weiterführenden Untersuchung wäre zu prüfen, ob eine Kombination aus verallgemeinertem Matchingkoeffizienten der nominalen Variablen und der normierten absoluten Merkmalsdifferenz metrischer Variablen tatsächlich die Stärken der beiden analysierten aggregierten Distanzmaße in sich vereint.

Die im Vorfeld der Analyse aufgestellte Hypothese, nach der die Mahalanobisdistanz und ein aggregiertes Distanzmaß besser in der Lage sind, Ähnlichkeiten und Unterschiede hinsichtlich der Determinanten des Arbeitsmarkterfolgs wiederzugeben als Propensity Score und Index Score, wird von den Simulationsergebnissen bestätigt.

Nach der zweiten Hypothese gelingt die Gewichtung der nominalen Variablen für die Gesamtdistanz entsprechend ihrer Bedeutung mit einem aggregierten Distanzmaß besser als mit der Mahalanobisdistanz. Diese Hypothese beinhaltet zwei Aspekte. Sie drückt die Erwartung aus, dass die „richtige“ Wiedergabe der Unterschiede in nominalen Variablen in einer geringeren Abweichung der Ausprägungen dieser Variablen nach dem Matching resultiert. Darüber hinaus sollte mit einem aggregierten Distanzmaß eine gleichmäßigere Verteilung der „Restabweichung“ über alle Variablen möglich sein, weil alle Variablen gleichermaßen – entsprechend ihres Skalenniveaus – in der Gesamtdistanz berücksichtigt werden. Die erwartete bessere Angleichung nominaler Variablen kann für die gewichtete Mahalanobis-Matching-Distanz bezüglich polytomer Variablen bestätigt werden. Bei dichotomen Variablen sind allerdings keine großen Unterschiede zwischen Mahalanobisdistanz und Mahalanobis-Matching-Distanz festzustellen. Dem Distanzmaß nach Gower dagegen gelingt die

[37] Die Mahalanobisdistanz und die gesonderte Betrachtung jedes einzelnen Merkmals – wie im Distanzmaß nach Gower – sind dagegen nicht so gut geeignet.

Angleichung der nominalen Variablen schlechter als der Mahalanobisdistanz. Die gleichmäßigere Verteilung der nach Matching noch bestehenden Unterschiede über alle Skalenniveaus (im Vergleich zur Mahalanobisdistanz) wird dagegen nur mit dem Distanzmaß nach Gower realisiert.

4.3.2 Analyse der Zuordnungsprozesse

Für die Analyse der Zuordnungsprozesse werden drei verschiedene Teilnehmerstichprobengrößen generiert (Stichproben mit 50, 100 und 300 Personen) und mit unterschiedlichen Nichtteilnehmerstichproben kombiniert. Die resultierenden Gesamtstichproben unterscheiden sich zum einen durch das vorgegebene Größenverhältnis, zum anderen durch eine unterschiedlich große Übereinstimmung der Merkmalsausprägungen in den Teilstichproben (Common-Support-Bereich).

In den folgenden Tabellen werden die wichtigsten Ergebnisse der Analyse präsentiert. Es werden Informationen über den Bias, die empirische Standardabweichung, die Wurzel des mittleren quadratischen Fehlers (RMSE) sowie die Summe der quadrierten Distanzen für die einzelnen Zuordnungsprozesse gegeben. Der Bias wird definiert als Abweichung des durchschnittlichen geschätzten vom durchschnittlichen „wahren“ Nichtteilnahmeeinkommen. Er ist ein Maß für die Fähigkeit der untersuchten Matchingverfahren, den „wahren“ Wert des Vergleichseinkommens mit den gegebenen Daten abzubilden.[38] Die empirische Standardabweichung wird als Wurzel der empirischen Varianz ermittelt und an ihrer Stelle ausgewiesen, da sie aufgrund ihrer Größe besser mit den anderen Kriterien vergleichbar ist. Sie ist ein Maß für die Streubreite der einzelnen Schätzwerte. Beide Größen, Bias und Varianz, sind Bestandteile des mittleren quadratischen Fehlers (MSE), dessen Wurzel (RMSE) ebenfalls in den Tabellen präsentiert wird.[39] Die Summe der quadrierten Distanzen gibt Auskunft über die Abweichung der Teilnehmer- und der Kontrollgruppe im verwendeten Distanzmaß. Unter der Annahme, dass das verwendete Distanzmaß

[38] Die Tabellen B.7 bis B.15 im Anhang B.3 enthalten zusätzlich Informationen über den Bias vor Matching und die Bias Reduzierung durch Matching.

[39] In den Tabellen B.7 bis B.15 werden zusätzlich der mittlere quadratische Fehler (MSE), die empirische Varianz und der prozentuale Anteil der Varianz am MSE für alle untersuchten Zuordnungsprozesse ausgewiesen.

die Unterschiede in den einzelnen Variablen adäquat zusammenfasst, kann sie als Näherungswert für die Übereinstimmung der Teilstichproben in der Verteilung der Merkmale betrachtet werden.

Zusätzlich dazu findet sich in der Tabelle 4.5 eine Information über die von der Analyse ausgeschlossenen Teilnehmer. Es wird der Anteil der ausgeschlossenen Personen an der Gesamtzahl der Teilnehmer in der Ausgangsstichprobe ausgewiesen. Da das Problem des Ausschlusses von Teilnehmern in der Simulation nur in Stichproben mit gleicher Teilnehmer- und Nichtteilnehmeranzahl auftritt, wird die entsprechende Zeile in den darauf folgenden beiden Tabellen nicht mehr ausgewiesen.

Neben den Gütemaßen ist auch das „wahre“ Nichtteilnahmeeinkommen, das anhand der im Abschnitt 4.2.1 beschriebenen Variablen für die Teilnehmer definiert wird, aus den Tabellen ersichtlich. Diese Größe stimmt für alle analysierten Zuordnungsprozesse überein, da für jeden Zuordnungsprozess dieselben Teilnehmerstichproben verwendet werden.[40] Ebenfalls in den Tabellen zu finden ist das geschätzte Nichtteilnahmeeinkommen. Diese Größe entspricht dem durchschnittlichen Einkommen in der durch Matching erzeugten Kontrollgruppe. Beide Größen sollen die Anschaulichkeit der Ergebnisse erhöhen.[41]

Die folgende Tabelle 4.5 fasst die Simulationsergebnisse für gleich große Teilnehmer- und Nichtteilnehmerstichproben zusammen.[42]

[40] Eine Ausnahme bildet das „wahre“ Einkommen der Teilnehmer beim Random Matching in der Tabelle 4.5, das vom Einkommen der beiden anderen Zuordnungsprozesse abweicht. Das erklärt sich aus dem Ausschluss von Teilnehmern bei diesem Zuordnungsprozess. Im Zusammenhang mit der Auswertung der Tabelle 4.5 wird dieses Problem näher erläutert.

[41] Das gleiche gilt für die in den Tabellen B.7 bis B.15 aufgeführten Angaben über das Teilnahmeeinkommen und das Nichtteilnahmeeinkommen in der Ausgangsstichprobe der Nichtteilnehmer (Nichtteilnahmeeinkommen vor Matching).

[42] Die Tabellen B.7, B.8 und B.9 im Anhang B.3 enthalten ergänzende Informationen zu den präsentierten Ergebnissen.

Tabelle 4.5: **Ergebnisse für Stichproben mit einem Größenverhältnis von 1:1**

	Ridge Matching	Zuordnung m. Zurücklegen	Random Matching
Stichproben mit je 50 Personen			
ähnliche Merkmale			
$Y_{NT_{wahr}}$ [a]	2653,58	2653,58	2651,04
$\hat{Y}_{NT}$ [b]	2681,79	2702,50	2675,16
Bias	-28,21	-48,92	-24,13
emp. Std.-abw. [c]	800,30	265,46	262,15
RMSE [d]	800,80	269,93	263,26
$\sum$ quad. Dist. [e]	–	0,58	1,88
ausgeschl. T [f]	0,00	0,00	0,03
eingeschränkt ähnliche Merkmale			
$Y_{NT_{wahr}}$ [a]	2665,95	2665,95	2665,24
$\hat{Y}_{NT}$ [b]	2883,23	2657,38	2662,82
Bias	-217,28	8,57	2,42
emp. Std.-abw. [c]	768,86	254,59	263,43
RMSE [d]	798,97	254,74	263,44
$\sum$ quad. Dist. [e]	–	0,62	1,76
ausgeschl. T [f]	0,00	0,00	0,02
unähnliche Merkmale			
$Y_{NT_{wahr}}$ [a]	2689,18	2689,18	2687,26
$\hat{Y}_{NT}$ [b]	2937,66	2798,20	2760,25
Bias	-248,48	-109,02	-72,99
emp. Std.-abw. [c]	960,52	297,73	297,79
RMSE [d]	992,14	317,06	306,60
$\sum$ quad. Dist. [e]	–	0,61	1,58
ausgeschl. T [f]	0,00	0,00	0,01
Stichproben mit je 100 Personen			
ähnliche Merkmale			
$Y_{NT_{wahr}}$	2682,13	2682,13	2681,16
$\hat{Y}_{NT}$ [b]	2723,89	2700,85	2702,68
Bias	-41,77	-18,72	-21,52
emp. Std.-abw. [c]	483,05	160,98	174,11
RMSE [d]	484,85	162,07	175,43
$\sum$ quad. Dist. [e]	–	0,76	5,06
ausgeschl. T [f]	0,00	0,00	0,03
eingeschränkt ähnliche Merkmale			
$Y_{NT_{wahr}}$ [a]	2676,76	2676,76	2676,38
$\hat{Y}_{NT}$ [b]	2917,53	2701,73	2684,34
Bias	-240,77	-24,98	-7,96
emp. Std.-abw. [c]	475,67	184,32	192,59
RMSE [d]	533,14	186,00	192,76
$\sum$ quad. Dist. [e]	–	0,73	2,76
ausgeschl. T [f]	0,00	0,00	0,01

Fortsetzung siehe nächste Seite

Fortsetzung Tabelle 4.5

	Ridge Matching	Zuordnung m. Zurücklegen	Random Matching
unähnliche Merkmale			
$Y_{NT_{wahr}}$ [a]	2665,16	2665,16	2667,91
$\hat{Y}_{NT}$ [b]	2932,12	2692,69	2665,32
Bias	-266,96	-27,52	2,59
emp. Std.-abw. [c]	621,18	236,71	234,93
RMSE [d]	676,11	238,30	234,94
$\sum$ quad. Dist. [e]	–	0,77	2,87
ausgeschl. T [f]	0,00	0,00	0,01
Stichproben mit je 300 Personen			
ähnliche Merkmale			
$Y_{NT_{wahr}}$ [a]	2694,71	2694,71	2695,24
$\hat{Y}_{NT}$ [b]	2697,61	2715,31	2712,19
Bias	-2,90	-20,60	-16,95
emp. Std.-abw. [c]	256,51	85,63	101,13
RMSE [d]	256,52	88,07	102,54
$\sum$ quad. Dist. [e]	–	1,09	11,72
ausgeschl. T [f]	0,00	0,00	0,00
eingeschränkt ähnliche Merkmale			
$Y_{NT_{wahr}}$ [a]	2692,13	2692,13	2692,01
$\hat{Y}_{NT}$ [b]	2788,32	2759,45	2743,05
Bias	-96,19	-67,32	-51,05
emp. Std.-abw. [c]	295,43	133,90	138,58
RMSE [d]	310,70	149,87	147,69
$\sum$ quad. Dist. [e]	–	1,08	6,77
ausgeschl. T [f]	0,00	0,00	0,00
unähnliche Merkmale			
$Y_{NT_{wahr}}$ [a]	2693,17	2693,17	2693,43
$\hat{Y}_{NT}$ [b]	2785,28	2808,96	2801,45
Bias	-92,11	-115,79	-108,02
emp. Std.-abw. [c]	457,11	180,81	197,69
RMSE [d]	466,30	214,70	225,28
$\sum$ quad. Dist. [e]	–	1,11	6,19
ausgeschl. T [f]	0,00	0,00	0,00

Anmerkungen:

Durchschnittsergebnisse aus 100 Stichproben.
Abweichung der Mittelwerte bzw. Häufigkeitsverteilungen:
ähnliche Merkmale 1% der merkmalsspezifischen Streuung,
eingeschränkt ähnliche 10%, unähnliche 25%.

[a] „wahres“ Nichtteilnahmeeinkommen;
[b] geschätztes Nichtteilnahmeeinkommen als Durchschnitt der Einkommen in der Kontrollgruppe;
[c] empirische Standardabweichung;
[d] Wurzel des mittleren quadratischen Fehlers;
[e] Summe der quadrierten Distanzen;
[f] Anteil der ausgeschlossenen Teilnehmer an der Gesamtzahl der Teilnehmer.

Die präsentierten Ergebnisse beinhalten – je nach betrachtetem Gütemaß – unterschiedliche Aussagen über die Qualität der Matchingergebnisse. So liefert Random Matching hinsichtlich der Abweichung des geschätzten zum „wahren“ Nichtteilnahmeeinkommen (Bias) die besten Ergebnisse, d.h. die geringste Abweichung. Allerdings ist für diesen Zuordnungsprozess ein Verlust von Teilnehmern zu verzeichnen. Bei der Zuordnung mit Zurücklegen tritt das Problem des Verlusts von Beobachtungen dagegen nicht auf. Darüber hinaus ist die Summe der quadrierten Distanzen in allen untersuchten Stichproben geringer. Auch hinsichtlich der Streuung der Schätzwerte um ihren Mittelwert (empirische Standardabweichung) ist die Zuordnung mit Zurücklegen das bessere Verfahren. Betrachtet man allerdings den mittleren quadratischen Fehler, ist kein klarer Vorzug gegenüber dem Random Matching festzustellen.

Deutlicher ist die Aussage der Ergebnisse für Ridge Matching. Bei diesem Algorithmus tritt ebenfalls kein Verlust von Teilnehmern auf, was im Vergleich zum Random Matching einen Vorteil darstellt. Allerdings sind eine vergleichsweise große Abweichung des geschätzten zum „wahren“ Nichtteilnahmeeinkommen ebenso wie eine größere Streuung als bei beiden anderen Zuordnungsprozessen festzustellen. Die Summe der quadrierten Distanzen ist für dieses Verfahren nicht verfügbar, da die Vergleichsgröße direkt aus mehreren unterschiedlich gewichteten Nichtteilnehmereinkommen generiert wird – ohne den Zwischenschritt der Konstruktion einer eigenen Kontrollgruppe für jeden Teilnehmer.[43]

Die Entscheidung, welcher Algorithmus besser für annähernd gleich große Teilnehmer- und Nichtteilnehmerstichproben geeignet ist, fällt also zwischen Zuordnung mit Zurücklegen und dem Random Matching. Bei homogenen Teilnehmerstichproben – wenn also keine Verzerrung des Ergebnisses durch den Verlust von Beobachtungen auftritt – könnte Random Matching angewendet werden. Da in empirischen Analysen aber eher selten homogene Stichproben vorausgesetzt werden können, legen die Simulationsergebnisse die Nutzung von Zuordnungen mit Zurücklegen nahe.

Die Simulationsergebnisse bestätigen die in der Literatur geäußerten Vorbehalte gegen die Anwendung von Matchingverfahren ohne Zurücklegen – zu denen das

[43] Eine detaillierte Beschreibung dieses Verfahrens findet sich im Abschnitt 4.2.2.

Random Matching zählt – auf annähernd gleich große Teilnehmer- und Nichtteilnehmerstichproben.[44]

Im Folgenden werden Ergebnisse für Stichproben, in denen mehr Nichtteilnehmer als Teilnehmer vorhanden sind, präsentiert. Für dieses Design werden zusätzlich zu den drei o.g. Zuordnungsprozessen Ergebnisse für ein Optimal Full Matching auf Grundlage eines Auktionsalgorithmus' und ein optimales Nearest Neighbor Matching auf Grundlage des Ungarischen Algorithmus' dargestellt. In der Tabelle 4.6 wird zunächst ein Größenverhältnis von Teilnehmeranzahl zu Nichtteilnehmeranzahl von 1:3 unterstellt.

Die oben getroffenen Aussagen über die Abhängigkeit der Beurteilung eines Zuordnungsprozesses vom betrachteten Gütemaß bestätigen sich auch in den Ergebnissen der Stichproben mit einem Größenverhältnis von 1:3. Eine Ausnahme bildet das Ridge Matching, das hinsichtlich aller verfügbaren Kriterien in den meisten Stichprobendesigns deutlich schlechter ist als die vier anderen Algorithmen (die Zuordnung mit Zurücklegen, ein optimales Nearest Neighbor Matching, Optimal Full Matching und Random Matching).

Für die Zuordnung mit Zurücklegen ist – wie für die gleich großen Stichproben – in allen Stichprobendesigns die geringste Summe der quadrierten Distanzen festzustellen. Am schlechtesten in diesem Kriterium ist Optimal Full Matching, dazwischen liegen Random Matching und Optimal Nearest Neighbor Matching mit sehr ähnlichen Werten. Eine umgekehrte Rangfolge ergibt sich bei Betrachtung des mittleren quadratischen Fehlers (bzw. des RMSE). Hier sind in den meisten Stichprobendesigns die besten Werte für das Optimal Full Matching, die schlechtesten für die Zuordnung mit Zurücklegen zu verzeichnen. Dazwischen liegen Optimal Nearest Neighbor Matching und Random Matching, wieder mit sehr ähnlichen Werten.

[44] Der Verlust von Beobachtungen wird vor allem in der anwendungsorientierten Literatur thematisiert (Lechner, 2001b), aber auch in theoretischen Studien nachgewiesen (Augurzky, 2000a).

Tabelle 4.6: **Ergebnisse für Stichproben mit einem Größenverhältnis von 1:3**

	Ridge Matching	Zuordnung m. Zurücklegen	opt.1:1 Matching	opt. Full Matching	Random Matching
Stichproben mit 50 Teilnehmern und 150 Nichtteilnehmern					
ähnliche Merkmale					
$Y_{NT_{wahr}}{}^{a}$	2675,75	2675,75	2675,75	2675,75	2675,75
$\hat{Y}_{NT}{}^{b}$	2657,14	2722,82	2717,17	2700,12	2714,03
Bias	18,61	-47,06	-41,42	-24,37	-38,28
emp. Std.-abw.^{c}	581,01	225,85	201,91	190,69	208,89
RMSE^{d}	581,31	230,70	206,11	192,24	212,37
$\sum$ quad. Dist.^{e}	–	0,40	0,45	0,62	0,49
eingeschränkt ähnliche Merkmale					
$Y_{NT_{wahr}}{}^{a}$	2679,64	2679,64	2679,64	2679,64	2679,64
$\hat{Y}_{NT}{}^{b}$	2753,41	2773,72	2770,26	2731,36	2773,40
Bias	-73,77	-94,08	-90,62	-51,72	-93,76
emp. Std.-abw.^{c}	563,79	234,05	229,64	228,37	219,53
RMSE^{d}	568,60	252,25	246,88	234,16	238,71
$\sum$ quad. Dist.^{e}	–	0,41	0,46	0,64	0,50
unähnliche Merkmale					
$Y_{NT_{wahr}}{}^{a}$	2674,75	2674,75	2674,75	2674,75	2674,75
$\hat{Y}_{NT}{}^{b}$	2937,34	2721,62	2716,39	2690,29	2722,31
Bias	-262,59	-46,86	-41,64	-15,54	-47,56
emp. Std.-abw.^{c}	752,88	276,95	269,61	274,22	262,92
RMSE^{d}	797,36	280,89	272,81	274,66	267,18
$\sum$ quad. Dist.^{e}	–	0,42	0,47	0,66	0,52
Stichproben mit 100 Teilnehmern und 300 Nichtteilnehmern					
ähnliche Merkmale					
$Y_{NT_{wahr}}{}^{a}$	2688,03	2688,03	2688,03	2688,03	2688,03
$\hat{Y}_{NT}{}^{b}$	2656,61	2664,34	2667,92	2660,59	2674,27
Bias	31,42	23,68	20,11	27,44	13,75
emp. Std.-abw.^{c}	429,55	131,82	121,35	126,52	124,85
RMSE^{d}	430,70	133,93	123,00	129,46	125,60
$\sum$ quad. Dist.^{e}	–	0,42	0,48	0,71	0,51
eingeschränkt ähnliche Merkmale					
$Y_{NT_{wahr}}{}^{a}$	2689,23	2689,23	2689,23	2689,23	2689,23
$\hat{Y}_{NT}{}^{b}$	2704,84	2757,10	2751,52	2737,96	2757,98
Bias	-15,61	-67,87	-62,28	-48,73	-68,74
emp. Std.-abw.^{c}	479,40	165,24	152,01	153,12	161,60
RMSE^{d}	479,65	178,63	164,27	160,68	175,61
$\sum$ quad. Dist.^{e}	–	0,40	0,46	0,72	0,50

Fortsetzung siehe nächste Seite

Fortsetzung Tabelle 4.6

	Ridge Matching	Zuordnung m. Zurücklegen	opt.1:1 Matching	opt. Full Matching	Random Matching
unähnliche Merkmale					
$Y_{NT_{wahr}}{}^{a}$	2684,05	2684,05	2684,05	2684,05	2684,05
$\hat{Y}_{NT}{}^{b}$	2781,63	2750,48	2764,76	2744,04	2764,80
Bias	-97,58	-66,43	-80,70	-59,98	-80,74
emp. Std.-abw.^{c}	581,38	224,33	227,13	231,00	226,60
RMSE^{d}	589,51	233,96	241,04	238,66	240,56
$\sum$ quad. Dist.^{e}	–	0,42	0,49	0,76	0,53
Stichproben mit 300 Teilnehmern und 900 Nichtteilnehmern					
ähnliche Merkmale					
$Y_{NT_{wahr}}{}^{a}$	2697,25	2697,25	2697,25	2697,25	2697,25
$\hat{Y}_{NT}{}^{b}$	2603,49	2707,44	2711,49	2697,13	2708,55
Bias	93,76	-10,19	-14,24	0,12	-11,30
emp. Std.-abw.^{c}	228,21	95,95	87,76	84,83	86,89
RMSE^{d}	246,72	96,49	88,90	84,83	87,62
$\sum$ quad. Dist.^{e}	–	0,49	0,61	0,98	0,65
eingeschränkt ähnliche Merkmale					
$Y_{NT_{wahr}}{}^{a}$	2696,75	2696,75	2696,75	2696,75	2696,75
$\hat{Y}_{NT}{}^{b}$	2596,81	2781,70	2784,36	2767,31	2786,07
Bias	99,94	-84,95	-87,61	-70,56	-89,32
emp. Std.-abw.^{c}	308,26	115,94	112,54	114,14	115,44
RMSE^{d}	324,05	143,73	142,62	134,19	145,96
$\sum$ quad. Dist.^{e}	–	0,48	0,60	0,99	0,64
unähnliche Merkmale					
$Y_{NT_{wahr}}{}^{a}$	2695,84	2695,84	2695,84	2695,84	2695,84
$\hat{Y}_{NT}{}^{b}$	2665,77	2796,74	2803,45	2781,35	2804,19
Bias	30,08	-100,89	-107,61	-85,51	-108,34
emp. Std.-abw.^{c}	481,27	172,78	175,73	176,47	174,94
RMSE^{d}	482,21	200,08	206,06	196,09	205,77
$\sum$ quad. Dist.^{e}	–	0,50	0,64	1,05	0,69

Anmerkungen:
Durchschnittsergebnisse aus 100 Stichproben.
Abweichung der Mittelwerte bzw. Häufigkeitsverteilungen: ähnliche Merkmale 1% der merkmalsspezifischen Streuung, eingeschränkt ähnliche 10%, unähnliche 25%.
[a] „wahres“ Nichtteilnahmeeinkommen;
[b] geschätztes Nichtteilnahmeeinkommen als Durchschnitt der Einkommen in der Kontrollgruppe;
[c] empirische Standardabweichung;
[d] Wurzel des mittleren quadratischen Fehlers;
[e] Summe der quadrierten Distanzen.

Weniger deutlich sind die Ergebnisse hinsichtlich der Bestandteile des MSE. In Bezug auf den Bias lässt sich eine Überlegenheit des Optimal Full Matching beobachten. Die Rangfolge der drei anderen Zuordnungsprozesse wechselt zwischen den verschiedenen Designs. In kleinen Stichproben verbleiben nach der Zuordnung mit Zurücklegen die größten Abweichungen zwischen „wahrem“ und geschätztem Nichtteilnahmeeinkommen. Mit zunehmender Größe der Stichproben verringert sich der Bias allerdings. Insgesamt liegen die ermittelten Werte aller drei Prozesse relativ dicht beieinander. Anders als in Dehejia und Wahba (2002) lässt sich für Stichproben mit unähnlichen Merkmalen kein Vorteil des Zuordnungsverfahrens mit Zurücklegen gegenüber den Verfahren ohne Zurücklegen feststellen.

Die Zuordnung mit Zurücklegen weist in den meisten Stichproben eine größere Streuung der Schätzwerte auf als die beiden optimalen Algorithmen und Random Matching. Für diese drei Prozesse lässt sich keine Rangfolge festlegen, da die ermittelten Werte auch für dieses Kriterium relativ dicht beieinander liegen und die Rangfolge zwischen den einzelnen Stichprobendesigns wechselt.

Tendenziell scheint das Optimal Full Matching hinsichtlich der betrachteten Effizienzkriterien am besten geeignet zu sein für die Zuordnung passender Partner zu den Teilnehmern. Allerdings werden nicht in allen Stichproben Unterschiede zum Optimal Nearest Neighbor Matching, dem Random Matching und der Zuordnung mit Zurücklegen deutlich.

Den letzten Schritt der Analyse der Zuordnungsprozesse bildet die Anwendung der Algorithmen auf Stichproben mit deutlich mehr Nichtteilnehmern als Teilnehmern. Die Ergebnisse dieses Analyseschritts werden in der Tabelle 4.7 zusammengefasst.

Die Vergrößerung der Anzahl potenzieller Partner für die Teilnehmer verändert die getroffenen Aussagen über die Güte der Zuordnungsprozesse nicht wesentlich. Auch in den Stichproben mit einem Größenverhältnis von 1:10 ist Ridge Matching deutlich schlechter als die anderen Zuordnungsprozesse. Die Beurteilung der Algorithmen anhand der Summe der quadrierten Distanzen ändert sich ebenfalls nicht – die Zuordnung mit Zurücklegen ist in allen Stichprobendesigns am besten, am schlechtesten sind die Ergebnisse des Optimal Full Matching, dazwischen liegen Random Matching und Optimal Nearest Neighbor Matching mit sehr ähnlichen Werten.

Tabelle 4.7: **Ergebnisse für Stichproben mit einem Größenverhältnis von 1:10**

	Ridge Matching	Zuordnung m. Zurücklegen	opt.1:1 Matching	opt. Full Matching	Random Matching
Stichproben mit 50 Teilnehmern und 500 Nichtteilnehmern					
ähnliche Merkmale					
$Y_{NT_{wahr}}{}^a$	2684,64	2684,64	2684,64	2684,64	2684,64
$\hat{Y}_{NT}{}^b$	2473,43	2725,25	2719,39	2693,55	2725,43
Bias	211,21	-40,61	-34,75	-8,91	-40,80
emp. Std.-abw.^{c}	457,13	189,83	185,88	156,87	187,33
RMSE^d	503,56	194,12	189,10	157,12	191,72
$\sum$ quad. Dist.^{e}	–	0,21	0,22	0,54	0,23
eingeschränkt ähnliche Merkmale					
$Y_{NT_{wahr}}{}^a$	2678,22	2678,22	2678,22	2678,22	2678,22
$\hat{Y}_{NT}{}^b$	2626,95	2768,10	2767,92	2732,58	2773,57
Bias	51,27	-89,88	-89,70	-54,35	-95,35
emp. Std.-abw.^{c}	512,03	205,33	199,03	191,98	198,56
RMSE^d	514,59	224,14	218,31	199,52	220,27
$\sum$ quad. Dist.^{e}	–	0,21	0,22	0,55	0,22
unähnliche Merkmale					
$Y_{NT_{wahr}}{}^a$	2675,69	2675,69	2675,69	2675,69	2675,69
$\hat{Y}_{NT}{}^b$	2576,75	2770,85	2775,91	2756,28	2781,66
Bias	98,93	-95,16	-100,23	-80,60	-105,97
emp. Std.-abw.^{c}	672,68	255,55	256,02	239,43	254,47
RMSE^d	679,92	272,69	274,94	252,64	275,65
$\sum$ quad. Dist.^{e}	–	0,21	0,23	0,57	0,23
Stichproben mit 100 Teilnehmern und 1000 Nichtteilnehmern					
ähnliche Merkmale					
$Y_{NT_{wahr}}{}^a$	2692,38	2692,38	2692,38	2692,38	2692,38
$\hat{Y}_{NT}{}^b$	2567,73	2674,11	2676,57	2650,13	2674,25
Bias	124,65	18,27	15,80	42,24	18,13
emp. Std.-abw.^{c}	267,15	127,63	122,74	98,62	126,06
RMSE^d	294,80	128,93	123,75	107,29	127,36
$\sum$ quad. Dist.^{e}	–	0,19	0,20	0,62	0,21
eingeschränkt ähnliche Merkmale					
$Y_{NT_{wahr}}{}^a$	2689,77	2689,77	2689,77	2689,77	2689,77
$\hat{Y}_{NT}{}^b$	2494,36	2760,96	2767,85	2751,14	2765,81
Bias	195,41	-71,19	-78,08	-61,38	-76,04
emp. Std.-abw.^{c}	399,55	150,61	151,74	148,71	152,35
RMSE^d	444,78	166,59	170,65	160,88	170,27
$\sum$ quad. Dist.^{e}	–	0,18	0,19	0,64	0,20

Fortsetzung siehe nächste Seite

Fortsetzung Tabelle 4.7

	Ridge Matching	Zuordnung m. Zurücklegen	opt.1:1 Matching	opt. Full Matching	Random Matching
unähnliche Merkmale					
$Y_{NT_{wahr}}{}^{a}$	2687,30	2687,30	2687,30	2687,3	2687,30
$\hat{Y}_{NT}{}^{b}$	2592,51	2775,41	2777,74	2756,93	2776,85
Bias	94,78	-88,11	-90,44	-69,63	-89,55
emp. Std.-abw.^{c}	506,73	188,14	187,93	205,05	184,86
RMSE^{d}	515,52	207,75	208,55	216,55	205,41
$\sum$ quad. Dist.^{e}	–	0,19	0,21	0,68	0,21
Stichproben mit 300 Teilnehmern und 3000 Nichtteilnehmern					
ähnliche Merkmale					
$Y_{NT_{wahr}}{}^{a}$	2697,43	2697,43	2697,43	2697,43	2697,43
$\hat{Y}_{NT}{}^{b}$	2503,47	2699,43	2699,72	2688,24	2700,19
Bias	193,96	-2,00	-2,29	9,19	-2,76
emp. Std.-abw.^{c}	156,34	77,53	75,55	66,34	77,81
RMSE^{d}	249,12	77,56	75,58	66,97	77,85
$\sum$ quad. Dist.^{e}	–	0,17	0,19	0,82	0,19
eingeschränkt ähnliche Merkmale					
$Y_{NT_{wahr}}{}^{a}$	2696,59	2696,59	2696,59	2696,59	2696,59
$\hat{Y}_{NT}{}^{b}$	2560,14	2752,46	2753,41	2727,41	2753,28
Bias	136,45	-55,87	-56,82	-30,82	-56,69
emp. Std.-abw.^{c}	264,92	106,09	101,50	94,88	104,77
RMSE^{d}	297,99	119,90	116,32	99,76	119,12
$\sum$ quad. Dist.^{e}	–	0,16	0,18	0,82	0,19
unähnliche Merkmale					
$Y_{NT_{wahr}}{}^{a}$	2697,28	2697,28	2697,28	2697,28	2697,28
$\hat{Y}_{NT}{}^{b}$	2485,07	2794,39	2798,29	2791,94	2797,50
Bias	212,21	-97,11	-101,01	-94,66	-100,22
emp. Std.-abw.^{c}	438,91	168,40	169,31	174,19	169,11
RMSE^{d}	487,52	194,39	197,16	198,24	196,58
$\sum$ quad. Dist.^{e}	–	0,18	0,19	0,89	0,20

Anmerkungen:
Durchschnittsergebnisse aus 100 Stichproben.
Abweichung der Mittelwerte bzw. Häufigkeitsverteilungen: ähnliche Merkmale 1% der merkmalsspezifischen Streuung, eingeschränkt ähnliche 10%, unähnliche 25%.
[a] „wahres“ Nichtteilnahmeeinkommen;
[b] geschätztes Nichtteilnahmeeinkommen als Durchschnitt der Einkommen in der Kontrollgruppe;
[c] empirische Standardabweichung;
[d] Wurzel des mittleren quadratischen Fehlers;
[e] Summe der quadrierten Distanzen.

Eine Veränderung ergibt sich insofern, als die Vorzüge des Optimal Full Matching, die sich hinsichtlich des MSE im vorangegangenen Analyseschritt angedeutet haben, nun deutlicher sichtbar sind. Das gleiche gilt für den Bias und die Streuung der Schätzwerte. In Stichproben mit großer relativer Nichtteilnehmerzahl ist für Optimal Full Matching in nahezu allen Stichprobendesigns die geringste Abweichung zwischen „wahrem“ und geschätztem Nichtteilnahmeeinkommen, die geringste Streuung der Schätzwerte – und demzufolge auch der kleinste MSE unter den analysierten Zuordnungsalgorithmen festzustellen. Die entsprechenden Werte für Zuordnung mit Zurücklegen, Optimal Nearest Neighbor Matching und Random Matching sind für alle drei Kriterien sehr ähnlich, sodass sich keine Rangfolge dieser drei Prozesse festlegen lässt.

Ein Vergleich der Ergebnisse für die verschiedenen Stichprobendesigns macht deutlich, dass die Anzahl der potenziellen Partner je Teilnehmer, die Ähnlichkeit zwischen den Personen der Ausgangsstichproben sowie die Stichprobengröße einen Einfluss auf die Güte der erzielten Matchingergebnisse haben. Die Wirkung auf die einzelnen betrachteten Gütemaße ist dabei unterschiedlich.

Mit zunehmender Zahl potenzieller Partner verringert sich die Summe der quadrierten Distanzen. Das erklärt sich aus der Verbesserung der einzelnen Zuordnungen, die mit der Erhöhung der Wahlmöglichkeiten einhergehen. Innerhalb eines Größenverhältnisses zwischen Teilnehmer- und Nichtteilnehmeranzahl verändert sich die Summe der quadrierten Distanzen mit Zunahme der Personenzahl nicht. Auch das ist kein überraschendes Ergebnis, da sich die Bedingungen für die Zuordnung passender Partner für jeden einzelnen Teilnehmer nicht verändern.

Mit der Zunahme der Wahlmöglichkeiten – also je größer die Nichtteilnehmerstichproben in Relation zu den Teilnehmerstichproben sind – würde man ebenfalls eine Verringerung der Abweichung der Schätzergebnisse zu den „wahren“ Nichtteilnahmeeinkommen erwarten. Ein Einfluss der Anzahl potenzieller Partner für einen Teilnehmer auf den Bias ist in den Ergebnissen allerdings nicht zu beobachten. Für die Streuung der Ergebnisse lässt sich dagegen ein positiver Zusammenhang zur Stichprobenrelation feststellen.

Innerhalb der generierten Stichproben einer Größe ist die Streuung der Schätzergebnisse umso geringer, je ähnlicher die Merkmale der Personen in den Ausgangsstichproben sind. Für die Abweichung des geschätzten vom „wahren“ Nichtteilnahmeeinkommen lässt sich die gleiche Beobachtung machen. Die Vergrößerung der Stichproben – bei konstantem Größenverhältnis – bewirkt ebenfalls eine Verringerung der Streuung und des Bias. Diese Beobachtung stimmt mit den Aussagen in Fröhlich (2004a) über die Verbesserung von Matchingergebnissen mit zunehmender Stichprobengröße überein.

Ähnliche Aussagen wie für die Streuung der Schätzergebnisse lassen sich für den mittleren quadratischen Fehler treffen.[45] Er ist umso geringer, je ähnlicher die Merkmale der Personen in den Teilstichproben innerhalb einer Stichprobengröße sind, je größer die Nichtteilnehmerstichproben in Relation zu den Teilnehmerstichproben sind, und je größer die Stichprobe insgesamt ist.

Die Simulationsergebnisse zeigen – übereinstimmend mit den Ergebnissen früherer Studien –, dass kein Zuordnungsprozess in allen Kriterien gleichermaßen gut oder schlecht und den anderen in allen Gütekriterien überlegen ist. Die Aussagen über die Eignung der Zuordnungsprozesse hängen vom betrachteten Gütemaß ab.

Eine Ausnahme bildet das Ridge Matching, das hinsichtlich aller Kriterien schlechter bewertet wird als Random Matching, optimale Zuordnungsprozesse und die Zuordnung mit Zurücklegen. Ein möglicher Grund für diese schlechtere Performance liegt in der Verwendung des Propensity Scores anstelle der Mahalanobis-Matching-Distanz. Wie aus dem ersten Teil der Analyse hervorgeht, gelingt mit diesem Score die Zusammenfassung der einbezogenen Variablen nur unzureichend – und bietet damit eine schlechte Grundlage für die Zuordnung von Personen.[46]

Für Optimal Nearest Neighbor Matching auf Grundlage des Ungarischen Algorithmus' und Random Matching werden sehr ähnliche Ergebnisse – hinsichtlich aller

[45] Dieser enge Zusammenhang lässt sich mit dem hohen prozentualen Anteil der Varianz am MSE erklären, der aus dem Vergleich der Größe von empirischer Standardabweichung und RMSE – bzw. dem in den Tabellen B.7 bis B.15 im Anhang ausgewiesenen prozentualen Anteil der Varianz am MSE – ersichtlich ist.

[46] Insofern steht dieses Ergebnis nicht im Widerspruch zu den Aussagen in Fröhlich (2004a), wo der Propensity Score auf Grundlage eines Merkmals ermittelt und direkt zur Definition der Outcomegrößen verwendet wird.

Gütemaße – beobachtet. Es ist also kein Vorteil des optimalen Nearest Neighbor Matching gegenüber dem Random Matching zu erkennen. Dieses Ergebnis stimmt mit der Aussage in Gu und Rosenbaum (1993) überein, die in der zitierten Studie damit begründet wird, dass beide Algorithmen tendenziell die gleichen Nichtteilnehmer auswählen.

Hinsichtlich des Bias, der Streuung – und damit des mittleren quadratischen Fehlers – scheint Optimal Full Matching am besten für die Zuordnung passender Partner in Stichproben mit unterschiedlich großen Teilnehmer- und Nichtteilnehmerzahlen geeignet zu sein. Dies ist umso deutlicher, je größer die Nichtteilnehmerstichprobe im Vergleich zur Teilnehmerstichprobe ist. Diese Aussage bildet eine Ergänzung zu den Ergebnissen von Gu und Rosenbaum (1993), die einen Vorteil von Full Matching im Vergleich zu anderen analysierten Zuordnungsprozessen hinsichtlich der Angleichung der Verteilungen der Merkmale in Teilnehmer- und Kontrollgruppe feststellen.

Betrachtet man die Summe der quadrierten Distanzen als „Repräsentanten“ der einzelnen Merkmalsabweichungen, steht die Aussage der zitierten Studie im Widerspruch zu den Ergebnissen der Simulation, in der für Optimal Full Matching in allen Stichprobendesigns die größten Distanzsummen festgestellt werden. In diesem Kriterium werden mit der Zuordnung mit Zurücklegen die besten Ergebnisse erzielt. Das ist allerdings nicht überraschend, da die Distanz zwischen einem Teilnehmer und seinem besten Partner erwartungsgemäß geringer als die zwischen einem Teilnehmer und dem Durchschnittswert aus mehreren – und damit evtl. auch nicht so ähnlichen – Personen ist.

Im Vorfeld der Analyse wird die Hypothese aufgestellt, dass in ähnlichen Ausgangsstichproben alle Algorithmen etwa gleich gute Matchingergebnisse liefern, da die Auswahl an möglichen Partnern relativ groß ist. Diese Hypothese wird von den Simulationsergebnissen nicht bestätigt. Das gleiche gilt für die zweite Hypothese, nach der bei der Zuordnung mit Zurücklegen der Bias geringer und die Streuung des Ergebnisses größer als bei anderen Zuordnungsverfahren ist.
Die Hypothese über Optimal Full Matching kann dagegen teilweise bestätigt werden. In den Analyseergebnissen ist eine geringere Streuung des Maßnahmeeffekts als für die anderen Zuordnungsverfahren zu beobachten. Allerdings ist auch der Bias geringer, was der Hypothese widerspricht.

Der erwartete Vorteil des optimalen Nearest Neighbor Matching gegenüber dem Random Matching hinsichtlich der Vermeidung des Verlusts von Teilnehmern aufgrund einer suboptimalen Zuordnung kann in den Simulationsergebnissen nicht beobachtet werden, da ein solcher Verlust auch beim Random Matching in keinem der Stichprobendesigns auftritt. Uneingeschränkte Bestätigung liefern die Analyseergebnisse für die letzte Hypothese, nach der ein Verlust von Teilnehmern aufgrund fehlender Partner vermieden werden kann, wenn in Stichproben mit nahezu gleich großer Teilnehmer- und Nichtteilnehmeranzahl Ridge Matching und Zuordnungen mit Zurücklegen angewendet werden.

4.4 Zusammenfassung

In der Simulation werden verschiedene Distanzmaße und Zuordnungsprozesse verglichen, um festzustellen, welches Matchingverfahren in der praktischen Anwendung die besten Ergebnisse (hinsichtlich geeigneter Gütekriterien) verspricht. Das Hauptaugenmerk liegt dabei auf kleinen Stichproben, d.h. Stichproben mit 100 Teilnehmern. Für die Simulation wird ein in der Arbeitsmarktforschung häufig verwendeter Datensatz, der Mikrozensus Deutschland, nachgebildet. Der Vorteil einer solchen „Nachbildung“ besteht in der dadurch erreichten Nähe zu real vorzufindenden Entscheidungssituationen durch die Einbeziehung unterschiedlich skalierter Variablen.

Die Auswahl der in die Analyse einbezogenen Distanzmaße und Zuordnungsprozesse ergibt sich zum einen aus theoretischen Überlegungen, zum anderen aus den Ergebnissen früherer Studien. In der Simulation werden der Propensity Score, der Index Score, die Mahalanobisdistanz sowie die Mahalnobis-Matching-Distanz und das Distanzmaß nach Gower miteinander verglichen. In der Analyse der Zuordnungsprozesse werden Random Matching, optimales Nearest Neighbor Matching, Zuordnung mit Zurücklegen, Optimal Full Matching und Rigde Matching berücksichtigt.

Die Simulation besteht aus zwei Teilanalysen, die die beiden grundlegenden Entscheidungen bei der Auswahl eines geeigneten Matchingverfahrens wiedergeben. Zunächst wird untersucht, welches Distanzmaß am besten in der Lage ist, unterschiedlich skalierte Daten zusammenzufassen. Im zweiten Schritt werden Zuordnungspro-

zesse hinsichtlich ihrer Fähigkeit, die „besten“ Partner auszuwählen, analysiert. In beiden Teilanalysen werden unterschiedliche Gütemaße zur Beurteilung der Ergebnisse eingesetzt. Für die Prüfung der Distanzmaße werden die Bias Reduzierung durch Matching und – als Alternative zu den bisher in der Literatur verwendeten Gütemaßen – nichtparametrische skalenspezifische Mittelwert- bzw. Häufigkeitsverteilungstests angewendet. Zur Beurteilung der Zuordnungsprozesse werden der mittlere quadratische Fehler (MSE), der Bias und die Varianz sowie die Summe der quadrierten Distanzen zwischen Teilnehmern und Nichtteilnehmern sowie die Anzahl der entfernten (Teilnehmer-)Beobachtungen betrachtet.

In jedem Analyseschritt werden jeweils 100 Simulationsläufe mit verschiedenen Kombinationen aus Teilnehmer- und Nichteilnehmerstichproben durchgeführt, die sich in ihrer Größe insgesamt, dem Zahlenverhältnis von Teilnehmern und Nichtteilnehmern sowie der Größe des Common-Support-Bereichs unterscheiden.

In der Analyse der Distanzmaße wird festgestellt, dass der Propensity Score und der Index Score in allen untersuchten Designs deutlich schlechter als die Mahalanobisdistanz und die beiden aggregierten Distanzmaße in der Lage sind, Ähnlichkeiten bzw. Unterschiede in unterschiedlich skalierten Merkmalen adäquat abzubilden. Insgesamt scheint die gewichtete Mahalanobis-Matching-Distanz von den analysierten Distanzmaßen dazu am besten in der Lage zu sein. Allerdings trifft das nicht für alle Skalenniveaus gleichermaßen zu. Die Angleichung nominaler (dichotomer und vor allem polytomer) Variablen gelingt sehr gut, dafür bleiben nach dem Matching häufig Unterschiede in der Verteilung der metrischen Variablen bestehen. Nahezu das Gegenteil gilt für das Distanzmaß nach Gower. Dieses Maß ist nicht bzw. nur bedingt geeignet zur Angleichung polytomer Variablen. Auch die Ergebnisse hinsichtlich der dichotomen Variablen sind schlechter im Vergleich zu Mahalanobisdistanz und Mahalanobis-Matching-Distanz. Die „Stärke“ dieses Distanzmaßes liegt bei metrischen Variablen. Es wäre zu prüfen, ob eine Kombination beider Distanzmaße – des verallgemeinerten Matchingkoeffizienten für nominale Variablen und der normierten absoluten Merkmalsdifferenz für metrische Variablen – besser in der Lage ist, die Ähnlichkeiten und Unterschiede der verschieden skalierten Merkmale adäquat zusammenzufassen.

Die Simulationsergebnisse des zweiten Teils der Analyse zeigen, dass kein Zuordnungsprozess in allen Kriterien gleichermaßen gut oder schlecht ist. Die Aussagen über die Eignung der Algorithmen zur Zuordnung passender Partner zu den Teilnehmern hängen vom betrachteten Gütemaß ab. Eine Ausnahme bildet das Ridge Matching, das hinsichtlich aller Kriterien schlechter bewertet wird als Random Matching, optimale Zuordnungsprozesse und die Zuordnung mit Zurücklegen.

Für Optimal Nearest Neighbor Matching auf Grundlage des Ungarischen Algorithmus' und Random Matching ergeben sich in der Simulation in allen Gütemaßen sehr ähnliche Ergebnisse. Es ist kein Vorteil des optimalen Nearest Neighbor Matching gegenüber dem Random Matching zu erkennen.

Hinsichtlich des Bias, der Streuung – und damit des mittleren quadratischen Fehlers – scheint in Stichproben mit unterschiedlich großen Teilnehmer- und Nichtteilnehmerzahlen Optimal Full Matching am besten für die Zuordnung passender Partner geeignet zu sein. Dies ist umso deutlicher, je größer die Nichtteilnehmerstichprobe im Vergleich zur Teilnehmerstichprobe ist.

Betrachtet man dagegen die Summe der quadrierten Distanzen, liefert die Zuordnung mit Zurücklegen die besten Matchingergebnissse.

Sind die Teilnehmer- und Nichtteilnehmerstichproben annähernd gleich groß, legen die Simulationsergebnisse die Nutzung von Zuordnungen mit Zurücklegen nahe – und bestätigen die in der Literatur geäußerten Vorbehalte gegen die Anwendung von Zuordnungen ohne Zurücklegen.

Nach der theoretischen Analyse verschiedener Matchingverfahren wird im folgenden Kapitel ein Anwendungsbeispiel für die Evaluation der Berufsausbildungsförderung gegeben. Zur Beantwortung der Frage, ob die Förderung der beruflichen Erstausbildung in den Neuen Bundesländern einen negativen Einfluss auf die Berufseinstiegschancen hat, ist die Berücksichtigung anderer Einflüsse, insbesondere persönlicher Merkmale der Jugendlichen oder des Ausbildungsberufes, notwendig. Zur Kontrolle solcher Einflüsse wird anhand der vorgestellten Gütekriterien ein geeigneter Matchingalgorithmus ausgewählt.

Kapitel 5

Anwendungsbeispiel: Evaluation der Förderung der Berufsausbildung in den Neuen Bundesländern

In den vergangenen Jahren hat sich das Berufsausbildungssystem in Deutschland deutlich verändert. Während noch vor einigen Jahren die Ausbildung ausschließlich im „klassischen“ dualen System oder als schulische Ausbildung stattfand, gibt es inzwischen verschiedene ergänzende oder alternative Formen der Berufsausbildung. Zu nennen wären hier u.a. die Auslagerung einzelner Ausbildungsbestandteile aus den Betrieben oder der Ersatz der einzelbetrieblichen Ausbildung durch Ausbildungsverbünde. Hinzu kommt eine Vielzahl von Ausbildungs- und Ausbildungsvorbereitungsmaßnahmen, die vom Bund und den Ländern finanziert werden.

Während sich die Veränderung des Berufsausbildungssystems in den alten Bundesländern eher langsam vollzieht, bestehen in den neuen Ländern schon seit Beginn der 1990er Jahre vielfältige Erfahrungen bei der Ausgestaltung von Alternativen zur dualen Berufsausbildung. Ostdeutschland besitzt damit in gewisser Weise „Modellcharakter“ für die zukünftige Gestaltung der Berufsausbildung in Deutschland. Infolge des massiven Ausbildungsplatzmangels nach der Wende, der zum einen durch die finanzielle Unterstützung der betrieblichen Ausbildung, zum großen Teil aber auch durch die Bereitstellung geförderter Ausbildungsplätze ausgeglichen wird, haben sich in der Region verschiedene Bildungsträger etabliert.[1] Zwischen diesen Bildungsträgern und den ostdeutschen Betrieben sind verschiedene Formen der Kooperation – teilweise unter Einbeziehung weiterer Akteure – entstanden, die von beiden Seiten als nützlich und zukunftsfähig angesehen werden (Grünert und Wiekert, 2005). Den Hauptinhalt dieser Ausbildungsnetzwerke bildet die Erweiterung der Möglichkeiten der beruflichen Erstausbildung von Jugendlichen, etwa durch die Auslagerung einzelner Ausbildungsbestandteile sowie überbetriebliche oder berufsübergreifende Angebote der Bildungsträger. Entsprechend groß ist die Bedeutung (und das Interesse an) einer Evaluierung dieser Ausbildungsplatzprogramme.[2]

[1] Mehr als ein Viertel der in Ostdeutschland jährlich abgeschlossenen Ausbildungsverträge wird über die Bund-Länder-Ausbildungsplatzprogramme Ost, ergänzende Länderprogramme oder die Agentur für Arbeit zur Verfügung gestellt (Berger et al., 2007, S. 49).

[2] Am Bundesinstitut für Berufsbildung (BiBB) werden seit Ende der 1990er Jahre verschiedene Forschungsprojekte zur Evaluierung der Ausbildungsförderung in Ostdeutschland durchgeführt, in denen verschiedene Aspekte der Förderung (Ausgestaltung der Maßnahmen, Zufriedenheit der Träger und der geförderten Jugendlichen sowie die Wirksamkeit der unterschiedlichen Ausbildungsformen) untersucht werden. Ergebnisse dieser Forschungsprojekte finden sich u.a. in Berger und Walden (2003); Berger et al. (2007).

Den Ergebnissen früherer Studien zufolge sind Absolventen geförderter Ausbildungsgänge beim Berufseinstieg benachteiligt gegenüber Absolventen „normaler“ Berufsausbildungen (Berger et al., 2007; Prein, 2005; Steiner et al., 2004). Dafür werden verschiedene Gründe genannt. Zum einen spielen Selektionseffekte eine bedeutende Rolle – unterschieden wird dabei zwischen systematischen Unterschieden in den beschäftigungsrelevanten Merkmalen der Absolventen (persönlicher Selektionseffekt) und der Förderung marktferner Ausbildungsberufe (beruflicher Selektionseffekt). Zum anderen resultiert aus dem negativen Image, das den Absolventen geförderter Ausbildungen in den Augen potenzieller Arbeitgeber anhaftet, ein Stigmatisierungseffekt (Prein, 2005, S. 192f.).

Die Aussagen in den genannten Studien stützen sich überwiegend auf deskriptive Analysen und die Ergebnisse parametrischer Schätzungen, in denen der Selektionseffekt nicht explizit berücksichtigt wird. Damit ist die Frage, wie groß die Beschäftigungschancen der Jugendlichen wären, wenn sie einen Ausbildungsplatz ohne staatliche Förderung erhalten hätten, noch nicht abschließend beantwortet. Ein weiterer Grund für die erneute Evaluation der geförderten Berufsausbildung in den Neuen Bundesländern ist die Verbesserung der Datenlage auf diesem Gebiet. Während sich die Studien von Prein (2005) und Steiner et al. (2004) auf die erste Welle des Jugendpanels stützen, stehen inzwischen zwei weitere Wellen dieser Befragung zur Verfügung – und damit ein längerer Beobachtungszeitraum und höhere Fallzahlen.[3]

Das Anliegen der empirischen Analyse in diesem Kapitel ist eine Ergänzung zu früheren Untersuchungen der Effekte einer Förderung der beruflichen Erstausbildung auf den Einstieg in das Berufsleben, speziell für die Ausbildungsförderung in Ostdeutschland.

5.1 Gegenstand der Analyse

Im Mittelpunkt der Untersuchung steht die Frage, ob die geförderte Berufsausbildung in den Neuen Bundesländern ein negatives Image besitzt, das die geförder-

[3]Das Panel entstand als Teil eines Projekts, das sich mit Untersuchungen zur Mobilität Jugendlicher auf dem ostdeutschen Arbeitsmarkt beschäftigt. Nähere Informationen sind auf der website des Zentrums für Sozialforschung zu finden. Vgl. Zentrum für Sozialforschung Halle (zsh) (2003).

ten Jugendlichen beim Berufseinstieg gegenüber Absolventen ungeförderter Ausbildungsgänge benachteiligt.

Dieses negative Image wird in der Literatur mit der engen Zielgruppe der Berufsausbildungsförderung der vergangenen Jahre in den alten Bundesländern erklärt. Gefördert wurden hier v.a. beachteiligte Jugendliche – d.h. Jugendliche, die aufgrund körperlicher oder geistiger Behinderungen oder schulischer Defizite (v.a. Jugendliche mit Migrationshindergrund und aus bildungsfernen Elternhäusern) – auf dem regulären Ausbildungsmarkt keinen Ausbildungsplatz gefunden haben. Nur zu einem geringen Teil – und auch erst seit wenigen Jahren – wird auch die Berufsausbildung marktbenachteiligter Jugendlicher – Jugendlicher, die nach Ablauf einer bestimmten Frist keinen Ausbildungsplatz bekommen haben – durch verschiedene vom Bund und den Ländern finanzierte Fördermaßnahmen unterstützt.[4]

Obwohl die Zielgruppe der Ausbildungsförderung in den Neuen Bundesländern sehr viel weiter gefasst ist – es werden i.d.R. Jugendliche, die zum Beginn eines Ausbildungsjahres keinen Ausbildungsplatz haben, gefördert – wäre es möglich, dass das negative Image der Ausbildungsförderung in den Alten Bundesländern auf die Absolventen geförderter Berufsausbildungsgänge in den Neuen Bundesländern übertragen wird und Jugendliche mit sonst gleichen Voraussetzungen schlechtere Berufseinstiegschancen haben, wenn sie eine geförderte Ausbildung absolviert haben (Prein, 2005, S. 193).

Vor Beginn der Untersuchung müssen einige Begriffe, die im Zusammenhang mit der Berufsausbildung gebräuchlich sind, näher definiert werden. In Deutschland wird zwischen der vertraglich gebundenen Berufsausbildung und der vollqualifizierenden schulischen Berufsausbildung unterschieden. Zur ersten Kategorie gehören die betriebliche und die außerbetriebliche Ausbildung für Berufe lt. Berufsbildungsgesetz (BBiG) oder Handwerksordnung (HwO). Die vollqualifizierende schulische Berufsausbildung umfasst die Ausbildung in Berufsfachschulen und im Gesundheitswesen.[5]

[4]Den Beginn dieser Entwicklung markiert das Programm zur Bekämpfung der Jugendarbeitslosigkeit („Jump“) der Bundesregierung aus dem Jahr 1999. Der Nachfolger dieses Programms wurde im Zuge der Reformen der Arbeitsmarktpolitik (durch die sog. Hartz-Gesetze) in den Maßnahmenkatalog der Agentur für Arbeit integriert.

[5]Innerhalb der Berufsfachschulen (BFS) wird noch einmal zwischen Berufsfachschulen für BBiG-/HwO-Berufe und BFS für Berufe außerhalb BBiG-/HwO unterschieden (Bundesministeri-

Für beide Ausbildungsformen – vertraglich gebundene und schulische Berufsausbildung – gibt es verschiedene Arten der Förderung, die auf unterschiedliche Initiativen des Bundes und der Länder zurückgehen – und auch sehr unterschiedliche Adressaten haben. Die Ausbildungsförderung in Ostdeutschland bezieht sich v.a. auf marktbenachteiligte Jugendliche.[6]

Die Förderung erfolgt i.d.R. in außerbetrieblichen und betriebsnahen Ausbildungsgängen. Die außerbetriebliche Ausbildung ist gekennzeichnet durch die Vermittlung theoretischer Kenntnisse in Berufsfachschulen und eine fachpraktische Ausbildung in überbetrieblichen Ausbildungsstätten – nur zu einem geringen Teil in Praktikumsbetrieben. Charakteristisch für eine betriebsnahe Ausbildung ist dagegen die Absolvierung des praktischen Teils der Ausbildung (mindestens 50%) in Praktikumsbetrieben – neben der Vermittlung theoretischer Kenntnisse in Berufsfachschulen (Berger, 2006, S. 7).

In der Analyse wird zwischen nicht geförderter und geförderter Ausbildung unterschieden, wobei unter der ersten Kategorie die „klassische“ Ausbildung im dualen Ausbildungssystem und die nicht geförderte schulische Berufsausbildung verstanden werden. Innerhalb der geförderten Ausbildung wird zwischen den beiden genannten Förderarten betriebsnahe und außerbetriebliche Ausbildung unterschieden.

Für beide Formen der geförderten Berufsausbildung wird die Hypothese aufgestellt, dass die Absolventen aufgrund des Stigmatisierungseffekts schlechtere Chancen auf einen qualifikationsadäquaten Berufseinstieg haben als Absolventen ungeförderter Berufsausbildungen.
Darüber hinaus wird erwartet, dass der Stigmatisierungseffekt in außerbetrieblichen Berufsausbildungen stärker als in betriebsnahen Ausbildungsgängen ist, da die Jugendlichen hier nicht – oder nur in geringerem Maße – die Möglichkeit haben, während der Praktikumsaufenthalte die Vorurteile der potenziellen Arbeitgeber zu widerlegen.

---

um für Bildung und Forschung, 2006). Daneben gibt es verschiedene Formen der Berufsvorbereitung, die für die Analyse allerdings keine Rolle spielen.

[6] In den Bund-Länder-Vereinbarungen der Ausbildungsplatzprogramme Ost wird die Schaffung zusätzlicher Ausbildungsplätze für Jugendliche, die unmittelbar vor Maßnahmebeginn bei der Bundesagentur für Arbeit als noch nicht vermittelte Ausbildungsplatzbewerber geführt sind, insbesondere in wirtschaftlich konkurrenzfähigen Branchen und Unternehmen, als primäres Ziel genannt (Berger et al., 2007, S. 54).

5.2 Datengrundlage

Als Datenbasis der empirischen Analyse dient eine Befragung, die vom Zentrum für Sozialforschung (zsh) an der Martin-Luther-Universität Halle-Wittenberg im Zeitraum zwischen 2002 und 2006 durchgeführt wurde, das Jugendpanel.

5.2.1 Jugendpanel des Zentrums für Sozialforschung Halle

Für dieses Panel werden Jugendliche der Geburtsjahrgänge 1980 bis 1985 mit Hauptwohnsitz in den Neuen Bundesländern (außer Berlin) insgesamt drei Mal befragt. Die Befragung deckt den Zeitraum 1995 bis 2006 ab. Von den Jugendlichen werden demografische Angaben wie Geburtsjahr, Geschlecht, Staatsbürgerschaft, Wohnort, Haushaltszugehörigkeit und die Existenz eigener Kinder erhoben.
Die Angaben zur Schulbildung beinhalten neben der Art der besuchten Schule und dem höchsten Schulabschluss auch die Abschlusszensur und den Abgangszeitpunkt.

Den Hauptinhalt der Befragung bildet die Bildungs- und Erwerbsbiografie der Jugendlichen im Anschluss an die allgemeinbildende Schule. Dazu wurden alle Ausbildungs- bzw. Arbeitsmarktepisoden nach Beendigung der Schule, dabei evtl. erworbene Abschlüsse und Zusatzqualifikationen ebenso wie Angaben zum erlernten Beruf,[7] der Branche und Betriebsgröße des Ausbildungs- bzw. Beschäftigungsbetriebes erfragt. Zusätzlich finden sich Angaben zur Art der staatlichen Förderung, zum monatlichen Nettoverdienst sowie eine Bewertung verschiedener Aspekte der Ausbildung bzw. Beschäftigung durch die befragten Jugendlichen. Darüber hinaus sind Informationen darüber, wie und mit wessen Unterstützung die Ausbildung bzw. Beschäftigung gefunden wurde, verfügbar.

In jeder der drei Wellen werden zusätzlich Fragen zu wechselnden Schwerpunktthemen gestellt (Umzugsbereitschaft und Mobilität, Einkünfte, Freizeitgestaltung und Zukunftsvorstellungen sowie soziale Netzwerke). Da diese Informationen nicht für alle befragten Jugendlichen vorliegen, können sie in der Analyse nicht berück-

[7] Der Angabe der Berufe liegt das Klassifikationssystem des Statistischen Bundesamtes von 1992 zugrunde. Für nähere Erläuterungen vgl. Tillmann (2004) S. 82.

sichtigt werden. Insgesamt liegen 32254 Episodeninformationen von 10665 befragten Jugendlichen vor.

Aufgrund seiner umfangreichen und detaillierten Informationen und der Fokussierung auf eine spezielle Personengruppe ist das Jugendpanel besser als andere verfügbare Datensätze zur Analyse des Berufseinstiegs von Jugendlichen und der Berücksichtigung verschiedener Einflüsse auf ihre Beschäftigungschancen geeignet.

5.2.2 Stichprobe Jugendlicher mit abgeschlossener Berufsausbildung

Aus diesem Datensatz werden Informationen über Jugendliche, die eine (vertraglich gebundene oder schulische) Berufsausbildung erfolgreich abgeschlossen haben, genutzt.[8] Betrachtet wird insbesondere die erste erfolgreich abgeschlossene Ausbildung. Mit dieser Einschränkung stehen für die Untersuchung des Einflusses einer Ausbildungsförderung auf den Berufseinstieg noch 9251 Episoden von 3048 Jugendlichen zur Verfügung.

Die Identifikation der geförderten Jugendlichen basiert auf Angaben, die die Befragten zu ihrem Förderstatus selbst machen. Diese Abgrenzung birgt das Problem der Unterschätzung des Anteils der geförderten Berufsausbildungen: Nicht allen Jugendlichen wird der Umstand einer staatlichen Förderung bewusst sein, etwa bei einer direkten finanziellen Unterstützung der ausbildenden Unternehmen (Steiner et al., 2004, S. 17).
Die Unterscheidung zwischen nicht geförderter und geförderter Ausbildung sowie betriebsnaher und außerbetrieblicher Förderung wird analog zu Prein (2005) und Steiner et al. (2004) vorgenommen.[9]

[8] Die Unterscheidung zwischen vertraglich gebundener und schulischer Ausbildung wird anhand der Berufsbezeichnungen getroffen und steht als Zusatzinformation im Datensatz zur Verfügung.

[9] Als geförderte Ausbildungen werden definiert: (1) Ausbildungen, bei denen der Ausbildungsvertrag mit einem Bildungsträger geschlossen wurde, (2) vertraglich gebundene Ausbildungen, bei denen die Befragten eine Förderung erhalten haben, sowie (3) schulische Ausbildungen, bei denen eine Förderung erfolgte. Von den geförderten Ausbildungen gelten als außerbetriebliche Ausbildung: (1) staatlich geförderte schulische Berufsausbildungen sowie (2) geförderte vertraglich gebundene Ausbildungen, bei denen die praktische Ausbildung nicht ausschließlich im Praktikumsbetrieb stattfand. Andere geförderte Ausbildungsverhältnisse werden als betriebsnahe Ausbildung angesehen.

Die Berufsangaben, die im Datensatz als Berufsklasseninformation vorliegen, werden für die weitere Untersuchung zu Berufsbereichen zusammengefasst. Diese Zusammenfassung basiert auf dem Klassifikationssystem des Statistischen Bundesamtes (Statistisches Bundesamt, 1992). Es liegen dann Angaben für 10 Berufsbereiche vor, die eine sehr unterschiedliche Verteilung der Anzahl absolvierter Berufsausbildungen aufweisen.[10]

Die wichtigsten Merkmale der Jugendlichen in der Stichprobe – insgesamt sowie für analyserelevante Teilstichproben – sind in der Tabelle 5.1 zusammengefasst.[11]

Die Jugendlichen in der Stichprobe insgesamt waren zum Zeitpunkt des Abschlusses der Berufsausbildung durchschnittlich 20 Jahre alt. Der Frauenanteil in der Stichprobe ist mit 46% etwas niedriger als der Anteil der Männer. Nur sehr wenige Jugendliche (3%) haben eigene Kinder. Knapp die Hälfte der befragten Jugendlichen (44%) lebt in einem eigenen Haushalt.[12] Alle befragten Jugendlichen haben die deutsche Staatsangehörigkeit, ebenso wie ihre Eltern. Nur ca. 10% der Jugendlichen besitzen die allgemeine Hochschulreife, 15% der Befragten absolvierten die Hauptschule erfolgreich. Der häufigste Abschluss ist der Realschulabschluss, den etwa zwei Drittel der Jugendlichen besitzen.

Der weitaus größte Teil der befragten Jugendlichen (90%) absolvierte die Berufsausbildung in den Neuen Bundesländern. Die vertraglich gebundene Ausbildung ist die dominierende Form der Berufsausbildung – 78% der Befragten absolvierten diese Art der Berufsausbildung.

Mehr als die Hälfte der Jugendlichen (58%) bekam ihren Ausbildungsplatz aufgrund eigener Bemühungen, ca. ein Fünftel mit Hilfe von Familie, Freunden oder Kollegen. Auch die Unterstützung öffentlicher Stellen wie Berufsberatungszentren oder der Agentur für Arbeit wurden von 27% der Jugendlichen in Anspruch genommen.

[10] In den Berufsbereichen „Bergbau und Mineralgewinnung“ sowie „sonstige Arbeitskräfte“ wurde keiner der Jugendlichen der Stichprobe ausgebildet. Sie werden deshalb in der Tabelle 5.1 nicht berücksichtigt, finden sich aber in der ausführlichen Beschreibung der Stichprobe (Tabelle C.1 im Anhang).

[11] Eine ausführliche deskriptive Statistik findet sich in Tabelle C.1 im Anhang.

[12] Dazu zählen eine eigene Wohnung, die Wohngemeinschaft mit einem festen Partner oder anderen Personen sowie Internatsaufenthalte.

Tabelle 5.1: **Zusammenfassung wichtiger Merkmale der Stichprobe Jugendlicher mit abgeschlossener Berufsausbildung**

Merkmale	gesamte Stichprobe	ungeförderte Jugendliche	geförderte Jugendliche		
			gesamt	betriebsnah	außerbetr.
Anzahl Personen	3048	2556	492	324	168
soziodemografische Faktoren					
Alter	19,96	20,00	19,73	19,67	19,85
Frau	0,46	0,44	0,55	0,60	0,46
eigene Kinder	0,03	0,03	0,02	0,02	0,02
eigener Haushalt	0,44	0,43	0,46	0,48	0,43
dt. Staatsbürger	1,00	1,00	1,00	1,00	1,00
Eltern dt. Staatsbürger	1,00	1,00	1,00	1,00	1,00
Hauptschulabschluss	0,15	0,12	0,27	0,24	0,33
Realschulabschluss	0,76	0,77	0,67	0,71	0,59
Abitur	0,09	0,10	0,03	0,03	0,04
Charakteristika der Berufsausbildung					
BAB NBL	0,90	0,90	0,90	0,87	0,96
vertragl. geb. BAB	0,78	0,79	0,68	0,68	0,68
Zusatzqualifikation	0,19	0,20	0,17	0,17	0,17
Land-, Forstwirtschaft	0,03	0,03	0,03	0,04	0,01
Metall-, Elektroberufe	0,20	0,22	0,11	0,11	0,10
Bau-, Ausbauberufe	0,08	0,09	0,07	0,04	0,13
sonst. Fertigungsberufe	0,11	0,12	0,11	0,10	0,11
technische Berufe	0,04	0,04	0,04	0,03	0,05
Waren-, DL-kaufleute	0,13	0,13	0,15	0,17	0,11
Org., Verwaltung, Büro	0,16	0,16	0,16	0,14	0,20
Gesundheitsdienst	0,09	0,09	0,12	0,13	0,11
Sozial-, Erziehungsberufe	0,05	0,05	0,06	0,07	0,04
sonst. Dienstleistungen	0,09	0,08	0,15	0,16	0,13
BAB auf eigene Initiative	0,58	0,62	0,40	0,45	0,31
BAB m.H. anderer	0,20	0,21	0,17	0,20	0,13
BAB m.H. öff. Stellen	0,27	0,23	0,50	0,45	0,60
Arbeitsmarktstatus direkt nach Ausbildungsabschluss					
Erwerbstätigkeit	0,46	0,50	0,26	0,27	0,23
Arbeitslosigkeit	0,36	0,32	0,53	0,52	0,54

Anmerkungen:
Anteil Jugendlicher mit entsprechendem Merkmal; Ausnahme: Alter (arithmetisches Mittel).
BAB – Berufsausbildung.

Quelle: *Jugendpanel des zsh.*

Im Bereich der Metall- und Elektroberufe wurden die meisten Ausbildungsabschlüsse erworben (20%).[13] Weitere häufige Berufsgruppen sind Organisations-, Verwaltungs- und Büroberufe (16%), Waren- und Dienstleistungkaufleute (13%) und sonstige Fertigungsberufe (11%).[14]
Direkt im Anschluss an die Berufsausbildung sind knapp die Hälfte der Jugendlichen erwerbstätig, ca. ein Drittel ist arbeitslos.

Ein Vergleich der dritten und vierten Spalte – also zwischen Jugendlichen, die eine ungeförderte Berufsausbildung absolvierten, und denen in einer geförderten Ausbildung – zeigt deutliche Unterschiede in einigen Merkmalen. Der Frauenanteil ist in den geförderten Ausbildungen deutlich höher (55% vs. 44%). Auch der Anteil Jugendlicher mit Hauptschulabschluss ist mit 27% ungefähr doppelt so hoch wie unter den Jugendlichen in nicht geförderten Ausbildungen. Trotzdem ist der Realschulabschluss mit 67% auch hier der häufigste Schulabschluss.

Auch hinsichtlich der Ausbildungsform lassen sich Unterschiede feststellen. Etwa drei Viertel der ungeförderten Jugendlichen absolvierten eine vertraglich gebundene Ausbildung, bei den geförderten Jugendlichen trifft dies auf ca. zwei Drittel der Personen zu.

In der Verteilung der Berufsausbildungen auf die einzelnen Berufsbereiche zeigen sich v.a. im Bereich der Metall- und Elektroberufe und im Bereich sonstige Dienstleistungen deutliche Unterschiede. Der Anteil der Ausbildungen im erstgenannten Bereich ist unter den Jugendlichen in ungeförderten Ausbildungsverhältnissen mit 22% doppelt so hoch wie bei den gefördert Ausgebildeten. Umgekehrt liegt der Anteil der Abschlüsse im Bereich der sonstigen Dienstleistungen[15] unter den geför-

[13] Zu diesem Bereich gehören bspw. Maschinen-, Anlagen- und Fahrzeugbauer, Monteure, Mechaniker, Dreher, aber auch Werkzeugmacher, Gas- und Wasserinstallateure, Zahntechniker, Optiker, Uhrmacher und Rundfunkmechaniker.

[14] Zum Berufsbereich Organisations-, Verwaltungs- und Büroberufe zählen u.a. Manager, Wirtschaftsprüfer, Steuerfachleute, kaufmännische Angestellte, Verwaltungsfachleute sowie Informatiker, Sachbearbeiter, Bürokaufleute, Sekretäre und Kassierer.
Der Berufsbereich Waren- und Dienstleistungskaufleute umfasst Groß- und Einzelhandelskaufleute, Fachverkäufer, Tankwarte sowie Bank- und Versicherungsfachleute.
Zum Bereich sonstige Fertigungsberufe zählen sehr unterschiedliche Berufe, u.a Steinmetze und Bildhauer, Töpfer, Glasbläser, Maler, Lackierer, Drucker, Berufe in der Papierherstellung und der Holzverarbeitung, Chemieberufe, Textil- und Bekleidungsberufe sowie Ernährungsberufe wie Bäcker, Fleischer, Koch.

[15] Im Bereich sonstige Dienstleistungen werden Berufe im Hotel- und Gaststättengewerbe, Friseure, Kosmetiker und Reinigungs- und Entsorgungsberufe zusammengefasst.

derten Jugendlichen bei 15%, bei ungeförderter Ausbildung nur bei 8%. Ebenso ist der Anteil der Berufe im Gesundheitsdienst[16] unter den geförderten Jugendlichen etwas höher (12% vs. 9%).

Der Anteil der Jugendlichen, die ihre Ausbildung auf eigene Initiative hin gefunden haben, ist unter den ungeförderten Personen deutlich höher (62% vs. 40%). Etwas höher ist in dieser Teilstichprobe ebenfalls der Anteil Jugendlicher, die Unterstützung durch andere Personen erhalten haben (21% vs. 17% der geförderten Jugendlichen). Dagegen wurde die Hälfte der geförderten Jugendlichen bei der Suche nach einem Ausbildungsplatz von einem Berufsberatungszentrum oder der Agentur für Arbeit unterstützt, von den ungeförderten Jugendlichen nahm nur etwa ein Viertel die Hilfe öffentlicher Beratungsstellen in Anspruch.

Direkt im Anschluss an die Ausbildung ist der Anteil der erwerbstätigen Jugendlichen bei den ungeförderten Personen mit 50% etwa doppelt so hoch wie der der geförderten Jugendlichen. Das deutet darauf hin, dass die Jugendlichen, die eine geförderte Ausbildung absolvierten, schlechtere Beschäftigungschancen haben als Jugendliche nach einer „normalen“ Ausbildung.

In den letzten beiden Spalten der Tabelle 5.1 werden die beiden Arten der Ausbildungsförderung getrennt ausgewiesen. Spalte 5 enthält Informationen über Jugendliche in einer betriebsnahen Berufsausbildung, Spalte 6 über Jugendliche in einer außerbetrieblichen Ausbildung. Beide Gruppen unterscheiden sich in einigen Merkmalen voneinander. Unter den außerbetrieblich ausgebildeten Jugendlichen sind weniger Frauen als in betriebsnahen Ausbildungen (46% vs. 60%). In dieser Teilstichprobe ist der Hauptschulabschluss häufiger als unter Jugendlichen in betriebsnahen Ausbildungen (33% vs. 24%). Auch hinsichtlich der Verteilung der Ausbildungsberufe lassen sich Unterschiede feststellen. So ist der Anteil Jugendlicher in Bau- und Ausbauberufen[17] mit 13% bei außerbetrieblichen Ausbildungen deutlich höher als bei betriebsnahen Ausbildungen mit nur 4%. Höher ist auch der Anteil von Ausbildungen in Organisations-, Verwaltungs- und Büroberufen (20% vs.

[16] Dazu zählen neben Ärzten und Apothekenberufen auch medizinisch-technische Assistenten, Krankenschwestern, Krankenpfleger und andere therapeutische Berufe.

[17] Zum Bereich Bau- und Ausbauberufe werden u.a. Maurer, Stahlbetonbauer, Pflasterer, Staßen- und Gleisbauer sowie Zimmerer, Dachdecker, Ofenbauer, Stukkateure, Glaser und Raumausstatter gezählt.

14%). Im Bereich der Waren- und Dienstleistungskaufleute sind dagegen betriebsnahe Ausbildungen häufiger (17% vs. 11%). Bei der Suche eines Ausbildungsplatzes sind Eigenitiative und die Hilfe von Familie oder Freunden unter den außerbetrieblich Ausgebildeten seltener (31% vs. 45% bzw. 13% vs. 20%), die Hilfe öffentlicher Beratungsstellen dagegen häufiger (60% vs. 45%).

Die Beschäftigungschancen sind nach abgeschlossener Ausbildung etwa gleich hoch (bzw. niedrig): Nur etwa ein Viertel der Jugendlichen in beiden Ausbildungsarten ist direkt im Anschluss an die Ausbildung erwerbstätig.

Aus der deskriptiven Analyse wird deutlich, dass die Merkmale in den betrachteten Teilstichproben nicht gleich verteilt sind. Die unterschiedlichen Beschäftigungschancen der Jugendlichen können also aus unterschiedlichen Charakteristika der Jugendlichen sowie der gewählten Berufsausbildung resultieren. Darüber hinaus ist es möglich, dass der Umstand der Förderung selbst einen Einfluss auf die Beschäftigungschancen der Jugendlichen hat. Um diesen als Stigmatisierung oder Selbstselektion bezeichneten Effekt zu untersuchen, bieten sich die in den vorangegangenen Kapiteln beschriebenen Matchingmethoden an, weil mit ihnen die Kontrolle anderer Effekte, bspw. durch persönliche Merkmale oder die Berufsausbildung, möglich ist.

5.3 Untersuchungsmethode

In der Analyse eines Stigmatisierungseffekts der Förderung wird zwischen betriebsnaher und außerbetrieblicher Ausbildung unterschieden. Für beide Formen der Berufsausbildung wird eine Vergleichsgruppe Jugendlicher aus der Teilstichprobe der ungeförderten Jugendlichen gesucht. Darüber hinaus werden beide Formen der Förderung miteinander verglichen.

Um den Effekt des negativen Images auf die Beschäftigungschancen der geförderten Jugendlichen schätzen zu können, müssen alle beschäftigungs- und förderungsrelevanten Merkmale der Jugendlichen beim Matching berücksichtigt werden. Die Auswahl der Matchingvariablen orientiert sich an theoretischen Überlegungen und früheren Studien zum Thema Beschäftigungsaussichten, insbesondere Berufs-

einstiegschancen von Jugendlichen.[18] Von Bedeutung sind v.a. soziodemografische Faktoren, Charakteristika der gewählten Berufsausbildung sowie die allgemeine Lage am Arbeitsmarkt. Aber auch die persönliche Arbeitsmarkt-Vorgeschichte sowie die Unterstützung durch andere Personen im sozialen Umfeld müssen bei der Erklärung unterschiedlicher Beschäftigungschancen berücksichtigt werden.[19]

5.3.1 Auswahl der Matchingvariablen

Aus den Daten werden Informationen über das Alter bei Abschluss der Beufsausbildung, das Geschlecht, die Art des Schulabschlusses (kein Abschluss/Hauptschulabschluss/Realschulabschluss/Abitur), die Haushaltsführung (eigener Haushalt oder Leben im Elternhaushalt) sowie die evtl. Existenz eigener Kinder als Matchingvariablen genutzt. Vom Alter eines Jugendlichen zum Zeitpunkt seines ersten Abschlusses einer Berufsausbildung wird ein negativer Einfluss auf die Beschäftigungschancen erwartet. Je jünger ein Bewerber ist, desto größer sind – ceteris paribus – seine Beschäftigungsaussichten. Ebenso wird berücksichtigt, dass Männer und Frauen unterschiedliche Aussichten auf dem Arbeitsmarkt haben, wie die geschlechtsspezifisch unterschiedlichen Arbeitslosenquoten im Beobachtungszeitraum belegen.[20] Bei gleicher beruflicher Qualifikation wird erwartet, dass ein höherer Schulabschluss einen positiven Einfluss auf die Beschäftigungschancen eines Jugendlichen hat, da der Schulabschluss als Signal für die Leistungsfähigkeit eines Bewerbers interpretiert werden kann.

Die Art der Haushaltsführung kann als Hinweis auf die Selbstständigkeit eines Jugendlichen verstanden werden. Darüber hinaus wird der Anreiz, das Leben aus eige-

[18] Da sich die Förderung der Ausbildung an den Aussichten der Jugendlichen auf dem Ausbildungsmarkt orientiert, sind mit Ausnahme der berufsspezifischen Merkmale alle beschäftigungsrelevanten Faktoren gleichzeitig relevant für die Ausbildungsförderung – mit umgekehrtem Vorzeichen. Die Bedeutung der Faktoren für die Ausbildungsförderung wird deshalb im Folgenden nicht gesondert erwähnt.

[19] In Heckman, Ichimura und Todd (1997) wird – unter Hinweis auf eigene Untersuchungen – darauf hingewiesen, dass frühere Arbeitsmarkterfahrungen von entscheidender Bedeutung für die Beurteilung von Arbeitsmarktentscheidungen und Beschäftigungsaussichten einer Person sind. Die Bedeutung sozialer Strukturen bzw. Netzwerke für die Erklärung ungleicher Erwerbsaussichten und Ausbildungsentscheidungen wird u.a. in Solga (2005) ausdrücklich betont.

[20] Zum Beginn des Beobachtungszeitraums 1995 lag die Arbeitslosenquote der Männer mit ca. 9,5% unterhalb der der Frauen (ca.12%), von 2001 bis 2006 war dagegen eine höhere Arbeitslosenquote bei den Männern zu beobachten. Für nähere Angaben vgl. Bundesagentur für Arbeit (2008).

nem Einkommen zu finanzieren, in einem eigenen Haushalt vermutlich höher sein. Eigene Kinder gelten insbesondere für junge Frauen als Beschäftigungshindernis und werden deshalb beim Matching ebenfalls berücksichtigt.
Informationen über die Staatsbürgerschaft und den Geburtsort der Jugendlichen sind in den Daten ebenfalls verfügbar, werden aber zum Matching nicht verwendet, da alle Jugendlichen in der Stichprobe die deutsche Staatsangehörigkeit besitzen und auch in Deutschland geboren worden sind. Das gleiche trifft auf die Staatsbürgerschaft der Eltern zu.
Die Angaben über die erzielte Schulabschlussnote sind nicht verwendbar, da sie nur für knapp die Hälfte der Jugendlichen in der Stichprobe vorliegt.

Beschäftigungsrelevante Charakteristika der Berufsausbildung sind u.a. die Art der Ausbildung, der Ausbildungsberuf sowie evtl. erworbene Zusatzqualifikationen. Die Angaben über den Ausbildungsberuf werden – wie in der Beschreibung der Stichprobe – aggregiert in 10 Berufsbereiche: Berufe in Land- und Forstwirtschaft, Metall- und Elektroberufe, Bau- und Ausbauberufe, sonstige Fertigungsberufe, Technische Berufe, Waren- und Dienstleistungskaufleute, Organisations-, Verwaltungs- und Büroberufe, Gesundheitsdienstberufe, Sozial-, Erziehungsberufe sowie sonstige Dienstleistungsberufe. Die Art der Ausbildung (vertraglich gebundene oder schulische Ausbildung) wird durch den Ausbildungsberuf vorgegeben. Sie wird als Zusatzinformation beim Matching berücksichtigt, da durch die Zusammenfassung der Berufe die Trennung zwischen beiden Arten der Ausbildung nicht mehr exakt möglich ist. Zusatzqualifikationen, die während der Ausbildung erworben werden, können die Beschäftigungsaussichten positiv beeinflussen und werden deshalb ebenfalls berücksichtigt. Die Information über den erworbenen Berufsabschluss ist nicht verwendbar, da sie ebenfalls nur für knapp die Hälfte aller Personen vorliegt.
Die Beurteilung der Ausbildungsinhalte durch die Jugendlichen – insbesondere die Bewertung der gestellten Anforderungen und der Belastung, die mit der Ausbildung und den Arbeitszeiten verbunden ist – gibt Auskunft darüber, wie gut ein Jugendlicher in dem erlernten Beruf, insbesodere mit den gestellten Anforderungen, zurecht kommt. Sie wird ebenfalls beim Matching berücksichtigt.

Zur Berücksichtigung der allgemeinen Lage auf dem Arbeitsmarkt werden Informationen über den Ausbildungsort und den Zeitpunkt des Ausbildungsabschlusses

verwendet. Der Ausbildungsort wird auf Bundesländerebene grob in drei Regionen zusammengefasst: Mitteldeutschland (Sachsen, Sachsen-Anhalt und Thüringen), Nordostdeutschland (Brandenburg und Mecklenburg-Vorpommern) und die Alten Bundesländer incl. Berlin.[21] Die Aussichten auf dem Arbeitsmarkt sind in der letztgenannten Region besser als in beiden anderen Regionen.[22] Der Zeitpunkt des Ausbildungsabschlusses fällt für die Jugendlichen der Stichprobe in den Zeitraum zwischen 1995 und 2006. Diese 12 Jahre werden zu drei Zeiträumen zusammengefasst. Die Arbeitsmarktsituation der Jahre 1995-1998 ist gekennzeichnet durch eine stark wachsende Arbeitslosenquote, in den Jahren 1999-2002 ist die Arbeitslosenquote auf hohem Niveau relativ stabil, in der Phase 2002-2006 steigt sie erneut leicht an.[23]

Neben den genannten Angaben stehen Informationen darüber zur Verfügung, ob die Berufsausbildung direkt im Anschluss an die Schule begonnen wurde – und wenn nicht, was die Jugendlichen vor Beginn der Ausbildung gemacht haben. Interessant für die Beschäftigungschancen sind v.a. Informationen darüber, ob im Vorfeld eine Ausbildung begonnen aber nicht erfolgreich abgeschlossen wurde, die Absolvierung des Wehr- oder Zivildienstes oder eines Berufsvorbereitungsjahres sowie evtl. Erwerbstätigkeits- und Arbeitslosigkeitsphasen. Diese Angaben werden beim Matching ebenfalls berücksichtigt.

Ein weiterer wichtiger Aspekt bei der Beurteilung von Berufseinstiegschancen ist das soziale Umfeld der Jugendlichen. In der Literatur zur Erklärung ungleicher Bildungschancen – als Voraussetzung unterschiedlicher Beschäftigungsaussichten von Jugendlichen – wird der Einfluss des familiären Hintergrundes auf Bildungs- und Ausbildungsentscheidungen ausdrücklich betont.[24] In der Analyse liegt der Fokus allerdings weniger auf der Erklärung unterschiedlicher Bildungsentscheidungen

[21] Trotz ihrer Heterogenität werden diese 11 Länder zu einer Region zusammengefasst, da der Anteil der Jugendlichen, die dort ihre Ausbildung absolvieren, vergleichsweise gering ist (10%), wie aus der deskriptiven Analyse der Stichprobe hervorgeht.

[22] Diese Erwartung wird durch den Vergleich der Arbeitslosenquoten im Zeitraum zwischen 1995 und 2006 in den einzelnen Bundesländern gestützt (Bundesagentur für Arbeit, 2008).

[23] Die Einteilung dieses Zeitraums in drei Phasen orientiert sich an Angaben der Bundesagentur für Arbeit (2008).

[24] Der Zusammenhang zwischen dem Bildungsstand und der „Lebensweise“ der Eltern (z.B. Arbeitsmarktstatus und berufliche Stellung der Eltern oder Anzahl Bücher im Elternhaushalt) und den Bildungsentscheidungen und -erfolgen von Kindern und Jugendlichen wird in der Literatur häufig thematisiert (von Below, 1999; Woessmann, 2004).

in der Vergangenheit als vielmehr auf der Erklärung unterschiedlicher Berufseinstiegschancen – unter Berücksichtigung des Ergebnisses früherer Bildungsentscheidungen (dem Schulabschluss). Ein weiterer Einflussfaktor auf den Arbeitsmarkterfolg ist die Unterstützung, die ein Jugendlicher bei der Suche nach einem Ausbildungsplatz oder einer Beschäftigung von seinem Umfeld erhält. Im Datensatz sind Informationen darüber verfügbar, die für das Matching zu einem Netzwerkindikator zusammengefasst werden. Dieser umfasst sowohl die Unterstützung durch die Familie als auch durch Freunde oder Kollegen.

Zusätzlich lässt sich aus den Daten ein Indikator für die Motivation und Aktivität der Jugendlichen auf dem Arbeitsmarkt konstruieren. Wenn ein Jugendlicher aufgrund eigener Bemühungen einen Ausbildungsplatz gefunden oder eine Zusatzqualifikation auf eigene Initiative hin erworben hat, wird unterstellt, dass dieser Jugendliche sein Berufsleben bewusst und aktiv gestalten will, was sich vorteilhaft auf den Berufseinstieg auswirken sollte.

Die Indikatoren für Netzwerk und Motivation ermöglichen es – zusammen mit den Informationen über die Arbeitsmarkt-Vorgeschichte und die Beurteilung der Ausbildung durch den Jugendlichen – auch unbeobachtbare Heterogentäten der Jugendlichen hinsichtlich ihres Verhaltens auf dem Arbeitsmarkt zu berücksichtigen und damit die Selektionsverzerrung vollständig durch Matching zu kontrollieren.

5.3.2 Auswahl der Matchingmethode

Für die Zuordnung von Partnern zu den Jugendlichen in den einzelnen Teilstichproben müssen die genannten Merkmale in einem aggregierten Distanzmaß zusammengefasst werden. Aus den Ergebnissen der Simulation im vorangegangenen Kapitel geht hervor, dass die gewichtete Mahalanobis-Matching-Distanz insbesondere zur Zusammenfassung nominal skalierter – dichotomer und polytomer – Merkmale geeignet ist. Da alle genannten Einflussfaktoren – mit Ausnahme der Altersangabe – in Form nominaler Variablen vorliegen, wird dieses Maß als Grundlage der Zuordnungsprozesse verwendet.

Für die Analysen stehen jeweils unterschiedlich große Stichproben mit potenziellen Partnern für die betrachteteten Teilstichproben der Jugendlichen zur Verfügung. Die

Untersuchung des Stigmatisierungseffekts einer betriebsnahen Ausbildung basiert auf dem Vergleich der so geförderten Personen mit Jugendlichen in ungeförderten Ausbildungen. Das gleiche gilt für die Analyse der außerbetrieblichen Ausbildung. Die Gruppe der potenziellen Partner ist in beiden Fällen relativ groß – die Stichprobe der ungeförderten Jugendlichen umfasst 2556 Personen im Vergleich zu 324 Jugendlichen in betriebsnahen und 168 in außerbetrieblichen Ausbildungen. Nach den Ergebnissen der Simulation würde man erwarten, dass mit einem Optimal Full Matching die besten Matchingergebnisse erzielt werden.

Im Fall des direkten Vergleichs beider Arten der Ausbildungsförderung unterscheidet sich die Anzahl der Personen in beiden Gruppen dagegen nicht sehr stark. Den Simulationsergebnissen zufolge wäre in diesem Fall eine Zuordnung mit Zurücklegen das beste Verfahren.

Allerdings stimmen die Simulationsergebisse nicht in allen betrachteten Gütemaßen überein, die Rangfolge der Zuordnungsprozesse variiert hinsichtlich der verschiedenen Maße. Deshalb werden zur Prüfung der Robustheit der Matchingergebnisse zusätzlich zu den genannten Verfahren zwei 1:1-Matchingalgorithmen (ein optimales Nearest Neighbor Matching und Random Matching) zur Ermittlung der Vergleichsgruppen eingesetzt. Die Ergebnisse der Zuordnungsprozesse werden miteinander verglichen, und die Vergleichsgruppe des besten Verfahrens wird zur Untersuchung des Stigmatisierungseffekts einer Förderung verwendet.
Vor dem Einsatz der verschiedenen Zuordnungsprozesse wird die Einhaltung der Common-Support-Bedingung für die betrachteten Teilstichproben überprüft. Dazu werden die Ausprägungen der einzelnen Merkmale in beiden Teilstichproben – wie in Abschnitt 4.2.2 beschrieben – miteinander verglichen.[25]

Zur Überprüfung der Matchingergebnisse werden die im gleichen Abschnitt vorgestellt Tests verwendet. Um diese Tests auch auf die Ergebnisse des Optimal Full Matching anwenden zu können, wird vorher eine „repräsentative Ausprägung“ der Merkmale in der jeweiligen Unter-Kontrollgruppe eines geförderten Jugendlichen ermittelt.[26] Für metrisch skalierte Merkmale wird dazu der Median, für nominal

[25]Im Fall des Vergleichs beider Förderarten werden zwei außerbetrieblich ausgebildete Jugendliche aufgrund der Nichteinhaltung der Common-Support-Bedingung ausgeschlossen.

[26]Beim Full Matching werden alle Personen der Vergleichsgruppe – alle potenziellen Partner – unter den geförderten Jugendlichen aufgeteilt. Jede Person kann also mehr als einen Partner haben.

skalierte der Modalwert verwendet. Darüber hinaus wird die Summe der quadrierten Distanzen zwischen der betrachteten Stichprobe und der Kontrollgruppe angegeben.
Beim direkten Vergleich der beiden Förderarten kann das Problem des Verlustes von Beobachtungen auftreten. Die Anzahl evtl. ausgeschlossener Jugendlicher wird deshalb im letzten Schritt der Analyse zusätzlich zu den genannten Gütemaßen berücksichtigt.

Die Reduzierung des Bias durch Matching in den einzelnen Merkmalen ist nur begrenzt aussagefähig, da sie auf dem Vergleich des arithmetischen Mittels der Merkmale beruht, der Großteil der Matchingvariablen aber nominal skalierte dichotome und polytome Variablen sind. Sie wird deshalb nicht zur Beurteilung der Güte der Ergebnisse eingesetzt. Die Ermittlung des mittleren quadratischen Fehlers ist nicht möglich, weil der „wahre“ Effekt der Förderung nicht bekannt ist.

Die Überprüfung der Matchingergebnisse anhand der genannten Gütemaße wird in den Tabellen C.2 bis C.4 im Anhang dokumentiert. Wie aus den Testergebnissen hervorgeht, gelingt – entgegen der Erwartung – die Angleichung der Merkmale in den zu vergleichenden Stichproben mit Hilfe des Optimal Full Matching nicht vollständig. In beiden Fällen (sowohl für die Stichprobe der außerbetrieblich ausgebildeten Jugendlichen als auch bei betriebsnahen Ausbildungen) sind die Häufigkeitsverteilungen einiger Merkmale nach dem Matching noch unterschiedlich.[27]
Dagegen können mit der Zuordnung mit Zurücklegen in beiden Fällen die Verteilungen der Merkmale in den betrachteten Teilstichproben angeglichen werden.[28]

Von der Menge aller einem Jugendlichen zugeordneten Personen wird für die Durchführung der Tests jeweils der Median bzw. Modalwert eines Merkmals mit der entsprechenden Ausprägung beim geförderten Jugendlichen verglichen.

[27] Für außerbetriebliche Ausbildungen trifft dies für verschiedene Berufsbereiche, den Zeitpunkt des Ausbildungsabschlusses, einige Aspekte der Arbeitsmarkt-Vorgeschichte sowie einige der zusätzlich gebildeten Indikatoren zur Berücksichtigung der unbeobachtbaren Heterogenität zu (vgl. Tabelle C.2). Im Fall der betriebsnahen Ausbildung gelingt die Angleichung des Alters, des Geschlechts, des Abiturientenanteils, der Art der Ausbildung, des Zeitpunktes des Ausbildungsabschlusses sowie einiger Indikatoren zur Kontrolle der unbeobachtbaren Heterogenität nicht (vgl. Tabelle C.3).

[28] Die Ergebnisse sprechen dafür, dass die Summe der quadrierten Distanzen – wie im vorangegangenen Kapitel vermutet – ein guter Indikator für die Ähnlichkeit der Verteilungen der Merkmale in beiden Teilstichproben ist.

Im Fall der betriebsnahen Ausbildung ist das Durchschnittsalter der geförderten Jugendlichen nach dem Matching um 0,3 Jahre höher als in der Vergleichsgruppe. Von diesem Unterschied wird allerdings kein Effekt auf die Beschäftigungschancen erwartet. Die Analyse des Stigmatisierungseffekts wird deshalb für beide Förderarten auf Grundlage der durch Ziehen mit Zurücklegen gebildeten Vergleichsgruppen durchgeführt.

Für den direkten Vergleich beider Förderarten gelingt die Angleichung der Merkmale auch mit der Zuordnung mit Zurücklegen nicht vollständig (vgl. Tabelle C.4). Die Jugendlichen in beiden Gruppen unterscheiden sich nach dem Matching in ihrer Einschätzung der Anforderungen der Berufsausbildung sowie im Alter noch geringfügig. Darüber hinaus sind Arbeitslosigkeitsphasen und abgebrochene Ausbildungen vor Beginn der Berufsausbildung unter den außerbetrieblich Ausgebildeten häufiger. Insbesondere von den Unterschieden in der Arbeitsmarkt-Vorgeschichte wird ein Effekt auf die Beschäftigungsaussichten erwartet. Sowohl Arbeitslosigkeitsphasen als auch Ausbildungsphasen ohne erfolgreichen Abschluss werden die Aussichten auf Erwerbstätigkeit eher verringern. Dieser Einfluss der Arbeitsmarkt-Vorgeschichte muss bei der Interpretation der Ergebnisse berücksichtigt werden.

5.4 Effekte der Berufsausbildungsförderung

Für die Untersuchung des Stigmatisierungseffekts der Ausbildungsförderung auf die Beschäftigungschancen werden verschiedene Kriterien eingesetzt. Zum einen wird der quantitative Effekt der Förderung auf den Anteil der Jugendlichen, die eine Beschäftigung finden, festgestellt. Zum anderen werden verschiedene qualitative Merkmale der aufgenommenen Erwerbstätigkeit verglichen. Neben der beruflichen Stellung werden dabei insbesondere Merkmale atypischer Beschäftigung (befristete Beschäftigung, Teilzeit- oder geringfügige Beschäftigung) berücksichtigt, da solche Charakteristika auf eine potenziell unsicherere Arbeitsmarktposition hindeuten als „Normalarbeitsverhältnisse“ (unbefristete Vollzeitstellen).

5.4.1 Effekte der außerbetrieblichen Ausbildung

Zunächst werden die Ergebnisse für die außerbetriebliche Ausbildung betrachtet. In der Abbildung 5.1 wird der quantitative Beschäftigungseffekt der außerbetrieblichen Förderung direkt nach Abschluss der Berufsausbildung sowie im gesamten Beobachtungszeitraum dargestellt.

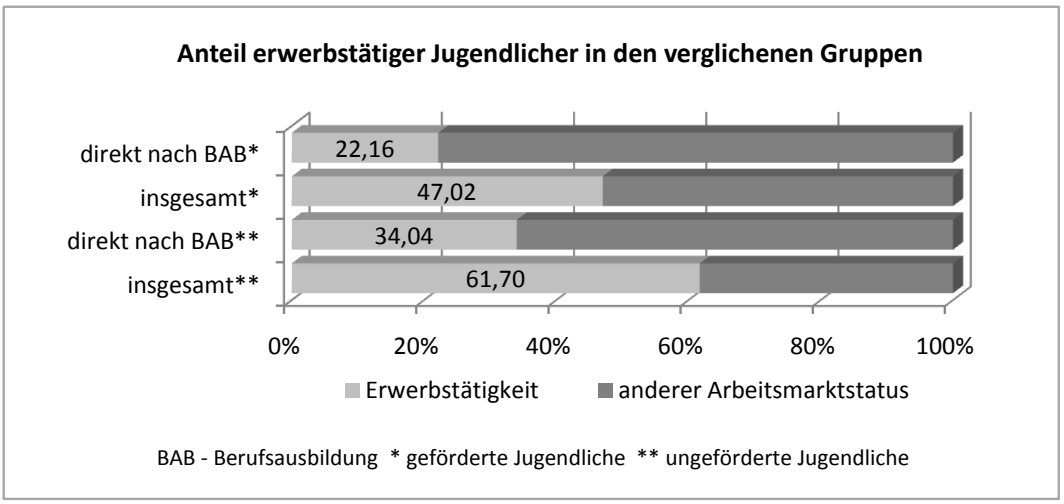

Abbildung 5.1: *Quantitativer Beschäftigungseffekt der außerbetrieblichen Ausbildung*
Quelle: Jugendpanel des zsh; eigene Berechnungen.

Direkt nach Abschluss der Berufsausbildung ist nur ca. ein Fünftel der außerbetrieblich ausgebildeten Jugendlichen erwerbstätig, in der Vergleichsgruppe ist es ca. ein Drittel.[29] Der Anteil erwerbstätiger Jugendlicher steigt bei den geförderten Jugendlichen auf knapp die Hälfte, bei den ungeförderten Jugendlichen auf ca. 60%. Sowohl direkt im Anschluss an die Berufsausbildung als auch im weiteren Verlauf des Beobachtungszeitraums sind die Beschäftigungschancen der außerbetrieblich geförderten Jugendlichen also geringer. Diese Beobachtung bestätigt den vermuteten Stigmatisierungseffekt der Förderung, der die Beschäftigungschancen der außerbetrieblich ausgebildeten Jugendlichen negativ beeinflusst.

In der Tabelle 5.2 werden einige Merkmale der von den Jugendlichen im Beobachtungszeitraum aufgenommenen Beschäftigungen näher betrachtet. Die geförderten

[29] In der Tabelle C.5 im Anhang finden sich detaillierte Angaben zum Arbeitsmarktstatus der Jugendlichen direkt nach Ausbildungsabschluss. Aus dieser Tabelle geht hervor, dass in beiden Gruppen der Anteil arbeitsloser Jugendlicher mit 55% und 48% sehr hoch ist.

Jugendlichen brauchen für den Übergang in Erwerbstätigkeit durchschnittlich 6 Monate, die Jugendlichen der Vergleichsgruppe nur 4 Monate. Hinsichtlich der aufgenommenen Beschäftigung lassen sich weitere interessante Unterschiede feststellen. So arbeiten nur ca. zwei Drittel der Absolventen einer außerbetrieblichen Ausbildung im erlernten Beruf, bei den ungeförderten Jugendlichen sind es vier Fünftel.

Tabelle 5.2: **Erwerbstätigkeit nach der Berufsausbildung – außerbetriebliche Ausbildung –**

Tätigkeitsmerkmale	geförderte Jugendliche	ungeförderte Jugendliche
Anzahl Personen	168	141
Anteil erwerbstätiger Personen	47,02	61,70
Art der Beschäftigung		
Vollzeit	90,63	83,75
Teilzeit	7,81	16,25
unterschiedlich	1,56	0,00
Art des Vertrages		
unbefristet	46,03	43,04
befristet	49,21	51,90
kein Vertrag / selbstständig	4,76	5,06
berufliche Stellung		
Beschäftigte ohne Abschluss[a]	19,04	16,25
Beschäftigte mit Abschluss[b]	73,01	71,25
höherqualifizierte Beschäftigte[c]	7,94	8,75
Führungskräfte[d]	0,00	1,25
Selbstständige[e]	0,00	2,50
regelmäßige Überstunden		
nein	53,13	51,90
ja	46,88	48,10
Beschäftigung im erlernten Beruf	66,67	82,05
Dauer bis zum Übergang (Monate)	6,19	4,03
monatl. Nettoentgelt (Euro)[f]	944,96	1033,50

Anmerkungen:
Angaben in Prozent.
[a] an-, ungelernte Arbeiter, Angestellte ohne Abschluss;
[b] Facharbeiter, Angestellte mit Abschluss, Beamte im einfachen Dienst;
[c] qualifizierte Angestellte, Beamte im mittleren Dienst;
[d] Führungskräfte, Beamte im höheren Dienst;
[e] Selbstständige, Freiberufler.
[f] Angabe für ca. 80% der Erwerbstätigen verfügbar.

Quelle: *Jugendpanel des zsh; eigene Berechnungen.*

Der überwiegende Teil der Jugendlichen – knapp drei Viertel in beiden Gruppen – ist als Facharbeiter oder mit vergleichbarem Stellenprofil (Beschäftigte mit Ab-

schluss) beschäftigt, knapp ein Fünftel (19% bzw. 16%) als ungelernt. Nur ein sehr geringer Anteil der Jugendlichen in beiden Gruppen ist in anpruchsvolleren Tätigkeiten beschäftigt. Das monatliche Nettoeinkommen unterscheidet sich um ca. 10% (944 Euro im Vergleich zu 1033 Euro).[30]

Hinsichtlich der beruflichen Stellung, der Befristung der Tätigkeit und zu leistender Überstunden unterscheiden sich die Beschäftigungsverhältnisse nur in geringem Maße. Etwa die Hälfte der erwerbstätigen Jugendlichen in beiden Gruppen hat einen befristeten Arbeitsvertrag. Der Anteil der Teilzeitbeschäftigungen ist mit 16% bei den ungeförderten Jugendlichen doppelt so hoch wie bei den geförderten Jugendlichen mit 8%. Der überwiegende Teil der Jugendlichen ist in beiden Gruppen allerdings vollzeitbeschäftigt (90% bzw. 84%).

Für die außerbetrieblich ausgebildeten Jugendlichen sind die Beschäftigungschancen insgesamt geringer. Wenn eine Erwerbstätigkeit aufgenommen wird, gelingt dies den geförderten Jugendlichen seltener im erlernten Beruf, was sich auf das Einkommen und – in geringerem Maße auch – auf die berufliche Stellung auswirkt. Ein Hinweis darauf, dass die geförderten Jugendlichen häufiger in atypischen Beschäftigungsverhältnissen tätig sind, findet sich in den Ergebnissen allerdings nicht.

5.4.2 Effekte der betriebsnahen Ausbildung

Die Ergebnisse der Analyse der betriebsnahen Ausbildung zeigen ein ähnliches Bild. Der quantitative Effekt der Förderung ist in der Abbildung 5.2 dargestellt.

Daraus geht hervor, dass der Anteil der Erwerbstätigen auch unter den betriebsnah ausgebildeten Jugendlichen direkt im Anschluss an die Berufsausbildung mit 27% geringer als der in der Vergleichsgruppe mit 44% ist.[31] Im Zeitverlauf erhöht sich der Anteil der erwerbstätigen Jugendlichen auf 57% bei den betriebsnah geförderten Jugendlichen, in der Vergleichsgruppe auf 67%.

[30] Angaben zum Nettoeinkommen liegen für ca. 80% der Erwerbstätigen in den Gruppen vor.

[31] Zu diesem Zeitpunkt ist der Anteil der Arbeitslosen mit über 50% der geförderten Jugendlichen deutlich höher als in der Vergleichsgruppe mit 40%. Vgl. Tabelle C.5 im Anhang.

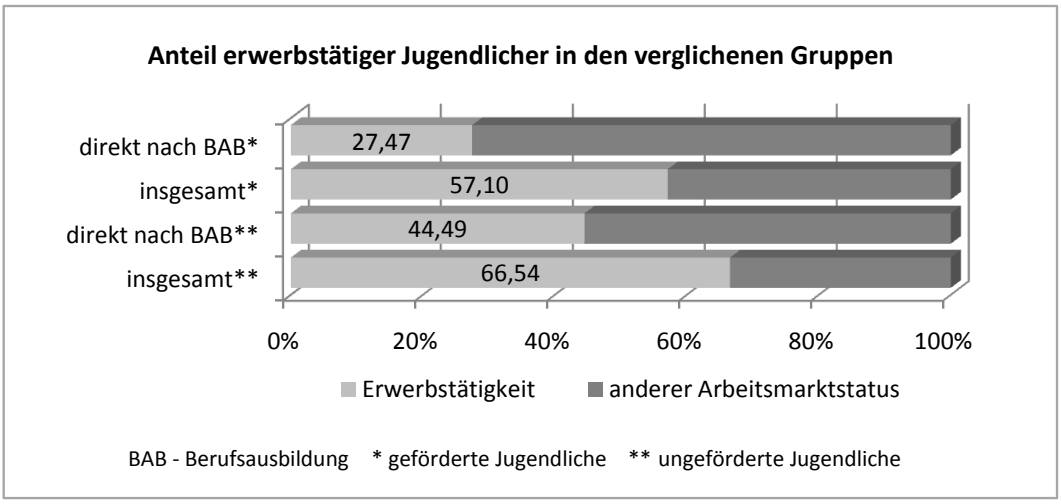

Abbildung 5.2: *Quantitativer Beschäftigungseffekt der betriebsnahen Ausbildung*
Quelle: Jugendpanel des zsh; eigene Berechnungen.

In der Tabelle 5.3 werden einige Merkmale der aufgenommen Beschäftigung zusammengefasst. Es ist festzustellen, dass der Übergang in Erwerbstätigkeit für die geförderten Jugendlichen durchschnittlich anderthalb mal so lange dauert (4,5 im Vergleich zu 3 Monaten). Der Anteil der Erwerbstätigen, die in ihrem erlernten Beruf tätig sind, ist in der Gruppe der betriebsnah Geförderten mit 70% um ca. 15%-Punkte geringer als unter den ungeförderten Jugendlichen. Das spiegelt sich auch in dem höheren Anteil Beschäftigter ohne Abschluss wider, der unter den geförderten Personen mit ca. 18% mehr als doppelt so hoch ist wie in der Vergleichsgruppe. Der Anteil Beschäftigter mit Abschluss liegt in beiden Gruppen bei ca. drei Viertel der Erwerbstätigen. Unter den betriebsnah Ausgebildeten ist der Anteil der höher qualifizierten Beschäftigten mit insgesamt ca. 9% halb so hoch wie unter den ungefördert ausgebildeten Jugendlichen. Auch das erzielte monatliche Nettoeinkommen liegt mit durchschnittlich ca. 910 Euro um 50 Euro unter dem der Absolventen ungeförderter Ausbildungsgänge.[32] Hinsichtlich der Art der Beschäftigung, der Art des Vertrages und zu leistender Überstunden unterscheiden sich die Beschäftigungsverhältnisse in beiden Gruppen dagegen nur in sehr geringem Maße.

[32] Auch hier liegen die Einkommensangaben nicht vollständig vor. Sie sind für ca. 75% der erwerbstätigen Jugendlichen verfügbar.

Tabelle 5.3: **Erwerbstätigkeit nach der Berufsausbildung – betriebsnahe Ausbildung –**

Tätigkeitsmerkmale	geförderte Jugendliche	ungeförderte Jugendliche
Anzahl Personen	324	254
Anteil erwerbstätiger Personen	57,10	66,54
Art der Beschäftigung		
Vollzeit	82,46	83,02
Teilzeit	15,79	15,09
geringfügig	0,58	0,63
unterschiedlich	1,17	1,26
Art des Vertrages		
unbefristet	42,11	47,80
befristet	57,89	48,43
kein Vertrag / selbstständig	0,00	3,78
berufliche Stellung		
Beschäftigte ohne Abschluss[a]	17,75	8,28
Beschäftigte mit Abschluss[b]	73,37	75,80
höherqualifizierte Beschäftigte[c]	8,88	12,10
hochqualifizierte Beschäftigte[d]	0,00	0,64
Führungskräfte[e]	0,00	1,27
Selbstständige[f]	0,00	1,91
regelmäßige Überstunden		
nein	50,29	50,00
ja	49,71	50,00
Beschäftigung im erlernten Beruf	69,59	85,99
Dauer bis zum Übergang (Monate)	4,55	2,99
monatl. Nettoentgelt (Euro)[g]	908,32	960,51

Anmerkungen:
Angaben in Prozent.
[a] an-, ungelernte Arbeiter, Angestellte ohne Abschluss;
[b] Facharbeiter, Angestellte mit Abschluss, Beamte im einfachen Dienst;
[c] qualifizierte Angestellte, Beamte im mittleren Dienst;
[d] Meister, hochqualifizierte Angestellte, Beamte im gehobenen Dienst;
[e] Führungskräfte, Beamte im höheren Dienst;
[f] Selbstständige, Freiberufler.
[g] Angabe für ca. 75% der Erwerbstätigen verfügbar.

Quelle: *Jugendpanel des zsh; eigene Berechnungen.*

Sowohl quantitativ als auch in Bezug auf die Merkmale der aufgenommenen Beschäftigung lässt sich über den gesamten Beobachtungszeitraum ein negativer Effekt der Förderung auf die Beschäftigungsaussichten der Absolventen betriebsnah geförderter Berufsausbildungen feststellen. Dies trifft in besonderem Maße auf die berufliche Stellung und das erzielte Einkommen zu. Allerdings findet sich auch in

dieser Gruppe kein Hinweis auf eine häufigere Beschäftigung der geförderten Jugendlichen in atypischen Beschäftigungsverhältnissen.

5.4.3 Effekte beider Arten der Ausbildungsförderung im Vergleich

Der Vergleich zwischen beiden Arten der Förderung ist nur eingeschränkt möglich, da sich die Arbeitsmarkt-Vorgeschichte der verglichenen Gruppen auch nach dem Matching noch unterscheidet. Von den unter den außerbetrieblich ausgebildeten Jugendlichen etwas häufiger auftretenden Arbeitslosigkeitsphasen und abgebrochenen Ausbildungen vor Beginn der geförderten Berufsausbildung wird ein negativer Effekt auf die Beschäftigungsaussichten erwartet, der den Stigmatisierungseffekt verstärken könnte. Bei der Interpretation der Ergebnisse wird dies berücksichtigt.

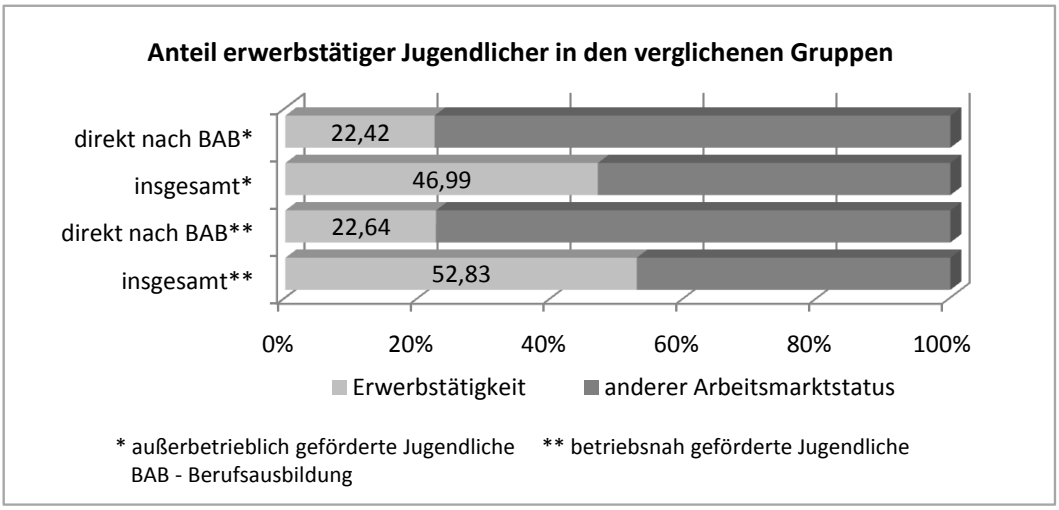

Abbildung 5.3: *Quantitativer Beschäftigungseffekt beider Förderarten im Vergleich*
Quelle: Jugendpanel des zsh; eigene Berechnungen.

Wie aus der Abbildung 5.3 ersichtlich ist, ist direkt im Anschluss an die Berufsausbildung in beiden Gruppen knapp ein Viertel der Jugendlichen erwerbstätig. Zu diesem Zeitpunkt ist der Unterschied im Beschäftigungserfolg also relativ gering. Der etwas höhere Anteil arbeitsloser Jugendlicher unter den außerbetrieblich geförderten Jugendlichen kann aus der Arbeitsmarkt-Vorgeschichte oder einem stärker ausgeprägten Stigmatisierungseffekt resultieren.[33] Der Anteil der Erwerbstätigen steigt

[33] Vgl. die Tabelle C.7 im Anhang.

auf 47% bzw. 53% der geförderten Jugendlichen. Die Beschäftigungssituation der betriebsnah geförderten Jugendlichen ist also auch im Zeitverlauf etwas positiver.[34]

In der Tabelle 5.4 werden die Charakteristika der aufgenommenen Beschäftigung dargestellt. Hinsichtlich der Dauer bis zur Aufnahme der Erwerbstätigkeit, der Beschäftigung im erlernten Beruf und des Nettoeinkommens unterscheiden sich die Jugendlichen in beiden Gruppen nur sehr wenig.[35] Auch die Tätigkeitsmerkmale der Beschäftigung (berufliche Stellung, Art der Beschäftigung und zu leistende Überstunden) sind für beide Gruppen sehr ähnlich. Es überwiegen Vollzeitbeschäftigungen (90% bei den außerbetrieblich Ausgebildeten, 87% bei den betriebsnah geförderten Jugendlichen) und Beschäftigungsverhältnisse als Facharbeiter oder mit vergleichbarem Stellenprofil (Beschäftigte mit Abschluss).

Bei den betriebsnah Geförderten überwiegen befristete Arbeitsverträge mit 60% im Vergleich zu 40% unbefristeten Verträgen, bei den außerbetrieblich geförderten Jugendlichen sind die Anteile befristeter und unbefristeter Beschäftigunsverhältnisse nahezu gleich groß (48% bzw. 47%).

Insgesamt sind die Unterschiede – sowohl quantitativ als auch in Bezug auf die Qualität der Beschäftigung – zwischen beiden Arten der Förderung nur sehr gering. Der erwartete stärkere Stigmatisierungseffekt in außerbetrieblichen Berufsausbildungen lässt sich mit den Ergebnissen der Untersuchung nicht bestätigen.

Die erste Hypothese, nach der die Absolventen geförderter Berufsausbildungen aufgrund des Stigmatisierungseffekts schlechtere Chancen auf einen qualifikationsadäquaten Berufseinstieg haben als Absolventen ungeförderter Berufsausbildungen, wird mit den Untersuchungsergebnissen bestätigt. Für beide Förderarten wird ein negativer Fördereffekt auf den Anteil der Jugendlichen, die eine Erwerbstätigkeit aufnehmen, beobachtet. Dieser Befund stimmt mit den Ergebnissen der Studien von Prein (2005) und Steiner et al. (2004) überein.

[34] Allerdings ist der Einfluss eines evtl. stärkeren Stigmatisierungseffekts auf dieses Ergebnis nicht von dem der Unterschiede in der Arbeitsmarkt-Vorgeschichte zu isolieren. Die geringfügig schlechteren Beschäftigunschancen der außerbetrieblich ausgebildeten Jugendlichen können deshalb nicht als Ergebnis eines stärkeren Stigmatisierungseffekts interpretiert werden.

[35] Informationen über das monatliche Nettoeinkommen liegen nur für ca. 65% der Erwerbstätigen vor. Die angegebenen Durchschnittswerte sind also nicht repräsentativ für die Gesamtgruppen und können deshalb hier nur als ergänzende Information angesehen werden.

Tabelle 5.4: **Erwerbstätigkeit nach der Berufsausbildung – Vergleich der Förderungen –**

Tätigkeitsmerkmale	außerbetriebliche Ausbildung	betriebsnahe Ausbildung
Anzahl Personen	166	106
Anteil erwerbstätiger Personen	46,99	52,83
Art der Beschäftigung		
Vollzeit	90,48	86,54
Teilzeit	7,94	13,46
unterschiedlich	1,59	0,00
Art des Vertrages		
unbefristet	46,77	40,38
befristet	48,39	59,62
kein Vertrag / selbstständig	4,84	0,00
berufliche Stellung		
Beschäftigte ohne Abschluss[a]	19,04	23,53
Beschäftigte mit Abschluss[b]	73,01	68,63
höherqualifizierte Beschäftigte[c]	7,94	7,84
regelmäßige Überstunden		
nein	52,38	51,92
ja	47,62	48,08
Beschäftigung im erlernten Beruf	66,13	61,54
Dauer bis zum Übergang (Monate)	5,44	5,20
monatl. Nettoentgelt (Euro)[d]	944,65	969,95

Anmerkungen:
Angaben in Prozent.
[a] an-, ungelernte Arbeiter, Angestellte ohne Abschluss;
[b] Facharbeiter, Angestellte mit Abschluss, Beamte im einfachen Dienst;
[c] qualifizierte Angestellte, Beamte im mittleren Dienst.
[d] Angabe für ca. 65% der Erwerbstätigen verfügbar.

Quelle: *Jugendpanel des zsh; eigene Berechnungen.*

Darüber hinaus wird festgestellt, dass die geförderten Jugendlichen längere Zeit bis zur Aufnahme einer Beschäftigung brauchen, seltener im erlernten Beruf tätig sind und ein geringeres Einkommen erzielen als Absolventen ungeförderter Berufsausbildungen.

Dagegen sind hinsichtlich der Art der Beschäftigung (Vollzeit-/Teilzeit-/geringfügige Beschäftigung) und der Art des Vertrages (befristet vs. unbefristet) keine nennenswerten Unterschiede zwischen den geförderten und den ungeförderten Jugendlichen zu finden. Dieses Ergebnis weicht von den Aussagen in Berger et al. (2007) ab, nach denen Absolventen geförderter Berufsausbildungsgänge deutlich häufiger in atypischen Beschäftigungsverhältnissen zu finden sind.

5.5 Zusammenfassung

In der empirischen Untersuchung dieses Kapitels wird die Förderung der Berufsausbildung in den Neuen Bundesländern evaluiert. Als geförderte Berufsausbildung werden dabei betriebsnahe und außerbetiebliche Ausbildungen angesehen.

Die Datenbasis der Analyse bildet das Jugendpanel des Zentrums für Sozialforschung Halle. Aus diesem Datensatz wird eine Stichprobe aller Jugendlichen, die eine Berufsausbildung erfolgreich abgeschlossen haben, gezogen.

Aus der deskriptiven Analyse der Stichprobe wird deutlich, dass die Merkmale in den Teilstichproben der ungeförderten und der (betriebsnah und außerbetrieblich) gefördeten Jugendlichen nicht gleich verteilt sind. Die unterschiedlichen Beschäftigungschancen der Jugendlichen in diesen Teilstichproben resultieren also – zumindest teilweise – aus unterschiedlichen Charakteristika der Jugendlichen und der gewählten Berufsausbildung. Ob darüber hinaus der Umstand der Förderung selbst einen Einfluss auf die Beschäftigungschancen der Jugendlichen hat, wird mit Hilfe des Matchingansatzes festgestellt.
Zur Ermittlung des Stigmatisierungseffekts gehört sowohl die Analyse des quantitativen Effekts der Förderung auf den Anteil der Jugendlichen, die eine Beschäftigung finden, als auch die Betrachtung verschiedener qualitativer Merkmale der aufgenommenen Erwerbstätigkeit.

Für beide Arten der Förderung wird ein negativer Beschäftigungseffekt festgestellt. Sowohl für die außerbetrieblich ausgebildeten Jugendlichen als auch die Absolventen betriebsnaher Ausbildungen sind die Beschäftigungschancen geringer als für die ungeförderten Jugendlichen. Die Jugendlichen, die nach der Ausbildung eine Erwerbstätigkeit aufgenommen haben, sind seltener im erlernten Beruf tätig, was sich v.a. auf das Einkommen und die berufliche Stellung auswirkt.

Der Vergleich zwischen beiden Arten der Förderung ergibt keinen Hinweis auf eine besonders starke Benachteiligung der außerbetrieblich geförderten Jugendlichen.

Kapitel 6

Zusammenfassung der wichtigsten Ergebnisse dieser Arbeit

Im Fokus der vorliegenden Arbeit steht die Frage nach der Eignung verschiedener Matchingverfahren für die empirische Evaluationsforschung unter unterschiedlichen Ausgangsbedingungen. Das bedeutet zum einen die besondere Berücksichtigung relativ kleiner Stichproben, zum anderen eine starke Orientierung an der Art der in der Praxis zur Verfügung stehenden Informationen.

Solche Informationen liegen in den zur Evaluation eingesetzten Datensätzen in Form unterschiedlich skalierter Variablen vor. Das macht die Anwendung von Distanz- oder Ähnlichkeitsindikatoren notwendig, mit deren Hilfe die Berücksichtigung verschieden skalierter Merkmale ohne Informationsverlust möglich ist. Neben den in empirischen Studien häufig angewendeten Balancing Scores werden im zweiten Kapitel dieser Arbeit aggregierte Distanzmaße aus anderen Wissenschaftsbereichen vorgestellt. Es wird erwartet, dass diese Maße in kleinen Stichproben besser in der Lage sind, die Informationen über Ähnlichkeiten und Unterschiede der betrachteten Personen zusammenzufassen als die bisher überwiegend verwendeten Scores.

Für die Zuordnung von Partnern auf Grundlage der ermittelten Distanzen bzw. Ähnlichkeiten werden in der Literatur sehr unterschiedliche Verfahren diskutiert und angewendet. Dabei ist kein Verfahren den anderen generell überlegen. Die Algorithmen lassen sich nach verschiedenen Aspekten unterscheiden. So ist die Anzahl der einer Person zugeordneten Partner ein Kriterium, nach dem Nearest Neighbor Matching von Zuordnungen einer festen oder variablen Anzahl von Personen – incl. der vollständigen Zuordnung der zur Verfügung stehenden potenziellen Partner – zu einem Teilnehmer unterschieden wird. Ein anderes Kriterium ist die Möglichkeit der Mehrfachzuordnung einer Person, nach dem Verfahren mit dieser Möglichkeit (Zuordnung mit Zurücklegen) von solchen ohne Mehrfachnutzung (Zuordnung ohne Zurücklegen) zu trennen sind. Zur erstgenannten Gruppe gehören auch die Verfahren der Local Polynomial Regression.
Eine besondere Bedeutung kommt den optimalen Zuordnungsprozessen zu, da mit ihnen die bestmögliche Zuordnung (hinsichtlich eines vorher festgelegten Kriteriums) erreicht werden kann. Die Vorstellung solcher – überwiegend aus der linearen Optimierung bzw. der Graphentheorie bekannten – Verfahren bildet einen weiteren Schwerpunkt des zweiten Kapitels dieser Arbeit.

Das dritte Kapitel gibt einen Überblick über den Stand der Forschung zur Entwicklung von „Standards“ bei der Wahl geeigneter Matchingverfahren in verschiedenen Situationen. In diesen Studien wird festgestellt, dass Matchingverfahren besser als andere nichtparametrische und parametrische Verfahren in der Lage sind, das Selektionsproblem zu lösen, wenn umfangreiche Informationen über die betrachteten Personen zur Verfügung stehen. Die Wahl eines geeigneten Algorithmus ist dabei abhängig von den verfügbaren Daten.
Unter den diskutierten Distanzmaßen werden der Propensity Score und der Index Score sowie die Mahalanobisdistanz für die empirische Forschung empfohlen. In Bezug auf Zuordnungsprozesse werden Optimal Full Matching, Ridge Matching und die Zuordnung mit Zurücklegen als vorteilhaft gegenüber anderen Verfahren angesehen.

Im vierten Kapitel wird eine Simulationsstudie vorgestellt, in der die empfohlenen Distanzmaße und Zuordungsprozesse miteinander verglichen werden. Zusätzlich zu den in der Literatur favorisierten Distanzmaßen werden zwei der vorgestellten aggregierten Distanzmaße, die Mahalanobis-Matching-Distanz und das Ähnlichkeitsmaß von Gower, in die Analyse einbezogen. Neben den in früheren Studien hervorgehobenen Zuordnungsalgorithmen und einem in der empirischen Literatur weit verbreiteten Verfahren, dem Random Matching, werden zwei Algorithmen aus der Gruppe der optimalen Zuordnungsprozesse betrachtet: der Ungarische Algorithmus für optimale 1:1-Zuordungen sowie ein Auktionsalgorithmus für Optimal Full Matching.

Als Datenbasis der Simulation dient eine Nachbildung des Mikrozensus Deutschland. Mit dieser engen Orientierung an einem häufig in der Arbeitsmarktforschung eingesetzten Datensatz wird eine realitätsnahe Verteilung der unterschiedlich skalierten Merkmale in den untersuchten Stichproben erreicht.
In jedem Schritt der Untersuchung werden verschiedene Teilnehmer- und Nichteilnehmerstichproben miteinander kombiniert, die sich in ihrer Größe insgesamt, dem Zahlenverhältnis von Teilnehmern und Nichtteilnehmern sowie dem Grad der Übereinstimmung der Merkmalsverteilungen in beiden Gruppen unterscheiden. In jedem Schritt werden jeweils 100 Simulationsläufe durchgeführt.
Die Ergebnisse werden anhand unterschiedlicher Gütemaße beurteilt. Zur Prüfung der Distanzmaße werden neben der Bias Reduzierung durch Matching nichtpara-

metrische skalenspezifische Tests der Übereinstimmung der Mittelwerte bzw. Häufigkeitsverteilungen der einzelnen betrachteten Variablen eingesetzt: für metrisch skalierte Merkmale der Vorzeichen-Rangtest von Wilcoxon, für dichotome der McNemartest und für polytome Variablen der χ^2-Homogenitätstest. Diese Tests stellen eine sinnvolle Alternative zu den bisher in der Literatur gebräuchlichen Verfahren der Gütemessung dar. Die Beurteilung der Zuodnunsprozesse erfolgt anhand des mittleren quadratischen Fehlers, des Bias und der empirischen Varianz sowie der Summe der quadrierten Distanzen zwischen Teilnehmern und Nichtteilnehmern.

Im ersten Teil der Analyse wird festgestellt, dass die Zusammenfassung unterschiedlich skalierter Merkmale mit den untersuchten Balancing Scores (Index Score und Propensity Score) deutlich schlechter gelingt als mit der Mahalanobisdistanz und den aggregierten Distanzmaßen. Die gewichtete Mahalanobis-Matching-Distanz scheint am besten zur Feststellung von Ähnlichkeiten bzw. Unterschieden der betrachteten Personen geeignet zu sein. Allerdings schwankt die Güte der erzielten Ergebnisse mit dem Skalenniveau der Variablen. Während die Angleichung der Verteilung nominaler (dichotomer und vor allem polytomer) Variablen sehr gut gelingt, treten nach dem Matching relativ häufig noch Unterschiede in der Verteilung der metrischen Variablen auf. Das Gegenteil gilt für das Distanzmaß nach Gower. In einer weiterführenden Analyse wäre zu prüfen, ob mit einer Verbindung beider Distanzmaße die Kombination ihrer jeweiligen Vorteile möglich ist. Dazu müsste der verallgemeinerte Matchingkoeffizient für nominale Variablen mit der normierten absoluten Merkmalsdifferenz metrischer Variablen verknüpft werden.

Im zweiten Teil der Analyse ergeben die verschiedenen betrachteten Gütemaße ein sehr heterogenes Bild. Kein Zuordnungsalgorithmus liefert in allen Qualitätskriterien gleichermaßen gute oder schlechte Ergebnisse. Die Rangfolge, die sich beim Vergleich der Prozesse ergibt, ist abhängig vom betrachteten Gütemaß.
Wird der mittlere quadratische Fehler betrachtet, gelingt mit Optimal Full Matching die Zuordnung der besten Partner in Stichproben mit unterschiedlich großen Teilnehmer- und Nichtteilnehmerzahlen. Dies ist umso deutlicher, je größer die Nichtteilnehmerstichprobe im Vergleich zur Teilnehmerstichprobe ist. Verwendet man dagegen die Summe der quadrierten Distanzen als Gütekriterium zur Beurteilung der Ähnlichkeit der Merkmalsverteilungen in Teilnehmer- und Kontrollgruppe,

liefert die Zuordnung mit Zurücklegen die besten Ergebnisse unter den betrachteten Zuordnungsprozessen.
Für die beiden 1:1-Zuordnungsprozesse ohne Mehrfachzuordnung werden sehr ähnliche Ergebnisse – hinsichtlich aller Gütemaße – beobachtet. Es lässt sich kein Vorteil des optimalen Nearest Neighbor Matching gegenüber dem Random Matching nachweisen.
Das Ridge Matching wird hinsichtlich aller Kriterien schlechter bewertet als die anderen analysierten Zuordnungsprozesse.

Mit der empirischen Untersuchung des fünften Kapitels soll die Frage, ob die Absolventen geförderter Berufsausbildungen in den Neuen Bundesländern beim Berufseinstieg gegenüber Absolventen ungeförderter Ausbildungsgänge benachteiligt sind, beantwortet werden. Hinsichtlich der Förderung wird zwischen außerbetrieblicher und betriebsnaher Ausbildung unterschieden. Die Analyse wird auf Basis des Jugendpanels des Zentrums für Sozialforschung Halle durchgeführt, aus dem Informationen über die Jugendlichen, die eine Berufsausbildung erfolgreich abgeschlossen haben, genutzt werden.

Aus der deskriptiven Analyse dieser Stichprobe wird deutlich, dass die Berufseinstiegschancen der geförderten Jugendlichen schlechter sind als die der ungefördert Ausgebildeten. Ebenfalls deutlich wird eine ungleiche Verteilung der Merkmale in den Teilstichproben, woraus sich die ungleichen Chancen auf dem Arbeitsmarkt zum Teil erklären lassen. Ob darüber hinaus der Umstand der Förderung selbst einen Einfluss auf die Beschäftigungschancen der Jugendlichen hat, wird in der Analyse mit Hilfe der Zuordnung mit Zurücklegen ermittelt.

Sowohl für die außerbetriebliche als auch die betriebsnahe Berufsausbildung wird ein negativer Effekt der Förderung auf die Berufseinstiegschancen der Jugendlichen festgestellt. Dies trifft sowohl auf den Anteil der Jugendlichen, die eine Beschäftigung aufnehmen, als auch auf qualitative Merkmale der aufgenommenen Berufstätigkeit zu. Der Vergleich beider Arten der Förderung ergibt keinen Hinweis darauf, dass außerbetrieblich geförderte Jugendliche schlechtere Berufseinstiegschancen haben als die Absolventen betriebsnaher Ausbildungen.

Literaturverzeichnis

Abadie, A.; Imbens, G. W. (2002): *Simple and Bias-Corrected Matching Estimators for Average Treatment Effects*. National Bureau of Economic Research, Cambridge: NBER Technical Working Paper T286.

Abbring, J. H.; van den Berg, G. J. (2003): The nonparametric identification of treatment effects in duration models. *Econometrica*, 71, Nr. 5, S. 1491–1517.

Angrist, J. D.; Hahn, J. (2004): When to control for covariates? Panel asymptotics for estimates of treatment effects. *The Review of Economics and Statistics*, 86, Nr. 1, S. 58–72.

Arntz, M.; Jacobebbinghaus, P.; Spermann, A. (2004): Minijobs, Midijobs und sozialversicherungspflichtige Beschäftigung in privaten Haushalten. In Hagen, T.; Spermann, A. (Hrsg.): *Hartz-Gesetze. Methodische Ansätze zu einer Evaluierung*. Baden-Baden: Nomos Verl.-Ges., *ZEW-Witrschaftsanalysen* 74, S. 171–189.

Augurzky, B. (2000a): *Evaluation Strategies in Labor Economics - An Application to Post-secondary Education*. Dissertation, Ruprecht-Karls-Universität, Heidelberg.

Augurzky, B. (2000b): *Matching the Extremes. A sensitivity Analysis Based on Real Data*. Ruprecht-Karls-Universität, Heidelberg: Discussion Paper No. 310.

Augurzky, B.; Kluve, J. (2004): *Assessing the Performance of Matching Algorithms When Selection Is Strong*. Forschungsinstitut zur Zukunft der Arbeit (IZA), Bonn: IZA Discussion Paper No. 1301.

Backhaus, K.; Erichsen, B.; Plinke, W.; Weiber, R. (2000): *Multivariate Analysemethoden*. 9. Auflage. Berlin: Springer-Verlag.

Bazaraa, M. S.; Jarvis, J. J.; Sherali, H. D. (1990): *Linear programming and network flows*. 2. Auflage. New York: Wiley.

Becker, S. O.; Ichino, A. (2002): Estimation of average treatment effects based on propensity scores. *The Stata Journal*, 2, Nr. 4, S. 358–377.

Bergemann, A.; Fitzenberger, B.; Speckesser, S. (2001): *Evaluating the Employment Effects of Public Sector Sponsored Training in East Germany: Conditional Difference-in-Differences and Ashenfelter's Dip*. Universität Mannheim; mimeo.

Bergemann, A.; Fitzenberger, B.; Speckesser, S. (2004): *Evaluating the Dynamic Employment Effects of Training Programs in East-Germany Using Conditional Difference-in-Difference*. Zentrum für Europäische Wirtschaftsforschung (ZEW), Mannheim: ZEW Discussion Paper No. 04-41.

Berger, K. (2006): *Evaluierung der Bund-Länder-Ausbildungsplatzprogramme Ost – Erwerbssituation der Programmabsolventinnen und Absolventen ein halbes Jahr nach Ausbildungsabschluss*. Bundesinstitut für Berufsbildung (BIBB), Bonn: http://www.bibb.de/de/wlk8305.htm, Juni 2008.

Berger, K.; Braun, U.; Drinkhut, V.; Schöngen, K. (2007): *Wirksamkeit staatlich finanzierter Ausbildung: Ausbildungsplatzprogramm Ost – Evaluation, Ergebnisse und Empfehlungen*. Bertelsmann Stiftung, Bielefeld: Schriftenreihe des Bundesinstituts für Berufsbildung.

Berger, K.; Walden, G. (2003): *Öffentliche Ausbildungsförderung in Ostdeutschland unter der Lupe: Ergebnisse aktueller Evaluationsstudien*. Bertelsmann Stiftung, Bielefeld: Berichte zur beruflichen Bildung Nr. 258.

Bernhard, S.; Hohmeyer, K.; Jozwiak, E.; Koch, S.; Kruppe, T.; Stephan, G.; Wolff, J. (2008): *Aktive Arbeitsmarktpolitik in Deutschland und ihre Wirkungen*. Institut für Arbeitsmarkt- und Berufsforschung (IAB), Nürnberg: IAB-Forschungsbericht 02/2008.

Bertsekas, D. P. (1981): A new algorithm for the assignment problem. *Journal of Mathematical Programming*, 21, Nr. 1, S. 152–171.

Bertsekas, D. P. (1992a): Auction Algorithms for Network Flow Problems: A Tutorial Introduction. *Computational Optimization and Applications*, 1, Nr. 1, S. 7–66.

Bertsekas, D. P. (1992b): *Linear Network Optimization. Algorithms and Codes*. 2. Auflage. Cambridge: MIT Press.

Bertsekas, D. P. (2001): Auction Algorithms. In Floudas, C. A.; Pardalos, P. M. (Hrsg.): *Encyclopedia of optimization*. Band I, Dordrecht: Kluwer Academic Publishers.

Bertsekas, D. P.; Castanon, D. A.; Tsaknakis, H. (1993): Reverse Auction and the solution of Inequality Constrained Assignment Problems. *SIAM Journal on Optimization*, 3, Nr. 2, S. 268–299.

Bikhchandani, S.; Ostroy, J. M. (2006): From the Assignment Model to Combinatorial Auctions. In Cramton, P.; Shoham, Y.; Steinberg, R. (Hrsg.): *Combinatorial Auctions*. 1. Auflage. Massachusetts: MIT Press. – Kapitel 8, S. 189–214.

Black, D. A.; Smith, J. A. (2004): How robust is the evidence on the effects of college quality? Evidence from matching. *Journal of Econometrics*, 121, Nr. 1-2, S. 99–124.

Bleymüller, J.; Gehlert, G.; Gülicher, H. (1996): *Statistik für Wirtschaftswissenschaftler*. 10. Auflage. München: Verlag Franz Vahlen.

Blundell, R.; Costa Dias, M. (2000): Evaluation Methods for Non-Experimental Data. *Fiscal Studies*, 21, Nr. 4, S. 427–468.

Büning, H.; Trenkler, G. (1994): *Nichtparametrische statistische Methoden*. 2. Auflage. Berlin, New York: Verlag de Gruyter.

Borlin, N. (1999): *MATLAB function hungarian*. http://www.mathworks.com/matlabcentral/fileexchange/loadFile.do?objectId=94&objectType=file, Mai 2004.

Brosius, F. (1998): *SPSS 8: professionelle Statistik unter Windows*. 1. Auflage. Bonn: MIT Press.

Bundesagentur für Arbeit (2008): *Statistik der Bundesagentur für Arbeit*. http://www.pub.arbeitsamt.de/hst/services/statistik/000000/html/start/schaubilder.shtml, Juni 2008.

Bundesministerium für Bildung und Forschung (2006): *Die Berufsausbildung stärken - Ausbildungschancen für jeden jungen Menschen. Berufsbildungsbericht 2006*. Berlin: http://www.bmbf.de/pub/bbb_2006.pdf, Juni 2008.

Cain, A. J.; Harrison, G. A. (1958): An analysis of the taxonomists judgement of affinity. *Proceedings of the Zoological Society*, 131, S. 85–98.

Caliendo, M.; Hujer, R. (2006): The Microeconometric Estimation of Treatment Effects – An Overview. *Allgemeines Statistisches Archiv*, 90, Nr. 1, S. 199–215.

Caliendo, M.; Kopeinig, S. (2005): *Some Practical Guidance for the Implementation of Propensity Score Matching*. Deutsches Institut für Wirtschaftsforschung (DIW), Berlin: DIW Discussion Paper No. 485.

Calmfors, L.; Forslund, A.; Hemström, M. (2002): *Does Active Labour Market Policy Work? Lessons from the Swedish Experiences*. Institute for International Economic Studies, Stockholm University, Stockholm: Seminar Paper No. 700.

Cheetham, A. H.; Hazel, J. E. (1969): Binary Presence-Absence Similarity Coefficients. *Journal of Paleontology*, 43, Nr. 5, S. 1130–1136.

Christensen, B. (2001): *Berufliche Weiterbildung und Arbeitsplatzrisiko: Ein Matching-Ansatz*. Institut für Weltwirtschaft (IfW), Kiel: Kieler Arbeitspapier Nr. 1033.

Clauss, G.; Ebner, H. (1967): *Grundlagen der Statistik für Psychologen, Pädagogen und Soziologen*. Berlin: Verlag Volk und Wissen.

Cochran, W. G. (1968): The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies. *Biometrics*, 24, Nr. 2, S. 295–313.

Cochran, W. G.; Rubin, D. B. (1973): Controlling Bias in Observational Studies: A Review. *Sakhya: The Indian Journal of Statistics, Ser. A*, 35, Nr. 4, S. 417–446.

Dehejia, R. H.; Wahba, S. (1999): Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs. *Journal of the American Statistical Association*, 94, Nr. 448, S. 1053–1062.

Dehejia, R. H.; Wahba, S. (2002): Propensity Score-Matching Methods for Nonexperimental Causal Studies. *The Review of Economics and Statistics*, 84, Nr. 1, S. 151–161.

Diday, E.; Simon, J. (1976): Clustering Analysis. In Fu, K. S. (Hrsg.): *Digital Pattern Recognition*. Berlin: Springer-Verlag. – Kapitel 3, S. 47–94.

Egervary, E. (1931): On combinatorial properties of matrices. *Matematikai Lapok*, 38, S. 16–28.

Eichler, M.; Lechner, M. (2001): Public Sector Sponsored Continuous Vocational Training in East Germany: Institutional Arrangements, Participants and Re-

sults of Empirical Evaluations. In Riphahn, R. T.; Snower, D. J.; Zimmermann, K. F. (Hrsg.): *Employment Policy in Transition: The Lessons from German Integration for the Labour Market*. Berlin: Springer-Verlag, S. 208–253.

Fan, J. (1992): Design-adaptive Nonparametric Regression. *Journal of the American Statistical Association*, 87, Nr. 420, S. 998–1004.

Fan, J.; Hall, P.; Martin, M.; Patil, P. (1996): On Local Smoothing of Nonparametric Curve Estimators. *Journal of the American Statistical Association*, 91, Nr. 1, S. 258–266.

Fitzenberger, B.; Prey, H. (2000): Evaluating public sector sponsored training in East Germany. *Oxford Economic Papers*, 52, Nr. 3, S. 497–520.

Fitzenberger, B.; Speckesser, S. (2002): *Weiterbildungsmaßnahmen in Ostdeutschland. Ein Misserfolg der Arbeitsmarktpolitik?* Zentrum für Europäische Wirtschaftsforschung (ZEW), Mannheim: ZEW Discussion Paper No. 02-16.

Fredriksson, P.; Johansson, P. (2003): *Program evaluation and random program starts*. Insititute for Labour Market Policy Evaluation (IFAU), Uppsala: IFAU Working Paper No. 2003:1.

Fröhlich, M. (2004a): Finite Sample Properties of Propensity-Score Matching and Weighting Estimators. *The Review of Economics and Statistics*, 86, Nr. 1, S. 77–90.

Fröhlich, M. (2004b): Programme Evaluation with Multiple Treatments. *Journal of Economic Surveys*, 18, Nr. 2, S. 181–224.

Gerfin, M.; Lechner, M. (2002): A Microeconomic Evaluation of the Active Labour Market Policy in Switzerland. *The Economic Journal*, 112, Nr. 482, S. 854–893.

Gowda, C. K.; Diday, E. (1992): Symbolic Clustering Using a New Similarity Measure. *IEEE Transactions on Systems, Man, and Cybernetics*, 22, Nr. 2, S. 368–378.

Gower, J. C. (1971): A General Coefficient of Similarity and Some of Its Properties. *Biometrics*, 27, Nr. 4, S. 857–871.

Greene, W. H. (2003): *Econometric Analysis*. 5. Auflage. Upper Saddle River: Prentice Hall.

Grünert, H.; Wiekert, I. (2005): Ostdeutschland als Labor zur Weiterentwicklung des dualen Systems der Berufsausbildung? In Jacob, M.; Kupka, P. (Hrsg.): *Perspektiven des Berufskonzepts: die Bedeutung des Berufs für Ausbildung und Arbeitsmarkt*. Nürnberg: Institut für Arbeitsmarkt- und Berufsforschung (IAB), *Beiträge zur Arbeitsmarkt- und Berufsforschung* 297, S. 123–142.

Gu, X. S.; Rosenbaum, P. R. (1993): Comparison of Multivariate Matching Methods: Structures, Distances, and Algorithms. *Journal of Computational and Graphical Statistics*, 2, Nr. 4, S. 405–420.

Hahn, J. (1998): On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects. *Econometrica*, 66, Nr. 2, S. 315–331.

Hall, P.; Turlach, B. A. (1997): Interpolation Methods for Adapting to Sparse Design in Nonparametric Regression. *Journal of the American Statistical Association*, 92, Nr. 438, S. 466–476.

Hansen, B. B.; Ohlsen-Klopfer, S. (2006): Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics*, 15, Nr. 3, S. 609–627.

Heckman, J. J. (1978): Dummy Endogenous Variables in a Simultaneous Equation System. *Econometrica*, 46, Nr. 6, S. 931–959.

Heckman, J. J. (1979): Sample Selection Bias as a Specification Error. *Econometrica*, 47, Nr. 1, S. 153–161.

Heckman, J. J. (1980): Addendum to Sample Selection Bias as a Specification Error. In Stromsdorfer, E. W.; Farkas, G. (Hrsg.): *Evaluation Studies Review Annual*. Band 5, Beverly Hills: Sage Publications, S. 69–74.

Heckman, J. J. (1997): Instrumental Variables. A Study of Implicit Behavioral Assumptions Used in Making Program Evaluation. *Journal of Human Resources*, 32, Nr. 3, S. 441–462.

Heckman, J. J.; Hotz, J. V. (1989): Choosing among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training. *Journal of the American Statistical Association*, 84, Nr. 408, S. 862–880.

Heckman, J. J.; Ichimura, H.; Smith, J. A.; Todd, P. E. (1998): Characterizing Selection Bias Using Experimental Data. *Econometrica*, 66, Nr. 5, S. 1017–1098.

Heckman, J. J.; Ichimura, H.; Todd, P. E. (1997): Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme. *Review of Economic Studies*, 64, Nr. 4, S. 605–654.

Heckman, J. J.; Ichimura, H.; Todd, P. E. (1998): Matching As An Econometric Evaluation Estimator. *Review of Economic Studies*, 65, Nr. 2, S. 261–294.

Heckman, J. J.; LaLonde, R. J.; Smith, J. A. (1999): The Economics and Econometrics of Active Labor Market Programs. In Ashenfelter, O.; Card, D. E. (Hrsg.): *Handbook of Labor Economics*. Band III, Amsterdam: Elsevier Science B.V., S. 1865–2097.

Heckman, J. J.; Navarro-Lozano, S. (2004): Using Matching, Instrumental Variables, and Control Functions to Estimate Economic Choice Models. *The Review of Economics and Statistics*, 86, Nr. 1, S. 30–57.

Heckman, J. J.; Robb, R. (1985): Alternative Methods for Evaluating the Impact of Interventions. In Heckman, J. J.; Singer, B. (Hrsg.): *Longitudinal Analysis of Labor Market Data*. Cambridge: Cambridge University Press, S. 156–245.

Heckman, J. J.; Smith, J. A. (1995): Assessing the Case for Social Experiments. *Journal of Economic Perspectives*, 9, Nr. 2, S. 85–110.

Heckman, J. J.; Smith, J. A. (1999): The Pre-Program Earnings Dip and the Determinants of Participation in a Social Program: Implications for Simple Program Evaluation Strategies. *Economic Journal*, 109, S. 313–348.

Heckman, J. J.; Vytlacil, E. J. (1999): Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects. *Proceedings of the National Academy of Sciences*, 96, Nr. 8, S. 4730–4734.

Hübler, O. (2001): Evaluation of policy interventions: Measurement and problems. *Allgemeines Statistisches Archiv*, 85, S. 103–126.

Hujer, R.; Caliendo, M. (2000): *Evaluation of Active Labour Market Policy: Methodological Concepts and Empirical Estimates*. Forschungsinstitut zur Zukunft der Arbeit (IZA), Bonn: IZA Discussion Paper No. 236.

Hujer, R.; Caliendo, M.; Radić, D. (2001): *Nobody Knows... How Different Evaluation Estimators Perform in a Simulated Labour Market Experiment*. Johann-Wolfgang-Goethe-Universität, Frankfurt (Main): http://www.wiwi.uni-frankfurt.de/Professoren/hujer/papers/HCR_Nobody_knows.pdf, November 2003.

Hujer, R.; Caliendo, M.; Thomsen, S. (2003): *New Evidence on the Effects of Job Creation Schemes in Germany - A Matching Approach with Threefold Heterogeneity*. Forschungsinstitut zur Zukunft der Arbeit (IZA), Bonn: IZA Discussion Paper No. 750.

Hujer, R.; Maurer, K.-O.; Wellner, M. (1997): *The Impact of Training on Unemployment Duration in West Germany*. Johann-Wolfgang-Goethe-Universität, Frankfurt (Main): Discussion Papers in Economics No. 74.

Hujer, R.; Thomsen, S. L. (2006): *How Do Employment Effects of Job Creation Schemes Differ with Respect to the Foregoing Unemployment Duration?* Zentrum für Europäische Wirtschaftsforschung (ZEW), Mannheim: ZEW Discussion Paper No. 06-47.

Hyvärinen, L. (1962): Classification of qualitative data. *BIT Numerical Mathematics*, 2, Nr. 2, S. 83–89.

Ichino, M.; Yaguchi, H. (1994): Generalized Minkowski Metrics for Mixed Feature-Type Data Analysis. *IEEE Transactions on Systems, Man, and Cybernetics*, 24, Nr. 4, S. 698–708.

Imbens, G. W. (2004): Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review. *The Review of Economics and Statistics*, 86, Nr. 1, S. 4–29.

Kaltenborn, B. (2002): *Hartz-Umsetzung und deren Evaluierung*. http://www.wipol.de/hartz/evaluierung.htm, Juli 2008.

Kaufmann, H.; Pape, H. (1996): Clusteranalyse. In Fahrmeir, L.; Hamerle, A.; Tutz, G. (Hrsg.): *Multivariate statistische Verfahren*. 2. Auflage. Berlin: Verlag de Gruyter, S. 437–536.

König, D. (1916): Über Graphen und ihre Anwendung auf Determinantentheorie und Mengenlehre. *Mathematische Annalen*, 77, Nr. 4, S. 453–465.

Konle-Seidl, R. (2005): *Lessons learned. Internationale Evaluierungsergebnisse zu Wirkungen aktiver und aktivierender Arbeitsmarktpolitik*. Institut für Arbeitsmarkt- und Berufsforschung (IAB): IAB-Forschungsbericht Nr. 9.

Kuhn, H. W. (1955): The hungarian method for solving the assignment problem. *Naval Research Logistics Quarterly*, 2, S. 83–97.

Kumar, A. (2007): *MATLAB function auction_match*. http://www.mathworks.com/matlabcentral/files/14251/auction_match.m, August 2007.

Lalive, R.; van Ours, J. C.; Zweimüller, J. (2002): *The Impact of Active Labor Market Programs on the Duration of Unemployment*. Institute for Empirical Research in Economics (IEW), Zürich: IEW Discussion Paper No. 41.

LaLonde, R. J. (1986): Evaluating the Econometric Evaluations of Training Programs with Experimental Data. *American Economic Review*, 76, Nr. 4, S. 604–620.

Larsson, L. (2003): Evaluation of Swedish Youth Labour Market Programmes. *Journal of Human Resources*, 38, Nr. 4, S. 891–927.

Le Sage, J. P. (1999): *Applied Econometrics using MATLAB*. http://www.spatial-econometrics.com/html/mbook.pdf, Mai 2004.

Lechner, M. (1998): *Training the East German Labour Force. Microeconometric Evaluations of Continuous Vocational Training after Unification*. Heidelberg: Physica-Verlag.

Lechner, M. (1999): Earnings and Employment Effects of Continuous Off-the-Job Training in East Germany after Unification. *Journal of Business & Economic Statistics*, 17, Nr. 1, S. 74–90.

Lechner, M. (2001a): *A note on the Commom Support Problem in applied evaluation studies*. Department of Economics, University of St. Gallen: Discussion Paper No. 2001-01.

Lechner, M. (2001b): Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In Lechner, M.; Pfeiffer, F. (Hrsg.): *Econometric Evaluation of Labour Market Policies*. Heidelberg: Physica/Springer-Verlag, *ZEW Economic Studies* 13, S. 43–58.

Lechner, M. (2004): *Sequential Matching Estimation of Dynamic Causal Models*. Forschungsinstitut zur Zukunft der Arbeit (IZA): IZA Discussion Paper No. 1042.

Lechner, M.; Miquel, R. (2001): *Identification of Effects of Dynamic Treatments by Sequential Conditional Independence Assumptions*. Department of Economics, University of St. Gallen: Discussion Paper No. 2001-07.

Lechner, M.; Miquel, R.; Wunsch, C. (2004): *Long-Run Effects of Public Sector Sponsored Training in West Germany*. Department of Economics, University of St. Gallen: Discussion Paper No. 2004-19.

Mahalanobis, P. C. (1936): On the generalized distance in statistics, for the classification problem. *Proceedings of the National Institute of Science India*, II, Nr. 1, S. 49–55.

Ming, K.; Rosenbaum, P. R. (2000): Substantial Gains in Bias Reduction from Matching with a Variable Number of Controls. *Biometrics*, 56, Nr. 1, S. 118–124.

Opitz, O. (1980): *Numerische Taxonomie*. Stuttgart, New York: Fischer-Verlag.

Prein, G. (2005): Die Maßnahme und die Folgen: Über die Konsequenzen der öffentliche Förderung der Berufsausbildung in Ostdeutschland für die Einmündung

in das Erwerbssystem. In Wiekert, I. (Hrsg.): *Zehn aus Achtzig. Burkart Lutz zum 80.* Berlin: Wissenschaftsverlag, *Berliner Debatte*, S. 191–207.

Qian, Y. (2004): *Do Additional National Patent Laws Stimulate Domestic Innovation In A Global Patenting Environment? A Cross-Country Analysis of Pharmaceutical Patent Protection: 1978-1999.* Dissertation, Harvard University, unpublished.

Reinowski, E. (2006): Mikroökonometrische Evaluation und das Selektionsproblem. Ein anwendungsorientierter Überblick über nichtparametrische Lösungsverfahren. *Zeitschrift für Evaluation*, 5, Nr. 2, S. 187–226.

Reinowski, E.; Schultz, B.; Wiemers, J. (2003): *Evaluation von Maßnahmen der aktiven Arbeitsmarktpolitik mit Hilfe eines iterativen Matching-Algorithmus. Eine Fallstudie über langzeitarbeitslose Maßnahmeteilnehmer in Sachsen.* Institut für Wirtschaftsforschung Halle (IWH), Halle: IWH-Diskussionspapier Nr. 173.

Reinowski, E.; Schultz, B.; Wiemers, J. (2005): Evaluation of Further Training Programmes with an Optimal Matching Algorithm. *Swiss Journal of Economics and Statistics*, 141, Nr. 4, S. 585–616.

Richardson, K.; van den Berg, G. J. (2001): The effect of vocational employment training on the individual transition rate from unemployment to work. *Swedish Economic Policy Review*, 8, Nr. 2, S. 175–213.

Rosenbaum, P. R. (1987): The Role of a Second Control Group in an Observational Study. *Statistical Science*, 2, Nr. 3, S. 292–316.

Rosenbaum, P. R. (1989): Optimal Matching for Observational Studies. *Journal of the American Statistical Association*, 84, Nr. 408, S. 1024–1032.

Rosenbaum, P. R. (2002): *Observational Studies.* 2. Auflage. New York: Springer-Verlag.

Rosenbaum, P. R.; Rubin, D. B. (1983): The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 70, Nr. 1, S. 41–55.

Rosenbaum, P. R.; Rubin, D. B. (1985): Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score. *The American Statistician*, 39, Nr. 1, S. 33–39.

Rubin, D. B. (1973): Matching to Remove Bias in Observational Studies. *Biometrics*, 29, Nr. 1, S. 159–183.

Rubin, D. B. (1986): Statistics and Causal Inference: Comment: Which Ifs have Causal Answers. *Journal of the American Statistical Association*, 81, Nr. 396, S. 961–962.

Ruppert, D. (1997): Empirical-Bias Bandwidths for Local Weighted Least Squares Regression. *Journal of the American Statistical Association*, 92, Nr. 439, S. 1049–1062.

Sachverständigenrat zur Begutachtung der gesamtwirtschaftlichen Entwicklung (2004): *Erfolge im Ausland - Herausforderungen im Inland. Jahresgutachten 2004/05.* Berlin: Verlag H. Heenemann GmbH & Co..

Schmidt, C. M. (1999): *Knowing what Works - the Case for Rigorous Program Evaluation*. Forschungsinstitut zur Zukunft der Arbeit (IZA), Bonn: IZA Discussion Paper No. 77.

Seifert, B.; Gasser, T. (1996): Finite-Sample Variance of Local Polynomials: Analysis and Solutions. *Journal of the American Statistical Association*, 91, Nr. 433, S. 267–275.

Seifert, B.; Gasser, T. (2000): Data Adaptive Ridging in Local Polynomial Regression. *Journal of Computational and Graphical Statistics*, 9, Nr. 2, S. 338–360.

Sianesi, B. (2001): *Diffential Effects of Swedish Active Labour Market Programmes for Unemployed Adults during the 1990s – revised version –*. The Institute for Fiscal Studies (IFS), London: IFS Working Paper No. 01/25.

Sianesi, B. (2002): *An Evaluation of the Swedish System of Active Labour Market Programmes in the 1990s*. The Institute for Fiscal Studies (IFS), London: IFS Working Paper No. 02/01.

Sianesi, B. (2004): An Evaluation of the Swedish system of Active Labor Market Programs. *The Review of Economics and Statistics*, 86, Nr. 1, S. 133–155.

Siegel, S. (1997): *Nichtparametrische statistische Methoden*. Band 4, Eschborn: Verlag Dietmar Klotz GmbH.

Smith, J. A.; Todd, P. E. (2005a): Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics*, 125, Nr. 1-2, S. 305–353.

Smith, J. A.; Todd, P. E. (2005b): Does matching overcome LaLonde's critique of nonexperimental estimators? Rejoinder. *Journal of Econometrics*, 125, Nr. 1-2, S. 365–75.

Sokal, R. R.; Sneath, P. H. (1963): *Principles of Numerical Taxonomy*. San Francisco & London: W. H. Freeman and Company.

Solga, H. (2005): *Ohne Abschluss in die Bildungsgesellschaft. Die Erwerbschancen gering qualifizierter Personen aus soziologischer und ökonomischer Perspektive*. Opladen: Verlag Barbara Budrich.

Statistisches Bundesamt (1992): *Personensystematik. Klassifizierung der Berufe*. Stuttgart: Verlag Metzler-Poeschel.

Steiner, C.; Böttcher, S.; Prein, G.; Terpe, S. (2004): *Land unter – Ostdeutsche Jugendliche auf dem Weg ins Beschäftigungssystem*. Zentrum für Sozialforschung Halle (zsh), Halle: Forschungsberichte aus dem zsh Nr. 04-1.

Steinhausen, D.; Langer, K. (1977): *Clusteranalyse: Einführung in Methoden und Verfahren der automatischen Klassifikation*. Berlin: Verlag de Gruyter.

Tillmann, F. (2004): Codierung offener Berufsangaben mit der KldB1992 – Ein Leitfaden zur computergestützten Vercodung mit SPSS. *RBS-Mitteilungen*, 1, S. 79–91.

von Below, S. (1999): Bildungschancen von Jugendlichen in Ost- und Westdeutschland. In Lüttinger, P. (Hrsg.): *Sozialstrukturanalysen mit dem Mikrozensus*. Band 6, Mannheim: ZUMA, S. 271–299.

Wilson, D. R.; Martinez, T. R. (1997): Improved Heterogeneous Distance Functions. *Journal of Artificial Intelligence Research*, 6, S. 1–34.

Woessmann, L. (2004): *How Equal Are Educational Opportunities? Family Background and Student Achivement in Europe and the United States*. Forschungsinstitut zur Zukunft der Arbeit (IZA), Bonn: IZA Discussion Paper No. 1284.

Zentrum für Sozialforschung Halle (zsh) (2003): *ostmobil*. http://www.ostmobil.de, Juni 2008.

Zhao, Z. (2004): Using Matching to Estimate Treatment Effects: Data Requirements, Matching Metrics, and Monte Carlo Evidence. *The Review of Economics and Statistics*, 86, Nr. 1, S. 91–107.

Zhao, Z. (2006): *Matching Estimators and the Data from the National Supported Work Demonstration Again*. Forschungsinstitut zur Zukunft der Arbeit (IZA), Bonn: IZA Discussion Paper No. 2375.

Zijl, M.; van den Berg, G. J.; Heyma, A. (2004): *Stepping Stones for the Unemployed: The Effect of Temporary Jobs on the Duration until Regular Work*. Forschungsinstitut zur Zukunft der Arbeit (IZA), Bonn: IZA Discussion Paper No. 1241.

Anhang A

Symbolverzeichnis

AI	...	Distanzmaß von Abadie/Imbens
B	...	Bias
BS	...	Balancing Score
C	...	Kontrollgruppe
$C(X_i)$	...	Menge aller Nichtteilnehmer mit ähnlichen Merkmalen wie ein Teilnehmer
Cov	...	Varianz-Kovarianz-Matrix
D	...	Teilnahmeindikator
DG	...	Distanzmaß nach Gower
$E(.)$	...	Erwartungswert
G	...	Glättungsmatrix der Local Polynomial Regression
HC	...	Ähnlichkeitsmaß von Hyvärinen
HD	...	heterogene Wertedifferenz
IN	...	latenter Modellteil der Probitschätzung
JC	...	Jaccardkoeffizient
K	...	Kernfunktion der Local Polynomial Regression Schätzung
M	...	Median
MC	...	Matchingkoeffizient
$MDMC$	...	gewichtete Matching-Mahalanobisdistanz
ME	...	Maßnahmeeffekt
MM	...	Minkowski-Metrik
P	...	Prüfgröße für statistische Tests
PN	...	Partizipationsneigung
Pr	...	Wahrscheinlichkeit des Eintritts eines Ereignisses
PS	...	Propensity Score
Q	...	Indikator für die Übereinstimmung von Merkmalsausprägungen
R	...	Ridge-Parameter
SC	...	Smirnoffkoeffizient
SG	...	Ähnlichkeitsmaß von Gower
U	...	Merkmalsraum

W	...	Gewichtung der Nichtteilnehmer für einen Teilnehmer
X	...	Menge der beobachteten Merkmale
Y	...	Einkommen
a	...	Konstante
b	...	Bandbreite der Local Polynomial Regression
d	...	Distanz
$diff$	...	Differenz
g	...	struktureller Zusammenhang zwischen Outcome und beobachtbaren Faktoren
gMC	...	verallgemeinerter Matchingkoeffizient
gMM	...	verallgemeinerte Minkowski-Metrik
k	...	Anzahl Stichproben; $k = 1, \ldots, K$
me	...	Anzahl metrisch skalierter Merkmale
mJC	...	modifizierter Jaccardkoeffizient
n	...	Anzahl beobachteter Merkmale; $n = 1, \ldots, N$
no	...	Anzahl nominal skalierter Merkmale
p	...	Wahrscheinlichkeitsverteilung der Ausprägungen eines Merkmals
r	...	polynomiale Ordnung des Minimierungsproblems der Local Polynomial Regression
s	...	Ähnlichkeit von Objekten
s^2	...	empirische Varianz
sd	...	standardisierte Differenz
v	...	Ausprägungen einer Variable
w	...	Gewichtung eines Teilnehmers für die Gesamtdistanz
x	...	beobachtete Merkmale
z	...	Häufigkeit des Auftretens einer Ausprägung
Δ	...	Änderung zwischen zwei Zeitpunkten
Φ	...	Verteilungsfunktion der Standardnormalverteilung
α	...	Signifikanzniveau
β	...	Koeffizienten der Merkmale im latenten Modellteil der Probitschätzung
ε	...	normalverteilter Störterm
ϵ	...	unbeobachtbare Faktoren
γ	...	Koeffizienten der Einkommensgleichung
λ	...	Gewichtungsfaktor für Übereinstimmungen von Merkmalsausprägungen
κ	...	Koeffizienten eines Minimierungsproblems
μ	...	Erwartungswert der Normalverteilung
ν	...	Störterm der Einkommensgleichung
θ	...	Gewichtungsfaktor für Nichtübereinstimmungen von Merkmalsausprägungen
ρ	...	Koeffizienten der Maßnahmeeffektschätzung

σ	...	Standardabweichung der Normalverteilung
σ^2	...	Varianz der Normalverteilung
υ	...	Störterm der Maßnahmeeffektschätzung
τ	...	aggregierter t-Test auf Mittelwertgleichheit
ψ	...	Toleranzgrenze für die Abweichung von Merkmalsausprägungen

häufig verwendete Indizes

$\cdot^{C}$	...	Kontrollgruppe
$\cdot^{NT}$	...	Nichtteilnehmergruppe
$\cdot^{T}$	...	Teilnehmergruppe
$\cdot_{BC}$	...	Bias-korrigiertes Matching
$\cdot_{C_i}$	...	Unterkontrollgruppe eines Teilnehmers
$\cdot_{M}$	...	Matchinggrößen
$\cdot_{RM}$	...	regression-adjusted Matching
$\cdot_{i}$	...	ein Teilnehmer; $i = 1, \dots I$
$\cdot_{j}$	...	ein Nichtteilnehmer; $j = 1, \dots J$
$\cdot_{n}$	...	ein Merkmal; $n = 1, \dots N$
$\cdot_{nach}$	...	Größen nach Matching
$\cdot_{p}$	...	polynomiale Ordnung eines Minimierungsproblems
$\cdot_{q}$	...	Zeitpunkt vor t
$\cdot_{t}$	...	Zeit; $t = 1, \dots T$
$\cdot_{vor}$	...	Größen vor Matching

häufig verwendete Zeichen

$\hat{\cdot}$	...	geschätzte Größen
$\tilde{\cdot}$	...	Residualgröße der Outcomeschätzung
$\bar{\cdot}$	...	arithmetisches Mittel
$[.]$	...	Intervall
$\lvert . \rvert$	...	absoluter Wert
$\langle . \rangle$	...	Distanz zwischen zwei Objekten/Personen

Anhang B

Ergänzende Informationen zur Simulation

B.1 Definition der Stichproben für die Analyse

Tabelle B.1: **Definition der Stichproben**

		generierte Variablen			Verwendung für ...		
	Merkmale[a]	Mittelwert	Std.dev.[b]	Anz. Ausp.[c]	Matching	ME[d]	Einkommen[d]
x_1	Alter	40,00	8,00		x	x	x
x_2	Kinderanzahl	0,70	1,00		x		x
x_3	Ausbildungsdauer	12,00	2,50		x		x
x_4	Betriebszugehörigkeit	10,00	9,00		x		x
x_5	Nettoeinkommen	1200,00	800,00		x		
x_6	Geschlecht	0,50		2	x		x
x_7	verheiratet	0,64		2	x		x
x_8	dt. Staatsbürgerschaft	0,91		2	x	x	x
x_9	öffentlicher Dienst	0,17		2			x
x_{10}	Neue Bundesländer	0,15		2	x		x
x_{11}	Wirtschaftszweig			3		x	
x_{12}	Schulbildungsniveau			4	x		
x_{13}	Ausbildungsniveau			4	x	x	
x_{14}	Beschäftigungstyp			4	x	x	
x_{15}	Betriebsgröße			4			x
x_{16}	quadrat. Term	(x_1^2)					x
x_{17}	quadrat. Term	(x_4^2)					x
x_{18}	Interaktionsterm	$(x_4 * x_{14})$				x	

Veränderung der Variablen zwischen Teilnehmer- und Nichtteilnehmerstichprobe

$x_1 - x_5$: $\bar{x}_{NT} = \bar{x}_T \pm (\breve{P}_t + k * \sigma)$

$x_6 - x_{10}$: $z_1^{NT} = \breve{P}_{\chi^2} * z_1^T \pm (1 + k)$

$z_0^{NT} = z^{NT} - z_1^{NT}$

x_{11} : $z_3^{NT} = \breve{P}_{\chi^2} * z_3^T \pm (k * s_M)$

$z_1^{NT} = \frac{z_3^{NT} * z_1^T}{z_3^T}$; $z_2^{NT} = z^{NT} - z_1^{NT} - z_3^{NT}$

$x_{12} - x_{15}$: $z_4^{NT} = \breve{P}_{\chi^2} * z_4^T \pm (k * s_M)$

$z_1^{NT} = \frac{z_4^{NT} * z_1^T}{z_4^T}$; $z_2^{NT} = z_2^T$; $z_3^{NT} = z^{NT} - z_1^{NT} - z_2^{NT} - z_4^{NT}$

Anmerkungen:

[a] entsprechende Merkmale im Mikrozensus;
[b] Standardabweichung des entsprechenden Merkmals im Mikrozensus;
[c] Anzahl möglicher Ausprägungen der Variable;
[d] Verwendung der entsprechenden Variablen zur Definition des Maßnahmeeffekts (ME) und der Einkommensgrößen für Teilnahme und Nichtteilnahme.

Skalenniveau der Variablen: x_1–x_5 metrisch (normalverteilt); x_6–x_{10} dichotom; x_{11}–x_{15} polytom.

$\bar{x}_T$, $\bar{x}_{NT}$ – Mittelwert der Variable in Teilnehmer- bzw. Nichtteilnehmerstichprobe;
z_v^T, z_v^{NT} – Häufigkeit des Auftretens einer Variablenausprägung v in Teilnehmer- bzw. Nichtteilnehmerstichprobe;
$\breve{P}_t$, $\breve{P}_{\chi^2}$ – Konstante, orientiert sich an der Prüfgröße statistischer Anpassungstests bei $\alpha = 5\%$ (für normalverteilte Variablen t-Test, für dicotome und polytome χ^2-Homogenitästest);
σ – Standardabweichung der Normalverteilung;
s_M – absolute Abweichung vom Median;
k – gewünschter Umfang der Abweichung (geringe Abweichung: 1% der Streuung eines Merkmals, mittlere 10%, große 25%).

B.2 Ausführliche Ergebnisse der Analyse der Distanzmaße

Tabelle B.2: **Analyse des Propensity Scores**

	Merkmalsmittelwerte			Testergebnisse[c]		Bias		
Merkmale[a]	T[b]	NT[b]	C[b]	spez. Tests[d]	t-Test	vor M.	nach M.	Reduktion
ähnliche metrische und nominale Merkmale								
1	40,05	39,80	37,15	0,67	0,67	1,16	2,96	-155,07
2	0,98	1,04	0,91	0,36	0,38	0,10	0,14	-46,76
3	12,02	11,99	11,12	0,65	0,73	0,51	1,02	-99,70
4	11,95	12,28	10,65	0,42	0,42	1,05	1,67	-59,02
5	1302,33	1323,95	1164,49	0,47	0,52	111,32	204,53	-83,74
6	0,50	0,51	0,51	0,46	0,52	0,10	0,15	-55,11
7	0,66	0,63	0,71	0,44	0,48	0,09	0,14	-49,00
8	0,90	0,89	0,90	0,26	0,32	0,06	0,07	-18,17
9	0,20	0,19	0,34	0,62	0,67	0,08	0,19	-144,62
10	2,84	2,86	2,65	0,79	0,34	0,04	0,23	-539,66
11	2,36	2,37	2,22	0,74	0,21	0,02	0,17	-576,42
12	3,15	3,16	2,94	0,72	0,33	0,04	0,26	-565,61
unähnliche metrische, ähnliche nominale Merkmale								
1	40,06	40,30	38,12	0,58	0,56	2,35	2,44	-3,59
2	0,98	1,08	0,94	0,31	0,39	0,21	0,14	32,52
3	12,03	11,80	11,73	0,41	0,43	1,08	0,67	38,31
4	11,97	12,99	10,89	0,50	0,56	2,50	1,93	23,06
5	1301,84	1383,92	1188,05	0,53	0,61	253,22	214,01	15,49
6	0,50	0,49	0,54	0,25	0,26	0,10	0,09	6,43
7	0,66	0,65	0,70	0,19	0,20	0,09	0,08	15,76
8	0,90	0,90	0,91	0,14	0,23	0,06	0,05	8,23
9	0,20	0,21	0,25	0,25	0,31	0,08	0,09	-16,62
10	2,84	2,85	2,77	0,76	0,19	0,04	0,18	-384,16
11	2,35	2,36	2,32	0,78	0,16	0,03	0,15	-449,91
12	3,16	3,16	3,05	0,76	0,26	0,04	0,23	-473,92
ähnliche metrische, unähnliche nominale Merkmale								
1	40,05	40,08	37,98	0,58	0,54	1,19	2,24	-87,46
2	0,98	1,05	0,92	0,24	0,37	0,10	0,14	-30,25
3	12,03	12,04	11,36	0,57	0,60	0,50	0,85	-69,51
4	11,96	12,42	10,78	0,44	0,47	1,10	1,78	-60,76
5	1302,72	1343,39	1182,57	0,46	0,47	115,95	194,89	-68,08
6	0,50	0,48	0,58	0,49	0,52	0,10	0,16	-57,32
7	0,66	0,65	0,69	0,44	0,49	0,10	0,13	-36,72
8	0,90	0,89	0,92	0,27	0,35	0,06	0,06	-3,41
9	0,20	0,20	0,31	0,53	0,57	0,08	0,17	-107,78
10	2,84	2,92	2,60	0,88	0,38	0,08	0,25	-215,19
11	2,36	2,49	2,09	0,91	0,52	0,14	0,27	-95,59
12	3,16	3,17	3,02	0,79	0,19	0,04	0,20	-386,00
unähnliche metrische und nominale Merkmale								
1	40,06	40,22	38,93	0,38	0,38	2,42	1,85	23,77
2	0,98	1,06	0,97	0,24	0,28	0,20	0,13	34,57
3	12,03	12,15	11,65	0,42	0,44	1,09	0,67	38,49
4	11,94	12,45	11,38	0,39	0,44	2,30	1,74	24,45
5	1301,08	1370,79	1232,12	0,50	0,51	245,21	174,63	28,79
6	0,50	0,50	0,53	0,14	0,16	0,10	0,09	16,15
7	0,65	0,64	0,69	0,24	0,31	0,10	0,10	1,72

Fortsetzung siehe nächste Seite

Fortsetzung Tabelle B.2

	Merkmalsmittelwerte			Testergebnisse[c]		Bias		
Merkmale[a]	T[b]	NT[b]	C[b]	spez. Tests[d]	t-Test	vor M.	nach M.	Reduktion
8	0,90	0,89	0,92	0,17	0,30	0,06	0,05	17,24
9	0,20	0,20	0,25	0,25	0,32	0,08	0,10	-20,19
10	2,84	2,91	2,77	0,83	0,20	0,07	0,16	-138,45
11	2,36	2,49	2,24	0,83	0,16	0,14	0,15	-11,42
12	3,16	3,17	3,12	0,84	0,17	0,04	0,18	-350,86

Anmerkungen:

Durchschnittsergebnisse aus 100 Stichproben.
Abweichung der Mittelwerte bzw. Häufigkeitsverteilungen: ähnliche Merkmale 1% der merkmalsspezifischen Streuung, unähnliche 25%.
[a] Skalenniveau der Variablen: 1-5 metrisch, 6-9 dichotom, 10-12 polytom;
[b] Durchschnitttliche Ausprägungen der Merkmale in den Stichproben (T – Teilnehmer, NT – Nichtteilnehmer, C – Kontrollgruppe);
[c] Durchschnittliche Ablehnungsrate der Nullhypothese gleicher Mittelwerte (Signifikanzniveau 5%);
[d] Skalenspezifische Tests (metrische Variablen: Wilcoxontest, dichotome: McNemartest, polytome: χ^2-Test).

Tabelle B.3: **Analyse des Index Scores**

	Merkmalsmittelwerte			Testergebnisse[c]		Bias		
Merkmale[a]	T[b]	NT[b]	C[b]	spez. Tests[d]	t-Test	vor M.	nach M.	Reduktion
ähnliche metrische und nominale Merkmale								
1	40,05	39,80	37,15	0,67	0,67	1,16	2,96	-155,07
2	0,98	1,04	0,91	0,34	0,38	0,10	0,14	-46,76
3	12,02	11,99	11,12	0,71	0,73	0,51	1,02	-99,70
4	11,95	12,28	10,65	0,42	0,42	1,05	1,67	-59,02
5	1302,33	1323,95	1164,49	0,48	0,52	111,32	204,53	-83,74
6	0,50	0,51	0,51	0,47	0,52	0,10	0,15	-55,11
7	0,66	0,63	0,71	0,44	0,48	0,09	0,14	-49,00
8	0,90	0,89	0,90	0,26	0,32	0,06	0,07	-18,17
9	0,20	0,19	0,34	0,61	0,67	0,08	0,19	-144,62
10	2,84	2,86	2,65	0,79	0,34	0,04	0,23	-539,66
11	2,36	2,37	2,22	0,74	0,21	0,02	0,17	-576,42
12	3,15	3,16	2,94	0,72	0,33	0,04	0,26	-565,61
unähnliche metrische, ähnliche nominale Merkmale								
1	40,06	40,30	38,12	0,54	0,56	2,35	2,44	-3,59
2	0,98	1,08	0,94	0,31	0,39	0,21	0,14	32,52
3	12,03	11,80	11,73	0,45	0,43	1,08	0,67	38,31
4	11,97	12,99	10,89	0,50	0,56	2,50	1,93	23,06
5	1301,84	1383,92	1188,05	0,53	0,61	253,22	214,01	15,49
6	0,50	0,49	0,54	0,22	0,26	0,10	0,09	6,43
7	0,66	0,65	0,70	0,19	0,20	0,09	0,08	15,76
8	0,90	0,90	0,91	0,13	0,23	0,06	0,05	8,23
9	0,20	0,21	0,25	0,26	0,31	0,08	0,09	-16,62
10	2,84	2,85	2,77	0,76	0,19	0,04	0,18	-384,16
11	2,35	2,36	2,32	0,78	0,16	0,03	0,15	-449,91
12	3,16	3,16	3,05	0,76	0,26	0,04	0,23	-473,92
ähnliche metrische, unähnliche nominale Merkmale								
1	40,05	40,08	37,98	0,53	0,54	1,19	2,24	-87,46
2	0,98	1,05	0,92	0,34	0,37	0,10	0,14	-30,25
3	12,03	12,04	11,36	0,60	0,60	0,50	0,85	-69,51
4	11,96	12,42	10,78	0,44	0,47	1,10	1,78	-60,76
5	1302,72	1343,39	1182,57	0,47	0,47	115,95	194,89	-68,08
6	0,50	0,48	0,58	0,47	0,52	0,10	0,16	-57,32

Fortsetzung siehe nächste Seite

Fortsetzung Tabelle B.3

	Merkmalsmittelwerte			Testergebnisse[c]		Bias		
Merkmale[a]	T[b]	NT[b]	C[b]	spez. Tests[d]	t-Test	vor M.	nach M.	Reduktion
7	0,66	0,65	0,69	0,44	0,49	0,10	0,13	-36,72
8	0,90	0,89	0,92	0,27	0,35	0,06	0,06	-3,41
9	0,20	0,20	0,31	0,54	0,57	0,08	0,17	-107,78
10	2,84	2,92	2,60	0,88	0,38	0,08	0,25	-215,19
11	2,36	2,49	2,09	0,91	0,52	0,14	0,27	-95,59
12	3,16	3,17	3,02	0,79	0,19	0,04	0,20	-386,00
unähnliche metrische und nominale Merkmale								
1	40,06	40,22	38,93	0,37	0,38	2,42	1,85	23,77
2	0,98	1,06	0,97	0,23	0,28	0,20	0,13	34,57
3	12,03	12,15	11,65	0,37	0,44	1,09	0,67	38,49
4	11,94	12,45	11,38	0,41	0,44	2,30	1,74	24,45
5	1301,08	1370,79	1232,12	0,47	0,51	245,21	174,63	28,79
6	0,50	0,50	0,53	0,13	0,16	0,10	0,09	16,15
7	0,65	0,64	0,69	0,27	0,31	0,10	0,10	1,72
8	0,90	0,89	0,92	0,16	0,30	0,06	0,05	17,24
9	0,20	0,20	0,25	0,25	0,32	0,08	0,10	-20,19
10	2,84	2,91	2,77	0,83	0,20	0,07	0,16	-138,45
11	2,36	2,49	2,24	0,83	0,16	0,14	0,15	-11,42
12	3,16	3,17	3,12	0,84	0,17	0,04	0,18	-350,86

Anmerkungen: siehe Tabelle B.2.

Tabelle B.4: **Analyse der Mahalanobisdistanz**

	Merkmalsmittelwerte			Testergebnisse[c]		Bias		
Merkmale[a]	T[b]	NT[b]	C[b]	spez. Tests[d]	t-Test	vor M.	nach M.	Reduktion
ähnliche metrische und nominale Merkmale								
1	40,05	39,80	40,03	0,15	0,00	1,16	0,50	56,85
2	0,98	1,04	1,00	0,08	0,01	0,10	0,04	55,29
3	12,02	11,99	12,00	0,18	0,00	0,51	0,21	59,67
4	11,95	12,28	11,98	0,16	0,00	1,05	0,52	50,29
5	1302,33	1323,95	1302,71	0,20	0,00	111,32	55,36	50,27
6	0,50	0,51	0,50	0,01	0,00	0,10	0,02	76,57
7	0,66	0,63	0,67	0,06	0,00	0,09	0,02	77,05
8	0,90	0,89	0,91	0,01	0,00	0,06	0,01	74,73
9	0,20	0,19	0,20	0,00	0,00	0,08	0,01	81,90
10	2,84	2,86	2,88	0,43	0,00	0,04	0,06	-65,71
11	2,36	2,37	2,34	0,50	0,00	0,02	0,04	-63,65
12	3,15	3,16	3,23	0,46	0,01	0,04	0,09	-137,31
unähnliche metrische, ähnliche nominale Merkmale								
1	40,06	40,30	40,20	0,65	0,03	2,35	1,05	55,55
2	0,98	1,08	1,01	0,62	0,03	0,21	0,10	52,52
3	12,03	11,80	11,94	0,62	0,08	1,08	0,42	60,74
4	11,97	12,99	12,33	0,62	0,03	2,50	1,04	58,56
5	1301,84	1383,92	1330,01	0,64	0,00	253,22	105,79	58,22
6	0,50	0,49	0,50	0,04	0,00	0,10	0,03	71,86
7	0,66	0,65	0,66	0,04	0,00	0,09	0,02	75,61
8	0,90	0,90	0,91	0,02	0,00	0,06	0,02	68,78
9	0,20	0,21	0,20	0,02	0,00	0,08	0,02	75,95

Fortsetzung siehe nächste Seite

Fortsetzung Tabelle B.4

	Merkmalsmittelwerte			Testergebnisse[c]		Bias		
Merkmale[a]	T[b]	NT[b]	C[b]	spez. Tests[d]	t-Test	vor M.	nach M.	Reduktion
10	2,84	2,85	2,89	0,49	0,00	0,04	0,06	-64,45
11	2,35	2,36	2,34	0,50	0,00	0,03	0,05	-72,42
12	3,16	3,16	3,23	0,49	0,00	0,04	0,09	-128,76
ähnliche metrische, unähnliche nominale Merkmale								
1	40,05	40,08	40,07	0,16	0,00	1,19	0,54	54,37
2	0,98	1,05	1,00	0,13	0,00	0,10	0,05	47,96
3	12,03	12,04	12,01	0,22	0,00	0,50	0,21	56,91
4	11,96	12,42	12,07	0,20	0,00	1,10	0,56	48,96
5	1302,72	1343,39	1308,43	0,20	0,00	115,95	52,34	54,86
6	0,50	0,48	0,50	0,02	0,00	0,10	0,02	77,72
7	0,66	0,65	0,67	0,05	0,00	0,10	0,02	75,55
8	0,90	0,89	0,91	0,00	0,00	0,06	0,02	70,98
9	0,20	0,20	0,20	0,00	0,00	0,08	0,02	79,79
10	2,84	2,92	2,92	0,35	0,00	0,08	0,08	-0,42
11	2,36	2,49	2,37	0,52	0,00	0,14	0,05	64,88
12	3,16	3,17	3,22	0,48	0,00	0,04	0,09	-107,85
unähnliche metrische und nominale Merkmale								
1	40,06	40,22	40,14	0,65	0,03	2,42	1,05	56,77
2	0,98	1,06	1,01	0,53	0,09	0,20	0,10	51,41
3	12,03	12,15	12,07	0,69	0,10	1,09	0,44	59,53
4	11,94	12,45	12,07	0,51	0,02	2,30	0,98	57,56
5	1301,08	1370,79	1324,59	0,64	0,02	245,21	104,70	57,30
6	0,50	0,50	0,50	0,07	0,00	0,10	0,03	72,11
7	0,65	0,64	0,67	0,06	0,00	0,10	0,02	75,40
8	0,90	0,89	0,92	0,01	0,00	0,06	0,02	68,44
9	0,20	0,20	0,20	0,02	0,00	0,08	0,02	77,23
10	2,84	2,91	2,91	0,52	0,02	0,07	0,08	-16,72
11	2,36	2,49	2,40	0,61	0,00	0,14	0,06	54,95
12	3,16	3,17	3,23	0,42	0,00	0,04	0,09	-135,21

Anmerkungen: siehe Tabelle B.2.

Tabelle B.5: **Analyse der gewichteten Mahalanobis-Matching-Distanz**

	Merkmalsmittelwerte			Testergebnisse[c]		Bias		
Merkmale[a]	T[b]	NT[b]	C[b]	spez. Tests[d]	t-Test	vor M.	nach M.	Reduktion
ähnliche metrische und nominale Merkmale								
1	40,05	39,80	39,94	0,20	0,11	1,16	1,06	8,99
2	0,98	1,04	1,00	0,15	0,02	0,10	0,07	26,83
3	12,02	11,99	12,07	0,15	0,06	0,51	0,35	31,10
4	11,95	12,28	12,30	0,19	0,12	1,05	0,95	9,56
5	1302,33	1323,95	1332,45	0,19	0,20	111,32	119,59	-7,43
6	0,50	0,51	0,50	0,00	0,00	0,10	0,01	89,64
7	0,66	0,63	0,66	0,02	0,00	0,09	0,01	87,39
8	0,90	0,89	0,93	0,12	0,03	0,06	0,03	51,25
9	0,20	0,19	0,18	0,08	0,00	0,08	0,02	68,78
10	2,84	2,86	2,87	0,00	0,00	0,04	0,04	-3,10
11	2,36	2,37	2,34	0,01	0,00	0,02	0,04	-48,47
12	3,15	3,16	3,21	0,00	0,00	0,04	0,05	-34,12

Fortsetzung siehe nächste Seite

Fortsetzung Tabelle B.5

	Merkmalsmittelwerte			Testergebnisse[c]		Bias		
Merkmale[a]	T[b]	NT[b]	C[b]	spez. Tests[d]	t-Test	vor M.	nach M.	Reduktion
unähnliche metrische, ähnliche nominale Merkmale								
1	40,06	40,30	40,34	0,71	0,62	2,35	2,02	14,20
2	0,98	1,08	1,03	0,53	0,31	0,21	0,14	31,45
3	12,03	11,80	11,94	0,65	0,54	1,08	0,70	35,23
4	11,97	12,99	12,77	0,72	0,64	2,50	2,13	14,83
5	1301,84	1383,92	1387,07	0,75	0,77	253,22	250,00	1,27
6	0,50	0,49	0,50	0,00	0,00	0,10	0,01	89,45
7	0,66	0,65	0,66	0,00	0,00	0,09	0,01	87,53
8	0,90	0,90	0,93	0,09	0,04	0,06	0,03	48,01
9	0,20	0,21	0,18	0,08	0,00	0,08	0,02	73,47
10	2,84	2,85	2,87	0,00	0,00	0,04	0,03	14,02
11	2,35	2,36	2,33	0,01	0,00	0,03	0,04	-32,41
12	3,16	3,16	3,21	0,00	0,00	0,04	0,06	-42,57
ähnliche metrische, unähnliche nominale Merkmale								
1	40,05	40,08	40,12	0,16	0,08	1,19	1,01	15,80
2	0,98	1,05	1,01	0,18	0,04	0,10	0,08	23,86
3	12,03	12,04	12,07	0,19	0,11	0,50	0,39	21,39
4	11,96	12,42	12,38	0,22	0,15	1,10	1,08	2,35
5	1302,72	1343,39	1348,28	0,22	0,21	115,95	128,91	-11,18
6	0,50	0,48	0,49	0,00	0,00	0,10	0,01	89,35
7	0,66	0,65	0,66	0,01	0,00	0,10	0,01	87,91
8	0,90	0,89	0,93	0,14	0,02	0,06	0,03	51,59
9	0,20	0,20	0,18	0,05	0,01	0,08	0,02	71,97
10	2,84	2,92	2,88	0,00	0,00	0,08	0,04	49,07
11	2,36	2,49	2,36	0,00	0,00	0,14	0,02	81,69
12	3,16	3,17	3,22	0,00	0,00	0,04	0,07	-62,65
unähnliche metrische und nominale Merkmale								
1	40,06	40,22	40,30	0,76	0,67	2,42	2,10	13,35
2	0,98	1,06	1,03	0,57	0,42	0,20	0,14	28,79
3	12,03	12,15	12,13	0,73	0,58	1,09	0,76	30,18
4	11,94	12,45	12,49	0,64	0,57	2,30	1,98	13,86
5	1301,08	1370,79	1378,85	0,69	0,74	245,21	241,30	1,60
6	0,50	0,50	0,50	0,00	0,00	0,10	0,01	89,16
7	0,65	0,64	0,67	0,00	0,00	0,10	0,01	84,88
8	0,90	0,89	0,93	0,14	0,06	0,06	0,03	52,35
9	0,20	0,20	0,18	0,07	0,02	0,08	0,02	70,35
10	2,84	2,91	2,88	0,00	0,00	0,07	0,04	32,83
11	2,36	2,49	2,36	0,00	0,00	0,14	0,03	78,21
12	3,16	3,17	3,22	0,00	0,00	0,04	0,07	-64,77

Anmerkungen: siehe Tabelle B.2.

Tabelle B.6: **Analyse des Distanzmaßes nach Gower**

	Merkmalsmittelwerte			Testergebnisse[c]		Bias		
Merkmale[a]	T[b]	NT[b]	C[b]	spez. Tests[d]	t-Test	vor M.	nach M.	Reduktion
ähnliche metrische und nominale Merkmale								
1	40,05	39,80	40,02	0,08	0,00	1,16	0,22	80,86
2	0,98	1,04	0,98	0,12	0,00	0,10	0,02	74,74
3	12,02	11,99	12,04	0,11	0,00	0,51	0,11	78,46
4	11,95	12,28	11,92	0,12	0,00	1,05	0,25	75,85

Fortsetzung siehe nächste Seite

Fortsetzung Tabelle B.6

	Merkmalsmittelwerte			Testergebnisse[c]		Bias		
Merkmale[a]	T[b]	NT[b]	C[b]	spez. Tests[d]	t-Test	vor M.	nach M.	Reduktion
5	1302,33	1323,95	1298,01	0,06	0,00	111,32	23,21	79,15
6	0,50	0,51	0,50	0,15	0,18	0,10	0,10	2,18
7	0,66	0,63	0,64	0,14	0,15	0,09	0,09	8,18
8	0,90	0,89	0,90	0,07	0,17	0,06	0,06	-1,87
9	0,20	0,19	0,19	0,15	0,18	0,08	0,08	0,57
10	2,84	2,86	2,86	0,59	0,00	0,04	0,08	-124,57
11	2,36	2,37	2,37	0,62	0,02	0,02	0,08	-236,62
12	3,15	3,16	3,15	0,64	0,01	0,04	0,08	-104,91
unähnliche metrische, ähnliche nominale Merkmale								
1	40,06	40,30	40,08	0,31	0,00	2,35	0,41	82,41
2	0,98	1,08	0,99	0,31	0,00	0,21	0,04	80,62
3	12,03	11,80	11,97	0,47	0,00	1,08	0,23	78,68
4	11,97	12,99	12,08	0,37	0,00	2,50	0,48	80,70
5	1301,84	1383,92	1308,83	0,41	0,00	253,22	43,82	82,69
6	0,50	0,49	0,49	0,16	0,21	0,10	0,10	-3,16
7	0,66	0,65	0,65	0,14	0,20	0,09	0,10	-4,60
8	0,90	0,90	0,90	0,06	0,19	0,06	0,06	3,80
9	0,20	0,21	0,21	0,14	0,19	0,08	0,08	-0,04
10	2,84	2,85	2,85	0,71	0,01	0,04	0,09	-134,32
11	2,35	2,36	2,36	0,63	0,00	0,03	0,07	-147,09
12	3,16	3,16	3,17	0,60	0,00	0,04	0,10	-141,09
ähnliche metrische, unähnliche nominale Merkmale								
1	40,05	40,08	40,06	0,05	0,00	1,19	0,21	82,08
2	0,98	1,05	0,98	0,05	0,00	0,10	0,02	81,10
3	12,03	12,04	12,01	0,15	0,00	0,50	0,11	78,91
4	11,96	12,42	11,93	0,06	0,00	1,10	0,21	80,80
5	1302,72	1343,39	1303,41	0,09	0,00	115,95	27,88	75,96
6	0,50	0,48	0,48	0,15	0,20	0,10	0,10	2,16
7	0,66	0,65	0,66	0,20	0,23	0,10	0,10	0,47
8	0,90	0,89	0,89	0,12	0,31	0,06	0,06	-9,76
9	0,20	0,20	0,20	0,12	0,19	0,08	0,08	4,12
10	2,84	2,92	2,93	0,68	0,03	0,08	0,11	-32,40
11	2,36	2,49	2,50	0,85	0,07	0,14	0,15	-9,40
12	3,16	3,17	3,18	0,61	0,02	0,04	0,09	-126,16
unähnliche metrische und nominale Merkmale								
1	40,06	40,22	40,09	0,21	0,00	2,42	0,39	83,94
2	0,98	1,06	0,98	0,25	0,00	0,20	0,04	81,19
3	12,03	12,15	12,03	0,48	0,00	1,09	0,22	79,83
4	11,94	12,45	11,98	0,25	0,00	2,30	0,44	81,01
5	1301,08	1370,79	1309,36	0,37	0,00	245,21	46,24	81,14
6	0,50	0,50	0,50	0,25	0,29	0,10	0,12	-13,26
7	0,65	0,64	0,65	0,14	0,20	0,10	0,10	1,52
8	0,90	0,89	0,90	0,15	0,25	0,06	0,06	-2,27
9	0,20	0,20	0,20	0,13	0,22	0,08	0,08	-4,35
10	2,84	2,91	2,91	0,69	0,02	0,07	0,08	-27,97
11	2,36	2,49	2,49	0,79	0,08	0,14	0,14	-2,25
12	3,16	3,17	3,18	0,58	0,00	0,04	0,09	-123,65

Anmerkungen: siehe Tabelle B.2.

B.3 Ausführliche Ergebnisse der Analyse der Zuordnungsprozesse

Tabelle B.7: **Stichproben mit 50 Teilnehmern und 50 Nichtteilnehmern**

	Ridge Matching	Replacement Matching	Random Matching
Stichproben mit ähnlichen Merkmalen			
Einkommen			
Teilnahme	2893,32	2893,32	2891,11
Nichtteilnahme „wahr“	2653,58	2653,58	2651,04
Nichtteilnahme vor Matching	2671,59	2671,59	2671,59
Nichtteilnahme nach Matching	2681,79	2702,50	2675,16
Bias			
vor Matching	-18,01	-18,01	-20,56
nach Matching	-28,21	-48,92	-24,13
Reduktion	-56,60	-171,61	-17,38
mittl. quad. Fehler			
MSE	641275,66	72862,31	69306,06
RMSE	800,80	269,93	263,26
emp. Varianz	640480,08	70469,00	68723,86
emp. Std.-abw.	800,30	265,46	262,15
Anteil Varianz am MSE (%)	99,88	96,72	99,16
ausgeschl. Teilnehmer			
Anzahl	0,00	0,00	1,54
Anteil	0,00	0,00	0,03
Summe quad. Distanzen	–	0,58	1,88
Stichproben mit eingeschränkt ähnlichen Merkmalen			
Einkommen			
Teilnahme	2903,77	2903,77	2902,98
Nichtteilnahme „wahr“	2665,95	2665,95	2665,24
Nichtteilnahme vor Matching	2634,02	2634,02	2634,02
Nichtteilnahme nach Matching	2883,23	2657,38	2662,82
Bias			
vor Matching	31,93	31,93	31,22
nach Matching	-217,28	8,57	2,42
Reduktion	-580,53	73,17	92,24
mittl. quad. Fehler			
MSE	638356,76	64890,14	69401,96
RMSE	798,97	254,74	263,44
emp. Varianz	591146,23	64816,75	69396,09
emp. Std.abw.	768,86	254,59	263,43
Anteil Varianz am MSE (%)	92,60	99,89	99,99
ausgeschl. Teilnehmer			
Anzahl	0,00	0,00	0,84
Anteil	0,00	0,00	0,02
Summe quad. Distanzen	–	0,62	1,76

Fortsetzung siehe nächste Seite

Fortsetzung Tabelle B.7

	Ridge Matching	Replacement Matching	Random Matching
Stichproben mit unähnlichen Merkmalen			
Einkommen			
Teilnahme	2928,02	2928,02	2926,04
Nichtteilnahme „wahr“	2689,18	2689,18	2687,26
Nichtteilnahme vor Matching	2734,19	2734,19	2734,19
Nichtteilnahme nach Matching	2937,66	2798,20	2760,25
Bias			
vor Matching	-45,00	-45,00	-46,92
nach Matching	-248,48	-109,02	-72,99
Reduktion	-452,14	-142,24	-55,54
mittl. quad. Fehler			
MSE	984344,04	100529,10	94004,00
RMSE	992,14	317,06	306,60
emp. Varianz	922603,36	88644,75	88676,90
emp. Std.abw.	960,52	297,73	297,79
Anteil Varianz am MSE (%)	93,73	88,18	94,33
ausgeschl. Teilnehmer			
Anzahl	0,00	0,00	0,63
Anteil	0,00	0,00	0,01
Summe quad. Distanzen	–	0,61	1,58

Anmerkungen:

Durchschnittsergebnisse aus 100 Stichproben.
Abweichung der Mittelwerte bzw. Häufigkeitsverteilungen: ähnliche Merkmale 1% der merkmalsspezifischen Streuung, eingeschränkt ähnliche 10%, unähnliche 25%.

Tabelle B.8: **Stichproben mit 100 Teilnehmern und 100 Nichtteilnehmern**

	Ridge Matching	Replacement Matching	Random Matching
Stichproben mit ähnlichen Merkmalen			
Einkommen			
Teilnahme	2918,72	2918,72	2917,43
Nichtteilnahme "wahr"	2682,13	2682,13	2681,16
Nichtteilnahme vor Matching	2700,58	2700,58	2700,58
Nichtteilnahme nach Matching	2723,89	2700,85	2702,68
Bias			
vor Matching	-18,46	-18,46	-19,43
nach Matching	-41,77	-18,72	-21,52
Reduktion	-126,29	-1,45	-10,78
mittl. quad. Fehler			
MSE	235082,90	26265,97	30776,58
RMSE	484,85	162,07	175,43
emp. Varianz	233338,36	25915,37	30313,35
emp. Std.-abw.	483,05	160,98	174,11
Anteil Varianz am MSE (%)	99,26	98,67	98,49

Fortsetzung siehe nächste Seite

Fortsetzung Tabelle B.8

	Ridge Matching	Replacement Matching	Random Matching
ausgeschl. Teilnehmer			
Anzahl	0,00	0,00	3,07
Anteil	0,00	0,00	0,03
Summe quad. Distanzen	–	0,76	5,06
Stichproben mit eingeschränkt ähnlichen Merkmalen			
Einkommen			
Teilnahme	2913,14	2913,14	2912,61
Nichtteilnahme "wahr"	2676,76	2676,76	2676,38
Nichtteilnahme vor Matching	2668,23	2668,23	2668,23
Nichtteilnahme nach Matching	2917,53	2701,73	2684,34
Bias			
vor Matching	8,52	8,52	8,15
nach Matching	-240,77	-24,98	-7,96
Reduktion	-2724,57	-192,99	2,35
mittl. quad. Fehler			
MSE	284235,41	34597,63	37154,55
RMSE	533,14	186,00	192,76
emp. Varianz	226264,72	33973,88	37091,24
emp. Std.-abw.	475,67	184,32	192,59
Anteil Varianz am MSE (%)	79,60	98,20	99,83
ausgeschl. Teilnehmer			
Anzahl	0,00	0,00	0,50
Anteil	0,00	0,00	0,01
Summe quad. Distanzen	–	0,73	2,76
Stichproben mit unähnlichen Merkmalen			
Einkommen			
Teilnahme	2900,15	2900,15	2902,96
Nichtteilnahme "wahr"	2665,16	2665,16	2667,91
Nichtteilnahme vor Matching	2652,76	2652,76	2652,76
Nichtteilnahme nach Matching	2932,12	2692,69	2665,32
Bias			
vor Matching	12,40	12,40	15,15
nach Matching	-266,96	-27,52	2,59
Reduktion	-2053,33	-122,02	82,88
mittl. quad. Fehler			
MSE	457128,79	56789,19	55197,08
RMSE	676,11	238,30	234,94
emp. Varianz	385863,43	56031,61	55190,35
emp. Std.-abw.	621,18	236,71	234,93
Anteil Varianz am MSE (%)	84,41	98,67	99,99
ausgeschl. Teilnehmer			
Anzahl	0,00	0,00	1,01
Anteil	0,00	0,00	0,01
Summe quad. Distanzen	–	0,77	2,87

Anmerkungen: siehe Tabelle B.7.

Tabelle B.9: **Stichproben mit 300 Teilnehmern und 300 Nichtteilnehmern**

	Ridge Matching	Replacement Matching	Random Matching
Stichproben mit ähnlichen Merkmalen			
Einkommen			
Teilnahme	2930,18	2930,18	2930,81
Nichtteilnahme "wahr"	2694,71	2694,71	2695,24
Nichtteilnahme vor Matching	2710,72	2710,72	2710,72
Nichtteilnahme nach Matching	2697,61	2715,31	2712,19
Bias			
vor Matching	-16,01	-16,01	-15,48
nach Matching	-2,90	-20,60	-16,95
Reduktion	81,87	-28,70	-9,50
mittl. quad. Fehler			
MSE	65803,57	7756,53	10514,84
RMSE	256,52	88,07	102,54
emp. Varianz	65795,14	7332,09	10227,58
emp. Std.-abw.	256,51	85,63	101,13
Anteil Varianz am MSE (%)	99,99	94,53	97,27
ausgeschl. Teilnehmer			
Anzahl	0,00	0,00	1,14
Anteil	0,00	0,00	0,00
Summe quad. Distanzen	–	1,09	11,72
Stichproben mit eingeschränkt ähnlichen Merkmalen			
Einkommen			
Teilnahme	2928,17	2928,17	2928,03
Nichtteilnahme "wahr"	2692,13	2692,13	2692,01
Nichtteilnahme vor Matching	2740,14	2740,14	2740,14
Nichtteilnahme nach Matching	2788,32	2759,45	2743,05
Bias			
vor Matching	-48,01	-48,01	-48,14
nach Matching	-96,19	-67,32	-51,05
Reduktion	-100,35	-40,22	-6,05
mittl. quad. Fehler			
MSE	96533,29	22461,19	21811,41
RMSE	310,70	149,87	147,69
emp. Varianz	87281,05	17929,27	19205,71
emp. Std.-abw.	295,43	133,90	138,58
Anteil Varianz am MSE (%)	90,42	79,82	88,05
ausgeschl. Teilnehmer			
Anzahl	0,00	0,00	0,47
Anteil	0,00	0,00	0,00
Summe quad. Distanzen	–	1,08	6,77
Stichproben mit unähnlichen Merkmalen			
Einkommen			
Teilnahme	2929,16	2929,16	2929,43
Nichtteilnahme "wahr"	2693,17	2693,17	2693,43
Nichtteilnahme vor Matching	2796,82	2796,82	2796,82
Nichtteilnahme nach Matching	2785,28	2808,96	2801,45

Fortsetzung siehe nächste Seite

Fortsetzung Tabelle B.9

	Ridge Matching	Replacement Matching	Random Matching
Bias			
vor Matching	-103,65	-103,65	-103,39
nach Matching	-92,11	-115,79	-108,02
Reduktion	11,13	-11,71	-4,48
mittl. quad. Fehler			
MSE	217436,44	46098,18	50748,83
RMSE	466,30	214,70	225,28
emp. Varianz	208951,35	32691,07	39080,53
emp. Std.-abw.	457,11	180,81	197,69
Anteil Varianz am MSE (%)	96,10	70,92	77,01
ausgeschl. Teilnehmer			
Anzahl	0,00	0,00	0,48
Anteil	0,00	0,00	0,00
Summe quad. Distanzen	–	1,11	6,19

Anmerkungen: siehe Tabelle B.7.

Tabelle B.10: **Stichproben mit 50 Teilnehmern und 150 Nichtteilnehmern**

	Ridge Matching	Replacement Matching	opt. 1 : 1 Matching	opt. Full Matching	Matching Matching
Stichproben mit ähnlichen Merkmalen					
Einkommen					
Teilnahme	2912,88	2912,88	2912,88	2912,88	2912,88
Nichtteilnahme "wahr"	2675,75	2675,75	2675,75	2675,75	2675,75
Nichtteilnahme vor Matching	2724,45	2724,45	2724,45	2724,45	2724,45
Nichtteilnahme nach Matching	2657,14	2722,82	2717,17	2700,12	2714,03
Bias					
vor Matching	-48,70	-48,70	-48,70	-48,70	-48,70
nach Matching	18,61	-47,06	-41,42	-24,37	-38,28
Reduktion	61,78	3,36	14,95	49,96	21,40
mittl. quad. Fehler					
MSE	337920,69	53221,29	42482,14	36955,52	45101,17
RMSE	581,31	230,70	206,11	192,24	212,37
emp. Varianz	337574,21	51006,26	40766,71	36361,74	43635,79
emp. Std.-abw.	581,01	225,85	201,91	190,69	208,89
Anteil Varianz am MSE (%)	99,90	95,84	95,96	98,39	96,75
ausgeschl. Teilnehmer					
Anzahl	0,00	0,00	0,00	0,00	0,00
Anteil	0,00	0,00	0,00	0,00	0,00
Summe quad. Distanzen	–	0,40	0,45	0,62	0,49
Stichproben mit eingeschränkt ähnlichen Merkmalen					
Einkommen					
Teilnahme	2917,05	2917,05	2917,05	2917,05	2917,05
Nichtteilnahme "wahr"	2679,64	2679,64	2679,64	2679,64	2679,64
Nichtteilnahme vor Matching	2712,52	2712,52	2712,52	2712,52	2712,52
Nichtteilnahme nach Matching	2753,41	2773,72	2770,26	2731,36	2773,40

Fortsetzung siehe nächste Seite

Fortsetzung Tabelle B.10

	Ridge Matching	Replacement Matching	opt. 1 : 1 Matching	opt. Full Matching	Random Matching
Bias					
vor Matching	-32,88	-32,88	-32,88	-32,88	-32,88
nach Matching	-73,77	-94,08	-90,62	-51,72	-93,76
Reduktion	-124,37	-186,15	-175,61	-57,31	-185,16
mittl. quad. Fehler					
MSE	323305,87	63631,10	60947,67	54830,02	56982,27
RMSE	568,60	252,25	246,88	234,16	238,71
emp. Varianz	317863,87	54779,41	52735,70	52154,91	48191,48
emp. Std.-abw.	563,79	234,05	229,64	228,37	219,53
Anteil Varianz am MSE (%)	98,32	86,09	86,53	95,12	84,57
ausgeschl. Teilnehmer					
Anzahl	0,00	0,00	0,00	0,00	0,00
Anteil	0,00	0,00	0,00	0,00	0,00
Summe quad. Distanzen	–	0,41	0,46	0,64	0,50
Stichproben mit unähnlichen Merkmalen					
Einkommen					
Teilnahme	2911,40	2911,40	2911,40	2911,40	2911,40
Nichtteilnahme "wahr"	2674,75	2674,75	2674,75	2674,75	2674,75
Nichtteilnahme vor Matching	2675,85	2675,85	2675,85	2675,85	2675,85
Nichtteilnahme nach Matching	2937,34	2721,62	2716,39	2690,29	2722,31
Bias					
vor Matching	-1,10	-1,10	-1,10	-1,10	-1,10
nach Matching	-262,59	-46,86	-41,64	-15,54	-47,56
Reduktion	-23879,18	-4179,55	-3702,48	-1319,05	-4243,09
mittl. quad. Fehler					
MSE	635782,07	78898,00	74424,76	75435,74	71387,60
RMSE	797,36	280,89	272,81	274,66	267,18
emp. Varianz	566828,12	76701,73	72690,87	75194,26	69125,62
emp. Std.-abw.	752,88	276,95	269,61	274,22	262,92
Anteil Varianz am MSE (%)	89,15	97,22	97,67	99,68	96,83
ausgeschl. Teilnehmer					
Anzahl	0,00	0,00	0,00	0,00	0,00
Anteil	0,00	0,00	0,00	0,00	0,00
Summe quad. Distanzen	–	0,42	0,47	0,66	0,52

Anmerkungen: siehe Tabelle B.7.

Tabelle B.11: **Stichproben mit 100 Teilnehmern und 300 Nichtteilnehmern**

	Ridge Matching	Replacement Matching	opt. 1 : 1 Matching	opt. Full Matching	Random Matching
Stichproben mit ähnlichen Merkmalen					
Einkommen					
Teilnahme	2922,27	2922,27	2922,27	2922,27	2922,27
Nichtteilnahme "wahr"	2688,03	2688,03	2688,03	2688,03	2688,03
Nichtteilnahme vor Matching	2675,65	2675,65	2675,65	2675,65	2675,65
Nichtteilnahme nach Matching	2656,61	2664,34	2667,92	2660,59	2674,27

Fortsetzung siehe nächste Seite

Fortsetzung Tabelle B.11

	Ridge Matching	Replacement Matching	opt. 1 : 1 Matching	opt. Full Matching	Random Matching
Bias					
vor Matching	12,37	12,37	12,37	12,37	12,37
nach Matching	31,42	23,68	20,11	27,44	13,75
Reduktion	-153,93	-91,42	-62,52	-121,78	-11,16
mittl. quad. Fehler					
MSE	185502,35	17936,60	15130,12	16760,63	15775,54
RMSE	430,70	133,93	123,00	129,46	125,60
emp. Varianz	184515,37	17375,69	14725,83	16007,74	15586,40
emp. Std.-abw.	429,55	131,82	121,35	126,52	124,85
Anteil Varianz am MSE (%)	99,47	96,87	97,33	95,51	98,80
ausgeschl. Teilnehmer					
Anzahl	0,00	-0,48	0,00	0,00	0,00
Anteil	0,00	0,00	0,00	0,00	0,00
Summe quad. Distanzen	–	0,42	0,48	0,71	0,51
Stichproben mit eingeschränkt ähnlichen Merkmalen					
Einkommen					
Teilnahme	2923,87	2923,87	2923,87	2923,87	2923,87
Nichtteilnahme "wahr"	2689,23	2689,23	2689,23	2689,23	2689,23
Nichtteilnahme vor Matching	2730,84	2730,84	2730,84	2730,84	2730,84
Nichtteilnahme nach Matching	2704,84	2757,10	2751,52	2737,96	2757,98
Bias					
vor Matching	-41,60	-41,60	-41,60	-41,60	-41,60
nach Matching	-15,61	-67,87	-62,28	-48,73	-68,74
Reduktion	62,48	-63,13	-49,71	-17,13	-65,24
mittl. quad. Fehler					
MSE	230063,33	31908,81	26984,83	25818,54	30839,86
RMSE	479,65	178,63	164,27	160,68	175,61
emp. Varianz	229819,73	27302,84	23105,58	23444,22	26114,25
emp. Std.-abw.	479,40	165,24	152,01	153,12	161,60
Anteil Varianz am MSE (%)	99,89	85,57	85,62	90,80	84,68
ausgeschl. Teilnehmer					
Anzahl	0,00	0,00	0,00	0,00	0,00
Anteil	0,00	0,00	0,00	0,00	0,00
Summe quad. Distanzen	–	0,40	0,46	0,72	0,50
Stichproben mit unähnlichen Merkmalen					
Einkommen					
Teilnahme	2919,35	2919,35	2919,35	2919,35	2919,35
Nichtteilnahme "wahr"	2684,05	2684,05	2684,05	2684,05	2684,05
Nichtteilnahme vor Matching	2730,59	2730,59	2730,59	2730,59	2730,59
Nichtteilnahme nach Matching	2781,63	2750,48	2764,76	2744,04	2764,80
Bias					
vor Matching	-46,54	-46,54	-46,54	-46,54	-46,54
nach Matching	-97,58	-66,43	-80,70	-59,98	-80,74
Reduktion	-109,67	-42,73	-73,41	-28,89	-73,50
mittl. quad. Fehler					
MSE	347524,46	54737,77	58100,58	56958,84	57868,66
RMSE	589,51	233,96	241,04	238,66	240,56

Fortsetzung siehe nächste Seite

Fortsetzung Tabelle B.11

	Ridge Matching	Replacement Matching	opt. 1 : 1 Matching	opt. Full Matching	Random Matching
emp. Varianz	338002,56	50325,23	51587,65	53360,96	51349,28
emp. Std.-abw.	581,38	224,33	227,13	231,00	226,60
Anteil Varianz am MSE (%)	97,26	91,94	88,79	93,68	88,73
ausgeschl. Teilnehmer					
Anzahl	0,00	0,00	0,00	0,00	0,00
Anteil	0,00	0,00	0,00	0,00	0,00
Summe quad. Distanzen	–	0,42	0,49	0,76	0,53

Anmerkungen: siehe Tabelle B.7.

Tabelle B.12: **Stichproben mit 300 Teilnehmern und 900 Nichtteilnehmern**

	Ridge Matching	Replacement Matching	opt. 1 : 1 Matching	opt. Full Matching	Random Matching
Stichproben mit ähnlichen Merkmalen					
Einkommen					
Teilnahme	2932,07	2932,07	2932,07	2932,07	2932,07
Nichtteilnahme "wahr"	2697,25	2697,25	2697,25	2697,25	2697,25
Nichtteilnahme vor Matching	2699,85	2699,85	2699,85	2699,85	2699,85
Nichtteilnahme nach Matching	2603,49	2707,44	2711,49	2697,13	2708,55
Bias					
vor Matching	-2,60	-2,60	-2,60	-2,60	-2,60
nach Matching	93,76	-10,19	-14,24	0,12	-11,30
Reduktion	-3504,48	-291,82	-447,57	95,54	-334,23
mittl. quad. Fehler					
MSE	60869,85	9309,68	7904,01	7195,79	7677,16
RMSE	246,72	96,49	88,90	84,83	87,62
emp. Varianz	52078,18	9205,79	7701,12	7195,77	7549,57
emp. Std.-abw.	228,21	95,95	87,76	84,83	86,89
Anteil Varianz am MSE (%)	85,56	98,88	97,43	100,00	98,34
ausgeschl. Teilnehmer					
Anzahl	0,00	0,00	-0,37	0,00	0,00
Anteil	0,00	0,00	0,00	0,00	0,00
Summe quad. Distanzen	–	0,49	0,61	0,98	0,65
Stichproben mit eingeschränkt ähnlichen Merkmalen					
Einkommen					
Teilnahme	2932,04	2932,04	2932,04	2932,04	2932,04
Nichtteilnahme "wahr"	2696,75	2696,75	2696,75	2696,75	2696,75
Nichtteilnahme vor Matching	2762,51	2762,51	2762,51	2762,51	2762,51
Nichtteilnahme nach Matching	2596,81	2781,70	2784,36	2767,31	2786,07
Bias					
vor Matching	-65,76	-65,76	-65,76	-65,76	-65,76
nach Matching	99,94	-84,95	-87,61	-70,56	-89,32
Reduktion	-51,98	-29,19	-33,23	-7,30	-35,83
mittl. quad. Fehler					
MSE	105011,09	20659,24	20341,48	18006,77	21304,23
RMSE	324,05	143,73	142,62	134,19	145,96
emp. Varianz	95023,30	13442,93	12665,97	13028,27	13326,53

Fortsetzung siehe nächste Seite

Fortsetzung Tabelle B.12

	Ridge Matching	Replacement Matching	opt. 1 : 1 Matching	opt. Full Matching	Random Matching
emp. Std.-abw.	308,26	115,94	112,54	114,14	115,44
Anteil Varianz am MSE (%)	90,49	65,07	62,27	72,35	62,55
ausgeschl. Teilnehmer					
Anzahl	0,00	0,00	0,00	0,00	0,00
Anteil	0,00	0,00	0,00	0,00	0,00
Summe quad. Distanzen	–	0,48	0,60	0,99	0,64
Stichproben mit unähnlichen Merkmalen					
Einkommen					
Teilnahme	2931,25	2931,25	2931,25	2931,25	2931,25
Nichtteilnahme "wahr"	2695,84	2695,84	2695,84	2695,84	2695,84
Nichtteilnahme vor Matching	2778,64	2778,64	2778,64	2778,64	2778,64
Nichtteilnahme nach Matching	2665,77	2796,74	2803,45	2781,35	2804,19
Bias					
vor Matching	-82,79	-82,79	-82,79	-82,79	-82,79
nach Matching	30,08	-100,89	-107,61	-85,51	-108,34
Reduktion	63,67	-21,86	-29,97	-3,28	-30,86
mittl. quad. Fehler					
MSE	232524,76	40031,25	42460,22	38451,95	42341,67
RMSE	482,21	200,08	206,06	196,09	205,77
emp. Varianz	231620,00	29851,96	30881,19	31140,34	30603,82
emp. Std.-abw.	481,27	172,78	175,73	176,47	174,94
Anteil Varianz am MSE (%)	99,61	74,57	72,73	80,99	72,28
ausgeschl. Teilnehmer					
Anzahl	0,00	0,00	0,00	0,00	0,00
Anteil	0,00	0,00	0,00	0,00	0,00
Summe quad. Distanzen	–	0,50	0,64	1,05	0,69

Anmerkungen: siehe Tabelle B.7.

Tabelle B.13: **Stichproben mit 50 Teilnehmern und 500 Nichtteilnehmern**

	Ridge Matching	Replacement Matching	opt. 1 : 1 Matching	opt. Full Matching	Random Matching
Stichproben mit ähnlichen Merkmalen					
Einkommen					
Teilnahme	2919,89	2919,89	2919,89	2919,89	2919,89
Nichtteilnahme "wahr"	2684,64	2684,64	2684,64	2684,64	2684,64
Nichtteilnahme vor Matching	2723,89	2723,89	2723,89	2723,89	2723,89
Nichtteilnahme nach Matching	2473,43	2725,25	2719,39	2693,55	2725,43
Bias					
vor Matching	-39,25	-39,25	-39,25	-39,25	-39,25
nach Matching	211,21	-40,61	-34,75	-8,91	-40,80
Reduktion	-438,14	-3,48	11,46	77,30	-3,95
mittl. quad. Fehler					
MSE	253572,94	37684,50	35760,40	24688,09	36756,32
RMSE	503,56	194,12	189,10	157,12	191,72
emp. Varianz	208964,64	36035,12	34552,73	24608,73	35091,96

Fortsetzung siehe nächste Seite

Fortsetzung Tabelle B.13

	Ridge Matching	Replacement Matching	opt. 1 : 1 Matching	opt. Full Matching	Random Matching
emp. Std.-abw.	457,13	189,83	185,88	156,87	187,33
Anteil Varianz am MSE (%)	82,41	95,62	96,62	99,68	95,47
ausgeschl. Teilnehmer					
Anzahl	0,00	0,00	0,00	0,00	0,00
Anteil	0,00	0,00	0,00	0,00	0,00
Summe quad. Distanzen	–	0,21	0,22	0,54	0,23
Stichproben mit eingeschränkt ähnlichen Merkmalen					
Einkommen					
Teilnahme	2914,28	2914,28	2914,28	2914,28	2914,28
Nichtteilnahme "wahr"	2678,22	2678,22	2678,22	2678,22	2678,22
Nichtteilnahme vor Matching	2736,61	2736,61	2736,61	2736,61	2736,61
Nichtteilnahme nach Matching	2626,95	2768,10	2767,92	2732,58	2773,57
Bias					
vor Matching	-58,38	-58,38	-58,38	-58,38	-58,38
nach Matching	51,27	-89,88	-89,70	-54,35	-95,35
Reduktion	12,18	-53,94	-53,64	6,90	-63,32
mittl. quad. Fehler					
MSE	264804,29	50239,44	47660,34	39809,69	48519,38
RMSE	514,59	224,14	218,31	199,52	220,27
emp. Varianz	262175,77	42161,69	39614,57	36855,26	39427,89
emp. Std.-abw.	512,03	205,33	199,03	191,98	198,56
Anteil Varianz am MSE (%)	99,01	83,92	83,12	92,58	81,26
ausgeschl. Teilnehmer					
Anzahl	0,00	0,00	0,00	0,00	0,00
Anteil	0,00	0,00	0,00	0,00	0,00
Summe quad. Distanzen	–	0,21	0,22	0,55	0,22
Stichproben mit unähnlichen Merkmalen					
Einkommen					
Teilnahme	2911,11	2911,11	2911,11	2911,11	2911,11
Nichtteilnahme "wahr"	2675,69	2675,69	2675,69	2675,69	2675,69
Nichtteilnahme vor Matching	2777,94	2777,94	2777,94	2777,94	2777,94
Nichtteilnahme nach Matching	2576,75	2770,85	2775,91	2756,28	2781,66
Bias					
vor Matching	-102,25	-102,25	-102,25	-102,25	-102,25
nach Matching	98,93	-95,16	-100,23	-80,60	-105,97
Reduktion	3,25	6,94	1,98	21,18	-3,64
mittl. quad. Fehler					
MSE	462284,86	74360,87	75591,75	63824,79	75983,99
RMSE	679,92	272,69	274,94	252,64	275,65
emp. Varianz	452497,15	65304,88	65546,37	57328,56	64753,66
emp. Std.-abw.	672,68	255,55	256,02	239,43	254,47
Anteil Varianz am MSE (%)	97,88	87,82	86,71	89,82	85,22

Fortsetzung siehe nächste Seite

Fortsetzung Tabelle B.13

	Ridge Matching	Replacement Matching	opt. 1 : 1 Matching	opt. Full Matching	Random Matching
ausgeschl. Teilnehmer					
Anzahl	0,00	0,00	0,00	0,00	0,00
Anteil	0,00	0,00	0,00	0,00	0,00
Summe quad. Distanzen	–	0,21	0,23	0,57	0,23

Anmerkungen: siehe Tabelle B.7.

Tabelle B.14: **Stichproben mit 100 Teilnehmern und 1000 Nichtteilnehmern**

	Ridge Matching	Replacement Matching	opt. 1 : 1 Matching	opt. Full Matching	Random Matching
Stichproben mit ähnlichen Merkmalen					
Einkommen					
Teilnahme	2926,00	2926,00	2926,00	2926,00	2926,00
Nichtteilnahme "wahr"	2692,38	2692,38	2692,38	2692,38	2692,38
Nichtteilnahme vor Matching	2663,96	2663,96	2663,96	2663,96	2663,96
Nichtteilnahme nach Matching	2567,73	2674,11	2676,57	2650,13	2674,25
Bias					
vor Matching	28,42	28,42	28,42	28,42	28,42
nach Matching	124,65	18,27	15,80	42,24	18,13
Reduktion	-338,65	35,72	44,39	-48,66	36,20
mittl. quad. Fehler					
MSE	86906,42	16624,04	15313,79	11510,81	16220,15
RMSE	294,80	128,93	123,75	107,29	127,36
emp. Varianz	71369,69	16290,42	15064,10	9726,31	15891,44
emp. Std.-abw.	267,15	127,63	122,74	98,62	126,06
Anteil Varianz am MSE (%)	82,12	97,99	98,37	84,50	97,97
ausgeschl. Teilnehmer					
Anzahl	0,00	0,00	0,00	0,00	0,00
Anteil	0,00	0,00	0,00	0,00	0,00
Summe quad. Distanzen	–	0,19	0,20	0,62	0,21
Stichproben mit eingeschränkt ähnlichen Merkmalen					
Einkommen					
Teilnahme	2923,67	2923,67	2923,67	2923,67	2923,67
Nichtteilnahme "wahr"	2689,77	2689,77	2689,77	2689,77	2689,77
Nichtteilnahme vor Matching	2759,67	2759,67	2759,67	2759,67	2759,67
Nichtteilnahme nach Matching	2494,36	2760,96	2767,85	2751,14	2765,81
Bias					
vor Matching	-69,90	-69,90	-69,90	-69,9	-69,90
nach Matching	195,41	-71,19	-78,08	-61,38	-76,04
Reduktion	-179,54	-1,84	-11,69	12,2	-8,78
mittl. quad. Fehler					
MSE	197828,91	27751,64	29122,79	25881,31	28992,52
RMSE	444,78	166,59	170,65	160,88	170,27
emp. Varianz	159643,92	22683,35	23026,42	22114,29	23210,29
emp. Std.-abw.	399,55	150,61	151,74	148,71	152,35
Anteil Varianz am MSE (%)	80,70	81,74	79,07	85,45	80,06

Fortsetzung siehe nächste Seite

Fortsetzung Tabelle B.14

	Ridge Matching	Replacement Matching	opt. 1 : 1 Matching	opt. Full Matching	Random Matching
ausgeschl. Teilnehmer					
Anzahl	0,00	0,00	0,00	0,00	0,00
Anteil	0,00	0,00	0,00	0,00	0,00
Summe quad. Distanzen	–	0,18	0,19	0,64	0,20
Stichproben mit unähnlichen Merkmalen					
Einkommen					
Teilnahme	2921,08	2921,08	2921,08	2921,08	2921,08
Nichtteilnahme "wahr"	2687,30	2687,30	2687,30	2687,3	2687,30
Nichtteilnahme vor Matching	2779,82	2779,82	2779,82	2779,82	2779,82
Nichtteilnahme nach Matching	2592,51	2775,41	2777,74	2756,93	2776,85
Bias					
vor Matching	-92,52	-92,52	-92,52	-92,52	-92,52
nach Matching	94,78	-88,11	-90,44	-69,63	-89,55
Reduktion	-2,44	4,77	2,25	24,74	3,22
mittl. quad. Fehler					
MSE	265761,40	43159,76	43494,95	46892,42	42193,30
RMSE	515,52	207,75	208,55	216,55	205,41
emp. Varianz	256777,43	35396,46	35315,92	42043,47	34174,35
emp. Std.-abw.	506,73	188,14	187,93	205,05	184,86
Anteil Varianz am MSE (%)	96,62	82,01	81,20	89,66	80,99
ausgeschl. Teilnehmer					
Anzahl	0,00	0,00	0,00	0,00	0,00
Anteil	0,00	0,00	0,00	0,00	0,00
Summe quad. Distanzen	–	0,19	0,21	0,68	0,21

Anmerkungen: siehe Tabelle B.7.

Tabelle B.15: **Stichproben mit 300 Teilnehmern und 3000 Nichtteilnehmern**

	Ridge Matching	Replacement Matching	opt. 1 : 1 Matching	opt. Full Matching	Random Matching
Stichproben mit ähnlichen Merkmalen					
Einkommen					
Teilnahme	2932,10	2932,10	2932,10	2932,10	2932,10
Nichtteilnahme "wahr"	2697,43	2697,43	2697,43	2697,43	2697,43
Nichtteilnahme vor Matching	2709,08	2709,08	2709,08	2709,08	2709,08
Nichtteilnahme nach Matching	2503,47	2699,43	2699,72	2688,24	2700,19
Bias					
vor Matching	-11,65	-11,65	-11,65	-11,65	-11,65
nach Matching	193,96	-2,00	-2,29	9,19	-2,76
Reduktion	-1564,25	82,83	80,33	21,18	76,33
mittl. quad. Fehler					
MSE	62060,44	6015,25	5713,02	4485,03	6061,34
RMSE	249,12	77,56	75,58	66,97	77,85
emp. Varianz	24441,62	6011,24	5707,76	4400,66	6053,73
emp. Std.-abw.	156,34	77,53	75,55	66,34	77,81
Anteil Varianz am MSE (%)	39,38	99,93	99,91	98,12	99,87

Fortsetzung siehe nächste Seite

Fortsetzung Tabelle B.15

	Ridge Matching	Replacement Matching	opt. 1 : 1 Matching	opt. Full Matching	Random Matching
ausgeschl. Teilnehmer					
Anzahl	0,00	-1,06	0,00	0,00	0,00
Anteil	0,00	0,00	0,00	0,00	0,00
Summe quad. Distanzen	–	0,17	0,19	0,82	0,19
Stichproben mit eingeschränkt ähnlichen Merkmalen					
Einkommen					
Teilnahme	2931,47	2931,47	2931,47	2931,47	2931,47
Nichtteilnahme "wahr"	2696,59	2696,59	2696,59	2696,59	2696,59
Nichtteilnahme vor Matching	2725,04	2725,04	2725,04	2725,04	2725,04
Nichtteilnahme nach Matching	2560,14	2752,46	2753,41	2727,41	2753,28
Bias					
vor Matching	-28,45	-28,45	-28,45	-28,45	-28,45
nach Matching	136,45	-55,87	-56,82	-30,82	-56,69
Reduktion	-379,70	-96,41	-99,74	-8,35	-99,30
mittl. quad. Fehler					
MSE	88799,65	14376,73	13531,25	9951,77	14189,89
RMSE	297,99	119,90	116,32	99,76	119,12
emp. Varianz	70180,61	11255,29	10303,06	9001,94	10976,11
emp. Std.-abw.	264,92	106,09	101,50	94,88	104,77
Anteil Varianz am MSE (%)	79,03	78,29	76,14	90,46	77,35
ausgeschl. Teilnehmer					
Anzahl	0,00	0,00	0,00	0,00	0,00
Anteil	0,00	0,00	0,00	0,00	0,00
Summe quad. Distanzen	–	0,16	0,18	0,82	0,19
Stichproben mit unähnlichen Merkmalen					
Einkommen					
Teilnahme	2932,27	2932,27	2932,27	2932,27	2932,27
Nichtteilnahme "wahr"	2697,28	2697,28	2697,28	2697,28	2697,28
Nichtteilnahme vor Matching	2793,53	2793,53	2793,53	2793,53	2793,53
Nichtteilnahme nach Matching	2485,07	2794,39	2798,29	2791,94	2797,50
Bias					
vor Matching	-96,25	-96,25	-96,25	-96,25	-96,25
nach Matching	212,21	-97,11	-101,01	-94,66	-100,22
Reduktion	-120,48	-0,89	-4,94	1,65	-4,12
mittl. quad. Fehler					
MSE	237676,46	37787,06	38870,21	39300,64	38642,55
RMSE	487,52	194,39	197,16	198,24	196,58
emp. Varianz	192644,14	28357,55	28667,43	30340,60	28598,89
emp. Std.-abw.	438,91	168,40	169,31	174,19	169,11
Anteil Varianz am MSE (%)	81,05	75,05	73,75	77,20	74,01
ausgeschl. Teilnehmer					
Anzahl	0,00	0,00	0,00	0,00	0,00
Anteil	0,00	0,00	0,00	0,00	0,00
Summe quad. Distanzen	–	0,18	0,19	0,89	0,20

Anmerkungen: siehe Tabelle B.7.

Anhang C

Ergänzende Informationen zur Evaluation der Förderung der Berufsausbildung in den Neuen Bundesländern

C.1 Deskriptive Analyse der Stichprobe

Tabelle C.1: **Detaillierte deskriptive Statistik der Jugendlichen mit abgeschlossener Berufsausbildung**

Merkmale	gesamte Stichprobe	ungeförderte Jugendliche	geförderte Jugendliche		
			gesamt	betriebsnah	außerbetr.
Anzahl Personen	3048	2556	492	324	168
Teilstichproben					
BAB ungefördert	0,84	–	0,00	–	–
BAB gefördert	0,16	–	1,00	–	–
BAB gbetriebsnah	0,11	–	0,66	–	–
BAB außerbetrieblich	0,06	–	0,34	–	–
soziodemografische Faktoren					
Alter	19,96	20,00	19,73	19,67	19,85
Mann	0,54	0,56	0,45	0,40	0,54
Frau	0,46	0,44	0,55	0,60	0,46
in BRD geboren	1,00	1,00	1,00	1,00	0,99
dt. Staatsbürger	1,00	1,00	1,00	1,00	1,00
Eltern dt. Staatsbürger	1,00	1,00	1,00	1,00	1,00
eigene Kinder	0,03	0,03	0,02	0,02	0,02
keine eigenen Kinder	0,97	0,97	0,98	0,98	0,98
Elternhaushalt	0,56	0,57	0,54	0,52	0,57
eigener Haushalt	0,44	0,43	0,46	0,48	0,43
kein Schulabschluss	0,01	0,01	0,03	0,02	0,04
Hauptschulabschluss	0,15	0,12	0,27	0,24	0,33
Realschulabschluss	0,76	0,77	0,67	0,71	0,59
Abitur	0,09	0,10	0,03	0,03	0,04
Abschlusszensur[a]	2,46	2,41	2,57	2,56	2,62
Charakteristika der Berufsausbildung					
vertragl. geb. BAB	0,78	0,79	0,68	0,68	0,68
schulische BAB	0,22	0,21	0,32	0,32	0,32
Zusatzqualifikation	0,19	0,20	0,17	0,17	0,17
keine Zusatzqualifikation	0,81	0,80	0,83	0,83	0,83
Facharbeiter[a]	0,44	0,44	0,43	0,43	0,44
Geselle[a]	0,13	0,15	0,10	0,09	0,14
Fachangestellter[a]	0,18	0,19	0,16	0,18	0,12
Assistent[a]	0,13	0,13	0,11	0,10	0,12
Techniker[a]	0,01	0,01	0,01	0,01	0,00
anderer Abschluss[a]	0,11	0,07	0,19	0,20	0,16
Land-, Forstwirtschaft	0,03	0,03	0,03	0,04	0,01
Bergbau, Mineralgew.	0,00	0,00	0,00	0,00	0,00
Metall-, Elektroberufe	0,20	0,22	0,11	0,11	0,10
Bau-, Ausbauberufe	0,08	0,09	0,07	0,04	0,13
sonst. Fertigungsberufe	0,11	0,12	0,11	0,10	0,11
technische Berufe	0,04	0,04	0,04	0,03	0,05
Waren-, DL-kaufleute	0,13	0,13	0,15	0,17	0,11
Org., Verwaltung, Büro	0,16	0,16	0,16	0,14	0,20
Gesundheitsdienst	0,09	0,09	0,12	0,13	0,11
Sozial-, Erziehungsberufe	0,05	0,05	0,06	0,07	0,04
sonst. Dienstleistungen	0,09	0,08	0,15	0,16	0,13
sonst. Arbeitskräfte	0,00	0,00	0,00	0,00	0,01

Fortsetzung siehe nächste Seite

Fortsetzung Tabelle C.1

Merkmale	gesamte Stichprobe	ungeförderte Jugendliche	geförderte Jugendliche		
			gesamt	betriebsnah	außerbetr.
allgemeine Lage auf dem Arbeitsmarkt					
BAB ABL	0,10	0,10	0,10	0,13	0,04
BAB NBL	0,90	0,90	0,90	0,87	0,96
BAB Brandenburg	0,13	0,13	0,12	0,13	0,10
BAB Mecklenburg-Vorpommern	0,10	0,11	0,09	0,09	0,10
BAB Sachsen	0,30	0,29	0,34	0,35	0,32
BAB Sachsen-Anhalt	0,15	0,16	0,14	0,12	0,17
BAB Thüringen	0,22	0,22	0,22	0,18	0,28
BAB-Ende 1995-1998	0,01	0,01	0,00	0,00	0,01
BAB-Ende 1999-2002	0,46	0,51	0,21	0,14	0,34
BAB-Ende 2003-2006	0,53	0,48	0,79	0,86	0,65
persönliche Arbeitsmarkt-Vorgeschichte vor Ausbildungsbeginn					
keine Unterbrechung	0,42	0,44	0,30	0,36	0,20
Zivildienst / Bund	0,02	0,02	0,01	0,01	0,02
Erwerbstätigkeit	0,01	0,01	0,00	0,00	0,00
Arbeitslosigkeit	0,05	0,04	0,07	0,05	0,11
abgebr. Ausbildung	0,04	0,03	0,05	0,06	0,05
zusätzliche Informationen					
BAB-Theorie (eher) gut	0,87	0,86	0,90	0,90	0,92
BAB-Theorie (eher) schlecht	0,13	0,14	0,10	0,10	0,08
BAB-Praxis (eher) gut	0,87	0,87	0,88	0,88	0,89
BAB-Praxis (eher) schlecht	0,13	0,13	0,12	0,12	0,11
BAB-Anford. (eher) hoch	0,71	0,71	0,69	0,73	0,63
BAB-Anford. (eher) niedrig	0,29	0,29	0,31	0,27	0,38
BAB-Belastung (eher) hoch	0,35	0,33	0,42	0,44	0,36
BAB-Belastung (eher) niedrig	0,65	0,67	0,58	0,56	0,64
BAB auf eigene Initiative	0,58	0,62	0,40	0,45	0,31
BAB m.H. anderer	0,20	0,21	0,17	0,20	0,13
BAB m.H. öff. Stellen	0,27	0,23	0,50	0,45	0,60
Motivation	0,60	0,64	0,42	0,47	0,33
Arbeitsmarktstatus direkt nach Ausbildungsabschluss					
Erwerbstätigkeit	0,47	0,51	0,26	0,27	0,22
Arbeitslosigkeit	0,35	0,31	0,53	0,52	0,56
neue Ausbildung	0,04	0,04	0,05	0,06	0,03
geförderte Beschäftigung	0,01	0,01	0,01	0,01	0,00
anderer Status	0,13	0,12	0,15	0,14	0,19

Anmerkungen:

Anteil Jugendlicher mit entsprechendem Merkmal; Ausnahmen: Alter, Abschlusszensur (arithmetisches Mittel).
BAB – Berufsausbildung.
[a] Angaben für ca. 40 % der Jugendlichen in der Stichprobe verfügbar.

Quelle: *Jugendpanel des zsh.*

C.2 Prüfung der Matchingergebnisse

Tabelle C.2: **Bewertung der Matchingergebnisse zur außerbetrieblichen Berufsausbildung**

Ausgangsdaten			Match Replacement			opt. Full Match			opt. 1 : 1 Match			Random Match		
Merkmale	T[a]	NT[a]	C[a]	Test[b]	p-value	C[a]	Test[b]	p-value	C[a]	Test[b]	p-value	C[a]	Test[b]	p-value
soziodemographische Faktoren														
Alter	19,85	20,00	19,90	0	0,63	19,95	0	0,09	19,88	0	0,91	19,89	0	0,71
Mann	0,54	0,56	0,52	0	0,61	0,55	0	0,63	0,52	0	0,39	0,52	0	0,61
kein Schulabschluss	0,04	0,01	0,02	0	0,13	0,01	0	0,06	0,01	0	0,06	0,01	0	0,06
Hauptschulabschluss	0,33	0,12	0,34	0	1,00	0,34	0	1,00	0,34	0	1,00	0,33	0	1,50
Realschulabschluss	0,59	0,77	0,61	0	0,25	0,61	0	0,25	0,61	0	0,13	0,61	0	0,13
Abitur	0,04	0,10	0,04	0	2,00	0,05	0	0,25	0,04	0	2,00	0,04	0	1,00
eigene Kinder	0,02	0,03	0,03	0	1,00	0,01	0	0,69	0,04	0	0,73	0,04	0	0,73
Elternhaushalt	0,57	0,57	0,61	0	0,10	0,59	0	0,45	0,61	0	0,10	0,61	0	0,10
Charakteristika der Berufsausbildung														
vertragl. geb. BAB	0,68	0,79	0,69	0	1,00	0,71	0	0,06	0,70	0	0,50	0,69	0	1,00
Zusatzqualifikation	0,17	0,20	0,16	0	0,50	0,18	0	0,50	0,17	0	1,00	0,17	0	1,00
Land-, Forstwirtschaft	0,01	0,03	0,04	0	0,18	0,02	0	1,00	0,04	0	0,18	0,05	0	0,11
Metall-, Elektroberufe	0,10	0,22	0,10	0	2,00	0,11	0	1,00	0,10	0	2,00	0,11	0	0,50
Bau-, Ausbauberufe	0,13	0,09	0,11	0	0,75	0,08	0	0,07	0,11	0	0,58	0,11	0	0,61
sonst. Fertigungsberufe	0,11	0,12	0,15	0	0,12	0,13	0	0,58	0,15	0	0,12	0,14	0	0,42
technische Berufe	0,05	0,04	0,03	0	0,38	0,01	1	0,04	0,04	0	1,00	0,03	0	0,45
Waren-, DL-kaufleute	0,11	0,13	0,08	0	0,27	0,07	1	0,04	0,08	0	0,15	0,09	0	0,42
Org., Verwaltungs, Büro	0,20	0,16	0,19	0	0,79	0,13	1	0,00	0,19	0	0,79	0,18	0	0,58
Gesundheitsdienst	0,11	0,09	0,14	0	0,23	0,10	0	0,51	0,15	0	0,11	0,14	0	0,18
Sozial-, Erziehungsberufe	0,04	0,05	0,05	0	0,63	0,06	0	0,29	0,05	0	0,63	0,05	0	0,38
sonst. Dienstleistungen	0,13	0,08	0,10	0	0,11	0,06	1	0,01	0,09	0	0,07	0,10	0	0,15
allgemeine Lage auf dem Arbeitsmarkt														
Region d. BAB	1,00	1,00	1,00	0	0,49	1,00	0	0,41	1,00	0	0,98	1,00	0	0,93
BAB-Ende 1995-1998	0,01	0,01	0,01	0	2,00	0,01	0	1,00	0,01	0	2,00	0,01	0	2,00
BAB-Ende 1999-2002	0,34	0,51	0,36	0	0,25	0,38	1	0,03	0,36	0	0,13	0,35	0	0,50
BAB-Ende 2003-2006	0,65	0,48	0,64	0	0,25	0,61	1	0,02	0,63	0	0,13	0,64	0	0,50

Fortsetzung siehe nächste Seite

Fortsetzung Tabelle C.2

Ausgangsdaten			Match Replacement			opt. Full Match			opt. 1 : 1 Match			Random Match		
Merkmale	T[a]	NT[a]	C[a]	Test[b]	p-value	C[a]	Test[b]	p-value	C[a]	Test[b]	p-value	C[a]	Test[b]	p-value
persönliche Arbeitsmarkt-Vorgeschichte														
keine Unterbrechung	0,20	0,44	0,20	0	0,77	0,27	1	0,00	0,20	0	1,00	0,21	0	0,79
Zivildienst / Bund	0,02	0,02	0,01	0	1,00	0,01	0	0,50	0,01	0	1,00	0,02	0	1,50
Arbeitslosigkeit	0,11	0,04	0,05	0	0,10	0,01	1	0,00	0,05	0	0,06	0,05	0	0,06
Erwerbstätigkeit	0,00	0,01	0,00	0	1,00	0,00	0	1,00	0,00	0	1,00	0,01	0	1,00
abgebr. Ausbildung	0,05	0,03	0,02	0	0,07	0,04	0	0,63	0,03	0	0,29	0,02	0	0,07
zusätzliche Indikatoren														
Netzwerk	0,13	0,21	0,13	0	0,75	0,23	1	0,00	0,13	0	0,75	0,13	0	1,00
Belastung (eher) hoch	0,36	0,33	0,36	0	1,27	0,35	0	0,69	0,37	0	1,00	0,36	0	1,00
Motivation	0,33	0,64	0,33	0	2,00	0,38	1	0,01	0,35	0	0,50	0,33	0	2,00
Summe quad. Distanzen					0,57			2,08			0,67			0,70

Anmerkungen:

[a] Anteil Personen mit entsprechendem Merkmal in der Stichprobe der geförderten Jugendlichen (T), der ungeförderten Jugendlichen (NT) bzw. der Kontrollgruppe (C);
[b] Skalenspezifische Tests (metrische Variablen: Wilcoxontest, dichotome: McNemartest, polytome: χ^2-Test); Signifikanzniveau 5%.

Tabelle C.3: **Bewertung der Matchingergebnisse zur betriebsnahen Berufsausbildung**

Ausgangsdaten			Match Replacement			opt. Full Match			opt. 1 : 1 Match			Random Match		
Merkmale	T[a]	NT[a]	C[a]	Test[b]	p-value	C[a]	Test[b]	p-value	C[a]	Test[b]	p-value	C[a]	Test[b]	p-value
soziodemographische Faktoren														
Alter	19,67	20,00	19,85	1	0,00	19,90	1	0,00	19,82	1	0,01	19,81	1	0,01
Mann	0,40	0,56	0,39	0	1,00	0,45	1	0,00	0,40	0	0,71	0,41	0	0,06
kein Schulabschluss	0,02	0,01	0,00	0	0,06	0,00	0	0,06	0,00	0	0,06	0,00	0	0,06
Hauptschulabschluss	0,24	0,12	0,23	0	1,00	0,23	0	1,00	0,23	0	0,50	0,23	0	0,25
Realschulabschluss	0,71	0,77	0,72	0	0,25	0,72	0	0,25	0,73	0	0,06	0,73	0	0,06
Abitur	0,03	0,10	0,04	0	0,25	0,06	1	0,00	0,04	0	0,50	0,04	0	0,25
eigene Kinder	0,02	0,03	0,05	0	0,08	0,02	0	0,77	0,04	0	0,10	0,03	0	0,33
Elternhaushalt	0,52	0,57	0,53	0	0,58	0,55	0	0,12	0,55	0	0,09	0,55	0	0,15

Fortsetzung siehe nächste Seite

Fortsetzung Tabelle C.3

Ausgangsdaten			Match Replacement			opt. Full Match			opt. 1 : 1 Match			Random Match		
Merkmale	T[a]	NT[a]	C[a]	Test[b]	p-value	C[a]	Test[b]	p-value	C[a]	Test[b]	p-value	C[a]	Test[b]	p-value
Charakteristika der Berufsausbildung														
vertragl. geb. BAB	0,68	0,79	0,69	0	1,00	0,76	1	0,00	0,69	0	0,13	0,74	1	0,00
Zusatzqualifikation	0,17	0,20	0,16	0	0,58	0,19	0	0,15	0,17	0	0,85	0,15	0	0,38
Land-, Forstwirtschaft	0,04	0,03	0,03	0	0,58	0,04	0	1,00	0,02	0	0,23	0,04	0	1,00
Metall-, Elektroberufe	0,11	0,22	0,13	0	0,23	0,12	0	0,29	0,12	0	0,34	0,12	0	0,55
Bau-, Ausbauberufe	0,04	0,09	0,04	0	1,00	0,06	0	0,15	0,05	0	0,25	0,04	0	0,73
sonst. Fertigungsberufe	0,10	0,12	0,11	0	1,00	0,12	0	0,52	0,14	0	0,06	0,14	0	0,06
technische Berufe	0,03	0,04	0,02	0	0,27	0,02	0	0,27	0,02	0	0,61	0,02	0	0,55
Waren-, DL-kaufleute	0,17	0,13	0,15	0	0,58	0,13	0	0,05	0,15	0	0,58	0,16	0	0,71
Org., Verwaltungs, Büro	0,14	0,16	0,15	0	1,00	0,15	0	0,22	0,16	1	0,04	0,14	0	1,50
Gesundheitsdienst	0,13	0,09	0,15	0	0,40	0,13	0	1,00	0,13	0	1,00	0,14	0	0,84
Sozial-, Erziehungsberufe	0,07	0,05	0,08	0	0,82	0,06	0	0,33	0,08	0	0,81	0,08	0	1,00
sonst. Dienstleistungen	0,16	0,08	0,15	0	0,77	0,11	1	0,01	0,12	1	0,02	0,13	0	0,07
allgemeine Lage auf dem Arbeitsmarkt														
Region d. BAB	1,00	1,00	1,00	0	0,51	1,00	0	0,96	1,00	0	0,89	1,00	0	0,10
BAB-Ende 1995-1998	0,00	0,01	0,00	0	1,00	0,01	0	0,25	0,00	0	1,00	0,01	0	0,50
BAB-Ende 1999-2002	0,14	0,51	0,14	0	2,00	0,27	1	0,00	0,15	0	0,25	0,14	0	1,00
BAB-Ende 2003-2006	0,86	0,48	0,86	0	1,00	0,78	1	0,00	0,85	0	0,13	0,85	0	0,25
persönliche Arbeitsmarkt-Vorgeschichte														
keine Unterbrechung	0,36	0,44	0,37	0	0,68	0,38	0	0,30	0,37	0	0,84	0,37	0	0,52
Zivildienst / Bund	0,01	0,02	0,01	0	1,50	0,01	0	1,00	0,01	0	1,50	0,01	0	0,63
Arbeitslosigkeit	0,05	0,04	0,03	0	0,50	0,02	0	0,07	0,03	0	0,50	0,04	0	0,84
Erwerbstätigkeit	0,00	0,01	0,01	0	0,63	0,00	0	1,50	0,01	0	1,00	0,00	0	1,50
abgebr. Ausbildung	0,06	0,03	0,04	0	0,06	0,03	1	0,02	0,04	1	0,03	0,03	0	0,07

Fortsetzung siehe nächste Seite

Fortsetzung Tabelle C.3

Ausgangsdaten			Match Replacement			opt. Full Match			opt. 1 : 1 Match			Random Match		
Merkmale	T[a]	NT[a]	C[a]	Test[b]	p-value	C[a]	Test[b]	p-value	C[a]	Test[b]	p-value	C[a]	Test[b]	p-value
zusätzliche, Indikatoren														
Netzwerk	0,20	0,21	0,20	0	2,00	0,21	0	0,29	0,20	0	2,00	0,19	0	0,69
Anford. (eher) hoch	0,73	0,72	0,72	0	0,93	0,86	1	0,00	0,73	0	1,00	0,73	0	1,00
Belastung (eher) hoch	0,44	0,33	0,46	0	0,57	0,44	0	0,85	0,45	0	1,00	0,44	0	0,85
Motivation	0,47	0,64	0,48	0	0,34	0,58	1	0,00	0,52	1	0,01	0,47	0	1,00
Summe quad. Distanzen					0,94			3,40			1,16			1,49

Anmerkungen: siehe Tabelle C.2.

Tabelle C.4: **Bewertung der Matchingergebnisse für den Vergleich der geförderten Berufsausbildungsgänge**

Ausgangsdaten			Match Replacement			opt. Full Match			opt. 1 : 1 Match			Random Match		
Merkmale	T[a]	NT[a]	C[a]	Test[b]	p-value	C[a]	Test[b]	p-value	C[a]	Test[b]	p-value	C[a]	Test[b]	p-value
soziodemographische Faktoren														
Alter	19,86	19,64	19,54	1	0,00	19,59	1	0,00	19,57	1	0,00	19,60	1	0,01
Mann	0,54	0,39	0,51	0	0,33	0,45	1	0,01	0,45	1	0,01	0,45	1	0,01
kein Schulabschluss	0,04	0,02	0,02	0	0,25	0,04	0	1,00	0,04	0	1,00	0,03	0	0,63
Hauptschulabschluss	0,33	0,23	0,34	0	1,00	0,31	0	0,25	0,31	0	0,25	0,32	0	0,63
Realschulabschluss	0,59	0,72	0,61	0	0,25	0,61	0	0,13	0,61	0	0,13	0,61	0	0,13
Abitur	0,04	0,03	0,03	0	1,00	0,04	0	1,50	0,04	0	1,50	0,04	0	1,50
eigene Kinder	0,02	0,02	0,05	0	0,23	0,03	0	0,73	0,04	0	0,51	0,02	0	1,00
Elternhaushalt	0,58	0,52	0,61	0	0,34	0,60	0	0,72	0,57	0	1,00	0,59	0	0,86
Charakteristika der Berufsausbildung														
vertragl. geb. BAB	0,68	0,68	0,69	0	0,50	0,68	0	2,00	0,68	0	2,00	0,70	0	0,25
Zusatzqualifikation	0,17	0,17	0,14	0	0,29	0,19	0	0,82	0,17	0	0,83	0,16	0	0,84
Land-, Forstwirtschaft	0,01	0,04	0,01	0	1,00	0,06	1	0,01	0,04	0	0,13	0,04	0	0,29
Metall-, Elektroberufe	0,10	0,11	0,11	0	0,75	0,10	0	1,00	0,08	0	0,45	0,11	0	0,77
Bau-, Ausbauberufe	0,13	0,03	0,10	0	0,13	0,06	1	0,00	0,06	1	0,00	0,05	1	0,00
sonst. Fertigungsberufe	0,11	0,11	0,13	0	0,50	0,13	0	0,63	0,13	0	0,63	0,11	0	1,31

Fortsetzung siehe nächste Seite

Fortsetzung Tabelle C.4

Ausgangsdaten			Match Replacement			opt. Full Match			opt. 1 : 1 Match			Random Match		
Merkmale	T[a]	NT[a]	C[a]	Test[b]	p-value	C[a]	Test[b]	p-value	C[a]	Test[b]	p-value	C[a]	Test[b]	p-value
technische Berufe	0,05	0,03	0,01	0	0,07	0,03	0	0,45	0,02	0	0,29	0,02	0	0,34
Waren-, DL-kaufleute	0,11	0,16	0,13	0	0,65	0,17	1	0,04	0,16	0	0,08	0,17	1	0,05
Org., Verwaltungs, Büro	0,20	0,15	0,18	0	0,22	0,17	0	0,07	0,17	0	0,07	0,16	1	0,04
Gesundheitsdienst	0,11	0,13	0,13	0	0,58	0,13	0	0,77	0,12	0	1,00	0,12	0	1,00
Sozial-, Erziehungsberufe	0,04	0,07	0,05	0	0,73	0,06	0	0,34	0,05	0	0,55	0,05	0	0,45
sonst. Dienstleistungen	0,13	0,17	0,15	0	0,65	0,17	0	0,17	0,16	0	0,50	0,14	0	0,81
allgemeine Lage auf dem Arbeitsmarkt														
Region	1,00	1,00	1,00	0	0,82	1,19	1	0,03	1,16	0	0,11	1,15	1	0,03
BAB-ende 1995-1998	0,00	0,00	0,00	0	1,00	0,00	0	1,00	0,00	0	1,00	0,00	0	1,00
BAB-ende 1999-2002	0,36	0,14	0,36	0	2,00	0,25	1	0,00	0,25	1	0,00	0,26	1	0,00
BAB-ende 2003-2006	0,64	0,86	0,64	0	2,00	0,75	1	0,00	0,75	1	0,00	0,74	1	0,00
persönliche Arbeitsmarkt-Vorgeschichte														
keine Unterbrechung	0,20	0,37	0,26	0	0,09	0,35	1	0,00	0,30	1	0,00	0,28	1	0,01
Zivildienst / Bund	0,02	0,00	0,00	0	0,25	0,00	0	0,25	0,00	0	0,25	0,00	0	0,25
Arbeitslosigkeit	0,10	0,04	0,01	1	0,00	0,04	1	0,02	0,02	1	0,00	0,02	1	0,00
Erwerbstätigkeit	0,00	0,00	0,00	0	1,00	0,00	0	1,00	0,00	0	1,00	0,00	0	1,00
abgebr. Ausbildung	0,05	0,04	0,01	1	0,02	0,03	0	0,51	0,02	0	0,34	0,02	0	0,34
zusätzliche Indikatoren														
Netzwerk	0,13	0,19	0,16	0	0,30	0,17	0	0,08	0,14	0	0,58	0,14	0	0,82
Anford. (eher) hoch	0,63	0,74	0,77	1	0,01	0,80	1	0,00	0,73	1	0,05	0,71	0	0,11
Belastung (eher) hoch	0,37	0,44	0,36	0	0,84	0,42	0	0,15	0,40	0	0,46	0,37	0	0,86
Motivation	0,33	0,47	0,38	0	0,17	0,43	1	0,01	0,40	0	0,06	0,39	0	0,09
Summe quad. Distanzen					1,13			1,86			1,68			2,00
ausgeschl. Beob.					2			2			2			2

Anmerkungen: siehe Tabelle C.2.

C.3 Arbeitsmarktstatus nach der Berufsausbildung

Tabelle C.5: **Arbeitsmarktstatus direkt nach der Berufsausbildung – außerbetriebliche Förderung**

Arbeitsmarktstatus	geförderte Jugendliche	ungeförderte Jugendliche
Anzahl Personen	168	141
Erwerbstätigkeit	22,16	34,04
Arbeitslosigkeit	55,69	48,23
Ausbildung / Studium	2,99	2,84
geförderte Beschäftigung[a]	0,00	0,71
sonstiger Status[b]	19,16	14,18

Anmerkungen:
Angaben in Prozent.
[a] Arbeitsbeschaffungs,- Weiterbildungsmaßnahmen, Kurzarbeit;
[b] Nichterwerbstätigkeit, Wehr-, Zivildienst.
Quelle: *Jugendpanel des zsh; eigene Berechnungen.*

Tabelle C.6: **Arbeitsmarktstatus direkt nach der Berufsausbildung – betriebsnahe Förderung**

Arbeitsmarktstatus	geförderte Jugendliche	ungeförderte Jugendliche
Anzahl Personen	324	254
Erwerbstätigkeit	27,47	44,49
Arbeitslosigkeit	52,16	40,16
Ausbildung / Studium	5,86	5,12
geförderte Beschäftigung[a]	0,93	0,79
sonstiger Status[b]	13,58	9,45

Anmerkungen: siehe Tabelle C.5.
Quelle: *Jugendpanel des zsh; eigene Berechnungen.*

Tabelle C.7: **Arbeitsmarktstatus direkt nach der Berufsausbildung – Vergleich der Förderungen**

Arbeitsmarktstatus	außerbetriebliche Ausbildung	betriebsnahe Ausbildung
Anzahl Personen	166	106
Erwerbstätigkeit	22,42	22,64
Arbeitslosigkeit	55,76	50,94
Ausbildung / Studium	3,03	6,60
geförderte Beschäftigung[a]	0,00	0,94
sonstiger Status[b]	18,79	18,87

Anmerkungen: siehe Tabelle C.5.
Quelle: *Jugendpanel des zsh; eigene Berechnungen.*

Eidesstattliche Erklärung

Ich versichere, dass ich meine Dissertation mit dem Titel

Matching kleiner Stichproben. Ein Vergleich verschiedener Verfahren

selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel benutzt habe. Die Stellen, die anderen Werken - dazu zählen auch Internetquellen - dem Wortlaut oder dem Sinn nach entnommen sind, wurden unter Angabe der Quelle als Entlehnung kenntlich gemacht.

Halle (Saale), Dezember 2008

Eva Reinowski