DIW Berlin

German Institute
for Economic Research

Discussion Papers

401

Markus Gangl
Thomas A. DiPrete

Kausalanalyse durch Matchingverfahren

Berlin, February 2004

Opinions expressed in this paper are those of the author and do not necessarily reflect views of the Institute.

DIW Berlin

German Institute
for Economic Research

Königin-Luise-Str. 5
14195 Berlin,
Germany

Phone +49-30-897 89-0
Fax +49-30-897 89-200

www.diw.de

ISSN 1619-4535

Kausalanalyse durch Matchingverfahren*

Markus Gangl
Wissenschaftszentrum Berlin für Sozialforschung (WZB)
Reichpietschufer 50, 10785 Berlin
Tel. (030) 25491-141; Fax (030) 25491-222
Email gangl@wz-berlin.de

und

Thomas A. DiPrete
Forschungsprofessor des DIW Berlin
Department of Sociology, Duke University
Box 90088, Durham NC 27708-0088, United States
Tel. +1.919.660.5612 / 5614; Fax. +1.919.660.5623
Email tdiprete@soc.duke.edu

Erste Fassung, 16. Februar 2004
Kommentare willkommen

WORD COUNT: 56.700 Zeichen (mit Leerzeichen), 1 Abbildung, 3 Tabellen

* Der Beitrag entstand im Rahmen des DFG-Projekts ‚Human Capital Effects of the Welfare State'. Ein Gastaufenthalt von Thomas DiPrete am Wissenschaftszentrum Berlin ermöglichte uns den Beginn der Arbeiten. Die im Beitrag verwendeten Daten des Sozio-ökonomischen Panels wurden freundlicherweise durch das Deutsche Institut für Wirtschaftsforschung (DIW), Berlin, zur Verfügung gestellt. Selbstverständlich liegt die Verantwortung für die hier durchgeführten Analysen allein bei den Autoren. Wir bedanken uns bei Ulrich Kohler für hilfreiche Anregungen und Kommentare.

Kausalanalyse durch Matchingverfahren

Zusammenfassung
Aufgrund ihrer Nähe zum Konzept kontrafaktischer Kausalität haben nichtparametrische Matchingverfahren in der neueren statistischen und ökonometrischen Literatur zur Kausalanalyse an Bedeutung gewonnen. Vor diesem Hintergrund führt der Beitrag das Rubin Causal Model (RCM) in die soziologische Methodendiskussion ein und diskutiert seine empirische Umsetzung im Rahmen des Propensity Score Matchings. Der Beitrag verdeutlicht die Relevanz dieser Verfahren für soziologische Fragestellungen sowie die ihnen im Vergleich zu üblichen Regressionsverfahren zugrundeliegenden Annahmen. Wir illustrieren die Anwendung von Matchingverfahren anhand einer Analyse des kausalen Effekts von Arbeitslosigkeit auf den weiteren Erwerbsverlauf.

Keywords
Matching, Kausalanalyse, Nichtparametrische Verfahren, Beobachtungsdaten, Rubin Causal Model, Kontrafaktische Analyse

Matching methods for the causal analysis of observational data

Abstract
Having close linkages with the counterfactual concept of causality, nonparametric matching estimators have recently gained in popularity in the statistical and econometric literature on causal analysis. Introducing key concepts of the Rubin causal model (RCM), the paper discusses the implementation of counterfactual analyses by propensity score matching methods. We emphasize the suitability of the counterfactual framework for sociological questions as well as the assumptions underlying matching methods relative to standard regression analysis. We then illustrate the application of matching estimators in an analysis of the causal effect of unemployment on workers' subsequent careers.

Keywords
Matching, causality, nonparametric estimators, observational data, Rubin causal model, counterfactual analysis

1 Einleitung

Die Analyse kausaler Zusammenhänge ist ohne Zweifel das Herzstück einer theoriegeleiteten, hypothesenprüfenden und kumulativen Sozialwissenschaft, in der sich theoretische Modellbildung und deren empirische Überprüfung verbinden: aus einem Modell abgeleitete Hypothesen oder gar konkurrierende Erklärungen können letztlich nur dadurch auf ihren Realitätsgehalt getestet werden, dass sie mit empirischen Daten konfrontiert werden, die potentiell falsifizierende Aussagen über eine unterstellte Beziehung zwischen Ursache und Wirkung ermöglichen. Dass Kausalschlüssen eine zentrale Rolle für die Verzahnung empirischer und theoretischer Forschung zukommt, ist eine alles andere als neue Erkenntnis.

In den Sozialwissenschaften stellt sich dabei allerdings das zentrale Problem, dass experimentell erhobene Daten, die eine unmittelbar kausale Interpretation empirischer Befunde erlauben, in der Forschungspraxis bestenfalls eine marginale Rolle zukommt. Abgesehen von speziellen Anwendungen wie etwa im Bereich der experimentellen Spieltheorie (z.B. Fehr und Gächter 2000; Falk und Fehr 2003), stoßen experimentelle Ansätze in aller Regel sehr schnell an ethische, forschungspraktische und methodologische Grenzen (vgl. Manski 1995; Heckman 1992). Nichtexperimentelle Daten haben ihrerseits den wichtigen Vorteil, dass mit ihnen das reale Verhalten von Personen sowie deren wichtigste Merkmale (repräsentativ) erfasst werden. Andererseits führt gerade die Beobachtung realen Verhaltens dazu, dass kausale Schlussfolgerungen über Ursache-Wirkungs-Zusammenhänge nicht uneingeschränkt möglich sind. In Beobachtungsdaten resultieren in aller Regel eben nicht nur die abhängigen, sondern vor allem auch die unabhängigen Variablen der Analyse aus intentionalem Handeln von Akteuren, und können damit eben nicht umstandslos als exogen gegeben angesehen werden.

Man kann vor diesem Hintergrund geneigt sein, den Schluss zu ziehen, dass ein eher naturwissenschaftlich inspiriertes Verständnis von Kausalität als Ursache-Wirkungs-

Zusammenhang dem Gegenstandsbereich der Soziologie – außer in Randbereichen wie etwa der Evaluationsforschung – grundsätzlich nur wenig angemessen ist, und die empirische Sozialforschung sich daher auf im wesentlichen beschreibende Analysen zu beschränken habe, die gegebenenfalls durch akteurszentrierte ‚causal narratives' unterfüttert werden können (vgl. Goldthorpe 2001). Wir werden im Folgenden dagegen die Position einnehmen, dass Kausalschlüsse auch in den Sozialwissenschaften nicht nur möglich und angemessen sind, sondern letzten Endes das wesentliche Ziel hypothesen- und mechanismenprüfender empirischer Sozialforschung darstellen. In unserem Verständnis beziehen sich Kausalschlüsse immer auf die empirische Überprüfung theoretisch erwarteter Ursache-Wirkungs-Zusammenhänge, bilden also mithin das zentrale Verbindungsglied zwischen theoretischer Modellbildung und theoriegeleiteter empirischer Sozialforschung.

Dass Kausalschlüsse innerhalb der Sozialwissenschaften auf besondere Schwierigkeiten stoßen, liegt dagegen angesichts ihres Gegenstandsbereichs auf der Hand: sozialwissenschaftliche Kausalanalysen können ausschließlich vor dem Hintergrund intentionalen Handelns von Akteuren durchgeführt werden, und beziehen sich auf die empirisch beobachtbaren Reaktionen der handelnden und interagierenden Akteure auf exogen variierende Rahmenbedingungen. Da letztere durch Beobachtungsdaten nur bedingt erfasst werden (können), beruhen kausale Schlussfolgerungen notwendigerweise immer auf einer (geglückten) Kombination von Untersuchungsdesign und statistischer Methodologie, die sich einer Experimentalsituation möglichst weit annähert. Innerhalb der statistischen und ökonometrischen Literatur hat sich in den vergangenen 15 Jahren ein Konsens zu den daraus für die statistische Analyse folgenden Implikationen herausgebildet, das so genannte Rubin Causal Model (RCM). Wir werden im folgenden dieses kontrafaktische Verständnis von Kausalität ausführlicher darstellen und verschiedene potenziell interessante und empirisch schätzbare kausale Parameter definieren.

Da sie eine direkte statistische Umsetzung des RCM darstellen, werden wir im Anschluß daran die wesentlichen Grundzüge der Propensity score matching (PSM)-Verfahren zur empirischen Analyse kausaler Effekte mit Hilfe von Beobachtungsdaten vorstellen. Matchingverfahren ermöglichen die direkte Schätzung der wichtigsten kausalen Parameter, und haben entsprechend der zunehmenden Akzeptanz des RCM in den letzten Jahren in der angewandten Statistik und in ökonometrischen Arbeiten deutlich an Bedeutung gewonnen. Im Unterschied zu üblichen Regressionsverfahren handelt es sich dabei zudem um nichtparametrische Verfahren, deren Anwendung keine oder nur vergleichsweise wenig restriktive statistische Annahmen voraussetzt. Neben einem Präzisierung der Voraussetzungen kausaler Schlußfolgerungen auf der Basis sozialwissenschaftlicher Beobachtungsdaten ist damit in den letzten Jahren auch eine alternative statistische Tradition entwickelt worden, die in der angewandten Forschung zumindest Sensitivitätsanalysen der mit Regressionsverfahren gewonnenen Ergebnisse ermöglicht. Nach der Darstellung ihrer wesentlichen Grundlagen werden wir die empirische Anwendung von Matchingverfahren am Beispiel einer Analyse der Karrierefolgen von Arbeitslosigkeit illustrieren.

2 Kontrafaktische Inferenz im Rubin Causal Model

Das kontrafaktische Verständnis von Kausalität wie es im Rubin Causal Model (RCM) zum Ausdruck kommt, ist die gemeinsame methodologische Grundlage aller Matchingverfahren (vgl. Rubin 1974, 1978; Holland 1986; Pratt und Schlaifer 1988; Manski 1995; Heckman 1997; Rosenbaum 2002; aus spezifischer soziologischem Blickwinkel Sobel 1995, 1996; Smith 1997; Winship und Morgan 1999; Winship und Sobel 2004). Gedanklicher Ausgangspunkt des RCM ist die Modellvorstellung potentieller Ereignisse Y, die jeweils

abhängig vom Auftreten eines kausal wirksamen Faktors *T* eintreten würden.[1] Für das einfachste Beispiel eines binären Faktors *T* ist der kausale Effekt dann nichts anderes als der Unterschied zwischen dem Ereignis $Y_{1i} = Y \mid T=1$, das bei Auftreten von *T* realisiert wird, und dem alternativen Ereignis $Y_{0i} = Y \mid T=0$. Vor diesem Hintergrund ist es dann möglich, den kausalen Effekt von *T* als Einheitseffekt (unit effect)

$$\delta_i = Y_i(X_i, T=1) - Y_i(X_i, T=0) = Y_{1i} - Y_{0i} \qquad (1)$$

auf der Ebene individueller Akteure mit spezifischen Merkmalen X_i, zu definieren.[2] Im unten ausführlicher dargestellten empirischen Beispiel entspräche Y_{1i} dem individuellen Erwerbsverlauf nach Eintritt einer Arbeitslosigkeit zu einem bestimmten Zeitpunkt T_0, Y_{0i} dagegen dem individuellen Erwerbsverlauf, in welchem zum Zeitpunkt T_0 keine Arbeitslosigkeit aufgetreten war. Der kausale Effekt der Arbeitslosigkeit besteht in der Differenz zwischen Y_{1i} und Y_{0i}, also dem auf die Arbeitslosigkeit zurückführbaren Unterschied im individuellen Erwerbsverlauf – gemessen etwa anhand des realisierten Erwerbseinkommens, der beruflichen Stellung oder der individuellen Arbeitsplatzsicherheit. Konzeptionell sind die Einheitseffekte δ_i zudem personenspezifisch; inwieweit die kausale

---

[1] Dementsprechend firmiert das RCM in der statistischen Literatur auch als potential outcome model. In der ökonometrischen Literatur sind auch die Bezeichnungen Roy model oder Switching (regression) model üblich (Manski 1995; Heckman et al. 1997, 1998), um denselben Sachverhalt anzudeuten. In der Soziologie wird häufig die Bezeichnung des interventionistischen bzw. manipulativen Kausalmodells gewählt, um die gedankliche Nähe zur experimentellen Forschung zu betonen (etwa Goldthorpe 2001). Leider entsteht dadurch oft auch der (falsche) Eindruck, der konzeptionelle Rahmen des RCM sei ausschließlich im experimentellen Bereich bzw. in der Evaluation expliziter Interventionen relevant.

[2] Wenn auch ursprünglich aus der experimentellen Forschung stammend, ist das RCM konzeptionell nicht auf die Wirkung binärer Faktoren beschränkt (vgl. die ausführlichere Darstellung zur Verbindung mit üblichen Regressionsverfahren bei Wooldridge 2002 oder Berk 2004 bzw. Strukturgleichungsmodellen bei Pearl 1998, 2000). Zur vereinfachenden Illustration verwenden wir im Beitrag jedoch durchgehend das binäre Ereignis Arbeitslosigkeit als potentiell kausal wirksamen Faktor. Ordinale Kausalfaktoren werden im Rahmen des Standardansatzes in multiple Paarvergleiche der Wirkung jeweils binärer Faktoren aufgelöst.

Wirkung des Faktors *T* tatsächlich individuell verschieden ausfällt, ist dann im wesentlichen eine empirische Frage.

2.1 Das Fundamentalproblem der Kausalanalyse

Die Definition des Einheitseffektes von *T* als Differenz potentieller Ereignisse *Y* ist einerseits unmittelbar intuitiv und entspricht andererseits auch gängigen Kausalitätskriterien (etwa Marini und Singer 1988; Pearl 2000): einmal ist das Auftreten von *T* dem Ereignis *Y* zeitlich vorgelagert, und zum zweiten besteht eine Kovariation zwischen *T* und *Y*, die nicht auf die Wirkung alternativer (vorgelagerter) Kovariaten *X* zurückführbar ist. Im Zentrum der kontrafaktischen Interpretation steht allerdings drittens, dass der oben definierte Einheitseffekt zudem der Wirkung einer bewussten Manipulation von bzw. der Intervention *T* auf das Ereignis *Y* entspricht. Kausale Effekte beschreiben damit Reaktionsfunktionen (response schedules in der Terminologie von Pearl 2000; Berk 2004) im Sinne der hypothetisch (kontrafaktisch) zu erwarteten Ereignisse infolge einer (wie auch immer praktisch durchführbaren) Setzung der Umstände $T=t$.

Aus der Definition kausaler Effekte über den Vergleich potentieller Ereignisse folgt allerdings auch unmittelbar, dass kausale Wirkungen nicht direkt beobachtbar, sondern immer nur unter bestimmten Annahmen empirisch erschlossen werden können. Das fundamentale Problem jedweden Kausalschlusses besteht in der Unmöglichkeit, dieselbe Person unter sonst gleichen Bedingungen gleichzeitig alternativen Rahmenbedingungen ausgesetzt zu sehen, und damit ihre empirische Reaktionen auf unterschiedliche Handlungskontexte direkt erfassen zu können. Einheitseffekte könnten nur dann umstandslos empirisch geschätzt werden, wenn die Erwerbsverläufe *Y* von Personen *i* tatsächlich sowohl nach einer Arbeitslosigkeit *T* zum Zeitpunkt T_0 als auch ohne eine Arbeitslosigkeit *T* zum Zeitpunkt T_0 beobachtbar wären – der Einheitseffekt δ_i ergäbe sich unmittelbar aus der Differenz der Zellen (a) und (b) in Abbildung 1.

- Abbildung 1 etwa hier -

Tatsächlich beobachtbar sind aber natürlich nur die Erwerbsverläufe $Y_{1i,i\in E} \mid T=1$, die empirischen Erwerbsverläufe von zum Zeitpunkt T_0 tatsächlich arbeitslos gewordenen Personen (Zelle (a)), und $Y_{0i,i\in C} \mid T=0$, also die empirischen Erwerbsverläufe von Personen, die zu einem bestimmten Zeitpunkt T_0 nicht arbeitslos geworden sind (Zelle (d)). Alle Methoden empirischer Kausalanalyse haben Annahmen darüber zu treffen, auf welche Weise aus den empirischen vorliegenden Beobachtungen $Y_{0i,i\in C} \mid T=0$ in Zelle (d) eine kontrafaktische Abschätzung des latenten Ereignisses $\{Y_{0i,i\in E} \mid T=0\}$, also des alternativen Erwerbsverlaufs von *i*, der ohne Arbeitslosigkeit beobachtet worden wäre (Zelle (b)), konstruiert werden kann, und dementsprechend daraus der kausale Effekt $\delta_i = Y_{1i,i\in E} \mid T=1 - \{Y_{0i,i\in E} \mid T=0\}$ empirisch zu bestimmen ist.

2.2 Grundlegende Annahmen der empirischen Analyse

Inwiefern der so vorgenommene Vergleich empirisch die kausale Wirkung von *T* auf *Y* ermittelt, hängt letzten Endes davon ab, ob drei zentrale Annahmen als erfüllt angesehen werden können. Zur Illustration dieser Annahmen genügt es, sich zu verdeutlichen, dass der beobachtbare Unterschied zwischen den Ereignissen $Y_{1i,i\in T} \mid T=1$ und $Y_{0i,i\in C} \mid T=0$ immer in die drei Komponenten

$$Y_{1i,i\in E} - Y_{0i,i\in C} = \delta T + \left(Y_{0i,i\in E} - Y_{0i,i\in C}\right) + \left(\delta_{i\in E} - \delta_{i\in C}\right) \qquad (2)$$

Unterschied zwischen Experimental- und Kontrollperson = „wahrer" kausaler Effekt δ des Faktors *T* + Unterschied im Basisereignis *Y* ohne Wirkung von *T* + Unterschied in der Wirkung von *T* zwischen Experimental- und Kontrollperson

zerlegt werden kann. Die erste Komponente in Gleichung (2), der „wahre" kausale Effekt von *T*, basiert auf der Annahme der Existenz einer stabilen kausalen Wirkung von *T*. Ohne diese

so genannte stable unit treatment value assumption (SUTVA) sind kausale Wirkungen nicht allein mit Daten auf derselben Beobachtungs- und Wirkungsebene identifizierbar. Sie ist immer dann nicht erfüllt, wenn Interdependenzen zwischen Experimental- und Kontrollgruppe, Spillover- oder allgemeine Gleichgewichtseffekte vorliegen, so dass die für die Kontrollgruppe beobachteten Ereignisse nicht unbeeinflusst von den sozialen Prozessen in der Experimentalgruppe sind (vgl. ausführlicher Garfinkel et al. 1992; Heckman, Lochner und Taber 1998). Kausalanalysen sowohl auf der Basis von Regressions- wie auch Matchingverfahren setzen zumindest implizit in aller Regel die SUTVA-Annahme voraus.

Im Kern empirischer Analysen steht dagegen zumeist die Frage, inwieweit die über SUTVA angenommene Wirkung δ tatsächlich durch den gewählten Untersuchungsansatz identifiziert wird, der auf jeweils spezifische Art und Weise für die Auswirkungen der zweiten und dritten Komponente statistisch bzw. designbasiert kontrolliert. Der mittlere Term in Gleichung (2), der Unterschied in den Basisereignissen ohne Intervention, betrifft dabei vor allem das Problem des Vergleichs vergleichbarer Einheiten bzw. die einem Kausalschluss zugrundeliegende Annahme der unit homogeneity. Es ist offensichtlich, dass δ nur dann konsistent geschätzt werden kann, wenn (typischerweise mithilfe von Kovariaten) die Vergleichbarkeit der „sonstigen Bedingungen“ in den Vergleichsgruppen hergestellt wird.

Zum zweiten ist in nichtexperimentellen Daten davon auszugehen, dass die Zuweisung in die Experimentalgruppe selektiv erfolgt, etwa weil sich eine bestimmte Intervention nur an eine spezifische (besonders vielversprechende) Zielgruppe wendet, oder weil sich Akteure aufgrund privater (unbeobachteter) Informationen für oder gegen eine Teilnahme an T entscheiden. Diesen Umständen trägt der dritte Term in Gleichung (2) Rechnung, der einen potentiellen Unterschied zwischen der Wirkung von T in der ausgewählten Experimentalstichprobe und in der Kontrollgruppe unterstellt. In der empirischen Analyse kann δ nur dann konsistent geschätzt werden, wenn dieser letzte Term entfällt, also die Zuweisung zum Experimentalstatus (unter Kontrolle von Kovariaten) konditional unabhängig

vom Ergebnis Y ist (conditional independence assumption (CIA), nach Rosenbaum und Rubin 1983 auch als strict ignorability assumption bezeichnet).[3]

2.3 Durchschnittliche kausale Effekte als empirische Schätzgrößen

Auch wenn sie die Grundlage des kontrafaktischen Verständnisses von Kausalität darstellen, können Einheitseffekte – abgesehen von Einzelfällen – weder plausibel geschätzt werden, noch stehen sie typischerweise im Zentrum des sozialwissenschaftlichen Interesses. In aller Regel ist die empirische Forschung implizit oder explizit vor allem an der Abschätzung des typischen, d.h. durchschnittlichen kausalen Effekts von *T* auf *Y* interessiert. Entsprechend der Dekomposition der Gruppenunterschiede in Gleichung (2) werden in der neueren ökonometrischen Literatur (vgl. Heckman und Vytlacil 2002; Heckman et al. 2003) im wesentlichen zwei zentrale Parameter betrachtet: einmal der durchschnittliche kausale Effekt von *T* in der untersuchten Gesamtpopulation, der average treatment effect (ATE), und zum zweiten der durchschnittliche kausale Effekt für die Personen, die empirisch der Wirkung des kausalen Faktors ausgesetzt waren, der average treatment effect on the treated (ATT).[4]

Der ATE entspricht dabei konzeptionell dem erwarteten Effekt von *T* für *Y* einer zufällig aus der Gesamtpopulation (d.h. Stichproben *E* und *C* in Abbildung 1) gezogenen Person, der ATT dagegen dem im Durchschnitt beobachteten Effekt von *T* für *Y* für eine zufällig aus der Experimentalstichprobe *E* gezogenen Person. Der ATE erfasst konzeptionell damit die typischen Folgen von *T* in der untersuchten Population, der ATT die typischen

---

[3] In der Literatur findet sich teilweise eine Gleichsetzung der beiden Annahmen mit den Begrifflichkeiten der „selection on observables" bzw. „selection on unobservables" (Heckman und Robb 1985). Auch wenn sie im konkreten Einzelfall sicher nicht trennscharf sind, werden wir heuristisch ebenfalls von prinzipiell beobachtbaren Hintergrundkovariaten und prinzipiell unbeobachtbarem privatem Wissen ausgehen.

[4] Ein weiterer zentraler Parameter in der ökonometrischen Diskussion ist der so genannte local average treatment effect (LATE, vgl. Imbens und Angrist 1994; Angrist, Imbens und Rubin 1996), der jeweils im Hinblick auf die Wirkung einer Instrumentalvariable definiert ist. Im Zusammenhang mit Matchingverfahren spielt dieser allerdings eine bestenfalls marginale Rolle, und wird von uns im Folgenden vernachlässigt.

Folgen von *T* für die (möglicherweise selektive) Gruppe der tatsächlich von *T* Betroffenen. Die aus der experimentellen Forschung abgeleitete Terminologie des kontrafaktischen Ansatzes sollte insgesamt aber nicht darüber hinwegtäuschen, dass „treatment" hier nicht primär im Sinne einer expliziten Manipulation, sondern allgemeiner als kausal wirksamer Faktor zu verstehen ist.[5] Zugunsten der Konsistenz mit der Literatur folgen wir hier durchgängig den inzwischen eingebürgerten Begrifflichkeiten.

ATE- und ATT-Parameter sind direkt aus den Grundannahmen des RCM definierbar (vgl. zum Folgenden Heckman 1997). Ausgangspunkt ist das folgende Modell potentieller Ergebnisse

$$\begin{aligned} Y_{0i} &= \mu_0(X_i) + U_{0i} \\ Y_{1i} &= \mu_1(X_i) + U_{1i} \end{aligned} \qquad (3)$$

das die potentiellen Ergebnisse Y_0 und Y_1 in Abhängigkeit von beobachteten Variablen *X* und unbeobachteten Größen *U* beschreibt. Empirisch beobachtet wird jeweils nur eines der beiden potentiellen Ereignisse *Y*, Y_{0i} im Fall, dass *i* dem kausalen Faktor *T* nicht ausgesetzt war (T_i=0), und Y_{1i} für den Fall, dass i der Wirkung von T ausgesetzt war (T_i=1). Die Wirkung der Kovariaten *X* wirkt sich vermittelt über die nichtparametrische Funktion $\mu(.)$ auf *Y* aus, wobei

---

[5] Aus Hollands (1986) Position, dass Kausalfaktoren prinzipiell manipulierbar sein müssten, hat sich in der soziologischen Rezeption der Konsens gebildet, das kontrafaktische Konzept sei sozialwissenschaftlichen Fragestellungen prinzipiell unangemessen, da in den Sozialwissenschaften die Frage nach den Auswirkungen von unveränderlichen Attributen – etwa Geschlecht, soziale Herkunft oder ethnische Zugehörigkeit – im Unterschied zu den Naturwissenschaften konstituierend sei. Wir teilen weder Hollands spezifische Position noch deren Rezeption in der Soziologie, sondern verstehen Kausaleffekte vielmehr in einem weiteren Sinne als Auswirkung exogen variierender Randbedingungen auf soziales Handeln. In diesem Sinn ist es vollkommen legitim, von einem kausalen Effekt des (individuell nicht beeinflussbaren) Geschlechts etwa auf individuelle Lebensverläufe zu sprechen und diese empirisch zu ermitteln (vgl. etwa Sobel 1998). In substanziellen Beiträgen steht bei genauerer Betrachtung in aller Regel jedoch gar nicht die Ermittlung (oft unbestrittener) kausaler Geschlechtereffekte per sé, sondern vielmehr der Nachweis der zugrundeliegenden Mechanismen im Mittelpunkt. Aus dem Verständnis des RCM folgt unmittelbar, dass dieser Nachweis nur über ein

die Wirkung sowohl der beobachteten Kovariaten *X* wie auch der unbeobachteten Einflußfaktoren *U* prinzipiell von T abhängig sein kann (d.h. im allgemeinen gilt $\mu_0(.) \neq \mu_1(.)$ und $U_0 \neq U_1$).

Unter diesen Annahmen können die beiden Parameter wie folgt definiert werden: zunächst der bedingte average treatment effect (ATE_X) als Differenz der alternativen Ereignisse *Y* an der Kovariatenstelle *X* als

$$ATE_X = E(\delta_i \mid X) = \mu_1(X) - \mu_0(X), \tag{4}$$

und daraus abgeleitet der average treatment effect (ATE) als Durchschnitt der bedingten average treatment effects ATE_X über die gesamte Verteilung der Kovariaten *X* als

$$ATE = E(\delta_i) = \int_X [\mu_1(X) - \mu_0(X)]\, dF(X) \approx \sum_X ATE_X \; f(X). \tag{5}$$

Äquivalent dazu ist der bedingte average treatment effect on the treated (ATT_X) an der Kovariatenstelle *X* definiert als

$$ATT_X = E(\delta_i \mid X, T=1) = \mu_1(X) - \mu_0(X) + E(U_1 - U_0 \mid X, T=1), \tag{6}$$

und daraus abgeleitet der average treatment effect on the treated (ATT) wiederum als Durchschnitt aller bedingten ATT_X über die gesamte Verteilung der Kovariaten *X* als

$$\begin{aligned} ATT = E(\delta_i \mid T=1) &= \int_X [\mu_1(X) - \mu_0(X) + E(U_1 - U_0 \mid X, T=1)]\, dF(X) \\ &\approx \sum_X ATT_X \; f(X). \end{aligned} \tag{7}$$

Danach unterscheiden sich ATE und ATT konzeptionell allein durch den Term $E(U_1 - U_0 \mid X, T=1)$, der für die Experimentalgruppe *E* die Auswirkung ungemessener Faktoren auf die Ergebnisvariable *Y* angibt. Der Term umfasst beispielsweise die mögliche Wirkung

---

Untersuchungsdesign zu führen ist, dass in einer nachvollziehbaren Weise exogene Variation in den vermuteten

privaten Wissens der Akteure, aber auch die Wirkung sonstiger, in der Analyse nicht beobachteter Faktoren, welche die empirische Zuweisung der Akteure *i* in die Experimentalgruppe *E* bestimmen. Immer wenn angenommen werden muß, dass solche unbeobachteten Einflussfaktoren relevant sind, liegt eine endogene Zuweisung zur Testgruppe vor (das so genannte Selbstselektionsproblem bzw. „selection on unobservables“ nach Heckman und Robb 1985; „hidden bias“ in der Terminologie von Rosenbaum 2002). Da in diesem Fall $ATT_X \neq ATE_X$, ist die empirische Abschätzung des kausalen Effekts von *T* in aller Regel – d.h. ohne zusätzliche Informationen – nicht möglich.

Ganz allgemein ergibt sich aus diesen Überlegungen, dass sich der Parameter ATT_X empirisch immer dann identifizieren lässt, wenn das Problem des Vergleichs homogener Einheiten lokal (d.h. für einen spezifischen Kovariatenvektor *X*) gelöst, und zusätzlich angenommen werden kann, dass die Zuweisung in die Kontrollgruppe *C* konditional unabhängig vom Ereignis *Y* ist (also $E(U_0)=0$ gilt; vgl. Heckman et al. 1998, 2003). Der Populationsparameter ATT kann im Anschluss daran immer dann identifiziert werden, wenn ATT_X über die gesamte Verteilung von *X* schätzbar ist. Auf die entsprechenden (und inhaltlich oft eigentlich interessierenden) ATE-Parameter kann schließlich nur dann generalisiert werden, wenn das Selbstselektionsproblem sowohl für die Experimentalgruppe *E* als auch die Kontrollgruppe *C* als gelöst betrachtet werden kann.

Durch die randomisierte Zuweisung von Test- und Kontrollbedingung sind beide Bedingungen in einer Experimentalsituation unmittelbar erfüllt (vgl. Rosenbaum 2002): die Zufallsauswahl der Testgruppe aus der Stichprobe der Versuchspersonen garantiert zum einen im Aggregat die Vergleichbarkeit von Experimental- und Kontrollgruppe anhand ihrer Merkmale *X*, und zum anderen ist eine endogene Zuweisung zu einer der beiden Vergleichsgruppen von vorneherein ausgeschlossen. Die experimentelle Versuchsanordnung

---

Mechanismen erfasst, und dadurch die Identifikation der (jeweils spezifischen) kausalen Effekte ermöglicht.

identifiziert dann durch einen einfachen Mittelwertvergleich sowohl ATT und – da bei exogener Zuweisung ATT=ATE gilt – auch den ATE für den manipulierten Faktor, und erlaubt, je nach Stichprobengröße, auch die Schätzung subgruppenspezifischer ATEs bzw. ATTs.

Dem partiellen Regressionskoeffizient β_T kommt eine vergleichbare Rolle im Rahmen parametrischer Regressionsmodelle zu. ATE und ATT-Parameter sind jedoch nur dann durch β_T identifiziert, wenn die zugrundeliegenden statistischen Annahmen des Regressionsmodells erfüllt sind. Für die ATE-Parameter trifft dies insbesondere nur dann zu, wenn der Faktor *T* unter Kontrolle der Kovariaten *X* exogen, d.h. konditional unabhängig vom Fehlerterm des Regressionsmodells ist, das Selbstselektionsproblem also durch geeignete Kovariatenauswahl gelöst werden kann (Heckman und Robb 1985).[6] Zudem hängt die Schätzung sowohl der ATT als auch der ATE-Parameter von der korrekten Modellspezifikation ab, also insbesondere davon, dass die Annahmen über die funktionale Form des Modells, mit deren Hilfe v.a. in sparsam besetzten Datenbereichen kontrafaktisch extrapoliert wird, den datengenerierenden Mechanismus zuverlässig abbilden (Wooldridge 2002; Berk 2004).[7] Soweit diese relativ weitreichenden Annahmen als erfüllt angesehen werden, können die bedingten ATT_X- bzw. ATE_X-Parameter durch geeignete Interaktionseffekte geschätzt werden.

Nichtparametrische Matchingverfahren sind vor diesem Hintergrund von Interesse, weil sie zur Schätzung kausaler Effekte nicht auf die Spezifikation eines mathematischen

---

[6] Wir gehen hier implizit von einem Eingleichungsmodell, d.h. einer üblichen OLS- oder auch logistischen Regression aus. Für explizite parametrische und nichtparametrische Modellierungen des Selbstselektionsproblems vgl. etwa Manski (1989), Heckman und Robb (1985), Vella und Verbeek (1992), Brüderl (2001) sowie Pötter (2004).

[7] Annahmen über die funktionale Form schließen aber auch inhaltlich weitergehende Annahmen wie etwa die Annahme unkorrelierter unbeobachteter Heterogenität (unkorrelierter Fehlervarianzen) über die Ebenen eines Mehrebenenmodells ein (vgl. die ausführlichere Darstellung in Engel 2004).

Modells rekurrieren, und durch entsprechende Spezifikationsfehler verzerrte Schätzungen mithin vermeiden. Der Verzicht auf parametrische Annahmen wird allerdings dadurch erkauft, dass Matchingverfahren kausale Wirkungen eines Faktors *T* nur innerhalb des Kovariatenbereichs *S* abschätzen können, in welchem sowohl Personen der Experimental- wie auch der Kontrollgruppe empirisch beobachtet werden. Für diesen Bereich des common support $X \in S$ identifizieren Matchingverfahren den bedingten ATT'$_X$ innerhalb von *S* entsprechend Gleichung (6), und den ATT' als

$$ATT' = E(\delta_i \mid X \in S, T = 1) = \frac{\int_{X \in S} [\mu_1(X) - \mu_0(X) + E(U_1 - U_0 \mid X, T = 1)] \, dF(X \mid T = 1)}{\int_{X \in S} dF(X \mid T = 1)} \quad (8)$$

(vgl. Heckman et al. 1997, 1998).

Da sie sich auf konzeptionell unterschiedliche Parameter beziehen, werden regressions- bzw. matchingbasierte Schätzwerte des ATT$^{(')}$ typischerweise unterschiedlich ausfallen. Eine Konvergenz von regressionsbasiertem ATT und matchingbasiertem ATT' ist immer nur dann zu erwarten, wenn der Bereich des common support praktisch die gesamte Matrix der empirischen Kovariatenkombinationen umfasst und die parametrische Spezifikation des Regressionsmodells empirisch (näherungsweise) korrekt ist, oder wenn die „wahre" parametrische Modellspezifikation so ist, dass sie aus den innerhalb des common support zur Verfügung stehenden Daten für die Gesamtpopulation korrekt geschätzt werden kann. Zur Identifikation der entsprechenden ATE$^{(')}$-Parameter muss wie in parametrischen Verfahren die (unter Kontrolle von Kovariaten) konditional unabhängige Zuweisung von *T* angenommen werden (vgl. Rubin 1974, 1978; Rosenbaum und Rubin 1983; Rosenbaum 2002; Heckman et al. 1998).

3 Kausalanalyse durch Propensity Score Matching

Der Verzicht auf parametrische Modellannahmen und die dadurch vergleichsweise größere Robustheit der empirischen Schätzergebnisse ist eine wichtige Eigenschaft von Matchingverfahren, die ihre Anwendung anstelle von oder zumindest zusätzlich zu üblichen Regressionsverfahren rechtfertigt. Ein zweiter Vorzug der Matchingverfahren liegt in ihrer konzeptionellen Klarheit aufgrund der methodologischen Nähe zum Rubin Causal Model: dementsprechend lassen sich die empirischen Ergebnisse einer Matchinganalyse unmittelbar zu den oben definierten zentralen Parametern des RCM in Beziehung setzen. Zudem sind auch die weitergehenden inhaltlichen Annahmen klar spezifiziert, die jeweils zur Identifikation bestimmter RCM-Parameter erfüllt sein müssen. Ein weiterer praktischer Vorzug der Matchingverfahren ist, dass sie als nichtparametrisches Verfahren konstruktionsbedingt kausale Effekte nur im überlappenden Kovariatenbereich *S* abschätzen können, und dadurch in der empirischen Anwendung auch dafür sensibilisieren, in welchem Ausmaß regressionsbasierte Schlussfolgerungen von den zugrundeliegenden Annahmen über die funktionale Form des Modells abhängig sind.

Substantiell betrachtet erzwingt die Verwendung von Matchingverfahren schließlich geradezu, sich in der Analyse auf einen (zentralen) kausal wirksamen Faktor zu konzentrieren, dessen Wirkung mit einer bestimmten Datenbasis empirisch getestet werden soll. Quasi als Nebenprodukt der statistischen Strategie wird dazu eine klare Spezifikation der Ausgangshypothese sowie der empirischen Testbedingungen notwendig. Im Unterschied zur gängigen Regressionspraxis ergibt sich dadurch zum Beispiel eine klar hierarchische Struktur der unabhängigen Variablen und ihrer Funktion in der Datenanalyse: Matchingverfahren verdeutlichen unmittelbar, dass nicht jeder unabhängigen Variable denselben kausalen Status besitzt. Vielmehr erfordert das Verfahren, den relevanten kausalen Faktor *T* und die Kovariaten *X* explizit zu benennen. Da letztere (lediglich) dazu dienen, die Annahmen der unit homogeneity und strict ignorability zu rechtfertigen, kommt ihnen in der Interpretation

erkennbar kein kausaler Status zu. Matchingverfahren führen einen Hypothesentest allein für die Wirkung von *T* auf *Y* durch und ermitteln damit kausale Effekte im Sinne der „effects of causes“ (vgl. Goldthorpe 2001); Matchingverfahren sind wenig geeignet, Varianzerklärungen im Sinne des Konzepts der „causes of effects“ zu erfassen.

3.1 Vorgehensweise

Allen Matchingverfahren gemeinsam ist die Grundidee, den kausalanalytischen Vergleich vergleichbarer Einheiten nur für diejenigen Merkmalskombinationen vorzunehmen, für die empirisch sowohl Fälle aus der Experimental- als auch der Kontrollstichprobe (= matches) vorliegen. Infolgedessen kann die statistische Analyse nichtparametrisch, d.h. unter Verzicht auf weitergehende Annahmen zur funktionalen Form des Modells, erfolgen. Die empirische Schätzung der ATT- und ATE-Parameter erfolgt in aller Regel durch einfache deskriptive Verfahren, etwa durch Mittelwertvergleiche oder weiterführende univariate Verteilungsanalysen in den beiden Vergleichsstichproben.

Tatsächlich können alle Matchingverfahren grundsätzlich einfach als Gewichtungsverfahren aufgefasst werden, die durch jeweils unterschiedliche Algorithmen für eine Anpassung der Verteilung der Hintergrundfaktoren *X* in der Kontrollstichprobe an die entsprechende Verteilung in der Experimentalstichprobe sorgen. Die Schätzfunktion für den ATT’ lässt sich beispielsweise allgemein angeben mit

$$ATT' = \sum_{i \in E} w_i \left[Y_{1i} - \sum_{j \in C} W_{i,j} Y_{0j} \right], \tag{9}$$

d.h. als (mit w_i gewichteter) Durchschnitt der Differenzen zwischen den Ereignissen Y_{1i} der Experimentalfälle *i* und der mit $W_{i,j}$ gewichteten Ereignisse Y_{0j} der Kontrollbeobachtungen *j* (Heckman et al. 1997, 1998). Einzelne Matchingalgorithmen unterscheiden sich in der Konstruktion der Vergleichsgewichte $W_{i,j}$, insbesondere darin, ob ein reales Matching von spezifischen Beobachtungen *i* und *j* aus beiden Stichproben vorgenommen wird, oder ob die

Vergleichsgruppe als neu gewichteter Durchschnitt aller Kontrollbeobachtungen konstruiert wird. Grundlage der Konstruktion der Vergleichsgruppe wiederum ist das so genannte Zuweisungsmodell, d.h. die empirische Beschreibung des Zusammenhangs zwischen der Zugehörigkeit zur Experimentalgruppe *E* und den beobachteten Kovariaten *X*. Wir betrachten im Folgenden die einzelnen Komponenten einer Matchinganalyse im Detail.

3.2 Zuweisungsmodell und Propensity scores

Die Spezifikation eines Zuweisungsmodells (sog. assignment model) bildet die zentrale Grundlage jeden Matchingverfahrens. Dem Zuweisungsmodell kommt im Rahmen einer Matchinganalyse dieselbe Funktion zu wie dem Einschluss von Kovariaten in einer Regressionsanalyse: das Zuweisungsmodell stellt die Vergleichbarkeit der Beobachtungseinheiten her, indem es die Ähnlichkeit bzw. Unähnlichkeit der beiden Stichproben beschreibt, und dadurch anschließend das Matching vergleichbarer Einheiten ermöglicht. Zugleich sollte ein sozialwissenschaftliches Zuweisungsmodell idealerweise auch für die potentiell endogene Zuweisung (Selbstselektion) von Akteuren in den Experimentalstatus kontrollieren, um dadurch ATT- und ATE-Parameter zu identifizieren. In das Zuweisungsmodell sollten deshalb alle beobachteten Kovariaten *X* eingehen, die sowohl den Experimentalstatus *T* als auch (direkt oder indirekt) das Ereignis *Y* beeinflussen. Die Kovariaten *X* müssen dabei ökonometrisch nicht strikt exogen von *Y* sein, dürfen aber natürlich nicht endogen von *T*, also nicht selbst eine Folge des Experimentalstatus *T* sein.[8]

[8] Wenn beispielsweise der kausale Effekt einer Arbeitslosigkeit auf den weiteren Erwerbsverlauf ermittelt werden soll, ist es weder im Rahmen einer Matching- noch innerhalb einer Regressionsanalyse sinnvoll, die Zahl der Stellenwechsel als Kovariate aufzunehmen. Eine im Vergleich zu durchgehend Beschäftigten höhere Zahl von Stellenwechseln unter Arbeitslosen dürfte gerade infolge der Arbeitslosigkeit hervorgerufen sein, ist also eine Komponente der untersuchten Wirkung und stellt keine Anpassung der Hintergrundfaktoren vor der Arbeitslosigkeit dar.

Das einfachste Zuweisungsmodell, das zudem selbst ebenfalls wieder nichtparametrisch ist, besteht darin, direkt den multidimensionalen Kovariatenvektor *X* als Grundlage eines Matchingalgorithmus zu betrachten, und dementsprechend ein exaktes Matching von Experimental- und Kontrollpersonen über alle Kovariaten *X* vorzunehmen. In der Praxis, und zudem insbesondere dann, wenn viele Kovariaten vorliegen, so dass das Zuweisungsmodell eigentlich inhaltlich zufriedenstellend geschätzt werden kann, wird diese Strategie aber selbst mit großen Datensätzen offensichtlich schnell an ihre Grenzen stoßen.

Von zentraler Bedeutung für die praktische Anwendung von Matchingverfahren ist deshalb das theoretische Ergebnis von Rosenbaum und Rubin (1983), dass es zur Anpassung von Kovariatendifferenzen zwischen beiden Stichproben ausreicht, ein Matching lediglich über einen Ähnlichkeitsindex, den so genannten Propensity score, durchzuführen. Der Propensity score ist dabei definiert als die bedingte Wahrscheinlichkeit einer Beobachtungsperson *i* mit Merkmalen X_i, der Wirkung des kausalen Faktors *T* ausgesetzt zu sein (= als Mitglied der Experimentalgruppe *E* beobachtet zu werden). Er wird empirisch üblicherweise als Vorhersagewahrscheinlichkeit *P(X)* z.B. aus einer logistischen oder einer Probit-Regression mit *T* als abhängiger und *X* als unabhängigen Variablen bestimmt,[9] und im Anschluss daran als Ähnlichkeitsmetrik für das Matching von Experimental- und Kontrollgruppe verwendet.

3.3 Matchingalgorithmen

Ausgehend von einer empirischen Schätzung des Zuweisungsmodells und der daraus generierten Propensity scores kann dann mithilfe unterschiedlicher Matchingalgorithmen versucht werden, die Verteilung der beobachteten Hintergrundfaktoren *X* in der Kontrollstichprobe *C* an die entsprechende Verteilung in der Experimentalstichprobe *E*

[9] In Simulationsstudien haben sich parametrische Modellannahmen in der Konstruktion der propensity scores als wenig problematisch erwiesen (vgl. Deheija und Wahba 2002; Smith und Todd 2002).

anzupassen, und dadurch statistisch eine Situation „sonst gleicher Bedingungen“ zu erzeugen, in der kausale Schlussfolgerungen gerechtfertigt werden können. In der angewandten Forschung lassen sich im wesentlichen drei Matchingstrategien unterscheiden: Stratifizierung und Nearest-neighbor-Verfahren (Rubin 1978; Rubin und Rosenbaum 1983, 1985; Rosenbaum 2002; Deheija und Wahba 1999, 2002), die vor allem aus dem Bereich der angewandten Statistik stammen, sowie andererseits Kernel Matching und verwandte Verfahren, die in den letzten Jahren in der Ökonometrie entwickelt wurden (v.a. Heckman et al. 1997, 1998; Lechner 1999, 2002; Smith und Todd 2002).

Ein wichtiger Unterschied zwischen diesen Strategien besteht darin, dass beim stratifiziertem Matching, aber auch beim Nearest-neighbor Matching nur ein Teil der Kontrollbeobachtungen C zur Konstruktion des kontrafaktischen Ereignisses $\{Y_{0i}\}$ herangezogen werden, im Kernel Matching und verwandten Verfahren (z.B. Local Linear oder Mahalanobis Matching) wird dagegen ein gewichteter Durchschnitt aus der gesamten Kontrollstichprobe gebildet. Ein zweiter Unterschied liegt darin, dass Nearest-neighbor und verwandte Verfahren zu einem spezifischen Matching zwischen Beobachtungen i der Experimentalgruppe und Beobachtungen j der Kontrollgruppe führen, während Stratifizierung und Kernel Matching die kontrafaktische Beobachtung $\{Y_{0i}\}$ jeweils als gewichtete Durchschnitte aus einer Population von Kontrollbeobachtungen bestimmen. Tabelle 1 fasst die wesentlichen Merkmale einiger gängiger Algorithmen zusammen.

- Tabelle 1 etwa hier -

Wie unser eigenes empirisches Beispiel unten zeigen wird, können die Vorzüge einzelner Ansätze zum Teil auch kombiniert werden, etwa indem innerhalb von Schichten ein kernelbasiertes Matching oder ein Nearest-neighbor Matching durchgeführt wird. Ebenso ist es möglich durch Angabe eines calipers, d.h. einer maximal zulässigen Distanz c, auch

innerhalb von Nearest-neighbor-Verfahren schlechte Matches mit hohen Differenzen im Propensity score auszuschließen, oder durch multiples Matching auch im Rahmen von Nearest-neighbor-Algorithmen die kontrafaktische Beobachtung als gewichteten Durchschnitt mehrerer „ähnlich guter" Matches zu bilden (z.B. Smith 1997). Welche Kombination dabei auch immer verwendet wird, zentral bleibt das Ziel, eine Kontrollstichprobe mit empirisch der Experimentalstichprobe entsprechenden Hintergrundfaktoren *X* zu erzeugen. Die Güte eines spezifischen Zuweisungsmodells und des darauf aufbauenden Matchingalgorithmus sollte deshalb durch so genannte balancing tests auf Mittelwertunterschiede (bias) auf *P(X)* oder einzelnen (zentralen) Kovariaten *X* für die Analysepopulation oder auch spezifische Subgruppen ermittelt werden (Rosenbaum und Rubin 1983, 1985; Rosenbaum 2002). Bei festgestellten Verstößen kann dann sowohl der Matchingalgorithmus wie auch das Zuweisungsmodell (etwa durch Aufnahme von Interaktionseffekten) jeweils spezifisch angepasst werden.[10]

Alle Unterschiede in der empirischen Vorgehensweise der verschiedenen Matchingalgorithmen sollten allerdings nicht darüber hinwegtäuschen, dass alle Verfahren asymptotisch äquivalent sind, da der kausale Vergleich bei großem N in jedem Verfahren ausschließlich auf exakten Matches basiert. In der Praxis ergeben sich natürlich Unterschiede zwischen den Verfahren, so dass es sich in der angewandten Literatur eingebürgert hat, vor allem bei kleineren Stichproben oder sehr spezifischen Kausaleffekten Sensitivitätsanalysen mit jeweils variablem Analyseaufbau durchzuführen. Als Faustregeln für die empirische Analyse lässt sich sagen, dass Single Nearest-neighbor-Verfahren, evtl. ergänzt um einen caliper, in aller Regel sehr gut geeignet sind, vergleichbare Kontrollgruppen mit minimalem bias zu erzeugen. Wenn jedoch entweder nur sehr wenige sehr gut geeignete Matches oder

[10] Balancing tests sind dabei ausschließlich als Tests des durch einen spezifischen Algorithmus erzielten Matchings für ein gegebenes Zuweisungsmodell zu verstehen. Sie stellen keinen statistischen Test für die Güte des Zuweisungsmodells dar.

auch sehr viele relativ gut bis sehr gute Matches zur Verfügung stehen, dann können Multiple Nearest-neighbor-Verfahren oder auch Kernel Matching dazu beitragen, (wenigstens bzw. zusätzlich) die Varianz der Parameterschätzungen zu verringern (z.B. Smith 1997; Smith und Todd 2001). Wie für nichtparametrische Verfahren üblich, werden beide Ziele vor allem in kleinen Stichproben mit wenigen Kontrollbeobachtungen nicht gleichzeitig zu erreichen sein.

3.4 Punktschätzungen und statistische Inferenz

Da der eigentliche Kern der Matchingverfahren in der Konstruktion der kontrafaktischen Vergleichsgruppe besteht, gestaltet sich die anschließende Schätzung kausaler Effekte sehr unaufwendig. Unter Gültigkeit der Annahme konditionaler Unabhängigkeit des Experimentalstatus lässt sich beispielsweise der average treatment effect on the treated (ATT') nach Gleichung (9) als einfacher Mittelwertvergleich für die Experimentalgruppe *E* und der gematchten bzw. neu gewichteten Kontrollbeobachtungen *C* berechnen. Entsprechend einfach können subgruppenspezifische (konditionale) Parameter als Mittelwertvergleich innerhalb von Subgruppen ermittelt werden (z.B. Lechner 1999). Und ebenso einfach können durch Vergleiche der univariaten Ergebnisverteilungen auch kausale Effekte auf andere Merkmale, etwa den Median oder die Quartile der Verteilung, bestimmt werden (vgl. Imbens 2003; Gangl und DiPrete 2004). Da zudem bei konditional unabhängiger Zuweisung des Experimentalstatus der konditionale ATT_X dem konditionalen ATE_X entspricht, ergibt sich der ATE'-Parameter aus Gleichung (5) als (der Struktur der Gesamtpopulation angepasster) gewichteter Mittelwert der konditionalen ATT_X-Parameter.

Da die gematchten Stichproben gleichzeitig eine Verteilung individueller ATT- bzw. ATE-Parameter erzeugen, können Signifikanztests für die ermittelten Effekte im Prinzip mithilfe gewöhnlicher t-Tests bzw. äquivalenten nichtparametrischen Rangsummentests durchgeführt werden (Rosenbaum 2002). Die Verwendung dieser Teststatistiken unterstellt implizit allerdings individuell gegebene Propensity scores. Verteilungstheoretische Modelle,

die zusätzlich berücksichtigen, dass empirisch lediglich geschätzte Propensity scores vorliegen, existieren bislang lediglich für einzelne (Klassen von) Matchingalgorithmen (vgl. etwa Heckman et al. 1998 für Kernel und Local Linear Matching-Verfahren). In der Praxis dominieren deshalb Bootstrapverfahren zur Ermittlung der Standardfehler (vgl. Efron und Tibshirani 1993; Mooney und Duval 1993).

4 Ein Anwendungsbeispiel: Kausaleffekte von Arbeitslosigkeit

Wir wollen die empirische Anwendung von Matchingverfahren im Folgenden an einer Analyse der Karriereeffekte von Arbeitslosigkeit mit Daten des Sozio-ökonomischen Panels beispielhaft illustrieren. Wir erwarten theoretisch, dass Arbeitslosigkeit kurz- und eventuell auch längerfristige Karrierenachteile auslöst, weil mit dem Verlust des Arbeitsplatzes auch eine Entwertung firmenspezifischen Humankapitals sowie unter Umständen die Notwendigkeit einer beruflichen Neuorientierung verbunden ist. Diese Forschungsfrage zielt eindeutig auf den (kurz- bzw. längerfristigen) kausalen Effekt einer Arbeitslosigkeit, d.h. auf die kontrafaktische Frage nach den „effects of causes“ ab, wie stark sich die Erwerbsverläufe von empirisch arbeitslos Gewordenen allein durch die Erfahrung der Arbeitslosigkeit verändert haben. Welcher Anteil etwa der Einkommensunterschiede in der Bundesrepublik auf Arbeitslosigkeitserfahrungen zurückzuführen ist, oder welche Auswirkung alternative Determinanten von Arbeitsmarkterfolg – etwa Bildung oder Geschlecht – besitzen, ist für die Antwort auf diese Frage nicht von Belang.

Wir betrachten stattdessen im Folgenden lediglich die Frage, welche Auswirkung ein Arbeitsplatzverlust auf die nachfolgenden individuellen Beschäftigungschancen, das erzielte Erwerbseinkommen, sowie die individuelle Arbeitsplatzsicherheit hat. Zur Vereinfachung de Analyse verwenden wir lediglich Arbeitsplatzverluste aus den Rezessionsjahren 1993 und 1994, d.h. alle Ereignisse, die zwischen den Befragungswellen J und K des Sozio-ökonomischen Panels beobachtet wurden (vgl. zum Datensatz SOEP-Group 2001). Wir

beobachten die anschließenden Erwerbsverläufe sowohl unmittelbar (T_0+1) als auch bis zu insgesamt fünf Jahren (T_0+5) nach einem Arbeitsplatzverlust. Für jeden der bis zu fünf Befragungszeitpunkten nach dem ursprünglichen Arbeitsplatzverlust erfassen wir die Beschäftigungschance als Wahrscheinlichkeit, (abhängig oder selbständig) erwerbstätig zu sein, das logarithmierte reale Bruttomonatseinkommen in dieser Beschäftigung sowie prospektiv aus den Folgewellen die tatsächliche weitere Beschäftigungsdauer beim gegenwärtigen Arbeitgeber als Maß der Arbeitsplatzsicherheit. Da wir uns für den kausalen Effekt eines Arbeitsplatzverlustes für die tatsächlich von Arbeitslosigkeit Betroffenen interessieren, wollen wir im Folgenden kausale Effekte im Sinne der ATT-Parameter ermitteln.

Aufgrund der methodologischen Trennung von Kovariatenkontrolle und Effektschätzung im Matchingverfahren können wir alle drei abhängigen Variablen auf der Basis jeweils desselben Zuweisungsmodells und Matchingalgorithmus bearbeiten, und lediglich im letzten Schritt der Punktschätzung dem Skalenniveau jeweils angemessene Verfahren (Mittelwert- bzw. Anteilsvergleich für Einkommen und Beschäftigungschancen, Kaplan-Meier-Schätzer für Ereignisdaten) einsetzen. Da der Median der weiteren Beschäftigungsdauer in aller Regel nicht beobachtet werden kann, beschreiben wir die Beschäftigungsstabilität durch den Wert der Überlebensfunktion $G(36)$ für jeweils den Zeitpunkt 3 Jahre nach dem betreffenden Welleninterview $T_0+k,\ k=1..5$.

Wir stellen für jede der drei abhängigen Variablen drei unterschiedliche Analysen vor, die jeweils unterschiedliche Matchingalgorithmen verwenden: zum einen einen 1x1 Nearest-neighbor-Algorithmus, der für jede Experimentalbeobachtung i innerhalb des common support S und des Calipers c ein Matching mit einer (der nächstliegenden) Beobachtung j aus der Kontrollgruppe erzeugt, sowie eine zweite (1x10) Nearest-neighbor-Variante, in der (soweit vorhanden) bis zu 10 verfügbare Kontrollbeobachtungen im common support und innerhalb des Calipers als matches für eine Experimentalbeobachtung herangezogen werden.

Wir verwenden in beiden Analysen einen sehr engen Caliper von c=0.05*Standardabweichung des Propensity scores, um nur sehr gute matches in der Analyse zu berücksichtigen, und dadurch eine sehr hohe Biasreduktion zu erreichen.[11] Wir kontrastieren die Nearest-neighbor-Verfahren mit einer Analyse auf der Basis eines Kernel Matching, in welchem wir einen Epanechnikov-Kernel mit einer Bandbreite von ebenfalls h=0.05* Standardabweichung des Propensity scores. Diese drei Analysen werden zusätzlich jeweils stratifiziert für die alten bzw. neuen Bundesländer durchgeführt, so dass matches jeweils nur innerhalb der Teilregionen gesucht werden.[12]

Allen Analysen liegt zudem ein gemeinsames Zuweisungsmodell zugrunde, welches das Risiko eines Arbeitsplatzverlustes zwischen den SOEP-Wellen J und K durch Bildung, Berufserfahrung, Geschlecht, ethnische Zugehörigkeit und Migrationsstatus, bisheriges Erwerbseinkommen, detaillierten Beruf und Branche der gegenwärtigen Erwerbstätigkeit, sowie die Firmengröße des gegenwärtigen Arbeitgebers. Wir schätzen dazu – schichtspezifisch – eine logistische Regression, mit der wir ein Pseudo-R^2 von insgesamt 7,6% (West: 7,8%, Ost 7,4%) erreichen. Da wir im SOEP über weiterreichende Längsschnittinformationen verfügen, würde es sich in einer inhaltlichen Analyse aber in

---

[11] Diese Strategie hängt offensichtlich von der empirischen Datenlage ab, und dabei insbesondere von der Verfügbarkeit sehr guter (sehr naher) Kontrollbeobachtungen möglichst über den gesamten common support. Diese Bedingung ist tendenziell natürlich in großen Stichproben, aber auch bei wenig speziellen Experimentalgruppen eher erfüllt. In unseren eigenen Analysen gelingt es uns mit jedem der gewählten Verfahren, zu jeweils allen Experimentalbeobachtungen innerhalb des common support adäquate Kontrollbeobachtungen zu finden. Zusätzlich umfasst der Bereich des common support immer über 95% aller Experimentalfälle, so dass eine hinreichende Generalisierbarkeit der Ergebnisse gegeben erscheint. Vor diesem Hintergrund ist zu erwarten, dass sich das 1x1 Nearest-neighbor-Verfahren als relativ ineffizient erweisen wird, da sowohl das multiple Nearest-neighbor als auch das Kernel Matching die reichlich vorhandenen Kontrollbeobachtungen vollständig(er) in die Parameterschätzung einbeziehen.

[12] Die regionale Stratifizierung entspricht einer fixed-effects-Strategie im Rahmen von Regressionsverfahren, mit der in diesem Fall alle beobachteten und unbeobachteten Effekte auf der regionalen Ebene konstant gehalten werden (vgl. Heckman et al. (1997) für eine ausführliche Sensitivitätsanalyse der Konsequenzen regional unstratifizierten Matchings im Falle von amerikanischen Arbeitsmarktdaten)

jedem Fall anbieten, zusätzlich etwa noch für die Anzahl der Arbeitslosigkeitsspells innerhalb der letzten fünf Jahre oder für ähnliche Längsschnittkovariaten zu kontrollieren, und dadurch die Plausibilität der conditional independence-Annahme zu erhöhen.

- Tabelle 2 etwa hier -

Wie die in Tabelle 2 dargestellten Balancing tests zeigen, gelingt uns in allen drei Matchingalgorithmen und für jeweils alle drei abhängigen Variablen eine sehr weitreichende Angleichung der Verteilung der Propensity scores in der Kontrollstichprobe an die entsprechende Verteilung in der Experimentalgruppe. Die verbleibenden Mittelwertunterschiede sind minimal, ebenso wie (dank der großen Stichprobe des SOEP) die Unterschiede zwischen den einzelnen Algorithmen. In keiner Analyse ist der verbleibende Bias (Zeile (1)) auch nur annähernd statistisch signifikant. Wir gehen also davon aus, dass das Matching der Stichproben jeweils erfolgreich war, und eine Abschätzung der ATT-Parameter möglich ist.[13] Es sollte an dieser Stelle aber noch einmal explizit betont werden, dass der Nachweis des erfolgreichen Matchings nicht gleichbedeutend mit einem Modelltest für das Zuweisungsmodell zu verstehen ist. Ob und inwieweit in einer konkreten empirischen Analyse tatsächlich die kausalen ATT- und ATE-Parameter identifiziert sind, hängt allein von der inhaltlichen Plausibilität des Zuweisungsmodells ab – die, wie oben angedeutet, in der konkreten Analyse durchaus noch verbesserungsfähig wäre.

- Tabelle 3 etwa hier -

[13] In einer inhaltlichen Analyse würde man u.U. zusätzliche Balancing tests für Subgruppen oder auch für spezifische Kovariaten *X* durchführen, um die Qualität des Matchings dezidiert jeweils innerhalb wichtiger Subgruppen bzw. für zentrale Kovariaten zu überprüfen.

Tabelle 3 schließlich gibt die durch die Matchingverfahren ermittelten ATT'-Parameter für den kausalen Effekt eines Arbeitsplatzverlustes 1993/94 auf den weiteren Erwerbsverlauf. Die in der Tabelle ausgewiesenen Koeffizienten sind jeweils als Mittelwertdifferenz in der gematchten Stichprobe bzw. für die Beschäftigungsstabilität als Differenz der Überlebensfunktion *G(36)* ermittelt. Alle Matchingalgorithmen sind konsistent dahingehend, dass ein Arbeitsplatzverlust unmittelbar, aber auch mittelfristig geringere Beschäftigungschancen, Einkommensverluste sowie eine geringere Beschäftigungsstabilität nach sich zieht. Für alle drei abhängigen Variablen fallen die kurzfristigen Effekte dabei deutlich negativer aus als im mittelfristigen Vergleich nach bis zu 5 weiteren (Berufs)Jahren. Zum Teil gelingt es den Betroffenen also, entstandene negative Karrierefolgen wieder auszugleichen, insbesondere im Hinblick auf die mittelfristige Beschäftigungssicherheit.

Im Vergleich der Ergebnisse der verschiedenen Matchingalgorithmen zeigen sich im Detail – und trotz nahezu äquivalenter Biasreduktion – zudem die erwarteten Unterschiede in der Performanz in vergleichsweise großen Stichproben. Während das 1x10 Nearest-neighbor und das Kernel Matching zu fast deckungsgleichen Ergebnissen gelangen, weichen die durch das 1x1 Nearest-neighbor-Verfahren ermittelten Schätzwerte zum Teil deutlich ab. Dies betrifft einmal die Parameterschätzungen des ATT' selbst, die vor allem in der Einkommensanalyse, aber zum Teil auch in der Analyse der Beschäftigungsstabilität offenbar auch in dieser großen Stichprobe noch relativ sensitiv sind.[14] Zudem liegen die Standardfehler der Koeffizienten im 1x1 Nearest-neighbor-Modell in allen Analysen deutlich über den entsprechenden Werten der alternativen Algorithmen, in welchen die Kontrollstichprobe umfassender genutzt wird.

---

[14] Beide Nearest-neighbor-Analysen wurden „mit Zurücklegen", d.h. mit potentiell mehrmaliger Verwendung derselben Kontrollbeobachtung durchgeführt, wodurch sich die Sensitivität der Ergebnisse bekanntermaßen erhöht. Wie das 1x10 Matching zeigt, kann diesem Nachteil der Nearest-neighbor-Verfahren aber durch multiples Matching abgeholfen werden (vgl. auch Smith 1997).

5 Ausblick und Perspektiven

Aus vielerlei Gründen stellen Matchingverfahren unserer Meinung nach eine attraktive und sinnvolle Erweiterung des Methodenspektrums der empirischen Sozialforschung dar. Matchingverfahren sind intuitiv einleuchtend, leicht handhabbar und relativ voraussetzungslos in den ihnen zugrundeliegenden statistischen und mathematischen Annahmen. Zudem besitzen Matchingverfahren eine solide konzeptionelle Fundierung durch das Rubin Causal Model und identifizieren direkt die relevanten Kausalparameter des Modells. Damit sind Matchinganalysen in aller Regel sowohl technisch relativ einfach durchführbar, als auch inhaltlich leicht kommunizierbar.

Gleichzeitig ist die Anwendung von Matchingverfahren auf sozialwissenschaftliche Fragestellungen noch ein relativ junges statistisches Feld, das sich in den letzten Jahren sehr schnell entwickelt hat. Es ist deshalb zu erwarten, dass der weiteren Verbreitung der Verfahren auch eine stärkere Standardisierung folgt, in welche zunehmend vielfältigere Erfahrungen in der Anwendung unterschiedlicher Algorithmen in bestimmten Daten- und Fragekontexten einfließen. Zudem entwickelt sich das Feld auch insofern dynamisch weiter, als die Überschneidungen und Unterschiede zu traditionellen ökonometrischen Verfahren zunehmend deutlich werden (Heckman und Navarro-Lozano, 2003), Matchingschätzer z.B. als bedingte Differenzen-in-Differenzen-Schätzer in der Analyse von Längsschnittdaten eingesetzt werden (Smith und Todd 2002) oder sensitivitätsanalytische Methoden entwickelt werden, um Verzerrungen durch Effekte ungemessener Variablen abschätzen zu können (Rosenbaum 2002; DiPrete und Gangl 2004). Zudem sind vielfältige neue Anwendungsmöglichkeiten denkbar, etwa auf makrosoziologische oder international vergleichende Anwendungen, während bislang in der statistischen und ökonometrischen Literatur ausschließlich mikroanalytische Fragestellungen betrachtet wurden.

Dennoch, auch Matchingverfahren sind natürlich kein Wundermittel der empirischen Sozialforschung, sondern immer nur mehr oder minder geeignetes statistisches Hilfsmittel bei

der – notwendigerweise annahmebehafteten – empirischen Überprüfung sozialwissenschaftlicher Hypothesen. Die Validität matchingbasierter kausaler Schlussfolgerungen hängt dabei von einer relativ weitgehenden Annahme ab: dass es gelungen ist, durch die beobachteten Kovariaten sowohl vergleichbare Kontextbedingungen in der Experimental- und Kontrollgruppe herzustellen als auch die potentiell endogene Zuweisung des Experimentalstatus zu kontrollieren. Die Plausibilität dieser Annahme ist letzten Endes nicht mit statistischen Methoden zu beantworten, sondern verlangt eine sorgfältige theoretische Spezifikation der sich in den Beobachtungsdaten wiederspiegelnden sozialen Mechanismen und Prozessen, und die inhaltlich begründete Abschätzung, inwieweit dieselben durch die beobachteten Kovariaten hinreichend erfasst werden.

Aus dieser Perspektive ist vielleicht die verfahrenslogisch notwendige Transparenz der Analysen und der ihnen zugrundeliegenden inhaltlichen Annahmen das stärkste Argument für die Verwendung von Matchingverfahren in der Soziologie: um überhaupt empirische Ergebnisse erhalten zu können, müssen die methodologischen Grundlagen – die Frage nach der Identifizierbarkeit eines bestimmten kausalen Effekts mit den zur Verfügung stehenden Daten, der kausale Status der Kovariaten, sowie die konkrete Operationalisierung der Testbedingung und des Zuweisungsmodells – notwendigerweise sehr viel expliziter geklärt werden, als dies bei (mangelhafter Anwendung) manch anderer Methoden der Fall ist. Insofern steht zu hoffen, dass die Beschäftigung mit kausalanalytischen Matchingverfahren zu einem verbesserten Verständnis der theoretischen und empirischen Grundlagen von Kausalschlüssen in den Sozialwissenschaften führt, so dass in empirischen Analysen die den Schlussfolgerungen jeweils zugrundeliegenden Annahmen unmittelbar offen gelegt und damit diskutabel werden.

Literatur

Abbott, Andrew, 1998: The Causal Devolution. Sociological Methods & Research 27: 148-181.

Angrist, Joshua, Guido W. Imbens und Donald Rubin, 1996: Identification of Causal Effects Using Instrumental Variables. Journal of the American Statistical Association 91: 444-472 (mit Diskussion).

Berk, Richard A., 2004: Regression Analysis. A Constructive Critique. Thousand Oaks: Sage.

Brüderl, Josef, 2000: Regressionsverfahren in der Bevölkerungswissenschaft. S. 589-642 in Ulrich Müller, Bernhard Nauck und Andreas Diekmann (Hrsg.) Handbuch der Demographie. Band I. Berlin: Springer.

Dehejia, Rajeev H., und Sadek Wahba, 1999: Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs. Journal of the American Statistical Association 94: 1053-1062.

Dehejia, Rajeev H., und Sadek Wahba, 2002: Propensity Score-Matching Methods for Nonexperimental Causal Studies. Review of Economics and Statistics 84: 151-161.

DiPrete, Thomas A., und Markus Gangl, 2004: Assessing Bias in the Estimation of Causal Effects: Rosenbaum Bounds on Matching Estimators and Instrumental Variables Estimation with Imperfect Instruments. Discussion paper SP I 2004-101. Berlin: Wissenschaftszentrum Berlin für Sozialforschung.

Efron, Bradley, und Robert J. Tibshirani, 1993: An Introduction to the Bootstrap. New York: Chapman & Hall.

Engel, Uwe, 2004: Sozialer Kontext in der Mehrebenenanalyse. In Andreas Diekmann (Hrsg.), Methoden der Sozialforschung. Sonderheft 43 der Kölner Zeitschrift für Soziologie und Sozialpsychologie. Opladen: Westdeutscher Verlag.

Falk, Armin, und Ernst Fehr, 2003: Why Labour Market Experiments? Labour Economics 10: 399-406.

Fehr, Ernst, und Simon Gächter, 2000: Cooperation and Punishment in Public Goods Experiments. American Economic Review 90: 980-94.

Gangl, Markus, und Thomas A. DiPrete, 2004: Unemployment Insurance and Post-unemployment Wages: Evidence from Matching Estimators. Berlin, Durham: mimeo.

Garfinkel, Irwin, Charles F. Manski und Charles Michalopolous, 1992: Micro Experiments and Macro Effects. S. 253-273 in Charles F. Manski und Irwin Garfinkel (Hrsg.), Evaluating Welfare and Training Programs. Cambridge, MA: Harvard University Press.

Goldthorpe, John H., 2001: Causation, Statistics, and Sociology. European Sociological Review 17: 1-20.

Heckman, James J., 1992: Randomization and Social Policy Evaluation. S. 201-230 in Charles F. Manski und Irwin Garfinkel (Hrsg.), Evaluating Welfare and Training Programs. Cambridge, MA: Harvard University Press.

Heckman, James J., 1997: Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations. Journal of Human Resources 32: 441-462.

Heckman, James J., Lance Lochner und Christopher Taber, 1998: General-Equilibrium Treatment Effects: A Study of Tuition Policy. American Economic Review 88: 381-386.

Heckman, James, 2001: Accounting for Heterogeneity, Diversity and General Equilibrium in Evaluating Social Programmes. Economic Journal 111: F654-F699.

Heckman, James, und Richard Robb, 1985: Alternative Methods for Evaluating the Impact of Interventions. S. 156-245 in James Heckman and Burton Singer (Hrsg.), Longitudinal Analysis of Labor Market Data. New York: Wiley.

Heckman, James J., Hidehiko Ichimura und Petra E. Todd, 1997: Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme. Review of Economic Studies 64: 605-654.

Heckman, James J., Hidehiko Ichimura und Petra E. Todd, 1998: Matching as an Econometric Evaluation Estimator. Review of Economic Studies 65: 261-294.

Heckman, James J., Robert J. LaLonde und Jeffrey A. Smith, 1999: The Economics and Econometrics of Active Labor Market Programs. S.1865-2097 in Orley Ashenfelter und David Card (Hrsg.), Handbook of Labor Economics. Band III. Amsterdam: North Holland.

Heckman, James, und Salvador Navarro-Lozano, 2003: Using Matching, Instrumental Variables and Control Functions to Estimate Economic Choice Models. NBER Working Paper 9497. Cambridge, MA: National Bureau of Economic Research.

Heckman, James, Justin L. Tobias und Edward Vytlacil, 2003: Simple Estimators for Treatment Parameters in a Latent Variable Framework. Review of Economics and Statistics 85: 748-755.

Holland, Paul W., 1986: Statistics and Causal Inference. Journal of the American Statistical Association 81: 945-960.

Imbens, Guido W., und Joshua D. Angrist, 1994: Identification and Estimation of Local Average Treatment Effects. Econometrica 62: 467-75.

Imbens, Guido W., 2003: Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review. NBER Technical Working Paper 294. Cambridge, MA: National Bureau of Economic Research.

Lechner, Michael, 1999: Earnings and Employment Effects of Continuous Off-the-Job Training in East Germany after Unification. Journal of Business & Economic Statistics 17: 74-90.

Lechner, Michael, 2002: Program Heterogeneity and Propensity Score Matching: An Application to the Evaluation of Active Labor Market Policies. Review of Economics and Statistics 84: 205-220.

Manski, Charles F., 1989: Anatomy of the Selection Problem. Journal of Human Resources 24: 343-360.

Manski, Charles F., 1995: Identification Problems in the Social Sciences. Cambridge, MA: Harvard University Press.

Marini, Margaret M., und Burton Singer, 1988: Causality in the Social Sciences. Sociological Methodology 18: 347-409.

Mooney, Christopher Z., und Robert D. Duval, 1993: Bootstrapping. A Nonparametric Approach to Statistical Inference. Newbury Park: Sage.

Pearl, Judea, 1998: Graphs, Causality, and Structural Equation Models. Sociological Methods & Research 27: 226-284.

Pearl, Judea, 2000: Causality. Models, Reasoning, and Inference. Cambridge: Cambridge University Press.

Pötter, Ulrich, 2004: Das Problem der Selektionsverzerrung bei nicht-experimentellen Daten. In Andreas Diekmann (Hrsg.), Methoden der Sozialforschung. Sonderheft 43 der Kölner Zeitschrift für Soziologie und Sozialpsychologie. Opladen: Westdeutscher Verlag.

Pratt, John W. und Robert Schlaifer, 1988: On the Interpretation and Observation of Laws. Journal of Econometrics 39: 23-52.

Rosenbaum, Paul R., 2002: Observational Studies. 2. Auflage. New York: Springer.

Rosenbaum, Paul R., und Donald B. Rubin, 1983: The Central Role of the Propensity Score in Observational Studies for Causal Effects. Biometrika 70: 41-55.

Rosenbaum, Paul R., und Donald B. Rubin, 1985: Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score. The American Statistician 39: 33-38.

Rubin, Donald B., 1974: Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. Journal of Educational Psychology 66: 688-701.

Rubin, Donald B., 1978: Bayesian Inference for Causal Effects. Annals of Statistics 6: 34 58.

Rubin, Donald B., 1991: Practical Implications of the Modes of Statistical Inference for Causal Effects and the Critical Role of the Assignment Mechanism. Biometrics 47: 1213-1234.

Smith, Herbert L., 1997: Matching with Multiple Controls to Estimate Treatment Effects in Observational Studies. Sociological Methodology 27: 325-353.

Smith, Jeffrey A., und Petra E. Todd, 2001: Reconciling Conflicting Evidence on the Performance of Propensity-Score Matching Methods. American Economic Review 91: 112-118.

Smith, Jeffrey A., und Petra E. Todd, 2002: Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators? University of Western Ontario und University of Pennsylvania, mimeo.

Sobel, Michael E., 1995: Causal Inference in the Social and Behavioral Sciences. S. 1-38 in Gerhard Arminger, Clifford C. Clogg und Michael E. Sobel (Hrsg), Handbook of Statistical Modeling for the Social and Behavioral Sciences. New York: Plenum.

Sobel, Michael E., 1996: An Introduction to Causal Inference. Sociological Methods & Research 24: 353-379.

Sobel, Michael E., 1998: Causal Inference in Statistical Models of the Process of Socioeconomic Achievement. Sociological Methods & Research 27: 318-348.

SOEP Group, 2001: The German Socio-Economic Panel (GSOEP) after more than 15 Years – Overview. Vierteljahreshefte zur Wirtschaftsforschung 70: 7-14.

Winship, Christopher, und Michael E. Sobel, 2004: Causal Inference in Sociological Studies. Im Erscheinen in Melissa Hardy und Alan Bryman (Hrsg.), Handbook of Data Analysis. Thousand Oaks: Sage.

Winship, Christopher, und Stephen L. Morgan, 1999: The Estimation of Causal Effects from Observational Data. Annual Review of Sociology 25: 659-707.

Wooldridge, Jeffrey M., 2002: Econometric Analysis of Cross Section and Panel Data. Cambridge, MA: MIT Press.

Tabellen und Abbildungen

Abbildung 1 – Das Fundamentalproblem der Kausalanalyse

		Potentiell endogene Zuweisung	
		Kausaler Faktor T = 1 (Testbedingung)	Kausaler Faktor T = 0 (Kontrollbedingung)
Potentiell fehlende Vergleichbarkeit	Intervention T=1, Experimentalstichprobe E	(a) $Y_{1i,i\in T} \mid X_i, T=1$	(b) $\{Y_{0i,i\in T} \mid X_i, T=0\}$
	Intervention T=0, Kontrollstichprobe C	(c) $\{Y_{1i,i\in C} \mid X_i, T=1\}$	(d) $Y_{0i,i\in C} \mid X_i, T=0$

Anmerkung: Latente Ereignisse in geschweiften Klammern.

Tabelle 1 – Zentrale Matchingalgorithmen

	Beschreibung	Gewichtungsfunktion $W_{i,j}$
Stratifizierung	Schichtung der Stichprobe über *P(X)*; Mittelwertvergleich innerhalb der Schichten *k*	$W_{i,j} = \begin{cases} 1/N_j & \text{für } j \in k_i \\ 0 & \text{sonst} \end{cases}$
Nearest-neighbor matching	Direkte Paarbildung aus der Experimental- und Kontrollstichprobe; Matching anhand größter Ähnlichkeit auf *P(X)*	$W_{i,j} = \begin{cases} 1/N_j & \text{für } j \in \min\left\|P_i(X) - P_j(X)\right\| \\ 0 & \text{sonst} \end{cases}$
Caliper matching	Direkte Paarbildung aus der Experimental- und Kontroll-stichprobe; Matching innerhalb eines Ähnlichkeitsradius *c* über *P(X)*	$W_{i,j} = \begin{cases} 1/N_j & \text{für } j \in \left\|P_i(X) - P_j(X)\right\| \le c \\ 0 & \text{sonst} \end{cases}$
Mahalanobis matching	Kontrollbeobachtung als gewichteter Durchschnitt aller Fälle der Kontrollstichprobe; Gewichtungs-faktoren definiert durch Mahalanobisdistanzen über *X*	$W_{i,j} = \begin{cases} \dfrac{\left[(X_j - X_i)' C_X (X_j - X_i)\right]^{-1}}{\sum_{k \in C}\left[(X_k - X_i)' C_X (X_k - X_i)\right]^{-1}} \end{cases}$ mit C_X – Kovarianzmatrix des Kovariatenvektors *X*
Kernel matching	Kontrollbeobachtung als gewichteter Durchschnitt aller Fälle der Kontrollstichprobe; Gewichtungs-faktoren ermittelt durch Distanzfunktion über *P(X)*	$W_{i,j} = \begin{cases} \dfrac{K\left[(P_j(X) - P_i(X))/h\right]}{\sum_{k \in C} K\left[(P_k(X) - P_i(X))/h\right]} \end{cases}$ mit *K* – Kernelfunktion, z.B. $\phi(.)$ und *h* – Bandbreitenparameter

Tabelle 2 – Balancing Tests der Modellschätzungen

		AV: Beschäftigung			AV: Logarithmiertes Realeinkommen			AV: Beschäftigungsstabilität		
		(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
Zeitpunkt		Nearest neighbor 1x1	Nearest neighbor 1x10	Kernel matching	Nearest neighbor 1x1	Nearest neighbor 1x10	Kernel matching	Nearest neighbor 1x1	Nearest neighbor 1x10	Kernel matching
T+1	(1)	-0,0001	-0,0001	<0,0001	-0,0001	-0,0001	<-0,0001	-0,0001	-0,0001	<0,0001
(N_i = 490, 344,	(2)	0,002	0,002	<0,001	0,002	0,002	<0,001	0,002	0,002	<0,001
393)	(3)	-99,8%	-99,8%	-100,0%	-99,8%	-99,8%	-100,0%	-99,8%	-99,8%	-100,0%
T+2	(1)	-0,0001	-0,0001	<0,0001	-0,0002	-0,0002	<0,0001	-0,0002	-0,0002	<0,0001
(N_i = 467, 295,	(2)	0,002	0,002	0,001	0,002	0,002	0,001	0,002	0,002	0,001
334)	(3)	-99,8%	-99,8%	-99,9%	-99,8%	-99,8%	-99,9%	-99,8%	-99,8%	-99,9%
T+3	(1)	-0,0001	-0,0001	<0,0001	-0,0002	-0,0002	<0,0001	-0,0002	-0,0002	<0,0001
(N_i = 451, 274,	(2)	0,002	0,002	0,001	0,002	0,002	<0,001	0,002	0,002	<0,001
296)	(3)	-99,8%	-99,8%	-99,9%	-99,8%	-99,8%	-100,0%	-99,8%	-99,8%	-100,0%
T+4	(1)	-0,0002	-0,0002	<0,0001	-0,0002	-0,0002	<0,0001	-0,0002	-0,0002	<0,0001
(N_i = 420, 234,	(2)	0,002	0,002	<0,001	0,003	0,003	<0,001	0,002	0,002	0,001
251)	(3)	-99,8%	-99,8%	-100,0%	-99,7%	-99,7%	-100,0%	-99,7%	-99,7%	-99,9%
T+5	(1)	-0,0002	-0,0002	<0,0001	-0,0002	-0,0002	<-0,0001	-0,0002	-0,0002	<0,0001
(N_i = 391, 197,	(2)	0,002	0,002	0,001	0,003	0,003	<0,001	0,002	0,002	<0,001
216)	(3)	-99,8%	-99,8%	-99,9%	-99,7%	-99,7%	-100,0%	-99,7%	-99,7%	-100,0%

Anmerkung: Stratifiziertes Nearest neighbor-Matching mit caliper $c = 0{,}05 \cdot \mathrm{sd}(P(X))$ und stratifiziertes Kernel matching mit Epanechnikov-Kernel, Bandbreite $h = 0{,}05 \cdot \mathrm{sd}(P(X))$; Stratifizierung jeweils für die alten bzw. neuen Bundesländer. Angaben in den Tabellenzellen jeweils für (1) Mittelwertdifferenz $P_E(X)$-$P_C(X)$ im Propensity score zwischen Experimental- und gematchter Kontrollgruppe, statistische Signifikanzangaben für * p<.05 (Bootstrap-Standardfehler, N=100 Replikationen); (2) Standardisierter Bias des Propensity score in der gematchten Stichprobe und (3) Relative Biasreduktion gegenüber den Ausgangsstichproben.

Datenquelle: Sozio-ökonomisches Panel, Wellen J-R, eigene Berechnungen.

Tabelle 3 – Der kausale Effekt eines Arbeitsplatzverlustes 1993/94 auf den weiteren Erwerbsverlauf (ATT')

	AV: Beschäftigung			AV: Logarithmiertes Realeinkommen			AV: Beschäftigungsstabilität (G_{36})		
	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
Zeitpunkt	Nearest neighbor 1x1	Nearest neighbor 1x10	Kernel matching	Nearest neighbor 1x1	Nearest neighbor 1x10	Kernel matching	Nearest neighbor 1x1	Nearest neighbor 1x10	Kernel matching
T+1	-0,553*	-0,563*	-0,563*	-0,107	-0,120*	-0,121*	-0,208*	-0,194*	-0,191*
	(0,030)	(0,024)	(0,027)	(0,057)	(0,042)	(0,042)	(0,045)	(0,035)	(0,033)
T+2	-0,261*	-0,289*	-0,289*	-0,153*	-0,094	-0,089	-0,138*	-0,124*	-0,125*
	(0,032)	(0,027)	(0,028)	(0,073)	(0,055)	(0,048)	(0,052)	(0,033)	(0,038)
T+3	-0,266*	-0,235*	-0,241*	-0,168*	-0,125*	-0,127*	-0,052	-0,089*	-0,083*
	(0,037)	(0,031)	(0,027)	(0,064)	(0,045)	(0,051)	(0,052)	(0,040)	(0,032)
T+4	-0,224*	-0,224*	-0,232*	-0,188*	-0,153*	-0,145*	-0,033	-0,066	-0,089*
	(0,034)	(0,025)	(0,026)	(0,064)	(0,050)	(0,052)	(0,049)	(0,039)	(0,043)
T+5	-0,184*	-0,168*	-0,172*	0,003	-0,089	-0,088	0,012	-0,033	-0,040
	(0,041)	(0,031)	(0,028)	(0,071)	(0,057)	(0,047)	(0,058)	(0,045)	(0,042)

Anmerkung: Stratifiziertes Nearest neighbor-Matching mit caliper $c = 0{,}05 \cdot \mathrm{sd}(P(X))$ und stratifiziertes Kernel matching mit Epanechnikov-Kernel, Bandbreite $h = 0{,}05 \cdot \mathrm{sd}(P(X))$; Stratifizierung jeweils für die alten bzw. neuen Bundesländer. Angaben in den Tabellenzellen jeweils für ATT' = Mittelwertdifferenz $Y_E(P(X)) - Y_C(P(X))$ in der abhängigen Variable zwischen Experimental- und gematchter Kontrollgruppe. Bootstrap-Standardfehler des ATT' in Klammern (N=100 Replikationen; statistische Signifikanzangaben für * p<.05).

Datenquelle: Sozio-ökonomisches Panel, Wellen J-R, eigene Berechnungen.