AgEcon SEARCH
RESEARCH IN AGRICULTURAL & APPLIED ECONOMICS

The World's Largest Open Access Agricultural & Applied Economics Digital Library

This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.

Help ensure our sustainability.

Give to AgEcon Search

AgEcon Search
http://ageconsearch.umn.edu
aesearch@umn.edu

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

No endorsement of AgEcon Search or its fundraising activities by the author(s) of the following work or their employer(s) is intended or implied.

Ökonometrische Methoden der Politikevaluation: Meilenstein für eine sinnvolle Agrarpolitik der 2. Säule oder akademische Fingerübung?

Econometric policy evaluation techniques: milestone for effective 2nd pillar policies or a pure academic exercise?

Christian H.C.A. Henning und Jerzy Michalek
Christian-Albrechts-University of Kiel

Zusammenfassung

In jüngster Zeit wird verstärkt die Bedeutung einer umfassenden Politikevaluierung von internationalen Organisationen wie der OECD, der Weltbank und der FAO und insbesondere auch von der Europäischen Kommission hervorgehoben. Speziell für die Agrarpolitiken der 2. Säule ist eine Evaluierung der Politikmaßnahmen obligatorisch. Allerdings ist festzustellen, dass die bisherige Evaluationspraxis deutlich hinter den formulierten Ansprüchen zurückbleibt. Die Diskrepanz zwischen theoretischem Anspruch und praktischer Umsetzung ist allerdings verständlich, da eine konsistente quantitative Politikevaluierung eine starke methodische Herausforderung darstellt und es bislang kaum praktisch anwendbare und konsistente methodische Ansätze in der Agrarökonomie gibt, die eine quantitative Evaluierung ländlicher Entwicklungspolitik oder Agrarumweltpolitik erlauben. In diesem Zusammenhang stellt das Papier unterschiedliche innovative ökonometrische Evaluierungstechniken vor, die im Rahmen des EU-Forschungsprojektes ADVANCED-EVAL weiterentwickelt worden sind. Insbesondere wird die Problematik der bisher angewandten EU-Evaluierungstechniken durch einen Vergleich mit innovativen ökonometrischen Methoden wie insbesondere dem Propensity-Score-Matching am Beispiel der SAPARD-Evaluierung in der Slowakei demonstriert.

Schlüsselwörter

Politikevaluation; propensity score matching; 2. Säule der Agrarpolitik

Abstract

Recently the importance of a comprehensive policy evaluation has been increasingly recognized by international organizations, e.g. the World Bank, OECD and FAO as well as especially by the EU. In particular, for agricultural policies of the second pillar a comprehensive evaluation is obligatory. However, presently applied evaluation techniques are clearly lagging behind the ambiguous evaluation targets set by the EU Commission. Given the fact that a comprehensive policy evaluation is a very complex methodological challenge this discrepancy is not really surprising. Thus, nowadays it is still fair to conclude that adequate evaluation tools applicable to EU rural development policies do not exist yet. In this context this paper derives microeconometric evaluation techniques which have been developed within the EU research project ADVANCED-EVAL to evaluate RD policy programmes. In particular, the methodological shortcomings of simple evaluation techniques currently applied to evaluate EU RD policy programmes are demonstrated via a comparison of empirical evaluation results of SAPARD policies in Slovakia derived from these simple EU techniques with evaluation results derived from advanced microeconometric methods, i.e. propensity score matching.

Key words

policy evaluation; propensity score matching; second pillar of the CAP

1. Einleitung: *Policy Learning* durch Politikevaluierung

Mit Beginn der Agrarreform 1992 hat eine kontinuierliche Verschiebung der finanziellen Mittel aus der ersten in die zweite Säule stattgefunden. Der Kommissionsvorschlag zum anstehenden *Health-Check* führt diese Verschiebung finanzieller Mittel in die zweite Säule fort. Während aus wohlfahrtstheoretischer Sicht der Abbau der Preis- und Marktpolitik in der ersten Säule positiv zu bewerten ist, hängt die Bewertung der Stärkung der zweiten Säule von der Wirkung der konkreten Politiken ab. Im Gegensatz zu den Preis- und Marktpolitiken der ersten Säule ist die Wirkung der ländlichen Entwicklungspolitik wie auch der Agrarumweltpolitikmaßnahmen, welche die zweite Säule dominieren, erheblich komplexer und entsprechend nicht so einfach zu bewerten. Insofern stellt die Evaluierung der Politikmaßnahmen der zweiten Säule auch eine methodische Herausforderung für die Agrarökonomie dar. Konkret sind Kosten-Nutzen-Analysen auf der Grundlage einfacher Partialmodelle, die das methodische Kernstück zur Evaluierung von Agrarmarkt- und Preispolitiken darstellen, nicht mehr das adäquate Mittel zur Evaluierung von Politiken der zweiten Säule. Dies folgt aus der Tatsache, dass diese Politiken anders als Preispolitik in der Regel nicht direkt in klassische mikroökonomische Verhaltensmodelle integrierbar sind. Entsprechend lassen sich Effekte spezieller Politikmaßnahmen auch nicht unmittelbar auf der Mikro- oder Makroebene mit entsprechenden ökonomischen Modellen abbilden. Insbesondere lassen sich die konkreten funktionalen Zusammenhänge nicht explizit spezifizieren. Insofern sind vor allem nicht parametrische Ansätze zur Ermittlung entsprechender Politikeffekte hilfreich. In dieser Hinsicht stellt die *Ex-post*-Evaluierung ein effektives Instrument des *Policy-Learning* gerade hinsichtlich der Politiken der 2. Säule (MICHALEK, 2007) wie auch allgemein der Regionalpolitik (OECD 2002) oder auch struktureller Arbeitsmarktpolitiken (CALIENDO 2006, HAGEN und SPERMANN, 2004) dar.

Erst auf der Grundlage empirisch gemessener Effekte einzelner Politikmaßnahmen kann *ex post* festgestellt werden, welche Maßnahmen erfolgreich waren und welche nicht. Insofern ist eine umfassende Politikevaluierung gerade für Politiken der 2. Säule, deren komplexe Wirkungsmechanismen oft noch gar nicht explizit abgebildet werden können, eine zentrale Voraussetzung für die Formulierung effektiver und effizienter Agrarpolitiken. Gleichzeitig wird es politisch zunehmend wichtiger, dem Steuerzahler aufzu-

zeigen, dass die knappen finanziellen Ressourcen der EU sinnvoll eingesetzt werden.

Eine quantitative Politikevaluation ist nicht nur ein wirksames Instrument zum *Policy-Learning*, sondern darüber hinaus ein effektives Kontrollmittel für Politiker und Beamte (*accountability*) sowie für ein effizientes Instrument einer erfolgreichen Politikplanung (OECD, 2004).

Gerade in jüngster Zeit wurde die Bedeutung einer umfassenden Politikevaluierung nicht nur von vielen internationalen Organisationen wie der OECD, der Weltbank und der FAO erkannt, sondern insbesondere auch von der Europäischen Kommission sowie vielen nationalen Regierungen. Unter anderem ist eine umfassende Politikevaluierung seit der Verwaltungsreform im Jahr 2000 Pflicht für alle Gemeinschaftsaktivitäten (TOULEMONDE et al., 2002). Ebenfalls wurde speziell für die Agrarpolitiken der 2. Säule die Bedeutung einer konsistenten Evaluierung der Politikmaßnahmen hervorgehoben (EU-MINISTERRAT, 2005).

Allerdings bleibt festzustellen, dass, obwohl die EU die Bedeutung einer konsistenten Politikevaluation allgemein sowie speziell der Agrarpolitik in der 2. Säule erkannt hat und diese auch verpflichtend eingeführt hat, die bisherige Evaluationspraxis deutlich hinter den formulierten Ansprüchen zurückbleibt. Im Kern hat die Kommission zwar einen konsistenten theoretischen Rahmen für eine konsistente Politikevaluierung formuliert (EU-KOMMISSION 2004), aber die praktizierten Evaluierungstechniken sind überwiegend qualitativ und genügen den Ansprüchen einer vernünftigen Politikevaluierung nicht (MICHALEK, 2007).

Die Diskrepanz zwischen theoretischem Anspruch und praktischer Umsetzung ist allerdings verständlich, da eine konsistente quantitative Politikevaluierung ohne Frage eine starke methodische Herausforderung darstellt und es bislang kaum praktisch anwendbare und konsistente methodische Ansätze in der Agrarökonomie gibt, die eine quantitative Evaluierung ländlicher Entwicklungspolitik oder Agrarumweltpolitik erlauben.

Dieser methodischen Problematik ist sich die EU-Kommission voll bewusst, und nicht zuletzt deshalb hat sie ein entsprechendes EU-Forschungsprojekt im 6. Rahmenprogramm ausgeschrieben, in dem adäquate Methoden zur Ex-post- und Ex-ante-Evaluierung der Agrarpolitiken in der 2. Säule entwickelt werden sollen.

In diesem Zusammenhang stellt das Papier unterschiedliche innovative ökonometrische Evaluierungstechniken vor, die im Rahmen des EU-Forschungsprojektes Advanced-Eval zur Evaluierung agrarpolitischer Maßnahmen in der 2. Säule (SAPARD-Politiken in der Slowakei und Polen) angewandt wurden. Insbesondere wird die Problematik der bisher angewandten EU-Evaluierungstechniken durch einen Vergleich mit den innovativen ökonometrischen Methoden am Beispiel der SAPARD-Evaluierung in der Slowakei demonstriert.

Hierzu werden im nächsten Abschnitt zunächst kurz das grundsätzliche Problem der Politikevaluierung dargestellt und mögliche Lösungsansätze diskutiert. Danach erfolgt ein empirischer Vergleich des bislang praktizierten Evaluierungsansatzes der EU mit einem mikroökonometrischen Ansatz anhand von SAPARD-Daten der Slowakei. Im letzten Abschnitt folgen eine kurze Zusammenfassung sowie ein Ausblick auf zukünftige Forschungsarbeiten.

2. Methoden der Politikevaluierung

Generell kann ein Politikevaluierungsprozess in drei Stufen zerlegt werden (vgl. CALIENDO, 2006). 1. Mikroökonomische Evaluierung, d.h. der Effekt eines Politikprogramms auf individuelle Wirtschaftsakteure. 2. Makroökonomische Evaluierung, d.h. aggregieren sich die individuellen Effekte des Programms zu einem sozialen Gewinn auf die Makroebene? 3. Kosten-Nutzenanalyse, d.h. sind die erzielten Effekte die besten Effekte, die für die eingesetzten finanziellen und personellen Ressourcen erzielt werden konnten?

Mikroökonomische Evaluierung

Die zentrale Frage hinsichtlich der mikroökonomischen Evaluierung ist die Ermittlung des Effektes, den die Teilnahme an einem Politikprogramm, z.B. an dem Einzelbetrieblichen Förderprogramm (EBF) oder der Förderung des ökologischen Landbaus, auf den individuellen Wirtschaftsakteur ausübt. In der Regel handelt es sich bei dem Effekt um eine gewünschte Performance, z.B. Gewinnsteigerung oder aber Steigerung der Umweltverträglichkeit. Sieht man einmal von der Messproblematik der Performance ab, so würde man idealerweise den Politikeffekt ermitteln, indem man die Performance eines Wirtschaftsakteurs einmal unter Teilnahme an dem Politikprogramm und einmal ohne Teilnahme vergleicht. In diesem Zusammenhang ergibt sich das fundamentale Evaluierungsproblem, da man in der Regel niemals für denselben Wirtschaftsakteur die Performance unter der Teilnahme und Nichtteilnahme an dem Politikprogramm beobachten kann. Um dieses fundamentale Evaluierungsproblem zu lösen, muss man die entsprechende Kontrollgruppe zu der Gruppe der Programmteilnehmer finden. Idealerweise entspricht die Kontrollgruppe der Gruppe der Teilnehmer in allen performance-relevanten Eigenschaften außer der Teilnahme bzw. Nichtteilnahme am Politikprogramm. In diesem idealen Fall könnte der Politikeffekt aus der Differenz zwischen der durchschnittlichen Performance der Teilnehmer und der Nichtteilnehmer ermittelt werden.

Die Frage ist nun, auf welche Weise man eine ideale Kontrollgruppe herstellen kann. Hier lassen sich experimentelle von nicht-experimentellen Designs unterscheiden. Im experimentellen Design werden die Teilnehmer und Nichtteilnehmer zufällig aus einer Grundgesamtheit aller Wirtschaftsakteure ausgewählt. In der Regel impliziert ein experimentelles Design eine ideale Kontrollgruppe. Allerdings stehen aus Kosten- und nicht zuletzt aus ethischen und politischen Gründen (siehe CALIENDO 2006) gerade für die Ex-post-Evaluierung für die meisten Politikmaßnahmen keine experimentellen Daten zur Verfügung. Insofern wurden gerade in jüngerer Zeit nichtexperimentelle Evaluierungsdesigns auf der Grundlage von ökonometrischen und statistischen Methoden entwickelt, die ebenfalls das fundamentale Evaluierungsproblem lösen (siehe vor allem die Arbeiten von HECKMAN mit verschiedenen Koautoren, z.B. HECKMAN und ROBB, 1985; HECKMAN und HOTZ, 1989; HECKMAN, LALONDE und SMITH, 1999).

Um diese methodischen Ansätze besser verstehen zu können, werden im Folgenden einige begriffliche Definitionen eingeführt. Grundsätzlich gehen diese begrifflichen Definitionen auf den so genannten *Potential Outcome Approach* (FISHER, 1935; NEYMANN, 1935; ROY, 1951) zurück, der in

der Literatur allgemein als Roy-Rubin-Model (RRM) abgehandelt wird (ROY 1951, RUBIN 1974). Entsprechend wird zwischen dem *„treatment"* (Teilnahme an einem politischen Programm wie z.B. EBF) und dem möglichen Ergebnis (*potential outcome*) unterschieden. In dem Grundmodell werden zwei mögliche Ergebnisse unterschieden, das Ergebnis, das sich unter Annahme eines *treatment* (Teilnahme) ergibt, Y_1, und das Ergebnis, das sich ohne *treatment* (Nichtteilnahme) ergibt, Y_0. *Treatment* und *non-treatment* werden mit Hilfe einer Dummy-Variablen D definiert, wobei D=1 *treatment* und D=0 *non-treatment* bezeichnet.

Weiterhin werden bestimmte Variablen X berücksichtigt, die annahmegemäß nicht durch das Treatment beeinflusst werden und somit auch als exogene Attribute der individuellen Wirtschaftsakteure bezeichnet werden (HOLLAND, 1986).

Der Treatment-Effekt ist dann als Differenz zwischen den potentiellen Ergebnissen definiert:

$$(1) \quad \Delta_i = Y_i^1 - Y_i^0$$

Das fundamentale Problem der Evaluierung folgt nun aus der Tatsache, dass für jeden individuellen Wirtschaftsakteur entweder das Ergebnis mit *treatment*, Y_1, oder aber ohne *treatment*, Y_0, beobachtet werden kann, aber niemals beide. Somit kann der Treatment-Effekt nicht direkt entsprechend Gl. (1) geschätzt werden. Die jeweils unbeobachtete Komponente in Gl. (1) wird auch als *counterfactoral outcome* (kontrafaktorisches Ergebnis) bezeichnet.

Da es unmöglich ist, für ein Individuum den individuellen Treatment-Effekt zu schätzen, konzentriert sich die Literatur darauf, durchschnittliche Treatment-Effekte für Teilnehmer- bzw. Nichtteilnehmergruppen zu schätzen. Entsprechend werden die folgenden durchschnittlichen Treatment-Effekte definiert. Erstens wird einfach die Differenz zwischen dem erwarteten Treatment-Effekt der Gruppe der Teilnehmer und dem erwarteten Treatment-Effekt der Gruppe der Nichtteilnehmer als *„Average Treatment Effect"* (ATE) berechnet:

$$(2) \quad \Delta_{ATE} = E(Y^1) - E(Y^0)$$

Inhaltlich gibt der ATE einfach an, welches Ergebnis zu erwarten wäre, wenn ein individueller Wirtschaftsakteur zufällig einem Treatment unterzogen werden würde. HECKMAN (1997) hat in diesem Zusammenhang richtig festgestellt, dass ATE oft nicht politisch relevant ist, da es gerade bei Programmen mit einer speziellen Zielgruppe, z.B. Arbeitslose, Effekte für individuelle Akteure mit berücksichtigt, auf die das Programm gar nicht zugeschnitten ist. Deshalb ist der prominenteste Evaluierungsparameter der so genannte *average treatment effect on the treated* (ATT):

$$(3) \quad \Delta_{ATT} = E(Y^1 | D=1) - E(Y^0 | D=1)$$

In Gl. (3) muss berücksichtig werden, dass die Variable D auf der Grundlage der tatsächlichen Teilnahme individueller Akteure definiert ist, d.h. $E(Y_0, D=1)$ ist das erwartete Ergebnis für ein Individuum, das an dem Programm teilgenommen hat, unter der Annahme, dass es nicht teilgenommen hätte. Geht man nun weiterhin davon aus, dass der individuelle Treatment-Effekt eines individuellen Akteurs nicht durch die Teilnahme anderer Akteure beeinflusst wird, so stellt ATT genau den individuellen Bruttogewinn des Programms dar, der dann in einer Kosten-Nutzen-Analyse mit den entsprechenden Programmkosten verglichen werden kann (HECKMANN, LALONDE und SMITH, 1999).

Weiterhin kann anhand von Gl. (3) sehr gut der Selektivitätsbias erläutert werden. Der zweite Term ist kontrafaktisch für die Teilnehmer, d.h. dieser kann nicht direkt beobachtet werden. Nun wird das o*utcome* ohne *treatment* für die Gruppe der Nichtteilnehmer beobachtet, so dass man unter der Annahme:

$$(4) \quad E(Y^0 | D=1) = E(Y^0 | D=0)$$

die Nichtteilnehmer als adäquate Kontrollgruppe verwenden könnte, d.h. der ATT würde sich unter der Annahme (4) empirisch auf der Grundlage der beobachteten Effekte für die Teilnehmer und Nichtteilnehmer schätzen lassen.

Unglücklicherweise gilt die Bedingung in Gl. (4) in der Regel nicht für nichtexperimentelle Daten, d.h. die Gruppe der Teilnehmer unterscheidet sich systematisch von der Gruppe der Nichtteilnehmer hinsichtlich der erfolgsbestimmenden Eigenschaften (X), d.h. es gilt:

$$(5) \quad E(Y^0 | D=1) \neq E(Y^0 | D=0)$$

Schätzt man unter der Annahme, dass Gl. (5) gilt, den ATT als Differenz der Mittelwerte der Teilnehmer und Nichtteilnehmer (als Kontrollgruppe), so ergibt sich der folgende Selektionsbias:

$$(6) \quad \begin{aligned} & E(Y^1 | D=1) - E(Y^0 | D=0) = E(Y^1 - Y^0 | D=1) \\ & + \left[E(Y^0 | D=1) - E(Y^0 | D=0) \right] \end{aligned}$$

Der Term in den eckigen Klammern auf der rechten Seite von Gl. (5) ist genau der Selektionsbias. Inhaltlich kann der Selektionsbias auf unterschiedliche Faktoren zurückgeführt werden, die beobachtbar oder aber unbeobachtbar sein können. Zum Beispiel kann ein Programm gerade auf wettbewerbsfähige Betriebe abgestellt sein, so dass die Teilnehmer und Nichtteilnehmer sich systematisch hinsichtlich ihres Erfolges auch ohne das politische Programm unterscheiden. In diesem Fall würde ein Vergleich der durchschnittlichen Performance zwischen Teilnehmern und Nichtteilnehmern den Programmeffekt (ATT) systematisch überschätzen. Andererseits können auch nicht unmittelbar beobachtbare Faktoren die Teilnahmeentscheidung beeinflussen, und diese Faktoren haben auch einen systematischen Einfluss auf die Performance. Zum Beispiel könnten ökologisch orientierte Betriebsleiter stärker motiviert sein, an einem Programm für umweltschonende Produktionsverfahren teilzunehmen, und diese ökologische Motivation beeinflusst ebenfalls die konsequente Umsetzung umweltschonender Produktionsverfahren.

Neben den ATE und ATT lässt sich noch eine Reihe weiterer interessanter Effekte[1] definieren, auf die wir an dieser

[1] Weitere Effekte sind unter anderem der *„average treatment effect of untreated"* (ATU), der *marginal treatment effect of treated* (MATE) oder the *local average treatment effect* (LATE) (IMBENS und ANGRIST, 1994) (vgl. auch HECKMAN, LALONDE und SMITH, 1999, HECKMAN, SMITH und CLEMENTS, 1997).

Stelle nicht weiter eingehen wollen, da unser Hauptfokus in diesem Papier auf der Berechnung des ATT mit Hilfe mikroökonometrischer Ansätze liegen soll.

2.1 Relevante ökonometrische Evaluierungsansätze

2.1.1 Grundlagen

In der Literatur wurde nun eine Reihe von ökonometrischen Ansätzen entwickelt, die das oben genannte Problem eines Selektionsbias bei nichtexperimentellem Datendesign lösen sollen. Diese Ansätze lassen sich grundsätzlich anhand von zwei Dimensionen einteilen (siehe CALIENDO, 2006). Eine erste Dimension entspricht dem benötigten Datendesign, konkret können hier Querschnitts- und Zeitreihenanalysen unterschieden werden. Eine zweite Dimension ist die Frage, wie der Selektionsbias konstruiert wird. Hier lassen sich Ansätze, die eine Selektion nach beobachtbaren Variablen, von Ansätzen, die eine Selektion anhand von unbeobachtbaren Variablen unterstellen, unterscheiden. Dabei sind die erste und die zweite Dimension miteinander verwoben, da Ansätze, die eine Selektion nach unbeobachtbaren Variablen unterstellen, grundsätzlich Paneldaten benötigen. Innerhalb der Ansätze, die eine Selektion nach beobachtbaren Variablen unterstellen, spielen die so genannten Matching-Verfahren, insbesondere das Propensity-Score-Matching, eine zentrale Rolle.

Jeder der ökonometrischen Ansätze macht spezielle Annahmen und setzt eine spezielle Datenverfügbarkeit voraus. Um die grundlegenden Annahmen der zentralen Ansätze besser herausarbeiten zu können, ist es hilfreich, das RRM mit dem klassischen ökonometrischen Ansatz zu verbinden. In der folgenden Darstellung folgen wir im Wesentlichen BLUNDELL und COSTAS DIAS (2002) bzw. CALIENDO (2006).

Zunächst definieren wir die folgenden *Outcome*-Regressionsgleichungen:

$$(7)\quad \begin{aligned} Y_{it}^1 &= g_t^1(X_i) + u_{it}^1 \\ Y_{it}^0 &= g_t^0(X_i) + u_{it}^0 \end{aligned}$$

wobei der Index t bzw. i die Zeitperiode t bzw. den individuellen Akteur i bezeichnet. Die Funktionen g^1 und g^0 bezeichnen die funktionale Beziehung zwischen dem „*potential outcome*“ und der Menge der beobachtbaren Eigenschaften eines Akteurs i. u^1 und u^2 sind Störterme, für die angenommen wird, dass sie einen Mittelwert von Null haben und jeweils unkorreliert mit den Regressorvariablen X sind.

Die Teilnahme an einem Programm ist nun nicht zufällig, sondern wird durch eine bestimmte Logik determiniert. Formal kann diese Logik mit Hilfe einer Indexfunktion IN_i abgebildet werden, wobei IN_i gerade den Netto-Nutzen, den der Akteur von der Teilnahme an einem Politikprogramm hat, repräsentiert. Wir nehmen weiterhin an, dass dieser Netto-Nutzen durch eine Reihe beobachtbarer Variablen, Z_i, sowie unbeobachtbarer Variablen, V_i, determiniert wird:

$$(8)\quad IN_i = f(Z_i) + V_i$$

Danach ergibt sich die Teilnahme eines Akteurs an einem Politikprogramm:

$$(9)\quad D_i = \begin{cases} 1 & IN_i > 0 \\ 0 & IN_i \le 0 \end{cases}$$

Unter den gegebenen Definitionen und unter der Annahme, dass das Treatment in der Periode k stattfindet, ergibt sich der individuelle Treatment-Effekt mit:

$$(10)\quad \Delta_{it}(X_i) = Y_{it}^1 - Y_{it}^0 = \left[g_t^1(X_i) - g_t^0(X_i)\right] + \left[u_{it}^1 - u_{it}^0\right] \quad \text{für } t > k$$

Entsprechend ergeben sich die relevanten Treatment-Effekte:

$$(11)\quad \begin{aligned} \Delta_{ATE} &= E\left(\Delta_{it} \middle| X = X_i\right) \\ \Delta_{ATT} &= E\left(\Delta_{it} \middle| X = X_i, D_i = 1\right) \\ \Delta_{ATU} &= E\left(\Delta_{it} \middle| X = X_i, D_i = 0\right) \end{aligned}$$

Wie bereits oben erklärt, ist die Teilnahme an einem Politikprogramm in der Regel nicht zufällig, d.h. es ergibt sich eine Korrelation zwischen der Teilnahme (D_i) und den Fehlertermen der *Outcome*-Regressionsgleichungen (u^1 und u^0). Diese Korrelation kann aus der stochastischen Abhängigkeit zwischen den Störtermen (u^1 und u^0) und V_i oder aus der stochastischen Abhängigkeit zwischen den Störtermen (u^1 und u^0) und Z_i resultieren. Im ersten Fall ergibt sich eine Selektion nach unbeobachtbaren Variablen (*selection on unobservables*) und im zweiten Fall eine Selektion nach beobachtbaren Variablen (*selection on observables*).

Es ergeben sich jeweils unterschiedliche ökonometrische Verfahren, die mit dem Problem „*selection on observables*“ bzw. „*selection on unobservables*“ umgehen. Wir werden uns in diesem Beitrag auf spezielle Matching-Verfahren konzentrieren, die grundsätzlich eine „*selection on observables*“ annehmen.

2.1.2 Ökonometrische Evaluationsschätzverfahren

Bevor wir das Propensity-Score-Matching-Verfahren zur Evaluation von Politikprogrammen genauer darstellen, sollen zunächst noch kurz drei grundlegende Evaluierungsstrategien diskutiert werden. Dies sind: (1) Before and After Estimator (BAE), (2) Cross-Section Estimator (CSE) und Difference-in-Difference Estimator (DID).

Die zentrale Annahme des BAE ist:

$$(12)\quad E(Y_{t'}^0 \mid D = 1) = E(Y_t^0 \mid D = 1)$$

Dabei bezeichnet t’ die Periode vor und t nach dem *treatment*. Unter dieser Annahme kann der folgende ATT-Schätzer abgeleitet werden:

$$(12a)\quad \Delta_{ATT}^{BAE} = E\left(Y_t^1 \middle| D = 1\right) - E\left(Y_{t'}^o \middle| D = 1\right)$$

Die Annahme Gl. (12) setzt dabei voraus, dass das individuelle Verhalten der Akteure nicht durch die Programmteilnahme beeinflusst wird, z.B., dass Akteure sich auf eine bestimmte Weise verhalten, um an dem Programm teilnehmen zu können. Weiterhin wird mit Gl. (12) implizit vorausgesetzt, dass keine zeitlich variierenden Faktoren die Performance zwischen der Periode t und t’ beeinflussen. Letztere schließt zum Beispiel grundlegende veränderte ökonomische Rahmenbedingungen in den Perioden t’ und t aus. Ein Vorteil dieses Verfahrens ist sicherlich, dass man lediglich Daten über Teilnehmer zu erheben braucht.

Das zweite grundlegende Verfahren CSE vergleicht nicht den gleichen Teilnehmer zu unterschiedlichen Zeitpunkten,

sondern zu einem Zeitpunkt Teilnehmer und Nichtteilnehmer. Voraussetzung ist allerdings, dass die relevanten X Charakteristika in der Gruppe der Teilnehmer und Nichtteilnehmer gleich verteilt sind, d.h. es gilt insbesondere:

$$(13)\quad E\left(Y_t^0 \middle| X, D=1\right) = E\left(Y_t^0 \middle| X, D=0\right)$$

Unter der Annahme Gl. (13) lässt sich der folgende ATT-Schätzer ableiten:

$$(14)\quad \Delta_{ATT}^{CES} = E\left(Y_t^1 \middle| X, D=1\right) - E\left(Y_{t'}^o \middle| X, D=0\right)$$

SCHMIDT (1999) hat gezeigt, dass die Annahme Gl. (13) nur korrekt ist, solange die Teilnahme nicht durch unbeobachtbare Faktoren (z.B. Motivation) beeinflusst wird. Weiterhin ist der CES-Schätzer im Gegensatz zu dem BAE-Schätzer nicht sensitiv gegenüber veränderten allgemeinen ökonomischem Rahmenbedingungen (Schocks) (HECKMAN, LALONDE und SMITH, 1999).

Der letzte grundlegende Ansatz ist der DID-Schätzer. Im Gegensatz zu den beiden vorangegangen Ansätzen umfasst der DID-Schätzer die Berücksichtigung von „selection on unabservables". Formal kann der DID-Schätzer als Erweiterung des BAE-Schätzers verstanden werden, da dieser die Differenz der Before-and-After-Differenz eines Teilnehmers von der Before-and-After-Differenz eines Nichtteilnehmers subtrahiert. Dadurch werden gemeinsame Zeittrends eliminiert:

$$(15)\quad \Delta^{DID} = \left[Y_t^1 - Y_{t'}^0 \middle| D=1\right] - \left[Y_t^0 - Y_{t'}^0 \middle| D=0\right]$$

Der DID-Schätzer basiert auf der Annahme eines zeitlich invarianten linearen Selektionseffekts. Um diese Annahme verständlich zu machen, nehmen wir folgende Outputbeziehung an:

$$(16)\quad Y_{it} = \pi_{it} + D_{it} Y_{it}^1 + (1 - D_{it}) Y_{it}^0$$

In Gl. (16) umfasst π_{it} Effekte der Selektion aufgrund unbeobachtbarer Faktoren („*effects of selection on unobservables*"). Der DID-Schätzer ist dann nur valide, wenn angenommen wird, dass diese Selektionseffekte zeitlich invariant sind, d.h. es gilt: $\pi_{it} = \pi_{it'}$.

Propensity Score Matching

Die grundlegende Annahme von statistischen Verfahren, d.h. Matching und Regressionsverfahren, zur Schätzung von Treatment-Effekten ist die so genannte *Conditional-independence*-Annahme, die auch als „*Unconfoundedness* oder „*selection on observables*" bezeichnet wird. Inhaltlich impliziert diese Annahme, dass systematische Unterschiede in der Performance für Individuen, für die alle relevanten beobachtbaren Kovariaten X identisch sind, eindeutig als Treatment-Effekte interpretiert werden können. Formal gilt:[2]

$$(17)\quad \begin{aligned} E\left(Y^0 \middle| X, D=1\right) &= E\left(Y^0 \middle| X, D=0\right) = E\left(Y^0 \middle| X\right) \\ E\left(Y^1 \middle| X, D=1\right) &= E\left(Y^1 \middle| X, D=0\right) = E\left(Y^1 \middle| X\right) \end{aligned}$$

Um zu gewährleisten, dass beide Seiten von Gl. (17) für alle X simultan definiert sind, wird zusätzlich die so genannte „*Common-Support*-Bedingung angenommen:

$$(18)\quad 0 < \Pr(D=1 \mid X) < 1 \quad \text{für alle X}.$$

Die Annahme in Gl. (18) impliziert, dass der Support, d.h. der Wertebereich, für den die Dichtefunktion nicht Null ist, für die Teilnehmer- und die Nichtteilnehmergruppe gleich ist.

Geht man nun davon aus, dass die *Common-Support*-Annahme (Gl. (18)) und die *Unconfoundedness*-Annahme erfüllt sind, so lässt sich der ATT durch den Vergleich der Teilnehmer und Nichtteilnehmer wie folgt berechnen:

$$(19)\quad \begin{aligned} \Delta_{ATT}^{MAT} &= E\left(Y^1 - Y^0 \middle| X, D=1\right) \\ &= E\left(Y^1 \middle| X, D=1\right) - E_X\left[E\left(Y^0 \middle| X, D=1\right)\right] \\ &= E\left(Y^1 \middle| X, D=1\right) - E_X\left[E\left(Y^0 \middle| X, D=0\right)\right] \end{aligned}$$

Der erste Term in Gl. (19) lässt sich mit Hilfe der Teilnehmergruppe und der zweite Term mit Hilfe der „gematchten" Nichtteilnehmergruppe schätzen. Dabei bezeichnet der Term „E_X" den Erwartungswert der Verteilung von X in der Teilnehmergruppe. Zusätzlich ist Gl. (19) nur in der *Common-Support-Region* „S" von X definiert. Matching ist nun ein statistisches Verfahren, dass sicherstellt, dass die Verteilung von X in der Gruppe der Nichtteilnehmer mit der Verteilung von X in der Gruppe der Teilnehmer für die *Common-Support-Region* S möglichst gut übereinstimmt.

Die Grundidee von Matching ist einfach. Man sucht zu jedem Teilnehmer einen Nichtteilnehmer, der mit diesem hinsichtlich aller relevanten Kovariaten X übereinstimmt. Technisch problematisch wird Matching, wenn die Dimension der relevanten Kovariaten groß wird. Geht man beispielsweise vereinfachend davon aus, dass alle Kovariate dichtom sind, so ergeben sich bereits 2^n mögliche Matches bei n Kovariaten. Insofern ist unmittelbar klar, dass in empirisch relevanten Studien ein exaktes „Zell-Matching" nicht möglich ist.

In diesem Zusammenhang haben ROSENBAUM und RUBIN (1983) Propensity–Score-Matching entwickelt, um Matching auch bei höheren Dimensionen technisch durchführbar zu gestalten. Technisch reduzieren Rosenbaum und Rubin die Matching-Dimensionen auf eine einzige Dimension, indem sie mit Hilfe einer logistischen Regression oder einer Probitschätzung die Wahrscheinlichkeit eines Akteurs, an einem Politikprogramm teilzunehmen, als Funktion seiner Charakteristika X schätzen: Pr(D=1)=b(x). Auf der Grundlage dieser Schätzung lässt sich für jeden Akteur ein Propensity Score, b(X), berechnen, auf dessen Grundlage das Matching dann einfach erfolgen kann.[3]

[2] Tatsächlich lässt sich die Bedingung in Gl. (17) auch durch eine schwächere Annahme als der *Unconfoundedness*, der so genannten *Mean-Independence*-Annahme, implizieren (vgl. HECKMAN et al., 1997).

[3] ROSENBAUM und RUBIN haben ihre Eigenschaften des Propensity-Score-Matching Verfahren auch formal abgeleitet. Der interessierte Leser wird auf ROSENBAUM und RUBIN (1983) verwiesen.

Im Vergleich zu Matching-Verfahren, die auf allen Kovarianten beruhen, ist Propensity–Score-Matching ein gröberes Matching-Verfahren, das aber in empirischen Studien gute Ergebnisse erzielt (DEHEJIA und WAHBA, 1999, 2002).

Hat man nun die Propensity Scores für alle Teilnehmer und Nichtteilnehmer berücksichtigt, so ist eine weitere technische Frage, mit welcher konkreten Technik das Matching individueller Teilnehmer zu entsprechenden Nichtteilnehmern erfolgt. Hier ergibt sich ebenfalls eine Vielzahl von unterschiedlichen Matching-Techniken, von denen wir die wichtigsten im Folgenden kurz darstellen wollen. Eine sehr gute Übersicht bieten HECKMAN et al. (1998) bzw. SMITH und TODD (2005) oder IMBENS (2004).

Um eine übersichtliche Darstellung zu gewährleisten, führen wir zunächst die folgende allgemeine Notation ein. Wir bezeichnen mit I_1 bzw. I_0 die Indexmenge der Teilnehmer bzw. Nichtteilnehmer.

Der geschätzte individuelle Treatment-Effekt eines Teilnehmers $i \in I_1$ ergibt sich dann aus der Differenz des beobachteten *outcoms* für den individuellen Teilnehmer und dem gewichteten durchschnittlichen *outcome* der Kontrollgruppe:

$$(20)\quad \Delta^{MAT} = \frac{1}{N_1}\sum_{i\in I_1}\left[Y_i^1 - \sum_{j\in I_0} W_{N_0}(i,j)Y_j^0\right]$$

, wobei N_1 bzw. N_0 die Anzahl der Beobachtungen in der Teilnehmer- und Nichtteilnehmergruppe bezeichnet.

Die einzelnen Matching-Verfahren unterscheiden sich nun in der Zuordnung der individuellen Gewichte, $W_{No}(i,j)$. Dabei ist für alle Verfahren die Summe der Gewichte aller Nichtteilnehmer für jeden individuellen Teilnehmer immer gleich eins.

Grundsätzlich ergibt sich die Logik der Zuordnung der Gewichte zum einzelnen Nichtteilnehmer, indem zunächst auf der Grundlage der Propensity Scores, b(X), für jeden individuellen Teilnehmer eine Nachbarschaft C(b) definiert wird. Ein individueller Teilnehmer i wird dann zu allen Nichtteilnehmer gematcht, die in seiner Nachbarschaft sind: $j \in C(b_i)$. Die einzelnen Verfahren unterscheiden sich nun dadurch, wie die Nachbarschaft definiert wird und wie die Gewichte innerhalb der Nachbarschaft berechnet werden.

Nearest Neighbour Matching

Der populärste Matching-Schätzer ist das *nearest neigbour* (NN) Matching. Dabei wird die Nachbarschaft wie folgt definiert:

$$(21)\quad C^{NN}(b_i) = \min_j \left\| b_i - b_j \right\|, \quad j \in N_0$$

Im Falle von Propensity-Score-Matching ist b_i bzw. b_j der Propensity Score von i bzw. j und $\| \ \|$ ist die Betragfunktion. Entsprechend wird beim NN-Matching der Nichtteilnehmer als „Match" gewählt, der einen Propensity Score aufweist, der möglichst dicht an dem Propensity Score des Teilnehmers i liegt. Somit folgt:

$$(22)\quad W_{N_0}^{NN}(i,j) = \begin{cases} 1, & if \left\| b_i - b_j \right\| = \min_j \left\| b_i - b_j \right\| \\ 0, & otherwise \end{cases}$$

Dabei gibt es unterschiedliche Varianten des NN-Matching. Unter anderem NN-Matching mit „*replacement*" und „*ohne replacement*", d.h. im ersten Fall kann ein Nichtteilnehmer mehrfach als „Match" für unterschiedliche Teilnehmer fungieren, während dies im letzten Fall ausgeschlossen ist. Während die Variante mit „*replacement*" die Qualität des Matching erhöht, impliziert diese gleichzeitig, dass die Anzahl der unterschiedlichen Matching-Partner verringert wird im Vergleich zum NN-Matching ohne *replacement*, wodurch die Varianz des Schätzers steigt, d.h. die Effizienz sinkt. Es ergibt sich somit ein Trade-off zwischen Effizienz und Verzerrung des Matching (vgl. SMITH und TODD, 2005). Als Kompromiss wird deshalb „*oversampling*" vorgeschlagen, d.h. es werden die dichtesten m>1 Nachbarn zum Matching herangezogen, wobei das Gewicht der einzelnen Nachbarn entweder uniform verteilt sein kann oder aber reziprok zu der Distanz zu dem Teilnehmer definiert wird (vgl. DAVIES and KIM, 2004).

Caliper and Radius Matching

NN-Matching birgt die Gefahr, dass schlechte Matches in die Analyse miteinbezogen werden, wenn nämlich der dichteste Nachbar immer noch absolut sehr weit von dem Teilnehmer entfernt ist. Diese Gefahr wird bei dem Caliper-Matching-Verfahren vermieden, indem eine absolute Grenze für die Entfernung des nächsten Nachbarn gesetzt wird (vgl. COCHRANE und RUBIN, 1973), d.h. es gilt:

$$(23)\quad C^{CM}(b_i) = \left\{ j \in N_0 \middle| \left\| b_i - b_j \right\| < \varepsilon \right\}$$

$$(23a)\quad W_{N_0}^{CM}(i,j) = \begin{cases} 1, & j \in C^{CM} \wedge \left\| b_i - b_j \right\| = \min_j \left\| b_i - b_j \right\| \\ 0, & sonst \end{cases}$$

Insofern ist Caliper-Matching eine Möglichkeit die "*Common Support*-Bedingung" einzuführen. Eine weitere Variante des Caliper-Matching wurde von DEHEJIA und WAHBA (2002) vorgeschlagen, indem diese analog zum „*oversampling*" nicht den nächsten Nachbarn sondern alle Nachbarn, die in der ε-Umgebung liegen, als Matching-Partner verwenden.

Stratification and Interval Matching

Diese Methode geht auf ROSENBAUM und RUBIN (1983) zurück. Technisch wird dabei die Gruppe aller Teilnehmer und Nichtteilnehmer anhand der geschätzten Propensity Scores in M Intervalle eingeteilt, und für jedes Intervall wird der ATT bzw. ATE separat berechnet. Der gesamte ATT bzw. ATE ergibt sich dann als gewichteter Durchschnitt aller Intervall-, ATT- bzw. ATE-Werte, wobei das Gewicht für den ATT gerade dem Anteil der Teilnehmer, die in einem Intervall liegen, entspricht. Für die Berechnung des ATE wird der Anteil der Teilnehmer und Nichtteilnehmer als Gewicht herangezogen.

Kernel and Local Polynomial Matching

Alle bisherigen Matching-Verfahren haben gemeinsam, dass immer nur ein kleiner Teil der gesamten Beobachtungen der Nichtteilnehmer zur Konstruktion des kontrafaktorischen *outcome* benutzt werden. Im Gegensatz dazu werden beim Kernel-Matching (KM) wie auch bei dem *Local-*

Linear-Matching (LLM) nicht parametrische Matching-Schätzer verwendet, die alle Nichtteilnehmer als Kontrollgruppe berücksichtigen. Entsprechend ist es ein Vorteil dieser Verfahren, dass sie eine geringere Varianz aufweisen, da sie alle verfügbaren Informationen nutzen. Nachteilig ist allerdings, dass diese Verfahren auch Beobachtungen zur Berechnung des kontrafaktorischen *outcome* verwenden, die tatsächlich sehr schlechte Matches sind. Insofern ist die Definition einer adäquaten Common-Support-Region essentiell für diese Verfahren (vgl. HECKMAN, ICHIMURA und TODD, 1998).

Kernel-Matching definiert I_0 als gesamte relevante Nachbarschaft und benutzt die folgenden Gewichte für die jeweiligen Nichteilnehmer:

$$W_{N_0}^{KM}(i,j) = \frac{G_{ij}}{\sum_{k \in I_0} G_{ik}} \qquad G_{ik} = \frac{G(b_i - b_j)}{a_{N_0}} \tag{24}$$

G bezeichnet dabei die Kernel-Funktion. Solange diese nicht negativ ist, nehmen die Kernel-Gewichte mit zunehmendem Abstand zwischen den Propensity Scores eines Nichtteilnehmers zu dem jeweiligen Teilnehmer ab.

Der LLM-Schätzer ist im Prinzip ein generalisierter Kernel-Schätzer (vgl. HECKMAN, ICHIMURA und TODD, 1998).

2.1.3 DID Matching Estimator

Alle bislang erwähnten Matching-Schätzer gehen davon aus, dass, solange man alle beobachtbaren relevanten Faktoren (X) kontrolliert, der bedingte Erwartungswert für die Gruppe der Teilnehmer und Nichtteilnehmer gleich ist. Diese Annahme ist in konkreten empirischen Anwendungen oft sehr restriktiv. Der DID-Matching-Schätzer geht von einer deutlich schwächeren Annahme aus, indem dieser zeitlich konstante Unterschiede in der Performance zwischen der Gruppe der Teilnehmer und Nichtteilnehmer zulässt. Beispielsweise können ökologisch motivierte Landwirte grundsätzlich erfolgreicher ökologisch wirtschaften, da sie motivierter sind. Gleichzeitig erhöht diese ökologische Motivation auch die Wahrscheinlichkeit, an einem ökologischen Förderprogramm teilzunehmen, so dass ein systematischer Unterschied in der ökologischen Wirtschaftsweise der Teilnehmer und Nichtteilnehmer erwartet werden kann, der aufgrund der nicht beobachtbaren unterschiedlichen ökologischen Motivation von Landwirten entsteht. In diesem Fall ist die „*conditional mean independence*-Annahme“ verletzt, d.h die o.g. Schätzer führen zu inkonsistenten Ergebnissen, während der DID-Matching-Schätzer konsistente Ergebnisse liefert, solange angenommen wird, dass die Motivationsunterschiede zeitlich konstant sind.

Formal basiert der DID-Matching-Schätzer auf der folgenden identifizierenden Annahme:

$$E\left(Y_t^0 - Y_{t'}^o \middle| b(X), D=1\right) = E\left(Y_t^0 - Y_{t'}^o \middle| b(X), D=0\right) \tag{25}$$

Wie aus Gl. (25) zu ersehen ist, erfordert der DID-Matching-Schätzer Paneldaten, wobei sich der ATT wie folgt ergibt:

$$\Delta_{ATT}^{DDM} = \frac{1}{N_1} \sum_{i \in I_1 \cap S_P} \left[\left(Y_{it}^1 - Y_{it'}^0\right) - \sum_{j \in I_0 \cap S_P} W_{N_0}(i,j) \left(Y_{tj}^0 - Y_{t'j}^0\right) \right] \tag{26}$$

S_P bezeichnet dabei die *Common Support Region* bzgl. der Propensity-Score-Verteilung der Teilnehmer und Nichtteilnehmer.

Die Gewichte hängen wie oben erklärt von dem konkreten gewählten Matching-Verfahren ab. Obwohl der DID-Matching-Schätzer klare Vorteile gegenüber den oben genannten Matching-Schätzern aufweist, gibt es bislang erst sehr wenige empirische Anwendungen dieses Schätzers (eine interessante agrarökonomische Anwendung bieten PUFAHL und WEISS, 2007). Entsprechende ökonometrische Evaluierungsstudien ländlicher Entwicklungspolitiken, u.a. auch unter Anwendung des DID-Matching-Schätzers, wurden im Rahmen des EU-Projektes Advanced-Eval durchgeführt. Erste empirische Ergebnisse, insbesondere ein Vergleich der Matching-Verfahren zu den bisher in der EU praktizierten naiven Evaluierungsansätzen, werden im nächsten Kapitel dargestellt.

3. Empirische Anwendung von Propensity Score Matching zur Evaluierung ländlicher Entwicklungspolitik

3.1 Angewandte Evaluierungsansätze der EU

Die EU hat gerade in jüngster Zeit verstärkt die Evaluierung von EU-Programmen entwickelt. Dabei versteht die EU unter Evaluierung einen Prozess, der eine Beurteilung erlaubt, inwieweit ein Politikprogramm seine proklamierten Ergebnisse und Ziele erreicht hat (vgl. „A practical guide for the Commission Services“, DG BUDGET, 2004).

Entsprechend der EU-Verordnung 1698/2005 umfasst die Evaluierung ländlicher Entwicklungspolitik eine Ex-ante-, Mid-term- und Ex-post-Evaluierung. Den methodischen Rahmen der Evaluierung stellt der so genannte *Logical-Framework-Matrix*-Ansatz dar, wonach systematisch Programm-Inputs und -Outputs mit den Programmeffekten, welche in unmittelbare *results* und mittelfristige *impacts* unterschieden werden, im Sinne einer Interventionslogik in Beziehung gesetzt werden. Neben der Herausarbeitung der Interventionslogik umfasst der *Logical-Framework-Matrix*-Ansatz vor allem die Definition geeigneter Indikatoren zur empirischen Messung der jeweiligen Programm-Inputs und -Outputs sowie der Programmeffekte.

Als relevante Evaluierungskriterien werden neben der Relevanz die Effektivität und Effizienz herangezogen. Programmrelevanz hebt dabei auf den Tatbestand ab, dass die jeweiligen Politikprogramme grundsätzlich auf identifizierte relevante Probleme (*needs*) abzielen sollten. Die Effektivität umfasst, dass ein Programm tatsächlich in der Lage ist, gewünschte Effekte zu bewirken, während Effizienz auf die Gegenüberstellung der Kosten und Nutzen eines Programms abzielt und diese im Sinne einer Kosten-Nutzen-Analyse vergleicht.

Die einzelnen Evaluierungsverfahren und methodischen Vorgehensweisen sind für die Evaluierung der ländlichen Entwicklungspolitik in EU-Richtlinien und -Verordnungen detailliert vorgegeben (siehe EUROPÄISCHE KOMMISSION

2006). Empirische Grundlage der Evaluierung sind dabei die so genannten „*evaluation questions*“, welche sich in *common evaluation questions* und *programme specific evaluation questions* unterteilen. Im Kern umfassen die *evaluation questions* die Erhebung von „Performance-Indikatoren“ auf der Mikroebene sowie der regionalen und nationalen Ebene mit Hilfe von Interviewdaten. Zum Beispiel lautet eine „*measure specific evaluation question*“ im Rahmen der Evaluierung der Förderung von Investitionen in den landwirtschaftlichen Betrieb: „To what extent have the supported investments contributed to improve the income of the beneficiary farmers?”.

Auf der Grundlage der erhobenen Performance-Indikatoren erfolgt dann eine Evaluierung des Politikprogramms.

Dabei werden in der Regel sehr einfache Evaluierungsverfahren angewendet. Im Einzelnen sind dies:

1. Before-and-After-Vergleich relevanter Performance-Indikatoren für die Programmteilnehmer (BAE).
2. Vergleich durchschnittlicher Performance-Indikatoren der Teilnehmer mit einer Vergleichsgruppe von Nichtteilnehmern bzw. dem Durchschnitt der Teilnehmer und Nichtteilnehmer (CSE).
3. Vergleich der zeitlichen Entwicklung eines Performance-Indikators der Teilnehmergruppe mit der zeitlichen Entwicklung des Performance-Indikators der Nichtteilnehmergruppe bzw. der gesamten Gruppe der Teilnehmer und Nichtteilnehmer (DID).

Berücksichtigt man die Ausführungen oben, so folgt unmittelbar, dass die Qualität der praktizierten einfachen Evaluierungsverfahren BAE, CSE und DID im Allgemeinen eher gering ist, wobei die Qualität von BAE über CSE zu DID zunimmt. Dies folgt insbesondere aus zwei Gründen. Erstens werden bei einfachen BAE-Verfahren weder der Einfluss anderer Faktoren, z.B. zeitliche Veränderung allgemeiner Rahmenbedingungen, noch entsprechende Selektionsprobleme berücksichtigt. CSE-Verfahren eliminieren zumindest zum Teil den Einfluss anderer Faktoren, da diese zu einem Zeitpunkt zwischen Teilnehmern und Nichtteilnehmern die Performance vergleichen. Allerdings bleiben, wie oben beschrieben, andere unbeobachtete Faktoren, die die Performance der Teilnehmer und Nichtteilnehmer beeinflussen, unberücksichtigt. Ebenso bleibt das Selektionsproblem, d.h. dass relevante beobachtbare Faktoren, die die Performance beeinflussen, unterschiedlich in der Gruppe der Teilnehmer und Nichtteilnehmer verteilt sind, unberücksichtigt. Der erste Punkt wird schließlich in den einfachen DID-Verfahren eliminiert, allerdings bleibt auch für diese Verfahren das Selektionsproblem unberücksichtigt. Dass die Nichtberücksichtigung der verzerrten Selektion hinsichtlich der Programmteilnahme zu einer erheblich verzerrten Evaluation führen kann, soll an dem folgenden einfachen Beispiel verdeutlich werden.

Man betrachte ein Informationsprogramm zur ökologischen Wirtschaftweise, an dem Landwirte teilnehmen können. Die Landwirte lassen sich dabei hypothetisch hinsichtlich ihrer ökologischen Grundeinstellung in zwei Gruppen enteilen. Eine „leicht überzeugbare“ und eine „normale“ Gruppe. Die Wahrscheinlichkeit, dass ein „leicht überzeugbarer“ Landwirt auf eine ökologische Produktionsweise im nächsten Jahr umstellt, sei der Einfachheit halber immer 1, unabhängig, ob dieser am Training teilnimmt oder nicht. Die Wahrscheinlichkeit, dass ein „normaler“ Landwirt im nächsten Jahr umstellt, sei ohne Training 0 und mit Training 0.5.

Geht man nun von 100 Teilnehmern aus, die in dem einen Implementationsszenario zu 50% aus normalen und zu 50% aus „leicht überzeugbaren“ Landwirten besteht, so wäre die beobachtete Performance, gemessen am Anteil der Landwirte, die nach dem Training umstellen, 75% (alle 50 „leicht überzeugbaren“ plus die Hälfte der normalen Landwirte). Geht man hingegen in einem anderen Implementationsszenario davon aus, dass alle 100 Teilnehmer normale Landwirte sind, so ergibt sich nur eine beobachtete Performance von 50% (die Hälfte aller normalen Teilnehmer). Gemessen an dem Performance-Indikator wäre also die erste Programmimplementation erfolgreicher als die zweite.

Will man hingegen den tatsächlichen „*Impact*“ der Programme messen, so muss man berücksichtigen, welche zusätzlichen Betriebsumstellungen durch das Programm impliziert wurden. Das heißt, man muss die beobachtete Performance mit der hypothetischen Situation vergleichen, dass die Teilnehmer nicht an dem Training teilgenommen hätten. Entsprechend unseren Annahmen würde sich im ersten Fall eine Umstellungsrate von 50% ergeben, da alle 50 „leicht überzeugbaren“ Teilnehmer auch ohne Training umstellen würden. Das heißt der tatsächliche Programm-*Impact* ist 75%-50%=25%. Hingegen würden sich im zweiten Fall ohne Training überhaupt keine Betriebsumstellungen ergeben, d.h. der Programm-*Impact* ist 50%-0%=50%. Es ergibt sich also ohne adäquate Berücksichtigung der Selektionsverzerrung der Teilnehmer und Nichtteilnehmer eine komplett verzerrte Bewertung des wahren Programm-*Impacts*. Dynamisch ist dies umso negativer zu beurteilen, wenn man berücksichtigt, dass Politiker bzw. Regierungsbeamte nach der erfolgreichen Bewertung der von ihnen durchgeführten Programme beurteilt und befördert werden. Bei Anwendung einer performance-indikator-basierten Evaluierung würden diese dann einen Anreiz haben, insbesondere Programmteilnehmer auszuwählen, die „leicht zu überzeugen“ sind und damit den wahren Programm-*Impact* gerade drastisch reduzieren.

Insofern ist eine adäquate Evaluierung, in der tatsächliche Programm-*Impacts* kalkuliert werden, d.h. insbesondere Selektionsverzerrungen der Teilnehmer explizit berücksichtigt werden, von entscheidender Bedeutung. Da in der Regel alle ländlichen Entwicklungsprogramme quasi per Design auf bestimmte Teilnehmer mit speziellen Eigenschaften, wie z.B. Betriebsgröße, Betriebsstruktur, Alter etc., abheben, führen die einfachen Evaluierungsmethoden in der Regel zu stark verzerrten und inkonsistenten Evaluierungsergebnissen.

Diese Diskrepanz zwischen einfachen praktizierten Evaluierungsmethoden der EU und den oben dargestellten Propensity-Score-Matching-Verfahren wird im nächsten Abschnitt noch einmal an einem empirischen Beispiel anhand der SAPARD-Programme in der Slowakei verdeutlicht.

3.2 Vergleich der praktizierten EU-Evaluierungsverfahren mit Propensity-Score-Matching-Verfahren

Um die Bedeutung der Verzerrung, die sich durch die Anwendung nicht adäquater Evaluierungsmethoden ergibt, zu

demonstrieren, werden im Folgenden die Ergebnisse der Ex-post-Evaluierung der SAPARD-Programme in der Slowakei für die klassischen EU-Verfahren (BAE, CSE und DID) sowie verschiedene korrespondierende Matching-Schätzer aufgeführt und diskutiert. Die Analysen wurden im Rahmen des EU-Projektes Advanced-Eval durchgeführt. Datengrundlage bildet ein *balanced* Panel von 293 FADN-Betrieben für die Jahre 2001, 2003 und 2005, von denen 51 Betriebe an SAPARD-Programmen teilgenommen haben (D=1) und von denen 181 Betriebe als Kontrollgruppe dienen, die grundsätzlich berechtigt waren an SAPARD-Programmen teilzunehmen, aber nicht teilgenommen haben. Konkret wurde die SAPARD-Maßnahme 1 „Investition in landwirtschaftliche Betriebe" als Beispiel gewählt.[4]

3.2.1 Methodisches Vorgehen

Die Evaluierung mit Hilfe der EU-Standardverfahren erfolgt, indem für die SAPARD-Teilnehmer bzw. die Kontrollgruppe der Nichtteilnehmer jeweils die durchschnittlichen Performance-Indikatoren berechnet werden. Konkret werden die Ergebnisse für (1) den Gewinn je Hektar Betriebsfläche in Slowakischen Kronen (SKK) pro Hektar, (2) die Arbeitsproduktivität gemessen als Deckungsbeitrag pro Arbeitskraftstunde dauerhaft beschäftigter Arbeitskräfte und (3) die Beschäftigung je Betrieb gemessen als dauerhaft beschäftigte Arbeitskräfte pro Betrieb gewählt.[5]

Die Evaluierung mit Hilfe des DID-Matching-Verfahrens erfolgte auf der Grundlage der geschätzten ATT. Entsprechend der oben beschriebenen Vorgehensweise wurde hierzu zunächst eine logistische Regression auf der Grundlage der gesamten Stichprobendaten für alle 232 Betriebe geschätzt, wobei die Teilnahme am SAPARD-Programm (D=1) die zu erklärende Variable war. Als erklärende Variable wurden 39 Betriebscharakteristika aus über 1500 Variablen ausgewählt (siehe Tabelle A1 im Anhang). Die Auswahl der erklärenden Variablen erfolgte entsprechend der empfohlenen methodischen Vorgehensweise (vgl. Caliendo 2006) einerseits anhand theoretischer Überlegungen sowie andererseits anhand von statistischen Signifikanztests. Schließlich wurden grundsätzlich nur Variablen in die letztendliche Schätzung aufgenommen, die den Balance-Test bestanden haben, d.h. nach dem Matching dürfen keine signifikanten Unterschiede der in die Schätzung aufgenommen Charakteristika X zwischen der Gruppe der SAPARD-Teilnehmer und der Kontrollgruppe der Nichtteilnehmer bestehen.

Auf der Grundlage der geschätzten Logitfunktion wurden die Propensity Scores berechnet, und es erfolgte die Festlegung der Common Support Region nach dem Verfahren von LEUVEN und SIANESI (2003). Das Matching erfolgte nach dem Kernel-Matching-Verfahren, wobei eine Epanechnikov-Kernel-Funktion verwendet wurde.

Zusätzlich wurden Sensitivitätsanalysen mit Hilfe der *Rosenbaum Bounds* durchgeführt (ROSENBAUM, 2002), um die Robustheit der Schätzergebnisse gegenüber *hidden bias* aufgrund von *unobserved heterogeneity* zu überprüfen. Für die Berechnung der *Rosenbaum Bounds* wurde eine andere Matching-Technik verwendet. Alle Schätzungen wurden mit dem Software-Paket von Stata durchgeführt.

3.2.2 Ergebnisse

Die Ergebnisse der einzelnen Evaluierungsmethoden sind in Tabelle 1 unten aufgeführt. Die Ergebnisse der Logitschätzung, auf deren Grundlage das Matching erfolgte und die DID-Matching-Schätzer berechnet wurden, sind in Tabelle A1 im Anhang aufgeführt.

Wie aus Tabelle 1 zu ersehen ist, ergeben sich für alle Performance-Indikatoren erhebliche Abweichungen zwischen den EU-Standardverfahren und den korrespondierenden Matching-Verfahren. Insbesondere erkennt man aus Tabelle 1, dass sich ein starker Selektionseffekt für die SAPARD-Teilnahme ergibt. Man beachte, dass sich die jeweiligen Mittelwerte der Teilnehmer und Nichtteilnehmer erheblich unterscheiden für die „nicht gematchten" und die korrespondierenden „gematchten" Gruppen.

Entsprechend führen die einfachen Vergleiche der Performance-Indikatoren zwischen nicht gematchten SAPARD-Teilnehmern und Nichtteilnehmern zu einer klaren Fehleinschätzung des Effektes von SAPARD-Programmen. Dies gilt selbst für das einfache DID-Verfahren ohne Matching. Beispielsweise ergibt sich für die Arbeitsproduktivität ein positiver SAPARD-Programmeffekt von 80 bzw. 62 SKK pro dauerhaft Beschäftigten für die einfache DID-Methode, während die ATT des DID-Matching-Verfahrens einen negativen Effekt des SAPARD-Programms aufweisen in der Höhe von –66 SKK pro dauerhaft Beschäftigten.

Hinsichtlich der beiden anderen Performance-Indikatoren, Beschäftigung je Betrieb und Gewinn pro Hektar, führen die EU-Standardverfahren und die DID-Matching-Verfahren zwar zu qualitativ gleichen Ergebnissen, wobei sich ein positiver Effekt hinsichtlich der Beschäftigung und ein negativer Effekt für den Gewinn ergibt. Allerdings ergeben sich auch für diese Performance-Indikatoren erhebliche quantitative Unterschiede. So führt das DID-Matching-Verfahren zu einer um 112% höheren Beschäftigungssteigerung, während die Teilnahme am SAPARD-Programm zu einer um mehr als 500% stärkeren Absenkung des Gewinns je Hektar nach dem DID-Matching-Verfahren im Vergleich zu dem einfachen EU-DID-Evaluierungsverfahren ohne Matching führt.

Allgemein nimmt die Güte der Evaluierung von einem einfachen *Before–and-After*-Vergleich für die Gruppe der SAPARD-Teilnehmer (dY-Spalte in Tabelle 1) über die Berechnung von Differenzen der BAE für Teilnehmer und Nichtteilnehmer (DID-Spalte in Tabelle 1) zu, da letztere, wie oben bereits ausgeführt, den Einfluss anderer Faktoren zumindest teilweise ausschließt. Optimal wird der Selektionsbias schließlich bei dem DID-Matching-Verfahren berücksichtigt. Dabei unterscheiden sich die ATT für das *Rosenbaum-Bound*-Verfahren (DID^{MAT}_{RB}) geringfügig von dem ATT des DID-Matching-Verfahrens (DID^{MAT}), da bei dem ersteren ein andere Matching-Technik (Nearest Neighbour anstatt Kernel-Matching) verwendet wird. Allgemein zeigen die durchgeführten Sensitivitätsanalysen mit Hilfe

[4] Tatsächlich wurden im Rahmen von Advanced-Eval alle SAPARD-Maßnahmen in der Slowakei wie auch in Polen mit Hilfe unterschiedlicher Propensity-Score-Matching-Verfahren wie auch der Standard-EU-Verfahren evaluiert (siehe MICHALEK, 2007).

[5] Tatsächlich wurden entsprechende Evaluierungen auch für eine Reihe weiterer Indikatoren berechnet, siehe MICHALEK und HENNING, 2008.

Tabelle 1. Vergleich der Evaluierungsergebnisse von SAPARD zwischen EU- und Matching-Verfahren

	Gewinn				Arbeitsproduktivität				Beschäftigung			
	Y			ΔY	Y			ΔY	Y			ΔY
EU-Standard	2001	2003	2005		2001	2003	2005		2001	2003	2005	
Supported (D=1)	1,1	-0,46	0,908	1,368	222	235	326	38%	101	100	106	6
Non-supported (D=0)	0,403	-2,32	-0,51	1,81	167	140	151	8%	90	80	77	-3
Average Ø	0,6	-1,94	-0,209	1,731	182	161	190	18%	93	84	84	0
DiD-Matching												0
Supported (D=1)	1,07	-0,46	0,27	0,73	224	238	247	4%	100,8	87,6	97,5	9,9
D=0	0,92	-4,46	-0,07	4,39	203	116	191	65%	112,9	82,4	75,4	-7

	Gewinn				Arbeitsproduktivität				Beschäftigung			
	CSE			DID	CSE			DID	CSE			DID
EU-Standard	2001	2003	2005		2001	2003	2005		2001	2003	2005	
Difference (1-0)	0,7	1,92	1,47	-0.48	54	95	175	80	11,8	20,4	28,3	7,9
Difference (1-Ø)	0,5	1,49	1.117	-0.37	40	74	136	62	8	16	22	6
DiD-Matching												
ATT^{MAT}	0,15	4.00*	0,97	-3.04	20,4	122*	56*	-66	-12,15	5,2	22	16,8
ATT^{MAT}_{RB}	0,18	3.77*	0.94*	-2,83	11,6	103*	51*	-52	-13,3	-0,58	12,7	13,28

$\Delta Y = Y_{2005} - Y_{2003,}$ $DID = CSE_{2005} - CSE_{2003}$

Quelle: eigene Berechnungen

der *Rosenbaum Bounds,* dass die DID-Matching-Ergebnisse stabil sind[6].

4. Zusammenfassung und Ausblick

Eine konsistente Evaluierung von Politikprogrammen ist hinsichtlich einer effizienten Politikformulierung, die eine effiziente Nutzung von Steuermitteln für die Bereitstellung relevanter öffentlicher Güter erlaubt, Grundvorausetzung. Insofern ist die zunehmende Bedeutung, die der Politikevaluierung in der EU wie auch in internationalen Institutionen wie der Weltbank, der OECD oder der FAO beigemessen wird, verständlich und folgerichtig. Allerdings konnte in diesem Beitrag aufgezeigt werden, dass eine fundierte Evaluierung ländlicher Entwicklungspolitik eine komplexe methodische Herausforderung darstellt, die deutlich über die bisherigen Kosten-Nutzen-Analysen der Preis- und Marktpolitik hinausgeht. Eine konsistente Evaluierung umfasst dabei drei Ebenen: Evaluierung der individuellen Effekte auf der Mikroebene, Aggregation der individuellen Effekte auf der Makroebene und schließlich die Kosten-Nutzen-Analyse. Neben einem konsistenten Evaluierungsrahmen, der bereits von der EU-Kommission entwickelt worden ist, ergibt sich das fundamentale methodische Evaluierungsproblem aus der Tatsache, dass die Performance eines Wirtschaftsakteurs oder einer Menge an Wirtschaftsakteuren grundsätzlich nicht simultan für den Fall einer Teilnahme und Nichtteilnahme an einem Politikprogramm beobachtet werden kann. Somit muss die jeweils nicht beobachtete Performance kontrafaktorisch aus der entsprechenden beobachteten Performance einer Kontrollgruppe abgeleitet werden. Hierbei ergibt sich in der Regel ein Selektionsproblem, da die Verteilung relevanter Faktoren, die die Performance beeinflussen, nicht identisch in der Teilnehmergruppe und entsprechenden Kontrollgruppe verteilt ist. Eine Möglichkeit, dieses Selektionsproblem statistisch zu korrigieren, bieten die so genannten Propensity-Score-Matching-Verfahren. Die grundlegenden Ansätze dieser Verfahren wurden im Rahmen eines Europäischen Forschungsprojektes Advanced-Eval weiterentwickelt und zur Evaluierung ländlicher Entwicklungspolitik angewendet. In diesem Beitrag werden diese entwickelten Ansätze dargestellt, und es wurde anhand der SAPARD-Daten der Slowakei ein empirischer Vergleich der Matching-Verfahren und der bislang in der EU praktizierten einfachen Evaluationsverfahren durchgeführt. Dieser Vergleich ergab eindeutig, dass die einfachen Verfahren zu einer stark verzerrten Evaluierung von SAPARD-Programmen führen. Dabei weichen die Evaluierungsergebnisse nicht nur quantitativ, sondern zum Teil sogar qualitativ ab, d.h. es ergeben sich unterschiedliche Vorzeichen hinsichtlich der Programmwirkung. Insofern unterstreichen die empirischen Ergebnisse die Bedeutung adäquater mikroökonometrischer Evaluierungsverfahren für eine konsistente Bewertung ländlicher Entwicklungspolitik wie auch für ein effektives *Policy Learning*, d.h. eine kontinuierliche Verbesserung und Entwicklung effizienter Maßnahmen und Programme.

Allerdings beschränkt sich dieses Papier auf mikroökonometrische Evaluierungsansätze, während eine umfassende Politikevaluierung auch Effekte auf der Makroebene sowie eine Kosten-Nutzen-Analyse umfasst. Die Ermittlung von Politikeffekten auf der Makroebene beinhaltet dabei zusätzliche Komplikationen. Insbesondere bewirken entsprechende Interaktionen und Interdependenzen zwischen den individuellen Wirtschaftsakteuren, dass eine einfache Hochrechnung individueller Politikeffekte zu verzerrten Ergebnissen führt. Interessanterweise können zur empirischen Ermittlung unverzerrter makroökonomischer Effekte ländlicher Entwicklungsprogramme grundsätzlich auch die hier präsentierten Matching-Verfahren eingesetzt werden. Erste methodische Verfahren zur Ermittlung makroökonomischer Effekte ländlicher Entwicklungspolitiken wurden ebenfalls im Rahmen von Advanced-Eval entwickelt (vgl. MICHALEK,

[6] Die detaillierten Ergebnisse der Sensitivitätsanalysen mit Hilfe der *Rosenbaum Bounds* sind bei Bedarf von den Autoren erhältlich.

2008). Allerdings geht eine detaillierte Darstellung dieser Verfahren über den anvisierten Rahmen dieses Beitrages hinaus, so dass wir an dieser Stelle den interessierten Leser auf unsere laufenden Forschungsarbeiten (MICHALEK und HENNING, 2008) verweisen.

Literatur

BLUNDELL, R. and M. COSTA DIAS (2002): Alternative Approaches to Evaluation in Empirical Microeconomics. In: Portuguese Economic Journal (1): 91-115.

CALIENDO, M. (2006): Microeconometric evaluation of labor market policies. Springer Verlag, Berlin.

COCHRANE, W. and D. RUBIN (1973): Controlling Bias in Observational Studies. In: Sankyha 35(4): 417-446.

DAVIES, R. and S. KIM (2004): Matching and the Estimated Impact of Interlisting. Working Paper. University of Reading.

DEHEJIA, R. H. and S. WAHBA (1999): Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs. In: Journal of the American Statistical Association, Dec 1999, 94 (448), 1053-1062.

– (2002): Propensity Score Matching Methods for Nonexperimental Causal Studies. In: The Review of Economics and Statistics 84 (1): 151-161.

EUROPEAN COMMISSION DG AGRI (2006): Guidelines for the midterm evaluation of rural development programmes funded by SAPARD 2007-2013.

EUROPEAN COMMISSION DG BUDGET (2004): A practical guide for the commission services. Brussels.

EU-MINISTERRAT (2005): Verordnung (EC) Nr. 1698/2005. Brüssel.

FISHER, R. (1935): Design of Experiments. Hafner, New York.

HAGEN, T. and A. SPERMANN (2004): Hartz-Gesetze – Methodische Ansätze zu einer Evaluierung. Nomos-Verlag, Baden-Baden.

HECKMAN, J. (1997): Instrumental Variables – A Study of the Implicit Behavioral Assumptions Used in Making Program Evaluations. In: The Journal of Human Resources 32 (3): 441-462.

HECKMAN, J. and J. HOTZ (1989): Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training. In: Journal of the American Statistical Association 84 (408): 862-874.

HECKMAN, J., H. ICHIMURA and P. TODD (1997): Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme. In: Review of Economic Studies 64 (4): 605-654.

– (1998): Matching as an Econometric Evaluation Estimator. In: Review of Economic Studies 65 (2): 261-294.

HECKMAN, J., R. LALONDE and J. SMITH (1999): The Economics and Econometrics of Active Labor Market Programs. In: Ashenfelter, O. and D. Card (eds.): Handbook of Labor Economics - Vol. III: 1865-2097. Elsevier, Amsterdam.

HECKMAN, J. and R. ROBB (1985): Alternative Methods for Evaluating the Impact of Interventions. In: Heckman, J. and B. Singer (eds.): Longitudinal Analysis of Labor Market Data: 156-245. Cambridge University Press.

HECKMAN, J. J., J. SMITH and N. CLEMENTS (1997): Making the Most out of Programme Evaluations and Social Experiments: Accounting for Hererogeneity in Programme Impacts. In: The Review of Economic Studies 64 (4): 487-535.

HOLLAND, P. (1986): Statistics and Causal Inference. In: Journal of the American Statistical Association 81 (396): 945–960.

IMBENS, G. (2004): Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review. In: The Review of Economics and Statistics 86 (1): 4-29.

IMBENS, G. and J. ANGRIST (1994): Identification and Estimation of Local Average Treatment Effects. In: Econometrica 62 (2): 467-490.

LEUVEN, E. and B. SIANESI (2003): PSMATCH2: Stata Module to Perform Full Mahalanobis and Propensity Score Matching, Common Support Graphing, and Covariate Imbalance Testing. Software, http://ideas.repec.org/c/boc/bocode/s432001.html.

MICHALEK, J. (2007): Quantitative Tools for ex-post evaluation of RD Programmes. Conceptual Report. Universität Kiel.

- (2008): Construction and measurement of rural development index. Applications to evaluation of RD policies. Conceptual Report. Universität Kiel.

MICHALEK, J. and C.H.C.A. HENNING (2008): Periodic Activity Report. Deliverable Periodic Report. Universität Kiel.

NEYMANN, J. (1935): Statistical Problems in Agricultural Experiments. In: The Journal of the Royal Statistical Society 2 (2): 107-180).

OECD (2002): Employment Outlook. Paris.

– (2004): Employment Outlook. Paris.

PUFAHL, A. und C. WEISS (2007): Evaluating the Effects of Farm Programs. Results from Propensity Score. Working Paper No. 13. Vienna University of Economics & BA.

ROSENBAUM, P. and D. RUBIN (1983): The Central Role of the Propensity Score in Observational Studies for Causal Effects. In: Biometrika 70 (1): 41-50.

ROSENBAUM, P.R. (2002): Observational Studies. Springer, New York.

ROY, A. (1951): Some Thoughts on the Distribution of Earnings. In: Oxford Economic Papers 3 (2): 135-145.

RUBIN, D. (1974): Estimating Causal Effects to Treatments in Randomised and Nonrandomised Studies. In: Journal of Educational Psychology 66: 688-701.

SCHMIDT, C. (1999): Knowing What Works – The Case for Rigorous Program Evaluation. Discussion Paper No. 77, Institute for the Study of Labor (IZA), Bonn.

SMITH, J. and P. TODD (2005): Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators? In: Journal of Econometrics 125 (1-2): 305-353.

TOULEMONDE, J., H. SUMMA and N. USHER (2002): Three layers of quality assurance: would this help provide EU policy makers with the evaluative information they need? The 2002 European Evaluation Society Conference, 10-12 October 2002.

Danksagung

Diese Arbeit entstand im Rahmen des EU-Forschungsprojekts Advanced-Eval (EU-022708). Die Autoren danken der EU für die Förderung der Forschungsarbeiten.

Kontaktautor:
PROF. DR. DR. CHRISTIAN HENNING
Institut für Agrarökonomie, Christian-Albrechts-Universität Kiel
Olshausenstr. 40-60, 24098 Kiel
Tel.: 04 31-880 44 53, Fax: 04 31-880 13 97
E-Mail: chenning@agric-econ.uni-kiel.de

Tabelle A1: Ergebnisse der Logitfunktionsschätzung

Variable	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
v1	.0010399	.0009505	1,09	0,274	-.0008231	.002903
v2	.0005844	.0001957	2,99	0,003	.0002009	.0009679
v3	-.0046042	.001309	-3,52	0	-.0071699	-.0020385
v4	-.0041709	.0013684	-3,05	0,002	-.006853	-.0014887
v5	.0002302	.0001432	1,61	0,108	-.0000504	.0005108
v6	-.0121612	.0049346	-2,46	0,014	-.0218328	-.0024897
v7	-.0035127	.0009677	-3,63	0	-.0054094	-.0016159
v8	.0016618	.0010218	1,63	0,104	-.0003409	.0036646
v9	-.2086804	.1039418	-2,01	0,045	-.4124027	-.0049582
v10	.0397408	.0222323	1,79	0,074	-.0038338	.0833153
v11	.0765484	.0574285	1,33	0,183	-.0360095	.1891062
v12	-.0398741	.0123417	-3,23	0,001	-.0640634	-.0156849
v13	.0015348	.0008284	1,85	0,064	-.0000888	.0031584
v14	-.0086651	.0031065	-2,79	0,005	-.0147538	-.0025764
v15	.0000276	.0000297	0,93	0,353	-.0000306	.0000858
v16	-.0779033	.050257	-1,55	0,121	-.1764053	.0205986
v17	.0001235	.0000782	1,58	0,114	-.0000297	.0002768
v18	.015079	.0048949	3,08	0,002	.0054853	.0246727
v19	-.0037461	.0018089	-2,07	0,038	-.0072916	-.0002006
v20	-.0072734	.0020818	-3,49	0	-.0113535	-.0031932
v21	-.0006416	.00029	-2,21	0,027	-.0012099	-.0000732
v22	.0061768	.002315	2,67	0,008	.0016394	.0107141
v23	.0003141	.0001875	1,68	0,094	-.0000533	.0006816
v24	.0007263	.0003638	2	0,046	.0000132	.0014394
v25	-.0838564	.0262439	-3,2	0,001	-.1352936	-.0324193
v26	-.0157825	.0072258	-2,18	0,029	-.0299448	-.0016202
v27	-.0023481	.0008059	-2,91	0,004	-.0039277	-.0007686
v28	-.0027955	.0015564	-1,8	0,072	-.005846	.000255
v29	-1.916601	.8770684	-2,19	0,029	-3.635624	-.19755787
v30	.0052822	.0026352	2	0,045	.0001174	.0104471
v31	-.0000109	.0000266	-0,41	0,681	-.0000631	.0000412
v32	.0003355	.0001033	3,25	0,001	.000133	.000538
v33	-.0082926	.0075219	-1,1	0,27	-.0230352	.00645
v34	-.0000159	.0006538	-0,02	0,981	-.0012974	.0012657
v35	.0003626	.0008618	0,42	0,674	-.0013266	.0020518
v36	-.0006339	.0002719	-2,33	0,02	-.0011668	-.000101
v37	.0225985	.010272	2,2	0,028	.0024657	.0427312
v38	.0120376	.0136046	0,88	0,376	-.0146269	.038702
v39	-.0162642	.0301497	-0,54	0,59	-.0753566	.0428281
v40	-3,517304	.9051905	-3,89	0	-5.291445	-1.743164

v1 = number of pigs for fattening, v2 = liabilities, v3 = initial stock wheat, v4 = costs of interest paid, v5 = amortization, v6 = stock of other sheep, v7 = assets total non-current receivables, v8 = costs of purchased feeding stuff, v9 = employment other manual workers, v10 = initial stock beans, peas, etc., v11 = employment tractor drivers and mechanics, v12 = production of oat, v13 = initial stock grass and hay in haylage, v14 = overhead costs (water), v15 = costs of interest and fees (total), v16 = net value of current assets, v17 = costs of production (raw materials, energy, services), v18 = interest income, v19 = costs of heating fuel, v20 = value adjustment against operating expenses, v21 = revenue from sale of merchandise, v22 = social expenses, v23 = costs of own feedstuff for pigs, v24 = production of grass and hay, v25 = stock of heifers and bulls for fattening (6-12 months), v26 = stock of bulls for fattening (1-2 years), v27 = area hired from others (grass land and pasture), v28 = costs of cars, v29 = employment (directors, chairman, representatives, etc.), v30 = production of industrial potatoes for starch, v31 = total liabilities (foreign sources), v32 = other operating income, v33 = production of grapes for wine, v34 = cost of consulting and service, v35 = cost of consumption of own seeds, v36 = cost of concentrates/feedstuff for livestock, v37 = number of breeding heifers > 2 years old, v38 = initial stock of oat, v39 = employment (other employees).

Quelle: eigene Berechnung