Quantitative Methoden der Internationalen Beziehungen

Constantin Ruhe[1], Gerald Schneider[2]/Gabriele Spilker[3]

Beitrag für *Handbuch der Internationalen Beziehungen*, Hrsg. Frank Sauer und Carlo Masala, Wiesbaden: Springer VS. (2. Auflage 2014)

[1] Doktorand, Graduate School of Decision Science, Fach 86, Universität Konstanz, Email: Constantin.Ruhe@uni-konstanz.de
[2] Ordinarius für Internationale Politik, geschäftsführender Herausgeber „European Union Politics“ und Ko-Herausgeber „International Interactions“, Fachbereich für Politik- und Verwaltungswissenschaft und Graduate School of Decision Sciences, Fach 86, Universität Konstanz, Email: Gerald.Schneider@uni-konstanz.de
[3] Senior Researcher ETH Zürich und Assistenzprofessorin Universität Salzburg, Email: gabriele.spilker@ir.gess.ethz.ch

1. Einleitung: Grundlagen der quantitativen Analyse [1]

Eines der Grundmerkmale wissenschaftlichen Arbeitens besteht darin, Informationen so zu verdichten, dass eine Struktur erkennbar ist. Ohne Theorien über den Prozess, in dem die Daten entstanden sind, bleibt jeder Forscher blind. Doch mit Theorien allein ist es in den angewandten Wissenschaften nicht getan. Ob die Erklärung zu den Daten passt, lässt sich nur mit Hilfe von Methoden prüfen, die zugleich transparent und unabhängig von der Theorie sind, die es zu überprüfen gilt.

Wenn eine theoriegeleitete Wissenschaftlerin ungeeignete Forschungsdesigns und Methoden verwendet, um die Datenstruktur zu beschreiben, kann ihr Vorhaben aufgrund dreier Probleme scheitern: der Storch-bringt-Kinder-Illusion, des Fata Morgana-Trugschlusses, sowie der Vernebelungsgefahr. Die erste Tücke besteht darin, dass ein ungeeignetes Forschungsdesign die Wissenschaftler nicht befähigt, eine Scheinkorrelation von einem kausalen Zusammenhang zu unterscheiden. Nur mit Hilfe solider Forschungsansätze lässt sich eine beobachtete Wirkung eindeutig auf eine (theoretisch einleuchtende) Ursache zurückführen. Die zweite Schwierigkeit besteht darin, dass die Methode den Effekt einer Ursache so überschätzt, dass der Forscher sich in falscher Sicherheit wiegt. Die dritte Folge einer ungeeigneten Methodenwahl ist umgekehrt, dass die Wirkungen systematisch unterschätzt werden. Die Auswahl des Forschungsdesigns und der Methode ist deshalb mitentscheidend, um schlechte und gute Forschung, wahre und falsche Erkenntnisse unterscheiden zu können. Generell gibt es drei Kriterien, um die Angemessenheit eines jeden Forschungsdesigns zu beurteilen, unabhängig davon, ob das Verfahren „quantitativer“ oder „qualitativer“ Natur ist. Diese Prüfsteine sind alle mit der wissenschaftlichen Kernerfordernis der Gültigkeit verknüpft und lassen sich über die Zusätze „interne“, „externe“ sowie

[1] Dieses Kapitel ist im Vergleich zur ersten Fassung (Schneider und Ruoff 2010) vollständig überarbeitet und deutlich erweitert worden. Die Autorenreihenfolge ist alphabetisch; alle Beteiligten haben gleichermaßen zu diesem Aufsatz beigetragen. Wir danken den Herausgebern für Hinweise. Für die Erarbeitung der ersten Fassung haben Schneider und Spilker von der Hilfe durch Glenn Palmer (University Park, PA) und J. David Singer (Ann Arbor, MI) bei der Bereitstellung der Daten sowie Aurelio Tobias (Madrid) für seine Übersendung eines STATA-Programms zur Schätzung eines Zeitreihenpoissonmodells profitiert.

„statistische“ Validität definieren (Shadish et al. 2002). [2] Ein Forschungsdesign ist erstens intern gültig, wenn das Design es ermöglicht, den realen Wirkungszusammenhang zu identifizieren. Externe Validität liegt zweitens vor, wenn die Ergebnisse auch über die jeweilige Untersuchung hinaus generalisierbar und replizierbar sind. Drittens besagt das Kriterium der statistischen Gültigkeit, dass die gewählte Analysemethode die Größe des identifizierten Wirkungszusammenhangs zuverlässig bestimmen soll. So muss ein geschätzter Zusammenhang („der Schätzer“) zum einen unverzerrt sein, was sich auch als Erwartungstreue oder Absenz einer Verzerrung (Bias) bezeichnen lässt. Zum anderen sollten die Resultate, die ein Verfahren bei Replikationen erbringt, eine geringe Varianz aufweisen und somit „effizient“ sein (King, Keohane und Verba 1994). Da das Unverzerrtheitskriterium nicht für alle Schätzer zu erreichen ist, gilt als asymptotischer Ersatzmaßstab die Konsistenz eines Schätzers. Bei einem konsistenten Schätzer nähert sich bei steigender Stichprobenfallzahl der Parameter, der geschätzt wird, dem wahren Wert an.

Die meisten methodologischen Diskussionen in der Politikwissenschaft seit den 1990er Jahren drehen sich um diese Kriterien. Im Bereich der statistischen Validität ist die Kompetenz der einschlägigen Methodenlehre dabei so weit vorangeschritten, dass Politologen mittlerweile selbständig effiziente, erwartungstreue Schätzverfahren zu entwickeln versuchen. Obwohl die fachliche Diskussion anspruchsvoll ist, erreicht sie einen immer größer werdenden Kollegenkreis. Im Zuge dieser Entwicklung hat sich auch die Diskussion in den Internationalen Beziehungen intensiviert. Dies ist anhand der einschlägigen Veröffentlichungen in der bislang einzigen Methodenzeitschrift, dem Quartalsheft *Political Analysis,* und den Spitzenjournalen der Disziplin ersichtlich. Einige der methodischen Neuerungen und der Debatten darum greifen wir hier auf.

[2] Shadish et al. (2002) definieren zudem die Konstruktvalidität als zentrales Kriterium. Dieses Kriterium besagt, wie weit ein Test ein für eine Untersuchung wichtiges Phänomen misst, so dass diese Messung der Konstruktdefinition entspricht. Obwohl Messprobleme nach unserem Erachten eine zu geringe Beachtung in den Internationalen Beziehungen erfahren, beschränken wir die Diskussion hier auf die Inferenzprobleme, die die drei anderen Validitätskriterien aufwerfen.

Während sich der Methodendiskurs der frühen 2000er Jahre vor allem um die Einführung neuer Schätzer drehte, hat sich in den vergangenen Jahren die Diskussion verstärkt auf die interne Validität von Forschungsdesigns und damit die Storch-bringt-Kinder Illusion fokussiert. Ziel dieser jüngeren Bestrebungen ist es, dass die gewählte Untersuchungsform erlaubt kausale Rückschlüsse zu ziehen und damit einen realen Wirkungszusammenhang zu identifizieren. Im Idealfall ermöglicht ein elaboriertes Forschungsdesign somit, dass ein Wirkungszusammenhang von alternativen Erklärungen isoliert wird, so dass die Ergebnisse einer Untersuchung mittels simpelster Verfahren, wie beispielsweise Mittelwertdifferenzen, ausgewertet werden können. Als Maßstab für eine Untersuchungsform, die diesem Ideal entspricht, gilt typischerweise die experimentelle Forschung. Interne Validität ist dann gegeben, wenn eine Wissenschaftlerin die Wirkung eindeutig auf die vermutete Ursache zurückführen kann, in dem sie sämtliche Störvariablen ausblendet (Diekmann 2007). Ein Experiment ist daher typischerweise durch eine hohe interne Validität gekennzeichnet, da die Wissenschaftlerin durch die Manipulation der Ursache eine klare Wirkungskette nachweisen kann. Wenn die Wissenschaftlerin nur über Beobachtungsdaten verfügt, d.h. Daten, bei denen sie die Ausprägung der mutmaßlichen Ursache nicht steuern kann, ist dies nicht möglich. Natürlich ist es den Forschern für die meisten Fragestellungen der Internationalen Beziehungen verwehrt, diese sog. Datengenerierung zu kontrollieren. In dieser Situation ergeben sich je nach Kontext verschiedene Alternativen, die die Wissenschaftlerin in unterschiedlichem Maße dem Ziel interner Validität näher bringen. Zusätzlich gilt zu bedenken, dass manches Experiment sowie einige quasi-experimentelle Forschungsdesigns jedoch teilweise die Generalisierbarkeit der Schlussfolgerungen auf bestimmte Teile der Grundgesamtheit begrenzen und somit die externe Validität der Ergebnisse einschränken.
In diesem Übersichtsartikel wollen wir zunächst ganz praktisch die Kriterien der internen, externen und statistischen Validität verdeutlichen und auf mögliche Forschungsansätze und Methoden verweisen. Anschließend gehen wir auf einige praktische Herausforderungen der

Datenanalyse sowie neuere Entwicklungen ein, die in den letzten Jahren die internationale Spitzenforschung in den Internationalen Beziehungen stark beeinflusst haben.

2. Methoden für experimentelle und Beobachtungsdaten

Kernherausforderung in den angewandten Sozialwissenschaften ist in jedem Fall das Streben nach kausaler Inferenz. Darunter ist zumindest nach dem kontrafaktischen Ansatz (potential outcomes framework) von Neyman-Rubin (Neyman 1923, Rubin 1971) die Identifikation eines kausalen Effektes zu verstehen, in dem die Wirkung - gegeben einer Bedingung - mit dem Nicht-Effekt in der Absenz dieser Bedingung verglichen wird. Ein Grundproblem der empirischen Forschung besteht angesichts dieses kontrafaktischen Ideals darin, dass kein Untersuchungsobjekt jemals gleichzeitig mit zwei verschiedenen Ausprägungen einer Ursache beobachtet werden kann. Dies hat zur Folge, dass eine Wissenschaftlerin in keinem Falle empirisch ein Untersuchungsobjekt mit sich selbst, zum gleichen Zeitpunkt, aber mit veränderter Ursache vergleichen kann (vgl. Holland 1986, für eine Einführung siehe Sekhon 2008).

Dieses fundamentale Problem der kausalen Inferenz wird deutlicher, wenn man es im Lichte einer historischen, politikwissenschaftlichen Fragestellung betrachtet. Als Beispiel kann die klassische, von Waltz (1979) im Rahmen des Strukturellen Realismus immer wieder beschworene Frage gelten, ob in einer bipolare Weltordnung weniger Konflikte entstehen als in andersartigen Polaritätsstrukturen wie Uni- oder Multipolarität. Das fundamentale Problem besteht nun darin, dass eine Wissenschaftlerin die Welt im Jahre 1970 nur einmal und mit nur einer Ausprägung der Ursache „Weltordnung“ beobachten kann. Empirisch ist es unmöglich, die Welt im Jahre 1970 mit sich selber, jedoch mit einer veränderten Weltordnung zu vergleichen. Ein solcher Vergleich ist rein kontrafaktisch. Als einzige Möglichkeit einer empirischen Untersuchung bleibt der Vergleich verschiedener Ausprägungen der Weltordnung zu unterschiedlichen Zeitpunkten.

Ganz ähnlich verhält es sich mit Ursachen, die zwar zeitgleich beobachtbar sind, jedoch in diesem Fall bei zwei unterschiedlichen Untersuchungsobjekten auftreten. So kann bei der Erforschung des Effekts von Demokratie auf das Kriegsrisiko zwar zeitgleich für demokratische und autokratische Länder das Kriegsrisiko beobachtet werden, jedoch nur für unterschiedliche Länder.

Unabhängig davon, ob eine Untersuchung das gleiche Untersuchungsobjekt an verschiedenen Zeitpunkten mit sich selber oder verschiedene Untersuchungsobjekte zum gleichen Zeitpunkt miteinander vergleicht - in jedem dieser idealtypischen Forschungsdesigns besteht die Gefahr, dass ein beobachteter Unterschied nicht durch die postulierte Ursache ausgelöst wurde, sondern auf andere Unähnlichkeiten zurückzuführen ist. Eine Schlussfolgerung, dass die beobachtete Differenz der Effekt der Ursache ist, wäre in diesem Falle nicht korrekt und die Ergebnisse somit nicht intern valide. Bei Ursache-Wirkungs-Fragen ist somit das Ziel jeglichen quantitativen Forschungsdesigns und jeder Methode, alternative Erklärungen so weit wie möglich auszuschließen und somit interne Validität sicherzustellen.[3]

2.1 Interne Validität durch Forschungsdesigns: Die Analyse experimenteller Daten

Kann die Wissenschaftlerin die vermutete Ursache manipulieren, so können alternative Erklärungen plausibel ausgeschlossen werden. Ist diese Möglichkeit gegeben, so kann ein Experiment entworfen werden, in dem verschiedene Untersuchungsobjekte zufällig einer Ausprägung der Ursache, auch Treatment genannt, ausgesetzt sind. Durch die zufällige Zuteilung eines Treatments (Randomisierung) kann bei einer großen Stichprobe sichergestellt werden, dass das Treatment der einzige systematische Unterschied zwischen den Untersuchungsgruppen ist. Somit sollten Differenzen bei den beobachteten Effekten zwischen den verschiedenen Gruppen alleine auf die manipulierte Ursache zurückzuführen sein.

[3] Reine Prognosemodelle, auf die wir im dritten Unterkapitel näher eingehen, bilden eine Ausnahme.

Sozialwissenschaftliche Experimente sind nicht ausschließlich der Psychologie oder der Verhaltensökonomie vorbehalten und können auch auf Fragen der Internationalen Beziehungen angewendet werden. Beispielsweise können Laborexperimente den Effekt von Todesopfern auf die Zustimmung zu Kriegseinsätzen untersuchen, in dem Probanden manipulierte Prognosen präsentiert werden (vgl. Gartner 2008). Ebenso ist es möglich Fragen der Abschreckungstheorie in Laborsituation zu simuliert und zu untersuchen. So zeigen etwa Blendin und Schneider (2012) in einer Computersimulation, dass Zeitdruck im Sinne der Groupthink-Thesen und ähnlicher theoretischer Erwartungen der Politischen Psychologie die Entscheidungsqualität mindert, nicht jedoch das Erzeugen von Stress über Kortisoltabletten. Biologische Ursachen des aggressiven Verhaltens in einem Krisenspiel glauben etwa McDermott et al. (2007) identifiziert zu haben, die ihren Probanden nach unterschiedlichem Testosteronniveau einteilen. Ob biologische oder auch neuronale Treatments eine vorübergehende Modeerscheinung der experimentellen Politikwissenschaft sind oder unser Verständnis von Entscheidungsprozessen in der internationalen Politik grundlegend verändern können, ist noch nicht abzusehen. Bedenkenswert scheint auf alle Fälle der Nachweis in verschiedenen Studien, dass verschiedene Versionen des „Kriegergens“ monoamine oxidase-A das Konfliktverhalten in bestimmten Experimentalsituationen befördern (McDermott et al. 2013).

Laborexperimente zeichnen sich durch umfassende Kontrolle des Untersuchungskontexts aus, sind jedoch häufig abstrakt. Der hohen internen Validität steht somit eine in manchen Fällen fragliche externe Validität gegenüber. Feldexperimente, wie beispielsweise die experimentelle Untersuchung von Maßnahmen in der Entwicklungszusammenarbeit (vgl. Fearon et al. 2009), greifen diese Problematik auf und testen den Effekt von Ursachen in einer „natürlichen“ Umgebung. Dadurch verringert sich die Kontrolle über mögliche Störfaktoren und das Treatment. Jedoch ermöglichen diese Untersuchungen bei einer gegebenenfalls verringerten internen Validität eine Schätzung der Effekte in einem empirisch relevanten Umfeld.

Eine Möglichkeit interne Validität zu sichern, ohne gleichzeitig die externe Validität stark aufs Spiel zu setzen, sind Umfrageexperimente (Jensen, Mukherjee und Bernhard 2014, Morton und Williams 2010, Mutz 2011). Wie der Name schon vermuten lässt, wird bei einem Umfrageexperiment ein typischerweise informationsbasiertes Experiment in eine Umfrage integriert. Durch die randomisierte Zuteilung der Information auf die befragten Personen, zeichnen sich Umfrageexperimente im Normalfall durch eine starke interne Validität aus. Da Umfrageexperimente zudem häufig in national repräsentative Umfragen integriert werden, kann dadurch auch eine bessere externe Validität erreicht werden, als dies mit Laborexperimenten der Fall ist. So können beispielsweise Jensen und Shin (2014) zeigen, dass die Zustimmungsrate für Agrarsubventionen stark zunimmt, wenn die befragte Person zuvor die Information erhält, dass diese Art der Subvention weniger hoch ausfällt als in vergleichbaren Ländern. Indem sie den befragten Personen randomisierte Informationstreatments zukommen lassen, können Jensen und Shin damit erklären, warum Agrarsubventionen trotz ihres wohlfahrtsmindernden Charakters typischerweise eine hohe Unterstützungsrate in der Bevölkerung genießen. Allerdings zeigt dieses Beispiel auch die Grenzen von Umfrageexperimenten auf. Diese eignen sich hervorragend für die empirische Forschung, wenn der zu testende Kausalmechanismus, wie im obigen Beispiel, als Informationstreatment formuliert werden kann. Ein weiteres Beispiel, das die Möglichkeiten und Grenzen von Umfrageexperimenten verdeutlicht, ist die Frage, wie Regierungschefs internationale Drohungen in Krisensituation glaubhaft machen können, indem sie Publikumskosten (audience costs) kreieren (Tomz 2007). Lässt sich die Fragestellung nicht durch das Formulieren („framing“) von verschiedenen Informationstreatments beantworten, bieten Umfrageexperimente keinen Ausweg, die externe Validität zu erhöhen, so dass Forscherteams wie Blendin und Schneider (2012) zu einem klassischen Experiment Zuflucht nehmen müssen. Diese Grenzen gelten oft auch für die Überprüfung von Thesen, wie sie im Rahmen der Kollektivgütertheorie für Fragen der internationalen Zusammenarbeit entwickelt

werden. Hier dominieren zumeist klassische Experimente, auch wenn etwa Nobelpreisträgerin Elinor Ostrom (1990) ihre Thesen zu Allmendgütern auch durch Feldforschung illustrieren konnte.

Die Analyse von experimentellen Daten erfordert grundsätzlich keinerlei komplexe Methoden. In den meisten Fällen reicht eine Auswertung aus, die auf Mittelwertdifferenzen basiert. Obwohl t-Tests oder ähnliche nicht-parametrische Verfahren für die Auswertung genügen, kann ebenfalls eine Auswertung mittels Regressionsverfahren erfolgen. Welche Vorgehensweise gewählt wird, hängt maßgeblich von den Präferenzen der Wissenschaftlerin ab. Mittelwertdifferenzen benötigen keinerlei Annahmen hinsichtlich einer funktionalen Form, wie sie üblicherweise in Regressionsmodellen getroffen werden. Allerdings ermöglichen es Regressionsmodelle, Kontrollvariablen aufzunehmen. Diese zusätzlichen Variablen erklären einen Teil der Varianz in der abhängigen Variablen und gestatten so eine effizientere Schätzung eines Effekts. Zudem können Veränderungen im geschätzten Effekt bei der Hinzunahme von Kontrollvariablen ein Hinweis darauf sein, dass die Randomisierung nicht vollständig erfolgreich war. Eine weitere Möglichkeit zur Analyse experimenteller Daten bieten sogenannte Permutationstests. Diese berechnen exakte Signifikanztests und benötigen keinerlei Annahmen hinsichtlich der Stichprobenziehung (siehe Edington und Onghena 2007).

2.2 Die Analyse von Beobachtungsdaten

Viele Fragestellungen der Internationalen Beziehungen münden, wie angedeutet, in der Analyse von Beobachtungsdaten. So ist es auf der einen Seite nicht möglich, jegliche theoretische Erwartung experimentell zu überprüfen. Umgekehrt weisen etwaige Experimente aufgrund ihrer Abstraktion eine fragwürdige externe Validität auf. Den Wissenschaftlern bleibt somit nur die Möglichkeit, systematisch Informationen zu den interessierenden empirischen Phänomenen zu erheben und mit größtmöglicher Sorgfalt auszuwerten.

Für die Auswertung von Beobachtungsdaten steht eine Vielzahl von Methoden zur Verfügung. Einige dieser Methoden nutzen dabei natürliche Randomisierung, so dass Arbeiten, die sich auf solche Verfahren stützen, der internen Validität von Experimenten sehr nahe kommen. Ein Beispiel für ein solches „natürliches Experiment“ ist die Studie von Kern und Hainmüller (2009) zum Einfluss der westdeutschen Fernsehanstalten auf die Systemzufriedenheit in der Endphase der Deutschen Demokratischen Republik. Diese quasi-experimentellen Designs ermöglichen jedoch zumeist nur Schlussfolgerung für einen Teil der Untersuchungsobjekte und beruhen auf starken Annahmen. In der großen Mehrheit aller sozialwissenschaftlichen Studien lässt sich die Ursache nicht manipulieren, und es liegt auch keine natürliche Randomisierung vor. So bleibt einer Wissenschaftlerin nur der Versuch, alle möglichen alternativen Erklärungen auszuschließen und die hiermit verbundenen Störfaktoren in der Analyse zu berücksichtigen. Regressionsmodelle sind in der quantitativen Sozialforschung die verbreitetste Vorgehensweise, um in Beobachtungsstudien für Störfaktoren zu kontrollieren. Für Ursachen mit lediglich zwei Ausprägungen bieten sich alternativ sog. Matchingverfahren an.

Regressionsmodelle sind effiziente Schätzverfahren, die bei korrekter Implementierung Korrelationen zwischen Variablen nutzen und hohe Vorhersagekraft besitzen (Greene 2008). Werden Regressionsmodelle zur Überprüfung kausaler Ursache-Wirkungszusammenhänge verwendet, so ist jedoch eine Reihe von Annahmen zu beachten.[4] Die wohl prominenteste Annahme besagt, dass der Fehlerterm der Regression nicht mit der unabhängigen Variablen, also der postulierten Ursache, korreliert ist. Praktisch bedeutet dies, dass alle Variablen, die sowohl die unabhängige Variable als auch die abhängige Variable beeinflussen, beobachtet und in das Regressionsmodell aufgenommen werden müssen. Ist dies nicht der Fall, kann der geschätzte Effekt verfälscht sein. Dieses Problem wird allgemein als Verzerrung aufgrund

[4] Eine kritische Diskussion von Regressionsmodellen als Methode für Kausale Inferenz findet sich bei Morgan und Winship (2007).

weggelassener Variablen (*omitted variable bias)* bezeichnet.[5] Im schlimmsten Fall führt eine fehlende Variable zu einer Storch-bringt-Kinder-Illusion, in der eine Scheinkorrelation als solider wissenschaftlicher Befund fehlinterpretiert wird[6]. Häufiger kommt dieser Bias jedoch eher einer Fata Morgana gleich: ein Effekt existiert zwar, wirkt aber viel naheliegender und größer, als er eigentlich ist. In beiden Fällen leugnet eine Forscherin den Beitrag von alternativen Erklärungen zur Kovarianz mit der abhängigen Variablen, in dem sie diese ausschließlich der präferierten erklärenden Variablen zuschlägt, indem sie diese anderen Prädiktoren gar nicht erst ins Regressionsmodell aufnimmt. Verzerrte Schätzungen einfach durch willkürliches Hinzufügen von Kontrollvariablen mit der Realität in Einklang bringen zu wollen, ist jedoch nicht unproblematisch. Achen (2002, siehe auch Clarke 2005 und Schrodt 2014) weist darauf hin, dass eine allzu komplexe Modellierung nicht unbedenklich sei, da die zusätzlichen Regressoren wiederum das Ergebnis verfälschen können. Aus diesem Grund seien wenig komplexe, aber theoriegeleitete Schätzverfahren oft den theorielosen „Spülbeckenverfahren“ (kitchen sink regression) vorzuziehen.

Eine weitere Annahme besteht in der rigiden funktionalen Form, die den meisten Regressionsmodellen zugrunde liegt. So geht etwa ein lineares Regressionsmodell davon aus, dass eine unabhängige Variable einen linearen Effekt auf die abhängige Variable hat. Ein Logitmodel hingegen modelliert einen S-förmigen Effekt auf die Wahrscheinlichkeit eines binären Ereignisses. Zwar lässt sich die funktionale Form durch transformierte Variablen anpassen, die hierfür zu treffenden Annahmen sollten jedoch mit den theoretisch angenommenen Zusammenhängen übereinstimmen. Der Anwenderin von Regressionsmodellen sollte daher besonders bei komplexeren Modellen bewusst sein, dass

[5] Sind Daten für dieselben Untersuchungsobjekte über einen gewissen Zeitraum verfügbar, sogenannte Panelstudien, ermöglicht dies der Wissenschaftlerin durch Fixed-Effects Modelle den *omitted variable bias* durch konstante, unbeobachtbare Unterschiede zwischen Untersuchungseinheiten zu eliminieren. Somit sind verlässlichere Kausalaussagen möglich. Jedoch sind die Schätzungen dieser Modelle weniger effizient als gepoolte Modelle oder Random-Effects Modelle (Wooldridge 2010).

[6] Höfer, Przyrembel und Verleger (2004) diskutieren in einem humorvollen Beitrag den Zusammenhang zwischen Störchen und Geburten als Alternative zur Theorie der sexuellen Reproduktion und schlagen unter anderem vor: „Supporting the stork population by organic farming may have a positive influence on the low birth rate in most European countries, at least on deliveries outside hospitals“ (S. 91).

dieser Analyseansatz auf einem statistischen Modell basiert, welches explizit einen gewissen, datengenerierenden Prozess zugrunde legt. Eine gute theoretische Kenntnis dieses Prozesses ist daher unumgänglich.[7] Wird der wahre Prozess falsch modelliert, so sind die Ergebnisse voraussichtlich verzerrt. Im dritten Abschnitt dieses Beitrages gehen wir daher nochmals explizit auf praktische Probleme bei diesen Modellierungsentscheidungen ein.

Sogenannte Matchingverfahren greifen häufig auf Regressionsmodelle für binäre abhängige Variablen zurück, um in großen Stichproben gleiche oder zumindest ähnliche Untersuchungsobjekte mit zwei verschiedenen Ausprägungen der Ursache zu identifizieren (Morgan und Winship 2007). Matching ist somit der Fallstudienmethode eines *Most Similar Systems Design* sehr ähnlich, nur dass sehr viel mehr Fälle mit einander verglichen werden. Das Ziel ist es, zwei Gruppen von Untersuchungsobjekten zu kreieren, die sich nur in der Ausprägung der Ursache unterscheiden und somit einem Experiment ähneln. Sind genügend ähnliche Fälle identifiziert, so können die beobachteten Differenzen durch einfache Methoden wie t-Tests oder auch Regressionsmodelle ausgewertet werden. Aufgrund der Paarbildung machen Matchingverfahren in der eigentlichen Auswertung keine Annahmen hinsichtlich einer funktionalen Form und bieten so gegenüber Regressionsmodellen einen Vorteil.[8] Wie bei Regressionsverfahren beruht die Validität der Ergebnisse jedoch ebenfalls darauf, dass alle für eine Selbstselektion relevanten Einflussfaktoren beobachtet und im Matchingverfahren berücksichtigt wurden. Somit besteht auch bei Matchingverfahren die Gefahr eines *omitted variable bias*.[9]

In Instrumentalvariablenansätzen wird häufig ein möglicher Ausweg aus den Schwierigkeiten der Regressionsanalyse gesehen. Instrumentalvariablen ermöglichen in einigen Situationen

[7] Eine politikwissenschaftlich orientierte Einführung in die statistische Modellierung mittels Maximum Likelihood-Schätzung bietet King (1989).

[8] Allerdings basieren Matchingverfahren auf der Annahme, dass sämtliche Unterscheidungen zwischen den Fällen beobachtbar sind. Geht die Wissenschaftlerin davon aus, dass auch nicht zu beobachtende Variablen eine entscheidende Rolle in der Unterscheidung zwischen den Fällen spielen, muss auf sogenannte Selektionsmodelle zurückgegriffen werden (Heckman 1979, siehe dazu auch von Stein 2005 und Simmons und Hopkins 2005).

[9] Eine Einführung in Matchingverfahren bieten Morgan und Winship (2007).

die Schätzung eines Effekts, auch wenn die unabhängige Variable, hier D genannt, mit dem Fehlerterm korreliert ist. Hierzu muss eine Instrumentalvariable gefunden werden, die die unabhängige Variable D monoton beeinflusst, jedoch keinerlei direkten Einfluss auf die abhängig Variable hat. Der Kausalmechanismus, der von der Instrumentalvariablen über die unabhängige Variable D ausgeht, muss zudem der einzige Wirkungspfad sein, durch den die Instrumentalvariable die abhängige Variable beeinflusst. Diese Anforderung stellen eine große Hürde für die Anwendung von Instrumentalvariablen in den Internationalen Beziehungen dar.[10] Eine nach dem Zufallsprinzip auftretende Variable ist dabei am ehesten geeignet, um als Instrumentalvariable zu dienen. Häufig werden hierfür natürliche Experimente genutzt, in denen eine natürliche, nicht durch eine Wissenschaftlerin herbeigeführte Randomisierung vorliegt. Jedoch muss auch in diesen Fällen glaubhaft dargelegt werden, dass die Instrumentalvariable alle Annahmen erfüllt. Sind diese Bedingungen erfüllt, ist es möglich, den lokalen Effekt der Variable D auf die abhängige Variable zu schätzen, der auf den Einfluss der Instrumentalvariablen zurückzuführen ist. Regressions-Diskontinuitäts-Analysen sind ein weiterer Ansatz, der auf der Idee eines natürlichen Experiments beruht. Hierzu werden Schwellenwerte genutzt, die bestimmen, ob eine Untersuchungseinheit in eine bestimmte Kategorie fällt oder nicht. Beispielsweise bekommen nur solche europäischen Regionen EU-Mittel aus Strukturfonds, die eine Wirtschaftsleistung unterhalb eines bestimmten Schwellenwerts aufweisen. Die zentrale Annahme einer Regressions-Diskontinuitäts-Analyse besteht darin, dass Fälle in unmittelbarer Nähe des Schwellenwerts mit annähernd gleicher Wahrscheinlichkeit oberhalb oder unterhalb des Schwellenwerts hätten liegen können. Ist dies korrekt und sind diese Fälle auch in anderen relevanten Bereichen ähnlich, so kann ein lokaler Effekt für die Untersuchungseinheiten in unmittelbarer Nähe des Schwellenwerts identifiziert werden, der eben rein auf die postulierte Ursache zurückzuführen ist. Der Effekt von EU-Mitteln aus Strukturfonds in Regionen mit

[10] Siehe Sovey und Green (2011) für eine umfangreiche, praxisorientierte Diskussion der Anforderungen an Instrumentalvariablen.

einer Wirtschaftsleistung nahe dem Schwellenwert kann so geschätzt werden (vgl. Becker et al. 2010). Regressions-Diskontinuitäts-Analysen sind jedoch mit Problemen behaftet, wenn die Untersuchungsobjekte ihre Position um den Schwellenwert manipulieren können, da diese Selbstselektion die zentrale Annahme der lokalen Randomisierung unterminiert.[11]

Sowohl Instrumentalvariablen-Ansätze wie auch Regressions-Diskontinuitäts-Analysen besitzen eine hohe interne Validität, wenn sie korrekt angewendet werden. Beide Ansätze benötigen jedoch natürliche Experimente und sind somit relativ selten anwendbar. In vielen Fällen bleibt die Regressionsanalyse somit die einzige Möglichkeit der Datenanalyse. Aus diesem Grund gehen wir im Folgenden anhand von Beispielen nochmals gezielt auf typische Probleme der Regressionsanalyse ein.

3. Typische Probleme der Datenanalyse anhand von Beispielen

In der Politikwissenschaft hat sich seit der Publikation von King (1989) die Auffassung durchgesetzt, dass ein statistisches Modell zum Messniveau der Daten passen muss, die es zu untersuchen gilt. So passt ein OLS-Regressionsmodell (d.h. ein Regressionsmodell, in dem von linearen Beziehungen ausgegangen wird und in dem nach der Methode der kleinsten Quadrate geschätzt wird) eigentlich nur auf intervallskalierte Daten, während sich für die Analyse von Ordinalskalen (ordinale) Logit- und Probitmodelle eignen. Für die Analyse von Nominalskalen mit mehr als zwei Kategorien empfiehlt sich etwa die multinominale logistische Regression. Für Häufigkeiten wie die Zahl von Streiks oder Konflikten werden am besten Poisson-, Negativ Binomial- oder verwandte Regressionstechniken eingesetzt, und Verläufe (d.h. die Zeit, bis ein bestimmtes Ereignis wie Demokratisierung einsetzt) lassen sich mit Verlaufsmodellen wie beispielsweise einem Weibull-Survival-Modell schätzen.

Ein weiterer zentraler Gesichtspunkt, der die Verfahrenswahl beeinflusst, besteht darin, ob die Daten längsschnitt- oder querschnittsorientiert sind. Eine Kombination dieser beiden

[11] Eine nicht-technische Einführung in die Regressions-Diskontinuitäts-Analyse findet sich bei Shadish et al. (2002), technische Details diskutieren Imbens und Lemieux (2008).

Möglichkeiten besteht in Paneldatensätzen, die in der Politikwissenschaft nach der Definition von Beck (2001) zumeist sog. TSCS-Datensätze (Times series cross section) sind, da die Zahl der Zeitpunkte häufig die Zahl der Einheiten (Länder, Staaten, etc.) übersteigt. Besonders bei der Analyse von Zeitreihen besteht ein Problem darin, dass die Fälle meist nicht unabhängig voneinander sind und dass somit eine zentrale Annahme der meisten Regressionsverfahren verletzt ist. Dies kann sich in Autoregression äußern (die Werte aufeinander folgender Schätzungen korrelieren untereinander; d. h. das BIP in diesem Jahr ist beeinflusst vom BIP des vergangenen Jahres) oder Autokorrelation (die Fehler aufeinander folgenden Schätzungen korrelieren untereinander). In Querschnitten ist die Annahme der Unabhängigkeit dadurch gefährdet, dass die Wirtschaft oder Politik von geographisch benachbarten Regionen oder Staaten miteinander korrelieren. In diesem Zusammenhang ist dann von räumlicher Autokorrelation die Rede, für deren Analyse in den letzten Jahren vor allem Ward und seine Ko-Autoren (z. b. Hoff und Ward 2004, für eine Einführung siehe Ward und Gleditsch 2008) zentrale Ergebnisse geliefert haben. Ein ähnliches Problem wie die Autokorrelation ist die Heteroskedastizität. Ergebnisse, die unter dieser Schwierigkeit leiden, verletzen die Annahme, dass die Varianz der Fehler für alle Werte von X gleich sein soll. Dies führt zu ineffizienten Schätzungen und lässt daher keine aussagekräftigen Hypothesentests zu.

Die pragmatische Haltung gegenüber solchen Problemen besteht darin, die Residuen nach einer ersten provisorischen Analyse graphisch und mit Hilfe von Testverfahren (White, Beusch-Pragan etc.) auf Heteroskedastizität hin zu überprüfen. Liegt das Problem vor, besteht der übliche Ausweg darin, auf geeignetere Verfahren auszuweichen. Zudem kann versucht werden, die Veränderung der Varianz explizit zu modellieren. In der Zeitreihenökonometrie bestehen beispielsweise Techniken, die bei Hochfrequenzdaten wie täglichen Börsenkursen eine Schätzung von Heteroskedastizität erlaubt. Die Clusterung der Fehler ist ja nicht einfach eine unappetitliche Begleiterscheinung der Daten, sondern unter Umständen ein Phänomen, das zu schätzen sich lohnt. In der Politikwissenschaft haben solche Verfahren – es handelt

sich im wesentlichen um GARCH-Modelle und ihre Erweiterungen[12] - über die Analyse der politischen Determinanten von Finanzmarkdaten eine gewisse Verbreitung erfahren (Leblang und Mukerjee 2004, Schneider und Tröger 2006, Bechtel und Schneider 2010). Liegt eine explizite Theorie zu der beobachteten Veränderung der Varianz vor, so lässt sich auch für andere Schätzverfahren die Heteroskedastizität modellieren, zum Beispiel bei Zähldaten (negatives Binomialregressionsmodell mit Modellierung des Dispersionsparameters) oder bei binären Daten (heteroskedastisches Probit Modell).[13]

Wer sich der Natur seiner Daten bewusst ist, hat bei der statistischen Modellierung bereits einen entscheidenden Schritt getan. Wohin aber unterschiedliche Modellierung und damit eine mutmaßlich falsche Methodenwahl führen kann, wollen wir anhand eines klassischen Textes zeigen. Der Artikel von Singer, Bremer und Stuckey (1972, fortan SBS) gilt als einer der ersten Veröffentlichungen, in dem in den Internationalen Beziehungen ein multivariates OLS-Verfahren zur Anwendung gelangte. Was retrospektiv durchaus als bahnbrechender Beitrag zur Theorie der Internationalen Beziehungen zu gelten hat, ist vom methodischen Anspruch heute veraltet. Wie zu zeigen ist, stehen so in heutigem Licht auch die Schlussfolgerungen von SBS auf tönernen Füßen.

In unserem Forschungsdesign halten wir uns, so gut es geht, an die Untersuchungsanlage, wie sie im Originaltext beschrieben ist. Leider lässt sich die Studie aber nicht vollständig replizieren, weil die Daten trotz unserer Rückfragen nicht archiviert und nicht alle Operationalisierungsschritte ausreichend dokumentiert sind. Daher weichen unsere Daten minimal von den Originaldaten ab, vor allem weil wir anstelle einer Fünfjahresperiode

[12] Die Abkürzung steht für Generalized Autoregressive Conditional Heteroskedasticity. Bei der Entwicklung dieser Modelle hat R. Engle, der 2003 zusammen mit C. Granger den Nobelpreis für Wirtschaftswissenschaft erhielt, pionierhafte Vorarbeiten geleistet.

[13] Viele dieser Modelle sind in moderner Statistiksoftware bereits implementiert. Theoretisch relevante Erweiterungen lassen sich jedoch auch mit etwas Kenntnissen und geringem Aufwand durch den Benutzer programmieren.

jährliche Daten verwenden.[14] Kernaussage von SBS ist in Anlehnung an die Diskussionen zwischen Waltz auf der einen und Deutsch und Singer auf der anderen Seite der 1960er Jahre, dass die Konzentration der Machtressourcen im internationalen System einen Einfluss auf die Kriegshäufigkeit hat. Es werden zwei konkurrierende Modelle getestet: zum einen das Vorherschafts- bzw. Stabilitätsmodell, welches besagt, dass die Kriegshäufigkeit ansteigt, sobald das Staatensystem sich von einer hohen Konzentration von Machtressourcen wegbewegt. Im Gegensatz dazu bedeutet das Paritäts- bzw. Fluiditätsmodell, dass die Kriegswahrscheinlichkeit abnimmt, wenn sich das Staatensystem von hoher Konzentration zu einem ausgeglicheneren Zustand der Verteilung von Machtressourcen bewegt. SBS verwenden drei unabhängige Variablen, um ihr abhängiges Konstrukt, die Anzahl der Kriegsmonate pro Jahr, zu schätzen. Dies sind i) die Konzentration von Machtressourcen (Konzentration), wobei Machtressourcen durch einen additiven Index gemessen wird, der aus einer demographischen, einer industriellen und einer militärischen Dimension besteht; ii) die Veränderung der Konzentration und iii) die Verschiebung der Konzentration. Während die Veränderung der Konzentration unverändert bleiben kann, da ein Machtverlust eines Staates sich in einem identischen Machtgewinn eines anderen Staates äußern kann, misst die Verschiebungsvariable die Summe aller Veränderungen.

Wir schätzen fünf Modelle. Während sich das erste Modell, eine OLS-Schätzung, auf die Vorgehensweise von SBS stützt, korrigieren die zwei nachfolgenden Schätzmethoden – Prais-Winsten und Cochrane Orcutt – die Zeitreihe um das Problem der Autokorrelation. Das vierte Modell schließlich berücksichtigt, dass die abhängige Variable nicht auf einer Intervallskala gemessen wird, sondern eine Häufigkeit auszählt. Wir verwenden als Häufigkeitsmodell ein Poission-Modell. Zusätzlich rechnen wir als letztes Modell eine Poisson-Regression, die für Autokorrelation korrigiert.

[14] Eine solche Nachlässigkeit würde heute bei den besten Journalen nicht mehr durchgehen und die Publikation eines Artikels verhindern. Die meisten führenden Zeitschriften haben mittlerweile eine strenge Replikationspolitik eingeführt und verlangen die Publikation der Daten, die für eine Untersuchung verwendet wurden. Siehe Bueno de Mesquita et al. (2003).

Tabelle 1: Modelle zur Erklärung der Anzahl der Kriegsmonate in einem Untersuchungsjahr von 1816 bis 2001.

	(1) OLS	(2) Prais-Winsten	(3) Cochrane-Orcutt	(4) Poisson Regression	(5) Poisson-Zeitreihen-modell
Konzentration von Machtressourcen	6.694 (9.030)	15.904 (17.628)	18.918 (18.042)	2.079 (1.016)**	3.490 (2.207)
Veränderung der Konzentration	56.698 (23.322)**	23.217 (15.931)	21.529 (16.091)	5.880 (1.935)***	0.857 (4.203)
Verschiebung der Konzentration	117.152 (27.989)***	-9.632 (23.987)	-10.686 (24.058)	22.093 (2.512)***	9.878 (5.543)*
Autoregressiver Term R1					0.585 (0.055)***
Konstantglied	-1.416 (3.767)	-3.485 (7.604)	-4.634 (7.744)	-0.153 (0.430)	-0.559 (0.931)
Beobachtungen	185	185	184	185	184
Angepasstes R-Quadrat	0.15	0.01	0.02	0.10[1]	0.46[1]
Durbin Watson	0.79	1.79	1.79		1.628

Anmerkungen: Standardfehler in Klammern; * signifikant auf dem 10%-Fehlerniveau; ** 5%-Fehlerniveau; *** 1%- Fehlerniveau. [1] Pseudo R^2

Die OLS-Regression zeigt, dass die Vorzeicheninterpretation der Koeffizienten eher für das Paritäts- bzw. Fluiditätsmodell sprechen. So folgt einem hohen Wert von Konzentration der Machtressourcen eine hohe Anzahl von Kriegsmonaten. Allerdings zeigt der geringe Wert der Durbin Watson-Statistik, dass in den Daten Autokorrelation eine große Rolle spielt und die Annahme identisch und unabhängig verteilter Fehler verletzt ist. Als Faustregel gilt, dass der Durbin Watson-Koeffizient nahe bei 2.0 liegen muss, damit wir von einem Fehlen von Autokorrelation ausgehen können. Ohne den grundlegenden Ansatz eines linearen Regressionsmodells zu ändern, könnte dies durch die beiden klassischen Verfahren Prais-Winsten oder Cochrane-Orcutt verbessert werden. Schätzt man den Zusammenhang mit diesen Methoden und berücksichtig somit eine mögliche Autokorrelation der Fehler, so übt keiner der Koeffizienten einen signifikanten Einfluss aus. Es ist somit fraglich, ob das Modell richtig spezifiziert wurde. Dabei ist anzunehmen, dass wichtige erklärende Variablen fehlen.

Aus der Sicht der heutigen Methodenforschung wäre es ferner angebracht, ein geeigneteres Modell als eine lineare Regression zu verwenden. Wie erwähnt berücksichtigen sog. Zähldatenmodelle (event count), dass die abhängige Variable nicht auf einer Intervallskala gemessen wird, sondern die Häufigkeit bestimmter Ereignisse wie Kriegsausbrüche oder Opfer bestimmter Konflikte auszählt. Darum haben wir zusätzlich noch zwei Poisson-Modelle geschätzt. Das erste Poisson-Modell weist stark signifikante Koeffizienten aus. Was zunächst wie eine Bestätigung der Hypothesen erscheinen mag, sollte jedoch zu großer Skepsis führen, ob ein einfaches Poisson-Modell angemessen ist. Die Poissonverteilung nimmt an, dass die Varianz identisch mit dem Mittelwert ist und schätzt daher keinen Varianzparameter. Ist die wahre Varianz jedoch größer, so ergibt das Poissonmodell zu kleine Schätzungen der Standardfehler und als Konsequenz davon eine „inflationierte“ Signifikanz. Eine solche Überdispersion kann unter anderem dadurch auftreten, dass einzelne Ereignisse, wie Kriegsmonate nicht unabhängig sind.[15] Das zweite Zähldaten-Modell korrigiert aus diesem Grund noch für Autokorrelation, was, wie in den vorherigen Modellen deutlich wurde, eine problematische Rolle in den zugrunde liegenden Daten spielt. Die Vorzeicheninterpretation spricht in diesen beiden Modellen wieder für das Paritäts- bzw. Fluiditätsmodell. So geht mit einem hohen Wert von Konzentration der Machtressourcen eine größere Wahrscheinlichkeit für eine hohe Anzahl von Kriegsmonaten einher. Trotzdem muss gesagt werden, dass bei allen Modellen, die Konfidenzintervalle so groß sind, dass eine vertrauenswürdige Interpretation der Ergebnisse nicht möglich ist. Bei einigen Koeffizienten kann aufgrund dieser “Großzügigkeit“ nicht ausgeschlossen werden, dass die Koeffizienten auch das umgekehrte Vorzeichen haben könnten.

Somit ist fraglich, ob das Modell richtig spezifiziert wurde. Dabei ist anzunehmen, dass wichtige erklärende Variablen fehlen und wir es daher mit dem oben diskutierten *omitted*

[15] Ein weiterer Grund könnte die Annahme des Poisson-Modells sein, dass Mittelwert und Varianz identisch sind. Schätz man jedoch ein Negativ-Binomial Modell, das diese Annahme nicht trifft und damit auch für Daten mit Überdispersion geeignet ist, erhält man substantiell die gleichen Ergebnisse.

variable bias zu tun haben. Die Replikation einer klassischen Studie der quantitativen Konfliktforschung ergibt so aus theoretischer Warte, dass bestimmte klassische realistische Hypothesen sich kaum für eine systematische Überprüfung eignen. Vor diesem Hintergrund überrascht es nicht, dass Beobachter wie Vasquez (1997) dieses klassische Forschungsprogramm als „degeneriert“ im Sinne des Wissenschaftsphilosophen Lakatos begreifen.

Neben dem *omitted variable bias* können allerdings noch andere Verzerrungsformen der Forscherin das Leben schwer machen: Ein Schätzer ist dann nicht erwartungstreu, wenn die Fälle, die untersucht werden, nicht zufällig ausgewählt sind, sondern unter einem systematischen Selektionsbias leiden. In der politikwissenschaftlichen Methodenlehre ist dieses Problem früh durch Achen (1987) diskutiert worden. Unter einem Selektionsbias haben in den Internationalen Beziehungen sowohl qualitative wie quantitative Analysen zur Effektivität der militärischen Abschreckung gelitten, wie die Aufsätze von Achen und Snidal (1989) und Fearon (1994) verdeutlichen. Bei den von diesen Forschern kritisierten Untersuchungen bestand die Verzerrung insofern, als sie sich nur auf Fälle gescheiterter Abschreckung oder auf militärische Krisen bezogen. Da die Episoden erfolgreicher Abschreckung, in denen ein potentieller Herausforderer den Status quo akzeptiert, ausgeschlossen sind, entsteht eine systematische Verzerrung. Die Berücksichtigung der gesamten Varianz ist oft nicht ausreichend, um das potentielle Problem des Selektionsbias in den Griff zu kriegen. So lässt sich etwa der Erfolg von Sanktionen nicht verlässlich analysieren, wenn als Grundlage der Untersuchung nur Fälle dienen, in denen ein Staat oder eine Staatengruppe gegenüber einem Land oder einer Regierung dieses außenpolitische Instrument gebraucht. Die Sanktionsfälle sind keine Zufallsstichprobe sämtlicher möglicher Situationen, in denen die Verhängung einer Sanktion möglich schien. Die Verzerrung rührt unter anderem daher, dass die mächtigen Schurken unter Umständen einer Sanktion entgehen, weil sie selber über ein glaubwürdiges Drohpotential verfügen, während schwache Sünder

nicht die Kraft haben, eine Sanktion abzuwenden. Wenn sich nun die Analyse nur auf die tatsächlichen und nicht auf die potentiellen Sanktionen bezieht, wird die Wirkung des Instrumentes Sanktion systematisch überschätzt.

Zur Analyse solcher Selektionsprozesse hat der Nobelpreisträger James Heckmann spezielle Selektionsmodelle entworfen, die vor allem bei intervallskalierten und bei binären abhängigen Variablen gebräuchlich sind. Nooruddin (2002) bietet etwa eine Studie zum Erfolg von Sanktionen, bei denen der Erfolg als Dummyvariable operationalisiert ist. Das verwendete Heckmann-Probit-Modell ist eine Spezialform von sog. Probitregressionsmodellen, bei denen die abhängige Variable – wie erwähnt - kategorial definiert ist. Bei einer einfachen Probitregression zeigt sich etwa, dass der Sanktionserfolg zunimmt, wenn die Kosten der Sanktion wachsen. Doch dieses Modell ist, wie erwähnt, der Datenstruktur nicht angemessen. Nooruddin (2002) zeigt im Vergleich des einfachen Probit- zum Heckmann-Probit-Modell, dass sich die Wirkung der Kostenvariablen halbiert, wenn der Selektionseffekt berücksichtigt ist.

Ein weiteres Problem verzerrter Schätzung kann durch den Aggregationsbias entstehen. Diese Form von Verzerrung ist typisch für Makroanalysen, in denen etwa aufgrund von Investitionsdaten einzelner Unternehmen auf das Verhalten ganzer Länder geschlossen wird. Solche Interpretationen sind nur sinnvoll, wenn die Investitionen im Vergleich der erfassten Firmen unimodal verteilt sind und nur geringfügig streuen oder bekannt ist, wie das Sample der Unternehmen gezogen wurde, so dass Rückschlüsse auf die Grundgesamtheit möglich sind. Oft werden diese Annahmen jedoch verletzt, so dass Studien, die solche hochaggregierten Indikatoren verwenden, den tatsächlichen Einfluss in einem Hypothesentest verfälschen. Ein Ausweg aus diesem Problem besteht in der Mikrofundierung der Forschung, welche die jeweils relevanten Akteure – im Beispiel die einzelnen Unternehmen – statt einen abstrakten Aggregationsakteur wie den Nationalstaat in den Vordergrund der Analyse rückt. Auf theoretischer Ebene entspricht die mikrofundierte Forschung dem Ideal des

methodologischen Individualismus, wonach Makrophänomene an individuelles Handeln zurückzubinden sind. Empirisch äußert sich der Trend hin zur Mikrofundierung in der Verwendung von zeitlich wie räumlich disaggregierten Daten oder von Umfragen (ausführlich dazu siehe Schneider 2013).

Besonders virulent ist die mikrobasierte Forschung in der Analyse von Bürgerkriegen, wo aufgrund neuer Datenquellen etwa gezeigt werden konnte, wie Ungleichheit das Bürgerkriegsrisiko fördert (Østby 2008, Cederman et al. 2013). So begrüßenswert dieser Wandel auch ist: einige Anwendungen ignorieren die Grundannahme jeglicher empirischer Überprüfung, sei sie nun qualitativer oder quantitativer Orientierung, dass die Beobachtungen eigentlich unabhängig voneinander sein sollten. Diese Voraussetzungen sind etwa in Studien verletzt, in denen für praktisch identische Regionen aufgrund derselben Kovariate das Risiko der politischen Gewalt errechnet wird. Die so praktizierte Vervielfachung der Fälle führt oft dazu, dass im Sinne einer Fata Morgana die Forschenden in ihren Daten Strukturen entdecken, die gar nicht vorhanden sind. Ein mittlerweile vielpraktizierter Ausweg besteht über sog. Mehrebenenmodelle, in denen der Zusammenhang zwischen Merkmalen einer höheren Aggregationsebene und dem Verhalten einer tieferen Aggregationsebene untersucht wird (Steenbergen und Jones 2002). Dieser Ansatz hat etwa in der länder- oder regionenvergleichenden Umfrageforschung Einzug gehalten, in der etwa aufgrund von Strukturmerkmalen einer politischen Einheit und dem soziodemographischen Profil der Befragten auf deren Antwortverhalten geschlossen wird. Wenn sich nun aber die Regionen oder die Individuen nicht genügend unterscheiden und somit die Disaggregation zu feinteilig erfolgt, liegt immer noch eine Verletzung der Annahme voneinander unabhängiger Beobachtungen vor.

Ein weiteres Problem der mikrofundierten Forschung besteht darin, dass für tieferliegende Einheiten oft die Messung problematisch ist. Dies gilt in bestimmten Regionen etwa für die Satellitendaten, die die Beleuchtung in der Nacht wiedergeben und die als Indikator für

wirtschaftliche Entwicklung verwendet werden (Nordhaus 2006). Schließlich öffnet sich den Verfechtern von disaggregierten Analysen ein Füllhorn von Daten, deren Einsatz bisweilen theoretisch unvermittelt erfolgt. Dies gilt etwa für die erste Phase der disaggregierten Bürgerkriegsforschung, in der geographische Maße wie die Entfernung der Gewalttätigkeiten zu den Zentralen der Kontrahenten als „unabhängige Variablen" in die Analysen eingingen. Dass die Lokalisierung der Gewalt auf bewussten Entscheidungen der militärischen Planer beruht und so die Endogenität von Entfernungsmaßen zumindest zu diskutieren wäre, haben spätere Studien über die Verwendung von Instrumentalregressionsansätzen wieder zurecht gerückt.

4. Schlußbetrachtung: Big Data und das Primat der Theorie - neue Herausforderungen

Die mikrofundierte Forschung ist aufs Engste mit der stärkeren Nutzung neuartiger und überaus umfangreicher Datensatzätze verknüpft, die oft unter dem Stichwort von „Big data" diskutiert werden. In einem provokativen Artikel meinte Anderson (2009), dass diese revolutionäre Informationsfülle und die Fortschritte, die besonders Informatiker in der Generierung von komplexen Algorithmen erzielt haben, um die Struktur der Datenberge zu enthüllen, die Theoriebildung überflüssig machen würden. Natürlich haben die sozialen Medien die Gesellschaft verändert. Darüber hinaus lassen sich aus dem Internet oder aus Facebook, Twitter und anderen virtuellen Netzwerken Daten generieren, die uns praktisch Prognosen in Echtzeit vermitteln. Ein Beispiel für die neue Datenwelt der Internationalen Beziehungen bietet der GDELT-Ereignisdatensatz (Leetaru und Schrodt 2013), der bereits im Frühjahr 2013 über 200 Millionen Ereignisse umfasste.[16] Doch obgleich diese Datenfülle unerhörte Möglichkeiten zu umfassenden Analysen wie auch zu Untersuchungen von

[16] Ereignisdaten haben in den Internationalen Beziehungen eine lange Tradition. Erste Datensätze wurden in den 1960er Jahren etwa zur Untersuchung einzelner Konflikte wie des 1. Weltkriegs entwickelt. Später folgten umfassendere Datensätze wie WEIS und COPDAP, die beispielsweise zur Analyse der Supermachtbeziehungen (Goldstein und Freeman 1990) und des Endes des Kalten Krieges (Schneider, Widmer und Ruloff 1993) verwendet wurden. Über die Renaissance der Ereignisdatensammlungen in den 2000er Jahren informiert ein Sonderheft der Zeitschrift *International Interactions* (Bernauer und Gleditsch 2012).

einzelnen Akteurspaaren eröffnet hat, geht dem Erklären der Internationalen Beziehungen immer noch das Entwickeln von kausalen Mechanismen voraus. Eine hochentwickelte, Algorithmen-basierte Suche nach Korrelationen in den neuen Datenquellen ersetzt nicht die Rolle von theoretisch postulierten Kausalannahmen und deren Überprüfung durch Forschungsdesigns mit hoher, interner Validität. Die sozialen Medien mögen zwar Entscheidungsprozesse beschleunigt haben, doch ob sie eine neue Qualität dieser Entscheidungen geschaffen haben, mit denen die etablierten Theorien nicht zu Recht kommen, muss sich noch erweisen.

Auch wenn große Datensätze wie GDELT die Theoriebildung natürlich entgegen Anderson (2009) und anderer induktiv orientierter Big Data-Enthusiasten nicht ersetzen können, lassen sich für die die Prognose internationaler Prozesse fruchtbar einsetzen. Frühe Beispiele dafür sind die Studien von Beck, King und Zeng (2000, siehe auch 2004) zur Prognose von zwischenstaatlichem Konflikt. Sie zeigen, dass die Standardmodelle der makro-quantitativen Kriegsursachenforschung – logistische und Probit-Regressionen - oft nur das für die Praxis weniger relevante Phänomen vorhersagen, die Jahre nämlich, in dem Frieden in einer Dyade herrschte (vgl. Russett und Oneal 2001). Der Informatik entlehnte Ansätze wie Neuronale Netzwerke hingegen vermögen es auch, das eigentlich interessierende Phänomen - den Ausbruch von Krieg - zu prognostizieren. Der Vorteil von neuronalen Netzwerken und verwandter komplexer Algorithmen ist es, dass sie flexibler sind als Standardregressionsverfahren, die für die Beziehung zwischen den unabhängigen Variablen und der abhängigen Variablen eine feste Funktion vorgeben – im Falle von Logit ist das, wie erwähnt, eine Funktion, die s-förmig verläuft. Neuronale Netzwerke sind nun als Prognoseverfahren konventionellen Logit- oder Probitansätzen insofern überlegen, als sie die simultane Schätzung unterschiedlicher Beziehungen gleichzeitig zulassen und eine sehr flexible, funktionale Form ermöglichen. Interessanterweise schneiden Standardansätze bei der Prognose von Bürgerkriegen nicht schlechter ab als neuronale Netzwerke, auch wenn sie eine

höhere Anzahl von fehlerhaft prognostizierten Konflikten produzieren als neuronale Netzwerke (Rost, Schneider und Kleibl 2009). Um die Prognosekraft von Modellen generell zu evaluieren, haben Ward et al. (2010) und andere in jüngster Zeit das Methodenrepertoire beträchtlich erweitert.

Natürlich ist noch einmal hervorzuheben, dass Prognose allein nicht das Ziel der empirischen Sozialwissenschaft sein kann. Aus diesem Grund sind auch die Anstrengungen wichtig, die es erlauben, rigorose theoretische Modelle direkt empirisch zu schätzen und damit die Relevanz von innovativen Erklärungen zu prüfen. Die übliche Vorgehensweise außerhalb der experimentellen Forschung ist es, dass die aus der formalen Theorie abgeleiteten Modelle mit Standardverfahren geschätzt werden, wie dies etwa Fearon (1994) für sein einflussreiches Krisenverhandlungsmodell getan hat. Dies ist aus verschiedenen Gründen fragwürdig: Erstens ist ein statistisches Modell wiederum an zusätzliche Annahmen geknüpft, die mit den theoretischen Aussagen darüber, wie die Daten entstanden sind, durchaus in Widerspruch stehen können (Morton 1999). Zweitens lassen sich aus den spieltheoretischen Modellen oft deterministische Prognosen ableiten. Im Falle des Gefangenendilemma läuft das auf die Vorhersage hinaus, dass sich die Akteure mit Wahrscheinlichkeit 1 nicht-kooperativ verhalten werden. Zu einem Test dieser Hypothese passen aber nicht die probabilistischen Modelle, die üblicherweise zur Evaluation von spieltheoretischen Prognosen Verwendung finden. Thomson et al. (2006) oder Schneider et al. (2010) gebrauchen deshalb einfache statistische Verfahren wie die Anzahl korrekter Punktprognosen oder des durchschnittlichen quadrierten Prognosefehlers. Drittens ist bei der Überprüfung eines strategischen Modells damit zu rechnen, dass die Handlungen der Akteure sich gegenseitig bedingen und dass die Untersuchung diese Interdependenz berücksichtigen muss. Dies ist beim Test der Modellprognosen von Krisenverhandlungsspielen problematisch. Ein einfaches Logit- oder Probit-Modell beschränkt die Analyse auf den letzten Ast des Spielbaums, die Entscheidung für oder gegen die Kriegsoption. Nicht berücksichtigt werden dabei die Züge, die

vorangegangen sind. Signorino (1999, 2003, siehe auch Lewis und Schultz 2003) zeigen, dass die Standardverfahren dabei zu substantiell inkorrekten Ergebnissen führen können. Signorino benutzt ein statisches Verfahren, dass die Spielstruktur einer militärischen Krise und damit die strategische Interaktion berücksichtigt. Das von Signorino vorgeschlagene Verfahren kann aber auch problemlos auf andere Situationen strategischen Handelns angewandt werden, wie beispielsweise zur Schätzung der Eskalationsdynamiken bei Streitverfahren in der Welthandelsorganisation (Sattler, Spilker und Bernauer 2014). Zu beachten ist allerdings, dass natürlich nicht für die Analyse jedes Phänomens ein neues Schätzverfahren gesucht werden sollte, da dies die Transparenz des Forschungsprozesses deutlich verringern und das Gebot unterminieren würde, Theorie und Schätzansatz auseinander zu halten. Kritik hat der besonders in den 2000er Jahren virulente Trend, avancierte theoretische Modelle möglichst direkt zu prüfen, durch den anti-theoretischen Reflex einiger seiner Verfechter gefunden. So weisen Clarke und Primo (2012) in einer viel beachteten Streitschrift zu Recht daraufhin, dass besonders Resultate der normativen Theorie wie Arrows Unmöglichkeitstheorem keiner Überprüfung bedürfen, da sie einzig den Kriterien der mathematischen Logik genügen müssen.

Zugleich ist es wohl auch nicht sinnvoll, barocke Modelle, die nicht dem Grundsatz des „non-fat modeling“ genügen wollen, 1 zu 1 testen zu wollen. Achen (2002) empfiehlt in seiner Diskussion des „omitted variable bias“ und anderer seiner Meinung nach lässlicher Sünden als Regel sogar, die Zahl der erklärenden Variablen auf drei zu reduzieren – dies aber immer auf der Basis eines klaren, möglichst mathematisierten Hypothesenfundaments. Die klare Verbindung zwischen theoretischen und statistischen Modellen wird es in Zukunft auch erlauben, vermehrt vergleichende Tests von theoretischen Modellen vorzunehmen, wie dies Bennett und Stam (2003) für die Kriegsursachenforschung sowie Thomson et. al. (2006) und Schneider et al. (2010) für die Analyse von Entscheidungsprozessen in der Europäischen Union getan haben.

In der qualitativen Politikforschung ist interessanterweise eine ähnliche Bewegung hin zur direkten Überprüfung der Theorie zu beobachten. Ein Problem bei diesen qualitativen komparativen Analysen (QCA) besteht hier allerdings darin, dass die Theorien oft komplex sind. Eine Möglichkeit zur Explizierung solcher Zusammenhänge besteht darin, mit Hilfe der Boolschen Logik die Variablen logisch miteinander zu verknüpfen, was über mengentheoretische Konzepte geschieht (Schneider und Wagemann 2012 für eine Einführung). Zugleich sind die Theorien aber oft auch deterministisch spezifiziert, so dass, um die Hypothesen zu widerlegen, deshalb bereits eine einzelne Fallstudie mit divergierenden Ergebnissen genügt (Lieberson 1991). Zudem reagieren die Modellprognosen sehr stark auf kleine Messfehler, wie Hug (2013) in Monte Carlo-Simulationen zeigt.

Eine subjektivistische Art und Weise der Theorieprüfung bieten bayesianische Modelle, die zunehmend den Weg in die Politikwissenschaft finden (Western 1996, Gill 2004). Dieser Ansatz verbindet die unbeobachtbaren Daten mit einer a priori-Wahrscheinlichkeit, die einer substantiellen Theorie oder vorherigen Analysen entnommen sein können oder schlicht auch auf der Erfahrung der Forscherin – sprich: ihrem Vorurteil – beruhen können. Dieses "Wissen" wird dann mit Hilfe der Regel von Bayes und aufgrund der beobachteten Daten in eine posteriore Einschätzung darüber verwandelt, wie der Datengenerierungsprozess tatsächlich verlaufen ist. Der Reiz dieses Ansatzes besteht in den Sensitivitätsanalysen. Sie erlauben es, die subjektive Einschätzung der Wirklichkeit gezielt mit der Realität zu vergleichen. Damit ist das "Fata Morgana"-Problem natürlich nicht gelöst, aber der Weg hin zur Etablierung von Zusammenhängen wird transparenter.

Abschließend wollen wir festhalten, dass die Internationalen Beziehungen in den letzten Jahren nicht nur theoretisch sondern auch methodisch enorme Fortschritte erzielt haben. Dies äußert sich neben den bereits diskutierten Entwicklungen unter anderem darin, dass die Zahl an Artikeln mit „Spülbecken“-Regressionen zumindest in den Spitzenzeitschriften zunehmend abnimmt. Auch die einseitige Fokussierung auf die bloße statistische „Signifikanz“ von

Schätzergebnissen ist an guten Konferenzen kaum mehr anzutreffen. Auch die Publikation unübersichtlicher Tabellen, die aufgrund der vielen Signifikanz-Sternchen an Miniaturmilchstraßen erinnern, ist eher noch in zweit- und drittrangigen Zeitschriften als in den wichtigsten Publikationsorganen anzutreffen. Viel mehr präsentieren eine wachsende Zahl von Autoren ihre Ergebnisse statt in, auf graphische Weise und zeigen mit Hilfe der komparativen Statik, wie Änderungen in den zentralen abhängigen Variablen praktisch relevante Reaktionen in den Erwartungswerten der abhängigen Variablen hervorrufen.[17] Wie der Rest in der Disziplin ist aber auch die quantitativ orientierte Forschung in den Internationalen Beziehungen bestimmten Methodentrends unterworfen. Die damit verbundenen Konjunkturen von Modewörtern wie „Endogenität" lassen dann bisweilen vergessen, dass für die Tests innovativer theoretischer Argumente oft einfache Methoden genügen. Dass diese Überprüfungen zugleich zentralen Validitätsansprüchen genügen müssen, versuchte dieses Kapitel zu zeigen.

5. Literaturverzeichnis

Achen, Christopher H. 1987. *The Statistical Analysis of Quasi-Experiments.* Berkeley: University of California Press.

Achen, Christopher H. 2002. Toward a New Political Methodology: Microfoundations and ART. *Annual Review of Political Science* (5): 423-50.

Achen, Christopher H. und Duncan Snidal. 1989. Rational Deterrence Theory and Comparative Case Studies. *World Politics* (41): 144-69.

Anderson, Chris. 2009. *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*. http://www.wired.com/science/discoveries/magazine/16-07/pb_theory, gesehen 30/12/2013

[17] Einführungen für die Produktion von Ergebnisgrafiken bieten Kastellec und Leoni (2007). Die Algorithmen von King, Tomz, und Jason Wittenberg (2000) helfen bei der komparativen Statik.

Bechtel, Michael M. und Gerald Schneider. 2010. Eliciting Substance from 'Hot Air': Financial Market Responses to EU Summit Decisions on European Defense. *International Organization* 64 (2): 199-223.

Beck, Nathaniel. 2001: Time-Series Cross-Section Data: What Have We Learned in the Past Few Years*? Annual Review of Political Science* 4: 271-93.

Beck, Nathaniel und Jonathan Katz. 2001. Throwing Out the Baby with the Bath Water: A Comment on Green, Kim, and Yoon. *International Organization* 55 (2): 487-98.

Beck, Nathaniel, Gary King und Langche Zeng. 2000. Improving Quantitative Studies of International Conflict: A Conjecture. *American Political Science Review* 94 (1): 21-36

Beck, Nathaniel, Gary King und Langche Zeng. 2004. Theory and Evidence in International Conflict: A Response to de Marchi, Gelpi and Grynaviski. *American Political Science Review* 98 (2): 379-389.

Becker, Sascha O., Peter H. Egger, und Maximilian von Ehrlich. 2010. Going NUTS: The Effect of EU Structural Funds on Regional Performance. *Journal of Public Economics* 94 (9-10): 578-590.

Bennett, D. Scott und Allan C. Stam. 2003. *The Behavioral Origins of War*. Ann Arbor: University of Michigan Press.

Bernauer, Thomas und Nils Petter Gleditsch. Hrsg. 2012. Event Data in the Study of Conflict. *International Interactions* 38 (4): 375-569.

Blendin, Hanja und Gerald Schneider. 2012*:* Nicht jede Form von Stress mindert die Entscheidungsqualität: ein Laborexperiment zur Groupthink-Theorie. *Jahrbuch für Handlungs- und Entscheidungstheorie* 7: 61-80.

Bueno de Mesquita, Bruce, Nils Petter Gleditsch, Patrick James, Gary King, Claire Metelitis, James Lee Ray, Bruce Russett, Håvard Strand, und Brandon Valerino. 2003.

Symposium on Replication in International Studies Research. *International Studies Perspectives* 4 (1): 72-107.

Cederman, Lars-Erik, Kristian Skrede Gleditsch, und Halvard Buhaug. 2013. *Inequality, Grievances, and Civil War*. Cambridge: Cambridge University Press.

Clarke, Kevin. 2005. The Phantom Menace: Omitted Variable Bias in Econometric Research. *Conflict Management and Peace Science* 22 (4): 341-52.

Clarke, Kevin A. und David M. Primo. 2012. *A Model Discipline: Political Science and the Logic of Representations*. Oxford: Oxford University Press.

Diekmann, Andreas. 2007. *Empirische Sozialforschung. Grundlagen, Methoden, Anwendungen*. 17te Auflage. Reinbek: Rowohlt.

Edgington, Eugene S. und Patrick Onghena. 2007. *Randomization tests*. Boca Raton: Chapman & Hall/CRC.

Fearon, James. 1994. Signalling versus the Balance of Power and Interests: an Empirical Test of a Crisis Bargaining Model. *Journal of Conflict Resolution* 38 (2): 236-69.

Fearon, James, Macartan Humphreys und Jeremy M. Weinstein. 2009. Can Development Aid Contribute to Social Cohesion after Civil War? Evidence from a Field Experiment in Post-Conflict Liberia. *American Economic Review* 99 (2): 287–291.

Gartner, Scott S. 2008. The Multiple Effects of Casualties on Public Support for War: An Experimental Approach. *American Political Science Review* 102 (1): 95-106.

Gill, Jeff. Hrsg. 2004. Special Issue on Bayesian Methods. *Political Analysis* 12 (4): 323-443.

Goldstein, Joshua S. und John R. Freeman. 1990. *Three-Way Street: Strategic Reciprocity in World Politics*, Chicago, Ill.: University of Chicago Press.

Greene, William. 2008. *Econometric Analysis*. Upper Saddle River: Prentice Hall.

Heckman, James. 1979. Sample selection bias as a specification error. *Econometrica* 47 (1): 153-61.

Höfer Thomas, Hildegard Przyrembel und Silvia Verleger. 2004. New evidence for the Theory of the Stork. *Paediatric & Perinatal Epidemiology* 18 (1): 88–92.

Hoff, Peter D. und Michael D. Ward. 2004. Modeling Dependencies in International Relations Networks. *Political Analysis* 12 (2): 160-175.

Holland, Paul W. 1986. Statistics and Causal Inference. *Journal of the American Statistical Association* 81 (396): 945-960.

Hug, Simon. 2013. Qualitative Comparative Analysis: How inductive use and measurement error lead to problematic inference. *Political Analysis* 21 (2): 252-265

Imbens, Guido W. und Thomas Lemieux. 2008. Regression discontinuity designs: A guide to practice. *Journal of Econometrics* 142 (2): 615-635.

Jensen, Nathan, Bumba Mukherjee, und William Bernhard. 2014. Introduction: Survey and Survey Experimental Work in IPE. *International Interaction* 40 (3). Im Erscheinen.

Jensen, Nathan und Mi Jeon Shing. 2014. Globalization and Domestic Trade Policy Preferences: Reciprocity and Mass Support for Agriculture Subsidies. *International Interaction* 40 (3). Im Erscheinen.

Kastellec, Jonathan P. und Eduardo L. Leoni. 2007. Using Graphs Instead of Tables in Political Science. *Perspectives on Politics* 5 (4): 755-771.

Kern, Holger und Jens Hainmüller. 2009. Opium for the Masses: How Foreign Media Can Stabilize Authoritarian Regimes. *Political Analysis* 17 (4): 377-99.

King, Gary. 1989. *Unifying political methodology: the likelihood theory of statistical inference.* 1. Auflage. Cambridge: Cambridge University Press.

King, Gary. 2011. Ensuring the Data Rich Future of the Social Sciences. *Science* 331 (11): 719-721.

King, Gary und Lanche Zeng. 2001. Explaining Rare Events in International Relations. *International Organization* 55 (3): 693-715.

King, Gary, Robert O. Keohane und Sidney Verba. 1994. *Designing social inquiry: scientific inference in qualitative research*. Princeton: Princeton University Press.

King, Gary, Michael Tomz und Jason Wittenberg. 2000. Making the Most of Statistical Analyses: Improving Interpretation and Presentation. *American Journal of Political Science* 44 (2): 341–355.

Leblang, David und Bumba Mukherjee. 2004. Presidential Elections and the Stock Market: Comparing Markov-Switching and Fractionally Integrated GARCH Models of Volatility. *Political Analysis* 12 (3): 296-322.

Leetaru, Kalev und Philip A. Schrodt. 2013. GDELT: Global Data on Events, Location and Tone. Paper präsentiert an der ISA-Konferenz, San Francisco April 2013.

Lewis, Jeffrey B. und Kenneth A. Schultz. 2003. Revealing Preferences: Empirical Estimation of a Crisis Bargaining Game with Incomplete Information. *Political Analysis* 11 (4): 345-67.

Lieberson, Stanley. 1991. Small Ns and Big Conclusions: An Examination of the Reasoning in Comparative Studies Based on a Small Number of Cases. *Social Forces* (70): 307-20.

McDermott, Rose, Dominic D. P. Johnson, Jonathan Cowden, Jonathan und Stephen Rosen. 2007. Testosterone and Aggression in a Simulated Crisis Game. *Annals of the American Academy of Political and Social Sciences* 614: 15-33.

McDermott, Rose, Dawes, Chris, Prom-Wormley, Elizabeth, Eaves, Lindon und Hatemi, Peter K. 2013. MAOA and Aggression: A Gene–Environment Interaction in Two Populations. *Journal of Conflict Resolution* 57 (6): 1043-1064.

Morgan, Stephen L und Winship, Christopher. 2007. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge: Cambridge University Press.

Morton, Rebecca. 1999. *Methods and models: a guide to the empirical analysis of formal models in political science.* Cambridge: Cambridge University Press.

Morton, Rebecca und Williams, Kenneth. 2010. *Experimental Political Science and the Study of Causality*. Cambridge: Cambridge University Press.

Mutz, Diana. 2011. *Population-Based Survey Experiments*. Princeton: Princeton University Press.

Nooruddin, Irfan. 2002. Modeling Selection Bias in Studies of Sanctions Efficacy. *International Interactions* 28 (1): 59-75.

Nordhaus William D. 2006. Geography and Macroeconomics: New Data and New Findings. *Proceedings of the National Academy of Science* 103 (10): 3510–3517.

Ostrom, Elinor. 1990. *Governing the Commons. The Evolution of Institutions for Common Action*. Cambridge: Cambridge University Press.

Østby, Gudrun. 2008. Polarization, Horizontal Inequalities and Violent Civil Conflict. *Journal of Peace Research* 45 (2): 143–162.

Rost, Nicolas, Gerald Schneider und Johannes Kleibl. 2009. A global risk assessment model for civil wars. *Social Science Research* 38 (4): 921-933.

Russett, Bruce M. und John R. Oneal. 2001. *Triangulating Peace: Democracy, Interdependence, and International Organizations.* New York: W. W. Norton.

Sattler, Thomas, Gabriele Spilker und Thomas Bernauer. 2014. Does WTO Dispute Settlement Enforce or Inform? *British Journal of Political Science*. Im Erscheinen.

Schneider, Carsten Q. und Claudius Wagemann. 2012. *Set-Theoretic Methods: A User's Guide for Qualitative Comparative Analysis and Fuzzy Sets in Social Science*. Cambridge: Cambridge University Press.

Schneider, Gerald. 2013. Von Makro zu Mikro: Grundlagen und Perspektiven der empirischen Forschung zur politischen Gewalt. Unveröffentlichtes Manuskript, Universität Konstanz.

Schneider, Gerald und Vera E. Troeger. 2006. War and the World Economy: Stock Market Reactions to International Conflicts. *Journal of Conflict Resolution* 50 (5): 623-645.

Schneider, Gerald und Gabriele Ruoff. 2010. Quantitative Methoden der Internationalen Politik. In *Handbuch der Internationalen Politik*, Hrsg. Carlo Masala, Frank Sauer und Andreas Wilhelm, 236-244. Wiesbaden: Verlag der Sozialwissenschaften.

Schneider, Gerald, Thomas Widmer und Dieter Ruloff. 1993. Personality, Unilateralism, or Bullying: What Caused the End of the Cold War? *International Interactions* 18 (4): 323-342.

Schneider, Gerald, Daniel Finke und Stefanie Bailer. 2010. Bargaining Power in the European Union: An Evaluation of Competing Game-Theoretic Models. Political Studies 58 (1): 85–103.

Schrodt, Phil. 2014. Seven Deadly Sins of Contemporary Quantitative Political Analysis. *Journal of Peace Research* 51 (2). Im Erscheinen.

Sekhon, Jasjeet S. 2008. The Neyman-Rubin Model of Causal Inference and Estimation via Matching Methods. In *The Oxford Handbook of Political Methodology*, Hrsg. Janet M. Box-Steffensmeier, Henry E. Brady and David Collier, S. 271–99, New York: Oxford University Press

Shadish, William R., Thomas D. Cook und Donald T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Belmont: Wadsworth.

Signorino, Curtis S. 1999. Strategic Interaction and the Statistical Analysis of International Conflict. *American Political Science Review* 93 (2): S. 279-97.

Signorino, Curtis S. 2003. Structure and Uncertainty in Discrete Choice Models. *Political Analysis* 11 (4): S. 316-44.

Simmons, Beth und Dan Hopkins. 2005. The constraining power of international treaties: Theory and methods. *American Political Science Review* 99 (4): S. 623–631.

Singer, J. David, Stuart Bremer und John Stuckey. 1972. Capability Distribution, Uncertainty, and Major Power War, 1820-1965. In *Peace, War, and Numbers*, Hrsg. Bruce Russett, 19-48. 1. Auflage. Beverly Hills und London: Sage Publications.

Sovey, Allison J. und Donald P. Green. 2011. Instrumental Variables Estimation in Political Science: A Readers' Guide. *American Journal of Political Science* 55(1): 188-200.

Steenbergen, Marco R. und Bradford S. Jones. 2002. Modeling Multilevel Data Structures. *American Journal of Political Science* 46 (2): 218-37.

Thomson, Robert, Frans N. Stokman, Christopher H. Achen und Thomas König. 2006. *The European Union Decides*. Cambridge: Cambridge University Press.

Tomz, Michael. 2007. Domestic Audience Costs in International Relations: An Experimental Approach. *International Organization* 61 (4): 821-40.

Vasquez, John. 1997. The Realist Paradigm and Degenerative versus Progressive Research Programs: An Appraisal of Neotraditional Research on Waltz's Balancing Proposition. *American Political Science Review* 91 (4): 899–912.

von Stein, Jana. 2005. Do Treaties Constrain or Screen? Selection Bias and Treaty Compliance. *American Political Science Review* 99 (4): 611–622.

Waltz, Kenneth J. 1979. *Theory of International Politics, Reading.* Mass: Addison-Wesley.

Ward, Michael D. und Kristian Skrede Gledisch. 2008. *Spatial Regression Models.* Thousand Oaks: Sage.

Ward, Michael D., Brian D. Greenhill, und Kristin M. Bakke 2010. The Perils of Policy by P-Value: Predicting Civil Conflicts. *Journal of Peace Research* 47 (4): 363-375.

Western, Bruce. 1996. Causal Heterogenity in Comparative Research: A Bayesian Hierarchical Modelling Approach. *American Journal of Political Science* 42 (4): 1233-59.

Wooldrigde, Jeffrey. 2010. *Econometric Analysis of Cross Section and Panel Data.* Cambridge, MA: MIT Press.