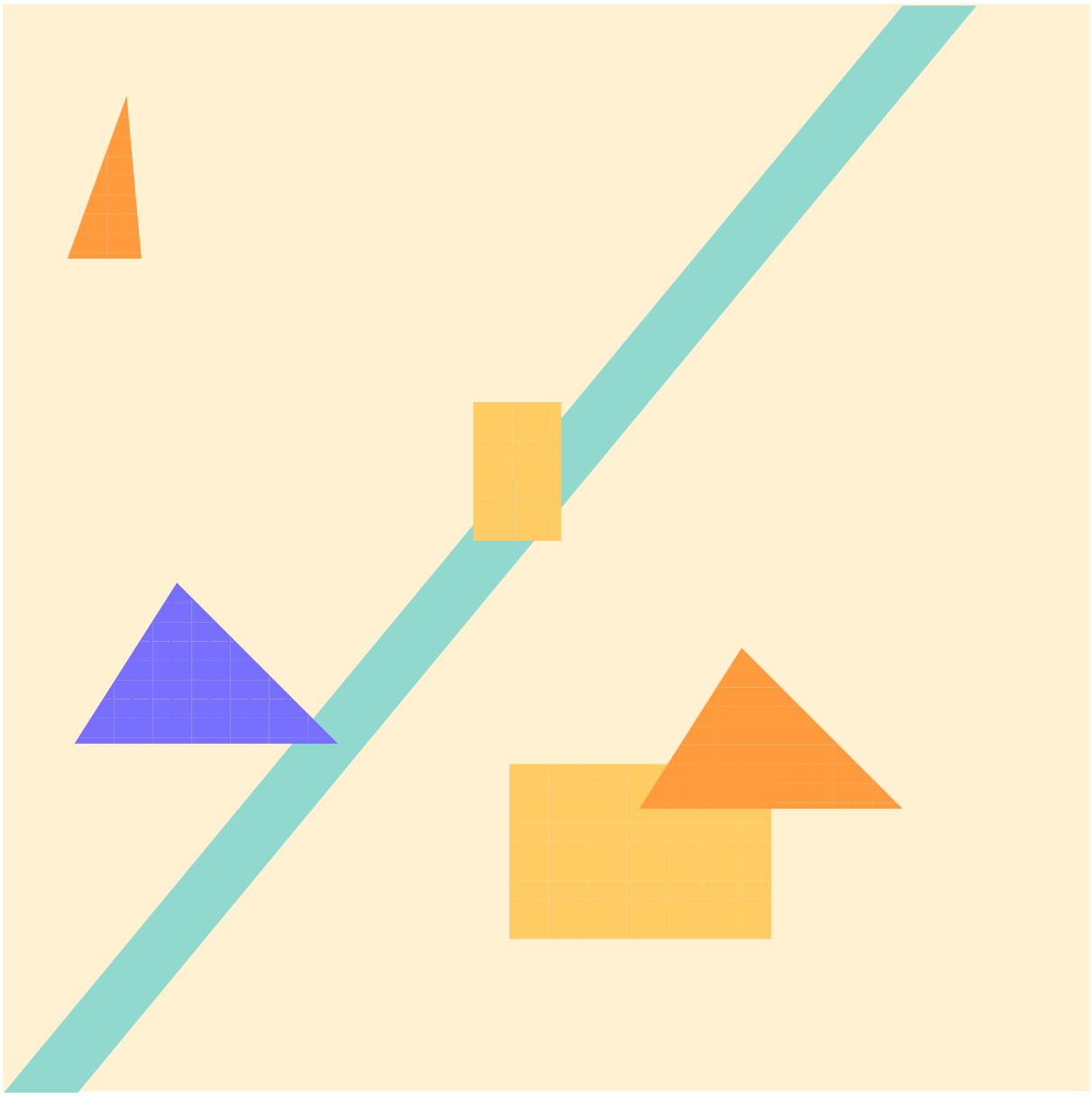
Dagmar Neubauer

Fuzzy-Regression bei Fehlern in den Daten

Modellierung und Analysepotentiale



Inaugural-Dissertation
zur Erlangung des Doktorgrades

Fuzzy-Regression bei Fehlern in den Daten

— Modellierung und Analysepotentiale

Vorgelegt von
Dipl.math.oec. Dagmar Neubauer

2. Dezember 2009

Fachbereich Wirtschaftswissenschaften
der Johann Wolfgang Goethe-Universität
Frankfurt am Main

Erstgutachter:	**Prof. Dr. Heinrich Rommelfanger**
Zweitgutachter	**Prof. Dr. Uwe Hassler**

Tag der Promotion: **3. März 2010**

Vorwort

Bei der Erstellung meiner Dissertation wurde ich von vielen Menschen unterstützt und begleitet. Sie alle haben Anteil an der vorliegenden Arbeit und ich möchte mich dafür ganz herzlich bedanken. Einigen Personen möchte ich besonderen Dank aussprechen:

Bedanken möchte ich mich zuallererst bei meinem Doktorvater, Herrn Prof. Dr. Heinrich Rommelfanger, für die verlässliche Unterstützung während meiner gesamten Zeit als Doktorandin und für seine stete Bereitschaft sich auf neue und ungewöhnliche Herangehensweisen einzulassen. Danken möchte ich außerdem meinem Zweitgutachter, Prof. Dr. Uwe Hassler, für sein Gutachten und seinen Langmut. Dem Kreis der Teilnehmer/innen an den Doktorand/innenkolloquien der Professur gilt Dank für das Feed Back und den kreativen Austausch, auch wenn ich Euch manchmal mit meiner Theorielastigkeit sehr geplagt habe. Nicht zuletzt ist dem Verein der Freunde und Förderer der Goethe-Universität für die Projektförderung zur Erstellung des empirischen Teils der Arbeit zu danken.

Für ihren besonderen Beitrag zum Inhalt der Arbeit möchte ich mich bedanken bei Prof. Dr. Renate Neubäumer (Universität Koblenz-Landau) für viele praktische Ratschläge und für die Heranführung an die Datenungenauigkeiten in der Arbeitsmarktstatistik; und ganz besonders bei PD Dr. Volker Krätschmer (Weierstraß-Institut für Angewandte Analysis und Stochastik, Berlin) für die engagierten Streitgespräche, die wesentlich zur Trennschärfe des mathematischen Begriffsapparates beigetragen haben.

Zugleich mit der Erstellung meiner Promotion hatte ich fast über die ganze Zeit ein „zweites Standbein“ in der Gleichstellungspolitik. Das war nur mit Konzentration und Beschränkungen in beiden Arbeitsbereichen zu leisten und war nur möglich mit dem solidarischen Rückhalt durch Kolleg/innen, Kooperationspartner/innen und Vorgesetzte. Herzlichen Dank an das Team im Frauenbüro: ich vermisse vor allem die Zertifizierung meiner Outfits und die mütterlichen Ratschläge. Ganz ausdrücklich möchte ich mich bei meiner Kollegin und Mit-Frauenbeauftragten, Prof. Dr. Ulla Wischermann, für die enge Zusammenarbeit in guten und in schlechten Zeiten und für die freundschaftliche Begleitung bedanken.

Für Lektorierungstätigkeiten in unterschiedlichen Stadien der Arbeit herzlichen Dank an Heidi für den anderen Blick, Nina für die Formulierungsvorschläge und insbesondere Dietmar für seine klugen Anmerkungen und die intensive Auseinandersetzung mit der mathematisch-technischen Argumentation.

Schließlich danke ich meinen Eltern und meinen Brüdern für vielfältige Hilfestellungen und Ratschläge; Martin und den Altenkirchenern für die ungezählten Gelegenheiten zum Tapetenwechsel und die unkomplizierte Aufnahme; Henri für die Lösung der technischen Probleme bei der Erstellung von Schwarz-Weiß-Grafiken; Karin für Freundschaft und Zuspruch; Vera für ein offenes Ohr bei allen Schreibkrisen; Kika für die erholsame Betreuung und die kulinarischen Köstlichkeiten während des Endspurts; Sabine M. für die Geschichten über das rasselnde Gespenst sowie Thomas und der gesamten Silvestergruppe für den letzten Anstoß. Außerdem danke ich meinem Schicksal von Herzen dafür, dass mein PC erst nach der Disputation kaputtgegangen ist.

Im Angedenken an Sabine K., die viel zu früh ihr Leben loslassen musste.

Inhaltsverzeichnis

Abbildungsverzeichnis

1 Einleitung

1.1 Einführung und Zielsetzung

Die Aussagekraft einer ökonometrischen Analyse hängt davon ab, ob die empirischen Daten, auf denen die Analyse aufbaut, tatsächlich die Phänomene und Begriffe abbilden, die der Modellbildung zugrundeliegen. Die Frage der Datenqualität ist von zentraler Bedeutung, weil die statistische Erfassung von Daten zu ökonomischen Fragestellungen stark von der Genauigkeit der sprachlichen Definitionen, von subjektiven Bewertungen sowie vom reibungslosen Ablauf der Messprozesse beeinflusst wird. Das Messergebnis ist somit sensibel für Störeinflüsse. Besondere Probleme können dabei nicht-zufällige Störeinflüsse bereiten, da ihre Auswirkungen häufig nur schwer beschrieben werden können. Trotz der Bedeutung der Datenqualität werden die möglichen Auswirkungen von Datenfehlern bei ökonometrischen Analysen häufig vernachlässigt. Begründet wird dies damit, dass die Fehler in den Daten nicht hinreichend genau bekannt sind, um sie zu korrigieren. Daher sei es letztlich besser, mit den verzerrten Daten zu arbeiten, als diese aufgrund einer Bereinigung mit falschen Vorannahmen möglicherweise weiter zu verfälschen und sich überdies dem Vorwurf auszusetzen, dass die Daten theoriegeleitet nachgesteuert wurden.

Die Fuzzy-Mengen-Theorie bietet eine Möglichkeit, vage und ungenaue Werte durch graduellen Zugehörigkeiten zu einer Menge abzubilden. Daher wird vorgeschlagen, die Behandlung fehlerhafter Daten in der Ökonometrie auf eine andere Basis zu stellen, indem Einschränkungen in der Datenqualität mit Hilfe von Fuzzy-Mengen modelliert werden. Eine zentrale Methode zur ökonometrischen Datenanalyse ist die lineare Regression. Im Mittelpunkt dieser Arbeit steht deshalb die Frage, wie Fehler in den Merkmalswerten in geeigneter Weise als Fuzzy-Menge modelliert werden können und welche Konsequenzen dies für die Analyse von funktionalen Zusammenhängen durch eine Regressionsschätzung hat.

Die ersten Arbeiten zur Erweiterung der Regressionsanalyse durch den Einsatz von Fuzzy Mengen-Theorie wurden ab 1985 vorgelegt.[1] Diese eröffneten eine Phase intensiver

[1] Zu den grundlegenden Arbeiten gehören u.a.: [Heshmaty und Kandel, 1985; Tanaka, 1987; Diamond, 1990] sowie der Sammelband [Kacprzyk und Fedrizzi, 1992].

Forschungstätigkeiten zu den Möglichkeiten der Fuzzy-Regression bis etwa 1997, in der eine Vielzahl von methodischen Weiterentwicklungen vorgestellt wurden, seitdem gibt es nur noch vereinzelte Veröffentlichungen dazu. Der Schwerpunkt der Forschung lag bisher überwiegend bei der technischen Weiterentwicklung der Fuzzy-Regression, praktische Relevanz haben die Methoden aber nicht erreicht. Als Ursache dafür ist zu sehen, dass es trotz der Methodenfülle nur wenige Ansätze gibt, in denen die Interpretation der Regressionsergebnisse systematisch hergeleitet und motiviert wird.

Tatsächlich gibt es in der Literatur keine einheitliche Definition und Abgrenzung von Fuzzy-Regression. Unter dem Begriff der Fuzzy-Regression werden sehr unterschiedliche Analysekonzepte diskutiert, die z.T. nur in einem sehr weiten Sinne als Regressionsansätze aufgefasst werden können. Die Vielfalt der Methoden ist u.a. im Sammelband von Kacprzyk und Fedrizzi [1992] vertreten. Die Bandbreite reicht hierbei von der Approximation eines Fuzzy-Reglers an scharfe Daten, wie bei der „weichen“ Regression (engl. „soft regression“) gemäß [Niskanen, 2001] über die Possibilistische Regression gemäß [Tanaka und Ishibuchi, 1992] bis hin zu induktiven Ansätzen einer Kleinsten Quadrate Regression auf der Basis von Fuzzy-Zufallsvariablen, vgl. dazu [Körner und Näther, 1998; Näther, 2000; Krätschmer, 2006a].

Als Hauptproblem bei der Modellierung von empirischen Analysezielen für die Methoden der Fuzzy-Regression stellt sich dar, dass die Fuzzy-Daten stark kontextabhängig modelliert werden, was seinerseits zu einer starken Kontextabhängigkeit der Interpretation führt und diese erschwert. Die Flexibilität der Modellierung und Interpretation von Fuzzy-Daten bringt Vorteile, aber auch Nachteile für die Datenanalyse. Dem Vorteil einer realistischeren Darstellung der Kenntnisse über die Daten und ihrer Qualität steht der Nachteil gegenüber, dass im extremen Fall jede Betrachtung als Einzelfall behandelt werden und somit die Vergleichbarkeit der Daten erheblich eingeschränkt sein kann. Die Modellierung von Fuzzy-Daten und deren Interpretation beziehen sich also stark auf den konkreten Untersuchungsgegenstand. Dieser interpretatorische Kontext wurde bei der Methodenentwicklung aber nur unbefriedigend berücksichtigt, häufig sogar völlig vernachlässigt. So werden die Methoden bisher nur anhand sehr weniger, in der Mehrzahl artifizieller Datenbeispiele vorgeführt.[2] Zudem beklagen Redden und Woodall [1996] im Bezug auf die

[2]Insgesamt werden etwa fünf Datenbeispiele immer wieder herangezogen. Gerne zitiert wird z.B. das Beispiel bei dem eine Regression von Ausstattungsvariablen wie Qualität des Baumaterials, Geschossfläche im 1. und im 2. Stock sowie Zahl der Räume auf den Kaufpreis japanischer Häuser aufgestellt wird, vgl. [Tanaka u. a., 1995]. Der Datensatz umfasst 15 Datensätze. Die Frage, ob die Qualität des Baumaterials als kategoriale Variable mit den drei Qualitätsstufen niedrig, mittel, hoch sinnvoll in die Regression einbezogen werden kann, wird nicht erörtert. Ein anderes vielzitiertes Beispiel stammt aus einer tatsächlich durchgeführten Untersuchung über die subjektive Bewertung der Wiedergabequalität von Videobildschirmen, die in [Chang u. a., 1996] präsentiert wird. Es handelt sich dabei um einen bivariaten Ansatz mit sechs Datenpunkten pro Bildschirmtyp. Die Daten dienen vor allem dazu, die Leistungsfähigkeit der jeweiligen Technik vorzuführen.

Possibilistische Regression das Fehlen einer Interpretation für die Fuzzy Regressionsparameter. Der Nutzen von Methoden zur Fuzzy-Regression für die praktische Anwendung bleibt damit ungeklärt.

Da bereits vielfältige Ansätze zur Fuzzy-Regression vorliegen, besteht das Ziel dieser Arbeit nicht darin, weitere Regressionsansätze für Fuzzy-Daten zu entwickeln. Stattdessen ist es das Hauptanliegen der vorliegenden Untersuchung, die interpretatorische Lücke zwischen den Datenproblemen in empirischen Datensätzen, deren Modellierung als Fuzzy-Daten und den Aussagen und Analysen zu schließen, die auf der Basis einer Fuzzy-Regression über die vorliegenden Daten getroffen werden können. Im Zentrum der Überlegungen steht also nicht die Konstruktion der Fuzzy-Methoden als solche, sondern vielmehr die fehlende Vorstellung über die mögliche Bedeutung der Fuzzy-Modellierungen im konkreten Anwendungsfall. Die Arbeit ist so angelegt, dass die einzelnen Arbeitsschritte von der Konstruktion der Fuzzy-Daten bis zur Interpretation der Fuzzy-Regression im Sinne einer Machbarkeitsstudie nachvollzogen werden. Dabei werden für jeden der Schritte exemplarische Beispiele ausgewählt, um die Denkweise der Fuzzy-Modellierung auszuarbeiten und die Möglichkeiten und Grenzen der Methoden auszuloten. In einer abschließenden Auswertung werden die einzelnen Fallbeispiele unter dem Blickwinkel reflektiert, inwieweit einzelne Ergebnisse auf andere Anwendungsfälle übertragen werden können und verallgemeinerbar sind. Auf diese Weise wird die Einzelfallbetrachtung mit den Fragestellungen der abstrakten Modellbildung verknüpft, und es können Anforderungen an die Modell- und die Methodenentwicklung formuliert werden, die aus einem konkreten Anwendungsbezug heraus motiviert sind.

Um ein möglichst vollständiges Bild über die Modellvorstellungen für den Zusammenhang zwischen den Fuzzy-Variablen zu erhalten, sollten bei der Aufarbeitung der Methoden zur Fuzzy-Regression möglichst viele Fuzzy-Ansätze einbezogen werden. Dabei ist die überwiegende Zahl von Fuzzy-Regressionsmethoden nur auf eine explorative Datenanalyse ausgerichtet. Außerdem bieten einige der Methoden ein Verfahren an, mit dem die Streuung der Daten um den funktionalen Zusammenhang subjektiv bewertet werden kann. Dies erfordert die Auseinandersetzung mit Ansätzen, die eine Auffassung von Häufigkeitsverteilungen abbilden, die nicht durch klassische Zufallsvariable beschrieben werden können. Es erschien daher dringend geboten, strikt zwischen zufallsinduzierten Störungen der Daten, die als Zufallsvariable abgebildet werden können, und nicht-zufälligen Störungen der Daten, die als Fuzzy-Mengen abgebildet werden können, zu unterscheiden. Dies wurde für notwendig erachtet, um in einem ersten Schritt zu analysieren, wie die Datenunschärfe, die mit Fuzzy-Daten abgebildet werden kann, sich bei der Funktionsapproximation auswirkt und welche Bedeutung dies für die Interpretation der Approximationsergebnisse hat. Dabei ist die Unterscheidung zwischen der Modellierung von zufälligen und von nicht-

zufälligen Einflüssen insbesondere für die Analyse von fehlerhaften Daten von zentraler Bedeutung. Die Untersuchung konzentriert sich daher zunächst auf explorative Regressionsmethoden und somit auf Methoden zur Approximation von linearen Modellfunktionen. Erst auf der Basis der Begrifflichkeiten und Mechanismen, die sich als Rahmen für die Interpretation der Analyse von Fuzzy-Daten mit einer explorativen Regression ergeben, ist die Einbettung in ein Verteilungsmodell und die Modellierung eines Schätzmodells für die Fuzzy-Regression bei Fehlern in den Daten sinnvoll möglich.

Als komplementäre Erweiterung zu den stochastischen Regressionsansätzen interessieren uns vor allem solche Ansätze, mit denen Strukturen in den Daten identifiziert werden können. Wir beschränken uns daher im Folgen auf die Methoden zur Fuzzy-Regression, bei denen eine Fuzzy-lineare Funktion an alle Datenpunkte zugleich anzupassen ist, d.h. bei denen die Daten global mit einer parametrisierten Funktion approximiert werden.

1.2 Aufbau der Arbeit

Nach einer Einführung in die allgemeinen Qualitätsprobleme bei der Erhebung von Wirtschaftsdaten und den möglichen Verzerrungen von Regressionsparametern aufgrund von Datenfehlern in Kapitel *2* werden in Kapitel *3* die vorliegenden Methoden zur Fuzzy-Regression kritisch erörtert. Dabei werden einerseits die impliziten Annahmen an die Daten herausgearbeitet, und andererseits wird untersucht, welche Informationen in den Daten durch die Fuzzy-Regressionsfunktion verdichtet werden.

Um den Besonderheiten der einzelfallbezogenen Modellierung von Fuzzy-Daten gerecht zu werden, wird in Kapitel *4* die Konstruktion von empirischen Fuzzy-Merkmalswerten anhand einer empirischen Datenquelle der amtlichen Statistik als Fallbeispiel vorgeführt. Dazu werden die Messfehler in den Merkmalen der Beschäftigtenstatistik analysiert. Ausgehend von den vorliegenden Messdaten werden dann prototypische Modellierungen von empirischen Fuzzy-Daten entwickelt. Die Modellierung erfolgt hierbei grundsätzlich ex post. Die Wahl fiel auf die Beschäftigtenstatistik, weil sie eine Totalstatistik ist. Das hat den Vorteil, dass die Zufallsfehler entfallen, die durch die Verteilung der Stichprobe in der Grundgesamtheit verursacht werden. Als Teilstatistik im Rahmen der amtlichen Arbeitsmarktstatistik hat die Beschäftigtenstatistik außerdem den Vorzug, dass eine ausführliche und kontinuierliche Dokumentation der Probleme bei der Datenerfassung und der Datenqualität bereitgestellt wird. Von zusätzlichem Interesse für die Fehlermodellierung ist die Tatsache, dass amtliche Statistiken häufig eine große Vielfalt von Unschärfequellen aufweisen, etwa durch den Einfluss administrativer Vorgaben auf Datenerhebung und Datengehalt oder durch Methodenwechsel in der Erhebung.

Um den Zusammenhang zwischen der Fehlermodellierung in den Fuzzy-Daten, dem Approximationsmodell und dem Approximationsergebnis zu verdeutlichen, wird in Kapitel *5* ein Benchmarking zwischen verschiedenen Varianten der Kleinste Quadrate Fuzzy-Regression durchgeführt, die mit der klassischen Kleinste Quadrate-Anpassung in den „wahren“ und den fehlerhaften Werten ins Verhältnis gesetzt werden. Da es normalerweise nicht möglich ist, die „wahren“ Werte zu ermitteln, die hinter einem fehlerhaften Messergebnis verborgen sind, werden für das Benchmarking simulierte Fuzzy-Daten verwendet. Für die Simulationen werden die Koinzidenzen zwischen „wahren“ und fehlerhaften Werten sowie deren Abbildung als Fuzzy-Zahl mit Hilfe von Fehlerszenarien konstruktiv beschrieben.

2 Theoretische Grundlagen

Im Rahmen der Wirtschafts- und Sozialstatistik werden Messdaten zur Deskription des realen Geschehens erhoben, bearbeitet und zur Verfügung gestellt. Die Qualität dieser Daten kann allerdings nur relativ beurteilt werden, wie Morgenstern in seiner grundlegenden Arbeit herausstellt: „Es ist klar, daß jeder, der Meßergebnisse und Daten benützt, wünscht, daß diese genau seien und (...) das Wesentliche erfassen. (...) Die Frage, ob ein Meßergebnis, eine Beobachtung, Beschreibung oder Zählung — worum immer es sich im konkreten Fall handeln mag — genau und brauchbar ist, läßt sich nur im Hinblick auf den angestrebten Verwendungszweck beantworten. Mit anderen Worten: will man sinnvoll über Genauigkeit diskutieren, so muß eine Statistik immer im Hinblick auf eine — wenn auch noch so grob formulierte — Theorie oder ein Modell oder einen bestimmten Zweck formuliert werden“.[1] Eine absolute Messgenauigkeit von Daten ist also ausgeschlossen. Der reale Zustand des Geschehens kann nur ungefähr beschrieben werden.

Grundsätzlich stellen Beobachtungen mit Hilfe unserer Sinnesorgane oder anderer Instrumente zur Wahrnehmung und Beschreibung der Realität nur ein unvollkommenes Abbild des realen Geschehens dar. Gründe dafür bestehen unter anderem in der Komplexitätsreduktion, die unsere Wahrnehmung prägt, in der Unvollkommenheit unserer Messinstrumente und darin, dass die Realität sich in ständiger Veränderung befindet. In diesem Sinne beschreiben Beobachtungsdaten die Realität immer nur ungefähr. Die Unvollkommenheit der Daten zur Beschreibung des realen Geschehens wird hier ganz allgemein als *Datenunschärfe* bezeichnet. Es gibt eine Vielzahl von Strategien, mit dieser Unschärfe der Daten umzugehen. In dieser Arbeit steht die quantitative Beschreibung von Daten-unschärfen im Mittelpunkt. Die entsprechenden Modellierungen lassen Zwischenwerte zu, die die Grenzsituation zwischen Existenz und Nicht-Existenz bzw. zwischen Beobachtung und Nicht-Beobachtung abbilden. Wir wollen dafür die folgende Unterscheidungen treffen[2]. Ein vertrautes und ausgearbeitetes Konzept von Datenunschärfe ist das der *Unsicherheit*. Unsicherheit über eine Merkmalsbeobachtung liegt vor, wenn wir den Zustand des Merkmals bei einem zufällig ausgewählten Subjekt aus einer Grundmenge messen, wobei der Merkmalszustand bei den Subjekten der Grundmenge in unterschied-

[1] Vgl. Morgenstern 1965, S. 2.

[2] Zur Unterscheidung der Begriffe vgl. [Bandemer, 1997].

lichen Ausprägungen vorliegt. Hier bewirkt der Einfluss des Zufalls, dass wir unsicher darüber sind, wie bei allen möglichen bzw. bei einem weiteren zufällig ausgewählten Subjekt der Merkmalswert ausfallen würde. Unsicherheit kann mit Wahrscheinlichkeiten gut beschrieben werden, sie kann durch Einholen besserer Information nicht verringert werden.

Im Unterschied zur Unsicherheit sprechen wir von *Ungenauigkeit*, wenn aufgrund von Wissenseinschränkungen nur ungefähr angegeben werden kann, welches der reale Zustand des Merkmals ist. Ungenauigkeit über den Merkmalswert liegt z.B. vor, wenn das Messverfahren unzuverlässig oder verzerrt ist oder wenn die Beobachtungsmöglichkeiten aus Kosten- oder technischen Gründen eingeschränkt sind. Ungenauigkeit kann durch Einholen von zusätzlicher Information bzw. von Verbesserungen im Messverfahren verringert werden.

Als ein weiterer Unschärfeaspekt soll außerdem *Vagheit* eingeführt werden. Vagheit ist Ausdruck der Unvollkommenheit der sprachlichen Beschreibung der Realität und liegt dann vor, wenn die quantitative Operationalisierung eines sprachlichen Begriffs nur in Form einer mehrwertigen Menge möglich ist, die Teilzugehörigkeiten umfassen kann. Z.B. kann nicht eindeutig angegeben werden, welche Farbwerte der Aussage entsprechen, dass eine Tomate „rot“ ist. Auch Vagheit kann durch Einholen zusätzlicher Information nicht verringert werden, sie enthält aber offensichtlich keine Zufallseinflüsse.

In Abschnitt *2.1* wird ein Überblick über die allgemeinen Qualitätsprobleme bei der Erhebung von Wirtschaftsdaten gegeben. Am Beispiel des klassischen Fehlermodells als einfachem Fehlerstrukturmodell werden dann in Abschnitt *2.2* die möglichen Verzerrungen des Regressionsergebnisses durch fehlerhafte Daten illustriert. Mit der Kritik am ökonometrischen Fehlermodell in Abschnitt *2.3* wird der Perspektivwechsel von den klassischen Schätzmodellen bei Fehlern in den Variablen zu Fuzzy-Modellen bei Fehlern in den Daten eingeleitet. Abschnitt *2.4* gibt zunächst eine knappe Einführung in die Grundlagen der Fuzzy-Mengen-Theorie. Schließlich werden in Abschnitt *2.5* Fuzzy-Merkmalswerte definiert und als Konzept zur Modellierung von fehlerhaften Daten beschrieben.

2.1 Datenqualität in der Wirtschafts- und Sozialstatistik

In diesem Abschnitt sollen die wesentlichen Probleme der Datenqualität in der Wirtschafts- und Sozialstatistik rekapituliert und die Probleme bei der Quantifizierung von Fehlereinflüssen charakterisiert werden.

Die Wirtschafts- und Sozialstatistik ist dadurch gekennzeichnet, dass Daten nahezu ausschließlich als Auskunftsdaten erhoben werden. Die statistischen Begriffe, die den Erfassungsvorgängen zugrundeliegen, werden daher im Bezug auf gesellschaftliche, politische und wirtschaftliche Anforderungen gebildet. Sie verkörpern oft Kompromisse, die eine zeitliche oder internationale Vergleichbarkeit sicherstellen oder eine vielseitigere Nutzung der Daten[3] ermöglichen sollen, die damit gemessen werden. Für die Konzeption von Statistiken bedeutet dies insgesamt, dass das, „was beobachtet, erhoben werden kann, (...) wesentlich von institutionellen Vorgaben ab[hängt]. Die Wirtschaftsverfassung, die Steuergesetzgebung und ähnliche Vorhaben prädeterminieren die Definition von Merkmalen, bestimmen die Möglichkeiten der Abgrenzung von Erhebungsmassen, etc.“[4]. Hinzu kommt, dass die Datengewinnung häufig in einem hoch arbeitsteiligen Prozess erfolgt, in dem „die elementare Rezeption der Wirklichkeit (...) an jene — die Respondenten — delegiert [wird], deren Tun auch Gegenstand der Beobachtung ist“[5]. Die Kommunikations- und Verständigungsprozesse bei der Erfassung der Rohdaten sind daher von essentieller Bedeutung für die Datenqualität.

Da für die Messung nur begrenzte Ressourcen zur Verfügung stehen, wird der Rahmen dessen, was faktisch beobachtet werden kann, außerdem von Kosten-Nutzen-Aspekten determiniert. Aus diesem Grund — aber auch deshalb, um eine zu starke Belastung und Übermüdung der Befragten zu vermeiden — wird ein großer Anteil der Daten im Zuge von administrativen Prozessen erhoben, z.B. der Steuer- oder Sozialverwaltung. Etwa die Hälfte der amtlichen Statistiken in Deutschland sind solche Sekundärstatistiken, die auf Register der öffentlichen Verwaltung basieren.[6] Diese Daten spiegeln meistens stärker die gesetzlichen Anforderungen wieder als die Zielgrößen aus der wirtschaftwissenschaftlichen Theoriebildung. Griliches [1986] charakterisiert solche Daten daher als „found data“[7].

Schließlich können wirtschaftsstatistische Messungen nur bedingt unter unveränderten Bedingungen wiederholt werden. Damit sind die Möglichkeiten stark eingeschränkt, mit denen die Richtigkeit der vorliegenden Daten überprüft und das Messverfahren justiert werden kann. Insgesamt ist zu konstatieren, dass der Erfassungsvorgang typischerweise weniger stabil als bei technischen Messverfahren und zugleich die Abbildung von Datenunschärfen deutlich erschwert ist.[8]

[3] Richter [2002] bezeichnet dies als „multi purpose“-Charakter der statistischen Begriffe (S. 14).

[4] Vgl. Richter 2002, S. 7.

[5] Vgl. ebd. S. 7.

[6] Vgl. Statistische Ämter des Bundes und der Länder 2006, S. 22.

[7] Vgl. ebd. S. 1466.

[8] Die Untersuchung von Morgenstern [1965] zu Messfehlern in der Wirtschafts- und Sozialstatistik gilt als grundlegend und wird weiterhin gerne zitiert. Als zentral für die deutsche Statistik können außerdem die Arbeiten von Strecker und Wiegert gelten, die u.a. Schätzverfahren zur Berücksichtigung der Antwortvariabilität untersuchen, hier etwa [Strecker und Wiegert, 1994; Strecker, 2004]. In neuerer Zeit wird darüber

Im Zuge der Qualitätsentwicklung für die amtlichen Statistiken stellen die statistischen Ämter des Bundes und der Länder seit Anfang 2006 Qualitätsberichte für alle Statistiken als Service für die Nutzer/innen zur Verfügung. Diese enthalten strukturierte Informationen über die Qualität der statistischen Ergebnisse sowie Angaben zu den verwendeten Methoden und Definitionen, die eine sachgerechte Nutzung der Statistiken ermöglichen sollen. Somit liegen nun systematische Darstellungen der Fehlereinflüsse vor, die trotz aller Einschränkungen eine bessere Vergleichbarkeit zwischen den Datenungenauigkeiten herstellen.

Den Qualitätsberichten werden einheitliche Qualitätsstandards zugrundegelegt, die es zulassen, die Aspekte der Datenunschärfe besser zu beschreiben.[9] Wir verwenden sie daher als Ausgangspunkt für die weiteren Untersuchungen. Die Qualitätsstandards beziehen sich nicht nur auf die Datenqualität der Statistiken im engeren Sinne sondern auch auf den institutionellen Rahmen, z.B. fachliche Unabhängigkeit, Neutralität, Objektivität sowie den Datenschutz, und auf die Prozesse zur Erfassung und Auswertung der Daten, z.B. Verwendung adäquater Verfahren und Methoden sowie vor allem auch die Vermeidung einer übermäßigen Belastung der Befragten. Insbesondere können die Fehlereinflüsse überwiegend lediglich qualitativ beschrieben werden. Als Qualitätskriterien für die statistischen Produkte im engeren Sinne werden die folgenden sechs Ziele definiert:[10] *Relevanz*, die sich darin ausdrückt, in welchem Maß die Daten den Anforderungen der Nutzer/innen entsprechen. *Genauigkeit*, d.h. der Nähe des gemessenen Wertes zum „wahren", aber unbekannten Wert. *Aktualität und Pünktlichkeit*, d.h. die Veröffentlichung der Daten soll möglichst zeitnah erfolgen. Dazu sind Veröffentlichungstermine festzulegen und bekanntzugeben, die eingehalten werden. Für wichtige Statistiken, bei denen die Aktualität höchste Priorität hat, werden vorläufige Ergebnisse veröffentlicht. *Verfügbarkeit und Transparenz*, d.h. neben der leichten Verfügbarkeit der Statistiken für die Nutzer/innen müssen die Ergebnisse auch hinsichtlich Konzept und Methoden vollständig dokumentiert sein. *Vergleichbarkeit*, d.h. die Ergebnisse sollen zeitlich, räumlich und fachlich vergleichbar sein. Insbesondere werden einheitliche Standards in Bezug auf die Definitionen, die Einheiten, die Merkmale und die Klassifikationen verwendet, die möglichst internationale Gültigkeit haben. *Kohärenz* soll in dem Sinne gewährleistet sein, dass unterschiedliche Statistiken, die sich auf die gleiche Grundgesamtheit beziehen, möglichst widerspruchsfrei untereinander in Beziehung gesetzt werden, damit Statistiken aus verschiedenen Quellen

hinaus auch der leichtfertige Umgang mit der „Theoriehaltigkeit" von amtlichen Daten problematisiert, hier vor allem bei der volkswirtschaftlichen Gesamtrechnung sowie bei internationalen Statistiken. Insbesondere Bereinigungen bzw. Glättungen von Daten werden in vielen Fällen mit Bezug auf theoretische Modelle durchgeführt. Zu diesem Problembereich vgl. [Holub und Tappeiner, 1995, 1997; Froeschl, 1999; Richter, 2002].

[9] Vgl. [Statistische Ämter des Bundes und der Länder, 2006].

[10] Vgl. ebd. S. 6ff.

kombiniert und gemeinsam verwendet werden können.[11] Die Statistiken sollen also als Gesamtsystem entwickelt werden, in dem die Einzelstatistiken eine bestimmte Funktion übernehmen.

Zwischen den Qualitätszielen gibt es Konkurrenzen. Häufig besteht ein wesentlicher Gegensatz zwischen Genauigkeit und Aktualität, etwa wenn Wartezeiten bei der Datensammlung zu berücksichtigen sind.[12] Um den Zielkonflikten zu begegnen, müssen bei der Gestaltung des Erhebungsverfahrens Prioritäten gesetzt werden, „die je nach Interessenschwerpunkt des Anwenders selbst innerhalb der einzelnen Forschungsdisziplinen unterschiedlich ausfallen dürften“[13]. Die Datenunschärfen sollen nun ausgehend von den Qualitätszielen der amtlichen Statistik genauer betrachtet werden.

Die Mehrzahl der Kriterien wie Relevanz, Aktualität, Vergleichbarkeit und Kohärenz sind auf die Konzeptualisierung der statistischen Einheiten und Begriffe ausgerichtet. Die Operationalisierung der zu messenden Merkmale sowie des Erhebungsverfahrens in einem Arbeitsmodell, in dem versucht wird, eine möglichst hohe Übereinstimmung mit den fachwissenschaftlichen Zielgrößen des Idealmodells sowie den darin enthaltenen theoretischen Vorstellungen zu erreichen, wird auch als *Adäquation* bezeichnet.[14] Ziel bei der Adäquation ist die Definition von eindeutigen Begriffen, so dass sowohl die Mess-Subjekte bzw. die statistischen Einheiten als auch deren Merkmalswerte konkret und zweifelsfrei bestimmt werden können.[15]. Krätschmer schlägt vor, Datenunschärfen, die durch die Adäquation induziert sind, als Vagheit der statistischen Begriffe aufzufassen. In seiner Habilitationsschrift stellt er ein entsprechendes Messmodell für Fuzzy-Variable vor, in weiteren Arbeiten hat er weitreichende Ergebnisse zur Regressionsanalyse bei vagen Fuzzy-Konzepten vorgelegt.[16] In dieser Arbeit soll der Schwerpunkt demgegenüber auf den Datenunschärfen liegen, die auf eine mangelnde Genauigkeit der Daten zurückgehen.

[11] Das betrifft nicht nur die Verwendung identischer Klassifikationssysteme sondern reicht auch soweit, dass bei Befragungen identische Frageformulierungen eingesetzt werden.

[12] In [Statistische Ämter des Bundes und der Länder, 2006] wird aber darauf hingewiesen, dass eine Verbesserung der Aktualität nicht zwangsläufig mit einem Verlust an Genauigkeit verbunden sein muss. Aktualitätsgewinne können im Rahmen einer Effizienzsteigerung der Prozesse bei der Statistikerstellung ggf. auch bei gleichbleibender Genauigkeit erreicht werden. Dazu tragen neue Erhebungstechniken, die Einführung flexiblerer Auswertungssoftware, Elektronischer Datenaustausch und das Instrument der Bereitstellung vorläufiger Ergebnisse bei (S. 16).

[13] Vgl. Löbbe 1993, S. 47.

[14] Vgl. Strecker 1993, S. 26ff. Zu den Dimensionen der Datenunschärfe vgl. Diagramm 2.1 auf Seite 25.

[15] Allerdings kommt es vor, dass Merkmale nicht bei allen interessierenden Mess-Subjekten der Grundgesamtheit beobachtbar sind. So können z.B. Aufwendungen bei unselbständigen regionalen Unternehmen ohne eigene Buchführung und bei kleinen Unternehmen, die von der Buchführung befreit sind, nicht erfasst werden. Um die Vollständigkeit der Daten für dieses Merkmal zu gewährleisten, werden bei der Erhebung des Betriebspanels daher plausible Ersatzdaten aufgenommen, die nach einem vorab definierten Verfahren zu bestimmen sind.

[16] Für weitere Verweise s. [Krätschmer, 2001a, 2006a].

Ein *Fehler*[17] im Messergebnis liegt definitionsgemäß dann vor, wenn eine Einzelbeobachtung im Messergebnis vom „wahren“ Wert bei einer fehlerfreien Durchführung des Arbeitssystems abweicht.[18] Dabei ist das Messergebnis als Zusammenfassung aller bzw. einer ausgewählten Reihe von Einzelmesswerten aufzufassen. Die allgemeine Abweichung zwischen dem Erhebungsergebnis und dem unbekannten „wahren“ Wert der Grundgesamtheit wird als *Gesamtfehler* des Ergebnisses bezeichnet. Je größer der Gesamtfehler ist, desto geringer ist die Genauigkeit des Messergebnisses. Als Maßstab für die Genauigkeit des Messergebnisses wird eine Operationalisierung des Gesamtfehlers verwendet, die üblicherweise durch eine reellwertige Aggregation der Fehlereinflüsse bestimmt wird.

In Abhängigkeit von der Art der Erhebung werden die Fehlereinflüsse zunächst in stichprobenbedingte Fehler und nicht-stichprobenbedingte Fehler unterschieden. Diese Unterscheidung soll verdeutlichen, dass bei Stichprobenerhebungen besondere Fehlereinflüsse zu berücksichtigen sind: Zum einen die Stichprobenzufallsfehler, die infolge der Zufallsauswahl einer Teilmenge von Individuen und aufgrund der darauf basierenden Hochrechnungen entstehen. Zum anderen die nicht-zufälligen Stichprobenfehler, die auf systematische Ursachen zurückzuführen sind, die nur bei Stichprobenerhebungen auftreten, wie z.B. wenn Mängel hinsichtlich der Erhebungsgesamtheit oder Verzerrungen durch die Auswahlmethode bestehen oder wenn das Hochrechnungsverfahren an sich fehlerhaft ist.[19] Nicht-stichprobenbedingte Fehler treten sowohl bei Stichprobenerhebungen als auch bei Vollerhebungen auf. Es handelt sich dabei häufig um systematische Fehler, d.h. ihr Auftreten kann zur Verzerrungen des Ergebnisses führen.

Sofern Aussagen über die Ungenauigkeit von Messwerten aufgrund von Fehlern gemacht werden können, ist es üblich, eine Reihe von Einzelmessungen[20] zusammenzufassen und für diese den „mittleren quadratischen Gesamtfehler“ MSE (engl. *mean square error*) anzugeben. Es wird in der Regel angenommen, dass sich der mittlere quadratische Gesamtfehler additiv aus der Zufallskomponente und der systematischen Fehlerkomponente des Ereignisses zusammensetzt, d.h.

$$(\text{MSE})^2 = (\text{Standardfehler})^2 + (\text{Bias})^2. \tag{2.1}$$

[17] Die Abweichung des Messergebnisses vom „wahren“ Wert wird allgemein als „Fehler“ bezeichnet. Davon sind die „Messfehler“ als eine bestimmte Fehlerart zu unterscheiden, die in der Datenerhebungsphase auftreten. Sie werden folglich durch den Fragebogen, die Interviewer/innen oder die Befragten verursacht. Vgl. dazu Statistische Ämter des Bundes und der Länder 2006, S. 14 und S. 87.

[18] Vgl. Strecker 1993, S. 26ff. Tatsächlich ist auch das im Prozess der Adäquation festgelegte Arbeitssystem noch nicht eindeutig bestimmt und konkretisiert sich erst bei der tatsächlichen Durchführung der Erhebung.

[19] Vgl. Krug u. a. 2001, S. 217.

[20] Diese werden entsprechend als das Messergebnis aufgefasst.

Der *Standardfehler* ist ein Maß zur Beurteilung des Stichprobenzufallsfehlers und gibt die sog. Präzision des Ergebnisses wieder.[21] Unter der Annahme, dass die gemessenen Merkmalswerte entsprechend der Auswahlwahrscheinlichkeit der Stichprobe für alle Merkmalswerte der Grundgesamtheit repräsentativ sind, kann daraus der Stichprobenzufallsfehler geschätzt werden. Überdies können die zufälligen Stichprobenfehler durch Vergrößerung der Stichprobe kontrolliert und verringert werden. Der Stichprobenzufallsfehler ist somit vergleichsweise einfach zu bestimmen. Für wichtigere Erhebungsergebnisse bei Stichprobenstatistiken werden differenzierte Fehlerrechnungen durchgeführt,[22] so dass relevante Stichprobenfehler als hinreichend gut dokumentiert gelten können.

Hingegen ist eine quantitative Beschreibung des *systematischen Fehlers*, der alternativ auch als *Verzerrung* oder als *Bias* bezeichnet wird, und damit der anderen Komponente des Gesamtfehlers häufig nur tendenziell oder gar nicht möglich. Zur Beschreibung des Gesamtfehlers werden daher auch alternative Kennzahlen herangezogen wie z.B. Größenordnung und Vorzeichen des Bias bzw. dessen relatives Verhältnis zum Standardfehler oder aber die qualitative Bewertung des Gesamtfehlers sowie die Beschreibung der Fehlerarten, die bei der Bewertung des Gesamtfehlers zu berücksichtigen sind.[23] Zur Bestimmung der Größenordnung des Gesamtfehlers eignet sich auch der Vergleich mit unabhängigen Datenressourcen.[24] Im Unterschied zum Standardfehler kann die systematische Verzerrung im Gesamtfehler nicht durch Vergrößerung der Stichprobe verringert werden, sondern ausschließlich durch Verbesserung des Messverfahrens.

Nicht-stichprobenbedingte Fehler treten in allen Phasen des Datenerhebungs- und Aufbereitungsprozesses auf. Die nicht-stichprobenbedingten Fehler können wie folgt charakterisiert werden:[25] Fehler durch die Erfassungsgrundlage liegen vor, wenn z.B. statistische Einheiten der Grundgesamtheit nicht erfasst werden, mehrfach erfasst werden oder aber Einheiten fälschlich erfasst werden. Messfehler sind Fehler, die während der Datenerhebungsphase auftreten, sie werden z.B. durch missverständliche Fragen des Fragebogens, durch Falschauskünfte seitens der Befragten oder durch die Interviewer/innen verursacht, wenn diese durch ihr Auftreten das Antwortverhalten beeinflussen. Aufbereitungsfehler sind Fehler, die in der Phase der Aufbereitung der Daten entstehen, z.B. bei der Daten-

[21] Vgl. Statistische Ämter des Bundes und der Länder 2006, S. 83.

[22] Vgl. Statistische Ämter des Bundes und der Länder 2006, S. 39.

[23] Vgl. ebd. S. 83f.

[24] Vgl. beispielsweise Federal Committee on Statistical Methodology [2001], Kapitel 8, S. 3f. In dem Arbeitspapier des Federal Committee on Statistical Methodology ist eine sehr umfangreiche Übersicht über den aktuellen Stand der Methoden zur Fehlerreduzierung und zur Fehlerschätzung sowie das Fehlerreporting in den USA zusammengestellt. Allerdings können die Ergebnisse nur analog übertragen werden, da in den Analysen die Ungenauigkeit von Stichprobenstatistiken im Zentrum steht.

[25] Die Fehlersystematik ist ebenfalls Statistische Ämter des Bundes und der Länder [2006] entnommen (S. 79ff.). Dort wird zu allen Fehlerarten auch angegeben, mit welchen Angaben die Relevanz der jeweiligen Fehlerarten beschrieben werden können.

erfassung, der Codierung oder bei der logischen Bereinigung aufgrund von Plausibilitätsprüfungen.[26] Außerdem entstehen Fehler durch Antwortausfälle, sog. „nonresponse“. Dabei können ganze Einheiten entfallen, weil sie bei der Zählung nicht angetroffen werden oder die Teilnahme an der Erhebung verweigern, sog. „missing units“. Zudem kommt es zu fehlenden Merkmalswerten, sog. „missing items“, wenn eine Bestimmung des Merkmalswerts nicht möglich ist, wenn keine Antwort erfolgt oder ggf. auch wenn besondere Datenschutzvorgaben zu berücksichtigen sind. Die Fehlerwirkung von Antwortausfällen hängt davon ab, ob ein systematischer Zusammenhang zwischen antwortenden und nicht-antwortenden Einheiten besteht.[27]

Im Hinblick auf die Fuzzyfizierung von Datenungenauigkeiten soll hier noch besonders auf die Bedeutung von *Klassifikationsfehlern* hingewiesen werden. Klassifikationsfehler sind stark interdependent mit dem Adäquationsproblem, da Missklassifikationen in dem Maße zunehmen, wie Zuordnungen den Respondent/innen uneindeutig erscheinen. Ihre Zahl steigt also mit zunehmender Gliederungstiefe bzw. mit wachsender Komplexität der Klassifizierungsvorschrift tendenziell an.[28] Von Bedeutung für die Datenanalyse sind vor allem Missklassifikationen, die zu systematischen Abweichungen bei den Zuordnungen führen. Ein Ausweg besteht hauptsächlich in der Vereinfachung des Klassifizierungssystems, insbesondere durch Beschränkung auf eine höhere Aggregationsstufe der Klassifikation. Dies erkauft man sich allerdings mit einer steigenden Inhomogenität bei den Mess-Subjekten, die einer der entsprechenden Klassen zugeordnet sind. Die Verwendung von Fuzzy-Klassifikationen würde hier die Möglichkeit anbieten, dass Ermessensentscheidungen bei den Klassifizierungsvorgängen ebenfalls abgebildet werden können, so dass — vermittelt über den Grad der Zugehörigkeit zu einer Klasse — eine größere Homogenität in den Klassen erreicht werden könnte. Dennoch wird der Aspekt von Klassifikationsfehlern im Weiteren nicht genauer betrachtet. Es wird auf die einschlägigen Arbeiten zur Fuzzy-Clusteranalyse verwiesen.[29]

Es ist hervorzuheben, dass das Konzept des Gesamtfehlers auf Aggregationsprozessen in zwei Dimensionen beruht. Zum einen wird der Gesamtfehler im Bezug auf eine Zusammenfassung von Einzelmessungen bewertet. Zum anderen ist über alle vorkommenden Fehlereinflüsse zu aggregieren. Insbesondere hinsichtlich des sog. „systematischen Fehlers“ ist auf eine Begriffsverwirrung hinzuweisen, denn von einigen Autor/innen werden auch nicht-zufällige Fehlereinflüsse als „systematische Fehler“ bezeichnet[30], im Zusam-

[26] Die Datenqualität kann u.a. nach der Quote der aufgrund der Plausibilitätsprüfung korrigierten Daten bewertet werden.

[27] Zufällige Datenausfälle führen u.a. auch zu einer Erhöhung des Stichprobenzufallsfehlers.

[28] Vgl. Neubauer 1993, S. 16f.

[29] So z.B. [Höppner u. a., 1997].

[30] So z.B. Krug u. a. 2001, S. 216f.

menhang mit dem Gesamtfehler ist als „systematischer Fehler“ aber immer der mittlere Fehler zu verstehen, der nicht auf den Stichprobenzufallsfehler zurückzuführen ist. In der Wirtschafts- und Sozialstatistik gelingt normalerweise weder eine quantitative Beschreibung aller Fehlerarten noch ist hinreichend genau bekannt, inwieweit sich die Einflüsse der einzelnen Fehlerarten bei der Aggregation im Gesamtfehler gegenseitig kompensieren oder verstärken. Um die Messfehler in den Daten annähernd abzuschätzen, sollte daher zunächst geprüft werden, welche Fehlerarten am stärksten zur Verzerrung der Daten beitragen, um die Betrachtung dann darauf zu beschränken. Die Untersuchung [Brinner, 2003] zu den Erhebungsungenauigkeiten beim Übergang von der DDR-Statistik zur gemeinsamen bundesdeutschen Statistik bei der amtlichen Mortalitätsmessung ist ein gutes Beispiel für dieses Vorgehen, stellt mit ihrer Ausführlichkeit aber eine positive Ausnahme dar.

Die Ausprägung des Gesamtfehlers kann danach charakterisiert werden, ob in den Merkmalswerten über eine Klasse von individuellen wirtschaftlichen Akteur/innen aggregiert wird oder nicht. Merkmalsbeobachtungen, die auf Individualebene vorliegen, werden als *Mikrodaten* bezeichnet. Aggregierte Daten, bei denen Teilpopulationen zusammengefasst betrachtet werden, werden *Makrodaten* genannt.

Ein Vorteil bei der Verwendung von Makrodaten besteht darin, dass die Messfehler aus den Einzelbeobachtungen sich zum Teil gegenseitig kompensieren. Grundsätzlich ist aber die Zusammenfassung und Verallgemeinerung der Beobachtungen in den Makrodaten nicht unproblematisch, wenn die Ausprägungen der Merkmalswerte inhomogen ausfallen und von Einzelfällen überformt werden. Insbesondere können Änderungen der Merkmalswerte, die durch eine veränderte Zusammensetzung innerhalb eines Aggregates entstehen, damit nicht analysiert werden. Häufig werden Makrodaten als Analogon für ein fiktives „repräsentatives Individuum“ interpretiert. Allerdings existiert nicht notwendig ein Mess-Subjekt mit den „repräsentativen“ Eigenschaften, so dass die Adäquation bei Makrodaten erschwert ist. Insgesamt gilt, dass Fehler in Makrodaten typischerweise systematische Fehler sind, die in vielen Fällen zusätzlich über die Zeit korreliert sind.[31]

Um die durch die Aggregation verursachten Adäquationsprobleme zu vermeiden, werden inzwischen verstärkt Mikrodaten für die Analyse herangezogen. Dadurch wird aber andererseits das Gewicht und die Auswirkungen von Messfehlern bei der Analyse verschärft und der Einfluss von fehlenden Daten steigt.[32] Zudem sind Inhomogenitäten unter den klassifizierten statistischen Einheiten von höherer Relevanz.[33]

[31] Vgl. Griliches 1986, S. 1476.

[32] Vgl. ebd. S. 1469.

[33] Vgl. Baltagi 1998, S. 105. Das Problem entsteht dann, wenn der Inhomogenität im Regressions- bzw. Schätzmodell nicht Rechnung getragen wird. Überdies ist anzumerken, dass Klassifizierungen bei Regressionsanalyse üblicherweise mittels (0, 1)-Variablen eingebunden werden, die aufgrund der Diskretisierung

Aus der Perspektive der Datenanalyse betrachtet, treten zu den Ungenauigkeiten, die bei Datengewinnung entstehen, noch weitere Ungenauigkeiten hinzu. Diese werden durch die Datentransformation oder durch Einschränkungen bei der Vergleichbarkeit sowie durch Unvereinbarkeiten zwischen den ökonometrischen Begriffsbildungen und den verfügbaren Daten induziert. Von besonderer Bedeutung sind aus dieser Perspektive die „*synthetisierten Daten*“[34], mit denen theoretische Begriffe approximiert werden, für die es keine direkte Entsprechung in der Realität gibt und die durch Verknüpfung von Daten anderer Merkmale konstruiert werden. Dies trifft auf viele wirtschaftliche Kennzahlen zu, wie z.B. den Preisindex für die private Lebenshaltung oder das reale Bruttosozialprodukt. Aber auch durch Fortschreibung mittels bekannter Bewegungsmassen bestimmte Bestandsmassen sowie Vergleichszahlen wie Relationsquotienten oder Differenzen werden durch Kombination von primären Daten berechnet. Fehler in den Eingangsdaten werden dabei entsprechend mittransformiert. Dies kann zur gegenseitigen Verstärkung oder zur gegenseitigen Kompensation der Fehlerbestandteile in den abgeleiteten Daten führen. Beispielsweise können konstante, additive Fehleranteile durch die Verwendung von Differenzen ausgeschaltet werden, wohingegen sich der Gesamtfehler bei Fortschreibungen über die Perioden kumuliert. Bei der Verwendung von synthetisierten Daten ist in besonderer Weise darauf zu achten, welche Theorieannahmen die Verknüpfungsoperationen motivieren. Überdies kann die Datensynthese Schwierigkeiten bei der empirischen Analyse bereiten, wenn durch die Transformation verdeckte, zusätzliche Abhängigkeiten zwischen einzelnen Merkmalsvariablen entstehen. Das reale Bruttosozialprodukt wird z.B. durch Division mit einem aggregierten Preisindex ermittelt. Es ist folglich nicht möglich, die Messfehler auf dem aggregierten Niveau als unabhängig von den im Preisindex enthaltenen Ungenauigkeiten bei der Messung der Preise und Mengenverhältnisse zu betrachten.[35]

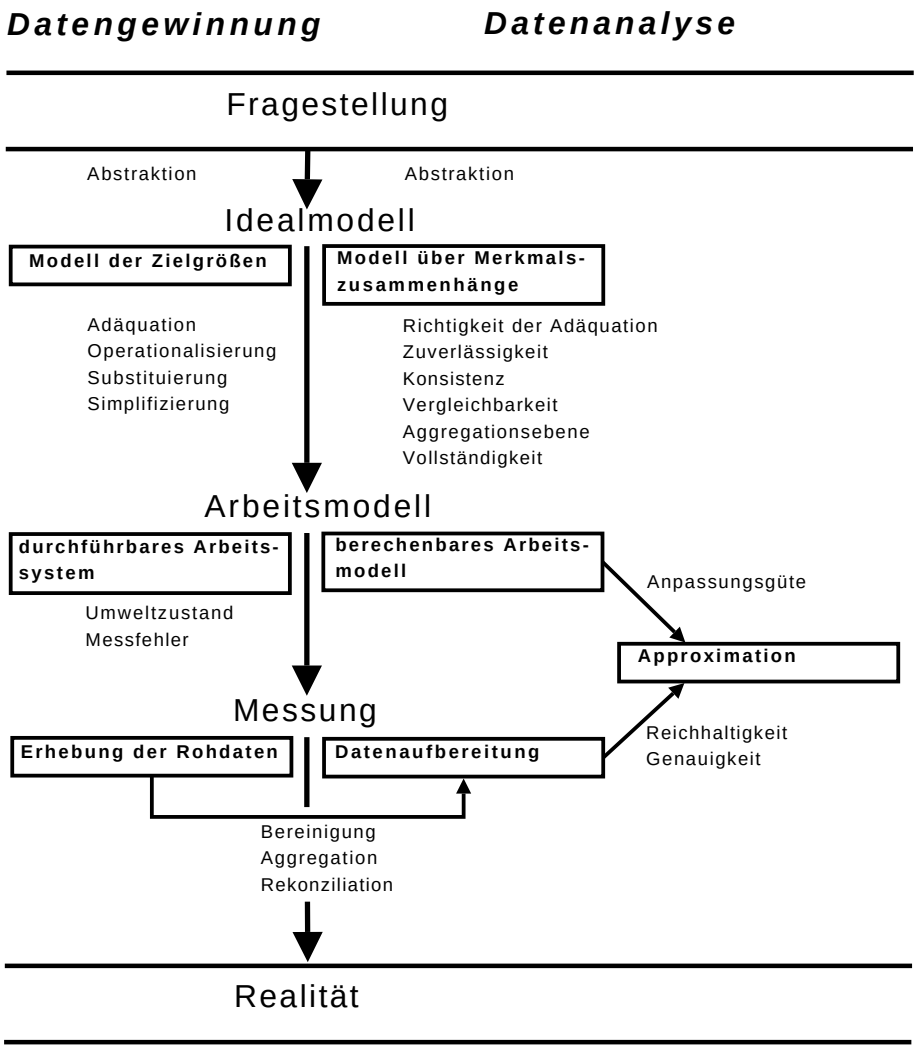
Zusammenfassend sind die ausgeführten Interpretationsebenen für die Bewertung von Datenunschärfen im Diagramm 2.1 gegenübergestellt. Da die Operationalisierungen im Messprozess und bei der Modellanpassung widersprüchliche Anforderungen an die statistischen Begriffe und die Aussagekraft der Daten stellen können, werden hierbei die Seite der *Datengewinnung* und der *Datenanalyse* gegenübergestellt. Als Stufen, die zwischen einem Erkenntnisinteresse bzw. einer Fragestellung und der Realität vermitteln sollen, können die Ebenen des Idealmodells, des Arbeitsmodells sowie der Messung betrachtet

---

des Lösungsraums zu Unstetigkeiten und Umschlagspunkten und damit zu einer größeren Sensitivität im Optimum führen können.

[34] Vgl. Neubauer 1993, S. 16f.

[35] Vgl. Griliches 1986, S. 1473: „The major serious problem (...) probably occurs in the measurement of ‚real‘ output, GNP or industry output in ‚constant prices‘ (...). Since most of the output measures are derived by deviding (‚deflating‘) current value totals by some price index, the quality of these measures is intimately connected to the quality of the available price data. Because of this, it is impossible to treat errors of measurement at the aggregate level as being independent across price and ‚quantity‘ measures“.



(Eigene Darstellung)[36]

Abbildung 2.1: Interpretationsebenen und Unschärfequellen bei der Datengewinnung und bei der Modellanpassung

werden. Zwischen den Abstraktionsebenen sind jeweils spezifische Qualitätsmerkmale zu verzeichnen, die die Unschärfe der Daten ausmachen und beschreiben.

Auslösend für die datengestützte Beobachtung und deren Analyse ist eine allgemeine Fragestellung bzw. ein Fragenkomplex. Bei einer Analyse der Beschäftigtenstruktur könnte beispielsweise die Ausgangsfrage formuliert werden: „Welches sind wichtige Einflussfaktoren auf die Entwicklung der Beschäftigung?“. Diese Fragestellung wird auf der Ebene eines *Idealmodells* zunächst in quantitative Zielgrößen X übertragen, die die hypothetischen Begriffe repräsentieren, bzw. auf der Seite der Datenanalyse in ein quantifizierbares Modell F übertragen, das funktionale Zusammenhänge zwischen Variablen von Analysegrößen formuliert. Das quantifizierbare Modell F schließt im Falle einer induktiven Analyse auch

[36] Die Modellierungszusammenhänge sind für die Seite der Datengewinnung in Anlehnung an Strecker und Wiegert 1994, S. 104f, und für die Seite der Datenanalyse in Anlehnung an Frohn 1993, S. 62, dargestellt.

ein Verteilungsmodell ein, das die Variabilitäten zwischen den Analysevariablen und dem Strukturmodell abbildet.[37]

Auf der Ebene des *Arbeitsmodells* werden die Zielmodelle dann so operationalisiert, dass sie für die Anwendung geeignet sind. Für die Datengewinnung werden die Zielgrößen dabei in ein Arbeitssystem umgesetzt, das einen durchführbaren Erhebungsplan darstellt.[38] Hierzu werden zusammen mit dem Erhebungsziel, der Adäquation der Zielgrößen und den operablen Merkmalen der definierten Einheiten auch die Erhebungstechnik einschließlich des Tabellenprogramms, der Erhebungsart, der Organisation der Feldarbeit sowie der Aufbereitung der erhobenen Daten definiert und festgelegt.[39] Durch das Arbeitssystem wird ein Vektor der „wahren" Merkmalswerte $\mathbf{X}_W$ repräsentiert. Analog wird bei der Datenanalyse ein Arbeitsmodell f spezifiziert, in dem die Qualität und der Aussagewert der erhältlichen, d.h. der vorliegenden oder der beobachtbaren, Daten berücksichtigt ist. Beim Übergang vom Idealmodell zum berechenbaren Arbeitsmodell müssen Analysevariable häufig durch ein oder mehrere Hilfsvariable substituiert werden, die operationalisierbar sind und bei denen eine zufriedenstellende Datenqualität erreicht werden kann. Zur Festlegung des berechenbaren Arbeitsmodell ist eine passende Methode zur Bestimmung der Relationen zwischen den Modellvariablen auszuwählen.

Die Ebene der *Messung* steht auf der Seite der Datengewinnung für die Realisation des Arbeitssystems, bei der ein Messwert $\mathbf{x}^*$ gewonnen wird. Das Ergebnis $\hat{\mathbf{x}}$ der Datenaufbereitung durch Datenbereinigung, Aggregation und ggf. auch Rekonziliation[40] wird hier bereits als Teil der Datenanalyse aufgefasst. Außerdem wird auch die Bildung von verknüpften Kennzahlen zu diesem Schritt gezählt. Die Aufbereitung der Daten mittels einer Fuzzy-Modellierung, die im Laufe dieser Arbeit diskutiert werden soll, ist somit ebenfalls auf dieser Stufe der Datenanalyse einzuordnen.

Bemerkung 2.1 (Genauigkeitsschwelle) *Durch die Bestimmung einer geeigneten Mess-Skala für die Fragestellung wird auch festgelegt, mit welcher Sensitivität die Messungen durchgeführt werden sollen. Also z.B. ob die Größe eines Objekts in der Einheit cm, m oder km erfasst werden soll. In Abhängigkeit davon verschiebt sich auch die Wahrneh-*

[37] Auch bei explorativen Studien wird häufig bereits eine Vorauswahl „relevanter" Zielgrößen durchgeführt. Ebenso werden die Parameter einer Approximations- bzw. Ausgleichsfunktion, die als Hinweis auf einen proportionalen Zusammenhang gewertet werden können, durch die Qualität und Repräsentativität der verwendeten Daten beeinflusst.

[38] Dies ist eine Verkürzung gegenüber der Darstellung in [Strecker und Wiegert, 1994]. Sie setzen eine *Variabilität des Arbeitssystems* und damit eine Menge von durchführbaren Erhebungsplänen voraus, von denen unter Kosten-Nutzen-Abwägungen einer ausgewählt wird bzw. sich im Zuge der Erhebung konstituiert.

[39] Vgl. Strecker und Wiegert 1994, S. 103.

[40] D.h. der Glättung von Daten, damit diese bestimmten Anforderungen genügen wie etwa der Identität gewisser Randsummen.

mungsgrenze für unwesentliche Veränderungen. So werden wir bei der Messung in km eine Abweichung von 1 m als zu vernachlässigen betrachten, bei einer Messung im m hingegen erst eine Abweichung im mm-Bereich. Zadeh bezeichnet dieses Phänomen als „granulation".[41] *Die Wahl des Maßstabes kann ein bewusst gewähltes Mittel zur Komplexitätsreduktion sein. Morgenstern [1965] verweist allerdings darauf, dass bei der Verwendung von sekundären Datenquellen, auf die man bei wirtschaftsstatistischen Analysen häufig angewiesen ist, besonders darauf zu achten ist, welche Genauigkeitsschwelle überhaupt für alle Merkmalsvariablen gewährleistet werden kann, um Scheingenauigkeiten zu vermeiden.*[42] *Grundsätzlich sind Messfehler daher erst dann relevant, wenn sie die modellimmanente Genauigkeitsschwelle überschreiten. Es soll allerdings nicht verschwiegen werden, dass diese Stabilität nur dann gilt, wenn keine komplexen Rückkopplungseffekte bestehen. Solche Effekte sind z.B. Gegenstand der Chaostheorie.*

Schließlich erhalten wir aus der Anpassung des Arbeitsmodells anhand der passend aufbereiteten Daten eine *Approximation* $\hat{f}(\hat{\mathbf{x}})$, die für die Interpretation von funktionalen Zusammenhängen in den Merkmalswerten zur Verfügung steht.

Auf der Seite der Datenanalyse werden wir uns im Folgenden auf die Regressionsanalyse konzentrieren. Diese wird im folgenden Abschnitt kurz vorgestellt, wobei vor allem die Ansätze zur Behandlung ungenauer Daten kritisch beleuchtet werden sollen.

2.2 Datenunschärfe bei einfacher, linearer Regression

Ökonometrische Analysen haben die Aufgabe, funktionale Beziehungen zwischen ökonomischen Variablen mit Hilfe von statistischen Methoden zu identifizieren und zu messen. Die von der Theorie vorgeschlagenen funktionalen Zusammenhänge sollen dabei mit Hilfe empirischer Daten statistisch geprüft werden. Dabei soll nicht nur das Vorzeichen des Zusammenhangs zwischen den ökonomischen Variablen bestätigt werden, sondern es sollen nach Möglichkeit auch dessen quantitative Intensitäten abgebildet werden. Da Beziehungen im sozialen Raum kontinuierlichen Änderungsprozessen unterliegen, ist das Ziel der Untersuchungen darauf ausgerichtet, relevante ökonomische Beziehungen zu ermitteln, die zumindest über einen gewissen Zeitraum stabil sind. Nicht zuletzt sollen aus den vorliegenden Daten auch quantitative Prognosen ermittelt werden, für die im Rahmen der Modellannahmen eine Abschätzung des Fehlerspielraums möglich ist.

Grundsätzlich stellt die Regressionsanalyse die zentrale Methode der Ökonometrie dar, mit der Modellparameter numerisch spezifiziert werden können. Typische Beispiele für

[41] Vgl. [Zadeh, 1997].

[42] Vgl. Morgenstern 1965, S. 95.

solche Parameter, die wesentliche Zusammenhänge zwischen den ökonomischen Variablen kennzeichnen, sind Grenzneigungen oder Elastizitäten.[43]

Am Anfang der Überlegungen steht die Formulierung des Schätzproblems. Darin werden die theoretische Vorüberlegungen in ein oder mehrere parametrisierte Funktionsgleichungen übersetzt, so dass die Parameter anhand von empirischen Daten ermittelt werden können. Im Schätzproblem ist außerdem zu unterscheiden, welche Variable für den funktionalen Zusammenhang als gegeben bzw. unabhängig erscheinen und welche Variable sich aus dem Zusammenhang ergeben und somit abhängige Variable sind. In den Begrifflichkeiten der Ökonometrie werden erstere als exogene Variable und zweitere als endogene Variable für den untersuchten Zusammenhang bezeichnet. Dabei wird angenommen, dass Abweichungen von der Modellfunktion dadurch entstehen, dass der funktionale Zusammenhang von zufälligen Störungen überlagert ist[44], die ebenfalls auf die abhängigen Variablen — und im einfachen Modellansatz nur auf diese — einwirken.

Die Modellbildung findet in der Ökonometrie meistens unter ungünstigen Bedingungen statt. Häufig ist der funktionale Zusammenhang im Zeitablauf nur beschränkt stabil. Insbesondere bei Mikrodaten können die Störeinflüsse auf den Zusammenhang relativ groß sein und stark variieren. Schließlich sind bei der allgemeinen Interdependenz ökonomischer Beziehungen Rückkopplungseffekte zwischen den beobachtbaren ökonomischen Variablen und den Störvariablen zu erwarten. Es ist daher sinnvoll, den ökonometrischen Zusammenhang nur annähernd abzubilden, so dass nur die wesentlichen Zusammenhangsparameter geschätzt werden können. Für die meisten Situationen ist eine Beschreibung mit Hilfe von linearen Modellfunktionen hinreichend genau. Beispielsweise können auch exponentielle Wachstumsentwicklungen durch Logarithmieren auf eine lineare Zusammenhangsgleichung transformiert werden. Die Kleinste Quadrate-Regression ist allerdings nur dann optimal, wenn die Störkomponente stochastisch unabhängig von den Variablen der Modellgleichung ist. Andernfalls kann die Verzerrung der Schätzparameter erheblich sein, wie anhand des einfachen klassischen Fehlermodells illustriert werden kann, das den Effekt von fehlerhaften Daten auf die Parameter der linearen Regression beschreibt.

Bei ökonometrischen Analysen kann keinesfalls davon ausgegangen werden, dass die Bedingungen für die Optimalität der Kleinste Quadrate-Regression hinreichend gut erfüllt sind. Es kommt häufig vor, dass die Störkomponente innere Abhängigkeiten (Autokorrela-

[43] Vgl. Schneeweiß 1990, S. 17.

[44] In der Literatur wird an dieser Stelle üblicherweise die Bezeichnung „stochastische Störung“ verwendet, um deutlich zu machen, dass die Störung mittels eines Verteilungsmodells beschrieben und durch eine Zufallsvariable abgebildet werden kann. Da Stochastik aus dem Altgriechischen auch als „zum Erraten gehörende Kunst“ übersetzt werden kann, gehört die Fuzzy-Theorie in einem weiteren Sinne auch zu den Gebieten der Stochastik. Vgl. Brockhaus, 21. völlig neu bearbeitete Auflage, Online-Ausgabe; Zugriff über Munzinger-Online am 25. November 2009. Aus diesem Grund wird hier immer der Begriff „zufällig“ verwendet, wenn eine Modellierung mit Hilfe von Zufallsvariablen möglich ist.

tionen)[45] aufweist oder durch Einflussfaktoren induziert ist, die nicht identifiziert werden können oder die nicht messbar sind. Eine zutreffende Messung von funktionalen Beziehungen ist daher nur dann möglich, wenn „das Zustandekommen der beobachteten Daten theoretisch verstanden und modellmäßig erfasst worden ist“[46]. Für das hinter den Daten stehende ökonometrische Modell reicht es folglich nicht aus, eine Funktion anzugeben, die gewisse Variable sinnvoll miteinander verknüpft, sondern es besteht die umfassendere Anforderung, dass auch die stochastische Natur dieser Variablen anzugeben ist, wobei die möglichen Interdependenzen der Störvariablen besonderes zu berücksichtigen sind. Darüber hinaus sind alle Variablen im Modell zu spezifizieren, mit denen möglicherweise Interdependenzen vorhanden sind.

Insgesamt bewegt man sich also bei der Spezifikation des Schätzmodells in dem Spannungsfeld, dass das Modell einerseits vollständig genug sein soll, so dass alle wesentlichen Effekte und Abhängigkeiten in die Schätzung einbezogen werden, andererseits soll die Zahl der Variablen nicht zu groß werden, um den approximativen Ansatz in der Modellbildung zu wahren und nicht einer Scheingenauigkeit zu unterliegen, die den Instabilitäten der ökonomischen Zusammenhänge nicht angemessen ist. Überdies können bei einer wachsenden Zahl von Variablen aussagekräftige Schätzergebnisse nur noch erreicht werden, wenn immer mehr Beobachtungen einbezogen werden. Außerdem steigt die Gefahr, dass die Parameter nicht mehr identifiziert werden können, d.h. dass das Schätzproblem keine oder keine eindeutige Lösung hat. Nicht-Identifizierbarkeit eines Strukturmodells kann aber grundsätzlich auch in der Natur der Fragestellung selbst liegen, wenn diese ungünstig gestellt ist. Von einer *Fehlspezifikation* des Schätzmodells sprechen wir immer dann, wenn wesentliche Einflussfaktoren oder Interdependenzen nicht modelliert wurden. Eine Schätzung nach einem fehlspezifizierten Modell führt definitionsgemäß zu verzerrten Parametern oder unzutreffenden Konfidenzmengen[47].

Die ökonometrische Modellkonstruktion basiert auf der Grundannahme, dass die Daten eindeutig und fehlerfrei beobachtet werden und somit auch die Merkmalsvariablen eindeutig und fehlerfrei sind. Fehler in den Variablen müssen demzufolge im Schätzmodell gesondert spezifiziert werden und die Verteilungsannahmen sind im Hinblick darauf zu modifizieren. Entsprechende Schätzmodelle werden häufig unter der Bezeichnung *Modelle mit Fehlern in den Variablen*[48] zusammengefasst. Als Ausgangspunkt für die weiteren Untersuchungen wird in Abschnitt *2.2.1* zunächst die „klassische“, lineare Regressions-

[45] Autokorrelationen sind häufig bei Zeitreihendaten festzustellen. Zu den unbekannten Einflussfaktoren sind u.a. auch Beobachtungsfehler in den Daten zu zählen.

[46] Vgl. Schneeweiß 1990, S. 22.

[47] Eine Verzerrung der Konfidenzmengen ist gleichbedeutend damit, dass der (Ko-)Varianzschätzer verzerrt ist. Dies ist z.B. bei Heteroskedastizität der Störkomponente, d.h. einer veränderlichen Streuung der Modellabweichungen zwischen den Stichproben, der Fall.

[48] Eine wichtige Publikation für den deutschsprachigen Raum ist z.B. [Schneeweiß und Mittag, 1986].

analyse eingeführt und die Kleinste Quadrate-Approximation als zentrale Schätzmethode vorgestellt. Im Abschnitt *2.2.2* werden dann einfache Modelle bei Fehlern in den Variablen als exemplarische Beispiele für die Verzerrungseffekte durch deterministische oder zufällige Fehler in den Daten diskutiert. Ein besonderer Augenmerk liegt dabei auf der Modellierung der Fehlereinflüsse.

2.2.1 Grundmodell der linearen Regression

Die einfache, lineare Regression mit einer abhängigen Variablen $Y : \Omega \longrightarrow \mathbb{R}$ und k unabhängigen Variablen $X_1, \ldots, X_k : \Omega \longrightarrow \mathbb{R}$ stellt den Prototyp und im weitesten Sinne auch das allgemeine Ideal der Regressionsanalyse dar. Für die Aufgabenstellung der linearen Regression wird ein Zusammenhang zwischen k unabhängigen Variablen $X_1, \ldots, X_k$ und einer abhängigen Variablen Y unterstellt, der linear in den k Parametern $a_0, a_1, \ldots, a_k \in \mathbb{R}$ ist.[49] Dabei wird angenommen, dass die Abweichung zwischen den Beobachtungswerten und der Modellgleichung durch eine Zufallsvariable ε modelliert werden kann. Im Mittelpunkt steht somit die folgende allgemeine Modellgleichung

$$Y = a_0 + \sum_{j=1}^{k} a_j X_j + \varepsilon. \tag{2.2}$$

Im Folgenden benötigen wir eine verkürzte Notation der Schar der Modellfunktionen. Wir setzen $f(\mathbf{z}, \mathbf{a}) = a_0 + \sum_j a_j z_j$, wobei $\mathbf{z} = (z_1, \ldots, z_k), \mathbf{a} = (a_0, a_1, \ldots, a_k) \in \mathbb{R}^{k+1}$. Hierbei soll der Buchstabe ‚z' auf eine unbestimmte Variable verweisen, damit in der weiteren Diskussion eine Verwechslung mit den Variablen bzw- den Beobachtungswerten für die abhängige und die unabhängigen Variablen vermieden wird.

Die Parameter der Modellgleichung sollen nun aufgrund von empirischen Daten spezifiziert werden. Dazu muss der Zustand der Modellvariablen für den Zusammenhang wiederholt gemessen werden. Es sollen daher n Beobachtungen zur Beschreibung des Zusammenhangs vorliegen. Im Schätzmodell wird angenommen, dass die gemessenen Daten als Ergebnis von Zufallsprozessen zustande kommen und durch Zufallsvariablen beschrieben werden können. Dabei werden für eine beliebige Modellvariable $Z \in \{Y, X_1, \ldots, X_k\}$ die Einzelmessungen unterschieden, so dass im Extremfall jede Einzelmessung einer anderen Wahrscheinlichkeitsverteilung folgen kann. Die Grundannahmen an den zufälligen Messprozess werden folgendermaßen präzisiert.

[49] D.h. es ist zulässig, dass die unabhängigen Variablen eine nicht-lineare Funktion wiedergeben, z.B. $X_1 = x, X_2 = x^2, X_3 = x^3$.

Definition 2.2 Sei $n \in \mathbb{N}$ die Zahl der durchgeführten Einzelmessungen für die Modellvariable Z. Für jedes $1 \leq i \leq n$ ist die Zufallsvariable $Z_i : \Omega \longrightarrow \mathbb{R}$ eine *Beobachtung*. Die *Stichprobe* wird dann vom n-Tupel $(Z_1, \ldots, Z_n)$ der Beobachtungen gebildet.

Für ein $\omega \in \Omega$ ist $z_i = Z_i(\omega)$ der *Beobachtungs- bzw. Merkmalswert* und das n-Tupel $(z_1, \ldots, z_n) = (Z_1(\omega), \ldots, Z_n(\omega))$ ist das *Stichprobenergebnis*.

Die Begriffe werden gleichfalls auf die Familie der Modellvariablen $(X_1, \ldots, X_k, Y)$ übertragen. Wir sprechen dann von einer Beobachtung der Modellvariablen, von einer Stichprobe der Modellvariablen und so fort. Die Bedeutung ist im Folgenden immer aus dem Kontext erkennbar.

In Anlehnung an die Begrifflichkeiten der ökonomischen Theorie werden im Folgenden die Beobachtungswerte für die abhängige Variable Y auch als *Outputs* und die Beobachtungswerte für die unabhängigen Variablen $X_1, \ldots, X_k$ als *Inputs* bezeichnet.

Wenn der funktionale Zusammenhang aufgrund der n Beobachtungswerte ermittelt werden soll, ist zu berücksichtigen, ob sich Veränderungen in den Modellannahmen zwischen den Beobachtungen ergeben und ob zwischen den n Beobachtungen Abhängigkeiten vorliegen. Das Schätzmodell der einfachen, linearen Regression ist dadurch gekennzeichnet, dass die Rahmenbedingungen sich zwischen den Beobachtungen nicht wesentlich ändern und dass keine Abhängigkeiten bei den Störvariablen bestehen. Im Modell wird dazu angenommen, dass die Störvariablen ε_i paarweise unkorreliert sind und dass alle Störvariablen dieselbe Streuung haben. Außerdem sollen die Störvariablen im Mittel verschwinden. Da der Zustand der unabhängigen Variablen, anders als bei experimentellen Situationen, normalerweise nicht vorgegeben werden kann, werden die unabhängigen Variablen ebenfalls als Zufallsvariable betrachtet.[50] Da nur der Zusammenhang zwischen den unabhängigen und der abhängigen Variablen geschätzt werden soll, kommt es auf die Verteilung der unabhängigen Variablen $X_{11}, \ldots, X_{nk}$ nicht an, sondern vielmehr auf die Verteilung der $Y_1, \ldots, Y_n$ unter gegebenen $X_{11}, \ldots, X_{nk}$. Die Modellgleichung ist dann als bedingte Modellgleichung zu sehen und die Verteilungsannahmen gelten ebenfalls bedingt unter gegebenem Stichprobenergebnis für die unabhängigen Variablen.

Das Schätzmodell der einfachen, linearen Regression[51] kann nun unter Einbeziehung der n Beobachtungen spezifiziert werden. Dabei werden die Zusammenhangsgleichungen

[50] Zu den steuerbaren Zuständen können allerdings auch die Zeitperioden bei Zeitreihen betrachtet werden, wenn ein Trend untersucht werden soll. Vgl. Schneeweiß 1990, S. 30.

[51] Hier steht „einfache, lineare Regression“ für das Modell der (engl.) „*ordinary least squares regression*“ Im Unterschied dazu kann im Schätzmodell der verallgemeinerten, linearen Regression bzw. (engl.) „*general least squares regression*“ nicht angenommen werden, dass die Varianzen für alle Beobachtungen gleich sind. In diesem Fall muss für eine unverzerrte Schätzung der Parameterschätzer mittels der Kovarianzmatrix transformiert werden. Dazu ist aber vorauszusetzen, dass die Kovarianzmatrix auch geschätzt werden kann.

für die Beobachtungen in Form einer Matrixgleichung zusammengefasst.

$$\mathbf{Y} = \mathbf{X}\mathbf{a} + \boldsymbol{\varepsilon}, \tag{2.3}$$

wobei

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{nk} \end{pmatrix}, \quad \mathbf{a} = \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_k \end{pmatrix} \quad \text{und} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Außerdem soll die Verteilung der Störvariablen $\boldsymbol{\varepsilon}$ die *Gauß-Markov-Bedingungen* erfüllen. Diese stellen die folgenden Anforderungen:

(GM.1) $\mathrm{E}(\boldsymbol{\varepsilon} \mid \mathbf{X}) = \mathbf{0}$, d.h. bei gegebenen Stichprobenwerten für die unabhängigen Variablen ist der Erwartungswert aller Störvariablen 0. Da die bedingte Erwartung konstant ist, folgt daraus insbesondere, dass $\varepsilon_1, \ldots, \varepsilon_n$ stochastisch unabhängig von $X_{11}, \ldots, X_{nk}$ ist.

(GM.2) $\mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T \mid \mathbf{X}) = \sigma^2\mathbf{I}$, d.h. bei gegebenen Stichprobenwerten für die unabhängigen Variablen gilt für alle $1 \leq i \leq n$, dass $\operatorname{Var}\varepsilon_i = \sigma^2 \in (0, \infty]$ und die Störgrößen sind paarweise unkorreliert für $h \neq i$.

(GM.3) Die Matrix $(\mathbf{X}^0)^T(\mathbf{X}^0)$ ist fast-sicher regulär. Dabei ist

$$\mathbf{X}^0 = \begin{pmatrix} X_{11} & \cdots & X_{1k} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{nk} \end{pmatrix}$$

die Stichprobenmatrix mit den Beobachtungswerten für die unabhängigen Variablen. Die Variation der unabhängigen Variablen ist also in dem Maße unabhängig, dass in der Stichprobe keine relevante Abhängigkeit zwischen ihnen besteht.

Die Unkorreliertheit von $\mathbf{X}, \boldsymbol{\varepsilon}$ in (GM.1) stellt sicher, dass der Einfluss der gemeinsamen Verteilung der $X_{11}, \ldots, X_{nk}$ gering ist und somit vernachlässigt werden kann.[52]

Der Kleinste Quadrate-Schätzer für die einfache, lineare Regression ist dadurch definiert, dass er die Summe der Abstandsquadrate zwischen den Beobachtungswerte für Y und der Modellfunktion $f(\cdot, \mathbf{a})$ in den Beobachtungswerten für $X_1, \ldots, X_k$ minimiert.

[52] Vgl. Schneeweiß 1990, S. 39f.

Definition 2.3 (Klassischer Kleinste Quadrate-Schätzer) Es sei das Schätzmodell (2.3) mit den Verteilungsbedingungen (GM.1), (GM.2) und (GM.3) gegeben.

Für das Schätzmodell liege nun ein Stichprobenergebnis $\big((x_{i1},\dots,x_{ik},y_i)_i; 1 \leq i \leq n\big)$ vor. Ein Parameter $\hat{\mathbf{a}} \in \mathbb{R}^k$, der das Funktional

$$D^2(\mathbf{a}) = \sum_{i=1}^{n}(y_i - f(\mathbf{x}_i,\mathbf{a}))^2, \tag{2.4}$$

d.h. die Summe der Abstandsquadrate bezüglich der Modellgleichung (2.2), minimiert, wird *Kleinste Quadrate-Schätzer* für das Schätzmodell der einfachen, linearen Regression genannt.

Die Differenzen

$$e_i(\mathbf{a}) = y_i - f(\mathbf{x}_i,\mathbf{a})$$

zwischen den Outputs und den Funktionswerten in den Inputs werden als *Residuen* bezeichnet. Dabei werden die Funktionswerte $\hat{y}_i = f(\mathbf{x}_i,\mathbf{a})$ als Schätzwerte für die Outputs betrachtet. Es gilt $D^2(\mathbf{a}) = \sum_i e_i^2(\mathbf{a})$. Die Residuen im Kleinste Quadrate-Schätzer werden vereinfachend auch als $e_i = e_i(\hat{\mathbf{a}})$ notiert.

Der Kleinste Quadrate-Schätzer stellt eine Orthogonalprojektion des Vektors der Stichprobenwerte für Y auf den Unterraum von $\mathbb{R}^k$ dar, der durch die Vektoren der Stichprobenwerte für $X_1,\dots,X_k$ aufgespannt wird.[53] Er kann nach der folgenden Formel berechnet werden

$$\hat{\boldsymbol{a}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}, \tag{2.5}$$

wobei $\mathbf{y} = (y_1,\dots,y_n)^T \in \mathbb{R}^k$ der Vektor der Stichprobenwerte für Y ist. Die Inverse von $\mathbf{X}^T\mathbf{X}$ existiert gemäß Bedingung (GM.3). Diese stellt außerdem sicher, dass der Kleinste Quadrate-Schätzer eindeutig ist.[54]

Es kann gezeigt werden, dass der Kleinste Quadrate-Schätzer im Schätzmodell der einfachen, linearen Regression ein Optimum darstellt. Dabei kann die Güte unter verschiedenen Perspektiven nachgewiesen werden. Diese Güteeigenschaften gelten später bei der Fuzzy Kleinste Quadrate-Regression häufig nur noch unter Einschränkungen. Im Hinblick auf die

[53] Als Orthogonalprojektion hat $\hat{\mathbf{a}}$ die günstige Eigenschaft, dass die Residuen sich in der Summe gerade ausgleichen, es gilt $\sum_i e_i = 0$. Außerdem sind die Residuen für jede exogene Variable X_j orthogonal zum Vektor der entsprechenden Stichprobenwerte $\mathbf{x}^{(j)} = (x_{ij})_i$, so dass $\mathbf{e}^T\mathbf{x}^{(j)} = 0$. Vgl. Schneeweiß 1990, S.95f.

[54] Tatsächlich kann der Kleinste Quadrate-Schätzer allgemeiner mit der Moore-Penrose-Inversen $\mathbf{X}^T\mathbf{X}$ berechnet werden, sofern die verallgemeinerte Inverse existiert. Der Kleinste Quadrate-Schätzer ist dann aber im allgemeinen nicht mehr eindeutig.

Untersuchung der Fuzzy-Regression bei einer Fuzzy-Modellierung von fehlerhaften Daten wird daher bei der Erörterung der Güteeigenschaften des Kleinste Quadrate-Schätzers eine Vorauswahl getroffen. Der Hauptaugenmerk liegt bei den Eigenschaften, die durch Datenfehler beeinflusst werden sowie bei den Eigenschaften, die ebenfalls auf einige der untersuchten Fuzzy-Methoden übertragen werden können. Da wir davon ausgehen müssen, dass die beobachtete Situation das Schätzmodell nur annähernd erfüllt, werden auch Aussagen über die Konvergenz der Schätzung benötigt. Deshalb werden lokale und asymptotische Güteeigenschaften unterschieden.

Als *lokale Güteeigenschaften* sollen solche Eigenschaften des Kleinste Quadrate-Schätzers gelten, die für ein und dieselbe Stichprobe nachgewiesen werden können. Eine wesentliche lokale Eigenschaft ist die Approximationseigenschaft. Denn die Aufgabenstellung der linearen Regression kann zunächst auf die Lösung des Kleinste Quadrate-Problems gemäß Gleichung (2.4) aus Definition 2.3 reduziert werden. Dies stellt eine Beschränkung auf das reine Approximationsproblem dar, bei dem eine lineare *Ausgleichsgerade* zu einer gegebenen Menge von Datenpunkten $((\mathbf{x}_i, y_i); 1 \leq i \leq n)$ zu bestimmen ist. Mit der Summe der Fehlerquadrate $D^2(\hat{\mathbf{a}}) = \sum_i |e_i|^2$ liegt darüber hinaus auch eine *Gütebewertung für die Approximation* vor, die beschreibt, wie gut die Outputs durch die Ausgleichsfunktion angenähert werden. Die Kleinste Quadrate-Ausgleichsgerade kann als eine Art Mittelwert über die möglichen Geraden betrachtet werden, die durch die Datenpunkte aufgespannt werden. Dies lässt sich damit motivieren, dass die Lösung des Optimierungsproblems $\min\{a \in \mathbb{R} \mid \sum_i (x_i - a)^2\}$ gerade durch den Mittelwert $\overline{x}_\cdot$ über die x_i gegeben ist.

Aus einer allgemeinen Warte betrachtet ist eine Ausgleichsfunktion eine Funktion aus einer festgelegten Menge von interessierenden Funktionen, die die gegebenen Daten bestmöglich annähert. Die bestmögliche Annäherung wird im allgemeinen durch Minimieren einer Distanzfunktion bestimmt. Im vorliegenden Fall der klassischen Kleinste Quadrate-Approximation ist die Distanzfunktion durch das Quadrat der euklidischen Norm $\|\cdot\|^2$ bestimmt. Mit Hilfe von Ausgleichsfunktionen kann untersucht werden, welche der vorgegebenen Funktionen die Punkte einer Datenwolke gut charakterisiert. Dabei werden keine weiteren Annahmen bzgl. der Struktur der Daten aufgestellt. Mit einer Ausgleichsfunktion können daher nur Zusammenhänge innerhalb der aktuell vorliegenden Daten identifiziert werden. Diese können als Hinweise für quantitative Strukturen in den Daten aufgefasst werden, sind aber nicht dazu geeignet, einen funktionalen Zusammenhang zu belegen oder gar zu überprüfen. Insbesondere können bei einer Ausgleichsgeraden keine Aussagen darüber getroffen werden, inwieweit die Kleinste Quadrate-Parameter den „wahren“ Zusammenhang beschreiben. Solche Analysen und die davon abzuleitenden Ergebnisse sind nur im Rahmen eines Schätzmodells möglich.[55] Es ist daher notwendig die unterschied-

[55] In einem Schätzmodell kann z.B. der „wahre“ Wert für einen Beobachtungswert modelliert werden,

lichen Ansätze genau zu unterscheiden. Für die Kleinste Quadrate-Ausgleichsgerade soll im Folgenden die Bezeichnung *Kleinste Quadrate-Approximation* und für die zugehörigen Parameter die Bezeichnung *Approximationsparameter* reserviert sein. Im Unterschied dazu liegt eine Kleinste Quadrate-Schätzung nur dann vor, wenn diese im Bezug auf ein Schätzmodell wie hier dem Schätzmodell der einfachen, linearen Regression (2.3) mit den Verteilungsbedingungen (GM.1), (GM.2) und (GM.3) bestimmt wurde. Nur dann werden wir von *Parameter-Schätzern* sprechen.

Die Verteilungseigenschaften des Schätzmodells stellen sicher, dass der Kleinste Quadrate-Schätzer in der Nähe der erwarteten Parameter liegt. Gemäß des Satzes von Gauß-Markov gilt für den Kleinste Quadrate-Schätzer unter den Gauß-Markov-Bedingungen die Eigenschaft, dass er BLUE ist, d.h. er ist erwartungstreu und unter allen linearen Schätzern im Schätzmodell derjenige mit der kleinsten Varianz (engl. „*best linear unbiased estimator*“).[56]

Ganz allgemein kann hinsichtlich der Empfindlichkeit der Parameter-Schätzer auf Änderungen in den Daten festgehalten werden, dass die Parameterschätzung tendenziell besser ausfällt, wenn die Stichprobenwerte in $X_1, \ldots, X_k$ stärker streuen. Hingegen steigt der Anpassungsfehler für den Achsenabschnitt a_0, je weiter die X-Werte vom Ursprung entfernt sind. Da in der ökonometrischen Praxis die X-Werte oft weit vom Nullpunkt entfernt sind, ist die Schätzung des absoluten Glieds meist sehr unzuverlässig.[57]

Weitere lokale Eigenschaften des Kleinste Quadrate-Schätzers liegen nur dann vor, wenn die Störvariablen normalverteilt sind, d.h. wenn gilt

$$\forall\ 1 \leq i \leq n: \quad \varepsilon_i \sim N(0, \sigma^2) \tag{2.6}$$

Unter der Normalverteilungsannahme gilt, dass die Kleinste Quadrate-Parameterschätzer außerdem mit den Maximum Likelihood-Schätzern identisch sind. Daraus kann insbesondere gefolgert werden, dass die Parameter-Schätzer unter dieser Voraussetzung sogar effizient sind, d.h. erwartungstreu und unter allen erwartungstreuen Schätzern haben sie kleinste Varianz.[58]

Bei gegebener Stichprobenmatrix $\mathbf{X}$ sind die Parameter-Schätzer, wenn wir sie als Funktionen von Zufallsvariablen betrachten, unter der Normalverteilungsannahme ebenfalls normalverteilt. Dies gilt ebenfalls für jede Linearkombination der Parameterschätzer.[59] Als wichtigste Folgerung daraus lässt sich die Varianz der Störvariablen aus den Da-

indem er mit dem Erwartungswert der Verteilung der Beobachtung identifiziert wird.

[56] Vgl. Schneeweiß 1990, S. 109.

[57] Vgl. Schneeweiß 1990, S. 60f.

[58] Vgl. ebd. S. 72.

[59] Vgl. Schneeweiß 1990, S. 111.

ten schätzen, indem die Verteilung der Teststatistik in Gestalt der F-Verteilung analytisch bestimmt und damit Konfidenzbereiche für die Schätzwerte berechnet werden. Auf dieser Grundlage können dann schließlich auch Tests für lineare Hypothesen wie z.B. $a_1 = \cdots = a_k = 0$ konstruiert werden.[60]

Die *asymptotischen Güteeigenschaften* beziehen sich auf das Konvergenzverhalten des Kleinste Quadrate-Schätzers bei Vergrößerung der Stichprobe, d.h. für $n \to \infty$. Eine wesentliche Forderung ist der Nachweis der Konsistenz des Parameter-Schätzers. Diese liegt dann vor, wenn der Parameter-Schätzer bei wachsender Stichprobengröße stochastisch gegen den zu schätzenden Parameter konvergiert, d.h. wenn gilt $\text{P-lim}_{n\to\infty} \hat{\mathbf{a}} = \mathbf{a}$. Durch Vergrößerung der Stichprobe kann also im Schätzmodell der zugrundeliegende Parameter beliebig nah approximiert werden. Der Kleinste Quadrate-Schätzer ist konsistent, wenn die Folge der unabhängigen Variablen einer Beschränktheitsbedingung genügen und wenn hinreichend viele Beobachtungen vom Stichprobenmittel abweichen.[61]

Meistens erfüllen die Beobachtungen bzw. die Störvariablen ε_i in den Modellgleichungen des Schätzmodells in der Realität die Normalverteilungsannahme (2.6) nicht, sondern weichen mehr oder weniger davon ab. Die Normalverteilung der Störvariablen kann dann aber noch als asymptotische Eigenschaft sichergestellt werden, sofern die ε_i die Bedingungen des zentralen Grenzwertsatzes erfüllen. Demnach stellt die Normalverteilung eine gute Approximation für die gemeinsame Verteilung der Summe einer Vielzahl von Zufallsvariablen dar, die je für sich unbedeutend und die weitgehend voneinander unabhängig sind.[62] Eine typische Anforderung ist z.B., dass die Störvariablen stochastisch unabhängig und identisch verteilt sind. Wenn der zentrale Grenzwertsatz für die ε_i gilt, dann hat das zur Folge, dass auch die Parameter-Schätzer asymptotisch normalverteilt sind. Wenn die Störeffekte unabhängig und eher klein oder überwiegend symmetrisch sind, dann kann häufig bereits bei relativ kleinen Stichproben — schon ab mehr als fünf Beobachtungen — davon ausgegangen werden, dass die Normalverteilungsannahme hinreichend gut erfüllt ist.[63] Die Konvergenz verschlechtert sich, je schiefer die Einzelverteilungen sind. Die Bedingungen des zentralen Grenzwertsatzes sind klar verletzt, wenn die Verteilungen „dicke Schwänze“ (engl. „fat tails“) haben, d.h. wenn beliebig große Ausschläge mit einer gewissen Häufigkeit vorkommen. Unter den Bedingungen des zentralen Grenzwertsatzes gilt, dass der Kleinste Quadrate-Schätzer asymptotisch gleich dem Maximum Likelihood-Schätzer und somit auch asymptotisch effizient ist. Mit Hilfe der asymptotischen Konvergenz kann auch die Approximation von Konfidenzbereichen und die Konstruktion von Tests durchgeführt wer-

[60] Allgemein können Test für Hypothesen $\mathbf{K}\boldsymbol{\nu} = \boldsymbol{\gamma}$ mit $\mathbf{K} \in \mathbb{R}^{l\times(k+1)}, \boldsymbol{\gamma} \in \mathbb{R}^l$ konstruiert werden.

[61] Schneeweiß [1990] führt beispielsweise den Nachweis unter der Bedingung, dass für die *Momentenmatrix* $M^{(n)} = \frac{1}{T}\mathbf{X}^T\mathbf{X}$ gilt $\lim_{n\to\infty} M^{(n)}$ existiert und fast sicher regulär ist (S. 92f).

[62] Vgl. Schneeweiß 1990, S. 40.

[63] Vgl. ebd. S. 62. Der Hinweis bezieht sich dort vor allem auf die lineare Regression über Zeitreihen.

den. Da zur Bestimmung der Konfidenzbereiche aber auch der Streuungsschätzer benötigt wird, verlangsamt sich bei den beiden letzteren Anwendungen allerdings die Konvergenzgeschwindigkeit. Unter günstigen Bedingungen kann ab mehr als 30 Beobachtungen von einer hinreichenden Konvergenz ausgegangen werden.[64]

Im Hinblick auf die Modellierung von Fehlern in den Daten stellt das Ausreißerproblem ein interessantes Beispiel für die Verletzung der Annahmen im Schätzmodell der einfachen, linearen Regression dar. Als *Ausreißer* werden Beobachtungswerte bezeichnet, die nicht in eine erwartete Messreihe passen oder allgemein nicht den Erwartungen entsprechen. Da die Abstände zu allen Stichprobenwerten mit dem gleichen Gewicht in die Kleinste Quadrate-Schätzung einfließen, reagiert diese sensitiv auf einzelne Punkte, die abseits von den anderen Beobachtungswerten liegen.[65] Ausreißer können daher einen starken Effekt auf die Qualität der Kleinste Quadrate-Schätzung haben. Der gleiche Effekt ist vor dem Hintergrund des Schätzmodells unterschiedlich zu beurteilen, je nachdem worin die Ursache für den Ausreißer besteht. Ausreißerwerte sind auch im Rahmen des Schätzmodells möglich, wenn die Stichprobenwerte seltene Werte der Verteilung annehmen oder wenn die Stichprobe im Verhältnis zur Streuung der Beobachtungen zu klein ist oder insgesamt die vorliegende Stichprobe besonders ungünstig ausgefallen ist. Andererseits können Ausreißer auch darauf zurückgehen, dass die statistischen Begriffe zu weit gefasst sind und daher starke Inhomogenitäten vorkommen, oder dass die Beobachtungswerte sehr ungenau sind. Schließlich können Ausreißer dadurch verursacht sein, dass die Normalverteilungsannahme sich auch approximativ nicht halten lässt oder dass die Umweltbedingungen sich im Laufe der Messreihe erheblich geändert haben. Ausreißerwerte werden uns im Folgenden unter anderem noch unter dem Stichwort von singulären Fehlern in den Daten beschäftigen. Das Hauptinteresse wird aber eher bei verschmutzten Daten liegen, bei denen von einer Ungenauigkeit jedes Beobachtungswertes auszugehen ist.

Zusammenfassend ist festzuhalten, dass die Kleinste Quadrate-Schätzung auf des Basis des einfachen, linearen Schätzmodells sich trotz der aufgeführten Schwächen in realen Situationen häufig als recht stabil bei Abweichungen von den Grundannahmen erweist. Die Vorzüge der Kleinste Quadrate-Regression speisen sich aber überwiegend daraus, dass die Normalverteilungsannahme zumindest asymptotisch gewährleistet ist. Wenn die Normalverteilungsannahme nicht aufrecht erhalten werden kann, können keine Konfidenzbereiche mehr bestimmt werden, so dass ein zuverlässiges Testen nicht mehr zu begründen ist.[66] In diesem Fall verbleiben im wesentlichen die lokalen Eigenschaften, d.h. die BLUE-

[64] Vgl. Schneeweiß 1990, S. 66f.

[65] Im ungünstigsten Fall kann ein einzelner, separat liegender Punkt sogar eine Drehung der Regressionsgeraden um 90 ° verursachen. Vgl. das Beispiel in Sen und Srivastava 1990, S. 12.

[66] Es kann daher sinnvoll sein, zu robusten Schätzverfahren überzugehen, die zwar bei normalverteilten ε_i nicht optimal sind, die aber bei Abweichungen von der Normalverteilung oft genauere Schätzungen liefern

Eigenschaften und die Bestimmtheit bzw. die Approximationsgüte, von den asymptotischen Eigenschaften verbleibt zunächst die Konsistenz. Da insbesondere gerade dann, wenn die Datenqualität schlecht ist, die Stichprobe im Bezug auf die Zahl der erforderlichen Variablen im Schätzmodell klein ist, hat der Nachweis lokaler Eigenschaften grundsätzlich Priorität, weil diese auch bei kleinen Stichproben gelten.

2.2.2 Auswirkungen von Fehlern in den Daten

Anhand des klassischen Fehlermodells für die lineare Regression kann einerseits gezeigt werden, wie Fehler in den Daten üblicherweise in das Schätzmodell integriert werden. Andererseits kann damit illustriert werden, welche Effekte additive Fehler in den unabhängigen Variablen auf den Kleinste Quadrate-Schätzer haben. Insgesamt gibt dies einige Hinweise darauf, welchen Einfluss Datenunschärfen auf die Aussagefähigkeit der linearen Regression haben.

Die klassische Fehlermodellierung basiert auf der Zerlegung in eine Zufallskomponente und eine mittlere deterministische Komponente, aus denen der mittlere quadratische Gesamtfehler MSE bei der Datenmessung ermittelt wird.[67] Fehler in den Daten liegen auf Modellebene dann vor, wenn alle Beobachtungen in der Stichprobe von den Fehlereinflüssen gestört sind. Singuläre Fehlereffekte sind nach Möglichkeit vorab bei der Bereinigung der Rohdaten durch entsprechende Transformationen oder durch Abschneiden von Ausreißerwerten auszuschalten.

Es wird zunächst der Fall betrachtet, dass der mittlere deterministische Fehler eine relevante Größenordnung erreicht und dass keine Zufallsfehler zu berücksichtigen sind. Per definitionem ist der mittlere deterministische Fehler höchstens eine Funktion des „wahren“ Beobachtungswertes. Der mittlere deterministische Fehler soll daher durch eine lineare Funktion approximiert werden.[68] Die fehlerhafte Beobachtung X_{ij} für ein Merkmal j ergibt sich damit als

$$X_{ij} = \nu_j + \varphi_j X^*_{ij}, \tag{2.7}$$

wobei $\nu_j, \varphi_j \in \mathbb{R}$ der absolute bzw. der relative Fehler und X^*_{ij} die Zufallsvariable der „wahren“ Beobachtung ist, die hinter X_{ij} verborgen ist. Was passiert nun, wenn weiterhin das einfache, lineare Schätzmodell ohne Fehler in den Variablen unterstellt und die klassische Kleinste Quadrate-Schätzung durchgeführt wird? Die Transformation der Parameter im Schätzmodell aufgrund der Verwendung der fehlerhaften Beobachtungen erhalten wir,

als die Kleinste Quadrate-Methode. Vgl. Schneeweiß 1990, S. 109.

[67] Vgl. Formel (2.1) auf Seite 20.

[68] Vgl. Schneeweiß 1990, S. 216ff.

wenn wir die Fehlergleichung (2.7) nach X_{ij}^* auflösen und in das ursprüngliche Schätzmodell einsetzen. Damit erhalten wir die verfälschten Parameter

$$a_0' = a_0 - \frac{a_j}{\varphi_j}\nu_j \quad \text{und} \quad a_j' = \frac{1}{\varphi_j}a_j$$

Falls die Daten zur abhängigen Variablen fehlerhaft gemessen werden, also falls $Y_i = \zeta + \xi Y_i^*$, dann werden die verfälschten Parameter

$$a_0' = \xi a_0 + \zeta \quad \text{und} \quad a_j' = \xi a_j \; (1 \leq j \leq k)$$

induziert. Außerdem wird bei fehlerhafter abhängiger Variable die Störvariable und insbesondere auch deren Streuung um den Proportionalitätsfaktor ξ bzw. ξ^2 transformiert. In beiden Fällen überträgt sich die lineare Transformation der Parameterschätzer analog auf die jeweiligen Streuungsschätzer der Koeffizienten. Die Strukturbedingungen des Schätzmodells werden also infolge von deterministischen Fehlern nicht verändert.

Da der Achsenabschnitt a_0 in der Regel sehr stark streut und daher nicht für die Interpretation herangezogen wird, sind insgesamt vor allem relative deterministische Fehler relevant, die außerhalb einer Umgebung von 1 liegen. Dann erhalten wir gegenüber dem „wahren“ Parameter verzerrte Parameter, die auch asymptotisch verzerrt sind. Die Schärfe von Tests auf der Basis der üblichen Streuungsschätzer wird hingegen nicht berührt, da die lineare Transformation ohne Verzerrungen übertragen wird. Die gesamte Herleitung basiert allerdings wesentlich darauf, dass Veränderungen der Fehler zwischen den Beobachtungen einer Merkmalsvariable vernachlässigt werden können.

Mit schwerwiegenderen Folgen ist zu rechnen, wenn davon auszugehen ist, dass die Fehler in den Daten von Beobachtung zu Beobachtung variieren, denn das zieht regelmäßig eine Verzerrung der Kleinste Quadrate-Parameter nach sich. Die Auswirkungen können nachgewiesen werden, indem ein Schätzmodell spezifiziert wird, das zufällige Fehler in den Variablen berücksichtigt. Eine fehlerhafte Beobachtung X_{ij} für ein Merkmal j sei nun beschrieben durch

$$X_{ij} = X_{ij}^* + \psi_{ij}^{\varepsilon}, \tag{2.8}$$

wobei die Zufallsvariable ψ_{ij}^{ε} ein zufälliger Messfehler und X_{ij}^* die Zufallsvariable der „wahren“ Beobachtung ist, die hinter X_{ij} verborgen ist.

Wansbeek und Meijer [2003] verwenden die Modellierung in Gleichung (2.8) gleichrangig sowohl für „wahre“ Beobachtungen X_{ij}^*, die grundsätzlich messbar wären, als auch für Theorievariable, die selber nicht operationalisierbar sind.[69] In beiden Fällen liegen im

[69] Vgl. Wansbeek und Meijer 2003, S. 162.

betrachteten Schätzmodell keine Beobachtungswerte vor, so dass sie „latente“ Variable im Modell darstellen. Im Licht der eingangs getroffenen Unterscheidung zwischen Ungenauigkeit und Vagheit auf Seite 16 sind aber Zweifel an dieser Gleichsetzung angebracht. Diese basiert auf der Annahme, dass die Differenz zwischen den vorliegenden Beobachtungswerten und den zugehörigen Werten der „latenten“ Variablen eindeutig bestimmt und auf einer reellwertigen Skala abgetragen werden kann. Die Bestimmung eines solchen reellwertigen Abstands mag auf der Grundlage des festgelegten Messverfahrens bei Fehlern in den Daten noch gelingen. Die Quantifizierung des Unterschieds zwischen theoretischen Begrifflichkeiten, von denen nur eine operationalisierbar ist, muss hingegen rein hypothetisch erscheinen. Im allgemeinen kann weder vorausgesetzt werden, dass der Unterschied skalierbar ist, noch dass die Unterschiede sich eindimensional als Abweichung nach oben bzw. nach unten erfassen lassen.

Im „klassischen Messfehlermodell“[70] als dem einfachsten Modell mit stochastischen Fehlern in den Variablen setzen wir voraus, dass die „wahren“, aber unbekannten Merkmalsvariablen das Grundmodell der linearen Regression erfüllen mit

$$\mathbf{Y} = (\mathbf{X}^*)^T\mathbf{a}^* + \boldsymbol{\varepsilon}^*,$$

wobei $\mathbf{Y}$ der Stichprobenvektor für die abhängige Variable, $\mathbf{X}^*$ die Stichprobenmatrix für die unabhängigen Variablen der „wahren“ Werte, $\mathbf{a}^*$ die „wahren“ Parameter und $\boldsymbol{\varepsilon}^*$ der Vektor der Störvariablen des „wahren“ Zusammenhangs sind. Das unverfälschte Schätzmodell entspreche der einfachen, linearen Regression, d.h. dafür gelten die Gauß-Markov-Bedingungen (GM.1), (GM.2) und (GM.3). Für die zufälligen Fehler ψ_{ij}^{ε} gelte, dass sie stochastisch unabhängig sind, dass für alle $1 \leq i \leq n, 1 \leq j \leq k$ gilt $\mathrm{E}\,\psi_{ij}^{\varepsilon} = 0$ und $\mathrm{Var}\,\psi_{ij}^{\varepsilon} = \rho_j^2$, wobei $\rho_j^2 = 0$ genau dann, wenn in den Beobachtungen für das Merkmal j kein Fehler vorliegt. Ferner gelte für alle i, j, dass $\mathrm{Cov}(X_{ij}^*, \psi_{ij}^{\varepsilon}) = \mathrm{Cov}(\varepsilon_i^*, \psi_{ij}^{\varepsilon}) = 0$.

Schon in diesem Modell, bei dem die möglichen Abhängigkeiten auf ein Minimum reduziert sind, führen die Fehler bei einer Schätzung aufgrund der vorliegenden Stichprobe $(\mathbf{X}, \mathbf{Y})$ zu einer Abhängigkeit der exogenen Variablen $\mathbf{X}$ und der Störvariablen. Denn es gilt

[70] Vgl. Angrist und Krueger 1999, S. 1340.

$$\mathbf{Y} = \mathbf{X}^T\mathbf{a} + \boldsymbol{\varepsilon} \quad \text{mit} \quad \boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}^* - \boldsymbol{\psi}^{\varepsilon}\mathbf{a}^*, \tag{2.9}$$

wobei

$$\boldsymbol{\psi}^{\varepsilon} = \begin{pmatrix} 1 & \psi^{\varepsilon}_{11} & \cdots & \psi^{\varepsilon}_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \psi^{\varepsilon}_{n1} & \cdots & \psi^{\varepsilon}_{nk} \end{pmatrix},$$

dabei seien $\mathbf{a}$ der Parametervektor und $\boldsymbol{\varepsilon}$ der Vektor der Störvariablen im transformierten Schätzmodell, in dem die Fehler in den Variablen berücksichtigt werden. Infolge der Abhängigkeit zwischen $\mathbf{X}$ und $\boldsymbol{\varepsilon}$ führt die Kleinste Quadrate-Approximation zu Parameterschätzungen, die nicht erwartungstreu und zudem nicht konsistent sind und somit auch asymptotisch nicht gegen den „wahren“ Parameter konvergieren.[71]

Bei einer univariaten Regression mit nur einer exogenen Variablen wird bei einer naiven Regression von Y auf X, wenn die Fehler in der unabhängigen Variable nicht beachtet werden, tatsächlich die folgende Modellgleichung geschätzt[72]

$$Y = a_0' + a_1^*\lambda X_1 + \varepsilon, \quad \text{wobei} \quad \lambda = \frac{\operatorname{Cov}(X_1^*, X_1)}{\operatorname{Var} X_1} = \frac{\operatorname{Var} X_1^*}{\operatorname{Var} X_1}.$$

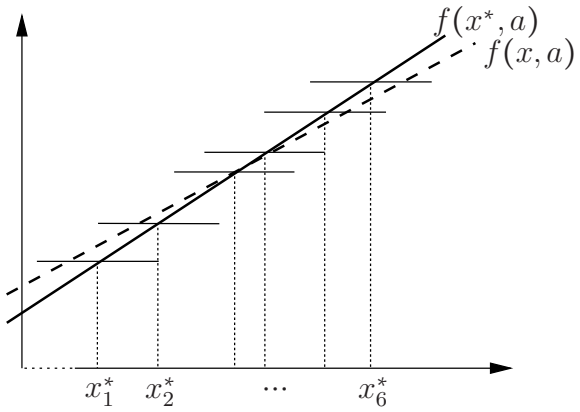
Für die verallgemeinerte Formulierung auf der Ebene der Modellgleichung ist zu beachten, dass für alle $1 \leq i \leq n$ gilt $\operatorname{Var}(\psi^{\varepsilon}_{i1}) = \rho_1^2 = \operatorname{Var}(\psi^{\varepsilon}_1)$, dies gilt entsprechend auch für $\operatorname{Var}(X_1^*)$[73]. Die Verzerrung im Steigungsparameter ergibt sich damit als $(1 - \lambda) < 1$[74]. Ein positiver Steigungskoeffizient $a_1 > 0$ wird also nach unten verzerrt und ein negativer Steigungskoeffizient $a_1 < 0$ nach oben. Komplementär dazu wird der Achsenabschnitt dann

[71] Zur Spezifikation des klassischen Fehlermodells vgl. Schneeweiß 1990, S. 218ff. Die Situation ist vergleichbar mit einer Miss-Spezifikation durch Ausschluss einer exogenen Variablen X_{k+1}, nämlich $X_{k+1} = -\frac{1}{a_{k+1}}(a_0^* + \sum_{j=1}^{k} a_j^*\psi_j^{\varepsilon})^T$, wobei ψ_j^{ε} die Variable für die Variablen in der allgemeinen Modellgleichung darstellt. Dort gilt: je stärker die Korrelation zwischen einer verbleibenden Variablen X_j und der ausgeschlossenen exogenen Variablen X_{k+1} ist und je stärker der Zusammenhang zwischen X_{k+1} und Y ist, desto stärker weicht der geschätzte Parameter $\hat{a}_j^*$ im fehlspezifizierten Modell vom Parameter $\hat{a}_j$ im voll spezifizierten Modell ab. Zudem ist der Kleinste Quadrate-Schätzer $\hat{a}_j^*$ nicht erwartungstreu. Im univariaten Modell gilt, dass bei positiver Korrelation von X_1, X_2 der Achsenabschnitt unterschätzt und die Steigung überschätzt wird; bei negativer Korrelation sind die Parameter in die jeweils entgegengesetzte Richtung verzerrt. Sind alle exogenen Variablen unkorreliert bzw. orthogonal, so verursacht eine fälschlich ausgeschlossene Variable keinen Effekt auf die Parameterkoeffizienten der verbleibenden Variablen. Vgl. Schneeweiß 1990, S. 148ff.

[72] Vgl. Angrist und Krueger 1999, S. 1341.

[73] Alternativ kann vorausgesetzt werden, dass die ψ_{ij}^{ε} bzw. die X_{ij}^* für alle i, j identisch verteilt sind.

[74] Die Verzerrung ist demnach umso größer, je größer der relative Fehler von X_1 gemessen an dem Verhältnis $\tau^2 = \frac{\operatorname{Var}(\psi_1^{\varepsilon})}{\operatorname{Var}(X_1^*)}$ ist. Für den verzerrten Parameter der univariaten Gleichung gilt dann $a_1' = a_1\lambda = a_1(1 + \tau^2)^{-1}$



Eigene Darstellung in Anlehnung an Schneeweiß 1990, S. 224.

Abbildung 2.2: Verzerrung der Parameterschätzer bei klassischen, zufälligen Fehlern — Jeder der „wahren“ Werte $x_1^*, x_2^*, \ldots, x_6^*$ wurde zweimal beobachtet und zwar jeweils einmal mit einem positiven und einmal mit einem negativen Fehler vom selben Betrag.

betragsmäßig zu hoch geschätzt. Dieses Phänomen wird wegen des Dämpfungseffekts auch *Attenuation*[75] genannt. Der Verzerrungseffekt wird in Abbildung 2.2 an einem einfachen Beispiel illustriert.

Im multivariaten Fall hat die Summe über alle Steigungsparameter ebenfalls asymptotisch eine Verzerrung gegen 0. Der Vektor der Schätzwerte liegt dabei in einem Halbraum, der durch eine Hyperebene durch den Vektor der „wahren“ Parameter nach oben begrenzt wird.[76] Insbesondere können einzelne Parameterschätzer weiter vom Nullpunkt entfernt sein als der „wahre“ Koeffizient. Dies gilt selbst dann, wenn die entsprechende Variable selber fehlerfrei beobachtet wurde. Die Varianz der „wahren“ Störvariablen $\boldsymbol{\varepsilon}^*$ wird hingegen asymptotisch überschätzt[77], demzufolge sind die Konfidenzbereiche für die Parameterwerte ebenfalls asymptotisch verzerrt und die Aussagekraft der Signifikanztests wird verfälscht.

Eine gesonderte Betrachtung des Effektes eines stochastischen Fehlers in der abhängigen Variablen Y kann entfallen, da dieser unter den Annahmen des „klassischen Fehlermodells“

[75] Der Begriff wird hauptsächlich in der englischsprachigen Literatur verwendet, vgl. Wansbeek und Meijer 2003, S. 164. Da die Wortbedeutung als solche auch im deutschen Sprachraum verwendet wird, wurde sie als Bezeichnung übernommen, vgl. Duden (1999–2004): Das Fremdwörterbuch, aktualisierte Online-Ausgabe, Dudenverlag: Mannheim; Zugriff über Munzinger-Online am 25. November 2009.

[76] Vgl. Wansbeek und Meijer 2003, S. 164. Es kann gezeigt werden, dass die Verzerrung in einer exogenen Variablen sich erhöht, wenn die Korrelation mit den übrigen beobachteten Variablen stark ist oder wenn sich die unabhängige Varianz des „wahren“ Anteils aufgrund der Korrelation mit den übrigen Variablen verringert. Gleichzeitig kann die Verzerrung aber auch dadurch abgeschwächt werden, dass die Ladung in der Variable den entgegengesetzten Bias in den übrigen Variablen, die durch Fehler in den Variablen verursacht werden, anteilig kompensiert, vgl. dazu Griliches 1986, S. 1479.

[77] Außerdem wird die Quadratsumme des erklärten Teils der Regression und damit auch R^2 asymptotisch unterschätzt. Vgl. Wansbeek und Meijer 2003, S. 164.

wegen der stochastischen Unabhängigkeit von $\boldsymbol{\psi}_Y^{\varepsilon}$ und $\boldsymbol{\varepsilon}^*$ zur Störvariable gerechnet werden kann und dort zunächst nur zu einer Erhöhung des Standardfehlers der Regression führt. Falls die Messfehler in $\mathbf{Y}$ nicht klassisch sind, dann sind die Regressionsparameter verzerrt, wobei die Verzerrung der Parameter gerade den Parametern einer virtuellen Regression des Messfehlers in Y auf die unabhängigen Variablen $X_1, \ldots, X_k$ entspricht.[78]

Liegen andere Fehlerstrukturen vor als im „klassischen Messfehlermodell", können völlig andere Verzerrungseffekte auftreten. So können, beispielsweise bei der Verwendung von „Pro Kopf"-Größen, sowohl die abhängige als auch eine unabhängige Variable denselben fehlerhaften Messwert im Nenner haben. Dies führt dazu, dass bei einem „wahren" Parameter 0 die geschätzten Parameter gegen 1 verzerrt sind.[79]

Zusammenfassend kann festgehalten werden, dass sowohl deterministische als auch zufällige Fehler in den Variablen zu asymptotischen Verzerrungen bei den Kleinste Quadrate-Parameter linearen Regression führen, wenn die Fehlereinflüsse vernachlässigt werden. Vor allem der Schätzer für den Achsenabschnitt wird durch Fehler verzerrt und zwar sowohl bei deterministischen als auch bei zufälligen Messungenauigkeiten. Da $\hat{a}_0$ auch auf andere Einflüsse sensitiv reagiert, hat er für die Interpretation des Schätzergebnisses allerdings wenig Bedeutung. Bei den Steigungsparametern ergeben sich Verzerrungen, wenn relative deterministische Fehler oder zufällige Fehler vorliegen. Dabei kann schon ein geringer zufälliger Fehler erhebliche Verzerrungen verursachen, wenn dieser stark mit den vorliegenden Beobachtungen für die Merkmalsvariablen korreliert ist. Darüber hinaus wirkt sich die Verzerrung auf die Schätzung aller Parameter im Modell aus, so dass auch die Zusammenhangsstruktur der Steigungsparameter berührt ist. Zudem verzerren Messfehler auch den Schätzer für die Parameterkovarianz und die damit ermittelten Konfidenzbereiche. Alle Schätzer sind nicht konsistent, so dass die Verzerrungen nicht ausgeräumt werden können, indem die Stichprobe vergrößert wird. Hingegen wirken sich deterministische Fehler nur lokal auf den zugehörigen Schätzer aus, wobei die Qualität des entsprechenden Signifikanztests unberührt bleibt.

Um zu unverzerrten Parameter-Schätzern bei Fehlern in den Variablen zu kommen, müssen zusätzliche Informationen vorliegen. Der Einfluss eines deterministischen Fehlers kann nur dann korrigiert werden, wenn dieser quantitativ bestimmt wurde, so dass das Modell entsprechend transformiert werden kann. Bei zufälligen Fehlern in den Variablen ist ein unverzerrten Schätzer für den Steigungsparameter dann bestimmbar, wenn obere Grenzen für die Fehlervarianzen angegeben werden können. Alternativ ist auch eine Schätzung mit Hilfe von Instrumentvariablen möglich[80].

[78] Vgl. Angrist und Krueger 1999, S. 1341.

[79] Vgl. Angrist und Krueger 1999, S. 1343.

[80] Wansbeek und Meijer [2003] umreißen die Möglichkeiten umfassender: „when searching for consistent

Das klassische Fehlermodell für die lineare Regression stellt eine starke Idealisierung dar, mit der insbesondere gezeigt werden kann, dass variierende Fehler in den Beobachtungen zu erheblichen Verzerrungen der geschätzten Parameter und der Signifikanztests führen können, sobald aufgrund der Fehlervariabilität eine Korrelation mit der Störvariablen entsteht. Dabei kann die Variabilität von Fehlern in den Variablen im Schätzmodell nur dadurch abgebildet werden, dass die Fehler als Zufallsvariable aufgefasst werden. Schneeweiß [1990] motiviert z.B. die Modellierung von zufälligen Fehlern so: „Folgen derartige Strukturbrüche [DN: aufgrund von sprunghaften Wechseln in den Fehlereinflüssen] rasch aufeinander, dann können sie u.U. den Charakter von Zufallsschwankungen annehmen“.[81] Grundsätzlich kann eine unverzerrte Schätzung nur dann erreicht werden, wenn das Modell richtig spezifiziert ist. Die Spezifikation von Fehlereinflüssen erfolgt aber in der Regel in Ergänzung zu der Spezifikation der „eigentlichen“ Modellzusammenhänge. Griliches [1986] weist daher darauf hin, dass die Korrektheit der Spezifikation von Messfehlern in den Daten davon abhängt, ob die Annahmen über die eigentlichen Modellvariablen korrekt sind.[82] In der Konsequenz stellt sich die Frage, ob die dichotome Abbildung von Fehlern in den Beobachtungen als deterministische Fehler einerseits und als zufällige Fehler andererseits reichhaltig genug ist, um den Einfluss von Fehlern bei der linearen Regression angemessen zu beschreiben. Die Frage nach den Unzulänglichkeiten bei der ökonometrischen Spezifikation von Fehlereinflüssen wird in Abschnitt *2.3* wieder aufgegriffen und erörtert.

2.3 Kritik am ökonometrischen Fehlermodell

In Abschnitt *2.1* wurde motiviert, dass grundsätzlich davon ausgegangen werden kann, dass wirtschaftsstatistische Beobachtungsdaten Fehler enthalten. Dabei steigt bei der Verwendung von Mikrodaten in mikroökonometrischen Modellen die Bedeutung von Messfehlern für die Güte der Parameterschätzungen — sowohl hinsichtlich ihres relativen Anteils als auch hinsichtlich ihrer Streuung. Dennoch werden Datenungenauigkeiten und deren Einfluss auf die Schätzgenauigkeit bei ökonometrischen Datenanalysen häufig vernachlässigt.[83] Anhand des klassischen Fehlermodells wurde in Abschnitt *2.2.2* illustriert, dass

estimators, additional information, instruments, nonnormality, or additional structure (...) are desirable“ (S. 166). Falls alle Stichprobenwerte im selben Orthanten liegen, dann liegt der unverzerrte Parameterschätzer in dem Polygon, das durch die $k+1$ Parametervektoren aufgespannt wird, die man erhält, wenn man den Kleinste Quadrate-Schätzer für die einfache, lineare Regression für jede Modellvariable auf die übrigen Variablen berechnet. Vgl. ebd. S. 165.

[81] Vgl. Schneeweiß 1990, S. 218.

[82] Vgl. Griliches 1986, S. 1485.

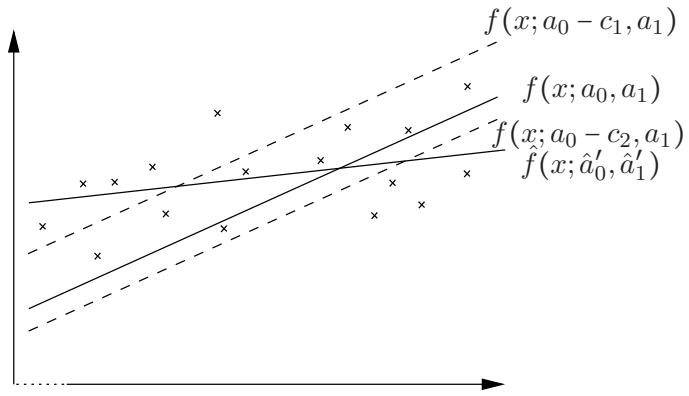
[83] Auf einen allzu sorglosen Umgang mit empirischen Daten bei ökonometrischen Analysen wird von verschiedenen Autor/innen aufmerksam gemacht. Als exemplarische Beispiele sei hier verwiesen auf [Pesaran

vor allem zufällig variierende Fehler in den Daten zu Korrelationen mit der Störvariablen sowie zu Verzerrungen der Parameter-Schätzer und der Signifikanztests führen. In diesem Abschnitt wurde bereits Kritik an der Modellierung der Fehlereinflüsse im ökonometrischen Schätzmodell angebracht, das nur eine Unterscheidung in vollständig funktional bestimmte, deterministische Fehler in den Merkmalsvariablen $\eta_{ij} = f_j(Z_{ij}(\omega))$ und in variierende Fehler kennt, die als Zufallsvariable ψ_{ij} modelliert werden können. Dabei kommt es wegen des approximativen Charakters des Schätzmodells weniger darauf an, ob die Fehler durch deterministische oder zufällige Einflüsse angetrieben sind, sondern ob sie sich im Gesamtergebnis ähnlich wie eine deterministische Variable oder wie eine Zufallsvariable darstellen. Somit sind Fehlereinflüsse für das Schätzmodell hauptsächlich dann relevant, wenn sie häufig auftreten. Dabei muss entweder gegeben sein, dass die Fehler gut durch einen funktionalen Zusammenhang von Modellvariablen beschrieben werden können, oder es muss gegeben sein, dass die Variation der Fehler beschränkt und im Verhältnis zur Stichprobengröße nicht zu groß ist.[84] Hingegen werden singuläre Fehlerereignisse nicht weiter betrachtet, diese sind normalerweise vorab bei der Bereinigung der Rohdaten zu bewerten und ggf. zu transformieren.

In der Praxis können nicht-zufällige Fehler meist nur mit einem Näherungswert im Bezug auf ein aggregiertes Messergebnis bestimmt werden. Eine Modellierung von deterministischen Fehlern ist also dann relevant, wenn diese im Verhältnis zur Skalierung groß sind und die Fehlerwerte nur unwesentlich um den mittleren Fehler variieren. Diese Modellierung von stetigen, deterministischen Fehlern als Funktion der Beobachtungswerte gibt zwar unsere alltägliche Erfahrung wieder, z.B. wenn eine Waage das Gewicht immer um 10g zu wenig anzeigt. Bei einem großen Teil ökonometrischer Datensätze ist es in der Praxis aber kaum möglich, den systematischen Fehler mit einem eindeutigen Wert abzubilden. Denn gerade die deterministischen Fehler können häufig nur ansatzweise bestimmt werden. In Abschnitt *2.1* wurde deutlich, dass Informationen über die Fehler in den Variablen üblicherweise in unterschiedlichen Formaten vorliegen, die von qualitativen Bewertungen über die Schwere einzelner Fehler oder den Einfluss einzelner Fehlerarten über Kennzahlen, wie z.B. Nichtbeantwortungsquoten, bis hin zu externen Abschätzungen reichen, wie z.B. Fehlerschätzungen auf der Basis von Kontrollstichproben oder Korrekturwerte anhand von unabhängigen Statistiken. Ein typischer Fall ist es beispielsweise, dass über die vorliegenden Beobachtungswerte bekannt ist, dass sie die tatsächlichen Werte tendenziell von unten annähern, d.h. die „wahren“ Werte liegen fast immer darüber. Statt einer

und Smith, 1992; Mayer, 1993; Richter, 2002].

[84] Zwar genügt auf der formalen Ebene für die asymptotische Betrachtung, dass die Streuung der Fehler in jeder Beobachtung σ^2 entspricht. Die Fehler könnten somit beliebig groß werden, solange die Fehlereinflüsse durch eine endliche Konstante nach oben beschränkt sind. Für eine konkrete Schätzung ist es aber relevant, dass die Zahl der Parameter, die zu bestimmen sind, im Verhältnis zur Stichprobengröße nicht zu groß wird. Die Fehler sollten sich also möglichst entsprechend derselben Fehlerverteilung verhalten.



(Eigene Darstellung)

Abbildung 2.3: Auswirkungen eines sprunghaften Wechsels im deterministischen Fehler — Dabei liegen fast alle Beobachtungswerte mit Fehler c_1 in einem anderen Sektor der X-Achse als die Beobachtungswerte mit Fehler c_2.

reellwertigen Fehlerabbildung wäre dann eher eine Abschätzung über den Fehlereinfluss geeignet, um die mögliche Lage des „wahren“ Wertes zu betrachten.

Eine besondere Rolle nehmen stetige Fehlereinflüsse mit sprunghaften Änderungen ein. Sie können als Zwischenstufe und Übergang zwischen deterministischen und zufälligen Fehlern angesehen werden.

Den einfachsten Fall stellt die sprunghafte Änderung eines deterministischen Fehlers dar, also z.B. $X_{ij} = X^*_{ij} + c_1$ für $1 \leq i \leq m$ und $X_{ij} = X^*_{ij} + c_2$ für $m+1 \leq i \leq n$, wobei $c_1 \neq c_2$. Solche Sprünge sind z.B. naheliegend, wenn in einer Zeitreihenstatistik die Erhebungsmethode umgestellt wurde. Jeder der Fehler wirkt sich direkt auf den Achsenabschnitt a_0 aus. Die sprunghafte Änderung in den Fehlereinflüssen ist daher wie ein Strukturbruch zu betrachten.[85] In Abbildung 2.3 ist dargestellt, dass es zu erheblichen Verzerrungen der Kleinste Quadrate-Schätzung kommen kann, wenn die Beobachtungswerte zu den Fehlereinflüssen c_1 bzw. c_2 sich in disjunkten Umgebungen häufen. Wenn Indizien über einen sprunghaften Wechsel im Fehlerbias vorliegen, kann dieser ähnlich wie ein Strukturbruch geschätzt und getestet werden, wenn eine Klassifizierung der Beobachtungswerte nach den Fehlereinflüssen möglich ist.[86] Dieser einfache Fall zeigt, dass die Klassifizierung der Beobachtungswerte nach den Fehlerwerten und zudem die Bestimmung des jeweiligen Fehlerwertes wesentlich ist, um sprunghafte deterministische Fehlereinflüsse im Schätzmodell zu berücksichtigen. Insbesondere ist die Modellierung ähnlich wie bei einfachen

[85] Schneeweiß und Mittag [1986] bezeichnen dieses Phänomen als „scheinbaren Strukturbruch“ (S. 38f).

[86] In [Schneeweiß, 1990] wird ein einfacher Test vorgestellt, der allerdings auf der Annahme beruht, dass die Varianz in der Stichprobe konstant bleibt (S. 118ff). Allerdings gibt es in der Regel kaum Informationen darüber, inwieweit ein Methodenwechsel die Fehlerstreuung tatsächlich unberührt lässt. Eine wichtige Motivation für die Änderung der Messmethoden besteht sogar in der Regel gerade darin, dass die Fehleranteile reduziert werden sollen.

deterministischen Fehlereinflüssen dadurch erschwert, dass die Fehlerwerte häufig nicht eindeutig bestimmt werden können.[87] Dazu kommt, dass zur Bestimmung von sprunghaften Fehlereinflüssen zudem die Beobachtungswerte nach den unterschiedlichen Fehlern zu klassifizieren sind. Diese Zuordnung ist aber nicht einfach, wenn die Fehler durch Änderungen in den gesellschaftlichen Einstellungen oder im administrativen Verfahren induziert sind.

Stetige Fehlereinflüsse mit sprunghaften Änderungen markieren den Übergang zwischen deterministischem und zufälligem Fehler im Schätzmodell. Wenn die Genauigkeit der Beobachtungen stark auf wechselnde gesellschaftliche Einstellungen oder auf administrative Änderungen reagiert, dann sind häufigere Sprünge in den Fehlern möglich. Bei zunehmender Zahl vom Sprüngen in den deterministischen Fehlern in der Stichprobe ist aber irgendwann eine Schätzung der einzelnen deterministischen Fehler nicht mehr möglich. Es liegt dann nahe zu einem Ansatz überzugehen, in dem die Fehler in den Variablen als Zufallsvariable modelliert sind.

Grundsätzlich kann Variabilität von Fehlern infolge der dichotomen Fehlermodellierung im klassischen Schätzmodell ausschließlich durch Zufallsvariable abgebildet werden. Es gibt daher eine gewisse Tendenz alle Fehlereinflüsse als Zufallsvariable aufzufassen, die nicht auf eine eindeutige Maßzahl verdichtet werden können. So kann eine Abschätzung für einen unbekannten Fehler durch eine Gleichverteilung abgebildet werden, oder mehrere sprunghafte Wechsel in den Fehlereinflüssen werden dadurch zusammengefasst und vereinfacht, indem man annimmt, dass sie ein und derselben Verteilung genügen. Eine Modellierung als Zufallsvariable ist im Schätzmodell aber nur unter bestimmten Voraussetzungen angemessen. Insbesondere erfordert die Modellierung, dass die Existenz einer festen Häufigkeitsverteilung für die modellierten Ergebnisse plausibel ist. Als Indizien für die Existenz einer festen Häufigkeitsverteilung für eine Fehlervariable sollten die folgenden Anforderungen erfüllt sein: die Stichprobe für den Fehlereinfluss sollte groß genug sein, um einerseits die Annahme einer stabilen Häufigkeitsverteilung zu stützen und andererseits das Streuungsintervall der Fehler möglichst auszuschöpfen, dabei sollten die Stichprobenwerte unter stabilen Bedingungen gezogen sein. Die Bedeutung einer ausreichenden Größe der Stichprobe steigt noch, wenn die Fehlerverteilung als sehr schief anzusehen ist. Die Verzerrungen im klassischen Fehlermodell der Regression ergeben sich als direkte Folge der Modellierung als reellwertige Zufallsvariable. Falls keine ausreichenden Informationen

[87] Grundsätzlich genügt es für eine unverzerrte Schätzung der Steigungsparameter, die Differenzen im Achsenabschnitt zu bestimmen, die durch die Sprünge im deterministischen Fehler induziert werden. Ein Wert für diesen Sprungabstand kann in einigen Fällen durch Vergleich der Messreihen zwischen zwei deterministischen Fehlereinflüssen abgeleitet werden. Aber auch dann ist zu fragen, ob dieser Fehlerwert tatsächlich so genau ist, wie er aufgrund der reellwertigen Abbildung erscheint, denn es verbleibt daneben ein ungenauer Anteil der deterministischen Fehler, der auf den Modellzusammenhang einwirkt. Außerdem bleibt das Problem bestehen, dass die Fehlereinflüsse noch mit Heteroskedastizität behaftet sind.

über Abhängigkeiten vorliegen, kann es sein, dass die vorliegenden Korrelationen damit überzeichnet werden und infolgedessen Scheinkorrelationen erzeugt werden.[88]

Die klassische Fehlermodellierung im Schätzmodell der Regression ist überdies auf einer grundsätzlichen Ebene unbefriedigend, weil Fehler nur indirekt durch Strukturaussagen modelliert werden können. Die üblichen reellwertigen Daten beinhalten als solche keine Informationen über ihre Genauigkeit und Richtigkeit und können in die statistischen Analysewerkzeuge eingespeist werden, ohne dass die Gültigkeit der Strukturannahmen hinterfragt wird. Dies verführt zum unkritischen Umgang mit dem Einfluss von Fehlern in den Daten auf die Güte der Modellanpassung. Zudem werden diejenigen Informationen, die über die Messfehler bekannt sind, nur als sekundäre Merkmale in der Modellgleichung abgebildet. Die Spezifikation der Messfehler ist damit von der Richtigkeit der Spezifikation des „eigentlichen" Strukturzusammenhangs abhängig. Die Spezifikation der Fehlereinflüsse erscheinen infolgedessen nachrangig gegenüber der Spezifikation der Variablen im Gesamtmodell. Darin kann vermutlich ein Grund dafür gesehen werden, dass häufig die Auswirkungen von Messfehlern in den Beobachtungsdaten und auf die Schätzergebnisse nicht weiter untersucht werden.

Aus der Sicht der Semantik wäre zudem zu problematisieren, ob angesichts der Unbestimmtheit der Fehlerbeträge dafür überhaupt eine klassische Zufallsverteilung wie für einfache, reellwertige Beobachtungen bestimmt werden kann. Diese Kritik kann grundsätzlich auf alle latenten Variablen in ökonometrischen Schätzmodellen ausgeweitet werden.[89]

In dieser Arbeit soll daher ein Perspektivwechsel vorgenommen werden, bei dem Fehler zuvorderst als Eigenschaft der Beobachtungswerte aufgefasst werden und nicht als Strukturkomponente des Modells. Dazu wird im Folgenden dem klassischen Modell zur Fehlerschätzung ein Modell mit Fuzzy-Daten gegenübergestellt, bei dem die möglichen fehlerinduzierten Verzerrungen in den Beobachtungen auf dem Stand des vorhandenen Wissens bewertet werden. Der Vorteil eines solchen Modells liegt darin, dass die vorliegenden Informationen über die Genauigkeit der Daten vollständig einbezogen werden können. Dies ist besonders dann hilfreich, wenn deterministische Fehler nur durch eine Abschätzung beschrieben werden oder wenn die Strukturzusammenhänge von zufälligen Fehlereinflüssen nicht genau genug spezifiziert werden können.[90] Der Fuzzy-Ansatz fokus-

[88] Zum Unterschied zwischen der Addition von stochastisch unabhängigen Zufallsvariablen und von Fuzzy-Werten vgl. das Beispiel in Abschnitt *2.5.2*, S. 74.

[89] Im Rahmen der Fuzzy-Theorie kann diese Fragestellung für Fuzzy-Zufallsvariablen untersucht werden. Dabei kommt man zu dem Ergebnis, dass tatsächlich für unterschiedliche Interpretationen der Fuzzy-Zufallsvariablen unterschiedliche Verteilungen relevant sind. Die Verteilungsmodelle für Fuzzy-Zufallsvariable werden in Abschnitt *6.1* dargestellt.

[90] Als ein Fernziel von Regressionsmodellen mit Fuzzy-Fehlern kann daher gesehen werden, dass auch unvollständige oder nur anteilige Korrelationen damit modelliert und geschätzt werden können. Dazu wird ein entsprechendes Schätzmodell mit Fuzzy-Zufallsvariablen benötigt.

siert dabei auf die Abbildung der tatsächlich verfügbaren Information und stellt damit im Unterschied zu den ökonometrischen Strukturannahmen den Grad des (Nicht-)Wissens und dessen Bewertung in den Mittelpunkt der Analyse. Insbesondere baut ein Fuzzy-Modell wesentlich auf der fachlichen Bewertung der Datenqualität auf und bietet somit die Möglichkeit, das Expert/innenwissen direkt im Modell abzubilden.

2.4 Fuzzy-Mengen und Fuzzy-Vektoren

Bei der Modellierung von Unsicherheit durch eine Zufallsvariable wird die unbekannte Lage eines Messwerts mit Hilfe eines Verhaltensmodells beschrieben, das die Variabilität der Werte in Analogie zu einem Zufallsexperiment abbildet, dessen relative Häufigkeiten sich bei häufiger Wiederholung stabilisieren. Diese Modellierung hat den Vorteil, dass die Verhaltensstabilitäten durch den Erwartungswert sowie ggf. höhere Momente der Verteilung auf der Grundlage von vorliegenden Daten charakterisiert werden können. Eine Beschreibung als Zufallsvariable ist aber nur dann sinnvoll, wenn ausreichendes Wissen für die Spezifikation des Verhaltensmodells verfügbar ist. Statistische Modellierungen sollten daher nach Möglichkeit durch eine ausreichende Anzahl von annahmegerechten Daten hinterlegt werden können. Hingegen sind Schlüsse, die sich in letzter Konsequenz nur auf eine Realisation beziehen, als eher fragwürdig zu betrachten.[91]

Die Fuzzy-Mengen-Theorie stellt einen weiteren Ansatz zur Beschreibung von Datenunschärfen dar, mit dem vor allem Wissenseinschränkungen und Ungenauigkeiten bei der numerischen Beschreibung eines Phänomens abgebildet werden können. Die Fuzzy-Modellierung geht von der subjektiven Abschätzung der möglichen Werte eines Objektes aus, die für jede Beobachtung individuell zu beschreiben sind, verallgemeinernde Verhaltensannahmen treten demgegenüber in den Hintergrund. Damit werden die aktuelle Wahrnehmung bzw. die verfügbare Information über den aktuellen Zustand des Phänomens zum zentralen Gegenstand der Abbildung.

Ein wichtiger Auslöser für die Entwicklung der Fuzzy-Mengen-Theorie war die Erkenntnis, dass Modelle zur Datenapproximation trotz einer zunehmenden Verfeinerung der Methoden dennoch mit wachsender Komplexität der Modelle häufig zu realitätsfremden Ergebnissen führen. Zadeh, der Gründungsvater der Fuzzy-Mengen-Theorie, hat daher in seinem Aufsatz über „Fuzzy sets“[92] die Beschreibung von Fuzzy-Mengen mit der programmatischen Aufgabenstellung verknüpft, dass mathematische Modelle für einen praktischen Sachverhalt nur so genau formuliert werden sollen, wie es für die Lösung

[91] Vgl. Bandemer 1997, S. 188.
[92] Vgl. [Zadeh, 1965].

des Problems gerade angemessen ist.[93] Als sehr fruchtbar hat sich dieser Ansatz der bewusst unscharfen Modellierung beispielsweise bei der Konstruktion von technischen Reglern erwiesen, wo es durch Nachbildung von menschlichem Kontrollverhalten in Form von einfachen Regeln gelungen ist, komplexe Steuerungsprobleme mit linguistischer Inferenz zu lösen. Im Standardbeispiel des sog. „umgekehrten Pendels", d.h. eines aufrecht stehenden Stabes, der durch seitliche Bewegung eines Wagens zu balancieren ist, genügen dazu wenige Regeln der Art „`wenn` der Stab `stark` nach `links` geneigt ist, `dann` fahre `stark` nach `rechts`". Zu den zahlreichen Anwendungsgebieten der Fuzzy-Mengen-Theorie gehören unter anderem auch Entscheidungsunterstützung mit Fuzzy-Optimierung oder mit Fuzzy-Expert/innensystemen, Künstliche Intelligenz sowie Data Mining mit Fuzzy-Clusteranalyse.

In Abschnitt *2.4.1* werden zunächst einige Grundbegriffe der Fuzzy-Theorie zur Verfügung gestellt, die zur Definition von Fuzzy-Vektoren und Fuzzy-Funktionen benötigt werden. Für Fuzzy-Vektoren können Addition und Skalarprodukt eingeführt werden. Allerdings gelten für Fuzzy-Vektoren arithmetische Besonderheiten, so dass nicht in der üblichen Weise eine Vektorraumstruktur sichergestellt wird. Es existiert dennoch eine lineare Struktur, für die allerdings Einschränkungen bei der Addition und der Skalarmultiplikation zu berücksichtigen sind, wie in Abschnitt *2.4.2* aufgezeigt wird.

2.4.1 Fuzzy-Mengen: Definitionen und Grundbegriffe

Mit Fuzzy-Mengen können vage Begriffe oder unscharfe Grenzen dadurch dargestellt werden, dass ein Bereich des allmählichen Übergangs zwischen Zugehörigkeit und Nicht-Zugehörigkeit zu einer Menge bzw. einer Kategorie zugelassen und mathematisch abgebildet wird. Fuzzy-Mengen eignen sich somit gut, die ungenaue menschliche Wahrnehmung sowie menschliches Entscheidungsverhalten für eine maschinelle Verarbeitung nachzubilden und mit mathematischen Methoden zu behandeln. Insbesondere ermöglicht eine Fuzzy-Modellierung es, den tatsächlichen, eingeschränkten Informationsstand besser wiederzugeben. Somit können Idealisierungen aufgegeben werden, die zur Gewinnung von reellwertigen Daten häufig notwendig sind. Denn diese vermitteln im allgemeinen eine falsche Sicherheit über das Ergebnis, falls die Messungen durch Fehler verzerrt sind.

In diesem Teilkapitel sollen die grundlegenden Begriffe kursorisch vorgestellt werden, so dass unkundige Leser/innen ein Gefühl für die Operationen auf Fuzzy-Mengen erhalten. Außerdem sollen wichtige Begriffe für die Arbeit in einer exakten Definition zur Verfügung gestellt werden. Für eine ausführliche Einführung in die Fuzzy-Theorie wird auf die

[93] Vgl. Bandemer 1997, S. 61.

einschlägige Literatur verwiesen.[94]

Die Definition einer Fuzzy-Menge ist abgeleitet von der Repräsentation einer Menge durch eine Indikatorfunktion. Wenn A eine Menge über der Grundmenge U ist, dann kann A äquivalent dargestellt werden durch die Funktion $1_A : U \longrightarrow \{0,1\}$ mit

$$1_A(u) = \begin{cases} 1, & \text{falls } u \in A, \\ 0, & \text{falls } u \notin A. \end{cases}$$

Bei einer unscharfen Menge wird die Menge der Zugehörigkeitswerte über die binäre Bewertung hinaus auf beliebige Werte im Intervall $[0,1]$ erweitert.

Definition 2.4 (Fuzzy-Menge) Sei U ein Grundmenge zulässiger Werte. Dann ist eine *Fuzzy-Menge* $\widetilde{A}$ über U definiert durch

$$\widetilde{A} = \{(u, \mu_{\widetilde{A}}(u)) \mid u \in U\},$$

wobei $\mu_{\widetilde{A}} : U \longrightarrow [0,1]$.

$\mu_{\widetilde{A}}$ heißt *Zugehörigkeitsfunktion* von $\widetilde{A}$. Wir interpretieren $\mu_{\widetilde{A}}(u)$ als den *Zugehörigkeitsgrad* von u zu $\widetilde{A}$.

Die Gesamtheit aller Fuzzy-Mengen über U ist die *Fuzzy-Potenzmenge*

$$\mathcal{F}(U) = \{\widetilde{A} \mid \widetilde{A} \text{ ist Fuzzy-Menge über } U\}.$$

Eine Fuzzy-Menge $\widetilde{A}$, die nur binäre Zugehörigkeitswerte annimmt, d.h. $\mu_{\widetilde{A}}(\mathbb{R}) \in \{0,1\}$, nennen wir zur Unterscheidung manchmal auch *unechte oder crispe Fuzzy-Menge*, weil sie dann einer gewöhnlichen Menge entspricht.

Die grundlegenden Mengenoperationen können mit den folgenden Definitionen auf Fuzzy-Mengen übertragen werden. Diese basieren auf dem min- und dem max-Operator[95].

Definition 2.5 Seien $\widetilde{A}, \widetilde{B} \in \mathcal{F}(U)$ Fuzzy-Mengen.

$\widetilde{A}$ ist *Teilmenge* von $\widetilde{B}$ genau dann, wenn gilt

$$\forall\, u \in U : \quad \mu_{\widetilde{A}}(u) \leq \mu_{\widetilde{B}}(u).$$

Wir schreiben $\widetilde{A} \subset \widetilde{B}$.

[94] Die Darstellungen in diesem Abschnitt stützen sich vor allem auf [Dubois und Prade, 1980; Bandemer und Näther, 1992; Kruse u. a., 1993; Rommelfanger, 1994].

[95] Dabei stellt min eine t-Norm und max eine t-Conorm dar. Unter Verwendung anderer Paare von t-Norm und t-Conorm können abweichende Mengenoperationen definiert werden, die für bestimmte Anwendungen ggf. besser geeignet sind. Vgl. auch hierzu [Zadeh, 1965].

Die *Schnittmenge* $\widetilde{A} \cap \widetilde{B}$ ist definiert durch die Zugehörigkeitsfunktion

$$\mu_{\widetilde{A}\cap\widetilde{B}}(u) = \min\{\mu_{\widetilde{A}}(u), \mu_{\widetilde{B}}(u)\}.$$

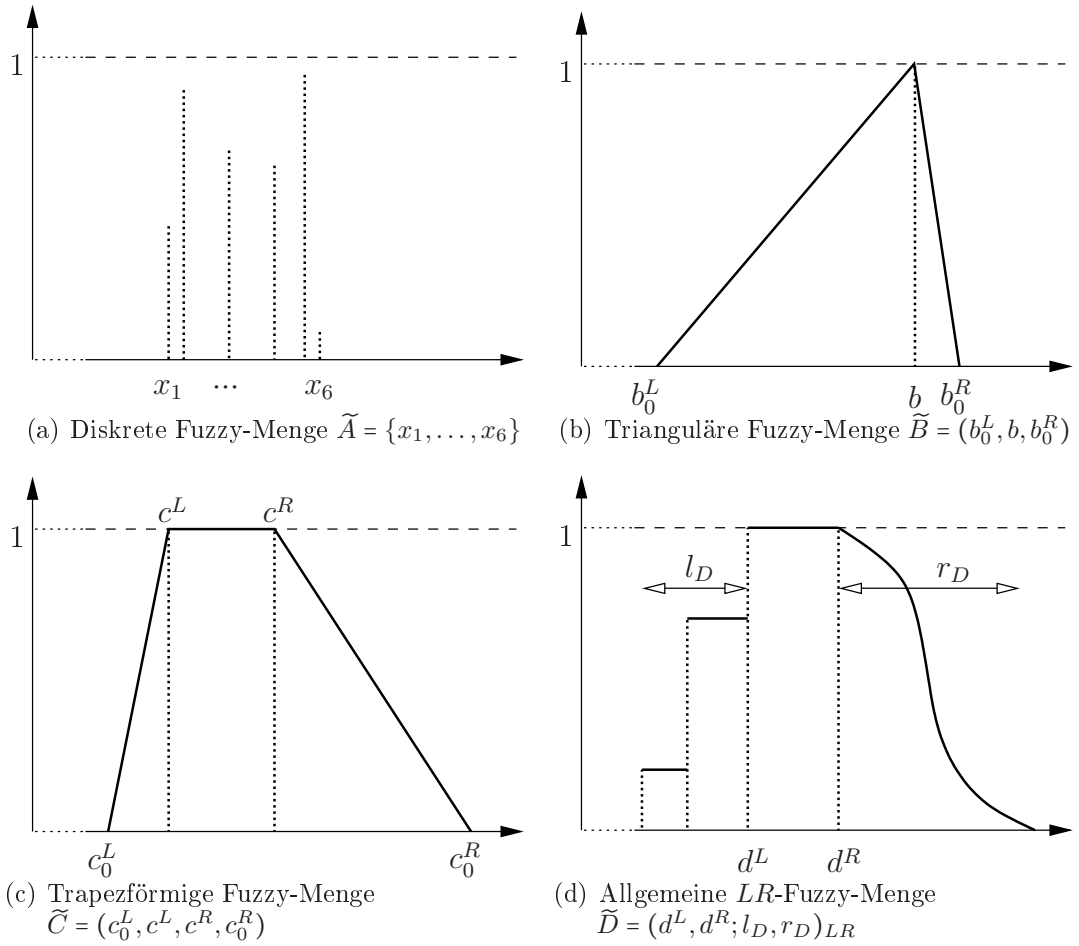
Die *Vereinigung* $\widetilde{A} \cup \widetilde{B}$ ist definiert durch die Zugehörigkeitsfunktion

$$\mu_{\widetilde{A}\cup\widetilde{B}}(u) = \max\{\mu_{\widetilde{A}}(u), \mu_{\widetilde{B}}(u)\}.$$

Das *Komplement* $\widetilde{A}^C$ in U ist definiert durch die Zugehörigkeitsfunktion

$$\mu_{\widetilde{A}^C}(u) = 1 - \mu_{\widetilde{A}}(u).$$

Anders in der klassischen Mengenlehre sind bei Fuzzy-Mengen eine Menge und ihr Komplement nicht mehr disjunkt. Es gilt im allgemeinen $\widetilde{A} \cap \widetilde{A}^C \neq \varnothing$.



(a) Diskrete Fuzzy-Menge $\widetilde{A} = \{x_1, \ldots, x_6\}$

(b) Trianguläre Fuzzy-Menge $\widetilde{B} = (b_0^L, b, b_0^R)$

(c) Trapezförmige Fuzzy-Menge $\widetilde{C} = (c_0^L, c^L, c^R, c_0^R)$

(d) Allgemeine LR-Fuzzy-Menge $\widetilde{D} = (d^L, d^R; l_D, r_D)_{LR}$

(Eigene Darstellung in Anlehnung an Rommelfanger 1994, S. 14f)

Abbildung 2.4: Einige typische Beispiele für Fuzzy-Mengen

In Abbildung 2.4 sind einige typische Beispiele von Fuzzy-Mengen dargestellt. Das Konstruktionsprinzip für Fuzzy-Mengen ist am Beispiel einer Fuzzy-Menge $\widetilde{A}$ über einer endli-

chen Menge $\{x_1, \ldots, x_n\} \subset \mathbb{R}$ gut nachvollziehbar. In Abbildung 2.4(a) ist dargestellt, dass dazu für jedes Element x_i ein Zugehörigkeitswert $\mu_i \in [0,1]$ anzugeben ist. Fuzzy-Mengen sollen unscharfe Beobachtungen oder Begriffe repräsentieren und in eine maschinenlesbare Form übertragbar sein. Exakte Zugehörigkeitsfunktionen zur Beschreibung von vagen oder ungenauen Begriffen können in der Regel nicht oder nur mit erheblichem Aufwand erfasst werden. Darüber hinaus ist eine übergroße Genauigkeit der Modellierung normalerweise nicht notwendig. In dieser Arbeit werden daher ausschließlich Fuzzy-Mengen für die Modellierung verwendet, die als LR-Fuzzy-Intervalle dargestellt werden können.

Definition 2.6 (LR-Fuzzy-Intervall) Seien $a^L, a^R \in \mathbb{R}$ und $l_A, r_A > 0$. Seien ferner $L, R : [0,\infty) \longrightarrow [0,1]$ Funktionen, für die gelte

(i) $L(0) = R(0) = 1$ und

(ii) L, R monoton fallend.

Dann ist durch

$$\mu_{\widetilde{A}}(x) = \begin{cases} L(\frac{a^L - u}{l_A}) & \text{falls } u \leq a^L, \\ 1 & \text{falls } u \in (a^L, a^R), \\ R(\frac{u - a^R}{r_A}) & \text{falls } u \geq a^R. \end{cases}$$

die *LR-Fuzzy-Menge* $\widetilde{A} = (a^L, a^R; l_A, r_A)_{LR}$ mit den *Spannweiten* l_A *und* r_A in $\mathcal{F}(\mathbb{R})$ definiert.

L und R sind die *linke (bzw. rechte) Referenzfunktion* von $\widetilde{A}$. Falls $L = R$ bzw. $a = a^L = a^R$ notieren wir verkürzt $\widetilde{A} = (a^L, a^R; l_A, r_A)_L$ bzw. $\widetilde{A} = (a; l_A, r_A)_{LR}$. Eine LR-Fuzzy-Menge ist in Abbildung 2.4(d) dargestellt, wobei die linke Referenzfunktion L treppenförmig und die rechte Referenzfunktion R stetig ist.

LR-Fuzzy-Mengen können als Fuzzy-Verallgemeinerung von klassischen Intervallen verstanden werden. Die Konstruktion einer LR-Fuzzy-Menge kann anhand weniger Profileigenschaften vorgenommen werden. Zum einen ist das Intervall aller Werte zu bestimmen, die positive Zugehörigkeit zur Menge haben. Zum anderen ist das Intervall aller Werte anzugeben, die sicher zur Menge gehören und somit eine Zugehörigkeit von 1 haben. Darüber hinaus kann noch eine Kurve spezifiziert werden, die den Übergang bzw. den Kontrast zwischen Nicht-Zugehörigkeit und sicherer Zugehörigkeit möglichst gut beschreibt. Eine besonders einfache Konstruktion ist möglich, wenn die Referenzfunktionen linear sind.

Definition 2.7 (Trapezförmige und trianguläre Fuzzy-Intervalle)
Seien $\widetilde{B} = (b^L, b^R; l_B, r_B)_{LL}, \widetilde{D} = (d^L, d^R; l_D, r_D)_{LL}$ LR-Fuzzy-Mengen in $\mathcal{F}(\mathbb{R})$ mit $L = R$.

Dann ist $\widetilde{D}$ eine *trapezförmige Fuzzy-Menge* genau dann, wenn $L(u) = 1 - u$. Ein Beispiel ist in Abbildung 2.4(c) dargestellt. Trapezförmige Fuzzy-Mengen werden häufig auch durch das Tupel der Intervallenden des Trägers und des Kerns angegeben, d.h.

$$\begin{aligned}\widetilde{D} &= (d^L - l_D, d^L, d^R, d^R + r_D) \\ &= (d_0^L, d^L, d^R, d_0^R)\end{aligned}$$

Eine trapezförmige Fuzzy-Menge $\widetilde{B}$ ist eine *triangluläre Fuzzy-Menge* genau dann, wenn die Menge der Punkte mit Zugehörigkeit 1 einelementig ist und somit $b = b^L = b^R$, vgl. Abbildung 2.4(b). Trianguläre Fuzzy-Mengen werden ebenfalls häufig in Tupelschreibweise notiert, d.h. $\widetilde{B} = (b - l_B, b, b + r_B) = (b_0^L, b, b_0^R)$.

Wenn man Fuzzy-Menge aus der Perspektive der Zugehörigkeit betrachtet, ist es interessant, welche Punkte einer Fuzzy-Menge für einen gegebene Zugehörigkeitsgrad $\alpha \in [0, 1]$ mindestens eine Zugehörigkeit von α aufweisen.

Definition 2.8 (α-Schnitte) Sei $\widetilde{A} \in \mathcal{F}(U)$ eine Fuzzy-Menge und $\alpha \in (0, 1]$. Dann heißt die Cantor-Menge

$$[\widetilde{A}]^\alpha = \{u \in U \mid \mu_{\widetilde{A}}(u) \geq \alpha\}$$

α-Schnitt bzw. *α-Niveau von $\widetilde{A}$*.

Für $\alpha = 0$ definieren wir

$$\operatorname{supp} \widetilde{A} = [\widetilde{A}]^0 = \overline{\{u \in U \mid \mu_{\widetilde{A}}(u) > 0\}}.$$

$[\mu_{\widetilde{A}}]^0$ ist der *Träger von $\widetilde{A}$* (engl. „support“).

Ein weiterer spezieller α-Schnitt ist das 1-Niveau $[\mu_{\widetilde{A}}]^1$. Dieses ist der *Kern von $\widetilde{A}$* (engl. „core“). Einen einelementigen Kern werden wir im Weiteren auch als *Modalwert* bezeichnen.

Eine Fuzzy-Menge $\widetilde{A}$ heißt *normal*, wenn sie einen nicht-leeren Kern hat, d.h. wenn gilt $[\mu_{\widetilde{A}}]^1 \neq \varnothing$.

Demnach sind alle LR-Fuzzy-Mengen als normale Fuzzy-Mengen definiert. Für beliebige $\alpha \in (0, 1]$ können die α-Schnitte einer LR-Fuzzy-Menge $\widetilde{A} \in \mathcal{F}(\mathbb{R})$ berechnet werden als

die Intervalle

$$\begin{aligned}[\widetilde{A}]^{\alpha} &= [a^L(\alpha), a^R(\alpha)] \\ &= [a^L - l_A L^{-1}(\alpha), a^R + r_A R^{-1}(\alpha)],\end{aligned} \tag{2.10}$$

wobei $L^{-1}(\alpha) = \sup\{u \in \mathbb{R} \mid L(u) \geq \alpha\}$ und $R^{-1}(\alpha)$ entsprechend.

Während trapezförmige und trianguläre Fuzzy-Mengen über ihre Trägermenge und ihren Kern bereits vollständig definiert sind, gilt dies im allgemeinen für LR-Mengen nicht. Es kann aber gezeigt werden, dass Fuzzy-Mengen durch ihre Familie von α-Schnitten eindeutig repräsentiert werden.[96] Der Repräsentationssatz ermöglicht es in vielen Fällen Berechnungen für Fuzzy-Mengen auf α-Niveaus übertragen und sie dort explizit durchzuführen. Die praktisch relevanten Fuzzy-Mengen über $\mathbb{R}$ können in aller Regel bereits hinreichend genau dargestellt werden, indem über die Definition von trapezförmigen oder triangulären Fuzzy-Mengen hinaus ein oder zwei weitere α-Niveaus angegeben werden.

Die folgenden nützlichen Eigenschaften können vermittels der α-Schnitte von klassischen Mengen auf Fuzzy-Mengen übertragen werden. Wir benötigen sie später bei der Konstruktion von Fuzzy-Daten in Abschnitt *2.5.1*, um wesentliche Eigenschaften von Datenvektoren auch bei Fuzzy-Daten sicherstellen zu können.

Definition 2.9 Sei $U \subset \mathbb{R}^m$ und $\widetilde{A} \in \mathcal{F}(U)$ eine Fuzzy-Menge.

$\widetilde{A}$ ist *konvex* genau dann, wenn

$$\forall\, \alpha \in (0,1]: \quad [\widetilde{A}]^{\alpha} \text{ ist konvex.}$$

$\widetilde{A}$ ist *kompakt* genau dann, wenn

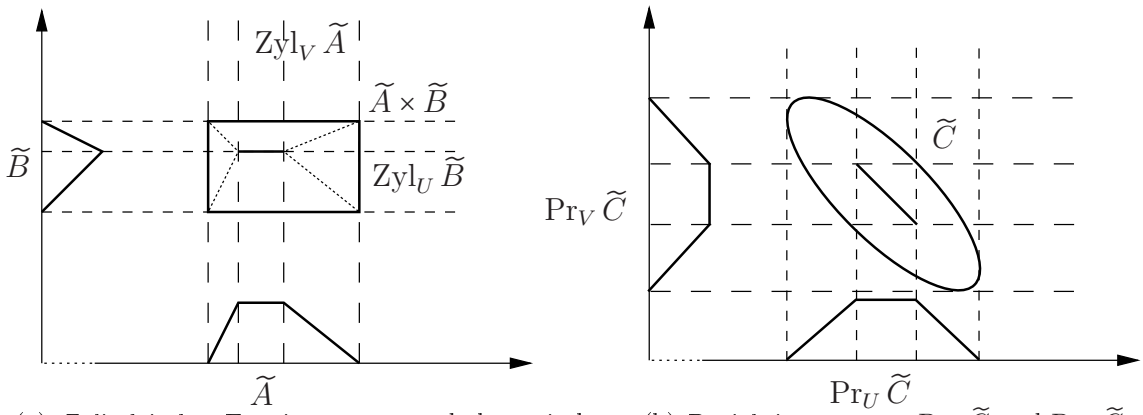
$$\forall\, \alpha \in (0,1]: \quad [\widetilde{A}]^{\alpha} \text{ ist kompakt.}$$

$\widetilde{A}$ ist *beschränkt* genau dann, wenn

$$[\widetilde{A}]^{0} \text{ ist beschränkt.}$$

Wenn Beobachtungen zu unabhängigen Variablen $X_1, \ldots, X_k$ und einer abhängigen Variablen Y untersucht werden sollen, liegen diese als Vektoren im $\mathbb{R}^{k+1}$ vor. Entsprechend werden auch Fuzzy-Vektoren im $\mathbb{R}^{k+1}$ benötigt. Dazu wird zunächst allgemein das kartesische Produkt von Fuzzy-Mengen definiert.

[96] Zum Repräsentationssatz von Fuzzy-Mengen durch α-Schnitte vgl. beispielsweise Kruse u. a. 1993, S. 17f.



(a) Zylindrische Erweiterungen und kartesisches Produkt von $\widetilde{A}$ und $\widetilde{B}$

(b) Projektionsmengen $\mathrm{Pr}_U\,\widetilde{C}$ und $\mathrm{Pr}_V\,\widetilde{C}$

(Eigene Darstellung in Anlehnung an Rommelfanger 1994, S. 35)

Abbildung 2.5: Fuzzy-Mengen im $\mathbb{R}^2$

Definition 2.10 (Mehrdimensionale Fuzzy-Mengen) Seien U, V Grundmengen. Seien nun $\widetilde{A} \in \mathcal{F}(U)$, $\widetilde{B} \in \mathcal{F}(V)$ sowie $\widetilde{C} \in \mathcal{F}(U \times V)$.

Die *zylindrische Erweiterung* $\mathrm{Zyl}_V\,\widetilde{A}$ von $\widetilde{A}$ auf $U \times V$ ist definiert durch

$$\mu_{\mathrm{Zyl}_V\,\widetilde{A}}(u,v) = \mu_{\widetilde{A}}(u).$$

Die zylindrische Erweiterung $\mathrm{Zyl}_U\,\widetilde{B}$ von $\widetilde{B}$ auf $U \times V$ ist analog definiert.

Das *kartesische Produkt* $\widetilde{A} \times \widetilde{B}$ auf $U \times V$ ist definiert als $\mathrm{Zyl}_V\,\widetilde{A} \cap \mathrm{Zyl}_U\,\widetilde{B}$, d.h.

$$\mu_{\widetilde{A}\times\widetilde{B}}(u,v) = \min\{\mu_{\widetilde{A}}(u), \mu_{\widetilde{B}}(v)\}.$$

In Abbildung 2.5 sind zylindrische Erweiterung und kartesisches Produkt im $\mathbb{R}^2$ illustriert.

Umgekehrt ist die *Projektion* $\mathrm{Pr}_U\,\widetilde{C}$ von $\widetilde{C}$ auf U definiert durch

$$\mu_{\mathrm{Pr}_U\,\widetilde{C}}(u) = \sup\{\mu_{\widetilde{C}}(u,v) \mid v \in V\}.$$

Eine analoge Definition gilt für $\mathrm{Pr}_V\,\widetilde{C}$.

Die Projektion macht die zylindrische Erweiterung wieder rückgängig und für jedes $\widetilde{A} \in \mathcal{F}(U)$ gilt

$$\mathrm{Pr}_U(\mathrm{Zyl}_V\,\widetilde{A}) = \widetilde{A}.$$

Für $\widetilde{C} \in \mathcal{F}(U \times V)$ gilt die umgekehrte Beziehung im allgemeinen aber nicht, sondern vielmehr

$$\mathrm{Zyl}_V(\mathrm{Pr}_U \widetilde{C}) \cap \mathrm{Zyl}_U(\mathrm{Pr}_V \widetilde{C}) \supset \widetilde{C},$$

wie in Abbildung 2.5(b) gezeigt wird. Gleichheit gilt hier insbesondere, wenn $\widetilde{C} = \widetilde{A} \times \widetilde{B}$ mit $\widetilde{A} \in \mathcal{F}(U)$ und $\widetilde{B} \in \mathcal{F}(V)$.

Dies gibt Anlass zu der folgenden Definition.

Definition 2.11 Eine Fuzzy-Menge $\widetilde{C} \in \mathcal{F}(U \times V)$ ist *interaktiv* genau dann, wenn gilt

$$\widetilde{C} \subsetneq \mathrm{Pr}_U \widetilde{C} \times \mathrm{Pr}_V \widetilde{C}.$$

Umgekehrt ist $\widetilde{C}$ *nicht-interaktiv* genau dann, wenn gilt

$$\widetilde{C} = \mathrm{Pr}_U \widetilde{C} \times \mathrm{Pr}_V \widetilde{C}.$$

Fuzzy-Mengen werden also dann als nicht-interaktiv aufgefasst, wenn die Gesamtmenge vollständig aus ihren Randmengen rekonstruiert werden kann. Interaktivität drückt in dem Sinne eine Abhängigkeit zwischen den Komponentenmengen aus, dass für manche möglichen Randwerte die Menge der möglichen zugehörigen Koordinatenpunkte verengt ist, d.h. es gibt in jedem $u \in \mathrm{Pr}_U \widetilde{C}$ bzw. in jedem $v \in \mathrm{Pr}_V \widetilde{C}$ nur eine „bedingte“ Fuzzy-Bildmenge.

Damit sind alle Begriffe eingeführt, mit denen wir Fuzzy-Vektoren über $\mathbb{R}^m$ definieren können. Grundsätzlich wäre es möglich, nur diejenigen Fuzzy-Mengen im $\mathbb{R}^m$ als Fuzzy-Vektoren aufzufassen, die einen einelementigen Kern haben. Allerdings kommt es nicht selten vor, dass ungenaue Beobachtungen nur als Fuzzy-Intervalle mit einem intervallförmigen Kern abgebildet werden können. Aus diesem Grund wird eine verallgemeinerte Definition vorgenommen. Mehrgipflige und unendliche Fuzzy-Mengen sollen aber keinesfalls als Fuzzy-Vektoren in Frage kommen.

Definition 2.12 (Fuzzy-Vektor) Sei $m \in \mathbb{N}$. Die Menge aller Fuzzy-Mengen über $\mathbb{R}^m$, die normal, beschränkt, konvex und kompakt sind, notieren wir mit $F_{coc}^{nob}(\mathbb{R}^m)$. Eine beliebige Fuzzy-Menge $\widetilde{C} \in F_{coc}^{nob}(\mathbb{R}^m)$ ist ein *m-dimensionaler Fuzzy-Vektor* über $\mathbb{R}^m$. Fuzzy-Vektoren $\widetilde{A} \in F_{coc}^{nob}(\mathbb{R})$ mit einelementigem Kern bezeichnen wir auch als *Fuzzy-Zahlen-Vektoren* oder *mehrdimensionale Fuzzy-Zahlen*.

Sei $\widetilde{C} \in F_{coc}^{nob}(\mathbb{R}^m)$ nun ein nicht-interaktiver Fuzzy-Vektor. Dann gibt es gemäß Definition 2.11 Fuzzy-Mengen $\widetilde{C}_1, \ldots, \widetilde{C}_m \in F_{coc}^{nob}(\mathbb{R})$, so dass gilt

$$\widetilde{C} = \widetilde{C}_1 \times \cdots \times \widetilde{C}_m.$$

Daher werden nicht-interaktive Fuzzy-Vektoren in Anlehnung an die übliche Vektorenschreibweise auch notiert als $\widetilde{C} = (\widetilde{C}_1, \ldots, \widetilde{C}_m)$.

Die Teilmenge aller nicht-interaktiven Fuzzy-Vektoren in $F_{coc}^{nob}(\mathbb{R}^m)$ bezeichnen wir zur Unterscheidung als $F_{coc}^{nob}(\mathbb{R})^m$.

Unabhängig von der inhaltlichen Information, die damit abgebildet wird, hat ein Fuzzy-Vektor $\widetilde{Z} \in F_{coc}^{nob}(\mathbb{R}^m)$ die Bedeutung „ungefähr z“. Das läßt sich aus den gegebenen Eigenschaften begründen: Als normale Fuzzy-Menge existiert $z \in \mathbb{R}^m$, so dass $\mu_{\widetilde{Z}}(z) = 1$, damit wird mindestens das Element z als sicher zugehörig betrachtet. Kompaktheit und Beschränktheit beinhalten, dass alle α-Schnitte sowie der Träger in $\mathbb{R}^m$ abgeschlossen und beschränkt sind. Aus der Konvexität folgt, dass $\widetilde{Z}$ eingipflig ist. Da für die α-Schnitte gilt, dass $\alpha \geq \beta$ genau dann wenn $[\widetilde{Z}]^\alpha \subset [\widetilde{Z}]^\beta$, zieht die Konvexität außerdem nach sich, dass der Akzeptanzgrad der Zahlen im Träger mit wachsender Entfernung vom Kern monoton abnimmt.

Um Fuzzy-Mengen als Erweiterung von gewöhnlichen Zahlen oder Mengen verwenden zu können, sind die Operationen auf reellwertigen Vektoren oder Mengen in geeigneter Weise zu übertragen. Ein wesentliches Prinzip zur Übertragungen von gewöhnlichen Funktionen auf Fuzzy-Eingangswerte ist das sog. *Erweiterungsprinzip* nach Zadeh.

Definition 2.13 (Erweiterung einer Funktion) Seien U, V beliebige Grundmengen und $f : U \longrightarrow V$ eine Funktion. Für $\widetilde{A} \in \mathcal{F}(U)$ ist durch

$$\mu_{f(\widetilde{A})}(v) = \sup\{\mu_{\widetilde{A}}(u) \mid u \in U : f(u) = v\}$$

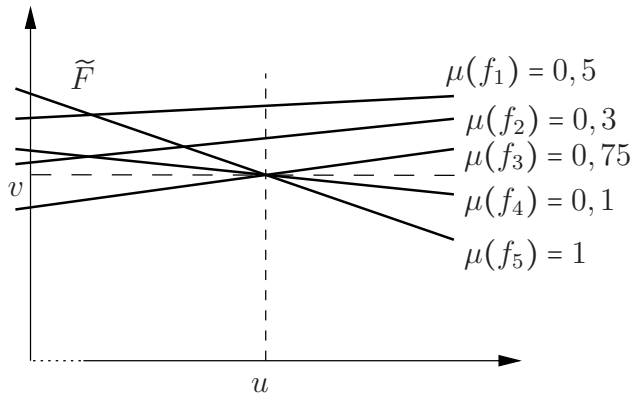
das (fuzzy-)erweiterte Bild $f(\widetilde{A})$ von $\widetilde{A}$ durch f definiert.

Die Funktion

$$\widetilde{f} : \mathcal{F}(U) \longrightarrow \mathcal{F}(V), \widetilde{A} \longmapsto f(\widetilde{A})$$

wird *erweiterte Funktion von* f genannt.

Aus Bequemlichkeit wird üblicherweise auch die erweiterte Funktion mit f bezeichnet. Dies lässt sich dadurch motivieren, dass die erweiterte Funktion für alle unechten Fuzzy-Eingangswerte, also für alle gewöhnlichen Eingangswerte, die ursprünglichen Bildwerte ergibt.



(Eigene Darstellung)

Abbildung 2.6: Fuzzy-Bündel von Funktionen

Die Menge der erweiterten Funktionen ist im Bezug auf die Funktionen auf Fuzzy-Mengen nicht sehr reichhaltig.[97] Für die Analyse von ungenauen Daten mit Fuzzy-Regression ist es aber zentral, dass der Rückbezug zu den reellwertigen Funktionen erhalten bleibt. Da erweiterte Funktionen direkt aus gewöhnlichen Funktionen abgeleitet werden, sind sie besonders gut als Modellfunktionen für die Fuzzy-Regression geeignet.

Andererseits ist es ebenfalls denkbar, dass ungenaue Daten eine ungenaue Beschreibung einer Menge möglicher reellwertiger Approximationsfunktionen darstellen.

Definition 2.14 (Fuzzy-Bündel von Funktionen) Seien U, V beliebige Grundmengen. Dann ist eine Fuzzy-Menge $\widetilde{F}$ ein *Fuzzy-Bündel von Funktionen*, wenn gilt

$$\widetilde{F} \in \mathcal{F}(\mathrm{Map}(U,V)) = \mathcal{F}(V^U),$$

wobei $\mathrm{Map}(U,V) = V^U = \{f \mid f : U \longrightarrow V\}$. D.h. jeder Funktion $f \in \operatorname{supp} \widetilde{F}$ wird der Zugehörigkeitswert $\mu_{\widetilde{F}}(f)$ zugeordnet.

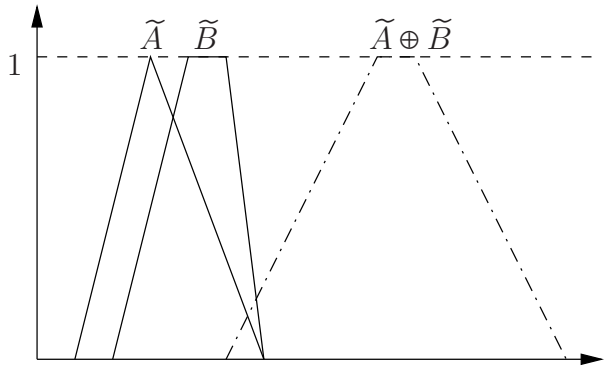
Sei $A \subset U \times V$. Für $(u,v) \in A$ kann dann definiert werden

$$M(u,v) = \{\mu_{\widetilde{F}}(f) \mid f \in \widetilde{F} \text{ mit } f(u) = v\}.$$

$M(u,v)$ ist im allgemeinen nicht einelementig wie Abbildung 2.6 zeigt. Wenn aus $\widetilde{F}$ eindeutige Informationen über die Zugehörigkeit von (u,v) zu A abgeleitet werden sollen, dann geht das nur unter Zuhilfenahme eines Aggregationsoperators, z.B. $\mu_{\widetilde{A}}(u,v) = \sup M(u,v)$.

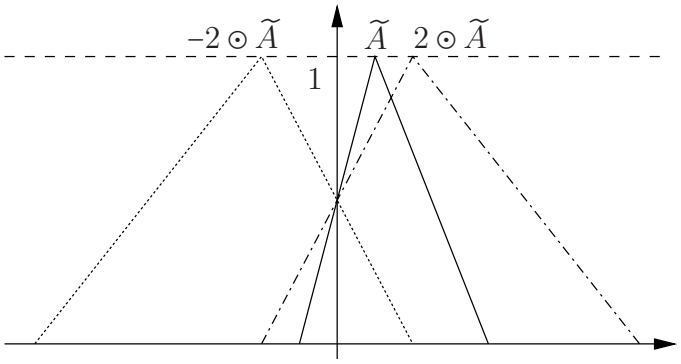
Bei der Erweiterung von Funktionen auf Fuzzy-Mengen entstehen dadurch, dass Fuzzy-Mengen eine positive Ausdehnung haben, arithmetische Besonderheiten gegenüber dem

[97] Vgl. Dubois und Prade [1980], S. 95ff.



(Eigene Darstellung in Anlehnung an Rommelfanger 1994, S. 41)

Abbildung 2.7: Addition der LR-Fuzzy-Intervalle $\widetilde{A}$ und $\widetilde{B}$



(Eigene Darstellung in Anlehnung an Rommelfanger 1994, S. 41)

Abbildung 2.8: Skalarmultiplikation der LR-Fuzzy-Zahl $\widetilde{A}$ mit $\lambda = 2$ bzw. $\lambda = -2$

reellwertigen Fall, bei dem Datenpunkte immer die Ausdehnung 0 haben. Insbesondere können die Linearitätseigenschaften, die bei der Berechnung der klassischen Kleinste Quadrate-Parameter und der Konstruktion von Tests von zentraler Bedeutung sind, nur mit Einschränkungen übertragen werden.

2.4.2 Lineare Strukturen über $F_{coc}^{nob}(\mathbb{R})$

Mit dem Erweiterungsprinzip in Satz 2.13 kann die gewöhnliche Addition und Multiplikation reeller Zahlen auf Fuzzy-Intervalle in $\mathbb{R}$ übertragen werden. Für Fuzzy-Intervalle in LR-Darstellung gelten die folgenden einfachen Formeln zur Berechnung der Addition und der Skalarmultiplikation.

Satz 2.15 Für $\widetilde{A} = (a^L, a^R; l_A, r_A)_{LR}$, $\widetilde{B} = (b^L, b^R; l_B, r_B)_{LR}$ und $\lambda \in \mathbb{R}$ gilt

$$\widetilde{A} \oplus \widetilde{B} = (a^L + b^L, a^R + b^R; l_A + l_B, r_A + r_B)_{LR}$$

sowie

$$\lambda \odot \widetilde{A} = \begin{cases} (\lambda a^L, \lambda a^R; |\lambda| l_A, |\lambda| r_A)_{LR} & \text{falls } \lambda \geq 0, \\ (\lambda a^R, \lambda a^L; |\lambda| r_A, |\lambda| l_A)_{LR} & \text{falls } \lambda < 0. \end{cases}$$

Allgemein gilt, dass stetige Funktionen bzw. stetige Operationen auch mittels der α-Schnitte direkt berechnet werden können, sofern die Zugehörigkeitsfunktionen der Fuzzy-Mengen von oben halbstetig sind. Demzufolge kann die Fuzzy-Arithmetik alternativ auch aus den Rechenregeln für Intervalle hergeleitet werden.

Satz 2.16 Seien $\widetilde{A}_i \in \mathcal{F}(\mathbb{R})$ $(1 \leq i \leq m)$ mit von oben halbstetigen Zugehörigkeitsfunktionen $\mu_{\widetilde{A}_i}$ und sei $f : \mathbb{R}^m \longrightarrow \mathbb{R}$ eine stetige Funktion, dann gilt

$$\forall\, \alpha \in [0,1]: \quad [f(\widetilde{A}_1, \ldots, \widetilde{A}_m)]^\alpha = f([\widetilde{A}_1]^\alpha, \ldots, [\widetilde{A}_m]^\alpha).$$

Ein Beweis dieses Satzes in sehr allgemeiner Form findet sich z.B. in [Kruse u. a., 1993].[98]

Nachfolgend nutzen wir den Zusammenhang zwischen Intervallmathematik und Fuzzy-Arithmetik aus und stellen stellvertretend nur die Rechenregeln für Fuzzy-Mengen zur Verfügung. Diese gelten analog auch für Intervalle in $\mathbb{R}^m$.[99]

Satz 2.17 (Rechenregeln) Seien $\widetilde{A}, \widetilde{B}, \widetilde{C} \in F_{coc}^{nob}(\mathbb{R}^m)$ Fuzzy-Mengen und $\lambda, \lambda' \in \mathbb{R}$ Skalare. Dann gelten die folgenden Rechenregeln

Assoziativität:

$$(\widetilde{A} \oplus \widetilde{B}) \oplus \widetilde{C} = \widetilde{A} \oplus (\widetilde{B} \oplus \widetilde{C}) \tag{2.11}$$

sowie

$$(\lambda \cdot \lambda') \odot \widetilde{A} = \lambda \odot (\lambda' \odot \widetilde{A}); \tag{2.12}$$

Kommutativität:

$$\widetilde{A} \oplus \widetilde{B} = \widetilde{B} \oplus \widetilde{A}; \tag{2.13}$$

[98] Vgl. ebd. S. 36ff. Dubois und Prade [1980] zufolge stammt der ursprüngliche Beweis des Satzes von Nguyen im folgenden Artikel: Nguyen, H.T. (1978): A note on the extension principle. J. Math. Anal. Appl., 64, No. 2, S. 369–380.

[99] Einführung und Beweise der Rechenregeln finden sich z.B. bei Dubois und Prade 1980, S. 49ff.

Existenz eines neutralen Elements der Addition:

$$\widetilde{A} \oplus \{\mathbf{0}\} = \widetilde{A}; \tag{2.14}$$

Existenz eines neutralen Elements der Skalarmultiplikation:

$$1 \odot \widetilde{A} = \widetilde{A}; \tag{2.15}$$

Verträglichkeit zwischen Addition und Skalarmultiplikation:

$$\lambda \odot (\widetilde{A} \oplus \widetilde{B}) = (\lambda \odot \widetilde{A}) \oplus (\lambda \odot \widetilde{B}). \tag{2.16}$$

Für die Skalarmultiplikation gilt im allgemeinen nur eine Richtung der Distributivität

$$(\lambda + \lambda') \odot \widetilde{A} \subseteq (\lambda \odot \widetilde{A}) \oplus (\lambda' \odot \widetilde{A}). \tag{2.17}$$

Der linke und der rechte Term in der Distributionsregel für die Skalarmultiplikation sind genau dann gleich, wenn zusätzlich $\widetilde{A}$ einelementig ist oder $\lambda\lambda' \geq 0$. Für diese letzte Regel ist die Konvexität von $\widetilde{A}$ tatsächlich notwendig. Alle anderen Regeln gelten auch für allgemeine Fuzzy-Vektoren.

Bemerkung 2.18 *Eine Inverse der Addition $\oplus$ existiert im allgemeinen nicht.*

Seien $A, B \subset \mathbb{R}^m$ Intervalle $A = [\mathbf{a}, \mathbf{b}], B = [\mathbf{c}, \mathbf{d}]$.

Gesucht ist $A \oplus B = [\mathbf{a}, \mathbf{b}] \oplus [\mathbf{c}, \mathbf{d}] = [\mathbf{a} + \mathbf{c}, \mathbf{b} + \mathbf{d}] \stackrel{!}{=} \{\mathbf{0}\}$.

Dies gilt nur dann, wenn $\mathbf{a} = -\mathbf{c}$ und $\mathbf{b} = -\mathbf{d}$. Damit A, B wohldefiniert sind, muss außerdem gelten $-\mathbf{d} \leq -\mathbf{c} = \mathbf{a} \leq \mathbf{b} = -\mathbf{d}$. Die Gleichung ist also nur dann erfüllbar, wenn A einelementig ist.

Daraus erhalten wir im Fall von Fuzzy-Mengen sofort: falls es ein $\alpha \in [0, 1)$ gibt, so dass $[\widetilde{A}]^\alpha$ nicht einelementig ist, dann existiert keine additive Inverse von $\widetilde{A}$. Insbesondere gilt

$$\widetilde{A} \oplus ((-1) \odot \widetilde{A}) \neq \{\mathbf{0}\}.$$

Dies ist ein Gegenbeispiel, welches zeigt, dass $(\lambda \odot \widetilde{A}) \oplus (\lambda' \odot \widetilde{A}) = (\lambda + \lambda') \odot \widetilde{A}$ nicht immer erfüllt ist.[100]

Die Menge der Fuzzy-Vektoren $F_{coc}^{nob}(\mathbb{R}^m)$ versehen mit der erweiterten Addition $\oplus$ und der erweiterten Skalarmultiplikation $\odot$ ist kein Vektorraum, da keine Inverse für $\oplus$

[100] Gleichermaßen gibt es keine Inverse für die Multiplikation. Diese ist für den Nachweis der Vektorraumstruktur aber unerheblich.

existiert und die Verträglichkeit mit der Skalarmultiplikation auf die positive Halbachse beschränkt ist, wie die Rechenregeln in Satz 2.17 zeigen. Infolgedessen wächst bei Verwendung der erweiterten Addition — und analog bei den anderen mit dem Zadehschen Erweiterungsprinzip übertragenen Rechenoperationen — die Unschärfe des Ergebnisses monoton mit jedem weiteren Fuzzy-Summanden. Einmal berücksichtigte Unschärfe ist somit irreversibel. Am Beispiel der Distributionsregel in Gleichung (2.17) wird deutlich, dass die Unschärfe des Ergebnisses im allgemeinen sogar abhängig von der Reihenfolge der Rechenschritte in der erweiterten Funktion ist.

Die Multiplikation zweier LR-Fuzzy-Intervalle ist zwar in $F_{coc}^{nob}(\mathbb{R})$ enthalten und somit wohldefiniert. Das Ergebnis ist aber nicht mehr vom selben LR-Typ wie die Fuzzy-Eingangswerte und kann daher insbesondere nicht mehr als geschlossene Formel ermittelt werden. Eine Darstellung ist nur auf den α-Niveaus möglich. Wir erhalten die folgenden Berechnungsformeln:

Satz 2.19 Seien L, R von oben halbstetige Referenzfunktionen und $\widetilde{A} = (a^L, a^R; l_A, r_A)_{LR}$, $\widetilde{B} = (b^L, b^R; l_B, r_B)_{LR}$. Wir notieren die Grenzen der α-Niveaus als $[\widetilde{A}]^\alpha = [a^L(\alpha), a^R(\alpha)]$ bzw. $[\widetilde{B}]^\alpha = [b^L(\alpha), b^R(\alpha)]$. Dann kann die Multiplikation gemäß Satz 2.16 für alle $\alpha \in [0,1]$ folgendermaßen berechnet werden.

Falls $b^L(\alpha) \geq 0$, dann

$$[a^L(\alpha), a^R(\alpha)] \odot [b^L(\alpha), b^R(\alpha)] = \begin{cases} [a^L(\alpha)b^L(\alpha), a^R(\alpha)b^R(\alpha)] & a^L(\alpha) \geq 0, \\ [a^L(\alpha)b^R(\alpha), a^R(\alpha)b^L(\alpha)] & a^R(\alpha) < 0, \\ [a^L(\alpha)b^R(\alpha), a^R(\alpha)b^R(\alpha)] & a^L(\alpha) < 0 \leq a^R(\alpha). \end{cases}$$

Falls $b^R(\alpha) < 0$, dann

$$[a^L(\alpha), a^R(\alpha)] \odot [b^L(\alpha), b^R(\alpha)] = \begin{cases} [a^R(\alpha)b^L(\alpha), a^L(\alpha)b^R(\alpha)] & a^L(\alpha) \geq 0, \\ [a^R(\alpha)b^R(\alpha), a^L(\alpha)b^L(\alpha)] & a^R(\alpha) < 0, \\ [a^R(\alpha)b^L(\alpha), a^L(\alpha)b^L(\alpha)] & a^L(\alpha) < 0 \leq a^R(\alpha). \end{cases}$$

Den Fall $b^L(\alpha) < 0 \leq b^R(\alpha)$ benötigen wir im Folgenden nicht. Daher stellen wir die Formeln hier nicht zur Verfügung.

Für die Multiplikation zweier Fuzzy-Mengen in $F_{coc}^{nob}(\mathbb{R}^m)$ gilt ebenfalls das Assoziativ- und das Kommutativgesetz. Das Distributivgesetz gilt im Allgemeinen nur in eine Rich-

tung,[101] d.h. für $\widetilde{A}, \widetilde{B}, \widetilde{C} \in F_{coc}^{nob}(\mathbb{R}^m)$ gilt

$$\widetilde{A} \odot (\widetilde{B} \oplus \widetilde{C}) \subset (\widetilde{A} \odot \widetilde{B}) \oplus (\widetilde{A} \odot \widetilde{C}). \tag{2.18}$$

Damit liegen nun alle Formeln vor, mit denen lineare Funktionen $\sum_{j=1}^{m} a_j x_j$ mit $(a_1, \ldots, a_m), (x_1, \ldots, x_m) \in \mathbb{R}^m$ auf *LR*-Fuzzy-Mengen übertragen werden können. Wir wollen zur sprachlichen Vereinfachung alle möglichen Varianten von erweiterten linearen Funktionen unter dem Begriff der Fuzzy-linearen Funktion zusammenfassen. Dabei soll die Bezeichnung „Fuzzy-linear" die Einschränkung gegenüber der Linearität im reellwertigen Fall hervorheben. Die Eigenschaften der Fuzzy-linearen Funktionen werden in Abschnitt *3.1* vorgestellt. Dort finden sich auch beispielhafte Abbildungen.

Definition 2.20 (Fuzzy-lineare Funktionen) Sei $A \subset \mathbb{R}^m$ und $f : \mathbb{R}^m \times A \longrightarrow \mathbb{R}$, $(\mathbf{u}, \mathbf{a}) \longmapsto f_{\mathbf{a}}(\mathbf{u}) = \mathbf{u}^T\mathbf{a}$ eine Schar linearer Funktionen.

Wir nennen jede der daraus mit dem Erweiterungsprinzip gewonnen Funktionen

$$f(\cdot, \widetilde{A}) : \mathbb{R}^m \longrightarrow F_{coc}^{nob}(\mathbb{R}), \mathbf{u} \longmapsto \bigoplus_{j=1}^{m} \widetilde{A}_j \odot u_j \qquad \text{für } \widetilde{A} \in A \cap F_{coc}^{nob}(\mathbb{R})^m,$$

$$f(\cdot, \mathbf{a}) : F_{coc}^{nob}(\mathbb{R})^m \longrightarrow F_{coc}^{nob}(\mathbb{R}), (\widetilde{U}_1, \ldots, \widetilde{U}_m) \longmapsto \bigoplus_{j=1}^{m} a_j \odot \widetilde{U}_j \quad \text{für } \mathbf{a} \in A$$

$$f(\cdot, \widetilde{A}) : F_{coc}^{nob}(\mathbb{R})^m \longrightarrow F_{coc}^{nob}(\mathbb{R}), (\widetilde{U}_1, \ldots, \widetilde{U}_m) \longmapsto \bigoplus_{j=1}^{m} \widetilde{A}_j \odot \widetilde{U}_j \quad \text{für } \widetilde{A} \in A \cap F_{coc}^{nob}(\mathbb{R})^m$$

eine *Fuzzy-lineare Funktion*. Wir symbolisieren die Funktionstypen durch $f[u, \widetilde{A}], f[\widetilde{U}, a]$ bzw. $f[\widetilde{U}, \widetilde{A}]$.

Die Funktionsschar $\mathbf{u}^T\mathbf{a}$ mit $\mathbf{a} \in A$ kann bei Bedarf auch auf interaktive Fuzzy-Vektoren über $\mathbb{R}^m$ erweitert werden. Diese sind für die Untersuchung der Fuzzy-Regression aber eher nachrangig. Für die Interpretation von interaktiven Fuzzy-Parametern liegen bislang nur wenige Ansätze vor.[102] Interaktive Fuzzy-Vektoren auf der anderen Seite liegen typischerweise als symmetrische, bohnenförmige Fuzzy-Daten vor. Wenn Fuzzy-Inputs als interaktive Fuzzy-Vektoren dargestellt werden, sind sie also im Vergleich zu nicht-interaktiven Fuzzy-Vektoren bei der Datenmodellierung bereits stark geglättet. Zur Vereinfachung werden wir uns in dieser Arbeit auf den Fall nicht-interaktiver Parameter und nicht-interaktiver Daten beschränken.

[101] Vgl. Kruse u. a. [1993], S. 39.

[102] Einige Ansätze zur Fuzzy-Regression beziehen sich auf Modelle mit interaktiven Parametern, so z.B. [Celmiņš, 1987b; Chang und Lee, 1994a]

2.5 Von der Fehlerschätzung zur Fehlerbewertung

Am Ende der kritischen Revision des ökonometrischen Fehlermodells in Abschnitt *2.3* stand der Vorschlag, zu einer alternativen Fehlermodellierung überzugehen, bei der die Möglichkeiten einer Fuzzy-Modellierung von ungenauen Werten ausgenutzt werden. Damit wird ein Perspektivwechsel vorgenommen. Während die Fehlereinflüsse im Schätzmodell mit Fehlern in den Variablen als Struktureigenschaft aufgefasst werden, werden sie bei der Fuzzy-Modellierung als Eigenschaft der Daten selbst betrachtet. Deren Ungenauigkeit ist dabei für jeden Beobachtungswert auf dem aktuellen Stand des Wissens zu bewerten. Um diesen Ansatz von den Modellen mit Fehlern in den Variablen abzugrenzen, soll er im Folgenden als Regressionsmodell mit *Fehlern in den Daten* bezeichnet werden.

Zunächst wird in Abschnitt *2.5.1* der Begriff von Fuzzy-Daten konkretisiert und es werden zwei wesentliche semantischen Zugänge vorgestellt. In Abschnitt *2.5.2* wird schließlich das Konzept der Fuzzy-Merkmalswerte eingeführt, mit dem die Abschätzung über die möglich Lage eines Einzelwertes abgebildet werden kann, der mit Fehlern beobachtet wird. Dabei ist die Abschätzung selber als Beobachtungswert zu betrachten.

2.5.1 Modellierung von Fuzzy-Daten

Gegenüber beliebigen Zahlen oder Werten zeichnen *Daten* sich dadurch aus, dass sie Informationen über einen gewissen, uns interessierenden, Gegenstand oder Zustand tragen. Zu dem schieren Wert tritt also ein semantischer Kontext hinzu. Üblicherweise wird ein einzelnes Datum als Realisierung einer Variablen über einer geeigneten Grundmenge betrachtet. Information trägt das Datum nur dann, wenn für den interessierenden Gegenstand mehr als ein abbildbarer Zustand existiert.[103] Allgemein ist ein *Fuzzy-Datum* $\widetilde{Z}$ die Realisation einer Variablen $U : \Omega \longrightarrow \mathcal{F}(\Upsilon)$ mit Fuzzy-Werten über einem Wertebereich Υ[104], so dass für ein $\omega \in \Omega$ gilt $U(\omega) = \widetilde{Z}$. Dabei muss man sich die Variable U eingebettet in einen Interpretationskontext vorstellen. Meistens liegt der Modellierung von Fuzzy-Daten die Vorstellung von menschlicher Wahrnehmung und Verständigung zugrunde, die sich häufig auf uneindeutige und ungefähre Begriffe bezieht. Mit Bezug auf die Abbildung von verbalen Begriffen spricht man daher oft auch von *linguistischen Variablen*.

Für die Messung von reellwertigen Daten werden verschiedene Typen von Skalen unterschieden. In erster Linie sind dabei ordinale und metrische Skalen zu unterscheiden. Allerdings werden auch kategoriale oder nominale Skalen häufig mit ganzen Zahlen wiedergegeben. Mit dem Skalentyp wird eine Unterscheidung getroffen, welche Operationen

[103] Vgl. Bandemer 1997, S. 24.

[104] Entsprechend der Einführung wird angenommen, dass Υ mehr als ein Element enthält.

grundsätzlich auf den Datenwerten erlaubt sind, d.h. Vergleichsoperationen auf einer ordinalen Skala, arithmetische Operationen auf einer metrischen Skala etc. Dabei verschwinden weitere wichtige Aspekte für die Auswertung der Daten, wie z.B. Verlässlichkeit und Genauigkeit der Messung, hinter den Beobachtungswerten. Sie sind bei der Konstruktion der Analysemodelle und der Interpretation der Ergebnisse gesondert abzuwägen. Fuzzy-Daten basieren auf klassischen reellwertigen Skalen und erweitern diese durch die Möglichkeit, bei Unschärfen in den Daten[105] graduelle Übergänge für die Zugehörigkeit zum Messwert anzugeben. Dazu werden alle Werte der zugrunde liegenden Skala danach bewertet, zu welchem Grad sie beim aktuellen Wissensstand zum jeweiligen Beobachtungswert zugehörig sind. Die Modellierung zielt darauf hin, dass der aktuelle Wissensstand über die Ungenauigkeit oder Uneindeutigkeit der Daten so gut und so korrekt wie möglich abgebildet wird. Dabei kann der aktuelle Wissensstand sich aus allen möglichen Quellen speisen, wie z.B. vorliegende Daten, ergänzende und alternative Statistiken oder Expertenmeinungen.

Die graduelle Zugehörigkeit zu einer Menge ist ein sehr offenes Konzept zur Abbildung von Datenunschärfen, das vielseitig ausgelegt und eingesetzt werden kann. In der Literatur zur Fuzzy-Mengen-Theorie gibt es infolgedessen eine kontinuierliche Auseinandersetzung mit der Frage, welche Interpretationen von Datenunschärfe mit Fuzzy-Mengen beschrieben werden können und welche Bedeutung dabei die Zugehörigkeitsgrade annehmen.[106] Es ist daher unerlässlich, dass die Semantik von Fuzzy-Daten explizit angegeben wird, wenn diese für weitergehende Analysen verwendet werden sollen. Denn eine sinnvolle Interpretation von Analyseergebnissen ist nur unter der Bedingung möglich, dass der Aussagegehalt der Fuzzy-Daten definiert und die angewendete Methodik mit dieser semantischen Bedeutung vereinbar ist.

In Abschnitt *2.1* wurde auf Seite 19 bereits angemerkt, dass die Qualität von Daten sich hauptsächlich im Bezug auf die Adäquation und auf die Genauigkeit der Messung begründet. Beim Adäquationsproblem stehen unter anderem Fragen der Uneindeutigkeit der statistischen Begriffe sowie der Homogenität und Vergleichbarkeit der Messwerte im Mittelpunkt. Demgegenüber beziehen sich Genauigkeitsprobleme direkt auf die Fehler in den Daten. Diese stellen sich als eindeutige Begriffe dar, die aufgrund eines eingeschränkten Wissensstandes ungenau oder uneindeutig erscheinen. In Anlehnung an Kruse u.a.

[105] Die Begriffe, mit denen in dieser Arbeit Unschärfen in den Daten differenziert werden, wurden zu Beginn von Kapitel *2* eingeführt, s. S. 15.

[106] Arbeiten, die Vorschläge zur Modellierung von Fuzzy-Mengen machen, sind z.B. [Zadeh, 1965; Dubois und Prade, 1980; Bandemer und Gottwald, 1993; Zimmermann, 1996]. Im Sammelband [Dubois und Prade, 2000] werden die Grundlagen der Fuzzy-Mengen-Theorie auf dem aktuellen Stand der Forschung vorgestellt. Dort nimmt die semantische Differenzierung zwischen unterschiedlichen Arten von Fuzzy-Mengen und ihre Eignung für bestimmte Anwendungsgebiete großen Raum ein. In [Avenhaus und Seising, 1999] wird die Rolle der Fuzzy-Mengen-Theorie im Verhältnis zur Wahrscheinlichkeitstheorie erörtert.

[1993] soll daher in dieser Arbeit zwischen den Interpretationskonzepten der physikalischen und der epistemischen Fuzziness unterschieden werden. Diese beiden Konzepte werden in den weiteren Untersuchungen zugrunde gelegt, um daran zu verdeutlichen, welche Konsequenzen die unterschiedlichen Perspektiven für die Regressionsanalyse bei Fuzzy-Daten haben. Krätschmer [2001a] hat in seiner Habilitationsschrift ein allgemeines Messmodell für Fuzzy-Daten entwickelt. Er bezieht sich aber in seiner Arbeit hauptsächlich auf eine physikalische Interpretation der Fuzziness.

Fuzzy-Daten in der physikalischen Interpretation erhält man, wenn Werte von Begriffen mit einer vagen Definition spezifiziert werden, d.h. wenn durch die Definition der Zielgröße bzw. des statistischen Begriffs eine Uneindeutigkeit gegeben ist, die sich auch durch Einholen von besserer Information nicht weiter ausräumen lässt. Häufig führen Adäquationsprobleme zu solchen vagen statistischen Begriffen. Ein einfaches anschauliches Beispiel ist die Beschreibung aller Oberflächenpunkte im Gesicht, die zur Nase gehören. Ein bekanntes Beispiel für eine vage Begriffsdefinition in der Wirtschaftsstatistik ist etwa das Bruttosozialprodukt[107]. Physikalische Fuzzy-Daten können als genaue Abbildung eines uneindeutigen, vagen Objektes oder Zustandes verstanden werden. Die Fuzzy-Menge wird hier also als unveränderliches mathematisches Objekt aufgefasst.[108]

Im Gegensatz dazu liegen *Fuzzy-Daten in der epistemischen Interpretation* vor, wenn ein eindeutiger, reellwertiger Wert für die Beobachtung existiert, dieser aber aufgrund von eingeschränkten Informationen oder fehlerhaften Messverfahren nur ungenau beschrieben werden kann. Die Fuzzy-Menge wird dann als Abschätzung für die mögliche Lage des verborgenen, eindeutigen Wertes spezifiziert. Dabei erfolgt die Abschätzung auf der Basis des aktuell verfügbaren Wissens, d.h. durch Einholen von besseren Informationen könnte die Qualität der Abschätzung erhöht werden. Ein einfaches anschauliches Beispiel stellt z.B. die sprachliche Beschreibung dar, eine Person sei „um die 35 Jahre alt“. Es ist klar, dass die betreffenden Person ein eindeutiges Alter hat, dieses kann aber ohne weitere Nachforschungen nur mit einer Menge von mehreren plausiblen Werten beschrieben werden. Ein Beispiel aus der Wirtschaftsstatistik für einen eindeutigen Wert, der in der Regel nur ungenau erfasst werden kann, ist etwa das Netto-Jahreseinkommen. In der Regel können nur einzelne individuelle Einkommensbestandteile oder das Jahreseinkommen in einer anderen Abgrenzung erfasst werden, das Netto-Einkommen kann hingegen nur annähernd bestimmt werden. Folglich ist ein Fuzzy-Datum in der epistemischen Interpretation kein unveränderliches mathematisches Objekt, sondern stellt eine kontextabhängige, ggf. subjektiv festgelegte, Begrenzung und Zugehörigkeitsbewertung der Grundmenge von möglichen Werten dar.

[107] Pflegetätigkeiten werden z.B. nur dann zum Bruttosozialprodukt gerechnet, wenn sie über den Markt vermittelt angeboten werden, familiäre Pflege wird hingegen nicht mitgerechnet.

[108] Vgl. Borgelt u. a. 1999, S. 371.

Aus den Beschreibungen wird deutlich, dass physikalische Fuzzy-Daten sich eher wie Punkte darstellen, mit denen ein uneindeutiger Zustand als Gesamtheit abgebildet wird, epistemische Fuzzy-Daten stellen sich hingegen eher als Intervallabschätzungen dar, die eine Menge möglicher Werte erfassen. Im Bezug darauf sprechen wir vom *Punkt-* und vom *Mengencharakter* von Fuzzy-Daten.

Zusammenfassend kann festgestellt werden, dass eine Modellierung von Fuzzy-Daten für Fehler in den Daten grundsätzlich dem epistemischen Interpretationskonzept genügt. Hingegen können Fuzzy-Modellierungen von Adäquationsfehlern idealtypisch als physikalische Fuzzy-Daten betrachtet werden. In der Praxis findet man allerdings häufig eine Vermischung von Adäquationsfehlern und Datenungenauigkeiten vor, so dass eine eindeutige Zuordnung zum physikalischen oder epistemischen Interpretationskonzept nicht möglich ist, weil keiner der Einflüsse zu vernachlässigen ist. Bei der Modellierung von Fuzzy-Merkmalswerten entspricht dies einem Changieren zwischen der physikalischen und der epistemischen Interpretationsperspektive. Dies ist für die Konstruktion empirischer Fuzzy-Daten nachrangig, wird aber bei der Approximation einer Regressionsgleichung dann zum Problem, wenn die jeweilige Semantik besondere Anforderungen an das arithmetische Kalkül oder den Optimierungsansatz stellt. Die Eigenschaften der physikalischen und der epistemischen Modellierung können bei solchen vagen statistischen Begriffen, deren Messung mit Fehlern behaftet ist, als Extrempunkte betrachtet werden. Während bei mehreren Werten für ein Merkmal in der physikalischen Fuzzy-Modellierung von einer Homogenität in der Qualität der Abschätzungen ausgegangen werden kann, ist eine Homogenität in der Qualität der Abschätzungen bei der epistemischen Fuzzy-Modellierung nicht mehr ohne Weiteres sichergestellt.[109] Eine semantische Interpretation kann daher auch im Bezug auf die Eigenschaften der beiden Extreme hergeleitet werden.

Häufig werden epistemische Fuzzy-Daten auch als „Possibility-Verteilungen“ bezeichnet, weil sie ähnlich wie Wahrscheinlichkeitsverteilungen ebenfalls eine Quantifizierung der Möglichkeit von Zuständen oder Ereignissen sind. Tatsächlich gibt es eine ausgearbeitete Theorie der Possibilitäts-Verteilungen. Diese unterscheidet sich semantisch aber deutlich vom epistemischen Interpretationskonzept. Während epistemische Fuzzy-Daten für jedes Zugehörigkeitsniveau eine Intervallabschätzung abgeben, trifft eine Possibility-Verteilung im Sinne eines Verteilungskonzeptes eine Aussage über Auftretenshäufigkeiten, wenn es eine Ambiguität des Auftretens gibt. Damit sollen Verteilungen für Situationen beschrieben werden, bei denen nicht eindeutig festgestellt werden kann, ob ein Ereignis eingetreten ist oder ob es nicht eingetreten ist, bei denen also die sog. „Regel des aus-

[109] Eine alternative Beschreibung könnte sein, dass die Ränder von physikalischen Fuzzy-Daten ebenfalls auf der verwendeten Skala in Bezug gesetzt werden können, während die Ränder von epistemischen Fuzzy-Daten allenfalls in einen ordinalen Bezug gesetzt werden können.

geschlossenen Dritten“ nicht gilt. Aufbauend auf den Possibility-Verteilungen wurde eine ganze Verteilungstheorie entwickelt, die Possibility-Theorie.[110]

Mit den klassischen Skalentypen erhalten wir somit bei Berücksichtigung der semantischen Konzepte für Fuzzy-Daten die folgenden Skalentypen für Fuzzy-Daten. Für Fuzzy-Daten mit physikalischer Interpretation können kategoriale, ordinale und metrische Skalen problemlos übertragen werden. Bei fuzzifizierten kategorialen Skalen kann es im Unterschied zum klassischen Fall Überschneidungen zwischen den Kategorien geben, wobei einzelne Elemente Teilzugehörigkeiten in mehr als einer Kategorie haben, als Operationen sind nur Schlussweisen der Fuzzy-Interferenz zulässig. Fuzzifizierte ordinale Skalen sind nur dann wohldefiniert, wenn eine geeignete Fuzzy-Präferenzrelation angegeben wird, die wieder eine ordinale Struktur etabliert. Für fuzzifizierte metrische Skalen können die arithmetischen Operationen mit dem Erweiterungsprinzip übertragen werden und gegenüber dem klassischen Fall gelten u.a. die Einschränkungen, die in Abschnitt *2.4.2* diskutiert wurden. Ebenso können epistemische Fuzzy-Daten auf der Basis von kategorialen, ordinalen und metrischen Skalen definiert werden.[111] Im Unterschied zu physikalischen Fuzzy-Skalen, deren Elemente als Objekte verstanden werden können, stellen die Elemente von epistemischen Fuzzy-Skalen Fuzzy-Sicherheitsmengen bzw. Fuzzy-Mengen-Abschätzungen für einen eindeutigen unbekannten Wert dar.[112]

In der Praxis wird ein Fuzzy-Datum $\widetilde{Z}$ über den reellen Zahlen $\mathbb{R}$ meistens durch die folgenden Angaben spezifiziert: festzulegen sind der Kern als Punkt bzw. Intervall der sicher möglichen Punkte, der Träger als Intervall der Zahlen, die zu einem positiven Grad noch als zum Datum zugehörig akzeptiert werden sollen sowie die Geschwindigkeit der monotonen Abnahme der Zugehörigkeiten vom Kern zum Rand des Trägers, die durch die Form der Referenzfunktion modelliert wird.[113] Die Zugehörigkeitsfunktion wird auf der Basis der vorliegenden Gegebenheiten und Kenntnisse festgelegt. Die Spezifizierungsvorschrift beinhaltet somit eine subjektive Bewertung des unscharfen Datums, die

[110] Vgl. [Dubois u. a., 2000]. Folgt man Borgelt u. a. [1999], dann können eher physikalische Fuzzy-Mengen in Possibility-Verteilungen „übersetzt“ werden. Bei der Fuzzy-Menge für die Nase würde die zugehörige Possibility-Verteilung beschreiben, wie plausibel es ist, dass in der Aussage „er hat einen Fleck auf der Nase“ mit der Angabe „auf der Nase“ bestimmte Raumpunkte im Gesicht eines Menschen beschrieben sind (S. 372).

[111] Die Übertragung leuchtet bei metrischen Skalen sofort ein, wobei Fuzzy-Intervall-Abschätzungen etabliert werden. Eine kategoriale, epistemische Fuzzy-Skala könnte beispielsweise sinnvoll sein, um Klassifikationsfehler abzubilden, die durch fehlerhafte Messwerte induziert sind. Eine ordinale, epistemische Fuzzy-Skala könnte z.B. genutzt werden, um Probleme in der zeitlichen Abgrenzung bei Stromgrößen darzustellen.

[112] Die Unterscheidung ist z.B. dafür relevant, welche Fuzzy-Präferenzrelation jeweils für die Fuzzy-ordinalen Skalen geeignet erscheint. Während für erstere Variante eher eine Defuzzifizierung herangezogen werden kann, erscheint bei der zweiteren Variante vielleicht eher ein ε-Vergleich angemessen. Die tatsächliche Festlegung kann allerdings nur unter Berücksichtigung der konkreten Modellierung erfolgen.

[113] Vgl. Bandemer 1997, S. 67.

die Ungenauigkeit oder Vagheit in der konkreten Situation ausdrückt. Die auf Basis der subjektiven Bewertungen spezifizierte Zugehörigkeitsfunktion stellt nur eine von vielen möglichen Vereinfachungen der Situation dar. Der modellierte Fuzzy-Beobachtungswert ist infolgedessen nicht eindeutig. Häufig wird aber die Eindeutigkeit eines Messwertes als Nachweis für seine Objektivität bzw. für seine Aussagekraft über die reale Situation wahrgenommen. In diesem Sinne erscheinen reellwertige Messwerte als objektiver als Fuzzy-Beobachtungswerte. Dabei wird übersehen, dass auch reellwertige Messwerte subjektiven Bearbeitungs- und Vereinfachungsprozessen unterworfen sind, die zu einer Veränderung des Messergebnisses führen können. Ganz im Gegenteil besteht ein wesentlicher Vorteil einer Fuzzy-Modellierung im Vergleich zu einer Darstellung des Bereichs möglicher Werte in Form einer Intervallabschätzung gerade darin, dass man von dem Dilemma befreit wird, eine scharfe Grenze zwischen möglichen und unmöglichen Werten zu spezifizieren. Statt einer scharfen Grenze wird ein allmählicher Übergang zwischen sicher möglichen und sicher unmöglichen Werten eingesetzt, den man sich wie die Einstellung des Kontrastes bei einem Schwarz-Weiß-Foto vorstellen kann, bei dem Weiß mit 0, Schwarz mit 1 und dazwischenliegende Zugehörigkeitswerte mit abgestuften Grautönen identifiziert werden. Veränderungen der Zugehörigkeitsfunktion unter Wahrung der lokalen Monotonie sind demnach vergleichbar mit Kontraständerungen bei Grautonbildern, die in der Regel den semantischen Gehalt nicht wesentlich ändern.[114]

Die subjektive Modellierung und die Tatsache, dass Vagheit und Ungenauigkeit von Messverfahren abgebildet werden sollen, bedeutet nicht, dass zur Spezifikation von Fuzzy-Daten weniger Informationen benötigt werden als bei der Spezifikation bzw. Messung von scharfen Daten. Ganz im Gegenteil verhilft eine Fuzzy-Modellierung dazu, dass zusätzliches Wissen über die individuellen und konkreten Erhebungsbedingungen quantitativ abgebildet wird und bei der kalkulatorischen Datenanalyse mit verarbeitet werden kann. Sie kann damit bessere Ergebnisse liefern als eine Analyse scharfer Daten, bei der die abschließende Interpretation von unausgesprochenen, subjektiven Daumenregeln überformt wird.

Damit eine Regressionsanalyse bei Fuzzy-Daten durchgeführt werden kann, müssen die arithmetischen Operationen definiert sein. Wir beziehen uns daher in dieser Arbeit ausschließlich auf metrische Fuzzy-Skalen mit Daten aus $F_{coc}^{nob}(\mathbb{R})$.

Definition 2.21 (Skalenniveau der Fuzzy-Daten für die Fuzzy-Regression)
Im Folgenden werden nur noch Fuzzy-Messwerte bzw. Fuzzy-Beobachtungen verwendet, für die die Ergebnisse der erweiterten Addition, des erweiterten Skalarproduktes und der

[114] Vgl. Bandemer 1999, S. 253.

erweiterten Multiplikation wohldefiniert sind. D.h. für die Mess-Skala $\mathcal{S} \subset \mathcal{F}(\mathbb{R})$ für eine Variable Z in der Modellfunktion gelte, dass

$$\mathcal{S} \subset F_{coc}^{nob}(\mathbb{R})$$

und für je zwei Skalenwerte $\widetilde{U}, \widetilde{V} \in \mathcal{S}$ und $\lambda \in \mathbb{R}$ sei erfüllt

$$\begin{aligned}
&\widetilde{U} \oplus \widetilde{V} \in \mathcal{S}, \\
&\lambda \odot \widetilde{U} \in \mathcal{S} \quad \text{oder} \quad -\lambda \odot \widetilde{U} \in \mathcal{S}, \\
&\widetilde{U} \odot \widetilde{V} \in F_{coc}^{nob}(\mathbb{R}).
\end{aligned}$$

Durch die Definition sollen insbesondere kategoriale Fuzzy-Daten, d.h. linguistische Bewertungen der Art „`niedrig`“, „`mittel`“, „`hoch`“, im Weiteren als mögliche Merkmale ausgeschlossen werden. Diese sind mit der Einbeziehung von unscharfen Dummy-Variablen in die Regressionsgleichung vergleichbar und wären im Hinblick auf mögliche Interpretationen und ihre Aussagekraft gesondert zu betrachten. Wir berücksichtigen daher auch kategoriale Klassifikationen, wie sie z.B. in der Beschäftigtenstatistik die Klassifikationen der Berufe oder der Wirtschaftszweige darstellen, und deren Messprobleme nicht weiter.

2.5.2 Fuzzy-Merkmalswerte als Alternative

In diesem Abschnitt werden die allgemeinen Anforderungen für die Konstruktion von Fuzzy-Merkmalswerten formuliert, mit denen Fehler in den Daten abgebildet werden können. Dies dient zur Vorbereitung auf die konstruktive Modellierung von Fuzzy-Daten am Beispiel der Beschäftigtenstatistik in Kapitel *4*.

Man kann sich zunächst exemplarisch vorstellen, dass für eine beliebige Modellvariable fehlerhafte Werte anstelle der „wahren“ Werte $Z^* : \Omega \longrightarrow \mathbb{R}$ beobachtet werden. Für die Konstruktion wird ein einzelner, konkreter Beobachtungswert $z^* = Z^*(\omega)$ herausgegriffen. Bei der Modellierung eines epistemischen Fuzzy-Datums $\widetilde{Z}$, das den Fehlereinfluss in z^* abbildet, ist für jedes Zugehörigkeitsniveau $\alpha \in [0,1]$ das Intervall der Werte zu spezifizieren, die aufgrund der fehlerhaften Messung der Daten gleichermaßen möglich sind. Im Unterschied zum ökonometrischen Fehlermodell wird hierbei also der Fehler nicht gesondert spezifiziert. Die Fuzzy-Intervall-Abschätzung für die Fehler in den Daten zielt im Gegenteil darauf hin, alle möglichen Werte für z unter den aktuell verfügbaren Informationen so scharf wie möglich einzugrenzen. Infolgedessen wird für die Fuzzy-Modellierung keine Unterscheidung in „wahren“ Wert und Messfehler benötigt. Das Fuzzy-Datum ist umgekehrt aber im allgemeinen auch nicht dazu geeignet, einen eindeutigen Wert zu spezifizieren, der den „wahren“ Wert des Modells repräsentiert.

Da bei realen Datenerhebungen bei Fehlern in den Daten der „wahre“ Wert unbekannt ist, soll im Folgenden für alle epistemischen Fuzzy-Daten die Annahme getroffen werden, dass diese den „wahren“ Wert tatsächlich enthalten.

Definition 2.22 Seien für $1 \leq i \leq n$ die Fuzzy-Daten $\widetilde{Z}_i$ ein Messergebnis, das die fehlerhaften Beobachtungen für ein Merkmal abbildet. Seien ferner z_i^* die zu $\widetilde{Z}_i$ gehörigen „wahren“ Werte. Dann ist das Fuzzy-Datum $\widetilde{Z}_i$ *günstig* für die i-te Beobachtung genau dann, wenn gilt

$$z_i^* \in \operatorname{supp}(\widetilde{Z}_i).$$

Im günstigen Fall gilt also für den „wahren“ Wert z_i^*, dass dieser tatsächlich von der Fuzzy-Beschreibung $\widetilde{Z}_i$ eingehüllt wird.

In die Konstruktion der Fuzzy-Merkmalswerte bei Fehlern in den Daten können grundsätzlich alle Informationen auf dem aktuellen Stand des Wissens einbezogen werden. Voraussetzung für die Modellierung von Informationen ist allerdings, dass sie für die Quantifizierung von Zugehörigkeitsgraden über dem Wertebereich herangezogen werden können. Einer Abbildung von qualitativen Einschätzungen über die Fehlereinflüsse sind durch die Anforderung der Quantifizierbarkeit Grenzen gesetzt.

Wenn die Modellierung von Fuzzy-Merkmalsvariablen für Fehler in den Daten ex post, d.h. nach Abschluss des Erhebungsverfahrens, vorgenommen wird, ergibt sich eine paradoxe Situation dadurch, dass der vorliegende, fehlerhafte Messwert x die sicherste Information über den unbekannten „wahren“ Wert z darstellt. Es ist also ausgehend vom fehlerhaften Wert x die Information über die mögliche Lage des „wahren“ Wertes z zu erschließen. Es liegt daher aus Gründen der Risikovermeidung nahe, für x eine Zugehörigkeit von 1 anzunehmen, sofern nicht starke Gründe dafür vorliegen, den Kern des Fuzzy-Datums so zu verschieben, dass er x nicht enthält. Andernfalls besteht die Gefahr, die verschmutzte Information aus dem falschen Wert x damit noch aktiv zu verschlimmern. Überdies ginge bei einer solchermaßen transformierten Modellierung die Information über den vorliegenden Wert x für weitere Betrachtungen vollständig verloren. Diese Herangehensweise unterscheidet sich diametral von der Modellierung der Fehlereinflüsse durch eine Zufallsvariable, in der die Abweichung zwischen Beobachtung und „wahrem“ Wert als reellwertige Zahl erscheint, die nach Belieben relativ zum „wahren“ Wert oder relativ zur fehlerhaften Beobachtung betrachtet werden kann. Dabei kann unter den Bedingungen des Strukturmodells der unbekannte „wahre“ Wert mit dem Erwartungswert identifiziert werden, so dass die beobachteten, fehlerhaften Werte als Streuungsumgebung

zum unbekannten „wahren“ Wert aufgefasst werden.[115] Wesentliche Voraussetzung dafür ist allerdings, dass das Verteilungsmodell zutrifft.

Bei einer Fuzzy-Modellierung von Messungenauigkeiten besteht aufgrund der normalisierten Darstellung der Fuzzy-Zahlen die Gefahr, dass diese als eine Gewichtungsfunktion aufgefasst werden, die sich ähnlich wie eine Wahrscheinlichkeitsdichte verhält. Dabei liegt ein Denkfehler vor, denn es wird übersehen, welche Kompensationseffekte bei der Verknüpfung von Wahrscheinlichkeitsdichten entstehen: durch die Kumulierung der Punkthäufigkeiten aus den Einzelverteilungen in den relativen Häufigkeiten der gemeinsamen Verteilung werden nämlich Verstärkungs- oder Abschwächungseffekte bei den Auftretenswahrscheinlichkeiten induziert. Insbesondere kann die Gesamtverteilung eine höhere Spezifität aufweisen als die Einzelverteilungen. Diese Kompensation gibt es bei der Verknüpfung von Fuzzy-Zahlen nicht. Am Beispiel der Addition von stochastisch unabhängigen Zufallsvariablen bzw. von Fuzzy-Mengen kann gut verdeutlicht werden, dass zwischen den beiden Betrachtungsweisen ein fundamentaler Unterschied besteht.

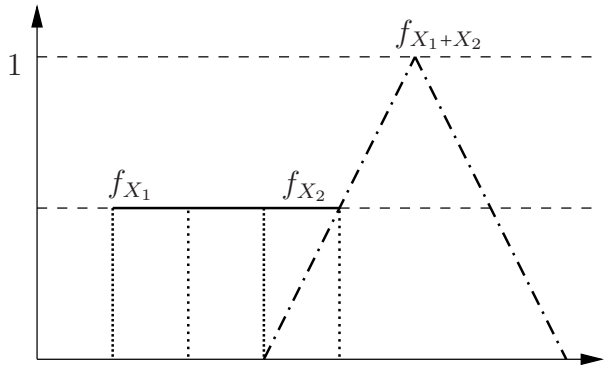
Wir betrachten dazu zwei gleichverteilte Zufallsvariablen $X_1 \sim G([1,3]), X_2 \sim G([2,4])$ mit den zugehörigen Wahrscheinlichkeitsdichten $f_{X_1} = \frac{1}{2}1_{[1,3]}, f_{X_2} = \frac{1}{2}1_{[2,4]}$. Falls X_1 und X_2 außerdem stochastisch unabhängig sind, kann die Verteilung von $X_1 + X_2$ als Faltung der zugehörigen Wahrscheinlichkeitsdichten berechnet werden, d.h.

$$\begin{aligned} f_{X_1+X_2}(t) &= (f_{X_1} * f_{X_2})(t) \\ &= \frac{1}{2}\int 1_{[1,3]}(\tau)1_{[2,4]}(t-\tau)\,d\tau \\ &= \frac{1}{2}\int 1_{[1,3]\cap[t-4,t-2]}(\tau)\,d\tau \\ &= \begin{cases} \frac{1}{2}(t-3) & \text{für } t \in [3,5] \\ \frac{1}{2}(7-t) & \text{für } t \in (5,7] \end{cases} \end{aligned}$$

Als Ergebnis erhalten wir eine Dreiecksverteilung, wie in Abbildung 2.9 dargestellt ist.

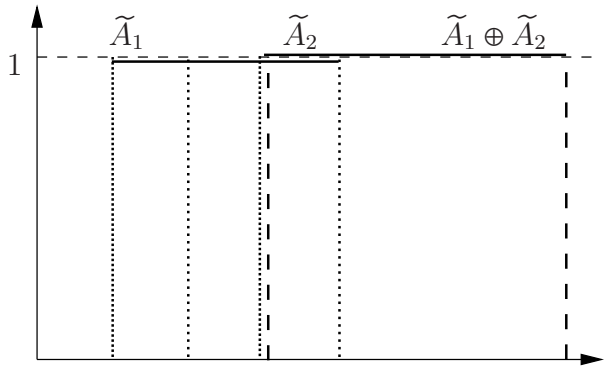
Wenn die Wahrscheinlichkeitsdichten als Zugehörigkeitsfunktionen interpretiert und normalisiert werden, erhalten wir in LL-Darstellung die Fuzzy-Mengen $\widetilde{A}_1 = (1,3;0,0)_L$ und $\widetilde{A}_2 = (2,4;0,0)_L$, die gewöhnliche Intervalle in $\mathbb{R}$ sind. Wie in Abschnitt *2.5.2* dargestellt wurde, können die Fuzzy-Zahlen als Zugehörigkeitsbeschreibung von ungenau beobachteten aber unbekannten reellwertigen Zuständen $a_1, a_2 \in \mathbb{R}$ betrachtet werden, deren Zugehörigkeit zum „wahren“ Zustand mit $\widetilde{A}_1$ bzw. $\widetilde{A}_2$ beschrieben wird. Sowohl die Wahrscheinlichkeitsbetrachtung als auch die Fuzzy-Bewertung bilden ab, dass die „wahren“

[115] Dies ist gerade die Modellvorstellung des statistischen Fehlermodells. Diese Transformation ist aber im Fuzzy-Kontext nicht darstellbar, so dass gewohnte Vorstellungen aus der Statistik nicht ohne Weiteres übertragbar sind.



(Eigene Darstellung)

Abbildung 2.9: Wahrscheinlichkeitsdichten bei der Addition zweier stochastisch unabhängiger gleichverteilter Zufallsvariablen



(Eigene Darstellung)

Abbildung 2.10: Addition von zwei Fuzzy-Mengen, die der Dichte der Gleichverteilungen entsprechen

Werte irgendwo im Intervall $[1,3]$ bzw. $[2,4]$ liegen. Die Zugehörigkeit der Summe $a_1 + a_2$ zur Summe der „wahren“ Zustände kann aber nur unspezifisch beschrieben werden als

$$\widetilde{A}_1 \oplus \widetilde{A}_2 = (a + c, b + d; 0, 0)_L = (3, 7; 0, 0)_L.$$

Eine Kompensation durch eine Häufung von Vorkommnissen findet dabei nicht statt, vgl. Abbildung 2.10. Dies ist ein wesentlicher Unterschied zum Wahrscheinlichkeitskalkül. Daher dient das Fuzzy-Kalkül eher als „Abschätzungstechnik“ dafür, wie die Ungenauigkeit der Beobachtungen sich in den Modellrechnungen auswirkt. Eine wesentliche Konsequenz des Perspektivwechsels vom Fehler in den Variablen- zu einem Fehler in den Daten-Ansatz ist darüber hinaus, dass die Korrelation der Fehlereinflüsse mit der Störvariablen aufgelöst werden kann.[116]

[116] Im Rahmen von Verteilungsansätzen für Fuzzy-Zufallsvariablen ist es vermutlich auch möglich, abgeschwächte, anteilige Korrelationen zu modellieren.

Eine Modellierung von epistemischen Fuzzy-Merkmalswerten bei fehlerhaften Daten ist vor allem dann sinnvoll, wenn die Evidenz für die Modellierung eines ökonometrischen Schätzmodells für die Fehlereinflüsse nicht ausreicht. Der Perspektivwechsel auf einen Ansatz mit Fehler in den Daten sollte also dann in Erwägung gezogen werden, wenn deterministische Fehler mit häufigeren Sprüngen im Fehlerbetrag vorliegen oder wenn die Stichprobe für den Fehlereinfluss im Verhältnis zur Schiefe der Fehlerverteilung und zur Fehlervarianz letztlich zu klein ausfällt. Eine Fuzzy-Modellierung der Fehlereinflüsse erscheint in jedem Fall dann geboten, wenn relevante Fehlereinflüsse nur ungefähr beschrieben und quantifiziert werden können.

Ein Vorteil des Fuzzy-Ansatzes besteht darin, dass damit grundsätzlich die Abbildung des aktuell vorliegenden Wissens über die Genauigkeit jedes einzelnen Beobachtungswertes möglich ist. Damit ist ein Instrumentarium verfügbar, mit dem der Informationsgrad über die Datenqualität in den Mittelpunkt der Analysen gerückt werden kann. Beispielsweise könnte mit Hilfe einer Fuzzy-Modellierung eine Landkarte bzw. eine Feldanalyse über die Genauigkeit der Datenwerte in einem Datensatz erstellt werden. Entsprechend bietet die Fuzzy-Modellierung einen Anlass dafür, nach besseren Bewertungsverfahren für die Qualität der Daten zu suchen und mit Bezug auf die Fuzzy-Abbildung weiter zu verbessern. Falls eine Fuzzy-Modellierung der individuellen Beobachtungswerte zu aufwändig erscheint, lässt die Fuzzy-Modellierung es ebenfalls zu, zu einer stärker vereinheitlichten Verwendung der Unschärfecharakteristika überzugehen.

Einige Konstruktionsstrategien für die empirische Anwendung werden in Kapitel *4* am Beispiel der Beschäftigtenstatistik herausgearbeitet.

3 Datenanalyse mit Fuzzy-Regression

Es gibt eine große Variationsbreite von Fuzzy-Methoden zur Datenanalyse, die in einem gewissen Sinne als „Fuzzy-Regression“ aufgefasst werden können. Im Folgenden soll untersucht werden, inwieweit Fuzzy-Daten genutzt werden können, um Ungenauigkeiten von Daten bei der einfachen Regression zu berücksichtigen. Wir beschränken uns daher auf solche Ansätze zur Fuzzy-Regression, bei denen eine lineare bzw. eine erweiterte lineare Funktion bei vorliegenden Fuzzy-Daten approximiert wird.

In der Literatur werden hauptsächlich zwei Motive für die Durchführung einer Fuzzy-Regression genannt. Einerseits wird Fuzzy-Regression benötigt, wenn die Merkmalswerte in Form von Fuzzy-Daten vorliegen und eine Regressionsanalyse durchgeführt werden soll. Für diese Situation sind weite Teile eines statistischen Instrumentariums bereits ausgearbeitet. Die entsprechenden Ansätze basieren auf Schätzmodellen mit Fuzzy-Zufallsvariablen. Für diese können wesentliche Güteeigenschaften der klassischen Kleinste Quadrate-Schätzer bestätigt werden. Allerdings ist die Modellierung durch die Verteilungsannahmen und deren technische Eigenschaften und Anforderungen stark vorstrukturiert, mit denen die Güte der Schätzer sichergestellt wird. Im Rahmen der epistemischen Semantik ist die Interpretation der Schätzparameter aber nicht trivial, weil sich in den Fuzzy-Daten alternative Konstruktionsstrategien zur Beschreibung der möglichen „wahren“ Werte niederschlagen können. Es bleibt daher unklar, über welche Mechanismen und bis zu welchem Grad die Güteaussagen über die Schätzparameter auf eine Situation mit epistemischen Fuzzy-Daten übertragen werden können. Neben diesen Ansätzen gibt es außerdem eine Reihe von Methoden zur Fuzzy-Regression, bei denen die epistemische Semantik stärker im Mittelpunkt steht. Für diese Ansätze wird allerdings in der Regel kein Verteilungsmodell angegeben. Sie sind daher zu den explorativen Regressionsmethoden zu rechnen.

Andererseits kann mit Hilfe einer Fuzzy-linearen Funktion ein ungefährer funktionaler Zusammenhang zwischen den unabhängigen und der abhängigen Variablen modelliert werden, wenn der Zusammenhang abgeschwächt ist, so dass er mit einer reellwertigen

Funktion nicht mehr adäquat abgebildet werden kann.[1] Ein abgeschwächter Zusammenhang ist z.B. naheliegend bei kleinen Stichproben oder bei Instabilitäten zwischen den Einzelmessungen. Mit Bezug auf die zweite Herangehensweise wird bei einer großen Anzahl von Methoden zur Fuzzy-Regression in Aussicht gestellt, dass die Bestimmung eines Strukturzusammenhangs auch in Situationen möglich ist, in denen keine wahrscheinlichkeitstheoretischen Verteilungsannahmen gemacht werden können. Stattdessen soll die Fuzziness der angepassten Fuzzy-Funktion die Variabilität der Beobachtungen zum funktionalen Zusammenhang abbilden. Leider werden, bis auf wenige Ausnahmen wie etwa Coppi u. a. [2006], bei keinem dieser Ansätze hinreichend genau festgelegt, welche Ziele mit der Datenanalyse angestrebt werden. Damit fehlt ein systematischer Zugang zur Interpretation der Regressionsfunktion und der Nutzen des Ansatzes für die Datenanalyse bleibt unklar.

Da Fuzzy-Daten kontextbezogene Informationen über die Ungenauigkeit in den Beobachtungen abbilden, ist es tatsächlich eine relevante Frage, welche Aussagen über die globale Ungenauigkeit daraus abgeleitet werden können. Die damit verbundenen Fragen der Interdependenz zwischen der Modellierungsstrategie für die Fuzzy-Daten und der Regressionsfunktion sollen in dieser Arbeit genauer betrachtet werden. Um der Verschiedenartigkeit der Ansätze gerecht zu werden, werden die Methoden zur Fuzzy-Regression hauptsächlich im Bezug auf ihre explorativen Eigenschaften untersucht und miteinander verglichen. Diese Herangehensweise wurde auch deshalb gewählt, weil dadurch die Unterscheidung der verschiedenen Ursachen von Datenunschärfe durch Vagheit, Ungenauigkeit oder Variabilität und deren Semantik erleichtert wird.

In Abschnitt *2.4.2* wurde bereits aufgezeigt, dass für Fuzzy-Vektoren nicht in der üblichen Weise eine Vektorraumstruktur sichergestellt werden kann. Es existiert dennoch eine lineare Struktur, für die allerdings Einschränkungen bei der Addition und der Skalarmultiplikation zu berücksichtigen sind. Diese arithmetischen Einschränkungen erfordern besondere Strategien bei der Datenapproximation. In Abschnitt *3.1.1* werden zunächst die besonderen Eigenschaften von Fuzzy-linearen Funktionen diskutiert, die bei der Datenanalyse in Betracht zu ziehen sind. Die Eigenschaften der Fuzzy-linearen Funktionen bedeuten nicht notwendig eine Einschränkung, sondern können unter Umständen auch gezielt für die Datenanalyse ausgenutzt werde. Dieser Aspekt wird später in Abschnitt *3.3* erneut aufgegriffen und genauer erörtert. In Abschnitt *3.1.2* wird schließlich beleuchtet, wie sich die eingeschränkte Linearität auf das Funktional auswirkt, das bei der Regression zu optimieren ist.

Ein Überblick über die Methoden zur Fuzzy-Regression wird in Abschnitt *3.2* gegeben. Zu den unterschiedlichen Ansätzen wird jeweils eine Methode ausführlich präsentiert, um

[1] Vgl. [Näther, 2006].

damit einen Einblick in die technischen Besonderheiten des Ansatzes zu vermitteln. Die Fuzzy-Regressionsmethoden werden in Abschnitt *3.3* schließlich in Auseinandersetzung mit unterschiedlichen Zielen in der empirischen Datenanalyse allgemein bewertet.

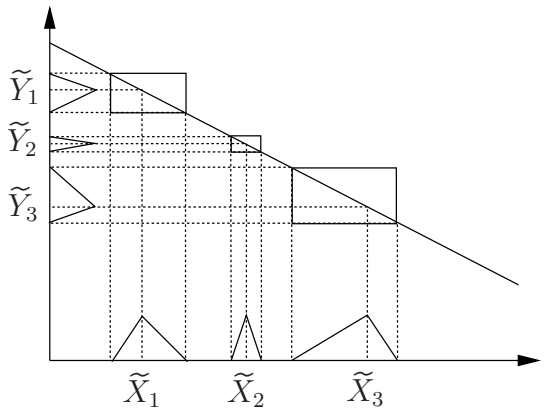
Zum Abschluss werden die vorgestellten Fuzzy-Regressionsmethoden danach verglichen, inwieweit die Modellvorstellungen, die den Methoden zugrundeliegen, tatsächlich einen Beitrag zur Beantwortung von ökonometrischen Fragestellungen leisten können. Um eine Systematik für die Bewertung der Fuzzy-Regression zu etablieren, werden für die Methodendiskussion formalisierte Approximations- bzw. Regressionsmodelle zugrundegelegt, wobei wir bei der Anpassung einer Ausgleichsfunktion zwischen einem defuzzifizierten Ausgleich und der Approximation einer fuzzifizierten Ausgleichsfunktion bzw. bei den Strukturmodellen zwischen einem Modell mit reellwertigen und mit Fuzzy-Parametern unterscheiden.[2] Ziel der Methodendiskussion ist die Entwicklung von Kriterien zur Bewertung und Auswahl eines Fuzzy-Regressionsansatzes bei praktischen Anwendungen. Konkret sollen als Vorbereitung auf das Benchmarking in Kapitel *5* die Regressionsmethoden identifiziert werden, die für die Regression bei epistemischen Fuzzy-Daten am besten geeignet sind.

3.1 Eigenschaften der Fuzzy-linearen Modellfunktionen

Die Fuzzy-linearen Funktionen vom Typ $f[\widetilde{U}, a], f[u, \widetilde{A}]$ und $f[\widetilde{U}, \widetilde{A}]$ sind die Kandidatinnen, die als Modellfunktionen in der Fuzzy-Regression verwendet werden sollen. Da die Funktionen wegen der Einschränkungen in der Linearität nicht in der gewohnten Weise wie lineare Funktionen interpretiert werden können, ist die Art des funktionalen Zusammenhangs zu charakterisieren, den diese beschreiben. Dazu werden in einem ersten Schritt die jeweiligen Funktionseigenschaften in einem gegebenen Parameter $\mathbf{a}$ bzw. $\widetilde{A}$ untersucht.

Bei der Funktionsanpassung sind umgekehrt die Eigenschaften der Bildmengen bei gegebenen Messwerten $(\widetilde{X}_i, \widetilde{Y}_i) \in F_{coc}^{nob}(\mathbb{R})^{m+1}$ für $1 \leq i \leq n$ und frei variierenden Parametern $\mathbf{a} \in A \subset \mathbb{R}^m$ bzw. $\widetilde{A} \in \mathcal{F}(A) \subset F_{coc}^{nob}(\mathbb{R})^m$ relevant. Hierbei sind Einschränkungen nötig, um die Wohldefiniertheit des Approximationsproblems zu sichern.

[2] Grundsätzlich kann bei den Modellansätzen sowohl die Annahme von reellwertigen Parametern als auch von Fuzzy-Parametern unabhängig davon gesehen werden, ob die Eingangsdaten Fuzzy-Daten sind oder nicht.



(Eigene Darstellung)

Abbildung 3.1: Fuzzy-lineare Funktion $f[\widetilde{U}, a]$ mit beliebigen Fuzzy-Inputs $\widetilde{X}_1, \widetilde{X}_2, \widetilde{X}_3$ und zugehörigen Bildwerten $\widetilde{Y}_1, \widetilde{Y}_2, \widetilde{Y}_3$

3.1.1 Abbildungscharakteristika der Fuzzy-linearen Funktionen

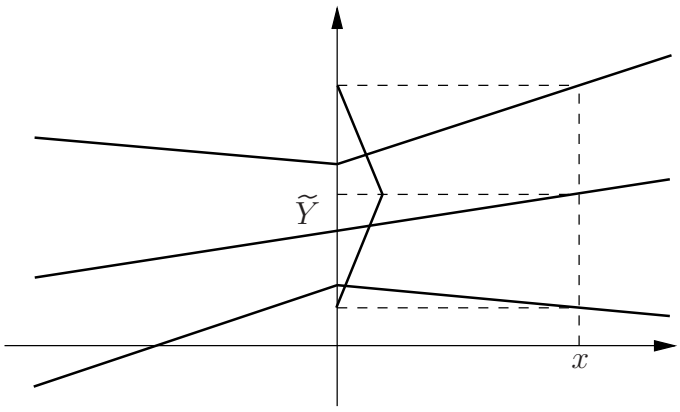
Im Folgenden werden einige Eigenschaften von Fuzzy-linearen Funktionen vorgestellt, die den Zusammenhang zwischen Input- und Outputwerten charakterisieren. Die Zusammenstellung verdeutlicht, dass die Modellierung einen funktionalen Zusammenhangs als Fuzzy-linearer Zusammenhang mit einigen Einschränkungen und Besonderheiten verbunden ist.

Den einfachsten Fall stellt die Fuzzy-lineare Funktion $f[\widetilde{U}, a]$ dar. Sei dazu $\mathbf{a} \in \mathbb{R}^{k+1}$ angenommen. Dann ist f die direkte Erweiterung einer klassischen linearen Modellfunktion auf ein Tupel von Fuzzy-Eingangswerten $(\widetilde{X}_1, \ldots, \widetilde{X}_k) \in F_{coc}^{nob}(\mathbb{R})^k$. Abbildung 3.1 zeigt, dass die Bildwerte ebenfalls Fuzzy-Daten sind und sich durch Projektion der Fuzzy-Inputs vermittelt über die reellwertige lineare Funktion auf die Y-Achse ergeben. Dabei wird sowohl die Lage der Fuzzy-Eingangswerte als auch deren Fuzziness proportional zu den Steigungsparametern a_j transformiert. Die Spannweite $y^R(0) - y^L(0)$ des Bildwertes $\widetilde{Y}(\widetilde{X}_1, \ldots, \widetilde{X}_k)$ ergibt sich als lineare Transformation der Spannweiten der Eingangswerte $\widetilde{X}_i$. Dabei sind $y^L(0)$ und $y^R(0)$ die linke und die rechte Intervallgrenze des Trägers von $\widetilde{Y}$, d.h. $[\widetilde{Y}]^0 = [y^L(0), y^R(0)]$[3]. Es gilt

$$y^R(0) - y^L(0) = \sum_{j=1}^{k} |a_j|(x_j^R(0) - x_j^L(0)).$$

Bei der Fuzzy-linearen Funktion $f[u, \widetilde{A}]$ mit reellwertigen Inputs und Fuzzy-Parametern $(\widetilde{A}_0, \ldots, \widetilde{A}_k) \in F_{coc}^{nob}(\mathbb{R})^{k+1}$ werden hingegen reellwertige Eingangswerte auf Fuzzy-Werte abgebildet. Die Funktion erzeugt also „zusätzliche“ Fuzziness. Ähnlich wie im reellwertigen

[3] Vgl. die Setzung in Satz 2.19 auf S. 63.



(Eigene Darstellung)

Abbildung 3.2: Fuzzy-lineare Funktion $f[u, \widetilde{A}]$ mit reellwertigem Input x und zugehörigem Bildwert $\widetilde{Y}$

Fall können auch die Fuzzy-Parameter aus dem Funktionsgraphen abgelesen werden. Sei dazu zunächst $k = 1$. Dann gilt $\widetilde{A}_0 = f_{\widetilde{A}}(0)$ und $\widetilde{A}_1 = f_{\widetilde{A}}(1) \ominus_H \widetilde{A}_0$, wobei $\ominus_H$ die *Hukuhara-Subtraktion* auf Fuzzy-Mengen ist. Die Hukuhara-Subtraktion $\widetilde{B} \ominus_H \widetilde{A}$ ist für zwei Fuzzy-Mengen $\widetilde{A}, \widetilde{B} \in F_{coc}^{nob}(\mathbb{R}^m)$ nur dann definiert wenn die Gleichung $\widetilde{A} \oplus \widetilde{X} = \widetilde{B}$ eine Lösung $\widetilde{X} \in F_{coc}^{nob}(\mathbb{R}^m)$ hat. Dann ist $\widetilde{B} \ominus_H \widetilde{A} = \widetilde{X}$.[4] Aufgrund der Konstruktion der Fuzzy-linearen Funktionen gibt es immer eine Lösung für die Subtraktion. Auch für höhere Dimensionen können die Parameter durch Einsetzen der Basiskoordinaten analog ermittelt werden. Im Fall $k = 2$ also durch Einsetzen von $(1,0), (0,1)$ u.s.f. Bei gleicher Spannweite der Inputwerte — die hier 0 ist — wachsen die Outputspannweiten mit wachsendem Abstand von x zum Koordinatenursprung. Für die Spannweiten gilt:

$$y^R(0) - y^L(0) = (a_0^R(0) - a_0^L(0)) + \sum_{j=1}^{k} (a_j^R(0) - a_j^L(0))|x_j|$$

Die Funktion hat ihre schmalste Stelle für $\mathbf{x} = \mathbf{0}$, so dass sich in jeder Koordinatenrichtung eine schmetterlingsförmige Gestalt abzeichnet, wie Abbildung 3.2 illustriert. Eine Translation aus der 0 kann nur durch Verschiebung der Eingangswerte erreicht werden. Die eingeschränkte Linearität bewirkt überdies, dass die Bilder auf der gegenüberliegenden Halbachse jeweils um den Kern gespiegelt sind.

Die Abbildung von Fuzzy-Inputs mit einer Fuzzy-linearen Funktion vom Typ $f[\widetilde{U}, \widetilde{A}]$ erzeugt ebenfalls zusätzliche Fuzziness gegenüber dem Fall mit reellwertigen Parametern. Bei Einsetzen von unechten Fuzzy-Vektoren $(\widetilde{X}_1, \ldots, \widetilde{X}_m) = \mathbf{x} \in \mathbb{R}^m$ ist die Funktion identisch mit dem Typ $f[u, \widetilde{A}]$. Die Eigenschaften, d.h. Spiegelung der Bildwerte bei einem Vorzeichenwechsel im Parameter sowie Wachstum der Outputspannweiten für wachsenden

[4] Zur Definition vgl. Diamond und Kloeden 1994, S. 8.

Entfernung von $\widetilde{X}_{ij}$ vom Ursprung, können daher analog übertragen werden. Zur Bestimmung der Outputspannweiten gibt es keine vergleichbare Formel wie in den beiden ersten Fällen, stattdessen können die Spannweiten der Bilder $\widetilde{Y}$ mit der folgenden Formel nach unten abgeschätzt werden.

$$y^R(0) - y^L(0) \geq (a_0^R(0) - a_0^L(0)) + \sum_{j=1}^{m} |a_j^R(0)x_j^R(0) - a_j^L(0)x_j^L(0)|.$$

Es gilt Gleichheit, wenn es kein $1 \leq j \leq m$ gibt, so dass $a_j^L(0) < 0 \leq a_j^R(0)$.

Schließlich kann jede Fuzzy-lineare Funktion $f[\cdot, \widetilde{A}]$ mit Fuzzy-Parametern auch als Fuzzy-Bündel $\widetilde{F}_f \in \mathcal{F}(\mathrm{Lin}(\mathbb{R}^k, \mathbb{R}))$ linearer Funktionen aufgefasst werden. Seien dazu $f^{\alpha,L}, f^{\alpha,R}$ die Funktionen, die die Intervallgrenzen von f auf dem α-Niveau beschreiben. D.h. $f^{\alpha,L}, f^{\alpha,R} : \mathbb{R}^k \longrightarrow \mathbb{R}$ sind definiert dadurch, dass für alle $\mathbf{x} \in \mathbb{R}^k$ ein $\widetilde{X} \in F_{coc}^{nob}(\mathbb{R})^k$ existiert, so dass $f^{\alpha,L}(\mathbf{x}) = \inf[f(\widetilde{X})]^\alpha$ und $f^{\alpha,R}(\mathbf{x}) = \sup[f(\widetilde{X})]^\alpha$. Damit wird definiert

$$\mathrm{supp}(\widetilde{F}_f) = \{f_{\mathbf{a}} \mid \mathbf{a} \in [\widetilde{A}]^0 \text{ und } f_{\mathbf{a}} \in \{f^{\alpha,L}, f^{\alpha,R}\} \text{ für } \alpha \in [0,1]\}$$

und

$$\mu_{\widetilde{F}_f}(f_{\mathbf{a}}) = \mu_{\widetilde{A}}(\mathbf{a}).$$

Eine umgekehrte Beziehung gilt nicht, da ein Fuzzy-Bündel von Funktionen im allgemeinen über einem Inputwert $\mathbf{x}$ verschiedene Zugehörigkeitswerte annimmt, d.h. es gibt $g_1, \ldots, g_l \in \widetilde{F}$ mit $g_1(\mathbf{x}) = \cdots = g_l(\mathbf{x})$, aber $\mu_{\widetilde{F}}(g_1) \neq \cdots \neq \mu_{\widetilde{F}}(g_l)$.[5] Im allgemeinen kann daher aus $\widetilde{F}$ eine fuzzifizierende Funktion nur in Gestalt einer aggregierten Randfunktion gewonnen werden. Beispielsweise ist $f_F : \mathbb{R}^k \longrightarrow F_{coc}^{nob}(\mathbb{R})$ bei Supremumsbildung definiert durch

$$\mu_{f_F(x)}(y) = \sup_{f \in \widetilde{F}, y = f(x)} \mu_{\widetilde{F}}(f). \tag{3.1}$$

3.1.2 Fuzzy-lineares Bild auf der Parametermenge

Nun soll der umgekehrte Fall betrachtet werden, bei dem die Inputwerte für die Modellfunktion gegeben, die Funktionsparameter aber unbestimmt sind. Seien dazu n Inputvektoren gegeben, d.h. für $1 \leq i \leq n$ seien $\mathbf{x}_i \in \mathbb{R}^m$ bzw. $(\widetilde{X}_{i1}, \ldots, \widetilde{X}_{im}) \in F_{coc}^{nob}(\mathbb{R})^m$ gegeben.

Durch Einsetzen der Inputvektoren in die parametrisierte Modellfunktion erhalten wir zu jedem Parameter ein n-Tupel von hypothetischen Funktionswerten. Diese stellen die

[5] Vgl. Dubois und Prade 1980, S. 99f.

Suchmenge für die Funktionsanpassung dar. Die Parameter sind zunächst unbestimmt, daher können wir die Eigenschaften der Suchmenge anhand der Auswertungsfunktionale

$$\chi_x : F_{coc}^{nob}(\mathbb{R})^m \longrightarrow F_{coc}^{nob}(\mathbb{R})^n, \chi_x(\widetilde{A}) = (f_{\widetilde{A}}(x_{i1}, \ldots, x_{im}); 1 \le i \le n) \tag{3.2}$$

für reellwertige Inputvektoren

bzw.

$$\chi_{\widetilde{X}} : \mathbb{R}^m \longrightarrow F_{coc}^{nob}(\mathbb{R})^n, \chi_{\widetilde{X}}(\mathbf{a}) = (f_{\mathbf{a}}(\widetilde{X}_{i1}, \ldots, \widetilde{X}_{im}); 1 \le i \le n) \tag{3.3}$$

für Fuzzy-Inputvektoren

über der Parametermenge untersuchen.

Im Fall der reellwertigen linearen Regression bilden die Bildmengen der Auswertungsfunktionale einen linearen Teilraum in der Parametermenge. Aufgrund der eingeschränkten Linearität können bei den Fuzzy-linearen Funktionen die Linearitätseigenschaften nur für die Skalarmultiplikation mit positiven Werten $\lambda \in [0, \infty)$ garantiert werden, wie Gleichung (2.17) zeigt.

Noch deutlicher wird der Charakter dieser Einschränkung, wenn man sich klarmacht, dass alle verwendeten Fuzzy-Mengen ihrerseits als Funktionen modelliert sind. Zur Vereinfachung wird hierbei angenommen, dass $\widetilde{A}_0 = \{0\}$. Das stellt keine Einschränkung dar, weil die nachfolgend diskutierten Einschränkungen durch die (Skalar-)Multiplikation von Fuzzy-Intervallen induziert sind und daher nicht für $\widetilde{A}_0$ gelten. Wir erhalten dann beispielsweise bei $m = 1$ für χ_x

$$\begin{aligned}
&\forall\ 1 \le i \le n, \alpha \in (0,1]: \\
&(a^L(\alpha), a^R(\alpha)) \in \mathbb{R}^2 \longmapsto [a^L(\alpha), a^R(\alpha)] \subset \mathbb{R} \longmapsto \\
&\qquad [f_{\widetilde{A}}(x_i)]^\alpha = \begin{cases} [x_i a^L(\alpha), x_i a^R(\alpha)] & \text{für } x_i \ge 0, \\ [x_i a^R(\alpha), x_i a^L(\alpha)] & \text{für } x_i < 0. \end{cases}
\end{aligned}$$

Für $m = 1$ und $\chi_{\widetilde{X}}$ erhalten wir bei reellwertigen Parametern in ähnlicher Weise

$$\begin{aligned}
&\forall\ 1 \le i \le n, \alpha \in (0,1]: \\
&a \in \mathbb{R} \longmapsto [f_a(\widetilde{X}_i)]^\alpha = \begin{cases} [a x_i^L(\alpha), a x_i^R(\alpha)] & \text{für } a \ge 0, \\ [a x_i^R(\alpha), a x_i^L(\alpha)] & \text{für } a < 0. \end{cases}
\end{aligned}$$

Aus diesen Darstellungen ist ersichtlich, dass bei gegebenem Input $\mathbf{x} \in \mathbb{R}^m$ das Auswertungsfunktional $\chi_{\mathbf{x}}$ für beliebige Fuzzy-Parameter $(\widetilde{A}_1, \ldots, \widetilde{A}_m) \in F_{coc}^{nob}(\mathbb{R})^m$ mit derselben Bestimmungsfunktion berechnet werden kann. Für $\chi_{\widetilde{X}}$ hingegen können die Funktionsauswertungen nur für die Teilmengen der Parametermenge mit derselben Bestimmungsfunktion berechnet werden, für die sich die Vorzeichen nicht ändern. Die Situation entspricht im Prinzip der in Abbildung 3.2, dabei sind allerdings die Fuzzy-Inputs mit den Fuzzy-Funktionsparametern $\widetilde{A}_j$ der abgebildeten Funktion zu identifizieren und die unbestimmten Parameter $\mathbf{a}$ des Auswertungsfunktionals mit den Eingangswerten $\mathbf{x}$ der Beispielfunktion. Dies macht deutlich, dass $\chi_{\widetilde{X}}$ sich bei einem Vorzeichenwechsel in der Menge der Parameter sprunghaft ändert.

Sei nun $A \subset \mathbb{R}^m$ eine m-dimensionale Menge. Dann können Berechnungen jeweils nur auf der Schnittmenge von A mit einem kartesischen Orthanten von $\mathbb{R}^m$ durchgeführt werden. Dabei werden die *kartesischen Orthanten* durch das m-fache Kreuzprodukt positiver bzw. negativer Halbachsen der reellen Zahlen gebildet.

Definition 3.1 (Kartesische Orthanten) Seien die positive und negative Halbachse von $\mathbb{R}$ indiziert durch

$$\mathbb{R}_b = \begin{cases} [0, \infty) & b = 1, \\ (-\infty, 0] & b = -1. \end{cases}$$

Dann sind die *kartesischen Orthanten* von $\mathbb{R}^m$ definiert durch

$$\forall\ \mathbf{q} = (b_1, \ldots, b_m) \in \{-1, 1\}^m : \quad \mathbb{R}_{\mathbf{q}} = \bigtimes_{j=1}^{m} \mathbb{R}_{b_j}.$$

Insgesamt erhalten wir also, dass $\chi_{\widetilde{X}}$ bei reellwertigen Parametern $A \subset \mathbb{R}^m$ über der Parametermenge in 2^m Bereiche zerfällt, die bei der Optimierung getrennt zu behandeln sind.[6]

Einen Sonderfall stellt es dar, wenn bei Fuzzy-Inputs $(\widetilde{X}_{i1}, \ldots, \widetilde{X}_{im}) \in F_{coc}^{nob}(\mathbb{R})^m$ eine Fuzzy-lineare Funktion mit Fuzzy-Parametern $\mathcal{A} \subset F_{coc}^{nob}(\mathbb{R})^m$ angenommen wird, wenn also $\chi_{\widetilde{X}} : F_{coc}^{nob}(\mathbb{R})^m \longrightarrow F_{coc}^{nob}(\mathbb{R})^n$ zu betrachten ist. Um die Berechnung nicht unnötig zu komplizieren soll davon ausgegangen werden, dass alle Inputs in einem Orthanten von $\mathbb{R}^m$ liegen. Das kann ggf. durch eine Datentransformation erreicht werden. Satz 2.19 zeigt, dass dann nicht nur die Parametervorzeichen in den verschiedenen Orthanten gesondert

[6] Dieser Sachverhalt lässt sich in streng formalisierter Form darüber herleiten, dass die Einbettungsfunktionale $\jmath_{F_{coc}^{nob}(\mathbb{R})^m}$ von $F_{coc}^{nob}(\mathbb{R})^m$ nach $L^2([0,1]\times\mathbb{S}^0)^m$ nur auf den kartesischen Orthanten linear sind. Vgl. Krätschmer 2006a, S. 2584.

zu betrachten sind, sondern darüber hinaus auch die Fälle bei denen $0 \in \widetilde{A}_j$ ist. Diese Übergänge über die Koordinatenachsen sind besonders aufwändig zu berechnen, da es im allgemeinen $\alpha \in (0,1)$ gibt, so dass $0 < a_j^L(\alpha)$ oder $a_j^R(\alpha) < 0$.

Bei Vorliegen von Fuzzy-Inputs bedarf es daher besonderer Maßnahmen, um die Fuzzy-lineare Approximation zu bestimmen. Insbesondere ist nachzuweisen, dass ein globales Optimum für das Regressionsproblem existiert. Der Suchaufwand zur Bestimmung der Parameter kann reduziert werden, wenn Dateneigenschaften nachgewiesen werden können, die die Lage des Steigungsparameters in bestimmten Orthanten nach sich zieht.[7] Darüber hinaus ist es sinnvoll, explorative Instrumente zur Reduzierung der Zahl der potentiellen Orthanten zu entwickeln.

3.2 Methoden der Fuzzy-Regression

Bei allen Methoden zur Fuzzy-Regression wird zu gegebenen Fuzzy-Daten eine Funktion aus einer Menge von interessierenden Funktionen bestimmt, die die gegebenen (Fuzzy-) Daten abhängig von den Annahmen des jeweiligen Regressionsansatzes „bestmöglich“ annähert. Dabei wird die bestmögliche Annäherung zwischen Daten und Modellfunktion im allgemeinen durch Minimieren einer Distanzfunktion bestimmt. Beispiele für eine solche Distanzfunktion sind u.a. Metriken und Quasimetriken auf Fuzzy-Zahlen, aber auch andere Funktionen, die die Abweichung zwischen zwei Fuzzy-Mengen abbilden. Wir notieren diese Anforderungen in allgemeiner Form:

Definition 3.2 Mit $\mathcal{M} = \mathrm{Lin}(F_{coc}^{nob}(\mathbb{R})^k, F_{coc}^{nob}(\mathbb{R}))$ sei die Menge der Fuzzy-linearen Funktionen bezeichnet. Seien nun $n, k \in \mathbb{N}$ mit $k < n$. Seien ferner für $1 \le i \le n$ die Fuzzy-Daten $(\widetilde{X}_{i1}, \ldots, \widetilde{X}_{ik}, \widetilde{Y}_i)$ in $F_{coc}^{nob}(\mathbb{R})^{k+1}$ gegeben.

Seien außerdem $\mathcal{A} \subset F_{coc}^{nob}(\mathbb{R})^{k+1}$ eine Menge von Fuzzy-Parametern. Die Menge $\mathcal{H} \subset F_{coc}^{nob}(\mathbb{R})^n$ sei nun durch die Vektoren der Auswertungsfunktionale auf $\mathcal{A}$ gegeben

$$\mathcal{H} = \{\big(\chi_{(\widetilde{X}_{11},\ldots,\widetilde{X}_{1k})}(\widetilde{A}), \ldots, \chi_{(\widetilde{X}_{n1},\ldots,\widetilde{X}_{nk})}(\widetilde{A})\big) \mid \widetilde{A} \in \mathcal{A} \text{ und } f(\cdot, \widetilde{A}) \in \mathcal{M}\}.$$

Eine *Ausgleichsfunktion* $f(\cdot, \widehat{\widetilde{A}}) \in \mathcal{M}$ mit Parameter $\widehat{\widetilde{A}} \in \mathcal{A}$ ist dann dadurch definiert, dass gilt

$$\forall\, \widetilde{H} \in \mathcal{H}: \quad \mathrm{dist}\left((\widetilde{Y}_1, \ldots, \widetilde{Y}_n), \widetilde{H}\right) \ge \mathrm{dist}\Big((\widetilde{Y}_1, \ldots, \widetilde{Y}_n), \big(\chi_{(\widetilde{X}_{11},\ldots,\widetilde{X}_{1k})}(\widehat{\widetilde{A}}), \ldots, \chi_{(\widetilde{X}_{n1},\ldots,\widetilde{X}_{nk})}(\widehat{\widetilde{A}})\big)\Big),$$

[7] In diese Kategorie sind die Eigenschaften der Dichte (engl. *tightness*) und des Zusammenhaltes (engl. *cohesiveness*) der Daten einzuordnen, die in [Diamond und Tanaka, 1999] eingeführt werden (S. 373ff.).

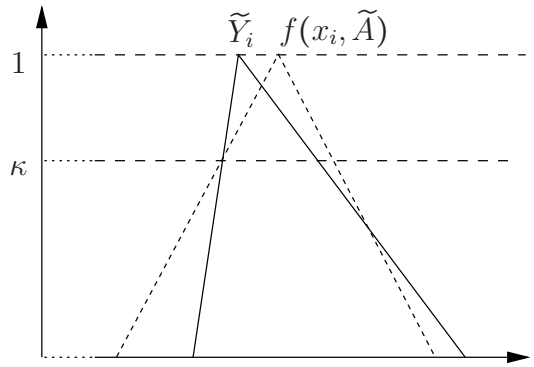
wobei dist eine beliebige Distanzfunktion sei.

Bei der Berechnung einer Ausgleichsfunktion ist nicht gefordert, dass die angepasste Funktion alle Datenpunkte berührt. Ziel der Methoden zur Fuzzy-Regression ist die globale Approximation einer Funktion an alle gegebenen Daten zugleich. Davon sind die Methoden der Interpolation zu unterscheiden, bei denen das Ziel ist, dass die angepasste Funktion die Datenpunkte enthält bzw. schneidet. Dies ist aber häufig nur dann erreichbar, wenn zugelassen wird, dass die Funktion lokal an eine Teilmenge von Punkten angepasst werden kann.

Unabhängig von der Semantik bei der Konstruktion von Fuzzy-Daten, die in Abschnitt *2.5.1* auf Seite 68 skizziert wurde, spielt der Dualismus zwischen Punktcharakter und Mengencharakter auch bei der Konstruktion der Methoden zur Fuzzy-Regression eine Rolle. Dabei bezieht sich der Punktcharakter auf die Lage eines Fuzzy-Vektors im Koordinatensystem, das durch die Merkmalsskalen aufgespannt wird, hingegen bezieht sich der Mengencharakter auf die Zugehörigkeitsumgebungen des Fuzzy-Vektors und somit auf die Unschärfecharakteristik. Je nachdem, welcher Blickwinkel gerade eingenommen wird, ist der Grad der Annäherung zwischen den Daten und der Approximationsfunktion durch eine Abstandsfunktion (Punktcharakter) bzw. durch eine Mengenbeziehung (Mengencharakter) zu beschreiben. Dieser Dualismus spiegelt sich bei den Methoden der Fuzzy-Regression u.a. darin wieder, dass sie bis auf wenige Ausnahmen entweder dem Ansatz der *Possibilistischen Regression* oder dem Ansatz der *Fuzzy-Regression vom Kleinste Quadrate-Typ* zuzurechnen sind.

In diesem Abschnitt wird ein Überblick über den Stand der Methodenentwicklung gegeben und es werden die zentralen Herangehensweisen zur Fuzzy-Regression vorgestellt. Bei der Aufarbeitung des Forschungsstandes stehen die technischen Eigenschaften der Methoden zunächst im Mittelpunkt, da die überwiegenden Zahl der Artikel zur Fuzzy-Regression sich darauf beschränkt. Die Verknüpfung mit einem Modellansatz und damit zu einer Interpretation des Approximationsergebnisses wird anschließend im Abschnitt *3.3* herausgearbeitet und diskutiert.

Zur Vereinfachung gilt im Weiteren für alle Fuzzy-Werte $\widetilde{B} \in F_{coc}^{nob}(\mathbb{R})$, dass sie einen eindeutigen Modalwert $b \in \mathbb{R}$ haben, sofern nicht explizit andere Angaben gemacht werden. D.h. unabhängig davon, ob es sich dabei um Fuzzy-Parameterwerte, Fuzzy-Inputs oder -Outputs handelt, sei der Kern aller Fuzzy-Werte einelementig. Zudem sei aus Vereinfachungsgründen immer vorausgesetzt, dass die Fuzzy-Daten vollständig im positiven Orthanten liegen.



(Eigene Darstellung)

Abbildung 3.3: Inklusionsbedingung $[\widetilde{Y}_i]^\kappa \subset [f(\mathbf{x}_i, \widetilde{A})]^\kappa$ für das Anspruchsniveau κ

3.2.1 Possibilistische Regression

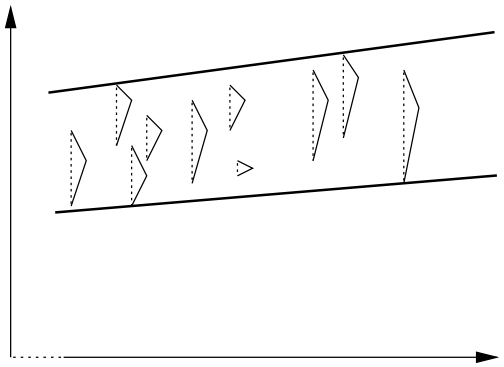
Der erste Ansatz der Possibilistischen Regression[8] wurde von Tanaka, Uejima und Asai entwickelt und von diesen 1982 vorgestellt. Als Grundmodell stellen wir hier eine ausgereifte Version vor, wie sie z.B. in Diamond und Tanaka [1999] beschrieben ist. An reellwertige Daten $(\mathbf{x}_i, y_i) \in \mathbb{R}^{k+1}$ oder Fuzzy-Daten vom Typ $(\mathbf{x}_i, \widetilde{Y}_i) \in \mathbb{R}^k \times F_{coc}^{nob}(\mathbb{R})$ wird dabei eine Fuzzy-Modellfunktion vom Typ $f[u, \widetilde{A}]$ angepasst. Die Approximation basiert auf der Idee, die bestmögliche Funktion unter den zulässigen Funktionen zu finden, für die eine Fuzzy-Ähnlichkeitsrelation zwischen den Daten und den Funktionswerten erfüllt ist. Der Vorteil dieses Ansatzes besteht darin, dass das Approximationsproblem in Form eines LP-Problems dargestellt und gelöst werden kann.

Bei der einfachen *Possibilistischen Regression* besteht die Menge der zulässigen Funktionen aus denjenigen Fuzzy-linearen Funktionen, die für ein vorgegebenes Zugehörigkeitsniveau $\kappa \in [0, 1]$ der Bedingung genügen, dass für alle Beobachtungswerte der Wert der abhängigen Variablen $\widetilde{Y}_i$ im Fuzzy-Funktionswert $f(\mathbf{x}_i, \widetilde{A})$ mit $\widetilde{A} = (\widetilde{A}_0, \ldots, \widetilde{A}_k) \in F_{coc}^{nob}(\mathbb{R})^{k+1}$ enthalten ist. D.h. für die gilt, dass die κ-Niveau-Menge zu $\widetilde{Y}_i$ eine Teilmenge der κ-Niveau-Menge zum Fuzzy-Funktionswert $f(\mathbf{x}_i, \widetilde{A})$ ist. Die Inklusionsbedingung ist in Abbildung 3.3 dargestellt. Durch diese Bedingung ist eine Ähnlichkeitsrelation beschrieben, die auch durch lineare Ungleichungen ausgedrückt werden kann, denn

$$[f(\mathbf{x}_i, \widetilde{A})]^\kappa \supset [\widetilde{Y}_i]^\kappa \quad \Leftrightarrow \quad f(\mathbf{x}_i, \widetilde{A})^L(\kappa) \leq y_i^L(\kappa) \text{ und } f(\mathbf{x}_i, \widetilde{A})^R(\kappa) \geq y_i^R(\kappa), \tag{3.4}$$

wobei $[\widetilde{Y}_i]^\kappa = [y_i^L(\kappa), y_i^R(\kappa)]$ und $[f(\mathbf{x}_i, \widetilde{A})]^\kappa = [f(\mathbf{x}_i, \widetilde{A})^L(\kappa), f(\mathbf{x}_i, \widetilde{A})^R(\kappa)]$ sei.

[8] Die Bezeichnung für diese Klasse von Methoden geht darauf zurück, dass die ermittelte Approximationsfunktion häufig mit Bezug auf die Semantik der Possibility-Verteilungen interpretiert wird. Diese mögliche Interpretation wird später in Abschnitt *3.3.2* kritisch hinterfragt.



(Eigene Darstellung in Anlehnung an Tanaka und Ishibuchi 1992, S. 57)

Abbildung 3.4: κ-Niveaudarstellung einer Possibilistischen Regression mit dem Anspruchsniveau $\kappa = 0$

Für die Approximation soll gelten, dass die Gesamtunschärfe bei Aggregation über die Funktionswerte im Parameter $\widetilde{A}$ minimal ist. Die Approximationsfunktion wird also als „minimale Einhüllende“ zum vorgegebenen Niveau κ konstruiert. In der Regel wird als Maß für die aggregierte Unschärfe G^κ die Summe der Unschärfen der Approximationswerte auf dem κ-Niveau verwendet[9]. Falls die Parameter $\widetilde{A}_0, \ldots, \widetilde{A}_k$ LR-Fuzzy-Zahlen mit einelementigen Kernen sind, ist G^κ also mit Satz 2.15 und Gleichung (2.10) definiert durch

$$\begin{aligned} G^\kappa(x, \widetilde{A}) &= \sum_{i=1}^{n} \Big(f(x_i, \widetilde{A})^R(\kappa) - f(x_i, \widetilde{A})^L(\kappa) \Big), \\ &= \sum_{i=1}^{n} \Big(\sum_{j=0}^{k} x_{ij} (r_{A_j} R^{-1}(\kappa) + l_{A_j} L^{-1}(\kappa)) \Big) \end{aligned}$$

dabei ist zur Vereinfachung der Schreibweise für alle $1 \le i \le n$ gesetzt $x_{i0} = 1$.

Da G^κ in den Variablen l_{A_j}, r_{A_j} linear ist, ergibt sich das folgende LP-Problem

$$\begin{array}{lll} \min & G^\kappa(x, \widetilde{A}) & \\ \text{unter NB} & f(\mathbf{x}_i, \widetilde{A})^L(\kappa) \le y_i^L(\kappa) \text{ und } f(\mathbf{x}_i, \widetilde{A})^R(\kappa) \ge y_i^R(\kappa) & (1 \le i \le n) \\ & l_{A_j}, r_{A_j} \ge 0 & (0 \le j \le k) \end{array} \tag{3.5}$$

Abbildung 3.4 zeigt die Intervallfunktion auf dem κ-Niveau, die sich bei einer Possibilistischen Regression an die vorliegenden Fuzzy-Daten $(x_i, \widetilde{Y}_i) \in \mathbb{R} \times F_{coc}^{nob}(\mathbb{R})$ ergibt. Dabei wurde als Anspruchsniveau $\kappa = 0$ gewählt. Die Intervallfunktion auf dem κ-Niveau berührt gerade die extremen Randwerte der Fuzzy-Datenpunkte auf dem κ-Niveau.

[9] In einigen früheren Arbeiten wird das einfachere Unschärfemaß $G'(x, \widetilde{A}) = \sum_{j=0}^{k} (l_{A_j} + r_{A_j})$ verwendet, vgl. z.B. [Savic und Pedrycz, 1991; Tanaka und Ishibuchi, 1992; Moskowitz und Kim, 1993]. Dieses muss allerdings als unbrauchbar angesehen werden, da mit G' als Zielfunktion die Lösungsparameter von der Skalierung der Merkmalswerte abhängig sind, wie [József, 1992] zeigt.

Da die Fuzzy-lineare Funktion bei der Anpassung nur auf dem κ-Niveau zu berechnen ist, kann das Verfahren ohne Weiteres auf Fuzzy-Daten vom Typ $(\widetilde{X}_i, \widetilde{Y}_i)$ ausgeweitet werden. Das Vorgehen wird in [Sakawa und Yano, 1992a,b] präsentiert.

Anstelle von (3.4) werden auch andere Ähnlichkeitsrelationen verwendet. Tanaka und Ishibuchi [1992] schlagen etwa auch die Bestimmung einer maximalen Ausfüllenden mit der Zulässigkeitsbedingung $[f(\mathbf{x}_i, \widetilde{A})]^\kappa \subset [\widetilde{Y}_i]^\kappa$ vor, die sog. *Necessity Regression*. Das zugehörige Approximationsproblem ist allerdings nur dann lösbar, wenn es ein $\mathbf{a} \in \mathbb{R}^{k+1}$ gibt, so dass für alle $1 \leq i \leq n$ gilt $f(\mathbf{x}_i, \mathbf{a}) \in \widetilde{Y}_i$.

Als dritte elementare Variante kann gefordert werden, dass Approximationsfunktion und Funktionswerte einen nicht-leeren Schnitt haben, d.h. $[f(\mathbf{x}_i, \widetilde{A})]^\kappa \cap [\widetilde{Y}_i]^\kappa \neq \varnothing$, diese Regressionsmethode wird auch als *Conjunction Regression*[10] bezeichnet. Mit Bezug auf verschiedene Indizes zum Vergleich von Fuzzy-Zahlen, die von Dubois und Prade entwickelt wurden, interpretieren Sakawa und Yano [1992a] den zulässigen Bereich der Conjunction Regression als die Menge der Fuzzy-linearen Funktionen für die die sogenannte Possibility für „$\widetilde{Y}_i = f(\mathbf{x}_i, \widetilde{A})$“ mindestens κ beträgt, wobei $Pos(\widetilde{A} = \widetilde{B}) = \sup\{\min\{\mu_{\widetilde{A}}(u), \mu_{\widetilde{B}}(u)\} \mid u \in \mathbb{R}\}$. Sie konstruieren andere possibilistische Regressionsmethoden auf der Basis von weiteren Ähnlichkeitsindizes für Fuzzy-Mengen.[11]

Als Alternativen zur Berücksichtigung der aggregierten Gesamtunschärfe G^κ in der Zielfunktion schlägt Bárdossy [1990] u.a. den Mittelwert über die Spannweiten der Approximationswerte sowie den Mittelwert über Flächeninhalte der Approximationswerte vor, der linearisiert werden kann. Tsaur und Wang [1999] erweitern die Funktionsanpassung um die zusätzliche Anforderung, dass die Spannweiten in den Approximationswerten die jeweiligen Spannweiten der gemessenen $\widetilde{Y}_i$ möglichst gut von oben annähern sollen, indem sie die folgende Zielfunktion verwenden:

$$G^\kappa_{TW} = \sum_{i=1}^{n} \Big(\Big(\Big[\sum_{j=0}^{k} x_{ij} r_{A_j} R^{-1}(\kappa) \Big] - r_{Y_i} R^{-1}(\kappa) \Big) + \Big(\Big[\sum_{j=0}^{k} x_{ij} l_{A_j} L^{-1}(\kappa) \Big] - l_{Y_i} L^{-1}(\kappa) \Big) \Big).$$

Dabei gilt unter den Nebenbedingungen von (3.5), dass $G^\kappa_{TW} \geq 0$.[12]

[10] Vgl. Tanaka und Ishibuchi 1991, S. 151. Die Ähnlichkeitsrelation des nicht-leeren Schnittes passt gut zur epistemischen Interpretation, da wir davon ausgehen, dass der „wahre“ Wert im Support der Daten enthalten ist und die Approximationsfunktion die „wahren“ Werte ebenfalls umfassen soll, so dass diese tendenziell im Schnittbereich liegen. Tatsächlich aber ist das Ergebnis der einfachen Conjunction Regression kaum interpretierbar, weil die Minimierung von G^κ entweder zu ∞-vielen reellwertigen Lösungen führt (in diesem Fall wäre eine Necessity Regression angemessener) oder die Lösung bei stark volatilen Daten die κ-Niveaus gerade noch an den Rändern berührt und damit nur die zulässigen Werte mit der geringstmöglichen Zugehörigkeit erfasst. Unabhängig davon kann die Maximierung der Schnitte zwischen Daten und Approximation grundsätzlich aber sehr wohl als Gütekriterium für die Anpassung betrachtet werden.

[11] Der interessanteste davon ist $Nec(\widetilde{Y}_i \subset f(\mathbf{x}_i, \widetilde{A})) = \inf\{\max\{1 - \mu_{\widetilde{Y}_i}(u), \mu_{f(\mathbf{x}_i, \widetilde{A})}(u)\} \mid u \in \mathbb{R}\}$. Es gilt $Nec(\widetilde{A} \subset \widetilde{B}) \neq 0$ genau dann, wenn der Kern von $\widetilde{A}$ einen nichtleeren Schnitt mit $\operatorname{supp} \widetilde{B}$ hat. In diesem Fall ist $Nec(\widetilde{A} \subset \widetilde{B}) = \mu_{\widetilde{B}}([\widetilde{A}]^1)$.

[12] Achtung, das lässt sich nicht mehr auf Fuzzy-Parameter $\widetilde{A}$ übertragen, deren Komponenten nichttriviale

Die Ähnlichkeit zwischen den Daten- und den Funktionswerten wird bei den Methoden der Possibilistische Regression für ein vorgegebenes κ-Niveau ausgewertet und bezieht ausschließlich dieses ein. Dies führt allerdings dazu, dass das 1-Niveau von $f(\cdot, \widehat{\widetilde{A}})$ nicht durch das Optimierungsproblem bestimmt ist, da das Unschärfemaß in der Zielfunktion nicht von der Lage der Modalwerte abhängt und die Abgrenzung des zulässigen Bereiches durch die Nebenbedingungen nur auf den Randwerten des κ-Niveaus beruht. Das Optimierungsproblem liefert demzufolge nur eine lineare Intervallfunktion, die die κ-Schnitte $(\mathbf{x}_i, [\widetilde{Y}_i]^\kappa)$ der Eingangsdaten annähert.

Das Approximationsproblem ist also insofern schlecht gestellt, dass die Modalwerte der Schätzfunktion $f(\cdot, \widehat{\widetilde{A}})$ nicht aus den Informationen in den Daten bestimmbar sind. Insbesondere ist die Lösung von (3.5) unabhängig von der Lage der Modalwerte der Eingangswerte $\widetilde{Y}_i$. Es sind daher zusätzliche Annahmen über die Eigenschaften der Anpassungsfunktion zu treffen, um eine echte Fuzzy Funktion[13] als Datenapproximation zu erhalten.

Die meisten Autor/innen umgehen das Problem der unbestimmten Modalwerte durch die Annahme, dass die Fuzzy-Parameter symmetrisch sind.[14] Diese Annahme wird allerdings in der Regel nicht inhaltlich motiviert. Meist beschränkt sich die Begründung darauf, dass dies der Vereinfachung der Berechnungen dient. Insbesondere bei der Anwendung der Fuzzy-Regression auf reellwertige Daten $(\mathbf{x}_i, y_i)$ drängt sich sogar häufig der Verdacht auf, dass Verteilungsannahmen aus der Statistik unhinterfragt auf die Fuzzy-Parameter übertragen werden.[15] Nur so ist zu erklären, warum in mehreren Arbeiten ein Benchmarking zwischen der Modalwertgeraden der Possibilistischen Regression und der Kleinste Quadrate-Regression durchgeführt wird.[16] Genauso gut könnte aber auch eine asymmetrische Lage der Modalwerte relativ zum linken und rechten Ende des Supports unterstellt werden. So kann z.B. vorgegeben werden, dass der Modalwert den Support im Verhältnis 1 : 2 unterteilt. Ein Nachteil einer solchen relativen Vorgabe der Modalwerte von $f(\cdot, \widetilde{A})$ ist aber, dass bei echten Fuzzy-Daten die Modalwertgerade von der Vorgabe von κ abhängt und bei einer Variation von κ im allgemeinen nicht stabil ist.

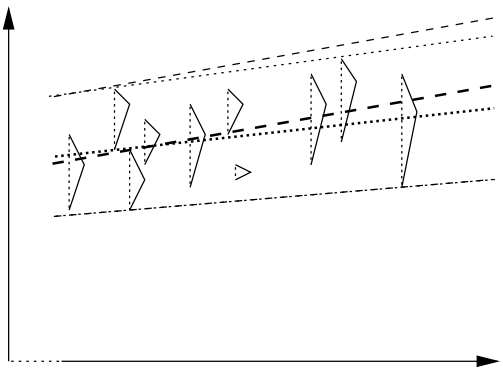
Fuzzy-Intervalle sind!

[13] Vergleiche dazu Definition 2.4.

[14] So z.B. [Sakawa und Yano, 1992a; Chang und Lee, 1994b; Diamond und Tanaka, 1999]. In diesem Fall gibt es häufig dennoch ∞-viele Lösungen.

[15] Als ein besonders augenfälliges Beispiel kann [Wang und Tsaur, 2000] gelten. Sie unterstellen eine Gleichverteilung von Informationen auf dem κ-Niveau und leiten daraus ab, dass die Intervallmitte aufgrund der Erwartungswerteigenschaft als besonders repräsentativ gewertet werden kann.

[16] Vgl. [Chang und Lee, 1994b; Kim u. a., 1996; Chang und Ayyub, 2001]. Die hohe Qualität der Anpassung ist in der Regel stark von der günstigen Verteilung der Beispieldaten beeinflusst, Artefakte werden meist nicht diskutiert. In der Simulationsstudie von Kim u. a. [1996] werden z.B. auch häufigkeitsbezogene Gütekriterien verwendet, die nicht geeignet sind, die Eigenschaften einer Fuzzy-Approximation angemessen darzustellen, da sich aus der Anpassungsfunktion keine weitergehenden Aussagen über die räumliche



(Eigene Darstellung)

Abbildung 3.5: Possibilistische Regression mit symmetrischen Parametern (unterste gestrichelte sowie die beiden gepunkteten Geraden) vs. hybride Possibilistische Regression mit Kleinste Quadrate Modalwertgerade (gestrichelte Geraden)

Es gibt mehrere Ansätze, bei denen Dateninformationen aus mehreren α-Niveaus berücksichtigt werden, um die Modalwerte von $f(\cdot, \widetilde{A})$ zu bestimmen. Dazu gehören die *hybriden Methoden zur Fuzzy-Regression*, bei denen zunächst eine klassische Kleinste Quadrate-Regression über die Modalwerte der Beobachtungen durchgeführt wird, d.h. es wird zunächst angenommen, dass wir reellwertige Beobachtungen $(\mathbf{x}_i, [\widetilde{Y}_i]^1)$ haben. Im zweiten Schritt werden die Unschärfen der Modellfunktion gemäß (3.5) minimiert. Dabei werden die Approximationswerte der Modalwertregression als Modalwerte der Fuzzy-linearen Approximation vorgegeben. Der Grundansatz der hybriden Fuzzy-Regression wird in [Savic und Pedrycz, 1991, 1992] vorgestellt, [Ishibuchi und Nii, 2001] erweitern den Ansatz auf Fuzzy-lineare Modellfunktionen mit asymmetrischen Fuzzy-Parametern. Außerdem wurden Methoden zur hybriden Fuzzy-Regression mit interaktiven Fuzzy-Parametern entwickelt und zwar für Parameter mit quadratischen Referenzfunktionen, vgl. [Tanaka und Ishibuchi, 1991], und mit exponentiellen Referenzfunktionen, vgl. [Tanaka u. a., 1995]; diese Ansätze führen allerdings zu nicht-linearen Nebenbedingungen. Abbildung 3.5 zeigt, dass die Trägermenge der Approximationsfunktion bei der hybriden Fuzzy-Regression im allgemeinen eine Obermenge der Trägermenge bei der Grundmethode der Possibilistischen Regression ist. Für Fuzzy-Daten $(\widetilde{X}_i, \widetilde{Y}_i) \in F_{coc}^{nob}(\mathbb{R})^{k+1}$, bei denen nicht nur die abhängige Variable Y sondern auch die unabhängigen Variablen $X_1, \ldots, X_k$ ungenau vorliegen, schlagen Redden und Woodall [1996] die Verwendung der orthogonalen Kleinste Quadrate-Regression zur Bestimmung der Modalwertgeraden vor, damit alle Dimensionen der Datenungenauigkeit in der Approximation berücksichtigt werden.[17]

Verteilung oder die relative Häufigkeit der Daten zwischen den extremen Randpunkten ableiten lassen.

[17] Vgl. ebd. S. 209.

Statt der Verknüpfung einer klassischen Regression für die defuzzifizierten Eingangsdaten mit einer Possibilistischen Regression können auch andere hybride Kombinationen verschiedener Fuzzy-Regressionsmethoden verwendet werden, die sich monoton zueinander verhalten und entsprechend angeordneten α-Niveaus zugeordnet werden. Wir fassen solche Methoden, zu denen auch die beschriebenen hybriden Ansätze gerechnet werden können, unter dem Oberbegriff *Bricolage*[18] bzw. *Designapproximation* zusammen. Beispiele dafür sind die Bestimmung von trapezoidalen Fuzzy-Parametern in [Tanaka und Ishibuchi, 1992], indem für Intervalldaten, die als κ-Schnitt über Fuzzy-Daten $(\mathbf{x}_i, \widetilde{Y}_i)$ betrachtet werden können, sowohl eine Possibility als auch eine Necessity Intervallapproximation berechnet werden, wobei die Possibility Approximation das 0-Niveau und die Necessity Approximation das 1-Niveau der Designapproximation bilden. [Ishibuchi und Nii, 2001] hingegen verknüpfen in den trapezoidalen Fuzzy-Parametern die possibilistische Anpassung für das κ-Niveau mit einer possibilistischen Anpassung für das 1-Niveau. Die so gewonnene Approximation hat die Eigenschaft, dass sie die Daten auf allen Niveaus $\alpha'' \geq \kappa$ einhüllt und überdies hat sie eine geringere Spannweite als die Approximation gemäß der einfachen Possibilistischen Regression.

Ein kritischer Punkt bei den Methoden der Possibilistischen Regression ist die Auswahl des Vergleichsniveaus κ. Offensichtlich nimmt die Spannweite der Approximationsfunktion $f(\cdot, \widehat{\widetilde{A}})$ zu, je näher κ bei 1 liegt. Dabei sei der triviale Fall ausgenommen, bei dem für alle Daten gilt $\widetilde{Y}_i = f(\mathbf{x}_i, \widetilde{A})$. Bei reellwertigen Daten $(\mathbf{x}_i, y_i)$ wird durch κ die Mindestzugehörigkeit der Datenpunkte zur Fuzzy-linearen Funktion nach unten begrenzt, d.h. $\mu_{f(\mathbf{x}_i, \widetilde{A})}(y_i) \geq \kappa$. Es wird also festgelegt, bis zu welchem Zugehörigkeitsgrad die Variationsbreite der Fuzzy-linearen Approximationsfunktion durch die Daten beschrieben wird. Fuzzy-Daten $(\mathbf{x}_i, \widetilde{Y}_i)$ beinhalten im Gegensatz dazu bereits eine Zugehörigkeitsbewertung.[19] Infolgedessen führt eine Erhöhung von κ zu einer besseren Approximation für die α''-Niveaus mit $\alpha'' \geq \kappa$, wobei gleichzeitig die Gesamtunschärfe der Anpassungsfunktion wächst und somit die Anpassung der α'-Niveaus mit $\alpha' < \kappa$ verschlechtert und durch eine vergröberte Abschätzung ersetzt wird.[20]

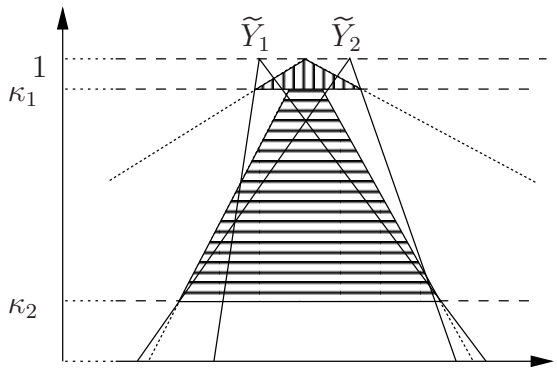
Insgesamt führt eine Erhöhung von κ zu einer besseren Anpassung der Modalwerte und damit auch einer stärkeren Gewichtung der Variation in der Lage der Daten, allerdings um den Preis zunehmender Funktionsunschärfe.[21] Demgegenüber führt eine Verringerung

[18] Diese Bezeichnung ist durch den französischen Begriff für „Bastelei“ inspiriert.

[19] [Ishibuchi und Nii, 2001] untersuchen auch das Übergangsmodell, bei dem reellwertige Daten mit einem Zugehörigkeitswert $\mu_{(\mathbf{x}_i, y_i)}(u, v) = \mu_{y_i}(v) = \mu_i 1_{y_i}(v)$ mit $\mu_i \in (0, 1]$ versehen sind.

[20] Diese zusätzliche Unschärfe bietet aber im allgemeinen keine Sicherheit dafür, dass auch die Werte auf dem α'-Niveau mit $\alpha' < \kappa$ von der Approximationsfunktion eingehüllt werden. Ein Gegenbeispiel lässt sich z.B. mit asymmetrischen Fuzzy-Daten konstruieren, die nicht gleichförmig sind.

[21] [Chang und Lee, 1994b] charakterisieren diesen Sachverhalt so: „die Verwendung eines hohen κ-Niveaus bedeutet, dass die Entscheidungsträger/in höheres Vertrauen in die Modalwerte der Daten [hat]“. Ähnlich



(Eigene Darstellung)

Abbildung 3.6: Anpassung der Lageschwankungen bei κ_1 vs. Anpassung der Datenungenauigkeit bei κ_2

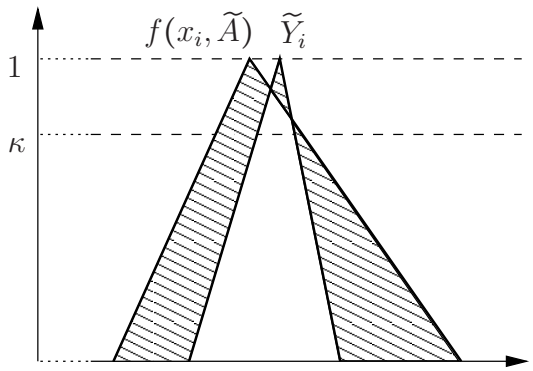
von κ zu einer stärkeren Gewichtung der Fuzziness der Datenpunkte, verbunden mit einer Verringerung der Parameterspannweiten. Der Zielkonflikt wird in Abbildung 3.6 illustriert. Zum Anspruchsniveau κ_1 ist die angepasste Intervallfunktion schmal und die Fuzzy-Approximation für die Niveaus oberhalb von κ_1 liegt nahe an den Modalwerten der Eingangswerte $\widetilde{Y}_1$ und $\widetilde{Y}_2$. Die Spannweite der Anpassung auf dem Niveau κ_1 ist hingegen sehr groß. Umgekehrt stellt der Fuzzy-Approximationswert zum Niveau κ_2 eine gute Approximation der Trägermengen von $\widetilde{Y}_1$ und $\widetilde{Y}_2$ dar, wohingegen die Modalwerte der Eingangswerte nicht gut approximiert werden.

Hinter der Festlegung von κ verbirgt sich somit ein komplexes Entscheidungsproblem bei dem die Bewertung des Beitrags der Fuzzy-Daten zur Lage der Approximationsfunktion nur im Widerspruch zur Bewertung des Beitrags der Fuzziness der Daten zur Fuzziness der Approximationsfunktion berücksichtigt werden kann. Vermutlich aus diesem Grund wird häufig der Kompromisswert $\kappa = \frac{1}{2}$ gewählt.

Eine bessere Alternative bietet die Methode von Sakawa und Yano [1992a,b] mit einer Mehrzieloptimierung, bei der sowohl das Anspruchsniveau κ maximiert als auch die Gesamtunschärfe der Approximation minimiert wird. Das Optimierungsproblem kann mit einem interaktiven Algorithmus gelöst werden, bei dem nach jedem Schritt der lokale Anstieg in der Gesamtunschärfe bei Erhöhung von κ als Entscheidungskriterium zur Verfügung gestellt wird. Als Ergebnis erhält man ein subjektiv und kontextabhängiges „bestes“ Anspruchsniveau κ.

Es gibt nur wenige Ansätze zur Bestimmung von Güteindizes für das Ergebnis einer Possibilistischen Regression, da üblicherweise davon ausgegangen wird, dass die Güte der Anpassung durch die Wahl von κ vorgegeben wird und κ somit das Anspruchsniveau für die Anpassung darstellt. Dieser Deutung steht aber der beschriebene Widerspruch für die

argumentieren auch [Chang und Ayyub, 2001].



(Eigene Darstellung)

Abbildung 3.7: Berechnung von D_i als Fläche der Mengendifferenz $\{(y, \mu_{\widetilde{Y}_i}(y)) \mid y \in \mathbb{R}\} \triangle \{(y, \mu_{f(\widetilde{X}_i, \widetilde{A})}(y) \mid y \in \mathbb{R}\}$

Anpassungsgüte bei der Auswahl eines geeigneten κ entgegen. Zudem ist die Güte der Anpassung auf dem κ-Niveau hoch abhängig davon, inwieweit die spezifizierten Fuzzy-Daten auf dem κ-Niveau vergleichbar sind. Denn im Unterschied zum üblichen Konstruktionsprinzip von α-Schnitten wird bei der Possibilistischen Regression mit der Auswahl des Anspruchsniveaus κ nicht zugleich sichergestellt, dass die erwünschte Eigenschaft — hier die Mengenähnlichkeit — ebenso für alle darüberliegenden Niveaus mit $\alpha'' \geq \kappa$ gilt. Es findet also faktisch eine Verengung der Betrachtung allein auf das Niveau κ statt. Von diesem Problem ist einzig der Ansatz von Ishibuchi und Nii [2001] ausgenommen, bei dem intervallförmige Fuzzy-Parameter als Parameter der Approximationsfunktion vorgesehen werden, die zusätzlich auf dem 1-Niveau minimale Einhüllende darstellen.

Das von Kim und Bishu [1998] vorgeschlagene Maß D für die Anpassungsgüte einer Fuzzy-linearen Funktion an Fuzzy-Daten eignet sich auch für die Possibilistische Regression. Dabei werden die außerhalb der Schnittmenge von $\widetilde{Y}_i$ und $f(\widetilde{X}_i, \widetilde{A})$ liegenden Flächen als Maßzahl für die Abweichung gewertet,[22] wobei für die Definition der allgemeinste Fall von Fuzzy-Daten mit Fuzzy-Inputs $\widetilde{X}_i = (\widetilde{X}_{i1}, \ldots, \widetilde{X}_{ik}) \in F_{coc}^{nob}(\mathbb{R})^k$ zugrundegelegt wird. Es sei

$$\begin{aligned} D_i &= \Delta(\widetilde{Y}_i, f(\widetilde{X}_i, \widetilde{A})) \\ &= \int_{\mathrm{supp}(\widetilde{Y}_i) \cup \mathrm{supp}(f(\widetilde{X}_i, \widetilde{A}))} |\mu_{\widetilde{Y}_i}(y) - \mu_{f(\widetilde{X}_i, \widetilde{A})}(y)| \, dx \end{aligned} \tag{3.6}$$

[22] Tatsächlich berechnen Kim und Bishu [1998] einen relativen Index $R = \sum_i R_i$ mit $R_i = D_i / \int_{\mathrm{supp}(\widetilde{Y}_i)} \mu_{\widetilde{Y}_i}(y) \, dy$. Der einfache Index D wird von Kao und Chyu [2002] verwendet, die die Arbeit von Kim und Bishu allerdings zitieren. R ist nicht auf 1 normiert, kann aber vielleicht als „relative Ungenauigkeitsverschlimmerung“ der Approximation gegenüber der Eingangsungenauigkeit in den Daten betrachtet werden.

und

$$D = \sum_{i=1}^{n} D_i. \tag{3.7}$$

Die Konstruktion wird in Abbildung 3.7 verdeutlicht.

D_i verringert sich, je größer die Schnittfläche von $\widetilde{Y}_i$ und $f(\widetilde{X}_i, \widetilde{Y}_i)$ ist und je größer der relative Anteil von $f(\widetilde{X}_i, \widetilde{Y}_i)$ ist, der $\widetilde{Y}_i$ überdeckt. Zusätzlich zur wünschenswerten größtmöglichen Überdeckung wird dabei also auch bewertet, wie stark die Kongruenz zwischen dem Approximationswert und dem Datenwert ist.

Wie in Abschnitt *3.1.1* dargestellt wurde, kann eine relative Verringerung der Spannweiten der $\widetilde{Y}_i$ bei wachsendem X durch eine Fuzzy-lineare Funktion nicht abgebildet werden. Chang und Lee [1994b] zeigen, dass diese funktionale Einschränkung eine wesentliche Ursache dafür ist, dass als Ergebnis der Possibilistischen Regression häufig keine Fuzzy-Parameter sondern reellwertige Parameter, d.h. mit Spannweite 0, bestimmt werden. Eine Verbesserung der Anpassung kann bei Daten mit einer fallenden Tendenz in den Spannweiten dadurch erreicht werden, dass in (3.5) auch negative Spannweiten $l_{A_j}, r_{A_j} \in \mathbb{R}$ zugelassen werden. Dies führt bei der Schreibweise als Fuzzy-lineare Funktion zu Fuzzy-Parametern $\widetilde{A}_j$, die nicht wohldefiniert sind. Die Gleichung kann aber weiterhin formal mit dem Erweiterungsprinzip sinnvoll berechnet werden, so dass negative Spannweiten als von interaktiven Parametern verursacht zu betrachten sind.[23] Georgescu [1998] schlägt mit dem „*decoupling principle*“ eine ähnliche Strategie vor, indem auf dem vorgegebenen κ-Niveau jeweils eine klassische Regressionsgerade an die linken bzw. rechten Intervallgrenzen der Fuzzy-Daten $(\mathbf{x}_i, \widetilde{Y}_i)$ angepasst wird, wobei zugelassen wird, dass die Parameter $\widetilde{A}$ nicht mehr wohldefiniert sind und $a_j^L(\kappa) > a_j^R(\kappa)$. Im Ergebnis erhalten wir weiterhin immer einen Intervallkorridor.[24]

Ein besonderer Nachteil der Possibilistischen Regression ist es, dass diese Methoden besonders sensitiv auf Änderungen in den Randwerten der Datenwolke reagieren. Ausreißer haben somit einen direkten Einfluss auf die Approximationsfunktion, der durch Vergrößerung der Zahl der Beobachtungen nicht verringert werden kann. Peters [1994] und Chen [2001] stellen Methoden vor, mit denen durch eine Transformation der Possibilistischen Regression in ein Fuzzy-Optimierungsproblem der Einfluss von extremen Werten abgeschwächt und potentielle Ausreißer explorativ bestimmt werden können. Einen anderen Ansatz entwickelt Inuiguchi [2003] — allerdings nur für Intervalldaten —, indem

[23] Vgl. Chang und Lee 1994b, S. 69.

[24] Die Erstellung einer Designapproximation durch eine Schichtung über mehrere κ-Niveaus ist im allgemeinen nicht möglich, weil diese die notwendige Monotonie nicht erfüllen. Hier ist das in der Arbeit vorgeführte Beispiel irreführend!

er Schlupfvariablen einerseits für zulässige und andererseits für unzulässige Abweichungen der Approximationswerte von den Intervalldaten einführt.[25] In der zu minimierenden Zielfunktion wird für jede Abweichung eine spezifische Strafbewertung in Abhängigkeit von der Schwere der Verletzung der Zulässigkeit berücksichtigt. Dabei werden zulässige Abweichungen mit einem Funktional zur robusten Approximation bewertet, so dass die Sensitivität für Ausreißer sinkt, unzulässige Abweichungen hingegen werden linear progressiv betraft. Der Vorteil dieses Ansatzes liegt u.a. darin, dass auch das Necessity-Problem immer lösbar ist.

Es gibt einen fließenden Übergang zwischen einigen Designansätze zur Fuzzy-Regression und den metrischen Ansätzen zur Fuzzy-Regression im nächsten Abschnitt. Beispielsweise bestimmen Kim und Bishu [1998] bei Fuzzy-Daten $(\mathbf{x}_i, \widetilde{Y}_i)$ auf dem κ-Niveau simultan klassische Regressionsgeraden zu den Modalwerten sowie den linken und rechten Intervallgrenzen. Dies führt im Wesentlichen zum selben Ergebnis wie die Kleinste Quadrate Fuzzy-Regression mit einer Modellfunktion vom Typ $f[u, \widetilde{A}]$.[26]

3.2.2 Kleinste Quadrate Fuzzy-Regression

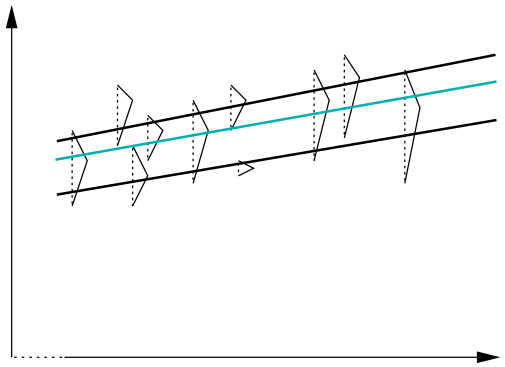
Eine weitere Klasse von Fuzzy-Regressionsmethoden zeichnet sich dadurch aus, dass der Anpassung der Fuzzy-linearen Funktion an die Fuzzy-Daten eine Metrik d auf $F_{coc}^{nob}(\mathbb{R})$ zugrunde gelegt wird, mit der die *Residuen* $e_i(\widetilde{A})$ zwischen den Datenwerten und den Approximationswerten der Modellfunktion gebildet werden können, die zu minimieren sind. Die Residuen sind definiert durch

$$e_i(\widetilde{A}) = d(\widetilde{Y}_i, f(\widetilde{X}_i, \widetilde{A})),$$

wobei wiederum $\widetilde{X}_i = (\widetilde{X}_{i1}, \ldots, \widetilde{X}_{ik}) \in F_{coc}^{nob}(\mathbb{R})^k$ und $\widetilde{A} = (\widetilde{A}_0, \ldots, \widetilde{A}_k) \in F_{coc}^{nob}(\mathbb{R})^{k+1}$ ist. Der Vorteil dieses Ansatzes liegt darin, dass die Residuen auch eine Aussage zur Güte der Modellanpassung treffen. Nahezu alle metrischen Ansätze zur Fuzzy-Regression sind darüber hinaus so konstruiert, dass die verwendeten Metriken bei reellwertigen Daten $(\mathbf{x}_i, y_i)$ zum Euklidischen Abstand entarten und das zugehörige Optimierungsproblem damit zur üblichen Kleinste Quadrate-Regression wird. Daher fassen wir diese Methoden hier unter dem Oberbegriff *Kleinste Quadrate Fuzzy-Regression* zusammen. Die direkte Übertragung der klassischen Kleinste Quadrate-Schätzer auf Fuzzy-Daten mittels des Erweiterungsprinzips, das in Definition 2.13 eingeführt wurde, führt zu einer Regressi-

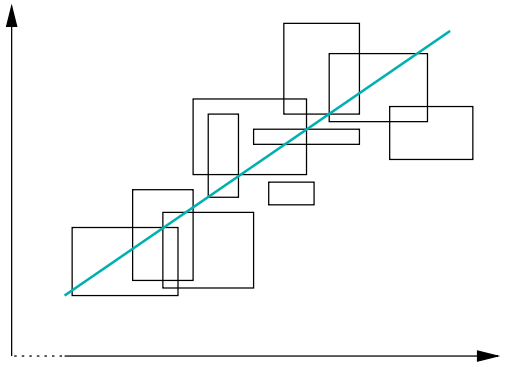
[25] Beispielsweise wird bei der Possibilistischen Regression eine Überschreitung der linken Grenze des κ-Niveaus durch die linke Grenze der Approximation um ε als unzulässige Abweichung gewertet.

[26] Vgl. Körner und Näther 1998, S. 114. Der Rechenalgorithmus ist auch in Gleichung (5.5) auf S. 165 zusammengefasst.



(Eigene Darstellung in Anlehnung an Körner und Näther 1998, S. 109)

Abbildung 3.8: Kleinste Quadrate Fuzzy-Regression für trianguläre Fuzzy-Daten $(x_i, \widetilde{Y}_i)$ mit $f(\cdot, \widetilde{A})$



(Eigene Darstellung)

Abbildung 3.9: Kleinste Quadrate Fuzzy-Regression für Fuzzy-Daten $(\widetilde{X}_i, \widetilde{Y}_i)$ mit $f(\cdot, \mathbf{a})$

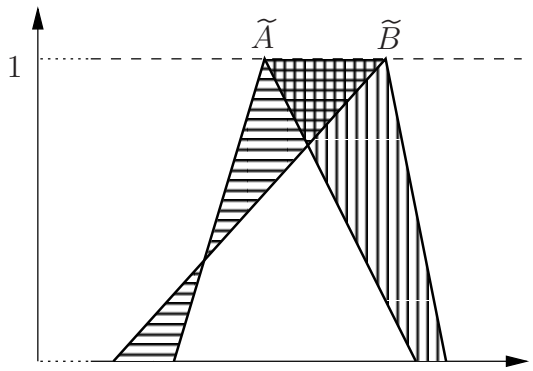
onsfunktion mit einer relativ großen Spannweite und ist daher im Vergleich mit den hier vorgestellten Methoden unbrauchbar, wie Körner und Näther [1998] zeigen.[27]

Als Grundmodell stellen wir die Kleinste Quadrate Fuzzy-Regression vor, wie sie z.B. in [Diamond und Tanaka, 1999] und [Körner und Näther, 1998] beschrieben ist. An beliebige Fuzzy-Daten $(\widetilde{X}_i, \widetilde{Y}_i)$ wird dabei eine Modellfunktion vom Typ $f[\widetilde{U}, a], f[u, \widetilde{A}]$ oder $f[\widetilde{U}, \widetilde{A}]$ angepasst. Analog wie bei der klassischen Kleinste Quadrate Regression ist derjenige Parameter $\widehat{\widetilde{A}} \in F_{coc}^{nob}(\mathbb{R})^k$ gesucht, für den die Summe der Abstandsquadrate bzw. der quadrierten Residuen e_i minimal wird, d.h.

$$D^2(\widetilde{A}) = \sum_{i=1}^{n} e_i^2(\widetilde{A}) = \sum_{i=1}^{n} d^2(\widetilde{Y}_i, f(\widetilde{X}_i, \widetilde{A})) \longrightarrow \min_{\widetilde{A}}. \tag{3.8}$$

Die Lösung des Regressionsproblems, für die die Summe der Abstandsquadrate D^2 minimal wird, ist offensichtlich von der Wahl der verwendeten Metrik d abhängig. Dabei lassen

[27] Vgl. ebd. S. 101ff.



(Eigene Darstellung)

Abbildung 3.10: Berechnung des δ_2-Abstands von $\widetilde{A}$ und $\widetilde{B}$

sich eine Vielzahl der Methoden zur Kleinste Quadrate Fuzzy-Regression mit Werten in $F_{coc}^{nob}(\mathbb{R})$ auf die Metrik δ_2 zurückführen, die von Diamond und Kloeden [1994] untersucht wird.[28] In den Abbildungen 3.8 und 3.9 werden einfache Beispiele für die mögliche Anwendung von Kleinste Quadrate Fuzzy-Regression gezeigt. Im Vergleich zur Possibilistischen Regression fällt auf, dass die Ausgleichsfunktion die Daten nicht einhüllt und ggf. sogar gar keine Schnittmengen mit den Eingangsdaten aufweisen kann.

Jede Fuzzy-Menge $\widetilde{A} \in F_{coc}^{nob}(\mathbb{R})$[29] lässt sich durch ihre Stützfunktion $s_{\widetilde{A}} : [0,1]\times\mathbb{S}^0 \longrightarrow \mathbb{R}$ eindeutig beschreiben, wobei $\mathbb{S}^0 = \{-1,1\}$ die Oberfläche der Einheitskugel im $\mathbb{R}^1$ sei. Bei der Abbildung $\jmath : \widetilde{A} \longmapsto s_{\widetilde{A}}$ werden außerdem die Monotonie und die lineare Struktur erhalten. Die Definition und die Eigenschaften der Stützfunktion sind für die weitere Diskussion nicht zentral und sind daher im Anhang *A.1* zusammengefasst.[30] Mit der Konstruktion $\langle \widetilde{A} \,|\, \widetilde{B} \rangle \stackrel{\text{def}}{=} \langle s_{\widetilde{A}} \,|\, s_{\widetilde{B}} \rangle$ lässt sich das Skalarprodukt auf $F_{coc}^{nob}(\mathbb{R})$ übertragen, das auf $L^2([0,1]\times\mathbb{S}^0)$ bzgl. $\lambda^{[0,1]}\otimes\lambda^{\mathbb{S}^0}$ definiert ist, wobei wieder die üblichen Einschränkungen für die Linearität zu beachten sind. Diese günstigen Eigenschaften der Stützfunktion werden bei der Definition der Metrik δ_2 auf $F_{coc}^{nob}(\mathbb{R})$ ausgenutzt.

[28] Dort wird sie allerdings unter der Bezeichnung ρ_2 diskutiert.

[29] In [Diamond und Kloeden, 1994] wird zunächst die allgemeinere Menge $F_{coc}^{no}(\mathbb{R})$ der konvexen, kompakten, normalen Fuzzy-Mengen über $\mathbb{R}$ eingeführt. Zusammen mit der Definition von δ_2 wird dann eine Einschränkung auf die Teilmenge der Fuzzy-Mengen $\widetilde{A} \in F_{coc2}^{no}(\mathbb{R})$ vorgenommen, deren Stützfunktionen quadratintegrierbar sind, d.h. für die gilt $s_{\widetilde{A}} \in L^2([0,1]\times\mathbb{S}^0)$. Dies ist die übliche Grundmenge für die Kleinste Quadrate Fuzzy-Regression mit Metriken, die mittels der Stützfunktionen definiert sind. Auf diese Feinheiten wird hier verzichtet. Es gilt ohnehin $F_{coc}^{nob}(\mathbb{R}) \subset F_{coc2}^{no}(\mathbb{R})$.

[30] Auf einige der Eigenschaften werden wir allerdings bei den Berechnungen in Kapitel *5* wieder zurückgreifen.

Definition 3.3 (δ_2-Metrik) Für Fuzzy-Mengen $\widetilde{A}, \widetilde{B} \in F_{coc}^{nob}(\mathbb{R})$ sei nun die Metrik δ_2 gegeben durch

$$\delta_2^2(\widetilde{A}, \widetilde{B}) = \int_{[0,1]\times\mathbb{S}^0} (s_{\widetilde{A}} - s_{\widetilde{B}})^2 \, d\lambda^1 \otimes \lambda^{\mathbb{S}^0}. \tag{3.9}$$

Da für jede Fuzzy-Zahl $\widetilde{Z}$ und jedes $\alpha \in [0,1]$ gilt, dass $s_{\widetilde{Z}}(\alpha, -1) = -z^L(\alpha)$ und $s_{\widetilde{Z}}(\alpha, 1) = z^R(\alpha)$ erhalten wir

$$\delta_2^2(\widetilde{A}, \widetilde{B}) = \frac{1}{2} \int_0^1 [(-a^L(\alpha) + b^L(\alpha))^2 + (a^R(\alpha) - b^R(\alpha))^2] \, d\alpha \tag{3.10}$$

mit $\lambda^{[0,1]}, \lambda^{\mathbb{S}^0}$ den zu $[0,1]$ bzw. $\mathbb{S}^0$ gehörigen Borel-Lebesgue-Maßen. Außerdem sei wie oben notiert $[\widetilde{A}]^\alpha = [a^L(\alpha), a^R(\alpha)], [\widetilde{B}]^\alpha = [b^L(\alpha), b^R(\alpha)]$. Die Flächen, auf die sich die Bestimmung des δ_2-Abstands bezieht, sind in Abbildung 3.10 dargestellt. Es ist zu erkennen, dass die Fläche, um die die Gipfel der Fuzzy-Mengen voneinander verschoben sind, dadurch doppelt gewertet wird.

Satz 3.4 Für LR-Fuzzy-Zahlen $\widetilde{A} = (a; l_A, r_A)_{LR}, \widetilde{B} = (b; l_B, r_B)_{LR}$ lässt sich die folgende Berechnungsformel für δ_2 herleiten, falls $L^{-1}, R^{-1} \in L^2(\mathbb{R})$:[31]

$$\begin{aligned} \delta_2^2(\widetilde{A}, \widetilde{B}) = {} & (a-b)^2 + \|R\|_2^2 (r_A - r_B)^2 + \|L\|_2^2 (l_A - l_B)^2 \\ & + 2(a-b)\Big[\|R\|_1 (r_A - r_B) - \|L\|_1 (l_A - l_B)\Big], \end{aligned} \tag{3.11}$$

wobei

$$\begin{aligned} \|L\|_1 &= \frac{1}{2} \int_0^1 L^{-1}(\alpha) \, d\alpha, \qquad \|L\|_2^2 = \frac{1}{2} \int_0^1 (L^{-1}(\alpha))^2 \, d\alpha \quad \text{und} \\ \|R\|_1 &= \frac{1}{2} \int_0^1 R^{-1}(\alpha) \, d\alpha, \qquad \|R\|_2^2 = \frac{1}{2} \int_0^1 (R^{-1}(\alpha))^2 \, d\alpha. \end{aligned}$$

Für trianguläre Fuzzy-Mengen gilt $\|L\|_1 = \|R\|_1 = \frac{1}{4}$ und $\|L\|_2^2 = \|R\|_2^2 = \frac{1}{6}$. Mit der Definition $\delta_{2,m}^2(\widetilde{A}, \widetilde{B}) = \sum_{l=1}^m \delta_2^2(\widetilde{A}_l, \widetilde{B}_l)$ für $\widetilde{A} = (\widetilde{A}_1, \ldots, \widetilde{A}_m), \widetilde{B} = (\widetilde{B}_1, \ldots, \widetilde{B}_m) \in F_{coc}^{nob}(\mathbb{R})^m$ können in der üblichen Weise entsprechende Ergebnisse für die Menge der m-dimensionalen Fuzzy-Vektoren $F_{coc}^{nob}(\mathbb{R})^m$ hergeleitet werden. Es gilt der folgende Satz:[32]

Satz 3.5 Sei das Skalarprodukt auf $F_{coc}^{nob}(\mathbb{R})^m$ auf einen Orthanten $\mathbb{R}_{\mathbf{q}}$ beschränkt. Dann ist der metrische Raum $(F_{coc}^{nob}(\mathbb{R})^m, \delta_{2,m})$ ein abgeschlossener, konvexer Kegel im separierbaren Hilbertraum $L^2([0,1] \times \mathbb{S}^{m-1})$.

[31] Vgl. Körner und Näther 1998, S. 100.

[32] Vgl. Krätschmer 2001b, S. 5.

Für den Nachweis, dass eine minimierende Lösung für das Kleinste Quadrate-Problem (3.8) existiert, kann nun der Projektionssatz für Hilberträume herangezogen werden.[33] Dieser besagt, dass für jede konvexe, abgeschlossene Teilmenge $\varnothing \neq G \subset \mathcal{H}$ eines Hilbertraumes $\mathcal{H}$ und für jedes Element $h \in \mathcal{H}$ ein eindeutiges Element $g_0 \in G$ existiert, so dass für die zugehörige Norm $\|\cdot\| = \|\cdot\|_{\mathcal{H}}$ gilt[34]

$$\|h - g_0\| = \min\{\|h - g\| \mid g \in G\}.$$

Eine Lösung der Kleinste Quadrate Fuzzy-Regression existiert also immer dann, wenn die Menge

$$\begin{aligned}\jmath\Big(\Big\{\big(f(\widetilde{X}_{11},\ldots,\widetilde{X}_{1k};\widetilde{A}),\ldots,f(\widetilde{X}_{n1},\ldots,\widetilde{X}_{nk},\widetilde{A})\big) \mid \widetilde{A} \in \mathcal{A}\Big\}\Big) \\ = \jmath\Big(\Big\{\big(\chi_{(\widetilde{X}_{11},\ldots,\widetilde{X}_{1k})}(\widetilde{A}),\ldots,\chi_{(\widetilde{X}_{n1},\ldots,\widetilde{X}_{nk})}(\widetilde{A})\big) \mid \widetilde{A} \in \mathcal{A}\Big\}\Big)\end{aligned}$$

konvex und abgeschlossen in $L^2([0,1]\times\mathbb{S}^0)^n$ ist.[35] Damit kann die Existenz einer Lösung in den folgenden Fällen gezeigt werden:

Satz 3.6 Notwendige Bedingungen dafür, dass das Kleinste Quadrate Fuzzy-Optimierungsproblem (3.8) lösbar ist, sind:

(i) Der Approximationsparameter ist ein reellwertiger Vektor $\mathbf{a} \in \mathbb{R}^{k+1}$;

(ii) der Fuzzy-Approximationsparameter ist von der Gestalt $\widetilde{A} = (\widetilde{A}_0, a_1, \ldots, a_k) \in F_{coc}^{nob}(\mathbb{R})\times\mathbb{R}^k$ oder

(iii) die Fuzzy-Beobachtungsdaten sind von der Gestalt $(x_i, \widetilde{Y}_i)$.

Wie Krätschmer [2006a] zeigt, sind im Fall (i) die gefundenen Approximationsparameter $\mathbf{a}$ eindeutig, wenn die sog. Regressionsfunktion $X^{(n)} : \mathbb{R}^{k+1} \longrightarrow F_{coc}^{nob}(\mathbb{R})^n$, $\mathbf{a} \longmapsto (\chi_{(\widetilde{X}_{11},\ldots,\widetilde{X}_{1k})}(\mathbf{a}),\ldots,\chi_{(\widetilde{X}_{n1},\ldots,\widetilde{X}_{nk})}(\mathbf{a}))$ injektiv ist (*Identifizierbarkeit*).[36]

Ein Vorteil bei der Verwendung von δ_2 ist, dass dadurch eine σ-Algebra über $F_{coc}^{nob}(\mathbb{R})$ induziert wird, für die wesentliche Definitionen und Ergebnisse der Wahrscheinlichkeitstheorie und Statistik rekonstruiert werden können. Auf dieser Grundlage lässt sich eine Schätztheorie für die Kleinste Quadrate Fuzzy-Regression gemäß (3.8) entwickeln, für die

[33] Krätschmer [2006a] präsentiert einen detaillierten Nachweis des Satzes. Eine frühe Version des Approximationssatzes für trianguläre Fuzzy-Vektoren ist z.B. bei Diamond [1992] zu finden.

[34] Als Projektionseigenschaft ist zu sehen, dass eine notwendige und hinreichende Bedingung dafür, dass $g_0 \in G$ das minimierende Element ist, darin besteht, dass für alle $g \in G$ gilt: $\langle x - g_0, g - g_0\rangle \leq 0$.

[35] Der in [Diamond und Körner, 1997; Körner und Näther, 1998] skizzierte Nachweis für eine allgemeine Lösbarkeit von (3.8) ist nur für konstante Funktionen $f[\cdot,\widetilde{A}] \equiv \widetilde{C} \in F_{coc}^{nob}(\mathbb{R})$ geeignet.

[36] Ein Nachweis der Identifizierbarkeit für die Fälle (ii) und (iii) steht noch aus.

wesentliche Ergebnisse der klassischen Regressionsanalyse bestätigt werden können.[37] Die Schätzeigenschaften werden in Abschnitt *3.3.3* genauer diskutiert.

Bei *LR*-Fuzzy-Zahlen die durch Tupel von reellwertigen Zahlen repräsentiert werden, lässt sich das Kleinste Quadrate-Problem mit Hilfe der Darstellung (3.11) in ein quadratisches Optimierungsproblem über reellwertigen Variablenvektoren $\boldsymbol{\beta}$ überführen, so dass die Approximationsparameter in Abhängigkeit von den Variablenvektoren $\boldsymbol{\beta}$ als $\widetilde{A}(\boldsymbol{\beta}) = (\widetilde{A}_0(\boldsymbol{\beta}), \ldots, \widetilde{A}_k(\boldsymbol{\beta})) \in F_{coc}^{nob}(\mathbb{R})^{k+1}$ bestimmt werden. In [Diamond und Körner, 1997; Körner und Näther, 1998] werden Algorithmen zur Berechnung der quadratischen Optimierung präsentiert. Dabei ist zu beachten, dass bei einer Modellfunktion mit Fuzzy-Parametern zusätzliche Nebenbedingungen aufgenommen werden müssen, die sicherstellen, dass die Spannweiten der $\widetilde{A}_j(\boldsymbol{\beta})$ nicht-negativ sind.[38] Bei Modellfunktionen mit Fuzzy-Inputs ist aufgrund der Nicht-Linearitäten der Modellfunktion, die in Abschnitt *3.1.2* dargestellt wurden, für jeden der Orthanten $\mathbb{R}_{\mathbf{q}}$ ein gesondertes Optimierungsproblem aufzustellen. Die Suche nach einem optimalen Variablenvektor $\widehat{\boldsymbol{\beta}}$ kann dann jeweils immer nur beschränkt auf dem Orthanten $\mathbb{R}_{\mathbf{q}}$ durchgeführt werden. Dabei kann es vorkommen, dass es für mehrere paarweise verschiedene Orthanten $\mathbb{R}_{\mathbf{q}} \neq \mathbb{R}_{\mathbf{p}}$ lokale Lösungen $\widehat{\boldsymbol{\beta}}_{\mathbf{q}}, \widehat{\boldsymbol{\beta}}_{\mathbf{p}}$ von (3.8) gibt. Wir erhalten dann als globale Lösung der Kleinste Quadrate Fuzzy-Approximation den Parametervektor, der die Summe der Abstandsquadrate minimiert, d.h.

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\mathbf{q}} \sum_{i=1}^{n} \delta_2^2\big(\widetilde{Y}_i, f(\widetilde{X}_i, \widetilde{A}(\widehat{\boldsymbol{\beta}}_{\mathbf{q}}))\big),$$

wobei $\widetilde{A}(\widehat{\boldsymbol{\beta}}_{\mathbf{q}}) = (\widetilde{A}_0(\widehat{\boldsymbol{\beta}}_{\mathbf{q}}), \ldots, \widetilde{A}_k(\widehat{\boldsymbol{\beta}}_{\mathbf{q}})) \in F_{coc}^{nob}(\mathbb{R})^{k+1}$ der zum optimalen reellwertigen Variablenvektor $\widehat{\boldsymbol{\beta}}_{\mathbf{q}}$ für den Orthanten $\mathbb{R}_{\mathbf{q}}$ gehörige Vektor von Fuzzy-Parametern in der Fuzzy-linearen Modellfunktion ist.[39]

Die Approximationseigenschaften der Kleinste Quadrate Fuzzy-Regression lassen sich auf eine ganze Familie von Metriken $\delta_{(\nu)}$ übertragen, die auf $F_{coc}^{nob}(\mathbb{R})$ topologisch äquivalent zu δ_2 sind. Seien dazu ν ein Wahrscheinlichkeitsmaß auf $[0,1]$ und $\widetilde{A}, \widetilde{B} \in F_{coc}^{nob}(\mathbb{R})$. Dann ist die zugehörige Metrik $\delta_{(\nu)}$ definiert durch

$$\delta_{(\nu)}^2(\widetilde{A}, \widetilde{B}) = \int_{[0,1]\times\mathbb{S}^0} (s_{\widetilde{A}} - s_{\widetilde{B}})^2 \, d\nu \otimes \lambda^{\mathbb{S}^0}.$$

$\delta_{(\nu)}$ unterscheidet sich somit von δ_2 durch die Einführung einer Gewichtsfunktion bei der Integration über die α-Niveaus der Stützfunktionen $s_{\widetilde{A}}, s_{\widetilde{B}}$.[40] Ein Großteil der Metriken,

[37] Vgl. [Krätschmer, 2001b, 2004, 2006a].

[38] Aus diesem Grund wird bei Optimierung der Nachweis einer Kuhn-Tucker-Bedingung benötigt.

[39] Da bei Fuzzy-Inputs in der Regel nur Modellfunktionen mit reellwertigen Steigungsparametern $(a_1, \ldots, a_k) \in \mathbb{R}^k$ vorgesehen werden, kann in der Regel identifiziert werden $a_j = \beta_j$ für $1 \leq j \leq k$.

[40] Zur Definition von $\delta_{(\nu)}$ vgl. [Diamond und Körner, 1997]. Die Familie $\delta_{(\nu)}$ stellt auf $F_{coc}^{nob}(\mathbb{R})$ einen

die für alternative Methoden zur Kleinste Quadrate Fuzzy-Regression verwendet werden, lassen sich auf die allgemeine Form $\delta_{(\nu)}$ zurückführen, so z.B. [Bárdossy u. a., 1992; Xu und Li, 2001] für $\nu = f d\lambda$, wobei f f.s. positiv und monoton nicht-fallend sei. Die Herleitung bei [Chang, 2001] zeigt, dass die Metrik bei [Bárdossy u. a., 1992] mit $f = \mathrm{id}|_{[0,1]}$ auch als Abstandsmessung der nach der Schwerpunkmethode defuzzifizierten Fuzzy-Zahlen gesehen werden kann.[41]

Es lassen sich aber nicht alle bekannten Methoden der Kleinste Quadrate Fuzzy-Regression ohne Weiteres auf die angesprochenen Metriken $\delta_{(\nu)}$ zurückführen. Als ein Beispiel kann die von Yang und Lin [2002] verwendete Metrik für LR-Fuzzy-Zahlen gelten. Sie ist für Fuzzy-Werte $\widetilde{A} = (a; l_A, r_A)_{LR}, \widetilde{B} = (b; l_B, r_B)_{LR} \in F_{coc}^{nob}(\mathbb{R})$ definiert durch

$$d_{YL}^2(\widetilde{A}, \widetilde{B}) = [a-b]^2 + [a - \|L\| l_A - (b - \|L\| l_B)]^2 + [a + \|R\| r_A - (b + \|R\| r_B)]^2,$$

wobei $\|L\| = \int_0^1 L^{-1}(\alpha)\, d\alpha$ und $\|R\| = \int_0^1 R^{-1}(\alpha)\, d\alpha$. Bei dieser Metrik werden die quadrierten Flächen des Abstandes der Intervallgrenzen berücksichtigt, d.h. am Beispiel der linken Spannweite $\left(\int \inf[\widetilde{A}]^\alpha - \inf[\widetilde{B}]^\alpha\, d\alpha\right)^2$, anstelle der Abstandsquadrate der Intervallgrenzen, wie bei $\delta_{(\nu)}$.[42]

Es werden mehrere Strategien für den Umgang mit Fuzzy-Daten vorgeschlagen, bei denen nur eine eingeschränkte Verträglichkeit zwischen dem funktionalen Zusammenhang der Datenpunkte und dem Verhalten der Ungenauigkeitsumgebungen besteht. In [Diamond, 1992] und [Diamond und Tanaka, 1999] werden z.B. Verträglichkeitsbedingungen dafür abgeleitet, wann die vorliegende Wolke von Fuzzy-Daten die gegebene Modellfunktion gut beschreibt und das Approximationsproblem somit als „gut gestellt“ aufzufassen ist. Falls die Spannweiten der Fuzzy-Daten bei wachsendem Abstand vom Koordinatenursprung fallen, anstelle zu wachsen, kann mit Hilfe der metrischen Erweiterung der Hukuhara-Subtraktion eine Fuzzy-Funktion mit schrumpfenden Spannweiten approximiert werden. Man gelangt so zu Fuzzy-linearen Modellen mit erweiterter Differenz (engl. „*extended LR-Fuzzy linear models*“).[43]

D'Urso und Gastaldi [2000] schlagen einen anderen Weg ein, um eine Fuzzy-Funktion zu bestimmen, die die Fuzzy-Daten $(\mathbf{x}_i, \widetilde{Y}_i)$ bestmöglich approximiert. Dazu geben sie die Forderung auf, dass die Approximationsfunktion eine Fuzzy-lineare Funktion ist. Statt-

Spezialfall der Metriken auf Fuzzy-Mengen dar, die Bertoluzza u. a. [1995] vorstellen. Vgl. dazu Satz A.6 in Anhang *A.1*.

[41] In Kapitel *5* bzw. in Anhang *A.1* wird deutlich, dass der dort definierte Schwerpunkt — analog wie der Steiner-Punkt $\sigma_{\widetilde{A}}$ für die δ_2-Metrik — einen minimierenden Wert für die Metrik δ_H darstellt.

[42] In [Yang und Liu, 2003] wird nach einer ähnlichen Philosophie eine Metrik für bohnenförmige Fuzzy-Zahlen $(a, C_{\widetilde{A}})_Q$ im $\mathbb{R}^k$ hergeleitet.

[43] Vgl. [Diamond und Körner, 1997].

dessen wird eine simultane Kleinste Quadrate-Regression für die Approximationsfunktion aufgestellt, bei dem die Spannweiten der Approximationswerte $\hat{\zeta}^{L/R} = \hat{\zeta}^{L/R}(\mathbf{x})$ eine lineare Abhängigkeit von den Modalwerten $\hat{y} = \hat{y}(\mathbf{x})$ aufweisen. Dieser Zusammenhang ist modelliert als:

$$
\begin{aligned}
y_i &= \mathbf{a}^T \mathbf{x}_i + \varepsilon_{y_i}, \\
\zeta_i^L &= b^L(\mathbf{a}^T \mathbf{x}_i) + d^L + \varepsilon_{\zeta_i^L}, \\
\zeta_i^R &= b^R(\mathbf{a}^T \mathbf{x}_i) + d^R + \varepsilon_{\zeta_i^R},
\end{aligned}
$$

wobei $\widetilde{Y}_i = (y_i, \zeta_i^L, \zeta_i^R)_{LR}$ sowie $\mathbf{a} \in \mathbb{R}^k$ und $b^{L/R}, d^{L/R}, \varepsilon_{y_i}, \varepsilon_{\zeta_i^L}, \varepsilon_{\zeta_i^R} \in \mathbb{R}$. Für dieses Regressionsproblem kann dann rekursiv eine optimale Lösung berechnet werden.[44]

Die Nähe zur klassischen Kleinsten Quadrate-Regression macht deutlich, dass auch die Kleinste Quadrate Fuzzy-Regression auf Änderungen in der relativen Lage der Eingangsdaten reagiert und daher keine robuste Methode ist. Dabei kann grundsätzlich eine beliebige Defuzzifizierung der Fuzzy-Daten als Lagecharakteristik betrachtet werden. Aufgrund ihrer Eigenschaft als minimierende Werte für die δ_2-Metrik, vgl. Satz A.8 im Anhang, empfehlen sich aber vor allem die Steiner-Punkte $\sigma_{\widetilde{A}}$ als Defuzzifizierung. Die Approximation ist ebenfalls sensitiv für Änderungen in der Unschärfecharakteristik, also den Spannweiten der Fuzzy-Daten. Bei der Approximation von physikalischen Fuzzy-Daten ist dies als erwünschter Effekt zu werten, weil somit die gesamte Information aus den Fuzzy-Datenpunkten bei der Approximation zu Geltung kommt. Bei epistemischen Fuzzy-Daten kann eine hohe Sensitivität für die Ränder der Trägermengen hingegen unerwünscht sein.

3.2.3 Weitere Ansätze

Neben den beiden großen Methodenklassen, die bisher vorgestellt wurden, gibt es weitere Ansätze zur Regression bei Fuzzy-Daten, die sich nicht oder nicht eindeutig einer der beiden Gruppen zuordnen lassen.

Solymosi und Dombi [1992] interpretieren Fuzzy-Daten $(\widetilde{X}_i, \widetilde{Y}_i) \in F_{coc}^{nob}(\mathbb{R})^2$ als Fehlerumgebungen von reellwertigen Messwerten und bestimmen dazu aus der Schar der Funktionen, die alle Fehlerumgebungen schneiden, diejenige, die aggregiert über alle Fehlerumgebungen die höchste Zugehörigkeit hat („*curve fitting problem*“). Die Güte der Anpassung wird dabei durch das Maximum der Schnittmenge jeder Fehlerumgebung mit der Fuzzy-Erweiterung der Anpassungsfunktion auf der Fehlerumgebung bestimmt, also

[44] Ein Beweis für die Konvergenz des Verfahrens wird allerdings nicht erbracht.

durch

$$\alpha_i(\mathbf{a}) = \sup\{\min\{\mu_{\widetilde{Y}_i}(y), \mu_{f(\widetilde{X}_i;\mathbf{a})}(y)\} \mid y \in \mathbb{R}\}. \tag{3.12}$$

Dabei sei $\mathbf{a} = (a_0, a_1) \in \mathcal{A} \subset \mathbb{R}^2$.

Diese Maßzahl wird auch als Possibility[45] für $\widetilde{Y}_i = f(\widetilde{X}_i; \mathbf{a})$ bezeichnet. Gesucht ist der Parameter $\widehat{\mathbf{a}} \in \mathcal{A}$, der bei Zusammenfassung über alle Daten $(\widetilde{X}_i, \widetilde{Y}_i)$ maximale Possibility für die Ähnlichkeit mit den Messdaten hat. Es soll folglich der Parameter bestimmt werden, für den gilt

$$\forall\, \mathbf{a} \in \mathcal{A}: \quad min_{1\le i\le n}\alpha_i(\mathbf{a}) \le min_{1\le i\le n}\alpha_i(\widehat{\mathbf{a}}). \tag{3.13}$$

Unter bestimmten Stetigkeitsbedingungen kann eine Lösung für das Problem mit numerischen Methoden approximativ bestimmt werden. Als Ergebnis erhalten wir einen zufriedenstellenden Parameter $\widehat{\mathbf{a}}$ zusammen mit einem Zugehörigkeitswert $\mu_{\widetilde{\mathbb{L}}}(\widehat{\mathbf{a}}) = \min_{1\le i\le n} \alpha_i(\widehat{\mathbf{a}})$ zur Fuzzy-Menge der möglichen Lösungen $\widetilde{\mathbb{L}}$. Der Funktionsanpassung liegt ein Konstruktionsprinzip zugrunde, bei dem eine vorgegebene parametrisierte Funktionenmenge anhand der Fuzzy-Eingangsdaten auf ihre Plausibilität überprüft wird. Dabei wird die Plausibilität einer Einzelfunktionen in der Zugehörigkeit ihres Parameterwerts zur Parametermenge abgebildet. Der funktionale Zusammenhang kann hier also mit einem Fuzzy-Bündel von Funktionen identifiziert werden. Für die Übertragung der Fuzziness aus den Daten auf die Parametermenge ist immer ein Aggregrationsoperator — hier Gleichung (3.13) —, der die Abhängigkeit von den Messungen $1 \le i \le n$ eliminiert, mit einem Integrationsoperator — hier Gleichung (3.12) —, der die Abhängigkeit von $(x, y) \in \mathbb{R}^2$ eliminiert, zu verknüpfen.

Diese „*transfer principles*“ werden u.a. von Bandemer und Näther [1992] ausführlich beschrieben.[46] Die Fuzzy-Parametermenge $\widetilde{\mathbb{L}}$, die sich als Ergebnis der Übertragungsprinzipien ergibt, ist auch unter den geforderten Stetigkeitsbedingungen im allgemeinen nicht normalisiert und lässt sich in der Regel auch nicht mit elementaren Referenzfunktionen, wie z.B. linearen oder quadratischen Referenzfunktionen, darstellen. Darüber hinaus existiert nur dann eine Lösung, wenn die Fuzzy-Daten auf dem funktionalen Zusammenhang liegen. Insgesamt stellen die Übertragungsprinzipien eher einen Ansatz zur Funktionsinterpolation bei Fuzzy-Daten dar und gehören nicht im engeren Sinne zur Fuzzy-Regression.

[45] Vgl. dazu auch Abschnitt *3.2.1*, S. 88.

[46] In weiteren Aufsätzen wird herausgearbeitet, welches der Übertragungsprinzipien für bestimmte Fragestellungen geeignet ist. Vgl. [Bandemer und Kudra, 1988; Kudra, 1990] sowie den Methodenüberblick bei [Näther, 2006].

Eine wegweisende Methode zur Fuzzy-Regression wurde Ende der 80er Jahre von Celmiņš ausgearbeitet.[47] Sie ist als quasi-metrischer Kleinste-Quadrate-Ansatz zu betrachten. Die Methode ist ausschließlich auf den Fall bohnenförmiger Fuzzy-Daten ausgelegt, die im allgemeinen auch interaktiv sind. Zu jedem Parameter werden Schlupfwerte $c_{ij,a}$ eingeführt, mit denen der funktionale Zusammenhang der Modellfunktion erfüllt ist. D.h. für die gilt:

$$f([\widetilde{X}_i]^1 + (c_{i1,a}, \ldots, c_{ik,a})^T; \mathbf{a}) - ([\widetilde{Y}_i]^1 + c_{i0,a}) = 0.$$

Gesucht ist derjenige Parameter $\widehat{\mathbf{a}}$ für den die Summe der quadratischen Abweichungen

$$d_C^2([(\widetilde{X}_i, \widetilde{Y}_i)]^1, [(\widetilde{X}_i, \widetilde{Y}_i)]^1 + (c_{i1,a}, \ldots, c_{ik,a}, c_{i0,a})^T) = \left(1 - \mu_{(\widetilde{X}_i, \widetilde{Y}_i)}\left([(\widetilde{X}_i, \widetilde{Y}_i)]^1 + (c_{i1,a}, \ldots, c_{ik,a}, c_{i0,a})^T\right)\right)^2$$

in den verschobenen Beobachtungen minimal ist. Für bohnenförmige Fuzzy-Daten $\widetilde{Z}_i = (\widetilde{X}_i, \widetilde{Y}_i)$ gibt es eine positiv semidefinite Matrix A_{Z_i} und einen Modalwert $z_i = [(\widetilde{X}_i, \widetilde{Y}_i)]^1 \in \mathbb{R}^{k+1}$, so dass $\mu_{\widetilde{Z}_i}(z) = 1 - \min\{1, \sqrt{(z_i - z)^T A_{Z_i}(z_i - z)}\}$. Folglich ist $\|\cdot\|_{A_{Z_i}} = 1 - \mu_{\widetilde{Z}_i}$ eine Quasinorm auf $\mathbb{R}^{k+1}$. Die resultierende Approximationsfunktion wird durch zwei Hyperbel-Äste gebildet, so dass der Funktionswert mit der geringsten Unschärfe nicht notwendig 0 ist. Der Ansatz wird nicht weiter verfolgt, da er auf den Fall bohnenförmiger Fuzzy-Daten und der dazugehörigen Quasinorm beschränkt ist. Ein bedenkenswerter Aspekt der Regression nach Celmiņš ist vor allem die relative Gütebewertung der Approximationsfunktion mit der Zugehörigkeit zu den Fuzzy-Messwerten. Daraus können weitere Güteindizes abgeleitet werden, wie z.B. der *Zusammenhalt* (engl. „*grade of collocation*“) zwischen zwei Fuzzy-Mengen $\widetilde{A}$ und $\widetilde{B}$, der gerade der Höhe der Schnittmenge $\widetilde{A} \cap \widetilde{B}$ entspricht.

Eine Designmethode im Grenzbereich zu den metrischen Ansätzen haben Kao und Chyu [2002] für Fuzzy-Daten $(\widetilde{X}_i, \widetilde{Y}_i) \in F_{coc}^{nob}(\mathbb{R})^{k+1}$ mit $\widetilde{X}_i = (\widetilde{X}_{i1}, \ldots, \widetilde{X}_{ik})$ vorgestellt. Die Daten werden zunächst defuzzifiziert zu $(\overline{\mathbf{x}}_i, \overline{y}_i)$, z.B. nach dem Schwerpunktverfahren, wobei $\overline{\mathbf{x}}_i = (\overline{x}_{i1}, \ldots, \overline{x}_{ik})$ der Vektor der Schwerpunkte von $\widetilde{X}_i$ und $\overline{y}_i$ der Schwerpunkt von $\widetilde{Y}_i$ ist. An die defuzzifizierten Werte $(\overline{\mathbf{x}}_i, \overline{y}_i)$ wird eine klassische Regressionsgerade approximiert (*Lageapproximation*), so dass wir reellwertige Steigungsparameter $\widehat{\mathbf{a}} = (\widehat{a}_0, \ldots, \widehat{a}_k) \in \mathbb{R}^{k+1}$ erhalten. Im zweiten Schritt wird der Fuzzy-Achsenabschnitt $\widehat{\widetilde{A}}_0$ so bestimmt, dass $[\widehat{\widetilde{A}}_0]^1 = \widehat{a}_0$ und dass für den Approximationswert $\widehat{\widetilde{Y}}_i = \widehat{\widetilde{A}}_0 + \sum_j \widehat{a}_j \widetilde{X}_{ij}$ gilt,

[47] Vgl. [Celmiņš, 1992, 1987a,b].

dass die Summe über die Mengendifferenzen

$$D_i = \int_{\mathrm{supp}(\widetilde{Y}_i)\cup\mathrm{supp}(\widehat{\widetilde{Y}}_i)} |\mu_{\widetilde{Y}_i}(y) - \mu_{\widehat{\widetilde{Y}}_i}(y)|\, dy,$$

die bereits in (3.6) eingeführt wurden, minimal ist (*Unschärfeapproximation*). $\widetilde{E} = \widehat{\widetilde{A}}_0 - \widehat{a}_0$ kann als globale Fehlerabweichung interpretiert werden. Dabei ist darauf hinzuweisen, dass eine wohldefinierte Lösung für die hypothetische Residuengleichung $\widetilde{e}_i = \widetilde{Y}_i - (\widehat{a}_0 + \sum_j \widehat{a}_j \widetilde{X}_{ij})$ für die Fuzzy-Arithmetik im allgemeinen nicht existiert.[48] Die Methode kann nur für Modellfunktionen mit Fuzzy-Parametern $(\widetilde{A}_0, a_1, \ldots, a_k)$ angewendet werden und führt zu einem ähnlichen Ergebnis wie die Kleinste Quadrate Fuzzy-Regression. Allerdings wird gemäß $\sum_i D_i$ die bessere Abdeckung erreicht,[49] so dass das Ergebnis dem intuitiven Verständnis der Approximation von Fuzzy-Ungenauigkeitsumgebungen in der epistemischen Interpretation sehr nahe kommt.

Eine weitere Variante stellt die Fuzzy-Regression nach Tran und Duckstein [2002] dar, für die eine Mehrzieloptimierung durchgeführt wird, die als Gütekriterien bei der Optimierung sowohl die Abstandsminimierung in einer Intervallmetrik als auch die Unschärfeanpassung nach dem Inklusionsansatz der Possibilistischen Regression auf einem gegebenen Anspruchsniveau κ berücksichtigt. Zusätzlich wird noch eine Kompensation für Ausreißer zugelassen. Es handelt sich dabei also um eine hybride Fuzzy-Regression. Aufgrund der Kompromissbildung in der Lösung ist es allerdings kaum möglich, die Approximationsfunktion inhaltlich zu interpretieren.

Von den hier beschriebenen Ansätzen werden wir in den folgenden Abschnitten nur noch auf die Funktionsevaluation zurückgreifen. Diese ist ein Beispiel für die Bestimmung des Zusammenhangs zwischen Fuzzy-Messdaten und der Menge der Approximationsparameter. In der klassischen Regression ist ein solcher Zusammenhang vor allem bei der Bestimmung eines Parameterkonfidenzintervalls relevant.

[48] Einen Ausweg bietet hier u.a. die metrische Vervollständigung der Hukuhara-Subtraktion, wie sie in [Diamond und Körner, 1997] bei der Konstruktion von Fuzzy-linearen Modellen mit erweiterter Differenz hergeleitet wird.

[49] Dies ist dem Gütevergleich für Diamond-, Kim-Bushu- sowie Kao-Chyu-Regression in [Kao und Chyu, 2002] zu entnehmen (S. 405).

3.3 Empirische Anwendung der Fuzzy-Regression — ein kritischer Methodenvergleich

Häufig bleibt bei der Darstellung von Fuzzy-Regressionsmethoden unklar, welche Ziele die Autor/innen mit der Methode verfolgen und wie die Ergebnisse interpretiert werden können. Dies gilt ganz besonders für Designansätze wie die Possibilistischen Regression und davon abgeleitete Ansätze. Es ist festzustellen, dass es gerade bei diesen Ansätzen häufig keine eindeutigen Modellvorstellungen über die Fuzziness in den Daten gibt. Wenn die Annahmen über die Fuzziness in den Daten und über die Fuzziness des Strukturzusammenhangs aber nicht sauber unterschieden werden, ist keine schlüssige Interpretation der Approximationsfunktion möglich. Es kann daher nicht verwundern, dass die Methoden noch keinen Eingang in die empirische Datenanalyse gefunden haben. Aber auch bei den Kleinste Quadrate-Ansätzen zur Fuzzy-Regression fehlt es an einer tiefergehenden Auseinandersetzung damit, wie die empirischen Fragestellungen in der Modellfunktion abgebildet werden können und nach welchen Kriterien ein geeigneter Modellansatz auszuwählen ist. Insbesondere bleibt daher offen, welche Anforderungen an die Interpretation der Kleinste Quadrate Fuzzy-Regression zu stellen sind, wenn die Abschätzungen, die in den Fuzzy-Daten abgebildet sind, eher unscharf sind und zudem nicht von einer Homogenität der Abschätzung in den Beobachtungswerten ausgegangen werden kann. Diese Aspekte sind für die Modellierung eines Approximations- oder Schätzmodells für Fehler in den Daten noch genauer zu untersuchen.

In diesem Abschnitt wird der Versuch unternommen, die Analysekriterien für die Fuzzy-Regression neu zu ordnen und konsequent auszuarbeiten. Dabei werden die Fuzzy-linearen Funktionen in den Mittelpunkt der Modellbildung gestellt. Ausgehend von den Methoden zur Fuzzy-Regression, die in den vorangehenden Abschnitten vorgestellt wurden, werden dann verschiedene Interpretationsansätze durchgespielt, um schließlich konkrete Analysekriterien abzuleiten. Dabei wird in der üblichen Weise grundsätzlich zwischen den explorativen Ansätzen mit reellwertigen Parametern und mit Fuzzy-Parametern in den Abschnitten *3.3.1* und *3.3.2*, sowie den Ansätzen mit induktive Struktur- und Verteilungsannahmen in Abschnitt *3.3.3* unterschieden. Ziel der Systematisierung ist es, die Definition der „Fuzzy-Regression“ weiter zu präzisieren und darüber hinaus die vorhandenen Methoden zur Fuzzy-Regression nach ihrer Eignung für die Datenanalyse zu bewerten.

Bei der Formulierung der Fuzzy-Regressionsmodelle gehen wir davon aus, dass Fuzzy-Daten $(\widetilde{X}_i, \widetilde{Y}_i) \in F_{coc}^{nob}(\mathbb{R})^{k+1}$ für Messvorgänge $1 \leq i \leq n$ vorliegen, deren funktionaler Zusammenhang analysiert werden soll. Bei der verallgemeinerten Betrachtung wird grundsätzlich vernachlässigt, wie diese Daten zustande gekommen sind und was sie beinhalten. Fuzzy-Daten geben in dieser Sicht ganz allgemein eine verschmutzte Datenlage wieder, wo-

bei angenommen wird, dass die Daten den Wissensstand über die Datenlage bestmöglich abbilden.

Wir führen darüber hinaus noch eine zusätzliche Dimension in die Modelldiskussion ein, nämlich die Unterscheidung in physikalische Fuzzy-Daten, als scharfe Beschreibung des abgebildetes vagen Zustandes, und in epistemische Fuzzy-Daten, als unscharfe Abschätzung des vorliegenden Zustandes, die zudem Wahrnehmungsschwankungen unterliegen kann. Der Mengencharakter der epistemischen Fuzzy-Daten zieht nach sich, dass die Fuzziness des Fuzzy-linearen Bildes auf der rechten Seite der Modellgleichung im Sinne einer Abschätzung nach oben zu interpretieren ist, da die erweiterten Operationen keine Kompensation berücksichtigen, sondern alle möglichen Kombinationen voll berücksichtigen. Diese vorsichtige Abschätzung ist mit der Fuzzy-Abschätzung für den „wahren“ Wert für die abhängige Variable ins Verhältnis zu setzen. Der „wahre“ Outputwert ist also sowohl innerhalb der Fuzzy Outputs als auch innerhalb der zulässigen Funktionswerte zu verorten. Geeignete Approximationsverfahren können daher entweder durch Maximieren der Schnittmenge $\widetilde{Y}_i \cap f(\widetilde{X}_i, \widetilde{A})$[50] oder durch Minimieren eines metrischen Abstands $d(\widetilde{Y}_i, f(\widetilde{X}_i, \widetilde{A}))$ gebildet werden. Die Unterschiede, die die Interpretationsperspektiven nach sich ziehen, sind ein wichtiger Ausgangspunkt für das Benchmarking der Fuzzy-Regression bei Fuzzy-Daten, mit denen ungenaue Daten repräsentiert werden, in Kapitel *5*.

3.3.1 Defuzzifizierter Ausgleich

Beim defuzzifizierten Ausgleich wird an die vorliegenden Fuzzy-Daten eine erweiterte Fuzzy-Funktion angepasst, deren Parameter alle reellwertig sind. Wir haben den folgenden Modellansatz.

Sei $\mathcal{A} \subset \mathbb{R}^{k+1}$ eine Menge reeller Vektoren. Dann sei die Menge der interessierenden Funktionen gegeben durch

$$f(\cdot, \mathbf{a}) : F_{coc}^{nob}(\mathbb{R})^k \longrightarrow F_{coc}^{nob}(\mathbb{R}),\ f(\widetilde{X}, \mathbf{a}) = a_0 \oplus \bigoplus_{j=1}^{k} a_j \odot \widetilde{X}_j \qquad \text{mit } \mathbf{a} \in \mathcal{A}.$$

Der Modellansatz führt zu einer Defuzzifizierung der Fuzzy-Daten in der Approximationsfunktion. Ein wesentlicher Vorteil der Bestimmung einer defuzzifizierten Ausgleichsgeraden liegt darin, dass diese einen direkten Vergleich mit den Ergebnissen einer Regressionsrechnung auf der Basis von reellwertigen Daten zulässt.

[50] Z.B. durch Maximieren der Schnittfläche, oder durch durch Maximieren der Höhe des Schnitts, falls es eine Schnittmenge gibt.

Da keine weiteren Annahmen an die Struktur der Daten gemacht werden, können mit der Ausgleichsfunktion nur lineare Zusammenhänge innerhalb der aktuell vorliegenden Fuzzy-Daten identifiziert werden. Die Anpassungsgerade selber liefert daher zunächst nur Aussagen über den Zusammenhang zwischen den vorliegenden Fuzzy-Inputs und -Outputs. Aussagen über den Zusammenhang zwischen den dahinter verborgenen „wahren“ Werten können demnach nicht ohne Weiteres direkt aus den Approximationsparametern abgelesen, sondern immer nur mittelbar aus der Abbildung der zugehörigen Fehlerumgebungen abgeleitet werden. Es ist zu beachten, dass die Ausgleichsfunktion den Zusammenhang in den Daten gemäß den technischen Eigenschaften der verwendeten Distanzfunktion charakterisiert. Anders als bei den Strukturmodellen in Abschnitt *3.3.3* ist hier also eine Interpretation des Approximationsergebnisses nicht methodenunabhängig möglich, sondern ist immer auf das jeweilige Anpassungsfunktional zu beziehen.

Als Methoden zum defuzzifizierten Ausgleich von Fuzzy-Daten stehen einerseits die Übertragungsprinzipien zur Evaluation von reellwertigen Funktionen mit Fuzzy-Daten aus Abschnitt *3.2.3* und andererseits die metrischen Ansätze zur Kleinste-Quadrate Fuzzy-Regression zur Verfügung.

Bei der Funktionsevaluation werden die Fuzzy-Daten als Fehlerumgebungen betrachtet, die den möglichen Definitionsbereich begrenzen und eine Genauigkeitsbewertung der „wahren“ Werte abbilden. Sie entsprechen somit nahezu idealtypisch der Vorstellung von epistemischen Fuzzy-Daten. Zu einem gegebenen Anspruchsniveau κ kann dann jeweils eine Menge von Parametern bestimmt werden, für die die aus den Daten übertragene Zugehörigkeit mindestens κ ist. Die Anpassungsfunktionen queren die Gesamtheit der Fuzzy-Daten mindestens mit der vorgegebenen Plausibilität. Das Anpassungsproblem ist aber bei den meisten Verfahren[51] nur dann lösbar, wenn die gemessenen Zustände keiner oder nur einer sehr geringen Variabilität unterliegen, d.h. wenn der zugrundeliegende Zusammenhang fast deterministisch ist. Anders ausgedrückt muss die Messungenauigkeit wesentlich größer als die Variabilität der Beobachtungen sein. Methodisch wird zunächst eine Zugehörigkeits-Bewertung von reellwertigen Funktionen in Gestalt eines Fuzzy-Bündels von Funktionen vorgenommen, aus denen anschließend eine günstige Funktion ausgewählt werden kann, z.B. wie bei Solymosi und Dombi [1992] durch Intervallschachtelung zur Maximierung des Zugehörigkeitswertes.

Demgegenüber erhalten wir bei der Kleinste Quadrate Fuzzy-Regression optimale Lösungen, für die keine gesonderte Zugehörigkeitsbewertung vorliegt. Bei der Approximation wird die Abstandssumme zwischen den Fuzzy-Werten der abhängigen Variablen und den erweiterten Funktionswerten der Fuzzy-Inputs minimiert. Die reellwertige Approxi-

[51] Es gibt auch Übertragungsprinzipien, für die immer mindestens eine Lösung existiert, vgl. z.B. Bandemer und Näther 1992, S. 192ff.

mationsfunktion beschreibt den funktionalen Zusammenhang zwischen den Bewertungs- bzw. den Ungenauigkeitsumgebungen, die durch die Fuzzy-Daten abgebildet werden. Dabei setzt die Anpassung beim Punktcharakter der Fuzzy-Mengen an und betrachtet die Fuzzy-Zahlen als Gesamtheit. Der defuzzifizierte Kleinste Quadrate Fuzzy-Ausgleich ist daher besonders gut für physikalische Fuzzy-Daten geeignet und liefert dann eine Approximationsfunktion im üblichen Sinne. Bei einer Kleinste Quadrate-Fuzzy-Regression von epistemischen Fuzzy-Mengen werden die subjektiven Bewertungsumgebungen bei der Approximation ebenso als Punkte behandelt. Die subjektive Bewertung eines möglichen Wertes geht dabei mit einer Gewichtung in die Approximation ein, die für jede Fuzzy-Metrik spezifisch ausfällt. Bei der δ_2-Metrik werden etwa die Integrale über die Abweichungsquadrate der linken und der rechten Seiten der α-Niveaus betrachtet. Relevant ist somit vor allem der Anstieg des Übergangs in den Zugehörigkeiten und nicht die Zugehörigkeitsbewertung eines einzelnen Punktes. Während bei der δ_2-Metrik alle α-Niveaus mit demselben Gewicht berücksichtigt werden bei den Metriken vom Typ $\delta_{(\nu)}$ abweichende Gewichte vorgegeben, mit denen die α-Niveaus in die Integration eingehen. Die Approximationsfunktion kann daher nur in Abhängigkeit von der Modellierung der Fuzzy-Daten und deren Aussagekraft interpretiert werden.[52]

Für die Verwendung der metrischen Ansätze und insbesondere der Kleinste Quadrate Fuzzy-Regression spricht, dass damit auch bei volatilen Fuzzy-Daten eine Approximationsfunktion bestimmt werden kann. Außerdem kann die Funktionsapproximation mit der üblichen Terminologie durchgeführt und untersucht werden. Dabei ist die Funktionsapproximation so konstruiert, dass sie eine Erweiterung der klassischen Kleinste Quadrate-Approximation bei reellwertigen Daten darstellt. Darüber hinaus können für die δ_2-Metrik wesentliche Ergebnisse der statistischen Regression auch für die Fuzzy-Regression nachgewiesen werden, so dass daraus leicht Hypothesen zur Konstruktion von ökonometrischen Modellen gewonnen werden können.

Insgesamt kommen für die defuzzifizierte Approximation bei Fuzzy-Daten hauptsächlich die metrischen Methoden der Kleinst Quadrate Fuzzy-Regression in Frage. Dabei ist bei Vorliegen von Fuzzy-Ungenauigkeitsumgebungen in der epistemischen Interpretation besonders zu betrachten, welche Metriken bei welchem Mess- und Modellierungsansatz besonders geeignet sind.

[52] Die genaue Art der Abhängigkeit wird dann im Fall von Fuzzy-Fehlern in Kapitel *5* untersucht.

3.3.2 Approximation einer Fuzzy-Charakteristik

Beim Modellansatz der Approximation einer Fuzzy-Charakteristik wird eine Fuzzy-lineare Modellfunktion an die Daten angepasst, so dass die Unschärfe in den Fuzzy-Daten sich auch in der Fuzziness der Parameter niederschlägt. Im Unterschied zum Ansatz in Abschnitt *3.3.1* wird die Fuzziness in den Messwerten der unabhängigen Variablen also nicht nur proportional mittels des Erweiterungsprinzips umgebrochen, sondern es wird eine darüber hinausgehende Zunahme der Fuzziness durch die Abbildung berücksichtigt.

Sei dazu $\mathcal{A} \subset F_{coc}^{nob}(\mathbb{R})^{k+1}$ eine Menge von Fuzzy-Vektoren. Dann sei die Menge der interessierenden Funktionen für jedes $\widetilde{A} = (\widetilde{A}_0, \ldots, \widetilde{A}_k) \in \mathcal{A}$ gegeben durch

$$f(\cdot, \widetilde{A}) : F_{coc}^{nob}(\mathbb{R})^k \longrightarrow F_{coc}^{nob}(\mathbb{R}),\ f(\widetilde{X}, \widetilde{A}) = \widetilde{A}_0 \oplus \bigoplus_{j=1}^{k} \widetilde{A}_j \odot \widetilde{X}_j.$$

Die Approximation einer Fuzzy-linearen Funktion mit Fuzzy-Parametern lässt sich einerseits aus rein technischen Gründen motivieren, wenn die Fuzzy-Daten mit einer linearen Funktion mit Fuzzy-Parametern bestmöglich approximiert werden können. Wie in Abschnitt *3.1* gezeigt wurde, ist dies dann der Fall, wenn die Spannweiten der Werte der abhängigen Variablen global gesehen größer sind als die Spannweiten der Fuzzy-Werte der unabhängigen Variablen nach der linearen Transformation gemäß dem Erweiterungsprinzip, die den Lagezusammenhang der Fuzzy-Daten charakterisiert.

Man stelle sich dazu etwa die folgende Situation vor, wenn die defuzzifizierten Fuzzy-Daten mit der Gerade $y = 0,5x + 1$ approximiert werden können, aber für die Spannweiten die folgenden funktionalen Zusammenhänge gelten:

$$l_Y = l_X \quad \text{bzw.} \quad r_Y = 2r_X.$$

Die Fuzziness der Approximationsfunktion kann in diesem Fall als „unerklärter additiver Rest“ der Abbildung der unabhängigen Variablen $X_1, \ldots, X_k$ auf die abhängige Variable Y betrachtet werden, der durch die gegenüber der Werteproportionalität erhöhte Fuzziness in Y induziert wird.[53]

Für eine Approximation nach diesem Ansatz eignen sich insbesondere die Methoden zur metrischen bzw. zur Kleinste Quadrate Fuzzy-Regression mit Fuzzy-Parametern. Da Fuzzy-Steigungsparameter nur dann auftreten, wenn die Unschärfe von Y mit wachsendem

[53] Dazu ist allerdings zu bemerken, dass mit der Zahl der Regressoren die eingebrachte Unschärfe bzw. Gesamtspannweite in der Tendenz ansteigt, so dass ab einer gewissen Dimension der Modellgleichung der beschriebene Fuzzy-Überschuss in der Praxis vermutlich nicht mehr auftreten wird. Hierauf hat die Wahl der Merkmalsskalierungen einen relevanten Einfluss.

Abstand des Variablenvektors $(X_1, \ldots, X_k)$ von 0 wächst, kann man sich häufig auf eine Modellfunktion mit Fuzzy-Parametern $(\widetilde{A}_0, a_1, \ldots, a_k) \in F_{coc}^{nob}(\mathbb{R}) \times \mathbb{R}^k$ beschränken. Für die Approximation wird dann eine *vereinfachte Fuzzy-Modellfunktion*

$$\widetilde{Y} = \widetilde{A}_0 \oplus \bigoplus_{j=1}^{k} a_j \odot \widetilde{X}_j \tag{3.14}$$

verwendet. Dabei nimmt die Fuzzy-Inhomogenität $\widetilde{A}_0$ die „zusätzliche Grundunschärfe in Y“ auf.

Bemerkung 3.7 *Eine Interpretation der Fuzziness im Ergebnis als „zusätzliche Grundunschärfe“ ist möglich, wenn die Approximationsfunktion in eine einfache lineare Approximationsfunktion und einen Fuzzy-linearen Rest dekomponiert werden kann. Es ist also zu fragen, wann eine solche Dekomposition konstruiert werden kann.*

Sei dazu ein Defuzzifizierungsoperator $O_F : F_{coc}^{nob}(\mathbb{R})^{k+1} \longrightarrow \mathbb{R}^{k+1}$ gegeben, der bzgl. der Addition homomorph ist, d.h. es gelte $O_F(\widetilde{A} \oplus \widetilde{B}) = O_F(\widetilde{A}) + O_F(\widetilde{B})$. Ähnlich wie bei der Definition der zufälligen Störung mit Erwartungswert 0 kann dann im univariaten Modell die Dekomposition folgendermaßen vorgenommen werden:

$$\widetilde{Y} = \left[O_F(\widetilde{A}_0) \oplus O_F(\widetilde{A}_1) \odot \widetilde{X}\right] \oplus \left[(\widetilde{A}_0 - O_F(\widetilde{A}_0)) \oplus (\widetilde{A}_1 - O_F(\widetilde{A}_1)) \odot \widetilde{X}\right].$$

Diese Darstellung soll mit der einfachen Modellgleichung identifiziert werden

$$\widetilde{Y} = \widetilde{A}_0 \oplus \widetilde{A}_1 \odot \widetilde{X}.$$

Da $O_F(\widetilde{A}_0) \in \mathbb{R}$, gilt offensichtlich $O_F(\widetilde{A}_0) \oplus (\widetilde{A}_0 - O_F(\widetilde{A}_0)) = \widetilde{A}_0$. Sei $\alpha \in (0, 1]$ und $o_1 = O_F(\widetilde{A}_1)$ gesetzt. Für die Defuzzifizierung o_1 von $\widetilde{A}_1$ gilt, dass $a_1^L(\alpha) - o_1 \leq 0 \leq a_1^R(\alpha) - o_1$. Daher können wir mit Satz 2.19 für den multiplikativen Teil des Ansatzes die Operationen auf der Ebene der α-Schnitte berechnen.

1. Fall: $o_1 > 0$.

$$\begin{aligned}
&o_1 \odot [x^L(\alpha), x^R(\alpha)] \oplus ([a_1^L(\alpha), a_1^R(\alpha)] - o_1) \odot [x^L(\alpha), x^R(\alpha)] \\
&= [o_1 x^L(\alpha), o_1 x^R(\alpha)] \oplus [a_1^L(\alpha) - o_1, a_1^R(\alpha) - o_1] \odot [x^L(\alpha), x^R(\alpha)] \\
&= [o_1 x^L(\alpha), o_1 x^R(\alpha)] \oplus [(a_1^L(\alpha) - o_1) x^R(\alpha), (a_1^R(\alpha) - o_1) x^R(\alpha)] \\
&= [a_1^L(\alpha) x^R(\alpha) + o_1(x^L(\alpha) - x^R(\alpha)), a_1^R(\alpha) x^R(\alpha) + o_1(x^R(\alpha) - x^R(\alpha))]
\end{aligned}$$

Das ist identisch mit $[\widetilde{A}_1 \odot \widetilde{X}]^\alpha$ genau dann wenn $o_1 = 0$ (und damit auch $a_1^L(0) \leq 0 \leq a_1^R(0)$) oder $x^L(\alpha) = x^R(\alpha)$.

2. Fall: $o_1 < 0$.

Die Identität ist wiederum nur dann erfüllt, wenn $o_1 = 0$ *oder* $x^L(\alpha) = x^R(\alpha)$.

Zusammengefasst wird deutlich, dass eine solche Dekomposition nur für Fuzzy-Daten vom Typ $(x_i, \widetilde{Y}_i)$ *existiert, also wenn alle Inputs reellwertig sind. Außerdem existiert immer eine einfache Dekomposition bei der Approximation mit vereinfachten Modellfunktionen wie in Gleichung* (3.14).

Allerdings ist es bei Fuzzy-Fehlerumgebungen häufig nicht sinnvoll, einen funktionalen Zusammenhang zwischen den Unschärfeumgebungen der Inputs und den Unschärfeumgebungen der Outputs zu unterstellen, beispielsweise wenn der Einfluss auf die Genauigkeit der Messungen sich sprunghaft ändert. Sinnvoll ist dies in den wenigen Fällen, in denen sich die Datengenauigkeit tatsächlich mit wachsenden bzw. fallenden Messwerten kongruent oder überproportional zur Steigung der Anpassungsfunktion verschlechtert. In solchen Fällen kann die Art des Unschärfezusammenhangs mit einer Kleinste Quadrate Fuzzy-Regression quantifiziert werden. Insbesondere kann es bei Fuzzy-Daten $(x_i, \widetilde{Y}_i)$ mit reellwertigen Inputs wünschenswert sein, den funktionalen Zusammenhang für die Messungenauigkeit der Outputs in der Approximationsfunktion mit zu berücksichtigen.

Dem gegenüber ist es bei physikalischen Daten grundsätzlich wünschenswert, die Anpassung der Modellfunktion durch die Verwendung von Fuzzy-Parametern ggf. noch zu verbessern, da hierbei ein vager Begriff „genau“ beschrieben wird.

Ein alternativer Ansatz besteht darin, Abweichungen in der Zurechenbarkeit der gemessenen Fuzzy-Daten zur Modellfunktion zuzulassen, d.h. die Abweichung von einer klassischen, reellwertigen Modellfunktion wird mit einem geringeren Zugehörigkeitsgrad zum Funktionszusammenhang bewertet.[54] Wenn wir uns die Datenwerte entlang der reellwertigen Modellfunktion auf die y-Achse projiziert vorstellen, kann dies als Dämpfung im Informationsgewinn bei einer Häufung von Daten interpretiert werden[55], die als Fuzziness der Approximationsfunktion abgebildet wird. Formal wird dabei ein Fuzzy-Bündel von parallelen linearen Funktionen an die Fuzzy-Daten angepasst, das in Form einer Fuzzy-linearen Funktion dargestellt wird.

[54] Auch dieser Interpretationsansatz wurde bereits im Zusammenhang mit der klassischen statistischen Regression diskutiert. In den Modellen für eine Regression mit stochastischen Parametern wird von der Annahme ausgegangen, dass aufgrund von beobachtungsspezifischen Einflüssen Abweichungen in den Modellparametern vorliegen, die durch eine Zufallsvariable beschrieben werden können. Man kann sich dies als Bündel von reellwertigen Geraden vorstellen, wobei für die Geraden eine Häufigkeitsverteilung existiert. Die Schätzverfahren ermöglichen es, die individuellen Abweichungen zu schätzen. Vgl. z.B. [Hildreth und Houck, 1968; Rosenberg, 1973; Newbold und Bos, 1985].

[55] Weiter außen liegende Werte gehören also möglicherweise eher zu einem weniger typischen Funktionszusammenhang. Dicht beieinander liegende Werte gehören möglicherweise zum selben, möglicherweise aber auch zu unterschiedlichen Funktionszusammenhängen.

Die Grundvorstellung des beschriebenen Approximationsansatzes lässt sich am besten anhand der Methoden zur Evaluation reellwertiger Funktionen anhand von Fuzzy-Daten aus Abschnitt *3.2.3* verdeutlichen. Durch die Anwendung eines Transfer-Operators wird die Fuzziness der Daten in die Menge der Funktionsparameter projiziert und wir erhalten eine Fuzzy-Menge der möglichen Parameter $\widetilde{\mathcal{A}}$. Durch den Zusammenhang $\mu_{\widetilde{F}}(f(\cdot, \mathbf{a})) = \mu_{\widetilde{\mathcal{A}}}(\mathbf{a})$ ist das dazugehörige Fuzzy-Bündel von Funktionen $\widetilde{F}$ definiert. Im allgemeinen ist $\widetilde{\mathcal{A}}$ aber nicht-normalisiert und interaktiv, so dass die Fuzzy-Parametermenge nicht in einen Vektor von Fuzzy-Parameter-Werten für die einzelnen Merkmalskomponenten zerlegt werden kann. Darüber hinaus ist $\widetilde{\mathcal{A}}$ typischerweise nicht von einfacher Gestalt und im allgemeinen nicht Element in der Menge der *LR*-Fuzzy-Zahlen. $\widetilde{F}$ lässt sich daher nicht wie gefordert als Approximationsfunktion darstellen.

Grundsätzlich kommen vor allem die Methoden zur Possibilistischen Regression dem zweiteren Approximationsprinzip nahe. Das wird bei der Possibilistischen Regression mit rein reellwertigen Daten (x_i, y_i) besonders deutlich. Dabei wird eine Fuzzy-lineare Funktion mit Fuzzy-Parametern so angepasst, dass die reellwertigen Datenpunkte von der Approximationsfunktion eingehüllt und mit einem Zugehörigkeitswert $\mu_{f(\cdot,\widetilde{A})}(x_i, y_i) \in [\kappa, 1]$ relativ zur Modalwertgeraden versehen werden. Mit der Vorgabe des Anspruchsniveaus κ wird eine Entscheidung darüber getroffen, zu welchem Grad die Einzelwerte mindestens zur charakterisierenden linearen Funktion zugehörig sind. Dabei geht die Possibilistische Regression von der extremen Annahme aus, dass jede Beobachtung potentiell die Abgrenzung des Bereiches möglicher Funktionswerte erweitert und dass eine Häufung von Punkten keinen Beitrag zur Bestätigung der Plausibilität eines Funktionsparameters leistet. Die Zugehörigkeitsbewertung erfolgt nach einem vorgegebenen Lagemuster. In der Regel wird angenommen, dass die Fuzzy-Parameter symmetrisch sind.

Im Gegensatz dazu haben Bárdossy u. a. [1992] gezeigt, dass die Kleinste Quadrate Fuzzy-Regression bei reellwertigen Daten (x_i, y_i) auch bei vorgegebenen Modellfunktionen mit Fuzzy-Parametern immer nur die klassischen reellwertigen Kleinste Quadrate-Parameter ergibt.[56] Bei einer Kleinste Quadrate Fuzzy-Approximation spielen folglich nur die Lage der Datenpunkte und die Form der Daten eine Rolle für die Approximationsfunktion. Eine Teilzugehörigkeit zum funktionalen Zusammenhang kann damit nicht bestimmt werden. Die Ergebnisse können auf alle Metriken gemäß Bertoluzza u. a. [1995] verallgemeinert werden.

Diese Beispiele können als Illustration dafür gesehen werden, dass das alternative Approximationsmodell nur dann lösbar ist, wenn eine zusätzliche Instanz angegeben werden kann, mit der die Inhomogenität der Messereignisse technisch bewertet wird. Z.B. zu welchem Grad die Volatilität der Daten als einfache Lagevariation von der Modellfunkti-

[56] Vgl. ebd. S. 185f. für den univariaten Fall sowie S. 189 für den multivariaten Fall.

on und zu welchem Grad sie als abweichende Zugehörigkeit zum Gesamtzusammenhang bewertet werden soll.[57]

Zusammenfassend ist zu konstatieren, dass bei Fuzzy-Daten in der epistemischen Interpretation die Approximation einer Fuzzy-Charakteristik nach dem ersten Modellansatz hauptsächlich bei Daten mit reellwertigen Inputs $(x_i, \widetilde{Y}_i)$ sinnvoll ist. Die Kleinste Quadrate Fuzzy-Regression ist ausschließlich für Modelle mit der ersten Interpretation geeignet. Die Possibilistische Regression bei Fuzzy-Daten gibt hingegen mit der Wahl von κ einen Kompromiss zwischen dem ersten und den zweiten Modellansatz vor, wie auf Seite 92 erörtert wurde. In diesem Sinne lässt sich auch ein Ergebnis von Bárdossy [1990] interpretieren. Dort wird gezeigt, dass bei Fuzzy-Daten vom Typ $(x_i, \widetilde{Y}_i)$ jeder Datenpunkt der possibilistischen Approximationsfunktion auf dem κ-Niveau auf einer Geraden mit Zugehörigkeitsgrad von mindestens κ liegt, die durch die möglichen Parameter definiert ist. Es gilt

$$\forall\ 1 \le i \le n\ \forall\ y_i^* \in [\widetilde{Y}_i]^\kappa\ \exists\ a(y_i^*) \in [\widetilde{A}]^\kappa :\ y_i^* = f(x_i, a(y_i^*)).$$

Im Unterschied dazu steht bei der überwiegenden Anzahl der Methoden zur Fuzzy-Designregression der erste Modellansatz im Vordergrund, da die Approximationsfunktion zunächst an Lagecharakteristika der Fuzzy-Daten angepasst werden, bevor die Unschärfecharakteristik bestimmt wird.

Der zweite Approximationsansatz ist grundsätzlich sowohl für physikalische als auch für epistemische Fuzzy-Daten denkbar. Wie in Abschnitt *3.2.1* bereits erörtert wurde, ist allerdings die Possibilistische Regression aufgrund ihrer technischen und interpretatorischen Schwächen nicht als Anpassungsmethode zu empfehlen.[58] Somit sind bislang noch keine tragfähigen Ansätze zur Approximation eines Fuzzy-Bündels von reellwertigen Funktionen bekannt.

[57] Bei den Regressionsansätzen mit stochastischen Parametern, die in der Statistik diskutiert werden, wird diese Bewertung z.B. durch die Annahme gewährleistet, dass die stochastischen Parameter autokorreliert sind, oder durch die Annahme, dass eine Korrelation zwischen Störvariabler und unabhängigen Variablen besteht, die auf die Zufallsverteilung der Koeffizienten zurückgeht. Dies sind Annahmen, die es ermöglichen, die Modellparameter sowie deren Varianzen aus den vorliegenden Daten zu schätzen. Als Fuzzy-Ansatz für variierende Koeffizienten kann man sich z.B. eine Methode mit gewichteten Abständen $d_{\mu_X}(\widetilde{Y}, f(\widetilde{X}, \widetilde{A})) = \mu_X d(\widetilde{Y}, f(\widetilde{X}, \widetilde{A}))$ oder $d_{\mu_X}(\widetilde{Y}, f(\widetilde{X}, \widetilde{A})) = d(\mu_X \widetilde{Y}, f(\widetilde{X}, \widetilde{A}))$ vorstellen, wobei die Zugehörigkeits-Gewichte μ_X exogen vorzugeben sind. Es wäre noch zu prüfen, inwieweit auch Methoden zur endogenen Bestimmung solcher Zurechnungsgewichte vorstellbar sind. Naheliegend wäre z.B. $\Gamma : F_{coc}^{nob}(\mathbb{R})^k \longrightarrow [0,1]$ Maß der Fuzziness und $\mu_X = \Gamma(\widetilde{X})$.

[58] Stattdessen wären Fuzzy-Ansätze mit geeigneten, formalisierten Verteilungsannahmen zu untersuchen.

3.3.3 Fuzzy-Schätzung

Von Parameterschätzung kann erst dann gesprochen werden, wenn der Modellanpassung ein ökonometrisches Schätzmodell zugrundeliegt. In der Einleitung zu Abschnitt *2.2* wurde bereits dargestellt, dass im Rahmen eines Schätzmodells einerseits der Strukturzusammenhang funktional zu beschreiben und andererseits die Abhängigkeiten der Modellvariablen und deren Konvergenzverhalten im Rahmen eines Verteilungsmodells zu erklären ist.

Bei Fuzzy-Daten ist dazu insbesondere zu beschreiben, in welcher Beziehung Wahrscheinlichkeitsverteilung und Fuzziness zueinander stehen. Ähnlich wie im reellwertigen Fall wird das gewählte Verteilungsmodell durch eine Fuzzy-Zufallsvariable verkörpert, also durch eine messbare Abbildung $\widetilde{Z} : \Omega \longrightarrow F_{coc}^{nob}(\mathbb{R})$. Dabei ist $F_{coc}^{nob}(\mathbb{R})$ mit einer passenden σ-Algebra zu versehen. Krätschmer [2001b] stellt einen verallgemeinerten Ansatz vor, mit dem Fuzzy-Zufallsvariable als gewöhnliche Zufallsvariable im Rahmen der klassischen Maß- und Wahrscheinlichkeitstheorie definiert werden können. Die Definition legt die Messbarkeit bezüglich der Borel-σ-Algebra der δ_2-Topologie zugrunde.[59] Er zeigt, dass die verallgemeinerte Definition eine Vereinheitlichung der Messbarkeitskonzepte für Fuzzy-Zufallsvariablen von Puri/Ralescu bzw. von Kwakernaak ist. Für den Erwartungswert einer Zufallsvariable kommen mehrere Definitionen in Frage. Für metrische Räume (Υ, d) kann der Erwartungswert $\mathrm{E}^{(d)}\, Z$ einer Zufallsvariable Z nach einem Prinzip von Fréchet definiert werden als Lösung des Minimierungsproblems

$$\mathrm{E}\, d^2(Z, \mathrm{E}^{(d)}\, Z) \stackrel{!}{=} \inf_{w \in \Upsilon} \mathrm{E}\, d^2(Z, w). \tag{3.15}$$

Es kann gezeigt werden, dass der Fréchet-Erwartungswert bezüglich der δ_2-Topologie identisch mit dem Aumann-Erwartungswert[60] ist. Da der Aumann-Erwartungswert linear bzgl. der Addition ist, ist somit gewährleistet, dass der definierte Erwartungswert $\mathrm{E}^{(\delta_2)}\, \widetilde{Z}$ linear bzgl. der Addition $\oplus$ ist, d.h. $\mathrm{E}^{(\delta_2)}(\lambda_1 \odot \widetilde{Z}_1 \oplus \lambda_2 \odot \widetilde{Z}_2) = (\lambda_1 \odot \mathrm{E}^{(\delta_2)}\, \widetilde{Z}_1) \oplus (\lambda_2 \odot \mathrm{E}^{(\delta_2)}\, \widetilde{Z}_2)$.

Entsprechend wird die Varianz für Fuzzy-Zufallsvariablen üblicherweise als zugehörige Fréchet-Varianz definiert, d.h. $\mathrm{Var}^{(\delta_2)}\, \widetilde{Z} = \mathrm{E}\, \delta_2^2(\widetilde{Z}, \mathrm{E}^{(\delta_2)}\, \widetilde{Z})$. Die so definierte Fréchet-Varianz ist reellwertig. Sie liegt auch dem Nachweis der statistischen Eigenschaften der Parameterschätzer in den nachfolgend beschriebenen Modellen zugrunde.

---

[59] Diese Definition ist topologisch äquivalent zu den weiteren Metriken $\delta_{(\nu)}$ auf der Basis der Stützfunktionen $s_{\widetilde{A}}$ gemäß [Bertoluzza u. a., 1995] bzw. [Diamond und Körner, 1997].

[60] Der Aumann-Erwartungswert $\mathrm{E}^{(A)}\, \widetilde{Z}$ einer Fuzzy-Zufallsvariable $\widetilde{Z}$ ist dadurch definiert, dass für alle $\alpha \in [0,1]$ gilt $\left[\mathrm{E}^{(A)}\, \widetilde{Z}\right]^\alpha = \mathrm{E}^{(A)}([\widetilde{Z}]^\alpha)$. Dabei ist der Aumann-Erwartungswert $\mathrm{E}^{(A)}[\widetilde{Z}]^\alpha$ der klassischen zufälligen Menge $[\widetilde{Z}]^\alpha$ definiert als $\mathrm{E}^{(A)}[\widetilde{Z}]^\alpha = \left\{\mathrm{E}\, \eta \mid \eta \in L^1(\mathbb{R}^m, \mathbb{B}^m) \text{ und } \eta(\omega) \in [\widetilde{Z}]^\alpha(\omega) \text{ f.s.}\right\}$. Die Identität zwischen Fréchet- und Aumann-Erwartungswert kann auf die $\delta_{(\nu)}$-Metriken erweitert werden, sofern die Dichte von ν symmetrisch und positiv definit ist, vgl. Näther 2000, S. 207ff.

Als lineare Modellgleichungen stehen auch im Strukturmodell grundsätzlich nur die Funktionsvarianten zur Verfügung, die in den vorangehenden Abschnitten diskutiert wurden. Zur Vereinfachung beschränken wir uns auf eine Situation mit deterministischen Regressoren $x_i \in \mathbb{R}^m$ bzw. $\widetilde{X}_i \in F_{coc}^{nob}(\mathbb{R})^m$. D.h. es wird vorerst ausgeschlossen, dass die Regressoren selber ebenfalls Zufallsvariable sind. Außerdem seien $\widetilde{Y}_i(\omega)$ Realisationen von Fuzzy-Zufallsvariablen die einem Verteilungsmodell für epistemische Fuzzy-Daten unterliegen, d.h. insbesondere gebe es eine reellwertige Zufallsvariable $U_i^* : \Omega \longrightarrow \mathbb{R}$, so dass $U_i^*(\omega) \in \widetilde{Y}_i(\omega)$ die „wahren“ Werte sind.

Als eine Grundfrage für die Konstruktion einer Strukturfunktion ist zu entscheiden, ob die ungenauen Messverhältnisse in das Strukturmodell integriert werden sollen oder nicht. Je nachdem wird entweder ein defuzzifiziertes Modell angepasst, das aufgrund der beobachteten Fuzzy-Umgebungen als plausibler Zusammenhang für die ursprüngliche Zufallsvariable U^* angesehen werden kann, oder es wird ein fuzzifizierendes Modell angepasst, in dem der lineare Zusammenhang zwischen den Messungenauigkeiten mit abgebildet ist. Eine explizite Berücksichtigung der Messverhältnisse ist dann sinnvoll, wenn Hypothesen über das Verhalten der Verschmutzungen vorliegen und die Verschmutzung durch das Messverfahren einen wesentlichen Einfluss auf die Qualität der Beobachtungen hat.[61] Man ist dann allerdings mit einigen Einschränkungen konfrontiert, die ihre Ursache in der Fuzzy-Arithmetik haben. Insbesondere muss man sich fragen, ob eine Zunahme der Messungenauigkeit mit wachsendem Abstand vom Koordinatenursprung plausibel ist, ob also auch Fuzzy-Steigungsparameter $\widetilde{A}_j$ für $j \neq 0$ angesetzt werden sollen. Die Hauptfrage ist hierbei, ob beide Aspekte in derselben Zusammenhangsgleichung geschätzt werden sollen, oder ob es aufgrund der starken arithmetischen Abhängigkeiten und Begrenzungen doch sinnvoller ist, beide Aspekte getrennt zu modellieren und zu schätzen.

Mit der folgenden Gleichung wird ein linearer Strukturzusammenhang mit reellwertigen Parametern $a_0, \ldots, a_k \in \mathbb{R}$ modelliert

$$\mathrm{E}\,\widetilde{Y} = a_0 \oplus \bigoplus_{j=1}^{k} a_j \odot \widetilde{X}_j. \tag{3.16}$$

Als Schätzmethode steht die Kleinste Quadrate Fuzzy-Regression in der δ_2-Metrik zur Verfügung.[62] Es kann gezeigt werden, dass die Parameterschätzer unter bestimmten Regularitätsanforderungen stark konsistent sind.[63] Außerdem kann die asymptotische Kon-

[61] Also z.B. dann, wenn die Unschärfe in Y in einem bestimmten Verhältnis zur Unschärfe in X steht, d.h. wenn die Modellierung faktisch zu einer Homogenisierung des Verhältnisses der Messungenauigkeiten zwischen den unabhängigen und der abhängiger Variablen beiträgt.

[62] Vgl. [Krätschmer, 2006a,c].

[63] Vgl. [Krätschmer, 2006c].

vergenz der Verteilung der Parameterschätzer gegen eine $\mathbb{R}^k$-wertige Normalverteilung gezeigt werden.[64]

Die Güteeigenschaften werden allerdings in einem Verteilungsmodell hergeleitet, das von „unscharfen Messkonzepten“ ausgeht und damit von einer physikalischen bzw. linguistischen Interpretation der Fuzzy-Zufallsvariablen, die den Punktcharakter der Fuzzy-Daten ausnutzt. Es bleibt daher ungeklärt, inwieweit die statistischen Eigenschaften der Schätzer auch auf das Strukturmodell für epistemische Fuzzy-Daten übertragbar sind. Bei epistemischen Fuzzy-Daten liegt es nahe, die Schätzfunktion im Sinne einer Plausibilitätsbewertung für den „wahren“ Regressionszusammenhang zu interpretieren. Dazu müsste die Qualität der Plausibilitätsbewertung, die durch die Regressionsfunktion repräsentiert wird, allerdings genauer beschrieben werden.

In einer Situation, in der nur die Outputs als Fuzzy-Daten beobachtet werden, so dass die zufällige Variabilität zum Teil unter der Messungenauigkeit verborgen bleibt, kann ein Strukturmodell mit Fuzzy-Parametern $\widetilde{A}_0, \ldots, \widetilde{A}_k \in F_{coc}^{nob}(\mathbb{R})$ hinterlegt werden:

$$\mathrm{E}\,\widetilde{Y} = \widetilde{A}_0 \oplus \bigoplus_{j=1}^{k} \widetilde{A}_j \odot x_j, \tag{3.17}$$

dabei seien die $x_j \in \mathbb{R}$. Der Zuwachs der Fuzziness in den Bildwerten gegenüber den Eingangswerten ist dann ebenfalls Gegenstand des Strukturmodells. Für diese Modellgleichung gibt es unter bestimmten Bedingungen sogar beste, lineare, erwartungstreue Schätzer (BLUE).[65] Dabei ist ein linearer Parameterschätzer $\widetilde{\beta}_j$ bei Fuzzy-Daten definiert durch

$$\widetilde{\beta}_j = \bigoplus_{i=1}^{n} \lambda_{ij} \odot \widetilde{Y}_i$$

mit $\lambda_{ij} \in \mathbb{R}$, und es wird die Erwartungstreue bzgl. des Aumann-Erwartungswertes gefordert. Die Existenz von BLUE-Schätzern ist eine nicht-asymptotische Eigenschaft und ist gerade bei Fuzzy-Daten wünschenswert, da in Situationen, in denen Fuzzy-Daten vorliegen, häufig keine großen Stichproben durchgeführt werden können.

Anders als beim klassischen Regressionsproblem mit reellwertigen i.i.d. Zufallsvariablen sind die Parameterschätzer der Kleinste Quadrate Fuzzy-Regression aber im allgemeinen nicht identisch mit den BLUE-Schätzern. In [Näther, 2000] wird gezeigt, dass es in einem Modell mit $k \geq 2$ im allgemeinen keinen BLUE-Schätzer gibt.[66] Entscheidend für die Nicht-

[64] Vgl. [Krätschmer, 2006b].

[65] Vgl. [Näther, 1994; Körner und Näther, 1998; Näther, 2001, 2006].

[66] Dabei wird ein Modell ohne Inhomogenität $\widetilde{A}_0$ angenommen, d.h. in unserer Notation existiert dennoch immer ein BLUE, falls $\widetilde{X}_0 = 1_{\{0\}}$ und falls gilt $\mathrm{Var}\,\widetilde{Y}_i = \sigma^2$.

Übertragbarkeit der Ergebnisse für reellwertige Zahlen auf Fuzzy-Zahlen sind wiederum die Einschränkungen für die Linearität von Fuzzy-Zahlen, die u.a. nur monoton nichtfallendes Verhalten in den Spannweiten zulässt. Der hauptsächliche Grund dafür, dass nur in seltenen Fällen BLUE-Schätzer existieren, ist aber darin zu sehen, dass die Fuzzy-Arithmetik bei linearen Fuzzy-Schätzern erzwingt, dass für Modalwerte und Spannweiten der Regressionsfunktion dieselbe Proportionalität zugrunde liegen.

Es werden daher auch schwächere Ansätze von BLUE-Schätzern untersucht.[67] So ist ein in der Fréchet-Varianz bester Schätzer *komponentenweise BLUE*, wenn es unterschiedliche lineare, erwartungstreue Schätzer in separierten Regressionen für die Modalwerte bzw. für die Spannweiten gibt. Bei *LR*-Fuzzy-Daten ergibt sich der komponentenweise BLUE aus der Lösung des Kleinste Quadrate-Problems für die Modalwerte und für die Spannweiten. Die Lösung ist allerdings nur dann wohldefiniert, wenn die Lösung des Schätzproblems für die Spannweiten keine negativen Parameter enthält.

Insgesamt ist es aufgrund der Einschränkungen für die Existenz eines BLUE günstiger, zu Kleinste Quadrate-Schätzern für (3.17) überzugehen, für die immer eine Lösung existiert. Zur statistischen Güte der Kleinste Quadrate-Schätzer für dieses Schätzmodell liegen allerdings keine gesonderten Ergebnisse vor.

Der Vollständigkeit halber sei hier darauf hingewiesen, dass es bislang keine Ansätze gibt, mit denen ein Verteilungsmodell für die Vagheit des Strukturzusammenhangs modelliert und geschätzt werden kann. Für die alternative Interpretation einer Modellfunktion mit Fuzzy-Parametern wie in Abschnitt *3.3.2*, Seite 112f, bei der die Beobachtungen nur lose einem gemeinsamen Zusammenhang genügen, gibt es also insgesamt noch kein zufriedenstellendes Approximationsverfahren.

Die Aufstellung einer Strukturgleichung mit Fuzzy-Parametern ist nur in den speziellen Fällen sinnvoll, in denen das Verhalten der Messungenauigkeiten mit einer Fuzzy-linearen Funktion annähernd beschrieben werden kann. Dies trifft hauptsächlich bei Fuzzy-Daten $(\mathbf{x}_i, \widetilde{Y}_i) \in \mathbb{R}^k \times F_{coc}^{nob}(\mathbb{R})$ zu. In den übrigen Fällen ist eine Modellfunktion mit reellwertigen Steigungsparametern, d.h. eine defuzzifizierte oder eine vereinfachte Modellfunktion, besser geeignet. Grundsätzlich stellt die Fuzzy-lineare Schätzfunktion aber keinen klassischen Funktionsgraphen im $\mathbb{R}^{k+1}$ dar, sondern eine Abbildungsvorschrift für den funktionalen Zusammenhang zwischen Fuzzy-Variablen, die die Fuzzy-Abschätzungsumgebungen für die mögliche Lage der ungenauen Beobachtungswerte repräsentieren. Je nachdem, welches Verteilungsmodell hinterlegt werden kann, ist es daher beispielsweise möglich, dass die geschätzte Proportionalität des Zusammenhangs nur auf einem gewissen (Un-) Genauigkeitsniveau zutrifft.

[67] Vgl. [Körner und Näther, 1998; Näther, 2000, 2006].

Falls die Eingangswerte $\widetilde{X}_i$ im wesentlichen genau sind, stellt ein reellwertiger Strukturzusammenhang eine Proportionalität zu den Fuzzy-Umgebungen $\widetilde{Y}_i(\omega_i)$ her, die den „wahren“ Wert $U_i^*(\omega_i)$ enthalten. Es wäre daher wünschenswert, wenn die reellwertigen Bildwerte als „bestmögliche“ Plausibilitätsbewertung für den „wahren“ Wert $U_i^*(\omega_i)$ betrachtet werden könnten. Dies wäre anhand verschiedener Verteilungsmodelle für epistemische Fuzzy-Daten noch genauer zu untersuchen. Als Vorbereitung auf die Auswahl eines geeigneten Verteilungsmodells für die Regression bei epistemischen Fuzzy-Daten wird in Kapitel *5* zunächst ein Benchmarking der klassischen Kleinste Quadrate-Approximation mit der Kleinste Quadrate Fuzzy-Approximation durchgeführt, dabei werden exemplarische Szenarien für ungenaue Messergebnisse verwendet. Mit den Ergebnisses des Benchmarking kann dann in Abschnitt *6.1* eine erster Ausblick auf die praktischen Anforderungen an das Verteilungsmodell gegeben werden.

Bei Daten mit deterministischen Fuzzy-Inputs $\widetilde{X}_i$ kann die Strukturgleichung darüber hinaus höchstens als Aussage zur Plausibilität des Zusammenhangs der „wahren“ $U_i^*(\omega)$ zum ungenauen Messwert $\widetilde{X}_i$ ausgewertet werden. Die Plausibilität des „wahren“ Zusammenhangs könnte nur dann geschätzt und beschrieben werden, wenn die $\widetilde{X}_i$ als Zufallsvariable modelliert werden, die ebenfalls einem Verteilungsmodell für epistemische Fuzzy-Daten genügen, wenn also zu einem Modell mit stochastischen Eingangswerten übergegangen wird.[68]

[68] Ansätze zur Fuzzy-Regression mit stochastischen Fuzzy-Inputs werden z.B. in [Wünsche und Näther, 2002; Näther, 2006] betrachtet.

4 Datenunschärfen in der Beschäftigtenstatistik

Die Beschäftigtenstatistik hat günstige Eigenschaften, die ausschlaggebend für ihre Auswahl als Ausgangspunkt der Analyse in der vorliegenden Arbeit waren: Da die Beschäftigtenstatistik eine Totalstatistik ist, entfallen Probleme, die durch die Verteilung der Stichprobe in der Grundgesamtheit verursacht werden. Insbesondere kann die Berücksichtigung von zusätzlichen stochastischen Unschärfen in den Daten vernachlässigt werden. Infolge der Herkunft der Daten aus dem Meldeverfahren für die Sozialversicherungen bietet sie andererseits zusätzlich die Grundlage für eine Untersuchung der Fuzzifizierbarkeit solcher Datenunschärfen, die durch den Einfluss der Gesetzgebung und andere administrative Rahmenbedingungen verursacht werden.

Für zusätzliche Informationen über wesentliche Datenfehler in der Beschäftigtenstatistik können die Stichprobenstatistiken der Erwerbsbevölkerung herangezogen werden, die häufig als Kontrollstatistiken verwendet werden. Solche komplementären Datenquellen sind vor allem Mikrozensus und SOEP sowie andererseits auf der Unternehmensseite das Betriebspanel. Zudem stehen im Verbund der Arbeitsmarktstatistik mit diesen und weiteren Statistiken vielfältige externe Informationen zur Verfügung, die bei der Datenmodellierung ergänzt werden können.

Bei der Modellierung von Fuzzy-Daten ist die Ungenauigkeit für jeden Beobachtungswert auf dem aktuellen Stand des Wissens zu bewerten. Die Modellierung von Fuzzy-Daten ist daher stark auf den Einzelfall bezogen. Sie entzieht sich dadurch weitgehend einer Vereinheitlichung, da jede Situation gesondert abzubilden und zu betrachten ist. Um dieser Offenheit des Modellierungsansatzes gerecht zu werden, sollen in diesem Kapitel am Beispiel der Beschäftigtenstatistik zunächst reale Fehlereinflüsse daraufhin untersucht werden, wie diese bei der Konstruktion von Fuzzy-Merkmalswerten berücksichtigt und operationalisiert werden können. Alle Fehlereinflüsse in der Beschäftigtenstatistik können als systematisch angesehen werden, weil sie als Totalerhebung keinem Stichprobeneinfluss unterliegt.

Die exemplarischen Modellierungen wurden so ausgewählt, dass sie bestimmten Typen von Fehlereinflüssen zugeordnet werden können. Dieses Vorgehen ermöglicht es, das kon-

struktive Verfahren zur Abbildung von Fuzzy-Merkmalsvariablen am konkreten Beispiel vorzuführen und mit einer allgemeinen Erörterung der Strategien bei der Modellierung zu verbinden. Anknüpfend daran werden in einem zweiten Schritt die beispielhaften Modellierungen danach ausgewertet, welche spezifischen Eigenschaften der Fuzzy-Daten bei der Fuzzy-Regression zu berücksichtigen sind und welche Anforderungen sich daraus für die Regressionsanalyse ergeben.

In Abschnitt *4.1* werden zunächst die Besonderheiten der Beschäftigtenstatistik vorgestellt und die relevanten Fehlerquellen beschrieben. Die Auswertung der Fehlereinflüsse für die Modellierung konkreter Fuzzy-Merkmalsvariablen erfolgt in Abschnitt *4.2*. Dort werden einige Beispiele und Prinzipien zur Konstruktion von empirischen Fuzzy-Daten vorgestellt. Abschließend werden in Abschnitt *4.3* die wesentlichen Eigenschaften der empirischen Fuzzy-Merkmalswerte charakterisiert, die bei der Anpassung von Fuzzy-Regressionsmodellen zu berücksichtigen sind.

4.1 Eigenschaften der Beschäftigtenstatistik

Die Beschäftigtenstatistik ist eine Sekundärstatistik, die auf einem integrierten Meldeverfahren basiert. Dabei ist jede sozialversicherungspflichtig beschäftigte Person durch die Arbeitgeber/innen zunächst an die zuständige Krankenkasse zu melden. Diese prüfen die Daten auf inhaltliche Richtigkeit und Vollständigkeit und nehmen — falls erforderlich — Korrekturen vor.[1] Die von den Krankenkassen geprüften Daten werden an die Datenstellen der Rentenversicherungsträger weitergeleitet. Diese leiten die für die Arbeitsverwaltung relevanten Daten nach einer weiteren Prüfung an die Bundesagentur für Arbeit (BA)[2] weiter, wo sie statistisch aufbereitet werden. Entscheidend für die Aussagefähigkeit der Beschäftigtenstatistik ist daher, dass die Meldungen möglichst vollständig und zügig an die BA übermittelt werden. Sie ist insofern hoch abhängig von Einflüssen auf die Meldedisziplin der Betriebe und der beteiligten Sozialversicherungsträger, z.B. durch Änderungen im Meldeverfahren sowie davon, wie die Sozialversicherungspflicht jeweils gesetzlich geregelt ist. Da die Beschäftigtenstatistik eine Totalauszählung darstellt, können die Ergebnisse regional, wirtschaftsfachlich und beruflich tief gegliedert werden. Zudem hat sie aufgrund der bestehenden Meldepflicht eine sehr hohe Abdeckung. Allerdings erfasst die Statistik nur die Personen, die sozialversicherungspflichtig beschäftigt sind. Sozialversicherungspflichtig sind alle Beschäftigten, die der Versicherungspflicht in der Krankenversicherung, der Rentenversicherung oder der Arbeitslosenversicherung unterliegen. „Beschäftigung“

[1] Diese Überprüfungspflicht für die Krankenkassen besteht seit der 2. Datenerhebungs- und Datenübermittlungs-Verordnung (DEVO/DÜVO) vom 29. Mai 1981.

[2] Bis Anfang 2004 Bundesamt für Arbeit.

ist dabei nach dem 4. Sozialgesetzbuch, §7 Abs. 1 definiert als „nichtselbständige Arbeit, insbesondere in einem Arbeitsverhältnis“. Betriebliche Berufsausbildung wird dazu gerechnet. Somit sind u.a. Beamte, Selbständige, mithelfende Familienangehörige ausgeschlossen. Geringfügig beschäftigte Personen werden erst seit 1999 von der Meldepflicht eingeschlossen und erfasst.[3] Nach den Ergebnissen der Repräsentativstatistik über die Bevölkerung und den Arbeitsmarkt (Mikrozensus) stellen sozialversicherungspflichtig Beschäftigte einen Anteil von über 75% an allen Erwerbstätigen, wobei die Gesamtheit der Beschäftigten eines Wirtschaftszweiges, bedingt durch die Beschäftigtenstruktur in den einzelnen Wirtschaftszweigen, durch die Ergebnisse der Beschäftigtenstatistik unterschiedlich stark repräsentiert wird.[4] Damit sind Ergebnisse aus der Beschäftigtenstatistik ein wesentlicher Faktor für die Darstellung des erwerbsstatistischen Gesamtbildes.[5]

Der Bestand an sozialversicherungspflichtig Beschäftigten wird vierteljährlich jeweils zum Quartalsende als Auswertungsstichtag — d.h. am 31. März, am 30. Juni[6], am 30. September und am 31. Dezember eines Jahres — nachgewiesen. Für das frühere Bundesgebiet liegen Bestandsdaten seit Juni 1974 vor, für die neuen Bundesländer seit März 1992.[7]

4.1.1 Meldeverfahren und Datenerfassung

Das Meldeverfahren wird zunächst nach der geltenden DEÜV vom 10. Februar 1998 dargestellt. Auf die Veränderungen im Zeitablauf wird dann gesondert hingewiesen. Demnach sind die Arbeitgeber/innen auskunftspflichtig für bestehende sozialversicherungspflichtige Beschäftigungsverhältnisse. Sie müssen über die sozialversicherungspflichtigen Arbeitnehmer/innen in ihren Betrieben an die Träger/innen der Sozialversicherung Meldungen erstatten. Meldetatbestände sind insbesondere *Anmeldung* sowie *Abmeldung* bei Aufnahme bzw. Beendigung einer sozialversicherungspflichtigen Beschäftigung, *Unterbrechung* der sozialversicherungspflichtigen Beschäftigung bei Wegfall des Anspruchs auf Arbeitsentgelt

[3] Seit der gesetzlichen Neuregelung zum Stichtag 1.4.1999 sind Arbeitgeber/innen verpflichtet, auch für Personen, die ausschließlich sogenannte geringfügig entlohnte Tätigkeiten ausüben, pauschalierte Beiträge zu Kranken- und Rentenversicherung zu entrichten. Die Ergebnisse sind Gegenstand eigener Statistiken, die die BA veröffentlicht.

[4] So ist im Verarbeitenden Gewerbe der Abdeckungsgrad hoch, während er z.B. in der Öffentlichen Verwaltung, im Handel oder in der Land- und Forstwirtschaft erheblich geringer ist.

[5] Vgl. Statistisches Bundesamt 2009, S. 4.

[6] Für den Stichtag 30. Juni werden die Daten mit einer stärkeren Aufgliederung ausgewiesen als zu den anderen Stichtagen. Dazu gehört beispielsweise die Auswertung nach Wirtschaftsabschnitten, nach Berufen sowie nach Altersgruppen.

[7] Für das Land Berlin können statistische Ergebnisse, infolge der Zusammenlegung von Arbeitsagenturen, nicht mehr getrennt nach Ost- und West-Berlin nachgewiesen werden. Vgl. [Statistisches Bundesamt, vierteljährlich] für den 31. März 2008.

für mindestens einen Monat[8] und *Jahresmeldung* für jedes am 31.12. eines Jahres bestehende Beschäftigungsverhältnis. Die Meldefristen wurden mittlerweile vereinheitlicht; alle Meldungen haben grundsätzlich mit der nächsten Lohn- und Gehaltsabrechnung zu erfolgen.[9] Außer bei der Anmeldung wird mit den aufgeführten Meldungen jeweils das Ende eines Lohnzahlungszeitraumes mit dem darin erzielten Arbeitsentgelt angezeigt.

Voraussetzung für die Meldung einer/eines sozialversicherungspflichtigen Beschäftigten ist, dass für den Betrieb eine Betriebsnummer vergeben wurde und dass eine Versichertennummer für den/die Beschäftigte existiert. Um sicherzustellen, dass eine *Versichertennummer*, die für eine Arbeitnehmer/in verwendet wird, eindeutig ist, werden Meldedaten ohne Versichertennummer nicht weitergeleitet und zunächst ein Prüfverfahren durchgeführt, bevor schließlich ggf. eine neue Versichertennummer vergeben wird. Die daraus entstehenden Verzögerungen führen vor allem bei Erstbeschäftigungsverhältnissen zu einem verspäteten Eingang der Meldungen bei der BA. Die *Betriebsnummern* werden von den Betriebsnummernstellen der Arbeitsagenturen auf Antrag des Betriebs vergeben. Zu den Betriebsnummern werden in einer zentralen Betriebsdatei der BA die wesentlichen Merkmale des Betriebs geführt, so dass darüber Wirtschaftsklasse, Regionalkennziffer des Betriebes und Dienststellennummer der zuständigen Arbeitsagentur identifiziert und in der Beschäftigtenstatistik zugespielt werden können.

Die eingehenden Meldungen für die Beschäftigten werden von der BA in eine Versichertendatei übernommen. Dabei werden alle Meldungen zu einer Versicherungsnummer chronologisch in einem Versichertenkonto abgelegt, das somit den Beschäftigungsverlauf abbildet. Zusätzlich zu den Angaben über Beschäftigungsbeginn und -ende sowie das zugehörige Bruttoentgelt stehen die folgenden auswertbaren Merkmale in der Beschäftigtenstatistik zur Verfügung: Alter; Geschlecht; Staatsangehörigkeit; allgemeiner und beruflicher Ausbildungsabschluss; ausgeübte Tätigkeit (Beruf); Auszubildende; Stellung im Betrieb als Facharbeiter/in, Meister/in oder Polier/in oder andere Vollzeitbeschäftigte; Voll-, Teilzeitbeschäftigung; Wirtschaftszweig des Betriebes; bis zum 31. Dezember 2004 auch Rentenversicherungsträger mit der Gliederung in Arbeiter/innen und Angestellte[10] sowie Arbeits- und Wohnort, aus denen die Ein- und Auspendler für Regionen ermittelt werden können.[11]

[8] Unterbrechungen im Sinne der DEÜV sind z.B. Zeiten der Arbeitsunfähigkeit mit Bezug von Krankengeld nach Ende der Lohnfortzahlung, Mutterschutzzeiten, Zeiten des Elternurlaubs mit Bezug von Erziehungsgeld sowie Wehrdienst- und Zivildienstzeiten.

[9] Vgl. Statistisches Bundesamt vierteljährlich vom 31. März 2008, S. 7.

[10] Das Merkmal wird aufgrund der Organisationsreform in der gesetzlichen Rentenversicherung seit dem 1. Januar 2005 nicht mehr erhoben.

[11] Vgl. Statistisches Bundesamt vierteljährlich vom 31. März 2008, S. 5b.

Die Auswertung der Quartalsbestände der Beschäftigtenstatistik erfolgt nach einer Wartezeit von sechs Monaten nach dem Stichtag. Mit der Wartezeit sollen die Verzögerungen im Meldeverfahren ausgeglichen werden. Der Abstand von sechs Monaten zwischen Stichtag und Auswertung ist dabei ein Kompromiss zwischen größtmöglicher Aktualität der Ergebnisse und einer möglichst vollständigen Erfassung aller für den Berichtsstichtag vorliegenden Meldungen. Erfahrungsgemäß liegen der BA nach sechs Monaten ca. 95% der Meldungen vor. Zur Ermittlung des Beschäftigtenbestandes wird jedes Versichertenkonto maschinell daraufhin abgefragt, ob der/die Versicherte am Stichtag in einem Beschäftigtenverhältnis stand oder nicht. Das versichertenbezogenen Vorgehen hat den Vorteil, dass nicht eine konsistente Abfolge von Jahresmeldungen, An- und Abmeldungen vorliegen muss. Fehlende Meldungen und logische Fehler können häufig unter Zuhilfenahme bereits vorliegender Meldungen interpoliert werden. Wenn z.B. die Aufnahme einer neuen Beschäftigung gemeldet wurde, kann die Beendigung des vorhergehenden Beschäftigungsverhältnisses ergänzt werden, obwohl die Abmeldung der vorhergehenden Arbeitgeber/in noch nicht vorliegt.[12]

Die vierteljährliche Berichterstattung zur Struktur des Beschäftigtenbestandes ist nicht die einzige Statistik, die auf der Basis der Meldungen zur Sozialversicherung erstellt wird. Einmal jährlich werden zusätzlich die zeitraumbezogene Angaben zur Beschäftigungsdauer und zur beitragspflichtigen Höhe des Bruttoentgelts ausgewertet, das sog. *Jahreszeitraummaterial*. Beim Jahreszeitraummaterial werden nur entgeltrelevante Meldungen einbezogen, so dass nur sicher bestehende Beschäftigungsverhältnisse einbezogen werden.[13] Das Jahreszeitraummaterial wird daher insbesondere auch für Bewegungs- und Verlaufsuntersuchungen von Beschäftigungsverhältnissen herangezogen. Aufgrund der Meldeverzögerungen liegt ein sicherer Bestand an Beschäftigungsmeldungen aber erst nach einer Wartezeit von etwa drei Jahren vor, so dass Statistiken und Auswertungen, die auf das Jahreszeitraummaterial aufbauen, erst nach einer Wartezeit von drei Jahren zur Verfügung stehen.[14] In der *Beschäftigten-Historik* werden die seit dem Bestehen des Meldeverfahrens zur Sozialversicherung abgegebenen Meldungen von der BA kontenbezogen abgelegt. Sie umfasst in bereinigter und vervollständigten Form alle Beschäftigungsfälle, für die seit 1975 mindestens einmal eine sozialversicherungspflichtige Beschäftigung vorlag.[15] Die Historik wurde 1990 aus den archivierten Datenbeständen aufgebaut und wird seitdem

[12] Vgl. Statistisches Bundesamt vierteljährlich vom 31. März 2008, S. 10.

[13] Der Sachverhalt wird in Abschnitt *4.2.1* genauer erläutert.

[14] Vgl. Lüken 2002, S. 22. Die Vollständigkeit der Meldungen hat allerdings aufgrund der Automatisierung des Meldeverfahrens erheblich zugenommen. So wird im aktuellen Qualitätsbericht zur Beschäftigtenstatistik, [Statistisches Bundesamt, 2009], darauf hingewiesen, dass seit 2001 eine nachträgliche Korrektur der als vorläufig anzusehenden Quartalsergebnisse nicht notwendig war. Somit können die Ergebnisse letztlich doch als „endgültig“ gelten (S. 7).

[15] Vgl. Lüken 2002, S. 39.

fortlaufend ergänzt. Sie bietet u.a. die Möglichkeit, die Struktur verspäteter Meldungen in den Quartalsauswertungen zu untersuchen.[16] Die Historik ist außerdem die Datengrundlage für die *Beschäftigtenstichproben* des IAB. Dabei werden 1%- bzw. 2%-Stichproben der Beschäftigten in der Historik gezogen, wobei den Versichertenkonten noch Daten aus der Leistungsempfängerdatei der BA zugespielt werden. Die Beschäftigtenstichproben bieten jeweils kontinuierliche Verlaufsdaten zur sozialversicherungspflichtigen Erwerbsgeschichte und zum Bezug von Arbeitslosengeld, -hilfe oder Unterhaltsgeld und eignen sich dementsprechend zur Analyse von Erwerbsverläufen und Leistungsbezugsbiographien.[17]

Die Betriebe, die Beschäftigte melden, können über die Betriebsnummer identifiziert werden. Es ist daher möglich, die sozialversicherungspflichtig Beschäftigten — sowie seit 1999 auch die geringfügig Beschäftigten — auf Betriebsebene zu aggregieren. Dies wird genutzt, um sekundäre Betriebsgrößenklassen zu ermitteln. Seit 1977 werden jeweils zum 30. Juni aus der Beschäftigtenstatistik die Angaben über Betriebe ausgewertet und als zusätzliche Merkmale in den Datensätzen ergänzt. Die Betriebe, für die mindestens eine Beschäftigte/r in der Beschäftigtenstatistik gemeldet ist, werden im *Betriebs-Historik-Panel* aufbereitet, das sich aus aus Querschnittsdatensätzen ab dem Jahr 1975 für Westdeutschland und ab 1992 für Ostdeutschland zusammensetzt. Jeder Querschnitt umfasst alle Betriebe des gesamtdeutschen Raumes, die zur Jahresmitte (Stichtag: 30.06.) in der Beschäftigten-Historik erfasst sind. Außerdem bilden sie die Grundlage für die Ziehung der Betriebe für das *Betriebspanel* der IAB. Das IAB-Betriebspanel stellt die zentrale Quelle für Analysen zur Arbeitskräftenachfrage auf dem Arbeitsmarkt in Deutschland dar. Diese repräsentative Arbeitgeberbefragung zu betrieblichen Bestimmungsgrößen der Beschäftigung wird seit 1993 in Westdeutschland und seit 1996 auch in Ostdeutschland jährlich vom Forschungsbereich „Betriebe und Beschäftigung“ des IAB durchgeführt. In den *Linked-Employer-Employee-Datensätzen (LIAB)* wird eine Verbindung zwischen den Daten des IAB-Betriebspanels und den Personendaten der Beschäftigtenstatistik hergestellt. Dazu werden die Daten des Betriebspanels mit den längsschnittaufbereiteten Meldungen der sozialversicherungspflichtig Beschäftigten verknüpft.

Das Meldeverfahren zur Sozialversicherung, das der Beschäftigtenstatistik zugrundeliegt wird seit dem 1.1.1973 durchgeführt.[18] Größere Änderungen im Meldeverfahren hat es zum

[16] Zu den ersten Ergebnissen über strukturelle Einflüsse auf verspätete Meldungen vgl. [Cramer und Majer, 1991].

[17] Komplementär zu den Beschäftigtenstichproben wird auch ein Beschäftigtenpanel erstellt, bei dem 2%-Stichproben aus dem Quartalsbeschäftigtenbestand gezogen werden. Bei dieser Stichprobe steht die Aktualität im Vordergrund. Sie ist repräsentativ für die amtlichen Bestandszahlen zu den Quartalsstichtagen.

[18] Infolge des 1969 verabschiedeten Arbeitsförderungsgesetzes wurde mit der 1. „Verordnung über die Erfassung von Daten für die Träger der Sozialversicherung und für die Bundesanstalt für Arbeit (Datenerfassungsverordnung — DEVO)“ vom 24.11.1972 und der 1. „Verordnung über die Datenermittlung auf maschinell verwertbaren Datenträgern im Bereich der Sozialversicherung und der Bundesanstalt für Arbeit

1.1.1981 mit der 2. DEVO und der 2. DÜVO[19] gegeben, mit denen u.a. sichergestellt wurde, dass nur noch eine Versicherungsnummer für jede Person vergeben wird. Außerdem wurden darin Zuständigkeiten in einem mehrstufigen Prüfverfahren geregelt, mit dem die Vollständigkeit und Korrektheit der Meldungen verbessert wurde. Mit Wirksamkeit zum 1.1.1999 wurde das bis dahin gültige Meldeverfahren (DEVO/DÜVO) schließlich durch seit dem 1.1.1999 gültige „Verordnung über die Erfassung und Übermittlung von Daten für die Träger der Sozialversicherung (Datenerfassungs- und -übermittlungsverordnung — DEÜV)“ abgelöst, die zu einer vollständigen Überarbeitung und Neugestaltung des Verfahrens führte und mit der auch geringfügige Beschäftigungsverhältnisse meldepflichtig wurden.[20] Insgesamt hat sich die Erfassung der geringfügig Beschäftigten im Zeitverlauf zweimal geändert. Bis zum 1. Quartal 1999 wurden in den Quartalsdaten nur die sozialversicherungspflichtigen Beschäftigten ausgewiesen. Ab dem 1.4.1999 sind auch die ausschließlich geringfügig Beschäftigten in den Daten enthalten. Zum 1.4.2003 wurde die letzte Änderung wirksam. Galt vorher das Prinzip, dass in den Quartalsdaten nur ein (Haupt-)Beschäftigungsverhältnis pro Person erfasst wurde, wird ab dem 2. Quartal 2003 nun insofern davon abgewichen, dass auch geringfügige Beschäftigungsverhältnisse, die neben einer sozialversicherungspflichtigen Hauptbeschäftigung ausgeübt werden, aufgenommen werden. Diese Änderungen in der Grundgesamtheit wirken sich auch auf die Berechnung der Betriebsgröße und der Beschäftigtenanteile in den Betrieben aus.[21]

Ferner ist bei vergleichenden Analysen zu beachten, dass der Entgeltbegriff sich über die Zeit verändert hat. Dabei wurde das Entgelt „sukzessive auf Bezüge, die neben dem Lohn gezahlt wurden, ausgedehnt, so dass im Zeitverlauf immer mehr Lohnzuschläge dem beitragspflichtigen Entgelt zugerechnet wurden“[22]. So ist beispielsweise seit 1984 vorgeschrieben, dass Lohnsonderzahlungen in das Bruttoentgelt einzubeziehen sind. Geringfügigkeitsgrenze und Beitragsbemessungsgrenze, als Mindestentgelt bzw. maximaler Entgeltbestandteil, die der Sozialversicherung unterliegen, wurden ebenfalls im Zeitverlauf häufig angepasst.[23] Weitere Änderungen im Meldeverfahren ergeben sich bei einer Umstellung der verwendeten Klassifikationen für die Wirtschaftszweige der Unternehmen oder die ausgeübten Berufe der Beschäftigten.

(Datenübermittlungsverordnung — DÜVO)“ vom 18.12.1972 ein neues Meldeverfahren zur Sozialversicherung eingeführt.

[19] Beide vom 29.5.1980.

[20] Vgl. Statistisches Bundesamt vierteljährlich vom 31. März 2008, S. 6.

[21] Vgl. Schmucker und Seth 2009, S. 45.

[22] Vgl. Schmähl und Fachinger 1994, S. 188.

[23] Eine tabellarische Übersicht findet sich z.B. in Schmucker und Seth 2009, S. 37ff.

4.1.2 Fehlereinflüsse in der Beschäftigtenstatistik

Die Merkmale, die in den Meldungen zur Sozialversicherung anzugeben sind, beziehen sich auf die Person oder den Betrieb oder das Beschäftigungsverhältnis. Bis auf die Merkmale Geschlecht und Geburtsjahr sind sie im Prinzip über die Zeit veränderlich; im Laufe ein und desselben Beschäftigungsverhältnisses sind sie jedoch überwiegend konstant.[24] Die Merkmale der Beschäftigtenstatistik können unterschieden werden in diejenigen, die versicherungsrechtlich relevant sind und solche, die ausschließlich statistischen Zwecken dienen.[25] Zu den *versicherungsrechtlich relevanten Merkmalen* gehören die Versicherungsnummer, die Beschäftigungszeit und das versicherungspflichtige Entgelt. Aufgrund dieser Angaben werden die Zahlungen für die Sozialversicherungen ermittelt. Es ist daher davon auszugehen, dass sie von den Arbeitgeber/innen genau angegeben und zudem von den Versicherten überprüft werden. Diese Genauigkeit ist einer der Vorzüge der Beschäftigtenstatistik. Dabei sind die Definitionen der einzelnen Merkmale gesetzlich festgelegt und erfüllen in erster Linie juristische und administrative Anforderungen, so dass eine Identifikation mit volkswirtschaftlichen Begriffsbestimmungen erschwert ist.[26]

Bei den *Meldungen für statistische Zwecke* sind Abstufungen in der Genauigkeit zu vermuten, da verschiedene Stellen für die Ermittlung der Merkmalswerte zuständig sind. Die wirtschaftsfachliche und regionale Zuordnung des Betriebes, also die Merkmale, die sich auf den Betrieb selber beziehen, werden von Fachkräften in den Betriebsnummernstellen der Arbeitsagenturen vorgenommen und in der hauseigenen Betriebsdatei geführt. Trotz der Unschärfe der Klassifizierung selber ist hier von einer zufriedenstellenden Einheitlichkeit und Genauigkeit der Angaben auszugehen.[27]. Die persönlichen Merkmale der Versicherten beruhen hingegen auf Angaben der Arbeitgeber/innen und werden auch dort verschlüsselt. Die Codierung dieser Merkmale wird zwar durch die Anleitungen des Meldeverfahrens gem. DEÜV vorgegeben, jedoch kann eine einheitliche Handhabung nicht erreicht werden. Betriebsspezifische Berufsbezeichnungen können z.B. zu abweichenden Klassifizierungen gleicher Tätigkeiten führen und insbesondere bei kleinen Betrieben werden die Meldungsvordrucke häufig von nicht fachlich geschulten Kräften ausgefüllt. Mit einem Betriebswechsel ändern sich überzufällig häufig auch die Ausprägungen von Variablen, die man irrtümlich als über die Zeit konstant ansehen könnte, z.B. die Ausbildung oder die Stellung im Beruf. Das kann einerseits auf Meldefehler des Betriebes zurückzuführen sein, andererseits auch eine tatsächliche Veränderung des jeweiligen Status be-

[24] Vgl. Bender u. a. 1996, S. 11.

[25] Vgl. Cramer 1985, S. 62ff

[26] Vgl. Bender u. a. 1996, S. 11.

[27] Dennoch gibt es Hinweise darauf, dass die Betriebsklassifizierungen in den AA-Regionen unterschiedlichen Tendenzen folgen können. Vgl. Fritsch u. a. 2002, S. 94.

deuten.[28] Beispielsweise werden Änderungen im Ausbildungsstatus häufig erst bei einem Betriebswechsel neu erfasst.[29]

Grundsätzlich können die Ergebnisse der Beschäftigtenstatistik nicht besser sein als die zum Auswertungszeitpunkt vorliegenden Meldungen. Wermter und Cramer [1988] stellen dazu fest: „Wie groß die Fehlermargen anzusetzen sind, diese Frage ist vor oder in einer angemessenen Zeit nach einer Auswertung nicht zu beantworten".[30] Typische Ursachen für Meldeverzögerungen und -ungenauigkeiten können z.B. sein: unvollständige und nicht fristgerechte Meldungen durch die Betriebe; Verzögerungen in der Datenübermittlung zwischen Krankenkassen, Datenstelle der Rentenversicherungsträger und BA — ggf. verursacht durch das Klärungsverfahren für eine unbekannte Versicherungsnummer — oder die unzutreffende Verwendung von Betriebsnummern außerhalb ihres Geltungsbereichs durch die Arbeitgeber/innen. Da Meldeverzögerungen bzw. -unterlassungen mit ungleichmäßiger Dauer und Verteilung auftreten, die nicht kurzfristig beobachtet werden können, können Konjunktureinflüsse auf die Entwicklung des Beschäftigungsstandes am aktuellen Rand in der Beschäftigtenstatistik nicht bzw. nur eingeschränkt beobachtet werden.

Darüber hinaus ist davon auszugehen, dass es systematische Einflüsse auf die Verzögerung von Meldungen gibt. Bei den ersten Auswertungen der Historik-Datei konnten Cramer und Majer [1991] im Bezug auf die Struktur verspäteter Meldungen zeigen, dass relativ kurze Beschäftigungszeiten nachgemeldet werden „und zwar in stärkerem Maße für Ausländer, Teilzeitkräfte, Frauen und Arbeiter"[31]. Dabei gingen die Meldeverzögerungen vermutlich überwiegend auf versicherungsrechtliche Probleme zurück, für die ein Klärungsverfahren erforderlich wurde. Eine weitere strukturelle Ursache für die Verspätung hing mit der Art des meldenden Betriebes zusammen: „Kleinere Betriebe mit weniger sachkundigem Personal, zudem aus Branchen, in denen sich wegen der dort typischen höheren Fluktuation versicherungstechnische Problemfälle häufen, sind vor allem für die Verzögerung verantwortlich".[32]

Aufgrund des Verfahrens für die Vergabe von Betriebsnummern gibt es Inhomogenitäten bei der Identifizierung und Klassifizierung der statistischen Einheit der „Betriebe". Prinzipiell war beabsichtigt, dass die Betriebsnummern die Einheit der „Arbeitsstätte" aus der Arbeitsstättenzählung 1970 und dementsprechend eine „örtliche Produktionseinheit einschließlich der in der Nähe befindlichen Verwaltungs- und Hilfsbetriebe" abbilden sollten. Verschiedene Niederlassungen eines Betriebes sollten also auch verschiedene Betriebsnummern erhalten. Aufgrund des Widerstandes der Arbeitgeber/innen wurde

[28] Vgl. Schmucker und Seth 2009, S. 94.
[29] Vgl. ebd. S. 55.
[30] Vgl. ebd. S. 480.
[31] Vgl. Cramer und Majer 1991, S. 83.
[32] Vgl. Cramer und Majer 1991, S. 83.

schließlich eine flexiblere Regelung für die Betriebsidentifizierung vorgesehen. Es ist daher möglich, dass Niederlassungen in derselben Gemeinde, die demselben Wirtschaftszweig angehören, unter derselben Betriebsnummer zusammengefasst sind. Ebenso ist es möglich, dass eine Niederlassung auf Antrag der Arbeitgeber/innen aus praktischen Gründen mehrere Betriebsnummern erhält, z.B. getrennt für Angestellte und Arbeiter/innen.[33] In den Vorstudien zum IAB-Betriebspanel wurde deutlich, dass die Vergabepraxen für die Betriebsnummer sich zwischen den Branchen unterscheiden. So bezieht sich etwa im verarbeitenden Gewerbe die Betriebsnummer in fast allen Fällen auf eine betriebswirtschaftlich sinnvoll interpretierbare Einheit. Vor allem in der öffentlichen Verwaltung, bei den Sozialversicherungsträgern sowie den Organisationen mit Erwerbscharakter treten hingegen häufig Abgrenzungen auf, die für Analysen auf Betriebsebene wenig Sinn machen.[34] Im Bezug auf Längsschnittbetrachtungen ist außerdem zu beachten, dass die Betriebsnummer im zeitlichen Verlauf auf unterschiedliche Betriebe verweisen kann. Im Grundsatz ist die Betriebsnummer persönlich an die Inhaber/in des Betriebes gebunden, nicht aber an einen Einzelbetrieb als solchen. Mögliche Verläufe können z.B. sein: Wechsel der Betriebsnummer bei Wechsel der Inhaber/in bzw. Wechsel der Rechtsform, Neugründung eines Betriebes unter derselben Betriebsnummer, Integration mehrerer Betriebe oder Ausgliederung von Unternehmensteilen sowie veränderte Zuordnung der Beschäftigten auf mehrere Nummern.[35]

Da die Beschäftigungsentwicklung in einer Region häufig von wenigen Großbetrieben dominiert wird, werden bei Veränderungen der Wirtschaftszugehörigkeit oder Betriebsverlagerungen erhebliche Effekte ausgelöst, die singulär auftreten. Dies ist besonders dann relevant, wenn solche Betriebe ungünstig abgegrenzt oder zugeordnet sind.

Mit der Einführung der DEÜV am 1.1.1999 wurde die Abgabe der Meldungen zur Sozialversicherung per EDV als Standard eingeführt. Außerdem wurde die Aufbereitung der Beschäftigtenstatistik in der BA in einem „data warehouse“ neu aufgebaut, wobei die zu verarbeitenden Einzeldaten zunächst zu Aggregaten verdichtet werden, die für die Standardauswertungen ausreichend sind. Dadurch können die großen Datenbestände schneller und flexibler als bisher ausgewertet werden.[36] Allgemein konnte durch die Automatisierung die Vollständigkeit und Korrektheit der Daten verbessert werden. Insgesamt wird im aktuellen Qualitätsbericht zur Beschäftigtenstatistik konstatiert, dass die Qualität der Daten — aufgrund der bestehenden Meldepflicht und der mehrstufigen Prüf- und Korrekturverfahren — für statistische Zwecke als „sehr gut eingeschätzt“ wird. Fehlerhafte Angaben in den Meldungen werden grundsätzlich nur bei den Angaben zum Beruf er-

[33] Vgl. [Cramer, 1985].
[34] Vgl. Fritsch 1997, S. 132.
[35] Vgl. [Fritsch, 1997].
[36] Vgl. Lüken 2002, S. 52.

wartet. Fehlende Einheiten (d.h. unit-non-response) gibt es nur in den Fällen, in denen die letzte Beschäftigungsmeldung schon länger zurückliegt und keine Abmeldung erfolgt ist. Solche unvollständigen Kontenverläufe existieren in geringem Umfang. Fehlende Meldungen (d.h. item-non-response) kommen hingegen häufiger vor. Relevant ist die Anzahl fehlender Meldungen für das Merkmal Ausbildungsstand: derzeit werden für über 15% der Beschäftigten keine konkreten Angaben zum Bildungsabschluss gemeldet.[37] Da die Fehlermodellierungen im nächsten Abschnitt als exemplarische Beispiele dienen sollen, wird die Relevanz der Fehlereinflüsse allerdings im Folgenden vernachlässigt. Ausschlaggebend für die Auswahl ist eher, ob die Merkmalsvariablen zur Illustration verschiedener Fehlertypen herangezogen werden können.

4.2 Fuzzy-Modellierung von Merkmalen

Aufbauend auf die verfügbaren Daten der Beschäftigtenstatistik werden in diesem Abschnitt Fuzzy-Merkmalsvariable konstruiert, die den tatsächlichen Informationsstand über den Fehler in den Beobachtungen möglichst gut wiedergeben sollen. Mit der Fuzzifizierung soll die Ungenauigkeit der Messwerte durch eine Zugehörigkeitsabschätzung der möglichen Lage des „wahren“ Merkmalswertes abgebildet werden. Die Fuzzy-Daten sind daher überwiegend epistemisch zu interpretieren. Es ist also z.B. konkret zu bewerten, wie sich die Meldeprobleme, die Inhomogenitäten bei der Betriebsnummernvergabe oder die Lerneffekte bei einer Umstellung des Meldeverfahrens bei der gegebenen Problemstellung in den möglichen Werten einer Merkmalsvariablen niederschlagen.

Wie in Definition 2.21 in Abschnitt *2.5.1* festgehalten wurde, sollen im Hinblick auf die Fuzzy-Regression zunächst nur Merkmale mit metrischer Skala betrachtet werden, für die die zugehörigen Fuzzy-Merkmale ebenfalls als metrisch skaliert angesehen werden können. Es werden aber nur einige Erhebungsmerkmale der Beschäftigtenstatistik auf einer metrischen Skala gemessen, nämlich Beschäftigungsbeginn, Beschäftigungsende, sozialversicherungspflichtiges Bruttoentgelt und Alter. Daher werden in dieser Arbeit überwiegend abgeleitete Merkmale, wie etwa Bestandszahlen, untersucht. Diese sind deutlich von der Güte der verwendeten Berufs- oder Wirtschaftszweigklassifikationen beeinflusst.[38] Die Modellierung sollte vor allem die relevanten Unschärfen wiedergeben, die die Genauigkeitsschwelle überschreiten. Ein Vorschlag, mit dem darüber hinaus die Genauigkeitsschwelle bei der Modellierung explizit einbezogen werden kann, wird in Abschnitt *4.2.6* vorgestellt.

Die Fuzzy-Merkmalswerte werden hier nicht in Form von LR-Fuzzy-Zahlen gemäß Definition 2.6 spezifiziert, sondern als gestuft trapezförmige Fuzzy-Mengen S-ter Ordnung,

[37] Vgl. Statistisches Bundesamt 2009, S. 5f.

[38] Darüber hinaus wird klassifiziert nach: Geschlecht, Staatsangehörigkeit, Beschäftigungsort, Wohnort, Stellung im Beruf, Ausbildung, Vollzeit/Teilzeit, Familienstand.

d.h. die Seiten der α-Niveaus können als lineare Polygonzüge mit (höchstens) $S-1$ Unstetigkeitspunkten dargestellt werden. Als prototypisches Beispiel einer trapezförmigen Fuzzy-Menge 2-ter Ordnung kann die Datenmodellierung in Abbildung 4.5 auf Seite 142 angesehen werden. Die Modellierung der Flanken der Fuzzy-Mengen als lineare Polygonzüge lassen eine flexible Annäherung an beliebige Referenzfunktionen zu und sind überdies leicht zu programmieren. Da die Fuzzy-Trapeze S-ter Ordnung im wesentlichen trapezförmige Fuzzy-Zahlen sind, können die Berechnungsmethoden der linearen Regression, die in Kapitel *3* vorgestellt wurden, ohne großen Aufwand auf die konstruierten Fuzzy-Merkmalswerte übertragen werden.

Zur Konstruktion der Fuzzy-Merkmalswerte muss der Gesamtfehler abgeschätzt und quantifiziert werden. Dazu werden die bekannten, wesentlichen Fehlereinflüsse identifiziert. Für die Konstruktion der Fuzzy-Daten wird zum einen die Menge der sicher möglichen Punkte bestimmt, die den Kern der Fuzzy-Menge ausmachen, und zum anderen die Menge der gerade noch möglichen Punkte, die die Trägermenge bilden. Der Kontrast für den Übergang zwischen möglichen und nicht-möglichen Punkten wird dann entweder durch die Vorgabe von Übergangspunkten auf den kritischen α-Niveaus für $1 < s < S$ oder durch die Vorgabe von Referenzfunktionen mit den zugehörigen Spannweiten festgelegt. In den folgenden Abschnitten werden verschiedene Vorschläge für die Konstruktion von empirischen Fuzzy-Daten vorgestellt. Schließlich werden die Eigenschaften der vorgeschlagenen Fuzzy-Modellierungen diskutiert, dabei werden unter anderem Ansätze für die Verallgemeinerung der Modellierungsverfahren erörtert. Das Ziel der Beispielsammlung liegt dabei weniger in der Güte der konstruierten Fuzzy-Merkmale selber als darin, verschiedene Konstruktionskonzepte aufzufächern und aufzubereiten, die zum Weiterdenken anregen sollen.

4.2.1 Auswirkung fehlender Meldungen — Individuelle Fuzzy-Zugehörigkeit zum Beschäftigtenbestand

Als Beispiel wird das Merkmal der *individuellen Zugehörigkeit zum Beschäftigtenbestand* bei der Quartalsauswertung betrachtet. Die Auswertung findet mit einer Wartezeit von mindestens sechs Monaten nach dem Stichtag statt.

Der Vollständigkeitsgrad der abgegebenen Meldungen in den Meldebeständen für die Stichtage unterliegt Schwankungen. Die Meldeprobleme führen auf der Ebene der Bestandskonten dazu, dass sozialversicherungspflichtig Beschäftigte fälschlich nicht im Bestand aufscheinen oder dass sie fälschlich weiterhin im Bestand geführt werden. Es gibt allerdings keine Möglichkeit, die Zahl der „unterlassenen Meldungen“ auf der Grundlage

des vorhandenen Datenmaterials festzustellen.[39] Dabei fallen die Meldeprobleme vor allem bei den Quartalsauswertungen und den Jahresauswertungen ins Gewicht, so dass von einem annähernd vollständigen Meldeeingang erst nach einer Wartezeit von drei Jahren ausgegangen werden kann. Zu beachten ist, dass verzögerte Meldungen zu Fehlern in der Wiedergabe der Struktur des Beschäftigtenbestandes führen. Auch diese können nicht aktuell identifiziert und somit nicht korrigiert werden.

Fehlende Meldungen können in den Versicherungskonten mit logischen Prüfalgorithmen zu einem sehr großen Teil ausgeglichen werden, indem Daten aus einer Ersatzmeldung ergänzt werden. Daher ist es sinnvoll, die Datenunschärfen durch die Meldeprobleme zunächst auf der Ebene der individuellen Versicherungskonten zu beschreiben und erst in einem weiteren Schritt die Aggregation zu einer Bestandszahl zu betrachten. Ansätze zur Fuzzy-Modellierung des aggregierten Beschäftigtenbestandes werden im nächsten Abschnitt *4.2.2* untersucht.

Grundsätzlich besteht nur dann Sicherheit über ein bestehendes Beschäftigtenverhältnis am Stichtag, wenn die vorliegenden Meldungen auf dem Versicherungskonto die Beschäftigung bis zu einem Datum anzeigen, das über den Stichtag hinausreicht. Falls die letzte vorliegende Meldung nur Zeiten vor dem Stichtag erfasst, kann die Beschäftigung somit nur bedingt festgestellt werden, wobei die Ungewissheit mit wachsendem Abstand zunimmt.[40] Eine Anmeldung ist nur eine unsichere Meldung, da sie nur den Beschäftigungsbeginn, nicht aber eine Beschäftigungsdauer anzeigt und es vorkommen kann, dass trotz Anmeldung keine Beschäftigung erfolgt.[41] Die Beschäftigung kann erst dann als sicher gelten, wenn eine Anmeldung durch eine nachfolgende zahlungsrelevante Meldung — eine der sog. „Entgeltmeldungen“ Unterbrechungs-, Jahres- oder Abmeldung — bestätigt wird. Diese fungieren daher auch als Ersatzmeldungen für fehlende Anmeldungen. Hingegen gibt es für fehlende Abmeldungen keine Ersatzmeldungen, die das Ende der Beschäftigung anzeigen. Von besonderer Relevanz sind dabei Beschäftigungsverhältnisse, die während oder zum Ende einer Unterbrechung beendet werden. Da eine Abmeldung dann keine Zahlungsrelevanz mehr hat, unterbleibt sie häufig. Unterbrecher/innen werden dann fälschlicherweise zu lange im Bestand mitgeführt. Durch fehlenden Abmeldungen nach Zeiten der Beschäftigungsunterbrechung, insbesondere bei Unterbrechungen durch Elternzeit z.T. auch durch Wehr- und Zivildienst, werden Verzerrungen in größerem Umfang verursacht.[42]

[39] Vgl. Wermter und Cramer 1988, S. 477.

[40] Die Bestandszugehörigkeit sinkt bei genauer Betrachtung also auch für real fortlaufende Beschäftigungen während des Jahres, bis sie durch die Jahresmeldung bestätigt werden und die Zugehörigkeit wieder auf 1 springt.

[41] Anmeldungen werden bei der Quartalsauswertung aufgrund der vergleichsweise kurzen Wartezeit mit einbezogen.

[42] Vgl. Wermter und Cramer 1988, S. 469ff. Im Jahr 1986 wurde der Erziehungsurlaub von sechs auf zehn

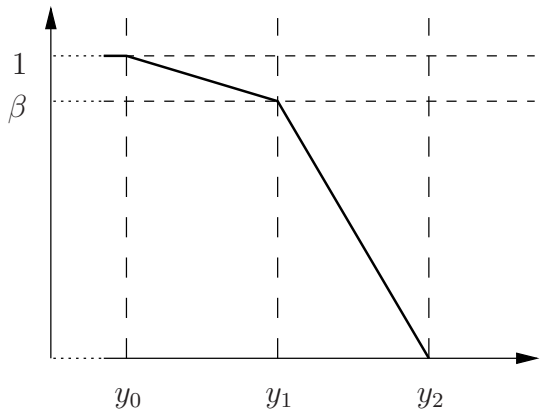
Für die Modellierung eines Fuzzy-Wertes ist zu beurteilen, wie die Verringerung der Gewissheit durch Meldefehler im individuellen Versicherungskonto identifiziert und quantifiziert werden können. Über fehlende Anmeldungen in einem Versicherungskonto, insbesondere über fehlende Anmeldungen von bisher nicht registrierten Personen, liegen am aktuellen Rand keine Informationen vor, sie können folglich nicht dargestellt werden. Fehlende Abmeldungen sind dann wahrscheinlich, wenn eine Jahresmeldung fehlt oder wenn eine neue Meldung durch eine andere Arbeitgeber/in erfolgt. Außerdem kann eine vorliegende Unterbrechungsmeldung als Indiz für die erhöhte Möglichkeit einer fehlenden Abmeldung gewertet werden. Meldeausfälle sind überwiegend mit Strukturkomponenten verknüpft, z.B. werden Beschäftigte in Kleinstbetrieben zu einem größeren Teil verspätet oder gar nicht gemeldet, und treten häufig nur punktuell auf, z.B. können Meldeverzögerungen durch einen Verarbeitungsfehler in einer Krankenkasse verursacht sein.

Die abnehmende Möglichkeit eine Beschäftigung mit wachsendem zeitlichen Abstand von der letzten vorliegenden Meldung kann quantitativ mit Hilfe von externen Informationen über die charakteristische Fortdauer der Beschäftigung beschrieben werden. Damit kann eine Teilzugehörigkeit der jeweiligen Beschäftigten zum Beschäftenbestand an einem Kalendertag bestimmt werden. In Anlehnung an die Datenbereinigung in der Historik-Datei wird zur Modellierung der Datenunschärfen bei einer Quartalsauswertung der Beschäftigtenstatistik der folgende Ansatz verwendet. Die Modellierung basiert auf den Annahmen, dass eine Meldeverzögerung nicht länger als ein Jahr dauert und dass höchstens eine (Jahres-)Meldung verloren geht.[43]

Eine unscharfe Modellierung wird dann ausgelöst, wenn im Versichertenkonto (noch) eine Jahresmeldung fehlt, der keine weiteren Meldungen folgen. Ist die letzte vorliegende Meldung eine Jahresmeldung für y_0, so kann angenommen werden, dass die Zugehörigkeit zum Bestand zwischen y_0 und dem Stichtag der fehlenden Jahresmeldung y_1 linear bis auf

Monate verlängert. Eine Aufstellung der Änderungen der Beurlaubungsfristen für die Elternzeit bis 1996 ist in Bender u. a. 1996, S. 25 zu finden.

[43] Die Bereinigung um „Altfälle“ erfolgte bis Ende der 80er Jahre zunächst mit einem mechanischen *Abschneideverfahren*. Dabei wurde von den Fällen, für die länger als ein Jahr keine Meldungen mehr eingegangen waren, in chronologischer Reihenfolge so viele ausgesondert bzw. ggf. auch wieder eingegliedert, wie es dem Saldo zwischen den neu gemeldeten Beschäftigten und den abgemeldeten Beschäftigten entsprach. Vgl. Wermter und Cramer [1988], S. 479.
Der Ansatz basierte auf der Beobachtung, dass die Zahl der Anmeldungen im Saldo die der Abmeldungen übersteigt und deren Summe über einen Zeitraum von zwei Jahren in etwa die Zahl der offengebliebenen Konten reproduziert. Die ersten Auswertungen mit der Historik-Datei zeigten aber, dass durch das Abschneideverfahren zusätzliche Verzerrungen in der Beschäftigtenstruktur verursacht wurden und es darüber hinaus dazu führte, dass Beschäftigtenanstiege während eines Konjunkturaufschwungs nicht erkennbar wurden. Eine wesentliche Stärke der Beschäftigtenstatistik ist aber gerade die Genauigkeit in der Abbildung der regionalen, wirtschaftsfachlichen und beruflichen Beschäftigtenstruktur. Das Interpolationsverfahren für fehlende Meldungen sollte daher möglichst keine zusätzlichen unbekannten Verzerrungen in den Bestandsdaten erzeugen. Infolgedessen wurde schließlich auf die Datenbereinigung mit dem Abschneideverfahren verzichtet.



(Eigene Darstellung)

Abbildung 4.1: Individuelle Zugehörigkeit zum Bestand bei fehlender Jahresmeldung in y_1 und y_2 — Zu beachten: Die Modellierung erfolgt erst für Stichtage $t > y_1$.

einen Schwellenwert β abnimmt, der im Zusammenhang mit dem Anteil „typischerweise" fehlender Folgemeldungen gewählt werden kann. Bis zum Ende des zweiten Folgejahres y_2 kann eine bis Null sinkende Zugehörigkeit unterstellt werden. Anhand der so konstruierten Fuzzy-Menge über der Zeitleiste, die die individuelle Bestandszugehörigkeit einer Beschäftigten darstellt, kann an jedem Stichtag s für jedes Individuum i eine Zugehörigkeit zum Beschäftigtenbestand $\mu_{s,i} \in [0,1]$ ausgewertet werden. Die Modellierung ist in Abbildung 4.1 illustriert.

Der 2-Jahres-Zeitraum für das Abklingen der Zugehörigkeit ist dadurch zu begründen, dass erst nach zwei Jahren die Kontrollzyklen des Meldeverfahrens überwiegend abgeschlossen sind und eine annähernde Vollständigkeit in den Meldeeingängen erreicht wird.[44] Es liegen Strukturstatistiken über Abweichungen in der Meldedisziplin bei spezifischen Beschäftigtengruppen bzw. Betriebstypen vor. Längere Laufzeiten und fehlende Meldungen sind demnach typisch für Beschäftigungsverhältnisse von Frauen und AusländerInnen sowie für Meldungen von Kleinstbetrieben. Die Modellierung des Effekts von fehlenden Abmeldungen kann damit ggf. noch modifiziert werden.

Ist die letzte vorliegende Meldung eine Unterbrechungsmeldung, dann kann entsprechend eine linear bis Null fallende Zugehörigkeit für das Jahr nach der Unterbrechungsmeldung unterstellt werden, sofern in einem Zeitfenster von einem Jahr danach keine Folgemeldungen vorliegen.[45]

[44] Es ist zu vermuten, dass sich durch die aktuellen Änderungen im Erhebungsverfahren (z.B. Ausweitung des elektronischen Meldeverfahrens, Verschmelzen der Rentenversicherungsträger) erhebliche Verbesserungen in der Meldegenauigkeit ergeben haben. In der Einführung zur Fachserie 1 „Bevölkerung und Erwerbstätigkeit", Reihe 4.2.1 „Struktur der sozialversicherungspflichtigen Beschäftigten" vom 31.12.2006 wird allerdings weiterhin darauf hingewiesen, dass die vierteljährlichen Bestandsergebnisse verfahrensbedingt für einen Zeitraum von drei Jahren als vorläufig gelten und von der BA gegebenenfalls noch korrigiert werden (S. 4).

[45] Bei Anmeldung durch eine neue Arbeitgeber/in innerhalb eines kürzeren Zeitraums kann gegebenenfalls

Der so bestimmte individuelle Fuzzy-Zugehörigkeitswert $\mu_{t,i}$ ist eine Bewertung dafür, zu welchem Grad das Individuum i zum gegebenen Stichtag t dem Beschäftigtenbestand angehört. Der Wert kann also als Klassenzugehörigkeit interpretiert werden. Im Unterschied zu einem Gewichtungsfaktor, der den Einfluss des Individuums auf einer kardinalen Skala abbildet, die eine rechnerische Übertragung der Verhältnisse ermöglicht, repräsentieren die Zugehörigkeitsgrade aber nur eine ordinale Anordnung. Die individuellen Bestandszugehörigkeiten können daher nicht sinnvoll arithmetisch verknüpft werden. Sie sind eher als Vorstufe für die Bestimmung von aggregierten Fuzzy-Merkmalen zu sehen, wie es etwa die Auszählung des Fuzzy-Beschäftigtenbestandes darstellt, die in Abschnitt *4.2.2* beschrieben wird. In ähnlicher Weise wie die üblichen Gewichtungsfaktoren können die individuellen Zugehörigkeitsgrade auch als Genauigkeitsinformation in die Bewertung von anderen Merkmalswerten übertragen werden. Dies kann besonders bei Merkmalsauswertungen interessant sein, die sich auf Teilbestände beziehen. So kann es etwa von Interesse sein, den Zugehörigkeitsgrad zum Beschäftigtenbestand auf das Merkmal „Ausbildungsstand“ zu übertragen, um diesen bei der Bestimmung von ausbildungsstandbezogenen, abgeleiteten Betriebsgrößenklassen ebenfalls zu berücksichtigen.[46]

Entsprechend wie bei dem Bereinigungsverfahren bei der Erstellung der Historik-Datei bleiben bei der individuellen Fuzzy-Bestandszugehörigkeit die Strukturinformationen über die gemeldeten Beschäftigten erhalten und werden darüber hinaus durch die Bewertung der Meldesicherheit von Abmeldungen ergänzt. Eine Ergänzung von Informationen über fehlende Anmeldungen ist auf der Ebene der individuellen Versicherungskonten nicht bzw. nur in der Tendenz[47] möglich. Diese können ggf. bei der Auszählung von aggregierten Teilbeständen abgeschätzt werden, dort aber zu Lasten der Differenzierung nach den Strukturmerkmalen.

Insgesamt ist die Modellierung der individuellen Fuzzy-Zugehörigkeit zum Beschäftigtenbestand ein Beispiel für die Fuzzifizierung einer binären Variablen, über deren aktuellen „Schaltzustand“ Ungewissheit besteht. Die Modellierung beruht dabei stark auf der Kontextinformation über die „Schaltvorgänge“, hier z.B. die typischen Eigenschaften des Meldeverfahrens. Die Zugehörigkeitswerte sind nur im Rahmen eines ordinalen Vergleichs interpretierbar und können bei der Bildung von Teilpopulationen als Information über die Klassenzugehörigkeit genutzt werden.

---

eine schneller fallende Zugehörigkeit unterstellt werden, falls nicht ein nahtloser Beschäftigungswechsel anzunehmen ist.

[46] Dabei wäre zu unterstellen, dass die Angaben für den „Ausbildungsstand“ unabhängig von den Strukturmerkmalen sind, die in die Bestandszugehörigkeit einfließen.

[47] Es ist z.B. möglich, durchschnittliche Verspätungsdaten als „tendenzielle Zahl von fehlenden Anmeldungen“ einzubeziehen. Der Vorteil einer Fuzzy-Modellierung ist dann, dass die aktuelle Bestandsinformation mit dem Expert/innenwissen gewichtet werden kann. Daumenregeln können so visualisiert und objektiviert werden.

4.2.2 „Beschäftigtenbestand“ — aggregierte Fuzzy-Bestandsauszählung

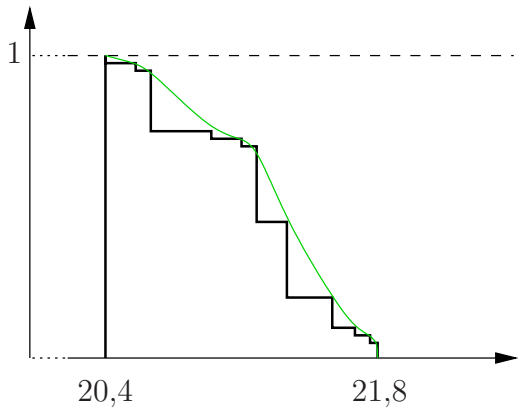
Als Beispiel wird das Merkmal der *aggregierten Zugehörigkeit zum Beschäftigtenstand* bei der Quartalsauswertung betrachtet. Die Auswertung findet mit einer Wartezeit von mindestens sechs Monaten nach dem Stichtag statt.

Aus der Beschäftigtenstatistik wird insbesondere der Beschäftigtenstand zu gegebenen Stichtagen ermittelt. Eine erste verlässliche Auszählung der Bestandsdaten wird bei den Quartalsauswertungen nach einer Wartezeit von sechs Monaten durchgeführt. Aufgrund verspäteter Meldeeingänge gelten die Bestandszahlen aber für weitere drei Jahre als vorläufig und werden gegebenenfalls angepasst, so dass das Wissen über die Fehlereinflüsse in den Bestandszahlen sich mit erhöhter Wartezeit verändert. Durch den unsicheren Eingang der Meldungen entstehen Niveaufehler in der Beschäftigtenentwicklung. *Positive Niveaufehler* entstehen durch die fehlenden oder verzögerten Abmeldungen. Aus den nicht eingegangenen Anmeldungen entstehen demgegenüber *negative Niveaufehler*. Für diese ist strukturell relevant, dass besonders häufig Anmeldungen von Berufsanfänger/innen von Verzögerungen betroffen sind, da hierbei noch das Kontroll- und Vergabeverfahren für fehlende Versicherungsnummern durchgeführt werden muss.[48] Verzögerte Anmeldungen bewirken, dass sich eine Zunahme der Beschäftigung bei einem Konjunkturaufschwung erst nach einer deutlichen Wartezeit in der Beschäftigtenstatistik niederschlägt. Eine Auswertung der Historik-Datei zeigte, dass zwischen den entstehenden negativen und positiven Niveaufehlern keine Korrelation beobachtet werden kann, d.h. der Saldo der Niveaufehler ist aktuell nicht bestimmbar und kann positiv oder negativ ausfallen.[49]

Durch die in Abschnitt *4.2.1* eingeführte individuelle Fuzzy-Zugehörigkeit zum Beschäftigtenbestand, werden die Auswirkungen der Meldefehler abgebildet, soweit sie im individuellen Versicherungsverlauf beschreibbar sind. Damit kann hauptsächlich der *Fortschreibungsfehler* korrigiert werden, der daraus entsteht, dass der Beschäftigtenbestand infolge von nicht erfolgten Abmeldungen systematisch überzeichnet wird und dessen Effekt sich über die Folgeperioden kumuliert. Um eine zunehmende Überzeichnung der Bestandsdaten zu verhindern, ist eine solche zeitnahe Korrektur des Fortschreibungsfehlers besonders wichtig.

Es ist zu beachten, dass verzögerte Meldungen auch zu Fehlern in der Wiedergabe der Struktur des Beschäftigtenbestandes führen können. Beispielsweise wenn die Wirtschaftszweigzugehörigkeit einer Person, die real nicht mehr beschäftigt ist, weiterhin ausgewiesen wird, hingegen diejenige einer Person, die eine Beschäftigung neu aufgenommen

[48] Vgl. Wermter und Cramer 1988, S. 477.
[49] Vgl. [Cramer und Majer, 1991].



(Eigene Darstellung)[51]

Abbildung 4.2: Fuzzy-Beschäftigtenbestand am 30.9.1984 in Mio. Beschäftigte

hat, noch nicht. Diese Fehler können kaum aktuell identifiziert und somit nicht korrigiert werden. Außerdem entstehen in einzelnen Stichtagen spezifische Effekte: die Bilanzfunktion der Jahresmeldung führt zu verstärkten Korrekturen für fehlende Meldungen, die häufig zu einer starken Erhöhung der gemeldeten Beschäftigten führt; Veränderungen in der Verteilung der Beschäftigten auf die Wirtschaftszweige, insbesondere aber in der Betriebsstruktur des sekundären Betriebspanels ergeben sich z.B. auch durch die regelmäßig durchgeführten Betriebsnummernrevisionen.

Zur Konstruktion des Fuzzy-Merkmalswertes werden hier zunächst die individuellen Zugehörigkeiten zum Stichtagsbestand ausgezählt. In einen zweiten Schritt können die Fuzzy-Bestandszahlen dann noch um weitere Fehlerinformationen modifiziert werden.

Für einen gegebenen Stichtag liegen uns für alle Individuen $i \in \{1, \dots N\}$, für die ein Versichertenkonto existiert, individuelle Bestandszugehörigkeiten μ_i vor. Die individuellen Zugehörigkeiten können ohne Einschränkung in fallender Reihenfolge angeordnet werden:

$$(\mu_i), \quad \text{so dass} \quad 1 \geq \mu_1 \geq \cdots \geq \mu_N > 0.$$

Damit können wir, wie in Abbildung 4.2 dargestellt, den Fuzzy-Bestand $\widetilde{B}$ definieren durch

$$\mu_{\widetilde{B}}(n) = 1_{[B_0,N]}(n)\,\mu_n, \qquad \text{wobei} \quad B_0 = \max\{i \,|\, \mu_i = 1\}.^{50}$$

Dabei sei allgemein für $a, b \in \mathbb{R}$ mit $a \leq b$ mit $1_{[a,b]}$ die Indikatorfunktion für das Intervall $[a, b]$ bezeichnet, die für Elemente von $[a, b]$ den Wert 1 annimmt und ansonsten verschwindet.

[50] Diese Definition entspricht — bis auf die Optimierung der Zugehörigkeitswerte für eine passgenaue Bewertung — der erweiterten Addition. Die Verwendung von $\oplus$ würde in diesem Fall zu einem sehr kleinen, konstanten Zugehörigkeitswert $\mu_c = \min\{\mu_i \,|\, 1 \leq i \leq N\}$ führen.

Die Treppenfunktion kann ähnlich wie in Abbildung 4.2 durch eine geglättete Form des Fuzzy-Bestandswertes ersetzt werden, die dann günstigere Eigenschaften für die Anpassung der Regressionsfunktion aufweist. Der so bestimmte aggregierte Fuzzy-Wert für den Rohbestand an Beschäftigten kann ggf. um weitere Einflüsse modifiziert werden, die nur für aggregierte Bestände modelliert werden können.

Bei Daten vor 1999[52] können z.B. die „typischen" Anteile von fehlenden Abmeldungen nach einer Unterbrechungsmeldung nachträglich korrigiert werden, so dass die Spezifika bestimmter Beschäftigtengruppen — wie Frauen, Ausländer/innen, Arbeiter/innen, Beschäftigte in Kleinstbetrieben — besser repräsentiert werden. Mit Hilfe von Expert/innenwissen werden dazu für den jeweiligen Teilbestand Modifizierungsfunktionen[53] m^L oder $m^R : [0,1) \longrightarrow [-1,1]$ modelliert, die für jedes α-Niveau angeben, wie der Verlauf der linksseitigen bzw. der rechtsseitigen Zugehörigkeiten im aggregierten Teilbestand gegenüber der Annahme für die individuelle Bestandszugehörigkeit in Abschnitt *4.2.1* zu korrigieren ist. Die Modifizierungsfunktion kann z.B. auf der Basis von historischen Daten über die chronologischen Häufigkeiten der fehlenden Abmeldungen nach einer Unterbrechung für den entsprechenden Teilbestand konstruiert werden. Die Bestandszugehörigkeit auf dem Niveau α wird dann auf der rechten Seite des Fuzzy-Bestandes um $m^R(\alpha) < 0$ nach unten korrigiert. D.h. für $w(\alpha) = \sup[\widetilde{B}]^\alpha$ setzen wir $\mu^R_{\widetilde{B}_{mod}}(w(\alpha)) = \max\{\min\{\mu_{\widetilde{B}}(w(\alpha))+m(\alpha), 1\}, 0\}$. Dabei muss allerdings ggf. durch ein Glättungsverfahren sichergestellt werden, dass $\widetilde{B}_{mod}$ wieder eine konvexe Fuzzy-Menge ist. Abbildung 4.3 zeigt in verallgemeinerter Form, wie Fuzzy-Merkmalswerte durch eine Modifikation mit m^L und m^R kalibriert werden können.

Falls in relevantem Umfang die Abmeldungen von Frauen unzuverlässig erfolgen, dann könnte beispielsweise die Bestandszahl für Beschäftigungsverhältnisse von Frauen mit der Modifikationsfunktion $m^R_F(\alpha) = -c_F\alpha^2$ mit positiver Dämpfungskonstante $c_F \in [0,1]$ angepasst werden. Der modifizierte Fuzzy-Bestand an beschäftigten Frauen ist dann gegeben durch

$$\mu_{\widetilde{B}_{mod}}(z) = \begin{cases} \mu_{\widetilde{B}}(z) + m^R_F(\mu_{\widetilde{B}}(z)) & \text{für } z > \sup[\widetilde{B}]^1 \\ \mu_{\widetilde{B}}(z) & \text{für } z \leq \sup[\widetilde{B}]^1. \end{cases}$$

[51] Die Grenzen des Supports sind [Cramer und Majer, 1991] entnommen (S. 84).

[52] Spätestens seit 1999 werden mehrere Typen von Unterbrechungen in den Meldungen erhoben, vgl. Anlageband zu Schmucker und Seth 2006, S. 95ff.

[53] In [Bandemer und Näther, 1992] werden *modifizierte Fuzzy-Daten* dadurch definiert, dass sie für ein „einfaches" Fuzzy-Datum $\widetilde{A} \in F^{nob}_{coc}(\mathbb{R})$ durch eine Modifizierung der Zugehörigkeitsfunktion $f(\mu_{\widetilde{A}})$ für $f : [0,1] \longrightarrow [0,1]$ gewonnen werden können, dabei gilt nicht notwendig f surjektiv. Die Definition bezieht sich allerdings hauptsächlich auf lingustische Fuzzy-Daten, um aus der Aussage „u ist $\widetilde{A}$" die modifizierte Aussage „u ist sehr $\widetilde{A}$" ableiten zu können (S. 99). Der hier entwickelte Begriff der Modifizierungsfunktion kann nicht damit identifiziert werden, sondern stellt demgegenüber eine Weiterentwicklung dar.

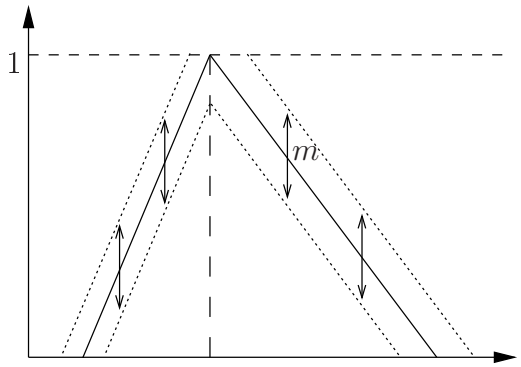


Abbildung 4.3: Kalibrieren von Fuzzy-Merkmalsdaten mit Modifizierungsfunktionen m

Es ist sinnvoll, für jeden relevanten Teilbestand eine spezifische Modifizierungsfunktion zu bestimmen, damit die jeweilige Mischung der Korrektureinflüsse abgebildet werden kann. Dieses Vorgehen hat aber im allgemeinen zur Folge, dass die Summe über die gruppenspezifischen Bestandszahlen nicht mehr gleich der Fuzzy-Zahl für den Gesamtbestand ist, sondern sich je nach Abschätzung der jeweiligen Korrekturfunktionen unterscheidet.[54] Eine Korrektur für verzögerte Anmeldungen im Fuzzy-Beschäftigtenbestand ist ebenfalls nur dann möglich, wenn externe Informationen über die aktuelle Entwicklung der Anmeldungen vorliegen. Sie sind in der Auswirkung auf den aggregierten Bestand vergleichbar mit der Korrektur um fehlende Abmeldungen nach Unterbrechungen.[55]

Bei den Quartalsbeständen, bei denen auch die Anmeldungen gewertet werden, kann es sinnvoll sein, den Anteil an „faulen“ Anmeldungen zu korrigieren. Damit sind Anmeldungen gemeint, auf die tatsächlich keine Beschäftigung folgt. Infolge von „faulen“ Anmeldungen verringert sich der Bestandswert gegenüber dem vorliegenden Wert. Die Zahl der „faulen“ Anmeldungen kann durch die unteren Intervallgrenzen der α-Niveaus modelliert werden und führt zu einer wachsenden linken Spannweite mit fallendem Zugehörigkeitsniveau.[56]

Die Modellierung der aggregierten Bestandszahl der Beschäftigten ist ein Beispiel für die Konstruktion eines Fuzzy-Auszählungswertes für die Elemente einer Fuzzy-Klasse. Das Verfahren zur Bestimmung der Fuzzy-Bestandszahl kann dadurch motiviert werden, dass zum Wert des sicheren Bestandes stückweise die Anzahl der Elemente mit jeweils der nächsthöheren Zugehörigkeit hinzuaddiert werden, so dass das Supremum der Trä-

[54] Alternativ zu diesem Vorgehen können auch im jeweiligen Bestandskonto spezifische Fuzzy-Verlaufscharakteristika für die entfallene Abmeldungen in Abhängigkeit zur Zugehörigkeit zu einer Teilgruppe ergänzt werden. Nach diesem Ansatz sind die Fuzzy-Werte der Teilbestände und des Gesamtbestandes per Definitionem miteinander verträglich. Das Verfahren ist jedoch sehr aufwändig.

[55] Dies wäre besonders interessant, um die Unterschätzung der Beschäftigungsspitzen im September zu korrigieren, die überwiegend durch Meldeverzögerungen verursacht wird.

[56] Dieses Beispiel ist vermutlich rein akademisch, da „faule“ Anmeldungen nicht so oft vorkommen.

germenge gerade der Anzahl aller Werte in der Fuzzy-Klasse entspricht. Falls Einflüsse auf die Eingangsdaten erst auf dem Aggregationsniveau der Bestandszählung quantifiziert werden können, kann dies durch eine Modifikation des aggregierten Fuzzy-Bestandes berücksichtigt werden. Dazu ist eine Modifikationsfunktion zu bestimmen, mit der die obere bzw. die unteren Intervallgrenzen auf den α-Niveaus entsprechend transformiert werden. Da die Fuzzy-Bestandswerte den Charakter von Abschätzungen des ungenau beobachteten Wertes haben, ist es zulässig und nützlich, die Werte schließlich zu glätten, um wohldefinierte Fuzzy-Beobachtungen zu erhalten, die sich als Eingangswerte für die Fuzzy-Regression eignen.

4.2.3 „Gehalt oberhalb der Beitragsbemessungsgrenze“ — Fuzzy-Informationsergänzung

Als Beispiel wird das Merkmal *Einkommen oberhalb der Beitragsbemessungsgrenze* betrachtet.

Durch die jeweils geltende Gesetzgebung über die Sozialversicherungspflicht wird u.a. vorgegeben, wonach das Bruttoentgelt zu bemessen ist, das für die Sozialversicherung gemeldet und erfasst wird.[57] Dabei wird das Entgelt nur bis zur Beitragsbemessungsgrenze exakt erfasst; liegt das Entgelt darüber, wird lediglich der Betrag der Bemessungsgrenze gemeldet. D.h. die gemeldeten Bruttoeinkommen werden oberhalb der geltenden Beitragsbemessungsgrenze G abgeschnitten. Im Datensatz führt das zu einer auffallenden Häufung von Einkommen in Höhe der Beitragsbemessungsgrenze.[58] Die Beitragsbemessungsgrenze kann zudem von Jahr zu Jahr variieren, so dass bei Längsschnittbetrachtungen auch die Effekte der Betragsänderungen zu beachten sind.

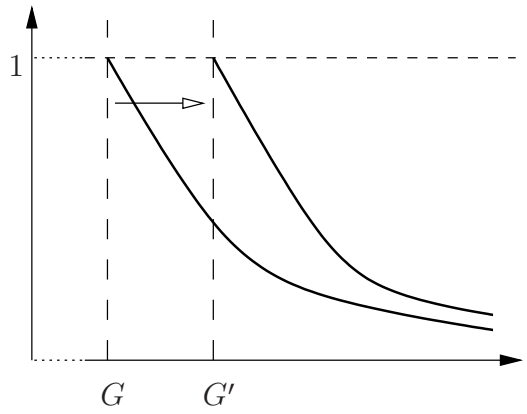
Die Vielfalt der Einkommen, die an der Beitragsbemessungsgrenze abgeschnitten wurden, kann durch einen Fuzzy-Einkommenswert dargestellt werden, der mit Hilfe externer Informationen die jeweils aktuelle Häufigkeitsverteilung von Einkommen oberhalb von G nachzeichnet.[59] Der Wert für die Beitragsbemessungsgrenze wird also durch eine Fuzzy-Wert ersetzt, der die mögliche Lage des tatsächlichen Einkommens abbildet. Dies ist eine einfache Übertragung des üblichen Verfahrens zur Imputation der Einkommen oberhalb von G in eine Fuzzy-Modellierung. In Abbildung 4.4 wird dargestellt, dass dadurch auch Veränderungen von G einfach nachvollzogen werden können.

Die beschriebene Modellierung von Einkommen oberhalb der Beitragsbemessungsgrenze hat im beschriebenen Fall allerdings zur Folge, dass reellwertige Einkommenswerte von

[57] Ab Juni 1999 sind auch geringfügig Beschäftigte enthalten.

[58] Vgl. Schmucker und Seth 2006, S. 49.

[59] Nötigenfalls muss eine konvexe Hülle dieser Verteilung herangezogen werden.



(Eigene Darstellung)

Abbildung 4.4: Erhöhung der Beitragsbemessungsgrenze von G nach G'

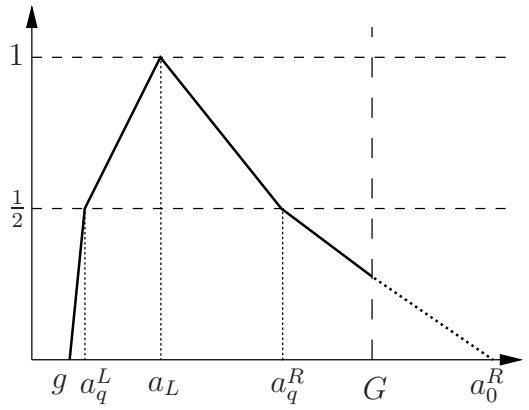
hoher Genauigkeit mit Fuzzy-Werten zu vergleichen sind. Die Instrumente der Datenanalyse sind daher so auszuwählen, dass die reellwertigen Bruttoentgelte unterhalb und die Fuzzy-Entgelte oberhalb von G miteinander ins Verhältnis gesetzt werden können. Im vorliegenden Fall, bei denen von einer hohen Genauigkeit der Einkommen unterhalb der Beitragsbemessungsgrenze ausgegangen werden kann, erscheinen die metrischen Ansätze als geeignet.

Die Fuzzy-Ergänzung der Einkommensverteilung oberhalb von G ist ein Beispiel dafür, wie externe Informationen zugespielt werden können, um die Ungenauigkeit eines Messwertes besser abzuschätzen. Sie stellt somit einen Standardfall für die Fuzzy-Modellierung von Datenungenauigkeiten dar.

4.2.4 Durchschnittliches Bruttoentgelt — Fuzzy repräsentativer Wert

Als Beispiel wird das Merkmal *durchschnittliches Bruttoentgelt* betrachtet.

Wenn die Datenanalyse auf einem höheren Aggregationsniveau durchgeführt werden soll, ist es meistens erforderlich, individuelle Rohdaten zu aggregierten Werten zusammenzufassen. Ein höheres Aggregationsniveau ermöglicht es dabei, die Komplexität des Zusammenhangs zu vereinfachen und die Zahl der verwendeten Variablen zu reduzieren. Außerdem kompensieren sich zufällige Fehler in den Ausgangsdaten bei einer Aggregation oft gegenseitig. Der aggregierte Wert soll relevante Eigenschaften der individuellen Rohdaten charakterisieren. Als Beispiel wollen wir hier die Bestimmung des durchschnittlichen Bruttoentgelts für bestimmte Teilpopulationen betrachten, also etwa das Durchschnittsentgelt für die weiblichen Beschäftigten nach AA-Region, Wirtschaftszweig und Hochschulabschluss. Das Durchschnittsentgelt soll das „typische“ Entgelt der individu-



(Eigene Darstellung)

Abbildung 4.5: Repräsentatives Bruttoentgelt als Quartilsprofil über Median, $\frac{1}{4}$- und $\frac{3}{4}$-Quartil

ellen Rohdaten charakterisieren. Da der Mittelwert reellwertig ist, gehen aber bei der Durchschnittsbildung Informationen über die Streuung der individuellen Punkte verloren. Insbesondere ist nicht mehr ersichtlich, welcher Wertebereich durch die Eingangsdaten aufgespannt wird und damit auch, wie gut die individuellen Beobachtungen im Durchschnittswert repräsentiert werden. Fuzzy-Daten stellen eine Möglichkeit dar, wie die Abweichungen vom Mittelwert für die Anpassung eines funktionalen Zusammenhangs zur Verfügung gestellt werden können.

Das „Fuzzy-repräsentative Bruttoentgelt“ $\overline{\widetilde{L}}_P$ für die Teilpopulation $P \subset \{1, \ldots, n\}$ kann aus den individuellen Merkmalswerten x_i folgendermaßen konstruiert werden, vgl. Abbildung 4.5. Zur Definition der Zugehörigkeitsfunktion definieren wir zunächst die Ecken

$$
\begin{aligned}
&a_L = \text{Median}\{x_i \mid i \in P\}, \\
&a_q^L = \tfrac{1}{4}\text{-Quartil}\{x_i \mid i \in P\}, && a_q^R = \tfrac{3}{4}\text{-Quartil}\{x_i \mid i \in P\}, \\
&a_0^L = \inf\{x_i \mid i \in P\}, && a_0^R = \sup\{x_i \mid i \in P\}.
\end{aligned}
$$

$\overline{\widetilde{L}}_P$ hat nun ihren Kern im Median der individuellen Rohdaten, das Niveau für $\alpha = \frac{1}{2}$ wird durch das $\frac{1}{4}$- und das $\frac{3}{4}$-Quartil aufgespannt und der Träger umfasst den gesamten Wertebereich der Individualdaten, d.h.

$$
\mu_{\overline{\widetilde{L}}_P}(t) = \begin{cases}
\frac{t - a_0^L}{2(a_q^L - a_0^L)} & \text{für } t \in [a_0^L, a_q^L) \\
\frac{1}{2} + \frac{t - a_q^L}{2(a_L - a_q^L)} & \text{für } t \in [a_q^L, a_L) \\
1 & \text{für } t = a_L \\
\frac{1}{2} + \frac{a_q^R - t}{2(a_q^R - a_L)} & \text{für } t \in (a_L, a_q^R] \\
\frac{a_0^R - t}{2(a_0^R - a_q^R)} & \text{für } t \in (a_q^R, a_0^R]
\end{cases}
$$

In die Konstruktion von $\overline{\widetilde{L}}_P$ kann die Fuzzy-Ergänzung für das Bruttoentgelt oberhalb von G problemlos einbezogen werden, da zur Berechnung der Quartilswerte nur die Anzahl der jeweiligen Fälle relevant ist. Sofern das $\frac{3}{4}$-Quartil oberhalb von G liegt, kann — aufgrund der vorgeschlagenen Modellierung auf der Basis der Häufigkeitsverteilung — dasjenige Entgelt herangezogen werden, dessen Flächenanteil dem verbleibenden Personenanteil für das Quartil entspricht, die ein Einkommen oberhalb von G haben.

Alternativ zum Quartilsansatz kann auch eine Konstruktion auf der Basis des Mittelwerts vorgenommen werden. Dabei werden rechte und linke Spannweite jeweils als Mittelwert der positiven bzw. der negativen Abweichungen der individuellen Rohdaten vom Mittelwert bestimmt[60], d.h. die LR-Fuzzy-Zahl $\overline{\widetilde{L}}_P$ sei definiert durch $m_L = \frac{1}{\mathrm{card}(P)} \sum_{i \in P} x_i$ sowie $l_L = \frac{1}{\mathrm{card}(m_L - x_i > 0)} \sum_{i \in P} \max(m_L - x_i, 0)$ und $r_L = \frac{1}{\mathrm{card}(x_i - m_L > 0)} \sum_{i \in P} \max(x_i - m_L, 0)$. Während die erstere Konstruktion die Streuung der Einzelwerte abbildet, stellt die zweitere eher die Ungenauigkeit des Mittelwerts im Verhältnis zu den Einzelwerten dar, die dieser repräsentiert.

Mit der Definition eines Fuzzy-repräsentativen Merkmalswertes (Fuzzy-Mittelwert) ist es möglich, Informationen über die Streuung der aggregierten Rohdaten in die weiteren Berechnungsverfahren einzubringen und zu visualisieren. Als Bestandteile der Fuzzy-Modellierung können gesetzt werden: der relevante Wertebereich als Trägermenge, charakterisierende Punkte als Kern der Fuzzy-Zahl sowie ein Übergangsprofil für den Beitrag der vorliegenden Datenpunkte zum Bereich der charakterisierenden Punkte, die als Zugehörigkeit zum charakterisierenden Bereich interpretiert werden. Da Fuzzy-Daten konvex sind, gilt dabei die Einschränkung, dass als Kern nur ein Intervall oder ein Punkt zulässig sind, mehrere lokale Häufungspunkte bei den individuellen Rohdaten können daher nur stark vereinfacht und reduziert abgebildet werden.

4.2.5 „Strukturbruch Bruttoentgelt“ — Fuzzy-Angleichung

Als Beispiel wird das Merkmal *sozialversicherungspflichtiges Bruttoentgelt* betrachtet, für das aufgrund einer Änderung der amtlichen Definition ein Strukturbruch zu berücksichtigen ist.

Lohnsonderzahlungen sind seit 1984 in das Bruttoentgelt einzubeziehen. Bis dahin war dies freigestellt und wurde von den Betrieben uneinheitlich gehandhabt. Die geänderte Regelung wirkte sich in der Tendenz wie eine Einkommenssteigerung aus und zwar vorzugsweise in den höheren Einkommen. So führte die Neuregelung bei den Beschäftigten, die ein Arbeitsentgelt oberhalb der Beitragsbemessungsgrenze bezogen, zu einem

[60] Vgl. Coppi u. a. 2006, S. 281f.

sprunghaften Anstieg um ca. 250.000 Personen.[61] Außerdem sind vermutlich in größerem Umfang Einkommen unterhalb der Einkommensgrenze in die Sozialversicherungspflicht „hineingewachsen“, auch ohne dass es zu einer realen Einkommenssteigerung gekommen ist. Im Vergleich mit den Angaben im Sozio-oekonomischen Panel (SOEP) konnte nachgewiesen werden, dass die Änderung der Definition für das sozialversicherungspflichtige Bruttoentgelt tatsächlich das Ausmaß eines Strukturbruchs annimmt.[62] Sofern die geltende Definition als korrektes Messverfahren gewertet wird, können die abweichenden Messungen vor dem Strukturbruch als Messfehler aufgefasst werden.

Ein Ausgleich für den Zuwachs von Personen, die vor der Definitionsänderung nicht sozialversicherungspflichtig waren, ist nicht ohne Weiteres möglich.[63] Bei Beschäftigten, die vor der Änderung bereits sozialversicherungspflichtig waren, äußert sich die Änderung in einer Änderung des Einkommensniveaus, so dass eine Fehlerabschätzung für bereits vorliegende Beobachtungen vorzunehmen ist.

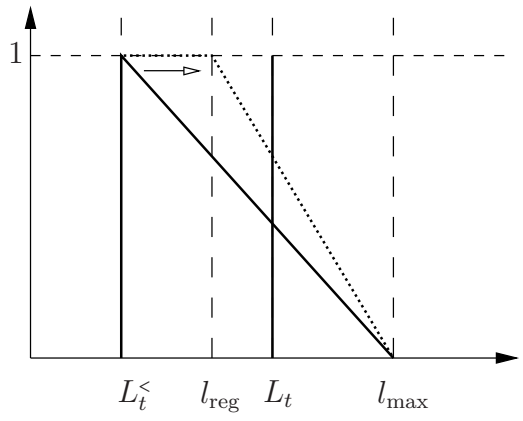
Die Abschätzung der Abweichung ist nur dann möglich, wenn eine gültige Definition der statistischen Einheit angegeben werden kann, die es ermöglicht die „fehlerhaften“ Abweichungen zu quantifizieren. Bei dem vorliegenden Strukturbruch soll die Berechnung des Bruttoentgelts in der Definition ab 1984 die maßgebende Definition darstellen, die wir mit L bezeichnen wollen. Die Wahl der „geltenden“ Definition wird im allgemeinen vom Zweck der Auswertung abhängen. Wir wählen sie hier u.a. deshalb, weil bei der Neudefinition des sozialversicherungspflichtigen Bruttoentgelts ab 1984 zusätzliche Entgeltbestandteile einbezogen wurden, mit denen eine Vereinheitlichung in der praktischen Handhabung erreicht wurde und infolge derer das Bruttoentgelt nicht kleiner als das Bruttoentgelt in der Definition vor 1984 (kurz: $L^<$) ist. Im allgemeinen gilt für einen Stichtag t also $L_t^< \leq L_t$. Bei den vor 1984 gemeldeten Bruttoentgelten ist folglich die mögliche, ungewisse Ausweitung des Umfangs bei Verwendung der Maßstabsdefinition L darzustellen: „möglicher Wert von L_t bei vorliegendem Messwert für $L_t^<$“. Für Zeitpunkte t vor 1984 kann daher das Bruttoentgelt in der damaligen Definition wie in Abbildung 4.6 in die Definition L überführt werden. Dabei liegt es nahe, den vorliegenden Messwert mit der Zugehörigkeit 1 zu bewerten und eine monoton fallende Zugehörigkeit für den Bereich der möglichen Werte nach der Definition L anzunehmen.

Eine Schwierigkeit bei der Konstruktion des Fuzzy-Ausgleichs für den Strukturbruch beim Bruttoentgelt besteht darin, dass die Messwerte $L_t^<$ in fast allen Fällen gegenüber

[61] Vgl. Bender u. a. 1996, S. 15.

[62] Vgl. Steiner und Wagner 1997, S. 640.

[63] Dies stellt einen der Selektionseffekte der Beschäftigtenstatistik dar. Aufgrund des bis 1999 verwendeten Meldeverfahrens liegen im Rahmen der Beschäftigtenstatistik keine Einkommensinformationen für geringfügig Beschäftigte vor. Der Übergang von der Nicht-Erfassung zur Erfassung kann daher für diesen Zeitraum nur durch Informationen aus alternativen Datenquellen ergänzt werden.



(Eigene Darstellung)

Abbildung 4.6: Fuzzy-Angleichung für den Strukturbruch im Bruttoentgelt — Gegebene Messwerte $L_t^<$ und L_t, sowie zwei Varianten zur Modellierung der Fuzzy-Anpassung $\widetilde{L}_t$ von $L_t^<$ an die Definition L. Dabei ist $l_{\max}$ der maximale Wert der möglichen Bruttoentgelte gemäß L. l_{reg} ist ein Regelungswert zur Korrektur des 1-Niveaus.

dem Maßstab L zu niedrig ausfallen. Das 1-Niveau des Fuzzy-Ausgleichs hat somit immer einen gewissen Abstand zum verborgenen „wahren“ Wert. Andererseits führt eine Transformation, die dem Messwert $L_t^<$ eine Zugehörigkeit von $\mu_{\widetilde{L}_t}(L_t^<) < 1$ zuweist, dazu, dass die Information über den ursprünglichen Messwert abgewertet wird. Eine Abwertung der Zugehörigkeit von $L_t^<$ sollte aber möglichst nur dann vorgenommen werden, wenn der Stand der Informationen einen anderen Wert als Wert mit der höchsten Möglichkeit nahelegen. Ansonsten ist eher eine vorsichtige Alternative zu empfehlen, bei der das 1-Niveau von $\widetilde{L}_t$ bis zu einem „typischen“ Ausgleichswert $l_{\text{reg}}(t)$ ausgeweitet wird. Dann erscheinen alle Werte im Intervall $[L_t^<, l_{\text{reg}}(t)]$ als ununterscheidbar sicher möglich.

Für ein rein quantitatives Verfahren zur Homogenisierung der Effekte durch die Einführung eines neuen Messverfahrens kann folgendermaßen vorgegangen werden: Seien $a_i^< \in \mathbb{R}$ die Messwerte nach dem alten Verfahren und $a_i \in \mathbb{R}$ die Messwerte nach dem neuen Verfahren. Sei ferner $\widetilde{\Delta}$ ein Fuzzy-repräsentativer Wert für die Menge der Differenzen $\{a_i - a_i^< \mid 1 \le i \le n\}$ (zur Konstruktion vergleiche Abschnitt *4.2.4*). Dann kann die Fuzzy-Angleichung $\widetilde{A}_i$ von $a_i^<$ berechnet werden als $\widetilde{A}_i = a_i^< \oplus \widetilde{\Delta}$. Dies ist gleichbedeutend damit, dass $\mu_{\widetilde{A}_i}(a_i^<) = \mu_{\widetilde{\Delta}}(0)$ und somit im allgemeinen kleiner als 1. Dieses Verfahren kann aber nur unter der Bedingung gerechtfertigt werden, dass der Veränderungssaldo überwiegend nur von der Definitionsänderung beeinflusst und nicht von anderen Rahmenbedingungen überlagert wird. Andernfalls ist eine Modellierung vorzuziehen, die inhaltlich motiviert werden kann.

Die hier beschriebenen Ansätze können nicht nur verwendet werden, um Änderungen im Messverfahren zu antizipieren und zu homogenisieren, ähnlich können auch Merk-

malsvariablen aus mehreren Datensätzen angeglichen werden, die bei der Datenanalyse miteinander verknüpft werden sollen. Besondere Beachtung verlangt dabei die Frage, was überhaupt über die möglichen „wahren“ Werte bekannt ist und was dies zur quantitativen Abschätzung beiträgt. Beispielsweise kommt es häufig vor, dass der vorliegende Messwert nur eine untere oder obere Grenze für die möglichen Werte darstellt und der vorliegende Messwert im Extremfall ggf. nie zutrifft. So kann die Zahl geringfügig Beschäftigter mit den Angaben aus der Beschäftigtenstatistik bzw. aus dem Mikrozensus nach unten abgeschätzt werden, liegt aber tendenziell um einen erheblichen Anteil darüber.[64] In einer solchen Situation kann die Modellierung eines Fuzzy-Ausgleichswertes eine, z.T. sogar erhebliche, verzerrte Abschätzung für den verborgenen „wahren“ Wert darstellen. Dieses Problem soll im Folgenden als *Randbeobachtungs-Problem* bezeichnet werden. Das Randbeobachtungs-Problem ist Ausdruck eines Zielkonfliktes zwischen der Datenwahrheit und -klarheit aufgrund der vorliegenden Messwerte einerseits und der Güte des modellierten Fuzzy-Wertes anhand des vorliegenden Expert/innenwissens andererseits. Hierbei hat eine Fuzzy-Modellierung der fehlerhaften Werte dennoch den Vorzug, dass der mögliche Wertebereich aufgespannt wird und nicht — wie häufig üblich — vernachlässigt wird.

4.2.6 Genauigkeitsschwelle und Relevanzschranke

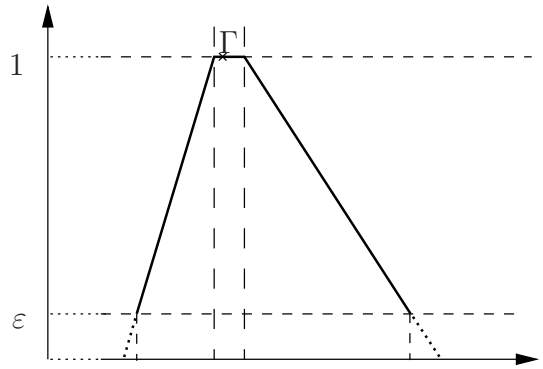
Aufgrund der Einschränkungen im Messverfahren können alle Punkte in einer Umgebung um den reellwertigen Beobachtungswert nicht vom eigentlichen Messwert unterschieden werden. Wie Bemerkung 2.1 verdeutlicht, ist die Genauigkeitsschwelle in erster Linie als bewusst gewählte Komplexitätsreduktion bzw. als Granulation in die Mess-Skala eingeschrieben, auf der auch die Abschätzung der Ungewissheit der Messung abgetragen wird. Sie tritt daher zunächst nur implizit in Erscheinung.

Grundsätzlich repräsentiert jedes Fuzzy-Datum schon an und für sich die (Un-)Genauigkeit des Messwertes, weil alle Punkte auf einem α-Niveau mindestens zum Grad α nicht vom Merkmalswert unterscheidbar sind.[65] Da die Zugehörigkeitsgrade nur ordinal skaliert sind, wird zudem deutlich, dass die Genauigkeitsschwelle selber ein vager Wert ist.

Da Fehlereinflüsse in den Daten vor allem in dem Maße relevant sind, wie sie die Genauigkeitsschwelle überschreiten, kann es sinnvoll sein, bei der Konstruktion von Fuzzy-Daten die Genauigkeitsschwelle explizit zu modellieren. Damit wird zwischen Messfehlern

[64] Dies hat sich mit Änderung im Meldesystem seit 1999 vermutlich wesentlich geändert, da nun auch eine Meldepflicht für geringfügig Beschäftigte besteht.

[65] Grundsätzlich können die Fuzzy-Genauigkeitsumgebungen durch eine Ähnlichkeitsrelation formalisiert werden. Diese bildet eine Fuzzy-Partition der Mess-Skala, die auf 1 normiert ist. Allerdings ermöglicht diese Darstellung keine Unterscheidung zwischen Genauigkeitsschwelle und Fehlereinfluss. Vgl. dazu Kruse u. a. 1993, S. 223ff.



(Eigene Darstellung)

Abbildung 4.7: Umgebung Γ für die Genauigkeitsschwelle des Messverfahrens und Relevanzschranke ε für die Fehlermodellierung

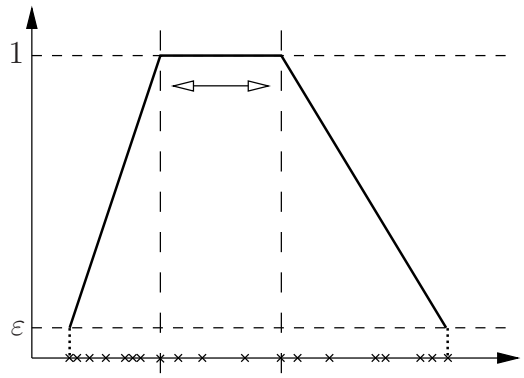
und Genauigkeitsschwelle unterschieden und deren Verhältnis hervorgehoben. Zu diesem Zweck kann die Genauigkeitsschwelle auf ein vorgegebenes γ-Niveau normiert werden. Unter der Annahme, dass der beobachtete Wert die sicherste Information über den eigentlichen Wert ist, können Punkte, die in der Nähe des Messwertes liegen, nicht bzw. so gut wie nicht vom eigentlichen Wert unterschieden werden. γ sollte somit nahe an 1 sein. Als Beispiel wurde in Abbildung 4.7 das 1-Niveau mit der Umgebung Γ für die Genauigkeitsschwelle identifiziert.

Dabei kann Γ aufgrund des Messwertes aus einer vorgegebenen, technisch motivierten Partition von Genauigkeitsumgebungen ausgewählt werden. Oder Γ kann als Kugel mit einem festen Radius und dem Messwert als Mittelpunkt, d.h. als Streuungsumgebung um den Messwert, modelliert werden.

Zusätzlich kann vorkommen, dass eine sichere Grenze für die möglichen Werte eines Merkmalswertes sehr weit vom Kern des Fuzzy-Datums entfernt ist. In einem solchen Fall ist es häufig sinnvoll, die Darstellung auf den Bereich relevanter Fehlerabweichungen zu beschränken und den Fuzzy-Wert nach unten abzuschneiden. Abbildung 4.7 zeigt einen Fuzzy-Merkmalswert mit Relevanzschranke ε.

Die Fuzzy-Modellierung kann daher dazu genutzt werden, einen *repräsentativen Wert bei ungünstiger Datenlage* zu bestimmen. Damit wird es z.B. möglich, schwach besetzte Teilpopulationen, die häufig ausgeschlossen werden, dennoch in die quantitative Analyse mit einzubeziehen.

Je nach der subjektiven Einschätzung der vorliegenden Messwerte können dazu manuell ein oder mehrere prototypische „mittlere“ Werte ausgewählt werden. In Abhängigkeit davon, wie gut die Repräsentation durch die ausgewählten Werte eingeschätzt wird, kann zusätzlich eine vorgegebene Anzahl von umgebenden Werten in das 1-Niveau einbezogen werden. Der Wertebereich an vorliegenden Messwerten spannt den Träger des Fuzzy-



(Eigene Darstellung)

Abbildung 4.8: Repräsentativer Wert bei ungünstiger Datenlage — 19 Messwerte liegen vor, davon werden die zentralen 5 zur Bestimmung des Intervalls „mittlerer“ Werte verwendet.

Datums auf. Falls — etwa aufgrund der geringen Anzahl von Werten — davon auszugehen ist, dass der Wertebereich dadurch nicht hinreichend genug beschrieben wird, kann dies durch die Angabe einer Relevanzschranke ε antizipiert werden (vgl. Abbildung 4.8).

Dadurch, dass in schwach besetzten Teilpopulationen nicht genug Information zur Identifizierung charakteristischer Merkmalsausprägungen vorliegen, hat die in den Aggregaten repräsentierte Information ein unterschiedliches Gewicht. Es ist daher ganz besonders bei der Verwendung von solchen Fuzzy-Intervall-Daten darauf zu achten, dass die Fuzzy-Genauigkeitsumgebungen durch die Analysemethoden nur soweit gewichtet werden, wie es der Aussagekraft der Datenmodellierung entspricht.

4.3 Konstruktionsprinzipien und allgemeine Eigenschaften der Fuzzy-Merkmalswerte

Im vorhergehenden Abschnitt wurden einige Vorschläge zur Modellierung von Fuzzy-Merkmalswerten vorgestellt, bei denen die Fuzzy-Modellierung dazu beiträgt, dass Ungenauigkeiten in den Daten besser abgebildet werden können. Die gemeinsamen Eigenschaften der Fuzzy-Daten und einige Besonderheiten im Vergleich zu reellwertigen Daten werden als Vorbereitung auf das Benchmarking der Fuzzy-Regression im nachfolgenden Kapitel *5* zusammengefasst und diskutiert.

Der hauptsächliche Nutzen bei der Verwendung von Fuzzy-Merkmalswerten liegt darin, dass Messfehler in den Beobachtungsdaten dokumentiert und visualisiert werden. Und zwar konkret mit den quantitativen Auswirkungen auf die jeweiligen Messwerte. Für die Fuzzy-Modellierung ist häufig der Preis zu zahlen, dass nur noch schwächere Aussagen

für die Datenanalyse hergeleitet werden können. Dem steht der Vorteil gegenüber, dass mit Fuzzy-Daten der Einfluss von Messfehlern auf die Analyseergebnisse systematisch abgebildet und somit objektiviert werden kann.

Es fällt auf, dass alle vorgestellten Merkmalsmodellierungen asymmetrisch sind. Sie unterscheiden sich damit von den üblichen ε-Fehlerumgebungen $[x-\varepsilon, x+\varepsilon]$. Der Unterschied kann einerseits dadurch erklärt werden, dass die vorgestellten Fuzzy-Daten nicht vom Ansatz einer Häufigkeitsverteilung ausgehen,[66] sondern eine nach dem Grad der Möglichkeit gewichtete Fehlerabschätzung sind. Ein wesentlicher Grund ist aber darin zu sehen, dass in fast allen Beispielen — mit Ausnahme der repräsentativen Fuzzy-Werte — systematische Fehlereinflüsse abgebildet werden. Für die weiteren Untersuchungen ist also davon auszugehen, dass epistemische Fuzzy-Daten typischerweise asymmetrisch sind. Je nach der Qualität der Beobachtungen können überdies auch intervallförmige Fuzzy-Daten vorkommen.

Die empirischen Ansätze zur Konstruktion von Fuzzy-Daten für ungenaue Beobachtungen, wie sie in diesem Kapitel vorgestellt wurden, basieren auf wenigen Prinzipien. Die Modellierungsbeispiele zeigen, dass verschiedene Eigenschaften der Fuzzy-Daten als Informationen über den verborgenen „wahren“ Wert relevant sind. Diese Eigenschaften können als Dimensionen der Güte der Fehlerabschätzung in epistemischen Fuzzy-Daten aufgefasst werden:

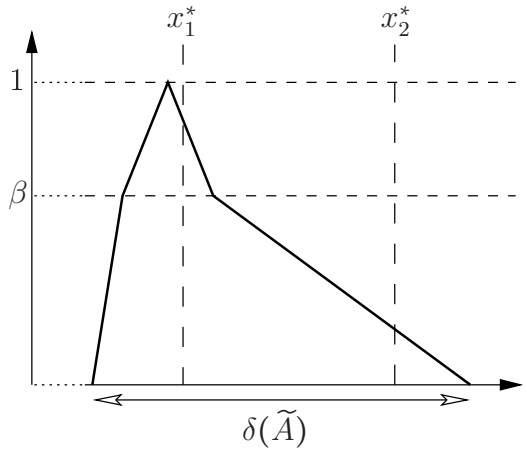
Ein wesentlicher Aspekt ist die *Schärfe der Abschätzung*, die sich bei einem reellwertigen „wahren“ Wert in der Ausdehnung der Trägermenge bzw. in der Intensität des Kontrastübergangs zwischen voller Zugehörigkeit und Nicht-Zugehörigkeit ausdrückt. Die Schärfe der Abschätzung ist bei günstigen Fuzzy-Daten identisch mit deren Fuzziness, sofern der „wahre“ Wert eindeutig und reellwertig ist.[67]

Zudem ist die *Nähe zwischen dem Kern des Fuzzy-Datums und dem „wahren“ Wert* ausschlaggebend für die Güte der Fehlermodellierung. Diese kann entweder als Abstand zwischen „wahrem“ Wert und Modalwert, d.h. $|z_i - z_i^*|$, gemessen werden oder alternativ als Zugehörigkeitswert des „wahren“ Wertes zum Fuzzy-Datum, d.h. $\mu_{\widetilde{Z}_i}(z_i^*)$.

Schließlich ist auch die *Gewichtung durch den Rand des Fuzzy-Datums* von Bedeutung für die Qualität der Fehlerdarstellung. Diese ist dann um so wichtiger, wenn die Lage des „wahren“ Wertes nur sehr ungenau und asymmetrisch beschrieben werden kann. Ein

[66] Mit dem zentralen Grenzwertsatz kann im klassischen Verteilungsmodell davon ausgegangen werden, dass bei einer Vermischung mehrerer Fehlereffekte eine Normalverteilung zugrundeliegt, die symmetrisch ist.

[67] Bei einem Fuzzy-Datum, das einen repräsentativen Wert für mehrere Subjekte wiedergibt, wird die Abbildung der Ungenauigkeit durch mehrere Werte gestützt, so dass die weiche Fehlerabschätzung inhaltlich schärfer ist als bei einem individuellen Messwert.



(Eigene Darstellung)

Abbildung 4.9: Dimensionen für die Güte der Fehlerabschätzung bei Fuzzy-Daten — Maß der Fuzziness $\delta(\widetilde{A})$ und Zugehörigkeit des „wahren“ Wertes x_1^* bzw. Inklusion des „wahren“ Wertes x_2^* durch den Rand.

wichtiges Beispiel dafür ist das Randbeobachtungs-Problem[68]. Also z.B. dann, wenn das Messverfahren einen Wert liefert, von dem bekannt ist, dass dieser in jedem Fall zu niedrig ausfällt, so dass der „wahre“ Wert nur mit einer Ausweitung des rechten Randes sicher in die Trägermenge eingeschlossen werden kann. In einem solchen Fall ist es wichtig, dass die Gewichtung durch die Ränder bei der Datenanalyse angemessen gewertet wird. Die beschriebenen Dimensionen für die Güte der Fehlerabschätzung sind in Abbildung 4.9 zusammengefasst.

Bei einer Modellierung von Fuzzy-Daten unter Hinzuziehung von Expert/innenwissen bestehen kaum Einschränkungen dafür, welche Datenungenauigkeiten abgebildet werden können, sofern irgendwelche quantifizierbare Informationen darüber vorliegen. Ein Einschränkung — insbesondere bei der Datenmodellierung für makroökonomische Analysen oder von synthetischen Begriffen — besteht allerdings in der erforderlichen Eingipfligkeit und Konvexität der Fuzzy-Daten. Hier ist zu prüfen, ob die kleinste konvexe Hülle eines Fehlerabschätzungsgebirges eine sinnvolle Abschätzung darstellt. Alternativ kann das Fehlerabschätzungsgebirge ggf. in Teilbeobachtungen aufgesplittet werden.

Ein Kriterium, das für die Datenanalyse mit Fuzzy-Daten nur mit Schwierigkeiten beurteilt werden kann, ist die Homogenität der Daten. Bei vagen statistischen Begriffen, d.h. bei linguistischen Variablen, kann ohne Zweifel die Homogenität der Daten bei einer Fuzzy-Modellierung erhöht werden, weil dadurch ein breiteres Spektrum der Begriffsfacetten in die Messwerte einfließt und somit im Vergleich mit einfachen reellwertigen Messwerten der Einfluss von Abgrenzungsproblemen reduziert ist. Was kann aber über die Homogenität von epistemischen Fuzzy-Daten ausgesagt werden, die eine subjektive Ab-

[68] Vgl. Abschnitt *4.2.5*, S. 146.

schätzung der Datenfehler unter Einbeziehung von Expert/innenwissen darstellen? Wann ist die Mess-Situation und der Aussagegehalt der vorliegenden Fuzzy-Daten als vergleichbar anzusehen?

Zunächst ist argumentativ zwischen der Ebene der Beobachtung der fehlerbehafteten Daten und deren Vergleichbarkeit sowie andererseits der Ebene der subjektiven Fehlerabschätzung zu unterscheiden. Inhomogenitäten bei der Datenbeobachtung können durch die Fehlerabschätzung ausgeglichen werden. So können singuläre Effekte auf die Datengenauigkeit, wie z.B. regionale Meldeprobleme oder die vorübergehend verringerte Datenqualität aufgrund der Einführung eines neuen Erhebungsverfahrens, sichtbar und damit überhaupt erst vergleichbar gemacht werden. Die Vergleichbarkeit bei der Abschätzung der Datenungenauigkeiten selber wird im Einzelfall davon abhängen, welche Informationen über die einzelnen Mess-Situationen aktuell verfügbar sind und wie diese bei der Konstruktion der Daten abgebildet werden. Grundsätzlich sind daher standardisierte Verfahren zur Modellierung von epistemischen Fuzzy-Daten zu bevorzugen. Bei den Methoden zur Fuzzy-Datenanalyse sollte darauf geachtet werden, dass sie einen geeigneten Kompromiss zwischen Sensitivität und Robustheit bei Änderungen in den Referenzfunktionen bzw. den Rändern der Fuzzy-Daten einhalten.

Grundsätzlich sollen die Beobachtungswerte für Modellvariablen, die in der Modellgleichung miteinander ins Verhältnis gesetzt werden, in ihrer Genauigkeit und Homogenität vergleichbar sein. Diese Anforderung gilt für die Fuzzy-Modellierung von Fehlern in den Daten umso mehr, da Fehlereinflüsse sich bei einer Aggregation gegenseitig kompensieren können. Falls Fuzzy-Daten nachträglich aggregiert werden, kann überschüssige Fuzziness erzeugt werden, weil die Fuzziness, die durch erweiterte Fuzzy-Funktionen erzeugt wird, irreversibel ist. Es ist daher zu fordern, dass die Datenungenauigkeiten nach Möglichkeit genau auf der Aggregationsebene abgebildet werden, die für das Zusammenhangsmodell benötigt werden.

5 Regression mit Fuzzy-Fehlern in den Daten

In diesem Kapitel soll die Fuzzy-Regression als eine Alternative zur klassischen Regression bei fehlerhaften Daten untersucht werden. Dazu nehmen wir anknüpfend an die Kritik in Abschnitt *2.3* an, dass die Merkmalswerte mit Fehlern behaftet sind, die im Rahmen eines klassischen Fehlermodells nicht angemessen abgebildet werden können, weil die Fehlereinflüsse zu stark sind, nicht spezifisch genug beschrieben werden können oder singulären Sprüngen unterliegen. Insbesondere gebe es keine Kenntnisse über Abhängigkeiten mit den Fehlereinflüssen in den Beobachtungen. Es wurde vorgeschlagen, solche Fehler in Form von Fuzzy-Merkmalswerten zu modellieren und somit für die Datenanalyse nutzbar zu machen. Anforderungen an die Modellierung und deren Eigenschaften wurden in Kapitel *4* bereits anhand konkreter Beispiele ausgearbeitet und untersucht. Messfehler, die den genannten Anforderungen entsprechen und die gut in Form von Fuzzy-Beobachtungen abgebildet werden können, werden im Folgenden auch kurz als *Fuzzy-Fehler* bezeichnet.

Wenn die fehlerbehafteten Daten in Form von Fuzzy-Daten vorliegen, können Fuzzy-Regressionsmodelle konstruiert werden, die den Modellen zur klassischen Regression bei Fehlern in den Daten gegenüber zu stellen sind. Für den Vergleich in diesem Kapitel werden die Fuzzy-Modelle nicht nach den Modellfunktionen, sondern nach drei Typen von Fehlereinflüssen unterschieden: Ansätze mit Fuzzy-Daten von Typ $(\widetilde{X}_i, y_i)$ stellen das direkte Fuzzy-Analogon zum klassischen Fehlermodell dar. Fuzzy-Fehler können, anders als im klassischen Schätzmodell, auch in der abhängigen Variablen Y auftreten. Sie sind in diesem Fall von der Variation der Beobachtungen vom funktionalen Zusammenhang zu trennen, und wir erhalten ein gesondertes Modell mit Fuzzy-Daten vom Typ $(x_i, \widetilde{Y}_i)$. Schließlich sind Fälle zu betrachten, in denen sowohl die Regressoren als auch die Regressand Fuzzy-Fehler enthalten, so dass wir Fuzzy-Daten vom Typ $(\widetilde{X}_i, \widetilde{Y}_i)$ haben. In diesem letzteren Fall werden die beiden vorhergehenden Fälle zusammengeführt.

Es soll im Folgenden immer vorausgesetzt sein, dass die Fuzzy-Daten günstig im Sinne von Definition 2.22 sind, so dass die „wahren“ Werte immer im Support der Fuzzy-Daten enthalten sind.

Die Eigenschaften der Fuzzy-Regression für die Datenanalyse werden im Vergleich mit einer „fehlerblinden“ klassischen Regression herausgearbeitet. Allerdings stößt ein analytischer Vergleich selbst im univariaten Modell schnell an seine Grenzen, da die Lösungsparameter im Fuzzy-Fall von Einflussgrößen, wie z.B. der Streuung der Spannweiten der Fuzzy-Daten[1], abhängen, die auch aus methodischen Gründen mit den Verzerrungen im klassischen Fehlermodell nicht vergleichbar sind. Deshalb wird anhand von Fallbeispielen mit simulierten Fuzzy-Daten ein Benchmarking zwischen den Methoden durchgeführt. Um die Zahl der Einflussgrößen nicht zu stark auszuweiten, werden die Untersuchungen zunächst in einem univariaten Modell durchgeführt. Sowohl im induktiven Modell mit Verteilungsannahmen als auch bei der explorativen Datenanalyse ist es zentral, dass das Ergebnis der Modellapproximation den Zusammenhang in den Daten möglichst gut wiedergibt. Wir werden uns daher bei der Untersuchung auf die Unterschiede bei den Approximationseigenschaften der ausgewählten Regressionstechniken konzentrieren und beschränken.

Die hauptsächliche Schwierigkeit für das Benchmarking besteht darin, geeignete Datensätze von Fuzzy-Daten mit Fuzzy-Fehlern zu finden. Die Testdaten sollen es ermöglichen, für einzelne Aspekte der Datenmodellierung zu kontrollieren, und zudem sollen die „wahren“ Werte als Vergleichsmaßstab zur Verfügung stehen. Deshalb wird hier der Versuch unternommen, Algorithmen zur maschinellen Erzeugung von Fuzzy-Datensätzen zu entwickeln, so dass deren Eigenschaften mit vorgegebenen Parametern gesteuert werden können. Die automatische Generierung von Fuzzy-Daten schafft zudem die Möglichkeit, eine beliebige Anzahl von vergleichbaren Datensätzen zu erzeugen, wie sie für die Durchführung von Simulationsstudien notwendig sind. Gegenüber der negierenden Abgrenzung von Fuzzy-Fehlern gegenüber Daten mit zufälligen Fehlern als „nicht-zufällig“, zwingt die Entwicklung von Verfahren zur Datengenerierung dazu, die Eigenschaften von Daten mit Fuzzy-Fehlern konstruktiv anzugeben. Die Entwicklung der Algorithmen hat somit als Voraussetzung, dass Szenarien von relevanten und typischen Fuzzy-Fehlern entwickelt werden. Dies wird als Chance gesehen, die Abgrenzung von zufälligen und Fuzzy-Fehlern für die empirische Praxis zu präzisieren und zu konkretisieren.

In Abschnitt *5.1* wird das Verfahren des Benchmarking motiviert und beschrieben, darunter speziell in Abschnitt *5.1.1* die verwendeten Verfahren zur Generierung von prototypischen Fuzzy-Fehler-Daten. Anhand von ausgewählten Fallbeispielen werden in Abschnitt *5.2* die Eigenschaften der Fuzzy-Regression bei Fuzzy-Fehlern illustriert. Deren Bedeutung für die Interpretation werden abschließend in Abschnitt *5.3* ausgewertet.

[1] Beispielsweise hängen die Lösungsparameter bei triangulären Fuzzy-Daten von der Streuung der linken Endpunkte der Träger und der Korrelation zwischen Kernen und linken Endpunkten ab. Vgl. dazu den Rechenalgorithmus (5.12) auf S. 174.

5.1 Benchmarking der Fuzzy-Regression

Fuzzy-Daten stellen gegenüber den „fehlerblinden“ reellwertigen Beobachtungen eine Zunahme an Information dar, die dann von Nutzen ist, wenn das Approximationsergebnis dadurch verbessert werden kann. Eine Verbesserung muss nicht zu einem vollständigen Ausgleich der Verzerrungen führen, sondern kann z.B. auch darin bestehen, dass die Approximationsparameter korrigiert werden oder dass der Wissensstand sich besser als bisher in der Approximationsfunktion widerspiegelt. Eine wesentliche Fragestellung des Benchmarkings bezieht sich folglich auf den Einfluss der Fehlermodellierung in den Fuzzy-Daten auf die Approximation. Darüber hinaus ist zu fragen, welcher Art die Sensitivität der Fuzzy-Approximationsfunktion im Bezug auf Änderungen in den Fehlerbewertungen ist.

In Abschnitt *3.3* wurde hervorgehoben, dass auch bei der Fuzzy-Regression die Kleinste Quadrate-Ansätze prominent sind, die für reellwertige Daten mit der klassischen Kleinste Quadrate-Regression zusammenfallen und die auf einer Metrik auf Fuzzy-Mengen basieren. Insbesondere die δ_2-Metrik ermöglicht es, wesentliche Ergebnisse der statistischen Kleinste Quadrate-Regression auch auf den Fuzzy-Fall zu übertragen und somit zu einer Fuzzy-Schätzung zu gelangen wie in Abschnitt *3.3.3*. Unter anderen wegen dieser Vergleichbarkeit mit der klassischen Regression wurde die δ_2-Metrik sowie die Klasse der davon abgeleiteten Metriken gemäß Bertoluzza u. a. [1995] für das Benchmarking herangezogen. Außerdem stehen für diese Klasse von Metriken Rechenalgorithmen zur Verfügung[2], so dass sie für die eigenen Zwecke leicht nachprogrammiert[3] werden konnten.

Die Kleinste Quadrate-Approximationsfunktion beschreibt den funktionalen Zusammenhang zwischen den Bewertungs- bzw. den Ungenauigkeitsumgebungen, die durch die Fuzzy-Daten beschrieben werden. Aussagen über den Zusammenhang zwischen den „wahren“ Werten können demnach nicht wie gewohnt direkt aus den Approximationsparametern abgelesen, sondern immer nur mittelbar aus der Abbildung der zugehörigen Fehlerumgebungen abgeleitet werden. Die Approximationsparameter lassen sich folglich nur dann in einfacher Weise als Hinweis auf eine lineare Proportionalität der Merkmalsvariablen interpretieren, wenn die Fuzzy-Daten eine scharfe Abschätzung der abgebildeten statistischen Begriffe wiedergeben. Das ist vor allem dann der Fall, wenn die Fuzzy-Daten der physikalische Interpretation entsprechen oder wenn die Fuzziness der Daten klein ist. Bei Fuzzy-Daten mit denen Fuzzy-Fehler abgebildet werden, ist hingegen überwiegend davon auszugehen, dass sie unscharfe Abschätzungen des „wahren“ Messwertes sind, die überdies noch semantische Inhomogenitäten sowie zusätzliche subjektive Einflüsse in der

[2] Vgl. [Diamond und Körner, 1997; Körner und Näther, 1998].

[3] Die Programme für die Datensimulation und die Berechnung der Fuzzy-Approximation wurden in GNU Octave geschrieben, der vergleichbaren Open Source-Entwicklung zu MATLAB.

Modellierung[4] aufweisen können. Eine wesentliche Frage ist also, in welcher Weise die verschiedenen Aspekte der Fuzzy-Modellierung die Approximation beeinflussen und ob diese Wirkungen mit der Fehlerbewertung in den Daten konform sind. Nur dann kann davon ausgegangen werden, dass Fehlereinflüsse durch eine entsprechende Fuzzy-Modellierung in der Tendenz korrigiert werden können. Interessant ist außerdem, wie stark die Korrektur sich auswirkt: Was kann in welchem Maße korrigiert werden und wann lohnt sich der Aufwand einer Fuzzy-Modellierung, weil der klassische Stabilitätsoptimismus ausgehebelt wird?

5.1.1 Simulation von Daten mit Fuzzy-Fehlern

Die Simulationsverfahren sollen Fuzzy-Werte liefern, die als typische Abbildung einer ungenauen Messung von reellwertigen, „wahren“ Daten gesehen werden können, die einen linearen Zusammenhang beschreiben. Dabei soll die Ungenauigkeit sich infolge eines typischen Szenarios von Fuzzy-Fehlern ergeben. Also einer Mess-Situation, in der die Fehler nicht in Form von Zufallsvariablen im Modell darstellbar sind. Mit der Konstruktion der Fehlerszenarien sollen typischerweise vorkommende Abweichungen des Messwertes vom „wahren“ Wert in empirischen Datensätzen charakterisiert und nachempfunden werden. Gewissermaßen werden damit verallgemeinerbare Anforderungsprofile generiert, an denen die Sensitivität der Regressionsmethoden zu erproben ist. Anhand der in Kapitel *4* konstruierten Fuzzy-Merkmalswerte wurde bereits herausgearbeitet, dass die Qualität der Fehlerdarstellung entlang von drei Eigenschaften eingeordnet werden kann, die sich in Teilen auch widersprechen: der Schärfe der Abschätzung, der Nähe zwischen dem Kern des Fuzzy-Datums und dem „wahren“ Wert sowie — insbesondere in Situationen, in denen ein Randbeobachtungs-Problem vorliegt — der Gewichtung durch den Rand des Fuzzy-Datums.[5]

Entsprechend dem Fazit in Abschnitt *4.3* sollen die Fuzzy-Daten überwiegend asymmetrisch sein. Außerdem sollen sie im Sinne von Definition 2.22 günstig sein, wobei wir für den Vergleich mit klassischen Methoden unterstellen, dass der Modalwert eines Fuzzy-Datums dem fehlerhaften, reellwertigen Messwert entspricht. Die zweitere Annahme hat den Zweck, die Konstruktion der Fehlerszenarien zu vereinfachen, gleichzeitig entfällt aber damit die Möglichkeit, mit der Lage des Modalwertes die Güte der Beschreibung des „wahren“ Wertes durch die Fuzzy-Fehlerumgebung nachträglich zu kalibrieren und zur Verbesserung der Abbildungsgüte zu nutzen. Gleichwohl wird es typisch für empirische Mess-Situationen sein, dass der fehlerbehaftete Messwert mangels besseren Wissens mit

[4] Es geht hier z.B. um Verzerrungen in der Wahrnehmung, wie wir sie beispielsweise von optischen Täuschungen her kennen.

[5] Vgl. Abschnitt *4.3* und dort v.a. S. 149.

dem Modalwert der zugehörigen Fuzzy-Beobachtung identifiziert wird.

Weitere Voraussetzungen an die Fuzzy-Daten dienen der technischen Vereinfachung bei der Berechnung der Fuzzy-linearen Approximationsfunktion. Insbesondere soll gelten, dass $0 \leq \inf_i \operatorname{supp}(\widetilde{X}_i)$, d.h. dass alle X-Werte strikt positiv sind, damit bei den arithmetischen Operationen nur die Vorzeichen der Funktionsparameter zu berücksichtigen sind.[6] Diese Vorgabe bedeutet gegenüber empirischen Fuzzy-Daten keine Einschränkung, weil sie durch eine Transformation der Daten immer erreicht werden kann. Tatsächlich kann aufgrund der besonderen Form von Fuzzy-linearen Funktionen die Analysequalität sogar verbessert werden, indem die Fuzzy-Daten so transformiert werden, dass alle Werte für ein exogenes Merkmal in der positiven bzw. der negativen Halbachse zu liegen kommen. O.E. gelte darüber hinaus, dass die unscharf beobachteten, „wahren" Werte aufsteigend sortiert sind, d.h. $0 \leq x_1^* \leq \cdots \leq x_n^*$. Außerdem werden für das Benchmarking nur trianguläre Fuzzy-Zahlen verwendet, weil diese die einfachste Form von asymmetrischen LR-Fuzzy-Mengen sind.

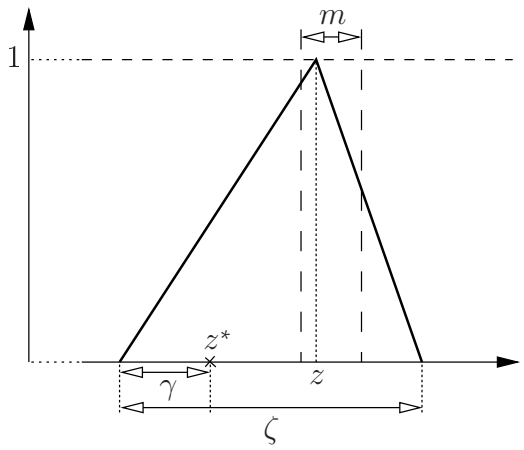
Für die Fuzzy-Fehler-Daten werden zunächst die „wahren" Werte bestimmt. Da auch die Streuung der X-Werte einen Einfluss auf die Approximationsparameter hat, werden diese im ersten Schritt durch eine Zufallsziehung aus dem relevanten Intervall — das ist in diesem Fall das Intervall $[0, 10]$ — vorgegeben. Daraus werden die Y-Werte mit der vorgegebenen „wahren" linearen Funktion berechnet. Wir berücksichtigen den Fall ungestörter Y-Werte und den Fall, dass die Y-Werte mit einer Normalverteilung um den „wahren" Zusammenhang variieren, also den einfachen Fall des klassischen Regressionsmodells.

Die Untersuchung der Fallbeispiele mit isolierten Fehlern in der unabhängigen Variablen bzw. in der abhängigen Variablen sollen eine Vorstufe für die Untersuchung der Fallbeispiele mit Fehlern sowohl in den Beobachtungswerten für die unabhängigen als auch für die abhängigen Variablen bilden. Dieses Vorgehen zielt darauf hin, dass die Fehlereffekte bei der Approximation mit Fuzzy-Fehlern $(\widetilde{X}_i, \widetilde{Y}_i)$ als Zusammenwirken von Fehlereffekten interpretiert werden können, die durch fehlerhafte Einzelvariablen verursacht sind. Dazu ist sicherzustellen, dass die Fehlereinflüsse in den Einzelvariablen sich bei der Zusammenführung nicht gegenseitig kompensieren, sondern eher gegenseitig verstärken und somit die Korrekturwirkung der untersuchten Methoden besser verdeutlicht werden können.

Für das Benchmarking werden zwei unterschiedliche Verfahren zur Erzeugung von günstigen Fuzzy-Merkmalswerten verwendet, die den Einfluss von typischen Fehlern auf die generierten „wahren" Werte beschreiben.

Mit dem ersten Verfahren soll eine Situation nachgebildet werden, in der die Messwerte einem additiven, systematischen Fehler unterliegen, der im wesentlichen konstant

[6] Vgl. dazu Abschnitt *3.1.*

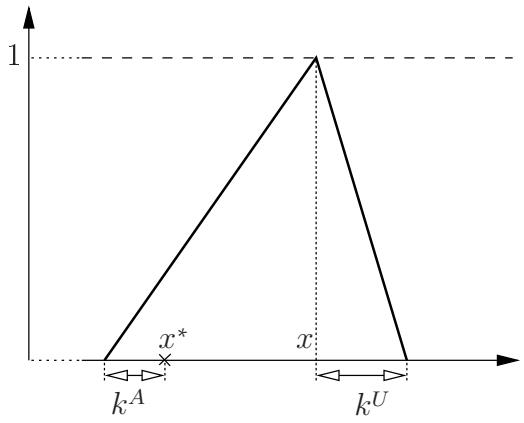


(Eigene Darstellung)

Abbildung 5.1: Konstruktion von Fuzzy-Daten mit einfachen, systematischen Fehlern

ist. Der Fehler in den Einzelmessungen soll aber dem systematischen Fehler nicht genau entsprechen, sondern von beliebigen Schwankungen überlagert sein, so dass er sich nur im Mittel über alle Messwerte abzeichnet. Bei einer vorliegenden Beobachtung z würde man in dieser Situation den Einfluss des systematischen Fehlers in einem Fuzzy-Datum abbilden, indem die Trägermenge so weit aufgespannt wird, dass sie den „wahren“ Wert sicher enthält. Die Fuzzy-Menge ist dann schief in Richtung von z, so dass z die Trägermenge proportional zum systematischen Fehler in einem festen Verhältnis teilt. Das Konstruktionsverfahren wird in Abbildung 5.1 verdeutlicht. Zur Generierung eines Datensatzes für dieses Fehlerszenario wird zunächst die Trägermenge des Fuzzy-Merkmalswertes $\widetilde{Z}_i$ bestimmt, indem die Breite der Trägermenge ζ_i mit einer Gleichverteilung über dem Intervall $[0, K]$ gezogen wird und die Lage von z_i^* relativ zum linken Rand ebenfalls gleichverteilt im Intervall durch $\gamma_i \in [0, \zeta_i]$ festgelegt wird. Dieses Vorgehen sorgt dafür, dass die Fuzzy-Daten eine beliebige Breite annehmen, die sich nicht monoton verhält. Die Spannweite der Messwerte ist dabei unabhängig von $d(\{0\}, \widetilde{X})$. Außerdem schöpfen die „wahren“ Werte die gesamte Breite der möglichen Werte aus, ohne dass eine Häufung auftritt. Die Modalwerte z_i werden nun so berechnet, dass sie die Trägermenge in einem festen, asymmetrischen Verhältnis m_Z teilen, dabei wird der Verhältniswert m_Z aus einem gegebenen Bereich $m \subset [0, 1]$ gezogen. Insbesondere ist die Lage der Modalwerte im Support damit nicht unmittelbar an die Lage der „wahren“ Werte gebunden, gleichwohl weichen sie systematisch davon ab. Der Wert des systematischen Fehlers ergibt sich ex post als Mittelwert der Differenzen $z_i - z_i^*$.

Damit die Fallbeispiele so aufeinander aufbauen, dass bei Fuzzy-Daten $(\widetilde{X}_i, \widetilde{Y}_i)$ die Fehlereffekte sich verstärken, werden die Modalwerte für alle $\widetilde{X}_i$ so bestimmt, dass sie die Trägermenge im Verhältnis 0.7 bis 0.9 teilen. Die Modalwerte für die $\widetilde{Y}_i$ teilen demgegenüber die Trägermenge im Verhältnis 0.2 bis 0.4. Dabei werden die Verhältniswerte gleichverteilt



(Eigene Darstellung)

Abbildung 5.2: Konstruktion von Fuzzy-Daten mit korrelierten Fehlern

aus den jeweils angegebenen Intervallen gezogen. In der Kombination der Messfehler wird dann der negative Fehlereffekt im Achsenabschnitt der Approximationsfunktion teilweise kompensiert, im Steigungsparameter verstärkt sich hingegen die Tendenz zur Verringerung der Steigung.

Mit dem zweiten Verfahren soll in stark vereinfachter Form eine Situation reproduziert werden, in der die Fehlerabweichungen in X mit der Entfernung $d(\{0\}, \widetilde{X})$ korreliert sind oder in der die Fehlerabweichungen in Y mit der Entfernung $d(\{0\}, \widetilde{X})$ — d.h. gleichfalls mit der Entfernung von X vom Koordinatenursprung — korreliert sind. In einem ersten Schritt werden Modalwerte zu den vorliegenden „wahren“ Werten bestimmt. Dabei wird die Korrelation mit Hilfe einer arctan-Funktion modelliert, die ihren Nullpunkt in der Mitte des Intervalls annimmt, das durch die Inputwerte aufgespannt wird. Auf diese Weise kann erreicht werden, dass die Modalwerte der Fuzzy-Daten bis zum Mittelpunkt der Datenwolke immer näher an den „wahren“ Werten zu liegen kommen, und in der Mitte außerdem ein Seitenwechsel relativ zu den „wahren“ Werten stattfindet. Zur Berechnung der Modalwerte werden hier die folgenden Funktionen verwendet:

$$\text{Für } X: \qquad [\widetilde{X}_i]^1 = x_i^* + c_x \cdot \arctan(x_i^* - \frac{x_1^* + x_n^*}{2}), \tag{5.1}$$

$$\text{für } Y: \qquad [\widetilde{Y}_i]^1 = y_i^* + c_y \cdot \arctan(x_i^* - \frac{x_1^* + x_n^*}{2}), \tag{5.2}$$

wobei $c_x, c_y \in \mathbb{R}$ geeignete Kontraktionsparameter seien. Damit die Fehlereffekte sich im Kombinationsfall bei Fuzzy-Daten $(\widetilde{X}_i, \widetilde{Y}_i)$ verstärken, wurden $c_x < 0$ und $c_y > 0$ gewählt; konkret $c_x = -0.4$ und $c_y = 0.2$. Tendenziell wird die Steigung der Approximationsgeraden dadurch nach oben verzerrt, wohingegen der Achsenabschnitt nach unten verzogen wird.

Die rechten und linken Spannweiten werden in Abhängigkeit davon konstruiert, auf welcher Seite der „wahre“ Wert eingeschlossen ist. Das Konstruktionsverfahren wird in Abbildung 5.2 gezeigt. Auf der freien Seite wird die Spannweite k^U gleichverteilt im Inter-

vall $[0, \zeta^U]$ bis zum vorgegebenen Maximalbetrag ζ^U gewählt. Auf der gebundenen Seite reicht die Spannweite gleichverteilt im Intervall $[0, \zeta^A]$ bis zum vorgegebenen Maximalbetrag ζ^A über den „wahren“ Wert hinweg, d.h. für X beträgt die Spannweite auf der gebundenen Seite $|[\widetilde{X}_i]^1 - x_i^*| + k_i^A$, wobei k_i^A der zufällige Wert für die zusätzliche Spannweite ist. Die Gesamtspannweiten der so konstruierten Fuzzy-Daten sind in der Mitte des Wertebereiches minimal und nehmen zu beiden Enden hin zu. Zudem haben die Modalwerte eine Drift relativ zu den Enden der Trägermengen und zwar z.B. bei den X-Werten von weit links nach weit rechts, wobei der „wahre“ Wert eine deutliche Entfernung vom Modalwert hat und eher am langen Ende des Fuzzy-Werts liegt. Das generierte Fehlerszenario ist somit eine gute Illustration für eine Situation bei der die Gewichtung durch den Rand der Fuzzy-Menge relevant ist. Die Schmetterlingsform der Datenungenauigkeiten ist plausibel für die konstruierte Situation, weil dort, wo der Abstand zwischen „wahren“ Wert und Messwert am geringsten ist, die Messgenauigkeit in der Tat am höchsten ist.

Weitere Fehlerszenarien mit besonderen Effekten werden durch eine manuelle Nachbearbeitung von simulierten Datensätzen erzeugt, da sie nur schwer automatisiert werden können. Dazu gehören Fehlerszenarien, die einzelne bzgl. ihrer Lage oder ihrer Unschärfe extreme Fuzzy-Mengen enthalten oder aber Datensätze mit großen und stark wechselnden Fehlerabweichungen.

5.1.2 Vergleichskriterien

Ein Vergleich zwischen einer klassischen Kleinste Quadrate-Approximation und einer Kleinste Quadrate Fuzzy-Approximation ist im eigentlichen Sinne nicht möglich, da es sich dabei um Approximationsverfahren handelt, die in der Qualität völlig verschieden sind. Im ersten Fall erhalten wir eine Funktion, die den Zusammenhang von reellwertigen Punkten wiedergibt, im zweiten Fall hingegen eine Funktion, die den Zusammenhang von Fuzzy-Werten abbildet, die wir als Fehlerumgebungen von reellwertigen Punkten interpretieren. Auch eine Kleinste Quadrate Fuzzy-Approximation, die sich als einfache lineare Funktion darstellt, hat mithin eine völlig andere Aussage als eine klassische Ausgleichsgerade, da sie auch den Zusammenhang zwischen den Ungenauigkeitscharakteristiken der Fuzzy-Daten einschließt und simultan ausgleicht. Die Fuzzy-Modellierung der Regression mit Fehlern in den Variablen erweist aber erst dann einen Nutzen für die Datenanalyse, wenn Rückschlüsse über die Zusammenhänge zwischen Messfehlern, Eigenschaften der Fuzzy-Daten und Eigenschaften der Fuzzy-Approximationsfunktion möglich sind.

Die reellwertigen Regressionsansätze werden daher als Extrembeispiele und als Vergleichsmaßstab für die Fuzzy-Approximationsfunktion eingesetzt. Dabei wird für verschiedene Szenarien von Fuzzy-Fehlerumgebungen durchgespielt, wie sich die Unterschiede

zwischen den Approximationsergebnissen darstellen und welche Schlüsse daraus für die Interpretation gezogen werden können. Ziel dabei ist es, Zusammenhänge bei der Fuzzy-Approximation zu erkennen, die zusätzliche Aussagen für die Datenanalyse ermöglichen sowie Einflussfaktoren im Bezug auf die Modellierung der Fuzzy-Daten zu identifizieren, die für die Güte der Approximationsfunktion besonders relevant sind. Das Benchmarking ist also als Illustration und als Vorstufe zu einem systematischen Vergleich zwischen verschiedenen Ansätzen zur Regression bei Fuzzy-Fehlern zu sehen, insbesondere sind die Ergebnisse nicht ohne weiteres verallgemeinerbar.

Für das Benchmarking werden die folgenden klassischen Kleinste Quadrate-Approximationen als Vergleichsmaßstab herangezogen:

Die *Kleinste Quadrate-Approximation an die „wahren" Werte* wird als Abbildung des „wahren" Zusammenhangs betrachtet. Der Grund für die Verwendung der Approximation anstelle der vorgegebenen linearen Funktion liegt darin, dass die Datensätze für die Fallbeispiele eher klein sind, so dass für die Beispiele, in denen normalverteilte Y unterstellt werden, die Ziehungseffekte aus dem Vergleich ausgeschaltet werden sollen.

Bei allen simulierten Fehlerszenarien werden die Modalwerte der Fuzzy-Daten mit den fehlerhaften Messungen identifiziert. Daher gibt die *Kleinste Quadrate-Approximation der Modalwerte* $([\widetilde{X}_i]^1, [\widetilde{Y}_i]^1)$ das Ergebnis einer naiven Regressionsrechnung wieder, bei der der Einfluss der Fehler nicht beachtet wird.

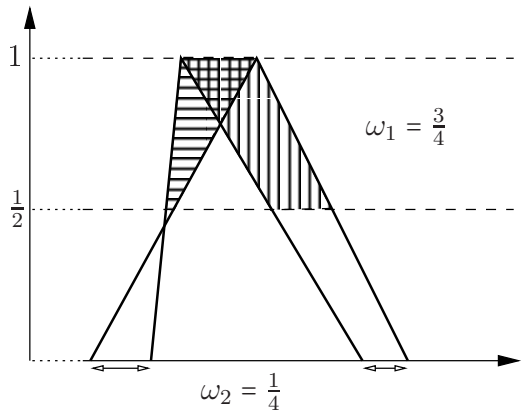
In den Modellen mit fehlerhaftem X wird außerdem eine *Kleinste Quadrate-Approximation über die Werte in den Endpunkten der Träger von* X durchgeführt. Konkret wird dazu im Modell mit genauen Outputs die Ausgleichsgerade für die erweiterte Menge von artifiziellen Beobachtungswerten

$$(x_1^L(0), y_1), \ldots, (x_n^L(0), y_n) \quad \text{und} \quad (x_1^R(0), y_1), \ldots, (x_n^R(0), y_n)$$

berechnet. Im Modell mit fehlerhaften In- und Outputs werden zur Vereinfachung anstelle von genauen y_i die Modalwerte $[\widetilde{Y}_i]^1$ sowie zum Vergleich die Steiner-Punkte und die „wahren" Werte eingesetzt. Dieser dritte Ansatz bildet ab, wie die Approximation ausfällt, wenn die X-Werte mit extremen Fehlern in beide Richtungen gemessen werden, so dass diese sich bei der Approximation z.T. gegenseitig ausgleichen. Mit diesem Ansatz lässt sich demonstrieren, wie stark die Abweichung bzw. Verzerrung ist, die sich aus der Fehlervariabilität bei der Messung von X ergibt.[7]

Die *Kleinste Quadrate Fuzzy-Regression* wird hauptsächlich auf der Basis der δ_2-Metrik durchgeführt. Mit Hilfe der Klasse der Fuzzy-Metriken gemäß Bertoluzza u. a. [1995] kann

[7] Vgl. das anschauliche Beispiel zur Attenuation in Schneeweiß 1990, S. 224 sowie die illustrierende Abbildung 2.2 auf S. 42.



(Eigene Darstellung)

Abbildung 5.3: Berechnung des Abstandes in der gewichteten Metrik δ_G

außerdem untersucht werden, welchen Effekt eine unterschiedliche Gewichtung der α-Niveaus bei der Abstandsmessung auf die Funktionsapproximation hat. Definition und einige wichtige Eigenschaften der Metriken nach Bertoluzza u.a. sind im Anhang *A.1*, Definitionen A.6 und A.9 beschrieben. Der besondere Vorzug dieser Metriken liegt darin, dass sie bei LR-Fuzzy-Mengen nur zu veränderten Gewichtungsfaktoren in den Rechenalgorithmen führen, so dass diese leicht angepasst werden können.[8]

Als Vergleichsmetriken wählen wir zum einen die *Hagaman-Metrik*[9] δ_H, in die wie bei der δ_2-Metrik alle α-Niveaus eingehen, wobei die Niveaus mit hoher Zugehörigkeit stärker gewichtet werden als Niveaus mit niedriger Zugehörigkeit. In Anlehnung an die allgemeine Definition für die erweitere Familie der Kleinste Quadrate Fuzzy-Metriken ist δ_H definiert durch

$$\begin{aligned}\delta_H^2(\widetilde{A},\widetilde{B}) &= \int_{[0,1]\times\mathbb{S}^0} (s_{\widetilde{A}}(\alpha,u) - s_{\widetilde{B}}(\alpha,u))^2 f(\alpha)\, d\lambda^1(\alpha) \otimes \lambda^{\mathbb{S}^0}(u) \\ &= \frac{1}{2}\int_0^1 [(-a^L(\alpha)+b^L(\alpha))^2 + (a^R(\alpha)-b^R(\alpha))^2] f(\alpha)\, d\alpha \end{aligned} \tag{5.3}$$

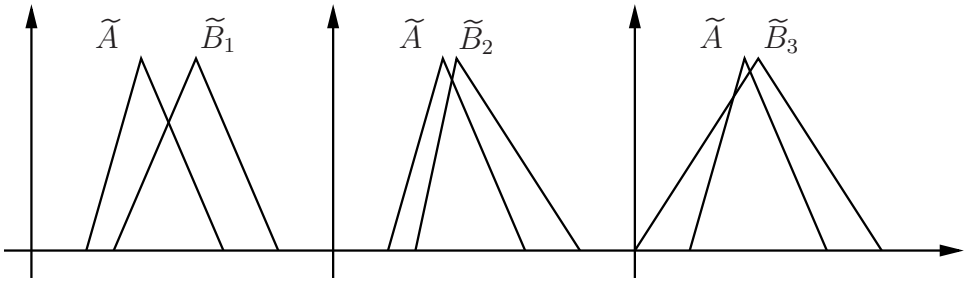
wobei f eine stetige Gewichtungsfunktion ist, für die gelte $f|_{(0,1]} > 0$ und $\int_0^1 f d\lambda^1 = 1$. Damit gilt für trianguläre Fuzzy-Mengen, dass nur die Gewichte $\|L\|_2^2$ und $\|L\|_1$ in der Formel zur Berechnung der Metrik angepasst werden müssen. Wir wählen $f(\alpha) = 2\alpha$ und erhalten damit $\|L_{H,f}\|_2^2 = \|R_{H,f}\|_2^2 = \frac{1}{12}$ und $\|L_{H,f}\|_1 = \|R_{H,f}\|_1 = \frac{1}{6}$.

Als weitere Variante wählen wir die folgende gewichtete Metrik

$$\delta_G^2(\widetilde{A},\widetilde{B}) = \frac{3}{4}\delta_2^2(\widetilde{A}|_{0,5},\widetilde{B}|_{0,5}) + \frac{1}{8}[(-a^L(0)+b^L(0))^2 + (a^R(0)-b^R(0))^2],$$

[8] Vgl. Diamond und Körner 1997, S. 30.

[9] Vgl. [Bárdossy u. a., 1992].



(Eigene Darstellung)

Abbildung 5.4: Unterschiede zwischen den Fuzzy-Metriken $\delta_2, \delta_H, \delta_G$ — Datenwerte wie im Text angegeben (S. 162).

	$\widetilde{A}, \widetilde{B}_1$	$\widetilde{A}, \widetilde{B}_2$	$\widetilde{A}, \widetilde{B}_3$
δ_2	0.7917	0.2917	0.3558
δ_H	0.8542	0.1979	0.1988
δ_G	0.8203	0.2695	0.3414

Tabelle 5.1: Abstandswerte für $\delta_2, \delta_H, \delta_G$

wobei $\widetilde{A}|_{0,5}$ die bei $\alpha = 0,5$ abgeschnittene Fuzzy-Menge sei, die durch $\mu_{\widetilde{A}|_{0,5}}(u) = \mu_{\widetilde{A}} \cdot 1_{[\widetilde{A}]^{0,5}}$ definiert ist. δ_G^2 bildet somit einen Kompromiss zwischen δ_2^2 für die obere Hälfte der Fuzzy-Mengen, die mit einem Gewicht von $\omega_1 = \frac{3}{4}$ gewertet wird, und dem zugehörigen Intervallabstand der Trägermengen, der mit einem Gewicht von $\omega_2 = \frac{1}{4}$ eingeht. Die Ermittlung von δ_G wird in Abbildung 5.3 verdeutlicht. In der üblichen Weise werden als Faktoren $\|L_G\|_2^2 = \|R_G\|_2^2 = \frac{5}{32}$ und $\|L_G\|_1 = \|R_G\|_1 = \frac{7}{32}$ berechnet. Die Unterschiede zwischen den Metriken können am Beispiel der Fuzzy-Mengen in Abbildung 5.4 verdeutlicht werden. Die abgebildeten Fuzzy-Mengen sind gegeben durch $\widetilde{A} = (2; 1, 1.5)_L, \widetilde{B}_1 = (3; 1.5, 1.5)_L, \widetilde{B}_2 = (2.25; 0,75, 2.25)_L, \widetilde{B}_3 = (2.25; 2.25, 2.25)_L$. Während $\widetilde{B}_1$ in relativ gleichmäßiger Entfernung zu $\widetilde{A}$ liegt, liegen die Modalwerte von $\widetilde{B}_2$ und $\widetilde{B}_3$ nahe beim Modalwert. Der einzige Unterschied von $\widetilde{B}_2$ und $\widetilde{B}_3$ besteht darin, dass die Trägermenge von $\widetilde{B}_3$ stärker von $\widetilde{A}$ abweicht. Die gemäß $\delta_2, \delta_H, \delta_G$ ermittelten Abstandswerte sind in Tabelle 5.1 aufgelistet. Da die δ_2-Metrik keine Gewichtungen zwischen den α-Niveaus vornimmt, kann sie als „neutrale“ Abstandsmessung betrachtet werden. Die Hagaman-Metrik δ_H legt demgegenüber starkes Gewicht auf die Modalwerte, entsprechend ist der Abstand zu $\widetilde{B}_2$ und $\widetilde{B}_3$ besonders klein. Die Metrik δ_G reagiert ebenfalls stärker auf Änderungen in den Modalwerten als die δ_2-Metrik. Im Unterschied zu δ_H ist sie aber sensibler für Abweichungen der Trägermenge, wie der Übergang von $\widetilde{B}_2$ zu $\widetilde{B}_3$ zeigt.

5.2 Benchmarking der Kleinste Quadrate Fuzzy-Regression bei Fuzzy-Fehlern

Die Ergebnisse einer Fuzzy-Approximation werden hier anhand der simulierten Fuzzy-Daten zu den alternativen Fehlerszenarien durchgespielt, die in Abschnitt *5.1* vorgestellt wurden. Dabei werden zunächst Datensätze mit genauen Inputs sowie mit genauen Outputs untersucht, weil dadurch die unterschiedlichen Effekte von Fehlern in Y und Fehlern in X isoliert betrachtet werden können, bevor sie im allgemeinen Ansatz zusammengeführt werden. Die Güte einer Approximationsfunktion wird dabei hauptsächlich danach beurteilt, wie gut der Steigungsparameter a_1 die Steigung der Kleinste Quadrate-Approximation der „wahren" Werte annähert. Dies ist auch dadurch begründet, dass in den Ansätzen zur Kleinste Quadrate-Regression der Achsenabschnitt üblicherweise in Abhängigkeit vom Steigungsparameter als „Saldierungsgröße" berechnet wird. Die Zahl der Varianten sollte nicht übermäßig erhöht werden, daher werden in allen Beispielen dieselben Strukturparameter vorgegeben und zwar als Achsenabschnitt $a_0 = 1{,}5$ und als Steigung $a_1 = 0{,}3$.

Die Datensätze, die für die untersuchten Fallbeispiele generiert wurden, sind in Anhang *A.2* dokumentiert. Es soll hier bereits für alle Fallbeispiele vorausgeschickt werden, dass die Fuzzy-Approximation mit δ_2 und den davon abgeleiteten Metriken zu Ergebnissen führen, die eine ähnliche Größenordnung haben, wie die üblichen Kleinste Quadrate-Ergebnisse. Beispielsweise liegen bei der Kleinste Quadrate Fuzzy-Approximation mit Fuzzy-Inputs die δ_2-Residuen der Fuzzy-Approximation immer zwischen dem Residuum des linken Endwertes und dem Residuum des rechten Endwertes zur Endwerteapproximation. Damit bewegen sich die Fuzzy-Approximationsverfahren auch bei den Gütekennziffern im wesentlichen im Rahmen der Dimensionen, die auch für die klassische Regressionsrechnung gelten. Es ist allerdings darauf hinzuweisen, dass das Residuum der Fuzzy-Approximation als metrischer Abstand von Fuzzy-Werten berechnet wird und somit vorzeichenfrei ist. Das Vorzeichen kann gesondert nach einer Ordnungsrelation für Fuzzy-Zahlen zugewiesen werden. Wir verwenden hierzu das Vorzeichen der Differenz $\sigma_{\widetilde{Y}_i} - \sigma_{f(\widetilde{X}_i,\widetilde{A})}$ zwischen den Steiner-Punkten von $\widetilde{Y}_i$ und dem zugehörigen Fuzzy-Approximationswert, weil die Steiner-Punkte die Schwerpunkte der Fuzzy-Mengen im Bezug auf die δ_2-Metrik darstellen.[10]

Es ist zu beachten, dass in diesem Abschnitt ausnahmsweise die Dezimalzahlen in amerikanischer Schreibweise mit einem Punkt verwendet werden, um die Lesbarkeit der Notation für die Fuzzy-Daten in LR-Schreibweise zu erhöhen. Wir schreiben also ab jetzt 1.5 statt 1,5.

[10] Für die alternativen Metriken δ_H und δ_G können angepasste minimierende Werte bestimmt werden, die wir als die zugehörigen Steiner-Punkte verwenden. Vgl. dazu Definition A.9 im Anhang *A.1*.

5.2.1 Modell mit genauen Inputs

Als Modell mit genauen Inputs betrachten wir eine Regression bei Fuzzy-Fehlern in den Daten für Y, d.h. es liegen Daten $(x_i, \widetilde{Y}_i)$ vor. Dieser Approximationsansatz ist dann sinnvoll, wenn die Messungenauigkeiten in Y gegenüber der zufälligen Variabilität auffällig sind. Er bietet sich u.a. an, wenn Einflussfaktoren auf die Ungenauigkeit in Y nicht eindeutig identifiziert werden können. In dieser Situation kann entweder eine defuzzifizierter Ausgleich durchgeführt werden (vgl. Abschnitt *3.3.1*) oder eine Fuzzy-Charakteristik angepasst werden (vgl. Abschnitt *3.3.2*).

Beim defuzzifizierten Ausgleich wird der Zusammenhang der genauen Inputs x_i mit den möglichen „wahren“ y_i^* in den Mittelpunkt gestellt. Dabei wird eine Funktion vom Typ

$$y = a_0 + a_1 \cdot x$$

approximiert, wobei die Parameter $a_0, a_1 \in \mathbb{R}$ sind. Da die Fuzzy-Daten LR-Fuzzy-Intervalle sind, kann die Kleinste Quadrate-Bedingung $\sum_i \delta_2^2(\widetilde{Y}_i, a_0 + a_1 x_i)$ gemäß Formel (3.11) als Gleichung in den Parametern a_0, a_1 umgeschrieben werden. Das Minimum kann wie üblich durch Nullsetzen der partiellen Ableitungen $\partial_{a_0}, \partial_{a_1}$ berechnet und unter Verwendung der Berechnungsformel (A.6) für $\sigma_{\widetilde{Y}_i}$ zusammengefasst werden zu:

$$\begin{aligned} \hat{a}_0 &= \overline{\sigma_{\widetilde{Y}\cdot}} - \hat{a}_1 \bar{x}_\cdot, \\ \hat{a}_1 &= \frac{c(x, \sigma_{\widetilde{Y}})}{c(x,x)}, \end{aligned} \tag{5.4}$$

wobei $c(u,v) = \frac{1}{n}\sum_{i=1}^n (u_i - \bar{u}_\cdot)(v_i - \bar{v}_\cdot)$ und $\overline{\sigma_{\widetilde{Y}\cdot}} = \frac{1}{n}\sum_i \sigma_{\widetilde{Y}_i}$ als Mittelwert der Steiner-Punkte von $\widetilde{Y}_i$ gesetzt ist.

Die Kleinste-Quadrate-Approximation in der δ_2 Metrik berücksichtigt demzufolge die Zugehörigkeitsbewertung für die „wahren“ Outputs dadurch, dass die plausible mögliche Lage von y_i^* mit $\sigma_{\widetilde{Y}_i}$ charakterisiert und verallgemeinert wird. Dabei ist $\sigma_{\widetilde{Y}_i}$ die reelle Zahl, die unter δ_2 minimalen Abstand zu $\widetilde{Y}_i$ hat.[11] Aus dem einfachen Zusammenhang $\sigma_{\widetilde{Y}_i} = y_i + ||L||_1 (r_{Y_i} - l_{Y_i})$ für trianguläre Fuzzy-Daten[12] ist ersichtlich, dass bei der defuzzifizierten Anpassung eine Korrektur des Modalwertes um die gewichteten Anteile der Spannweiten berücksichtigt wird.

Falls stattdessen eine Fuzzy-lineare Funktion mit Fuzzy-Parametern zugrundegelegt wird, bildet der lineare Zusammenhang ab, wie die reellwertigen x_i mit den Fehlerumgebungen $\widetilde{Y}_i$ für den „wahren“ Wert y_i^* verknüpft sind. Der funktionale Zusammenhang

[11] Vgl. Satz A.8.
[12] Vgl. Satz A.9.

erzeugt dabei gegenüber der Genauigkeit der x_i zusätzliche Fuzziness in den Bildwerten. Die Fuzzy-lineare Funktion hat demzufolge Fuzzy-Parameter $\widetilde{A}_0, \widetilde{A}_1$ und wir approximieren Funktionen vom Typ:

$$\widetilde{Y} = \widetilde{A}_0 \oplus \widetilde{A}_1 \odot x.$$

Die linken (bzw. rechten) Spannweiten des Steigungsparameters verschwinden, wenn die linken (bzw. rechten) Spannweiten von $\widetilde{Y}_i$ in i eine nicht-wachsende Tendenz haben.[13]

Im Rahmen seiner Dissertation hat Körner [1997] Rechenverfahren für die Kleinste Quadrate Fuzzy-Regression hergeleitet und zur Verfügung gestellt.[14] Aus den dort vorgestellten Formeln ist direkt ersichtlich, dass die Approximationsparameter im wesentlichen durch die Kleinste Quadrate-Lösung jeweils für die Modalwerte sowie die linken und rechten Spannweiten berechnet werden können, d.h.

$$\begin{pmatrix}\hat{a}_0\\ \hat{a}_1\end{pmatrix} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{y}, \quad \begin{pmatrix}\hat{l}_{A_0}\\ \hat{l}_{A_1}\end{pmatrix} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{l}_Y, \quad \begin{pmatrix}\hat{r}_{A_0}\\ \hat{r}_{A_1}\end{pmatrix} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{r}_Y, \tag{5.5}$$

wobei $\mathbf{y} = (y_1, \ldots, y_n)^T, \mathbf{l}_Y = (l_{Y_1}, \ldots, l_{Y_n})^T, \mathbf{r}_Y = (r_{Y_1}, \ldots, r_{Y_n})^T$ und $\mathbf{Z} = (1_{(n)}, \mathbf{x})$ mit $1_{(n)} = (1, \ldots, 1) \in \mathbb{R}^n$ ist.

Eine Abweichung ergibt sich dann, wenn einer der Kleinste-Quadrate-Parameter für die Spannweiten < 0 wird. Dann ist eine quadratische Optimierung unter Vorzeichennebenbedingungen durchzuführen, die insbesondere auch eine Veränderung der Parameter für die Modalwertgerade nach sich zieht.

Als dritte Variante wurde der vereinfachter Modellansatz mit

$$\widetilde{Y} = \widetilde{A}_0^s \oplus a_1^s \cdot x$$

angepasst. Dieser führt zum selben Steigungsparameter a_1 gemäß Gleichung (5.4) wie die defuzzifizierte Approximation. Der Modalwert des Fuzzy-Achsenabschnitts $\widetilde{A}_0^s$ liegt in der Regel nahe beim Achsenabschnitt des defuzzifizierten Ansatzes a_0. Seine Spannweiten gleichen die empirische Tendenz in den Spannweiten aus, d.h. wenn bei der Anpassung der Fuzzy-Charakteristik eine Fuzziness im Steigungsparameter $\widetilde{A}_1$ verbleibt, dann ist die Spannweite von $\widetilde{A}_0^s$ größer als die Spannweite von $\widetilde{A}_0$, wie in Tabelle 5.2 zum Beispiel mit einfachem, systematischem Fehler abzulesen ist.

[13] Ein Ausweg besteht darin, dass die Forderung eines Fuzzy-linearen Zusammenhangs durch die schwächere Forderung eines linearen Zusammenhangs in den Modalwerten sowie eines davon unabhängigen linearen Zusammenhangs in den in den Spannweiten ersetzt wird, vgl. als Beispiele [Chang und Lee, 1994b; Diamond und Körner, 1997; D'Urso und Gastaldi, 2000].

[14] Vgl. [Körner und Näther, 1998; Diamond und Körner, 1997].

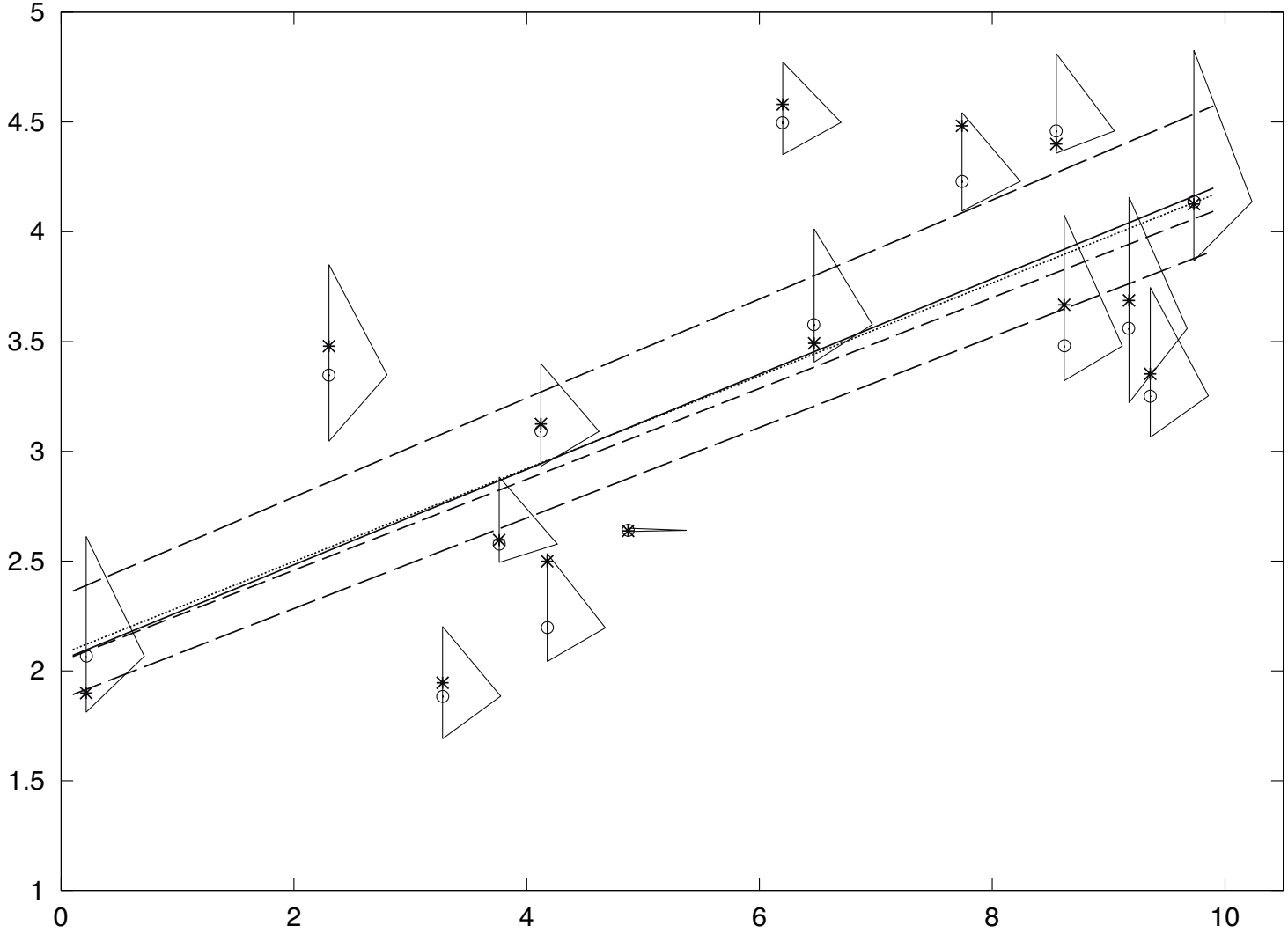


Abbildung 5.5: Modell mit genauen Inputs und einfachem systematischen Fehler: Die „wahren“ Werte der Outputs weichen mit Varianz $s^2 = 0.5$ zufällig vom ursprünglichen Zusammenhang ab.

Legende: „wahre“ Werte (*), fehlerhafte Beobachtungen (∘), „wahre“ Gerade (——), defuzzifizierte Approximation (·····), Modalwertgerade der Fuzzy-Approximation (– –), Begrenzungsgeraden des Trägers der Fuzzy-Approximation (— —)

	„wahre“ Werte	defuzz. Modell	vereinf. Modell	Fuzzy-Modell
a_0	2.0490	2.0752	$(2.0183; 0.1776, 0.4054)_L$	$(2.0444; 0.1730, 0.2963)_L$
a_1	0.2172	0.2114	0.2114	$(0.2070; 0.0008, 0.0185)_L$

Tabelle 5.2: Approximationsparameter des Datenbeispiels mit genauen Inputs und einfachem systematischen Fehler

Das Datenbeispiel in Abbildung 5.5 ist mit dem Algorithmus erzeugt, der Fuzzy-Daten mit einfachen systematischen Fehlern simulieren soll. Der systematische Fehler in den Beobachtungen ergibt sich „ex post“ als $\psi_\cdot = \frac{1}{n}\sum_i (y_i - y_i^*) = -0.06472$. Die Approximationsfunktion mit Fuzzy Parametern kann direkt gemäß den Kleinste Quadrate-Formeln in Gleichung (5.5) berechnet werden, so dass die Spannweiten sich in etwa als „Mittel“ über die Ergebnisspannweiten ergeben, und die Modalwertgerade der Fuzzy-Approximation mit der Kleinste Quadrate-Approximation der Modalwerte zusammenfällt. Es sind keine besonderen Effekte aus der zufälligen Variabilität der Y-Werte zu beobachten. Im Fallbeispiel weichen die Kleinste Quadrate-Parameter der linearen Approximation der „wahren“ Werte stark von den vorgegebenen Parametern ab. Die defuzzifizierte Approximation nähert die „wahre“ Gerade insgesamt gut an, insbesondere gibt sie den Steigungsparameter besser wieder als die Gerade der Modalwerte der Regression mit Fuzzy-Parametern, die — wie oben beschrieben — mit der Kleinste Quadrate-Approximation an die fehlerhaften Beobachtungen zusammenfällt. Die Güte der defuzzifizierten Approximation kann damit erklärt werden, dass die Messfehler in den Beobachtungen in den Steiner-Punkten $\sigma_{\widetilde{Y}_i}$, die gemäß Gleichung (5.4) der Berechnung vom a_1 zugrundeliegen, durch eine Lageanpassung relativ zum Verhältnis der linken und der rechten Spannweite antizipiert und damit proportional mit der gewählten Metrik korrigiert werden. Im vereinfachten Fuzzy-Modell erhält man zusätzlich zum Steigungsparameter noch eine Approximation der mittleren zusätzlichen Ungenauigkeit der Outputs Y gegenüber den Inputs X. Die Approximation der Fuzzy-Charakteristik zeichnet sich dadurch aus, dass sie die Fuzzy-Umgebungen in den Eingangsdaten gut annähert und zudem die leichte Drift in den Modalwerten der Fuzzy-Daten verdeutlicht. Allerdings ist darauf hinzuweisen, dass die gute Approximation des Achsenabschnitts im ausgewählten Beispiel eine Ausnahme ist und dadurch entsteht, dass der kleinste Fuzzy-Wert isoliert liegt und damit großen Einfluss auf alle Approximationsvarianten hat. Typischerweise liegt die Gerade der Modalwerte der Fuzzy-Regression hingegen parallel zur „wahren“ Gerade, was auch dem Einfluss des einfachen, systematischen Fehlers entspricht. Insgesamt sind aber die Abweichungen zwischen den verschiedenen Steigungsparametern eher klein. Die fehlerhaften Beobachtungen in Y verursachen somit also eine leichte Verringerung in der Steigung der Approximationsfunktion.

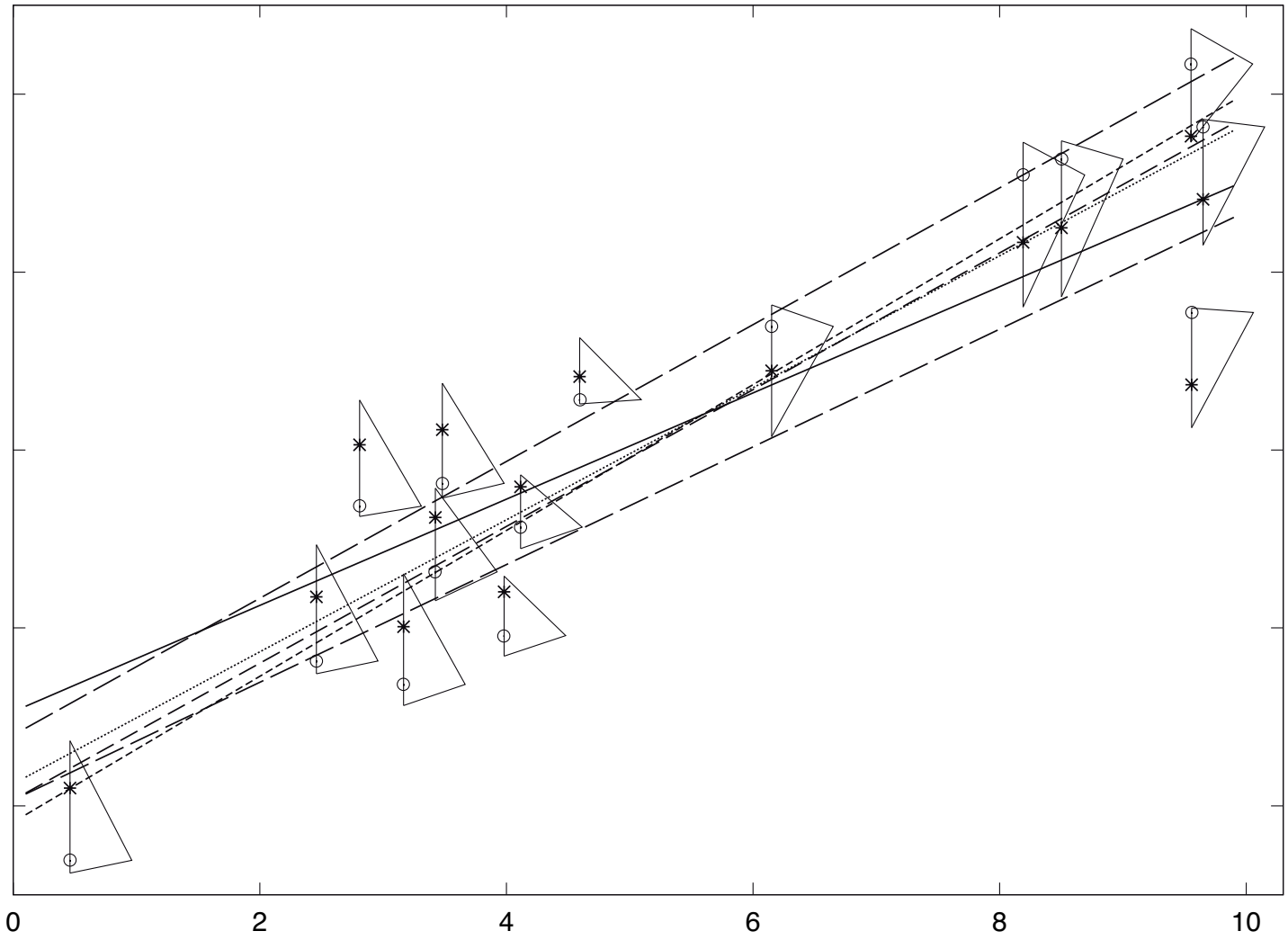


Abbildung 5.6: Modell mit genauen Inputs und korreliertem Fehler: Die „wahren“ Werte der Outputs weichen mit Varianz $s^2 = 0.3$ zufällig vom ursprünglichen Zusammenhang ab.

Legende: „wahre“ Werte (*), fehlerhafte Beobachtungen (∘), „wahre“ Gerade (——), Modalwerteapproximation (----), defuzzifizierte Approximation (.....), Modalwertgerade der Fuzzy-Approximation (– –), Begrenzungsgeraden des Trägers der Fuzzy-Approximation (— —)

	„wahre“ Werte	Modal- werte	defuzz. Modell	vereinf. Modell	Fuzzy- Modell
a_0	1.5299	0.9084	1.1242	$(1.1155; 0.3139, 0.3486)_L$	$(1.0328; 0, 0.3654)_L$
a_1	0.2986	0.4098	0.3710	0.3710	$(0.3843; 0.0535, 0)_L$

Tabelle 5.3: Approximationsparameter des Datenbeispiels mit genauen Inputs und korreliertem Fehler

Die Fuzzy-Daten in Abbildung 5.6 sind mit dem zweiten Algorithmus generiert, der korrelierte Fehler nachempfinden soll. Dabei fallen Modalwerteapproximation und die Modalwertgerade der Fuzzy-Approximation nicht zusammen. Dies ist bei der Fuzzy-Approximation mit reellwertigen Inputs nur dann der Fall, wenn bei der Kleinste Quadrate-Anpassung der Spannweiten von $\widetilde{\widetilde{A}_1}$ ein negativer Wert ermittelt wurde. Die dadurch ausgelöste Optimierung führt immer zu einer kleineren Steigung als bei der Modalwerteapproximation. Bei diesem Fehlerszenario führt das zu einer Verbesserung des Steigungsparameters in den Modalwerten der Approximation.

Auffällig ist außerdem der Verwringungseffekt bei der Approximation des Modells mit Fuzzy-Parametern, der durch die Drift in den Modalwerten relativ zu den Endpunkten entsteht. Dieser Effekt wird dadurch erzeugt, dass die linken Spannweiten in Richtung 0 schnell abnehmen und infolgedessen die linke Spannweite des Achsenabschnitts verschwindet. Hingegen nehmen für wachsende X die rechten Spannweiten schnell ab, was die rechte Spannweite des Steigungsparameters verschwinden lässt. Wie Tabelle 5.3 zeigt, wird hinsichtlich der reellwertigen Benchmarkgeraden auch hier wieder das beste Approximationsergebnis bei der defuzzifizierten Approximation mit einer Steigung von $a_1 = 0.3710$ erzielt, gefolgt von der Modalwertgerade der Fuzzy-Approximation. Schließlich ist die Modalwerteapproximation am stärksten nach oben verzerrt.

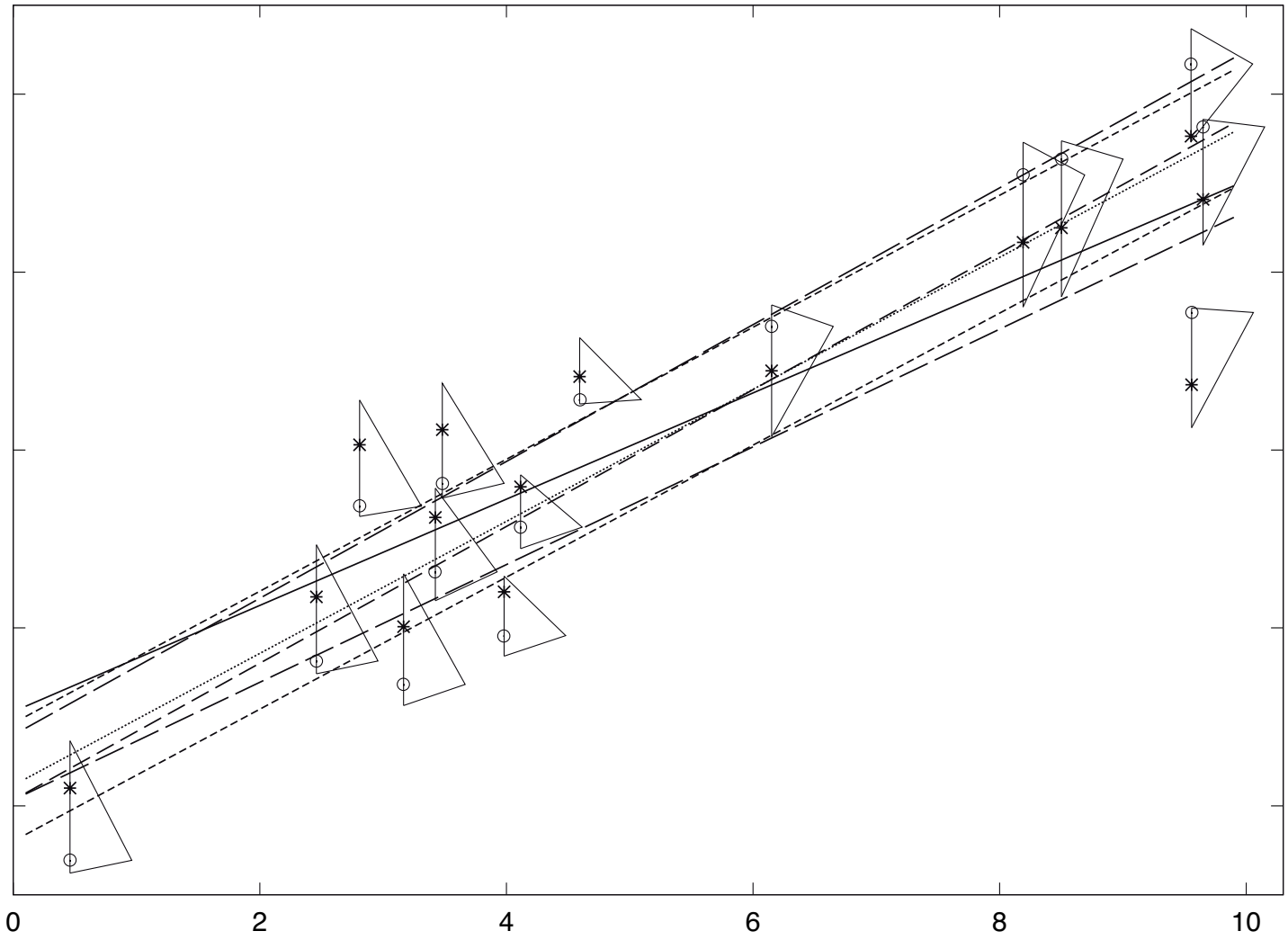


Abbildung 5.7: Modell mit genauen Inputs und korreliertem Fehler — Vergleich von Fuzzy-Approximation und vereinfachter Fuzzy-Approximation

Legende: „wahre“ Werte (*), fehlerhafte Beobachtungen (∘), „wahre“ Gerade (——), Modalwertgerade der Fuzzy-Approximation (– –), Begrenzungsgeraden des Trägers der Fuzzy-Approximation (— —), Modalwertgerade der vereinfachten Approximation (.....), Begrenzungsgeraden des Trägers der vereinfachten Approximation (----)

Abbildung 5.7 zeigt, dass sowohl Fuzzy-Approximation als auch die Approximation des vereinfachten Modells den „wahren“ Zusammenhang zu einem großen Teil einhüllen, so dass der Korridor des 0-Niveaus eine gute Abschätzung für den „wahren“ Zusammenhang darstellt. Allerdings darf dies nicht als Konfidenzintervall missverstanden werden. Die Fuzzy-Approximation ist eher als eine Abbildung der Fehlersensitivität durch die vorliegenden Daten und deren Ungenauigkeiten zu verstehen. Dabei werden allerdings nicht die maximalen Fehlerausschläge abgebildet, sondern deren Kleinste Quadrate-Approximation und mithin so etwas wie eine „mittlere“ oder „wesentliche“ Fehlersensitivität. In einem induktiven Modell wäre es hingegen möglich, über die Fuzzy-lineare Funktion hinaus noch Konfidenzintervalle für die Parameterschätzer bzw. die Approximationsfunktion zu bestimmen. Bei der vereinfachten Approximation tritt im Gegensatz zur Fuzzy-Approximation der Verwringungseffekt nicht auf, der einen Hinweis darauf gibt, wie sich die Modalwerte der Fuzzy-Daten im Verhältnis zu den Endpunkten der jeweiligen Trägermengen verschieben.

Die Verwendung einer der alternativen Metriken würde bei diesem Datenbeispiel eine Änderung in den Fuzzy-Parametern bewirken, da sich bei beiden alternativen Metriken die Modalwerte stärker auswirken. Z.B. verschlechtern sich die Parameter der Fuzzy-Approximation mit δ_H zu $\widetilde{A}_0 = (0.9930; 0, 0.3558)_L, \widetilde{A}_1 = (0.3932; 0.0561, 0)_L$.

Zusammenfassend kann festgestellt werden, dass die defuzzifizierte Approximation bei den betrachteten Beispielen für die beiden Fehlerszenarien mit genauen Inputs die „wahre“ Gerade am besten angenähert hat. Dies zeigt, dass mit dem Übergang zum Steiner-Punkt — und zwar derjenige für die δ_2-Metrik — eine geeignete und wohlinformierte Korrektur für die fehlerbehafteten Outputs vorgenommen werden kann, sofern die Fehlereinflüsse nach dem Ansatz der vorgestellten Fehlerszenarien modelliert werden können. Der vereinfachte Fuzzy-Ansatz liefert als einzige zusätzliche Information die mittlere Gesamtunschärfe in den Fuzzy-Daten. Demgegenüber sind zwar die Parameter der Modalwertgerade der Fuzzy-Approximation im Vergleich zur defuzzifizierten Approximation stärker verzerrt, sie geben aber Auskunft über lineare Tendenzen in den Fuzzy-Daten, die in der Form der Fuzzy-linearen Funktion abgebildet werden.[15] Da bei der Modellierung von Fuzzy-Daten typischerweise eine Fuzzy-Umgebung für den fehlerbehafteten Messwert konstruiert wird, so dass dieser zum Maßstab wird und nicht der unbekannte „wahre“ Wert, und da Unschärfe im Fuzzy-linearen Modell irreversibel ist, d.h. negative Unschärfe ist auszuschließen, ist die Interpretation der Modellierungstendenz in den Daten allerdings erschwert. Im zweiten Fallbeispiel wird im Verwringungseffekt nur die Tendenz in den linken Spannweiten vollständig abgebildet. In den rechten Spannweiten der Approximationsfunktion kann das

[15] Bei den untersuchten Fallbeispielen gab die Summe der Abstandsquadrate R^2 keinen Hinweis darauf, ob die Fuzzy-Approximation oder die vereinfachte Fuzzy-Approximation grundsätzlich zu bevorzugen wäre. Ein validierter Nachweis bleibt einer umfangreicheren Simulationsstudie überlassen.

Verschwinden der rechten Spannweiten nur dadurch abgebildet werden, dass die rechte Spannweite des Fuzzy-Steigungsparameters 0 wird, d.h. dass die rechte Spannweite der Funktion parallel zur Modalwertgerade ist. Eine vollständige Abbildung des Effektes würde es hingegen erfordern, dass zu negativen Spannweiten übergegangen werden kann.

5.2.2 Modell mit genauen Outputs

Als weiteren isolierten Ansatz betrachten wir das Fuzzy-Modell, in dem umgekehrt die Inputs ungenau und die Outputs genau sind. Wir haben somit Fuzzy-Daten $(\widetilde{X}_i, y_i)$. Dieser Ansatz ist vergleichbar mit dem klassischen Messfehlermodell der Regression. Der lineare Zusammenhang bildet ab, wie die X-Fehlerumgebungen mit den genauen Y zusammenhängen. Da Fuzzy-linearen Funktionen Fuzzy-Eingangswerte aufgrund des Erweiterungsprinzips immer in Fuzzy-Werte abbilden, kann nur eine lineare Funktion mit Y-Fehlerumgebungen als Bildwerten berücksichtigt werden. Diese sollen möglichst gut an die reellwertigen y_i angenähert werden. Wir approximieren entsprechend die folgenden Fuzzy-lineare Funktionen:

$$\widetilde{Y}(y) = a_0 \oplus a_1 \odot \widetilde{X}.$$

Die Notation $\widetilde{Y}(y)$ soll verdeutlichen, dass der Zusammenhang nur zu unbekannten, „virtuellen“ Fehlerumgebungen der genauen Beobachtungen y_i besteht, auch wenn die Approximation die Abstände zu den Eingangswerten minimiert.

Die Lösung der Kleinste Quadrate Fuzzy-Regression kann für die Metrik δ_2 in geschlossener Form berechnet werden. Dabei ist wegen der eingeschränkten Linearität der Auswertungsfunktionale $\chi_{\widetilde{X}_1,\ldots,\widetilde{X}_n}$ auf $\mathbb{R}$ zwischen dem Fall mit positiven und mit negativen Steigungsparameter a_1 zu unterscheiden, wie in Abschnitt *3.1.2* dargestellt wurde.

1. Fall: Für $a_1 \geq 0$ gilt mit Satz A.5 in Anhang *A.1*, dass

$$\begin{aligned}\sum_i \delta_2^2(y_i, a_0 \oplus a_1 \odot \widetilde{X}_i) &= \sum_i (y_i - a_0 - a_1\sigma_{\widetilde{X}_i})^2 + \sum_i \delta_2^2(0, a_1\widetilde{X}_i^\circ) \\ &= \sum_i (y_i - a_0 - a_1\sigma_{\widetilde{X}_i})^2 + a_1^2 \sum_i \delta_2^2(0, \widetilde{X}_i^\circ),\end{aligned} \tag{5.6}$$

wobei $\widetilde{X}_i^\circ = \widetilde{X}_i \ominus \sigma_{\widetilde{X}_i}$, die im Steiner-Punkt $\sigma_{\widetilde{X}_i}$ zentrierte Fuzzy-Menge ist.

Wir setzen nun $q = \sum_i \delta_2^2(0, \widetilde{X}_i^\circ)$. Offensichtlich gilt $q \geq 0$.

Aus (5.6) erhalten wir die folgenden Normalengleichungen:

$$\partial_{a_0} = -2n(\bar{y}_\cdot - a_0 - a_1\bar{\sigma}_{\widetilde{X}\cdot}), \tag{5.7}$$

$$\partial_{a_1} = \sum_i 2(y_i - a_0 - a_1\sigma_{\widetilde{X}_i})(-\sigma_{\widetilde{X}_i}) + 2qa_1. \tag{5.8}$$

Setzen wir nun ∂_{a_0} gleich 0, so folgt

$$a_0 = \bar{y}_\cdot - a_1\bar{\sigma}_{\widetilde{X}\cdot}.$$

Eingesetzt in (5.8) ergibt sich daraus

$$\partial_{a_1} = 2[(q + \sum_i(\sigma_{\widetilde{X}_i} - \overline{\sigma_{\widetilde{X}\cdot}})^2)a_1 - \sum_i(y_i - \bar{y}_\cdot)(\sigma_{\widetilde{X}_i} - \overline{\sigma_{\widetilde{X}\cdot}})]. \tag{5.9}$$

Insgesamt erhalten wir folglich als Fuzzy Kleinste Quadrate-Lösung unter der Bedingung, dass $a_1 \geq 0$:

$$\hat{a}_1 = \frac{\sum_i(y_i - \bar{y}_\cdot)(\sigma_{\widetilde{X}_i} - \overline{\sigma_{\widetilde{X}\cdot}})}{q + \sum_i(\sigma_{\widetilde{X}_i} - \overline{\sigma_{\widetilde{X}\cdot}})^2} = \frac{\frac{1}{n}\sum_i(y_i - \bar{y}_\cdot)(\sigma_{\widetilde{X}_i} - \overline{\sigma_{\widetilde{X}\cdot}})}{\frac{q}{n} + \frac{1}{n}\sum_i(\sigma_{\widetilde{X}_i} - \overline{\sigma_{\widetilde{X}\cdot}})^2} \tag{5.10}$$

und

$$\hat{a}_0 = \bar{y}_\cdot - \hat{a}_1\overline{\sigma_{\widetilde{X}\cdot}}. \tag{5.11}$$

2. Fall: Sei $a_1 < 0$.

Dann erhalten wir mit Satz A.8 auf S. 221, dass für alle $1 \leq i \leq n$ gilt

$$\delta_2^2(0, (a_1\widetilde{X}_i)^\circ) = a_1^2\,\delta_2^2(0, (-\widetilde{X}_i)^\circ) = a_1^2\,\delta_2^2(0, \widetilde{X}_i^\circ).$$

Für trianguläre Fuzzy-Zahlen $\widetilde{A}$ gilt darüber hinaus auch, dass $\sigma_{-\widetilde{A}} = -\sigma_{\widetilde{A}}$. Durch Ausrechnen erhalten wir daher, dass die Approximationsparameter ebenfalls mit den Formeln (5.10) und (5.11) bestimmt werden können. Die Approximationsparameter $\hat{a}_1, \hat{a}_2$ können also bei der defuzzifizierten Regression mit ungenauen Inputs nicht durch eine klassische Kleinste Quadrate-Approximation in den Steiner-Punkten bestimmt werden, wie im Fall mit genauen Inputs und Fuzzy-Outputs in Gleichung (5.4). Stattdessen verringert sich die Steigung mit steigender Fuzziness der Inputs, die im Wert q gemessen wird.

Mit dieser Berechnung können wir zeigen, dass der Steigungsparameter bei Fuzzy-Daten $(\widetilde{X}_i, y_i)$ mit triangulären $\widetilde{X}_i$ im Betrag immer kleiner oder gleich dem Kleinste Quadrate-Wert ohne Nebenbedingungen ist. Diamond und Körner [1997] geben die folgenden Formel

zur Berechnung der optimalen Steigungsparameter ohne Nebenbedingungen an[16]

$$\hat{\hat{a}}_1 = \frac{C_{XY}}{C_{XX}} \tag{5.12}$$

wobei

$$\begin{aligned} C_{XX} &= c(x,x) + \|R\|_2^2 c(r_X, r_X) + \|L\|_2^2 c(l_X, l_X) + 2[\|R\|_1 c(x, r_X) - \|L\|_1 c(x, l_X)] \\ C_{XY} &= c(x,y) + \|R\|_2^2 c(r_X, r_Y) + \|L\|_2^2 c(l_X, l_Y) \\ &\quad + \|R\|_1 [c(x, r_Y) + c(r_X, y)] - \|L\|_1 [c(x, l_Y) + c(l_X, y)] \end{aligned}$$

wobei $c(\cdot,\cdot)$ definiert ist wie auf S. 164. Bei triangulären Fuzzy-Mengen ist $\|R\|_2^2 = \|L\|_2^2 = \frac{1}{6}$ und $\|R\|_1 = \|L\|_1 = \frac{1}{4}$.

Damit erhalten wir, dass für reellwertige y_i gilt

$$C_{XY} = \frac{1}{n} \sum_i (y_i - \bar{y}_\cdot)(\sigma_{\widetilde{X}_i} - \overline{\sigma_{\widetilde{X}}}_\cdot)$$

und

$$C_{XX} = \Big[\frac{q}{n} + \frac{1}{n} \sum_i (\sigma_{\widetilde{X}_i} - \overline{\sigma_{\widetilde{X}}}_\cdot)^2\Big] - \Big[\frac{5}{48}\overline{r_X}_\cdot^2 + \frac{1}{8}\overline{r_X}_\cdot \overline{l_X}_\cdot + \frac{5}{48}\overline{l_X}_\cdot^2\Big].$$

und folglich $|\hat{a}_1| \leq |\hat{\hat{a}}_1|$.

[16] Vgl. Diamond und Körner 1997, S. 18ff.

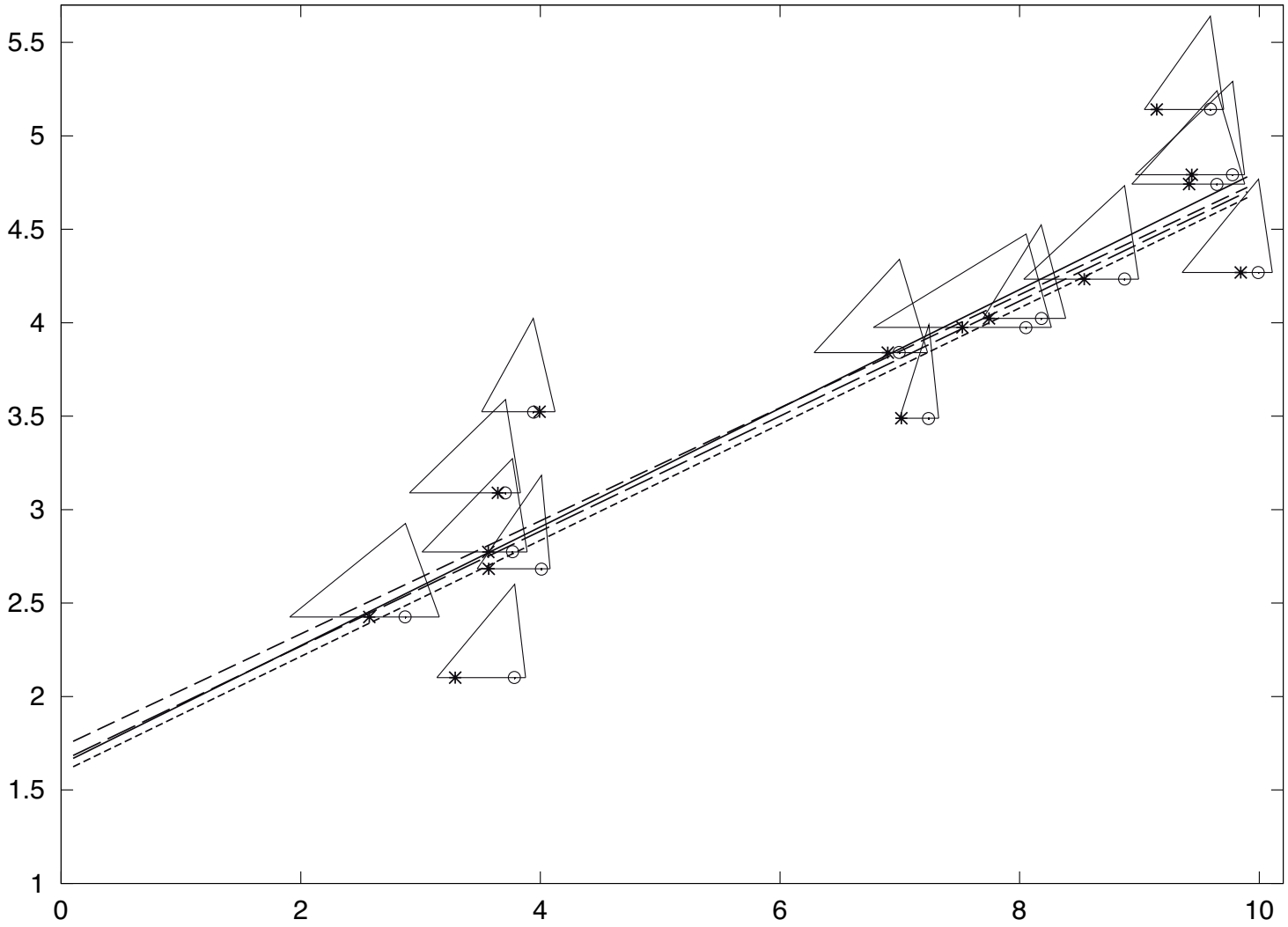


Abbildung 5.8: Modell mit genauen Outputs und einfachem systematischen Fehler ($s^2 = 0.3$)

Legende: „wahre“ Werte (*), fehlerhafte Beobachtungen (∘), „wahre“ Gerade (——), Modalwerteapproximation (----), Endwerteapproximation (– –), Fuzzy-Approximation (— —)

	„wahre“ Werte	Modalwerte	Endwerte	Fuzzy-Daten
a_0	1.6377	1.5928	1.7304	1.6542
a_1	0.3175	0.3107	0.3024	0.3078

Tabelle 5.4: Approximationsparameter des Datenbeispiels mit genauen Outputs und einfachem systematischen Fehler

In Abbildung 5.8 ist ein Fallbeispiel mit genauen Outputs und einfachem systematischen Fehler in den Inputs zu sehen. Der „ex post“ bestimmte systematische Fehler beträgt $\psi_\cdot = \frac{1}{n}\sum_i (x_i - x_i^*) = 0.28361$. Die Parameter der Benchmark-Approximationen sind in Tabelle 5.4 zusammengefasst. Dabei nähert die Modalwerteapproximation den „wahren“ Steigungsparameter am besten an, ist aber im Vergleich zu den anderen Benchmark-Geraden am stärksten nach unten verschoben. Das entspricht im wesentlichen dem Effekt einer konstanten Verschiebung aller X-Werte um den Betrag des systematischen Fehlers nach rechts. Die Endwerteapproximation hat den am weitesten abweichenden Steigungsparameter, der am kleinsten ausfällt, wie es aufgrund der Attenuation bei variierenden Fehlereinflüssen — in diesem Fall dem linken und rechten Rand der Trägermenge — zu erwarten ist, die in Abschnitt *2.2.2* beschrieben ist. Die Fuzzy-Approximationsgerade liegt zwischen Modalwerte- und Endwerteapproximation. Durch einen Wechsel zu den alternativen Fuzzy-Metriken kann in diesem Fehlerszenario eine Verbesserung der Steigungsparameter erreicht werden. Dies ist damit zu erklären, dass beide ein größeres Gewicht auf die Modalwerte der Fuzzy-Daten legen und daher die Approximationsfunktion stärker zur Modalwerteapproximation ziehen. Im Vergleich zu den Datensätzen ohne zufällige Störungen fallen die Fehlereffekte bei Datensätzen mit zufälligen Störungen in den „wahren“ Werten anscheinend in etwas abgeschwächter Form aus.

Bei manchen Datenbeispielen, die für dieses Fehlerszenario erzeugt wurden, fällt abweichend zu hier die Modalwerteapproximation steiler aus als die „wahre“ Gerade,[17] dennoch liefert der Steigungsparameter der Modalwerteapproximation in der Regel die beste Annäherung an die „wahre“ Steigung. Insgesamt wirken sich bei der Approximation der Benchmark-Geraden sowohl der „konstante“ systematische Fehler $\psi_\cdot$ als auch die variierenden Fehler aus, die $\psi_\cdot$ überlagern. Dies lässt sich anhand der klassischen Modelle mit Fehler in den Variablen, die in Abschnitt *2.2.2* dargestellt wurden, gut erklären: der Anteil des additiven, deterministischen Fehlers, der in etwa $\psi_\cdot$ entspricht, führt zu einer negativen Verschiebung des Achsenabschnitts im Vergleich zur „wahren“ Approximationsgeraden und die variierenden Fehleranteile — die aufgrund der Konstruktion in der Summe verschwinden — führen bei der positiven Steigung des Zusammenhangs zu einer

[17] Dies ist auf Abhängigkeiten zwischen den Fehlereinflüssen mit den „wahren“ Werten zurückzuführen. Etwa trifft eine steilere Verzerrung bei der Modalwertgeraden häufig mit einer negativen deskriptiven Kovarianz der Fehlerschwankungen mit den „wahren“ Werten zusammen. Dies ist aber nicht hinreichend.

Verzerrung des Steigungsparameters nach unten. Die Fuzzy-Approximation bildet hierbei einen Kompromiss zwischen der Modalwerte- und der Endwerteapproximation und reagiert robuster auf eine Veränderung in der Modellierung der Fehlerintervalle bzw. der Modalwerte.

Aus der Approximationsfunktion kann rückwirkend eine virtuelle Fehlerumgebung für die y_i berechnet werden, die ausdrückt, welche Fuzziness des Datenpunktes zum selben Approximationsergebnis geführt hätte. Seien dazu $\hat{a}_0, \hat{a}_1$ die Lösungsparameter der Approximation. Wir unterstellen, dass die virtuelle Fehlerumgebung $\widetilde{Y}_i^R$ von y_i die folgende Gestalt hat

$$\widetilde{Y}_i^R(y_i) = \hat{a}_0 \oplus \hat{a}_1 \widetilde{X}_i \oplus h_{y_i}.$$

Die virtuelle Fehlerumgebung soll das Residuum in y_i reproduzieren, d.h. es soll gelten

$$\delta_2^2(y_i, \hat{a}_0 \oplus \hat{a}_1 \widetilde{X}_i) = \delta_2^2(\widetilde{Y}_i^R(y_i), \hat{a}_0 \oplus \hat{a}_1 \widetilde{X}_i).$$

Dies ist dann der Fall, wenn gilt

$$h_{y_i}^2 = \delta_2^2(y_i, \hat{a}_0 \oplus \hat{a}_1 \widetilde{X}_i).$$

Das Vorzeichen von h_{y_i} kann gemäß einer passenden Ordnungsrelation auf Fuzzy-Mengen bestimmt werden. Wir wählen das Vorzeichen entsprechend dem Vorzeichen von $y_i - \sigma_{\hat{a}_0 \oplus \hat{a}_1 \widetilde{X}_i}$, da der Steiner-Punkt einer Fuzzy-Menge $\widetilde{A}$ die Eigenschaft hat, dass er den Abstand zu $\widetilde{A}$ minimiert, vgl. Satz A.8.

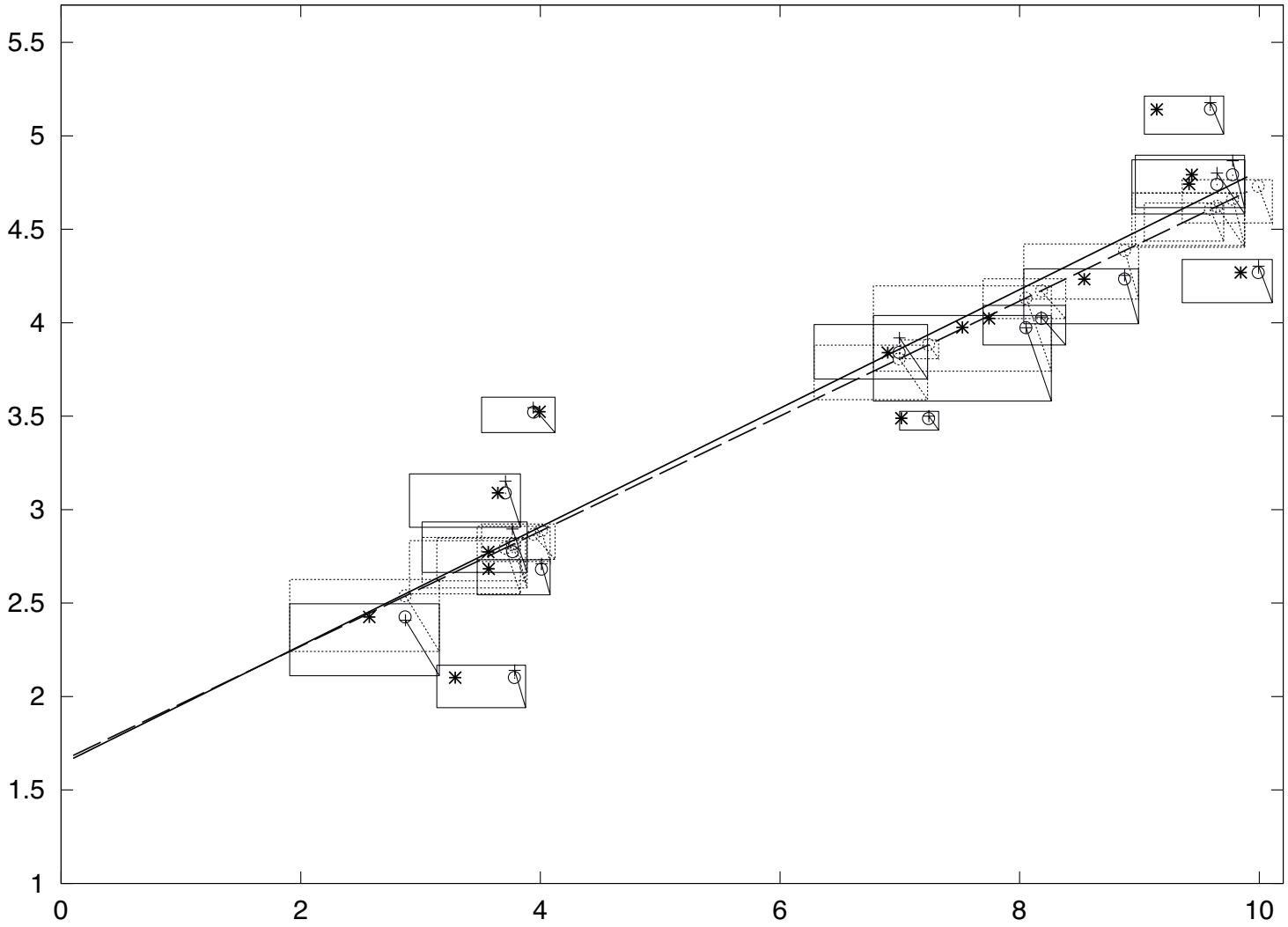


Abbildung 5.9: Bildwerte der Fuzzy-Approximation und die projizierten virtuellen Fehlerumgebungen des Datenbeispiels mit einfachem systematischen Fehler

Legende: „wahre“ Werte (*), fehlerhafte Beobachtungen (∘), Modalwerte der virtuellen Fehlerumgebungen (+), virtuelle Fehlerumgebungen (—), Bildwerte der Fuzzy-Approximation (.....), „wahre“ Gerade (__), Fuzzy-Approximation (_ _ _)

Die Fuzzy-Approximationsgerade ist, anders als bei den bekannten reellwertigen Funktionen, nicht identisch mit dem Funktionsgraphen, sondern repräsentiert die Abbildungsvorschrift für beliebige Fuzzy-Mengen aus $F_{coc}^{nob}(\mathbb{R})$ als Inputs durch eine Fuzzy-Relation. Abbildung 5.9 zeigt die Fuzzy-Approximationswerte in den $\widetilde{X}_i$, in denen die Genauigkeit der y_i verschwimmt. Außerdem sind dort die virtuellen Fehlerumgebungen in y_i eingezeichnet. Da die Modalwerte der virtuellen Fehlerumgebungen im allgemeinen nicht mit y_i übereinstimmen sind diese gesondert hervorgehoben, sie kennzeichnen gemäß dem Berechnungsverfahren gerade den Residualabstand von der Approximationsgeraden. Die Bildwerte $\hat{a}_0 \oplus \hat{a}_1 \odot \widetilde{X}_i$ überdecken in allen Punkten die „wahre“ Gerade. Der „wahre“ Wert y_i^* ist im allgemeinen aber nicht Element des zugehörigen Fuzzy-Approximationswertes. Darüber hinaus ist auch die Projektion des „wahren“ Wertes auf den „wahren“ Zusammenhang nicht notwendig im Fuzzy-Approximationswert enthalten.[18]

[18] Dies ist u.a. auch abhängig von der Qualität der Fehlerbeschreibung durch die Fuzzy-Daten, sowohl global als auch im betreffenden individuellen Fuzzy-Merkmalswert.

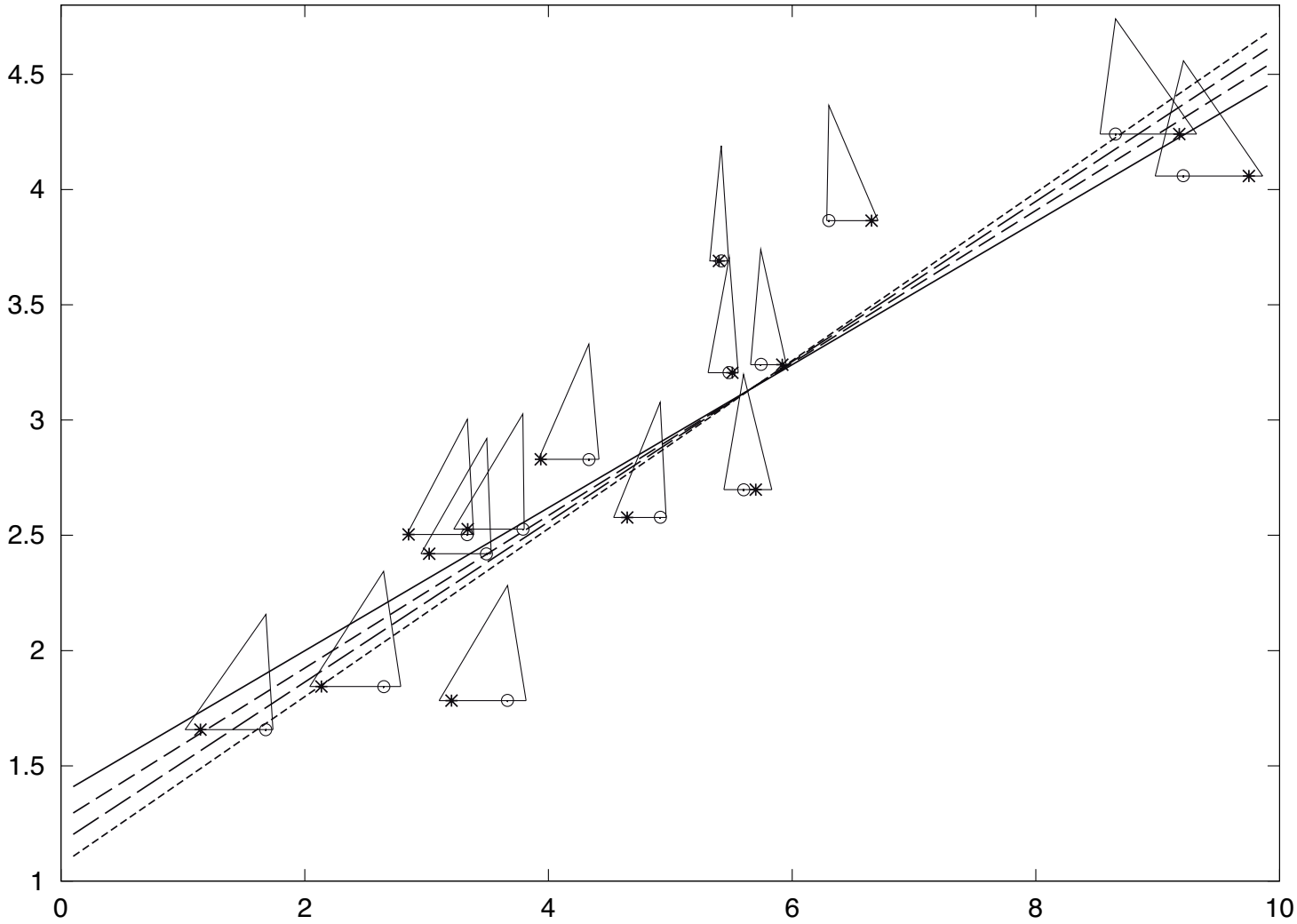


Abbildung 5.10: Modell mit genauen Outputs und korreliertem Fehler ($s^2 = 0.3$)

Legende: „wahre“ Werte (*), fehlerhafte Beobachtungen (∘), „wahre“ Gerade (——), Modalwerteapproximation (----), Endwerteapproximation (– –), Modalwertgerade der Fuzzy-Approximation (— —)

	„wahre“ Werte	Modalwerte	Endwerte	Fuzzy-Daten
a_0	1.3788	1.0715	1.2624	1.1678
a_1	0.3102	0.3644	0.3309	0.3476

Tabelle 5.5: Approximationsparameter des Datenbeispiels mit korrelierten Fehlern

In Abbildung 5.10 sind die Ergebnisse der Fuzzy-Approximation bei Fuzzy-Inputs mit korrelierten Fehlern dargestellt. Hier hat umgekehrt die Modalwerteapproximation die größte Steigung, die am weitesten von der „wahren“ Steigung abweicht, und die Endwerteapproximation nähert den „wahren“ Zusammenhang am besten an.[19] Die Güte der Endwerteapproximation ist vermutlich dadurch zu erklären, dass die Spannweite, in der der „wahre“ Wert liegt im Vergleich zur gegenüberliegenden Spannweite lang ist, so dass die Kombination der Endpunkte in Richtung einer Korrektur zieht.[20] Die Fuzzy-Approximation liegt zwischen den reellwertigen Approximationen der fehlerhaften Werte. Mit diesem Beispiel lässt sich gut illustrieren, dass Modalwerte- und Endpunkteapproximation im univariaten Fall tatsächlich extreme Fälle markieren, zwischen denen die Kleinste Quadrate Fuzzy-Approximation einen Kompromiss bildet. Die Übersicht der Approximationsparameter aus Abbildung 5.10 in Tabelle 5.5 verdeutlicht diesen Zusammenhang.

[19] Das Bild ist typisch für das Modell mit korrelierten Fehlern in X.

[20] Leider sind die vorgeführten Beispiele nicht so gut geeignet, um zu zeigen, das dieser Effekt nicht wesentlich durch die Attenuation induziert ist. Dies ist dem Interesse geschuldet, die Bespiele so auszuwählen, dass die Effekte sich bei der Zusammenführung der Betrachtungen in Abschnitt *5.2.3* für Daten mit fehlerhaften Inputs und Outputs gegenseitig nicht kompensieren. Tatsächlich ist die Steigung der Endwerteapproximation aber auch in solchen Fehlerszenarien am besten, in denen eine negative Korrelation der Beobachtungsfehler mit den „wahren“ Werten vorliegt, d.h. wenn in Gleichung (5.1) gewählt wird $c_x = 0.4 > 0$. Dort führt die Endwerteapproximation entgegen der allgemeinen Attenuation zu einer Erhöhung der Steigung.

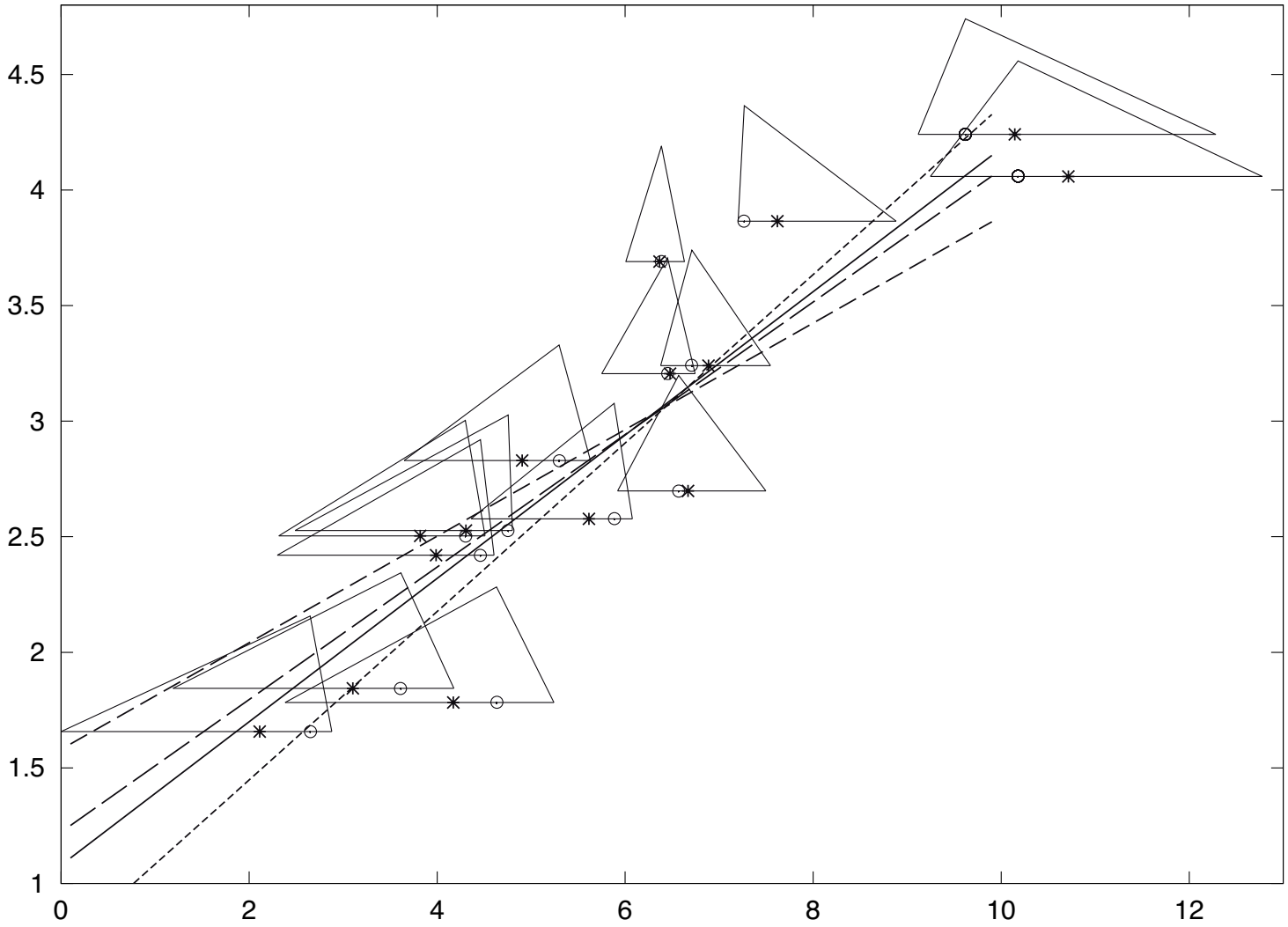


Abbildung 5.11: Modell mit genauen Outputs und korreliertem Fehler bei Überstreckung der Spannweiten um den Faktor 4

Legende: „wahre“ Werte (*), fehlerhafte Beobachtungen (∘), „wahre“ Gerade (——), Modalwerteapproximation (----), Endwerteapproximation (– –), Modalwertgerade der Fuzzy-Approximation (— —)

	„wahre“ Werte	Modalwerte	Endwerte	Fuzzy-Daten
a_0	1.3788	1.0715	1.8033	1.4991
a_1	0.3102	0.3644	0.2305	0.2867

Tabelle 5.6: Approximationsparameter des Datenbeispiels mit genauen Outputs und korrelierten Fehlern und bei Überstreckung der Spannweiten um den Faktor 4

Ein interessanter Effekt entsteht, wenn die linken und rechten Spannweiten des Datensatzes um den Faktor 4 verlängert werden, um einen Datensatz mit sehr ungenauen Daten zu erzeugen. Wenn also die folgenden Transformationen durchgeführt werden

$$l'_{X_i} = 4\, l_{X_i} \quad \text{sowie} \quad r'_{X_i} = 4\, r_{X_i}.$$

Wie Abbildung 5.11 zeigt, ist in diesem Fall wieder die Endwerteapproximation am ungünstigsten. Hingegen kommt der Steigungsparameter der Fuzzy-Approximation dem des „wahren“ Zusammenhangs am nächsten. Die Modalwerteapproximation ändert sich aufgrund der Konstruktion gegenüber dem nicht-transformierten Datensatz nicht. Offenbar gelingt es hier durch die Fuzzy-Daten den Verzerrungseffekt der Modalwerte durch die Breite und die Gewichtung durch die Spannweiten auszugleichen, wobei gleichzeitig der erhöhte Effekt der Attenuation durch die starke Wertung der Modalwerte unterbunden wird. Insbesondere führt die Datentransformation im Bezug auf die Qualität der Fuzzy-Modellierung dazu, dass der „wahre“ Wert relativ näher an den Modalwert rückt. Die Approximationsparameter werden in Tabelle 5.6 dargestellt. Das Approximationsergebnis für die Fuzzy-Daten kann bei diesem Datenbeispiel bei Verwendung von δ_H so verbessert werden, dass die Fuzzy-Approximation nahezu mit der „wahren“ Geraden zusammenfällt: $a_0^H = 1.0594$, $a_1^H = 0.3122$. Die Verwendung von δ_G führt hingegen zu einem schlechteren Ergebnis. Wird ein einzelner Ausreißer dadurch erzeugt, dass nur bei einem Wert die Spannweiten verändert wird, dann reagiert die Endwerteapproximation stärker als die Fuzzy-Approximation.

Insgesamt zeigen die untersuchten Fallbeispiele, dass die Kleinste Quadrate Fuzzy-Regression, mit der bei genauen Outputs ausschließlich ein defuzzifizierter Zusammenhang angepasst werden kann, unter den Benchmarkvarianten nur im Ausnahmefall das beste Ergebnis liefert. Gleichzeitig wird deutlich, dass die Modalwerteapproximation und die Endwerteapproximation für die bisher betrachteten Fehlerszenarien tatsächlich geeignete Benchmarks darstellen, die das Variationsspektrum der Fuzzy-Regression markieren. Dabei nimmt die Fuzzy-Regression die Funktion eines Kompromisses an, der eine Mischung zwischen den extremen Ansätzen darstellt. Insbesondere weist die Fuzzy-Regression eine größere Robustheit gegenüber Änderungen in der Fehlermodellierung auf, dies allerdings zu dem Preis, dass die alternativen Methoden im Einzelfall zu einer besseren Lösung kom-

men. Zudem ist der Korrektureffekt einer Fuzzy-Regression umso geringer, je stärker die Fehler in den Daten von einem deterministischen Anteil dominiert sind.

5.2.3 Modell mit fehlerhaften In- und Outputs

Wenn unter den Inputs die Messwerte für ein Merkmal nur ungenau vorliegen und auch die Outputs ungenau vorliegen, sind Fuzzy-Daten $(\widetilde{X}_i, \widetilde{Y}_i)$ zu berücksichtigen. Der lineare Zusammenhang bildet dann ab, wie die Fehlerumgebungen der exogenen Merkmale mit den Fehlerumgebungen der endogenen Merkmale zusammenhängen. Da es sich hierbei um die Zusammenführung der beiden vorhergehenden Modelle handelt, erhalten wir zunächst allgemein den Funktionsansatz:

$$\widetilde{Y} = \widetilde{A}_0 \oplus \widetilde{A}_1 \odot \widetilde{X}.$$

Eine technische Schwierigkeit besteht allerdings darin, dass das Produkt $\widetilde{A}_1 \odot \widetilde{X}$ keine LR-Fuzzy-Menge mehr ist und die Metrik daher alternativ berechnet werden muss. Darüber hinaus ist diese Berechnung erheblich erschwert, wenn der Support eines Steigungsparameters 0 enthält, d.h. $0 \in \widetilde{A}_1$, weil dann gemäß Satz 2.19 die Multiplikation für die α-Niveaus unterschiedlich zu berechnen ist, je nachdem ob $a^L(\alpha) \geq 0$, $a^L(\alpha) < 0 < a^R(\alpha)$ oder $a^R(\alpha) < 0$.[21] Fuzzy-Steigungsparameter sind vor allem dann angebracht, wenn die Ungenauigkeitsumgebungen der $\widetilde{Y}_i$ in i tendenziell linear wachsen. Bei vielen Datensätzen ist ein solches Wachstum in der Ungenauigkeit aber vernachlässigbar. Auch bei den simulierten Daten mit korrelierten Fehlern wie in Abbildung 5.6 liegt kein eindeutiges monotones Wachstum der Datenungenauigkeiten in Y vor, so dass auch in einer solchen Situation ein Ansatz mit reellwertigen Steigungsparametern begründet werden kann.

Wir beschränken uns hier daher auf den vereinfachten Ansatz[22]

$$\widetilde{Y} = \widetilde{A}_0 \oplus a_1 \odot \widetilde{X}.$$

Daneben kann hier ähnlich wie in Abschnitt *5.2.1* auch eine vollständig defuzzifizierte

[21] In der Literatur wird verschiedentlich diskutiert, ob ein Steigungsparameter mit Element 0 überhaupt als sinnvoll anzusehen ist, da ein solcher als „ungefähr 0“ interpretiert werden kann und somit der Gleichung nur „zusätzliche Unschärfe“ hinzufüge. Insbesondere bei den metrischen Ansätzen ist diese Argumentation m.E. nicht stichhaltig, da es hier um die bestmögliche Approximation der Abbildungsrelation zwischen X und Y geht.

[22] Die vereinfachte Fuzzy-Modellfunktion ist in Abschnitt *3.3.2*, S. 111 definiert. Gemäß Bemerkung 3.7 kann $\widetilde{A}_0$ als zusätzliche Grundunschärfe gegenüber einer einfachen linearen Abbildung von Fuzzy-Werten interpretiert werden.

Modellfunktion angepasst werden, d.h.

$$\widetilde{Y} = a_0 \oplus a_1 \odot \widetilde{X}.$$

Als wichtige Benchmarks werden zudem die Kleinste Quadrate-Regression in den Steiner-Punkten der Fuzzy-Daten $(\sigma_{\widetilde{X}_i}, \sigma_{\widetilde{Y}_i})$ sowie Fuzzy-Regressionen in den Randdaten $(\xi_i, \widetilde{Y}_i)$ bzw. $(\widetilde{X}_i, \zeta_i)$ betrachtet, wobei $\xi_i \in \mathrm{supp}(\widetilde{X}_i)$ und $\zeta_i \in \mathrm{supp}(\widetilde{Y}_i)$. Als markante Punkte für ξ_i und ζ_i werden im Folgenden die Steiner-Punkte herausgegriffen. Die Steiner-Punkte können aus den Fuzzy-Daten abgeleitet werden und zudem rücken sie die Approximationen auf das Lageniveau der Fuzzy-Approximation, wie in den isolierten Betrachtungen in den Abschnitten *5.2.1* und *5.2.2* herausgearbeitet wurde.

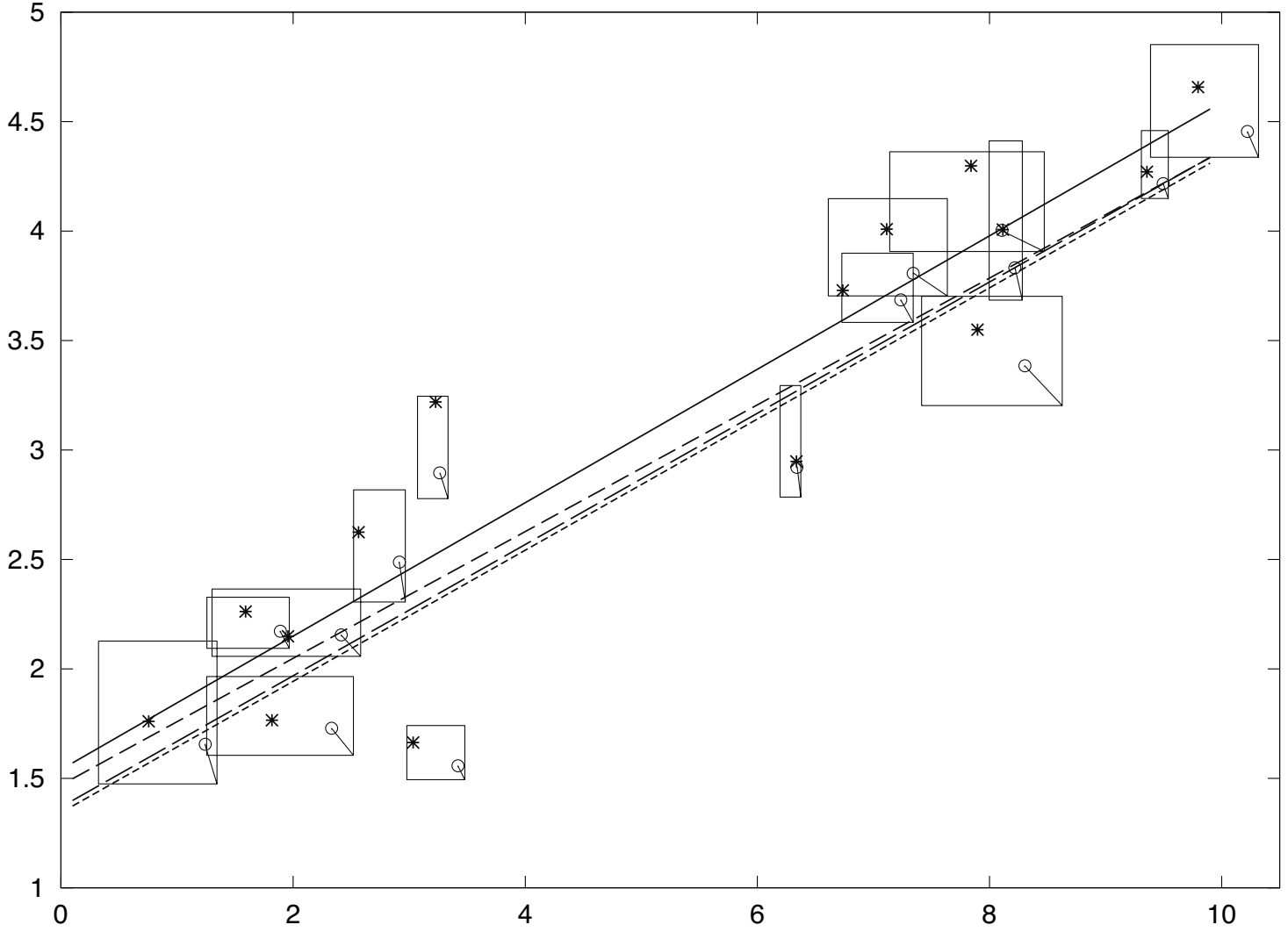


Abbildung 5.12: Modell mit Fuzzy-Daten und einfachem systematischen Fehler: Benchmarking

Legende: „wahre“ Werte (*), fehlerhafte Beobachtungen bzw. Modalwerte der Fuzzy-Daten (∘), Fuzzy-Beobachtungen (—), „wahre“ Gerade (—), Modalwerteapproximation (----), Endwerteapproximation (– –), Modalwertgerade der Fuzzy-Approximation (— —)

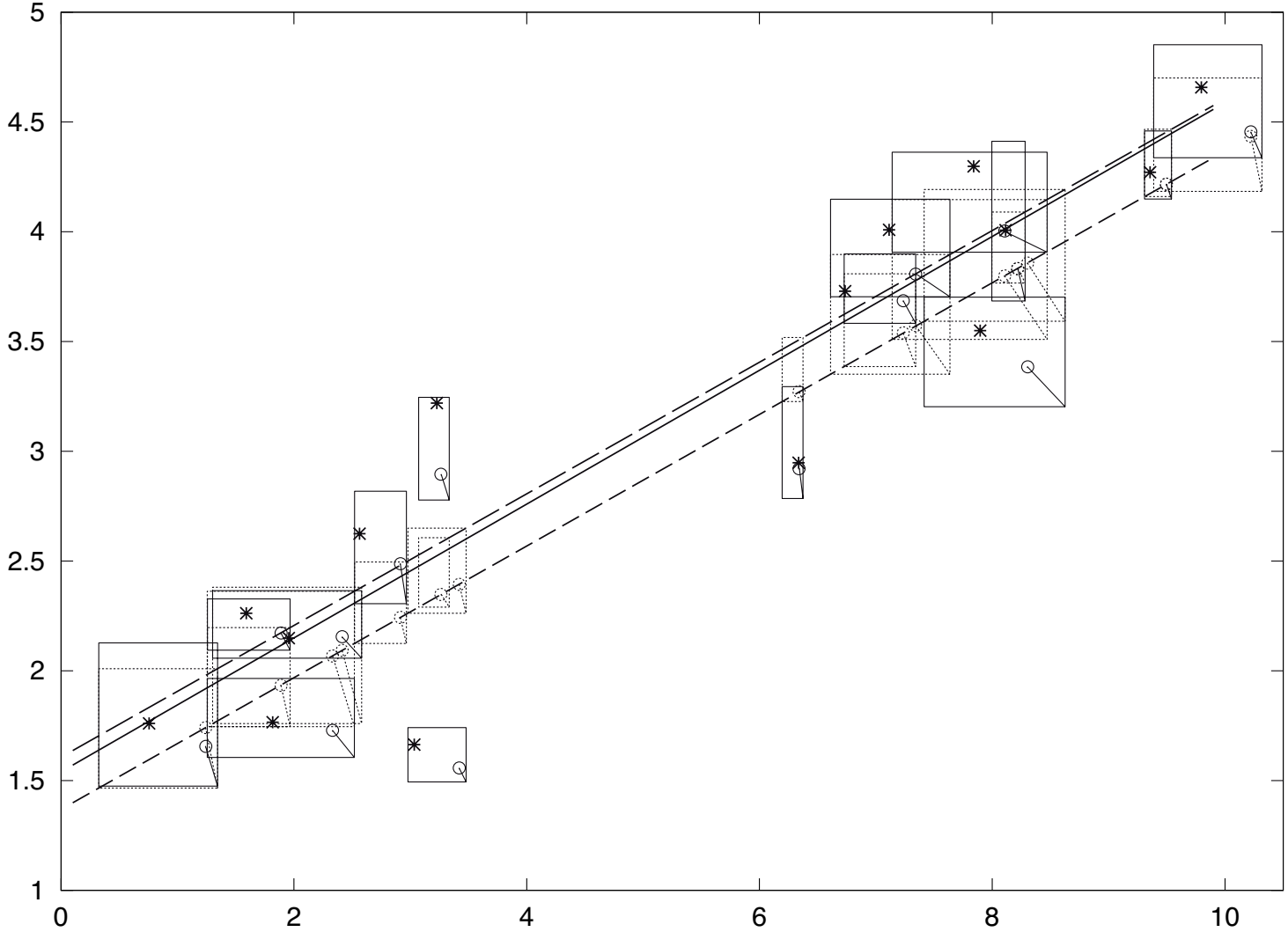


Abbildung 5.13: Modell mit Fuzzy-Daten und einfachem systematischen Fehler: Bildwerte der Fuzzy-Approximationsfunktion mit Fuzzy-Parameter $\widetilde{A}_0$

Legende: „wahre“ Werte (*), fehlerhafte Beobachtungen bzw. Modalwerte der Fuzzy-Daten (∘), Fuzzy-Messwerte (—), Bildwerte der Fuzzy-Approximation (.....), „wahre“ Gerade (—), Modalwertgerade der Fuzzy-Approximation (– –), Begrenzungsgerade des Trägers der Fuzzy-Approximation (— —)

	„wahre“ Werte	Modalwerte	Endwerte	Fuzzy-Daten
a_0	1.5397	1.3442	1.4693	$(1.3689; 0, 0.2376)_L$
a_1	0.3048	0.2996	0.2895	0.2998

Tabelle 5.7: Approximationsparameter des Datenbeispiels mit Fuzzy Daten und mit einfachem systematischen Fehler und Fuzzy $\widetilde{A}_0$ — Benchmarking

Zunächst werden wieder einige Datenbeispiele mit einfachen systematischen Fehlern in X und Y vorgeführt. Ähnlich wie auch schon in den Szenarien mit genauen Inputs bzw. genauen Outputs ist auch hier die Endwerteapproximation insgesamt die schlechteste Approximation.[23] Mit der Modalwerte- und der Fuzzy-Approximation wird eine ähnliche Steigung bestimmt, die nah an der „wahren“ Steigung liegt. Allerdings ist der Achsenabschnitt der Modalwerteapproximation deutlich niedriger als bei der „wahren“ Approximation. Hingegen wird bei der Fuzzy-Approximation ein Korridor zur rechten Seite hin aufgespannt, der die „wahre“ Gerade enthält, wie in Abbildung 5.13 ersichtlich ist. Dabei spiegelt sich in der Fuzziness von $\widetilde{A}_0$ wieder, dass die breiten linken Spannweiten der $\widetilde{X}_i$ mit schmalen linken Spannweiten der $\widetilde{Y}_i$ verknüpft sind und die schmalen rechten Spannweiten der $\widetilde{X}_i$ mit breiten rechten Spannweiten der $\widetilde{Y}_i$.

[23] Die Nähe zwischen der Endwerteapproximation und der „wahren“ Gerade ändert sich bei diesem Szenario, wenn statt der Modalwerte der $\widetilde{Y}_i$ die Steiner-Punkte, die ebenfalls zusammen mit den Fuzzy-Werten vorliegen, als repräsentative Punkte gewählt werden. Die Steigung wird dadurch aber nicht wesentlich beeinflusst.

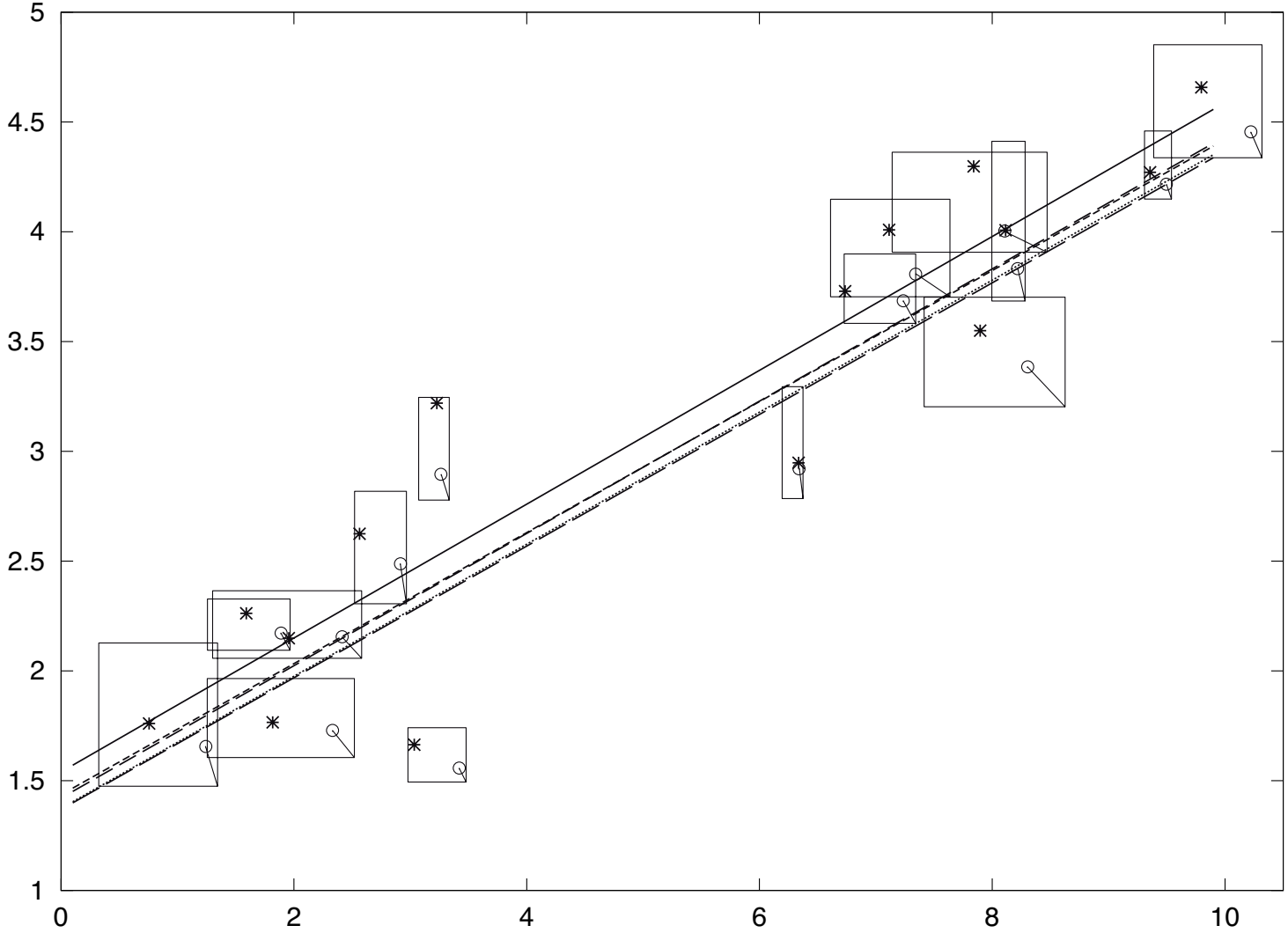


Abbildung 5.14: Modell mit Fuzzy-Daten und einfachem systematischen Fehler: Vergleich mit dem defuzzifizierten Modell und den Randapproximationen

Legende: „wahre“ Werte (*), fehlerhafte Beobachtungen bzw. Modalwerte der Fuzzy-Daten (∘), Fuzzy-Messwerte (—), „wahre“ Gerade (—), defuzzifizierte Approximation (– –), Modalwertgerade der Fuzzy-Approximation (— —), Modalwertgerade der vereinfachten Fuzzy-Approximation bei genauen Inputs $\sigma_{\widetilde{X}_i}$ (.....); Fuzzy-Approximation bei genauen Outputs $\sigma_{\widetilde{Y}_i}$ (----)

	„wahre“ Werte	Fuzzy-Modell	defuzz. Modell	genaue Inputs	genaue Outputs
a_0	1.5397	$(1.3689; 0, 0.2376)_L$	1.4209	$(1.3743; 0.1195, 0.3178)_L$	1.4358
a_1	0.3048	0.2998	0.3012	0.3007	0.2985

Tabelle 5.8: Approximationsparameter des Datenbeispiels mit Fuzzy Daten und mit einfachem systematischen Fehler — Vergleich mit dem defuzzifizierten Modell und den Randapproximationen

Im Vergleich der Fuzzy-Approximation mit der defuzzifizierten Fuzzy-Regression und den Randapproximationen mit genauen Inputs bzw. genauen Outputs kommt insgesamt die defuzzifizierte Anpassung der „wahren“ Gerade am nächsten. Hierbei unterscheiden sich die Steigungsparameter der Fuzzy-Ansätze nur graduell, wie Tabelle 5.8 zeigt. Auch die einfache Kleinste Quadrate-Approximation zu den Steiner-Punkten entspricht im wesentlichen den Vergleichsergebnissen: $a_0 = 1.4238$ und $a_1 = 0.3007$, wobei a_1 wie in Abschnitt *5.2.1* identisch mit der Steigung bei Approximation mit genauen Inputs ist. Die beiden reellwertigen Approximationsfunktionen sind nahezu deckungsgleich, ebenso sind die Modalwertgeraden der Fuzzy-Approximationsfunktionen in Abbildung 5.14 kaum zu unterscheiden. Die Berücksichtigung einer Fuzziness der Approximationsfunktion bildet den Einfluss der Modalwerte deutlicher ab und schließt die „wahre“ Gerade in der Trägermenge ein. Bei Unterstellung von reellwertigen Inputs ergibt sich im Unterschied zum Fuzzy-Modell bei Fuzzy-Inputs eine Approximation mit positiver linker und rechter Spannweite, da auch die linken Spannweiten der $\widetilde{Y}_i$ eine Zunahme der Fuzziness darstellen.

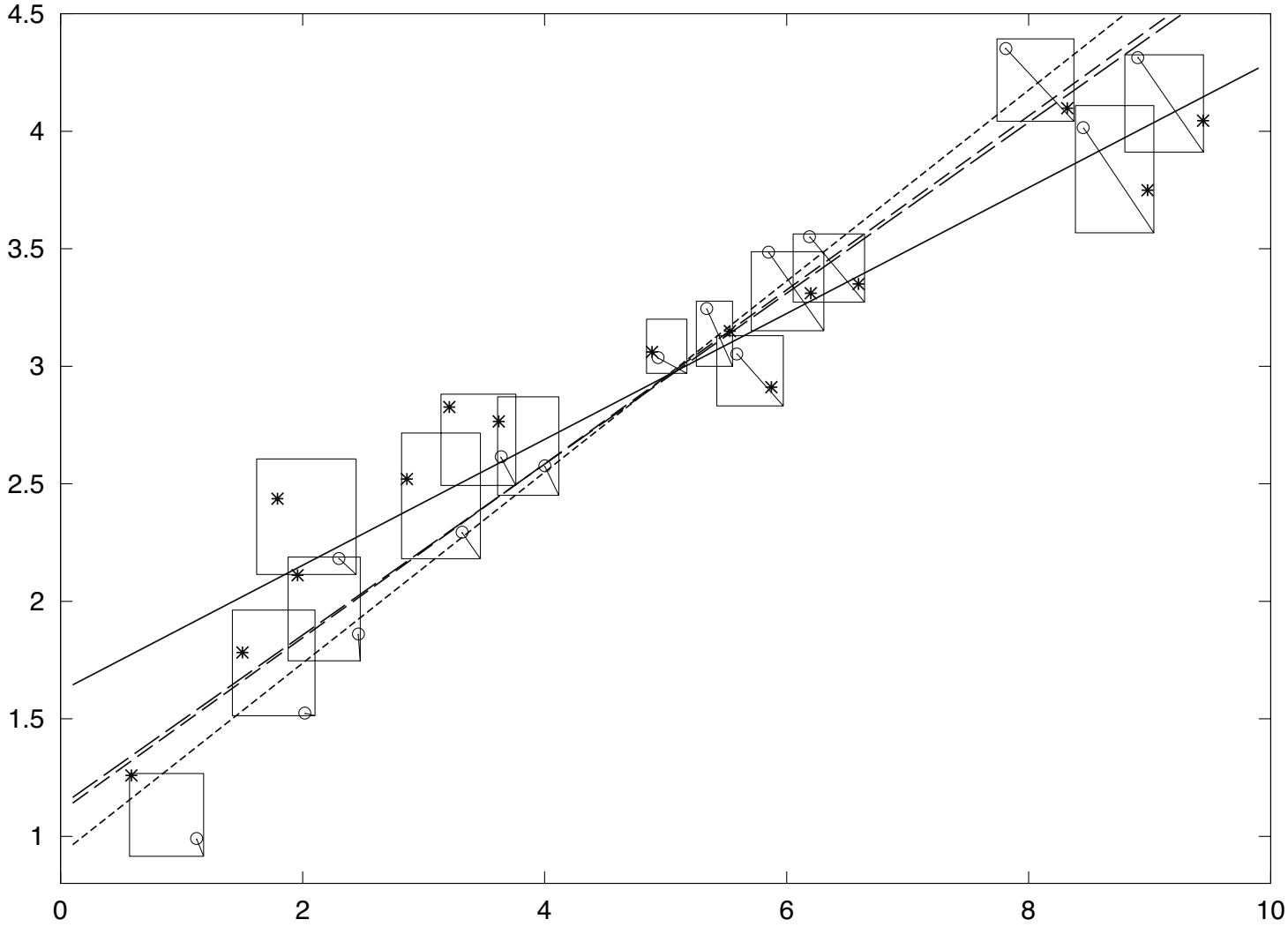


Abbildung 5.15: Modell mit Fuzzy-Daten und korreliertem Fehler: Benchmarking

Legende: „wahre“ Werte (*), fehlerhafte Beobachtungen bzw. Modalwerte der Fuzzy-Daten (∘), Fuzzy-Messwerte (—), „wahre“ Gerade (—), Modalwerteapproximation (----), Endwerteapproximation (— —), Modalwertgerade der Fuzzy-Approximation (— —)

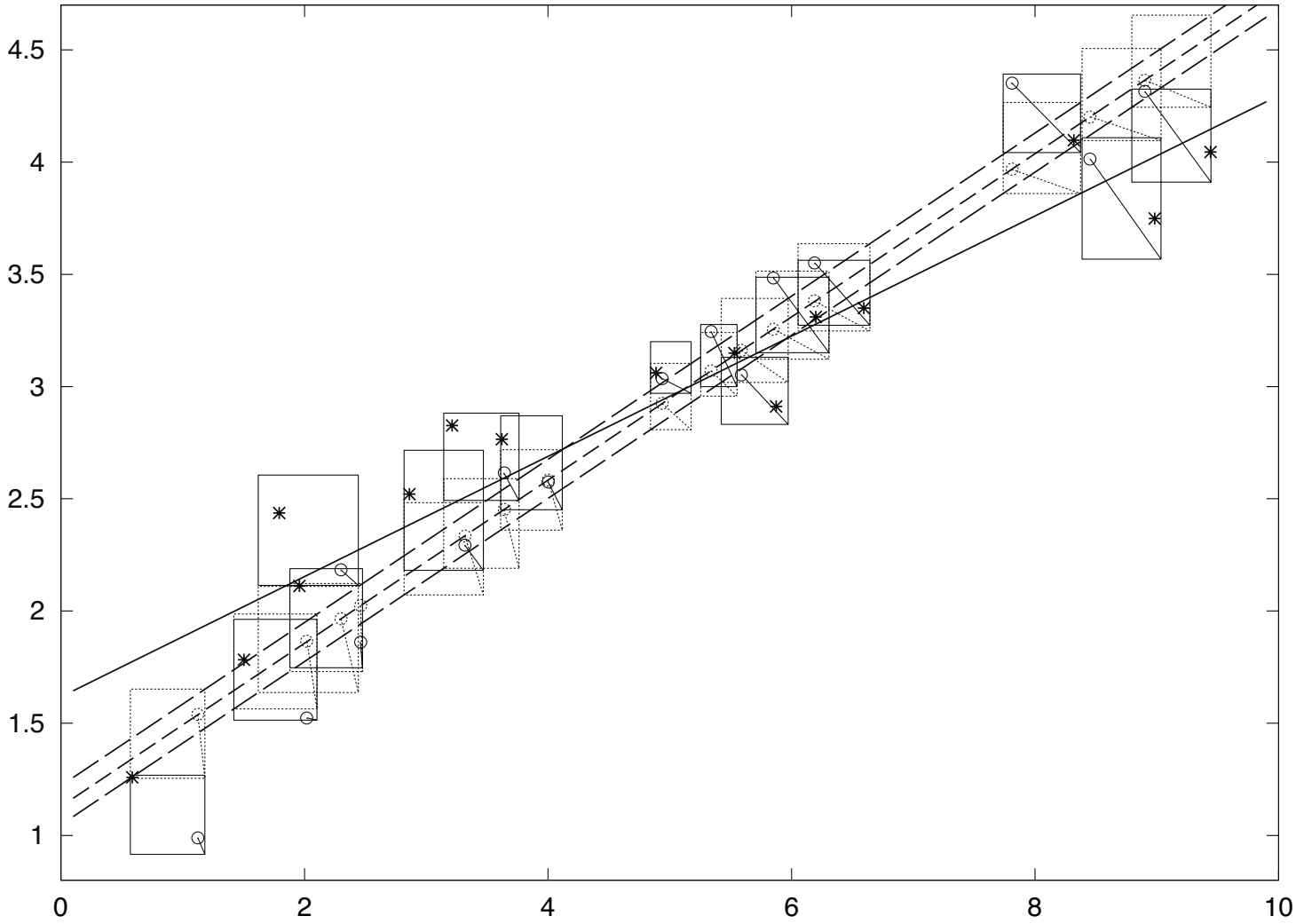


Abbildung 5.16: Modell mit Fuzzy-Daten und korreliertem Fehler: Bildwerte der Fuzzy-Approximationsfunktion mit Fuzzy-Parameter $\widetilde{A}_0$

Legende: „wahre“ Werte (*), fehlerhafte Beobachtungen bzw. Modalwerte der Fuzzy-Daten (∘), Fuzzy-Messwerte (—), Bildwerte der Fuzzy-Approximation (.....), „wahre“ Gerade (—), Modalwertgerade der Fuzzy-Approximation (– –), Begrenzungsgeraden des Trägers der Fuzzy-Approximation (— —)

	„wahre“ Werte	Modalwerte	Endwerte	Fuzzy-Daten
a_0	1.6175	0.9237	1.1048	$(1.1292; 0.0822, 0.0924)_L$
a_1	0.2678	0.4064	0.3700	0.3636

Tabelle 5.9: Approximationsparameter des Datenbeispiels mit Fuzzy Daten und mit korreliertem Fehler und Fuzzy $\widetilde{A}_0$ — Benchmarking

Im zweiten Fehlerszenario mit korreliertem Fehler in den Daten weicht die Modalwerteapproximation am stärksten von der „wahren“ Gerade ab, wie in Tabelle 5.9 zu sehen ist. Demgegenüber liegen die Modalwertgerade der Fuzzy-Approximation und Endwerteapproximation nahe beisammen, wobei als markante Punkte für die Endwerteapproximation zunächst gemäß den Vorüberlegungen in Abschnitt 5.1.2 die Modalwerte von $\widetilde{Y}_i$ gewählt wurden. Falls stattdessen Steiner-Punkte als markante Punkte der $\widetilde{Y}_i$ gewählt werden, wird die Endwerteapproximation näher an die „wahre“ Gerade gedreht, so dass sie zur besten Annäherung im Vergleich mit den Benchmarkapproximationen wird. Dies ist konsistent mit der Schlussfolgerung in Abschnitt *5.2.2*, dass im gewählten Fehlerszenario die Fehler in den Messdaten in den Steiner-Punkte gut korrigiert werden.

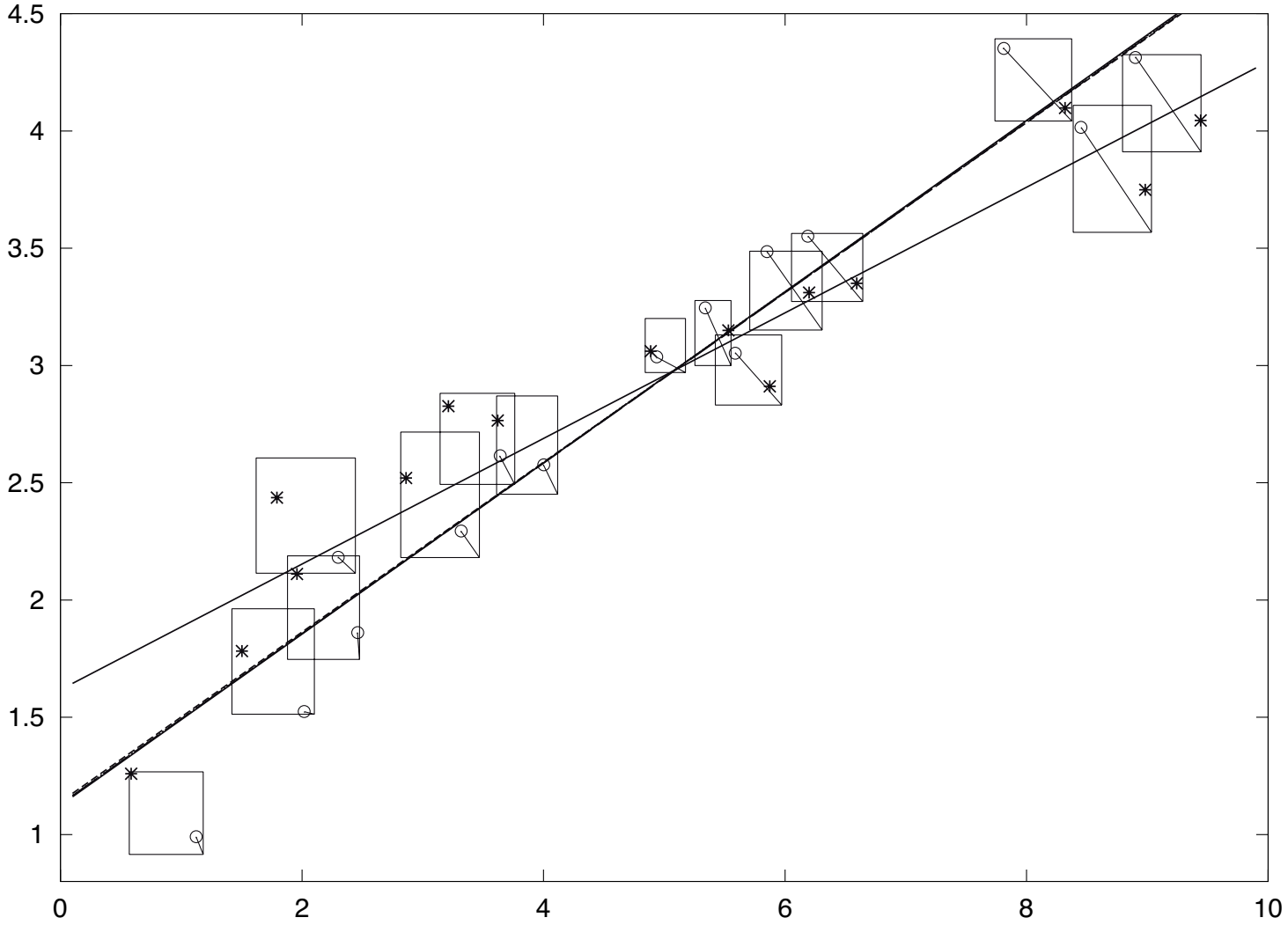


Abbildung 5.17: Modell mit Fuzzy-Daten und korreliertem Fehler: Vergleich mit dem defuzzifizierten Modell und den Randapproximationen

Legende: „wahre“ Werte (*), fehlerhafte Beobachtungen bzw. Modalwerte der Fuzzy-Daten (∘), Fuzzy-Messwerte (—), „wahre“ Gerade (—), defuzzifizierte Approximation (– –), Modalwertgerade der Fuzzy-Approximation (— —), Modalwertgerade der vereinfachten Fuzzy-Approximation bei genauen Inputs $\sigma_{\widetilde{X}_i}$ (.....); Fuzzy-Approximation bei genauen Outputs $\sigma_{\widetilde{Y}_i}$ (----)

	„wahre“ Werte	Fuzzy-Modell	defuzz. Modell	genaue Inputs	genaue Outputs
a_0	1.6175	$(1.1292; 0.0822, 0.0924)_L$	1.1248	$(1.1298; 0.1953, 0.1924)_L$	1.1391
a_1	0.2678	0.3636	0.3650	0.3641	0.3620

Tabelle 5.10: Approximationsparameter des Datenbeispiels mit Fuzzy Daten und mit korreliertem Fehler — Vergleich mit dem defuzzifizierten Modell und den Randapproximationen

Beim Vergleich mit dem defuzzifizierten Modell und den Randapproximationen lassen sich keine wesentlichen Unterschiede identifizieren. Auch hier hat die einfache Kleinste Quadrate-Approximation an die Steiner-Punkte dieselbe Steigung wie die Approximation bei genauen Inputs: $a_0 = 1.1291$ und $a_1 = 0.3641$. Die Spannweite der Fuzzy-Approximation fällt im Vergleich zum Modell mit genauen Outputs schmal aus, da die einfließende Fuzziness in den $\widetilde{X}_i$ nicht zusätzlich durch die Abbildungsvorschrift zu generieren ist (vgl. Abbildung 5.10 bzw. Tabelle 5.10 in diesem Abschnitt).

	Endwerte		Fuzzy-Daten		defuzz. Modell	
	a_0	a_1	a_0	a_1	a_0	a_1
X-Dehnung	1.6248	0.2642	1.3891	0.3116	1.3891	0.3116
Y-Dehnung	1.1048	0.3700	$(1.5096; 0.6926, 0.6915)_L$	0.2842	1.4540	0.2958
XY-Dehnung	1.6248	0.2642	$(1.6823; 0.4722, 0.4966)_L$	0.2482	1.5733	0.2724

Tabelle 5.11: Approximationsparameter des Beispieldatensatzes bei Dehnung der Spannweiten der Fuzzy-Werte um den Faktor 4

Die Veränderungen durch eine Überstreckung der Spannweiten um das 4-fache wurde wieder anhand des Fallbeispiels mit korrelierten Fehlern in Abbildung 5.15 untersucht.[24] Die Ergebnisse für die Parameterwerte sind in Tabelle 5.11 dargestellt. Wie die Zusammenstellung der Approximationsparameter für alle Überstreckungsvarianten zeigt, führt eine Dehnung der Daten in der X-Richtung, ähnlich wie bei den Modellen mit genauen Outputs, zu einer flacheren Steigung der Endwerteapproximation, die damit die „wahre“ Gerade besser approximiert, wohingegen die Fuzzy-Approximation ihre Steigung nur wenig ändert. Die defuzzifizierte Approximation ist hier identisch mit der Fuzzy-Approximation, da die X-Spannweiten die Y-Spannweiten deutlich übersteigen.

Bei Dehnung der Fuzzy-Werte in Y-Richtung ändert sich die Endwerteapproximation gegenüber dem Ausgangsdatensatz nicht. Die Modalwertgerade der Fuzzy-Approximation approximiert die „wahre“ Gerade recht gut. Dabei vergrößert sich die Spannweite um

[24] Bei den Datensätzen mit einfachen systematischen Fehlern kommt es nur zu den offensichtlichen Effekten.

etwa das 7-fache.[25] Die Approximation des defuzzifizierten Modells fällt steiler als die Fuzzy-Approximation aus. Insgesamt wirkt sich hier offenbar aus, dass die Intervallendwerte eine gute Korrektur der fehlerhaften Messwerte darstellen. Bei einer Kombination der Dehnungen in X- und Y-Richtung liegt die defuzzifizierte Fuzzy-Approximation näher an der „wahren“ Gerade als das vereinfachte Fuzzy-Modell. Die Steigung der Fuzzy-Approximation ist sogar flacher als die „wahre“ Gerade.

Durch die Überstreckung erhalten die Endwerte im Verhältnis zu den Modalwerten der Fuzzy-Daten ein höheres Gewicht. Aufgrund der Besonderheit in der Konstruktion der Fuzzy-Fehlerszenarien, dass die Modalwerte identisch mit den fehlerhaften Beobachtungen sind, wird v.a. bei den Szenarien mit korrelierten Fehler eine Verbesserung in der Approximation erreicht. Dies wird bei der Endwerteapproximation besonders deutlich. Einen Korrektureffekt für den korrelierten Fehler in den Modalwerte bei einer Überstreckung gibt es sowohl in X- als auch in Y-Richtung. Die Korrektureffekte wirken aufgrund der gewählten Konstruktion der Fehlerszenarien in dieselbe Richtung und bei einer Verlängerung in X- und Y-Richtung addieren sie sich auch zu einem gewissen Anteil.

5.3 Ergebnisse und Interpretation

Die vorgestellten Fallbeispiele zeigen, dass mit der Fuzzy-Regression in fast allen untersuchten Fällen ein günstigerer Steigungsparameter ermittelt wird als mit der Modalwerteapproximation, die die fehlerhaft gemessenen Werte ohne Korrekturen berücksichtigt. Die Endwerteapproximation als das entgegengesetzte Benchmark stellt hingegen ein eher artifizielles Beispiel dar, bei dem ausgenutzt wird, dass die Berücksichtigung der Endpunkte des Fehlerintervalls bei positiver Steigung des ursprünglichen Zusammenhangs in der Regel zu einer geringeren als der tatsächlichen Steigung führt. Wie die Fallbeispiele mit korreliertem Fehler zeigen, ist diese einfache Interpretation der Endwerteapproximation nicht immer zu halten, da einerseits die Fehlerbewertung bei der Konstruktion der Fuzzy-Daten und andererseits Korrelationen in den fehlerhaften Beobachtungen sich darin auswirken. Die Eignung als Benchmark begründet sich daher im Rückblick vor allem aufgrund der geometrischen Eigenschaft, Approximationsgerade der Endpunkte der Trägermengen $\widetilde{X}_i$ zu sein.

Als weitere Benchmark hat sich im Laufe der Untersuchungen die Kleinste Quadrate-Approximation an die Steiner-Punkte $\sigma_{\widetilde{X}_i}$ bzw. $\sigma_{\widetilde{Y}_i}$ herauskristallisiert, da sie sehr nah

[25] Dieses Verhalten ist erklärungsbedürftig: Erwarten würde man zunächst etwa das 4-fache, da rechte und linke Spannweiten der $\widetilde{Y}_i$ jeweils um den Faktor 4 steigen. Der Wachstumsfaktor entsteht vermutlich dadurch, dass die Spannweiten nach außen hin schnell wachsen und bei Multiplikation des Dehnungsfaktors die langen Spannweiten entsprechend vervielfacht werden und sich infolgedessen stärker auswirken.

bei der defuzzifizierten Approximation liegt. Im Fall mit genauen Inputs sind die beiden Approximationen identisch, im Fall mit genauen Outputs zeigt Gleichung (5.10) auf Seite 173, dass die Steigung der defuzzifizierten Gerade für echte Fuzzy-Daten immer kleiner ist als bei der Steiner-Punkte-Approximation.[26]

Die Ansätze der Kleinste Quadrate Fuzzy-Regression bewähren sich unterschiedlich gut in den verschiedenen Fehlerszenarien. Insgesamt betrachtet, schneiden das vereinfachte und das defuzzifizierte Modell ähnlich gut ab, wenn man berücksichtigt, dass im Fall mit genauen Outputs alle untersuchten Fuzzy-Ansätze zusammenfallen. Es ist zu konstatieren, dass der defuzzifizierte und der vereinfachte Fuzzy-Ansatz den verborgenen „wahren“ Zusammenhang zufriedenstellend approximieren und zwar in dem Sinne, dass die Approximationsgeraden die „wahre“ Gerade immer besser annähern als die Modalwerteapproximation und sich dabei im Rahmen der Grenzen bewegen, die in die Modellierung der Fuzzy-Daten eingeflossen ist. Es erscheint daher zulässig, die defuzzifizierte Approximation bzw. die Modalwertgerade der vereinfachten Fuzzy-Approximation als plausiblen Ort für die verborgene „wahre“ Gerade anzusehen. Die Plausibilitätsbewertung erhält allerdings dadurch einen besonderen Charakter, dass sie sich immer auf den funktionalen Zusammenhang von Fuzzy-Fehlerumgebungen bezieht. Die Approximationsfunktionen erscheinen daher im Wesentlichen als Kompromiss zwischen Modalwerteapproximation und Endwerteapproximation, der zu einer höheren Robustheit der Approximationsparameter gegenüber Änderungen in den Daten verhilft. Dabei gilt die Robustheit sowohl im Bezug auf Änderungen in den fehlerbehafteten, reellwertigen Beobachtungen, solange diese innerhalb der Trägermenge der Fuzzy-Daten verbleiben, als auch im Bezug auf Änderungen in der Modellierung der Fuzzy-Daten, sofern diese den Steiner-Punkt nur wenig beeinflussen. Hingegen ist die Kleinste Quadrate Fuzzy-Regression in vergleichbarer Weise wie die klassische Regressionsrechnung sensitiv für Veränderungen in der Lage der Fuzzy-Werte, wenn diese zu einer wesentlichen Transformation der Steiner-Punkte führt.

Im Fall von Fuzzy-Daten mit genauen Outputs schneidet die defuzzifizierte Approximation, die identisch mit der Modalwertgerade der vereinfachten Approximation ist, am besten ab. Dies gilt ebenfalls für die Fallbeispiele bei denen sowohl Inputs als auch Outputs Fuzzy-Mengen sind. Bei dem Szenario mit korrelierten Fehlern ist die Abweichung von der Fuzzy-Approximation aber nur gering. Die Anwendung des Fuzzy-Modells mit Fuzzy-Steigungsparameter nimmt ein Sonderposition ein und wurde aus technischen Gründen auf den Fall mit genauen Inputs beschränkt. Die Modalwertgerade weicht hierbei in Richtung der Approximation über die fehlerhaften Messwerte von der defuzzifizierten Approximation ab. Falls eine der Spannweiten negativ ausfallen würde, wird dies dadurch

[26] Insbesondere im Fallbeispiel mit einfachen systematischen Fehlern hat die Steiner-Punkte-Approximation eine Steigung, die höher ist als die der „wahren“ Gerade und als die der Modalwerteapproximation.

kompensiert, dass ein flacherer Steigungsparameter angepasst wird[27]. Stattdessen gibt die Fuzzy-Approximation Hinweise auf funktionale Tendenzen in der Modellierung der Fuzzy-Daten. Also beispielsweise: wie entwickeln sich die Relationen zwischen Modalwert, linker und rechter Spannweite im Verhältnis zum zugrundeliegenden funktionalen Zusammenhang. Allerdings führen die Einschränkungen durch die Irreversibilität von Fuzziness dazu, dass die Interpretation der Entwicklungstendenzen wenig intuitiv erscheint. Die Schwierigkeiten können anhand des Fallbeispiels in Abbildung 5.6 gut illustriert werden, bei dem die linke Spannweite der $\widetilde{Y}_i$ mit steigenden Werten wächst, wohingegen die rechte Spannweite fällt. Dies führt zu einer verschwindenden linken Spannweite im Achsenabschnitt $\widetilde{A}_0$ nicht aber zu einer verschwindenden rechten Spannweite, die rechte Spannweite verschwindet stattdessen beim Steigungsparameter $\widetilde{A}_1$.

Sowohl bei der Approximation eines Modells mit Fuzzy-Steigungsparameter als auch bei der Approximation eines vereinfachten Fuzzy-Modells ist zu beachten, dass die Approximationsfunktion nur in denjenigen Komponenten eine positive Spannweite aufweist, in denen die Fuzziness der abhängigen Variablen Y tendenziell größer ist als die Fuzziness der unabhängigen Variablen X. Wenn die Spannweiten proportional zur Steigung der Approximationsgeraden zusammenhängen, kompensieren sie sich, bei größerer Spannweite in Y entsteht eine positive Spannweite.

Defuzzifizierte bzw. vereinfachte Fuzzy-Approximation und Fuzzy-Approximation können daher als komplementäre Anwendungen aufgefasst werden, wobei mit ersteren Aussagen über die plausible Lage der „wahren“ Gerade getroffen und mit Hilfe von zweiterem Tendenzen in der Bewertung und Konstruktion der Fuzzy-Daten hervorgehoben werden können. Bei einer solchen Anwendung sollte aber auf Methoden übergangen werden, die zumindest eine abnehmende Entwicklung der Spannweiten abbilden können.

Bei einem Vorliegen von Fehlern in den Variablen — d.h. im Fall von Fuzzy-Daten mit genauen Outputs —, dem unsere besondere Aufmerksamkeit galt, schneidet die (defuzzifizierte) Fuzzy-Approximation, die hier mit den anderen Ansätzen zur Fuzzy-Regression zusammenfällt, im Vergleich mit den Benchmarks eher mittelmäßig ab. Aufgrund der höheren Trägheit gegenüber Änderungen in den Eingangsdaten haben die Parameter der Fuzzy-Approximation aber den Vorteil, dass sie robust sind. Demgegenüber scheint sich die relative Performance beim Übergang zu den Fehlerszenarien mit Fuzzy-Daten $(\widetilde{X}_i, \widetilde{Y}_i)$ zu verbessern. Insbesondere bei korrelierten Fehlern sind die Fuzzy-Ansätze eindeutig besser als eine naive Regression, die die Fehlereinflüsse nicht berücksichtigt. In diesem Fall erhalten wir bei der vereinfachten Approximation nicht mehr dieselbe Steigung wie im defuzzifizierten Modell. Die Güte der Approximation steigt insbesondere im Überstreckungsbeispiel, bei dem die Ungenauigkeit der Daten sehr hoch wird und der Einfluss des Rands der Trägermenge relativ zum Modalwert hoch ist.

[27] Das ist natürlich einfach die Auswirkung der Quadratischen Optimierung unter Nebenbedingungen.

Anhand der verwendeten Benchmarks kann gut gezeigt werden, welchen Einfluss die Modellierung der Fuzzy-Daten auf die Güte der Approximationen hat. Im Fall der gewählten Fehlerszenarien, bei denen die Modalwerte mit den fehlerbehafteten Messwerten identifiziert werden, ist das Verhältnis zwischen rechter und linker Spannweite in den Fuzzy-Daten die wesentliche Einflussgröße für die Anpassungsgüte.[28] Dies hängt mit der zentralen Rolle der Steiner-Punkte zusammen, die bei triangulären Fuzzy-Daten $\widetilde{Z}$ durch Transformation der Modalwerte aus dem Größenverhältnis zwischen den Spannweiten berechnet wird: $\sigma_{\widetilde{Z}} = z + \|L\|_1(r_Z - l_Z)$.

Der Einfluss der Schärfe der Fehlerabschätzung in den Fuzzy-Daten wurde bei korrelierten Fehlern mit den Überstreckungsbeispielen eingeschränkt erprobt. Demnach kann die Fuzzy-Approximation im Vergleich mit den Benchmarks die beste Approximation sein. Allerdings wurde die Modalwerteapproximation von der Überstreckung nicht berührt und die Endwerteapproximation verschlechtert sich. Ein besonderer Zusammenhang zwischen der Fuzziness der Daten und der Güte der Approximation ist hingegen nicht zu erkennen. Der Grad der Fuzziness — bzw. umgekehrt die Schärfe der Abschätzung — ist somit für die Approximation nur mittelbar relevant. Dennoch ist es eine wesentliche Information für die relative Güte der Approximationsfunktion, ob noch weit entfernte Punkte der Fehlerumgebungen in die Parameter einfließen. Die Approximation gilt gewissermaßen nur zu einem gewissen Grad an Fuzziness. Kennzahlen über die Fuzziness bzw. die Schärfe der Fuzzy-Daten sind daher eine wichtige Zusatzinformation bei der Interpretation der Kleinste Quadrate Fuzzy-Regression.

Bei einer vorliegenden Einschätzung darüber, wie nah die Modalwerte der Fuzzy-Daten in der Gesamtbetrachtung bei den „wahren“ Werten liegen, können ungünstige Effekte ggf. durch die Wahl einer geeigneten Bertoluzza-Metrik korrigiert werden, die den Kompromiss zwischen Modalwerten und Endwerten gewichtet. Im Fall der verwendeten Fehlerszenarien, in denen der Modalwert eindeutig falsch ist, wäre es evtl. sinnvoll zu einer Bertoluzza-Metrik überzugehen, die kein Gewicht auf eine Umgebung der Modalwerte legt.

Mit den Fehlerszenarien, die für das Benchmarking konstruiert wurden, sollten einfache systematische Fehler und korrelierte Fehler exemplarisch untersucht werden. Aus diesem Grund wurden die Generierungsalgorithmen so einfach wie möglich angelegt und sollten möglichst plakativ sein. Die Fuzzy-Modellierung der Fehlerszenarien sollte einerseits typische Messfehler wiedergeben und andererseits auch eine typische Fuzzy-Modellierung beschreiben, mit der man eine entsprechenden Mess-Situation abbilden könnte. Es ist

[28] Tatsächlich macht die Berechnungsformel für die Kleinste Quadrate Fuzzy-Regression deutlich, dass die Korrelationen zwischen linken und rechten Spannweiten und die Korrelationen der Spannweiten mit den Modalwerten für die Berechnung des Steigungsparameters relevant sind. Vgl. Diamond und Körner 1997, S. 18f.

festzustellen, dass die klassischen Benchmark-Approximationen auf die Fehlereffekte so reagieren, wie es aufgrund der theoretischen Überlegungen zu erwarten war.[29] In diesem Sinne können die beiden Fehlerszenarien als stimmig betrachtet werden.

Allerdings konnten bei den einfachen, systematischen Fehlern außer der Translation durch den deterministischen Anteil $\psi_\cdot$ nur geringe Effekte in den Benchmarkgeraden beobachtet werden. Diese sind auf die variierenden Fehleranteile in den Modalwerten zurückzuführen. Insgesamt erscheint die Konstruktion von unbekannten, systematischen Einflüssen noch zu stark von klassischen Wahrscheinlichkeitsverteilungen induziert zu sein, da die Konstruktion letztlich auf einer einfachen Verknüpfung von Gleichverteilungen basiert. Es wäre wünschenswert, wenn Konstruktionsalgorithmen entwickelt werden könnten, die die Einzeleffekte besser trennen.

Umgekehrt ist die Korrelation und ihre Verzerrungswirkung im zweiten Fallbeispiel extrem stark, so dass mögliche andere Einflüsse davon vollständig überformt werden, insbesondere unterscheiden sich die Ergebnisse verschiedener Simulationsgänge so gut wie nicht. Hierzu wäre eher zu überlegen, wie abgeschwächte Korrelationseffekte erzeugt werden könnten.

Ein möglicher Lösungsansatz liegt darin, von Fehlerszenarien auszugehen, bei denen die Ungenauigkeit funktional beeinflusst ist. Also beispielsweise, dass die Ungenauigkeit in den Inputs mit dem Abstand zum Ursprung ansteigt. Außerdem könnten Fuzzy-Daten $(\widetilde{X}_i, \widetilde{Y}_i)$ genutzt werden, um unterschiedliche Fehlereinflüsse in den Komponenten zu berücksichtigen und zu mischen.

Bisher wurde Fuzzy-Regression nur aus der Perspektive der explorativen Datenanalyse betrachtet, wobei die Untersuchung von mechanischen Eigenschaften der möglichen Approximationsmodelle und der Parameterapproximationen im Mittelpunkt stand. Dabei wurde durch die Simulation von Fehlerszenarien von Fuzzy-Daten eine Verbindung zwischen den „wahren“ Werten und den Fehlereinflüssen in den Beobachtungen hergestellt, die als Anhaltspunkt für eine Gegenüberstellung der Methoden genutzt werden konnte. Aufgrund dieses Konstruktionsprinzips eignen sich die simulierten Fehlerszenarien als Instrument für den Übergang zu Verteilungsmodellen für Fuzzy-Fehler. Sie bieten den Vorteil, dass damit die besonderen Bedingungen bei der Modellierung von Fehlern in den Daten operationalisiert und eingefangen werden können. Zum Abschluss sollen daher im folgenden Kapitel *6* noch einige Überlegungen zur Einbettung von Fehlerszenarien in die Verteilungsmodelle von Fuzzy-Zufallsvariablen angeknüpft werden.

[29] Vgl. Abschnitt *2.2.2*.

6 Potentiale der Fuzzy-Regression

In diesem Kapitel werden die Ergebnisse der vorangegangenen Untersuchungen im Hinblick auf die Anwendung von Fuzzy-Regression zur Datenanalyse bei fehlerbehafteten Daten zusammengefasst und ausgewertet. Die Fragestellung wurde aus verschiedenen Perspektiven beleuchtet. Ausgangspunkte dafür war der Vorschlag fehlerbehaftete Daten als Fuzzy-Daten zu modellieren, die nicht als Strukturmodell dargestellt werden können, da nicht genug Informationen über die Art der Fehler vorliegt. Anhand der exemplarischen Modellierung von Fuzzy-Merkmalsdaten am Beispiel der Beschäftigtenstatistik konnten die folgenden typischen Zielstellungen für die Modellierung von Fuzzy-Fehlern herausgearbeitet werden, die z.T. miteinander konkurrieren: Schärfe der Abschätzung, Nähe zwischen dem Kern des Fuzzy-Datums und dem „wahren“ Wert, Gewichtung durch den Rand des Fuzzy-Datums. Beim kritischen Vergleich der Fuzzy-Approximationsmethoden wurde der Schwerpunkt auf die Interpretation der verschiedenen Ansätze für Approximationsfunktionen gelegt. Als zentrale Methode, die bei allen Modellansätzen verwendet werden kann, erwies sich hierbei die Kleinste Quadrate Fuzzy-Regression. Sie führt vor allem bei Fuzzy-Inputs im Fehlerszenario mit korreliertem Fehler durch die Kompromissbildung zwischen Modalwerte- und Endwerteapproximation zu einer Korrektur gegenüber einer naiven Approximation über die fehlerbehafteten Werte.

Es liegt nahe, die Approximationsfunktion im defuzzifizierten Modell — bzw. die Modalwertgerade im vereinfachten Fuzzy-Modell — als plausible Approximation über die verborgenen „wahren“ Werte zu sehen. Eine Bestätigung dieser Betrachtungsweise ist aber nur dann möglich, wenn Verteilungsannahmen gemacht werden, die abbilden, welche Informationen über die „wahren“ Werte in den Fuzzy-Daten beschrieben ist. Die Auswertung der Fallbeispiele soll daher dadurch abgerundet werden, dass im ersten Abschnitt *6.1* einige Verteilungsmodelle für Fuzzy-Zufallsvariable vorgestellt werden und eine Einordnung der Fehlerszenarien, die bei den Fallbeispielen verwendet wurden, in die Verteilungsmodelle versucht wird. Damit wird der Übergang zu den Fuzzy-Schätzmodellen im Forschungsausblick vorbereitet. In Abschnitt *6.2* wird für die Untersuchungen in dieser Arbeit ein abschließendes Fazit über die Einsatzmöglichkeiten von Fuzzy-Regression als Instrument der Datenanalyse gezogen, die in den Ausblick auf die Forschungsperspektiven in Abschnitt *6.3* münden.

6.1 Verteilungsmodelle für die Fuzzy-Regression bei Fuzzy-Fehlern

In diesem Abschnitt werden die Fehlerszenarien der Fallstudien aus Kapitel *5* den Verteilungsmodellen für Fuzzy-Zufallsvariable in der epistemischen Interpretation gegenübergestellt. Da die Fehlerszenarien selber eher holzschnittartig gewählt wurden, geht es hierbei nicht darum, deren Verteilungseigenschaften zu analysieren. Stattdessen soll der Vergleich Hinweise dafür liefern, wie die Beobachtungen aus den Fallstudien vor dem Kontext eines Verteilungsmodells interpretiert werden könnten. Ziel ist es, einen Ausblick auf die Parameterschätzung im Fuzzy-linearen Modell zu geben, für das die Untersuchungen in dieser Arbeit eine Vorstufe darstellen, die aber in dieser Arbeit nicht abschließend geleistet werden.

Die Verteilungsmodelle für Fuzzy-Zufallsvariable werden nur knapp skizziert, dabei liegt der Augenmerk vor allem auf den Unterscheidungsmerkmalen. Die Darstellung folgt im wesentlichen Couso u. a. [2007]. Ihnen zufolge sind drei Verteilungsmodelle für physikalische und für epistemische Fuzzy-Daten zu unterscheiden, die mit einem spezifischen Interpretationsansatz für Fuzzy-Daten verknüpft werden können. Wesentlicher Grund für die Unterscheidung der Verteilungsmodelle ist, dass damit unterschiedliche Vorstellungen über die Streuung respektive die Dispersion der Fuzzy-Zufallsvariablen repräsentiert werden.

Die erste Variante einer Fuzzy-Zufallsvariable in der Definition von Puri und Ralescu [1986] ist hauptsächlich auf eine Situation bezogen, bei der die Beobachtungen nicht einem numerischen Wert entsprechen, sondern einen vagen linguistischen Begriff ausdrücken. In Anlehnung an die Definition der reellwertigen zufälligen Menge wird dabei angenommen, dass die Fuzzy-Zufallsvariable $\widetilde{Z} : \Omega \longrightarrow F_{coc}^{nob}(\mathbb{R})$ jeder Beobachtung eine Fuzzy-Menge zuordnet. Über einem gewöhnlichen Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ ist eine zufällige Menge als messbare Abbildung $M : \Omega \longrightarrow 2^{\Omega}$ definiert, formal ist sie nichts anderes als eine mengenwertige Zufallsvariable. Eine Fuzzy-Zufallsvariable $\widetilde{Z} : \Omega \longrightarrow F_{coc}^{nob}(\mathbb{R})$ wird dann dadurch definiert, dass alle α-Niveaus gewöhnliche Zufallsmengen sind, d.h. wenn für alle $\alpha \in [0,1]$ gilt $[\widetilde{Z}]^{\alpha} : \Omega \longrightarrow 2^{\mathbb{R}}$ ist eine zufällige Menge. Es lässt sich zeigen, dass $\widetilde{Z}$ eine Funktion ist, die im klassischen Sinne für eine σ-Algebra auf Ω und eine σ-Algebra auf $F_{coc}^{nob}(\mathbb{R})$ messbar ist, die durch eine passende Metrik d auf $F_{coc}^{nob}(\mathbb{R})$ erzeugt wird. In der Wahrscheinlichkeitsverteilung, die durch die Fuzzy-Zufallsvariable induziert wird, kann also die Zufallsinformation, die die Zufallsvariable modelliert, in der üblichen Weise zusammengefasst werden. Als Beispiel für diesen Ansatz kann eine Zufallsvariable $\widetilde{Z}$ herangezogen werden, die die vagen linguistischen Bewertungen „`hoch`“,

„`mittel`“, „`niedrig`“ abbildet.[1] Dann kann die folgende Verteilung angenommen werden: $P(\text{„}\texttt{hoch}\text{“}) = 0{,}5,\ P(\text{„}\texttt{mittel}\text{“}) = P(\text{„}\texttt{niedrig}\text{“}) = 0{,}25$.

Bei den gebräuchlichen Metriken für Fuzzy-Daten kann die Varianz im Verteilungsmodell von Puri/Ralescu auf die Varianz einer reellwertigen Zufallsvariablen Z' zurückgeführt werden. Diese Varianz wird als „klassische“ Varianz bezeichnet und wird berechnet als Fréchet-Varianz gemäß $\mathrm{Var}(\widetilde{Z}) = \int d^2(\widetilde{Z}, \mathrm{E}(\widetilde{Z}))\, dP$, wobei $\mathrm{E}(\widetilde{Z}) \in F_{coc}^{nob}(\mathbb{R})$ der Fréchet-Erwartungswert von $\widetilde{Z}$ bzgl. d sei, die in Gleichung (3.15) auf S. 115 definiert ist. Sie ist folglich ebenfalls reellwertig. Dieser Sachverhalt lässt sich so interpretieren, dass die Varianz nur auf Veränderungen der defuzzifizierten Werte der Zufallsvariablen reagiert und diese somit nur in ihrem Punktcharakter wahrnimmt.[2]

Bei einem weiteren Ansatz von Kwakernaak bzw. von Kruse/Mayer[3] zur Definition einer Fuzzy-Zufallsvariablen wird eine Fuzzy-Menge als Modellierung für das unvollständige Wissen über eine Beobachtung angesehen, die tatsächlich reellwertig ist. Dabei wird angenommen, dass die Fuzzy-Zufallsvariable $\widetilde{Z} : \Omega \longrightarrow F_{coc}^{nob}(\mathbb{R})$ ungenaue Informationen über eine klassische Zufallsvariable $U^* : \Omega \longrightarrow \mathbb{R}$ abbildet, die als *ursprüngliche Zufallsvariable* bezeichnet wird. Folglich beschreibt für $\omega \in \Omega$ die Zugehörigkeit $\mu_{\widetilde{Z}(\omega)}(z)$ die Bewertung der Zugehörigkeit des ursprünglichen Ereignisses $\{U^*(\omega) = z\}$ zur ungenauen Beobachtung $\widetilde{Z}(\omega)$.

In formaler Beschreibung ist eine Abbildung $\widetilde{Z} : \Omega \longrightarrow F_{coc}^{nob}(\mathbb{R})$ genau dann eine Fuzzy-Zufallsvariable, wenn für jedes $\alpha \in (0, 1]$ die Funktionen $\inf[\widetilde{Z}]^\alpha$ sowie $\sup[\widetilde{Z}]^\alpha$ $\mathcal{A}$-$\mathbb{B}(\mathbb{R})$-messbar sind. Damit ist sichergestellt, dass für jedes $\omega \in \Omega$ und jedes $\alpha \in (0, 1]$ eine reellwertige Zufallsvariable $U_\alpha : \Omega \longrightarrow \mathbb{R}$ existiert, so dass $\mu_{\widetilde{Z}(\omega)}(U_\alpha(\omega)) \geq \alpha$. Die U_α heißen Selektoren von $\widetilde{Z}$.

Da die ursprünglichen Werte $U^*(\omega)$ hinter den Fuzzy-Werten $\widetilde{Z}(\omega)$ „verborgen“ sind, können letztere als Aggregation über alle möglichen Selektoren $U : \Omega \longrightarrow \mathbb{R}$ betrachtet werden, für die $\omega \in \Omega$ existiert, so dass $U(\omega) \in \widetilde{Z}(\omega)$. Bei vorliegender Fuzzy-Variable $\widetilde{Z}$ kann daher die Plausibilität, dass ein möglicher Selektor U der ursprünglichen Zufallsvariable U^* entspricht, durch Aggregation über die Punktzugehörigkeiten zur Fuzzy-Zufallsvariable

[1] Tatsächlich gibt es auch nicht-diskrete Verteilungen für linguistische Variable, wenn beispielsweise das Adäquationsproblem bei der Erfassung statistischer Begriffe modelliert wird.

[2] Die Varianz im „klassischen“ Modell reagiert auch auf Veränderungen in der Genauigkeit der Werte, wie Couso u. a. 2007, S. 5 zeigen.

[3] Die Definition wird in der überarbeiteten Version von [Kruse und Meyer, 1987] vorgestellt. Als Originalarbeit von Kwakernaak wird meistens zitiert: H. Kwakernaak (1978): Fuzzy random variables I. *Inform. Sci.*, 15:1–29.

$\widetilde{Z}$ bewertet werden. Z.B. kann der *Plausibilitätsgrad*[4] von U bestimmt werden durch

$$\operatorname{acc}(U) = \inf_{\omega\in\Omega} \mu_{\widetilde{Z}(\omega)}(U(\omega)).$$

Dabei wird angenommen, dass die Ähnlichkeit zu U^* mit $\inf_\omega \mu_{\widetilde{Z}}(U)$ monoton wächst.

Der Zulässigkeitsgrad acc kann nun auf die Menge der zugehörigen Wahrscheinlichkeitsverteilungen P_U der zulässigen Zufallsvariablen U projiziert werden und wir erhalten

$$\begin{aligned}
\operatorname{supp}(P_{\widetilde{Z}}) = \{P_U \mid\ & U : \Omega \longrightarrow \mathbb{R} \text{ ist Zufallsvariable, so dass} \\
& \exists\, \alpha \in (0,1], \omega \in \Omega :\ \mu_{\widetilde{Z}(\omega)}(U(\omega)) \geq \alpha\}, \\
\mu_{P_{\widetilde{Z}}}(Q) = & \sup_{P_U = Q} \operatorname{acc}(U) \\
= & \sup_{P_U = Q} \big(\inf_{\omega\in\Omega} \mu_{\widetilde{Z}(\omega)}(U(\omega)) \big).
\end{aligned}$$

$P_{\widetilde{Z}}$ ist eine Fuzzy-Menge von Wahrscheinlichkeitsverteilungen und stellt eine Vereinigung der Wahrscheinlichkeitsverteilungen P_U der zulässigen Zufallsvariablen dar. Sie beschreibt den bestmöglichen Grad, mit dem eine der Zufallsvariablen U zu $\widetilde{Z}$ gehört, die die Wahrscheinlichkeitsverteilung Q erzeugen. Auf Basis dieses Modells können nur bewertete Aussagen gemacht werden, die für ein Ereignis B aus der Zusammenhangsformel $\mu_{P_{\widetilde{Z}}}(Q(B))$ abzuleiten sind. Durch Anwendung von Minkowski-Operationen erhalten wir so Aussagen der Art: „Der Grad der Zulässigkeit, dass die Wahrscheinlichkeit des „wahren“ Ereignisses $U^* = z$ gleich 0,25 ist, ist mindestens 0,1 und höchstens 0,7“.

Die Ergebnisse der reellwertigen ursprünglichen Zufallsvariable U^* sind bei diesem Ansatz deterministisch vorgegeben, sie sind aber vermittels der Fuzzy-Zufallsvariable $\widetilde{Z}$ nur ungenau wahrnehmbar. Diese stellt sich also für jedes $\omega \in \Omega$ als Fuzzy-Bündel von reellwertigen Zufallsvariablen dar. In diesem Modell wird die Varianz von $\widetilde{Z}$ als Fuzzy-Menge der Varianzen der möglichen reellwertigen Selektoren berechnet.[5] Der Wert $\mu_{\operatorname{Var}(\widetilde{Z})}(z)$ gibt die Möglichkeit wieder, mit der z beim ungenauen Messprozess $\widetilde{Z}$ die Varianz der ursprünglichen Variablen ist.

Schließlich schlagen Couso u. a. [2007] eine weitere Variante als Verteilungsmodell für eine Fuzzy-Zufallsvariable $\widetilde{Z}$ vor. Dabei wird angenommen, dass die Zufallsvariable die Verknüpfung von zwei Teilexperimenten $U^* : (\Omega, \mathcal{A}) \longrightarrow (\Gamma, \mathcal{C})$ und $V : \mathbb{B}(\mathbb{R}) \times \Gamma \longrightarrow [0,1]$ ist, wobei die Wahrscheinlichkeitsverteilung $P = P^{U^*}$ des ersten Teilexperimentes bekannt

[4] Die Bezeichnung acc steht für den englischen Begriff „acceptability degree“. Vgl. Couso u. a. 2007, S. 136.

[5] D.h. für alle $\alpha \in (0,1]$ sei $[\operatorname{Var}(\widetilde{Z})]^\alpha = \{\operatorname{Var}(U) \mid U \text{ ist Zufallsvariable und } \exists\ \omega \in \Omega \text{ so dass } \mu_{\widetilde{Z}(\omega)}(U(\omega)) \geq \alpha\}$. Die so definierte Varianz ist im allgemeinen nicht eindeutig definiert. Vgl. dazu ausführlich [Miranda u. a., 2005].

sei und für jedes $\gamma \in \Gamma$ gelte $V(\cdot, \gamma) = V(\cdot\,|\,\gamma) \in [0,1]^{\mathbb{B}(\mathbb{R})}$ ist eine bedingte Possibilitätsverteilung.[6] Die beiden Teilexperimente sind dadurch verknüpft, dass für $\omega \in \Omega$ mit dem Zwischenergebnis $\gamma' = U^*(\omega)$ gelte $V(A|\gamma') = V_{\widetilde{Z}(\omega)}(A) = \sup_{z \in A} \mu_{\widetilde{Z}(\omega)}(z)$ für alle $A \in \mathbb{B}(\mathbb{R})$.[7] Dabei seien die möglichen Übergangswahrscheinlichkeiten $Q(A|\,\gamma)$ für das Ereignis, dass das Gesamtergebnis der verknüpften Teilexperimente in $A \in \mathbb{B}(\mathbb{R})$ liegt, durch den Wert $V(A\,|\,\gamma)$ nach oben beschränkt.[8] Das Verteilungsmodell, das durch eine klassische Zufallsvariable induziert wird, wird also dadurch generalisiert, dass unterstellt wird, dass die beobachteten ungenauen Fuzzy-Werte abhängig vom aktuell realisierten Wert der „ursprünglichen" Zufallsvariablen U^* sind. Da die bedingte Possibilitätsverteilung die Verteilung des Gesamtexperimentes nach oben beschränkt, können wiederum unter Einsatz von Minkowski-Operationen Aussagen gemacht werden wie: die Wahrscheinlichkeit, dass die Beobachtung $U^* = z$ in das Intervall $[a, b]$ fällt, ist mindestens 0,3 und maximal 0,7.

Die Varianz einer Fuzzy-Zufallsvariable erscheint in diesem Modell im Unterschied zum zweiten Ansatz als gewöhnliche (scharfe) Menge in $\mathbb{R}$. Dabei werden für die Varianz des Verteilungsmodells die einfachen Varianzen für alle Wahrscheinlichkeiten aus der beschränkten Menge der möglichen Übergangswahrscheinlichkeiten Q zusammengefasst.

Wie Krätschmer [2001b] gezeigt hat, können sowohl Zufallsvariablen nach dem Verteilungsansatz von Puri/Ralescu als auch Zufallsvariablen nach dem Ansatz von Kwakernaak als einfache Zufallsvariable $\widetilde{Z} : (\Omega, \mathcal{A}, P) \longrightarrow (F_{coc}^{nob}(\mathbb{R}), \mathcal{F})$ mit kompakten Fuzzy-Werten aufgefasst werden, die mit der Borel-σ-Algebra $\mathcal{F}$ bzgl. δ_2 versehen ist. Das Konstruktionsverfahren nach dem Ansatz von Couso u.a. ergibt nur dann eine Fuzzy-Zufallsvariable, wenn zusätzliche Anforderungen an die Messbarkeit gestellt werden. Wir setzen daher für dieses Modell immer voraus, dass $\widetilde{Z}$ eine Fuzzy-Zufallsvariable nach dem zusammenfassenden Konzept gemäß Krätschmer [2001b] ist. Da der Aumann-Erwartungswert und der Fréchet-Erwartungswert für eine Fuzzy-Zufallsvariable zusammenfallen,[9] sind der Erwartungswert im Verteilungsmodell von Puri/Ralescu und im Modell von Kwakernaak identisch. Derselbe Erwartungswert kann entsprechend auch für das Modell nach Couso u.a. verwendet werden.[10]

[6] Die eigentlichen Anforderungen an V sind Messbarkeits- und Normalisierungseigenschaften.

[7] Bei der Darstellung von Couso u. a. [2007] wird auf die Zufallsvariable U^* verzichtet. Damit die beiden Darstellungen sich entsprechen, muss gelten dass U^* surjektiv ist bzw. dass P^{U^*}-fast sicher auf $\Gamma \setminus U^*(\Omega)$ gilt $V(A, \cdot) = 0$ für alle $A \in \mathbb{B}(\mathbb{R})$. Mit der hier gewählten Darstellung soll die Vergleichbarkeit mit den anderen Definitionen von Fuzzy-Zufallsvariablen erreicht werden.

[8] Im Rahmen der Theorie der Possibilitätsverteilungen kann dies auch direkt aus den Eigenschaften einer Possibilitätsverteilung gefolgert werden.

[9] S. dazu Gleichung (3.15) auf Seite 115.

[10] Dies ist sinnvoll, da auch hier per Definitionem eine Fuzzy-Zufallsvariable vorliegt.

Das Verteilungsmodell von Puri/Ralescu geht vom Punkt-Charakter der Fuzzy-Zahlen aus und ist daher für Fuzzy-Fehler kaum geeignet. Die beiden letzteren Ansätze hingegen geben jeweils eine Situation wieder, in der bei reellwertigen Beobachtungen nur Fuzzy-Fehlerumgebungen beobachtet werden können, die die Zugehörigkeit der möglichen Werte zur Beobachtung beschreiben, also bei epistemischen Fuzzy-Daten. Das Modell nach Kwakernaak ist dann angemessen, wenn zu jedem Zugehörigkeitsniveau α angegeben werden kann, in welchem Intervall $[\widetilde{Z}]^\alpha$ sich die möglichen Beobachtungswerte bewegen. Dabei wird angenommen, dass der „wahre" Zustand bereits feststeht, der Realisierungswert zu einem Ereignis $\omega \in \Omega$ liegt also schon vor. Im Verteilungsmodell nach Couso u.a. wird demgegenüber angenommen, dass für jeden beobachteten Zustand $\omega \in \Omega$ die zugehörige Messungenauigkeit $V(\cdot \mid U(\omega))$ angegeben werden kann. Dabei sei allerdings nur die Wahrscheinlichkeitsverteilung für den „wahren" Zustand bekannt. Ein Vorteil des Verteilungsmodells von Couso u.a. besteht darin, dass die Berechnung der Varianzmenge weniger Irregularitäten aufweist als bei der Fuzzy-Varianz im Kwakernaak-Modell.[11]

Um die Fehlerszenarien aus Kapitel *5* besser einordnen zu können, werden die verwendeten Konstruktionsverfahren mit Hilfe von gewöhnlichen Zufallsvariablen formalisiert. Dabei sollen die eingesetzten Zufallsvariablen stochastisch unabhängig sein. Wir beschränken uns zunächst auf die Betrachtung der Verteilungsmodelle für epistemische Fuzzy-Fehler in der abhängigen Variablen Y.[12]

Die Konstruktion des Fehlerszenarios mit einfachem, systematische Fehler kann folgendermaßen beschrieben werden. Seien dazu $m \in (0,1)$ das feste Verhältnis, mit dem der Modalwert die Trägermenge teilt. m wird hier zur Vereinfachung abweichend zu den simulierten Fuzzy-Daten als konstant angenommen. Seien ferner U^L, U^R, U stochastisch unabhängige reellwertige Zufallsvariable. Dabei seien U^L und U^R gleichverteilt mit $G([0, \frac{K}{2}])$ und U sei normalverteilt mit $N(\mu(x), \sigma^2)$. U repräsentiert das Verhalten der „wahren" Werte. Dann ergeben sich die Randwerte der Trägermenge und der Modalwert wie

$$Y^L(0) = U - U^L, \tag{6.1}$$

$$Y = U - U^L + m(U^L + U^R) = U + (m-1)U^L + mU^R, \tag{6.2}$$

$$Y^R(0) = U + U^R. \tag{6.3}$$

Diese Beschreibung ist möglich, weil es im Ergebnis nicht unterscheidbar ist, ob zunächst die Breite ζ der Trägermenge beliebig in $[0, K]$ und danach die Positionierung des „wahren" Wertes beliebig im Intervall $[0, \zeta]$ festgelegt wird, oder ob ausgehend vom „wahren" Wert

[11] Vgl. dazu [Kruse und Meyer, 1987] sowie [Miranda u. a., 2005].

[12] Diese Einschränkung führt allerdings dazu, dass der Fall von Fuzzy-Fehlern in den Variablen und reellwertigen Outputs hier ausgeschlossen wird. S. dazu weiter unten.

die rechte und die linke Spannweite beliebig im Intervall $[0, \frac{K}{2}]$ gezogen wird. Wir erhalten damit als Beschreibung der Fuzzy-Fehler des Szenarios

$$\mu_{\widetilde{Y}}(t) = \begin{cases} \frac{t-Y^L(0)}{Y-Y^L(0)} & \text{für } t \in [Y^L(0), Y] \\ \frac{Y^R(0)-t}{Y^R(0)-Y} & \text{für } t \in (Y, Y^R(0)] \\ 0 & \text{sonst} \end{cases}$$

$$= \begin{cases} \frac{t-U+U^L}{m(U^L+U^R)} & \text{für } t \in [U-U^L, U+(m-1)U^L+mU^R] \\ \frac{U+U^R-t}{(1-m)(U^L+U^R)} & \text{für } t \in (U+(m-1)U^L+mU^R, U+U^R] \\ 0 & \text{sonst} \end{cases}$$

In ähnlicher Weise kann auch das Generierungsverfahren für das Fehlerszenario mit korreliertem Fehler in Y formalisiert werden. Seien dazu $g(x) = c_y \arctan(x+5)$ wie in Gleichung (5.2), falls $x_1^* = 0$ und $x_n^* = 10$, und seien ferner U, U_1, U_2 stochastisch unabhängige, reellwertige Zufallsvariable, wobei U normalverteilt mit $N(\mu(x), \sigma^2)$ und U_1 gleichverteilt mit $G([0, \zeta^A])$, U_2 gleichverteilt mit $G([0, \zeta^U])$. Dann erhalten wir die Randwerte der Trägermenge und den Modalwert als

$$Y^L(0) = \begin{cases} U + g(x) - U_1 & \text{für } g(x) < 0 \\ U - U_2 & \text{für } g(x) \geq 0 \end{cases} \tag{6.4}$$

$$Y = U + g(x) \tag{6.5}$$

$$Y^R(0) = \begin{cases} U + U_2 & \text{für } g(x) < 0 \\ U + g(x) + U_1 & \text{für } g(x) \geq 0 \end{cases} \tag{6.6}$$

Entsprechend kann die Beschreibung der Fehlerszenarios hergeleitet werden als

$$\mu_{\widetilde{Y}}(t) = \begin{cases} \frac{t-Y^L(0)}{Y-Y^L(0)} & \text{für } t \in [Y^L(0), Y] \\ \frac{Y^R(0)-t}{Y^R(0)-Y} & \text{für } t \in [Y, Y^R(0)] \\ 0 & \text{sonst} \end{cases}$$

$$= \begin{cases} \frac{t-U-g(x)+U_1}{U_1} & \text{für } g(x) < 0, t \in [U+g(x)-U_1, U+g(x)] \\ \frac{U+U_2-t}{U_2-g(x)} & \text{für } g(x) < 0, t \in (U+g(x), U+U_2] \\ \frac{t-U+U_2}{U_2+g(x)} & \text{für } g(x) \geq 0, t \in [U-U_2, U+g(x)] \\ \frac{U+g(x)+U_1-t}{U_1} & \text{für } g(x) \geq 0, t \in (U+g(x), U+g(x)+U_1] \\ 0 & \text{sonst} \end{cases}$$

Das Verteilungsmodell nach Kwakernaak verlangt, dass für alle α-Niveaus die Randfunktionen $\inf[\widetilde{Y}]^\alpha$ und $\sup[\widetilde{Y}]^\alpha$ $\mathcal{A} - \mathbb{B}(\mathbb{R})$-messbar sind. Aus den Formeln der Zugehörigkeitsfunktionen $\mu_{\widetilde{Y}}$ ist erkennbar, dass die Randfunktionen als stetige Funktionen über Zufallsvariable ebenfalls Zufallsvariable sind. Allerdings ist dieses Verteilungsmodell blind dafür, dass U die „wahren“ Werte repräsentiert. U erhält damit für die Interpretation die Bedeutung eines nicht weiter spezifizierbaren Normalverteilungseinflusses in der Mischung, die sich durch das Bündel der möglichen Zufallsvariablen ergibt. Aussagen über die „wahren“ Werte können nur vermittelt über den Plausibilitätsgrad $\mathrm{acc}(W)$ bzw. über den besten Plausibilitätsgrad $\sup_{W \epsilon P_{\widetilde{Z}}} \mathrm{acc}(W)$ getroffen werden.

Bei den gewählten Konstruktionsverfahren werden zunächst die „wahren“ Werte und die Fuzzy-Umgebungen in Abhängigkeit davon bestimmt. Dies kommt dem Aufbau des Verteilungsmodells nach Couso et al. sehr nahe. Allerdings ist es für das Verteilungsmodell erforderlich, dass Zufallseinflüsse und Fuzziness so separiert werden können, dass die Possibilitätsverteilungen, die die Messungenauigkeit abbilden, nur noch von den Werten auf der Mess-Skala bzw. von den „wahren“ Werten abhängen. Gegenüber der allgemeinen Formulierung in Gleichungen (6.1)–(6.3) und (6.4)–(6.6) sind dazu die Werte für die Spannweiten als unabhängig von $\omega \in \Omega$ zu betrachten. D.h. es ist zu setzen $U^L = u^L \circ U, U^R = u^R \circ U$ mit $u^L, u^R : \mathbb{R} \longrightarrow \mathbb{R}$, so dass die Spannweiten nur noch von den Werten abhängen, die U annimmt. Analog ist vorauszusetzen, dass $U_1 = u_1 \circ U, U_2 = u_2 \circ U$ mit $u_1, u_2 : \mathbb{R} \longrightarrow \mathbb{R}$, so dass die Spannweiten nur noch von den Werten abhängen, die U_1 bzw. U_2 annehmen. Diese Annahmen stellen bei nicht-diskreten Wahrscheinlichkeitsverteilungen im allgemeinen keine große Einschränkung dar. Da die so eingeschränkte Fuzzy-Variable einen Spezialfall einer Fuzzy-Zufallsvariablen in der Definition von Kwakernaak darstellt, ist sie ebenfalls eine Fuzzy-Zufallsvariable in der üblichen Definition.

Das Verteilungsmodell von Couso u.a. ermöglicht es, den Zufallseinfluss von der Ungenauigkeit der Messung zu trennen, so dass einerseits der „wahre“ Wert explizit im Modell vorliegt, der im Modell von Kwakernaak nur implizit erfasst werden kann, und andererseits die Mechanik für den Zusammenhang zwischen „wahrem“ Wert und Messungenauigkeit angegeben werden kann. Man bezahlt die Trennung der Effekte allerdings mit einer Varianz, die gegenüber dem Ansatz von Kwakernaak weniger spezifisch ist. Insgesamt rückt die Modellierung nach Couso u.a. stärker in die Nähe des klassischen Fehlermodells, bei dem ebenfalls die Zufallsvariable der „wahren“ Werte mit Fehlerumgebungen versehen wird. Hier wäre noch genauer zu untersuchen, welche spezifischen Vorzüge das Fuzzy-Modell gegenüber dem klassischen Fehlermodell hat.

Die Ergebnisse zur Konsistenz und zum Zentralen Grenzwertsatz gelten für das Puri/ Ralescu-Modell bei Regression mit reellwertigen Parametern.[13] Sie lassen sich nicht ohne Weiteres auf die Verteilungsmodelle nach Kwakernaak bzw. nach Couso u.a. übertragen. Darüber hinaus sind die unterschiedlichen Verteilungsmodelle relevant, wenn die Varianz der Schätzparameter zu bestimmen ist, d.h. insbesondere bei der Ermittlung von Konfidenzbereichen und bei der Konstruktion von Tests. Gerade für den Ansatz nach Kwakernaak ist die Berechnung durch Irregularitäten erschwert.[14] Ein wesentliches Ziel des Verteilungsmodells von Couso u.a. ist es, ein Verteilungsmodell für epistemische Fuzzy-Daten zur Verfügung zustellen, das leichter zu berechnen und leichter zu interpretieren ist.

Allgemein kann das Verteilungsmodell nach Kwakernaak als gute Illustration für die Schätzung eines defuzzifizierten Modells bei Fuzzy-Fehlern betrachtet werden, da die Fuzzy-Daten als Bewertung eines Bündels von reellwertigen Zufallsvariablen aufgefasst werden.[15] Es ist davon auszugehen, dass die defuzzifizierte Approximationsgerade nur dann eine zufriedenstellende Plausibilitätsbewertung für den verborgenen „wahren“ Zusammenhang darstellen kann, wenn die Abschätzung der „wahren“ Werte gut angepasst ist. Die zufälligen Realisationen der „wahren“ Outputs, die im Couso-Modell von outputabhängigen Ungenauigkeitsumgebungen verwischt werden, können grundsätzlich mit einer vereinfachten Modellgleichung mit Fuzzy-Achsenabschnitt besser beschrieben werden. Eine Fuzzy-Charakteristik mit Fuzzy-Steigungsparameter erscheint hingegen nicht angemessen, da damit keine Aussage mehr über die plausible Lage der „wahren“ Werte getroffen werden kann.

Unser besonderes Interesse galt den Regressionsmodellen mit Fuzzy-Fehlern in den Inputs. Falls hierbei genaue Outputs vorliegen, ergibt sich allerdings eine Friktion in der Modellgleichung: der Erwartungswert über die Outputvariable ist immer reellwertig und die rechte Seite der Gleichung ergibt immer eine Fuzzy-Zahl mit positiven Spannweiten.[16] Die Gleichung ist daher nicht im eigentlichen Sinne erfüllbar. Es ist somit anzugeben, was unter „Gleichheit“ im Sinne der Modellgleichung zu verstehen ist.

Wenn die Modellgleichung defuzzifiziert wird, kann die abhängige Variable Y als klassische, reellwertige Zufallsvariable dargestellt werden, deren Verteilung mit einem der klassischen Modelle beschrieben werden kann. Es seien also Y eine reellwertige Zufallsvariable und $O_F : F_{coc}^{nob}(\mathbb{R}) \longrightarrow \mathbb{R}$ ein Defuzzifizierungsfunktional mit den Eigenschaften wie

[13] Vgl. Fußnoten 63f auf Seite 116.

[14] Vgl. [Kruse und Meyer, 1987] sowie [Miranda u. a., 2005].

[15] Es ist aber auch die Anpassung einer Fuzzy-Charakteristik mit Fuzzy-Parametern gut mit dem Kwakernaak-Ansatz vereinbar, da damit die gewichtete Abschätzung über die möglichen Bildwerte auf die Regressionsfunktion übertragen werden kann.

[16] Vgl. das empirische Beispiel in Abschnitt *5.2.2*.

in Bemerkung 3.7[17], dann gelte

$$\mathrm{E}\,Y = O_F(a_0 \oplus a_1 \odot \widetilde{X}) = a_0 + O_F(a_1 \odot \widetilde{X}).$$

Unter diese Modelle kann z.B. die Kleinste Quadrate-Regression über die Steiner-Punkte der Inputs gerechnet werden.

Um zu einem geschlossenen Fuzzy-Schätzmodell zu gelangen, kann umgekehrt aber auch Y als ursprüngliche Zufallsvariable aufgefasst werden, die die „wahren“ Werte abbildet und die in eine unbekannte Fuzzy-Zufallsvariable $\widetilde{Y}$ eingebettet ist, deren Fuzziness durch die Fuzziness der Inputwerte induziert wird. Dazu sei $\widetilde{Y}$ eine Fuzzy-Zufallsvariable und es gebe einen Selektor V_0 von $\widetilde{Y}$, so dass

$$\mathrm{E}\,\widetilde{Y} = a_0 \oplus a_1 \odot \widetilde{X} \quad \text{und} \quad Y = V_0 \;\; P\text{-fast sicher.}$$

Falls für die Zufallsvariable $\widetilde{Y}$ das Verteilungsmodell nach Kwakernaak hinterlegt wird, ist eine eindeutige Verortung von Y zu $\widetilde{Y}$ nicht möglich. Stattdessen kann z.B. gefordert werden, dass gilt

$$\mathrm{acc}_{\widetilde{Y}}(V_0) = \underset{V \text{ Selektor von } \widetilde{Y}}{\arg\max} \mathrm{acc}(V),$$

wobei $\mathrm{E}\,\widetilde{Y} = a_0 \oplus a_1 \odot \widetilde{X}$.

Falls ein Verteilungsmodell gemäß Couso u.a. unterstellt wird, kann Y mit der Zufallsvariablen der „wahren“ Werte identifiziert werden. Aussagen über die bedingte Possibilitätsverteilung V sind dann aus den aktuell vorliegenden Fuzzy-Daten abzuleiten. Beispielsweise könnte eine zum unterstellten Fehlerszenario passende Defuzzifizierung für das Verhältnis von „wahren“ Werten und Fuzzy-Fehlerumgebungen in den Inputs gewählt werden, um die Lage von Y relativ zum Funktional $\chi_{\widetilde{X}_1,\ldots,\widetilde{X}_n}(\mathbf{a})$ in den vorliegenden Fuzzy-Inputs zu bestimmen. Typischerweise entspricht das bei einem Kleinste Quadrate-Ansatz über δ_2 zunächst dem Steiner-Punkt $\sigma_{a_0 \oplus a_1 \odot \widetilde{X}_i}$, , da dieser den Abstand zu $a_0 \oplus a_1 \odot \widetilde{X}_i$ minimiert. Dabei ist der Steiner-Punkt als neutraler Ansatz zur Defuzzifizierung zu sehen, wenn keine weiteren Informationen über das Fehlerszenario vorliegen, da bei der Bestimmung kein Zugehörigkeitsniveau bevorzugt wird. Falls davon auszugehen ist, dass die „wahren“ Werte bei den Inputs in einem anderen Verhältnis zu den Fuzzy-Daten stehen, das sich auf das Regressionsfunktional überträgt, kann ggf. eine andere Defuzzifizierung vorgezogen werden. So kann z.B. zu einer anderen Defuzzifizierung übergegangen werden, wenn aufgrund eines Randbeobachtungs-Problems das Gewicht des Modalwertes reduziert oder wenn das Gewicht der linken Spannweite erhöht ist.

[17] Vermutlich ist es sinnvoll, darüber hinaus vorauszusetzen, dass O_F ähnliche Linearitätsbedingungen wie die Stützfunktion einer Fuzzy-Menge erfüllt.

Um das Verhältnis zwischen „wahren“ Werten und den abgebildeten Fehlereinflüssen in den Inputs besser zu beschreiben, kann schließlich zu Modellen übergegangen werden, bei denen auch Verteilungsannahmen über die Inputvariablen gemacht werden.

6.2 Datenanalyse mit Fuzzy-Regression

Wenn der praktische Nutzen einer Fuzzy-Modellierung von fehlerbehafteten Daten für die Datenanalyse zusammenfassend bewertet werden soll, ergibt sich vor dem Hintergrund der Untersuchungen in dieser Arbeit eher ein gemischtes Bild.

Mit der Fuzzy-Modellierung von fehlerhaften Daten wird ein Perspektivwechsel vorgenommen. Statt einer Modellierung von wesentlichen Strukturzusammenhängen wird in den Fuzzy-Merkmalswerten die Ungenauigkeit aufgrund von Fehlereinflüssen einzelfallbezogen und auf dem aktuellen Stand des Wissens abgebildet. Dabei lässt die Modellierung in Form von LR-Fuzzy-Intervallen es zu, dass auch qualitative Einflüsse dargestellt werden, wenn sie quantitativ zumindest ungefähr beschrieben werden können. Insbesondere können damit auch systematische Fehler operationalisiert werden, die nur ungenau bekannt sind. Die Fuzzy-Perspektive bietet somit die Möglichkeit, ein reichhaltigeres Spektrum von Fehlereinflüssen abzubilden, als bei den klassischen Verfahren zur Fehlerschätzung. Aufgrund der einzelfallbezogenen Modellierung können auch singuläre Effekte aufbereitet und bei den anschließenden Analysen einbezogen werden.

Da bei der Modellierung der Fuzzy-Daten eine subjektive Bewertung der Datenungenauigkeiten vorgenommen wird, ist dieser Vorteil allerdings mit der Gefahr verbunden, dass Artefakte durch Täuschungen in der Wahrnehmung induziert werden. Falls fehlerhafte, reellwertige Messdaten vorliegen, deren Ungenauigkeiten ex post als Fuzzy-Daten modelliert werden sollen, muss zudem der Paradoxie begegnet werden, dass für die nicht-zutreffenden Werte die höchste Sicherheit im Bezug auf ihre Zugehörigkeit zum Fuzzy-Merkmalswert besteht.

Insgesamt können Fuzzy-Merkmalswerte als ein Instrument gesehen werden, mit dem eine vorsichtige Bestandsaufnahme von Fehlereinflüssen durchgeführt und eine vorschnelle Vereinfachung bzw. Idealisierung von Verzerrungseffekten in den Daten vermieden werden kann. Die einzelfallbezogene Modellierung der Fehlereinflüsse und deren Visualisierung in den Fuzzy-Daten bietet außerdem einen Ansatzpunkt zur Analyse von weichen Faktoren der Datenqualität, wie z.B. das Expert/innenwissen, das in die Modellierung eingeflossen ist, oder die vorgenommene quantitative Bewertung von qualitativen Fehlereffekten. Denn auf der Basis der Fuzzy-Daten können auch die Strukturen in den Ungenauigkeitsumgebungen untersucht werden. Die Ergebnisse können evaluiert und zur Verbesserung der

Fuzzy-Modellierung sowie zur Generierung von weiterem Fehlerwissen, also von Wissen über die Fehlereinflüsse, verwendet werden.

Ein ähnlich positiver Befund kann für die Methoden zur Fuzzy-Regression bei epistemischen Fuzzy-Daten hingegen nicht abgegeben werden.

Der Vorzug der Kleinste Quadrate Fuzzy-Approximation gegenüber anderen Ansätzen zur Fuzzy-Regression besteht vor allem darin, dass damit sowohl bei Modellfunktionen mit reellwertigen Parametern als auch bei Modellfunktionen mit Fuzzy-Parametern immer eine Lösung des Approximationsproblems existiert. Da die δ_2-Metrik als Erweiterung der euklidischen Metrik definiert und stetig in $L^2([0,1]\times\mathbb{S}^0)$ eingebettet ist, kann die Lösung zudem als Fortsetzung der klassischen Kleinste Quadrate-Approximation aufgefasst werden.

Da die metrische Approximation vom Punktcharakter der Fuzzy-Daten ausgeht, kann die Approximationsgerade grundsätzlich nur als Abbildungsvorschrift für Fuzzy-Ungenauigkeitsumgebungen gelesen und interpretiert werden, die ähnlich genau bzw. ungenau sind wie die Eingangsdaten, auf deren Basis sie ermittelt wurden. Sie kann somit nur auf einem gegebenen Genauigkeitsniveau interpretiert werden.

Das Benchmarking in Kapitel *5* zeigt, dass die Kleinste Quadrate Fuzzy-Approximation auf der Basis der modellierten Fuzzy-Merkmalswerte tatsächlich zu einer Korrektur gegenüber einem naiven Kleinste Quadrate-Ausgleich führt, bei dem die Fehlereinflüsse nicht berücksichtigt werden. Dies gilt sowohl für die Approximationsgerade im defuzzifizierten Modell als auch für die Modalwertgerade der Approximation im vereinfachten Fuzzy-Modell. Zudem erweisen sich die Parameter der Fuzzy-Approximation als robuster gegenüber Veränderungen in den Modalwerten oder der Trägermengen der Fuzzy-Daten als die reellwertigen Benchmarks der Modalwerteapproximation oder der Endwerteapproximation. In diesem Sinne kann die Kleinste Quadrate Fuzzy-Approximation im Rahmen der explorativen Verfahren zur Fuzzy-Regression auch als Annäherung an einen plausiblen funktionalen Zusammenhang der verborgenen „wahren“ Werte aufgefasst werden.

Während für die explorative Approximation einer Ausgleichsgeraden das Abbildungsverhältnis der Fuzzy-Ungenauigkeitsumgebungen unproblematisch ist, wird hingegen im Schätzmodell eine Interpretation für den Zusammenhang zwischen Input-Werten und Output-Werten benötigt, die auch das Verhältnis zwischen den Fuzzy-Ungenauigkeitsumgebungen mit einbezieht. So ist es in einem Schätzmodell z.B. erklärungsbedürftig, in welchem Verhältnis eine genaue Outputvariable zum Fuzzy-Bild von ungenauen Input-Variablen steht. Einige Beispiele für entsprechende Schätzmodelle wurden in Abschnitt *6.1* auf S. 209 präsentiert. Mit einer ähnlichen Problemstellung ist man konfrontiert, wenn in einem Schätzmodell die Fuzziness der $\widetilde{Y}$ wesentlich von der Fuzziness von $f(\widetilde{X}_1,\ldots,\widetilde{X}_k;$

$\widetilde{A}_0, a_1, \ldots, a_k$) abweicht. Falls singuläre Fehlereinflüsse in den Daten bekannt sind, wird dies bei einer Modellierung als Fuzzy-Merkmalsvariable in der Regel durch eine auffallend große Spannweite der Ungenauigkeitsumgebung abgebildet. Infolge der Robustheit der Kleinste Quadrate Fuzzy-Regression bei Änderungen in den Spannweiten der Fuzzy-Daten werden die Auswirkungen des singulären Fehlers durch die Modellierung abgeschwächt.

Gleichförmige Strukturen in den epistemischen Fuzzy-Daten, die Auswirkungen auf das Approximationsergebnis haben und für dessen Interpretation relevant sind, können mit Hilfe von Fehlerszenarien angegeben werden. Der Begriff der Fehlerszenarien wurde in dieser Arbeit eingeführt, weil für die Simulation von Fuzzy-Merkmalswerten eine konstruktive Beschreibung von möglichen Fehlereinflüssen in den Daten gemeinsam mit den daran anknüpfenden Modellierungsstrategien benötigt wird, die den Stand des Wissens über den „wahren“ Wert abbilden. Umgekehrt können Fehlerszenarien für vorliegende Fuzzy-Daten auch Auskunft darüber geben, welche Aspekte der Fuzzy-Daten zur Beschreibung der „wahren“ Werte relevant sind. Infolgedessen können damit auch Annahmen über Artefakte verdichtet und operationalisiert werden, die aufgrund der Modellierungsstrategie entstehen, so dass sie bei der Auswahl der Approximationsmethode Berücksichtigung finden können. Die Fehlerszenarien können, aber müssen nicht, in ein Verteilungsmodell von Fuzzy-Zufallsvariablen eingebettet sein. Durch die Einführung von Fehlerszenarien als zusätzlichem Instrument ist es also möglich, sowohl die Bewertung der Zugehörigkeitsgrade für die Messwerte als auch eine Metabewertung für die Güte der Fuzzy-Abschätzungen zu formalisieren und in das Approximationsverfahren einzubeziehen.

Die Fehlerszenarien können bei der Approximation durch die Auswahl einer Kleinste Quadrate-Metrik $\delta_{(\nu)}$ antizipiert werden, die eine Gewichtung zwischen den Zugehörigkeitsniveaus der Fuzzy-Daten vornimmt, die sich in Richtung der „wahren“ Werte auswirkt. Die Verbesserung der Approximationsgüte der Fuzzy-Approximation in den Abbildungen 5.10 und 5.11 kann auch so gelesen werden, dass eine bessere Approximation erzielt wird, je enger die Nähe zwischen den „wahren“ Werten und den Modalwerten der Fuzzy-Daten relativ zu den Rändern der Trägermenge ist. Dahinter tritt die Fuzziness als Qualitätskriterium im Hinblick auf die Approximationsgüte zunächst zurück. Die Fehlerszenarien, die der Simulation der Fallbeispiele zugrundegelegt wurden, beschreiben Situationen, in denen die vorliegenden fehlerhaften Werte die sicherste Information über die „wahren“ Werte darstellen. In Reaktion darauf wird eine Modellierungsstrategie für die Fuzzy-Daten gewählt, bei der der Beobachtungswert den Modalwert bildet und die Ränder der Trägermenge so aufgespannt werden, dass der „wahre“ Wert sicher eingeschlossen ist. Hierbei ist der Modalwert nur in Ausnahmefällen identisch mit dem „wahren“ Wert. Gegen die verwendeten Fehlerszenarien, die die Paradoxie bei der Bewertung der fehlerhaften Beobachtungswerte in den Mittelpunkt stellen, lässt sich kritisch einwenden, dass diese im-

mer noch eine verzerrte Darstellung der Datenungenauigkeit sind. Trotz des Aufwands für die Fuzzy-Modellierung könne demnach keine Verbesserung in der Abbildung erreicht werden, so dass besser darauf verzichtet werden sollte. Dem kann entgegengehalten werden, dass die Modellierung von Fehlereinflüssen durch Zufallsvariable in einem Schätzmodell zwar das offenere Konzept ist, weil dafür nur Annahmen über die Häufigkeitsverteilung der Abweichungen getroffen werden müsse und auf die örtliche Fixierung von fehlerhafter Messung und „wahrem“ Wert verzichtet werden kann, so dass die beschriebene Paradoxie entfällt. Es darf aber nicht vergessen werden, dass das Wahrscheinlichkeitskalkül, dessen man sich dann bedient, mit starken Annahmen versehen ist. Wenn die Verteilungsannahmen nicht zutreffen, können die Abbildungsfehler daher erheblich sein. Im Vergleich dazu erfordert eine Fuzzy-Modellierung in viel stärkerem Maße, sich die Verzerrungen bei der Abbildung bewusst zu machen und zu analysieren. Sie kann damit zu einer höheren Transparenz im Umgang mit Fehlern in den Daten beitragen.

Die δ_2-Metrik wertet alle Zugehörigkeitsniveaus gleich und führt zu einer Approximation in der die Abweichungen auf allen α-Niveaus gleichmäßig berücksichtigt werden. Somit wird in der δ_2-Metrik kein Niveau der Fuzzy-Daten besonders gewichtet, außerdem wird kein Zugehörigkeitsbereich ausgeschlossen. Die beiden alternativen Kleinste Quadrate Fuzzy-Metriken δ_H und δ_G, die für die Fallbeispiele verwendet wurden, gewichten die Modalwerte stärker als die δ_2-Metriken. Daher kann damit nur in den Fällen eine Verbesserung der Approximation erreicht werden, in denen die Modalwerte in Richtung der „wahren“ Geraden ziehen. Die δ_2-Metrik zeichnet sich also dadurch aus, dass sie neutral gegenüber beliebigen Fehlerszenarien ist. Da die Modalwerte bei den simulierten Fuzzy-Daten keine gute Annäherung an den „wahren“ Wert sind, ist grundsätzlich ein Fuzzy-Abstandsfunktional besser für die Approximation geeignet, bei dem eine Umgebung um den Modalwert nicht gewichtet wird. Beispielhaft wird hier die folgenden Definition vorgeschlagen: Sei $\kappa \in (0,1)$. Dann sei für $\widetilde{A}, \widetilde{B} \in F_{coc}^{nob}(\mathbb{R})$ die folgende Quasi-Metrik definiert $\delta_\kappa(\widetilde{A}, \widetilde{B}) = \delta_2(\widetilde{A}^{[\kappa]}, \widetilde{B}^{[\kappa]})$, wobei $\widetilde{A}^{[\kappa]}$ durch $\mu_{\widetilde{A}^{[\kappa]}}(z) = \min\{\mu_{\widetilde{A}}(z), \kappa\}$ gegeben sei, entsprechend für $\widetilde{B}^{[\kappa]}$). Für δ_κ gilt aufgrund der Konstruktion allerdings im allgemeinen nicht mehr, dass $\delta_\kappa(\widetilde{A}, \widetilde{B}) = 0 \Rightarrow \widetilde{A} = \widetilde{B}$. Insbesondere ist δ_κ keine Erweiterung der euklidischen Metrik mehr.

Bei der Approximation eines Modells mit Fuzzy-Steigungsparametern können keine Aussagen über die plausible Lage der „wahren“ Werte abgeleitet werden. Stattdessen gibt die Approximation Hinweise auf globale Strukturen in den Ungenauigkeitsumgebungen der Fuzzy-Daten, wenn man von der Vorstellung ausgeht, dass die ungenauen Outputs bei Angabe von genauen Inputs möglichst gut reproduziert werden sollen. Aufgrund der eingeschränkten Linearität von Fuzzy-Zahlen ist die Approximation einer Fuzzy-Charakteristik aber nur in wenigen Fällen sinnvoll möglich. Insbesondere muss die Monotonie in den

Spannweiten der Fuzzy-Outputs möglichst zur Schmetterlingsform der Fuzzy-linearen Modellfunktion passen. Eine Abnahme von Spannweiten kann mit diesem Modellansatz nicht oder nur mit wenig intuitiven Methoden entdeckt werden. Aus diesem Grund ist der Ansatz nicht für eine Evaluation von Fehlerstrukturen geeignet. Für die Untersuchung der Abschätzungsstrukturen in den Fuzzy-Daten ist es unwichtig, ob die Fuzzy-Daten als Einheiten linear verknüpft sind. Es wäre daher sinnvoller, zu alternativen Approximationsmodellen überzugehen, die eine größere Flexibilität in der Anpassung haben. Daher ist eher zu empfehlen, dass die Betrachtung auf die geometrischen Eigenschaften im Koordinatenraum ausgerichtet wird, wie es z.B. der Ansatz von D'Urso und Gastaldi [2000] vorsieht. Die Anpassung von Fuzzy-linearen Modellen mit Fuzzy-Steigungsparametern ist hingegen interessant, wenn ein entsprechendes Fuzzy-Bündel von Funktionen angepasst werden soll, d.h. wenn nur ein vager Strukturzusammenhang vorliegt.

Abschließend ist aus den Ergebnissen dieser Arbeit das Fazit zu ziehen, dass der Vorteil einer Fuzzy-Modellierung bei Fehlern in den Daten in erster Linie darin liegt, dass die Fehlereinflüsse einzelfallbezogen und explizit abgebildet werden können und somit für weitere Analysen zur Verfügung stehen. Die Modellierung von Fuzzy-Merkmalswerten kostet einigen Aufwand und lohnt sich vor allem dann, wenn die Qualität der Daten sowie die Qualität der Abbildung der Datenungenauigkeiten durch die Fuzzy-Daten im Vordergrund der Untersuchungen stehen. Die Methoden zur Fuzzy-Regression stellen ein Instrumentarium zur Funktionsapproximation zur Verfügung, wenn solche Fuzzy-Daten vorliegen. Beim derzeitigen Stand der Entwicklung sind sie als explorative Methoden zu charakterisieren, mit denen eine Ausgleichsgerade bestimmt werden kann, die für das Genauigkeitsniveau der vorliegenden Daten auch als plausible Gerade für die verborgenen „wahren“ Werte interpretiert werden kann. Die Güte der Approximation kann verbessert werden, wenn die Modellierungsstrategie für die Fuzzy-Daten durch ein Fehlerszenario beschrieben und durch die Auswahl einer passenden Metrik antizipiert werden kann.

6.3 Forschungsperspektiven

Aus den Ergebnissen der vorangegangenen Untersuchungen ergeben sich einige weiterführende Fragestellungen. Als besonders aussichtsreich erscheint hierbei die Entwicklung von Messverfahren, mit der die Möglichkeit zur einzelfallbezogenen Abbildung von Fehlereinflüssen durch eine Fuzzy-Modellierung besser ausgenutzt werden kann. Ergänzend zu der bisherigen Modellierung von Fehlereinflüssen ex post zu bereits vorliegenden Daten wäre es hilfreich, auch über Verfahren zur Generierung von epistemischen Fuzzy-Beobachtungen im Zuge des Messprozesses zu verfügen. Dabei wäre zunächst über Verfahren zur Hebung von bereits vorhandenem Wissen für die Bewertung der Datenqualität nachzudenken, z.B.

durch begleitende, qualitative Fragebögen zur Erhebung von Informationen über Besonderheiten im laufenden Messprozess oder durch eine Reaktivierung und Einbindung von dokumentierten Wissensbeständen.

Die Fuzzifizierung von Klassifizierungungssystemen, wie z.B. der Wirtschaftszweige oder der Berufe, wurde in dieser Arbeit nicht genauer behandelt. Fuzzy-Klassifikationen ermöglichen es, dass die Abwägung für die Zuordnung einer statistischen Einheit in die eine oder die andere Klasse durch Teilzugehörigkeiten abzubilden. Dadurch kann die Zahl von Missklassifikationen reduziert und Erosionsprozesse innerhalb von Klassifikationssystemen können explizit abgebildet und im Zeitverlauf nachvollzogen werden. Forschungsbedarf besteht im Hinblick auf Fuzzy-Klassifikationen zum einen in der Frage, wie ein praktikables Erhebungsverfahren bei Verwendung von Fuzzy-Klassifikationen gestaltet werden könnte. Zum anderen wären für die Verwendung von Daten, die eine Fuzzy-Klassifizierung beinhalten, andere Regressionsmethoden erforderlich, als in dieser Arbeit vorgestellt wurden.[18]

Da die Fuzzy-Daten auf der Basis einer subjektiven Bewertung der Datenungenauigkeiten modelliert werden, ist es möglich, dass bei einer ungünstigen Modellierung Scheinzusammenhänge durch die Modellierung der Daten verursacht werden, die das Approximationsergebnis verschlechtern. Um günstige Strategien für die Modellierung von Fuzzy-Fehlern zu finden, wäre daher noch genauer zu untersuchen, welche Interdependenzen zwischen den Eigenschaften der Fuzzy-Daten und den Approximationsparametern bestehen und ob es Dateneigenschaften gibt, auf die das Approximationsverfahren besonders sensitiv reagiert. Insbesondere ist abzugrenzen, inwieweit eine ungünstige Beschaffenheit der Fuzzy-Daten dem Stand des Wissens entspricht oder ob sie durch unbewusste Prozesse erzeugt wurden und zu korrigieren sind.

Die Fehlerszenarien, die in dieser Arbeit für die Simulation von epistemischen Fuzzy-Daten konstruiert wurden, sind relativ einfach angelegt. Das Bild, das sie zeichnen, ist im Fall der sog. einfachen, systematischen Fehler noch zu stark von klassischen Zufallseinflüssen geprägt, im Fall der sog. korrelierten Fehler sind die Effekte so stark ausgeprägt, dass bei Fuzzy-Daten $(\widetilde{X}_i, \widetilde{Y}_i)$ kaum noch Unterschiede zwischen den Benchmarks zu erkennen sind. Mit den Fehlerszenarien werden die Modellierungsstrategien für fehlerhafte Daten vereinheitlicht und konzentriert, daher stellen sie ein wesentliches Instrument dar, mit dem die Auswirkungen eines Modellierungsverfahrens auf das Approximationsergebnis analysiert werden kann. Es besteht daher großer Bedarf für die Entwicklung von realistischeren Fehlerszenarien. Weiteres Forschungspotential ist vor allem im Hinblick auf die Mischung von Modellierungsstrategien zu sehen, insbesondere dann, wenn zu multivariaten Ansätzen übergegangen wird. In diese Richtung geht auch schon die Frage nach

[18] Es ist zu vermuten, dass die „Profilzuordnungsverfahren“ nach Manton u. a. [1994] in diese Richtung gehen.

der Kompensation oder Verstärkung von Fehlereffekten bei Daten mit Fuzzy-Inputs und Fuzzy-Outputs, die in Abschnitt *5.1.1* angesprochen wurde. Nicht zuletzt bilden die Fehlerszenarien eine wichtiges Bindeglied für die Konstruktion von Verteilungsmodellen für epistemische Fuzzy-Daten.

Im Bezug auf die Methoden zur Fuzzy-Regression steht eine genauere Sensitivitätsanalyse für bi- und multivariate Kleinste Quadrate-Ansätze noch aus. Es ist damit zu rechnen, dass dann, ähnlich wie bei den klassischen Fehlermodellen, die Verzerrungs- und Korrektureffekte komplexer sind, und nicht mehr so leicht interpretiert werden können, wie in den vorgeführten Fallbeispielen. Der Forschungsbedarf in Bezug auf die Entwicklung von Schätzmodellen bei epistemischen Fuzzy-Merkmalswerten wurde bereits in Abschnitt *6.1* skizziert. Bei einem Modell mit Fuzzy-Inputs kann in einem Schätzmodell, mit dem eine Aussage über den Zusammenhang zwischen den „wahren“ Werten getroffen werden soll, nicht auf eine Modellierung mit stochastischen Inputs verzichtet werden.

A Anhang

A.1 δ_2-Metrik und Steiner-Punkt

Der Steiner-Punkt stellt im Bezug auf die δ_2-Metrik auf dem Raum der Fuzzy-Zahlen $F_{coc}^{nob}(\mathbb{R})$ in ähnlicher Weise einen Schwerpunkt dar, wie dies der Mittelwert im Bezug auf die Euklidische Metrik ist. Die technischen Eigenschaften, die bei der Kleinste Quadrate Fuzzy-Regression Verwendung finden, werden hier zusammengestellt.

Das Konzept der Stützfunktion einer Menge und die weiteren Eigenschaften können von gewöhnlichen konvexen, kompakten Mengen in $\mathbb{R}^m$ auf konvexe, kompakte Fuzzy-Mengen über $\mathbb{R}^m$ übertragen werden.[1] Zunächst wird die Definition für gewöhnliche Mengen über $\mathbb{R}^m$ angegeben.

Definition A.1 (Stützfunktion einer konvexen, kompakten Menge) Sei $K \subset \mathbb{R}^m$ konvex und kompakt. Dann ist die *Stützfunktion* $s_K : \mathbb{R}^m \longrightarrow \mathbb{R}$ von K definiert durch

$$s_K(u) = \sup\{\langle u, z\rangle \mid z \in K\}.$$

Da K konvex und kompakt ist, wird das Supremum immer angenommen und s_K ist wohldefiniert.

Die Definition kann mit Hilfe der α-Schnitt-Darstellung auf Fuzzy-Mengen übertragen werden. Wir benötigen im Weiteren nur die Definition der Stützfunktion einer Fuzzy-Menge für $m = 1$.

Definition A.2 (Stützfunktion einer Fuzzy-Menge) Sei $\widetilde{A} \in F_{coc}^{nob}(\mathbb{R})$ eine konvexe, kompakte Fuzzy-Menge. Die *Stützfunktion* $s_{\widetilde{A}} : [0,1] \times \mathbb{S}^0 \longrightarrow \mathbb{R}$ von $\widetilde{A}$ ist dann definiert durch

$$s_{\widetilde{A}}(\alpha, u) = \begin{cases} \sup\{uy \mid y \in [\widetilde{A}]^\alpha\} & \text{für } \alpha \in (0,1], \\ 0 & \text{für } \alpha = 0, \end{cases}$$

[1] Vgl. Diamond und Kloeden 1994, S. 13ff.

wobei $\mathbb{S}^0 = \{-1, 1\}$ die euklidische Einheitskugel in $\mathbb{R}^1$ sei. Das Produkt uy entspricht dem Skalarprodukt $\langle u, y\rangle$ über $\mathbb{R}^1$.

Eine konvexe, kompakte Fuzzy-Menge wird eindeutig durch ihre Stützfunktion charakterisiert.

Satz A.3 (Eigenschaften der Stützfunktion) Die Stützfunktionen sind linear bezüglich der erweiterten Addition und des positiven Skalarprodukts auf Fuzzy-Mengen. Seien $\widetilde{A}, \widetilde{B} \in F_{coc}^{nob}(\mathbb{R})$ und $\lambda \geq 0$. Dann gilt

$$s_{\widetilde{A} \oplus \widetilde{B}} = s_{\widetilde{A}} + s_{\widetilde{B}} \quad \text{und} \quad s_{\lambda \odot \widetilde{A}} = \lambda s_{\widetilde{A}}.$$

Für negative Skalare $\lambda < 0$ kann der folgende Zusammenhang berechnet werden. Für alle $\alpha \in (0, 1]$

$$\begin{aligned} s_{\lambda \odot \widetilde{A}}(\alpha, u) &= s_{-\lambda \odot (-\widetilde{A})}(\alpha, u) \\ &= -\lambda s_{-\widetilde{A}}(\alpha, u) \qquad \text{(A.1)} \\ &= -\lambda \sup\{uy \mid y \in [-a^R(\alpha), -\alpha^L(\alpha)]\} \\ &= -\lambda \sup\{u(-y') \mid y' \in [a^L(\alpha), \alpha^R(\alpha)]\} \\ &= -\lambda s_{\widetilde{A}}(\alpha, -u). \qquad \text{(A.2)} \end{aligned}$$

Sei nun die Funktion $\jmath : F_{coc}^{nob}(\mathbb{R}) \longrightarrow \mathrm{Map}([0,1] \times \mathbb{S}^0, \mathbb{R}), \widetilde{A} \longmapsto s_{\widetilde{A}}$ gegeben, mit der die konvexen, kompakten Fuzzy-Mengen in ihre Stützfunktionen abgebildet werden. Dann gilt, dass $\jmath$ injektiv ist. Außerdem überträgt $\jmath$ die lineare Struktur mit Addition und positiver Skalarmultiplikation von $F_{coc}^{nob}(\mathbb{R})$ in die Menge der Funktionen $\mathrm{Map}([0,1] \times \mathbb{S}^0, \mathbb{R})$.[2]

Die δ_2-Metrik wird so definiert, dass der Abstand zwischen Fuzzy-Mengen in $F_{coc}^{nob}(\mathbb{R})$ mit dem Abstand der zugehörigen Stützfunktionen identifiziert wird.

Definition A.4 (δ_2-Metrik) Für Fuzzy-Mengen $\widetilde{A}, \widetilde{B} \in F_{coc}^{nob}(\mathbb{R})$ sei die Metrik δ_2 gegeben durch

$$\delta_2^2(\widetilde{A}, \widetilde{B}) = \int_{[0,1] \times \mathbb{S}^0} (s_{\widetilde{A}} - s_{\widetilde{B}})^2 \, d\lambda^1 \otimes \lambda^{\mathbb{S}^0}$$

mit $\lambda^{[0,1]}, \lambda^{\mathbb{S}^0}$ den zu $[0,1]$ bzw. $\mathbb{S}^0$ gehörigen Borel-Lebesgue-Maßen.

[2] Die Skalarmultiplikation für $\jmath$ kann durch Definition einer „negativen“ Skalarmultiplikation für negative Skalare $\lambda < 0$ gemäß Gleichung (A.2) auf ganz $\mathbb{R}$ erweitert werden. Vgl. Krätschmer 2006a, S. 2581.

Satz A.5 Da für jede Fuzzy-Zahl $\widetilde{Z}$ und jedes $\alpha \in [0,1]$ gilt, dass $s_{\widetilde{Z}}(\alpha,-1) = -z^L(\alpha)$ und $s_{\widetilde{Z}}(\alpha,1) = z^R(\alpha)$ kann die Definition von δ_2 umformuliert werden. Seien $\widetilde{A},\widetilde{B} \in F_{coc}^{nob}(\mathbb{R})$, dann gilt

$$\delta_2^2(\widetilde{A},\widetilde{B}) = \frac{1}{2}\int_0^1 [(-a^L(\alpha)+b^L(\alpha))^2 + (a^R(\alpha)-b^R(\alpha))^2]\,d\alpha.$$

Aus dieser Darstellung und der Konstruktion des „negativen“ Skalarproduktes gemäß Gleichung (A.2) kann gefolgert werden, dass δ_2 homogen in beliebigen Skalaren $\lambda \in \mathbb{R}$ ist. Es gilt

$$\delta_2^2(0,\lambda\widetilde{X}) = \lambda^2\delta_2^2(0,\widetilde{X}). \tag{A.3}$$

Bei LR-Fuzzy-Intervallen kann der δ_2-Abstand mit der folgenden Formel direkt berechnet werden, falls gilt $L^{-1}, R^{-1} \in L^2(\mathbb{R})$:

$$\begin{aligned}\delta_2^2(\widetilde{A},\widetilde{B}) = {} & (a-b)^2 + \|R\|_2^2(r_A-r_B)^2 + \|L\|_2^2(l_A-l_B)^2 \\ & + 2(a-b)\Big[\|R\|_1(r_A-r_B) - \|L\|_1(l_A-l_B)\Big],\end{aligned} \tag{A.4}$$

wobei

$$\begin{aligned}\|L\|_1 &= \frac{1}{2}\int_0^1 L^{-1}(\alpha)\,d\alpha, \qquad \|L\|_2^2 = \frac{1}{2}\int_0^1 (L^{-1}(\alpha))^2\,d\alpha \quad \text{und}\\ \|R\|_1 &= \frac{1}{2}\int_0^1 R^{-1}(\alpha)\,d\alpha, \qquad \|R\|_2^2 = \frac{1}{2}\int_0^1 (R^{-1}(\alpha))^2\,d\alpha.\end{aligned}$$

Es gibt eine ganze Klasse von Metriken, die eine äquivalente Topologie auf $F_{coc}^{nob}(\mathbb{R})^m$ induzieren wie δ_2. Um zu einer Kleinste Quadrate Fuzzy-Regression mit diesen Metriken überzugehen, müssen nur einige Gewichtsfaktoren im Optimierungsproblem verändert werden.

Satz A.6 (Erweiterte Familie von Kleinste Quadrate Fuzzy-Metriken)
Sei ein endliches Wahrscheinlichkeitsmaß ν auf $\mathbb{B}([0,1])$ gegeben. Für $\widetilde{A},\widetilde{B} \in F_{coc}^{nob}(\mathbb{R})$ ist die erweiterte Metrik $\delta_{(\nu)}$ dann gegeben durch[3]

$$\delta_{(\nu)}^2(\widetilde{A},\widetilde{B}) = \int_{[0,1]\times\mathbb{S}^0} (s_{\widetilde{A}}(\alpha,u) - s_{\widetilde{B}}(\alpha,u))^2\,d\nu(\alpha)\otimes\lambda^{\mathbb{S}^0}(u).$$

Unter Verwendung eines Beweises von Bertoluzza u. a. [1995] kann gezeigt werden, dass bei Einschränkung auf die LR-Fuzzy-Intervalle alle $\delta_{(\nu)}$-Metriken topologisch äquivalent zur δ_2-Topologie sind.

[3] Vgl. Diamond und Körner 1997, S.30.

Der $\delta_{(\nu)}$-Abstand von LR-Fuzzy-Intervallen kann mit derselben Formel wie der δ_2-Abstand direkt berechnet werden, falls gilt $L^{-1}, R^{-1} \in L^2(\nu)$. Dabei sind nur die Gewichtsfaktoren $\|R\|_2^{(\nu)}, \|R\|_1^{(\nu)}, \|L\|_2^{(\nu)}, \|L\|_1^{(\nu)}$ anzupassen. Dabei werden die gegenüber Gleichung (A.4) veränderten Gewichtsfaktoren bestimmt als

$$\|L\|_1^{(\nu)} = \frac{1}{2}\int_0^1 L^{-1}(\alpha)\, d\nu(\alpha), \qquad \|L\|_2^{(\nu)} = \frac{1}{2}\int_0^1 (L^{-1}(\alpha))^2\, d\nu(\alpha) \quad \text{und}$$
$$\|R\|_1^{(\nu)} = \frac{1}{2}\int_0^1 R^{-1}(\alpha)\, d\nu(\alpha), \qquad \|R\|_2^{(\nu)} = \frac{1}{2}\int_0^1 (R^{-1}(\alpha))^2\, d\nu(\alpha).$$

Der Steiner-Punkt stellt im Bezug zur δ_2-Metrik so etwas wie den reellwertigen Schwerpunkt einer Fuzzy-Menge dar. Ihm kommt daher eine ähnliche Funktion zu wie dem Mittelwert im Bezug zur Euklidischen Metrik.

Definition A.7 (Steiner-Punkt einer Fuzzy-Menge) Sei $\widetilde{A} \in F_{coc}^{nob}(\mathbb{R})$ eine Fuzzy-Menge. Dann ist der *Steiner-Punkt* von $\widetilde{A}$ definiert durch

$$\sigma_{\widetilde{A}} = \int_{[0,1]\times\mathbb{S}^0} u\, s_{\widetilde{a}}(\alpha, u)\, d\lambda^1(\alpha) \otimes \lambda^{\mathbb{S}^0}(u)$$

Satz A.8 Der Steiner-Punkt ist ein charakterisierender Punkt für $\widetilde{A}$ in dem Sinne, dass gilt

$$\delta_2(\{\sigma_{\widetilde{A}}\}, \widetilde{A}) = \inf_{v\in\mathbb{R}} \delta_2(\{v\}, \widetilde{A}).$$

Es gilt die folgende Zerlegung des Abstands von Fuzzy-Mengen in den euklidischen Abstand zwischen den Steiner-Punkten und den verbleibenden Abstand der Ungenauigkeitsumgebungen.

$$\delta_2^2(\widetilde{A}, \widetilde{B}) = \|\sigma_{\widetilde{A}} - \sigma_{\widetilde{B}}\|^2 + \delta_2^2(\widetilde{A}^\circ, \widetilde{B}^\circ),$$

wobei $\widetilde{A}^\circ = \widetilde{A} \ominus \sigma_{\widetilde{A}}$ und $\widetilde{B}^\circ = \widetilde{B} \ominus \sigma_{\widetilde{B}}$ die in 0 zentrierten Fuzzy-Mengen sind.

Die Steiner-Punkte erben von den Stützfunktionen auch die lineare Struktur, d.h. für $\widetilde{A}, \widetilde{B} \in F_{coc}^{nob}(\mathbb{R})$ und $\lambda \geq 0$ gilt

$$\sigma_{\widetilde{A}\oplus\widetilde{B}} = \sigma_{\widetilde{A}} + \sigma_{\widetilde{B}} \quad \text{und} \quad \sigma_{\lambda\odot\widetilde{A}} = \lambda\sigma_{\widetilde{A}}.$$

Für negative Skalare $\lambda < 0$ gilt mit Gleichung (A.1)

$$\sigma_{\lambda\odot\widetilde{A}} = -\lambda\sigma_{-\widetilde{A}} = |\lambda|\sigma_{-\widetilde{A}}.$$

Für die in 0 zentrierten Fuzzy-Mengen kann daraus die folgende Umformung für negative Skalare $\lambda < 0$ abgeleitet werden, die wir für Berechnungen benötigen.

$$\begin{aligned}(\lambda\widetilde{A})^\circ &= \lambda\widetilde{A} - (-\lambda\sigma_{-\widetilde{A}}) \\ &= (-\lambda)(-\widetilde{A}) - (-\lambda)\sigma_{-\widetilde{A}} \\ &= (-\lambda)(-\widetilde{A})^\circ.\end{aligned} \tag{A.5}$$

Da der Steiner-Punkt einer Fuzzy-Menge $\widetilde{A}$ den δ_2-Abstand minimiert, kann bei LR-Fuzzy-Intervallen $\widetilde{A} = (a; l_A, r_A)_{LR}$ durch Einsetzen und Minimieren von Formel (A.4) die folgende einfache Bestimmungsgleichung berechnet werden.

$$\sigma_{\widetilde{A}} = a + \|R\|_1 r_A - \|L\|_1 l_A. \tag{A.6}$$

Über den Weg der Minimierungseigenschaft kann auch ein Steiner-Punkt für die erweiterte Familie von Kleinste Quadrate Fuzzy-Metriken definiert werden.

Definition A.9 Der Steiner-Punkt für eine erweiterte Kleinste Quadrate Fuzzy-Metrik $\delta_{(\nu)}$ ist definiert, wenn es einen eindeutigen Wert $\sigma_{\widetilde{A}}^{(\nu)}$ gibt[4], so dass gilt

$$\delta_{(\nu)}(\{\sigma_{\widetilde{A}}^{(\nu)}\}, \widetilde{A}) = \inf_{v \in \mathbb{R}} \delta_{(\nu)}(\{v\}, \widetilde{A}).$$

Für LR-Fuzzy-Intervalle kann somit für eine erweiterte Kleinste Quadrate Fuzzy-Metrik $\delta_{(\nu)}$ durch Minimieren der folgende abgeleitete Steiner-Punkt bestimmt werden

$$\sigma_{\widetilde{A}}^{(\nu)} = a + \|R\|_1^{(\nu)} r_A - \|L\|_1^{(\nu)} l_A.$$

[4]Die Definition gilt also insbesondere dann nicht, wenn $\delta_{(\nu)}$ eine Quasi-Metrik ist, d.h. wenn aus $\delta_{(\nu)}(\widetilde{A}, \widetilde{B}) = 0$ nicht folgt $\widetilde{A} = \widetilde{B}$.

A.2 Dokumentation der abgebildeten Datensätze

Auf den folgenden Seiten werden die Fallbeispiele dokumentiert, die in Abschnitt *5.2* präsentiert wurden. Dabei werden nur die Datensätze selber sowie die Residuen und die Bestimmtheitsmaße für die Benchmarkapproximationen dargestellt. Die Parameterwerte wurden bereits im Text angegeben. Bei den Fuzzy-Approximationen werden nur die Ergebnisse für die δ_2-Metrik ausführlich dokumentiert. Das Bestimmtheitsmaß für die verschiedenen Modellansätze der Kleinste Quadrate Fuzzy-Regression werden gemäß der folgenden Formel berechnet, die hier in der allgemeinsten Variante notiert wird:[5]

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \delta_2\big(\widetilde{Y}_i, f(\widetilde{X}_i, \widetilde{A})\big)^2}{\sum_{i=1}^{n} \delta_2(\widetilde{Y}_i, \overline{\widetilde{Y}}_\cdot)^2},$$

wobei $\overline{\widetilde{Y}}_\cdot = \frac{1}{n}\sum_{i=1}^{n} \widetilde{Y}_i$ den Mittelwert der Fuzzy-Messwerte bezeichnet. Je näher R^2 bei 1 liegt, desto besser ist die Anpassung.

Die Datenwerte sind mit steigenden „wahren“ Werten x_i^* angeordnet. Die Auflistung der Residuen erfolgt in der Reihenfolge der Datenwerte. Sie können daher entsprechend zugeordnet werden.

Für die Benchmarkregressionen werden die folgenden Kürzel verwendet:

- „WAHR“ für die „wahre“ Gerade,
- „MOD“ für die Modalwerteapproximation,
- „END“ für die Endwerteapproximation,
- „DFUZ“ für die Approximation im defuzzifizierten Modell,
- „VFUZ“ für die Approximation im vereinfachten Fuzzy-Modell,
- „FFUZ“ für die Approximation im Modell mit Fuzzy-Parametern.
- „RINP“ bzw. „ROUT“ für die Randapproximationen mit genauen, reellwertigen In- bzw. Outputs bei Fuzzy-Daten $(\widetilde{X}_i, \widetilde{Y}_i)$, wobei als markante Punkte die Steiner-Punkte $\sigma_{\widetilde{X}_i}$ bzw. $\sigma_{\widetilde{Y}_i}$ ausgewählt wurden.

[5] Vgl. z.B. Diamond und Körner 1997, S. 18.

A.2.1 Fallbeispiele mit genauen Inputs

Outputs mit einfachen, systematischen Fehlern

Datenwerte

x_i^*	y_i^*	$y_i^L(0), \quad y_i \quad , y_i^R(0)$
0.2148	1.8986	1.8115, 2.0670, 2.6126
2.3010	3.4795	3.0461, 3.3480, 3.8500
3.2772	1.9462	1.6910, 1.8854, 2.2020
3.7644	2.5957	2.4941, 2.5773, 2.8830
4.1217	3.1249	2.9321, 3.0905, 3.3997
4.1764	2.4992	2.0431, 2.1964, 2.5348
4.8715	2.6378	2.6346, 2.6404, 2.6498
6.2007	4.5804	4.3511, 4.4980, 4.7736
6.4699	3.4919	3.4055, 3.5779, 4.0131
7.7410	4.4822	4.0944, 4.2303, 4.5430
8.5502	4.4000	4.3582, 4.4594, 4.8112
8.6171	3.6671	3.3216, 3.4798, 4.0769
9.1746	3.6870	3.2216, 3.5591, 4.1577
9.3576	3.3532	3.0634, 3.2527, 3.7463
9.7320	4.1263	3.8677, 4.1371, 4.8262

Bestimmtheitsmaß und Residuen

	WAHR	DFUZ	VFUZ	FFUZ
R^2	0.5279	0.5115	0.5078	0.5083
	-0.1970	0.0189	0.0664	0.0993
	0.9308	0.8364	0.8388	0.8406
	-0.8145	-0.8521	-0.8525	-0.8522
	-0.2708	-0.2382	-0.2447	-0.2423
	0.1807	0.1816	0.1849	0.1832
	-0.4568	-0.7155	-0.7161	-0.7157
	-0.4691	-0.4638	-0.4930	-0.4909
	1.1848	1.1440	1.1450	1.1451
	0.0378	0.2005	0.2007	0.2006
	0.7521	0.5627	0.5640	0.5649
	0.4941	0.6392	0.6403	0.6413
	-0.2534	-0.3075	-0.3130	-0.3103
	-0.3545	-0.3905	-0.4037	-0.3994
	-0.7280	-0.7248	-0.7255	-0.7249
	-0.0362	0.1093	0.1564	0.1409

Outputs mit korrelierten Fehlern

Datenwerte

x_i^*	y_i^*	$y_i^L(0)$, y_i , $y_i^R(0)$
0.4610	1.1004	0.6222, 0.6935, 1.3656
2.4582	2.1749	1.7422, 1.8140, 2.4678
2.8096	3.0295	2.6269, 2.6839, 3.2806
3.1652	2.0073	1.5640, 1.6820, 2.3047
3.4241	2.6215	2.1531, 2.3153, 2.7851
3.4807	3.1147	2.7300, 2.8132, 3.3758
3.9813	2.2028	1.8413, 1.9565, 2.2917
4.1143	2.7931	2.4460, 2.5666, 2.8604
4.5940	3.4129	3.2589, 3.2832, 3.6318
6.1510	3.4448	3.0746, 3.6941, 3.8159
8.1899	4.1672	3.8036, 4.5458, 4.7307
8.5004	4.2493	3.8611, 4.6358, 4.7376
9.5519	4.7641	4.7332, 5.1697, 5.3679
9.5575	3.3670	3.1250, 3.7726, 3.7977
9.6498	4.4094	4.1522, 4.8163, 4.8586

Bestimmtheitsmaß und Residuen

	WAHR	MOD	DFUZ	VFUZ	FFUZ
R^2	0.7868	0.8736	0.8450	0.8409	0.8419
	-0.5671	-0.4038	-0.4515	-0.4594	-0.4644
	-0.0888	-0.1017	-0.0766	-0.1115	-0.1129
	0.6608	0.6243	0.6524	0.6564	0.6553
	-0.4676	-0.5233	-0.4902	-0.4954	-0.4958
	0.0694	0.0039	-0.0022	-0.0404	-0.0302
	0.5456	0.4786	0.5176	0.5216	0.5202
	-0.5157	-0.5832	-0.5896	-0.5934	-0.5909
	0.0349	-0.0276	-0.0405	-0.0847	-0.0641
	0.5115	0.4924	0.5359	0.5440	0.5411
	0.0786	0.2654	0.1637	0.1823	0.1815
	0.1922	0.2815	0.2441	0.2697	0.2564
	0.1816	0.2443	0.1901	0.2244	0.2085
	0.3824	0.3473	0.4425	0.4444	0.4482
	-1.0163	-1.0520	-1.0527	-1.0569	-1.0566
	-0.0015	-0.0461	-0.0430	-0.1049	-0.0951

A.2.2 Fallbeispiele mit genauen Outputs

Inputs mit einfachen, systematischen Fehlern

Datenwerte

x_i^*	$x_i^L(0)$, x_i , $x_i^R(0)$	y_i^*
2.5725	1.9086, 2.8736, 3.1575	2.4260
3.2887	3.1356, 3.7861, 3.8768	2.1009
3.5655	3.0134, 3.7667, 3.8892	2.7733
3.5688	3.4730, 4.0098, 4.0820	2.6839
3.6461	2.9080, 3.7099, 3.8336	3.0898
3.9933	3.5096, 3.9405, 4.1231	3.5238
6.8990	6.2850, 6.9973, 7.2322	3.8402
7.0135	6.9978, 7.2431, 7.3245	3.4893
7.5218	6.7794, 8.0532, 8.2648	3.9751
7.7442	7.6955, 8.1801, 8.3849	4.0236
8.5395	8.0355, 8.8759, 8.9922	4.2336
9.1450	9.0412, 9.5913, 9.7036	5.1410
9.4140	8.9367, 9.6481, 9.8798	4.7418
9.4362	8.9660, 9.7782, 9.8770	4.7920
9.8449	9.3574, 9.9935, 10.1087	4.2687

Bestimmtheitsmaß und Residuen

	WAHR	MOD	END $x_i^L(0)$, $x_i^R(0)$	FFUZ
R^2	0.8612	0.8442	0.8223	0.8365
	-0.0283	-0.0597	0.1184 , -0.2592	-0.1298
	-0.5808	-0.6683	-0.5777, -0.8018	-0.6791
	0.0037	0.0101	0.1317 , -0.1332	0.0831
	-0.0867	-0.1548	-0.0967, -0.2809	-0.1783
	0.2946	0.3443	0.4801 , 0.2002	0.3569
	0.6184	0.7066	0.7321 , 0.5466	0.6783
	0.0124	0.0732	0.2092 , -0.0772	0.1110
	-0.3749	-0.3541	-0.3572, -0.4560	-0.3827
	-0.0505	-0.1200	0.1946 , -0.2545	-0.1594
	-0.0726	-0.1109	-0.0339, -0.2424	-0.1413
	-0.1151	-0.1171	0.0733 , -0.2160	-0.1327
	0.6001	0.5680	0.6765 , 0.4762	0.5719
	0.1155	0.1511	0.3089 , 0.0237	0.1775
	0.1586	0.1609	0.3503 , 0.0748	0.2028
	-0.4944	-0.4293	-0.2914, -0.5186	-0.4271

Inputs mit korrelierten Fehlern

Datenwerte

x_i^*	$x_i^L(0)$, x_i , $x_i^R(0)$	y_i^*
1.1446	1.0194, 1.6816, 1.7393	1.6573
2.1369	2.0418, 2.6478, 2.7886	1.8442
2.8526	2.8374, 3.3338, 3.3861	2.5033
3.0218	2.9536, 3.4937, 3.5299	2.4199
3.2043	3.1028, 3.6648, 3.8173	1.7828
3.3382	3.2227, 3.7894, 3.8004	2.5270
3.9368	3.9188, 4.3311, 4.4145	2.8294
4.6469	4.5358, 4.9167, 4.9648	2.5769
5.3990	5.3238, 5.4180, 5.4791	3.6900
5.5095	5.3094, 5.4843, 5.5576	3.2058
5.7011	5.4394, 5.6014, 5.8335	2.6983
5.9191	5.6597, 5.7425, 5.9517	3.2400
6.6518	6.2840, 6.3005, 6.7043	3.8645
9.1768	8.5272, 8.6533, 9.3190	4.2402
9.7485	8.9787, 9.2115, 9.8607	4.0582

Bestimmtheitsmaß und Residuen

	WAHR	MOD	END $x_i^L(0)$, $x_i^R(0)$	FFUZ
R^2	0.8413	0.8317	0.8241	0.8296
	-0.0767	-0.0271	0.0576 , -0.1806	-0.0891
	-0.1975	-0.1922	-0.0937, -0.3408	-0.2181
	0.2396	0.2169	0.3022 , 0.1206	0.2234
	0.1036	0.0752	0.1803 , -0.0104	0.1031
	-0.5900	-0.6242	-0.5061, -0.7425	-0.6277
	0.1126	0.0746	0.1984 , 0.0073	0.1109
	0.2294	0.1796	0.2705 , 0.1065	0.1921
	-0.2434	-0.2863	-0.1862, -0.3281	-0.2749
	0.6364	0.6441	0.6663 , 0.6149	0.6421
	0.1178	0.1357	0.1868 , 0.1046	0.1428
	-0.4491	-0.4145	-0.3638, -0.4941	-0.4245
	0.0250	0.0758	0.1051 , 0.0085	0.0717
	0.4223	0.4970	0.5231 , 0.3840	0.4754
	0.0147	0.0154	0.1566 , -0.1054	0.0858
	-0.3447	-0.3701	-0.1749, -0.4667	-0.3593

A.2.3 Fallbeispiele mit Fuzzy In- und Outputs

In- und Outputs mit einfachen, systematischen Fehlern

Datenwerte

x_i^*	$x_i^L(0)$, x_i , $x_i^R(0)$	y_i^*	$y_i^L(0)$, y_i , $y_i^R(0)$
0.7565	0.3231, 1.2406 , 1.3447	1.7601	1.4746, 1.6555, 2.1275
1.5919	1.2573, 1.8882 , 1.9680	2.2621	2.0940, 2.1728, 2.3278
1.8179	1.2556, 2.3321 , 2.5194	1.7656	1.6058, 1.7282, 1.9650
1.9567	1.3018, 2.4139 , 2.5818	2.1494	2.0573, 2.1543, 2.3650
2.5648	2.5211, 2.9140 , 2.9664	2.6245	2.3058, 2.4877, 2.8181
3.0342	2.9805, 3.4213 , 3.4805	1.6640	1.4937, 1.5565, 1.7410
3.2275	3.0720, 3.2665 , 3.3348	3.2191	2.7777, 2.8949, 3.2456
6.3359	6.1947, 6.3425 , 6.3749	2.9478	2.7850, 2.9219, 3.2946
6.7360	6.7275, 7.2360 , 7.3431	3.7295	3.5826, 3.6857, 3.8988
7.1145	6.6109, 7.3448 , 7.6373	4.0084	3.7039, 3.8085, 4.1479
7.8404	7.1400, 8.1058 , 8.4711	4.2982	3.9067, 4.0014, 4.3622
7.8960	7.4153, 8.3024 , 8.6258	3.5498	3.2025, 3.3850, 3.7018
8.1148	7.9959, 8.2206 , 8.2831	4.0070	3.6848, 3.8306, 4.4116
9.3564	9.3091, 9.4971 , 9.5408	4.2708	4.1486, 4.2151, 4.4590
9.7960	9.3866, 10.2224, 10.3177	4.6574	4.3367, 4.4535, 4.8523

Bestimmtheitsmaß und Residuen

	WAHR	MOD	END $x_i^L(0)$, $x_i^R(0)$	DFUZ	FFUZ	RINP	ROUT
R^2	0.8605	0.8763	0.8613	0.8672	0.8735	0.8754	0.8677
	-0.0103	-0.0603	0.0927 , -0.2030	-0.1263	-0.0543	-0.0641	-0.0963
	0.2371	0.2629	0.3395 , 0.1337	0.2462	0.2467	0.2493	0.2426
	-0.3282	-0.3146	-0.1046, -0.4705	-0.3047	-0.3138	-0.3025	-0.3296
	0.0132	0.0870	0.3082 , -0.0624	0.1225	0.1379	0.1113	0.1524
	0.3029	0.2705	0.2885 , 0.1596	0.2768	0.2516	0.2514	0.2481
	-0.8006	-0.8126	-0.7757, -0.9205	-0.8370	-0.8397	-0.8388	-0.8428
	0.6957	0.5722	0.5362 , 0.4602	0.5705	0.5567	0.5569	0.5525
	-0.5232	-0.3223	-0.3409, -0.3930	-0.3683	-0.3463	-0.3420	-0.3397
	0.1366	0.1738	0.2687 , 0.0905	0.1516	0.1485	0.1486	0.1579
	0.3001	0.2641	0.4253 , 0.1281	0.2752	0.2720	0.2683	0.2869
	0.3687	0.2290	0.4650 , 0.0796	0.2593	0.2613	0.2523	0.2833
	-0.3967	-0.4464	-0.2311, -0.5816	-0.4644	-0.4579	-0.4596	-0.4655
	-0.0062	0.0238	0.0464 , -0.0367	0.2050	0.1350	0.1063	0.0673
	-0.1208	0.0259	0.0507 , -0.0164	-0.0772	-0.0062	-0.0378	-0.0209
	0.1317	0.0470	0.2666 , -0.0030	0.1276	0.0943	0.0861	0.1265

In- und Outputs mit korrelierten Fehlern

Datenwerte

x_i^*	$x_i^L(0)$, x_i , $x_i^R(0)$	y_i^*	$y_i^L(0)$, y_i , $y_i^R(0)$
0.5850	0.5701, 1.1244, 1.1823	1.2590	0.9154, 0.9893, 1.2676
1.5032	1.4205, 2.0205, 2.1029	1.7823	1.5133, 1.5236, 1.9625
1.7921	1.6209, 2.3000, 2.4418	2.4363	2.1139, 2.1824, 2.6059
1.9582	1.8796, 2.4600, 2.4771	2.1109	1.7468, 1.8600, 2.1884
2.8622	2.8174, 3.3165, 3.4683	2.5206	2.1811, 2.2934, 2.7160
3.2124	3.1440, 3.6380, 3.7625	2.8268	2.4929, 2.6140, 2.8813
3.6208	3.6122, 4.0000, 4.1174	2.7654	2.4509, 2.5758, 2.8699
4.8865	4.8427, 4.9370, 5.1749	3.0606	2.9699, 3.0354, 3.2002
5.5303	5.2535, 5.3394, 5.5530	3.1493	2.9994, 3.2447, 3.2767
5.8729	5.4233, 5.5889, 5.9727	2.9113	2.8315, 3.0533, 3.1305
6.1996	5.7074, 5.8515, 6.3071	3.3109	3.1512, 3.4850, 3.4870
6.5940	6.0541, 6.1913, 6.6434	3.3498	3.2731, 3.5512, 3.5630
8.3185	7.7379, 7.8077, 8.3742	4.0977	4.0427, 4.3531, 4.3925
8.9830	8.3865, 8.4534, 9.0342	3.7490	3.5680, 4.0138, 4.1089
9.4416	8.7951, 8.9021, 9.4440	4.0448	3.9107, 4.3145, 4.3255

Bestimmtheitsmaß und Residuen

	WAHR	MOD	END $x_i^L(0)$, $x_i^R(0)$	DFUZ	FFUZ	RINP	ROUT
R^2	0.9031	0.9503	0.9370	0.9337	0.9368	0.9388	0.9439
	-0.5151	-0.3914	-0.3264, -0.5530	-0.4545	-0.4584	-0.4541	-0.4660
	-0.2378	-0.2213	-0.1068, -0.3593	-0.2127	-0.2080	-0.1979	-0.2075
	0.3388	0.3239	0.4778 , 0.1740	0.3688	0.3606	0.3585	0.3595
	-0.0310	-0.0634	0.0597 , -0.1614	-0.1056	-0.0862	-0.0693	-0.0947
	0.1365	0.0219	0.1461 , -0.0948	0.1260	0.0964	0.0905	0.0943
	0.3490	0.2118	0.3458 , 0.1170	0.2397	0.2330	0.2314	0.2376
	0.1781	0.0264	0.1343 , -0.0526	0.0971	0.0700	0.0629	0.0778
	0.1344	0.1052	0.1386 , 0.0156	0.1244	0.1220	0.1295	0.1259
	0.0506	0.1510	0.1959 , 0.0851	0.1224	0.1137	0.1152	0.1123
	-0.2791	-0.1418	-0.0584, -0.2617	-0.1731	-0.1712	-0.1700	-0.1751
	0.0330	0.1831	0.2682 , 0.0463	0.1342	0.1331	0.1243	0.1331
	-0.0337	0.1112	0.2061 , -0.0120	0.0925	0.0968	0.0865	0.0988
	0.2523	0.2562	0.3849 , 0.1495	0.2753	0.2784	0.2711	0.2838
	-0.2744	-0.3455	-0.1944, -0.4341	-0.3514	-0.3373	-0.3346	-0.3281
	-0.1014	-0.2272	-0.0449, -0.2850	-0.2192	-0.2079	-0.2020	-0.1984

Literaturverzeichnis

[Angrist und Krueger 1999] Angrist, Joshua D. ; Krueger, Alan B.: Empirical strategies in labor economics. In: Ashenfelter, O. (Hrsg.) ; Card, D. (Hrsg.): *Handbook of labor economics* Bd. 3. North Holland : Elsevier, 1999, Kap. 23, S. 1277–1367

[Avenhaus und Seising 1999] Avenhaus, Rudolf ; Seising, Rudolf: Fuzzy Theorie — eine Alternative zur Stochastik? Eine Podiumsdiskussion. In: [Seising, 1999], Kap. 12, S. 1–37

[Baltagi 1998] Baltagi, Badi H.: *Econometrics*. Springer-Verlag, 1998

[Bandemer 1997] Bandemer, Hans: *Ratschläge zum mathematischen Umgang mit Ungewißheit*. Leipzig : Teubner, 1997

[Bandemer 1999] Bandemer, Hans: Unscharfe Analyse unscharfer Daten. In: [Seising, 1999], Kap. 11, S. 251–267

[Bandemer und Gottwald 1993] Bandemer, Hans ; Gottwald, Siegfried: *Einführung in Fuzzy-Methoden*. 4. Berlin : Akademie Verlag, 1993

[Bandemer und Kudra 1988] Bandemer, Hans ; Kudra, Matthias: Some methods of evaluating functional relationships from fuzzy observations. In: Bandemer, Hans (Hrsg.): *Some applications of fuzzy set theory in data analysis*. Leipzig : VEB Deutscher Verlag für Grundstoffindustrie, 1988 (Freiberger Forschungshefte D 187), S. 46–61

[Bandemer und Näther 1992] Bandemer, Hans ; Näther, Wolfgang: *Fuzzy data analysis*. Dordrecht : Kluwer, 1992

[Bárdossy 1990] Bárdossy, András: Note on fuzzy regression. In: *Fuzzy sets and systems* 37 (1990), S. 65–75

[Bárdossy u. a. 1992] Bárdossy, András ; Hagaman, Roberta ; Duckstein, Lucien ; Bogárdi, István: Fuzzy least squares regression: theory and application. In: [Kacprzyk und Fedrizzi, 1992], S. 181–193

[Bender u. a. 1996] Bender, Stefan ; Hilzendegen, Jürgen ; Rohwer, Götz ; Rudolph, Helmut: *Die IAB-Beschäftigtenstichprobe 1975–1990*. Nürnberg : Institut für Arbeitsmarkt- und Berufsforschung, 1996 (Beiträge zur Arbeitsmarkt- und Berufsforschung 197)

[Bertoluzza u. a. 1995] Bertoluzza, Carlo ; Corral, Norberto ; Salas, Antonia: On a new class of distances between fuzzy numbers. In: *Mathware & soft computing* 2 (1995), S. 71–84

[Borgelt u. a. 1999] Borgelt, Christian ; Gebhardt, Jörg ; Kruse, Rudolf: Fuzzy Methoden in der Datenanalyse. In: [Seising, 1999], Kap. 17, S. 370–386

[Brinner 2003] Brinner, Karin: Auswirkungen von Erhebungsungenauigkeiten auf die amtliche Mortalitätsmessung. In: *Jahrbücher für Nationalökonomie und Statistik* 223 (2003), Nr. 4, S. 385–402

[Celmiņš 1987a] Celmiņš, Aivars: Least squares model fitting to fuzzy vector data. In: *Fuzzy sets and systems* 22 (1987), S. 245–269

[Celmiņš 1987b] Celmiņš, Aivars: Multidimensional least-squares fitting of fuzzy models. In: *Mathematical modelling* 9 (1987), Nr. 9, S. 669–690

[Celmiņš 1992] Celmiņš, Aivars: Nonlinear least squares regression in fuzzy vector spaces. In: [Kacprzyk und Fedrizzi, 1992], S. 152–180

[Chang und Lee 1994a] Chang, Ping-Teng ; Lee, E. Stanley: Fuzzy least absolute deviations regression and the conflicting trends in fuzzy parameters. In: *Computers and mathematics with applications* 28 (1994), Nr. 5, S. 89–101

[Chang und Lee 1994b] Chang, Ping-Teng ; Lee, E. Stanley: Fuzzy linear regression with spreads unrestricted in sign. In: *Computers and mathematics with applications* 28 (1994), Nr. 4, S. 61–70

[Chang u. a. 1996] Chang, Ping-Teng ; Lee, E. Stanley ; Konz, Stephan A.: Applying fuzzy linear regression to VDT legibility. In: *Fuzzy sets and systems* 80 (1996), S. 197–204

[Chang 2001] Chang, Yun-Hsi O.: Hybrid fuzzy least-squares regression analysis and its reliability measures. In: *Fuzzy sets and systems* 119 (2001), S. 225–246

[Chang und Ayyub 2001] Chang, Yun-Hsi O. ; Ayyub, Bilal M.: Fuzzy regression methods — a comparative assessment. In: *Fuzzy sets and systems* 119 (2001), S. 187–203

[Chen 2001] Chen, Yun-Shiow: Outliers detection and confidence interval modification in fuzzy regression. In: *Fuzzy sets and systems* 119 (2001), S. 259–272

[Chlumsky u. a. 1993] Chlumsky, Jürgen ; Wiegert, Rolf ; et al.: *Qualität statistischer Daten: Beiträge zum wissenschaftlichen Kolloquium am 12./13. November 1992 in Wiesbaden.* Stuttgart : Metzler-Poeschel, 1993

[Coppi u. a. 2006] Coppi, Renato ; D'Urso, Pierpaolo ; Giordani, Paolo ; Santoro, Adriana: Least squares estimation of a linear regression model with *LR* fuzzy response. In: *Computational statistics & data analysis* 51 (2006), S. 267–286

[Couso u. a. 2007] Couso, Ines ; Dubois, Didier ; Montes, Susana ; Sanchez, Luciano: On various definitions of the variance of a fuzzy random variable. In: De Cooman, Gert (Hrsg.) ; Vejnarová, Jiřina (Hrsg.) ; Zaffalon, Marco (Hrsg.): *ISIPTA '07 — Proceedings of the Fifth International Symposium on Imprecise Probability: Theory and Applications.* Prag : Action M Agency for SIPTA (Society for Imprecise Probability: Theories and Applications), 2007

(http://www.sipta.org/isipta07/proceedings/056.html (Zugriff am 31.8.2007)), S. 135–144

[Cramer 1985] Cramer, Ulrich: Probleme der Genauigkeit der Beschäftigtenstatistik. In: *Allgemeines statistisches Archiv* 69 (1985), S. 56–68

[Cramer und Majer 1991] Cramer, Ulrich ; Majer, Wolfgang: Ist die Beschäftigtenstatistik revisionsbedürftig? In: *Mitteilungen aus der Arbeitsmarkt- und Berufsforschung* 24 (1991), Nr. 1, S. 81–90

[Diamond 1990] Diamond, Phil: Higher level fuzzy numbers arising from fuzzy regression models. In: *Fuzzy sets and systems* 36 (1990), Nr. 2, S. 265–275

[Diamond 1992] Diamond, Phil: Least squares and maximum likelihood regression for fuzzy linear models. In: [Kacprzyk und Fedrizzi, 1992], S. 137–151

[Diamond und Kloeden 1994] Diamond, Phil ; Kloeden, Peter: *Metric spaces of fuzzy sets: theory and applications.* Singapore : World scientific, 1994

[Diamond und Körner 1997] Diamond, Phil ; Körner, Ralph: Extended fuzzy linear models and least squares estimates. In: *Computers and mathematics with applications* 33 (1997), Nr. 9, S. 15–32

[Diamond und Tanaka 1999] Diamond, Phil ; Tanaka, Hideo: Fuzzy regression analysis. In: Słowiński, Roman (Hrsg.): *Fuzzy sets in decision analysis, operations research and statistics.* Boston : Kluwer Academic Publishers, 1999 (The handbooks of fuzzy sets), Kap. 11, S. 349–387

[Dubois u. a. 2000] Dubois, Didier ; Nguyen, Hung T. ; Prade, Henri: *The handbook of Fuzzy sets series.* Bd. 7: *Possibility-Theory, probability, and fuzzy sets: misunderstandings, bridges and gaps.* Kap. 7, S. 343–438. Siehe [Dubois und Prade, 2000]

[Dubois und Prade 1980] Dubois, Didier ; Prade, Henri: *Fuzzy sets and systems: theory and applications.* New York : Academic Press, 1980

[Dubois und Prade 2000] Dubois, Didier ; Prade, Henri: *The handbook of Fuzzy sets series.* Bd. 7: *Fundamentals of fuzzy sets.* Boston : Kluwer, 2000

[Federal Committee on Statistical Methodology 2001] Federal Committee on Statistical Methodology: *Measuring and reporting sources of error in surveys.* Statistical working paper 31. 2001

[Fritsch 1997] Fritsch, Michael: Die Betriebsdatei der Beschäftigtenstatistik: Ansatz und Analysepotentiale. In: Hujer, Reinhard (Hrsg.) ; Rendtel, Ulrich (Hrsg.) ; Wagner, Gerd (Hrsg.): *Wirtschafts- und sozialwissenschaftliche Panel-Studien: Datenstrukturen und Analyseverfahren.* Göttingen : Vandenhoeck & Ruprecht, 1997 (Sonderhefte zum Allgemeinen statistischen Archiv 30), S. 127–147

[Fritsch u. a. 2002] Fritsch, Michael ; Grotz, Reinhold ; Brixy, Udo ; Niese, Michael ; Otto, Anne: Die statistische Erfassung von Gründungen in Deutschland: ein Vergleich von Beschäftigtenstatistik, Gewerbeanzeigenstatistik, und den Mannheimer Gründungspanels. In: *Allgemeines statistisches Archiv* 86 (2002), S. 87–96

[Froeschl 1999] Froeschl, Karl A.: Metadata Management in official statistics: an IT-based methodology approach. In: *Austrian journal of statistics* 28 (1999), Nr. 2, S. 49–79

[Frohn 1993] Frohn, Joachim: Anforderungen an die Qualität statistischer Daten. In: *Qualität statistischer Daten: Beiträge zum wissenschaftlichen Kolloquium am 12./13. November 1992 in Wiesbaden.* [Chlumsky u. a., 1993], S. 55–63

[Georgescu 1998] Georgescu, Vasile: New estimation methods in fuzzy regression analysis, based on projection theorem and decoupling principle. In: *Fuzzy economic review* III (1998), Nr. 1, S. 21–38

[Griliches 1986] Griliches, Zvi: Economic data issues. In: Griliches, Zvi (Hrsg.) ; Intriligator, Michael D. (Hrsg.): *Handbook of econometrics* Bd. III. North Holland : Elsevier, 1986, Kap. 25, S. 1465–1514

[Heshmaty und Kandel 1985] Heshmaty, Behrooz ; Kandel, Abraham: Fuzzy linear regression and its applications to forecasting in uncertain environment. In: *Fuzzy sets and systems* 15 (1985), Nr. 2, S. 159–191

[Hildreth und Houck 1968] Hildreth, Clifford ; Houck, James P.: Some estimators for a linear model with random coefficients. In: *Journal of the american statistical association* 63 (1968), Nr. 322, S. 584–595

[Holub und Tappeiner 1995] Holub, Hans W. ; Tappeiner, Gottfried: Ex post, ex ante und dazwischen: Überlegungen zu einem wenig beachteten Problem der Empirischen Wirtschaftsforschung. In: *Jahrbücher für Nationalökonomie und Statistik* 214 (1995), Nr. 6, S. 663–673

[Holub und Tappeiner 1997] Holub, Hans W. ; Tappeiner, Gottfried: Modeling on the basis of models. In: *Review of income and wealth* 43 (1997), Nr. 4, S. 505–510

[Höppner u. a. 1997] Höppner, Frank ; Klawonn, Frank ; Kruse, Rudolf: *Fuzzy Clusteranalyse.* Vieweg, 1997

[Inuiguchi 2003] Inuiguchi, Masahiro: Robust interval regression methods based on Minkowski difference. In: Ramík, Jaroslav (Hrsg.) ; Novák, Vilém (Hrsg.): *Methods for decision support in environment with uncertainty — applications in economics, business and engineering: Proceedings of the Czech-Japan Seminar held in Valtice 20.9.–23.9.2003.* Ostrava : University of Ostrava, Institute for Research and Applications of Fuzzy Modeling, 2003 (Research report No. 49), S. 26–31

[Ishibuchi und Nii 2001] Ishibuchi, Hisao ; Nii, Manabu: Fuzzy regression using asymmetric fuzzy coefficients and fuzzified neural networks. In: *Fuzzy sets and systems* 119 (2001), S. 273–290

[József 1992] József, Sándor: On the effect of linear data transformations in possibilistic fuzzy linear regression. In: *Fuzzy sets and systems* 45 (1992), S. 185–188

[Kacprzyk und Fedrizzi 1992] Kacprzyk, Janusz (Hrsg.) ; Fedrizzi, Mario (Hrsg.): *Fuzzy regression analysis.* Heidelberg : Physika-Verlag, 1992

[Kao und Chyu 2002] Kao, Chiang ; Chyu, Chin-Lu: A fuzzy linear regression model with better explanatory power. In: *Fuzzy sets and systems* 126 (2002), S. 401–409

[Kim und Bishu 1998] Kim, Byungjoon ; Bishu, Ram R.: Evaluation of fuzzy linear regression models by comparing membership functions. In: *Fuzzy sets and systems* 100 (1998), S. 343–352

[Kim u. a. 1996] Kim, Kwang J. ; Moskowitz, Herbert ; Koksalan, Murat: Fuzzy versus statistical regression. In: *European journal of operational research* 92 (1996), S. 417–434

[Körner 1997] Körner, Ralph: *Linear models with random fuzzy variables*, Fakultät für Mathematik und Informatik der Technischen Bergakademie Freiberg, Dissertation, 1997

[Körner und Näther 1998] Körner, Ralph ; Näther, Wolfgang: Linear regression with random fuzzy variables: extended classical estimates, best linear estimates, least squares estimates. In: *Journal of information sciences* 109 (1998), S. 95–118

[Krätschmer 2001a] Krätschmer, Volker: *Induktive Statistik auf Basis unscharfer Messkonzepte am Beispiel linearer Regressionsmodelle*, Universität des Saarlandes, Habilitationsschrift, 2001

[Krätschmer 2001b] Krätschmer, Volker: A unified approach to fuzzy-random-variables. In: *Fuzzy sets and systems* 123 (2001), S. 1–9

[Krätschmer 2004] Krätschmer, Volker: Probability theory in fuzzy sample spaces. In: *Metrika* 60 (2004), S. 167–189

[Krätschmer 2006a] Krätschmer, Volker: Least-squares estimation in linear regression models with vague concepts. In: *Fuzzy sets and systems* 157 (2006), S. 2579–2592

[Krätschmer 2006b] Krätschmer, Volker: Limit distributions of least squares estimators in linear regression models with vague concepts. In: *Journal of multivariate analysis* 97 (2006), S. 1044–1069

[Krätschmer 2006c] Krätschmer, Volker: Strong consistency of least-squares estimation in linear regression models with vague concepts. In: *Journal of multivariate analysis* 97 (2006), S. 633–654

[Krug u. a. 2001] Krug, Walter ; Nourney, Martin ; Schmidt, Jürgen: *Wirtschafts- und Sozialstatistik: Gewinnung von Daten*. 6. München : Oldenbourg, 2001

[Kruse u. a. 1993] Kruse, Rudolf ; Gebhardt, Jörg ; Klawonn, Frank: *Fuzzy-Systeme*. Stuttgart : Teubner, 1993

[Kruse und Meyer 1987] Kruse, Rudolf ; Meyer, Klaus D.: *Statistics with vague data*. Dordrecht : Reidel, 1987

[Kudra 1990] Kudra, Matthias: Comparison of methods to evaluate fuzzy parameters in explicit functional relationships. In: Bandemer, Hans (Hrsg.): *Some applications of fuzzy set theory in data analysis II*. Leipzig : VEB Deutscher Verlag für Grundstoffindustrie, 1990 (Freiberger Forschungshefte D 197), S. 43–58

[Löbbe 1993] Löbbe, Klaus: Qualitätsansprüche der angewandten Wirtschaftsforschung. In: *Qualität statistischer Daten: Beiträge zum wissenschaftlichen Kolloquium am 12./13. November 1992 in Wiesbaden.* [Chlumsky u. a., 1993], S. 46–54

[Lüken 2002] Lüken, Stephan: *Spektrum Bundesstatistik.* Bd. 19: *Zur Fortentwicklung des Systems der Erwerbstätigenstatistiken: Bericht im Auftrag des Statistischen Beirats.* Reutlingen : Metzler & Poeschel, 2002

[Manton u. a. 1994] Manton, Kenneth G. ; Woodbury, Max A. ; Tolley, H. Dennis: *Statistical applications using fuzzy sets.* New York : Wiley, 1994

[Mayer 1993] Mayer, Thomas: *Truth versus precision in economics.* Aldershot : Edward Elgar, 1993

[Miranda u. a. 2005] Miranda, Enrique ; Couso, Inés ; Gil, Pedro: Random sets as inprecise random variables. In: *Journal of mathematical analysis and applications* 307 (2005), S. 32–47

[Morgenstern 1965] Morgenstern, Oskar: *Über die Genauigkeit wirtschaftlicher Beobachtungen.* 2. Wien : Physika, 1965

[Moskowitz und Kim 1993] Moskowitz, Herbert ; Kim, Kwangjae: On assessing the H value in fuzzy linear regression. In: *Fuzzy sets and systems* 58 (1993), S. 303–327

[Näther 1994] Näther, Wolfgang: *Linear statistical inference with random fuzzy data.* Preprint. 1994

[Näther 2000] Näther, Wolfgang: On random fuzzy variables of second order and their application to linear statistical inference with fuzzy data. In: *Metrika* 51 (2000), S. 201–221

[Näther 2001] Näther, Wolfgang: Random fuzzy variables of second order and applications to statistical inference. In: *Information sciences* 133 (2001), S. 69–88

[Näther 2006] Näther, Wolfgang: Regression with fuzzy random data. In: *Computational statistics & data analysis* 51 (2006), S. 235–252

[Neubauer 1993] Neubauer, Werner: Qualität statistischer Daten: zur Diskussion um einen Kriterienkatalog. In: *Qualität statistischer Daten: Beiträge zum wissenschaftlichen Kolloquium am 12./13. November 1992 in Wiesbaden.* [Chlumsky u. a., 1993], S. 12–25

[Newbold und Bos 1985] Newbold, Paul ; Bos, Theodore: *Quantitative applications in the social sciences.* Bd. 51: *Stochastic parameter regression models.* Beverly Hills : Sage Publications, 1985

[Niskanen 2001] Niskanen, Vesa A.: Prospects for soft statistical computing: describing data and inferring from data with words in the Human Sciences. In: *Information sciences* 132 (2001), S. 83–131

[Pesaran und Smith 1992] Pesaran, Hashem M. ; Smith, Ron: The interaction between theory and observation in economics. In: *The economic and social review* 24 (1992), Nr. 1, S. 1–23

[Peters 1994] Peters, Georg: Fuzzy linear regression with fuzzy intervals. In: *Fuzzy sets and systems* 63 (1994), S. 45–55

[Puri und Ralescu 1986] Puri, Madan L. ; Ralescu, Dan A.: Fuzzy Random Variables. In: *Journal of mathematical analysis and applications* 114 (1986), S. 409–422

[Redden und Woodall 1996] Redden, David T. ; Woodall, William H.: Further examination of fuzzy linear regression. In: *Fuzzy sets and systems* 79 (1996), S. 203–211

[Richter 2002] Richter, Josef: *Kategorien und Grenzen der empirischen Verankerung der Wirtschaftsforschung*. Stuttgart : Lucius und Lucius, 2002

[Rommelfanger 1994] Rommelfanger, Heinrich: *Fuzzy Decision Support-Systeme: Entscheiden bei Unschärfe*. 2. Berlin : Springer, 1994

[Rosenberg 1973] Rosenberg, Barr: Linear regression with randomly dispersed parameters. In: *Biometrika* 60 (1973), Nr. 1, S. 65–72

[Sakawa und Yano 1992a] Sakawa, Masatoshi ; Yano, Hitoshi: Fuzzy linear regression and its applications. In: [Kacprzyk und Fedrizzi, 1992], S. 61–80

[Sakawa und Yano 1992b] Sakawa, Masatoshi ; Yano, Hitoshi: Multiobjective fuzzy linear regression analysis for fuzzy input-output data. In: *Fuzzy sets and systems* 47 (1992), S. 173–181

[Savic und Pedrycz 1991] Savic, Dragan A. ; Pedrycz, Witold: Evaluation of fuzzy linear regression models. In: *Fuzzy sets and systems* 23 (1991), S. 51–63

[Savic und Pedrycz 1992] Savic, Dragan A. ; Pedrycz, Witold: Fuzzy linear regression models: construction and evaluation. In: [Kacprzyk und Fedrizzi, 1992], S. 91–100

[Schmähl und Fachinger 1994] Schmähl, Winfried ; Fachinger, Uwe: Prozessproduzierte Daten als Grundlage für sozial- und verteilungspolitische Analysen: Erfahrungen mit Daten der Rentenversicherungsträger für Längsschnittananalysen. In: Hauser, Richard (Hrsg.) ; Ott, Notburga (Hrsg.) ; Wagner, Gerd (Hrsg.): *Mikroanalytische Grundlagen der Gesellschaftspolitik: Erhebungsverfahren, Analysemethoden und Mikrosimulation* Bd. 2. Berlin : Akademie Verlag, 1994, S. 179–200

[Schmucker und Seth 2006] Schmucker, Alexandra ; Seth, Stefan: *BA-Beschäftigtenpanel 1998–2005: Codebuch*. Bundesagentur für Arbeit, 2006 (FDZ-Datenreport: Dokumentation zu Arbeitsmarktdaten 05)

[Schmucker und Seth 2009] Schmucker, Alexandra ; Seth, Stefan: *BA-Beschäftigtenpanel 1998–2007: Codebuch*. Bundesagentur für Arbeit, 2009 (FDZ-Datenreport: Dokumentation zu Arbeitsmarktdaten 01)

[Schneeweiß 1990] SCHNEEWEISS, Hans: *Ökonometrie*. 4. Heidelberg : Physika-Verlag, 1990

[Schneeweiß und Mittag 1986] SCHNEEWEISS, Hans ; MITTAG, Hans-Joachim: *Lineare Modelle mit fehlerbehafteten Daten*. Wien : Physika-Verlag, 1986

[Seising 1999] SEISING, Rudolf (Hrsg.): *Fuzzy-Theorie und Stochastik: Modelle und Anwendungen in der Disskussion*. Braunschweig : Vieweg, 1999

[Sen und Srivastava 1990] SEN, Ashish ; SRIVASTAVA, Muni: *Regression analysis: theory, methods, and applications*. New York : Springer, 1990

[Solymosi und Dombi 1992] SOLYMOSI, Tamás ; DOMBI, József: Fitting functions to data with error bounds: fuzzy regression with ERRGO. In: [Kacprzyk und Fedrizzi, 1992], S. 101–115

[Statistische Ämter des Bundes und der Länder 2006] STATISTISCHE ÄMTER DES BUNDES UND DER LÄNDER (Hrsg.): *Die Qualitätsstandards der amtlichen Statistik*. Wiesbaden : Statistisches Bundesamt, 2006

[Statistisches Bundesamt 2009] STATISTISCHES BUNDESAMT (Hrsg.): *Qualitätsbericht: Statistik der sozialversicherungspflichtig Beschäftigten — vierteljährliche Beschäftigtenstatistik*. Wiesbaden : Statistisches Bundesamt, Januar 2009

[Statistisches Bundesamt vierteljährlich] STATISTISCHES BUNDESAMT (Hrsg.): *Bevölkerung und Erwerbstätigkeit: Struktur der sozialversicherungspflichtigen Beschäftigten*. Wiesbaden : Statistisches Bundesamt, vierteljährlich (Fachserie 1, Reihe 4.2.1)

[Steiner und Wagner 1997] STEINER, Viktor ; WAGNER, Kersten: Entwicklung der Ungleichheit der Erwerbseinkommen in Westdeutschland: eine Einführung zu den Workshop-Beiträgen. In: *Mitteilungen aus der Arbeitsmarkt- und Berufsforschung* 30 (1997), Nr. 3, S. 638–641

[Strecker 1993] STRECKER, Heinrich: Fehler in statistischen Erhebungen, deren Messung und ihr Einfluss auf die Datenqualität. In: *Qualität statistischer Daten: Beiträge zum wissenschaftlichen Kolloquium am 12./13. November 1992 in Wiesbaden*. [Chlumsky u. a., 1993], S. 26–35

[Strecker 2004] STRECKER, Heinrich: Der Nicht-Stichprobenfehler und seine Zerlegung in die beiden Komponenten „glatter Wert“ — wahrer Wert plus systematischer Fehler — und „Zufallsfehler“: Schätzung mit Hilfe der Variate Difference-Methode und Analyse der horizontalen Aggregation individueller Zufallsfehler. In: *Jahrbücher für Nationalökonomie und Statistik* 224 (2004), Nr. 1+2, S. 198–230

[Strecker und Wiegert 1994] STRECKER, Heinrich ; WIEGERT, Rolf: Wirtschaftsstatistische Daten und ökonomische Realität. In: STRECKER, Heinrich (Hrsg.) ; WIEGERT, Rolf (Hrsg.): *Stichproben, Erhebungsfehler, Datenqualität*. Göttingen : Vendenhoeck & Ruprecht, 1994, S. 101–123. – (Nachdruck von Strecker, H., Wiegert, R. (1989). Wirtschaftsstatistische Daten und ökonomische Realität. *Jahrbücher für Nationalökonomie und Statistik*, 206(4–5):487–509.)

[Tanaka 1987] Tanaka, Hideo: Fuzzy data analysis by possibilistic linear models. In: *Fuzzy sets and systems* 24 (1987), Nr. 3, S. 363–375

[Tanaka und Ishibuchi 1991] Tanaka, Hideo ; Ishibuchi, Hisao: Identification of possibilistic linear systems by quadratic membership functions of fuzzy parameters. In: *Fuzzy sets and systems* 41 (1991), Nr. 2, S. 145–160

[Tanaka und Ishibuchi 1992] Tanaka, Hideo ; Ishibuchi, Hisao: Possibilistic regression analysis based on linear programming. In: [Kacprzyk und Fedrizzi, 1992], S. 47–60

[Tanaka u. a. 1995] Tanaka, Hideo ; Ishibuchi, Hisao ; Yoshikawa, S.: Exponential possibility regression analysis. In: *Fuzzy sets and systems* 69 (1995), S. 305–318

[Tran und Duckstein 2002] Tran, Liem ; Duckstein, Lucien: Multiobjective fuzzy regression with central tendency and possibilistic properties. In: *Fuzzy sets and systems* 130 (2002), S. 21–31

[Tsaur und Wang 1999] Tsaur, Ruey-Chyn ; Wang, Hsiao-Fan: An alternative approach to fuzzy regression model. In: IFSA (Hrsg.): *IFSA Proceedings* Bd. 1, 1999, S. 420–424

[D'Urso und Gastaldi 2000] D'Urso, Pierpaolo ; Gastaldi, Tommaso: A least-squares approach to fuzzy linear regression analysis. In: *Computational statistics & data analysis* 34 (2000), S. 427–440

[Wang und Tsaur 2000] Wang, Hsiao-Fan ; Tsaur, Ruey-Chyn: Insight of a fuzzy regression model. In: *Fuzzy sets and systems* 112 (2000), S. 355–369

[Wansbeek und Meijer 2003] Wansbeek, Tom ; Meijer, Erik: Measurement errors and latent variables. In: Baltagi, Badi H. (Hrsg.): *A companion to theoretical econometrics*. Malden, MA, USA : Blackwell Publishing, 2003, Kap. 8, S. 162–179

[Wermter und Cramer 1988] Wermter, Winfried ; Cramer, Ulrich: Wie hoch war der Beschäftigtenanstieg seit 1983? In: *Mitteilungen aus der Arbeitsmarkt- und Berufsforschung* 21 (1988), Nr. 4, S. 468–482

[Wünsche und Näther 2002] Wünsche, Andreas ; Näther, Wolfgang: Least-squares fuzzy regression with fuzzy random variables. In: *Fuzzy sets and systems* 130 (2002), S. 43–50

[Xu und Li 2001] Xu, Ruoning ; Li, Chulin: Multidimensional least-squares fitting with a fuzzy model. In: *Fuzzy sets and systems* 119 (2001), S. 215–223

[Yang und Lin 2002] Yang, Miin-Shen ; Lin, Tzu-Shun: Fuzzy least-squares linear regression analysis for fuzzy input-output data. In: *Fuzzy sets and systems* 126 (2002), S. 389–399

[Yang und Liu 2003] Yang, Miin-Shen ; Liu, Hsien-Hsiung: Fuzzy least-squares algorithms for interactive fuzzy linear regression models. In: *Fuzzy sets and systems* 135 (2003), S. 305–316

[Zadeh 1965] Zadeh, Lotfi A.: Fuzzy sets. In: *Information and control* 8 (1965), S. 338–353

[Zadeh 1997] Zadeh, Lotfi A.: Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. In: *Fuzzy sets and systems* 90 (1997), S. 111–127

[Zimmermann 1996] Zimmermann, Hans-Jürgen: *Fuzzy set theory and its applications*. 3. Boston : Kluwer, 1996

Symbolverzeichnis

$\|\cdot\|^2$	euklidische Norm im $\mathbb{R}^n$, Seite 32
$\odot$	erweiterte (Skalar-)Multiplikation, Seite 59
$\ominus_H$	Hukuhara-Subtraktion, Seite 78
$\oplus$	erweiterte Addition, Seite 59
1_A	Indikatorfunktion für die Menge A, Seite 49
$(a^L, a^R; l_A, r_A)_{LR}$	LR-Fuzzy-Intervall mit Spannweiten l_A und r_A, Seite 51
(a_0^L, a^L, a^R, a_0^R)	Tupelschreibweise für trapezförmige Fuzzy-Mengen, Seite 52
$[\widetilde{A}]^\alpha$	α-Niveau der Fuzzy-Menge $\widetilde{A}$, Seite 52
$[a^L(\alpha), a^R(\alpha)]$	Intervall-Darstellung des α-Niveaus $[\widetilde{A}]^\alpha$ eines LR-Fuzzy-Intervalls mit linkem Rand $a^L(\alpha)$ und rechtem Rand $a^R(\alpha)$, Seite 53
$\widetilde{A}, \widetilde{B}$	Fuzzy-Mengen, Seite 49
$\widetilde{A} \times \widetilde{B}$	kartesisches Produkt von $\widetilde{A} \in \mathcal{F}(U)$ und $\widetilde{B} \in \mathcal{F}(V)$, Seite 54
$\hat{\mathbf{a}} = (\hat{a}_0, \ldots, \hat{a}_k)$	Approximations- bzw. Schätzparameter, Seite 31
$\mathrm{Cov}(X, Y)$	Kovarianz der Zufallsvariablen X und Y, Seite 38
$\widetilde{C} = (\widetilde{C}_1, \ldots, \widetilde{C}_m)$	Vektorenschreibweise für nicht-interaktive Fuzzy-Vektoren in $F_{coc}^{nob}(\mathbb{R}^m)$, Seite 56
δ_2	L^2-Metrik der Kleinste Quadrate Fuzzy-Regression, Seite 96
$\delta_{(\nu)}$	erweiterte Familie von Kleinste Quadrate Fuzzy-Metriken, Seite 98
$D^2(\mathbf{a})$	Summe der Abstandsquadrate, Seite 31
$\mathrm{E}(\cdot \mid Z)$	bedingte Erwartung bei Vorliegen von Z, Seite 30

$\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)$	Vektor der Störvariablen, mit denen die zufälligen Abweichungen im Gleichungsmodell abgebildet werden, Seite 30
$e_i(\widetilde{A})$	Residuen der Kleinste Quadrate Fuzzy-Regression, Seite 93
$e_i(\mathbf{a}), e_i$	Residuen im Stichprobenergebnis $(\mathbf{x}_1, \ldots, \mathbf{x}_k, \mathbf{y})$, Seite 31
$F_{coc}^{nob}(\mathbb{R}^m)$	Menge aller Fuzzy-Mengen über $\mathbb{R}^m$, die normal, beschränkt und kompakt sind, Seite 55
$F_{coc}^{nob}(\mathbb{R})^m$	Menge aller nicht-interaktiven Fuzzy-Vektoren in $F_{coc}^{nob}(\mathbb{R}^m)$, Seite 56
$\mathcal{F}(U)$	Fuzzy-Potenzmenge über der Grundmenge U, Seite 49
$\widetilde{F} \in \mathrm{Map}(U, V)$	Fuzzy-Bündel von Funktionen, Seite 57
$f(\mathbf{z}, \mathbf{a})$	Schar der (linearen) Modellfunktionen in den allgemeinen Variablen $\mathbf{z} = (z_1, \ldots, z_k)$ und den Parametern $\mathbf{a} = (a_0, \ldots, a_k)$, Seite 28
$f[u, \widetilde{A}], f[\widetilde{U}, a], f[\widetilde{U}, \widetilde{A}]$	Kurzschreibweise für die Funktionstypen von Fuzzy-linearen Funktionen, Seite 62
$f^{\alpha,L}, f^{\alpha,R}$	Funktion der linken bzw. der rechten Intervallgrenze der erweiterten Funktion f auf dem Niveau α, Seite 79
$G^{\kappa}(x, \widetilde{A})$	Maßzahl für die aggregierte Gesamtunschärfe der Approximationsfunktion im Niveau κ, Seite 85
$\mathbf{I} \in \mathbb{R}^{n \times n}$	n-dimensionale Einheitsmatrix, Seite 30
$\jmath$	Abbildung zur Einbettung der Fuzzy-Mengen in $F_{coc}^{nob}(\mathbb{R})$ in die Menge ihrer Stützfunktionen, Seite 95
$k \in \mathbb{N}$	Anzahl der Modellvariablen, Seite 28
$\mathrm{Lin}(\mathbb{R}^k, \mathbb{R})$	Menge aller linearen Funktionen $f : \mathbb{R}^k \longrightarrow \mathbb{R}$, Seite 79
L, R	linke und rechte Referenzfunktion eines LR-Fuzzy-Intervalls, Seite 51
$\mathrm{Map}(U, V)$	Menge aller Funktionen $f : U \longrightarrow V$, Seite 57
$\mu_{\widetilde{A}}$	Zugehörigkeitsfunktion der Fuzzy-Menge $\widetilde{A}$, Seite 49
m^L, m^R	linke bzw. rechte Modifizierungsfunktion zur Kalibrierung von Fuzzy-Merkmalsdaten, Seite 136

$\mathbb{N}$	Menge der natürlichen Zahlen, Seite 29
$N(\mu, \sigma^2)$	Normalverteilung mit Erwartungswert μ und Varianz σ^2, Seite 33
$n \in \mathbb{N}$	Stichprobenumfang, Seite 29
Ω	Grundmenge der Ereignisse einer Zufallsvariablen, Seite 29
O_F	Defuzzifizierungsoperator, Seite 109
$\mathrm{Pr}_U \widetilde{C}$	Projektion von $\widetilde{C} \in \mathcal{F}(U \times V)$ auf U, Seite 54
$\mathbb{R}$	Menge der reellen Zahlen, Seite 28
$\mathbb{R}_{\mathbf{q}}$	kartesischer Orthant in Richtung $\mathbf{b}$, Seite 81
$\mathbb{S}^0$	Oberfläche der Einheitskugel im $\mathbb{R}^1$, Seite 95
$\sigma^2 = \mathrm{Var}\,\varepsilon$	Varianz der Störvariablen ε, Seite 30
$\sigma_{\widetilde{A}}$	Steiner-Punkt der Fuzzy-Menge $\widetilde{A}$, Seite 100
$\mathrm{supp}\,\widetilde{A}$	Trägermenge der Fuzzy-Menge $\widetilde{A}$, Seite 52
$s_{\widetilde{A}}$	Stützfunktion der Fuzzy-Menge $\widetilde{A}$, Seite 95
U, V	Grundmengen für die Konstruktion von Fuzzy-Mengen, Seite 49
U_i^*	ursprüngliche Variable, die durch die Fuzzy-Zufallsvariable $\widetilde{Y}_i$ beschrieben wird, Seite 114
$\chi_X, \chi_{\widetilde{X}}$	Auswertungsfunktional für reellwertige bzw. für Fuzzy-Inputvektoren, Seite 80
$\widetilde{X}^\circ$	im Steiner-Punkt zentrierte Fuzzy-Menge zu $\widetilde{X}$, Seite 170
$\mathbf{X}$	Stichprobenmatrix, Seite 30
$\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})$ $(1 \le j \le k)$	Stichprobenergebnis für die Beobachtungen $X_{ij} : \Omega \longrightarrow \mathbb{R}$ $(1 \le i \le n)$, Seite 29
$X_j, \widetilde{X}_j$ $(1 \le j \le k)$	unabhängige (Zufalls-)Variable bzw. Inputs, Seite 28
$X_j^*, \mathbf{x}_j^*, Y^*, \mathbf{y}^*$	(Zufalls-)Variable bzw. Stichprobenergebnis der „wahren" Werte, Seite 36

$\mathbf{y} = (y_1, \ldots, y_n)$	Stichprobenergebnis für die Beobachtungen $Y_i : \Omega \longrightarrow \mathbb{R}$ $(1 \leq i \leq n)$, Seite 29
$Y, \widetilde{Y}$	abhängige (Zufalls-)Variable bzw. Output, Seite 28
$\overline{z}.$	Mittelwert über das (beliebige) Stichprobenergebnis $\mathbf{z}$, Seite 32
$\mathrm{Zyl}_V\, \widetilde{A}$	zylindrische Erweiterung von $\widetilde{A} \in \mathcal{F}(U)$ auf $U \times V$, Seite 54