Alternative Methoden zur Biomasseschätzung auf Einzelbaumebene unter spezieller Berücksichtigung der *k*-Nearest Neighbour (*k*-NN) Methode

Dissertation zur Erlangung des Doktorgrades
der Fakultät für Forstwissenschaften und Waldökologie
der Georg-August-Universität Göttingen

vorgelegt von
Lutz Fehrmann
geboren in Göttingen

Göttingen, Oktober 2006

D7

1. Berichterstatter: Prof. Dr. Christoph Kleinn
2. Berichterstatter: Prof. Dr. Dr. h.c. Klaus von Gadow
3. Berichterstatter: Prof. Dr. Dr. h.c. Branislav Sloboda

Tag der mündlichen Prüfung: 07.12.2006

Diese Arbeit ist durch die Niedersächsische Staats- und Universitätsbibliothek, SUB-Göttingen, unter folgender Internetadresse elektronisch veröffentlicht:

http://resolver.sub.uni-goettingen.de/purl/?webdoc-1390
bzw.:
http://webdoc.sub.gwdg.de/diss/2007/fehrmann/

Inhalt

Vorwort

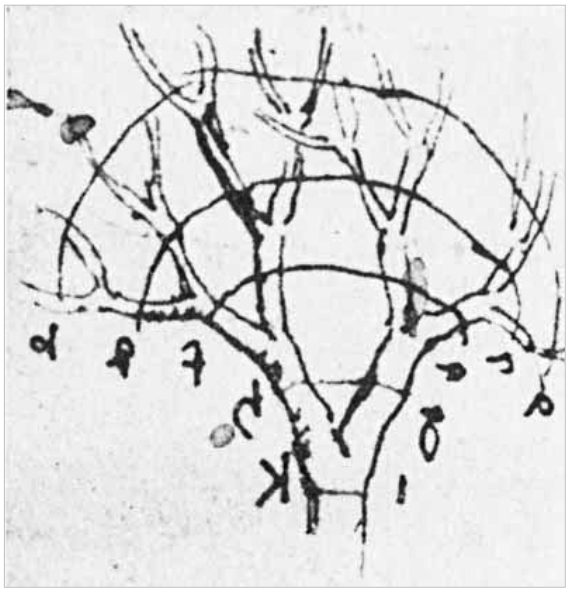
Die Durchführbarkeit und Umsetzung der vorliegenden Arbeit ist in unterschiedlichen Aspekten durch die Unterstützung und Mithilfe verschiedener Personen ermöglicht worden. Eine grundlegende Voraussetzung für die erfolgreiche Umsetzung einer wissenschaftlichen Arbeit ist eine langfristig gesicherte Finanzierung, die den nötigen Freiraum bietet eine Idee, sowie die damit verbundenen Schwierigkeiten und Konsequenzen zu durchdenken. Die Finanzierung meiner Tätigkeit am Institut für Waldinventur und Waldwachstum der Universität Göttingen wurde über einen Zeitraum von insgesamt drei Jahren durch die DFG gesichert.

Mein besonderer Dank gilt Herrn Prof. Dr. C. Kleinn, der maßgeblich an der Formulierung der Idee, die dieser Arbeit zugrunde liegt, sowie der Antragstellung bei der DFG beteiligt war und mich in vielerlei Hinsicht ermutigt und unterstützt hat.

Die wissenschaftliche Bearbeitung der gegebenen Fragestellungen erforderte zunächst die Sammlung geeigneter und ausreichender Datengrundlagen. Von den sich heirbei ergebenden Rückschlägen und Erfolgserlebnissen seien nur die letzteren genannt. Mein besonderer Dank gilt hierbei Dr. C. Wirth, Prof. Dr. E. Tomppo, Dr. A. Lehtonen, Dr. M. Ter-Mikaelian, Dr. V. Zakrzewski, Dr. A. Mench, Dr. R. Joosten sowie der Forschungsanstalt für Waldökologie und Forstwirtschaft Rheinland Pfalz (Dr. J. Block) für die Bereitstellung teilweise sehr umfangreicher Datengrundlagen.

Für wissenschaftlichen Gedankenaustausch und hilfreiche Anmerkungen möchte ich mich weiterhin bei Kai Staupendahl, Dr. M. Katila, Dr. A. Lehtonen, Prof. Dr. E. Tomppo, Prof. Dr. B. Sloboda und Prof. Dr. K.v. Gadow herzlich bedanken. Besonderer Dank gilt auch Lars Hinrichs, der als Kollege und Freund jederzeit für inhaltliche und mindestens ebenso wichtige anderweitige Diskussion zur Verfügung stand. Nicht zuletzt sei allen Mitarbeitern des Instituts für Waldinventur und Waldwachstum der Universität Göttingen, speziell Ulrike Dockter, Reinhard Schlote, Hendrik Heydecke sowie Sonja Rüdiger für kollegiales Miteinander und eine angenehme Arbeitsatmosphäre gedankt.

Für ihre Geduld mit mir und ihre Unterstützung bei der Korrektur dieser Arbeit danke ich Anja aus vollstem Herzen.



"… all the branches of a tree at every stage of its height when put together are equal in thickness to the trunk …"

(Aufzeichnung von Leonardo da Vinci aus dem Jahr 1510, zitiert nach RICHTER, 1970)

1 Einleitung

Die Vorhersage und Modellierung des Waldwachstums ist seit jeher durch wechselnde Ansprüche und Fragestellungen geprägt. Während die frühe Forschung im Bereich der Waldwachstumskunde zumeist das Ziel hatte, Planungsgrundlagen für die ökonomische Nachhaltigkeit der Holzerträge bereitzustellen, sind die Zielgrößen der Modellierung heute vielfältig. Abgesehen von der unterschiedlichen Motivation der Modellierung liegt jedoch allen quantitativen Untersuchungen die Idee zugrunde, aus bestehendem Wissen über einen Ausgangszustand und gegebenen Zusammenhängen eine Vorhersage für eine zukünftige Veränderung einer bestimmten Zielgröße abzuleiten. Grundsätzlich handelt es sich hierbei also um einen Lernprozess, in dem meist deduktiv auf Grundlage eines unterstellten Prinzips eine Vorhersage abgeleitet wird. Das zugrunde gelegte Prinzip bzw. der als Referenz verwendete Ursache-Wirkung- Zusammenhang wird hierbei durch eine bestimmte Modellformulierung vorgegeben, die entweder auf Grundlage von empirischen Daten abgeleitet wird (*empirische Modelle*), oder auf einem gesicherten Vorwissen über den Wirkungszusammenhang beruht (*Prozessmodelle*). Beide Ansätze haben dabei bestimmte Vor- und Nachteile.

Im Bereich der Modellierung der Biomasse von Wäldern, die in den letzten Jahrzehnten besonders durch die zunehmende globale Klimaveränderung angetrieben wird, hat die empirische Forschung zu einer kaum übersehbaren Fülle von Regressionsmodellen geführt, die in unterschiedlichster Formulierung und mathematischer Form vorliegen (VALENTINE und MÄKELÄ, 2005). Sie beschreiben die oberirdische trockene Biomasse auf Einzelbaumebene in Abhängigkeit einfach zu erfassender Variablen wie dem Brusthöhendurchmesser (BHD) oder der Baumhöhe. Die trockene Biomasse ermöglicht dann im weiteren Aussagen über den Kohlenstoffgehalt.

Im Gegensatz dazu bauen Prozessmodelle auf biologisch, physikalisch oder hydraulisch plausiblen Zusammenhängen niedrigerer Hierarchiestufen auf und fassen diese zu einem Erklärungsansatz zusammen. Das Ziel das Zusammenspiel der einzelnen Einflussgrößen aufzuklären kann dabei zu einer hohen Komplexität solcher Modelle führen, was ihre Handhabbarkeit für praktische Anwendungen einschränkt. Da nur der explizit erklärbare Einfluss der verwendeten Messgrößen berücksichtigt wird, haben sich Prozessmodelle teilweise als zu unflexibel erwiesen, wenn es darum geht unterschiedliche standörtliche Bedingungen oder Umweltfaktoren zu berücksichtigen (KRAMER et

al., 2002; LEHTONEN, 2005a). Andererseits besteht in der Vereinheitlichung der in empirischen Studien verwendeten Modellformulierung auf Grundlage der durch Prozessmodelle vorgegebenen biologisch plausiblen Wirkungszusammenhänge, eine Möglichkeit zur Generalisierung von Modellierungsansätzen. Bei der Formulierung solcher sog. *Hybridmodelle* (MÄKELÄ et al. 2000, MONSERUD, 2003) wie auch beim Vergleich von Prozessmodellen und empirischen Modellen sind jedoch die grundsätzlichen Unterschiede der Sichtweise beider Ansätze zu berücksichtigen. Einige generelle Aspekte, die sich aus der Verwendung empirischer Einzelbaumdaten ergeben, werden bisher jedoch kaum diskutiert.

Eine andere Form des Lernens besteht in induktiven Ansätzen, die eventuelle Wirkungszusammenhänge ignorieren und Entscheidungen oder Prognosen fallbasiert auf Grundlage von vorhandenen Beispielen ableiten. Das Lernen aus Beobachtungen ist besonders im Bereich der künstlichen Intelligenz und Mustererkennungsverfahren weit verbreitet. Die in diesem Forschungsfeld erarbeiteten Grundlagen werden mittlerweile in vielen Bereichen angewendet. Beispielsweise stützt sich eine tägliche Vorhersage einer Regenwahrscheinlichkeit des Wetterdienstes darauf, wie oft es in einer vergleichbaren bekannten Wettersituation in der Vergangenheit geregnet hat. Auch die Recherche im Internet mit Hilfe einer Suchmaschine oder die Vorhersage von Aktienkursen beruhen oftmals auf induktiven Methoden, nicht zuletzt, da die Modellierung und Vorhersage mit deduktiven Ansätzen zu komplex wäre.

Obwohl die Vorhersage einer unbekannten Zielgröße für einzelne Waldbäume theoretisch eine prädestinierte Aufgabe für Mustererkennungsverfahren zu sein scheint, ist die Anwendung entsprechender Verfahren im Vergleich zu herkömmlichen Modellierungsansätzen bisher wenig verbreitet. In Bezug auf die Schätzung der Biomasse einzelner Bäume ist bisher keine konkrete Verwendung instanzenbasierter Verfahren bekannt. Dies mag damit zusammenhängen, dass induktive Lernverfahren zwar keine Kenntnis über die zugrunde liegenden Wirkungszusammenhänge voraussetzen, für eine effiziente Anwendung jedoch eine relativ große Anzahl von bekannten Fallbeispielen erforderlich ist. Da destruktive Biomasseuntersuchungen mit einem enormen Zeit- und Kostenaufwand verbunden sind, ist die Anzahl untersuchter Bäume in einzelnen Studien typischerweise relativ gering. Die Zusammenstellung einer ausreichend großen Basis von Fallbeispielen aus verschiedenen Untersuchungen ist daher als Grundvoraussetzung zur

Implementierung der hier verwendeten k-Nearest Neighbour (k-NN) Methode anzusehen.

1.1 Hintergrund

Die internationale Gemeinschaft hat auf die Herausforderung des globalen Klimawandels mit einer Reihe von Vereinbarungen reagiert. Die UN Klimarahmenkonvention (UNFCCC) wurde auf der UN Konferenz 1992 in Rio unterzeichnet und ist 1994 in Kraft getreten. Sie setzt Ziele und definiert grundlegende Mechanismen als Vorgaben für eine zukünftige globale Klimapolitik und fordert die Unterzeichnerstaaten zur Erarbeitung von Grundlagen zur Umsetzung der angestrebten Ziele im Rahmen von jährlichen Konferenzen auf (Conferences of the Parties; COPs) (ROSENBAUM et al., 2004; WGBU, 1998).

Das Produkt der dritten COP (COP-3) ist das Kyoto Protokoll (1997), das nach der Unterzeichnung Russlands am 16.02.2005 in Kraft getreten ist. Im Rahmen der Umsetzung der im Kyoto Protokoll definierten Mechanismen zur Reduzierung von Treibhausgasen, ist eine Quantifizierung von Kohlenstoffquellen und Senkeneffekten, die durch forstwirtschaftliches Handeln beeinflusst werden, notwendig. Hierzu ist eine verlässliche Abschätzung der Biomassevorräte in Waldökosystemen und deren Veränderungen notwendig, wobei die rechtlich bindenden Vorgaben explizit eine statistisch abgesicherte Fehlereinschätzung vorschreiben (BROWN, 1997; 2001; SCHÖNE und SCHULTE, 1999; JOOSTEN et al., 2004; ROSENBAUM et al., 2004).

In Bezug auf die vorhandenen methodischen Grundlagen ergeben sich hierbei einige Probleme, die die Wirkungsweise dieses Politikprozesses hemmen. Zwar gibt es eine Vielzahl empirischer Biomassestudien in denen Biomassefunktionen abgeleitet wurden, die Zielsetzung solcher Untersuchungen war jedoch meist nicht die Quantifizierung von Kohlenstoffbilanzen auf regionaler oder gar nationaler Ebene. Beispielsweise umfasst der Puplikationszeitraum der in dieser Arbeit verwendeten Daten nahezu 90 Jahre. Die meisten Studien hatten dabei zumeist das Ziel, die außer dem Derbholz vorhandenen Biomassevorräte und Zuwächse von Wäldern abzuschätzen (Z.B. BURGER, 1925-1953; GRUNDNER und SCHWAPPACH, 1952; FIEDLER, 1986, MARKLUND, 1988; HAKKILA, 1989; AKÇA und MENCH, 1993; WIRTH et al., 2004).

Die ausschließliche Betrachtung des Waldes als Stammholzproduzenten änderte sich mit dem Bewusstsein, dass die gesamte organische Produktion von Wäldern zu berücksichtigen ist (MALENDE, 1997). Dabei haben die vermehrte Nachfrage nach Waldprodukten, die Suche nach erneuerbaren Rohstoffquellen sowie ein gesteigertes Interesse für ein intaktes Waldökosystem die Untersuchungen von Waldbiomasse vorangetrieben (WHITTAKER et al., 1974; MADWICK, 1976). Weitere Fragestellungen waren z.B. die Quantifizierung von Stoffflüssen in ökologischen Studien (z.B. HELLER und GÖTTSCHE, 1986; Rademacher, 2002) oder die Herleitung von Allokationsregeln im Wachstum von Einzelbäumen (z.B. BASKERVILLE, 1965; BARTELINK, 1998). Mit dem aufkommen neuartiger Waldschäden wurden weiterhin verschiedene Studien zur Standfestigkeit von Waldbäumen in Abhängigkeit der Biomasseverteilung durchgeführt (z.B. NIELSEN, 1990).

Erst neuere Untersuchungen beschäftigen sich ausdrücklich mit der Biomasseschätzung zum Zweck der Kohlenstoffbudgetierung. Die Anzahl dieser meist auf bestimmte Baumarten bezogenen Untersuchungen ist weltweit gesehen kaum zu überblicken. Dementsprechend finden sich zumindest für gut untersuchte und wirtschaftlich bedeutende Baumarten viele verschiedene, meist allometrische Biomassefunktionen. Die unübersichtliche Vielfalt der vorhandenen Modelle führte dazu, dass zahlreiche Überblicksstudien zur systematischen Sammlung von Biomassefunktionen durchgeführt wurden (z.B. ART und MARKS, 1971; YOUNG, 1976; CANELL, 1982; TRITTON und HORNBECK, 1982; MONSERUD et al., 1995; TER MIKAELIAN und KORZUKHIN, 1997; EASMUS, 2000; GIFFORD, 2000; KEITH et al., 2000; GRIESON et al., 2000; SNOWDON, 2000; JENKINS et al., 2004; ZIANIS et al., 2005).

Problematisch hierbei ist, dass die abgeleiteten Funktionen aufgrund der oftmals kleinen Datensätze, die aus destruktiven Untersuchungen aus eng begrenzten Untersuchungsgebieten stammen, in ihrem Geltungsbereich eingeschränkt sind und sich daher nur begrenzt zur Kohlenstoffbudgetierung für größere Gebiete verwenden lassen. Die Zielsetzung vieler Untersuchungen ist es daher allgemeingültige Modelle abzuleiten, die einen weiteren Geltungsbereich haben (beispielhaft seien genannt: MONSERUD et al., 1995; EASMUS et al., 2000; KETTERINGS et al., 2001; ENQUIST, 2002; VAN NOORDWIJK und MULIA, 2002; WIRTH und SCHUMACHER, 2002; JENKINS, 2003; LEHTONEN et al., 2003; WIRTH et al., 2003; ZIANIS und MENCUCCINI, 2003; CHAVE et al. 2005; ZIANIS et al., 2005). Dabei gibt es verschiedene Möglichkeiten zur Herleitung solcher Ansätze.

Eine Möglichkeit besteht darin, anhand von Meta-Analysen von vorhandenen spezifischen Funktionen eine Gruppierung ähnlicher Baumarten herzuleiten. Bei solchen Analysen stellt sich jedoch das Problem der Streuung zwischen verschiedenen Untersuchungen, die im Folgenden einen direkten Vergleich der Ergebnisse erschwert. JENKINS et al. (2003) konnten bei einer Meta-Analyse von nord-amerikanischen Biomassefunktionen große Streuungen zwischen den Untersuchungen feststellen, die teilweise innerhalb einer Baumart größer waren als zwischen den Baumarten. Sie schlagen daher vor, die Ursprungsdaten zu sammeln und neu auszuwerten.

Für die Abschätzung von Kohlenstoffbilanzen in Waldökosystemen auf regionaler oder nationaler Ebene wird aufgrund der Unsicherheit bezüglich bestehender Biomassemodelle normalerweise auf die Verwendung von Biomasse Expansionsfaktoren (BEFs) zurückgegriffen, obwohl die nötigen Datengrundlagen für eine Modellierung der Einzelbaum Biomasse im Rahmen von Waldinventuren erhoben werden (CANELL, 1995; BARITZ und STRICH, 2000; WIRTH et al., 2004; VAN CAMP et al., 2004; FEHRMANN und KLEINN, 2005). Biomasse Expansionsfaktoren beschreiben das Verhältnis zwischen Derbholz- oder Stammvolumen und der Gesamtbiomasse eines Baumes. Sie stellen daher das Ergebnis zweier Schätzungen, nämlich einer Volumenschätzung und einer Schätzung der Biomasse dar, und können insofern als ein Derivat von Biomassefunktionen angesehen werden.

Da sich das Verhältnis verschiedener Biomasse Kompartimente mit zunehmendem Alter verschiebt, sind BEFs auch innerhalb einer bestimmten Baumart nicht konstant, sondern stark vom Alter bzw. der Dimension der Bäume abhängig (siehe z.B. WHITTAKER et al., 1974; BARTELINK, 1998; LEHTONEN et al., 2004; WIRTH et al., 2004). Die Unsicherheit dieses Schätzverfahrens wird als eines der größten Hemmnisse für die Anrechnung von Kohlenstoff-Senkeneffekten im Rahmen des Kyoto Protokolls angesehen (SCHÖNE und SCHULTE, 1999). Da der Kyoto Prozess als Reaktion auf die weltweite Klimaveränderung eine globale Anstrengung ist, gilt es zur Implementierung der verschiedenen Mechanismen nicht nur die Kohlenstoffbindung in den Teilen der Erde zu quantifizieren, in denen gesicherte Forschungsergebnisse und wissenschaftliche Grundlagen vorhanden sind. Während man es in unseren Breiten mit einer geringen Anzahl relativ gut untersuchter Baumarten zu tun hat, ergeben sich hierbei in tropischen Regionen viel größere Probleme. Die Ableitung baumartenspezifischer Biomassefunktionen scheint hier kaum umsetzbar zu sein.

1.2 Zielsetzung

Ziel der vorliegenden Untersuchung ist es, alternative Methoden der Biomasseschätzung auf Einzelbaumebene zu untersuchen. Hierbei stehen zwei Gesichtspunkte im Vordergrund: Zum einen soll die bestehende Diskrepanz zwischen den Vorhersagen vorhandener Prozessmodelle und empirisch abgeleiteten Biomassefunktionen untersucht werden. Hierbei soll die grundlegend unterschiedliche Sichtweise der verschiedenen Ansätze herausgestellt und Möglichkeiten der Integration aufgezeigt werden. Zum anderen soll die Eignung nicht-parametrischer bzw. instanzenbasierter Verfahren zur Biomasseschätzung am Beispiel der k-NN Methode evaluiert werden.

Die gegebenen Ziele lassen sich in Form grundlegender Hypothesen bzw. Fragestellungen wie folgt formulieren:

i. Die aus Prozessmodellen abgeleiteten Gesetzmäßigkeiten sind mit empirischen Datengrundlagen nur eingeschränkt nachzuweisen, da generelle Unterschiede in der Sichtweise beider Ansätze bisher weitgehend unberücksichtigt bleiben.

ii. Um eine Generalisierung allometrischer Biomassefunktionen bzw. die Integration von Prozessmodellen und empirischer Datenanalyse zu ermöglichen, müssen die theoretischen Grundlagen allometrischer Beziehungen stärker berücksichtigt werden.

iii. Neben parametrischen Modellierungsansätzen eignet sich auch die k-NN Methode als instanzenbasiertes Verfahren zur Vorhersage der Einzelbaumbiomasse.

Zur Überprüfung dieser grundlegenden Fragestellungen und im Besonderen zur effizienten Umsetzung der k-NN Methode ist der Aufbau einer Biomassedatenbank für Einzelbäume zwingend erforderlich. Eine weitere Zielsetzung besteht daher in der Entwicklung einer geeigneten Datenbankstruktur, die in der Lage ist, die Voraussetzung eines instanzenbasierten Lernverfahrens zu erfüllen. Weiterhin ist die Sammlung von geeigneten Datengrundlagen aus einzelnen Biomasseuntersuchungen sowie deren Vereinheitlichung zu einer umfassenden Datengrundlage ein Hauptaspekt dieser Arbeit.

Da die Umsetzung der k-NN Methode, im Gegensatz zur Anwendung eines Regressionsmodells, die Implementierung des nötigen Algorithmus in einer Softwareumgebung erfordert, ist die Entwicklung einer geeigneten Anwendung unerlässlich. Hierzu müssen verschiedene theoretische Grundlagen des Verfahrens erarbeitet werden und in Form konkreter Handlungsanweisungen bzw. einer funktionsfähigen Softwarearchitektur ausgedrückt werden.

Die vorliegende Arbeit ist durch die Evaluierung der k-NN Methode als neuer Beitrag zur Biomasseforschung anzusehen. Weiterhin ergibt sich aus der Zielsetzung die Aspekte der Verwendung allometrischer Funktionen auf Einzelbaumebene zu hinterfragen, ein Gesichtspunkt der im Bereich der Biomassemodellierung bisher weitgehend undiskutiert bleibt.

2 Methodische Ansätze der Biomassemodellierung

2.1 Stand der Forschung

Die Biomasse von Bäumen wird schon seit langem wissenschaftlich untersucht (z.B. KUNZE, 1873). Frühe systematische Stichprobenuntersuchungen zur Schätzung der oberirdischen Biomasse fanden bereits 1929 an *Pinus strobus* in der Schweiz statt und wurden später durch Untersuchungen weiterer Baumarten ergänzt (BURGER 1925-1953). Biomasseuntersuchungen waren lange Zeit hauptsächlich für ökologische Fragestellungen, beispielsweise zur Modellierung von Nährstoffkreisläufen in Waldökosystemen, von Bedeutung. Im Gegensatz zur Schätzung des Holzvolumens, war die Schätzung der holzigen Biomasse von Bäumen in der Vergangenheit kein wirtschaftlich relevantes Thema, da ein Großteil des Nutzholzes traditionsgemäß in Volumen- und selten in Masseneinheiten gehandelt wird. Seit sich durch technischen Fortschritt und veränderte Holzverwendung neue Nutzungsmöglichkeiten ergeben, steigt jedoch das Interesse die Produktivität der Ressourcen auch in Masseneinheiten zu bewerten (KRAMER und AKÇA, 1995). Neben der notwendigen Bewertung von Kohlenstoff-Senkeneffekten, ergeben sich solche Fragestellungen z.B. im Bereich des Energieholzsektors oder bei der Nutzung von Hackschnitzeln (GADOW et al., 2006).

Die Verwendung von Wachstumsfunktionen im Rahmen der Biomasseschätzung weist dabei grundlegend verschiedene Motivationen auf. Einerseits ist die möglichst genaue Schätzung und statistische Beschreibung der Zielgröße erwünscht, andererseits versuchen Wachstumsfunktionen eine biologische Erklärung auf Grundlage physiologischer Auf- und Abbauprozesse zu liefern (PRETZSCH, 2001). Dementsprechend finden sich in der Literatur grundlegend verschiedene Erklärungsansätze in Form von *empirischen Modellen* und *Prozessmodellen*.

2.1.1 Empirische Ansätze der Biomassemodellierung

Im Gegensatz zu Prozessmodellen, bei denen auf der Grundlage biologischer oder physikalischer Gesetzmäßigkeiten eine erklärbare Modellformulierung abgeleitet wird, wird in empirischen Studien eine Modellformulierung als Grundlage der Regressionsanalyse gewählt. In diesem Fall ist das Ziel der Untersuchung also nicht die Erklär-

barkeit der biologischen Grundlagen oder Zusammenhänge, sondern vielmehr die Erklärung der beobachteten Varianz in der Datengrundlage. In der Literatur finden sich verschiedenste Modellformulierungen, wobei allometrische Funktionen deutlich überwiegen (PARRESOL, 1999). Die theoretischen Hintergründe und Eigenschaften allometrischer Funktionen werden wegen ihrer Bedeutung für die Biomassemodellierung nachfolgend ausführlicher behandelt.

2.1.2 Allometrische Biomassefunktionen

Allometrie ist als das Studium der mit Größenveränderungen einhergehenden Proportionsänderungen der Teile eines Organismus, entweder angewendet auf das Wachstum von Individuen (*ontogenetisch*) oder zum Vergleich unterschiedlich großer verwandter Organismen (*phylogenetisch*), definiert. Für vergleichende Darstellungen von Wachstumsvorgängen einzelner Teile eines Organismus bedient man sich häufig der ontogenetischen Allometrie (oder Wachstumsallometrie). Obwohl die dem Baumwachstum zugrunde liegenden physiologischen Auf- und Abbauprozesse sehr komplex sind, resultieren sie in oftmals sehr stabilen Wachstums und Zuwachskurven (PRETZSCH, 2001). Bis heute ist daher die klassische Allometrieformel, die auf SNELL (1892) zurückgeht, das meist verwendete Modell in empirischen Biomasseuntersuchungen. Weitere frühe Anwendungen allometrischer Beziehungen finden sich bei SPENCER (1864) und THOMPSON (1917) (PRETZSCH und BIBER, 2005).

Hierbei wird die relative Größenveränderung eines Teiles eines Organismus zur relativen Größenveränderung eines anderen Teils oder dem Gesamtorganismus in der Form $\delta y/y = b^{*}\delta x/x$ ausgedrückt (HUXLEY, 1924; 1932; BERTALANFFY, 1951; LABARBERA, 1989). Integriert man diesen Ausdruck, so erhält man die bekannte Grundgleichung der Allometrie:

$$y = a \cdot x^{b} \tag{1}$$

Der Exponent b ist hierbei ein Maß für das Verhältnis zweier relativer Wachstumsgeschwindigkeiten (absolute Wachstumsgeschwindigkeit dividiert durch die Wachstumsgröße zum selben Zeitpunkt) und wird auch als Allometriekonstante (oder Proportionalitätskonstante) bezeichnet. Die zugrunde liegende Differentialgleichung lautet:

$$\frac{\delta y}{y} = b * \frac{\delta x}{x} \qquad \text{bzw.} \qquad b = \frac{\frac{\delta x}{x}}{\frac{\delta y}{y}} \tag{2}$$

Da b in der zugrunde gelegten Allometrieformel über den gesamten Geltungsbereich konstant ist, wird somit unterstellt, dass eine relative Durchmesserveränderung zu jedem Zeitpunkt eine gleiche relative Änderung der Biomasse bedingt.

Ist b=1 spricht man von isometrischem, bei b<1 von negativ- und bei b>1 von positiv-isometrischen Wachstum. Allerdings gilt diese Grenze nur bei Maßen gleicher Dimension (PRETZSCH, 2001). Durch die Multiplikation mit dem konstanten Faktor a (Integrationskonstante) werden alle Ordinatenwerte im Verhältnis 1/a gestaucht (für $a < 1$). Da gilt:

$$y = a \cdot x^b = e^{b \cdot \ln x + \ln a} \Rightarrow \ln y = b \cdot \ln x + \ln a \tag{3}$$

wird im doppelt logarithmierten Koordinatensystem aus der Potenzkurve eine Gerade mit der Steigung b (Tangens des Winkels). Umgekehrt ausgedrückt, liegt ein allometrisches Wachstumsgesetz nur dann vor, wenn die Wertepaare (ln (x)/ln (y)) einen linearen Zusammenhang aufweisen. Allometrie ist in diesem Sinne als Studium des Wachstums *eines Teils* eines Organismus im Verhältnis zum Wachstum des *Gesamtorganismus* zu verstehen.

Dieser Zusammenhang kann auch verdeutlicht werden, indem man die Grundgleichung in Formel (1) transformiert:

$$\frac{\delta y}{\delta x} = b \cdot \frac{y}{x} \tag{4}$$

Hierbei wird die allometrische Beziehung als Ergebnis eines Verteilungsprozesses zwischen verschiedenen Organen eines Organismus interpretiert (PRETZSCH und BIBER, 2005).

Obwohl die lineare Beziehung in Formel (3) in mathematischer Hinsicht mit der Form (1) identisch ist, ergibt sich im Rahmen der Regressionsanalyse ein wesentlicher Unterschied. Durch die Logarithmierung wird eine Homogenisierung der Fehlervarianzen erreicht und eine unverzerrte Einschätzung der Güte erlaubt, da Heteroskedastizität vermieden wird (FURNIVAL, 1961; BASKERVILLE, 1972; PARRESOL, 1999; JOOSTEN et al., 2004; ZIANIS UND MENCICCINI, 2004). Aus diesem Grund werden üblicherweise die logarithmisch transformierten Wertepaare der unabhängigen und abhängigen Variablen als Eingangsgrößen der Regressionsanalyse verwendet. Dies hat gleichzeitig die Vorteile, dass eine einfache lineare Regression mit der Methode der kleinsten Quadrate angewendet werden kann (SPRUGEL, 1983) und der Einfluss von Ausreißern abgeschwächt wird.

Allerdings muss beachtet werden, dass sich bei der Rücktransformation der logarithmischen Schätzwerte auf das metrische Skalenniveau eine systematische Abweichung (Unterschätzung) ergibt, da der Mittelwert der transformierten y-Werte, der in der Regressionsanalyse zur Anpassung einer Funktion genutzt wird, dem Median und nicht dem Mittelwert der Rücktransformierten metrischen Verteilung entspricht (NIKLAS, 2004). Zur Bias-Korrektur bei der Rücktransformation von Schätzwerten aus logarithmischen Regressionen sind in der Literatur verschiedene Ansätze zu finden (siehe z.B. FINNEY, 1941; BASKERVILLE, 1972; BEAUCHAMP und OLSON, 1973; SPRUGEL, 1983). Eine Approximation des Korrekturfaktors kann dabei z.B. aus dem Standardfehler der Schätzung der Regression (*SEE*) berechnet werden:

$$CF = exp\left(S_{y,x}^2 / 2\right) \tag{5}$$

Für CF = Korrekturfaktor und $S_{y,x}^2$ = Standardfehler der Schätzung (SEE).

Bemerkenswert in Bezug auf die verwendete Methode der Regression ist, dass es bei der Untersuchung von allometrischen Beziehungen innerhalb eines Organismus (ontogen-

etische Allometrie) theoretisch keine unabhängigen Variablen gibt, da alle einbezogenen Größen durch bestimmte Verhältnisregeln in Beziehung zueinander stehen. Dies ist beispielsweise der Fall, wenn der BHD oder der BHD und die Baumhöhe als unabhängige Variable für die Schätzung der Biomasse verwendet werden und diese Größen durch allometrische Verhältnisregeln voneinander abhängig sind. Hierbei ist zu bedenken, dass nicht nur die Residuen in Richtung der y-Achse, sondern auch in jede andere Richtung zur Herleitung der Regressionskoeffizienten in Betracht gezogen werden müssen. Im Falle hoher Korrelationen führt diese Aufteilung des Fehlers jedoch nur zu geringen Abweichungen der geschätzten Koeffizienten (NIKLAS, 2004).

2.1.3 Theoretische Ansätze

Neben den oben beschriebenen empirischen Modellansätzen gibt es eine Reihe von Prozessmodellen, die die Beziehung zwischen Biomasse und messbaren Variablen auf der Grundlage biologischer, hydraulischer oder mechanischer Gesetzmäßigkeiten beschreiben. Einer der ersten grundlegenden und allgemeingültigen Ansätze zur Beschreibung der Baumarchitektur und der daraus folgenden Verteilungsmuster innerhalb von Baumindividuen ist die Pipe-Modell Theorie (SHINOZAKI, 1964; OOHATA und SHINOZAKI, 1979). Sie gibt einen Erklärungsansatz für die Verteilung der Biomasse innerhalb eines Individuums auf Grundlage von Annahmen über die hydraulische Architektur in Gefäßpflanzen.

Bereits im Jahr 1510 bemerkte Leonardo da Vinci in seinen Überlegungen zur fraktalen Geometrie eines Baumes: "… *all the branches of a tree at every stage of its height when put together are equal in thickness to the trunk* …" (zitiert nach RICHTER, 1970). Obwohl dieser Zusammenhang in seinen Aufzeichnungen aus dem Blickwinkel des Malers beschrieben wird, findet er sich später als Grundidee der Pipe Modell Theorie wieder. Auf die hydraulische Baumarchitektur bezogen, beschreibt das Modell den Zusammenhang zwischen transpirierender Blattfläche und dem leitendem Gefäßgewebe, das zum Transport von Wasser und Nährstoffen von den Wurzeln bis zu den Blättern dient. Aufbauend auf dieser Grundannahme wurden später verschiedene Prozessmodelle auf Einzelbaum- und Bestandesebene entwickelt. Bestandesbezogene Modelle haben dabei die Zielsetzung eine Kohlenstoffbilanz unter Berücksichtigung von Zuwächsen

und Verlusten bzw. Aufbau- und Abbauprozessen zu beschreiben. Exemplarisch für diesen Ansatz sei hier die *pipestem theory* (VALENTINE, 1988; 1999; MACFARLANE et al., 2000) genannt.

Ein sehr umfassender Ansatz zur Erklärung allgemeingültiger allometrischer Verhältnisregeln im Wachstum von einzelnen Pflanzen wurde später von WEST et al. (1997; 1999), ENQUIST et al. (1999), BROWN und WEST (2000), ENQUIST (2002) (NIKLAS, 1994; 2004) gegeben. Dieser hier nach ENQUIST (2002) im Folgenden als *WBE Modell* bezeichnete Ansatz beruht auf der Grundlage fraktaler Geometrie in Bäumen. Auf diesen Modellansatz soll im Weiteren näher eingegangen werden, da er im Rahmen von Versuchen zur Verallgemeinerung von Biomassefunktionen oft als Vergleich herangezogen wird (siehe z.B. CHAVE et al., 2001; JENKINS, 2003; ZIANIS und MENCUCCINI, 2004).

Unter der Grundannahme, dass die Gesamtheit der Querschnittsflächen der leitenden Gefäßbahnen auf jeder Verzweigungsebene konstant ist (die Pipe Modell Annnahme), berücksichtigt das WBE Modell die Tatsache, dass nicht die gesamte Querschnittsfläche aus leitendem Gewebe besteht. Vielmehr wird hier ein Verhältnis zwischen leitender und nicht-leitender Querschnittsfläche in Abhängigkeit der jeweiligen Verzweigungsebene unterstellt. Unter der Grundannahme, dass Bäume eine sich selbst wiederholende fraktale Struktur besitzen, in der sich bestimmte Verhältnisregeln auf verschiedenen Ebenen dieser Struktur wiederholen, wird angenommen, dass das Verhältnis zwischen Leitgewebe und nicht leitendem Gewebe in einem allometrischen Verhältnis steht. Unter weiteren Annahmen über die Beziehungen zwischen Gefäßdurchmesser, Astdurchmesser und Astlänge innerhalb der fraktalen Struktur, gibt der Ansatz Erklärungen für die Verteilungsmuster und die resultierende Baumarchitektur.

Für die Beziehung zwischen Durchmesser (D) und Gesamtbiomasse (M) resultiert hieraus $M \propto D^{8/3}$, was einer Allometriekonstante von 2,667 entspricht. Dieser Exponent wird dabei aus $D \propto M^{3a/2(a+3)}$ abgeleitet, wobei $a = 1$ ist, wenn das Verhältnis zwischen Astdurchmesser und Astlänge, oder unter der Annahme der fraktalen Struktur auch Baumhöhe und Durchmesser, mit dem erwarteten allometrischen Exponenten von 2/3 beschrieben werden kann (ENQUIST et al., 1998; WEST et al., 1999a).

2.2 Gegenüberstellung von Prozessmodellen und empirischen Ansätzen

Zwar sind die Divergenzen zwischen den theoretisch hergeleiteten und den in empirischen Studien geschätzten Exponenten in vielen Fällen nicht statistisch signifikant (Enquist, 2002; Niklas und Enquist, 2002), jedoch ist die Abweichung der Schätzungen zu empirisch beobachteten Parametern so hoch, dass theoretisch begründete Koeffizienten in der praktischen Anwendung bisher keine Verwendung finden.

Während Prozessmodelle wie das WBE Modell feste Verhältnisregeln vorhersagen, führen empirische Untersuchungen in nahezu jedem Fall zu variierenden Exponenten des allometrischen Modells. Diese Variation wird dabei Unterschieden zwischen den Baumarten, dem Bestandesalter, der Standortsqualität, der Bestandesdichte oder den klimatischen Gegebenheiten zugeschrieben. Zudem wird die alleinige Verwendung des BHD als unabhängige Variable kritisiert, da angenommen wird, dass weitere Faktoren, wie z.B. die mechanische Belastung der Stammquerschnittsfläche durch Winddruck und Eigengewicht (Wilson und Archer, 1979; Mitchell und Myers, 1995; Gaffrey und Sloboda, 2001) oder die Konkurrenz als Folge der Bestandesstruktur (Pretzsch, 2001), einen entscheidenden Einfluss auf die Biomasse eines Einzelbaumes bzw. die Allokation der gegebenen Ressourcen haben (siehe z.B. Sprugel, 2002; Vanninen und Mäkelä, 2005).

In einer Meta-Analyse von Biomassemodellen aus insgesamt 277 Studien finden Zianis und Mencuccinci (2004) signifikante Unterschiede zwischen dem theoretischen Exponent 2,667 und den empirisch geschätzten Koeffizienten, wobei der Mittelwert für b in ihrer Studie bei 2,37 liegt. Weiterhin stellen sie eine starke negative Korrelation zwischen den Koeffizienten a und b des allometrischen Regressionsmodells fest. Dieser negative Zusammenhang wird erwartet, da eine gleichzeitige Zunahme beider Koeffizienten zu einer extremen Zunahme der Biomasse mit steigendem Durchmesser führen würde, und so die mechanischen Restriktionen der Baumstabilität verletzt (Niklas, 1994).

Jenkins et al. (2003) stellen in einer Meta Analyse nord amerikanischer Biomassestudien eine hohe Variabilität der empirischen Exponenten fest, die teilweise innerhalb einer Baumart höher als zwischen verschiedenen Baumarten ist. Hierbei bleibt jedoch

zu bedenken, dass solche Vergleichsanalysen oftmals dadurch eingeschränkt bleiben, das nicht auf die Ausgangsdaten zurückgegriffen werden kann, sondern lediglich die Regressionskoeffizienten veröffentlichter Studien einbezogen werden (Montagu et al., 2004). Eine Möglichkeit die Diskrepanz zwischen empirischen Studien und Prozessmodellen aufzulösen wird in der Verwendung von Hybridmodellen gesehen (Hinrichs, 2006), welche die Verknüpfung biologisch begründbarer Kausalzusammenhänge mit empirischen Datengrundlagen ermöglichen (z.B. Johnsen et al., 2001; Mäkelä et al., 2000; Valentine und Mäkelä, 2005; Monserud, 2003).

Die unterschiedlichen Ergebnisse der verschiedenen Erklärungsansätze auf Grundlage von Prozessmodellen und empirischen Modellen kann dabei grundlegend zwei Ursachen haben:

- Die verwendeten empirischen Daten sind nicht geeignet um theoretische Modellannahmen nachzuweisen, oder
- Bäume haben bestimmte Eigenschaften, die in den theoretischen Ansätzen nicht ausreichend berücksichtigt werden.

So kann beispielsweise die im WBE Modell unterstellte Höhenbeziehung $H \propto D^{2/3}$ in realen Waldbeständen nur in den seltensten Fällen nachgewiesen werden. Empirische Untersuchungen führen hierbei zu mehr oder weniger großen Abweichungen mit einer großen Variabilität (Zucchini et al., 2001; Temesgen und Gadow, 2004).

Weiterhin ist zu berücksichtigen, dass die Motivation der Datenaufnahme in empirischen Studien normalerweise darauf beschränkt ist Regressionsmodelle anzupassen, mit deren Hilfe die gesuchte Zielgröße geschätzt werden kann. Die Überprüfung theoretischer Annahmen, wie sie beispielsweise durch Prozessmodelle vorgegeben werden, ist selten eine Zielsetzung solcher Studien. So wird z.B. bei der Verwendung allometrischer Funktionen als Regressionsmodell in den meisten Fällen die biologische Implikation bzw. der von einer solchen Modellformulierung unterstellte Ursache-Wirkung Zusammenhang selten berücksichtigt (Niklas, 2004).

Ein grundlegender Unterschied zwischen Prozessmodellen und empirischen Untersuchungen ist darin zu sehen, dass theoretische Annahmen wie z.B. das WBE Modell einen Erklärungsansatz für Verhältnisregeln des ontogenetischen Wachstums (innerhalb

eines Individuums) liefern. Datengrundlagen aus empirischen Untersuchungen basieren jedoch auf destruktiven Stichproben. Da es nicht möglich ist, die Masse eines Baumes zu mehreren Zeitpunkten zu erfassen ohne sein Wachstum zu stören, stammen die Daten aus *unechten* Zeitreihen-Untersuchungen (sog. *Chronosequenzen*), in denen Bäume verschiedener Dimension zu einem Zeitpunkt untersucht werden (FITZHUGH, 1976; GADOW, 2003). Hierbei wird also eine Wiederholungsmessung in der Zeit durch simultane Punkt-Messungen an verschiedenen Orten ersetzt. Diese Methode wird seit langem zur Modellierung des Waldwachstums verwendet (ASSMANN, 1953; KRAMER, 1988; WENK et al., 1990; GADOW et al., 2006). Als Folge können die so gewonnenen Daten lediglich eine Aussage über die Haupteigenschaften einer mittleren Wachstumskurve, nicht aber über Verhältnisregeln innerhalb eines Einzelbaumes ermöglichen (GILLE, 1889). Da die Entwicklung des Einzelbaumes durch Zeitreihen-Untersuchungen nicht erfasst werden kann, werden solche Daten als ungeeignet für die Wuchsmodellierung auf Einzelbaumebene angesehen.

Da im Fall der Biomasseschätzung keine zerstörungsfreie Datenaufnahme möglich ist, kommt der Auswahl der Probebäume aus einem gegebenen Baumkollektiv eine erhöhte Bedeutung zu. Die Zielsetzung der meisten empirischen Untersuchungen ist es, einen möglichst weiten Durchmesserbereich abzudecken, um den Geltungsbereich der abgeleiteten Regressionsfunktion zu erweitern. Hierbei wird bei der Auswahl der Probebäume die Variabilität der Höhen-Durchmesser Verhältnisse nicht in jedem Fall berücksichtigt.

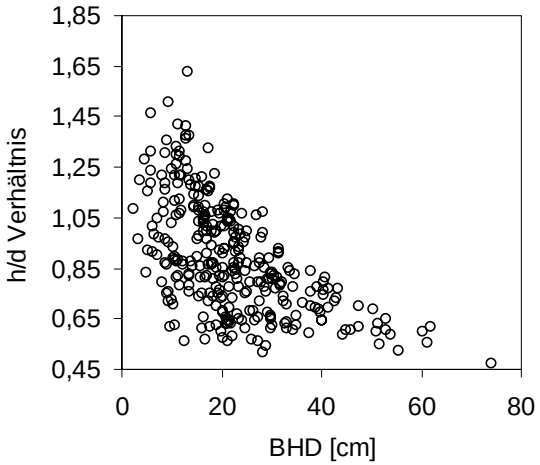


Abbildung 2-1. Höhen-Durchmesser (h/d-) Verhältnisse von Fichten ($n = 350$) über dem BHD in einem zusammengefassten Biomassedatensatz.

Wie aus Abbildung 2-1, in der die h/d Verhältnisse von 350 Fichten eines zusammengesetzten Datensatzes über dem BHD dargestellt sind deutlich wird, liegt vor allem im mittleren und unteren Durchmesserbereich eine hohe Streuung der Werte von verschiedenen Standorten vor. Eine ähnliche, wenn auch geringere Variabilität ist jedoch normalerweise auch in Einzeldatensätzen auf Bestandesebene zu beobachten.

Da Biomasseuntersuchungen sehr zeit- und kostenintensiv sind, sind die resultierenden Datensätze einzelner Untersuchungen oft sehr klein. Weiterhin stammen die Daten meist aus kleinräumigen Untersuchungsflächen. Dies führt dazu, dass innerhalb solcher Datensätze eine hohe Korrelation zwischen den für die Biomasse entscheidenden Variablen festgestellt werden kann. Besonders deutlich wird dieser Effekt z.B. bei der Anpassung von Regressionsfunktionen auf der Grundlage von Einzeluntesuchungen. In den meisten Fällen kann hierbei der Erklärungsanteil des Modells durch die Integration der Baumhöhe in die Modellformulierung nur geringfügig verbessert werden (siehe z.B. MADGWICK und SATOO, 1975; WIANT et al., 1979; JENKINS et al., 2003), obwohl theoretische Überlegungen nahe legen, dass die Höhe eines Baumindividuums einen entscheidenden Einfluss auf seine Gesamtmasse hat bzw. durch die Integration der Höhe ein Ansatz zur Verallgemeinerung des Modells gefunden werden kann. Die bestehende multiple Kolinearität zwischen den Variablen erschwert hierbei die getrennte Quantifizierung des Einflusses dieser Messgrößen.

Im Rahmen der Regressionsanalyse führt die oftmals hohe Korrelation zwischen BHD und Baumhöhe in bestandesweisen und kleinräumigen Untersuchungen dazu, dass die Trennschärfe bzw. Diskriminanz durch die Einbeziehung der Höhe nicht signifikant erhöht werden kann. Eine Untersuchung des Einflusses dieser Variablen ist daher nur möglich, wenn Daten von verschiedenen Standorten zusammengefasst werden. Hierbei ist jedoch zu berücksichtigen, dass durch die Aggregation von Datensätzen aus verschiedenen Untersuchungen in der weiteren Analyse möglicherweise eine Streuung aufgrund unterschiedlicher methodischer Grundlagen zwischen einzelnen Studien entstehen kann, die nicht getrennt quantifizierbar ist (WIRTH et al., 2003).

2.2.1 Der BHD als Eingangsgröße allometrischer Funktionen

Verwendet man das Allometriekonzept als theoretischen Erklärungsansatz für die Verteilungs- und Wachstumsverhältnisse der Baum- Biomasse, muss noch ein weiterer Aspekt der verfügbaren Datengrundlagen berücksichtigt werden, der in der vorhandenen Literatur bisher weitgehend undiskutiert bleibt. Wie unter 2.1.2 beschrieben, stellt ein allometrisches Wachstumsgesetz das Verhältnis zweier relativer Wachstumsraten dar. In empirischen Biomasseuntersuchungen wird dieses Modell verwendet, um das Verhältnis der relativen Änderung des BHD zur relativen Änderung der Biomasse zu beschreiben. Problematisch in Bezug auf die biologische Implikation allometrischer Beziehungen ist hierbei, dass der BHD in einer festgelegten Höhe von 1,3 m erfasst wird (PRETZSCH und BIBER, 2005). Da Bäume Objekte sind, die aufgrund des sekundären Dickenwachstums mit zunehmender Größe auch einer Formveränderung unterliegen, wird mit der Änderung des BHD also nicht nur die relative Änderung eines *funktionalen* Durchmessers, sondern gleichzeitig die relative Formveränderung in der festgelegten Messhöhe erfasst (siehe Abbildung 2-2).

Anders ausgedrückt, führt die Messung des BHD in einer absoluten Höhe dazu, dass bei einem Vergleich unterschiedlich großer Bäume Durchmesser aus unterschiedlicher relativer Stammhöhe Verwendung finden. Die Eigenschaft der Variablen BHD ist in dieser Hinsicht jedoch nicht konform mit den theoretischen Grundlagen allometrischer Beziehungen (siehe hierzu auch Anhang I, Seite 132).

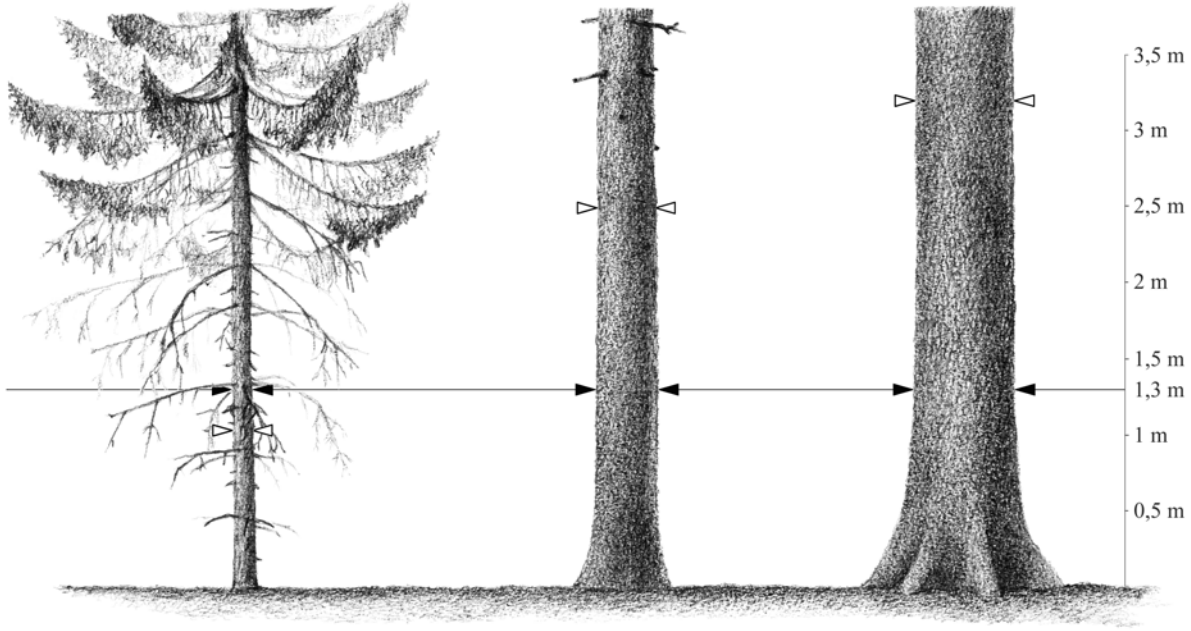


Abbildung 2-2. Lage des BHD und eines Durchmessers in relativer Stammhöhe. Hier ist die Lage des Durchmessers in 10 % der Stammhöhe ($D_{0,1}$) dargestellt (><) (Zeichnung W. Tambour).

Die Verwendung des BHD als unabhängige Variable ist jedoch auch in Bezug auf die dargestellten theoretischen Ansätze (siehe 2.1.3) durchaus kritisch zu betrachten. Unterstellt man eine konstante Querschnittsfläche der leitenden Gefäßbahnen (oder Tracheiden) als Folge eines flüssigkeitsgefüllten Gefäßnetzes von den Wurzeln bis zu den Blättern, führt die Verwendung des BHD als erklärende Variable für die Biomasse zu unterschiedlichen Verhältnissen zwischen leitendem und nicht leitendem Gewebe in unterschiedlichen relativen Stammhöhen (Abbildung 2-3). Die Formulierung des WBE Modells berücksichtigt zwar eine Änderung dieses Verhältnisses zwischen verschiedenen Verzweigungsstufen, schließt aber eine Veränderung des Verhältnisses *innerhalb* einer Verzweigungsstufe explizit aus.

Die Veränderung dieses Verhältnisses ist dabei eine Folge des sekundären Wachstums des Stammquerschnittes, die gleichzeitig auch von einer erhöhten mechanischen Belastung des Stammquerschnittes als Folge der Erhöhung der darüber liegenden Masse beeinflusst wird (BARTELINK, 1996; SLOBODA und GAFFREY, 1999; GAFFREY und SLOBODA, 2001).

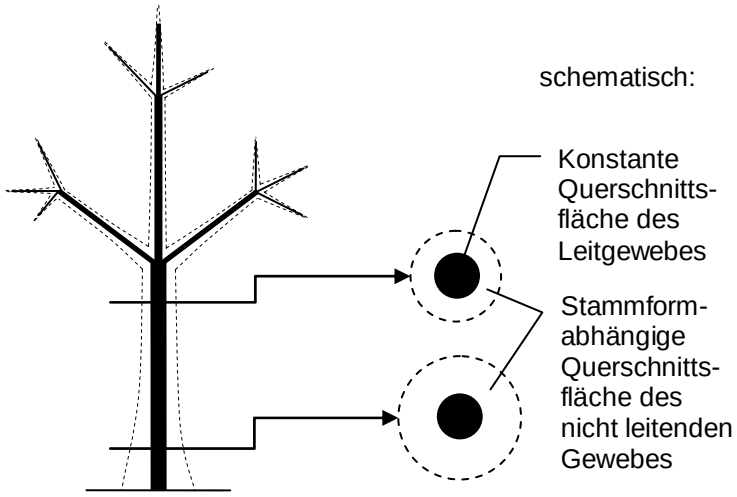


Abbildung 2-3. Schematische Darstellung der Veränderung des Verhältnisses zwischen leitendem und nichtleitendem Gewebe auf Stammquerschnittsflächen in unterschiedlicher Höhe.

Um den Effekt unterschiedlicher relativer Messhöhen an verschieden großen Bäumen zu kompensieren, besteht die Möglichkeit den gegebenen BHD mit Hilfe von Schaftformmodellen auf einen Durchmesser in einheitlicher relativer Schafthöhe umzurechnen. Mit Hilfe der so transformierten Eingangsgröße kann dann die Auswirkung auf die geschätzten Regressionskoeffizienten eines allometrischen Modells untersucht werden. Im Rahmen der vorliegenden Untersuchung soll hierfür auf einen zusammengestellten Fichtendatensatz zurückgegriffen werden (siehe 2.8).

Ein von PAIN und BOYER (1996) entwickeltes Schaftformmodell (im weiteren als *Pain-Funktion* bezeichnet) hat sich als geeignet erwiesen, um die Schaftformvariabilität von Fichten mit Hilfe der Einzelbaumvariablen BHD, Höhe und h/d-Wert zu beschreiben (SCHMIDT, 2001; GADOW, 2003). Da aufgrund der Anpassung der Schaftkurve mit den gegebenen Parametern des Modells, der Funktionswert der Pain-Funktion in 1,3 m Höhe nicht zwingend mit dem BHD übereinstimmt, schlägt Schmidt (2001) ein modifiziertes Modell vor, das diese Abweichung minimiert. Da hier der Effekt der Durchmessertransformation und seine Auswirkung in Bezug auf die Koeffizienten des allometrischen Modells nur beispielhaft dargestellt wird, soll hier auf die Originalversion der Pain-Funktion zurückgegriffen werden:

$$r\left(h_{rel}\right) = \alpha\left(1 - h_{rel}^{3}\right) + \beta\left(\ln\left(h_{rel}\right)\right) \tag{6}$$

Wobei:
$r\left(h_{rel}\right)$ = Schaftradius in relativer Baumhöhe;
α, β = dimensions- und formbeschreibende Parameter.

Die dimensions- und formbeschreibenden Parameter des Modells werden dabei in Abhängigkeit der formbeeinflussenden Variablen BHD und Baumhöhe wie folgt geschätzt:

$$\alpha = a_0 + a_1\left(\frac{1}{\ln\left(H^{1/D}\right)}\right) + a_2\left(\frac{1}{\left(H/D\right)^2}\right) \tag{7}$$

und:

$$\beta = b_0 + b_1\left(\frac{1}{\ln\left(H^{1/D}\right)}\right) + b_2\left(\frac{1}{\left(H/D\right)^2}\right) \tag{8}$$

Wobei:
α, β = dimensions- und formbeschreibende Parameter;
H = Baumhöhe (m);
D = BHD (cm);
$a_0,\ a_1,\ a_2,\ b_0,\ b_1,\ b_2$ = Parameter.

Im Rahmen dieser Untersuchung wurde auf Parameter zurückgegriffen, die für 827 Fichten aus dem Bereich Nordwest-Deutschland angepasst wurden (Schmidt, 2001). Zwar deckt somit der Parametrisierungsbereich des Modells nicht die breite geografische Herkunft der verwendeten Daten (siehe 2.8) ab, dies ist jedoch aufgrund der unterschiedlichen Datengrundlagen in diesem Fall auch von keinem anderen Modell zu erwarten. Die geschätzten Koeffizienten des verwendeten Modells sind in Tabelle 2-1 aufgeführt.

Tabelle 2-1. Koeffizienten und jeweiliger Standardfehler der Schätzung zur Herleitung der dimensions- und formbeschreibenden Parameter α und β der Pain-Funktion.

Koeffizient	Wert	Std. Fehler
a_0	-0,223	0,0615
a_1	1,595	0,0138
a_2	-3,155	0,0667
b_0	0,512	0,0333
b_1	-0,158	0,0075
b_2	-0,502	0,0362

Wie am Beispiel in Abbildung 2-4 zu sehen ist, scheint der berechnete Durchmesser in 10 % der Baumhöhe ($D_{0,1}$) sehr viel weniger von der individuellen Form des Stammanlaufes beeinflusst als der BHD. Da somit ein Großteil der Formvariabilität, die auf einer festen Messhöhe erfasst würde, vermieden wird, scheint der Durchmesser in relativer Stammhöhe als Eingangsgröße allometrischer Funktionen theoretisch besser geeignet, da er eine relativ formunabhängige Beschreibung der Dimensionsvariabilität erlaubt.

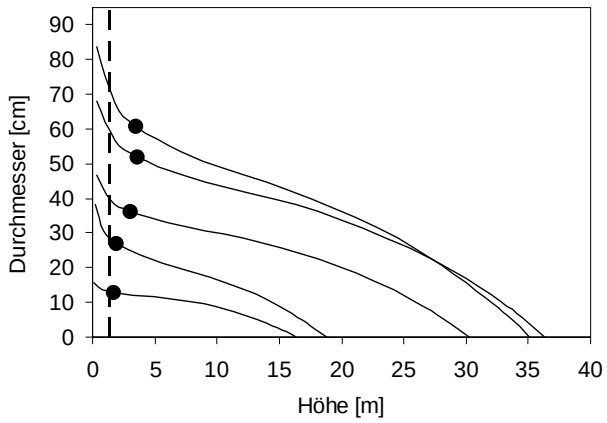


Abbildung 2-4. Lage des BHD (gestrichelte Linie) und des $D_{0,1}$ (Punkte) an generierte Schaftformprofilen von 5 Fichten.

Dieser Argumentation folgend stellt somit erwartungsgemäß die Baumhöhe eine weitere wichtige Eingangsgröße zur Biomassemodellierung dar.

2.3 Die *k*-Nearest Neighbour Methode

Die *k*-Nearest Neighbour (*k*-NN) Methode ist ein nicht-parametrisches, instanzenbasiertes Klassifizierungsverfahren und eine der einfachsten und ältesten Methoden der Mustererkennung im Bereich des maschinellen Lernens (ALTMAN, 1992, KOTZ et al., 1998). Sie wurde schon früh als ein Verfahren der nicht-parametrischen Diskriminanzanalyse beschrieben (ATKESON et al., 1997). Frühe Beispiele für einen „lazy similarity learning algorithm" finden sich z.B. bei FIX und HODGES (1951) oder COVER und HART (1967). Besonders mit der Weiterentwicklung und dem Fortschritt im Bereich der computergestützten Datenanalyse steigerte sich das Interesse an nicht-parametrischen Analysemethoden erheblich (FAN, 2000).

Die Grundidee des Verfahrens besteht darin, ein unbekanntes Merkmal eines Objektes aufgrund dessen Ähnlichkeit zu bekannten Objekten zu schätzen. Forstliche Anwendungen der *k*-NN Methode finden sich beispielsweise bei HAARA et al. (1997), MALTAMO und KANGAS (1998), NIGGEMEYER (1999), NIGGEMEYER und SCHMIDT (1999), TOMMOLA et al. (1999), HESSENMÖLLER (2001), MALINEN (2003a, 2003b), NIESCHULZE et al. (2005), wobei der methodische Ansatz hier im Wesentlichen zur Schätzung von Bestandesparametern oder als Alternative zu parametrischen Wachstumsmodellen verwendet wird. SIRONEN et al. (2001, 2003) verwenden den nicht-parametrischen Ansatz zur Ableitung von Wachstumsmodellen auf Einzelbaumebene. Weitere Ansätzen zur Schätzung von Einzelbaumvariablen finden sich außerdem bei HOLM et al. (1997) oder KORHONEN und KANGAS (1997).

Einen Vergleich verschiedener nicht-parametrischer Ansätze zur Schätzung von Einzelbaumvariablen findet sich bei MALINEN et al. (2003). Hier wird unter anderem auch die *k*-NN Methode verwendet, um Rückschlüsse auf innere Holzeigenschaften und –Qualität abzuleiten. LEMM et al. (2005) verwenden die *k*-NN Methode zur Vorhersage von Produktivitäten in der Holzernte. Weiterhin gibt es seit langem weitreichende Anwendungen im Bereich der Klassifizierung von Rasterdaten aus digitalen Luftbildern oder Satellitenbildern bzw. in mehrphasigen Stichprobeninventuren wie z.B. der „Finnish multisource National Forest Inventory (MS-NFI)" (z.B. KILKKI und PÄIVINEN, 1987; MOER, 1987; TOMPPO, 1991; MOER und STAGE, 1995; MOER und HERSHEY, 1999; ANTTILA, 2002; TEMESGEN, 2003; LEMAY und TEMESGEN, 2004) oder der US-Amerikanischen Nationalen Waldinventur (siehe z.B. McRoberts et al., 2002).

Hierbei wird die k-NN Methode genutzt, um eine Integration von Fernerkundungsdaten und Feldaufnahmen durch einen Vergleich der spektralen Signaturen von Fernerkundungssensoren und den terrestrischen Daten zu ermöglichen (HOLMSTRÖM et al., 2001; MALINEN, 2003; STÜMER und KÖHL, 2005).

2.3.1 Theoretischer Hintergrund der k-NN Methode

Der Grundlegende Unterschied des k-NN Ansatzes zu anderen Lernverfahren besteht darin, dass a priori keine explizite Hypothese oder Beschreibung der Zielfunktion zugrunde gelegt wird, sondern jegliche Generalisierungsentscheidung nur zum Zeitpunkt einer konkreten Suchanfrage stattfindet (sog. *lazy learning*). Die k-NN Methode ist daher nicht, wie z.B. Regressionsansätze, auf eine explizite Modellbildung angewiesen. Die Funktionsweise des k-NN Algorithmus beruht darauf, zunächst alle empirischen Datengrundlagen zu speichern (sog. Trainingsdaten[1]). Ein unbekanntes Merkmal eines neuen Individuums (Instanz[2]) wird dann durch einen Vergleich mit den bekannten Trainingsinstanzen klassifiziert.

Zur Schätzung einer unbekannten Instanz werden diejenigen in allen Merkmalen (Attribute, Variablen) bekannten Trainingsbeispiele herangezogen, die in Bezug auf die bekannten Merkmale des neuen Individuums den geringsten Abstand aufweisen. D.h. es werden aus den vorhandenen Trainingsdaten „ähnliche“ Merkmalsträger identifiziert und unter der Annahme, dass sie sich auch in Bezug auf das gesuchte Merkmal „ähneln“, zur Schätzung der Zielgröße verwendet. Die an allen Instanzen bekannten

1 Synonyme Begriffe sind z.B. Lerndatenpunkte, Referenzdaten, Trainingsinstanzen, Trainingsbeispiele.
2 Objekte oder Individuen einer bestimmten Grundform (hier Baum), deren Ausprägung durch beschreibende Variablen oder Attribute definiert sind, werden im Folgenden als Instanzen bezeichnet.

bzw. messbaren Attribute werden auch als *Indikatorattribute* bezeichnet. Die lediglich von den Trainingsinstanzen bekannten Zielgrößen bilden die Gruppe der sog. *Designattribute* (MOER und STAGE, 1995).

Die Prognose erfolgt dann, indem die Merkmalswerte derjenigen k Trainingsbeispiele, die zu der zu prognostizierenden Eingangsraumposition am nächsten liegen, entweder im Sinne eines Mehrheitsentscheides oder einer gewichteten Mittelwertbildung ausgewertet werden (HAENDEL, 2003). Hierbei wird also ein funktionaler Zusammenhang zwischen den bekannten Merkmalen und der Zielgröße unterstellt jedoch nicht explizit in einem Modell formuliert.

Im Gegensatz zu parametrischen Verfahren, bei denen zur Beschreibung eines Modells ein Vorwissen in Bezug auf die Zusammenhänge der wirkenden Faktoren bestehen muss (deduktives Lernen), beruht die k-NN Methode auf einem induktiven Lernansatz. Induktive Lernmethoden benötigen wenig Vorkenntnis über eventuelle Zusammenhänge, dafür jedoch eine relativ hohe Anzahl an Trainingsbeispielen.

Anders als bei einer Regressionsanalyse, die eine Funktionsangleichung über den gesamten Wertebereich zum Ziel hat, ist das Ergebnis des k-NN Algorithmus eine lokale Approximation, die über die k- nächsten (ähnlichsten) Nachbarn abgeleitet wird und sich somit bei jeder neuen Berechnung ändert (MITCHELL, 1997). Hierdurch ergibt sich im Fall von sehr komplexen Zusammenhängen der Vorteil, dass sich die Zielfunktion durch eine Auswahl weniger komplexer lokaler Approximationen beschreiben lässt. Im Folgenden soll näher auf die Eigenschaften und Grundlagen der k-NN Methode eingegangen werden. Hierzu gehört vor allem die Quantifizierung der Ähnlichkeit von Instanzen aufgrund ihrer Merkmale.

2.4 Quantifizierung von Ähnlichkeiten

Um die Ähnlichkeit bzw. Unähnlichkeit von Instanzen auf der Grundlage metrischer Variablen quantifizieren zu können, müssen sie als Punkte im euklidischen Raum ($\Re^n$) darstellbar sein. Auf diese Weise kann ihre Distanz zueinander als Maß ihrer Unähnlichkeit berechnet werden und durch eine einfache Transformation in ein Ähnlichkeitsmaß (Proximitätsmaß) überführt werden. Aus dem Bereich der multivariaten Analysemethoden wie z.B. der Diskriminanz- oder Clusteranalyse sind

verschiedene Distanzmaße bekannt, die sich zur Verwendung im Rahmen der k-NN Methode eignen. Ein oft verwendetes Distanzmaß ist die Euklidische Distanz:

$$d\left(x_i, x_j\right) = \sqrt{\sum_{r=1}^{R}\left(x_{ir} - x_{jr}\right)^2} = \sqrt{\left(\vec{x}_i - \vec{x}_j\right)^T\left(\vec{x}_i - \vec{x}_j\right)} \tag{9}$$

mit:
$d(x_i,x_j)$ = Distanz zwischen den Instanzen x_i und x_j (hier euklidische);
x_{ir}, x_{jr} = Wert der r-ten Variable bei Instanz i, j; ($r = 1,2,\ldots,R$) bzw.
$\vec{x}_i, \vec{x}_j$ = Ein geordneter Vektor der diskreten Variablen.

Durch eine Gewichtung der einzelnen Variablendifferenzen, auf die später näher eingegangen wird, erhält man die sog. diagonal gewichtete euklidische Distanz mit:

$$d_w\left(x_i, x_j\right) = \sqrt{\sum_{r=1}^{R}\left(w_r\left(x_{ir} - x_{jr}\right)\right)^2} = \sqrt{\left(\vec{x}_i - \vec{x}_j\right)^{\mathbf{T}}\mathbf{M}^{\mathbf{T}}\mathbf{M}\left(\vec{x}_i - \vec{x}_j\right)} \tag{10}$$

Wobei hier w_r ein Gewichtungsfaktor für die r-te Variable ist. M ist eine Diagonalmatrix in der $M_{rr} = w_r$.

In der praktischen Anwendung wird oft die Verallgemeinerung dieser Distanz, die sog. Minkowski-Metrik oder L-Norm verwendet, die sich im Fall ungewichteter Distanzen wie folgt berechnet (BACKHAUS et al., 2005):

$$d\left(x_i, x_j\right) = \left[\sum_{r=1}^{R}\left|x_{ir} - x_{jr}\right|^c\right]^{\frac{1}{c}} \tag{11}$$

mit:
$d(x_i,x_j)$ = Distanz der Instanzen i und j;
x_{ir}, x_{jr} = Wert der Variablen r bei Instanz i, j ($r = 1,2,\ldots,R$);
$c \geq 1$ = Minkowski-Konstante.

Dabei ergibt sich für eine Minkowski-Konstante von c =1 die sog. City-Block-Metrik (auch Manhattan- oder Taxifahrer-Metrik), die der Summe der Varieblendifferenzen entspricht. Hierbei werden alle Differenzwerte unabhängig von ihrer Größe gleich gewichtet. Für c =2 ergibt sich die euklidische Distanz (L2-Norm), bei der größere Differenzen ein stärkeres Gewicht erhalten als kleine (Abbildung 2-5).

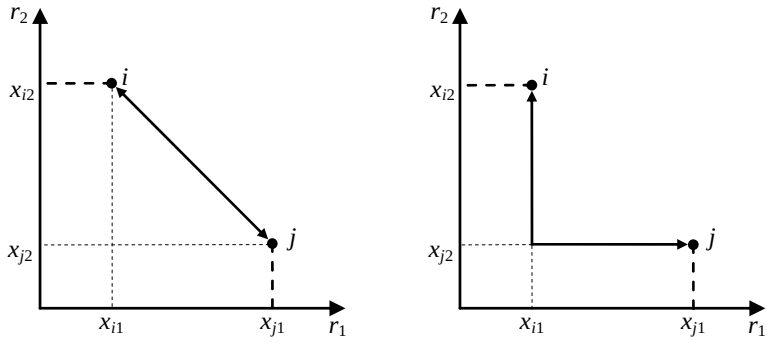


Abbildung 2-5. Euklidische Distanz (links) und City-Block-Distanz (rechts) zwischen zwei Punkten i und j im zweidimensionalen Raum.

Im Rahmen der k-NN Methode wird im Allgemeinen die euklidische Distanz verwendet. In der Literatur finden sich jedoch auch zahlreiche Beispiele zur Verwendung des einfachen Differenzbetrages als Distanzmaß oder vergleichende Gegenüberstellungen beider Maße (z.B. SIRONEN, 2003; HESSENMÖLLER, 2002). Die Wahl der City-Block-Metrik kann in den Fällen sinnvoll sein, in denen mit zufällig überhöhten Merkmalsdifferenzen (Ausreißern) gerechnet werden muss. Bemerkenswert ist weiterhin, dass der Unterschied zwischen euklidischer- und Manhattan Distanz mit zunehmender Anzahl von Dimensionen größer wird. Während die Wahl des Parameters c bei Verwendung von nur zwei Dimensionen noch einen relativ geringen Einfluss hat, erhöht sich die Auswirkung bei Verwendung von mehreren Variablen.

Es ist nicht ungewöhnlich, dass die beschreibenden Variablen, deren Distanzen zu einem Gesamtabstand summiert werden, untereinander korreliert sind (HOLMSTRÖM, 2001). In diesem Fall erhöht sich der Einfluss dieser korrelierten Variablen in der Abstandsberechnung im Vergleich zu unkorrelierten. Sind beispielsweise zwei Variablen vollkommen korreliert, führt dies in der Distanzberechnung zu einer doppelten Bewertung ihres Einflusses auf den Gesamtabstand zwischen zwei Instanzen. Ein

Distanzmaß, das die Kovarianz zwischen Variablen berücksichtigt und korrelierten Variablen aufgrund ihres geringen Trennungscharakters ein kleineres Gewicht in der Abstandsberechnung zuordnet, ist die Mahalanobis Distanz (BORTZ, 2004), auf die unter 2.4.1 weiter eingegangen werden soll. Hierbei sind jedoch einheitliche Mittelwerte der Variablen eine Voraussetzung (BACKHAUS, 2005). Abbildung 2-6 gibt einen Überblick über ausgewählte Distanzmaße.

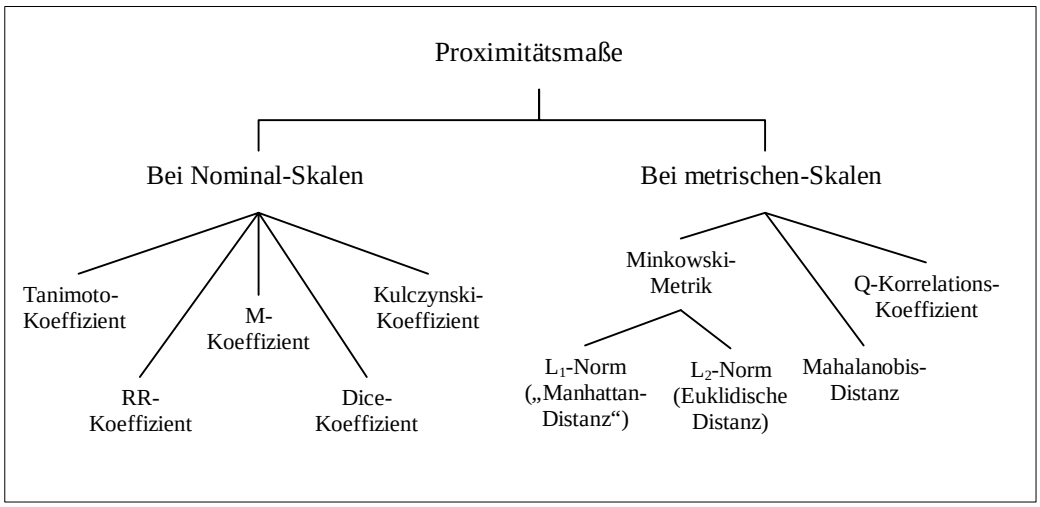


Abbildung 2-6. Überblick über ausgewählte Proximitätsmaße für nominal und metrisch skalierte Daten (nach BACKHAUS et al., 2005).

Neben den Distanzmaßen für metrische Variablen gibt es eine Reihe von Proximitätsmaßen für nominal skalierte Merkmale, auf die hier jedoch nicht weiter eingegangen werden soll.

2.4.1 Mahalanobis Distanz

Da im vorliegenden Fall ein Distanzmaß zur Berechnung des Abstandes zwischen Baumindividuen verwendet werden soll, wird hierbei teilweise auf Einzelbaumvariablen zurückgegriffen. Da diese Variablen naturgemäß oftmals eine hohe Korrelation untereinander aufweisen, kann dies wie bereits unter 2.4 angemerkt, zu einer Verzerrung der Gewichtung führen. Mit der Mahalanobis Distanz (Mahalanobis, 1963) erhält man ein generalisiertes euklidisches Distanzmaß, das die korrelative Beziehung

zwischen den Merkmalen berücksichtigt (Bortz, 2005). Dieses Distanzmaß wird auch im speziellen Fall des als MSN (Most similar neighbour) Methode bezeichneten Ansatzes von MOER und STAGE (1995) verwendet. Weitere Anwendungen dieses Distanzmaßes finden sich z.B. in ANTTILA (2002), HOLMSTRÖM et al. (2001), MUINONEN et al. (2001), MALTAMO et al. (2003), TEMESGEN (2003) und NIESCHULZE et al. (2005).

Dieses Distanzmaß entspricht einer voll gewichteten euklidischen Distanz, in der M nicht länger eine Diagonalmatrix der Gewichte darstellt, sondern durch eine kanonische Korrelation hergeleitet wird. Die Gewichtungsmatrix K^{-1} ist in diesem Fall die Inverse der Kovarianzmatrix der verwendeten Variablen und kann auch nicht-diagonale Elemente enthalten. Die Mahalanobis Distanz wird dementsprechend auch als voll gewichtete Distanz bezeichnet (ATKESON et al., 1996).

$$d_M^2\left(x_i, x_j\right) = \left(x_i - x_j\right)^{\mathrm{T}} \mathbf{K}^{-1} \left(x_i - x_j\right) \tag{12}$$

Wobei x_i und x_j die m-dimensionalen Spaltenvektoren der verwendeten Variablen zweier Instanzen sind und K^{-1} die Inverse der symmetrischen Kovarianzmatrix darstellt (und T für transponiert steht). Die Mahalanobis Distanz ist für solche Anwendungen geeignet, bei denen starke Merkmalskorrelationen auftreten, diese Korrelationen selbst aber keine inhaltliche Bedeutung für die Proximität haben. Sie nimmt zu, wenn die Korrelation bzw. die empirische Kovarianz zwischen den Variablen abnimmt. In speziellen Fällen entspricht die Mahalanobis Distanz der einfachen quadrierten Euklidischen Distanz. Werden z.B. die metrischen Merkmalsvariablen durch eine z-Transformation normiert (siehe hierzu auch 2.5.1), so bildet die Kovarianzmatrix unter der Annahme gleicher Streuungen die Einheitsmatrix E. Dies entspricht wiederum einer einheitlichen diagonalen Gewichtung aller einbezogenen Variablen. Weiterhin sind beide Distanzmaße identisch, wenn keinerlei Korrelationen zwischen den Variablen vorliegen.

2.4.2 Der Q-Korrelationskoeffizient

Die bisher beschriebenen Distanzmaße sind in der Lage, die Unähnlichkeit von Instanzen in Abhängigkeit der normierten Niveauunterschiede der Variablen zu quantifizieren. Zu berücksichtigen bleibt jedoch, dass eine Mustererkennung in Bezug auf die Profilverläufe der Variablen nur eingeschränkt möglich ist. Der Q-Korrelationskoeffizient (bzw. Pearson Korrelationskoeffizient) ist im Gegensatz zu den dargestellten Distanzmaßen ein Proximitätsmaß, das die Ähnlichkeit zwischen Objekten bzw. Instanzen ausschließlich aufgrund des Verlaufs der Variablenprofile beschreibt.

$$r_{i,j} = \frac{\sum_{r=1}^{R}\left(x_{ir} - \overline{x}_i\right)\left(x_{jr} - \overline{x}_j\right)}{\sqrt{\sum_{r=1}^{R}\left(x_{ir} - \overline{x}_i\right)^2 \cdot \sum_{r=1}^{R}\left(x_{jr} - \overline{x}_j\right)^2}} \tag{13}$$

wobei:
r_{ij} = Q-Korrelationkoeffizient;
x_{ir}, x_{jr} = Wert der Variablen r bei Objekt i, j ($r = 1,2,\ldots,R$);
$\overline{x}_i, \overline{x}_j$ = Mittelwert aller Variablen bei Instanz i oder j.

Zur Verdeutlichung des Unterschiedes bzw. der Interpretation des Q-Korrelationskoeffizienten im Vergleich zu den dargestellten Distanzmaßen, ist in Abbildung 2-7 der Profilverlauf von 4 hypothetischen Variablen an 3 verschiedenen Instanzen dargestellt. Die Distanzen bzw. der Q-Korrelationskoeffizient werden hier relativ zu einer Abfrageinstanz berechnet.

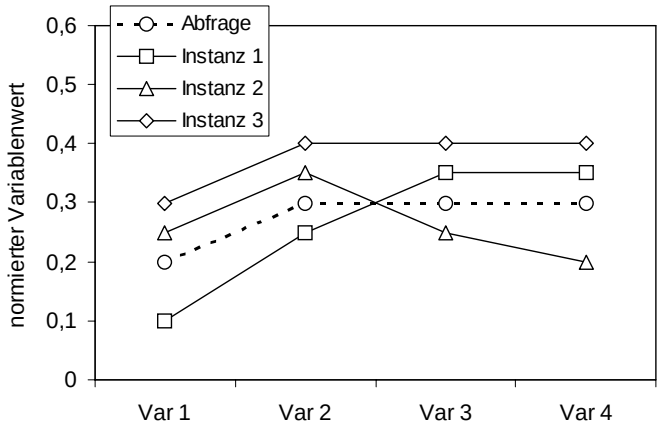


Abbildung 2-7. Unterschiedliche Profilverläufe von 4 Variablen (Var 1 – Var 4) bei verschiedenen Instanzen. Die Abfrageinstanz ist mit gestrichelter Linie dargestellt.

Würde man die Ähnlichkeit der oben dargestellten Instanzen zur gegebenen Eingangsraumposition bzw. Abfrageinstanz unter der Annahme der Gleichgewichtung der Variablen auf Grundlage der Minkowski Metrik quantifizieren, so hätten die Instanzen 1 und 2 den gleichen Abstand und würden somit in der Reihenfolge der k nächsten Nachbarn den gleichen Rang belegen. Offensichtlich unterscheiden sie sich jedoch erheblich in Bezug auf den Profilverlauf der berücksichtigten Variablen. Da die euklidische Distanz durch die Quadrierung der Variablendifferenzen zwar in der Lage ist, Abweichungen von Differenzbeträgen einzelner Variablen aufzudecken, hierbei jedoch die Richtung der Abweichung nicht berücksichtigt wird, können unterschiedliche Profilverläufe der Variablen kaum nachgewiesen werden. In Bezug auf das dargestellte Beispiel würde eine unterschiedliche Gewichtung der Variablen zwar zu verschiedenen Abständen der Instanzen 1 und 2 führen, ob jedoch die resultierende Reihenfolge in diesem Fall der implizierten Beziehung zwischen Suchvariablen und der zu schätzenden Zielgröße gerecht wird, bleibt fraglich. Tabelle 2-2 zeigt die für das gegebene Beispiel berechneten Distanzen und Q-Korrelationskoeffizienten.

Tabelle 2-2. Gegenüberstellung der euklidischen Distanz und des Q-Korrelationskoeffizienten in Bezug auf eine Abfrageinstanz für 3 hypothetische Trainingsinstanzen mit unterschiedlichem Profilverlauf der berücksichtigten Variablen.

Instanz	Ungewichtete euklidische Distanz	Q-Korrelationskoeffizient
Instanz 1	0,132	0,917
Instanz 2	0,132	0,132
Instanz 3	0,200	1,000

Da die Abfrageinstanz und Instanz 3 ein kongruentes Variablenprofil aufweisen, werden sie auf Grundlage des Q-Korrelationskoeffizienten als „am ähnlichsten“ bewertet, obwohl die ungewichtete euklidische Distanz hier am größten ist.

Die Aussagekraft dieses Proximitätsmaßes bei der abstandsgemäßen Sortierung von Instanzen ist also stark von einer eventuellen Kovarianz der verwendeten Variablen abhängig. Um den Einfluss unterschiedlicher Variablenprofile auf die Schätzung der Zielgröße zu überprüfen, kann die Beziehung der Residuen der Schätzung aus einer Kreuzvalidierung in Abhängigkeit des jeweiligen Q-Korrelationskoeffizienten untersucht werden. Falls sich die Streuung der Residuen in bemerkenswertem Umfang durch Unterschiede in den Variablenprofilen erklären lässt, könnte der Q-Korrelationskoeffizient z.B. als zusätzlicher Gewichtungsfaktor in die Abstandsberechnung einbezogen werden.

2.5 Der *k*-NN Algorithmus

Eine Instanz x wird als Vektor im n-dimensionalen Vektorraum abgebildet, wobei die Anzahl der Dimensionen (Achsen) durch die Anzahl der betrachteten Attribute einer Instanz vorgegeben ist:

$$< a_1(x), a_2(x), ...a_n(x) > \tag{14}$$

wobei hier $a_r(x)$ eine Bezeichnung für das r-te Attribut der Instanz x ist.

Im Fall einer konkreten Suchanfrage werden die Distanzen zu allen bekannten Instanzen in der Trainingsdatenbank berechnet und durch Invertierung der Distanz die k-nächsten Nachbarn identifiziert. Der k-NN Ansatz eignet sich dabei gleichermaßen zur Klassifizierung von diskreten und stetigen Zielfunktionen, wobei im Fall diskreter Variablen eine Zuordnung der Zielgröße nach der maximalen Häufigkeit eines Merkmals innerhalb der k-nächsten Nachbarn erfolgt, und im stetigen Fall ein (gewichteter) Mittelwert aus den k-nächsten Nachbarn gebildet wird (HESSENMÖLLER, 2001).

Im Fall einer diskreten Zielfunktion der Form $f : \Re^n \rightarrow V$, in der V eine endliche Liste der Funktionswerte ist, gibt der k-NN Algorithmus also den häufigsten Wert der k Nachbarn zurück:

$$\hat{f}(x_q) \leftarrow \arg\max_{v \in V} \sum_{i=1}^{k} \delta(v, f(x_i)) \qquad \text{mit} \qquad \delta(a,b) = \begin{cases} 1 & a = b \\ 0 & a \neq b \end{cases} \tag{15}$$

Wobei:

x_q = zu klassifizierende Instanz;
x_i = k nächste Nachbarn der Trainingsinstanzen;
V = endliche Liste von Funktionswerten $\{v_1, ...v_s\}$.

Zur lokalen Approximation einer stetigen Zielfunktion $f : \Re^n \rightarrow \Re$ wird stattdessen der Mittelwert aus den Werten der k nächsten Nachbarn der Trainingsinstanzen berechnet:

$$\hat{f}(x_q) \leftarrow \frac{\sum_{i=1}^{k} f(x_i)}{k} \tag{16}$$

Im Fall der Suche nach nur einem nächsten Nachbarn (k =1) ergibt sich im zweidimensionalen Fall für jeden Punkt p eine endliche Menge von Punkten, die p als nächsten Nachbarn haben. Die Polygonflächen die die Menge dieser Punkte bilden werden als Voronoi Regionen und ihre Darstellung (Abbildung 2-8) als Voronoi Diagramm[3] bezeichnet.

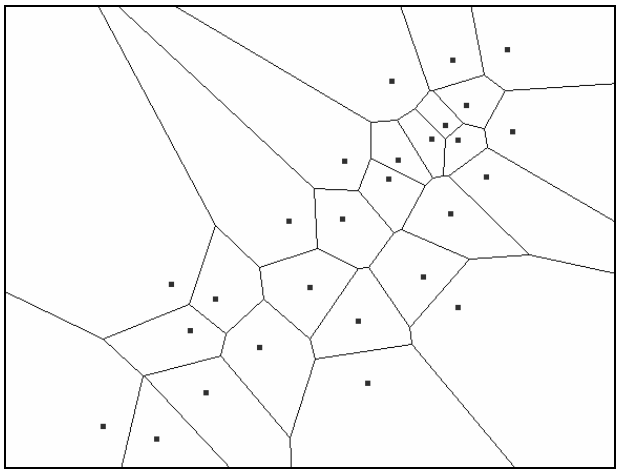


Abbildung 2-8. Voronoi Diagramm der L2-Norm (Euklidische Norm) mit Voronoi Regionen für jede Instanz.

Wie aus Formel (9) deutlich wird, berechnet sich die euklidische Distanz zwischen zwei Instanzen aus der Summe der Einzelabstände ihrer Attribute zueinander. Hierbei

[3] eine verbreitete Bezeichnung ist auch Diriclet- Region oder Mosaik bzw. Thiessen Polygon.

werden diese Einzelabstände ungewichtet und ohne Beachtung ihres jeweiligen Einflusses auf die Zielgröße verwendet und zu einem Gesamtabstand zwischen den Instanzen aufsummiert. Diese Vorgehensweise kann bei Betrachtung aller Attribute dazu führen, dass es aufgrund des Einflusses von Variablen, die für die Schätzung der Zielgröße weitgehend unbedeutend sind, zu Fehlklassifikationen kommt.

Sind beispielsweise aus einem Satz von 20 beschreibenden Attributen einer Instanz nur wenige für die Ausprägung der Zielgröße relevant, so führen die zusätzlich in die Berechnung einbezogenen Variablen zu einer Verzerrung des Abstands. Um den Einfluss relativ unbedeutender Variablen auf den Gesamtabstand zwischen den Instanzen abzuschwächen, bzw. den Einfluss bedeutender Variablen zu verstärken, müssen die Einzelabstände der betrachteten Attribute gewichtet werden (MITCHELL, 1997; TOMMOLA et al., 1999; HESSENMÖLLER, 2002).

$$d_w(x_i, x_j) = \left[\sum_{r=1}^{R} \left(w_r \left| x_{ir} - x_{jr} \right| \right)^c \right]^{\frac{1}{c}} \tag{17}$$

wobei:
d_w = gewichteter Gesamtabstand zwischen Instanzen;
w_r = Gewichtungsfaktor für den Abstand der r-ten Variablen;
x_{ir}, x_{jr} = Wert der r-ten Variablen der Instanz i, j ($r = 1,2,\ldots,R$);
c = Minkowski Konstante.

Weiterhin ist zu beachten, dass die verwendeten Variablen sehr unterschiedliche Wertebereiche aufweisen können und daher durch ihre unterschiedliche Skalierung einen unausgeglichenen Einfluss auf die Abstandsberechnung haben können. Da die Gewichte, die für einzelne Variablen in der Abstandsberechnung verwendet werden, in ihrer Summe 1 ergeben (vgl. 2.5.2), können sie die Verzerrung der Distanzen nicht ausgleichen. Üblicherweise geht man daher von vereinheitlichten Maßstäben aus, bei denen durch eine Normierung des Merkmalraums aller beteiligten Variablen der Skaleneinfluss eliminiert wird.

2.5.1 Normierung des Merkmalraums

Da angestrebt wird, dass alle betrachteten Merkmale gleichermaßen in den resultierenden Gesamtabstand eingehen, ist eine vorherige Normierung des Merkmalraums der Variablen notwendig (ALTMAN, 1992; BACKHAUS et al., 2005; HAENDEL, 2003), um den Skaleneinfluss zu eliminieren.

Zur Normierung der Daten bieten sich verschiedene Verfahren an. Um z.B. die Abweichung zweier Variablen vom jeweiligen Mittelwert vergleichbar zu machen, müssen sie zunächst an der Unterschiedlichkeit aller Ausprägungen dieses Merkmals innerhalb der Trainingsdaten relativiert werden. Hierzu kann eine z-Transformation oder Standardisierung vorgenommen werden.

$$x'_{ir} = \frac{x_{ir} - \overline{x}_{ir}}{\sigma_{ir}} \tag{18}$$

Eine andere Möglichkeit besteht darin, den Wertebereich der Variablen auf das Intervall [0, 1] zu normieren, was der Berechnung von Prozenträngen gleichkommt. Bei dieser Vorgehensweise können jedoch Ausreißer einen starken Einfluss auf die Skalierung haben.

HESSENMÖLLER (2001) bzw. HESSENMÖLLER UND ELSENHANS (2002) führen eine Normierung mit der Standardabweichung der Variablen durch, wobei hierbei zwar die Breite der Verteilung, nicht aber ihre absolute Lage (wie in Formel (17) durch den Mittelwert gegeben) normiert wird. Da durch die Normierung letztendlich eine Vergleichbarkeit der Variablendifferenzen hergestellt werden soll, können statt der Ausgangsdaten auch die, auf der Grundlage der Ausgangsdaten berechneten, Distanzwerte der einzelnen Variablen auf ein einheitliches Intervall normiert werden. In der dargestellten Distanzfunktion treten die Abstände der berücksichtigten Variablen explizit auf, so dass hier eine Integration der Normierung in die Abstandsfunktion möglich ist:

$$d'_w(x_i, x_j) = \left[\sum_{r=1}^{R} \left(w_r \frac{|x_{ir} - x_{jr}|}{\delta_r} \right)^c \right]^{\frac{1}{c}} \tag{19}$$

Hierbei ist δ_r ein Normierungsfaktor, der einfach an die Spannweite der jeweiligen Variablen in den Trainingsdaten gekoppelt werden kann. Setzt man $\delta_r = x_{r\ max}$-$x_{r\ min}$ liegen die Abstände der einzelnen Variablen im Intervall [0,1]. Da eine Normierung nur aufgrund der in der Datenbank enthaltenen Trainingsinstanzen vorgenommen wird, kann die Distanz zu einer unbekannten Instanz dieses Intervall jedoch verlassen. Dies ist davon abhängig, inwieweit die Trainingsdaten den gesamten möglichen Merkmalraum der betreffenden Variablen abdecken.

HEANDEL (2003) schlägt zur Abmilderung des Einflusses von Ausreißern eine Kopplung von δ_r an die Standardabweichung oder ein Vielfaches davon vor. Bei Verwendung der vierfachen Standardabweichung ($\delta_r = 4\sigma_r$) liegen so die Abstände je nach Grad ihrer Normalverteiltheit meistens im gewünschten Intervall [0, 1].

2.5.2 Gewichtung der Variablendifferenzen

Im einfachen k-NN Ansatz werden die Verschiedenartigkeiten in den einzelnen Dimensionen in einer „naiven“ Weise mit einer gleichen Gewichtung für alle Variablen zu einem Gesamtabstand zwischen zwei Instanzen summiert (bzw. die Gewichtung ist hierbei für alle Merkmale w_r =1). Dieser Ansatz ist deswegen unrealistisch, weil auf diese Weise auch irrelevante Merkmale die Abstandsberechnung beeinflussen. Dieses Problem wird auch als „Fluch der Dimensionalität“ bezeichnet (BELLMANN, 1961). Es gibt eine Vielzahl von verschiedenen Ansätzen um realistische Gewichtungsfaktoren für die einzelnen Merkmale in den k-NN Ansatz zu integrieren. Die Gewichtung der Variablenabstände ist dabei klar von der Normierung der Variablen zu trennen. Während die Normierung die Vergleichbarkeit der Differenzwerte verschiedener Variablen gewährleisten soll, ist die Gewichtung Ausdruck der Relevanz bzw. Irrelevanz einer Variablen für die gesuchte Zielgröße.

Obgleich die k-NN Methode generell ein nicht-parametrischer Ansatz im Hinblick auf die Suche nach den k nächsten Nachbarn ist, gilt dies nicht gleichermaßen für die Distanzfunktion, die für die einzelnen Variablen gebildet wird. Die Gewichtung der Variablendifferenzen sowie alle weiteren Parameter der Abstandsmetrik werden global definiert und somit deterministisch festgelegt. Die Gewichtung der Einzelabstände kommt dabei einer Skalierung aller Achsen im n-dimensionalen Raum gleich, wobei die Achsen entsprechend ihres Einflusses auf die gesuchte Zielgröße gestreckt oder gestaucht werden. Variablen, die keinen nachweislichen Einfluss auf die Zielgröße haben, können auf diese Weise auch ganz eliminiert und somit von der Abstandsberechnung ausgeschlossen werden (AHA, 1998; WETTSCHERECK, 1995).

Die Bestimmung der Gewichtungsfaktoren für die einzelnen Variablenabstände kann einen großen Einfluss auf die Erkennung der k nächsten Nachbarn haben, und ist somit entscheidend für die Qualität der Schätzung der Zielgröße. Zur Bestimmung geeigneter Gewichtungsfaktoren bieten sich verschiedene Möglichkeiten. Die Gewichtung der Abstände einzelner Variablen in der Distanzfunktion kann durch einen iterativen Prozess mit Hilfe einer Kreuzvalidierung hergeleitet werden. Hierbei wird für jede Instanz der Trainingsdaten eine Schätzung der Zielgröße aus den N-1 verbleibenden Instanzen abgeleitet. Durch die Modifikation der beeinflussenden Gewichte in der Schätzung kann eine Annäherung an ein optimales Gewichtungsverhältnis erzielt werden. Bedenkt man hierbei allerdings die Vielzahl der möglichen Gewichtungsverhältnisse bei einer steigenden Anzahl von berücksichtigten Variablen, erfordert dieses Vorgehen ohne Vorkenntnisse über deren Einfluss einen enormen Rechenaufwand, der allenfalls durch iterative Optimierungsalgorithmen wie z.B. dem Simulated Annealing oder Genetic Algorithm verringert werden kann (siehe z.B. TOMPPO und HALME, 2004).

Eine weitere Alternative zur Herleitung einer Distanzfunktion ist die Verwendung eines Regressionsmodells zur Bestimmung des Distanzmaßes (NIGGEMEYER, 1999; HOLMSTRÖM et al., 2001). Hierbei wird ein geeignetes Regressionsmodell angepasst und die geschätzten Regressionskoeffizienten, die letztendlich das Verhältnis des Einflusses der Variablen auf die Zielgröße widerspiegeln, zur Herleitung der Gewichtungsfaktoren für die k-NN Suche verwendet. Da alle Einzelgewichte in ihrer Summe 1 ergeben sollen, ist das individuelle Gewicht einer Variablen r:

$$w_r = \frac{\beta_r}{\sum_{r=1}^{n} \beta_r} \tag{20}$$

Mit:
w_r = Gewicht der Variablen r;
β_r = Regressionskoeffizient der Variablen r;
n = Anzahl der berücksichtigten Variablen.

TOMPPO et al. (1999) schlagen eine Gewichtung der Variablen nach ihren Korrelationskoeffizienten in Bezug auf die Zielgröße vor. Bei diesem Vorgehen ist jedoch zu bedenken, dass auf diese Weise nur Gewichtungen für Variablen bestimmt werden können, die einen linearen Bezug zur gesuchten Zielgröße haben. Im Fall der Beziehung zwischen Einzelbaumvariablen wie z.B. dem BHD der Baumhöhe oder dem Alter und der Baum-Biomasse als Zielgröße muss davon ausgegangen werden, dass es sich um nicht-lineare Beziehungen handelt. In diesem Fall handelt es sich oft um allometrische Beziehungen, die jedoch wie in Kapitel 2.1.2 beschrieben, durch eine logarithmische Transformation linearisiert werden können.

2.5.3 Distance weighted *k*-Nearest Neighbour

Eine Modifikation des k-NN Algorithmus besteht darin, die k Trainingsbeispiele, die zur Klassifizierung herangezogen werden, nach ihrem berechneten Gesamtabstand so zu gewichten, dass diejenigen mit dem geringsten Abstand das höchste Gewicht erhalten. Da der Einfluss von Trainingsbeispielen mit zunehmendem Abstand sinkt, können bei dieser Vorgehensweise theoretisch alle Trainingsdaten in die Schätzung der Zielgröße einbezogen werden. Im Gegensatz zur ausschließlichen Verwendung der k nächsten Nachbarn ergibt sich in diesem Fall statt einer lokalen Approximation ein globaler Ansatz, der auch als Shepards Methode (SHEPARD, 1968) bezeichnet wird (MITCHELL, 1997). In welchem Umfang der Abstand der Trainingsbeispiele Einfluss auf die Schätzung hat wird durch eine Kernel-Funktion festgelegt, die bestimmt, in welcher

Proportion der Einfluss der Nachbarn mit zunehmendem Abstand sinkt. Die Gewichtung der gefundenen k nächsten Nachbarn bei der Schätzung der Zielgröße ist dabei klar von der Gewichtung der Variablen in der Abstandsberechnung zu unterscheiden.

Eine Gewichtung, die mit zunehmender Distanz proportional abnimmt, erhält man durch Verwendung des einfachen Kehrwertes der berechneten Distanz:

$$w_k = \frac{\frac{1}{d_{q,k}}}{\sum_{i=1}^{k} \frac{1}{d_{q,k}}} \tag{21}$$

Wobei:
w_k = Gewicht des Nachbarn k;
$d_{q,i}$ = Berechneter Abstand zwischen Abfrageinstanz x_q und Nachbar x_k.

SIRONEN et al. (2003) und in ähnlicher Weise MALTAMO und KANGAS (1998) schlagen eine Gewichtungsfunktion vor, in der ein Gewichtungsparameter bestimmt, wie stark der Einfluss eines Referenzbaumes mit zunehmendem Abstand abnimmt:

$$w'_k = \frac{\left(\frac{1}{d_{q,k}}\right)^t}{\sum_{i=1}^{k} \left(\frac{1}{d_{q,k}}\right)^t} \tag{22}$$

Wobei:
w_k = Gewicht des Nachbarn k;
$d_{q,i}$ = Berechneter Abstand zwischen Abfrageinstanz x_q und Nachbar x_k;
t = Gewichtungsparameter.

Zu beachten ist hierbei, dass bei einem Abstand $d_{q,i} = 0$ ein zum Abfragepunkt gleicher Nachbar das Gewicht 1 erhält und alle weiteren Trainingsinstanzen nicht mehr berücksichtigt werden. Insofern ist die k-NN Methode in diesem Fall als ein exaktes Interpolationsverfahren zu bezeichnen.

Abbildung 2-9 zeigt die Auswirkung des Gewichtungsparameters t auf die Form der Kernelfunktion.

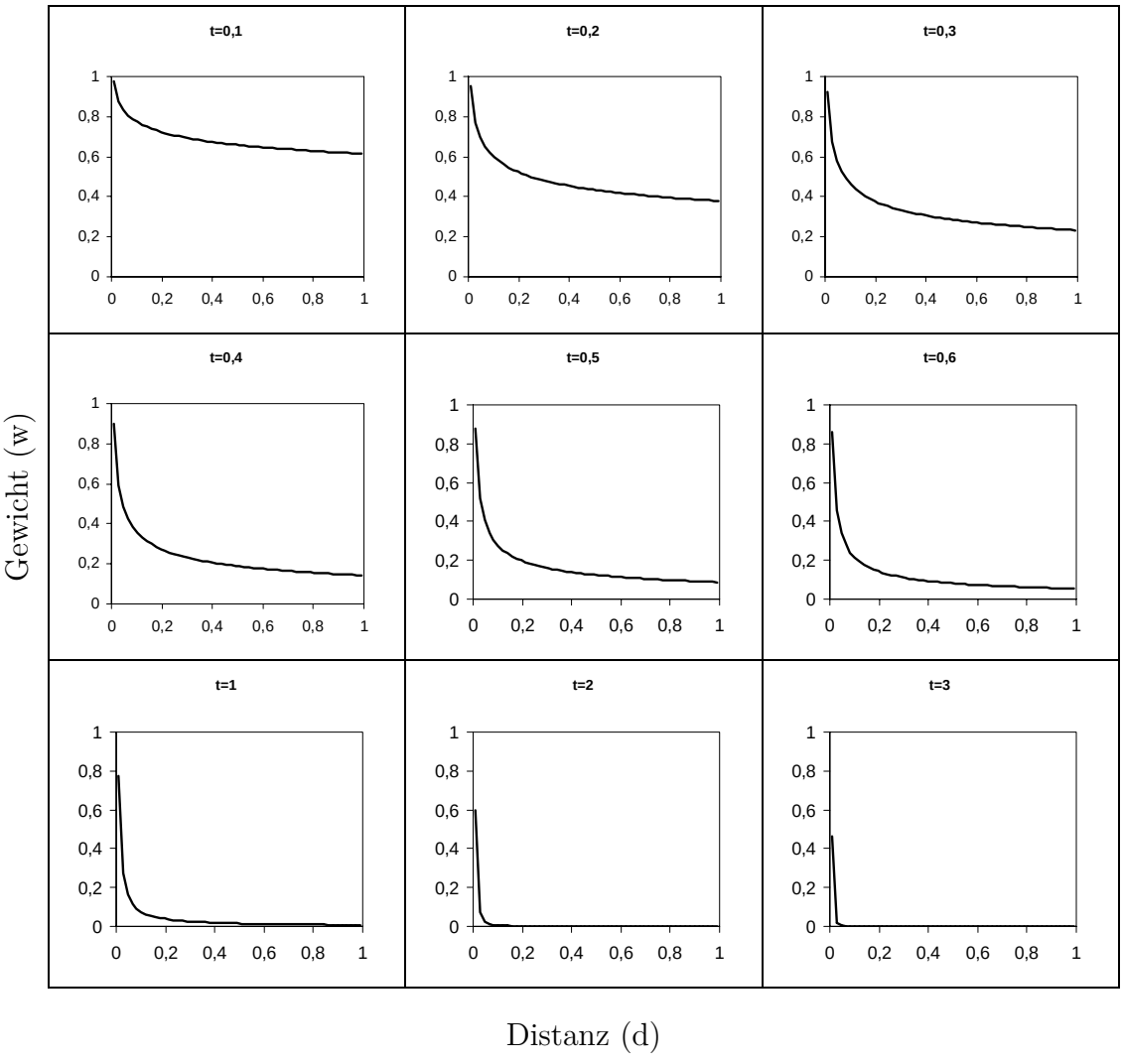


Abbildung 2-9. Einfluss des Gewichtungsparameters t auf die Form der Kernelfunktion.

Durch die Einbeziehung dieser Gewichtung in den Schätzer (vgl. (15)) für die Zielgröße ergibt sich somit:

$$\hat{f}\left(x_q\right) \leftarrow \frac{\sum_{i=1}^{k} w_k f\left(x_k\right)}{\sum_{i=1}^{k} w_k} \tag{23}$$

Im Fall des so berechneten gewichteten Mittelwertes entspricht dieser Schätzer dem *Nadaraya-Watson-Estimator* (NADARAYA, 1964; WATSON, 1964; ATKESON et al., 1996; HAENDEL, 2003).

Das in Abbildung 2-10 dargestellte Beispiel einer Kreuzvalidierung innerhalb eines Fichtendatensatzes zeigt die Verteilung der oberirdischen Gesamtbiomasse mit zunehmendem Abstand zu einer bestimmten Suchanfrage.

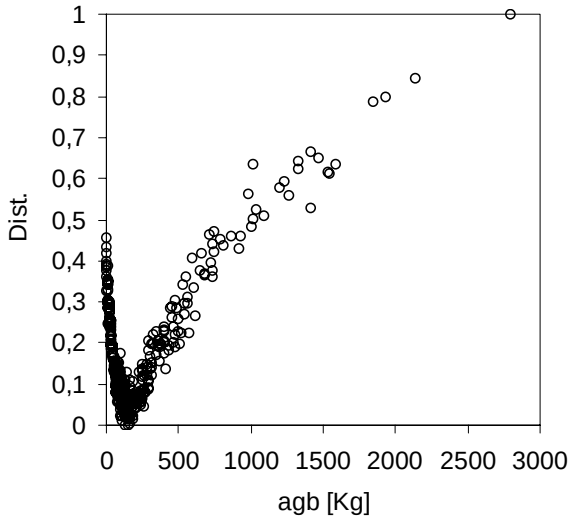


Abbildung 2-10. Normierter Gesamtabstand aller Trainingsinstanzen im Fall einer konkreten Suchanfrage in einem Fichtendatensatz, dargestellt über der oberirdischen Gesamtbiomasse (*agb*).

2.5.4 Fehlereinschätzung und Gütemaße

Eine Einschätzung der Prognosefehler ist im Rahmen der *k*-NN Methode für zwei Gesichtspunkte notwendig. Zum einen dienen Gütemaße als Kriterium für die Adjustierung der Parameter der Distanz- und Gewichtungsfunktion innerhalb der vorhandenen Trainingsdaten. Hierbei wird eine Minimierung des jeweiligen Gütemaßes durch eine Anpassung oder Veränderung der Parameter angestrebt. Zum anderen ist der Vergleich der Prognosen aus der *k*-NN Schätzung mit denen herkömmlicher Verfahren nur durch vergleichbare Einschätzungen der Prognosegüte möglich.

Da das Ergebnis der *k*-NN Methode immer eine Punktschätzung und nicht wie bei einem parametrischen Verfahren eine über den gesamten Wertebereich definierte Funktion ist, ist die Einschätzung der erwarteten Güte einer Schätzung erschwert. Im Gegensatz zu einem Regressionsmodell, für das ein Vertrauensintervall über die gesamte Spannbreite der Zielgröße geschätzt werden kann, ist die Sicherheit der Schätzung bei der *k*-NN Methode in nahezu jedem Punkt des Wertebereiches unterschiedlich. Zur Bestimmung des erwarteten Fehlers einer Schätzung wird bei nichtparametrischen Verfahren im Allgemeinen die Standardabweichung der Schätzung (RMSE = root mean square error) zwischen geschätzten und beobachteten Werten der Zielgröße innerhalb der Trainingsdaten herangezogen.

$$RMSE = \sqrt{\frac{\sum_{n=1}^{N}\left(\widehat{x_i} - x_i\right)^2}{N}} = \sqrt{MSE} \tag{24}$$

Wobei:

N = Anzahl der Beobachtungen;
x_i = beobachteter Wert für die Instanz i;
$\widehat{x_i}$ = geschätzter Wert für die Instanz i.

Die Abfrageinstanz ist in diesem Fall eine Trainingsinstanz, deren Wert für die Zielgröße bekannt ist. Mit Hilfe einer Kreuzvalidierung wird auf diese Weise für jede

Instanz in der Datenbank eine Schätzung für die Zielgröße auf der Grundlage der verbleibenden *N*-1 Instanzen abgeleitet und den beobachteten Werten gegenübergestellt. Üblicherweise wird der RMSE% als relative Abweichung vom Mittelwert der geschätzten Zielgröße verwendet (siehe z.B. MALINEN et al., 2003; SIRONEN et al., 2003).

$$RMSE\% = \frac{RMSE}{\bar{\hat{x}}_i} \cdot 100 \tag{25}$$

Der RMSE wird hauptsächlich aufgrund der quadratischen Fehlergewichtung, der einfachen Interpretation und der hohen Bekanntheit dieses Gütemaßes verwendet. Er liefert die Information, inwieweit die geschätzten Werte im Durchschnitt von den Beobachtungen abweichen, ohne dass sich negative und positive Abweichungen ausgleichen. Größere Abweichungen werden hierbei stärker gewichtet als kleine (WEBER, 1998).

Ein Nachteil bei der Verwendung des RMSE ist darin zu sehen, dass das Ausgangsniveau der Daten unberücksichtigt bleibt. D.h., Abweichungen werden nur aufgrund ihrer absoluten Größe, jedoch nicht in Bezug auf die Höhe der Zielgröße bewertet. Besonders im vorliegenden Fall, bei dem Schätzungen auf Einzelbaumebene abgeleitet werden, würde das dazu führen, dass ein bestimmter Fehlerbetrag bei kleinen oder besonders großen Bäumen gleich gewichtet werden würde. Ein Gütemaß das die Prognosefehler in Abhängigkeit des Ausgangsniveaus beschreibt, ist der mittlere absolute prozentuale Fehler (Mean Absolute Percentage Error = MAPE) bzw. die Wurzel des mittleren quadratischen prozentualen Fehlers (Root Mean Square Percentage Error = RMSP).

$$MAPE = \frac{1}{N} \sum_{n=1}^{N} \left| \frac{\widehat{x_i} - x_i}{x_i} \right| \cdot 100\% \tag{26}$$

$$RMSP = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\frac{\left(\widehat{x_i} - x_i\right)^2}{x_i^2}} \cdot 100\% \tag{27}$$

Der RMSE sowie auch der mittlere quadratische Fehler (MSE) lassen sich in einen systematischen und unsystematischen Fehleranteil zerlegen, wodurch eine weitere Untersuchung der Fehlerstruktur ermöglicht wird.

Der systematische Fehleranteil (Bias-Anteil) hat hierbei eine hohe Sensitivität gegenüber Niveau-Fehlerprognosen. Weicht der Bias von Null ab, so weist dies auf eine systematische Über- bzw. Unterschätzung der beobachteten Werte durch die Prognosen hin.

$$\textit{Bias-Anteil des MSE: } Bias = \frac{\left(\overline{x_i} - \overline{\widehat{x_i}}\right)^2}{MSE}\textit{; für MSE } (RMSE^2) \neq 0 \tag{28}$$

Weicht der Varianz-Anteil des RMSE von Null ab, so wird zwar die Streuung der Beobachtungswerte um ihren Mittelwert durch die Prognosen erfasst, das Ausmaß dieser Erfassung ist jedoch zu stark oder zu schwach (systematische Abweichungs-Fehlprognosen).

$$\textit{Varianz-Anteil des MSE: } Var = \frac{\left(S_{x_i} - S_{\widehat{x_i}}\right)^2}{MSE}\textit{; für MSE } (RMSE^2) \neq 0 \tag{29}$$

wobei S die Streuung der beobachteten bzw. geschätzten Werte ist.

Prognosen, die weder systematische Niveaufehler noch Abweichungsfehler aufweisen, können sich von den beobachteten Werten nur noch unsystematisch unterscheiden.

Solche Abweichungsfehler können durch den Kovarianz-Anteil des Fehlers aufgedeckt werden.

$$\textit{Kovarianz-Anteil: } Kov = \frac{2 \cdot (1 - R) \cdot S_{x_i} \cdot S_{\hat{x}_i}}{MSE} \textit{; für MSE (RMSE}^2\textit{)} \neq 0 \qquad (30)$$

Bias-, Varianz- und Kovarianz-Anteil sind in diesem Fall auf ein Intervall zwischen 0 und 1 beschränkt. Je kleiner der Bias- bzw. Varianz-Anteil und je weiter der Kovarianz-Anteil bei Eins liegt, desto besser ist die Prognose. WEBER (1998) schlägt daher zur Aufdeckung verschiedener Fehler einen Gütemaßmix vor.

2.5.5 Optimale Größe der Nachbarschaft (Bandbreite)

Nicht-parametrische instanzenbasierte Schätzverfahren sind typischerweise mit einem sog. Bias-Varianz-Dilemma behaftet. Ein methodischer Nachteil solcher Verfahren besteht darin, dass die Zielgröße nicht durch eine Extrapolation über die Trainingsdaten heraus hergeleitet werden kann. Befinden sich die Werte einzelner Variablen einer unbekannten Instanz am Rand der Werteverteilung der Trainingsdaten oder liegen sie sogar außerhalb, kommt es zu einer systematischen Über- bzw. Unterschätzung.

Die Bestimmung der optimalen Anzahl k nächster Nachbarn, die bei einer Schätzung der Zielgröße berücksichtigt werden, wird normalerweise mit Hilfe des, auf Grundlage verschieden großer Nachbarschaften berechneten, Fehlers einer Kreuzvalidierung innerhalb der Trainingsdaten ermittelt und danach für alle Schätzungen festgelegt. In vielen Fällen ist mit ansteigendem k zunächst ein Abfallen, bei weiterer Erhöhung von k aber ein Ansteigen des RMSE zu beobachten. In vielen Veröffentlichungen ist daher ein typisches lokales Minimum im Verlauf des RMSE über der steigenden Anzahl von Nachbarn zu finden (siehe z.B. MCROBERTS et al., 2002; MALINEN et al., 2003; MALINEN, 2003; MALINEN und MALTAMO, 2003; STÜMER und KÖHL, 2005; LEMM et al., 2005).

Das Ansteigen des Fehlers bei steigender Anzahl von Nachbarn ist hierbei hauptsächlich eine Folge des ansteigenden Bias und somit des beschriebenen Randeffektes (siehe auch LAWRENCE et al., 1996 oder LOADER, 1999, S.7). Besonders die großen absoluten Fehlerwerte die mit steigender Anzahl berücksichtigter Nachbarn im Bereich der Extreme des Wertebereiches der wichtigsten Variablen entstehen, verursachen ein Ansteigen des RMSE % im Falle einer Kreuzvalidierung.

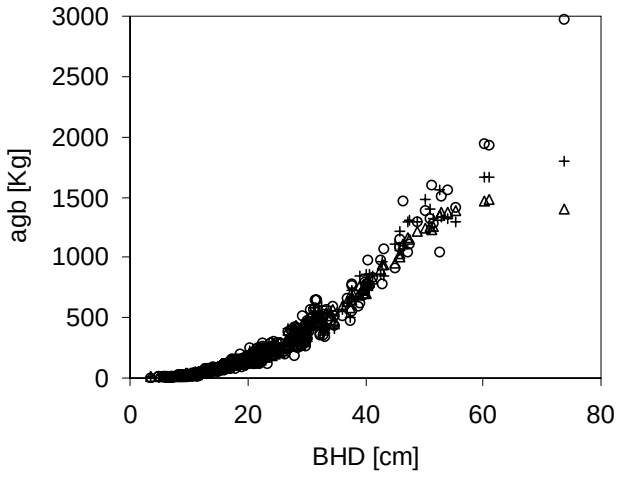


Abbildung 2-11. Aus einer Kreuzvalidierung innerhalb eines Fichtendatensatzes abgeleitete Biomasseschätzungen über $k = 3$ (Plus) bzw. 15 (Dreieck) Nachbarn und beobachtete Werte über dem BHD (Kreis).

Wie aus dem in Abbildung 2-11 dargestellten Beispiel deutlich wird, werden zur Schätzung der größten Bäume (hier bezogen auf eine Schätzung mit Hilfe des BHD und der Baumhöhe) überdurchschnittlich mehr kleinere Nachbarn herangezogen, was für die größten Individuen zu einer Unterschätzung bzw. für die kleinsten Individuen zu einer Überschätzung ihrer Biomasse führt.

Das eigentliche Problem bei der Determinierung der Größe der Nachbarschaft ist darin zu sehen, dass ein bestimmter Wert für k gesucht wird, der dann für jede Punktschätzung verwendet werden soll. Hierdurch steigt im speziellen Fall von Vergleichen zwischen Baumindividuen mit steigender Anzahl von Nachbarn besonders der Fehler für die Schätzung der größten Individuen. Während hier eine geringe Anzahl von Nachbarn den Fehler minimieren würde, kann bei einer ausgeglichenen Verteilung der

Nachbarn auf eine höhere Anzahl zurückgegriffen werden. Die Forderung die berücksichtigten Nachbarn nur aus einer „symmetrischen“ Nachbarschaft zu wählen, in der sich $k/2$ größere Nachbarn sowie kleinere Nachbarn befinden, findet sich z.B. auch bei KOTZ et al. (1998, S.472).

Die Frage nach einer optimalen Größe der Nachbarschaft, die auch als Bandbreite bezeichnet wird, ist daher abhängig von der Lage der Trainingsdaten im n-dimensionalen Merkmalraum. Die Entscheidung über eine feste Größe der Nachbarschaft ist hierbei ein typisches Bias–Varianz-Dilemma, das in vielen Fällen nicht zufrieden stellend gelöst werden kann. Durch die Erhöhung der Anzahl der berücksichtigten Nachbarn steigt zwar die Reliabilität der lokalen Approximation, gleichzeitig steigt jedoch der Bias der Schätzung (ALTMAN, 1990; MCROBERTS, 2002; KATILA, 2004; FINLEY et al., 2006). Eine Verringerung von k führt andererseits zu einer erhöhten Varianz der Schätzung, da hierdurch der ausgleichende Effekt der Mittelwertsbildung innerhalb der Nachbarschaft kleiner wird. Dementsprechend besteht hierbei die Gefahr des Overfittings.

Grundlegend lassen sich verschiedene Ansätze zur Eingrenzung der Nachbarschaft unterscheiden. Um die verschiedenen Herangehensweisen zu verdeutlichen, kann ein Glättungs- oder Bandbreitenparameter (h) verwendet werden, der über die Spannweite der Trainingsdaten aus denen eine Schätzung abgeleitet wird, bestimmt. Die Wahl dieses Parameters kann auf verschiedene Weise erfolgen (ATKESON et al., 1996; MALINEN, 2003):

- Feste Bandbreite: Der Parameter h ist hierbei ein konstanter Wert (Kernel Methode). Die nächsten Nachbarn werden nur bis zu einer festgelegten Distanz berücksichtigt. Die Größe der Nachbarschaft (k) ist in diesem Fall von der Anzahl der Trainingsbeispiele in der Nähe des Abfragepunktes abhängig.
- Nearest Neighbour Bandbreite: Der Parameter h wird als Distanz zum k-ten Nachbarn definiert (k-NN Methode) und ist damit von der Verteilung der Trainingsdaten um den Abfragepunkt abhängig. Die Größe der Nachbarschaft ist hierbei durch k festgelegt.

Beide Varianten können dabei entweder global für alle Schätzungen definiert, oder lokal (bzw. adaptiv) für jeden Abfragepunkt bestimmt werden. Eine Möglichkeit zur Optimierung von k ist dabei die Verwendung von iterativen Optimierungsalgorithmen die z.B. durch die globale oder lokale Veränderung von k den RMSE einer (Leave-One-Out-) Kreuzvalidierung minimieren.

Die Wahl einer festen Bandbreite kann dabei zu einer Erhöhung der Varianz in Bereichen mit einer geringen Anzahl von Trainingsdaten führen. Im Extremfall finden sich überhaupt keine Nachbarn innerhalb der vorgegebenen Distanz und eine Schätzung ist nicht möglich (CLEVELAND und LOADER, 1994; ATKESON et al., 1997). Dieses Problem kann allerdings durch eine Normierung der berechneten Distanzen auf ein festes Intervall (z.B. [0,1]) umgangen werden. Bestimmt man eine maximale (normierte) Distanz bis zu der Nachbarn berücksichtigt werden sollen, kann es auf diese Weise nicht dazu kommen, dass sich keine Trainingsdaten in diesem Bereich befinden (siehe z.B. Abbildung 2-10). In der Literatur finden sich zahlreiche Ansätze zur lokal adaptiven Wahl der Bandbreite (siehe z.B. CLEVELAND und LOADER, 1994; WETTSCHERECK und DIETTERICH, 1994; ATKESON et al., 1997; MCROBERTS et al., 2002 oder MALINEN, 2003).

Neben der Größe der Nachbarschaft hat jedoch auch die Integration einer Kernelfunktion, die zu einer Gewichtung innerhalb der gefundenen Nachbarn verwendet wird, einen entscheidenden Einfluss auf das beschriebene Bias-Varianz-Dilemma (siehe 2.5.3). Durch die Wahl des Gewichtungsparameters t kann der Einfluss von Nachbarn bei einer festgelegten Bandbreite entsprechend ihrer Distanz abgeschwächt werden. Hierbei entstehen daher Wechselwirkungen, die sich im speziellen Fällen gegenseitig aufheben können. Liegt zum Beispiel keine Beschränkung der Bandbreite vor, so dass alle vorhandenen Nachbarn (k=N) einbezogen werden, führt die Wahl eines hohen Gewichtungsparameters dazu, dass trotzdem nur eine gewisse Anzahl nächster Nachbarn einen entscheidenden Einfluss auf die Klassifizierung haben. Im Extremfall $t \rightarrow \infty$ wird hier nur der nächste Nachbar für eine Schätzung herangezogen.

Im Rahmen dieser Untersuchung soll überprüft werden, ob sich aus der Verteilung der berechneten Distanzen zwischen einem Abfragepunkt und allen Trainingsinstanzen ein Anhaltspunkt für eine individuelle Bestimmung von k für jede einzelne Abfrage ableiten lässt. Ansätze in diese Richtung finden sich z.B. mit der sog. *Locally Adaptable*

Neighbourhood (LAN) MSN Methode bei MALINEN (2003) oder in MCROBERTS et al. (2002).

Ziel dieser Untersuchung soll es sein, k variabel zu halten und je nach Lage des Abfragepunktes im Verhältnis zur Lage aller Trainingsdaten im n-dimensionalen Merkmalraum anzupassen.

2.6 Umsetzung der *k*-NN Methode

Im Gegensatz zu einem Regressionsmodell, das auf einfache Weise auf unabhängige Variablen eines Datensatzes angewendet werden kann, bedarf die Umsetzung instanzenbasierter Verfahren einer Softwareanwendung bzw. der Implementierung des Algorithmus. Im Rahmen der zugrunde liegenden Studie wurde in Kooperation mit der Firma Argus Forstplanung (Staupendahl, 2006) das Softwaremodul kNN-Biomass entwickelt, das in weiten Teilen eine Umsetzung der oben beschriebenen theoretischen Grundlagen darstellt.

Da die technische Umsetzung des k-NN Algorithmus zwar nicht Inhalt dieser Arbeit ist, andererseits aber die zugrunde gelegte Anwendung teilweise zur Generierung der im folgenden dargestellten Ergebnisse verwendet wurde, soll hier nur kurz auf dieses Softwaremodul eingegangen werden. Abbildung 2-12 zeigt einen Ausschnitt der grafischen Benutzeroberfläche.

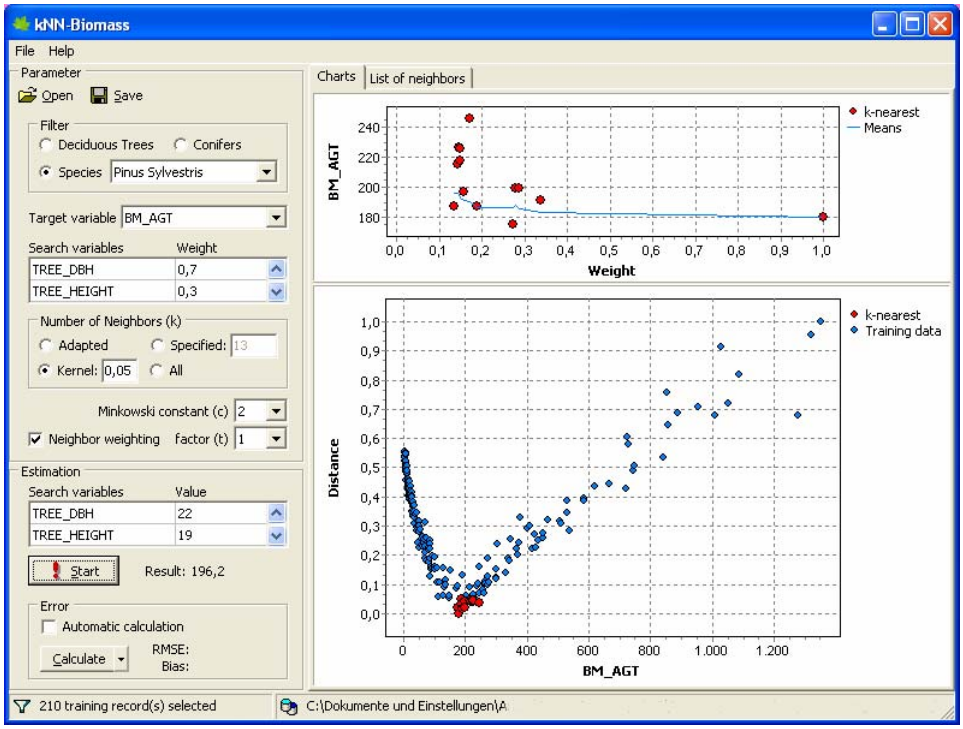


Abbildung 2-12. Benutzeroberfläche des Softwaremoduls kNN-Biomass (STAUPENDAHL, 2006).

Nach Anbindung einer geeigneten Trainingsdatenbank (siehe 2.7) im beschriebenen Format, können die Baumart bzw. Baumartengruppe sowie die zu schätzende Zielgröße ausgewählt werden. Weiterhin können hier die zur Distanzberechnung berücksichtigten Variablen übernommen werden, wobei diese entweder aus den in der Datenbank vorhandenen Einzelbaumvariablen oder vorhandenen Meta-Daten der Untersuchungsbestände bzw. aus zusätzlichen baumartenspezifischen Variablen, wie z.B. einer mittleren Holzdichte der betreffenden Art, gewählt werden können.

Neben den Gewichtungsfaktoren der einzelnen Dimensionen können hier alle weiteren notwendigen Parametereinstellungen zur Adjustierung der implementierten Distanzmetrik sowie die Distanzgewichtung vorgenommen werden. Um eine angepasste Wahl der Bandbreite (Größe der berücksichtigten Nachbarschaft) zu ermöglichen, wurde ein Zusatzmodul zur Berechnung ausgewählter Fehlermaße aus einer Kreuzvalidierung des selektierten Trainingsdatensatzes mit wechselnder Anzahl von Nachbarn implementiert (siehe Abbildung 2-13).

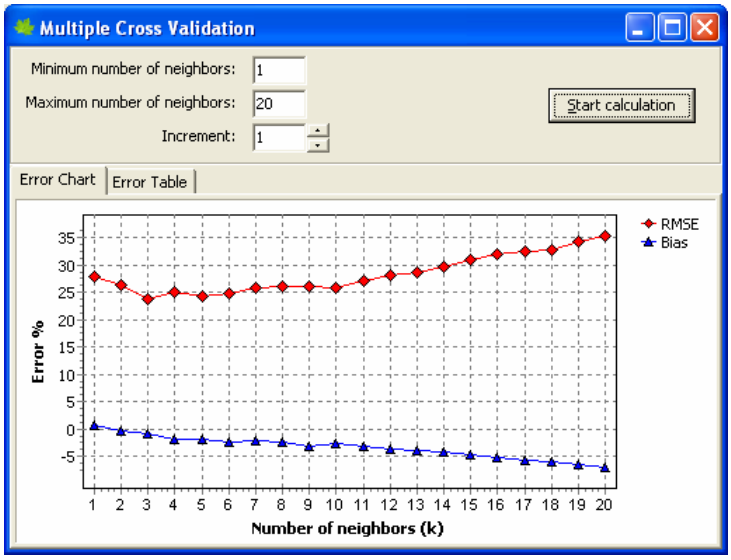


Abbildung 2-13. Modul zur Berechnung von Fehlermaßen aus Kreuzvalidierungen des Trainingsdatensatzes mit unterschiedlicher Größe der berücksichtigten Nachbarschaft.

Hierdurch ist eine Approximation angepasster Gewichtungsfaktoren sowie aller weiteren Einstellungen durch eine iterative Minimierung der Fehlermaße möglich.

Alle Parametereinstellungen können anschließend in baumarten- bzw. baumartengruppen -spezifischen Parameterdateien abgelegt werden, so dass eine Neukalibrierung der Parameter nur im Fall bemerkenswerter Änderungen der Datengrundlage notwendig sind. Um eine Schätzung der unbekannten Zielgröße neuer Instanzen durchzuführen, können entweder die verwendeten Variablen manuell eingegeben, oder eine komplette Baumliste mit den nötigen Werten eingelesen werden. Neben den für eine Schätzung benötigten Einstellungen ermöglicht eine Datenbankschnittstelle den Zugriff auf die zugrunde gelegte Datenbank. Hierbei kann sowohl auf die vorhandenen Einzelbaumdaten sowie die dazugehörigen Meta Informationen zugegriffen werden. Des Weiteren können die dem jeweiligen Datensatz zugehörigen Publikationen, soweit digital vorhanden, eingesehen werden.

2.7 Datenbankstruktur

Wie bereits erläutert ist die Effektivität der k-NN Methode in hohem Maße von der zugrunde gelegten Datenbank abhängig. Neben der Anzahl der Trainingsbeispiele ist besonders der Umfang an relevanten Informationen für jede Instanz ausschlaggebend für die Qualität der Schätzung.

Ziel der Entwicklung einer Datenbankstruktur zur Ablage von Trainingsdaten ist es Einzelbaumdaten mit möglichst vielen Informationen logisch zu verknüpfen um nach verschiedenen Fragestellungen jeweils die passenden Bäume finden zu können. Hierfür wurden verschiedene Tabellen zur Speicherung von Einzelbaum-, Bestandes- und weiterer Meta- Informationen angelegt. Die hier verwendete Datenbank wurde in MS Access umgesetzt. Tabelle 2-3 gibt einen Überblick über die einzelnen Datenbanktabellen.

Tabelle 2-3. Übersicht der einzelnen Datenbanktabellen.

Datenbanktabelle	Inhalt
COUNTRIES	Länderliste mit Abkürzungen
PROJECTS	Aufführung der Literaturquellen und Datenherkunft
SITES	Beschreibung der Bestände aus denen die Daten stammen
SOILTEXTURE	Klassifizierung der Bodenart
SOILTYPES	Liste von 181 Bodentypen (FAO Klassifizierung)
SPECIES	Auflistungen der Spezies zu denen Daten vorliegen
TREES	Tabelle zur Speicherung der Einzelbauminformationen

In der Tabelle TREES können die direkt am Baum gemessenen Variablen wie Höhe, BHD und die Biomasse der einzelnen Baumkompartimente eingetragen werden. Diese Tabelle ist einmal mit der Tabelle SPECIES über eine n:1 Beziehung verknüpft um jeden Baum einer Art zuzuordnen. Weiterhin besteht eine Verknüpfung mit der Tabelle SITES (n:1) um jedem Baum einem Bestand zuzuordnen.

Tabelle 2-4. Struktur der Tabelle TREES zur Ablage von Einzelbauminformationen.

Feldbezeichnung	Inhalt	Verknüpfungen
SITE_NO	Eindeutige Bestandesnummer	n:1 mit Tab. SITES
TREE_NO	Nummer des Baumes	
TREE_SPEC	Botanischer Name der Art	n:1 mit Tab. SPECIES
TREE_AGE	Einzelbaumalter	
TREE_DBH	BHD [cm]	
TREE_HEIGHT	Baumhöhe [m]	
TREE_CROWNLNG	Kronenlänge [m]	
TREE_BARKTHIK	Rindenstärke [mm]	
BM_STEM	Stamm-Biomasse [Kg]	
BM_BRANCH	Ast-Biomasse [Kg]	
BM_DRYBRANCH	Totholz-Biomasse [Kg]	
BM_LEAVES	Blatt-Biomasse [Kg]	
BM_ROOT	Wurzel-Biomasse [Kg]	
BM_BARK	Rinden-Biomasse [Kg]	
BM_THICKWOOD	Derbholz-Biomasse [Kg]	
TREE_COMMENT	Kommentar	

Die Tabelle SITES ist mit vier weiteren Tabellen (COUNTRIES, PROJECTS, SOILTEXTURE, SOILTYPES) jeweils durch eine n:1 Beziehung verknüpft. So wird für jeden Bestand das Land in dem er sich befindet, das Projekt in welchem die Daten aufgenommen wurden, sowie die dortige Bodenart und der Bodentyp angegeben.

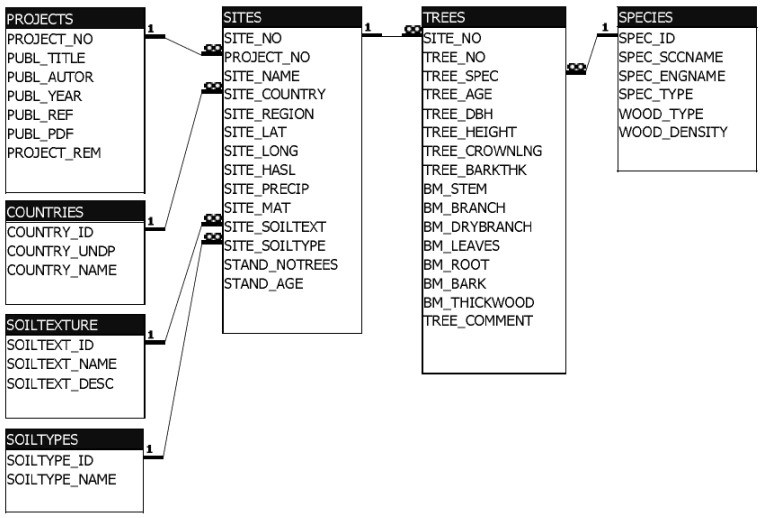


Abbildung 2-14. Datenbankstruktur.

2.8 Datengrundlagen

Die im Rahmen dieser Studie verwendeten Datengrundlagen entstammen zum Teil der verfügbaren Literatur. Einige Teildatensätze wurden bisher nicht veröffentlicht, werden hier jedoch nicht separat ausgewertet, sondern zu einer umfassenden Datenbasis zusammengefasst. Die Sammlung von Biomassedaten aus destruktiven empirischen Studien ist dadurch beschränkt, dass sich die Originaldaten einzelner Studien selten in veröffentlichten Publikationen finden lassen. Hierzu muss in den meisten Fällen auf die graue Literatur oder unveröffentlichte Projektberichte zurückgegriffen werden. Für eine Vielzahl von Studien lassen sich zwar die abgeleiteten Regressionsmodelle finden, die zugrunde gelegten Ausgangsdaten sind jedoch meist nicht mehr vorhanden. Oftmals ist daher die Kontaktaufnahme zu den Autoren einzelner Studien notwendig, wobei dies vor allem bei älteren Untersuchungen nur selten aussichtsreich ist (Wirth et al., 2004).

Hinzu kommt, dass aufgrund des enormen Kosten- und Zeitaufwands der für empirische Biomassestudien typisch ist, die gewonnenen Daten nur selten bereitwillig zur Verfügung gestellt werden. Die hier verwendeten Datengrundlagen bestehen daher teilweise aus Datensätzen, die lediglich im Rahmen dieser Studie verwendet werden dürfen. Unter diesem Gesichtspunkt sei z.B. die Bereitstellung umfangreicher Datensätze von Fichten durch Dr. C. Wirth (siehe Wirth et al., 2004) sowie Daten von Fichten und Kiefern aus der finnischen nationalen Waldinventur durch das Finnish Forest Research Institute (METLA) besonders hervorzuheben. Diese Daten wurden im Zeitraum von 1988 bis 1990 im Rahmen des national tree research project (VAPU) auf Stichprobenpunkten der Nationalen Finnischen Waldinventur (NFI 8) erhoben (Korhonen und Maltamo, 1990).

Weitere Datensätze von Kiefer, Eiche, Lärche und Buche konnten durch die Forschungsanstalt für Waldökologie und Forstwirtschaft in Kooperation mit dem Lehrstuhl für Waldwachstumskunde der TU München zur Verfügung gestellt werden (Grote et al., 2003). Ein großer Buchendatensatz wurde durch Joosten und Schulte (2003) beigesteuert. Weiterhin wurden umfagreiche Biomassedaten aus Nord-Amerika von Ter-Mikaelian (Ter-Mikaelian und Korzukhin, 1997) bereitgestellt, auf die jedoch nur in einer Teilauswertung eingegangen werden soll. Einen Überblick der verwendeten Datensätze der am häufigsten vertretenen Baumarten zeigt Tabelle 2-5.

Tabelle 2-5. Anzahl der Probebäume und Untersuchungsbestände einzelner Studien und Datenquellen der hier verwendeten Baumarten.

Baumart	Land	n	Bestände	Quelle
Fichte	Deutschland	38	7	Mund et al., 2002
	Deutschland	18		Heller und Göttsche, 1986
	Tschechische Republik	40	3	Vyskot, 1981
	Deutschland	5		Droste zu Hülshoff, 1969
	Deutschland	44	3	Sharma, 1992
	Belgien	6		Duvigneaud et al., 1970
	Deutschland	22	2	Poeppel, 1989
	Tschechische Republik	17	3	Cerny, 1990
	Deutschland	36	6	Rademacher, 2004
	Deutschland	28		Akça und Mench, 1993
	Schweden	32		Johansson, 1999
	Österreich	24		Neumann und Jandl, 2005
	Deutschland	3		Hesse, 1990
	Deutschland	20	2	Dietrich, 1968
	Deutschland	9	5	Raisch, 1983
	Deutschland	17	2	Poeppel, 1989
	Finnland	203		VAPU
	Deutschland	6		Grote, 2002
	Deutschland	19		LWF, 2002 (Dietrich)
Buche	Deutschland	59	6	FAWF
	Deutschland	116	4	Joosten und Schulte, 2003
	Deutschland	19		Pellinen, 1986
	Tschechische Republik	20	3	Cienciala et al., 2004
	Spanien	7		Santa Regina und Tarazona, 2001
Lärche	Deutschland	10		FAWF
Kiefer	Belgien	8		Xiao und Ceulemans, 2004
	Belgien	9		Xiao et al., 2003
	Deutschland	30	2	LWF
	Finnland	18		Vanninen et al., 1996
	Finnland	205		VAPU
Eiche	Deutschland	13		FAWF
	Deutschland	15		Grote et al., 2003

Da die Anzahl und Aufteilung in einzelne Kompartimente bei verschiedenen Untersuchungen oft nicht einheitlich ist bzw. die Biomasse in unterschiedlichem Grad differenziert vorliegt, sind die Datengrundlagen zunächst auf einen gleichen Satz von Attributen vereinheitlicht worden. Hierbei werden höhere Differenzierungsgrade, wie z.B. getrennt aufgenommene Nadeljahrgänge bei Fichten oder eine Unterteilung der Astbiomasse in einzelne Astdurchmesserklassen, zu der jeweils kleinstmöglichen Unterteilung auf Ebene aller Datensätze zusammengefasst. Die Einteilung der Biomasse erfolgt daher in den Kompartimenten Wurzeln, Stamm, Äste, Blätter oder Nadeln. Des Weiteren wird eine eventuell getrennt aufgenommene Rindenbiomasse bzw. die Aufteilung in Äste und Totäste erhalten. Da für viele Buchendatensätze eine Unterteilung in Derbholzbiomasse und weitere Kompartimente vorliegt, wird speziell für Laubbaumarten dieses zusätzliche Kompartiment eingeführt.

Im Rahmen der Auswertung werden hier generell nur Einzelbaumdatensätze verwendet, in denen sich die kompartimentweise geschätzten Biomassewerte mindestens zu einer oberirdischen holzigen Gesamtbiomasse (above ground woody biomass = *agwb*) aggregieren lassen. D.h. der Vektor der bekannten Designattribute (siehe 2.3.1) ist so auf eine Mindestlänge festgelegt, dass alle oberirdischen holzigen Kompartimente (also Stammbiomasse, Ast- und Totast-Biomasse sowie Rindenbiomasse) enthalten sind. Eine je nach Baumart geringeren Umfang haben die Datensätze, in denen sich die einzelnen Kompartimente zur oberirdischen Gesamtbiomasse (aboveground biomass = *agb*) zusammenfassen lassen. Hierfür sind neben den holzigen Kompartimenten auch Informationen über die Blatt- bzw. Nadelbiomasse nötig, die speziell für die verwendeten Buchen dieses Datensatzes nur in einigen Fällen vorhanden sind. Aufgrund der besonderen Schwierigkeiten, die mit der Beprobung von Wurzelsystemen größerer Bäume verbunden sind, ist der Anteil der Datensätze für die eine Gesamtbiomasse (total biomass = *tbm*) aggregiert werden kann am geringsten (FEHRMANN et al., 2003). Tabelle 2-6 gibt einen Überblick der zum jetzigen Stand der Datenbank vorhandenen Baumanzahlen in den einzelnen Aggregationsstufen.

Tabelle 2-6. Überblick der Baumanzahlen in den einzelnen Aggregationsstufen *agwb* (aboveground woody biomass), *agb* (aboveground biomass) und *tbm* (Gesamtbiomasse).

Baumart	Anzahl in Aggregationsstufe		
	agwb	*agb*	*tbm*
Abies balsamea	30	30	
Acer rubrum	53	53	
Acer saccharum	55	55	
Betula alleghaniensis	49	49	
Betula papyrifera	37	37	
Fagus grandifolia	14	14	
Fagus sylvatica	221	26	8
Larix decidua	10		
Picea abies	578	578	83
Picea glauca	24	24	
Picea mariana	24	24	
Picea rubens	37	37	
Pinus sylvestris	270	270	26
Populus grandidentata	30	30	
Populus Tremuloides	46	46	
Quercus petrea	28		
Summe	1499	1266	117

Ausführliche Literaturstudien zeigen, dass sich für einzelne Kompartimente weitaus größere Datensätze zusammentragen lassen. Alleine die sehr umfangreichen Untersuchungen von Burger (1925 – 1953) enthalten mehr als 600 Datensätze von Kiefern, Douglasien, Tannen, Eichen, Buchen und Fichten, die jedoch zumeist nur Nadel- oder Blattbiomasse und Astbiomassen enthalten. Des Weiteren finden sich zahlreiche ökologische Studien in denen ebenso lediglich Nadel- und Blattkompartimente untersucht wurden.

Da zu einer aussagekräftigen Evaluierung der k-NN Methode möglichst große Datensätze verwendet werden sollen, wird im Rahmen der Auswertung zumeist auf Fichten, Buchen oder Kiefern Bezug genommen. Abbildung 2-15 zeigt die Durchmesserverteilungen für diese größten Teildatensätze.

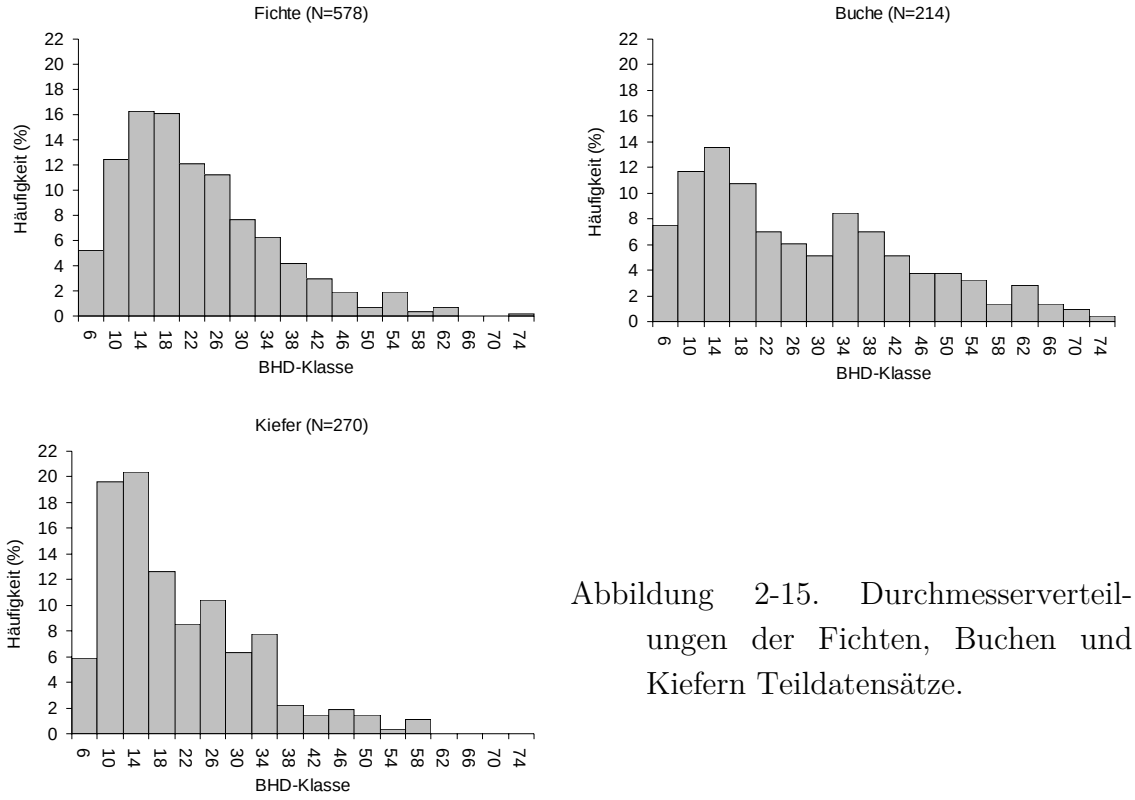


Abbildung 2-15. Durchmesserverteilungen der Fichten, Buchen und Kiefern Teildatensätze.

Einen Überblick der Durchmesserbereiche aller Baumarten ist in Abbildung 9-2 (Anhang II, Seite 138) dargestellt.

2.8.1 Zur Verwendung empirischer Biomassedaten

Da Biomassedaten aus destruktiven Untersuchungen typischerweise Schätzungen sind, die durch Hochrechnung aus Stichproben von verschiedenen Kompartimenten der betreffenden Bäume gewonnen werden, können Unterschiede in der Methodik zwischen verschiedenen Untersuchungen bei der Zusammenstellung einzelner Datensätze zu einem unbekannten Fehler in der Regressionsanalyse eines kombinierten Datensatzes führen. Zur Hochrechnung der Stichproben, die aus einzelnen Kompartimenten eines Baumes entnommen werden, haben sich gemischt lineare Modelle als besonders geeignet erwiesen (MCCULLOCH und SEARLE, 2000; MUUKKONEN and LEHTONEN, 2004; LEHTONEN, 2005b).

Daher ist zu bedenken, dass die in der Literatur als „beobachtete“ Biomasse dargestellten Einzelbaumwerte durch ihre Herleitung bereits mit einem Fehler unbekannter Größe behaftet sind. Der Standardfehler der Schätzung auf Ebene der einzelnen Bäume, die in einer Untersuchung herangezogen wurden, wird in der gegebenen Literatur in keinem der vorliegenden Fälle aufgeführt. Aus diesem Grund wird die Einzelbaumbiomasse wie sie in der jeweiligen Untersuchung angegeben ist im folgenden Ergebnisteil als beobachtete („wahre“) Biomasse interpretiert. Bei dieser Vorgehenswiese muss jedoch beachtet werden, dass die Residuen des jeweiligen Regressionsmodells bzw. der sich ergebende Standardfehler der Schätzung implizit auch diesen „Einzelbaumfehler“ enthalten. Bemerkenswert in Bezug auf die im Kyoto Protokoll verbindlich vorgeschriebene Fehlerberechnung für Kohlenstoffbilanzen ist, dass dieser Fehler der vorhandenen Datengrundlagen bisher weitgehend unberücksichtigt bleibt.

3 Ergebnisse

Die im Folgenden dargestellten Ergebnisse beziehen sich auf die formulierte Zielsetzung und die sich daraus ergebenen Fragestellungen. Hierbei soll zunächst auf die im Methodenteil beschriebenen Unterschiede zwischen Prozessmodellen und empirischer Datenanalyse sowie die Auswirkung einer Veränderung der Datengrundlage in Hinblick auf den als Eingangsgröße für allometrische Modelle verwendeten BHD eingegangen werden. Anschließend folgen beispielhaft verschiedene Auswertungen der mit Hilfe des k-NN Ansatzes generierten Prognosen.

Wie oben dargestellt, diente als Grundlage der Auswertungen eine Datenbank in Verbindung mit einer Softwareanwendung (kNN Biomass). Weitere Auswertungen wurden mit Hilfe von S-Plus 4.5, R und MS Excel durchgeführt. Die bereits dargestellten Schwierigkeiten in Bezug auf die Aquisition geeigneter Datengrundlagen führen dazu, dass eine Biomassedatenbank auf Einzelbaumebene niemals „fertig“ sein kann. Sie wird ständig um neue Datensätze ergänzt und ist somit immer im Wachstum begriffen. Ein Hauptvorteil instanzenbasierter Verfahren ist daher darin zu sehen, die Veränderungen der Datengrundlage in Echtzeit zu berücksichtigen. Gleichzeitig ergibt sich hieraus aber das Problem, dass Ergebnisse, die zu einem früheren Zeitpunkt generiert wurden, mit einer erweiterten Datengrundlage nicht mehr nachvollzogen werden können. Unter anderem aus diesem Grund werden zur Untersuchung der einzelnen Fragestellungen jeweils Teildatensätze verwendet, die sich jedoch in ihrem Umfang unterscheiden.

3.1 Sensitivität der Allometriekoeffizienten

Eine grundlegende Fragestellung in Bezug auf die Generalisierung von Biomassefunktionen ist, ob die im Bereich der Prozessmodelle abgeleiteten übergreifenden Annahmen mit empirischen Datengrundlagen nachgewiesen werden können, bzw. eine Integration beider Ansätze möglich ist.

Die Autokorrelation zwischen den Parametern der Grundgleichung der Allometrie wurde bereits in verschiedenen Zusammenhängen nachgewiesen. Hierbei wurden vielfach die Parameter veröffentlichter Biomassefunktionen gegenüber gestellt und die

vorhandenen Zusammenhänge untersucht (z.B. ZIANIS und MENCUCCINI, 2004). Solche Meta-Analysen zeigen eine hohe negative Korrelation zwischen den Allometriekoeffizienten a und b, die Unterschiede zwischen Laub- und Nadelbäumen erkennen lassen. Um diesen Zusammenhang anhand der vorliegenden Datengrundlagen nachzuweisen, wurde auf Datensätze von insgesamt 310 Fichten aus 12 verschiedenen destruktiven Untersuchungen zurückgegriffen. Die zugrunde gelegten Daten entstammen dabei zum Teil einer Literaturstudie von Wirth et al. (2004) sowie der im Rahmen dieser Studie durchgeführten Literaturrecherche bzw. Datensammlung.

Die Originaldaten der einzelnen Untersuchungen wurden soweit möglich in kleinere Teildatensätze aufgeteilt, um möglichst viele einzelne Schätzungen für die Regressionskoeffizienten des allometrischen Modells zu erhalten. Grundlage für die Aufteilung der Daten waren dabei die einzelnen Untersuchungsbestände. Auf Grundlage der Teildatensätze wurde dann eine einfache lineare Regression mit den logarithmisch transformierten Durchmesser- und Biomassewerten durchgeführt.

In der folgenden Zusammenstellung der Ergebnisse werden nur solche Datensätze berücksichtigt, die nach der beschriebenen Aufteilung eine Mindestgröße von $n = 5$ Bäumen nicht unterschreiten und zu einem R^2 größer 0,9 führen. Zwar ist dieser Stichprobenumfang als Grundlage einer Regressionsanalyse als relativ gering anzusehen, dies wird hier jedoch zugunsten einer möglichst hohen Anzahl an Parameterschätzungen toleriert. Tabelle 3-1 zeigt die geschätzten Regressionskoeffizienten für die insgesamt 30 Teildatensätze und den gesamten Zusammengefassten Datensatz.

Tabelle 3-1. Empirische Regressionskoeffizienten des einfachen allometrischen Modells ($agb=a\ BHD^b$) für die einzelnen Teildatensätze sowie für den Gesamtdatensatz mit Standardfehler der Residuen (RSE) und Korrekturfaktor KF (KF wurde nach Sprugel (1983) als *KF= exp(SEE2/2)* berechnet).

Datenherkunft	Land	n	a	b	RSE	KF[1]	R^2
Mund et al. (2002)	Deutschland	5	0,094	2,355	0,090	1,004	0,99
		5	0,032	2,859	0,118	1,007	0,97
		6	0,026	2,769	0,149	1,011	0,99
		7	0,234	2,102	0,105	1,006	0,97
		5	0,091	2,423	0,106	1,006	0,98
		5	0,199	2,295	0,050	1,001	0,99
		5	0,074	2,522	0,054	1,001	0,99
Heller u. Göttsche (1986)	Deutschland	18	0,040	2,706	0,165	1,014	0,99
Vyskot (1981)	Tschechische	15	0,088	2,569	0,118	1,007	0,99
	Republik	10	0,122	2,446	0,116	1,007	0,98
		15	0,176	2,340	0,203	1,021	0,95
Droste zu Hülshoff (1969)	Deutschland	5	0,495	2,011	0,199	1,020	0,94
Sharma (1992)	Deutschland	14	0,187	2,242	0,153	1,012	0,94
		15	0,146	2,320	0,124	1,008	0,97
		15	0,195	2,210	0,083	1,003	0,98
Duvigneaud et al. (1970)	Belgien	6	0,062	2,557	0,164	1,014	0,95
Poeppel (1989)	Deutschland	13	0,181	2,269	0,121	1,007	0,98
		9	0,718	1,866	0,108	1,006	0,97
Cerny (1990)	Tschechische	5	0,180	2,327	0,179	1,016	0,93
	Republik	5	0,453	2,030	0,091	1,004	0,98
		7	0,362	2,091	0,087	1,004	0,99
Rademacher (2004)	Deutschland	6	0,091	2,382	0,215	1,023	0,90
		6	0,118	2,369	0,128	1,008	0,99
		6	0,051	2,565	0,076	1,003	0,99
		6	0,018	2,852	0,113	1,006	0,96
		6	0,031	2,718	0,076	1,003	0,99
		6	0,056	2,545	0,089	1,004	0,99
Akca u. Mench (1993)	Deutschland	28	0,113	2,391	0,138	1,010	0,97
Johansson (1999)	Sweden	32	0,239	2,087	0,192	1,019	0,96
Neumann u. Jandl (2005)	Deutschland	24	0,121	2,368	0,168	1,014	0,99
Gesamtdatensatz		310	0,1095	2,402	0,224	1,025	0,97

Die geschätzten Werte für die Allometriekonstante b bewegen sich zwischen den Extremen 1,86 und 2,86; Der Mittelwert ist 2,38. Etwa 69 % der b-Werte fallen in das Konfidenzintervall zwischen 2,135 und 2,637. Damit sind die geschätzten Koeffizienten

im Mittel deutlich kleiner als der von WEST, BROWN und ENQUIST (WEST et al. 1999) theoretisch hergeleitete Exponent von 2,667.

Die hier abgeleiteten Ergebnisse decken sich gut mit einer Studie von ZIANIS und MENCUCCINI (2004), die im Rahmen einer Meta-Analyse von 277 Biomassemodellen, ohne auf die Ursprungsdaten zurückzugreifen, ähnliche Werte berichten. Auch die hohe negative Korrelation zwischen den Koeffizienten a und b kann anhand der vorliegenden Datengrundlage nachgewiesen werden. Der sich hier ergebende Zusammenhang ist wie in Abbildung 3-1 dargestellt relativ straff. Die angepasste Exponentialfunktion (hier durch logarithmische Skalierung der Ordinate in linearisierter Form dargestellt) hat mit einem R^2 von 0,95 einen hohen Erklärungsanteil.

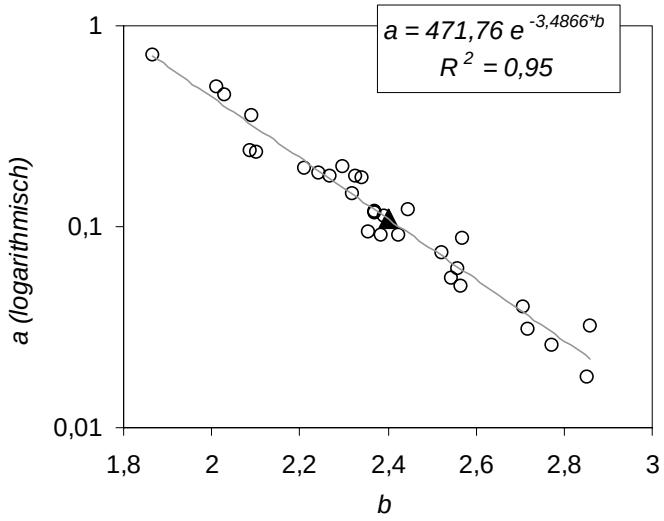


Abbildung 3-1. Autokorrelation der Regressionskoeffizienten a und b eines allometrischen Regressionsmodells ($agwb = a*BHD^b$) bei der Auswertung getrennter Teilmengen (Kreise) und des gesamten Fichtendatensatzes (Dreieck).

Um einen Vergleich der empirisch ermittelten Koeffizienten mit den theoretisch hergeleiteten Exponenten von WEST et al. (1999) zu ermöglichen, soll die Analyse im Folgenden nach der unter 2.2.1 beschriebenen Modifikation des BHD wiederholt werden.

3.1.1 Modifikation des BHD

Wie beschrieben, kann die Verwendung des BHD als Eingangsgröße für allometrische Funktionen kritisch betrachtet werden. Die festgelegte Messhöhe widerspricht dem Grundprinzip der Allometrie, die das Verhältnis relativer Messgrößen zueinander beschreibt. Durch die individuell unterschiedliche relative Messhöhe des BHD bei unterschiedlich hohen Bäumen, wird auf diese Weise nicht nur eine relative Durchmesserveränderung, sondern ebenso der Effekt einer Formveränderung im Bereich des Stammanlaufs erfasst. Um diesen Effekt abzumildern und die Veränderung der Parameter des allometrischen Modells bei Verwendung eines Durchmessers in relativer Baumhöhe zu untersuchen, wurde für den ausgewählten Fichtendatensatz der BHD mit Hilfe der Pain-Funktion (siehe 2.2.1) auf einen Durchmesser in 10% der Baumhöhe umgerechnet. Die daraus resultierenden Unterschiede der Eingangsgröße der Regression reichen dabei von nur einigen Millimetern für kleinere Bäume bis hin zu mehr als 13 cm für besonders große Individuen. Für diesen modifizierten Durchmesser wurden wiederum die Regressionskoeffizienten für ein lineares Modell mit den logarithmisch transformierten Eingangsgrößen ln $D_{0,1}$ und ln *agb* geschätzt.

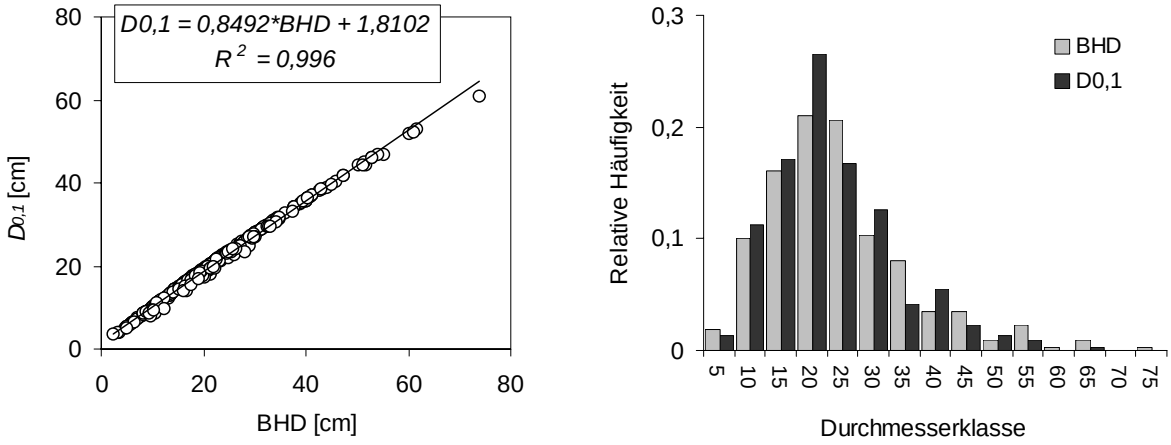


Abbildung 3-2. Beziehung zwischen dem BHD und dem mit Hilfe der Pain-Funktion berechneten Durchmesser in relativer (10%) Stammhöhe $D_{0,1}$ (links) und Durchmesserverteilung für den BHD und $D_{0,1}$.

Abbildung 3-2 zeigt die nur leicht unterproportional verlaufende Zunahme des $D_{0,1}$ mit steigendem BHD. Für alle Bäume mit einer Höhe von mehr als 13 m ist der $D_{0,1}$ kleiner als der BHD. Umgekehrt übersteigt der Durchmesser in relativer Höhe den BHD bei kleineren Bäumen, da sich der Referenzpunkt in Richtung Stammfuß verschiebt. Das Histogramm zeigt die Auswirkung dieser Änderungen auf die Durchmesserverteilung des Gesamtdatensatzes. Die auf Grundlage dieser modifizierten Eingangsgröße geschätzten Regressionskoeffizienten sind in Tabelle 3-2 zusammengefasst.

Tabelle 3-2. Geschätzte Regressionskoeffizienten für das allometrische Modell basierend auf dem Durchmesser in 10 % der Stammhöhe ($D_{0,1}$) für die einzelnen Teildatensätze sowie für den Gesamtdatensatz, Standardfehler der Residuen und Korrekturfaktor (KF wurde nach Sprugel (1983) als *KF= exp(SEE2/2)* berechnet).

Datenherkunft	n	$a_{D0.1}$	$b_{D0.1}$	RSE	KF	R^2
Mund et al. (2002)	5	0,044	2,700	0,145	1,011	0,99
	5	0,031	2,881	0,081	1,003	0,99
	6	0,009	3,261	0,118	1,007	0,99
	7	0,105	2,399	0,097	1,005	0,97
	5	0,037	2,748	0,107	1,006	0,98
	5	0,118	2,506	0,049	1,001	0,99
	5	0,027	2,879	0,047	1,001	0,99
Heller und Göttsche (1986)	18	0,031	2,837	0,157	1,012	0,99
Vyskot (1981)	15	0,049	2,819	0,129	1,008	0,99
	10	0,087	2,594	0,125	1,008	0,97
	15	0,132	2,472	0,219	1,024	0,94
Droste zu Hülshoff (1969)	5	0,355	2,159	0,195	1,019	0,95
Sharma (1992)	14	0,128	2,398	0,159	1,013	0,93
	15	0,107	2,446	0,131	1,009	0,96
	15	0,129	2,374	0,083	1,003	0,98
Duvigneaud et al. (1970)	6	0,030	2,828	0,158	1,013	0,96
Poeppel (1989)	13	0,103	2,487	0,119	1,007	0,98
	9	0,440	2,055	0,106	1,006	0,97
Cerny (1990)	5	0,109	2,528	0,193	1,019	0,91
	5	0,288	2,213	0,092	1,004	0,98
	7	0,216	2,295	0,089	1,004	0,98
Rademacher (2004)	6	0,033	2,752	0,176	1,016	0,93
	6	0,060	2,630	0,131	1,009	0,99
	6	0,027	2,836	0,090	1,004	0,99
	6	0,007	3,213	0,114	1,007	0,96
	6	0,011	3,093	0,067	1,002	0,99
	6	0,041	2,726	0,033	1,001	0,99
Akca und Mench (1993)	28	0,048	2,734	0,117	1,007	0,98
Johansson (1999)	32	0,282	2,099	0,185	1,017	0,96
Neumann und Jandl (2005)	24	0,042	2,787	0,168	1,014	0,99
Gesamtdatensatz	310	0,064	2,63	0,214	1,023	0,98

Abbildung 3-3 zeigt den negativen Zusammenhang zwischen $a_{0,1}$ und $b_{0,1}$. Da die Beziehung zwischen BHD und $D_{0,1}$ in diesem Fall einen nahezu linearen Verlauf zeigt

(siehe Abbildung 3-2), unterscheiden sich die Exponenten der angepassten Funktion zwischen a und b bzw. $a_{0,1}$ und $b_{0,1}$ nur geringfügig.

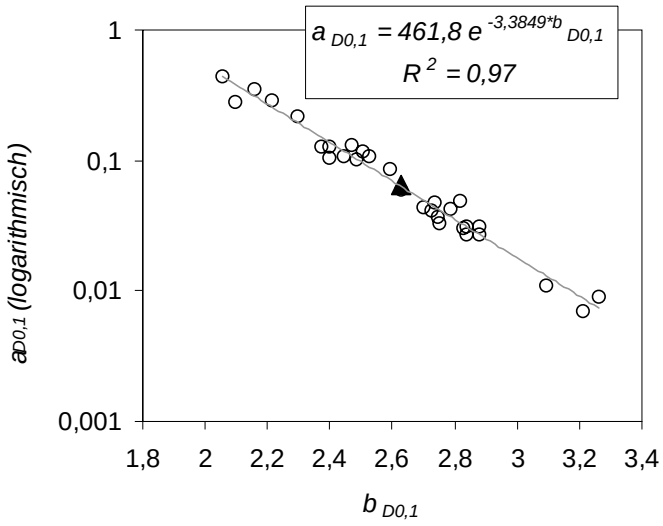


Abbildung 3-3 Autokorrelation der Regressionskoeffizienten $a_{D0.1}$ und $b_{D0.1}$ eines allometrischen Regressionsmodells bei der Auswertung getrennter Teilmengen (Kreis) und des gesamten Fichtendatensatzes (Dreieck) auf Grundlage des Durchmessers in relativer Stammhöhe ($D_{0,1}$).

Abbildung 3-4 zeigt die Häufigkeitsverteilung der Allometriekoeffizienten b und $b_{0,1}$. Die Verteilung ist aufgrund der Durchmessertransformation deutlich nach rechts verschoben.

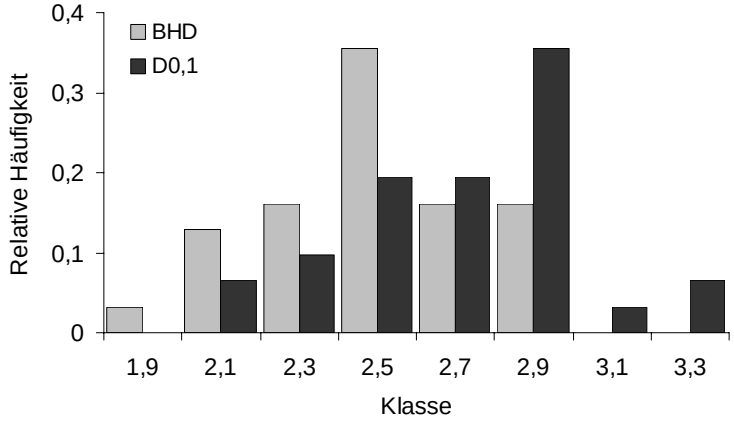


Abbildung 3-4. Relative Häufigkeit der Allometriekoeffizienten b und $b_{0,1}$ für die ausgewählten Datensätze.

Wie aus Tabelle 3-2 deutlich wird, erhöht sich das Bestimmtheitsmaß der einzelnen Regressionen nach der Durchmessertransformation nicht in jedem Fall, ist aber im Mittel mit 0,972 höher als bei der Verwendung des BHD als Eingangsgröße. Weiterhin ist für die einzelnen Datensätze eine bessere strukturelle Anpassung des Modells festzustellen.

Diese Verbesserung ist auch für die Modellanpassung auf Grundlage des Gesamtdatensatzes zu beobachten. Wie in Abbildung 3-5 dargestellt, scheint die strukturelle Anpassung des allometrischen Modells mit dem $D_{0,1}$ als unabhängiger Variablen besonders im Bereich großer Durchmesser gegenüber dem Modell mit dem BHD als Eingangsgröße verbessert zu sein.

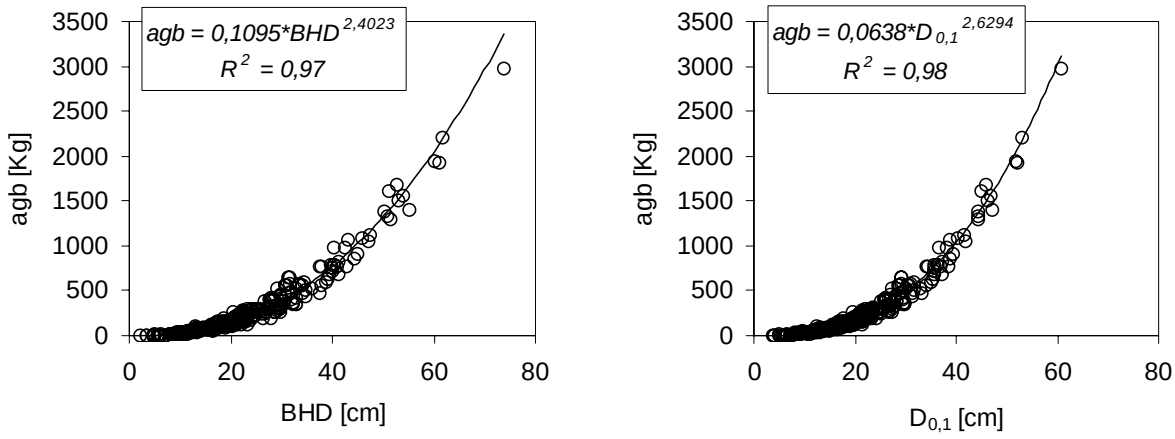


Abbildung 3-5. Oberirdische Gesamtbiomasse (*agb*) über dem BHD (links) und dem relativen Durchmesser in 10% der Stammhöhe $D_{0,1}$ (rechts).

Der geschätzte Exponent $b_{0,1}$ ist mit 2,63 zwar immer noch geringer als der theoretisch hergeleitete, dies wird allerdings auch erwartet, da Bäume im Laufe der Zeit durch das Absterben von Ästen an Masse verlieren (Enquist, 2002). Abbildung 3-6 zeigt die sich aus der Verwendung des $D_{0,1}$ ergebende Verschiebung der beiden Regressionskoeffizienten.

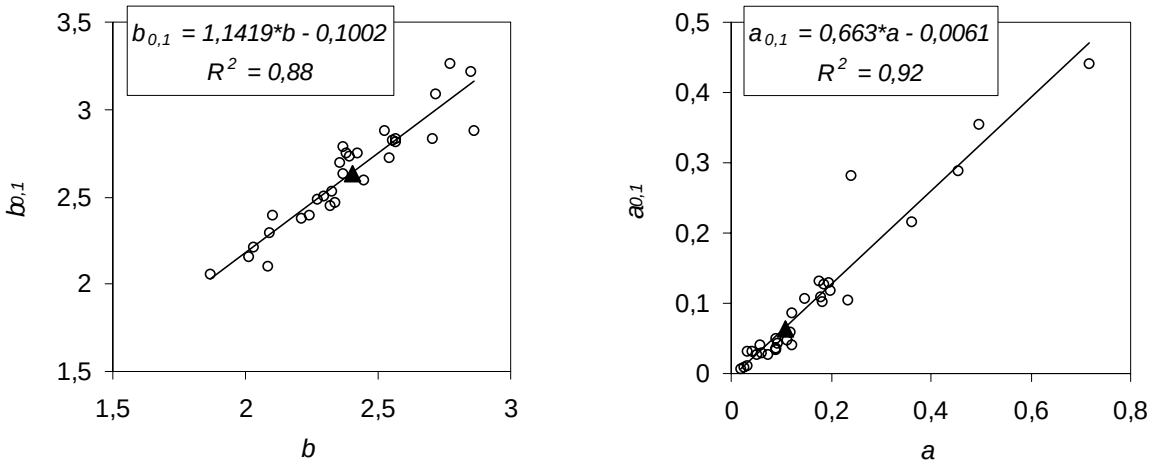


Abbildung 3-6. Beziehung der Koeffizienten b zu $b_{0,1}$ (links) und a zu $a_{0,1}$ (rechts).

KETTERINGS et al. (2001) schlagen vor, den Parameter b der allometrischen Biomassefunktion aus dem Allometriekoeffizienten der Höhenbeziehung (b') zu schätzen. Ein ähnlicher Ansatz zur Berücksichtigung der Höhe als weiterer Einflussgröße ergibt sich implizit auch aus dem vorgestellten Prozessmodell von WEST, BROWN und ENQUIST (ENQUIST, 2002). Auch ZIANIS und MENCUCCINI (2004) können in ihrer Meta- Analyse einen solchen Zusammenhang mit relativ hoher Sicherheit ableiten. Die Untersuchung dieses Zusammenhangs in dem hier verwendeten Teildatensatz kann diese Korrelation jedoch kaum nachweisen. Hierbei muss jedoch berücksichtigt bleiben, dass sich die erweiterte Potenzfunktion als Grundgleichung der Allometrie in den meisten Fällen als relativ ungeeignet zur Modellierung der Baumhöhe in Abhängigkeit des BHD erweist. Das Modell ist zwar für kleinere Durchmesserbereiche geeignet die Entwicklung der Baumhöhe adäquat abzubilden, für größere Durchmesserbereiche werden allerdings typischerweise eher logarithmische Funktionen verwendet. Abbildung 3-7 zeigt die Beziehung dieser Koeffizienten für die 24 ausgeschiedenen Einzeldatensätze dieser Teiluntersuchung.

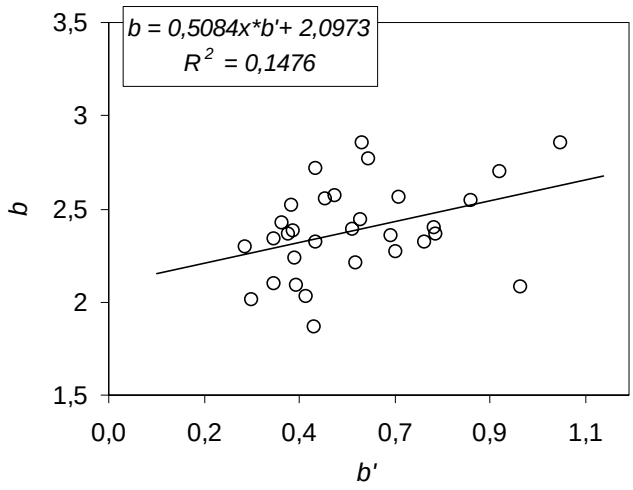


Abbildung 3-7. Beziehung zwischen dem Skalierungskoeffizienten b der allometrischen Biomassefunktion $agb = a*BHD^b$ und dem Allometriekoeffizienten der Höhenfunktion b' ($h = k*BHD^{b'}$) für die 24 untersuchten Einzeldatensätze.

Da davon auszugehen ist, dass sich die Variabilität der Regressionskoeffizienten zum Teil durch Unterschiede in der Baumhöhe erklären lässt, soll dieser Einfluss im Folgenden am Beispiel eines größeren Datensatzes näher untersucht werden.

3.2 Ableitung von Referenzmodellen

Um eine objektive Einschätzung der Effizienz der k-NN Methode im Vergleich zu parametrischen Verfahren zu ermöglichen, soll im Folgenden näher auf die hierfür nötige Modellbildung eingegangen werden. Hierbei wurde exemplarisch auf einen Teildatensatz von Fichten zurückgegriffen, da für diese Baumart besonders viele Einzeldatensätze vorhanden sind (n=578). Im Weiteren werden außerdem für die vorliegende Datengrundlage geeignete Regressionsmodelle für Kiefer und Buche dargestellt, die zur Überprüfung des k-NN Ansatzes für diese Baumarten verwendet werden.

Die Scatterplots in Abbildung 3-8 geben einen Überblick des Zusammenhangs zwischen der oberirdischen Gesamtbiomasse (agb) und den unabhängigen Variablen BHD und Baumhöhe in metrischer und logarithmischer Skalierung.

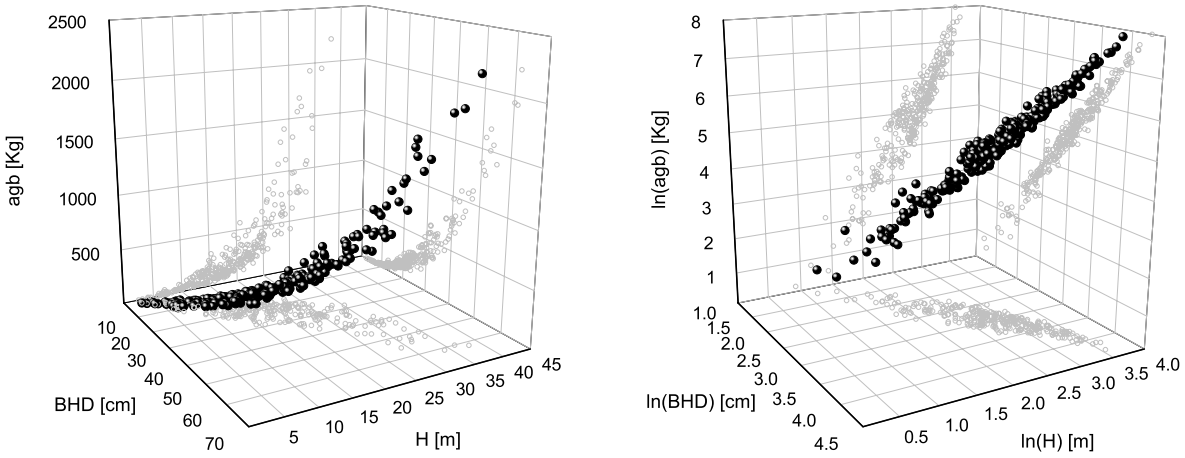


Abbildung 3-8. Dreidimensionale Scatterplots der oberirdischen Gesamtbiomasse (*agb*) in Abhängigkeit des BHD und der Baumhöhe (links), bzw. der jeweils logarithmisch transformierten Variablen (rechts). Zusätzlich sind die jeweiligen Projektionen der Daten auf die zweidimensionalen Ebenen dargestellt.

Um eine möglichst objektive Auswahl geeigneter Modellformulierungen zu treffen, wurden unterschiedliche Modellvarianten zugrunde gelegt und mit Hilfe verschiedener Gütemaße verglichen. Um eine Gleichverteilung der Fehlervarianzen zu gewährleisten, wurde hierzu wie unter 2.1.2 beschrieben auf die logarithmisch transformierten unabhängigen und abhängigen Variablen zurückgegriffen.

Eine andere Möglichkeit der bestehenden Heteroskedastizität der Datengrundlage zu begegnen besteht darin, die Proportionalität der Residuen zur verwendeten Zielgröße durch eine gewichtete Regression zu eliminieren und so eine tendenzfreie Schätzung der Regressionskoeffizienten zu ermöglichen (PARRESOL, 2001). Der Vorteil dieser Vorgehensweise wird darin gesehen, dass hierbei die nötige Bias-Korrektur bei der Rücktransformierung der Werte auf das originäre Skalenmaß entfällt (siehe z.B. LAMBERT et al., 2005).

Die verwendeten Kandidatenmodelle unterscheiden sich in der Anzahl der unabhängigen Variablen sowie in der Komplexität der Modellformulierung. Neben dem BHD als unabhängiger Variablen mit dem höchsten Erklärungsanteil, wurde die Baumhöhe als zweite Variable verwendet. Der Argumentation in 2.2 folgend, wird die Höhe besonders im Fall von zusammengesetzten Datensätzen, in denen standortsabhängige Unterschiede in der Durchmesser-Höhenbeziehung vorliegen, als wichtig angesehen.

Wie aus Abbildung 3-8 deutlich wird, ist die Interpretation der Regressionskoeffizienten durch die vorliegende multiple Kolinearität der logarithmisch transformierten Variablen erschwert. Um einen Eindruck über den vorliegenden Ursache-Wirkung Zusammenhang zwischen dem Durchmesser- bzw. Höhenzuwachs und der Biomasse zu erhalten, kann man jeweils eine der Variablen fixieren bzw. in möglichst kleine Klassen einteilen, um die partielle Auswirkung der jeweils anderen Variablen zu untersuchen. Die Scatterplots in Abbildung 3-9 zeigen die Entwicklung der Biomasse über dem BHD für 9 separate Höhenklassen. Da sich die Wertebereiche des BHDs aufgrund der vorliegenden Durchmesser - Höhenverteilung in den gebildeten Höhenklassen überlagern würden, sind die Klassen hier jeweils einzeln dargestellt.

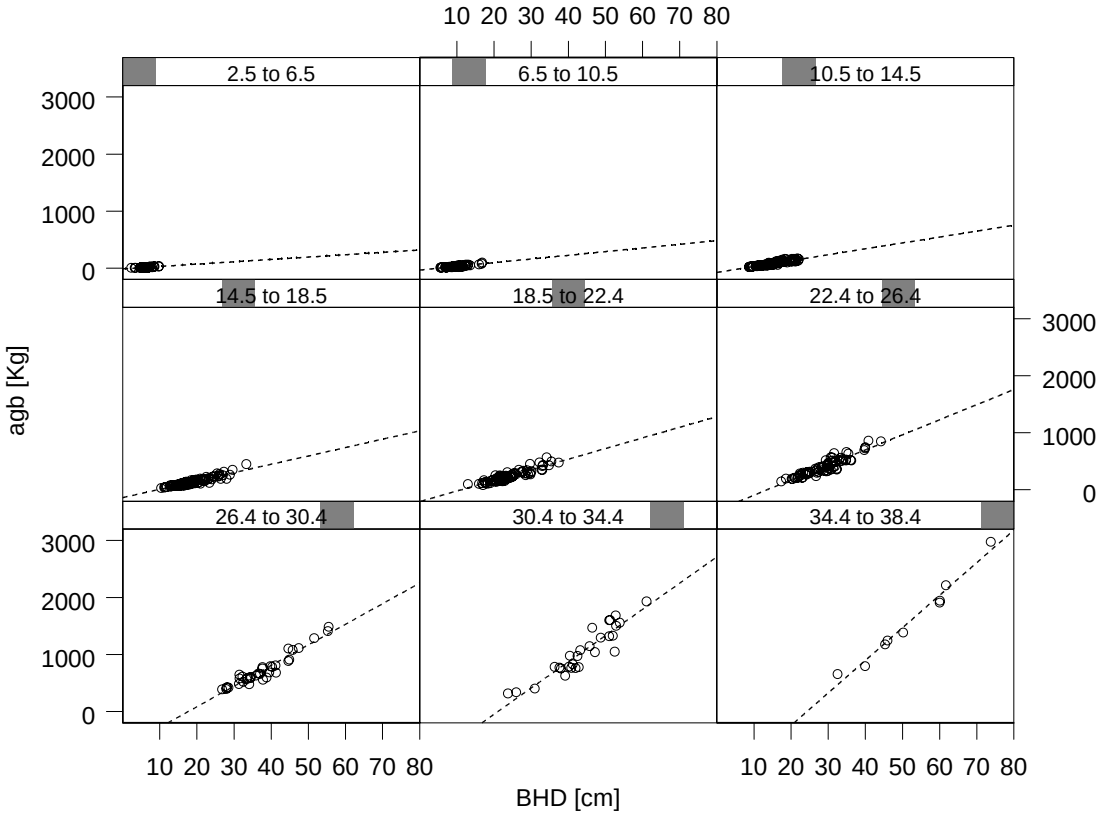


Abbildung 3-9. Scatterplots der oberirdischen Gesamtbiomasse (*agb*) über dem BHD für 9 separate Höhenklassen.

Bemerkenswert erscheint, dass sich innerhalb der so gebildeten Klassen ein nahezu linearer Trend zwischen dem BHD und der oberirdischen Gesamtbiomasse abzeichnet. Eine Ausnahme bildet die größte Höhenklasse (34,4 – 38,4 m). Da sie aufgrund der vor-

liegenden Höhenverteilung einen relativ großen Durchmesserbereich abdeckt, ist hier die typische überproportionale Zunahme der Biomasse mit steigendem BHD besser zu erkennen. Auch für alle anderen Klassen liefert zwar die erweiterte Potenzfunktion eine bessere Anpassung, die einzelnen Allometriekoeffizienten b sind jedoch nach der partiellen Elimination des Einflusses der Baumhöhe erwartungsgemäß weitaus kleiner als für die unklassierten Daten.

Entgegen einer in der Literatur häufig dargestellten, auf einfachen geometrischen Überlegungen basierenden Meinung, die Biomasse sowie das Volumen müssten sich tendenziell proportional zur Stammgrundfläche bzw. zu (BHD^2*H) entwickeln (siehe z.B. ÈERNY, 1990; KETTERINGS et al., 2001; XIAO et al., 2003; ALBERTI et al., 2005; CHAVE et al., 2005), wird anhand der vorliegenden Datengrundlage deutlich, dass sich die Masse in den einzelnen Höhenklassen unterproportional zur Stammquerschnittsfläche entwickelt. Von einer Fixierung des wechselseitigen Einflusses dieser Variablen durch die Verwendung der Kombination D^2*H als unabhängige Variable soll daher im Rahmen der folgenden Regressionsanalyse abgesehen werden.

Ein ähnliches Bild ergibt sich, wenn man die Beziehung zwischen Baumhöhe und Biomasse in einzelnen Durchmesserbereichen untersucht (Abbildung 3-10). Die Breite der Durchmesserklassen ist in dieser Darstellung nicht einheitlich, sondern mit dem Ziel möglichst gleiche Anzahlen an Bäumen pro Klasse zu berücksichtigen, verschieden gewählt.

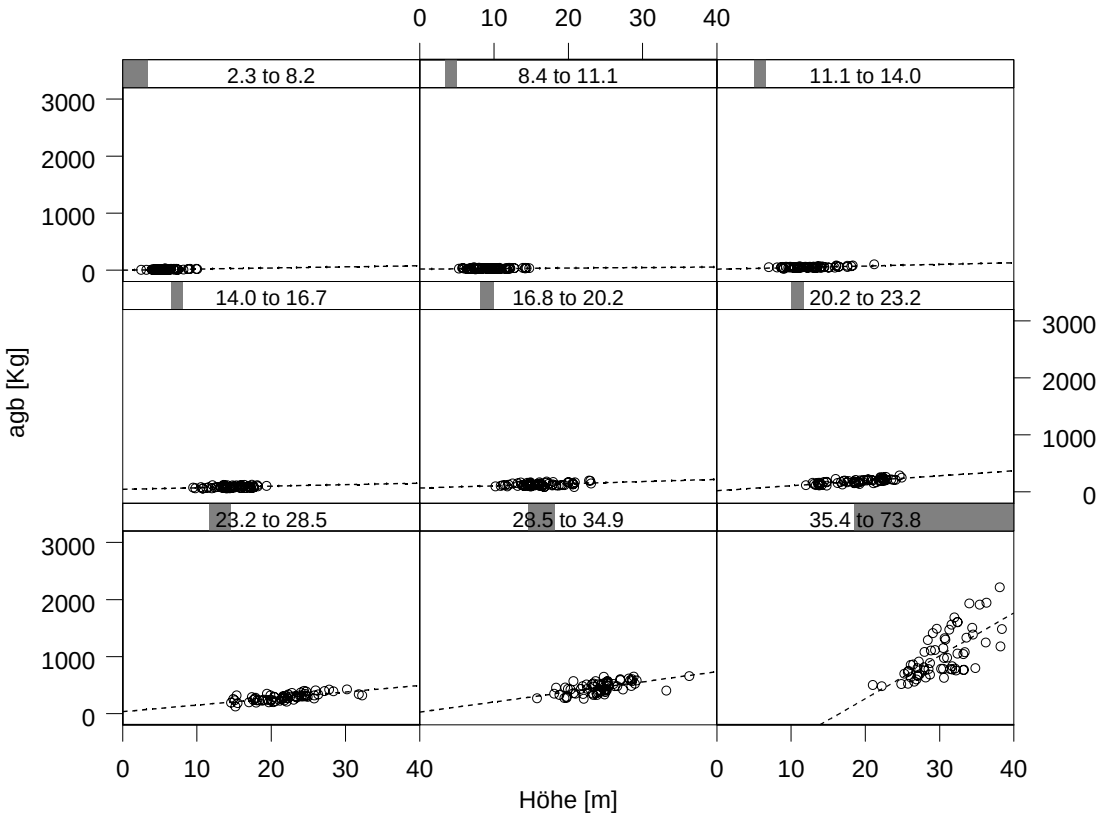


Abbildung 3-10. Separate Scatterplots der oberirdischen Gesamtbiomasse (*agb*) über der Baumhöhe nach Einteilung des Gesamtdatensatzes in 9 ungleich breite Durchmesserklassen mit jeweils gleicher Anzahl Bäume.

Während sich hier für den Gesamtdatensatz ein funktionaler Zusammenhang zwischen Baumhöhe und Biomasse nur mit einem, im Vergleich zum BHD, relativ geringen Bestimmtheitsmaß (R^2 = 0,86) nachweisen lässt, ergeben sich für ungleiche Durchmesserklassen gleicher Baumanzahl sehr straffe lineare Beziehungen. Lediglich in der größten Durchmesserklasse (35,4 – 73,8 cm) scheint die Baumhöhe alleine die Variabilität der Biomassewerte nur ungenügend erklären zu können. Ähnlich wie bei der vorangegangenen Betrachtung muss hierbei berücksichtigt werden, dass diese Durchmesserklasse aufgrund der vorliegenden Durchmesser-Höhenverteilung einen besonders großen Höhenbereich abdeckt. Es ist daher zu erwarten, dass sich bei einer weiteren Einteilung dieser Klasse ebenfalls straffere lineare Trends ergeben könnten.

In der empirischen Datenanalyse führt die Beziehung zwischen den logarithmisch transformierten Werten zur Wahl eines Modells erster Ordnung, dass in der Lage ist,

die unterschiedlichen Steigungen der linearen Beziehungen zwischen BHD, Höhe und Biomasse abzubilden (MENDENHALL und SINCICH, 1993; WILHELMSEN und VESTJORDET, 1974). Die hohe Korrelation zwischen BHD und Höhe kann aber auch die Verwendung des Schlankheitsgrades bzw. h-d- Verhältnisses als Regressorvariable rechtfertigen. Die resultierenden Modellformulierungen sind in Tabelle 3-3 aufgeführt

Ein weiterer Aspekt der vorliegenden Datengrundlage ist, dass sie aus verschiedenen Einzeldatensätzen zusammengesetzt ist. Diese Daten einzelner Studien bestehen oftmals aus Probebäumen, die in separaten Probeplots oder Beständen aufgenommen wurden. Dies gilt insbesondere für die Daten der Finnischen Nationalen Waldinventur, die für eine Teiluntersuchung zur Verfügung gestellt wurden. FEHRMANN et al. (2006) verwendeten zur Modellierung dieser Daten ein gemischt lineares Regressionsmodell[4] (MCCULLOCH und SEARLE, 2000; LAPPI et al., 2006). Aufgrund des Aufnahmeverfahrens, bei dem mehrere Bäume pro NFI-Probeplot aufgenommen wurden, die einzelnen Plots aber gleichzeitig so weiträumig verteilt sind, dass eine geringe Kovarianz zwischen den Daten zu erwarten ist, können die Probebäume aus einzelnen Plots als eine Art Subpopulation angesehen werden. Es wird also erwartet, dass die verwendeten Variablen innerhalb der Plots höher korreliert sind als in der Gesamtpopulation (LAPPI et al., 2006). Die generelle Ausdrucksform bei den gegebenen unabhängigen Variablen BHD und Höhe ist dann:

$$\ln agb_{ki} = \ln \alpha + \ln a_k + \beta \ln BHD_{ki} + \chi \ln h_{ki} + \varepsilon_{ki}$$

wobei agb_{ki}, BHD_{ki}, h_{ki}, h/d_{ki} und ε_{ki} die oberirdische Gesamtbiomasse, der BHD, die Baumhöhe sowie der Schlankheitsgrad und das Residuum von Baum i auf Plot k ist.

[4] Synonyme sind „mixed linear model“ oder „mixed effect model“.

Vorhersagen für den zusätzlichen Effekt ln a_k werden hierbei für jeden Plot einzeln geschätzt (mit einem Erwartungswert von 0), so dass gewissermaßen einzelne plotspezifische Biomassemodelle abgeleitet werden. Solche Modelle können also ebenso wie eine gewichtete Regression nur dann sinnvoll verwendet werden, wenn geeignete Datengrundlagen ausgewertet werden sollen. In der Praxis sind regionalisierte Daten nur selten vorhanden. Trotzdem werden die beschriebenen Modelle hier sozusagen als bestmögliche Referenz verwendet, um die Effizienz der *k*-NN Methode zu evaluieren. Die Zusammenstellung aller verwendeten Referenzmodelle ist in Tabelle 3-3 aufgeführt.

Tabelle 3-3. Verschiedene Modellformulierungen unterschiedlicher Komplexität, die im Folgenden als Referenz zu den durchgeführten *k*-NN Schätzungen verwendet werden.

Modell	Regressionsform	Beschreibung
1	$\ln agb_i = \ln \alpha_i + \beta \ln BHD_i + \varepsilon_i$	Einfaches lineares Modell
2	$\ln agb_i = \ln \alpha_i + \beta \ln BHD_i + \chi \ln h_i + \varepsilon_i$	Einfaches lineares Modell
3	$\ln agb_i = \ln \alpha_i + \beta \ln BHD_i + \chi \left[\frac{h}{BHD} \right]_i + \varepsilon_i$	Einfaches lineares Modell
4	$\ln agb_{ki} = \ln \alpha + \ln a_k + \beta \ln BHD_{ki} + \chi \ln h_{ki} + \varepsilon_{ki}$	Gemischt lineares Modell
5	$\ln agb_{ki} = \ln \alpha + \ln a_k + \beta \ln d_{ki} + \chi \left[\frac{h}{BHD} \right]_{ki} + \varepsilon_{ki}$	Gemischt lineares Modell

Das Modell 1 entspricht auf dem metrischen Skalenniveau dem einfachen allometrischen Ansatz $agb = a\ BHD^b$ (Modell 2 respektive $agb = a\ *BHD^b\ h^c$).

Eine Anwendung ausgewählter Modellformulierungen auf den gesamten Datenbestand findet sich in Anhang VI ab Seite 148.

3.3 Verfahrensvergleich und Evaluation

Wie bereits dargestellt, wird der Vergleich der *k*-NN Methode mit den gegebenen Regressionsmodellen anhand verschiedener Teildatensätze durchgeführt. Hierzu wurden die zugrunde gelegten Datensätze jeweils zufällig aufgeteilt, so dass ein Datensatz (im Folgenden jeweils als *modelling* bezeichnet) einerseits zur Schätzung der Regressions-

koeffizienten der aufgeführten Modelle und andererseits zur Parametrisierung der Distanz- bzw. Gewichtungsfunktion des *k*-NN Algorithmus verwendet wird. Der verbleibende, nicht an der Modellbildung beteiligte Datensatz (im Folgenden als *test* bezeichnet) wird zur Evaluierung der Schätzergebnisse mit Hilfe der beschriebenen Gütemaße herangezogen. Da einige Datensätze nur als Grundlage für bestimmte Veröffentlichungen (z.B. FEHRMANN et al., 2005; FEHRMANN et al., 2006) zur Verfügung gestellt wurden, werden im Folgenden einige separate Auswertungen dargestellt.

3.3.1 Teilauswertung I

Diese Teilauswertung bezieht sich auf die vom Finnish Forest Research Institute (METLA) bereitgestellten Fichten und Kieferndaten. Die vorhandenen Datensätze wurden jeweils zufallsbasiert in ein *modelling*- Subset (n=143 für Fichte und n=145 für Kiefer) und ein *test*- Subset (jeweils n=60) unterteilt. Abbildung 3-11 zeigt die Durchmesser-verteilungen in den einzelnen Datensätzen.

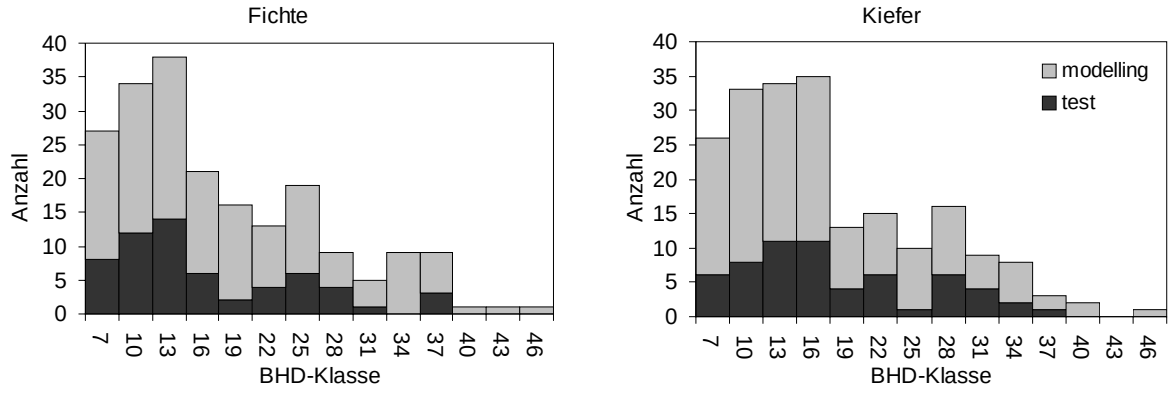


Abbildung 3-11. Histogramme der Durchmesserklassen der zufällig ausgeschiedenen *test*- Datensätze (jeweils n=60) und der zur Modellierung und *k*-NN Schätzung verbleibenden *modelling*- Daten für Fichte (n=143) und Kiefer (n=145).

Die zu schätzende Zielgröße ist in diesem Fall die oberirdische Gesamtbiomasse (*agb*). Die geschätzten Regressionskoeffizienten für die einfachen linearen Modelle sowie die gemischt linearen Modelle mit dem jeweiligen Standardfehler sind in Tabelle 3-4 aufgeführt.

Tabelle 3-4. geschätzte Regressionskoeffizienten des einfachen linearen Modells (Modell 2 für Kiefer, Modell 3 für Fichte) sowie eines gemischt linearen Modells (Modell 4 und 5) für die zufällig ausgeschiedenen *modelling*- Datensätze für Kiefer (n=145) und Fichte (n=143).

Baumart	Koeffizient	Schätzung	Std. Fehler	t-Wert	p-Wert
Einfache lineare Modelle:					
Kiefer	α	-2,355	0,048	-48,88	<0,0001
	β	2,202	0,041	53,30	<0,0001
	χ	0,272	0,042	6,39	<0,0001
Fichte	α	-1,973	0,064	-16,61	<0,0001
	β	2,345	0,024	94,40	<0,0001
	χ	0,055	0,079	0,46	0,4830
Gemischt lineare Modelle					
Kiefer	α	-2,36	0,054	-43,47	<0,0001
	β	2,19	0,042	52,00	<0,0001
	χ	0,29	0,045	6,36	<0,0001
Fichte	α	-2,13	0,139	-15,33	<0,0001
	β	2,36	0,030	78,05	<0,0001
	χ	0,111	0,093	1,19	0,2358

Eine detaillierte Übersicht der Regressionsanalyse für die einfachen linearen Modelle findet sich in Anhang III auf Seite 140. Die gemischt linearen Modelle wurden aus FEHRMANN et al. (2006) übernommen und wurden im Rahmen dieser Zusammenarbeit von A. LEHTONEN parametrisiert.

Zur Parametrisierung des k-NN Algorithmus bzw. der Variablengewichtung wurden mit Hilfe von 50 Iterationen ein angepasstes Gewichtungsverhältnis der Variablen bestimmt. Die besten Ergebnisse wurden hierbei mit der euklidischen Distanzmetrik (Minkowski Konstante $c = 2$) erzielt. Für die Gewichtungsfaktoren der Variablen BHD (w_{bhd}) und Baumhöhe (w_h) wurden gute Ergebnisse mit w_{bhd}=0,8 und w_h =0,2 für Fichte und w_{bhd}=0,75 und w_h =0,25 für Kiefer erzielt. Der Gewichtungsparameter t wurde hierbei für Fichte = 0 und für Kiefer = 1 gesetzt.

Zur angepassten Wahl der Größe der Nachbarschaft wurden mit dem unter 2.6 beschriebenen Modul zur multiplen Kreuzvalidierung verschiedene Bandbreiten ausgetestet. Abbildung 3-12 zeigt den Verlauf des RMSE % und des Bias für unter-

schiedlich große Nachbarschaften (k) im Fall der zugrunde gelegten *modelling*- Datensätze.

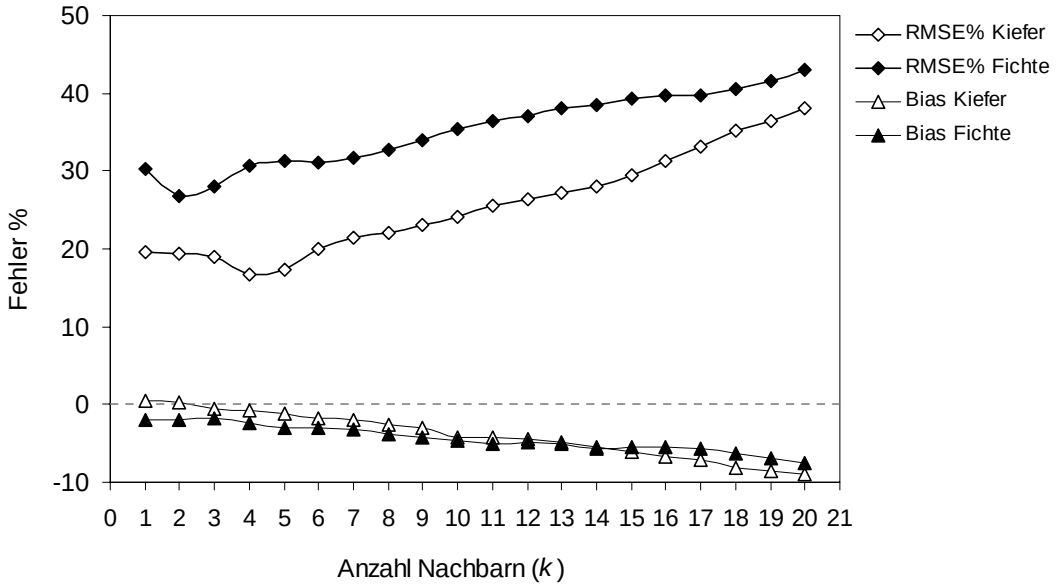


Abbildung 3-12. Entwicklung des RMSE % und des Bias bei unterschiedlicher Größe der Nachbarschaft (k) für die gegebenen *modelling*- Datensätze beider Baumarten. Berechnungsgrundlage für jedes k ist eine komplette Kreuzvalidierung der jeweiligen Trainingsdaten.

Der typische Fehlerverlauf ist durch ein Absinken am Anfang, das Durchlaufen eines lokalen Minimums und einem mit größer werdender Nachbarschaft steigendem Fehler, der zum größten Teil durch einen zunehmenden Bias verursacht wird, gekennzeichnet. Im Fall des vorliegenden Teildatensatzes kann eine Fehlerminimierung durch die Verwendung von nur 2 Nachbarn bei Fichte (RMSE % = 21) und 4 Nachbarn bei Kiefer (RMSE % = 16,5) erreicht werden. Eine minimale Verringerung des Fehlers konnte für Fichten durch die Wahl der Kernel-Methode (siehe 2.5.5) erzielt werden. Hierzu wurde der maximale normierte Abstand bis zu dem Nachbarn berücksichtigt wurden, auf 0,05 gesetzt (siehe Abbildung 3-13 links). Die resultierende Anzahl der berücksichtigten Nachbarn ist in Abbildung 3-13 (rechts) über dem BHD, der in diesem Fall die am höchsten gewichtete Variable ist, dargestellt.

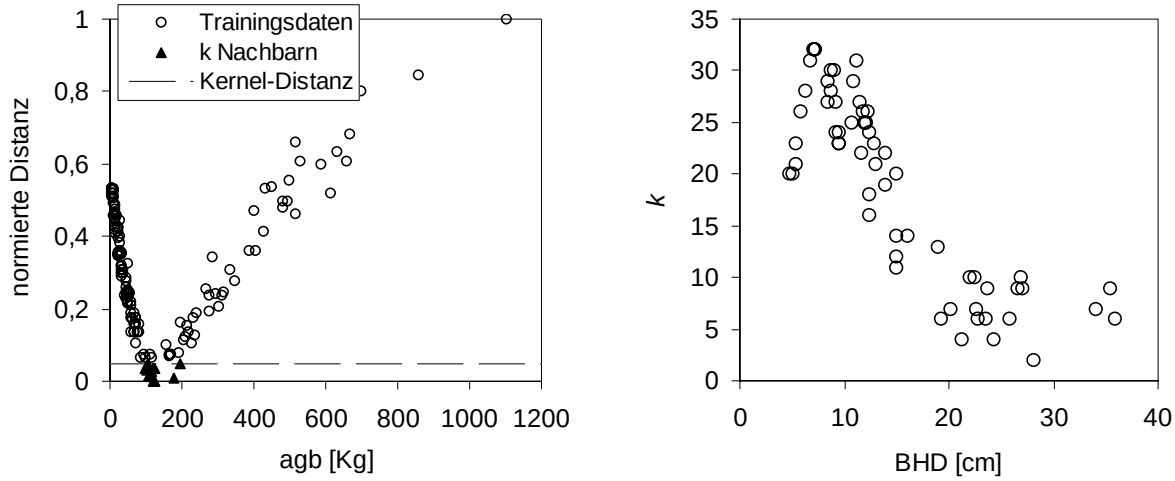


Abbildung 3-13. Berücksichtigte Nachbarn für einen bestimmten Abfragepunkt (links) und resultierende unterschiedliche Größe der berücksichtigten Nachbarschaft (k) über dem BHD durch die Verwendung einer festen Bandbreite (Kernel-Distanz = 0,05) für den zugrunde gelegten Fichtendatensatz.

Hierbei ist deutlich zu erkennen, dass die Größe der berücksichtigten Nachbarschaft einen ähnlichen Verlauf wie die Durchmesserverteilung des Datensatzes zeigt. D.h., dass in den Bereichen der Durchmesserverteilung, in denen viele Datenpunkte vorhanden sind, mehr Nachbarn zur Schätzung verwendet werden, wobei in Bereichen mit geringer Anzahl von Trainingsinstanzen die berücksichtigte Nachbarschaft kleiner ist.

Ob die auf Basis der euklidischen Distanzmetrik berechneten Abstände hierbei auch Ähnlichkeiten im Profilverlauf der beiden verwendeten Variablen erfassen, wurde durch eine zusätzliche Berechnung des Q-Korrelationskoeffizienten überprüft (siehe Anhang III). Da die hier verwendeten Variablen BHD und Höhe typischerweise eine hohe Kovarianz aufweisen, ist in diesem Fall davon auszugehen.

Die parametrisierten Regressionsmodelle sowie die k-NN Schätzung wurden anschließend auf die jeweiligen *test-* Datensätze (jeweils n=60) beider Baumarten angewendet. Die Beziehung zwischen beobachteter und geschätzter Biomasse für beide Verfahren ist in Abbildung 3-14 dargestellt.

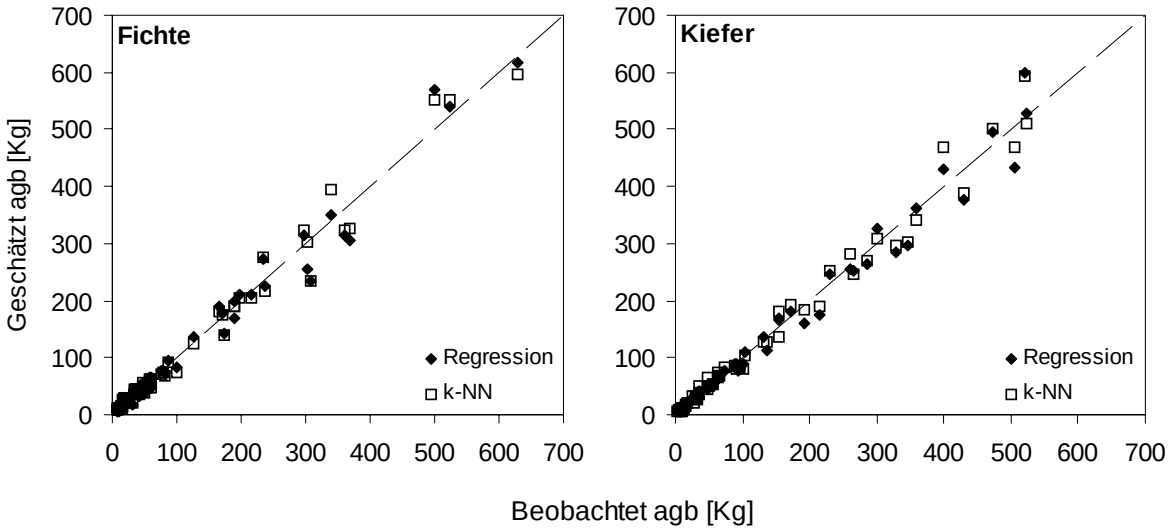


Abbildung 3-14. Beziehung zwischen beobachteter und geschätzter oberirdischer Gesamtbiomasse (*agb*) mit dem zugrunde gelegten Regressionsmodell (einfaches lineares Modell) sowie dem *k*-NN Ansatz für die *test*- Datensätze (n=60) von Fiche (links) und Kiefer (rechts).

Augenscheinlich sind die Unterschiede der Schätzungen im vorliegenden Fall sehr gering. Die Residuen beider Schätzungen sind gleichmäßig verteilt und weisen für keines der Verfahren bemerkenswerte Ausreißer auf. Dass die *k*-NN Schätzung hier in der Lage ist auch die Extremwerte der Verteilung angemessen abzubilden, lässt darauf schließen, dass die zufällig ausgeschiedenen Trainingsdaten (*modelling*- Subsets) einen weiteren Durchmesser- und Höhenbereich abdecken als die *test*- Daten. Um die Fehlerstruktur beider Voraussagen genauer zu untersuchen, wurde ein Gütemaß-mix (siehe 2.5.4) berechnet. Die einzelnen Fehlermaße sind in Tabelle 3-5 aufgeführt.

Tabelle 3-5. Gütemaße der Regressions- sowie des *k*-NN Ansatzes auf Grundlage der jeweiligen *test*- Datensätze (n=60) für Kiefer.

Modellierungsansatz	RMSE	RMSE %	MAPE	ME
Kiefer:				
$\ln agb_i = \ln \alpha + \beta \ln BHD_i + \chi \ln h_i + \varepsilon_i$	20,68	15,79	9,67	-2,562
$\ln agb_{ki} = \ln \alpha + \ln a_k + \beta \ln BHD_{ki} + \chi \ln h_{ki} + \varepsilon_{ki}$	19,76	15,00	9,21	-1,718
k-NN	19,41	14,54	12,61	0,009

Tabelle 3-6. Gütemaße der Regressions- sowie des *k*-NN Ansatzes auf Grundlage der jeweiligen *test*- Datensätze (n=60) für Fichte.

Modellierungsansatz	RMSE	RMSE %	MAPE	ME
Fichte:				
$\ln agb_i = \ln \alpha + \beta \ln BHD_i + \chi \left[\frac{h}{BHD} \right]_i + \varepsilon_i$	22,39	19,19	13,61	-0,938
$\ln agb_{ki} = \ln \alpha + \ln a_k + \beta \ln BHD_{ki} + \chi \left[\frac{h}{BHD} \right]_{ki} + \varepsilon_{ki}$	20,31	17,36	13,73	-0,398
k-NN	19,19	16,42	13,98	-0,493

Während der RMSE sowie der mittlere Fehler (ME) der *k*-NN Schätzungen für beide Baumarten geringer ist als für das einfache lineare Modell, zeigt der mittlere absolute prozentuale Fehler (MAPE) einen leichten Anstieg gegenüber der Regressionsfunktion. D.h., die mit dem Ausgangsniveau der Daten gewichteten Fehler sind für die *k*-NN Schätzung im mittel höher als für das verwendete Regressionsmodell. Abbildung 3-15 zeigt die beobachteten und geschätzten Werte über dem BHD. Aus Darstellungsgründen ist hierbei eine unterschiedliche Skalierung der oberirdischen Gesamtbiomasse (*agb*) für jeweils zwei Hälften des Durchmesserbereiches gewählt.

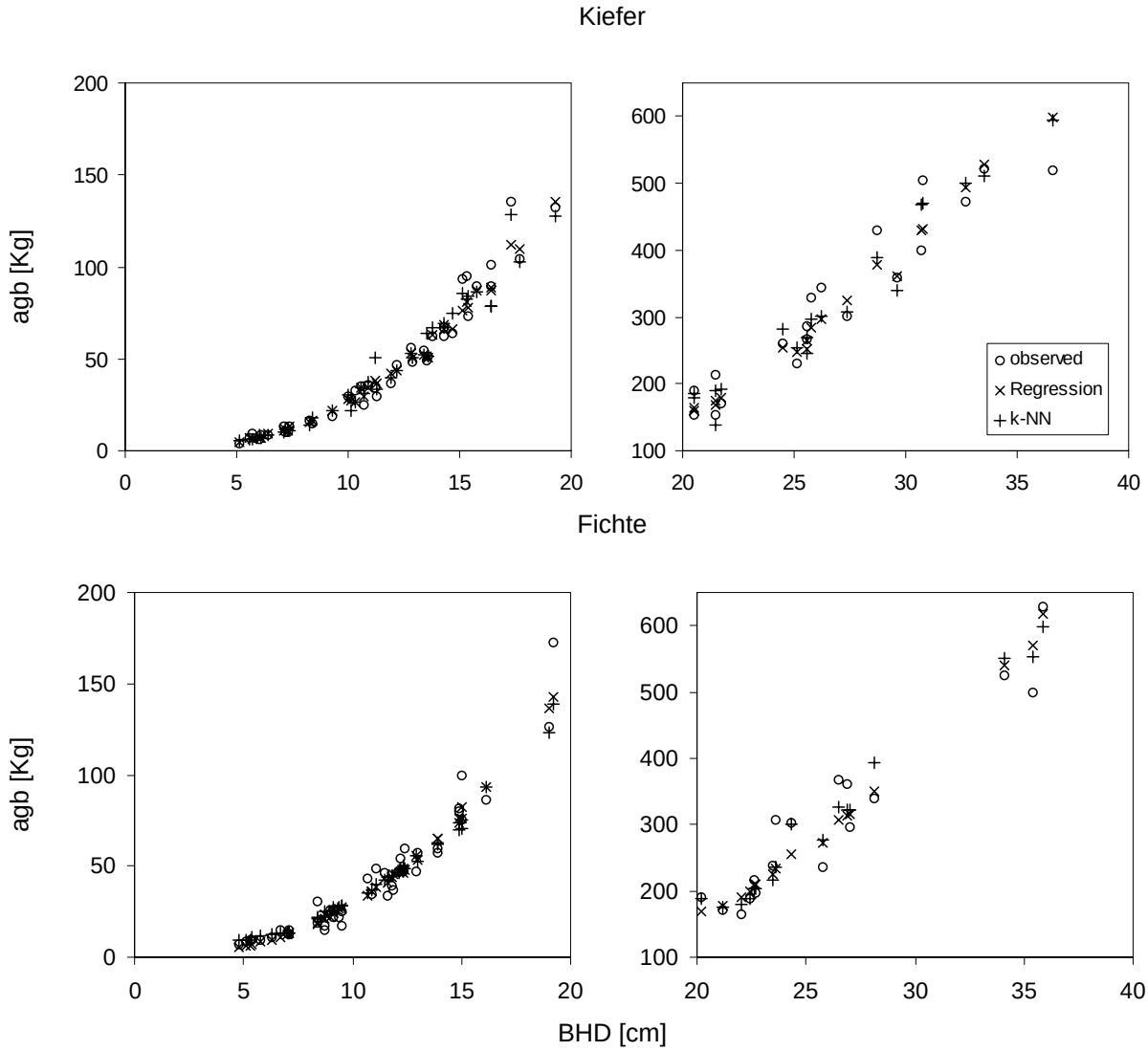


Abbildung 3-15. Beobachtete und geschätzte oberirdische Gesamtbiomasse (*agb*) beider Modellierungsansätze (einfaches lineares Regressionsmodell und *k*-NN Schätzung) über dem BHD. Aus Darstellungsgründen ist die Skalierung der Biomassewerte für kleine und große Durchmesser unterschiedlich gewählt.

Aus den in Abbildung 3-15 unten dargestellten Schätzwerten wird unter Anderem die Auswirkung der Verwendung einer mit Hilfe der Kernel-Distanz festgelegten Bandbreite deutlich. Während die Verwendung einer relativ großen Nachbarschaft im mittleren Durchmesserbereich aufgrund des ausgleichenden Effektes der Mittelwertsbildung zu einer starken Glättung führt, sind im oberen Bereich der Durchmesserskala aufgrund einer geringeren Anzahl berücksichtigter Nachbarn größere Sprünge in den Schätzwerten zu beobachten. Hierdurch wird die Varianz der Datengrundlage in Abhängigkeit

der Verteilung der Trainingsdaten in unterschiedlichem Maß berücksichtigt. Abbildung 3-16 zeigt den Verlauf der verschiedenen Schätzungen der Fichten über dem BHD.

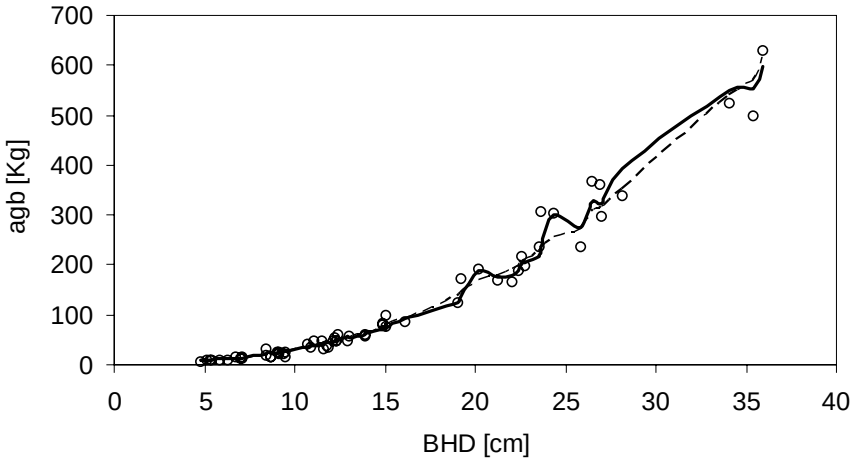


Abbildung 3-16. Verlauf der Regressionsfunktion (gestrichelt) und der k-NN Schätzungen (durchgezogene Linie) über dem BHD am Beispiel der n=60 Testdatensätze für Fichte (hier schematisch als kontinuierlicher Verlauf dargestellt).

Hierbei ist deutlich zu erkennen, dass die k-NN Schätzungen, da sie lokale Approximationen darstellen und nicht an den vordefinierten Verlauf einer Modellannahme gebunden sind, eine höhere Varianz aufweisen als die aus dem Regressionsansatz abgeleiteten Vorhersagen. Eine zusätzliche Untersuchung für einzelne Kompartimente, die typischerweise eine sehr viel höhere Varianz aufweisen würde als die Gesamtmasse eines Baumes, findet sich im Anhang (Seite 143).

3.3.2 Teilauswertung II

Neben Daten von Kiefern und Fichten enthält die hier aufgebaute Datenbank einen umfangreichen Buchendatensatz (n= 221). Die Daten stammen, ähnlich wie die der zuvor ausgewerteten Fichten, von verschiedenen Versuchsflächen, die hauptsächlich verschiedene Standorte in Deutschland abdecken (PELLINEN, 1986; GROTE et al., 2003; JOOSTEN et al., 2004). Ein kleinerer Teil der Daten stammt aus der Tschechischen Republik sowie aus Spanien (SANTA REGINA und TARAZONA, 2001; CIENCIALA et al.,

2004). Da für die vorliegenden Daten nur in einigen Fällen die Blattbiomasse bekannt ist, wird in dieser Auswertung lediglich die oberirdische holzige Biomasse (*agwb*) betrachtet. Wie in der ersten Teilauswertung werden zur besseren Vergleichbarkeit mit den verwendeten Referenzmodellen zunächst lediglich der BHD und die Baumhöhe als unabhängige Variablen für die Regression bzw. als Designattribute für die Suche nach den nächsten Nachbarn verwendet. Der Datensatz wurde hierzu zufallsbasiert in zwei Unterdatensätze aufgeteilt. Der zur Modellierung und Anpassung des *k*-NN Algorithmus verwendete (*modelling* -) Datensatz umfasst 161 Bäume. Der Umfang des zur Evaluation der Schätzergebnisse verwendete Referenzdatensatz (*test*) enthält 60 Bäume, die somit von der Regressionsanalyse sowie den Trainingsdaten ausgeschlossen wurden. Abbildung 3-17 zeigt das Ergebnis der zufälligen Aufteilung des Datensatzes.

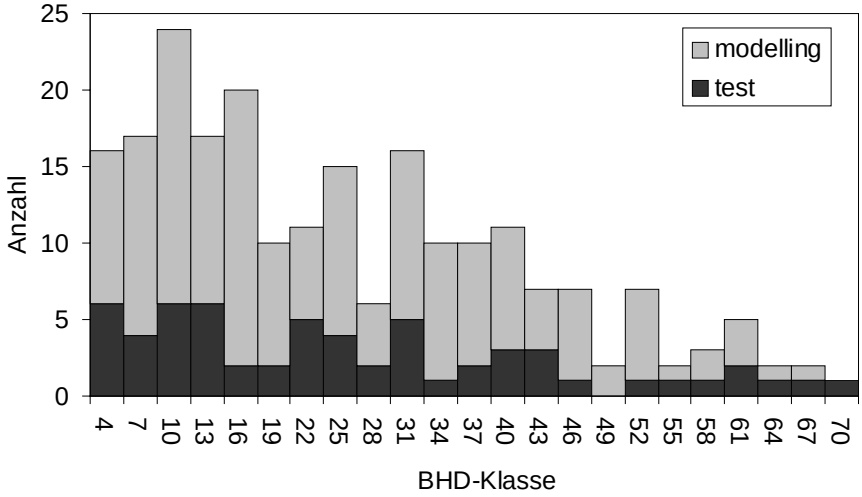


Abbildung 3-17. Histogramm der Durchmesserklassen der verwendeten Buchendaten (n=221) und deren Aufteilung auf den zur Modellierung verwendeten *modelling*-Datensatz (n=161) bzw. *test*-Datensatz (n=60).

Hierbei ist deutlich zu erkennen, dass der *test*-Datensatz in diesem Fall auch Extremwerte der Durchmesserverteilung enthält. Wie bereits unter 2.5.5 dargelegt, stellt das besonders für die *k*-NN Schätzung ein Problem dar, da zur Schätzung der Zielgröße der durchmesserstärksten Bäume nur kleinere Trainingsinstanzen zur Verfügung stehen und somit keine Extrapolation der Daten möglich ist. Die iterative Herleitung einer geeigneten Parametereinstellung für die Distanz- sowie Gewichtungsfunktion führt auf Grundlage des vorliegenden Datensatzes zu den folgenden Einstellungen: Es wurde eine euklidische Distanzmetrik (c=2) gewählt, wobei die Gewichtung zwischen den

Variablen BHD und Höhe mit w_{bhd}=0,7 und w_h =0,3 festgelegt wurde. Wie aus Abbildung 3-18 deutlich wird, kann das Ergebnis der kompletten Kreuzvalidierung des Datensatzes durch die Wahl eines höheren Gewichtungsparameters (t) verbessert werden (siehe hierzu 2.5.3). Hierdurch wird der Einfluss der zu jedem Abfragepunkt gefundenen Nachbarn mit zunehmender Distanz abgeschwächt. Diese Vorgehensweise stellte sich besonders im Fall unsymmetrischer Nachbarschaften als vorteilhaft heraus.

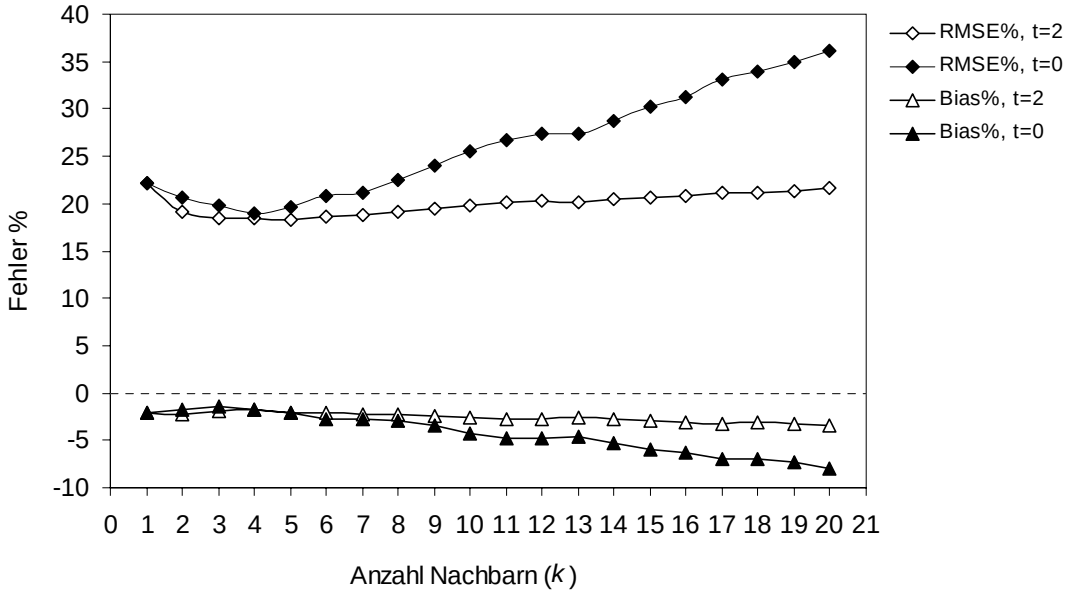


Abbildung 3-18. Entwicklung des RMSE % und des Bias % bei unterschiedlicher Größe der Nachbarschaft (k) und unterschiedlichem Gewichtungsparameter (t=0; t=2) für den gegebenen *modelling*- Datensatz.

Für eine festgelegte Anzahl von Nachbarn ergibt die Fehleranalyse innerhalb der Trainingsdaten in diesem Fall ein lokales Minimum bei $k = 5$, wobei der Gewichtungsparameter $t = 2$ gesetzt wurde. Der berechnete RMSE% liegt hier bei 18,26 mit einem relativen Bias von -1,73 % (im Folgenden wird diese Schätzung mit k-NN 1 bezeichnet). Da davon auszugehen ist, dass die vorliegende Verteilung der Durchmesserklassen im *test*-Datensatz besonders für die größten Bäume einen hohen Fehler bewirken wird, wurde zusätzlich eine feste Kernel-Distanz, bzw. eine adaptive Größe der Nachbarschaft überprüft (im Folgenden als k-NN 2 bezeichnet). Hierzu wurden Trainingsinstanzen bis zu einer normierten Distanz von 0,05 berücksichtigt. Die resul-

tierende Anzahl der verwendeten Nachbarn über dem BHD ist in Abbildung 3-19 dargestellt.

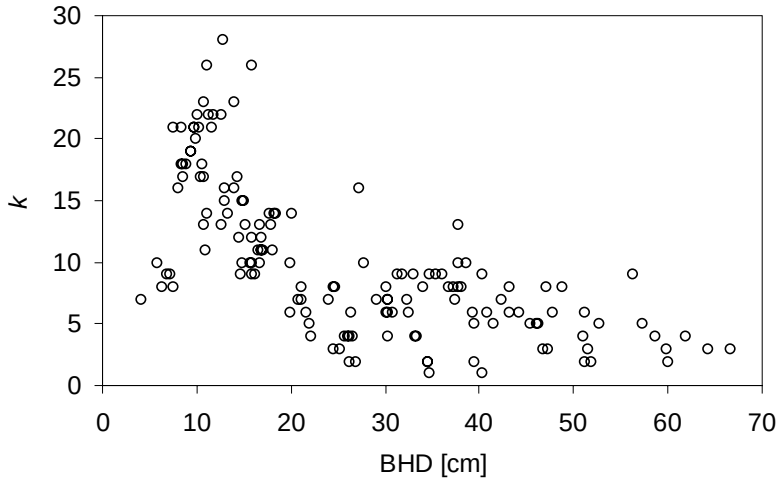


Abbildung 3-19. Unterschiedliche Größe der berücksichtigten Nachbarschaft (k) durch Verwendung einer festen Bandbreite (Kernel-Distanz = 0,05) über dem BHD.

Der RMSE% kann durch die Verwendung der festen Bandbreite geringfügig auf 18 % gesenkt werden. Als Referenzmodelle wird in diesem Fall das einfache lineare Modell mit der Eingangsgröße BHD bzw. BHD und Baumhöhe verwendet (Modell 1 und 2 in 3.2) verwendet. Das einfache allometrische Modell mit lediglich dem BHD als unabhängiger Variablen stellt zwar keine wirkliche Referenz dar, da im k-NN Ansatz die Baumhöhe als zusätzliche Eingangsgröße verwendet wird, es soll hier aber der Vollständigkeit halber trotzdem dargestellt werden. Die geschätzten Regressionskoeffizienten für beide Modellformulierungen mit den jeweiligen Standardfehlern sind in Tabelle 3-7 aufgeführt.

Tabelle 3-7. Geschätzte Regressionskoeffizienten und deren Standardfehler für die als Referenz verwendeten Modellformulierungen (Modell 1 und 2 in 3.2).

Modell	Koeffizient	Schätzung	Std. Fehler	t-Wert
1	α	-2,261	0,071	-31,67
	β	2,508	0,023	107,84
2	α	-3,137	0,137	-22,76
	β	2,105	0,060	34,96
	χ	0,704	0,099	7,12

Das Bestimmtheitsmaß beider Regressionsfunktionen ist mit R^2=0,986 für das Modell 1 und R^2=0,989 für das Modell 2 sehr hoch. Die für alle vier Schätzungen berechneten Fehlermaße sind in Tabelle 3-8 aufgeführt.

Tabelle 3-8. Auf Grundlage des *test*-Datensatzes (n=60) berechnete Fehlermaße für die zwei verwendeten Regressionsmodelle und die k-NN Schätzungen.

Fehlermaß	Modell 1	Modell 2	k-NN 1	k-NN 2
ME	0,02	52,21	79,32	72,82
MAPE	16,34	14,28	199,51	141,41
RMSE	144,81	162,90	257,89	241,40
MSE	20971	26536	66508	58276
Bias-Anteil	0,000	0,103	0,095	0,091
Var-Anteil	0,003	0,439	0,397	0,365
Kov-Anteil	0,997	0,458	0,508	0,544
Korrelation	0,991	0,994	0,982	0,984
RMSP	28,03	19,03	1333,18	903,51
RMSE%	19,46	23,54	38,78	35,95

Aus obiger Tabelle wird zunächst deutlich, dass die für das Modell 2 zugrunde gelegte Modellformulierung zwar in Bezug auf einige Fehlermaße eine Verbesserung gegenüber dem Modell 1 aufweist, andere Gütemaße jedoch schlechtere Ergebnisse aufweisen. So weist das Modell 1 trotz der alleinigen Einbeziehung des BHDs einen geringeren mittleren Fehler und RMSE% auf. Weiterhin deutet der hohe Kovarianz-Anteil des Fehlers darauf hin, dass das Modell 1 besser in der Lage ist, die gegebene Datengrundlage abzubilden.

In Bezug auf die beiden k-NN Schätzungen ist festzustellen, dass ausnahmslos alle berechneten Fehlermaße auf eine schlechtere Prognosegüte als die der Regressionsmodelle hinweisen, wobei der k-NN Ansatz mit fester Bandbreite (Kernel-Methode) geringfügig bessere Ergebnisse liefert als bei einer festen Nachbarschaft von $k = 5$. Besonders die hohen Werte des RMSP sowie des MAPE, die am Ausgangsniveau der Daten relativiert sind, deuten auf hohe relative Prognosefehler hin. Um die Ursache dieser Abweichung genauer zu untersuchen, wurden in Abbildung 3-20 die „beobachtete“ oberirdische holzige Biomasse sowie die Prognosen der k-NN (k-NN 1) Schätzung über dem BHD aufgetragen.

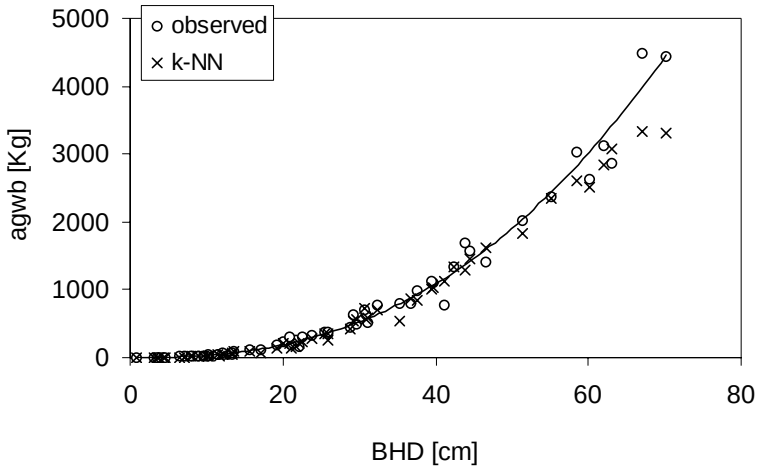


Abbildung 3-20. Beobachtete oberirdische holzige Biomasse (*agwb*), Regressionsfunktion (Modell 1) und *k*-NN Prognosen über dem BHD der Bäume des *test*-Datensatzes.

Hierbei ist deutlich zu erkennen, dass besonders die Biomasse der beiden größten Bäume des *test*-Datensatzes massiv unterschätzt wird, da ihre Dimensionen die aller Trainingsinstanzen übersteigt. Die hierdurch bedingten sehr hohen absoluten Fehlerbeträge von über 1100 Kg fallen bei der Berechnung der verwendeten Gütemaße besonders stark ins Gewicht. Um diesen Fehler zu eliminieren und eine Überprüfung der generellen Prognosegüte der *k*-NN Methode zu ermöglichen, wurden die beiden größten Bäume des *test*-Datensatzes entnommen und nochmals die wichtigsten Fehlermaße berechnet. In diesem Fall wurde somit der Prognosehorizont auf die in den Trainingsdaten vorhandenen Wertebereiche begrenzt.

Tabelle 3-9. Ausgewählte Fehlermaße für die abgeleiteten Regressionsmodelle sowie die *k*-NN Schätzungen auf Grundlage eines reduzierten Datensatzes.

Fehlermaß	Modell 1	Modell 2	k-NN 1	k-NN 2
ME	7,95	31,27	37,50	34,69
RMSE	128,65	114,74	117,19	121,02
RMSE%	20,60	19,65	20,29	20,85
MAPE	16,14	13,81	198,57	140,55

Trotz einer Verbesserung der Prognosegüte im Vergleich zum kompletten Datensatz weisen die *k*-NN Schätzungen in diesem Fall nach wie vor einen höheren absoluten

RMSE sowie MAPE auf. Der berechnete RMSE% liegt für die feste Nachbarschaft (k=5) jedoch nur unwesentlich über dem des Regressionsmodells.

3.4 Zur Einbeziehung weiterer Variablen

Die oben dargestellten Beispiele dienen vor allem dazu, die generelle Funktionsweise und Prognosegüte der k-NN Methode im direkten Vergleich zu Regressionsmodellen zu untersuchen. Hierbei wurden die zur Determinierung der nächsten Nachbarn verwendeten Variablen auf den BHD und die Baumhöhe begrenzt. Wie bereits im Methodenteil beschrieben, zeichnet sich die k-NN Methode als typisches Mustererkennungsverfahren gerade dadurch aus, dass weitere beschreibende Variablen und Attribute ohne genaue Kenntnis über ihre Wirkungsweise in Bezug auf die gesuchte Zielgröße einbezogen werden können. Hierbei können auch gleichzeitig kategoriale und metrisch skalierte Variablen verwendet werden. Während die messbaren Einzelbauminformationen, wie z.B. der BHD und die Baumhöhe, metrische Variablen sind, können aus vorhandenen Meta-Informationen weitere oftmals auch ordinal skalierte Daten abgeleitet werden. Dies könnten z.B. die mittlere Holzdichte einer Baumart, formbeschreibende Parameter oder taxonomische Daten wie z.B. die Gattung oder Familie der betreffenden Baumart sein.

Neben den verfügbaren Informationen auf Einzelbaumebene stehen weiterhin Bestandesinformationen zur Verfügung. Für die meisten in dieser Arbeit zusammengetragenen Einzelbaumdaten sind Informationen über die geografische Lage der Untersuchungsfläche (geografische Koordinaten) sowie Daten über die Höhe üNN, die jährliche Niederschlagsmenge sowie eine Klassifikation der Böden vorhanden.

3.4.1 Teilauswertung III

Um die Möglichkeit der Einbeziehung weiterer Variablen zu überprüfen, soll für diese Teilauswertung auf einen größeren Datenbestand zurückgegriffen werden. Der verwendete Datensatz enthält alle verfügbaren Einzelbaumdaten der europäischen Koniferen (Fichten, Kiefern und Lärchen) und umfasst 858 Einzelbäume. Mit Hilfe einer

kompletten Kreuzvalidierung dieses Datensatzes soll vor allem überprüft werden, inwieweit die *k*-NN Methode in der Lage ist, Schätzungen auch über verschiedene Baumarten abzuleiten. Als zusätzliche Variable wird hierzu die mittlere Holzdichte der betreffenden Baumart herangezogen, die als Diskriminanzvariable einen gewissen Trennungscharakter zwischen den Baumarten aufweist. Hierbei ist zu bedenken, dass die hier verwendeten Kennzahlen keine einzelbaumspezifischen Daten darstellen. Die mittlere Holzdichte in einem Baumindividuum ist stark von der Jahrringbreite, also somit den herrschenden Wuchsbedingungen in Bezug auf Klima, Boden und der gegebenen Konkurrenzsituation abhängig (NIEMZ, 1993). Die hier verwendeten Holzdichten sind daher nur bedingt geeignet, um als unabhängige Variable im Rahmen einer Regressionsanalyse verwendet zu werden. Hinzu kommt, dass die mittlere Dichte eines einzelnen Baumes stark von den in der Zeit wechselnden Anteilen der einzelnen Biomassekompartimente abhängig ist. Da solche Daten nicht für alle Baumarten vorhanden sind, wird hier auf Dichtekennzahlen aus dem Bereich der Materialkunde zurückgegriffen (SACHSSE, 1984). Neben der mittleren Holzdichte für eine bestimmte Baumart könnte zusätzlich auch der Elastizitätsmodul als weiteres Attribut verwendet werden. Der Elastizitätsmodul (E) als Materialkennwert aus der Werkstoffanalyse ist der Kehrwert der Dehnungszahl. Er entspricht der Zugkraft die nötig ist, um einen Stab mit definiertem Querschnitt elastisch in seiner Länge zu verdoppeln (ausgedrückt in Newton/mm^2).

Da die maximale Höhe und Verteilung der Biomasse eines Baumes zum Teil von der Stabilität des Stammes beeinflusst wird, kann diese Kennzahl eventuell als weitere Suchvariable verwendet werden. Die verwendeten Kennzahlen sind in Tabelle 3-10 aufgeführt.

Tabelle 3-10. Mittlere Holzdichte und Elastizitätsmodul der verwendeten Baumarten.

Baumart	Darrdichte [g/cm^3]	E-Modul [N/mm^2]
Fichte	0,43	107873
Kiefer	0,49	117000
Lärche	0,55	135331

Für die gegebenen Eingangsgrößen wurden, auf Grundlage von 20 Iterationen der kompletten Kreuzvalidierung, angepasste Gewichtungsverhältnisse hergeleitet. Wie

schon unter 2.5.2 dargelegt, führt die Erhöhung der Anzahl der berücksichtigten Variablen zu einer enorm großen Anzahl möglicher Kombinationen an Gewichtungsfaktoren für die einzelnen Dimensionen. Zur Bestimmung einer optimalen Kombination der Variablengewichtung sowie der Parameter der Distanzgewichtungsfunktion ist daher die Verwendung eines Optimierungsalgorithmus angebracht. Andererseits hat sich im Rahmen dieser Analyse herausgestellt, dass marginale Veränderungen der Gewichtungsfaktoren nur einen relativ geringen Einfluss auf das Fehlerkriterium RMSE haben. Vielmehr geht es darum, die Bedeutung der verwendeten Variablen auf die zu schätzende Zielgröße untereinander zu relativieren. Dies kann bereits nach einer relativ geringen Anzahl von Iterationen erreicht werden. Abbildung 3-21 zeigt den Verlauf des Fehlerkriteriums RMSE% sowie des Bias über der unterschiedlichen Größe der berücksichtigten Nachbarschaft. Die Zielgröße dieser Untersuchung ist die oberirdische holzige Biomasse (*agwb*).

Die Gewichtungsfaktoren für die berücksichtigten Variablen BHD (w_{bhd}), Höhe (w_h) und Holzdichte (w_d) sind dabei wie folgt gewählt: w_{bhd}=0,7; w_h=0,2 und w_d=0,1. Die Minkowski Konstante ist auf 2 festgelegt (euklidische Distanzmetrik). In diesem Beispiel wurde keine Distanzgewichtung durchgeführt, bzw. der Gewichtungsparameter t=0 gesetzt. Für die so gewählten Parametereinstellungen liegt der geringste RMSE der Kreuzvalidierung mit 26,7% bei einer Anzahl von 4 Nachbarn.

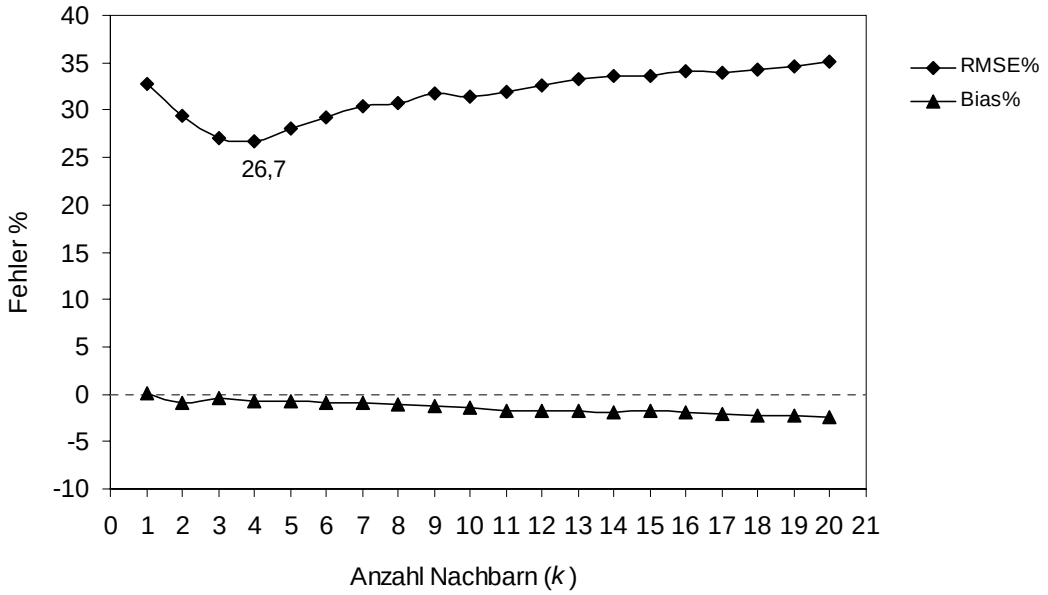
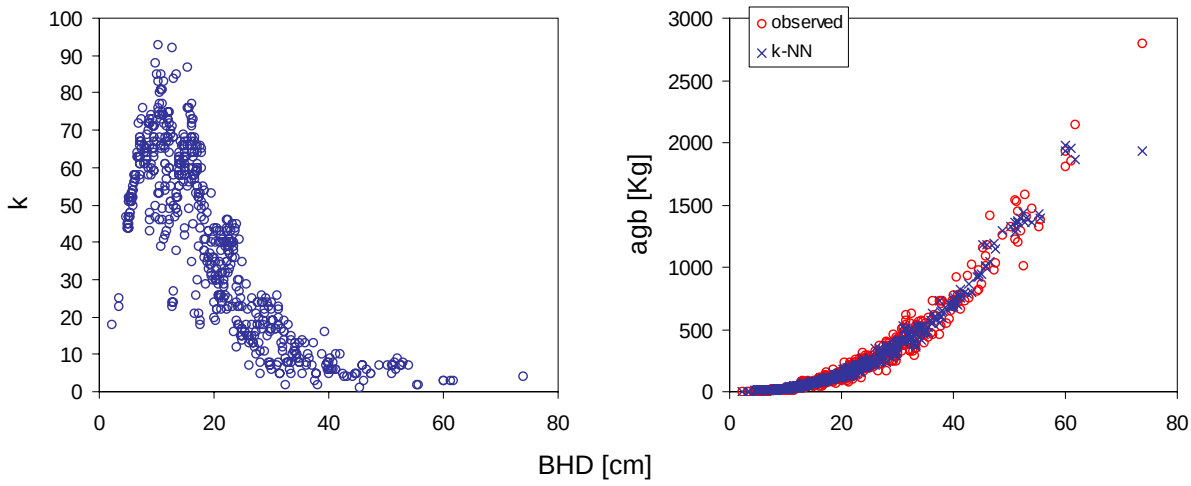


Abbildung 3-21. Entwicklung des RMSE% sowie des BIAS% über einer unterschiedlichen Anzahl berücksichtigter Nachbarn (k) für alle europäischen Nadelbäume des Datenbestandes (n=858).

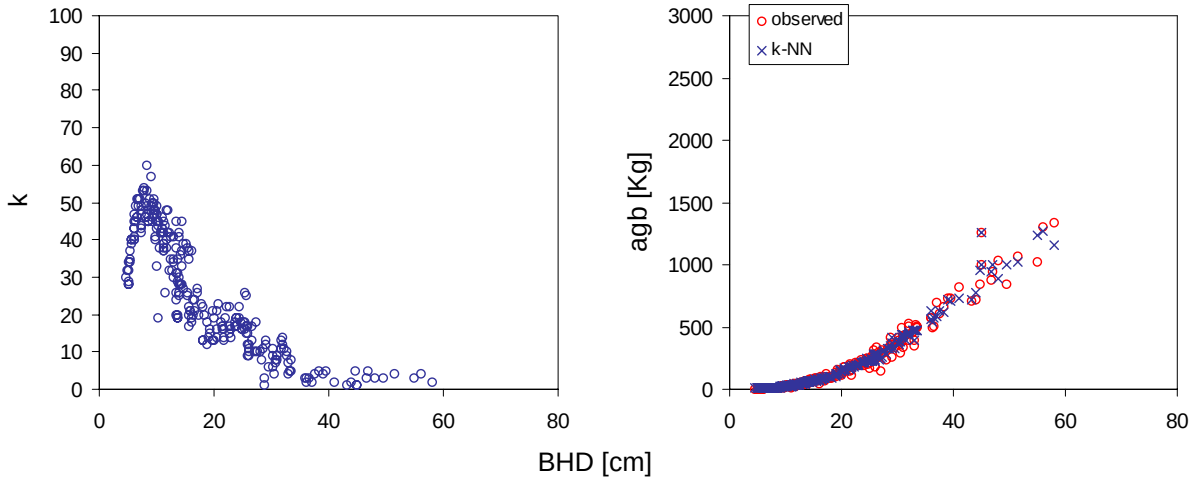
Ähnlich wie in der Teilauswertung III wurde als Alternative zu einer festgelegten Anzahl von Nachbarn eine feste Bandbreite überprüft. Hierbei konnte durch die Wahl einer Kernel-Distanz von 0,03 der RMSE% weiter auf 26% gesenkt werden.

Anders als beispielsweise in Teilauswertung I, in der verschiedene Prognoseansätze für die Baumarten Fichte und Kiefer angepasst wurden, wird in diesem Fall ein generalisierter Ansatz, der für alle beteiligten Baumarten gilt, abgeleitet. Abbildung 3-22 zeigt die „beobachtete“ und geschätzte Biomasse über dem BHD getrennt nach den Baumarten Fichte, Kiefer und Lärche. Weiterhin sind für jede Baumart die Anzahlen der verwendeten Nachbarn über dem BHD dargestellt.

Fichte (n=578):



Kiefer (n=270):



fortgesetzt

Lärche (n=10):

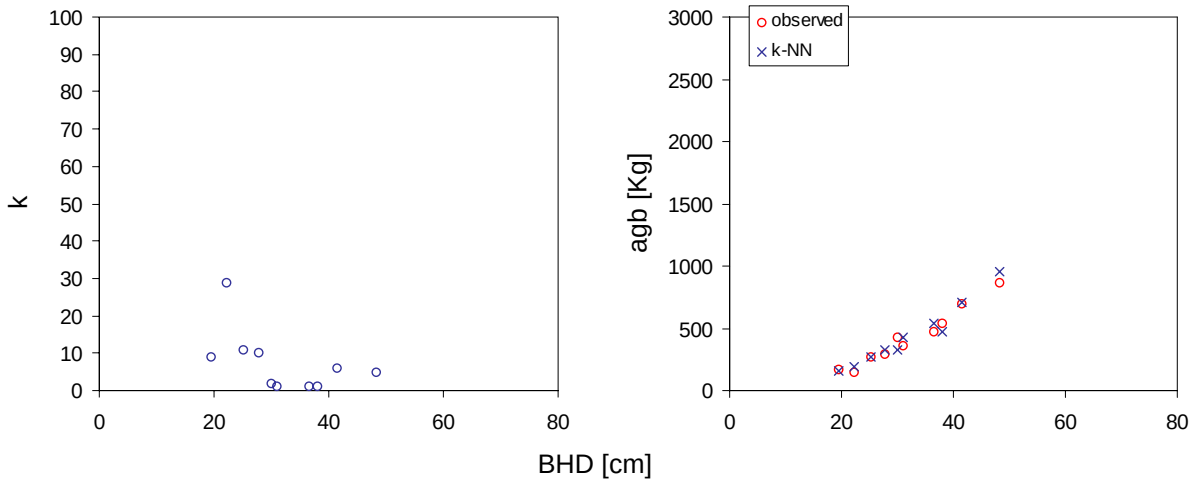


Abbildung 3-22. Anzahl berücksichtigter Nachbarn (k) (links) und beobachtete bzw. geschätzte oberirdische holzige Biomasse ($agwb$) über dem BHD (rechts).

Wie aus der Einzelbetrachtung der verschiedenen Baumarten deutlich wird, scheint der generalisierte Ansatz in diesem Fall relativ gut in der Lage zu sein, die Beobachtungswerte abzubilden. Lediglich bei den verwendeten Fichtendaten ist, ähnlich wie in der ersten Teilauswertung, eine beträchtliche Unterschätzung des größten Baumes zu beobachten, da dieser auch in diesem artübergreifenden Datensatz das Extrem der Durchmesserverteilung darstellt. Bei Kiefern und Lärchen bleibt dieser Effekt in diesem Fall aus, da größere Bäume anderer Baumarten zur Schätzung herangezogen werden können. So wird z.B. die Biomasse der größten Lärche aufgrund von 5 Nachbarn geschätzt, die alle Kiefern sind, da sich diese Baumarten in Bezug auf ihre Holzdichte ähnlicher sind als Lärchen und Fichten. Eine detailliertere Darstellung der Durchmesserbereiche der verwendeten Baumarten findet sich in Abbildung 9-2 auf Seite 138.

Neben der Holzdichte kann als weitere Variable auch das Alter der Bäume in die Schätzung einbezogen werden. Hierbei reduziert sich jedoch die Anzahl der berücksichtigten Trainingsdaten im vorhandenen Datensatz auf 327 Bäume, da nur für diese das Einzelbaumalter bekannt ist. Eine Minimierung des RMSE% führt hierbei zu einer angepassten Anzahl von 8 Nachbarn (siehe auch den Anhang auf Seite 146).

3.4.2 Teilauswertung IV

Für diese Teilauswertung wurde ein Beispieldatensatz verwendet, der in KETTERINGS et al. (2001) veröffentlicht ist (siehe Anhang, Seite 147). Die Datengrundlage stammt aus einer destruktiven Untersuchung, die in einem Sekundärwald auf Sumatra (Sepunggur) durchgeführt wurde. Hierbei wurden insgesamt 29 Bäume verschiedener Arten untersucht. Da nur für 25 dieser Bäume die Baumart bzw. die Holzdichte bekannt ist, wurden hier nur diese verwendet. Aus den gegebenen Kompartimenten wurde die oberirdische holzige Biomasse berechnet und der BHD, die Baumhöhe sowie die berichteten Holzdichten der betreffenden Arten als Eingangsgröße für eine *k*-NN Schätzung verwendet.

Das Ziel dieser Teiluntersuchung besteht darin zu evaluieren, inwieweit der verwendete Ansatz in der Lage ist, mit Hilfe der vorhandenen Datengrundlagen europäischer und nord-amerikanischer Einzelbäume, gänzlich andere und in sofern „unbekannte“ Bäume in Bezug auf ihre Biomasse einzuschätzen. Im Vergleich zu den vorhergehenden Auswertungen hat diese Auswertung einen eher experimentellen Charakter. KETTERINGS et al. (2001) schlagen als angepasstes Regressionsmodell ein einfaches allometrisches Modell auf Grundlage des BHD vor. Da im vorliegenden Fall 4 Bäume aufgrund fehlender Information über die Holzdichte eliminiert wurden und die Zielgröße in diesem Fall lediglich die holzige oberirdische Biomasse ist, wurden die Parameter des Modells hierzu neu geschätzt. Das resultierende Modell ist 0,0459*BHD^2,6692 mit einem R^2 von 0,96.

Die Kreuzvalidierung aller in der Datengrundlage vorhandenen Laubbäume (n=463) führt nach ca. 20 Iterationen zu folgenden Gewichtungsverhältnissen: w_{bhd}=0,6, w_h=0,3, w_d=0,1. In diesem Fall lieferte eine L1-Norm (Manhattan Distanz, c=1) bessere Ergebnisse als eine euklidische Distanzmetrik. Der Gewichtungsparameter t wurde auf 0 gesetzt und Nachbarn wurden bis zu einer Kernel-Distanz von 0,05 berücksichtigt. Die Ergebnisse der *k*-NN Schätzung und des verwendeten Regressionsmodells sind in Abbildung 3-23 über dem BHD dargestellt.

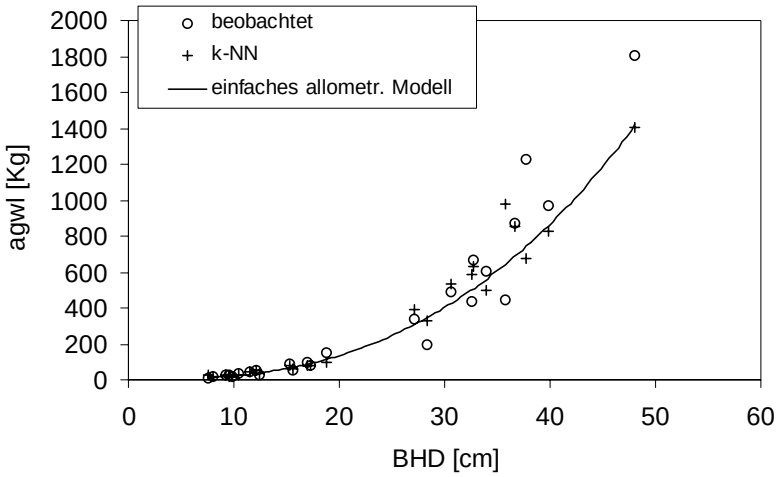


Abbildung 3-23. Ergebnisse der *k*-NN Schätzung sowie des verwendeten Regressionsmodells und beobachtete oberirdische holzige Biomasse für den gegebenen Datensatz.

Der aus der Gegenüberstellung der beobachteten und geschätzten Werte berechnete relative RMSE liegt im Fall des Regressionsmodells bei 38,3%, wobei der mittlere Fehler -22,43 ist. Der RMSE% der *k*-NN Schätzung liegt mit 49,29% zwar etwas höher, jedoch beträgt hier der mittlere Fehler lediglich 8,32. Dies ist darauf zurückzuführen, dass zwar einzelne Bäume einen hohen Fehlerbetrag aufweisen, die geschätzten Datenpunkte im Mittel die Varianz der Biomasse jedoch angemessen beschreiben. Da Biomasseschätzungen zumeist auf eine Flächeneinheit bezogen werden, um Aussagen für räumlich abgegrenzte Behandlungseinheiten oder sonstige Flächen abzuleiten, ist hierbei eventuell der mittlere Fehler der Schätzung höher zu bewerten als die Fehleinschätzung auf Einzelbaumebene.

4 Diskussion

Die im ersten Teil erarbeiteten Ergebnisse der vorliegenden Arbeit haben gezeigt, dass bestehende empirische, meist allometrische Biomassemodelle in Hinblick auf ihre Herleitung und Anwendung durchaus kritisch zu betrachten sind. Sie lassen sich aufgrund der unterschiedlichen Definition der Eingangsgrößen sowie der Tatsache, dass empirische Modelle aufgrund destruktiver Zeitreihenuntersuchungen an unterschiedlichen Organismen durchgeführt werden müssen, nur bedingt mit Vorhersagen von Prozessmodellen vergleichen. Die Überprüfung der Sensitivität der Allometriekoeffizienten an dem unter 3.1 ausgewerteten zusammengefassten Datensatz von Fichten hat gezeigt, dass sich ein Großteil der Varianz der oberirdischen Biomasse durch die Variablen BHD und Höhe erklären lässt. Bedenkt man, dass die hier zusammengetragenen Einzeldatensätze über einen Zeitraum von mehr als 35 Jahren auf sehr unterschiedlichen Standorten in Europa unter Verwendung verschiedener Aufnahmeverfahren erhoben wurden, scheint eine Generalisierung von Biomassefunktionen durchaus möglich zu sein.

4.1.1 Sensitivität der Allometriekoeffizienten

Einige Autoren (z.B. ZIANIS und MENCUCCINI, 2004; KETTERINGS et al., 2001) sehen in der bekannten Autokorrelation der Koeffizienten des allometrischen Modells einen Ansatzpunkt zur Verallgemeinerung von Biomassefunktionen. Die vorliegende Untersuchung hat gezeigt, dass sich nach Aufteilung der Datenbasis die starke Korrelation zwischen den Parametern a und b auch in der vorhandenen Datengrundlage nachweisen lässt. Die Aufteilung der einzelnen Datensätze in möglichst kleine Subsets wurde in diesem Fall auf Grundlage der einzelnen Untersuchungsflächen durchgeführt. D.h., die Datengrundlagen einzelner Studien, die typischerweise aus mehreren Beständen stammen, wurden nach den Bestandesnummern aufgeteilt, um möglichst viele (und möglichst unabhängige) Schätzungen der Regressionskoeffizienten des zugrunde gelegten einfachen allometrischen Modells durchführen zu können. Als Folge dieser Aufteilung, sind die Einzeldatensätze sehr klein. Bei der Unterteilung wurden Einzeldatensätze bis zu einer Mindestgröße von n=5 ausgewählt. Verwendet man diese

kleinen Datensätze als Grundlage für eine Regressionsanalyse, ist eine relativ hohe Streuung der geschätzten Regressionskoeffizienten daher schon aus diesem Grund zu erwarten. Ebenso ist eine Autokorrelation der Parameter zu erwarten, wenn man die Bedeutung dieser Koeffizienten in einer empirischen Datenanalyse bedenkt.

Die Autokorrelation zwischen den Parametern würde es im Rahmen einer Regressionsanalyse nahe legen, einen der Koeffizienten aus der Modellformulierung zu eliminieren. Gleichzeitig lässt der straffe Zusammenhang darauf schließen, dass trotz der unterschiedlichen Datengrundlagen, zumindest im vorliegenden Fall, der sich auf eine Baumart beschränkt, Zusammenhänge gefunden werden können, die eventuell zur Verallgemeinerung von Biomassefunktionen genutzt werden können.

Wie bereits unter 2.1.2 dargestellt, wird die Grundgleichung der Allometrie bevorzugt verwendet, da sie eine biologische Erklärung der Wirkungszusammenhänge ermöglicht. In verschiedenen empirischen Studien an unterschiedlichen Objekten konnte die hohe Korrelation zwischen den Allometriekoeffizienten a und b festgestellt werden. Dieser Autokorrelation wird allerdings gemeinhin keine biologische Implikation zugesprochen. Vielmehr geht man davon aus, dass dieser Zusammenhang aus dem allometrischen Modell bzw. der vorliegenden Datengrundlage selbst entsteht und somit eine Eigenart der erweiterten Potenzfunktion ist, die die allometrische Grundgleichung beschreibt.

Da die Regressionskoeffizienten $ln(a)$ und b im vorliegenden Fall Schätzungen für die Parameter einer Gradengleichung sind, wäre ein funktionaler Zusammenhang dieser Parameter nur dadurch zu erklären, dass sich alle Regressionsgeraden der Form $ln(agb)=b*ln(D)+ln(a)$ in genau einem Punkt schneiden. Nur so ist es möglich, aus der Steigung der Geraden (b) auf den Achsenabschnitt ($ln(a)$) zu schließen und umgekehrt. Dieser Zusammenhang würde unter der Annahme, die auf Ebene der Teildatensätze abgeleiteten Regressionen seien in der Lage die Beziehung zwischen Durchmesser und Biomasse vollkommen zu beschreiben, zu der Aussage führen, dass es einen Durchmesser geben muss, an dem alle untersuchten Bäume theoretisch die gleiche Biomasse hätten. Dieser fiktive Schnittpunkt der Regressionsgeraden muss sich dabei nicht notwendigerweise im Wertebereich der jeweiligen Datengrundlage befinden, sondern ergibt sich für einige Fälle eventuell nur durch Extrapolation.

Die Implikation dieser Aussage alleine scheint daher keine biologische Begründung zu ermöglichen, da neben dem BHD weitere Einflussgrößen, wie z.B. die Baumhöhe, zu einer Ausdifferenzierung der Biomasse zwischen verschiedenen Teildatensätzen führen. Der vorzufindende Zusammenhang zwischen den Parametern ist daher eher auf die

hohe Korrelation zwischen der Zielgröße und der unabhängigen Variablen im Gesamtdatensatz und der daraus resultierenden relativ schmale Punktwolke zurückzuführen, die dazu führt, dass alle angepassten Regressionsgeraden einen relativ engen Wertebereich durchlaufen.

Dass die beobachtete Variabilität der Allometriekoeffizienten in diesem Fall jedoch nur ein Zufallseffekt darstellt, scheint relativ unwahrscheinlich. Es muss eher vermutet werden, dass der vorzufindende Zusammenhang auf ein bestehendes Gleichgewicht zwischen der aktuellen Masse und dem Massenzuwachs hindeutet. In den vorliegenden Datengrundlagen konnte kein hinreichender Erklärungsansatz für die Varianz der Parameter gefunden werden. Die Regression zwischen der Allometriekonstanten b der Biomassefunktion und dem Exponenten b' der Durchmesser-Höhen Beziehung wies im Fall der vorliegenden Teildatensätze einen relativ schwachen Zusammenhang auf.

Bemerkenswert erscheint die Auswirkung der Umrechnung des BHDs zu einem Durchmesser in relativer Stammhöhe. Basierend auf der Überlegung, dass diese Messgröße, da es sich um einen funktionalen Durchmesser handelt, besser mit dem theoretischen Konzept allometrischer Beziehungen im Einklang steht, wurden alle Regressionsanalysen mit dieser modifizierten Eingangsgröße wiederholt. Aufgrund dieser Umrechnung des BHDs, entsteht wie in Abbildung 3-2 auf Seite 65 zu sehen ist, eine völlig veränderte Durchmesserverteilung. Bei Bäumen, die eine Höhe von weniger als 13 m aufweisen, führt diese Umrechnung zu einer Verringerung des Durchmessers, wohingegen der Durchmesser größerer Bäume entsprechend erhöht wird.

Hierbei zeigt sich, dass die für den Gesamtdatensatz geschätzte Allometriekonstante näher an der Vorhersage der beschriebenen Prozessmodelle liegt. Die geringe Abweichung des Koeffizienten $b_{0,1}$ (2,63) zum Erwartungswert (2,667) lässt sich hierbei eventuell dadurch erklären, dass Bäume in ihrer Entwicklung Totäste verlieren und so eine etwas geringere Masse aufweisen als theoretisch angenommen. Inwieweit die Veränderung der geschätzten Koeffizienten lediglich ein Zufallseffekt ist, oder auch auf Grundlage anderer Datensätzen nachvollzogen werden kann, konnte im Rahmen dieser Arbeit nicht überprüft werden. Zu bedenken bleibt, dass durch die Umrechnung der Durch-messer mit Hilfe der Pain-Funktion in diesem Fall ein Fehler entsteht, der in den folgenden Analysen nicht mehr quantifiziert werden kann. Die nahezu lineare Beziehung zwischen BHD und $D_{0,1}$ ist in diesem Fall darauf zurückzuführen, dass diese beiden Größen in den meisten Fällen relativ nahe beieinander liegen. Für Bäume im mittleren Durchmesserbereich befinden sie sich oberhalb des Wurzelanlaufes und somit

in einem relativ gleichmäßig konisch verlaufenden Stammabschnitt, der eventuell auch durch die vereinfachte geometrische Form eines Kegelstumpfes abgebildet werden kann. Ob die Verwendung einer spezifischen Schaftformfunktion angebracht ist, kann daher in Frage gestellt werden. Eventuell kann die Umrechnung des BHD zum relativen Durchmesser in 10% der Stammhöhe auch mit dieser vereinfachten Annahme (siehe hierzu auch den Exkurs in Anhang I) mit ausreichender Genauigkeit erzielt werden.

Die aus der Umrechnung des BHDs zum $D_{0,1}$ resultierenden Differenzen der Eingangsgröße für die durchgeführten Regressionsanalysen lagen in einem Bereich von nur einigen Millimetern bei kleineren Bäumen bis hin zu mehr als 13 cm für Bäume aus dem oberen Bereich des Durchmesserspektrums. Diese Unterschiede sind daher keinesfalls als geringfügig oder unbedeutend in Bezug auf die Konformität der unabhängigen Variable Durchmesser mit den Grundlagen allometrischer Beziehungen anzusehen. Allometrische Modelle werden zur Modellierung von Wachstumsprozessen nicht nur in der Waldwachstumskunde verwendet. Als Begründung für Wahl dieser Modellformulierung wird die lineare Beziehung der logarithmisch transformierten Wertepaare herangezogen. Hierbei ist zu bedenken, dass der ausgleichende Effekt der logarithmischen Transformation dazu führt, dass eine eventuelle Abweichung von der Linearität in empirischen Datengrundlagen nur schwer nachzuweisen ist (siehe hierzu auch Anhang I). Im Zweifelsfall wird erst durch die Rücktransformation der Werte auf ein metrisches Skalenniveau deutlich, dass das verwendete Modell nur unzureichend in der Lage ist, den bestehenden Zusammenhang adäquat abzubilden. Die theoretischen Überlegungen die im Rahmen dieser Auswertung angestellt wurden, führen in ihrer Konsequenz zu der provokanten Folgerung, dass keine allometrischen Beziehungen zwischen dem BHD und irgendeiner anderen Einzelbaumvariablen bestehen können, da der BHD als nicht-funktionale Messgröße eine ungeeignete Eingangsgröße darstellt.

Die in diesem Beispiel erzielte Verbesserung der strukturellen Anpassung des einfachen allometrischen Modells ist in Bezug auf Einzelbäume zwar nicht gravierend (der Erklärungsanteil der Gesamtstreuung konnte von 97% auf 98% erhöht werden), es bleibt jedoch zu bedenken, dass die einzelbaumbasierten Schätzungen im Rahmen der Kohlenstoffbudgetierung auf große Flächen hochgerechnet werden. Auch eine vergleichsweise geringe Verbesserung der Modellanpassung kann aufgrund der Fehlerfortschreibung für die Sicherheit der Biomasseschätzung auf einer Gesamtfläche oder Behandlungseinheit große Auswirkung haben.

4.2 Zur Anwendung der *k*-NN Methode

Die verschiedenen Teilauswertungen, die zur Evaluation des *k*-NN Anstzes durchgeführt wurden zeigen zunächst, dass die Umsetzung der Methode in der dargestellten Form funktionsfähig ist. Die entwickelte Datenbankstruktur der aufgebauten Einzelbaumdatenbank sowie die Umsetzung des hierauf zugreifenden *k*-NN Moduls sind geeignet, um Biomasseschätzungen für Einzelbäume bzw. deren Biomassekompartimente durchzuführen.

Am Beispiel der Teilauswertung I, in der jeweils ein Fichten- und Kieferndatensatz mit Hilfe des *k*-NN Algorithmus und den abgeleiteten Referenzmodellen durchgeführt wurde, ist erkennbar, dass die Schätzungen der *k*-NN Methode in Bezug auf fast alle verwendeten Gütemaße bessere Ergebnisse liefert als die abgeleiteten Regressionsansätze. Obgleich die auf Grundlage der zufällig ausgeschiedenen *modelling*- Datensätze angepassten Regressionsfunktionen mit einem R^2 von 0,986 für Fichte und 0,99 für Kiefer (siehe Anhang III) einen beachtlich hohen Erklärungsanteil haben, sind die Fehler der Schätzungen für die jeweiligen *test*- Datensätze auf Basis der *k*-NN Methode geringer. Die Aufteilung des Datensatzes wurde durchgeführt, um eine Überprüfung beider Ansätze anhand einer an der Modellbildung nicht beteiligten Datengrundlage zu ermöglichen. Dies währe für die *k*-NN Methode auch anhand einer (Leave-One-Out-) Kreuzvalidierung des gesamten Datensatzes möglich gewesen, da ja in diesem Fall anders als bei der Regressionsanalyse keine Modellanpassung stattfindet, sondern für jeden einzelnen Baum eine Schätzung über die *N*-1 verbleibenden Bäume des Datenbestandes abgeleitet wird. Um jedoch eine Vergleichsanalyse auf Grundlage der gleichen Datenbasis durchführen zu können, wurde hier mit denselben Datensätzen gearbeitet.

Die berechneten Fehlermaße RMSE (bzw. RMSE%) sowie der mittlere Fehler deuten darauf hin, dass die *k*-NN Schätzungen in diesem Fall sogar bessere Ergebnisse lieferten, als die angepassten gemischt-linearen Modelle, die einen zusätzlichen zufälligen Effekt beinhalten. Die zusätzliche Zufallskomponente dieses Modells wurde für die vorliegenden Daten auf Grundlage ihrer Herkunft aus einzelnen Aufnahmepunkten der Nationalen Waldinventur geschätzt. Hierbei wird der Interzept der Funktion für jeden Plot einzeln adjustiert. Zur Schätzung eines bestimmten Baumes des *test*- Datensatzes werden auf diese Weise die vorhandenen Daten aus diesem Plot stärker berücksichtigt, indem der Interzept des Modells für jeden Plot einzeln geschätzt wird. Die einzelnen Aufnahmeplots werden hierbei als eine Subpopulation angesehen, in

der aufgrund der räumlichen Nähe der Bäume eine höhere Korrelation der Variablen unterstellt wird als im Gesamtdatensatz. Solche Regressionsansätze sind natürlich nur dann sinnvoll, wenn ein eindeutiger räumlicher Bezug in der Datengrundlage vorliegt.

Die vergleichende Teiluntersuchung II hat gezeigt, dass die *k*-NN Methode jedoch nicht in jedem Fall Vorteile gegenüber Regressionsansätzen hat. Die zufällige Unterteilung der vorhandenen Datengrundlage in einen von der Modellierung ausgeschlossenen *test*-Datensatz führte in diesem Beispiel dazu, dass diese Datengrundlage Extremwerte enthält, die den Wertebereich der verbleibenden Trainingsdaten überschreiten. Solche Situationen sind realitätsnah, da die Spannweite der vorhandenen Datengrundlage, die zur Modellierung herangezogen werden kann, oftmals eingeschränkt ist. Während dies im Fall einer Regressionsfunktion durch die vorhandene Extrapolationsfähigkeit nur eingeschränkt problematisch ist, führt die Anwendung der *k*-NN Methode aufgrund der instanzenbasierten Schätzung zu massiven Fehleinschätzungen im Bereich der Extreme der Werteverteilung. Dementsprechend weisen alle berechneten Gütemaße im Fall des verwendeten Buchen-Datensatzes auf eine geringere Prognosegüte des *k*-NN Algorithmus hin.

Auf diesen methodischen Nachteil des Verfahrens kann auf unterschiedliche Weise eingegangen werden. Eine Möglichkeit besteht darin, den Wertebereich der zu schätzenden Bäume so einzugrenzen, dass er die Spannweite der vorhandenen Designattribute der Trainingsdaten nicht überschreitet. Normalerweise sollte diese Prämisse ebenso für die Anwendung eines Regressionsmodells gewährleistet sein, da eine statistisch abgesicherte Fehlerwahrscheinlichkeit nur innerhalb der Spannweite der Datengrundlage angegeben werden kann. Im vorliegenden Beispiel wurden die Bäume, die größer als alle Trainingsinstanzen waren (dies waren hier zwei Bäume), aus dem *test*-Datensatz eliminiert und anschließend die wichtigsten Fehlermaße der Schätzung nochmals berechnet. Hierdurch konnten der RMSE, der RMSE% sowie der MAPE unter das Fehlerniveau eines einfachen linearen Modells auf Grundlage des BHD abgesenkt werden. Nach wie vor waren sie aber etwas höher als für ein Regressionsmodell das auf dem BHD und der Baumhöhe basiert. Besonders der relativ hohe MAPE zeigte, dass die *k*-NN Schätzungen in diesem Fall im Mittel höhere absolute prozentuale Abweichungen in Bezug auf das Ausgangsniveau der Daten aufweisen. Dass diese Fehler im Fall der *k*-NN Schätzungen bei über 100% liegen, zeigt, dass hier vor allem kleine Bäume, bei denen bereits betragsmäßig geringe Abweichungen zu einem hohen absoluten prozentualen Fehler führen, relativ schlecht eingeschätzt wurden.

Im Rahmen der Teilauswertungen III und VI wurden weitere Variablen als Designattribute in die *k*-NN Schätzung einbezogen. Hierzu wurde zunächst auf Meta-Informationen über die mittlere Holzdichte der Baumarten zurückgegriffen. Die Holzdichte ist ein diskret verteiltes Merkmal, das in der Abstandsberechnung einen gewissen Trennungscharakter zwischen den einzelnen Baumarten aufweist. Die Gewichtungsverhältnisse der einzelnen Variablen, die durch einen iterativen Prozess zur Minimierung des RMSE% in der Teilauswertung III hergeleitet wurden, zeigen, dass die Holzdichte im Vergleich zum BHD und der Baumhöhe relativ gering gewichtet wird (w_d=0,1). Der RMSE% wurde in dieser Auswertung mit Hilfe einer kompletten Kreuzvalidierung der 858 vorhandenen Einzelbäume berechnet. Durch die Wahl einer adaptiven Nachbarschaft unter Verwendung der Kernel-Methode konnte der RMSE% auf 26 gesenkt werden. Wie auch die Gegenüberstellung der „beobachteten“ und geschätzten Werte zeigt, ist die baumartübergreifende Prognose des *k*-NN Algorithmus hier als sehr gut anzusehen.

In einer weiteren Auswertung (Teilauswertung IV) wurde auf einen Datensatz von tropischen Bäumen (KETTERINGS et al., 2001) zurückgegriffen, um zu überprüfen, inwieweit die verwendeten Variablen hinreichend sind, um generalisierte Biomasseschätzungen auf Basis einer völlig unterschiedlichen Datengrundlage abzuleiten. Hierbei wurden Laubbäume aus europäischen und nord-amerikanischen Untersuchungsgebieten als Trainingsdaten verwendet, um die Biomasse der Einzelbäume einer Studie zu schätzen, die in Indonesien durchgeführt wurde. Der für diesen Datensatz berechnete RMSE% der *k*-NN Schätzung liegt hierbei zwar leicht über dem Fehler, der für ein angepasstes Regressionsmodell berechnet wurde, der mittlere Fehler ist hierbei jedoch geringer. Hierbei muss bedacht werden, dass in diesem Fall aufgrund des geringen Umfangs der Datengrundlage der Datensatz nicht aufgeteilt wurde. D.h., die Regressionsanalyse wurde hier nicht wie in Teilauswertung I an einem unabhängigen Datensatz durchgeführt. Gleichzeitig waren die Testdaten jedoch nicht Bestandteil der Trainingsdatenbank auf die der *k*-NN Algorithmus zugreift. Aus diesem Grund sind die vergleichsweise guten Prognosen der *k*-NN Methode in diesem Fall hervorzuheben.

4.2.1 Zur Bestimmung der Größe der Nachbarschaft

Die Bestimmung der Größe der berücksichtigten Nachbarschaft ist ein zentraler Aspekt der k-NN Anwendung. Wie im Methodenteil dargelegt, wird die Entscheidung darüber wie viele Nachbarn für eine Schätzung herangezogen werden sollen von verschiedenen Gesichtspunkten geleitet. Zur Einschätzung der Güte einer Prognose wurden im Rahmen der Auswertung zumeist Gütemaße verwendet, die den Prognosefehler auf Ebene der Einzelbaumschätzungen entweder quadratisch oder einfach quantifizieren und hierbei teilweise das Ausgangsniveau der Daten berücksichtigen. Diese Gütemaße müssen, wie beispielsweise in Teilauswertung IV gezeigt wurde, nicht notwendigerweise zur gleichen Aussage führen. Welches Gütekriterium zur Optimierung des k-NN Algorithmus (und ebenso zur Bewertung eines Regressionsmodells) verwendet wird, ist daher stark von der Zielsetzung der Prognose abhängig. In den vorliegenden Auswertungen ging es zunächst darum, die Einzelbaummasse möglichst korrekt abzubilden. Im Rahmen der Biomassebestimmung ganzer Bestände oder Behandlungseinheiten bzw. zur Ermittlung der Kohlenstoffsenkeneffekte auf regionaler Ebene, ist eventuell der mittlere Fehler einer Schätzung sehr viel wichtiger als die Fehlprognosen für einzelne Bäume. Hierbei hat die k-NN Methode den Vorteil, dass die Schätzungen den Wertebereich der Trainingsdaten nicht verlassen können. Die fehlende Extrapolationsfähigkeit des Verfahrens kann in diesem Fall also auch einen Vorteil darstellen.

Die hier in einigen Teilauswertungen verwendete Modifikation des einfachen k-NN Ansatzes durch die Verwendung der adaptiven Nachbarschaft hat gezeigt, dass das typische Bias- Varianz- Dilemma zwar nicht zufriedenstellend gelöst werden kann, die negativen Effekte jedoch verringert werden konnten. Voraussetzung hierfür ist, dass die im konkreten Fall einer Suchanfrage berechneten Distanzen zu allen Trainingsdaten auf ein festgelegtes Intervall normiert werden. Durch die Festlegung einer maximalen normierten Distanz, bis zu welcher Trainingsdaten für die Schätzung berücksichtigt werden, kann die Größe der Nachbarschaft in Abhängigkeit der Verteilung der Trainingsdaten adaptiv bestimmt werden. Hierdurch werden extrem unsymmetrische Nachbarschaften am Rand der Werteverteilung vermieden.

Wird die Bandbreite der Nachbarschaft durch eine feste Anzahl von Nachbarn vorgegeben, so kann des weiteren anhand der Gewichtung der Nachbarn mit ihrem Abstand der Nachteil einer zu großen Nachbarschaft abgemildert werden.

4.3 Räumliche Komponenten

Im Rahmen der hier dargestellten Ergebnisse wurden Einzelbaumvariablen sowie Bestandes- und Meta-Informationen über die verschiedenen Baumarten verwendet. Zusätzlich zu diesen Informationen können auch räumliche Informationen in den k-NN Ansatz integriert werden. Hierzu können auf einfache Weise die Koordinaten der Untersuchungsgebiete als zusätzliche Indikatorattribute verwendet werden. So können räumlich nähere Trainingsinstanzen bei der Schätzung der Zielgröße stärker gewichtet werden, um auf diese Weise eventuell vorhandene naturräumliche Gegebenheiten bzw. Umweltfaktoren, die durch die berücksichtigten Variablen noch nicht beschrieben werden, einzubeziehen. Leider sind diese Informationen in der vorhandenen Datengrundlage nur für relativ wenige Datensätze bekannt, so dass eine solche Auswertung nicht durchgeführt werden konnte.

Theoretisch kann jedoch angenommen werden, dass z.B. die geografische Breite einen entscheidenden Einfluss auf das Wuchsverhalten der Bäume und auch ihre Biomasse bei einer gegebenen Dimension hat, da schlechtere klimatische Bedingungen bzw. eine kürzere Wachstumsperiode tendenziell zur Ausbildung engerer Jahresringe und somit einer höheren mittleren Holzdichte führen. Auch der klimatische Einflussauf auf die Entwicklung einer spezifischen Wuchsform bzw. Allokation der Biomasse auf die einzelnen Baumkompartimente ist sicherlich nicht unerheblich.

Würde man die räumliche Information in den k-NN Ansatz einbeziehen, könnten auch lokal unterschiedliche Biomasseschätzungen abgeleitet werden. Durch die Integration eines Geo-Informtions Systems, das durch eine geeignete Schnittstelle ohne weiteres an die bestehende Datenbank angebunden werden kann, würden dann eine räumlich explizite Analyse ermöglicht werden.

4.3.1 Generelle Bewertung des Verfahrens

Als eine Methode des induktiven Lernens setzt der k-NN Ansatz nur eingeschränkte Kenntnisse über die genauen Zusammenhänge zwischen den Einflussgrößen, ihren Wechselwirkungen und der Zielvariablen voraus. Die Biomasse eines Baumes kann von sehr vielen Einflussfaktoren bestimmt sein, die gleichzeitig vielen Wechselwirkungen

unterworfen sind. Zwar lassen sich für Daten aus Einzeluntersuchungen oft Modelle mit einem sehr hohen Erklärungsanteil finden, vergleicht man aber Bäume aus verschiedenen Untersuchungen, die unterschiedlichen Umweltbedingungen unterliegen, kann die Ableitung eines allgemeinen Modells aufgrund der Vielzahl an Einflussgrößen sehr kompliziert werden. Die k-NN Methode als Verfahren des maschinellen Lernens bietet hier eine Alternative, die nach dem Grundsatz „die einfachste Lösung ist oft die Beste“ die genaue Wirkungsweise der Einflussgrößen zunächst weitgehend ignoriert und alleine durch die Suche nach ähnlichen Merkmalsträgern eine Schätzung der Zielgröße erlaubt. Die k-NN Methode hat dabei eine Reihe von Vorteilen gegenüber parametrischen Verfahren (HESSENMÖLLER, 2001; WAGACHA, 2003):

- Der k-NN Algorithmus ist in der Lage sehr komplexe Zielfunktionen durch weniger komplexe Approximationen zu modellieren,
- der Algorithmus ist einfach zu implementieren und umzusetzen,
- es ist eine Lernmethode, die keinerlei Training oder Optimierung benötigt,
- der k-NN Algorithmus ist relativ unempfindlich gegenüber streuenden Trainingsdaten, da durch die Verwendung eines gewichteten Mittelwertes der Einfluss von Ausreißern in den Daten abgeschwächt wird,
- im Gegensatz zu anderen Verfahren ist es problemlos möglich, zu jeder Zeit neue Daten in Echtzeit in die Schätzung einzubeziehen,
- der Informationsgehalt der Trainingsdaten geht nicht verloren, wie z.B. bei der Ableitung einer Regressionsfunktion, sondern ist langfristig verfügbar,
- im Gegensatz zu einer globalen Funktion, die über einen großen Wertebereich abgeleitet wird, können lokale Schätzungen durchgeführt werden.

Gleichzeitig haben instanzenbasierte Lernverfahren jedoch generell auch einige Nachteile gegenüber regressionsanalytischen Methoden:

- Da jegliche Generalisierungsentscheidung auf den Zeitpunkt einer konkreten Suchanfrage verschoben wird, kann der Rechenaufwand für jede Klassifizierung

sehr hoch sein (in Abhängigkeit der Größe des Trainingsdatensatzes und der verwendeten Anzahl von Variablen),

- da generell alle bekannten Variablen unabhängig von ihrem tatsächlichen Einfluss auf die Zielgröße für eine Klassifizierung herangezogen werden können, kann es durch die Verwendung von irrelevanten Merkmalen zu Fehleinschätzungen kommen. Dieser Nachteil kann durch die alleinige Verwendung nachweislich relevanter Instanzen oder eine, dem Einfluss der Variablen entsprechenden Gewichtung der Einzelabstände, kompensiert werden (HESSENMÖLLER, 2001),
- die Allgmeingültigkeit der Prognosen wird durch den Umfang und die Verteilung der Beobachtungswerte bestimmt. Die empirischen Beobachtungen sind letztendlich entscheidend für die Schätzgenauigkeit (GADOW, 2003),
- aufgrund der nötigen Datenbank und einer angebundenen Anwendung, in der der k-NN Algorithmus umgesetzt wird, ist die Methode nicht so leicht zu verwenden wie ein abgeleitetes Regressionsmodell, welches leicht aus der Literatur entnommen werden kann,
- der nötige Umfang der Datengrundlage ist sehr viel höher als bei regressionsanalytischen Ansätzen,
- eine Extrapolation über den bekannten Wertebereich der Trainingsbeispiele hinaus ist nicht möglich,
- die Bestimmung der optimalen Größe der Nachbarschaft kann bei festem k aufgrund des dargestellten Bias-Varianz-Dilemmas immer nur ein Kompromiss sein. Die adaptive Anpassung der Größe der Nachbarschaft könnte dieses Problem lösen, ist jedoch sehr rechenaufwendig.

Bei der Verwendung instanzenbasierter Verfahren sind die dargestellten Argumente abzuwägen. Da im Fall von Biomasseschätzungen auf Einzelbaumebene aufgrund des relativ geringen Datenumfangs der Rechenaufwand begrenzt bleibt, werden die negativen Argumente hier abgeschwächt.

5 Schlussfolgerung

Das Ziel der vorliegenden Arbeit war es, alternative Ansätze zur Schätzung der Biomasse einzelner Bäume zu entwickeln und mit gegebenen Verfahren zu vergleichen. Die Ausrichtung der Arbeit wurde dabei auf zwei zentrale Fragestellungen aufgebaut, die sich aus den gegebenen Problemen bzw. der vorhandenen Unsicherheit bezüglich der vorhandenen Methoden ergeben. Eine Hauptmotivation der Untersuchung bestand darin, die grundlegenden Unterschiede zwischen empirischen Biomassemodellen und vorhandenen Prozessmodellen näher zu betrachten, um im Folgenden Ansätze zur Generalisierung von Biomassemodellen ableiten zu können. Im Rahmen der vermehrten Forschung auf dem Gebiet der Biomassemodellierung werden im Zuge der Umsetzung des Kyoto Protokolls vielerorts destruktive Biomasseuntersuchungen durchgeführt, um die rechtlich bindenden Vorgaben zur Quantifizierung von Kohlenstoff Senkeneffekten bzw. von Netto Emission zu erfüllen. Die Auswertung dieser Datensätze wird in ähnlicher Weise wie die bereits vorhandenen Forschungsergebnisse der letzten Jahrzehnte zu einer Vielzahl unterschiedlicher Biomassemodelle führen, die aufgrund der räumlich begrenzten Datenbasis lediglich für lokale, im besten Fall aber für regionale Schätzungen verwendet werden können. Hinzu kommt, dass die unterschiedliche mathematische Formulierung der verwendeten Regressionsmodelle den direkten Vergleich der Ergebnisse oder gar eine Generalisierung der Ansätze erschwert. Die Frage die sich hieraus ergibt ist, ob es sinnvoll ist beispielsweise 10 verschiedene Biomassemodelle für Fichten in unterschiedlichen Regionen Deutschlands abzuleiten, oder ob eine Generalisierung der Modellierungsansätze auf der Grundlage von Annahmen aus Prozessmodellen zielführender ist. Berücksichtigt man die dargestellten Unterschiede in der Sichtweise dieser verschiedenen Modellierungsansätze, könnte durch eine Kombination in Form von Hybridmodellen ein Ansatz für eine Verallgemeinerung gefunden werden.

Bisher stützen sich die für die Kohlenstoffbudgetierung auf nationaler Ebene verwendeten Biomassemodelle auf relativ kleine Datensätze. Oftmals wird eine überregionale Kohlenstoffbilanz auf Grundlage von weniger als 50 Testbäumen einer betreffenden Baumart erstellt, da keine allgemeingültigeren Berechnungsgrundlagen zur Verfügung stehen. Für manche Baumarten sind nicht einmal in den gut untersuchten europäischen Waldökosystemen, mit ihrer relativ überschaubaren Artenzahl, angepasste Berechnungsgrundlagen vorhanden. Bedenkt man, dass in einer nationalen Kohlenstoffbilanz

die Emissionen verschiedener Sektoren, die sich über den Verbrauch fossiler Brennstoffe relativ genau quantifizieren lassen, den Senkeneffekten gegenübergestellt werden, so führt die gegebene Unsicherheit bezüglich der Biomassemodellierung in Waldökosystemen dazu, dass die vorhandenen Berechnungsgrundlagen der steigenden wirtschaftlichen Bedeutung der Kohlenstoffbindung nicht Rechnung tragen können.

Eine Erweiterung der vorhandenen Datenbasis durch eine Zusammenfassung der Ergebnisse einzelner Studien, könnte in diesem Fall helfen, bessere und gleichzeitig allgemeingültigere Aussagen zu treffen. Der Hauptvorteil hierbei besteht neben der Verwendung größerer Datensätze, darin, dass weitere Einflussgrößen sinnvoll in die Modellformulierung integriert werden können. Da die als unabhängigen Variablen geeigneten Messwerte in räumlich abgegrenzten Untersuchungsgebieten untereinander korreliert sind, ist ihre Verwendung oft nur in zusammengesetzten Datensätzen sinnvoll.

Als Alternative zu gegebenen Modellierungsansätzen wurden in dieser Arbeit verschiedene Modifikationen der k-NN Methode als Beispiel für ein instanzenbasiertes Verfahren verwendet. Die Motivation hierfür bestand darin, ein generalisiertes Schätzverfahren zu entwickeln, dass im Gegensatz zu empirischen- und Prozessmodellen keinerlei modellhafte Annahmen benötigt.

Zu Beginn dieser Untersuchung stellte sich die provokative Frage: Könnte man Biomassedaten von verschiedenen Baumarten aus europäischen Wäldern verwenden, um Aussagen über relativ unbekannte Baumarten in anderen Teilen der Welt abzuleiten, wenn sich diese in ihren Haupteigenschaften sehr ähnlich sind? Bäume werden in diesem Ansatz als Instanzen verschiedener Ausprägung aufgefasst, die sich z.B. evolutionsbedingt in ihrem Grundmuster gleichen. Ein Mustererkennungsverfahren, welches gleichzeitig Einzelbaumdaten sowie auch verfügbare Meta-Informationen nutzt um Ähnlichkeiten in der vorhandenen Wissensbasis zu quantifizieren, scheint unter dieser Grundannahme geeignet zu sein, um die Gesamtheit der vorhandenen Informationen über die gesuchte Zielgröße in einer empirischen Datengrundlage so zu „sortieren“, dass eine Schätzung über die ähnlichsten Merkmalsträger möglich ist.

6 Zusammenfassung

Die vorliegende Arbeit beschäftigt sich mit der Biomasseschätzung auf Einzelbaumebene. Berechnungsmöglichkeiten für die trockene Biomasse von Bäumen sind vor allem im Rahmen der Umsetzung des Kyoto Protokolls zur Einschätzung der Kohlenstoffbindung in Waldökosystemen von Interesse. Das Problem bei der Anwendung vorhandener Biomassefunktionen besteht hauptsächlich darin, dass die empirische Forschung der letzten Jahrzehnte, zumindest für gut untersuchte Baumarten, eine kaum überschaubare Fülle von Schätzfunktionen unterschiedlichster mathematischer Formulierung hervorgebracht hat. Diese Modelle stützen sich, aufgrund des destruktiven Charakters der Datenaufnahme in Biomasseuntersuchungen, oftmals auf relativ kleine Datensätze aus lokal begrenzten Untersuchungsgebieten. Für die Erstellung von regionalen oder nationalen Kohlenstoffbilanzen werden allerdings generellere Modelle mit einem größeren Geltungsbereich benötigt.

Die Untersuchung befasst sich im Wesentlichen mit zwei Gesichtspunkten, die eine Generalisierung von Biomassefunktionen ermöglichen könnten. Zum einen kann die Integration von empirischer Forschung und Prozessmodellen zu einer Vereinheitlichung der abgeleiteten Modellformulierungen in Biomasseuntersuchungen in Form von Hybridmodellen genutzt werden. Zum anderen können auf Grundlage einer erweiterten Datengrundlage, die durch eine Zusammenstellung der vorhandenen Datensätze aus einzelnen Biomassestudien erstellt werden kann, auch instanzenbasierte Verfahren zur Biomasseschätzung verwendet werden.

Im Rahmen dieser Arbeit wird zunächst überprüft, welche grundlegenden Unterschiede beim Vergleich zwischen Vorhersagen aus Prozessmodellen und empirischen Biomassefunktionen zu beachten sind. Hierbei führt die unterschiedliche Motivation der Ansätze bisher dazu, dass Vorhersagen aus Prozessmodellen in der praktischen Anwendung keine Rolle spielen, da sie sich anhand empirischer Forschungsergebnisse nicht ausreichend bestätigen lassen. Als eine mögliche Hauptursache für diese Diskrepanz konnte die Verwendung des BHDs, der im Prinzip keine funktionale Messgröße von Bäumen darstellt, in allometrischen Biomassefunktionen identifiziert werden. Während Prozessmodelle Verhältnisregeln für relative Wachstumsraten innerhalb eines Organismus vorhersagen, stützt sich die empirische Forschung auf die Verwendung absoluter Messgrößen als unabhängige Variablen. Am Beispiel eines zusammengesetzten Fichten-

datensatzes konnte, durch die Umrechnung des BHDs zu einem Durchmesser in relativer Stammhöhe, eine Annäherung der Vorhersagen aus empirischen Regressionsmodellen und den theoretischen Verhältnisregeln eines Prozessmodells erzielt werden.

Als weiterer Erklärungsansatz für die Unterschiede beider Ansätze wird vermutet, dass destruktive Biomasseuntersuchungen im Gegensatz zu Prozessmodellen, nur eingeschränkte Aussagen über das ontogenetische Wachstum einzelner Bäume ermöglichen, da sie nur in Zeitreihenuntersuchungen (sog. Chronosequenzen) gewonnen werden können. Vielmehr muss bedacht werden, dass bei der Untersuchung von Bäumen unterschiedlicher Dimension zu einem bestimmten Zeitpunkt, implizit auch Effekte der jeweiligen Bestandesgeschichte erfasst werden, was wiederum dazu führt, dass ihre Anwendung auf regionaler Ebene kritisch zu betrachten ist.

Das Hauptziel dieser Arbeit liegt in der Überprüfung der Anwendbarkeit eines instanzenbasierten Prognoseverfahrens auf Einzelbaumebene. Hierbei wurde die k-Nearest-Neighbour (k-NN) Methode, ein nicht-parametrisches Klassifizierungsverfahren, zur Biomasseschätzung verwendet. Im Gegensatz zu Prozessmodellen sowie empirischer Datenanalyse, setzt dieses Verfahren nur eingeschränkte Kenntnisse über die bestehenden Wirkungszusammenhänge der einzelnen Einflussgrößen voraus und erfordert daher keine explizite Modellbildung.

Für die Umsetzung wurde eine angepasste Einzelbaumdatenbank aufgebaut, die die Trainingsbeispiele für den k-NN Algorithmus enthält. Mit Hilfe von unterschiedlichen Distanzmaßen aus der multivariaten Statistik, wird aus dieser Datenbasis eine gewisse Anzahl ähnlicher Merkmalsträger identifiziert, die unter der Annahme, dass sie sich auch in Bezug auf das gesuchte Merkmal ähneln, zur Schätzung der Zielgröße herangezogen werden. Hierzu werden die Merkmalswerte dieser k Trainingsbeispiele durch eine gewichtete oder ungewichtete Mittelwertbildung zur lokalen Approximation der Zielgröße verwendet. Eine Anpassung der zur Schätzung verwendeten Größe der berücksichtigten Nachbarschaft sowie der Distanzfunktion, wird jeweils durch die Minimierung ausgewählter Fehlermaße mit hilfe eines iterativen Prozesses bzw. einer multiplen Kreuzvalidierung der Trainingsdaten erzielt.

In verschiedenen Teilauswertungen, die sich auf unterschiedlich große Datensätze beziehen, wurde die Prognosegüte der k-NN Schätzungen durch den Vergleich verschiedener Fehlermaße mit denen von Regressionsmodellen verglichen, die jeweils auf Grundlage der gleichen Datenbasis angepasst wurden. Hierbei konnte für einzelne

Teiluntersuchungen eine Reduktion verschiedener Prognosefehler durch die Verwendung der k-NN Methode nachgewiesen werden.

Weiterhin konnte das für instanzenbasierte Verfahren typische Bias-Varianz-Dilemma, das hauptsächlich durch die mangelnde Extrapolationsfähigkeit des Ansatzes entsteht, dadurch abgemildert werden, dass wahlweise auch eine adaptive Größe der Nachbarschaft zur Schätzung verwendet wird. Hierbei wird die Bandbreite, der zur Schätzung verwendeten Trainingsinstanzen, nicht durch eine betsimmte Anzahl (k) festgelegt, sondern Nachbarn werden bis zu einer bestimmten normierten Distanz berücksichtigt. Hierdurch werden in Abhängigkeit der Verteilung der Trainingsinstanzen in Bereichen hoher Dichte mehr Nachbarn zur Schätzung verwendet als in Bereichen mit geringer Anzahl von Trainingsbeispielen.

7 Literatur

AHA, D.W., 1998. Feature weighting for lazy learning algorithms. Navy Center for Applied Research in Artificial Intelligence. Unveröffentlichter Bericht.

AKÇA, A., MENCH, A., 1993. Biomasseentwicklung in umweltbelasteten Fichtenbeständen des Einzugsgebietes Lange Bramke. Abschlussbericht. In: Berichte des Forschungszentrums Waldökosysteme, Reihe B, Bd. 37, 14 S.

ALBERTI, G., CANDIDO, P., PERESSOTTI, A., TURCO, S., PIUSSI, P., ZERBI, G., 2005. Aboveground biomass relationships for mixed ash (*Fraxinus excelsior* L. and *Ulmus glabra* Hudson) stands in Eastern Prealps of Friuli Venezia Giulia (Italy). Ann. For. Sci. 62, 831-836.

ALTMAN, N.S., 1990. Kernel Smoothing of Data With Correlated Errors. Journal of the American Statistical Association 85, No.411, 749-759.

ALTMAN, N.S., 1992. An introduction to kernel and nearest neighbour nonparametric regression. Am. Stat. 46, 175-184.

ANTTILA, P., 2002. Nonparametric estimation of stand volume using spectral features of aerial photographs. Can. J. For. Res. 32, 1849-1857.

ART, H. W., MARKS, P. L., 1971. A summary table of biomass and net annual primary production in forest ecosystems of the world. IUFRO: In: Young, H. E. (ed) Forest Biomass Studies, University of Maine at Orono, USA. Life science and Agricultural Experiment Station, 3-32.

ASSMANN, E., 1953. Zur Bonitierung süddeutscher Fichtenbestände. AFZ 10, 61-64.

ATKESON, C.G, MOORE, A.W., SCHAAL, S., 1996. Locally Weighted Learning. College of Computing, Georgia Institute of Technology, Atlanta. 52 S. http://www.cc.gatech.edu/fac/Chris.Atkeson/local-learning/.

ATKESON, C.G., MOORE, A.W., SCHAAL, S., 1997. Locally weighted learning. Artificial Intelligence Review 11 (1-5), 11-73.

BACKHAUS, K., ERICHSON, B., PLINKE, W., WEIBER, R., 2005. Multivariate Analysemethoden. Eine anwendungsorientierte Einführung. 11. Auflage, Springer-Verlag, 831 S.

BARITZ, R., STRICH, S., 2000. Forests and the National Greenhouse Gas Inventory of Germany. Biotechnol. Agron. Soc. Environ. 4 (4), 267-271.

BARTELINK, H.H., Allometric relationships for biomass and leaf area of beech (*Fagus sylvatica* L). Ann. Sci. For. 54, 39-50.

BARTELINK, H.H., 1998. A model of dry matter partitioning in trees. Tree Physiology 18, 91-101.

BASKERVILLE, G.L., 1965: Estimation of dry weight of tree components and total standing crop in conifer stands. Ecology 46; 867-869.

BASKERVILLE, G.L., 1972. Use of logarithmic regression in the estimation of plant biomass. Can. J. For. Res. 2, 49-53.

BEAUCHAMP, J.J., OLSON J.S., 1973. Corrections for bias in regression estimates after logarithmic transformation, Ecology 54, 1403–1407.

BELLMANN, R., 1961. Adaptive control processes: aguided tour. Princeton University Press. 255 S.

BERTALANFFY, V.L., 1951. Theoretische Biologie II. Band, Stoffwechsel, Wachstum. 72 S.

BOLTE, A., RAHMANN, T., KUHR, M., POGODA, P., MURACH, D., GADOW, K.V., 2004. Relationships between tree dimensions and coarse root biomass in mixed stands of European beech (*Fagus sylvatica* L.) and Norway spruce (*Picea abies* [L.] Karst.). Plant and Soil 264, 1-11.

BORTZ, J., 2004. Statistik für human- und Sozialwissenschaftler. 6. Auflage. Springer Verlag, 900 S.

BROWN, J.H., WEST, G.B., 2000. Scaling in Biology. Oxford University Press, New York.

BROWN, S., 1997. Estimating biomass and biomass change of tropical forests. A Primer. FAO Forestry Paper 134. FAO Rome.

BROWN, S., 2001. Measuring carbon in forests: current status and future challenges. Environm. Pollut. 116 (3), 363-372.

BURGER, H., 1925. Holz-, Laub- und Nadeluntersuchungen. Schweiz. Zeitschr. f. Forstwesen.

BURGER, H., 1929. Holz, Blattmenge und Zuwachs. Die Weymouthsföhre. Mitteilungen der schweiz. Anst. F. forstl. Versuchswesen Bnd. XV, 2. Heft.

BURGER, H., 1935. Holz, Blattmenge und Zuwachs. Die Douglasie. Mitteilungen der Schweiz. Anst. F. forstl. Versuchswesen Bnd. XIX, 1. Heft.

Burger, H., 1939. Baumkrone und Zuwachs in zwei hiebsreifen Fichtenbeständen. Mitteilungen der Schweiz. Anst. F. forstl. Versuchswesen Bnd. XXI, 1. Heft.

BURGER, H., 1940. Holz, Blattmenge und Zuwachs. Ein 80 jähriger Buchenbestand. Mitteilungen der schweiz. Anst. F. forstl. Versuchswesen Bnd. XXI, 2. Heft.

BURGER, H., 1945. Holz, Blattmenge und Zuwachs. VII: Die Lärche. Mitt. Schw. Anst. f. d. Forstl. Versw. 24, 7-103.

BURGER, H., 1947. Holz, Blattmenge und Zuwachs. Die Eiche. Mitteilungen der Schweiz. Anst. F. forstl. Versuchswesen Bnd. XXV, 1. Heft.

BURGER, H., 1948. Holz, Blattmenge und Zuwachs. Die Föhre. Mitteilungen der Schweiz. Anst. F. forstl. Versuchswesen Bnd. XXV, 2. Heft.

BURGER, H., 1950. Holz, Blattmenge und Zuwachs. Die Buche. Mitteilungen der Schweiz. Anst. F. forstl. Versuchswesen Bnd. XXVI, 2. Heft.

BURGER, H., 1951. Holz, Blattmenge und Zuwachs. Die Tanne. Mitteilungen der Schweiz. Anst. F. forstl. Versuchswesen Bnd. XXVII.

BURGER, H., 1953. Holz, Blattmenge und Zuwachs. XIII: Fichten im gleichaltrigen Hochwald. Mitteilungen der Schweiz. Anst. F. forstl. Versuchswesen 29, 38-130.

CANELL, M., 1982: World Forest Biomass and Primary Production Data. ACADEMIC PRESS. London.

CANELL, M., 1995. Forest and the global carbon cycle in the past, present and future. European Forest Institute Research Report 2, Joensuu, 66 S.

CERNY, M., 1990. Biomass of *Picea abies* (L.) Karst. In Midwestern Bohemia. Scan. J. For. Res. 5, 83-95.

CHAVE, J., RIERA, B., DUBOIS, M.-A., 2001. Estimation of biomass in a neotropical forest of French Guiana: spatial and temporal variability. Journal of Tropical Ecology 17, 79-96.

CHAVE, J., ANDALO, C., BROWN, S., CAIRNS, M.A., CHAMBERS, J.Q., EAMUS, D., FÖLSTER, H., FROMARD, F., HIGUCHI, N., KIRA, T., LESCURE, J.-P., NELSON, B.W., OGAWA, H., PUIG, H., RIÉRA, B., YAMAKURA, T., 2005. Tree allometry and improved estimation of carbon stocks and balance in tropical forests. Oecologia 145 (1), 87-99.

CIENCIALA, E., CERNY, M., APLTAUER, J., EXNEROVA, Z., 2003. Biomass functions applicable to European beech. J. For. Sci. 51 (4), 147-154.

CLEVELAND, W.S., LOADER, C., 1994. Local fitting for semiparametric (nonparametric) regression: Comments on a paper of Fan and Marron. Technical Report 8, AT&T Bell Laboritories, Statistics Department, Murray Hill, NJ.

COVER, T., HART, P., 1967. Nearest neighbour pattern classification. IEEE Transactions on Information Theory, 13:1, 21-27.

DIETRICH, H., 1968. Untersuchungen zur Nährstoffdynamik eines Fichtenbestandes. I. Mitteilungen: Messwerte des Fichtenbestandes und Einfluss einer Bestandeskalkung. Archiv für Forstwesen 17 (4), 391-412.

DROSTE ZU HÜLSHOFF, B., 1969. Struktur und Biomasse eines Fichtenbestandes aufgrund einer Dimensionsanalyse an oberirdischen Baumorganen. Inaugural Dissertation, Ludwig-Maximilians-Universität München, 209 S.

DUVIGNEAUD, P., DENAEYER-DE SMET, S., 1970. Biological Cycling of Minerals in Temperate Decidous Forests. Springer-Verlag, Berlin, Heidelberg, New York.

EAMUS, D., MCGUINESS, K., BURROWS, W, 2000. Review of Allometric Relationships for Estimating Woody Biomass for Qld, the NT and WA. National Carbon Accounting System Technical Report No. 5a, Australian Greenhouse Office.

ENQUIST, B.J., 2002. Universal scaling in tree and vascular plant allometry: toward a general quantitative theory linking plant form and function from cells to ecosystems. Tree Physiology 22, 1045-1064.

ENQUIST, B.J., BROWN, J.H., WEST, G.B., 1998. allometric scaling of plant energetics and population density. Nature 395, 163-165.

ENQUIST, B.J., WEST, G.B., CHARNOW, E.L., BROWN, J.H., 1999. Allometric scaling of production and life-history variation in vascular plants. Nature 401, 907-911.

ENQUIST, B.J., NIKLAS, K.J., 2002. Global allocation rules for patterns of biomass partitioning across seed plants. Science 295, 1517-1520.

FAN, J., 2000. Prospects of nonparametric modelling. Invited review article. Journal of American Statistical Association 95, 1296-1300.

FEHRMANN, L., KUHR, M., GADOW, K.v., 2003. Untersuchung der Wurzelstruktur großer Waldbäume an Fichte und Buche. Forstarchiv 74, 96-102.

FEHRMANN, L., KLEINN, C., 2005. Vergleich von Biomasseuntersuchungen in Hinblick auf die verwendeten Variablen sowie die abgeleiteten Modelle. Projektbericht im Auftrag der Forschungsanstalt für Waldökologie und Forstwirtschaft des Landes Rheinland-Pfalz, 27 S.

FEHRMANN, L., KLEINN, C., KIRCHNER, T., KRAPF, C., 2006. Entwicklung einer Einzelbaum-Biomassedatenbank als Grundlage zur Generalisierung von Biomassefunktionen. Projektbericht im Auftrag der Forschungsanstalt für Waldökologie und Forstwirtschaft des Landes Rheinland-Pfalz, 30 S.

FEHRMANN, L., KLEINN, C., 2006. Using a k-Nearest Neighbour (k-NN) Approach for single tree biomass estimation. In: Proceedings of the 7$^{\text{th}}$ Annual Forest Inventory and Analysis Symposium. October 3.-6., 2005, Portland, Maine.

FEHRMANN, L., KLEINN, C., 2006. General considerations about the use of allometric equations for biomass estimation on the example of Norway spruce in central Europe. For. Ecol. Man. 236, 412-421.

FEHRMANN, L., LEHTONEN, A., TOMPPO, E., KLEINN, C., 2006. Comparison of simple linear and mixed effect regression models and a *k*-Nearest Neighbour (*k*-NN) approach for estimation of single tree biomass. in Preparation.

FIEDLER, F., 1986. Die Dendromasse eines hiebsreifen Fichtenbestandes. Beiträge für die Forstwirtschaft 20, 171-180.

FINLEY, A. O., MCROBERTS, E., EK, A. R., 2006. Applying an Efficient k-Nearest Neighbour Search to Forest Attribute Imputation. For. Sci 52(2), 130-135.

FINNEY, D.J., 1941. On the distribution of a variate whose logarithm is normally distributed. J. Roy. Stat. Soc. S. B70, 155-161.

FITZHUGH, H.A. JR., 1976. Analysis of growth curves and strategies for altering their shape. J. Anim. Sci. 42, 1036-1050.

FIX, E., HODGES, J.L., 1951. Discriminatory analysis, nonparametric discrimination, consistency properties. Technical Report 4, United States Air Force, School of Aviation Medicine.

FURNIVAL, G.M., 1961. An Index for Comparing Equations Used In Constructing Volume Tables. Forest Science 7, 337-341.

GADOW, K.V., 2003. Waldstruktur und Waldwachstum. Universitätsverlag Göttingen, 246 S.

GADOW, K.V., FEHRMANN, L., MURACH, D., WALOTEK, P., 2006. Collating available information and Designing Field Trials for Energy Plantations. Manuscript prepared for the International IUFRO Conference, 3.-7. April 2006, University of Valladolid, Spain. 15 S.

GAFFREY, D., SLOBODA, B., 2001. Tree mechanics, hydraulics and needle-mass distribution as a possible basis for explaining the dynamics of stem morphology. J. For. Sci. 47 (6), 241-254.

GIFFORD, R.M., 2000. Carbon Contents of Above-Ground Tissues of Forest and Woodland Trees. National Carbon accounting System Technical Report No. 22. Australian Greenhouse Office. 27 S.

GILLE, U., 1989. Vergleichende Betrachtung zum postnatalen Wachstum der Körpermasse und ausgewählter Extremitätenmaße verschiedener Haus- und Laborspezies. Dissertation zur erlangung des Doktorgrades, Universität Leipzig, 97 S.

GRIESON, P., WILLIAMS, K., ADAMS, M., 2000. Review of Unpublished Biomass-Related Information: Western Australia, South Australia, New South Wales and Queensland. National Carbon Accounting System Technical Report No.25, Australian Greenhouse Office, 96 S.

GROTE, J., 2002. Foliage and branch biomass estimation of coniferous and deciduous tree species. Silva Fennica 36, 779-788.

GROTE, R., SCHUCK, J., BLOCK, J., PRETZSCH, H., 2003. Oberirdische holzige Biomasse in Kiefern-/Buchen- und Eichen-/Buchen-Mischbeständen. Fortw. Cbl. 122, 287-301.

GRUNDNER, F., SCHWAPPACH, A., 1952. Massentafeln zur Bestimmung des Holzgehaltes stehender Waldbäume und Waldbestände. Paul Parey, Berlin.

HAARA, A., MALTAMO, M., TOKOLA, T., 1997. The k-nearest neighbour method for estimating basal-area distribution. Scand. J. For. Res. 12, 200-208.

HAENDEL, L., 2003. Clusterverfahren zur datenbasierten Generierung interpretierbarer Regeln unter Verwendung lokaler Entscheidungskriterien. Dissertation an der Fakultät für Elektrotechnik und Informationstechnik der Universität Dortmund, 120 S.

HAKKILA, P., 1989. Utilization of Residual Forest Biomass. Springer Verlag, Berlin Heidelberg.

HELLER, H., GÖTTSCHE, D., 1986. Biomasse-Messungen an Buche und Fichte. In: H., E. (Ed.): Ökosystemforschung – Ergebnisse des Solling-Projekts. Eugen Ulmer, Stuttgart, 507 S.

HESSE, C., 1990. Inventur der Bestandesbiomasse und ausgewählter chemischer Elemente in einem 63-jährigen Fichtenbestand im Sauerland. Diplomarbeit, Georg-August Universität Göttingen, 230 S.

HESSENMÖLLER, D., 2001. Modelle zur Wachstums- und Durchforstungssimulation im Göttinger Kalkbuchenwald. Dissertation zur Erlangung des Doktorgrades der Fakultät für Forstwissenschaften und Waldökologie der Georg-August Universität Göttingen. Logos Verlag Berlin, 163 S.

HESSENMÖLLER, D., ELSENHANS, A.S., 2002. Zur Schätzung des Zuwachses bei Rotbuche Fagus sylvatica L.. Ein Vergleich parametrischer Verfahren mit der k-nearest neighbour Methode. Allgemeine Forst- und Jagdzeitung 173 (11/12), 216-223.

HINRICHS, L., 2006. Konstruktion eines Modells zur automatisierten Generierung von Behandlungspfaden als Grundlage zur Steuerung der Nutzungsplanung für Buchen Fichten Mischbestände. Dissertation zur Erlangung des Doktorgrades der

Fakultät für Forstwissenschaften und Waldökologie der Georg-August Universität Göttingen.

HOLM, S., HÄGGLUND, B., MÅRTENSON, A., 1997. A method for generalization of sample tree data from the Swedish National Forest Survey. Swedish University of Agricultural Sciences. Department of Forest Survey. Report No. 25, 94 S. Schwedisch mit englischer Zusammenfassung.

HOLMSTRÖM, H., 2002. Estimation of single-tree characteristics using the kNN method and plotwise aerial photograph interpretation. For. Ecol. Man. 167, 303-314.

HOLMSTRÖM, H., NILSSON, M., STÅHL, G., 2001. Simultaneous Estimations of Forest Parameters using Aerial Photograph Interpreted Data and the k Nearest Neighbour Method. Scan. J. For. Res. 16 (1), 67-78.

HUXLEY, J.S., 1924. Constant differential growth-ratios and their significance. Nature 114; 895.

HUXLEY, J.S., 1932. Problems of Relative Growth. Methuen & Co., Ltd London.

JENKINS, J.C., CHOJNACKY, D.C., HEATH, L.S., BIRDSEY, R.A., 2003. National-Scale Biomass Estimators for United States Tree Species. For. Sci. 49 (1), 12-35.

JENKINS, J.C., CHOJNACKY, D.C., HEATH, L.S., BIRDSEY, R.A., 2004. Comprehensive database of diameter-based biomass regressions for North American tree species. Gen. Tech. Rep. GTR-NE-319, Northeastern Research Station, USDA Forest Service, 203 S.

JOHANSSON, T., 1999. Biomass Production of Norway Spruce (Picea abies (L.) Karst.) Growing on Abandoned Farmland. Silva Fennica 3384), 261-280.

JOHNSEN, K., SAMUELSON, L., TESKEY, R., MCNULTY, S., FOX, T., 2001. Process models as tools in forestry and management. For. Sci. 47, 2-8.

JOOSTEN, R., SCHULTE, A., 2003. Schätzung der Reisigbiomasse und ihrer Nutzungspotentiale in Buchenbeständen (*Fagus sylvatica* L.) auf Grundlage von Inventurdaten. Forstarchiv 74, 159-165.

JOOSTEN, R., SCHUMACHER, J., WIRTH, C., SCHULTE, A., 2004. Evaluating tree carbon predictions for beech (Fagus sylvatica L.) in western Germany. For. Ecol. Manag. 189, 87-96.

KATILA, M., 2004. Error variations at the pixel level in the k-nearest neighbour predictions of the Finnish multi-source forest inventory. Proceedings of the 1st GIS & Remote Sensing Days (GGRS), 07.-08. 10.2004, Göttingen. 408 S.

KEITH, H., BARRETT, D., KEENAN, R., 2000. Review of allometric relationships for estimating woody biomass for New South Wales, the Australian Capital

Territory, Victoria, Tasmania, and South Australia. National carbon accounting system, technical report no. 5b. Australian Greenhouse Office, Canberra, 114S.

KETTERINGS, Q.M., NOORDWIJK, C.M.Y., AMBAGAU, R., PALM, C.A., 2001. Reducing uncertainty in the use of allometric biomass equations for predicting above-ground tree biomass in mixed secondary forests. For. Ecol. Man. 146, 199-209.

KILKKI, P., PÄIVINEN, R., 1987. Reference sample plots to compare field measurements and satellite data in forest inventory. Proceedings of Seminar on Remote Sensing-aided Forest Inventory, Finnland 10-12 December 1986. University of Helsinki, Department of Forest Mensuration and Management, 209-215.

KORHONEN, K.T., KANGAS, A., 1997. Application of nearest-neighbour regression for generalizing sample tree information. Scan. J. For. Res. 12, 97-101.

KORHONEN, K.T., MALTAMO, M., 1990. Männyn maanpäällisten osien kuivamassat Etelä-Suomessa. Metsäntutkimuslaitoksen tiedonantoja, Joensuun tutkimusasema 371, 1-29. In Finnisch.

KOTZ, S., READ, C.B., BANKS, D.L.,(eds) 1998: Encyclopedia Of Statistical Science. Updated Volume 2. John Wiley & Sons, New York. 745 S.

KRAMER, H., 1988. Waldwachstumslehre. Verlag Paul Parey, Hamburg und Berlin.

KRAMER, K., LEINONEN, I., BARTELINK, H.H., BERGBIGIER, P., BORGHETTI, M., BERNHOFER, C., CIENCIALA, E., DOLMAN, A.J., FROER, O., GRACIA, C.A., GRANIER, A., GRÜNWALD, T., HARI, P., JANS, W., KELLOMÄKI, S., LOUSTAU, D., MAGNANI, F., MARKKANEN, T., MATTEUCCI, G., MOHREN, G.M.J., MOORS, E., NISSINEN, A., PELTOLA, H., SABATÉ, S., SANCHEZ, A., SONTAG, M., VALENTINI, R., VESALA, T., 2002. Evaluation of six process based forest growth models using eddy-covariance measurements of CO2 and H2O fluxes at six forest sites in Europe. Global Change Biology 8 (3), 213-230.

KRAMER, H. U. AKÇA, A., 1995. Leitfaden zur Waldmesslehre. JD Sauerländer's Verlag. Frankfuhrt am Main.

KUNZE, M., 1837. Lehrbuch der Holzmesskunst. Zweiter Band, Wiegandt und Hempel, Berlin.

LABARBERA, M., 1989. Analyzing body size as a factor in ecology and evolution. Annu. Rev. Ecol. Syst. 20, 97-117.

LAMBERT, M.-C., UNG, C.-H., RAULIER, F., 2005. Canadian national tree aboveground biomass equations. Can. J. For. Res. 35, 1996-2018.

LAPPI, J., MEHTÄTALO, L., KORHONEN, K.T., 2006. Generalizing sample tree information. In: KANGAS, A., MALTAMO, M. (eds.). Forest inventory methodology and application. Managing forest ecosystems 10. Springer, Dordrecht. 362 S.

LAWRENCE, S., CHUNG TSOI, A., BACK, A.D., 1996. Function approximation with neural networks and local methods: bias, variance and smoothness. In: Barlett, P., Burkitt, A., Williamson, R. (eds): Australian Conference on Neural Networks. Australian National University, 16-21.

LEMAY, V., TEMESGEN, H., 2004. Comparison of Nearest Neighbour Methods for Estimating Basal Area and Stems per Hectare Using Aerial Auxiliary Variables. For. Sci. 51(2), 109-119.

LEHTONEN, A., MÄKIPÄÄ, R., HEIKKINEN, J., SIEVÄNEN, R., LISKI, J., 2004. Biomass expansion factors (BEFs) for Scots pine, Norway spruce and birch according to stand age of boreal forests. For. Ecol. Man. 188, 211-224.

LEHTONEN, A., 2005a. Carbon stocks and flows in forest ecosystems based on forest inventory data. Academic dissertation, Faculty of Agriculture and Forestry of University of Helsinki, 51 S.

LEHTONEN, A., 2005b. Estimating foliage biomass for Scots pine (Pinus sylvestris L.) and Norway spruce (Picea abies (L.) Karst.) plots. Tree Physiology 25(7): 803-811.

LEMM, R., VOGEL, M., FELBER, A., THEES, O., 2005. Eignung der k-Nearest Neighbour (kNN-) Methode zur Schätzung von Produktivitäten in der Holzernte – Grundsätzliche Überlegungen und erste Erfahrungen. Allg. Forst- u. J.-Ztg., 176, 189-200.

LOADER, C., 1999. Local Regression and Likelihood. Springer Verlag, 290 S.

MADWICK, H. A. I., 1976. Mensuration of forest biomass. In: Young, A. (ed), Oslo Biomass Studies, University of Maine at Orono, USA.

MÄKELÄ, A., LANDSBERG, J.J., EK, A.R., BURK, T.E., TER-MIKAELIAN, M., ÅGREN, G.I., OLIVER, C.D., PUTTONEN, P., 2000. Process-based models for forest ecosystem management: current state of the art and challenges for practical implementation. For. Sci. 20, 289-298.

MCCULLOCH, C.E., SEARLE, S.R., 2000. Generalized, Linear and Mixed Models. Wiley Series in Probability & Statistics 1. John Wiley and Sons Ltd, New York. 358 S.

MACFARLANE, D.W., GREEN, E.J., VALENTINE, H.T., 2000. Incorporating uncertainty into the parameters of a forest process model. Ecological Modelling 134, 27-40.

MADGWICK, H., SATOO, T., 1975. On estimating the aboveground weights of tree stands. Ecology 56, 1446-1450.

MAHALANOBIS, P.C., 1936. On the generalized distance in statistics. Proc. Natl. Insts. Sci. India 2, 49-55.

MALENDE, Y., 1997. Biomasseinventur auf der Basis von terrestrischen Aufnahmen und Luftbild-Stichproben – Dargestellt in einem Tansanischen Miombo-Waldgebiet. Dissertation. Fakultät für Forstwissenschaften und Waldökologie. Universität Göttingen.

MALINEN, J., 2003a. Prediction of characteristics of marked stand and metrics for similarity of log distribution for wood procurement management. Academic Dissertation, Faculty of Forestry of the University of Johensuu, Finnland.

MALINEN, J., 2003b. Locally Adaptable Non-parametric Methods for Estimating Stand Characteristics for Wood Procurement Planning. Siva Fennica 37(1), 109-118.

MALINEN, J., MALTAMO, M., HARSTELA, P., 2003. Application of Most Similar Neighbour Inference for Estimating Marked Stand Characteristics Using Harvester and Inventory Generated Stem Database. International Journal of Forest Engineering 33

MALINEN, J., MALTAMO, M., VERKASALO, E., 2003. Predicting the internal quality and value of Norway spruce trees by using two non-parametric nearest neighbour methods. Forest Products Journal 53(4), 85-94.

MALTAMO, M., KANGAS, A., 1998. Methods based on k-nearest neighbour regression in the prediction of basal area diameter distribution. Can. J. For. Res. 28, 1107-1115.

MALTAMO, M., MALINEN, J., KANGAS, A., HÄRKÖNEN, S., PASANEN, A.-M., 2003. Most similar neighbour-based stand variable estimation for use in inventory by compartments in Finland. Forestry 76(4), 449-463.

MARKLUND, L.G., 1988. Biomassfunktioner för tall, gran och björk I Sverige. Sveriges Lantbruksuniversitet, Rapporter-Skog 45, 1-73.

MCCULLOCH, C.E., SEARLE, S.R., 2000. Generalized, Linear and Mixed Models. Wiley Series in Probability & Statistics 1. John Wiley and Sons Ltd, New York. 358 p.

MCROBERTS, R., NELSON, M.D., WENDT, D.G., 2002. Stratified estimation of forest area using satellite imagery, inventory data, and the k-Nearest Neighbour technique. Remote Sensing of Environment 82, 457-468.

MENCUCCINI, M., ZIANIS, D., 2002. Biological Bases of Allometric Studies in Forest Research. Presentation. COST E21 Workshop, 7-8 October, Valencia, Spain.

MENDENHALL, W., SINCICH, T., 1993. A Second Course in Statistics: Regression Analysis. 5th ed. Prentice Hall, 766 S.

MITCHELL, C.A., MYERS, P.N., 1995. Mechanical stress regulation of plant growth and development. Horticultural Reviews 17, 1-42.

MITCHELL, T., 1997: Machine Learning. McGraw Hill College Div., 432 S.

MOER, M., 1987. Nearest neighbour inference for correlated multivariate attributes. Proceedings of IUFRO Conference on Forest Growth Modelling and Prediction, Minneapolis, 23-27 August 1987. USDA Forest Service, General Technical Report NC-120 (Saint Paul, MN, US Department of Agriculture, Forest Service, North Central Forest Experiment Station), 716-723.

MOER, M., HERSHEY, R.R., 1999. Preserving spatial and attribute correlation in the interpolation of forest inventory data. In: Lowell, K., Jaton, A. (eds). Spatial accuracy assessment: land information uncertainty in natural resources. Third International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, Quebec City, Canada, 20-22 Mai 1998. Ann Arbor Press, Chelsa, Michigan, 419-430.

MOER, M., STAGE, A.R., 1995. Most similar neighbour: an improved sampling inference procedure for natural resource planning, For. Sci. 41, 337-359.

MONSERUD, R.A., ONUCHIN, A.A., TCHEBAKOVA, N.M., 1996. Needle, crown, stem and root phytomass of *Pinus sylvestris* stands in Russia. For. Ecol. Man. 82, 59-67.

MONSERUD, R.A., 2003. Evaluating Forest Models in a sustainable Forest Management context. FBMIS Vol 1, 35-47.

MONSERUD, R.A., ONUCHIN, A.A., TCHEBAKOVA, N.M., 1995. Needle, crown, stem, and root Phytomass of *Pinus sylvestris* stands in Russia. For. Ecol. Man. 82, 59-67.

MONTAGU, K.D., DÜTTMER, K., BARTON, C.V.M., COWIE, A.L., 2004. Developing general allometric relationships for regional estimates of carbon sequestration – an example using *Eucalyptus pilularis* from seven contrasting sites. For. Ecol. Man. 204, 113-127.

MUINONEN, E., MATAMA, M., HYPPÄNEN, H., VAINIKAINEN, V., 2001. Forest stand characteristics estimation using a most similar neighbour approach and image spatial structure information. Remote Sensing of Environment 78, 223-228.

MUND, M., KUMMETZ, E., HEIN, M., BAUER, G.A., SCHULZE, E.D., 2002. Growth and carbon stocks of a spruce forest chronosequence in central Europe. For. Ecol. Man. 171 (3), 275-296.

MUUKKONEN, P., LEHTONEN, A., 2004. Needle and branch biomass turnover rates of Norway spruce (Picea abies). Canadian Journal of Forest Research 34, 2517-2527.

NADARAYA, E. A., 1964. On estimating regression. Theory of Probability and Ist Applications, 9, 141-142.

NEUMANN, M., JANDL, R., 2005. Derivation of locally valid estimators of the aboveground biomass of Norway spruce. Eur. J. Forest Res. 124, 125-131.

NIELSEN, C. C. N., 1990. Einflüsse von Pflanzenabstand und Stammzahlhaltung auf Wurzelform, Wurzelbiomasse, Verankerung sowie auf die Biomassenverteilung im Hinblick auf die Sturmfestigkeit der Fichte. Schriften aus der Forstlichen Fakultät Göttingen und der Niedersächsischen Forstlichen Versuchsanstalt, Band 100. Sauerländer's Verlag, Frankfuhrt a.M.

NIEMZ, P., 1993. Physik des Holzes und der Holzwerkstoffe. DRW-Verlag Leinfelden-Echterdingen. 243 S.

NIESCHULZE, J., 2003. Regionalization of Variables of Sample Based Forest Inventories at the District Level. Dissertation zur Erlangung des Doktorgrades der Fakultät für Forstwissenschaften und Waldökologie der Georg-August Universität Göttingen. 131 S. http://webdoc.sub.gwdg.de/diss/2003/nieschulze/

NIESCHULZE, J., BÖCKMANN, T.H., NAGEL, J., SABOROWSKI, J., 2005. Herleitung von einzelbestandesweisen Informationen aus Betriebsinventuren für die Zwecke der Forsteinrichtung. Allg. Forst- u. J.-Ztg. 176, 169-176.

NIGGEMEYER, P., 1999. Schätzung von Durchmesserverteilungen mit der *k*-nearest neighbour Methode. Diplomarbeit an der Fakultät für Forstwissenschaften und Waldökologie der Georg-August-Universität Göttingen, 71 S.

NIGGEMEYER, P., SCHMIDT, M., 1999. Estimation of the diameter distributions using the k-nearest neighbour method. In: Pukkala, T., Eerikäinen, K. (eds). Growth and yield modelling of tree plantations in South and East Africa. University of Johensuu, Faculty of Forestry. Research Notes 97, 195-209.

NIKLAS, K.J., 1994. Plant Allometry. University of Chicago Press, Chicago.

NIKLAS, K.J., 1994. The allometry of safety-factors for plant height. American Journal of Botany 81(3), 345-351.

NIKLAS, K.J., 2004. Plant allometry: is there a grand unifying theorem? Biol. Rev. 79, 871-889.

NIKLAS, K.J., ENQUIST, B.J., 2002. On the Vegetative Biomass Partitioning of Seed Plant Leaves, Stems, and Roots. Am. Nat. 159, 482-497.

OOHATA, S., SHINOZAKI, K., 1979. A statistical model of plant form – further analysis of the pipe model theory. Jap. J. Ecol. 29, 323-335.

PAIN, O., BOYER, E., 1996. A whole individual tree growth model for Norway Spruce. In: NEPVEU, G. (Ed.), 1996. Proceedings of the Second Workshop "Connection between Silviculture and Wood Quality through Modelling Approaches and Simulation Softwares". Berg-en-Dal, Kruger National Park, South Africa, August 26-31, 1996. Publication Equipe de Recherches sur la Qualité des Bois 1997/7, December. INRA-Nancy, France, 450 S., 13 - 23.

PARRESOL, B.R., 1999. Assessing Tree and Stand Biomass: A Review with Examples and Critical Comparisons. For. Sci. 45 (4), 573-593.

PARRESOL, B.R., 2001. Additivity of nonlinear biomass equations. Can. J. For. Res. 31, 865-878.

PELLINEN, P., 1986. Biomasseuntersuchungen im Kalkbuchenwald. Dissertation zur Erlangung des Doktorgrades der Fakultät für Forstwissenschaften und Waldökologie der Georg-August Universität Göttingen.

POEPPEL, B., 1989. Untersuchungen der Dendromasse in mittelalten Fichtenbeständen. Diplomarbeit, Technische Universität Dresden, 66 S.

PRETZSCH, H., 2001. Modellierung des Waldwachstums. Parey Buchverlag Berlin, 341 S.

PRETZSCH, H., BIBER, P., 2005. A Re-Evaluation of Reineke's Rule and Stand Density Index. For. Sci. 51(4), 305-320.

PROTOKOLL von Kyoto zum Rahmenübereinkommen der Vereinten Nationen über Klimaänderungen, 1997. 23 S.

RADEMACHER, P., 2002: Ermittlung der Ernährungssituation, der Biomasseproduktion und der Nährelementakkumulation mit Hilfe von Inventurverfahren sowie Quantifizierung der Entzugsgrößen auf Umtriebsebene in forstlich genutzten Beständen. Habilitationsschrift, Univ. Göttingen.

RADEMACHER, P., 2004. Abschlussbericht 1999 - 2003 zum BMBF-Verbund-forschungsvorhaben: Indikatoren und Strategien für eine nachhaltige, multifunktionelle Waldnutzung - Fallstudie Waldlandschaft Solling. Berichte des Forschungszentrums Waldökosysteme: Reihe B, 71.

RAISCH, W (Ed.), 1983. Bioelementverteilung in Fichtenökosystemen der Bärhalde (Südschwarzwald). Selbstverlag, Freiburg.

RICHTER, J.P., 1970. The notebooks of Leonardo da Vinci 1452–1519.

ROSENBAUM, K.L., SCHOENE, D., MEKOUAR, A., 2004. Climate change and forest sector. Possible national and subnational legislation. FAO Forestry Paper 144. Food and Agriculture Organisation of the United Nations. Rome.

ROSENBAUM, K.L., SCHOENE, D., MEKOUAR, A., 2004. Climate Change and the forest sector. Possible national and subnational legislation. FAO Forestry Paper 144 (2004), 73 S.

SANTA REGINA, I., TARAZONA, T., 2001. Organic matter and nitrogen dynamics in a mature forest of common beech in the Sierra de la Demanda, Spain. Ann. For. Sci. 58, 301-314.

SACHSSE, H., 1984. Einheimische Nutzhölzer und ihre Bestimmung nach makroskopischen Merkmalen. Pareys Studientexte 44. Verlag Paul Parey, Hamburg und Berlin.

SCHÖNE, D., SCHULTE, A., 1999. Forstwirtschaft nach Kyoto: Ansätze zur Quantifizierung und betrieblichen Nutzung von Kohlenstoffsenken. Forstarchiv 70, 167-176.

SHARMA, S.C., 1992. Untersuchungen über die Dendromasse der Baumart Fichte (*Picea abies* (L.) Karsten) im Tharandter Wald. Dissertation, Technische Universität Dresden, 150 S.

SHEPARD, D., 1968. A two-dimensional interpolation function for irregularly spaced data. Proceedings of the 23rd National Conference of the ACM, 517-523.

SIRONEN, S., KANGAS, A., MALTAMO, M., KANGAS, J., 2001. Estimating individual tree growth with the k-nearest neighbour and k-most similar neighbour methods. Silva Fennica 35(4), 453-467.

SIRONEN, S., KANGAS, A., MALTAMO, M., KANGAS, J., 2003. Estimating individual tree growth with nonparametric methods. Can. J. For. Res. 33, 444-449.

SLOBODA, B., GAFFREY, D., 1999. Dynamik der Stammorphologie. Project report. Institute of Forest Biometry and Informatics, University of Göttingen.

SNELL, O., 1892. Die Abhängigkeit des Hirngewichts von dem Körpergewicht und den geistigen Fähigkeiten. Arch. Psychiatr. 23, 436-446.

SNOWDON, P., EASMUS, D., GIBBONS, P., KHANNA, P., KEITH, H., RAISON, J., KIRSCHBAUM, M., 2000. Synthesis of Allometrics, Review of Root Biomass and Design of Future Woody Biomass Sampling Strategies. National carbon accounting system, technical report no. 17. Australian Greenhouse Office, Canberra, 136 S.

SPENCER, H., 1864. The principles of biology, vol. 1. Williams and Norgate, London.

SPRUGEL, D.G., 1983. Correcting for bias in log-transformed allometric equations. Ecology 64, 209-210.

SPRUGEL, D.G., 2002. When branch autonomy fails: Milton's Law of resource availability and allocation. Tree Physiology 22, 1119–1124.

STAUPENDAHL, K., 2006. kNN-Biomass. *k*-NN Softwaremodul zur Schätzung variabler Zielgrößen auf Einzelbaumebene. Argus Forstplanung.

STÜMER, W., KÖHL, M., 2005. Kombination von terrestrischen Aufnahmen und Fernerkundungsdaten mit Hilfe der k-Nächste-Nachbarn-Methode zur Klassifizierung und Kartierung von Wäldern. Fotogrammetrie Fernerkundung Geoinformation 1/2005, 23-36.

TEMESGEN, H., 2003. Estimating Stand Tables from Aerial Attributes: a Comparison of Parametric and Most Similar Neighbour Methods. Scan. J. For. Res. 18(3), 1-10.

TEMESGEN, H., GADOW, K.V., 2004. Generalized height–diameter models—an application for major tree species in complex stands of interior British Columbia. Eur. J. Forest Res. 123(1), 45-51.

TER-MIKAELIAN, M.T., KORZUKHIN, M.D., 1997: Biomass equations for sixty-five north American tree species. For. Ecol. Manage. 97; 1-24.

THOMPSON, D.W., 1917. On growth and form. Cambridge University Press, Cambridge.

TOMMOLA, M., TYNKKYEN, M., LEMMETTY, J. HARSTELA, P., SIKANEN, L., 1999. Estimating the Characteristics of a Marked Stand Using k-Nearest-Neighbour Regression. Journal of Forest Engeneering vol. 10 no. 2, 75-81.

TOMPPO, E., 1991. Satellite imagery-based national inventory of Finland. International Archives of Photogrammetry and Remote Sensing. 28: 7-1, 419-424.

TOMPPO, E., GOULDING, C., KATILA, M., 1999. Adapting Finnish multi-source forest inventory techniques to the New Zealand preharvest inventory. Scan. J. For. Res. 14, 182-192.

TOMPPO, E., HALME, M., 2004. Using coarse scale forest variables as ancillary information and weighting of variables in k-NN estimation: a genetic algorithm approach. Rem. Sens. of Env. 92, 1-20.

TRITTON, L.M., HORNBECK, J.W., 1982. Biomass Equations for Major Tree Species of the Northheast. Northeastern Forest Experiment Station,USDA Forest Service, General Technical Report NE-69, 46 S.

UNFCCC, 1992. Rahmenübereinkommen der Vereinten Nationen über Klimaänderungen, 1992.

VALENTINE, H.T., 1988. A carbon-balance model of stand growth: a derivation employing pipe-model theory and the self thinning rule. Ann. Botany 62, 389-396.

VALENTINE, H.T., 1999. Estimation of net primary productivity of even-aged stands with a carbon-allocation model. Ecological Modelling 122, 139-149.

VALENTINE, H.T., MÄKELÄ, A., 2005. Bridging process-based and empirical approaches to modelling tree growth. Tree Physiology 25, 769-779.

VAN CAMP, N., VANDE WALLE, I., MERTENS, J., DE NEVE, S., SAMSON, R., LUST, N., LEUMER, R., BOECKX, P., LOOTENS, P., BEHEYDT, D., MESTDAGH, I., SLEUTEL, S., VERBEECK, H., VAN CLEEMPUT, O., HOFMAN, G., CARLIER, L., 2004. Inventory-based carbon stock of Flemish forests: a comparison of European biomass expansion factors. Ann. For. Sci. 61 (7), 677-682.

VAN NOORDWIJK, M. MULIA, R., 2002. Functional branch analysis as tool for fractal scaling above- and belowground trees for their additive and non-additive properties. Ecological Modelling, Vol. 149 (1-2), 41 – 51.

VANNINEN, P., MÄKELÄ, A., 2005. Carbon budget for Scots pine trees: effects of size, competition and site fertility on growth allocation and production. Tree Physiology 25, 17-30.

VANNINEN, P., YLITALO, H., SIEVÄNEN, R., MÄKELÄ, A., 1996. Effects of age and site quality on the distribution of biomass in Scots pine (*Pinus sylvestris* L.). Trees 10, 231-238.

VYSCOT, M., 1981. Biomass of the tree layer of spruce forest in the Bohemian Uplands. Academia Publishing House of Czechoslovak Academy of Science, Praha.

WAGACHA, P. W., 2003. Instance-Base Learning: k-Nearest Neighbour. Notes for ICS320 Foundations of Learning and Adaptive Systems. Institute of Computer Science University of Nairobi.

WATSON, G. S., 1964. Smooth regression analysis. Sankhva: The Indian Journal of Statistics, Series A, 26, 359-372.

WEBER, R., 1998. Statische und dynamische Evaluation von Prognosen. In: ZA-Information 43, 111-123.

WENK, G., ANTANAITIS, V., SMELKO, S., 1990. Waldertragslehre. Deutscher Landwirtschaftsverlag. Berlin.

WEST, G.B., BROWN, J.H., ENQUIST, B.J., 1997. A general model for the origin of allometry scaling laws in biology. Science 276, 122-126.

WEST, G.B., BROWN, J.H., ENQUIST, B.J., 1999. The fourth dimension of live: fractal geometry and allometric scaling of organisms. Science 284, 167-169.

WEST, G.B., BROWN, J.H., ENQUIST, B.J., 1999a. A general model for the structure and allometry of plant vascular systems. Nature 400, 664-667.

WETTSCHERECK, D., AHA, D.W., 1995. Weighting Features. In: VOLOSO, M., AAMODT, A. (Ed.), 1995. Proceedings of the First International Conference on Case-Based Reasoning (ICCBR). Springer. 347-358.

WETTSCHERECK, D., DIETTERICH, T., 1994. Locally Adaptive Nearest Neighbour Algorithms. In: Advances in Neural Information Processing Systems 6. Morgan Kaufmann Publishers, San Mateo, CA.

WIANT, H.J., CASTANEDA, F., SHEETZ, C., COLANINNO, A., DEMOSS, J., 1979. Equations for predicting weights of some Appalachian hardwoods. West Virginia Univ. Agric. and For. Exp. Sta., Coll. of Agric. and For. West Virgina For. Notes, No 7.

WILHELMSEN, G., VESTJORDET, E., 1974. Preliminary Dry Wood Weight Tables for Merchendable Stems and Stands of Norway Spruce (*Picea abies* (L.) Karst.) in Norway. Reports of the Norwegian Forest Research Institute 31.5, 184-240.

WILSON, B.F., ARCHER, R.R., 1979. Tree design: Some biological solutions to mechanical problems. Bioscience 29 (5), 293-298.

WIRTH, C., SCHULZE, E.-D., SCHWALBE, G., TOMCZYK, S., WEBER, G., WELLER, E., 2004. Dynamik der Kohlenstoffvorräte in den Wäldern Thüringens. Mitteilungen 23/2004. Thüringer Ministerium für Landwirtschaft, Naturschutz und Umwelt.

WIRTH, C., SCHUMACHER, J. 2002. Allometric relationships for Norway spruce in Central Europe - a meta-analysis approach towards age- and site-specific expansion factors and their uncertainties. COST E21 expert group meeting on biomass expansion factors, Besalu, Spain, Presentation.

WIRTH, C., SCHUMACHER, J., SCHULZE, E.-D., 2003. Generic biomass functions for Norway spruce in Central Europe – a meta analysis approach toward prediction and uncertainty estimation. Tree Physiology 24, 121-139.

WISSENSCHAFTLICHER BEIRAT DER BUNDESREGIERUNG GLOBALE UMWELTVERÄNDERUNG (WGBU), 1998. Die Anrechnung biologischer Quellen und Senken im Kyoto-Protokoll: Fortschritt oder Rückschlag für den globalen Umweltschutz? Sondergutachten. 86 S.

WITTHAKER, R.H., BORMANN, F.H., LIKENS, G.E., SICCAMA, T.G., 1974. The Hubbard Brook Ecosystem Study: Forest Biomass and Production. Ecological Monographs 44(2), 233-254.

XIAO, C.-W., CEULEMANS, R., 2004. Allometric relationships for below- and aboveground biomass of young Scots pines. For. Ecol. Man. 203, 177-186.

Xiao, C.-W., Yuste, J.C., Janssens, I.A., Roskams, P., Nachtergale, L., Carrara, A., Sanchez, B.Y., Ceulemans, R., 2003. Above- and belowground biomass and net primary production in a 73-year-old Scots pine forest. Tree Physiology 23, 505-516.

Young, H. E., 1976: A summary and analysis of weight tables studies. In: Young, H.E. (Hrsg.) IUFRO Oslo Biomass Studies. Univ. of Maine at Orno.

Zianis, D., Mencuccini, M., 2004. On simplifying allometric analyses of forest biomass. For. Ecol. Man. 187, 311-332.

Zianis, D., Mukkonen, P., Mäkipää, R., Mencuccini, M., 2005. Biomass and Stem Volume Equationns for Tree Species in Europe. Silva Fennica Monographs 4, 63 S.

Zucchini, W., Schmidt, M., Gadow, K.v., 2001. A Model fort he Diameter-Height Distribution in an Uneven-Aged Beech Forest and a Method to Assess the Fit of Such Models. Silva Fennica 35 (2), 169-183.

226 Einträge

8 Anhang I, Exkurs

Als einfaches Beispiel zur Veranschaulichung der Diskrepanz zwischen einem Prozessmodell und empirischer Datenanalyse mit einer für diesen Vergleich ungeeigneten Messgröße soll hier eine vereinfachte Modellannahme am Beispiel der geometrischen Form eines Kegels dargestellt werden.

Angenommen die Höhe eines Kegels sei durch eine allometrische Beziehung der Form

$$H = k\,D^{b'} \qquad \text{bzw.:} \quad \ln H = b' \cdot \ln D + \ln k \tag{31}$$

gegeben, (für D in Zentimeter und H in Meter).

Das Volumen eines Kegels ist gegeben mit:

$$V_{cone} = \frac{\pi}{12} D^2\, h \tag{32}$$

Hieraus ergibt sich als funktionaler Zusammenhang zwischen Durchmesser und Volumen:

$$V_{cone} = \frac{\pi}{12} k\, D^{2+b'} \tag{33}$$

Nimmt man weiter an, dass die Raumdichte des beschriebenen Körpers ρw gleichmäßig verteilt sei, so führt dies zu:

$$M_{cone} = \frac{\pi}{12} k\, \rho w\, D^{2+b'} \tag{34}$$

Soll M_{cone} bei den gegebenen Eingangsgrößen in Kg ausgedrückt werden, muss das Ergebnis um den Faktor 0,1 korrigiert werden. In Form der Grundgleichung der Allometrie ausgedrückt ergibt sich somit:

$$M_{cone}[Kg] = a\, D^{b}, \quad \text{wobei} \qquad a = \frac{\pi}{12} k\, \rho w \cdot 0{,}1 \qquad \text{und} \qquad b = 2 + b' \tag{35}$$

Durch Logarithmierung linearisiert ergibt sich die Geradengleichung

$$\ln M = b \cdot \ln D + \ln a\,, \tag{36}$$

wobei die Allometriekonstante b das Verhältnis zweier relativer Wachstumsraten als $\delta M/M = b^{*}\delta D/D$ beschreibt und ln a der jeweilige Achsenabschnitt ist. Durch die Multiplikation mit dem konstanten Faktor a (Integrationskonstante) werden in Funktion (34) alle Ordinatenwerte im Verhältnis $1/a$ gestaucht (für $a < 1$). Die Grundgleichung der Allometrie beschreibt insofern das Verhältnis des Wachstums einer funktionalen Messgröße zum Wachstum einer anderen funktionalen Messgröße.

Wichtig sei anzumerken, dass diese Zusammenhänge ausschließlich für den Durchmesser der Grundfläche eines Kegels (D_0) gelten! Würde man an Kegeln unterschiedlicher Größe einen Durchmesser der Querschnittsflächen in einer absoluten Höhe als unabhängige Variable verwenden, so würde man die obigen Annahmen nicht bestätigt finden, da dieser Durchmesser keine funktionale Messgröße des Kegels ist. Man würde mit diesem Durchmesser nicht die Grundflächenänderung alleine, sondern auch die Formveränderung des Körpers in einer absoluten Höhe erfassen. Hierzu sei ein konkretes Beispiel mit folgenden Eingangsgrößen gegeben: $k = 1{,}9$; $b' = 2/3$ und $\rho w = 0{,}48$ g/m^3. Hieraus folgt für die Integrationskonstante a der Funktion (35):

$$a = \pi / 12 \cdot 1{,}9 \cdot 0{,}48 \cdot 0{,}1 = 0{,}02388 ,$$

sowie für die Allometriekonstante b:

$$b = 2 + 2 / 3 = 2{,}6667 .$$

Abbildung 8-1 stellt diesen Zusammenhang grafisch dar:

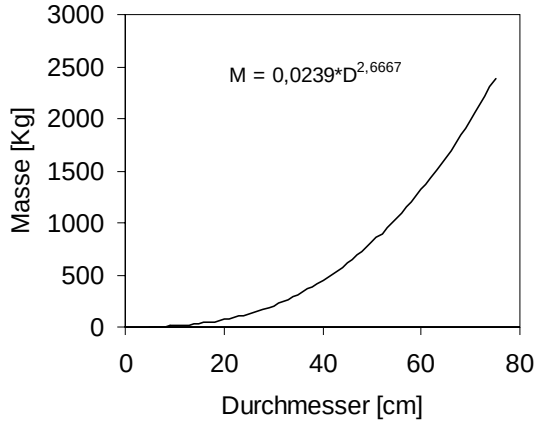


Abbildung 8-1. Funktionaler Zusammenhang zwischen Durchmesser und Masse eines Kegels bei einer gegebenen allometrischen Höhenfunktion.

Berechnet man für verschieden große Kegel den Durchmesser der Querschnittsflächen in einer absoluten Messhöhe (1,3 m) mit:

$$D_{1,3m} = D_0 \cdot \frac{(H - 1{,}3)}{H} \tag{37}$$

und verwendet diesen Bezugsdurchmesser als unabhängige Variable, so erhält man als Schätzung für die Koeffizienten einer einfachen linearen Regressionsgeraden auf Grundlage der logarithmisch transformierten Wertepaare b = 2,4817 und ln a = -2,8861 bzw. a = 0,0558 (siehe Abbildung 8-2)

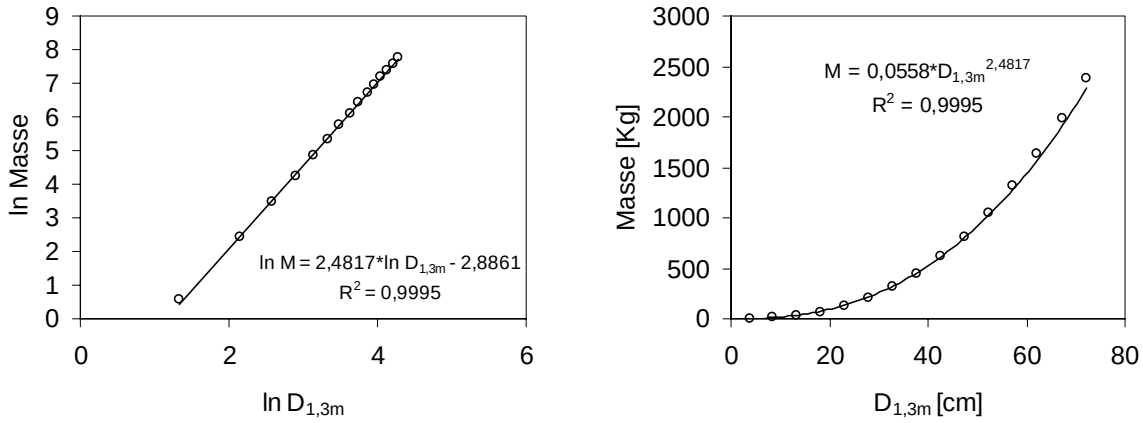


Abbildung 8-2. Lineare Regression zwischen den logarithmisch transformierten Wertepaaren eines Durchmessers in absoluter Höhe ($D_{1,3m}$) und Masse (links) und Darstellung auf dem metrischen Skalenniveau (rechts).

Aus obiger Abbildung wird deutlich, dass die resultierende Abweichung von der Linearität bei Verwendung eines Durchmessers in absoluter Messhöhe aufgrund des homogenisierenden Effektes der logarithmischen Transformation der Werte in einer empirischen Datenanalyse kaum festzustellen ist. Während in Abbildung 8-2 eine generierte Punktwolke dargestellt ist, würde in einer empirischen Datengrundlage die vorhandene Streuung der Werte diesen Nachweis zusätzlich erschweren. Der hohe Erklärungsanteil des Regressionsmodells von über 99 % würde daher ohne Kenntnis über den genauen Funktionalen Zusammenhang kaum Anlass geben die Modellformulierung in Frage zu stellen.

Der Vorteil der logarithmischen Transformation ist in diesem Fall dadurch gegeben, dass eine eventuelle Heteroskedastizität der Datengrundlage eliminiert wird und gleichzeitig eine einfache lineare Regression verwendet werden kann. Allerdings entsteht durch die Rücktransformation der Werte auf das metrische Skalenniveau ein systema-

tischer Fehler, der die Berechnung eines Korrekturfaktors für die Integrationskonstante a notwendig macht. Abgesehen davon ist jedoch offensichtlich, dass die Schätzung für die Allometriekonstante b wie erwartet nicht mit dem theoretisch hergeleiteten Skalierungsfaktor übereinstimmt, da die verwendete unabhängige Variable nicht konform mit den zugrunde gelegten Modellannahmen ist.

Verwendet man anstelle des Durchmessers in absoluter Messhöhe jedoch eine relative Messhöhe (z.B. D_{01} in 10 % der Gesamthöhe), so wäre dieser Durchmesser wiederum eine funktionale Messgröße und würde somit nicht gegen die Interpretation der Grundgleichung der Allometrie verstoßen. Gleichzeitig würde jedoch die Annahme (34) verletzt werden, so dass zwar die erwartete Allometriekonstante b nachgewiesen werden kann, die Integrationskonstante a jedoch nicht wie oben beschrieben nachvollziehbar ist (siehe Abbildung 8-3).

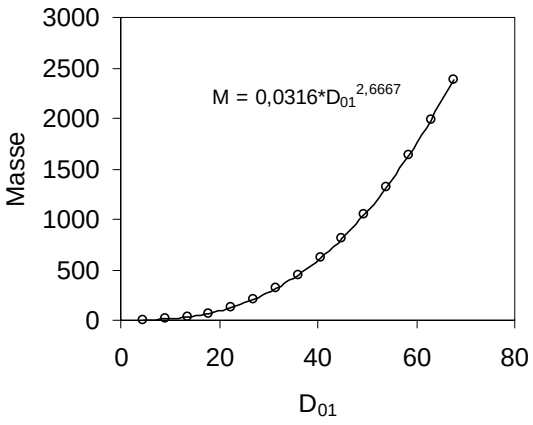


Abbildung 8-3. Allometrische Beziehung zwischen dem Durchmesser in relativer Höhe (hier 10 % der Höhe) D_{01} und der Masse.

Um den Aspekt der Autokorrelation der Parameter b und a darstellen zu können, sei eine folgende erweiterte Annahme zu Formel (30) getroffen: Aufgrund mechanischer bzw. physikalischer Zwänge sei die Höhenzunahme des Kegels durch eine Grenzbeziehung limitiert, so dass b' und ln k in einer linearen Beziehung zueinander stehen:

$$\ln k = m\,b' + n \tag{38}$$

Die Implikation dieses Zusammenhangs ist evident wenn man unterstellt, dass der Höhenzuwachs eines Kegels durch die absolute Höhe bei einem gegebenen Durchmesser limitiert ist. Dieses wiederum kann als angepasste Allokation des Volumenzuwachses unter der Einschränkung einer Erhaltung der Stabilität des Körpers mit zunehmender Masse verstanden werden. Hieraus folgt nach (34) dass ebenso b und ln a einen

funktionalen linearen Zusammenhang aufweisen. Abbildung 8-4 zeigt den Zusammenhang zwischen den Parametern der Höhenfunktion und der Massenfunktion bei einer unterstellten linearen Beziehung zwischen ln k und b'.

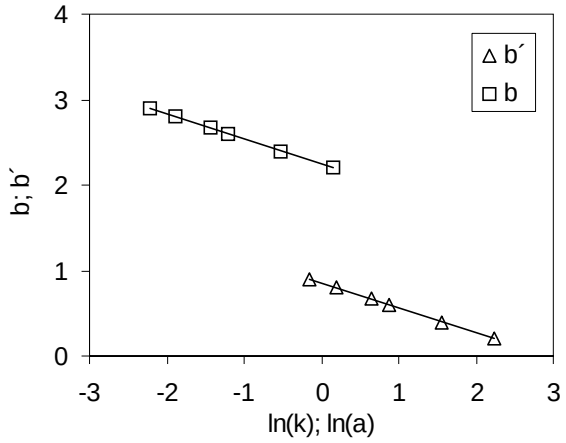


Abbildung 8-4. Beziehung zwischen den Parametern der Höhen- und Massenfunktion bei einer unterstellten linearen Abhängigkeit zwischen ln k und b'.

Der Zusammenhang zwischen der Allometriekonstante b und der Integrationskonstante a lässt sich dementsprechend mit Hilfe einer Exponentialfunktion beschreiben, wobei der Exponent hierbei der unterstellten Gradensteigung m in (38) entspricht. Abbildung 8-5 zeigt die Beziehung der Parameter in metrischer Skalierung (links), sowie die resultierenden Massenfunktionen (rechts).

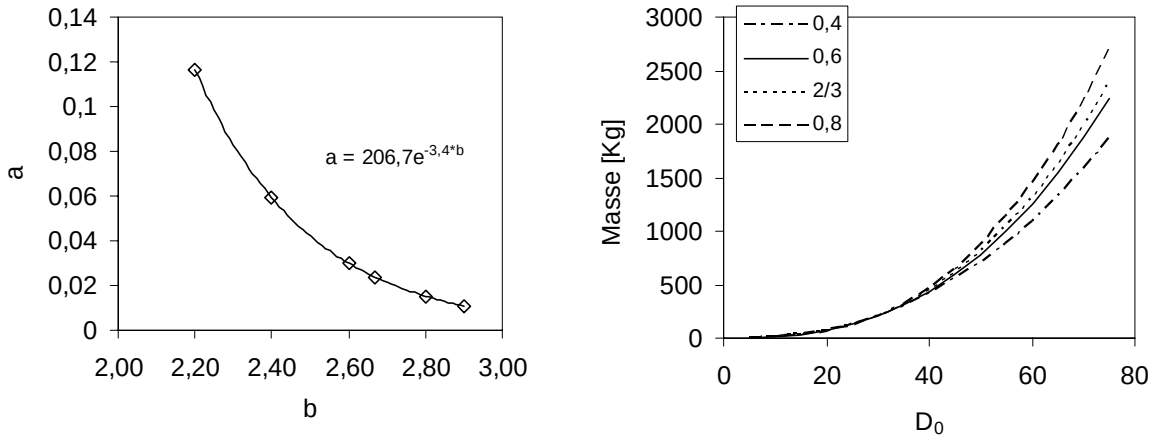


Abbildung 8-5. Negativ-exponentieller Zusammenhang zwischen den Parametern a und b der allometrischen Massenfunktion (links) und resultierende Beziehung zwischen Durchmesser (D_0) und Masse bei verschieden gewählten Parametern b' der Höhenfunktion.

9 Anhang II, Datengrundlagen

Baumart	N	BHD			Höhe		
		Min	Max	Mittel	Min	Max	Mittel
Abies Balsamea	30	2,5	28,3	14,99	2,9	17,0	10,22
Acer Rubrum	53	1,3	42,1	18,40	3,0	24,6	15,49
Acer Saccharum	55	1,2	55,2	19,36	2,3	26,5	15,68
Betula Alleghaniensis	49	1,6	72,0	20,68	2,8	26,0	15,03
Betula Papyrifera	37	1,1	34,1	15,21	2,7	20,1	12,68
Fagus Grandifolia	14	8,5	43,0	24,76	9,8	23,4	19,20
Fagus Sylvatica	221	0,8	70,2	25,33	1,9	39,1	20,75
Larix Decidua	10	19,5	48,3	32,02	20,4	25,9	23,89
Picea Abies	578	2,3	73,8	20,67	2,5	38,4	17,16
Picea Glauca	24	1,5	29,5	14,88	1,9	21,2	13,04
Picea Mariana	24	2,2	30,2	15,28	2,1	17,5	11,17
Picea Rubens	37	1,2	31,3	15,18	1,6	19,5	11,94
Pinus Sylvestris	270	4,6	58,0	18,56	3,2	32,0	13,68
Populus Grandidentata	23	7,8	33,8	19,30	9,7	19,4	13,95
Populus Tremuloides	46	0,8	36,0	13,84	2,0	20,4	10,72
Quercus Petraea	28	9,8	77,1	35,79	11,0	33,6	24,74

Abbildung 9-1. Übersicht der vorhandenen Einzelbaumdaten, mit Durchmesser- und Höhenbereich für verschiedene Baumarten.

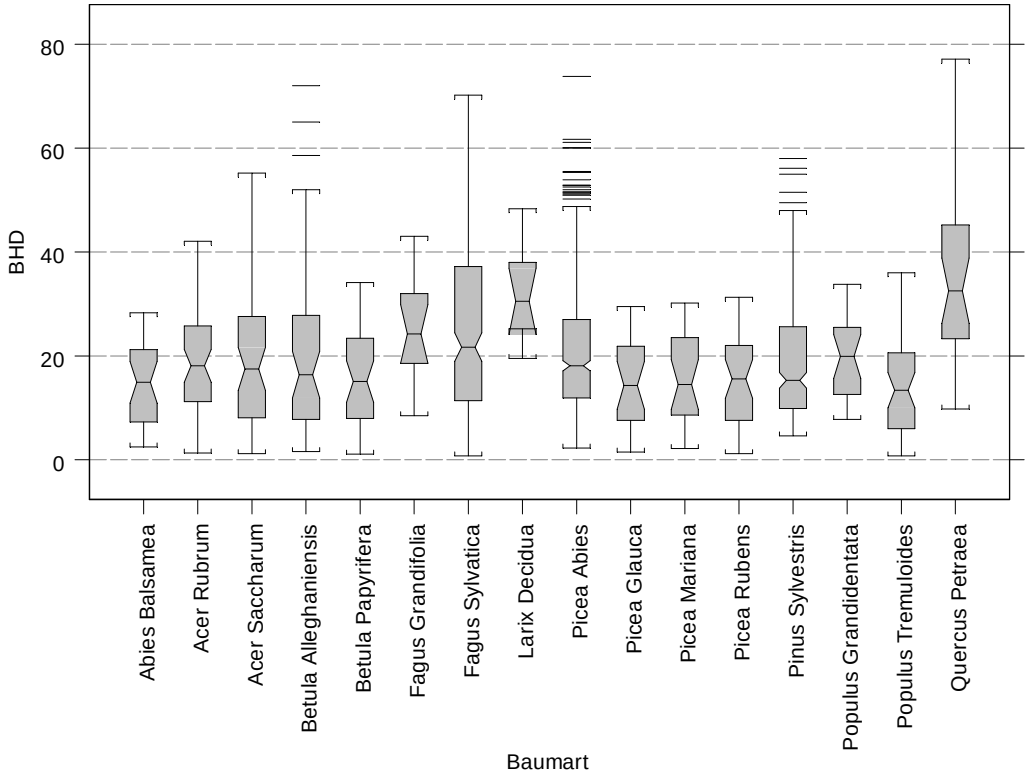


Abbildung 9-2. Box-Plots (Spannweite, 25- und 75 % Quartil, Median) der Durchmesserbereiche der vorhandenen Datengrundlagen getrennt nach den einzelnen Baumarten.

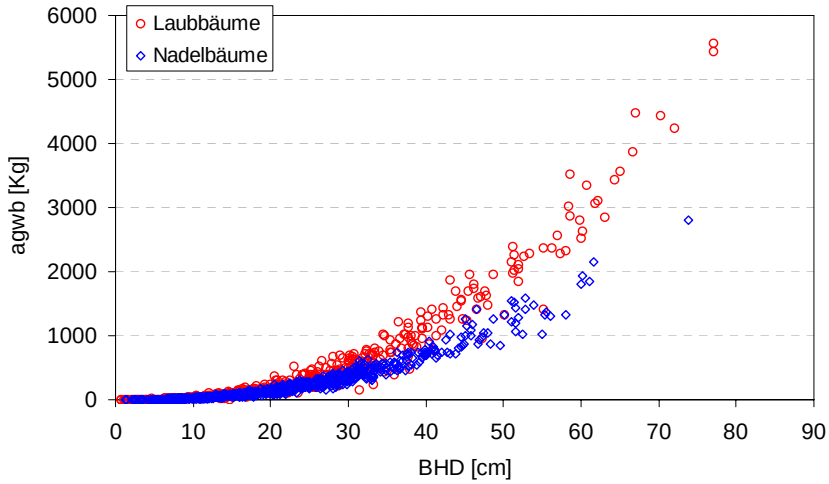


Abbildung 9-3. Oberirdische holzige Biomasse (*agwb*) über dem BHD getrennt nach Nadel- und Laubbäumen (n= 973; n= 528).

10 Anhang III

Für die unter 3.3.1 durchgeführten Teilauswertungen von Fichten und Kieferndaten sollen im Folgenden noch einige zusätzliche Datenanalysen sowie die Ergebnisse der Regressionsanalyse zur Ableitung der verwendeten Referenzmodelle dargestellt werden. Die hier aufgeführten Analysen beziehen sich zum einen auf die im Methodenteil dargestellten grundsätzlichen Möglichkeiten eine Ähnlichkeit zwischen Instanzen zu quantifizieren und zum andere auf eine zusätzliche beispielhafte Teilauswertung, in der anstelle einer oberirdischen Gesamtmasse ein einzelnes Biomassekompartiment untersucht wurde.

10.1 Ableitung der Referenzmodelle

Tabelle 10-1. Geschätzte Regressionskoeffizienten für ein einfaches lineares Modell mit dem BHD und Baumhöhe als unabhängige Variablen für auf Grundlage der zufällig ausgeschiedenen *modelling*-Daten für Fichte (n=142). Die Zielgröße ist hier die oberirdische Gesamtbiomasse (agb).

Modell	Koeffizient	Schätzung	Std. Fehler	t-Wert
$\ln agb_i = \ln \alpha_i + \beta \ln BHD_i + \chi \left[\frac{h}{BHD}\right]_i + \varepsilon_i$	α	-1,973	0,046	-16,61
	β	2,345	0,024	94,40
	χ	0,055	0,079	0,46

Standardfehler der Residuen: 0,157 bei 140 Freiheitsgraden,
Multiples R^2: 0,986,
F-Statistik: 5106 bei 2; p-Wert=0

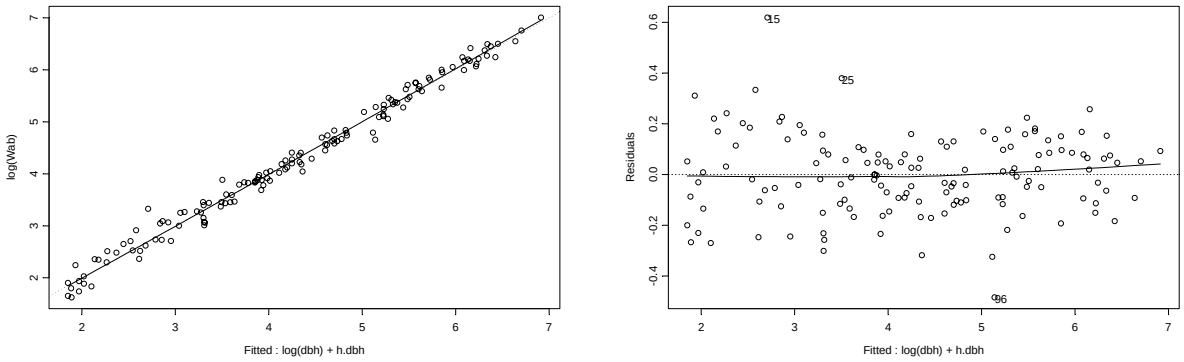


Abbildung 10-1. Beobachtete und geschätzte Biomasse (links) und Residuenplot der Regression (rechts).

Tabelle 10-2. Geschätzte Regressionskoeffizienten für ein einfaches lineares Modell mit dem BHD und Baumhöhe als unabhängige Variablen für auf Grundlage der zufällig ausgeschiedenen *modelling*-Daten für Kiefer (n=145). Die Zielgröße ist hier die oberirdische Gesamtbiomasse (agb).

Modell	Koeffizient	Schätzung	Std. Fehler	t-Wert
$\ln agb_i = \ln \alpha_i + \beta \ln BHD_i + \chi \ln h_i + \varepsilon_i$	α	-2,355	0,048	-48,88
	β	2,202	0,041	53,30
	χ	0,272	0,042	6,39

Standardfehler der Residuen: 0,1187 bei 142 Freiheitsgraden,
Multiples R^2: 0,9923,
F-Statistik: 9197 bei 2; p-Wert=0

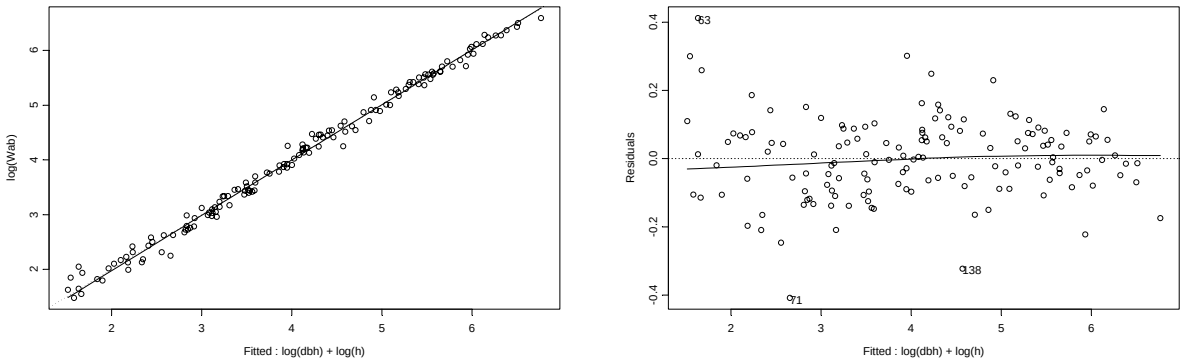


Abbildung 10-2. Beobachtete und geschätzte Biomasse (links) und Residuenplot der Regression (rechts).

10.2 Distanz und Korrelation

Im Methodenteil (2.4.2) wurde bereits darauf hingewiesen, dass eine einfache euklidische Distanzmetrik zwar in der Lage ist, aus der Summe der Variablendifferenzen zweier Instanzen einen Gesamtabstand zu berechnen, die Richtung der einzelnen Abstände dabei jedoch unberücksichtigt bleibt. Hierdurch kann es im Zweifelsfall dazu kommen, dass Instanzen mit einem komplementären Variablenprofil bei gleichen Einzeldifferenzen der Variablen ein gleicher Abstand zugeordnet wird. Eine

Möglichkeit die Kovarianzstruktur der Datengrundlage zu berücksichtige wurde bereits in Form der Mahalanobis Distanz vorgestellt. Diese entspricht jedoch nur in einem normierten und standardisierten Diskriminanzraum der euklidischen Distanz. Im der hier verwendeten Umsetzung der *k*-NN Methode wurde auf eine Normierung und Standardisierung der einzelnen Variablen vor der Distanzberechnung verzichtet. Stattdessen wurden die berechneten Variablenabstände mit einem Vielfachen der Standardabweichung der betreffenden Variablen relativiert.

Ob durch dieses Vorgehen eine Verzerrung innerhalb des Diskriminanzraumes entsteht, oder anders ausgedrückt, ob die so berechneten Abstände konform zum Variablenprofil der Instanzen sind, kann überprüft werden, indem man die Distanzen dem Q-Korrelationskoeffizienten gegenüberstellt. Hierzu wurde beispielhaft der *modelling*-Datensatz der Fichten verwendet (n=142). Ausgehend von dem Abfragepunkt 0;0 (BHD =0 und Höhe =0) wurden die Distanzen zu allen Trainingsdaten mit der gegebenen Parametereinstellung berechnet. Gleichzeitig wurde der Q-Korrelationskoeffizient ausgehend vom Ähnlichsten Nachbarn für alle Instanzen berechnet. Die Gegenüberstellung dieser Werte ist in Abbildung 10-3 dargestellt.

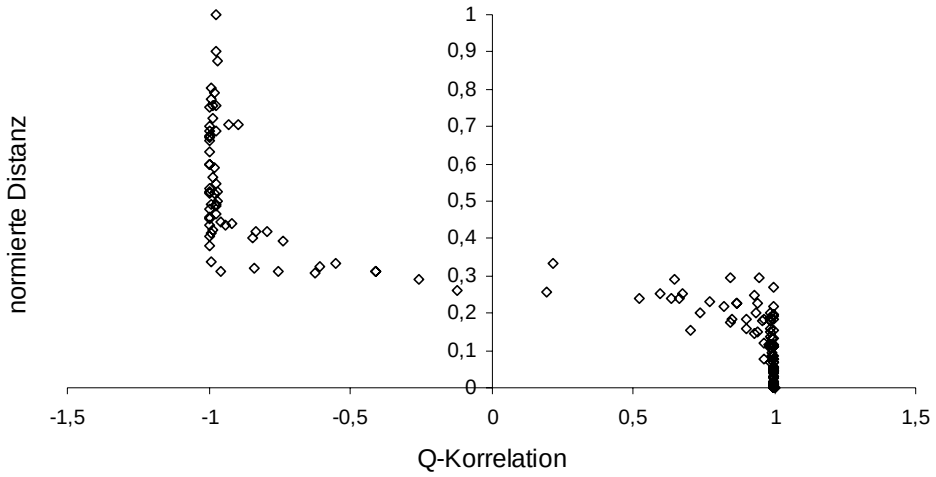


Abbildung 10-3. Beziehung zwischen berechneter (euklidischer-) Distanz und dem Q-Korrelationskoeffizienten der einzelnen Instanzen.

Hierbei ist deutlich zu erkennen, dass der Q-Korrelationskoeffizient für kleine Distanzen nahe bei 1 liegt, was auf ein ähnliches Variablenprofil der Instanzen hindeutet. Mit zunehmendem Abstand verringert sich die Korrelation und kehrt sich für größere

Distanzen um (-1), was auf ein komplementäres Variablenprofil hindeutet. In diesem Fall, in dem lediglich zwei Variablen betrachtet werden, scheint die euklidische Distanz also in der Lage zu sein, auch in Bezug auf das Variablenprofil eine Mustererkennung zu ermöglichen, wenn Nachbarn bis zu einem bestimmten Abstand bzw. eine angepasste Anzahl von Nachbarn verwendet wird.

10.3 Auswertung einzelner Kompartimente

In den im Ergebnisteil dargestellten Teilauswertungen der verschiedenen Datengrundlagen beziehen sich auf aggregierte Biomassen von Einzelbäumen in denen die Einzelmassen der verschiedenen Kompartimente entweder zu einer oberirdischen holzigen Biomasse (*agwb*) oder zur oberirdischen Gesamtbiomasse (*agb*) aufsummiert wurden. Diese zusammengefasste Gesamtmasse entwickelt sich typischerweise sehr viel stabiler als z.B. die Masse einzelner Kompartimente, da die Allokation der Masse auf verschiedene Organe unberücksichtigt bleibt. Unter der Grundannahme, dass Baumindividuen aufgrund der auf ihre Masse einwirkenden physikalischen Kräfte ein kritisches Verhältnis zwischen Masse und Dimension nicht überschreiten können, ist ihre Gesamtmasse stärker limitiert als die Verteilung dieser Masse auf einzelne Kompartimente.

In dieser zusätzlichen Auswertung soll der unter 3.3.1 beschriebene Fichten- und Kieferndatensatz aus der Nationalen Finnischen Waldinventur in Hinblick auf einzelne Kompartimente genauer untersucht werden. Hierzu wird beispielhaft auf die Nadelbiomasse eingegangen, da diese eine besonders große Streuung aufweist und sich daher mit herkömmlichen Regressionsmodellen relativ schwierig abbilden lässt. Abbildung 10-4 zeigt die Biomassewerte des Kieferndatensatzes der einzelnen Kompartimente über dem BHD.

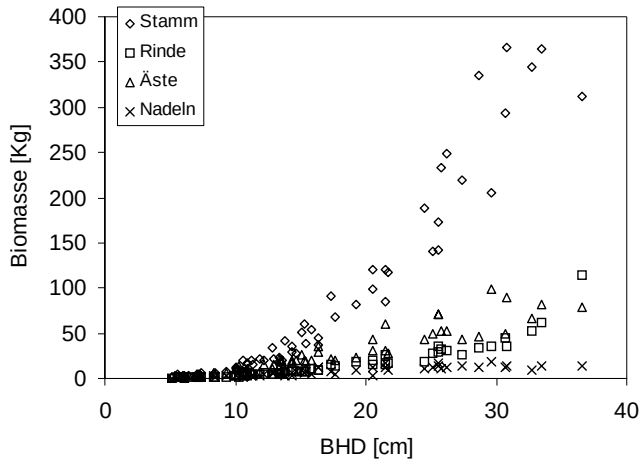


Abbildung 10-4. Biomassewerte der einzelnen Kompartimente Stamm, Äste, Rinde, und Nadeln über dem BHD für den unter 3.3.1 beschriebenen Kiefern *test*-Datensatz (n=60).

Eine Auswertung der Nadelbiomasse der 60 *test*-Bäume mit der k-NN Anwendung führt unter Berücksichtigung der Variablen BHD (w_{bhd}=0,5) und Baumhöhe (w_h=0,5) für den Kiefern zu den in Abbildung 10-5 dargestellten Ergebnissen.

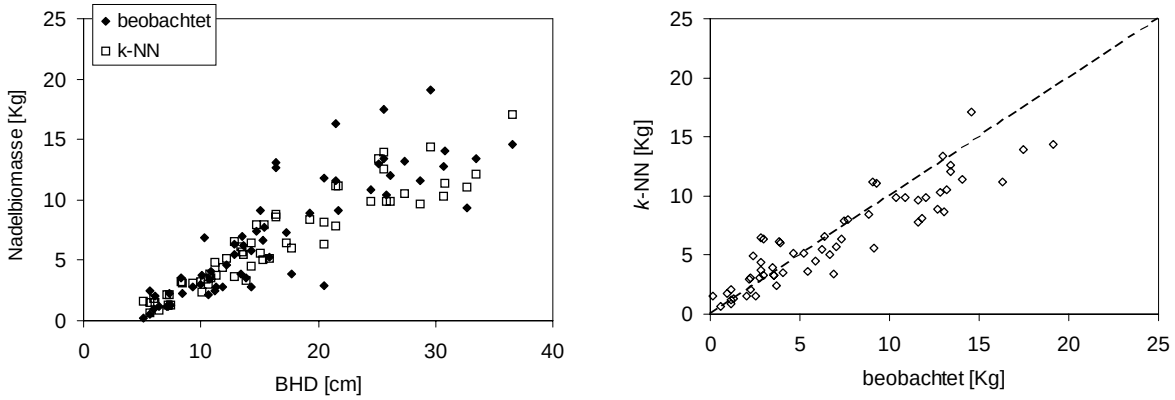


Abbildung 10-5. Beobachtete und geschätzte Nadelbiomasse (links) der *test*-Bäume des Kieferndatensatzes (n=60) über dem BHD und Verhältnis zwischen Schätzungen und Beobachtungen (rechts).

Der erwartete RMSE% aus der Kreuzvalidierung des zufällig ausgeschiedenen Trainingsdatensatzes (n=145) beträgt bei Verwendung von 8 Nachbarn 45,43% (für

c=2 und t=0). Tatsächlich wurde für den *test*-Datensatz ein RMSE% von 32,48% mit einem mittleren Fehler von 0,56 berechnet.

Die hohe Variabilität der Nadelbiomasse über dem BHD lässt erwarten, dass ein einfaches allometrisches Modell auf Grundlage des BHDs kaum in der Lage währe die Varianz dieser Datengrundlage adäquat abzubilden bzw. zu erklären. Die Verwendung der k-NN Methode hat hierbei zwei grundlegende Vorteile. Zum einen ist die Varianz der Schätzung höher und somit realitätsnäher, da in diesem Fall kein funktionaler Verlauf unterstellt wird, und zum anderen können die Schätzungen den Wertebereich der Trainingsdaten nicht verlassen, da der k-NN Algorithmus nicht in der Lage ist Extrapolationen über die vorhandenen Trainingsdatenpunkte hinaus zu produzieren.

11 Anhang IV

Dargestellt ist der Verlauf des RMSE% und des relativen Bias für den unter 3.4.1 verwendeten Datensatz (Teilauswertung III). Hierbei wurde zusätzlich zum BHD, der Baumhöhe und der Holzdichte auch das Alter der Einzelbäume berücksichtigt. Der Umfang der Trainingsdaten reduziert sich dadurch auf 327 Bäume, da nur für diese Individuen das Alter bekannt ist. Der RMSE% liegt bei einer Anzahl von 8 Nachbarn bei 23,6%.

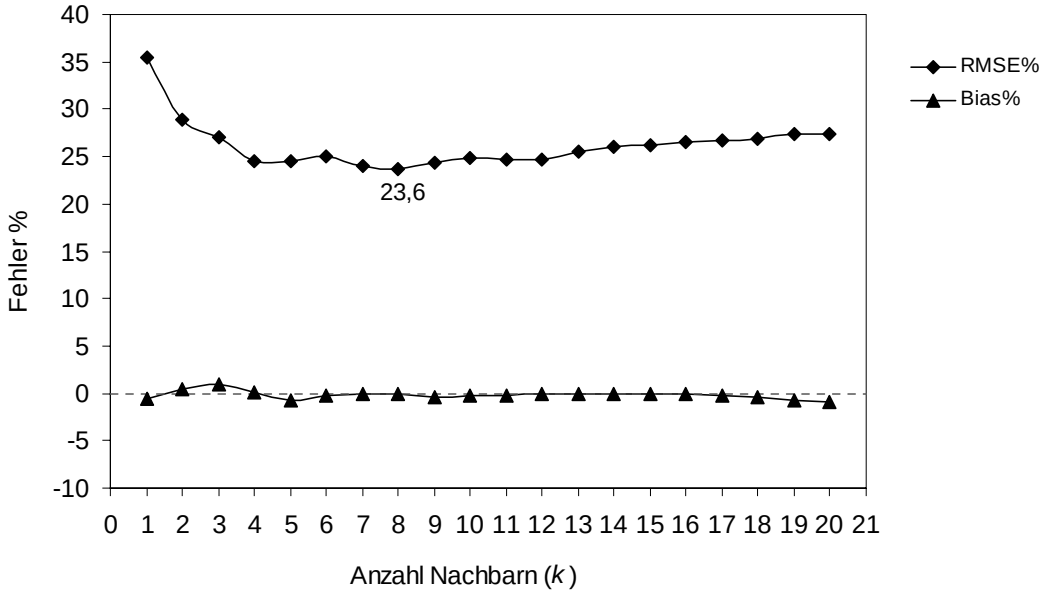


Abbildung 11-1. Verlauf des RMSE% und des relativen Bias über der Anzahl berücksichtigter Nachbarn bei Verwendung des BHD (w_{bhd}=0,5), der Höhe (w_h=0,2), der Holzdichte (w_d=0,1) und des Baumalters (w_a=0,2) für alle europäischen Koniferen (n=327).

12 Anhang V

Tabelle 12-1. Auflistung der in Teilauswertung VI verwendeten Bäume aus Ketterings et al. (2001).

Lokaler Name	Genus	BHD	H	*agwl*	*agb*
Buluh	*Ginetroshesia* spp.	7,6	12,3	8,6	13,3
Kubung	*Mastixia* spp.	8,0	10,5	16,5	18,7
Kayu kacang	*Strombosia* spp.	9,2	10,1	21,6	24,1
Balik angin	*Mollotus* spp.	9,6	8,5	23,5	26,3
Mahang	*Macaranga* spp.	9,9	7,2	14,3	15,9
Nilao	unknown	10,5	11,0	30,7	34,4
Kelat	*Eugenia* spp.	11,5	14,0	43,6	48,5
Mahang	*Macaranga* spp.	12,1	14,6	52,0	56,0
Balik angin	*Mollotus* spp.	12,4	10,5	19,6	25,4
Meranti	*Shorea* spp.	12,4	12,5	24,0	27,2
Mahang	*Macaranga* spp.	15,3	14,3	82,0	90,4
Patang buah	*Baccaurea* spp.	15,6	8,6	47,6	56,9
Balik angin	*Mollotus* spp.	16,9	11,8	92,4	97,3
Mahang	*Macaranga* spp.	18,8	11,0	138,6	147,1
Mahang	*Macaranga* spp.	27,1	22,9	303,5	337,7
Buluh	*Ginetroshesia* spp.	28,3	11,4	178,5	197,4
Buluh	*Ginetroshesia* spp.	30,6	20,2	465,3	491,3
Medang	*Dactylocladus* spp.	32,5	20,6	413,0	435,3
Kedungdung	*Pentaspadon* spp.	32,8	25,4	619,2	662,4
Kubung	*Mastixia* spp.	33,9	20,2	577,0	608,4
Medang	*Dactylocladus* spp.	35,7	32,4	417,7	446,8
Medang	*Dactylocladus* spp.	36,6	28,2	822,5	867,6
Medang	*Dactylocladus* spp.	37,7	23,6	1132,0	1224,9
Medang	*Dactylocladus* spp.	39,8	26,1	924,5	965,5
Maribungan	*Parinari* spp.	48,1	24,0	1670,5	1800,7

13 Anhang VI

In den folgenden Tabellen sind die Regressionsanalysen für verschiedene Modellformulierungen unterschiedlicher Komplexität für alle Nadel- bzw. Laubbäume aufgeführt. Die linearen Regressionen wurden mit Hilfe der Methode der kleinsten Quadrate (Funktion lm in R bzw. S-Plus) berechnet.

13.1 Nadelbäume

Tabelle 13-1. Geschätzte Regressionskoeffizienten für ein einfaches lineares Modell mit dem BHD als unabhängiger Variablen für alle vorhandenen Nadelbäume (n=963). Die Zielgröße ist hier die oberirdische Gesamtbiomasse (*agb*).

Modell	Koeffizient	Schätzung	Std. Fehler	t-Wert
$\ln agb_i = \ln \alpha_i + \beta \ln BHD_i + \varepsilon_i$	α	-2,4144	0,041	-58,68
	β	2,4642	0,014	171,11

Standardfehler der Residuen: 0,2804 bei 961 Freiheitsgraden,
Multiples R^2: 0,9682,
F-Statistik: 29280 bei 1; p-Wert=0

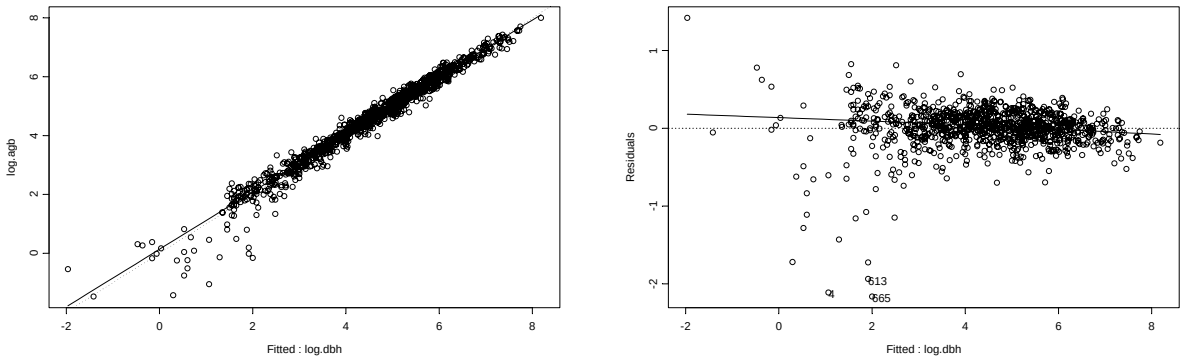


Abbildung 13-1. Beobachtete und geschätzte Biomasse (links) und Residuenplot der Regression (rechts).

Tabelle 13-2. Geschätzte Regressionskoeffizienten für ein einfaches lineares Modell mit dem BHD und der Baumhöhe als unabhängige Variablen für alle vorhandenen Nadelbäume (n=963). Die Zielgröße ist hier die oberirdische Gesamtbiomasse (*agb*).

Modell	Koeffizient	Schätzung	Std. Fehler	t-Wert
$\ln agb_i = \ln \alpha_i + \beta \ln BHD_i + \chi \ln h_i + \varepsilon_i$	α	-2,5040	0,041	-60,76
	β	2,1704	0,038	56,98
	χ	0,3492	0,042	8,28

Standardfehler der Residuen: 0,2711 bei 960 Freiheitsgraden,
Multiples R^2: 0,9703,
F-Statistik: 15710 bei 2; p-Wert=0

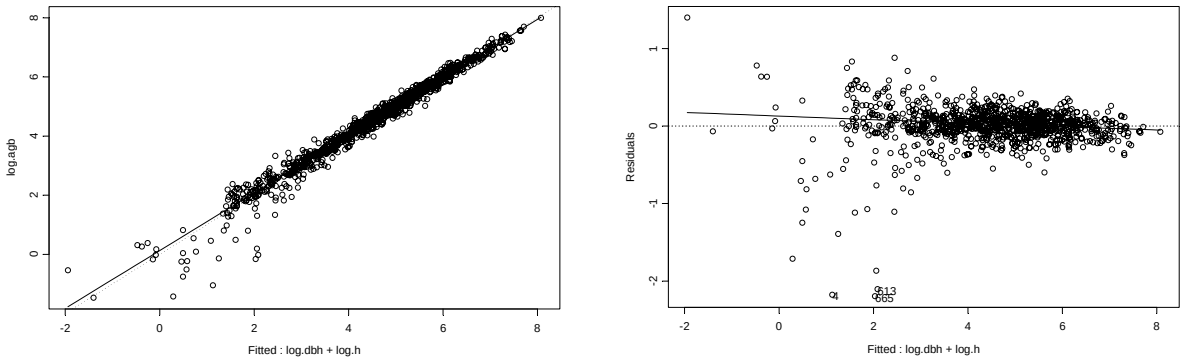


Abbildung 13-2. Beobachtete und geschätzte Biomasse (links) und Residuenplot der Regression (rechts).

Tabelle 13-3. Geschätzte Regressionskoeffizienten für ein einfaches lineares Modell mit dem BHD, der Baumhöhe und der Holzdichte (*wd*) als unabhängige Variablen für alle vorhandenen Nadelbäume (n=963). Die Zielgröße ist hier die oberirdische Gesamtbiomasse (*agb*).

Modell	Koeffizient	Schätzung	Std. Fehler	t-Wert
$\ln agb_i = \ln \alpha_i + \beta \ln BHD_i + \chi \ln h + \delta wd + \varepsilon_i$	α	-2,8966	0,153	-18,87
	β	2,1449	0,039	54,76
	χ	0,3805	0,043	8,72
	δ	0,8580	0,323	2,65

Standardfehler der Residuen: 0,2702 bei 959 Freiheitsgraden,
Multiples R^2: 0,9706,
F-Statistik: 10540 bei 3; p-Wert=0

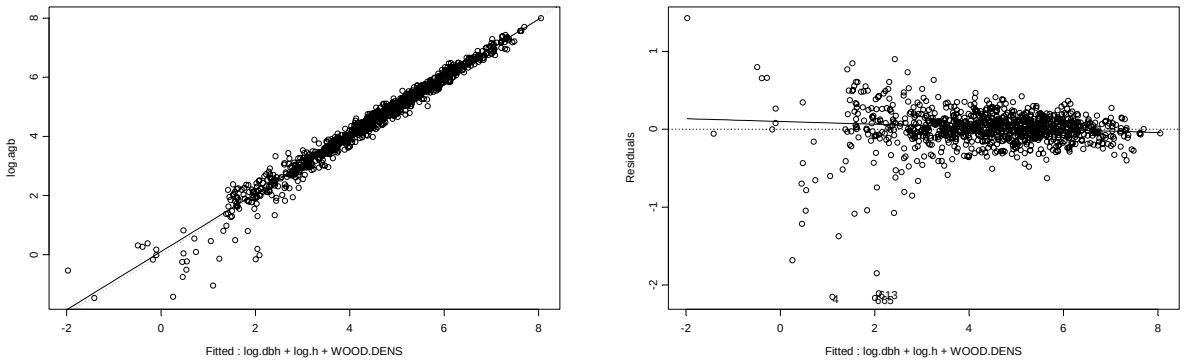


Abbildung 13-3. Beobachtete und geschätzte Biomasse (links) und Residuenplot der Regression (rechts).

13.2 Laubbäume

Tabelle 13-4. Geschätzte Regressionskoeffizienten für ein einfaches lineares Modell mit dem BHD als unabhängiger Variablen für alle vorhandenen Laubbäume (n=528). Die Zielgröße ist hier die oberirdische holzige Biomasse (*agwb*).

Modell	Koeffizient	Schätzung	Std. Fehler	t-Wert
$\ln agwb_i = \ln \alpha_i + \beta \ln BHD_i + \varepsilon_i$	α	-2,3170	0,045	-50,77
	β	2,5008	0,015	161,11

Standardfehler der Residuen: 0,2985 bei 526 Freiheitsgraden,
Multiples R^2: 0,9801,
F-Statistik: 25960 bei 1; p-Wert=0

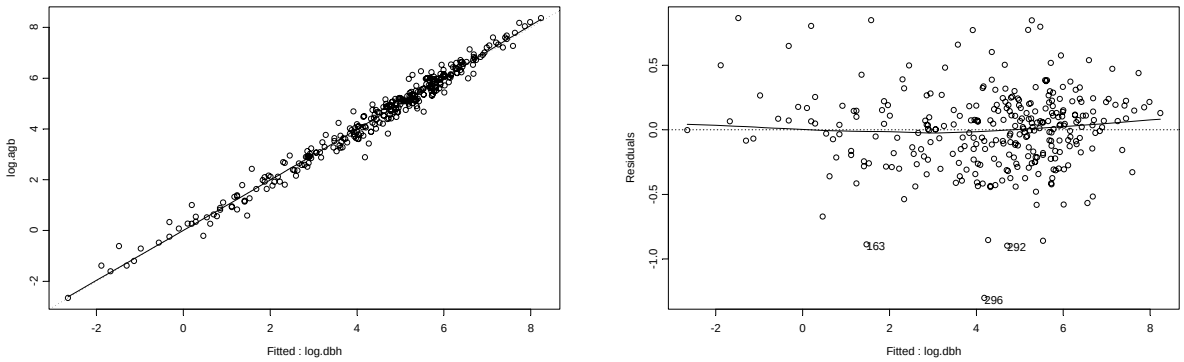


Abbildung 13-4. Beobachtete und geschätzte Biomasse (links) und Residuenplot der Regression (rechts).

Tabelle 13-5. Geschätzte Regressionskoeffizienten für ein einfaches lineares Modell mit dem BHD und Baumhöhe als unabhängige Variablen für alle vorhandenen Laubbäume (n=528). Die Zielgröße ist hier die oberirdische holzige Biomasse (*agwb*).

Modell	Koeffizient	Schätzung	Std. Fehler	t-Wert
$\ln agb_i = \ln \alpha_i + \beta \ln BHD_i + \chi \ln h_i + \varepsilon_i$	α	-3,1257	0,075	-41,62
	β	2,0476	0,038	53,68
	χ	0,7589	0,059	12,71

Standardfehler der Residuen: 0,2612 bei 525 Freiheitsgraden,
Multiples R^2: 0,9848,
F-Statistik: 17030 bei 2; p-Wert=0

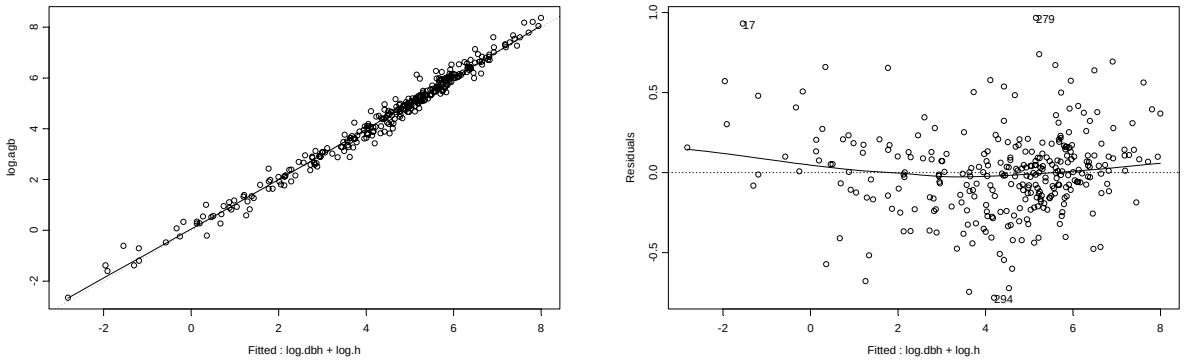


Abbildung 13-5. Beobachtete und geschätzte Biomasse (links) und Residuenplot der Regression (rechts).

Tabelle 13-6. Geschätzte Regressionskoeffizienten für ein einfaches lineares Modell mit dem BHD, Baumhöhe und Holzdichte (*wd*) als unabhängige Variablen für alle vorhandenen Laubbäume (n=528). Die Zielgröße ist hier die oberirdische holzige Biomasse (*agwb*).

Modell	Koeffizient	Schätzung	Std. Fehler	t-Wert
$\ln agb_i = \ln \alpha_i + \beta \ln BHD_i + \chi \ln h + \delta wd + \varepsilon_i$	α	-3,4863	0,081	-42,77
	β	2,1881	0,039	55,89
	χ	0,4844	0,064	7,55
	δ	1,1697	0,1341	8,72

Standardfehler der Residuen: 0,2444 bei 524 Freiheitsgraden,
Multiples R^2: 0,9867,
F-Statistik: 13000 bei 3; p-Wert=0

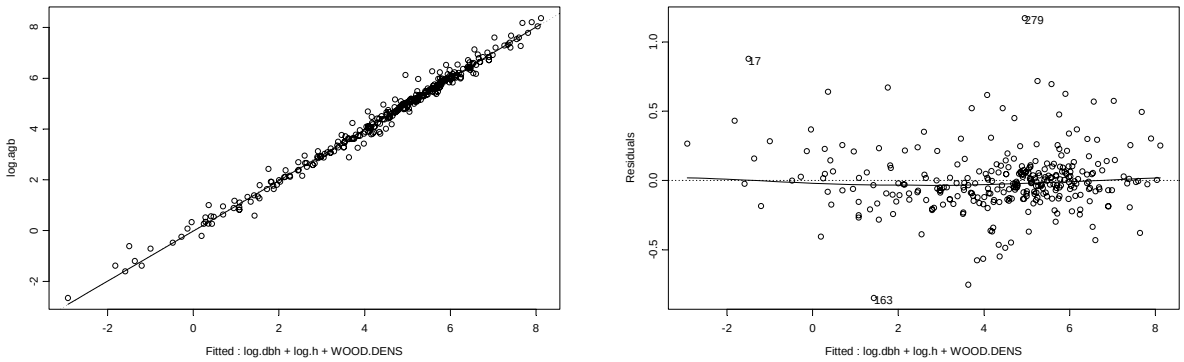


Abbildung 13-6. Beobachtete und geschätzte Biomasse (links) und Residuenplot der Regression (rechts).

Lebenslauf

Persönliche Daten

Name	Lutz Fehrmann
Geburtsdatum	17.10.1973
Geburtsort	Göttingen
Familienstand	ledig

Ausbildung

2004 – 2006	Promotion am Institut für Waldinventur und Waldwachstumskunde, Georg-August-Universität Göttingen
1999 – 2001	Masterstudium Tropical and International Forestry (Abschluss M.Sc.), Georg-August-Universität Göttingen
1996 – 1999	Bachelorstudium der Forstwissenschaften, Georg-August-Universität Göttingen

Berufliche Tätigkeit

seit 2004	Wissenschaftlicher Mitarbeiter am Institut für Waldinventur und Waldwachstum
2002 – 2003	Wissenschaftliche Hilfskraft am Institut für Forstliche Biometrie und Informatik
2001 – 2002	Wissenschaftliche Hilfskraft am Institut für Waldinventur und Waldwachstum