Herleitung einer generischen Biomassefunktion für die Buche

Ausarbeitung zum Projektteil der Vorlesung „Statistische Verfahren“

Jonas Franke* Gregor Romanus† Julian Schlichtholz‡

20. März 2016

*jonas.franke@uni-jena.de
†gregor.romanus@uni-jena.de
‡julian.schlichtholz@uni-jena.de

Inhaltsverzeichnis

1 Einleitung

Historisch waren Studien bezüglich der Biomasse von Bäumen in erster Linie dadurch motiviert, die ökonomische Nachhaltigkeit der Forstwirtschaft zu sichern. In der jüngeren Vergangenheit wurde, insbesondere durch die zunehmende Veränderung des globalen Klimas, eine große Fülle von Regressionsmodellen entwickelt. [5, S. 1]

Die Sammlung großer Datenmengen zur Biomasse von Buchen ist sehr aufwändig. Deshalb beschränken sich die zugrunde liegenden Studien meist auf Baumbestände mit wenigen Bäumen in kleinen Gebieten. Somit sind sie allein nicht repräsentativ. [8, S. 1] Bei Wutzler et al. [8] wurden deshalb die Datensätze verschiedener Studien zusammengefasst.

Unter Verwendung eines Teils dieser Datensätze sollen in der vorliegenden Arbeit Modelle zur Bestimmung der Biomasse, in Abhängigkeit verschiedener Einflussgrößen, ermittelt werden. Zunächst werden Modelle zu den Daten einzelner Autoren bestimmt. Im zweiten Schritt wird ein gemeinsames Modell für alle Beobachtungen der drei Studien ermittelt und dieses auf seine Besonderheiten untersucht. Abschließend wird ein alternatives Modell zur Bestimmung der Biomasse vorgestellt und seine statistischen Eigenschaften mit denen des ursprünglich verwendeten verglichen.

Dazu werden zunächst in Kapitel 2 die angewendeten Methoden sowie die damit gefunden abstrakten Modelle vorgestellt. In Kapitel 3 findet eine knappe Vorstellung der konkret errechneten Modelle statt. Abschließend werden die Ergebnisse in Kapitel 4 diskutiert.

2 Material und Methoden

2.1 Daten

Die dieser Arbeit zugrunde liegenden Datensätze stammen aus Studien von Bartelink [2], Joosten et. al. [7] und Heller [6]. Die Kombination der Datensätze stammt aus einer Studie von Wutzler et al. [8]. Die Datensätze enthalten jeweils die in Tabelle 2.1 erläuterten Variablen.

Variable	Erläuterungen
author	Der Autor der Studie (Bartelink, Joosten oder Heller)
hsl	Die Höhe über dem Meeresspiegel (am Standort des Baumes) in Metern
age	Das Alter des Baumes in Jahren
dbh	Der Brusthöhenumfang (Brusthöhe entspricht 1,3 m) des Baumes in Zentimetern
height	Die Höhe des Baumes in Metern
biom	Die oberirdische Biomasse in Kilogramm

Tabelle 2.1: Die vorliegenden Variablen

2.2 Statistische Methodik

2.2.1 Vor- und Rückwärtsselektion

Für die Modellwahl wurden die iterativen Verfahren der Vor- und Rückwärtsselektion angewendet.

Bei der Vorwärtsselektion wird ausgehend von einem (tendenziell kleinen) Anfangsmodell in jedem Iterationsschritt diejenige Einflussgröße in das Modell aufgenommen, welche die größte Verbesserung eines Informationskriteriums erreicht. Der Algorithmus terminiert, wenn keine Verbesserung mehr möglich ist. [4, S. 164]

Die Rückwärtsselektion funktioniert weitestgehend analog, jedoch wird zunächst von einem Startmodell mit vielen Einflussgrößen ausgegangen. In jedem Schritt wird eine Einflussgröße entfernt. [4, S. 164]

Es existieren verschiedene Informationskriterien, welche für die oben genannten Verfahren verwendet werden können. Im folgenden wird Akaikes Informationskriterium (AIC) [1] benutzt.

$$AIC = -2\ell\left(\hat{\beta}\right) + 2p \tag{2.1}$$

ℓ bezeichnet hier die Log-Likelihood-Funktion und p die Anzahl der Parameter. $2p$ beschreibt die „Bestrafung“ von Modellen mit zu vielen Einflussgrößen. [4, S. 206]

Für das maximale Modell wurde eine allometrische Funktion verwendet. Bei der Allometrie handelt es sich um „das Studium der mit Größenveränderungen einhergehenden Proportionsänderungen der Teile eines Organismus“ [5, S. 9]. Die Grundgleichung der Allometrie lautet

$$Y = ax^b = e^{b\ln x + \ln a} \Rightarrow \ln Y = b\ln x + \ln a \tag{2.2}$$

Nach dem Logarithmieren kann statt der exponentiellen Abhängigkeit eine lineare untersucht werden. Auch ein Blick auf den Datensatz lässt diese Modellwahl sinnvoll erscheinen. Abbildung 2.1 zeigt einen deutlichen Zusammenhang zwischen logarithmierter Biomasse und logarithmiertem Brusthöhendurchmesser.

2.2.2 Logarithmische Normalverteilung

In dieser Arbeit wird von einer Abhängigkeit der Zielgröße Y (die Biomasse) von den in Abschnitt 2.1 aufgeführten Einflussgrößen entsprechend einer allometrischen Funktion ausgegangen.

$$Y = \beta_0 X_{hsl}^{\beta_1} X_{age}^{\beta_2} X_{dbh}^{\beta_3} X_{height}^{\beta_4} \tag{2.3}$$

Die Koeffizienten β_i der allometrischen Funktion werden dabei in der logarithmierten Form ermittelt.

$$\ln Y = \beta_0 + \beta_1 \ln X_{hsl} + \beta_2 \ln X_{age} + \beta_3 \ln X_{dbh} + \beta_4 \ln X_{height} \tag{2.4}$$

Der Erwartungswert der logarithmierten Biomasse wird als normalverteilt mit folgenden Parametern angenommen.

$$\ln Y \sim \mathcal{N}(\mu, \sigma^2) \tag{2.5}$$

$$\mathbb{E}\ln Y = \mu \tag{2.6}$$

$$Var(\ln Y) = \sigma^2 \tag{2.7}$$

Wird die gefundene Funktion für den Erwartungswert durch Exponieren zurück transformiert, so ist zu berücksichtigen, dass die erhaltene Funktion für die Biomasse logarithmisch normalverteilt ist. Die logarithmische Normalverteilung ist rechtsschief und

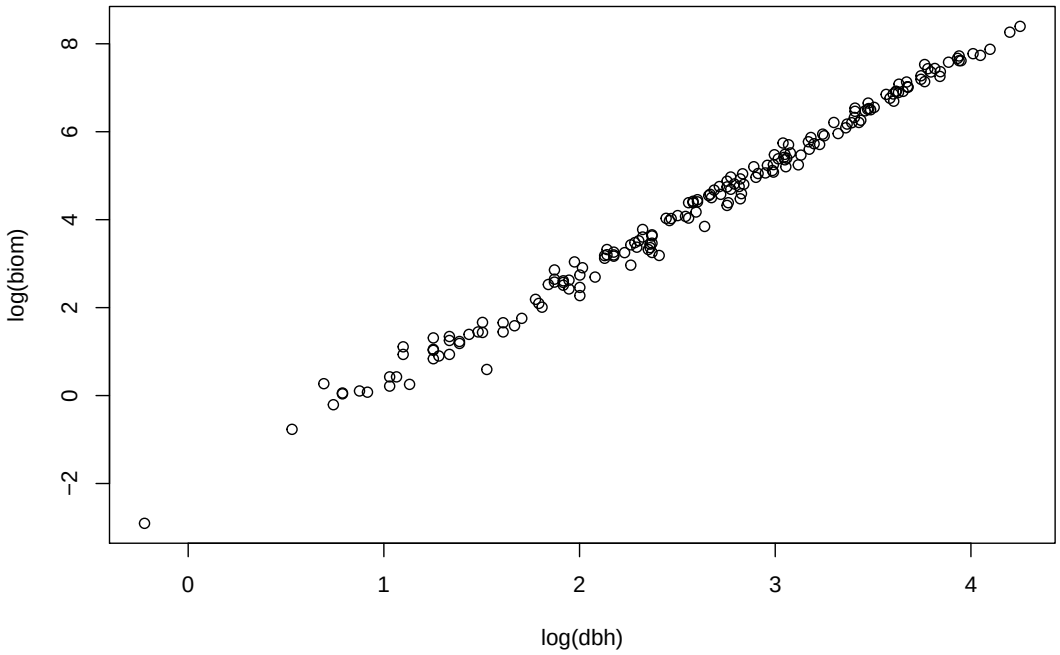

Abbildung 2.1: Zusammenhang zwischen logarithmierter Biomasse und logarithmiertem Brusthöhendurchmesser

hat als Parameter ebenfalls μ und σ^2 mit denselben Werten wie vor der Transformation. Aufgrund dessen wird die logarithmische Normalverteilung in der Regel über die Normalverteilung definiert [3, S. 57]. Die Parameter spiegeln allerdings nicht den Erwartungswert und die Varianz wieder. Diese werden definiert durch

$$\mathbb{E}Y = e^{\mu + \frac{\sigma^2}{2}} \tag{2.8}$$

$$Var(Y) = (e^{\sigma^2} - 1)e^{2\mu + \sigma^2} \tag{2.9}$$

In dieser Arbeit wird eine Umformung von der logarithmierten in die unlogarithmierte Form vereinfacht dargestellt, indem bei dem Erwartungswert der Biomasse ein zufälliger Anteil $\epsilon \in \mathbb{R}$ verwendet wird. Eine genauere Umformung würde Methoden wie *smearing estimate* erfordern, da durch die Rücktransformation Verzerrungen entstehen [7, S. 4].

2.2.3 Interaktionen

Unter einer Interaktion wird die Wechselwirkung des Effektes einer Einflussgröße mit mindestens einer weiteren verstanden. In dem beispielhaften Modell

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 \tag{2.10}$$

nennt man die Teile, welche nur von einer Einflussgröße abhängig sind, Haupteffekte und den Teil $\beta_3 X_1 X_2$ Interaktionseffekt. In dieser Arbeit werden zugunsten der Interpretierbarkeit des Ergebnisses lediglich Interaktionen 2. Ordnung, also mit nur zwei Einflussgrößen betrachtet. Es treten sowohl Wechselwirkungen zwischen einer binären und einer metrischen, als auch zwischen zwei metrischen Einflussgrößen auf. Der erste Fall kann unidirektional interpretiert werden. Der Term ändert sich bei Vorhandensein einer bestimmten Voraussetzung im Gegensatz zum Nichtvorhandensein dieser Voraussetzung. In dem Beispiel aus Gleichung 2.10 wären so bei Annahme von X_2 als binäre Variable folgende Gleichungen möglich

$$\text{Voraussetzung vorhanden:} \quad Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1 \tag{2.11}$$

$$\text{Voraussetzung nicht vorhanden:} \quad Y = \beta_0 + \beta_1 X_1 \tag{2.12}$$

Im Gegensatz dazu stehen die Interaktionen von zwei metrischen Einflussgrößen. Hier ist eine Interpretation und Modellierung des Einflusses deutlich schwieriger [4, S. 83 ff.]. Gleichung 2.10 würde in diesem Fall unverändert bleiben und eine Interpretation könnte bidirektional erfolgen.

2.3 Modelle für die einzelnen Autoren

Im ersten Schritt sollten lineare Modelle für die jeweiligen Teildatensätze der einzelnen Autoren gefunden werden. Die Suche erfolgte mittels der zuvor beschriebenen Vor- und Rückwärtsselektion. Als kleinstes Modell wurde dabei

$$\ln Y = \beta_0 \tag{2.13}$$

verwendet. Als größtes Modell kam

$$\begin{aligned} \ln Y = \beta_0 &+ \beta_1 \ln X_{hsl} + \beta_2 \ln X_{age} + \beta_3 \ln X_{dbh} + \beta_4 \ln X_{height} \\ &+ \beta_5 \ln X_{hsl} \ln X_{age} + \beta_6 \ln X_{hsl} \ln X_{dbh} + \beta_7 \ln X_{hsl} \ln X_{height} \\ &+ \beta_8 \ln X_{age} \ln X_{dbh} + \beta_9 \ln X_{age} \ln X_{height} \\ &+ \beta_{10} \ln X_{dbh} \ln X_{height} \end{aligned} \tag{2.14}$$

zur Anwendung.

Durch die Anwendung von Vor- und Rückwärtsselektion mittels der R-Funktion `step` ergaben sich (unter Verwendung des Minimalmodells aus Gleichung 2.13 und des Maximalmodells aus Gleichung 2.14) die folgenden Funktionen: das Modell

$$\mathbb{E} \ln Y = \beta_0 + \beta_1 \ln X_{dbh} + \beta_2 \ln X_{age} + \beta_3 \ln X_{dbh} \ln X_{age} \tag{2.15}$$

für Bartelink, das Modell

$$\mathbb{E} \ln Y = \beta_0 + \beta_1 \ln X_{dbh} + \beta_2 \ln X_{height} + \beta_3 \ln X_{dbh} \ln X_{height} \tag{2.16}$$

für Heller und das Modell

$$\begin{aligned} \mathbb{E} \ln Y = \beta_0 &+ \beta_1 \ln X_{dbh} + \beta_2 \ln X_{height} + \beta_3 \ln X_{hsl} + \beta_4 \ln X_{age} \\ &+ \beta_5 \ln X_{dbh} \ln X_{age} + \beta_6 \ln X_{height} \ln X_{age} \\ &+ \beta_7 \ln X_{dbh} \ln X_{height} + \beta_8 \ln X_{height} \ln X_{hsl} \end{aligned} \tag{2.17}$$

für Joosten.

Durch Exponieren und Umformen entstanden die Modelle

$$\mathbb{E}Y = e^{\beta_0} e^{\beta_3 \ln X_{dbh} \ln X_{age}} X_{dbh}^{\beta_1} X_{age}^{\beta_2} \pm \epsilon \tag{2.18}$$

für Bartelink,

$$\mathbb{E}Y = e^{\beta_0} e^{\beta_3 \ln X_{dbh} \ln X_{height}} X_{dbh}^{\beta_1} X_{height}^{\beta_2} \pm \epsilon \tag{2.19}$$

für Heller und

$$\begin{aligned} \mathbb{E}Y = &e^{\beta_0} e^{\beta_5 \ln X_{dbh} \ln X_{age}} e^{\beta_6 \ln X_{height} \ln X_{age}} \\ &\cdot e^{\beta_7 \ln X_{dbh} \ln X_{height}} e^{\beta_8 \ln X_{height} \ln X_{hsl}} \\ &\cdot X_{dbh}^{\beta_1} X_{height}^{\beta_2} X_{hsl}^{\beta_3} X_{age}^{\beta_4} \pm \epsilon \end{aligned} \tag{2.20}$$

für Joosten.

2.4 Gemeinsames Modell für alle Autoren

Um die Unterschiede zwischen den Autoren in den Daten berücksichtigen zu können, wurden 2 Indikatorvariablen, eine für die Datensätze von Bartelink und eine für die Datensätze von Heller, eingeführt. Hierdurch sollte deutlicher der Unterschied zu der umfangreicheren Studie Joostens herausgestellt werden können. Letztere hatte eine größere Vielfalt von Beobachtungen und Variabilität der Einflussgrößen. Die beiden Indikatorvariablen können durch folgende Indikatorfunktionen beschrieben werden.

$$\begin{aligned} X_{bart} &= \begin{cases} 1, & \text{wenn } X_{author} = Bartelink \\ 0, & \text{sonst} \end{cases} \\ X_{hell} &= \begin{cases} 1, & \text{wenn } X_{author} = Heller \\ 0, & \text{sonst} \end{cases} \end{aligned} \tag{2.21}$$

Zur Bestimmung eines gemeinsamen Modells wurde erneut eine Vorwärtsselektion und eine Rückwärtsselektion verwendet. Als Minimalmodell wurde das Modell aus Gleichung 2.13 übernommen. Als Maximalmodell wurde das Modell aus Gleichung 2.14 um die beiden oben genannten Indikatorvariablen und deren Interaktionen mit den anderen Einflussgrößen erweitert. Die AIC-Werte der so erhaltenen Modelle wurden miteinander verglichen und das Modell mit dem niedrigeren AIC-Wert als Lösung verwendet.

Dabei wurde durch die Vorwärtsselektion folgendes Modell mit einem AIC-Wert von $-132,18$ gefunden.

$$\begin{aligned}\mathbb{E}\ln Y = \beta_0 &+ \beta_1 \ln X_{dbh} + \beta_2 \ln X_{height} + \beta_3 \ln X_{hsl} + \beta_4 \ln X_{age} + \beta_5 X_{bart} \\ &+ \beta_6 \ln X_{dbh} \ln X_{height} + \beta_7 \ln X_{dbh} \ln X_{hsl} + \beta_8 \ln X_{height} \ln X_{hsl}\end{aligned} \tag{2.22}$$

Durch die Rückwärtsselektion wurde folgendes Modell mit einem AIC-Wert von $-134,74$ gefunden.

$$\begin{aligned}\mathbb{E}\ln Y = \beta_0 &+ \beta_1 \ln X_{dbh} + \beta_2 \ln X_{height} + \beta_3 \ln X_{hsl} \\ &+ \beta_4 X_{hell} + \beta_5 X_{bart} + \beta_6 X_{hell} \ln X_{dbh} \\ &+ \beta_7 \ln X_{dbh} \ln X_{height} + \beta_8 \ln X_{height} \ln X_{hsl}\end{aligned} \tag{2.23}$$

2.5 Vergleich mit verallgemeinertem linearen Modell

In obigen Betrachtungen wurde ein lineares Modell nach logarithmischer Transformation verwendet und die logarithmierte Biomasse als normalverteilt angenommen. Wird der Brusthöhendurchmesser als einzige Einflussgröße betrachtet, so nimmt dieses folgende Form an:

$$\text{(M1)}\ \mathbb{E}\ln Y_i = \mu_i = \beta_1 + \beta_2 \cdot \ln X_{dbh,i}\ ,\ Y_i \sim N(\mu, \sigma^2) \tag{2.24}$$

Dies ist jedoch keineswegs die einzig mögliche Betrachtungsweise. So könnte alternativ auch ein verallgemeinertes lineares Modell unter Annahme einer Gammaverteilung verwendet werden:

$$\text{(M2)}\ \ln \mathbb{E}Y_i = \ln \mu_i = \beta_1 + \beta_2 \cdot \ln X_{dbh,i},\ Y_i \text{ gammaverteilt mit } \mathbb{E}Y_i = \mu_i \tag{2.25}$$

Um das für den gegebenen Datensatz bessere Modell wählen zu können, sollte das Verhalten beider Modelle und insbesondere die Genauigkeit der durch sie vorhergesagten Biomassen verglichen werden. Da es sich beim Autor, wie bereits erläutert, um eine nicht vernachlässigbare Einflussgröße handelt, wurden die Betrachtungen nur für den Datensatz des Autors Joosten vorgenommen (116 Messwerte).
Zunächst wurden die ML-Schätzer $\beta_1^{M1}, \beta_2^{M1}$ für das Modell (M1) (mit der R-Funktion `lm`) bzw. $\beta_1^{M2}, \beta_2^{M2}$ für das Modell (M2) (mit der R-Funktion `glm`) bestimmt. Anhand

dieser konnten mit Rücktransformation von Gleichung 2.24 bzw. Gleichung 2.25 die vom jeweiligen Modell vorhergesagten Biomassen y_i^{M1} bzw. y_i^{M2}, $i = 1, ..., 116$, berechnet werden. Zur Untersuchung der Genauigkeit dieser Vorhersagen wurden ihre relativen Abweichungen zu den gemessenen Werten betrachtet:

$$d_i^{M1} = \frac{|y_i - y_i^{M1}|}{y_i} \text{ bzw. } d_i^{M2} = \frac{|y_i - y_i^{M2}|}{y_i}, i = 1, ..., 116 \tag{2.26}$$

Hier wurde die relative einer absoluten Betrachtungsweise vorgezogen, da die absoluten Schwankungen des Brusthöhendurchmessers dickerer Bäume naturgemäß größer als die dünnerer sind. Eine bloße Betrachtung der absoluten Abweichungen hätte somit zu einem überhöhten Einfluss von Bäumen mit größerem Brusthöhendurchmesser geführt, was nicht erwünscht ist. Als Maß für die Genauigkeit der Modelle wurden die Summen der relativen Differenzen aller Bäume

$$\sum_{i=1}^{116} d_i^{M1} \text{ bzw. } \sum_{i=1}^{116} d_i^{M2} \tag{2.27}$$

sowie die Standardabweichungen

$$\sqrt{\frac{1}{115}\sum_{i=1}^{116}\left(d_i^{M1}\right)^2} \text{ bzw. } \sqrt{\frac{1}{115}\sum_{i=1}^{116}\left(d_i^{M1}\right)^2} \tag{2.28}$$

verwendet. Für beide Maße gilt: kleine Werte sprechen für eine größere Genauigkeit des jeweils betrachteten Modells.

Von großer Bedeutung für ein Modell ist außerdem das statistische Verhalten des Schätzers. Relevant sind vor allem seine Genauigkeit und ob es sich um einen erwartungstreuen Schätzer handelt. Hier sollte nun die Schätzung des Parameters β_2 durch die Modelle (M1) und (M2) untersucht werden. Zu diesem Zweck wurden basierend auf Modell (M2) Pseudobeobachtungen unterschiedlichen Umfangs auf folgende Weise simuliert: Für einen vorgegebenen Stichprobenumfang N wurde die entsprechende Anzahl an Brusthöhendurchmessern zufällig aus dem Datensatz des Autors Joosten ausgewählt. Dafür wurden die jeweils nach (M2) erwarteten Biomassen

$$\mu_i = e^{\beta_1^{M2}} \cdot x_{dbh,i}^{\beta_2^{M2}} \tag{2.29}$$

bestimmt, wobei β_1^{M2} bzw. β_2^{M2} wie oben beschrieben mit der R-Funktion `glm` berechnet wurden. Diese Funktion lieferte außerdem den Dispersionsparameter ϕ der Gammaverteilung $\Gamma(a, s)$ mit Erwartungswert μ_i. Über

$$a = \frac{1}{\phi}\ ,\ s_i = \mu_i \cdot \phi \tag{2.30}$$

konnten somit die Parameter der Gamma-Verteilung berechnet werden. Entsprechend dieser Parameter wurde zu jedem der gewählten Brusthöhendurchmesser ein Biomassewert simuliert. Anhand der so bestimmten Werte konnten nun mit `lm` bzw. `glm` die

ML-Schätzer für (M1) bzw. (M2) berechnet werden. Dieser Prozess von Simulation und Berechnung der ML-Schätzer wurde für einen festen Stichprobenumfang N 1000-mal wiederholt. Als Stichprobenumfänge wurden $N = 1, ..., 100$ verwendet, wobei für jedes N die im Schritt $N-1$ verwendeten Brusthöhendurchmesser beibehalten und ein weiterer zufälliger hinzugefügt wurde. Die Arbeitsweise des Programms zur Bestimmung von Schätzern für β_2 nach (M1) bzw. (M2) ist in Algorithmus 1 beschrieben.

Algorithmus 1 Bestimmung der Schätzer basierend auf Pseudobeobachtungen

1: **for** $N \leftarrow 1$ *to* 100 **do**
2: $X_{dbh,N} = (X_{dbh,N-1}$, zufälliges X_{dbh} aus Originaldaten)
3: $\mu = e^{\beta_1^{M2}} \cdot X_{dbh,N}^{\beta_2^{M2}}$
4: **for** $M \leftarrow 1$ *to* 1000 **do**
5: $Y_{N,M} = \text{RANDOM}(\Gamma(\frac{1}{\phi}, \mu_i \cdot \phi))$
6: $\beta_{2,N,M}^{M1}$ =Koeffizient β_2 von $M1(X_{dbh,N}, Y_{N,M})$
7: $\beta_{2,N,M}^{M2}$ =Koeffizient β_2 von $M2(X_{dbh,N}, Y_{N,M})$
8: **return** $\left(\beta_{2,N,M}^{M1}\right)_{N,M}$, $\left(\beta_{2,N,M}^{M2}\right)_{N,M}$

Für jeden Stichprobenumfang N ergaben sich also für (M1) und (M2) jeweils 1000 Schätzwerte des Parameters β_2. Anhand dieser sollten Genauigkeit und Erwartungstreue der ML-Schätzer für β_2 beider Modelle untersucht werden. Der Erwartungswert selbiger für ein festes N wurde dabei geschätzt über

$$\frac{1}{1000} \sum_{M=1}^{1000} \beta_{2,N,M}^{M1} \text{ bzw. } \frac{1}{1000} \sum_{M=1}^{1000} \beta_{2,N,M}^{M2} \tag{2.31}$$

3 Ergebnisse

3.1 Ergebnisse der Modelle der einzelnen Autoren

Die Ergebnisse der in Abschnitt 2.3 vorgestellten Modelle sind

$$\mathbb{E}Y = 0,4362e^{0,2650 \ln X_{dbh} \ln X_{age}} X_{dbh}^{1,4226} X_{age}^{-0,3152} \pm \epsilon \tag{3.1}$$

mit Varianz $\sigma^2 = 0,037$ für Bartelink,

$$\mathbb{E}Y = 0,1102e^{0,2072 \ln X_{dbh} \ln X_{height}} X_{dbh}^{1,2684} X_{height}^{0,6040} \pm \epsilon \tag{3.2}$$

mit Varianz $\sigma^2 = 0,0293$ für Heller und

$$\begin{aligned} \mathbb{E}Y =& 0,0393e^{0,1812 \ln X_{dbh} \ln X_{age}} e^{-0,3498 \ln X_{height} \ln X_{age}} \\ & \cdot e^{0,0927 \ln X_{dbh} \ln X_{height}} e^{0,0471 \ln X_{height} \ln X_{hsl}} \\ & \cdot X_{dbh}^{1,1673} X_{height}^{1,5158} X_{hsl}^{-0,1946} X_{age}^{0,49601} \pm \epsilon \end{aligned} \tag{3.3}$$

mit Varianz $\sigma^2 = 0,021$ für Joosten.

3.2 Ergebnis des gemeinsamen Modells aller Autoren

Durch Einsetzen der Koeffizienten in das gemäß AIC-Wert bessere gemeinsame Modell der drei Autoren ergibt sich für den Erwartungswert der logarithmischen Biomasse

$$\begin{aligned} \mathbb{E} \ln Y = & -0,581 + 1,748 \ln X_{dbh} - 0,071 \ln X_{height} - 0,328 \ln X_{hsl} \\ & + 0,307 \ln X_{hell} - 0,157 \ln X_{bart} - 0,103 \ln X_{dbh} \ln X_{hell} \\ & + 0,126 \ln X_{dbh} \ln X_{height} + 0,092 \ln X_{height} \ln X_{hsl} \end{aligned} \tag{3.4}$$

mit der Varianz $\sigma^2 = 0,026$. Durch Rücktransformation und Umformungen ergibt sich für den Erwartungswert der Biomasse

$$\begin{aligned} \mathbb{E}Y = & 0,600 X_{dbh}^{1,748} X_{height}^{-0,071} X_{hsl}^{-0,328} 1,359^{X_{hell}} 0,855^{X_{bart}} X_{dbh}^{-0,103 X_{hell}} \\ & \cdot e^{0,126 \ln X_{dbh} \ln X_{height}} e^{(0,092 \ln X_{height} \ln X_{hsl}} \pm \epsilon \end{aligned} \tag{3.5}$$

3.3 Ergebnis des Vergleichs mit verallgemeinertem linearen Modell

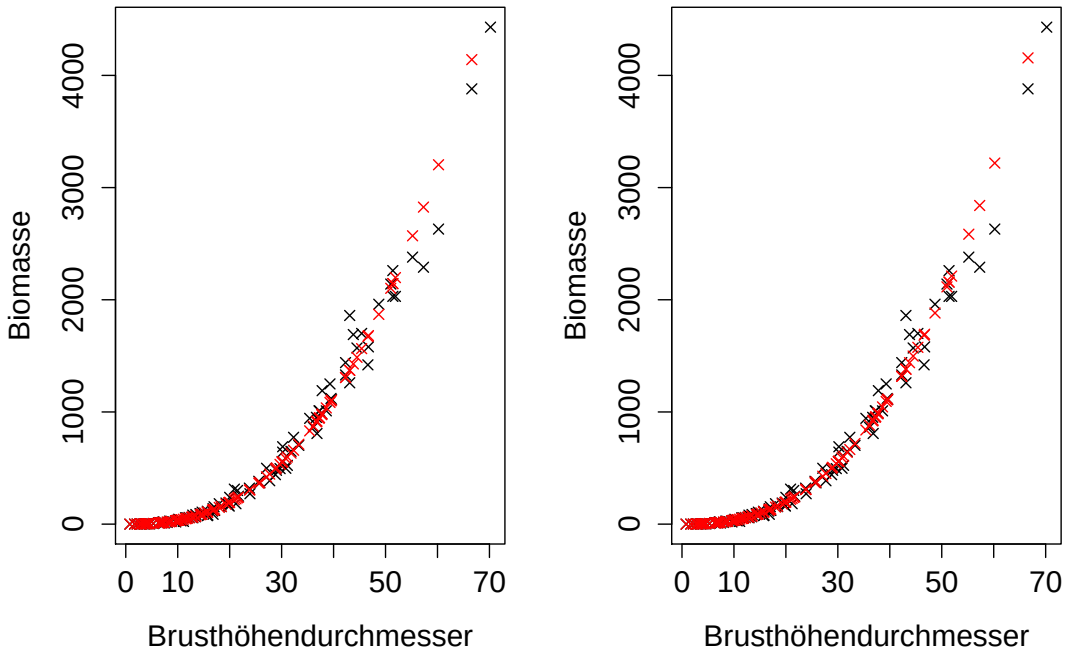

Abbildung 3.1: Gemessene (schwarz) und durch (M1) (linke Abbildung, rot) bzw. (M2) (rechte Abbildung, rot) vorhergesagte Biomassen

Die Summe der relativen Differenzen (Gleichung 2.27) beträgt für (M1) 16,44 und für (M2) 16,56. Modell (M1) hat eine Standardabweichung (Gleichung 2.28) von 114,69, bei Modell (M2) ergibt diese 116,05.

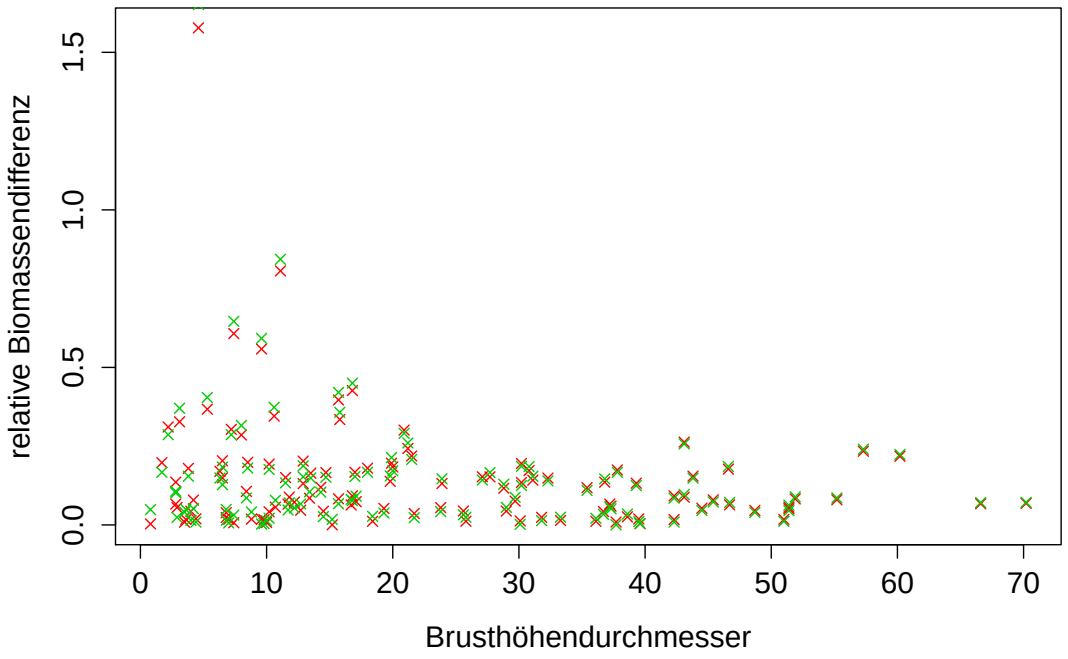

Abbildung 3.2: Relative Differenzen der durch (M1) (rot) bzw. (M2) (grün) vorhergesagten Biomassen zu den Messwerten (entsprechend Gleichung 2.27)

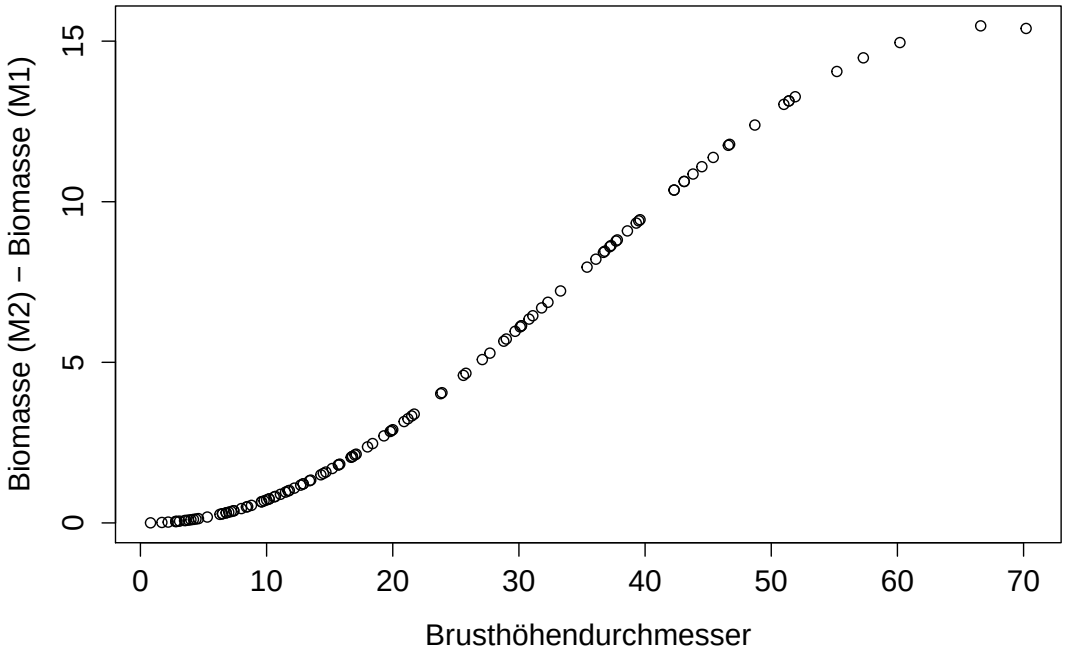

Abbildung 3.3: Differenz zwischen durch (M2) und durch (M1) vorhergesagter Biomassen

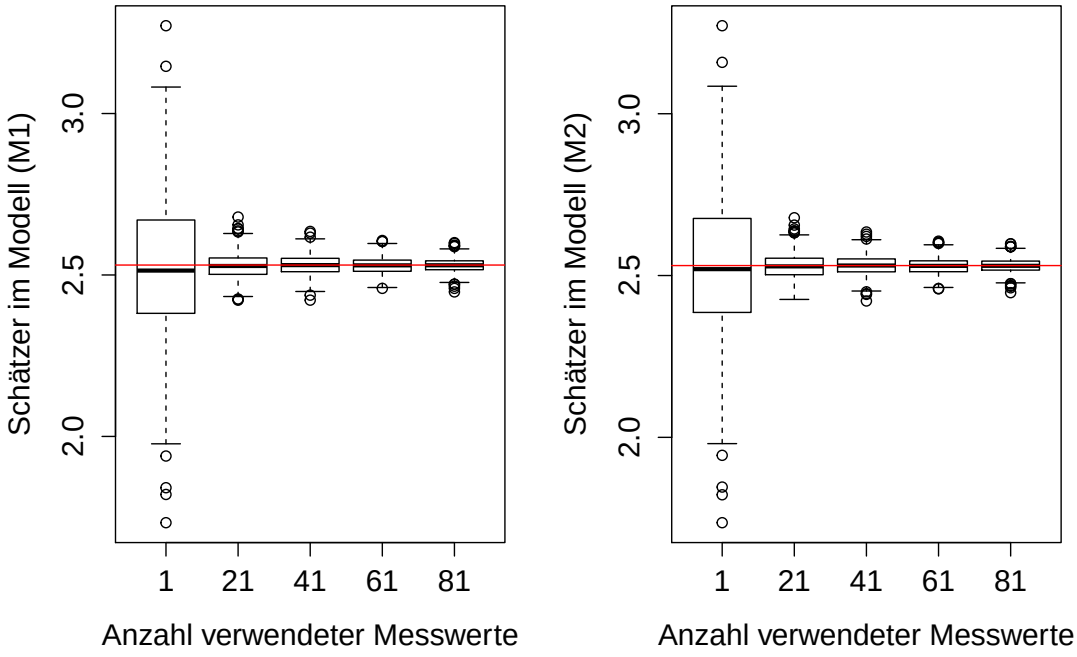

Abbildung 3.4: Boxplots auf Basis von 1000 Pseudobeobachtungen durch (M1) (links) bzw. (M2) (rechts) berechneter Schätzer für β_2 bei 1, 21, 41, 61 und 81 verwendeten Ursprungswerten (entsprechend Algorithmus 1). Die rote Linie kennzeichnet den wahren Wert β_2.

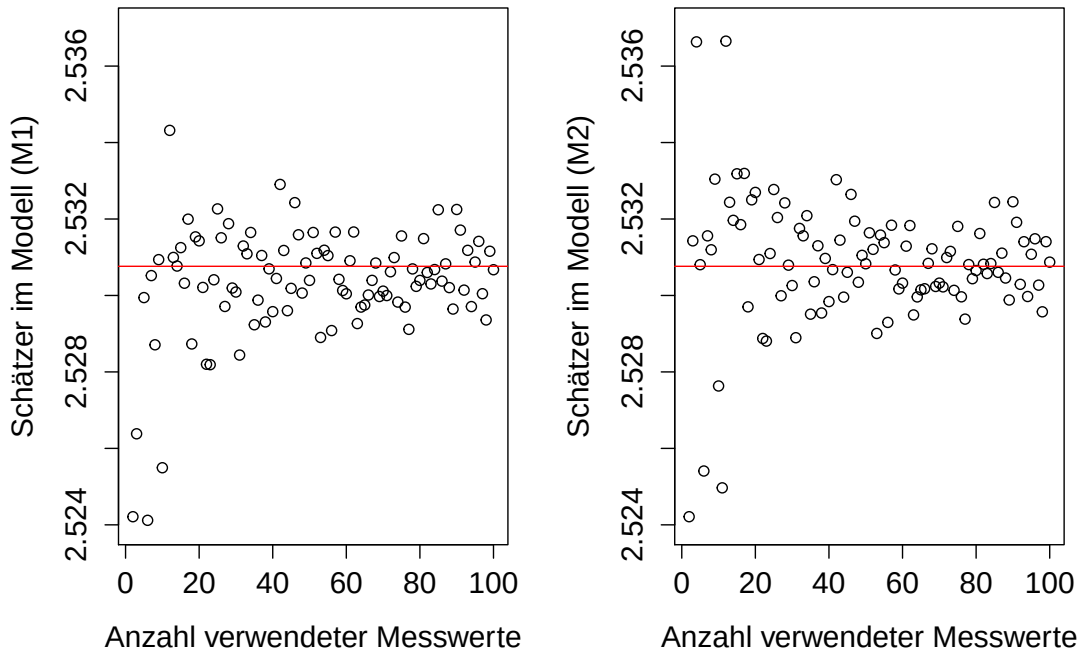

Abbildung 3.5: Erwartungswerte auf Basis von 1000 Pseudobeobachtungen durch (M1) (links) bzw. (M2) (rechts) berechneter Schätzer für β_2 (rote Linie)

4 Diskussion

4.1 Diskussion der Modelle der einzelnen Autoren

Die mittels Vor- und Rückwärtsselektion gefundenen Modelle für die einzelnen Autoren sind in vielen Punkten schlüssig. So fehlt bei Bartelink und Heller die Einflussgröße X_{hsl}, da jedes Datum in diesen Datensätzen denselben Wert hat. Bei Heller fehlt darüber hinaus die Einflussgröße X_{age}. Dies liegt vermutlich daran, dass die Variabilität von X_{age} bei Heller deutlich geringer ausfällt als bei den anderen beiden Autoren.

Im Modell zur Studie von Bartelink fehlt die Einflussgröße X_{height}. Sie hätte zwar die Vorhersage verbessert, jedoch nicht stark genug um das AIC insgesamt zu verbessern.

Das Vorhandensein aller Einflussgrößen bei Joosten lässt sich am ehesten auf den großen Umfang der Studie zurückführen. Durch die Verschiedenheit der betrachteten Bäume kommen die einzelnen Einflussgrößen insgesamt stärker zur Geltung.

4.2 Diskussion des gemeinsamen Modells aller Autoren

In Anbetracht des starken Zusammenhangs bestimmter Einflussgrößen, wie beispielsweise Höhe und Brusthöhendurchmesser eines Baumes, ist es schwierig, genaue Aussagen bei einer Interpretation des Modells und der einzelnen Koeffizienten zu treffen. Hinzu kommt, dass bereits Modelle mit nur dem Brusthöhendurchmesser als einzige Einflussgröße schon recht gute Näherungen liefern (wie bereits in Abbildung 2.1 zu sehen ist) und die Hinzunahme weiterer Einflussgrößen zum Teil nur geringe Auswirkungen hat. Allgemein müssen die jeweiligen Einflussgrößen so interpretiert werden, dass ohne Änderung der anderen Einflussgrößen ihr Effekt am deutlichsten erkennbar ist.

Der Brusthöhendurchmesser hat einen positiven Einfluss auf die Biomasse, während die anderen beiden metrischen Einflussgrößen, die Höhe und die Höhe über NN, singulär gesehen einen negativen Einfluss haben. Dies könnte dadurch begründet sein, dass Bäume in dichten bzw. hohen Wäldern durch den Wachstumsdruck eher in die Höhe wachsen und keine große Krone ausprägen können. Bäume in höheren Lagen könnten allgemein schlechtere Wachstumsbedingungen haben.

Diesen beiden negativen Einflüssen stehen zwei Wechselwirkungen entgegen. Zum einen die positive Wechselwirkung zwischen Brusthöhendurchmesser und Höhe, welche darin

begründet liegen könnte, dass Bäume, sobald sie eine bestimmte Höhe und Dicke erreicht haben, anfangen können eine breitere Krone auszubilden. Zum anderen die positive Interaktion zwischen Höhe und Höhe über NN, welche eventuell mit der langsameren Wachstumsgeschwindigkeit der Bäume in Höhenlagen und der somit höheren Dichte des Holzes zusammenhängen könnte, was sich erst bei großen Bäumen bemerkbar macht.

Betrachtet man die Unterschiede zwischen den Modellen, so fällt auf, dass bei den Datensätzen von Heller und Bartelink jeweils ein positiver Einfluss vorhanden ist. Die Bäume dieser Studien haben ohne die Betrachtung der anderen Einflussgrößen eine leicht höhere Biomasse, was besonders bei kleinen Bäumen in den logarithmierten Daten in Abbildung 4.1 erkennbar ist. Bei den Daten von Heller kommt außerdem noch eine negative Interaktion mit dem Brusthöhendurchmesser hinzu. Das heißt, dass obwohl die Biomasse von dünnen Bäume bei Heller eventuell höher ist als die bei Joosten, diese nicht so stark mit größerer Dicke des Stammes zunimmt. Dieser Effekt ist tendenziell in den nicht logarithmierten Daten in Abbildung 4.1 erkennbar. Die Gründe für diese Abweichung zwischen den Studien können vielfältig sein. Möglich sind sowohl natürliche Faktoren wie etwa der Boden oder die umliegende Flora, als auch artifizielle Faktoren wie die unterschiedlichen Arbeitsweisen der Forschungsgruppen.

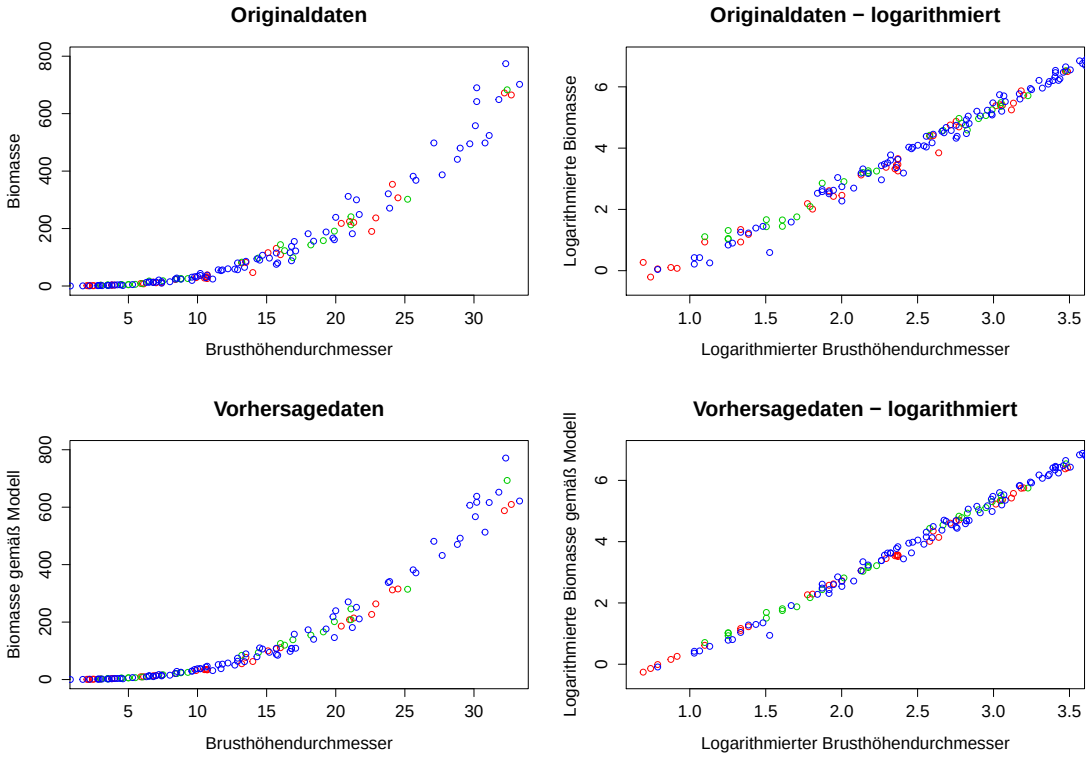

Abbildung 4.1: Vergleich der Datensätze der 3 Autoren Bartelink (rot), Heller (grün) und Joosten(blau)

4.3 Diskussion des Vergleichs mit verallgemeinertem linearen Modell

Für die Daten des Autors Joosten wurde das Verhalten zweier verschiedener Modelle (Gleichung 2.24, 2.25) untersucht. In Abbildung 3.1 sind die durch Modell (M1) und Modell (M2) vorhergesagten Biomassen zusammen mit den Originaldaten dargestellt. Obwohl es sich in beiden Fällen um sehr einfache Modelle handelt, nähern sie die gemessenen Werte bereits gut an, die Abweichungen sind minimal. Um die Genauigkeiten der beiden Modelle zu vergleichen, wurden in Abbildung 3.2 nun die relativen Differenzen der originalen zu den vorhergesagten Biomassen für beide Modelle verglichen. Auch hier zeigt sich kaum ein Unterschied zwischen (M1) und (M2), keines der beiden Modelle liefert eine durchgehend geringere relative Differenz. Die Summe der relativen Differenzen (Gleichung 2.27) war für (M1) geringer als für (M2). Gleiches gilt für die Standardabweichungen entsprechend Gleichung 2.28. Nach diesen Maßen nähert also (M1) die Originaldaten besser an und wäre somit zu bevorzugen. Allerdings sind die Unterschiede der beiden Größen für (M1) und (M2) so gering, dass die Genauigkeit der Modelle als nahezu gleich betrachtet werden kann. Ein klarer Unterschied zwischen ihnen zeigt sich in Abbildung 3.3: hier wird deutlich, dass (M2) durchgehend größere Biomassen als (M1) voraussagt. Da beide Modelle fast die gleiche Genauigkeit aufweisen, sollte bei der Modellwahl daher das Augenmerk eher darauf gelegt werden, ob eine Über- oder eine Unterschätzung als kritischer angesehen wird. Modell (M2) wird die Biomasse eher über-, Modell (M1) sie eher unterschätzen.

Weiterhin wurde das Schätzverhalten der Modelle (M1) und (M2) für den Parameter β_2 untersucht. In Abbildung 3.4 sind für (M1) bzw. (M2) Histogramme der bei verschiedenen Stichprobenumfängen bestimmten ML-Schätzer dargestellt. Darin wird deutlich, dass die Varianz selbiger für geringe Stichprobenumfänge sehr groß ist und mit wachsender Anzahl rapide fällt. Ein bedeutender Unterschied zwischen (M1) und (M2) ist in den Grafiken nicht erkennbar. Außerdem wird ersichtlich, dass für geringere Stichprobenumfänge der Erwartungswert des Schätzers weiter vom wahren Wert entfernt ist. Das wird in Abbildung 3.5 noch einmal verdeutlicht. Dort sind für (M1) bzw. (M2) die Erwartungswerte der Schätzer für β_2 bei verschiedenen Stichprobenumfängen N dargestellt. In beiden Grafiken zeigt sich dasselbe Verhalten: Für geringe N ist der Erwartungswert z.T. weit vom realen Wert (rote Linie) entfernt, für wachsende N nähert er sich diesem immer stärker an. Weder der mit (M1) noch der mit (M2) bestimmte Schätzer für β_2 sind also erwartungstreu, jedoch scheinen beide für wachsende Stichprobenumfänge gegen einen erwartungstreuen Schätzer zu konvergieren. Das steht im Einklang mit der Theorie: Der ML-Schätzer im linearen Modell (M1) konvergiert für $N \to \infty$ bekanntermaßen gegen einen erwartungstreuen Schätzer. Bei Modell (M2) handelt es sich um das zur Simulation der Pseudobeobachtungen verwendete, also ist auch hier diese Konvergenz gegeben. Eine unterschiedliche Konvergenzgeschwindigkeit der Erwartungswerte der beiden Schätzer β_2^{M1} und β_2^{M2} lässt sich in Abbildung 3.5 nicht erkennen.
Für möglichst genaue Schätzungen von β_2, d.h. solche mit geringen Varianzen und ei-

nem Erwartungswert nahe an β_2, sollten bei beiden Verfahren möglichst viele Messwerte verwendet werden. Dem steht die Bestrebung, möglichst wenige Messwerte aufzunehmen entgegen, da dies in der Praxis sehr aufwendig sein kann. Als maximaler Stichprobenumfang wurde $N = 100$ verwendet. Dieser liefert bereits sehr geringe Varianzen und einen Erwartungswert des Schätzers für β_2 nahe dem wahren Wert.

Literatur

[1] H. Akaike. „Information theory and an extension of the maximum likelihood principle“. In: *Selected Papers of Hirotugu Akaike*. Springer, 1998, S. 199–213.

[2] H. Bartelink. „Allometric relationships for biomass and leaf area of beech (Fagus sylvatica L)“. In: *Annales des Sciences Forestieres*. Bd. 54. 1. EDP Sciences. 1997, S. 39–50.

[3] C. Dormann. *Parametrische Statistik: Verteilungen, maximum likelihood und GLM in R*. Statistik und ihre Anwendungen. Springer Berlin Heidelberg, 2013.

[4] L. Fahrmeir, T. Kneib und S. Lang. *Regression: Modelle, Methoden und Anwendungen*. Springer-Verlag, 2009.

[5] L. Fehrmann. „Alternative Methoden zur Biomasseschätzung auf Einzelbaumebene unter spezieller Berücksichtigung der k-nearest neighbour (k-NN) Methode“. Diss. Georg-August-Universität Göttingen, 2006.

[6] H. Heller und D. Göttsche. „Biomasse-Messungen an Buche und Fichte“. In: *Ökosystemforschung, Ergebnisse des Sollingprojekts 1966–1986*. Hrsg. von H. Ellenberg. Stuttgart: E. Ulmer Verlag, 1986, S. 507.

[7] R. Joosten, J. Schumacher, C. Wirth und A. Schulte. „Evaluating tree carbon predictions for beech (Fagus sylvatica L.) in western Germany“. In: *Forest Ecology and Management* 189.1 (2004), S. 87–96.

[8] T. Wutzler, C. Wirth und J. Schumacher. „Generic biomass functions for Common beech (Fagus sylvatica) in Central Europe: predictions and components of uncertainty“. In: *Canadian Journal of Forest Research* 38.6 (2008), S. 1661–1675.

Abbildungsverzeichnis

Tabellenverzeichnis

Algorithmenverzeichnis