Ein Instrument zur Schätzung von Holzernteproduktivitäten mittels der kNN-Methode

Fabian Kostadinov Eidgenössische Forschungsanstalt für Wald, Schnee und Landschaft (CH)*
Renato Lemm Eidgenössische Forschungsanstalt für Wald, Schnee und Landschaft (CH)
Oliver Thees Eidgenössische Forschungsanstalt für Wald, Schnee und Landschaft (CH)

A software tool for the estimation of wood harvesting productivity using the kNN method

For operational planning and management of wood harvests it is important to have access to reliable information on time consumption and costs. To estimate these efficiently and reliably, appropriate methods and calculation tools are needed. The present article investigates whether use of the method of the k nearest neighbours (kNN) is appropriate in this case. The kNN algorithm is first explained, then is applied to two sets of data "combined cable crane and processor" and "skidder", both containing wood harvesting figures, and thus the estimation accuracy of the method is determined. It is shown that the kNN method's estimation accuracy lies within the same order of magnitude as that of a multiple linear regression. Advantages of the kNN method are that it is easy to understand and to visualize, together with the fact that estimation models do not become out of date, since new data sets can be constantly taken into account. The kNN Workbook has been developed by the Swiss Federal Institute for Forest, Snow and Landscape Research (WSL). It is a software tool with which any data set can be analysed in practice using the kNN method. This tool is also presented in the article.

Keywords: multiple regression, k nearest neighbours (kNN), timber harvesting, cost estimation, forest management
doi: 10.3188/szf.2012.0119

* Zürcherstrasse 111, CH-8903 Birmensdorf, E-Mail fabian.kostadinov@wsl.ch

Die Wirtschaftlichkeit der Waldbewirtschaftung wird massgeblich durch die Kosten der Holzernte beeinflusst. Ihr Anteil an den Gesamtkosten im Forstbetrieb liegt in der Schweiz zwischen 40 und 60 Prozent. Für die betriebliche Planung und Steuerung der Holzernte werden verlässliche Informationen über ihre Zeitbedarfe und Kosten benötigt. Aber auch für strategische und konzeptionelle Überlegungen zum Holzproduktionsprozess stellen diese Informationen eine wichtige Grundlage dar. Um die Zeitbedarfe und Kosten der Holzernte effizient und verlässlich schätzen zu können, bedarf es zweckmässiger Kalkulationsgrundlagen beziehungsweise IT-gestützter Instrumente, wie zum Beispiel der Holzernteproduktivitätsmodelle «HeProMo» der Eidgenössischen Forschunganstalt für Wald, Schnee und Landschaft (WSL; Frutig et al 2009).

Solche Modelle basieren auf der Zerlegung des gesamten Ernteprozesses in einzelne Aktivitäten und deren mathematischer Beschreibung mittels Regressionen, welche den Zusammenhang zwischen der Produktivität des Erntesystems und seinen Einsatzbedingungen quantifizieren. Die Erstellung praxistauglicher Modelle ist allerdings aufwendig – nicht zuletzt wegen der umfangreichen Datenerhebung – und erfordert eine professionelle IT-Umsetzung. Wegen des raschen technischen Fortschritts in der Holzernte veralten die Modelle relativ schnell, was wiederum Aktualisierungen notwendig machen kann.

Eine Alternative zu regressionsbasierten Methoden bildet die Methode der k nächsten Nachbarn (kNN-Methode). Sie erlaubt es, eine unbekannte abhängige Variable eines Datensatzes über die Ähnlichkeit zu Referenzdatensätzen mit bekannten Werten zu schätzen. Aus einer Holzschlagdatenbank eines Forstbetriebes oder eines Forstunternehmens lässt sich so für einen anstehenden Holzschlag die zu erwartende Produktivität bei der Holzernte schätzen, indem die dem neuen Holzschlag ähnlichsten früheren Holzschläge der Datenbank ermittelt werden und aus diesen ein Durchschnittswert der erzielten Produktivitäten berechnet wird. Die Methode lässt eine hohe Praxistauglichkeit erwarten. Für den Fall der Kalkulation von Selbstkosten auf betrieblicher Ebene kann auf eigene und aktuelle Holzschlagdaten zurückgegriffen werden. Dies verspricht eine

Abb 1 *Kombiseilgerät im Einsatz.*
Foto: Fritz Frutig

treffsichere Prognose und ein erhöhtes Vertrauen in diese. Methodische Probleme, wie sie sich bei der Anwendung der Regressionsrechnung ergeben können (z.B. Verletzung der Annahme, dass die Daten einer bestimmten Verteilung folgen), lassen sich vermeiden. Eine erste Untersuchung des Einsatzes der kNN-Methode zur Schätzung von Holzernteproduktivitäten bestätigte am Beispiel einer Holzschlagdatenbank für Vollernter (Lemm et al 2005) die Vorteilhaftigkeit des Ansatzes.

Allerdings fehlte bisher ein einfach zu bedienendes Kalkulationsinstrument, welches die Anwendung der kNN-Methode in der Praxis überhaupt erst ermöglicht. Daher wurde an der WSL eine zweckmässige Software als Prototyp entwickelt. Der vorliegende Beitrag

- erläutert die Funktionsweise der kNN-Methode und der erstellten Software,
- vergleicht anhand von je einer Holzschlagdatenbank für Kombiseilgeräte (Abbildung 1) und Seilschlepper die Prognoseergebnisse der kNN-Methode mit denen von Regressionsrechnungen,
- zieht Schlussfolgerungen für die Eignung und die Weiterentwicklung des Ansatzes in der Praxis.

Material und Methoden

Die kNN-Methode

Ziel der kNN-Methode ist es, für ein Ereignis, bestehend aus einem Vektor von m unabhängigen, erklärenden Variablen, eine Schätzung für eine ausgewählte, abhängige Variable zu erzeugen. Dazu werden die unabhängigen Variablen des Schätzereignisses mit denjenigen einer Menge von Referenzereignissen verglichen. Für alle Referenzereignisse wird mittels eines geeigneten Distanzmasses bestimmt, wie «nahe» sie zum Schätzereignis stehen. Referenzereignisse mit einer geringen Distanz zum Schätzereignis sind dem Schätzereignis ähnlicher als Referenzereignisse mit einer grösseren Distanz. Aus allen Nachbarn werden die k nächsten Nachbarn ausgewählt. Aus den abhängigen Variablen der zuvor ausgewählten benachbarten Referenzereignisse kann nun ein Mittelwert für die abhängige Variable des Schätzereignisses errechnet werden. Üblicherweise wird dabei die relative Nähe der Nachbarn zum Schätzereignis mittels eines geeigneten Gewichtes berücksichtigt.

Als Mass für die Distanz zwischen zwei Ereignissen wird bei der kNN-Methode oft die euklidische Distanz gewählt, wobei jedoch auch die Wahl eines anderen Distanzmasses möglich wäre, beispielsweise die Mahalanobis- (Fehrmann 2006) oder die Manhattan-Distanz (Duda et al 2001). Formal lautet die Berechnung der euklidischen Distanz:

$$d'_{ij} = \sqrt{\sum_{p=1}^{m} \frac{\alpha_p^2}{\beta_p^2}(x_{ip} - x_{ip})^2} \qquad (1)$$

wobei

d'_{ij} die euklidische Distanz zwischen den unabhängigen Variablen der Referenzereignisse $\vec{x}_i$ und dem Schätzereignis $\vec{x}_j$,

$\vec{x}_j = (x_{j1}, x_{j2}, \ldots, x_{jm}, \hat{y}_j)$ das Schätzereignis j mit m unabhängigen Variablen x_{j1} bis x_{jm} und der zu schätzenden abhängigen Variablen $\hat{y}_j$,

$\vec{x}_i = (x_{i1}, x_{i2}, \ldots, x_{im}, y_i)$ ein Referenzereignis i mit m unabhängigen Variablen x_{i1} bis x_{im} und der abhängigen Variablen y_i,

α_p $(p = 1, 2, \ldots, m)$ Gewichtungsfaktor der unabhängigen Variablen p,

β_p $(p = 1, 2, \ldots, m)$ Skalierungsfaktor der unabhängigen Variablen p

ist. Eine Gewichtung der unabhängigen Variablen mittels eines Faktors $\frac{\alpha_p}{\beta_p}$ ist nötig, weil die einzelnen unabhängigen Variablen verschiedene rechnerische Einheiten und Grössenordnungen aufweisen können (z.B. CHF/h, m³ usw.). Deshalb müssen sie standardisiert werden. Als Gewichtungsfaktor α_p kann beispielsweise der Korrelationskoeffizient der unabhängigen Variablen p mit der abhängigen Variablen y dienen und als Skalierungsfaktor β_p die Standardabweichung der unabhängigen Variablen p (Felber 2005). Es sei also

$$\alpha_p = \frac{\sum_{i=1}^{n}(x_{ip} - \bar{x}_p)(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_{ip} - \bar{x}_p)^2 \times \sum_{i=1}^{n}(y_i - \bar{y})^2}},\ p = 1, 2, \ldots, m \qquad (2)$$

der Korrelationskoeffizient zwischen der unabhängigen Variablen p und der abhängigen Variablen y_i (bei n Referenzereignissen) und

$$\beta_p = \sqrt{\frac{1}{n-1} \times \sum_{i=1}^{n}(x_{ip} - \bar{x}_p)^2} \qquad (3)$$

die Standardabweichung der unabhängigen Variablen p über alle Referenzereignisse. Weiter gelten

$$\bar{x}_p = \frac{1}{n}\sum_{i=1}^{n} x_{ip} \qquad (4)$$

als Mittelwert der unabhängigen Variablen p über alle Referenzereignisse und

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad (5)$$

als Mittelwert der abhängigen Variablen über alle Referenzereignisse.

Nachdem die Distanzen zwischen allen Referenzereignissen und dem Schätzereignis berechnet wurden, werden unter den Referenzereignissen die zum Schätzereignis k nächstgelegenen Nachbarn ausgewählt. Die Wahl eines geeigneten k erfordert einige Übung und Kenntnis der Datengrundlage. Loftsgarden & Quesenberry (1965) schlagen für k eine Wahl von $\sqrt{n}$ vor, Enas & Choi (1986) $n^{3/8}$ respektive $n^{2/8}$, wobei n die Anzahl Referenzereignisse darstellt.

Eine von den Autoren nicht weiterverfolgte Alternative besteht darin, statt die Anzahl k nächster Nachbarn vorzugeben, eine bestimmte maximale euklidische Distanz festzulegen. Alle zur Schätzung verwendeten Nachbarn müssen dann innerhalb dieser Maximaldistanz liegen. Der Schätzwert wird schliesslich über alle ausgewählten k nächsten Nachbarn hinweg berechnet:

$$\hat{y}_j = \sum_{i=1}^{k} g_{ij} y_i \qquad (6)$$

wobei

$\hat{y}_j$ Schätzwert der abhängigen Variablen des Schätzereignisses j,

y_i Messwerte der abhängigen Variablen der Referenzereignisse i,

k Anzahl berücksichtigte nächste Nachbarn,

g_{ij} Funktion zur Gewichtung der k nächsten Nachbarn.

Die Funktion g wird eingeführt, um die unterschiedlichen Distanzen zwischen dem Schätzereignis und den Referenzereignissen zu berücksichtigen. Auf diese Weise bekommen Referenzereignisse, die näher beim Schätzereignis liegen, ein höheres Gewicht als solche, die weiter entfernt sind. Soll beispielsweise allen ausgewählten Referenzereignissen dasselbe Gewicht beigemessen werden, so könnte

$$g_{ij} = \frac{1}{k} \qquad (7)$$

gewählt werden. Eine Funktion, die nahe Referenzereignisse stärker berücksichtigt als weiter entfernte, ist beispielsweise

$$g_{ij} = \frac{\left(\frac{1}{1+d_{ij}}\right)}{\sum_{i=1}^{k}\left(\frac{1}{1+d_{ij}}\right)} \qquad (8)$$

und folglich

$$\hat{y}_j = \frac{\sum_{i=1}^{k}\left(\frac{1}{1+d_{ij}}\right) \times y_i}{\sum_{i=1}^{k}\left(\frac{1}{1+d_{ij}}\right)} \qquad (9)$$

In sämtlichen weiteren Berechnungen wird die in (8) aufgeführte Gewichtungsfunktion eingesetzt.

Um nun für eine Reihe von Datensätzen die Güte der Anpassung des Schätzwertes an den realen Messwert beurteilen zu können, werden häufig die Quadratwurzel des mittleren quadratischen Fehlers (Root Mean Square Error, RMSE), der Bias und der mittlere absolute prozentuale Fehler (Mean Absolute Percentage Error, MAPE) herangezogen (Fehrmann 2006, Fehrmann et al 2008).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \qquad (10)$$

$$Bias = \frac{1}{n} \times \sum_{i=1}^{n}(\hat{y}_i - y_i) \qquad (11)$$

$$MAPE = \frac{1}{n} \times \sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \qquad (12)$$

wobei

$(\hat{y}_i - y_i)$ das i-te Residuum,

$\hat{y}_i$ geschätzte abhängige Variable,

y_i gemessene abhängige Variable,

n Anzahl Referenzereignisse.

Der RMSE hat dieselbe Einheit wie die abhängige Variable. Er kann nur Werte ≥ 0 annehmen. Der Bias beschreibt den systematischen Fehler. Er hat dieselbe Einheit wie die abhängige Variable. Ein positiver Bias bedeutet gemäss Formel (11) ein systematisches Überschätzen der abhängigen Variablen, ein negativer Bias ein systematisches Unterschätzen derselben. Der MAPE beschreibt die mittleren absoluten prozentualen Abweichungen zwischen Schätzung und Messwert. Der MAPE ist somit ein Mass für das Verhältnis der Residuen zu den tatsächlichen Messwerten. Ein Wert von 0 besagt, dass sämtliche Schätzungen exakt mit den tatsächlichen Messwerten übereinstimmen. Gegen oben sind dem MAPE hingegen keine Grenzen gesetzt, er kann auch Werte > 1 annehmen. Bei einer optimalen Wahl von *k* sind RMSE und MAPE minimal, und der Bias liegt nahe bei 0.

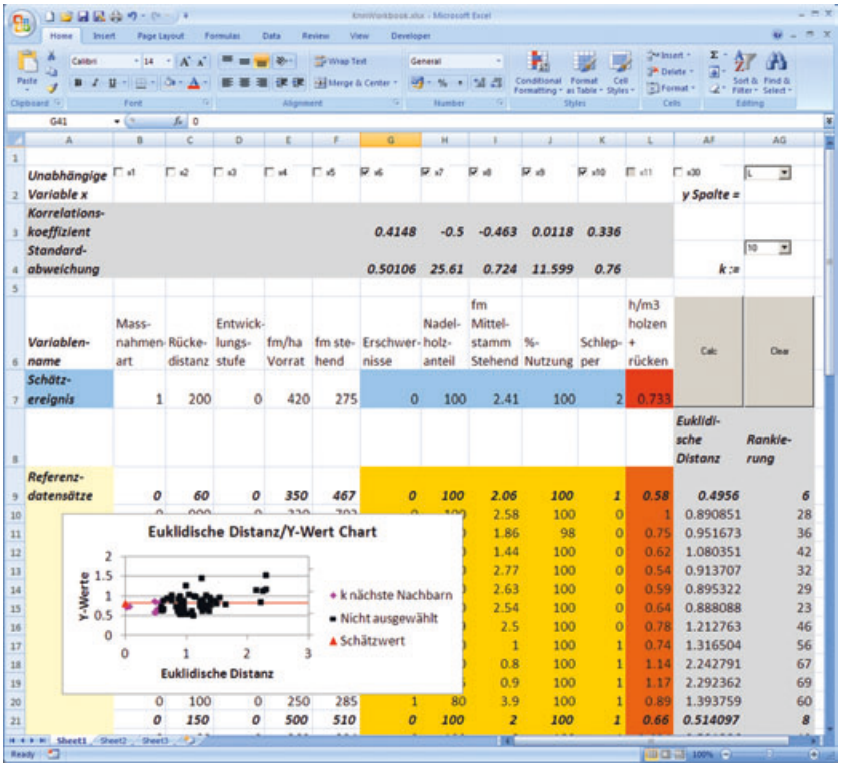

Abb 2 *Benutzeroberfläche des kNN-Workbooks.*

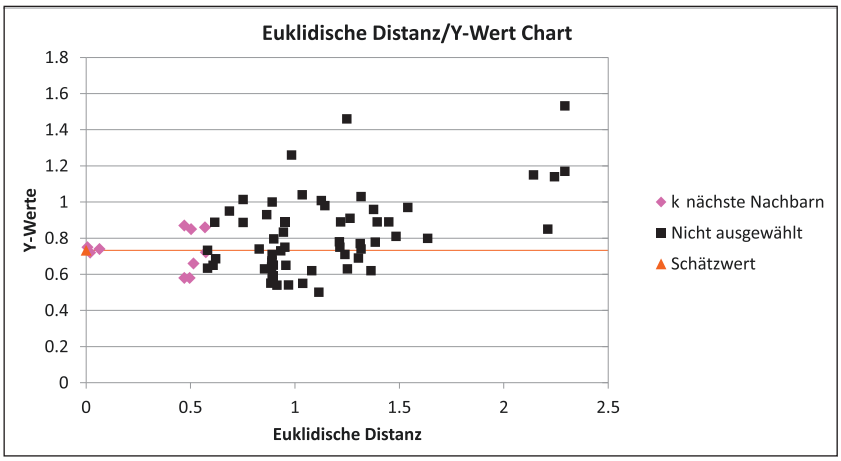

Abb 3 *Grafik einer kNN-Auswertung.*

Das kNN-Workbook

Um die kNN-Methode der Praxis zugänglich zu machen, wurde das kNN-Workbook entwickelt. Es handelt sich dabei um eine Software, mit welcher sich kNN-Auswertungen durchführen lassen.

Nutzung

Das kNN-Workbook ist in der Lage, mit einer potenziell grossen und dynamisch wachsenden Menge von Datensätzen umzugehen. Neu hinzugefügte Datensätze werden bei einer wiederholten Auswertung ebenfalls berücksichtigt. Daten können entweder im kNN-Workbook selbst gespeichert oder zur Auswertung aus einer anderen Excel-Datei herauskopiert und eingefügt werden. Die Installation eines Datenbanksystems entfällt.

Abbildung 2 zeigt einen Screenshot der Benutzeroberfläche des kNN-Workbooks, nachdem eine kNN-Auswertung vorgenommen wurde:

1. Um die abhängige Variable $\hat{y}$ eines Schätzereignisses zu berechnen, muss das Schätzereignis mit den konkreten Werten seiner unabhängigen Variablen in Zeile 7 der Benutzeroberfläche eingegeben werden. Das kNN-Workbook bietet die Möglichkeit zur Eingabe von maximal 29 unabhängigen und einer abhängigen Variablen.
2. Die Referenzereignisse trägt der Benutzer ab Zeile 9 ein, beispielsweise durch Kopieren und Einfügen der Daten aus einer bereits bestehenden Excel-Datei.
3. Dann wählt er mittels der Checkboxen in Zeile 2 die zur Berechnung vorgesehenen unabhängigen Variablen (Spalten G, H, I, J, K) und mittels des Dropdown-Menüs in Zelle AG2 die Spalte mit der abhängigen Variablen (Spalte L) aus.
4. Schliesslich legt er aus dem Drop-down-Menü in Zelle AG4 einen passenden Wert für *k* fest und betätigt den «Calc-Knopf».
5. Das kNN-Workbook wertet die eingegebenen Daten aus und präsentiert dem Benutzer folgende Ergebnisse:
   - Die geschätzte, abhängige Variable $\hat{y}$ des Schätzereignisses wird angezeigt (Zelle L7).
   - Jede ausgewählte, unabhängige Variable x_i wird mit der abhängigen Variablen korreliert, und es werden die dazugehörigen Bravais-Pearson-Korrelationskoeffizienten berechnet und angezeigt (Zeile 3). Für jede ausgewählte, unabhängige Variable wird zudem die Standardabweichung berechnet (Zeile 4).
   - Für jeden Datensatz werden die euklidische Distanz sowie die Rangzuordnung bezüglich der Nähe zum Referenzdatensatz berechnet (Spalten AF und AG). Die zu den *k* nächsten Nachbarn gehörenden Datensätze werden fett markiert (z.B. Zeile 9 und 21).
   - Es wird automatisch eine Grafik erstellt, in der auf der y-Achse der Wert der abhängigen Va-

Variablenname	Beschreibung	Einheit	Kombiseilgerät	Seilschlepper
Unabhängige Variablen				
Massnahmenart	Art des forstlichen Eingriffes in einen Bestand. Reduziert auf zwei Ausprägungen unterschiedlicher Arbeitsintensität: «Durchforstung» und «Räumung»	–	x	x
Rückedistanz	Durchschnittliche Rückedistanz	m	x	x
Erschwernisse	Besondere Erschwernisse bei der Holzernte durch Gelände, Witterung, unvorhersehbare Ereignisse etc. Reduziert auf die Ausprägung «vorhanden» oder «nicht vorhanden»	–	x	x
Entwicklungsstufe	Die durchschnittliche Entwicklungsstufe des Bestandes. Reduziert auf zwei Kategorien «Baumholz» und «stufig»	–	x	x
fm/ha Vorrat	Holzvorrat im Bestand vor Eingriff	fm/ha	x	x
Nadelholzanteil	Der Nadelholzanteil des angezeichneten Holzes	%	x	x
fm stehend	Entnahmemenge gemäss Anzeichnungsprotokoll insgesamt	fm	x	x
fm Mittelstamm stehend	Durchschnittliches Mittelstammvolumen (gemäss Anzeichnungsprotokoll)	fm	x	x
%-Nutzung	Anteil Entnahme am stehenden Vorrat	%	x	x
Kombiseilgerättyp	Zwei eingesetzte Kombiseilgerättypen: «Typ 1», «Typ 2»	–	x	
Anzahl Seillinien	Anzahl Seillinien des Holzschlages	–	x	
Anzahl Stützen	Anzahl Stützen aller Seillinien im Holzschlag	–	x	
Seillinienlänge	Gesamtlänge aller Seillinien des Holzschlages	m	x	
Rückerichtung	Rückerichtung reduziert auf zwei Kategorien unterschiedlicher Arbeitsintensität: «bergauf», «bergab/eben»*	–	x	
Seilschleppertyp	Drei eingesetzte Seilschleppertypen mit zunehmender Arbeitsleistung: «Typ 1», «Typ 2», «Typ 3»	–		x
Abhängige Variablen				
h/m³ Holzhauerei und Rücken	Stunden Holzhauerei und Rücken pro geernteten Kubikmeter Holz. Enthält die Arbeitsschritte Fällen, Aufrüsten, Rücken, Einschneiden und Poltern, beim Kombiseilgerät überdies Montage und Demontage	h/m³	x	x

Tab 1 *Aufgenommene unabhängige und abhängige Variablen. * Für alle aufgestellten Seillinien eines Holzschlages wurde in dieselbe Richtung gerückt. Aufgrund der eingeschränkten Vergleichbarkeit der Rückerichtungen wurde nicht weiter zwischen «bergab» oder «eben» unterschieden. Hingegen wurden diese beiden gemeinsam von der weniger arbeitsintensiven Rückerichtung «bergauf» abgegrenzt.*

riablen und auf der x-Achse dessen euklidische Distanz aufgetragen sind (Abbildung 3). Die Referenzdatensätze werden rosa und schwarz dargestellt. Rosa Rauten repräsentieren die für die Berechnung des Schätzwertes verwendeten *k* nächsten Nachbarn, schwarze Quadrate hingegen Referenzdatensätze, die nicht mehr zu den *k* nächsten Nachbarn gehören. Der Schätzwert selbst wird als rotes Dreieck mit einer horizontalen, roten Linie in der Grafik angezeigt.

Durch diese Informationen gewinnen die Benutzerinnen und Benutzer einen Überblick über ihre Daten, die Güte der Schätzung und insbesondere darüber, welches nun die zu ihrem Referenzdatensatz ähnlichen Datensätze sind. Die Erfahrungswerte bereits abgeschlossener Holzschläge werden so zu einer nützlichen Orientierungshilfe bei der Schätzung geplanter Schläge.

Technische Implementierung

Die Umsetzung des kNN-Workbooks erfolgte auf der Basis der Visual-Studio-Office-Tools(VSTO)-Add-in-Technologie für Excel 2007. Aufgrund der weiten Verbreitung der Microsoft-Office-Produktpalette kann davon ausgegangen werden, dass auf diese Weise die grösste Zahl von Nutzern erreicht werden kann. Ein Nachteil dieser Wahl liegt darin, dass damit keine Open-Source-Lösung unterstützt wird. Über die COM[1]-Schnittstelle von Excel 2007 werden Daten an eine .NET-Laufzeitumgebung[2] übergeben, welche diese Daten auswertet. Die Resultate werden wiederum via COM-Schnittstelle an Excel übergeben, welches die Daten anschliessend darstellt. Das kNN-Workbook wurde in der Programmiersprache C# geschrieben.

Datengrundlage des Fallbeispiels

Die kNN-Methode wurde mittels des kNN-Workbooks auf in den Jahren 2005 bis 2010 erhobene Holzschlagdatensätze für Seilschlepper und Kombiseilgerät angewendet. Die Daten stammen alle aus demselben schweizerischen Forstbetrieb. Sämtliche erfassten Holzschläge wurden in Eigenregie des Betriebes durchgeführt. Alle Holzschlagdatensätze wurden vom Betrieb mit denselben Variablen erstellt und den nach den Ernteverfahren getrennten Datensatzmengen zugeordnet. Anstelle von Produktivitätskennzahlen (z.B. m³/h) wurden vom Betrieb die dazu inversen Effizienzkennzahlen (h/m³) erhoben. Je kleiner die Effizienzkennzahl ausfällt, desto effizienter ist ein Verfahren.

1 COM steht für Component Object Model, eine von Microsoft entwickelte Middleware zur Kommunikation von auf Windows aufsetzenden Softwarekomponenten.

2 Common Language Runtime (CLR): die Laufzeitumgebung des Microsoft .NET-Frameworks.

Variablenname	Mittelwert	Minimum	Maximum	Standard-abweichung
Unabhängige Variablen				
Massnahmenart	66 Datensätze «Durchforstung», 30 Datensätze «Räumung»			
Rückedistanz	235	80	650	99
Erschwernisse	52 Datensätze «ohne», 44 Datensätze «mit Erschwernissen»			
Entwicklungsstufe	73 Datensätze «Baumholz», 23 Datensätze «stufig»			
fm/ha Vorrat	392	180	750	90
Nadelholzanteil	86	2	100	23
fm stehend	597	173	1600	297
fm Mittelstamm stehend	1.59	0.35	3.37	0.66
%-Nutzung	89	27	100	21
Kombiseilgerättyp	28 Datensätze «Typ 1», 68 Datensätze «Typ 2»			
Anzahl Seillinien	2.4	1.0	7.0	1.2
Anzahl Stützen	3.0	0.0	10.0	2.3
Seillinienlänge	694	150	2700	487
Rückerichtung	46 Datensätze «eben/bergab», 50 Datensätze «bergauf»			
Abhängige Variablen				
h/m³ Holzhauerei und Rücken	0.69	0.27	1.62	0.29

Tab 2 *Kennzahlen der unabhängigen und abhängigen Variablen beim Kombiseilgerät.*

Variablenname	Mittelwert	Minimum	Maximum	Standard-abweichung
Unabhängige Variablen				
Massnahmenart	32 Datensätze «Durchforstung», 38 Datensätze «Räumung»			
Rückedistanz	177	40	1500	212
Erschwernisse	39 Datensätze ohne, 31 Datensätze «mit Erschwernissen»			
Entwicklungsstufe	59 Datensätze «Baumholz», 11 Datensätze «stufig»			
fm/ha Vorrat	389	180	600	84
Nadelholzanteil	89	2	100	26
fm stehend	352	120	804	146
fm Mittelstamm stehend	2.03	0.64	3.90	0.72
%-Nutzung	98	30	100	12
Seilschleppertyp	31 Datensätze «Typ 1», 25 Datensätze «Typ 2», 14 Datensätze «Typ 3»			
Abhängige Variablen				
h/m³ Holzhauerei und Rücken	0.81	0.50	1.53	0.21

Tab 3 *Kennzahlen der unabhängigen und abhängigen Variablen beim Seilschlepper.*

Aus der Vielzahl der erhobenen unabhängigen Variablen wurden jene ausgewählt, welche einen signifikanten Einfluss auf die Holzernteproduktivität beziehungsweise -effizienz haben. Die aufgeführte Liste unabhängiger Variablen erhebt nicht den Anspruch auf Vollständigkeit, da durch den Forstbetrieb nur eine Auswahl möglicher Eingangsgrössen erhoben wurde. Unvollständige Datensätze wurden entfernt. Übrig blieben 96 Datensätze für das Holzernteverfahren mittels Kombiseilgerät und 70 Datensätze für dasjenige mittels Seilschlepper.

Die Tabellen 1 bis 3 zeigen die aufgenommenen unabhängigen und abhängigen Variablen und ihre Kennzahlen für die Holzernteverfahren Seilschlepper und Kombiseilgerät.

Ergebnisse und Diskussion

Gestützt auf die Daten des Fallbeispiels wurden als Erstes schrittweise Analysen mittels multipler linearer Regression durchgeführt. Die signifikanten unabhängigen Variablen (Prediktorvariablen) wurden in die Modellrechnung aufgenommen und Kennwerte bezüglich ihrer Schätzgüte berechnet (Tabellen 4 und 5).

In einem zweiten Schritt wurden die in der Regressionsmodellrechnung aufgenommenen unabhängigen Variablen pro Holzerntemethode für die Analyse mittels der kNN-Methode verwendet. Um das optimale *k* zu ermitteln, wurde für jeden möglichen Wert von *k* eine eigene Analyse durchgeführt, dabei wurden die Kennwerte RMSE, Bias und MAPE berechnet (Abbildungen 4 und 5).

In einem dritten Schritt wurde die Schätzgüte der kNN-Analyse bei optimalem *k* mit derjenigen der multiplen linearen Regressionsmodellrechnung verglichen (Abbildungen 6 und 7).

Regressionsanalyse

Sowohl für das Kombiseilgerät als auch für den Seilschlepper wurde zuerst ein Gesamtmodell einer multiplen linearen Regression mit sämtlichen für das Holzernteverfahren zur Verfügung stehenden unabhängigen Variablen erstellt. Anschliessend wurden in einem rückwärts gerichteten Eliminationsverfahren schrittweise jene unabhängigen Variablen mit dem geringsten Erklärungsbeitrag zum Modell entfernt, bis ein Gesamtmodell gefunden war, bei welchem alle unabhängigen Variablen das Signifikanzkriterium von 0.05 erfüllten.[3] Als abhängige Variable wurde die Effizienz des Holzschlags (h/m³ Holzhauerei und Rücken) betrachtet. Eine Überprüfung der Korrelationsmatrizen der unabhängigen Variablen ergab für zwei Variablenkombinationen beim Kombiseilgerät (fm stehend und Anzahl Seillinien sowie Anzahl Seillinien und Länge) eine erhöhte Korrelation (> 0.6). Durch das Weglassen einer der korrelierenden unabhängigen Variablen wurde das Modell jedoch nicht signifikant schlechter oder besser.

Für das Kombiseilgerät blieben am Ende die unabhängigen Variablen Kombiseilgerättyp, Erschwernisse, Nadelholzanteil, fm stehend, %-Nutzung, Anzahl Stützen und Rückerichtung übrig. Wie Tabelle 4 entnommen werden kann, handelt es sich insgesamt um ein plausibles Schätzmodell mit einem multiplen Korrelationskoeffizienten von 0.81. Es werden 65.1% der Gesamtvarianz durch die Regression erklärt.

Für den Seilschlepper blieben die unabhängigen Variablen Nadelholzanteil, fm Mittelstamm stehend, %-Nutzung und Seilschleppertyp übrig. Die Kennzahlen des Modells sind in Tabelle 5 dargestellt.

3 Für eine genaue Dokumentation des Verfahrens siehe die Benutzerdokumentation von SAS v9.1 (SAS Institute Inc. 2004).

Regressionskennzahlen	
Multipler Korrelationskoeffizient R	0.8071
Determinationskoeffizient R^2	0.6513
Adjustiertes R^2	0.6235
Standardfehler	0.1761
Beobachtungen	96

ANOVA					
Quelle der Variation	Freiheitsgrade	Summe d. Quadrate	Mittel d. Quadrate	F-Wert	Wahrsch. > F
Modell	7	5.0985	0.7284	23.4767	1.11E-17
Fehler	88	2.7302	0.0310		
Gesamtsumme	95	7.8286			

	Koeffizienten	Standardfehler	t-Statistik	p-Wert	95%-Vertrauensintervall	
					Untergrenze	Obergrenze
Achsenabschnitt	1.8463	0.1900	9.7155	1.38E-15	1.4687	2.2240
Kombiseilgerättyp	–0.1376	0.0481	–2.8618	0.0053	–0.2331	–0.0420
Erschwernisse	0.1061	0.0431	2.4610	0.0158	0.0204	0.1918
Nadelholzanteil	–0.0048	0.0009	–5.2102	1.23E-06	–0.0067	–0.0030
fm stehend	–0.0002	7.02E-05	–3.2682	0.0015	–0.0004	–8.99E-05
%-Nutzung	–0.0023	0.0010	–2.4156	0.0178	–0.0043	–0.0004
Anzahl Stützen	0.0361	0.0093	3.9067	0.0002	0.0178	0.0545
Rückerichtung	–0.0880	0.0380	–2.3128	0.0231	–0.1636	–0.0124

Tab 4 *Statistische Kennzahlen der multiplen linearen Regression für das Kombiseilgerät.*

Regressionskennzahlen	
Multipler Korrelationskoeffizient R	0.6866
Determinationskoeffizient R^2	0.4714
Adjustiertes R^2	0.4389
Standardfehler	0.1539
Beobachtungen	70

ANOVA					
Quelle der Variation	Freiheitsgrade	Summe d. Quadrate	Mittel d. Quadrate	F-Wert	Wahrsch. > F
Modell	4	1.3729	0.3432	14.4933	1.63E-08
Fehler	65	1.5393	0.0237		
Gesamtsumme	69	2.9123			

	Koeffizienten	Standardfehler	t-Statistik	p-Wert	95%-Vertrauensintervall	
					Untergrenze	Obergrenze
Achsenabschnitt	0.9227	0.1623	5.6853	3.34E-07	0.5986	1.2468
Nadelholzanteil	–0.0040	0.0008	–4.8262	8.79E-06	–0.0057	–0.0023
fm Mittelstamm stehend	–0.0827	0.0274	–3.0137	0.0037	–0.1374	–0.0279
%-Nutzung	0.0037	0.0018	2.0643	0.0430	0.0001	0.0072
Seilschleppertyp	0.0706	0.0247	2.8621	0.0057	0.0213	0.1199

Tab 5 *Statistische Kennzahlen der multiplen linearen Regression für den Seilschlepper.*

Dieses Schätzmodell besitzt einen multiplen Korrelationskoeffizienten von 0.69. Mit diesem Modell werden 47.1% der Gesamtvarianz erklärt.

Berechnung von RMSE, Bias und MAPE

Die Berechnung der Kennzahlen RMSE, Bias und MAPE zur Bewertung der Schätzgüte erfolgte anhand der Formeln (10) bis (12).

Der Bias liegt für beide Holzernteverfahren wie zu erwarten nahe bei 0 h/m³ Holzhauerei und Rücken, während der RMSE für das Kombiseilgerät bei 0.169 h/m³ und für den Seilschlepper bei 0.148 h/m³ Holzhauerei und Rücken liegt. Der MAPE liegt für das Kombiseilgerät bei circa 20.2% und beim Seilschlepper bei 14.7% (Tabelle 6, Zeilen a und e).

Diskussion der Variablen

Wie den p-Werten in Tabelle 4 und 5 entnommen werden kann, ist die unabhängige Variable Nadelholzanteil für beide Holzernteverfahren besonders massgebend, beim Kombiseilgerät ausserdem auch noch die Variable Anzahl Stützen. Die Variable %-Nutzung wurde ebenfalls für beide Verfahren in die Regressionsformel aufgenommen, allerdings mit unterschiedlichen Vorzeichen. Die beiden Gerätetyp-Variablen Kombiseilgerättyp und Seilschleppertyp wurden für beide Verfahren berücksichtigt. Ausserdem wurde für den Seilschlepper die Variable fm Mittelstamm stehend aufgenommen, für das Kombiseilgerät waren es die Variablen Erschwernisse, fm stehend und Rückerichtung.

Die Rückedistanz erwies sich für beide Holzernteverfahren als nicht aussagekräftige unabhängige Variable. Daraus sollte jedoch keine verallgemeinerte Aussage abgeleitet werden, da diese mit zunehmender Distanz immer stärker ins Gewicht fallen könnte. Ebenfalls zu wenig Aussagekraft besassen die Variablen Massnahmenart und Entwicklungs-

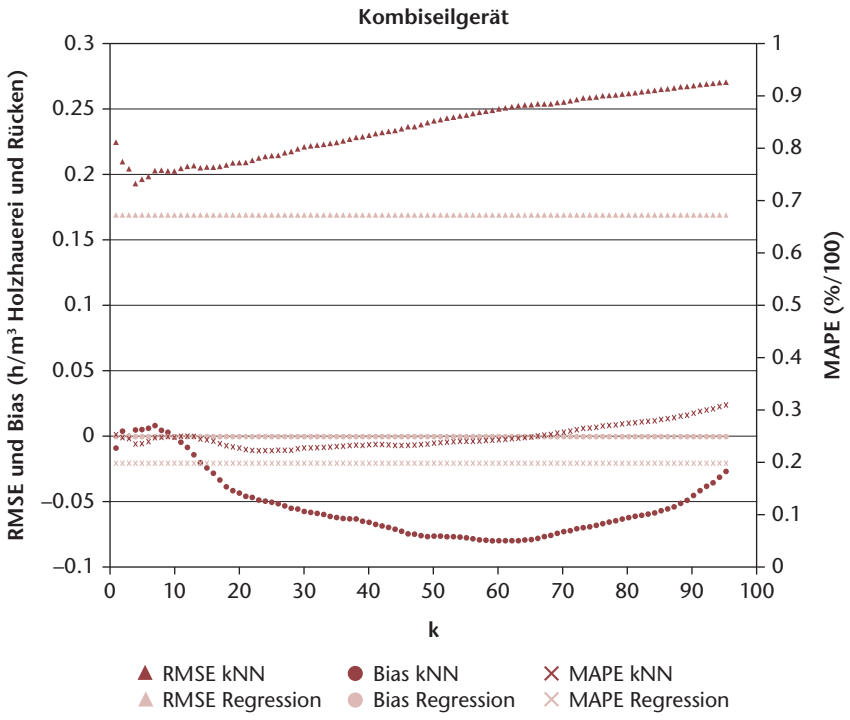

Abb 4 RMSE, Bias und MAPE in Abhängigkeit der Wahl von k für das Kombiseilgerät.

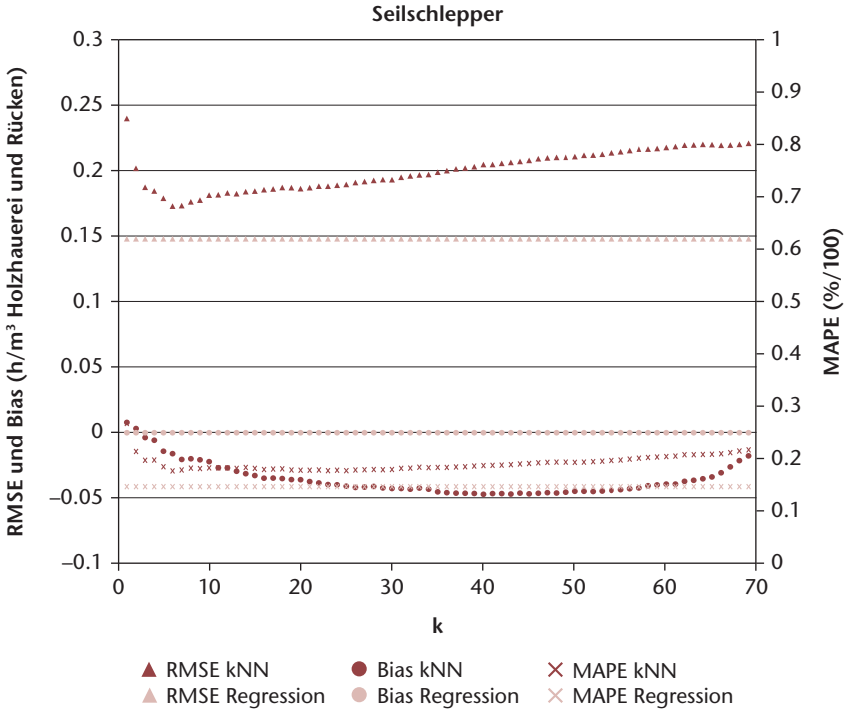

Abb 5 RMSE, Bias und MAPE in Abhängigkeit der Wahl von k für den Seilschlepper.

stufe, was womöglich an der unscharfen Abgrenzung der verschiedenen Eingriffe und Entwicklungsstufen lag, sowie die Variable fm/ha Vorrat. Für das Kombiseilgerätverfahren fanden zudem die beiden Variablen Anzahl Seillinien und Seillinienlänge keine Aufnahme in das Schätzmodell.

kNN-Methode

Nachdem Regressionsformeln für beide Holzerntemethoden hergeleitet wurden und eine Auswahl aus allen verfügbaren unabhängigen Variablen getroffen wurde, soll nun mithilfe derselben Variablen die kNN-Methode auf ihre Schätzgenauigkeit geprüft werden.

Um die Güte der Anpassung der Schätzwerte an die realen Messwerte zu beurteilen, werden der RMSE, der Bias und der MAPE herangezogen (Formeln [10] bis [12]). Diese Werte sind bei der kNN-Methode jedoch abhängig von der Wahl der Anzahl nächster Nachbarn *k*.

Die genannten Kennwerte können somit bei der kNN-Methode umgekehrt auch zur Bestimmung des optimalen *k* verwendet werden. Gesucht ist ein *k*, bei welchem alle drei Kennwerte RMSE, Bias und MAPE möglichst nahe bei 0 zu liegen kommen, beziehungsweise die Summe der im Intervall 0–1 normalisierten Kennwerte minimal ist. Für dieses *k* soll demnach

$$f(k) = RMSE'_k + |Bias_k|' + MAPE'_k \qquad (13)$$

minimal sein, wobei

$$RMSE'_k = \frac{RMSE_k - RMSE_{min}}{RMSE_{max} - RMSE_{min}} \qquad (14)$$

der normalisierte Kennwert des RMSE für ein gegebenes *k* ist ($RMSE_k$ ist der RMSE für ein gegebenes *k*, $RMSE_{max}$ ist der Maximalwert und $RMSE_{min}$ der Minimalwert aller über *k* variierten RMSE). Die Formeln der normalisierten Kennwerte $|Bias_k|'$ und $MAPE'_k$ sind analog.

Um diese Grössen für die kNN-Methode und ein bestimmtes *k* zu berechnen, ist es nötig, ein sogenanntes Kreuzvalidierungsverfahren durchzuführen.[4] Dabei werden die vorhandenen *n* Datensätze in *m* Teilmengen aufgeteilt *(m ≤ n)*. Dann wird in *m* Durchläufen unter Ausschluss einer Testteilmenge T_i mittels aller verbleibenden *m–1* Trainingsteilmengen die i-te Teilmenge T_i geschätzt. Einen Spezialfall bildet die hier gewählte Leave-one-out-Kreuzvalidierung, bei welcher die Zahl der Durchläufe der Zahl der Teilmengen entspricht (m = n), sodass jede Teilmenge genau einmal als Testteilmenge auftritt.

In Abbildung 4 sind RMSE, Bias und MAPE in Abhängigkeit ihrer Wahl von *k* für das Kombiseilgerät und in Abbildung 5 für den Seilschlepper dargestellt. Zum Vergleich wurden überdies die entsprechenden Kennwerte der Regressionsanalyse eingezeichnet.

Für RMSE und Bias liegt das Optimum bei verschiedenen *k*, für den RMSE bei *k = 4* (Kombiseilgerät) beziehungsweise *k = 6* (Seilschlepper) und für den Bias bei *k = 10* (Kombiseilgerät) beziehungsweise *k = 2* (Seilschlepper). Der MAPE nimmt beim Kombiseilgerät einen minimalen Wert bei *k = 23* an (Abbil-

4 Um eine noch bessere Vergleichbarkeit der beiden Schätzverfahren zu erreichen, müsste das Kreuzvalidierungsverfahren auch auf die Regressionsanalyse angewendet werden. Darauf wurde jedoch im Rahmen dieser Studie verzichtet, da von den Autoren im Hinblick auf die Kennziffern RMSE, Bias und MAPE für die Regression keine signifikante Abweichung erwartet wurde.

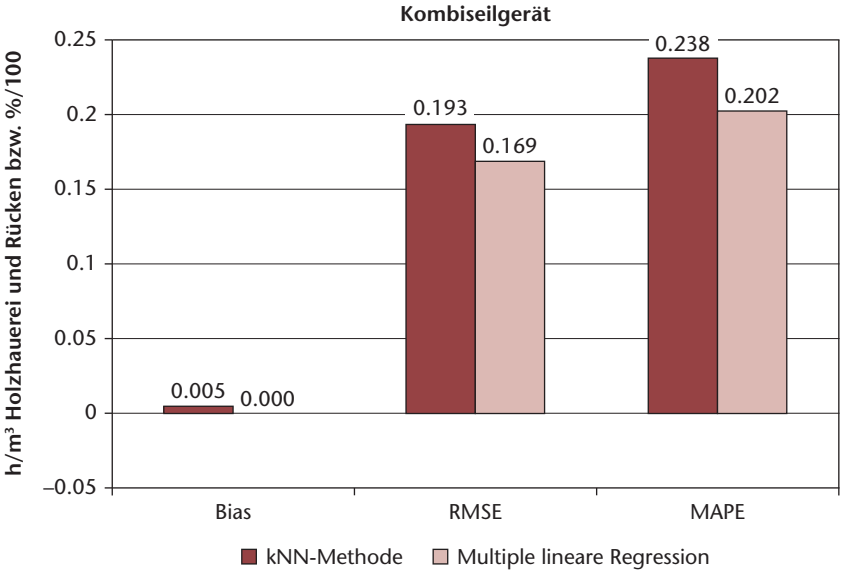

Abb 6 Bias, RMSE und MAPE für das Holzernteverfahren Kombiseilgerät bei k_{opt} = 4.

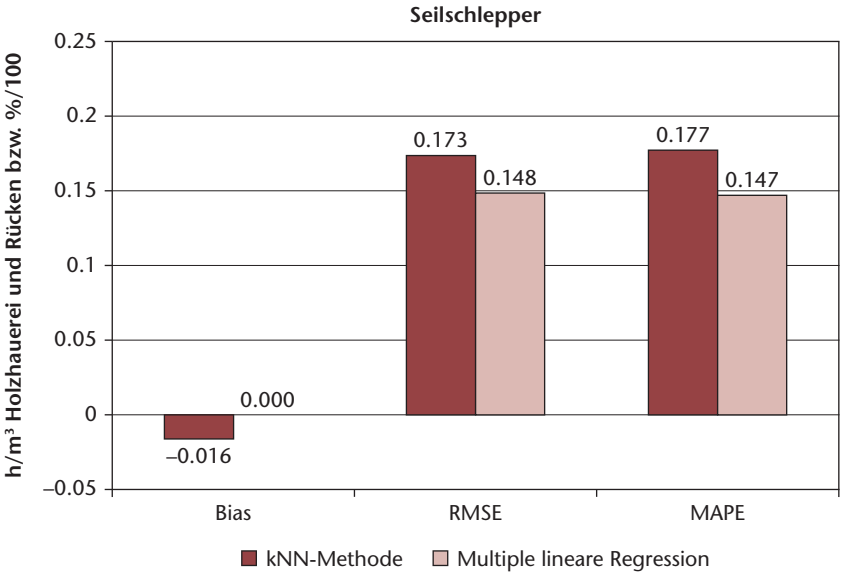

Abb 7 Bias, RMSE und MAPE für das Holzernteverfahren Seilschlepper bei k_{opt} = 6.

dung 4). Für eine gute Wahl von *k* wäre jedoch auch das Zwischenoptimum bei *k = 10* eine nähere Betrachtung wert. Für den Seilschlepper liegt das Optimum in Hinsicht auf den MAPE bei *k = 6* (Abbildung 5).

Das gemäss Formel (13) über alle drei Kennwerte hinweg optimale *k* liegt demnach in unseren Beispielen für das Kombiseilgerät bei *k = 4*, was nahe bei der Empfehlung von Enas & Choi (1986) liegt. Eine Wahl von *k = 10* würde allerdings kein signifikant schlechteres Resultat ergeben und läge näher bei der Empfehlung von Loftsgarden & Quesenberry (1965). Für den Seilschlepper ist das optimale *k = 6*. Damit liegen wir ebenfalls nicht weit von den Empfehlungen von Loftsgarden & Quesenberry (1965) sowie Enas & Choi (1986).

Der Bias liegt beim Kombiseilgerätverfahren bei 0.005 h/m³ (Tabelle 6, Zeile b) und beim Seilschlepperverfahren bei –0.016 h/m³ Holzhauerei und Rücken (Tabelle 6, Zeile f). Der RMSE liegt für das Kombiseilgerätverfahren bei 0.193 h/m³ Holzhauerei und Rücken und der MAPE bei 23.8% (Tabelle 6, Zeile b). Beim Seilschlepperverfahren liegt der RMSE bei 0.173 h/m³ Holzhauerei und Rücken und der MAPE bei 17.7% (Tabelle 6, Zeile f).

Vergleich der Schätzgüte der kNN-Methode und der multiplen Regressionsanalyse

Um zu überprüfen, inwieweit die kNN-Methode als Schätzmethode geeignet ist, werden die Kennzahlen Bias, RMSE und MAPE der multiplen Regression mit denjenigen der kNN-Methode verglichen (Abbildungen 6 und 7).

Es zeigt sich, dass trotz optimalen *k* die kNN-Methode bezüglich der drei Kennwerte in keinem Fall so gut abschneidet wie die Regressionsanalyse. Die Unterschiede zwischen beiden statistischen Verfahren sind allerdings nicht gravierend.[5]

Zur Wahl der Modellvariablen

Eine Schwierigkeit sowohl bei der Regression als auch bei der kNN-Methode besteht in der Wahl eines geeigneten Modells. Das rückwärtsgerichtete Eliminationsverfahren, mit welchem aus sämtlichen möglichen Variablenkombinationen das geeignete Schätzmodell extrahiert wurde, ist zeitaufwendig und bedarf einiger statistischer Kenntnisse, womit es in der Praxis nicht problemlos verwendet werden kann. Durch eine weitere Untersuchung mit sämtlichen pro Holzernteverfahren zur Verfügung stehenden unabhängigen Variablen (Tabelle 1) konnte jedoch gezeigt werden, dass derartige Schätzmodelle sowohl für die Regression als auch für die kNN-Methode nur unwesentlich bessere Resultate in Bezug auf Bias, RMSE und MAPE erzeugen (Tabelle 6).

Natürlich wird unter Einbezug auch nicht signifikanter unabhängiger Variablen in das Modell ein gewisses Rauschen in Kauf genommen. Da sich die Schätzresultate in Bezug auf die drei genannten Kennzahlen in unserer Untersuchung nicht wesentlich verschlechterten oder verbesserten, können Praktiker und Praktikerinnen, welche keinerlei Anhaltspunkte für die Modellwahl haben, in einem ersten Schritt sämtliche unabhängigen Variablen in ihr Modell einbeziehen.

Fazit und Ausblick

Beide Methoden, die kNN-Methode und die multiple lineare Regression, können zur Schätzung von Holzernteproduktivitäten herangezogen wer-

5 Eine nachträglich durchgeführte Residuenanalyse zeigte einen Trend der Residuen sowohl der multiplen linearen Regression als auch der kNN-Methode. Für beide Verfahren werden kleine Werte der unabhängigen Variablen generell über- und grosse Werte generell unterschätzt. Die Autoren sind der Meinung, dass dieses Phänomen auf die zugrunde liegenden Daten zurückzuführen ist. Weitere Untersuchungen sind geplant.

		Bias (h/m³)	RMSE (h/m³)	MAPE (%)
Kombiseilgerät (Modell mit signifikanten Variablen)				
a)	Multiple lineare Regression	–2.822 E-16	0.169	20.2
b)	kNN-Methode (k = 4)	0.005	0.193	23.8
Kombiseilgerät (Modell mit allen verfügbaren Variablen)				
c)	Multiple lineare Regression	–5.447 E-16	0.163	20.0
d)	kNN-Methode (k = 4)	0.004	0.193	23.7
Seilschlepper (Modell mit signifikanten Variablen)				
e)	Multiple lineare Regression	–7.296 E-17	0.148	14.7
f)	kNN-Methode (k = 6)	–0.016	0.173	17.7
Seilschlepper (Modell mit allen verfügbaren Variablen)				
g)	Multiple lineare Regression	–1.967 E-16	0.141	14.3
h)	kNN-Methode (k = 6)	–0.013	0.182	18.4

Tab 6 *Vergleich von Bias, RMSE und MAPE der Schätzmodelle, in welche alle pro Holzernteverfahren zur Verfügung stehenden unabhängigen Variablen eingeflossen sind, mit den optimalen Schätzmodellen.*

den. Die Prognosegenauigkeit liegt in derselben Grössenordnung. Die Gründe für die lediglich in einem akzeptablen Bereich liegende Schätzgüte sind nach Meinung der Autoren weniger in den Schätzverfahren zu suchen als in einer zu schmalen Variationsbreite der unabhängigen Variablen der zugrunde liegenden Daten.

Beide Schätzmethoden haben Vor- und Nachteile und ihre eigenen Anwendungsschwerpunkte.

Regression

- Die Annahmen über die funktionalen Zusammenhänge zwischen den unabhängigen und der abhängigen Variablen werden durch die Regressionsformel explizit gemacht. Aufgrund der einfachen Formel ist die Regression leicht anzuwenden und besonders geeignet für den Einsatz in Simulationsmodellen und für Sensitivitätsanalysen.
- Die Anwendung der Regressionsformel erfordert keinen Zugriff auf die zugrunde liegenden Daten.

kNN-Methode

- Aufgrund der einfachen Verständlichkeit der kNN-Methode sowie der Visualisierbarkeit der *k* nächsten Nachbarn ist sie besonders geeignet für Praktiker und Praktikerinnen, die einzelne Ereignisse schätzen wollen.
- Die kNN-Methode ist robust und nicht parametrisch, das heisst, sie setzt im Gegensatz zur Regressionsanalyse nicht die Wahl eines bestimmten Modelltyps (z.B. linear, exponentiell, logarithmisch) voraus. Im Gegensatz zur Regression müssen bei der Anwendung der kNN-Methode funktionale Zusammenhänge zwischen den unabhängigen und der abhängigen Variablen nicht bekannt sein (Fehrmann et al 2008), sie werden allerdings durch die kNN-Methode auch nicht aufgezeigt. Weiter können zusätzliche unabhängige Variablen nachträglich ins Modell aufgenommen werden, ohne dass das Schätzmodell neu gerechnet werden muss. Allerdings erfordert die kNN-Methode eine gute Wahl von *k*. Richtlinien können hierbei hilfreich sein, z.B. $k = \sqrt{n}$ *oder* $k = n^{3/8}$, wobei *n* die Anzahl Referenzereignisse darstellt.
- Die kNN-Methode nutzt lokale Abweichungen von einem grösseren Trend besser aus als eine lineare Regression (Fehrmann et al 2008).
- Die kNN-Methode hat die Tendenz, systematisch kleinste Werte zu über- und grösste Werte zu unterschätzen. Eine Verringerung von *k* verkleinert zwar diesen Bias, erhöht jedoch gleichzeitig die Varianz (Maltamo et al 2003, Malinen 2003).
- Die Anwendung der kNN-Methode berücksichtigt immer sämtliche vorhandenen Datensätze. Sie kann so auf eine laufend aktualisierte Datenbasis zugreifen, sie setzt aber den Zugriff auf sämtliche Daten notwendigerweise voraus. Die Schätzmodelle veralten aufgrund neu hinzugefügter Datensätze im Gegensatz zur Regression nicht.

Eine Extrapolation über den durch Daten unterlegten Bereich hinaus ist weder bei der Regressionsanalyse noch bei der kNN-Methode ohne Weiteres möglich. Eine Schwierigkeit, die überdies bei beiden Schätzmethoden zu bewältigen ist, ist die Auswahl geeigneter unabhängiger Variablen.

Mit der kNN-Methode beziehungsweise mit der vorgestellten Software kNN-Workbook lassen sich natürlich nicht nur Holzerntedaten auswerten sowie diesbezügliche Leistungen und Kosten schätzen. Grundsätzlich sind die Methode und das Instrument auch auf andere Datensätze und Schätzgrössen anwendbar. Zum Beispiel liesse sich auch die Sortimentszusammensetzung von Holzschlägen (Anteile an Säge-, Industrie- und Energieholz) gestützt auf Erfahrungszahlen prognostizieren.

Für eine Weiterentwicklung des Ansatzes wäre es überlegenswert, eine zentrale Holzerntedatenbank aufzubauen. Repräsentativ über die Schweiz verteilte Testbetriebe würden sich gegebenenfalls gegen Entschädigung oder Einräumung spezieller Nutzungsrechte dazu verpflichten, zuvor festgelegte Variablen ihrer Holzschläge zu erfassen und diese Daten in die Datenbank einzuspeisen (Frutig et al 2009). Die Datenerfassung sowie die Schätzung von Produktivitäten und Kosten mittels der kNN-Methode würden über Internet und Webservices[6] erfolgen.

Es ist vorgesehen, die auf Excel 2007 basierende Software kNN-Workbook einem breiteren Publikum, insbesondere den Forstbetriebsleitern, über die Website der WSL gratis zur Verfügung zu stellen. Die Autoren sind interessiert an Rückmeldungen von Personen, die Instrument und Methode getestet haben. ■

Eingereicht: 23. Februar 2011, akzeptiert (mit Review): 31. August 2011

6 Mithilfe von Webservices können Daten automatisiert ausgetauscht oder Funktionen auf entfernten Rechnern aufgerufen werden.

Literatur

DUDA RO, HART RE, STORK DG (2001) Pattern classification. New York: John Wiley, 2 ed. 654 p.

ENAS GG, CHOI SC (1986) Choice of the smoothing parameter and efficiency of k-nearest neighbour classification. Comput Math Appl 12: 235–244.

FEHRMANN L (2006) Alternative Methoden zur Biomasseschätzung auf Einzelbaumebene unter spezieller Berücksichtigung der k-Nearest Neighbour (k-NN) Methode. Göttingen: Georg-August-Univ, Fakultät Forstwissenschaften Waldökologie, Dissertation. 155 p.

FEHRMANN L, LEHTONEN A, KLEINN CH, TOMPPO E (2008) Comparison of linear and mixed-effect regression models and a k-nearest neighbour approach for estimation of single-tree biomass. Can J For Res 38: 1–9. doi: 10.1139/X07-119

FELBER A (2005) Prognose der lokalen Waldbrandgefahr im Tessin durch Modellierung mit Hilfe der k-Nearest Neighbors (kNN) Methode. Basel: Univ Basel, Naturwissenschaftl Fakultät, PhD Thesis. 143 p.

FRUTIG F, THEES O, LEMM R, KOSTADINOV F (2009) Holzernteproduktivitätsmodelle HeProMo – Konzeption, Realisierung, Nutzung und Weiterentwicklung. In: Thees O, Lemm R, editors. Management zukunftsfähige Waldnutzung. Zürich: Vdf. pp. 441–466.

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

LOFTSGAARDEN D, QUESENBERRY C (1965) A nonparametric estimate of multivariate density function. Ann Math Stat 36: 1049–1050.

MALINEN J (2003) Locally adaptable non-parametric methods for estimating stand characteristics for wood procurement planning. Silva Fenn 37: 109–120.

MALTAMO M, MALINEN J, KANGAS A, HÄRKÖNEN S, PASANEN AM (2003) Most similar neighbour-based stand variable estimation for use in inventory by compartments in Finland. Forestry 76: 449–464.

Ein Instrument zur Schätzung von Holzernteproduktivitäten mittels der kNN-Methode

Für die betriebliche Planung und Steuerung der Holzernte ist es wichtig, zuverlässige Informationen über ihre Zeitbedarfe und Kosten zu haben. Um diese effizient und verlässlich schätzen zu können, bedarf es entsprechender Methoden und Kalkulationsinstrumente. In vorliegendem Beitrag wird geprüft, ob sich die Methode der k nächsten Nachbarn (kNN-Methode) hierzu eignet. Dazu wird die Funktionsweise der kNN-Methode erläutert und anhand zweier Datensätze zu den Holzernteverfahren «Kombiseilgerät» und «Seilschlepper» die Schätzgenauigkeit der kNN-Methode geprüft. Es zeigt sich, dass ihre Schätzgenauigkeit in derselben Grössenordnung wie diejenige einer multiplen linearen Regression liegt. Vorteile der kNN-Methode sind die einfache Verständlichkeit und gute Visualisierbarkeit sowie der Umstand, dass die Schätzmodelle nicht veralten, weil laufend neue Datensätze berücksichtigt werden können. An der Eidgenössischen Forschungsanstalt für Wald, Schnee und Landschaft wurde daher das kNN-Workbook entwickelt, ein Softwareinstrument, mit welchem beliebige Datensätze in der Praxis mit der kNN-Methode analysiert werden können. Dieses Instrument wird im Artikel ebenfalls vorgestellt.

Un instrument pour évaluer la productivité de la récolte des bois selon la méthode kNN

Pour la planification et la gestion opérationnelles de la récolte des bois, il importe de disposer d'informations fiables sur le temps requis et les coûts engendrés. Des méthodes et des instruments de calcul appropriés sont dès lors nécessaires afin que ces paramètres soient estimés avec efficacité et fiabilité. Le présent article examine dans quelle mesure la méthode des k plus proches voisins (méthode kNN) s'adapte à ce dessein. Pour ce faire, l'algorithme des k plus proches voisins est d'abord décrit avant d'être appliqué à deux sets de données sur les méthodes de récolte des bois «câble-grue et processeur combiné» et «débusqueur». Il s'avère que l'exactitude de l'estimation de la méthode kNN se situe dans le même ordre de grandeur que celle d'une régression linéaire multiple. Les avantages de la méthode kNN résident dans sa compréhensibilité et la bonne visualisation ainsi que dans le fait que les modèles ne deviennent pas obsolètes, car continuellement complétés par de nouvelles données. A l'Institut fédéral de recherches WSL a ainsi été développé un workbook kNN, un instrument logiciel qui permet d'analyser des données de la pratique avec la méthode kNN. Cet instrument est également présenté dans cette contribution.