Reconstructing Frames of Reference

Auf den Spuren der Datenanalyse Lewis R. Binfords

Clemens Schmid

SS 17

Inhaltsverzeichnis

Einleitung und Kontextualisierung

Lewis R. Binfords umstrittenes Spätwerk *Constructing Frames of Reference* ist eines der wichtigsten Standardwerke der New Archaeology. Es versteht sich selbst als Methodikstudie zur induktiven Ableitung allgemeiner Regeln menschlichen Verhaltens aus ethnographischen Umwelt- und Sozialdatensätzen, erarbeitet aber am Fallbeispiel eines Datensatzes zu Jäger- und Sammlergruppen gleichermaßen inhaltlich relevante Ergebnisse.

Binford fasst seine methodischen Ambitionen folgendermaßen zusammen:

> I cannot emphasize strongly enough, that the major problem this book addresses is *the development of a method for productively using ethnographic data in the servive of archaeological goals*. […] This book is unapologetically written from a scientific perspective. It is largely an exercise in inductive reasoning, in that it asks questions regarding the character of the world of organized variability among ethnographically documented hunter-gatherer groups. […] And, since one of the goals of this book is to explain variability among hunter-gatherers, the explanatory theory that I have developed is available for archaeologists to use deductively by reasoning to or simulating changing conditions and thereby providing patterns of change that can be expected to occur in the archaeological record at specific locations.
>
> – Binford (2001), 2-3.

Die systematische Suche nach übergeordneten, wiederkehrenden Strukturen in der Mensch-Umwelt und Mensch-Mensch Beziehung ist eines der wesentlichen Themen des Buches. Die Komplexität dieser Aufgabe erklärt seinen bemerkenswerten Umfang und den dennoch teilweise fragmenthaften Charakter. Binford konstruiert allgemeine Lehrsätze, die für alle Jäger- und Sammlergruppen Gültigkeit beanspruchen. Erarbeitet auf Grundlage eines begrenzten Datensatzes und eines limitierten, analytischen Methodensets sind diese Axiome, qualitative und quantitative Thesen wissenschaftliche Aussagen. Die ihnen zugrundeliegenden Analysen sollten reproduzierbar und falsifizierbar sein. Zur Illustration ein Beispiel einer zufällig ausgewählten These, die im zehnten Kapitel formuliert wird:

> Proposition 10.19: As packing increases, groups that are dependent upon aquatic resources should resort to more complex subsistence technology. Increasing complexity in the design of weapons should also be associated with hunter-gatherer groups that are not primarily dependent upon aquatic resources as a function of their more specialized exploitation of a reduced number of high-yield species (see generalizations 10.15 and 10.16).
>
> – Binford (2001), 392.

Hier wird zunächst ein Rahmen definiert (“groups that are dependent upon aquatic resources”) und für den Fall eines Veränderungsprozesses (“as packing increases”) eine Vorhersage (“should resort to more complex subsistence technology”) getroffen, die dann noch weiter kontextualisiert und präzisiert wird.

Da sowohl die Daten als auch eine – unterschiedlich ausführliche – Beschreibung des Methodenset publiziert sind, die der Entwicklung dieser und aller anderen Aussagen in *Constructing Frames of Reference* zugrunde liegen, sollte es möglich sein, …

1. … auf Grundlage der selben Daten unter Anwendung der selben Methode zu den gleichen Aussagen zu kommen.
2. … Aussagen mit anderen Daten und anderen Methoden zu rekonstruieren und gegebenenfalls durch verbesserte Aussagen zu ersetzen.

Der vorliegenden Aufsatz dient auch dazu, Binfords wissenschaftlichen Selbstanspruch zu prüfen. Wie gut ist das Buch für einen zeitgemäßen Reproducible Research Workflow zugänglich? Gleichzeitig muss der Schwerpunkt dieser Arbeit ein didaktischer sein: Ich werde den Umgang mit Grundlagen des wissenschaftlichen Arbeitens üben und – im Kontext des Hauptseminars – meine Methodenkompetenz zur Modelleinpassung mittels Regressionsanalyse[1] ausbauen.

Vor diesem Hintergrund möchte ich weder versuchen die archäologischen oder archäologietheoretischen Leitgedanken des Buches nachzuzeichnen,[2] noch seine kontroverse Rezeption aufzugreifen. Kritiker haben Binford neben übermäßigem Naturdeterminismus und Funktionalismus u.a. die Verwendung widersprüchlicher und mangelhaft definierter Neologismen, Rechenfehler und Mängel in der Datenaufnahme bis hin zum Übersehen zentraler Trends vorgeworfen.[3] Die fachtheoretischen Grundlagen des Buches in Kulturökologie und Middle-Range-Theorie gehören in ein ohnehin umstrittenes Feld. Eine Auseinandersetzung mit dem Gesamtwerk ist im zeitlich stark begrenzten Rahmen einer Seminararbeit nicht sinnvoll möglich. Stattdessen möchte ich das Opus höchst selektiv betrachten, berechtigte Kritik vorerst beiseite lassen und mich voll auf einen kleinen Aspekt der explorativen Datenanalyse konzentrieren. Eine erste Abenteuerreise auf den Spuren Binfords, wie sie noch häufig unternommen werden sollte.

> Does it [Binfords *Constructing Frames of Reference*] help us to better suppose hunter-gatherer variability, and to conceptualize variability beyond what we currently know? The answer to the first question is a definite yes, but its generalizations require rigorous testing and replication, the normal scientific process that Binford himself has advocated for four decades.
>
> Ames (2004), 372.

[1] Nakoinz und Knitter (2016), 87-105.

[2] Für einen kurzen Abriss siehe z.B. Donald Pate (2005) oder Browman (2005)

[3] Hill (2002), Ames (2004)

Datensätze

Binfords komplexes Unterfangen erfordert Forschungsdaten, die sowohl für Kultur- als auch Naturphänomene ein möglichst breites Set an Observationen enthalten. Dabei muss einerseits die Anzahl als auch die Variabilität der Beobachtungen ausreichend groß, andererseits auch ein sinnvolles Set an Kennwerten und Proxies erfasst sein. Die Zusammenstellung der anthropologischen Daten ist besonders problematisch, da der Vergleich eine selten angewandte Systematisierung ethnographischer Datenaufnahme erfordert.

> It took me two years to develop the data bases dealing with the world's environments and the geographical distribution of documented hunter-gatherers. Once this aspect of the work was completed, it became clear that the limited range of hunter-gatherer characteristics upon which cross-cultural studies had focused was not really relevant to most of the issues that I hoped to address in my book.
>
> – (Binford 2001, 2.)

Ein großer Teil des Buches beschäftigt sich mit der Beschreibung und Kontextualisierung von Variablen, die Binford aus der geographischen oder anthropologischen Literatur entnommen und anschließend gesammelt oder gegebenenfalls berechnet hat. Das Ergebnis ist ein komplexer Datenbestand, der im Verlauf des Buches immer weiter in die Breite wächst. Es ist Amber Johnson, Doug White und Anthon Eff zu verdanken, dass der Datensatz heute in gegenüber der abgedruckten Version noch einmal deutlich erweiterter Form digitalisiert und leicht zugänglich vorliegt.[4] Über ein Paket der Statistikprogrammiersprache R, das Ben Marwick zusammengestellt hat, lässt sich auf die Daten besonders bequem zugreifen.[5]

Binford hat mit zwei Hauptdatensätzen gearbeitet: Eine Tabelle mit Informationen zu 339 ethnographisch aufgenommenen Jäger- und Sammlergruppen und eine Tabelle mit Atmosphärendaten von 1429 weltweit verteilten Wetterstationen, die anhand ihrer Position in verschiedenen Vegetationszonen ausgewählt wurden. Johnson, White und Eff haben ersteren Datensatz strukturell überarbeitet und eine Auswahl von 507 gesammelten und berechneten Variablen zusammengestellt. Dieser Datensatz **LRB** (im folgenden auch "Gruppendatensatz") liegt entsprechend in Form einer .csv-Tabelle mit 339 Zeilen und 507 Spalten vor. Der Metadatensatz **LRBkey** steht ebenfalls als .csv-Tabelle zur Verfügung und enthält zu jeder der 507 Variablen Informationen wie semantische Kurzbeschreibung, Skalenniveau und Fehlstellen. Auch der Wetterstationendatensatz ist in dieser Form zugänglich.

Gerade letzterer ist relativ einfach systematisch erweiterbar: Für die Berechnung der von Binford hinzugefügten, abhängigen Größen sowohl im Wetterstationen- als auch im Gruppendatensatz kann auf die Java-Software EnvCalc2.1 zurückgegriffen werden. Sie ist seit 2001 aus dem von Binfords Arbeitsgruppe entwickelten Programmcode hervorgegangen und wurde zuletzt 2014 aktualisiert.

Modelle zur Beschreibung der Ausbreitungsarealgröße von Jäger- und Sammlergruppen

Problemstellung

Im 5. Kapitel "Designing Frames of Reference and Exploring Projections" beschreibt Binford unter anderem eine Methode, Vorhersagen zu Attributen von Jäger- und Sammlergruppen in globalem Maßstab auf Grundlage von ethnographischen und naturräumlichen Daten treffen und über Projektion auf Karten visualisieren zu können. Im Abschnitt "Projecting Hunter-Gatherer Populations to the Entire Earth" gibt es wiederum einen Unterabschnitt "Using Relational Projections as Frames of Reference", der das Vorgehen anhand eines Beispiels illustriert. Binford schreibt:

> If I can develop continously scaled equations that summarize the relationship between the properties of hunter-gatherer systems and suites of environmental variables, it is likely that these equations could be used to project estimates for habitats from which there are few, if any, actual cases of hunter-gatherers documented in the resent past. But since such equations summarize interactive ecological relationships that are not confined to particular time periods, they may furnish strong clues about hunter-gatherer organizational variability that will provide a strong platform for subsequent theory building.
>
> – Binford (2001), 154.

Das Beispiel konzentriert sich auf die Variable *area* – die Größe des Areals, das von einer Jäger- und Sammlergruppe relativ exklusiv genutzt wird gemessen in Vielfachen von 100km². Mittels multipler Regression auf Grundlage des Gruppendatensatzes kommt Binford zu folgender Gleichung 1 (bzw. umgeformt Gleichung 2), die die abhängige Variable *area* in Relation zu mehreren unabhängigen Variablen (siehe Tabelle 1) beschreibt:

$$
\begin{aligned}
area = \\
&10 \hat{} [3.421431+ \\
&(0.004732 * hunting)+ \\
&(-0.387229 * lbio5)+ \\
&(0.186574 * lcoklm)+ \\
&(-0.110286 * lrunoff)+ \\
&(0.175157 * watrgrc)+ \\
&(-0.164604 * medstab)+ \\
&(-0.743144 * perwltg)+ \\
&(0.004706 * rlow)+ \\
&(-0.080339 * rungrc)+ \\
&(0.024755 * sdtemp)]
\end{aligned}
\tag{1}
$$

Binford hat seine Analyse in SPSS (Version 6.1.2) ausgeführt. Ein Skriptprotokoll der Analysesession liegt mir nicht vor. Um die Ergebnis in Form der Modellgleichung zu reproduzieren, werde ich nun also zunächst versuchen, das Vorgehen so gut wie möglich nachzuvollziehen. Dafür steht mir eine hoffentlich gleiche oder zumindest hochgradig ähnliche Version des oben beschriebene Gruppendatensatz zu Verfügung. Ich weiß weiterhin, dass Binford sein Modell mit der Methode schrittweiser, Multipler Regression ermittelt hat. Unabhängige Variablen, die sich kollinear zu anderen unabhängigen Variablen verhalten, hat er entfernt. Die abschließende Entscheidung über das beste Modell hat er unter Beachtung der Indikatorgrößen R^2 und Standardfehler getroffen. Diese Angaben sind leider nicht hinreichend präzise.

Datensatz

Ein wesentliches Wissensdefizit besteht bezüglich der Information, in welcher Reihenfolge und mit welcher Rechtfertigung in den Schritten der Multiplen Regression Variablen jeweils entfernt wurden. Schon die Angabe, welche Variablen im Ausgangsdatensatz berücksichtigt wurden, ist unscharf. Im-

[4] http://ajohnson.sites.truman.edu/data-and-program/ [14.8.2017]

[5] Marwick u. a. (2016)

merhin: Da Multiple Regression nur auf Variablen der Intervall- oder Verhältnisskala (zusammen auch Kardinalskala oder metrisch skalierte Variablen) anwendbar ist, ist eine erste Eingrenzung möglich. Der Metadatensatz LRBkey verfügt hierzu über die Spalte *type*, die zu jeder Variable ein Skalenniveau angibt. Leider wird nur zwischen "categorical" und "ordinal" unterschieden. Dabei werden alle Variablen jenseits der Nominalskala als "ordinal" angesprochen. Das genügt nicht, um automatisiert alle intervall- und verhältnisskalierten Variablen auszuwählen. Aus diesem Grund habe ich selbst die Spalte *type_exp* im Metadatensatz hinzugefügt, und nach meiner Einschätzung auf Grundlage des Wertebereichs und der Beschreibung eine Zuordnung zu einem der vier Skalenniveaus "nominal", "ordinal", "interval" und "ratio" vorgenommen. Tabelle 2 illustriert die Unterschiede zwischen der vorhandenen und meiner neu unternommenen Zuordnung für ein paar zufällig ausgewählte Variablen. Abbildung 1 zeigt, wie sich die Reevaluation durch die in *type_exp* deutlich akzentuiertere Verteilung der Skalenniveaus auswirkt.

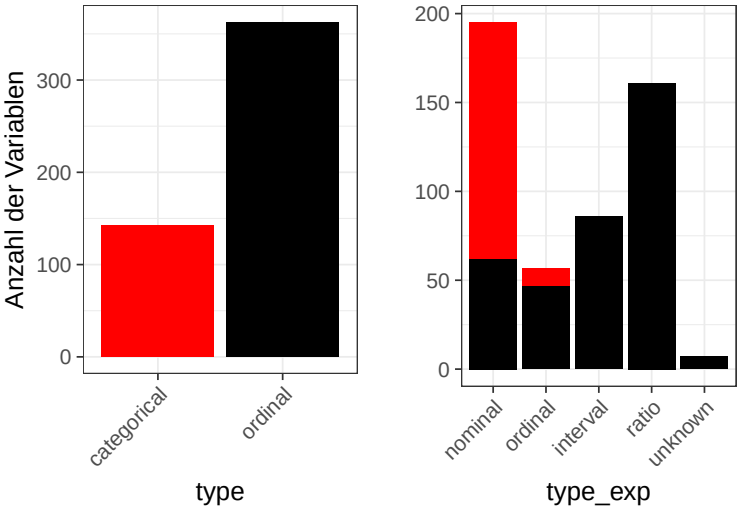

Abbildung 1: Verteilung der Skalenniveauzuordung in den Variablen *type* und *type_exp* des Metatdatensatzes. *type_exp* habe ich hinzugefügt, um Variablen automatisiert nach ihrer Skalenniveauzuordnung auswählen zu können. Die Klassenzuordnung in *type* ist farblich auf *type_exp* abgetragen.

Auf dieser Grundlage ist es jetzt also möglich, einen Ausgangsdatensatz zusammenstellen, der zwar alle 339 Gruppen aber nur die 247 interval- und ratioskalierten Variablen enthält. Hier fällt allerdings gleich ein erstes Defizit dieser Selektion auf: Einige Variablen haben sehr wenige Einträge, d.h. der Wert der entsprechenden Variable wurde nur für wenige Gruppen aufgenommen. Abbildung 2 enthält ein Histogramm der Fehlstellenanzahl.

Immerhin 207 der 247 ($\approx$ 84%) Variablen besitzen überhaupt keine Lehrstellen. Der Datensatz ist bemerkenswert vollständig. Ich habe mich entschieden, alle 33 Variablen, bei denen mehr als 1/3 der Werte fehlen aus der Analyse auszuschließen, um Problemen bei der Regressionsanalyse vorzubeugen. In Abbildung 2 ist die Demarkationslinie rot eingetragen. Bei den gruppenbezogenen Beobachtungen ist das Bild insgesamt ausgeglichener: Für alle Gruppen liegt eine große Menge an Werten vor. Hier sind keine Änderungen erforderlich.

In einem letzten Schritt muss nun noch die Variable *area* durch die Variable *larea* ausgetauscht werden. Binford gibt im Buch zwar die Modellergebnisgleichung für *area* an, das errechnete Modell bezieht sich aber auf den Logarithmus zur Basis 10 von *area* (siehe Umformung von Gleichung 1 zu 2). Der damit vorbereitete Arbeitsdatensatz wird im folgenden als `sel3` bezeichnet.

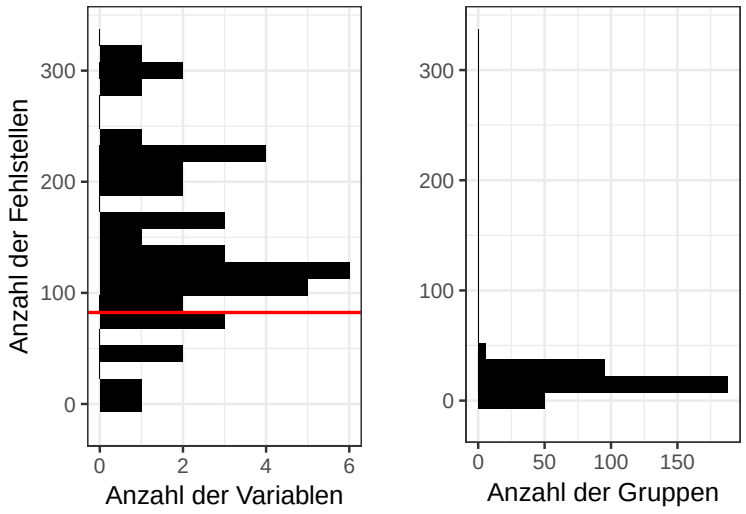

Abbildung 2: Verteilung der Fehlstellenanzahl in metrisch skalierten Variablen und den zugehörigen Beobachtungen (Gruppen). Die Klassenbreite der Histogramme beträgt 15. Variablen und Gruppen ohne Fehlstellen wurden für die Visualisierung ausgeschlossen. Die horizontale, rote Linie im Variablenhistogramm markiert die Grenze oberhalb der Variablen aus der weiteren Analyse entfernt wurden.

$$
\begin{aligned}
\log_{10} area &= \\
larea &= \\
&\quad 3.421431+ \\
&\quad (0.004732 * hunting)+ \\
&\quad (-0.387229 * lbio5)+ \\
&\quad (0.186574 * lcoklm)+ \\
&\quad (-0.110286 * lrunoff)+ \\
&\quad (0.175157 * watrgrc)+ \\
&\quad (-0.164604 * medstab)+ \\
&\quad (-0.743144 * perwltg)+ \\
&\quad (0.004706 * rlow)+ \\
&\quad (-0.080339 * rungrc)+ \\
&\quad (0.024755 * sdtemp)
\end{aligned}
\tag{2}
$$

Multiple Regression

Multiple Lineare Regression ist ein Verfahren der Multivariaten Statistik, das die Erklärung und Vorhersage einer abhängigen Variable durch mehrere unabhängige Variablen erlaubt.[6] Die Regressionsparameter (Koeffizienten) können wie bei der Einfachen Regressionsanalyse durch Reduktion der Fehlerquadrate berechnet werden. In einem ersten Schritt möchte ich das Modell von Binford nachstellen, indem ich seine Auswahl an Eingabevariablen übernehme und die Regression mittels der Funktion `lm()` aus dem R Basispaket stats darauf anwende.

```
binford_model <- lm(
  larea ~ hunting + lbio5 + lcoklm +
          lrunoff + watrgrc + medstab + perwltg +
          rlow + rungrc + sdtemp,
  data = sel3
)
```

[6] Backhaus u. a. (2008), 52-53/64-65., De Veaux, Velleman und Bock (2012), 784-812.

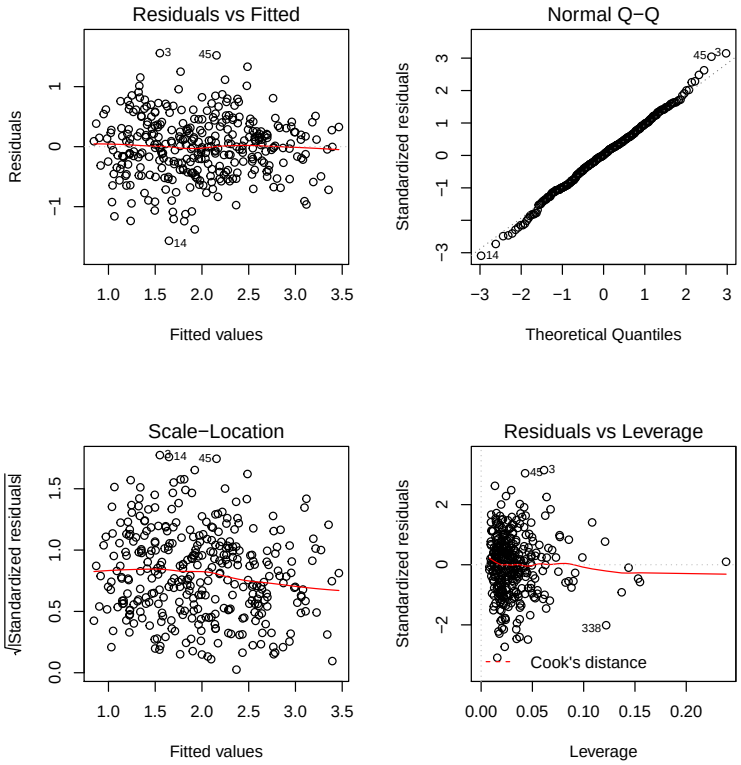

Abbildung 3: Diagnostische Plots für das mit Binfords Variablenauswahl reproduzierte Modell.
Residuals vs Fitted: Streuung der Residuen in Abhängigkeit von der Modellvorhersage. Kann als Indikator für nicht lineare Trends dienen.
Normal Q-Q: Sortierte Residuenwerte abgetragen auf die theoretischen Quantile einer Normalverteilung. Zeigt, inwiefern die Residuenverteilung der Normalverteilung entspricht.
Scale-Location: Vergleiche Residuals vs Fitted. Die Umskalierung erlaubt es, die Homogenität der Residuenvarianz (Homoskedastizität) besser zu beurteilen.
Residuals vs Leverage: Standardisierte Residuenwerte abgetragen auf *Leverage*, ein Maß zur Einschätzung des Einflusses auf das Modellergebnis. Dient zum Identifizieren von einflussreichen Ausreißern.[7]

Tabelle 3 enthält Koeffizienten und zugehörige Kennwerte des erzeugten Modells.[8] Ein Vergleich der Werte mit jenen in Gleichung 2 ergibt, dass die Koeffizienten von den von Binford errechneten abweichen. Ich nehme an, dass der Regressionsalgorithmus in SPSS geringfügig anders implementiert ist als derjenigen in `lm()` oder SPSS die Regression nicht mittels einfacher Reduktion der Fehlerquadrate durchführt. Unter dem Hyperonym *Robuste Regression* werden verschiedene andere Verfahren diskutiert.[9] Da die Werte nur geringfügig divergieren und Größenordnung sowie Vorzeichen übereinstimmen, gehe ich von einer nicht relevanten Abweichung aus, die ich hier nicht weiter diskutieren möchte. Tabelle 4 gibt einige Kennwerte der Modellgüte und Abbildung 3 die vier diagnostischen Standardplots wieder, die die Funktionen `summary.lm()` und `plot.lm()` aus stats bereitstellen. Das Modell scheint ein hohes Erklärungspotential zu besitzen, die abhängige Variable *area* also gut zu beschreiben. Starke Ausreißer gibt es nicht – die überwiegende Mehrzahl der Beobachtungen wird durch das Modell gut erklärt. Nur einzelne Beobachtungen zeigen Auffälligkeiten: 3: Kubu, 14: Agta (North Luzon), 45: Tehuelche, 338: Tareumiut. *watrgrc* ist nicht gut zur Beschreibung von *area* geeignet.

Ich möchte nun versuchen, selbst ein Modell für die Variable *area* zu erstellen. Dabei sind eigentlich einige Aspekte zu beachten,[10] mit denen ich in diesem Experiment aber bewusst sehr inkonsequent umgehen möchte:

1. **Linearität:** Zwischen der abhängigen und jeder einzelnen unabhängigen Variable sollte eine lineare, geradlinige Beziehung bestehen (“straight enough condition”).
2. **Unabhängigkeit:** Die Fehlerverteilung der unabhängigen Variablen sollte unabhängig voneinander sein (“randomization condition”).
3. **Varianzäquivalenz:** Die Variabilität der Fehler innerhalb der Beobachtungen der unabhängigen Variablen sollte näherungsweise gleich sein. Die Fehler sollten gleichmäßig streuen und keine Trends ausbilden.
4. **Normalität:** Die Abweichungen der Messwerte sollten rund um das eingepasste Ergebnismodell normalverteilt sein (“nearly normal condition”).

Für jedes dieser Kriterien gibt es diagnostische Plots, die visuell ausgewertet werden müssen. Auf dieser Grundlage kann dann eine Entscheidung über Variablen getroffen werden, die in das Modell aufgenommen werden sollen. Mir stehen allerdings 213 potentielle Eingabevariablen zur Verfügung und die Vorabprüfung dieser Variablen hätte viel Zeit in Anspruch genommen. Hinzu kommt, dass das Ergebnis der Multiplen Regression zwar tatsächlich im wesentlichen von den Eingabevariablen abhängig ist, andererseits aber auch – zumindest wenn sie nicht vollständig unkorreliert sind – von deren Eingabereihenfolge. Da im Fall der vorliegenden Analyse keine theoretischen oder sachlogischen Überlegungen Eingang finden sollen, die die Variablenauswahl determinieren würden, müssten eigentlich alle Permutationen von Auswahl und Reihenfolge betrachtet werden. Bei 213 unabhängigen Variablen ist schon die Anzahl der Permutationen weit größer als praktisch in irgendeiner Form verarbeitbar ($200! \approx 7.887 * 10^{374}$).

Ich habe mich entschieden, hier auf eine händische Variablenvorauswahl komplett zu verzichten und stattdessen auf ein automatisches Verfahren der Schrittweisen Regressionsanalyse zurückzugreifen. Diese treffen mittels Prüfgrößen selbständig und in verhältnismäßig wenigen Iterationsschritten eine Variablenauswahl.[11]

Die Funktion `stepAIC()` des R Pakets MASS[12] ist eine Implementierung eines solchen Verfahrens. `stepAIC()` erreicht die Modellreduktion durch schrittweise Minimierung der Prüfgröße AIC (Akaike Information Criterion – $AIC = -2 * maximierte\ LogLikelihood + 2 * \#Parameter$), die man vereinfacht als Maß für die Passgenauigkeit eines statistischen Modells verstehen kann. `stepAIC()` benötigt dafür ein berechnetes Eingangsmodell, das nach Möglichkeit nah am besten Ergebnismodell liegen sollte, sowie Modelldefinitionen jeweils eines maximal und minimal komplexen Ergebnismodells. Da das Ausgangsmodell mit 339 Beobachtungen und 214 Variablen in diesem Kontext als komplex gelten darf, möchte ich mich an folgende Empfehlung der Autoren halten, und dem Algorithmus stattdessen nur dieses Initialmodell zur Verfügung stellen.

> If a large model is selected as the starting point, the scope and scale arguments have generally reasonable defaults, but for a small model where the process is probably to be one of adding terms, they will usually need both to be supplied.
>
> – Venables und Ripley (2002), 175.

Zunächst erfolgt die Berechnung des Ausgangsmodells:

[7] Zum Verständnis der diagnostischen Plots: http://data.library.virginia.edu/diagnostic-plots [16.8.2017]

[8] Ich habe versucht möglichst viele Kennwerte nachzuvollziehen. Siehe dazu für die Tabellen 3 und 4 De Veaux, Velleman und Bock (2012), 785 & 792-794. und https://feliperego.github.io/blog/2015/10/23/Interpreting-Model-Output-In-R [23.8.2017] sowie für Tabelle 6 Gordon (2015) und De Veaux, Velleman und Bock (2012), 792-793.

[9] Jann (2010)

[10] De Veaux, Velleman und Bock (2012), 788-790.

[11] Backhaus u. a. (2008), 100-105.

[12] Venables und Ripley (2002), 172-177.

```
initial_model <- lm(larea ~ ., data = sel3)
```

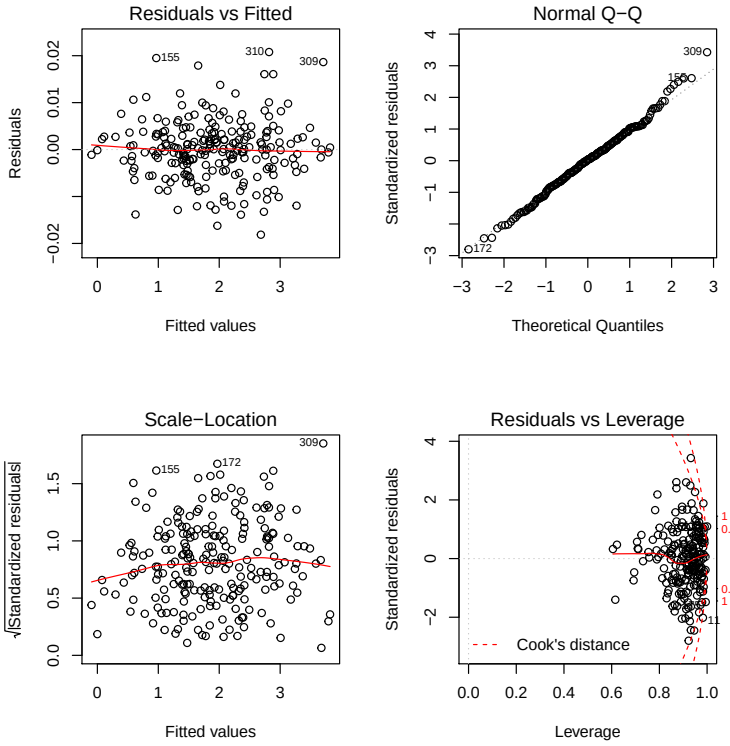

Abbildung 4: Diagnostische Plots für das Ausgangsmodell mit allen Variablen.

Abbildung 4 und Tabelle 7 zeigen, dass dieses Modell auf Grundlage aller Variablen herausragende Vorhersagefähigkeiten für die Variable *area* besitzt. Gleichermaßen entbehrt es jedoch jeder fachwissenschaftlichen Aussage, da sich eine Gesamtheit von 214 semantisch höchst unterschiedlichen Variablen jeder integrativen Interpretation entzieht. Das Modell ist nicht geeignet, ein besseres Verständnis der zugrundeliegenden naturräumlichen, kulturellen und sozioökonomischen Zusammenhänge zu generieren. Erst die Vereinfachung des Modells wird die Ableitung klarer, prüfbarer Hypothesen ermöglichen.

Entsprechend nun also ein erster Durchlauf der automatischen, schrittweisen Modellreduktion:

```
model1 <- MASS::stepAIC(
  initial_model,
  # trace option: don't show intermediate steps
  trace = FALSE
)
```

In Tabelle 5 werden die Reduktionsschritte zeilenweise dokumentiert. Mit dem Entfernen von Variablen nimmt die Anzahl der Freiheitsgrade schrittweise zu, während der AIC-Wert langsam abnimmt. 47 Variablen wurden von `stepAIC()` entfernt, dann allerdings kam der Prozess in diesem Durchlauf zum Halten. Der Vergleich des *Residuals vs Leverage* Plots in Abbildung 4 und Abbildung 5 legt nahe, dass sich die durchschnittliche Wirkung individueller Beobachtungen auf das Gesamtergebnis verringert hat. Anzunehmen ist, dass Variablen mit großer Variabilität entfernt wurden. Abbildung 3 lässt die Vermutung zu, dass sich dieser Trend bei weiterer Reduktion des Modells fortsetzen wird. 167 Variablen sind tatsächlich angesichts der deskriptiven Einfachheit des Binford-Modells hier noch kein befriedigendes Ergebnis. Die Minimierung des AIC-Werts ist scheinbar kein ausreichend starkes Optimierungskriterium:

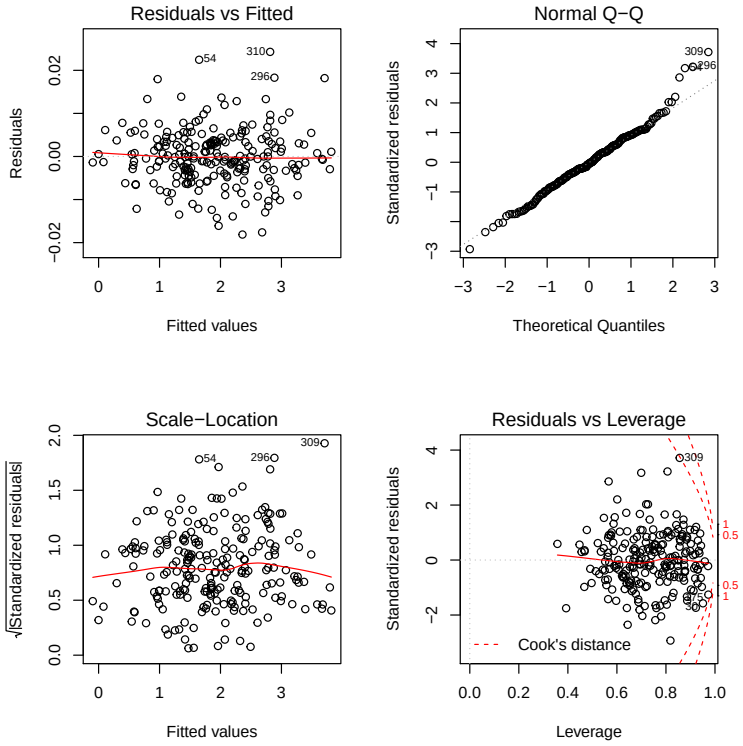

Abbildung 5: Diagnostische Plots für das Modell nach dem ersten Durchlauf von `stepAIC()`.

> This suggests, correctly, that selecting terms on the basis of AIC can be somewhat permissive in its choice of terms, being roughly equivalent to choosing an F-cutoff of 2.
>
> – Venables und Ripley (2002), 176.

Wir können nun fortfahren, indem wir entweder das k-Attribut in `stepAIC()` (*the multiple of the number of degrees of freedom used for the penalty*) erhöhen und den Prozess erneut starten, oder direkt mittels F-Statistik schrittweise jene Variablen entfernen, die nicht signifikant mit der abhängigen Variable verknüpft sind. Um diese zu identifizieren steht in MASS die Funktion `dropterm.lm()` bereit, die die Modelleinpassung für alle möglichen Modelle durchführt, die eine Variable weniger inkorporieren als das Ausgangsmodell.

```
MASS::dropterm(model1, test = "F")
```

Tabelle 6 gibt Einblick in die mit `dropterm(test = "F")` berechneten Prüfgrößen. Offensichtlich verändert sich die Modellgüte je nach dem, welche Variablen ausgeschlossen werden. Die F-Statistik gibt Antwort auf die Frage, ob es gegenüber dem Null-Modell – also dem Modell, in dem alle Koeffizienten außer dem Y-Achsenabschnitt (Intercept) auf Null gesetzt werden – einen Vorteil für die Vorhersagequalität des Ergebnismodells bringt, die jeweilige Variable mit in das Modell aufzunehmen. Große Werte für *Pr(F)* deuten also auf Variablen hin, die wahrscheinlich nicht in einer direkten Relation zur abhängigen Variablen stehen und entfernt werden können.[13]

Ich möchte das Vorgehen, mit `dropterm()` schrittweise Variablen zu entfernen, automatisieren und dann so oft wiederholt zur Anwendung bringen, bis die Anzahl der Variablen im Ergebnismodell der in Binfords Modell entspricht. Dafür habe ich einen simplen Algorithmus formuliert, der in einer Schleife die `dropterm()`-Funktion ausführt, die Variable mit dem größten p-Wert identifiziert und diese dann dann für den nächsten Schleifendurchlauf aus dem immer einfacheren Modell entfernt.

[13] Venables und Ripley (2002), 176., De Veaux, Velleman und Bock (2012), 792-793.

```
# dublicate model object
model2 <- model1

# determine number of vars to drop
to_drop <- (ncol(sel3) - nrow(model1$anova) - 10)

# drop loop
for (i in 1:to_drop) {
  # determine variable with highest
  # p-value of the F-Test
  victimvar <- MASS::dropterm(
      model2, test = "F"
    ) %>%
    tibble::as.tibble() %>%
    tibble::rownames_to_column() %>%
    dplyr::top_n(
     1, `Pr(F)`
    )

  # remove this variable from the model
  model2 <- update(
    model2,
    as.formula(paste(
      ". ~ . - ", victimvar$rowname
    ))
  )
}
```

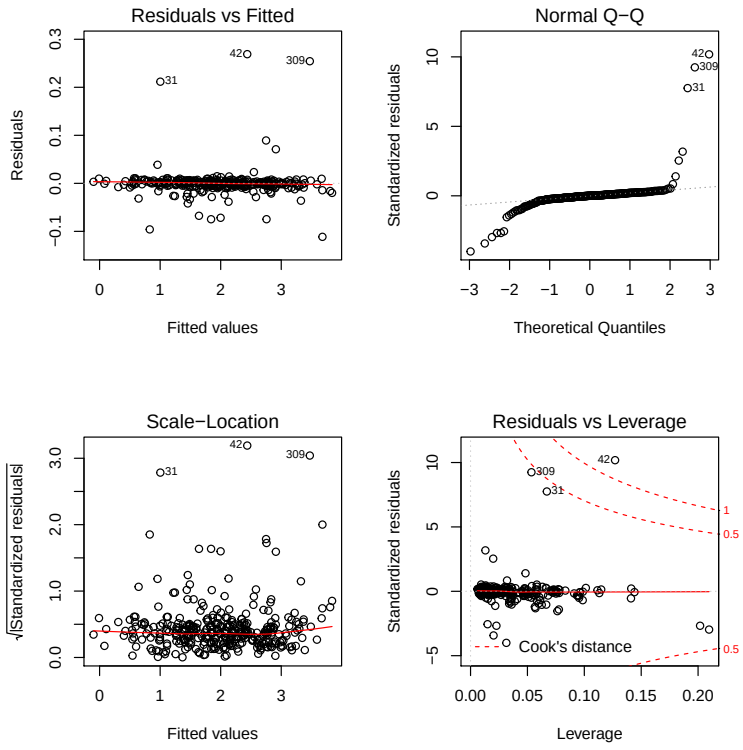

Abbildung 6: Diagnostische Plots für das Modell nach Anwendung des `dropterm()`-Algorithmus.

Das Modell, das aus diesem Algorithmus hervorgeht, erscheint im ersten Moment äußert vielversprechend. Abbildung 6 zeigt, dass es äußerst präzise Vorhersagen für *larea* erlaubt und – abgesehen von wenigen, starken Ausreißern – die Mehrzahl der Beobachtungen hervorragend erklärt. Ein Blick in Tabelle 8 offenbart jedoch, dass das Modell im wesentlichen auf den beiden Variablen *lnpop* und *lpackinx* beruht. Obgleich die Koeffizientenwerte bei der Multiplen Regression nicht so unmittelbar verstanden werden können, wie das bei der Einfachen Regression möglich ist,[14] ist doch klar, dass die anderen Variablen verschwindend wenig Einfluss auf das Ergebnis haben. *packinx* ist eine umskalierte Variante der Bevölkerungsdichte,[15] die sich selbst als abhängige Größe aus Bevölkerungszahl und Arealgröße definiert. Es ist also nicht verwunderlich, dass sich die logarithmisch skalierte Variante *lpackinx* gut zur Vorhersage von *larea* eignet. Um ein Ergebnis zu erhalten, dass mehr wissenschaftliche Relevanz besitzt, muss ich jene Variablen aus dem Ausgangsdatensatz entfernen, die direkt von der Arealgröße abhängig sind. Eine kurze Durchsicht der Schlüsseldatei LRBkey reduziert auf die oben getroffene Selektion metrisch skalierter Variablen ergibt dazu folgende Auswahl: *density*, *packinx*, *prindx*, *lden*, *lpackinx*. Freilich könnte man argumentieren, dass auch die Beziehung zwischen *larea* und der logarithmisch skalierten Bevölkerungszahl *lnpop*, die in Tabelle 8 ebenfalls deutlich als relevante Vorhersagegröße angeführt wird, trivial ist und entsprechend ausgeschlossen werden könnte. Das führt allerdings zu einer fortgeschrittenen, manuellen Variablenvorauswahl. Eine solche kann fragestellungsbezogen durchaus sinnvoll sein, wurde hier aber bewusst vermieden. Die Abwesenheit von *lpackinx* und *lnpop* in Binfords Ergebnismodell (siehe 2) spricht dafür, dass Binford hier Hand angelegt hat, ohne das explizit zu kommunizieren. Möglicherweise waren diese Variablen aber auch überhaupt nicht Teil der Arbeitsversion des Gruppendatensatzes, die ihm zum Zeitpunkt der Erstellung dieses Modells zur Verfügung stand.

Hier nun also ein Blick auf das Modell, das sich ergibt, wenn man die beschriebene Variablenvorauswahl trifft und noch einmal die Arbeitsschritte des `stepAIC()`- und `dropterm()`-Algorithmus wiederholt.

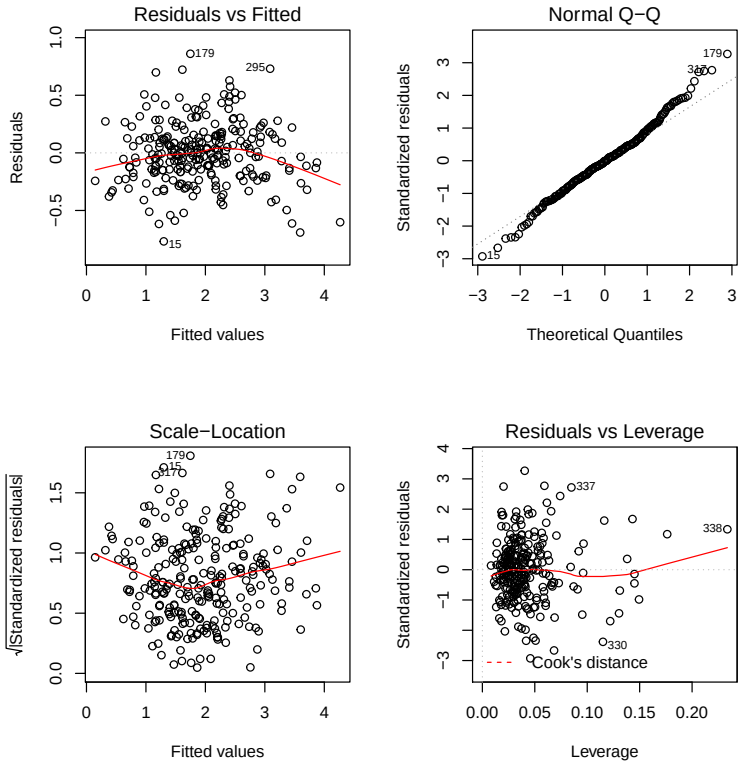

Abbildung 7: Diagnostische Plots für das finale Modell.

[14] De Veaux, Velleman und Bock (2012), 787-788 & 794.
[15] Binford (2001), 117.

$$
\begin{aligned}
\log_{10} area &= \\
larea &= \\
& 2.439+ \\
& (-0.021 * temp)+ \\
& (-0.132 * medstab)+ \\
& (0.259 * perwret)+ \\
& (-0.498 * perwltg)+ \\
& (-0.760 * lnagp)+ \\
& (0.372 * lnpop)+ \\
& (0.005 * gatherin)+ \\
& (0.001 * kmov)+ \\
& (-0.002 * nicheff)+ \\
& (-0.006 * lati)
\end{aligned}
\tag{3}
$$

Aus Tabelle 9 (und Gleichung 3) wird ersichtlich, dass dieses Modell nun auf einer wesentlich diverseren Auswahl unabhängiger Variablen fußt. Abbildung 7 und Tabelle 10 belegen, dass es sich um ein solides Modell handelt. Den Kennwerten in Tabelle 4 nach zu urteilen, scheint es wesentlich bessere Vorhersagen treffen zu können als Binfords Modell: Während das Modell, das ich auf Grundlage von Binfords Variablenauswahl ermittelt habe nur ($R^2 \approx$) 58.6% der Variabilität in *larea* erklären konnte, kommt dieses Modell auf 89.3%! Dieses Ergebnis ist eindeutig – selbst wenn man die Unzulänglichkeiten von R^2 und $adjusted R^2$ für den Modellvergleich in Betracht zieht.[16] Ich möchte damit also den Reproduktionsversuch abschließen und die Ergebnisse diskutieren.

Diskussion

Binford ist zuversichtlich, mit Gleichung 3 ein sinnvolles und relevantes Modell formuliert zu haben. Das Narrativ, das er zu seiner Erklärung konstruiert, entbehrt allerdings noch jener axiomatischen Schlüsse, die er im weiteren Verlauf des Buches generieren wird. Ich möchte seine Interpretation kurz wiedergeben (siehe dazu Abbildung 8 und Tabelle 1):

Die Variablen *hunting* und *lcoklm* sind negativ mit einer Abhängigkeit von marinen Ressourcen und positiv mit einer Abhängigkeit von landgebundenem Jagdwild verknüpft. Man kann nun schließen, dass die Nutzung terrestrischer Nahrungsquellen größere Streifgebiete für die Jäger- und Sammlergruppen erfordert. Damit wären *hunting* und *lcoklm* Anzeiger für die Arealgröße. Dieser Zusammenhang offenbart sich beispielhaft in Küstenarealen z.B. in Mexiko, Australien und der Nordamerikanischen Westküste. Dort konnten Jäger- und Sammlergemeinschaften mit kleinen Verbreitungsarealen beobachtet werden, deren Subsistenz stark von marinen Ressourcen abhängt. Zum Landesinneren hin nehmen die Arealgrößen zu.

Die primäre Biomasse, die in der Variablen *lbio5* gemessen wird, erhöht sich mit der Niederschlagsmenge. Niederschlagsüberschuss, wie er sich in *lrunoff* und *rungrc* abbildet, ist ein Indikator für ausreichende Wasserverfügbarkeit. Hohe Werte von *lbio5*, *lrunoff* und *rungrc* sind damit Anzeiger für eine Umgebung, in der Jäger- und Sammlergruppen sich aufgrund der hohen Dichte verfügbarer Biomasse aus nur kleinen Arealen versorgen können. Im diesem Kontext lässt sich auch die Negativkorrelation von *larea* mit *medstab* und *perwltg* verstehen. Wasserversorgung ist essentiell für Aufbau und stabile Verfügbarkeit von Biomasse und erlaubt damit kleinere Streifgebiete. Stabilität drückt sich auch in einer geringen Standardabweichung der Monatstemperatur *sdtemp* aus. Höhere Werte der Niederschlagsgebundenen Variablen *watrgrc* und *rlow* deuten darauf hin, dass es im Untersuchungsareal keine echte, jahreszeitliche Trockenphase gibt.

> The factors that appear correlated with small ethnic areas are the presence of marine coasts in the region, high plant biomass, and environmental stability in seasonality of temperature and rainfall variability. When these factors all have negative values indicating opposite conditions, large ethnic areas are unlikely.
>
> Binford (2001), 155.

Die von mir entwickelte Modellgleichung 3 erlaubt ebenfalls einen solchen Interpretationsversuch. Die Variablen *medstab* und *perwltg* finden auch in diesem Modell mit in Vorzeichen und Größenordnung gleichem Koeffizienten Berücksichtigung. Entsprechend lässt sich die von Binford vorgeschlagene, kausale Deutung zur Anwendung bringen. Die Variablen *temp*, *perwret* und *lnagp* passen gut in dieses Narrativ. Es liegt auf der Hand, inwiefern *lnpop* und *kmov* eine Vorhersage der Arealgröße von Jäger- und Sammlergruppen erlauben. Abbildung 9 legt nahe, dass diese beiden Variablen nicht geringen Anteil an der gegenüber dem von Binford erhöhten Güte dieses Modells haben. Die negative Korrelation mit *lati* überrascht zunächst, da auf der Nordhalbkugel insgesamt mehr nutzbare Landfläche zur Verfügung steht. Ein genauerer Blick auf den arithmetischen Mittelwert der Variable (26.23) eröffnet allerdings die Perspektive, dass ein überwiegender Teil der in die Analyse aufgenommenen Gruppen eben von der Nordhalbkugel stammt und *lati* dadurch als Indikator für Äquatornähe zu verstehen ist. Binford hat diesen Zusammenhang ebenfalls beobachtet:

> In both graphs [figure 5.14], hunter-gatherer cases occupying small ethnic areas are clustered in low latitudes that are characterized by high plant productivity.
>
> Binford (2001), 155.

Weder Binfords noch mein Modell können alle Aspekte erklären oder auch nur benennen, die die Größe des Ausbreitungsareals einer Jäger- und Sammlergruppe beeinflussen. Einige Trends, wie die Verringerung in Küstennähe, in durchsatzreichen Ökosystemen in niederen Breiten oder bei kleinen Bevölkerungszahlen lassen sich aber gut erkennen. Der Versuch, Binfords Modell zu rekonstruieren, hat also zur Bestätigung einiger Grundüberlegungen geführt. Freilich ist weitere Forschung erforderlich, um den Einfluss dieser Größen besser quantitativ zu fassen.

Wie im Rahmen der Modellbildung oben angedeutet, ist weder der von Binford noch der von mir applizierte Algorithmus zufriedenstellend. Eine Vorauswahl nach den oben eingeführten vier Kriterien (Linearität, Unabhängigkeit, Varianzäquivalenz und Normalität) wäre sicher sinnvoll, um völlig ungeeignete Variablen auszuschließen. Zumindest die Variablen im Ergebnismodell hätte ich gerne einer solchen Prüfung unterzogen, ich musste aber aus Zeitgründen darauf verzichten. Im Prozess der schrittweisen Vereinfachung des Ausgangsmodells führt die Reduktion wie ich sie vorgenommen habe (und Binford lässt nicht erkennen, dass er dieses Problem besser gelöst hätte) zu einem “Ziehen-ohne-Zurücklegen”. Diese selektive Tiefensuche im Baum der Variablenkombinationen kann den Verlust wesentlicher Einflussgrößen aus dem Analysekontext zur Konsequenz haben. `stepAIC()` ist zwar theoretisch in der Lage, ein “Ziehen-mit-Zurücklegen” durchzuführen, in meinen beiden Durchläufen hat die Funktion allerdings in keinem Schritt die Entscheidung getroffen, eine vormals entfernte Variable wieder hinzuzufügen. Möglicherweise würde sich das bei einer besseren Vorauswahl der Variablen verändern. Abbildung 10 eröffnet den Blick dafür, dass sich schon mit einfachen Korrelationsmaßen leicht potentiell vielversprechende Variablen für die Regressionsanalyse identifizieren lassen.

[16] De Veaux, Velleman und Bock (2012), 799-800.

Abschließende Gedanken

Die Ziele, die ich mir für diese Arbeit gesetzt habe, sind weitestgehend erreicht worden. Die Methode der Multiplen Regression habe ich besser verstanden und sinnvoll zur Anwendung bringen können – obgleich wie zu erwarten viele Fragen offen geblieben sind und ich gerne mehr Zeit in die Weiterentwicklung auf der erreichten Grundlage investiert hätte. Ich konnte ein Modell formulieren, das nach objektiven, statistischen Kriterien besser zur Erklärung der Variable *area* geeignet ist, als das von Binford vorgeschlagene. Ich erlaube mir jedoch kein Urteil darüber, ob Binfords Modell, in das mittels händischer Auswahl von Variablen bewusst oder unbewusst Fachwissen verarbeitet wurde, dennoch das wissenschaftlich wertvollere ist. Sicher bin ich jedoch, dass Binfords Vorgehen zur Erstellung des Modells nicht ausreichend dokumentiert wurde, um eine exakte Reproduktion zu ermöglichen. Das ist in Anbetracht der Tatsache, dass andere Teile des Buches dahingehend wesentlich besser vorbereitet sind, schade. Auch deswegen lohnt es sich, weiter mit *Constructing Frames of Reference* zu arbeiten und aus dem reichen Schatz an Hypothesen jene herauszuarbeiten, die tatsächlich als solide Grundlage für weitere Forschung dienen können.

Tabelle 1: Kurzbeschreibung der Variablen in Binfords Ergebnismodell und dem von mir erarbeiteten, finalen Modell. Die Spalte Referenz enthält die Seitenzahl in *Constructing Frames of Reference*, wo die jeweilige Variable eingeführt wird.

	Beschreibung	Einheit	Referenz
area	Größe des Areals, das von einer Jäger- und Sammlergruppe relativ exklusiv genutzt wird	100km²	117
larea	siehe area	log10(100km²)	
lbio5	Primäre (pflanzliche) Biomasse	log(kg/m^2)	85
lcoklm	Distanz zur nächstgelegenen, marinen Küste	log10(km)	154
gatherin	Ernährungsanteil pflanzlicher, terrestrischer Ressourcen	%	117
hunting	Ernährungsanteil tierischer, terrestrischer Ressourcen	%	117
kmov	Summe der Distanz, die eine durchschnittliche Familieneinheit in einem Jahr zurücklegt	km/yr	117
lati	Breitengrad auf einer Idealkugelprojektion (rectifying latitude)	°	
rlow	Niederschlagsmenge im trockensten Monat des Jahres	Millimeter	70
medstab	Indikatorgröße für die Ähnlichkeit zu mediterranem Klima berechnet aus Proxies zu Temperatur und Niederschlag	keine Einheit	72
lnagp	Net above-ground productivity – Zuwachs der Biomasse in einem Habitat durch Photosynthese und Wachstum	g/m^2/yr	79
nicheff	Niche effectiveness – Verhältnis der tatsächlichen Bevölkerungsdichte zu einer modellbasierten Vorhersage der Bevölkerungsdichte, die Binford in Kapitel 10 formuliert	keine Einheit	373
lnpop	Größe der untersuchten Bevölkerung	log(Anzahl)	117
perwret	Anteil des Wachstumszeitraum in dem der Boden Wasser gespeichert vorhält	%	79
perwltg	Anteil des Wachstumszeitraum in dem die Wasserverfügbarkeit unter dem pflanzlichen Welkepunkt liegt	%	79
rungrc	Anzahl der Monate im Wachstumszeitraum in denen der RUNOFF-Wert > 0	Anzahl	79
lrunoff	Wasser, das durch Abfluss für die Nutzung durch Pflanzen verloren geht	log(mm)	79
sdtemp	Standardabweichung der mittleren Monatstemperatur	keine Einheit	70
temp	Temperateness – Indikatorgröße für die Ausgeglichenheit der Monatstemperatur	keine Einheit	59
watrgrc	Anzahl der Monate im Wachstumszeitraum, in denen Wasser im Boden gespeichert bleibt	Anzahl	79

Literatur

Ames, Kenneth M. 2004. Supposing Hunter-Gatherer Variability. *American Antiquity* 69, Nr. 2: 364–374.

Backhaus, Klaus, Bernd Erichson, Wulff Plinke und Rolf Weiber, Hrsg. 2008. *Multivariate Analysemethoden: eine anwendungsorientierte Einführung*. 12., vollst. überarb. Aufl. Springer-Lehrbuch. Berlin: Springer.

Binford, L. R. 2001. *Constructing Frames of Reference: An Analytical Method for Archaeological Theory Building Using Hunter-Gatherer and Environmental Data Sets*. Berkeley/Los Angeles.

Browman, D. L. 2005. Constructing Frames of Reference: An Analytical Method for Archaeological Theory Building Using Hunter-Gatherer and Environmental Data Sets. *American Anthropologist* 107, Nr. 2: 277–279.

De Veaux, R. D., P. F. Velleman und D. Bock. 2012. *Stats: Data and models*. 3. Aufl. Upper Saddle River, NJ: Addison-Wesley.

Donald Pate, F. 2005. Review of „Constructing Frames of Reference: An Analytical Method for Archaeological Theory Building Using Ethnographic and Environmental Data Sets“ by Lewis R. Binford. *Australian Archaeology* 60: 82–83.

Gordon, R.A. 2015. *Regression Analysis for the Social Sciences*. Taylor & Francis.

Hill, K. 2002. Constructing Frames of Reference: An Analytical Method for Archeological Theory Building Using Ethnographic and Environmental Data Sets. Lewis R. Binford. *Journal of Anthropological Research* 58, Nr. 3: 416–419. doi:10.1086/jar.58.3.3631188,.

Jann, B. 2010. Robuste Regression. In: *Handbuch Der Sozialwissenschaftlichen Datenanalyse*, hg. von C. Wolf und H. Best, 707–740. Wiesbaden: VS Verlag für Sozialwissenschaften.

Marwick, B., A. Johnson, D. White und E. A. Eff. 2016. *Binford: Binford's Hunter-Gatherer Data*. http://github.com/benmarwick/binford.

Nakoinz, Oliver und Daniel Knitter. 2016. *Modelling Human Behaviour in Landscapes: Basic Concepts and Modelling Elements*.

Venables, W. N. und B. D. Ripley. 2002. *Modern Applied Statistics with S*. 4. Aufl. New York: Springer.

Tabelle 2: Auszug aus der Metatabelle zum Gruppendatensatz. Skalenniveauklassifizierung in den Spalten *type* und *type_exp*.

variable	description	type	type_exp
...	...	...	...
minlaw	The presence or absence of behavioral mother-in-law avoidance and other restrictions on behavior. Binary variable.	ordinal	nominal
male.mm	Male height in millimeters; (Table: 6.03); (Binford 2001:183)	ordinal	ratio
female.mm	Female height in millimeters; (Table: 6.03); (Binford 2001:183)	ordinal	ratio
male.kg	Male weight in kilograms; (Table: 6.03); (Binford 2001:183)	ordinal	ratio
female.kg	Female weight in kilograms; (Table: 6.03); (Binford 2001:183)	ordinal	ratio
termhnt	Proportion of food in a-cultural terrestrial model from hunting	ordinal	interval
termgath	Proportion of food in a-cultural terrestrial model from gathering	ordinal	interval
termh2	Terrestrial model hunting density: Number of persons per 100 sqkm who could be supported from ungulate resources alone; (Equation: 6.13); (Binford 2001:187)	ordinal	ratio
termg2	Terrestrial model gathering density: Number of persons per 100 sqkm who could be supported from terrestrial plant foods alone; (Equation: 6.14); (Binford 2001:187)	ordinal	ratio
termd2	Terrestrial model population density: Population density (adjusted for body size) expected at a particular location, based on terrestrial model (persons per 100 sqkm); (Equation: 6.15); (Binford 2001:187)	ordinal	ratio
subspx	Dominant food source predicted by Binford's Terrestrial Model; (Binford 2001:203)	categorical	nominal
nicheffg	Ratio of niche effectiveness (measured density compared to terrestrial model density) in exploitation of terrestrial plant resources; (Equation: 10.02); (Binford 2001:373)	ordinal	ratio
nicheffh	Ratio of niche effectiveness (measured density compared to terrestrial model density) in exploitation of terrestrial animal resources; (Equation: 10.03); (Binford 2001:373)	ordinal	ratio
...	...	...	...

Tabelle 3: Ergebniszusammenfassung des mit Binfords Variablenauswahl reproduzierten Modells für die einzelnen Koeffizienten. Alle Werte sind auf drei Nachkommastellen gerundet

estimate – Koeffizient: Koeffizienten(schätzung) des Ergebnismodells (Intercept und Variablenslopes).

std.error – Standardfehler: Maß für die Präzision der Schätzung für den Wert des Koeffizienten. Bei einer Variable, die gut für die Vorhersage der abhängigen Variable geeignet ist, sollte der Standardfehler in Relation zum Koeffizienten klein sein.

statistic – t-Wert: Anzahl der Standardabweichungen, die den Koeffizienten von Null trennt. Der Betrag des Wertes sollte (auch in Relation zum Standardfehler) groß sein um die Nullhypothese 'keine Relation der Variablen' verwerfen zu können.

p.value – p-Wert: p-Wert der t-Statistik. Wahrscheinlichkeit, dass eine Beobachtung auftritt, die gleich oder größer als der t-Wert ist. Ein kleiner p-Wert zeigt an, dass die Wahrscheinlichkeit einer zufälligen Enstehung dieses Modellergebnisses gering ist. Die Relation zwischen abhängiger und unabhängiger Variable ist damit signifikant.

	estimate	std.error	statistic	p.value
(Intercept)	3.024	0.343	8.804	0.000
hunting	0.007	0.002	3.355	0.001
lbio5	-0.337	0.086	-3.899	0.000
lcoklm	0.144	0.050	2.889	0.004
lrunoff	-0.074	0.030	-2.476	0.014
watrgrc	0.054	0.041	1.312	0.190
medstab	-0.128	0.038	-3.329	0.001
perwltg	-0.710	0.146	-4.868	0.000
rlow	0.004	0.001	3.736	0.000
rungrc	-0.103	0.044	-2.338	0.020
sdtemp	0.032	0.012	2.720	0.007

Tabelle 4: Ergebniszusammenfassung des mit Binfords Variablenauswahl reproduzierten Modells für das Gesamtmodell. Alle Werte sind auf drei Nachkommastellen gerundet.

r.squared – R^2 – Bestimmtheitsmaß: R^2 ist ein Maß für die Güte der Modelleinpassung. Es gibt den Anteil der Varianz der abhängigen Variablen wieder, der durch das lineare Modell erklärt wird. Der Wert liegt zwischen 0 (kein linearer Zusammenhang) und 1 (perfekter linearer Zusammenhang). Ein hoher Wert ist ein Indikator für ein gutes Modell.

adj.r.squared – korrigiertes Bestimmtheitsmaß: R^2 korrigiert unter Beachtung der Anzahl unabhängiger Variablen. R^2 wird bei zunehmender Anzahl an Variablen größer und muss entsprechend bei Multipler Regression normiert werden.

sigma – Residual Standard Error: Durchschnittliche Abweichung der Modellvorhersage für die abhängige Variable von den tatsächlich gemessenen Werten. Bei einem guten Modell sollte der Wert gering sein.

statistic und p.value Siehe Tabelle 3. Hier beziehen sich die Werte auf das Gesamtmodell.

logLik – LogLikelihood, AIC – Akaike Information Criterion und BIC – Bayesian Information Criterion Maße für die Güte der Modelleinpassung.

deviance Maß für die Distance zweier Modelle. Hier als Maß für die Güte der Modelleinpassung indem das Ergebnismodell mit dem Null-Modell (s.u.) verglichen wird.

df.residual – Anzahl der Freiheitsgrade Berechnet sich aus der Anzahl der Beobachtungen abzüglich der Anzahl der schätzbaren Koeffizienten. Insofern handelt es sich um die Anzahl der 'überflüssigen' Messwerte, die zur Berechnung der Modellparameter nicht erforderlich wären.

	value		value
r.squared	0.586	**logLik**	-247.781
adj.r.squared	0.574	**AIC**	519.563
sigma	0.511	**BIC**	565.475
statistic	46.47	**deviance**	85.624
p.value	0	**df.residual**	328

Tabelle 5: ANOVA Komponente des Ergebnismodelldatentyps von `stepAIC()`. Zeigt die schrittweise Entfernung von Variablen zur Reduktion der Modellkomplexität und zugehörige, diagnostische Prüfgrößen.
Step – Reduktionsschritt: Variable, die in diesem Schritt entfernt oder wieder hinzugefügt wurde.
Dev – Deviance, Resid. Df – Anzahl der Freiheitsgrade, AIC – Akaike information criterion: Siehe Tabelle 4.
Resid. Dev – residual deviance: Eine Konstante abzüglich zwei mal die maximierte Log-Likelihood. Ist nur bei gesättigten Modellen aussagekräftig und kann hier ignoriert werden.

Step	Dev	Resid. Df	Resid. Dev	AIC
	NA	21	0.0091203	-1894.864
- t_usda_tex_class	0.00e+00	21	0.0091203	-1894.864
- trange	0.00e+00	21	0.0091203	-1894.864
- lbar5	0.00e+00	21	0.0091203	-1894.864
- elev	0.00e+00	21	0.0091203	-1894.864
- termd2	0.00e+00	21	0.0091203	-1894.864
- watd	0.00e+00	21	0.0091203	-1894.864
- rrcorr2	0.00e+00	21	0.0091203	-1894.864
- longitude	2.00e-07	22	0.0091205	-1896.859
- nicheff	5.00e-07	23	0.0091210	-1898.847
- mtemp	5.00e-07	24	0.0091215	-1900.835
- l25	6.00e-07	25	0.0091221	-1902.820
- meanalt	7.00e-07	26	0.0091228	-1904.803
- bio.11	2.00e-06	27	0.0091247	-1906.754
- bio.14	2.40e-06	28	0.0091271	-1908.694
- bio.19.sd	1.90e-06	29	0.0091290	-1910.647
- nomov	2.70e-06	30	0.0091317	-1912.579
- cmat	2.10e-06	31	0.0091338	-1914.528
- t_bs	2.70e-06	32	0.0091365	-1916.460
- lptoae	4.70e-06	33	0.0091412	-1918.342
- sdalt	6.30e-06	34	0.0091475	-1920.186
- fishing	7.70e-06	35	0.0091552	-1921.994
- lsnowac	9.30e-06	36	0.0091645	-1923.763
- siltyclayloam	9.70e-06	37	0.0091742	-1925.521
- defper	1.18e-05	38	0.0091860	-1927.228
- anntotprecip	1.25e-05	39	0.0091985	-1928.918
- lden	1.58e-05	40	0.0092143	-1930.527
- sucstab2	1.37e-05	41	0.0092279	-1932.189
- lwaccess	1.21e-05	42	0.0092401	-1933.890
- evmmod2a	1.45e-05	43	0.0092546	-1935.531
- bio.9.sd	1.33e-05	44	0.0092679	-1937.204
- bio.16	1.69e-05	45	0.0092848	-1938.789
- lgather	1.96e-05	46	0.0093044	-1940.307
- bio.15	1.87e-05	47	0.0093231	-1941.849
- snowdepth	1.76e-05	48	0.0093407	-1943.420
- nicheffg	2.99e-05	49	0.0093706	-1944.693
- lammod3a	2.09e-05	50	0.0093915	-1946.184
- bio.6	5.85e-05	51	0.0094500	-1946.769
- bio.7	9.30e-06	52	0.0094593	-1948.544
- lsstab2	5.17e-05	53	0.0095110	-1949.301
- bio.12.sd	4.99e-05	54	0.0095609	-1950.108
- bio.15.sd	4.55e-05	55	0.0096063	-1951.026
- watret	5.19e-05	56	0.0096582	-1951.798
- alt.sd	8.31e-05	57	0.0097414	-1951.844
- lwatrgrc	8.40e-05	58	0.0098253	-1951.888
- bar5	7.89e-05	59	0.0099042	-1952.065
- s_ph_h2o	8.03e-05	60	0.0099844	-1952.225

Tabelle 6: dropterm Tabelle: Maße für die Qualität des Beitrags einzelner Variablen zum Gesamtmodell. Alle Werte sind auf vier Nachkommastellen gerundet, der AIC-Wert auf ganze Zahlen.
Sum of Sq – ESS – explained sum of squares: Summe der Abweichungsquadrate der Modellvorhersage zum arithmetischen Mittel der abhängigen Variable für das Modell ohne diese Variable.
RSS – residual sum of squares: Wie ESS, hier aber Summe der Abweichungsquadrate der Modellvorhersage zu *allen* Werten der abhängigen Variable.
AIC – Akaike information criterion: Siehe Tabelle 4.
F Value: Eingangswert des F-Tests.
Pr(F) – probability of F-Value: Maß für die Wahrscheinlichkeit (p-Wert) des F-Werts im F-Test der Gesamtsignifikanz. Der F-Test basiert auf dem Vergleich mit einem Modell, in dem alle Koeffizienten außer dem Intercept und dem der aktuellen Variable den Wert Null annehmen. Ist die Wahrscheinlichkeit klein, kann die Nullhypothese, dass kein wirklicher Zusammenhang zwischen abhängiger und unabhängigen Variablen besteht, verworfen werden.

rowname	Sum of Sq	RSS	AIC	F Value	Pr(F)
<none>	NA	0.01	-1952	NA	NA
latitude	0.0015	0.0115	-1922	9.2562	0.0035
dposit	7e-04	0.0107	-1939	4.0738	0.048
headwat	7e-04	0.0107	-1939	4.0505	0.0487
drain	0.0016	0.0116	-1921	9.5223	0.0031
h10	1e-04	0.0101	-1952	0.6275	0.4314
h25	5e-04	0.0105	-1943	3.0034	0.0882
h50	7e-04	0.0106	-1940	3.9323	0.052
l10	1e-04	0.0101	-1951	0.8298	0.366
l50	0.0022	0.0122	-1908	13.5196	5e-04
maxrange	1e-04	0.0101	-1952	0.7087	0.4032
et	3e-04	0.0103	-1947	1.8911	0.1742
mcm	0.0025	0.0125	-1903	15.0788	3e-04
mwm	6e-04	0.0106	-1942	3.4072	0.0698
temp	0.0011	0.0111	-1930	6.6874	0.0122
crr	0.008	0.018	-1820	48.0953	0
rhigh	6e-04	0.0106	-1941	3.5067	0.066
rlow	0.0015	0.0115	-1922	9.1231	0.0037
reven	0.002	0.012	-1913	11.8344	0.0011
mrain	0.004	0.014	-1877	24.1767	0
sdtemp	0.0021	0.0121	-1911	12.5376	8e-04
sdrain	9e-04	0.0109	-1934	5.5938	0.0213
rrcorr	0.002	0.012	-1913	11.8524	0.0011
rrcorr3	4e-04	0.0104	-1946	2.2447	0.1393
medstab	6e-04	0.0106	-1941	3.67	0.0602
growc	0.0017	0.0117	-1919	10.088	0.0024
pet	3e-04	0.0103	-1947	1.8189	0.1825
ae	0.0018	0.0118	-1917	10.6196	0.0018
snowac	0.0067	0.0167	-1837	40.1892	0
ptoae	0.0021	0.0121	-1911	12.4525	8e-04
hirx	0.0035	0.0134	-1887	20.7499	0
ptowatd	1e-04	0.0101	-1951	0.8299	0.366
watdgrc	8e-04	0.0108	-1937	4.7941	0.0325
perwret	3e-04	0.0103	-1948	1.5977	0.2111
perwltg	0.0036	0.0136	-1884	21.6046	0
wltgrc	0.0017	0.0117	-1917	10.4888	0.002
nagp	0.0011	0.0111	-1930	6.8501	0.0112
lnagp	0.0032	0.0132	-1891	19.1916	0
bio5	0.0017	0.0116	-1919	10.0043	0.0025
lbio5	0.0077	0.0177	-1824	46.2111	0
...	...	...	...	...	...

Tabelle 7: Ergebniszusammenfassung des initialen Modells. Alle Werte sind auf drei Nachkommastellen gerundet.

	value		value
r.squared	1	**logLik**	830.914
adj.r.squared	0.999	**AIC**	-1245.828
sigma	0.021	**BIC**	-532.524
statistic	1680.567	**deviance**	0.009
p.value	0	**df.residual**	21

Tabelle 8: Ergebniszusammenfassung für die einzelnen Koeffizienten des mittels `stepAIC()` und `dropterm()` aus der Gesamtvariablenmenge reduzierten Modells. Alle Werte sind auf drei Nachkommastellen gerundet.

term	estimate	std.error	statistic	p.value
(Intercept)	-0.969	0.012	-82.991	0
lnpop	0.436	0.001	329.518	0
lpackinx	-1.002	0.002	-404.206	0
claylight	-0.001	0.000	-3.987	0
siltloam	-0.001	0.000	-4.367	0
loam	-0.001	0.000	-5.570	0
sandyclayloam	-0.001	0.000	-3.917	0
sandyloam	-0.001	0.000	-5.366	0
loamysand	-0.001	0.000	-3.953	0
sand	-0.001	0.000	-4.972	0
t_ref_bulk_density	0.053	0.011	4.824	0

Tabelle 9: Ergebniszusammenfassung für die einzelnen Koeffizienten des finalen Modells, das durch erneute Anwendung des `stepAIC()`- und `dropterm()`-Algorithmus auf ein Modell mit leicht reduzierter Variablenauswahl ermittelt wurde. Alle Werte sind auf drei Nachkommastellen gerundet.

term	estimate	std.error	statistic	p.value
(Intercept)	2.439	0.250	9.773	0.000
temp	-0.021	0.002	-9.685	0.000
medstab	-0.132	0.022	-5.909	0.000
perwret	0.259	0.081	3.207	0.002
perwltg	-0.498	0.084	-5.961	0.000
lnagp	-0.760	0.065	-11.745	0.000
lnpop	0.372	0.014	26.068	0.000
gatherin	0.005	0.001	4.385	0.000
kmov	0.001	0.000	12.188	0.000
nicheff	-0.002	0.000	-6.185	0.000
lati	-0.006	0.001	-6.360	0.000

Tabelle 10: Ergebniszusammenfassung des finalen Modells. Alle Werte sind auf drei Nachkommastellen gerundet.

	value		value
r.squared	0.893	**logLik**	-21.666
adj.r.squared	0.889	**AIC**	67.332
sigma	0.269	**BIC**	110.107
statistic	209.03	**deviance**	18.041
p.value	0	**df.residual**	250

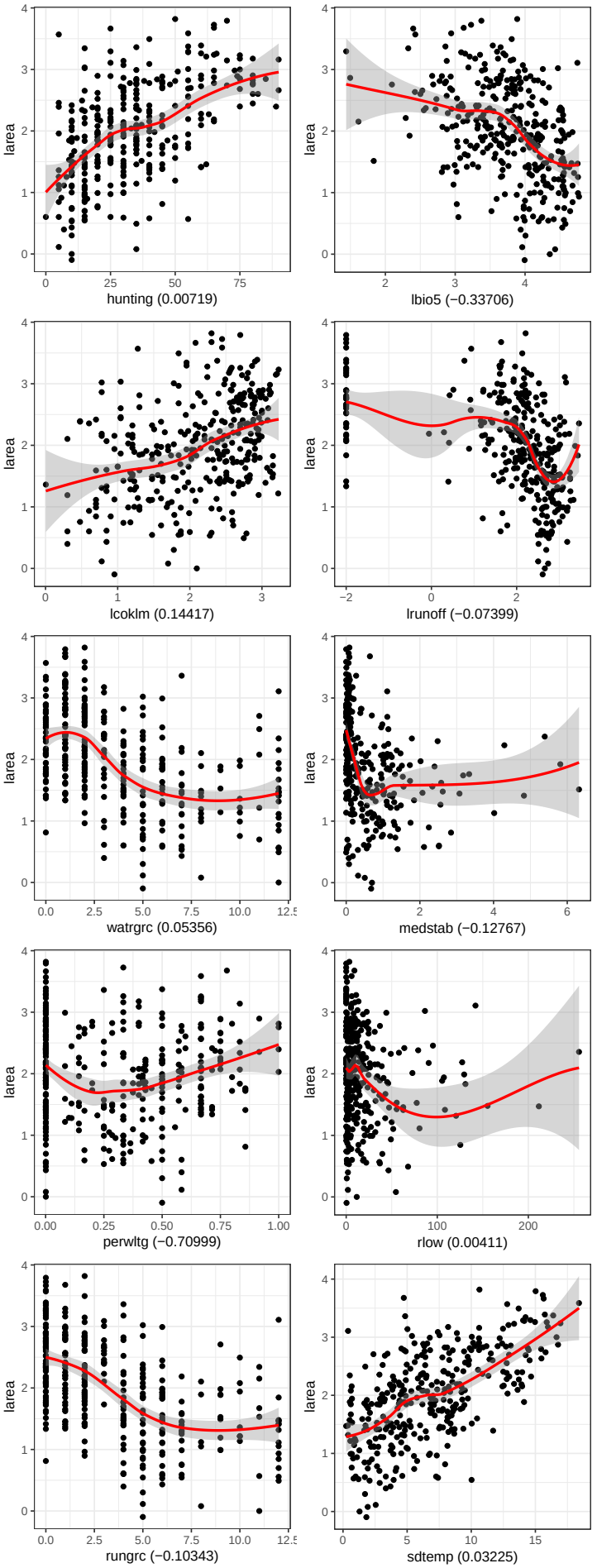

Abbildung 8: Bivariate Relationen der abhängigen Variable *larea* und allen unabhängigen Variablen in Binfords Ergebnismodell. Mit angegeben ist der Koeffizient der jeweiligen Variable auf fünf Nachkommastellen gerundet.

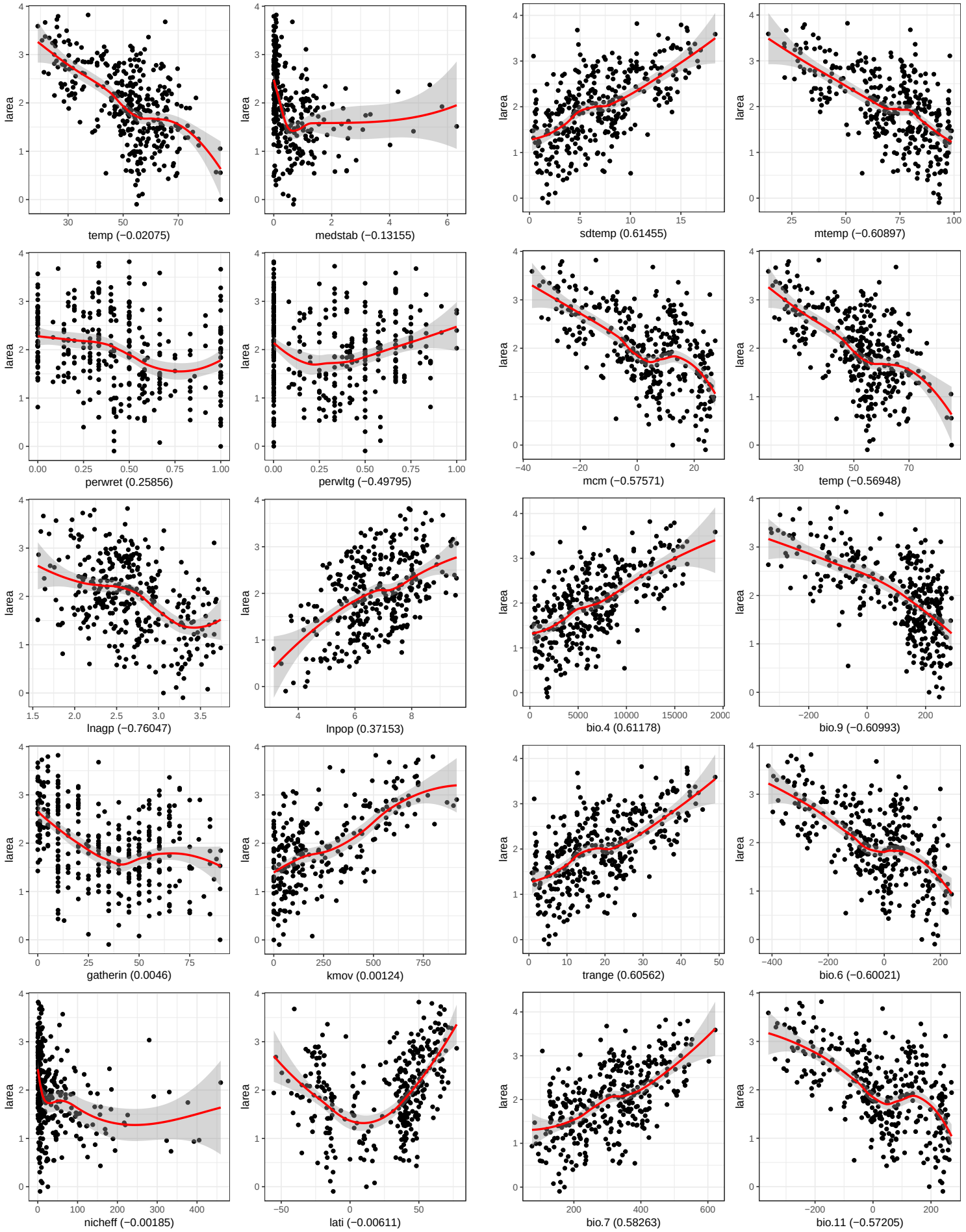

Abbildung 9: Bivariate Relationen der abhängigen Variable *larea* und allen unabhängigen Variablen in meinem finalen Ergebnismodell. Mit angegeben ist der Koeffizient der jeweiligen Variable auf fünf Nachkommastellen gerundet.

Abbildung 10: Bivariate Relationen der Variable *larea* und den 10 Variablen, die mit *larea* den höchsten Korrelationskoeffizienten nach Pearson teilen. Mit angegeben ist der Koeffizient auf fünf Nachkommastellen gerundet.