Berufsverläufe von Juristen

Statistisches Consulting

Hannah Busen (hannah.busen@googlemail.com)
Eva-Maria Müntefering (evamariam1102@gmx.de)

Projektpartner: Dr. Maike Reimer, Christina Müller
Bayerisches Staatsinstitut für Hochschulforschung und Hochschulplanung
Betreuer: Prof. Dr. Helmut Küchenhoff, Institut für Statistik

Institut für Statistik

Ludwig - Maximilians - Universität München

München, 16.03.2016

Zusammenfassung

Der vorliegende Bericht befasst sich mit der Frage, ob es Unterschiede zwischen den verschiedenen Berufsverläufen von Juristen gibt und welche Faktoren diese Unterschiede beeinflussen. Dafür wurden zunächst die Berufsverläufe der Rechtswissenschaftsabsolventen der Jahrgänge 2005 und 2006 in Bayern mit Hilfe der *Sequenzmusteranalyse* deskriptiv analysiert. Hierbei war vor allem auffällig, dass die Verläufe klar strukturiert sind und es nach dem Referendariat nur wenige Zustandswechsel gibt. Danach folgte eine *Clusteranalyse*, in deren Rahmen das *Optimal Matching* und der Algorithmus nach *Ward* angewandt wurden. Hierbei wurde ersichtlich, dass die Absolventen anhand ihrer Berufsverläufe zu sieben verschiedenen Gruppen zusammengefügt werden können. Abschließende, *multinomiale Regressionen* sowohl zur *Clusterzugehörigkeit* als auch zur *Art der Anstellung* ergaben, dass ein Großteil der einbezogenen Kovariablen keinen signifikanten Einfluss auf die Zielgröße hatten. Einzig die *Abschlussnote*, das *Geschlecht*, die *Art der Stellenfindung* und die Tatsache, dass bereits die Eltern der Absolventen ein Studium der Rechtswissenschaften abgeschlossen haben, scheinen die *Clusterzugehörigkeit* zu beeinflussen. Die *Art der Anstellung* blieb davon unberührt. Die Höhe des *Einkommens* betreffend ergab die durchgeführte *lineare Regression*, dass sich dieses in Abhängigkeit von der *Note*, der Anzahl der *Studienkontakte* und der *Clusterzugehörigkeit*, also dem eingeschlagenen Berufsweg, unterscheidet.

Inhaltsverzeichnis

1 Einführung

Der Arbeitsmarkt für Jura-Absolventen wird immer enger und der Leistungsdruck dadurch immer größer. Was passiert mit denjenigen Studenten, die ihr Studium erfolgreich oder auch weniger erfolgreich beendet haben? Welchen beruflichen Werdegang schlagen sie ein? Gibt es Muster innerhalb der Berufsverläufe? Und unterscheiden sich diese aufgrund von Faktoren wie der Abschlussnote, den sozialen, vor allem familiären Netzwerken der Absolventen? Macht es, den beruflichen Erfolg betreffend, tatsächlich einen Unterschied, ob etwa die Eltern bereits ein rechtswissenschaftliches Studium abgeschlossen haben? Und von welchen weiteren möglichen Faktoren, wie zum Beispiel dem Geschlecht oder einer Promotion, hängen Einkommen und die Art der Anstellung ab?

Mithilfe statistischer Methodik sollen diese und weitere Fragestellungen beantwortet werden. Dazu wurden die Absolventen der Rechtswissenschaften der Jahrgänge 2005/2006 in Bayern sowohl anderthalb als auch fünf Jahre nach ihrem Abschluss befragt. Um die Berufsverläufe genauer zu betrachten und eventuelle Muster und sowohl Gemeinsamkeiten als auch Unterschiede innerhalb dieser aufzudecken, kann man sich der *Sequenzmusteranalyse* bedienen. Diese, vor allem in den Sozialwissenschaften angewandte, Analysemethode eignet sich als Werkzeug für entsprechende Daten longitudinaler Struktur. Mit ihrer Hilfe lässt sich die Abfolge der einzelnen Tätigkeiten der Absolventen deskriptiv darstellen. In Kombination mit einer Clusteranalyse besteht die Möglichkeit, diese Verläufe für homogene Gruppen zu betrachten und somit weitere interessante Abhängigkeiten aufzudecken. (Stegmann et al. (2013))

Weiter soll mithilfe von Regressionsmodellen der Einfluss bestimmter Faktoren wie zum Beispiel dem Geschlecht, vorhandene soziale/familiäre Netzwerke, der Abschlussnote, dem eingeschlagenen Berufsweg, ob Kinder mit im Haushalt leben und der Promotionsstatus auf das Einkommen und die Art der Anstellung untersucht werden.

Ebenso soll mit Hilfe einer multinomialen Regression der Einfluss der zum Teil bereits zuvor genannten Faktoren auf den späteren Berufsweg quantifiziert werden.

Zunächst werden in Kapitel 2 die vorliegenden Daten genauer beschrieben, ihr Hintergrund erklärt und auf eventuelle Probleme dieser eingegangen. Danach folgt in Kapitel 3 eine Einführung in die statistische Methodik, welche der Analyse zugrunde liegt.

Hierzu zählen die *Sequenzmusteranalyse*, das *Optimal Matching*, die *Clusteranalyse*, die Verwendung *repräsentativer Sequenzen* und *lineare/multinomiale Regression*. Die mit diesen Methoden erhaltenen Ergebnisse werden in Kapitel 4 präsentiert und erklärt, bevor es in Kapitel 5 eine abschließende Zusammenfassung gibt, auf Probleme der Analyse hingewiesen wird und weitere ergänzende Möglichkeiten der Analyse in Ausblick gestellt werden.

2 Die Daten

In diesem Abschnitt werden die Daten genauer beschrieben, auf welche sich die Analysen stützen. Dazu folgt zunächst in Abschnitt 2.1 eine Beschreibung der Ausgangsdaten, gefolgt von einer Beschreibung der Datenbearbeitung zur Vorbereitung auf die Analysen in Abschnitt 2.2. In Abschnitt 2.3 verschaffen erste deskriptive Analysen einen ersten Überblick über die Datenbeschaffenheit.

2.1 Datenbeschreibung

Wie in der Einleitung bereits kurz beschrieben, umfassen die Daten die Ergebnisse der Befragungen der Absolventen des Abschlussjahrgangs 2005/2006. Hierzu zählen sowohl die Absolventen des Wintersemester 2005/2006 als auch die des Sommersemesters 2006. Für die folgenden Analysen ist jedoch nur der Subdatensatz der Juraabsolventen interessant. Die Absolventen wurden das erste Mal im Jahr 2007 und somit etwa ein Jahr nach ihrem Abschluss befragt. Eine zweite Befragung fand sieben Jahre nach dem Abschluss statt. Die gesamte Befragung wurde retrospektiv durchgeführt, das heißt bei der Beantwortung des zweiten Fragebogens sollten die Absolventen ihre jeweiligen Tätigkeiten für jeden Monat der vergangenen sieben Jahre seit ihrem Abschluss angeben. Neben personenbezogenen Daten wie dem Geschlecht, dem Alter, dem Familienstand und der Anzahl der Kinder, sind im weiteren Verlauf der Analyse vor allem die Antworten auf die Fragen nach dem monatlichen Einkommen, der Abschlussnote und der Art der Anstellung von Interesse. Ebenso ist interessant, ob die Absolventen promoviert oder bereits andere Familienmitglieder ein Jurastudium absolviert haben.

Insgesamt stehen fünf Ausgangsdatensätze zur Verfügung. Diese beinhalten jeweils folgende Daten:

1. *het12*: Datensatz longitudinaler Struktur, welche alle Beobachtungen aller Absolventen aller Variablen zum Zeitpunkt der ersten und letzten Beschäftigung enthält. Es gibt insgesamt 226.083 Beobachtungen zu 351 Variablen.

2. *05061Quer*: Querschnittsdatensatz aller Absolventen zum Zeitpunkt der ersten Befragung. Es gibt 6819 Beobachtungen zu insgesamt 304 Variablen.

3. *quer8final*: Querschnittsdatensatz aller Absolventen zum Zeitpunkt der 2. Befragung. Es gibt 3482 Beobachtungen zu insgesamt 319 Variablen.

4. *jurazustand*: Datensatz longitudinaler Struktur, welcher als Subdatensatz aus 1.) entstanden ist und nur Juraabsolventen enthält. Zusätzlich werden die Beschäftigungen monatsweise aufgeführt. Während im *het12*-Datensatz noch 242 Juraabsolventen aufgelistet sind, sind es in diesem nur noch 211. Dies liegt daran, dass bei der zweiten Befragung nach sieben Jahren nicht mehr alle Studenten, die bei der ersten Befragung dabei waren, mitgemacht haben.

5. *juraquer03*: Querschnittsdatensatz, welcher als Subdatensatz aus 2.) entstanden ist. Er enthält die Antworten aller Juraabsolventen, zusätzlich ist der aktuelle Zustand (zum Zeitpunkt der 2. Befragung) angegeben.

2.2 Datenbearbeitung

Bevor mit der Analyse begonnen werden kann, müssen zunächst die Daten auf diese vorbereitet werden. Letztendlich ist der wirklich interessierende Datensatz *jurazustand*. Dieser enthält sowohl alle Beschäftigungen aller Juraabsolvten monatsweise als auch, bis auf die Information über das Geschlecht, alle zusätzlichen Variablen, die von Interesse sind. Die Variable *Geschlecht* wird somit noch dem Datensatz *jurazustand* hinzugefügt. Alle zuvor beschriebenen Datensätze enthalten eine Vielzahl an Variablen, welche im Folgenden nicht mehr von Interesse sein werden. Die Variable *Promotion* besitzt die fünf Kategorien „Ja, ich habe meine Promotion beendet“, „Nein“, „Ja, noch dabei“, „Ja, habe aber zur Zeit unterbrochen“ und „Ja, habe aber abgebrochen“. Diese werden, je nachdem, ob die Promotion bereits abgeschlossen wurde, zu den zwei Kategorien „Ja“ und „Nein“ zusammengefasst. Auch die Variable *Kinder* wurde auf die zwei Kategorien „Ja“ und „Nein“ reduziert. Ebenso wird für die Monate, in denen nur einige Befragten eine Beschäftigung angegeben haben, ein *NA* bei denjenigen eingefügt, die hier keine Angabe gemacht haben. Hierdurch werden die Sequenzen auf dieselbe Länge gebracht, was dazu dient, den Datensatz später aus dem longitudinalen Format ins Querschnittsformat zu überführen. Weiter gibt es einen Absolventen, dessen Angaben zeitlich betrachtet weiter in die Gegenwart reichen als bei allen übrigen. Da es sich um einen Einzelfall handelt, werden die „überstehenden“Zustände abgeschnitten. Auf die Analyse hat dies keinen

Einfluss. Aus den Variablen, welche angeben, ob der Vater oder die Mutter ein Jurastudium absolviert hat, wird eine Variable erzeugt, welche diese Information für beide Elternteile zusammenfasst. Die so entstehenden Variablen *juEltern* gibt demnach Auskunft darüber, ob mindestens ein Elternteil Jura studiert hat. Für die Anzahl der Kontakte (Studium, Privat, Beruf) gilt, dass hier alle Werte > 100 in *NA's* umgewandelt wurden, da diese sehr unrealistisch erscheinen. Abschließend wird die Variable *stelfindung* erzeugt. Hierzu werden die Absolventen anhand der Art, wie sie ihre Stelle gefunden haben, in folgene Kategorien eingeteil:

1. Studiennahe Kontakte
2. Referendariat
3. Persönliche Kontakte
4. formelle Stellenfindung
5. Sonstiges/Selbstständig

Falls mehrere Wege angegeben wurden, so gilt, dass die nach dieser Auflistung ranghöhere Kategorie gewählt wird. Hat ein Absolvent beispielsweise angegeben, dass er seine Stelle sowohl über persönliche als auch studiennahe Kontakte gefunden hat, so wird er der Kategorie *Studiennahe Kontakte* zugeordnet.

2.3 Deskription interessierender Variablen

Dieses Unterkapitel soll einen besseren Eindruck der interessierenden Variablen *Abschlussnote, Stundenlohn, Geschlecht, juEltern, Kinder, Promotion, Privatkontakte, Studienkontakte* und zur *Art der Stellenfindung* liefern. Abbildung 2.1 zeigt Balkendiagramme der Variablen *Geschlecht* und *Kinder*. Es ist zu sehen, dass es unter den Absolventen nur wenig mehr Frauen ($n = 108$) als Männer ($n = 99$) gibt. Desweiteren gibt es mehr als doppelt so viele Absolventen, die noch kein Kind haben ($n = 142$) als Absolventen mit einem oder mehreren Kindern ($n = 64$).

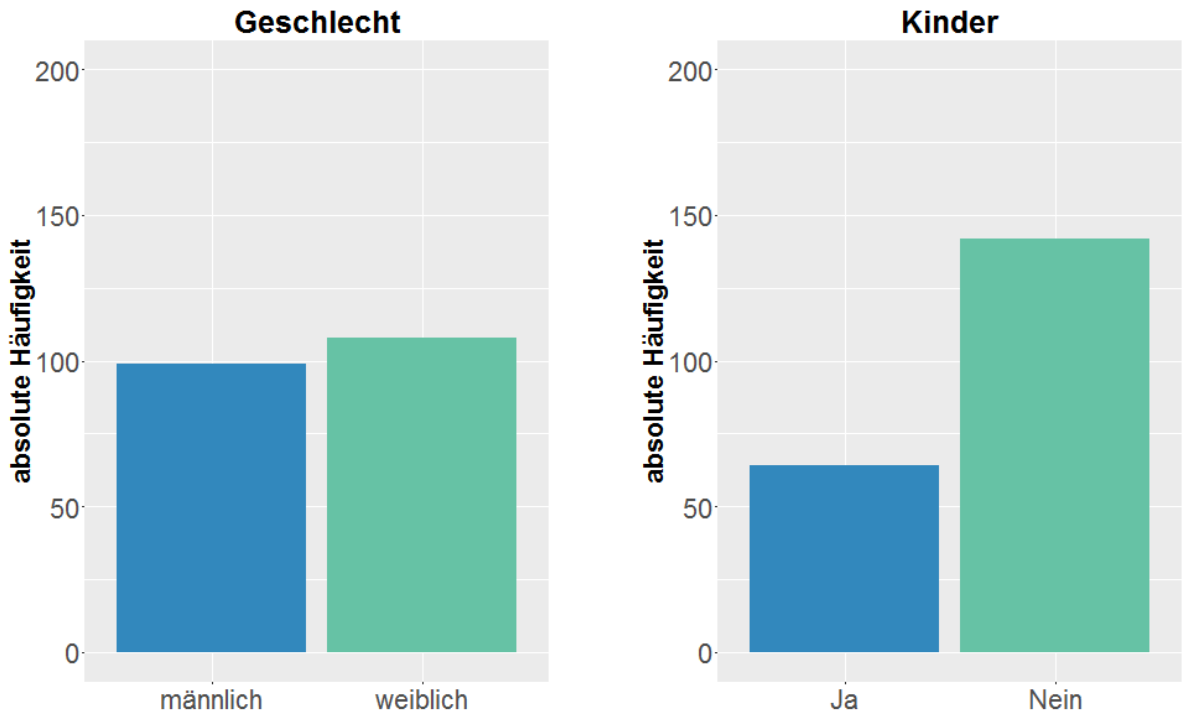

Abbildung 2.1: Welches Geschlecht haben die Absolventen und haben sie Kinder?

Abbildung 2.2 gibt sowohl Informationen zur Abschlussnote als auch der Anzahl der Privat- und Studienkontakte der Befragten. Die Grafik zeigt, dass der schlechteste Absolvent vier Punkte erreicht hat, der beste hingegen 14.1. Der Median liegt bei 8.07 Punkten, das heißt, dass 50% der Befragten sowohl mit einer besseren als auch schlechteren Punktzahl abgeschnitten haben. Die durchschnittliche Punktzahl beträgt 8.12. Der Absolvent mit den meisten Studienkontakten hat 70 von diesen, der zu dem Minimum gehörende Absolvent hat keine Studienkontakte. Im Mittel haben alle Absolventen 12.1 Studienkontakte, während 50% weniger als zehn haben. 100 Privatkontakte hat hingegen der Befragte mit den meisten Privatkontakten, während der Absolvent mit den wenigsten privaten Kontakte keine zu besitzen scheint. Im Mittel verfügt jeder Absolvent der Rechtswissenschaften über 31.02 private Kontakte. 50% von ihnen haben dabei weniger als 25.

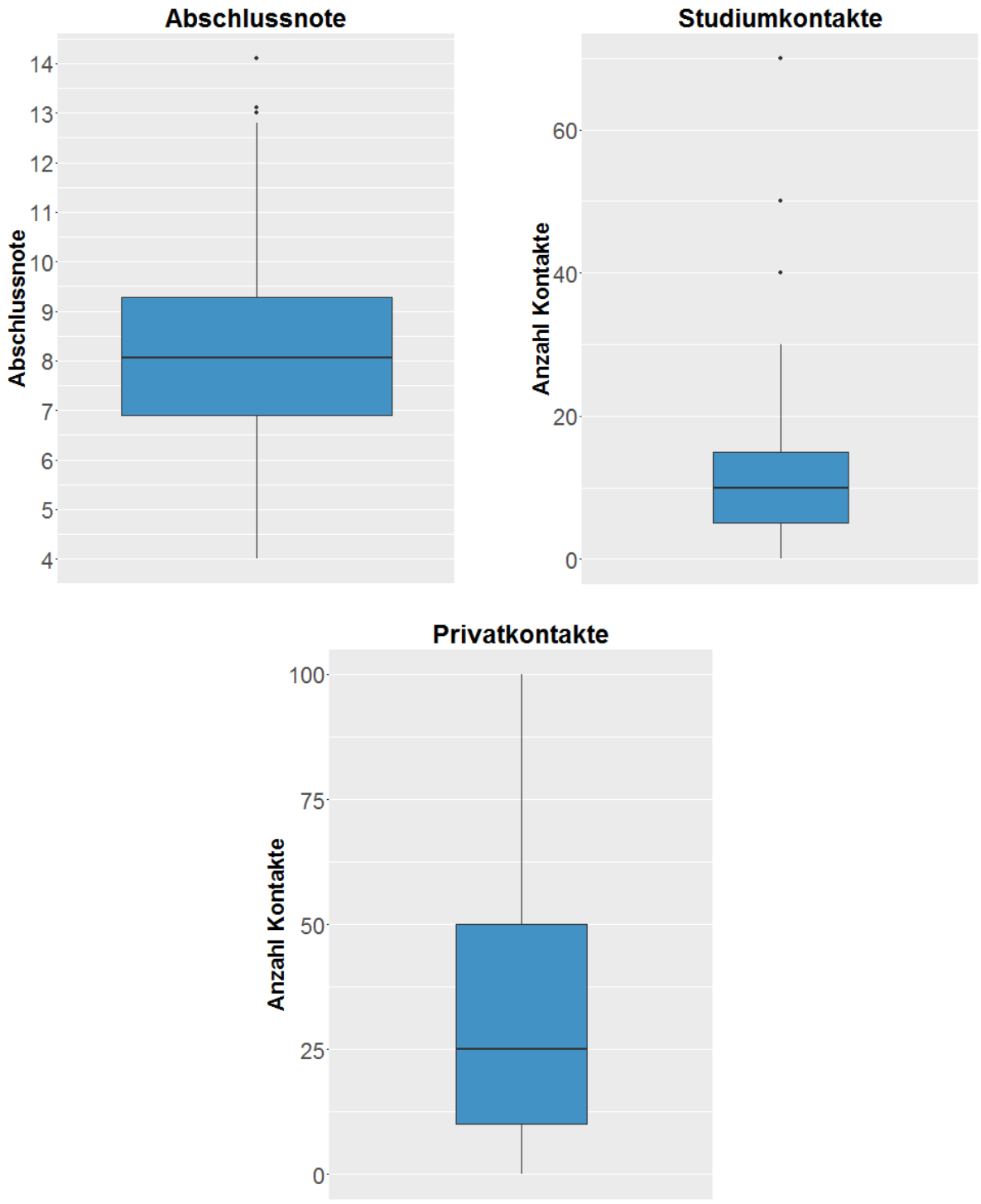

Abbildung 2.2: Verteilung der Abschlussnote sowie Anzahl der Studien- und Privatkontakte.

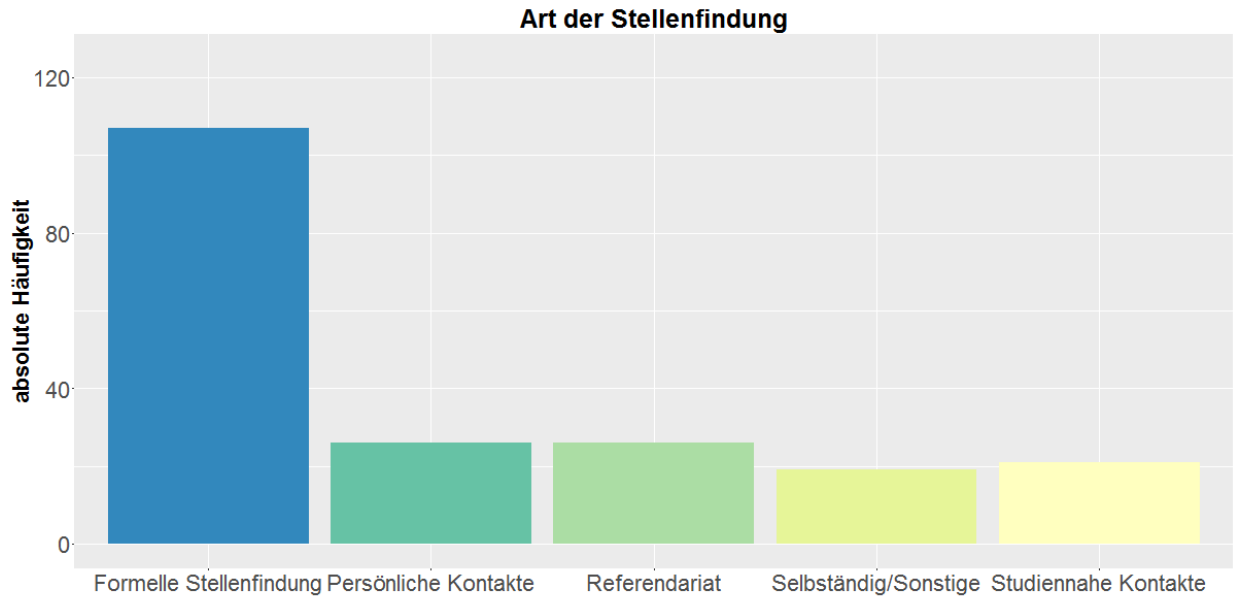

Abbildung 2.3: Art der Stellenfindung

Abbildung 2.3 zeigt, dass die meisten Stellen über fomelle Wege gefunden wurden ($n = 107$). Es folgen sowohl die Kategorien *Referendariat* und *Persönliche Kontakte* ($n = 26$) als auch *Studiennahe Kontakte* ($n = 21$). Die wenigsten Absolventen haben ihre Stelle über sonstige Wege oder selbständig gefunden ($n = 19$). Zwölf der Befragten haben keine Angabe dazu gemacht, wie sie ihre Stelle gefunden haben.

Abschließend können anhand von Abbildung 2.4 noch Aussagen darüber getroffen werden, ob die Absolventen Eltern mit abgeschlossenem Jurastudium oder eine abgeschlossene Promotion haben. Hierzu ist zu sagen, dass nur ein eher kleiner Anteil der Absolventen Eltern haben, welche ebenso ein Studium der Rechtswissenschaften abgeschlossen haben ($n = 30$). Der Großteil hat keine Eltern aus dem juristischen Bereich ($n = 167$). 14 Befragte machten hierzu keine Angabe. Die Promotion betreffend gibt mehr als dreimal soviele Absolventen ohne (abgeschlossene) Promotion ($n = 162$) wie solche mit Promotion ($n = 48$). Ein befragter Absolvent machte zu seinen Promotionsstatus keine Angabe.

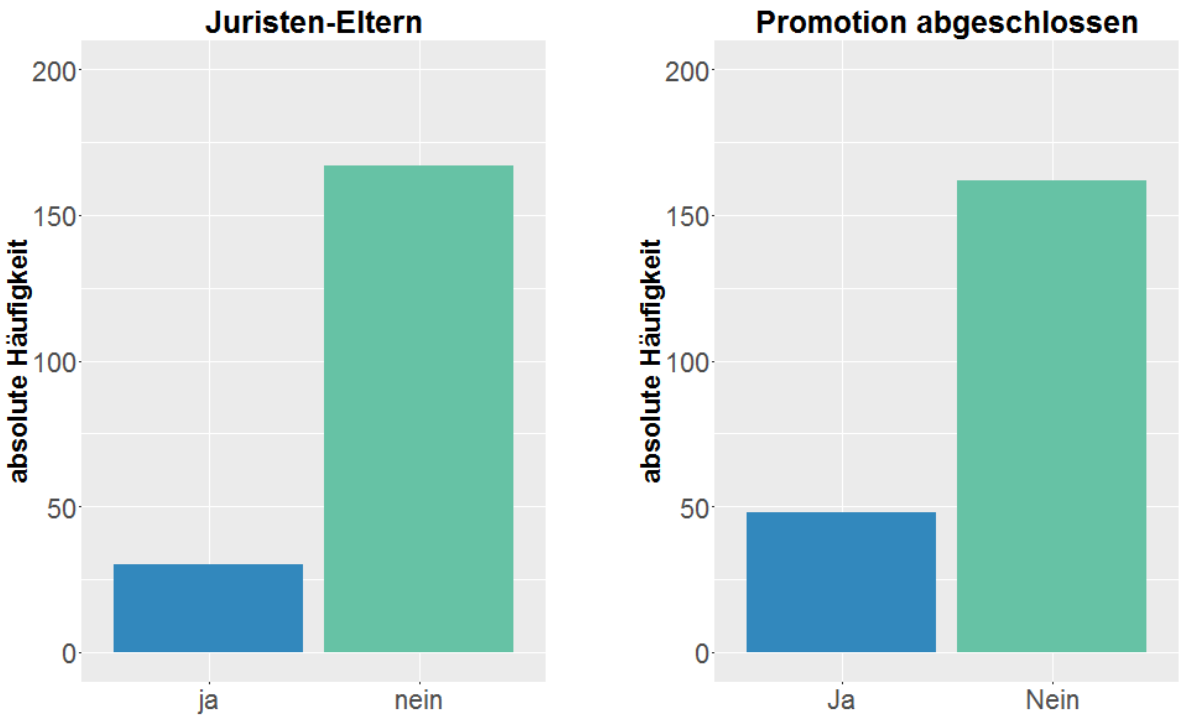

Abbildung 2.4: Absolute Häufigkeiten der Variablen zu Juristen-Eltern und Promotion

3 Statistische Methodik

In diesem Kapitel werden die für die Analyse verwendenten statistischen Methoden vorgestellt. Dazu folgt in Abschnitt 3.1 neben einem kurzen Überblick über die *Sequenzmusteranalyse* eine kurze Einführung in das `R`-Paket `TraMineR`, welches im Zuge der Sequenzmusteranalyse verwendet wurde.

3.1 Sequenzmusteranalyse

Die Sequenzmusteranalyse ist eine (überwiegend) deskriptive Analysemethode für Längsschnittdaten. Sie wird zum einen in den Sozialwissenschaften angewandt, um Lebensverläufe wie die hier vorgestellten Berufsverläufe der Juristen zu analysieren. Zum anderen findet sie aber auch Anwendung in der Genetik bei der Analyse von Gensequenzen. Bei der Verwendung der Sequenzmusteranalyse soll überwiegend untersucht werden, ob Muster oder Ähnlichkeiten in den Verläufen vorliegen. Damit dies möglich ist, müssen die Daten zunächst in ein geeignetes Format gebracht werden. Ein Beispiel wie diese Sequenzdaten auszusehen haben ist mit Tabelle 3.1 gegeben.

id	Monat 1	Monat 2	Monat 3	Monat 4	Monat 5	Monat 6
1	Re	Re	A	S	Ri	Ri
2	Re	Re	Re	Ö	Ö	Ö

Tabelle 3.1: Zustände: Re=Referendariat, A=Arbeitslos, S=Staatsanwalt, Ri=Richter, Ö=Öffentliche Verwaltung

Tabelle 3.1 zeigt die Berufsverläufe für die Personen mit den ID's 1 und 2. Die Person mit der ID 1 befand sich die ersten zwei Monate im Referendariat. Nach einem Monat der Arbeitslosigkeit folgte zunächst ein Monat, in dem sie als Staatsanwalt tätig war. Die letzten zwei Monate arbeitete sie als Richter. Bei der Person mit der ID 2 standen zu Beginn drei Monate Referendariat nach welchen sie direkt in die öffentliche Verwaltung wechselte und dort auch blieb.

Das vorrangegangene Beispiel soll nur zu einem besseren Verständnis beitragen und repräsentiert nur grundlegend die vorliegenden Daten!

Die Darstellung der Sequenzen erfolgt mit Hilfe des `R`-Pakets `TraMineR` (s. Gabadinho et al. (2016)). Möglichkeiten der grafischen Darstellungen sind in Abschnitt 4.1 zu finden. Die Informationen über die Sequenzen,welche man von `TraMineR` erhält, sind vielfältig. Einige ausgewählte Möglichkeiten sind die Ausgabe von

- Zustands-Sequenz-Objekten: Objekt, in welchem nur noch die Sequenzen gespeichert sind und als Basis für die weiteren Analysen dient
- den Namen der Zustände
- Tabellen der absoluten und relativen Sequenzhäufigkeiten
- etc.

Weiter können mit Hilfe des Pakets Grafiken erzeugt werden, welche

- alle Sequenzen,
- die x ersten bzw. häufigsten Sequenzen,
- die durchschnittlich in einem Zustand verbrachte Zeit,
- die relative Zustandshäufigkeit und
- viele weitere Möglichkeiten

abbilden.

3.2 Optimal Matching

Als Vorbereitung auf die Clusteranalyse (siehe Abschnitt 3.3) müssen die paarweisen Distanzen zwischen den Sequenzen bestimmt werden. Hierfür gibt es zunächst mehrere Möglichkeiten. Eine simple und leicht verständliche Vorgehensweise ist zum Beispiel die sogenannte Hamming-Distanz (HD). Zur Bestimmung dieser werden die einzelnen Elemente a_t und b_t der Sequenzen A und B an jeder Stelle $t = 1, \ldots, T$ miteinander verglichen. Hierbei werden die Substitutionskosten c_{st}, welche den Aufwand des Austauschens

von Zustand a_t durch Zustand b_t quantifizieren, konstant auf $c_{st} = 1$ festgelegt. Sind die verglichenen Zustände identisch, so betragen die Substitutionskosten $c_{st} = 0$. Gilt nun $a_t = b_t$, so erhöht sich das Distanzmaß nicht. Anders ausgedrückt: es ergeben sich Kosten von 0. Liegt jedoch ein Unterschied zwischen den Sequenzen vor ($a_t \neq b_t$), so erhöht sich das Distanzmaß um 1 oder anders gesagt: es ergeben sich Kosten von 1. Somit reicht die mögliche resultierende HD von 0 (die beiden Sequenzen sind identisch) bis T (die beiden Sequenzen sind sich komplett unähnlich). Es gilt:

$$\text{HAMMING}(\mathbf{a}, \mathbf{b}) = \sum_{a_t \neq b_t} 1, \ t = 1, \ldots, T \tag{3.1}$$

Generell können die Substitutionskosten auch auf einen anderen konstanten Wert gesetzt werden. Für jeden Wert von c_{st} ergibt sich dann für die HD:

$$\text{HAMMING}(\mathbf{a}, \mathbf{b}) = \sum_{a_t \neq b_t} c_{st}, \ t = 1, \ldots, T \tag{3.2}$$

(Bacher (2008), S.479 f.)

Die Problematik bei Verwendung der HD liegt darin, dass sie bei komplett unterschiedlichen Sequenzen den gleichen Wert annehmen wird wie bei Sequenzen, innerhalb derer nur die Subsequenzen verschoben sind. Infolgedessen wird der „. . . ganzheitliche[n] Charakter der Sequenzen . . . “ ignoriert (Bacher (2008), S.248). Veranschaulicht werden kann dieser Punkt mit Hilfe von Tabelle 3.2. Da bei Verwendung der HD nur Substitutionen möglich sind, ergibt sich für die HD bei allen drei Sequenzvergleichen eine Distanz von sechs.

Gelöst werden kann dieses Problem, indem an Stelle der HD die Levenshtein-Distanz (LD) verwendet wird. Im Gegensatz zur HD sind nun nicht nur Substitutionen (ersetzen eines Zustands durch einen anderen) möglich, sondern auch sogenannte *indel*-Operationen. Eine solche indel-Operation besteht entweder aus dem Einfügen (*insert*) oder dem Löschen (*delete*) eines Zustands. Eine Substitution kann somit durch zwei indel-Operationen ersetzt werden weswegen die indel-Kosten mit dem Faktor 0.5 zu gewichten sind (Stegmann et al. (2013), S.53). Für festgelegte Substitutionskosten von $c_{st} = 2$ würden sich somit für die *indel*-Operationen Kosten von $c_{indel} = 0.5 \cdot c_{st} = 1$ ergeben. Bei der Bestimmung der LD wird sowohl mit Substitutionen als auch mit *indel*-Operationen gearbeitet und diejenige Kombination aus beiden gewählt, welche die geringsten Kosten verursacht. Auch

dies erklärt Tabelle 3.2. Angenommen die Substiutionskosten liegen bei $c_{st} = 1$ und die *indel*-Kosten, nicht mit dem Faktor 0.5 gewichtet, ebenso ($c_{indel} = 1$). Wie in der Tabelle zu sehen ist, ergibt sich bei Vergleich 1 eine Levenshtein-Distanz von 6. Diese ergibt sich aus den Kosten für sechs Substitutionen $6 \cdot c_{st} = 6$. Eine Kombination aus Substitutionen und *indel*-Operationen würde zu einer Distanz > 6 führen. Würden nur *indel*-Operationen durchgeführt, so ergäbe sich eine Distanz von $12 \cdot c_{indel} = 12$, da hier zunächst sechsmal A entfernt und danach sechsmal eingefügt werden müsste. Die Levenshtein-Distanz ist nun die minimal mögliche Distanz (hier: 6). Auf die gleiche Weise ergibt sich für Vergleich 2 ebenso eine Levenshtein-Distanz von 6 (sechs Substitutionen oder sechs *indel*-Operationen) und für Vergleich 3 eine LD von 2 (zwei *indel*-Operationen). Hier werden zwar immerhin schon die letzten zwei Sequenzen als zueinander ähnlicher angesehen als es die HD tut, jedoch erhält man für Vergleich 1 die gleiche Distanz wie für Vergleich 2, obwohl sich die Sequenzen bei letzterem viel ähnlicher sind. Hier kommt nun die Gewichtung der *indel*-Operationen mit dem Faktor 0.5 ins Spiel. Jetzt erhält man für Vergleich 1 eine LD von $6 \cdot c_{st} = 6$, für Vergleich 2 ergibt sich $6 \cdot c_{indel} = 3$ und für den dritten Vergleich $2 \cdot c_{indel} = 1$.

Mit der Einführung der indel-Operationen werden die „…analytischen Schwächen der ‚naiven' Distanzmaße überwunden"(Bacher (2008), S.481). Das bedeutet, dass nun berücksichtigt wird, falls sich Sequenzen nur aufgrund von verschobenen Subsequenzen unterscheiden, die Abfolge der Zustände jedoch weitestgehend identisch ist. Auch diese Tatsache wird anhand von Tabelle 3.2 verdeutlicht. Festzuhalten ist deshalb, dass der Distanzmatrix die Levenshtein-Distanz zugrunde liegt. Dies ist ein wichtiger Punkt, der später im Zuge der Clusteranalyse noch zum Tragen kommen wird (vgl. Abschnitt 3.3).

	Vergleich 1	Vergleich 2	Vergleich 3
Sequenz 1	ÖÖÖÖÖÖ	AAAÖÖÖ	FÖASRÖ
Sequenz 2	AAAAAA	ÖÖÖAAA	ÖASRÖK
Hamming-Distanz	6	6	6
Levensthein-Distanz 1 ($c_s = 1; c_i = 1$)	6	6	2
Levenshtein-Distanz 2 ($c_s = 1; c_i = 0.5$)	6	3	1

Tabelle 3.2: Vergleich der Distanzen (Bacher (2008), S.483)

Bevor die Distanzen jedoch bestimmt werden können, müssen die Substitutionskosten festgelegt werden. Hierbei gibt es mehrere Möglichkeiten. Zum einen können die Kosten konstant auf einen bestimmten Wert c_{st} festgesetzt werden. Der Nachteil hierbei ist jedoch, dass diese Festlegung eher willkürlich geschieht und weiter nicht berücksichtigt wird, dass die Substitution von Zustand A durch Zustand B eventuell weniger kostenaufwändig ist als die Substitution von Zustand A durch Zustand C. Aus diesem Grund empfiehlt sich die Berechnung der Substitutionskosten aus den Daten. Hierfür werden diese in der Regel zunächst konstant auf den Wert 2 festgesetzt. Im nächsten Schritt werden die Übergangswahrscheinlichkeiten (auch Transitionswahrscheinlichkeiten) (ÜW) berechnet und diese vom zuvor festgesetzten Wert 2 abgezogen. Es ergeben sich dadurch die finalen Substitutionskosten für den paarweisen Austausch der Sequenzen. (Stegmann et al. (2013), S.54f.) Die Berechnung der Transitionsmatrix geschieht dabei wie folgt:

$$w(a,b) = \begin{cases} 2 - p(a,b) - p(b,a), & \text{wenn } a \neq b \\ 0, & \text{wenn } a = b \end{cases} \tag{3.3}$$

mit $w(a,b)$ ist T-Rate für die Zustände a und b und $p(a,b)$ bzw. $p(b,a)$ Übergangsrate von Zustand a nach b bzw. b nach a. Die Übergangsraten $p(a,b)$ bzw. $p(b,a)$ ergeben sich als

$$p(a,b) = \frac{\sum_{t=1}^{T-1} N_{t,t+1}(a,b)}{\sum_{t=1}^{T-1} N_t(a)} \tag{3.4}$$

mit $N_t(a)$ als Anzahl der Sequenzen im Zustand a zum Zeitpunkt t und $N_{t,t+1}(a,b)$ als Anzahl der Sequenzen im Zustand a zum Zeitpunkt t und im Zustand b zum Zeitpunkt $t+1$ (Stegmann et al. (2013), S.55). $T-1$ steht bei T Zuständen für die Gesamtzahl der möglichen Zustandswechsel. Mit dieser so erzeugten Kostenmatrix und den entsprechend festgelegten Kosten für eine *indel*-Operation lässt sich die Levenshtein-Distanz paarweise berechnen und dadurch die Distanzmatrix erzeugen.

3.3 Clusteranalyse

Eine Möglichkeit innerhalb der Sequenzmusteranalyse ist das Clustern der einzelnen Sequenzen. Dadurch sollen Gruppen (Cluster) gebildet werden, die aus möglichst ähnlichen Sequenzen bestehen. Dabei gibt es verschiedene Vorgehensweisen. Eine ist das hierarchisch

agglomerative Clustern mit der Methode nach *Ward*. Hierbei bilden die einzelnen Sequenzen den Ausgangspunkt und werden Schritt für Schritt anhand eines zu minimierenden Kriteriums zusammengefügt, bis es nur noch einen einzigen Cluster gibt, der aus allen Sequenzen besteht. Im Fall der *Ward*-Methode ist der Anstieg der Fehlerquadratsumme, der entsteht, wenn eine neue Sequenz mit einem bereits bestehenden Cluster fusioniert wird, zu minimieren. Eine andere Beschreibung wäre, dass man versucht, die Varianz innerhalb des Clusters zu minimieren, während gleichzeitig „die Streuung zwischen den Clusterzentren maximiert wird"(Bacher (2008), S.150). Diese Art der Clustermethode hat den Vorteil, dass die entstehenden Cluster von ähnlicher Größe sind (Stegmann et al. (2013), S.61). Da das *Ward-Verfahren* „Clusterzentren als Repräsentanten bei der Clusterbildung verwendet" (Bacher (2008), S.61), sollten die Daten möglichst quantitativ sein, da nur für diese Mittelwertbildungen möglich sind (Bacher (2008), S.61).
Liegen als (Un-)Ähnlichkeitsmaße euklidische Distanzen vor, so lässt sich die Fehlerquadratsumme innerhalb eines Clusters q auch schreiben als Summe der quadrierten Distanzen $d^2(i, q^*)$ jeder Beobachtung $i \in q$ zum Clustermittelpunkt q^*, also

$$\sum_{i \in q} d^2(i, q^*) \tag{3.5}$$

und

$$q^* = \frac{1}{|q|} \sum_{i \in q} i. \tag{3.6}$$

(Murtagh und Legendre (2014), S.277)
Um die Berechnung des Clustermittelpunktes zu umgehen, kann die Fehlerquadratsumme auch als Summe der paarweisen Distanzen aller Elemente innerhalb eines Clusters q dargestellt werden. Es gilt die folgende Äquivalenz:

$$\sum_{i \in q} d^2(i, q^*) = \frac{1}{|q|} \sum_{i,i' \in q, i<i'} d^2(i, i') \tag{3.7}$$

Die Äquivalenz wird in Murtagh und Legendre (2014) (S.279) wie folgt bewiesen:

$$
\begin{aligned}
\frac{1}{|q|}\sum_{i,i'\in q, i<i'} d^2(i,i') &= \frac{1}{|q|}\sum_{i,i'\in q, i<i'} (i-i')^2 \\
&= \frac{1}{|q|}\sum_{i,i'\in q, i<i'} (i-q^*-(i'-q^*))^2 \\
&= \frac{1}{|q|}\sum_{i,i'\in q, i<i'} ((i-q^*)^2+(i'-q^*)^2-2(i-q^*)(i'-q^*)) \\
&= \frac{1}{2}\frac{1}{|q|}\sum_{i\in q}\sum_{i'\in q} ((i-q^*)^2+(i'-q^*)^2-2(i-q^*)(i'-q^*)) \\
&= \frac{1}{2}\frac{1}{|q|}\left(2|q|\sum_{i\in q}(i-q^*)^2\right) - \underbrace{\frac{1}{2}\frac{1}{|q|}\left(\sum_{i\in q}\sum_{i'\in q} 2(i-q^*)(i'-q^*)\right)}_{=0} \\
&= \sum_{i\in q} d^2(i,q^*)
\end{aligned}
$$

Die in `R` innerhalb der Funktion `agnes()` implementierte sogenannte *Update*-Formel lautet:

$$
\delta(I\cup J,K) = \sqrt{\frac{|I|+|K|}{|I|+|J|+|K|}\delta^2(I,K) + \frac{|J|+|K|}{|I|+|J|+|K|}\delta^2(J,K) - \frac{|K|}{|I|+|J|+|K|}\delta^2(I,J)} \tag{3.8}
$$

$|.|$ entspricht jeweils der Kardinalität des Clusters. Als Input werden hier euklidische Distanzen vorausgesetzt. Diese werden innerhalb der Funktion noch quadriert. Es gibt auch Funktionen, welche bereits quadrierte euklidische Distanzen benötigen. Eine Überprüfung, welche Art der *Update*-Formel innerhalb der benutzten Funktion implementiert ist, ist daher empfehlenswert. (Murtagh und Legendre (2014), S.277ff.)

Nach der *Update*-Formel 3.8 werden folglich die neuen Distanzen zwischen dem neu entstandenen und den übrigen Clustern berechnet. Einfach ausgedrückt funktioniert die *Ward*-Methode wie folgt:

1. Fusionierung der zwei Objekte mit der geringsten Distanz.
2. Berechne die neuen Distanzen anhand von Gleichung 3.8.

Da die berechnete Distanz bei Verwendung euklidischer Distanzen der Fehlerquadratsumme entspricht (vgl. Gleichung 3.7), wird der Anstieg dieser mit Durchführung des ersten Schrittes minimiert. Diese zwei Schritte werden, wie zu Beginn bereits erwähnt, so lange wiederholt bis nur noch ein einziger Cluster übrig ist. Dieser enthält alle Objekte.

Da hier jedoch Sequenzen vorliegen, ist es nicht möglich euklidische Distanzen zu bestimmen. Wie in Abschnitt 3.2 bereits ausführlich beschrieben, liegt der Distanzmatrix die Levenshtein-Distanz zugrunde. Die Anwendung der *Ward*-Methode ist zwar grundsätzlich kein Problem, jedoch muss man sich darüber im Klaren sein, dass aufgrund der Levenshtein-Distanzen nicht die *Ward*-Methode im eigentlichen Sinn ausgeführt wird. Aufgrund des Fehlens der euklidischen Distanzen wird mit der *Update*-Formel nicht die Varianz innerhalb der Cluster minimiert, sondern lediglich die Summe der quadrierten paarweisen Distanzen. Dies ist jedoch nicht weiter problematisch. Vielmehr werden durch die Quadrierung Sequenzen, welche sich sehr unähnlich sind, noch stärker gewichtet, das heißt als noch unähnlicher angesehen.

Danach muss man sich für eine dieser entstandenen Clusterlösungen entscheiden, d.h. festlegen, wieviele Cluster man als Ergebnis erhalten möchte. Diese Entscheidung ist eher willkürlich, es gibt jedoch einige Kennzahlen, welche einem bei der Wahl der Clusteranzahl helfen können. Bei der Verwendung von Sequenzdaten ist davon jedoch nur die *within-between-Cluster-Distanz* empfehlenswert. Bei der regulären Anwendung auf metrische Daten wird hier für jede Clusterlösung die Varianz innerhalb der Cluster durch die Varianz zwischen den verschiedenen Clustern dividiert. Analog zum minimierenden Kriterium beim Clustern, wird auch hier nicht mit der Varianz gearbeitet, sondern mit den paarweisen, quadrierten Distanzen. Es werden also die paarweisen, quadrierten Distanzen innerhalb der Cluster durch die paarweisen, quadrierten Distanzen zwischen den Clustern dividiert. Es ist ratsam sich für diejenige Clusterlösung zu entscheiden, bei welcher der Wert 0.5 das erste Mal unterschritten wird. Zusätzlich zu diesem Kriterium sollte eine Clustervalidierung anhand inhaltlicher Merkmale stattfinden. (Stegmann et al. (2013), S.68f.)

3.4 Repräsentative Sequenzen

Ziel der repräsentativen Sequenzen ist es, ein möglichst kleines Subset der vorliegenden Sequenzen zu finden, welches alle Sequenzen (oder zumindest einen festgelegten Anteil) bezüglich eines zuvor bestimmten Genauigkeitslevels repräsentiert. Die Idee dabei ist, diejenigen Sequenzen einer repräsentativen Sequenz zuzuordnen, die in der Nachbarschaft dieser liegen, das heißt, die dieser bzgl. eines bestimmten Kriteriums ähnlich sind. (Gabadinho und Ritschard (2013))
Dies hat den Nutzen, dass auch Sequenzen, welche sich nur durch um einige Zeiteinheiten verschobene Zustände unterscheiden, als sich untereinander ähnlich eingestuft werden.
Nach Gabadinho et al. (2011) verwendet die heuristische Methode, um die repräsentativen Sequenzen zu erhalten, folgende Schritte:

1. Berechnung eines Repräsentativitäts-Scores für alle Sequenzen nach einem bestimmten Kriterium
2. Sortierung der Sequenzen bzgl. dieses Scores (repräsentativste Sequenz an erster Stelle)
3. Festlegung einer Regel, um die Kandidatenliste für die repräsentativen Sequenzen einzuschränken (Anzahl der repräsentativen Sequenzen oder Coverage-Trade-Off)
4. Iteratives Vorgehen, um redundante Sequenzen von der Kandidatenliste zu entfernen, sodass nur Sequenzen verbleiben, die sich bzgl. des festgelegten Grenzwertes unähnlich sind

Um die Kandidatenliste zu erstellen, wird zunächst anhand eines bestimmten Kriteriums für jede einzelne vorliegende Sequenz ein Repräsentativitäts-Score berechnet, nach welchem die Sequenzen anschließend sortiert werden. Nach Gabadinho und Ritschard (2013) und Gabadinho et al. (2011) gibt es verschiedene Möglichkeiten das Kriterium zur Berechnung des Scores zu wählen:

- **Frequency:** Auftretenshäufigkeit einer Sequenz als Kriterium. Je häufiger eine Sequenz vorliegt, desto repräsentativer ist sie.
- **Neighborhood density:** Anzahl der Sequenzen in der Nachbarschaft als Kriterum. Für die Nachbarschaftsdichte werden in der Distanzmatrix die Zeilen oder Spalten gezählt, die zu der betrachteten Sequenz eine Distanz kleiner oder gleich eines

festgelegten Wertes haben. Hierzu ist die Wahl eines Nachbarschaftsradius notwendig. Eine geeignete Wahl für diesen ist ein bestimmter Anteil (in %) der theoretisch maximalen Distanz zwischen zwei Sequenzen.

- **Centrality:** Zentralität einer Sequenz als Kriterium. Das zentralste Objekt ist als jenes mit der kleinsten Summe der Distanzen zu allen anderen Objekten definiert. Je zentraler eine Sequenz ist, desto repräsentativer ist sie.

Ist der Repräsentativitäts-Score für alle Sequenzen berechnet, werden die Sequenzen so nach diesem sortiert, dass die repräsentativste Sequenz an erster Stelle steht, die zweitrepräsentativste an zweiter Stelle und so weiter. Außerdem muss entweder die Anzahl an gewünschten Repräsentanten oder ein so genannter *Coverage Trade-Off*-Wert gewählt werden. Dieser gibt an, wie viel Prozent aller Sequenzen in der Nachbarschaft der repräsentativen Sequenzen liegen sollen.
Hat man nun eine sortierte Kandidatenliste der Sequenzen, werden die repräsentativen Sequenzen aus dieser so ausgewählt, dass keine redundanten Sequenzen mehr vorliegen. Es wird mit der ersten Sequenz in der sortierten Liste (die beste Sequenz bzgl. des festgelegten Kriteriums) begonnen und mit den weiteren Sequenzen in der festgelegten Reihenfolge fortgefahren. Für jede Sequenz wird betrachtet, ob sie bzgl. des festgelegten Schwellenwertes ähnlich zu einer der bereits gewählten repräsentativen Sequenzen ist, das heißt, ob sie bzgl. der Distanz weit oder nah entfernt liegen und ordnet sie der nächstgelegensten zu. Kann sie keiner der vorhandenen Repräsentanten zugeordnet werden, wird sie als neue repräsentative Sequenz aufgenommen. Der dazu nötige Schwellenwert berechnet sich wie der Nachbarschaftsradius für das oben genannte *Neigborhood-Density*-Kriterium als Anteil der theoretisch maximalen Distanz. Im Fall von *OM* Distanzen ist diese für zwei Sequenzen (s_1, s_2) mit den Längen (l_1, l_2) definiert als

$$D_{max} = min(l_1, l_2) \cdot min(2C_I, max(S)) + |l_1 - l_2| \cdot C_I. \tag{3.9}$$

Dabei sind C_I die Indelkosten und $max(S)$ die maximalen Substitutionskosten. Sequenz s_1 ist redundant zu Sequenz s_2, wenn sie in der Nachbarschaft von s_1 liegt, das heißt $d(s_1, s_2)$ nicht größer als der Nachbarschaftsradius ist.
Jedes mal, wenn eine neue Sequenz dem Set der Repräsentanten hinzugefügt wird, wird der *Coverage*-Anteil neu berechnet. Wurde durch den *Coverage Trade-Off* ein bestimmter Anteil vorgegeben, wird der Auswahlprozess gestoppt, sobald dieser erreicht ist. Andernfalls terminiert der Algorithmus, wenn die gewünschte Anzahl an repräsentativen Sequenzen

gefunden wurde. Auf diese Weise erhält man letztendlich die Liste mit der gewünschten Anzahl bzw. dem gewählten *Coverage* der repräsentativen Sequenzen. (Gabadinho et al. (2011))
Zusätzlich können Qualitätsmaße für die gefundenen repräsentativen Sequenzen berechnet werden. Die zwei hier vorgestellten Maße werden auch in den in Abschnitt 4.2 gezeigten Grafiken abzulesen sein. Zum einen berechnet sich die mittlere Distanz zur repräsentativen Sequenz (Mean distance to representative sequenz) als

$$MD_i = \frac{SD_i}{a_i}, \text{ mit } SD_i = \sum_{j \in R_i} d(x_j, r_i) \tag{3.10}$$

und zum anderen die mittlere Distanz zum Zentrum (Mean distance to center) durch

$$V_i = \frac{SC_i}{a_i}, \text{ mit } SC_i = \sum_{j \in R_i} d(x_j, c_i). \tag{3.11}$$

Dabei sind $r_i,\ i = 1, \ldots, k$, die repräsentativen Sequenzen, R_i das Set der Indizes der zu r_i zugeordneten Sequenzen und $a_i = |R_i|$ die Anzahl dieser zugeordneten Sequenzen. x_i bzw. x_j steht für die Sequenz, die einem Repräsentant zugeordnet werden soll und c_i für das Zentrum innerhalb einer repräsentativen Sequenz. Je kleiner diese Maße, desto ähnlicher sind sich im Allgemeinen die Sequenzen, die einer repräsentativen Sequenz zugewiesen werden. (Gabadinho und Ritschard (2013))
Außerdem lässt sich die *Overall-Quality* des gewählten Sets an repräsentativen Sequenzen berechnen. Diese ist definiert als

$$Q = \frac{\sum_i^k DC_i - \sum_i^k SD_i}{\sum_i^k DC_i} = \sum_{i=1}^{k} \frac{DC_i}{\sum_{j=1}^k DC_j} Q_i \tag{3.12}$$

mit

$$DC_i = \sum_{j \in R_i} d(x_j, c) \text{ und } Q_i = \frac{DC_i - SD_i}{DC_i}$$

als einzelnes Qualitätsmaß für die repräsentative Sequenz r_i. Die Idee ist hierbei, zu messen wie viel näher die repräsentativen Sequenzen an den einzelnen Sequenzen liegen als das wahre Zentrum c. (Gabadinho und Ritschard (2013))

3.5 Regression

"Regression ist die wohl am häufigsten eingesetzte statistische Methode zur Analyse empirischer Fragestellungen in Wirtschafts-, Sozial- und Lebenswissenschaften"(Fahrmeir et al. (2009), Vorwort zur 1. Auflage). Um auch die in diesem Projekt vorliegenden Daten durch Regression näher zu analysieren, werden in Unterabschnitt 3.5.1 und 3.5.2 kurz die grundlegenden Ideen und die zugrundeliegende Theorie der Linearen bzw. der Multinomialen Regression erläutert.

3.5.1 Lineare Regression

Interessiert man sich für den Einfluss mehrerer Kovariablen $x_1, \ldots, x_k$ auf die interessierende Zielgröße y, so kann der Zusammenhang zwischen diesen durch klassische lineare Regression modelliert werden. Dabei wird das Modell definiert als:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \epsilon_i = \mathbf{x_i}^T \boldsymbol{\beta} + \epsilon_i, \quad i = 1, \ldots, n. \tag{3.13}$$

In Matrixnotation kann dies auch geschrieben werden als:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{3.14}$$

wobei gilt, dass

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \ \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1k} \\ \vdots & \vdots & \ldots & \vdots \\ 1 & x_{n1} & \ldots & x_{nk} \end{pmatrix}, \ \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \text{ und } \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

Weiter müssen folgende Annahmen gelten:

1. $E(\epsilon) = \mathbf{0}$
2. $Cov(\epsilon) = E(\epsilon\epsilon') = \sigma^2 \boldsymbol{I}$
3. $rg(\mathbf{X}) = k + 1 = p$, d.h. Designmatrix $\mathbf{X}$ hat vollen Spaltenrang
4. $\epsilon \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I})$ (bei klassischer Normalregression)

(Fahrmeir et al. (2009))
Die durch das Modell geschätzten Regressionkoeffizienten $\hat{\boldsymbol{\beta}}$ können für die Interpretation verwendet werden. Dabei unterscheidet man zwischen metrischen und kategorialen Einflussgrößen:

- Metrisch: Steigt x_{ij} um eine Einheit, so verändert (steigt/sinkt) sich y_i um $\hat{\beta}_j$ bei Konstanthalten aller anderen Variablen (c.p.).
- Kategorial (binäre/mehrkategorial): Beim Übergang von $x_{ij} = 0$ zu $x_{ij} = 1$ (im mehrkategorialen Fall $x_{ij} \in \{0, 1, 2, \ldots\}$), verändert sich y_i um $\hat{\beta}_j$, c.p.

3.5.2 Multinomiale Regression

Möchte man ein Regressionsmodell mit einer nominalskalierten, mehrkategorialen ($r \in \{1, \ldots, c\}$) Einflussgröße (z.B. die Variable *Art der Anstellung* mit den Kategorien {*unbefristet*, *befristet*, *selbstständig*}) schätzen, bietet es sich an, ein multinomiales Logit-Modell zu verwenden. Dies stellt eine Verallgemeinerung des Logit-Modells (im Fall einer binären Zielgröße) dar (Theorie zum binären Logit-Modell siehe Fahrmeir et al. (2009), Kapitel 4.1). Die Interpretation ist analog zum binären Logit-Modell, jedoch mit dem Unterschied, dass nun die Auftretenswahrscheinlichkeit der Kategorie r jeweils immer ins Verhältnis zur Referenzkategorie c gesetzt wird. (Fahrmeir et al. (2009))
Nach Fahrmeir et al. (2009) ist für die Zielvariable $Y_i \in \{1, \ldots, c\}$ und gegebene Kovariablen $\mathbf{x}_i$ die Auftretenswahrscheinlichkeit für Kategorie r mit $r = 1, \ldots, q$ $(q = c - 1)$ definiert als:

$$P(Y_i = r|\mathbf{x}_i) = \pi_{ir} = \frac{exp(\mathbf{x}_i^T \hat{\boldsymbol{\beta}}_r)}{1 + \sum_{s=1}^{q} exp(\mathbf{x}_i^T \hat{\boldsymbol{\beta}}_s)}. \tag{3.15}$$

Für die Referenzkategorie c hingegen gilt:

$$\pi_{ic} = 1 - \pi_{i1} - \ldots - \pi_{iq} = \frac{1}{1 + \sum_{s=1}^{q} exp(\mathbf{x}_i^T \hat{\boldsymbol{\beta}}_s)}. \tag{3.16}$$

Äquivalent dazu ist die Darstellung durch die Linkfunktion:

$$log\left(\frac{P(y_i = r|\mathbf{x}_i)}{P(y_i = c|\mathbf{x}_i)}\right) = log\left(\frac{\pi_r}{\pi_c}\right) = \mathbf{x}_i^T\hat{\boldsymbol{\beta}}_r$$

$$\text{bzw.} \quad \frac{P(y = r|\mathbf{x}_i)}{P(y = c|\mathbf{x}_i)} = exp\{\mathbf{x}_i^T\hat{\boldsymbol{\beta}}_r\}. \tag{3.17}$$

Hierbei sind sowohl die Parameter $\hat{\boldsymbol{\beta}}_r = (\hat{\beta}_{r0}, \hat{\beta}_{r1}, \ldots, \hat{\beta}_{rk})^T$, als auch die linearen Prädiktoren $\eta_{ir} = \mathbf{x}_i^T\hat{\boldsymbol{\beta}}_r = \hat{\boldsymbol{\beta}}_{r0} + x_{i1}\hat{\boldsymbol{\beta}}_{r1} + \ldots + x_{ik}\hat{\boldsymbol{\beta}}_{rk}$ kategorienspezifisch mit $r = 1, \ldots, q$. Die Interpretation geschieht daraufhin über die logarithmierten Chancen wie beim binären Logit-Modell, jedoch jetzt immer in Bezug auf die Referenzkategorie c. (Fahrmeir et al. (2009))
Zum Beispiel wäre die Interpretation eines linearen Prädiktors der Form

$$log\left(\frac{\text{Wahrscheinlichkeit für Cluster r}}{\text{Wahrscheinlichkeit für Cluster c}}\right) =$$
$$\hat{\beta}_0 + x_{Geschlecht}\hat{\beta}_{r1} + x_{Juristen-Eltern}\hat{\beta}_{r2} + x_{Note}\hat{\beta}_{r3} + \ldots \tag{3.18}$$

folgendermaßen:

- **Geschlecht** (binär): Beim Übergang von *"männlich"* zu *"weiblich"* ändert sich die Chance von $Y = r$ zu $Y = c$ multiplikativ um $exp(\hat{\beta}_{r1})$, c.p.
- **Note** (metrisch): Steigt x_{Note} um eine Einheit, so ändert sich die Chance von $Y = r$ zu $Y = c$ multiplikativ um $exp(\hat{\beta}_{r3})$, c.p.

4 Ergebnisse

4.1 Sequenzmusteranalyse

Wie bereits in Abschnitt 3.1 kurz erwähnt, liefert das Paket `TraMineR` zahlreiche Grafiken als Mittel der Deskription der Sequenzen. Im Folgenden sollen einige ausgewählte Ergebnisse vorgestellt werden. Weitere, hier nicht genannte Ergebnisgrafiken, sind im Abschnitt A.1 zu finden. Um zunächst einen ersten Eindruck der Sequenzen zu bekom-

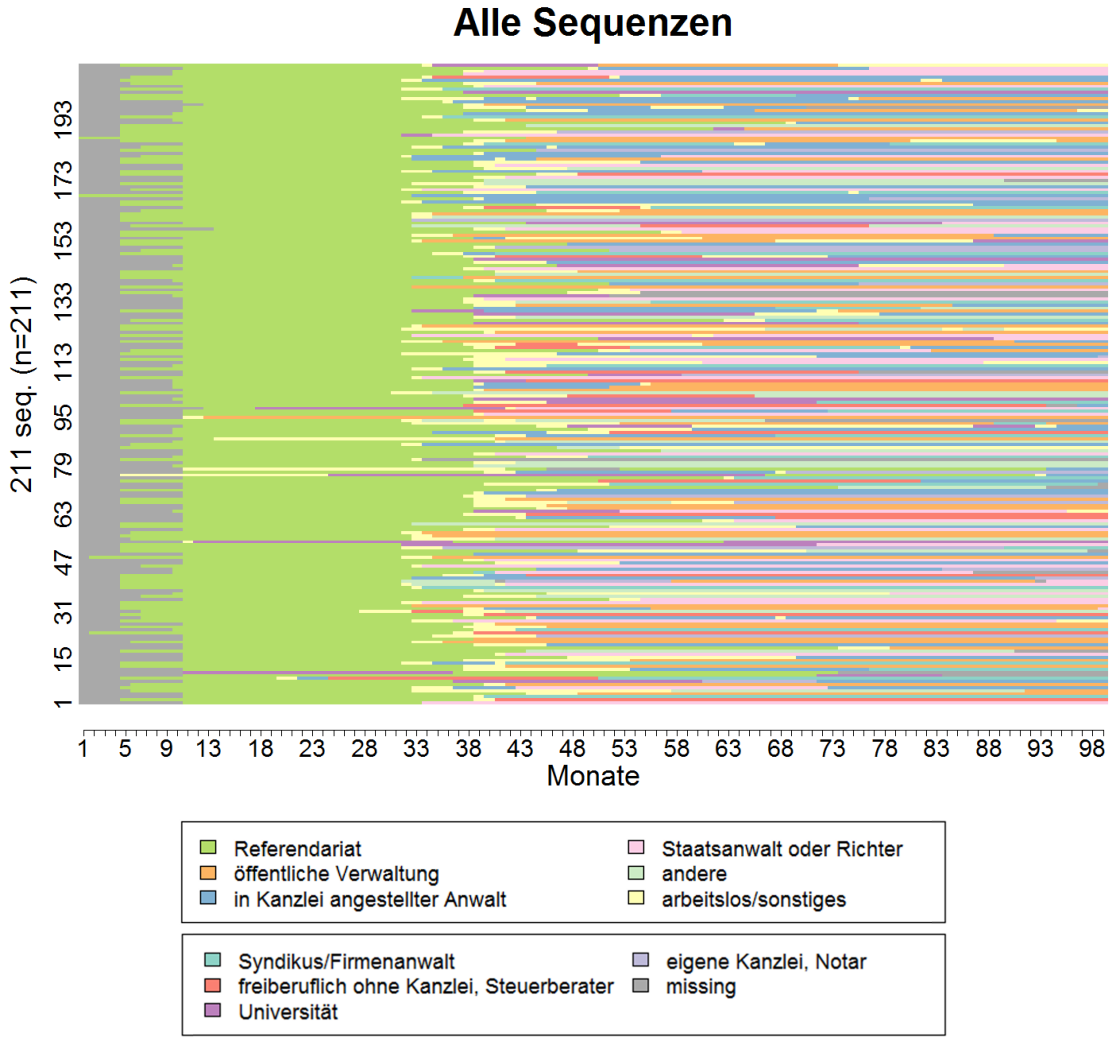

Abbildung 4.1: Alle Sequenzen.

men, ist es empfehlenswert, sich diese in einer Grafik anzusehen. Abbildung 4.1 zeigt alle Sequenzen. Während auf der y-Achse dabei die zu den Sequenzen gehörigen IDs abgetragen sind, zeigt die x-Achse den Zeitpunkt des Zustands in Monaten an. Am Anfang der Sequenzen steht ein großer, grüner Bereich. Dieser Bereich entspricht dem etwa zweijährigen Referendariat, welches Absolventen der Rechtswissenschaften nach ihrem Examen absolvieren müssen. Der unterschiedliche Zeitpunkt, zu welchem das Referendariat begonnen wurde, erklärt sich dadurch, dass die Daten die Berufsverläufe zweier Abschlussjahrgänge beinhalten. Nach dem Referendariat folgt ein großer, farbiger Block. Dieser stellt die unterschiedlichen Zustände der Absolventen nach dem Referendariat dar.

Während die Abbildung aller Sequenzen insgesamt nicht sehr übersichtlich ist, sind die einzelnen Sequenzen in Abbildung 4.2 gut zu erkennen. Hier sind nun die zehn häufigsten Sequenzen abgetragen. Je breiter die abgebildete Sequenz, desto mehr weitere Sequenzen gibt es, welche identisch zu der gezeigten sind. Weiter kann man auf der y-Achse den Anteil ablesen, den die gezeigten Sequenzen an der Gesamtzahl der Sequenzen ausmachen. In diesem Fall machen die zehn häufigsten Sequenzen 9% der gesamten Sequenzen aus. Zunächst kommt dadurch die Schlussfolgerung auf, dass die Berufsverläufe nicht homogen, das heißt, sehr unähnlich zueinander, sind. Betrachtet man die Verläufe der Sequenzen jedoch genauer, so fällt auf, dass sich die Sequenzen meist nur um wenige Zeitpunkte voneinander unterscheiden. So sind zum Beispiel die Verläufe der ersten, dritten und vierten Sequenz ähnlich: Zu Beginn steht das Referendariat, es folgt eine kurze Zeit der Arbeitslosigkeit, während die Absolventen danach als Staatsanwalt oder Richter tätig sind. Der Unterschied zwischen diesen Sequenzen liegt dabei hauptsächlich in der Dauer der Arbeitslosigkeit. Ebenso ähneln sich die Sequenzen zwei und sechs, sowie Sequenzen sieben, acht und neun. Hier können weitergehend *repräsentative Sequenzen* wie in Abschnitt 4.2 betrachtet werden, um sich ähnelde Sequenzen zusammenzufassen. Eine weitere Möglichkeit sich einen Überblick über alle Sequenzen zu verschaffen, ist die Darstellung der relativen Zustandshäufigkeit zu jedem Zeitpunkt wie in Abbildung 4.3. Diese Abbildung ist nicht wie die vorherigen von links nach rechts zu lesen, sondern von oben nach unten. Auf der y-Achse kann dabei die relative Zustandshäufigkeit abgelesen werden. In den ersten Monaten, in denen sich alle Absolventen im Referendariat befinden, beträgt die relative Häufigkeit des Zustands *Referendariat* dementsprechend 100%. Mit den Monaten kommen weitere Zustände hinzu, wodurch die relative Häufigkeit von *Referendariat* abnimmt und die der hinzukommenden Zustände zunimmt. Dabei machen die

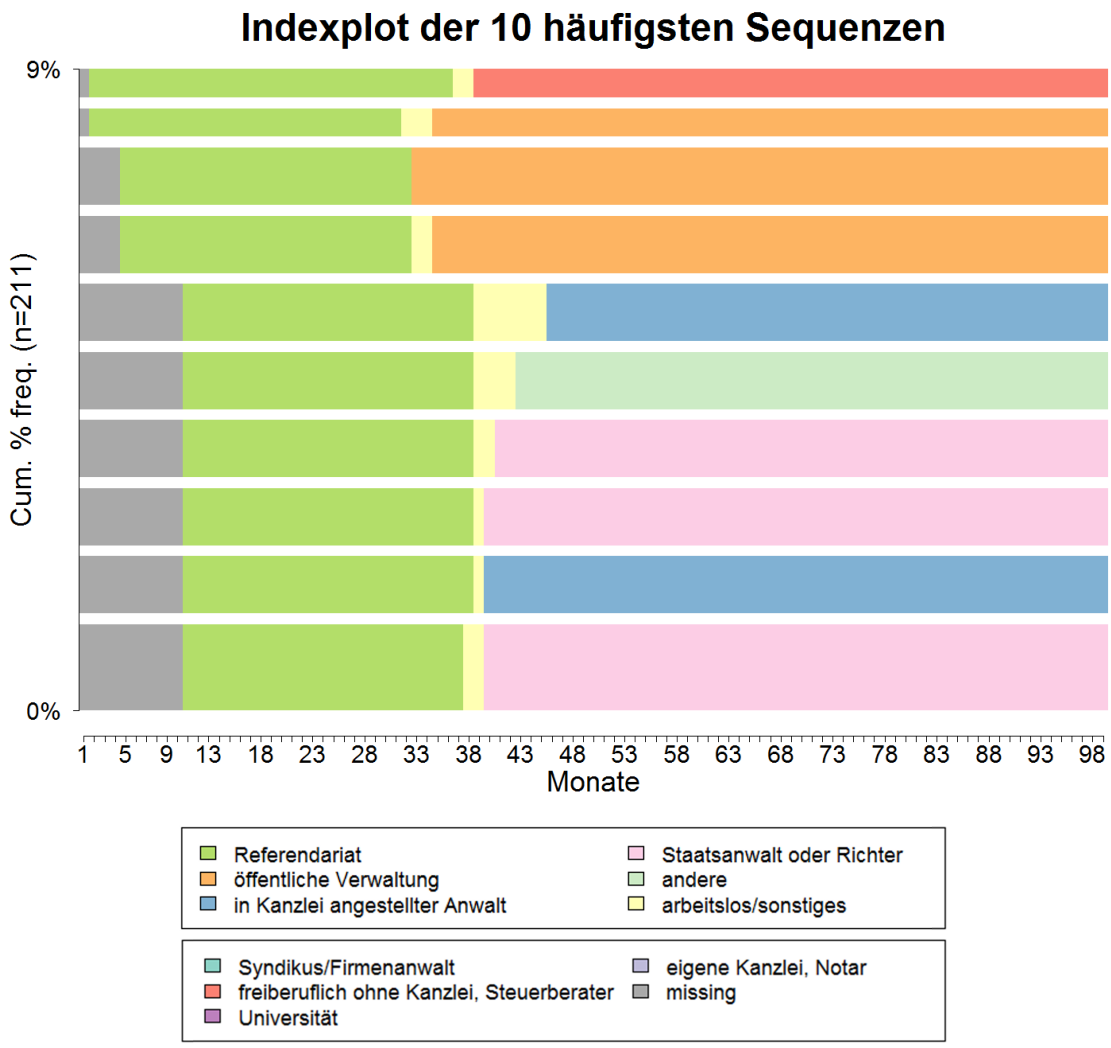

Abbildung 4.2: Die zehn häufigsten Sequenzen.

Zustände *Öffentliche Verwaltung*, *in Kanzlei angestellter Anwalt* und *Staatsanwalt oder Richter* über die Monate nach dem Referendariat den größten Anteil aus. Insgesamt ist jeder der Zustände ab circa dem 30. Monat nach Abschluss vertreten. Diese relative Zustandshäufigkeit kann nun auch nach Gruppen getrennt betrachtet werden. So zeigt Abbildung 4.4 diese einmal für Männer und einmal für Frauen. Wie in Abbildung 4.3 zeigt auch hier die y-Achse die relative Häufigkeits der Zustände an. Diese unterscheiden sich nicht groß von denen für alle Sequenzen, jedoch ist zu erwähnen, dass die *Öffentliche Verwaltung* bei den Frauen einen größeren Teil ausmacht als bei den Männern. Bei den Männern hingegen treten sowohl die Zustände *Staatsanwalt oder Richter* und *in Kanzlei angestellter Anwalt*

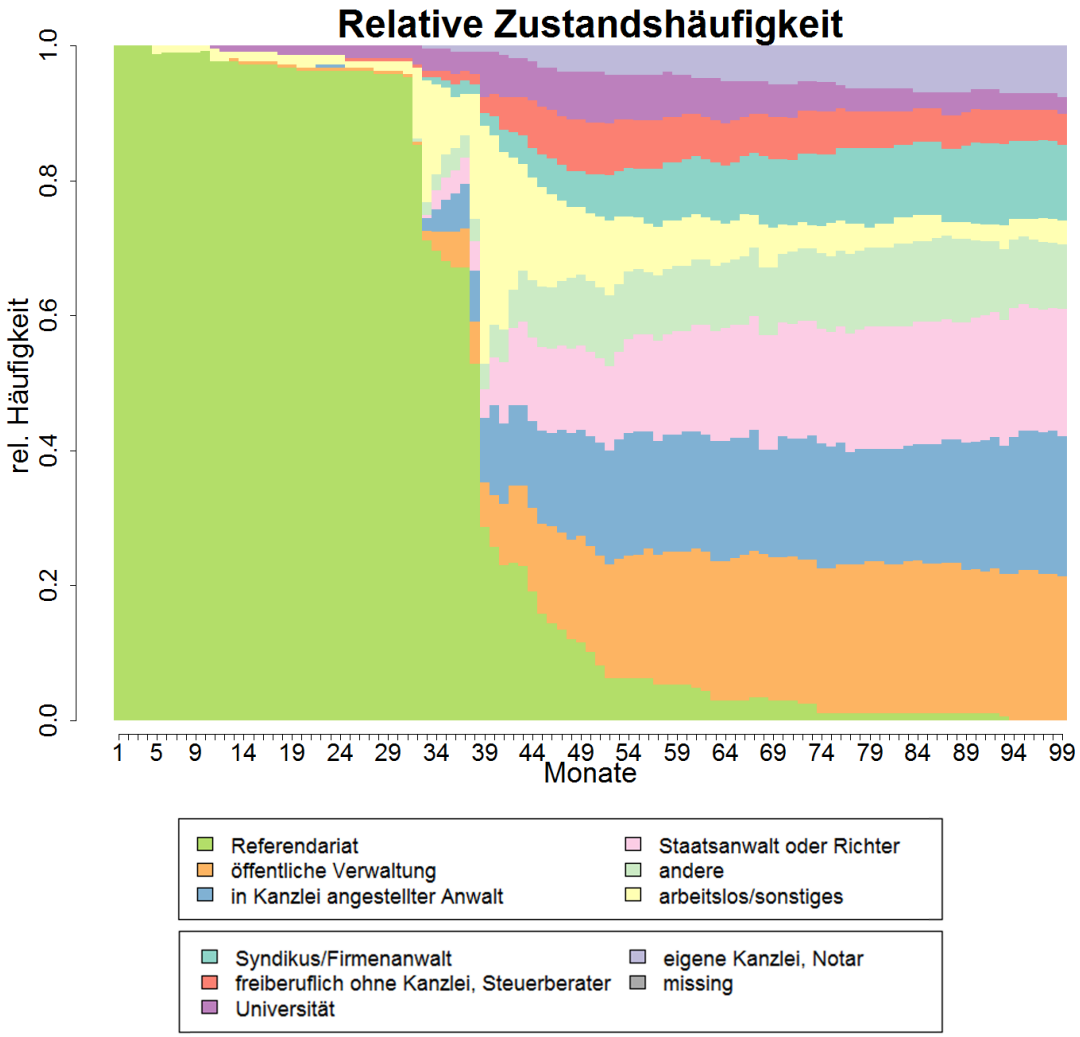

Abbildung 4.3: Relative Zustandshäufigkeit aller Sequenzen

als auch *eigene Kanzlei, Notar* häufiger auf als bei den Berufsverläufen der Frauen. Die relative Zustandshäufigkeit getrennt nach der *Art der Stellenfindung* ist in Abbildung 4.5 zu sehen. Hier fällt vor allem die Gruppe derjenigen Absolventen auf, welche ihre Stelle durch *studiennahe Kontakte* gefunden haben. Sie befinden sich häufiger als alle anderen im Zustand *Universität*, das heißt, sind an der Universität beschäftigt. Bei der *formellen Stellenfindung* überwiegt die *öffentliche Verwaltung*, bei der Gruppe, die ihre Stelle über *persönliche Kontakte* gefunden hat, überwiegt der Zustand *in Kanzlei angestellter Anwalt*. Während bei Stellenfindung über das Referendariat die Zustände *in Kanzlei angestellter Anwalt* und *Staatsanwalt oder Richter* den höchsten Anteil über die Monate hinweg ausma-

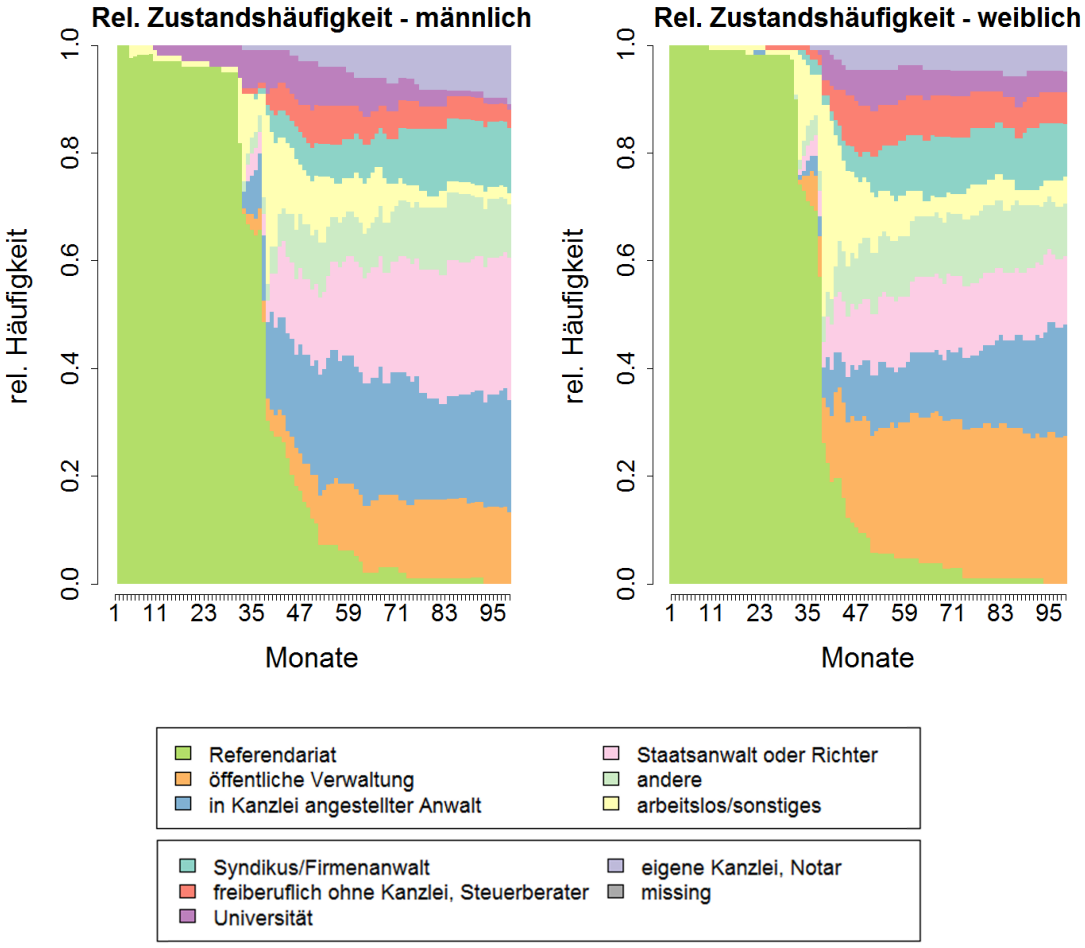

Abbildung 4.4: Relative Zustandshäufigkeit getrennt nach Geschlecht

chen, so ist es bei selbständiger bzw. sonstiger Art der Stellenfindung der Zustand *eigene Kanzlei, Notar*.

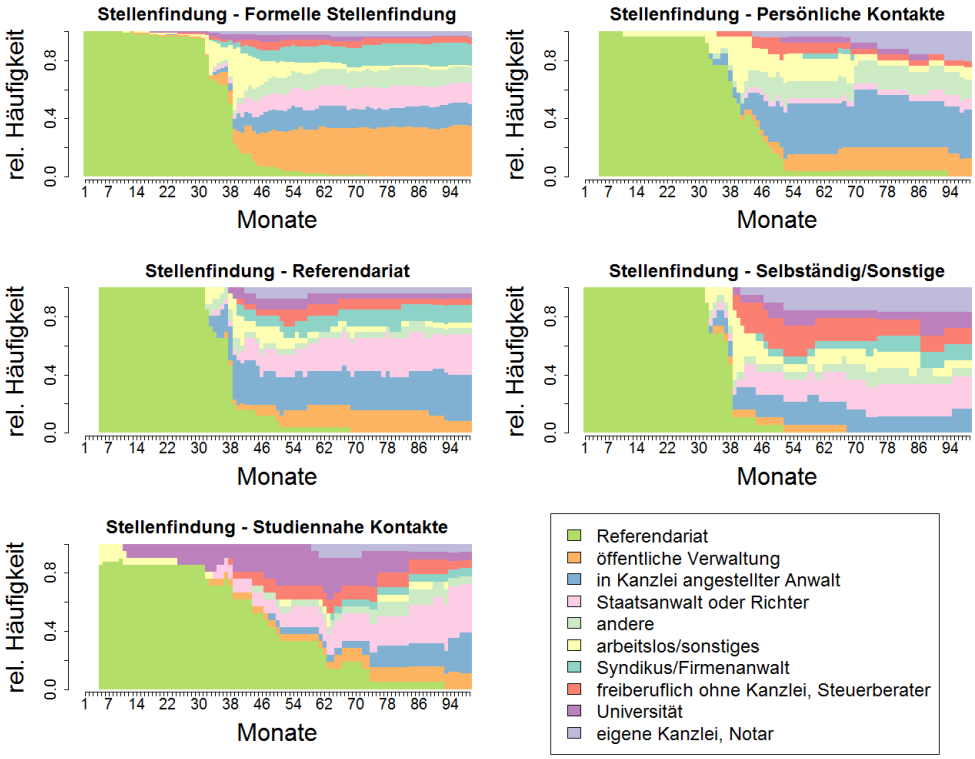

Abbildung 4.5: Relative Zustandshäufigkeit getrennt nach Stellenfindung

4.2 Repräsentative Sequenzen

Wie bereits in Abschnitt 4.1 angesprochen, liegen in dem Fall der Jura-Daten viele Sequenzen vor, die sich zwar prinzipiell ähnlich sind, jedoch unterschiedliche Längen von „Übergangszuständen“ wie zum Beispiel *Arbeitslosigkeit* aufweisen. Von Interesse ist allerdings hauptsächlich die Reihenfolge der Zustände und weniger die exakte Länge jedes einzelnen Zustands. Um diese Sequenzen trotzdem als ähnlich zu erkennen, werden nun die repräsentativen Sequenzen berechnet und das *Neighborhood-Density*-Kriterium zur Berechnung des Repräsentativitäts-Scores gewählt. Als Nachbarschaftsradius wird 20% (tsim = 0.2) der theoretisch maximalen Distanz gewählt, welche bei den vorliegenden Daten 197.7 beträgt. Dadurch werden zwei Sequenzen s_1 und s_2 als redundant angesehen, wenn die Distanz $d(s_1, s_2)$ zwischen ihnen 39.54 oder weniger beträgt. Außerdem wird festgelegt, dass insgesamt sechs (nrep = 6) repräsentative Sequenzen ausgewählt werden sollen. Die Wahl dieser Werte ist sowohl durch inhaltliche und grafische Überlegungen,

als auch durch die Berechnung der *Overall-Quality* für verschiedene Kombinationen aus Nachbarschaftsradius (tsim=$(0.1, 0.15, 0.2, ..., 0.45, 0.5)$) und Anzahl der Repräsentanten (nrep=$(1, 2, ...5, 6)$) entstanden.
Mit diesen Werten und den erhält für die vorliegenden Daten man eine Overall-Quality von 0.2507654 und die sechs repräsentativen Sequenzen wie in Abbildung 4.6 zu sehen. Der *Coverage*-Wert beträgt 62.1%. Dies bedeutet, dass 62.1% aller Sequenzen in der Nach-

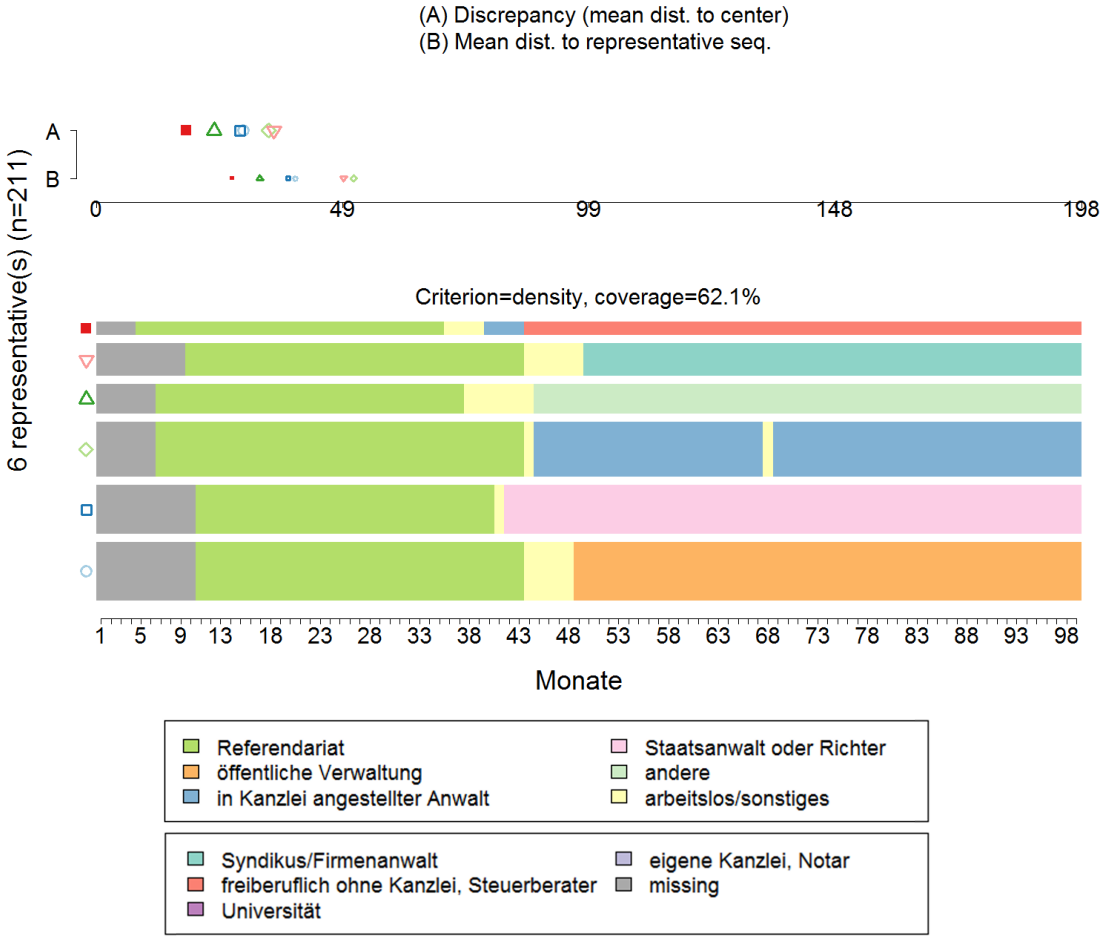

Abbildung 4.6: Sechs repräsentative Sequenzen mit Nachbarschaftsradius 20%und Coverage 62.1%

barschaft dieser sechs Sequenzen liegen, das heißt ähnlich zu diesen sind. Die Breite der einzelnen Sequenzen ergibt sich durch die Anzahl der Sequenzen, die dem jeweiligen Repräsentant zugewiesen werden. Die unterste Sequenz ist die nach dem gewählten Kriterium Repräsentativste, die Oberste die am wenigsten repräsentative. Man erkennt, dass

	na	na(%)	nb	nb(%)	SD	MD	DC	V	Q
r1	52	24.64	37	17.54	2076	39.9	2769	29.5	25.03
r2	43	20.38	27	12.80	1656	38.5	2430	28.9	31.85
r3	49	23.22	24	11.37	2530	51.6	2600	34.6	2.66
r4	26	12.32	18	8.53	854	32.9	1576	23.7	45.80
r5	29	13.74	16	7.58	1440	49.7	1729	35.7	16.70
r6	12	5.69	9	4.27	327	27.2	753	18.1	56.60
Total	211	100.00	131	62.09	8883	42.1	11857	56.2	25.08

Tabelle 4.1: Zu den sechs repräsentativen Sequenzen gehörige Werte und Qualitätsmaße:
na: number of assigned objects
nb: number of objects in the neighborhood
SD: sum of the na distances to the representative
MD: mean of the na distances to the representative
DC: sum of the na distances to the center of the complete set
V: discrepancy of the subset
Q: quality of the representative

die dargestellten Sequenzen einer einfachen Struktur folgen. Nach dem Abschnitt des *Referendariats* zu Beginn, enthält jede Sequenz einen mehr oder weniger kurzen Abschnitt der *Arbeitslosigkeit* bzw. *Sonstiges* und geht anschließend fast ohne weitere Zustandswechsel in einen weiteren Berufsbereich über, der bis zum Ende anhält.
Im oberen Bereich der Grafik sind zu jeder repräsentativen Sequenz die beiden Qualitätsmaße abgetragen, die in Abschnitt 3.4 erklärt wurden. Jeder Sequenz wird ein Symbol zugewiesen, welches den berechneten Wert auf den beiden Achsen A und B kennzeichnet. Achse A zeigt die jeweiligen Werte der mittleren Distanz zum Zentrum und Achse B die mittlere Distanz zur zugehörigen repräsentativen Sequenz.
Außerdem können zusätzlich noch die Anzahl der zugeordneten Sequenzen, die Anzahl der Sequenzen in der Nachbarschaft und verschiedene Qualitätsmaße für jeden einzelnen Repräsentanten ausgegeben werden. Diese sind in Tabelle 4.1 aufgelistet.
Die Grafik der repräsentativen Sequenzen kann zusätzlich nach Variablen getrennt betrachtet werden, wie zum Beispiel in Abbildung 4.7 getrennt nach der Variable *Juristen-Eltern (Ja/Nein)*. Auf der linken Seite der Grafik sind die sechs repräsentativen Sequenzen für Absolventen mit Juristen-Eltern zu sehen, auf der rechten Seite diejenigen der Absolventen ohne Juristen-Eltern. Die Anzahl der gewünschten repräsentativen Sequenzen und der Nachbarschaftsradius sind jeweils wie vorher gewählt. Der Coverage-Wert der repräsentativen Sequenzen beträgt in der linken Grafik 63.3%, in der rechten 59.3%. Unterschiede in

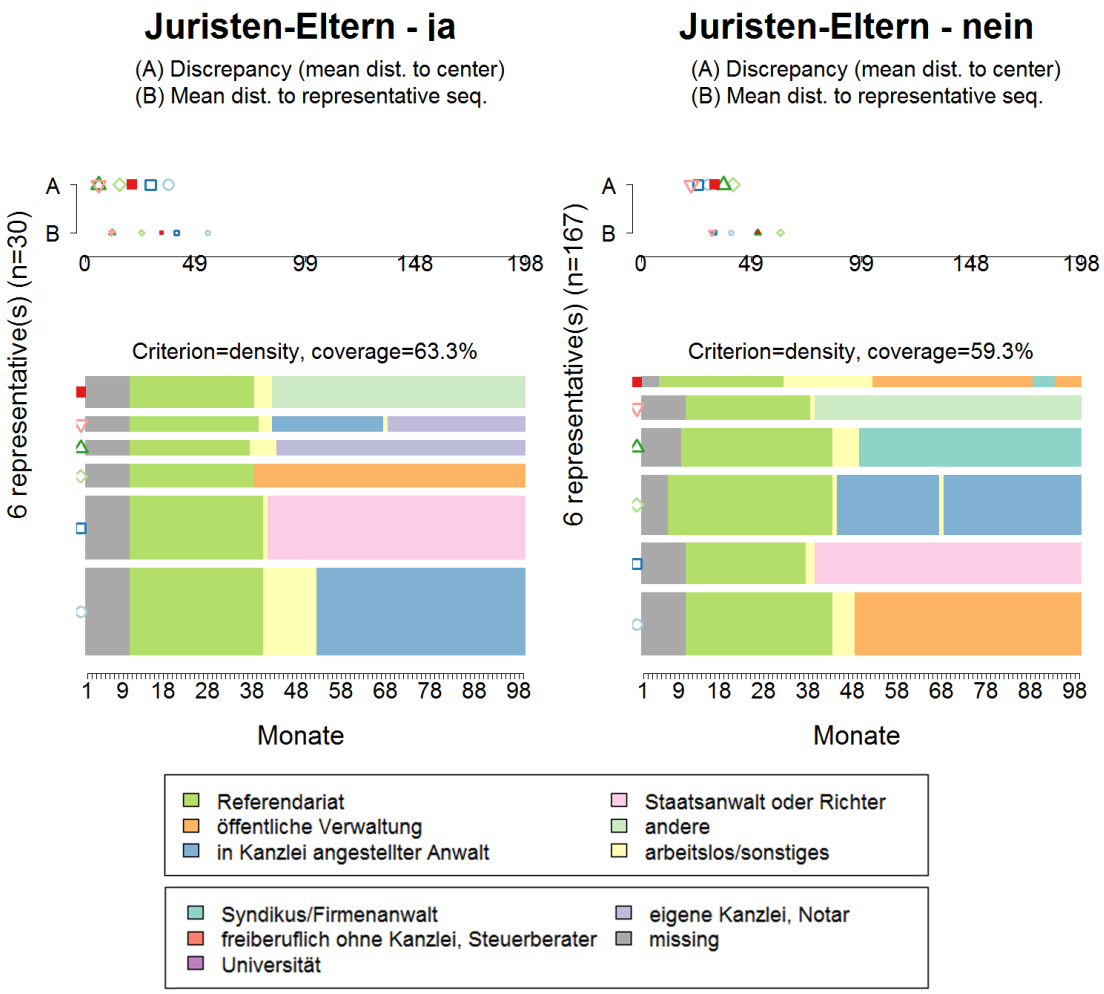

Abbildung 4.7: Sechs repräsentative Sequenzen getrennt nach der Variable *Juristen-Eltern*

den Sequenzen sind vor allem darin zu erkennen, dass in den Sequenzen der Absolventen mit Juristen-Eltern Sequenzen mit langen Abschnitten des Zustands *eigene Kanzlei, Notar* zu finden sind, welche in der rechten Grafik nicht vorhanden sind. Andersherum taucht bei den Absolventen ohne Juristen-Eltern eine Sequenz mit dem langen Abschnitt des Zustands *Syndikus, Firmenanwalt* auf, welche wiederum auf der linken Seite fehlt. Außerdem vertauscht sich die Reihenfolge der ersten und dritten Sequenz der beiden Kategorien, welche auch unterschiedliche Breiten haben.
Zusätzliche Grafiken zu Aufteilungen nach anderen Variablen befinden sich in Abschnitt A.1.

4.3 Clusteranalyse

Nachdem wie in Abschnitt 3.2 und 3.3 beschrieben die Distanzmatrix berechnet und die Clusteranalyse nach der Methode von *Ward* durchgeführt wurde, muss nun die Anzahl der zu verwendenden Cluster festgelegt werden. Aufgrund des *within-between-Kriteriums*, welches in Abschnitt 3.3 eingeführt wurde, und inhaltlichen bzw. grafischen Überlegungen zu den einzelnen Clustern mit verschiedenen Lösungen zur Clusteranzahl, wird in diesem Fall eine Clusterlösung mit sieben Clustern gewählt. In Abbildung 4.8 sind die einzelnen Quotienten des Kriteriums zu allen Clusterlösungen zwischen 2 und 15 Clustern, als auch die 0.5-Marke zur Entscheidung abzulesen.

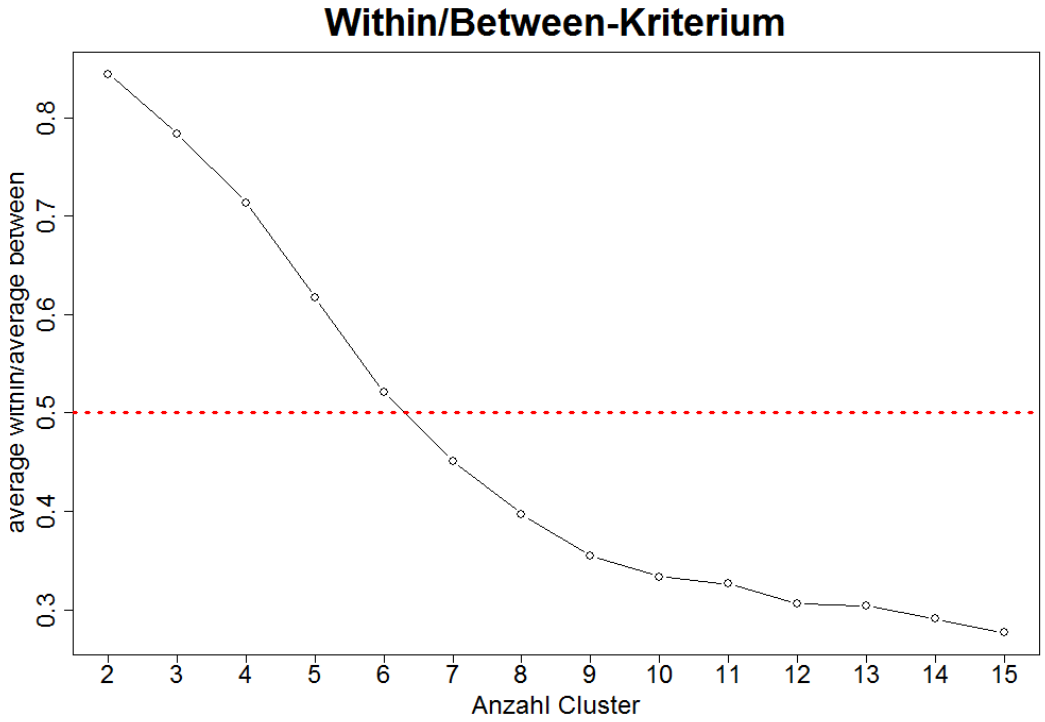

Abbildung 4.8: within-between-Kriterim zur Wahl der Cluster-Anzahl

Durch diese Analysen und die beschriebene Wahl der Clusteranzahl ergeben sich nun die in Abbildung 4.9 abgebildeten sieben Cluster, wobei hier jeweils die relative Zustandshäufigkeit der Sequenzen abgetragen ist.

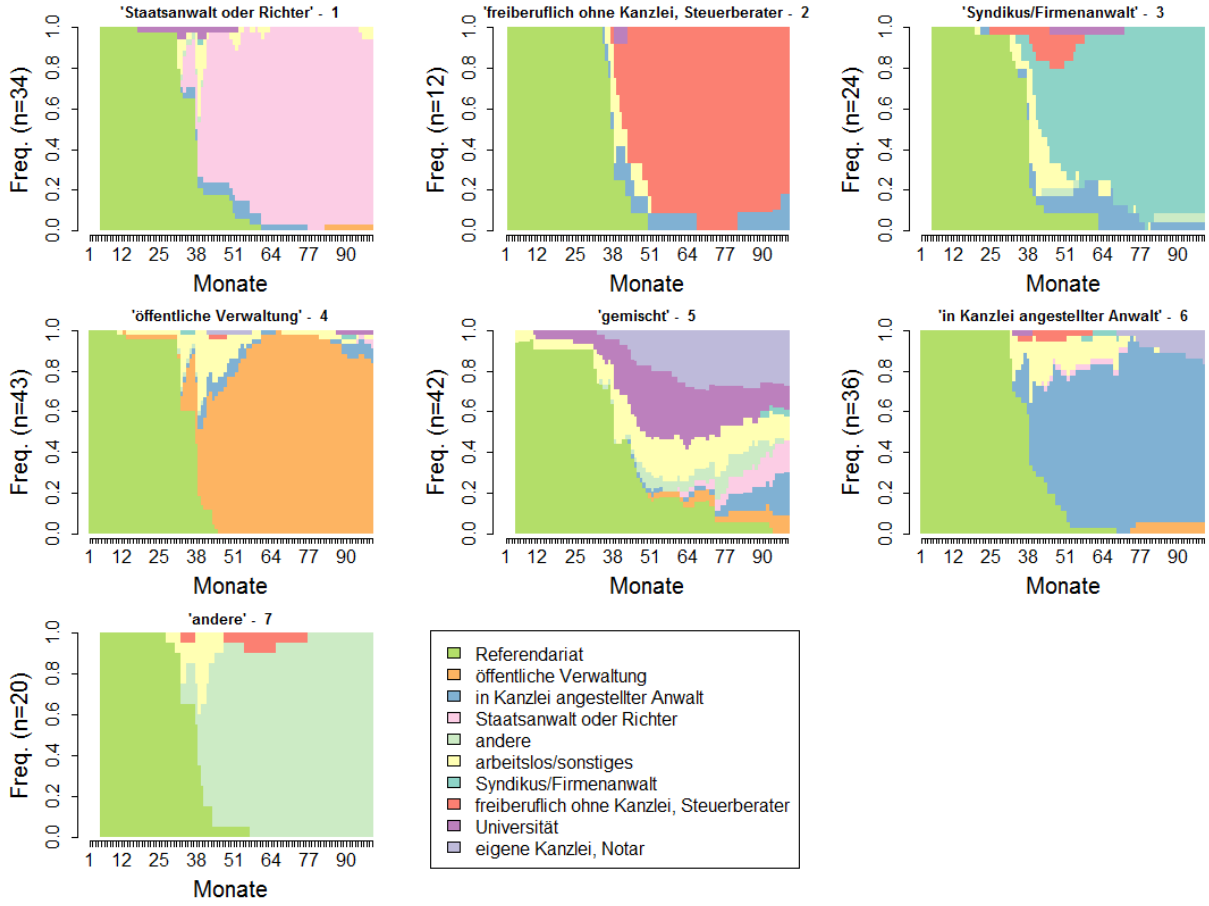

Abbildung 4.9: Relative Zustandshäufigkeit der sieben Cluster

Man erkennt in allen Clustern außer dem Fünften, dass zu Beginn der ca. zweijährige Bereich des *Referendariats* liegt, dem ein großer Anteil jeweils einer Farbe, das heißt einem speziellen Berufsbereich, folgt. Andere Berufsbereiche sind hierbei nur mit sehr kleinen Anteilen vertreten. Die einzelnen Cluster wurden nun auch nach der Berufsgruppe mit dem größten Anteil benannt. Die einzige Ausnahme bildet *Cluster 5*, der statt eines großen Anteil einer Berufsgruppe viele kleinere Anteile verschiedener Bereiche enthält. Dies bedeutet, dass hier noch viele unterschiedliche Sequenzen vorliegen. Hätte man eine Clusterlösung mit zum Beispiel acht oder neun Clustern gewählt, hätte sich dieser Cluster noch weiter in die einzelnen Bereiche aufgeteilt, die anderen sechs Cluster wären jedoch gleich geblieben. Da dies allerdings zu sehr kleinen Fallzahlen geführt hätte, wurde diese Lösung verworfen. (Grafiken zu alternativen Clusterlösungen sind in Abschnitt A.1 zu finden.) Letztendlich erhält man schließlich die sieben, nach dem größten Anteil benannten

Cluster

- 1 - ‚Staatsanwalt oder Richter'
- 2 - ‚freiberuflich ohne Kanzlei, Steuerberater'
- 3 - ‚Syndikus/Firmenanwalt'
- 4 - ‚öffentliche Verwaltung'
- 5 - ‚gemischt'
- 6 - ‚in Kanzlei angestellter Anwalt'
- 7 - ‚andere'

Die Fallzahlen der Cluster liegen zwischen $n = 12$ und $n = 43$.
Auch zu den sieben Clustern können jeweils die repräsentativen Sequenzen dargestellt werden, welche in Abbildung 4.10 zu sehen sind.

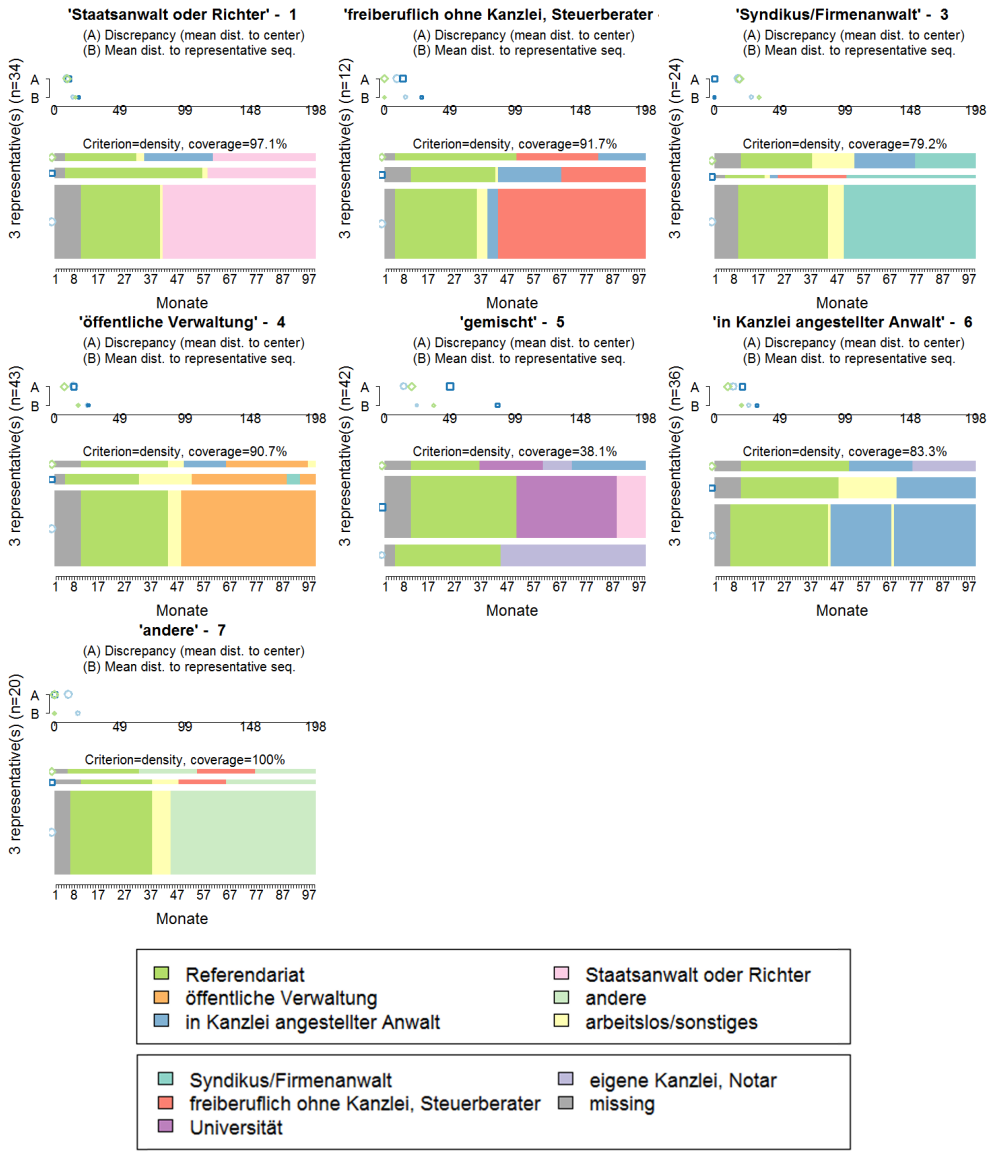

Abbildung 4.10: Repräsentative Sequenzen der sieben Cluster

Hier wurde erneut ein Nachbarschaftsradius von 20% gewählt, allerdings die Anzahl der gewünschten Sequenzen auf drei gesetzt, da hier aufgrund der Cluster bereits recht ähnliche Sequenzen vorliegen sollten und man nicht erwartet, viele sehr unterschiedliche Sequenzen zu erhalten. Es ist gut zu erkennen, dass bis auf *Cluster 5*, in allen Clustern die erste (repräsentativste) Sequenz sehr breit ist und die zu erwartende Abfolge der Zustände enthält. Aber auch die zweite und dritte Sequenz hat jeweils eine zur ersten Sequenz ähnlichen Struktur: der „Hauptzustand“ des jeweiligen Clusters ist immer enthalten. Der

gemischte Cluster unterscheidet sich von den anderen in dieser Hinsicht, da hier viele unterschiedliche Sequenzen enthalten sind. Dies ist auch durch den im Vergleich niedrigen Coverage-Wert von 38.1% zu erkennen. Die Coverage-Werte der einzelnen Sets von repräsentativen Sequenzen der anderen Cluster liegen zwischen 97.2% und 100%. Diese hohen Werte ergeben sich dadurch, dass die Sequenzen bereits geclustert wurden, d.h. bereits ähnliche Sequenzen zueinander sortiert wurden.

Zusätzlich werden die sieben Cluster deskriptiv betrachtet, um einen Eindruck dafür zu bekommen, welche Absolventen mit welchen Merkmalen den einzelnen Clustern zugeordnet wurden.

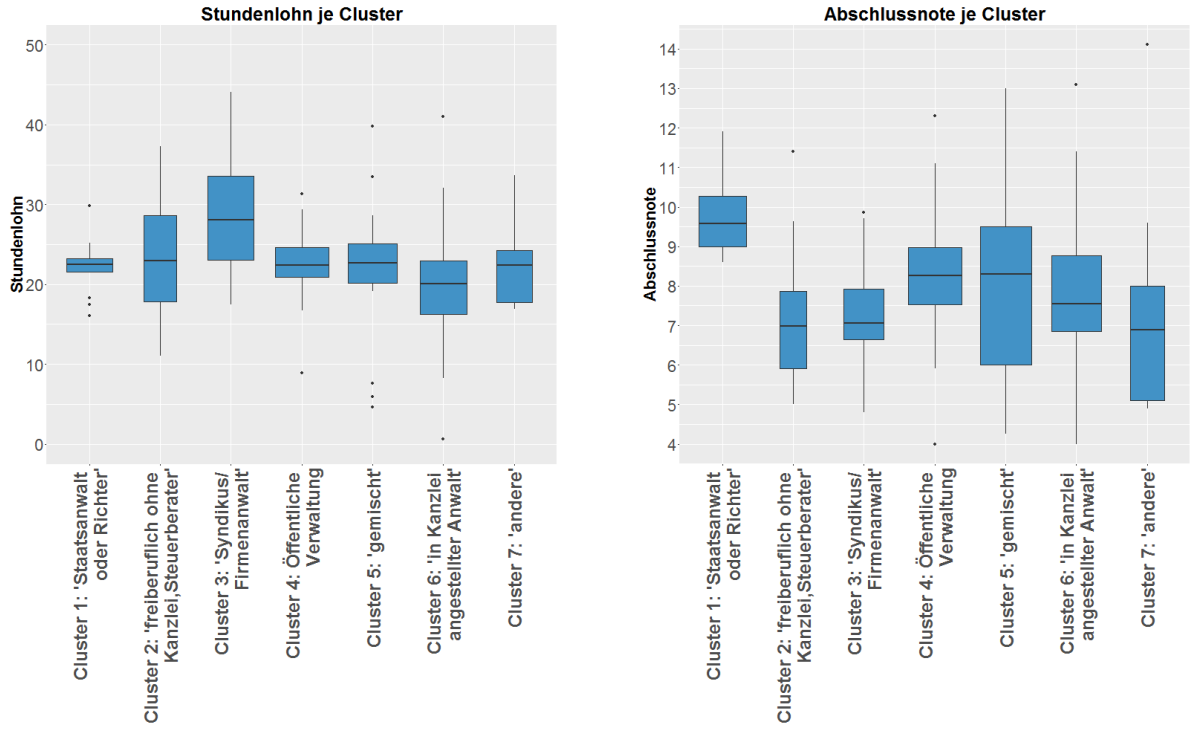

Abbildung 4.11: Boxplots zu Stundenlohn und Abschlussnote je Cluster

In Abbildung 4.11 sind auf der linken Seite Boxplots der Variable *Stundenlohn*, auf der rechten die der Variable *Abschlussnote* jeweils für alle Cluster zu sehen. Die Whiskers der Boxplots sind jeweils als das 1.5-fache des Interquartilsabstands (IQR) definiert. Es ist zu erkennen, dass beispielsweise *Cluster 3* (Syndikus/Firmenanwalt) mit einem Median von 28.05 den höchsten Stundenlohn und *Cluster 1* (Staatsanwalt oder Richter) mit einem

IQR von 1.7 die geringste Streuung des Stundenlohns aufweist. Die Abschlussnote ist durchschnittlich in *Cluster 1* (Staatsanwalt oder Richter) am besten und in *Cluster 7* (andere) am schlechtesten. Die Mediane in diesen beiden Clustern liegen bei 9.59 bzw. 6.88.

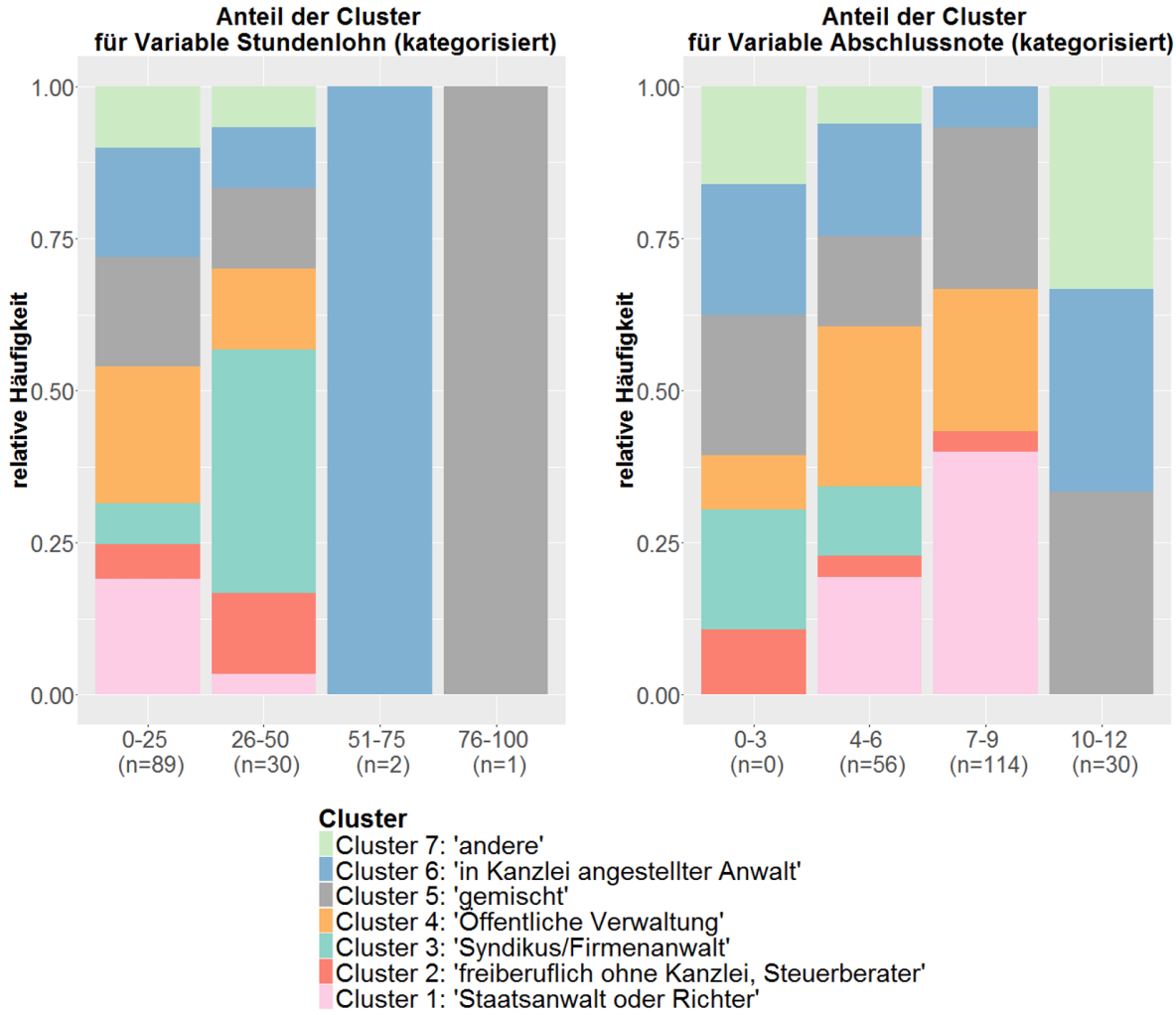

Abbildung 4.12: Gestapelte Balkendiagramme zu den Anteilen der Cluster in den Kategorien der kategorisierten Variablen Stundenlohn und Note

Abbildung 4.12 zeigt zusätzlich gestapelte Balkendiagramme für die kategorisierten Variablen *Stundenlohn* und *Abschlussnote*, wobei jeweils der Anteil der sieben Cluster pro Kategorie abgebildet wird. Da zum besseren Vergleich die relativen Häufigkeiten dargestellt sind, sollte hier allerdings die Fallzahl der einzelnen Kategorien beachtet werden. In Abbildung 4.13 werden die normierten Balkendiagramme für die Kategorien der Varia-

blen *Juristen-Eltern* und *Promotion* dargestellt, welche die Anteile der einzelnen Clustern in den Kategorien veranschaulichen.

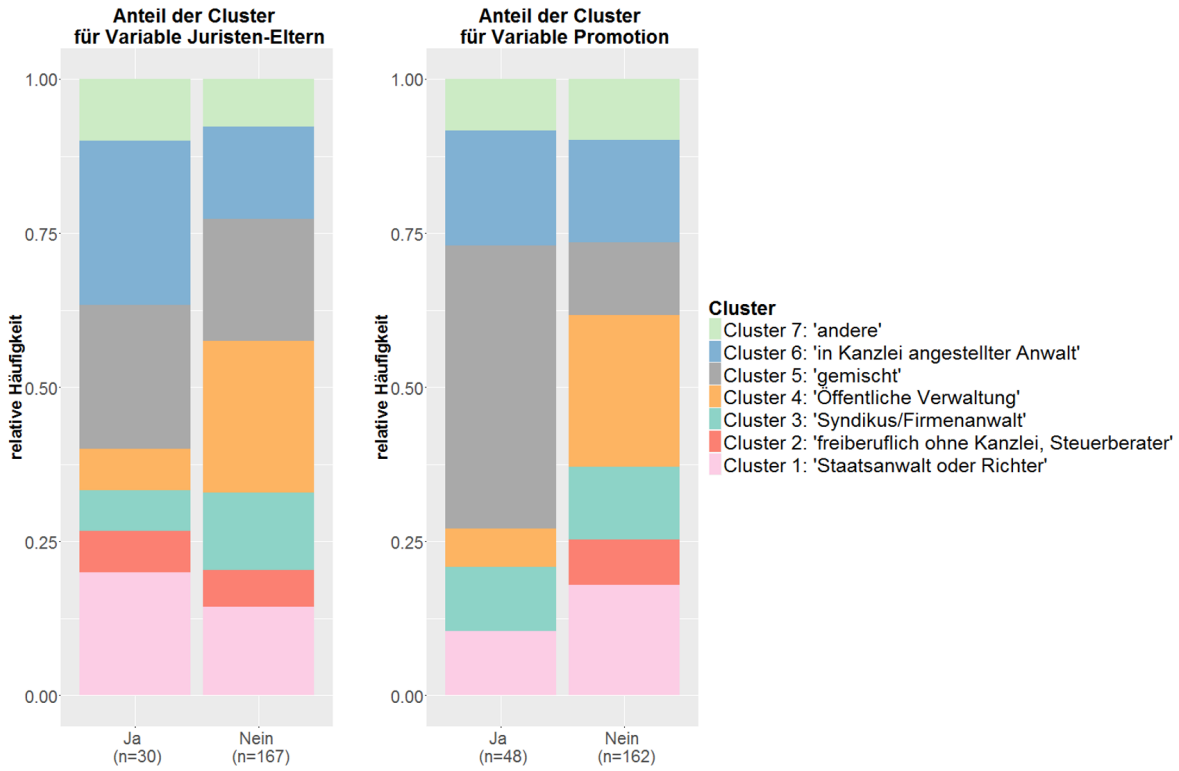

Abbildung 4.13: Gestapelte Balkendiagramme zu den Anteilen der Cluster in den Kategorien der Variablen Juristen-Eltern und Promotion

Auch hier ist die Fallzahl je Kategorie zu beachten. Für die Variable *Juristen-Eltern* fallen beispielsweise 167 Absolventen in die Kategorie *Nein*, haben also keine Juristen-Eltern, aber nur 30 Absolventen in Kategorie *Ja* und haben somit Juristen-Eltern. In der linken Grafik machen *Cluster 5* und *6* den größten Anteil der Kategorie *Ja* aus, *Cluster 4*, *3* und *2* allerdings sind nur mit einem kleinen Anteil vertreten. In Kategorie *Nein* hingegen beanspruchen *Cluster 4* und *5* den größten Anteil, *Cluster 2* und *7* sind wenig vertreten. Darüber hinaus ist in der Grafik zur Variable *Promotion* zu erkennen, dass insgesamt 48 Absolventen ihre Promotion abgeschlossen haben und den größten Anteil in dieser Kategorie *Cluster 4* ausmacht und *Cluster 2* gar nicht vorhanden ist. In der Kategorie *Nein*, d.h. Absolventen ohne abgeschlossene Promotion ($n = 162$), steigen im Vergleich die Anteile der *Cluster 1* und *4*.

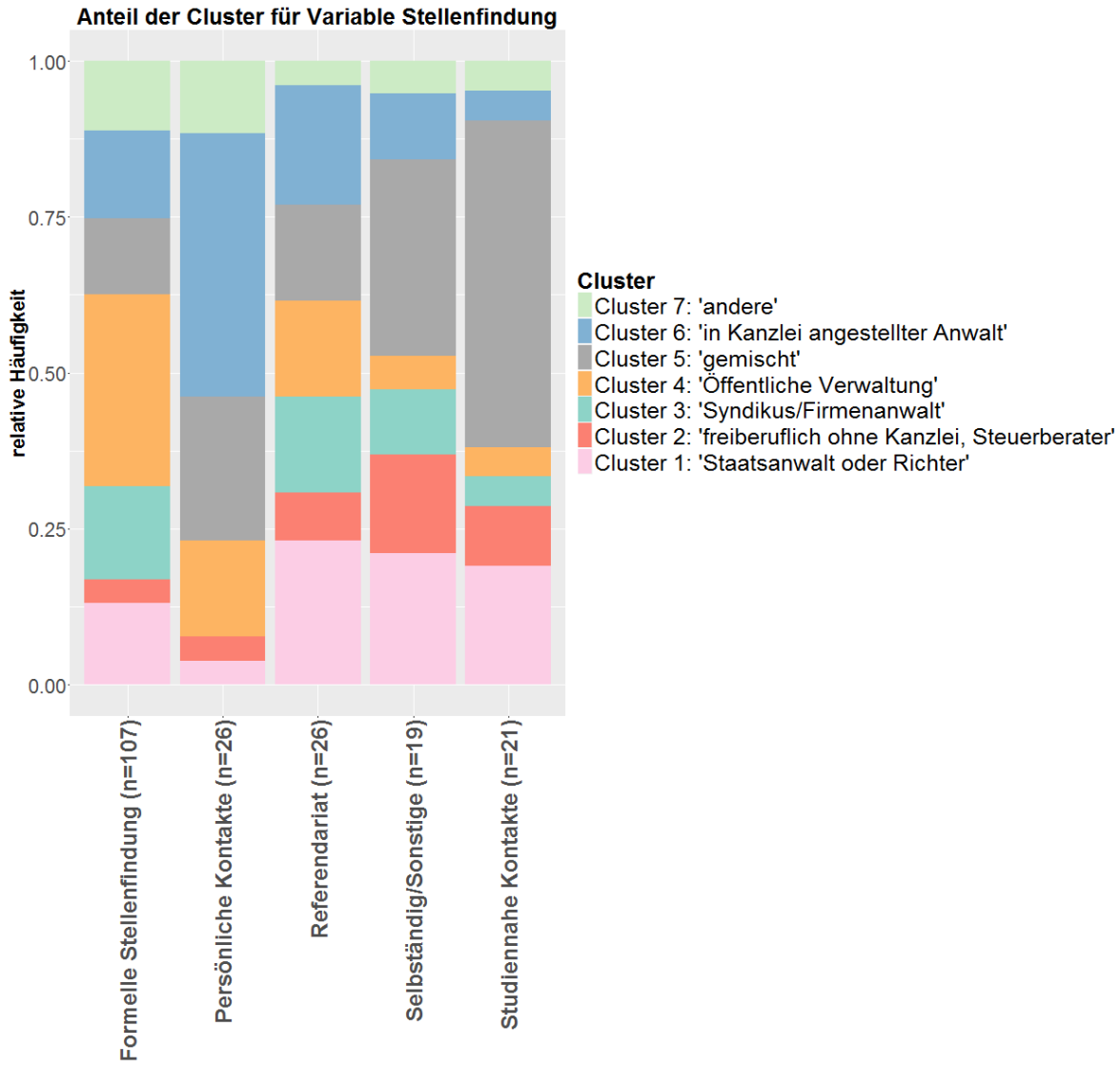

Abbildung 4.14: Gestapelte Balkendiagramme zu den Anteilen der Cluster in den Kategorien der Variable Art der Stellenfindung

Für die Variable *Art der Stellenfindung* sind die gestapelten Balkendiagramme zu allen fünf Kategorien in Abbildung 4.14 zu sehen. In der Kategorie *Formelle Stellenfindung* wird der größte Anteil durch *Cluster 4* (Öffentliche Verwaltung) repräsentiert, in der Kategorie *Persönliche Kontakte* durch *Cluster 6* (in Kanzlei angestellter Anwalt). Die Kategorie *Referendariat* hat nur kleine Anteile von *Cluster 2* und *7*, der Rest ist relativ ausgeglichen. In den Kategorien *Selbstständig/Sonstige* und *Studiennahe Kontakte* hingegen dominieren vor allem die *Cluster 5* (gemischt) und *1* (Staatsanwalt oder Richter).

Weitere deskriptive Grafiken zu den sieben Clustern sind in Abschnitt A.1 zu finden.

4.4 Regression

Die in Abschnitt 3.5 vorgestellten Regressionsmodelle werden nun auf die vorliegenden Daten angewendet und die resultierenden Ergebnisse im folgenden Kapitel aufgeführt.

4.4.1 Clusterregression

Zunächst wird ein multinomiales Logit-Modell angewendet, um zu untersuchen, welche Kovariablen einen (signifikanten) Einfluss auf die Clusterzugehörigkeit haben. Hierbei wird die Variable der durch die Clusteranalyse entstandenen Clusterzugehörigkeit als Zielvariable verwendet ($y_i \in \{1, \ldots, 7\}$), wobei *Cluster 4* als Referenzkategorie gewählt wird, da dieser mit $n = 43$ der größte Cluster ist. Als Einflussgrößen werden folgende Variablen verwendet und auf Wunsch der Projektpartner zunächst das volle Modell mit diesen berechnet:

- Geschlecht (männlich/weiblich)
- Juristen-Eltern (Nein/Ja)
- Note (metrisch)
- Studiums-Kontakte (Anzahl)
- Privat-Kontakte (Anzahl)
- Art der Stellenfindung (Formelle Stellenfindung/Persönliche Kontakte/Referendariat/(Selbständig/Sonstige)/Studiennahe Kontakte)

Bei den kategorialen Einflussgrößen wird jeweils die erste aufgelistete Kategorie als Referenz verwendet.
In Abbildung 4.15 sind die geschätzten Regressionskoeffizienten $\hat{\beta}_{ij}$, ($i = \{1, 2, 3, 5, 6, 7\} \mathrel{\widehat{=}}$ Cluster, $j = 1, \ldots, 6 \mathrel{\widehat{=}}$ Kovariablen) des Modells inklusive dem 95%-Konfidenzintervall

abgetragen. Zu einem Signifikanzniveau von $\alpha = 0.05$ gilt der Einfluss einer Variable als signifikant, wenn die 0 nicht im 95%-Konfidenzintervall enthalten ist. Grafisch erkennt man dies daran, dass das Intervall die Nulllinie nicht schneidet. Die Fallzahl von $n = 141$ erklärt sich durch Beobachtungen mit fehlenden Werten, die für die Berechnung des Modells nicht berücksichtigt wurden. In Tabelle 4.2 sind die genauen Koeffizienten des Modells abzulesen.

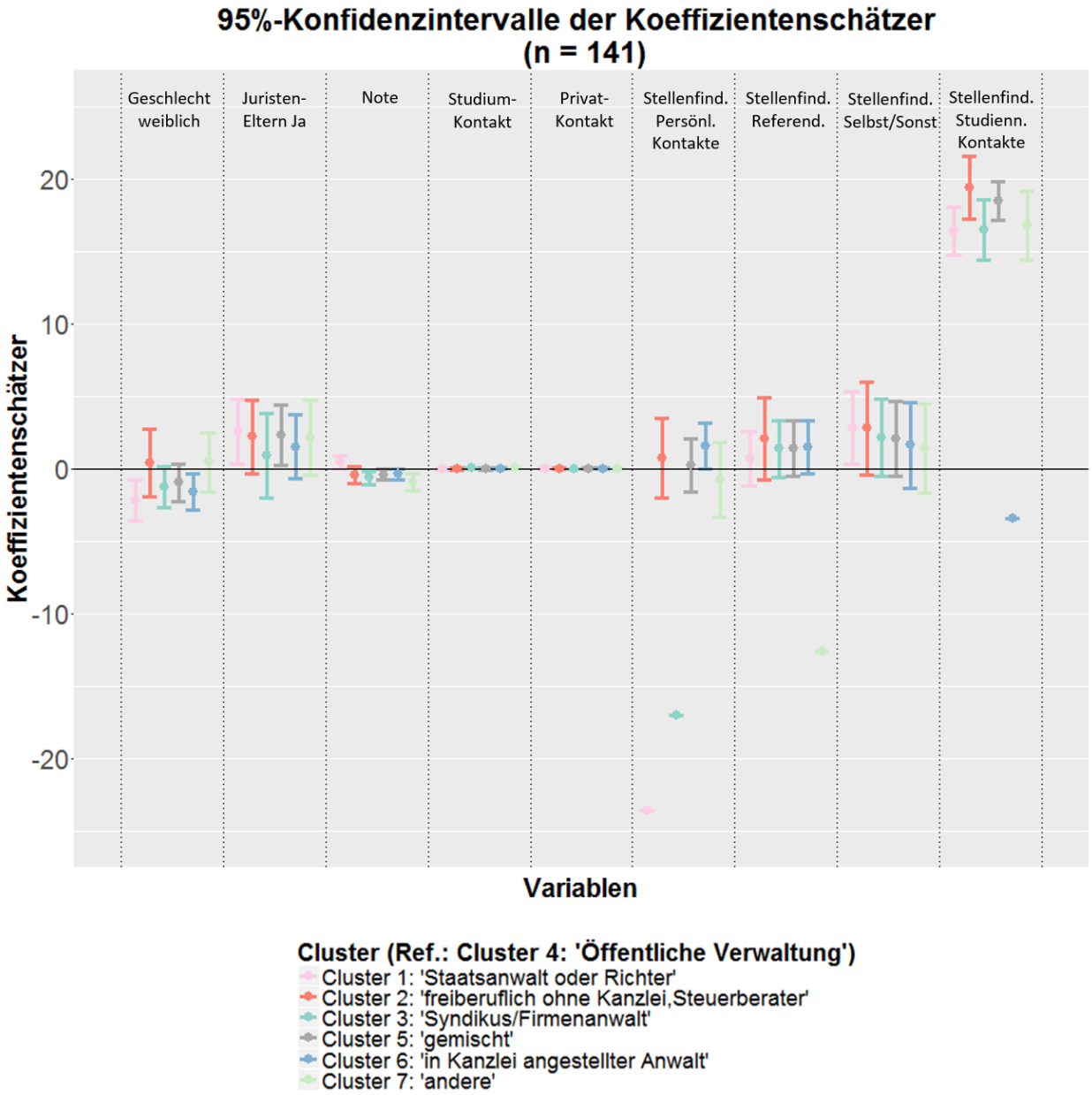

Abbildung 4.15: Regressionskoeffizienten des multinomialen Modells zur Clusteranalyse mit 95% - Konfidenzintervall

In Abbildung 4.15 ist deutlich zu erkennen, dass einige Koeffizienten der Variable *Art der Stellenfindung* sehr extreme Werte haben und weit von der Nulllinie entfernt liegen. Auf den ersten Blick würde man erwarten, dass diese Kategorien der Variable einen signifikanten Einfluss auf die Clusterzugehörigkeit haben. Diese Extremwerte ergeben sich in diesem

Cluster	(Intercept)	Geschlechtweiblich	juElternja	Note	studiumkont.
1	-3.62	-2.19	2.55	0.45	-0.04
2	0.76	0.39	2.18	-0.47	-0.02
3	5.17	-1.28	0.89	-0.64	0.01
5	3.24	-0.98	2.30	-0.42	-0.03
6	2.81	-1.60	1.48	-0.40	-0.01
7	4.81	0.44	2.15	-0.93	0.02

Cluster	privatkont.	stelf.Pers. Kont.	stelf.Ref.	stelf.Selbst/Sonst	stelf.Stu.Kont.
1	-0.00	-23.61	0.68	2.83	16.40
2	-0.00	0.74	2.03	2.76	19.40
3	-0.03	-17.03	1.35	2.13	16.45
5	-0.01	0.22	1.37	2.09	18.44
6	0.01	1.57	1.45	1.61	-3.43
7	-0.00	-0.77	-12.61	1.39	16.77

Tabelle 4.2: Regressionskoeffizienten des multinomialen Logit-Modells zur Clusterzugehörigkeit

Fall jedoch dadurch, dass die insgesamt $n = 141$ Beobachtungen zwischen den sieben Clustern und den jeweiligen Kategorien der Variablen aufgeteilt werden und in einigen Kategorien somit keine Beobachtungen vorliegen, wie in Tabelle 4.3 zu sehen ist. Das führt dazu, dass für diese Kategorien der Variable kein Maximum-Likelihood-Schätzer, der für die Angabe des Regressionskoeffizienten benötigt wird, berechnet werden kann. (Allison (2008))

Allison (2008) empfiehlt in diesem Fall, das geschätzte Modell so beizubehalten, jedoch zusätzlich einen Likelihood-Ratio-Test (LR-Test) für die problematische Variable durchzuführen. Dieser testet die Nullhypothese $H_0 : \hat{\boldsymbol{\beta}}_j = \mathbf{0}$. Demnach hat die zugehörige Variable keinen Einfluss auf die Zielgröße, wenn die Nullhypothese angenommen wird.

Im vorliegenden Fall ergibt der LR-Test zwischen dem Modell mit der Variable *Art der Stellenfindung* und dem Modell ohne diese Variable einen p-Wert von 0.00018. Die Nullhypothese wird zum Signifikanzniveau von $\alpha = 0.05$ abgelehnt. Es kann also nicht davon ausgegangen werden, dass die Variable keinen Einfluss auf die Clusterzugehörigkeit hat.

Eine andere Möglichkeit, dieses Problem zu lösen, wäre nach Allison (2008) einzelne Kategorien der problematischen Variablen zusammenzufassen. Diese Lösung wird hier allerdings aus Gründen der Konsistenz nicht angewendet, da das gleiche Problem auch in dem multinomialen Logit-Modell in Unterabschnitt 4.4.2 auftritt und dort unter anderem die binäre Variable *Juristen-Eltern* betroffen ist. Da diese Variable jedoch nur die

Kategorien *Ja* und *Nein* enthält, ist hier keine Zusammenlegung der Kategorien möglich. Bei den Variablen *Studiumkontakt* und *Privatkontakt* schneiden alle Konfidenzintervalle die

	Art der Stellenfindung				
Cluster	Form. Stellenf.	Pers. Kont.	Referend.	Selbst./Sonst.	Stud. Kont.
4	30	4	3	1	**0**
1	11	**0**	4	4	4
2	2	1	1	1	2
3	10	**0**	3	2	1
5	10	3	3	2	7
6	9	8	4	1	**0**
7	7	1	**0**	1	1

Tabelle 4.3: Kreuztabelle zwischen Clusterzugehörigkeit und *Art der Stellenfindung*

Nulllinie, was darauf hindeutet, dass diese Variablen keinen signifikanten Einfluss auf die Clusterzugehörigkeit haben (siehe auch Abbildung A.17 und A.18 in Abschnitt A.1).
Bei den Variablen *Geschlecht*, *Juristen-Eltern* und *Note* hingegen sind signifikante Unterschiede zu erkennen. Zur besseren Interpretation werden die Konfidenzintervalle zu diesen Variablen in Abbildung 4.16 und 4.17 zusätzlich separat dargestellt.

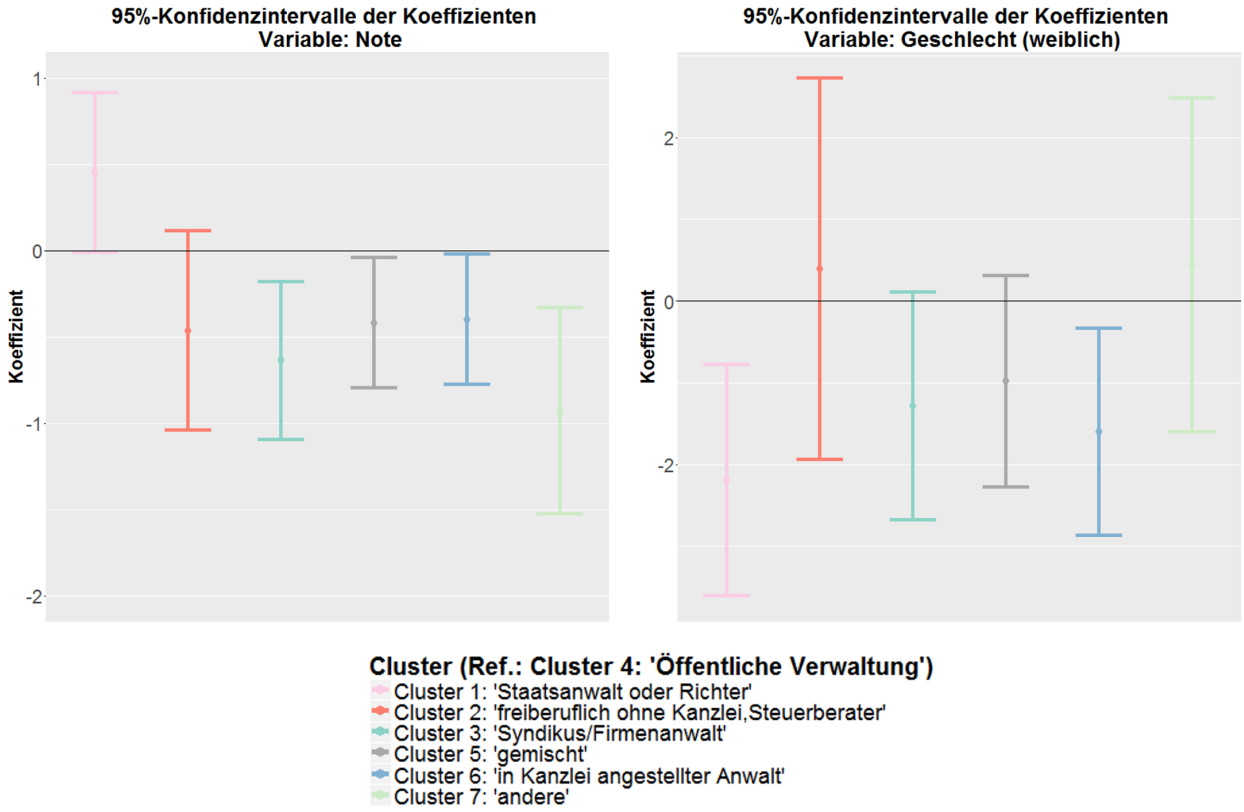

Abbildung 4.16: Regressionskoeffizienten der Variablen *Note* und *Geschlecht* des multinomialen Modells zur Clusteranalyse mit 95% - Konfidenzintervall

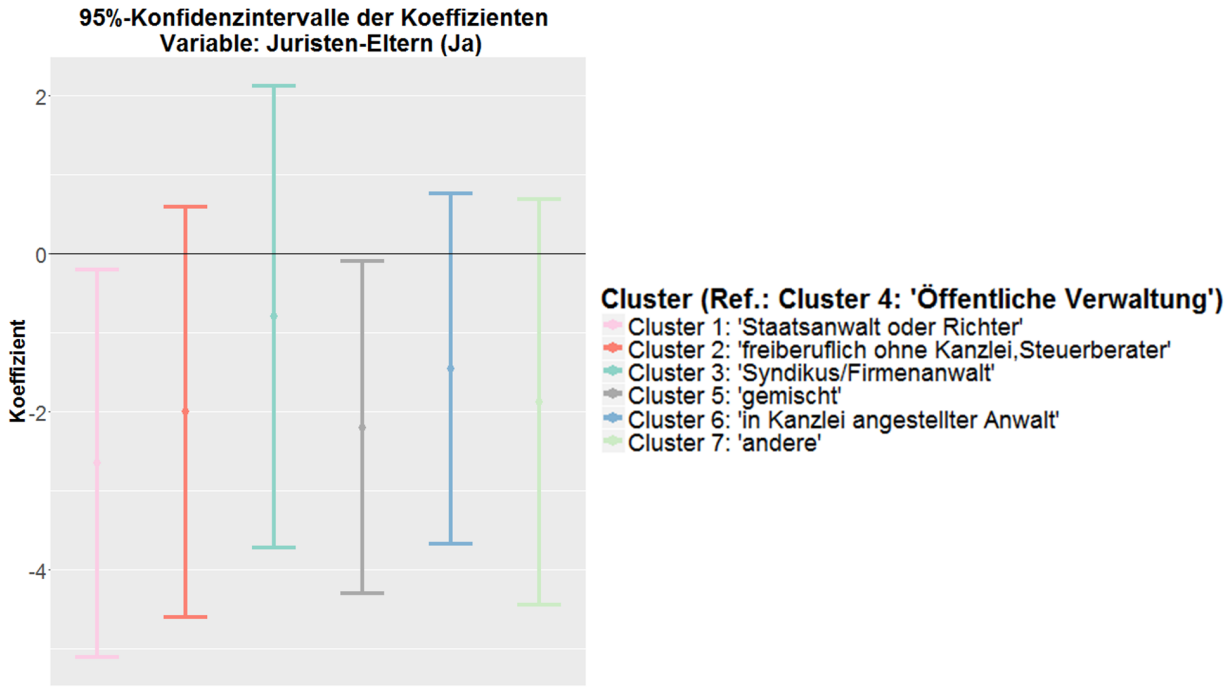

Abbildung 4.17: Regressionskoeffizienten der Variable *Juristen-Eltern* des multinomialen Modells zur Clusteranalyse mit 95% -Konfidenzintervall

Für alle Cluster, deren Konfidenzintervalle für die einzelnen Variablen die Nulllinie nicht schneiden, liegt ein signifikanter Einfluss der jeweiligen Variable auf die Clusterzugehörigkeit vor. Dies betrifft für die Variable *Note* die Cluster 3, 5, 6 und 7, für die Variable *Geschlecht* die Cluster 1 und 6 und für die Variable *Juristen-Eltern* die Cluster 1 und 5. Um die erhaltenen Regressionkoeffizienten zu interpretieren, würde man nun zum Beispiel für die Variable *Note* sagen, dass die Chance, in Cluster 3 (bzw. 5/6/7) statt in Cluster 4 (Referenzcluster) zugeordnet zu werden, multiplikativ um $exp(\hat{\beta}_{3,Note}) = exp(-0.64) = 0.527$ sinkt, wenn die Note um eine Einheit steigt und alle anderen Kovariablen konstant bleiben. Für die Variable *Juristen-Eltern* hingegen würde man sagen, dass sich die Chance in Cluster 1 (bzw. 5) statt in Cluster 4 zugeordnet zu werden für Absolventen mit Juristen-Eltern im Vergleich zu Absolventen ohne Juristen-Eltern multiplikativ um $exp(\hat{\beta}_{1,Juristen-Eltern}) = exp(2.55) = 12.807$ erhöht, wenn alle anderen Kovariablen konstant bleiben.

In Abschnitt A.1 ist als alternative Darstellung (Abbildung A.19) für die Regressionskoeffizienten der sogenannte Stern-Plot zu finden, der das Verhältnis zwischen den Clustern

besser darstellt, jedoch keine Rückschlüsse auf die genauen Schätzwerte oder Signifikanz zulässt.

Zusätzlich zu dem von den Projektpartnern gewünschten vollen Modell wird ein durch Variablenselektion reduziertes Modell geschätzt. Hierzu wird Rückwärtsselektion mit AIC-Kriterium angewendet und dadurch ein Modell erhalten, welches nur noch die Kovariablen *Geschlecht*, *Note* und *Art der Stellenfindung* beinhaltet. Die zugehörigen Regressionskoeffizienten sind in Tabelle 4.4 aufgelistet. Wie auch in Abbildung 4.18 zu erkennen,

	(Intercept)	Geschl.weib.	Note
1	-4.16	-1.96	0.48
2	-0.29	0.63	-0.37
3	4.17	-1.17	-0.58
5	2.14	-0.74	-0.34
6	2.32	-1.59	-0.32
7	4.49	0.29	-0.83

	stelf.Persönliche K.	stelf.Ref.	stelf.Selbst/Sonst	stelf.Studiennahe.K.
1	-24.56	0.66	2.77	18.56
2	1.22	1.99	2.76	21.65
3	-17.87	1.27	2.13	18.53
5	0.74	1.23	2.01	20.49
6	1.85	1.44	1.62	-5.71
7	-0.46	-12.71	1.69	19.62

Tabelle 4.4: Regressionskoeffizienten des reduzierten multinomialen Logit-Modells zur Clusterzugehörigkeit

ergeben sich dadurch keine großen Veränderungen für die Regressionskoeffizienten der verbliebenen Variablen.

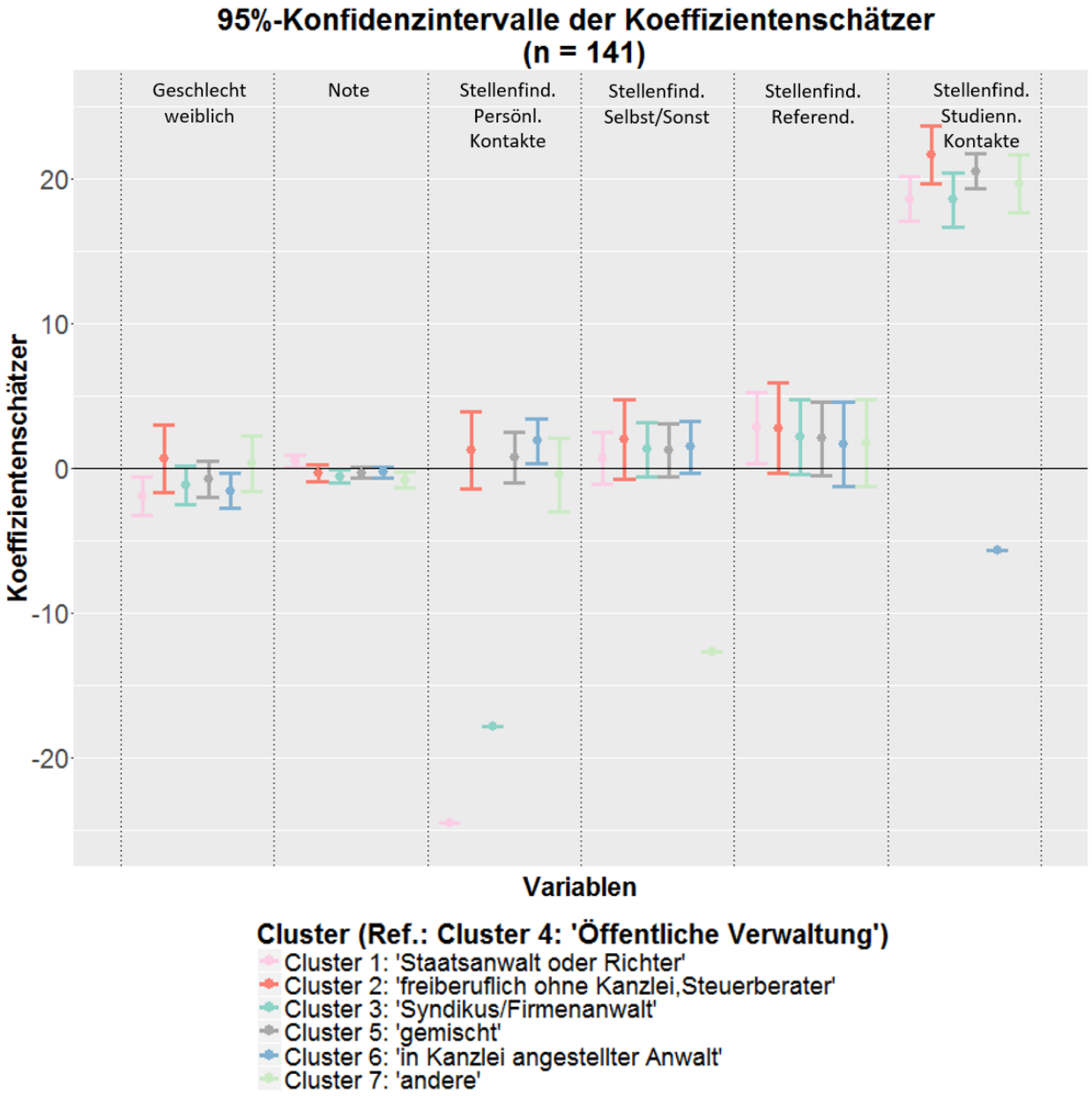

Abbildung 4.18: Regressionskoeffizienten des reduzierten multinomialen Modells zur Clusteranalyse mit 95% - Konfidenzintervall

4.4.2 Art der Anstellung

Um der Frage nachzugehen, welche Variablen einen Einfluss auf die Art der Anstellung haben, wurde ebenfalls ein multinomiales Logit-Modell gefittet. In diesem Fall wird die Variable zur *Art der Anstellung* (*unbefristet/ befristet/selbstständig*) als Zielgröße verwendet, wobei die Kategorie *unbefristet* als Referenz angesehen wird. Die Einflussgrößen sind in diesem Fall:

- Geschlecht (männlich/weiblich)
- Juristen-Eltern (Nein/Ja)
- Note (metrisch)
- Studiums-Kontakte (Anzahl)
- Privat-Kontakte (Anzahl)
- Kinder (Nein/Ja)
- Promotion (Nein/Ja)

Auch für diesen Fall wird auf Wunsch der Projektpartner zunächst das volle Modell geschätzt. Abbildung 4.19 zeigt die für dieses Modell erhaltenen Regressionskoeffizienten und die zugehörigen 95%-Konfidenzintervalle für alle Variablen und Tabelle 4.5 die genauen Werte der Regressionskoeffizienten. Die Anzahl der aufgenommenen Beobachtungen beträgt in diesem Modell $n = 90$.

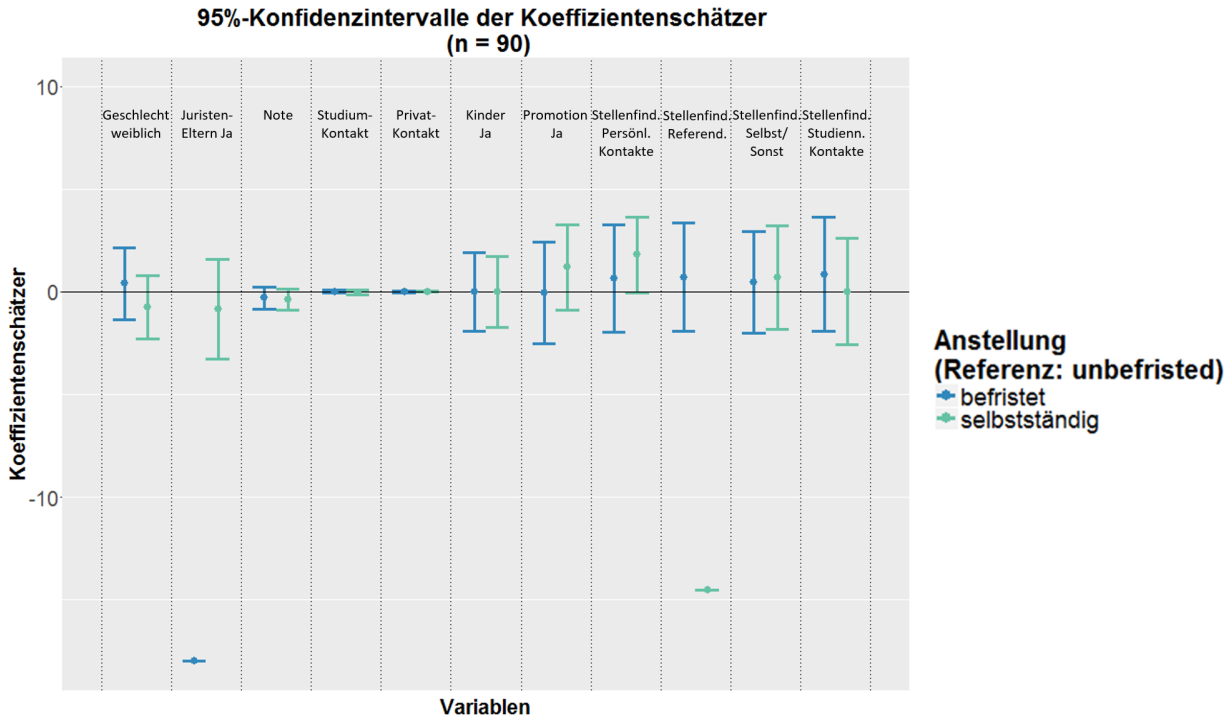

Abbildung 4.19: Regressionskoeffizienten des multinomialen Modells zur *Art der Anstellung* mit 95% - Konfidenzintervall

	(Intercept)	Geschlecht weibl.	juElternja	Note
befristet	-0.30	0.40	-17.98	-0.31
selbständig	1.26	-0.75	-0.85	-0.38

	studiumkontakt	privatkontakt	kinderJa	promoJa
befristet	0.02	-0.00	-0.01	-0.04
selbständig	-0.03	0.00	-0.01	1.19

	stelf.Persönl. K.	stelf.Ref.	stelf.Selbst/Sonst	stelf.Studienn. K.
befristet	0.65	0.71	0.45	0.85
selbständig	1.81	-14.54	0.68	0.01

Tabelle 4.5: Regressionskoeffizienten des multinomialen Logit-Modells zur *Art der Anstellung*

Auch in diesem Fall kommt es vor, dass für die Variablen *Juristen-Eltern* und *Art der Stellenfindung* in jeweils einer der drei Kategorien der *Art der Anstellung* keine Beobachtungen liegen, wie in Tabelle 4.6 überprüft werden kann. Deshalb werden hier ebenfalls LR-Tests

	Juristen-Eltern	
	Nein	Ja
unbefristet	60	11
befristet	8	**0**
selbständig	10	1

	Art der Stellenfindung				
	Form. Stellenf.	Pers. Kont.	Referend.	Selbst./Sonst.	Stud. Kont.
unbefristet	38	6	10	6	11
befristet	4	1	1	1	1
selbständig	5	4	**0**	1	1

Tabelle 4.6: Kreuztabelle zwischen *Art der Anstellung* und den Variablen *Juristen-Eltern* und *Art der Stellenfindung*

durchgeführt, um zu testen, ob der Koeffizient der jeweiligen Variablen Null ist. Hierzu wird das angegebene Modell zum einen ohne die Variable *Juristen-Eltern* und zum anderen ohne die Variable *Art der Stellenfindung* jeweils mit dem vollen Modell verglichen. Dabei ergeben sich p-Werte von 0.1819 bzw. 0.5563, d.h. die Nullhypothese wird zum Signifikanzniveau von $\alpha = 0.05$ nicht abgelehnt und man kann davon ausgehen, dass die beiden Variablen keinen Einfluss auf die Art der Anstellung haben.
Für die zusätzliche Darstellung der Regressionskoeffizienten durch den Stern-Plot siehe Abbildung A.20 in Abschnitt A.1.
Auf dieses Modell wird ebenfalls die Rückwärtsselektion mit AIC-Kriterium als Variablenselektion angewendet und dadurch ein reduziertes Modell erhalten, welches nun nur noch die Einflussgröße *Note* vorsieht. Tabelle 4.7 zeigt die zugehörigen Koeffizienten und Abbildung 4.20 die Konfidenzintervalle der Variable *Note*. Man erkennt, dass sich die Werte der Regressionkoeffizienten dieser Variable im Vergleich zu denen des vollen Modells wenig verändert haben und der Einfluss auf die *Art der Anstellung* weiterhin nicht signifikant ist.

	(Intercept)	Note
befristet	-0.21	-0.25
selbständig	0.85	-0.35

Tabelle 4.7: Regressionskoeffizienten des reduzierten multinomialen Logit-Modells zur *Art der Anstellung*

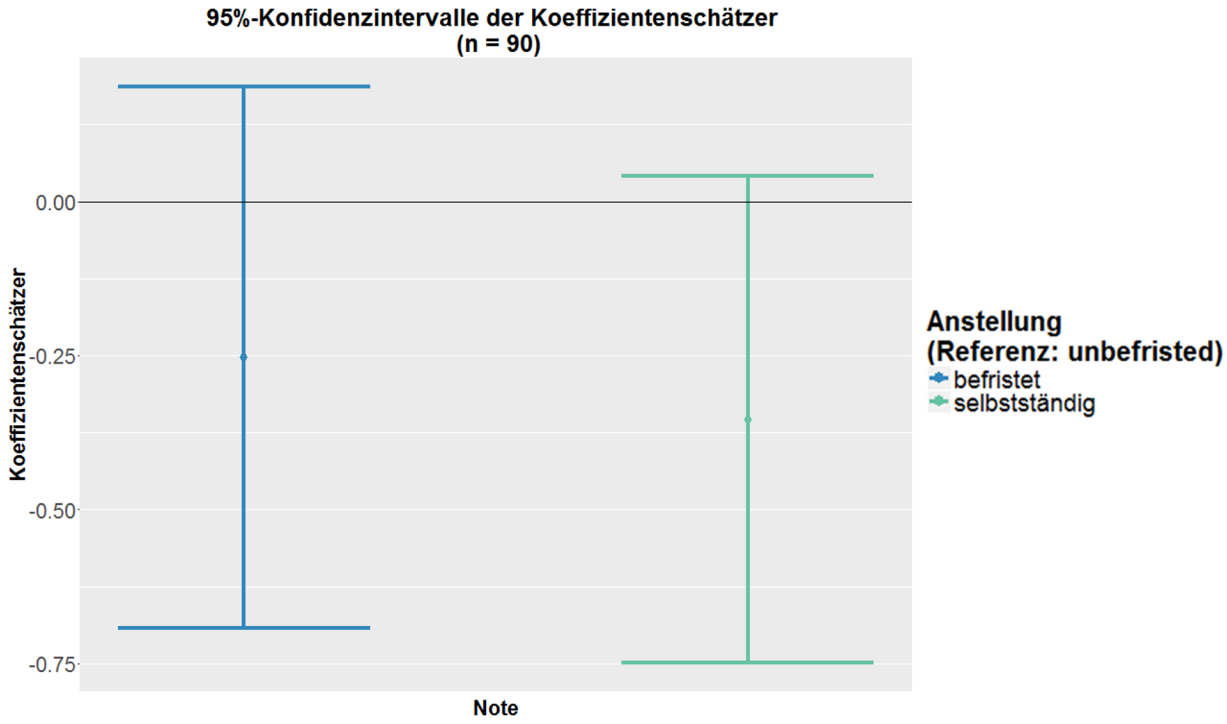

Abbildung 4.20: Regressionskoeffizienten des reduzierten multinomialen Modells zur *Art der Anstellung* mit 95% - Konfidenzintervall

4.4.3 Einkommen

Zuletzt wird die Fragestellung betrachtet, welchen Einfluss verschiedene Variablen auf das Einkommen der Absolventen haben. Hierzu wird ein lineares Regressionmodell geschätzt. Für die Zielgröße y wird der Stundenlohn berechnet und logarithmiert aufgenommen. Zusätzlich zu den in Unterabschnitt 4.4.2 genannten Einflussgrößen, werden folgende Variablen in das Modell mit aufgenommen:

- Cluster (Cluster 1,..., Cluster 7) (Referenz: Cluster 4)
- Interaktion Geschlecht:Note
- Interaktion Geschlecht:Promotion

Die durch dieses Modell berechneten Regressionskoeffizienten und die zugehörigen Standardfehler und p-Werte zu allen Variablen sind in Tabelle 4.8 zu sehen. In Abbildung 4.21

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.60	0.48	5.46	0.00
juElternnein	0.01	0.14	0.05	0.96
stelfindungPersönliche Kontakte	-0.17	0.14	-1.25	0.22
stelfindungReferendariat	0.15	0.13	1.15	0.25
stelfindungSelbständig/Sonstige	0.01	0.15	0.05	0.96
stelfindungStudiennahe Kontakte	0.15	0.15	0.98	0.33
kinderJa	0.05	0.10	0.48	0.63
studiumkontakt	0.00	0.00	1.23	0.22
privatkontakt	0.00	0.00	0.17	0.86
Geschlechtweiblich	0.03	0.55	0.05	0.96
Note	0.05	0.05	0.98	0.33
promoNein	-0.02	0.15	-0.16	0.87
cluster1	-0.13	0.15	-0.83	0.41
cluster2	0.18	0.19	0.95	0.34
cluster3	0.28	0.15	1.91	0.06
cluster5	-0.17	0.14	-1.20	0.24
cluster6	0.06	0.14	0.46	0.65
cluster7	0.16	0.18	0.87	0.39
Geschlechtweiblich:Note	-0.01	0.06	-0.12	0.91
Geschlechtweiblich:promoNein	-0.00	0.23	-0.00	1.00

Tabelle 4.8: Output der Regressionskoeffizienten des linearen Modells bzgl. Stundenlohn.

sind die Regressionskoeffizienten und die 95%-Konfidenzintervalle zu allen Einflussgrößen des vollen Modells abgetragen. Die Fallzahl beträgt hier $n = 84$.

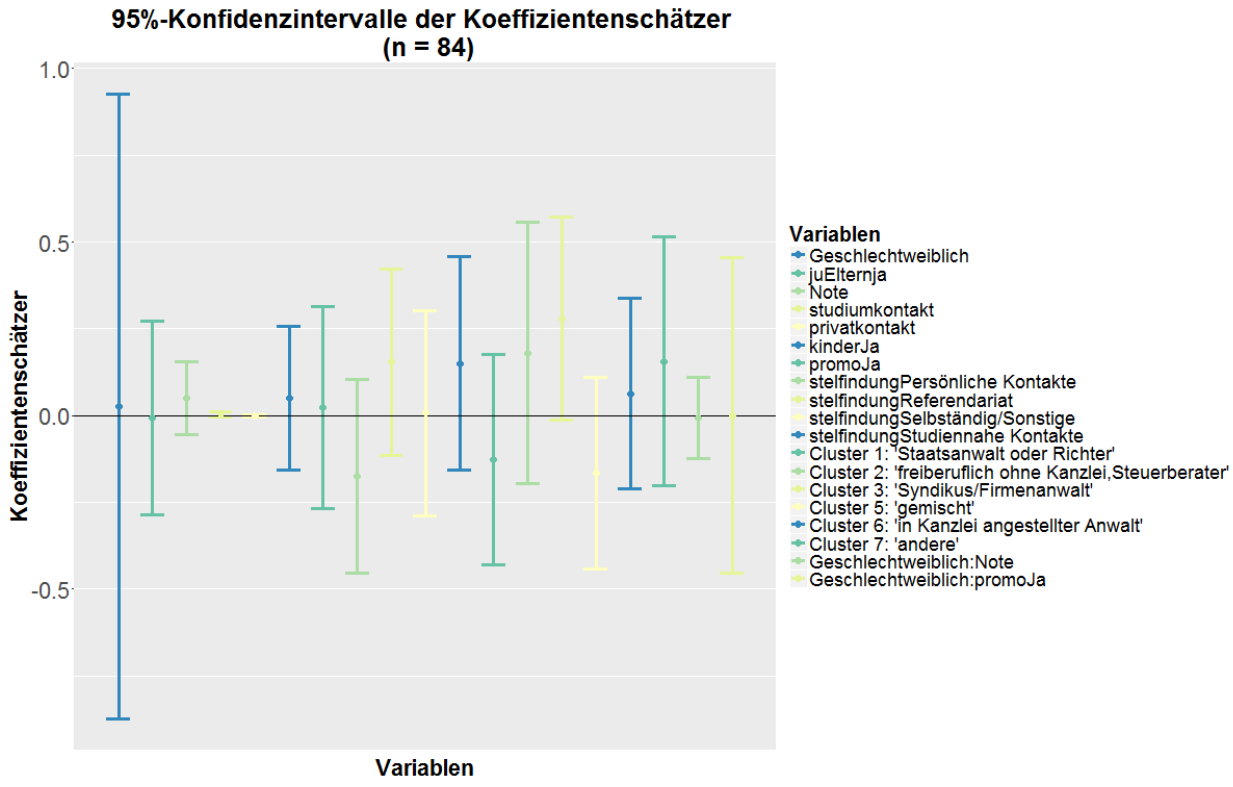

Abbildung 4.21: Regressionskoeffizienten des linearen Modells zum Stundenlohn mit 95% - Konfidenzintervall.

Wie zu erkennen ist, schneidet jedes der Konfidenzintervalle die Nulllinie, d.h. keine Variable hat einen signifikanten Einfluss auf den Stundenlohn der Absolventen. Dies ist auch an den p-Werten in Tabelle 4.8 zu erkennen, die alle größer als $\alpha = 0.05$ sind.
Nach der Berechnung des vollen Modells wird eine Variablenselektion wie in Unterabschnitt 4.4.1 und 4.4.2 durchgeführt und dadurch ein reduziertes Modell erhalten, welches die Einflussgrößen *Note*, *Studiumkontakt* und *Cluster* enthält. Die für dieses Modell erhaltenen Regressionsschätzer, Standardfehler und p-Werte sind in Tabelle 4.9 aufgeführt. Wie man am p-Wert erkennt, haben nun die Variablen *Note* und *Studiumkontakt* und die Kategorie *cluster3* der Variable *Cluster* zum Signifikanzniveau 0.05 einen signifikanten Einfluss auf den Stundenlohn, da jeweils $p < 0.05$. Steigt beispielsweise die Note um eine Einheit, so steigt der erwartete Stundenlohn bei Konstanthaltung der anderen Kovariablen um 0.06 €.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.49	0.21	11.70	0.00
Note	0.06	0.02	2.64	0.01
studiumkontakt	0.01	0.00	2.06	0.04
cluster1	-0.06	0.12	-0.54	0.59
cluster2	0.22	0.17	1.31	0.19
cluster3	0.34	0.13	2.62	0.01
cluster5	-0.11	0.12	-0.89	0.37
cluster6	0.05	0.12	0.45	0.65
cluster7	0.13	0.16	0.81	0.42

Tabelle 4.9: Regressionskoeffizienten des reduzierten linearen Modells zum Stundenlohn mit 95% - Konfidenzintervall

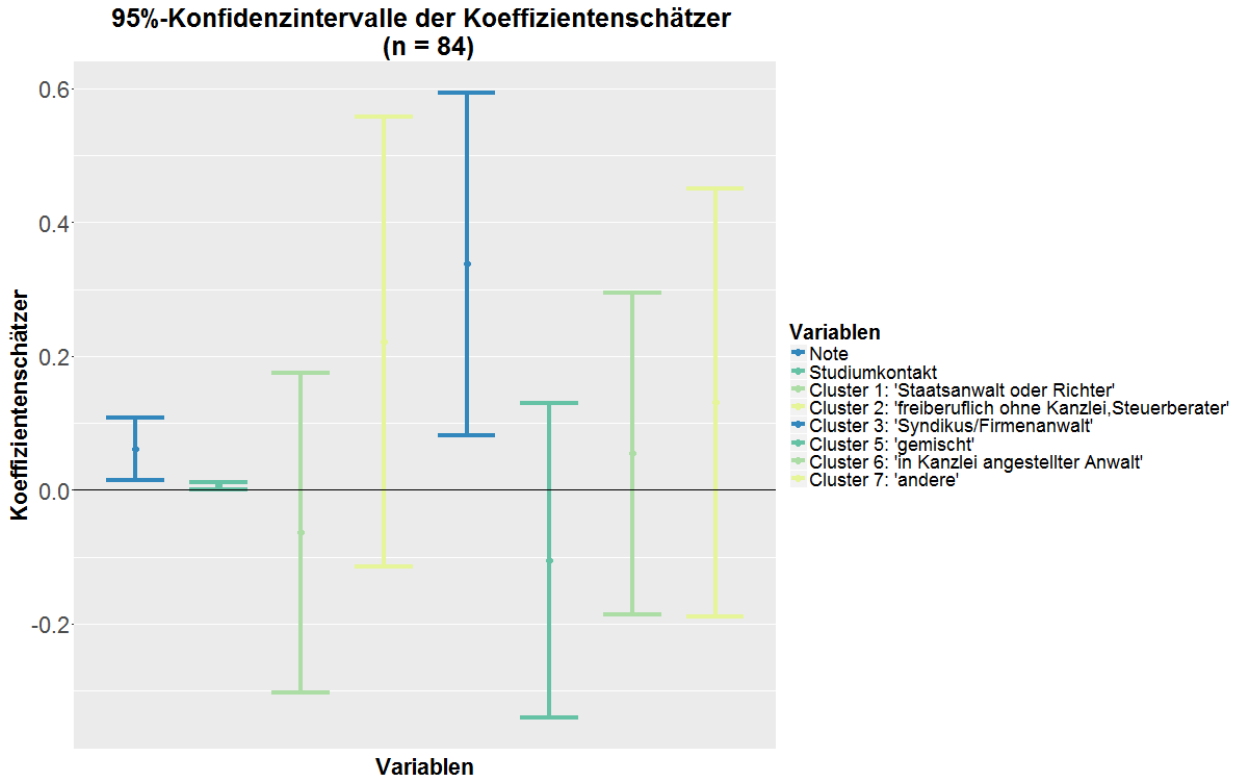

Abbildung 4.22: Regressionskoeffizienten des reduzierten linearen Modells zum Stundenlohn mit 95% - Konfidenzintervall

5 Zusammenfassung, Probleme der Analyse und Ausblick

Im folgenden, abschließenden Kapitel werden in Abschnitt 5.1 zunächst noch einmal die Ergebnisse zusammengefasst und abschließende Erkenntnisse formuliert. Ebenso soll in Abschnitt 5.2 auf die Probleme, welche während der Analyse entstanden sind, hingewiesen werden. Diese können zeitgleich als Kritikpunkte an die Analyse von Sequenzen angesehen werden. In Abschnitt 5.3 werden einige Lösungsansätze für diese Probleme genannt sowie auf weitere Analysemöglichkeiten hingewiesen.

5.1 Zusammenfassung

Haben Faktoren wie die Abschlussnote oder der familiäre Hintergrund Einfluss auf die Art der Anstellung oder das Einkommen der Absolventen? Folgen die Berufsverläufe einem bestimmten Muster? Macht es einen Unterschied, den beruflichen Werdegang betreffend, wie die Stelle gefunden wurde? Auf der Suche nach Antworten auf diese und weitere in Kapitel 1 genannte Fragestellungen wurden statistische Methoden angewandt, allen voraus die Sequenzmusteranalyse. Die Sequenzmusteranalyse ermöglichte in Abschnitt 4.1 einen Überblick über die vorliegenden Berufsverläufe. Dabei war zu erkennen, dass die Absolventen nach dem Referendariat und eventuell einer kurzen Periode der Arbeitslosigkeit eine Anstellung gefunden haben, in welcher sie auch verblieben. Bei Betrachtung der Sequenzen nach Gruppenzugehörigkeit, wie etwa dem Geschlecht oder der Art der Stellenfindung, fielen keine Besonderheiten auf. Einzig eine leichte Abweichung innerhalb der Berufsverläufe war zu bemerken.

Abschnitt 4.2 ließ erkennen, dass es unter den Berufsverläufen Sequenzen gibt, welche als redundant angesehen werden können. Dies bedeutet, dass alle zueinander redundanten Sequenzen durch eine Sequenz, welcher sie alle ähnlich sind, repräsentiert werden können. Diese sogenannten repräsentativen Sequenzen ähneln stark dem grundsätzlichen Verlauf der in Abschnitt 4.3 vorgestellten Cluster. Während zu Beginn eines jeden Clusters bzw. einer jeden repräsentativen Sequenz das Referendariat steht, folgt nach einer kurzen Periode der *Arbeitlosigkeit* entweder eine Anstellung in der *öffentlichen Verwaltung*,

als *Staatsanwalt oder Richter*, als *in einer Kanzlei angestellter Anwalt*, als *Syndikus* oder als *freiberuflich ohne Kanzlei arbeitender Steuerberater*. Eine weitere Gruppe der Absolventen trat nach dem *Referendariat* und der darauffolgenden *Arbeitslosigkeit* bzw. Übergangsphase eine Stelle außerhalb des juristischen Bereichs (*andere*) an.

Mit Hilfe der in Abschnitt 4.4 durchgeführten Regressionen, können die zu Beginn gestellten Fragen nach dem Einfluss bestimmter Variablen auf die Zielgrößen *Clusterzugehörigkeit, Einkommen* und *Art der Anstellung* nun beantwortet werden. Aufgrund der in Unterabschnitt 4.4.1 vorgestellten Clusterregression kann festgehalten werden, dass sowohl die Variable *Note* als auch *Geschlecht, Art der Stellenfindung* und *Juristen-Eltern* einen signifikanten Einfluss auf die Clusterzugehörigkeit hat. So steigt die Chance für Absolventen in *Cluster 1* (*Staatsanwalt/Richter*) statt in *Cluster 4* (*öffentliche Verwaltung*) zu kommen mit einem Anstieg der bei der Abschlussnote erreichten Punktzahl. Im Gegensatz dazu sinkt die Chance den *Clustern 3, 5, 6* und *7* zugeordnet zu werden mit steigender Punktzahl. Bei Betrachtung der Note scheinen diejenigen mit einer höhreren Punktzahl eher als Staatsanwalt/Richter zu arbeiten als in der öffentlichen Verwaltung. Dagegen finden Absolventen mit einer eher geringeren Punktzahl eher einer Anstellung als Firmenanwalt oder in allen zu *Cluster 5* gehörigen Bereiche (z.B. Universität), als ein in einer Kanzlei angestellter Anwalt oder im außerjuristischen Bereich. Ebenso sinkt die Chance für weibliche Absolventen in *Cluster 1* oder *6* statt in *Cluster 4* zu kommen. Das lässt vermuten, dass Frauen eher in der öffentlichen Verwaltung angestellt werden als dass sie als Staatsanwältin/Richterin oder eine in einer Kanzlei angestellte Anwältin arbeiten. Steigende Chancen für Absolventen mit Juristen-Eltern in *Cluster 1* oder *5* statt in *Cluster 4* zugeordnet zu werden, deuten darauf hin, dass diese eher einen Job als Staatsanwalt/Richter oder in einem *Cluster 5* zugeodneten Bereich ausüben. Dem durchgeführten LR-Test zu Folge hat auch die *Art der Stellenfindung* einen signifikanten ($\alpha = 0.05$) Einfluss auf die Clusterzugehörigkeit. Diese ist jedoch aufgrund der fehlenden Fallzahlen innerhalb mancher Kategorien nicht genau quantifizierbar. Dieses Ergebnis ergab sich sowohl für das volle als auch das reduzierte Modell.

Die multinomiale Regression bzgl. der *Art der Anstellung* ergab keine signifikanten Unterschiede in den Variablen. Eventuell denkbare Einflüsse auf die Art der Anstellung durch die erreichte Abschlussnote, das Geschlecht, die Art der Stellefindung oder ob die Eltern ebenfalls Juristen sind, scheinen nicht vorzuliegen.

Anders bei den Ergebnissen der Einkommensregression. Während beim vollen Modell

keine der Variablen einen signifikanten Einfluss zu haben schien, änderte sich dies beim reduzierten Modell. In diesem waren nur noch die Variablen *Note, Studiumkontakt* und *Cluster* enthalten. Nach dem gefitteten Modell steigt der Stundenlohn mit zunehmender Punktzahl in der Abschlussnote. Der Stundenlohn steigt ebenso, wenn die Absolventen mehr Studienkontakte besitzen oder sie *Cluster 3* (Syndikus/Firmenanwalt) zugeordnet wurden.

5.2 Probleme der Analyse

An dieser Stelle sollen einige Probleme genannt werden, die während der Analyse aufgetreten sind. Eine der Methoden, bei welcher es zu Problemen kam, ist das Optimal Matching. Durch fehlende Zustände in den Berufsverläufen wird dieses, und damit die Bestimmung der Distanzen zwischen den einzelnen Sequenzen, verzerrt. Ein weiterer Punkt ist die Quantifizierung qualitativer Merkmale, wie der Substitutionskosten. Hier empfiehlt sich dringend, diese aus den Daten zu berechnen (vgl. Abschnitt 3.2). Andernfalls müssen diese willkürlich bestimmt werden. Die korrekte Abschätzung, ob die Substitution von Zustand A durch Zustand B nun mehr Kosten verursacht als jene von Zustand C durch Zustand D, ist nahezu unmöglich. Da das Ergebnis der Clusteranalyse maßgeblich von den vorher festgelegten Kosten abhängt, ist auf diese willkürliche Art der Bestimmung der Kostenmatrix zu verzichten. Ein zusätzliches Problem des Optimal Matchings, welches hier jedoch nicht zum Tragen kam, ist die Tatsache, dass die genauen Zeitpunkte der Zustände beim Optimal Matching nicht berücksichtigt werden. Ein weiteres, im Rahmen der Clusteranalyse aufgetretenes Problem äußert sich in der Wahl der Clusterlösung. Während bei der Verwendung metrischer Daten eine Vielzahl sogenannter *stopping rules* zu Rate gezogen werden können, ist dies beim Clustern mit Sequenzen nicht möglich. Einzig das auch während der Analyse benutzte *within-between*-Kriterium kann empfohlen werden (Stegmann et al. (2013), S.68). (Stegmann et al. (2013)) Auch bei den Regressionen mussten aufgrund vieler fehlender Werte in den Zielgrößen und unvollständiger Fälle Einbußen bezüglich der Fallzahlen gemacht werden. Die Regressionsergebnisse sind somit mit Vorsicht zu genießen.

5.3 Ausblick

Zu denen in Abschnitt 5.2 genannten Problemen sollen hier Lösungsansätze sowie weiterführende Analysemöglichkeiten gegeben werden. Hier sei zum einen die Möglichkeit der Erweiterung des Optimal Matchings durch Gewichtung der Übergänge der Transitionsmatrix zu nennen. Während es in Abschnitt 3.2 irrelevant war, ob ein Zustandswechsel von A nach B oder von B nach A stattfindet, so können diese Wechsel durch die Gewichtung unterschieden werden. Damit erreicht man zum Beispiel eine Differenzierung zwischen „Arbeit finden“ (*arbeitslos→Staatsanwalt*)und „Arbeit verlieren“ (*Staatsanwalt→arbeitslos*).

Eine weitere Möglichkeit ist die Verwendung einer leicht veränderten Variante des Optimal Matchings (OMv). Hierbei wird die Berechnung der Distanzen an die Verweildauern in den Zuständen angepasst. Anwendungen zu Folge sollen sich die Ergebnisse hier jedoch nicht sehr von denen bei Verwendung des Optimal Matchings unterscheiden. Eine Alternative zum Optimal Matching ist der Algortihmus von *Elzinga*. Beim Elzinga-Algorithmus wird ausschließlich die Abfolge der Zustände betrachtet, während die zeitliche Dauer in den Hintergrund gestellt wird. (Wolf und Best (2010), S.1040)

Um das Problem der Verzerrung des Optimal Matchings aufgrund fehlender Zustände zu beheben, könnten diese imputiert werden (Stegmann et al. (2013)). Da die Berufsverläufe jedoch alle individuell sind und es somit fast unmöglich ist, den fehlenden Zustand korrekt zu bestimmen, ist diese Möglichkeit zwar theoretisch möglich jedoch praktisch nicht unbedingt empfehlenswert.

Literaturverzeichnis

Allison, P. D. (2008). Convergence failures in logistic regression. *SAS Global Forum 2008* (Paper 360-2008).

Bacher, J. (2008). *Clusteranalyse: Anwendungsorientierte Einführung* (3., Aufl. ed.). München: Oldenbourg, R.

Fahrmeir, L., T. Kneib, und S. Lang (2009). *Regression: Modelle, Methoden und Anwendungen*. Berlin/Heidelberg: Springer Berlin Heidelberg.

Fred, A., J. L. G. Dietz, K. Liu, und J. Filipe (Eds.) (2011). *Knowledge Discovery, Knowlege Engineering and Knowledge Management: First International Joint Conference, IC3K 2009, Funchal, Madeira, Portugal, October 6-8, 2009, Revised Selected Papers*, Volume 128 of *Communications in Computer and Information Science*. Berlin, Heidelberg: Springer Berlin Heidelberg.

Gabadinho, A., G. Ritschard, M. Studer, und N. S. Müller (2008). Mining sequence data in r with the traminer package: A user's guide: University of geneva.

Gabadinho, A. und G. Ritschard (2013). Searching for typical life trajectories applied to childbirth histories. In R. Levy and E. Widmer (Eds.), *Gendered life courses between individualization and standardization. A European approach applied to Switzerland*, S. 287–312. Vienna: Vienna : LIT Verlag.

Gabadinho, A., G. Ritschard, N. S. Müller, und M. Studer (2011). Analyzing and visualizing state sequences in r with traminer. *Journal of Statistical Software 40*(4), 1–37.

Gabadinho, A., G. Ritschard, M. Studer, und N. S. Müller (2011). Extracting and rendering representative sequences. In A. Fred, J. L. G. Dietz, K. Liu, and J. Filipe (Eds.), *Knowledge Discovery, Knowlege Engineering and Knowledge Management*, Volume 128 of *Communications in Computer and Information Science*, S. 94–106. Berlin, Heidelberg: Springer Berlin Heidelberg.

Gabadinho, A., Ritschard, G., Studer, M., Müller, N. S., und Buergin, R. (2016). Package 'traminer'.

Levy, R. und E. Widmer (Eds.) (2013). *Gendered life courses between individualization and standardization. A European approach applied to Switzerland*. Vienna: LIT Verlag.

Maechler, M., P. Rousseeuw, A. Struyf, M. Hubert, und K. Hornik (2015). *cluster: Cluster Analysis Basics and Extensions*. R package version 2.0.1 — For new features, see the 'Changelog' file (in the package source).

Murtagh, F. und P. Legendre (2014). Ward's hierarchical agglomerative clustering method: Which algorithms implement ward's criterion? *Journal of Classification 31*(3), 274–295.

Neuwirth, E. (2014). *RColorBrewer: ColorBrewer Palettes*. R package version 1.1-2.

R Core Team (2015a). *foreign: Read Data Stored by Minitab, S, SAS, SPSS, Stata, Systat, Weka, dBase, ...* R package version 0.8-63.

R Core Team (2015b). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Stegmann, M., H. Müller, und J. Werner (2013). *Sequenzmusteranalyse: Einführung in Theorie und Praxis*, Volume 5 of *Sozialwissenschaftliche Forschungsmethoden*. München and Mering: Hampp.

Studer, M. und G. Ritschard (2015). What matters in differences between life trajectories: A comparative review of sequence dissimilarity measures. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*.

Studer, M., G. Ritschard, A. Gabadinho, und N. S. Müller (2010). Discrepancy analysis of complex objects using dissimilarities. In F. Guillet, G. Ritschard, D. A. Zighed, and H. Briand (Eds.), *Advances in Knowledge Discovery and Management*, Studies in Computational Intelligence, S. 3–19. Berlin: Springer. 292.

Studer, M., G. Ritschard, A. Gabadinho, und N. S. Müller (2011). Discrepancy analysis of state sequences. *Sociological Methods and Research 40*(3), 471–510.

Venables, W. N. und B. D. Ripley (2002). *Modern Applied Statistics with S* (Fourth ed.). New York: Springer. ISBN 0-387-95457-0.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wickham, H. (2015). *lazyeval: Lazy (Non-Standard) Evaluation*. R package version 0.1.10.

Wickham, H. und R. Francois (2015). *dplyr: A Grammar of Data Manipulation*. R package version 0.4.3.

Wolf, C. und H. Best (2010). *Handbuch der sozialwissenschaftlichen Datenanalyse* (1. Aufl. ed.). Wiesbaden: VS Verlag für Sozialwissenschaften / Springer Fachmedien Wiesbaden GmbH Wiesbaden.

Abbildungsverzeichnis

Tabellenverzeichnis

A Anhang

Im Anhang finden sich ergänzende, bisher nicht gezeigte Grafiken (Abschnitt A.1), sowie eine Dokumentation und Beschreibung des angefügten elektronischen Anhangs (Abschnitt A.2).

A.1 Weitere Grafiken

A.1.1 Sequenzmusteranalyse

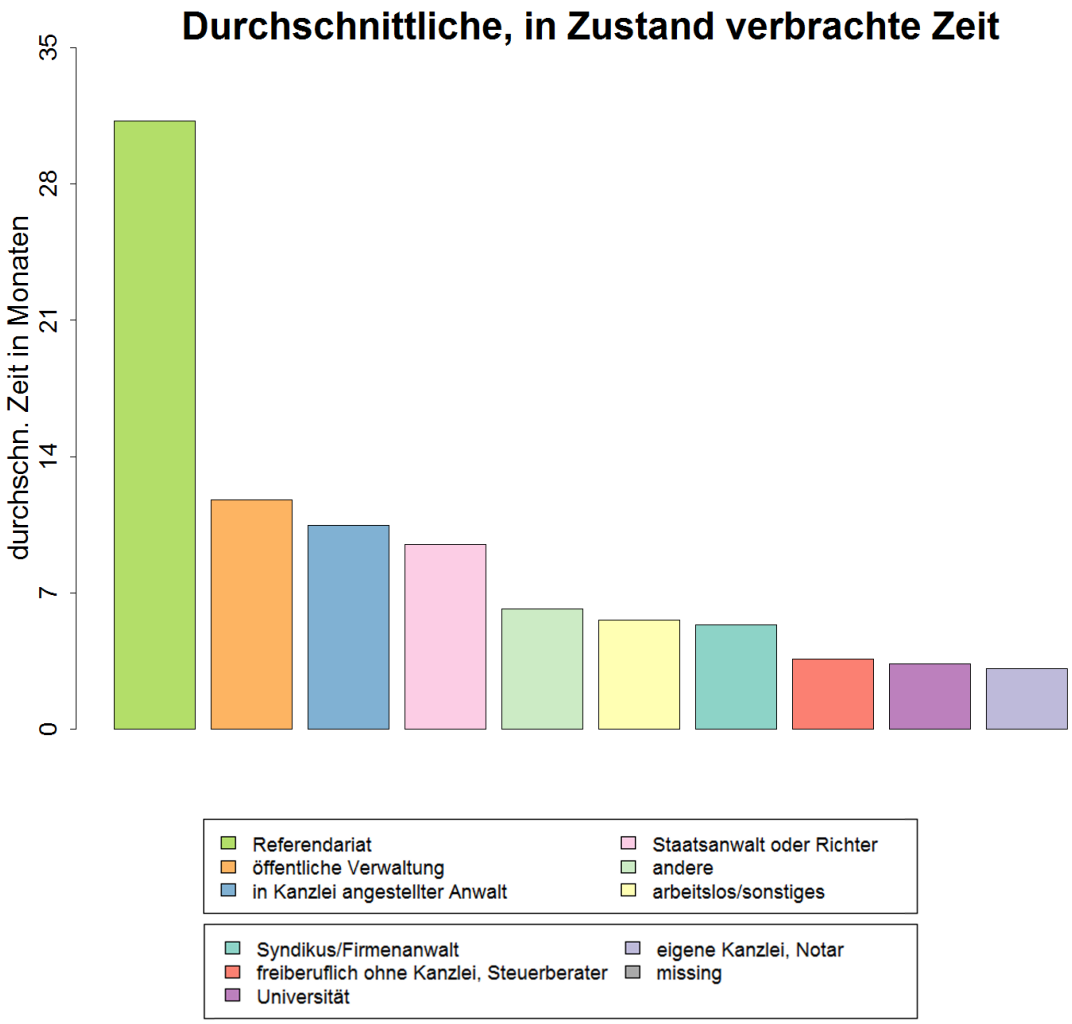

Abbildung A.1: Durchschnittliche, in Zustand verbrachte Zeit

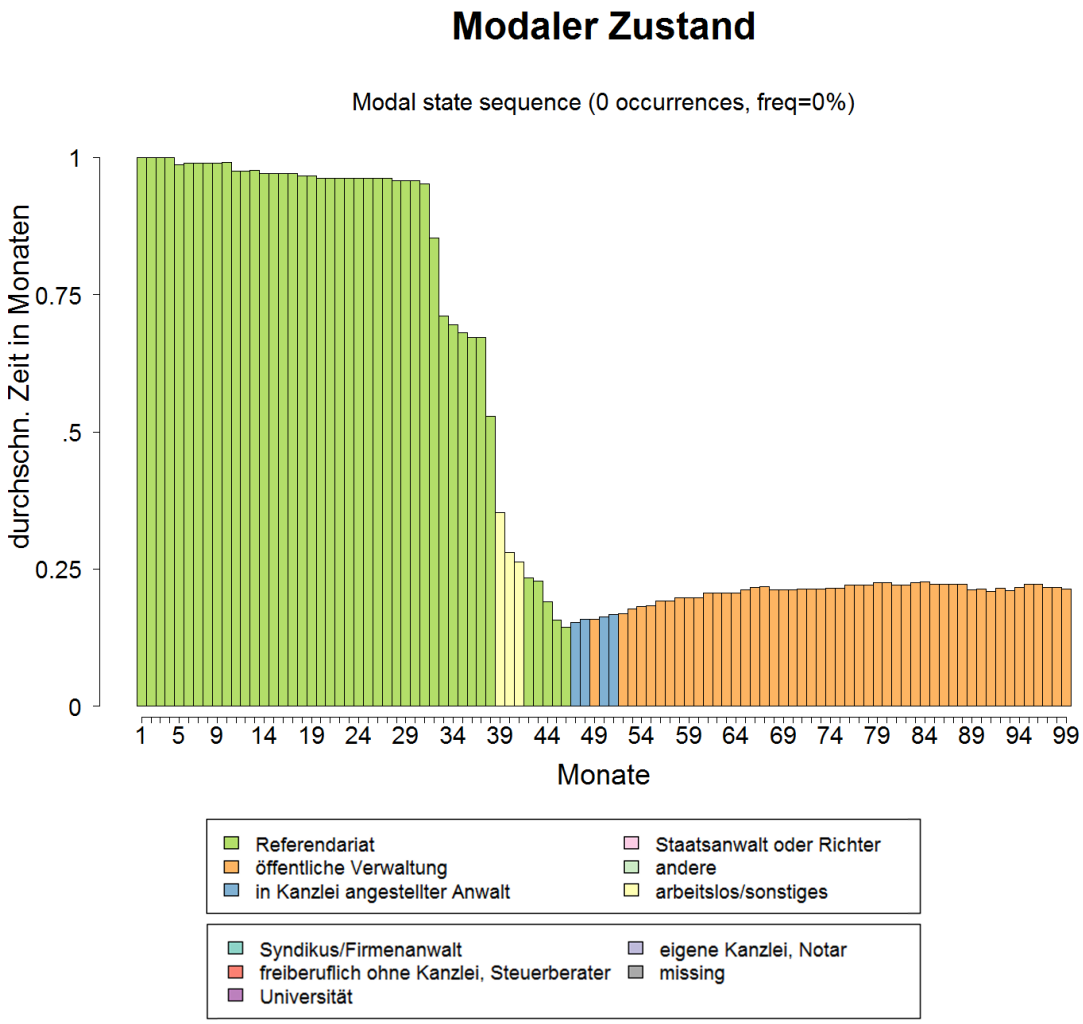

Abbildung A.2: Modaler Zustand

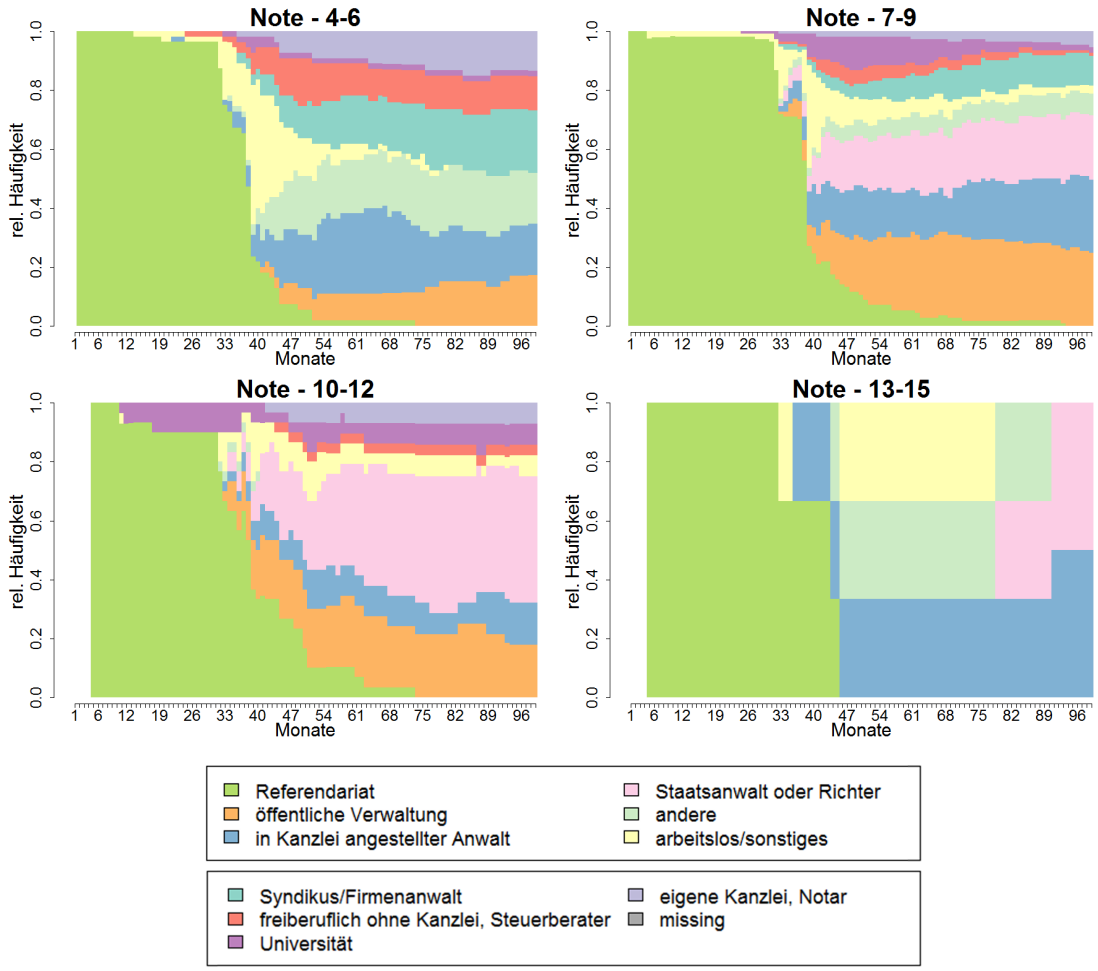

Abbildung A.3: Relative Zustandshäufigkeit (Note)

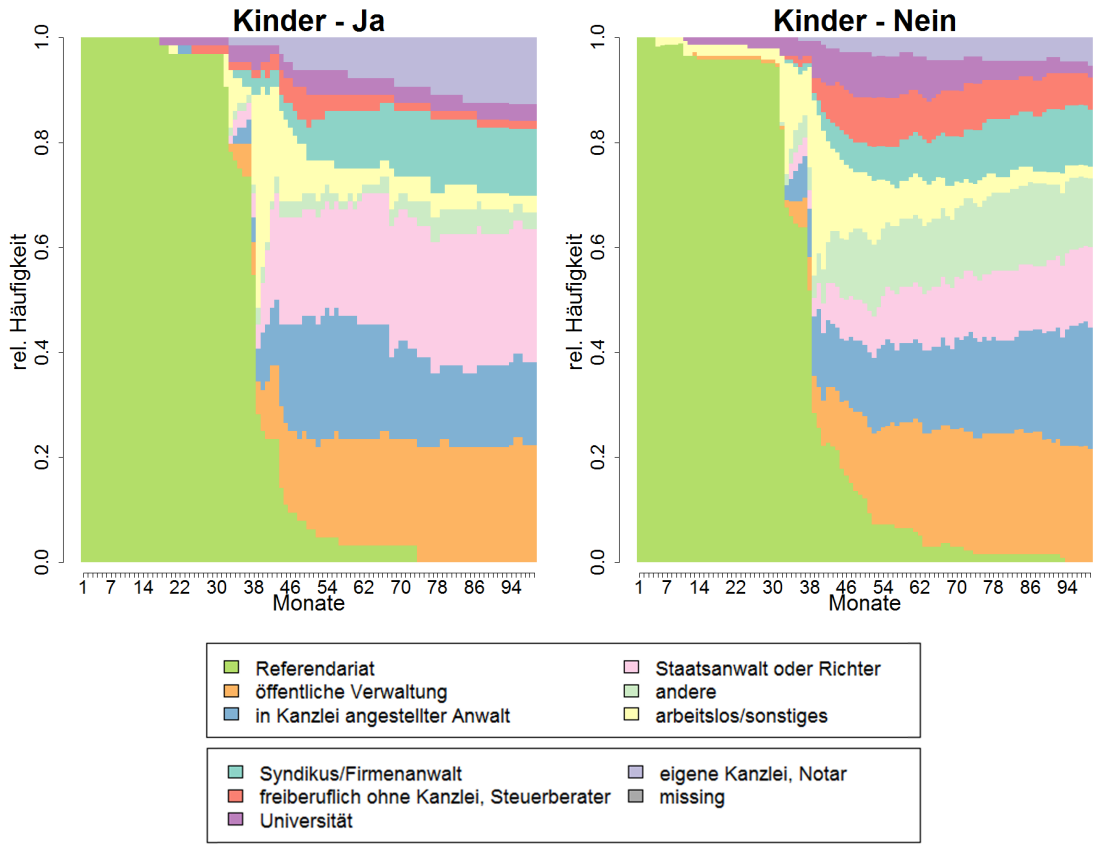

Abbildung A.4: Relative Zustandshäufigkeit (Kinder)

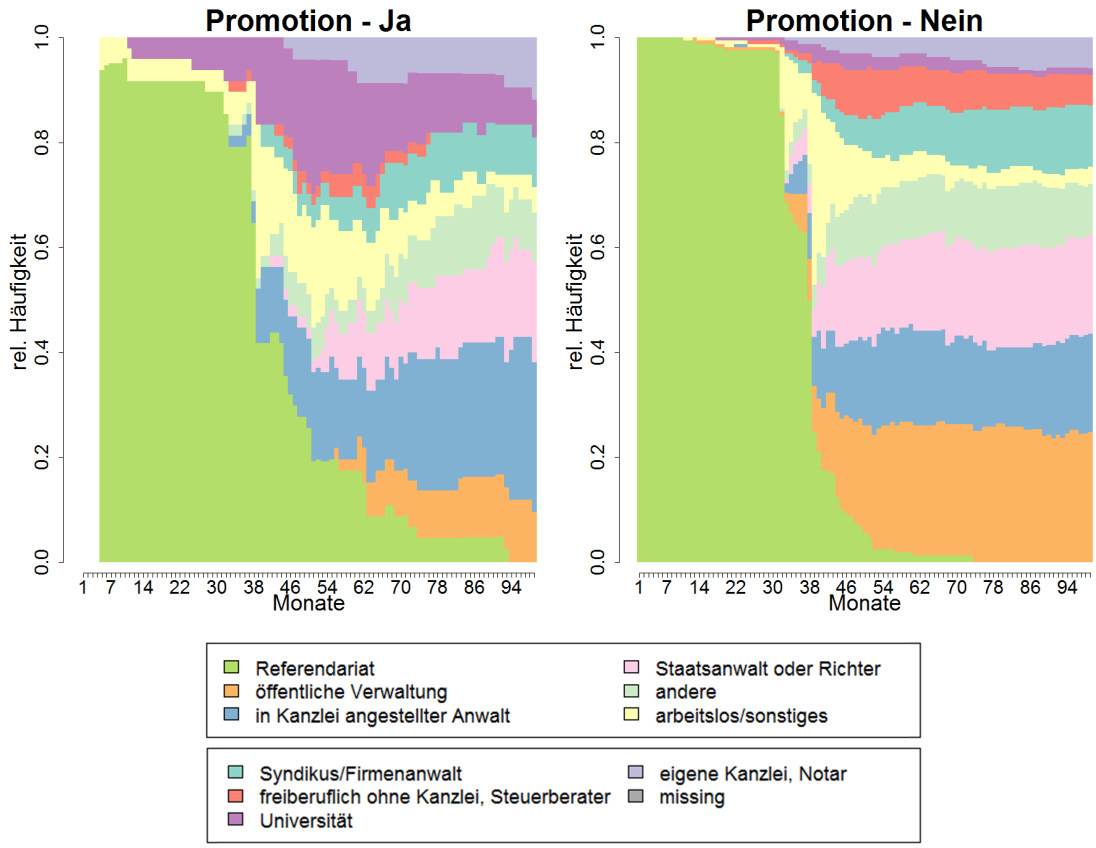

Abbildung A.5: Relative Zustandshäufigkeit (Promo)

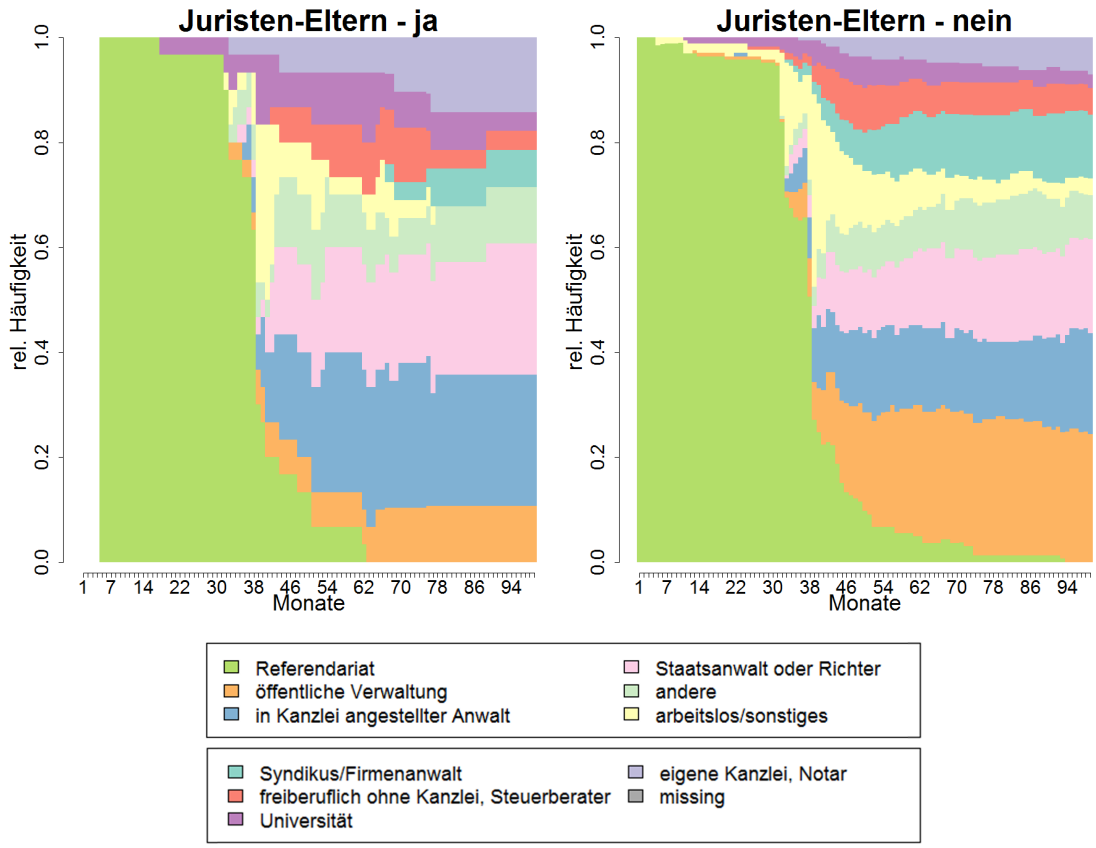

Abbildung A.6: Relative Zustandshäufigkeit (Juristen-Eltern)

A.1.2 Repräsentative Sequenzen

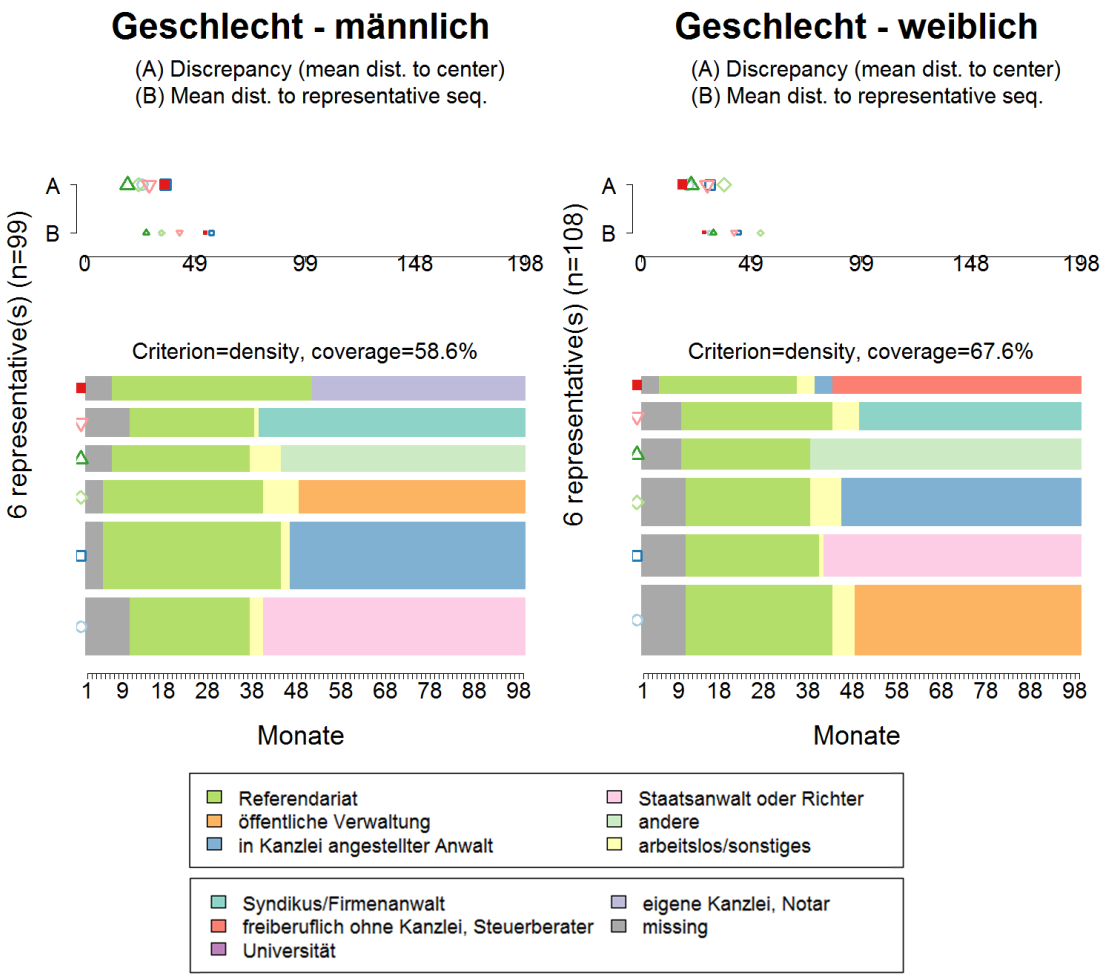

Abbildung A.7: Sechs repräsentative Sequenzen getrennt nach der Variable Geschlecht

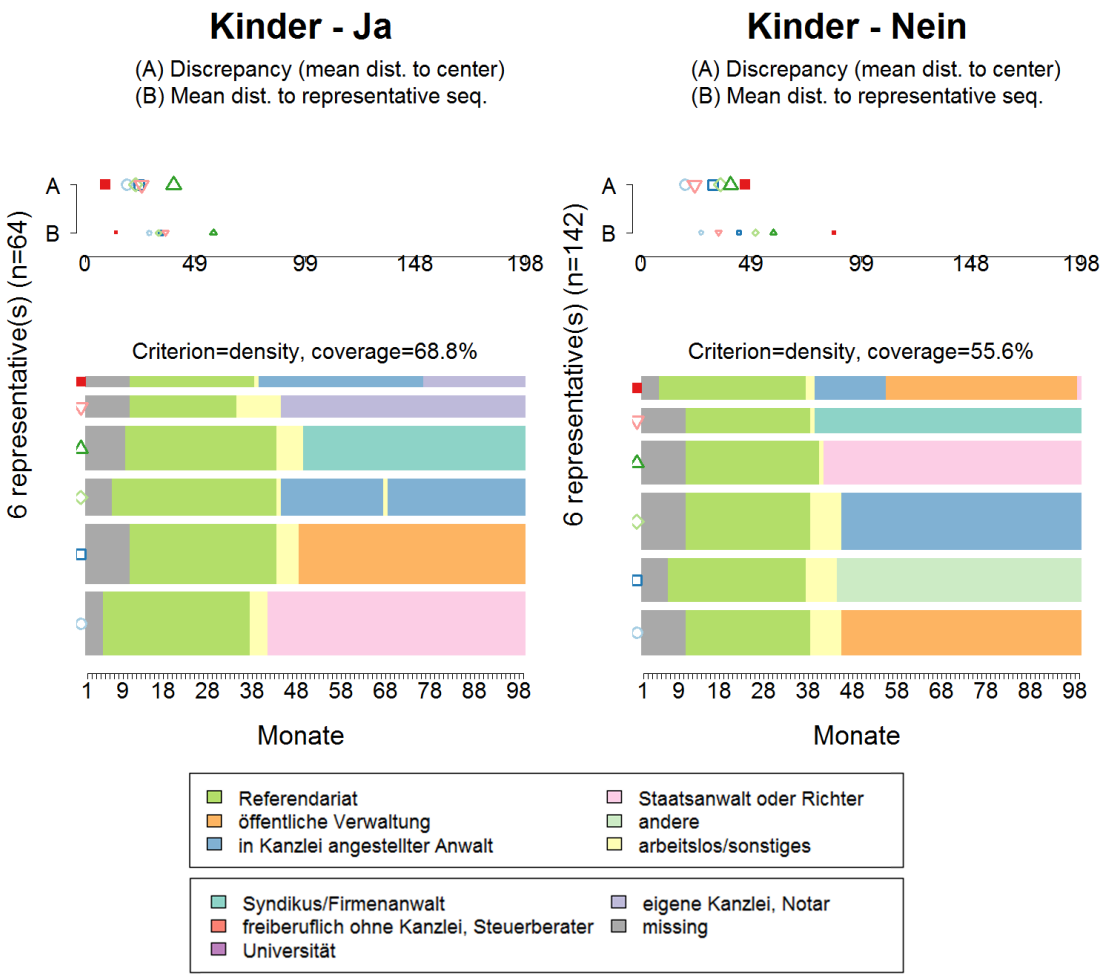

Abbildung A.8: Sechs repräsentative Sequenzen getrennt nach der Variable Kinder

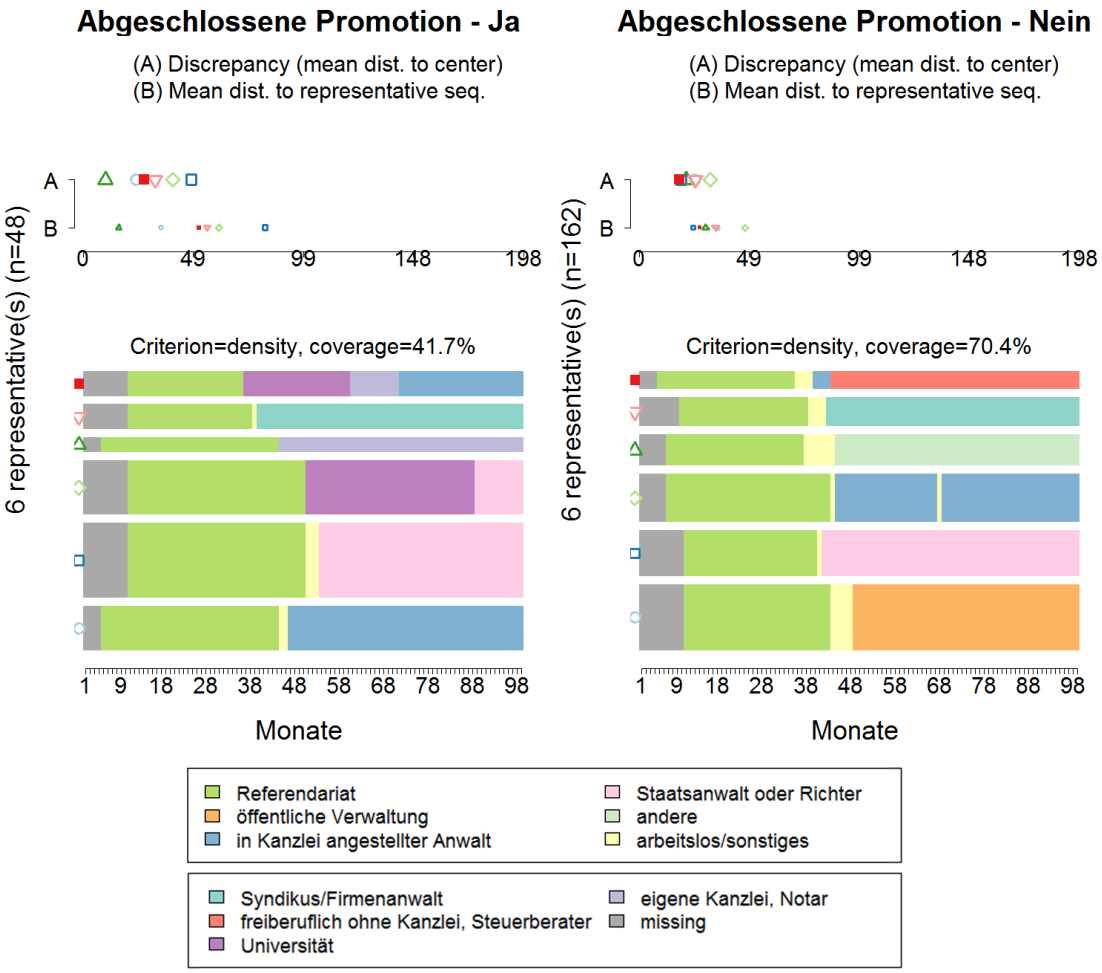

Abbildung A.9: Sechs repräsentative Sequenzen getrennt nach der Variable Promotion

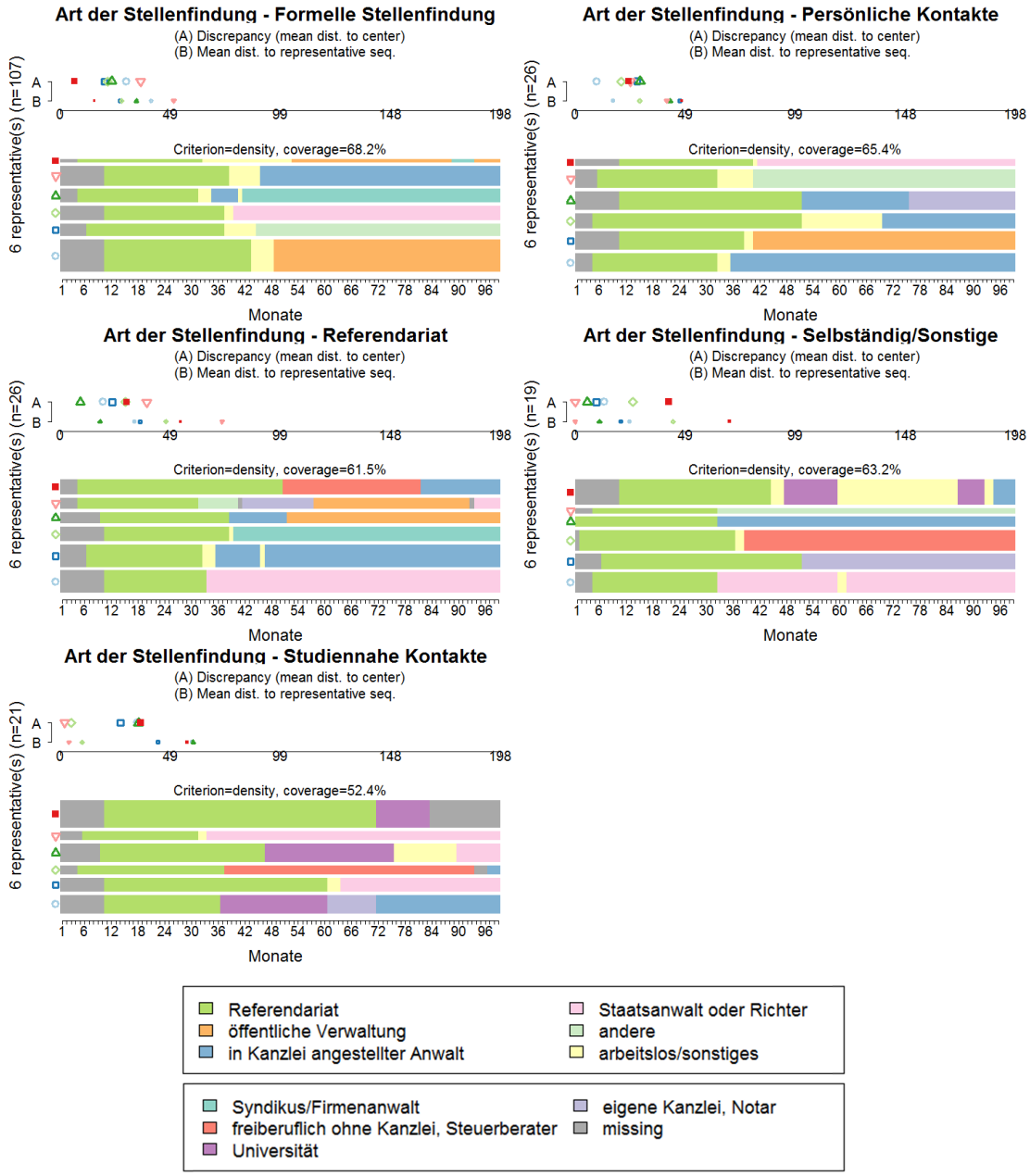

Abbildung A.10: Sechs repräsentative Sequenzen getrennt nach der Variable Stellenfindung

A.1.3 Clusteranalyse - deskriptiv

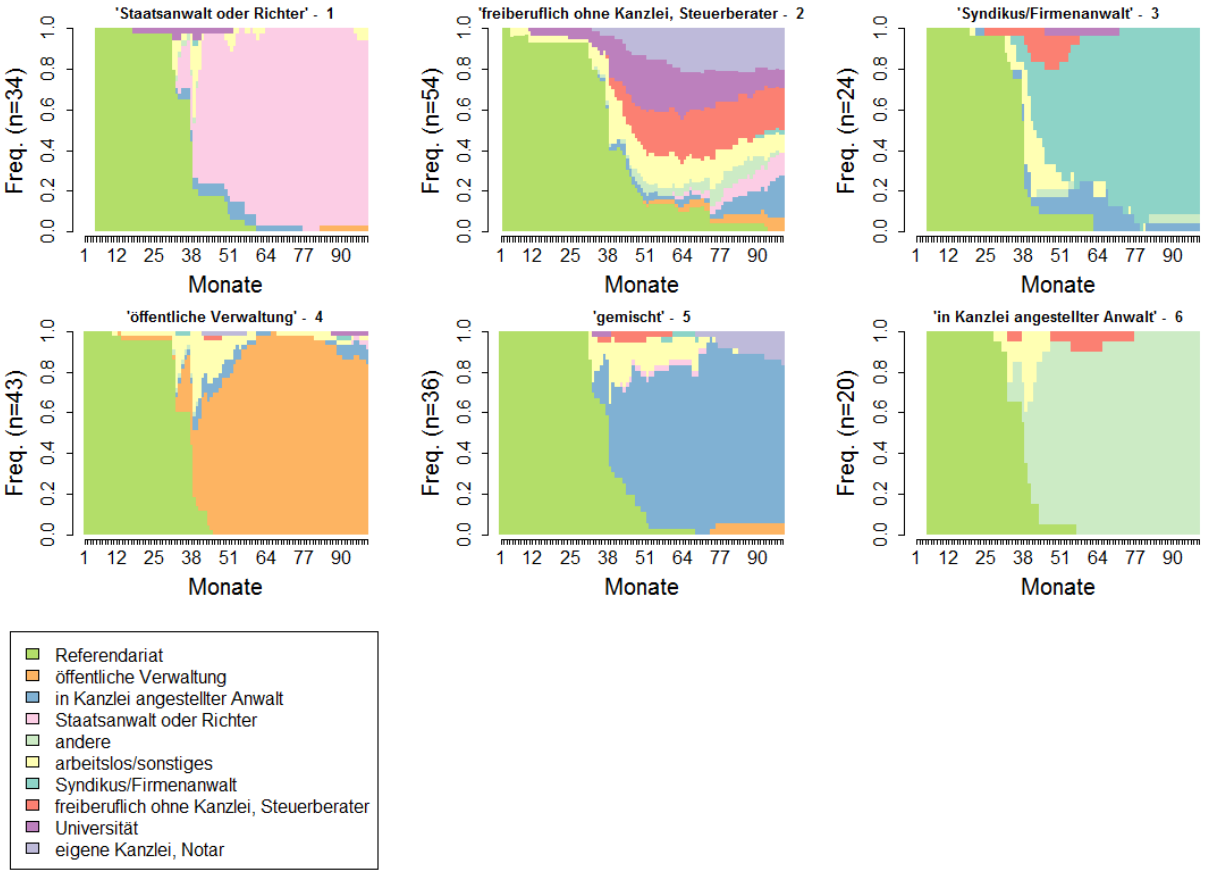

Abbildung A.11: Alternative Clusterlösung mit sechs Clustern (Relative Zustandshäufigkeiten)

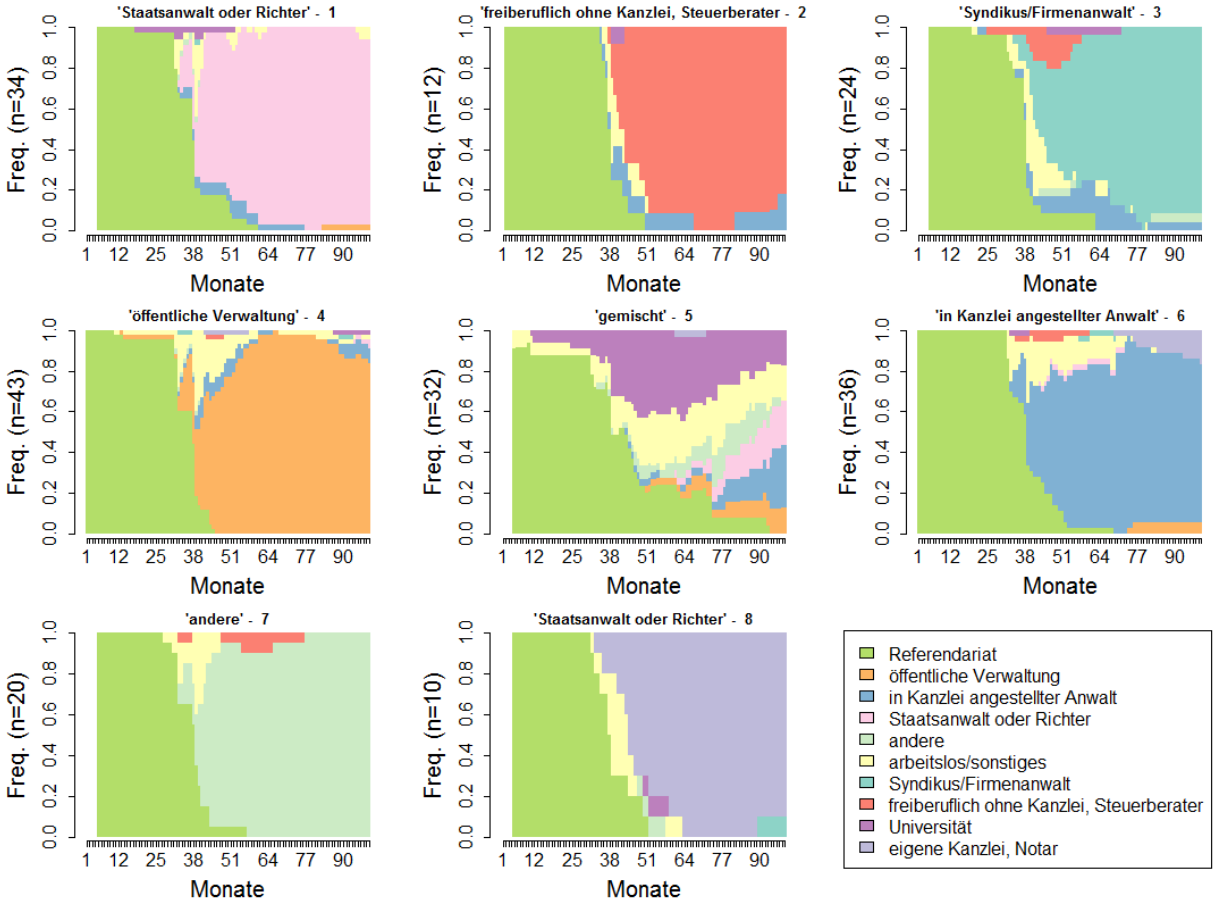

Abbildung A.12: Alternative Clusterlösung mit acht Clustern (Relative Zustandshäufigkeiten)

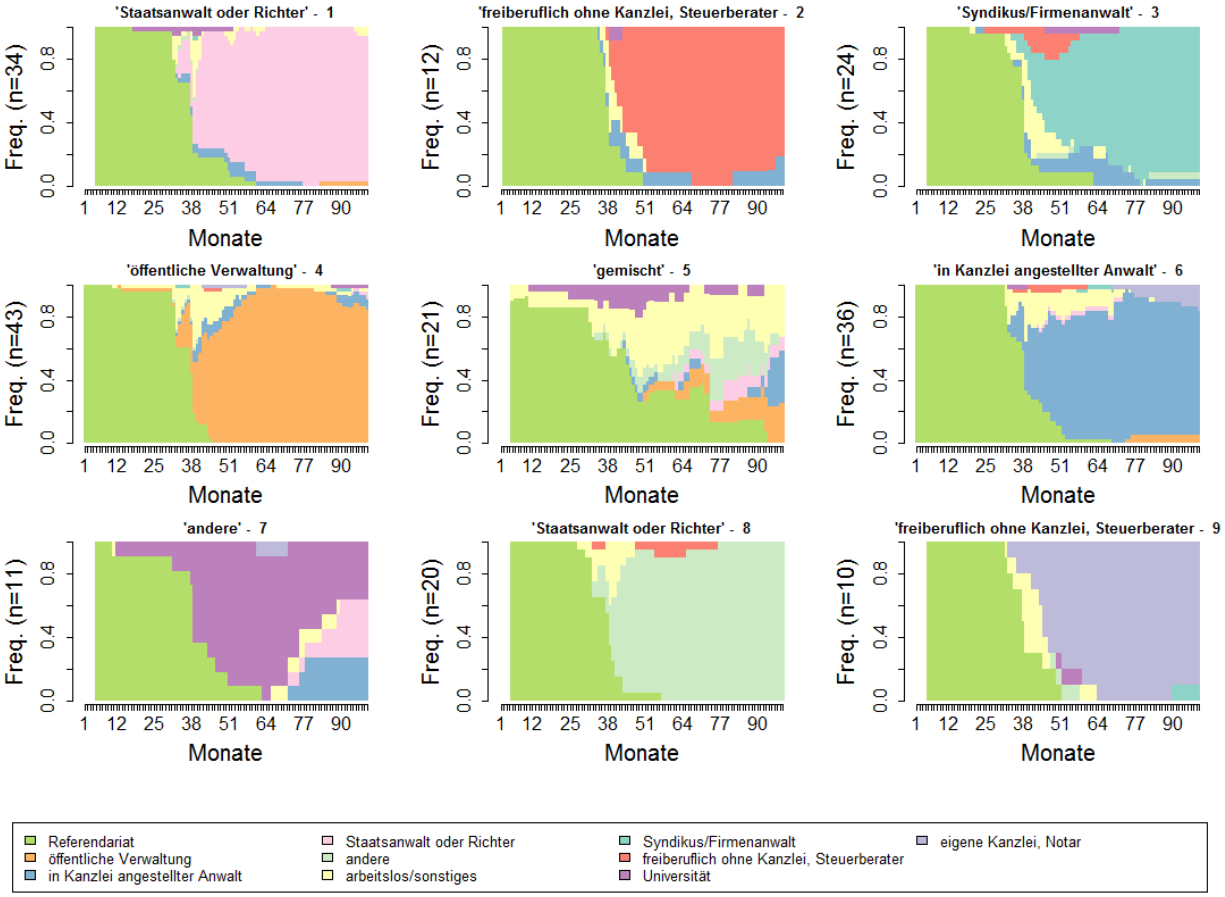

Abbildung A.13: Alternative Clusterlösung mit neun Clustern (Relative Zustandshäufigkeiten)

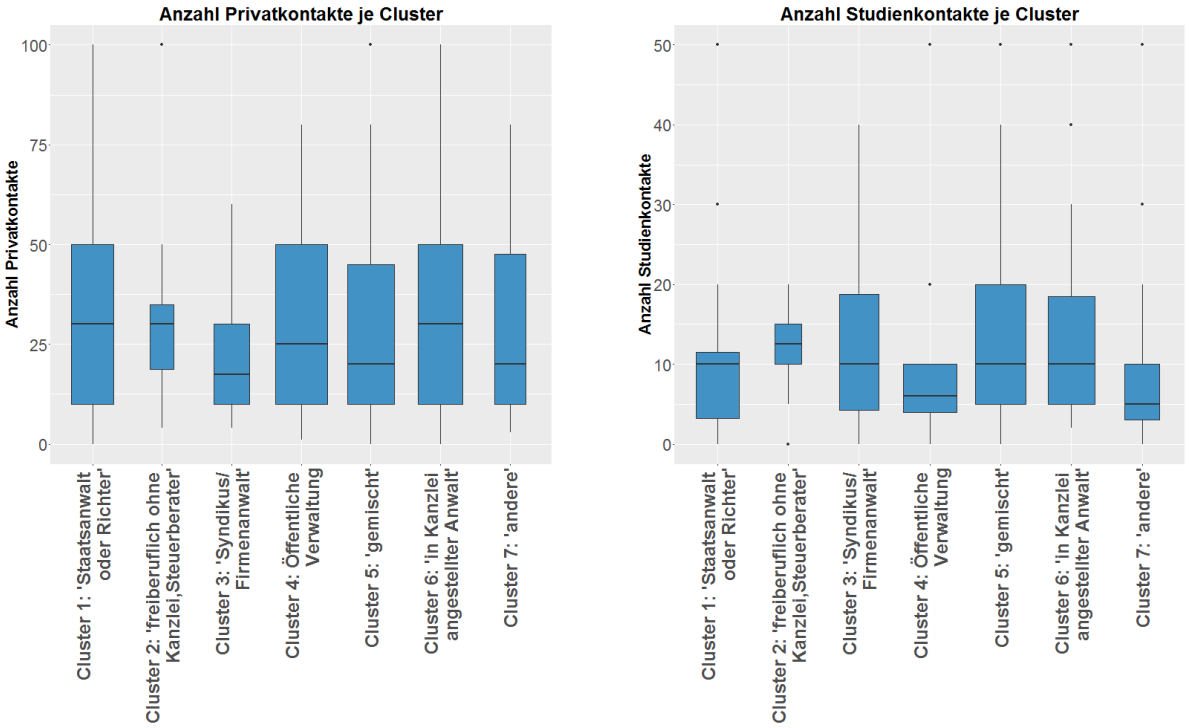

Abbildung A.14: Boxplots zu Privat- und Studiumskontakten je Cluster

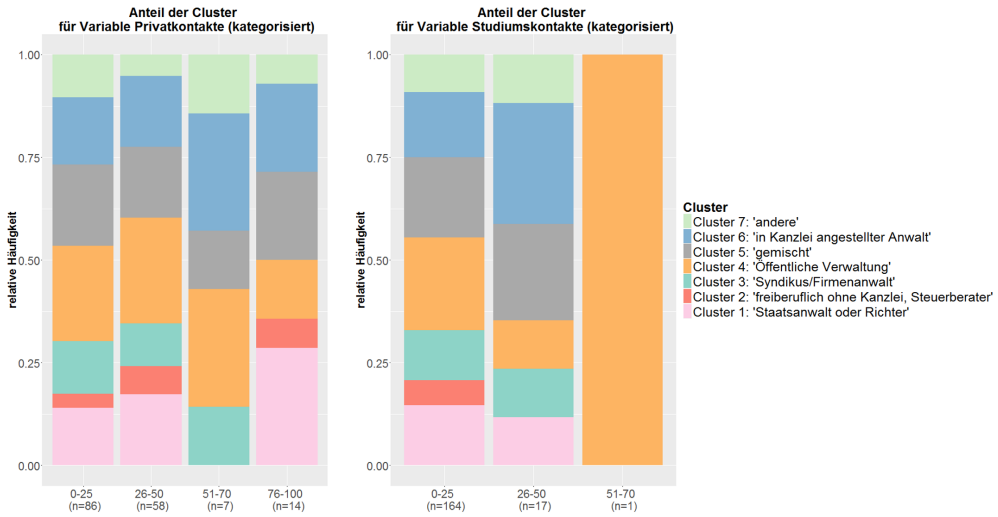

Abbildung A.15: Gestapelte Balkendiagramme zu den Anteilen der Cluster in den Kategorien der kategorisierten Variablen Privat- und Studiumskontakte

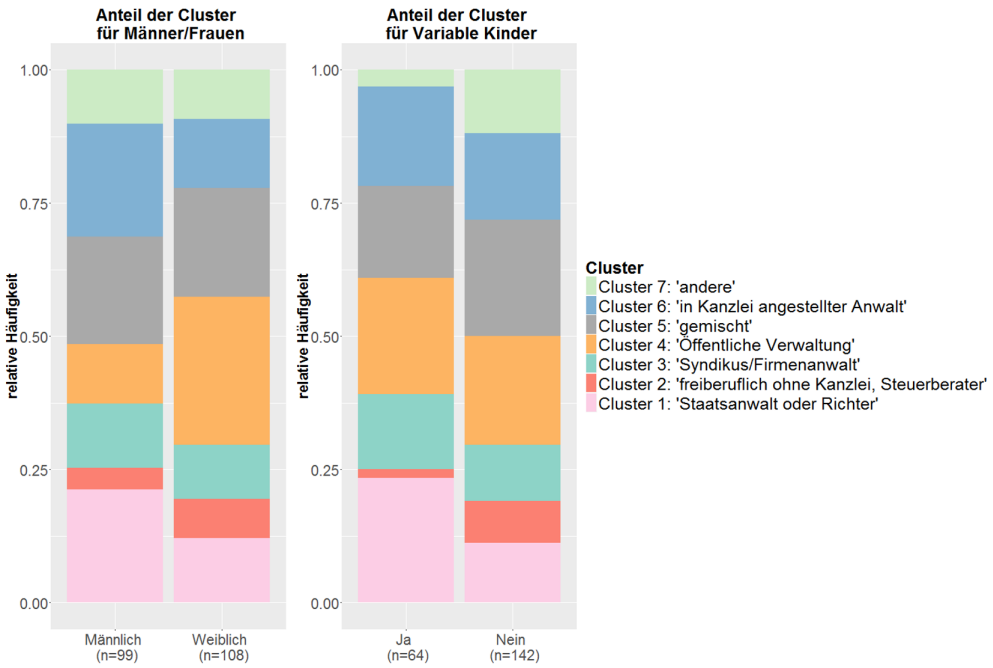

Abbildung A.16: Gestapelte Balkendiagramme zu den Anteilen der Cluster in den Kategorien der Variablen Geschlecht und Kinder

A.1.4 Regression

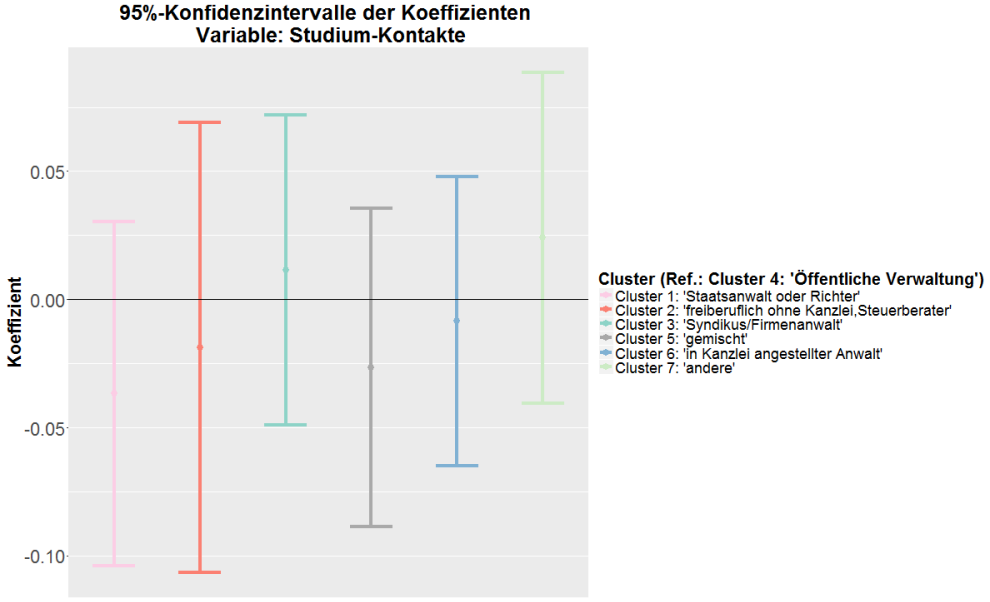

Abbildung A.17: Regressionskoeffizienten der Variable *Studium-Kontakt* des multinomialen Modells zur Clusteranalyse mit 95% -Konfidenzintervall

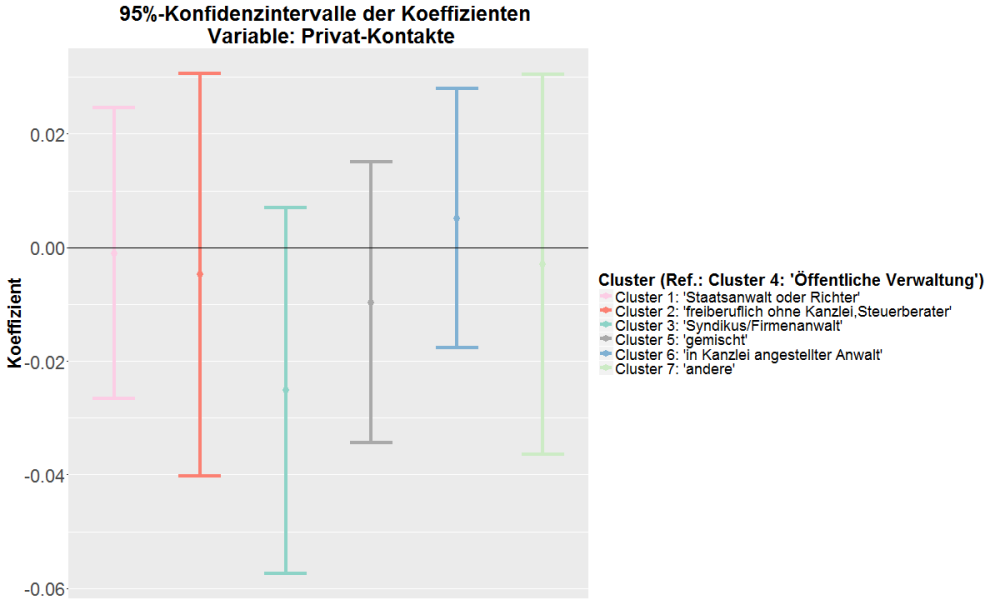

Abbildung A.18: Regressionskoeffizienten der Variable *Privat-Kontakt* des multinomialen Modells zur Clusteranalyse mit 95% -Konfidenzintervall

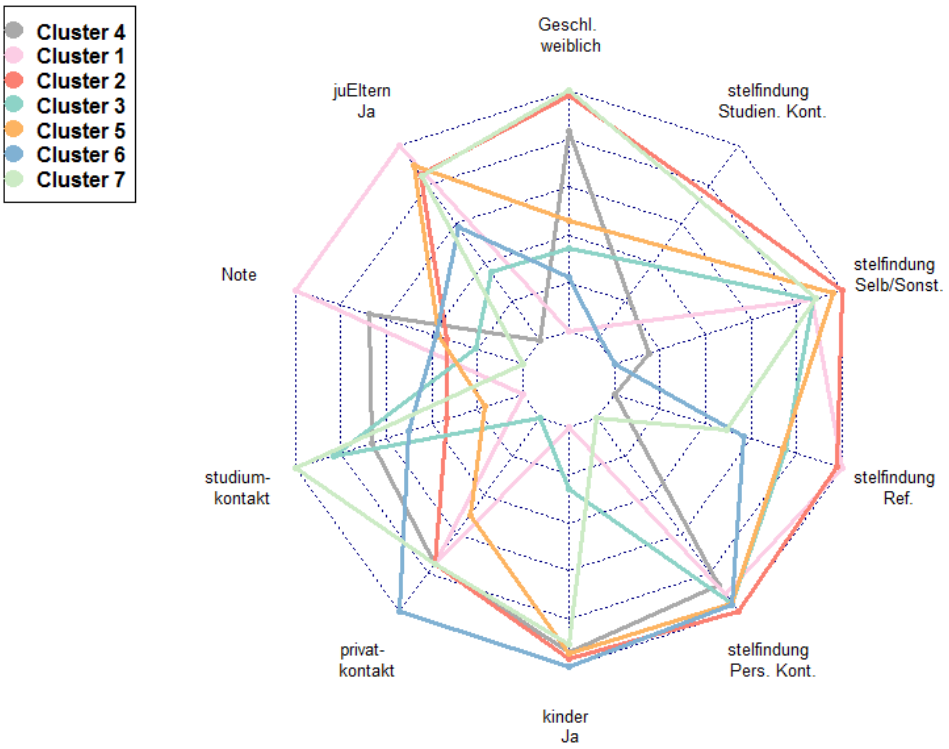

Abbildung A.19: Sternplot mit den Regressionskoeffizienten des multinomialen Modells zur Clusterzugehörigkeit

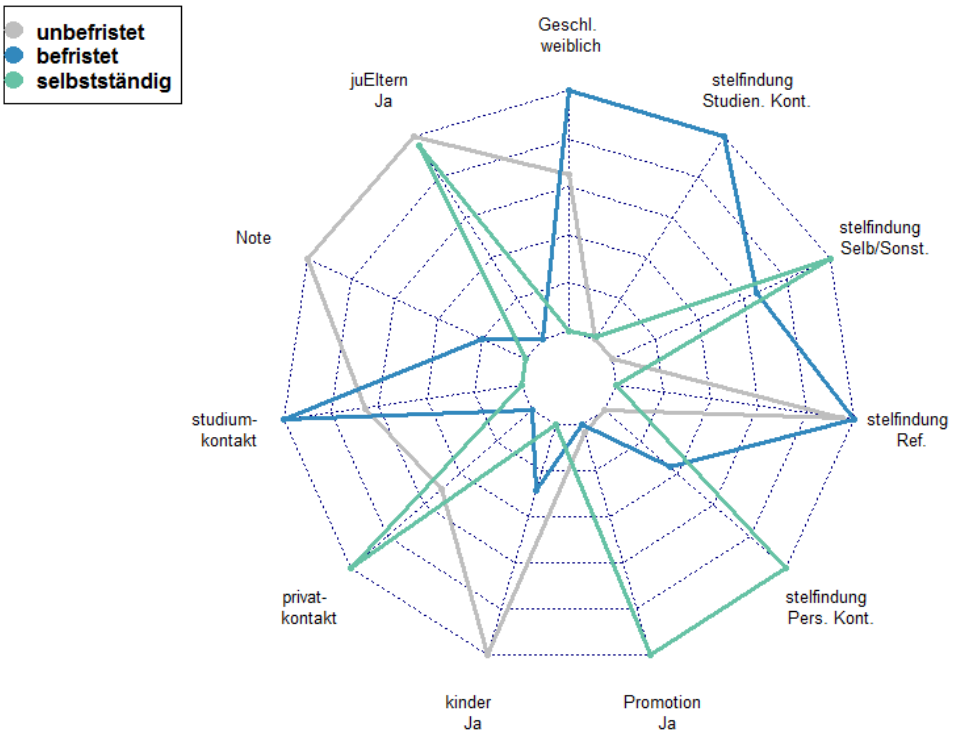

Abbildung A.20: Sternplot mit den Regressionskoeffizienten des multinomialen Modells zur Art der Befristung

A.2 Elektronischer Anhang

Die Analyse wurde mit R (Version 3.2.3) und dem R-Studio (Version 0.99.489) durchgeführt.

Die beigefügte DVD umfasst folgende Ordner mit den aufgelisteten Dateien:

- **R**:
  - **Datensätze**
  - "05061_Quer.dta“
  - "het12.dta“
  - "jura_zustand.dta“
  - "juraquer03.dta“
  - "quer8_final.dta“

  - **Workspace** (gespeicherte Version nach komplettem Code-Durchlauf)
  - **R-Code**:
  - Datenbearbeitung
  - Erstellung des Sequenz-Objekts
  - Clusteranalyse und -regression
  - Einkommensregression
  - Regression zur Art der Anstellung
  - Code zur Erstellung der Grafiken
- **Grafiken**: Dieser Ordner enthält alle Grafiken des Berichts (+ Anhang).

Ebenso enthält sie den Bericht in elektronischer Form.

Um einen problemlosen Durchlauf des Codes zu gewährleisten, müssen die R-Code-Dateien in der gespeicherten Reihenfolge ausgeführt werden $(1 - 9)$. Zudem müssen folgende Pakete installiert und für R zugreifbar sein:

- `TraMineR`
- `dplyr`
- `lazyeval`
- `fpc`

- `plotrix`
- `ggplot2`
- `labeling`
- `lmtest`
- `zoo`
- `fmsb`

Weitere benötigte Pakete sind bereits in R enthalten.

Damit es zu keinen Fehlern aufgrund von Überschreibung eines Pakets durch ein anderes kommt, sollten die Pakete so geladen werden, wie sie in den R-Code eingebunden sind!