Litzka, Leonie:

Resampling-basierte Variablenselektion: Wahl der Subsampling-Ratio

Bachelorarbeit, Sommersemester 2017

Fakultät für Mathematik, Informatik und Statistik

Ludwig-Maximilians-Universität München

https://doi.org/10.5282/ubm/epub.41009

BACHELORARBEIT

Resampling-basierte Variablenselektion: Wahl der Subsampling-Ratio

Verfasser: Leonie Friederike Litzka

Betreuer: Prof. Dr. Anne-Laure Boulesteix

Institut für Statistik
Institut für Medizinische Informationsverarbeitung, Biometrie und Epidemiologie

Ludwig-Maximilians-Universität München

14. Juli 2017

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

Von großem Interesse ist in vielen Fragestellungen der Zusammenhang zwischen einer Zielgröße und mehreren möglichen Einflussgrößen. Die Modellierung dieses Zusammenhangs erfolgt häufig mithilfe statistischer Regressionstechniken [Fahrmeir u. a., 2009]. Ein Kernthema ist dabei die Auswahl der Einflussgrößen. Oft sollen in das Modell nur die Einflussgrößen aufgenommen werden, die tatsächlich einen Einfluss auf die Zielgröße haben [de Bin u. a., 2016]. Hierfür gibt es mehrere etablierte Variablenselektionsverfahren, unter anderem sogenannte schrittweise Methoden [Royston u. Sauerbrei, 2008].
Ein großer Kritikpunkt dieser schrittweisen Methoden ist eine gewisse Instabilität ihrer Ergebnisse [Sauerbrei u. Schumacher, 1992]. In den letzten Jahren wurde die Resampling-basierte Variablenselektion als eine Möglichkeit vorgestellt, mit dieser Instabilität umzugehen [Sauerbrei u. Schumacher, 1992]. Das Resampling wurde dabei über das Ziehen von Bootstrap-Stichproben realisiert [Sauerbrei u. Schumacher, 1992]. Als Alternative für das Ziehen per Bootstrap wurde das sogenannte Subsampling ins Spiel gebracht. Dies entspricht dem Ziehen ohne Zurücklegen [de Bin u. a., 2016]. Es gibt Untersuchungen, die einige Vorteile des Subsamplings gegenüber dem Bootstrap aufzeigen [de Bin u. a., 2016; Janitza u. a., 2016; Rospleszcz u. a., 2016].
Es stellt sich die Frage, mit welchem Verhältnis, der sogenannten Ratio, bei Verwendung des Subsamplings bei Resampling-basierter Variablenselektion gezogen werden sollte [de Bin u. a., 2016]. Bisher wurde als Ratio häufig der Wert 0.632 verwendet. Dies entspricht im Mittel dem Anteil an per Bootstrap aus einem Datensatz gezogenen Beobachtungen, die dann auch teilweise mehrfach in der Bootstrap-Stichprobe vorkommen, an der Gesamtzahl der Beobachtungen des Datensatzes [de Bin u. a., 2016]. Es kommen allerdings auch andere Ratios in Frage.
Ziel dieser Arbeit ist es zu untersuchen, ob es eine Subsampling-Ratio gibt, deren Verwendung stabilere Ergebnisse der Resampling-basierten Variablenselektion liefert als die Verwendung anderer Ratios. Weitere Aspekte sind, ob die Wahl der Subsampling-Ratio bezüglich der Stabilität der Ergebnisse einer Resampling-basierten Variablenselektion von dem Stichprobenumfang des Datensatzes sowie von der Verteilung des Datensatzes abhängt.
Im Folgenden wird in Kapitel 2 anhand von Fachliteratur ein Überblick darüber gegeben, weshalb Variablenselektion durchgeführt wird und welche gängigen Methoden sich dafür etabliert haben. Anschließend wird die Resampling-basierte Variablenselektion in Kapitel 3 genauer vorgestellt. Dabei wird auch auf die zwei Varianten des Ziehens - per Bootstrap bzw. per Subsampling - eingegangen. Die für die Untersuchungen dieser Arbeit verwendeten Datensätze werden in Kapitel 4 vorgestellt. In Kapitel 5 wird anschließend basierend auf der Theorie, die in den vorherigen Abschnitten vorgestellt wurde, die Methodik der Untersuchungen dargelegt. Es werden zwei verschiedene Ansätze, die Ratios bezüglich der Stabilität zu beurteilen, eingeführt - die sogenannten Diskordanzkriterien und die Betrachtung der Korrelation von Rankings. Zudem wird kurz auf die Implementierung in der Software R [R Core Team, 2014] eingegangen. Die Ergebnisse der Untersuchungen werden schließlich in Kapitel 6 präsentiert. Zum Schluss wird in Kapitel 7 noch ein abschließendes Fazit gezogen sowie ein Ausblick gegeben.

2 Variablenselektion

Die Variablenselektion ist ein wesentlicher Schritt in der Modellbildung [de Bin u. a., 2016]. In Studien werden häufig viele Variablen, die möglicherweise einen Einfluss auf die interessierende Zielgröße haben könnten, erhoben. Anschließend sollen diejenigen Variablen identifiziert werden, die tatsächlich Einfluss auf die Zielgröße haben [de Bin u. a., 2016]. Im folgenden Kapitel wird die Variablenselektion anhand bereits existierender Literatur motiviert.

2.1 Motivation der Variablenselektion

Der Modellbildung liegen zwei wesentliche Ziele zugrunde, zum einen das der Vorhersage, zum anderen das der Erklärung der Zusammenhänge zwischen Zielgröße und Einflussgrößen [Sauerbrei u. a., 2007].
Laut Sauerbrei u. a. [2015] fittet ein gutes Modell die Daten adäquat gut, ist nicht zu komplex und sollte akkurate Vorhersagen für neue Daten liefern. Modellselektions-Kriterien sollten also eine Balance zwischen der Anpassung an die Daten und Sparsamkeit liefern [Sauerbrei u. a., 2015]. Der Zweck einer Studie beeinflusst stark, welches oben genannte Ziel einem wichtiger ist - das der Vorhersage oder das der Erklärung der Zusammenhänge. Falls ein Modell mit guter Vorhersagegüte angestrebt wird, ist ein komplexeres Modell, in dem auch Kovariablen mit eher schwachem Einfluss auf die Zielgröße enthalten sind, akzeptabel [Sauerbrei u. a., 2015]. Ist das Ziel jedoch ein Modell, das alle Größen mit einem Einfluss auf die Zielgröße identifiziert, also die Zusammenhänge zwischen Ziel- und Einflussgrößen erklären soll, könnte es laut Sauerbrei u. a. [2015] sein, dass ein einfacheres, weniger komplexes Modell bevorzugt wird. Dann nämlich ist die Vorhersagegüte eher zweitrangig, stattdessen gewinnen die Generalisierbarkeit und praktische Anwendbarkeit an Bedeutung. Bei dieser Art von Modellen ist die Modellstabilität von besonderem Interesse [Sauerbrei u. a., 2015], welche in dieser Arbeit genauer betrachtet werden soll. Aus diesem Grund wird hier der Fokus auf die zweite Gruppe von Modellen, die der Erklärung der Zusammenhänge dienen sollen, gelegt.
Fahrmeir u. a. [2009] schreiben hierzu: „Will man zwischen mehreren konkurrierenden statistischen Modellen mit verschiedenen Prädiktoren und Parametern auswählen, muss ein Kompromiss zwischen möglichst guter Datenanpassung und zu großer Modellkomplexität, d.h. einer hohen effektiven Anzahl von Parametern getroffen werden. So wird etwa in linearen Regressionsmodellen das Bestimmtheitsmaß R^2 durch Einbeziehen zusätzlicher Kovariablen, Interaktionen etc. immer weiter erhöht, jedoch ist dies meist mit einer Überanpassung (overfitting) an den vorliegenden (Lern-) Datensatz verbunden und geht mit einem Verlust an Prognosefähigkeit und der Generalisierbarkeit für neue Daten einher.“ [Fahrmeir u. a., 2009, S.477]
Miller [1984] nennt mehrere Gründe, weshalb nur ein paar und nicht alle der zur Verfügung stehenden Kovariablen ins Modell aufgenommen werden sollten. Zum einen werde durch die Reduzierung der Kovariablen eine Kostenersparnis des Aufwands beim Schätzen und Vorhersagen gewonnen [Miller, 1984]. Außerdem würden genaue Vorhersagen durch das Entfernen uninformativer Variablen erhalten werden. Des Weiteren werde eine sparsame Beschreibung des multivariaten Datensatzes erreicht. Zusätzlich ergäben sich kleine Stan-

dardfehler der Parameterschätzer, vor allem wenn einige Einflussgrößen hoch korreliert seien [Miller, 1984].
Royston u. Sauerbrei [2008] merken an, dass, bevor bestimmte Verfahren zur Variablenselektion durchgeführt werden, einige Dinge bezüglich des Modellbildungs-Prozesses bedacht werden sollten. Dazu zählen die Definition der möglichen Einflussvariablen sowie die Wahl der Kodierung der kategorialen Variablen. Des Weiteren sollte laut Royston u. Sauerbrei [2008] ein Umgang mit fehlenden Werten gewählt worden sein. Eine wichtige Entscheidung ist auch die Wahl der Modellklasse. Können die Daten beispielsweise mithilfe eines generalisierten linearen Modells modelliert werden, müssen eine passende Verteilung der Zielgröße und eine geeignete Link-Funktion ausgewählt werden. Diese und andere Vorüberlegungen haben entscheidenden Einfluss auf das später resultierende Modell [Royston u. Sauerbrei, 2008].
Nun sollen im nächsten Abschnitt einige populäre Methoden der Variablenselektion vorgestellt werden.

2.2 Methoden der Variablenselektion

Die Methoden zur Variablenselektion können laut Royston u. Sauerbrei [2008] in zwei große Bereiche eingeteilt werden. Einen Bereich bilden die schrittweisen Methoden, den anderen die sogenannten All-Subset Strategien [Royston u. Sauerbrei, 2008].
Als Alternative zu einer Variablenselektion gibt es beispielsweise Shrinkage-Methoden wie die Ridge-Regression, welche alle Kovariablen verwendet [Miller, 2002].

2.2.1 All-Subset-Strategien

Es wird mit der Vorstellung der All-Subset-Strategien begonnen. Stehen p Kovariablen zur Auswahl, die möglicherweise einen Effekt auf die Zielgröße haben, gibt es 2^p verschiedene mögliche Modelle, um den Zusammenhang zwischen der Zielgröße und den Einflussgrößen zu beschreiben [Royston u. Sauerbrei, 2008]. Für jede Kovariable muss entschieden werden, ob sie im Modell enthalten ist oder nicht, wodurch sich die 2^p möglichen Modelle ergeben.
Für jedes dieser 2^p Modelle wird bei den All-Subset-Strategien ein Modellwahlkriterium berechnet. Anschließend wird das Modell mit dem besten Kriterium ausgewählt [Royston u. Sauerbrei, 2008]. Um das Kriterium zu berechnen, müssen die in Frage kommenden Modelle geschätzt werden. Die zwei populärsten Kriterien sind laut Royston u. Sauerbrei [2008] das AIC und das BIC. Mithilfe dieser Kriterien werden die Modelle zum einen bezüglich der Güte des Fits und zum anderen bezüglich der Komplexität des Modells verglichen [Royston u. Sauerbrei, 2008].
Da in dieser Arbeit generalisierte lineare Modelle und Cox-Modelle für Überlebenszeitdaten verwendet werden, werden im Folgenden die Formeln des AIC und BIC für diese Modellarten kurz dargestellt. Die zugehörigen Formeln stammen aus Fahrmeir u. a. [2009], jedoch ist ihre Notation dieser Arbeit angepasst. AIC und BIC können für generalisierte lineare Modelle, die mithilfe der Maximum-Likelihood-Methode geschätzt werden, wie folgt berechnet werden. Das Akaike Informationskriterium (AIC) ist allgemein definiert als

$$AIC = -2l(\hat{\beta}) + 2r \tag{1}$$

[Fahrmeir u. a., 2009]. $l(\hat{\beta})$ ist der Wert der log-Likelihood an der Stelle des ML-Schätzers $\hat{\beta}$. Mit diesem Term wird die Modellanpassung beschrieben, die nun noch durch einen Term ergänzt wird, der eine „Überanpassung an den Datensatz durch Bestrafung zu hoher Komplexität, d.h. einer zu hohen (effektiven) Anzahl von Parametern“ [Fahrmeir u. a., 2009, S.477], verhindert. r ist die Anzahl der Parameter im Modell, was nicht zwingend der Anzahl der Kovariablen entsprechen muss, sondern der Anzahl der Modellparameter, also inklusive eines Intercepts und eventueller Dummyvariablen [Fahrmeir u. a., 2009].
Das Bayes'sche Informationskriterium (BIC) ist dem AIC formal sehr ähnlich [Fahrmeir u. a., 2009]. Die Formel lautet

$$BIC = -2l(\hat{\beta}) + log(n)r \tag{2}$$

[Fahrmeir u. a., 2009]. Das BIC bestraft ab einer Anzahl von $n = 8$ Beobachtungen die Komplexität eines Modells stärker als das AIC, weshalb das „BIC in der Regel weniger komplexe Modelle“ [Fahrmeir u. a., 2009, S.489] auswählt als das AIC. Es werden jeweils die Modelle mit dem kleinsten AIC- oder BIC-Wert ausgewählt.
Wird statt eines generalisierten linearen Modells ein Cox-Modell für Überlebenszeitdaten geschätzt, so werden AIC und BIC laut Moore [2016] über dieselben Formeln berechnet wie bei einem generalisiertem linearen Modell. Der Unterschied liegt darin, dass das Cox-Modell über Partielle Maximum-Likelihood-Schätzung geschätzt wird, weshalb bei einem Cox-Modell in Formel (1) und Formel (2) $l(\hat{\beta})$ den Wert der partiellen log-Likelihood darstellt [Moore, 2016]. Bei der Berechnung des BIC kann zusätzlich die Stichprobengröße n durch die Anzahl an Events ersetzt werden [Royston u. Sauerbrei, 2008].
Je mehr Kovariablen zur Auswahl stehen, desto größer wird die Anzahl der möglichen Modelle. Für $r = 10$ Kovariablen ergibt sich beispielsweise die Anzahl an zu schätzenden Modellen als $2^r = 2^{10} = 1024$. Aus diesem Grund sind All-Subset-Strategien nur für kleine Anzahlen an Kovariablen sinnvoll durchführbar, ansonsten verdoppeln sich die Kosten der Berechnung mit jeder zusätzlichen Kovariable [Miller, 2002]. Eine Alternative bieten die schrittweisen Methoden, die im folgenden Abschnitt näher betrachtet werden.

2.2.2 Schrittweise Methoden

Wichtige schrittweise Methoden sind Backward Elimination, Forward Selection und Stepwise Regression [Royston u. Sauerbrei, 2008]. Sie werden in diesem Abschnitt anhand von Royston u. Sauerbrei [2008] näher erläutert.
Alle Methoden können entweder auf der Basis von Hypothesentests durchgeführt werden oder mithilfe von Modellwahlkriterien wie dem AIC oder dem BIC [Royston u. Sauerbrei, 2008], die im vorherigen Abschnitt näher beschrieben sind. Es werden beide Vorgehensweisen vorgestellt.

Die Backward Elimination beginnt laut Royston u. Sauerbrei [2008] mit dem sogenannten vollen Modell, in dem alle Kovariablen enthalten sind. Im ersten Schritt wird für jede der p Kovariablen geprüft, ob ihre Entfernung aus dem Modell eine Modellverbesserung darstellen würde [Royston u. Sauerbrei, 2008]. Dies geschieht entweder, wie oben erwähnt, über Hypothesentests oder Modellwahlkriterien.
Bei der Verwendung von Hypothesentests wird ein nominales Signifikanzniveau α festgelegt, um Variablen aus dem Modell zu entfernen [Royston u. Sauerbrei, 2008]. Royston u.

Sauerbrei [2008] erklären, dass nun die Variable, deren p-Wert am wenigsten signifikant ist und gleichzeitig größer gleich dem nominalen Signifikanzniveau ist, aus dem Modell entfernt wird. Sollten die p-Werte aller Variablen kleiner α sein, so wird keine Variable entfernt, und die Backward Elimination ist mit diesem Schritt beendet [Royston u. Sauerbrei, 2008].
In den folgenden Schritten wird das gleiche Vorgehen auf der Basis des im vorherigen Schritt gewählten Modells ausgeführt. Wie schon im Schritt 1 beschrieben, endet die Backward Elimination mit dem Schritt, in dem kein p-Wert mehr größer gleich dem zuvor festgelegten Niveau α ist [Royston u. Sauerbrei, 2008].
Anstatt die Entscheidung, ob bzw. welche Variable in dem aktuellen Schritt aus dem Modell entfernt werden sollte, mithilfe der p-Werte und einem vorab festgelegten Signifikanzniveau α durchzuführen, können auch Modellwahlkriterien wie das AIC oder BIC verwendet werden [Royston u. Sauerbrei, 2008]. In diesem Fall wird die Variable aus dem Modell genommen, deren Entfernen das Modellwahlkriterium am meisten verbessert, also im Falle von AIC oder BIC am meisten reduziert [Fahrmeir u. a., 2009]. Sollte in einem Schritt bei jeder noch im Modell vorhandenen Variable das Entfernen dieser aus dem Modell das Kriterium nicht mehr verbessern, so wird keine Variable mehr entfernt, und die Backward Elimination ist beendet [Royston u. Sauerbrei, 2008].

Die Forward Selection beginnt im Gegensatz zur Backward Elimination mit dem sogenannten Intercept-Modell [Royston u. Sauerbrei, 2008]. In diesem sind keine Kovariablen enthalten. Der einzige zu schätzende Parameter ist der Intercept. Im ersten Schritt wird für jede der p Kovariablen geprüft, ob ihre Aufnahme in das Modell eine Modellverbesserung darstellen würde [Royston u. Sauerbrei, 2008]. Das Prinzip ist das gleiche wie bei der Backward Elimination.
Bei der Verwendung von Hypothesentests wird die Variable, deren p-Wert am kleinsten ist und gleichzeitig kleiner gleich α ist, in das Modell aufgenommen [Royston u. Sauerbrei, 2008]. Falls keine der Variablen einen p-Wert kleiner gleich α hat, so wird keine Variable in das Modell aufgenommen, und die Forward Selection ist mit diesem Schritt beendet [Royston u. Sauerbrei, 2008].
In den folgenden Schritten wird das gleiche Vorgehen auf der Basis des im vorherigen Schritt gewählten Modells ausgeführt. Die Forward Selection endet mit dem Schritt, in dem kein p-Wert mehr kleiner gleich dem zuvor festgelegten Niveau α ist [Royston u. Sauerbrei, 2008].
Auch hier kann die Entscheidung über die Aufnahme einer Variable mithilfe eines Modellwahlkriteriums wie dem AIC oder BIC geschehen [Royston u. Sauerbrei, 2008]. Es wird dann in jedem Schritt die Variable in das Modell aufgenommen, deren Aufnahme das AIC oder BIC am meisten verbessert [Fahrmeir u. a., 2009]. Sollte in einem Schritt bei keiner Variable deren Aufnahme in das Modell das Kriterium verbessern, so wird keine Variable mehr aufgenommen, und die Forward Selection ist beendet.
Die Stepwise Regression oder auch Stepwise Selection genannt ist eine „Kombination“ [Fahrmeir u. a., 2009, S.164] aus Forward Selection und Backward Elimination. Hier gibt es die Möglichkeit, schon entfernte Variablen wieder in das Modell aufzunehmen bzw. schon in das Modell aufgenommene Variablen wieder zu entfernen [Fahrmeir u. a., 2009].

2.2.3 Schwächen der schrittweisen Methoden

Die schrittweisen Methoden sind für Datensätze mit vielen Variablen besser geeignet als die All-Subset-Strategien, da jene ab einer gewissen Anzahl an Variablen immer größer werdende Berechnungskosten haben [Miller, 2002]. Laut Fahrmeir u. a. [2009] muss allerdings beachtet werden, dass die schrittweisen Methoden nicht unbedingt das beste Modell identifizieren, aber meist ein sehr gutes. Sauerbrei u. Schumacher [1992] stellen zudem fest, dass die schrittweisen Methoden immer mit einem unkorrekten Modell starten. Entweder beginnen sie mit dem Modell, das nur den Intercept enthält, oder mit dem vollen Modell [Sauerbrei u. Schumacher, 1992].
Copas u. Long [1991] fügen einen weiteren Gesichtspunkt, der beachtet werden sollte, hinzu. Die Modellwahl beruht auf den Schätzungen der Koeffizienten und nicht auf den wahren Werten der Koeffizienten. Deshalb hat eine Variable eine höhere Wahrscheinlichkeit in das Modell aufgenommen zu werden, wenn der Betrag ihres Koeffizienten überstatt unterschätzt ist [Copas u. Long, 1991]. Daher sind die Regressionskoeffizienten der selektierten Variablen nach oben verzerrt, weshalb der durch die Regression erklärte Anteil der Streuung ebenfalls nach oben verzerrt, also zu groß ist [Copas u. Long, 1991]. Damit geht einher, dass die Residualstreuung, die die nicht selektierten Variablen beinhaltet, nach unten verzerrt, also zu klein ist [Copas u. Long, 1991]. Copas u. Long [1991] bieten eine Möglichkeit, die Residualstreuung nach dem Durchführen einer Variablenselektion approximativ unverzerrt zu schätzen.
Henderson u. Velleman [1981] empfehlen, sich nicht nur auf automatische Prozeduren wie die oben vorgestellten schrittweise Methoden zur Variablenselektion zu verlassen. Henderson u. Velleman [1981] begründen dies damit, dass ein Datenanalyst oder Experte im Anwendungsgebiet Wissen und Informationen besitzt, die der Computer nicht hat. Es sollte also gemeinsam mit dem Computer gearbeitet werden, um manche Entscheidungen besser selbst zu treffen, anstatt sie dem Rechner zu überlassen. Ebenfalls stellen Henderson u. Velleman [1981] fest, dass aus automatischen Prozeduren entstandene Modelle auch schwierig interpretierbar sein können.

Die für diese Arbeit wichtigste Schwäche ist jedoch die der Modellinstabilität. Das letztendlich durch ein Verfahren der Variablenselektion ausgewählte Modell ist laut Royston u. Sauerbrei [2008] instabil in der Hinsicht, dass es sensitiv gegenüber kleinen Änderungen in den Daten ist. Diese Änderungen können die selektierten Variablen des Modells wieder ändern [Royston u. Sauerbrei, 2008].
Die Modellstabilität bezeichnet also eine Replikations-Stabilität in Bezug auf die Aufnahme von Variablen in ein Modell [Sauerbrei u. Schumacher, 1992]. Replikations-Stabilität bedeutet laut Gifi [1990], dass die Ergebnisse eines Experiments, das unter denselben Konditionen auf neuen Daten durchgeführt wird, sich nicht dramatisch ändern.
Kovariablen eines Datensatzes, deren p-Werte sehr gering sind, sind in der Hinsicht stabil, dass sie mit hoher Wahrscheinlichkeit in ähnlichen Datensätzen auch in das Modell gewählt würden [Royston u. Sauerbrei, 2008]. Für weniger signifikante Kovariablen beschreiben Royston u. Sauerbrei [2008], dass es eher eine Art Chance sei, in das Modell aufgenommen zu werden oder nicht. In der Realität existieren meist mehrere verschiedene Modelle, die die Daten ungefähr ähnlich gut fitten. Kleine Änderungen in den Daten können dann schon zu einem anderen Modell führen [Royston u. Sauerbrei, 2008].
Diese Instabilität der Ergebnisse von Variablenselektionsverfahren wird seit einigen Jah-

ren näher untersucht. Es wird versucht, zu einer höheren Modellstabilität zu gelangen. Welche Methoden hierzu entwickelt wurden bzw. werden, wird im Kapitel 3 ausführlich erläutert.
Im Folgenden werden noch zwei Beispiele für die Instabilität der Ergebnisse von Variablenselektion präsentiert.
Altman u. Andersen [1989] haben basierend auf 100 Bootstrap-Stichproben und auf diesen durchgeführten Variablenselektionen untersucht, wie häufig die Variablen in das Modell selektiert wurden. Der verwendete Datensatz enthielt 17 mögliche Kovariablen und hatte als Zielgröße die Überlebenszeit von Patienten mit PBC. Das Cox-Modell, das auf dem gesamten Datensatz geschätzt wurde, enthielt sechs Kovariablen. In den 100 Modellen, die auf den Bootstrap-Stichproben mithilfe von Stepwise Regression geschätzt wurden, waren alle 17 Variablen mindestens sechsmal in einem Modell enthalten. Hier zeigt sich die Unsicherheit der Ergebnisse der Variablenselektion. Die sechs Variablen, die in dem Modell basierend auf dem kompletten Datensatz enthalten waren, waren in den Modellen basierend auf den Bootstrap-Stichproben häufiger selektiert als die übrigen elf Kovariablen. Für genauere Information vergleiche Altman u. Andersen [1989].
Chen u. George [1985] haben ebenfalls basierend auf 100 Bootstrap-Stichproben und auf diesen durchgeführten Variablenselektionen untersucht, wie häufig die Variablen in das Modell selektiert wurden. Der verwendete Datensatz enthielt neun mögliche Einflussfaktoren, die Zielgröße war die Zeit von vollkommener Genesung bis zum ersten Rückfall von Kindern mit ALL. Auch hier wurde auf dem gesamten Datensatz ein Modell mithilfe von Stepwise Regression geschätzt, das sechs Kovariablen enthält. Anschließend wurden 100 Bootstrap-Stichproben gezogen und für jede dieser Stichproben ein Modell mit Stepwise Regression ermittelt. Die Anzahl an Kovariablen dieser Modelle schwankt zwischen zwei und acht Kovariablen. Auch hier zeigt sich die Unsicherheit der Ergebnisse der Variablenselektion. Für genauere Information vergleiche Chen u. George [1985].

3 Resampling-basierte Variablenselektion

Wie schon erwähnt, ist die Stabilität eines Modells im Sinne von Replikations-Stabilität in Bezug auf die Aufnahme von Variablen in das Modell zu verstehen [Sauerbrei u. Schumacher, 1992]. Oft gibt es laut Sauerbrei u. Schumacher [1992] mehrere Modelle, deren Fit ähnlich gut ist, die sich allerdings in ihren Kovariablen unterscheiden. Welches Modell von dem Variablenselektionsverfahren ausgewählt wird, hängt oft von den Charakteristiken einer kleinen Anzahl an Beobachtungen des Datensatzes ab. Kleine Änderungen in den Daten können dann zu einem anderen Modell führen [Sauerbrei u. a., 2015].
Resampling-basierte Variablenselektion wird nun mit dem Ziel angewandt, die Stabilität der Ergebnisse von Variablenselektionsverfahren zu verbessern [de Bin u. a., 2016]. Dies kann auf der Basis von Bootstrap-Stichproben mit Zurücklegen oder auf der Basis des sogenannten Subsamplings ohne Zurücklegen geschehen.

Wie de Bin u. a. [2016] erklären, beruht das Prinzip darauf, mehrere sogenannte Pseudostichproben aus dem Originaldatensatz zu erzeugen. Auf jede dieser Pseudostichproben wird dann ein Variablenselektionsverfahren angewandt. Aufgrund der Unterschiede in den Datensätzen der Pseudostichproben unterscheiden sich die auf den Pseudostichproben ge-

schätzten Modelle [de Bin u. a., 2016]. Die Pseudostichproben realisieren also die kleinen Änderungen in den Daten. Nun kann berechnet werden, wie oft die zur Auswahl stehenden Kovariablen in den Modellen der Pseudostichproben enthalten sind. Diese relativen Häufigkeiten werden auch als Inklusionshäufigkeiten bezeichnet. Anhand dieser können nun die Kovariablen danach beurteilt werden, wie relevant sie für das Modell sind [de Bin u. a., 2016].
Wie de Bin u. a. [2016] darlegen, ist die Erwartung, dass Variablen, die Einfluss auf die interessierende Zielgröße haben und demnach für das Modell relevant sind, bei allen oder zumindest fast allen Pseudostichproben im Modell enthalten sind. Von Variablen hingegen, die keinen Einfluss haben, wird angenommen, diese nicht in den Modellen der Pseudostichproben zu finden [de Bin u. a., 2016]. Das entspricht bei den relevanten Variablen jeweils einer Inklusionshäufigkeit von 1, bei den Variablen ohne Einfluss einer Inklusionshäufigkeit von 0. de Bin u. a. [2016] machen jedoch klar, dass in der Realität nicht so klare Ergebnisse zu erwarten sind. Zum einen gibt es Variablen, die zwar einen Einfluss auf die Zielgröße haben, der jedoch schwach ist [Royston u. Sauerbrei, 2008]. Für solche Variablen basiert es im Prinzip auf einer gewissen Chance, ob sie in das Modell der entsprechenden Pseudostichprobe selektiert werden oder nicht [Royston u. Sauerbrei, 2008]. Zum anderen können Variablen laut de Bin u. a. [2016] ohne jeglichen Einfluss auf die Zielgröße aufgrund von Type-I-Fehlern irrtümlich selektiert werden. Außerdem müssen die Korrelationen zwischen den Variablen beachtet werden, da diese Korrelationen das Ergebnis der Variablenselektionsverfahren beeinflussen und damit auch die Inklusionshäufigkeiten [de Bin u. a., 2016]. Im Extremfall von zwei hoch korrelierten Variablen kann es sein, dass immer eine von beiden in das Modell selektiert wird, aber beide nur eine Inklusionshäufigkeit von 50% haben [Sauerbrei u. a., 2011]. In dieser Arbeit werden die Korrelationen zwischen Kovariablen allerdings nicht weiter berücksichtigt.

3.1 Nonparametrischer Bootstrap

Das oben genannte Ziehen der Pseudostichproben kann zum einen über das Verfahren des Nonparametrischen Bootstrap geschehen. Die Idee des Nonparametrischen Bootstraps wird in Efron u. Tibshirani [1986] ausführlich erklärt.
Laut Miller [2002] soll auf der Basis eines Datensatzes mit Stichprobenumfang n eine Statistik θ geschätzt werden. Der Datensatz, der eine Stichprobe aus der kompletten Population darstellt, wird nun selbst als die komplette Population behandelt, indem mit Zurücklegen mehrere Stichproben der Größe n aus dem Datensatz gezogen werden [Miller, 2002]. Für jede der so entstandenen sogenannten Bootstrap-Stichproben wird der Schätzer $\hat{\theta}$ ermittelt und anschließend die empirische Verteilung der Statistik θ berechnet [Miller, 2002]. Wie Miller [2002] feststellt, werden vor dem Durchführen des Bootstraps keine Verteilungsannahmen getätigt. Das Verfahren arbeitet komplett datensatzbasiert. Wird das Ziehen einer Bootstrap-Stichprobe betrachtet, liegt die Wahrscheinlichkeit, dass eine Beobachtung des Datensatzes nicht der Bootstrap-Stichprobe landet, laut Miller [2002] bei $(1 - 1/n)^n$, was approximativ $e^{-1} = 0.3679$ entspricht. Demnach landen etwa 36 bis 37% der Beobachtungen nicht in der Bootstrap-Stichprobe [Miller, 2002].
Im Zusammenhang mit der Resampling-basierten Variablenselektion werden also die Pseudostichproben über das Bootstrap-Verfahren gezogen, die Stichprobengröße entspricht der

Größe des Datensatzes [de Bin u. a., 2016]. Auf Basis jeder Pseudostichprobe wird ein Variablenselektionsverfahren durchgeführt und anschließend für jede Variable die Inklusionshäufigkeit berechnet [Sauerbrei u. Schumacher, 1992]. Aufgrund des Ziehens per Bootstrap sollte laut Sauerbrei u. Schumacher [1992] jede Pseudostichprobe die den Daten unterliegende Struktur widerspiegeln, da die Bootstrap-Stichproben zufällig gezogen wurden. Deshalb kann die Inklusionshäufigkeit als Kriterium für die Relevanz der Variable für das Modell interpretiert werden [Sauerbrei u. Schumacher, 1992].

Es gibt einige Schwachpunkte des Anwendens einer Resampling-basierten Variablenselektion mittels Bootstrapping. Zum einen stellen Janitza u. a. [2016] fest, dass die p-Werte von statistischen Tests, die auf Bootstrap-Stichproben unter der Annahme, dass diese Stichproben die Originaldaten seien, durchgeführt werden, problematisch sind. Denn die Bootstrap-Stichproben werden, wie schon beschrieben, aus einer anderen Stichprobe durch Ziehen mit Zurücklegen generiert. Die p-Werte solcher Tests sind nicht valide und können deshalb laut Janitza u. a. [2016] nicht als p-Werte interpretiert werden, da die Typ-I-Fehler solcher Tests vergrößert sind. Im Kontext der Resampling-basierten Variablenselektion bedeutet das, dass die Tests im Zuge der Variablenselektionsverfahren zu kleine p-Werte aufweisen, d.h. zu viele Variablen in das Modell selektiert werden [Janitza u. a., 2016].
Auch de Bin u. a. [2016] haben diesen Punkt untersucht und das Ergebnis erhalten, dass Variablenselektionsverfahren basierend auf Bootstrap-Stichproben zu hohen Inklusionshäufigkeiten von Störvariablen führen. Die Inklusionshäufigkeiten sind in den Untersuchungen von de Bin u. a. [2016] bedeutend höher als die theoretisch bei einem Signifikanzniveau von $\alpha = 0.05$ aufgrund des Typ-I-Fehlers zu erwartende Häufigkeit 0.05. Zudem wurde festgestellt, dass dieses Problem nicht auftritt, wenn die Pseudostichproben statt über Bootstrap über Subsampling gezogen werden [de Bin u. a., 2016].
Rospleszcz u. a. [2016] haben festgestellt, dass bei Bootstrap-Stichproben bevorzugt kategoriale Variablen, die mehr als zwei Kategorien haben, im Vergleich zu metrischen Variablen ausgewählt werden. Insbesondere steigt die Inklusionshäufigkeit einer Kovariable deutlich mit der Anzahl ihrer Kategorien an [Rospleszcz u. a., 2016]. Dieser Effekt konnte bei der Ziehung der Pseudostichproben über Subsampling nicht beobachtet werden [Rospleszcz u. a., 2016]. Als Grund geben Rospleszcz u. a. [2016] an, dass die über Subsampling gezogenen Pseudostichproben echte Teilmengen des Datensatzes sind, sodass angenommen werden kann, dass sowohl die Pseudostichproben als auch der Datensatz von derselben unterliegenden Population gezogen sind, wo die Nullhypothese gilt.

3.2 Subsampling

Eine Alternative zum Ziehen der Pseudostichproben per Bootstrap stellt das sogenannte Subsampling dar. Das Subsampling ist auch unter dem Namen d-delete jackknife bekannt [Wu, 1986]. Im Gegensatz zum Bootstrap wird hier ohne Zurücklegen gezogen [de Bin u. a., 2016]. Wie in Wu [1986] beschrieben, entspricht das einem Entfernen von Beobachtungen aus dem Datensatz. Die Pseudostichproben haben einen kleineren Stichprobenumfang als der Datensatz, aus dem sie gezogen werden. Er hängt von der Ratio ab, mit der gezogen wird [de Bin u. a., 2016].

Das Ziehen der Pseudostichproben mittels Subsampling wird in neueren Untersuchungen als Alternative zum Bootstrap empfohlen, da einige der beim Bootstrap beobachteten Probleme beim Verwenden von Subsampling nicht auftreten [Rospleszcz u. a., 2016]. Laut Davison u. a. [2003] ist Subsampling valider als Bootstrap, insbesondere unter minimalen Bedingungen.
Die Wahl der Subsampling-Ratio, mit der die Pseudostichproben gezogen werden, ist ein wichtiger Parameter der Resampling-basierten Variablenselektion mittels Subsampling. Oft wurde in bisherigen Arbeiten als Ratio der Wert 0.632 gewählt, da dies im Mittel dem Anteil an aus dem Datensatz per Bootstrap gezogenen Beobachtungen, die dann teilweise mehrfach in der Bootstrap-Stichprobe vorkommen, entspricht [de Bin u. a., 2016]. Im Rahmen der Stability selection verwenden Meinshausen u. Bühlmann [2010] die Subsampling-Ratio 0.5. Es ist allerdings auch die Wahl einer anderen Ratio denkbar.
Ziel dieser Arbeit ist es, zu untersuchen, ob es eine bestimmte Subsampling-Ratio gibt, bei der die Ergebnisse der Resampling-basierten Variablenselektion stabiler im Sinne der Replikations-Stabilität sind als bei der Verwendung anderer Ratios. Bevor die Vorgehensweise der Untersuchung genau beschrieben wird, werden im folgenden Abschnitt zunächst die verwendeten Datensätze eingeführt.

4 Datensätze

Zur Untersuchung der Stabilität der Ergebnisse Resampling-basierter Variablenselektion basierend auf verschiedenen Subsampling-Ratios werden drei Datensätze aus dem medizinischen Kontext genutzt. Wie schon zuvor erwähnt, werden in allen Datensätzen Korrelationen zwischen den in Frage kommenden Kovariablen nicht untersucht.

4.1 Ozone

Der erste Datensatz stammt aus einer Studie von Ihorst u. a. [2004]. Untersucht wurden die mittel- und langfristigen Effekte von Ozon auf das Lungenwachstum von Kindern [de Bin u. a., 2016; Sauerbrei u. a., 2015]. In dieser Arbeit wird die Teilmenge der Daten, die auch in Sauerbrei u. a. [2015] verwendet wird, genutzt.
Die Zielgröße (FVC) ist metrisch und bezieht sich auf den Herbst 1997. Die Stichprobengröße beträgt $n = 496$ Kinder, der Datensatz enthält 24 Kovariablen, die möglicherweise einen Einfluss auf die Zielgröße haben. In den Kovariablen enthalten sind Angaben zum Alter der Kinder (ALTER) sowie zum Geschlecht (SEX), Geburtsgewicht (AGEBGEW), zu Allergien (ADHEU, FMILB, FTIER, FPOLL, FSPT) und zu Husten bzw. Atemproblemen (FSNIGHT, FSATEM, FSPFEI, FSHLAUF). Zudem existieren Variablen dazu, ob das Kind an einem Ort mit hohen Ozon-Werten lebt (HOCHOZON), ob eine mütterliche oder väterliche Atopie vorliegt (AMATOP, AVATOP), ob ein Ekzem diagnostiziert wurde (ADEKZ), ob das Kind juckende oder wässrige Augen hat (FSAUGE), und ob das Kind zuhause Tabakrauch ausgesetzt ist (ARAUCH). Zu einem pulmonalen Funktionstest gibt es Angaben über die Größe und das Gewicht des Kindes beim Test (FLGROSS, FLGEW), über den maximalen NO_2-Wert der letzten 24 Stunden vor dem Test (FNOH24), über den maximalen O_3-Wert der letzten 24 Stunden vor dem Test (FO3H24), über die Maximaltemperatur der letzten 24 Stunden vor dem Test (FTEH24) und über die Ge-

samtanzahl an Medikationen während des Tests (FLTOTMED) [Sauerbrei u. a., 2015]. Ein paar Kovariablen sind metrisch, die anderen binär. Der Datensatz enthält keine fehlenden Werte. Als volles Modell wird hier ein generalisiertes lineares Modell mit normalverteilter Zielgröße und der Identität als Linkfunktion geschätzt. Die Kovariablen werden alle wie in de Bin u. a. [2016] linear aufgenommen.

4.2 Glioma

Ebenfalls verwendet wird ein Datensatz aus Royston u. Sauerbrei [2008] zur Untersuchung der Überlebenszeit von Patienten, die an einem malignem Gliom erkrankt sind. Die Studie, für die der Datensatz erhoben wurde, ist eine randomisierte kontrollierte Studie zum Vergleich von zwei verschiedenen Chemotherapien.
Die Zielgröße ist die Überlebenszeit der Patienten. Der Stichprobenumfang ist $n = 411$, im Datensatz enthalten sind neben der Überlebenszeit ein Indikator für die Zensierung und zusätzlich 13 Kovariablen. In den Daten sind 274 Todesfälle, die sogenannten Events, vorhanden. Zu den Kovariablen gehören das Geschlecht (sex) und Alter (age) der Patienten. Des Weiteren gibt es die Zeit (time), die von den ersten Symptomen bis zur Diagnose vergangen ist, mit den Ausprägungen „kurz“ bzw. „lang“. Der Grad des Tumors (gradd), der Karnofsky Index (kard) und der Resektionstyp (surgd) sind ordinalskaliert mit je drei Kategorien. Sie werden jeweils durch zwei Dummyvariablen bei der Modellschätzung repräsentiert. Als Kodierung wird in Royston u. Sauerbrei [2008] die Splitkodierung verwendet. Mehrere binäre Variablen besitzen die Kategorien „ja“ und „nein“. Dazu zählen das Auftreten von Krämpfen (convul), Epilepsie (epi), Amnesie (amnesia), Organischem Psychosyndrom (ops) und Aphasie (aph) sowie die Gabe von Kortison (cort). Ob der Patient der Behandlungsgruppe mit der Standard-Therapie zugeteilt war oder die neue Therapie bekam, wird durch die binäre Variable trt angegeben. Das volle Modell dieses Datensatzes ist ein Cox-Modell, in dem alle Kovariablen enthalten sind.

4.3 PBC

Der PBC-Datensatz aus Royston u. Sauerbrei [2008] untersucht die Überlebenszeit von Patienten mit Primär Biliärer Zirrhose (auch Primär Biliäre Cholangitis). An der randomisierten kontrollierten Studie zum Vergleich von zwei Therapien nahmen $n = 312$ Patienten teil, von diesen starben 125. Eine zusätzliche Kohorte mit noch einmal 106 Patienten wird hier nicht mit betrachtet, da bei diesen Patienten nur sechs der 17 Kovariablen erhoben wurden. Die Zielgröße ist die Überlebenszeit der Patienten. Außerdem ist ein Indikator für die Zensierung in den Daten enthalten. Unter den Kovariablen sind die Behandlungsgruppe (trt) - es wurde entweder D-Penicillamin oder Placebo gegeben - und das Alter (age) der Patienten sowie das Geschlecht (sex). Außerdem wurden einige binäre Variablen mit den Kategorien „ja“ und „nein“ erhoben, sie betreffen das Auftreten von Aszites (asc), Hepatitis (hep) und sogenannten Gefäßspinnen (spider). Weitere metrisch gemessene Kovariablen erfassen Bilirubin (bil), Cholesterol (chol), Albumin (alb), den Kupfergehalt im Urin (cu), alkalische Phosphatase (ap), Serum-Glutamat-Oxalacetat-Transaminase (sgot), Triglyceride (trig), Plättchen (plt) und Thromboplastinzeit (pro). Zwei mehrkategoriale Variablen wurden wie in Royston u. Sauerbrei [2008] metrisch in das volle Modell aufgenommen. Es handelt sich um das Histologische Stadium (stage) mit den

Kategorien 1, 2, 3 und 4 sowie um die Variable edema mit den drei Kategorien 0, 0.5 und 1. Das volle Modell ist auch hier ein Cox-Modell, in dem alle Kovariablen enthalten sind. Im nun folgenden Abschnitt wird die Methodik der Untersuchungen, die im Rahmen dieser Arbeit anhand der gerade vorgestellten Datensätze durchgeführt wurden, beschrieben.

5 Methodik

Das Ziel dieser Arbeit ist, wie schon zuvor erwähnt, die Wahl der Subsampling-Ratio bei Resampling-basierter Variablenselektion hinsichtlich der Stabilität der Ergebnisse zu untersuchen. Dies wird anhand dreier realer Datensätze mithilfe der in diesem Kapitel beschriebenen Methodik durchgeführt. Diese Methodik basiert auf den Strategien, die in Kapitel 2 und Kapitel 3 beschrieben sind. Die Ergebnisse der Untersuchungen werden anschließend im Kapitel 6 präsentiert.

5.1 Grundidee

Variablenselektion auf der Basis von Resampling durchzuführen ist laut de Bin u. a. [2016] ein Versuch, die Stabilität der Ergebnisse zu verbessern. Die Entscheidung, welche der möglichen Variablen einen Einfluss auf die Zielgröße haben und in das Modell aufgenommen werden, fällt nicht auf Basis einer Variablenselektion, die auf dem gesamten Datensatz durchgeführt wird. Stattdessen werden aus dem Datensatz mehrere Pseudostichproben mittels Subsampling gezogen, auf denen jeweils eine Variablenselektion gerechnet wird [de Bin u. a., 2016]. Anschließend kann für jede der in Frage kommenden Kovariablen berechnet werden, wie häufig sie in das Modell selektiert wurde [Sauerbrei u. Schumacher, 1992]. Diese Häufigkeiten können dann als Entscheidungskriterium genutzt werden, welche Variablen einen Einfluss auf die Zielgröße haben und in das endgültige Modell aufgenommen werden sollten [Sauerbrei u. Schumacher, 1992]. Mithilfe der Inklusionshäufigkeiten kann dann ein Ranking der Variablen bezüglich ihrer Relevanz für das Modell gebildet werden.
Um die Stabilität dieses Verfahrens in dieser Arbeit untersuchen zu können, werden die Datensätze zufällig halbiert. Die Resampling-basierte Variablenselektion wird dann auf beiden so entstandenen Subdatensätzen durchgeführt. Anschließend können die Ergebnisse - insbesondere die Inklusionshäufigkeiten - der Subdatensätze verglichen werden.
Die Idee ist, dass sich die Subdatensätze gegenseitig validieren. Die beste Ratio ist diejenige, bei der sich die Ergebnisse der Resampling-basierten Variablenselektionen auf den Subdatensätzen entsprechend einem Kriterium am wenigsten unterscheiden. Diese Ratio kann als die Subsampling-Ratio interpretiert werden, die das Ranking der Variablen nach den Inklusionshäufigkeiten am ehesten in unabhängigen Daten validiert. Genaueres hierzu findet sich unter Kapitel 5.2.1.
Es werden vier Ratios betrachtet, das Ziehen im Verhältnis von 9-1, 4-1, 2-1 und 1-1. Das Verhältnis 2-1 wird durch 0.632 ersetzt. Dies entspricht im Mittel dem Anteil der aus dem Datensatz per Bootstrap gezogenen Beobachtungen, die dann teilweise mehrfach in der Bootstrap-Stichprobe vorkommen, an der Gesamtzahl der Beobachtungen des Datensatzes. Deshalb wurde 0.632 bisher oft als Subsampling-Ratio verwendet [de Bin u. a., 2016].

Im Folgenden wird teilweise von kleinen und großen Ratios gesprochen. Das soll die Größe der durch eine Ratio entstehenden Pseudostichproben widerspiegeln. Bei der Ratio 9-1 werden Pseudostichproben mit einer Größe von 90% des Umfangs der Subdatensätze gezogen, bei der Ratio 4-1 entsprechend 80%, bei 2-1 wie oben schon erwähnt 63.2% und bei der Ratio 1-1 50%. Die Pseudostichproben der Ratio 9-1 sind also größer als die der Ratio 4-1 usw. Deshalb wird in dieser Arbeit manchmal von größeren Ratios wie 9-1 oder auch 4-1 und im Vergleich dazu eher kleineren Ratios wie 2-1 oder 1-1 gesprochen.
Zusätzlich werden die Inklusionshäufigkeiten mit den Teststatistiken der Wald-Tests der vollen Modelle auf den Subdatensätzen anhand von Korrelationen verglichen. Eine detaillierte Beschreibung des Vorgehens findet sich unter Kapitel 5.2.2.

5.2 Aufbau der Analysen

Für jede Ratio wird mit jedem Datensatz folgendes Vorgehen angewandt. Das Halbieren des entsprechenden Datensatzes wird 50 Mal zufällig durchgeführt. Für jede der Halbierungen wird dann auf beiden Subdatensätzen mittels Subsampling eine Resampling-basierte Variablenselektion durchgeführt. Dabei wird aus beiden Subdatensätzen 1000 Mal eine Stichprobe mit der entsprechenden Ratio gezogen und auf dieser eine Backward Elimination gerechnet. Lediglich für den Datensatz Ozone mit Ratio 9-1 beträgt die Zahl der Ziehungen aus beiden Subdatensätzen statt 1000 Mal jeweils 10000 Mal. Aufgrund der langen Rechenzeit ist bei allen anderen Untersuchungen die Anzahl der Ziehungen auf 1000 reduziert.
Als Ergebnisse ergeben sich für jede der 50 Halbierungen für beide Subdatensätze die Inklusionshäufigkeiten der Variablen. Seien $k = 1, ..., niter$ die Halbierungen, die im Folgenden als Iterationen bezeichnet werden. Für einen Datensatz sind die Subdatensätze der Iteration k bei allen vier Ratios gleich. Auf diese Weise können für jede Iteration die Ergebnisse der Resampling-basierten Variablenselektionen auf den beiden Subdatensätzen für die verschiedenen Ratios verglichen werden. Technisch wird das über einen seed realisiert, wie unter Kapitel 5.3 noch genauer erläutert wird.

5.2.1 Vergleich über Diskordanzkriterien

Um die Ergebnisse der Variablenselektion bezüglich der Stabilität bewerten zu können, wird ein Diskordanzkriterium berechnet. Dieses Kriterium kann sowohl auf Basis der Iterationen als auch auf Basis der Kovariablen betrachtet werden. Je größer das Kriterium einer Iteration ist, desto stärker sind in dieser Iteration die Unterschiede der Inklusionshäufigkeiten der Variablen zwischen den zwei Subdatensätzen. Je größer das Kriterium einer Variable ist, desto stärker sind die Unterschiede der Inklusionshäufigkeiten aller Iterationen dieser Variable zwischen den zwei Subdatensätzen.
Seien $k = 1, ..., niter$ die Iterationen, um den Datensatz *niter*-mal zufällig zu halbieren, und $j = 1, ..., p$ die Kovariablen des vollen Modells. Dann bezeichnen in Formel (3) und Formel (4) $\pi_{kj}^{(1)}$ und $\pi_{kj}^{(2)}$ die Inklusionshäufigkeit der j-ten Kovariable in der k-ten Iteration in dem ersten bzw. zweiten Subdatensatz.
Wird das Diskordanzkriterium für die Iterationen berechnet, lautet die Formel für das

Kriterium der k-ten Iteration

$$\frac{1}{p}\sum_{j=1}^{p}|\pi_{kj}^{(1)}-\pi_{kj}^{(2)}|. \tag{3}$$

Es wird für jede der p Variablen der Absolutbetrag der Differenz der zwei Inklusionshäufigkeiten aus den zwei Subdatensätzen der k-ten Iteration berechnet. Diese absoluten Abweichungen werden aufsummiert und ergeben das Kriterium, indem die so entstehende Summe noch durch p geteilt wird. Je stärker sich in einer Iteration die Inklusionshäufigkeiten einer Variable in den zwei Subdatensätzen unterscheiden, desto weniger stabil ist das Ergebnis der Resampling-basierten Variablenselektion für diese Variable in dieser Iteration. Das Ergebnis in dem einen Subdatensatz wird in diesem Fall in dem anderen Subdatensatz nicht validiert und umgekehrt. Demzufolge ist das Ergebnis einer Iteration desto stabiler, je kleiner ihr Diskordanzkriterium ist. Ein großes Diskordanzkriterium bedeutet große Abweichungen zwischen den Inklusionshäufigkeiten der Variablen der zwei Subdatensätze.
Wird das Diskordanzkriterium für die Variablen berechnet, lautet die Formel für das Kriterium der j-ten Variable

$$\frac{1}{niter}\sum_{k=1}^{niter}|\pi_{kj}^{(1)}-\pi_{kj}^{(2)}|. \tag{4}$$

Es wird für jede der *niter* Iterationen der Absolutbetrag der Differenz der zwei Inklusionshäufigkeiten aus den zwei Subdatensätzen der j-ten Kovariable berechnet. Diese absoluten Abweichungen werden aufsummiert, durch *niter* geteilt und ergeben so das Kriterium. Je stärker sich für eine Variable ihre Inklusionshäufigkeiten in den zwei Subdatensätzen einer Iteration unterscheiden, desto weniger stabil ist das Ergebnis der Resampling-basierten Variablenselektion für diese Iteration dieser Variable. Das Ergebnis in dem einen Subdatensatz wird in dem anderen Subdatensatz nicht validiert und umgekehrt. Folglich ist das Ergebnis für eine Variable desto stabiler, je kleiner ihr Diskordanzkriterium ist. Ein großes Diskordanzkriterium bedeutet große Abweichungen zwischen ihren Inklusionshäufigkeiten der zwei Subdatensätze der *niter* Iterationen.
Der Wertebereich der Diskordanzkriterien ist bei beiden Berechnungsarten nach unten durch die 0 begrenzt. Im Falle der Berechnung für die Iterationen bedeutet das, dass in der entsprechenden Iteration alle Variablen in beiden Subdatensätzen dieselben Inklusionshäufigkeiten aufweisen, sodass die Differenzen dieser gleich 0 sind. Bei der Betrachtung aus dem Blickwinkel der Variablen heißt ein Diskordanzkriterium von 0, dass für die aktuell interessierende Variable in allen Iterationen in beiden Subdatensätzen dieselben Inklusionshäufigkeiten berechnet wurden. Nach oben ist der Wertebereich durch die 1 begrenzt. Dieser Fall tritt ein, wenn sich die Inklusionshäufigkeiten der beiden Subdatensätze für alle Variablen einer Iteration bzw. für alle Iterationen einer Variable maximal unterscheiden. Das heißt, eine der zwei zusammengehörigen Inklusionshäufigkeiten hat jeweils den Wert 1, die andere den Wert 0.

Für jede Subsampling-Ratio werden $niter = 50$ Iterationen durchgeführt. Über den Vergleich der zu den jeweiligen Ratios gehörigen Diskordanzkriterien kann die Stabilität der Ergebnisse der Resampling-basierten Variablenselektion basierend auf verschiedenen Subsampling-Ratios untersucht werden.

Mithilfe der Betrachtung des Diskordanzkriteriums auf Basis der Iterationen kann global für alle Kovariablen des Datensatzes beurteilt werden, wie stabil die Ergebnisse der Resampling-basierten Variablenselektionen in den einzelnen Iterationen sind. In den meisten Datensätzen sind verschiedene Variablentypen enthalten. Es gibt metrische, binäre oder mehrkategoriale Größen. Außerdem nehmen die verschiedenen Kovariablen unterschiedlich stark Einfluss auf die Zielgröße. Deshalb ist es sinnvoll, das Diskordanzkriterium für die einzelnen Iterationen zu berechnen, um für diese unterschiedlichen Typen und verschieden starken Einflüsse die Resampling-basierte Variablenselektion mit einer bestimmten Subsampling-Ratio global bezüglich der Stabilität beurteilen zu können.
Das Berechnen des Diskordanzkriteriums für die einzelnen Variablen liefert noch einen anderen Blickwinkel auf die Stabilität der Ergebnisse der Resampling-basierten Variablenselektion. Hier wird die Stabilität nicht global betrachtet, sondern für eine bestimmte Variable mit ihrem spezifischen Variablentyp und Einfluss. Dadurch kann untersucht werden, ob sich die Stabilität der Ergebnisse basierend auf den verschiedenen Ratios für bestimmte Variablen anders verhält als für die übrigen Variablen.
Für einen Datensatz und eine Ratio ergeben sich je nach Berechnungsart p bzw. *niter* Diskordanzkriterien. Um, wie in den vorigen zwei Absätzen beschrieben, die Ergebnisse der verschiedenen Ratios desselben Datensatzes vergleichen zu können, wird der Mittelwert dieser p respektive *niter* Kriterien gebildet. Auf diese Weise ergeben sich pro Datensatz und Ratio ein Diskordanzkriterium aus Sicht der Iterationen sowie ein Kriterium aus Sicht der Variablen. Der Wertebereich beider Kriterien ist aufgrund des Mittelns das Intervall $[0, 1]$. Als Alternative zur Mittelwertbildung könnte auch der Median betrachtet werden. In den Ergebnissen hat sich gezeigt, dass Mittelwert und Median der Diskordanzkriterien beinahe identisch sind, weshalb hier im Folgenden der Mittelwert betrachtet wird.

5.2.2 Vergleich über Korrelationen

Eine weitere Möglichkeit der Beurteilung der Ratios bezüglich der Stabilität ihrer Ergebnisse zusätzlich zu den Diskordanzkriterien bietet die Betrachtung der Korrelation der Rankings der Variablen. Als weiterer Aspekt kommt hinzu, dass auf Basis von Waldstatistiken ebenfalls Rankings gebildet werden können, sodass die Ergebnisse einer Resampling-basierten Variablenselektion basierend auf verschiedenen Subsampling-Ratios zusätzlich mit den Ergebnissen einer einzelnen Analyse basierend auf Waldstatistiken verglichen werden kann. Das Vorgehen hierzu wird in diesem Kapitel genauer beschrieben.
Zusätzlich zur Berechnung der Inklusionshäufigkeiten über Resampling-basierte Variablenselektion mit verschiedenen Subsampling-Ratios wird in jeder Iteration auf beiden Subdatensätzen ein Modell mit allen in Frage kommenden Kovariablen geschätzt und die Waldstatistiken der Komponenten dieser vollen Modelle dokumentiert. Für jede der r Komponenten ergibt sich eine sogenannte Waldstatistik, es ist die Teststatistik eines Wald-Tests mit der Nullhypothese $H_0 : \beta_i = 0$ mit $i = 1, ..., r$ [Fahrmeir u. a., 1996]. Eine Kovariable kann auch durch mehrere Komponenten in dem vollen Modell vertreten sein, wenn sie mehrkategorial ist und mehr als zwei Kategorien besitzt. In diesem Falle wird sie durch Dummyvariablen - also mehrere Komponenten - dargestellt [Fahrmeir u. a., 2009]. Für jede Dummyvariable ergibt sich dann eine Waldstatistik.
Laut Fahrmeir u. a. [1996] stellt die Nullhypothese $H_0 : \beta_j = 0$ einen Spezialfall des Wald-Tests dar, dessen Nullhypothese allgemein eine lineare Hypothese ist. Die Teststatistik des

Wald-Tests ist allgemein unter H_0 asymptotisch χ^2-verteilt. „Für den Spezialfall [...] ist die Wald-Statistik gleich dem quadrierten „t-Wert““ [Fahrmeir u. a., 1996, S.88]. Oft wird in den Outputs von Modellschätzungen der statistischen Software R [R Core Team, 2014] die nicht-quadrierte Teststatistik ausgegeben [Schlittgen, 2013]. Da die Analysen dieser Arbeit mithilfe der Software R durchgeführt wurden, werden hier die nicht-quadrierten Teststatistiken betrachtet.

Mit diesen hier als Waldstatistiken bezeichneten Teststatistiken ergibt sich eine Möglichkeit, die Stabilität der Ergebnisse der auf verschiedenen Ratios basierenden Resampling-basierten Variablenselektionen mit der Stabilität einer Analyse, die nicht Resampling-basiert arbeitet, zu vergleichen. Dies geschieht schließlich über Korrelationen und Rankings.
Bezogen auf die Inklusionshäufigkeiten und die Waldstatistiken werden nun Rankings der Variablen betrachtet. In jeder Iteration können die Variablen einmal nach den Inklusionshäufigkeiten geordnet werden sowie nach den Waldstatistiken. Das heißt, in einer Iteration kommt die Variable mit der höchsten Inklusionshäufigkeit ganz oben in das Ranking gefolgt von der Variable mit zweithöchster Inklusionshäufigkeit usw. Das gleiche gilt für die Waldstatistiken, mit der kleinen Ergänzung, dass die Variablen nicht direkt nach den Waldstatistiken, sondern nach deren Absolutbeträgen geordnet werden. Dies liegt daran, dass die nicht-quadrierten Teststatistiken betrachtet werden.
Die Idee, mit den Variablen Rankings zu bilden, wird in vielen komplexeren Variablenselektionsverfahren angewandt, wie beispielsweise in Quevedo u. a. [2007]. Hier liegt der Fokus allerdings darauf, anhand der Rankings eine Variablenselektion durchzuführen. In dieser Arbeit jedoch soll anhand der Rankings die Stabilität bereits stattgefundener Variablenselektionen verglichen werden.
Da in jeder Iteration Inklusionshäufigkeiten und Waldstatistiken jeweils auf den zwei Subdatensätzen berechnet werden, ergeben sich also zwei Rankings für jeden Subdatensatz einer Iteration, jeweils eines bezüglich der Inklusionshäufigkeiten und eines bezüglich der Waldstatistiken.
Wie Sauerbrei u. Schumacher [1992] erklären, können die Inklusionshäufigkeiten als Kriterien für die Relevanz der Variablen interpretiert werden. Die Rankings basierend auf den Inklusionshäufigkeiten können also so verstanden werden, dass die Variablen bezüglich ihrer Relevanz geordnet werden. Die Rankings basierend auf den Waldstatistiken können ebenfalls in diesem Sinn verstanden werden. Eine Variable mit betragsmäßig großer Waldstatistik hat mit hoher Wahrscheinlichkeit einen Koeffizienten, der sich von 0 unterscheidet [Fahrmeir u. a., 2011], und ist damit von Relevanz für das Modell.

Mithilfe des Korrelationskoeffizienten nach Spearman kann betrachtet werden, ob die zwei Subdatensätze der Iterationen ähnliche Rankings bezüglich der Inklusionshäufigkeiten und der Waldstatistiken liefern. Je ähnlicher die Rankings sind, desto höher ist der entsprechende Korrelationskoeffizient. Es ist hier von Interesse, ob die Variablen in den Rankings der zwei Subdatensätze an ähnlichen Stellen stehen. Also handelt es sich um den gleichsinnigen monotonen Zusammenhang zwischen den Rankings der zwei Subdatensätze - jeweils basierend auf den Inklusionshäufigkeiten und auf den Waldstatistiken. Deshalb wird der Korrelationskoeffizient nach Spearman berechnet, da dieser eine Maßzahl für den monotonen Zusammenhang angibt [Fahrmeir u. a., 2011]. Bei dessen Berechnung werden

statt der ursprünglichen Werte deren Ränge verwendet [Fahrmeir u. a., 2011].
Zuerst werden die Formeln von Spearman's Korrelationskoeffizient zu den Inklusionshäufigkeiten eingeführt. Auch hier kann wie bei den Diskordanzkriterien der Blickwinkel entweder aus Sicht der Iterationen oder der Variablen gewählt werden. Die Rankings werden, wie oben beschrieben, für jede Iteration gebildet. Es macht keinen Sinn, für eine Variable ein Ranking der Inklusionshäufigkeiten oder Waldstatistiken der einzelnen Iterationen zu bilden. Seien $k = 1, ..., niter$ die Iterationen, um den Datensatz $niter$-mal zufällig aufzuteilen, und $j = 1, ..., p$ die Kovariablen des vollen Modells. Dann bezeichnen $rg_{kj}^{(1)}$ und $rg_{kj}^{(2)}$ den Rang der Inklusionshäufigkeit der j-ten Kovariable im Ranking basierend auf dem Subdatensatz 1 der k-ten Iteration bzw. den Rang der Inklusionshäufigkeit der j-ten Kovariable im Ranking basierend auf dem Subdatensatz 2 der k-ten Iteration. Die nachfolgenden vier Formeln des Korrelationskoeffizienten nach Spearman stammen aus Fahrmeir u. a. [2011], lediglich die Notation ist dieser Arbeit angepasst.
Aus Sicht der Iterationen wird für jede Iteration k ein Korrelationskoeffizient über alle Variablen über die Formel

$$\frac{\sum_{j=1}^{p}(rg_{kj}^{(1)} - \overline{rg_k^{(1)}})(rg_{kj}^{(2)} - \overline{rg_k^{(2)}})}{\sqrt{\sum_{j=1}^{p}(rg_{kj}^{(1)} - \overline{rg_k^{(1)}})^2 \sum_{j=1}^{p}(rg_{kj}^{(2)} - \overline{rg_k^{(2)}})^2}} \tag{5}$$

berechnet. $\overline{rg_k^{(1)}}$ und $\overline{rg_k^{(2)}}$ stehen in Formel (5) für den Mittelwert der Ränge aller Variablen basierend auf dem Subdatensatz 1 der k-ten Iteration bzw. den Mittelwert der Ränge aller Variablen basierend auf dem Subdatensatz 2 der k-ten Iteration.
Aus dem Blickwinkel der Variablen lautet basierend auf denselben Rängen für eine Variable j die Formel

$$\frac{\sum_{k=1}^{niter}(rg_{kj}^{(1)} - \overline{rg_j^{(1)}})(rg_{kj}^{(2)} - \overline{rg_j^{(2)}})}{\sqrt{\sum_{k=1}^{niter}(rg_{kj}^{(1)} - \overline{rg_j^{(1)}})^2 \sum_{k=1}^{niter}(rg_{kj}^{(2)} - \overline{rg_j^{(2)}})^2}}. \tag{6}$$

Hierbei bezeichnen $\overline{rg_j^{(1)}}$ und $\overline{rg_j^{(2)}}$ in Formel (6) den Mittelwert der Ränge der j-ten Variable basierend auf dem Subdatensatz 1 aller Iterationen bzw. den Mittelwert der Ränge der j-ten Variable basierend auf dem Subdatensatz 2 aller Iterationen.
Die Formeln für die Waldstatistiken werden analog berechnet. Allerdings existieren im Falle mehrkategorialer Variablen für eine Variable mehrere Dummyvariablen und damit auch mehrere Waldstatistiken. Bei der Berechnung des Korrelationskoeffizienten aus Sicht der Iterationen stellt dies kein Problem dar. Es bezeichnen $i = 1, ..., r$ die Komponenten des vollen Modells und $k = 1, ..., niter$ wie gehabt die Iterationen. Des Weiteren bezeichnen $rg_{ki}^{(1)}$ und $rg_{ki}^{(2)}$ die Ränge der i-ten Komponente in der k-ten Iteration jeweils in dem ersten bzw. zweiten Subdatensatz. Aus Sicht der Iterationen lautet die Formel für den Korrelationskoeffizienten nach Spearman für jede Iteration k

$$\frac{\sum_{i=1}^{r}(rg_{ki}^{(1)} - \overline{rg_k^{(1)}})(rg_{ki}^{(2)} - \overline{rg_k^{(2)}})}{\sqrt{\sum_{i=1}^{r}(rg_{ki}^{(1)} - \overline{rg_k^{(1)}})^2 \sum_{i=1}^{r}(rg_{ki}^{(2)} - \overline{rg_k^{(2)}})^2}}, \tag{7}$$

wobei $\overline{rg_k^{(1)}}$ und $\overline{rg_k^{(2)}}$ in Formel (7) für den Mittelwert der Ränge aller Komponenten basierend auf dem Subdatensatz 1 der k-ten Iteration bzw. den Mittelwert der Ränge aller

Komponenten basierend auf dem Subdatensatz 2 der k-ten Iteration stehen.
Um den Korrelationskoeffizienten der Waldstatistiken aus Sicht der Variablen zu berechnen, wird aufgrund mehrkategorialer Variablen eine Möglichkeit benötigt, für solche Variablen auf den Subdatensätzen jeweils nur einen Rang zu bilden. Eine Möglichkeit wäre, für diese Variablen einen Test wie den Likelihood-Ratio-Test über alle jeweils zugehörigen Dummyvariablen durchzuführen [Rospleszcz u. a., 2016]. In dieser Arbeit werden stattdessen die Ränge der Waldstatistiken der zu einer Variable gehörenden Dummyvariablen gemittelt, sodass für jede Kovariable ein einziger Rang berechnet werden kann. Es seien $j = 1, ..., p$ die Kovariablen des vollen Modells und $k = 1, ..., niter$ die Iterationen. Nun bezeichnen also $rg_{kj}^{(1)}$ und $rg_{kj}^{(2)}$ die Ränge der j-ten Variable in der k-ten Iteration jeweils in dem ersten bzw. zweiten Subdatensatz. Im Falle einer mehrkategorialen Kovariable bezeichnen $rg_{kj}^{(1)}$ und $rg_{kj}^{(2)}$ die Mittelwerte der Ränge der Dummyvariablen der Kovariable in den jeweiligen Subdatensätzen. Die Formel aus dem Blickwinkel der Variablen lautet für eine Variable j dementsprechend

$$\frac{\sum_{k=1}^{niter}(rg_{kj}^{(1)} - \overline{rg_j^{(1)}})(rg_{kj}^{(2)} - \overline{rg_j^{(2)}})}{\sqrt{\sum_{k=1}^{niter}(rg_{kj}^{(1)} - \overline{rg_j^{(1)}})^2 \sum_{k=1}^{niter}(rg_{kj}^{(2)} - \overline{rg_j^{(2)}})^2}}. \tag{8}$$

In Formel (8) stellen $\overline{rg_j^{(1)}}$ und $\overline{rg_j^{(2)}}$ den Mittelwert der Ränge der j-ten Variable basierend auf dem Subdatensatz 1 aller Iterationen bzw. den Mittelwert der Ränge der j-ten Variable basierend auf dem Subdatensatz 2 aller Iterationen dar. Der Wertebereich der Korrelationskoeffizienten umfasst sowohl für die Inklusionshäufigkeiten als auch für die Waldstatistiken für beide Berechnungsarten das Intervall $[-1, 1]$.

Wie werden die Korrelationskoeffizienten der Ränge der Inklusionshäufigkeiten aus Sicht der Iterationen interpretiert? Für jede Ratio werden 50 Iterationen gerechnet, deren Ergebnis jeweils zum einen die auf dem ersten Subdatensatz berechneten Inklusionshäufigkeiten der Variablen und zum anderen die auf dem zweiten Subdatensatz berechneten Inklusionshäufigkeiten der Variablen beinhaltet. Wird für jede Iteration ein Korrelationskoeffizient berechnet, so gibt dieser eine Maßzahl dafür an, wie stabil die Ergebnisse der Resampling-basierten Variablenselektion in der jeweiligen Iteration sind. Sind im Extremfall die Rankings basierend auf den Inklusionshäufigkeiten der Variablen in den zwei Subdatensätzen identisch, so hat der Korrelationskoeffizient den Wert 1. Es handelt sich dann um einen gleichsinnigen monotonen Zusammenhang zwischen den Inklusionshäufigkeiten des einen Subdatensatzes und denen des anderen Subdatensatzes. Haben die Rankings hingegen im anderen Extremfall genau die jeweils umgekehrte Reihenfolge, wird ein negativer Koeffizient mit Wert -1 erwartet. Der monotone Zusammenhang ist dann gegensinnig. Gibt es keinen monotonen Zusammenhang, ist der Korrelationskoeffizient 0. Diese Korrelationskoeffizienten der Iterationen können für alle Ratios berechnet werden. Bei positiven Koeffizienten nahe 1 können die Ergebnisse der entsprechenden Iteration mit zugehöriger Ratio als stabil eingeschätzt werden, da dann die Resampling-basierte Variablenselektion auf den beiden Subdatensätzen ähnliche Rankings liefert. Bei Koeffizienten nahe 0 oder gar negativen Koeffizienten hingegen unterscheiden sich die Rankings stark, sodass in diesem Fall nicht von Stabilität gesprochen werden kann.
Wie ist der Korrelationskoeffizient der Ränge der Inklusionshäufigkeiten aus dem Blickwinkel der Variablen zu verstehen? Wie schon oben erwähnt, werden zur Berechnung des

Koeffizienten aus Sicht der Variablen dieselben Ränge verwendet wie bei demjenigen für die Iterationen. Es ändert sich lediglich die Betrachtung von einer Iteration über alle Variablen hin zur Betrachtung einer Variable über alle Iterationen. Hat eine Variable in allen Iterationen dieselben Ränge in den Rankings basierend auf den Inklusionshäufigkeiten der zwei Subdatensätze, sind die Rankings dieser Variable sehr stabil. In diesem Falle hat der Korrelationskoeffizient dieser Variable den Wert 1. Unterscheiden sich die Ränge dieser Variable in den einzelnen Iterationen jedoch stark, wird der Koeffizient um 0 herum schwanken oder bei gegensinnigem Zusammenhang der Ränge sogar negativ. Die Korrelationskoeffizienten aus Sicht der Variablen betrachten also die Stabilität der Rankings für eine spezielle Variable. Auch hier kann also wie bei den Diskordanzkriterien auch der spezifische Variablentyp und Einfluss einer Variable mitberücksichtigt werden.
Die Interpretation der Korrelationskoeffizienten auf Basis der Waldstatistiken ist analog zu der Interpretation der Inklusionshäufigkeiten. Auch hier wird die Stabilität der Ergebnisse betrachtet - diesmal jedoch der Ergebnisse einer einzelnen Analyse. Auch hier liefern identische Rankings auf den beiden Subdatensätzen einen Koeffizienten vom Wert 1. Eine Iteration liefert also dann sehr stabile Ergebnisse, wenn die Rankings aller Komponenten in beiden Subdatensätzen dieser Iteration identisch sind. Aus dem Blickwinkel der Variablen sind die Teststatistiken einer Variable dann sehr stabil, wenn die Ränge dieser Variable in allen Iterationen in den beiden Subdatensätzen übereinstimmen.

5.2.3 Abhängigkeit von der Stichprobengröße

Die Erkenntnisse bezüglich der Stabilität der Resampling-basierten Variablenselektion, die auf der Basis von drei realen Datensätzen anhand der bisher beschriebenen Diskordanzkriterien und Korrelationskoeffizienten erlangt werden, sind jeweils auf Datensätze mit derselben Verteilung und - wegen des Halbierens - mit einer Stichprobengröße von entsprechend $n/2$ anwendbar. Um die Rolle der Stichprobengröße bei der Wahl der optimalen Subsampling-Ratio zu untersuchen, werden die 50 Iterationen bei dem Datensatz Ozone für alle vier Ratios noch zwei weitere Male durchgeführt. Beim ersten Mal wird, anstatt bei jeder Iteration den Datensatz zufällig zu halbieren, zufällig eine Stichprobe der Größe $n/3$ gezogen. Aus den verbleibenden Beobachtungen wird eine zweite Stichprobe derselben Größe gezogen. Auf den beiden so entstandenen Subdatensätzen werden dann wieder vier Resampling-basierte Variablenselektionen mit den vier betrachteten Ratios durchgeführt. Beim zweiten Mal wird eine Stichprobengröße von $n/4$ für die Subdatensätze verwendet. Würden sich bei den Untersuchungen mit kleineren Subdatensätzen deutlich andere Ergebnisse hinsichtlich der Stabilität der Ratios ergeben, würde das die Verallgemeinerbarkeit der zuvor erlangten Erkenntnisse in Frage stellen.
In der Praxis ist die Größe des Datensatzes, auf dem eine Resampling-basierte Variablenselektion durchgeführt werden soll, ein wichtiger Faktor dafür, ob kleine Ratios überhaupt in Frage kommen. Ist der zur Verfügung stehende Datensatz schon sehr klein, könnte die Schätzung bei Verwendung der Ratio 1-1 möglicherweise sehr in Mitleidenschaft gezogen werden, während eine größere Ratio eventuell noch gute Schätzungen liefern könnte. In einem solchen Fall ist die Wahl der Subsampling-Ratio schon aufgrund der Qualität der Schätzung eingeschränkt.

5.3 Implementierung in R

Das oben beschriebene Vorgehen ist in der Software R [R Core Team, 2014] implementiert. Dafür werden die Pakete MASS [Venables u. Ripley, 2002], survival [Therneau, 2015; Therneau u. Grambsch, 2000] und testit [Xie, 2016] verwendet. Um die Ergebnisse ansprechend darzustellen werden zusätzlich die Pakete ggplot2 [Wickham, 2009], gridExtra [Auguie, 2015], colorspace [Ihaka u. a., 2015; Zeileis u. a., 2009] und gtable [Wickham, 2012] genutzt. Es existieren zwei Versionen der benötigten Funktionen. Eine Version ist für Daten geschrieben, auf denen ein generalisiertes lineares Modell geschätzt werden soll, und eine Version für Überlebenszeitdaten, auf denen ein Cox-Modell geschätzt werden soll. Der Aufbau dieser zwei Versionen ist identisch, es sind jeweils zwei Funktionen, von denen eine die andere aufruft. Das Grundgerüst dieser zwei Funktionen stammt von Frau Prof. Dr. Anne-Laure Boulesteix.
In der äußeren Funktion werden die Ergebnisse der Iterationen erstellt. Außerdem gibt es eine Schleife, in der *niter*-mal der Datensatz zufällig in die Subdatensätze aufgeteilt wird. In jeder Iteration k wird über die Funktion set.seed ein neuer Startwert k für den Zufallsgenerator gesetzt, sodass die Ergebnisse reproduzierbar sind. Das Ergebnisse der k-ten Iterationen von Analysen verschiedener Ratios desselben Datensatzes mit demselben Umfang der Subdatensätze sind deshalb vergleichbar. Die k-te Iteration eines Datensatzes mit einem bestimmten Stichprobenumfang der Subdatensätze der Ratio 9-1 arbeitet nämlich mit denselben Subdatensätzen wie die k-te Iteration der anderen Ratios desselben Datensatzes mit demselben Stichprobenumfang der Subdatensätze. Die Waldstatistiken der Variablen werden basierend auf den zwei Subdatensätzen ebenfalls in der Schleife mithilfe der Funktion glm und der Funktion coxph aus dem Paket survival [Therneau, 2015; Therneau u. Grambsch, 2000] berechnet.
Die Resampling-basierte Variablenselektion wird in der zweiten Funktion, die im Schleifenrumpf der ersten Funktion aufgerufen wird, durchgeführt. Auch hier gibt es wieder eine Schleife, die das Resampling realisiert. Wegen der Reproduzierbarkeit wird auch hier in jedem neuen Schleifendurchlauf ein neuer seed gesetzt. Dann wird mittels Subsampling und der entsprechenden Ratio eine Pseudostichprobe gezogen. Auf dieser wird mittels der Funktion stepAIC aus dem Paket MASS eine Backward Elimination durchgeführt. StepAIC entscheidet in jedem Schritt über das AIC, ob und welche Variable aus dem Modell entfernt wird oder nicht. Die innere Funktion berechnet die Inklusionshäufigkeiten der Variablen und gibt diese zurück. Das Schätzen der vollen Modelle geschieht mit der Funktion glm und der Funktion coxph aus dem Paket survival [Therneau, 2015; Therneau u. Grambsch, 2000]. Auf diesen basierend startet stepAIC die Backward Elimination.

Aufgrund von Warnmeldungen, die von der Funktion coxph bei der Analyse der Datensätze Glioma und PBC ausgegeben werden, sind die Funktionen der Version für die Überlebenszeitmodelle leicht modifiziert im Vergleich zu den Funktionen der anderen Version, bei der bei der Analyse des Datensatzes Ozone keine Warnungen auftreten. Diese Warnungen melden, dass die geschätzten Koeffizienten unendlich sein könnten. Die Erklärung dafür, warum diese Warnmeldungen hier beim Schätzen der Modelle der Überlebenszeitdaten auftreten, könnte wie folgt sein. Bei Überlebenszeitdaten kann es passieren, dass die Aufteilung des Datensatzes in die Subdatensätze oder das Ziehen aus diesen beim Resampling unglücklich in dem Sinne geschieht, dass das Verhältnis von kategorialen Kovariablen zu der Aufteilung der Beobachtungen in Events und Nicht-Events unausgewogen

ist. Beispielsweise kann es sein, dass in einer gezogenen Pseudostichprobe keine Events mit einer bestimmten Faktorausprägung einer Kovariable enthalten sind. Wegen dieser unausgewogenen Verteilung der Beobachtungen könnte die Schätzung der Koeffizienten Probleme bereiten. Iterationen, in denen solche Warnungen auftreten, werden in dieser Arbeit verworfen.
Des Weiteren treten bei den Datensätzen Glioma und PBC dieselben Warnungen beim Ausführen der stepAIC-Funktion bezüglich der Schätzung der Modelle in Zuge des Variablenselektionsprozesses auf. Dies geschieht bei kleinen Ratios deutlich häufiger als bei großen Ratios. Die Vermutung ist, dass je kleiner das Ratio und damit die Pseudostichproben sind, desto schwieriger die Schätzung wird. Iterationen mit solchen Warnmeldungen werden jedoch nicht verworfen. Zum einen gibt es Bedenken bezüglich der Laufzeit. Der wichtigere Grund ist jedoch der, dass kleine Ratios nicht bevorteilt werden sollen, indem Iterationen verworfen werden, in denen bei der Variablenselektionsprozedur stepAIC Warnungen auftreten. In Zukunft sollte untersucht werden, ob sich die Ergebnisse bezüglich der Stabilität Resampling-basierter Variablenselektion auf Basis verschiedener Ratios ändern, wenn auch die Iterationen mit Warnmeldungen der Funktion stepAIC verworfen würden.
Eine Frage, die sich im Rahmen der Implementierung in R noch stellt, ist, wie stepAIC mit Variablen, die mehr als zwei Kategorien besitzen, umgeht. Solche mehrkategorialen Variablen werden durch zwei oder mehr Dummyvariablen repräsentiert [Fahrmeir u. a., 2009]. Bei der Variablenselektion ist hier erwünscht, dass entweder alle Dummyvariablen der entsprechenden Kovariable aus dem Modell ausgeschlossen werden sollen oder alle Dummyvariablen im Modell verbleiben sollen. Wie sich stepAIC bei mehrkategorialen Variablen mit mehr als zwei Kategorien verhält, kann anhand einer kleinen Simulation herausgefunden werden. Das Ergebnis ist, dass stepAIC wie gewünscht entweder alle Dummyvariablen einer Variable oder keine aus dem Modell ausschließt.

6 Ergebnisse

Nun werden die Ergebnisse der Untersuchungen, die mithilfe der unter Kapitel 4 vorgestellten Datensätze und der unter Kapitel 5 erklärten Methodik durchgeführt wurden, präsentiert. Die Methodik basiert auf der Theorie, die in Kapitel 3 sowie Kapitel 2 beschrieben wird.

6.1 Vergleich über Diskordanzkriterien

6.1.1 Betrachtung der Iterationen

Zuerst wird zur Beurteilung der Stabilität der Ergebnisse der Resampling-basierten Variablenselektion mit verschiedenen Subsampling-Ratios das unter Kapitel 5.2.1 beschriebene Diskordanzkriterium für die Iterationen betrachtet. Bei allen drei Datensätzen liefert in jeder Iteration die Ratio 1-1 die stabilsten Ergebnisse. Ebenso ist bei allen drei Datensätzen die Ratio 2-1 am zweitbesten hinsichtlich der Stabilität, dann folgt die Ratio 4-1 und schließlich die Ratio 9-1. Demzufolge ist der Mittelwert der Diskordanzkriterien aller

Iterationen bei Ratio 1-1 auch bei allen Datensätzen am geringsten. Diese Mittelwerte sind in Tabelle 1 dargestellt. Wie schon unter Kapitel 5.2.1 beschrieben, sind die Inklu-

	Ozone	**Glioma**	**PBC**
Ratio 9-1	0.327	0.352	0.397
Ratio 4-1	0.281	0.311	0.346
Ratio 2-1	0.213	0.249	0.261
Ratio 1-1	0.163	0.203	0.193

Tabelle 1: Tabelle der gemittelten Diskordanzkriterien aus Sicht der Iterationen der Datensätze Ozone, Glioma und PBC für alle Ratios bei einem Stichprobenumfang der Subdatensätze von jeweils $n/2$. Bei allen Datensätzen liefert Ratio 1-1 das kleinste mittlere Diskordanzkriterium.

sionshäufigkeiten der zwei Subdatensätze einer Iteration je ähnlicher, desto kleiner das Diskordanzkriterium dieser Iteration ist. Je ähnlicher die Inklusionshäufigkeiten der zwei Subdatensätze sind, desto stabiler ist das Ergebnis der verwendeten Resampling-basierten Variablenselektion mit entsprechender Subsampling-Ratio in dieser Iteration.

In Abbildung 1 auf Seite 26 ist zu erkennen, dass in allen 50 Iterationen des Daten-

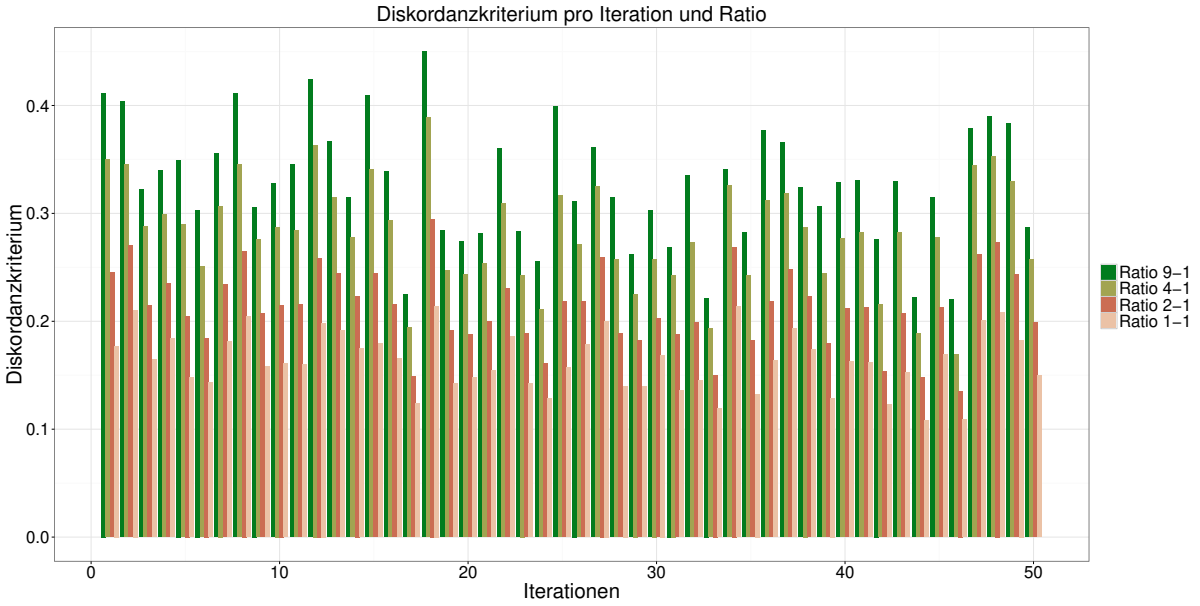

Abbildung 1: Übersicht der Diskordanzkriterien aus Sicht der Iterationen des Datensatzes Ozone für alle Ratios bei einem Stichprobenumfang der Subdatensätze von jeweils $n/2$. In allen Iterationen liefert Ratio 1-1 das kleinste Diskordanzkriterium.

satzes Ozone das Diskordanzkriterium der Ratio 1-1 am kleinsten ist, am zweitkleinsten das Kriterium der Ratio 2-1, gefolgt von Ratio 4-1 und zuletzt Ratio 9-1. Das bedeutet, wie in Kapitel 5 erklärt, dass die Ergebnisse der Resampling-basierten Variablenselektion mit der Subsampling-Ratio 1-1 am ehesten in einem unabhängigen Datensatz validiert

werden, weil die Unterschiede der Inklusionshäufigkeiten der beiden Subdatensätze bei dieser Ratio am geringsten sind. In manchen Iterationen, wie beispielsweise in Iteration 18, ist der Unterschied zwischen den Kriterien der verschiedenen Ratios recht deutlich zu erkennen. Bei anderen Iterationen hingegen liegen die Kriterien näher beieinander wie zum Beispiel in den Iterationen 17 oder 33.

Das gleiche Bild zeigt sich beim Datensatz Glioma. In Abbildung 2 auf Seite 27 ist zu

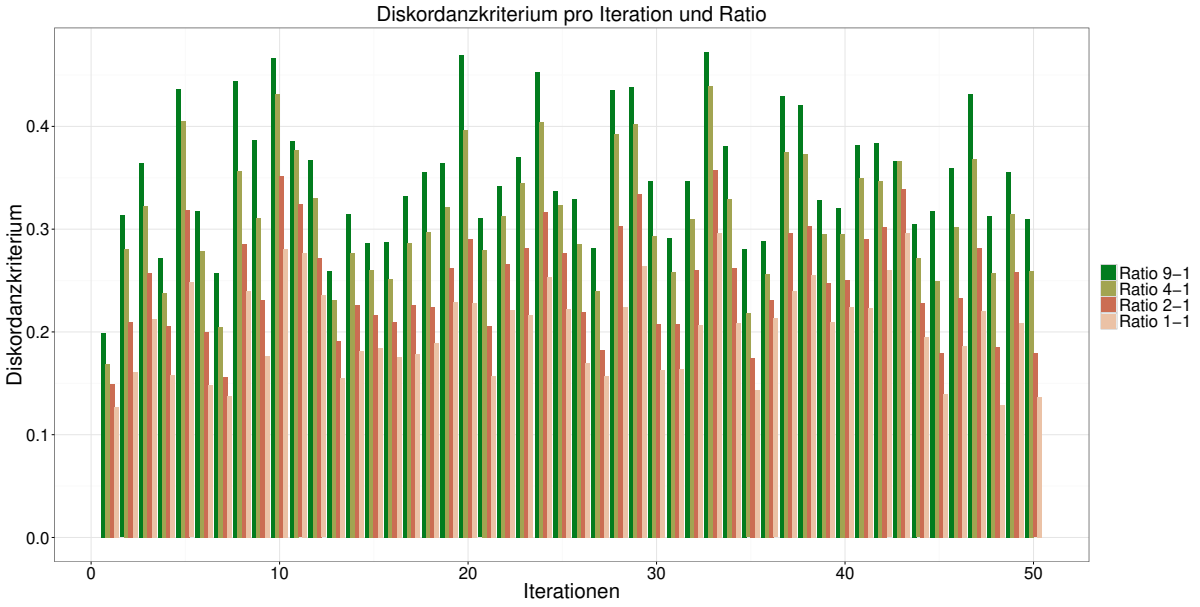

Abbildung 2: Übersicht der Diskordanzkriterien aus Sicht der Iterationen des Datensatzes Glioma für alle Ratios bei einem Stichprobenumfang der Subdatensätze von jeweils $n/2$. In allen Iterationen liefert Ratio 1-1 das kleinste Diskordanzkriterium.

erkennen, dass hier ebenfalls in allen Iterationen die Diskordanzkriterien der Ratio 1-1 am kleinsten sind. Auch hier unterscheiden sich die Kriterien der Ratios verschieden stark in den unterschiedlichen Iterationen.

Der Datensatz PBC bestätigt in Abbildung 3 auf Seite 28 ebenso die Ratio 1-1 als diejenige mit den kleinsten Diskordanzkriterien in allen Iterationen und damit als diejenige, die die stabilsten Ergebnisse liefert.

Das Diskordanzkriterium wird aus den Inklusionshäufigkeiten berechnet. Letztere werden nun pro Iteration für den Datensatz Ozone betrachtet, um die Diskordanzkriterien nachvollziehen zu können. In Abbildung 4 auf Seite 29 ist für jede der 50 Iterationen ein Streudiagramm der Inklusionshäufigkeiten der Ratio 1-1 gezeichnet. Jeder Punkt entspricht einer Variable in der jeweiligen Iteration, der x-Wert ist der Wert der Inklusionshäufigkeit des ersten Subdatensatzes, der y-Wert ist entsprechend derjenige des zweiten Subdatensatzes. Abbildung 5 auf Seite 30 veranschaulicht den gleichen Sachverhalt für die Ratio 9-1 des Datensatzes Ozone. Wie in Abbildung 1 auf Seite 26 zu sehen ist, unterscheiden sich die Diskordanzkriterien der Ratios 1-1 und 9-1 in jeder Iteration am meisten, weshalb hier die Streudiagramme dieser beiden Ratios betrachtet werden. Die entsprechenden Grafiken der Ratios 2-1 und 4-1 befinden sich im Anhang.

Beispielsweise wird in diesen Streudiagrammen die Iteration 8 näher studiert. Bei Ite-

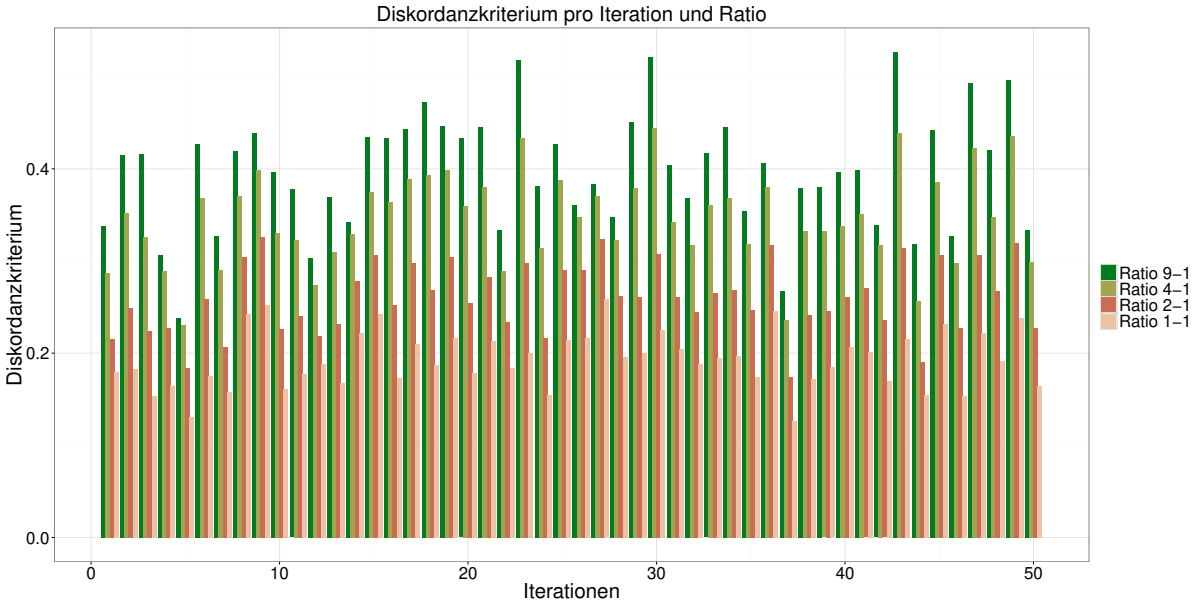

Abbildung 3: Übersicht der Diskordanzkriterien aus Sicht der Iterationen des Datensatzes PBC für alle Ratios bei einem Stichprobenumfang der Subdatensätze von jeweils $n/2$. In allen Iterationen liefert Ratio 1-1 das kleinste Diskordanzkriterium.

ration 8 ist wie bei den anderen Iterationen auch zu erkennen, dass die Punkte in der Grafik zu Ratio 9-1 näher an den Koordinatenachsen liegen als bei Ratio 1-1. Die Punkte streuen in Abbildung 5 stärker. Es gibt also mehrere Variablen in Iteration 8, deren Inklusionshäufigkeiten der beiden Subdatensätze der Ratio 9-1 sich stark unterscheiden. Einige Punkte liegen beispielsweise nahe an der y-Achse, sie haben also eine kleine Inklusionshäufigkeit in dem ersten Subdatensatz, in Subdatensatz zwei jedoch teilweise eine relativ große Inklusionshäufigkeit. Ebenso gibt es einige Punkte nahe der x-Achse. Bei diesen ist es umgekehrt. In Abbildung 4 hingegen liegen die Punkte nicht so nahe an den Koordinatenachsen und generell näher beieinander. Hier gibt es also in Iteration 8 nicht so viele Variablen mit stark unterschiedlichen Inklusionshäufigkeiten in den beiden Subdatensätzen. Deshalb ist das Diskordanzkriterium dieser Iteration auch bei Ratio 1-1 kleiner als bei Ratio 9-1. Es beträgt jeweils 0.20 und 0.41.
Was zuvor für Iteration 8 exemplarisch beschrieben wurde, lässt sich in den anderen Iterationen ebenfalls beobachten. Aus diesem Grund ist das Diskordanzkriterium für den Datensatz Ozone und Ratio 1-1 mit 0.16, was durch das Mitteln der Diskordanzkriterien der 50 Iterationen des Datensatzes Ozone der Ratio 1-1 berechnet wird und in Abbildung 1 erscheint, kleiner als dasjenige für den Datensatz Ozone und Ratio 9-1 mit dem Wert 0.33.

Bei den beiden anderen Datensätzen Glioma und PBC unterscheiden sich ebenfalls in jeder Iteration die Diskordanzkriterien der Ratios 1-1 und 9-1 am meisten. In den entsprechenden Grafiken der Ratios mit den Streudiagrammen der Inklusionshäufigkeiten der Ratios 1-1 bzw. 9-1 zeigt sich das gleiche Szenario, das gerade für den Datensatz Ozone beschrieben wurde. Für beide Datensätze befinden sich diese Grafiken zusammen mit den Grafiken der Ratios 2-1 und 4-1 ebenfalls im Anhang.
Extreme Unterschiede zwischen den Inklusionshäufigkeiten einer Variablen in einer Ite-

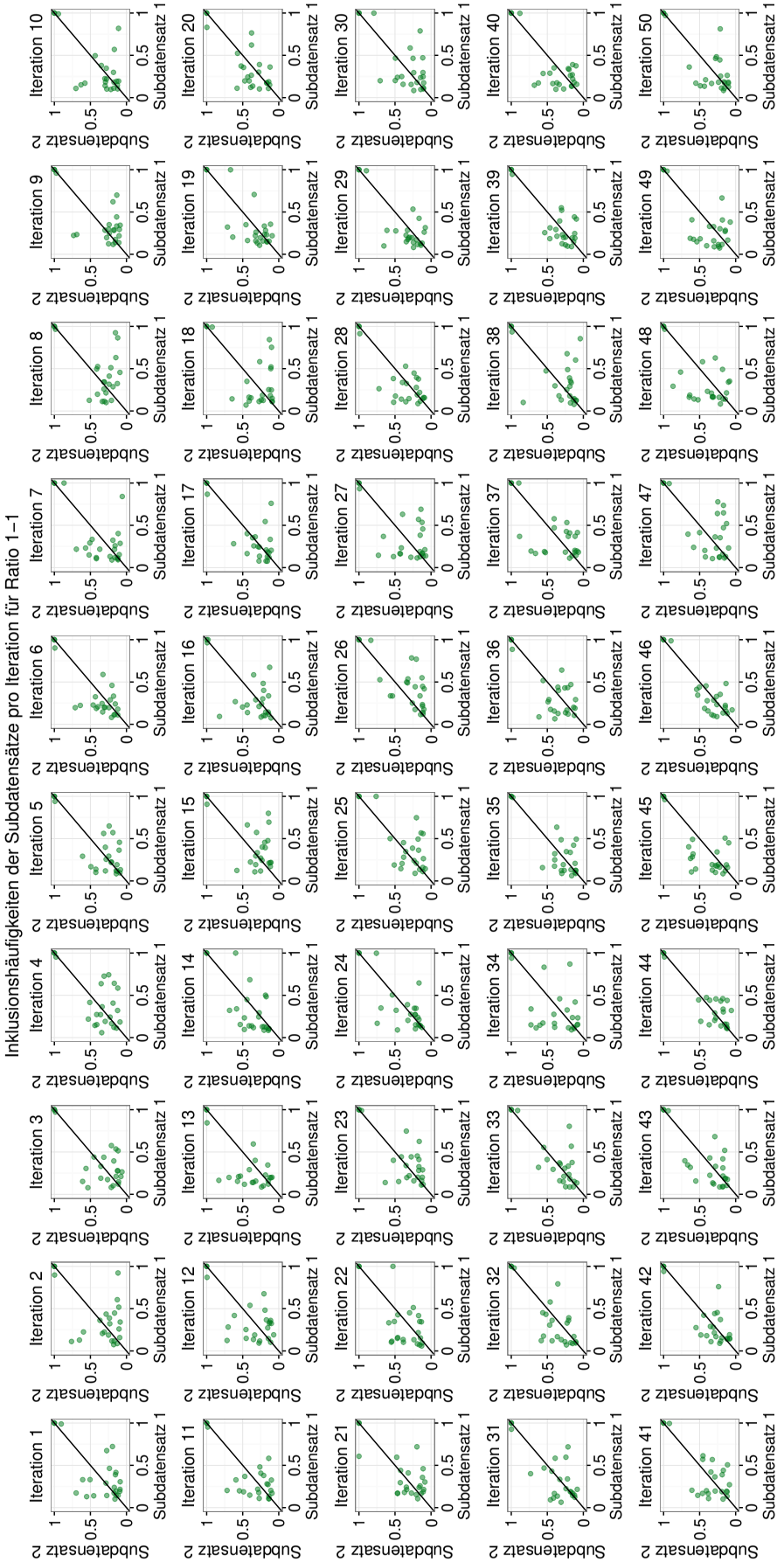

Abbildung 4: Übersicht der Inklusionshäufigkeiten pro Iteration des Datensatzes Ozone für die Ratio 1-1 bei einem Stichprobenumfang der Subdatensätze von jeweils $n/2$. Die Inklusionshäufigkeiten streuen hier weniger stark als in der entsprechenden Übersicht der Ratio 9-1.

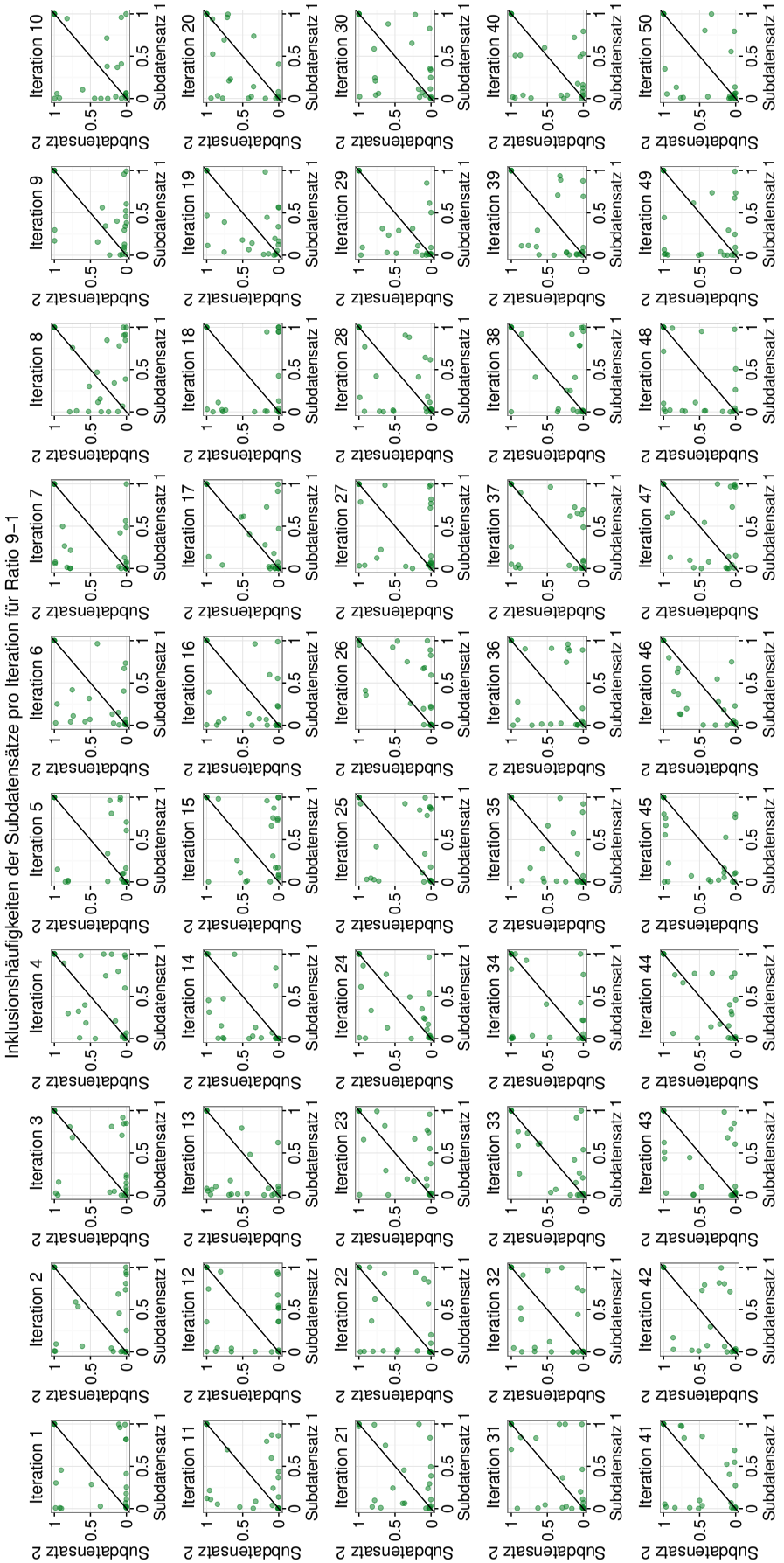

Abbildung 5: Übersicht der Inklusionshäufigkeiten pro Iteration des Datensatzes Ozone für die Ratio 9-1 bei einem Stichprobenumfang der Subdatensätze von jeweils $n/2$. Die Inklusionshäufigkeiten streuen hier stärker als in der entsprechenden Übersicht der Ratio 1-1.

ration lassen sich wahrscheinlich dadurch erklären, dass das Halbieren des Datensatzes in dieser Iteration zwei sehr unterschiedliche Subdatensätze generiert hat. Beim Ziehen der Pseudostichproben mit der Subsampling-Ratio 9-1 sind die so entstehenden Pseudostichproben den ursprünglichen Subdatensätzen, aus denen sie entstanden sind, noch sehr ähnlich. Je kleiner die Ratio gewählt wird, mit der gezogen wird, desto unähnlicher sind die Pseudostichproben den Subdatensätzen. Insbesondere wird der Umfang der Pseudostichproben mit kleineren Ratios kleiner. Deshalb werden durch die Halbierung verursachte Besonderheiten in den Subdatensätzen bei der Verwendung kleinerer Ratios wahrscheinlich weniger in die Pseudostichproben weitergegeben als bei größeren Ratios. Übertragen würde das bedeuten, dass sich Besonderheiten, die durch eine eventuell unglückliche Ziehung der Stichprobe verursacht werden, bei Anwendung eines Resampling-basierten Variablenselektionsverfahrens mit einer kleineren Subsampling-Ratio nicht so stark in der Modellauswahl niederschlagen würden. Das heißt, bei der Verwendung der Ratio 1-1 ziehen Änderungen im Datensatz anscheinend nicht so stark Änderungen des ausgewählten Modells nach sich wie bei den anderen Ratios.
Es gibt in einem Datensatz nicht nur Besonderheiten, die durch eine eventuell unglückliche Ziehung der Stichproben verursacht werden. Jeder Datensatz enthält per se Kovariablen, die jeweils einen spezifischen Einfluss auf die Zielgröße haben. Dies führt zu der Betrachtung der Diskordanzkriterien und Inklusionshäufigkeiten aus dem Blickwinkel der Variablen hin.

6.1.2 Betrachtung der Variablen

Es stellt sich die Frage, ob die Beurteilung der Ratios aus Sicht der Variablen zu anderen Schlüssen führen würde als aus Sicht der Iterationen. Deshalb wird das Diskordanzkriterium hier, wie in Kapitel 5.2.1 beschrieben, für die Variablen berechnet. In Abbildung 6 auf Seite 32 ist ein Überblick über die Diskordanzkriterien der Variablen des Datensatzes Ozone gegeben. Bei vielen Variablen ist ebenfalls das Kriterium der Ratio 1-1 das kleinste und somit ist bei diesen Variablen Ratio 1-1 diejenige, die nach Definition des Diskordanzkriteriums die stabilsten Ergebnisse liefert. Bei diesen Kovariablen ist ebenso fast immer die Ratio 2-1 die zweitkleinste, gefolgt von Ratio 4-1 und schließlich Ratio 9-1. Drei Variablen tanzen jedoch deutlich aus der Reihe. Bei der Variable FLGEW ist die Reihenfolge der Ratios genau umgekehrt, hier liefert Ratio 9-1 das kleinste Diskordanzkriterium und somit die stabilsten Ergebnisse. Die Variablen SEX und FLGROSS weisen bei allen Ratios außer der Ratio 1-1 ein Diskordanzkriterium von 0 auf. Das Kriterium der Ratio 1-1 beträgt bei ihnen 0.00044 bzw. 0.00004, ist also jeweils sehr nahe bei 0. Ein Diskordanzkriterium einer Variable mit dem Wert 0 bedeutet, dass die Inklusionshäufigkeiten dieser Variable in den beiden Subdatensätzen in allen Iterationen jeweils identisch sind. Die drei Variablen FLGEW, SEX und FLGROSS haben alle gemeinsam, dass die Ratio 1-1 bei Ihnen das größte Diskordanzkriterium liefert, also damit per Definition des Kriteriums die am wenigsten stabilen Ergebnisse.

Werden die Inklusionshäufigkeiten dieser drei Variablen betrachtet, stellt sich heraus, dass sie die Variablen mit den höchsten Inklusionshäufigkeiten des Datensatzes Ozone sind. In Abbildung 7 auf Seite 33 ist deutlich zu erkennen, dass die Variablen SEX und FLGROSS bei jeder Ratio fast ausschließlich den Wert 1 als Inklusionshäufigkeiten besitzen. FLGEW hat bei allen Ratios teilweise auch etwas geringere Inklusionshäufigkeiten

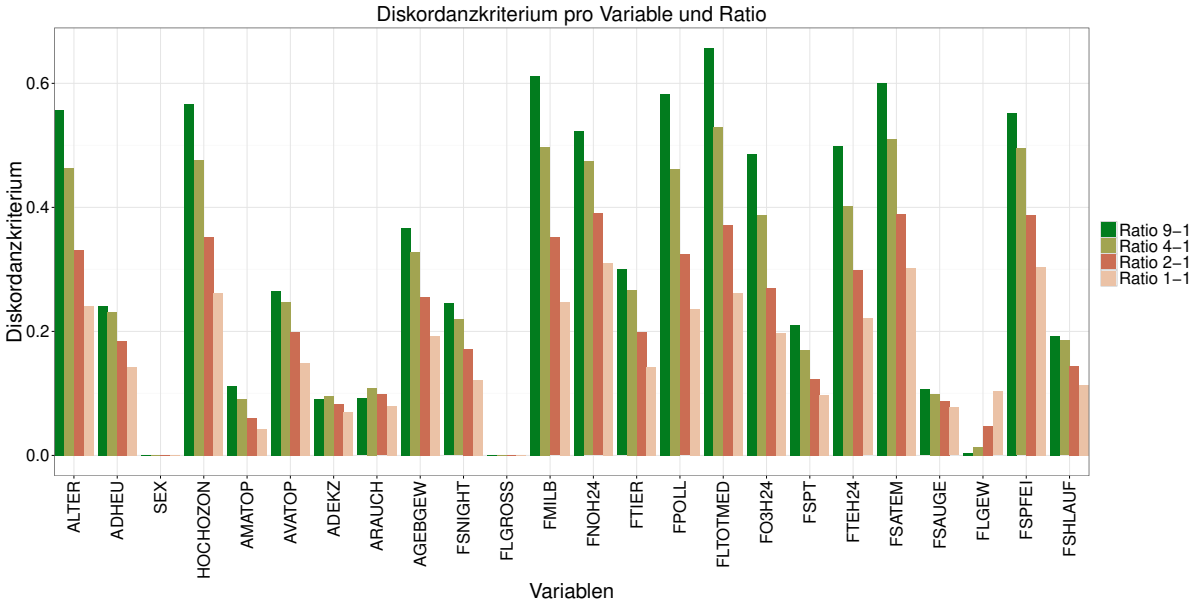

Abbildung 6: Übersicht der Diskordanzkriterien aus Sicht der Variablen des Datensatzes Ozone für alle Ratios bei einem Stichprobenumfang der Subdatensätze von jeweils $n/2$. Bei den Variablen SEX, FLGROSS und FLGEW liefert Ratio 1-1 nicht das kleinste Diskordanzkriterium.

als 1 in einem der beiden Subdatensätze aufzuweisen, scheint aber dennoch neben SEX und FLGROSS von deutlich stärkerer Relevanz für das Modell zu sein als die übrigen Variablen. Das heißt, diese drei Variablen scheinen am deutlichsten einen Einfluss auf die Zielgröße zu besitzen und scheinen deshalb von deutlicher Relevanz für das Modell zu sein. Sie müssen aber deshalb nicht den stärksten Einfluss im Sinne eines hohen betragsmäßigen Koeffizienten auf die Zielgröße haben. Hohe Inklusionshäufigkeiten einer Variable sagen lediglich aus, dass sie häufig in das Modell selektiert wurde, nicht jedoch, wie groß der Betrag ihres Koeffizienten ist.

Für Variablen, die deutlich Einfluss auf die Zielgröße haben, scheint also eine größere Ratio stabilere Ergebnisse zu liefern als kleinere Ratios, wie bei der Variable FLGEW zu erkennen ist. Für Variablen mit sehr deutlichem Einfluss wie SEX und FLGROSS betragen die Inklusionshäufigkeiten bei allen Ratios fast immer den Wert 1. Hier macht die Wahl der Ratio also keinen großen Unterschied in Bezug auf die Stabilität der Inklusionshäufigkeiten. Das zeigt sich insbesondere auch daran, dass die Diskordanzkriterien dieser Variablen 0 bzw. fast 0 sind, was in Abbildung 6 auf Seite 32 zu erkennen ist. Bei FLGEW jedoch macht die Wahl der Subsampling-Ratio etwas aus. Vermutlich hängt der Grund dafür damit zusammen, weshalb die Ratio 1-1 aus Sicht der Iterationen und bei den Variablen mit weniger deutlichem Einfluss auf die Zielgröße die stabileren Ergebnisse liefert als die anderen betrachteten Ratios. Je kleiner der Stichprobenumfang der Pseudostichproben im Vergleich zur Größe der Subdatensätze ist, desto mehr könnte eventuell ein gewisser „Verwaschungseffekt“ auftreten. Die oben schon angesprochenen Besonderheiten aus den Subdatensätzen werden mit schrumpfendem Stichprobenumfang der Pseudostichproben immer weniger auch in den Pseudostichproben auftreten. Deshalb könnte es sein, dass Variablen wie FLGEW, die einen deutlichen Einfluss haben, durch eine kleinere Ratio

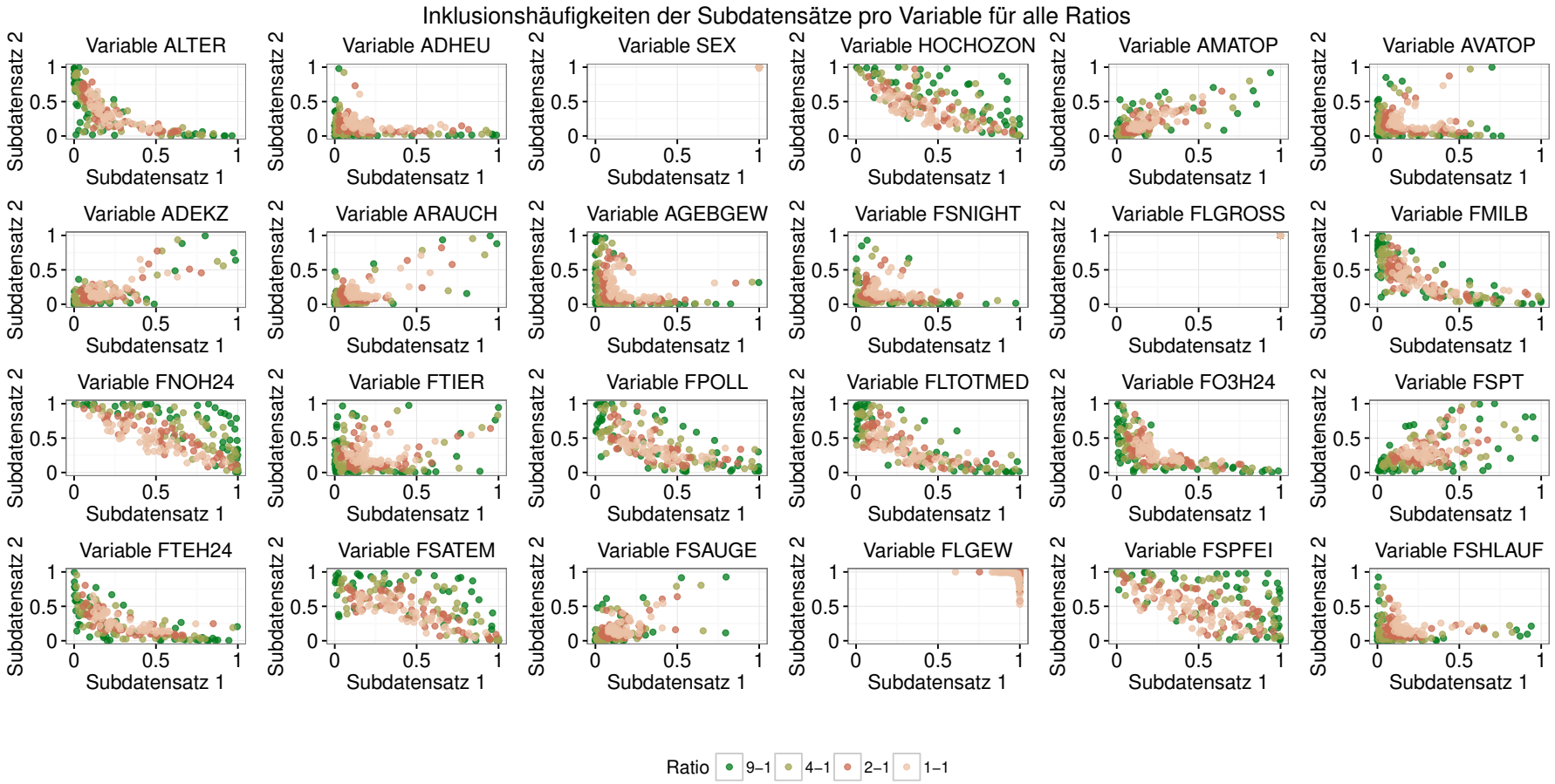

Abbildung 7: Übersicht der Inklusionshäufigkeiten pro Variable des Datensatzes Ozone für alle Ratios bei einem Stichprobenumfang der Subdatensätze von jeweils $n/2$. Die Variablen SEX, FLGROSS und FLGEW weisen bei allen Ratios hohe Inklusionshäufigkeiten auf.

ebenfalls ein wenig „verwaschen“ werden, sodass ihre Inklusionshäufigkeiten etwas instabiler werden. Variablen mit sehr deutlichem Einfluss wie SEX oder FLGROSS sind davon anscheinend nicht so stark betroffen. Inklusionshäufigkeiten von Variablen mit weniger deutlichem Einfluss hingegen sind bei einer niedrigen Ratio wie 1-1 stabiler, da sich der „Verwaschungseffekt“ hier wahrscheinlich positiv auswirkt. Mithilfe großer Ratios wie 9-1 generierte Pseudostichproben sind den Subdatensätzen sehr ähnlich und transportieren zufällige Besonderheiten solcher Variablen, die aufgrund eventuell unglücklicher Stichprobenziehung entstehen, wahrscheinlich eher in das Ergebnis der Modellauswahl. Je kleiner die Ratio ist, desto kleiner ist der Stichprobenumfang der Pseudostichproben und desto weniger tauchen wahrscheinlich solche Besonderheiten in den Pseudostichproben auf und beeinflussen somit dann weniger das Ergebnis der Modellauswahl.

Es macht also einen Unterschied, ob die Ratios anhand des Diskordanzkriteriums aus Sicht der Iterationen oder aus dem Blickwinkel der Variablen beurteilt werden. Bezüglich der Iterationen ist die Ratio 1-1 diejenige, welche die stabilsten Ergebnisse liefert. Werden die Variablen betrachtet, spielt die Deutlichkeit deren Einflusses eine Rolle. Für Variablen mit deutlichem Einfluss sind die Ergebnisse unter Verwendung der Ratio 9-1 stabiler, für Variablen mit sehr deutlichem Einfluss ist die Wahl der Ratio anscheinend egal. In Abbildung 7 auf Seite 33 ist allerdings am Beispiel der Variable FLGEW des Datensatzes Ozone zu erkennen, dass auch bei der aus Sicht dieser Variable ungeeignetsten Ratio 1-1 die Inklusionshäufigkeiten von FLGEW trotzdem in einem sehr hohen Bereich liegen und damit den Einfluss der Variable wohl gut widerspiegeln. In Abbildung 6 auf Seite 32 ist außerdem zu erkennen, dass das Diskordanzkriterium der Ratio 1-1 von FLGEW immer noch kleiner ist als das kleinste Diskordanzkriterium vieler anderer Variablen wie beispielsweise HOCHOZON, FSATEM oder FSPFEI. Daher scheint auch aus Sicht der Variablen beim Datensatz Ozone die Ratio 1-1 eine gute Wahl in Hinsicht auf die Stabilität zu sein.
Für den Datensatz Glioma können die Diskordanzkriterien ebenfalls aus Sicht der Variablen berechnet werden. Einen Überblick über die Diskordanzkriterien der Ratios für alle Variablen des Glioma-Datensatzes gibt Abbildung 8 auf Seite 35. Hier gibt es ebenfalls ein paar Variablen, bei denen nicht die Ratio 1-1 das kleinste Diskordanzkriterium liefert, sondern die Ratio 9-1 gefolgt von 4-1, 2-1 und 1-1 in dieser Reihenfolge. Es handelt sich um die Variablen gradd, ops und trt.
Wenn nun wie bei dem Ozone-Datensatz die Inklusionshäufigkeiten der Variablen in die Betrachtung miteinbezogen werden, ist in Abbildung 9 auf Seite 36 zu erkennen, dass die Variablen gradd, ops und trt sehr hohe Inklusionshäufigkeiten aufweisen. Die Verteilung der Punkte der Streudiagramme dieser Variablen unterscheidet sich stark von denen der übrigen Variablen. Die Variablen gradd, ops und trt scheinen einen klar deutlicheren Einfluss auf die Zielgröße zu haben als die anderen Variablen.
Auch beim Datensatz Glioma hängt also das Ergebnis, welche Ratio bei einer Variable die stabilsten Ergebnisse liefert, mit der Deutlichkeit des Einflusses der Variablen zusammen, wobei hier ebenfalls die Diskordanzkriterien der Ratio 1-1 der Variablen gradd, ops und trt kleiner sind als die geringsten Diskordanzkriterien von ein paar anderen Variablen des Datensatzes Glioma. Deshalb könnte auch bei der Betrachtung der Diskordanzkriterien aus Sicht der Variablen des Datensatzes Glioma die Ratio 1-1 als ein guter Kompromiss angesehen werden.
Bei dem PBC-Datensatz liefert in Abbildung 10 auf Seite 37 nur bei einer Variable eine

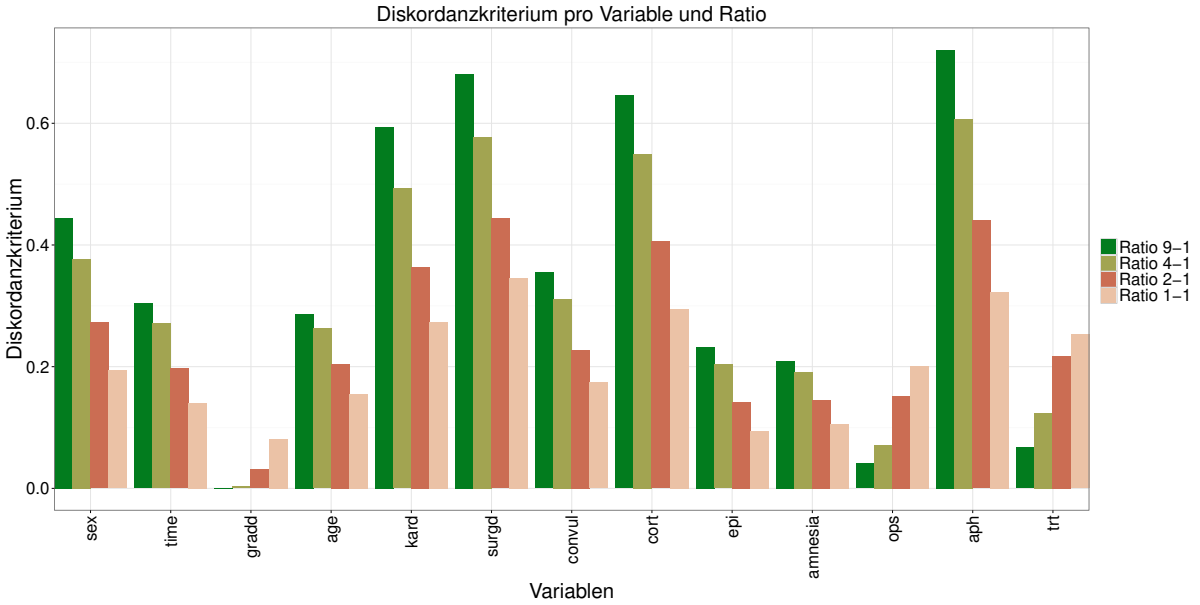

Abbildung 8: Übersicht der Diskordanzkriterien aus Sicht der Variablen des Datensatzes Glioma für alle Ratios bei einem Stichprobenumfang der Subdatensätze von jeweils $n/2$. Bei den Variablen gradd, ops und trt liefert Ratio 1-1 nicht das kleinste Diskordanzkriterium.

andere Ratio als Ratio 1-1 als die stabilsten Ergebnisse. Die Kovariable age besitzt mit der Ratio 9-1 das kleinste Diskordanzkriterium. Auch bei dieser Variable zeigt sich im Streudiagramm der Inklusionshäufigkeiten in Abbildung 11 auf Seite 38, dass age sehr hohe Inklusionshäufigkeiten aufweist und damit einen deutlichen Einfluss auf die Zielgröße zu haben scheint. Auch das Diskordanzkriterium der Ratio 1-1 der Variable age ist kleiner als die geringsten Diskordanzkriterien viele anderer Variablen des PBC-Datensatzes. Somit könnte auch hier die Ratio 1-1 als ein guter Kompromiss aus Sicht der Variablen angesehen werden.

6.2 Vergleich über Korrelationen

Die Inklusionshäufigkeiten können, wie in Kapitel 5.2.2 beschrieben, über die Betrachtung von Korrelationen auch mit den Waldstatistiken verglichen werden. Es werden auf den zwei Subdatensätzen jeder Iteration zum einen Waldstatistiken ermittelt und zum anderen Inklusionshäufigkeiten über die Resampling-basierte Variablenselektion berechnet. Also kann in jeder Iteration sowohl die Korrelation der Waldstatistiken der beiden Subdatensätze, als auch die Korrelation der Inklusionshäufigkeiten der beiden Subdatensätze betrachtet werden. Das gleiche kann ebenfalls für jede Variable durchgeführt werden.

6.2.1 Betrachtung der Iterationen

Zunächst werden die Korrelationskoeffizienten aus Sicht der Iterationen untersucht. In Abbildung 12 auf Seite 39 ist für den Datensatz Ozone ein Überblick über die Korrela-

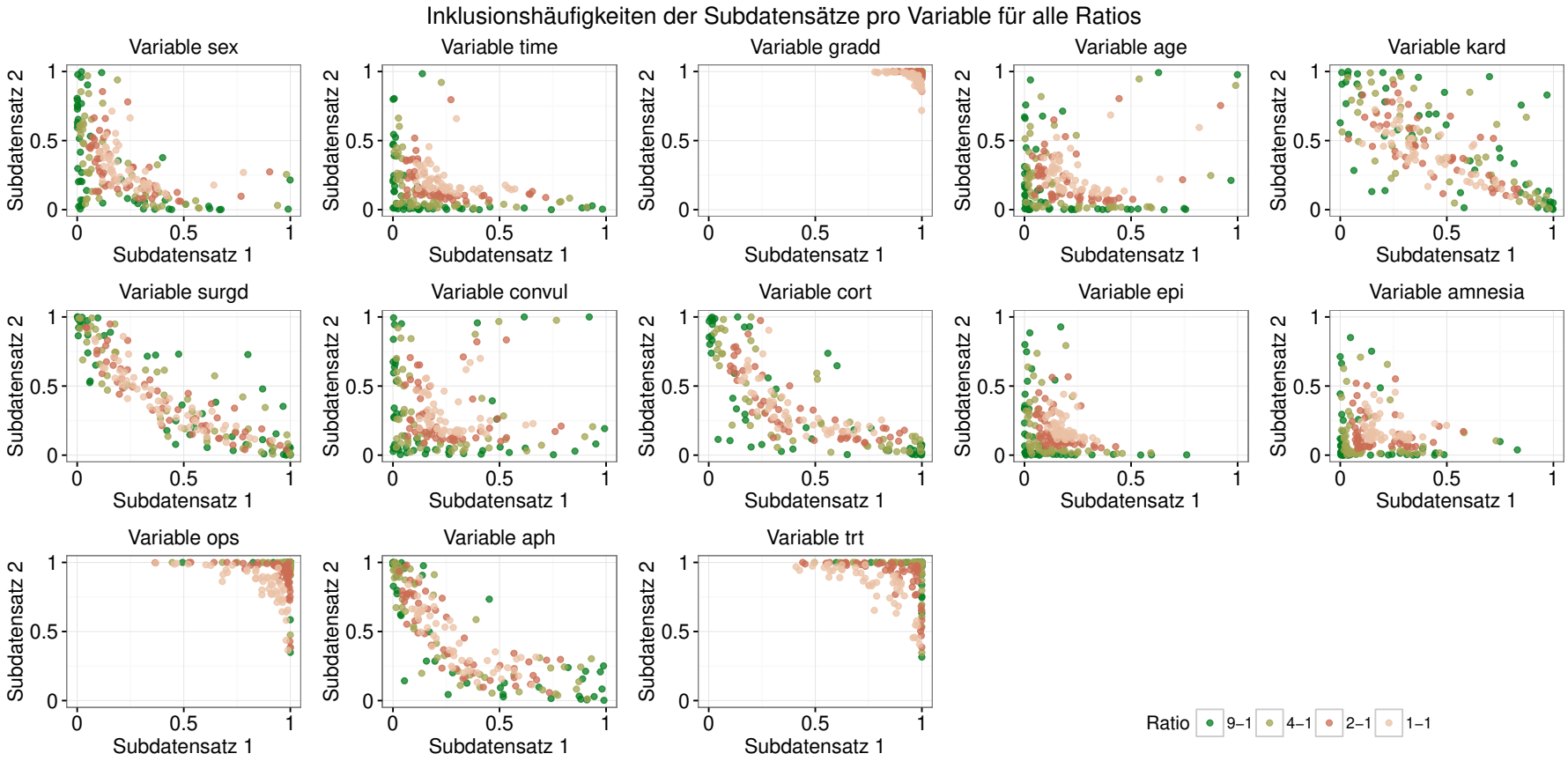

Abbildung 9: Übersicht der Inklusionshäufigkeiten pro Variable des Datensatzes Glioma für alle Ratios bei einem Stichprobenumfang der Subdatensätze von jeweils $n/2$. Die Variablen gradd, ops und trt weisen bei allen Ratios hohe Inklusionshäufigkeiten auf.

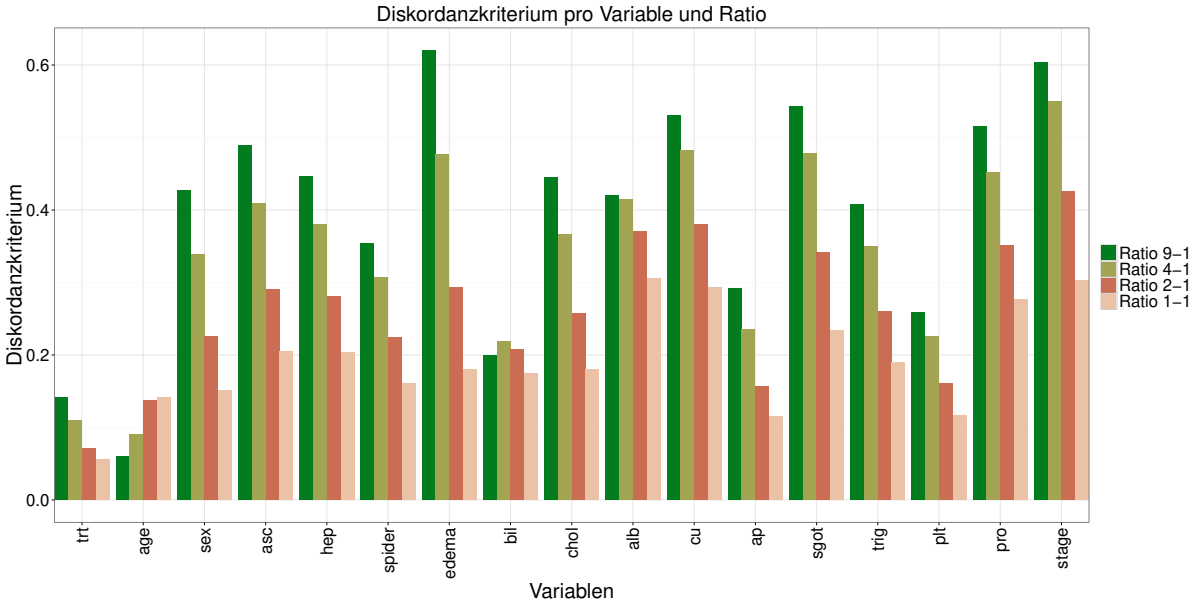

Abbildung 10: Übersicht der Diskordanzkriterien aus Sicht der Variablen des Datensatzes PBC für alle Ratios bei einem Stichprobenumfang der Subdatensätze von jeweils $n/2$. Bei der Variable age liefert Ratio 1-1 nicht das kleinste Diskordanzkriterium.

tionskoeffizienten gegeben. Es fällt auf, dass die meisten Korrelationskoeffizienten positiv sind.

Es gibt einige Iterationen, in denen die Korrelation der Inklusionshäufigkeiten der Ratio 1-1 am größten ist. Das bedeutet im Sinne der Stabilität, dass die Ratio 1-1 in diesen Iterationen die stabilsten Rankings liefert. Es gibt allerdings auch viele Iterationen, in denen nicht die Korrelation der Inklusionshäufigkeiten der Ratio 1-1 am größten ist. Werden also die Korrelation zwischen den Inklusionshäufigkeiten der beiden Subdatensätze als Kriterium verwendet, um die Ratios zu vergleichen, ist das Ergebnis nicht so eindeutig wie bei der Verwendung des Diskordanzkriteriums zuvor. Denn laut diesem war in allen Iterationen die Ratio 1-1 diejenige, die die stabilsten Ergebnisse liefert, während hier nur in 17 der 50 Iterationen die Ratio 1-1 den größten Korrelationskoeffizienten aufweist und demnach die stabilsten Ergebnisse liefert. Dennoch ist die Ratio 1-1 in den meisten Iterationen am höchsten korreliert. Insbesondere die Waldstatistiken sind nur in 13 Iterationen am höchsten korreliert. Für den Datensatz Ozone unterstützt also auch die Betrachtung der Korrelationen aus Sicht der Iterationen die Vermutung, dass die Ratio 1-1 die stabilsten Ergebnisse zu liefern scheint, insbesondere stabilere als die Waldstatistiken.

Beim Datensatz Glioma sind es in Abbildung 13 auf Seite 40 nur 14 Iterationen, in denen die Ratio 1-1 die höchste Korrelation besitzt. Insbesondere sind die Waldstatistiken in 29 Iterationen am höchsten korreliert. Damit ist die Ratio 1-1 immer noch diejenige, die unter den vier betrachteten Ratios in den meisten Iterationen die höchste Korrelation besitzt. Die Waldstatistiken sind allerdings noch häufiger am höchsten korreliert. Hier würde aus Sicht der Korrelationen geschlossen werden, dass eine einzige, auf dem gesamten Subdatensatz ausgeführte Analyse stabilere Ergebnisse liefert als eine Resampling-basierte Variablenselektion. Allerdings kann der Grund für das schlechte Abschneiden der Resampling-basierten Variablenselektion aus dem Blickwinkel der Korrelationen auch

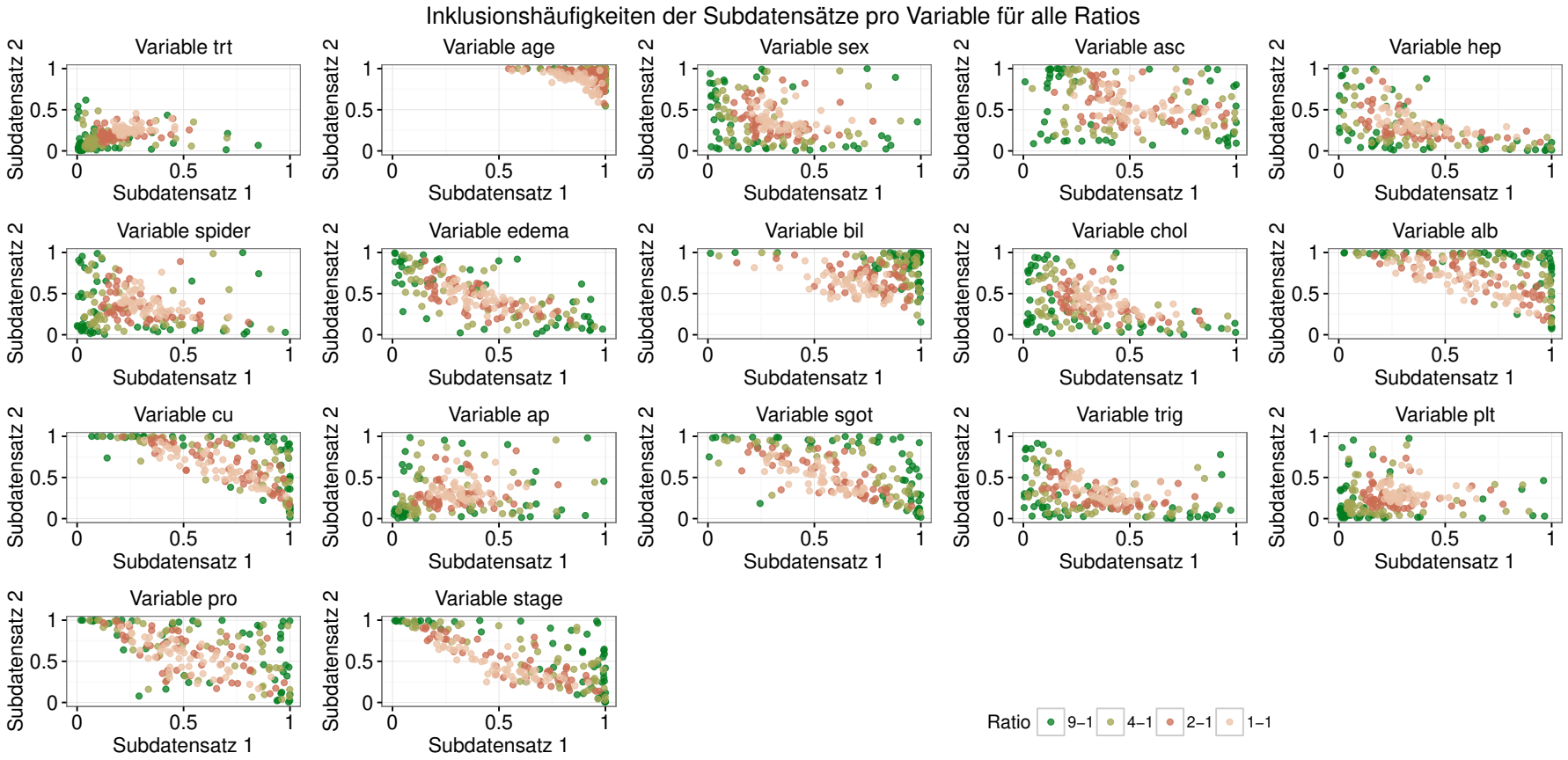

Abbildung 11: Übersicht der Inklusionshäufigkeiten pro Variable des Datensatzes PBC für alle Ratios bei einem Stichprobenumfang der Subdatensätze von jeweils $n/2$. Die Variable age weist bei allen Ratios hohe Inklusionshäufigkeiten auf.

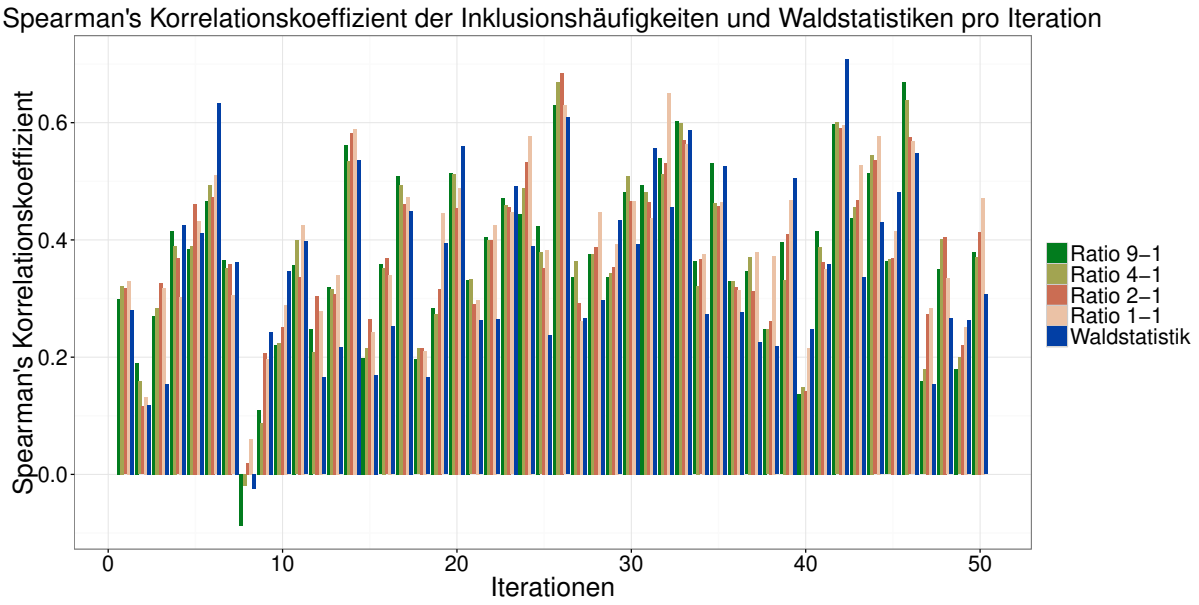

Abbildung 12: Übersicht der Korrelationskoeffizienten nach Spearman aus Sicht der Iterationen des Datensatzes Ozone für alle Ratios und die Waldstatistiken bei einem Stichprobenumfang der Subdatensätze von jeweils $n/2$. Ratio 1-1 liefert in den meisten Iterationen den höchsten Korrelationskoeffizienten.

sein, dass, wie in Kapitel 5.3 angedeutet, die Umfänge der Pseudostichproben eventuell so klein sind, dass Probleme beim Schätzen der Modelle aufgetreten sein können. Das könnte zu instabileren Rankings führen.

Sollte dies der Fall sein, wäre zu erwarten, dass auch beim PBC-Datensatz in den meisten Iterationen die Waldstatistiken am höchsten korreliert wären. Der Stichprobenumfang des PBC-Datensatzes ist noch geringer als derjenige des Glioma-Datensatzes, weshalb der Effekt eventuell sogar noch stärker zum Tragen kommen sollte. Wird jedoch Abbildung 14 auf Seite 41 betrachtet, ist zu erkennen, dass hier die Ratio 1-1 in 25 Iterationen am höchsten korreliert ist, während die Waldstatistiken nur in drei der 50 Iterationen die höchste Korrelation aufweisen. Die Korrelationen des PBC würden also auf den ersten Blick mitnichten die These stützen, dass bei den Ergebnissen der Überlebenszeitdatensätze die kleinen Stichprobenumfänge der Pseudostichproben zu schlechten Schätzungen und damit instabileren Rankings der Resampling-basierten Variablenselektion führen.
Werden jedoch die Korrelationen der Waldstatistiken des Datensatzes PBC genauer betrachtet, fällt einem ins Auge, dass - anders als bei den Datensätzen Ozone und Glioma - diese häufig negativ korreliert sind. Wie in Kapitel 5.2.2 erklärt, spielen die Vorzeichen der Waldstatistiken bei der Berechnung der Korrelationskoeffizienten keine Rolle, da der Betrag genommen wird. Es kann also sein, dass es durch den kleineren Stichprobenumfang des Datensatzes PBC im Vergleich zum Datensatz Glioma beim PBC-Datensatz bereits bei den Schätzungen der Waldstatistiken zu Problemen kommt. Die Waldstatistiken werden basierend auf den Subdatensätzen geschätzt, welche bei PBC kleiner sind als bei Glioma. Die Inklusionshäufigkeiten werden mit noch kleineren Stichprobenumfängen geschätzt. Bei der Ratio 1-1 insbesondere halbiert sich der Umfang noch einmal im Ver-

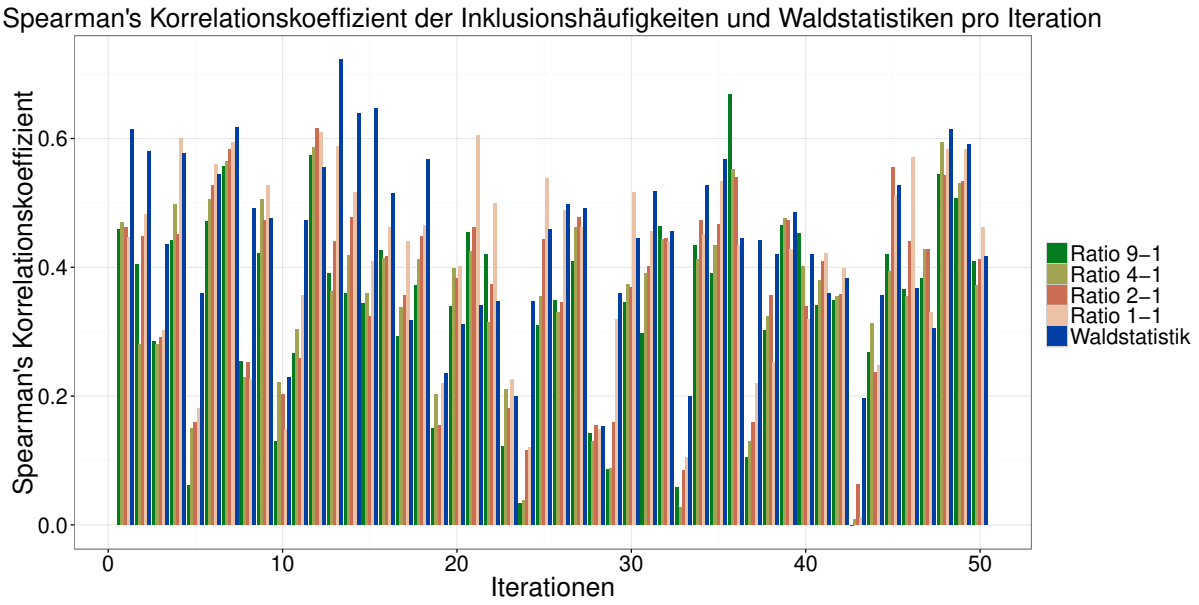

Abbildung 13: Übersicht der Korrelationskoeffizienten nach Spearman aus Sicht der Iterationen des Datensatzes Glioma für alle Ratios und die Waldstatistiken bei einem Stichprobenumfang der Subdatensätze von jeweils $n/2$. Die Waldstatistiken liefern in den meisten Iterationen den höchsten Korrelationskoeffizienten.

gleich zu den Subdatensätzen. Möglicherweise sind die Umfänge der Pseudostichproben des PBC-Datensatzes dann so klein, dass die Schätzungen der Modelle der Resampling-basierten Variablenselektion so schwierig werden, dass die Variablen ziemlich zufällig in den Modellen landen. Wird Abbildung 11 auf Seite 38 betrachtet, ist zu erkennen, dass bei einigen Variablen die Punkte mit abnehmender Ratio immer mehr in die Mitte des Streudiagramms wandern, also in Richtung des Punktes $(0.5, 0.5)$.
In diesem Falle würden die Inklusionshäufigkeiten des Glioma-Datensatzes und insbesondere des PBC-Datensatzes generell aufgrund zu kleiner Stichprobenumfänge nicht viel zur Beurteilung der Ratios aussagen. Auch die Diskordanzkriterien dieser Datensätze müssten in Frage gestellt werden. Dieser Aspekt sollte in Zukunft genauer untersucht werden, wenn möglich mit Überlebenszeitdatensätzen mit größerem Stichprobenumfang.

6.2.2 Betrachtung der Variablen

Die Korrelationen können ebenso wie die Diskordanzkriterien auch aus der Sicht der Variablen berechnet und betrachtet werden. Bei den Variablen des Datensatzes Ozone ist in Abbildung 15 auf Seite 42 zu erkennen, dass hier bei vielen Variablen die Inklusionshäufigkeiten und Waldstatistiken der beiden Subdatensätze negativ korreliert sind.
Wird beispielsweise die Variable FLGEW betrachtet, ist zu erkennen, dass alle Korrelationskoeffizienten dieser Variable negativ sind. Das Streudiagramm von FLGEW in Abbildung 7 auf Seite 33 zeigt, dass die Ränge der Inklusionshäufigkeiten der Variable FLGEW des einen Subdatensatzes teilweise recht entgegengesetzt zu den Rängen der Inklusionshäufigkeiten des anderen Subdatensatzes sind, weshalb die Korrelationskoeffizienten der Variable FLGEW für die verschiedenen Ratios negativ werden. Mit entgegengesetzten

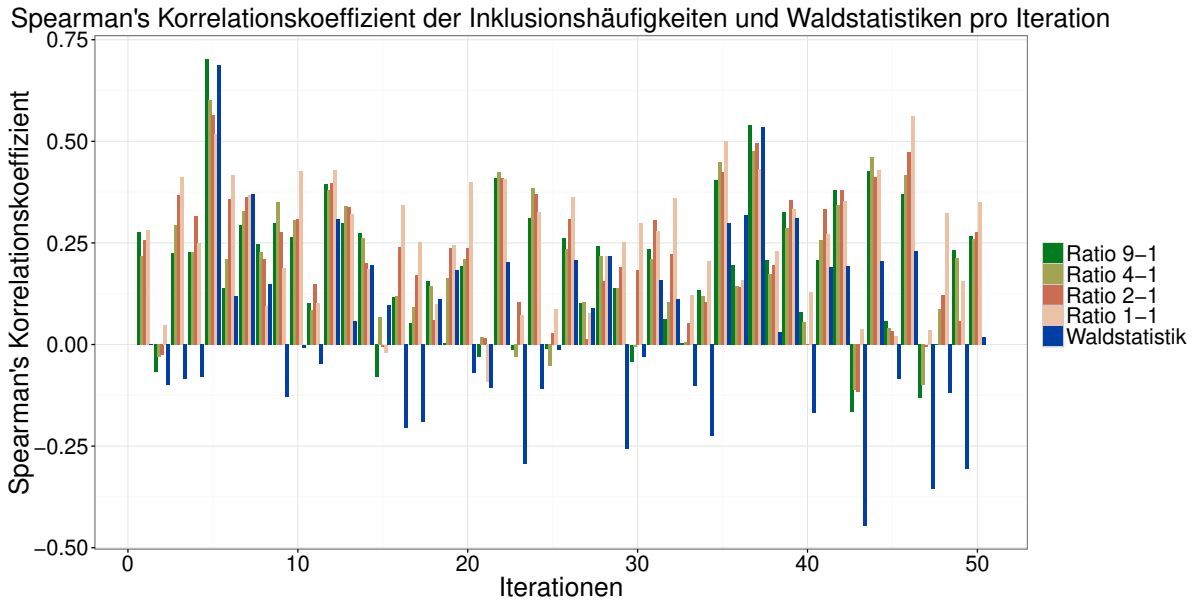

Abbildung 14: Übersicht der Korrelationskoeffizienten nach Spearman aus Sicht der Iterationen des Datensatzes PBC für alle Ratios und die Waldstatistiken bei einem Stichprobenumfang der Subdatensätze von jeweils $n/2$. Ratio 1-1 liefert in den meisten Iterationen den höchsten Korrelationskoeffizienten.

Rängen ist gemeint, dass die Punkte mit einem y-Wert von 1 und x-Werten, die etwas geringer als 1 sind, bei den y-Werten die höchsten Ränge besitzen, aber bei den x-Werten die niedrigsten Ränge. Genauso gibt es einige Punkte mit einem x-Wert von 1 und y-Werten, die etwas geringer als 1 sind. Diese Punkte haben bei den x-Werten die höchsten Ränge, aber bei den y-Werten die niedrigsten Ränge. Deshalb wird Spearmans's Korrelationskoeffizient negativ.
Trotzdem sind die Inklusionshäufigkeiten der Variable FLGEW objektiv betrachtet einigermaßen stabil, da die Inklusionshäufigkeiten der zusammengehörenden Subdatensätze größtenteils beide in einem sehr hohen Bereich liegen. Die Beurteilung der Stabilität der Ratios über den Korrelationskoeffizient nach Spearman bestraft eventuell durch die Verwendung der Ränge zu stark. Diese Vermutung wirft auch ein anderes Licht auf die obige Betrachtung der Iterationen.
Betrachtet werden nun zusätzlich noch die Variablen SEX und FLGROSS, welche den deutlichsten Einfluss auf die Zielgröße zu haben scheinen und insbesondere laut der Diskordanzkriterien sehr stabile Ergebnisse aufweisen. Ihre Korrelationskoeffizienten sind jedoch nahe 0 bzw. negativ. Ein Korrelationskoeffizient von 0 würde sich ergeben, wenn alle Punkte einer Variable aufgrund derselben Ränge in allen Iterationen dieselben x- und y-Werte bei der Berechnung des Korrelationskoeffizienten besitzen. Der negative Koeffizient der Variable SEX bei Ratio 1-1 ergibt sich zum Beispiel daher, dass die Variablen SEX, FLGROSS und FLGEW bei dieser Ratio die höchsten Ränge in fast allen Iterationen unter sich aufteilen, jedoch nicht in jeder Iteration in den Subdatensätzen untereinander dieselbe Reihenfolge haben. Dadurch ergeben sich verschiedene - zwar sehr hohe - Ränge für SEX in den Rankings der zwei Subdatensätze, die dann jedoch bei der Berechnung einen negativen Koeffizienten ergeben.

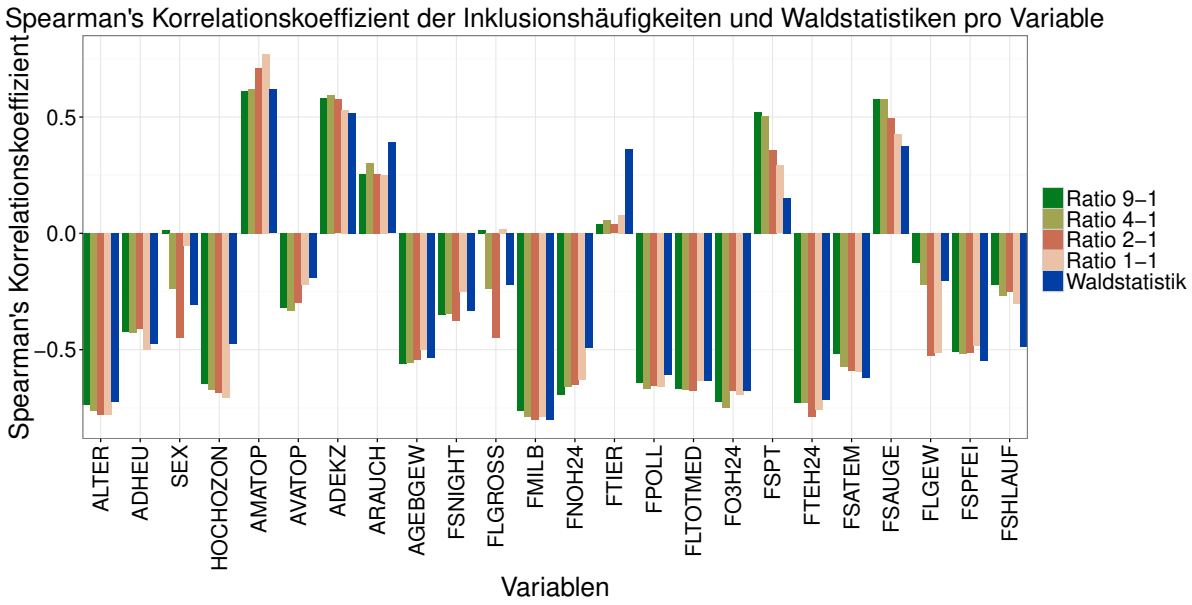

Abbildung 15: Übersicht der Korrelationskoeffizienten nach Spearman aus Sicht der Variablen des Datensatzes Ozone für alle Ratios und die Waldstatistiken bei einem Stichprobenumfang der Subdatensätze von jeweils $n/2$. Es existieren viele negative Korrelationskoeffizienten.

Intuitiv würde bei den Variablen SEX und FLGROSS allerdings ein Korrelationskoeffizient nahe bei 1 erwartet werden, da die Inklusionshäufigkeiten dieser Variablen objektiv betrachtet sehr stabil sind, wie in Abbildung 7 auf Seite 33 zu sehen ist. Aufgrund der Berechnung von Spearman's Korrelationskoeffizienten über die Ränge ergeben sich jedoch eher kontraintuitive Werte. Auch dies ist ein Punkt, über die Verwendung der Korrelationen der Rankings zur Beurteilung der Ratios nachzudenken.

Bei den Überlebenszeitdatensätzen Glioma und PBC sind ebenfalls die meisten Korrelationen negativ, die Grafiken hierzu befinden sich im Anhang. In den zugehörigen Streudiagrammen streuen die Punkte entsprechend stark. Teilweise kann in den Streudiagrammen dieser Datensätze auch recht gut ein gegensinniger Zusammenhang erkannt werden wie beispielsweise in Abbildung 9 auf Seite 36 in dem Streudiagramm der Variable surgd des Glioma-Datensatzes. Bei der Betrachtung der Korrelationen aus Sicht der Iterationen kommen Bedenken hinsichtlich der Stichprobenumfänge der Datensätze Glioma und PBC auf. Durch die Zweifel an der Eignung des auf den Rankings berechneten Korrelationskoeffizienten nach Spearman, die bei der Betrachtung aus dem Blickwinkel der Variablen aufkommen, werden diese Bedenken eventuell etwas relativiert. Insgesamt scheint die Betrachtung der Diskordanzkriterien ein sinnvolleres Maß zur Beurteilung der Stabilität der Ergebnisse Resampling-basierter Variablenselektion zu sein als die Untersuchung der Korrelationen der Rankings. Dennoch bleibt unabhängig davon das Problem zu kleiner Stichprobengrößen bei den Überlebenszeitdatensätzen bestehen.

6.3 Abhängigkeit von der Stichprobengröße

Anhand des Datensatzes Ozone wird noch ein weiterer Aspekt bezüglich der Stabilität Resampling-basierter Variablenselektion näher betrachtet. Der interessierende Gesichtspunkt ist, ob die Schlüsse bezüglich der Stabilität der Ergebnisse basierend auf verschiedenen Subsampling-Ratios auch auf Datensätze mit anderen Stichprobenumfängen als dem hier untersuchten Datensatz Ozone verallgemeinerbar sind. Dafür wird in zwei weiteren Untersuchungen der Datensatz Ozone in den 50 Iterationen nicht halbiert, sondern es werden zwei sich nicht überschneidende Subdatensätze der Größe $n/3$ bzw. $n/4$ gezogen. Auf diesen Subdatensätzen wird dann wie zuvor auf den Subdatensätzen des Umfangs $n/2$ jeweils eine Resampling-basierte Variablenselektion durchgeführt, wiederum jeweils für die vier betrachteten Ratios.

6.3.1 Stichprobenumfang n/3

Die Ergebnisse der Ziehung mit Stichprobenumfang $n/3$ bestätigen die Ergebnisse des Ozone-Datensatzes mit Halbierung des Datensatzes. Bei Betrachtung der Diskordanzkriterien aus Sicht der Iterationen weist die Ratio 1-1 stets das kleinste Kriterium auf, gefolgt von Ratio 2-1, Ratio 4-1 und Ratio 9-1. Die aus dem Blickwinkel der Variablen berechneten Diskordanzkriterien zeigen ebenfalls das gleiche Bild wie die Untersuchung basierend auf Subdatensätzen der Größe $n/2$. Die Variablen SEX, FLGROSS und FLGEW haben anscheinend den deutlichsten Einfluss auf die Zielgröße und bei ihnen liefert im Gegensatz zu den anderen Variablen nicht die Ratio 1-1 die stabilsten Ergebnisse. Die Grafiken der Diskordanzkriterien aus Sicht der Iterationen sowie der Variablen befinden sich im Anhang.

Werden die Streudiagramme der Inklusionshäufigkeiten der Variablen dieser Untersuchung mit denen der Untersuchung basierend auf den Subdatensätzen der Größe $n/2$ verglichen, ist zu erkennen, dass die Punkte in Abbildung 16 auf Seite 44 stärker streuen als in Abbildung 7 auf Seite 33. Besonders stark ist dies bei Variablen wie HOCHOZON, FNOH24, FPOLL oder FSPFEI zu sehen, deren Stabilität schon bei der Untersuchung basierend auf den Subdatensätzen der Größe $n/2$ nicht besonders groß ist. Grund für die stärkere Streuung ist wahrscheinlich der kleinere Stichprobenumfang, der zu etwas schlechteren Schätzungen der Modelle und damit zu etwas stärker streuenden Inklusionshäufigkeiten führt.

Bei den Korrelationen aus Sicht der Iterationen weist auch hier die Ratio 1-1 mit 20 von 50 in den meisten Iterationen die höchste Korrelation auf, gefolgt von den Waldstatistiken mit 13 Iterationen. Bis auf die Waldstatistiken von drei Iterationen sind alle Korrelationskoeffizienten positiv. Wird Spearman's Korrelationskoeffizienten für die Variablen berechnet, gibt es auch hier wie bei der Untersuchung der Subdatensätzen der Größe $n/2$ viele negative Korrelationen. Auch die Grafiken der Korrelationskoeffizienten aus dem Blickwinkel der Iterationen sowie der Variablen befinden sich im Anhang.

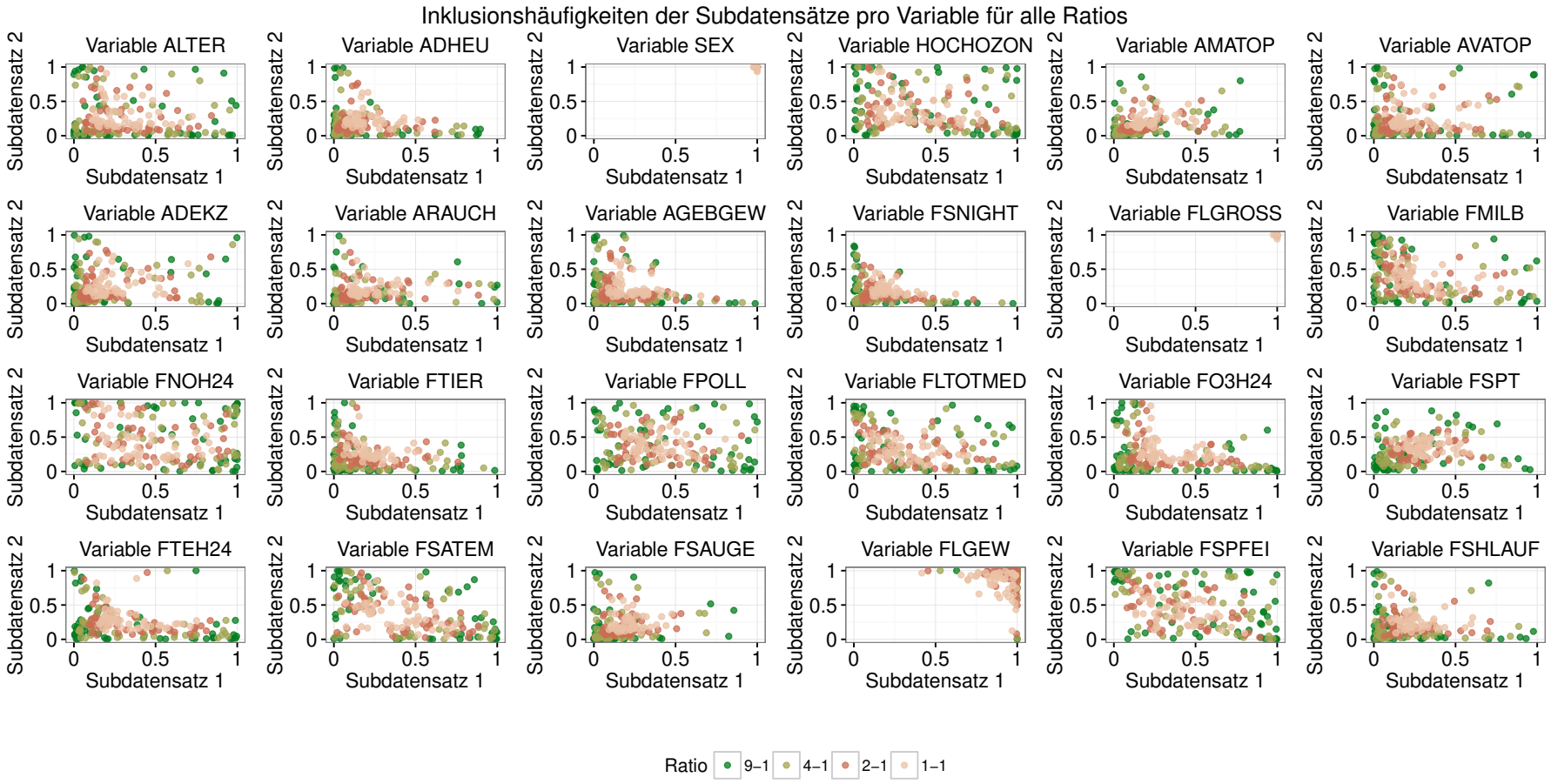

Abbildung 16: Übersicht der Inklusionshäufigkeiten pro Variable des Datensatzes Ozone für alle Ratios bei einem Stichprobenumfang der Subdatensätze von jeweils $n/3$. Die Inklusionshäufigkeiten streuen stärker als die der Analyse des Datensatzes Ozone mit einem Stichprobenumfang der Subdatensätze von jeweils $n/2$

6.3.2 Stichprobenumfang n/4

Auch bei den Subdatensätzen mit Stichprobenumfang $n/4$ ist nach den Diskordanzkriterien aus Sicht der Iterationen die Ratio 1-1 diejenige, die die stabilsten Ergebnisse liefert. Aus dem Blickwinkel der Variablen gesehen ist ebenfalls die Ratio 1-1 diejenige mit dem kleinsten Diskordanzkriterium außer bei den Variablen SEX, FLGROSS und FLGEW. Bei diesen weist wie auch bei den beiden anderen Untersuchungen die Ratio 1-1 das größte Diskordanzkriterium auf. Insbesondere liefert bei diesen dreien die Ratio 9-1 die stabilsten Ergebnisse.
In Abbildung 17 auf Seite 46 streuen die Punkte der Streudiagramme wie in Abbildung 16 auf Seite 44 stärker als bei der Untersuchung basierend auf den Subdatensätzen der Größe $n/2$. Insbesondere ist jetzt auch bei den Variablen SEX und FLGROSS eine Streuung erkennbar, während bei den anderen Untersuchungen diese Variablen fast überhaupt nicht streuen. Auch hier sind die meisten aus Sicht der Iterationen berechneten Korrelationskoeffizienten positiv. Hier liefern die Waldstatistiken mit 18 von 50 Iterationen am häufigsten die höchsten Korrelationskoeffizienten. Der Grund hierfür kann der nochmals deutlich kleinere Stichprobenumfang der Pseudostichproben im Vergleich zur Analyse mit einem Stichprobenumfang der Subdatensätze von $n/3$ sein, wodurch die Schätzung der Modelle auf den Pseudostichproben leiden kann. Für die Variablen berechnet sind wieder viele Koeffizienten negativ. Die Grafiken der Untersuchung mit Stichprobenumfang $n/4$ der Subdatensätze befinden sich ebenfalls im Anhang. Sie zeigen die Diskordanzkriterien aus Sicht der Iterationen sowie der Variablen und die Korrelationskoeffizienten aus dem Blickwinkel der Iterationen und der Variablen.

Insgesamt werden die Ergebnisse der Untersuchung des Datensatzes Ozone basierend auf Subdatensätzen der Größe $n/2$ von den Untersuchungen basierend auf Subdatensätzen mit kleineren Stichprobenumfängen bestätigt. Dies lässt die Vermutung zu, dass die Erkenntnisse über die Stabilität der Ergebnisse einer Resampling-basierten Variablenselektion unter Verwendung verschiedener Subsampling-Ratios nicht von dem Stichprobenumfang des Datensatzes abhängen. Das würde bedeuten, dass die Ratio 1-1 auch in Datensätzen mit anderem Stichprobenumfang als Ozone stabile Ergebnisse liefern sollte. Die Verteilung des Datensatzes ist ein weiterer Punkt, der bei der Verallgemeinerung der Ratio 1-1 auf andere Datensätze beachtet werden muss. In dieser Arbeit werden drei verschiedene Datensätze betrachtet, die jeweils ihrer eigenen Verteilung folgen. Die Ratio 1-1 liefert in allen drei Datensätzen bezüglich des Diskordanzkriteriums die stabilsten Ergebnisse. Allerdings sind die Ergebnisse der zwei Überlebenszeitdatensätze Glioma und PBC, wie schon erwähnt, etwas mit Vorsicht zu behandeln, sodass hier eine weitere Untersuchung stattfinden sollte. Zudem wurden Korrelationen zwischen den Kovariablen der Datensätze bei den Untersuchungen nicht beachtet.

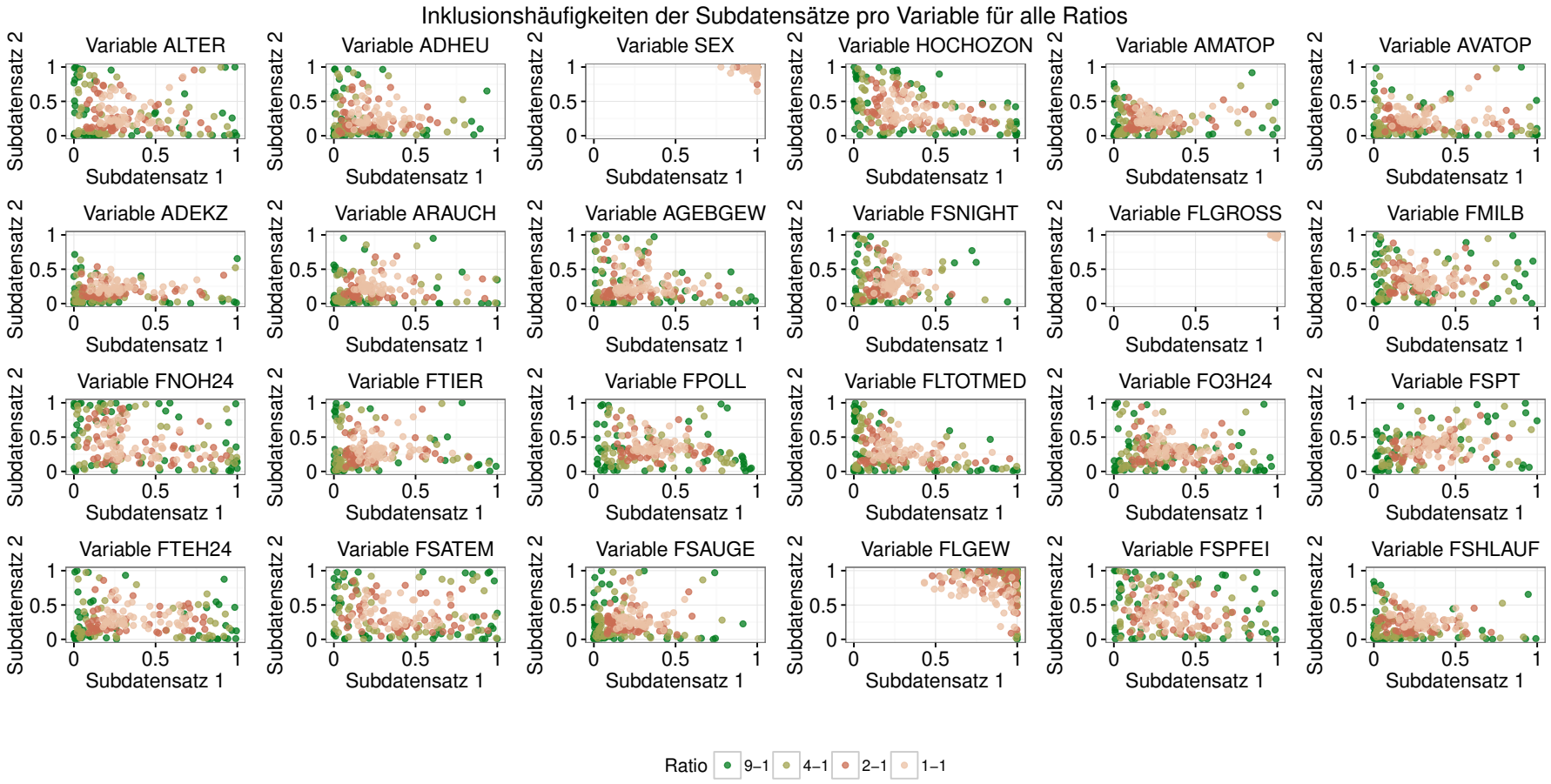

Abbildung 17: Übersicht der Inklusionshäufigkeiten pro Variable des Datensatzes Ozone für alle Ratios bei einem Stichprobenumfang der Subdatensätze von jeweils $n/4$. Die Inklusionshäufigkeiten streuen stärker als die der Analyse des Datensatzes Ozone mit einem Stichprobenumfang der Subdatensätze von jeweils $n/2$

7 Fazit

In dieser Arbeit wurde die Stabilität der Ergebnisse Resampling-basierter Variablenselektionen in Abhängigkeit unterschiedlicher Subsampling-Ratios betrachtet. Dies geschah über das Bilden von Subdatensätzen zur gegenseitigen Validierung. Die sogenannten Inklusionshäufigkeiten der Variablen wurden als Ergebnisse der Resampling-basierten Variablenselektionen auf ihre Stabilität hin betrachtet. Die untersuchten Ratios waren 9-1, 4-1, 2-1 und 1-1.
Zum einen wurde die Stabilität über Diskordanzkriterien beurteilt. Diese wurden auf zwei verschiedene Weisen berechnet, aus Sicht der Iterationen sowie aus dem Blickwinkel der Variablen. Die Kriterien, die sich aus Sicht der Iterationen ergaben, sprechen bei allen drei untersuchten Datensätzen dafür, dass die Ratio 1-1 die stabilsten Ergebnisse liefert. Aus dem Blickwinkel der Variablen spielt anscheinend die Deutlichkeit des Einflusses der Variablen eine Rolle. In allen Datensätzen scheinen bei Variablen, die einen deutlichen Einfluss auf die Zielgröße haben, größere Ratios stabilere Ergebnisse zu produzieren. Letztlich war bei diesen Variablen das Kriterium der Ratio 1-1 trotzdem noch besser als das beste Kriterium vieler anderer Variablen des jeweiligen Datensatzes. Aus diesem Grund scheint auch aus Sicht der Variablen Ratio 1-1 allgemein recht stabile Ergebnisse zu liefern.
Eine andere Möglichkeit, die Stabilität zu beurteilen, ist die Betrachtung von Rankings über Spearman's Korrelationskoeffizienten. Die Rankings wurden auf Basis der Inklusionshäufigkeiten der Resampling-basierten Variablenselektionen bzw. auf Basis der Waldstatistiken gebildet. Auch hier konnten die Korrelationskoeffizienten aus Sicht der Iterationen bzw. der Variablen berechnet werden. Die Resultate dieser Betrachtungsweise waren nicht so eindeutig wie die Ergebnisse der Diskordanzkriterien. Die Korrelationskoeffizienten aus Sicht der Iterationen sind bei allen Datensätzen meist positiv. Eine Ausnahme bilden die Koeffizienten der Ränge der Waldstatistiken des Datensatzes PBC. Aus dem Blickwinkel der Variablen gibt es bei allen Datensätzen viele negative Koeffizienten. Insgesamt ist die Eignung dieser Methode zur Beurteilung der Stabilität in Frage zu stellen, da auch für Variablen mit sehr stabilen Inklusionshäufigkeiten in einem sehr hohen Bereich negative Koeffizienten berechnet werden.
Für einen Datensatz wurden zusätzlich Subdatensätze mit kleineren Stichprobenumfängen betrachtet. Ziel war es zu untersuchen, ob die Stabilität und die damit verbundene Wahl der Ratio vom Stichprobenumfang des Datensatzes abhängen, auf dem eine Resampling-basierte Variablenselektion mittels Subsampling durchgeführt wird. Die Untersuchungen bestätigten die vorherigen Ergebnisse dieses Datensatzes. Es war in den Inklusionshäufigkeiten lediglich eine größere Streuung aufgrund des kleineren Stichprobenumfangs zu beobachten. Das legt die Vermutung nahe, dass unabhängig vom Stichprobenumfang die Ratio 1-1 bezüglich der Diskordanzkriterien die stabilsten Ergebnisse liefern sollte.
Ein weiterer Aspekt ist die Abhängigkeit der Stabilität der Ergebnisse einer Resampling-basierten Variablenselektion und der damit verbundenen Wahl der Subsampling-Ratio von der Verteilung der Daten. Aus diesem Grund wurden in dieser Arbeit drei verschiedene Datensätze betrachtet. Hinsichtlich der Diskordanzkriterien wiesen alle drei Datensätze die Ratio 1-1 als diejenige aus, die bei Resampling-basierter Variablenselektion die stabilsten Ergebnisse liefern sollte. In Zukunft sollte dieser Aspekt noch weiter untersucht werden. Wenn möglich sollten auch Überlebenszeitdatensätze mit einem ausreichend großen Stichprobenumfang gewählt werden, damit auch Untersuchungen bezüglich der Abhängigkeit

der Stabilität vom Stichprobenumfang durchgeführt werden können. Die Miteinbeziehung möglicher Korrelationen zwischen den Kovariablen der Datensätze ist ein weiterer Aspekt, der in Zukunft näher betrachtet werden sollte.

Literatur

[Altman u. Andersen 1989] Altman, Douglas G. ; Andersen, Per K.: Bootstrap investigation of the stability of a Cox regression model. In: *Statistics In Medicine* 8 (1989), Nr. 7, S. 771–783

[Auguie 2015] Auguie, Baptiste: *gridExtra: Miscellaneous Functions for "Grid"Graphics*, 2015. `http://CRAN.R-project.org/package=gridExtra`. – R package version 2.0.0

[de Bin u. a. 2016] Bin, Riccardo de ; Janitza, Silke ; Sauerbrei, Willi ; Boulesteix, Anne-Laure: Subsampling versus Bootstrapping in Resampling-Based Model Selection for Multivariable Regression. In: *Biometrics* 72 (2016), Nr. 1, S. 272–280

[Chen u. George 1985] Chen, Chen-Hsin ; George, Stephen L.: The bootstrap and identification of prognostic factors via Cox's proportional hazards regression model. In: *Statistics In Medicine* 4 (1985), Nr. 1, S. 39–46

[Copas u. Long 1991] Copas, J. B. ; Long, Tianyong: Estimating the Residual Variance in Orthogonal Regression with Variable Selection. In: *Journal Of The Royal Statistical Society: Series D (The Statistician)* 40 (1991), Nr. 1, S. 51–59

[Davison u. a. 2003] Davison, A. C. ; Hinkley, D. V. ; Young, G. A.: Recent Developments in Bootstrap Methodology. In: *Statistical Science* 18 (2003), Nr. 2, S. 141–157

[Efron u. Tibshirani 1986] Efron, B. ; Tibshirani, R.: Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. In: *Statistical Science* 1 (1986), Nr. 1, S. 54–75

[Fahrmeir u. a. 1996] Fahrmeir, Ludwig (Hrsg.) ; Hamerle, Alfred (Hrsg.) ; Tutz, Gerhard (Hrsg.): *Multivariate statistische Verfahren*. 2., überarb. Aufl. Berlin : Walter de Gruyter & Co., 1996

[Fahrmeir u. a. 2009] Fahrmeir, Ludwig ; Kneib, Thomas ; Lang, Stefan: *Regression: Modelle, Methoden und Anwendungen*. 2. Aufl. Berlin und Heidelberg : Springer, 2009

[Fahrmeir u. a. 2011] Fahrmeir, Ludwig ; Künstler, Rita ; Pigeot, Iris ; Tutz, Gerhard: *Statistik: Der Weg zur Datenanalyse*. 7. Aufl., korr. Nachdr. Berlin und Heidelberg : Springer, 2011

[Gifi 1990] Gifi, Albert: *Nonlinear Multivariate Analysis*. 1. Aufl. Chichester : Wiley, 1990

[Henderson u. Velleman 1981] Henderson, Harold V. ; Velleman, Paul F.: Building Multiple Regression Models Interactively. In: *Biometrics* 37 (1981), Nr. 2, S. 391–411

[Ihaka u. a. 2015] Ihaka, Ross ; Murrell, Paul ; Hornik, Kurt ; Fisher, Jason C. ; Zeileis, Achim: *colorspace: Color Space Manipulation*, 2015. `http://CRAN.R-project.org/package=colorspace`. – R package version 1.2-6

[Ihorst u. a. 2004] Ihorst, G. ; Frischer, T. ; Horak, F. ; Schumacher, M. ; Kopp, M. ; Forster, J. ; Mattes, J. ; Kuehr, J.: Long- and medium-term ozone effects on lung growth including a broad spectrum of exposure. In: *European Respiratory Journal* 23 (2004), Nr. 2, S. 292–299

[Janitza u. a. 2016] Janitza, Silke ; Binder, Harald ; Boulesteix, Anne-Laure: Pitfalls of hypothesis tests and model selection on bootstrap samples: Causes and consequences in biometrical applications. In: *Biometrical Journal* 58 (2016), Nr. 3, S. 447–473

[Meinshausen u. Bühlmann 2010] Meinshausen, Nicolai ; Bühlmann, Peter: Stability selection. In: *Journal Of The Royal Statistical Society: Series B (Statistical Methodology)* 72 (2010), Nr. 4, S. 417–473

[Miller 1984] Miller, Alan J.: Selection of Subsets of Regression Variables. In: *Journal Of The Royal Statistical Society: Series A (General)* 147 (1984), Nr. 3, S. 389–425

[Miller 2002] Miller, Alan J.: *Subset Selection in Regression.* 2. Aufl. Boca Raton : Chapman & Hall / CRC, 2002

[Moore 2016] Moore, Dirk F.: *Applied Survival Analysis Using R.* 1. Aufl. Schweiz : Springer, 2016

[Quevedo u. a. 2007] Quevedo, José R. ; Bahamonde, Antonio ; Luaces, Oscar: A simple and efficient method for variable ranking according to their usefulness for learning. In: *Computational Statistics & Data Analysis* 52 (2007), Nr. 1, S. 578–595

[R Core Team 2014] R Core Team: *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing, 2014. `http://www.R-project.org/`

[Rospleszcz u. a. 2016] Rospleszcz, Susanne ; Janitza, Silke ; Boulesteix, Anne-Laure: Categorical variables with many categories are preferentially selected in bootstrap-based model selection procedures for multivariable regression models. In: *Biometrical Journal* 58 (2016), Nr. 3, S. 652–673

[Royston u. Sauerbrei 2008] Royston, Patrick ; Sauerbrei, Willi: *Multivariable model-building: A pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables.* 1. Aufl. Chichester : Wiley, 2008

[Sauerbrei u. a. 2011] Sauerbrei, Willi ; Boulesteix, Anne-Laure ; Binder, Harald: Stability investigations of multivariable regression models derived from low- and high-dimensional data. In: *Journal Of Biopharmaceutical Statistics* 21 (2011), Nr. 6, S. 1206–1231

[Sauerbrei u. a. 2015] Sauerbrei, Willi ; Buchholz, Anika ; Boulesteix, Anne-Laure ; Binder, Harald: On stability issues in deriving multivariable regression models. In: *Biometrical Journal* 57 (2015), Nr. 4, S. 531–555

[Sauerbrei u. a. 2007] Sauerbrei, Willi ; Royston, Patrick ; Binder, Harald: Selection of important variables and determination of functional form for continuous predictors in multivariable model building. In: *Statistics In Medicine* 26 (2007), Nr. 30, S. 5512–5528

[Sauerbrei u. Schumacher 1992] Sauerbrei, Willi ; Schumacher, Martin: A bootstrap resampling procedure for model building: Application to the cox regression model. In: *Statistics In Medicine* 11 (1992), Nr. 16, S. 2093–2109

[Schlittgen 2013] Schlittgen, Rainer: *Regressionsanalysen mit R*. 1. Aufl. München : Oldenbourg Wissenschaftsverlag GmbH, 2013

[Therneau 2015] Therneau, Terry M.: *A Package for Survival Analysis in S*, 2015. `https://CRAN.R-project.org/package=survival`. – R package version 2.38

[Therneau u. Grambsch 2000] Therneau, Terry M. ; Grambsch, Patricia M.: *Modeling Survival Data: Extending the Cox Model*. 1. Aufl. New York : Springer, 2000

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

[Wickham 2009] Wickham, Hadley: *ggplot2: Elegant Graphics for Data Analysis*. 1. Aufl. New York : Springer, 2009

[Wickham 2012] Wickham, Hadley: *gtable: Arrange grobs in tables*, 2012. `http://CRAN.R-project.org/package=gtable`. – R package version 0.1.2

[Wu 1986] Wu, C.F.J.: Jackknife, bootstrap and other resampling methods in regression analysis. In: *The Annals Of Statistics* 14 (1986), Nr. 4, S. 1261–1295

[Xie 2016] Xie, Yihui: *testit: A Simple Package for Testing R Packages*, 2016. `http://CRAN.R-project.org/package=testit`. – R package version 0.5

[Zeileis u. a. 2009] Zeileis, Achim ; Hornik, Kurt ; Murrell, Paul: Escaping RGBland: Selecting Colors for Statistical Graphics. In: *Computational Statistics & Data Analysis* 53 (2009), Nr. 9, S. 3259–3270

Anhang

Inhalt der CD

Die beigelegte CD enthält vier Ordner.
Im Ordner R-Code sind alle im Rahmen dieser Arbeit erstellte R-Skripte enthalten. Die Datei clean_data.R enthält den Code zur Aufbereitung der Datensätze. In simulation_stepAIC.R ist die in Kapitel 5.3 angesprochene Simulation zu stepAIC enthalten. Die Datei functions.R definiert die Funktionen, mit deren Hilfe in analysis.R dann die Analysen durchgeführt werden. Für die Ergebnisse dieser Analysen werden Grafiken mithilfe der Dateien in results_ozone.R, results_glioma.R, results_pbc.R, results_ozone_size_1_2.R und results_ozone_size_1_3.R erstellt. In den R-Skripten analysis_warnungen.R und functions_warnungen.R werden die in Kapitel 5.3 angesprochenen Warnmeldungen untersucht.
Der Ordner Ergebnisse ist in mehrere Unterordner gegliedert. Diese Unterordner Ozone, Glioma, PBC, Ozone_1_2 und Ozone_1_3 beinhalten jeweils alle Grafiken, die für die Datensätze Ozone, Glioma und PBC erstellt wurden, im PDF-Format. Dabei sind in Ozone die Grafiken der Untersuchung mit einem Stichprobenumfang der Subdatensätze von $n/2$ gespeichert, während in Ozone_1_2 und Ozone_1_3 die Grafiken der Analysen mit Stichprobenumfang $n/3$ bzw. $n/4$ der Subdatensätze enthalten sind. Zusätzlich zu den Ordnern mit den Grafiken sind in dem Ordner Ergebnisse alle RData-Dateien mit den Ergebnissen der Analysen in R abgespeichert.
Die RData-Dateien bzw. txt-Dateien, die als Log-Dateien während der Analysen in R bzw. zur Untersuchung von Warnmeldungen erstellt wurden, befinden sich im dritten Ordner der CD, dem Ordner Log.
Im vierten Ordner der CD mit dem Namen Daten sind die Datensätze Ozone, Glioma und PBC gespeichert. Die Dateien ozone_short.csv, glioma_short.csv und pbc_short.csv enthalten die aufbereiteten und bei den Analysen verwendeten Datensätze Ozone, Glioma und PBC. Die Originaldaten des Datensatze Ozone sind die Datei ozone_reduced.txt, die Originaldaten der Datensätze Glioma und PBC sind in den Ordnern glioma und pbc enthalten, welche aus den zip-Archiven pbc.zip und glioma.zip stammen. Die Ordner glioma und pbc wurden aus den zip-Archiven extrahiert, um das Einlesen der Daten in R [R Core Team, 2014] zu ermöglichen.

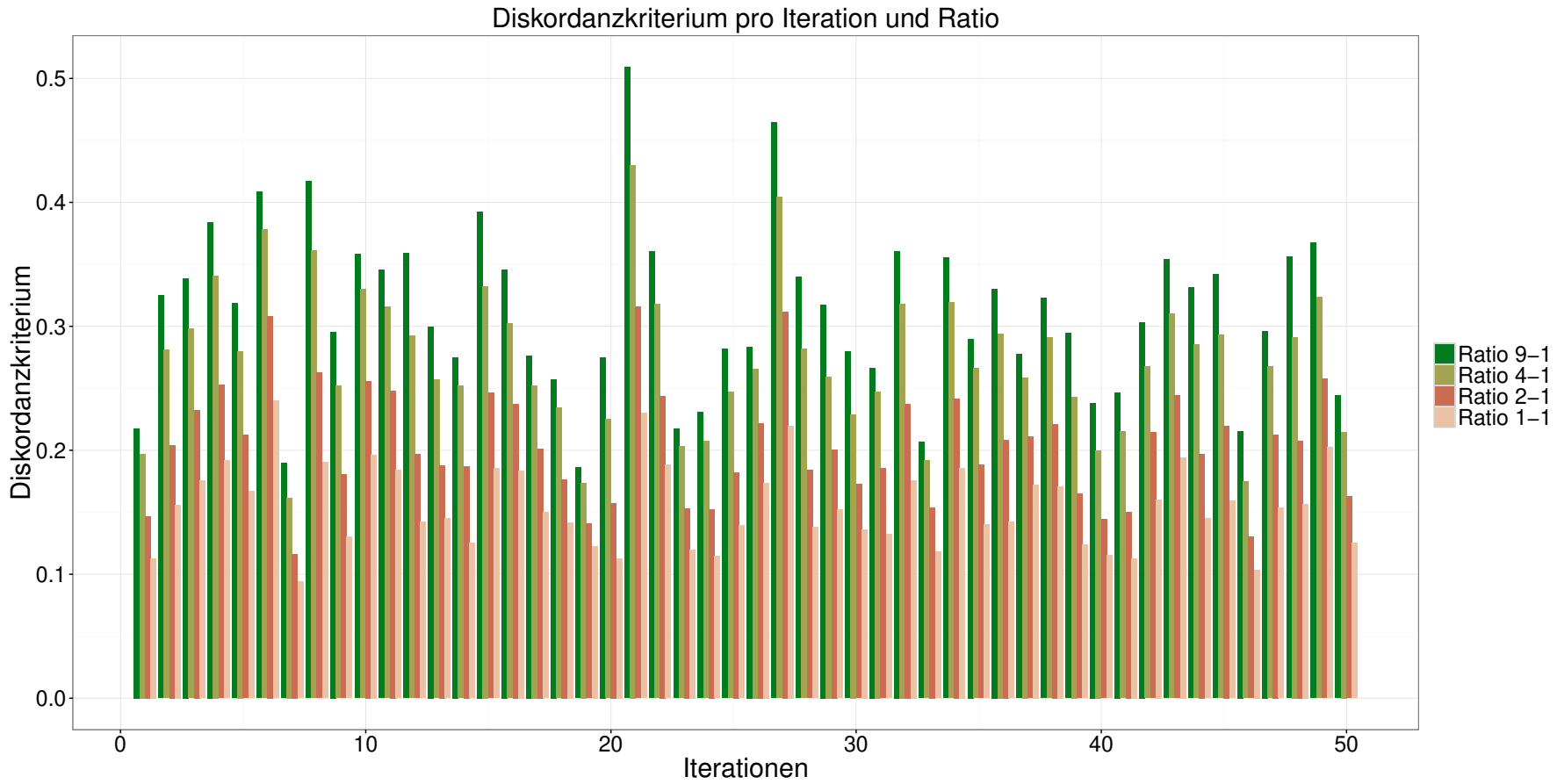

Abbildung 18: Übersicht der Diskordanzkriterien aus Sicht der Iterationen des Datensatzes Ozone für alle Ratios bei einem Stichprobenumfang der Subdatensätze von jeweils $n/3$. In allen Iterationen liefert Ratio 1-1 das kleinste Diskordanzkriterium.

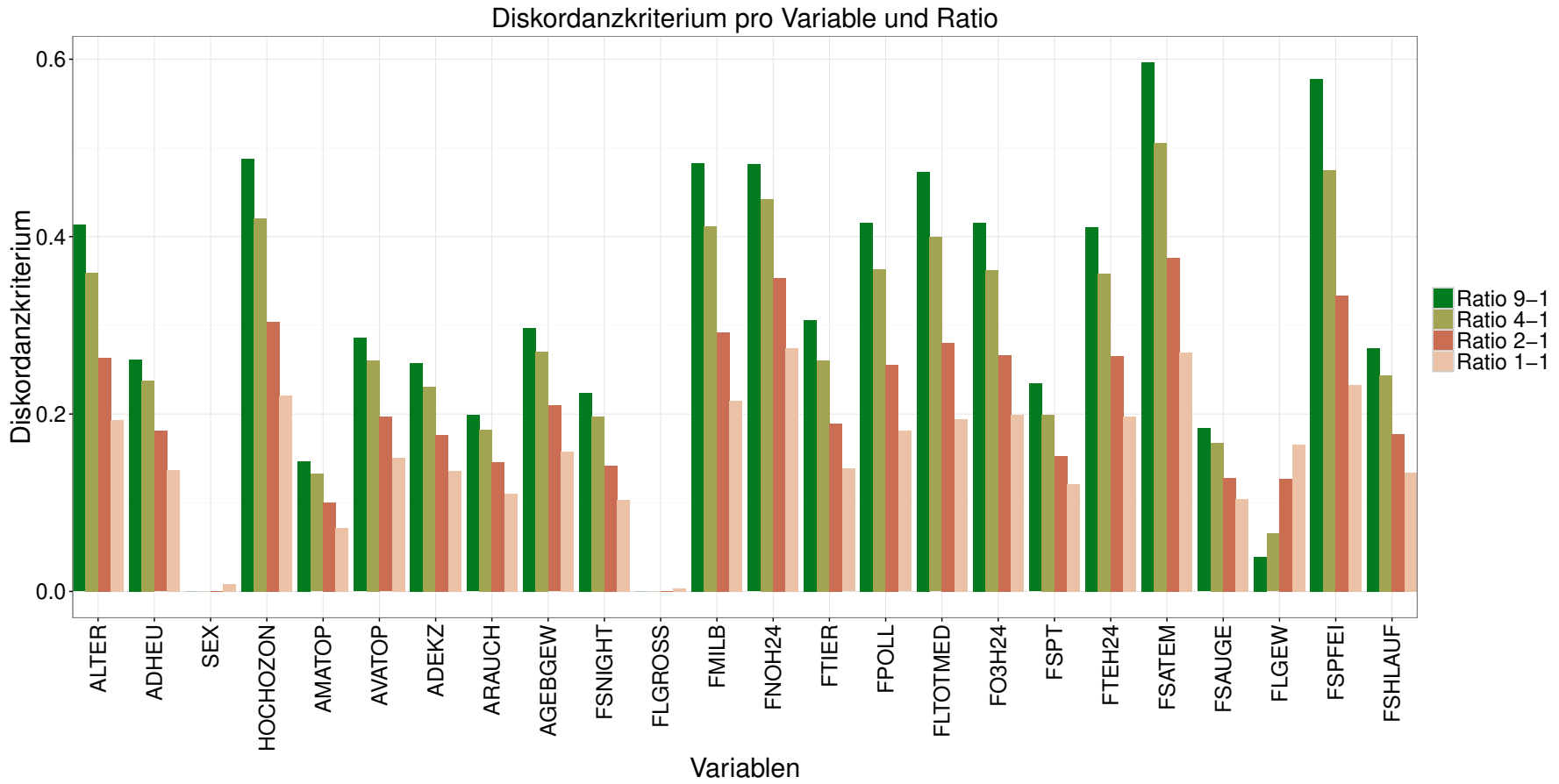

Abbildung 19: Übersicht der Diskordanzkriterien aus Sicht der Variablen des Datensatzes Ozone für alle Ratios bei einem Stichprobenumfang der Subdatensätze von jeweils $n/3$. Bei den Variablen SEX, FLGROSS und FLGEW liefert Ratio 1-1 nicht das kleinste Diskordanzkriterium.

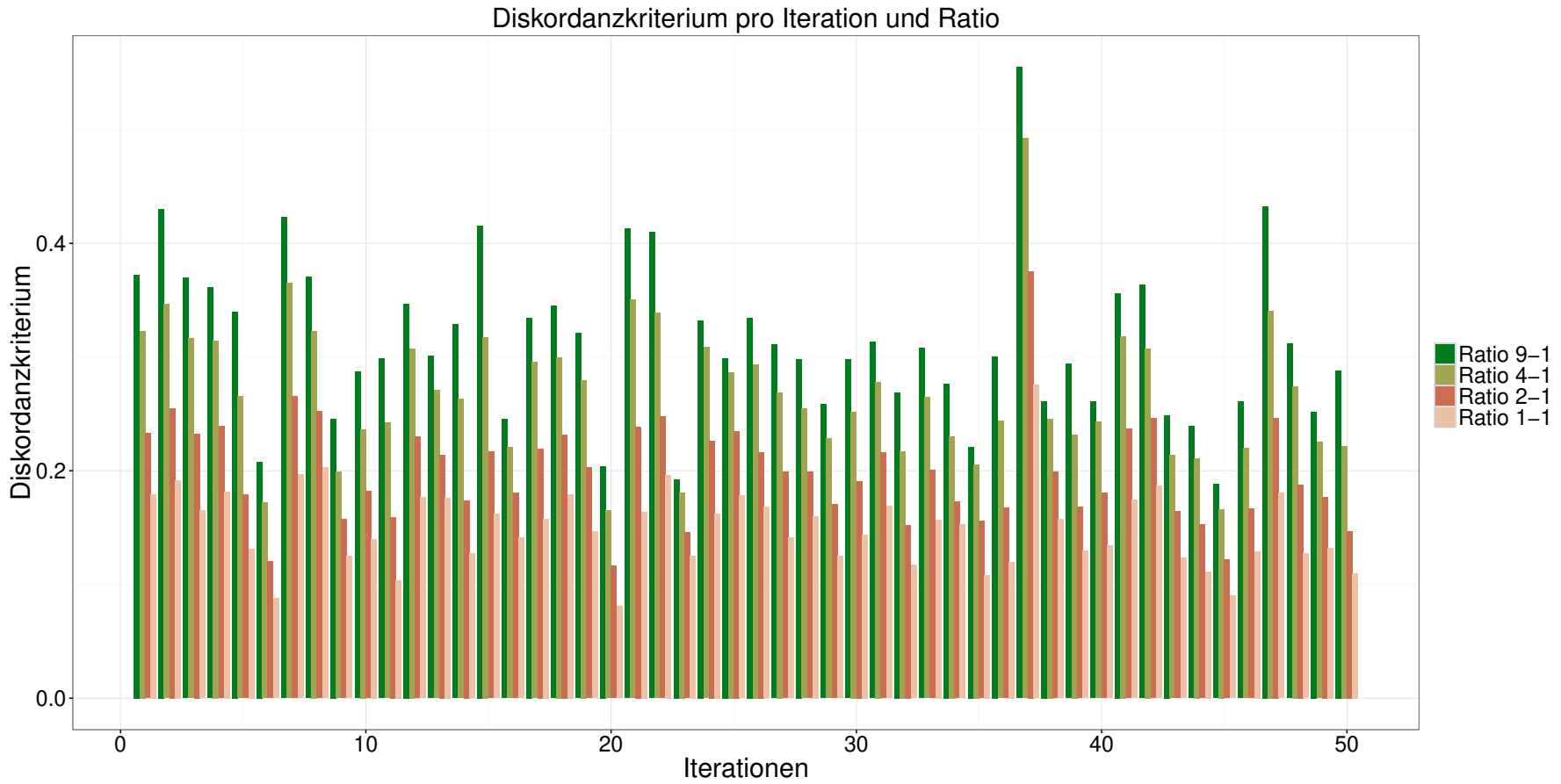

Abbildung 20: Übersicht der Diskordanzkriterien aus Sicht der Iterationen des Datensatzes Ozone für alle Ratios bei einem Stichprobenumfang der Subdatensätze von jeweils $n/4$. In allen Iterationen liefert Ratio 1-1 das kleinste Diskordanzkriterium.

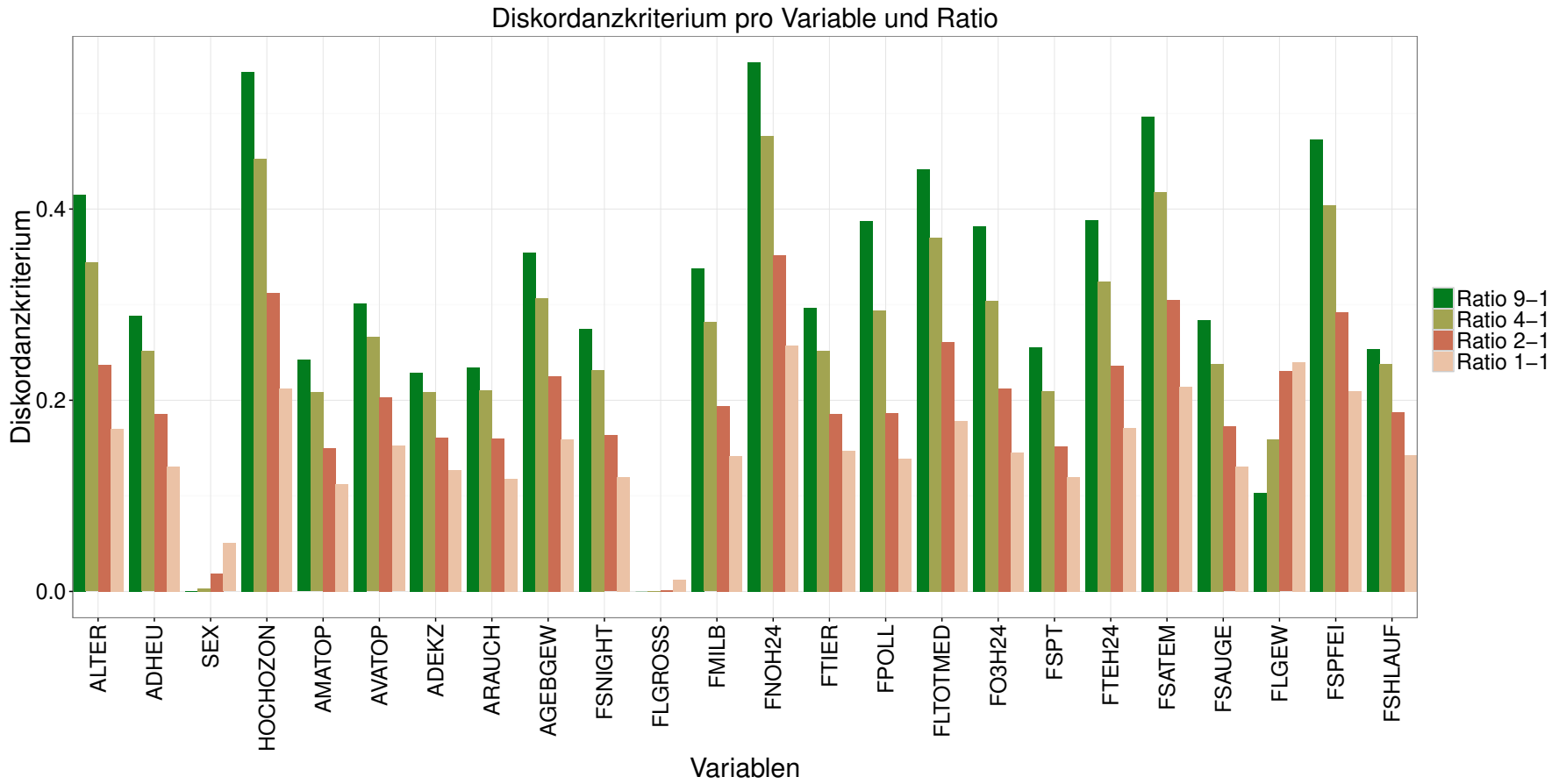

Abbildung 21: Übersicht der Diskordanzkriterien aus Sicht der Variablen des Datensatzes Ozone für alle Ratios bei einem Stichprobenumfang der Subdatensätze von jeweils $n/4$. Bei den Variablen SEX, FLGROSS und FLGEW liefert Ratio 1-1 nicht das kleinste Diskordanzkriterium.

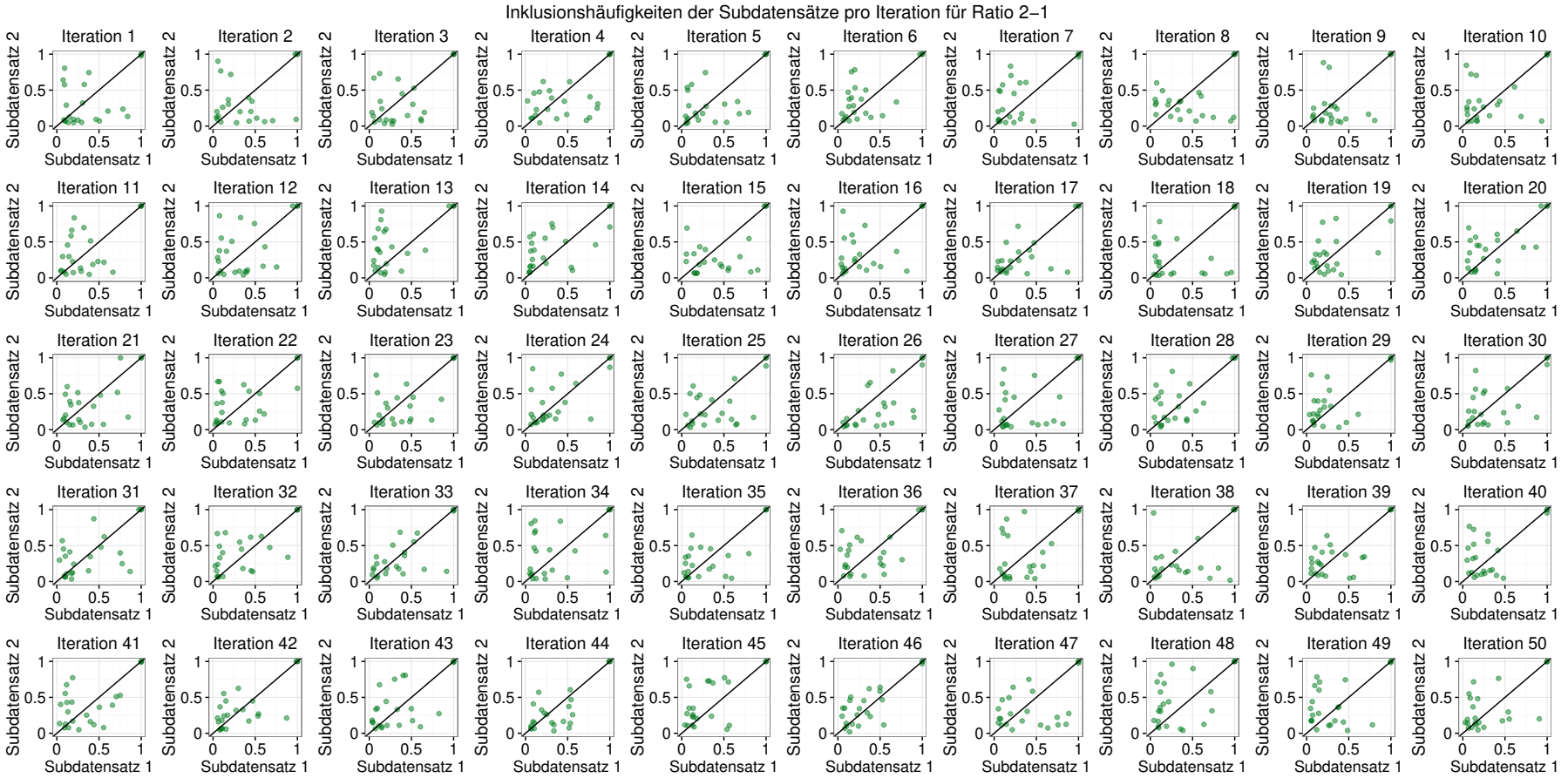

Abbildung 22: Übersicht der Inklusionshäufigkeiten pro Iteration des Datensatzes Ozone für die Ratio 2-1 bei einem Stichprobenumfang der Subdatensätze von jeweils $n/2$. Die Inklusionshäufigkeiten streuen hier stärker als in der entsprechenden Übersicht der Ratio 1-1.

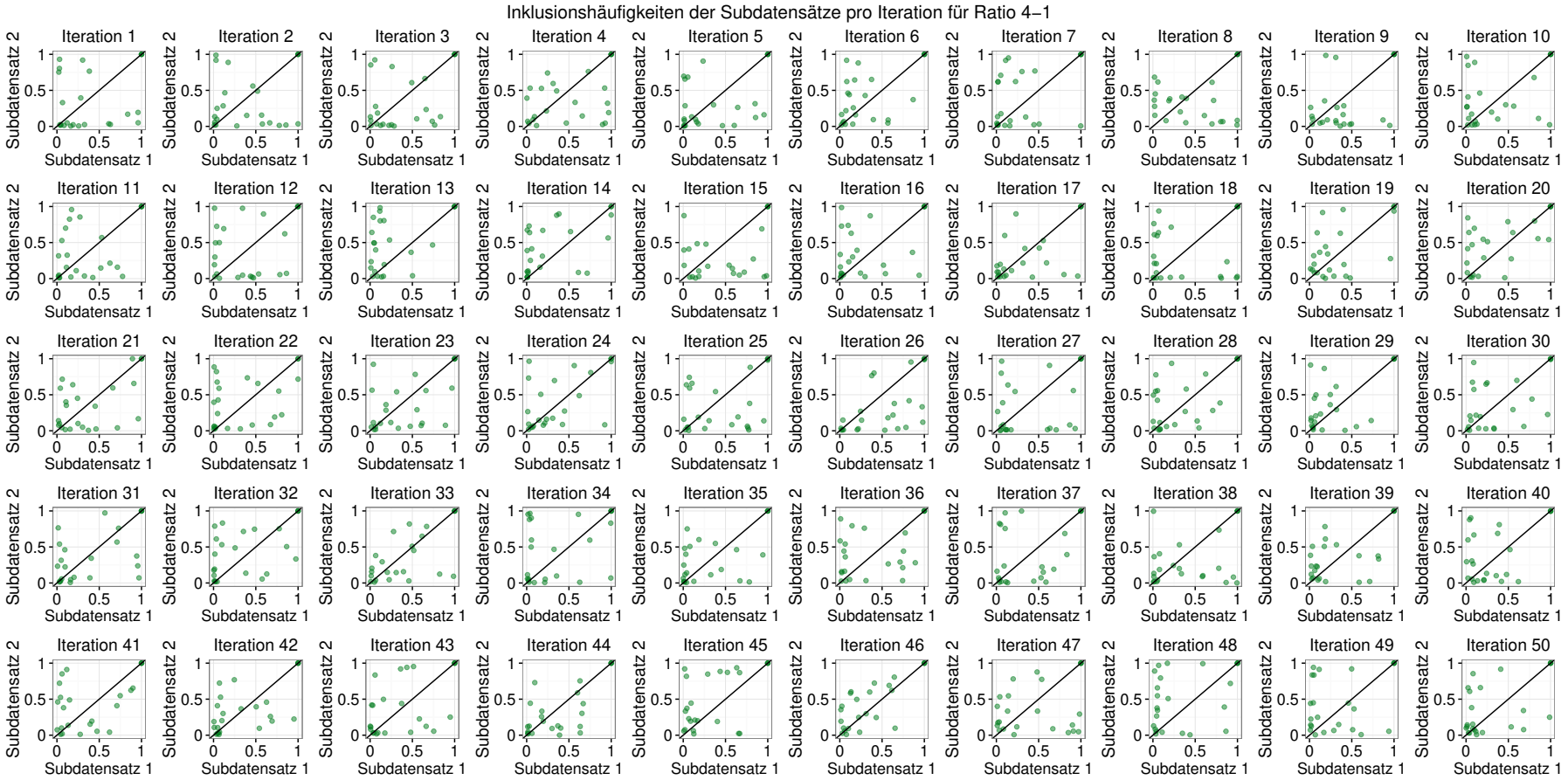

Abbildung 23: Übersicht der Inklusionshäufigkeiten pro Iteration des Datensatzes Ozone für die Ratio 4-1 bei einem Stichprobenumfang der Subdatensätze von jeweils $n/2$. Die Inklusionshäufigkeiten streuen hier weniger stark als in der entsprechenden Übersicht der Ratio 9-1.

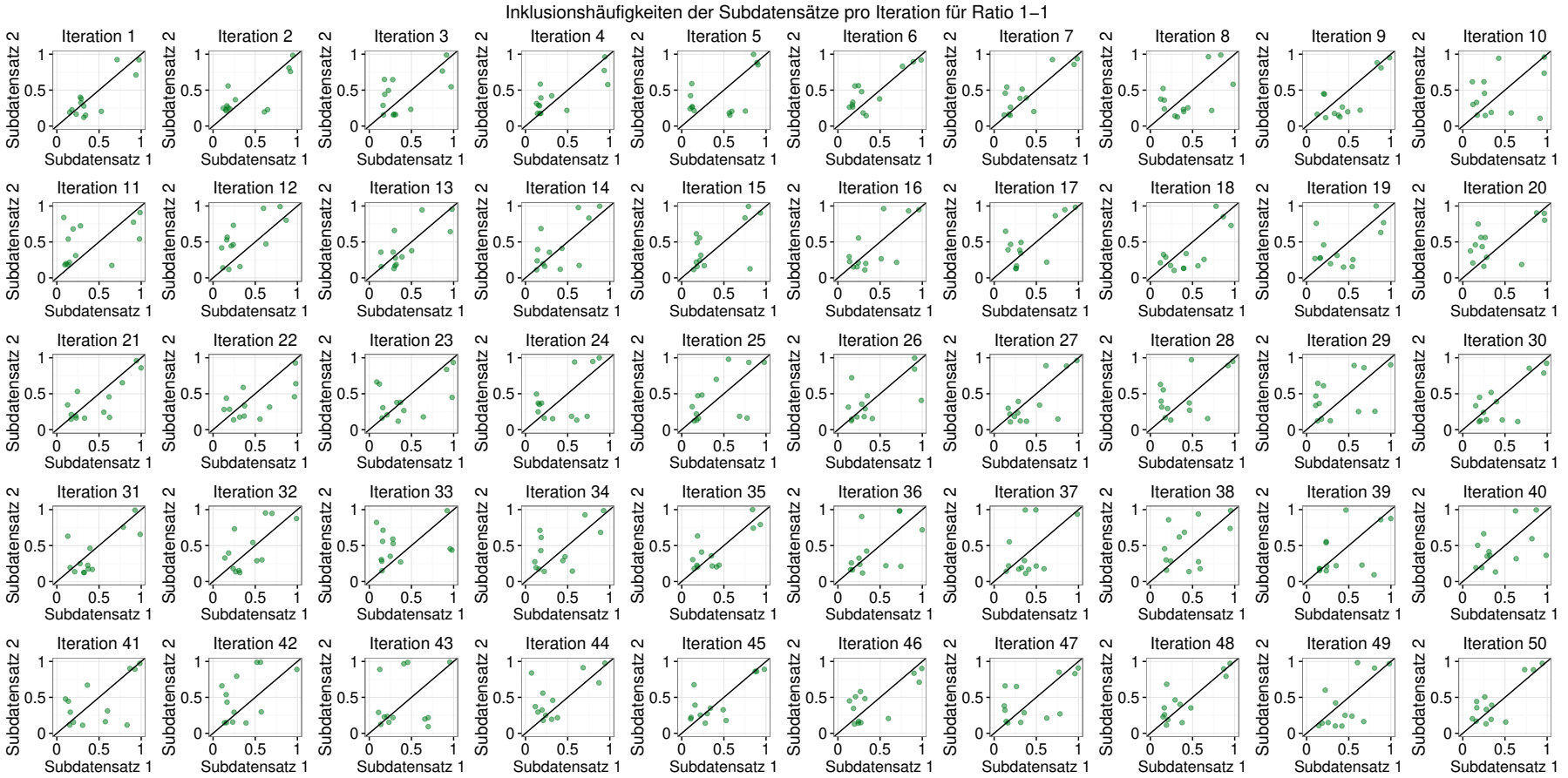

Abbildung 24: Übersicht der Inklusionshäufigkeiten pro Iteration des Datensatzes Glioma für die Ratio 1-1 bei einem Stichprobenumfang der Subdatensätze von jeweils $n/2$. Die Inklusionshäufigkeiten streuen hier weniger stark als in der entsprechenden Übersicht der Ratio 9-1.

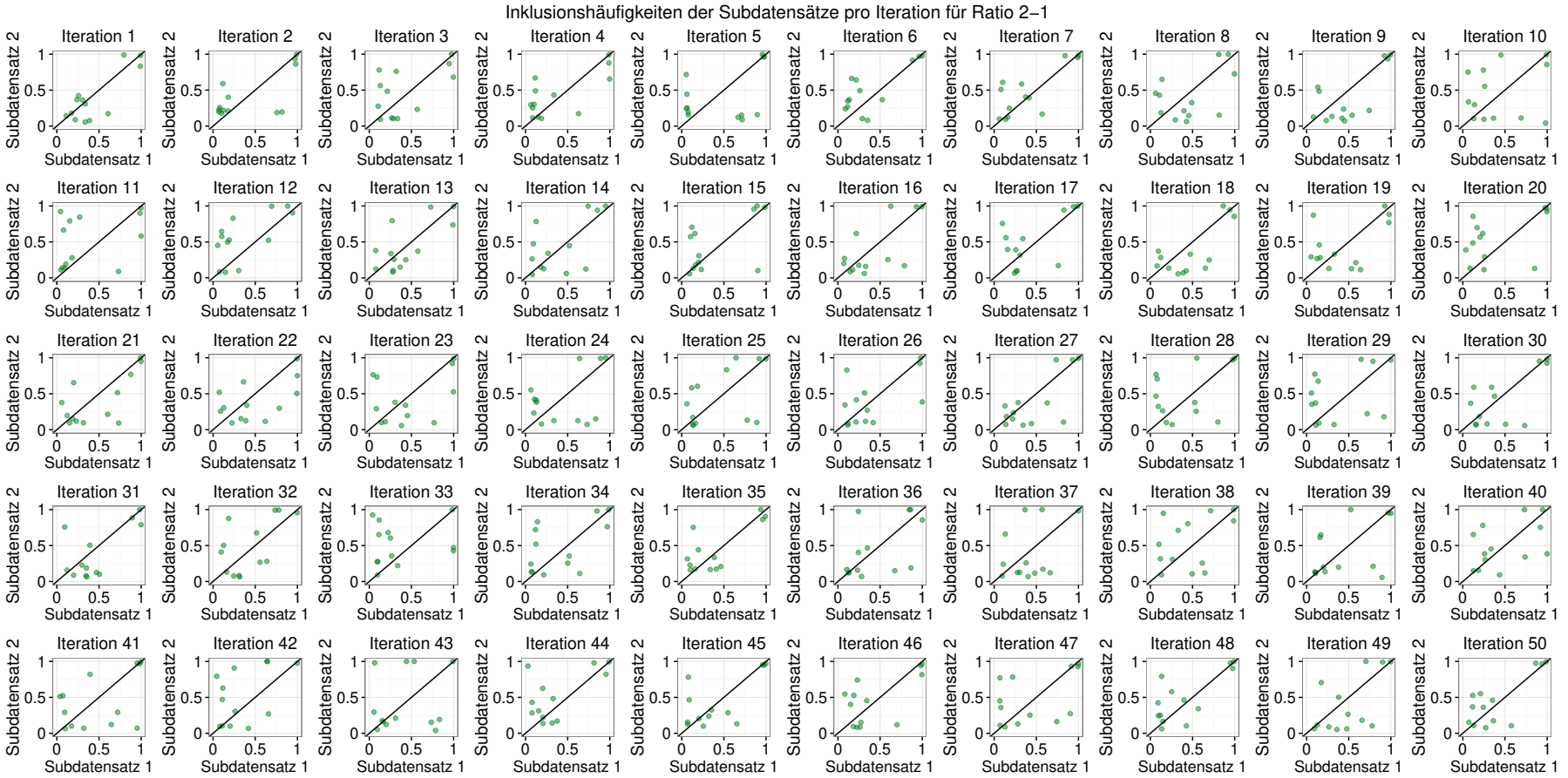

Abbildung 25: Übersicht der Inklusionshäufigkeiten pro Iteration des Datensatzes Glioma für die Ratio 2-1 bei einem Stichprobenumfang der Subdatensätze von jeweils $n/2$. Die Inklusionshäufigkeiten streuen hier stärker als in der entsprechenden Übersicht der Ratio 1-1.

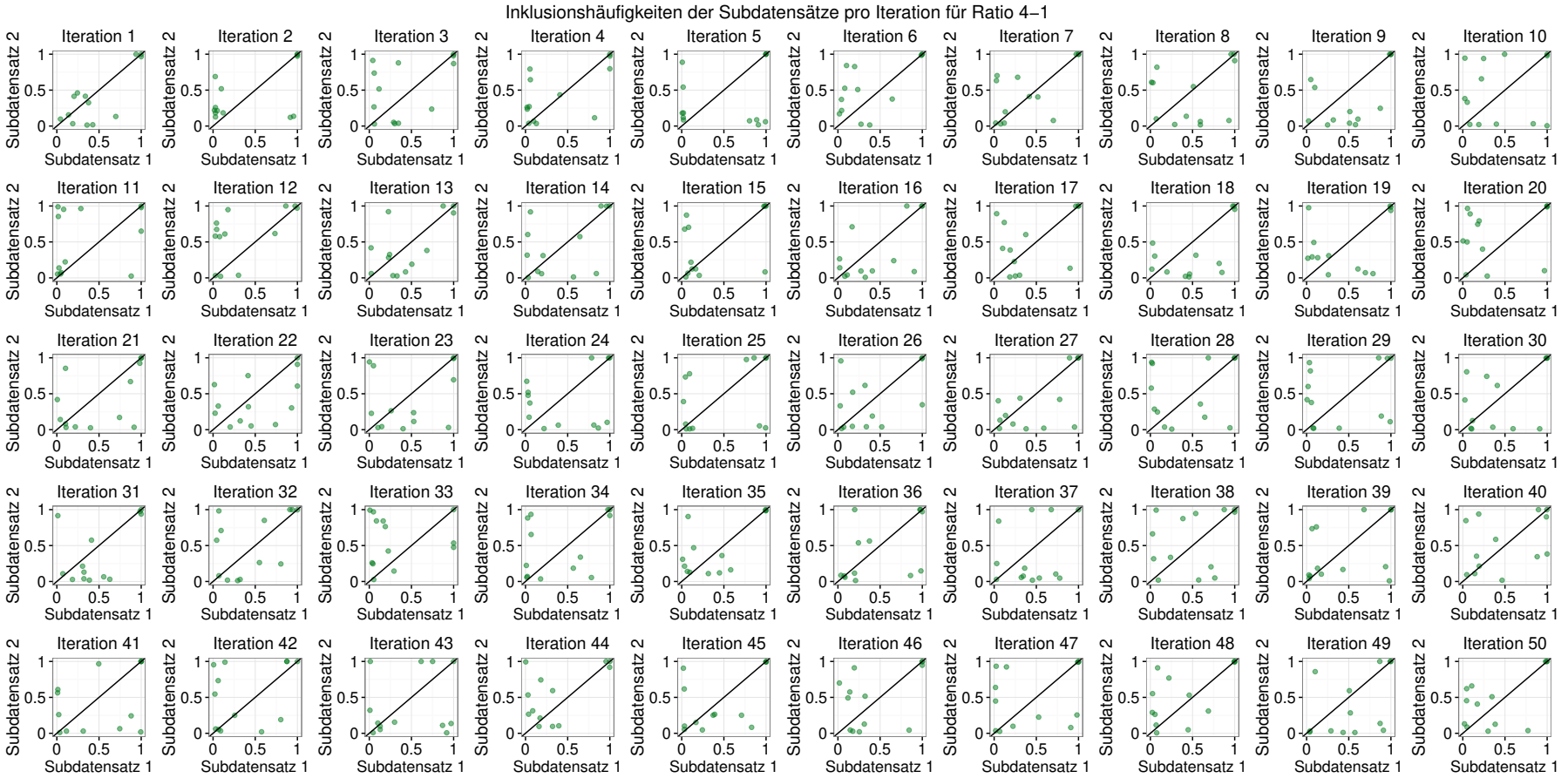

Abbildung 26: Übersicht der Inklusionshäufigkeiten pro Iteration des Datensatzes Glioma für die Ratio 4-1 bei einem Stichprobenumfang der Subdatensätze von jeweils $n/2$. Die Inklusionshäufigkeiten streuen hier weniger stark als in der entsprechenden Übersicht der Ratio 9-1.

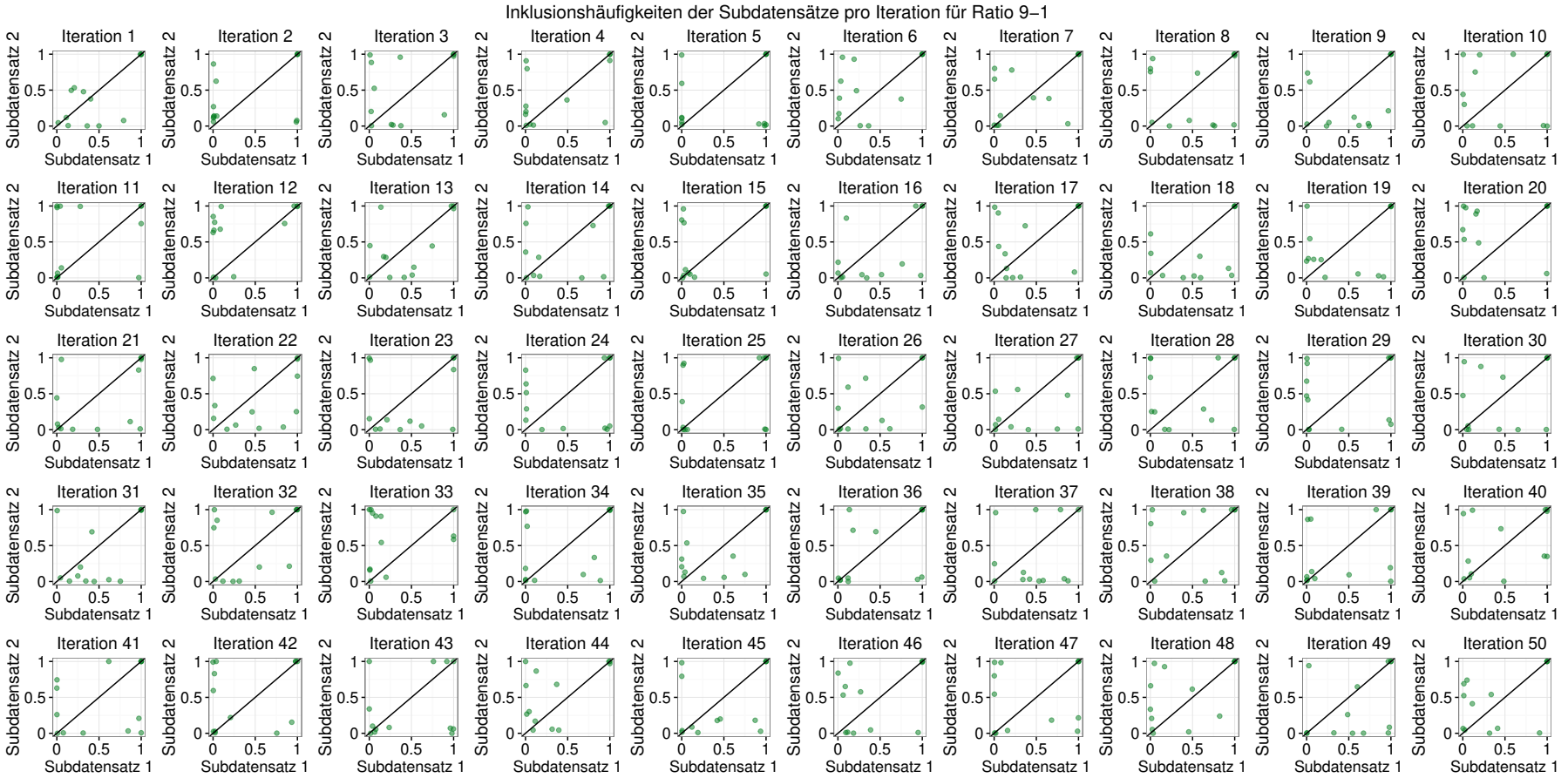

Abbildung 27: Übersicht der Inklusionshäufigkeiten pro Iteration des Datensatzes Glioma für die Ratio 9-1 bei einem Stichprobenumfang der Subdatensätze von jeweils $n/2$. Die Inklusionshäufigkeiten streuen hier stärker als in der entsprechenden Übersicht der Ratio 1-1.

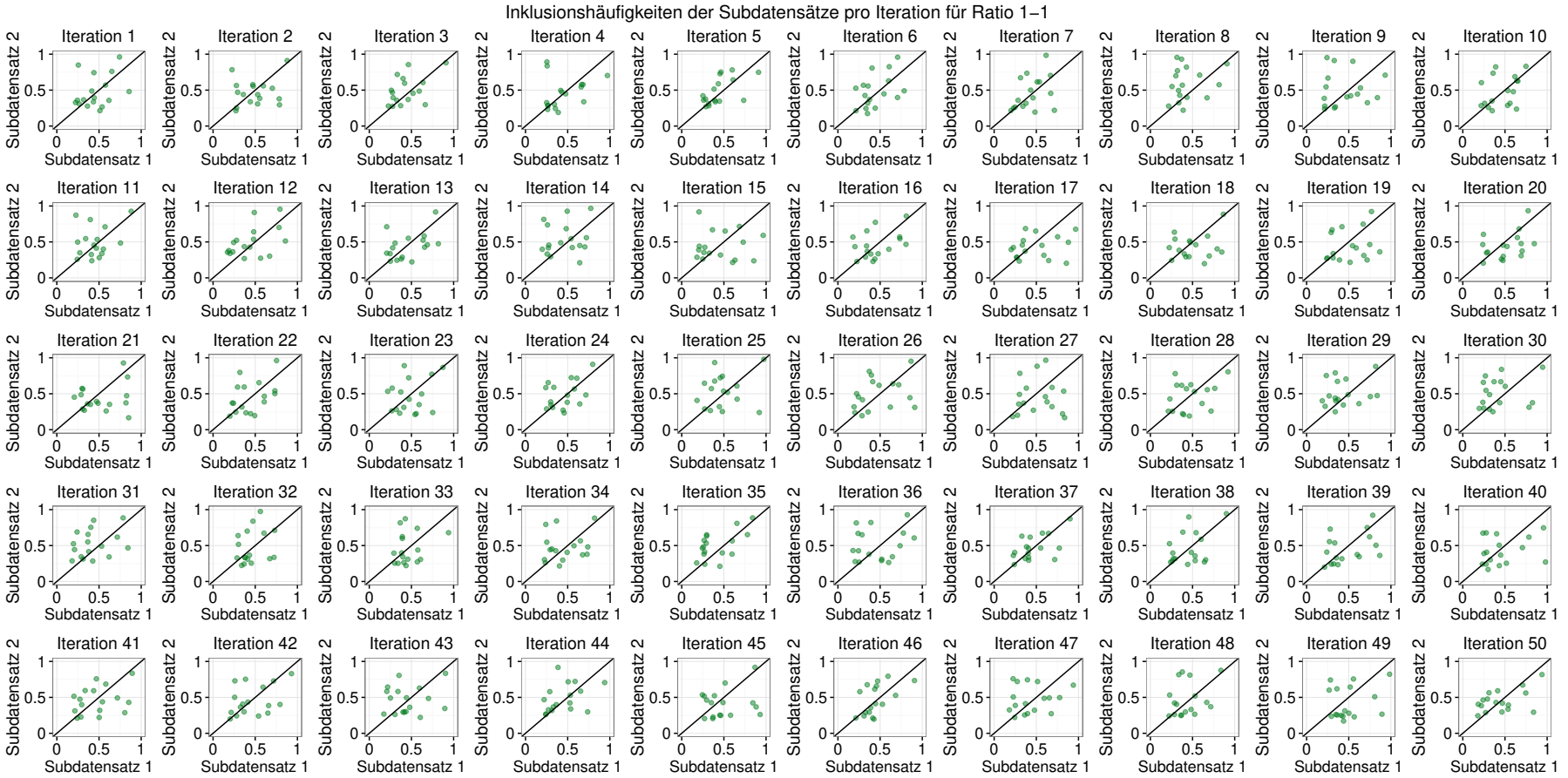

Abbildung 28: Übersicht der Inklusionshäufigkeiten pro Iteration des Datensatzes PBC für die Ratio 1-1 bei einem Stichprobenumfang der Subdatensätze von jeweils $n/2$. Die Inklusionshäufigkeiten streuen hier weniger stark als in der entsprechenden Übersicht der Ratio 9-1.

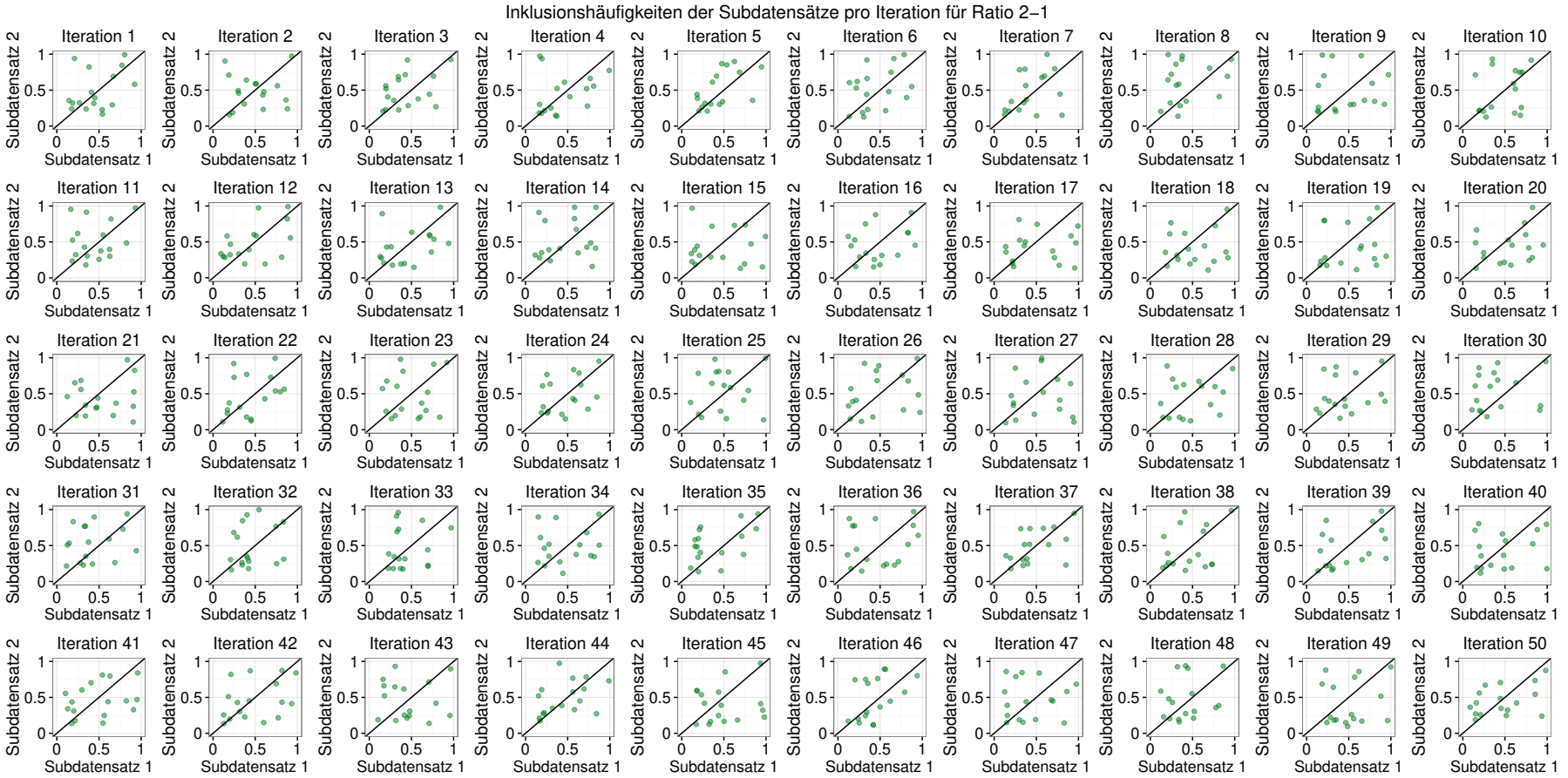

Abbildung 29: Übersicht der Inklusionshäufigkeiten pro Iteration des Datensatzes PBC für die Ratio 2-1 bei einem Stichprobenumfang der Subdatensätze von jeweils $n/2$. Die Inklusionshäufigkeiten streuen hier stärker als in der entsprechenden Übersicht der Ratio 1-1.

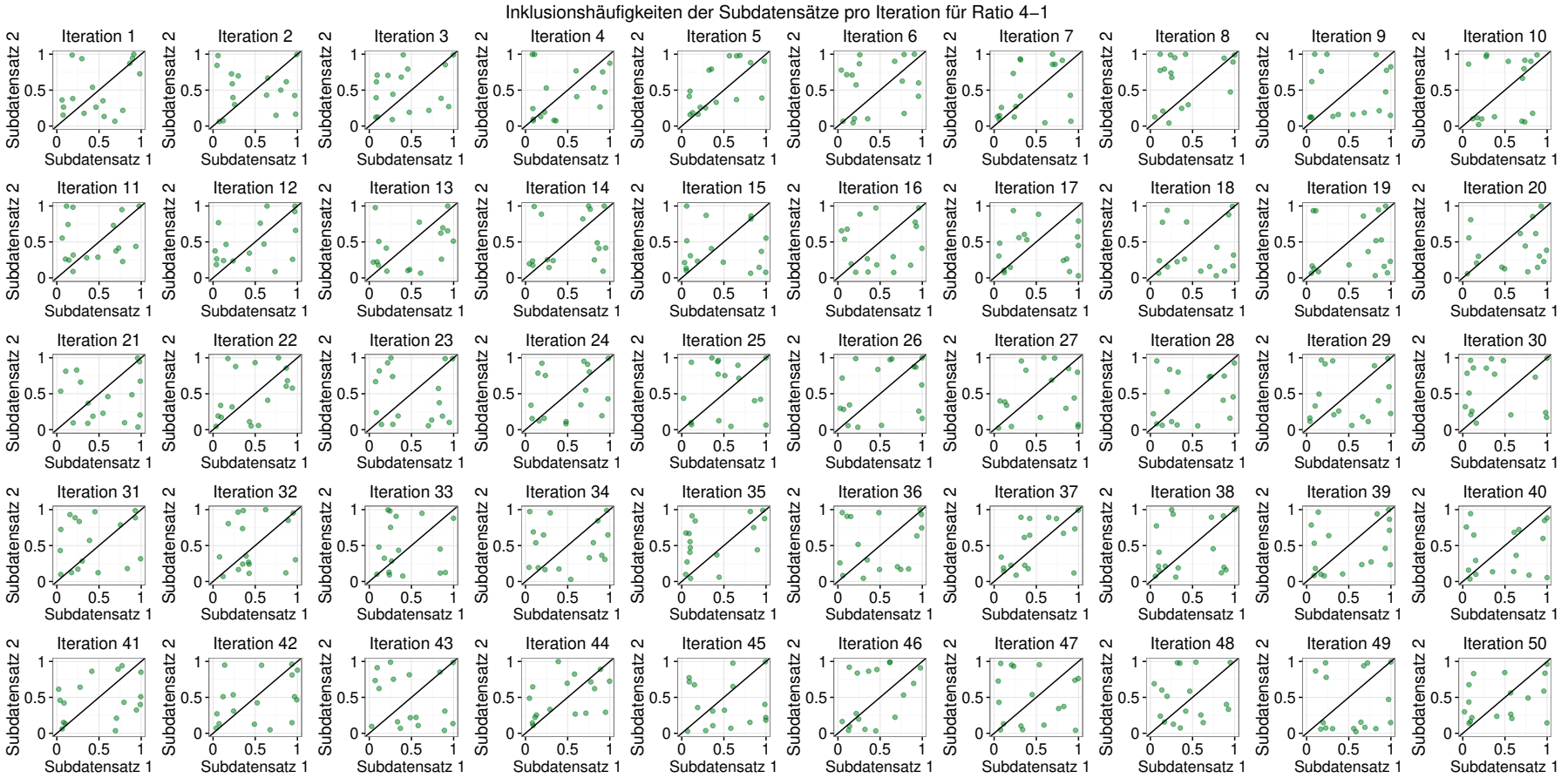

Abbildung 30: Übersicht der Inklusionshäufigkeiten pro Iteration des Datensatzes PBC für die Ratio 4-1 bei einem Stichprobenumfang der Subdatensätze von jeweils $n/2$. Die Inklusionshäufigkeiten streuen hier weniger stark als in der entsprechenden Übersicht der Ratio 9-1.

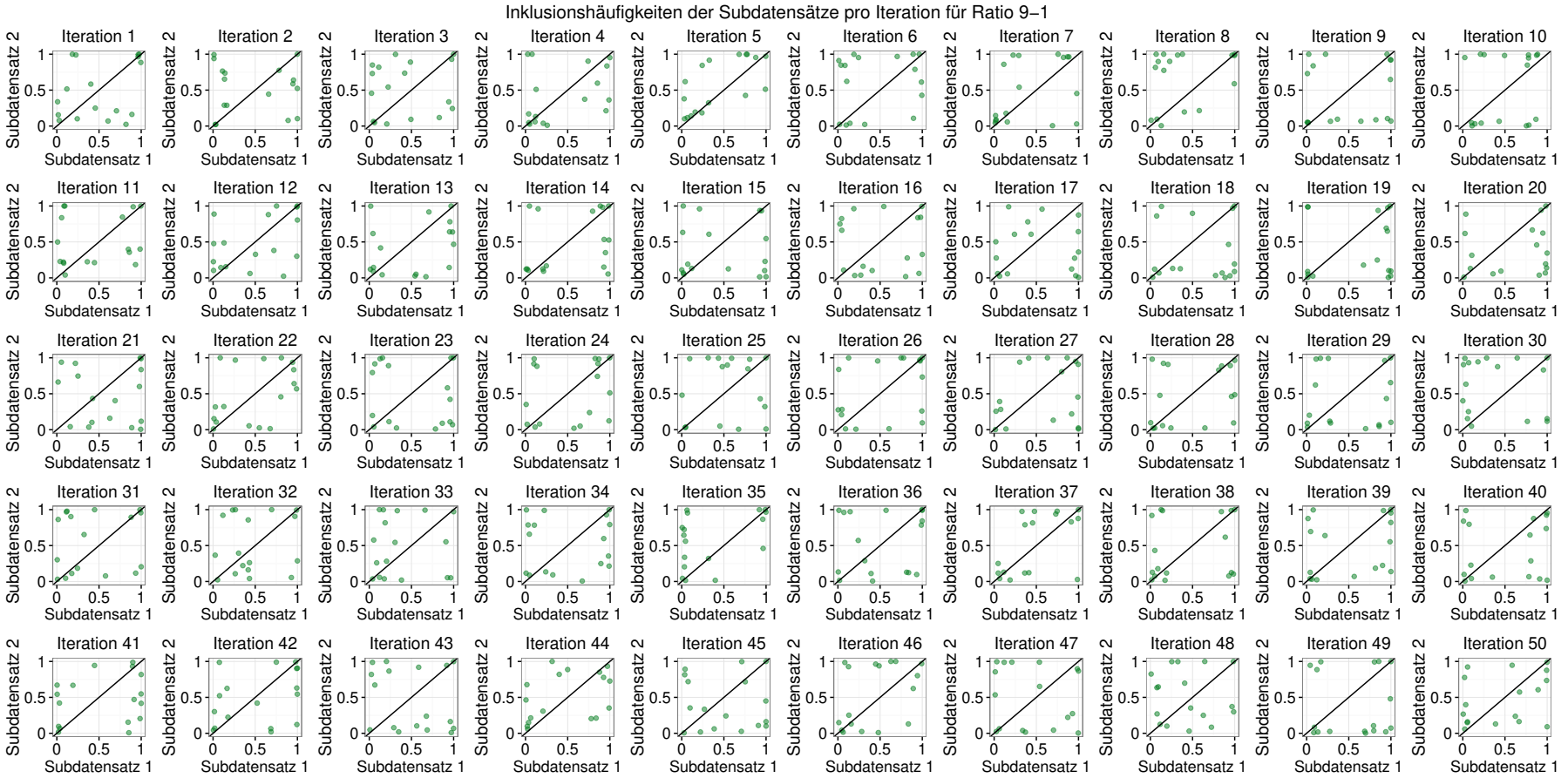

Abbildung 31: Übersicht der Inklusionshäufigkeiten pro Iteration des Datensatzes PBC für die Ratio 9-1 bei einem Stichprobenumfang der Subdatensätze von jeweils $n/2$. Die Inklusionshäufigkeiten streuen hier stärker als in der entsprechenden Übersicht der Ratio 1-1.

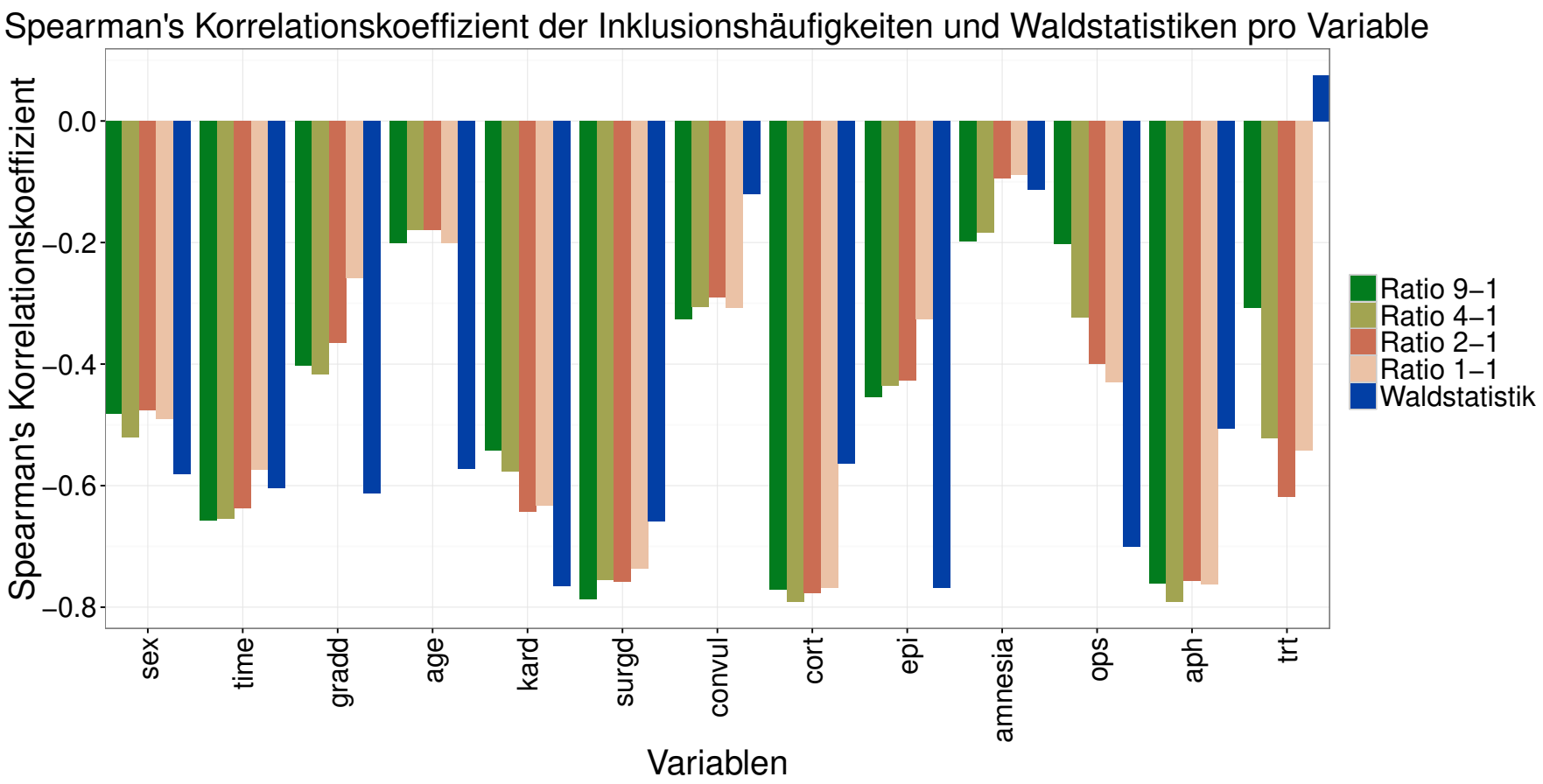

Abbildung 32: Übersicht der Korrelationskoeffizienten nach Spearman aus Sicht der Variablen des Datensatzes Glioma für die Inklusionshäufigkeiten aller Ratios und die Waldstatistiken bei einem Stichprobenumfang der Subdatensätze von jeweils $n/2$. Es existieren viele negative Korrelationskoeffizienten.

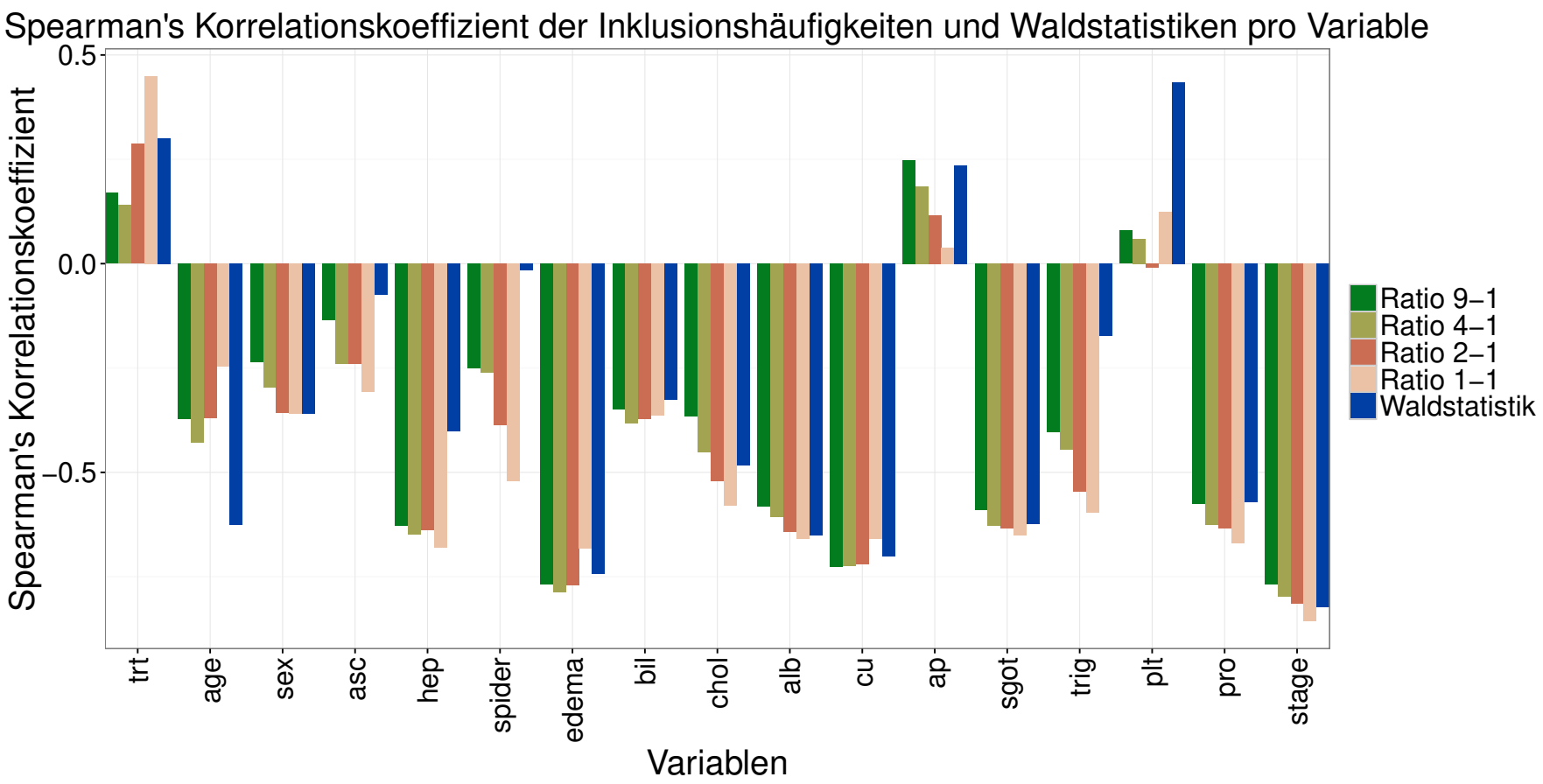

Abbildung 33: Übersicht der Korrelationskoeffizienten nach Spearman aus Sicht der Variablen des Datensatzes PBC für die Inklusionshäufigkeiten aller Ratios und die Waldstatistiken bei einem Stichprobenumfang der Subdatensätze von jeweils $n/2$. Es existieren viele negative Korrelationskoeffizienten.

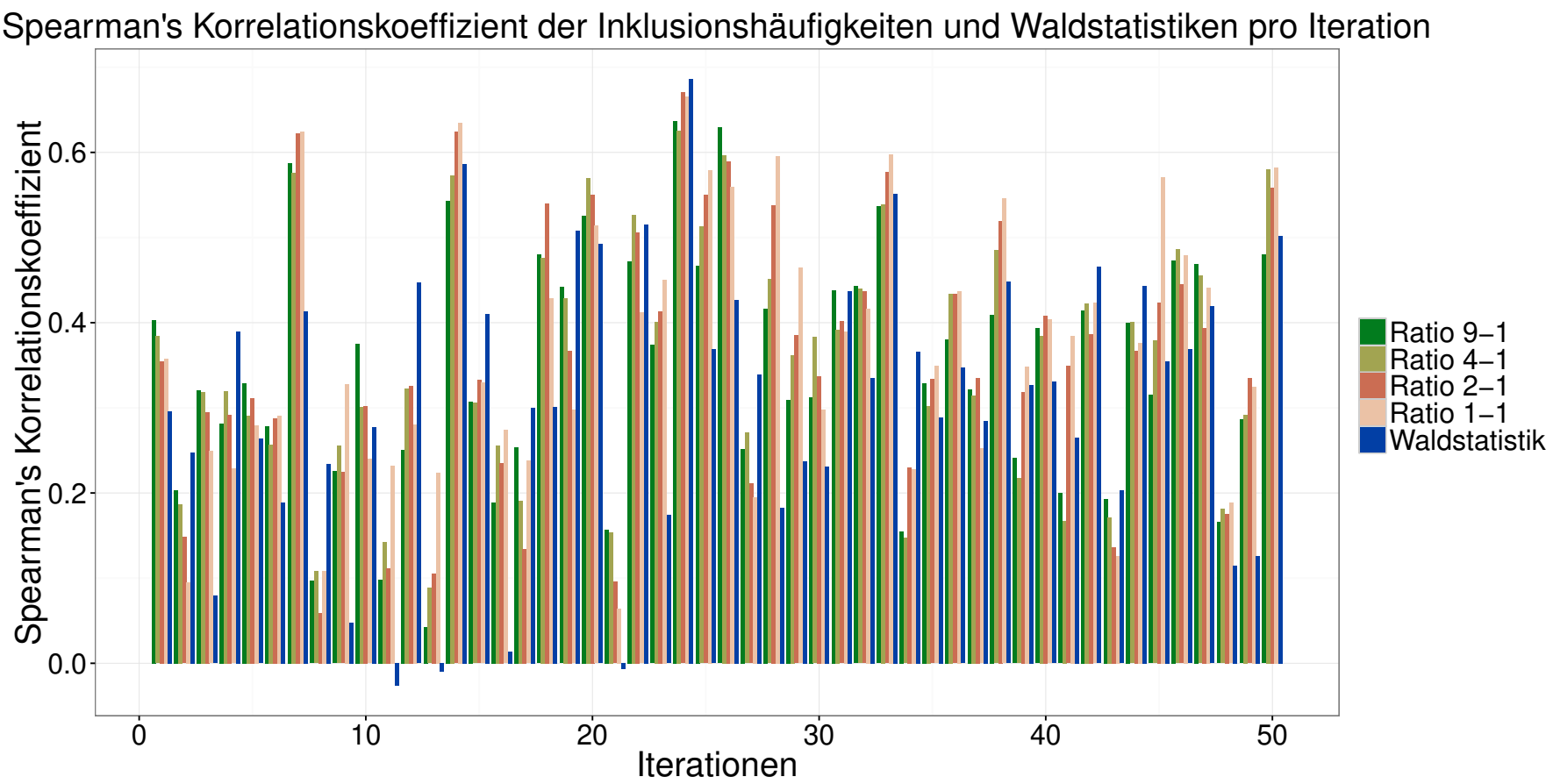

Abbildung 34: Übersicht der Korrelationskoeffizienten nach Spearman aus Sicht der Iterationen des Datensatzes Ozone für die Inklusionshäufigkeiten aller Ratios und die Waldstatistiken bei einem Stichprobenumfang der Subdatensätze von jeweils $n/3$. Ratio 1-1 liefert in den meisten Iterationen den höchsten Korrelationskoeffizienten.

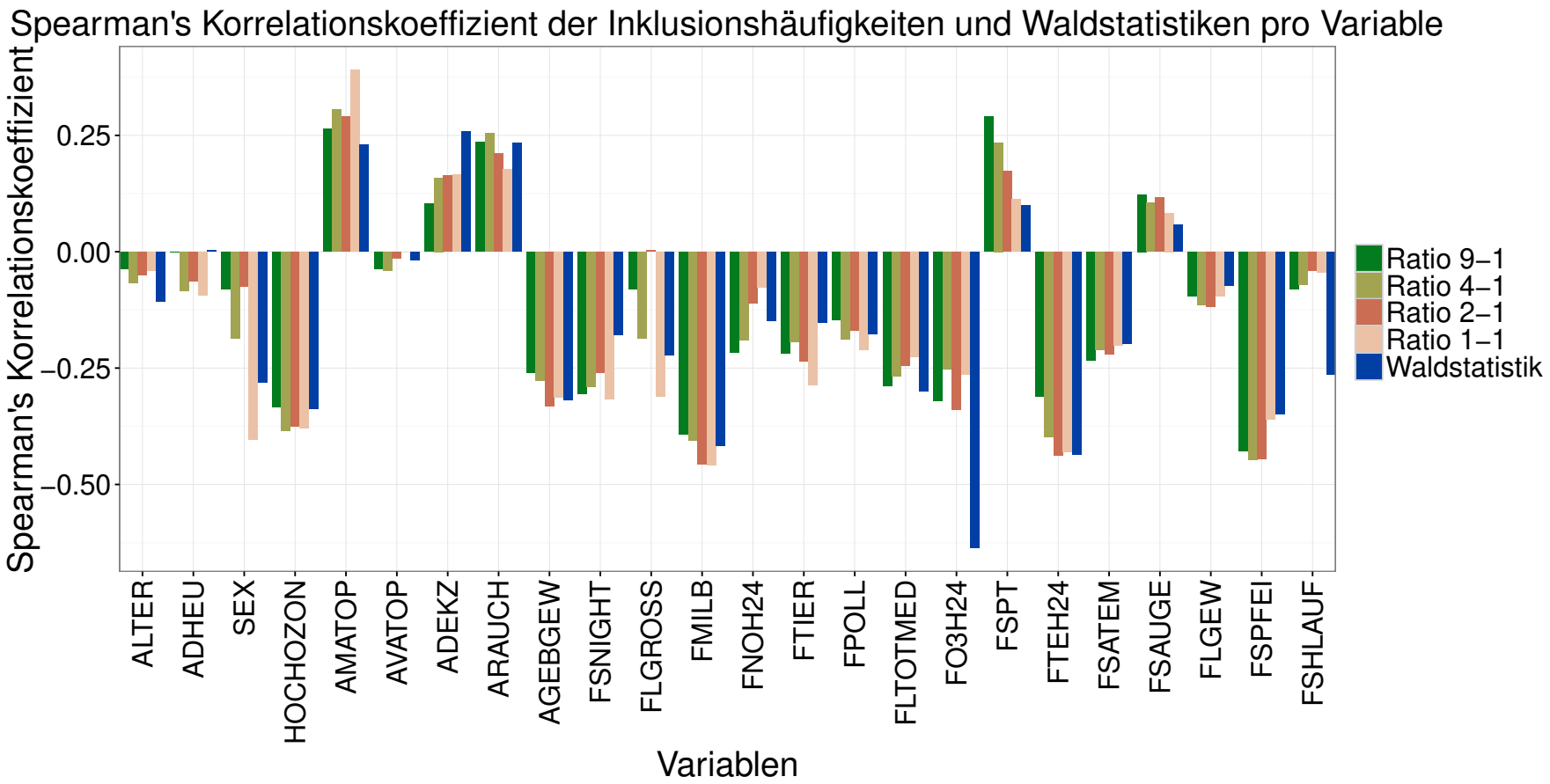

Abbildung 35: Übersicht der Korrelationskoeffizienten nach Spearman aus Sicht der Variablen des Datensatzes Ozone für die Inklusionshäufigkeiten aller Ratios und die Waldstatistiken bei einem Stichprobenumfang der Subdatensätze von jeweils $n/3$. Es existieren viele negative Korrelationskoeffizienten.

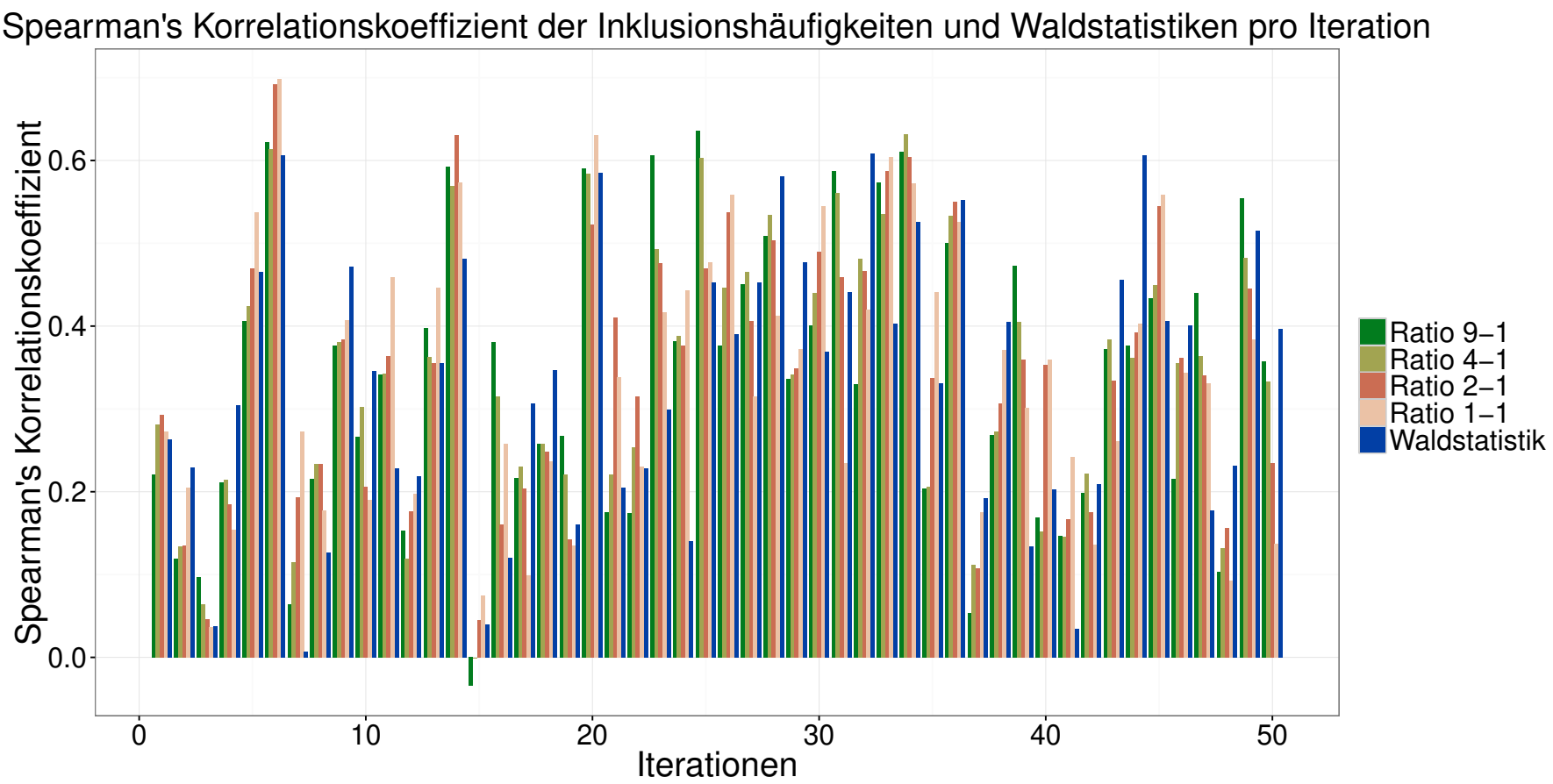

Abbildung 36: Übersicht der Korrelationskoeffizienten nach Spearman aus Sicht der Iterationen des Datensatzes Ozone für die Inklusionshäufigkeiten aller Ratios und die Waldstatistiken bei einem Stichprobenumfang der Subdatensätze von jeweils $n/4$. Die Waldstatistiken liefern in den meisten Iterationen den höchsten Korrelationskoeffizienten.

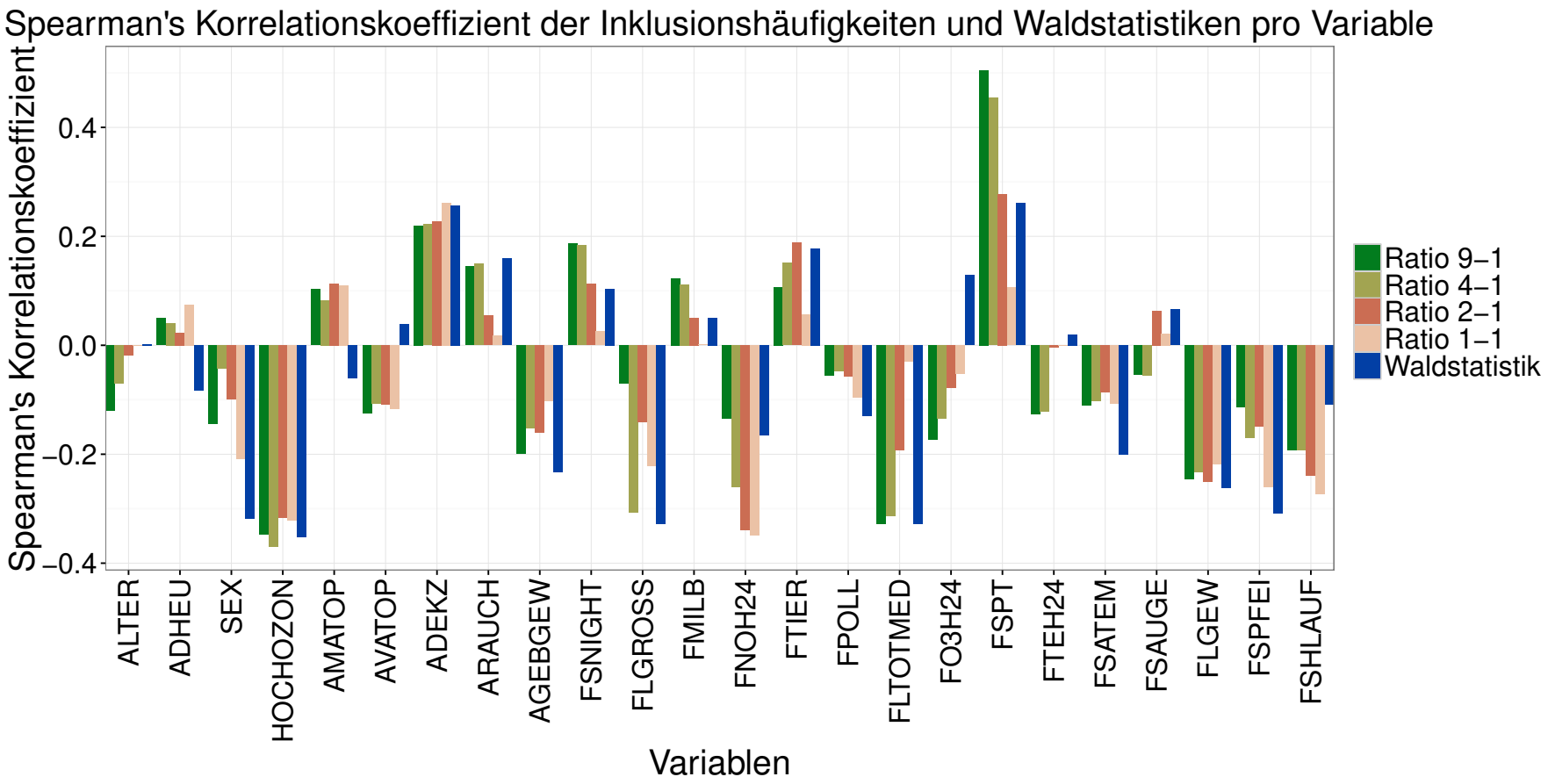

Abbildung 37: Übersicht der Korrelationskoeffizienten nach Spearman aus Sicht der Variablen des Datensatzes Ozone für die Inklusionshäufigkeiten aller Ratios und die Waldstatistiken bei einem Stichprobenumfang der Subdatensätze von jeweils $n/4$. Es existieren viele negative Korrelationskoeffizienten.

Eidesstattliche Erklärung

Ich erkläre, dass ich, Leonie Friederike Litzka, die vorliegende Bachelorarbeit selbstständig ohne fremde Hilfe angefertigt und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Ort, Datum	Unterschrift