LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

MASTERARBEIT

Vergleichsstudie von Random Forest-Varianten für ordinalen Response bezüglich verschiedener Variablenwichtigkeits- und Klassifikationsgütemaße

---

Verfasserin: Jasmin Wulf

Betreuer: Dr. Roman Hornung

Institut für Statistik

Institut für Medizinische Informationsverarbeitung,
Biometrie und Epidemiologie

14. März 2022

Abstract

Gegenstand dieser Arbeit ist eine Vergleichsstudie verschiedener Random Forest-Varianten für ordinalen Response hinsichtlich der Qualität der Vorhersagen und Einstufung der Prädiktoren nach ihrer Wichtigkeit für die Vorhersagen. Zu diesem Zweck werden unterschiedliche Ordinal Forest-Varianten (Hornung, 2020), darunter eine kürzlich entwickelte Version basierend auf dem Ranked Probability Score, und Random Forests aus Conditional Inference-Bäumen (Hothorn et al., 2006b) anhand von simulierten und realen ordinalen Daten gegenübergestellt.

Die ausgewählten Vorhersagemethoden behandeln den ordinalen Response als stetigen Response, indem optimierte Score-Werte (bei Ordinal Forests) oder beliebige Score-Werte (bei Random Forests aus Conditional Inference-Bäumen) anstelle der Klassenwerte des ordinalen Response verwendet werden.

In Bezug auf die Klassifikationsgüte zeigt die Vergleichsstudie, dass die bestehende Ordinal Forest-Variante tendenziell besser abschneidet als die Konkurrenz. Darüber hinaus legen die Ergebnisse in Bezug auf die Qualität der Variablenwichtigkeit nahe, dass der Random Forest aus Conditional Inference-Bäumen häufig bessere Leistungen erzielt als die konkurrierenden Methoden.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1. Einleitung

Statistische Methoden mit dem Ziel, einen Response vorherzusagen oder relevante Prädiktoren zu identifizieren, sollten mit einem ordinalen Response adäquat umgehen können, sodass die inhärente Ordnung berücksichtigt wird. Im biomedizinischen Bereich handelt es sich beispielsweise bei den Tumorstadien I-IV oder bei der Schwere einer Erkrankung (leicht, mittelschwer, schwer) um einen ordinalen Response mit geordneten Klassen.

Einige existierende Vorhersagemethoden ignorieren die Ordnung und behandeln den ordinalen Response als nominalskaliert (Janitza et al., 2016). Dieses Vorgehen kann zu einem Informationsverlust führen und in weniger zuverlässigen Vorhersagen resultieren. Dies bestätigt eine Untersuchung von Janitza et al. (2016), die sich auf den Vergleich zwischen ordinalen und nominalen Responses von Random Forests aus Conditional Inference-Bäumen fokussiert. Dabei wird der Frage nachgegangen, ob die Berücksichtigung der Ordnung zu einer verbesserten Klassifikationsgüte und Variablenselektion führt. Die Ergebnisse für simulierte und reale Daten zeigen, dass die Forests mit ordinalem und nominalem Response sehr ähnlich in ihrer Klassifikationsgüte sind, wenngleich mit ordinalem Response überwiegend minimal bessere Vorhersagen erzielt werden können. Dies lässt sich mit simulierten Daten insbesondere bei einer höheren Anzahl an Klassen (von 6 und 9 Klassen versus 3 Klassen) beobachten. Die Gegenüberstellung der beiden Forests zeigt außerdem, dass die Einbeziehung der Ordnung für die Variablenselektion von Vorteil ist.
Motiviert durch die Ergebnisse dieser Studie wird in der vorliegenden Arbeit die etablierte Methode der Random Forests aus Conditional Inference-Bäumen (im Folgenden als Conditional Inference Forests bezeichnet) für ordinalen Response hinsichtlich Variablenwichtigkeit und Klassifikationsgüte wiederholt in den Mittelpunkt der Betrachtung gerückt. In Konkurrenz mit dieser Methode stehen Ordinal Forests, die kürzlich von Hornung (2020) für den Umgang von ordinalen Responses eingeführt wurden. Ordinal Forests verfolgen das Konzept eines latenten stetigen Response, der die Klassen des beobachteten ordinalen Response in Form von optimierten Score-Werten einbezieht.

Beide vorgestellten Random Forest-Methoden haben gemeinsam, dass sie mit einem ordinalskalierten Response als stetigen Response verfahren, der die Ordnung implizit berücksichtigt. Zusätzlich können die Methoden einflussreiche Prädiktoren erkennen und nach ihrer Wichtigkeit für die Vorhersage ordnen.

Eine bestehende Variante des Ordinal Forest wird in dem Beitrag von Hornung (2020) mit konkurrierenden Random Forest-Methoden hinsichtlich der Leistungen bei der Vorhersage und Variablenselektion für simulierte und reale ordinale Daten verglichen. Auf Grundlage der stärkeren Leistungsfähigkeit, die von der Ordinal Forest-Variante ausgeht, schlägt Hornung (2021) eine weitere Variante vor, die in Ergänzung zu Punktvorhersagen auch Wahrscheinlichkeitsvorhersagen für Klassen ermöglicht.

In dieser Arbeit wird anhand von simulierten und realen Daten eine umfangreiche Benchmarkstudie zwischen den kürzlich entwickelten Ordinal Forests und den etablierten Conditional Inference Forests für einen ordinalen Response durchgeführt. Die verschiedenen Ansätze werden in Bezug auf zwei wesentliche Aspekte miteinander verglichen: die Qualität der Variablenwichtigkeit und die Klassifikationsgüte. Dieses Vorgehen ist durch die Arbeiten von Janitza et al. (2016) und Hornung (2020) motiviert. So basiert das hier verwendete Maß für die Bewertung der Variablenwichtigkeit auf dem gleichen Maß, das in der Arbeit von Janitza et al. (2016) eingesetzt wird. Ferner ist die Vorgehensweise für die Beurteilung der Klassifikationsgüte an den Beitrag von Hornung (2020) angelehnt.

Die Arbeit orientiert sich an dem folgenden Aufbau. Zunächst wird in Kapitel 2 eine theoretische Einführung in Klassifikations- und Regressionsbäume gegeben. Zusätzlich umfasst das entsprechende Kapitel den Aufbau von Conditional Inference-Bäumen.
In Kapitel 3 werden die Grundlagen für die Erstellung von Random Forests beschrieben. Daneben erfolgt in diesem Kapitel die Darstellung der Prinzipien der von Hornung (2020) entwickelten Methode der Ordinal Forests. Nachdem der zugrundeliegende Algorithmus dargelegt wurde, wird der für Ordinal Forests zentrale Parameter, nämlich die Performance-Funktion, vorgestellt. Außerdem beinhaltet das Kapitel die Berechnung der Variablenwichtigkeit für Random Forests. Schließlich lassen sich aus dem Zusammenspiel der beiden Parameter Performance-Funktion und Variablenwichtigkeit unterschiedliche Varianten eines Ordinal Forest generieren.

Das darauffolgende Kapitel 4 bildet den eigentlichen Kern dieser Arbeit, nämlich die Durchführung einer Benchmarkstudie mit den verschiedenen Random Forest-Varianten hinsichtlich Variablenwichtigkeit und Klassifikation für simulierte und reale ordinale Daten. Nach Beschreibung der Anwendung von den Random Forest-Varianten in dem Statistik-Programm `R` (R Core Team, 2021) werden die drei Simulationsdesigns, die für die Benchmarkstudie verwendet werden, vorgestellt. Die Sammlung an verschiedenen Simulationsdesigns umfasst dasjenige von Janitza et al. (2016), Hornung (2020) sowie Buri und Hothorn (2020). Anschließend werden die mit den simulierten Daten erzielten Ergebnisse präsentiert. Des Weiteren erfolgt die Vorstellung der fünf ausgewählten realen Datensätze, die ebenso einen Teil der Datengrundlage für die Benchmarkstudie ausmachen. Das Kapitel schließt mit dem Ergebnisbericht für die Analyse der realen Datensätze.
Zuletzt werden in Kapitel 5 die wichtigsten Ergebnisse zusammengefasst und ein Ausblick gegeben.

2. Einführung in Klassifikations-, Regressions- und Conditional Inference-Bäume

2.1. Klassifikations- und Regressionsbäume

Klassifikations- und Regressionsbäume (CART) wurden von Breiman et al. (1984) eingeführt und dienen der Vorhersage der Werte eines Response für neue Beobachtungen basierend auf deren Prädiktorenwerten. Für den Fall eines stetigen Response resultieren Regressionsbäume und im Falle eines nominalen Response ergeben sich Klassifikationsbäume. Der Konstruktion von Regressions- und Klassifikationsbäumen liegt die Idee der rekursiven binären Partitionierung zugrunde. Dabei wird ein Datensatz kontinuierlich aufgeteilt, sodass eine disjunkte Zerlegung des Messraumes $\mathcal{X}$ resultiert, die als invertierte Baumstruktur dargestellt werden kann. Der Messraum $\mathcal{X}$ ist die Menge aller möglichen Vektoren $\boldsymbol{x}$ mit den Werten für die Prädiktoren. Schließlich entspricht die Zerlegung von $\mathcal{X}$ in disjunkte Teilmengen – sodass jedem Vektor $\boldsymbol{x}$ in $\mathcal{X}$ eine Klasse oder ein Wert zugewiesen wird – einer Klassifizierungs- oder Vorhersageregel.
Die nachstehende Auflistung dient als Überblick über die einheitliche Notation in diesem Kapitel.

- p: Anzahl an Prädiktoren,
- $\boldsymbol{x} = (x_1, ..., x_p)^T$: Vektor mit den Werten der Prädiktoren für eine beliebige Beobachtung,
- $\boldsymbol{x}_i$: Vektor mit den Werten der Prädiktoren für die i-te Beobachtung, $i = 1, ..., n$,
- J: Anzahl an Klassen, $j = 1, ..., J$,
- C: Menge an möglichen Klassen,
- y_i: Wert des Response für die i-te Beobachtung, $y_i \in \{1, ..., J\}$.

Die nachfolgenden Erläuterungen beziehen sich, falls nicht anders angegeben, auf die Literatur von Breiman et al. (1984).

2.1.1. Erstellung von Klassifikationsbäumen

Ein Klassifikationsbaum hat die Konstruktion einer Klassifizierungsregel zum Ziel, anhand dieser einer neuen Beobachtung eine Klassenzugehörigkeit auf Grundlage ihrer Prädiktorenwerte zugewiesen werden kann.
Der Klassifikationsbaum startet mit dem Wurzelknoten, in dem alle Beobachtungen eines Trainingsdatensatzes, also einer Stichprobe des Originaldatensatzes, enthalten sind. Für den Trainingsdatensatz ist die korrekte Klassenzugehörigkeit der Beobachtungen bekannt. Ausgehend von diesem obersten Knoten wird für jede Aufteilung des Messraumes $\mathcal{X}$ aus allen p Prädiktoren der Prädiktor, und innerhalb dieses Prädiktoren der Grenzwert, ausgewählt, der zu der besten Aufteilung im Sinne von möglichst homogenen Teilmengen bezüglich des Response führt. Die Homogenität der Beobachtungen in einem Knoten wird durch die minimale Unreinheit, z.B. mittels des Gini-Index (Breiman et al., 1984) oder der Entropie (Sutton, 2005), gemessen.
Der oberste Knoten wird anhand einer einfachen Ja/Nein-Frage an einem gewissen (Grenz-)Wert aus dem Wertebereich des gewählten Prädiktoren in zwei Teilmengen, den hier benannten Entscheidungsknoten, aufgeteilt. Beispielsweise wird in Abbildung 2.1 der Messraum $\mathcal{X}$ mit allen möglichen Werten von $(x_1, x_2, x_3)^T$ zunächst in $\{\boldsymbol{x} \mid x_3 \leq 60.5\}$ und $\{\boldsymbol{x} \mid x_3 > 60.5\}$ zerlegt. Für die Aufteilung wird der numerische Prädiktor x_3 ausgewählt. Ein Vektor $\boldsymbol{x}^*$ einer neuen Beobachtung würde in die linke Teilmenge gelangen für den Fall, dass $x_3 \leq 60.5$, und in die rechte Teilmenge für den Fall, dass $x_3 > 60.5$ gilt.
Falls in einem oder beiden Entscheidungsknoten eine noch größere Homogenität erreicht werden kann, werden die entsprechenden Knoten ein weiteres Mal in zwei – noch homogenere – Entscheidungsknoten aufgeteilt. Beispielsweise kann in der Abbildung 2.1 die resultierende Menge $\{\boldsymbol{x} \mid x_3 \leq 60.5\}$ aus der oben durchgeführten Aufteilung weiter in die Knoten $A_1 = \{\boldsymbol{x} \mid x_3 \leq 60.5, x_1 \leq 25.5\}$ und $A_2 = \{\boldsymbol{x} \mid x_3 \leq 60.5, x_1 > 25.5\}$ aufgespalten werden, wobei für diese Aufspaltung der numerische Prädiktor x_1 verwendet wird. Gleichzeitig kann die andere Menge $\{\boldsymbol{x} \mid x_3 > 60.5\}$ aus der ersten Zerlegung durch den Prädiktor x_1 weiter in die Knoten $A_3 = \{\boldsymbol{x} \mid x_3 > 60.5, x_1 \leq 70.5\}$ und $A_4 = \{\boldsymbol{x} \mid x_3 > 60.5, x_1 > 70.5\}$ unterteilt werden. Es wäre ebenso möglich, die Mengen $\{\boldsymbol{x} \mid x_3 \leq 60.5\}$ und $\{\boldsymbol{x} \mid x_3 > 60.5\}$ durch unterschiedliche Prädiktoren anstatt durch denselben Prädiktor (hier: x_1) zu teilen (Sutton, 2005).
Dieses Verfahren wird sukzessive durchgeführt, bis die Knoten maximale Homogenität bezüglich des Response erreicht haben und nicht mehr weiter aufgespalten werden können. Diese Knoten werden schließlich als Endknoten des Baumes bezeichnet, denen eine bestimmte Klasse zugewiesen werden kann. Jedoch ist zu beachten,

dass mit einer hohen Komplexität des Baumes, d.h. mit einer hohen Anzahl an Aufteilungen, die Gefahr der Überanpassung an die Daten, sogenanntes Overfitting, einhergeht. Falls in den Endknoten nur noch eine einzige Klasse vertreten ist – die Knoten vollkommen homogen sind – kann eine zu starke Anpassung an die Daten bestehen, womit der Baum für die Vorhersage der Klassenzugehörigkeit neuer Beobachtungen nicht mehr geeignet wäre. Das Problem von Overfitting kann durch sogenanntes Pruning verhindert werden. Beim Pruning wird der Baum mit maximaler Anzahl an Aufteilungen solange gekürzt, bis keine Überanpassung mehr besteht.

In dem Beispiel aus Abbildung 2.1 können die Knoten A_1 bis A_4 als Endknoten definiert werden, insofern diese nicht noch einmal zerlegt werden können. Es ergibt sich ein symmetrischer invertierter Baum. Allerdings sind auch Bäume mit asymmetrischer Struktur möglich, wenn nur einer von zwei Entscheidungsknoten weiter zerlegt würde (Sutton, 2005).
Falls der Klassifikationsbaum einen binären Response verwendet, besteht die Menge an möglichen Klassen $\mathcal{C} = \{1, ..., J\}$, zu denen die Beobachtungen zugeordnet werden können, aus zwei Klassen, d.h. $\mathcal{C} = \{1, 2\}$. Einem Endknoten wird diejenige Klasse zugewiesen, die in diesem Knoten am häufigsten auftritt (sogenannte Mehrheitsentscheidung). Neue Beobachtungen, deren Vektor $\boldsymbol{x}^*$ zu den im obigen Beispiel deklarierten Endknoten A_1 oder A_3 gehört, werden sodann als Klasse 1 klassifiziert, während Beobachtungen mit $\boldsymbol{x}^*$ zu den Endknoten A_2 oder A_4 gehörend, als Klasse 2 klassifiziert werden.

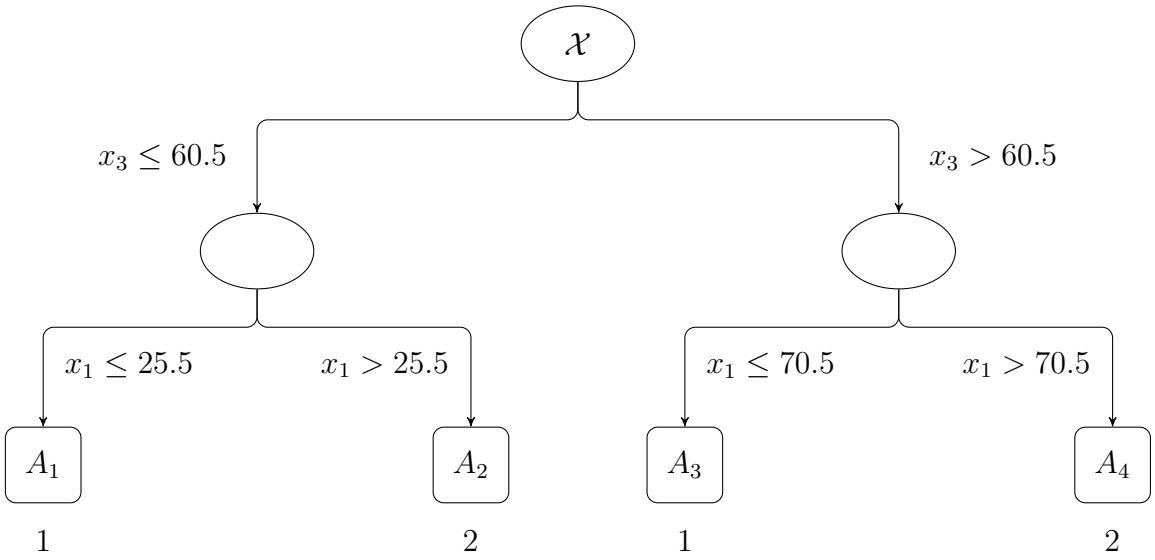

Abb. 2.1.: Darstellung einer Partitionierung von $\mathcal{X}$ als Klassifikationsbaum.

Die Art der Aufteilung eines Knotens ist abhängig von dem Skalenniveau des Prädiktoren, der die größtmögliche Homogenität in den beiden resultierenden Entscheidungsknoten gewährleistet und deshalb für die Aufteilung ausgewählt wird. Bei einem numerischen Prädiktor mit m eindeutigen Werten, wie in obigem Beispiel, gibt es $m-1$ mögliche binäre Zerlegungen. Eine Zerlegung ist definiert durch Werte kleiner oder gleich (erster Entscheidungsknoten) bzw. größer (zweiter Entscheidungsknoten) als ein bestimmter Grenzwert aus dem Wertebereich des ausgewählten Prädiktoren. Mit einem ordinalen Prädiktor wird auf gleiche Weise verfahren wie mit einem numerischen Prädiktor. Im Falle eines kategorialen Prädiktoren mit J Klassen können $2^{J-1}-1$ Zerlegungen durchgeführt werden.

2.1.2. Erstellung von Regressionsbäumen

Ein Regressionsbaum kann aus Daten mit einem stetigen Response resultieren und hat die Konstruktion einer Vorhersageregel zum Ziel, die prognostiziert, welcher numerische Wert einer neuen Beobachtung zugewiesen werden soll (Breiman et al., 1984). Der numerische Wert entspricht dem vorhergesagten Response-Wert für eine neue Beobachtung. Die Erstellung eines Regressionsbaumes unterscheidet sich primär in den folgenden zwei Aspekten von der Erstellung eines Klassifikationsbaumes aus Abschnitt 2.1.1, nämlich in der Evaluation der Homogenität in einem Knoten und in dem Prognoseverfahren.
Beim CART-Ansatz wird für jede Aufteilung von $\mathcal{X}$ aus allen p Prädiktoren derjenige Prädiktor ausgewählt, der die beste Aufteilung im Sinne von möglichst homogenen Teilmengen hinsichtlich des Response gewährleistet. Im Gegensatz zu einem Klassifikationsbaum wird bei einem Regressionsbaum maximale Homogenität in einem Knoten durch maximale Reduktion der Summe der quadrierten Abweichungen der einzelnen Response-Werte in dem Knoten von ihrem Mittelwert gemessen. Zur Erfüllung dieses Kriteriums muss für die Aufteilung einer Teilmenge des Messraumes $\mathcal{X}$ in die Knoten A_1 und A_2 die Summe

$$\sum_{i:\boldsymbol{x}_i \in A_1} (y_i - \overline{y}_{A_1})^2 + \sum_{i:\boldsymbol{x}_i \in A_2} (y_i - \overline{y}_{A_2})^2$$

minimiert werden. Dabei bezeichnet $\overline{y}_{A_1}$ den Mittelwert aus den Response-Werten der Beobachtungen in dem Knoten A_1, während $\overline{y}_{A_2}$ der Mittelwert aus den Response-Werten der Beobachtungen in dem Knoten A_2 ist.

Die Art der Prognose, die sich zwischen den beiden Typen von Bäumen unterscheidet, besteht bei einem Regressionsbaum in der Zuordnung eines reellen Vorhersa-

gewertes zu jedem Endknoten und bei einem Klassifikationsbaum in der Zuweisung einer prognostizierten Klasse zu jedem Endknoten. Der zu einem Endknoten eines Regressionsbaumes zugeordnete numerische Wert entspricht dem Mittelwert aus den Response-Werten der Beobachtungen in diesem Knoten.

Für eine detailliertere Anweisung zu dem Umgang der drei zentralen Elemente bei der Konstruktion eines CART, die von Breiman et al. (1984) ausgearbeitet wurden, nämlich

- die Auswahl der Prädiktoren für die Erstellung der Aufteilungen,
- die Entscheidung, wann ein Knoten als Endknoten deklariert wird, und
- die Zuweisung einer Klasse oder eines Wertes zu jedem Endknoten,

wird auf die Originalliteratur von Breiman et al. (1984) verwiesen.

2.2. Conditional Inference-Bäume

Hinter der Methodik von Conditional Inference-Bäumen (kurz CIT), die von Hothorn et al. (2006b) eingeführt wurde, steht das Konzept der Durchführung bedingter Inferenztests und deren statistische Signifikanz. Die nachfolgenden Erläuterungen orientieren sich im Wesentlichen an der Literatur von Hothorn et al. (2006b).
Vor jeder Aufteilung von $\mathcal{X}$ werden alle p Prädiktoren hinsichtlich ihres Zusammenhanges mit dem Response, wobei die Prädiktoren und der Response beliebig skaliert (z.B. nominal, stetig, ordinal, etc.) sein können, geprüft. Die beste Aufteilung im Sinne von maximaler Homogenität innerhalb der Teilmengen und maximaler Heterogenität zwischen den Teilmengen wird durch denjenigen Prädiktor gewährleistet, der den kleinsten p-Wert und damit die stärkste Assoziation mit dem Response aufweist. Fortführend wird in jedem Knoten ein Signifikanztest mit der Nullhypothese der Unabhängigkeit zwischen dem Response und allen Prädiktoren durchgeführt, solange bis die Nullhypothese nicht mehr abgelehnt werden kann. Falls die Nullhypothese nicht abgelehnt wird, wird der jeweilige Knoten nicht weiter verzweigt und entspricht einem Endknoten.
Es soll beachtet werden, dass dieses Verfahren des simultanen Testens von p Hypothesen (p = Anzahl an Prädiktoren) für jede Aufteilung das Problem des multiplen Testens birgt, weshalb eine Adjustierung der p-Werte durch z.B. die Bonferroni-Korrektur notwendig ist.

Für den Fall von kategorialen Prädiktoren mit vielen Klassen und somit vielen möglichen Aufteilungen ist die Verwendung von CIT gegenüber CART zu bevorzugen, da im Allgemeinen lediglich mit erstgenannter Methode die Prädiktoren bei jeder Aufteilung unverzerrt ausgewählt werden können (Hothorn et al., 2006b; Strobl et al., 2007). Ebenso vorteilhaft bei dem Einsatz von CIT ist, dass durch dessen zugrundeliegenden Algorithmus das Problem von Overfitting nicht besteht (Hothorn et al., 2006b). Außerdem ist der CIT-Ansatz dem klassischen CART-Ansatz in der Hinsicht überlegen, dass für den Fall eines ordinalen Response dessen inhärente Ordnung berücksichtigt wird, während die Ordnung bei CART ignoriert wird (Janitza et al., 2016). Dieser Umgang führt zu einem Informationsverlust und kann in weniger zuverlässigen Vorhersagen resultieren.

Für den Fall eines ordinalen Response resultieren im Rahmen von Conditional Inference-Bäumen sogenannte ordinale Regressionsbäume, indem den geordneten Klassen j ($j \in \{1, ..., J\}$) des ordinalen Response metrische Scores $s(j) \in \mathbb{R}$ zugeordnet werden, die die Distanz zwischen den Klassen widerspiegeln. Dadurch wird die Ordnungsskala in eine metrische Skala umgewandelt. Gemäß Janitza et al. (2016) haben spezifische Werte für die Scores keinen signifikanten Einfluss auf die Vorhersage, weshalb beliebige Werte oder die Standardwerte für die Scores verwendet werden können. Die Standard-Scores entsprechen den Klassenwerten des Response, d.h. $s(j) = j$ ($j \in \{1, ..., J\}$), womit die Distanzen zwischen allen Klassen gleich sind.
Vor jeder Aufteilung eines Knotens in zwei Teilmengen werden alle p Prädiktoren in Bezug auf ihre Assoziation mit dem umgewandelten Response getestet. Der am stärksten mit dem Response assoziierte Prädiktor wird ausgewählt und dessen Grenzwert für die Zerlegung stellt maximale Diskrepanz in den Scores zwischen den beiden resultierenden Teilmengen sicher.

3. Einführung in Random Forests

3.1. Grundlagen von Random Forests

Random Forests wurden von Breiman (2001) eingeführt und kombinieren mehrere möglichst unkorrelierte Bäume, die jeweils aus unterschiedlichen Bootstrap-Stichproben (Zufallsstichproben mit Zurücklegen) von Beobachtungen aus dem originalen Datensatz erstellt werden. Dabei wird eine möglichst hohe Anzahl von Bäumen empfohlen.
Random Forests können beispielsweise Ensembles aus Klassifikationsbäumen, Regressionsbäumen (im Folgenden als Regression Forests bezeichnet) oder Conditional Inference-Bäumen (Conditional Inference Forests) bilden.
Um eine verringerte Korrelation zwischen den Bäumen zu gewährleisten, erfolgt für jede Aufteilung die Auswahl eines Prädiktoren aus einer bestimmten Anzahl von $mtry \leq p$ zufällig gezogenen Prädiktoren. Im Allgemeinen reduziert sich die Korrelation zwischen den Bäumen mit einer geringeren Anzahl $mtry$ an Prädiktoren, es entsteht jedoch ein Bias bei einer zu geringen Prädiktorenanzahl. Dies wird von Koch (2016) damit begründet, dass für den Fall von wenigen einflussreichen Prädiktoren die Auswahl für einige Aufteilungen höchstwahrscheinlich auf nicht-einflussreiche Prädiktoren beschränkt wird mit der Folge einer schlechten Aufteilung. Als Standardwerte für $mtry$ gelten $\sqrt{p}$ für den Klassifikationsfall und $\frac{p}{3}$ für den Regressionsfall (Liaw und Wiener, 2002).

Diejenigen Beobachtungen aus dem originalen Datensatz, die nicht für die Erstellung eines Baumes verwendet werden, sind für diesen Baum sogenannte „Out-Of-Bag"- oder OOB-Beobachtungen (Breiman, 2001). Jede Beobachtung ist für mehrere Bäume eine OOB-Beobachtung, da die einzelnen Bäume auf verschiedenen Bootstrap-Stichproben basieren. Eine (OOB-)Beobachtung kann durch die Aggregation der Vorhersagen von denjenigen Bäumen, an deren Konstruktion diese Beobachtung nicht beteiligt war, vorhergesagt werden. Das Aggregationsverfahren für die Vorhersage einer Beobachtung unterscheidet sich abhängig davon, aus welcher Art von Bäumen der Random Forest besteht (Liaw und Wiener, 2002). So werden bei einem Regression Forest zu diesem Zweck die Vorhersagen über alle entsprechenden Bäume

gemittelt (Breiman, 2001). Daneben wird bei einem Random Forest aus Klassifikationsbäumen eine Mehrheitsentscheidung durchgeführt, d.h. die von den Bäumen am häufigsten vorhergesagte Klasse entspricht der Klassenvorhersage (Malley et al., 2012). Ferner wird bei der ordinalen Version eines Conditional Inference Forest (aus ordinalen Regressionsbäumen, in Abschnitt 2.2 eingeführt) eine Klassenvorhersage durch die Zuordnung in die wahrscheinlichste Klasse ermittelt (Janitza et al., 2016). Überdies ergeben sich die Wahrscheinlichkeitsvorhersagen für die erwähnten Random Forests, indem die Häufigkeiten der Klassen in den entsprechenden Endknoten über alle Bäume hinweg gemittelt werden (Malley et al., 2012; Janitza et al., 2016).

Für die Benchmarkstudie in Kapitel 4 sind ordinale Conditional Inference Forests und Ordinal Forests (siehe nachfolgenden Abschnitt 3.2) zentral. Für die Conditional Inference Forests werden, begründet durch die Arbeit von Janitza et al. (2016), standardmäßig die Klassenwerte $1, ..., J$ als Score-Werte für die Klassen des ordinalen Response verwendet.

Insgesamt liefern Random Forests zuverlässigere Vorhersagen als einzelne Bäume, allerdings ist die Interpretation von Random Forests schwierig, da der Einfluss einzelner Prädiktoren schwer bewertet werden kann (Strobl et al., 2007). Um dieser Problematik zu begegnen, können Messgrößen für die Wichtigkeit von Prädiktoren ermittelt werden. Die Wichtigkeit von Prädiktoren spiegelt deren Einfluss bei der Vorhersage des Response wider. Solche Messgrößen werden in Abschnitt 3.3 für Random Forests vorgestellt.

3.2. Ordinal Forests

Die Methodik der Ordinal Forests, die von Hornung (2020) entwickelt wurde, dient der Vorhersage eines ordinalen Response unter Berücksichtigung dessen Ordnung und kann zusätzlich die Prädiktoren sowohl für niedrig- als auch hochdimensionale Daten nach ihrer Wichtigkeit für die Vorhersage ordnen. Die nachfolgenden Erläuterungen beziehen sich im Wesentlichen auf die Originalliteratur von Hornung (2020).

Die Vorhersagemethode der Ordinal Forests weist Ähnlichkeiten zu der klassischen Methode der Regression Forests (Breiman, 2001) für stetige Responses auf, jedoch mit dem konzeptionellen Unterschied, dass Ordinal Forests optimierte Score-Werte und nicht die Klassenwerte $1, ..., J$ für die entsprechenden Klassen des ordinalen Response verwenden. Die optimierten Score-Werte können durch Maximierung der

OOB-Vorhersageleistung während der Erstellung des Ordinal Forest erhalten werden und führen bei deren Verwendung zu einer erhöhten Vorhersageleistung.
In Übereinstimmung mit Regressionsmodellen für einen ordinalskalierten Response, die von McCullagh (1980) vorgestellt wurden, basieren Ordinal Forests auf dem Konzept eines latenten (nicht gemessenen oder unbekannten) stetigen Response y^*, der die Werte des beobachteten ordinalen Response y bestimmt. Nach diesem Konzept wird mit dem ordinalen Response als stetigen Response verfahren, wobei die Klassen des ordinalen Response implizit einbezogen werden.
Laut Hornung (2020) ist das methodische Vorgehen bei Ordinal Forests, verglichen mit dem Vorgehen bei Regression Forests, für den Fall eines unbekannten stetigen Response y^* besser für die Prognose eines ordinalen Response y geeignet. Dies liegt insbesondere daran, dass bei dem erstgenannten Konzept die unterschiedlichen Klassenbreiten des ordinalen Response y, also die Variation in den Ausmaßen seiner geordneten Klassen J, durch die Breiten von J benachbarten Intervallen berücksichtigt werden. Die Breiten der J Intervalle werden geschätzt bzw. optimiert durch die Maximierung der OOB-Vorhersageleistung, wodurch optimierte Score-Werte resultieren. Entgegen der Feststellung von Janitza et al. (2016), dass spezifische Werte für die Scores keinen signifikanten Einfluss auf die Vorhersage haben, zeigt Hornung (2020) in seinen Analysen, dass die Verwendung von optimierten Score-Werten für die Werte der Klassen von y in einer verbesserten Vorhersageleistung resultiert. Trotz dieser starken Eigenschaft liefern die Score-Werte bzw. die geschätzten Intervallbreiten keine relevanten Informationen über die wahren Intervallbreiten, was Hornung (2020) in seinen Analysen umfangreich untersuchte.

Zusammenfassend definieren sich Ordinal Forests als Regression Forests, die sich durch den Einsatz von optimierten Score-Werten auszeichnen, wodurch die Vorhersageleistung verbessert wird.

3.2.1. Ordinal Forest-Algorithmus

Mit dem Ordinal Forest-Algorithmus wird eine Vorhersageregel zur Vorhersage des ordinalen Response für neue Beobachtungen konstruiert. In dem Algorithmus sind die folgenden zwei Schritte zentral:

1. die Optimierung der Score-Werte und
2. die Erstellung eines Ordinal Forest als Regression Forest.

Dabei werden die Vektoren $\boldsymbol{x}_i, i = 1, ..., n$, mit den Prädiktoren für die Beobachtungen und die Klassenwerte $y_i, i = 1, ..., n$, des ordinalen Response angenommen.

1. Optimierung der Score-Werte

(a) Für die Optimierung einer Menge an Score-Werten $\{s_1, ..., s_J\}$ wird mehrfach eine Menge $\{s_{b,1}, ..., s_{b,J}\}$ ($b \in \{1, ..., B_{\text{sets}}\}$) an zufälligen Score-Werten als mögliche Kandidaten für die optimale Score-Menge, die die OOB-Vorhersageleistung maximiert, generiert. Dabei sollen möglichst viele und voneinander unterschiedliche Score-Mengen (z.B. $B_{\text{sets}} = 1000$) erzeugt werden, damit die Auswahl der besten Menge, die der optimalen Menge am nächsten ist, gewährleistet wird. Für die Unterschiedlichkeit der Mengen wird ein Algorithmus angewendet, der in dem Beitrag von Hornung (2020) detailliert ausgeführt wird.

Die nachfolgenden Schritte (b) - (e) werden für jede generierte Score-Menge $b = 1, ..., B_{\text{sets}}$ aus Schritt (a) wiederholt.

(b) Jede Score-Menge besteht aus den J Mittelpunkten von J benachbarten Intervallen, die eine Partition des Intervalls $[0, 1]$ darstellen. Die $[0, 1]$-Partition $\{d_{b,1}, ..., d_{b,J+1}\}$ wird generiert, indem zunächst $J - 1$ Werte von einer gleichverteilten Zufallsvariable gezogen werden. Nach anschließendem Sortieren der Werte werden diese als $d_{b,2}, ..., d_{b,J}$ bezeichnet und $d_{b,1} := 0$ und $d_{J+1} := 1$ festgelegt. Schließlich ergibt sich die b-te Menge $\{d_{b,1}, ..., d_{b,J+1}\}$.

(c) Ein stetiger Response $\boldsymbol{z}_b := z_{b,1}, ..., z_{b,n}$ wird gebildet, indem anstelle jedes Klassenwertes j ($j \in \{1, ..., J\}$) des ordinalen Response $\boldsymbol{y} := y_1, ..., y_n$ der j-te Wert aus der Score-Menge $\boldsymbol{s}_b := \{s_{b,1}, ..., s_{b,J}\}$ verwendet wird, wobei $s_{b,j} := \Phi^{-1}(c_{b,j})$, $c_{b,j} := \frac{(d_{b,j}+d_{b,j+1})}{2}$ mit Φ^{-1} als Quantilsfunktion der Standardnormalverteilung.

(d) Ferner erfolgt die Konstruktion eines Ordinal Forest f_{s_b} als Regression Forest für $\boldsymbol{z}_b$ als Response und mit einer Anzahl an Bäumen von $B_{\text{ntreeprior}}$ (z.B. $B_{\text{ntreeprior}} = 100$).

(e) Daran anschließend wird die Messung der OOB-Vorhersageleistung des in (d) konstruierten Ordinal Forest unter Verwendung einer bestimmten Performance-Funktion durchgeführt, deren Wahl abhängig davon ist, was der Ordinal Forest in Bezug auf die Genauigkeit der Vorhersage leisten soll (siehe Abschnitt 3.2.2). Dieser Schritt kann in drei Teilschritte untergliedert werden:

i Zunächst werden die OOB-Vorhersagen $\hat{z}_{b,1}, ..., \hat{z}_{b,n}$ von $z_{b,1}, ..., z_{b,n}$ gemessen.

ii In einem zweiten Schritt werden die OOB-Vorhersagen $\hat{y}_{b,1}, ..., \hat{y}_{b,n}$ von $y_1, ..., y_n$ wie folgt geschätzt: $\hat{y}_{b,i} := j$, falls $\hat{z}_{b,i} \in]\Phi^{-1}(d_{b,j}), \Phi^{-1}(d_{b,j+1})]$ ($i \in \{1, ..., n\}$). Das heißt, wenn $\hat{z}_{b,i}$ in dem j-ten Intervall liegt, nimmt die Klasse des ordinalen Response $\hat{y}_{b,i}$ den Wert j an.

iii Anschließend wird dem Ordinal Forest f_{s_b} ein Performance-Score $sc_b := g(\boldsymbol{y}, \hat{\boldsymbol{y}}_b)$ zugewiesen, wobei $\hat{\boldsymbol{y}}_b := \hat{y}_{b,1}, ..., \hat{y}_{b,n}$ und g eine bestimmte Performance-Funktion ist.

(f) Die erste Phase des Algorithmus schließt mit der Erstellung der finalen Menge von Score-Werten durch die Zusammenfassung der als optimal deklarierten Score-Mengen, wobei Optimalität mit dem Erreichen der höchsten Schätzungen für die OOB-Vorhersageleistung, also den höchsten sc_b-Werten verbunden ist. Dabei bezeichnet B_{bestsets} (z.B. $B_{\text{bestsets}} = 10$) die Anzahl der als optimal deklarierten Score-Mengen und S_{best} die Menge von deren entsprechenden Indizes. Die Zusammenfassung der B_{bestsets} erfolgt für jedes $j \in \{1, ..., J+1\}$ durch die Berechnung des Durchschnitts aus denjenigen Werten $d_{b,j}$, für die $b \in S_{\text{best}}$ ist. Der berechnete Durchschnitt entspricht einer Menge von Werten, die als $d_1, ..., d_{J+1}$ bezeichnet werden.

Durch die Zusammenfassung von mehreren Score-Mengen zu einer endgültigen Score-Menge kann gemäß Hornung (2020) mit hoher Sicherheit ausgeschlossen werden, dass die Schätzungen für die OOB-Vorhersageleistung aller zusammengefassten Score-Mengen rein zufällig hoch sind, anders als bei der Verwendung von lediglich einer Score-Menge – die mit der höchsten Schätzung – als finale Score-Menge.

2. Erstellung eines Ordinal Forest als Regression Forest

(a) Ein neuer stetiger Response $\boldsymbol{z} := z_1, ..., z_n$ wird gebildet, indem anstelle jedes Klassenwertes j ($j \in \{1, ..., J\}$) des ordinalen Response $\boldsymbol{y} := y_1, ..., y_n$ der j-te Wert aus der optimierten Score-Menge $\{s_1, ..., s_J\}$ verwendet wird, wobei $s_j := \Phi^{-1}(c_j)$ und $c_j := \frac{(d_j + d_{j+1})}{2}$.

(b) Schließlich erfolgt die Erstellung des Ordinal Forest f_{final} als Regression Forest für $\boldsymbol{z}$ als Response und mit einer Anzahl an Bäumen von B_{ntree} (z.B. $B_{\text{ntree}} = 5000$). Der Ordinal Forest f_{final} entspricht der Vorhersageregel.

Vorhersage mit einem Ordinal Forest

Der oben vorgestellte Algorithmus dient der Konstruktion einer Vorhersageregel, d.h. eines Ordinal Forest f_{final}, der den Wert des Response für eine i^*-te unabhängige Beobachtung auf Grundlage ihres Prädiktorenvektors $\boldsymbol{x}_{i*}$ vorhersagt. Zu diesem Zweck dient der nachfolgende Algorithmus.

1. Für $b = 1, ..., B_{\text{ntree}}$ mit B_{ntree} als Anzahl an Bäumen in f_{final} gilt:

(a) Für die i^*-te Beobachtung wird mit dem b-ten Baum in f_{final} eine Vorhersage $\hat{z}_{i^*,b}$ geschätzt.

(b) Ferner wird eine Klassenvorhersage mit dem b-ten Baum wie folgt ermittelt: $\hat{y}_{i^*,b} := j$, falls $\hat{z}_{i^*,b} \in]\Phi^{-1}(d_j), \Phi^{-1}(d_{j+1})]$. Das heißt, wenn $\hat{z}_{i^*,b}$ in dem j-ten Intervall liegt, nimmt die Klasse des ordinalen Response $\hat{y}_{i^*,b}$ den Wert j an.

2. Schließlich erfolgt eine endgültige Klassenvorhersage von $\hat{y}_{i^*,1}, ..., \hat{y}_{i^*,B_{\text{ntree}}}$ durch eine Mehrheitsentscheidung, d.h. die Klassenvorhersage entspricht derjenigen Klasse, die von den Bäumen am häufigsten vorhergesagt wird – mit Ausnahme von dem Fall, dass die Performance-Funktion basierend auf dem Ranked Probability Score angewendet wird (siehe nachfolgenden Abschnitt 3.2.2). In dieser Situation entspricht die Klassenvorhersage derjenigen Klasse mit der höchsten Klassenwahrscheinlichkeit.

3.2.2. Performance-Funktion

Ordinal Forests werden in dem Programm `R` (R Core Team, 2021) für statistische Berechnungen mit dem Paket `ordinalForest` von Hornung (2021) und der darin enthaltenen Funktion `ordfor` erstellt. Die Funktion ermöglicht mit dem inhärenten Argument `importance` die Auswahl eines geeigneten Maßes für die Variablenwichtigkeit (siehe Abschnitt 3.3) und daneben mit `perffunction` die Selektion einer passenden Performance-Funktion. In diesem Abschnitt liegt der Fokus auf letzterer Auswahlmöglichkeit.

Die Performance-Funktion dient der Messung der Vorhersageleistung der OOB-Beobachtungen, die nicht für die Konstruktion des Ordinal Forest genutzt wurden (siehe Schritt (e) des Ordinal Forest-Algorithmus in Abschnitt 3.2.1). Abhängig von der Motivation, was der Ordinal Forest leisten soll, wird eine bestimmte Variante der Performance-Funktion angewendet. Diese Arbeit fokussiert sich auf zwei Varianten, die später in der Vergleichsstudie für die Ordinal Forests eingesetzt werden. Bei den ausgewählten Versionen handelt es sich um die bestehende Performance-Funktion Equal und die neue Performance-Funktion Ranked Probability Score. In dem Beitrag von Hornung (2020) werden weitere Performance-Funktionen zur Verfügung gestellt, die allerdings auf spezifische – und hier nicht intendierte – Situationen zugeschnitten sind und deshalb nicht erläutert werden.

3.2.2.1. Performance-Funktion Equal

Die Performance-Funktion Equal basiert, anders als die Performance-Funktion Ranked Probability Score, auf dem Youden-Index. Die allgemeine Form der Performance-Funktion g basierend auf dem Youden-Index ist

$$g(\boldsymbol{y}, \hat{\boldsymbol{y}}) := \sum_{j=1}^{J} w_j \ \text{Yind}(\boldsymbol{y}, \hat{\boldsymbol{y}}, j) \tag{3.1}$$

$$\text{mit} \quad \sum_{j} w_j = 1 \quad \text{und} \quad \text{Yind}(\boldsymbol{y}, \hat{\boldsymbol{y}}, j) := \text{sens}(\boldsymbol{y}, \hat{\boldsymbol{y}}, j) + \text{spec}(\boldsymbol{y}, \hat{\boldsymbol{y}}, j) - 1,$$

$$\text{wobei} \quad \text{sens}(\boldsymbol{y}, \hat{\boldsymbol{y}}, j) := \frac{\#\{y_i = j \wedge \hat{y}_i = j \mid i \in \{1, ..., n\}\}}{\#\{y_i = j \mid i \in \{1, ..., n\}\}} \quad \text{und}$$

$$\text{spec}(\boldsymbol{y}, \hat{\boldsymbol{y}}, j) := \frac{\#\{y_i \neq j \wedge \hat{y}_i \neq j \mid i \in \{1, ..., n\}\}}{\#\{y_i \neq j \mid i \in \{1, ..., n\}\}}.$$

Das Symbol # bezeichnet die Mächtigkeit einer Menge und $\hat{\boldsymbol{y}} := \{\hat{y}_1, ..., \hat{y}_n\}$ die Schätzung von $\boldsymbol{y}$. Ferner stellt $\text{Yind}(\boldsymbol{y}, \hat{\boldsymbol{y}}, j)$ den Youden-Index bezüglich der Klasse j ($j \in \{1, ..., J\}$) dar. Die durch diesen Index gemessene Fähigkeit eines Ordinal Forest zwischen Beobachtungen zu differenzieren, die zur Klasse j und die nicht zur Klasse j gehören, steigt mit dem Gewicht w_j für die Klasse j.

Die Performance-Funktion Equal ergibt sich durch das Gewicht $w_j := \frac{1}{J}$:

$$g_{\text{clequal}}(\boldsymbol{y}, \hat{\boldsymbol{y}}) := \sum_{j=1}^{J} \frac{1}{J} \ \text{Yind}(\boldsymbol{y}, \hat{\boldsymbol{y}}, j). \tag{3.2}$$

Diese Performance-Funktion wird verwendet, wenn der resultierende Ordinal Forest die Beobachtungen aus jeder Klasse mit der gleichen Genauigkeit klassifizieren soll, unabhängig von der Klassengröße. Die Klassenvorhersagen für die Beobachtungen ergeben sich durch eine Mehrheitsentscheidung bezüglich der Vorhersagen von den einzelnen Bäumen in dem Ordinal Forest. Demnach entspricht eine Klassenvorhersage demjenigen Klassenwert, der von den Bäumen am häufigsten vorhergesagt wird.

3.2.2.2. Performance-Funktion Ranked Probability Score

Der Ranked Probability Score, kurz RPS, stellt eine Erweiterung des Brier-Scores (Brier, 1950) von zwei Klassen für binäre Responses auf mehrere geordnete Klassen für ordinale Responses dar (Epstein, 1969; Murphy, 1970).
Der (negative) RPS als Performance-Funktion kann angewendet werden, wenn neben den Klassenvorhersagen auch die vorhergesagten Klassenwahrscheinlichkeiten von Interesse sind. Der RPS misst die Diskrepanz zwischen den vorhergesagten Klassenwahrscheinlichkeiten und den wahren Klassenwerten wie folgt (Ebert, 2005):

$$g_{\text{clrps}}(\boldsymbol{y}, \hat{\boldsymbol{\pi}}) = \frac{1}{J-1} \sum_{j=1}^{J} \sum_{i=1}^{n} \Big(\hat{\pi}_i(j) - I(y_i \leq j) \Big)^2. \tag{3.3}$$

Dabei bezeichnet J erneut die Anzahl an Klassen, $\hat{\boldsymbol{\pi}}(j) := \{\hat{\pi}_1(j), ..., \hat{\pi}_n(j)\}$ die vorhergesagten Klassenwahrscheinlichkeiten und $\boldsymbol{y} := \{y_1, ..., y_n\}$ die wahren Klassenwerte. Der Operator $I(y_i \leq j)$ mit $I(\cdot)$ als Indikatorfunktion resultiert bei dem wahren Klassenwert j in eine vereinfachte Treppenfunktion mit einer Stufe, die von 0 bis 1 reicht (Janitza et al., 2016).
Die vorhergesagten Klassenwahrscheinlichkeiten für einen Ordinal Forest ergeben sich als Durchschnitt aus den vorhergesagten Klassenwahrscheinlichkeiten der einzelnen Bäume (Hornung, 2021). Die Klassenwahrscheinlichkeiten der einzelnen Bäume entsprechen den relativen Häufigkeiten der Beobachtungen aus den unterschiedlichen Klassen in den Endknoten, in denen sich die neue Beobachtung befindet. Für die Messung der Vorhersageleistung von OOB-Beobachtungen werden jeweils nur die Bäume verwendet, für die die Beobachtungen nicht in den Trainingsdaten für die Baumkonstruktion enthalten sind.
Bei Verwendung des RPS können die Klassenvorhersagen durch die vorhergesagten Klassenwahrscheinlichkeiten erhalten werden. Die Klassenvorhersagen entsprechen den Klassen mit den höchsten Klassenwahrscheinlichkeiten (Hornung, 2021).

Der RPS in obiger Gleichung (3.3) wird durch $\frac{1}{J-1}$ auf dem Intervall von 0 bis 1 skaliert und nimmt einen Wert näher an 0 bei einer besseren Vorhersage an, die sich aus geringen Distanzen zwischen den vorhergesagten Klassenwahrscheinlichkeiten und wahren Klassenwerten ergibt (Ebert, 2005). Umgekehrt nimmt der RPS einen höheren Wert bei einer schlechteren Vorhersage an, die daraus resultiert, dass die Wahrscheinlichkeiten weiter von dem wahren Ergebnis weg liegen. Aus diesem Grund wird der negative RPS als Performance-Funktion verwendet (Hornung, 2021).

Laut Hornung (2021) kann für den Fall, dass keine Wahrscheinlichkeitsvorhersagen, sondern einfache Klassenvorhersagen von Interesse sind, die Performance-Funktion RPS weniger gut und die Performance-Funktion Equal besser geeignet sein. Diese Annahme wird in der Benchmarkstudie, die in Kapitel 4 durchgeführt wird, geprüft.

3.3. Variablenwichtigkeit

Wie zu Beginn des vorherigen Abschnitts 3.2.2 erwähnt, wird mit der Funktion `ordfor` für die Konstruktion eines Ordinal Forest über den Parameter `importance` ein Maß für die Variablenwichtigkeit, kurz VIM, ausgewählt. Das VIM ist ein Maß, das für die Unterscheidung zwischen einflussreichen und nicht-einflussreichen Prädiktoren verwendet wird. Ordinal Forests in `R` bieten die Auswahl zwischen dem bestehenden „Misclassification Error"-VIM, kurz ER-VIM, und dem neuen „Ranked Probability Score"-VIM, kurz RPS-VIM. Die Berechnung der beiden VIM-Varianten unterscheidet sich im Hinblick auf das verwendete Fehlermaß. So basiert das ER-VIM auf dem Klassifikationsfehler und das RPS-VIM auf dem Ranked Probability Score. Beide Maße sind für die Vergleichsstudie in Kapitel 4 sowohl für Ordinal Forests als auch Conditional Inference Forests von zentraler Bedeutung und werden für die Random Forest-Varianten in gleicher Weise berechnet.
Damit liegt der Fokus auf Variablenwichtigkeitsmaßen mit Permutationen, für deren Berechnung die OOB-Beobachtungen verwendet werden und die im Allgemeinen eine unverzerrte Variablenauswahl gewährleisten.

3.3.1. Variablenwichtigkeit mit dem Klassifikationsfehler

Die Berechnung des ER-VIM basierend auf dem Klassifikationsfehler als Fehlermaß erfolgt für einen Random Forest mit einem Prädiktor m und der Anzahl an Bäumen B_{ntree} wie folgt (Breiman, 2001):

$$VI_m := \frac{1}{B_{\text{ntree}}} \sum_{b=1}^{B_{\text{ntree}}} \text{ER}(\boldsymbol{y}_{OOB,b,m}, \hat{\boldsymbol{y}}'_{OOB,b,m}) - \text{ER}(\boldsymbol{y}_{OOB,b,m}, \hat{\boldsymbol{y}}_{OOB,b,m}), \tag{3.4}$$

wobei $\boldsymbol{y}_{OOB,b,m}$ den Vektor der Klassenwerte für die OOB-Beobachtungen des b-ten Baumes, also für diejenigen Beobachtungen, die für die Konstruktion des b-ten Baumes nicht verwendet wurden, angibt. Weiterführend bezeichnet $\hat{\boldsymbol{y}}'_{OOB,b,m}$ die Vorhersagen der Klassenwerte für die OOB-Beobachtungen des b-ten Baumes nach zufälliger Permutation der Werte eines Prädiktoren m, während $\hat{\boldsymbol{y}}_{OOB,b,m}$ diejenigen Vorhersagen ohne Permutation der Prädiktorenwerte angibt. Dementsprechend ist

der Minuend der Subtraktion, $\text{ER}(\boldsymbol{y}_{OOB,b,m}, \hat{\boldsymbol{y}'}_{OOB,b,m})$, der Fehler des b-ten Baumes bei der Vorhersage der Klassenwerte für die OOB-Beobachtungen nach zufälliger Permutation der Werte des m-ten Prädiktoren. Daneben ist der Subtrahend, $\text{ER}(\boldsymbol{y}_{OOB,b,m}, \hat{\boldsymbol{y}}_{OOB,b,m})$, der Vorhersagefehler vor zufälliger Permutation der Werte des m-ten Prädiktoren.

Die Fehlerfunktion ER verwendet für das ER-VIM in Gleichung (3.4) den Klassifikationsfehler:

$$\text{ER} = \frac{1}{n} \sum_{i=1}^{n} I(y_i \neq \hat{y}_i) \tag{3.5}$$

mit y_i als wahren Klassenwert für die i-te Beobachtung, $i = 1, ..., n$, und $\hat{y}_i$ als vorhergesagten Klassenwert für diese Beobachtung. Hierbei wird nur zwischen einer richtigen Vorhersage ($y_i = \hat{y}_i$) und einer falschen Vorhersage ($y_i \neq \hat{y}_i$) differenziert (Janitza et al., 2016). Dadurch kann die Reihenfolge der Klassen nicht berücksichtigt werden und zusätzlich werden alle Fehlklassifikationen als gleich schlecht angesehen, d.h. es findet keine Unterscheidung statt zwischen einer Zuordnung in eine Klasse, die weit entfernt von der wahren Klasse y ist, und in eine Klasse, die nah an y ist.

Die mittlere Differenz der Vorhersagefehler oder vielmehr der Vorhersagegenauigkeit der einzelnen Bäume vor und nach Permutation in Gleichung (3.4), die für jeden Prädiktor separat berechnet wird, spiegelt die Wichtigkeit der einzelnen Prädiktoren wider (Janitza et al., 2016). Durch die Permutation wird die Assoziation zwischen den Prädiktoren und dem Response aufgelöst. Falls ein Prädiktor keinen Einfluss auf den Response hat und somit ein Rauschprädiktor ist, ist sein Ergebnis vor und nach Permutation ähnlich und seine Wichtigkeit für die Vorhersagegenauigkeit nahe Null. Wenn demgegenüber ein Prädiktor einen Einfluss hat und dementsprechend ein Signalprädiktor ist, ergibt sich für die Differenz der Fehler vor und nach Permutation seiner Werte ein positives Ergebnis, wodurch dem Prädiktor eine hohe Wichtigkeit zugeschrieben wird. Die Wichtigkeit des Prädiktoren steigt mit der Abweichung der Fehler vor und nach Permutation seiner Werte. Das positive Ergebnis resultiert daraus, dass mit dem Ausschalten der Assoziation eine schlechtere Vorhersage einhergeht, die sich in einem größeren Fehler widerspiegelt.
Demnach bilden VIMs eine Rangfolge für die Prädiktoren, ein sogenanntes Prädiktoren-Ranking, entsprechend ihrer Assoziation mit dem Response ab und unterscheiden dabei zwischen Rauschprädiktoren, also Prädiktoren ohne Effekt auf den Response, und Signalprädiktoren, die einen Effekt auf den Response haben.

3.3.2. Variablenwichtigkeit mit dem Ranked Probability Score

In Gleichung (3.4) kann anstatt des Klassifikationsfehlers ebenso der Ranked Probability Score als Fehlermaß für das VIM angewendet werden. Wie bereits erwähnt, wird das RPS-VIM für die verschiedenen Random Forest-Varianten in dieser Arbeit auf die gleiche Weise berechnet.
Für das RPS-VIM werden die Klassifikationsfehler der einzelnen Bäume in Gleichung (3.4) durch die jeweiligen RPS-Werte, die durch Gleichung (3.3) erhalten werden, ersetzt. Die Berechnung der RPS-Werte wird in Gleichung (3.3) an Ordinal Forests demonstriert, ist aber auf Conditional Inference Forests übertragbar.
Das resultierende RPS-VIM misst die Abweichung der Wahrscheinlichkeitsvorhersagen vor und nach Permutation, getrennt für jeden Prädiktor. Einem Prädiktor wird eine umso höhere Wichtigkeit zugeschrieben, je schlechter seine Wahrscheinlichkeitsvorhersagen nach Ausschalten der Assoziation mit dem Response, also nach zufälliger Permutation seiner Werte sind, was sich in einem höheren RPS-Wert widerspiegelt.

Das ER-VIM und RPS-VIM unterscheidet der Umgang mit dem ordinalen Response. Bei Verwendung des RPS-VIM wird die inhärente Ordnung der Klassen des ordinalen Response einbezogen, während dies bei Verwendung des ER-VIM nicht der Fall ist (Hornung, 2021). Demnach besteht die Vermutung, dass das RPS-VIM bessere Prädiktoren-Rankings ermöglicht als das ER-VIM. Dies wird in der Vergleichsstudie in Kapitel 4 überprüft. Dazu kann ergänzt werden, dass das RPS-VIM zwar die Ordnung der Klassen berücksichtigt, jedoch keine Annahmen über die Abstände zwischen den Klassen trifft (Janitza et al., 2016).

3.4. Ableitung verschiedener Ordinal Forest-Varianten

Aus dem Zusammenspiel der beiden zentralen Parameter für die Erstellung eines Ordinal Forest, nämlich Performance-Funktion und Variablenwichtigkeitsmaß, können unterschiedliche Varianten eines Ordinal Forest generiert werden. Dazu zählen die bestehende Variante mit Performance-Funktion Equal und ER-VIM und die neue Variante mit Performance-Funktion RPS und RPS-VIM. Zusätzlich ergibt sich eine weitere essenzielle Version mit Performance-Funktion Equal und RPS-VIM.

Laut Hornung (2020) besteht die Möglichkeit, dass die Verwendung desselben Fehlermaßes sowohl für das VIM als auch für die Performance-Funktion in einer verbesserten Variablenselektion resultiert. Dies würde dazu führen, dass mit der Kom-

bination aus Performance-Funktion RPS und RPS-VIM einflussreiche Prädiktoren besser identifiziert werden können als mit den restlichen Kombinationen. Unabhängig davon werden der Kombination mit ER-VIM aufgrund der Missachtung der Ordnung in den Klassen schlechtere Ergebnisse hinsichtlich der Variablenauswahl zugeschrieben (Hornung, 2021). Damit geht die Annahme einher, dass der Ordinal Forest mit Performance-Funktion Equal und ER-VIM schlechter als die beiden Ordinal Forest-Varianten mit RPS-VIM abschneidet, wenn es um die Qualität des Prädiktoren-Rankings geht. Die Möglichkeit aus Performance-Funktion RPS und ER-VIM wird aus den genannten Gründen nicht betrachtet.

Darüber hinaus sind in der Vergleichsstudie in Kapitel 4 einfache Klassenvorhersagen und keine Wahrscheinlichkeitsvorhersagen für die Klassen von Interesse, weshalb der Ordinal Forest mit Performance-Funktion Equal im Vergleich zu dem Ordinal Forest mit Performance-Funktion RPS bessere Ergebnisse hinsichtlich der Vorhersageleistung erzielen könnte, unabhängig von der Auswahl des VIM.

Auf theoretischer Basis lassen sich die obigen Annahmen schwer verifizieren, weshalb eine Benchmarkstudie mit den verschiedenen Ordinal Forest-Varianten und zusätzlich Conditional Inference Forests (Hothorn et al., 2006b) für einen ordinalen Response durchgeführt wird.

4. Benchmarkstudie der Forests hinsichtlich Variablenwichtigkeit und Klassifikation

4.1. Anwendung der Forests in R

Die statistischen Berechnungen für die Benchmarkstudie der Random Forest-Varianten anhand von simulierten und realen ordinalen Daten werden mit dem Programm R (R Core Team, 2021) durchgeführt. Wie bereits in Abschnitt 3.2.2 dargelegt, sind Ordinal Forests in dem R-Paket `ordinalForest` (Hornung, 2021) implementiert und verwenden die Funktion `ordfor`. Der für die Benchmarkstudie ebenso relevante Conditional Inference Forest ist in dem R-Paket `party` (Hothorn et al., 2006a; Strobl et al., 2007; Strobl et al., 2008) implementiert und kann mit der Funktion `cforest` konstruiert werden.

Damit für die Konstruktion der zu vergleichenden Forest-Varianten die gleichen Bedingungen gelten, werden dieselben Werte für die Hyperparameter verwendet. Dahingehend wird die Anzahl an Bäumen bei jedem Ordinal Forest und dem Conditional Inference Forest auf 1000 gesetzt. Zusätzlich wird der entsprechende Parameter, der die minimale Anzahl an Beobachtungen in einem Endknoten angibt, einheitlich auf 10 festgelegt (`min.node.size` bei `ordfor` und `minbucket` bei `cforest`). Daneben wird für den Parameter `mtry`, der die Anzahl an zufällig gezogenen Prädiktoren als Kandidaten für eine Aufteilung spezifiziert, der Standardwert, d.h. die Quadratwurzel aus der Prädiktorenanzahl verwendet: $\lfloor\sqrt{p}\rfloor$, wobei $\lfloor\cdot\rfloor$ die nächstkleinere ganze Zahl bezeichnet.

Für eine unverzerrte Konstruktion des Conditional Inference Forest wird in Anlehnung an die Empfehlung von Strobl et al. (2007) der funktionsspezifische Parameter `mincriterion`, der über das `controls`-Argument gesteuert wird, auf Null gesetzt. Dementsprechend muss der p-Wert keinen Schwellenwert überschreiten, damit eine Aufteilung durchgeführt wird. Außerdem wird aus den Analysen von Janitza et al. (2016) die Wahl von `minsplit = 0` übernommen, womit kein Abbruchkriterium bei der Knotenaufteilung angewendet wird.

Schließlich bleiben die restlichen Hyperparameterwerte der Funktion `ordfor` in ihren Standardeinstellungen. In diesem Sinne ist `nsets = 1000` (entspricht B_{sets} in dem

Ordinal Forest-Algorithmus, siehe Abschnitt 3.2.1), `ntreeperdiv = 100` (entspricht $B_{\text{ntreeprior}}$) und `nbest = 10` (entspricht B_{bestsets}).

4.2. Simulationsstudien

4.2.1. Simulierte Daten

Im Folgenden werden drei Simulationsdesigns vorgestellt, die allesamt als Grundlage für die Benchmarkstudie der Random Forest-Varianten hinsichtlich Variablenwichtigkeit und Klassifikation dienen. Die verschiedenen Simulationsdesigns umfassen dasjenige von Janitza et al. (2016), Hornung (2020) und Buri und Hothorn (2020).

4.2.1.1. Simulationsdesign von Janitza et al. (2016)

Die Datenbasis, gestützt auf die Simulationsumgebung von Janitza et al. (2016), bilden simulierte Daten, die aus einer Mischung von zwei unabhängigen Proportional-Odds Modellen gezogen werden. Die kumulative Wahrscheinlichkeit $\mathbb{P}(y \leq j|\boldsymbol{x})$ für das Eintreten einer Klasse des Response, die höchstens der j-ten Klasse entspricht, ergibt sich für ein Individuum mit Prädiktorenvektor $\boldsymbol{x}$ aus einer Mischung von zwei Proportional-Odds Modellen mit Mischungsanteil ζ, wie folgt dargestellt:

$$\mathbb{P}(y \leq j|\boldsymbol{x}) = \zeta\mathbb{P}_1(y \leq j|\boldsymbol{x}) + (1-\zeta)\mathbb{P}_2(y \leq j|\boldsymbol{x}). \tag{4.1}$$

Das entsprechende Modell für die Mischungskomponente $g \in \{1, 2\}$ hat die Form

$$\mathbb{P}_g(y \leq j|\boldsymbol{x}) = \frac{\exp(\gamma_{0jg} + \boldsymbol{x}^T\boldsymbol{\gamma}_g)}{1 + \exp(\gamma_{0jg} + \boldsymbol{x}^T\boldsymbol{\gamma}_g)}, \qquad j = 1, ..., J. \tag{4.2}$$

Die Intercepts γ_{0jg} sind abhängig von j, d.h. sie variieren über die Klassen hinweg, und erfüllen die Bedingung $\gamma_{01g} \leq ... \leq \gamma_{0Jg} = \infty$. Demgegenüber sind die Koeffizienten $\boldsymbol{\gamma}_g$ unabhängig von j, womit die kumulativen Odds von zwei Individuen $\frac{\mathbb{P}_g(y \leq j|\boldsymbol{x})}{\mathbb{P}_g(y > j|\boldsymbol{x})}$ für die Mischungskomponente g proportional zueinander sind. Aus dem Beitrag von Janitza et al. (2016) wird übernommen, dass die Intercepts für beide Mischungskomponenten $g = 1$ und $g = 2$ identisch sind, d.h. die Intercepts $\gamma_{0j1} = \gamma_{0j2} = \gamma_{0j}$ sind unabhängig von g. Die Werte für die Intercepts sind in Tabelle A.1 im Anhang dieser Arbeit angegeben.
Für die Bewertung der Fähigkeit eines VIM zwischen Prädiktoren zu unterscheiden, die mit dem Response assoziiert sind oder keine Assoziation aufweisen, werden 15 Signalprädiktoren und 50 Rauschprädiktoren in die Simulation eingeschlossen.

Dabei unterscheiden sich die Effekte der meisten Prädiktoren zwischen den beiden Mischungskomponenten. Für die erste Mischungskomponente betragen die Effektgrößen $0, 0.5, 0.75$ oder 1, während diese für die zweite Mischungskomponente Werte von $-1, 0$ oder 1 annehmen. Die Tabelle A.2 im Anhang dieser Arbeit enthält die Effektgrößen für alle Prädiktoren und beide Mischungskomponenten.

Darüber hinaus werden unterschiedliche Einstellungen für bestimmte Parameter generiert, auf die im Folgenden eingegangen wird. Der Wert für den Mischungsanteil variiert zwischen $\zeta = 0.6, \zeta = 1$ und $\zeta = 0$. Mit dem Wert von $\zeta = 0.6$ werden Daten basierend auf einer Mischung aus zwei Proportional-Odds Modellen simuliert, während der Wert von $\zeta = 1$ die Daten auf Grundlage des Proportional-Odds Modells, das durch die Mischungskomponente $g = 1$ spezifiziert ist, generiert. Ferner wird $\zeta = 0$ für die Datengenerierung auf Basis des Proportional-Odds Modells, das durch die Mischungskomponente $g = 2$ definiert ist, verwendet. Außerdem wird die Anzahl an Klassen des ordinalen Response auf $J = 3$, $J = 6$ und $J = 9$ gesetzt. Zuletzt generieren Janitza et al. verschiedene Settings mit unabhängigen und korrelierten Prädiktoren. Für den ersten Fall von unabhängigen Prädiktoren werden $\boldsymbol{x}_i, i = 1, ..., n$, aus einer Normalverteilung $N(\mathbf{0}_p, \boldsymbol{I}_p)$ mit Erwartungswertvektor $\mathbf{0}_p$ und Kovarianzmatrix $\boldsymbol{I}_p$, die der Identitätsmatrix mit der Dimension $(p \times p)$ entspricht, gezogen. Demgegenüber werden für den Fall von korrelierten Prädiktoren $\boldsymbol{x}_i, i = 1, ..., n$, aus einer Normalverteilung $N(\mathbf{0}_p, \boldsymbol{\Sigma}_p)$ mit Erwartungswertvektor $\mathbf{0}_p$ und folgender Blockdiagonalmatrix $\boldsymbol{\Sigma}_p$ als Kovarianzmatrix gezogen:

$$\boldsymbol{\Sigma}_p = \begin{bmatrix} \boldsymbol{A}_{signal} & 0 & 0 & 0 & 0 & 0 \\ 0 & \boldsymbol{A}_{noise_1} & 0 & 0 & 0 & 0 \\ 0 & 0 & \boldsymbol{A}_{noise_2} & 0 & 0 & 0 \\ 0 & 0 & 0 & \boldsymbol{A}_{noise_3} & 0 & 0 \\ 0 & 0 & 0 & 0 & \boldsymbol{A}_{noise_4} & 0 \\ 0 & 0 & 0 & 0 & 0 & \boldsymbol{A}_{noise_5} \end{bmatrix}.$$

Mit der Blockmatrix $\boldsymbol{A}_{signal} \in \mathbb{R}^{(15 \times 15)}$ werden die Korrelationen von den 15 Signalprädiktoren beschrieben. Sechs dieser Signalprädiktoren korrelieren mit einer Stärke von 0.8, während die restlichen neun Signalprädiktoren keine Korrelation aufweisen. Die Einträge a_{kl} von $\boldsymbol{A}_{signal}$ sind wie folgt:

$$a_{kl} = \begin{cases} 1, & k = l \\ 0.8, & k \neq l; k, l \in \{1, 3, 6, 8, 11, 13\} \\ 0, & \text{sonst.} \end{cases}$$

Die Matrizen $\boldsymbol{A}_{noise_l} \in \mathbb{R}^{(10\times 10)}$ für $l = 1, ..., 5$ bestimmen die Korrelationen der 10 Rauschprädiktoren mit den Korrelationsstärken $\rho_1 = 0.8, \rho_2 = 0.6, \rho_3 = 0.4$, $\rho_4 = 0.2$ und $\rho_5 = 0$, und sind wie folgt gegeben:

$$\boldsymbol{A}_{noise_l} = \begin{bmatrix} 1 & \rho_l & \cdots & \rho_l \\ \rho_l & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho_l \\ \rho_l & \cdots & \rho_l & 1 \end{bmatrix}.$$

Zusammenfassend ergeben sich aus dem Zusammenspiel von korrelierten oder unabhängigen Prädiktorendaten und Stichprobengröße die folgenden vier Situationen für jeden Wert des Mischungsanteils sowie jede Klassenanzahl:

1) `"korreliert_n200"` ($n = 200$),
2) `"korreliert_n400"` ($n = 400$),
3) `"unabhängig_n200"` ($n = 200$),
4) `"unabhängig_n400"` ($n = 400$).

Für die einzelnen Varianten an verschiedenen Parametereinstellungen (d.h. insgesamt $2 \times 2 \times 3 \times 3 = 36$) erfolgt die Simulation anhand von 100 Trainingsdatensätzen und 100 Testdatensätzen, letztere mit je $n = 10000$.

4.2.1.2. Simulationsdesign von Hornung (2020)

Die Simulationsumgebung von Hornung (2020) verwendet für die Generierung der Prädiktoren die gleichen Einstellungen wie Janitza et al. (2016). Demnach werden 15 Signalprädiktoren und 50 Rauschprädiktoren in die Simulation eingeschlossen. Bezüglich der Generierung des zugrundeliegenden Modells und des Response unterscheiden sich die beiden Simulationsdesigns jedoch voneinander. Während Janitza et al. (2016) eine Mischung aus zwei Proportional-Odds Modellen als Grundlage für ihre Simulation heranziehen, basiert die Datengenerierung bei Hornung (2020) auf dem Modell für Ordinal Forests mit einem latenten stetigen Response, das in Abschnitt 3.2 vorgestellt wurde.
Die Effekte der Prädiktoren stimmen mit den Effekten überein, die von Janitza et al. (2016) für die erste Mischungskomponente $g = 1$ generiert werden. Somit betragen die Effektgrößen der Prädiktoren $0, 0.5, 0.75$ oder 1 (siehe Tabelle A.2 im Anhang). Diese Werte werden für die Berechnung des linearen Prädiktoren und schließlich für die Ermittlung des latenten stetigen Response benötigt. Genauer können die

Werte des latenten stetigen Response erhalten werden, indem zunächst die Werte des linearen Prädiktoren berechnet werden und anschließend Gaußsches Rauschen mit konstanter Varianz (`sigmanoise = 1`) hinzugefügt wird. Die Werte des latenten stetigen Response werden sodann vergröbert, sodass die (optimierten) Klassenwerte des ordinalen Response resultieren.

In Übereinstimmung mit den Simulationen von Janitza et al. (2016) werden die Simulationen von Hornung (2020) ebenso für die Stichprobengrößen $n = 200$ und $n = 400$ durchgeführt sowie für unterschiedliche Einstellungen einiger Parameter, auf die im Folgenden eingegangen wird. Die Intervalle, die den J Klassen des ordinalen Response entsprechen, werden sowohl mit gleicher als auch zufälliger Breite simuliert. Die Intervalle mit gleicher Breite reichen vom 0.001-Quantil bis zum 0.999-Quantil der marginalen (Normal-)Verteilung des latenten stetigen Response. Daneben werden die zufälligen Intervallbreiten in jeder Iteration it der Simulation wie folgt neu generiert. Zunächst werden $J - 1$ Werte von einer gleichverteilten Zufallsvariable gezogen. Ferner werden für die entsprechenden sortierten Werte $d_{it,2}, ..., d_{it,J}$ die (approximativen) Quantilsfunktionen $Q(d_{it,j}) = l_{it,j}, j = 2, ..., J$, der marginalen Verteilung des stetigen Response berechnet. Die Quantilsfunktionen werden durch Stichprobenquantile aus simulierten Datensätzen (je $n = 50.000$) mit korrelierten und unabhängigen niedrigdimensionalen sowie korrelierten hochdimensionalen Prädiktorendaten (siehe unten) approximiert. Die untere Grenze $l_{it,1}$ und die obere Grenze $l_{it,J+1}$ werden auf $-\infty$ und $+\infty$ gesetzt. Schließlich resultieren die Intervalle $]l_{it,1}, l_{it,2}], ...,]l_{it,J}, l_{it,J+1}]$ in jeder Iterationen it der Simulation.

Insgesamt generiert Hornung (2020) unterschiedliche Settings mit korrelierten oder unabhängigen niedrigdimensionalen sowie korrelierten hochdimensionalen Prädiktorendaten. Daraus ergeben sich die folgenden fünf Settings, die sich zusätzlich in der Stichprobengröße unterscheiden. Die Settings werden für jede Klassenanzahl des ordinalen Response von $J = 3$, $J = 6$, $J = 9$ und für jede Art von Intervall simuliert.

1) `"korreliert_n200"` ($n = 200$),
2) `"korreliert_n400"` ($n = 400$),
3) `"unabhängig_n200"` ($n = 200$),
4) `"unabhängig_n400"` ($n = 400$),
5) `"hochdimensional"` ($n = 200$).

Für die Fälle 1) und 2) wird wie bei Janitza et al. (2016) vorgegangen, d.h. sechs einflussreiche Prädiktoren korrelieren mit einer Stärke von 0.8, während die restlichen neun einflussreichen Prädiktoren nicht korrelieren. Außerdem weisen zehn Rauschprädiktoren Korrelationen in den Stärken von $\rho_1 = 0.8,\ \rho_2 = 0.6,\ \rho_3 = 0.4,\ \rho_4 = 0.2$ und $\rho_5 = 0$ auf. Die Fälle 1) und 2) unterscheiden sich lediglich in der Stichprobengröße voneinander. In den Fällen 3) und 4) sind alle 65 Prädiktoren unabhängig. Auch diese Fälle weichen nur bezüglich der Stichprobengröße voneinander ab. Im letzten Fall 5) wird das Korrelationsmuster der ersten beiden Fälle auf hochdimensionale Prädiktorendaten mit 1015 Prädiktoren angewendet. Dieses Szenario wird lediglich mit einer Stichprobengröße von $n = 200$ simuliert.
Für jede Kombination an verschiedenen Parametereinstellungen (d.h. insgesamt $2 \times 3 \times 5 = 30$) werden 100 Trainingsdatensätze erzeugt. Zusätzlich wird für jeden Trainingsdatensatz ein unabhängiger Testdatensatz ($n = 10000$) generiert.

4.2.1.3. Simulationsdesign von Buri und Hothorn (2020)

Die simulierten Daten von Buri und Hothorn (2020) basieren auf einem Transformationsmodell mit bedingter Verteilungsfunktion, das wie folgt dargestellt werden kann:

$$\begin{aligned} \mathbb{P}(y \leq j|\boldsymbol{x}) &= \text{expit}\{\xi(\boldsymbol{x}) \cdot a(j)^T \boldsymbol{\vartheta}(\boldsymbol{x}) + \alpha(\boldsymbol{x})\} \\ &= \frac{\exp\{\xi(\boldsymbol{x}) \cdot a(j)^T \boldsymbol{\vartheta}(\boldsymbol{x}) + \alpha(\boldsymbol{x})\}}{1 + \exp\{\xi(\boldsymbol{x}) \cdot a(j)^T \boldsymbol{\vartheta}(\boldsymbol{x}) + \alpha(\boldsymbol{x})\}}, \quad j = 1, ..., J. \end{aligned} \tag{4.3}$$

Dabei ist $a \in \mathbb{R}^{J-1}$ eine Basisfunktion und $\boldsymbol{\vartheta} = (\vartheta_1, ..., \vartheta_{J-1})^T \in \mathbb{R}^{J-1}$ eine bedingte Parameterfunktion. Der ordinale Response wird mit $J = 4$ geordneten Klassen und $\boldsymbol{\vartheta} = (\vartheta_1, ..., \vartheta_3)^T = \text{logit}(0.15, 0.5, 0.85)^T$ generiert. Die weiteren bedingten Parameterfunktionen ξ (Skalenterm) und α (Verschiebungsterm) sind Funktionen der Prädiktoren $\boldsymbol{x}$, deren Modellierung durch die Friedman-Funktion von Friedman (1991) erfolgt:

$$F(x_1, x_2, x_3, x_4, x_5) = 10 \sin(\pi x_1, x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5.$$

Das Ergebnis der Friedman-Funktion F wird nach Skalierung auf das Intervall $[-2.5, 2.5]$ in das Ergebnis von F^* umbenannt. Die Skalierung auf das genannte Intervall führt dazu, dass die Odds Ratios $\exp(\alpha(\boldsymbol{x})) = \exp(F^*)$ lediglich Werte zwischen 0.08 und 12.18 annehmen können, da $\exp(-2.5) = 0.08$ und $\exp(2.5) = 12.18$.

Mit den beiden Funktionen $\xi(\boldsymbol{x})$ und $\alpha(\boldsymbol{x})$ werden vier verschiedene Effekte der Prädiktoren simuliert. Durch die Präsenz und/oder Absenz von Skalenterm und/oder Verschiebungsterm werden Daten ohne Effekt, mit proportionalen Odds („PO“), nicht-proportionalen Odds („Nicht-PO“) und Daten mit einer Kombination aus proportionalen und nicht-proportionalen Odds („Kombination“) generiert. Die nachfolgende Tabelle 4.1 zeigt die vier Effekte mit den entsprechenden Einstellungen für den Skalenterm $\xi(\boldsymbol{x})$ und Verschiebungsterm $\alpha(\boldsymbol{x})$.

Effekt	$\xi(\boldsymbol{x})$		$\alpha(\boldsymbol{x})$	
Kein Effekt	1	d.h. ohne $\xi(\boldsymbol{x})$	0	d.h. ohne $\alpha(\boldsymbol{x})$
PO	1	d.h. ohne $\xi(\boldsymbol{x})$	$F^*(x_6, ..., x_{10})$	d.h. mit $\alpha(\boldsymbol{x})$
Nicht-PO	$\exp(F^*(x_1, ..., x_5))$	d.h. mit $\xi(\boldsymbol{x})$	0	d.h. ohne $\alpha(\boldsymbol{x})$
Kombination	$\exp(F^*(x_1, ..., x_5))$	d.h. mit $\xi(\boldsymbol{x})$	$F^*(x_6, ..., x_{10})$	d.h. mit $\alpha(\boldsymbol{x})$

Tab. 4.1.: Effekte von den Prädiktoren für das Transformationsmodell in (4.3).

Im Vergleich zu dem Proportional-Odds Modell in Gleichung (4.2), das dem Simulationsdesign von Janitza et al. (2016) zugrunde liegt, wird der lineare Prädiktor $\boldsymbol{x}^T\boldsymbol{\gamma}$ durch eine flexible Funktion der Prädiktoren $\boldsymbol{x}$, weiterhin unter der Annahme von proportionalen Odds, ersetzt.

Zusätzlich simulieren Buri und Hothorn (2020) die vier Arten von Effekten jeweils für niedrig- und hochdimensionale Prädiktorendaten, sodass sich insgesamt acht unterschiedliche Szenarien ergeben. Für den Fall von niedrigdimensionalen Prädiktorendaten werden $p = 15$ unabhängige Prädiktoren, darunter 10 Signal- und 5 Rauschprädiktoren aus einer Gleichverteilung, entsprechend $\boldsymbol{x} = (x_1, ..., x_p) \sim U[0, 1]^p$ generiert. Für den Fall von hochdimensionalen Prädiktorendaten werden $p = 60$ unabhängige gleichverteilte Prädiktoren simuliert, davon 10 Signal- und 50 Rauschprädiktoren.
Jede der acht Varianten von verschiedenen Parametereinstellungen wird für 100 Trainingsdatensätze (je $n = 250$) sowie Testdatensätze (je $n = 500$) wiederholt.

4.2.2. Bewertung der Variablenwichtigkeit

Für die Bewertung der Qualität des Prädiktoren-Rankings, das durch die Berechnung eines VIM gewonnen wird, wird die „Area under the curve“ (kurz AUC) herangezogen. Das AUC beurteilt hier die Fähigkeit eines Ordinal Forest oder Conditional Inference Forest, zwischen Signalprädiktoren und Rauschprädiktoren richtig zu dif-

ferenzieren und basiert auf folgender Berechnung (Hanley und McNeil, 1982):

$$AUC = \frac{1}{|M_0||M_1|} \sum_{j \in M_0} \sum_{i \in M_1} I(VI_j < VI_i) + 0.5 I(VI_j = VI_i). \tag{4.4}$$

Dabei stellen M_0 und M_1 zwei disjunkte Mengen der Prädiktoren dar, d.h. $M = M_0 \cup M_1$. Während $|M_0|$ die Mächtigkeit der Menge an Rauschprädiktoren bezeichnet, beschreibt $|M_1|$ die Mächtigkeit der Menge an Signalprädiktoren. Ferner ist VI_j die Variablenwichtigkeit des j-ten Rauschprädiktoren und VI_i die Wichtigkeit des i-ten Signalprädiktoren. Die VI-Werte können für einen Ordinal Forest und Conditional Inference Forest mit der Gleichung (3.4) berechnet werden.
Das AUC schätzt die Wahrscheinlichkeit, dass einem zufällig gezogenen Signalprädiktor ein höherer Wert für die Wichtigkeit zugewiesen wird als einem zufällig gezogenen Rauschprädiktor. Wenn das AUC einen Wert von 1 annimmt, hat jeder Signalprädiktor eine höhere Wichtigkeit als jeder Rauschprädiktor, d.h. das VIM unterscheidet perfekt zwischen Prädiktoren mit und ohne Effekt auf den Response. Demgegenüber bedeutet ein AUC von 0.5, dass nur in der Hälfte der Fälle ein zufällig gezogener Signalprädiktor eine höhere Wichtigkeit hat als ein zufällig gezogener Rauschprädiktor, d.h. das VIM weist keine Fähigkeit zur Unterscheidung von Prädiktoren mit und ohne Effekt auf. Zusammenfassend bedeuten höhere AUC-Werte eine bessere Leistung des VIM bezüglich der Identifikation relevanter Prädiktoren.

4.2.3. Bewertung der Klassifikationsgüte

Das Ziel, die Klassifikationsgüte eines Random Forest zu bewerten, verfolgten bereits Janitza et al. (2016) in ihrem Beitrag, der in der Einleitung kurz vorgestellt wurde. In dem Beitrag ist der Vergleich der Klassifikationsgüte von einem Conditional Inference Forest mit ordinalem Response und einem Conditional Inference Forest mit nominalem Response anhand des RPS-Ratio zentral. Das RPS-Ratio beschreibt das Verhältnis des Ranked Probability Score für den Forest mit ordinalem Response zu dem Forest mit nominalem Response. Anders als bei Janitza et al. (2016) erfolgt in dieser Arbeit die Bewertung der Klassifikationsgüte anhand des gewichteten Kappa mit quadratischen Gewichten und des gewichteten Kappa mit linearen Gewichten (Cohen, 1968) sowie durch Cohen's Kappa (Cohen, 1960) von Random Forests, die einen ordinalskalierten Response vorhersagen. Die Auswahl dieser Gütemaße ist durch die Arbeit von Hornung (2020) motiviert. Der zentrale Unterschied zwischen dem RPS-Ratio und dem (gewichteten) Kappa besteht darin, dass Ersteres als Gütemaß für Wahrscheinlichkeitsvorhersagen und Letzteres als Gütemaß für Punktvorhersagen gilt. Da nicht mit allen Methoden, die Gegenstand der

Benchmarkstudie in Kapitel 4 sind, Wahrscheinlichkeitsvorhersagen erhalten werden können, wird die Qualität der Punktvorhersagen aller Methoden miteinander verglichen, wofür auf das (gewichtete) Kappa zurückgegriffen wird. Allgemein ist das gewichtete Kappa wie folgt definiert (Cohen, 1968):

$$\kappa_w := \frac{\sum_{i=1}^{J}\sum_{j=1}^{J} w_{ij}p_{oij} - \sum_{i=1}^{J}\sum_{j=1}^{J} w_{ij}p_{cij}}{1 - \sum_{i=1}^{J}\sum_{j=1}^{J} w_{ij}p_{cij}}. \tag{4.5}$$

Dabei bezeichnet p_{oij} den beobachteten Anteil von Fällen mit i als wahre Klasse und j als vorhergesagte Klasse und p_{cij} den zufällig erwarteten Anteil von Fällen mit i als wahre Klasse und j als vorhergesagte Klasse. Ferner werden über w_{ij} die sogenannten Übereinstimmungsgewichte definiert, die in dieser Arbeit eine lineare oder quadratische Form annehmen können. Für den Fall des linear gewichteten Kappa ist $w_{ij} = 1 - \frac{|i-j|}{J-1}$ (Cicchetti und Allison, 1971), während für den Fall des quadratisch gewichteten Kappa die Übereinstimmungsgewichte durch $w_{ij} = 1 - \frac{|i-j|^2}{(J-1)^2}$ (Fleiss und Cohen, 1973) gegeben sind. Darüber hinaus ist für Cohen's Kappa $w_{ij} = 1$, insofern die wahre und vorhergesagte Klasse identisch ist (d.h. wenn $i = j$), andernfalls ist $w_{ij} = 0$ (d.h. wenn $i \neq j$). Demnach wird bei Verwendung von Cohen's Kappa lediglich denjenigen Vorhersagen einen Nutzen zuerkannt, die den wahren Klassenwerten entsprechen, und ansonsten keinen Nutzen. In Kontrast dazu weist das gewichtete Kappa allen Vorhersagen einen bestimmten Nutzen zu, womit es ein geeignetes Maß für die Bewertung von Vorhersagen ordinaler Responses ist, da diese nicht notwendigerweise mit den wahren Klassenwerten auf der Ordinalskala übereinstimmen müssen, um ihnen einen Nutzen zuschreiben zu können (Ben-David, 2008). Der Nutzen von den Vorhersagen, die weiter von den wahren Klassenwerten entfernt liegen, ist für den Fall von quadratischen Gewichten größer als für den Fall von linearen Gewichten. Gleichzeitig ist der Nutzen von den Vorhersagen, die (annähernd) gleich zu den wahren Klassenwerten sind, für den Fall von quadratischen Gewichten geringer als für den Fall von linearen Gewichten. Zusammenfassend stellt das linear gewichtete Kappa einen Kompromiss zwischen dem quadratisch gewichteten Kappa und Cohen's Kappa dar.

Die drei eingeführten Kappa-Varianten ergeben Werte zwischen 0 und 1, wobei mit einem höheren Wert eine bessere Vorhersage einer Methode bezüglich der Leistung der verwendeten Kappa-Variante indiziert wird.
Die unterschiedlichen Kappa-Varianten können sich beispielsweise folgendermaßen verhalten: Cohen's Kappa nimmt einen hohen Wert und das gewichtete Kappa einen geringen Wert bei gleichen Daten an, wenn eine Vorhersagemethode meistens exakt

klassifiziert und selten Vorhersagen trifft, die von den wahren Klassenwerten weit entfernt sind.

Laut Jakobsson und Westergren (2005) sind die drei vorgestellten Maße abhängig von den Klassengrößen und der Klassenanzahl. Dies muss für die Vergleichsstudie jedoch nicht berücksichtigt werden, da die Vergleiche jeweils auf den gleichen simulierten und realen Daten durchgeführt werden.

4.2.4. Ergebnisse

4.2.4.1. Ergebnisse zur Variablenwichtigkeit

In diesem Abschnitt werden das ER-VIM des CIF (Conditional Inference Forest) und OF (Ordinal Forest) mit Equal-Perff sowie das RPS-VIM des CIF, OF mit Equal-Perff und OF mit RPS-Perff in Bezug auf deren Qualität bei der Unterscheidung zwischen Signal- und Rauschprädiktoren verglichen. Die Abbildungen 4.1 - 4.6 zeigen die AUC-Werte, berechnet aus den VI-Werten der Prädiktoren, für alle Methoden und simulierten Datensätze.

AUC-Werte für die Simulationen von Janitza et al.

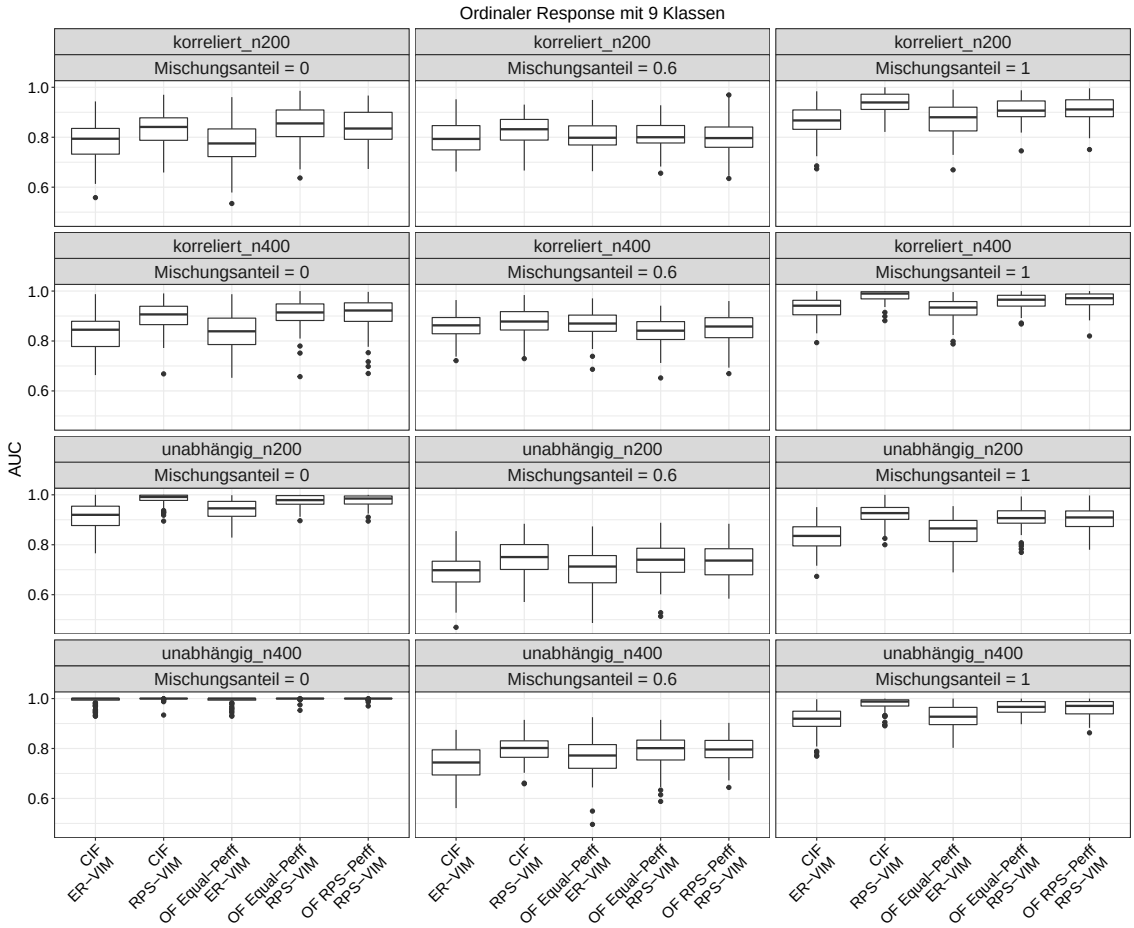

Abb. 4.1.: AUC-Werte zur Beurteilung der Leistung von VIMs für alle betrachteten Methoden in jedem Setting für einen ordinalen Response mit 9 Klassen basierend auf den Simulationen von Janitza et al. (2016). Die Boxplots zeigen die Werte für die 100 Iterationen.

Die Boxplots in den Abbildungen 4.1 - 4.3 zeigen die AUC-Werte für das ER-VIM und RPS-VIM von den betrachteten Methoden, gestützt auf die Simulationsumgebung von Janitza et al. (2016). Die Abbildung 4.1 stellt die Resultate für 9 Klassen,

die Abbildung 4.2 für 6 Klassen und die Abbildung 4.3 für 3 Klassen des ordinalen Response dar. Zunächst liegt der Fokus auf der Beschreibung von den Ergebnissen für 9 Klassen und danach wird auf wichtige Unterschiede in den Ergebnissen für 6 und 3 Klassen aufmerksam gemacht. Die Spalten von allen Abbildungen unterscheiden zwischen einem Mischungsanteil von $0, 0.6$ und 1. Ferner differenzieren die Zeilen zwischen korrelierten und unabhängigen Prädiktorendaten sowie der Stichprobengröße. So bilden die ersten beiden Zeilen die Ergebnisse für die Simulationen von korrelierten Prädiktorendaten und einer Stichprobengröße von $n = 200$ (erste Zeile) sowie einer Stichprobengröße von $n = 400$ (zweite Zeile) ab. Daneben präsentieren die letzten beiden Zeilen die AUC-Werte für die Simulationen von unabhängigen Prädiktorendaten und erneut den Stichprobengrößen $n = 200$ (dritte Zeile) sowie $n = 400$ (vierte Zeile).

Die Abbildung 4.1 zeigt für die meisten Simulationen von 9 Klassen, dass das RPS-VIM von den Methoden CIF, OF mit Equal-Perff und OF mit RPS-Perff sehr ähnlich ist und gleichzeitig bessere Leistungen als das ER-VIM von dem CIF und OF mit Equal-Perff erzielt. Die Ausnahmen sind die folgenden Einstellungen. Bei unabhängigen Prädiktorendaten, $n = 400$ und einem Mischungsanteil von $\zeta = 0$ weisen das RPS-VIM und ER-VIM von allen Methoden besonders ähnliche AUC-Werte auf. Daneben schneidet bei korrelierten Prädiktorendaten, $n = 200$ und $\zeta = 0.6$ das RPS-VIM von dem OF mit RPS-Perff minimal schlechter als das ER-VIM von dem OF mit Equal-Perff ab, und bei korrelierten Prädiktorendaten, $n = 400$ und $\zeta = 0.6$ ist das RPS-VIM des OF mit Equal-Perff am schlechtesten, gefolgt von dem RPS-VIM des OF mit RPS-Perff.
In den meisten Fällen schneidet das RPS-VIM von dem CIF am besten ab. In den wenigen Ausnahmen, wo dies nicht der Fall ist, zeigen der OF mit RPS-Perff und RPS-VIM („korreliert_n400“, $\zeta = 0$) oder der OF mit Equal-Perff und RPS-VIM („korreliert_n200“, $\zeta = 0$) die besten Leistungen. Insgesamt sind die Unterschiede zwischen den beiden OF-Varianten mit RPS-VIM jedoch gering und inkonsistent.
Die Tatsache, dass das ER-VIM in fast allen Einstellungen für 9 Klassen am schwächsten ist, legt den Schluss nahe, dass die Ordnung in den Klassen – für den Fall von 9 Klassen – bei der Berechnung des VIM einbezogen werden sollte, um genauere Prädiktoren-Rankings zu erhalten.
Darüber hinaus weist das ER-VIM des OF mit Equal-Perff in mehr als der Hälfte aller Einstellungen etwas bessere Ergebnisse als das ER-VIM des CIF auf.
In allen Einstellungen für 9 Klassen fallen die AUC-Werte für die Stichprobengröße von $n = 200$ geringer als für $n = 400$ aus. Bis auf die oben dargelegten Ausnahmen

gibt es jedoch keine Unterschiede in den Ergebnissen für die beiden Stichprobengrößen in Bezug auf die Leistungen der Methoden relativ zueinander.

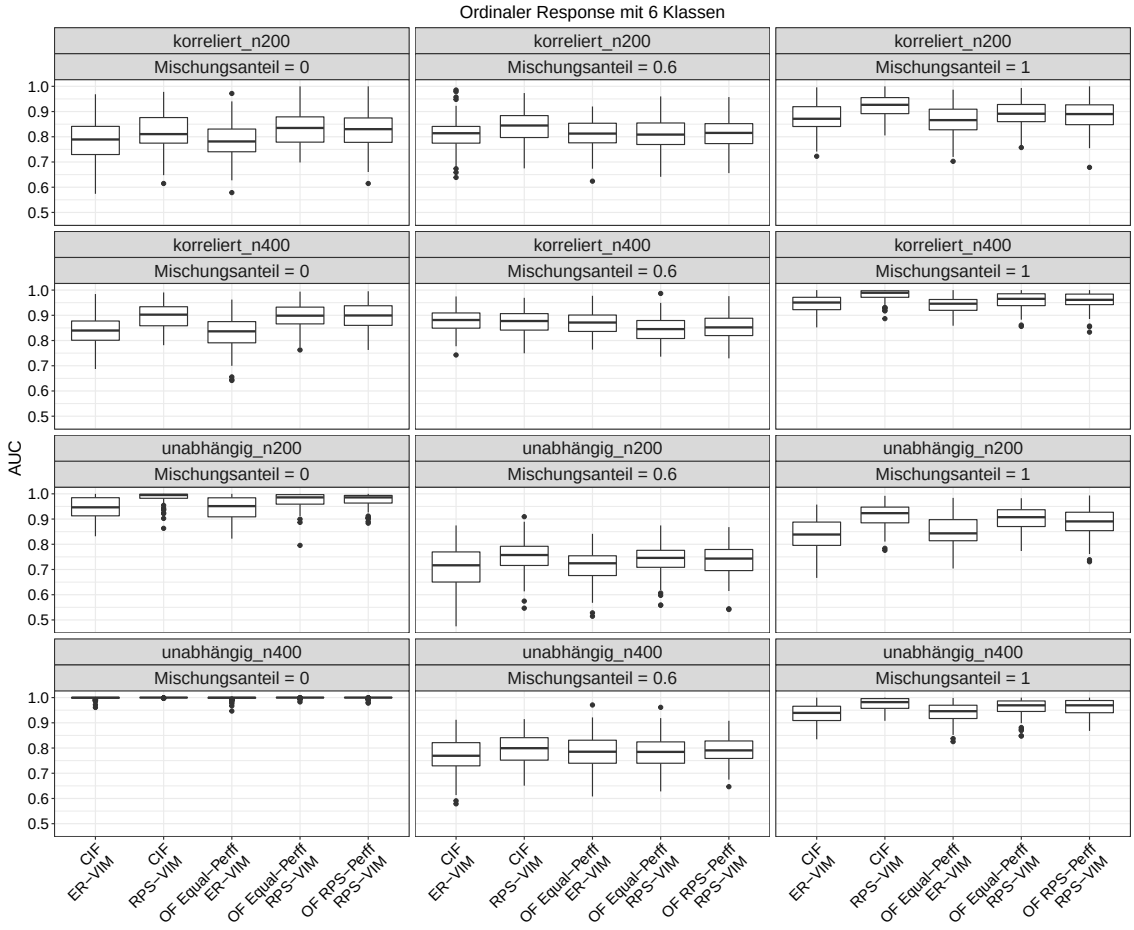

Abb. 4.2.: AUC-Werte zur Beurteilung der Leistung von VIMs für alle betrachteten Methoden in jedem Setting für einen ordinalen Response mit 6 Klassen basierend auf den Simulationen von Janitza et al. (2016). Die Boxplots zeigen die Werte für die 100 Iterationen.

Die Ergebnisse für die Simulationen von 6 Klassen in Abbildung 4.2 sind sehr ähnlich zu den Ergebnissen für die Simulationen von 9 Klassen. Allerdings nehmen in den meisten Einstellungen für 6 Klassen die Unterschiede zwischen allen fünf Methoden ab. Anders als für die Simulationen von 9 Klassen ist in nahezu allen Simulationen von 6 Klassen das ER-VIM von dem CIF und OF stark übereinstimmend und beide Methoden abwechselnd am schlechtesten.

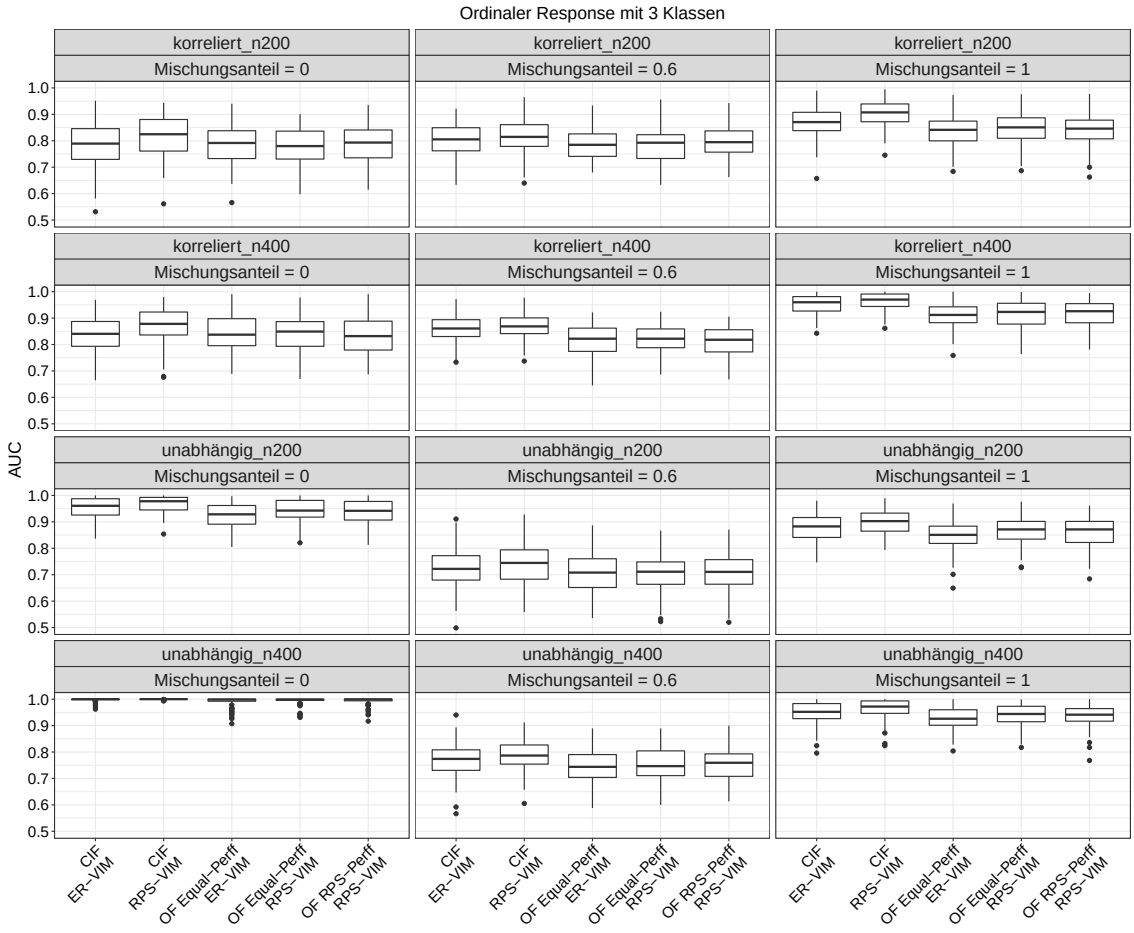

Abb. 4.3.: AUC-Werte zur Beurteilung der Leistung von VIMs für alle betrachteten Methoden in jedem Setting für einen ordinalen Response mit 3 Klassen basierend auf den Simulationen von Janitza et al. (2016). Die Boxplots zeigen die Werte für die 100 Iterationen.

Die Ergebnisse für die Simulationen von 3 Klassen in Abbildung 4.3 zeigen, dass die Abweichungen nochmals insbesondere zwischen dem ER-VIM und RPS-VIM von den verwendeten OF-Varianten abnehmen und zusätzlich dessen Leistungen schlechter werden, verglichen mit den Leistungen des ER-VIM und RPS-VIM von dem CIF. Somit übertreffen in nahezu allen Einstellungen die beiden VIMs von dem CIF, vor allem das RPS-VIM, die Leistungen der beiden VIMs von den OF-Varianten. Allerdings ist dabei auffällig, dass in der Hälfte dieser Einstellungen insbesondere das RPS-VIM von den OF-Varianten stark vergleichbare Resultate wie das ER-VIM von dem CIF erzielt.

Für das Simulationsdesign von Janitza et al. (2016) kann zusammengefasst werden, dass das RPS-VIM von allen verwendeten Methoden ähnliche Ergebnisse erzielt. Mit abnehmender Klassenanzahl nehmen zusätzlich die Ähnlichkeiten zwischen dem ER-VIM und RPS-VIM von den OF-Varianten zu.

Für die meisten Einstellungen von 6 und 9 Klassen schneidet das RPS-VIM von allen verwendeten Methoden, am häufigsten von dem CIF, am besten ab. Währenddessen sind für die meisten Einstellungen von 3 Klassen sowohl das RPS-VIM als auch das ER-VIM von den OF-Varianten am schlechtesten.
Darüber hinaus zeigt der Vergleich von beiden OF-Varianten mit RPS-VIM, dass die Unterschiede zwischen den beiden Methoden für jede Klassenanzahl geringfügig und inkonsistent sind.
Überdies sind in der Gegenüberstellung von beiden Methoden mit ER-VIM die Leistungen des OF gegenüber den Leistungen des CIF für den Fall von 9 Klassen minimal besser, für den Fall von 6 Klassen nahezu gleich und für den Fall von 3 Klassen eindeutig schlechter.
Schließlich sind die beiden neuen OF-Varianten mit RPS-VIM in den meisten Einstellungen für jede Klassenanzahl leistungsfähiger als die bestehende OF-Variante mit ER-VIM in Bezug auf die Qualität des Prädiktoren-Rankings.

AUC-Werte für die Simulationen von Hornung

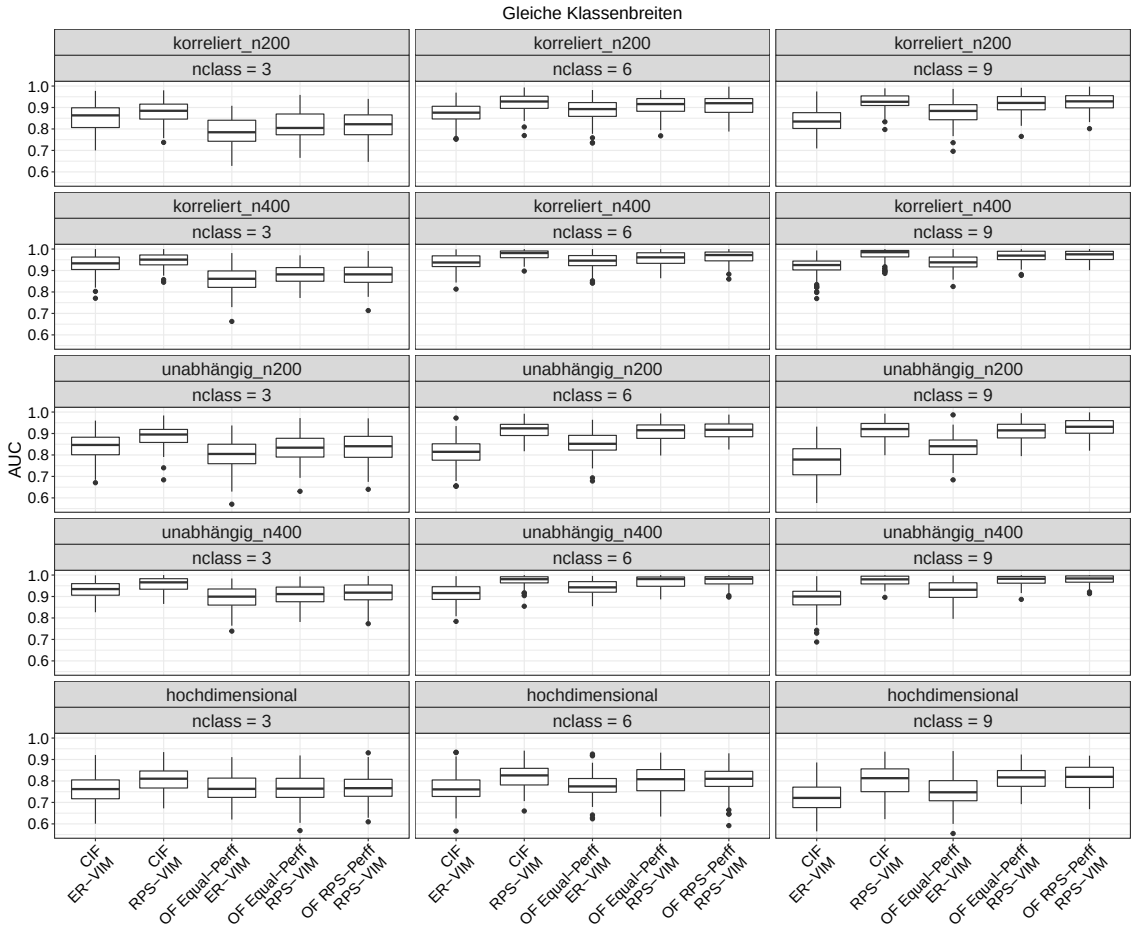

Abb. 4.4.: AUC-Werte zur Beurteilung der Leistung von VIMs für alle betrachteten Methoden in jedem Setting mit gleichen Klassenbreiten basierend auf den Simulationen von Hornung (2020). Die Boxplots zeigen die Werte für die 100 Iterationen.

Die Boxplots in den Abbildungen 4.4 und 4.5 zeigen die AUC-Werte, welche die Qualität des Prädiktoren-Rankings durch das ER-VIM und RPS-VIM von den verwendeten Methoden basierend auf den simulierten Daten von Hornung (2020) widerspiegeln. Beide Abbildungen unterscheiden zwischen gleichen Klassenbreiten (Abb. 4.4) und zufälligen Klassenbreiten (Abb. 4.5). Der Schwerpunkt liegt zunächst auf der Beschreibung der Ergebnisse für gleiche Klassenbreiten. Danach werden zentrale Abweichungen in den Ergebnissen für zufällige Klassenbreiten herausgearbeitet. In den Abbildungen teilen sich die Spalten nach der Anzahl an Klassen auf, während die Zeilen in korrelierte, unabhängige und hochdimensionale Prädiktorendaten sowie gleichzeitig in die Stichprobengröße unterteilen. Es wurden Daten mit den Stichprobengrößen von $n = 200$ (erste, dritte und fünfte Zeile) und $n = 400$ (zweite und vierte Zeile) simuliert.

In den Simulationen von gleichen Klassenbreiten (Abb. 4.4) ist für den Fall von 6 und 9 Klassen das RPS-VIM von dem CIF und den beiden OF-Varianten sehr ähnlich und durchweg besser als das ER-VIM von dem CIF und OF. Währenddessen weisen für den Fall von 3 Klassen das RPS-VIM von dem CIF eindeutig und das ER-VIM von dem CIF minimal bessere Leistungen als die VIMs von den OF-Varianten auf mit Ausnahme der Einstellung von hochdimensionalen Prädiktorendaten, wo das RPS-VIM von den beiden OF-Varianten höhere AUC-Werte als das ER-VIM von dem CIF aufzeigt. Demnach ist ein Trend in Bezug auf die Klassenanzahl des ordinalen Response erkennbar, dahingehend, dass mit abnehmender Klassenanzahl das RPS-VIM von den beiden OF-Varianten stetig schlechter als das RPS-VIM von dem CIF abschneidet. Umgekehrt erzielt das RPS-VIM von dem OF mit RPS-Perff in fast allen Einstellungen von 9 Klassen die höchsten AUC-Werte. Eine Ausnahme ist die Einstellung von korrelierten Prädiktorendaten und $n = 400$, wo das RPS-VIM von dem OF mit RPS-Perff minimal schlechter als von dem CIF abschneidet.
Überdies sind für jede Klassenanzahl die Leistungen des RPS-VIM von dem OF mit RPS-Perff besser als von dem OF mit Equal-Perff mit Ausnahme von einer Einstellung („korreliert_n400“, „nclass = 3“), in der die Leistungen besonders ähnlich sind. Des Weiteren zeigt der Vergleich des ER-VIM zwischen CIF und OF mit Equal-Perff eindeutig bessere Leistungen seitens des OF für den Fall von 9 Klassen, minimal bessere Leistungen für den Fall von 6 Klassen und demgegenüber eindeutig schlechtere Leistungen für den Fall von 3 Klassen.
In allen Einstellungen sind höhere AUC-Werte für die größere Stichprobe gegenüber der kleineren Stichprobe zu erkennen. Abgesehen davon lassen die wenigen und geringfügigen Unterschiede in den Ergebnissen für die beiden Stichprobengrößen auf keinen eindeutigen Trend schließen, d.h. für beide Stichprobengrößen können die gleichen Rückschlüsse gezogen werden.

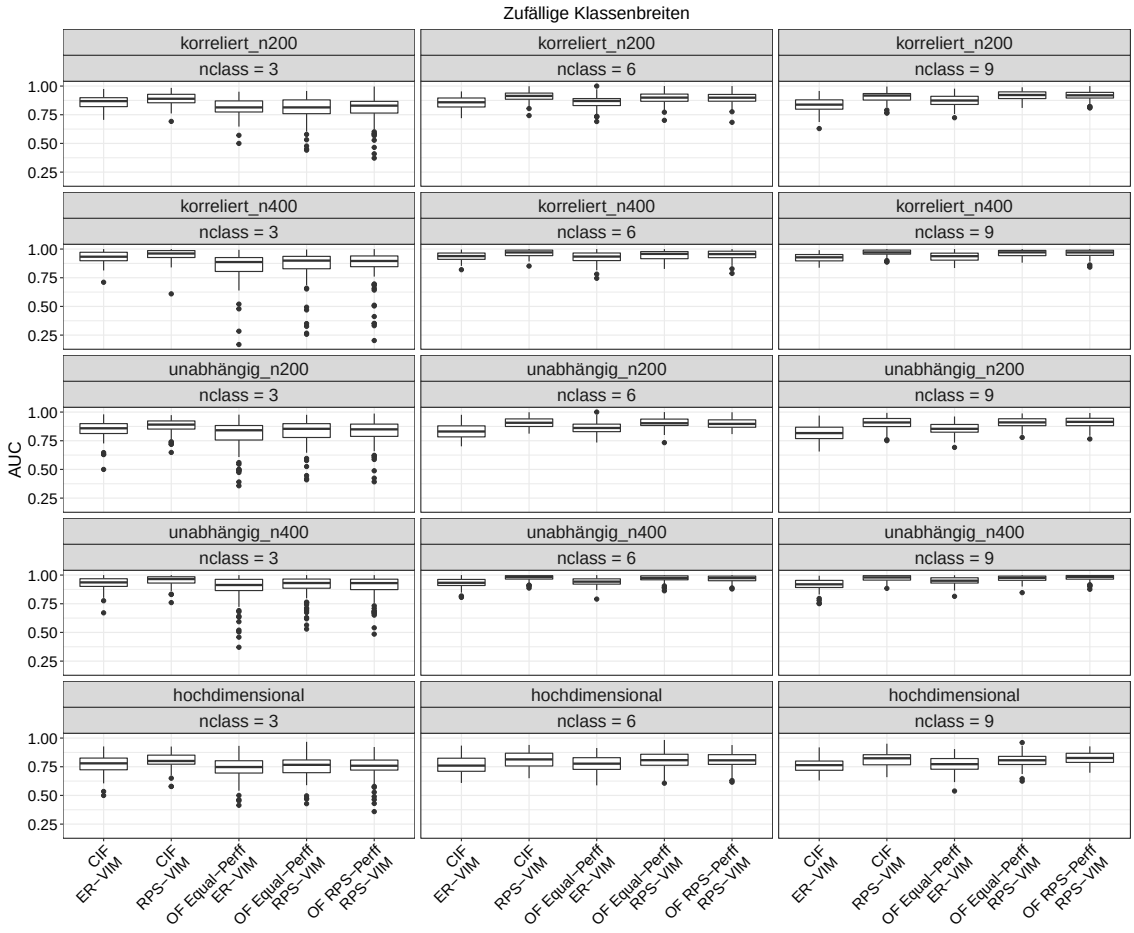

Abb. 4.5.: AUC-Werte zur Beurteilung der Leistung von VIMs für alle betrachteten Methoden in jedem Setting mit zufälligen Klassenbreiten basierend auf den Simulationen von Hornung (2020). Die Boxplots zeigen die Werte für die 100 Iterationen.

Die Simulationen von zufälligen Klassenbreiten (Abb. 4.5) zeigen wie die vorherigen Simulationen von gleichen Klassenbreiten, dass das RPS-VIM von allen verwendeten Methoden sehr ähnlich ist und für alle Einstellungen von 6 und 9 Klassen besser als das ER-VIM von den entsprechenden Methoden abschneidet. Außerdem weist der OF mit RPS-Perff und RPS-VIM weiterhin in den meisten Einstellungen von 9 Klassen die besten Prädiktoren-Rankings auf, jedoch nun mit zwei Ausnahmen anstatt einer Ausnahme. Die Ausnahmen für 9 Klassen sind die beiden Einstellungen von korrelierten Prädiktorendaten für beide Stichprobengrößen, für die der OF mit Equal-Perff und RPS-VIM am besten abschneidet. Daneben bleibt unverändert, dass für alle Simulationen von 3 Klassen der CIF, unabhängig von dem VIM, bessere Leistungen als alle OF-Versionen erzielt.

In den meisten Einstellungen für zufällige Klassenbreiten erreicht das RPS-VIM von dem OF mit Equal-Perff etwas höhere AUC-Werte als das gleiche VIM von dem OF mit RPS-Perff. Dies steht im Gegensatz zu den Einstellungen von gleichen Klas-

senbreiten, wo sich das RPS-VIM von beiden OF-Varianten auf umgekehrte Weise verhält. Dennoch sind sowohl für gleiche als auch zufällige Klassenbreiten beide OF-Varianten mit RPS-VIM sehr ähnlich.
Schließlich bleibt die Beziehung des ER-VIM zwischen CIF und OF mit Equal-Perff im Vergleich zu den Simulationen von gleichen Klassenbreiten unverändert.

Für das Simulationsdesign von Hornung (2020) kann zusammengefasst werden, dass beide OF-Varianten mit RPS-VIM in allen Einstellungen für gleiche sowie zufällige Klassenbreiten sehr ähnlich sind. Während dennoch der OF mit RPS-Perff und RPS-VIM in allen Einstellungen für gleiche Klassenbreiten minimal bessere Ergebnisse als der OF mit Equal-Perff und gleichem VIM erzielt, ist in den meisten Einstellungen für zufällige Klassenbreiten letztere OF-Variante knapp besser. Außerdem sind in den Einstellungen von 6 und 9 Klassen für beide Arten von Klassenbreiten die starken Ähnlichkeiten zwischen CIF mit RPS-VIM und den beiden OF-Varianten mit RPS-VIM auffällig, die jedoch in den Einstellungen von 3 Klassen deutlich geringer ausfallen.
Schließlich sind die beiden neuen OF-Varianten mit RPS-VIM in nahezu allen Einstellungen für beide Arten von Klassenbreiten leistungsfähiger als die bestehende OF-Variante mit ER-VIM.

AUC-Werte für die Simulationen von Buri und Hothorn

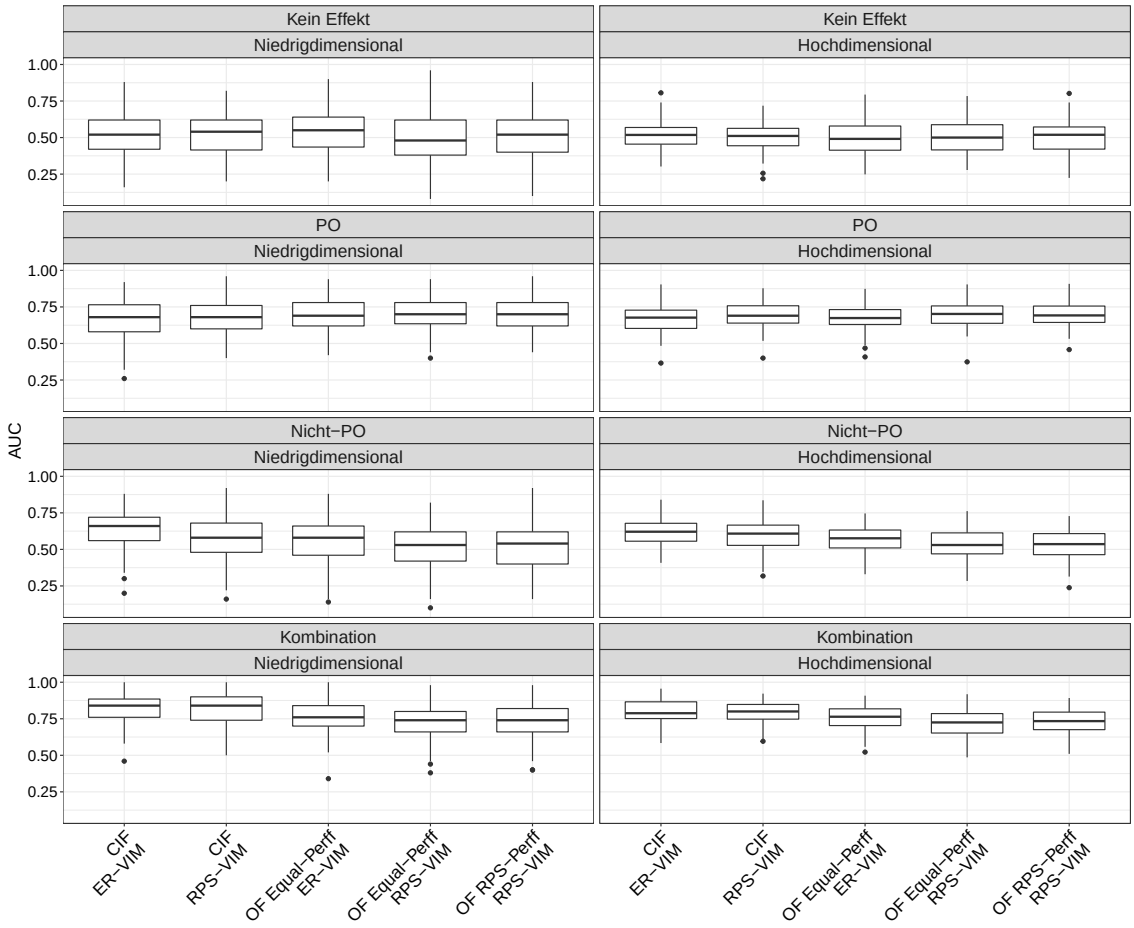

Abb. 4.6.: AUC-Werte zur Beurteilung der Leistung von VIMs für alle betrachteten Methoden in jedem Setting basierend auf den Simulationen von Buri und Hothorn (2020). Die Boxplots zeigen die Werte für die 100 Iterationen.

Die Boxplots in Abbildung 4.6 zeigen die AUC-Werte für das ER-VIM und RPS-VIM von den betrachteten Methoden basierend auf dem Simulationsdesign von Buri und Hothorn (2020). Die erste Spalte in der Abbildung zeigt die Ergebnisse für die niedrigdimensionalen Prädiktorendaten (mit 5 Rauschprädiktoren) und die zweite Spalte die Ergebnisse für die hochdimensionalen Prädiktorendaten (mit 50 Rauschprädiktoren). Daneben unterteilen die Zeilen in die vier verschiedenen Effekttypen (Kein Effekt; PO, d.h. proportionale Odds; Nicht-PO, d.h. nicht-proportionale Odds; Kombination aus PO und Nicht-PO).

Die Unterschiede zwischen den Ergebnissen zur Qualität der Prädiktoren-Rankings sind für die unterschiedlichen Methoden in den meisten Fällen sehr gering.
Wie zu erwarten, schwanken die AUC-Werte in den Settings ohne Effekt um 0.5, d.h. Prädiktoren ohne Effekt werden von den VIMs korrekterweise nicht identifiziert.

In den Settings mit proportionalen Odds (PO) schneidet das RPS-VIM von beiden OF-Varianten sowohl für die niedrig- als auch die hochdimensionalen Prädiktorendaten gleichmäßig am besten ab, verglichen mit der Konkurrenz. Allerdings bestehen zu den anderen Methoden kaum Unterschiede.
In Kontrast zu den Einstellungen mit PO erzielt in den Einstellungen mit Nicht-PO das RPS-VIM von beiden OF-Varianten die schlechtesten Leistungen, während das ER-VIM von dem CIF die besten Leistungen erreicht. Dabei ist das RPS-VIM von dem OF mit RPS-Perff minimal besser als von dem OF mit Equal-Perff.
Entsprechend der Ergebnisse für die Einstellungen von Nicht-PO weist in den Einstellungen von der Kombination aus PO und Nicht-PO das RPS-VIM von beiden OF-Varianten die schlechtesten Prädiktoren-Rankings auf, gefolgt von dem ER-VIM des OF mit Equal-Perff. Außerdem fällt bei der Kombination für die niedrigdimensionalen Prädiktorendaten auf, dass die Ergebnisse für die beiden CIF-Methoden deutlich besser sind als für die restlichen Methoden.

Für das Simulationsdesign von Buri und Hothorn (2020) kann zusammengefasst werden, dass sich die Methoden OF mit Equal-Perff und OF mit RPS-Perff, jeweils mit RPS-VIM, in allen Einstellungen sehr ähnlich sind. Ebenso stark vergleichbare Leistungen werden in den meisten Einstellungen für die beiden VIMs von dem CIF erzielt.
Schließlich sind die neuen OF-Varianten mit RPS-VIM lediglich in den Einstellungen von PO leistungsfähiger als die bestehende OF-Variante mit ER-VIM. Daneben ist die bestehende OF-Variante mit ER-VIM in fast allen Einstellungen weniger leistungsfähig als der CIF, unabhängig von dem verwendeten VIM. Die beiden Ausnahmen sind die Einstellungen von PO und Nicht-PO, jeweils für die niedrigdimensionalen Prädiktorendaten, denn für die entsprechende Einstellung von PO schneidet der OF mit Equal-Perff und ER-VIM etwas besser als der CIF sowohl mit ER-VIM als auch RPS-VIM ab, und für die jeweilige Einstellung von Nicht-PO erzielt der OF mit Equal-Perff und ER-VIM sehr ähnliche Werte wie der CIF mit RPS-VIM.

Vergleich von allen Ergebnissen bezüglich der Variablenwichtigkeit

Im Folgenden werden die Ergebnisse in Bezug auf die Qualität des Prädiktoren-Rankings für alle Simulationsdesigns verglichen, um relevante Unterschiede sowie Übereinstimmungen hervorzuheben.
Die Simulationsdesigns von Janitza et al. (2016) und Hornung (2020) teilen die Gemeinsamkeit, dass für die meisten Einstellungen von 6 und 9 Klassen das RPS-VIM, verglichen mit dem ER-VIM, von den verwendeten Methoden die besten Leistungen

erzielt, die allesamt ähnlich sind. Demnach ist für diese Einstellungen die Berücksichtigung der Ordnung bei Berechnung der Variablenwichtigkeit durchaus vorteilhaft. Demgegenüber ist in beiden Simulationsdesigns für die meisten Einstellungen von 3 Klassen der CIF mit den zwei VIMs besser als die Konkurrenz. Zu diesem Ergebnis kommt ebenso das Simulationsdesign von Buri und Hothorn (2020), das ausschließlich einen ordinalen Response mit 4 Klassen simuliert.
Während in den Simulationen von Janitza et al. der CIF mit RPS-VIM für jede Klassenanzahl häufiger minimal besser als die konkurrierenden Methoden abschneidet, ist dies in den Simulationen von Hornung für 3 und 6 Klassen der Fall. Im Gegensatz dazu ermöglicht in den Simulationen von Buri und Hothorn der CIF mit ER-VIM öfter geringfügig bessere Prädiktoren-Rankings als die Konkurrenz.
Zusammenfassend liefert der CIF für alle Simulationsdesigns die besten Prädiktoren-Rankings.

In allen drei Simulationsdesigns sind sich die beiden neuen OF-Varianten mit RPS-VIM sehr ähnlich und für die Simulationsdesigns von Janitza et al. und Hornung in nahezu allen Einstellungen leistungsfähiger als die bestehende OF-Variante mit ER-VIM. In Bezug auf das Simulationsdesign von Buri und Hothorn trifft dies lediglich für wenige Einstellungen zu.

Überdies stimmen die Tendenzen zwischen den beiden Methoden mit ER-VIM für die Simulationsdesigns von Janitza et al. und Hornung größtenteils überein. Demnach ist die Leistung des OF mit ER-VIM gegenüber der Leistung des CIF mit gleichem VIM für den Fall von 9 Klassen etwas besser, für den Fall von 6 Klassen nahezu gleich und für den Fall von 3 Klassen eindeutig schlechter. Letzteres gilt auch überwiegend für das Simulationsdesign von Buri und Hothorn, das einen ordinalen Response mit 4 Klassen simuliert.

Im Hinblick auf die erzielten Ergebnisse muss allerdings beachtet werden, dass die Unterschiede häufig nur sehr schwach sind und auch Zufallsschwankungen geschuldet sein können, womit ihre praktische Relevanz teilweise fraglich ist.

4.2.4.2. Ergebnisse zur Klassifikationsgüte

In diesem Abschnitt werden die Werte für das linear gewichtete Kappa, quadratisch gewichtete Kappa und Cohen's Kappa von den Methoden CIF, OF mit Equal-Perff und OF mit RPS-Perff verglichen. Die Abbildungen 4.7 - 4.12 in der Hauptarbeit sowie die ergänzenden Abbildungen A.1 - A.12 im Anhang zeigen die (gewichteten) Kappa-Werte für die betrachteten Methoden und simulierten Datensätze.

Kappa-Werte für die Simulationen von Janitza et al.

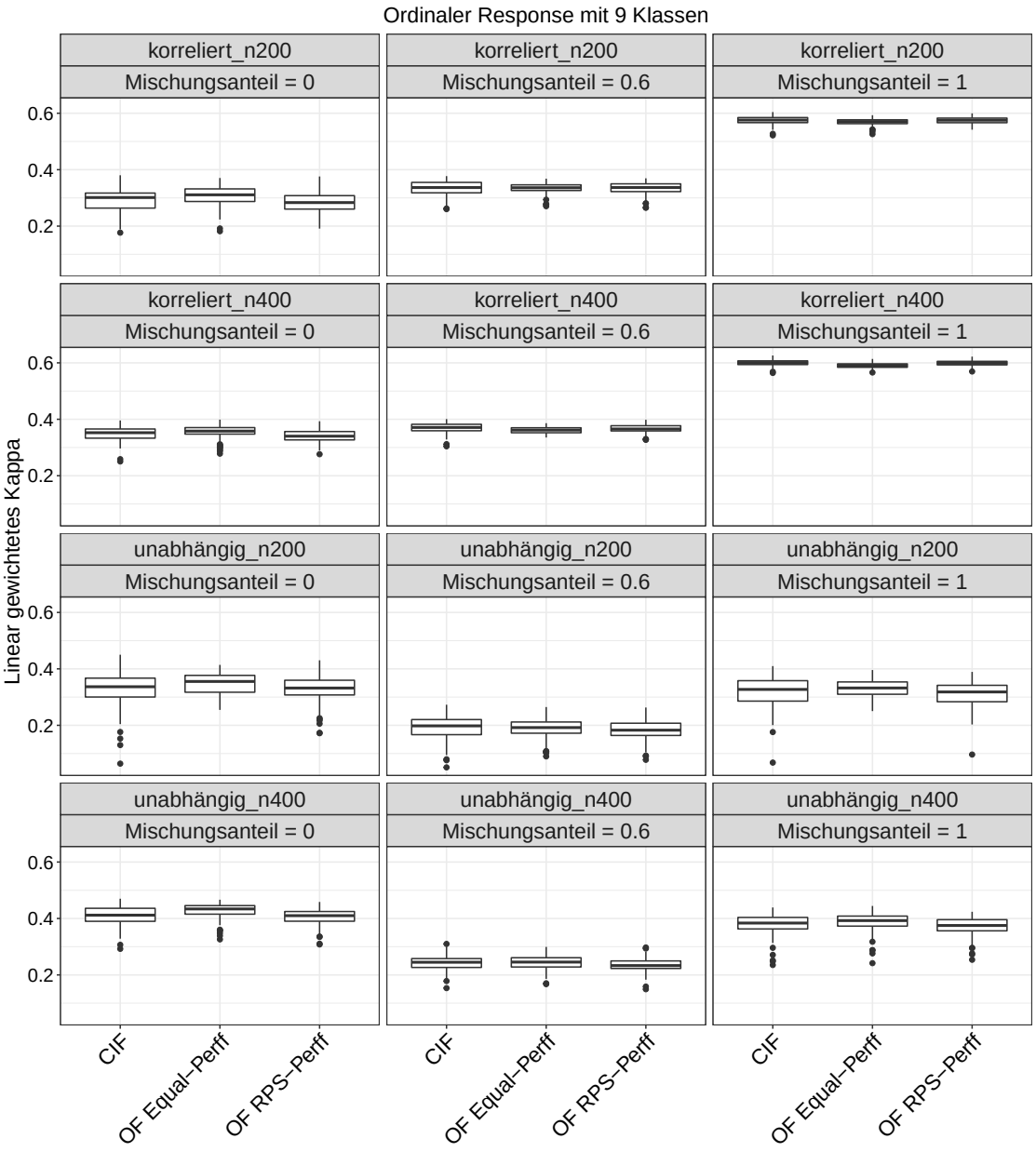

Abb. 4.7.: Werte für das linear gewichtete Kappa von den betrachteten Methoden in jedem Setting für einen ordinalen Response mit 9 Klassen basierend auf den Simulationen von Janitza et al. (2016). Jeder Boxplot zeigt die Werte für die 100 Iterationen.

Die Boxplots in den Abbildungen 4.7 - 4.9 zeigen die Werte des linear gewichteten Kappa für die Methoden CIF, OF mit Equal-Perff und OF mit RPS-Perff basierend auf dem Simulationsexperiment von Janitza et al. (2016). Die Abbildung 4.7 stellt die entsprechenden Ergebnisse für 9 Klassen, die Abbildung 4.8 für 6 Klassen und die Abbildung 4.9 für 3 Klassen des ordinalen Response dar. Alle Darstellungen mit den Werten für das quadratisch gewichtete Kappa (Abb. A.1, A.3, A.5) und Cohen's Kappa (Abb. A.2, A.4, A.6) befinden sich im Anhang dieser Arbeit. Damit liegt der Fokus zunächst auf den Ergebnissen für das linear gewichtete Kappa, da dieses Maß einen Kompromiss zwischen den anderen beiden Kappa-Maßen schließt. Im weiteren Verlauf werden zentrale Unterschiede zwischen allen drei Maßen herausgearbeitet, separat für jede Klassenanzahl. Die Zeilen und Spalten von jeder Abbildung unterscheiden sich auf gleiche Weise voneinander, wie bereits zu den Abbildungen für die AUC-Werte basierend auf den Simulationen von Janitza et al. (2016) erläutert.

In den Simulationen von 9 Klassen in Abbildung 4.7 sind die Werte für das linear gewichtete Kappa von allen Methoden sehr ähnlich. Dennoch sind die linearen Kappa-Werte des OF mit Equal-Perff den linearen Kappa-Werten des OF mit RPS-Perff und des CIF in den meisten Situationen minimal überlegen. Gleichzeitig weist der OF mit RPS-Perff am häufigsten die schlechteste Leistung bezüglich des linear gewichteten Kappa auf. Dennoch ist dieser OF in den Simulationen von korrelierten Prädiktorendaten und einem Mischungsanteil größer als Null dem OF mit Equal-Perff überlegen, gleichzeitig dem CIF jedoch unterlegen. Demnach ist in diesen Situationen der OF mit Equal-Perff etwas schlechter als der CIF. Eine weitere Einstellung, wo dies noch der Fall ist, ist diejenige von unabhängigen Prädiktoren, $n = 200$ und einem Mischungsanteil von $\zeta = 0.6$.
Darüber hinaus kann in allen Simulationen für 9 Klassen beobachtet werden, dass die linear gewichteten Kappa-Werte für die Stichprobengröße $n = 200$ kleiner als für die Stichprobengröße $n = 400$ sind. Jedoch gibt es für die beiden Stichprobengrößen bis auf die wenigen, oben dargelegten, Ausnahmen keine Unterschiede in den Leistungen der drei Methoden im Vergleich zueinander.

Die Werte für das quadratisch gewichtete Kappa in der ergänzenden Abbildung A.1 sind allesamt höher als für das linear gewichtete Kappa, was allgemein häufig in der Praxis beobachtet werden kann (Warrens, 2013). Außerdem unterscheiden sich in den Ergebnissen für das quadratische und lineare Kappa die Leistungen der drei Methoden relativ zueinander, dahingehend, dass der CIF für das quadratische Kappa bis auf zwei Einstellungen („korreliert_n200“ und „unabhängig_n400“, jeweils

für $\zeta = 0$) jetzt auch etwas besser als der OF mit Equal-Perff abschneidet. Demnach kann angenommen werden, dass der CIF seltener als die OF-Varianten Vorhersagen trifft, die viele ordinale Einheiten von den wahren Klassenwerten entfernt liegen.

Die Werte von Cohen's Kappa in der ergänzenden Abbildung A.2 sind deutlich kleiner als die Werte von dem Kappa mit linearen und quadratischen Gewichten. Diese Ungleichheit kann häufig in der Praxis beobachtet werden (Warrens, 2013). Außerdem gibt es kaum Unterschiede zwischen den verschiedenen Methoden. Dennoch verändern sich erneut die Leistungen der Methoden relativ zueinander, verglichen mit dem gewichteten Kappa. Anders als beim gewichteten Kappa ist gemäß Cohen's Kappa in den Einstellungen von unabhängigen Prädiktoren und gleichzeitig $\zeta = 0$ für beide Stichprobengrößen der OF mit RPS-Perff dem CIF minimal überlegen, dem OF mit Equal-Perff jedoch weiterhin unterlegen. Daneben ist der OF mit RPS-Perff in den Einstellungen von korrelierten Prädiktoren und gleichzeitig $\zeta = 0.6$ für beide Stichprobengrößen sowie $\zeta = 1$ für die Stichprobengröße $n = 200$ beiden Methoden geringfügig überlegen.
Darüber hinaus ändern sich bei Cohen's Kappa, verglichen mit dem gewichteten Kappa, ebenso die Leistungen des OF mit Equal-Perff und des CIF zueinander, zugunsten des OF mit Equal-Perff. Dabei handelt es sich um die Einstellungen von korrelierten und unabhängigen Prädiktoren, jeweils für $n = 200$ und $\zeta = 0.6$, wo die Leistungen des OF mit Equal-Perff den Leistungen des CIF in Bezug auf Cohen's Kappa minimal überlegen sind, anders als in Bezug auf das lineare Kappa. Daneben ist diese Veränderung, verglichen mit dem quadratischen Kappa, für fast alle Einstellungen zu beobachten.
Da die Leistungen des OF mit Equal-Perff bezüglich Cohen's Kappa überwiegend am besten sind, kann von diesem OF erwartet werden, leistungsfähiger bezüglich exakten Vorhersagen zu sein, verglichen mit den anderen Vorhersagemethoden.

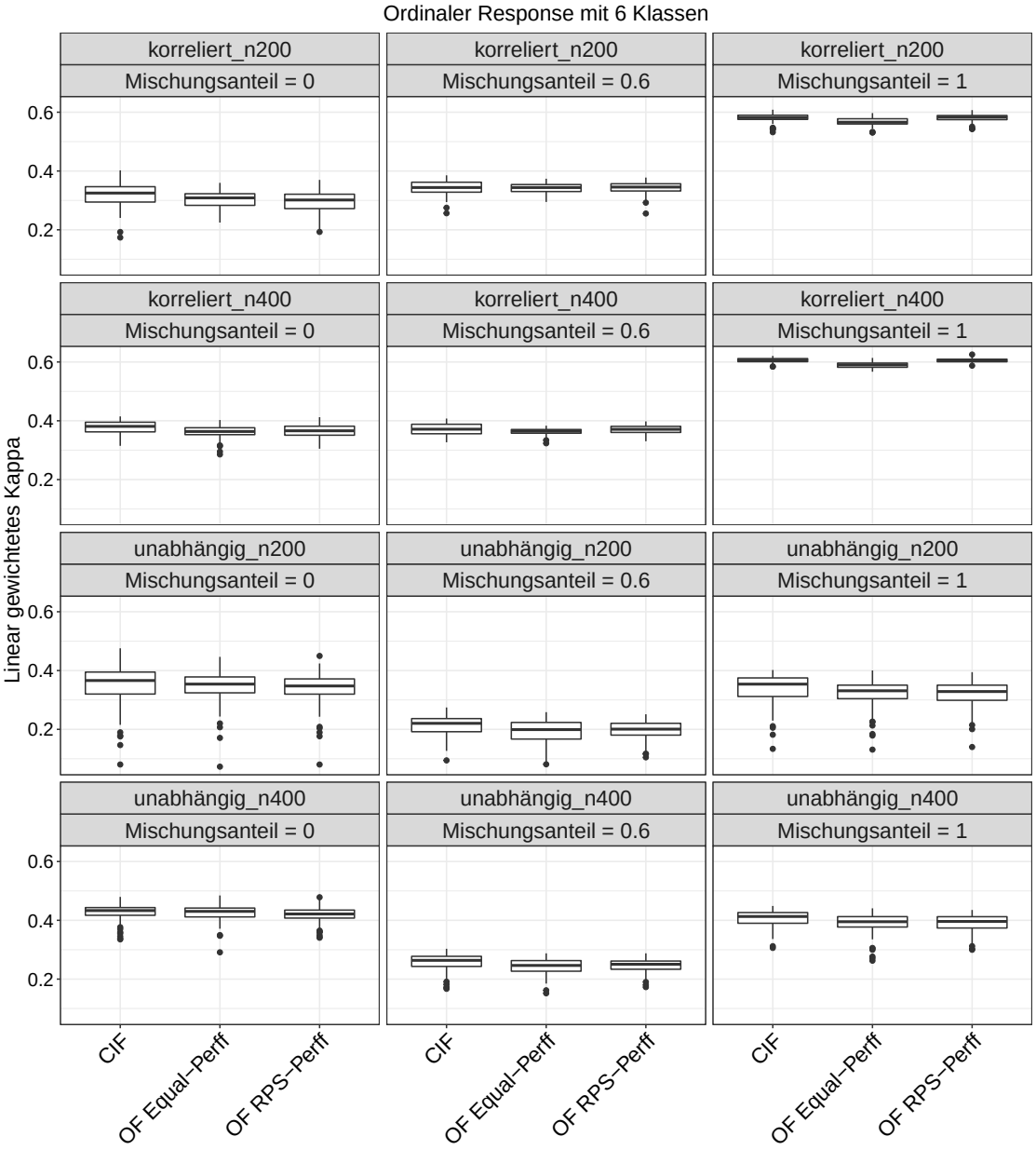

Abb. 4.8.: Werte für das linear gewichtete Kappa von den betrachteten Methoden in jedem Setting für einen ordinalen Response mit 6 Klassen basierend auf den Simulationen von Janitza et al. (2016). Jeder Boxplot zeigt die Werte für die 100 Iterationen.

In den Simulationen von 6 Klassen in Abbildung 4.8 zeigen die Werte des linear gewichteten Kappa erneut sehr ähnliche Ergebnisse für alle Methoden. Im Vergleich zu 9 Klassen verringern sich in fast allen Simulationen von 6 Klassen die Unterschiede in den linearen Kappa-Werten nochmals zwischen den beiden OF-Varianten. Dennoch sind die Werte des OF mit RPS-Perff häufiger minimal besser als die Werte des OF mit Equal-Perff, was auf eine Verbesserung im Vergleich zu den Simulationen von 9 Klassen hindeutet. Die Ausnahmen, in denen der OF mit RPS-Perff keine besseren Leistungen als der OF mit Equal-Perff erzielt, sind weiterhin fast alle Einstellungen

mit einem Mischungsanteil von $\zeta = 0$ (außer „korreliert_n400“) und die Einstellung „unabhängig_n200“ mit einem Mischungsanteil von $\zeta = 1$.
Darüber hinaus ist die Vorhersageleistung des CIF derjenigen des OF mit RPS-Perff weiterhin in den meisten Fällen geringfügig überlegen. Die Ausnahmen sind die Einstellungen von korrelierten Prädiktoren und $n = 200$, jeweils mit einem Mischungsanteil größer als Null. Zusätzlich ist die Leistung des CIF derjenigen des OF mit Equal-Perff sogar in allen Fällen minimal überlegen, sodass allgemein der CIF bei einer Anzahl von 6 Klassen besser als der OF mit Equal-Perff abschneidet.
Überdies sind erneut die linearen Kappa-Werte für die Stichprobengröße von $n = 400$ größer als von $n = 200$. Jedoch sind, zumindest für die Einstellungen von unabhängigen Prädiktorendaten, keine Unterschiede in den Ergebnissen für beide Stichprobengrößen in Bezug auf die Leistungen der Methoden relativ zueinander zu beobachten. Die geringfügigen Abweichungen, die teilweise für die Einstellungen von korrelierten Prädiktorendaten festgestellt werden können, wurden bereits oben dargelegt.

In Bezug auf 6 Klassen stimmen die Ergebnisse für das quadratisch gewichtete Kappa (ergänzende Abb. A.3) mit den Ergebnissen für das linear gewichtete Kappa größtenteils überein, abgesehen von den größeren Werten seitens des quadratisch gewichteten Kappa. Die wenigen relevanten Abweichungen bestehen darin, dass die Unterschiede einerseits zwischen dem CIF und OF Equal-Perff und andererseits zwischen dem CIF und OF mit RPS-Perff in den meisten Simulationen minimal zunehmen. Außerdem ist der OF mit RPS-Perff in den Einstellungen von korrelierten Prädiktoren und $\zeta = 1$ für $n = 400$ sowie erneut für $n = 200$ die beste Methode.

Die Werte für Cohen's Kappa (ergänzende Abb. A.4) schwanken in fast allen Simulationen von 6 Klassen wie von 9 Klassen um 0.1, außer in den beiden Einstellungen mit korrelierten Prädiktorendaten und gleichzeitig $\zeta = 1$. Außerdem verändern sich erneut die Leistungen der Methoden relativ zueinander, im Vergleich zum gewichteten Kappa. Gemäß Cohen's Kappa ist demnach beispielsweise der OF mit RPS-Perff in der Einstellung von korrelierten Prädiktorendaten, $n = 400$ und $\zeta = 0.6$ die beste Methode, verglichen mit der Konkurrenz. Zusätzlich schneidet diese OF-Variante in der Einstellung von korrelierten Prädiktorendaten, $n = 200$ und gleichzeitig $\zeta = 1$ für alle drei Kappa-Maße bei einer Anzahl von 6 Klassen konstant am besten ab.

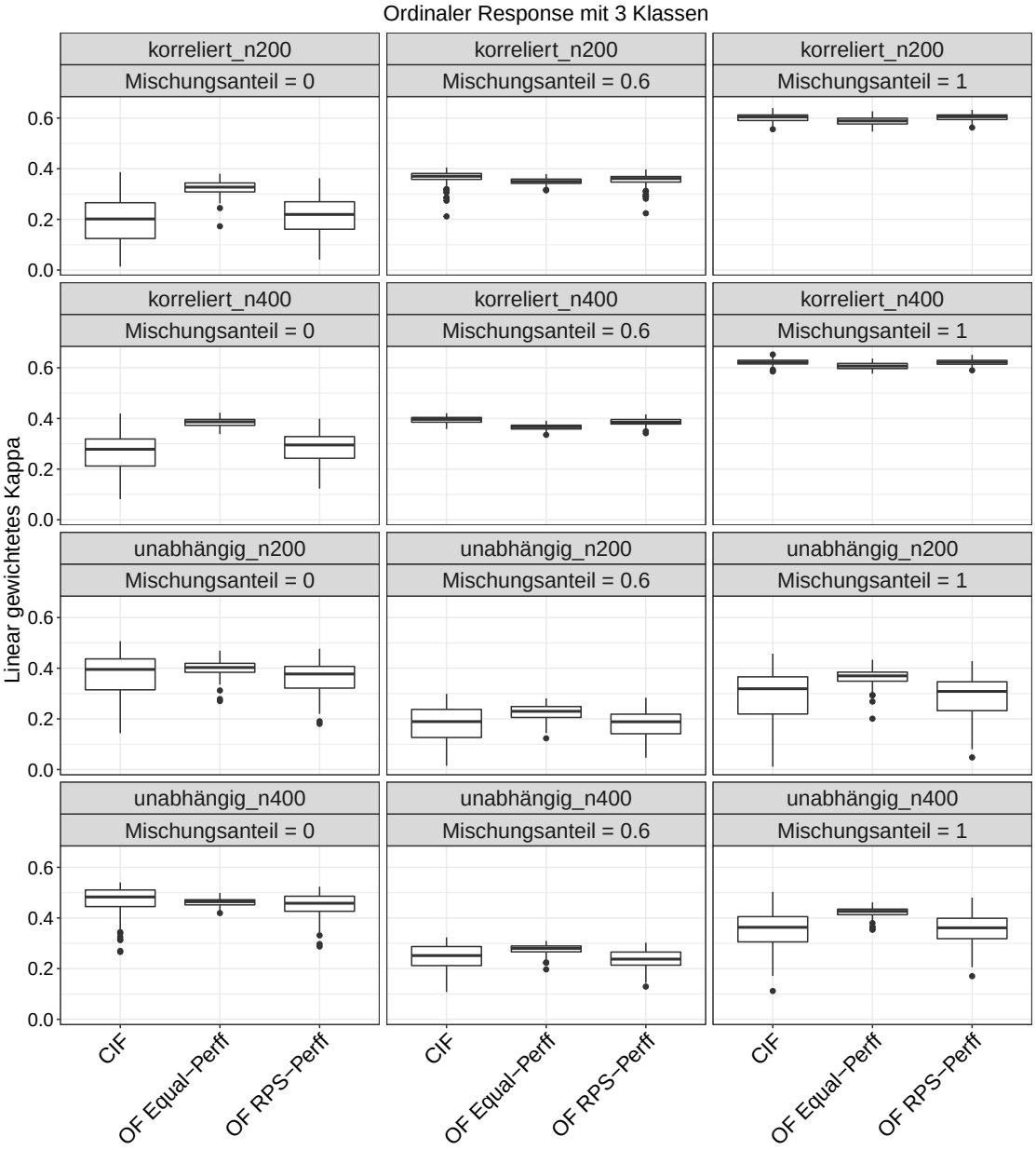

Abb. 4.9.: Werte für das linear gewichtete Kappa von den betrachteten Methoden in jedem Setting für einen ordinalen Response mit 3 Klassen basierend auf den Simulationen von Janitza et al. (2016). Jeder Boxplot zeigt die Werte für die 100 Iterationen.

In den Simulationen von 3 Klassen in Abbildung 4.9 erzielt der OF mit Equal-Perff überwiegend die besten Resultate bezüglich des linearen Kappa, gefolgt von dem CIF. Dies widerspricht den Ergebnissen für das lineare Kappa bei einer Anzahl von 6 Klassen, entspricht allerdings den Ergebnissen für das lineare Kappa bei einer Anzahl von 9 Klassen. Außerdem belegen die Ergebnisse für das lineare Kappa in den Simulationen von 3 Klassen eine Abnahme der Ähnlichkeiten zwischen allen betrachteten Methoden. Die Einstellungen von korrelierten Prädiktoren und $\zeta = 0$ für beide Stichprobengrößen sowie von unabhängigen Prädiktoren und $\zeta = 1$ ebenso für

beide Stichprobengrößen zeigen die stärksten Verbesserungen von OF mit Equal-Perff gegenüber dem CIF, verglichen mit allen Ergebnissen zum linearen Kappa für 3, 6 und 9 Klassen. In diesen Fällen sowie zusätzlich in beiden Einstellungen von unabhängigen Prädiktoren und gleichzeitig $\zeta = 0.6$ verbessert sich der OF mit Equal-Perff auch gegenüber dem OF mit RPS-Perff am stärksten, verglichen mit allen Ergebnissen zum linearen Kappa für 3, 6 und 9 Klassen.
Darüber hinaus zeigen die Simulationen von 3 Klassen, dass erneut die linearen Kappa-Werte für die geringere Stichprobengröße kleiner als für die größere Stichprobe sind. Die Leistungen der betrachteten Methoden relativ zueinander bleiben jedoch unverändert.

In den Simulationen von 3 Klassen zeigen die Werte für das quadratische Kappa (ergänzende Abb. A.5) sehr ähnliche Tendenzen zu den Werten für das lineare Kappa bei gleicher Klassenanzahl auf, wobei die quadratischen Kappa-Werte höher sind.

Die Werte von Cohen's Kappa für 3 Klassen (ergänzende Abb. A.6) sind allesamt etwas geringer als die Werte vom gewichteten Kappa für 3 Klassen, lassen jedoch die gleichen Entwicklungen erkennen.

Für das Simulationsdesign von Janitza et al. (2016) kann zusammengefasst werden, dass mit abnehmender Klassenanzahl eine Abnahme der Ähnlichkeiten zwischen den Methoden beobachtet werden kann. Darüber hinaus schneidet der OF mit RPS-Perff in den Einstellungen von 6 Klassen häufiger minimal besser als der OF mit Equal-Perff ab, jedoch meistens etwas schlechter als der CIF. Daneben ist der OF mit RPS-Perff in der Mehrzahl aller Einstellungen von 3 und 9 Klassen am schlechtesten, während der OF mit Equal-Perff häufig am besten abschneidet.

Kappa-Werte für die Simulationen von Hornung

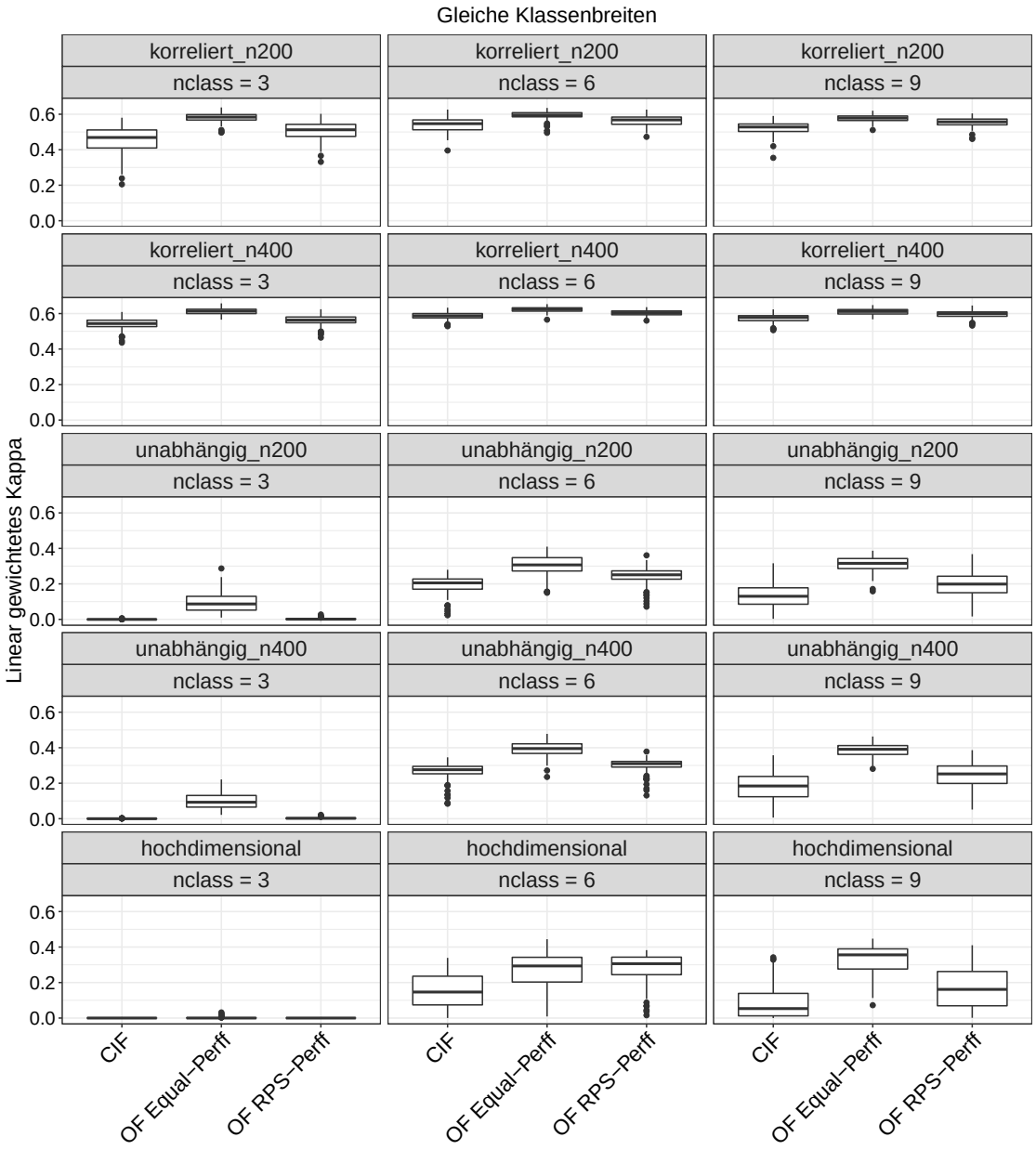

Abb. 4.10.: Werte für das linear gewichtete Kappa von den betrachteten Methoden in jedem Setting mit gleichen Klassenbreiten basierend auf den Simulationen von Hornung (2020). Jeder Boxplot zeigt die Werte für die 100 Iterationen.

Die Ergebnisse für das linear gewichtete Kappa von den Methoden CIF, OF mit Equal-Perff und OF mit RPS-Perff sind für alle Simulationen mit gleichen Klassenbreiten in Abbildung 4.10 und mit zufälligen Klassenbreiten in Abbildung 4.11 basierend auf dem Simulationsdesign von Hornung (2020) dargestellt. Die entsprechenden Abbildungen mit den Werten für das quadratisch gewichtete Kappa (Abb. A.7 und A.9) und Cohen's Kappa (Abb. A.8 und A.10) befinden sich im Anhang

dieser Arbeit. Nach Beschreibung der Ergebnisse für das linear gewichtete Kappa wird auf relevante Abweichungen in den Ergebnissen zwischen allen drei Kappa-Varianten aufmerksam gemacht, separat für jede Art von Klassenbreite. Die Zeilen und Spalten aller Abbildungen unterscheiden sich erneut auf gleiche Weise voneinander, wie zu den Abbildungen für die AUC-Werte, die ebenso für die simulierten Daten von Hornung erhalten wurden, dargelegt.

In nahezu allen Simulationen mit gleichen Klassenbreiten weist der OF mit Equal-Perff höhere Werte für das linear gewichtete Kappa als der OF mit RPS-Perff und als der CIF auf (Abb. 4.10). Die Ausnahmen sind die Einstellungen von hochdimensionalen Prädiktorendaten mit 3 und 6 Klassen. Für den entsprechenden Fall mit 3 Klassen weisen alle Vorhersagemethoden Werte um Null auf und für den jeweiligen Fall mit 6 Klassen ist der OF mit RPS-Perff minimal besser als der OF mit Equal-Perff und beide Methoden deutlich besser als der CIF. Neben diesem letzten Fall schneidet der OF mit RPS-Perff in vielen weiteren Einstellungen besser als der CIF ab. Die Ausnahmen sind alle Simulationen von 3 Klassen mit unabhängigen Daten sowie von 3 Klassen mit hochdimensionalen Daten. In diesen Einstellungen sind die Kappa-Werte aller Methoden nahe Null. Die schlechte Leistung in den benannten Simulationen von 3 Klassen könnte gemäß Hornung (2020) damit begründet werden, dass dabei der latente stetige Response stark vergröbert ist, was deutlich weniger Signal in den Daten zur Folge hat.
Die Simulationen von hochdimensionalen Prädiktorendaten und einem Response mit 6 sowie 9 Klassen zeigen die stärkste Verbesserung von dem OF mit RPS-Perff gegenüber dem CIF. In diesen Fällen sowie zusätzlich in den Simulationen von unabhängigen Prädiktoren und gleichzeitig einem Response mit 9 Klassen ist auch die Verbesserung von dem OF mit Equal-Perff gegenüber dem CIF am größten. Daneben verbessert sich die Leistung des OF mit Equal-Perff gegenüber der Leistung des OF mit RPS-Perff in den beiden Einstellungen mit unabhängigen Prädiktoren und 9 Klassen am deutlichsten.
In fast allen Simulationen, außer in denjenigen mit Kappa-Werten um Null für alle Methoden, kann beobachtet werden, dass die Werte für $n = 400$ größer als für $n = 200$ sind. Dennoch gibt es zwischen den Ergebnissen für beide Stichprobengrößen keine Unterschiede in den Leistungen der betrachteten Methoden im Vergleich zueinander. Bis auf die wenigen, hier dargelegten, Ausnahmen scheint insgesamt kein Trend dahingehend vorhanden zu sein, dass durch die Anzahl der Klassen die Leistungen der Vorhersagemethoden relativ zueinander beeinflusst werden.

Die Ergebnisse für das quadratisch gewichtete Kappa in der ergänzenden Abbildung A.7 weisen höhere Werte auf als für das linear gewichtete Kappa, stimmen jedoch in Bezug auf die Leistungen der Methoden relativ zueinander überein. Dennoch nehmen die Verbesserungen von beiden OF-Varianten gegenüber dem CIF in fast allen Einstellungen mit unabhängigen und hochdimensionalen Daten, jeweils sowohl für 6 als auch 9 Klassen, deutlich zu, außer in der Einstellung von unabhängigen Prädiktoren, $n = 400$ und 6 Klassen, wo die Leistung des OF mit RPS-Perff gegenüber der Leistung des CIF lediglich minimal verbessert ist. In den gleichen Einstellungen verbessert sich außerdem die Leistung des OF mit Equal-Perff gegenüber der Leistung des OF mit RPS-Perff mit Ausnahme von dem hochdimensionalen Fall mit 6 Klassen, in dem der OF mit RPS-Perff erneut eine minimal bessere Leistung als der OF mit Equal-Perff erzielt. Insgesamt nehmen also die Unterschiede zwischen den Methoden insbesondere in den Einstellungen, in denen die Unterschiede bereits für das linear gewichtete Kappa groß waren, weiterhin zu.
Auf Grundlage dieser Ergebnisse kann angenommen werden, dass der OF mit Equal-Perff weniger häufig als die Konkurrenz zu Vorhersagen führt, die viele ordinale Einheiten von den wahren Klassenwerten entfernt sind. Gleichzeitig kann davon ausgegangen werden, dass der OF mit RPS-Perff seltener als der CIF, aber häufiger als der OF mit Equal-Perff, Vorhersagen liefert, die weit von den wahren Klassenwerten weg liegen.

Im Gegensatz zu den Kappa-Werten mit Gewichten fallen die Werte für Cohen's Kappa in allen Simulationen von gleichen Klassenbreiten (ergänzende Abb. A.8) kleiner aus. Dennoch bleiben die Leistungen der Methoden relativ zueinander bestehen, wobei eine Abnahme der Unterschiede zwischen allen Methoden auffällig ist, vor allem in den Situationen, in denen die Unterschiede für das Kappa mit Gewichten besonders groß waren. Demnach ist in diesen Situationen die Verbesserung von OF mit Equal-Perff gegenüber den konkurrierenden Methoden hinsichtlich annähernd richtigen Vorhersagen stärker als hinsichtlich exakten Vorhersagen.

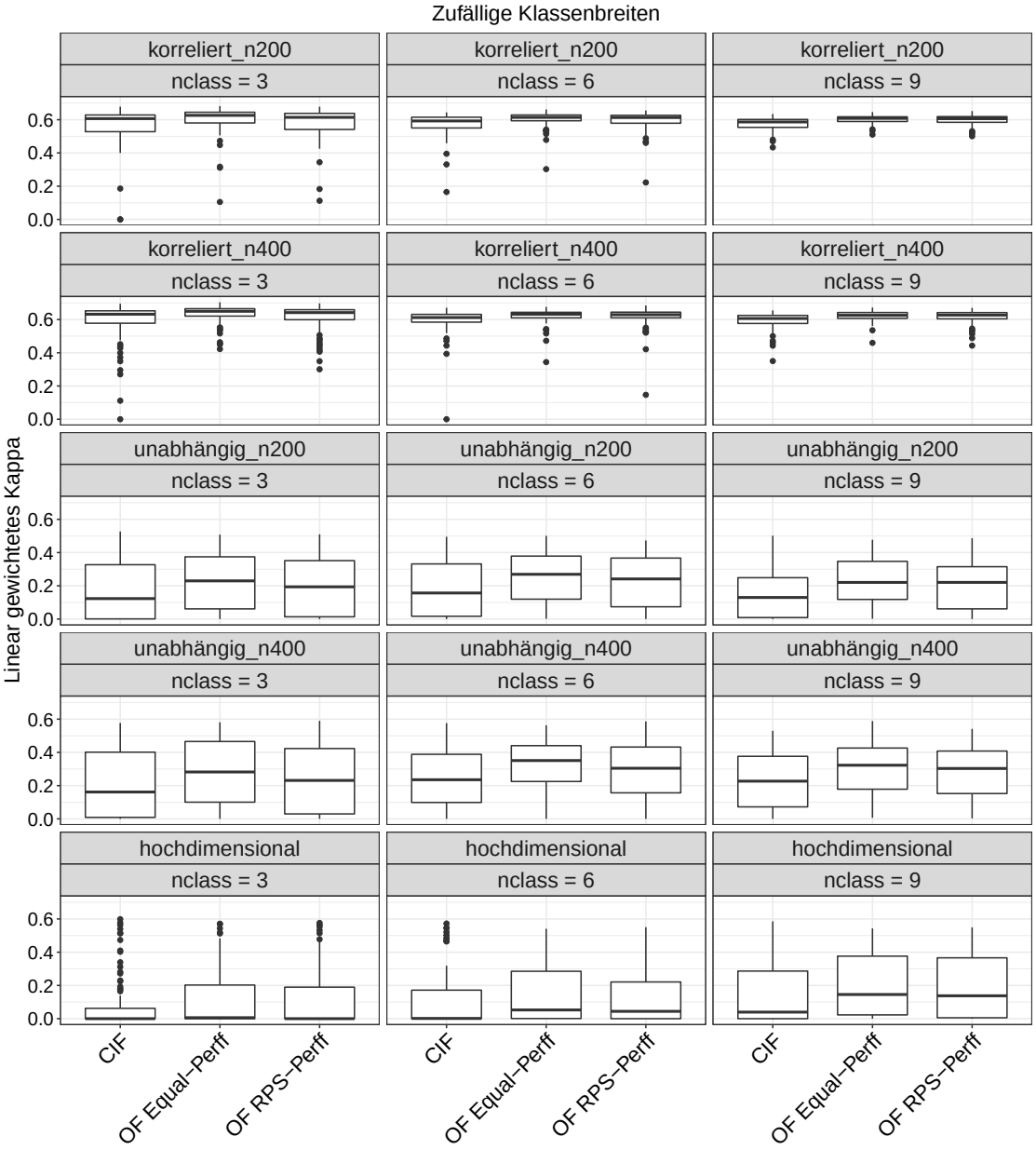

Abb. 4.11.: Werte für das linear gewichteten Kappa von den betrachteten Methoden in jedem Setting mit zufälligen Klassenbreiten basierend auf den Simulationen von Hornung (2020). Jeder Boxplot zeigt die Werte für die 100 Iterationen.

Im Folgenden werden die linearen Kappa-Werte für die Simulationen von zufälligen Klassenbreiten (Abb. 4.11) mit den vorherigen linearen Kappa-Werten für die Simulationen von gleichen Klassenbreiten hinsichtlich relevanter Unterschiede verglichen. In den Simulationen von zufälligen Klassenbreiten sind die Werte für das lineare Kappa von dem OF mit RPS-Perff in zwei Fällen, anstatt nur einem Fall wie in den Simulationen von gleichen Klassenbreiten, etwas besser als von dem OF mit Equal-Perff. Bei diesen Fällen handelt es sich um die Simulationen von korrelierten Prädiktoren und $n = 400$ sowie von unabhängigen Prädiktoren und $n = 200$, jeweils

für 9 Klassen. Die minimalen Leistungsunterschiede können jedoch auch Zufallsschwankungen geschuldet sein. Ansonsten gibt es keine Diskrepanzen in den linearen Kappa-Werten für zufällige und gleiche Klassenbreiten bezüglich der Leistungen der Methoden relativ zueinander. Allerdings sind in allen Simulationen für zufällige Breiten geringere Abweichungen zwischen den Werten von beiden OF-Varianten erkennbar. Darüber hinaus sind in den meisten Einstellungen ebenso abnehmende Unterschiede zwischen den Werten des OF mit Equal-Perff und des CIF zu beobachten. Die Ausnahmen bilden die Einstellungen der unabhängigen und hochdimensionalen Prädiktorendaten, jeweils mit einem Response von 3 Klassen, sowie der unabhängigen Prädiktorendaten mit $n = 200$ und 6 Klassen, wo die Abweichungen zwischen OF mit Equal-Perff und CIF leicht zunehmen. Schließlich ist auffällig, dass die Einstellungen mit unabhängigen und die Einstellungen mit hochdimensionalen Prädiktorendaten deutlich höhere Varianzen aufweisen.

Die Ergebnisse für das quadratisch gewichtete Kappa von zufälligen Klassenbreiten (ergänzende Abb. A.9) ähneln den Ergebnissen für das linear gewichtete Kappa von zufälligen Klassenbreiten mit den Ausnahmen, dass der OF mit RPS-Perff nun in zwei anderen Einstellungen für 9 Klassen mit minimalem Abstand am besten von allen Methoden abschneidet: in den Einstellungen mit korrelierten Prädiktoren und $n = 200$ sowie unabhängigen Prädiktoren und $n = 400$. Außerdem sind hier, verglichen mit den Ergebnissen für das linear gewichtete Kappa, stärkere Verbesserungen seitens beider OF-Varianten gegenüber dem CIF insbesondere in den Einstellungen von unabhängigen Prädiktorendaten für jede Klassenanzahl sowie von hochdimensionalen Prädiktorendaten für 9 Klassen auffällig.

Im Fall von Cohen's Kappa (ergänzende Abb. A.10) sind erneut die geringeren Werte, verglichen mit den gewichteten Kappa-Werten, zu beobachten. Außerdem zeigt sich, dass der OF mit RPS-Perff erstmals in zwei Simulationen von 6 Klassen sowie in zwei Simulationen von 9 Klassen am besten von allen Methoden abschneidet. Bei diesen Simulationen handelt es sich um „korreliert_n400“ und „nclass = 6“, „hochdimensional“ und „nclass = 6“, „korreliert_n200“ und „nclass = 9“ (im Einklang mit quadratisch gewichtetem Kappa) sowie „korreliert_n400“ und „nclass = 9“ (im Einklang mit linear gewichtetem Kappa). Zusätzlich ist in allen Einstellungen für unabhängige Prädiktoren eine deutliche Abnahme der Abweichungen zwischen den drei Methoden im Vergleich zum gewichteten Kappa zu beobachten. In den anderen Einstellungen bleiben die Abstände zwischen den Methoden jedoch weitestgehend konstant.

Für alle Simulationen von gleichen und zufälligen Klassenbreiten basierend auf dem Simulationsdesign von Hornung (2020) kann zusammengefasst werden, dass der OF mit Equal-Perff überwiegend besser als der OF mit RPS-Perff und als der CIF abschneidet, während der OF mit RPS-Perff meistens besser als der CIF ist.
Für gleiche Klassenbreiten zeigen die Ergebnisse zwischen $n = 200$ und $n = 400$ keine Unterschiede in Bezug auf die Leistungen der Methoden relativ zueinander. Demgegenüber gibt es für zufällige Klassenbreiten minimale Abweichungen zwischen den Ergebnissen für die verschiedenen Stichprobengrößen, die allerdings auf kein bestimmtes Muster hindeuten. Außerdem ist für beide Klassenbreiten kein eindeutiger Trend dahingehend zu beobachten, dass durch die Anzahl der Klassen die Leistungen der Methoden relativ zueinander beeinflusst werden.
Der Vergleich von Cohen's Kappa mit dem gewichteten Kappa zeigt für beide Arten von Klassenbreiten, dass im Allgemeinen bei Ersterem sowohl die einzelnen Werte als auch die Unterschiede zwischen den Methoden geringer sind.

Kappa-Werte für die Simulationen von Buri und Hothorn

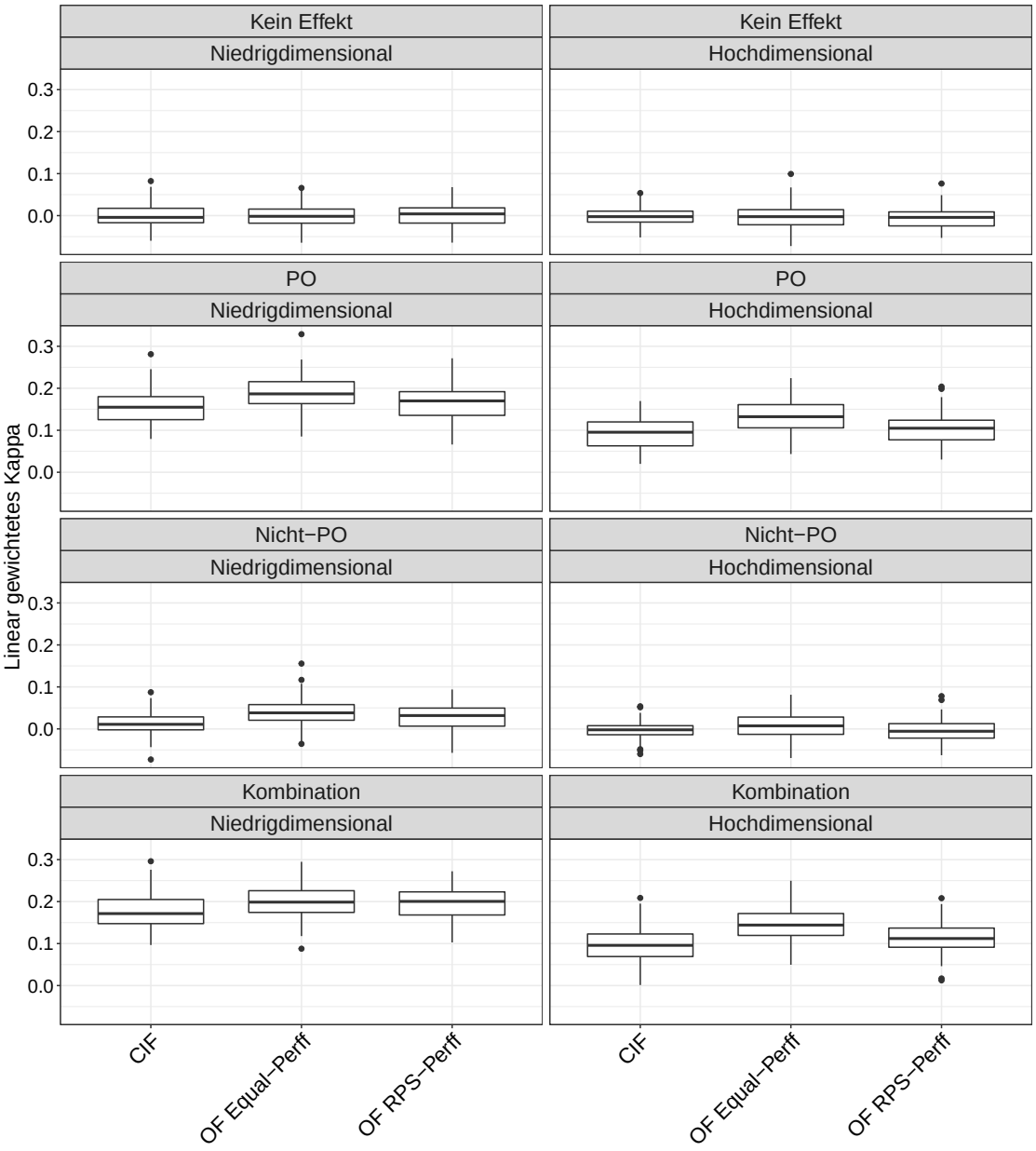

Abb. 4.12.: Werte für das linear gewichtete Kappa von den betrachteten Methoden in jedem Setting basierend auf den Simulationen von Buri und Hothorn (2020). Jeder Boxplot zeigt die Werte für die 100 Iterationen.

Die Abbildung 4.12 zeigt die Werte des linear gewichteten Kappa von den Vorhersagemethoden CIF, OF mit Equal-Perff und OF mit RPS-Perff für alle Settings basierend auf der Simulationsumgebung von Buri und Hothorn (2020). Die entsprechenden Abbildungen mit den Ergebnissen für das quadratisch gewichtete Kappa (Abb. A.11) und für Cohen's Kappa (Abb. A.12) sind im Anhang dieser Arbeit zu finden. Erneut liegt der Fokus zunächst auf den Ergebnissen für das lineare Kappa,

gefolgt von zentralen Unterschieden zwischen dem linearen Kappa, dem quadratischen Kappa und Cohen's Kappa. Die Zeilen und Spalten von allen Abbildungen unterscheiden sich gleichermaßen voneinander, wie bereits für die Abbildungen zu den AUC-Werten basierend auf derselben Simulationsumgebung dargelegt.

Die Werte für das linear gewichtete Kappa in Abbildung 4.12 sind allesamt sehr gering und für die Einstellungen ohne Effekt sowie für die hochdimensionale Einstellung mit Nicht-PO schwanken die Leistungen von allen Methoden um 0. Der Grund für die geringen Kappa-Werte liegt darin, dass in den simulierten Daten fast kein Signal vorhanden ist.
Abgesehen von den Einstellungen, in denen die Kappa-Werte um 0 schwanken, weist der OF mit Equal-Perff durchweg etwas höhere Werte für das lineare Kappa auf als der CIF und gleichzeitig mit Ausnahme einer weiteren Einstellung auch etwas höhere Werte als der OF mit RPS-Perff. Die Ausnahme ist die Kombination aus PO und Nicht-PO für die niedrigdimensionalen Prädiktorendaten, wo der OF mit RPS-Perff eine minimal bessere Leistung als der OF mit Equal-Perff aufweist. Darüber hinaus ist die Leistung des OF mit RPS-Perff in fast allen Einstellungen, außer in denen mit linearen Kappa-Werten um Null, geringfügig stärker als die Leistung des CIF.
Die Simulationen von PO und der Kombination aus PO und Nicht-PO, jeweils für die niedrig- als auch die hochdimensionalen Prädiktorendaten, sowie von Nicht-PO für die niedrigdimensionalen Prädiktorendaten zeigen die stärkste Verbesserung von OF mit Equal-Perff gegenüber dem CIF. Daneben ist die Verbesserung von OF mit RPS-Perff gegenüber dem CIF in den Einstellungen von PO und der Kombination, jeweils lediglich für die niedrigdimensionalen Daten, am größten. Für diese Effekttypen, allerdings diesmal für die hochdimensionalen Prädiktorendaten, zeigt sich außerdem die stärkste Verbesserung in der Leistung von OF mit Equal-Perff gegenüber der Leistung von OF mit RPS-Perff.
Schließlich kann in den Ergebnissen kein klarer Trend beobachtet werden, d.h. die Leistungen der Methoden relativ zueinander steigen oder sinken nicht tendenziell durch spezifische Einstellungen.

Die Ergebnisse für das quadratisch gewichtete Kappa in der ergänzenden Abbildung A.11 zeigen ein übereinstimmendes Bild mit den Ergebnissen für das linear gewichtete Kappa, abgesehen von den höheren Werten. Bezüglich der höheren Werte, die vom quadratisch gewichteten Kappa ausgehen, bildet die Einstellung von Nicht-PO für die niedrigdimensionalen Prädiktorendaten eine Ausnahme, wo die Werte für das quadratisch gewichtete Kappa von allen Methoden etwas geringer sind als die Werte

für das linear gewichtete Kappa. Ein kleinerer Wert für das quadratisch gewichtete Kappa als für das linear gewichtete Kappa wird in der Praxis selten beobachtet (Warrens, 2013). Demnach ist es in dieser Einstellung wahrscheinlich, dass ein großer Teil der Vorhersagen weit von den wahren Klassenwerten entfernt liegt.

Zusätzlich sind die Ergebnisse für Cohen's Kappa (ergänzende Abb. A.12) den Ergebnissen für das Kappa mit Gewichten sehr ähnlich. Allerdings ist nun in allen Einstellungen mit Kappa-Werten größer als Null der OF mit Equal-Perff die beste Methode, verglichen mit der Konkurrenz. Demnach kann erwartet werden, dass der OF mit Equal-Perff leistungsfähiger als die Konkurrenz in Bezug auf exakte Vorhersagen ist und gleichzeitig seltener als die konkurrierenden Methoden Vorhersagen trifft, die weit von den wahren Klassenwerten entfernt liegen.

Vergleich von allen Ergebnissen bezüglich der Klassifikationsgüte

Im Folgenden werden die Ergebnisse bezüglich der Klassifikationsgüte für alle Simulationsdesigns miteinander verglichen, um relevante Unterschiede sowie Übereinstimmungen herauszuarbeiten.
Die Ergebnisse für die Simulationsumgebungen von Hornung (2020) sowie von Buri und Hothorn (2020) teilen die Gemeinsamkeit, dass der OF mit Equal-Perff überwiegend die beste Methode darstellt, gefolgt von dem OF mit RPS-Perff als zweitbeste Methode, unter den drei betrachteten Methoden in Bezug auf alle Kappa-Varianten. Mit dem Simulationsdesign von Janitza et al. (2016) werden dagegen unterschiedliche Ergebnisse für jede Klassenanzahl sowie teilweise für die verschiedenen Kappa-Varianten erreicht. So ist für 9 Klassen die Leistung von dem OF mit Equal-Perff ebenso meistens am besten, die Leistung von dem OF mit RPS-Perff jedoch überwiegend am schlechtesten von allen betrachteten Methoden in Bezug auf das linear gewichtete Kappa und Cohen's Kappa. Dieses Muster wiederholt sich für 3 Klassen bezüglich aller Kappa-Varianten. Hinsichtlich des quadratisch gewichteten Kappa für 9 Klassen sowie hinsichtlich aller Kappa-Varianten für 6 Klassen ermöglicht der CIF öfter bessere Vorhersagen als beide OF-Varianten. Für alle Kappa-Varianten von 6 Klassen erzielt gleichzeitig der OF mit RPS-Perff häufiger bessere Leistungen als der OF mit Equal-Perff. Dabei darf nicht außer Acht gelassen werden, dass die Werte, getrennt für jede Kappa-Variante und jedes Simulationsdesign, in den meisten Einstellungen sehr ähnlich sind, womit lediglich geringfügige Unterschiede zwischen den Methoden bestehen, die auch Zufallsschwankungen geschuldet sein können. Bei dem Simulationsdesign von Janitza et al. nehmen die Unterschiede zwischen fast allen Methoden für 6 Klassen, verglichen mit 9 Klassen nochmals ab,

während die Unterschiede zwischen allen Methoden für 3 Klassen wieder zunehmen. Für das Simulationsdesign von Hornung ergibt sich ein inkonsistentes Bild in Bezug auf eine mögliche Abnahme oder Zunahme der Unterschiede zwischen den Methoden bei sinkender Klassenanzahl. Für das Simulationsdesign von Buri und Hothorn kann diesbezüglich keine Aussage gemacht werden, da für alle Einstellungen lediglich ein ordinaler Response mit 4 Klassen verwendet wird.

Im Hinblick auf die erzielten Ergebnisse muss jedoch beachtet werden, dass viele Unterschiede nur sehr gering sind und unklar ist, ob diese nicht auch zufälligen Schwankungen geschuldet sind.

4.3. Reale Datenanalyse

4.3.1. Reale Daten

Der Vergleich zwischen den ausgewählten Random Forest-Ansätzen wird zusätzlich für fünf reale Datensätze mit ordinalem Response durchgeführt. Zu diesem Zweck werden die von Hornung (2020) vorverarbeiteten Versionen von diesen Datensätzen verwendet, die im elektronischen Anhang dieser Arbeit enthalten sind. In der Tabelle 4.2 sind die realen Datensätze mit ihren wichtigsten Eigenschaften für die Analyse aufgelistet.

Die Ergebnisse für die reale Datenanalyse können lediglich bezüglich der Klassifikationsgüte und nicht in Bezug auf die Qualität des Prädiktoren-Rankings berichtet werden, denn anders als mit simulierten Daten kann mit realen Daten nicht beurteilt werden, ob das VIM von einer Methode relevante Prädiktoren identifizieren kann (Hornung, 2020). Der Grund liegt darin, dass reale Daten keine Auskunft darüber geben, welche Prädiktoren einen tatsächlichen Effekt haben und welche nicht. Damit bleibt unbekannt, ob ein Prädiktor relevant ist.
Die Klassifikationsgüte wird wie bei den simulierten Daten anhand von dem Kappa mit linearen Gewichten und dem Kappa mit quadratischen Gewichten (Cohen, 1968) sowie anhand von Cohen's Kappa (Cohen, 1960) eingeschätzt. Diese drei Maße wurden bereits in Abschnitt 4.2.3 erläutert.

Um zu gewährleisten, dass keine übermäßig optimistischen Ergebnisse erhalten werden, die durch die Verwendung derselben Daten sowohl für die Konstruktion der Bäume in dem Random Forest als auch für die Vorhersagen und die anschließende Evaluation der Vorhersagen resultieren würden, wird zunächst wie in Hornung (2020) eine k-fach stratifizierte Kreuzvalidierung mit $k = 10$ durchgeführt. Dabei wird jeder Datensatz zufällig in $k = 10$ ähnlich große disjunkte Teilmengen aufgeteilt (Refaeilzadeh et al., 2009). Von den resultierenden Teilmengen dienen $k-1 = 9$ Teilmengen zusammen als Trainingsdatensatz für die Erstellung der Bäume in dem Random Forest, während die restliche Teilmenge als Testdatensatz für die Vorhersage des ordinalen Response verwendet wird. Diese Vorgehensweise durchläuft so viele Iterationen, bis jede der k Teilmengen einmal als Testdatensatz eingesetzt wurde. Bei $k = 10$ sind dies insgesamt 10 Iterationen. Für zuverlässigere Resultate wird der gesamte Prozess 10-mal wiederholt.

Bezeichnung des Datensatzes	Größe der Stichprobe	Anzahl an Prädiktoren	Betrachteter Response	Anzahl an Klassen	Bezeichnung und Größe der Klassen
mammography	412	5	„Last mammography visits“ (deutsch: Letzte Besuche bei der Mammographie)	3	1 – „never“ (deutsch: niemals), $n = 234$ 2 – „within a year“ (deutsch: innerhalb eines Jahres), $n = 104$ 3 – „over a year“ (deutsch: über ein Jahr), $n = 74$
nhanes	1914	26	„Self-reported health status“ (deutsch: Selbst eingeschätzter Gesundheitsstatus)	5	1 – „excellent“ (deutsch: ausgezeichnet), $n = 198$ 2 – „very good“ (deutsch: sehr gut), $n = 565$ 3 – „good“ (deutsch: gut), $n = 722$ 4 – „fair“ (deutsch: mittelmäßig), $n = 346$ 5 – „poor“ (deutsch: schlecht), $n = 83$
supportstudy	798	15	„Functional disability“ (deutsch: Funktionsunfähigkeit)	5	1 – „patient lived 2 months, and from an interview (taking place 2 months after study entry) there were no signs of moderate to severe functional disability“ (deutsch: der Patient lebte 2 Monate, und bei einer Befragung (die 2 Monate nach Studienbeginn stattfand) gab es keine Anzeichen für eine mittlere bis schwere Funktionsunfähigkeit), $n = 310$ 2 – „patient was unable to do 4 or more activities of daily living 2 months after study entry; if the patient was not interviewed but the patient's surrogate was, the cutoff for disability was 5 or more activities“ (deutsch: der Patient war 2 Monate nach Studienbeginn nicht in der Lage, 4 oder mehr Aktivitäten des täglichen Lebens auszuführen; wenn der Patient nicht befragt wurde, aber der Stellvertreter des Patienten, war der Grenzwert für die Behinderung 5 oder mehr Aktivitäten), $n = 104$ 3 – „Sickness Impact Profile total score is at least 30 2 months after study entry“ (deutsch: der Gesamtwert des Sickness Impact Profile liegt 2 Monate nach Studienbeginn bei mindestens 30), $n = 57$ 4 – „patient intubated or in coma 2 months after study entry“ (deutsch: Patienten, die 2 Monate nach Studienbeginn intubiert sind oder im Koma liegen), $n = 7$ 5 – „patient died before 2 months after study entry“ (deutsch: der Patient ist vor 2 Monaten nach Studienbeginn gestorben), $n = 320$
vlbw	218	10	Apgar Score (Score für den körperlichen Gesundheitszustand eines Neugeborenen)	9	1 – „life-threatening“ (deutsch: lebensgefährlich), $n = 33$ 2, $n = 16$ 3, $n = 19$ 4, $n = 15$ 5, $n = 25$ 6, $n = 27$ 7, $n = 35$ 8, $n = 36$ 9 – „optimal physical condition“ (deutsch: optimale körperliche Verfassung), $n = 12$
winequality	4893	11	„Wine quality score“ (deutsch: Score für die Weinqualität)	6	3 – „moderate quality“ (deutsch: moderate Qualität), $n = 20$ 4, $n = 163$ 5, $n = 1457$ 6, $n = 2198$ 7, $n = 880$ 8 – „high quality“ (deutsch: hohe Qualität), $n = 175$

Tab. 4.2.: Übersicht über die Datensätze für die reale Datenanalyse.

4.3.2. Ergebnisse

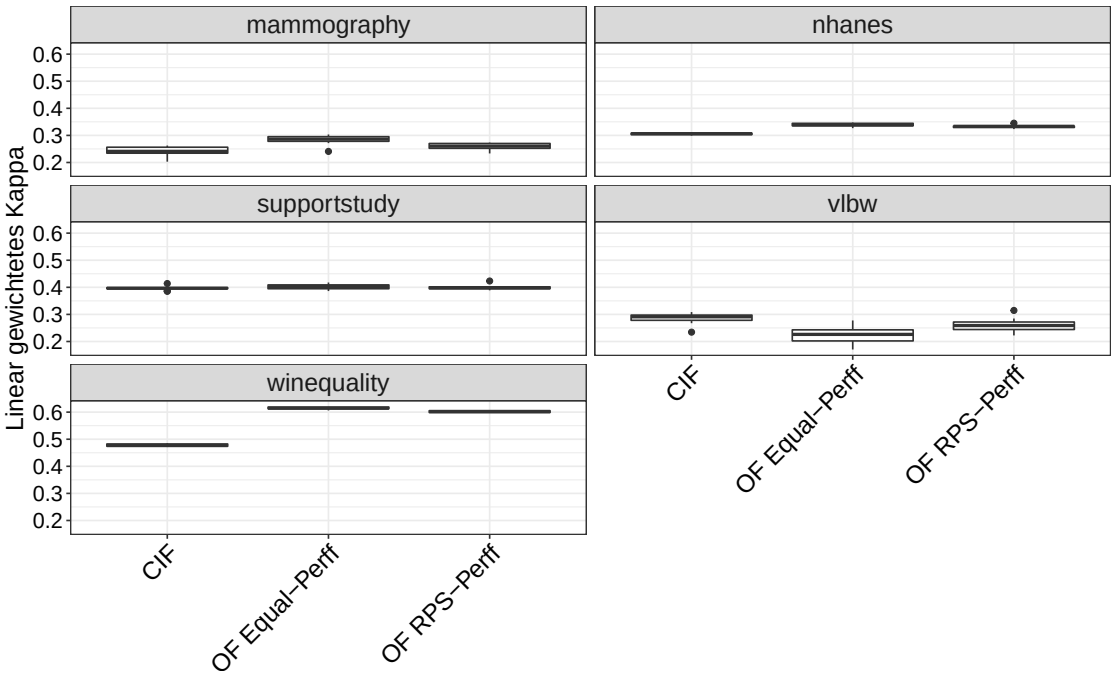

Abb. 4.13.: Werte für das linear gewichtete Kappa von den betrachteten Methoden für jeden realen Datensatz. Jeder Boxplot zeigt die Werte für die Iterationen der 10-fach stratifizierten Kreuzvalidierung.

Die Abbildung 4.13 zeigt die Werte für das linear gewichtete Kappa von den Methoden CIF, OF mit Equal-Perff und OF mit RPS-Perff basierend auf den fünf realen Datensätzen. Die entsprechenden Abbildungen mit den Werten für das quadratisch gewichtete Kappa (Abb. A.13) und Cohen's Kappa (Abb. A.14) befinden sich im Anhang dieser Arbeit. Damit liegt wie bei den simulierten Daten der Schwerpunkt zunächst auf den Ergebnissen für das linear gewichtete Kappa. Nachfolgend werden relevante Unterschiede zwischen allen drei Kappa-Varianten hervorgehoben.

Mit Ausnahme von zwei Datensätzen („supportstudy“ und „vlbw“) schneidet der OF mit Equal-Perff bezüglich des linear gewichteten Kappa (Abb. 4.13) eindeutig besser als der CIF und minimal besser als der OF mit RPS-Perff ab. Gleichzeitig ist die Leistung des OF mit RPS-Perff deutlich besser als die Leistung des CIF. In Bezug auf die beiden Ausnahmen kann beobachtet werden, dass in einem Datensatz („supportstudy“) die Leistungen von allen drei Vorhersagemethoden sehr ähnlich sind und in dem übrigen Datensatz („vlbw“) der CIF die beste Methode ist und gleichzeitig der OF mit RPS-Perff ebenso eindeutig besser als der OF mit Equal-Perff.

Die Ergebnisse für das quadratisch gewichtete Kappa in der ergänzenden Abbildung A.13 zeigen ein ähnliches Bild wie die vorherigen Ergebnisse für das linear gewichtete Kappa. Dennoch ist eine deutliche Annäherung der beiden OF-Varianten in einem Datensatz („vlbw“) auffällig. Außerdem sind die Werte für das quadratische Kappa durchweg höher als für das lineare Kappa. Dementsprechend ist hier die Verbesserung von dem OF mit Equal-Perff gegenüber dem CIF tendenziell stärker als für das linear gewichtete Kappa – außer bei dem Datensatz „vlbw“, in dem der CIF konstant am besten ist. Deshalb kann angenommen werden, dass der OF mit Equal-Perff seltener als der CIF Vorhersagen trifft, die von den wahren Klassenwerten weit entfernt liegen.

Die Ergebnisse für Cohen's Kappa in der ergänzenden Abbildung A.14 stimmen größtenteils mit den Ergebnissen für das gewichtete Kappa überein. Allerdings ist für Cohen's Kappa erstmals der OF mit RPS-Perff in einem Datensatz („nhanes“) etwas besser als die anderen Methoden. Während der OF mit Equal-Perff dennoch in den meisten Datensätzen minimal bessere Leistungen als der OF mit RPS-Perff und als der CIF erzielt, kann von dem OF mit Equal-Perff erwartet werden, etwas leistungsfähiger als die konkurrierenden Methoden in Bezug auf exakte Vorhersagen zu sein und gleichzeitig seltener als die Konkurrenz Vorhersagen zu treffen, die von den wahren Klassenwerten weit entfernt sind.

Zusammenfassend für die reale Datenanalyse erzielt der OF mit Equal-Perff überwiegend die beste Vorhersageleistung bezüglich aller Kappa-Varianten, verglichen mit der Konkurrenz. Gleichzeitig ist der OF mit RPS-Perff dem CIF häufiger überlegen. Allerdings müssen die Ergebnisse mit Vorsicht interpretiert werden, da diese auf nur fünf Datensätzen beruhen. Um wirklich belastbare Rückschlüsse ziehen zu können, wäre die Analyse basierend auf mehreren Datensätzen notwendig.

Vergleich der Ergebnisse von realen Daten mit simulierten Daten

Die Ergebnisse für die realen Daten basierend auf den fünf Datensätzen stehen in Einklang mit den Ergebnissen für die simulierten Daten von Hornung (2020) sowie Buri und Hothorn (2020). Die Tatsache, dass die simulierten Daten von Janitza et al. (2016) zu unterschiedlichen Ergebnissen für jede Klassenanzahl sowie teilweise für die verschiedenen Kappa-Varianten führen, wurde bereits für die Zusammenfassung der Ergebnisse von allen Simulationsdesigns herausgearbeitet.

5. Zusammenfassung und Ausblick

Die simulierten Daten von Janitza et al. (2016) und Hornung (2020) ergeben in Bezug auf die Qualität des Prädiktoren-Rankings, dass der CIF mit RPS-VIM meistens etwas besser als die Konkurrenz abschneidet. Die Ausnahmen sind einige Einstellungen von 9 Klassen für das Simulationsdesign von Hornung, wo der OF mit RPS-Perff und RPS-VIM häufiger etwas genauere Prädiktoren-Rankings liefert als die Konkurrenz. Diese Ergebnisse bestätigen die Vermutung, dass das RPS-VIM besser zwischen einflussreichen und nicht-einflussreichen Prädiktoren differenzieren kann, indem das VIM die inhärente Ordnung des ordinalen Response einbezieht. Allerdings entgegen dieser Vermutung ist für das Simulationsdesign von Buri und Hothorn (2020) der CIF mit ER-VIM häufiger die bessere Methode hinsichtlich der Qualität des Prädiktoren-Rankings. Darüber hinaus kann die Annahme, dass die Verwendung desselben Fehlermaßes sowohl für das VIM als auch für die Performance-Funktion des Ordinal Forest, also der OF mit RPS-Perff und RPS-VIM eine bessere Variablenselektion als der OF mit Equal-Perff und RPS-VIM gewährleistet, durchweg für die simulierten Daten von Hornung (2020) mit gleichen Klassenbreiten und teilweise für die simulierten Daten von Buri und Hothorn (2020) bestätigt werden. Dabei ist jedoch zu beachten, dass die Unterschiede in den Leistungen von beiden Vorhersagemethoden nur minimal sind.

Der Vergleich von allen betrachteten Vorhersagemethoden in Bezug auf die Klassifikationsgüte zeigt für die simulierten und realen Daten, dass der OF mit Equal-Perff meistens die genauesten Ergebnisse erzielt, im Vergleich zu den konkurrierenden Vorhersagemethoden. Die Ausnahmen treten vor allem in den Simulationen von Janitza et al. (2016) für 6 Klassen auf. Abgesehen von diesen Ausnahmen bestätigen die Ergebnisse die Annahme von einer stärkeren Leistung bezüglich einfacher Klassenvorhersagen seitens des OF mit Equal-Perff gegenüber dem OF mit RPS-Perff.

Es sei nochmals betont, dass die Ergebnisse mit Vorsicht behandelt werden müssen, da die Unterschiede zwischen den Methoden häufig nur schwach sind und fraglich ist, ob diese sehr geringen Unterschiede nicht auch auf Zufallsschwankungen zurückzuführen sind. In Bezug auf die Ergebnisse für die reale Datenanalyse muss

zusätzlich berücksichtigt werden, dass anhand von nur fünf Datensätzen keine besonders sicheren Rückschlüsse zulässig sind, weshalb in einer möglichen zukünftigen Analyse weitere Datensätze betrachtet werden sollten. Dennoch zeigt die Arbeit, dass die Ergebnisse für die verschiedenen Simulationsdesigns teilweise deutlich voneinander abweichen. Dies könnte andeuten, dass einzelnen Simulationen nicht immer vertraut werden kann, da die Ergebnisse stark von dem gewählten datengenerierenden Prozess abhängen können. Somit beweist diese Arbeit, dass mit verschiedenen Simulationsdesigns unterschiedliche Rückschlüsse folgern können.

Gegenüber der bestehenden OF-Variante mit Equal-Perff ermöglicht die neue OF-Variante mit RPS-Perff zusätzlich Wahrscheinlichkeitsvorhersagen für Klassen. Eine zukünftige Untersuchung könnte darin bestehen, die neue OF-Variante mit anderen Random Forests für ordinalen Response hinsichtlich Wahrscheinlichkeitsvorhersagen, beispielsweise anhand des Ranked Probability Score, zu vergleichen. Geeignete Random Forests für solch einen Vergleich könnten erneut CIFs oder daneben Transformation Forests für ordinalen Response darstellen.
Transformation Forests wurden von Hothorn und Zeileis (2017) eingeführt und kürzlich von Buri und Hothorn (2020) um zwei neue Varianten für die Vorhersage eines ordinalen Response erweitert. Im Gegensatz zu OFs und CIFs verwenden Transformation Forests keine Score-Werte für die geordneten Klassen des ordinalen Response, wodurch dieser nicht als stetig behandelt wird. Stattdessen verwenden diese Forests ein parametrisches Modell, zugeschnitten auf ordinale Daten.
So basieren die neuen Random Forest-Varianten im Rahmen von Transformation Forests auf dem Transformationsmodell in Gleichung (4.3), welches für das Simulationsdesign von Buri und Hothorn (2020) vorgestellt wurde, und können durch spezielle Kriterien für die Knotenaufteilung in den Bäumen Abweichungen der (nicht-) proportionalen Odds erkennen. Eine der beiden Varianten nimmt explizit den Einfluss von proportionalen Odds an und ist bei Vorhandensein dieser Art von Einfluss besonders leistungsfähig. Demgegenüber nimmt die andere Variante den Einfluss von nicht-proportionalen Odds an und kann Änderungen der Modellparameter $\boldsymbol{\vartheta}(\boldsymbol{x})$ in Gleichung (4.3) erkennen, womit diese Variante bei Vorhandensein eines variierenden Einflusses über die Klassen hinweg besonders leistungsfähig ist.

Mit kritischer Distanz den Ergebnissen gegenüber zeigt die Arbeit, dass OFs mit Equal-Perff die bessere Methode für die (Punkt-)Prognose eines ordinalen Response sind, während CIFs die geeignetere Methode für das Ranking von Prädiktoren nach ihrer Wichtigkeit für die Prognose sind.

Literaturverzeichnis

Ben-David, A. (2008). Comparison of classification accuracy using Cohen's Weighted Kappa. *Expert Systems with Applications*, 34:825–832.

Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., und Stone, C. J. (1984). *Classification and Regression Trees*. Monterey: Wadsworth International Group.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78:1–3.

Buri, M. und Hothorn, T. (2020). Model-based random forests for ordinal regression. *The International Journal of Biostatistics*, 16.

Cicchetti, D. V. und Allison, T. (1971). A new procedure for assessing reliability of scoring EEG sleep recordings. *American Journal of EEG Technology*, 11:101–110.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Cohen, J. (1968). Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70:213.

Ebert, E. (2005). WWRP/WGNE Joint Working Group on Verification. Forecast Verification – Issues, Methods and FAQ.

Epstein, E. S. (1969). A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology (1962-1982)*, 8(6):985–987.

Fleiss, J. L. und Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33:613–619.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, Seiten 1–67.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., und Hothorn, T. (2021). *mvtnorm: Multivariate Normal and t Distributions*. R package version 1.1-2.

Hanley, J. A. und McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143:29–36.

Hornung, R. (2020). Ordinal forests. *Journal of Classification*, 37:4–17.

Hornung, R. (2021). *ordinalForest: Ordinal Forests: Prediction and Variable Ranking with Ordinal Target Variables*. R package version 2.4-2.

Hothorn, T., Buehlmann, P., Dudoit, S., Molinaro, A., und Van Der Laan, M. (2006a). Survival Ensembles. *Biostatistics*, 7:355–373.

Hothorn, T., Hornik, K., und Zeileis, A. (2006b). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15:651–674.

Hothorn, T. und Zeileis, A. (2017). Transformation Forests. *arXiv preprint arXiv:1701.02110*.

Jakobsson, U. und Westergren, A. (2005). Statistical methods for assessing agreement for ordinal data. *Scandinavian Journal of Caring Sciences*, 19:427–431.

Janitza, S., Tutz, G., und Boulesteix, A.-L. (2016). Random forest for ordinal responses: prediction and variable selection. *Computational Statistics and Data Analysis*, 96:57–73.

Koch, D. (2016). Klassifikationsverfahren. In *Verbesserung von Klassifikationsverfahren*, Seiten 19–60. Wiesbaden: Springer.

Liaw, A. und Wiener, M. (2002). Classification and Regression by randomForest. *R News*, 2:18–22.

Malley, J. D., Kruppa, J., Dasgupta, A., Malley, K. G., und Ziegler, A. (2012). Probability machines. *Methods of Information in Medicine*, 51:74–81.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42:109–127.

Murphy, A. H. (1970). The ranked probability score and the probability score: A comparison. *Monthly Weather Review*, 98(12):917–924.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Refaeilzadeh, P., Tang, L., und Liu, H. (2009). Cross-Validation. *Encyclopedia of Database Systems*, 5:532–538.

Revelle, W. (2021). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois. R package version 2.1.9.

Sing, T., Sander, O., Beerenwinkel, N., und Lengauer, T. (2005). ROCR: visualizing classifier performance in R. *Bioinformatics*, 21:7881.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., und Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, 9:1–11.

Strobl, C., Boulesteix, A.-L., Zeileis, A., und Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, Sources and a Solution. *BMC Bioinformatics*, 8:1–21.

Sutton, C. D. (2005). Classification and regression trees, bagging, and boosting. *Handbook of Statistics*, 24:303–329.

Warnes, G. R., Bolker, B., und Lumley, T. (2021). *gtools: Various R Programming Tools*. R package version 3.9.2.

Warrens, M. J. (2013). Conditional inequalities between Cohen's kappa and weighted kappas. *Statistical Methodology*, 10:14–22.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer.

Wickham, H., François, R., Henry, L., und Müller, K. (2021). *dplyr: A Grammar of Data Manipulation*. R package version 1.0.7.

A. Anhang

A.1. Simulationsstudien

A.1.1. Simulierte Daten

Simulationsdesign von Janitza et al. (2016)

Anzahl an Klassen des Response	γ_{01}	γ_{02}	γ_{03}	γ_{04}	γ_{05}	γ_{06}	γ_{07}	γ_{08}	γ_{09}
$J = 3$	-1.80	1.80	∞	-	-	-	-	-	-
$J = 6$	-4.50	-1.50	0.00	1.50	4.50	∞	-	-	-
$J = 9$	-5.90	-3.41	-1.55	-0.31	0.31	1.55	3.41	5.90	∞

Tab. A.1.: Intercepts für das Proportional-Odds Modell in (4.2) mit $\gamma_{0jg} = \gamma_{0j}$.

Mischungs-komponente	Koeffizientenvektor $\boldsymbol{\gamma}_g^T = (\gamma_{g,1}, ..., \gamma_{g,65})$																
$g = 1$	$(1$	1	1	1	1	0.75	0.75	0.75	0.75	0.75	0.5	0.5	0.5	0.5	0.5	0	$0, ..., 0)$
$g = 2$	$(1$	1	-1	-1	0	1	1	-1	-1	0	1	1	-1	-1	0	0	$0, ..., 0)$

Tab. A.2.: Effekte von den Prädiktoren für das Proportional-Odds Modell in (4.2) für die Mischungskomponenten $g = 1, 2$.

A.1.2. Ergebnisse

Kappa-Werte für die Simulationen von Janitza et al.

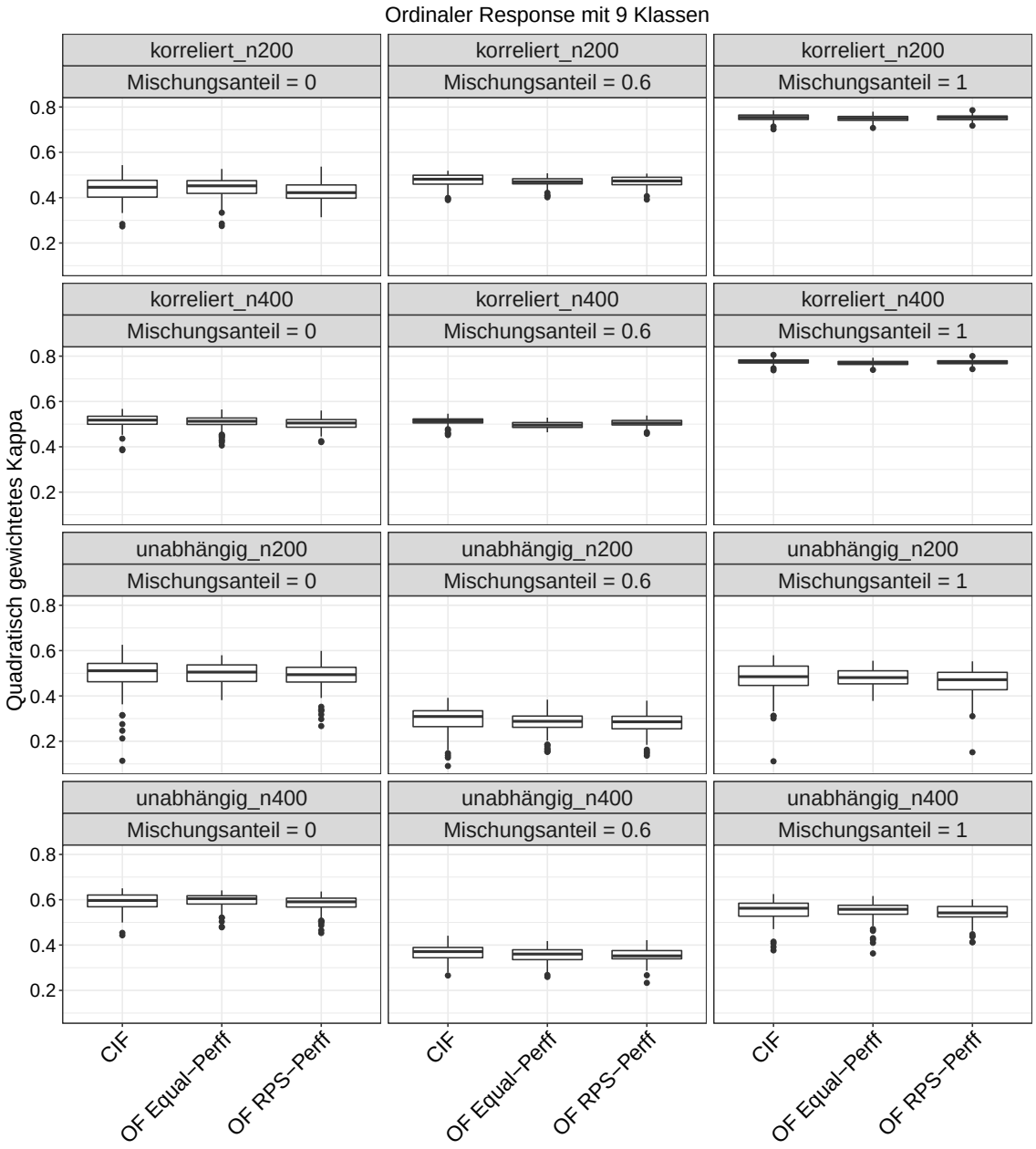

Abb. A.1.: Werte für das quadratisch gewichtete Kappa von den betrachteten Methoden in jedem Setting für einen ordinalen Response mit 9 Klassen basierend auf den Simulationen von Janitza et al. (2016). Jeder Boxplot zeigt die Werte für die 100 Iterationen.

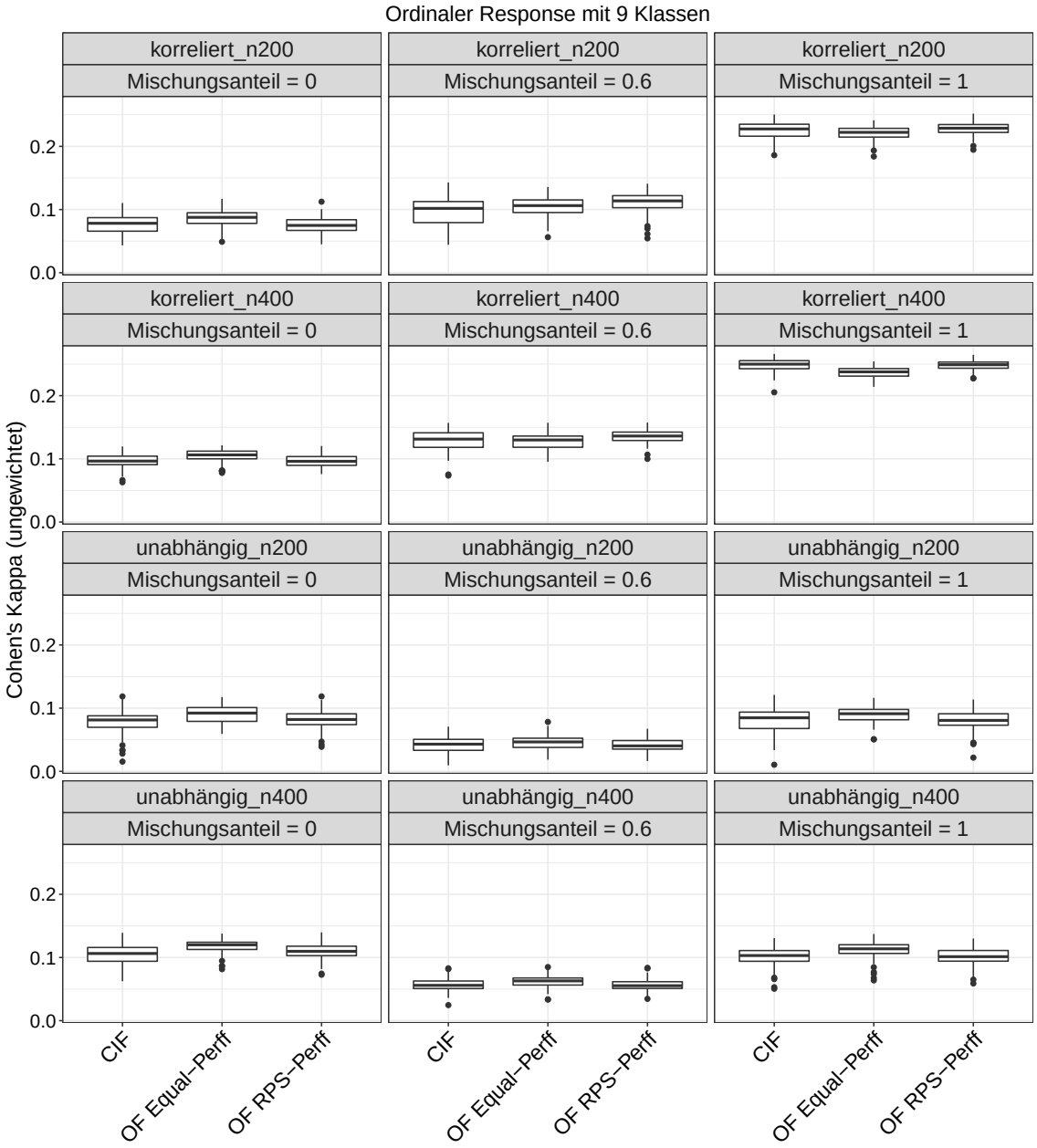

Abb. A.2.: Werte für Cohen's Kappa von den betrachteten Methoden in jedem Setting für einen ordinalen Response mit 9 Klassen basierend auf den Simulationen von Janitza et al. (2016). Jeder Boxplot zeigt die Werte für die 100 Iterationen.

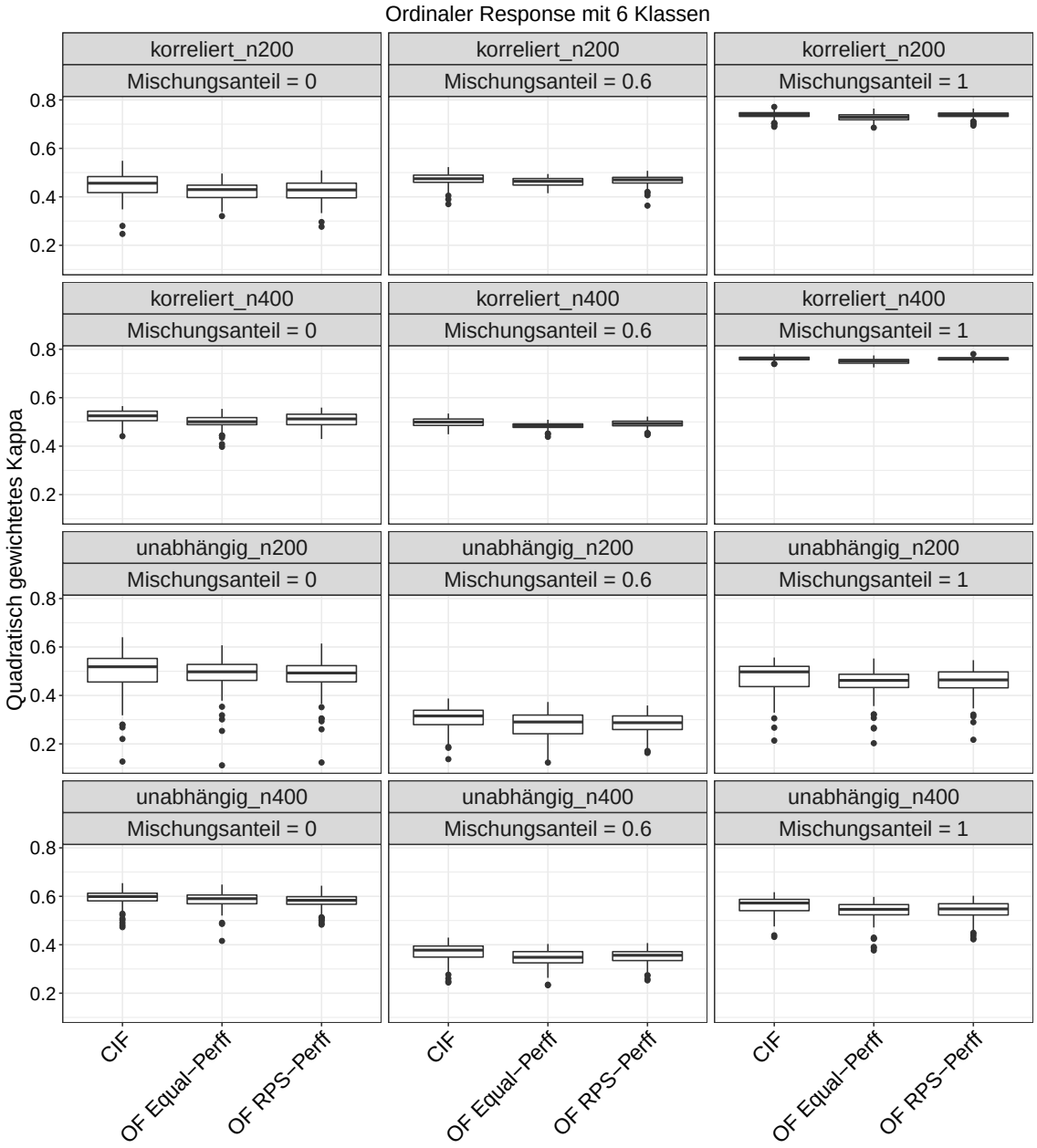

Abb. A.3.: Werte für das quadratisch gewichtete Kappa von den betrachteten Methoden in jedem Setting für einen ordinalen Response mit 6 Klassen basierend auf den Simulationen von Janitza et al. (2016). Jeder Boxplot zeigt die Werte für die 100 Iterationen.

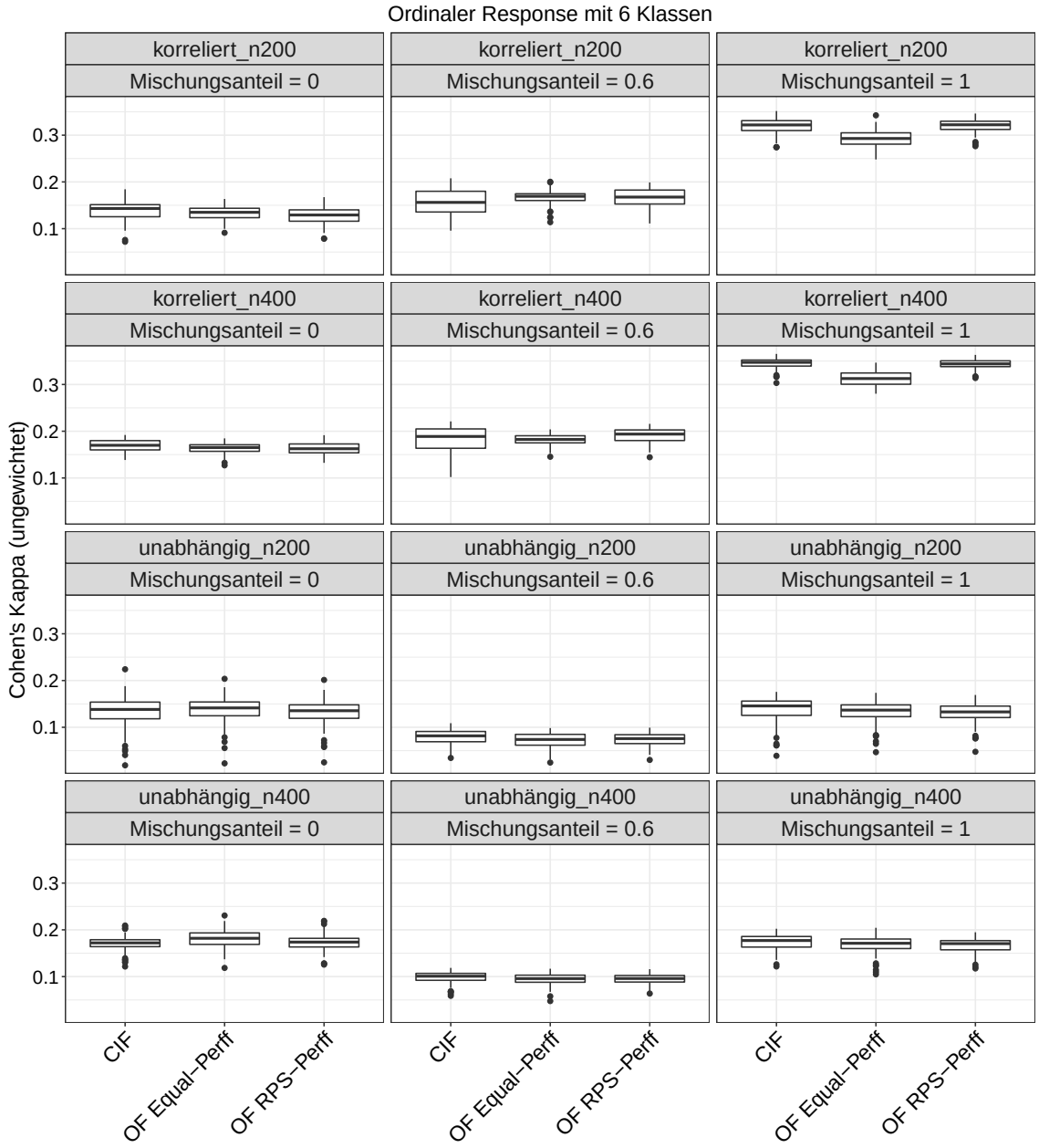

Abb. A.4.: Werte für Cohen's Kappa von den betrachteten Methoden in jedem Setting für einen ordinalen Response mit 6 Klassen basierend auf den Simulationen von Janitza et al. (2016). Jeder Boxplot zeigt die Werte für die 100 Iterationen.

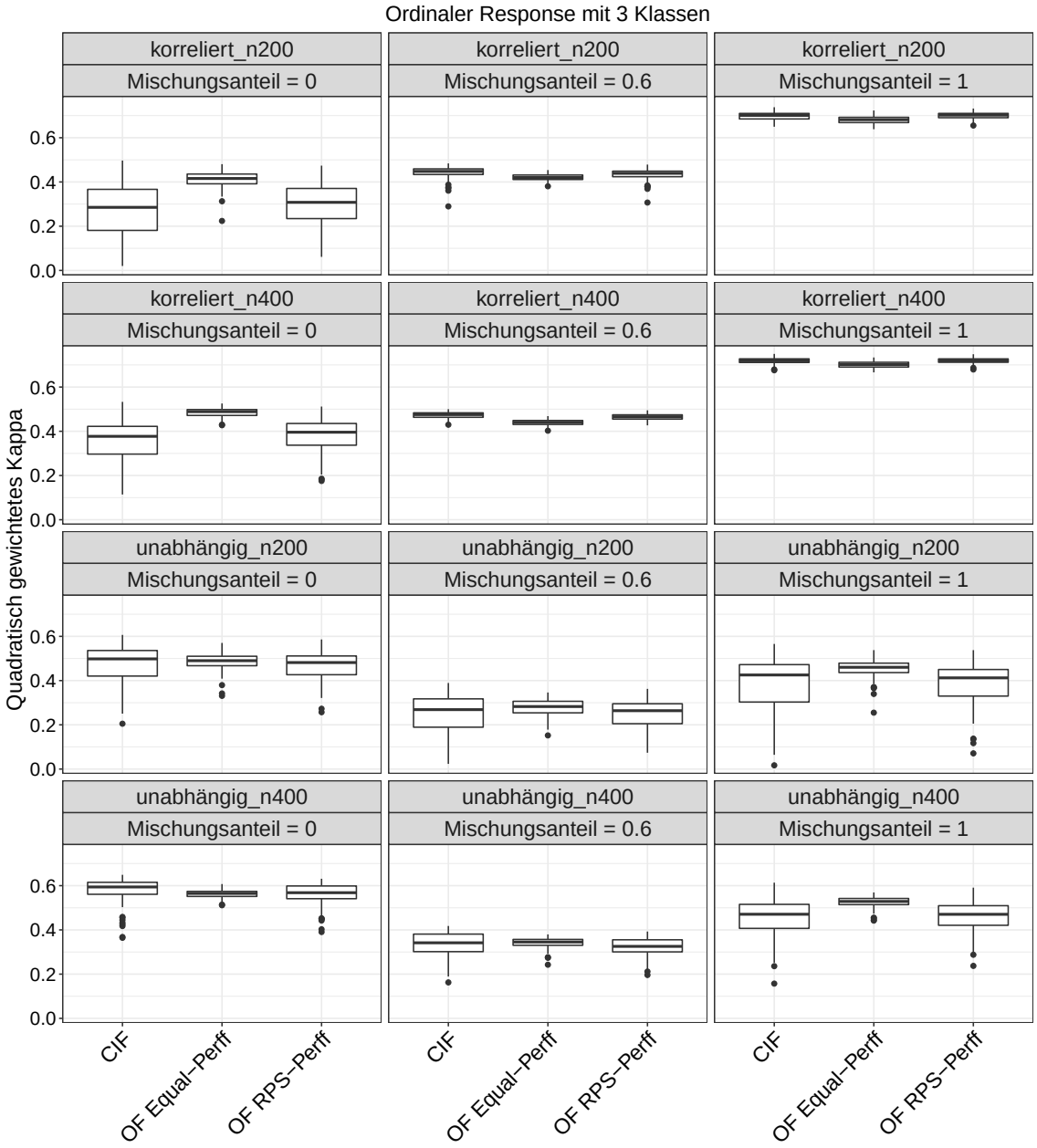

Abb. A.5.: Werte für das quadratisch gewichtete Kappa von den betrachteten Methoden in jedem Setting für einen ordinalen Response mit 3 Klassen basierend auf den Simulationen von Janitza et al. (2016). Jeder Boxplot zeigt die Werte für die 100 Iterationen.

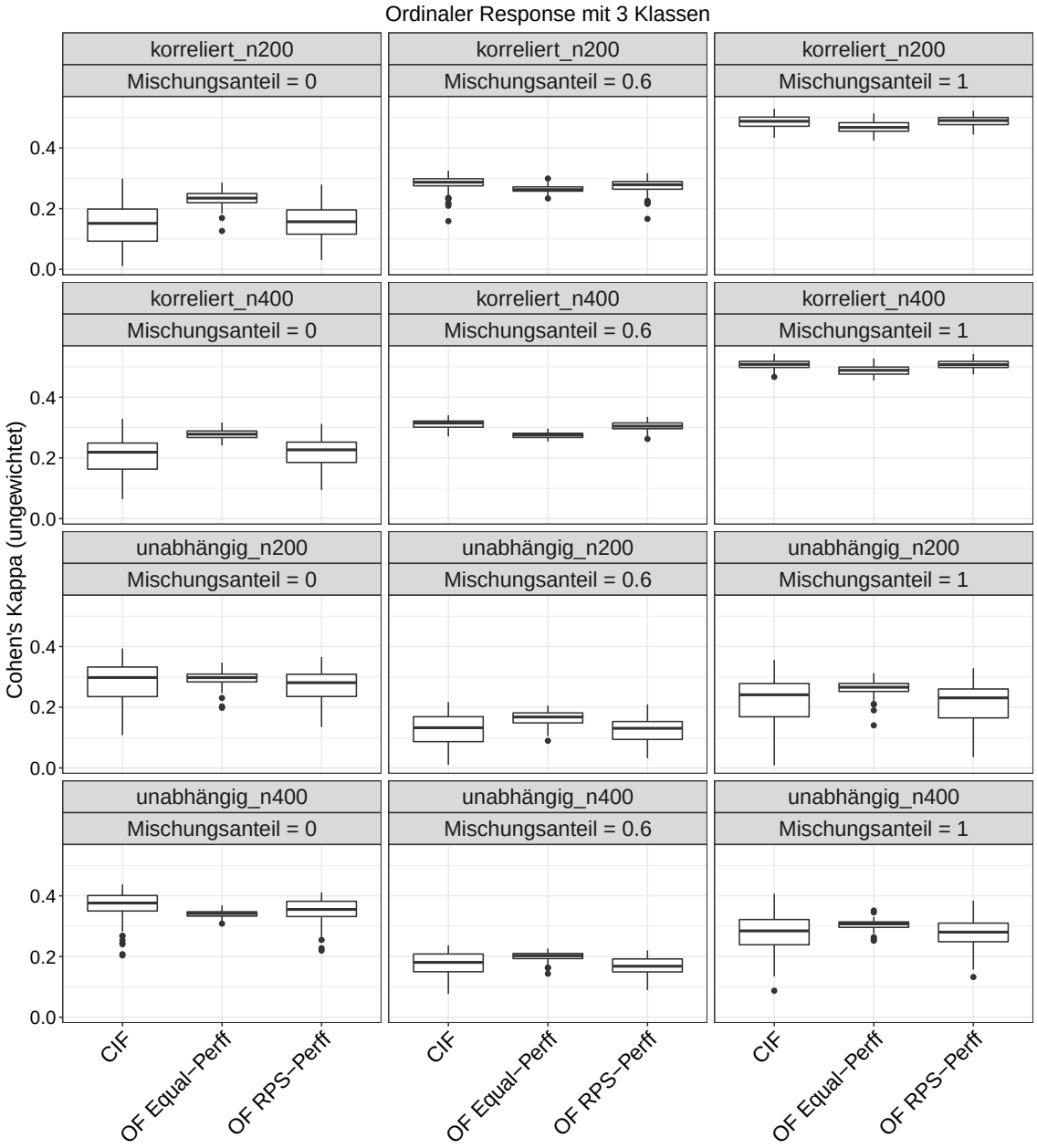

Abb. A.6.: Werte für Cohen's Kappa von den betrachteten Methoden in jedem Setting für einen ordinalen Response mit 3 Klassen basierend auf den Simulationen von Janitza et al. (2016). Jeder Boxplot zeigt die Werte für die 100 Iterationen.

Kappa-Werte für die Simulationen von Hornung

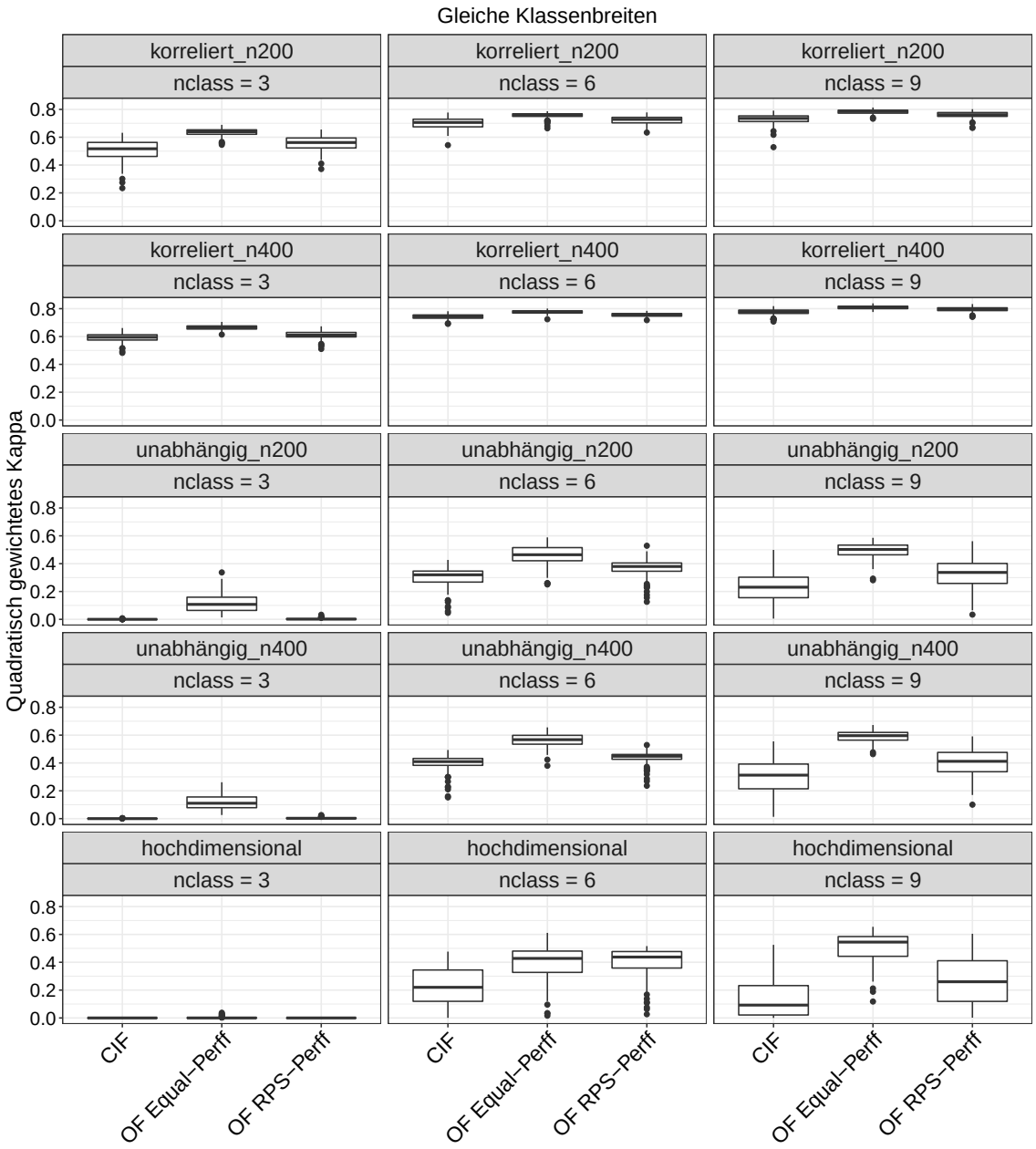

Abb. A.7.: Werte für das quadratisch gewichtete Kappa von den betrachteten Methoden in jedem Setting mit gleichen Klassenbreiten basierend auf den Simulationen von Hornung (2020). Jeder Boxplot zeigt die Werte für die 100 Iterationen.

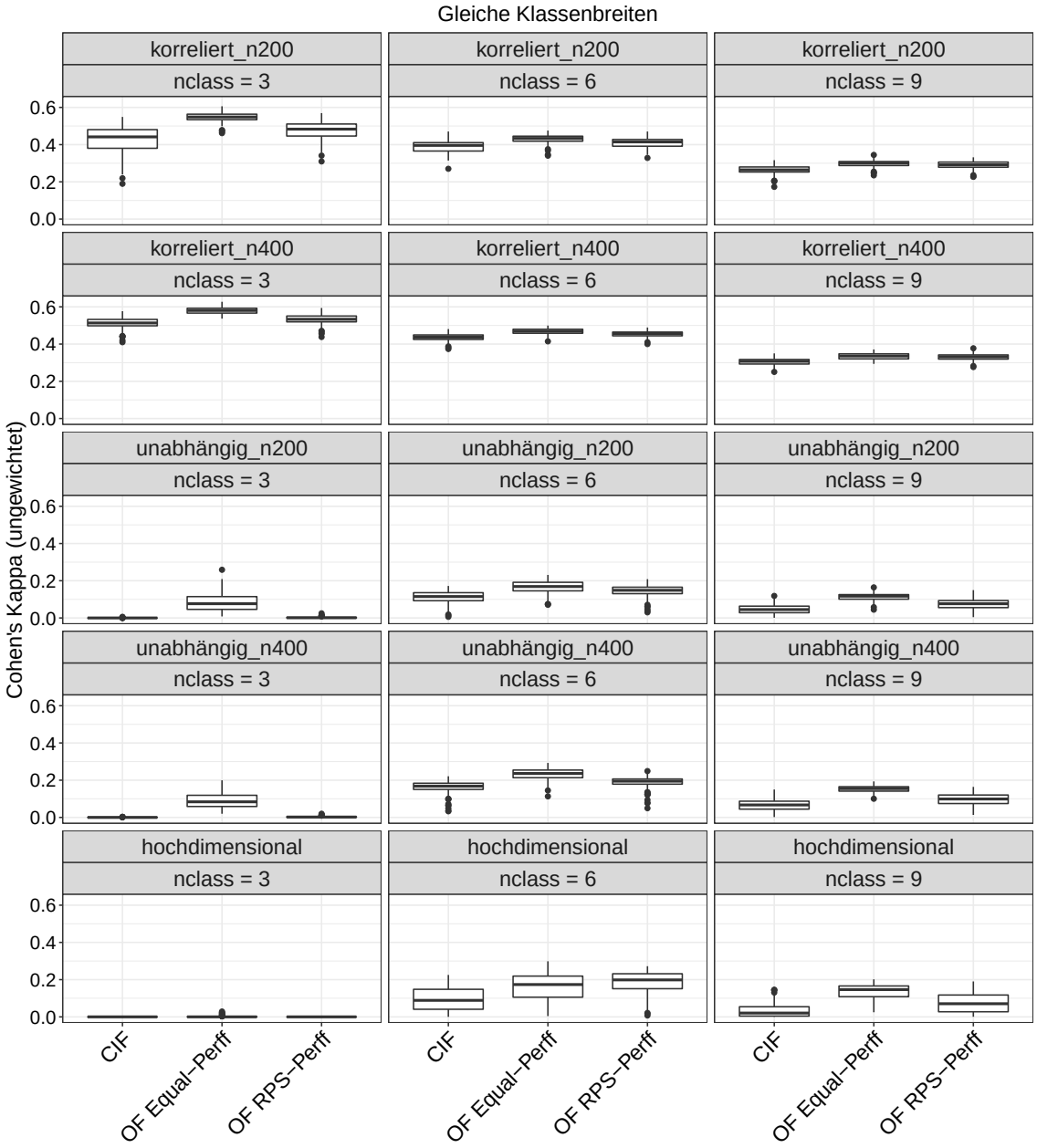

Abb. A.8.: Werte für Cohen's Kappa von den betrachteten Methoden in jedem Setting mit gleichen Klassenbreiten basierend auf den Simulationen von Hornung (2020). Jeder Boxplot zeigt die Werte für die 100 Iterationen.

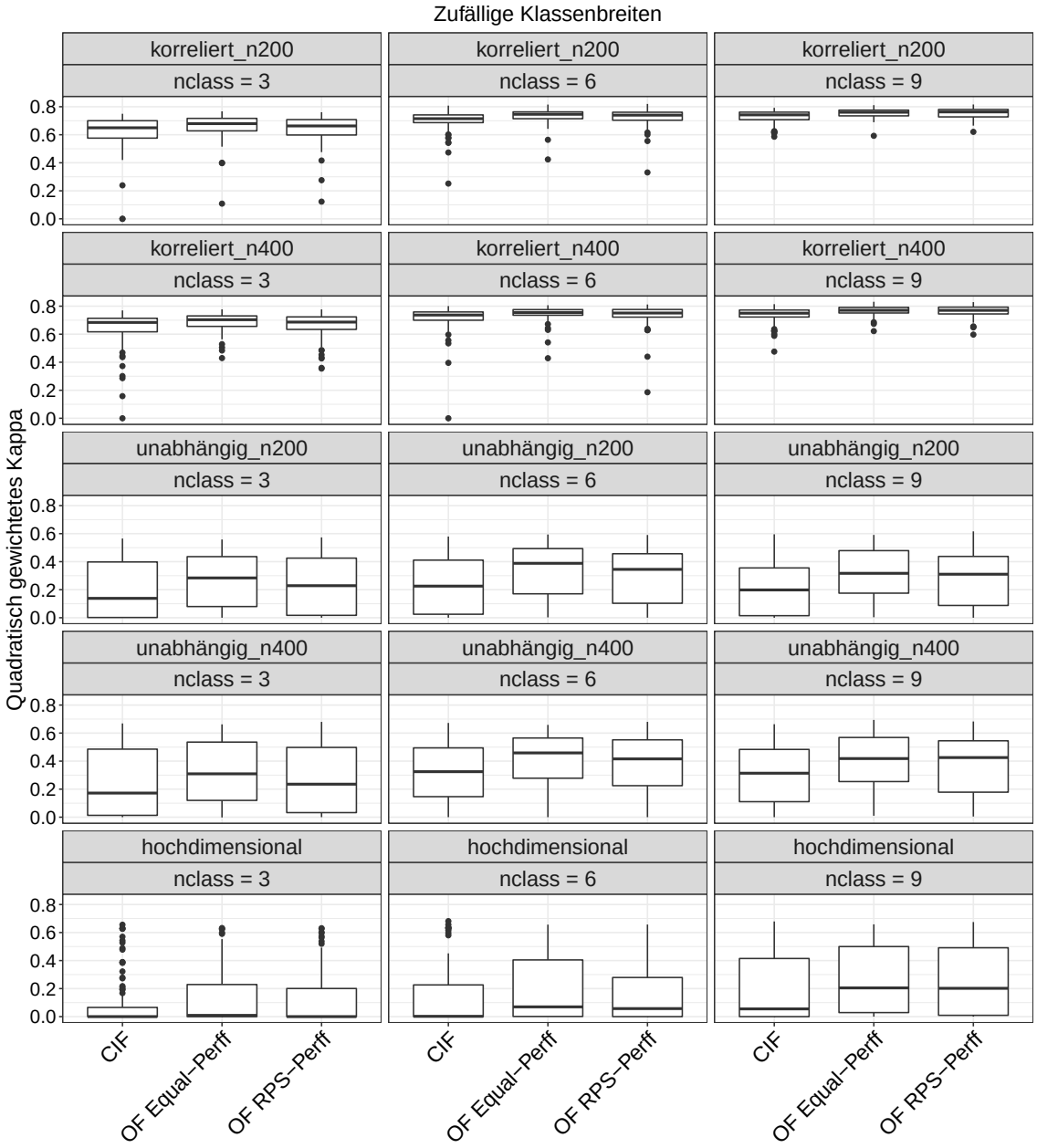

Abb. A.9.: Werte für das quadratisch gewichtete Kappa von den betrachteten Methoden in jedem Setting mit zufälligen Klassenbreiten basierend auf den Simulationen von Hornung (2020). Jeder Boxplot zeigt die Werte für die 100 Iterationen.

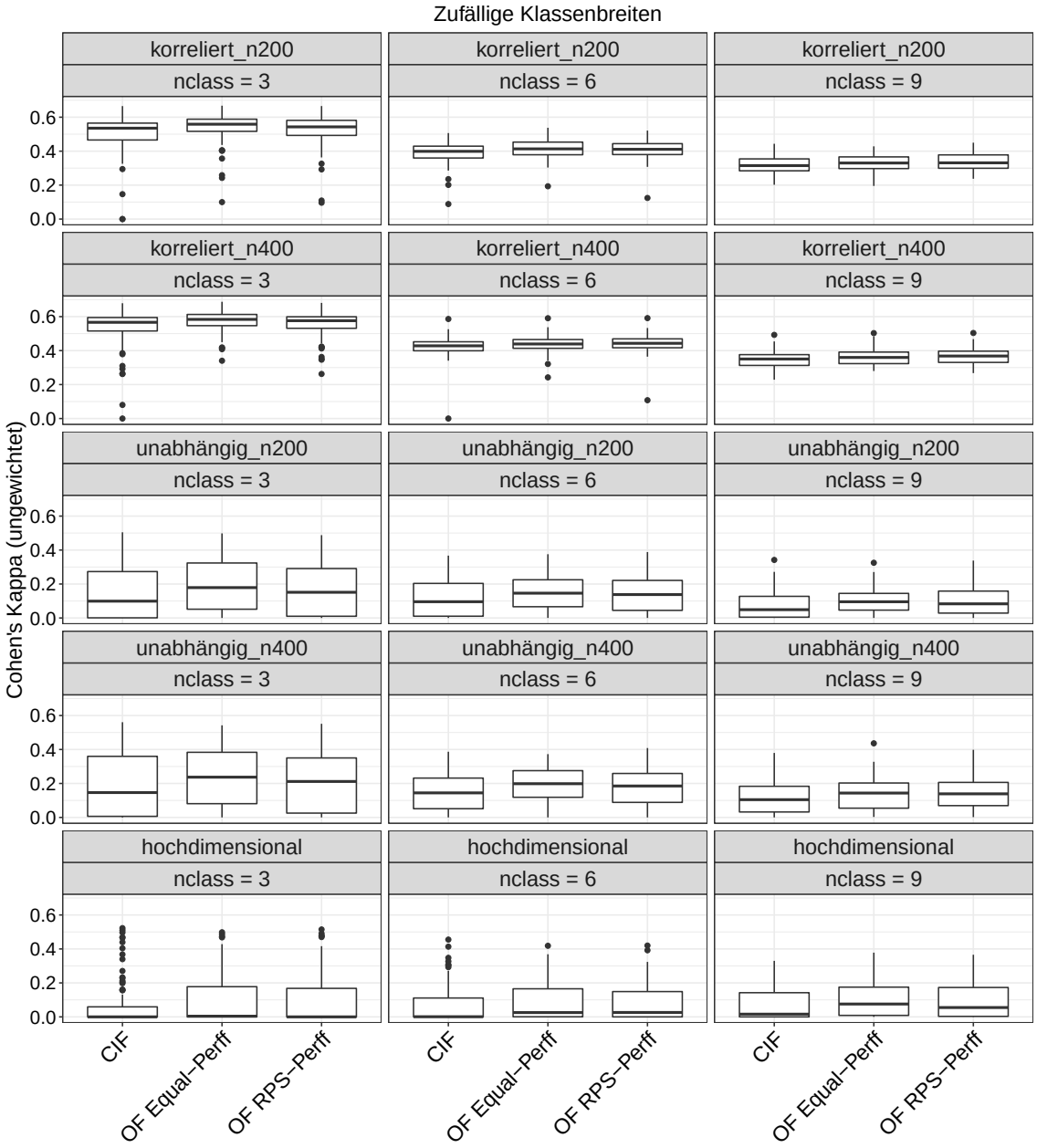

Abb. A.10.: Werte für Cohen's Kappa von den betrachteten Methoden in jedem Setting mit zufälligen Klassenbreiten basierend auf den Simulationen von Hornung (2020). Jeder Boxplot zeigt die Werte für die 100 Iterationen.

Kappa-Werte für die Simulationen von Buri und Hothorn

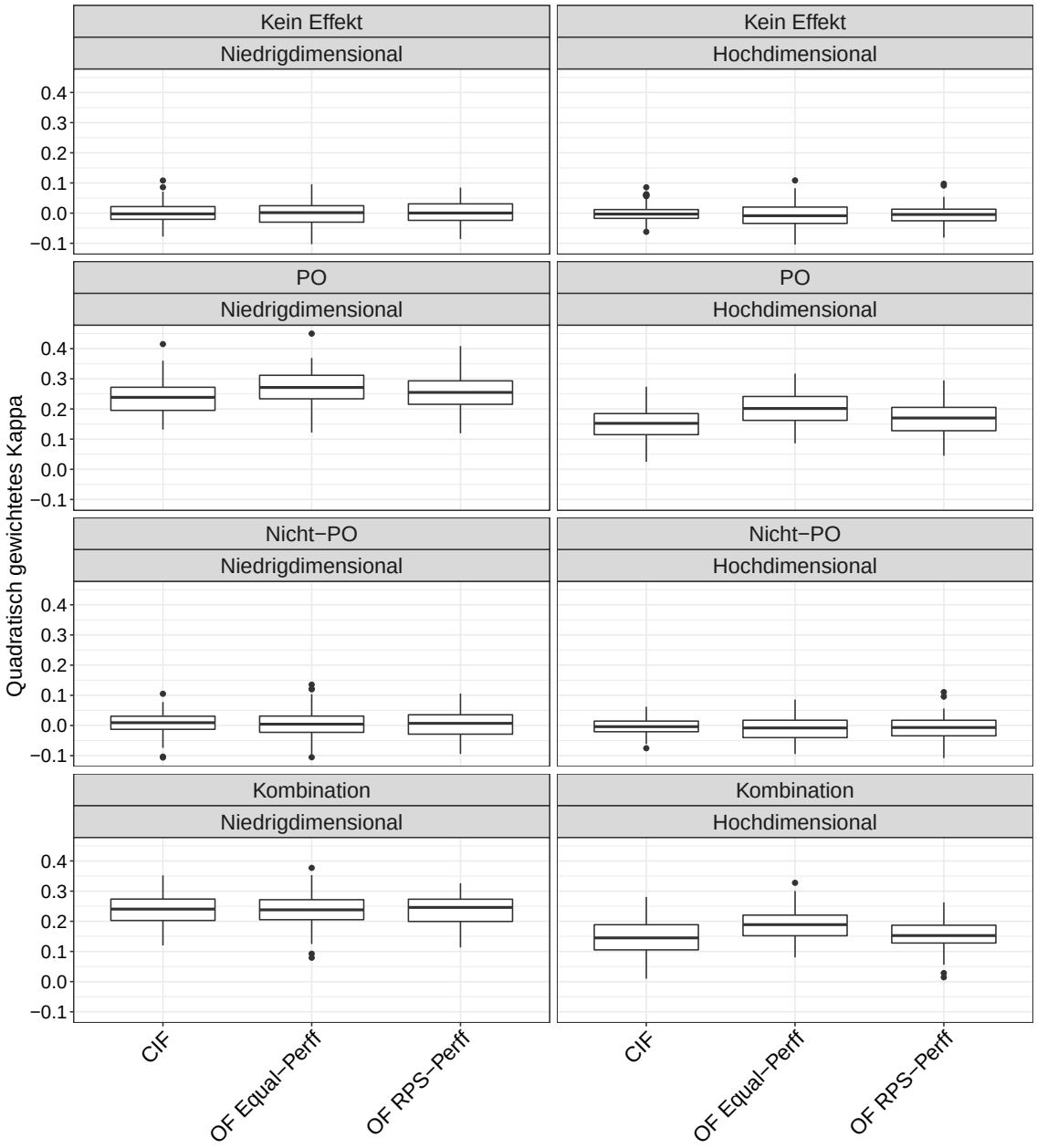

Abb. A.11.: Werte für das quadratisch gewichtete Kappa von den betrachteten Methoden in jedem Setting basierend auf den Simulationen von Buri und Hothorn (2020). Jeder Boxplot zeigt die Werte für die 100 Iterationen.

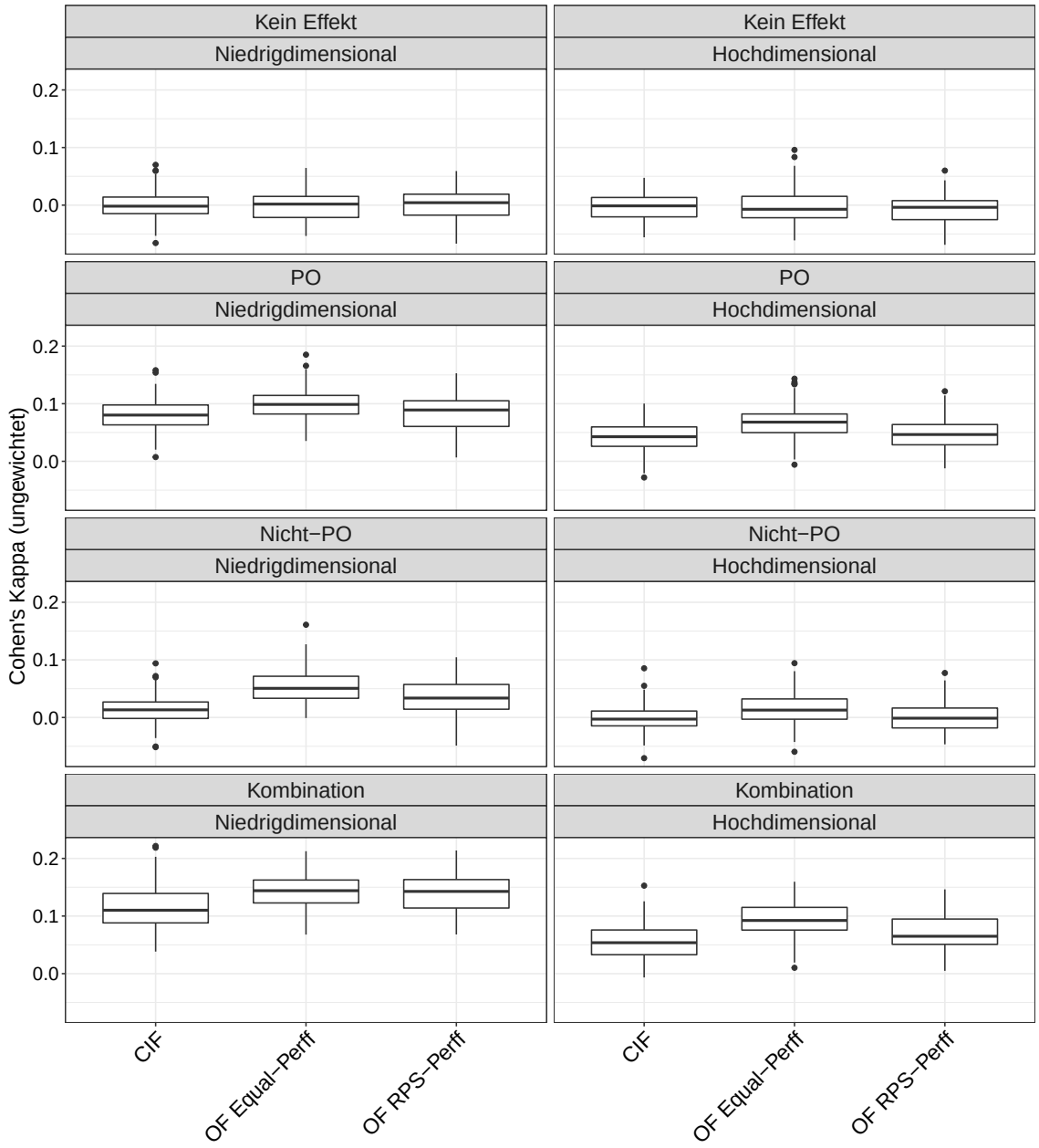

Abb. A.12.: Werte für Cohen's Kappa von den betrachteten Methoden in jedem Setting basierend auf den Simulationen von Buri und Hothorn (2020). Jeder Boxplot zeigt die Werte für die 100 Iterationen.

A.2. Ergebnisse der realen Datenanalyse

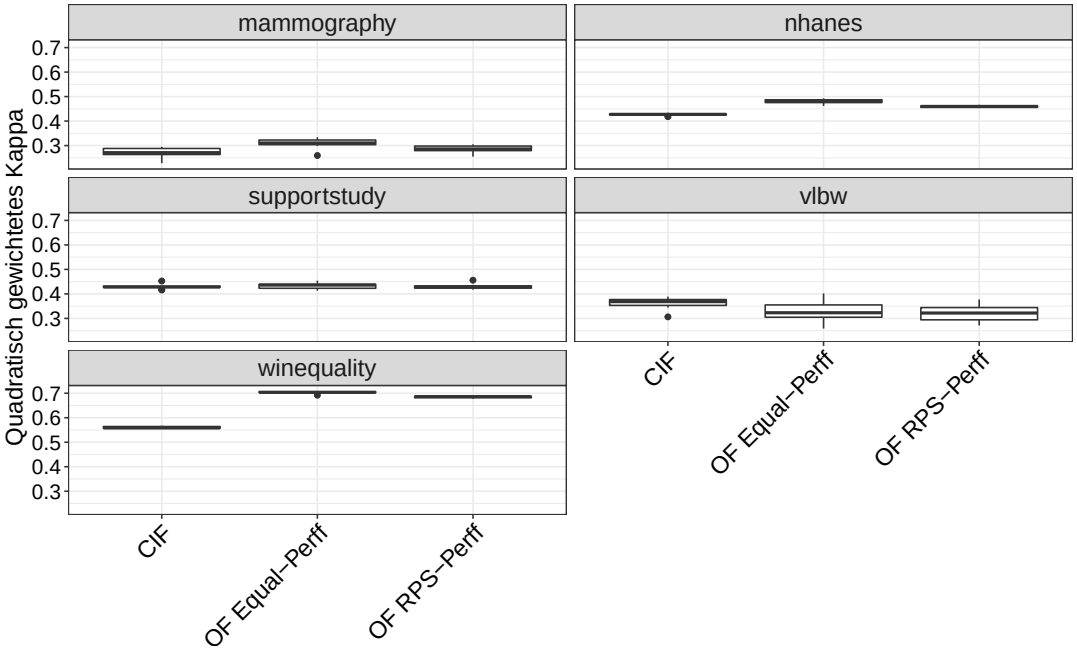

Abb. A.13.: Werte für das quadratisch gewichtete Kappa von den betrachteten Methoden für jeden realen Datensatz. Jeder Boxplot zeigt die Werte für die Iterationen der 10-fach stratifizierten Kreuzvalidierung.

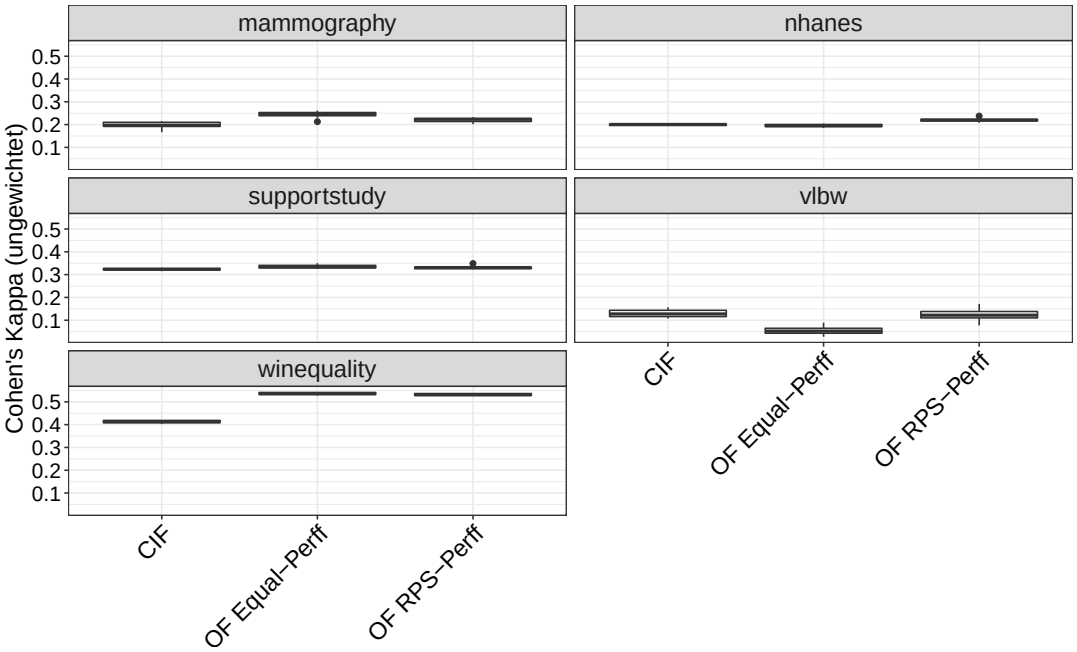

Abb. A.14.: Werte für Cohen's Kappa von den betrachteten Methoden für jeden realen Datensatz. Jeder Boxplot zeigt die Werte für die Iterationen der 10-fach stratifizierten Kreuzvalidierung.

B. Elektronischer Anhang

Die R-Dateien und Ergebnisse für die Analysen in dieser Arbeit sind in dem Ordner „RandomForest_ordinalerResponse“ enthalten, der auf dem beigefügten Datenstick gespeichert ist. Dieser Ordner ist wie folgt gegliedert:

- „Masterarbeit_JasminWulf.pdf“: Hierbei handelt es sich um die elektronische Version dieser Arbeit.
- „RealeDatenanalyse“: Dieser Ordner enthält entsprechende Unterordner mit den vorverarbeiteten Datensätzen, R-Codes, Ergebnissen und Grafiken für die reale Datenanalyse.
- „Simulationsstudien“: In diesem Ordner befinden sich weitere Unterordner für die einzelnen Simulationsstudien. Alle Unterordner sind in gleicher Weise aufgebaut und enthalten die entsprechenden R-Codes, Ergebnisse und Grafiken für die Simulationsstudien.

Detaillierte Anweisungen befinden sich in der README-Datei, die ebenso Inhalt des Ordners „RandomForest_ordinalerResponse“ ist.

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die vorliegende Masterarbeit selbstständig angefertigt habe. Es wurden nur die in der Arbeit ausdrücklich benannten Quellen und Hilfsmittel benutzt. Wörtlich oder sinngemäß übernommenes Gedankengut habe ich als solches kenntlich gemacht.

Ort, Datum	Unterschrift