LUDWIG-MAXIMILIANS-UNIVERSITÄT
MÜNCHEN

Institut für Statistik

---

# Eine empirische Studie zum Einfluss von Ausreißern auf Resampling-basierte Variablenselektion bei multipler Regression

---

## Bachelorarbeit


Eingereicht von: ALMA SEHIC

Betreuerin: PROF. DR. ANNE-LAURE BOULESTEIX

MÜNCHEN, DEN 04.08.2015

# Abstract

Im vergangenen Jahr wurden Resampling-Methoden auf ihre Modellstabilität über Resampling-basierte Variablenselektion untersucht. Aufgrund der Tatsache, dass in diesem Zusammenhang der Bezug zu möglichen Ausreißern fehlte, soll diese Bachelorarbeit untersuchen, inwiefern sich das Ergebnis durch die Präsenz von Ausreißern verändert. Diese Arbeit befasst sich unter anderem mit der Aufgabe modifizierte Datensätze zu generieren, die dem realen Datensatz ähneln, aber Ausreißer für verschiedene Szenarien enthalten. Durch Simulation haben die modifizierten Datensätze sowie der original Datensatz eine Resampling-basierte Variablenselektion durchlaufen, die zwischen den Resampling-Methoden Bootstrap und Subsampling differenziert. Dabei handelt es sich, um eine Rückwärtsselektion mittels $BIC$ in einem multiplen linearen Regressionsmodell. Mit Beendigung der Simulation wurde der Gini-purity und die relativen Inklusions-Häufigkeiten der Variablen für jeden Datensatz berechnet. Diese Ergebnisse wurden schließlich hergenommen, um den Einfluss der Ausreißer zu untersuchen und Informationen über die Modellstabilität zu erhalten, insbesondere im Bezug zu den Resampling-Methoden. Diese Analysen liefern teils überraschende Ergebnisse. Die Annahme, dass sich die Ausreißer negativ auf die Modellstabilität mittels Bootstrap-Verfahren auswirken würden, konnte nicht bestätigt werden. Die Ausreißer tragen entscheidend zu der Stabilität der Modellselektion bei. Insbesondere bewirken sie, dass andere Variablen ins Modell selektiert werden, als es beim original Datensatz der Fall ist.

# Inhaltsverzeichnis

# Abbildungsverzeichnis

# Tabellenverzeichnis

# 1. Einleitung

Bei Datenerhebungen kommt es in vielen Fällen vor, dass sich unter den Beobachtungen auffällig kleine bzw. große Werte befinden. Insbesondere wenn diese als unrealistisch empfunden werden, werden sie häufig aus dem Datensatz entfernt. Denn sie würden aufgrund ihrer hohen Auswirkung die Richtigkeit der Daten in Frage stellen. Solche Werte werden in der Statistik als *Ausreißer* bezeichnet, wenngleich eine präzise Definition des Begriffs nicht existiert. (Fahrmeir, Kneib & Lang, 2009, S. 173)

In der vorliegenden Arbeit sind Datensätze durch Simulation für unterschiedliche Szenarien generiert worden. Ausgehend von einem realen Teildatensatz wurden je Szenario zufällig Beobachtungen ausgewählt, die durch berrechnete Ausreißer-Werte ersetzt worden sind. Das bedeutet, dass absichtlich Ausreißer-Werte in unterschiedlichsten Stärken und Mengen in die Variablen eingefügt worden sind und dass dadurch viele unterschiedliche modifizierte Datensätze entstanden sind. Damit schließlich eine Aussage über den Einfluss von Ausreißern auf Resampling-basierte Variablenselektion bei multipler Regression getroffen werden konnte, sollten die modifizierte Datensätze den gleichen Prozess durchlaufen wie der original Datensatz. Auf diese Weise konnten die Ergebnisse des ursprünglichen Datensatzes mit denen der modifizierten Datensätze verglichen werden und der Einfluss der Ausreißer veranschaulicht werden.

Diese Bachelorarbeit baut auf den Forschungen von De Bin, Janitza, Sauerbrei und Boulesteix (2014) auf. Die Forschungen dienten der Untersuchung der beiden Resampling-Methoden *Bootstrap* und *Subsampling* auf Resampling-basierte Variablenselektion bei multivariabler Regression. Für diese Analyse wurde unter anderem derselbe Datensatz hergenommen, auf den sich auch meine Arbeit bezieht (De Bin et al., 2014, S. 4).
Zum einen wurde die relative Inklusions-Häufigkeit der Variablen untersucht, die nach Resampling-basierter Variablenselektion je Resampling-Methode berrechnet wurde. Daraus wurden sowohl Schlüsse bezüglich der Modellstabilität, als auch über die Bedeutung der Variablen im Modell gezogen (De Bin et al., 2014, S. 1).
Zum anderen konnten die Ergebnisse dieser Analyse auch durch die Untersuchung des $AUC$ bestätigt werden (De Bin et al., 2014, S. 23). Aus den Ergebnissen dieser Simulationsstudie kam unter anderem hervor, dass die definierten Störvariablen eine verhältnismäßig hohe relative Inklusions-Häufigkeit bei dem Bootstrap-Verfahren hatten, während bei

der Subsampling-Methode bessere Modelle ausgewählt wurden, dadurch dass häufiger die relevanten Variablen selektiert wurden (De Bin et al., 2014, S. 23). In dem Zusammenhang der Analysen von De Bin et al. (2014) erwies sich die Subsampling-Methode vorteilhafter als die Bootstrap-Methode (De Bin et al., 2014, S. 1).

Das Ziel meiner Bachelorarbeit ist es zum einen den Einfluss von Ausreißern auf die Stabilität der Modellselektion zu quantifizieren und zum anderen zu untersuchen inwiefern sich die Ausreißer auf die Ergebnisse der Bootstrap-Methode und die der Subsampling-Methode auswirken werden. Die Kriterien anhand dessen die Ergebnisse der Resampling-basierten Variablenselektion hinsichtlich der Auswirkungen der Ausreißer auf die Modellstabilität verglichen worden sind, sind wiederum die relativen Inklusions-Häufigkeiten der Variablen und außerdem der Gini-purity.

Im folgenden Kapitel wird kurz auf die ursprünglichen Daten der empirischen Studie eingegangen. Die darauffolgenden drei Kapitel sollen die wesentlichen Eigenschaften der multiplen Regression, der Resampling-Methoden und den Prozess der Variablenselekektion erläutern. Anschließend wird in Kapitel 6 erklärt wie die modifizierten Datensätze erzeugt wurden. Dies beinhaltet die Generierung der Ausreißer und die Unterscheidung der Szenarien. Das Kapitel 7 beschäftigt sich mit der Definition der Vergleichskriterien. Insbesondere soll geklärt werden wie die Ergebnisse entstanden sind, die schließlich in Kapitel 8 veranschaulicht werden und zum Vergleich der Methoden bezüglich der Auswirkung der Ausreißer dienen. Das Kapitel 8 ist wesentlicher Hauptteil meiner Arbeit und umfasst den Einfluss von Ausreißern für unterschiedliche Szenarien auf die Modellstabilität der Resampling-Methoden, die nach dem Gini-purity und den relativen Inklusions-Häufigkeiten beurteilt werden. Zum Schluss wird über die erhaltenen Ergebnisse disskutiert und es werden alternativen Vorgeschlagen die eventuell zu anderen bzw. eindeutigeren Ergebnissen geführt hätten.

# 2. Original Datensatz

Die behandelten Daten basieren auf einer empirischen Studie von Ihorst et al. (2004). Ursprünglich diente die Studie dazu mittel- und langfristige Auswirkungen des Ozons auf die Lungenfunktion von Schulkindern zu untersuchen (De Bin et al., 2014, S. 4). Unteranderem auf das Atemvolumen, dass nach maximaler Einatmung eines Kindes forciert ausgeatmet wurde, die sogenannte *forcierte Vitalkapazität* (FVC) (Kellner, 2009). Im Folgenden wird ein Teildatensatz *"forced vital capacity in autumn 1997"* dieser empirischen Studie behandelt, der deutlich weniger Kinder einbezieht (De Bin et al., 2014, S. 4).

Explizit wurden für diesen original Datensatz *Ozon* 496 Schulkinder einbezogen. Außer den gemessenen Werten des gesundheitlichen Zustandes wurden auch spezifische Merkmale der Kinder erhoben, wie Alter und Geschlecht. Insgesamt enthält der Datensatz für $25$ Merkmale $n = 496$ Beobachtungen. Diese Merkmale sind in der Tabelle 2 dargestellt. Die Bedeutung dieser Variablen ist für die folgenden Untersuchungen unrelevant. Die $25$ Merkmale setzen sich aus einer Zielvariablen $Y$ und den $p = 24$ erklärenden Variablen zusammen.

In den Forschungsberichten von De Bin et al. (2014) wurden die Variablen SEX, FLGROSS und FLGEW, aufgrund ihrers starken Effekts, als *Core*-Variablen definiert (De Bin et al., 2014, S. 12). Alle anderen Kovariablen wurden daher in dieser Arbeit als *Non-Core*-Variablen bezeichnet.

| **Typ** | **Variable** | **Skalierung** |
|---|---|---|
| Response $Y$ | FFVC | metrisch |
| *Core*-Variablen | SEX | binär |
| | FLGROSS | metrisch |
| | FLGEW | metrisch |
| *Non-Core*-Variablen | ALTER | metrisch |
| | AGEBGEW | metrisch |
| | FNOH24 | metrisch |
| | FO3H24 | metrisch |
| | FTEH24 | metrisch |
| | ADHEU | metrisch |
| | HOCHOZON | binär |
| | AMATOP | binär |
| | AVATOP | binär |
| | ADEKZ | binär |
| | ARAUCH | binär |
| | FSNIGHT | binär |
| | FMILB | binär |
| | FTIER | binär |
| | FPOLL | binär |
| | FLTOTMED | binär |
| | FSPT | binär |
| | FSATEM | binär |
| | FSAUGE | binär |
| | FSPFEI | binär |
| | FSHLAUF | binär |

Tabelle 2.1.: Variablen des original Datensatzes *Ozon*

# 3. Multiple lineare Regression

Bei dieser empirischen Studie handelt es sich um eine multiple lineare Regression. Die einfache lineare Regression gilt als Spezialfall der multiplen Regression. Im Gegensatz zum linearen Regressionsmodell tragen bei der multiplen Regression mehrere Einflussgrößen $X_1, \ldots, X_p$ zur Erklärung des Response $Y$ bei. Eine multiple lineare Regression kann angewendet werden, falls: (Vgl.: Groß, 2010, S. 205)

- bei $p$ Kovariablen $X_1, \ldots, X_p$ und einem metrischen Response $Y$
  $n$ Beobachtungstupel $(x_{i1}, \ldots, x_{ip}, y_i)$, $i = 1, \ldots, n$ gegeben sind,
- für $k$ Regressionskoeffizienten $n > k$ gilt und
- jede Variable $X_j$ sich nicht als Linearkomination anderer Variablen im Modell bilden lässt.

Wobei $x_{ij}$ die Beobachtungen der Kovariablen $X_j$, $j = 1, \ldots, p$ (Vgl.: Groß, 2010, S. 205) und $y_i$ die Beobachtungen des Response $Y$ sind (Vgl.: Groß, 2010, S. 191).

## 3.1. Definition

Das multiple lineare Regressionsmodell ist, mit oben genannten Annahmen, gegeben durch: (Vgl.: Fahrmeir et al., 2009, S. 24)

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i \,, \qquad i = 1, \ldots, n \,.$$

Die Fehlerterme $\epsilon_1, \ldots, \epsilon_n$ sind dabei unabhängig und identisch verteilt mit $E(\epsilon_i) = \mathbf{0}$ und $Var\left(\sigma^2\right) = \mathbf{0}$ (Vgl.: Fahrmeir et al., 2009, S.21). Die abhängigen Variablen sind bei gegebenen Kovariablenwerten unter der Normalverteilungsannahme (bedingt) unabhängig und normalverteilt: (Vgl.: Fahrmeir et al., 2009, S. 24f.)

$$y_i \sim N\left(\mu_i, \sigma^2\right) \,,$$

dabei ist

$$\mu_i = E(y_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} \,, \qquad i = 1, \ldots, n \,.$$

Bei einem multiplen linearen Regressionsmodell mit Intercept ist die Anzahl der Regressionskoeffizienten $k = p + 1$ und es gilt: (Vgl.: Groß, 2010, S. 206)

$$\underbrace{\begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix}}_{\mu} = \underbrace{\begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix}}_{X} \cdot \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}}_{\beta} ,$$

mit bekannter Modellmatrix $X$ der Dimension $(n \times k)$ und unbekanntem Vektor $\beta$ der Dimension $(k \times 1)$.

## 3.2. Modell des original Datensatzes

Das multiple lineare Regressionsmodell kann für den original Datensatz *Ozon* mit $p = 24$ Kovariablen $X_1, \dots, X_{24}$ und einem metrischen Response $Y$ angewendet werden, da die in Kapitel 3 genannten Annahmen wie folgt erfüllt sind:

- $n = 496$ Beobachtungstupel $(x_{i1}, \dots, x_{i24}, y_i)\,,\ i = 1, \dots, 496$ sind gegeben,
- für $k = 25$ Regressionskoeffizienten gilt $496 = n > k = 25$ und
- keine der Kovariablen $X_j,\ j = 1, \dots, 24$ lässt sich als Linearkomination anderer Variablen im Modell bilden.

Damit lässt sich das volle multiple lineare Regressionsmodell des original Datensatzes durch: (Vgl.: Fahrmeir et al., 2009, S. 24)

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{i24} + \epsilon_i\ , \qquad i = 1, \dots, 496\ ,$$

erklären.

# 4. Resampling-Methoden

Bei empirischen Studien werden Resampling-Methoden zur Untersuchung von statistischen Daten immer häufiger herangezogen. Vor allem über die letzten Jahre haben sie gegenüber Standardverfahren an Durchsetzungskraft gewonnen. Während klassische Verfahren theoretische Annahmen fordern, wie die Normalverteilungsannahme, können Resampling-Methoden ohne großen Analyse- und Modellierungsaufwand zur Untersuchung von komplexen, trunkierten oder abhängigen Datensätzen herangezogen werden. (Albers, Klapper, Konradt, Walter & Wolf, 2009, S. 521)

Sogar bei unkomplizierten statistischen Problemen resultieren in vielen Fällen mittels Resampling-Methoden präzisere Ergebnisse. Wie der Begriff *Resampling* schon deuten lässt, beruhen statistische Schlussfolgerungen solcher Methoden auf wiederholten Stichprobenziehungen der analysierten Daten und deren empirischen Verteilungsmerkmalen, die bei der Analyse herausgekommen sind. (Albers et al., 2009, S. 521)

In den folgenden Unterkapiteln sollen die Grundideen zweier Resampling-Methoden, die bei der behandelten empirischen Studie angewandt wurden, vermittelt werden. Die daraus resultierenden Ergebnisse werden in Kapitel 8 verglichen und diskutiert.

## 4.1. Bootstrap

Das wohl wichtigste und flexibelste Resampling-Verfahren (Albers et al., 2009, S. 522) ist der von Efron (1979) eingeführte und in Zusammenarbeit von Efron und Tibshirani (1993) weiterentwickelte Bootstrap (Albers et al., 2009, S. 527).

Aus der Sage des Baron von Münchhausen, der sich an den eigenen Haaren aus dem Sumpf gezogen haben soll, siehe Abbildung 4.1, oder wie es im Englischen formuliert wird, an der eigenen Stiefelschlaufe (engl.: *Bootstrap*), ist sowohl der Begriff als auch der Prozess auf die Statistik abgeleitet worden. Im Grunde bedeutet Bootstrap, das wiederholte Zufallsziehen mit Zurücklegen aus einer Stichprobe, sodass mehrere neue Stichproben erzeugt werden, die den gleichen Stichprobenumfang wie die Originalstichprobe haben. (Albers et al., 2009, S. 527)

Abbildung 4.1.: Baron von Münchhausen mit Pferd im Sumpf (Hosemann, 1807-1875)

Aus dem Paper von De Bin et al. (2014) ist bereits bekannt, dass es sich bei dieser empirischen Studie um einen *nichtparametrischen Bootstrap* handelt (De Bin et al., 2014, S. 8f.). Insbesondere unterscheidet sich dieses Verfahren vom parametrischen Bootstrap dadurch, dass sich die erzeugten Pseudo-Stichproben aus den Komponenten der Originalstichprobe ergeben (Wittmann, 2010, S. 58).

Das nichtparametrische Bootstrap-Verfahren wurde wie im Folgenden erläutert in dieser Studie verwendet:
Aus einer Menge von $i = \{1, \ldots, n\}$ Beobachtungen wurde $n$-mal mit Zurücklegen gezogen, dadurch wurde eine Pseudo-Stichprobe vom Umfang $n$ generiert, die sich aus den gezogenen Beobachtungen für jede der $p$ Einflussgrößen zusammensetzt. Dieser neu gewonnene Datensatz wurde für die Variablenselektion verwendet. Anschließend wurde dieses Verfahren $B$-mal wiederholt, wodurch sich für einen Datensatz $B$ Pseudo-Stichproben generierten.

Das bedeutet, dass einige Beobachtungswerte aus dem original Datensatz in einer Pseudo-Stichprobe mehrfach vorkommen könnten und folglich andere Beobachtungswerte aus der original Stichprobe ausgeschlossen werden würden. In einer Bootstrap-Stichprobe sind im Mittel $0,632n$ voneinander verschiedene Beobachtungen enthalten.
(De Bin et al., 2014, S. 8)

Durch die Möglichkeit von wiederholten Beobachtungen könnten beim nichtparametrischen Bootstrap inkonsistente Schätzer hervorgehen (Albers et al., 2009, S. 534). Dies stellt einen Nachteil für den Bootstrap dar und zieht andere Resampling-Methoden in Betracht. Eine wichtige Alternative ist die im folgenden Kapitel geschilderte *Subsampling*-Methode (De Bin et al., 2014, S. 9).

## 4.2. Subsampling

Subsampling, oder auch *delete-d jackknife* genannt, wurde erstmals von Wu (1986) eingeführt und hat sich durch seine asymptotische Konsistenz gegenüber dem Bootstrap bewiesen (De Bin et al., 2014, S. 9).

Im Gegensatz zum Bootstrap wird beim Subsampling $m$-mal ohne Zurücklegen aus einer Originalstichprobe vom Umfang $n$ gezogen, wobei $m < n$ gilt. Dadurch wird eine Pseudo-Stichprobe generiert, die einen kleineren Stichprobenumfang hat und deren Beobachtungen einmalig sind. Für diese Studie wurde $m$ gleich der durchschnitllichen Anzahl der einmaligen Beobachtungen im Bootstrap gewählt, d.h.: $m = \lfloor 0,632n \rfloor$, sodass sich ein Vergleich der beiden Resampling-Methoden machen lässt.
(De Bin et al., 2014, S. 9)

Die Eigenschaften, der in diesem Kapitel vorgestellten Resampling-Methoden, spielten eine wichtige Rolle bei der Simulationsstudie. Im folgenden Kapitel soll der Umgang der Resampling-Methoden in dieser empirischen Studie erläutert werden, insbesondere im Kapitel 5.3.

# 5. Variablenselektion

Je weniger Parameter ein Modell besitzt, desto geringere Standardfehler haben ihre Koeffizientenschätzer. Viele Parameter in einem Modell führen zu einem breiten Prognoseintervall (Schlittgen, 2013, S. 40). Eines der wichtigsten Anwendungsbereiche der Regressionsanalyse ist die Variablenselektion. Ihre Aufgabe ist es aus einer Menge $p$ von möglichen Einflussgrößen $X_1, \ldots, X_p$ herauszufinden, ob eine geringe Anzahl an Kovariablen genügen würde, um die Zielvariable $Y$ zu erklären (Pruscha, 2006, S. 119). Koeffizientenschätzer ermöglichen bereits eine grobe Einschätzung über den Einflussgrad der Kovariablen auf den Response. Bei großem Absolutbetrag der Koeffizienten kann man einen großen Einfluss der Kovariablen erwarten, sofern ihre Zielvariablen auf der gleichen Skala gemessen wurden. Es bieten sich für die Art und Weise der Variablenselektion mehrere Möglichkeiten an. (Schlittgen, 2013, S. 40)

Bei dieser Arbeit wurde eine *Rückwärts-Selektion* mittels des Modellbewertungskriteriums des *BIC* durchgeführt. In den folgenden Unterkapiteln wird dieser Prozess beschrieben.

## 5.1. Rückwärts-Selektion

Die Rückwärts-Selektion, bekannter unter dem englischen Begriff *backward selection*, hat gewisse Vorzüge im Vergleich zu anderen Selektionsverfahren, welche in Mantel (1970) genauer untersucht wurden (De Bin et al., 2014, S. 7).

In einem Anfangsschritt wird mit dem vollen Modell gestartet. Sukzessive wird diejenige Kovariable aus dem Modell eliminiert, die gemäß Modellwahlkriterium zum schlechtesten Wert führt (Fahrmeir et al., 2009, S. 164). Die aus dem Modell entfernte Variable sollte diejenige sein, die am wenigsten für die Erklärung des Response $Y$ beigetragen hätte. Dieser Schritt wird wiederholt bis keine Verbesserung mehr möglich ist. Das heißt, dass es bei einer weiteren Reduktion zu viel Informationsverlust bedeuten würde. Damit ist die Rückwärts-Selektion beendet. (Vgl.: Schneider, Hommel & Blettner, S. 780)

## 5.2. BIC

Der $BIC$, die Abkürzung für das *Bayesianische Informationskriterium*, stellt ein wichtiges Modellwahlkriterium dar, welches für diese Studie hergenommen wurde. Weiterhin sei $n$ die Anzahl der Beobachtungen und $k$ die Anzahl der Parameter, so wird dieses Informationskriterium definiert durch: (Vgl.: Unkel, 2013, S. 22)

$$BIC = -2 \cdot l(\hat{\theta}) + k \cdot log(n) \ ,$$

wobei $\theta \in \mathbb{R}^k$ ein $k$-dimensionaler Parametervektor ist, mit Log-Likelihood $l(\theta)$ und Maximum-Likelihood-Schätzer $\hat{\theta}$. Bei der Modellwahl ist dasjenige Modell zu wählen, dass die größte Posteriori-Wahrscheinlichkeit besitzt. Das entspricht dem Modell, bei welchem der kleinste $BIC$-Wert berechnet wurde. (Fahrmeir et al., 2009, S. 489)

Bei dieser Studie wurde die Rückwärts-Selektion mittels $BIC$ durchgeführt. Das bedeutet, dass mit dem vollen Modell gestartet wurde und der $BIC$ bei jedem Schritt, d.h.: nach jeder Reduktion einer Kovariablen, berrechnet wurde. Dasjenige Modell mit dem kleinsten $BIC$-Wert wurde ausgewählt.

Diese Variablenselektion wurde auf Resampling-Stichproben laufen gelassen. Dieser Prozess sowie die weitere Vorgehensweise nach Erhalt des geeignetsten Modells soll im Folgenden veranschaulicht werden.

## 5.3. Resampling-basierte Variablenselektion

Bei dieser weiterführenden Studie wurde die Rückwärtsselektion mittels $BIC$ für die Resampling-Stichproben benutzt. Einerseits wurden Resampling-Stichproben aus der Originalstichprobe gezogen, andererseits aus den modifizierten Stichproben, siehe Kapitel 6. Die Abbildung 5.1 soll das Verfahren für die Bootstrap-Methode demonstrieren.

Sei nun die Ausgangsstichprobe $X$ die Originalstichprobe *Ozon*. Der *Ozon*-Datensatz hat zur Erinnerung $n = 496$ Beobachtungen, $p = 24$ Kovariablen und ein Response $Y$. Zu Beginn wurde eine leere Matrix definiert mit $B = 1000$ Zeilen für die Iterationen und $p = 24$ Spalten für die Kovariablen:

$$\text{B=1000}\left\{ \underbrace{\begin{pmatrix} 0 & \dots & 0 \\ \vdots & & \vdots \\ 0 & \dots & 0 \end{pmatrix}}_{p=24} \right.$$

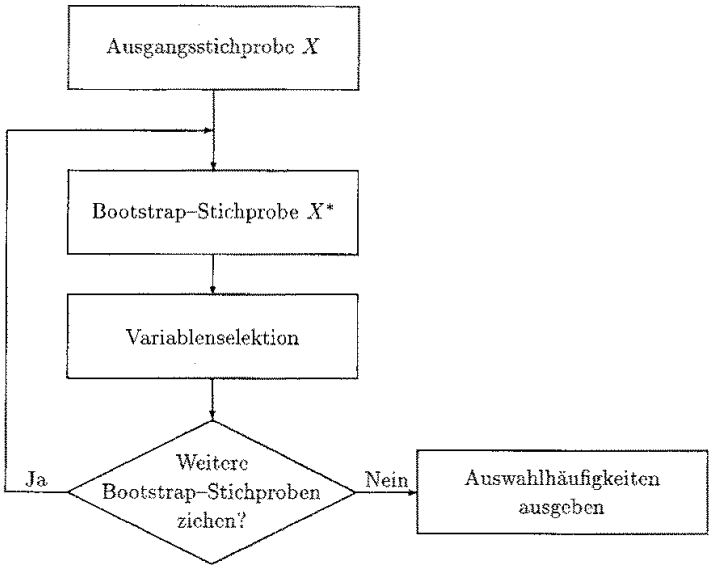



Abbildung 5.1.: Variablenselektion mit Bootstrap-Stichprobe (Fleischer & Folda, 1996, S. 109)

Mit Hilfe des R-Paketes *MASS* von Ripley et al. (2014) wurde aus $\{1,\ldots,496\}$ Beobachtungen $496$-mal mit Zurücklegen gezogen. Die dadurch gewonnene Bootstrap-Pseudo-Stichprobe setzt sich wiederum aus $496$ Beobachtungen zusammen, wobei einige Beobachtungen aus der Originalstichprobe nun gar nicht, einmal, oder mehrfach vorkommen. Die Rückwärtsselektion mittels $BIC$ wurde anschließend auf die Pseudo-Stichprobe laufen gelassen, um das beste Modell dieses Resampling-Datensatzes zu finden.

Für jede Kovariable $X_j$ mit $j = 1,\ldots,24$, die zufolge des Modellwahlkriteriums zum besten Modell gehört, wird in $j$-ter Spalte und $i$-ter Zeile für jeden $i$-ten Durchlauf mit $i = 1,\ldots,B$ eine Eins in die zum Anfang leere Matrix gesetzt. Beispielsweise ergab sich bei der Rückwärts-Selektion mittels $BIC$ im ersten Durchgang $i = 1$ folgender R-Output:

Step: $AIC = -1546.44$
$Y \sim SEX + FLGROSS + FMILB + FNOH24 + FSATEM + FLGEW$

oder anders formuliert: $Y \sim X_3 + X_{11} + X_{12} + X_{13} + X_{20} + X_{22}$ . Das würde bedeuten, dass für jede $j$-te Spalte mit $j = \{3,\ 11,\ 12,\ 13,\ 20,\ 22\}$ in $i$-ter Zeile die Matrix mit Einsen vervollständigt wird, hier für $i = 1$:

| | 1 | 2 | **3** | 4 | … | 10 | **11** | **12** | **13** | 14 | … | 19 | **20** | 21 | **22** | 23 | 24 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 0 | 0 | **1** | 0 | ⋯ | 0 | **1** | **1** | **1** | 0 | ⋯ | 0 | **1** | 0 | **1** | 0 | 0 |
| 2 | 0 | 0 | 0 | 0 | ⋯ | 0 | 0 | 0 | 0 | 0 | ⋯ | 0 | 0 | 0 | 0 | 0 | 0 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋱ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋱ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 1000 | 0 | 0 | 0 | 0 | ⋯ | 0 | 0 | 0 | 0 | 0 | ⋯ | 0 | 0 | 0 | 0 | 0 | 0 |

Damit war der erste Durchlauf beendet und beim zweiten Durchlauf wurde wieder aus der Ausgangsstichprobe eine neue Bootstrap-Pseudo-Stichprobe generiert, die dann wiederum einer Variablenselektion ausgesetzt war. Analog wurde dieser Prozess $B = 1000$-mal wiederholt und die Ergebnisse wurden für die darauffolgenden $i = 2, \ldots, 1000$-Durchläufe in die nicht mehr leere Matrix eingetragen, bis schließlich die Matrix vollkommen $0-1$-kodiert war.

Der Prozess war somit für diese Ausgangsstichprobe beendet, das heißt es wurden keine weitere Bootstrap-Stichproben mehr aus dieser Ausgangsstichprobe gezogen und die Auswahlhäufigkeiten wurden ausgegeben. In diesem Fall sind das die Gini-purity und die relativen Inklusions-Häufigkeiten, die in Kapitel 7 behandelt werden.

Sowohl die Abbildung 5.1 als auch der eben beschriebene Prozess gilt analog für die Subsample-Stichproben. Der einzige Unterschied ist, dass beim Subsampling *m-mal ohne Zurücklegen* gezogen wurde, mit $m = \lfloor 0,632 \cdot n \rfloor = \lfloor 0,632 \cdot 496 \rfloor = \lfloor 313,472 \rfloor = 313$.

Gleiches gilt für jede der im folgenden Kapitel vorgestellten modifizierten Datensätze. Jeder Datensatz hat für $1000$-Iterationen die Rolle der Ausgangsstichprobe angenommen.

# 6. Modifizierte Datensätze

Damit für die empirische Studie die Auswirkung von Ausreißern untersucht werden konnte, mussten zusätzliche Ausreißer durch eine Simulation in den *Ozon*-Datensatz eingebaut werden. Insgesamt kamen dabei 1.800 unterschiedliche durch Ausreißer modifizierte Datensätze zustande. Die folgenden Unterkapitel sollen einen Überblick geben, wie die Ausreißer eingebaut wurden und inwiefern sich die modifizierten Datensätze unterscheiden.

## 6.1. Generierung von Ausreißern

Der *Ozon*-Datensatz enthält 24 Kovariablen, davon sind 17 binär und 7 metrisch. Die Möglichkeit Ausreißer in den Datensatz einzufügen bot sich daher nur für diese 7 Kovariabeln an: *FLGROSS, FLGEW, ALTER, AGEBGEW, FNOH24, FO3H24, FTEH24*. Diese Kovariablen wurden weiterhin differenziert in folgende Variablentypen:

a) *Core*-Variablen: *FLGROSS, FLGEW*

b) *Non-Core*-Variablen: *ALTER, AGEBGEW, FNOH24, FO3H24, FTEH24*

Nach Vorgabe in welchen Variablentyp Ausreißer generiert werden sollen, wurden zufällig innerhalb der Typen Ausreißer-Werte für eine oder mehrere Variablen berechnet. Danach wurde für die entsprechende Variable zufällig eine Beobachtung gewählt, die durch den Ausreißer-Wert ersetzt wurde.

Die Literatur bietet unterschiedlichste Auffassungen für den Begriff des *Ausreißers*, daher gibt es keine allgemeine Definition (Fahrmeir et al., 2009, S.173). Basierend auf der von Tukey (1997) vorgeschlagenen Definition des Ausreißers wurden Ausreißer-Werte berechnet. Alle Werte, die:

- größer sind als $Q_3 + IQR \cdot 1,5$ bzw.
- kleiner sind als $Q_1 - IQR \cdot 1,5$

sind nach Tukey *Ausreißer*. Wobei $Q_1$ das untere Quartil (25%-Quartil), $Q_3$ das obere Quartil (75%-Quartil) und $IQR$ den Interquartilsabstand definiert.
(Vgl.: Geßler, 1993, S. 99)

Das bedeutet, dass ein Ausreißer-Wert generiert werden kann, wenn statt dem Faktor 1,5 eine Zahl größer als 1,5 gesetzt wird. Mit diesem Hintergrund wurden Ausreißer-Werte für die unterschiedlichen Variablen nach diesem Muster definiert:

$$Q_3 + IQR \cdot \textbf{Faktor}\ ,\ mit\ \textbf{Faktor} \in \{2, 5, 10\},$$

Dadurch folgten unterschiedlich starke Ausreißer-Werte:

a) *moderate* Ausreißer, die mit Faktor 2,

b) *mittel-starke* Ausreißer, die mit Faktor 5 und

c) *starke* Ausreißer, die mit Faktor 10

berechnet wurden.

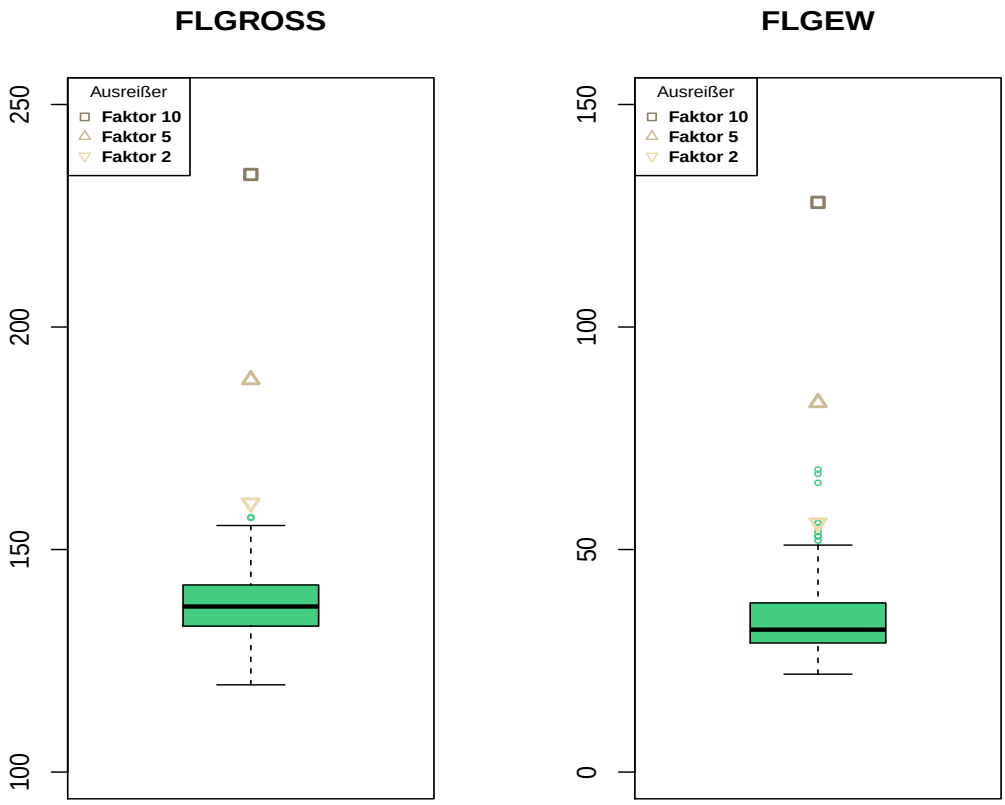



Abbildung 6.1.: Durch Ausreißer modifizierte *Core*-Variablen

An den Abbildungen 6.1 und 6.2 lässt sich erkennen, wie sehr die berechneten Ausreißer-Werte von der Gesamtstreuung der jeweiligen Variablen abweichen. Die Ausreißer-Werte je Variable wurden unbeachtet davon, ob diese sinnvoll bzw. möglich sind erzeugt. An den Box-Plots sieht man, dass die *Core*-Variablen FLGROSS und FLGEW, sowie die *Non-Core*-Variablen ALTER und AGEBGEW bereits natürliche Ausreißer hatten.

Allerdings spielt dies für die Untersuchungen keine so große Rolle, da die modifizierten Datensätze später mit dem original Datensatz verglichen werden und diese sich lediglich durch die berechneten Ausreißer-Werte unterscheiden. Außerdem ist zu beachten, dass es sich bei der Skalierung der $y$-Achse um unterschiedliche Maßeinheiten handelt.

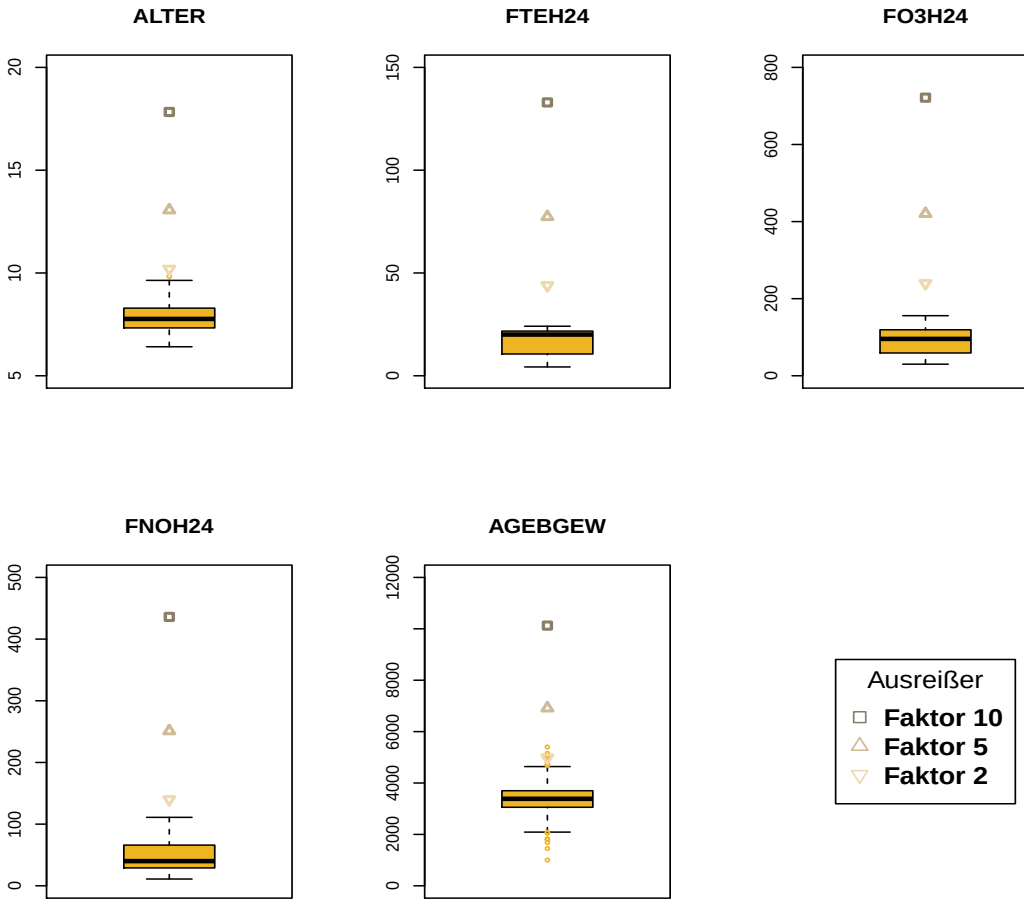



Abbildung 6.2.: Durch Ausreißer modifizierte *Non-Core*-Variablen

## 6.2. Szenarien

Der Einfluss von Ausreißern bei der Datenanalyse und speziell bei der Resampling-basierten Variablenselektion könnte von folgenden Kriterien abhängen:

a) Die Stärke der Ausreißer (moderat, mittel-stark oder stark)

b) Die Anzahl der Ausreißer (1, 5, 10 oder 50 Ausreißer)

c) Die Lage der Ausreißer (Core-Variable, Non-Core-Variable oder in alle Metrischen)

Deswegen wurden verschiedene Szenarien untersucht, die in Tabelle 6.2 dargestellt sind.

| Szenario | Variablentyp | Ausreißer-Menge | Faktor $f \in \{2, 5, 10\}$ |
|---|---|---|---|
| 1 | *core* | | |
| 2 | *non-core* | 1 | $f$ |
| 3 | alle 7 | | |
| 4 | *core* | | |
| 5 | *non-core* | 5 | $f$ |
| 6 | alle 7 | | |
| 7 | *core* | | |
| 8 | *non-core* | 10 | $f$ |
| 9 | alle 7 | | |
| 10 | *core* | | |
| 11 | *non-core* | 50 | $f$ |
| 12 | alle 7 | | |

Tabelle 6.1.: Zwölf untersuchte Szenarien einer Ausreißer-Stärke mit Faktor $f$

Je Ausreißer-Faktor $f$ wurden 12 Szenarien untersucht. Beispielsweise sind im 9. Szenario 10 Ausreißer zufällig in die 7 metrischen Variablen generiert worden. Dabei ist es möglich, dass alle, einige, einer oder gar keiner der 10 Ausreißer in die *Core*-Variablen gesetzt worden sind. Analog sind die anderen untersuchten Szenarien von der Tabelle 6.2 abzulesen.

Für jedes Szenario und je Ausreißer-Stärke wurden 50 variierende Datensätze erzeugt. Wie beim *Ozon*-Datensatz wurde für jede der $50 \cdot 12 \cdot 3 = 1.800$ modifizierten Datensätze Resampling-Stichproben gezogen und analog wie in Kapitel 5.3 beschrieben durch Rückwärts-Selektion mittels $BIC$ Auswahlhäufigkeiten angegeben.

Im folgenden Kapitel werden die Eigenschaften der Auswahlhäufigkeiten, die bei diesem Prozess berechnet wurden, beschrieben.

# 7. Vergleichskriterien

Die Auswahlhäufigkeiten, die durch die Variablenselektion für jeden Datensatz und je Resampling-Methode ausgegeben wurden, sind in vollständig $0-1$-kodierterte Matrizen zusammengefasst worden, daraus wurden schließlich die Gini-purity und die relativen Inklusions-Häufigkeiten der Variablen berechnet. Weichen diese Werte, die aus den modifizierten Datensätzen berechnet wurden, sehr von den Ergebnissen des original Datensatzes ab, so kann eine Auswirkung der Ausreißer auf die Stabilität der Variablenselektion unterstellt werden. Daher stellen die Gini-purity und die realtiven Inklusions-Häufigkeiten Vergleichskriterien in dieser Studie dar. In den folgenden Unterkapiteln werden diese Vergleichskriterien definiert und erläutert.

## 7.1. Gini-purity

Der *Gini-purity*, das Gegenteil zum Gini-impurity, bezeichnet ein Reinheitsmaß und wird in dieser Studie als Maß hergenommen, um die Stabilität der Variablenselektion der unterschiedlichen Stichproben zu vergleichen. Seien in einer Trainingsmenge $T$ mit disjunkten Trainingsobjekten $T_i$, mit $i = 1, \ldots, n$ Klassen gegeben und $g_i$ bezeichnet die relative Häufigkeit der Klasse $c_i$ in $T$, dann ist der Gini-purity gegeben durch: (Vgl.: Aßfalg et al., 2003, S. 103ff.)

$$gini(T) = \sum_{i=1}^{C} (g_i)^2 \ .$$

Die durch Variablenselektion an Resampling-Stichproben erzeugten $0-1$-kodierten Matrizen stellen, wie in Kapitel 5.3 erwähnt, in jeder $i$-ten Zeile das in dieser Iteration gewählte Modell dar. Dabei umfasst eine Klasse alle identischen Modelle. Somit entspricht der *Gini-purity* in diesem Fall, der Summe der quadrierten relativen Häufigkeiten der gegebenen Modelle. Würde beispielsweise in allen 1000 Iterationen das gleiche Modell ausgewählt werden, wäre $gini(T) = 1$. Dies würde eine maximale Konzentration eines Modells und hohe Reinheit bedeuten (Vgl.: Aßfalg et al., 2003, S. 105), da nur eine Klasse vorhanden ist, jedoch wäre dies bei diesem Umfang der Iterationen höchst unrealistisch. Ist der Gini-purity jedoch nahe Null, so bedeutet das, dass viele unterschiedliche Modelle selektiert wurden und damit eine geringe Reinheit gegeben ist. Dies soll in einem kleinen Beispiel mit zwei statt 24 Kovaribalen und mit drei statt mit 1000 Iterationen demonstriert werden:

Bei der Variablenselektion sei in jeder Iteration die erste Kovariable nie und die zweite Kovariable immer ausgewählt worden. Das heißt, dass es nur eine Klasse $"01"$ gibt, welche drei mal ausgewählt wurde und damit ist:

$$\begin{pmatrix} 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \Rightarrow gini(T)_{Reinh.hoch} = \sum_{i=1}^{1} (g_i)^2 = \left(\frac{3}{3}\right)^2 = 1$$

Sei dagegen in allen Iterationen ein anderes Modell gewählt worden, gäbe es drei unterschiedliche Klassen $"01", "10"$ und $"11"$, die je einmal ausgewählt wurden und damit:

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{pmatrix} \Rightarrow gini(T)_{Reinh.gering} = \sum_{i=1}^{3} (g_i)^2 = \left(\frac{1}{3}\right)^2 + \left(\frac{1}{3}\right)^2 + \left(\frac{1}{3}\right)^2 = \frac{1}{3} = 0,3\overline{3} \; .$$

Durch den Gini-purity als Vergleichskriterium kann erkannt werden, bei welcher Resampling-Methode die Modellstabilität stärker ausgeprägt ist und für welche Szenarien eine höhere Konzentration der Modelle gegeben ist.

## 7.2. Relative Inklusions-Häufigkeiten

Eine sehr wichtige weitere Information, die durch die Resampling-basierte Variablenselektion erhalten wurde, sind die relativen Inklusions-Häufigkeiten der Variablen. Damit ist die relative Häufigkeit der Male in dem diese Variable in ein Modell einbezogen wurde definiert, d.h.: das ihre Werte alle zwischen 0 und 1 liegen. Ist die relative Inklusions-Häufigkeit einer Variable 0, bedeutet das, dass die Variable nie in ein Modell einbezogen wurde. Dementsprechend bedeutet eine relative Inklusions-Häufigkeit von 1, dass diese Variable immer (Vgl.: De Bin et al., 2014, S. 8) in das durch Rückwärts-Selektion mittels $BIC$ gewählte Modell einbezogen wurde. Die Forschungen von De Bin et al. (2014) haben bereits gezeigt, dass die *Core*-Variablen des *Ozon*-Datensatzes immer hohe relative-Inklusionshäufigkeiten bewiesen. (De Bin et al., 2014, S. 12)

Konkreter hat jede *Core*-Variable für $B = 1000$ Iterationen beim original Datensatz eine relative Inklusions-Häufigkeit von 1. Im folgenden Kapitel, explizit in Kapitel 8.2, wird unteranderem untersucht, ob dies für die modifizierten Stichproben auch gilt, oder ob durch die hinzugefügten Ausreißer in den *Core*-Variablen sich die relativen Inklusions-Häufigkeiten verringern werden.

# 8. Vergleich der Resampling-Methoden

## 8.1. Methodenvergleich bei Betrachtung der Gini-purity

In den folgenden Unterkapiteln werden die Ergebnisse der Gini-purity verglichen. In Kapitel 8.1.1 wird der Einfluss der Ausreißer-Stärke für jedes Szenario thematisiert und mit den ursprünglichen Ergebnissen verglichen und in Kapitel 8.1.2 wird insbesondere auf die Modellstabilität der Resampling-Methoden in Abhängigkeit der Ausreißer-Anzahl eingegangen.

### 8.1.1. Einfluss der Ausreißer-Stärke

Beim original Datensatz wurden zwei *Gini-purity*-Werte berechnet, da nur zwei $0-1$-kodierte Matrizen bei der Variablenselektion erzeugt wurden, das heißt für jede Resampling-Methode eine Matrix. Dementsprechend werden in den folgenden Abbildungen zwei Geraden dargestellt. Die pinke Gerade repräsentiert den berechneten *Gini-purity* der durch Bootstrap-Verfahren erzeugten Matrix und die türkise Gerade den Wert des *Gini-purity* der durch die Subsampling-Methode zustande kam. Die folgenden Abbildungen veranschaulichen die Ergebnisse der *Gini-purity* für jede der drei Ausreißer-Stärken. Dabei stellt die x-Achse für jede Grafik die Szenarien dar, welche zusätzlich durch die gestrichelten vertikalen Linien gruppiert sind, sodass sie in gleicher Reihenfolge wie in Tabelle 6.2 dargestellt sind. Für jedes Szenario und je Resampling-methode sind 50 modifizierte Datensätze erzeugt worden, dementsprechend sind je 50 Gini-purity-Werte berechnet worden und daher sind die Ergebnisse für jedes Szenario in Boxplots dargestellt. Analog zum original Datensatz sind die Ergebnisse der modifizierten Datensätze in den entsprechenden Farben der Resampling-Methode gekennzeichnet.

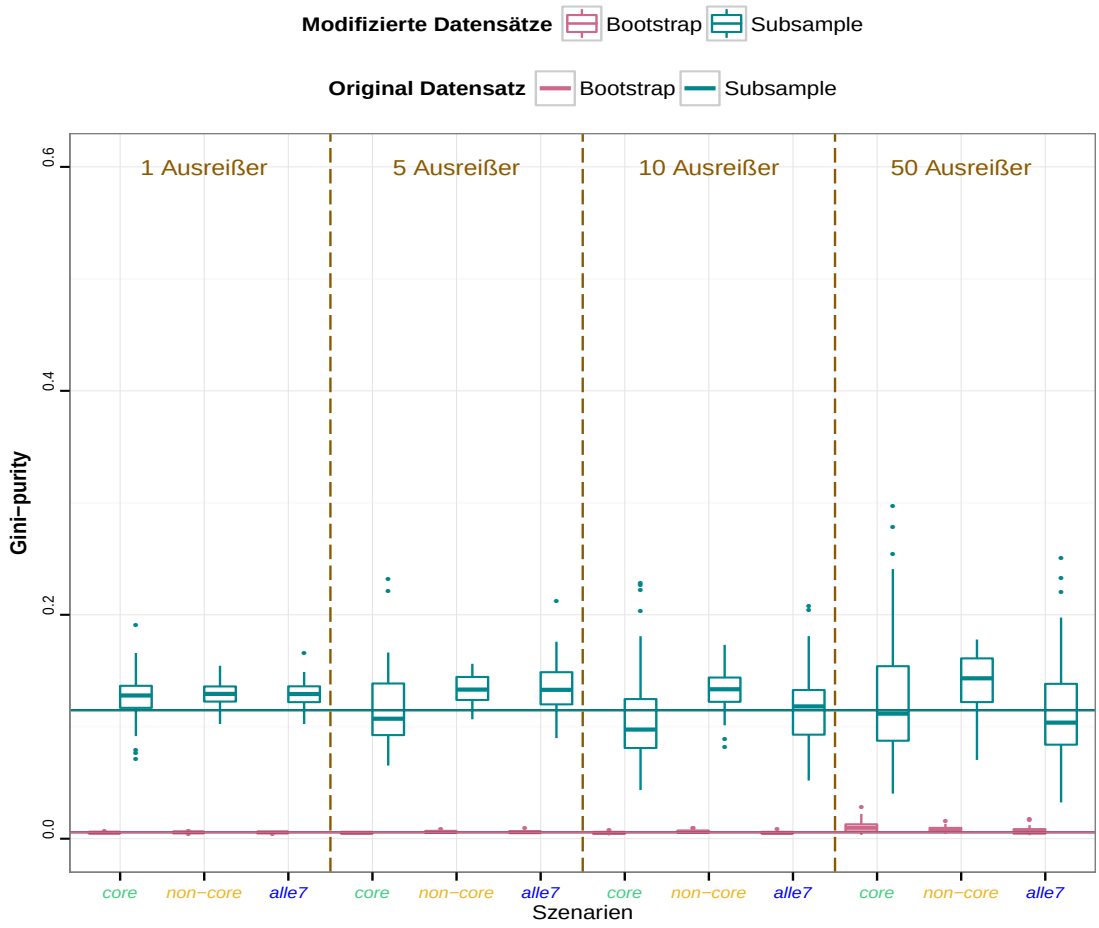



Abbildung 8.1.: Die Streuung der *Gini-purity* der modifizierten Daten für alle Szenarien durch **moderate** Ausreißer und die *Gini-purity* des original Datensatzes je Resampling-Methode

In Abbildung 8.1 ist dies für moderate Ausreißer dargestellt. Wie bei den Ergebnissen des original Datensatzes wurden auch bei den modifizierten Datensätzen höhere Gini-purity-Werte gemessen. Der Großteil der Ergebnisse liegt unter einem Gini-purity-Wert von $0,2$. Bei den Ergebnissen durch Subsampling-Methode kann man eine größere Streuung der Werte für Ausreißer die in Core-Variablen liegen erkennen. Dabei vergrößert sich die Streuung mit ansteigender Ausreißer-Anzahl. Die Ergebnisse der modifizierten Datensätze durch die Bootstrap-Methode sind mit dem Ergebnis vom original Datensatz fast identisch, sodass die Boxplots erst bei 50 Ausreißern in der Grafik zu erkennen sind. Dabei ist, wie beim Subsampling, die Streuung der Gini-purity bei 50 Ausreißern in den Core-Variablen am größten.

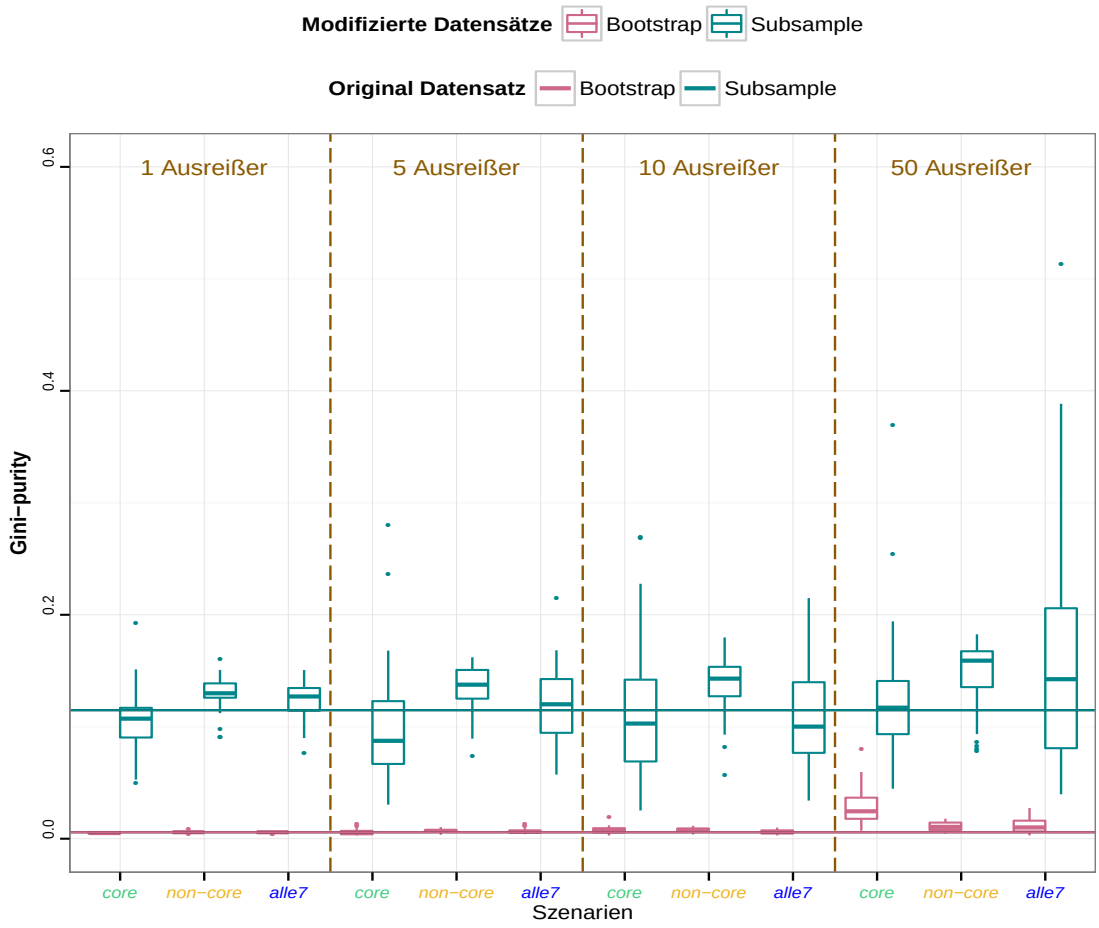



Abbildung 8.2.: Die Streuung der *Gini purity* der modifizierten Daten für alle Szenarien durch **mittel-starke** Ausreißer und die *Gini-purity* des original Datensatzes je Resampling-Methode

In Abbildung 8.2 sind die Ergebnisse der Gini-purity bei mittel-starken Ausreißern dargestellt. Im Vergleich zur vorherigen Abbildung hat sich die Lage der Boxplots verändert. Dies gilt besonders für Ausreißer in den Core-Variablen bei der Subsampling-Methode, denn die Boxen verlagern sich allmählich unterhalb des gemessenen Gini-purity-Wertes der original Daten. Das bedeutet, dass öfters unterschiedliche Modell selektiert wurden. Im Gegensatz zur Abbildung 8.1 besitzt das Szenario bei dem die Lage der 50 Ausreißer nicht berücksichigt wurde, sodass zufällig in die sieben metrischen Variablen Ausreißer eingefügt worden sind, die größte Streuung der Gini-purity-Werte bei der Subsampling-Methode. Währenddessen steigen die Gini-purity-Werte bei der Boostrap-Methode an. Die Modelle sind mit mittel-starken Ausreißern stabiler geworden, dies gilt vor allem bei den Szenarien mit 50 Ausreißern.

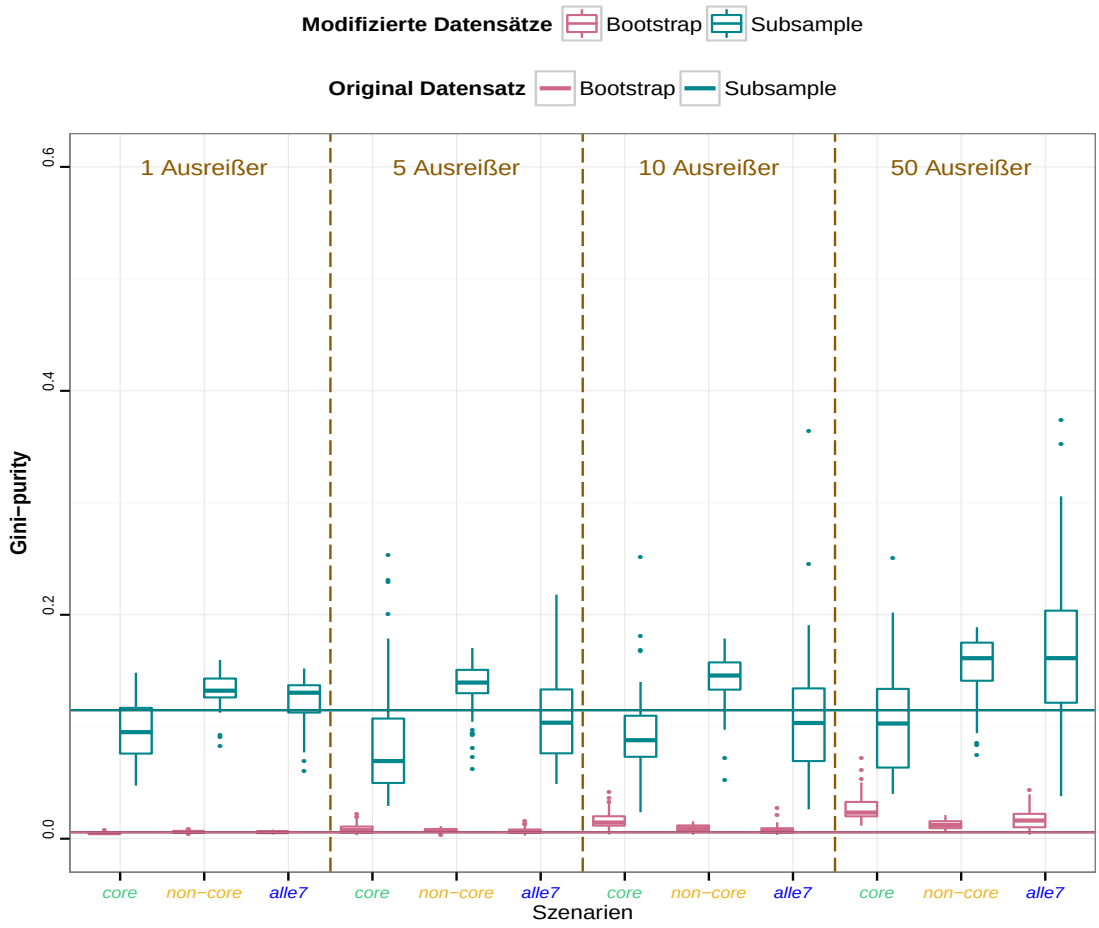



Abbildung 8.3.: Die Streuung der *Gini-purity* der modifizierten Daten für alle Szenarien durch **starke** Ausreißer und die *Gini-purity* des original Datensatzes je Resampling-Methode

Die Abbildung 8.3 bestätigt die Tendenz aus vorangegangener Grafik. Die Modelle der Stichproben mit Ausreißern in Core-Variablen sind bei der Subsampling-Methode noch unstabiler geworden. Die Konzentration der Modelle aus den Szenarien der Non-Core-variablen ist dagegen leicht angestiegen. Für die Bootstrap-Methode ist zu bemerken, dass durch die Anwesenheit der Ausreißer die Stabilität des Modells sich zu verbessern scheint. Mit Anstieg der Ausreißer-Stärke haben sich in allen Szenarien die Werte des Gini-purity erhöht.

Im folgenden Unterkapitel wird untersucht, ob sich die Modellstabilität tatsächlich auch mit der Anzahl der Ausreißer verbessert.

### 8.1.2. Einfluss der Ausreißer-Menge auf die Gini-purity

Das Verhalten der Gini-purity bei den Untersuchung der Ausreißer-Stärke ließen darauf hin deuten, dass die Konzentration der Modelle sich vor allem bei den Ergebnissen der Bootstrap-Methode durch steigende Ausreißer-Anzahl verbessert.
Dies wird in Abbildung 8.4 offensichtlicher. Diese Grafik veranschaulicht die Gini-purity in Abhängigkeit der Ausreißer-Anzahl. Dabei sind die Ergebnisse aus den Resampling-Methoden in den gleichen Farben dargestellt wie in den vorherigen Grafiken, wobei in dieser Grafik die Boxplots mit den entsprechenden Farben für die Ausreißer-Stärke gefüllt sind.

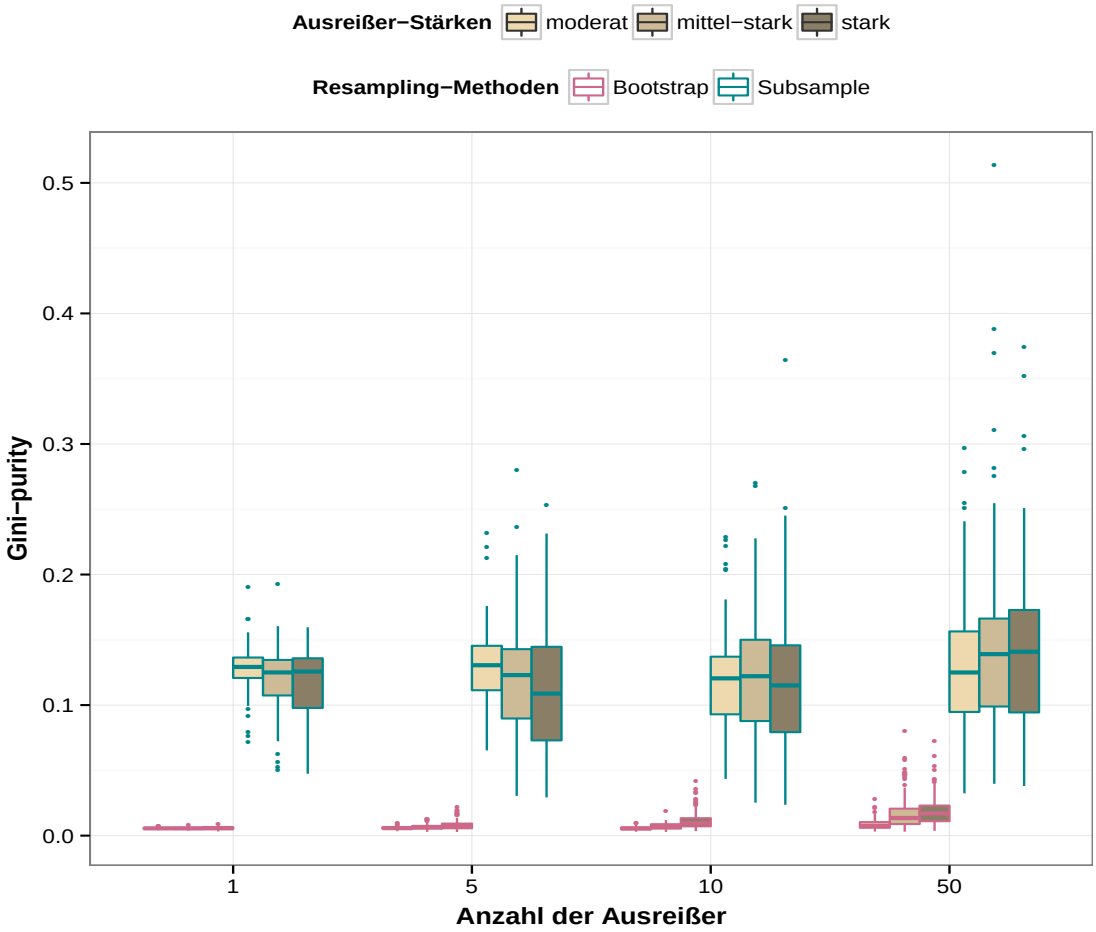



Abbildung 8.4.: Die Streuung der *Gini-purity* der modifizierten Daten in Abhängigkeit von der Ausreißer-Menge für je Ausreißer-Stärke und Resampling-Methode

Durch diese Darstellung ist zu erkennen, dass bei der Bootstrap-Methode die Werte des Gini-purity mit größerer Anzahl der Ausreißer ansteigen. Das heißt, dass sich die Stabilität der Modellwahl durch mehrere Ausreißer in der Stichprobe verbessert. Dies wird noch durch die Ausreißer-Stärke verstärkt. Bei der Subsampling-Methode ist nur ein leichter Anstieg der Werte vor allem für mittel-starke und starke Ausreißer zu erkennen.

Für eine größere Ausreißer-Menge ist insbesondere eine größere Streuung der Gini-purity-Werte zu beobachten. Die Größe der Gini-purity-Werte für moderate Ausreißer bei der Subsampling-Methode nehmen für fünf Ausreißer gegenüber einem Ausreißer zu, dann sinken die Werte für zehn Ausreißer und steigen bei 50 eingefügten Ausreißern wieder an. Das sich allgemein die Stabilität bei der Bootstrap-Methode in Abhängigkeit der Ausreißer-Anzahl verbessert und im Vergleich dazu bei der Subsampling-Methode kaum eine Verbesserung zu sehen ist, hätte man nach den Forschungsberichten von De Bin et al. (2014) zu urteilen, nicht erwartet. Nichtsdestotrotz liegen die Gini-purity-Werte der Subsampling-Methode über denen der Bootstrap-Methode und damit ist bei dieser Untersuchung eine größere Konzentration der Modelle durch die Subsampling-Methoden feststellbar.

## 8.2. Methodenvergleich bei Betrachtung der relativen Inklusions-Häufigkeiten

Die Untersuchungen der relativen Inklusions-Häufigkeiten, welche je Ausreißer-Stärke und je Szenario betrachtet wurden, haben sowohl beim Vergleich der Ausreißer-Stärke bei gegebenem Szenario als auch beim Vergleich der Ausreißer-Menge bei gegebener Ausreißer-Stärke Abweichungen in jeglicher Hinsicht ergeben. Aufgrund der großen Vielfalt, die sich durch die große Anzahl der Szenarien je Ausreißer-Stärke ergab, stellten sich viele Möglichkeiten die Ergebnisse zu vergleichen. Im Unterkapitel 8.2.2 werden die Ergebnisse des original Datensatzes mit denen der modifizierten Datensätze verglichen. Dabei handelt es sich bei den modifizierten Datensätzen immer um mittel-starke Ausreißer die in eine *Core*-Variable des *Ozon*-Datensatzes eingefügt wurden. Die Ergebnisse der übrigen Szenarien sind im Anhang zu finden.

### 8.2.1. Bedeutung der Lage der Ausreißer

Bei Betrachtung aller Ergebnisse wurde festgestellt, dass auch die Lage der Ausreißer einen Einfluss auf die Inklusions-Häufigkeiten der Variablen ausübten. So wurden bei den Ergebnissen größere Unterschiede festgestellt, wenn sich der Ausreißer in einer *Core*-Variablen befand als in einer *Non-Core*-Variablen. Aus diesem Grund war die Entscheidung für die Darstellung der *Core*-Variablen gefallen. Sind die Ausreißer in den *Non-Core*-Variablen so hat die Ausreißer-Stärke einen geringen Einfluss auf die relativen Inklusions-Häufigkeiten und im Wesentlichen veränderte sich meist nur die Streuung der *Non-Core*-Variablen. Während bei Szenarien, die Ausreißer in Core-Variablen enthalten, sowohl sich der Einfluss der Ausreißer-Stärke als auch der Einfluss der Ausreißer-Menge tendenziell bei allen Variablen auswirkte.

### 8.2.2. Auswirkung der Ausreißer-Menge auf die Inklusions-Häufigkeiten

Wie beim Gini-purity werden in den folgenden Abbildungen sowohl die Resampling-Methoden als auch die Abweichung der modifizierten Datensätze zum original Datensatz verglichen. Dabei kennzeichnet pink wieder die Ergenisse, die durch Boostrap-Verfahren entstanden sind und türkis die durch Subsampling-Verfahren entsprungen sind. Selbstverständlich sind als Vergleich die Ergebnisse des original Datensatzs zu jeder Variable eingezeichnet, wobei diese Ergebnisse durch Schneeflocken bzw. Zielscheiben gekennzeichnet sind. Die x-Achse bildet dabei immer die Kovariablen ab und zwar zuerst die *Core*-Variablen in denen sich die Ausreißer befinden, dann die metrischen *Non-Core*-Variablen und schließlich alle binären Variablen. Die y-Achse stellt somit die relative Inklusions-Häufigkeit dar.

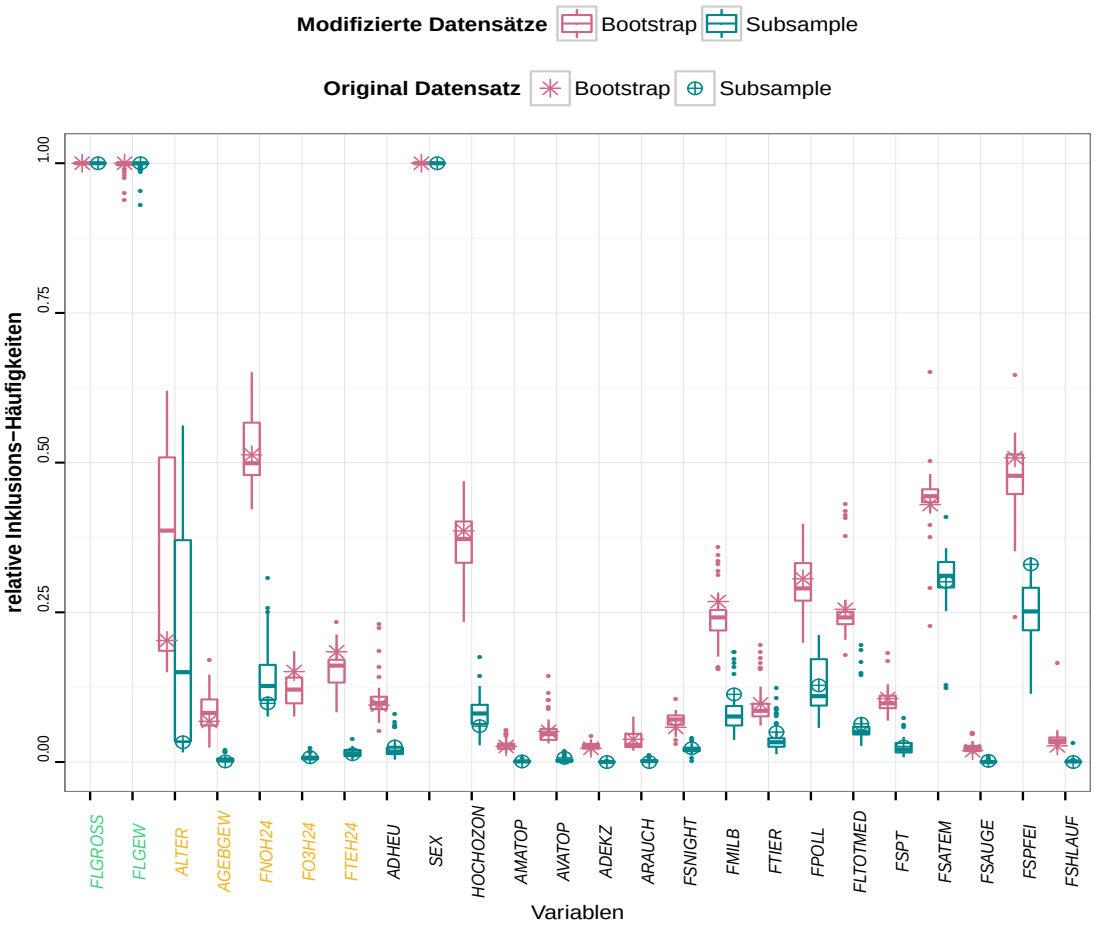



Abbildung 8.5.: Szenario 1 mit Faktor 5: **Ein mittel-starker Ausreißer in einer Core-Variablen**

Die Abbildung 8.5 zeigt die Ergebnisse der relativen Inklusions-Häufigkeiten bei der nur ein mittel-starker Ausreißer in eine der beiden *Core*-Variablen eingefügt wurde. Die *Core*-Variablen, welche dadurch definiert wurden, dass sie für 1000 Iterationen immer zum besten Modell gehören, dass man anhand der Symbole für den original Datensatz beider Resampling-Methoden erkennen kann, geben bei einem enthaltenden moderaten Ausreißer ein anderes Bild ab. Denn nur die Variablen FLGROSS und SEX wurden immer ins Modell gewählt, während die Variable FLGEW bei der Variablenselektion der modifizierten Datensätze nicht immer zum besten Modell gehört hat. Dafür gewinnt die Variable ALTER im Modell der modifizierten Datensätze mehr an Bedeutung. Die Boxplots der Variable ALTER besitzen die größte Streuung im Vergleich zu den anderen Variablen. Wobei der durch Subsampling-Verfahren berrechnete Boxplot eine etwas größere Streuung hat, als der der Bootstrap-Methode. Jedoch liegt der Median beim Subsampling-Verfahren näher am original Ergebniss. Allgemein ist zu bemerken, dass bei der Variablen ALTER die Mediane der Boxplots beider Resampling-Methoden mehr von dem original Ergebniss abweichen als bei anderen Variablen.

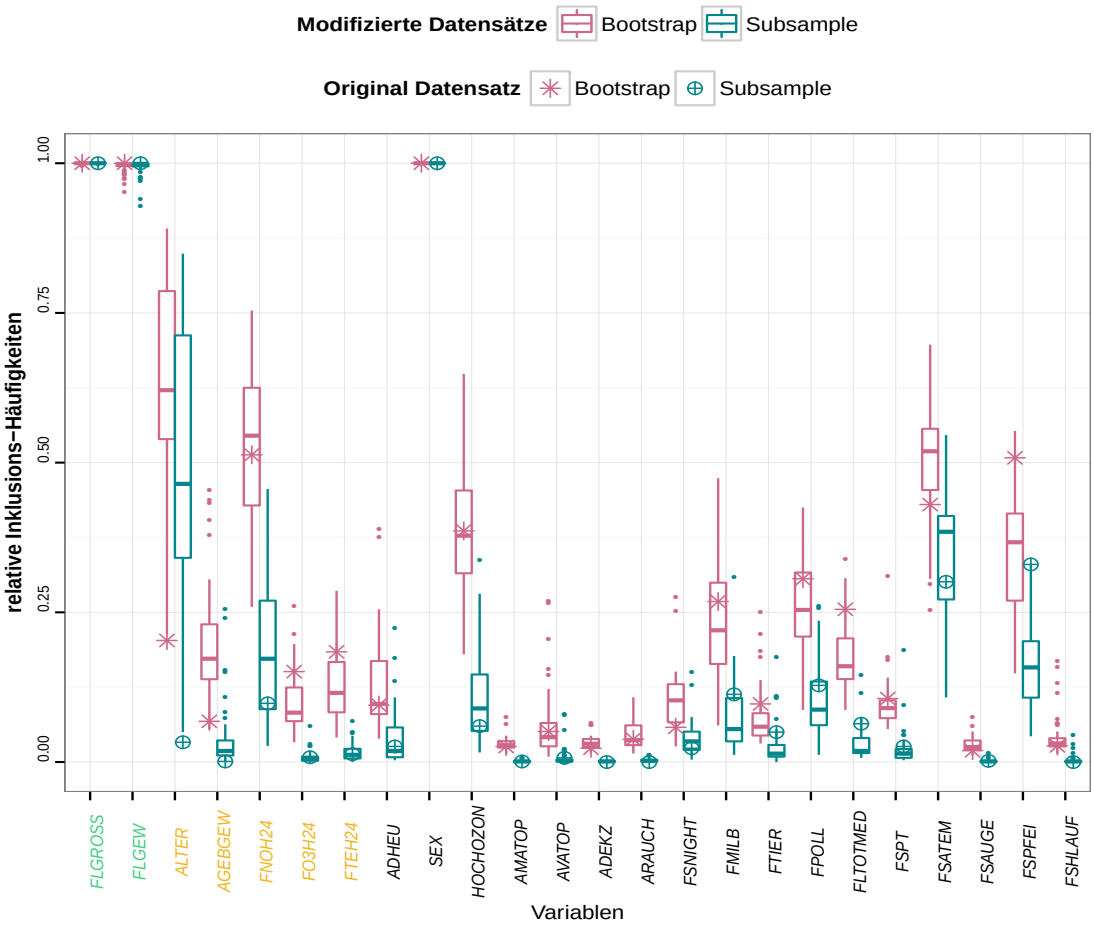



Abbildung 8.6.: Szenario 4 mit Faktor 5: **Fünf mittel-starke Ausreißer in Core-Variablen**

Durch den eingefügten Ausreißer ist auch bei der Variablen FSPFEI eine eher größere Abweichung der Ergebnisse gegenüber dem Ergebniss des original Datensatzes zu bemerken. Dies ist insbesondere bei der Subsampling-Methode sichtbar. Dabei wird die Variable FSPFEI aufgrund des Ausreißers in einer *Core*-Variablen seltener ins Modell gewählt. Dies wird in Abbildung 8.6 deutlicher. Nun sind fünf mittel-starke Ausreißer in einer *Core*-Variablen oder in beide *Core*-Variablen verteilt. Dabei wird vor allem die Streuung der Variablen ausgeprägter. Die Variable ALTER wurde noch häufiger ins Modell gewählt und hat weiterhin die größte Streuung gegenüber den anderen Variablen. Der Median entfernt sich nun für beide Resampling-Methoden gleichermaßen vom Ergebniss des original Datensatzes.Durch die fünf mittel-starken Ausreißer wurden außer der Variablen ALTER vor allem die Variablen AGEBGEW, FNOH24 und FSATEM öfters ins Modell gewählt. Während die Variable FSPFEI noch weniger zur Erklärung des Response beitrug und daher seltener in Modell gewählt wurde. Die Resampling-Methoden verhalten sich durch das Einfügen der Ausreißer ähnlich. Entweder steigt die Anzahl der Inklusion-Häufigkeiten der modifizierten Datensätze gegenüber dem original Datensatz an, oder sie fällt bei beiden Resampling-Methoden.

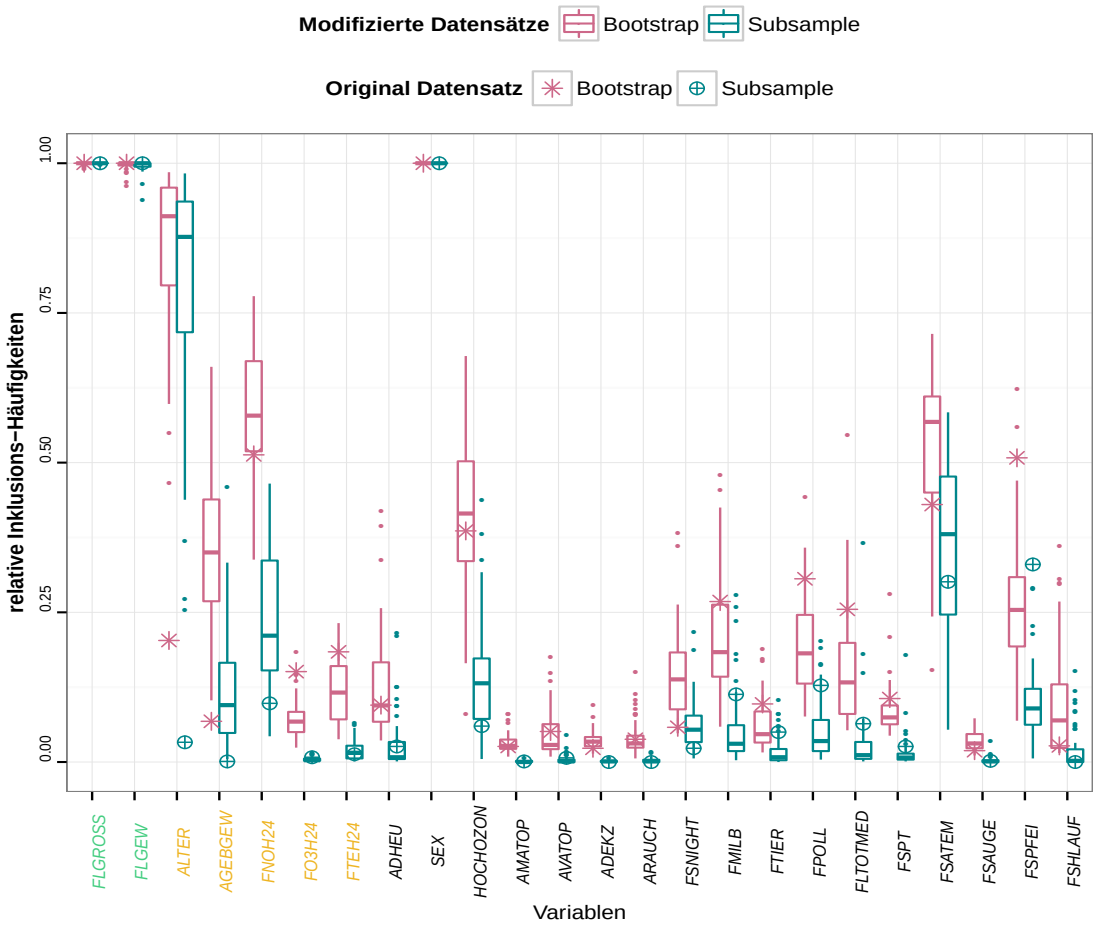



Abbildung 8.7.: Szenario 7 mit Faktor 5: **Zehn mittel-starke Ausreißer in Core-Variablen**

Bei Betrachtung der Abbildung 8.7, wird die Abweichung der Ergebnisse der modifizierten Datensätze aufgrund der zehn mittel-starken Ausreißer zum original Datensatz stärker zum Ausdruck gebracht. Die Variable ALTER hat im Vergleich zum Ergebnis mit fünf eingefügten Ausreißern eine geringere Streuung, allerdings unterscheiden sich die Ergebnisse des modifizierten Datensatzes von den original Datensatz Werten mit deutlichem Abstand. Außerdem weichen auch die Ergebnisse der modifizierten Datensätze der Variablen AGEBGEW und FSPFEI von den original Ergebnissen klarer ab.

Ein eindeutigeres Bild über die Wichtigkeit der Variablen im Modell spiegelt die Abbildung 8.8 wider. Bei 50 mittel-starken Ausreißern, die in die *Core*-Variablen eingefügt worden sind, wurden durch die Rückwärts-Selektion mittels $BIC$ öfters andere Variablen ins Modell gewählt bzw. ausgeschlossen als es beim *Ozon*-Datensatz der Fall war. Die Core-Variablen FLGEW und FLGROSS verlieren bei den modifizierten Datensätzen häufiger ihre Position im geeignetsten Modell. Während die Variable ALTER bei allen 50 modifizierten Datensätzen je Resampling-Methode für 1000 Iterationen eine relative Inklusions-Häufigkeit von 1 besitzt.

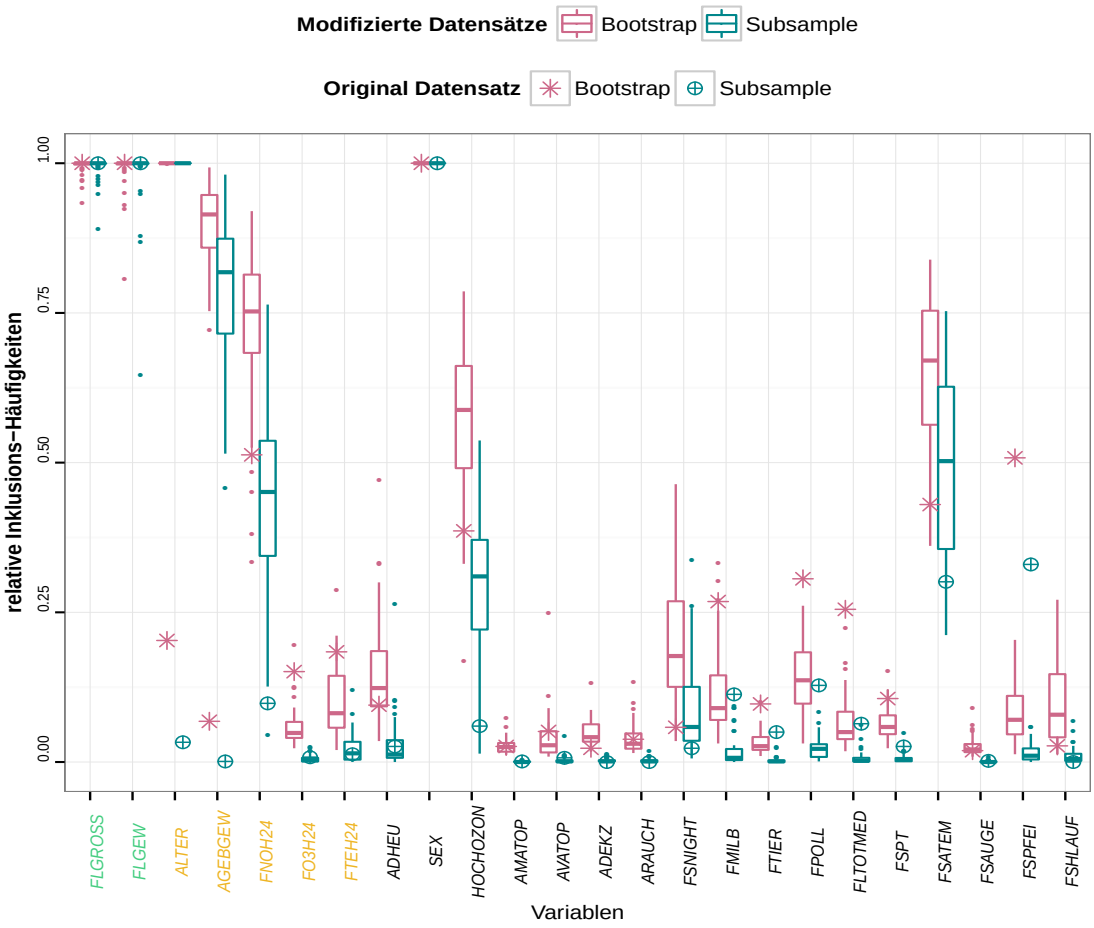



Abbildung 8.8.: Szenario 10 mit Faktor 5: **Fünfzig mittel-starke Ausreißer in Core-Variablen**

Das heißt die Ergebnisse der modifizierten Datensätze und die des original Datensatzes unterscheiden sich im Maximum für diese Variable. Im Vergleich der steigenden Anzahl der Ausreißer, die in die *Core*-Variablen eingefügten wurden lässt sich erkennen, dass die Variablen ALTER, AGEBGEW, FNOH24, ADHEU, HOCHOZON, FSNIGHT, FSATEM und FSHLAUF eine immer wichtigere Rolle zur Erklärung des Response spielen. Dabei tragen besonders die Variablen FLGROSS, FLGEW, FMILB, FPOLL, FLTOTMED und FSPFEI im Vergleich zum original Datensatz weniger zur Erklärung bei.

Insbesondere streuen die Variablen FNOH24, HOCHOZON und FSATEM im Vergleich zur Abbildung 8.7 mehr. Diese Variablen weisen vor allem durch die Subsample-Methode eine größere Streuung auf. Interessant ist dabei, dass nur die Variable SEX von der Anzahl der Ausreißer nicht beeinflusst wurde. Dies gilt auch für die Ausreißer-Stärke und die Lage der Ausreißer. Für jedes mögliche Szenario war die relative Inklusions-Häufigkeit der Variablen SEX stets 1.

Für alle Szenarien gilt, dass die Ausreißer-Stärke, die Ausreißer-Menge und die Ausreißer-Lage die Wahl der Variablen, die in das Modell gewählt wurden, beeinflusst haben. Je mehr Ausreißer in den Datensatz eingefügt worden sind und je größer die Ausreißer-Stärke, desto mehr weichen die Ergebnisse von denen des original Datensatzes ab. Die relative Inklusions-Häufigkeit der *Non-Core*-Variablen steigt tendenziell an, während die *Core*-Variablen seltener ins Modell gewählt wurden. Eine große Bedeutung spielt dabei, welches Szenario man betrachtet. Die Ergebnisse der relativen Inklusions-Häufigkeiten der Variablen unterscheiden sich von Szenario zu Szenario und insbesondere bei den Extremfällen. Wie man in den Abbildungen im Anhang erkennen kann gibt es einige Besonderheiten. Die größten Unterschiede sind in den Abbildungen A.26 (Core), A.29 (Non-Core) und A.32(Core/Non-Core) zu erkennen, die 50 starke Ausreißer enthalten. Bei Betrachtung dieser Abbildungen wird nochmal deutlich, wie sehr sich die Ergebnisse unterscheiden, wenn die Ausreißer sich in unterschiedlichen Variablentypen befinden. Zusammenfassend ist zu erkennen, dass bei wenigen bzw. moderaten Ausreißern die Boxen sich im Bereich der original Ergebnisse befinden, bei Zunahme der Ausreißer-Stärke bzw. Ausreißer-Menge wird die Streuung größer und bei einer weiteren Erhöhung entfernen sich die Ergebnisse der modifizierten Datensätze drastischer von den original Datensatz Ergebnissen.

# 9. Diskussion und Ausblick

In dieser Arbeit wurde durch Resampling-basierte Rückwärtsselektion mittels $BIC$ bei multipler Regression die Gini-purity und die relative Inklusions-Häufigkeit der Variablen berrechnet, um den Einfluss der Ausreißer zu untersuchen. Bei Gegenüberstellung der Ergebnisse des original Datensatzes und der Ergebnisse der modifizierten Datensätze, welchen Ausreißer-Werte übergeben worden sind, konnten einige Unterschiede festgestellt werden. Aufgrund der Studien von De Bin et al. (2014) war zu vermuten, dass sich die Ausreißer negativ auf die Modellstabilität insbesondere auf die der Bootstrap-Methode auswirken würden.

Entgegen den Erwartungen wurde eine Verbesserung der Stabilität der Modellselektion durch Bootstrap-Verfahren in Abhängigkeit der Ausreißer-Anzahl und der Ausreißer-Stärke festgestellt. Dies konnte durch die Ergebnisse der Gini-purity beurteilt werden. Eine Aussage über die Modellstabilität die durch die Subsampling-Methode erzielt wurde, kann in diesem Zusammenhang nicht konkretisiert werden, da aufgrund der schwankenden Ergebnisse eine Tendenz schwer zu erkennen ist. Beim Vergleich der Ergebnisse aus den Analysen der relativen Inklusions-Häufigkeiten der Variablen ist durchaus ein Einfluss der Ausreißer auf die Resampling-basierte Variablenselektion zu erkennen. Durch die eingefügten Ausreißer gewinnen andere Variablen als beim ursprünglichen Datensatz zur Erklärung des Response an Bedeutung. In dieser Arbeit wurden 12 Szenarien je Ausreißer-Stärke untersucht. Es gibt natürlich die Möglichkeit andere Szenarien oder mehrere zu vergleichen. Desweiteren könnten die Ausreißer-Stärken auch anders gewählt werden, als es hier der Fall ist. Dieser Aspekt hätte wohlmöglich andere bzw. eindeutigere Ergebnisse zum Vorschein gebracht.

Die Variablenselektion mittels $BIC$ durchzuführen hatte zu einfache Modelle als Konsequenz. Tatsächlich hätte sich das Akaike Informationskriterium $AIC$ mehr angeboten, damit mehr Parameter in das Modell selektiert werden und dass somit der starke Strafterm des $BIC$ vermieden wird. Dieser wesentliche Unterschied dieser beiden Informationskriterien ist ein Grund, dass in der Praxis häufiger der $AIC$ verwendet wird.
(Fahrmeir et al., 2009, S. 489)

Die Rückwärtsselektion mittels $AIC$ würde wahrscheinlich andere Ergebnisse liefern, die vermutlich in diesem Kontext interessanter wären, um die Ergebnisse der Resampling-Methoden zu vergleichen. Ein weiterer Grund, dass beispielsweise die Tendenz der Subsampling-Methode auf die Stabilität der Modellselektion in Abhängigkeit der Ausreißer-Anzahl schwer zu erkennen ist, ist auch die Anzahl der modifizierten Datensätze je Szenario. Für diese Analysen wurden 50 modifizierte Datensätze je Szenario erzeugt, es würde sich empfehlen mehr als das Doppelte an modifizierten Datensätzen für jedes Szenario zu erzeugen. Je mehr Ergebnisse berechnet werden, desto aussagekräftiger sind die Auswirkungen der Ausreißer auf die Modellstabilität der Resampling-basierten Variablenselektion.

Weiterhin könnten andere Kennzahlen berrechnet werden, um zu untersuchen, wie sich Ausreißer in dieser empirischen Studie auswirken. Beispielsweise könnte der $AUC$, wie in (De Bin et al., 2014), zur Analyse herangezogen werden.

# Literaturverzeichnis


Albers, S., Klapper, D., Konradt, U., Walter, A. & Wolf, J. (Hrsg.) (2009). *Methodik der empirischen Forschung*. (3., überarb. u. erw. Aufl.). Wiesbaden: Gabler.

Aßfalg, J., Böhm, C., Borgwardt, K., Ester, M., Januzaj, E., Kailing, K., Kröger, P., Sander, J. & Schubert, M. (2003). Skript zur Vorlesung Knowledge Discovery in Databases im Wintersemester 2009/2010. Kapitel 3: Klassifikation. Download am [06.07.15] von http://www.dbs.ifi.lmu.de/Lehre/KDD/WS0910/skript/kdd-3-klassifikation.pdf

De Bin, R., Janitza, S., Sauerbrei, W. & Boulesteix, A.-L. (2014). Subsampling versus bootstrapping in resampling-based model selection for multivariable regression. *Technical Report* **171**, *Biometrics* (akzeptiert) im Druck.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics* **7**, 1-26.

Efron, B. & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. New York: Chapman & Hall/CRC.

Fahrmeir, L., Kneib, T. & Lang, S. (2009). *Regression. Modelle, Methoden und Anwendungen*. (2. Aufl.). Berlin/Heidelberg: Springer.

Fleischer, K. & Folda, R. (1996). Einsatzmöglichkeiten des Bootstrap-Verfahrens bei der Bonitätsprüfung. *Operations-Research-Spektrum* **18**, 107-115. **DOI: 10.1007/BF01539735**

Geßler, J.R. (1993). *Statistische Graphik*. Basel: Springer.

Groß, J. (2010).*Grundlegende Statistik mit R. Eine anwendungsorientierte Einführung in die Verwendung der Statistik Software R*. Wiesbaden: Vieweg+Teubner.

Hosemann, T. (1807-1875) Urheber. Download am [06.07.15] von http://3.bp.blogspot.com/-dJEQZ0Cr-fI/TcvYKkmziGI/AAAAAAAAFjE/7mike3i6Wh4/s400/M%2525C3%2525BCnchhausen-Sumpf-Hosemann.PNG

Ihorst, G., Frischer, T., Horak, F., Schumacher, M., Kopp, M., Forster, J., Mattes, J. & Kühr, J. (2004). Long-and medium-term ozone effects on lung growth including a broad spectrum of exposure. *European Respiratory Journal* **23**, 292-299.

Kellner, W. (2009). Wichtige Lungenfunktionswerte. Letzter Abruf: [09.07.2015] von http://www.asthma.medhost.de/lungenfunktion.html

Mantel, N. (1970). Why stepdown procedures in variable selection. *Technometrics* **12**, 621-625.

Pruscha, H. (2006). *Statistisches Methodenbuch. Verfahren, Fallstudien, Programmcodes*. Berlin/Heidelberg/New York: Springer.

Ripley, B., Venables, B., Bates, D. M., Hornik, K., Gebhardt, A. & Firth, D. (2014). *MASS: Support Functions and Datasets for Venables and Ripley's MASS*. R Paket Version 7.3-35.

Schlittgen, R. (2013). *Regressionsanalysen mit R*. München: Oldenbourg.

Schneider, A., Hommel, G. & Blettner, M. (2010). Linear regression analysis. Part 14 of a series on evaluation of scientific publications. *Dtsch Arztebl Int* **107(44)**, 776–82. **DOI: 10.3238/arztebl.2010.0776**

Tukey, J. (1977). *Exploratory data analysis*. New York: Addison-Wesley.

Unkel, S. (2013). Kapitel 2: Likelihood-Inferenz (Fortsetzung). Download am [06.07.15] von http://www.statistik.lmu.de/institut/ag/biostat/vorlesungen/WS1314/StatistikIIINebenfach/vorlesung/kap2/Kap2Folien2.pdf

Wittmann, P., (2010). Das Testen der Martingaleigenschaft. In E. Bomsdorf, W. Kösters, W. Matthes & M. Trede (Hrsg.), *Quantitative Ökonomie*. (Bd. 162). Köln: JOSEF EUL.

Wu, C.-F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics* **14**, 1261-1295.

# Anhang

# A. Abbildungen

Abbildung A.1.: Szenario 1 mit Faktor 2

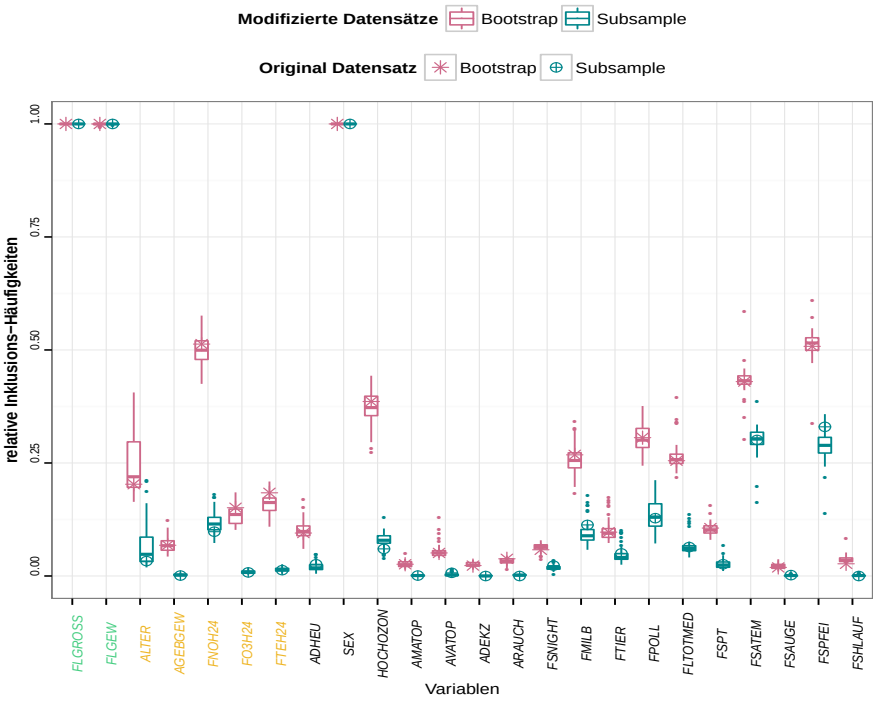



Abbildung A.2.: Szenario 1 mit Faktor 10

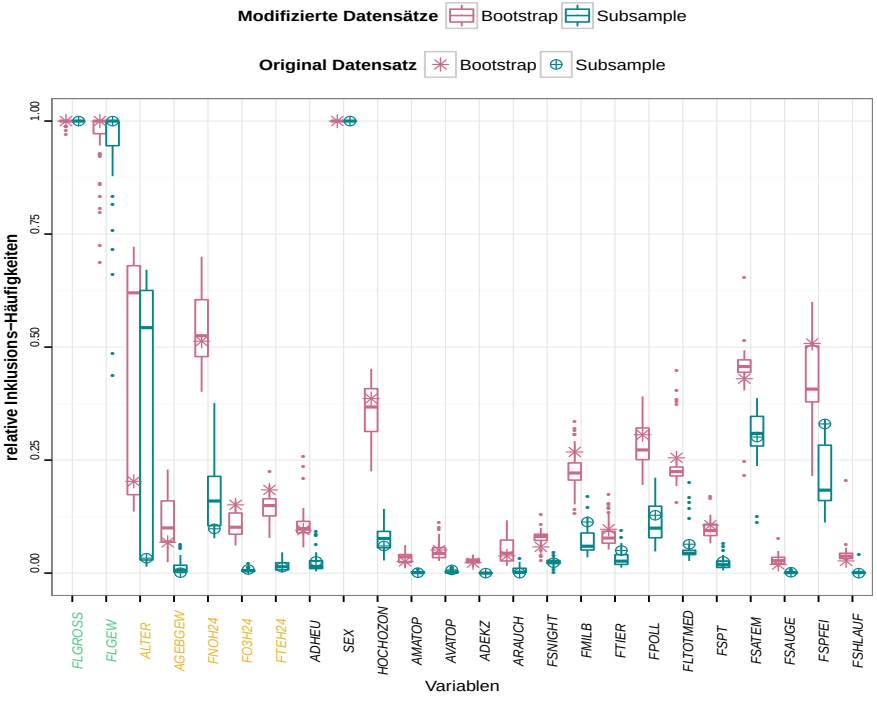


Abbildung A.3.: Szenario 2 mit Faktor 2

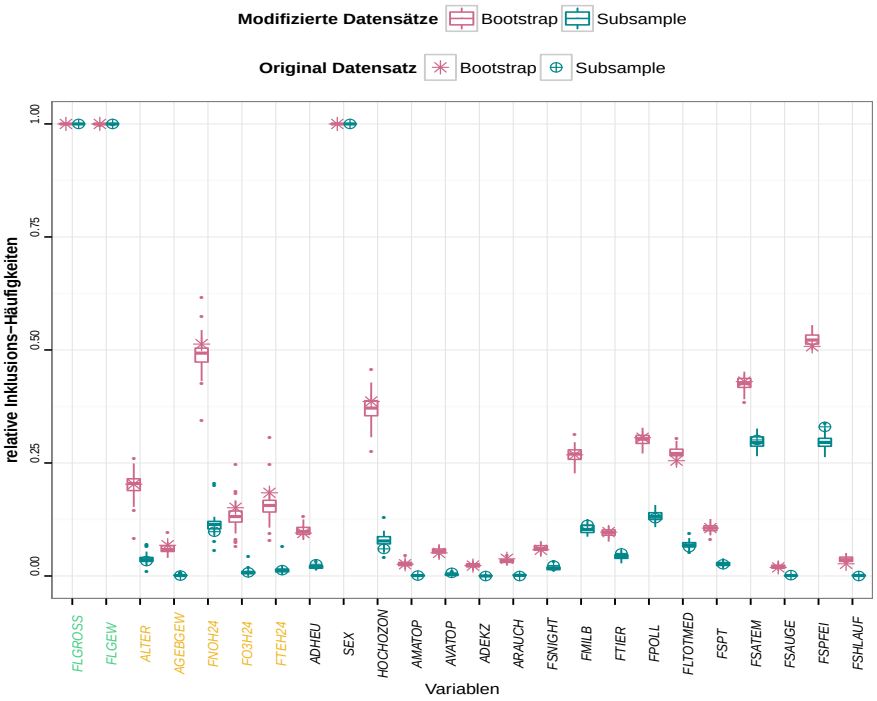



Abbildung A.4.: Szenario 2 mit Faktor 5

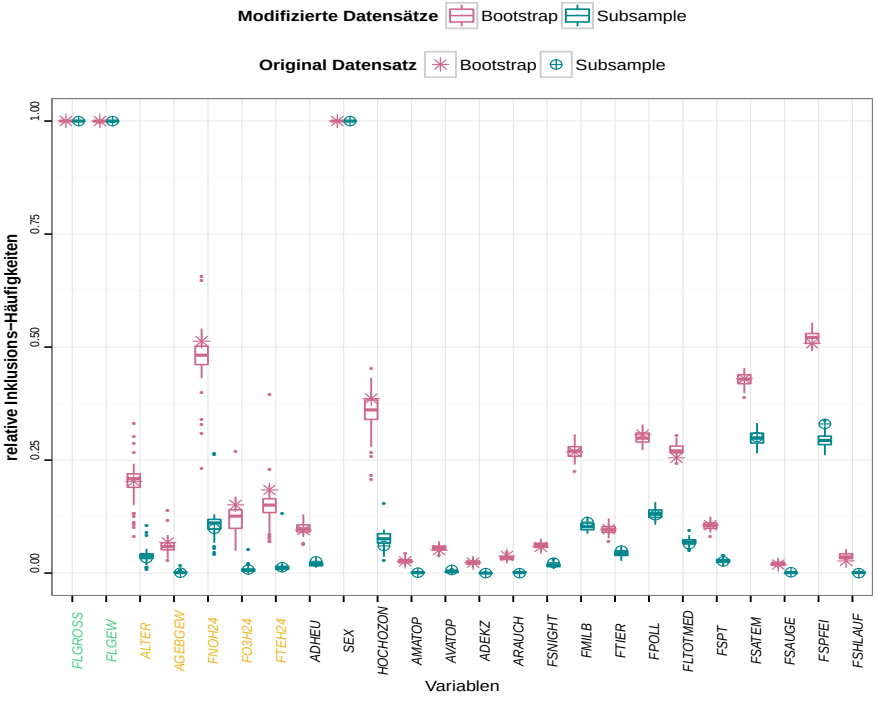


Abbildung A.5.: Szenario 2 mit Faktor 10

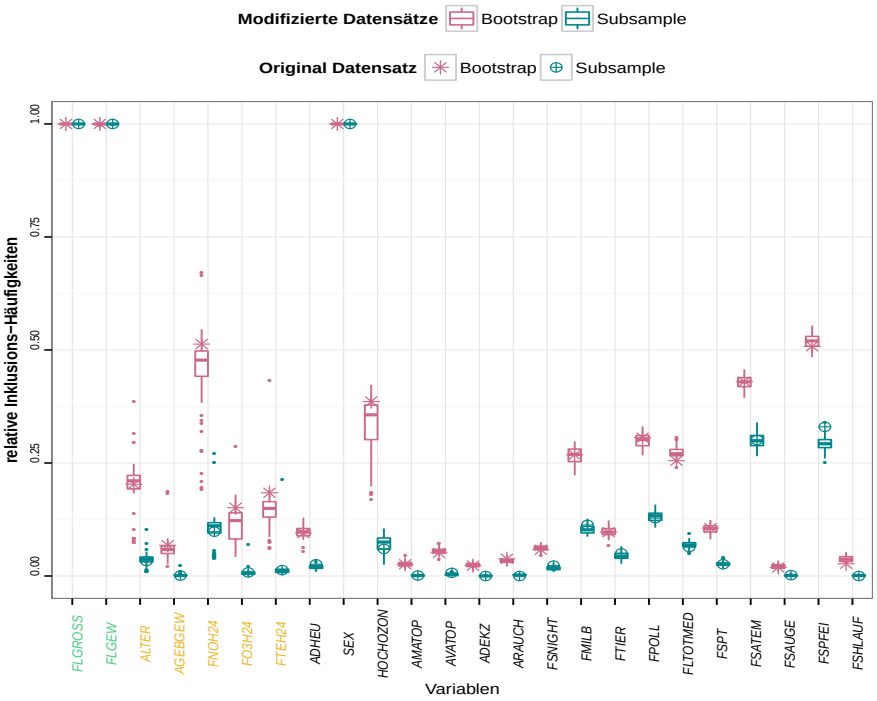



Abbildung A.6.: Szenario 3 mit Faktor 2

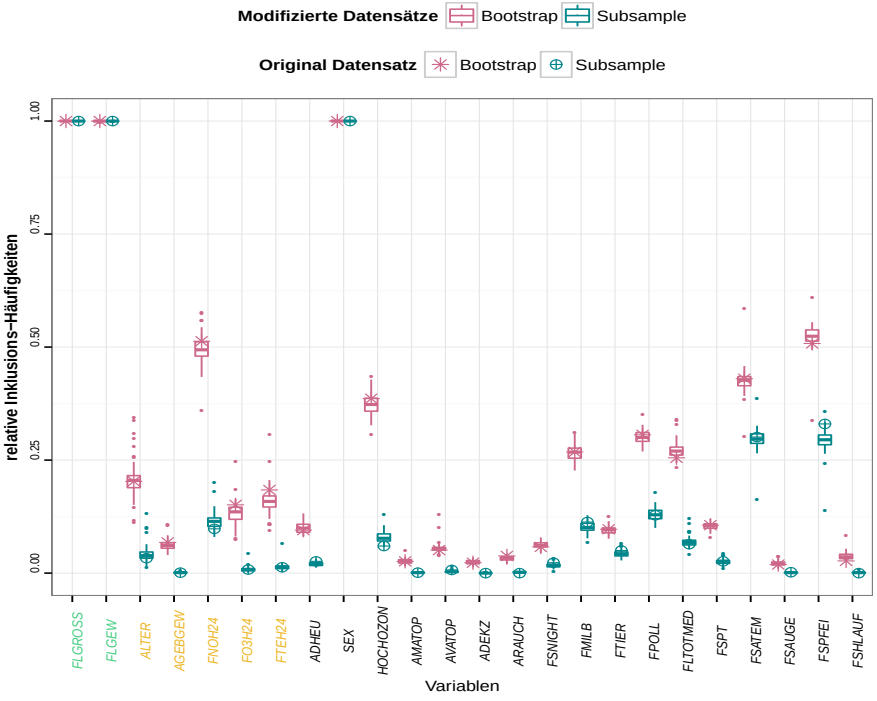


Abbildung A.7.: Szenario 3 mit Faktor 5

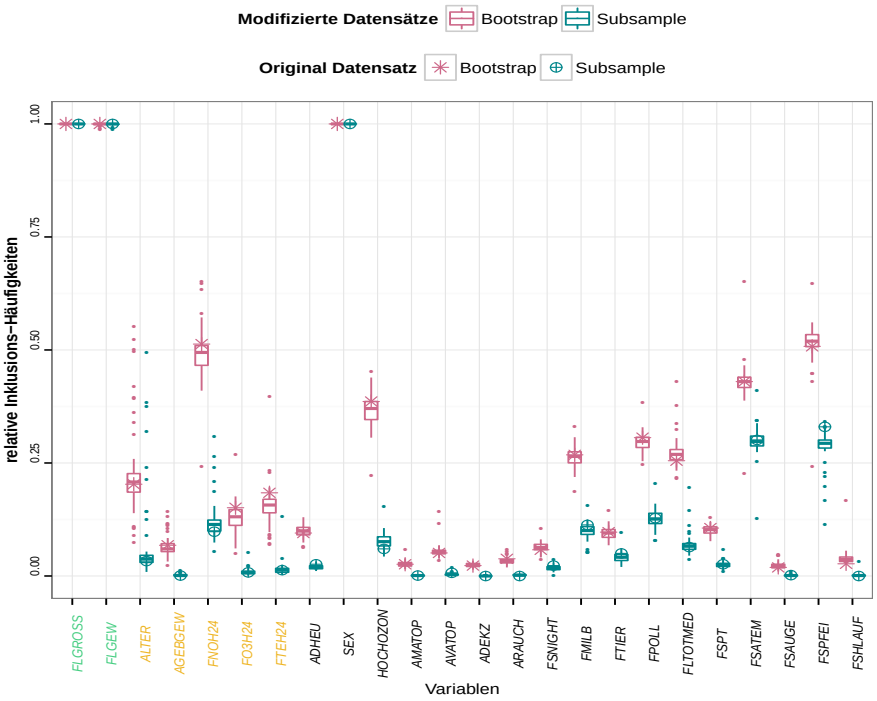



Abbildung A.8.: Szenario 3 mit Faktor 10

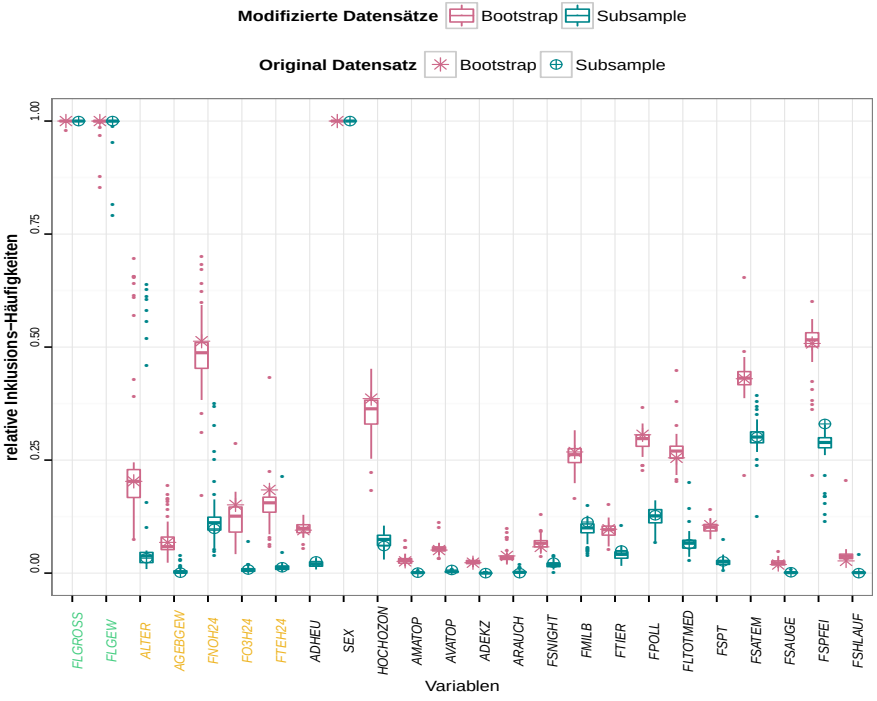


Abbildung A.9.: Szenario 4 mit Faktor 2

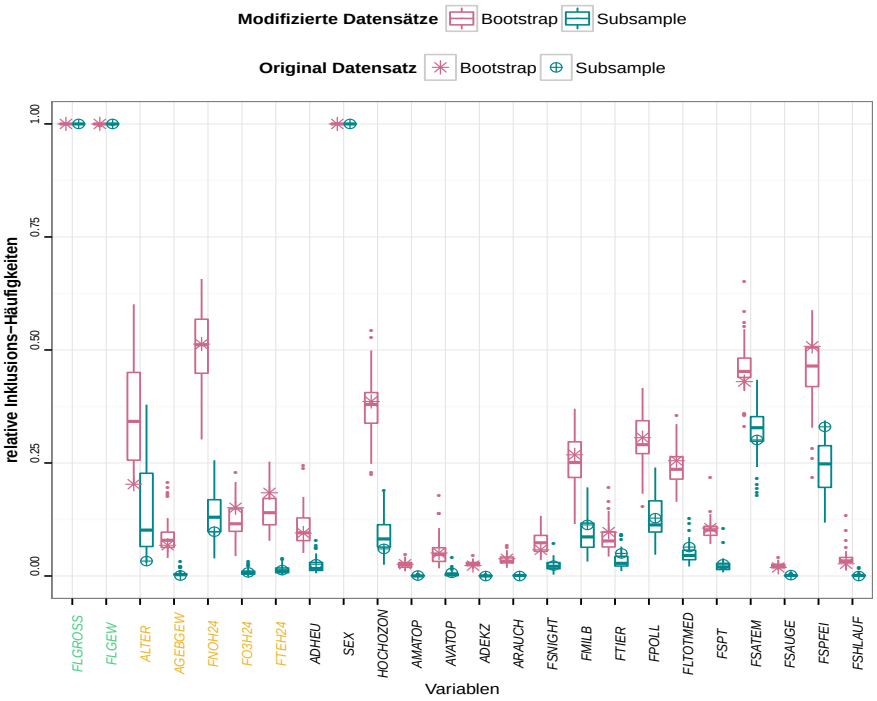



Abbildung A.10.: Szenario 3 mit Faktor 10

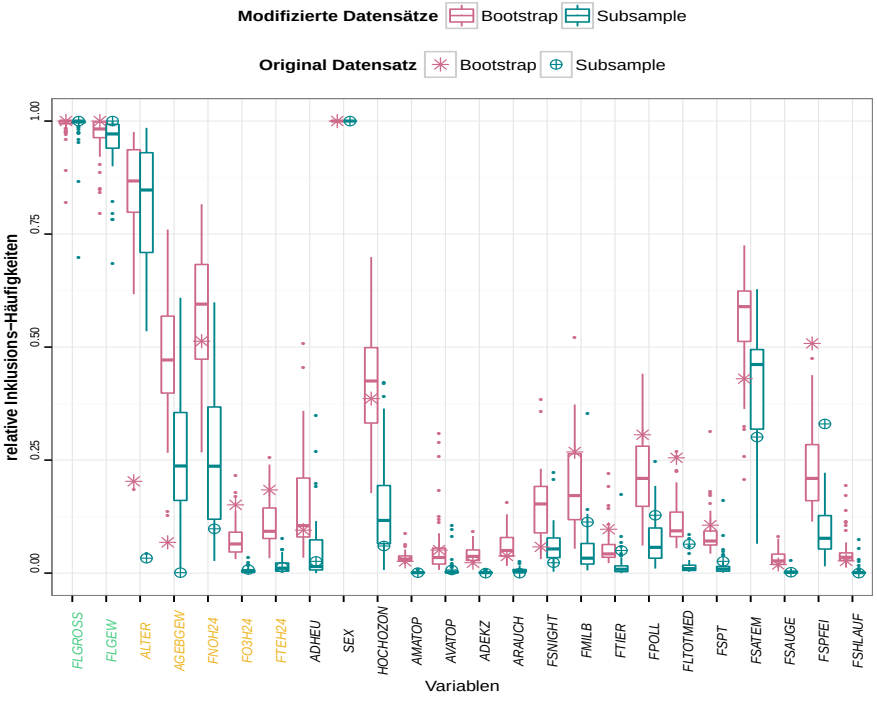


Abbildung A.11.: Szenario 5 mit Faktor 2

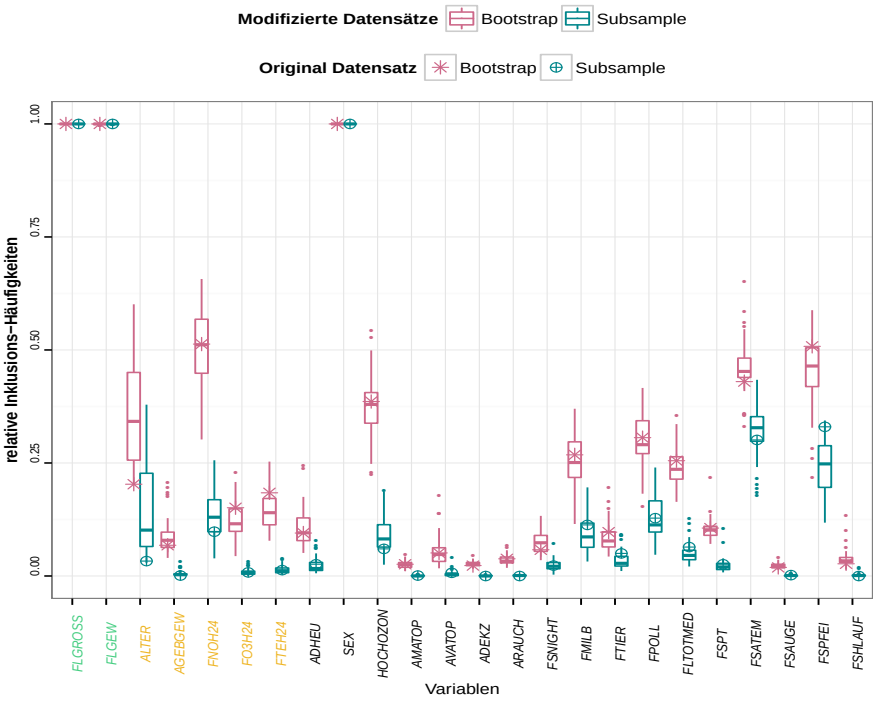



Abbildung A.12.: Szenario 5 mit Faktor 5

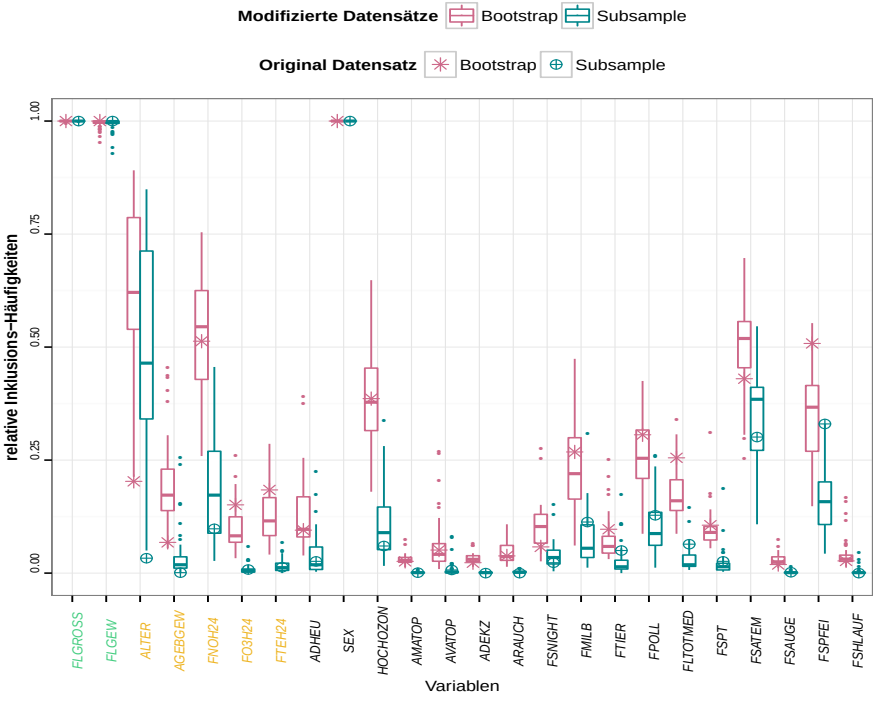


Abbildung A.13.: Szenario 5 mit Faktor 10

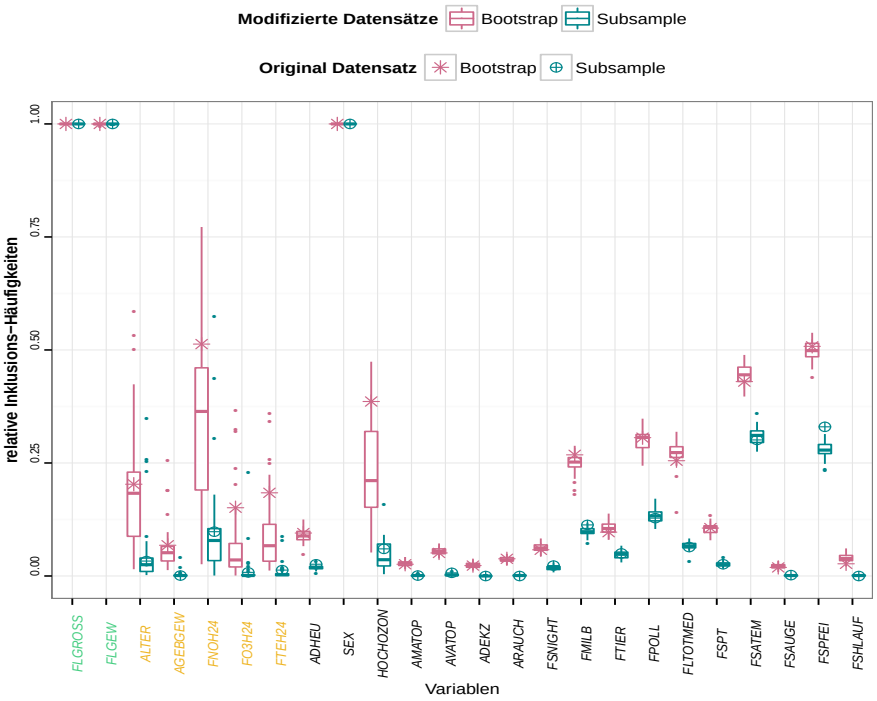



Abbildung A.14.: Szenario 6 mit Faktor 2

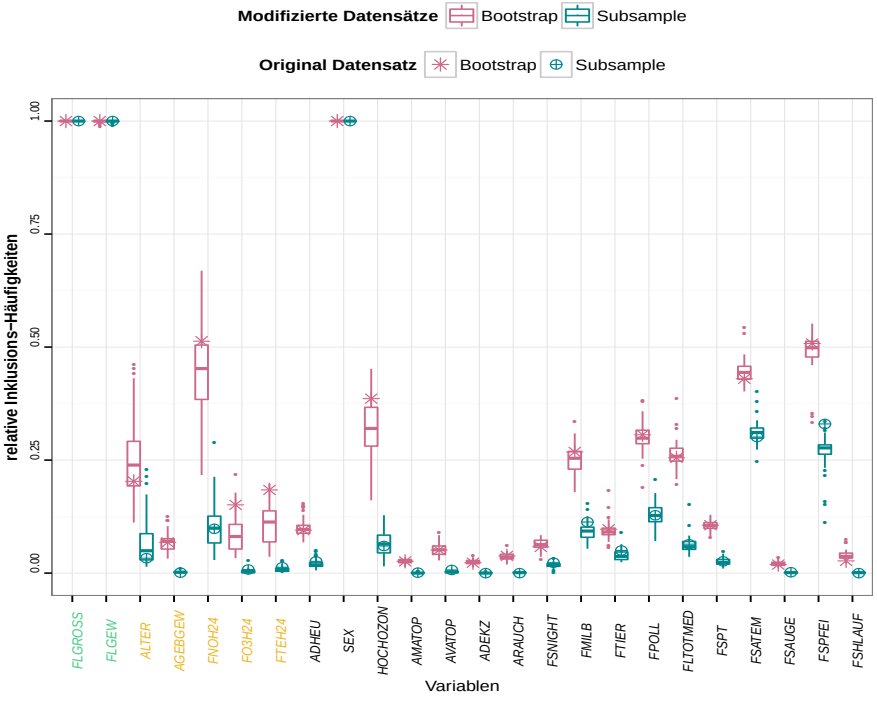


Abbildung A.15.: Szenario 6 mit Faktor 5

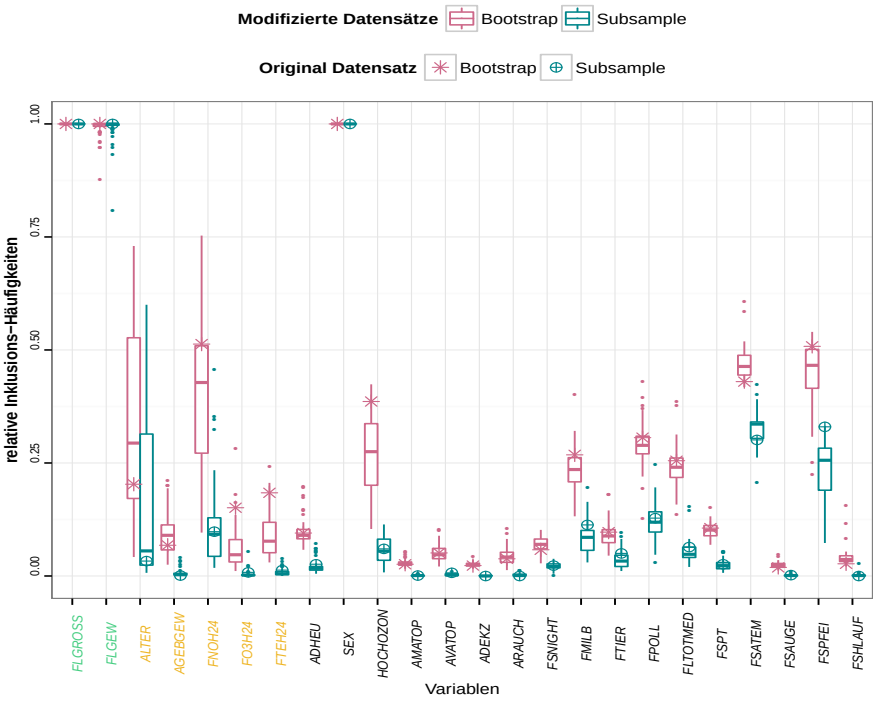



Abbildung A.16.: Szenario 6 mit Faktor 10

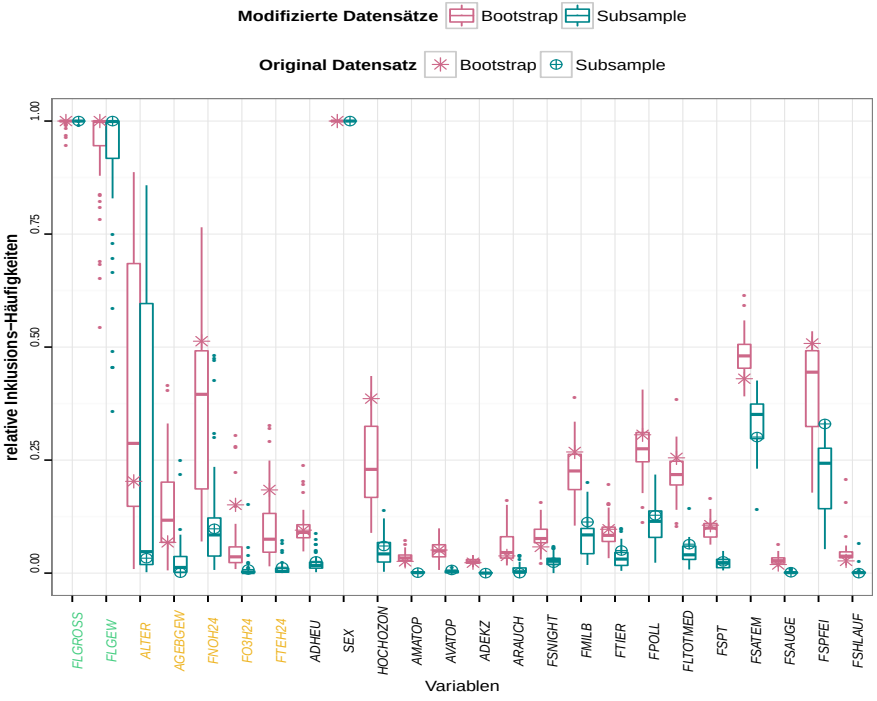


Abbildung A.17.: Szenario 7 mit Faktor 2

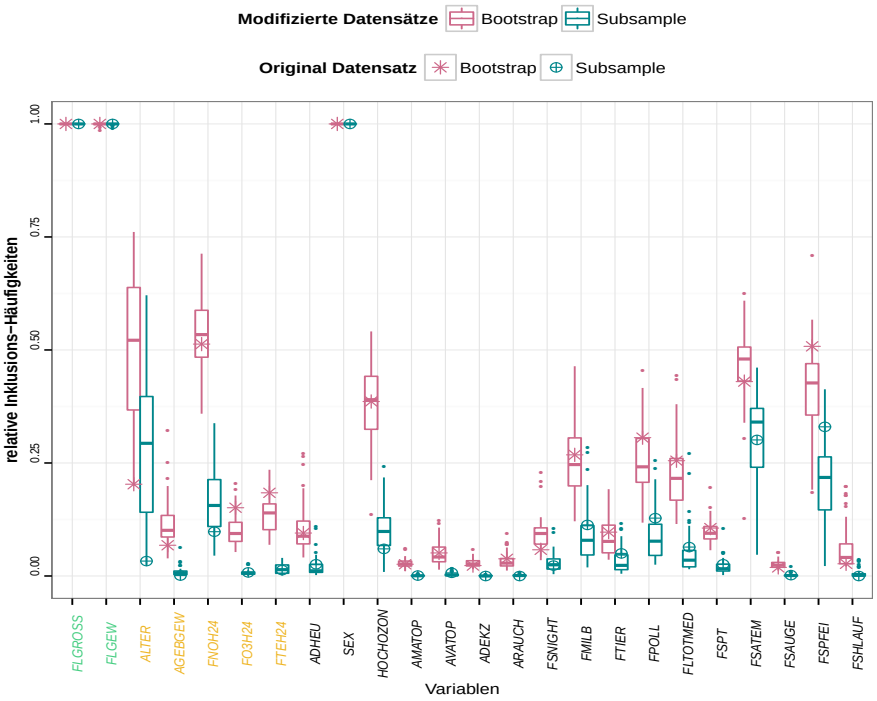



Abbildung A.18.: Szenario 7 mit Faktor 10

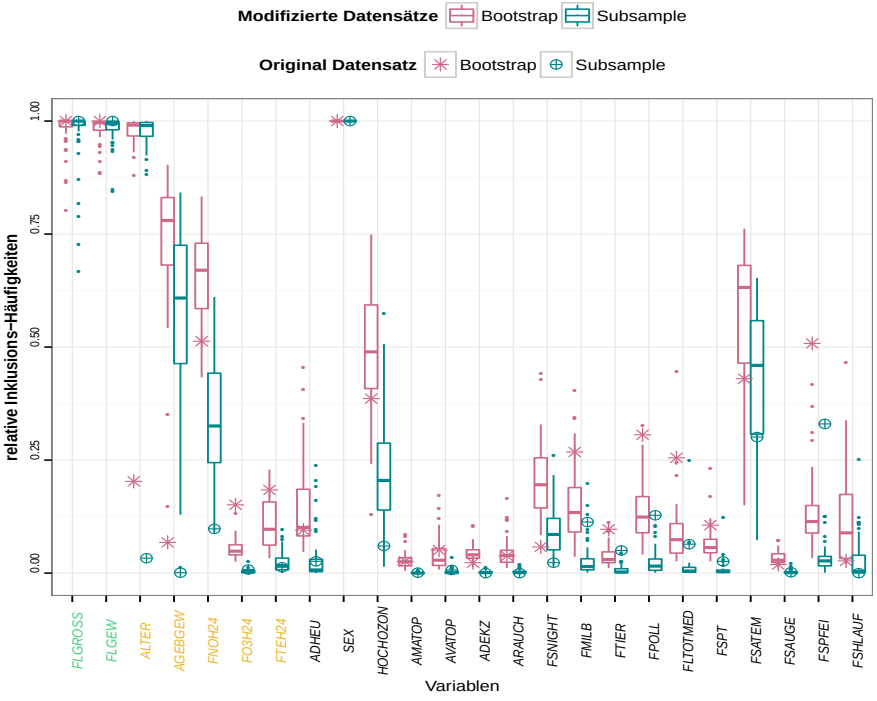


Abbildung A.19.: Szenario 8 mit Faktor 2

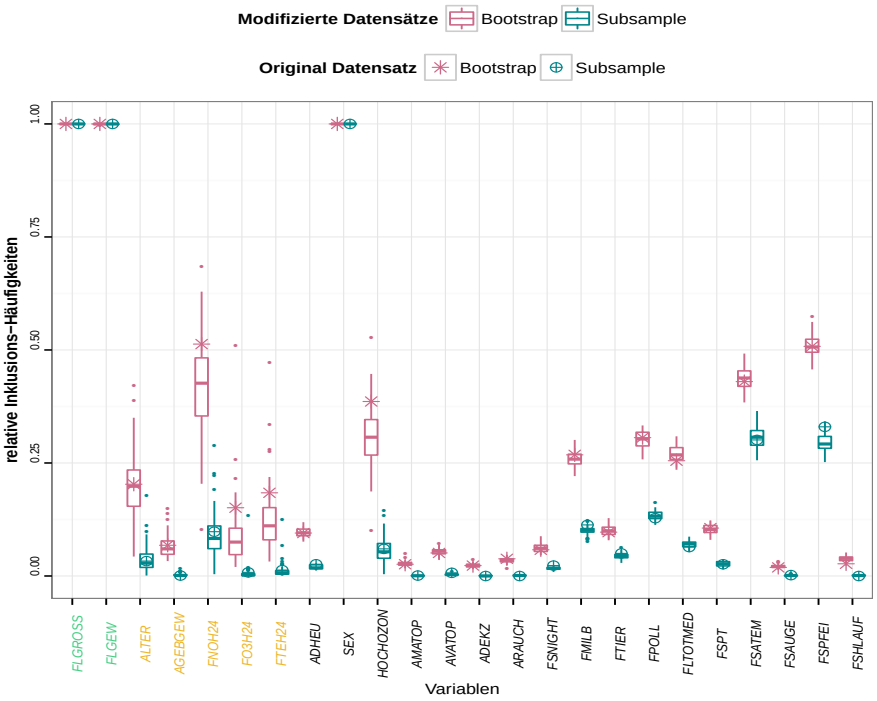



Abbildung A.20.: Szenario 8 mit Faktor 5

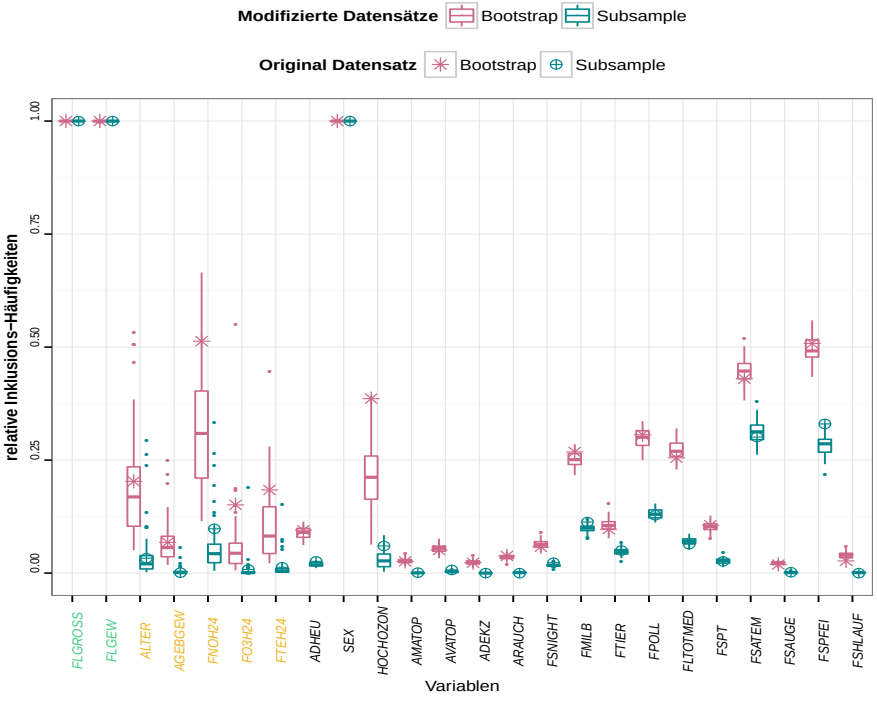


Abbildung A.21.: Szenario 8 mit Faktor 10

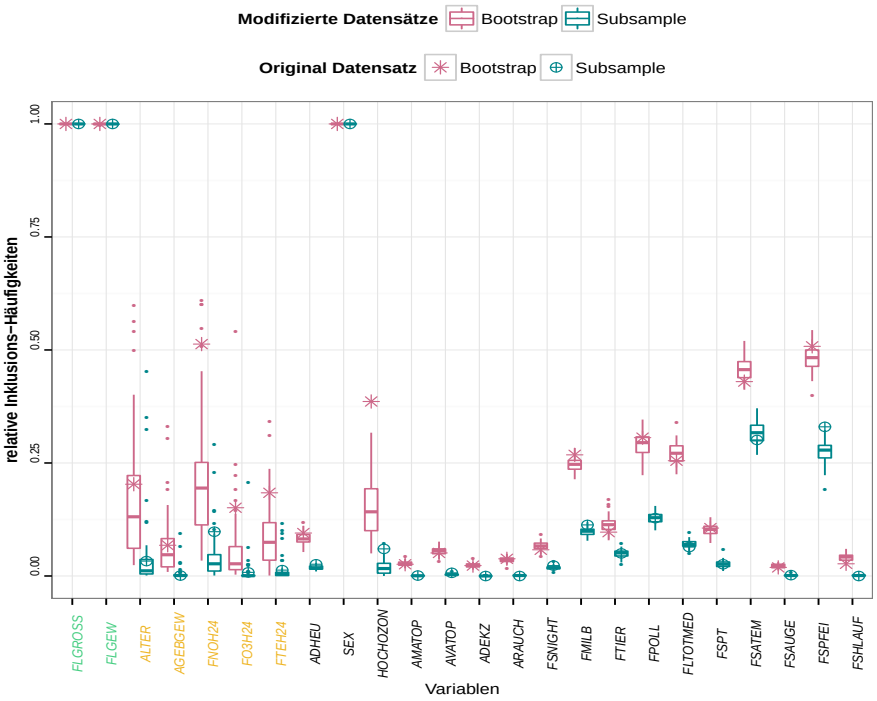



Abbildung A.22.: Szenario 9 mit Faktor 2

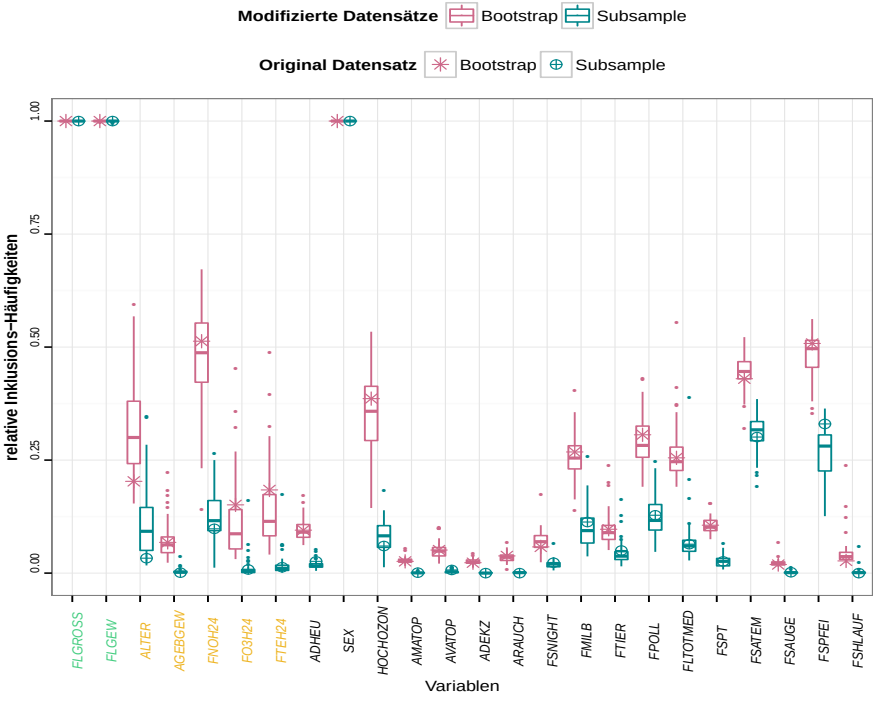


Abbildung A.23.: Szenario 9 mit Faktor 5

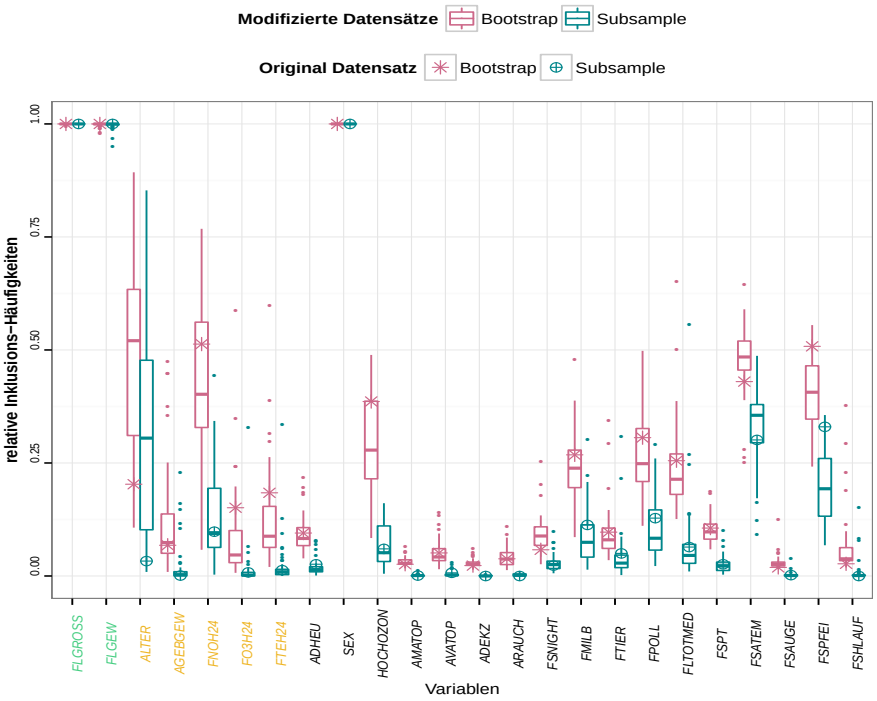



Abbildung A.24.: Szenario 9 mit Faktor 10

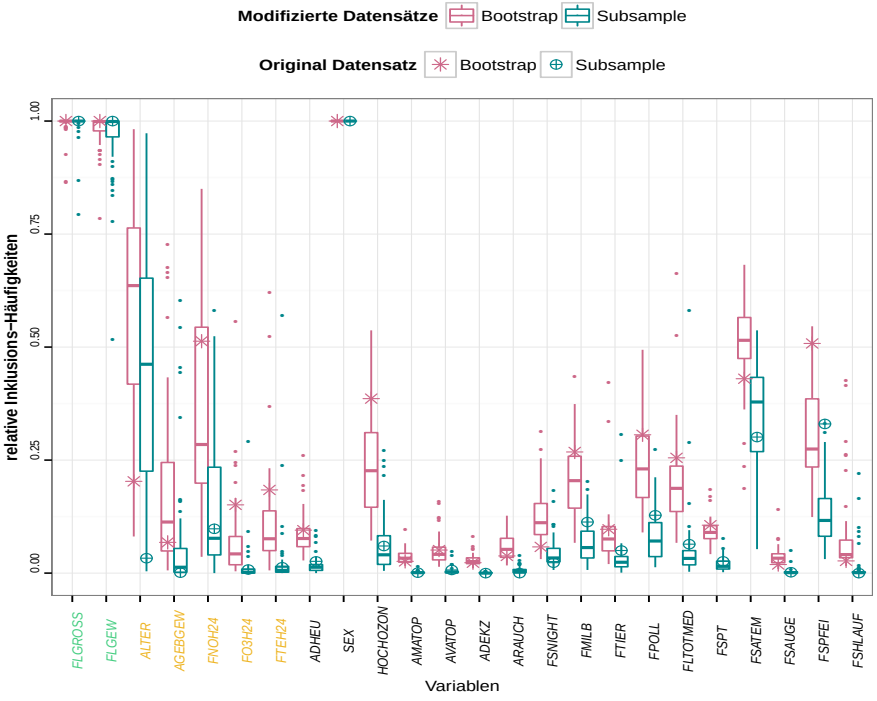


Abbildung A.25.: Szenario 10 mit Faktor 2

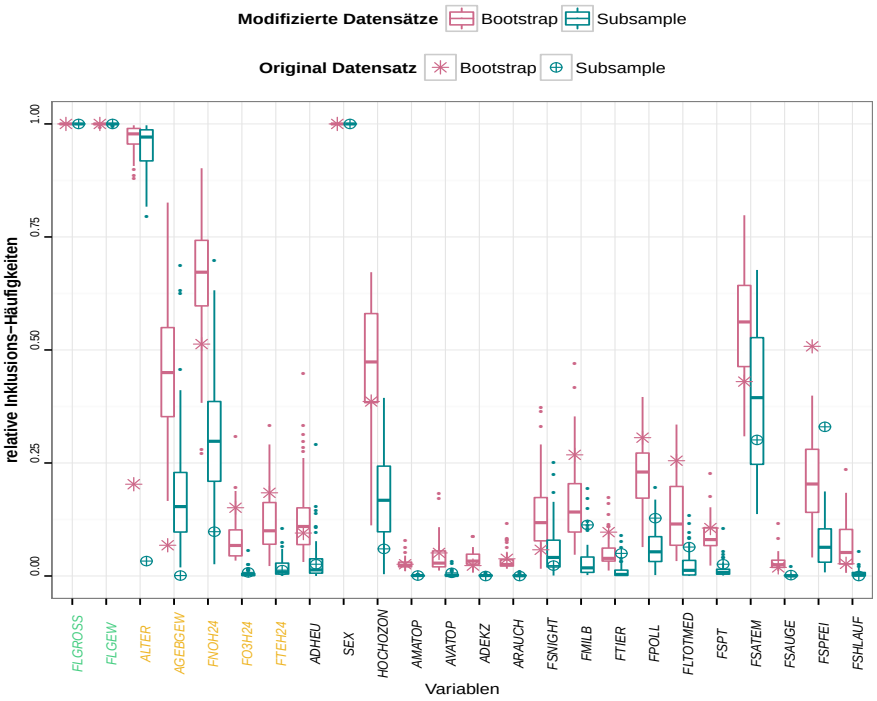



Abbildung A.26.: Szenario 10 mit Faktor 10

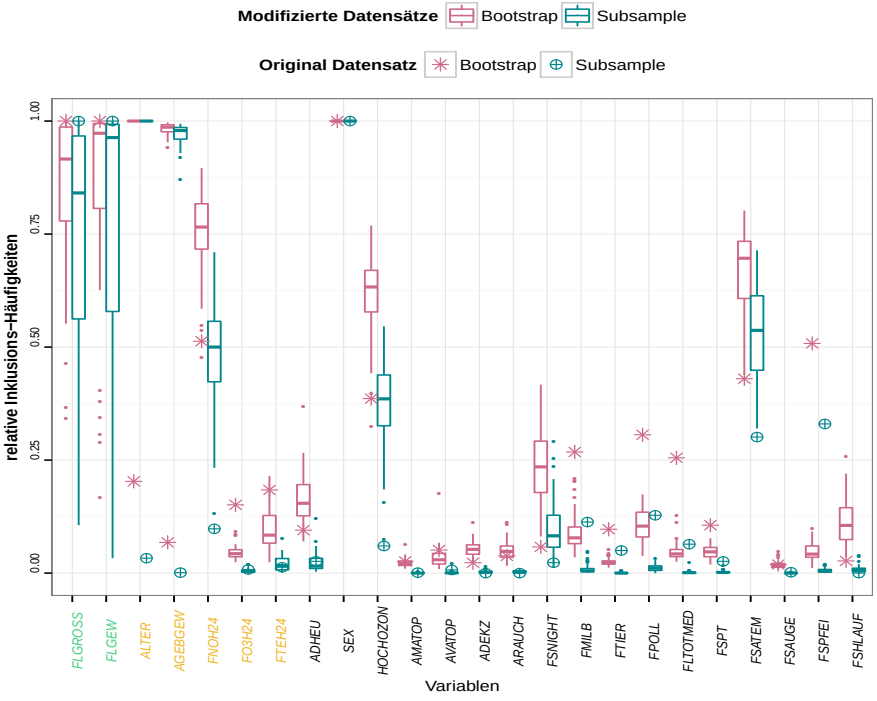


Abbildung A.27.: Szenario 11 mit Faktor 2

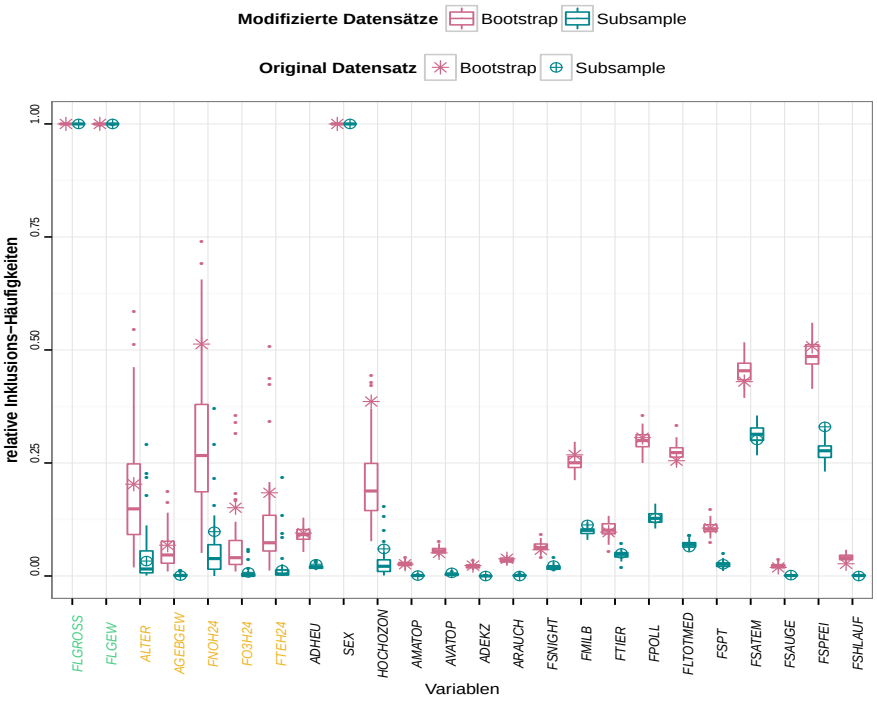



Abbildung A.28.: Szenario 11 mit Faktor 5

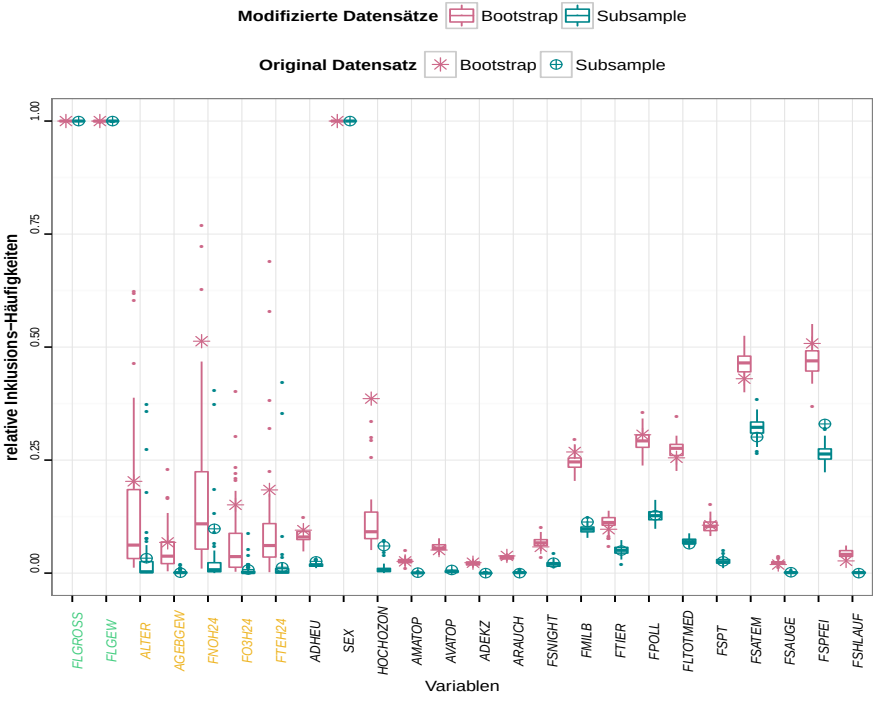


Abbildung A.29.: Szenario 11 mit Faktor 10

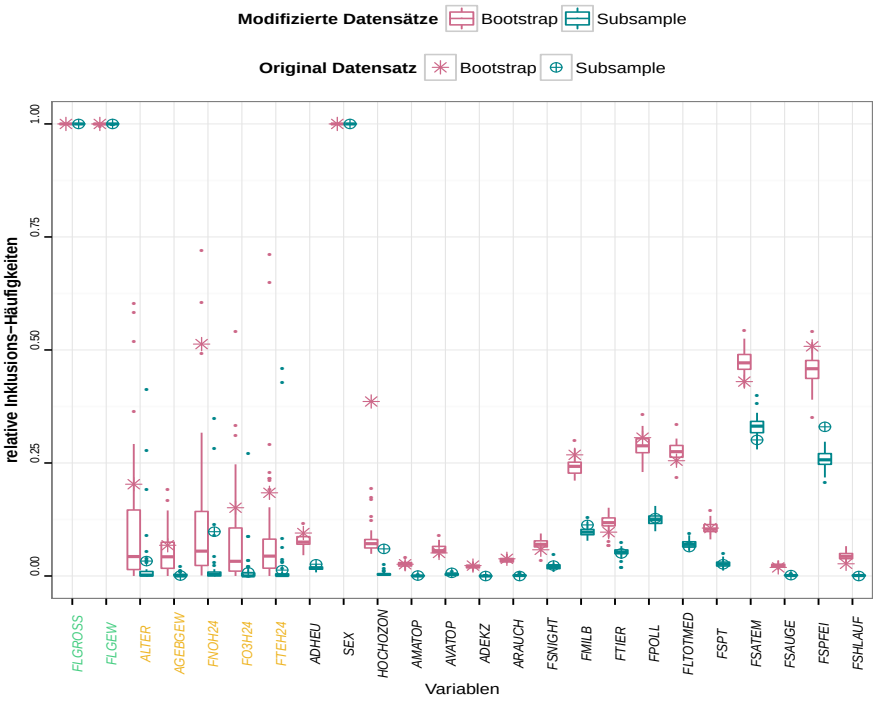



Abbildung A.30.: Szenario 12 mit Faktor 2

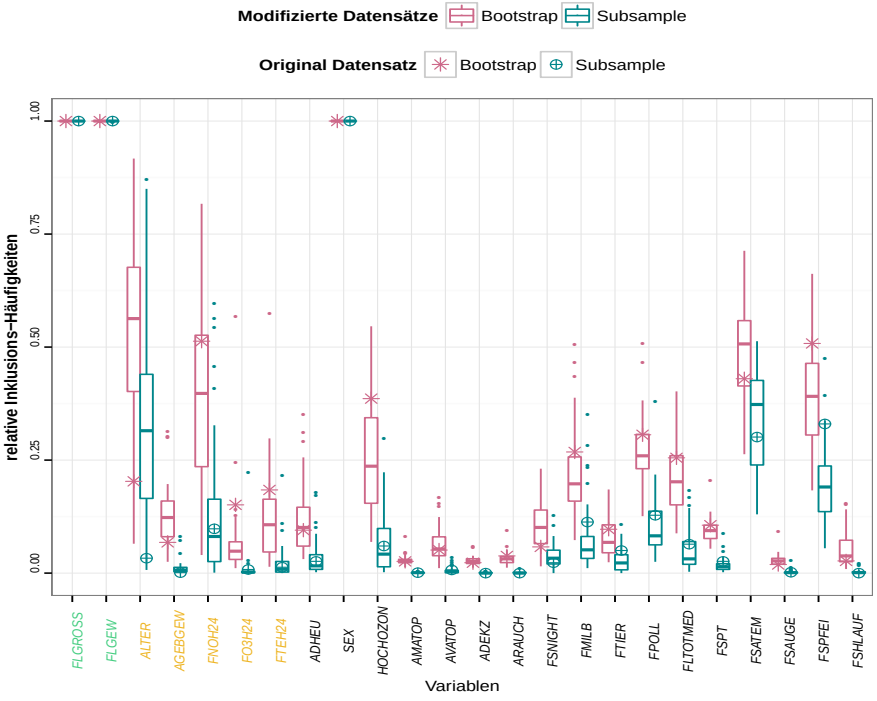


Abbildung A.31.: Szenario 12 mit Faktor 5

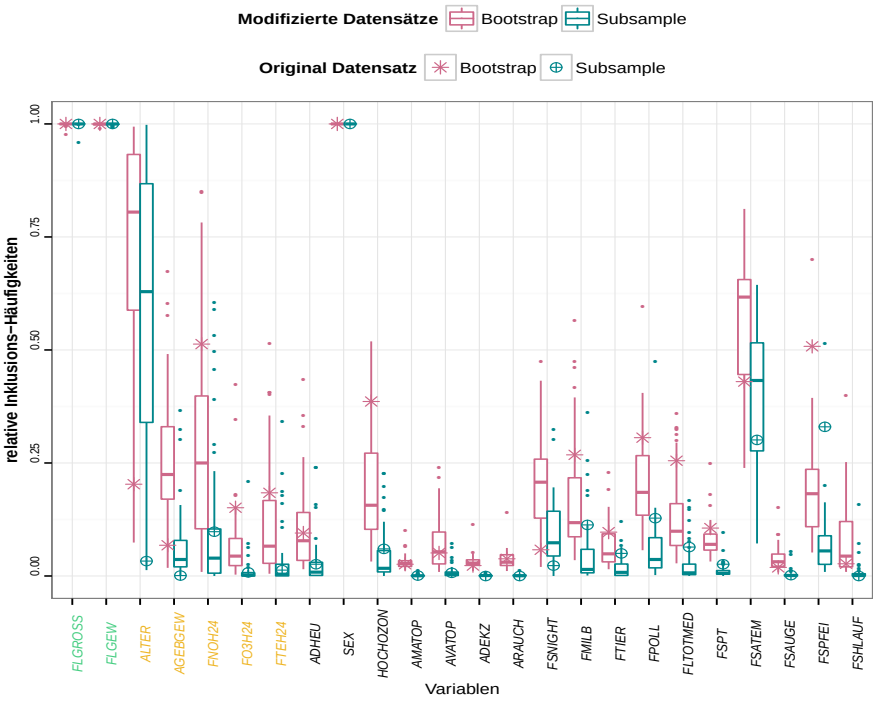
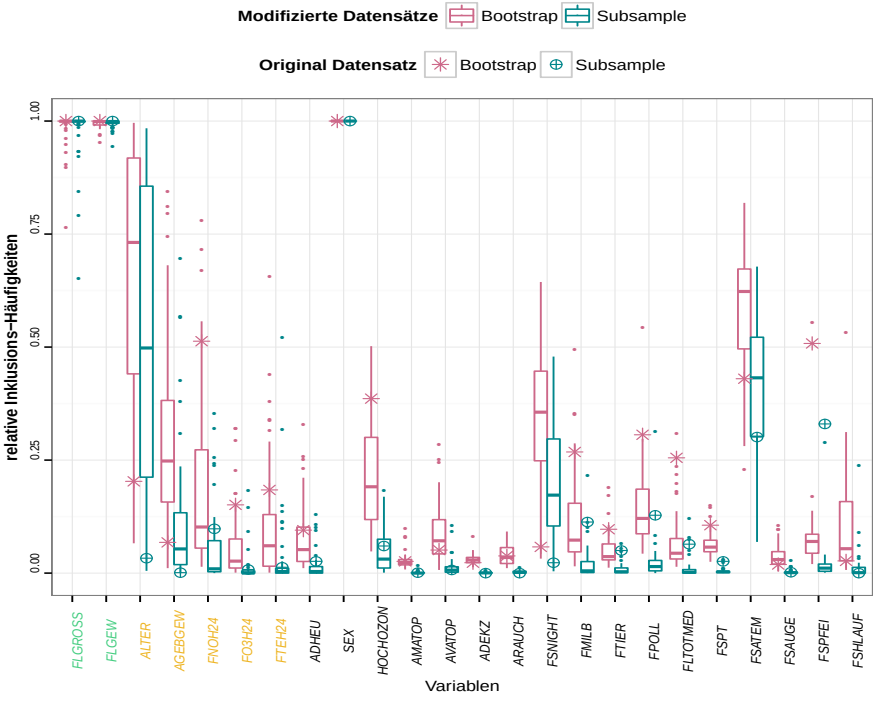



Abbildung A.32.: Szenario 12 mit Faktor 10



# B. Digitaler Anhang

Auf der beigefügten CD-ROM befindet sich folgender Inhalt:

1.) **Ozon Datensatz, Ergebnisse und Ausreißer-Abbildungen:**

- *ozone_reduced.txt*: zur Verfügung gestellter Datensatz
- *original.R*: beinhaltet die Resampling-basierte Variablenselektion für den ursprünglichen Datensatz und den Code für die Grafiken
- *original_ergebnis_n1000.csv*: Ergebnisse des Ozon-Datensatzes
- Grafiken in .pdf Format:

  core-Variablen

  non-core-Variablen

2.) **Szenario R Skripte:** Programmcode für die Simulation der modifizierten Datensätze

- *faktor2szenario1.R*
- *faktor2szenario2.R*
- *faktor2szenario3.R*
- *faktor2szenario4.R*
- *faktor2szenario5.R*
- *faktor2szenario6.R*
- *faktor2szenario7.R*
- *faktor2szenario8.R*
- *faktor2szenario9.R*
- *faktor2szenario10.R*
- *faktor2szenario11.R*
- *faktor2szenario12.R*
- *faktor5szenario1.R*
- *faktor5szenario2.R*
- *faktor5szenario3.R*
- *faktor5szenario4.R*
- *faktor5szenario5.R*
- *faktor5szenario6.R*
- *faktor5szenario7.R*
- *faktor5szenario8.R*
- *faktor5szenario9.R*
- *faktor5szenario10.R*
- *faktor5szenario11.R*
- *faktor5szenario12.R*
- *faktor10szenario1.R*
- *faktor10szenario2.R*
- *faktor10szenario3.R*
- *faktor10szenario4.R*
- *faktor10szenario5.R*
- *faktor10szenario6.R*
- *faktor10szenario7.R*
- *faktor10szenario8.R*
- *faktor10szenario9.R*
- *faktor10szenario10.R*
- *faktor10szenario11.R*
- *faktor10szenario12.R*

3.) **Szenario R Ergebnisse:** Ergebnisse der modifizierten Datensätze als csv.-Dateien

- *szenario1 mit faktor2*
- *szenario2 mit faktor2*
- *szenario3 mit faktor2*
- *szenario4 mit faktor2*
- *szenario5 mit faktor2*
- *szenario6 mit faktor2*
- *szenario7 mit faktor2*
- *szenario8 mit faktor2*
- *szenario9 mit faktor2*
- *szenario1 mit faktor5*
- *szenario2 mit faktor5*
- *szenario3 mit faktor5*
- *szenario4 mit faktor5*
- *szenario5 mit faktor5*
- *szenario6 mit faktor5*
- *szenario7 mit faktor5*
- *szenario8 mit faktor5*
- *szenario9 mit faktor5*
- *szenario10 mit faktor2*
- *szenario11 mit faktor2*
- *szenario12 mit faktor2*
- *szenario10 mit faktor5*
- *szenario11 mit faktor5*
- *szenario12 mit faktor5*
- *szenario1 mit faktor10*
- *szenario2 mit faktor10*
- *szenario3 mit faktor10*
- *szenario4 mit faktor10*
- *szenario5 mit faktor10*
- *szenario6 mit faktor10*
- *szenario7 mit faktor10*
- *szenario8 mit faktor10*
- *szenario9 mit faktor10*
- *szenario10 mit faktor10*
- *szenario11 mit faktor10*
- *szenario12 mit faktor10*
- *melt*-Datensätze zu allen Szenarien wurden mittels folgender R-Skripte erstellt.

4.) **R Skripte für die Abbildungen der Inklusions-Häufigkeiten:**

- *Grafik Inklusion Faktor 2.R*
- *Grafik Inklusion Faktor 5.R*
- *Grafik Inklusion Faktor 10.R*

5.) **Abbildungen Inklusions-Häufigkeiten der Szenarien:** Grafiken in pdf.-Format

- Szenario1 und faktor2 häufigkeiten
- Szenario2 und faktor2 häufigkeiten
- Szenario3 und faktor2 häufigkeiten
- Szenario4 und faktor2 häufigkeiten
- Szenario5 und faktor2 häufigkeiten
- Szenario6 und faktor2 häufigkeiten
- Szenario7 und faktor2 häufigkeiten
- Szenario8 und faktor2 häufigkeiten
- Szenario9 und faktor2 häufigkeiten
- Szenario10 und faktor2 häufigkeiten
- Szenario11 und faktor2 häufigkeiten
- Szenario12 und faktor2 häufigkeiten
- Szenario1 und faktor5 häufigkeiten
- Szenario2 und faktor5 häufigkeiten
- Szenario3 und faktor5 häufigkeiten
- Szenario4 und faktor5 häufigkeiten

- Szenario5 und faktor5 häufigkeiten
- Szenario6 und faktor5 häufigkeiten
- Szenario7 und faktor5 häufigkeiten
- Szenario8 und faktor5 häufigkeiten
- Szenario9 und faktor5 häufigkeiten
- Szenario10 und faktor5 häufigkeiten
- Szenario11 und faktor5 häufigkeiten
- Szenario12 und faktor5 häufigkeiten
- Szenario1 und faktor10 häufigkeiten
- Szenario2 und faktor10 häufigkeiten
- Szenario3 und faktor10 häufigkeiten
- Szenario4 und faktor10 häufigkeiten
- Szenario5 und faktor10 häufigkeiten
- Szenario6 und faktor10 häufigkeiten
- Szenario7 und faktor10 häufigkeiten
- Szenario8 und faktor10 häufigkeiten
- Szenario9 und faktor10 häufigkeiten
- Szenario10 und faktor10 häufigkeiten
- Szenario11 und faktor10 häufigkeiten
- Szenario12 und faktor10 häufigkeiten

6.) **Gini-purity Dateien, Skripte und Abbildungen:**

- *Grafik Gini alle Faktoren.R*: verknüpft je Faktor alle Szenarien-Datensätze und enthält Code für die Erstellung der Grafiken
- *faktor2.csv*: Datensatz der Ergebnisse für Faktor 2
- *faktor5.csv*: Datensatz der Ergebnisse für Faktor 5
- *faktor10.csv*: Datensatz der Ergebnisse für Faktor 10
- Grafiken in pdf.-Format:

  gini faktor2

  gini faktor5

  gini faktor10

  ginialle

Die Reihenfolge der aufgelisteten Inhalte soll bitte auch für das Laden der R-Skripte eingehalten werden.

# C. Eigenständigkeitserklärung

Hiermit versichere ich, Alma Sehic, die vorliegende Bachelorarbeit selbstständig und ausschließlich unter Benutzung der angegebenen Quellen und Hilfsmittel verfasst habe.
Die aus fremden Quellen (einschließlich elektronischer Quellen) direkt oder indirekt übernommenen Gedanken sind ausnahmslos als solche kenntlich gemacht. Die Arbeit ist in gleicher oder ähnlicher Form oder auszugsweise im Rahmen einer anderen Prüfung noch nicht vorgelegt worden.

München, den 04.08.2015

Alma Sehic