technische universität
dortmund

fakultät
statistik

Statistische Analyse von MCC-IMS-Messungen

Dissertation

zur Erlangung des akademischen Grades
Doktor der Naturwissenschaften
an der Fakultät Statistik
der Technischen Universität Dortmund

Salome Horsch

Mündliche Prüfung:
Dortmund, 15. Juni 2020

Erstgutachter: Prof. Dr. Jörg Rahnenführer
Zweitgutachterin: Prof. Dr. Katja Ickstadt

Inhaltsverzeichnis

1 Einleitung

Die Untersuchung der Atemluft oder der Gerüche von Patientinnen und Patienten, um Krankheiten zu erkennen, reicht weit bis ins alte Griechenland zurück und wird seit einigen Jahrzehnten vermehrt erforscht (Cao und Duan, 2006). Es wurden zahlreiche Untersuchungen durchgeführt, um Krankheiten mit Hilfe von Atemluft zu detektieren. Dabei wurden unter anderem viele Lungenerkrankungen (beispielsweise Asthma, Lungenentzündungen oder Krebs), Stoffwechselerkrankungen (Diabetes), Erkrankungen des Magen-Darm-Traktes (Infektion mit dem Bakterium Helicobacter pylori), Lebererkrankungen (Leberzirrhose) und Nierenversagen untersucht (Di Francesco u. a., 2005; Cao und Duan, 2006; Kim u. a., 2012; Fink u. a., 2014).

Die Atemluft eines Menschen zu analysieren, hat verschiedene Vorteile gegenüber anderen Methoden, wie beispielsweise Untersuchungen des Blutes. Die Atemluft ist stets verfügbar und ihre Gewinnung ist sicher, da kein Eingriff in den Körper notwendig ist. Wird zur Analyse der Atemluft die Multikapillarsäulen-Ionenmobilitätsspektrometrie (MCC-IMS) verwendet, so ist die Messung innerhalb weniger Minuten abgeschlossen und könnte theoretisch direkt ausgewertet werden. Damit dies möglich wird, müssen die entstehenden Rohmessungen jedoch automatisch verarbeitet werden. Dies geschieht im Augenblick noch weitgehend manuell.

Um dieses Problem zu lösen, wurde im Sonderforschungsbereich 876 der TU Dortmund in Kooperation mit dem Dortmunder Unternehmen B&S Analytik ein Transferprojekt (TB1: Ressourcen-beschränkte Analyse von Spektrometriedaten) ins Leben gerufen. Die Firma entwickelt MCC-IMS-Geräte und unterstützte das Projekt insbesondere durch die Bereitstellung von Datenmaterial sowie die Durchführung und manuelle Auswertung neuer Messungen.

Zunächst wurden im Sonderforschungsbereich Methoden entwickelt, mit denen die Rohmessungen automatisch ausgewertet werden können. Diese Auswertung betrifft die Merkmalsextraktion aus den Rohmessungen, welche als Matrizen (ungefähr der Größenordnung 1500×2500) vorliegen. Im Gegensatz zu bestehenden Methoden wurde dabei in Betracht gezogen, dass die Auswertung auch auf sehr kleinen Geräten mit eingeschränkten Energieressourcen möglich sein soll. Die automatische Merkmalsextraktion für einen Datensatz aus mehreren Rohmessungen besteht aus zwei Schritten, der „Peakauswahl“ und dem „Peakclustern“. Zusammen

werden sie hier „Peakerkennung“ genannt. Für beide existieren verschiedene Ansätze. In dieser Arbeit werden verschiedene Algorithmen miteinander kombiniert und auf drei Datensätze angewendet. Jeder Datensatz enthält gesunde und kranke Personen, welche durch statistische Klassifikationsalgorithmen unterschieden werden sollen. Die Güte der Algorithmen zur Peakerkennung wird anhand der Klassifikationsgüte beurteilt. Es wird die Frage beantwortet, ob die automatischen Methoden hinsichtlich der Klassifikationsgüte genauso gut sind wie die aktuell manuelle Auswertung. Gleichzeitig werden verschiedene Klassifikationsalgorithmen angewandt, um auch die Frage zu beantworten, welche Algorithmen sich für die Klassifikation der MCC-IMS-Daten eignen. Anschließend wird eine Kombination von Algorithmen für den Gesamtanalyseprozess empfohlen, mit der zukünftig Klassifikationsprobleme nach einer automatischen Peakerkennung gelöst werden können.

Im zweiten Themenkomplex werden einige praktische Probleme der Atemluftanalyse mit MCC-IMS-Messungen betrachtet. Da für die MCC-IMS noch kontrollierte Studien fehlen, um Einflussfaktoren auf die Atemluft zu identifizieren, wurden Messungen an 49 Personen durchgeführt. In dieser Arbeit wird untersucht, ob das Geschlecht und der Raucherstatus einer Person einen Einfluss auf die Messungen haben. Darüber hinaus wurden die Messungen auf zwei Geräten durchgeführt, um herauszufinden, ob sich die Messungen auf verschiedenen Geräten unterscheiden. Um die dabei festgestellten Effekte zu reduzieren, werden zwei Korrekturmaßnahmen vorgeschlagen. Die erste betrifft die automatische Peakerkennung, bei der durch einen Zwischenschritt die Peakpositionen des einen Geräts an die Positionen des anderen Gerätes angeglichen werden. Die zweite Maßnahme entspricht einer nach Geräten getrennten Skalierung der Messwerte. Um künstlich ein Klassifikationsproblem zu schaffen, wurde jede Person zweifach gemessen, wobei die zweite Messung jeweils nach dem Konsum eines Glases Orangensaft durchgeführt wurde. Dieser Effekt kann ebenfalls als Einflussfaktor auf die Atemluft interpretiert werden. Anhand dieses Beispiels wird zusätzlich demonstriert, welchen Effekt das Ignorieren des Geräte-Effekts auf die Klassifikationsgüte haben kann. Insgesamt werden in dieser Arbeit also vier Störfaktoren, das Geschlecht und der Raucherstatus einer Person, das verwendete Gerät sowie Nahrungsmittel, betrachtet. Im Rahmen der zugehörigen Klassifikationsprobleme wird untersucht, ob die Veränderung des Wahrscheinlichkeits-Schwellenwerts von üblicherweise 0.5 verbessert werden kann, indem der Schwellenwert so gewählt wird, dass die Prävalenz im Trainingsdatensatz erhalten bleibt. Alle Analysen werden sowohl für die manuelle Auswertung der Rohmessungen als auch für die zuvor empfohlene automatische Algorithmenkombination durchgeführt.

In Kapitel 2 werden die Hintergründe der Arbeit erläutert. Dafür wird zunächst die Technologie der MCC-IMS vorgestellt. Anschließend werden die Problemstellungen für Algorithmen im Gesamtanalyseprozess und den Einfluss von Störgrößen dargelegt. Die verwendeten Datensätze werden in Kapitel 3, die angewandten Methoden zur Analyse der MCC-IMS-Daten in Kapitel 4 beschrieben. Dieses umfasst die Darstellung von Rohmessungen, Methoden zur Peakerkennung, der Alignierung von Peakpositionen bei zwei verschiedenen Geräten, Möglichkeiten der Skalierung von Messwerten bei zwei Geräten sowie Methoden, die im Zusammenhang mit der Klassifikation verwendet werden. In Kapitel 5 findet die Untersuchung der Algorithmen im Gesamtanalyseprozess von den Rohmessungen bis zur Klassifikation statt. Im Mittelpunkt steht dabei die Auswertung der drei Schritte Peakauswahl, Peakclustern und Klassifikation, welche sowohl gemeinsam als auch getrennt bewertet werden. Am Ende wird eine Kombination aus drei Algorithmen für das weitere Vorgehen empfohlen. Diese wird auch in Kapitel 6 verwendet, in dem die Analyse von Störgrößen für die MCC-IMS behandelt wird. Zunächst wird die Planung der durchgeführten Studie erläutert, bevor die gewonnenen Daten deskriptiv vorgestellt werden. Anschließend wird die Peakerkennung durchgeführt (manuell und automatisch) und die Eigenschaften der resultierenden Datensätze untersucht. Da hierbei große Unterschiede zwischen den Geräten festgestellt werden, wird für die automatische Peakerkennung anschließend ein Korrekturschritt angewandt („mit Alignierung“), sodass drei Arten der Peakerkennung vorliegen. Die Peakpositionen dieser drei Methoden werden verglichen. Um nach wie vor vorhandene Geräteunterschiede zu reduzieren, wird der Effekt mehrerer Skalierungen untersucht. Ob es Unterschiede zwischen den Beobachtungen beim Konsum von Saft, beim verwendeten Gerät, beim Geschlecht und beim Rauchen gibt, wird durch statistische Tests auf Mittelwertsunterschiede sowie durch Klassifikation überprüft. Anschließend wird demonstriert, welche Auswirkung die Vernachlässigung des Geräte-Effekts auf die Klassifikation haben kann. In Kapitel 7 werden die Ergebnisse der Arbeit eingeordnet und mögliche Erweiterungen vorgestellt. Kapitel 8 fasst die zentralen Ergebnisse der Arbeit zusammen.

2 Problemstellung und technische Grundlagen

Im folgenden Kapitel wird die Motivation dieser Arbeit erläutert. Da für das Verständnis der Problematiken grundlegende Kenntnisse der verwendeten technischen Gerätschaften notwendig sind, wird zunächst deren Funktionsweise in ihren Grundzügen vorgestellt. Anschließend werden die beiden inhaltlichen Themenkomplexe der Arbeit erörtert.

2.1 Multikapillarsäulen-Ionenmobilitätsspektrometrie (MCC-IMS)

Um die menschliche Atemluft zu analysieren, wird in dieser Arbeit die Technologie der Multikapillarsäulen-Ionenmobilitätsspektrometrie (MCC-IMS) betrachtet. Durch sie können flüchtige organische Verbindungen (VOCs, nach dem Englischen Begriff „volatile organic compounds“) in der Atemluft detektiert werden. Zur Anwendung kommen hier ausschließlich Daten, welche durch Geräte der Firma B&S Analytik erhoben wurden. Die Untersuchung der Atemluft erfolgt dabei in zwei aufeinanderfolgenden Schritten, welche namensgebend für die Technologie sind. Zunächst wird die Atemluft durch die Multikapillarsäule (MCC) geleitet, anschließend durch das Ionenmobilitätsspektrometer (IMS). Die folgende Beschreibung orientiert sich an der Darstellung von Baumbach (2009) und Kopczynski (2017).

Zunächst gelangt die Atemluft gemeinsam mit einem Trägergas (Stickstoff oder synthetische Luft) in die MCC. Die MCC ist eine Röhre, welche sich aus etwa 1000 kleineren Röhrchen zusammensetzt. Jede dieser einzelnen „Kapillaren“ hat einen Durchmesser von 40 µm, die gesamte MCC hat einen Durchmesser von etwa 3 mm. Dies erlaubt eine Trägergasgeschwindigkeit von bis zu 150 mL/s, welche optimal für das nachfolgende IMS ist. Die MCC wird beheizt und hat dabei häufig eine Temperatur von 30 oder 40 °C. Die Innenwände der Kapillaren sind mit einer sogenannten „stationären Phase“ ausgekleidet. Diese kann aus verschiedenen Materialien bestehen. Die hier verwendete ist eine OV-5-Phase (sie besteht zu 5% aus Biphenyl und zu 95% aus Dimethylpolysiloxan). Wird die Atemluft mit Hilfe des Trägergases durch die MCC geleitet, so bindet die stationäre Phase Moleküle aus der Atemluft zeitweise an sich. Wie lange die Moleküle an der stationären Phase haften, ist von ihrer chemischen Beschaffenheit

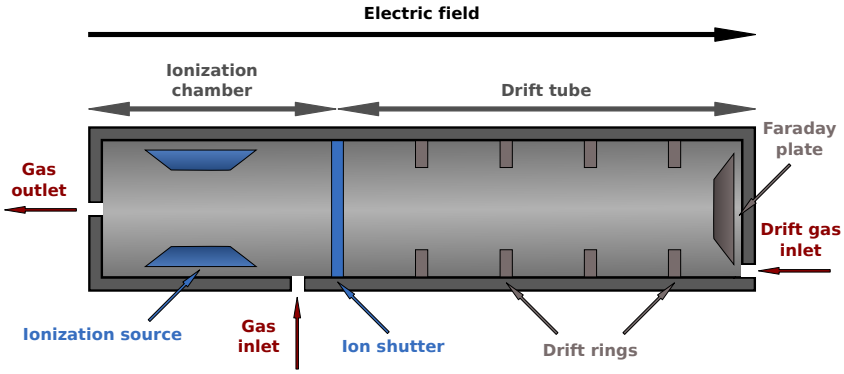


Abbildung 2.1: Schema eines Ionenmobilitätsspektrometers (Kopczynski, 2017). Mit freundlicher Genehmigung von Dr. Dominik Kopczynski.

abhängig. Während sich die Moleküle durch die MCC bewegen, wechseln sie immer wieder zwischen der mobilen und der stationären Phase. Während die gleichen Moleküle aus diesem Grund ungefähr die gleiche Zeit brauchen, um die MCC zu passieren, unterscheiden sich die Zeiten für Moleküle anderer Struktur. Auf diese Weise findet eine Vortrennung der Atemluft statt. Die Zeit, die ein Molekül zum Durchqueren der MCC benötigt, wird *Retentionszeit* genannt. Sie wird in Sekunden gemessen.

Nachdem die Moleküle die MCC durchquert haben, werden sie schubweise in das IMS (schematisch in Abbildung 2.1 dargestellt) geleitet. Gelangt ein Molekülschwarm in das IMS, so kommt es zunächst in der Ionisationskammer an. Dort werden die Trägergasmoleküle durch eine radioaktive Ionisationsquelle (hier das Nickel-Isotop ^{63}Ni) ionisiert. Diese Ionen reagieren mit den Molekülen der Atemluftprobe, sodass diese ebenfalls ionisiert werden. Anschließend öffnet sich ein Gitter, sodass einige Ionen in den nächsten Bereich, die Driftröhre, gelangen. Übrig gebliebene Moleküle werden aus der Ionisationskammer geleitet, wenn das Gitter wieder geschlossen ist. Durch ein elektrisches Feld in der Driftröhre (welches durch Driftringe entlang der Röhre stabilisiert wird) bewegen sich die Ionen vorwärts. Ein Driftgas (Stickstoff oder synthetische Luft) strömt ihnen entgegen. Durch die ständige Kollision der Ionen mit den Molekülen des Driftgases stellt sich für die Moleküle, erneut abhängig von ihren chemischen Eigenschaften, eine annähernd konstante Geschwindigkeit ein. Die Zeit, welche die Ionen benötigen, um die Driftröhre zu durchqueren, heißt „Driftzeit“. Da die Driftzeit von äußeren Einflüssen wie beispielsweise der Temperatur abhängig ist, wird diese nicht direkt verwendet sondern liegt als transformierte Größe vor. Diese sogenannte *Inverse reduzierte Mobilität* (1/K0, hier auch IRM) bezieht verschiedene Faktoren ein und kann als proportional zur Driftzeit angenommen werden. Sie wird in $\mathrm{V\,s/cm^2}$ angegeben. Am Ende der Driftröhre

treffen die Ionen auf eine Faraday-Platte, an welcher sie sich neutralisieren und gleichzeitig ein elektrisches Potenzial erzeugen. Dieses ist ein Indikator für die Menge der ankommenden Ionen. Das zeitabhängige Potenzial (jeweils für eine halbe Gitter-Öffnungszeit, also eine feste Retentionszeit) wird auch „Spektrum“ genannt. Das Potenzial für einen festen IRM-Wert in Abhängigkeit von der Retentionszeit, heißt auch „Chromatogramm“.

In dieser Arbeit kam das Gerät stets im positiven Modus zum Einsatz, was bedeutet, dass nur die positiv geladenen Ionen betrachtet werden. Die Ionen des Trägergases werden auch „Reaktand-Ionen“ genannt. Sie treten, da sie in jedem Gasgemisch enthalten sind, in jedem Spektrum auf und erzeugen dort den sogenannten „Reaktand-Ionen-Peak“ (RIP).

Das an der Faraday-Platte kontinuierlich gemessene Potenzial wird mit Hilfe eines Analog-Digital-Umsetzers in diskrete Werte (-2048 bis 2048) zu diskreten Zeitpunkten umgewandelt. Im positiven Modus ergeben sich hauptsächlich negative Werte. Damit, intuitiv plausibler, höhere Werte einer größeren Molekülmenge entsprechen, werden diese Werte hier mit -1 multipliziert. Diese werden dann im Folgenden als *Signalintensität* bezeichnet. Die vollständige Analyse einer Atemluftprobe benötigt etwa 10–12 Minuten.

Nachdem eine Messung mit dem MCC-IMS-Gerät abgeschlossen ist, liegt eine Reihe von Retentionszeiten, IRM-Werten und Signalintensitäten vor. Diese werden als Matrizen angeordnet, wobei für jede Kombination aus Retentionszeit und IRM-Wert genau eine Signalintensität vorliegt. Die Kombinationen von Retentionszeiten und zugehörigen IRM-Werten werden genutzt, um die unterschiedlichen Stoffe in der Atemluft voneinander unterscheidbar zu machen. Einer der beiden Werte allein würde keine ausreichend spezifische Unterscheidung erlauben. Die Signalintensitäten geben Informationen über die Menge der Analyte. Werden die Matrizen als Heatmap dargestellt (vergleiche Kapitel 4.1), bilden sich sichtbare Areale (Peaks), welche durch die einzelnen Komponenten entstehen und im Folgenden durch ihre Position charakterisiert werden. Die Matrix, welche zu einer einzelnen Atemluftprobe gehört, wird hier als *Rohmessung* bezeichnet. Jede Zeile der Rohmessung enthält ein Spektrum, jede Spalte ein Chromatogramm.

2.2 Algorithmen im Gesamtanalyseprozess

In dieser Arbeit steht das Erkennen von Hinweisen auf eine Krankheit mit Hilfe einer Atemluftuntersuchung im Vordergrund. Die Analyse der Atemluft hat gegenüber anderen Untersuchungsmethoden, wie beispielsweise einer Blutuntersuchung, den Vorteil, dass kein

Eingriff in den Körper notwendig ist, um das Analysematerial zu erhalten. Die ausgeatmete Atemluft steht als Abfallprodukt des Körpers permanent zur Verfügung. Eine solche Untersuchung kann somit ohne Folgen für die untersuchte Person durchgeführt und praktisch beliebig oft wiederholt werden. Auch eine durchgängige Überwachung, beispielsweise bei narkotisierten Personen, ist möglich. Da die MCC-IMS-Geräte fortlaufend weiterentwickelt und verkleinert werden, ist mit ihnen eine Untersuchung in einem gewissen Rahmen auch mobil denkbar. Da die Analyse einer Probe nur wenige Minuten in Anspruch nimmt, stehen die Ergebnisse schnell zur Verfügung.

Die Untersuchung der Atemluft mit der MCC-IMS-Technologie bietet Vor- und Nachteile. Im Vergleich zu anderen Methoden (wie beispielsweise der GC-MS, Gaschromatographie gekoppelt mit Massenspektrometrie) benötigt das MCC-IM-Spektrometer beispielsweise kein Hochvakuum, ist kleiner, leichter und somit auch mobiler (Baumbach, 2009). Auf der anderen Seite ist die Identifikation der in der Luftprobe enthaltenen Stoffe schwieriger. Die Peakpositionen lassen sich nicht ohne bestätigende Experimente konkreten chemischen Verbindungen zuordnen. Zudem erfordert die Detektion der Peaks bisher die (semi-)manuelle Auswertung eines Experten. Zu diesem Zweck wird aktuell die kommerzielle Software VisualNow (vergleiche Kapitel 4.2.1) verwendet. Dieses Vorgehen kann als Goldstandard angesehen werden.

Um die Atemluft mobil auswerten zu können, muss die Peakerkennung jedoch automatisch erfolgen. Die Analyse der Rohmessungen erfolgt dabei in zwei Schritten. Im ersten Schritt (Peakauswahl) wird jede Rohmessung einzeln betrachtet. Dabei ist es das Ziel, alle Positionen zu identifizieren, an denen sich das Zentrum eines Peaks befindet. Im zweiten Schritt (Peakclustern) werden die Peaks, welche in den einzelnen Messungen gefunden wurden, verglichen. Die Peakpositionen, welche dem gleichen Stoff in der Atemluft entspringen, befinden sich nicht immer exakt an der gleichen Position sondern weichen leicht voneinander ab. Aus diesem Grund werden über die verschiedenen Rohmessungen hinweg diejenigen Peaks zusammengefasst, welche wahrscheinlich durch die gleiche Komponente in der Atemluft verursacht werden. Eine Übersicht über in der Literatur vorgeschlagene Methoden zur Peakerkennung (auch über die hier verwendeten hinaus) ist in Kopczynski (2017) enthalten.

Das Erkennen von Krankheiten entspricht aus statistischer Sicht dem Lösen eines Klassifikationsproblems. Es gilt, zwei Klassen (gesunde und kranke Personen) mit Hilfe eines Klassifikationsalgorithmus möglichst gut zu unterscheiden. Anschließend können mit dem entsprechenden Modell Personen, bei denen die Klassenzugehörigkeit unbekannt ist, einer der beiden Klassen zugeordnet werden. Die Frage, welche Klassifikationsalgorithmen sich eignen, wurde bisher nur auf einzelnen Datensätzen überprüft (Hauschild, Baumbach u. a.,

2012; Hauschild u. a., 2013). Eine gleichzeitige Untersuchung von Peakerkennungsmethoden wurde ebenfalls in Hauschild u. a. (2013) durchgeführt. Diese wird hier erweitert, indem mehrere Datensätze, mehr Klassifikationsverfahren und neuere Peakerkennungsverfahren betrachtet werden, von denen einige eine Analyse der Rohmessungen erlauben, ohne diese komplett speichern zu müssen. Dies ist insbesondere für die Miniaturisierung der Geräte wünschenswert, da die Auswertung der Rohmessungen auf diese Weise weniger Speicher und weniger Energie benötigt.

Das erste Ziel dieser Arbeit ist herauszufinden, welche Algorithmen im Gesamtanalyseprozess eine gute Kombination darstellen. Da die wahren Inhaltsstoffe der Atemluft nicht bekannt sind, kann die unmittelbare Qualität der Peakerkennung nicht ohne eine manuelle Bewertung überprüft werden. Aus diesem Grund wird der Erfolg der Peakerkennung in dieser Arbeit daran gemessen, ob das nachfolgende Klassifikationsproblem gut gelöst werden kann. Als Bewertungskriterium wird hier also die Klassifikationsgüte verwendet. Dabei stellt sich die Frage, ob die automatischen Algorithmen eine erfolgreiche Klassifikation der vorliegenden Datensätze erlauben beziehungsweise ob sie dabei deutlich schlechter abschneiden als der (semi-)manuelle Goldstandard. Gleichzeitig kann die Wahl des Klassifikationsalgorithmus einen erheblichen Einfluss auf die Resultate haben. Aus diesem Grund wird eine Auswahl verschiedener Algorithmen für die drei Aufgaben (Peakauswahl, Peakclustern, Klassifikation) getroffen und kombiniert. Anschließend werden die Klassifikationsergebnisse verwendet, um geeignete Kombinationen zu identifizieren.

2.3 Einfluss von Störgrößen

Die Analyse der Atemluft ermöglicht nicht nur die Detektion von Krankheiten. Auch andere Informationen können in der Atemluft enthalten sein. Dies können Informationen über die Person sein, deren Atemluft analysiert wird oder sogar über den Ort, an dem sich die Person zuletzt aufgehalten hat, da die Atemluft neben den Metaboliten aus der Lunge auch Moleküle aus der Raumluft enthalten kann. Problematisch wird dies in der Klassifikation, wenn diese Informationen mit der Zielvariable in Zusammenhang stehen. Dies könnte der Fall sein, wenn gesunde und kranke Personen an unterschiedlichen Tagen oder Orten untersucht werden. Dann können sich rasch Unterschiede ergeben, beispielsweise wenn an den Standorten unterschiedliche Geräte verwendet werden oder in einer Klinik an den verschiedenen Tagen unterschiedliche Nahrungsmittel gereicht werden, bevor die Untersuchung stattfindet. Solche Einflussfaktoren werden hier als *Störgrößen* bezeichnet. Wird trotz solcher Störgrößen eine

Klassifikation durchgeführt, kann nicht mehr unterschieden werden, ob die Klassifikation anhand der Zielgröße (krank oder gesund) oder anhand der Störgröße (beispielsweise „an welchem Tag/Standort fand die Messung statt“) gelingt. Um derartige Probleme zu vermeiden, ist es unerlässlich zu wissen, welche Faktoren einen Einfluss auf die Atemluft haben.

Zu möglichen Faktoren, welche die Atemluft beeinflussen können, gibt es bereits verschiedene Studien, diese beziehen sich jedoch meist auf andere Technologien als die MCC-IMS. Eine große Studie (Blanchet u. a., 2017) unter Verwendung eines GC-TOF-MS (Gaschromatograph gekoppelt mit einem Flugzeitmassenspektrometer) untersuchte verschiedene Variablen. Insbesondere beim Raucherstatus, dem Alter, dem BMI (Body-Mass-Index) und dem Geschlecht fielen (bei Anwendung eines t-Tests) Unterschiede zwischen den 1417 untersuchten Personen auf. Die Klassifikation (mit einem Random Forest) gelang bei der Unterscheidung zwischen rauchenden und nichtrauchenden Personen (82% werden auf einem unabhängigen Datensatz korrekt klassifiziert). Keine Unterschiede fanden sich beim Cholesterinspiegel, dem Einsatz von Verhütungsmitteln und der Anzahl der weißen Blutkörperchen. Bei diversen Medikamenten konnten signifikante Unterschiede gefunden werden, jedoch weisen die Autoren darauf hin, dass diese auch auf die Krankheit zurückzuführen sein könnten, die der Medikation zu Grunde liegt.

Auch andere Studien zeigen Unterschiede für diese Variablen. So wurden bei der SPME-GC-MS (Festphasenmikroextraktions-GC-MS) signifikante Unterschiede für die Variablen Rauchen, Geschlecht, Alter (unter 40-Jährige im Vergleich mit älteren Personen) gefunden (Kischkel u. a., 2010). Für die PTR-MS (Proton-Transfer-Reaktions-MS) wurden ebenfalls Stoffe detektiert, die sich für Rauchende und Nichtrauchende unterscheiden (Jordan u. a., 1995; Kushch u. a., 2008). McWilliams u. a. (2015) führten Untersuchungen mit der Elektronischen Nase durch und fanden eine Komponente in der Atemluft, die sich für die Geschlechter signifikant unterschied. Zudem schlossen sie, dass das Klassifikationsergebnis für die Detektion von Lungenkrebs zu einem gewissen Grad vom Geschlecht und dem Raucherstatus abhängt. Auch Cheng u. a. (2009) verwendeten die Elektronische Nase und stellten Unterschiede zwischen rauchenden und nichtrauchenden Personen fest. Unterschiede zwischen den Geschlechtern können nicht nur generell vorhanden sein, offenbar verstoffwechseln Männer und Frauen einige Stoffe auch verschieden. So zeigten Ernstgård u. a. (2003), dass bei Frauen, die Propanol ausgesetzt wurden, nach 10 Minuten eine viermal höhere 2-Propanol-Konzentration in der Atemluft gemessen wurde als bei den Männern der Vergleichsgruppe.

Für die Technologie der MCC-IMS fehlen derartige Studien über die Effekte des Geschlechts und des Raucherstatus nach Wissen der Autorin bisher. Dass der Konsum bestimmter Nahrungsmittel (Süßigkeit, Orangensaft) die Atemluft beeinflusst, wurde von Vautz u. a. (2009) genutzt, um beispielhaft die Analyse und Auswertung von MCC-IMS-Messungen darzustellen. Dass es Unterschiede bei den Messungen auf verschiedenen Geräten geben kann, wurde ebenfalls bereits festgestellt (Cumeras u. a., 2012). Aus diesem Grund wird in dieser Arbeit eine Studie durchgeführt, bei der beispielhaft die Störfaktoren Geschlecht, Rauchen und das Gerät untersucht werden. Da viele Anwendungen der Atemluftanalyse auf die Klassifikation eines Krankheitsstatus zielen, werden dabei einige Messungen gezielt beeinflusst, um eine Krankheit, also eine abweichende Atemluftzusammensetzung, zu simulieren. Da bekannt ist, dass Nahrungsmittel einen Einfluss auf die Atemluft haben, wird der Konsum eines bestimmten Nahrungsmittels eingesetzt, um die Zusammensetzung der Atemluft kontrolliert zu verändern. Auch dieser Effekt kann gleichzeitig als Störfaktor betrachtet werden.

3 Datenmaterial

In diesem Kapitel werden die verwendeten Datensätze vorgestellt. Für die Bearbeitung der Problemstellung des Gesamtanalyseprozesses werden drei Datensätze verwendet, die aus den Datenbanken von B&S Analytik stammen. Für die Beantwortung der Frage nach Einflussfaktoren wurden neue Messungen durchgeführt.

3.1 Datensätze für Algorithmen im Gesamtanalyseprozess

In diesem Abschnitt werden die drei Datensätze vorgestellt, welche für die Untersuchung des Gesamtanalyseprozesses verwendet werden. In allen drei Fällen handelt es sich um diagnostische Fragestellungen in Form von Zwei-Klassen-Problemen, bei denen jeweils eine Gruppe erkrankter Personen sowie eine gesunde Kontrollgruppe enthalten sind. Alle Daten lagen bereits vor, es wurden in diesem Abschnitt keine Daten für diese Arbeit erhoben.

Datensatz 1 Der erste Datensatz enthält Daten einer noch laufenden klinischen Studie der Lungenklinik Hemer (Start: 2006, voraussichtliches Ende: 2023). Die vorliegenden Messungen stammen aus den Jahren 2006/2007. Die Studie ist auf www.clinicaltrials.gov in einer Datenbank der US-amerikanischen Nationalbibliothek für Medizin unter der Kennung NCT00632307 registriert. Sie wurde von der Ethikkommission der Universität Münster genehmigt. In dieser Arbeit werden lediglich gesunde Kontrollen und an COPD (chronisch obstruktive Lungenerkrankung) erkrankte Personen betrachtet. Die Studie selbst umfasst noch weitere Krankheiten. Insgesamt waren zum Zeitpunkt der Erstellung des Datensatzes für die interessierenden Gruppen die Atemluftmessungen von 127 Personen enthalten. Von ihnen waren 92 an COPD erkrankt und 35 gesunde Kontrollen. Erste Ergebnisse zu dieser Studie finden sich unter anderem in Westhoff u. a. (2010).

Datensatz 2 Der zweite Datensatz enthält die Daten einer Studie der zum Universitätsklinikum Essen gehörenden Ruhrlandklinik aus dem Jahr 2010. Die Studie wurde von der Ethikkommission des Universitätsklinikums Essen genehmigt. Es liegen Atemluftmessungen aus zwei Gruppen vor. Die Atemwege von 30 Teilnehmenden waren durch einen sogenannten „Krankenhauskeim“, das Bakterium Pseudomonas aeruginosa, infiziert. Als Kontrolle wurden die Daten von 37 nichtrauchenden Personen des Krankenhauspersonals erhoben. Univariate Unterschiede einzelner Peaks zwischen den beiden Gruppen wurden auf einem ähnlichen Datensatz bereits in Rabis u. a. (2011) untersucht.

Datensatz 3 Der dritte Datensatz wurde 2012 vom Knappschaftskrankenhaus Dortmund erhoben. Die Studie wurde von einer lokalen Ethikkommission genehmigt. Die Studie umfasst Daten von 39 Personen mit diagnostizierter Asbestose (eine durch Einatmung von asbesthaltigen Stäuben entstehende Krankheit) und 30 gesunde Kontrollen.

Datenqualität Die Qualität der vorliegenden Messdaten ist aufgrund geringer Hintergrundinformationen schwierig einzuschätzen. Die Daten wurden in geplanten Studien erhoben, welche von Ethik-Kommissionen genehmigt wurden. Dies erfordert die Einhaltung gewisser Standards, um die Qualität der Studien zu sichern. Über die Atemluftmessungen hinaus liegen hier jedoch keinerlei Zusatzinformationen, beispielsweise über die Probandinnen und Probanden, vor. Mögliche Einflüsse von Störgrößen sind hier dementsprechend nicht untersuchbar. Alle Datensätze wurden von der Autorin auf Auffälligkeiten untersucht, wobei selten Anomalien wie Mehrfachmessungen (teilweise auch mit leicht verschiedenen Messwerten) entdeckt wurden. Derartige Unklarheiten wurden durch Rückfragen bereinigt, lassen jedoch den Schluss zu, dass die Daten im Vorfeld noch nicht vollständig aufbereitet waren oder im Laufe der Zeit (die Atemluftmessungen wurden teilweise von verschiedenen Personen zu unterschiedlichen Zeitpunkten mit der Software VisualNow ausgewertet) verunreinigt worden sein könnten.

Da in diesem Kapitel die Untersuchung des Gesamtanalyseprozesses, also die optimale Auswahl von Auswertungsverfahren im Fokus steht, haben kleine Mängel in den Daten jedoch voraussichtlich keinen entscheidenden Einfluss auf die Ergebnisse. Für Aussagen über die Möglichkeiten, die konkret benannten Krankheiten an Personen erkennen zu können, wären detailliertere Informationen über die Datensätze notwendig.

3.2 Datensatz zur Analyse von Störgrößen

Für die Analyse von Störgrößen wurde im Mai 2017 eine neue Studie bei B&S Analytik in Dortmund durchgeführt. Insgesamt wurde die Atemluft von 49 Personen gemessen. Die Teilnehmenden waren hauptsächlich wissenschaftliche Mitarbeiterinnen und Mitarbeiter der Fakultät Statistik der TU Dortmund. Jede Person wurde zweimal gemessen, jeweils einmal vor und einmal nach dem Konsum eines Glases Orangensaft. Es standen zwei MCC-IMS-Geräte zur Verfügung, sodass die Messung mit Orangensaft für jede Person auf dem anderen Gerät gemessen wurde als die Messung ohne Orangensaft (das Startgerät wurde nach Geschlecht stratifiziert randomisiert). Zusätzlich wurden zu jeder Atemluftmessung eine Raumluft- und eine Spülmessung durchgeführt. Jede Person beantwortete in einem Fragebogen Fragen zu Alter und Geschlecht sowie zu Konsumverhalten von Nahrungsmitteln und Getränken vor der Studie sowie dem Rauchverhalten. Jede teilnehmende Person stimmte der Verwendung der erhobenen Daten schriftlich zu. Details zur Planung und Durchführung der Studie werden in Kapitel 6.1 erläutert.

Datenqualität Die Daten wurden kontrolliert und eigens zum Zweck der Beantwortung der dargelegten Problemstellung erhoben. Mögliche Einflussfaktoren wurden ausgewählt und in den Datensatz aufgenommen. Der Laborraum wurde allerdings zeitweise von anderen Angestellten des Labors für Arbeiten verwendet, sodass gelegentliche Kontaminationen der Raumluft nicht auszuschließen sind. Es gibt nur einen fehlenden Wert (eine Person gab ihr Alter beim ersten Rauchen nicht an), sonst wurden alle Fragebögen vollständig ausgefüllt. Eine Person ordnete sich selbst als aktiv rauchend ein, obwohl die im Fragebogen angegebenen Kriterien rechnerisch nicht erfüllt sein konnten. Die Selbsteinschätzung wurde beibehalten.

4 Methoden zur Analyse von MCC-IMS Daten

Dieses Kapitel beschreibt die Methoden, welche für die Analyse der MCC-IMS-Daten verwendet werden. Dies umfasst die Darstellung von Rohmessungen, Methoden zur Peakerkennung (automatisch und manuell), das vorgeschlagene Verfahren zur Alignierung der Peakpositionen bei zwei Geräten sowie zur Skalierung der Werte und die Methoden, die im Rahmen der Klassifikation verwendet werden.

4.1 Darstellung von MCC-IMS-Rohmessungen

Die Rohmessungen eines MCC-IMS-Geräts werden in der Regel als Heatmap dargestellt. Dabei wird die Matrix der Rohdaten, welche die Retentionszeiten in den Zeilen und die IRM-Werte in den Spalten enthält, als Bild interpretiert. Die Matrixeinträge werden hier daher auch als „Pixel“ bezeichnet. Für jede Kombination aus Retentionszeit und IRM-Wert wird das entsprechende Pixel gemäß des zugehörigen Intensitätswerts eingefärbt. Eine weiße Färbung entspricht einem Wert von Null. Mit steigender Intensität gehen die Pixelfarben abgestuft in blau, dann rot und zuletzt gelb über. Der Farbverlauf setzt sich hier aus insgesamt 101 möglichen Farbwerten zusammen.

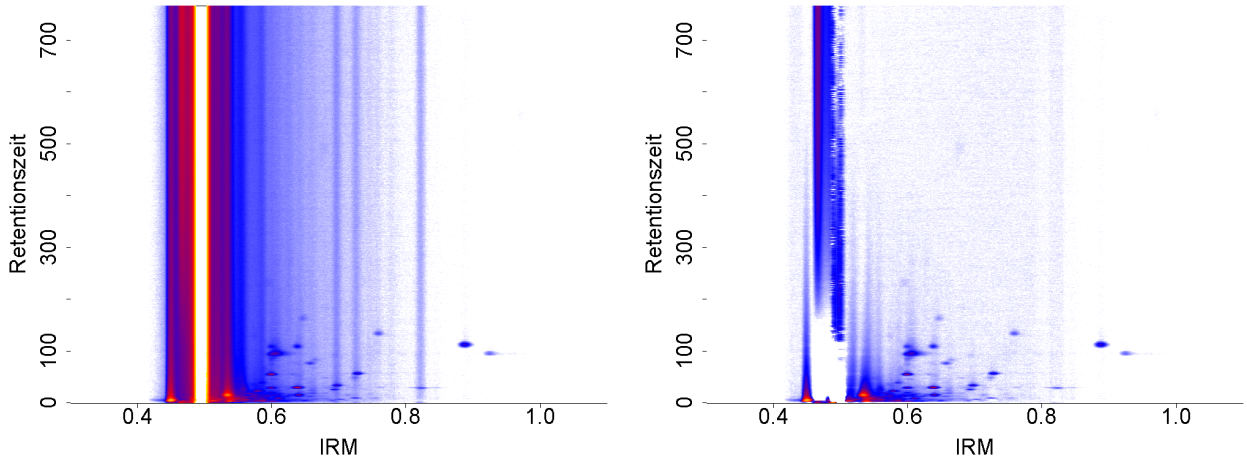
Um den RIP (vergleiche Kapitel 2.1) und gleichmäßiges Rauschen auf der Rohmessung grob zu entfernen, wird in dieser Arbeit vor der Darstellung ein einfacher Filter angewendet, welcher spaltenweise das 20%-Quantil der beobachteten Werte abzieht (Werte kleiner als Null werden auf Null gesetzt). Auf diese Weise werden uninformative Werte, die zu allen Retentionszeiten vorkommen, reduziert. Dies schließt insbesondere das Rauschen und den RIP mit ein.

Um anschließend optisch Peaks erkennen zu können, muss die verwendete Farbskala geeignet transformiert werden. Signalintensitäten mit Werten größer als 500 werden nicht mehr abgestuft sondern mit dem gleichen gelben Farbwert dargestellt. Die Intensitäten zwischen

Null und 500 werden logarithmisch dargestellt, sodass kleine Intensitäten gut unterschieden werden können. Für größere Signalintensitäten unterscheiden sich die Farben hingegen nicht mehr so stark.

In dieser Arbeit werden die Rohmessungen nicht einzeln sondern als mittlere Rohmessungen dargestellt. Diese zeigen komprimiert, wie die Rohmessungen aussehen. Variierende Peakpositionen sowie das Hintergrundrauschen werden dadurch geglättet und unterliegen somit weniger den natürlichen Schwankungen als die einzelnen Rohmessungen. Um die Rohmessungen mitteln zu können, ist es notwendig, dass sich die einzelnen Pixel auf die gleichen IRM- und Retentionszeitwerte beziehen. Da die Messungen jedoch nicht auf exakt identischen Intervallen durchgeführt werden, müssen die Rohmessungen zunächst vereinheitlicht (normiert) werden. Dafür werden zunächst die zweidimensionalen Intervalle für die resultierende mittlere Rohmessung festgelegt, anschließend werden die einzelnen Rohmessungen an diese Intervalle angepasst. Als Spannweite für die mittlere Rohmessung werden für beide Dimensionen die Extrema der Beobachtungen verwendet (gegebenenfalls werden Ausreißermessungen, bei denen in eine Dimension deutlich länger gemessen wurde, verkleinert). Die neuen Pixel-Intervalle werden äquidistant zwischen den äußeren Grenzen berechnet, sodass die Anzahl der Pixel etwa der der einzelnen Rohmessungen entspricht. Für die einzelnen Rohmessungen werden dann die Intensitäten für die neu definierten Pixel berechnet, indem für jede Messung die in dem Intervall liegenden Intensitäten gemittelt (arithmetisches Mittel) und anschließend wieder auf ganze Zahlen gerundet werden, um Speicherplatz zu sparen. Liegen in einem Intervall keine Intensitäten, wird ein fehlender Wert eingetragen. Um die mittlere Rohmessung zu erhalten, werden die derart auf gleiche Dimensionen normierten Rohmessungen anschließend je Pixel gemittelt (arithmetisches Mittel). Fehlende Werte werden dabei weggelassen.

Als Beispiel ist in Abbildung 4.1 die mittlere Rohmessung für alle Messungen eines der beiden Geräte für den Datensatz aus Kapitel 3.2 dargestellt. Links ist bei IRM≈ 0.5 der RIP zu sehen. Rechts ist spaltenweise das 20%-Quantil abgezogen worden, dadurch ist es dort weniger stark zu sehen. Das Abziehen höherer Quantile würde mehr Rauschen entfernen, wobei jedoch die Sichtbarkeit von Peaks mit niedrigen Signalintensitäten abnehmen könnte. Um dies zu vermeiden, wird hier auf stärkere Filter verzichtet und im Gegenzug ein etwas ausgeprägteres Hintergrundrauschen toleriert. Komplexere Filtermethoden werden in Hauschild, Schneider u. a. (2012) zusammengefasst, weitere in D'Addario u. a. (2014) vorgestellt.



(a) Ohne Filter

(b) Mit 20%-Quantil-Filter

Abbildung 4.1: Ein Beispiel für eine mittlere Rohmessung.

4.2 Peakerkennung

In diesem Kapitel werden die Methoden zur Erkennung von Peaks vorgestellt. Darunter befinden sich die automatischen Verfahren, welche sich in zwei Schritte unterteilen. Im ersten Schritt werden Peaks in einzelnen Rohmessungen annotiert (Peakauswahl). Im zweiten Schritt wird bestimmt, welche Peaks aus verschiedenen Messungen der gleichen Komponente in der Atemluft zugeordnet werden sollen (Peakclustern). Außerdem wird der Goldstandard der Peakerkennung, die (semi-)manuelle Methode mit der Software VisualNow, beschrieben. Diese unterscheidet die beiden Schritte nicht. Mit „Peakerkennung“ ist in dieser Arbeit stets der gesamte Prozess, welcher von den einzelnen Rohmessungen zur Peakliste gemeinsamer Peaks führt, bezeichnet.

4.2.1 Goldstandard: VisualNow (manuell)

Der manuelle Goldstandard der Peakerkennung für MCC-IMS-Messungen verwendet die kommerzielle Software VisualNow (Bödeker u. a., 2008). Da die Analyse Software-gestützt erfolgt, aber im Wesentlichen eine manuelle Annotation der Peaks erfordert, wird das Verfahren hier auch als (semi-)manuell bezeichnet. Die Software bietet unter anderem die Möglichkeit, die einzelnen Rohmessungen als Heatmap zu betrachten, Filter (beispielsweise zur Glättung) anzuwenden und manuell Peakpositionen zu markieren. Peaks werden durch eine Position sowie ein den Punkt umgebendes Rechteck charakterisiert. Der Rechteckausschnitt eines Peaks

kann für alle Rohmessungen gleichzeitig betrachtet werden. Peaks können beliebig hinzugefügt, entfernt und modifiziert werden. Zahlreiche Funktionen, wie beispielsweise die vergrößerte Darstellung eines Ausschnitts, erlauben eine detaillierte Analyse der Rohmessungen.

Zu den Ergebnissen der (semi-)manuellen Peakerkennung gehört der Datensatz, welcher für alle gefundenen Peaks die zugehörigen Intensitäten der Rohmessungen (höchster Wert im Rechteck) auflistet. Zudem wird ein sogenanntes „Layer“ erstellt. Dieses enthält die Peakpositionen (Zentrum des Rechtecks) sowie die Ausdehnung des zugehörigen Rechtecks (jeweils Abstand vom Zentrum für beide Dimensionen). Diese Layer können auch bei einer späteren Auswertung eines anderen Datensatzes zur Anwendung kommen.

4.2.2 Automatische Peakauswahl

In diesem Kapitel werden die Ideen der verwendeten Methoden zur automatischen *Peakauswahl* vorgestellt. Die Algorithmen detektieren Peaks auf *einzelnen* Rohmessungen und geben die Positionen der gefundenen Peaks sowie die Signalintensität des Peaks für die Rohmessung als Peakliste aus. Die ausgegebenen Peaks einer einzelnen Rohmessung heißen *Single Peaks*. Die Methoden können individuelle Vorverarbeitungsschritte (beispielsweise zur Glättung der Rohmessungen) beinhalten, welche hier nicht beschrieben werden. Ebenso werden in manchen Methoden auf individuelle Art Schwellenwerte bestimmt, die ein Peak überschreiten muss, um sich vom Grundrauschen der Rohmessungen abzusetzen. Dies kann zum Beispiel geschehen, indem ein Bereich der Rohmessung analysiert wird, in dem sich normalerweise keine Peaks befinden (z.B. vor dem RIP) und der dementsprechend das Rauschen beschreibt. Der Schwellenwert kann dann beispielsweise auf drei Mal die Standardabweichung des Rauschens gesetzt werden, nachdem der Mittelwert des Rauschens von allen Werten abgezogen wurde.

Local Maxima (LM) Die Methode *Local Maxima* (LM) stammt aus dem Peax-Framework (D’Addario u. a., 2014). Zunächst werden alle Matrixeinträge der betrachteten Rohmessung daraufhin überprüft, ob sie als „Peakkandidat“ in Frage kommen. Dafür muss die zugehörige Signalintensität einen bestimmten Schwellenwert überschreiten. Zusätzlich müssen alle acht umliegenden Einträge Signalintensitäten kleiner oder gleich dem potenziellen Peakkandiaten aufweisen, wobei der Schwellenwert weiterhin überschritten werden muss.

Abschließend werden Peakkandidaten, die zu dicht beieinander liegen, zu einem Peak zusammengefügt. Dafür wird das Cluster Editing Problem gelöst (die Cluster-Editing-Methode wird in Kapitel 4.2.3 vorgestellt). Dabei sind die Knoten die Peakkandidaten und jedes Cluster entspricht einem Single Peak.

Peak Model Estimation (PME) Die Methode *Peak Model Estimation* (PME) stammt ebenfalls aus dem Peax Framework (D'Addario u. a., 2014) und entspricht einer speziellen Kombination der dort aufgeführten Algorithmen. Eine detailliertere Beschreibung des Algorithmus findet sich in Kopczynski (2017).

Auch hier werden alle Matrixeinträge der Rohmessung zunächst daraufhin überprüft, ob sie als Peakkandidat eingestuft werden können. Das Verfahren („Cross-Finding") ist jedoch komplexer als bei Local Maxima. Jede Rohmessung wird zuerst zeilenweise (= spektrumweise) betrachtet und die erste Ableitung berechnet (entspricht jeweils der Differenz zweier aufeinanderfolgender Einträge). Die Position, bei der eine negative Steigung auf eine positive Steigung folgt (dies spricht für ein lokales Maximum dazwischen), wird in einer Liste von „aktiven" Positionen für das Spektrum gespeichert. Anschließend werden die aktiven Positionen über mehrere Spektren verbunden, wenn sie nahe beieinander liegen. Die aktive Position eines Spektrums kann mit aktiven Positionen des folgenden Spektrums zusammengefasst werden, wenn sich die Indizes der IRM-Positionen um nicht mehr als neun unterscheiden. Auf diese Art entstehen Spektrums-Peakkandidaten, die aus einer Liste von Positionen (RT und IRM) bestehen. Anschließend wird analog spaltenweise (=chromatogrammweise) vorgegangen, also wird ebenfalls die erste Ableitung gebildet und werden aktive Positionen für die Chromatogramme bestimmt, welche dann über die Chromatogramme hinweg zu Listen von Chromatogramm-Peakkandidaten zusammengefasst werden. Die finalen Peakkandidaten werden an der Stelle annotiert, an der sich Peakpositionen aus den Listen der Spektren-Peakkandidaten und Chromatogramm-Peakkandidaten überschneiden. Überlappen sich mehrere Positionen zweier Listen, so werden Position und Intensität des finalen Peakkandidaten von der Position übernommen, welche die größte Intensität aufweist. Es werden nur Peakkandidaten in den zweiten Schritt übergeben, deren Signalintensität einen bestimmten Schwellenwert überschreitet.

Um dicht beieinander liegende Peakkandidaten zu einem Single Peak zu verschmelzen, wird ein Expectation Maximization (EM) Algorithmus verwendet, der hier auch als „EM Clustering" bezeichnet wird (siehe Kapitel 4.2.3). Dabei sind die Single Peaks hier die Peakkandidaten und die Consensus Peaks die daraus entstehenden Single Peaks.

Peak Detection by Slope Analysis (PDSA) Die Methode *Peak Detection by Slope Analysis* (PDSA) (Egorov u. a., 2014) ist eine Variante, die Peaks in einer Rohmessung detektiert, ohne dass dabei die gesamte Rohmessung gespeichert werden muss („online Auswertung“). Dies erlaubt die (in Hinblick auf Speicherplatz und Zeit) ressourcenschonende Auswertung der Rohmessung im laufenden Betrieb. Dabei werden die Spektren nacheinander jeweils einzeln eingelesen. Für das einzelne Spektrum werden Peakkandidaten gesucht, welche anschließend gespeichert werden, bevor das komplette Spektrum gelöscht und das nächste eingelesen wird.

Die Peakerkennung im Spektrum erfolgt über ein gleitendes Fenster. Zunächst wird das eingelesene Spektrum geglättet, indem innerhalb des gleitenden Fensters (10 Indizes breit, wobei zuvor die Auflösung reduziert wurde, indem immer fünf aufeinanderfolgende Spektren zu einem Eintrag gemittelt wurden) die Summe der Signalintensitäten gebildet wird. Die Summe der Signalintensitäten wird mit der erwarteten Summe unter der Annahme, dass das Fenster nur Rauschen enthält, verglichen. Das Rauschniveau für einen Matrixeintrag wird hier als Mittelwert plus zwei Mal die Standardabweichung der Rohmessung angenommen. Der Mittelwert und die Standardabweichung werden online approximiert, indem geeignete Schätzer je Spektrum berechnet und am Schluss über die Spektren hinweg zusammengefasst werden. Die Summe der Signalwerte im Fenster muss also das Rauschniveau (multipliziert mit der Fensterbreite) überschreiten. Ein Peakkandidat im Spektrum wird als Folge von Indizes gebildet, welche bestimmte Eigenschaften erfüllen muss. Zu Beginn muss die im Fenster gebildete Summe der Signalintensitäten über die Indizes *steigen*. Ist die Summe erstmals niedriger als die vorige, werden der Mittelpunkt des vorigen Fensters als Position und der größte Wert im Fenster als Intensität des Maximums gespeichert. Zusätzlich wird die zugehörige Signalintensitäts-Summe im Fenster, für den späteren Vergleich über die Retentionszeiten hinweg, gespeichert. Sie kann als die Fläche des Peaks interpretiert werden. Ein Peakkandidat gilt nur dann als vollständig, wenn (mindestens) die erste niedrigere Summe noch das erwartete Rauschen im Fenster übersteigt. Andernfalls wird der Peakkandidat verworfen. Die folgenden Indizes mit *fallenden* Summen werden so lange dem Peak zugeordnet, bis die Summe erstmals wieder steigt oder das erwartete Rauschniveau unterschritten wird. Für jeden Peakkandidaten werden sein Start- und Endpunkt sowie die Position des Maximums mit dem lokalen Maximum sowie die Summe der Intensitäten des Fensters und die Retentionszeit des Spektrums gespeichert.

Um aus den Peakkandidaten des Spektrums finale Peaks zu bilden, müssen sie über die Retentionszeiten hinweg verbunden werden. Die Peakkandidaten werden dabei iterativ über die Retentionszeiten verschmolzen. Dafür müssen für die Positionen der beiden zu verschmelzenden Maxima zwei Bedingungen erfüllt sein. Die IRM-Werte dürfen nicht weiter als $0{,}003\,\mathrm{V\,s/cm^2}$ auseinanderliegen, die Retentionszeiten dürfen sich um nicht mehr als $0{,}1r + 3\,\mathrm{s}$ unterscheiden. Werden zwei Kandidaten verschmolzen, aktualisieren sich die zugehörige Position und ihre Intensität auf die Werte des größeren der beiden Maxima. Wird für einen Peakkandidat in der folgenden Retentionszeit kein passender Peakkandidat gefunden, gilt der Peak als abgeschlossen. Ein Peakkandidat, der nur in einer Retentionszeit vorliegt, also nicht verschmolzen wird, wird nicht als ausreichend für das Vorliegen eines Peaks angesehen und entfernt. Die Größe eines Peaks bezeichnet die Anzahl an Retentionszeiten, über die sich ein Peak beim Verschmelzen erstreckt. Üblicherweise steigen die gespeicherten Fenstersummen der Peakkandidaten über die Retentionszeiten erst an, erreichen das Maximum und fallen dann wieder ab. Es wird eine Toleranz für die Anzahl einzelner steigender Fenstersummen im abfallenden Teil des Peaks erlaubt, welche 10% der Peakgröße beträgt (ein Peak der Größe 23 darf also bis zu zwei ansteigende Fenstersummen im abfallenden Teil des Peaks aufweisen).

Savitzky-Golay Laplace-operator filter thresholding regions (SGLTR) Das Verfahren *Savitzky-Golay Laplace-operator filter thresholding regions* (SGLTR) wurde ebenfalls von Egorov u. a. (2014) vorgestellt. Die Idee basiert darauf, einen Peak durch die zweite Ableitung zu detektieren, also Positionen auf der Rohmessung zu finden, die eine hohe Krümmung aufweisen. Dies geschieht durch die Approximation des Laplace-Operators, welcher die Summe der partiellen zweiten Ableitungen in die beiden Dimensionen darstellt (ohne gemischte Ableitungen). Die zweiten Ableitungen beziehen sich auf eine zugrunde liegende polynomiale Regression vom Grad 3×2. Die Approximation des Laplace-Operators wird durch die Anwendung eines auf zwei Dimensionen erweiterten Savitzky-Golay-Filters der Dimension 11×11 durchgeführt. Der Filterkern kann vorberechnet werden, sodass die Approximation einfach durchführbar ist. Der Filterkern wird für jeden Matrixeintrag angewendet. Ist die Schätzung für die Krümmung an der Stelle hoch (wird anhand eines experimentell bestimmten Schwellenwerts bestimmt), so wird die Position als Peakkandidat in Betracht gezogen.

In der online-Variante des Verfahrens werden nur 11 aufeinanderfolgende Spektren zwischengespeichert, um den Filter anwenden zu können. Anschließend werden für die einzelnen Spektren nur die Positionen der Peakkandidaten gespeichert. Berührt ein aktueller Peakkandidat im folgenden Spektrum einen Peakkandidaten, so wird der neue Peakkandidat zur Region des

alten Peakkandidaten hinzugefügt. So werden angrenzende Peakpositionen zu Peakregionen zusammengefügt. Eine Peakregion muss mindestens zehn Pixel umfassen, alle übrigen werden entfernt. Jede Peakregion gilt als ein Peak, welcher durch die Position repräsentiert wird, welche den höchsten Intensitätswert aufweist. Diese Intensität wird auch als Peakintensität verwendet.

Online Peak Model Estimation (OPME) Die Methode *Online Peak Model Estimation* (OPME) (Kopczynski und Rahmann, 2014a) ist ebenfalls eine Variante, die Peaks in einer Rohmessung detektiert, ohne dass dabei die gesamte Rohmessung gespeichert werden muss.

Zunächst wird die Rohmessung zeilenweise (spektrumweise) verarbeitet und auf Peakkandidaten untersucht. Dazu wird die Zeile in einem gleitenden Fenster betrachtet. Innerhalb des Fensters wird ein Polynom zweiten Grades angepasst. Ist das geschätzte Extremum ein Maximum, liegt es innerhalb des Fensters und überschreitet die Signalintensität einen festgelegten Schwellenwert, so wird die Position zusammen mit der Intensität als Peakkandidat gespeichert. Wird im Fenster kein Peakkandidat gefunden, gleitet es nur einen Index weiter, bei Vorliegen eines Peakkandidaten gleitet das Fenster seine halbe Breite weiter, um den gleichen Kandidaten nicht mehrfach hintereinander zu detektieren. Die Form des Peaks wird zuvor durch eine inverse Gauß-Verteilung approximiert und vom betrachteten Spektrum abgezogen. Für jedes Spektrum resultiert eine Liste mit den geschätzten Parametern aller Peakkandidaten in diesem Spektrum.

Um die Peakkandidaten über die Spektren hinweg zu Peaks zusammenzufügen, werden im nächsten Schritt iterativ zwei aufeinanderfolgende Spektren betrachtet. Die Peaks $P_i, i = 1, ..., I$ des aktuellen Spektrums sollen mit den Peaks $P_j^+, j = 1, ..., J$ des folgenden Spektrums aligniert werden, das heißt, Peaks aus dem aktuellen Spektrum sollen entweder einem Peak aus dem nächsten Spektrum oder keinem Peak zugeordnet werden. Im letzteren Fall gilt der Peak als abgeschlossen. Um die optimale Zuordnung der Peaks des aktuellen Spektrums zu den Peaks des Folgespektrums zu finden, wird ein Algorithmus ähnlich dem Needleman-Wunsch-Algorithmus (Needleman und Wunsch, 1970) angewendet. Dabei wird zunächst für jede mögliche Peakkombination eine Bewertung der Passung vorgenommen. Für eine Zuordnung des Peaks P_j^+ zum Peak P_i wird die geschätzte inverse Gaußfunktion des Peaks P_i an der Position des Modus des Peaks P_j^+ ausgewertet (f_j^+). Diese Wahrscheinlichkeit wird in Relation zur Wahrscheinlichkeit des Erwartungswerts derselben inversen Gaußfunktion (f_i) gesetzt und der Logarithmus des Verhältnisses berechnet (log-Odds= $\log(\frac{f_j^+}{f_i})$). Eine schlechte Passung erzeugt negative, eine gute Passung positive Werte. Wird ein Peak nicht

einem anderen zugeordnet, wird der Wert Null vergeben. Anhand dieser Werte wird iterativ eine Bewertungsmatrix berechnet, die alle möglichen Zuordnungen berücksichtigt und aus welcher die optimale Alignierung über Backtracking (Zurückverfolgen des besten Pfades in der Matrix) bestimmt werden kann.

Alignierte Peakkandidaten entlang der Retentionszeit werden anschließend in gleitenden Fenstern ebenfalls durch quadratische Polynome angepasst. Ähnlich wie zuvor wird bei Vorliegen eines Peaks im Fenster eine inverse Gauß-Verteilung entlang der Retentionszeit modelliert. Die alignierten Peakkandidaten können auf diese Weise noch in mehrere Peaks aufgeteilt werden. Anschließend wird für jeden der resultierenden Peaks noch die Verteilung in IRM-Richtung bestimmt, sodass jedes Peakmodell durch eine zwei-dimensionale inverse Gaußverteilung dargestellt wird. Die Peakhöhe entspricht hier nicht einer an einer bestimmten Position gemessenen Intensität, sondern einem aus dem Modell geschätzten Wert.

Automatisches VisualNow (VN^a) Die in der Software VisualNow implementierte Peakauswahl basiert auf dem Algorithmus von Bader u. a. (2006). Die Signalintensitäten aller Pixel werden zunächst durch den Clusteralgorithmus k-Means (vergleiche auch das zugehörige Clusterverfahren in Kapitel 4.2.3) mit $k = 2$ in die beiden Gruppen „Peak“ und „kein Peak“ eingeteilt. Niedrige Werte sprechen dabei für „kein Peak“ wohingegen hohe Werte auf einen Peak hindeuten. Auch Regionen, welche keinen Peak aufweisen, können durch Ausreißer im Hintergrundrauschen vereinzelt hohe Werte annehmen. Aus diesem Grund werden alle Zuordnungen „Peak“, deren acht umgebenden Pixel nicht ebenfalls als Peak eingeschätzt werden, entfernt, also auf „kein Peak“ gesetzt. Dies kommt einer Glättung der Zuordnungen gleich. Anschließend werden die verbleibenden Positionen mit Peak-Zuordnung verbunden. Zunächst werden zeilenweise alle benachbarten Peak-Pixel verbunden. Anschließend werden alle entstandenen Ketten über die Spalten hinweg verbunden, wenn mindestens zwei Pixel benachbart sind. Wenn sich dabei Pixel von zwei verschiedenen Peak-Regionen benachbarn oder überlappen, so werden auch sie zusammengefasst. Die Position eines Peaks wird durch einen Zentroid repräsentiert. Dieser bildet sich aus den häufigsten Retentions- beziehungsweise IRM-Werten, welche von den Peak-Pixeln einer Region realisiert werden.

4.2.3 Automatisches Peakclustern

In diesem Kapitel werden die automatischen Algorithmen vorgestellt, welche für den Schritt des *Peakclusterns* eingesetzt werden können. Das Ziel dabei ist es, aus den Peaklisten der einzelnen Rohmessungen (welche aus den im vorigen Kapitel vorgestellten Methoden resultieren) eine neue Liste von Peaks zu generieren, welche die Komponenten im gemessenen Gasgemisch repräsentieren sollen. Da die Positionen der einzelnen Rohmessungen leicht variieren können, auch wenn sie von der gleichen Komponente im Gasgemisch erzeugt werden, werden ähnliche Positionen von Peaks verschiedener Rohmessungen durch Clusterverfahren zu sogenannten *Consensus Peaks* zusammengefügt.

Grid Squares (GS) Die Methode *Grid Squares* (GS) in einer Variante von Horsch u. a. (2015) stellt eine sehr einfache Form des Clusterns dar. Dabei werden die Rohmessungen mit Hilfe eines einfachen Gitters in rechteckige Segmente unterteilt, welche die möglichen Consensus Peaks darstellen. Die Breite der Rechtecke ist fix und beträgt $0{,}006\,\mathrm{V\,s/cm^2}$, die Höhe nimmt mit zunehmenden IRM-Werten linear zu. Die Position des Consensus Peaks wird als Mittelwert der Peakpositionen im Feld berechnet. Die Intensität des Consensus Peaks für eine Rohmessung wird auf das Maximum der Intensitäten der zugehörigen Single Peaks in dem Feld gesetzt.

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) Der Algorithmus DBSCAN (Ester u. a., 1996) ist ein bekannter Clusteralgorithmus, für den die Anzahl der Cluster nicht im Voraus festgelegt werden muss. Zunächst wird eine zufällige Beobachtung (Peakposition eines Single Peaks) aus dem Datensatz ausgewählt. Es wird geprüft, ob eine bestimmte Mindestanzahl an Beobachtungen (hier 10) in der Umgebung des Punktes liegt. Die Umgebung eines Punktes wird hier durch eine ihn umgebende Ellipse beschrieben, welche analog zu den Rechtecken bei Grid Squares eine feste Breite und eine vom IRM-Wert des Peaks linear abhängige Höhe aufweist. Wird die Anzahl der notwendigen Beobachtungen nicht erreicht, wird der Punkt verworfen und zufällig ein neuer ausgewählt. Andernfalls stellt der Punkt die erste Beobachtung eines neuen Clusters dar und wird als ein „Kernpunkt“ bezeichnet. Alle Beobachtungen in seiner Umgebung werden ebenfalls auf das Kriterium überprüft, ob sie Kernpunkte darstellen. Iterativ werden dem Cluster so Kernpunkte hinzugefügt, bis keine neuen Kernpunkte mehr erreicht werden. Abschließend werden für alle Punkte die in der Umgebung liegenden Punkte dem Cluster hinzugefügt, auch wenn diese selbst nicht genug

Nachbarn besitzen. Diese Punkte heißen „Dichte-erreichbare Punkte“ und liegen am Rand des Clusters. Ist ein Cluster abgeschlossen, wird zufällig ein noch nicht betrachteter Punkt gezogen. Dies geschieht so lange, bis jeder Punkt als Clustermitglied in Betracht gezogen wurde.

Cluster Editing (CE) Das Verfahren für gewichtetes Cluster Editing wurde von Rahmann u. a. (2007) vorgestellt und von D'Addario u. a. (2014) für das PEAX Framework auf die Problematik der Peakerkennung angewendet. Das Cluster Editing Problem betrachtet jeden Peak als Knoten in einem gewichteten, ungerichteten Graphen ohne Schleifen mit einer symmetrischen Gewichtsfunktion. Der Graph wird durch Hinzufügen oder Entfernen von Kanten verändert, wobei jeder Kante Kosten zugeordnet sind. Ziel ist es, den Graphen bei minimalen Gesamtkosten so zu modifizieren, dass disjunkte Cliquen (also Teilmengen von Knoten, die alle untereinander, aber mit keinem Knoten außerhalb der Clique verbunden sind) entstehen. Jede Clique repräsentiert einen Consensus Peak. Als Signalintensität des Consensus Peaks wird das Maximum der Signalintensitäten aller Peaks in der Clique gewählt.

Die dabei verwendete Gewichtsfunktion basiert auf einer Ähnlichkeitsfunktion und gewichtet Kanten zwischen ähnlichen Peaks mit positiven und Kanten zwischen unähnlichen Peaks mit negativen Werten. Liegen die Peaks genau auf der Grenze der maximal erwünschten Distanz in Retentionszeit und IRM, so ist das Gewicht Null. Die Ähnlichkeitsfunktion ist so gewählt, dass Peaks mit geringem Abstand als ähnlich gelten und weit entfernte als unähnlich. Dabei wird berücksichtigt, dass sich die IRM-Werte zweier Peaks, die zusammengeclustert werden können, um weniger als $0{,}003\,\mathrm{V\,s/cm^2}$ unterscheiden sollen. Für das Clustern eines Peaks an der Stelle (r, t) soll der potenzielle zweite Peak eine Retentionszeit aufweisen, die sich um weniger als $0{,}1r + 3\,\mathrm{s}$ von der ersten unterscheidet. Die Kosten für eine Änderung im Graphen entstehen durch das Hinzufügen von Kanten, die nicht ähnliche Peakkandidaten verbinden und das Entfernen von Kanten, die ähnliche Peaks verbanden.

EM Clustering (EM) Das Verfahren des EM Clusterings (Kopczynski und Rahmann, 2014b; Kopczynski, 2017) ist ebenfalls im Peax Framework (D'Addario u. a., 2014) enthalten und basiert auf dem EM Algorithmus (Dempster u. a., 1977). Dieser besteht aus zwei Schritten, dem E-Schritt („Expectation“-Schritt) und dem M-Schritt („Maximization“-Schritt), welche iterativ wiederholt werden. Für die Anwendung auf Peakpositionen wird zunächst davon ausgegangen, dass sich die Positionen der Consensus Peaks als Stichprobe einer gewichteten Summe von Wahrscheinlichkeitsverteilungen (zwei-dimensionale Normalverteilungen mit

diagonalen Kovarianzmatrizen) beschreiben lassen, wobei jede Wahrscheinlichkeitsverteilung einer im Gasgemisch enthaltenen Komponente zugeordnet ist. Es müssen sowohl die Gewichte als auch die Parameter der Wahrscheinlichkeitsverteilungen geschätzt werden. Zudem muss in jedem Schritt die in Wahrheit unbekannte Anzahl der Komponenten im Gasgemisch angepasst werden. Initial wird davon ausgegangen, dass jeder Peakkandidat einer Komponente im Gasgemisch entspricht.

Zunächst wird im E-Schritt für jeden Single Peak geschätzt, wie groß die Wahrscheinlichkeiten sind, zu den einzelnen Komponenten zu gehören und wie groß die jeweiligen Gewichte sind (gegeben die Parameterschätzungen). Wird die Peakposition eines Single Peaks von den Wahrscheinlichkeitsverteilungen zweier Komponenten gut erklärt (das heißt, die Wahrscheinlichkeit ist für beide Komponenten hoch), so verschieben sich die Lageparameter beider Verteilungen im Laufe der Iterationen aufeinander zu. Wird ein festgelegter Abstand zweier Komponenten unterschritten, so werden sie zu einer zusammengefügt, indem die Verteilung mit der niedrigeren Signalintensität entfernt wird und ihr Gewicht auf das Gewicht des anderen Peaks addiert wird. Im M-Schritt werden dann die Parameterschätzer mittels Maximum Likelihood aktualisiert, indem die geschätzten Zugehörigkeiten aus dem vorigen E-Schritt verwendet werden. Der E- und der M-Schritt werden iteriert, bis ein Konvergenzkriterium erreicht ist.

Automatisches VisualNow (VN^a) Das in der Software VisualNow implementierte Clusterverfahren wird in Bader u. a. (2006) vorgestellt. Die Basis des Algorithmus bildet das k-Means-Clusterverfahren. Dieses beginnt normalerweise mit k zufällig gewählten Beobachtungen als Clusterzentren. Jede Beobachtung wird anschließend dem Cluster zugeordnet, zu dessen Zentrum sie den kleinsten (euklidischen) Abstand aufweist. Anschließend wird das Clusterzentrum neu berechnet, indem die Beobachtungen (Peakpositionen) eines Clusters gemittelt werden. So wird fortgefahren, bis sich die Cluster nicht mehr verändern. Da die initialen Clusterzentren einen großen Einfluss auf die resultierenden Cluster haben können, werden sie in dieser Variante nicht zufällig gewählt, sondern durch einen vorgeschalteten Algorithmus bestimmt. Dafür wird zunächst ein agglomeratives hierarchisches Clusterverfahren („Ward-Methode“) verwendet, welches mit jedem Peak als Cluster beginnt und iterativ Cluster zusammenfügt. Dabei werden die Cluster vereinigt, deren Verschmelzung die geringste Erhöhung der Gesamtfehlerquadratsumme (Summe der quadratischen Abstände aller Beobachtungen zu ihrem zugehörigen Clusterzentrum) zur Folge hat. Die resultierenden Clusterzentren werden als Startwerte für die k-Means-Methode verwendet. Abschließend werden alle Peaks, die einem Cluster angehören, von einem Rechteck umschlossen, welches

die finale Peakregion repräsentiert. Die Intensitäten der einzelnen Rohmessungen eines Consensus Peaks ergeben sich daraus, dass alle Intensitätswerte im entsprechenden Rechteck der Rohmessung gemittelt werden.

4.3 Alignierung der Peakpositionen bei zwei Geräten

In der Praxis fällt auf, dass die Peaks gleicher Komponenten in der Atemluft bei unterschiedlichen Geräten auch etwas unterschiedliche Positionen aufweisen können. Dieser Effekt ist systematisch und geht über die gewöhnliche Variabilität der Peakpositionen hinaus. Um diesen Umstand zu berücksichtigen, wird hier vorgeschlagen, wie ein Referenzgemisch (welches bekannte Stoffe enthält), verwendet werden kann, um die Positionen zweier Geräte auf eine vergleichbare Skala zu transformieren. Auf die gleiche Art können zu einem späteren Zeitpunkt dann die Peakpositionen von Peaks, welche nicht aus dem Referenzgemisch stammen, transformiert werden.

Es sei ein Referenzgemisch, welches p Stoffe enthalte, auf zwei Geräten, A und B, gemessen worden. Seien die zugehörigen Peakpositionen $P_i^A = (r_i^A, t_i^A), i = 1, ..., p$ die Positionen auf Gerät A und analog die Positionen P_i^B für die Positionen auf Gerät B. Es wird ein Referenzgerät gewählt, dessen Peakpositionen nicht verändert werden. Sei hier ohne Beschränkung der Allgemeinheit das Referenzgerät Gerät A. Ziel ist es, die Peakpositionen von Gerät B so zu verschieben, dass sie sich möglichst wenig von denen von Gerät A unterscheiden (die mittlere quadratische Abweichung soll gering sein) unter der Annahme, dass in jede Dimension ein linearer Zusammenhang zwischen den Positionen bestehen soll. Dazu wird in beide Dimensionen der Peakpositionen unabhängig voneinander eine lineare Regression angewendet. Die resultierende Regressionsfunktion, um beispielsweise die Retentionszeiten von Gerät B zu verschieben, lautet entsprechend

$$r^A = a_r + b_r \cdot r^B.$$

Sie wird aus den Daten des Referenzgemischs berechnet. Anhand dieser Gleichung können anschließend auch die Retentionszeiten anderer Messungen von Gerät B auf die Skala von Gerät A transformiert werden. Analog wird für die IRM-Dimension vorgegangen.

Die lineare Gleichung bedeutet, dass sich die Positionen in den beiden Dimensionen um einen additiven sowie einen multiplikativen Faktor unterscheiden können. Dementsprechend sind die Entfernungen zwischen den Peaks der beiden Geräte nicht konstant sondern abhängig von den realisierten Werten.

4.4 Skalierungen bei zwei Geräten

Werden MCC-IMS-Messungen auf zwei verschiedenen Geräten durchgeführt, so kann es sein, dass die Messungen der beiden Geräte nicht auf der gleichen Skala liegen. Um zu prüfen, ob die Messungen beider Geräte auf die gleiche Skala transformiert werden können, werden verschiedene Methoden angewendet.

Im Folgenden bezeichnen $x_1, ..., x_{n_A}$ die Intensitäts-Werte eines von Gerät A gemessenen Metaboliten und $y_1, ..., y_{n_B}$ die Werte desselben Metaboliten, gemessen von Gerät B. Außerdem bezeichnen $\bar{x} = \frac{1}{n_A} \sum_{i=1}^{n_A} x_i$ und $\bar{y} = \frac{1}{n_B} \sum_{i=1}^{n_B} y_i$ die arithmetischen Mittel der beiden Geräte, $\hat{\sigma}_x = \sqrt{\frac{1}{n_A - 1} \sum_{i=1}^{n_A} (x_i - \bar{x})^2}$ und $\hat{\sigma}_y = \sqrt{\frac{1}{n_B - 1} \sum_{i=1}^{n_B} (y_i - \bar{y})^2}$ die Standardabweichungen für beide Geräte. Darüber hinaus seien $\bar{x}^{(\neq 0)}$, $\bar{y}^{(\neq 0)}$, $\hat{\sigma}_x^{(\neq 0)}$, $\hat{\sigma}_y^{(\neq 0)}$ die gleichen Größen, bei denen zuvor alle Werte, die Null sind, aus dem Datensatz entfernt wurden.

Mittelwert-Skalierung

Bei der Mittelwert-Skalierung werden beide Geräte unabhängig voneinander auf den gleichen Mittelwert (Null) skaliert. Diese einfache Skalierung eignet sich, wenn nur die Mittelwerte der beiden Geräte verschoben sind. Die transformierten Werte für die beiden Geräte berechnen sich zu

$$x_i^{(\mathrm{m0})} = x_i - \bar{x}, \quad y_i^{(\mathrm{m0})} = y_i - \bar{y}.$$

Varianz-Skalierung

Bei der Varianz-Skalierung werden die Varianzen der beiden Geräte unabhängig voneinander auf den gleichen Wert (Eins) skaliert. Diese einfache Skalierung eignet sich, wenn die Mittelwerte der beiden Geräte gleich sind, die Streuung eines Gerätes aber stärker ist als des anderen. Die transformierten Werte für die beiden Geräte berechnen sich zu

$$x_i^{(\mathrm{v1})} = \frac{x_i}{\hat{\sigma}_x}, \quad y_i^{(\mathrm{v1})} = \frac{y_i}{\hat{\sigma}_y}.$$

Mittelwert-Varianz-Skalierung

Die Mittelwert-Varianz-Skalierung kombiniert die beiden vorigen Skalierungen und skaliert beide Geräte unabhängig voneinander auf den gleichen Mittelwert (Null) und die gleiche Varianz (Eins). Diese Skalierung ist ein Standardverfahren, das auch „Normalisierung“ oder „Standardisierung“ genannt wird. In Kapitel 6 wird diese Methode auch mit *Standardisierung* bezeichnet. Die normalisierten Werte für die beiden Geräte berechnen sich zu

$$x_i^{(\text{m0v1})} = \frac{x_i - \bar{x}}{\hat{\sigma}_x} \quad , \quad y_i^{(\text{m0v1})} = \frac{y_i - \bar{y}}{\hat{\sigma}_y}.$$

Erweiterte Mittelwert-Varianz-Skalierung

In der praktischen Anwendung der automatisierten Peakerkennung können die beobachteten Intensitäten nicht mehr als stetige Variable angenommen werden, da häufig der Wert Null gemessen wird. In diesen Fällen lassen sich die vorigen Skalierungen nicht mehr plausibel interpretieren (vergleiche dazu Kapitel 6.4). Um diese Problematik zu adressieren, wird die Standardisierung hier erweitert. Die Werte beider Geräte, die nicht Null sind, werden wie zuvor unabhängig voneinander auf Mittelwert Null und Varianz Eins skaliert, indem alle Null-Werte im Datensatz vernachlässigt werden. Die Null-Werte beider Geräte werden hingegen *gemeinsam* transformiert, da sie nach der Skalierung die gleiche Bedeutung haben sollen (auch zuvor hatten sie die gleiche Interpretation, dass der Metabolit nicht in der Luft gemessen wurde). Um sicherzustellen, dass kein Wert, der zuvor Null war (und damit kleiner als alle anderen Werte ungleich Null) nach der Skalierung einen höheren Wert als andere Messungen haben kann, wird von den Null-Werten beider Geräte der größere der beiden Mittelwerte $\bar{x}^{(\neq 0)}$ und $\bar{y}^{(\neq 0)}$ abgezogen und durch die kleinere Standardabweichung von $\hat{\sigma}_x^{(\neq 0)}$ und $\hat{\sigma}_y^{(\neq 0)}$ geteilt. In Kapitel 6 wird diese Methode auch mit *Standardisierung*$^+$ bezeichnet. Die transformierten Werte für die beiden Geräte berechnen sich entsprechend zu

$$x_i^{(\text{m0v1}^+)} = \begin{cases} \frac{x_i - \bar{x}^{\neq 0}}{\hat{\sigma}_x^{\neq 0}}, & \text{für } x_i \neq 0,\ \hat{\sigma}_x^{\neq 0} \neq 0 \\ x_i - \bar{x}^{\neq 0}, & \text{für } x_i \neq 0,\ \hat{\sigma}_x^{\neq 0} = 0 \\ \frac{-\max\{\bar{x}^{\neq 0}, \bar{y}^{\neq 0}\}}{\min\{\hat{\sigma}_x^{\neq 0}, \hat{\sigma}_y^{\neq 0}\}}, & \text{für } x_i = 0,\ \hat{\sigma}_x^{\neq 0} \neq 0 \text{ and } \hat{\sigma}_y^{\neq 0} \neq 0 \\ \frac{-\max\{\bar{x}^{\neq 0}, \bar{y}^{\neq 0}\}}{\max\{\hat{\sigma}_x^{\neq 0}, \hat{\sigma}_y^{\neq 0}\}}, & \text{für } x_i = 0,\ \text{entweder } \hat{\sigma}_x^{\neq 0} \neq 0 \text{ oder } \hat{\sigma}_y^{\neq 0} \neq 0 \\ -\max\{\bar{x}^{\neq 0}, \bar{y}^{\neq 0}\}, & \text{für } x_i = 0,\ \hat{\sigma}_x^{\neq 0} = \hat{\sigma}_y^{\neq 0} = 0 \end{cases},$$

$y_i^{(\text{m0v1}^+)}$ ist analog definiert.

Erweiterte Median-MAD-Skalierung

Eine Variante der vorigen Skalierung, die robust gegenüber Ausreißern ist, ist die erweiterte Median-MAD-Skalierung. Die transformierten Beobachtungen werden mit $x_i^{(\mathrm{m0m1^+})}$ und $y_i^{(\mathrm{m0m1^+})}$ bezeichnet. Sie berechnen sich analog zur erweiterten Mittelwert-Varianz-Skalierung, wobei arithmetische Mittelwerte durch den Median und Standardabweichungen durch die mediane absolute Abweichung (MAD) ersetzt werden.

4.5 Klassifikation

Im folgenden Kapitel werden die verwendeten Klassifikationsalgorithmen beschrieben. Ein Klassifikationsalgorithmus stellt für einen Datensatz mit kategoriellem Zielkriterium (dieses gibt die Klassenzugehörigkeit an), eine Entscheidungsregel auf, mit welcher die Beobachtungen möglichst gut ihrer wahren Klasse zugeordnet werden. Mit dieser Entscheidungsregel können anschließend neue Beobachtungen, für welche der Wert der Zielvariable nicht bekannt ist, einer Klasse zugeordnet werden. Da in dieser Arbeit nur Zwei-Klassen-Probleme behandelt werden, beschränkt sich die Darstellung der Methoden auf den Fall von nur zwei Klassen. Die Algorithmen umfassen die Support Vector Machine (linear und mit RBF-Kern), das k-Nächste-Nachbarn-Verfahren, den Klassifikationsbaum, den Random Forest und das Gradienten Basierte Boosting.

4.5.1 Support Vector Machine (SVM)

Die Support Vector Machine (SVM) ist ein populärer Klassifikationsalgorithmus, welcher hier nach Hastie u. a. (2009, S. 417–424) beschrieben wird. Die Klassifikation wurde in R mit dem Paket `e1071` (Meyer u. a., 2014) durchgeführt.

Es seien n Beobachtungen $\boldsymbol{x_1}, ..., \boldsymbol{x_n} \in \mathbb{R}^k$ gegeben. Zusätzlich sei für jede Beobachtung eine Klassenzugehörigkeit $y_i \in \{-1, 1\}$ gegeben.

Lineare SVM Im Fall der linearen SVM ist es das Ziel, den Stichprobenraum durch eine Hyperebene zu trennen, sodass die Beobachtungen auf je einer Seite der Hyperebene einer Klasse zugeordnet werden. Die Hyperebene wird beschrieben durch

$$\{\boldsymbol{x} : f(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta} + \beta_0 = 0\} \quad , \quad \boldsymbol{\beta} \in \mathbb{R}^k, \|\boldsymbol{\beta}\| = 1.$$

Die zugehörige Entscheidungsregel lautet dann:

$$F(\boldsymbol{x}) = \text{sign}[\boldsymbol{x}^T\boldsymbol{\beta} + \beta_0].$$

Die Hyperebene soll so bestimmt werden, dass es einen möglichst großen Rand („Margin“) mit Abstand C gibt, in dem keine oder nur wenige Beobachtungen liegen. Dies wird erreicht, indem Beobachtungen innerhalb des Margins bestraft werden. Der Margin soll also maximiert werden, wobei die Beobachtungen mindestens um C (auf der richtigen Seite) von der Hyperebene entfernt sein sollen. Allerdings erlauben die sogenannten „Schlupfvariablen“ ξ_i, dass die Beobachtungen um einen gewissen Anteil $1 - \xi_i$ innerhalb des Margins oder sogar auf der falschen Seite der Hyperebene (für $\xi_i > 1$) liegen dürfen. Die Summe der ξ_i wird dabei beschränkt. Der (vorzeichenbehaftete) Abstand der Beobachtung $\boldsymbol{x_i}$ zur Hyperebene ist ${\boldsymbol{x_i}}^T\boldsymbol{\beta} + \beta_0$, sodass sich das Optimierungsproblem folgendermaßen formulieren lässt:

$$\max_{\boldsymbol{\beta},\beta_0,\|\boldsymbol{\beta}\|=1} C \qquad \text{sodass} \quad y_i({\boldsymbol{x_i}}^T\boldsymbol{\beta} + \beta_0) \geqslant C(1 - \xi_i) \ \forall i, \quad \xi_i \geqslant 0, \quad \sum_{i=1}^{n} \xi_i \leqslant \text{const}$$

$$\Leftrightarrow$$

$$\min_{\boldsymbol{\beta},\beta_0} \frac{1}{2}\|\boldsymbol{\beta}\|^2 + \gamma \sum_{i=1}^{n} \xi_i \qquad \text{sodass} \quad y_i({\boldsymbol{x_i}}^T\boldsymbol{\beta} + \beta_0) \geqslant (1 - \xi_i) \ \forall i, \qquad \xi_i \geqslant 0.$$

In der zweiten Zeile ist eine alternative Notation des Optimierungsproblems dargestellt, bei der die Summe der ξ_i nicht länger festgelegt ist sondern mit γ als Bestrafungsterm eingeht. Werden die Nebenbedingungen integriert, so ergibt sich die Lagrange-Funktion zu

$$L_P = \frac{1}{2}\|\boldsymbol{\beta}\|^2 + \gamma \sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i(y_i({\boldsymbol{x_i}}^T\boldsymbol{\beta} + \beta_0) - (1 - \xi_i)) - \sum_{i=1}^{n} \mu_i\xi_i.$$

Anschließend kann aus dieser primalen Lagrangefunktion die duale Lagrangefunktion bestimmt werden, indem die partiellen Ableitungen bezüglich $\boldsymbol{\beta}$, β_0 und ξ_i berechnet und gleich Null gesetzt werden und die resultierenden Bedingungen (4.1-4.3) in die primale Lagrangefunktion

eingesetzt werden. Die duale Lagrangefunktion lautet dann

$$L_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \boldsymbol{x_i}^T \boldsymbol{x_j}.$$

Sie wird bezüglich der α_i optimiert, wobei $0 \leqslant \alpha_i \leqslant \gamma$ sowie $\sum_{i=1}^{n} \alpha_i y_i = 0$ gelten müssen. Mit den eben berechneten Bedingungen und den zusätzlich geltenden Karush-Kuhn-Tucker-Bedingungen (4.4-4.6) wird die Lösung eindeutig festgelegt und ist für das primale und duale Lagrange-Problem identisch. Das duale Problem ist ein einfacheres konvexes Optimierungsproblem als das primale und wird numerisch gelöst.

Die notwendigen Nebenbedingungen lauten:

$$\sum_{i=1}^{n} \alpha_i y_i = 0, \tag{4.1}$$

$$\boldsymbol{\beta} = \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x_i}, \tag{4.2}$$

$$\alpha_i = \gamma - \mu_i \quad \forall i, \tag{4.3}$$

$$\alpha_i (y_i (\boldsymbol{x_i^t} \boldsymbol{\beta} + \beta_0) - (1 - \xi_i)) = 0, \tag{4.4}$$

$$\mu_i \xi_i = 0, \tag{4.5}$$

$$y_i (\boldsymbol{x_i^T} \boldsymbol{\beta} + \beta_0) - (1 - \xi_i) \geqslant 0. \tag{4.6}$$

Liegen nach der Optimierung die Schätzer $\hat{\alpha}_i$ vor, so kann der Vektor $\hat{\boldsymbol{\beta}}$ entsprechend der Nebenbedingung 4.2 berechnet werden als

$$\hat{\boldsymbol{\beta}} = \sum_{i=1}^{n} \hat{\alpha}_i y_i \boldsymbol{x_i}.$$

Dabei ist zu beachten, dass $\hat{\alpha}_i$ nur dann ungleich Null ist, wenn $y_i(\boldsymbol{x_i}^T \boldsymbol{\beta} + \beta_0) = (1 - \xi_i)$ (resultiert aus 4.4). Diese Punkte, die entweder auf dem Margin liegen ($\hat{\xi}_i = 0$) oder innerhalb des Margins oder auf der falschen Seite liegen ($\hat{\xi}_i > 0$), heißen darum auch „Stützvektoren“ oder „Support Vectors“, weil die Schätzung der Hyperebene nur auf ihnen basiert. Die Konstante β_0 kann für jede Beobachtung geschätzt werden, die genau auf dem Margin liegt, indem die Nebenbedingung 4.4 nach β_0 umgestellt wird. Für eine stabile Lösung werden alle möglichen Schätzer für β_0 gemittelt.

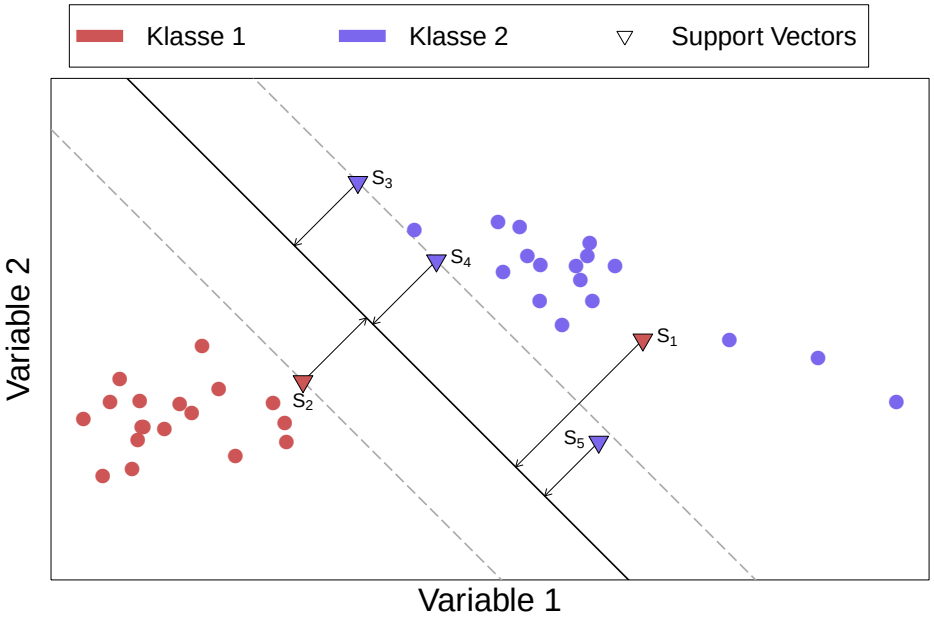


Abbildung 4.2: Beispiel für die lineare SVM im zweidimensionalen Raum. Die durchgezogene Linie entspricht der Geraden der Entscheidungsregel (Punkte unterhalb der Geraden werden Klasse 1 zugeordnet, Punkte oberhalb der Klasse 2). Die gestrichelten Linien beschreiben den Margin. Die Gerade wird nur durch die Support Vectors S_1-S_5 festgelegt.

Ein Beispiel für die lineare SVM ist in Abbildung 4.2 dargestellt. Die durchgezogene Linie entspricht $f(x)$, die gestrichelten Linien markieren den Margin. Die als Dreiecke dargestellten Beobachtungen sind die Stützvektoren (für sie alle gilt $\hat{\alpha}_i > 0$). Die Punkte S_2, S_3 und S_4 liegen jeweils auf dem Margin und werden korrekt klassifiziert, daher gilt für sie $\xi_i = 0$. Der Punkt S_1 liegt auf der falschen Seite außerhalb des Margins und wird dementsprechend falsch klassifiziert. Für ihn gilt $\xi_i > 1$. Der Punkt S_5 liegt auf der richtigen Seite innerhalb des Margins (dies entspricht $0 < \xi_i \leqslant 1$) und wird dementsprechend korrekt klassifiziert.

Eine neue Beobachtung $\boldsymbol{x_i}$ wird anhand der folgenden Regel klassifiziert:

$$F(\boldsymbol{x}^*) = \text{sign}[\boldsymbol{x}^{*T}\hat{\boldsymbol{\beta}} + \hat{\beta}_0].$$

Um stattdessen Wahrscheinlichkeiten zu erhalten, wird in R eine logistische Regression auf die Abstände zur Hyperebene gerechnet.

RBF-SVM In der linearen Variante der SVM steht nur eine Hyperebene zur Verfügung, um die Daten in zwei Klassen zu teilen. Um eine nicht-lineare Trennung zu bewirken, können die Beobachtungen mit einer Funktion $h(\boldsymbol{x_i}) = (h_1(\boldsymbol{x_i}), h_2(\boldsymbol{x_i}), ..., h_M(\boldsymbol{x_i}))$ zunächst in eine höhere Dimension transformiert werden, in der eine Trennung möglicherweise besser erfolgt. Wird die duale Lagrangefunktion der linearen Variante für die transformierten Beobachtungen betrachtet, ergibt sich folgende Funktion:

$$L_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \langle h(\boldsymbol{x_i}), h(\boldsymbol{x_j}) \rangle.$$

Die Lösungsfunktion $f(x)$ lässt sich mit Nebenbedingung 4.2 umformen zu:

$$\begin{aligned} f(\boldsymbol{x}) &= h(\boldsymbol{x})^T \boldsymbol{\beta} + \beta_0 \\ &= \sum_{i=1}^{n} \alpha_i y_i \langle h(\boldsymbol{x_i}), h(\boldsymbol{x_j}) \rangle + \beta_0. \end{aligned}$$

Beide Gleichungen hängen nur vom Skalarprodukt der Funktion $h(\boldsymbol{x})$ ab, nicht aber von der Funktion selbst. Daher werden die Transformationen der Beobachtungen nicht berechnet, sondern ausschließlich ihr Skalarprodukt, welches dann auch als „Kern“ oder „Kernel“ bezeichnet wird. Dieses Vorgehen wird deshalb auch als „Kernel-Trick“ bezeichnet.

Der hier verwendete Kern ist der „Radial Basis Function“-Kern (RBF-Kern):

$$\begin{aligned} K(\boldsymbol{x_i}, \boldsymbol{x_j}) &= \langle h(\boldsymbol{x_i}), h(\boldsymbol{x_j}) \rangle \\ &= \exp(-a \left\| x_i - x_j \right\|^2). \end{aligned}$$

Die dazugehörige Transformations-Funktion $h(\boldsymbol{x})$ stellt eine Transformation in eine unendliche Dimension dar, wird aber selbst nicht berechnet. Die Konstante a (in der Literatur häufig mit γ bezeichnet) wird als zu optimierender Parameter behandelt.

4.5.2 k-Nächste-Nachbarn (kNN))

Das Verfahren k-Nächste-Nachbarn (kNN) ist ein sehr einfacher Ansatz zur Klassifikation, welcher zunächst durch E. Fix und J.L. Hodges in in einem unveröffentlichten Manuskript (dieses wurde von Silverman und Jones (1989) in einer kommentierten Version veröffentlicht) erwähnt wurde. In R wurde die Implementierung aus dem Paket `kknn` (Schliep und Hechenbichler, 2014) verwendet.

Um eine Beobachtung zu klassifizieren, werden ihre k Nachbarn betrachtet. Die Beobachtung wird dann der Klasse zugeordnet, der die meisten Nachbarn angehören (Mehrheitsentscheid). Sollte es dabei zu einem Gleichstand kommen, wird zufällig eine der Klassen ausgewählt. Sollten mehrere Beobachtungen gleich weit entfernt sein wie die k-te, so werden diese ebenfalls in den Mehrheitsentscheid miteinbezogen. Die Wahl des Parameters k hat einen großen Einfluss auf den Erfolg der Klassifikation. Wird k zu klein gewählt, kann durch leichte Variation in den Daten eine andere Klassifikationsentscheidung getroffen werden, wird k hingegen zu groß gewählt, kann die Klasse durch zu weit entfernte Beobachtungen beeinflusst werden. Als Distanzmaß für die Bestimmung der nächsten Nachbarn wird hier der Euklidische Abstand verwendet.

4.5.3 Klassifikationsbaum (CT)

Der Klassifikationsbaum (CT, nach dem englischen Begriff „Classification Tree“) ist ein populärer Klassifikationsalgorithmus, der auch häufig in Ensemblemethoden (beispielsweise dem Random Forest) verwendet wird. Die ersten baumbasierten Klassifikations- (oder zunächst Regressions-) Methoden reichen in die 60er Jahre des 20. Jahrhunderts zurück (Morgan und Sonquist, 1963). Die Grundidee wurde über die folgenden Jahre weiterentwickelt, bis der hier verwendete Algorithmus CART (englische Abkürzung: „classification and regression trees“) (Breiman u. a., 1984) entstand. Die Darstellung orientiert sich an Breiman u. a. (1984) und Hastie u. a. (2009, S. 308–312) und wird in R durch das Paket `rpart` (Therneau u. a., 2015) realisiert, das die meisten Ideen aus Breiman u. a. (1984) implementiert.

Gegeben seien n Beobachtungen mit jeweils p Ausprägungen und einer Klasse $y_i \in \{1, 2\}$, $(\boldsymbol{x_i}, y_i) = ((x_{i1}, x_{i2}, ..., x_{ip}), y_i)$, $i = 1, ..., n$. Da in dieser Arbeit ausschließlich Zwei-Klassen-Probleme betrachtet werden, beschränkt sich die Darstellung hier auf diesen Spezialfall. Die Methode ist jedoch auch für mehr als zwei Klassen anwendbar.

Der Entscheidungsbaum teilt den Stichprobenraum $\mathbb{R}^p$ schrittweise durch binäre Schnitte in Untermengen ein. Anschließend wird jeder entstehenden Region eine Klasse zugeordnet. Neue Beobachtungen werden anhand ihrer Zugehörigkeit zu einer Region klassifiziert.

Zunächst wird eine einzelne Variable $X_j, j \in \{1, ..., p\}$ ausgewählt und ihr Wertebereich anhand eines einfachen Schwellenwertes s_1 in zwei Teile geteilt. Anhand dieses Schwellenwertes wird der Stichprobenraum $\mathbb{R}^p$ in die Mengen $R^1 = \{\boldsymbol{x_i} \in \mathbb{R}^p | x_{ij} < s_1\}$ und $R^2 = \{\boldsymbol{x_i} \in \mathbb{R}^p | x_{ij} \geqslant s_1\}$ eingeteilt. Anschließend können diese Mengen durch erneute binäre Schnitte einzelner

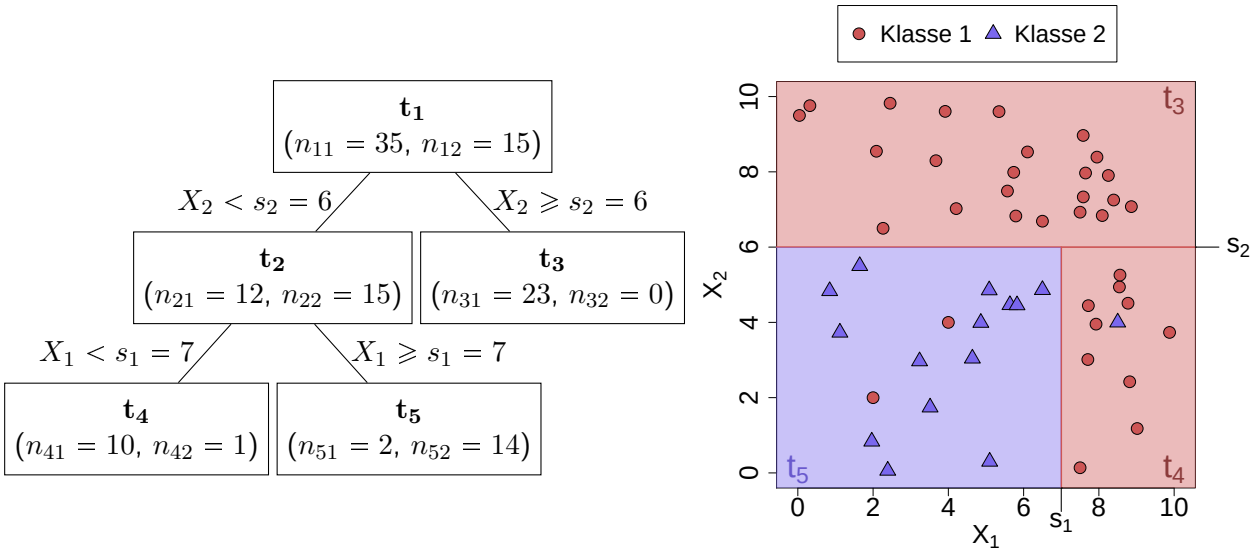


Abbildung 4.3: Beispiel für einen Klassifikationsbaum. Links: Darstellung als Baumdiagramm. Jeder Knoten wird durch ein Rechteck dargestellt. Rechts: Zugehörige Darstellung des Datensatzes mit Kennzeichnung der binären Schnitte sowie Einfärbung der Bereiche der Endknoten entsprechend der Klassifikationsregel.

Variablen weiter unterteilt werden. Die durch binäre Schnitte entstandenen Mengen heißen auch „Knoten“. Der erste Knoten wird auch „Wurzelknoten“ genannt, die letzten Knoten (also die Bereiche, die nicht durch weitere binäre Schnitte aufgeteilt werden), heißen auch „Endknoten“. Für jeden Knoten können die direkt verbundenen Nachbarknoten auch relativ als „Elternknoten“ (die unmittelbar *oberhalb* des Knotens liegenden Knoten, die eine Übermenge des aktuellen Knotens sind) oder „Kindknoten“ (die unmittelbar *unterhalb* des Knotens liegenden Knoten, die eine Untermenge des aktuellen Knotens darstellen) bezeichnet werden. Ein Beispiel für einen Baum ist in Abbildung 4.3 (links) dargestellt. In diesem Fall bezeichnet t_1 den Wurzelknoten, t_3, t_4 und t_5 sind die Endknoten. t_2 ist ein Kindknoten von t_1 und gleichzeitig Elternknoten von t_4 und t_5.

Ist ein Baum fertig gestellt, werden also keine weitere Knoten gebildet, werden für die M Endknoten Klassenzuordnungen vorgenommen. Dies geschieht anhand der Beobachtungen, die, würde man sie anhand ihrer Realisierungen von oben nach unten durch den Baum führen, im entsprechenden Knoten enthalten wären. Einem Endknoten wird diejenige Klasse zugeordnet, die bei den Beobachtungen in diesem Knoten am häufigsten auftritt (bei Gleichstand wird die Klasse zufällig gewählt). Der Anteil der insgesamt N_m Beobachtungen in Endknoten m

(beschrieben durch die Region R_m), die der Klasse k zugehörig sind, wird mit $\hat{p}_{mk}$ bezeichnet:

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{\boldsymbol{x}_i \in R_m} \mathbb{1}(y_i = k) \quad , \quad k \in \{1, 2\}, m \in \{1, ..., M\}.$$

Die Klassifikationsregel für eine Beobachtung x_i in Region R_m lautet entsprechend

$$k(m) = \arg\max_{k \in \{1,2\}} \hat{p}_{mk}.$$

Im Beispiel aus Abbildung 4.3 wird die zu Endknoten t_5 gehörige Region beschrieben durch $R = \{\boldsymbol{x_i} \in \mathbb{R}^2 | x_{i1} < 7 \wedge x_{i2} < 6\}$ (vgl. Abbildung 4.3, rechts). Die geschätzte Wahrscheinlichkeit für Klasse 1 ist gemäß ihrer relativen Häufigkeit im Knoten $\frac{1}{8}$, die für Klasse 2 entsprechend $\frac{7}{8}$. Dem Knoten wird daher die Klasse 2 zugeordnet. Jede neue Beobachtung, welche in die Region R fällt, wird in Klasse 2 eingeordnet.

Die Konstruktion eines Klassifikationsbaumes erfolgt schrittweise durch einen Greedyalgorithmus. Um einen bestehenden Knoten weiter aufzuteilen, werden eine Trennvariable X_j und ein Schwellenwert $s \in \mathbb{R}$ benötigt. Ziel dabei ist es, einen Knoten so aufzuteilen, dass die resultierenden Kindknoten jeweils möglichst rein sind, das heißt, möglichst viele Beobachtungen *einer* Klasse und möglichst wenige Beobachtungen aus anderen Klassen enthalten.

Um die Reinheit eines Knotens K beurteilen zu können, wird ein Unreinheitsmaß verwendet. Das hier verwendete Unreinheitsmaß ist der Gini-Index. Er berechnet sich zu

$$I_{\text{gini}(K)} = \sum_{k=1}^{2} \hat{p}_{mk}(1 - \hat{p}_{mk}).$$

Der Gini-Index könnte als Schätzer für die Fehlklassifikationswahrscheinlichkeit in diesem Knoten interpretiert werden, wenn die Klassenlabel (anstatt des Mehrheitsentscheids) in dem Knoten gemäß der relativen Häufigkeiten der beobachteten Klassen vergeben würden.

Es wird das Paar (j, s) als bester Split bezeichnet, das die größte Reduktion der Unreinheit eines Knotens, durch Splitten der Variable X_j mit Schwellenwert s, bewirkt. Hierbei bezeichne $I_{\text{gini}}(K)$ die Unreinheit im Knoten K. Die durch den Split $S = (j, s)$ entstehenden linken und rechten Knoten werden mit K_l und K_r, die Anteile der Beobachtungen in den jeweiligen Knoten mit p_l und p_r bezeichnet. Dann lautet die bezüglich S zu minimierende Reduktion des Gini-Indexes:

$$\Delta I_{\text{gini}}(K, S) = I_{\text{gini}}(K) - p_l I_{\text{gini}}(K_l) - p_r I_{\text{gini}}(K_r).$$

Um eine Überanpassung zu vermeiden, sollte ein Klassifikationsbaum nicht zu viele Splits enthalten. Grundsätzlich wird ein Knoten nicht weiter aufgeteilt, wenn alle Beobachtungen in einem Knoten nur noch einer Klasse angehören. Jedoch ist es nicht grundsätzlich erstrebenswert, dass ein fertiger Baum nur reine Knoten enthält, da hierbei oft Splits verwendet werden, die nur im bestehenden Datensatz zur Klassifikation beitragen, aber nicht auf neue Daten generalisierbar sind. Aus diesem Grund sollten Bäume in der Regel gestutzt werden. Zu diesem Zweck gibt es verschiedene Möglichkeiten. Um nur Splits zu berücksichtigen, welche die Klassifikationsgüte des Baumes angemessen verbessern, kann gefordert werden, dass ein Knoten nur dann weiter aufgeteilt wird, wenn sich die Unreinheit des Knotens durch die Aufteilung um einen bestimmten Wert verringert (bspw. wenn die Reduktion des Gini-Indexes mindestens 0.01 beträgt). Eine weitere Möglichkeit ist es, die Tiefe des Baumes (also die maximal hintereinander auszuführenden Splits, um einen Endknoten zu beschreiben) zu beschränken. Darüber hinaus kann ein Aufteilen in sehr viele kleine Knoten verhindert werden, indem die Mindestgröße eines zu teilenden Knotens oder die Mindestgröße eines Endknotens festgelegt werden.

Da in Kapitel 4.5.5 auch *Regressions*bäume (Hastie u. a., 2009, S. 305–307) verwendet werden, werden hier kurz die Unterschiede zum *Klassifikations*baum dargestellt. Bei Regressionsbäumen ist die Zielvariable y stetig und nicht mehr kategoriell. Der Stichprobenraum wird ebenfalls durch binäre Schnitte eingeteilt. Für alle Beobachtungen in einem Endknoten wird der gleiche Wert geschätzt, indem der Mittelwert der darin enthaltenen Beobachtungen gebildet wird. Die optimale Kombination für die Splitvariable j und den Split s wird bestimmt, indem für alle Kombinationen die daraus resultierende Summe der quadratischen Abstände der Schätzung zum wahren Wert berechnet wird und die Kombination mit dem kleinsten Ergebnis ausgewählt wird.

4.5.4 Random Forest (RF)

Klassifikationsbäume haben einen geringen Bias, wenn sie ausreichend tief gebildet werden, jedoch eine große Varianz (wenn früh andere Splits durchgeführt werden, ändert sich schnell der ganze Baum). Um die geringe Verzerrung zu erhalten und die große Varianz zu minimieren, wurden verschiedene Ansätze entwickelt, bei denen viele Bäume gebildet und auf unterschiedliche Arten aggregiert werden. Eine von ihnen ist der weit verbreitete Random

Forest (RF, „Entscheidungswald“). Er wurde von Breiman (2001) definiert und wird hier nach Hastie u. a. (2009, S. 587–589) dargestellt. In R wird die Implementierung aus dem Paket `randomForest` (Liaw und Wiener, 2002) verwendet.

Die Bezeichnungen seien wie in Kapitel 4.5.3. Um einen Random Forest zu bilden, werden zunächst B Bootstrap-Stichproben vom Umfang n gezogen. Es wird also B Mal eine Stichprobe vom Umfang n mit Zurücklegen aus den n vorhandenen Beobachtungen gezogen. Für jede dieser Stichproben wird ein Entscheidungsbaum erstellt, wobei die Trennvariable für jeden Split eines Knotens nur aus zufälligen $p^* < p$ der Variablen ausgewählt wird. Unter diesen p^* Variablen wird dann, wie in Kapitel 4.5.3 beschrieben, der beste Split bestimmt. Für den nächsten Knoten werden erneut p^* Variablen zufällig ausgewählt und nur unter diesen wird der beste Split ausgewählt. Ein typischer Wert für p^* ist $\lfloor\sqrt{p}\rfloor$.

Insgesamt liegen schlussendlich B Klassifikationsbäume $T_1, ..., T_B$ vor. Eine neue Beobachtung $\boldsymbol{x}$ wird gemäß des Mehrheitsentscheids der Klasse zugeordnet, die unter den Vorhersagen der Einzelbäume am häufigsten vorkommt. Bei gleicher Häufigkeit beider Klassen erfolgt die Zuordnung zufällig.

Die einzelnen Bäume innerhalb des Random Forests werden hier nicht gestutzt. Nach Hastie u. a. (2009, S. 596) kann darauf verzichtet werden, ohne dass zu große Verluste (in Form von Varianzerhöhung) durch Überanpassung entstehen.

Variablenwichtigkeit Um beurteilen zu können, wie wichtig die einzelnen Variablen für die Klassifikation durch den Random Forest sind, kann die *Variablenwichtigkeit* als Maß für die Wichtigkeit aller Variablen (Hastie u. a., 2009, S. 593) berechnet werden. Dafür wird für jede Variable berücksichtigt, wie oft sie in den Bäumen verwendet wurde und wie groß die Reduktion des Unreinheitsmaßes eines Splits ist, indem diese Reduktionen über alle Bäume hinweg summiert werden. Wird eine Variable in vielen Bäumen verwendet und erreicht sie dort stets große Unreinheitsreduktionen, so wird die Wichtigkeit dieser Variable hoch bewertet im Vergleich zu Variablen, die nur selten verwendet werden und/oder nur geringe Reduktionen des Unreinheitsmaßes erzielen.

4.5.5 Gradienten Basiertes Boosting (GBM)

Das Gradienten Basierte Boosting (GBM) stellt ebenfalls ein Verfahren dar, welches viele Bäume nutzt, um eine Klassifikationsregel aufzustellen. Die Herangehensweise ist jedoch mit der des Random Forests nicht zu vergleichen und im Unterschied dazu werden statt Klassifikations- hier Regressionsbäume verwendet. Die Darstellung orientiert sich an Ridgeway (2019), welche die im verwendeten R-Paket `gbm` (Ridgeway, 2013) implementierte Version beschreibt.

Die Grundidee des Boostings besteht darin, viele schwache Klassifikatoren zu einem guten Klassifikator zu verbinden. Die Beobachtungen seien $\boldsymbol{x_i}, i = 1, ..., n$, die zugehörige beobachtete Klasse $y_i \in \{0, 1\}$. Die Güte der Klassifikation wird über eine Verlustfunktion bewertet. Hier wird die Bernoulli-Verlustfunktion verwendet:

$$\Psi(\boldsymbol{y}, f(\boldsymbol{x})) = -\sum_{i=1}^{n} \big(y_i f(\boldsymbol{x_i}) - \log(1 + \exp(f(\boldsymbol{x_i})))\big).$$

Sie entspricht der negativen der log-Likelihood der Daten mit $f(\boldsymbol{x}) = \log(\frac{p}{1-p})$ (entspricht den log-Odds), wobei p die Wahrscheinlichkeit der positiven Klasse beschreibt.

Im ersten Schritt des Algorithmus wird die Funktion $f(\boldsymbol{x})$ als Konstante geschätzt. Um die Verlustfunktion zu minimieren, wird die Verlustfunktion $\Psi\big(y_i, \hat{f}^{(0)}(\boldsymbol{x})\big)$ bezüglich $\hat{f}^{(0)}(\boldsymbol{x})$ abgeleitet und gleich Null gesetzt. Auf diese Weise ergibt sich der initiale Wert zu

$$\hat{f}^{(0)}(\boldsymbol{x}) = \log\left(\frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n}(1 - y_i)}\right).$$

Würde man diese log-Odds in Wahrscheinlichkeiten umrechnen, so entspräche diese Wahrscheinlichkeit der relativen Häufigkeit der Klasse.

Anschließend wird diese Schätzung iterativ verbessert. Dazu wird zunächst der negative Gradient der Verlustfunktion (bezüglich $f(\boldsymbol{x})$) berechnet und an der Stelle der aktuellen Schätzung ausgewertet:

$$\begin{aligned} z_i^{(t)} &= -\frac{\partial}{\partial f(\boldsymbol{x_i})}\Psi(y_i, f(\boldsymbol{x_i}))\bigg|_{f(\boldsymbol{x_i})=\hat{f}^{(t-1)}(\boldsymbol{x_i})} \\ &= y_i - \frac{1}{1 + \exp\left(-\hat{f}^{(t-1)}(\boldsymbol{x_i})\right)}. \end{aligned}$$

Dabei ist $t = 1, ..., T$ die aktuelle der insgesamt auf T festgelegten Iterationen. Dies entspricht dem Vorgehen beim Gradientenabstiegsverfahren. Setzt man für $\hat{f}^{(t-1)}(\boldsymbol{x})$ die aktuell geschätzten log-Odds ein, so ergeben sich die $z_i = y_i - \hat{p}_i^{(t-1)}$. Diese können auch als „Pseudo-Residuen" bezeichnet werden, da sie Abstände zwischen den wahren Klassenzugehörigkeiten und den geschätzten Wahrscheinlichkeiten darstellen. Diese Residuen werden anschließend durch einen Regressionsbaum beschrieben, welcher auf einer zufälligen Auswahl (ohne Zurücklegen) von $n^* < n$ Beobachtungen berechnet wird. Die Tiefe des Regressionsbaums wird durch eine festgelegte Anzahl an Endknoten K beschränkt.

Sei S_k die Menge der Beobachtungen, welche im Endknoten k liegen. Die Vorhersage für die Beobachtungen in diesem Knoten wird so berechnet, dass sie den zugehörigen Verlust minimiert, wenn auf die bisherige Vorhersage ein konstanter Wert addiert wird:

$$\begin{aligned}\rho_k^{(t)} &= \arg\min_{\rho} \sum_{\boldsymbol{x_i} \in S_k} \Psi(y_i, \hat{f}^{(t-1)}(\boldsymbol{x_i}) + \rho) \\ &= \frac{\sum_{\boldsymbol{x_i} \in S_k} (y_i - p_i^{(t-1)})}{\sum_{\boldsymbol{x_i} \in S_k} p_i^{(t-1)}(1 - p_i^{(t-1)})} \quad , \quad \text{wobei } p_i^{(t-1)} = \frac{1}{1 + \exp\left(-\hat{f}^{(t-1)}(\boldsymbol{x_i})\right)}.\end{aligned}$$

Im letzten Schritt wird die Funktion (für die zufällig ausgewählten Beobachtungen, die im Baum repräsentiert sind) aktualisiert, wobei die Geschwindigkeit des Gradientenabstiegs durch eine Lernrate λ bestimmt wird:

$$\hat{f}^{(t)}(\boldsymbol{x_i}) = \hat{f}^{(t-1)}(\boldsymbol{x_i}) + \lambda \rho_{k(\boldsymbol{x_i})}.$$

Dabei bezeichne $k(\boldsymbol{x_i})$ den Index des Endknotens, in den die Beobachtung $\boldsymbol{x_i}$ fällt.

Anschließend werden die Schritte (beginnend mit der Berechnung des negativen Gradienten) wiederholt, bis T Iterationen durchgeführt wurden. Für die Klassifikation einer neuen Beobachtung $\boldsymbol{x_{i*}}$ wird diese durch alle Bäume geführt und $\hat{f}^{(T)}(\boldsymbol{x_{i*}})$ berechnet, was einen Schätzer für die log-Odds darstellt. Der Schätzer für die Wahrscheinlichkeit, dass eine Beobachtung zur positiven Klasse gehört, ist $p_{i*}^{(T)}$.

4.5.6 Gütemaßzahlen

Im Zwei-Klassen-Klassifikationsproblem kann der Erfolg eines Klassifikators durch verschiedene Gütemaßzahlen bewertet werden (die Darstellung orientiert sich an Wehberg u. a. (2007)). In dieser Arbeit wird primär die Fläche unter der ROC-Kurve (AUC) betrachtet, sekundär

		Wahrheit		
		$W{=}1$	$W{=}2$	$\sum$
Klassifikation	$K{=}1$	n_{11}	n_{12}	n_{1+}
	$K{=}2$	n_{21}	n_{22}	n_{2+}
	$\sum$	n_{+1}	n_{+2}	N

Tabelle 4.1: Vierfeldertafel des Klassifikationsergebnisses im Vergleich zum wahren Status.

werden zusätzlich die Korrektklassifikationsrate (accuracy, ACC), die Richtig-positiv-Rate (auch: Sensitivität, true positive rate, TPR), die Richtig-negativ-Rate (auch: Spezifität, true negative rate, TNR), der positive prädiktive Wert (positive predictive value, PPV) und der negative prädiktive Wert (negative predictive value, NPV) angegeben. Werden die Vorhersagen der Klassifikation (Klasse 1 oder Klasse 2) mit den wahren Klassen (wahre Klasse 1 oder wahre Klasse 2) für alle Beobachtungen verglichen, können sie in einer Vierfeldertafel (Tabelle 4.1) zusammengefasst werden. Klasse 1 sei dabei die „positive“ und Klasse 2 die „negative“ Klasse.

Die ACC bezeichnet den Anteil aller richtig klassifizierter Beobachtungen. Die TPR beschreibt den Anteil der richtig als positiv klassifizierten Beobachtungen unter allen in Wahrheit positiven Beobachtungen. Die TNR ist der Anteil richtig als negativ klassifizierten Beobachtungen an allen in Wahrheit negativen Beobachtungen. Der PPV beschreibt umgekehrt den Anteil richtig positiver Beobachtungen an allen positiv klassifizierten Beobachtungen, der NPV den Anteil richtig negativer Beobachtungen an allen negativ klassifizierten Beobachtungen. Diese Maßzahlen können als Schätzer der entsprechenden Wahrscheinlichkeiten angesehen werden (beispielsweise ist die TPR ein Schätzer für die Wahrscheinlichkeit, eine Beobachtung positiv zu klassifizieren, wenn sie in Wahrheit der positiven Klasse angehört). Im Fall einer perfekten Klassifikation nehmen alle Maßzahlen den Wert Eins an. Die Maßzahlen berechnen sich wie folgt:

$$\text{ACC} = \frac{n_{11} + n_{22}}{N} \quad \text{TPR} = \frac{n_{11}}{n_{+1}} \quad \text{TNR} = \frac{n_{22}}{n_{+2}} \quad \text{PPV} = \frac{n_{11}}{n_{1+}} \quad \text{NPV} = \frac{n_{22}}{n_{2+}}.$$

Die Klassifikation einer Beobachtung erfolgt anhand einer Klassifikationsfunktion, welche beispielsweise Wahrscheinlichkeiten für die positive Klasse ausgibt. Um die Beobachtungen einer Klasse zuzuordnen, muss diese Wahrscheinlichkeit in eine binäre Zuordnung zu beiden Klassen umgewandelt werden. Dazu wird oft der Schwellenwert 0.5 verwendet. Beobachtungen mit einer höheren Wahrscheinlichkeit als der Schwellenwert werden der positiven Klasse

zugeordnet. Wird der Schwellenwert nicht als fest angenommen, so kann die ROC-Kurve betrachtet werden. Dafür werden für alle möglichen Schwellenwerte die TPR und die TNR berechnet. Anschließend werden in einer Grafik alle zugehörigen Paare $((1 - \text{TNR}), \text{TPR})$ durch lineare Interpolation verbunden und somit als Kurve dargestellt. Die Kurve beginnt im Punkt $(0, 0)$ (alle Beobachtungen werden der negativen Klasse zugeordnet, also ist TPR=0 und TNR=1) und endet im Punkt $(1, 1)$ (alle Beobachtungen werden der positiven Klasse zugeordnet, also ist TPR = 1 und TNR = 0). Dazwischen verläuft sie monoton wachsend. Eine Maßzahl, welche alle Schwellenwerte berücksichtigt, wird durch die Fläche unter der Kurve (AUC) beschrieben. Auf diese Weise kann das Klassifikationsverfahren beurteilt werden, ohne dass dabei ein konkreter Schwellenwert ausgewählt werden muss. Gibt es einen Schwellenwert, für den alle Beobachtungen korrekt klassifiziert werden, so ist der AUC-Wert Eins. Erfolgt die Klassifikation zufällig, verläuft die Kurve auf der Winkelhalbierenden und der AUC-Wert wird ungefähr bei 0.5 liegen, wenn beide Klassen gleich stark besetzt sind.

4.5.7 Wahl des Wahrscheinlichkeits-Schwellenwerts

Die in Kapitel 4.5 vorgestellten Klassifikationsalgorithmen geben für die Klassifikation einer Beobachtung Wahrscheinlichkeitsschätzungen, zu den beiden Klassen zu gehören, aus. Anhand eines Schwellenwerts wird festgelegt, ab welcher Wahrscheinlichkeit die Beobachtungen der positiven Klasse zugeordnet werden. Die im vorigen Kapitel genannten Maßzahlen ACC, TPR, TNR, PPV und NPV hängen somit von einem zuvor festgelegten Schwellenwert ab. Als Standardwert wird 0.5 verwendet, das heißt, eine Beobachtung wird der Klasse zugeordnet, für welche die geschätzte Wahrscheinlichkeit mehr als 50% beträgt. Diese Zuordnung ist zwar intuitiv logisch, jedoch nicht immer optimal (Freeman und Moisen, 2008) und kann auch anders gewählt werden.

Dies ist beispielsweise sinnvoll, wenn die Klassen sehr ungleichmäßig besetzt sind. So kann vermieden werden, dass häufige Klassen durch die Klassifikationsregel bevorzugt werden und sehr niedrige MMCE-Werte erreicht werden, obwohl die kleinere Klasse oft falsch klassifiziert wird (was sich in sehr niedriger Sensitivität oder Spezifität ausdrückt). Aber auch, wenn beispielsweise die Prävalenz der Klassen bei der Klassifikation erhalten bleiben soll, kann der Schwellenwert auf eine andere Weise bestimmt werden.

Entsprechend kann der Schwellenwert so gewählt werden, dass die geschätzte Prävalenz im Modell der Prävalenz im Datensatz entspricht (Manel u. a., 2001). Dies wird erreicht, indem das Klassifikationsmodell zunächst auf alle Beobachtungen, die für das Training des Modells

verwendet wurden, angewendet wird. Die geschätzten Wahrscheinlichkeiten für die positive Klasse im Datensatz werden anschließend geordnet. Der Schwellenwert wird dann so gewählt, dass genau so viele Beobachtungen der positiven Klasse zugeordnet werden, wie in Wahrheit in der positiven Klasse enthalten sind. Sind in der negativen Klasse n_{neg} Beobachtungen, so wird der Schwellenwert auf den Mittelwert der Beobachtungen mit den Rängen n_{neg} und $(n_{\text{neg}} + 1)$ gesetzt. Es ist zu beachten, dass der Schwellenwert nicht auf eine realisierte Wahrscheinlichkeit gesetzt wird. Auf diese Art wird eine Verzerrung des Schwellenwerts verhindert (bei einfachen Klassifikationsproblemen können die Wahrscheinlichkeiten beider Gruppen sehr weit auseinander liegen, sodass der Schwellenwert auf diese Art mittig dazwischen und nicht näher an der einen Gruppe als an der anderen liegt). Im Folgenden wird das beschriebene Vorgehen als *Prävalenz-Methode* bezeichnet.

5 Algorithmen im Gesamtanalyseprozess

In diesem Kapitel wird die Frage beantwortet, ob automatische Algorithmen die Problematik der Peakerkennung für MCC-IMS-Daten gut genug bearbeiten können, um eine nachfolgende Klassifikationsaufgabe erfolgreich zu lösen. Darüber hinaus wird aus den betrachteten automatischen Algorithmenkombinationen die beste Kombination als Empfehlung für zukünftige Anwendungen ausgewählt. Als Goldstandard der Peakerkennung dient die (semi-)manuelle Auswertung der Rohmessungen mit VisualNow durch geübte Experten. Die verschiedenen automatischen Verfahren sollen in der Klassifikation idealerweise also mindestens ebenso gute Resultate erzielen wie die (semi-)manuelle Variante. Um gleichzeitig eine Abhängigkeit vom Klassifikationsalgorithmus auszuschließen, beziehungsweise um das am besten geeignete Klassifikationsverfahren auszuwählen, werden auch verschiedene Klassifikationsalgorithmen getestet.

Ergebnisse dieser Studie sind bereits in Horsch u. a. (2015), Horsch u. a. (2017) und Kopczynski (2017) beschrieben. Die Darstellung ist hier jedoch deutlich ausführlicher. Dabei werden unter anderem die drei einzelnen Schritte aus Peakauswahl, Peakclustern und Klassifikation detaillierter betrachtet. Zudem wird die Güte der Peakerkennung nicht nur durch die Analyse der Klassifikationsergebnisse bewertet sondern auch anhand gemittelter Rohmessungen auf Plausibilität überprüft.

5.1 Datensätze

Insgesamt liegen drei Datensätze aus verschiedenen Anwendungsgebieten der Atemluftforschung vor, wie in Kapitel 3.1 beschrieben. Jeder Datensatz repräsentiert ein Zwei-Klassen-Problem, bei welchem gesunde Personen („Kontrollen“) von kranken Personen („Fällen“) unterschieden werden sollen. In Tabelle 5.1 sind die Anzahlen der beiden Klassen für die drei Datensätze zusammengefasst. Beim ersten Datensatz sind die beiden Klassen sehr unterschiedlich groß, die Fälle sind mit 92 Personen deutlich stärker vertreten als die Kontrollen

Tabelle 5.1: Anzahl der Beobachtungen in den beiden Klassen für die drei Datensätze.

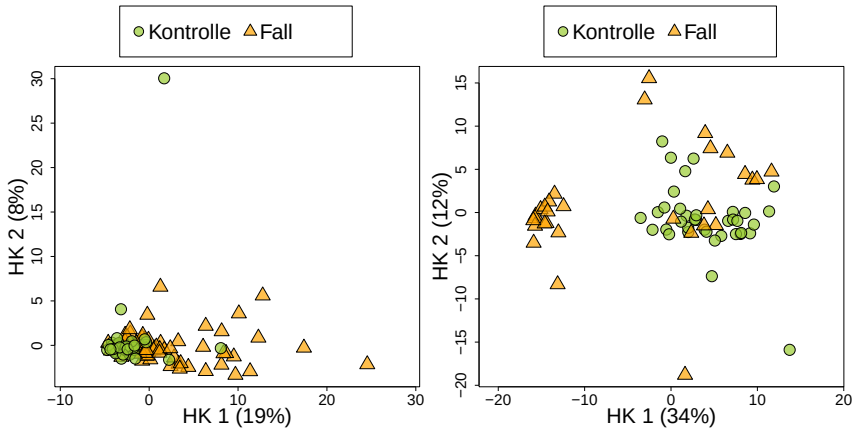
	Kontrollen	Fälle	$\sum$
Datensatz 1	35	92	127
Datensatz 2	37	30	67
Datensatz 3	30	39	69

mit 35 Personen. In den beiden anderen Datensätzen sind etwa 70 Personen enthalten, wobei die beiden Klassen gleichmäßiger vertreten sind (37 Kontrollen und 30 Fälle für Datensatz 2 sowie 30 Kontrollen und 39 Fälle bei Datensatz 3).

Für einen kurzen deskriptiven Überblick über die drei Datensätze wird eine Hauptkomponentenanalyse für die mit dem (semi-)manuellen Goldstandard VN^{m} ausgewerteten Datensätze durchgeführt. Da die Messungen schon früher verwendet wurden, liegen diese bereits vor. Dabei wurden für den ersten Datensatz 120, für den zweiten 224 und für den dritten Datensatz 60 Peaks annotiert. In Abbildung 5.1 sind die Beobachtungen jeweils in ihrer Darstellung als die ersten Hauptkomponenten (HK) in Form von Streudiagrammen dargestellt (Individuenplot). Die Beobachtungen wurden vor der Hauptkomponentenanalyse zentriert und standardisiert.

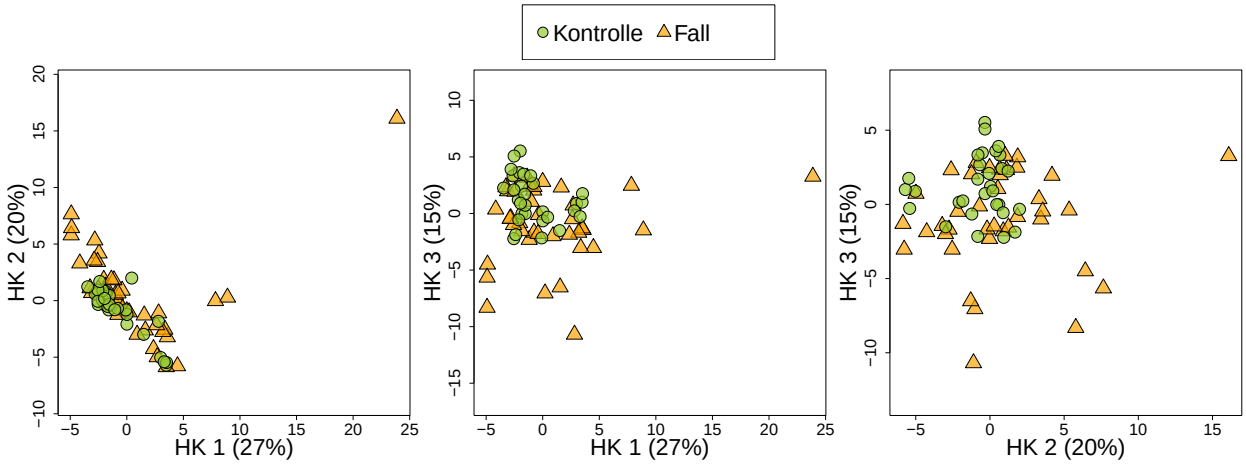
Für den ersten Datensatz (Abbildung 5.1a) sind die ersten beiden Hauptkomponenten dargestellt, wobei die erste etwa 19% der Variabilität in den Daten erklärt, die zweite nur noch 8%. Die erste Hauptkomponente trennt einige der Fälle von den Kontrollen, wobei höhere Werte für einen Fall sprechen. Die zweite Hauptkomponente enthält keine relevante Information über die Klassen und trennt in erster Linie eine Beobachtung von den übrigen ab. Die ersten beiden Hauptkomponenten erklären nur recht wenig Varianz (27%) in den Daten und alle folgenden Hauptkomponenten erklären jeweils weniger als 8%. Das bedeutet, dass viele Hauptkomponenten benötigt werden, um den Großteil der Varianz erklären zu können.

Für den zweiten Datensatz sind ebenfalls die ersten beiden Hauptkomponenten dargestellt (Abbildung 5.1b). Die erste Hauptkomponente, welche 34% der Variabilität erklärt, trennt etwa die Hälfte der Fälle von den übrigen Beobachtungen. Niedrige Werte für diese Hauptkomponente sprechen für Fälle. Die zweite Hauptkomponente, die noch 12% der Variabilität in den Daten erklärt, nimmt für die Fälle der übrigen Beobachtungen im Mittel etwas höhere Werte an als für Kontrollen, auch wenn hier keine vollständige Trennung vorliegt.



(a) Datensatz 1

(b) Datensatz 2



(c) Datensatz 3

Abbildung 5.1: Die wichtigsten Hauptkomponenten für die drei Datensätze bei Verwendung der (semi-)manuellen Peakerkennung.

Da für den dritten Datensatz auch die dritte Hauptkomponente noch 15% der Variabilität erklärt, sind hier die ersten drei Hauptkomponenten, jeweils paarweise, dargestellt (Abbildung 5.1c). Während die erste Hauptkomponente keine Informationen über die Klasse zu enthalten scheint, gibt es bei der zweiten und dritten Hauptkomponente einige Fälle, die sich bezüglich ihrer Werte von den übrigen Beobachtungen unterscheiden. Dies ist im dritten Bild (rechts) am deutlichsten zu erkennen. Insgesamt gibt es dabei für den Großteil der Beobachtungen jedoch keine klare Trennung.

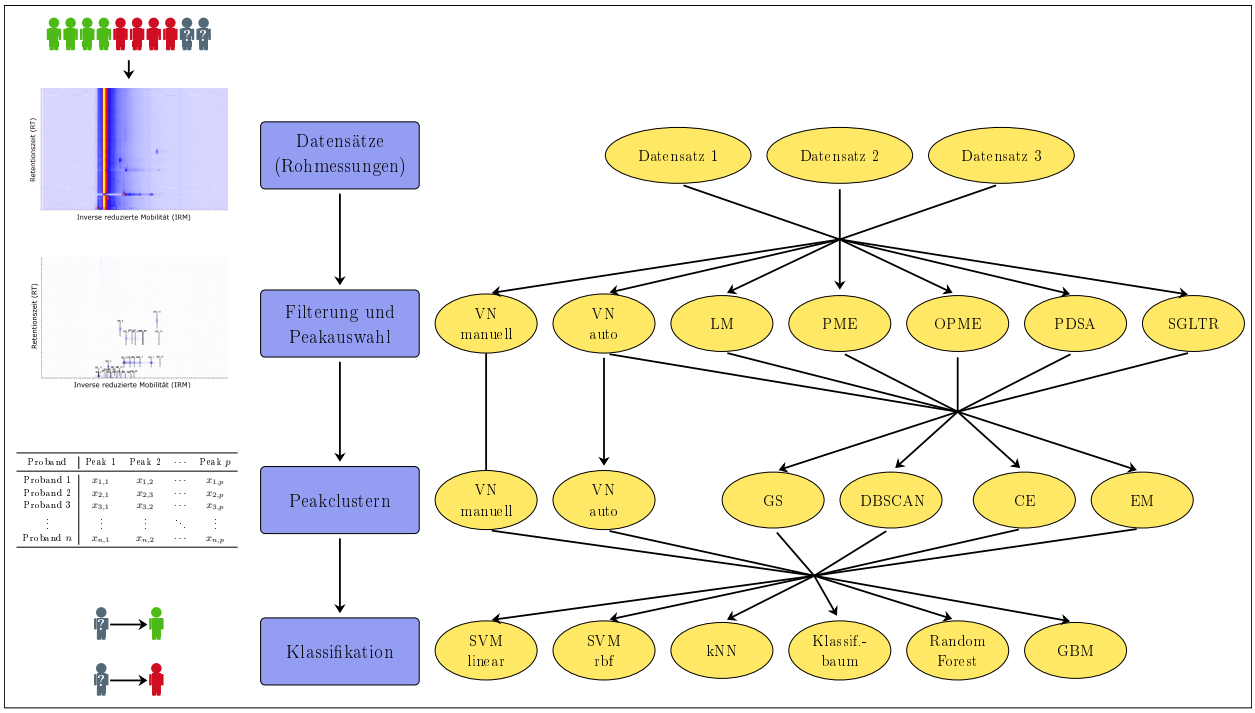


Abbildung 5.2: Schematische Darstellung des Studienaufbaus. Mögliche Kombinationen der Datensätze und der Algorithmen für Peakauswahl, Peakclustern und Klassifikation sind durch Pfeile dargestellt. Für die manuelle Peakerkennung mit VisualNow (VN^m) ist zu beachten, dass Peakauswahl und Peakclustern nicht getrennt voneinander sondern in einem gemeinsamen Schritt durchgeführt werden.

5.2 Aufbau der Studie

Um die automatische Peakerkennung mit der (semi-)manuellen vergleichen zu können, wird auf jeden Datensatz eine große Anzahl an automatischen Algorithmenkombinationen angewendet. Die Peakauswahl- und Peakcluster-Methoden wurden von Dominik Kopczynski im Rahmen eines gemeinsamen Forschungsprojekts im SFB 876 (Teilprojekt TB1: „Ressourcen-beschränkte Analyse von Spektrometriedaten“) auf die Rohdaten angewendet. Unter anderem wurde dabei auf Software-Umsetzungen und entwickelte Methoden einer Projektgruppe der Fakultät Informatik an der TU Dortmund zurückgegriffen (Egorov u. a., 2014). Die anschließenden Klassifikationen sowie die weiterführenden Analysen wurden von der Autorin durchgeführt.

Der Aufbau der durchgeführten Vergleichsstudie ist in Abbildung 5.2 schematisch dargestellt. Zunächst wird für jede Rohmessung der drei Datensätze einzeln eine Peakauswahl durchgeführt. Dies beinhaltet die in Kapitel 4.2.2 vorgestellten automatischen Algorithmen Local

Maxima (LM), Peak Model Estimation (PME), Peak Detection by Slope Analysis (PDSA), Savitzky-Golay Laplace-operator filter thresholding regions (SGLTR), Online Peak Model Estimation (OPME) und automatisches VisualNow (VN^a). Das Ergebnis jedes automatischen Peakauswahlalgorithmus ist eine sogenannte *Peakliste*. Sie enthält für jede einzelne Rohmessung Informationen über die gefundenen Peaks dieser Messung. Diese umfassen insbesondere die Retentionszeit und die inverse reduzierte Mobilität (IRM), welche die Position eines Peaks beschreiben, sowie die Signalintensität des Peaks. Die Peaks, die in den einzelnen Rohmessungen gefunden werden, werden auch *Single Peaks* genannt.

Diese Positionen werden anschließend verwendet, um das Peakclustern durchzuführen. Dabei werden alle Rohmessungen eines Datensatzes betrachtet. Nahe beieinander liegende Peaks verschiedener Messungen werden geclustert und als ein gemeinsamer Consensus Peak interpretiert, welcher für die gleiche gemessene Substanz in der Atemluft steht. Die Position eines Consensus Peaks ergibt sich aus der Zusammenfassung der Positionen der Single Peaks, welche im Cluster enthalten sind (dies wird für die verschiedenen Methoden auf unterschiedliche Weise erreicht). Die automatischen Peakcluster-Methoden Grid Squares (GS), Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Cluster Editing (CE), EM Clustering (EM) und automatisches VisualNow (VN^a) aus Kapitel 4.2.3 werden also im Anschluss für alle möglichen Kombinationen mit den automatischen Peakauswahlmethoden durchgeführt. Dabei ist zu beachten, dass das Programm VN^a beide Schritte (Peakauswahl und Peakclustern) durchführt. Die resultierenden Peaklisten können nach dem ersten Schritt exportiert werden, sodass sie mit anderen Peakcluster-Methoden kombiniert werden können. Es ist allerdings nicht möglich, die Peaklisten anderer Peakauswahlmethoden in das Programm zu importieren, sodass das zugehörige Cluster-Programm nicht mit anderen Peakauswahlalgorithmen kombiniert werden kann.

Neben den automatischen Algorithmen wurde bereits die (semi-)manuelle Peakerkennung VN^m als Goldstandard durchgeführt. Bei dieser Methode laufen die beiden Schritte Peakauswahl und Peakclustern gleichzeitig ab und können nicht voneinander getrennt werden. Aus diesem Grund kann bei dieser Methode nicht mit den automatischen Algorithmen der Peakauswahl oder des Peakclusterns kombiniert werden. Da das Endresultat der (semi-)manuellen Peakerkennung die gleiche Struktur aufweist, wie sie bei den übrigen Kombinationen im Anschluss an das Peakclustern vorliegen, wird VN^m stets bei den Ergebnissen der Peakcluster-Methoden behandelt.

Insgesamt ergeben sich 26 Kombinationen aus Peakauswahl und Peakclustern (einschließlich der manuellen Peakerkennung). Das Resultat des Peakclusterns sind die sogenannten „Layer“. Diese enthalten insbesondere die Peakpositionen der Consensus Peaks. Außerdem wird beim Peakclustern ein Datensatz erstellt, der die Consensus Peaks als Variablen sowie die dazugehörige Peakintensität für alle Rohmessungen des Datensatzes enthält. Für jede Beobachtung wird außerdem die Klasse (Fall oder Kontrolle) als Variable aufgenommen.

Die resultierenden Datensätze werden anschließend mit den Klassifikationsalgorithmen Support Vector Machine mit linearem (SVM^{lin}) und rbf-Kern (SVM^{rbf}), dem k-Nächste-Nachbarn-Verfahren (kNN), einem Klassifikationsbaum (CT), dem Gradienten Basierten Boosting (GBM) und dem Random Forest (RF) klassifiziert. Jede Klassifikation wird in einer 10-fachen stratifizierten Kreuzvalidierung durchgeführt, um überoptimistische Resultate zu vermeiden und unbalancierte Klassenverteilungen zu berücksichtigen. Um außerdem die Abhängigkeit der Ergebnisse einer Kreuzvalidierung von der zufälligen Einteilung in Trainings- und Testdaten zu berücksichtigen, wird jede Kreuzvalidierung 50 Mal mit unterschiedlichen Partitionen wiederholt. Da in einigen Klassifikationsalgorithmen Parameter optimiert werden, wird in diesen Fällen eine geschachtelte Kreuzvalidierung (ebenfalls 10-fach und stratifiziert) durchgeführt. Auf diese Weise werden auch die Parameter nicht überoptimistisch gewählt, wodurch die Beeinträchtigung der Güte der Klassifikation vermieden wird. Die optimierten Parameter sowie die verwendeten R-Pakete für die Klassifikation sind im Anhang auf den Seiten 164 bis 165 aufgeführt.

Für die Auswertung werden in Kapitel 5.3 die Ergebnisse der Peakerkennung betrachtet. Dafür wird deskriptiv untersucht, wie viele Peaks die verschiedenen Verfahren für die drei Datensätze finden. Anschließend werden die Ergebnisse der Klassifikation genutzt, um die Güte der einzelnen Verfahren und deren Kombinationen zu beurteilen. In Kapitel 5.4 werden alle Algorithmenkombinationen gleichzeitig betrachtet. Anschließend werden die Ergebnisse noch auf Stabilität geprüft, indem die drei Schritte Peakauswahl, Peakclustern und Klassifikation in Kapitel 5.5 einzeln betrachtet werden. Die Ergebnisse werden in Kapitel 5.6 zusammengefasst, um eine automatische Algorithmenkombination für zukünftige Anwendungen zu empfehlen.

5.3 Gefundene Peaks

Zunächst werden die Ergebnisse der Peakauswahl für die verschiedenen Algorithmen betrachtet. Da dieser Schritt nur bei den automatischen Methoden existiert, kommt der Goldstandard VN^m hier nicht vor. Es wird hier deskriptiv verglichen, wie unterschiedlich die Anzahlen gefundener Peaks je Rohmessung sind. Die Qualität der Peakerkennung wird hier nicht direkt betrachtet, da hierfür jede Rohmessung einzeln manuell gesichtet und für alle Peakauswahlverfahren ausgewertet werden müsste. Stattdessen wird die Güte der Peakerkennung anhand der Klassifikationsergebnisse beurteilt. Die Positionen der Consensus Peaks des Goldstandards, der automatischen Peakerkennung, die in Kapitel 5.6 empfohlen wird sowie die Positionen einer in der Klassifikation ungünstigen automatischen Peakerkennung werden in Kapitel 5.6 miteinander verglichen, indem sie in mittlere Rohmessungen eingezeichnet werden. Auf diese Art kann die Qualität grob eingeschätzt werden.

In Tabelle 5.2 sind die minimalen, medianen und maximalen Anzahlen gefundener Peaks der einzelnen Peakauswahlverfahren über alle Rohmessungen eines Datensatzes hinweg dargestellt. Die Anzahlen schwanken zwischen Null und 143, es gibt also sowohl Rohmessungen, in denen einzelne Peakauswahlverfahren keinen einzigen Peak finden als auch Rohmessungen in denen sehr viele Peaks annotiert werden. Die größten medianen Anzahlen für die drei Datensätze werden von SGLTR (42 Peaks bei Datensatz 1 und 49 bei Datensatz 2) und VN^a (56 Peaks bei Datensatz 3) erzielt. Dabei ist zu beachten, dass viele gefundene Peaks nicht zwingend einer guten Peakerkennung entsprechen müssen, da auch Peaks an Stellen annotiert werden können, an denen in Wahrheit keines vorhanden ist (falsch positive Peaks). Die Anzahlen unterscheiden sich stark zwischen den drei Datensätzen, wobei Datensatz 2 fast immer die höchsten Maximalwerte erzielt, Datensatz 3 hingegen häufig die höchsten medianen Anzahlen. Der dritte Datensatz ist der einzige, bei dem alle Peakauswahlmethoden in jeder Rohmessung mindestens einen Peak entdecken. Teilweise werden sehr starke Unterschiede zwischen den Rohmessungen festgestellt, beispielsweise findet PDSA im zweiten Datensatz im Median nur sieben Peaks, mindestens einmal gar keine, aber auch in einer Messung 90 Peaks. Im dritten Datensatz sind die Ergebnisse am stabilsten. Minimale und maximale Anzahlen liegen hier am nächsten beieinander und in jeder einzelnen Rohmessung wurden unabhängig von der Peakauswahlmethode mindestens 19 Peaks gefunden.

Im Anschluss an die Anwendung der Peakauswahlalgorithmen findet das Peakclustern statt. Dabei wird festgelegt, welche Single Peaks der einzelnen Rohmessungen eines Datensatzes wahrscheinlich dem gleichen Analyt in der Atemluftmessung entstammen. Diese Peaks werden

Tabelle 5.2: Anzahl der Single Peaks in den einzelnen Rohmessungen der drei Datensätze für die verschiedenen Peakauswahlmethoden. Dargestellt sind die medianen, minimalen und maximalen Anzahlen über die Messungen der Datensätze hinweg. Da VN^m über keinen separaten Peakauswahl-Schritt verfügt, sind für diese Methode keine Werte angegeben.

	LM	PME	PDSA	SGLTR	OPME	VN^a	VN^m
Datensatz 1							
Median	13	14	15	42	31	30	-
Minimum	5	5	3	15	5	0	-
Maximum	34	54	47	115	69	62	-
Datensatz 2							
Median	12	17	7	49	14	10	-
Minimum	5	6	0	15	3	0	-
Maximum	69	69	90	137	61	143	-
Datensatz 3							
Median	29	32	33	45	29	56	-
Minimum	20	23	22	34	19	25	-
Maximum	38	53	46	84	41	93	-

zu einem Cluster zusammengefasst. Allen Rohmessungen, welche einen zugehörigen Single Peak enthalten, wird für diesen Consensus Peak eine Intensität ungleich Null zugeordnet. Umgekehrt wird die Intensität eines Consensus Peaks auf Null gesetzt, wenn in der Rohmessung kein Single Peak detektiert wurde, der diesem Consensus Peak zugeordnet wird.

Analog zu den Single Peaks werden in Tabelle 5.3 die Anzahlen der Consensus Peaks für die verschiedenen Kombinationen aus Peakauswahl- und Clusteralgorithmen dargestellt. Da VN^m die beiden Schritte nicht voneinander trennt, ist diese Methode nicht kombinierbar. Ebenso kann das Clusterverfahren VN^a nicht mit anderen Peakauswahlmethoden kombiniert werden. Die höchsten Anzahlen werden durch das Peakauswahlverfahren VN^a erzielt, wenn auch das zugehörige Clusterverfahren verwendet wird. Von dieser Methode abgesehen erzeugen die Peakcluster-Methoden GS und EM im ersten und dritten Datensatz die höchsten beiden Peakanzahlen, wenn das Peakauswahlverfahren als fest angesehen wird. Je nach Peakauswahlverfahren sind dies etwa 40–140 Consensus Peaks. Im zweiten Datensatz sind es häufig CE und EM (hier liegen die höchsten Anzahlen zwischen 20 und 70 Consensus Peaks). Der (semi-)manuelle Goldstandard ergibt auf den drei Datensätzen 120, 224 und 60 Peaks. Diese stark unterschiedlichen Zahlen können sich dadurch ergeben, dass für die manuelle Analyse spezielle Layer (d.h. Listen mit Peakpositionen, an denen bereits in anderen Anwendungen Peaks gefunden wurden) verwendet werden. Diese liegen für verschiedene Gerätestandorte separat vor und werden über die Zeit weiterentwickelt. Es ist also möglich, dass hier Peaks aus anderen Anwendungen annotiert werden, obwohl sie in diesem Datensatz nicht vorkommen. Insgesamt schwanken die Anzahlen der Consensus Peaks stark, je nach Peakauswahl- und Clusterverfahren.

5.4 Gleichzeitige Analyse aller Algorithmen

Da die Anzahl der gefunden Peaks noch keine Auskunft über die Qualität der Peakerkennung gibt, wird in den folgenden Kapiteln die Güte der nachfolgenden Klassifikationsaufgaben bewertet. Hier werden zunächst alle Algorithmen gleichzeitig betrachtet, anschließend wird die Stabilität des Ergebnisses überprüft, indem die einzelnen Schritte (Klassifikation, Peakauswahl, Peakclustern) getrennt untersucht werden (Kapitel 5.5).

Für jede Kombination aus Peakauswahl, Peakclustern und Klassifikation liegen die AUC-Werte von 50 Mal wiederholten Kreuzvalidierungen vor. Das AUC-Kriterium dient als die entscheidende Maßzahl, um die Ergebnisse zu vergleichen. Um einen Überblick über den

Tabelle 5.3: Anzahl der Consensus Peaks in den drei Datensätzen für die verschiedenen Peakauswahl- und Peakcluster-Methoden. Da in das Programm VN[a] keine externen Single Peaks importiert werden können, ist diese Peakcluster-Methode nicht mit anderen Peakauswahlalgorithmen kombinierbar.

Peakauswahl-Methoden	Clustermethoden					
	GS	DBSCAN	CE	EM	VN[a]	VN[m]
Datensatz 1						
LM	42	25	26	42	-	-
PME	51	26	23	56	-	-
PDSA	43	26	22	44	-	-
SGLTR	138	28	52	142	-	-
OPME	124	66	35	116	-	-
VN[a]	112	29	25	98	239	-
VN[m]	-	-	-	-	-	120
Datensatz 2						
LM	23	16	18	19	-	-
PME	25	14	23	26	-	-
PDSA	17	23	26	27	-	-
SGLTR	46	43	66	67	-	-
OPME	11	16	35	19	-	-
VN[a]	20	30	37	51	239	-
VN[m]	-	-	-	-	-	224
Datensatz 3						
LM	62	40	34	47	-	-
PME	53	23	25	57	-	-
PDSA	62	43	35	56	-	-
SGLTR	75	52	54	59	-	-
OPME	54	42	38	56	-	-
VN[a]	132	56	34	105	265	-
VN[m]	-	-	-	-	-	60

Tabelle 5.4: Lagemaße für die AUC-Werte der jeweils 50 Kreuzvalidierungsiterationen für die drei Datensätze.

Datensatz	Minimum	1. Quartil	Median	Mittelwert	3. Quartil	Maximum
D_1	0.560	0.878	0.933	0.911	0.965	0.997
D_2	0.407	0.780	0.827	0.825	0.874	0.999
D_3	0.328	0.643	0.721	0.711	0.792	0.935

Schwierigkeitsgrad der Klassifikationsaufgaben zu bekommen, sind die Maßzahlen aller Kombinationen, getrennt nach Datensatz, in Tabelle 5.4 zusammengefasst. Es sind die extremsten beobachteten AUC-Werte, die Quartile, der Median und der Mittelwert aller Kombinationen und Wiederholungen dargestellt. Es ist deutlich zu erkennen, dass der erste Datensatz das einfachste, der zweite das mittlere und der dritte Datensatz das schwierigste Klassifikationsproblem darstellt. Die entsprechenden mittleren AUC-Werte liegen jeweils etwas über 0.9, 0.8 und 0.7. Die besten Werte liegen bei allen drei Datensätzen oberhalb von 0.9, die niedrigsten liegen für D2 und D3 unterhalb von 0.5. Dass es zu diesen extremen Ergebnissen kommen kann, ist bei der großen Anzahl an Kombinationen und Wiederholungen nicht überraschend. Das untere Quartil liegt sogar beim schwierigsten Klassifikationsproblem mit 0.643 noch deutlich von 0.5 entfernt, was bedeutet, dass auch schwächere Kombinationen noch teilweise sinnvolle Klassifikationsergebnisse liefern.

In den Abbildungen 5.3 bis 5.5 sind die Klassifikationsergebnisse für alle Kombinationen, getrennt nach Datensatz, als Boxplots detailliert dargestellt. Jede Kombination aus Peakauswahlmethode, Peakcluster-Methode und Klassifikationsalgorithmus ist durch einen Boxplot dargestellt, welcher sich aus den 50 AUC-Werten der Kreuzvalidierungs-Wiederholungen zusammensetzt. Jede der drei Abbildungen trägt auf der y-Achse die AUC-Werte ab. Die Kombinationen sind zunächst in 6 Felder für die Klassifikationsalgorithmen eingeteilt. In jedem Feld sind die Peakauswahlmethoden auf der x-Achse dargestellt und die Peakcluster-Methoden durch Farben gekennzeichnet. Für eine vereinfachte visuelle Einschätzung der Höhe der Boxen über die drei Zeilen von Feldern hinweg, ist für jeden Datensatz eine individuelle horizontale Linie in die Felder eingezeichnet. Diese liegt für D1 bei 0.9, für D2 bei 0.8 und für D3 bei 0.7.

Für den ersten Datensatz (Abbildung 5.3) fällt in Bezug auf die Klassifikationsalgorithmen auf, dass die AUC-Werte für die Boxen der SVMs und die von GBM meist oberhalb von 0.9 liegen, die von kNN und CT häufig unterhalb. Die Boxen des RF liegen alle oberhalb dieser Grenze und weisen im Gegensatz zu den übrigen Klassifikationsverfahren eine deutlich

geringere Varianz auf. Die Güte der Peakauswahlverfahren ist für diesen Datensatz abhängig vom verwendeten Clusteralgorithmus. Kein Peakauswahlalgorithmus erzielt gleichmäßig die besten Ergebnisse. Es fällt hingegen auf, dass die Kombinationen mit den Clusteralgorithmen GS und CE häufig die ungünstigsten für einen Auswahlalgorithmus sind. Der Goldstandard VN^m (in jedem Feld ganz rechts dargestellt) erzielt im Vergleich mit den automatischen Algorithmen mittlere bis sehr gute Ergebnisse.

Für den zweiten Datensatz (Abbildung 5.4) ergibt sich für die Klassifikationsalgorithmen ein ähnliches Bild. Die SVMs schneiden hier ebenso wie CT mittelmäßig ab (die Boxen verteilen sich etwa gleichmäßig um den Wert 0.8), kNN etwas schlechter. Die Boxen von GBM und RF liegen mehrheitlich oberhalb von 0.8. Während für die SVMs und CT kaum Unterschiede zwischen den Peakauswahlalgorithmen bestehen, schwanken die Ergebnisse für KNN stark in Abhängigkeit vom Clusteralgorithmus (GS schneidet am schlechtesten ab). Für GBM und RF stechen insbesondere SGLTR und LM als Peakauswahlmethoden heraus (für SGLTR ist GS dabei eine ungünstige, für LM eine günstige Kombination). Sie erzielen die besten Ergebnisse und übertreffen dabei den manuellen Goldstandard deutlich. Für die anderen Klassifikationsalgorithmen erzielt VN^m sehr gute Ergebnisse im Vergleich mit den automatischen Methoden.

Der dritte Datensatz weist sehr unterschiedliche Klassifikationsergebnisse auf. Die SVMs und GBM erzielen stark variierende Ergebnisse, von denen viele oberhalb, doch auch einige deutlich unterhalb der 0.7-Markierung liegen. Für kNN liegen die meisten Boxen auf oder unterhalb der Markierung. RF erzielt auch hier für die meisten Peakauswahlmethoden AUC-Werte oberhalb von 0.7. Unabhängig vom Klassifikationsalgorithmus schneidet hier der Peakauswahlalgorithmus PME besonders schlecht ab, auch PDSA erzielt oft weniger gute Resultate. SGLTR erzielt häufig die am höchsten liegenden Boxen der AUC-Werte. Der Clusteralgorithmus DBSCAN erzielt (insbesondere für SGLTR) bessere Ergebnisse als die übrigen Clusteralgorithmen, welche sonst keine eindeutigen Unterschiede aufweisen. Der manuelle Goldstandard schneidet im Vergleich mit den automatischen Algorithmen oft gut bis sehr gut ab.

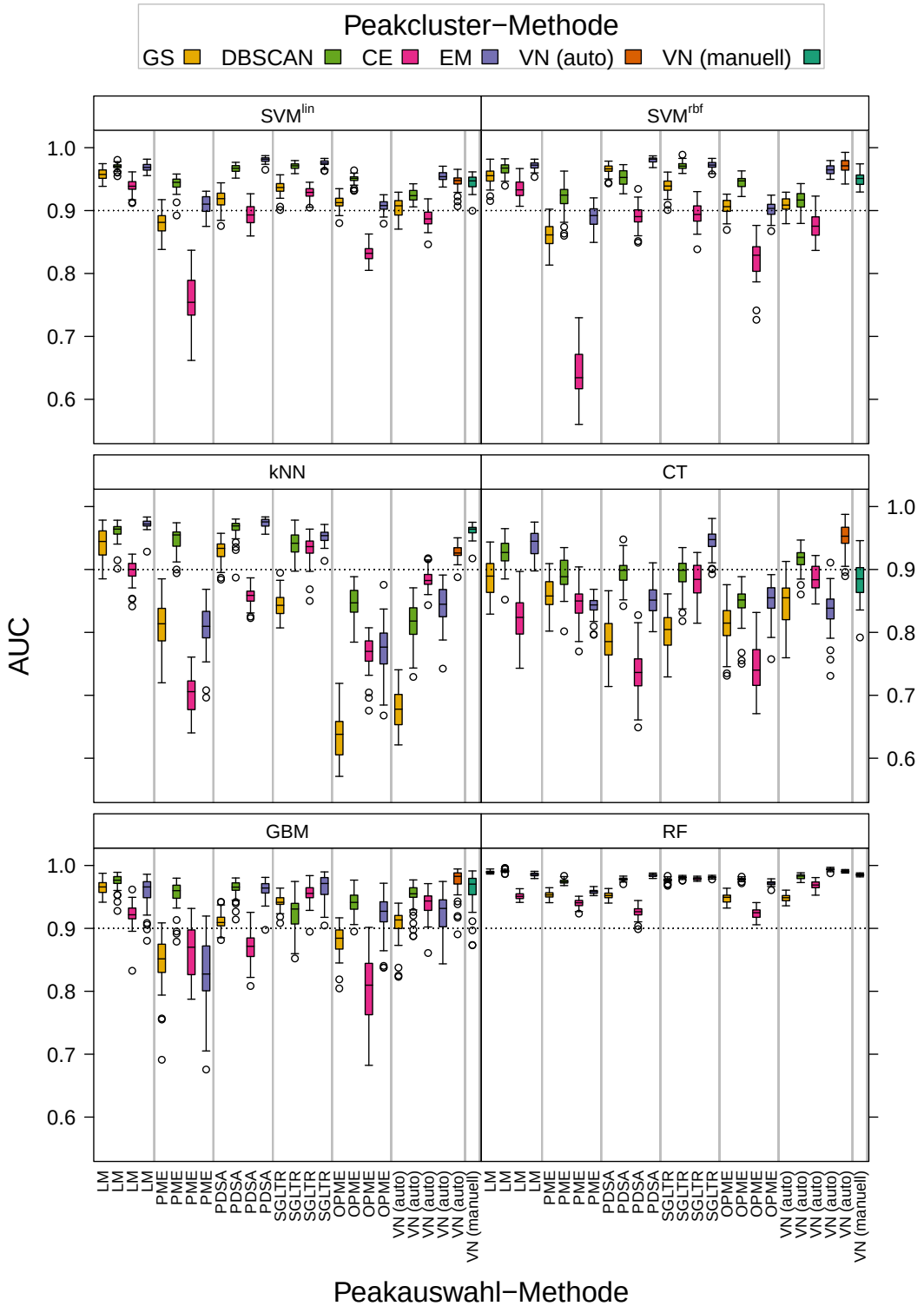


Abbildung 5.3: Boxplots der AUC-Werte für die jeweils 50 Kreuzvalidierungswiederholungen aller möglicher Kombinationen aus Peakauswahl, Peakclustern und Klassifikation für den **ersten Datensatz**.

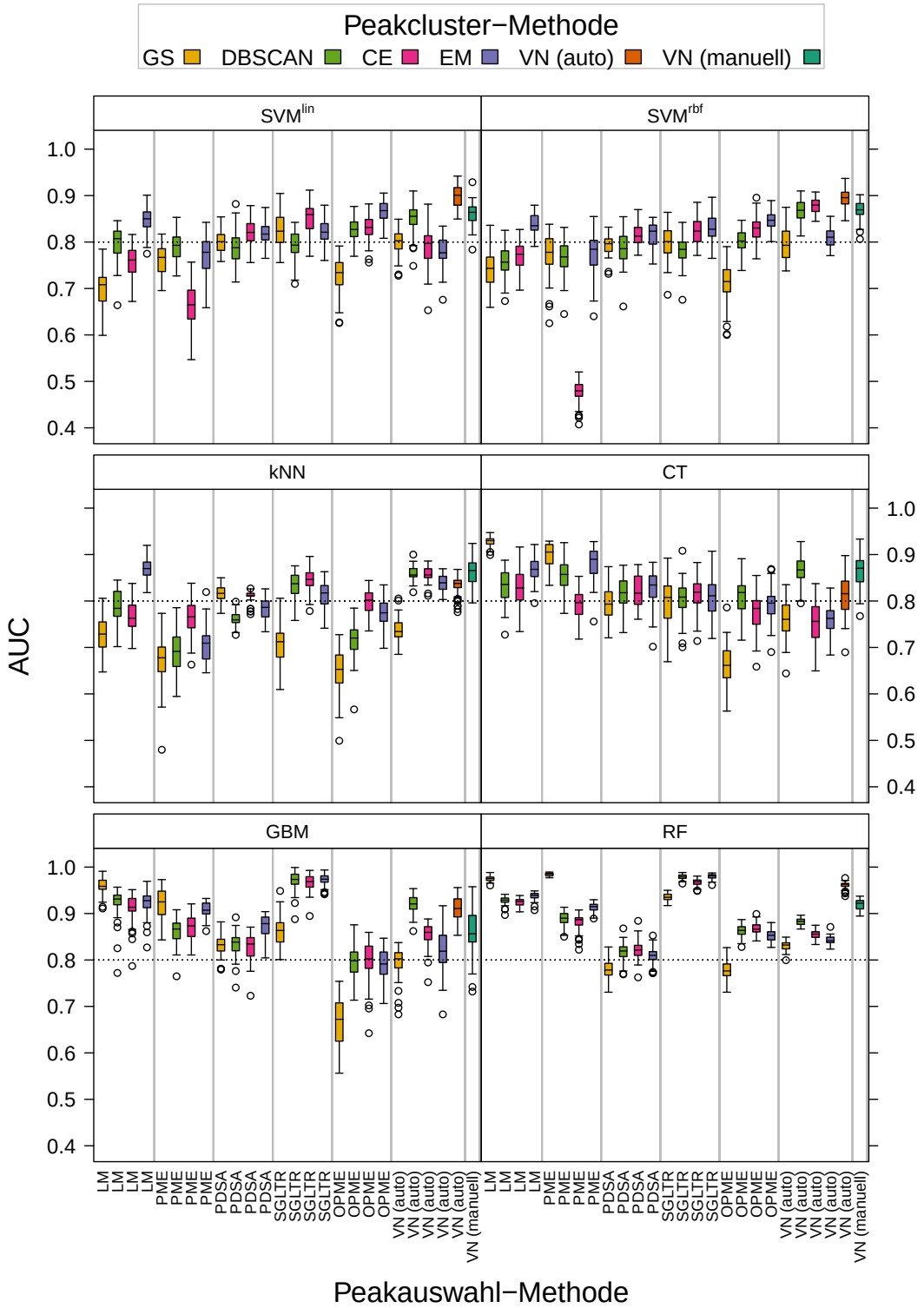


Abbildung 5.4: Boxplots der AUC-Werte für die jeweils 50 Kreuzvalidierungswiederholungen aller möglicher Kombinationen aus Peakauswahl, Peakclustern und Klassifikation für den **zweiten Datensatz**.

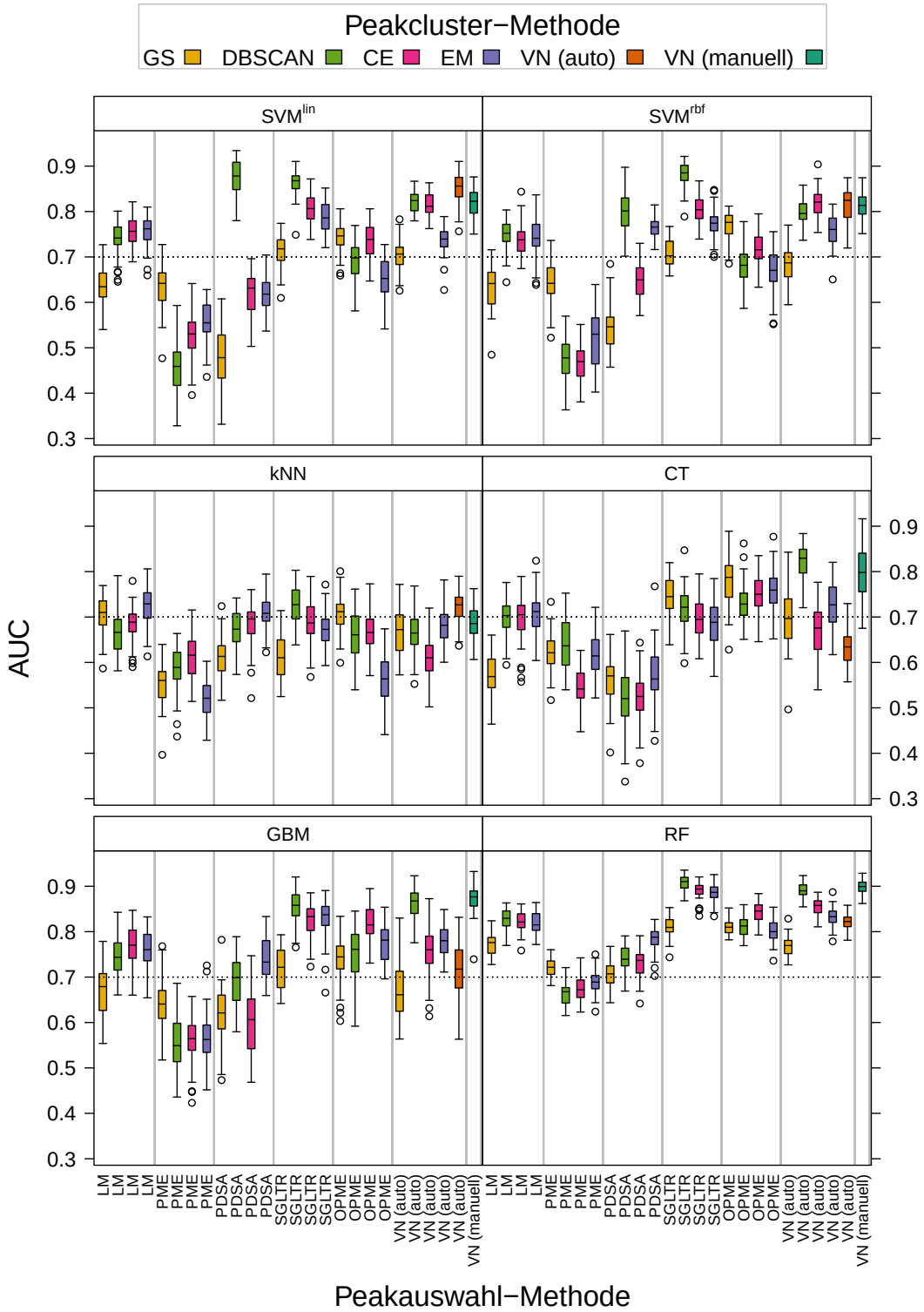


Abbildung 5.5: Boxplots der AUC-Werte für die jeweils 50 Kreuzvalidierungswiederholungen aller möglicher Kombinationen aus Peakauswahl, Peakclustern und Klassifikation für den **dritten Datensatz**.

Um die beste Kombination aus Peakauswahl-, Peakcluster- und Klassifikationsalgorithmus zu finden, werden die mittleren AUC-Werte aller Kombinationen für die drei Datensätze verglichen. Für jeden Datensatz getrennt wird für jede Kombination das arithmetische Mittel der AUC-Werte der Kreuzvalidierungswiederholungen berechnet. Es sei darauf hingewiesen, dass sich die folgenden zusammengefassten Maßzahlen leicht von bisher publizierten Ergebnissen (beispielsweise in Horsch u. a. (2017)) unterscheiden, da in dieser Arbeit stets das arithmetische Mittel über die Kreuzvalidierungswiederholungen berechnet wird, wohingegen in vorigen Publikationen der Median gebildet wurde. Die Interpretation der Ergebnisse ändert sich dadurch jedoch nicht. Für jeden Datensatz werden die 156 Kombinationen absteigend geordnet und jeder Kombination wird der entsprechende Rang zugeordnet (die Kombination mit dem höchsten AUC-Wert erhält den niedrigsten Rang). Anschließend werden die Ränge der drei Datensätze für jede Kombination addiert. Die resultierenden Rangsummen werden ebenfalls absteigend geordnet. Diese Rangsummen geben an, wie gut die Klassifikationsergebnisse der jeweiligen Kombination über die Datensätze hinweg sind. Die 20 Kombinationen mit den niedrigsten Rangsummen sind in Tabelle 5.5 gelistet. Neben der Rangsumme sowie den erzielten Rängen auf den einzelnen Datensätzen ist zusammenfassend der Mittelwert der drei arithmetischen Mittel der Datensätze dargestellt.

Die beste Kombination, mit einer Rangsumme von 15, besteht aus dem Peakauswahlverfahren SGLTR, dem Peakclusteralgorithmus DBSCAN und dem Random Forest als Klassifikationsverfahren. Der zweite und dritte Platz unterscheiden sich nur im Clusteralgorithmus (EM und CE). Die beste Kombination war auf dem dritten Datensatz die beste, auf dem zweiten Datensatz die drittbeste und auf dem ersten Datensatz die elftbeste Kombination, bei jeweils 156 möglichen Kombinationen. Somit erzielt diese Kombination auf allen drei Datensätzen sehr gute Ergebnisse. Über die drei Datensätze wird im Mittel ein AUC-Wert von 0.956 erzielt. Der vierte Platz wird von der manuellen Peakerkennung erreicht, ebenfalls mit RF als Klassifikationsalgorithmus. Sie erzielt auf dem dritten Datensatz den zweiten, auf dem ersten Datensatz den sechsten und auf dem zweiten Datensatz den 20sten Platz. Insgesamt fällt auf, dass in den Top 20 meist RF als Klassifikationsalgorithmus verwendet wird. Bei den Positionen, bei denen GBM enthalten ist, ist die ansonsten gleiche Kombination mit RF immer bereits auf einer höheren Platzierung in der Tabelle enthalten. Nur einmal wird SVM^{rbf} als Klassifikationsalgorithmus verwendet und auch in diesem Fall erzielte die entsprechende Kombination mit RF ein besseres Ergebnis. Bei den Peakauswahlalgorithmen schneidet SGLTR am besten ab, aber auch VN^{m}, VN^{a} und LM sind mehrfach in den Top 20 enthalten. Bei den Peakcluster-Methoden zeichnet sich kein eindeutiges Bild, DBSCAN

Tabelle 5.5: Die 20 besten Kombinationen aus Peakauswahl, Peakclustern und Klassifikation über die drei Datensätze hinweg. Für jeden der drei Datensätze wurden für die Kombinationen die absteigenden Ränge der mittleren AUC-Werte über die 50 Wiederholungen der Kreuzvalidierungen gebildet und für die drei Datensätze aufsummiert. Die Kombinationen sind entsprechend dieser absteigenden Rangsummen (RS) dargestellt. Zusätzlich wurde das arithmetische Mittel der über die CV-Wiederholungen gemittelten AUC-Werte der drei Datensätze gebildet ($\overline{\text{AUC}}$).

Nummer	Peak	Cluster	Klassif.	$\overline{\text{AUC}}$	D_1	D_2	D_3	RS
1	SGLTR	DBSCAN	RF	0.956	11	3	1	15
2	SGLTR	EM	RF	0.949	9	2	5	16
3	SGLTR	CE	RF	0.946	13	7	3	23
4	VN^{m}	VN^{m}	RF	0.934	6	20	2	28
5	LM	DBSCAN	RF	0.915	2	13	17	32
5	VN^{a}	VN^{a}	RF	0.925	3	9	20	32
7	LM	EM	RF	0.913	5	11	23	39
8	VN^{a}	DBSCAN	RF	0.918	8	29	4	41
9	LM	GS	RF	0.912	4	4	46	54
10	SGLTR	EM	GBM	0.922	39	5	16	60
10	SGLTR	GS	RF	0.907	16	12	32	60
12	SGLTR	CE	GBM	0.916	47	8	18	73
13	VN^{a}	EM	RF	0.890	1	60	15	76
14	VN^{a}	VN^{a}	SVM^{rbf}	0.894	27	27	27	81
15	VN^{a}	DBSCAN	GBM	0.911	55	19	10	84
16	OPME	DBSCAN	RF	0.884	14	47	29	90
17	LM	CE	RF	0.899	56	15	21	92
18	LM	DBSCAN	GBM	0.882	18	16	59	93
18	VN^{a}	CE	RF	0.893	30	52	11	93
20	VN^{m}	VN^{m}	GBM	0.898	40	50	8	98

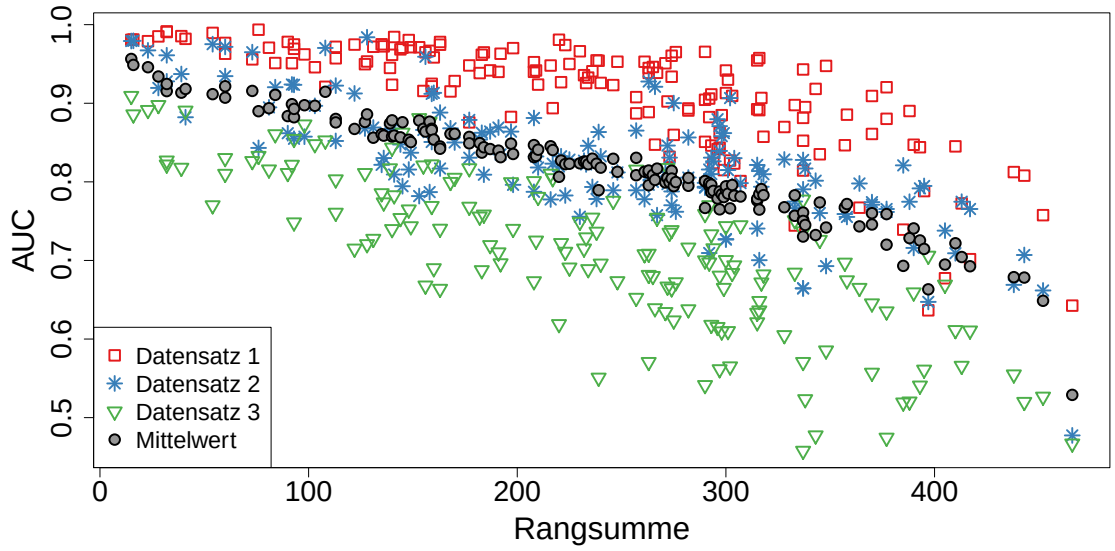


Abbildung 5.6: Zusammenhang zwischen der Rangsumme über die drei Datensätze und dem AUC-Wert. Für jede Kombination aus Peakauswahl-, Peakcluster- und Klassifikationsalgorithmus sind die einzelnen medianen AUC-Werte für die drei Datensätze farbig dargestellt, das arithmetische Mittel der drei medianen AUC-Werte grau. Jede Kombination ist also viermal mit der gleichen Rangsumme in der Grafik enthalten.

wird sechs Mal verwendet, EM und CE jeweils vier Mal, GS sowie die nicht mit anderen Peakauswahlverfahren kombinierbaren Methoden VN^a und VN^m jeweils zwei Mal. Dabei ist jedoch keine der Peakcluster-Methoden den anderen deutlich überlegen.

Der Zusammenhang zwischen der Rangsumme und dem mittleren AUC-Wert der drei einzelnen Datensätze, beziehungsweise dem mittleren AUC-Wert über alle drei Datensätze, ist in Abbildung 5.6 in einem Streudiagramm dargestellt. Jede der 156 Algorithmenkombinationen ist in der Grafik dementsprechend vier Mal (immer mit der gleichen Rangsumme) enthalten. Es ist erkennbar, dass die Rangsumme stark mit den AUC-Werten, insbesondere dem mittleren AUC-Wert über die Datensätze korreliert. Außerdem ist erkennbar, dass beim ersten Datensatz die Unterschiede zwischen den sehr guten Algorithmen nicht sehr groß sind. Beim zweiten Datensatz sind die Unterschiede etwas größer und beim dritten Datensatz zeigen sich auch bei den insgesamt sehr guten Kombinationen schon große Unterschiede. Dass die beste Algorithmenkombination SGLTR, DBSCAN und RF aus Tabelle 5.5 beim zweiten und dritten Datensatz besonders gut und beim ersten Datensatz am schwächsten abgeschnitten hat, ist insgesamt also zum Vorteil dieser Kombination zu werten, da der AUC-Wert des ersten Datensatzes nicht viel niedriger ist als der der besseren Kombinationen

Tabelle 5.6: Mediane AUC-Werte für die **Klassifikationsalgorithmen** und die drei Datensätze über die Kreuzvalidierungswiederholungen der Kombinationen mit den Peakauswahl- und Peakcluster-Methoden.

Datensatz	SVM^{lin}	SVM^{rbf}	kNN	CT	GBM	RF
D_1	0.937	0.935	0.889	0.863	0.939	0.977
D_2	0.808	0.806	0.789	0.819	0.874	0.890
D_3	0.727	0.735	0.664	0.685	0.742	0.808

auf diesem Datensatz. Der mittlere AUC-Wert von 0.965 ist auch der höchste mittlere Wert über alle Kombinationen und stimmt im Ergebnis dementsprechend mit der Rangsumme überein.

5.5 Analyse der einzelnen Schritte

Insgesamt zeigte sich im vorangegangenen Kapitel, dass RF das beste Klassifikationsverfahren zu sein scheint, welches insbesondere mit SGLTR Peakauswahl zu guten Ergebnissen führt. Da bei der sehr großen Anzahl an Kombinationen die konkrete Reihenfolge der Kombinationen wegen zufälliger Schwankungen nicht als stabil angenommen werden kann, werden die einzelnen Schritte (Peakauswahl, Peakclustern, Klassifikation) im folgenden Kapitel noch einmal einzeln untersucht.

5.5.1 Klassifikation

Zunächst werden in Tabelle 5.6 die AUC-Werte der Klassifikationsmethoden für die einzelnen Datensätze betrachtet. Dabei wird der Median über alle Kreuzvalidierungswiederholungen der Peakauswahl- und Peakcluster-Methoden (1300 Werte) gebildet. Für alle drei Datensätze ist RF im Median der beste Klassifikationsalgorithmus, GBM ist stets der zweitbeste. Am schlechtesten schneidet kNN ab und auch CT erzielt keine guten Ergebnisse. Beide Varianten der SVM erzielen im Vergleich mittlere Ergebnisse, die sich im Median nicht stark unterscheiden.

In Abbildung 5.7 sind die arithmetischen Mittel der AUC-Werte für alle Kombinationen aus Peakauswahl, Peakclustern und Klassifikation über die Kreuzvalidierungswiederholungen dargestellt. Die Werte sind die gleichen, welche in Tabelle 5.5 verwendet wurden, um die

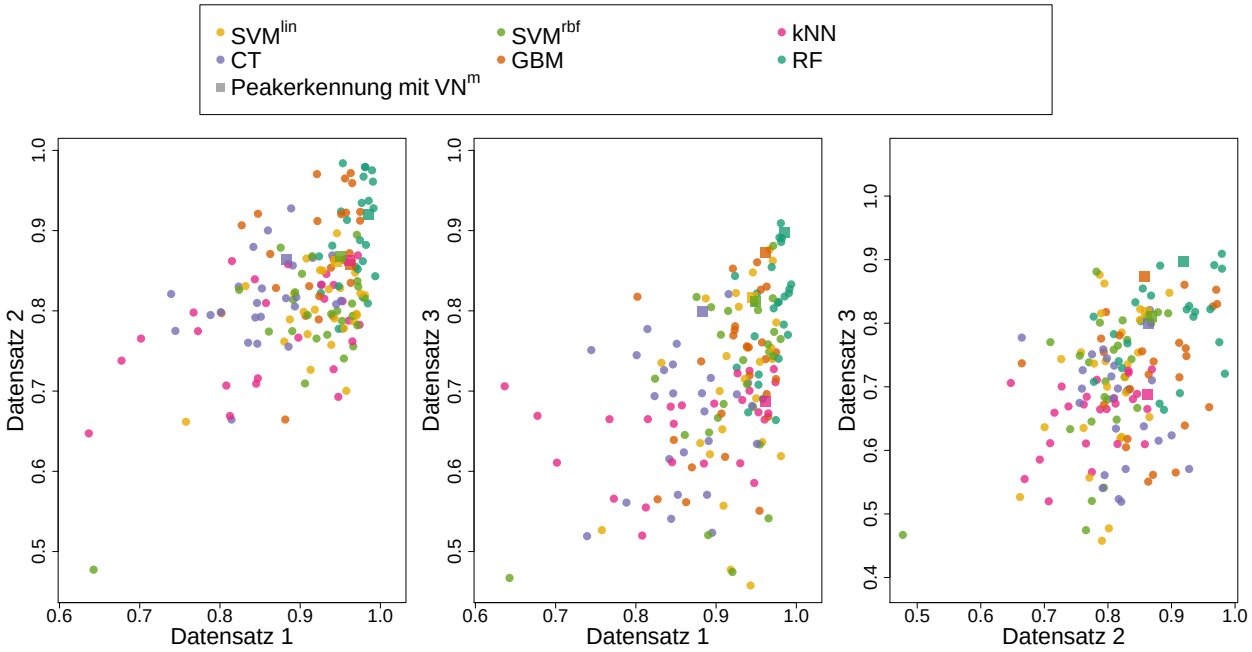


Abbildung 5.7: Paarweise Streudiagramme der mittleren AUC-Werte über die Kreuzvalidierungswiederholungen für jeweils zwei der drei Datensätze. Jeder Punkt steht für eine Kombination aus Peakauswahl-, Peakcluster- und Klassifikationsmethode. Die verschiedenen **Klassifikationsalgorithmen** sind farblich markiert. Die Kombinationen, die den Goldstandard VN^m für die Peakerkennung nutzen, sind durch ein Quadrat gekennzeichnet.

Tabelle 5.7: Mediane AUC-Werte für die **Peakauswahlalgorithmen** und die drei Datensätze über die Kreuzvalidierungswiederholungen der Kombinationen mit den Peakcluster-Methoden und Klassifikationsalgorithmen.

Datensatz	LM	PME	PDSA	SGLTR	OPME	VN^a	VN^m
D_1	0.959	0.880	0.940	0.948	0.891	0.925	0.955
D_2	0.851	0.812	0.811	0.839	0.798	0.843	0.874
D_3	0.731	0.587	0.664	0.777	0.741	0.764	0.828

Ränge innerhalb eines Datensatzes zu bilden. In drei Streudiagrammen sind die Werte für jeweils zwei Datensätze gegeneinander aufgetragen. Dabei sind die Klassifikationsalgorithmen farbig markiert. Punkte, die jeweils oben rechts liegen, gehören zu Kombinationen, die auf beiden Datensätzen gute Ergebnisse erzielen. Ist nur auf einem Datensatz ein gutes Ergebnis erzielt worden, so liegen die Punkte am rechten oder oberen Rand. Die zum Goldstandard VN^m gehörenden Werte sind durch ein Quadrat dargestellt. Auch in dieser Abbildung zeigt sich, dass die besten Ergebnisse von RF und GBM erzielt werden und dass die Punkte von kNN und CT häufig unten links liegen, diese im Vergleich mit den anderen Methoden also auf keinem der dargestellten Datensätze gute Ergebnisse liefern.

Insgesamt ergibt sich auch hier deutlich, dass RF das am besten geeignete Klassifikationsverfahren ist.

5.5.2 Peakauswahl

Die AUC-Werte der Peakauswahlmethoden für die einzelnen Datensätze werden in Tabelle 5.7 betrachtet. Dabei wird der Median über alle Kreuzvalidierungswiederholungen der Peakcluster- und Klassifikationsmethoden (300 Werte für VN^m, 1500 für VN^a, sonst 1200 Werte) gebildet. Für zwei der drei Datensätze erzielt VN^m den höchsten medianen AUC-Wert, für den anderen Datensatz den zweithöchsten. Der (semi-)manuelle Goldstandard ist demnach eine gute Wahl, um Peaks zu detektieren. Da er jedoch nicht mit Clusterverfahren kombiniert werden kann, hat diese Methode den Vorteil, dass ungünstige Kombinationen den Median nicht nach unten verschieben können. Es ist daher möglich, dass bestimmte Kombinationen aus Peakauswahl- und Peakcluster-Methode ebenso gute Ergebnisse erzielen. Wird jedoch über alle möglichen Kombinationen verglichen, so erzielt bei den automatischen Peakauswahlverfahren LM die besten Ergebnisse bezüglich der Rangfolge. LM erzielt auf einem Datensatz das beste, auf einem das zweitbeste und auf dem dritten jedoch nur das fünftbeste Ergebnis. SGLTR,

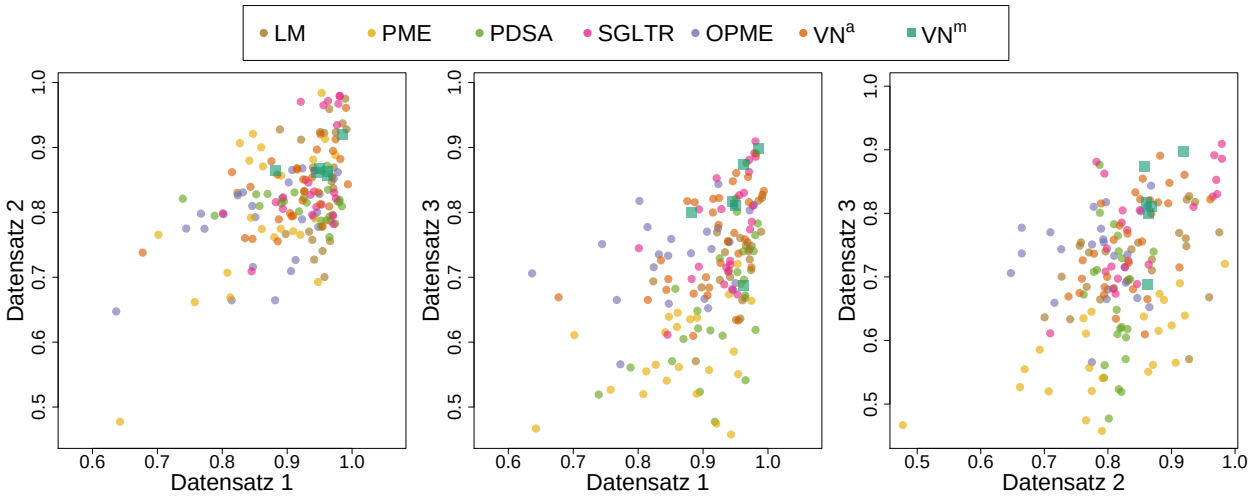


Abbildung 5.8: Paarweise Streudiagramme der mittleren AUC-Werte über die Kreuzvalidierungswiederholungen für jeweils zwei der drei Datensätze. Jeder Punkt steht für eine Kombination aus Peakauswahl-, Peakcluster- und Klassifikationsmethode. Die verschiedenen **Peakauswahlmethoden** sind farblich markiert.

welches in Kapitel 5.4 sehr gut abgeschnitten hat, erzielt hier das zweitbeste automatische Ergebnis, es belegt jeweils auf einem Datensatz den zweiten, den dritten und vierten Rang. Am schlechtesten schneidet PME ab, welches auf zwei von drei Datensätzen den niedrigsten medianen AUC erzielt.

In Abbildung 5.8 sind analog zu Abbildung 5.7 die arithmetischen Mittelwerte der in den Kreuzvalidierungswiederholungen erzielten AUC-Werte jeweils für zwei Datensätze als Streudiagramm dargestellt, jedoch sind hier die Peakauswahlalgorithmen farblich markiert. Auch hier fällt auf, dass PME außer auf dem zweiten Datensatz eher schlechte Ergebnisse erzielt. SGLTR, LM und VN^a hingegen fallen durch gute Ergebnisse auf. VN^m schnitt in Tabelle 5.7 zwar am besten ab, hier zeigt sich jedoch, dass einzelne Kombinationen aus Peakauswahl-, Peakcluster- und Klassifikationsalgorithmen bessere Ergebnisse erzielen.

Da sich im vorigen Unterkapitel sowie in Kapitel 5.4 deutlich herausgestellt hat, dass RF der geeignetste Klassifikationsalgorithmus ist, werden die Ergebnisse aus Tabelle 5.7 noch einmal dargestellt, wobei hier nur der Median über alle Kombinationen gebildet wird, die RF als Klassifikationsalgorithmus verwenden. Die resultierenden medianen AUC-Werte sind in Tabelle 5.8 dargestellt. Die Werte sind durch die Einschränkung auf RF als Klassifikationsalgorithmus deutlich größer als zuvor. VN^m schneidet erneut am besten ab, indem es auf den drei Datensätzen jeweils die Ränge eins bis drei annimmt. Als bestes automatisches

Tabelle 5.8: Mediane AUC-Werte für die **Peakauswahlalgorithmen** und die drei Datensätze bei Anwendung des **Random Forests** über die Kreuzvalidierungsiterationen der Kombinationen mit den Peakcluster-Methoden.

Datensatz	LM	PME	PDSA	SGLTR	OPME	VN^a	VN^m
D_1	0.988	0.956	0.967	0.980	0.962	0.983	0.985
D_2	0.934	0.904	0.809	0.973	0.856	0.856	0.922
D_3	0.811	0.685	0.738	0.889	0.815	0.838	0.899

Peakauswahlverfahren erreicht hier SGLTR die Ränge eins, zwei und vier. LM schneidet als zweitbestes automatisches Verfahren mit den Rängen eins, zwei und fünf ab. PME und PDSA erzielen die schlechtesten Ergebnisse.

5.5.3 Peakclustern

In diesem Unterkapitel werden die verschiedenen Clusteralgorithmen betrachtet. Die Frage, welches der Verfahren besonders geeignet ist, wird hier mit den Peakauswahlalgorithmen verknüpft. Da beide Verarbeitungsschritte eng miteinander verbunden sind, wird die Fragestellung adressiert, welches Clusterverfahren zu welchem Peakauswahlverfahren passt.

Die Güte der Clusterverfahren wird vorab anhand von Abbildung 5.9 nur kurz einzeln betrachtet. Es handelt sich dabei um die gleichen Punkte wie in den Abbildungen 5.7 und 5.8, mit dem Unterschied, dass hier die Clusterverfahren farblich markiert sind. Dabei fällt auf, dass die Ergebnisse der Clusterverfahren sehr unterschiedlich sind. In der linken Grafik ist der obere rechte Rand (wo gute Verfahren liegen) nicht durch bestimmte Clusterverfahren gekennzeichnet. Aus den Abbildung 5.7 ist bereits bekannt, dass diese Punkte zu Kombinationen mit RF oder GBM Klassifikation gehören. Das Clustern scheint dabei keine übergeordnete Rolle zu spielen. Im mittleren (und in geringerer Deutlichkeit im rechten) Bild erzielt neben VN^m DBSCAN die besten Ergebnisse. Insgesamt ergibt sich jedoch kein eindeutiges Bild. Aus diesem Grund werden die Clusterverfahren im Folgenden in Kombination mit den Peakauswahlalgorithmen betrachtet.

In Tabelle 5.9 wird für die drei Datensätze getrennt dargestellt, wie gut die jeweiligen Peakcluster-Methoden in Kombination mit den einzelnen Peakauswahlmethoden funktionieren. Da für jede Kombination sechs verschiedene Klassifikationsmethoden untersucht wurden, werden für jede Klassifikationsmethode die arithmetischen Mittel über die Kreuzvalidierungswiederholungen der Kombinationen aus einer Peakauswahlmethode und allen

Tabelle 5.9: Für alle Kombinationen aus **Peakauswahl- und Clusterverfahren** die mittleren Ränge ($\bar{R}$) der über die Kreuzvalidierungswiederholungen gemittelten AUC-Werte der einzelnen Klassifikationsalgorithmen. Außerdem das arithmetische Mittel dieser mittleren AUC-Werte über die Klassifikationsalgorithmen. Die niedrigsten mittleren Ränge für ein Peakauswahlverfahren sind für die einzelnen Datensätze jeweils grau hinterlegt, um das am besten passende Clusterverfahren zu markieren.

		D_1		D_2		D_3	
Peakausw.	Clusterverf.	AUC	$\bar{R}$	AUC	$\bar{R}$	AUC	$\bar{R}$
LM	GS	0.949	2.667	0.838	2.500	0.663	3.667
	DBSCAN	0.965	1.500	0.837	2.500	0.737	2.333
	CE	0.911	4.000	0.827	3.333	0.744	2.167
	EM	0.966	1.833	0.880	1.667	0.752	1.833
PME	GS	0.869	2.667	0.835	2.000	0.636	1.500
	DBSCAN	0.938	1.000	0.809	2.833	0.562	3.000
	CE	0.791	3.500	0.741	3.333	0.563	3.000
	EM	0.872	2.833	0.825	1.833	0.578	2.500
PDSA	GS	0.911	2.833	0.802	3.000	0.586	3.500
	DBSCAN	0.953	1.833	0.801	3.167	0.718	2.000
	CE	0.863	4.000	0.819	2.000	0.634	3.000
	EM	0.956	1.333	0.821	1.833	0.698	1.500
SGLTR	GS	0.907	3.500	0.822	3.500	0.718	3.500
	DBSCAN	0.946	2.333	0.861	2.833	0.824	1.167
	CE	0.929	3.167	0.878	1.833	0.786	2.333
	EM	0.965	1.000	0.871	1.833	0.772	3.000
OPME	GS	0.850	2.833	0.698	4.000	0.757	1.833
	DBSCAN	0.918	1.167	0.802	2.333	0.722	3.000
	CE	0.815	3.833	0.816	1.667	0.755	1.833
	EM	0.888	2.167	0.820	2.000	0.704	3.333
VN^a	GS	0.865	4.500	0.786	4.500	0.698	4.333
	DBSCAN	0.917	2.833	0.875	1.667	0.810	2.000
	CE	0.907	3.667	0.832	3.167	0.755	3.000
	EM	0.919	2.667	0.808	3.833	0.753	2.833
	VN^a	0.960	1.333	0.885	1.833	0.759	2.833
VN^m	VN^m	0.948	1.000	0.872	1.000	0.814	1.000

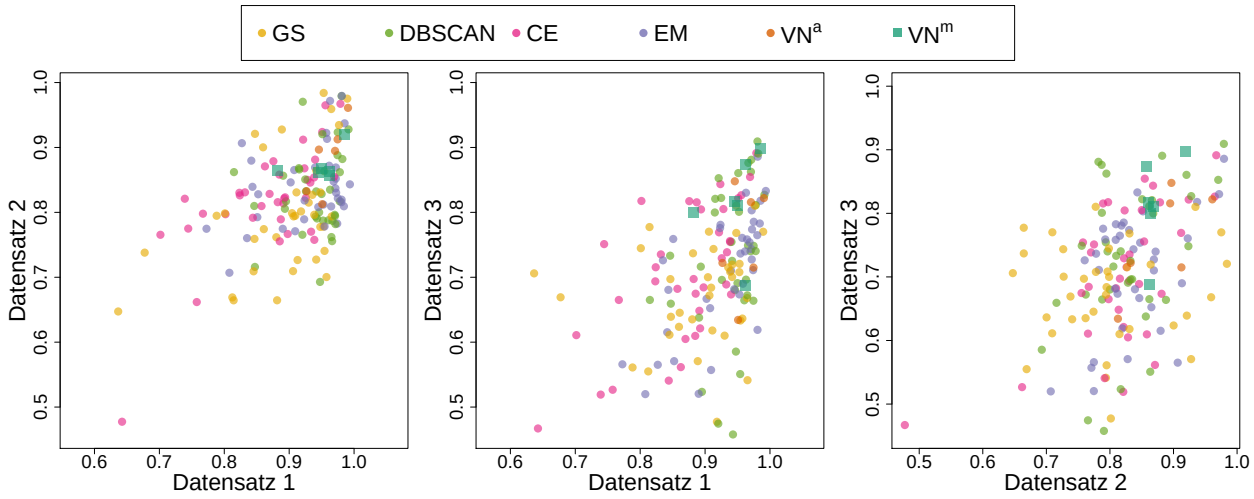


Abbildung 5.9: Paarweise Streudiagramme der mittleren AUC-Werte über die Kreuzvalidierungswiederholungen für jeweils zwei der drei Datensätze. Jeder Punkt steht für eine Kombination aus Peakauswahl-, Peakcluster- und Klassifikationsmethode. Die verschiedenen **Peakclustermethoden** sind farblich markiert.

möglichen Peakcluster-Methoden absteigend geordnet und Ränge gebildet (der größte mittlere AUC-Wert erhält Rang 1). Die Ränge für die Peakcluster-Methoden werden dann über die Klassifikationsmethoden addiert und durch die kleinst-mögliche Rangsumme geteilt.

Beispielhaft wird der Eintrag für LM Peakauswahl mit GS Clustern beim ersten Datensatz erläutert. Für die Peakauswahlmethode LM sind vier Peakcluster-Methoden verfügbar. Die sechs Klassifikationsmethoden werden nacheinander betrachtet. Für jede Klassifikationsmethode werden die Ränge 1–4 vergeben, je nach dem, welche Peakcluster-Methode in Kombination mit LM die höchsten mittleren AUC-Werte erzielt. Die Ränge werden für jede Peakcluster-Methode über die Klassifikationsmethoden addiert. GS erzielt in Kombination mit LM bei SVM^{lin}, SVM^{rbf} und kNN den zweiten, bei CT und GBM den dritten und bei RF den vierten Rang. Als Rangsumme ergibt sich demnach $\text{RS} = 2 + 2 + 2 + 3 + 3 + 4 = 16$. Dies wird in Relation zur minimal möglichen Rangsumme gesetzt, die (bei sechsmal dem ersten Rang) 6 ist. Es wird dementsprechend ausgegeben: $\bar{\text{R}} = \frac{16}{6} \approx 2.667$. Dies entspricht dem mittleren Rang. Für jede Kombination aus Peakauswahl- und Peakclusterverfahren wird außerdem das arithmetische Mittel der sechs Mittelwerte der Klassifikationsmethoden gebildet.

Auf den ersten Blick ergibt sich in Tabelle 5.9 kein für alle Datensätze und Peakauswahlmethoden überlegenes Clusterverfahren. Nur für PDSA ist auf allen drei Datensätzen die gleiche Peakcluster-Methode (EM) am besten geeignet. Darüber hinaus schneidet EM für

die Peakauswahlmethoden LM und SGLTR auf jeweils zwei Datensätzen am besten ab. Für OPME ist CE ebenfalls zweimal die beste Methode und für VN^a erzielt DBSCAN auch zweimal das beste Ergebnis. Für PME erzielt auf jedem Datensatz ein anderes Clusterverfahren den niedrigsten mittleren Rang. Die Auswahl des Clusterverfahrens sollte also passend zur Peakauswahlmethode getroffen werden. Am häufigsten sind EM (acht Mal) und DBSCAN (sechs Mal) die beste Wahl.

Da sich in in den vorherigen Untersuchungen bereits gezeigt hat, dass unter den Klassifikationsmethoden RF in der Regel deutlich besser abschneidet als die anderen Klassifikationsmethoden, sind in Tabelle 5.10 die mittleren AUC-Werte über die Kreuzvalidierungswiederholungen aller Peakauswahl- und Peakcluster-Methoden nur für die Klassifikation mit RF dargestellt. Für jede Peakauswahlmethode ist wieder die Zelle grau hinterlegt, die den mittleren AUC-Wert für das beste zugehörige Clusterverfahren beinhaltet. Beim ersten Datensatz ist für die einzelnen automatischen Peakauswahlmethoden immer entweder DBSCAN oder EM das beste Clusterverfahren. Beim zweiten Datensatz ist das Clusterverfahren stark von den Peakauswahlverfahren abhängig. Ebenso verhält es sich beim dritten Datensatz, wobei drei Mal DBSCAN am besten abschneidet. Das Ergebnis aus Kapitel 5.4, insbesondere aus Tabelle 5.5, dass die Peakauswahl durch SGLTR und das Klassifikationsverfahren RF sehr gute Ergebnisse liefern, wird hier detaillierter verdeutlicht. Während die Peakcluster-Methoden DBSCAN und EM auf den ersten beiden Datensätzen die gleichen mittleren AUC-Werte erreichen (in Tabelle 5.10 auf drei Nachkommastellen gerundet, wohingegen die Ränge in Tabelle 5.5 unterschiedlich sind, da dort nicht gerundet wurde), ist auf dem dritten Datensatz DBSCAN minimal besser. Insgesamt stehen diese Ergebnisse in Einklang mit der Rangfolge der drei besten Kombinationen aus Tabelle 5.5, welche sich nur in der Peakcluster-Methode unterscheiden.

5.6 Empfehlung einer automatischen Algorithmenkombination

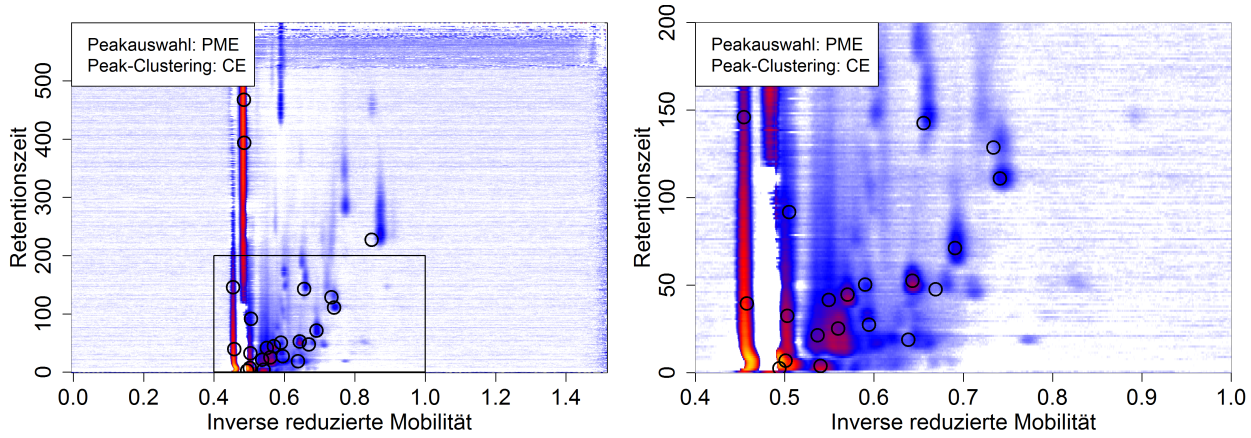
In diesem Kapitel wird das Abschneiden der automatischen Algorithmenkombinationen mit dem Abschneiden des (semi-)manuellen Goldstandards VN^m verglichen und eine automatische Algorithmenkombination für die drei Schritte Peakauswahl, Peakclustern und Klassifikation für zukünftige Anwendungen empfohlen.

Tabelle 5.10: Mittlere AUC-Werte für die Kombinationen aus **Peakauswahl- und Clusterverfahren** über die Kreuzvalidierungswiederholungen des **Random Forests** für die drei Datensätze. Der höchste Wert für jedes Peakauswahlverfahren ist jeweils grau hinterlegt, um das am besten passende Clusterverfahren zu markieren.

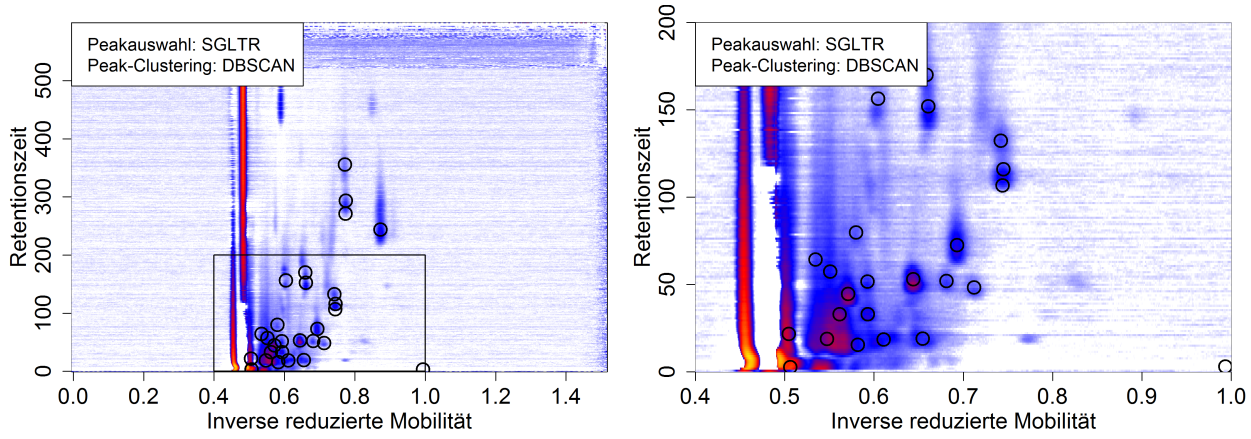
	GS	DBSCAN	CE	EM	VN[a]	VN[m]
Datensatz 1						
LM	0.990	0.991	0.951	0.985	-	-
PME	0.953	0.974	0.940	0.958	-	-
PDSA	0.953	0.978	0.926	0.984	-	-
SGLTR	0.977	0.981	0.979	0.981	-	-
OPME	0.948	0.978	0.924	0.972	-	-
VN[a]	0.948	0.982	0.969	0.994	0.991	-
VN[m]	-	-	-	-	-	0.985
Datensatz 2						
LM	0.975	0.928	0.924	0.937	-	-
PME	0.984	0.888	0.881	0.913	-	-
PDSA	0.778	0.817	0.822	0.809	-	-
SGLTR	0.935	0.979	0.967	0.979	-	-
OPME	0.778	0.862	0.868	0.852	-	-
VN[a]	0.831	0.882	0.855	0.843	0.961	-
VN[m]	-	-	-	-	-	0.920
Datensatz 3						
LM	0.770	0.827	0.822	0.818	-	-
PME	0.721	0.664	0.673	0.690	-	-
PDSA	0.708	0.740	0.729	0.783	-	-
SGLTR	0.810	0.909	0.891	0.886	-	-
OPME	0.810	0.811	0.844	0.803	-	-
VN[a]	0.768	0.891	0.854	0.833	0.822	-
VN[m]	-	-	-	-	-	0.898

Insgesamt wurde in den vorigen Kapiteln gezeigt, dass einige automatische Verfahren im Vergleich mit dem Goldstandard VN^m ähnlich gute und teilweise sogar bessere Ergebnisse erzielen, wenn die Klassifikationsgüte als Merkmal für gute Peakerkennung herangezogen wird. Weitgehend unabhängig von den Peakerkennungsalgorithmen hat RF als Klassifikationsalgorithmus deutlich die besten Resultate erzielt. Aus diesem Grund wird auch für zukünftige Klassifikations-Aufgaben der Random Forest als Klassifikationsalgorithmus empfohlen. Als Peakauswahlalgorithmus erzielten LM und SGLTR die besten Ergebnisse. Über alle möglichen Kombinationen mit Clusterverfahren und Klassifikationsalgorithmen schneidet LM etwas besser ab als SGLTR. Wird die Wahl des Klassifikationsalgorithmus auf RF eingeschränkt, so schneidet SGLTR etwas besser ab als LM. Über alle möglichen Kombinationen aus Peakauswahl, Peakclustern und Klassifikation enthalten die drei besten Kombinationen jeweils RF Klassifikation und SGLTR Peakauswahl. Insgesamt wird aus diesem Grund in folgenden Anwendungen SGLTR verwendet. Auch LM wäre als Alternative für die Peakauswahl denkbar. Die Clusterverfahren wurden hauptsächlich entsprechend ihrer Passung zu den einzelnen Peakauswahlalgorithmen betrachtet. Für LM und SGLTR schneidet EM am besten ab, wenn alle Klassifikationsverfahren einbezogen werden. Eingeschränkt auf RF als Klassifikationsalgorithmus ergibt sich für SGLTR Peakauswahl DBSCAN (welches auch in anderen Kombinationen gute Ergebnisse erzielte) als bestes Clusterverfahren. Auch für LM und RF ergibt sich DBSCAN als beste Kombination (vergleiche Tabelle 5.5). Für die zukünftige Anwendung wird in dieser Arbeit DBSCAN verwendet. Alternativ wäre jedoch auch EM eine gute Wahl als Clusterverfahren zur Kombination mit SGLTR oder LM Peakauswahl.

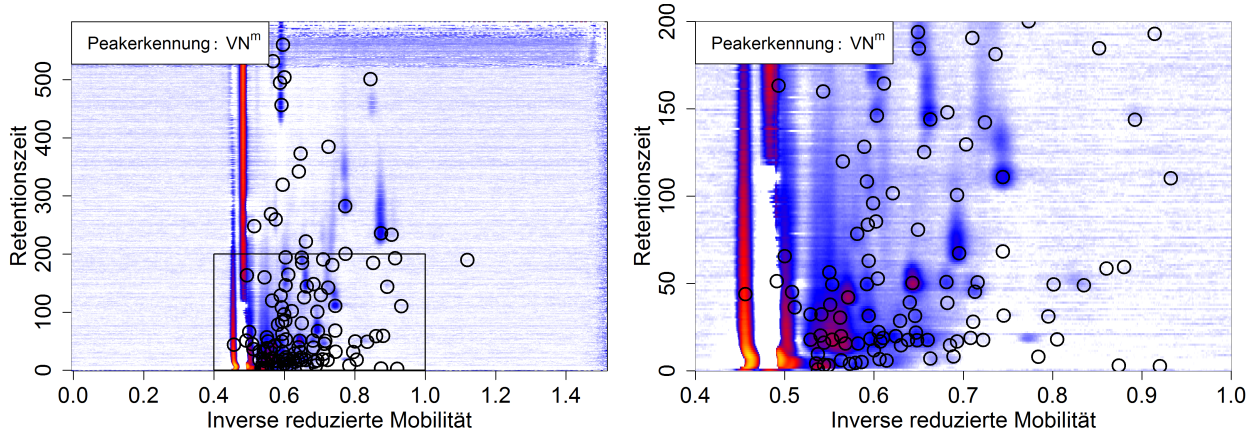
Um einen groben Überblick über die Güte der gewählten Peakerkennung über das Klassifikationsergebnis hinaus zu geben, sind in den Abbildungen 5.10 bis 5.12 die mittleren Rohmessungen der drei Datensätze abgebildet und zum Vergleich verschiedene Consensus Peaks dargestellt. Gezeigt sind jeweils die Peakpositionen einer automatischen Peakerkennung mit weniger guten Klassifikationsergebnissen, die Peakpositionen der empfohlenen automatischen Kombination SGLTR Peakauswahl mit DBSCAN Peakclustern und die Peakpositionen des Goldstandards VN^m jeweils einmal auf der vollständigen Rohmessung und einmal auf einem Ausschnitt, der den Teil der Rohmessung zeigt, auf dem besonders viele Peaks annotiert sind. Die ungünstigen Peakerkennungen sind anhand der Boxplots in den Abbildungen 5.3 bis 5.5 ausgewählt worden, wobei auch das Ziel berücksichtigt wurde, die Positionen verschiedener Methoden zu zeigen.



(a) Ungünstige automatische Kombination: PME Peakauswahl und CE Peakclustern

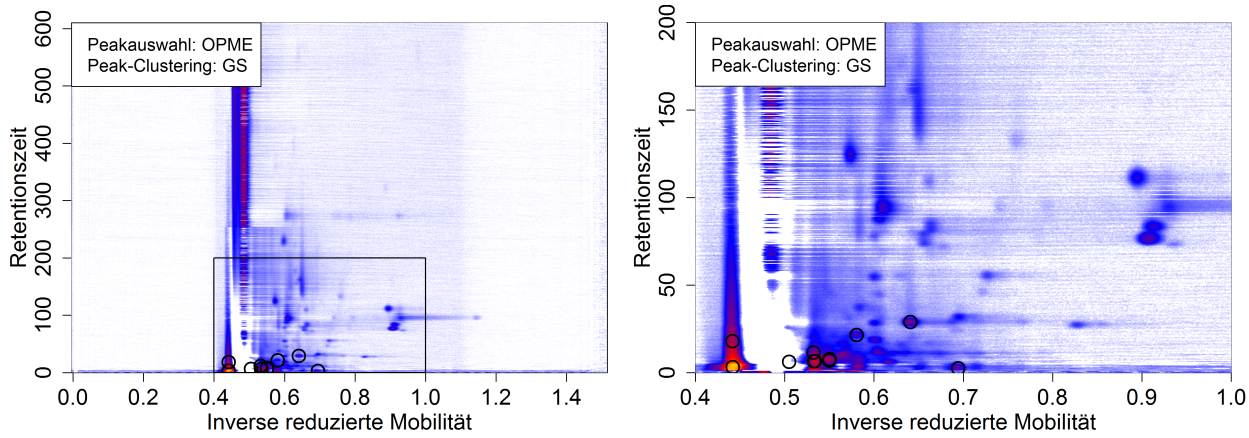


(b) Empfohlene automatische Kombination: SGLTR Peakauswahl und DBSCAN Clustern

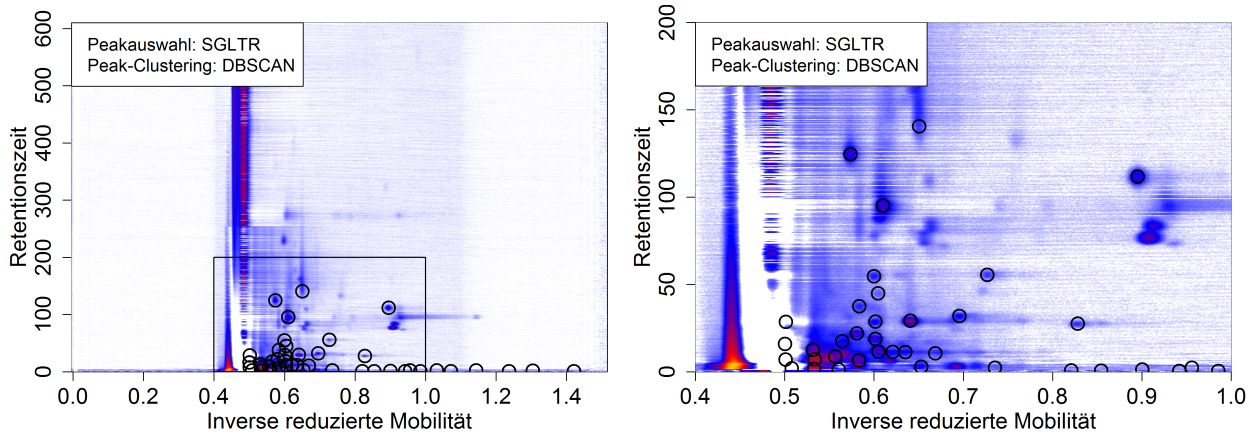


(c) (Semi-)manueller Goldstandard VN^m

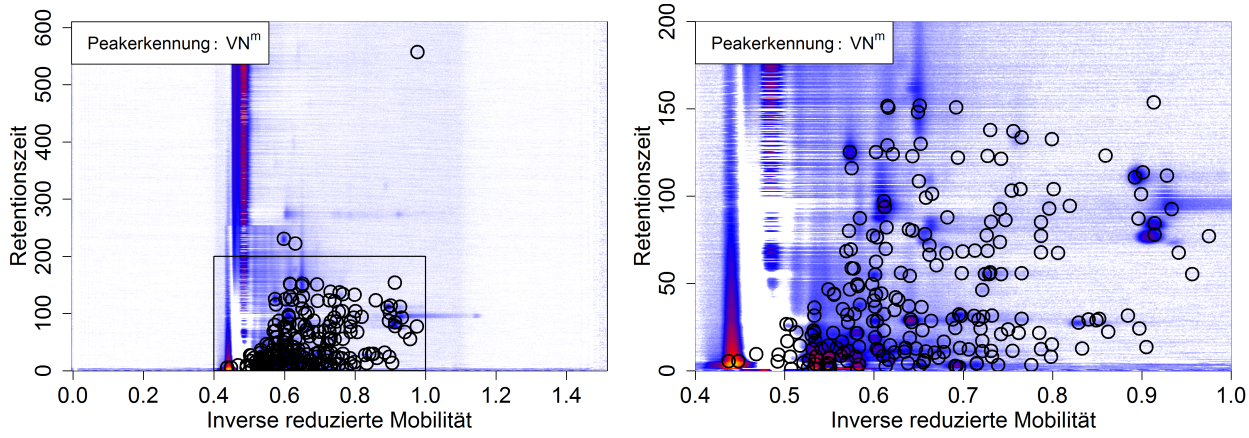
Abbildung 5.10: Peakpositionen der Consensus Peaks einer ungünstigen automatischen Peakerkennung, der empfohlenen automatischen Kombination SGLTR Peakauswahl mit DBSCAN Clustern und des (semi-)manuellen Goldstandards VN^m auf den gemittelten Rohmessungen für den **ersten Datensatz**. Links jeweils die vollständige Rohmessung, rechts ein Ausschnitt.



(a) Ungünstige automatische Kombination: OPME Peakauswahl und GS Peakclustern

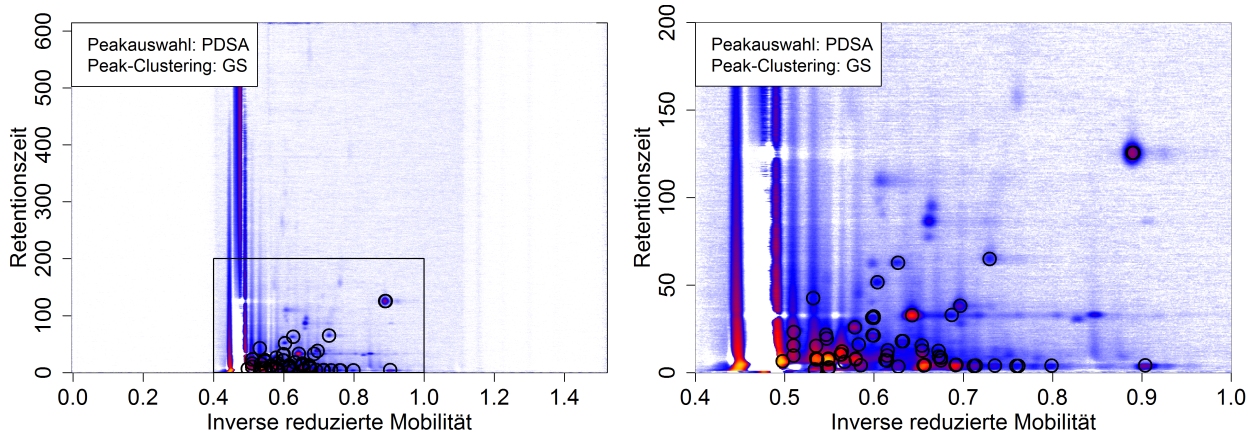


(b) Empfohlene automatische Kombination: SGLTR Peakauswahl und DBSCAN Clustern

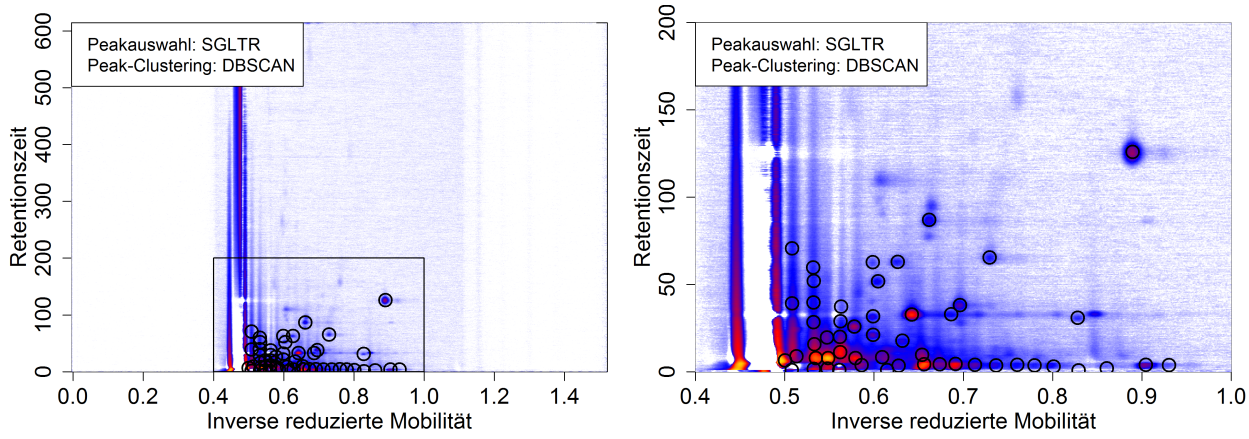


(c) (Semi-)manueller Goldstandard VN^m

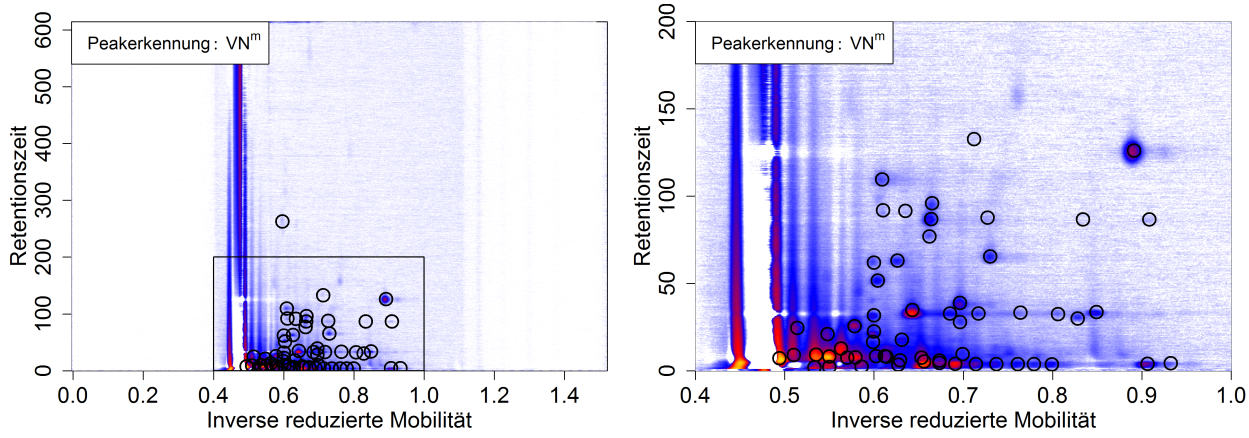
Abbildung 5.11: Peakpositionen der Consensus Peaks einer ungünstigen automatischen Peakerkennung, der empfohlenen automatischen Kombination SGLTR Peakauswahl mit DBSCAN Clustern und des (semi-)manuellen Goldstandards VN^m auf den gemittelten Rohmessungen für den **zweiten Datensatz**. Links jeweils die vollständige Rohmessung, rechts ein Ausschnitt.



(a) Ungünstige automatische Kombination: PDSA Peakauswahl und GS Peakclustern



(b) Empfohlene automatische Kombination: SGLTR Peakauswahl und DBSCAN Clustern



(c) (Semi-)manueller Goldstandard VN^m

Abbildung 5.12: Peakpositionen der Consensus Peaks einer ungünstigen automatischen Peakerkennung, der empfohlenen automatischen Kombination SGLTR Peakauswahl mit DBSCAN Clustern und des (semi-)manuellen Goldstandards VN^m auf den gemittelten Rohmessungen für den **dritten Datensatz**. Links jeweils die vollständige Rohmessung, rechts ein Ausschnitt.

Die hier gezeigten von den automatischen Algorithmen ausgegebenen Peakpositionen der Consensus Peaks werden für die Darstellung in diesem Kapitel korrigiert dargestellt. Die ausgegebenen Peaklisten der Single Peaks einiger automatischer Peakauswahlverfahren enthalten jeweils die Peakpositionen als Indizes der Zeilen und Spalten auf der Rohmessung. Die zusätzlich angegebenen IRM- und RT-Werte liegen in der verwendeten Implementierung nur (annähernd linear) transformiert vor, sodass die eingezeichneten Peaks nicht an der richtigen Stelle der Rohmessung eingezeichnet werden würden. Diese Positionen werden daher korrigiert, indem für jeden Peak die zu den Indizes gehörigen (bekannten) IRM- und RT-Werte aus der Rohmessung abgelesen werden. Die Positionen der Consensus Peaks sind allerdings nicht mehr als Indizes angegeben, da nicht jede Rohmessung die gleichen Pixeleinteilungen aufweist, sondern sie werden von der Implementierung anhand der (nicht korrekt skalierten) Positionen der Single Peaks bestimmt. Die Positionen der Consensus Peaks werden daher für die Abbildung korrigiert, indem eine lineare Regression der neuen Positionen der Single Peaks auf die alten Positionen durchgeführt wird und mit Hilfe dieses Modells die Positionen der Consensus Peaks auf die Skala der korrigierten Single Peaks verschoben werden. Die Korrektur wird für alle Auswahlverfahren durchgeführt, die den Index der Rohmessung angeben, wobei sich nur für SGLTR und PDSA deutliche Unterschiede in den Peakpositionen ergeben. Für die Ergebnisse in Kapitel 6 liegt eine korrigierte SGLTR-Implementierung vor, bei der die ausgegebenen Positionen korrekt bestimmt werden.

Bei der Interpretation der Positionen muss beachtet werden, dass die Darstellung *mittlere* Rohmessungen nach Anwendung eines spaltenweisen Filters zeigt (es wird spaltenweise das 20%-Quantil subtrahiert, um das Rauschen zu reduzieren). Es könnte also sowohl ein Peak, der mit einer hohen Intensität nur in einer Rohmessung vorhanden ist (und damit von den Clusterverfahren möglicherweise nicht berücksichtigt wird) auf dem Bild sichtbar sein als auch ein Peak, der nicht häufig und nur mit niedrigen Intensitäten vorkommt, auf der gemittelten und gefilterten Messung nicht mehr mit bloßem Auge zu erkennen sein. Die Bilder stellen also nur eine grobe Orientierung dar, um einen allgemeinen Eindruck von der Qualität der Peakerkennung zu vermitteln.

Für den ersten Datensatz (Abbildung 5.10) findet die ungünstige Kombination (PME Peakauswahl und CE Peakclustern) 23, die empfohlene automatische Kombination SGLTR Peakauswahl mit DBSCAN Peakclustern 28 und VN^m 120 Consensus Peaks. Beiden automatischen Verfahren entgehen einige deutlich sichtbare Peaks. Dies trifft beispielsweise bei IRM≈ 0.6, RT≈ 450, bei IRM≈ 0.9, RT≈ 450 (jeweils im linken Bild erkennbar) oder bei IRM≈ 0.85, RT≈ 50 und IRM≈ 0.9, RT≈ 150 (jeweils im rechten Bild sichtbar) zu. Diese werden bei der

manuellen Peakerkennung alle annotiert. Die empfohlene automatische Kombination wiederum erkennt einige Peaks, welche die ungünstige Kombination nicht detektiert, beispielsweise bei IRM$\approx$ 0.8, RT$\approx$ 300, bei IRM$\approx$ 0.6, RT$\approx$ 150 oder bei IRM$\approx$ 0.725, RT$\approx$ 50. Umgekehrt ist dies nur selten der Fall, beispielsweise bei IRM$\approx$ 0.55, RT$\approx$ 5. Insgesamt wirkt SGLTR in Kombination mit DBSCAN optisch plausibler als PME in Kombination mit CE, allerdings entdeckt der (semi-)manuelle Goldstandard noch mehr sichtbare Peaks. Obwohl VN^m auch viele Peaks an Stellen detektiert, an denen in der gemittelten Rohmessung keine zu sehen sind, ist er den automatischen Verfahren hier optisch deutlich überlegen.

Für den zweiten Datensatz wird als ungünstige Kombination OPME Peakauswahl mit GS Peakclustern dargestellt. Diese Kombination annotiert auf dem gesamten Datensatz nur 11 Consensus Peaks. In Abbildung 5.11a sind die Positionen auf der mittleren Rohmessung dargestellt. Es ist deutlich ersichtlich, dass diese Kombination eine Vielzahl an Peaks nicht erkennt. Die Kombination aus SGLTR und DBSCAN (Abbildung 5.11b) hingegen deckt mit seinen 43 Consensus Peaks viele der sichtbaren Peaks ab, jedoch sind auch hier noch viele klar erkennbare Peaks nicht detektiert, beispielsweise bei IRM$\approx$ 0.9, RT$\approx$ 75, bei IRM$\approx$ 0.65, RT$\approx$ 75 oder IRM$\approx$ 0.6, RT$\approx$ 250. VN^m annotiert insgesamt 224 Peaks, deckt dabei fast alle in der mittleren Rohmessung sichtbaren Peaks ab, weist jedoch auch an vielen Stellen annotierte Peaks auf, an denen keine sichtbar sind, beispielsweise bei IRM$\approx$ 0.9, RT$\approx$ 150 oder IRM$\approx$ 0.8, RT$\approx$ 125. Insgesamt erscheint die Anzahl gefundener Peaks dabei sehr hoch, teilweise liegen mehrere Peaks sehr dicht beieinander. Insgesamt versagt die ungünstige Kombination bei der automatischen Peakerkennung, sodass überraschend ist, dass bei der Klassifikation im Kombination mit RF noch AUC-Werte von 0.7–0.8 erzielt werden. Im Vergleich mit SGLTR/DBSCAN annotiert VN^m zwar sehr viele Peaks mehr, bei der Klassifikation erringt die Methode dadurch jedoch keinen Vorteil. Möglicherweise wirkt sich die Anzahl (bezüglich der Klasse) uninformativer Peaks negativ auf die Klassifikationsgüte aus.

Für den dritten Datensatz wird in Abbildung 5.12a als ungünstige Kombination PDSA Peakauswahl mit GS Peakclustern dargestellt. Diese Kombination war in der Klassifikation nicht die schlechteste, hat aber im Vergleich mit den anderen Kombinationen nicht gut abgeschnitten. Es fällt auf, dass von den 62 detektierten Peaks mehrfach zwei Peaks sehr dicht nebeneinander liegen, sodass optisch nicht immer auch zwei Peaks erkennbar sind (beispielsweise bei IRM$\approx$ 0.6, RT$\approx$ 30 oder IRM$\approx$ 0.72, RT$\approx$ 5). Dies könnte bedeuten, dass das Clustern möglicherweise einen wahren Peak als zwei detektiert, es ist jedoch auch möglich, dass sich zwei wahre Peaks überlappen und in der Darstellung optisch nicht erkennbar sind.

Insgesamt werden recht viele sichtbare Peaks nicht detektiert, jedoch ist die Peakerkennung hier deutlich besser als bei der ungünstigen Kombination in Datensatz 2. Die empfohlene Kombination SGLTR/DBSCAN erkennt auch hier wieder mehr der fehlenden Peaks der ungünstigen automatischen Kombination (siehe Abbildung 5.12b), wenngleich die (semi-)manuelle Peakerkennung (Abbildung 5.12c) die meisten sichtbaren Peaks annotiert. Hier entdeckt auch VN^m nur 60 Peaks und somit nur acht Peaks mehr als die empfohlene automatische Kombination.

Insgesamt zeigt sich für alle drei Datensätze, dass die vorgeschlagene automatische Kombination im Vergleich zu einer automatischen Kombination mit schlechten Klassifikationsergebnissen auch bei der optischen Beurteilung der Peakerkennung bessere Resultate erzielt. Der Goldstandard jedoch erzielt optisch noch deutlich bessere Ergebnisse bei der Peakerkennung. Dass die Klassifikationsergebnisse dies nicht widerspiegeln, ist möglicherweise auf korrelierte Peaks zurückzuführen, sodass ein nicht detektierter Peak bei der automatischen Peakerkennung durch einen anderen Peak mit ähnlicher Information „ausgeglichen“ werden kann. Möglich ist auch, dass bei der häufig deutlich größeren Anzahl Peaks bei der (semi-)manuellen Peakerkennung mehr (bezüglich der Klasse) uninformative Peaks enthalten sind, welche das Klassifikationsproblem erschweren.

Die Kombination aus SGLTR Peakauswahl, DBSCAN Clustern und RF Klassifikation wird für die folgenden Auswertungen in Kapitel 6 verwendet. Die Kombination aus SGLTR und DBSCAN wird dabei als automatische Alternative zur (semi-)manuellen Auswertung der Rohmessungen verwendet, wobei die beiden Auswertungsansätze (manuell und automatisch) weiterhin verglichen werden.

6 Analyse von Störgrößen

In diesem Kapitel werden mögliche Störgrößen für die Atemluft im Rahmen von MCC-IMS-Messungen untersucht. Zu diesem Zweck wurden neue Daten erhoben. In Kapitel 6.1 wird die Planung dieser Erhebung beschrieben. Die von den teilnehmenden Personen ausgefüllten Fragebögen werden in Kapitel 6.2 deskriptiv ausgewertet. Anschließend werden in Kapitel 6.3 die Ergebnisse der Peakerkennung vorgestellt, wobei neben der (semi-)manuellen Auswertung die in Kapitel 5.6 empfohlene automatische Kombination aus SGLTR Peakauswahl und DBSCAN Peakclustern verwendet wird. Zusätzlich wird die automatische Peakerkennung als Variante mit einem Korrekturschritt bezüglich des Geräte-Effekts betrachtet. Die Auswirkung verschiedener Skalierungen bezüglich der beiden Geräte auf die vorliegenden Daten wird in Kapitel 6.4 analysiert. Die Effekte der Beeinflussung durch ein Nahrungsmittel, die Geräte, das Geschlecht und den Raucherstatus werden in Kapitel 6.5 mit Hilfe univariater Tests untersucht, in Kapitel 6.6 im Rahmen einer Klassifikation. Abschließend wird in Kapitel 6.7 verdeutlicht, welche Risiken die Vernachlässigung des Geräte-Effekts bei der automatischen Peakerkennung für die Klassifikationsergebnisse birgt.

Die Ergebnisse wurden teilweise bereits in Horsch u. a. (2019) veröffentlicht. In dieser Arbeit wird deutlich stärker auf die Datengewinnung (Planung der Studie, Pilotversuch) eingegangen. Der erhobene Datensatz wird hier ausführlich vorgestellt, wobei auch auf nicht verwendete Variablen (Zeitpunkt der letzten Mahlzeit, Getränke, Alter der Personen, Raumluftmessungen) eingegangen wird, die für den Fall erhoben wurden, dass starke Effekte auftreten, die nicht im Fokus der Untersuchung standen. Die Beurteilung der Störgrößen im Rahmen eines Klassifikationsproblems wird hier von der Beeinflussung durch ein Nahrungsmittel auch auf die Variablen Geschlecht, Raucherstatus und Gerät ausgeweitet. Darüber hinaus wird der Effekt einer alternativen Wahl des Wahrscheinlichkeits-Schwellenwerts (als Alternative zum fixen Wert 0.5) untersucht.

6.1 Planung der Studie

Im folgenden Abschnitt werden das Ziel der Studie, ein durchgeführter Pilotversuch, die Erstellung des Fragebogens sowie Details zur Durchführung der Studie erläutert.

6.1.1 Ziel

Das Ziel der durchgeführten Studie war das beispielhafte Untersuchen von möglichen Störgrößen an gesunden Testpersonen, an eigens zu diesem Zweck erhobenen Daten. Einige mögliche Störgrößen werden in Studien zur Atemluft oft nur als Nebenprodukt erhoben. Hier war das Ziel jedoch, die Daten in einem möglichst kurzen Zeitraum zu erheben und bei der Durchführung auf einheitliche Rahmenbedingungen zu achten. Gleichzeitig sollte der Effekt von Störgrößen auf Klassifikationsprobleme untersucht werden, wie sie in klassischen angewandten Fragestellungen zur Diagnose von Krankheiten relevant sind. Um künstlich ein Klassifikationsproblem zu schaffen, wurde jede Person zweifach gemessen, einmal vor dem Konsum eines speziellen Nahrungsmittels und einmal unmittelbar danach.

Der Fokus der Studie lag auf der Untersuchung der Effekte der Nahrungsmittel, des Geschlechts, des Raucherstatus sowie des Geräts auf die Zusammensetzung der Atemluft. Alle diese Effekte wurden für die Atemluftanalyse in der Literatur bereits beschrieben (vergleiche Kapitel 2.3). Für den Fall, dass die Effekte auch in dieser Studie unter Verwendung von MCC-IMS nachgewiesen werden können, sollen die Auswirkungen auf Klassifikationsprobleme diskutiert und Möglichkeiten zur deren Kompensation aufgezeigt werden.

Die Resultate der Studie werden sowohl für die (semi-)manuelle Peakerkennung mit Hilfe der Software VisualNow als auch für die automatische Peakerkennung SGLTR-DBSCAN bestimmt, um beide Methoden gezielt miteinander zu vergleichen.

6.1.2 Pilotversuch

Um den Umgang mit den Gerätschaften zu erproben sowie für die Auswahl des Nahrungsmittels zur Beeinflussung der Atemluft, wurde im Vorfeld eine kleine Pilotstudie durchgeführt (Februar 2017). Als Beispiel aus der Literatur diente Vautz u. a. (2009). Dort wurde an einer Person über einen Zeitraum von insgesamt etwa vier Stunden die Atemluft gemessen. Zehn Minuten bevor ein Bonbon verzehrt wurde und anschließend alle 30 Minuten wurden

die Atemluftmessungen durchgeführt. Nach etwa zwei Stunden wurde außerdem ein Glas Orangensaft konsumiert. In dem Versuch wurden mehrere Metaboliten entdeckt und durch Datenabgleiche zugeordnet (z.B. Eucalyptol, Menthol) oder bereits durch experimentelle Versuche identifiziert (Limonen).

Im durchgeführten Pilotversuch wurde auf mentholhaltige Bonbons verzichtet, da Erfahrungswerte des Labors ergaben, dass die hohen Konzentrationen möglicherweise noch in Folgemessungen detektierbar sein könnten. Aus diesem Grund wurden stattdessen die Effekte von Orangensaft und Schokolade getestet, um anschließend für die größer angelegte Studie das besser geeignete der beiden auszuwählen. Es wurden an einem Tag drei Personen jeweils vier Mal gemessen. Die erste Messung fand ohne gezielte Beeinflussung statt, d.h. ohne Einschränkungen wie Nahrungskarenz an die Testpersonen, außer dem Verzicht auf die getesteten Nahrungsmittel an diesem Tag. Im Abstand von jeweils ca. 40 Minuten wurden Messungen mit den beiden beeinflussenden Nahrungsmitteln (erst Saft, dann Schokolade) durchgeführt, um längerfristige Effekte (z.B. Hunger), zu vermeiden. Am Nachmittag wurde von jeder Person noch eine weitere Messung ohne erneute Beeinflussung, allerdings nach dem Mittagessen, durchgeführt. Diese Messung diente als zusätzliche Kontrollmessung. Der Ablauf der Pilotstudie ist in Tabelle B.1 im Anhang auf Seite 166 dargestellt. Anschließend an jede Atemluftmessung wurden zusätzlich eine Messung der Raumluft sowie eine Messung während des Spülvorgangs mit einem Reinigungsgas durchgeführt, um den Ablauf identisch zur größer angelegten Studie zu gestalten. Die Verarbeitung der drei Teilmessungen benötigt insgesamt etwa 35 Minuten, bevor die nächste Atemluftmessung gestartet werden kann.

Die Auswertung der Rohmessung erfolgte in der Pilotstudie nur mit SGLTR-DBSCAN. Dabei wurden für die Peakerkennung alle Rohmessungen (Atemluft-, Raumluft- und Spülmessungen) verwendet, um den Prozess so stabil wie möglich zu gestalten. Insgesamt detektierte der Algorithmus 31 Peaks. Die Verteilung der Intensitäten wurden anschließend deskriptiv für die drei Gruppen „keine Beeinflussung“, „mit Orangensaft“ und „mit Schokolade“ verglichen.

Für zwei Metaboliten konnten deutliche Unterschiede zwischen den Gruppen beobachtet werden. Diese sind in Abbildung 6.1 dargestellt. Für Metabolit C26 streuen zwar die Beobachtungen ohne Beeinflussung recht stark, jedoch sind die Beobachtungen nach Schokoladenkonsum im Mittel größer als für die Kontrollen. Für die Messungen nach der Aufnahme von Orangensaft sind sogar alle größer als jede Kontroll-Beobachtung. Insbesondere kann beobachtet werden, dass die Messwerte, für jede Person einzeln betrachtet, mit Schokolade oder Orangensaft größer sind als die zugehörigen Kontrollen. Für C29 sind die Beobachtungen sehr ähnlich, allerdings streuen die Kontrollen weniger stark, sodass die Unterschiede

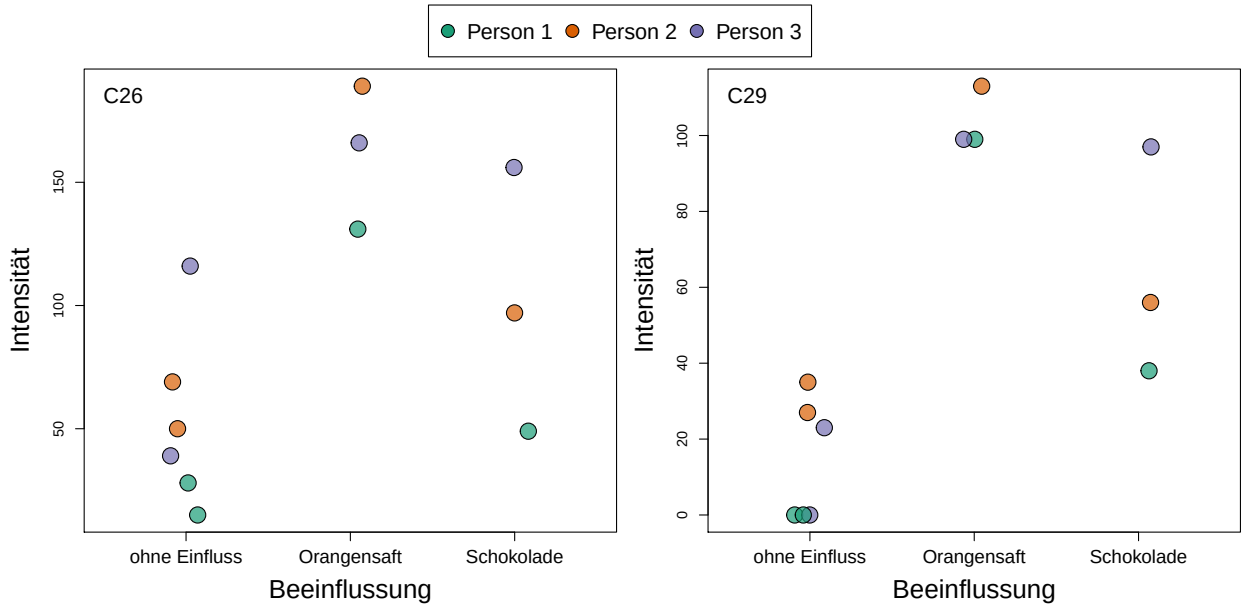


Abbildung 6.1: Ausprägungen der Atemluftmessungen für zwei optisch differenzierende Peaks (C26 und C29).

Tabelle 6.1: Automatisch ermittelte Peakpositionen der zwei deskriptiv ermittelten Kandidaten für relevante Peaks.

Peakname	IRM	RT
C26	0.640	29.841
C29	0.600	29.443

zwischen Kontrollen und beeinflussten Messungen stärker hervortreten. Auch hier sind die Unterschiede zwischen Kontrollen und durch Orangensaft beeinflussten Messungen größer als die zwischen Kontrollen und durch Schokolade beeinflussten Messungen. Aus diesem Grund wurde für die nachfolgende Studie Orangensaft als Nahrungsmittel ausgewählt, um ein Klassifikationsproblem zu schaffen.

In Tabelle 6.1 sind die Positionen der beiden Peaks C26 und C29 aufgeführt. Bei B&S Analytik wurde für beide Peaks ein Datenbank-Abgleich durchgeführt, wonach Eucalyptol und D-Limonen als mögliche Stoffe in Frage kommen. Die zugehörigen Peakpositionen sind in Tabelle 6.2 dargestellt. Die CAS-Nummer (Chemical Abstracts Service Nummer) gibt die Bezeichnung des chemischen Stoffs nach internationalem Standard an. Ohne Kontrollmessun-

Tabelle 6.2: Peakpositionen chemischer Stoffe, die mit C26 und C29 übereinstimmen könnten.

Stoff	CAS-Nummer	IRM	RT
Eucalyptol	470-82-6	0.643	29.200
D-Limonen	5989-27-5	0.594	28.500

gen dieser Stoffe am Gerät können diese Vermutungen allerdings nicht als gesichert gelten. Dass diese Stoffe in der Studie von Vautz u. a. (2009) ebenfalls entdeckt wurden, untermauert jedoch die Vermutung, dass es sich um diese Stoffe handeln könnte.

Für die durchgeführte Studie ist die Fragestellung, anhand welcher Stoffe Orangensaft bzw. Schokolade in der Atemluft gemessen werden können, nicht entscheidend. Da außerdem die Ergebnisse des Datenbankabgleichs mit den Erkenntnissen aus Vautz u. a. (2009) übereinstimmen und sich die Messungen mit Orangensaft in dieser Pilotstudie deutlich von den Kontrollen unterscheiden, wurde auf bestätigende Messungen von reinem D-Limonen oder Eucalyptol verzichtet.

6.1.3 Erstellung des Fragebogens

Zur Gewinnung notwendiger Informationen wurde den Testpersonen ein Fragebogen ausgehändigt, der anschließend beantwortet werden sollte. Gemäß dem Ziel, den Einfluss klinischer Variablen beispielhaft an den Attributen „Geschlecht“ und „Raucherstatus“ zu bewerten, wurden diese für alle Teilnehmenden erhoben. Zusätzlich wurden Variablen erhoben, die möglicherweise zusätzlichen Einfluss haben könnten, insbesondere die Nahrungsaufnahme der letzten Stunden. Diese Variablen dienten nicht der Beantwortung der primären Fragestellungen, sondern zur Identifikation und Einbeziehung möglicher unerwarteter Effekte.

Der Fragebogen untergliederte sich in vier thematische Abschnitte. Im ersten Abschnitt wurden persönliche Angaben, das Geschlecht (männlich/weiblich) sowie das Alter, erfragt. Der zweite Abschnitt befasste sich mit der Ernährung. Es wurden der Zeitpunkt der letzten vollwertigen Mahlzeit sowie generell der letzten Nahrungsaufnahme (bspw. auch Obst oder Joghurt, die nicht als vollwertige Mahlzeit gelten sollten) erfragt. Die wählbaren Zeitintervalle lauteten (wie für die übrigen Fragen dieser Art) „< 30 Minuten“, „30–60 Minuten“, „1–3 Stunden“, „3–6 Stunden“, „6–12 Stunden“ und „> 12 Stunden“. Mit diesen Fragen sollten eventuell auffällig abweichende Ergebnisse (bspw. direkt nach dem Mittagessen) erklärt werden können. Anschließend wurde nach dem Konsum von Zitrusfrüchten in den letzten 24

Stunden gefragt. Wenn die Frage bejaht wurde, folgte eine Frage nach dem Zeitintervall mit den oben aufgeführten Kategorien. Die Teilnehmenden wurden zwar im Vorfeld gebeten, 12 Stunden vor der Messung auf Zitrusfrüchte zu verzichten, jedoch sollte hier erhoben werden, ob eine Testperson dies vergessen hat (oder kurzfristig für nicht Erscheinende eingesprungen ist und daher nicht vorbereitet war). Diese Frage diente somit der Eruierung möglicher Kontamination von Kontrollmessungen mit Stoffen, die erst in der Messung mit Saft auftreten sollten. Die letzte Frage im Ernährungsblock zielte auf konsumierte Getränke innerhalb der letzten Stunde ab. Auswahlmöglichkeiten (Mehrfachauswahl möglich) waren „Leitungswasser“, „Abgefülltes Wasser“, „Saft“, „Limonade/Cola“, „Tee“, „kaffeehaltige Getränke“, „Kakao“, „alkoholische Getränke“, „keine“, sowie eine Freifeldeingabe „Sonstige“. Auch hier sollten Voraussetzungen zur Erklärung unerwarteter Effekte geschaffen werden. Der dritte Fragenabschnitt beinhaltete Fragen zu den Rauchgewohnheiten. Die Fragen zum Zigarettenrauchen entsprechen denen der „Ultrakurzversion zur Erhebung des Aktivrauchens“ der vom Robert-Koch-Institut herausgegebenen Schrift „Erhebung, Quantifizierung und Analyse der Rauchexposition in epidemiologischen Studien“ zur Vereinheitlichung der Erhebung der Rauchexposition in epidemiologischen Studien (Latza u. a., 2005). Sie wurden lediglich um die Frage, wann die letzte Zigarette geraucht wurde (Kategorien wie oben) erweitert. Gemäß dem Fragebogen wurden Zigarettenrauchende so in „Nichtrauchende“, „Ex-Rauchende“ und „Rauchende“ eingeteilt. Zusätzlich wurde eine Frage nach dem Rauchen anderer Tabakwaren als Zigaretten aufgenommen, die mit den Kategorien (Mehrfachauswahl möglich) „E-Zigarette“, „Zigarillo“, „Zigarre“, „Pfeife“, „Wasserpfeife/Shisha“ sowie eine Freifeldangabe „Sonstige“ abgedeckt wurde. Der letzte Abschnitt bestand aus einer Freifeldangabe für Kommentare zum Fragebogen.

Neben dem Fragebogen unterzeichneten alle Testpersonen eine Einwilligungserklärung zur Teilnahme an der Studie, in der unter anderem über das Ziel der Studie informiert wurde und die Person ihr Einverständnis zur Nutzung der Daten für wissenschaftliche Zwecke gab. Als Vorlage wurde die „Musterinformation - Beobachtung“[1] des Forums Österreichischer Ethikkommissionen verwendet und für diese Studie angepasst.

Der Fragebogen sowie die Einwilligungserklärung sind im Anhang auf den Seiten 187ff. bzw. 191f. abgebildet, wie sie den Teilnehmenden ausgehändigt wurden.

[1] Version 1.1 vom 02.6.2016. Abrufbar unter https://www.medunigraz.at/ethikkommission/Forum/index.htm, zuletzt abgerufen am 19.09.2018.

6.1.4 Durchführung der Studie

Für die Studie standen zwei MCC-IMS-Geräte der Firma B&S Analytik vier Tage lang zur Verfügung (02.–05. Mai 2017). Die Messungen fanden in einem Labor der Firma statt, das Ausfüllen der Fragebögen und der Einverständniserklärung sowie der Konsum des Orangensafts in einem benachbarten Büroraum. Jede Messung bestand aus drei Teilmessungen (Atemluft, Raumluft, Spülvorgang), für deren Auswertung das MCC-IMS Gerät insgesamt circa 40 Minuten brauchte. Etwa alle 45 Minuten wurde ein neuer Messvorgang gestartet. Die erste Person wurde täglich um 08:30, die letzte um 15:15 Uhr aufgenommen, am letzten Tag um 16 Uhr. Insgesamt wurde die Atemluft von 49 Personen jeweils zweimal analysiert.

Die Teilnehmenden der Studie waren hauptsächlich wissenschaftliche Mitarbeiterinnen und Mitarbeiter der Fakultät Statistik an der TU Dortmund. Vereinzelt nahmen auch Bekannte von Teilnehmenden oder der Autorin teil. Spontane Ausfälle durch Nichterscheinen einiger Teilnehmenden wurden durch die Rekrutierung von Angestellten der Labore vor Ort kompensiert. Die einzigen Ausschlusskriterien für die Teilnehmenden waren Schwangerschaft (aufgrund von Regularien des Labors), sowie eine Allergie gegen Zitrusfrüchte (da ein Glas Orangensaft getrunken werden musste).

Jede Person füllte zunächst den Fragebogen sowie die Einverständniserklärung zur Teilnahme an der Studie aus. Anschließend wurde die erste Atemluftmessung durchgeführt. Das Startgerät wurde im Vorfeld nach Geschlechtern stratifiziert randomisiert ausgewählt. Bei Ausfällen wurde das Startgerät manuell derart bestimmt, dass jedes Geschlecht möglichst gleich oft auf beiden Geräten zuerst gemessen wurde. Anschließend nahm jede teilnehmende Person ein Glas Orangensaft (0.2 l) zu sich. Direkt im Anschluss wurde die Atemluft am zweiten Gerät gemessen.

Die beiden Geräte werden hier mit „Gerät A“ und „Gerät B“ bezeichnet. Gerät A basiert auf der Software VOCan, Version v2.7, Gerät B auf Version v3.4.1. Beide Geräte wurden mit denselben Einstellungen verwendet. Es wurden stets Atemluftproben mit einem Volumen von 10 ml in die Geräte geleitet. Die OV5-Multikapillarsäule arbeitete bei einer Temperatur von 40 °C. Die IMS-Komponente mit einer Driftlänge von 120 mm arbeitete bei Umgebungsdruck und -temperatur, die Ionengitter-Öffnungszeit betrug 300 µs, die Trägergasflussrate betrug 150 ml/min, die Mess- und Driftgasgeschwindigkeit jeweils 100 ml/min.

Tabelle 6.3: Anzahl an Männern und Frauen je Startgerät.

	männlich	weiblich
Gerät A	13	12
Gerät B	13	11

Beide Geräte wurden im selben Raum verwendet, sodass beide Messungen einer Person bei gleichen Raumbedingungen gemessen wurden. Da das Labor unerwartet teilweise von anderen Mitarbeitern genutzt wurde, können inkonsistente Raumluft-Bedingungen über die vier Tage hinweg jedoch nicht ausgeschlossen werden.

6.2 Deskriptive Auswertung des Fragebogens

Die Fragebögen der 49 Versuchspersonen wurden vollständig ausgefüllt, bis auf eine Raucherin, die ihr Alter beim ersten regelmäßigen Rauchen nicht angab. Insgesamt nahmen an der Studie 26 Männer und 23 Frauen teil. In Tabelle 6.3 ist dargestellt, bei wie vielen Männern und Frauen jeweils Gerät A oder Gerät B als Startgerät verwendet wurde, also auf welchem Gerät die Messung *vor* dem Verzehr des Orangensafts durchgeführt wurde. Durch die Stratifizierung je Geschlecht sind die Geräte je Geschlecht gleichmäßig verteilt, bei den Frauen wurde eine Person mehr zuerst auf Gerät A gemessen als auf Gerät B.

Da die Studienpopulation hauptsächlich aus wissenschaftlichen Mitarbeiterinnen und Mitarbeitern der TU Dortmund besteht, ist die Altersverteilung recht homogen. Die Altersverteilung ist in Abbildung 6.2 als Balkendiagramm dargestellt. Die häufigste Gruppe ist die der 28-Jährigen mit 7 von 49 Personen. Etwa 85% der Teilnehmenden waren zum Zeitpunkt der Studie zwischen 25 und 35 Jahre alt. Die jüngste teilnehmende Person war 23, die älteste 49 Jahre alt.

In Abbildung 6.3 ist dargestellt, wann die Probandinnen und Probanden zum letzten Mal eine vollwertige Mahlzeit und wann sie generell das letzte Mal Nahrung (also auch in geringeren Mengen) zu sich genommen hatten. Am häufigsten wurde 1–3 Stunden vor der Befragung das letzte Mal etwas gegessen bzw. die letzte vollwertige Mahlzeit zu sich genommen (22, bzw. 16 Personen). Die Kategorien „$<$ 0.5 Stunden“, „0.5–1 Stunde“ und „3–6 Stunden“ sind in etwa gleich stark besetzt mit jeweils ca. 8–9 Personen. Nur bei wenigen lag die

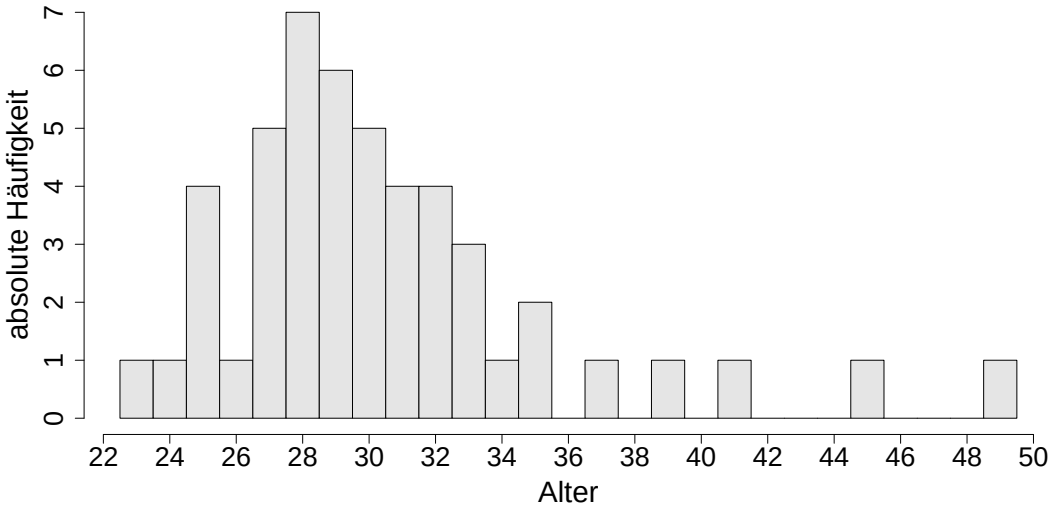


Abbildung 6.2: Verteilung des Alters der Studienteilnehmenden.

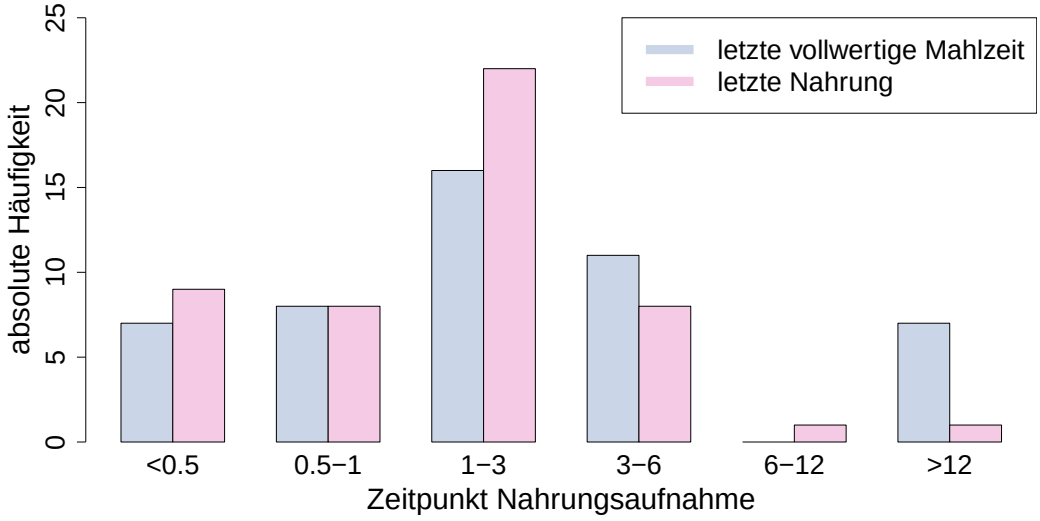


Abbildung 6.3: Zeitpunkte der letzten vollwertigen Mahlzeit (blau) und der letzten Nahrungsaufnahme (rosa) aller Studienteilnehmenden.

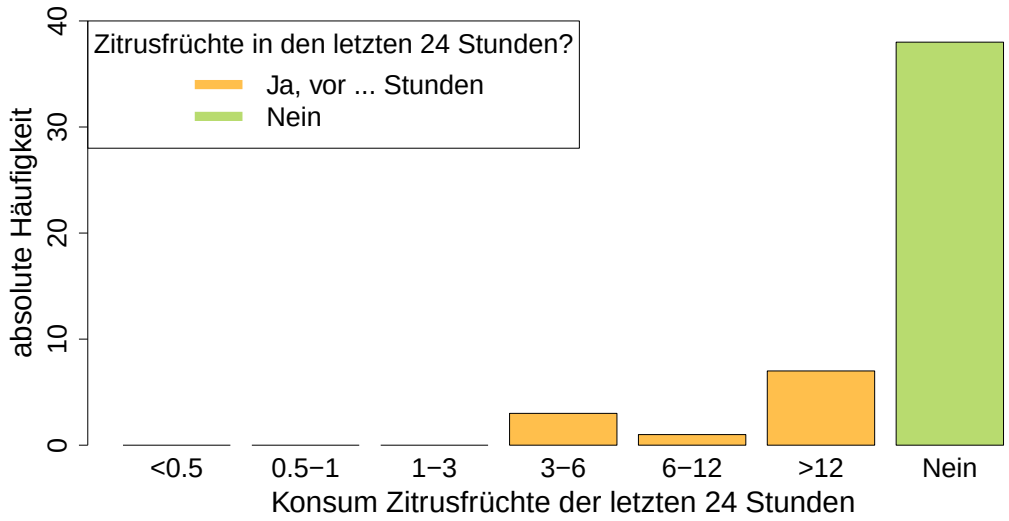


Abbildung 6.4: Anzahl der Studienteilnehmenden, die in den letzten 24 Stunden Zitrusfrüchte konsumiert haben, bzw. die Anzahl je Zeitfenster.

letzte Nahrungsaufnahme mehr als sechs Stunden zurück. Bei sieben Personen lag die letzte vollwertige Mahlzeit jedoch mehr als 12 Stunden zurück. Bei 39 Personen stimmte der Zeitraum der letzten Nahrungsaufnahme mit dem der letzten vollwertigen Mahlzeit überein.

In Abbildung 6.4 ist in Form eines Balkendiagrammes dargestellt, wie viele Teilnehmende innerhalb der letzten 24 Stunden Zitrusfrüchte konsumiert hatten und wenn ja, wann. Die große Mehrheit (38 von 49 Personen) hat bis 24 Stunden vor der Befragung keine Zitrusfrüchte konsumiert, bei sieben Personen lag der letzte Zitrusfruchtkonsum mehr als 12 Stunden zurück. In den drei Stunden vor der Befragung konsumierte niemand Zitrusfrüchte. Drei Personen nahmen in den 3–6 Stunden vor der Studie und eine in den 6–12 Stunden vor der Studie noch Zitrusfrüchte zu sich.

Die Getränke, die die teilnehmenden Personen innerhalb einer Stunde vor der Befragung konsumierten, sind in Abbildung 6.5 mit ihren absoluten Häufigkeiten dargestellt. Das mit 34 Angaben am häufigsten konsumierte Getränk ist Wasser (abgefüllt oder aus der Leitung), gefolgt von kaffeehaltigen Getränken (12 Personen) und Tee (9 Personen). Limonade und Saft wurden von jeweils zwei Personen getrunken, jedoch gab eine Person unter „Sonstiges“ als Getränk „Apfelsaft“ an, was ebenso zu Saft gezählt werden sollte. Niemand hatte Kakao oder alkoholische Getränke konsumiert, unter „Sonstiges“ gab eine Person „Milch“ an. Sieben Personen hatten innerhalb der letzten Stunde keine Getränke zu sich genommen.

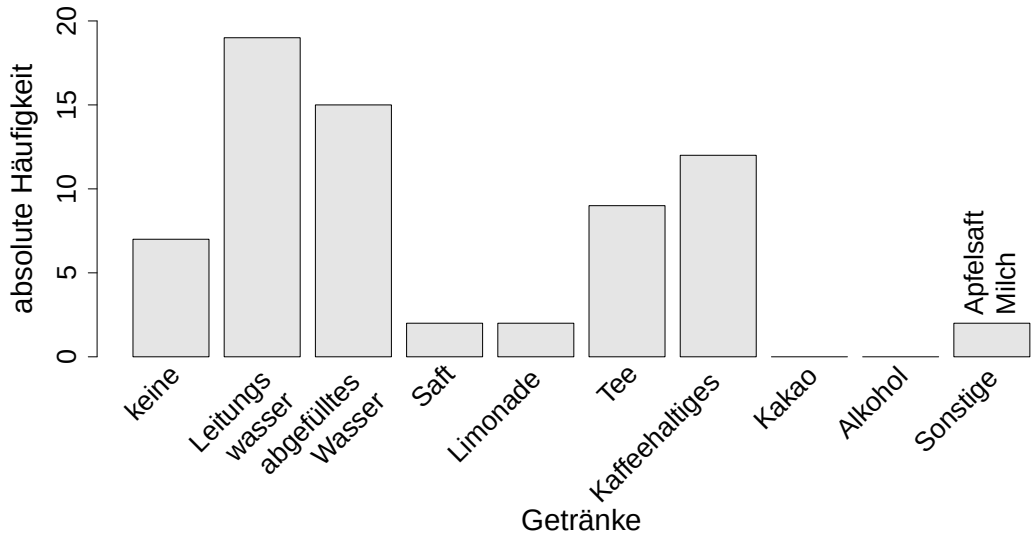


Abbildung 6.5: Anzahl von Personen, die innerhalb der letzten Stunde bestimmte Getränke verzehrt haben.

Abbildung 6.6 enthält drei Grafiken zum Rauchverhalten der Teilnehmenden. Abbildung 6.6A veranschaulicht die absoluten Häufigkeiten für die drei Klassen „Nichtrauchende“, „Ex-Rauchende“ und „Rauchende“. 40 der 49 Teilnehmenden (ca. 80%) sind Nichtrauchende, 5 sind Ex-Rauchende und 4 sind Rauchende. In Abbildung 6.6B ist die Historie der Ex-Rauchenden und Rauchenden abgebildet. Für jede Person ist auf einer Altersskala dargestellt, wann die Person angefangen hat, regelmäßig zu rauchen, für Ex-Rauchende, wann sie aufgehört haben zu rauchen und das aktuelle Alter. Durch Pfeile ist die Zeit, in der eine Person geraucht hat, markiert. Eine Person hat ihr Alter beim Beginn des Rauchens nicht angegeben, dort fehlt entsprechend ein Pfeil. Im Durchschnitt begannen die Personen mit 15 Jahren, zu rauchen. Mit im Durchschnitt 26.5 Jahren haben die Ex-Rauchenden aufgehört zu rauchen. Zum Zeitpunkt der Studie rauchten die Ex-Rauchenden im Schnitt seit 6.2 Jahren nicht mehr. In Abbildung 6.6C ist die Anzahl der Zigaretten pro Tag dargestellt, die Rauchende aktuell rauchen und Ex-Rauchende in der Vergangenheit rauchten. Rauchende und Ex-Rauchende rauchen/rauchten zwischen 0.1 und 50 Zigaretten am Tag, im Mittel etwa 11 Zigaretten. Mit 0.1 Zigaretten am Tag sind die Kriterien für „regelmäßiges Rauchen“ (1 Zigarette pro Tag oder mindestens 5 Zigaretten pro Woche oder mindestens 1 Packung Zigaretten pro Monat für mindestens 6 Monate), wie im Fragebogen angegeben, für eine Person nicht erfüllt, die Angabe wurde hier jedoch beibehalten und nicht auf „Nichtrauchende“ verändert. Die aktiv Rauchenden rauchten ihre letzte Zigarette in der letzten halben Stunde vor der Beantwortung

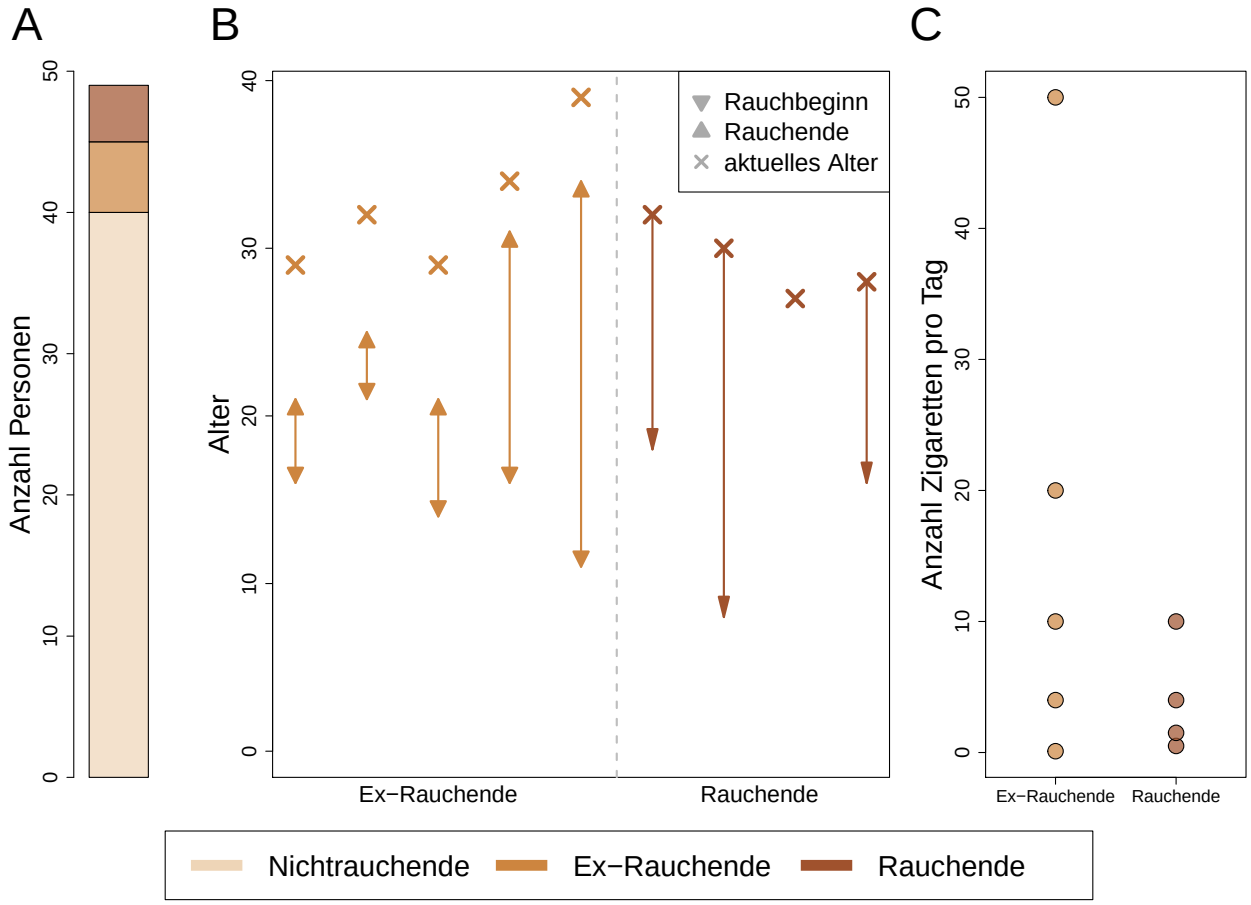


Abbildung 6.6: Rauchverhalten der Testpersonen. A: Anzahlen an Nichtrauchenden, Ex-Rauchenden und Rauchenden in der Studie. B: Zeit, in der Ex-Rauchende und Rauchende regelmäßig geraucht haben. Die vorletzte Raucherin gab ihr Alter beim ersten regelmäßigen Rauchen nicht an. C: Anzahl der Zigaretten pro Tag für Ex-Rauchende und Rauchende.

des Fragebogens (eine Person), 1–3 Stunden vor der Befragung (eine Person) oder mehr als 12 Stunden vor der Befragung (zwei Personen). Keine Testperson rauchte andere Tabakwaren wie beispielsweise Zigarren, E-Zigaretten oder Wasserpfeife.

6.3 Peakdetektion

Im folgenden Abschnitt werden die gemessenen MCC-IMS-Daten genauer betrachtet. Insgesamt werden drei verschiedene Methoden zur Peakerkennung verwendet. Neben der Anwendung des Goldstandards (manuelle Erkennung mit der Software VisualNow) kommt

der vollautomatische Algorithmus SGLTR-DBSCAN zum Einsatz. Dieser wird als dritte Variante mit einem zusätzlichen Schritt zur Kompensation des Geräte-Effekts kombiniert. Die entstehenden Variablen werden durchnummeriert, wobei den Peaknummern der manuellen Peakerkennung der Buchstabe „P“, der automatischen Peakerkennung der Buchstabe „Q“ und der automatischen Peakerkennung mit Zusatzschritt der Buchstabe „R“ vorangestellt wird.

Zusätzlich zu den in der Auswertung im Vordergrund stehenden Messungen der Atemluft werden für die Peakerkennung die Messungen der Raumluft sowie der Spülmessungen verwendet. Dies unterstützt eine möglichst stabile Peakerkennung und erlaubt, bei Bedarf Raumluftmessungen zu einem späteren Zeitpunkt einzubeziehen.

Insgesamt wurden 49 Personen jeweils zweimal gemessen, wobei jede Messung aus drei Einzelmessungen bestand. Die erste Messung betraf die Atemluft, die zweite die Raumluft und die dritte diente als Spülmessung, bevor die Atemluft der nächsten Testperson in das Gerät eingeleitet wurde. Zwischen den Tagen wurden unregelmäßig Spülmessungen durchgeführt (in der ersten Nacht durchgängig, an den übrigen Tagen nur vereinzelt morgens), die zusätzlich in den Datensatz aufgenommen wurden. Insgesamt liegen 345 MCC-IMS-Messungen vor, davon sind 98 Atemluftmessungen, 98 Raumluftmessungen, 98 Spülmessungen im Anschluss an die Raumluftmessungen und 51 Messungen außerhalb des Messplans, von denen 31 sogenannte „feuchte Nullmessungen“ sind (entspricht den Spülmessungen nach der Raumluft) und 20 sogenannte „trockene Nullmessungen“, welche nur das Trägergas enthalten und ebenfalls der Reinigung des Gerätes dienen.

6.3.1 Manuelle Peakerkennung

Die manuelle Peakerkennung aller 345 Rohmessungen wurde von B&S Analytik in Dortmund durchgeführt. Insgesamt wurden dabei 124 Peaks gefunden. Bei dieser Art der Peakerkennung wird an jeder Stelle, an der in einer Messung ein Peak detektiert wurde, automatisch auch in allen anderen Messungen an dieser Stelle eine Intensität für diesen Peak bestimmt, indem das Maximum aus einem definierten Umkreis um die Peakposition ermittelt wird. Dies geschieht unabhängig davon, ob visuell ein Peak in dieser Messung erkennbar ist, sodass nur wenige Werte exakt Null sind (insgesamt 30 Werte, keiner davon auf Atemluftmessungen). Durch interne Verarbeitungsschritte der MCC-IMS-Geräte gibt es allerdings negative Intensitäten in den Rohmessungen, die durch das beschriebene Vorgehen auch als Realisationen der Peaks im manuellen Datensatz auftreten. Dies betrifft insgesamt 725 Werte, davon 172 auf

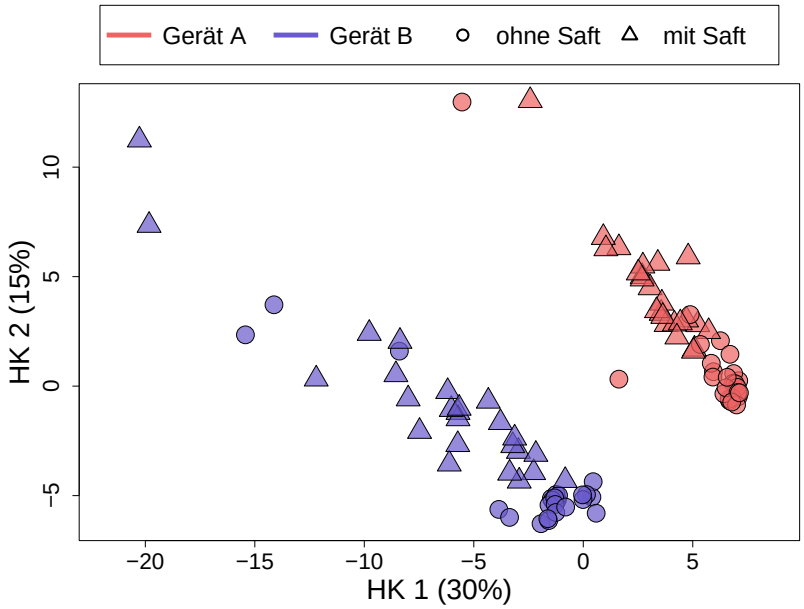


Abbildung 6.7: Die ersten beiden Hauptkomponenten der metabolischen Variablen aus den Atemluftdaten der **manuellen** Peakerkennung. Die Farben geben an, auf welchem Gerät die jeweilige Messung durchgeführt wurde (rot: Gerät A, blau: Gerät B), die Symbole, ob es sich um eine Messung mit oder ohne zuvor konsumierten Orangensaft handelt (Kreis: ohne Orangensaft, Dreieck: mit Orangensaft)

Atemluftmessungen. Da bei dieser Methode also kein Schwellenwert verwendet wird, um Werte unterhalb eines Rauschens auf Null zu setzen, kann hier nicht wie bei der automatischen Peakerkennung verglichen werden, für wie viele Beobachtungen die Peaks gefunden wurden.

Für einen deskriptiven Überblick über die manuell ausgewerteten Atemluftmessungen (Raum- und Spülluft werden hier nicht betrachtet) ist in Abbildung 6.7 der Individuenplot für die ersten beiden Hauptkomponenten (HK) abgebildet. Dabei sind die Messungen der beiden Geräte durch Farben, die Messungen des künstlich geschaffenen Haupteffekts durch Symbole gekennzeichnet. Es zeigen sich deutliche Unterschiede zwischen Beobachtungen auf den beiden Geräten sowie Messungen mit oder ohne vorigen Saftkonsum. Die Messungen der beiden Geräte verlaufen hier parallel zueinander, jedoch nicht parallel zu einer der Hauptkomponenten. Die erste Hauptkomponente enthält Informationen über das verwendete Gerät. Je Gerät enthalten beide Hauptkomponenten Informationen über den Saft-Effekt (niedrige Werte für die erste und hohe Werte für die zweite Hauptkomponente sprechen jeweils eher für Messungen

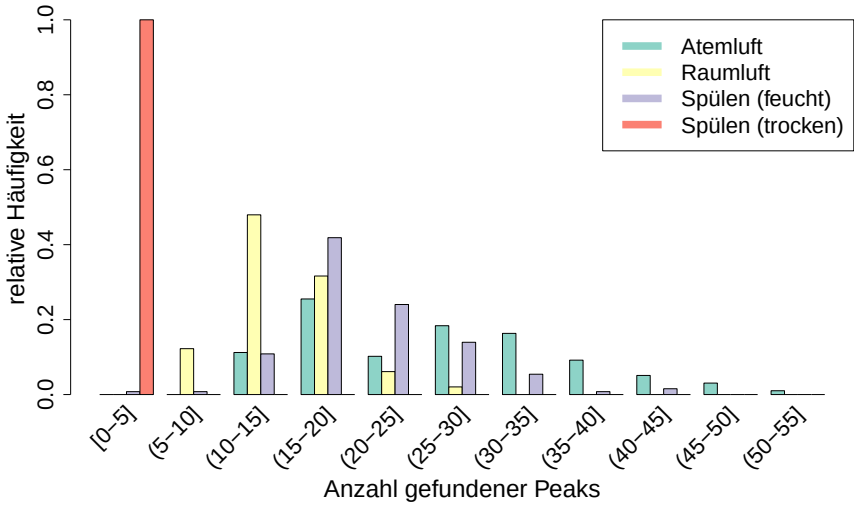


Abbildung 6.8: Anzahl gefundener Single Peaks in den Atemluft-Rohmessungen unter Verwendung des SGLTR-Algorithmus.

mit Saft). Es gibt demnach große Unterschiede zwischen den Geräten. Insgesamt erklären die beiden Hauptkomponenten etwa 45% der Variabilität in den Daten, davon entfallen 30 Prozentpunkte auf die erste Hauptkomponente und 15 auf die zweite.

6.3.2 Automatische Peakerkennung

In diesem Abschnitt werden die Ergebnisse der Anwendung der automatischen Peakerkennung, bestehend aus Peakauswahl (durch den Algorithmus SGLTR) und Peakclustern (durch DBSCAN) dargestellt. Summiert über alle 345 Messungen wurden nach Anwendung von SGLTR insgesamt 6811 Single Peaks gefunden, dies entspricht im Durchschnitt etwa 20 je Messung. Abbildung 6.8 zeigt, dass die Art der Messung dabei eine große Rolle spielt. In den trockenen Nullmessungen wurde nur in einer Messung ein Peak gefunden, sonst überhaupt keine. In den Raumluftmessungen wurden deutlich mehr Peaks gefunden (etwa die Hälfte der Messungen enthält 10–15 Peaks, der Median beträgt 14) und in den feuchten Spülmessungen noch etwas mehr (ca. 40% enthalten 15–20 Peaks, der Median beträgt 20). Die Anzahl der Single Peaks in den Atemluftmessungen streut am stärksten, bei insgesamt größeren Anzahlen. Es wurden dort zwischen 12 und 55 Peaks detektiert, der Median beträgt 27. In der Atemluft werden im Schnitt also mehr Peaks gefunden als in den übrigen Luftproben.

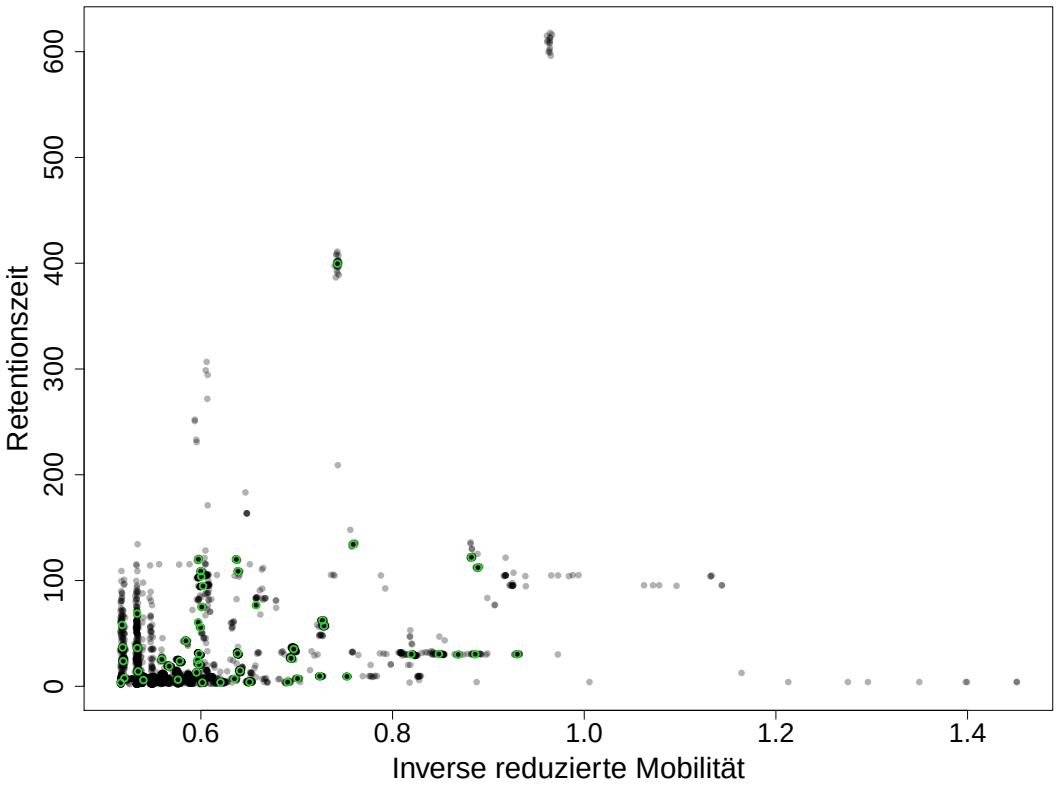
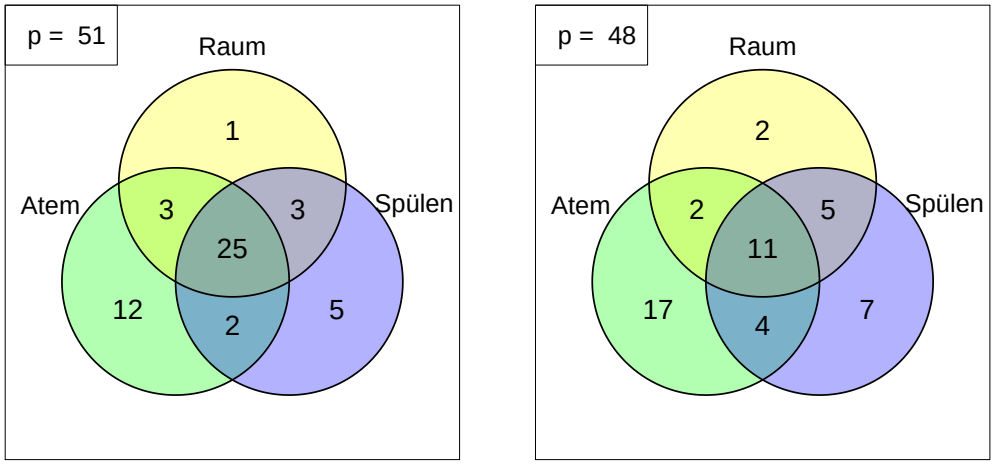


Abbildung 6.9: Alle gefundenen Single Peaks (graue Punkte) durch SGLTR und die aus der nachgeschalteten Anwendung von DBSCAN resultierenden Positionen der Consensus Peaks (grüne Kreise).

Im Anschluss an die Erkennung der Single Peaks wird der Clusteralgorithmus DBSCAN angewendet, um die Single Peaks über mehrere Messungen hinweg zu einheitlichen Einflussgrößen zusammenzufassen. Insgesamt entstehen dabei 51 Consensus Peaks, also weniger als halb so viele wie bei der manuellen Peakerkennung. In Abbildung 6.9 ist dargestellt, an welchen Positionen auf den entsprechenden Rohmessungen Single Peaks gefunden wurden (alle 6811 Single Peaks sind als graue Punkte dargestellt) und an welchen Stellen nach dem Clustern Consensus Peaks entstehen (grüne Kreise). Es ist gut zu erkennen, dass an den meisten Stellen, an denen sich viele graue Kreise überlappen, nach dem Cluster-Schritt ein Consensus Peak entstanden ist. An einigen Stellen fehlen nach visuellem Eindruck jedoch Consensus Peaks, z.B. zwei Peaks bei IRM ≈ 0.95, RT ≈ 100 oder ein Peak bei RT ≈ 600. Im ersten Fall ist die eingestellte Zahl von 10 Peaks pro Cluster nicht erreicht, sodass diese Single Peaks übergangen werden, im zweiten Fall könnten die Peaks über eine zu lange Strecke verteilt sein, um zusammengeordnet zu werden. Die Herausforderung für den Clusteralgorithmus wird besonders für niedrige IRM- und RT-Werte deutlich. Mit bloßem Auge sind getrennte



(a) Die Metaboliten haben für mindestens *eine* Messung Werte ungleich Null (trifft auf 51 Messungen zu).

(b) Die Metaboliten haben für mehr als 10% der Messungen Werte ungleich Null (trifft auf 48 Messungen zu).

Abbildung 6.10: Venn-Diagramm der Anzahlen an Metaboliten, die in den drei Messarten Atemluft, Raumluft und Spülluft (feucht) vorkommen.

Peaks kaum auszumachen, in diesem Bereich liegen viele Single Peaks derart gestreut, dass sie sich nicht klar voneinander abgrenzen. Hier ist es möglich, dass große Bereiche zu einem Peak zusammengefasst werden. Dies kann hier jedoch nicht nachvollzogen werden, da der vorliegende Algorithmus nicht ausgibt, welche Single Peaks welchem Consensus Peak zugeordnet werden.

Es gilt zu beachten, dass der Clusteralgorithmus die Peakintensität eines Consensus Peaks für eine Beobachtung auf Null setzt, wenn in der entsprechenden Rohmessung kein Single Peak detektiert wurde, das anschließend diesem Consensus Peak zugeordnet wurde. Dementsprechend lässt sich für jede Beobachtung die Frage, ob ein Metabolit in der Luft enthalten ist, für den Algorithmus eindeutig beantworten. Im Gegensatz dazu wird bei der manuellen Peakerkennung für eine festgelegte Peakposition in jeder Rohmessung derjenige Wert als Intensität festgelegt, der das Maximum in einer festgelegten Umgebung darstellt, unabhängig davon, ob visuell ein Peak erkennbar ist.

Die eindeutige Beantwortung der Frage, ob ein Peak in einer Rohmessung enthalten ist, ermöglicht für den automatischen Algorithmus einen direkten Vergleich, in welcher Luft (Atem-, Raum-, Spülluft) welche Metaboliten enthalten sind. Das zugehörige Venn-Diagramm (Abbildung 6.10) gibt an, in welchem Probenursprung wie viele Metaboliten detektiert wurden und wie viele Metaboliten spezifisch für bestimmte Proben sind. Da die trockenen Nullmessungen

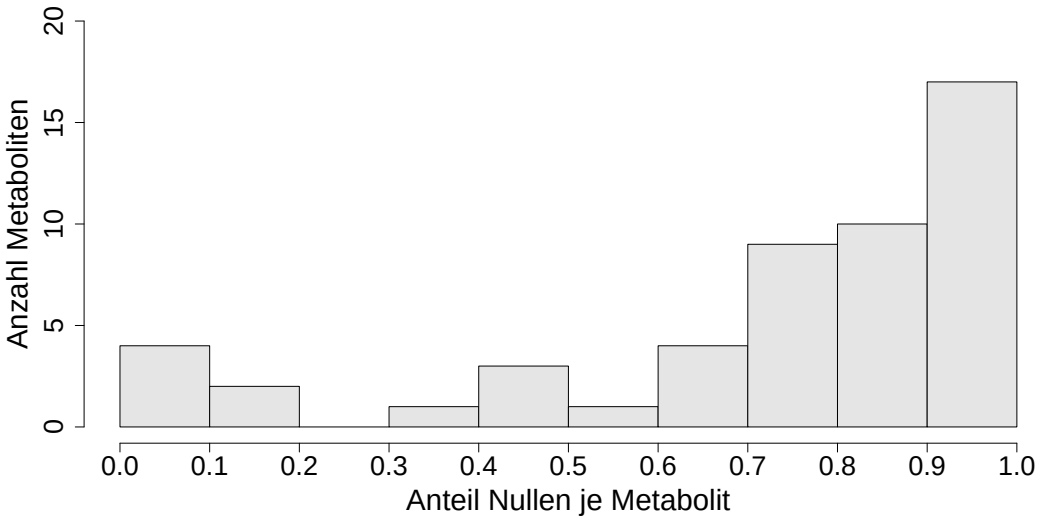


Abbildung 6.11: Histogramm der Anteile an Werten, die für einen Metaboliten genau Null sind.

praktisch keine Peaks enthalten, werden sie hier vernachlässigt. Im Venn-Diagramm 6.10a wird ein Peak einer Probenart zugeordnet, sobald der Peak in einer einzigen zugehörigen Messung enthalten ist (z.B. reicht eine Atemluftmessung aus, in der der entsprechende Metabolit nicht den Wert Null annimmt). In dieser Darstellung werden viele Peaks in allen drei Probenarten gefunden (25 von 51), zwölf wurden ausschließlich in der Atemluft entdeckt, einer nur in der Raumluft und fünf nur in der Spülluft. Da diese Darstellung anfällig für Fehler in der Peakerkennung ist, wurde in der Darstellung 6.10b ein Metabolit erst dann als in der Messart enthalten gezählt, wenn der Metabolit in mehr als 10% der Messungen gefunden wurde. Nach dieser Rechnung sind nur noch elf Metaboliten in allen Probenursprüngen enthalten, 17 sind demnach spezifisch für Atemluft, zwei für die Raumluft und sieben für die Spülluft. Drei Metaboliten fallen ganz aus der Wertung, weil sie für keine der drei Probenarten in mehr als 10% der Messungen gefunden wurden.

In Abbildung 6.11 wird der spezielle Charakter der automatisch generierten Peakintensitäten deutlich. Dargestellt ist, für wie viele Metaboliten bestimmte relative Häufigkeiten an Null-Werten in den Daten enthalten sind. Da im Folgenden nur die Atemluftwerte verwendet werden, basiert dieses Histogramm ausschließlich auf den Atemluftmessungen. Nur sehr wenige Metaboliten wurden in fast allen Atemluftmessungen gefunden (nur vier Metaboliten mit über 90% an Werten ungleich Null) und nur neun mit über 50% Werten ungleich Null.

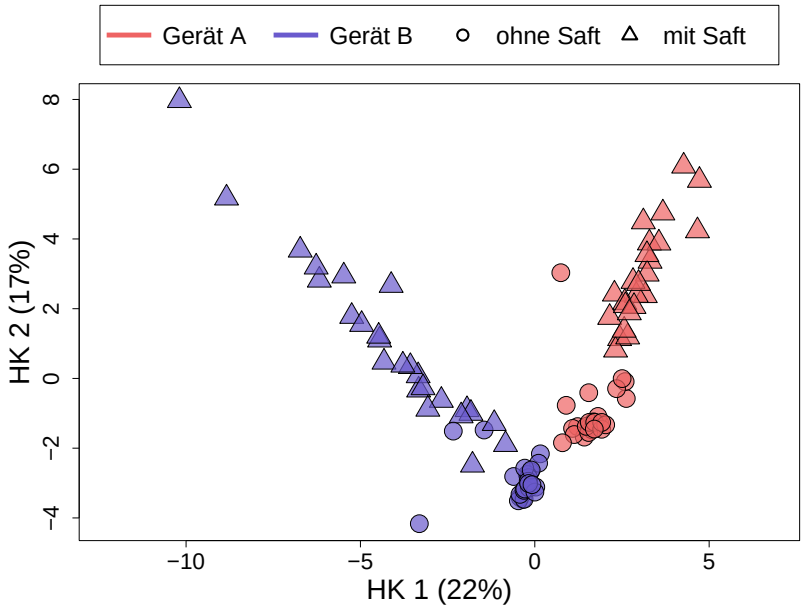


Abbildung 6.12: Die ersten beiden Hauptkomponenten der metabolischen Variablen aus den Atemluftdaten der **automatischen** Peakerkennung. Die Farben geben an, auf welchem Gerät die jeweilige Messung durchgeführt wurde (rot: Gerät A, blau: Gerät B), die Symbole, ob es sich um eine Messung mit oder ohne zuvor konsumierten Orangensaft handelt (Kreis: ohne Orangensaft, Dreieck: mit Orangensaft)

Dafür wurden 17 Metaboliten in weniger als 10% der Atemluftmessungen detektiert. Hier sind auch die Metaboliten enthalten, die speziell nur in der Raumluft oder der Spülluft gefunden wurden.

In allen folgenden Abschnitten werden ausschließlich die Atemluft-Daten verwendet, außer es wird explizit darauf hingewiesen, dass auch die übrigen Luft-Proben verwendet werden.

Analog zu Abbildung 6.7 sind in Abbildung 6.12 die ersten beiden Hauptkomponenten dargestellt. Gemeinsam erklären die beiden Hauptkomponenten 39% der Variabilität in den Daten (die erste Hauptkomponente 22%, die zweite 17%). Die erste HK trennt die Messungen der beiden Geräte, die zweite trennt jeweils Messungen ohne bzw. mit Saft, wobei sich die Beobachtungen jeweils nicht genau parallel zu den Achsen verteilen sondern etwas schräg stehen.

Offensichtlich unterscheiden sich die Messungen zwischen den beiden verwendeten Geräten auch hier stark. Da bekannt ist, dass sich Peakpositionen auf verschiedenen Geräten unterscheiden können (Cumeras u. a., 2012), werden die Peakpositionen der Consensus Peaks

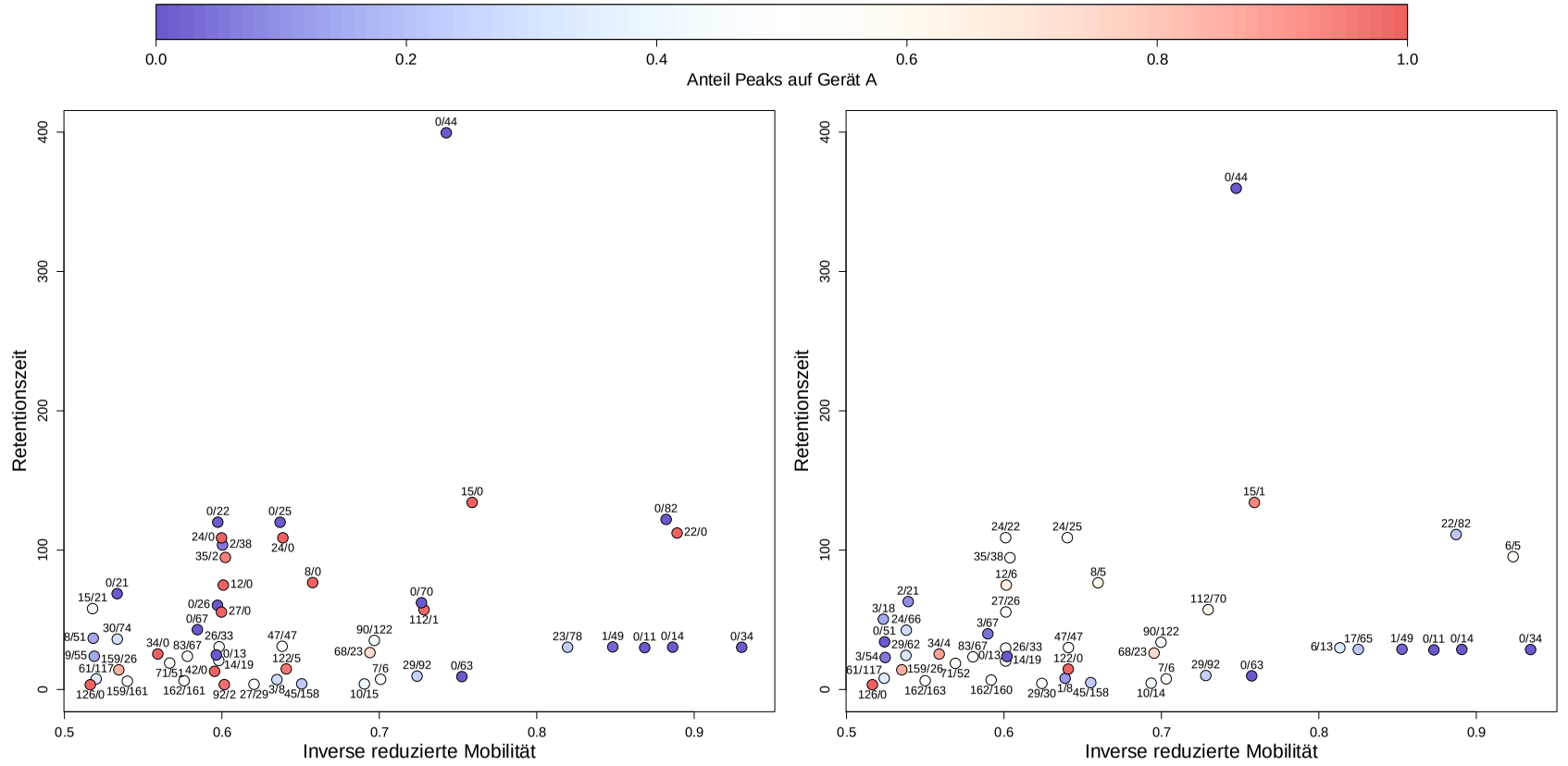
in Abbildung 6.13A erneut dargestellt. Durch Farben wird dargestellt, ob die Metaboliten anteilig häufiger auf Gerät A (rot) oder Gerät B (blau) detektiert werden. Bei ausgeglichenen Anteilen ist die Färbung weiß. Die absoluten Anzahlen der Rohmessungen, bei denen der Wert des entsprechenden Metaboliten ungleich Null ist, sind als Beschriftung beigefügt. Hierbei werden alle Messungen (nicht nur die Atemluftmessungen) berücksichtigt, da für die Peakerkennung alle Messungen verwendet wurden. Daraus resultiert, dass für jeden Peak maximal 172 Werte auf Gerät A und maximal 173 Werte auf Gerät B ungleich Null sein können. Da die Raumluft für beide Geräte die gleiche ist und jede Person jeweils einmal auf jedem Gerät gemessen wurde und somit viele mögliche Störgrößen gleichmäßig verteilt sind, ist davon auszugehen, dass beide Geräte im Mittel vergleichbare Messungen durchführten und somit im Mittel die gleichen Metaboliten jeweils vergleichbar oft gemessen werden sollten und die Färbung der Peaks somit hell sein müsste. Bei Betrachtung der Abbildung fällt jedoch auf, dass viele Peaks hauptsächlich auf einem der Geräte gefunden wurden. Zudem liegen oft Peaks mit gegengleicher kräftiger Färbung dicht beieinander. Eine plausible Erklärung ist, dass die gleichen Peaks auf den beiden Geräten leicht unterschiedliche Positionen annehmen (insbesondere bei IRM$\approx$ 0.6). Aus diesem Grund werden diese Peaks im Cluster-Schritt nicht als ein gemeinsames, sondern als zwei unterschiedliche Peaks erkannt und geclustert. Dabei scheinen die Retentionszeiten auf Gerät B etwas höher und die IRM-Werte etwas niedriger zu sein als auf Gerät A.

Im folgenden Kapitel wird ein Korrektur-Schritt in die automatische Peakerkennung eingefügt (zwischen Peakauswahl und Peakclustern), um diese Problematik zu adressieren.

6.3.3 Automatische Peakerkennung mit Alignierung

Für die Korrektur des Geräte-Effekts wurden in der Vergangenheit Referenzgemische entwickelt, die eine begrenzte Anzahl bekannter Stoffe enthalten. Diese können dann auf verschiedenen Geräten gemessen und die zugehörigen ermittelten Peakpositionen verglichen werden. Für beide Geräte liegen manuell ausgewertete Datenbank-Messungen von Referenzgemischen vor. Für Gerät A wurden fünf Komponenten gemessen, für Gerät B dieselben fünf, jedoch zusätzlich fünf weitere. Da die (manuell bestimmten) Peakpositionen der beiden Geräte verglichen werden sollen, beschränkt sich die Analyse auf die fünf gemeinsamen Stoffe.

Gerät A wird hier willkürlich als Referenzgerät gewählt. Die Peakpositionen, die auf Gerät B ermittelt wurden, sollen demnach auf die Skala der Peakpositionen von Gerät A verschoben werden (vergleiche Kapitel 4.3). Dies wird für Retentionszeit und IRM getrennt durchgeführt,



(a) Consensus Peaks ohne Alignierung der Peakpositionen.

(b) Consensus Peaks nach Alignierung der Peakpositionen der Single Peaks.

Abbildung 6.13: **Automatisch** detektierte Positionen der Consensus Peaks mit den Anzahlen von Rohmessungen (von insgesamt 172 Messungen auf Gerät A und 173 Messungen auf Gerät B), in denen dieser Peak gefunden wurde (# Gerät A/# Gerät B). Die Farben indizieren die Anteile auf beiden Geräten (rot: Peak wurde hauptsächlich auf Gerät A gefunden, blau: Peak wurde hauptsächlich auf Gerät B gefunden, weiß: ausgeglichene Verteilung).

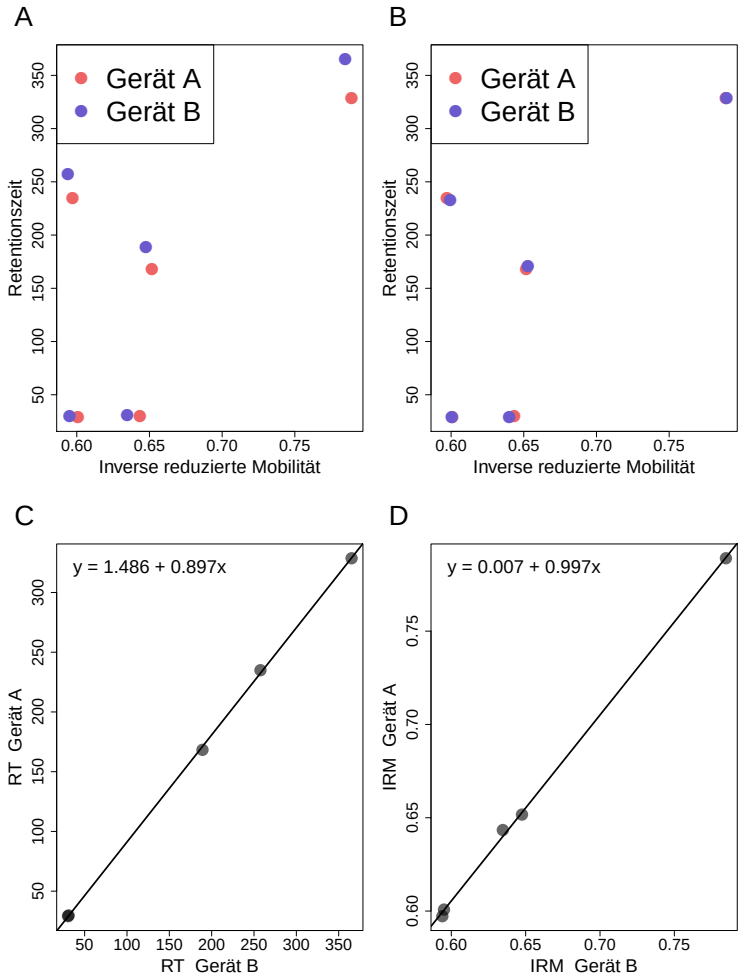


Abbildung 6.14: Manuelle Peakpositionen der Komponenten im Referenzgemischs. Gerät A dient als Referenzgerät. A: Positionen der Referenzstoffe. B: Positionen nach der Alignierung. C: Regression der Retentionszeiten. D: Regression der Inversen reduzierten Mobilitäten.

indem das Modell einer linearen Regression der Positionen angewendet wird und die Positionen von Gerät A auf die Positionen von Gerät B regressiert werden. Abbildung 6.14A zeigt die manuell bestimmten Peakpositionen der fünf Peaks jeweils für beide Geräte. Die Stoffe verteilen sich über den Bereich, an dem häufig Peaks gefunden werden (vergleiche dazu Abbildung 6.9). Die Beobachtung, dass Peaks auf Gerät B häufig größere Retentionszeiten, aber kleinere IRM-Werte annehmen, bestätigt sich hier. Die Abbildungen 6.14C und D zeigen die Ergebnisse der Regressionen für die beiden Dimensionen. Neben der Regressionsgeraden ist auch die Funktionsgleichung angegeben. Sowohl für die Retentionszeiten als auch für die IRM-Werte beschreiben die Regressionsgeraden die Datenpunkte sehr gut. Anschließend können die

Retentionszeiten und IRM-Werte der Peaks von Gerät B anhand der Regressionsgleichungen auf die Skala der Positionen der Peaks von Gerät A verschoben werden, welche dabei unverändert bleiben. Das Ergebnis ist in Abbildung 6.14B dargestellt. Die Positionen der Peaks auf beiden Geräten liegen deutlich näher aneinander als zuvor.

Für die Verbesserung der Peakerkennung werden diese Regressionsgeraden nun verwendet, um die Single Peaks von Gerät B äquivalent zu verschieben, *bevor* die Peaks geclustert werden. Dieser Schritt wird im Folgenden mit *(Peak) Alignierung* bezeichnet, beziehungsweise zur Abgrenzung zur Peakerkennung ohne diesen Zwischenschritt, die Kombination aus Peakauswahl, Peak Alignierung und Peakclustern als *Automatische Peakerkennung mit Alignierung*.

Nach dem Alignierungs-Schritt wird das Peakclustern durchgeführt, wobei 46 Consensus Peaks entstehen. Dies sind fünf Peaks weniger als ohne die Alignierung. In Abbildung 6.13B ist für die resultierenden Consensus Peaks erneut dargestellt, an welchen Positionen sie liegen und auf welchem Gerät sie absolut und anteilig häufiger gefunden wurden. Im Vergleich zur automatischen Peakerkennung ohne Alignierung (Abbildung 6.13A) sind die Verhältnisse für die Consensus Peaks deutlich ausgeglichener, was an der deutlich helleren Färbung vieler Peaks erkennbar ist. Dies trifft besonders auf die Peaks um eine IRM von etwa 0.6 zu, wo zuvor viele Peaks hauptsächlich auf einem der beiden Geräte gefunden worden waren. Dabei ist zu beachten, dass nicht immer die Peaks miteinander verschmelzen, die am nächsten aneinander lagen. Beispielsweise liegen bei der Variante ohne Alignierung (Abbildung 6.13A) bei IRM≈ 0.6 und RT≈ 100) zwei Peaks mit 24/0 beziehungsweise 2/38 Beobachtungen auf Gerät A/B sehr dicht nebeneinander. Diese werden jedoch durch die Alignierung nicht zu einem Cluster, sondern der Peak, der hauptsächlich auf Gerät B gefunden wurde, verschmilzt mit einem Peak von etwas niedrigerer Retentionszeit, das hauptsächlich auf Gerät A gefunden wurde (erkennbar an den absoluten Häufigkeiten in den beiden Abbildungen). Dies verdeutlicht, dass die konkrete Beziehung der Peakpositionen auf beiden Geräten von Bedeutung ist und es nicht ausreichen würde, beispielsweise in einem Korrekturschritt im Anschluss an das Peakclustern nachträglich Consensus Peaks zu verschmelzen, die nahe beieinander liegen und beide hauptsächlich auf unterschiedlichen Geräten gefunden wurden. Betrachtet man die absoluten Häufigkeiten, so fällt auf, dass sich die Unterschiede zwischen beiden Methoden nicht auf die Verschmelzung von Consensus Peaks beschränken, sondern dass sich auch die Anzahlen von Messungen, in denen ein Peak gefunden wurde, verändern, da einige Single Peaks nicht mehr oder erst dann in ein Cluster fallen. Es kann also auch passieren, dass

Consensus Peaks verschwinden, an völlig neuen Positionen entstehen oder sich Peaks aufteilen. Dies ist hier jedoch nur selten der Fall, z.B. entsteht bei RT$\approx$ 100 und IRM$>$ 0.9 ein neuer Peak.

Obwohl sich durch den Alignierungs-Schritt an vielen Stellen Consensus Peaks, die zuvor nur auf einem der Geräte gefunden wurden, zu plausibleren Peaks verschmolzen haben, so bleibt auch danach noch eine nicht zu vernachlässigende Anzahl an Peaks übrig, die hauptsächlich auf einem der Geräte gefunden werden. Dies tritt insbesondere zu frühen IRM-Werten (< 0.55) auf, wo in Abbildung 6.9 optisch nur schwer Consensus Peaks auszumachen waren und zu späten IRM-Werten (> 0.8) bei gleichzeitig frühen Retentionszeiten (< 50). Dort wurden systematisch zur gleichen Retentionszeit sechs Consensus Peaks gefunden, die hauptsächlich auf Gerät B Werte ungleich Null annehmen.

Als mögliche Ursache kommen tatsächliche Geräteunterschiede infrage. Dies können Metaboliten in den Geräten sein, die zum Beispiel aus den verbauten Kunststoffen ausgasen (dass Kunststoffe messbar ausgasen wird beispielsweise in Cumeras u. a. (2012) verwendet). Eine weitere Möglichkeit für tatsächliche Geräteunterschiede können auch technische Ursachen sein, die erhöhte Intensitäten erzeugen, welche dann fälschlicherweise als Peaks erkannt und als Metaboliten interpretiert werden. Letzteres ist besonders für die Peaks in Abbildung 6.13 jeweils unten rechts (RT$<$ 50, IRM$>$ 0.8) plausibel. Diese Peaks werden fast ausschließlich auf Gerät B gefunden und liegen in einem Bereich, in dem nach Einschätzung erfahrener Auswerter von MCC-IMS Messungen bei B&S Analytik keine Peaks zu erwarten sind.

Die Atemluftdaten mit Alignierung sind in Abbildung 6.15 durch die ersten beiden Hauptkomponenten der zugehörigen Hauptkomponentenanalyse dargestellt. Im Vergleich zu den nicht alignierten Daten (Abbildung 6.12) fällt auf, dass sowohl die Geräte als auch die (Nicht-) Saftmessungen auch hier zu großen Teilen deutlich voneinander getrennt liegen. Der Geräte-Effekt wurde also durch die Alignierung nicht vollständig beseitigt. Hier stehen die Beobachtungen jedoch parallel zu den Achsen, sodass die erste Hauptkomponente, die 21% der Varianz erklärt, eindeutig dem Saft-Effekt zuzuschreiben ist und die zweite Hauptkomponente, die 15% der Variabilität erklärt, die beiden Geräte voneinander unterscheidet.

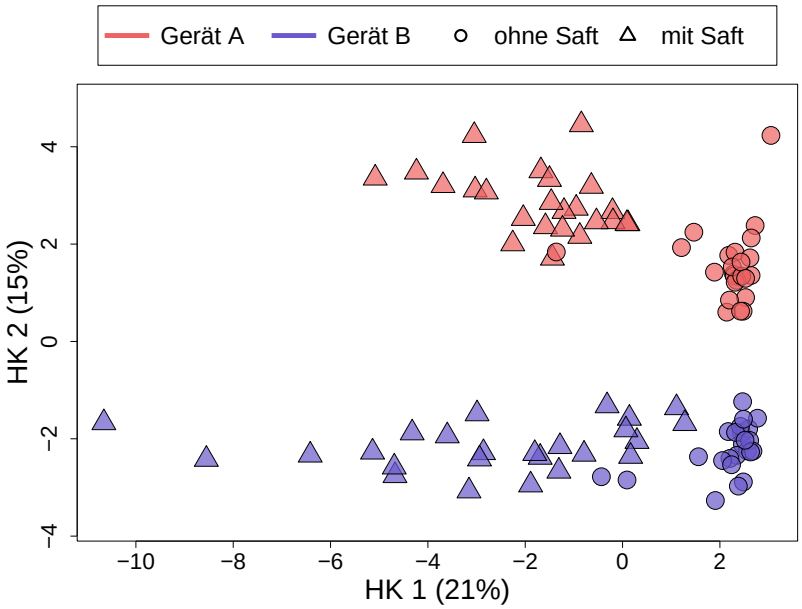


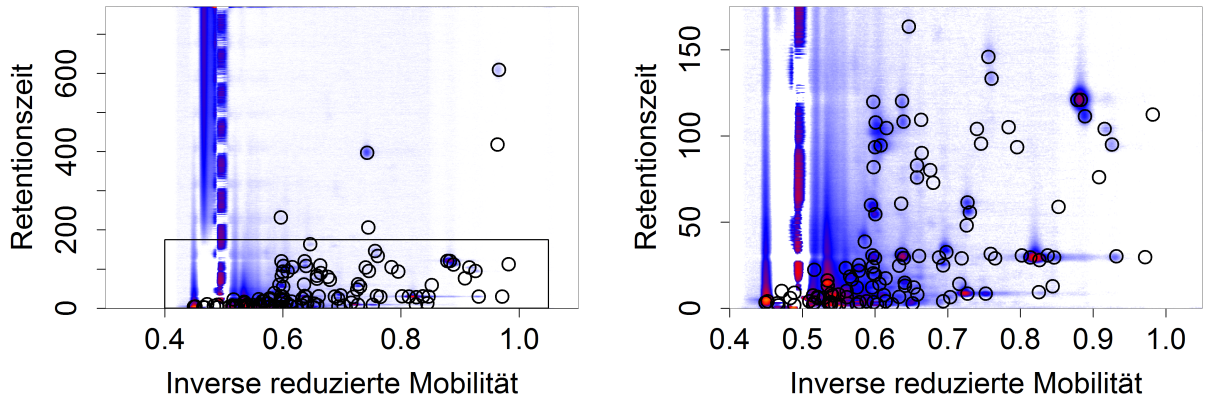
Abbildung 6.15: Die ersten beiden Hauptkomponenten der metabolischen Variablen aus den Atemluftdaten der automatischen Peakerkennung **mit Peak Alignierung**. Die Farben geben an, auf welchem Gerät die jeweilige Messung durchgeführt wurde (rot: Gerät A, blau: Gerät B), die Symbole, ob es sich um eine Messung mit oder ohne zuvor konsumierten Orangensaft handelt (Kreis: ohne Orangensaft, Dreieck: mit Orangensaft)

6.3.4 Vergleich der Peakpositionen

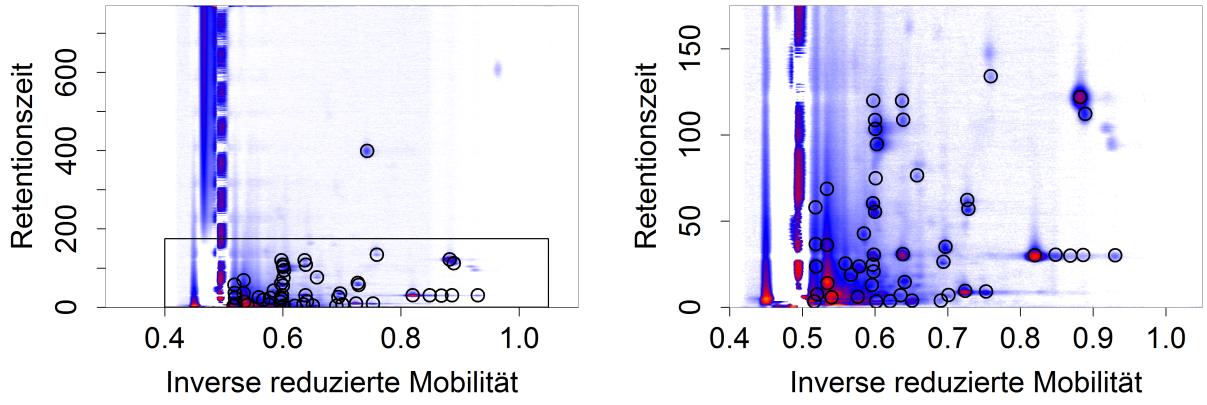
In Abbildung 6.16 sind die Positionen der resultierenden Consensus Peaks der drei vorgestellten Peakerkennungsmethoden (manuell, automatisch, automatisch mit Alignierung) auf allen gemittelten Rohmessungen (nicht nur Atemluftmessungen) dargestellt. Für die automatische Peakerkennung mit Alignierung wird dabei die Skala der Rohmessungen, die auf Gerät B erhoben wurden, mithilfe des linearen Regressionsmodells aus Kapitel 6.3.3 auf die Skala der Rohmessungen von Gerät A transformiert. Dies bedeutet, dass die Retentions- und IRM-Zeiten der Pixel von Gerät B transformiert werden, bevor die Normierung der Pixel und die anschließende Mittelung stattfinden. Auf diese Weise wird sichtbar, wie die Rohmessungen für den Cluster-Schritt der automatischen Peakerkennung aussehen würden. Da die anderen beiden Verfahren jedoch nicht auf diesen Transformierungen basieren, werden die Rohmessungen dort ohne die Verschiebung dargestellt.

In Abbildung 6.16a sind die Peakpositionen der manuellen Peakerkennung dargestellt. Auf den ersten Blick sind keine Peaks nicht gefunden worden, dafür sind an einigen Stellen Peaks notiert, an denen optisch kein Peak sichtbar ist (z.B. vier mit IRM um 0.75 und RT um 100). Dies könnte jedoch auch an der Darstellung liegen, da bei der manuellen Peakerkennung das Vorkommen in einer Rohmessung ausreicht. Da hier jedoch arithmetische Mittel der Rohmessungen dargestellt sind, sind seltene Peaks hier möglicherweise nicht erkennbar. Darüber hinaus fallen Stellen auf, an denen mehrere Peaks dargestellt sind, an denen optisch nur ein Peak erkennbar ist (IRM$\approx$ 0.875, RT$\approx$ 120; IRM$\approx$ 0.825, RT$\approx$ 25). Es ist möglich, dass die grobe Darstellung hier nicht ausreicht, um beispielsweise überlappende Peaks zu erkennen. Es ist jedoch auch möglich, dass Peakverschiebungen, wie die verschiedener Geräte, manuell nicht erkannt werden und so mehrere Peaks detektiert werden. Beispielsweise werden an der Stelle IRM$\approx$ 0.925 und RT$\approx$ 100, wo die automatische Peakerkennung (Abbildung 6.16b) keinen Peak und die automatische Erkennung mit Peakalignierung (Abbildung 6.16c) einen Peak findet, hier zwei Peaks annotiert.

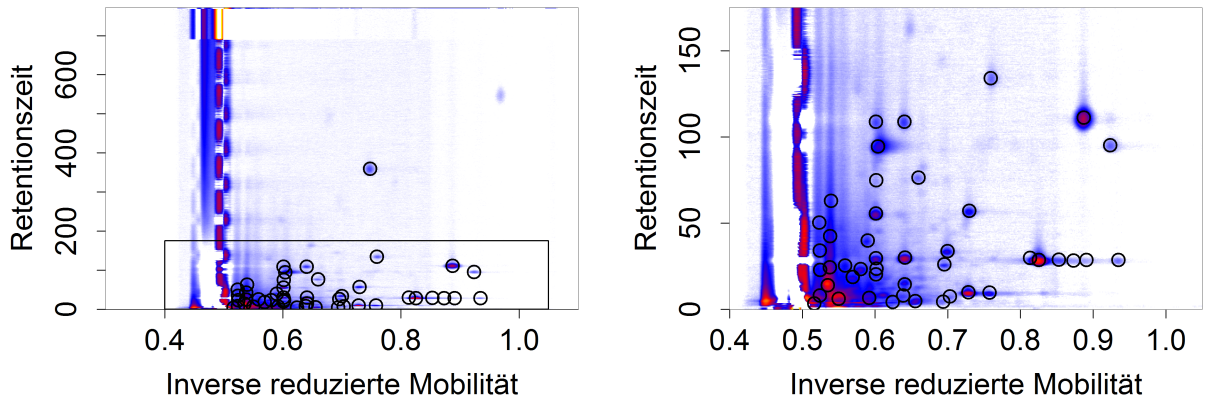
Bei der automatischen Peakerkennung (Abbildung 6.16b) fallen auf den ersten Blick einige nicht erkannte Peaks auf (IRM$\approx$ 0.95, RT$\approx$ 600 oder IRM$\approx$ 0.925, RT$\approx$ 100) sowie einige dicht nebeneinander liegende Peaks (IRM$\approx$ 0.6, RT$\approx$ 100 oder IRM$\approx$ 0.875, RT$\approx$ 120). Durch die Alignierung der Peakpositionen sind letztere in Abbildung 6.16c nicht mehr zu sehen. Durch die entsprechende Transformation der Rohmessungen sind an diesen Stellen auch optisch nicht mehr zwei Peaks erkennbar. Während der Peak bei IRM$\approx$ 0.95 und RT$\approx$ 600 auch hier nicht gefunden wurde, so ist der Peak bei IRM$\approx$ 0.925 und RT$\approx$ 100 durch die



(a) Manuelle Peakerkennung.



(b) Automatische Peakerkennung.



(c) Automatische Peakerkennung mit Alignierung. Hier wurden zusätzlich die Achsen der Rohmessungen von Gerät B auf die von Gerät A aligniert, bevor die Messungen gemittelt wurden.

Abbildung 6.16: Consensus Peaks der drei Verfahren auf allen gemittelten Rohmessungen. Links jeweils die vollständige Rohmessung, rechts ein Ausschnitt.

Alignierung erkannt worden. Aus optischer Sicht ist die automatische Peakerkennung mit Alignierung der automatischen Erkennung ohne Alignierung überlegen. Insgesamt werden deutlich weniger als halb so viele Peaks annotiert wie bei der manuellen Peakerkennung. Dies kann bedeuten, dass die automatische Peakerkennung viele Peaks übersieht, mit den genannten Erklärungen ist es jedoch auch möglich, dass einige dieser Peaks in der manuellen Peakerkennung Duplikate sind oder aus einem Layer stammen, das Metaboliten enthält, die in den Messungen nicht enthalten sind.

6.4 Skalierungen

Um die vorhandenen Unterschiede zwischen den beiden Geräten zu verringern und die Messungen so vergleichbarer zu machen, werden im Folgenden verschiedene Skalierungen durchgeführt. Diese beziehen sich ausschließlich auf die Atemluftmessungen. Die übrigen Messungen werden zur Skalierung nicht verwendet und auch nicht skaliert.

Die verschiedenen Skalierungen (Kapitel 4.4) berücksichtigen unterschiedliche mögliche Ursachen für Unterschiede zwischen den Geräten. Rein additive Unterschiede könnten durch eine Mittelwert-Skalierung ausgeglichen werden, rein multiplikative durch eine Varianz-Skalierung. Eine Kombination aus beiden führt zur häufig angewendeten Standardisierung (auch: „z-Transformation“), bei der Mittelwert und Varianz auf 0 bzw. 1 skaliert werden. Da bei der automatischen Peakerkennung viele Werte Null sind, wird die Standardisierung erweitert, indem Werte, die exakt Null sind, gesondert betrachtet werden. Dabei werden die Werte beider Geräte, die nicht Null sind, getrennt voneinander auf Mittelwert 0 und Varianz 1 skaliert. Die Werte, die Null sind, werden gemeinsam verschoben. Im Fall von exakt gleichen Mittelwerten und Varianzen auf beiden Geräten, würden die Werte, die Null sind, analog transformiert werden. Im Regelfall, dass die beiden Kennzahlen unterschiedlich sind, werden die Null-Werte derart verschoben, dass sie auch nach wie vor die kleinsten Werte bleiben (indem das Maximum der Mittelwerte abgezogen und durch die kleinere Standardabweichung geteilt wird). Um mögliche Ausreißer in Betracht zu ziehen, wird diese Variante auch mit einer robusten Skalierung auf Median 0 und MAD 1 durchgeführt mit analoger Behandlung der Null-Werte.

Unter der Annahme, dass beide Geräte in etwa die gleichen Beobachtungen messen (jede Person wurde auf beiden Geräten genau einmal gemessen, ob zuvor Saft getrunken wurde oder nicht ist auf beiden Geräten etwa gleich verteilt), müssten sich die Werte derart

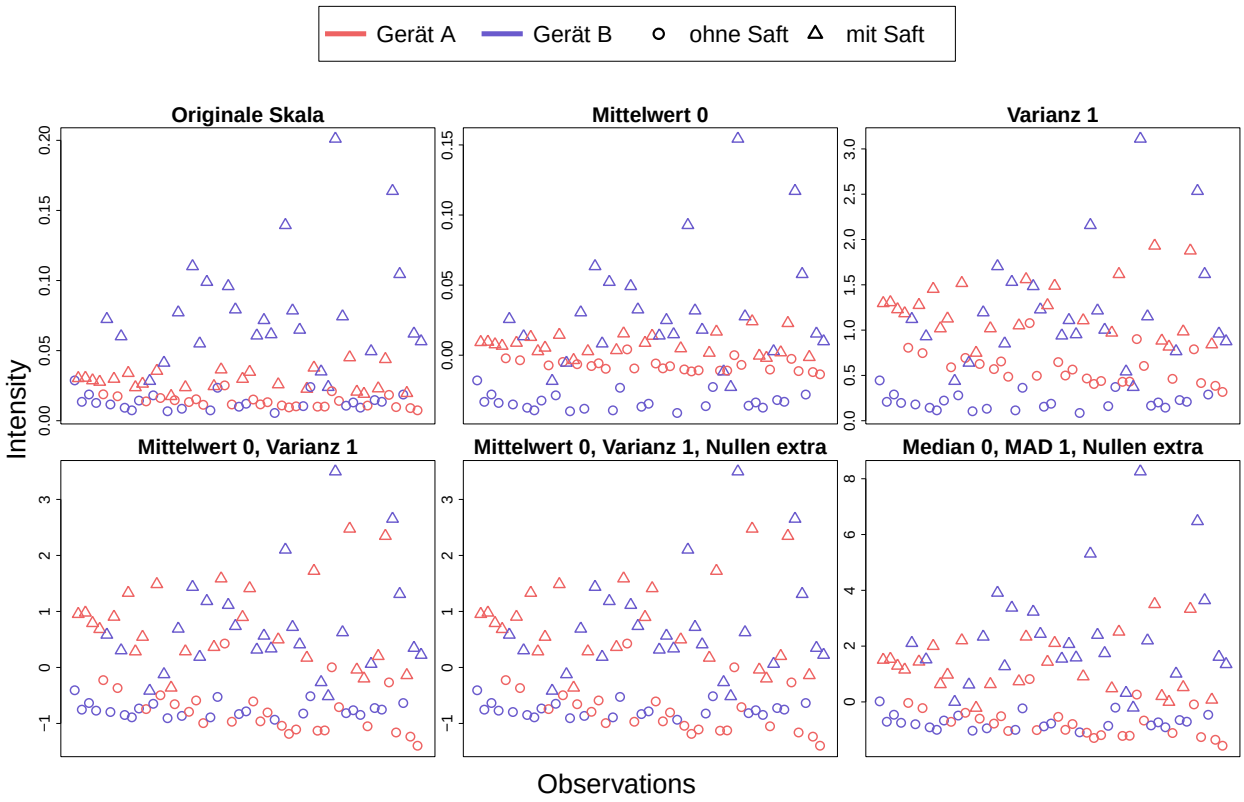


Abbildung 6.17: Beispiel für die Effekte der Skalierungen an einem Metaboliten der **manuellen** Peakerkennung.

skalieren lassen, dass sie sich für die Geräte gemäß ihrer Verteilung nicht mehr wesentlich unterscheiden. Die Auswirkungen der verschiedenen Skalierungen werden in den Abbildungen 6.17 und 6.18 an Beispielen illustriert. Abbildung 6.17 zeigt oben links die unskalierten Atemluft-Beobachtungen eines Metaboliten der manuellen Peakerkennung. Die Geräte sind durch Farben gekennzeichnet. Zusätzlich ist hier der künstliche Haupteffekt mit/ohne Saft durch Symbole kenntlich gemacht. Getrennt betrachtet unterscheiden sich Messungen ohne und mit Saft auf den Messungen je Gerät. Messungen mit Saft erzielen im Mittel höhere Werte als Messungen ohne Saft. Die Werte der Messungen mit Saft sind für Gerät B im Allgemeinen jedoch deutlich größer als die Werte der Messungen mit Saft für Gerät A. Eine einfache Mittelwert-Skalierung (oben mittig) bewirkt die Verschiebung der Beobachtungen von Gerät A zwischen die Beobachtungen ohne und mit Saft von Gerät B. Die Verteilungen beider Geräte unterscheiden sich demnach deutlich aufgrund stark verschiedener Varianzen. Findet nur eine Varianzskalierung statt (oben rechts), so liegen die Beobachtungen mit Saft beider Geräte in vergleichbaren Bereichen, jedoch liegen die meisten Beobachtungen ohne

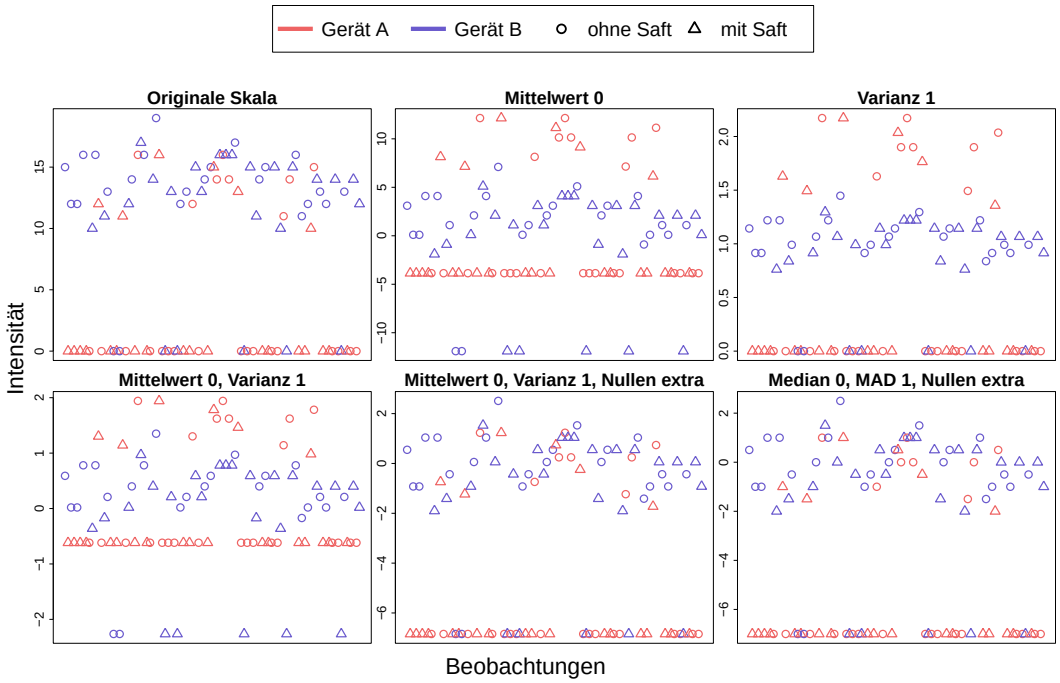


Abbildung 6.18: Beispiel für die Effekte der Skalierungen an einem Metaboliten der automatischen Peakerkennung **mit Alignierung**.

Saft von Gerät A oberhalb der Beobachtungen von Gerät B ohne Saft. Die Mittelwerte der beiden Geräte unterscheiden sich stark. Einzeln reicht für diesen Metabolit also keine der beiden Skalierungen aus. Die Kombination aus Mittelwert- und Varianz-Skalierung (unten links) erreicht eine Angleichung der beiden Verteilungen, die sogar innerhalb der Subgruppen ohne/mit Saft Bestand hat, ohne dass die Subgruppeninformation in die Skalierung eingegangen wäre. Da in den Atemluftmessungen bei der manuellen Peakerkennung keine Nullen enthalten sind, ist die Standardisierungen mit separater Behandlung der Nullen (unten mittig) identisch zur gewöhnlichen Standardisierung. Die robuste Skalierung (unten rechts) unterscheidet sich nicht stark von der gewöhnlichen Skalierung. Die Subgruppen ohne/mit Saft liegen optisch etwas näher beieinander, ansonsten ähneln sich die beiden Skalierungen.

Das zweite Beispiel zeigt einen Metaboliten der automatischen Peakerkennung mit Alignierung (Abbildung 6.18). Im Gegensatz zu den Daten der manuellen Peakerkennung liegen hier viele Werte vor, die exakt Null sind. In diesem Beispiel sind diese Anzahlen nicht gleichmäßig auf beide Geräte verteilt, auf Gerät A liegen deutlich mehr Nullen vor als auf Gerät B. Die Werte, die nicht Null sind, sind auf beiden Geräten in etwa gleich verteilt. Daraus ergibt sich bei der Mittelwert-Skalierung (oben mittig), dass die Werte, die zuvor Null waren, für beide Geräte stark unterschiedliche Werte annehmen (Werte von Gerät A deutlich größer

als Werte von Gerät B). Gleichzeitig schieben sich die Werte, die zuvor nicht Null waren, von einer vorher ähnlichen Skala ebenfalls auseinander (Werte von Gerät A größer als Werte von Gerät B). Insgesamt führt dies zu unplausiblen Ergebnissen. Beobachtungen auf beiden Geräten, bei denen ein Metabolit zuvor nicht detektiert wurde und welche somit den Wert Null erhielten, resultieren nach der Skalierung in unterschiedlichen Werten je Gerät. Es entstehen so noch deutlich größere Unterschiede zwischen den beiden Geräten. Der ursprüngliche Zweck der Skalierung wird somit verfehlt. Wird nur eine Varianz-Skalierung durchgeführt (oben rechts), so bleiben die Werte beider Geräte, die zuvor Null waren, auch nach der Skalierung Null, da sich das Dividieren einer Konstante naturgemäß nicht auswirkt. Die Werte, die nicht Null waren, haben nach der Skalierung auf beiden Geräten jedoch unterschiedliche Wertebereiche, was ebenfalls nicht plausibel erscheint. Werden im Zuge der Standardisierung (unten links) beide Skalierungen durchgeführt, so ähnelt das Ergebnis stark den Ergebnissen der reinen Mittelwert-Skalierung. Somit führt die Standardisierung bei der automatischen Peakerkennung durch das gehäufte Auftreten von Nullen nicht zum gewünschten Resultat. Werden die Nullen jedoch separat behandelt (in der Abbildung unten mittig auf Mittelwert- und Varianz-Skalierung basierend sowie unten rechts für die Skalierung basierend auf Median und MAD), so verschieben sich konsequenterweise die Nullen beider Geräte um den gleichen Wert und sind somit auch nach der Skalierung vergleichbar. Die Werte, die zuvor nicht Null waren, bleiben auf einem vergleichbaren Niveau.

Die ersten beiden Hauptkomponenten der manuellen Peakerkennung, analog zu Abbildung 6.7, hier jedoch auf Mittelwert 0 und Varianz 1 skaliert, sind in Abbildung 6.19 dargestellt. Da die Varianzen der beiden Geräte für alle Metaboliten künstlich angeglichen wurden, ist es nicht überraschend, dass der Geräte-Effekt hier nicht mehr sichtbar ist. Die erste Hauptkomponente erklärt 25% der Variabilität in den Daten und trennt grob Messungen ohne und mit Saft. Wird statt der Standardisierung die robuste Variante verwendet, so ändert sich dies nicht. Die zugehörigen ersten beiden Hauptkomponenten sind in Abbildung B.1 auf Seite 167 im Anhang dargestellt. Beide Abbildungen ähneln sich sehr stark. Die Skalierungen bewirken hier, dass sich die verwendeten Geräte nicht mehr stark auf die Variabilität der Messungen auswirken.

Bei den Daten der automatischen Peakerkennung mit Alignierung werden die Unterschiede zwischen der gewöhnlichen Standardisierung und der Standardisierung mit separater Behandlung der Null-Werte in Abbildung 6.20 anhand ihrer ersten beiden Hauptkomponenten deutlich. Die gewöhnliche Standardisierung (links) zeigt auf den ersten Blick keine großen Geräteunterschiede. Die erste Hauptkomponente, welche 22% der Variabilität der Daten

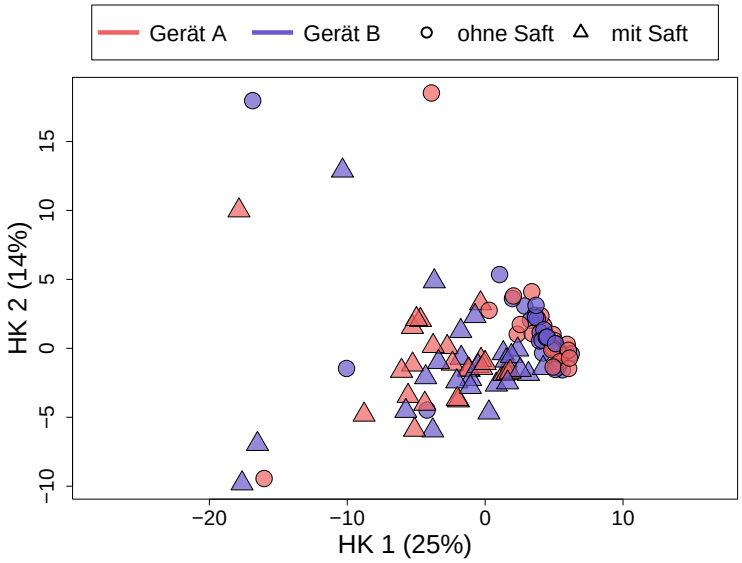


Abbildung 6.19: Die ersten beiden Hauptkomponenten der metabolomischen Atemluftdaten der **manuellen** Peakerkennung mit Skalierung auf Mittelwert 0 und Varianz 1. Die Farben markieren, auf welchem Gerät die jeweilige Messung durchgeführt wurde (rot: Gerät A, blau: Gerät B), die Symbole, ob es sich um eine Messung mit oder ohne zuvor konsumierten Orangensaft handelt (Kreis: ohne Orangensaft, Dreieck: mit Orangensaft).

erklärt, trennt vor allem Messungen ohne und mit Saft. Bei genauerem Hinsehen jedoch fällt auf, dass sich die Messungen ohne Saft in der ersten Hauptkomponente doch bezüglich der beiden Geräte unterscheiden.

Die Skalierung, bei der die Nullen separat behandelt werden (rechts) weist deutliche Unterschiede zwischen den Geräten auf. Die zweite Hauptkomponente, die 10% der Variabilität der Daten erklärt, trennt die beiden Geräte deutlich. Dies erklärt sich dadurch, dass die Varianzen der beiden Geräte durch die separate Behandlung der Nullen eben nicht gleich werden, sondern sich deutlich unterscheiden können, wenn die Anzahl der Nullen auf beiden Geräten sehr verschieden ist. Für die robuste Alternative (siehe Abbildung B.2 im Anhang auf Seite 168) zeigen sich im Vergleich hierzu keine großen Unterschiede.

Insgesamt scheint sich die gewöhnliche Standardisierung für die manuellen Daten zu eignen, um Geräteunterschiede zu mindern. Für die automatischen Daten erzielt diese Skalierung teilweise unplausible Resultate. Für eine intuitive Interpretation der Daten eignet sich die Standardisierung, bei der Nullen getrennt behandelt werden, allerdings sind dann nach wie vor Unterschiede zwischen den Geräten in den Daten erkennbar. Grund für diese Unterschiede können zwei verschiedene Kategorien von Metaboliten in den Daten sein. Null-Werte in den

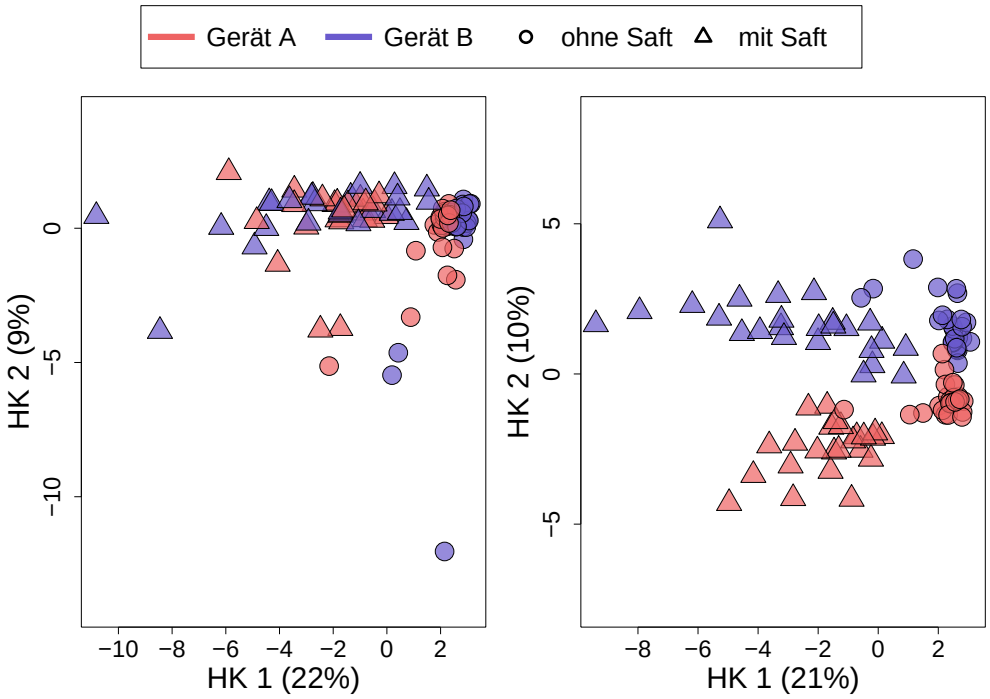


Abbildung 6.20: Die ersten beiden Hauptkomponenten der metabolomischen Atemluftdaten der automatischen Peakerkennung **mit Positionsalignierung**. Die Farben markieren, auf welchem Gerät die jeweilige Messung durchgeführt wurde (rot: Gerät A, blau: Gerät B), die Symbole, ob es sich um eine Messung mit oder ohne zuvor konsumierten Orangensaft handelt (Kreis: ohne Orangensaft, Dreieck: mit Orangensaft). Links: Skalierung auf Mittelwert 0 und Varianz 1. Rechts: Skalierung auf Mittelwert 0 und Varianz 1 mit separater Verschiebung der Nullen.

Daten entstehen überall dort, wo die Peakerkennungsmethoden keinen Peak detektierten. Schwächen der Algorithmen können also dazu führen, dass Peaks nicht entdeckt werden, obwohl der entsprechende Metabolit in der Luftprobe enthalten ist. Dass dies spezifisch für ein Gerät auftritt, kann beispielsweise daran liegen, dass die Detektion eines Peaks durch gerätespezifisches Rauschen gestört wird. Dies tritt im ersten Schritt der Peakerkennung, in der Peakauswahl, auf. Eine weitere mögliche Ursache dafür, dass ein in Wahrheit vorhandener Metabolit nicht entdeckt wird, ist, dass die Peakpositionen auf einem Gerät schwanken oder sich gleichmäßig von den Positionen des anderen Geräts unterscheiden, sodass die Peaks nicht zusammengeclustert werden. Dieser Fall betrifft den Cluster-Schritt der automatischen Peakerkennung. Die zweite Kategorie von Metaboliten, die Unterschiede zwischen den beiden Geräten verursachen, sind Metaboliten, die tatsächlich spezifisch für ein Gerät sind, weil sie beispielsweise aus den verbauten Materialien ausgasen und somit nicht Teil der Atemluftmessung sind.

Die beiden genannten Kategorien von Metaboliten sind in den Daten unterschiedlich zu interpretieren. Ein Metabolit, der in einer Atemluftprobe enthalten ist, aber nicht detektiert wird, entspricht einem fehlenden Wert. Für die Peakerkennung ist die Zuordnung des Wertes Null jedoch folgerichtig, da der entsprechende Metabolit nicht gefunden wurde. Im Fall, dass ein Metabolit aus nur einem Gerät austritt und anschließend detektiert wird, sind allerdings tatsächliche Unterschiede zwischen den Geräten vorhanden. In Bezug auf die Skalierungen bedeutet die Anwendung der gewöhnlichen Standardisierung, dass eher davon ausgegangen wird, dass die Werte fehlen, also der erste Fall zutrifft. Wären alle Werte eines Geräts Null und die Werte des anderen Geräts nicht, so würden nach der Skalierung die Werte des zweiten Geräts um Null herum schwanken. Die Werte des ersten Geräts blieben bei Null und lägen mittig zwischen den Beobachtungen des zweiten. Würde die erweiterte Standardisierung angewendet werden, so würde der Abstand zwischen den Werten des ersten und zweiten Geräts erhalten bleiben. Die Werte des ersten Geräts wären weiterhin niedriger als die des zweiten Geräts. Diese Interpretation entspricht der Vorstellung, dass der entsprechende Metabolit hauptsächlich auf einem der Geräte auftritt. Die Anwendung dieser Skalierung bedeutet also anzuerkennen, dass einige Metaboliten nicht auf allen Geräten gleichermaßen auftreten und diese Tatsache in den Daten erhalten bleiben soll.

In den folgenden Kapiteln werden die unskalierten Daten mit der Standardisierung und der erweiterten Standardisierung (ab jetzt auch bezeichnet mit $\textit{Standardisierung}^{+}$) weiter verglichen. Die einzelnen Mittelwerts- und Varianz-Skalierungen werden im Folgenden nicht weiter betrachtet, da gezeigt wurde, dass diese einzeln nicht ausreichend sind. Die robuste

Skalierung wird aus Gründen der Übersichtlichkeit ab hier nicht weiter betrachtet, da sich keine großen Unterschiede zur erweiterten Standardisierung gezeigt haben. Für Anwendungen, in denen Ausreißer bekanntermaßen auftreten, sollte diese Skalierung jedoch in Betracht gezogen werden.

6.5 Univariate Tests

Um herauszufinden, welche der interessierenden Variablen (Saft, Gerät, Geschlecht, Rauchen) einen Effekt auf die Daten haben und somit auch als Störgröße in Frage kommen, werden zunächst univariate Tests auf Lageunterschiede durchgeführt (zweiseitige Tests). Die p-Werte werden je Fragestellung mit der Methode nach Bonferroni-Holm (Holm, 1979) adjustiert, um dem Umstand des multiplen Testens Rechnung zu tragen. Um die drei verschiedenen Peakerkennungsmethoden weiterhin zu vergleichen, werden die Tests für alle drei Datensätze durchgeführt. Die zuvor ausgewählten Skalierungen (gewöhnliche Standardisierung und erweiterte Standardisierung) werden darüber hinaus mit den unskalierten Daten verglichen.

Die vier jeweils zu unterscheidenden Gruppen sind somit ohne/mit Saft, Gerät A/B, männlich/weiblich sowie (Ex-)Rauchende/Nichtrauchende. Bei der Fragestellung nach dem Rauchen wurden die Personen gemäß des Fragebogens in drei Gruppen eingeteilt: Nichtrauchende, Ex-Rauchende und aktiv Rauchende. Um ein Zwei-Klassen-Problem zu schaffen und da die Gruppe der Rauchenden mit vier Individuen nur sehr dünn besetzt ist, werden im Folgenden die aktiv Rauchenden und die Ex-Rauchenden in eine gemeinsame Gruppe, die „(Ex-)Rauchenden“, zusammengefasst, welche neun von 49 Personen beinhaltet. Da die Ex-Rauchenden schon seit mehreren Jahren nicht mehr rauchen, ist es jedoch denkbar, dass ein potenzieller Effekt des Rauchens nach der Zusammenfassung der beiden Gruppen vermindert oder gar nicht auftritt.

Zum Vergleich werden für jede der vier Fragestellungen ein parametrischer und ein nicht-parametrischer Test durchgeführt. Für die Variablen Saft und Gerät gilt zu beachten, dass jede Person zweimal gemessen wurde, jeweils einmal mit jeder möglichen Ausprägung der Variable. Somit können diese Fragestellungen als gepaarte Testprobleme verstanden werden, bei welchen die Messungen der Testpersonen als Differenzen („mit Saft“−„ohne Saft“, „Gerät B“−„Gerät A“) aufgefasst werden und auf Unterschiede zum Wert Null getestet werden. Als parametrischer Test wird hier ein gepaarter t-Test und als nicht-parametrischer ein gepaarter Wilcoxon-Test verwendet. Für die beiden Variablen Geschlecht und Rauchen werden

ungepaarte Tests, je ein Welch-Test (bei dem im Gegensatz zum t-Test die Varianzen nicht als gleich angenommen werden) und ein Wilcoxon-Test angewendet. Hierbei ist zu beachten, dass jede Person zweimal gemessen wurde und somit die Varianz in den Daten unterschätzt werden könnte. Dies würde zu niedrigeren p-Werten und somit einer erhöhten Wahrscheinlichkeit für den Fehler 1. Art führen. Da die Personen jedoch nicht unter identischen Bedingungen gemessen wurden (jeweils unterschiedliche Geräte und jeweils eine Messung mit und eine ohne Saft), ist nicht davon auszugehen, dass dieser Effekt hier sehr stark ist.

In Tabelle 6.4 ist für alle geschilderten Testprobleme dargestellt, wie viele Metaboliten univariat signifikante Mittelwertsunterschiede aufweisen (jeweils ohne und mit Adjustierung bezüglich der Fragestellung). Im Folgenden werden die einzelnen Fragestellungen anhand der Ergebnisse aus der Tabelle diskutiert. Wird nicht explizit auf etwas anderes hingewiesen, werden die Anzahlen basierend auf den *adjustierten* p-Werten betrachtet. Für die Zielvariablen Saft und Gerät ist in den Abbildungen B.3 bis B.5 im Anhang auf den Seiten 169 bis 171 für je alle Metaboliten der drei Peakerkennungen abgebildet, bei welchen Standardisierungen und bei welchen Tests sie signifikante Unterschiede aufweisen. Signifikante Metaboliten sind rot, nicht signifikante grau unterlegt. Wurde kein Test durchgeführt (dies kommt vor, wenn zu wenige verschiedene Werte realisiert wurden, also zu viele Werte Null sind), ist das entsprechende Feld weiß. Somit kann verglichen werden, bei welchen Standardisierungen und Tests die gleichen Metaboliten signifikant sind.

6.5.1 Saft

Die Fragestellung, ob sich Metaboliten vor und nach dem Trinken eines Glases Orangensaft unterscheiden, kann anhand der Tabelle eindeutig bejaht werden. Unabhängig von der verwendeten Peakerkennung, der Standardisierung oder des Tests werden stets mindestens 14 Peaks gefunden, deren Mittelwerte sich signifikant unterscheiden. Bei der manuellen Peakerkennung werden im Vergleich zu den automatischen Verfahren mindestens doppelt so viele (mindestens 30) signifikante Peaks gefunden, wobei anzumerken ist, dass die automatische Peakerkennung auch insgesamt mehr als doppelt so viele Peaks detektiert hatte. Bei den beiden automatischen Peakerkennungsmethoden sind unabhängig von der Standardisierung und dem verwendeten Test stets etwa 15 Peaks signifikant. Bei der manuellen Peakerkennung steigt die Anzahl der signifikanten Metaboliten durch die Anwendung der Skalierungen (die hier wie in Kapitel 6.4 erläutert, identisch sind) um 9–10 Peaks an.

Tabelle 6.4: Anzahl univariat signifikanter Metaboliten für Mittelwertsunterschiede verschiedener Zielvariablen (Saft, Gerät, Geschlecht, Rauchen). Die Tests wurden für die drei Peakerkennungsmethoden sowie jeweils die unskalierten Daten und für zwei Skalierungen durchgeführt, Standardisierung und Standardisierung mit separater Behandlung der Nullen (= Standardisierung$^+$). Die Anzahlen basieren jeweils auf unadjustierten (= unad.) oder je Zielvariable adjustierten (= ad.) p-Werten. Für die Variablen Saft und Gerät wurden gepaarte Tests (mit * gekennzeichnet), ein Wilcoxon-Test und ein t-Test, durchgeführt, für Geschlecht und Rauchen ungepaarte Tests, ein Wilcoxon und ein Welch-Test.

	Ohne Skalierung				Standardisierung				Standardisierung$^+$			
	Welch/t		Wilcox.		Welch/t		Wilcox.		Welch/t		Wilcox.	
	ad.	unad.	ad.	unad.	ad.	unad.	ad.	unad.	ad.	unad.	ad.	unad.
Manuelle Peakerkennung (insgesamt 124 Peaks)												
Saft*	30	(42)	34	(47)	39	(59)	44	(60)	39	(59)	44	(60)
Gerät*	70	(99)	83	(108)	0	(0)	2	(6)	0	(0)	2	(6)
Geschlecht	0	(4)	0	(1)	0	(6)	0	(5)	0	(6)	0	(5)
Rauchen	0	(5)	0	(1)	0	(19)	0	(6)	0	(19)	0	(6)
Automatische Peakerkennung (insgesamt 51 Peaks)												
Saft*	15	(20)	15	(22)	16	(20)	18	(24)	16	(21)	14	(22)
Gerät*	21	(27)	24	(27)	0	(0)	7	(13)	18	(25)	18	(23)
Geschlecht	0	(0)	0	(1)	0	(1)	0	(2)	0	(0)	0	(2)
Rauchen	1	(4)	0	(2)	1	(4)	0	(2)	1	(6)	0	(2)
Automatische Peakerkennung mit Alignierung (insgesamt 46 Peaks)												
Saft*	15	(19)	14	(20)	15	(19)	15	(21)	15	(20)	14	(20)
Gerät*	11	(17)	11	(20)	0	(0)	8	(14)	8	(14)	7	(14)
Geschlecht	0	(1)	0	(1)	0	(2)	0	(1)	0	(1)	0	(1)
Rauchen	2	(3)	0	(2)	2	(5)	0	(3)	2	(5)	0	(3)

In den Abbildungen B.3 bis B.5 im Anhang auf den Seiten 169 bis 171 jeweils in der linken Abbildung (A) ist zu sehen, dass die signifikanten Metaboliten häufig übereinstimmen, unabhängig von der verwendeten Skalierung oder dem Test unter der Berücksichtigung, dass bei der manuellen Peakauswertung ohne Skalierung grundsätzlich weniger Metaboliten signifikant sind. Die Positionen der signifikanten Peaks sind in Abbildung 6.21 für die drei Peakerkennungsmethoden auf Ausschnitten der gemittelten Rohmessungen dargestellt. Außerhalb des Ausschnitts liegen nur Peaks, für die in keinem Fall ein signifikantes Ergebnis erzielt wurde. Für die automatische Peakerkennung mit Alignierung sind die Rohmessungen von Gerät B auf die von Gerät A aligniert worden, bevor sie gemittelt wurden. Jeweils links sind nur die Atemluftmessungen ohne Saft, rechts nur mit Saft gemittelt worden. Peaks, die für alle Kombinationen (also die ganze Zeile in den Abbildungen B.3 bis B.5) signifikant sind, sind mit ✕ markiert, die bei denen nur einzelne Kombinationen signifikant sind mit + und die, bei denen keine Kombination signifikant ist, mit ∘. Es zeigen sich deutliche Unterschiede zwischen den gemittelten Rohmessungen ohne und mit Saft. Insbesondere bei IRM≈ 0.6 sind jeweils rechts im Bild zwei große Peaks zu sehen (etwa bei RT≈ 30 und RT≈ 60). Aber auch an anderen Stellen (insbesondere die mit ✕ markierten, bei denen alle Kombinationen signifikant sind) lassen sich Unterschiede ausmachen. Auch die beiden Peaks aus der Pilotstudie (vgl. Tabelle 6.1) sind offenbar bei allen Peakerkennungsmethoden detektiert worden und sind in der Abbildung mit ✕ markiert und somit für alle Kombinationen signifikant.

6.5.2 Gerät

Ohne Skalierung des Geräte-Effekts sind viele Peaks univariat signifikant für die Geräte (Tabelle 6.4). Bei der automatischen Peakerkennung ohne Alignierung sind 21–24 von 51 Peaks signifikant. Werden die Peakpositionen vor dem Cluster-Schritt der automatischen Peakerkennung aligniert, so reduziert sich diese Anzahl auf 11 von 46 Peaks. Dies kann eine direkte Folge der Reduzierung von Peaks sein, die jeweils hauptsächlich auf einem der beiden Geräte detektiert wurden (bei der Zusammenlegung zweier solcher Peaks können im Optimalfall gleich zwei Peaks wegfallen, die sich bezüglich der beiden Geräte unterscheiden). Bei der manuellen Peakerkennung sind ohne die Skalierung 70–83 von insgesamt 124 Peaks signifikant, also mehr als die Hälfte.

Für die gewöhnliche Standardisierung sind in Kombination mit dem gepaarten t-Test für keine der drei Peakerkennungsmethoden noch Metaboliten signifikant. Da Mittelwerte und Varianzen in diesem Fall für beide Geräte künstlich angeglichen wurden, ist dies ein zu

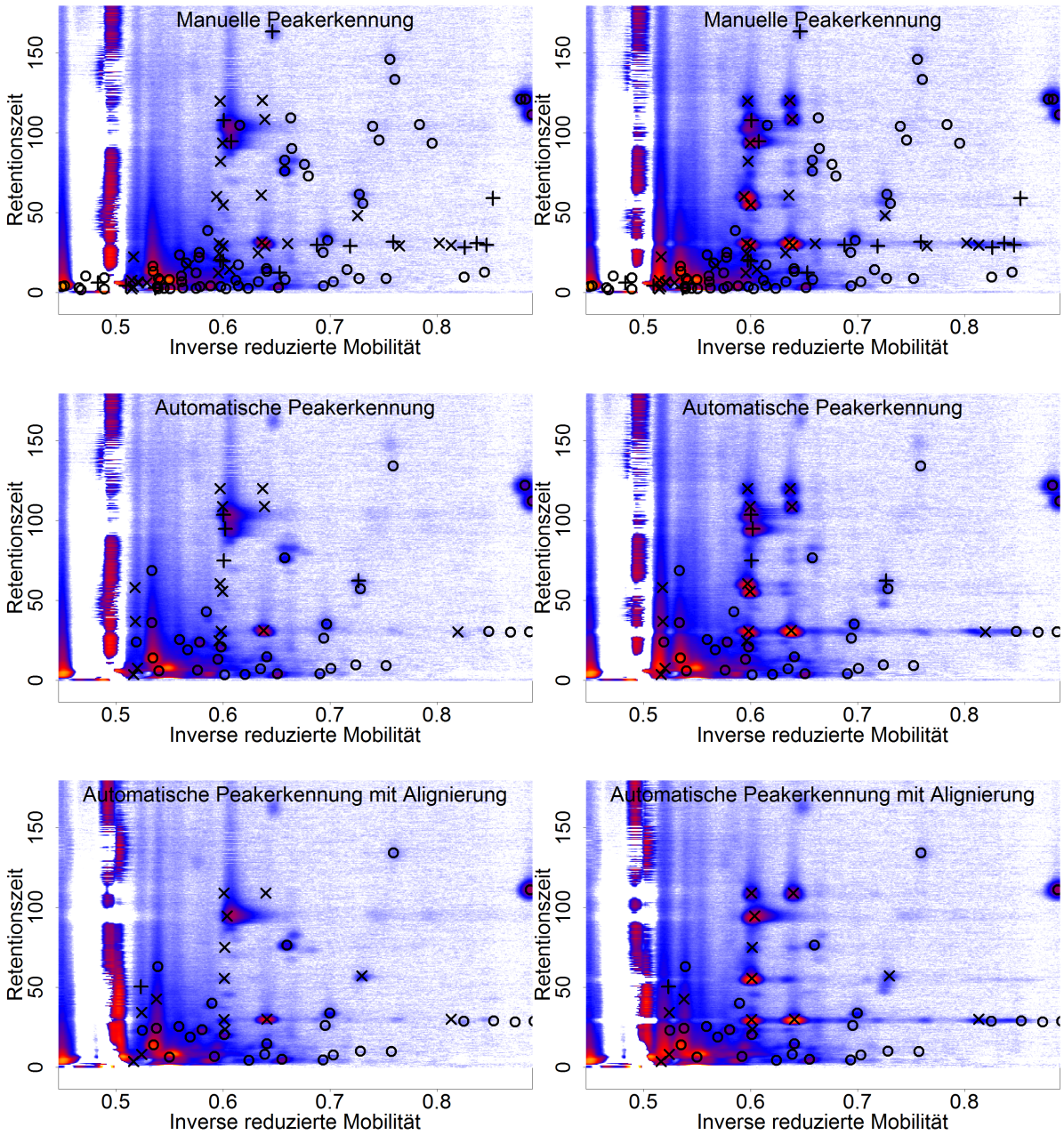


Abbildung 6.21: Ausschnitt der gemittelten Rohmessungen (für die automatische Peakerkennung mit Alignierung wurden zusätzlich die Achsen der Rohmessungen von Gerät B auf die von Gerät A aligniert, bevor die Messungen gemittelt wurden). Jeweils links sind die Atemluftmessungen **ohne Saft**, jeweils rechts die Messungen **mit Saft** gemittelt worden. Die Position eines Peaks, der für alle Kombinationen aus verwendetem Test und den Skalierungen signifikante Unterschiede bezüglich ohne/mit Saft aufweist, ist durch $\times$ markiert, Peakpositionen bei denen nur einzelne oder keine Kombinationen signifikant sind (oder kein Test durchgeführt wurde), sind mit $+$ beziehungsweise $\circ$ dargestellt.

erwartendes Resultat. Im Fall des gepaarten Wilcoxon-Tests sind für die beiden automatischen Peakerkennungen sieben bis acht Peaks und für die manuelle Peakerkennung nur zwei Peaks signifikant verschieden. Dass es beim Wilcoxon-Test zu Signifikanzen kommt, bedeutet, dass sich die Ränge auf beiden Geräten unterscheiden. Die Verteilungen der in diesem Fall signifikanten Metaboliten sind (am Beispiel der automatischen Peakerkennung mit Alignierung) in Abbildung B.6 im Anhang auf Seite 172 dargestellt. Sie zeichnen sich durch eine hohe Anzahl an identischen Werten auf beiden Geräten aus, die im unskalierten Fall alle Null waren. Durch die Skalierung, die Nullen nicht separat behandelt, werden die Werte für beide Geräte auf unterschiedliche Werte verschoben, wie in Kapitel 6.4 erläutert. Da der Wilcoxon-Test ausschließlich Ränge beachtet und viele Werte betroffen sind, unterscheiden sich die Geräte konsequenterweise signifikant.

Wird die Standardisierung$^+$ angewendet, so sind bei gleichem Test (Wilcoxon) bei der automatischen Peakerkennung mit Alignierung sieben Variablen signifikant, es gibt jedoch keine Überlappungen zu den signifikanten Metaboliten bei Verwendung der gewöhnlichen Skalierung (siehe dazu auch Abbildung B.5 auf Seite 171). Die in diesem Fall signifikanten Metaboliten sind in Abbildung B.7 im Anhang auf Seite 173 aufgeführt. Sie sind insbesondere gekennzeichnet durch unterschiedlich viele (unskaliert) Null-Werte auf beiden Geräten. Die Ränge der früheren Null-Werte sind nach der Skalierung identisch. Demnach unterscheiden sich die Ränge der beiden Geräte hauptsächlich durch deutlich mehr gleiche Werte auf einem der beiden Geräte, was zu den Signifikanzen führt. Wird statt des Wilcoxon-Tests ein t-Test durchgeführt, so sind alle sieben Variablen auch in diesem Fall signifikant (eine kommt zusätzlich noch hinzu). Im Vergleich zur gewöhnlichen Standardisierung, für die bei einer Anwendung des t-Tests keine Metaboliten signifikant waren, können hier also Geräteunterschiede festgestellt werden. Analog zu den Überlegungen am Ende von Kapitel 6.4 wird hier deutlich, dass Standardisierung$^+$ bedeutet, anzunehmen, dass es in Wahrheit Geräteunterschiede gibt.

Für die manuellen Daten, bei denen sich Standardisierung und Standardisierung$^+$ nicht unterscheiden, sind die zahlreichen Geräteunterschiede ohne Standardisierung beim t-Test nicht mehr auffindbar, lediglich der Wilcoxon-Test führt noch zu zwei Signifikanzen.

Die Positionen signifikanter Peaks sind analog zur Zielvariable Saft im vorigen Unterkapitel für die Geräte in Abbildung 6.22 auf den mittleren Rohmessungen dargestellt, wobei links die Atemluftmessungen von Gerät A und rechts die Messungen von Gerät B gemittelt wurden. In diesem Fall gibt es jedoch keine Variablen, die für alle Kombinationen aus Test und Skalierung signifikant sind, daher sind keine Peaks mit $\times$ markiert. Für die manuelle Peakerkennung

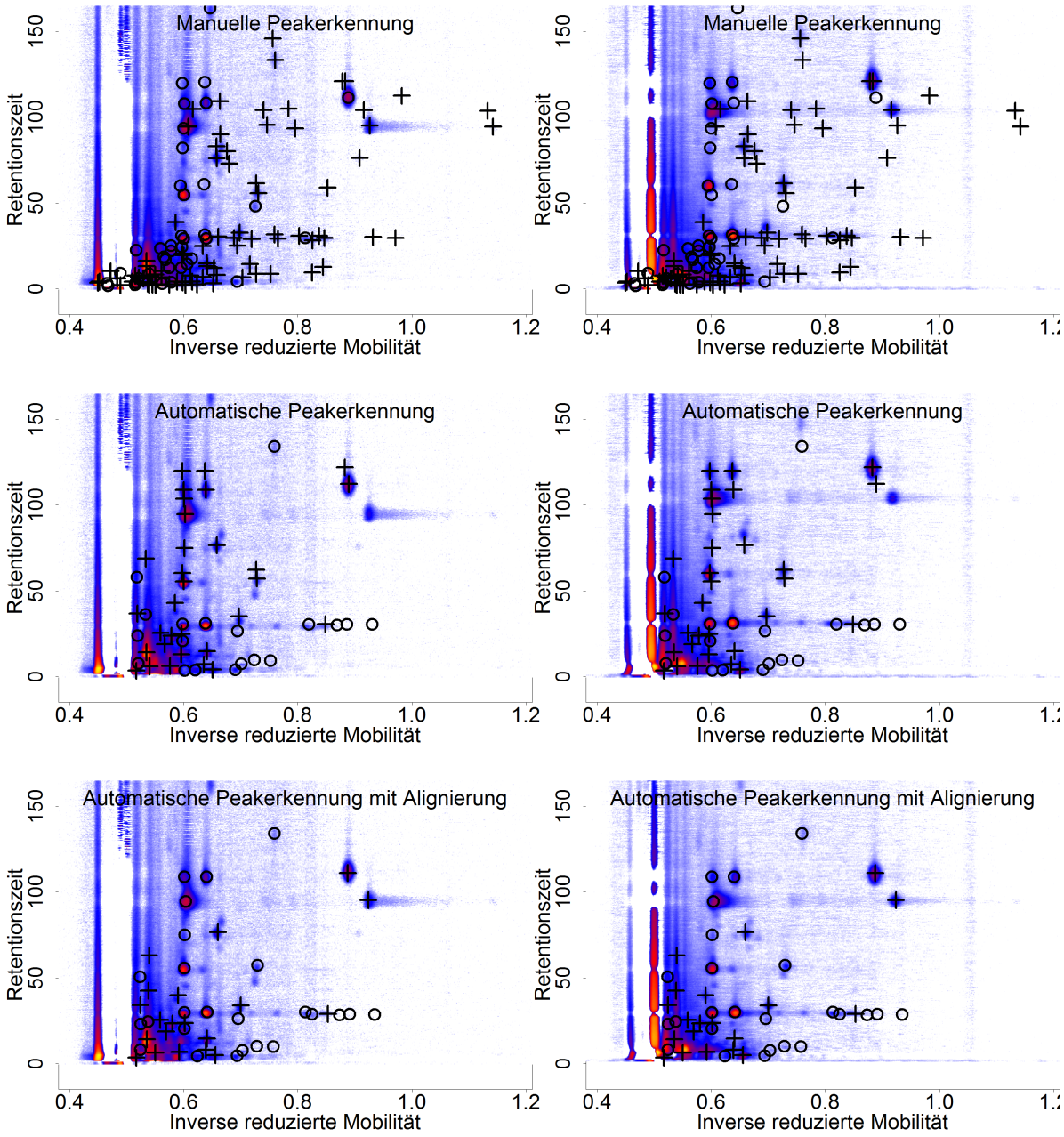


Abbildung 6.22: Ausschnitt der gemittelten Rohmessungen (für die automatische Peakerkennung mit Alignierung wurden zusätzlich die Achsen der Rohmessungen von Gerät B auf die von Gerät A aligniert, bevor die Messungen gemittelt wurden). Jeweils links sind die Atemluftmessungen von **Gerät A**, jeweils rechts die Messungen von **Gerät B** gemittelt worden. Die Positionen eines Peaks, bei dem wenigstens für eine Kombination aus dem verwendeten Test und den Skalierungen ein signifikanter Unterschied zwischen den Geräten festgestellt wird, ist mit + gekennzeichnet. Peakpositionen bei denen kein signifikantes Ergebnis vorliegt (oder kein Test durchgeführt wurde), sind durch ∘ dargestellt.

fällt auf, dass für die meisten Peaks Signifikanzen auftreten. Dies trifft sogar auf Stellen zu, an denen auf den gemittelten Rohmessungen optisch keine Peaks sichtbar sind (beispielsweise bei IRM≈ 0.95 und RT≈ 30). Optische Unterschiede sind bei den gemittelten Rohmessungen besonders zu frühen IRM-Zeiten sichtbar. Der RIP ist bei Gerät B stärker zu sehen, da er offenbar bei frühen Retentionszeiten deutlich höhere Intensitäten annimmt als zu späteren Zeiten und somit durch den Filter (spaltenweises Abziehen des 20%-Quantils) schlechter entfernt wird als bei Gerät A. Bei der manuellen Peakerkennung fällt auf, dass beispielsweise bei IRM≈ 0.6 und RT≈ 60 zwei Peaks durch Kreise markiert sind (also für keine Kombination signifikant sind), obwohl jeder Peak nur auf einem der beiden Geräte sichtbar ist und sie sich demnach stark unterscheiden müssten. Diese waren auch bei der Zielvariable Saft signifikant. Warum diese Metaboliten hier nicht signifikant sind, wird in Kapitel 6.6.1 (Seite 135) deutlich, wo auf diese Metaboliten genauer eingegangen wird. Insgesamt fallen die Geräteunterschiede auf den gemittelten Rohmessungen optisch nicht so stark auf. In Kapitel 6.6.2 wird jedoch noch beispielhaft auf die Positionen einzelner Peaks eingegangen, wobei in den dort betrachteten kleineren Bildausschnitten Unterschiede besser zu erkennen sind.

6.5.3 Geschlecht und Rauchen

Die beiden klinischen Variablen, Geschlecht und Rauchen, weisen in den univariaten Tests kaum Unterschiede auf (Tabelle 6.4). Für das Geschlecht sind nach dem Adjustieren der p-Werte keine Metaboliten (mehr) signifikant, unabhängig von der Datenvorverarbeitung und dem angewendeten Test.

Beim Rauchen werden für die Daten der automatischen Peakerkennung unter Anwendung des t-Tests unabhängig von der Skalierung ein signifikanter Metabolit, für die Daten der automatischen Peakerkennung mit Alignierung zwei signifikante Metaboliten festgestellt. Für die manuellen Daten oder bei Anwendung des Wilcoxon-Tests sind keine Variablen signifikant. Die Ausprägungen der signifikanten Variablen sind in Abbildung 6.23 mit Standardisierung$^+$ dargestellt. Für alle drei Variablen sind die Werte für die Raucher bis auf wenige Ausnahmen für Metabolit R14 identisch (Null vor der Skalierung). Für die Nichtrauchenden hingegen treten andere Werte (vor der Skalierung ungleich Null) auf, sodass der t-Test Mittelwertsunterschiede feststellt. Es ist zu beachten, dass durch den Umstand, dass die Null-Werte nicht nur bei den Nichtrauchenden vorkommen sondern auch in großer Zahl bei den (Ex-)Rauchenden, die entsprechenden Variablen keine diskriminative Stärke aufweisen. Um (Ex-)Rauchende von Nichtrauchenden zu unterscheiden, sind diese Metaboliten (zumindest univariat) ungeeignet,

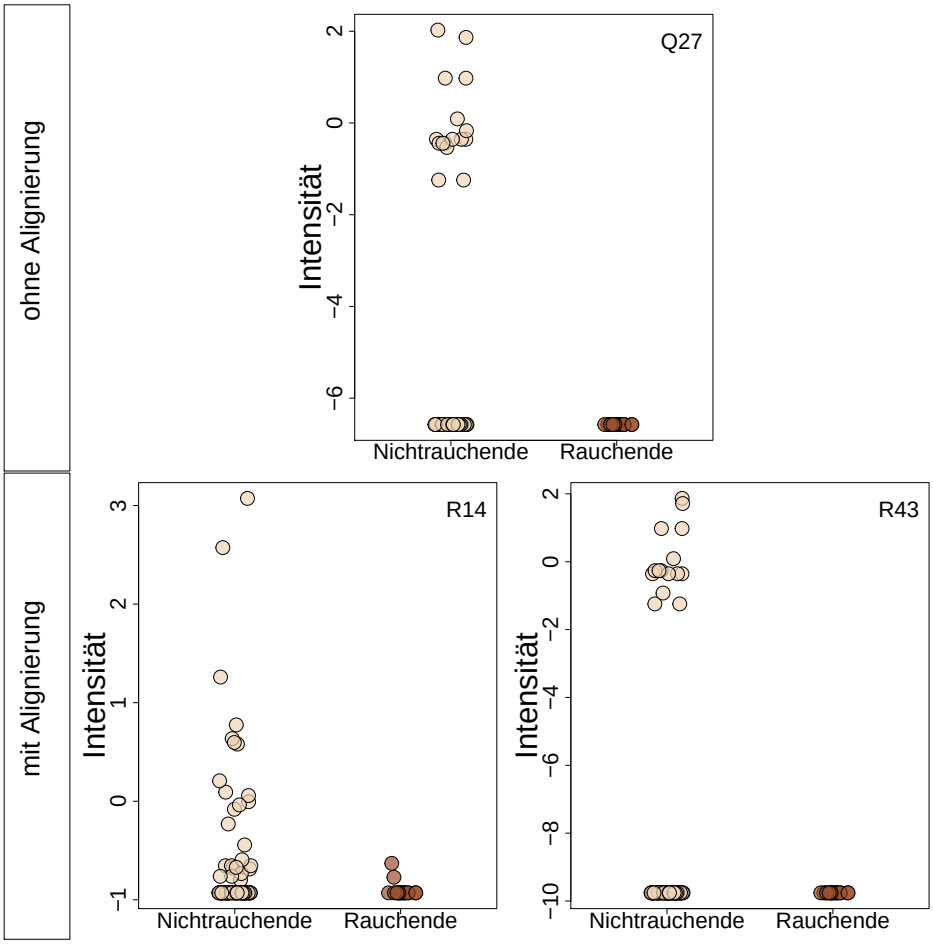


Abbildung 6.23: Verteilung der Metaboliten der **automatischen** Datensätze, die unabhängig von der verwendeten Skalierung für das Rauchen unter Anwendung des t-Tests univariat (adjustiert) signifikant sind. Hier sind die Beobachtungen beispielhaft mit Standardisierung$^+$ dargestellt.

da sie beim Festlegen eines Schwellenwertes entweder alle Null-Werte den (Ex-)Rauchenden oder Nichtrauchenden zuordnen würden und somit eine große Anzahl falscher Klassifikationen zur Folge hätten. In letzterem Fall würden alle Beobachtungen zu den (Ex-)Rauchenden zählen, was einer trivialen und unbrauchbaren Diskriminationsregel entspräche.

6.5.4 Tests auf Lageunterschiede

Es wurden verschiedene Tests auf Lageunterschiede durchgeführt, für die Variablen Saft und Gerät jeweils gepaarte Tests (t- und Wilcoxon-Test), für die Variablen Geschlecht und Rauchen jeweils ungepaarte Tests (Welch- und Wilcoxon-Test). Beide liefern häufig ähnliche Resultate, es werden nicht generell mit einem Test mehr oder weniger Metaboliten als signifikant detektiert als mit dem anderen Test. Eine Ausnahme stellen hierbei die Daten der beiden automatischen Peakerkennungen mit der Zielvariable Gerät mit der Standardisierung dar. Wie im entsprechenden Abschnitt ab Seite 119 erläutert, führt die Skalierung in diesem Spezialfall zu unterschiedlichen Resultaten bei t- und Wilcoxon-Test. Insgesamt verringert sich die Anzahl der signifikanten Metaboliten durch die Adjustierung nicht sehr stark. Lediglich in den Fällen, bei denen von Vornherein nur wenige Metaboliten unadjustiert signifikant waren, sind nach der Adjustierung häufig keine Signifikanzen mehr vorhanden.

6.5.5 Skalierung

Welche Skalierung gewählt wird, hat in erster Linie bei der Fragestellung des Geräts einen Einfluss, da diese Methode durch die Skalierungen unmittelbar beeinflusst wird. Insbesondere die Standardisierung reduziert die Anzahl signifikanter Metaboliten deutlich. Bei Anwendung von Standardisierung$^+$ sind bei den automatischen Peakerkennungen meist mehr Metaboliten signifikant. Da sich bei der manuellen Peakerkennung Standardisierung und Standardisierung$^+$ nicht unterscheiden, sind dort keine Unterschiede vorhanden. Für die Fragestellung ohne/mit Saft unterscheiden sich die Skalierungen in Hinsicht auf die Anzahl signifikanter Metaboliten bei den automatischen Verfahren nicht stark, bei der manuellen Peakerkennung sind nach der Skalierung mehr Metaboliten signifikant. Für die Fragestellungen Rauchen und Geschlecht unterscheiden sich die Skalierungen ebenfalls kaum. Für beide Fragestellungen werden fast keine Signifikanzen erkannt, unabhängig von der angewandten Skalierung.

6.6 Klassifikation

Ob Metaboliten univariat signifikant für Mittelwertsunterschiede zwischen Gruppen sind, hat nur teilweise Aussagekraft darüber, ob diese Gruppen auch mit Hilfe von statistischen Klassifikationsregeln voneinander zu trennen sind. In der Diagnostik von Krankheiten ist aber besonders diese Trennschärfe von Algorithmen notwendig, da Aussagen über Individuen getroffen werden müssen und somit Aussagen über Mittelwerte nicht ausreichen. Darüber hinaus sind Einflussfaktoren besonders dann problematisch für Klassifikationsprobleme, wenn sie selbst klassifiziert werden können. Dann kann eine Zielvariable, die auch mit der Störgröße zusammenhängt, unter Umständen durch Realisationen einer Störgröße (besser) klassifiziert werden, auch wenn die Störgröße nicht ursächlich mit der Zielvariable zusammenhängt. In Kapitel 6.7 wird beispielhaft demonstriert, welche Schwierigkeiten bei der Vernachlässigung von Störgrößen bei der durchgeführten Analyse auftreten können. Im Folgenden wird ein Klassifikationsverfahren angewendet, um zu testen, ob die Beobachtungen bezüglich der vier Zielvariablen Saft, Gerät, Geschlecht und Rauchen klassifizierbar sind. Insbesondere für die beiden letzteren Variablen ist hier von Interesse, ob durch den Einsatz multivariater Verfahren im Gegensatz zu den vorigen univariaten Tests, Unterschiede feststellbar sind.

Den Resultaten aus Kapitel 5 folgend, wird hier ausschließlich der Random Forest als Klassifikationsalgorithmus angewandt. Die Anzahl der Klassifikationsbäume wird an dieser Stelle von 500 auf 1000 erhöht, da die Rechenlaufzeiten hier gering sind und eine höhere Anzahl an Bäumen keine Nachteile mit sich bringt, sondern die Konvergenz verbessert (Breiman, 2001). Es wird eine 50 Mal wiederholte 10-fache stratifizierte Kreuzvalidierung durchgeführt, außer wenn Rauchen die Zielvariable ist. In diesem Fall wird nur eine 9-fache stratifizierte Kreuzvalidierung durchgeführt, damit im Testdatensatz immer zwei Beobachtungen von (Ex-)Rauchenden stammen und nicht nur eine.

Um außerdem zu verhindern, dass die Klasse der Nichtrauchenden durch den Algorithmus durch ihre viel stärkere Klassengröße bevorzugt wird, wird beim Klassifikationsproblem des Rauchens auf sogenanntes „Undersampling“ zurückgegriffen. Für die Kreuzvalidierung werden zunächst, wie bei den übrigen Klassifikationsproblemen auch, stratifizierte Trainings- und Testdatensätze gebildet. Im Testdatensatz sind dementsprechend Messungen von acht oder neun Nichtrauchenden und 2 (Ex-)Rauchenden enthalten. Anschließend werden die Trainingsdatensätze (71 oder 72 Nichtrauchende und 16 (Ex-)Rauchende) durch zufälliges Weglassen von Beobachtungen der größeren Klasse verkleinert, bis die Anzahl der Beobachtungen in beiden Gruppen gleich groß (also 16) ist. Das Modell wird dann auf dem verkleinerten

Trainingsdatensatz gebildet und anschließend auf den (unveränderten) Testdatensatz angewendet. Auf diese Weise bleibt das Prinzip der Kreuzvalidierung, dass für die Schätzung der Gütemaßzahlen jede Beobachtung genau einmal verwendet wird, erhalten.

Es werden die Maßzahlen AUC, ACC, TPR, TNR, PPV und NPV betrachtet, wobei der AUC-Wert von primärem Interesse ist. Der AUC-Wert kann mit Hilfe der geschätzten Wahrscheinlichkeiten für die verschiedenen Klassen eindeutig bestimmt werden, die übrigen Maßzahlen basieren jedoch auf einem zuvor festgelegten Schwellenwert. Dieser Schwellenwert definiert die Grenze für die Wahrscheinlichkeit, die für die Zuordnung einer Beobachtung zur positiven Klasse überschritten werden muss. In Kapitel 5 wurde der naive Standardwert 0.5 verwendet, das heißt, eine Beobachtung wird der Klasse zugeordnet, für welche die geschätzte Wahrscheinlichkeit größer ist als 50%. Da die Prävalenz im Testdatensatz durch die stratifizierte Kreuzvalidierung etwa der Prävalenz im Trainingsdatensatz entspricht (außer beim Rauchen), wird als Alternative die Prävalenzmethode aus Kapitel 4.5.7 angewendet. Dadurch sollte die Prävalenz im Testdatensatz erhalten werden. Die Anwendung dieser Methode in der Kreuzvalidierung hat zur Folge, dass in jeder Kreuzvalidierungsiteration ein anderer Schwellenwert verwendet wird. Für die Berechnung der vom Schwellenwert abhängigen Maßzahlen (also alle mit Ausnahme der AUC) einer Kreuvalidierungswiederholung werden die auf unterschiedlichen Schwellenwerten beruhenden Vorhersagen der einzelnen Kreuzvalidierungsiterationen jeweils zusammengefasst. Die Maßzahlen der 50 Kreuzvalidierungswiederholungen werden abschließend gemittelt. Für die Berechnung des AUC-Werts werden die Wahrscheinlichkeiten direkt verwendet.

Für alle Fragestellungen werden die unskalierten Daten sowie die Daten mit Standardisierung und Standardisierung$^+$ betrachtet, jeweils für die drei Peakerkennungsmethoden manuell, automatisch ohne Alignierung und automatisch mit Alignierung.

6.6.1 Saft

Für die Zielvariable Saft sind die berechneten Maßzahlen bei Verwendung des Schwellenwerts 0.5 in Tabelle 6.5 dargestellt. Unabhängig von der verwendeten Peakerkennungsmethode und der Skalierung ist die Klassifikation nahezu fehlerfrei. Der beste Wert wird für die automatische Peakerkennung mit Alignierung unter Anwendung von Standardisierung$^+$ erzielt. Mit einem AUC-Wert, TNR und PPV von 1, bei einer Accuracy von 0.999, sowie TPR und NPV von 0.998, spricht dies für nur sehr seltene Fehlklassifikationen in einzelnen Kreuzvalidierungswiederholungen. Insgesamt sind die meisten Werte, insbesondere AUC und

Tabelle 6.5: Gemittelte Gütemaßzahlen der kreuzvalidierten Klassifikationsergebnisse für die Zielvariable **Saft** bei Verwendung des **Schwellenwerts 0.5**. Die Klassifikation wurde für die unskalierten Daten und für zwei Skalierungen durchgeführt, Standardisierung und Standardisierung mit separater Behandlung der Nullen (= Standardisierung$^+$).

Skalierung	AUC	ACC	TPR	TNR	PPV	NPV
Manuelle Peakerkennung						
ohne	1.000	0.987	0.992	0.981	0.981	0.992
Standard.	0.999	0.989	0.978	1.000	1.000	0.979
Standard.$^+$	0.999	0.989	0.978	1.000	1.000	0.978
Automatische Peakerkennung						
ohne	1.000	0.990	0.981	1.000	1.000	0.981
Standard.	0.999	0.990	0.980	1.000	1.000	0.980
Standard.$^+$	1.000	0.990	0.980	1.000	1.000	0.980
Automatische Peakerkennung mit Alignierung						
ohne	1.000	0.991	0.982	1.000	1.000	0.983
Standard.	1.000	0.991	0.981	1.000	1.000	0.982
Standard.$^+$	1.000	0.999	0.998	1.000	1.000	0.998

ACC, für die automatische Peakerkennung mit Alignierung für die Skalierungen besser als die Werte der manuellen Peakerkennung und der automatischen Peakerkennung ohne Alignierung bei den jeweiligen Skalierungen. Da die Klassifikation jedoch insgesamt bei allen Varianten fast perfekt durchgeführt wird, sind keine großen Unterschiede festzustellen.

Welche Beobachtungen gut und welche schwerer klassifizierbar sind, kann in den Abbildungen B.8 bis B.10 im Anhang auf den Seiten 174 bis 176 abgelesen werden. Dort ist für die drei Peakerkennungsmethoden für jede Beobachtung der Abstand ihrer geschätzten Wahrscheinlichkeit für die korrekte Klasse zum Schwellenwert dargestellt. Eine Beobachtung, die mit Wahrscheinlichkeit 1 ihrer wahren Klasse zugeordnet wird, erhält den Wert 0.5, eine Beobachtung, die mit Wahrscheinlichkeit 0.8 der falschen, also mit Wahrscheinlichkeit 0.2 der korrekten Klasse zugeordnet wird, erhält den Wert -0.3. Auf diese Art erhält jede fehlklassifizierte Beobachtung einen negativen Wert. Die Boxen entstehen durch die Wiederholungen der Kreuzvalidierung. Jede Box besteht dementsprechend aus 50 Punkten.

Über die drei Peakerkennungsmethoden hinweg werden nur drei Beobachtungen fehlklassifiziert, davon ist eine Beobachtung ohne und zwei mit Saftkonsum. Die Beobachtung ohne Saft wird nur für die manuelle Peakerkennung ohne Skalierung fehlklassifiziert. Eine der Beobachtungen mit Saft wird nur selten fehlklassifiziert (für einzelne Kreuzvalidierungswiederholungen bei der manuellen Peakerkennung sowie der automatischen Peakerkennung mit Alignierung). Die andere Beobachtung mit Saft wird bei allen Peakerkennungsmethoden und Skalierungen mindestens in einigen, häufig in allen Kreuzvalidierungswiederholungen fehlklassifiziert. Die exzellenten Maßzahlen der automatischen Peakerkennung mit Alignierung bei Standardisierung$^+$ erklären sich hier dadurch, dass nur diese eine Beobachtung selten fehlklassifiziert wird, wohingegen alle übrigen Beobachtungen stets korrekt klassifiziert werden.

Wird der Schwellenwert nicht auf 0.5 festgesetzt, sondern entsprechend der Prävalenz-Methode in jedem Trainingsdatensatz aller Kreuzvalidierungen bestimmt, ergeben sich die Maßzahlen wie in Tabelle 6.6. Der über alle Trainingsdatensätze der Kreuzvalidierungen gemittelte Schwellenwert ist in der letzten Spalte angegeben. Außer für die manuelle Peakerkennung ohne Standardisierung, liegen die gemittelten Schwellenwerte für alle Peakerkennungsmethoden und Standardisierungen weit unterhalb von 0.5 (0.357–0.444). Dies ist ein Hinweis darauf, dass die Wahrscheinlichkeiten für die positive Klasse nicht symmetrisch um 0.5 verteilt sind, sondern auch für in Wahrheit positive Beobachtungen Wahrscheinlichkeiten unterhalb von 0.5 geschätzt werden bzw. eine perfekte Trennung vorliegt, deren Wahrscheinlichkeitsmittelpunkt nicht 0.5 ist. Für die manuelle Peakerkennung sowie die automatische Peakerkennung ohne

Tabelle 6.6: Gemittelte Gütemaßzahlen der kreuzvalidierten Klassifikationsergebnisse für die Zielvariable **Saft**, sowie das arithmetische Mittel der verwendeten Schwellenwerte s (entsprechend der **Prävalenz-Methode**). Die Klassifikation wurde für die unskalierten Daten und für zwei Skalierungen durchgeführt, Standardisierung und Standardisierung mit separater Behandlung der Nullen (= Standardisierung$^+$).

Skalierung	AUC	ACC	TPR	TNR	PPV	NPV	s
Manuelle Peakerkennung							
ohne	1.000	0.982	0.984	0.980	0.981	0.984	0.514
Standard.	0.999	0.973	0.979	0.968	0.969	0.979	0.401
Standard.$^+$	0.999	0.972	0.978	0.967	0.967	0.978	0.399
Automatische Peakerkennung							
ohne	1.000	0.982	0.980	0.984	0.985	0.980	0.444
Standard.	0.999	0.972	0.979	0.966	0.967	0.979	0.385
Standard.$^+$	1.000	0.986	0.980	0.992	0.992	0.980	0.426
Automatische Peakerkennung mit Alignierung							
ohne	1.000	0.998	0.997	1.000	1.000	0.997	0.359
Standard.	1.000	1.000	0.999	1.000	1.000	0.999	0.357
Standard.$^+$	1.000	1.000	0.999	1.000	1.000	0.999	0.391

Alignierung verschlechtern sich die Maßzahlen bei Anwendung der Prävalenzmethode im Allgemeinen etwas (insbesondere TNR und PPV). Da der AUC-Wert nicht vom Schwellenwert abhängt, bleibt er konstant. Für die automatische Peakerkennung mit Alignierung verbessern sich die Maßzahlen hingegen noch weiter, sodass die mittlere Accuracy bei Anwendung der Standardisierung$^+$ bei 1 liegt.

Diese Unterschiede sind auch in den Abbildungen der Differenzen zum Schwellenwert (Abbildungen B.11 bis B.13 im Anhang auf den Seiten 177 bis 179) erkennbar. Insgesamt sind die Boxen hier breiter, das heißt, dass die Wahrscheinlichkeiten für die korrekte Klasse stärker schwanken. Dies ist möglicherweise dadurch bedingt, dass die Schwellenwerte hier nicht mehr konstant sind, sondern selbst schwanken und zusätzlich die maximale Differenz zum Schwellenwert s nicht mehr auf den Bereich $[-0.5, 0.5]$ beschränkt ist, sondern bis zu $\max\{s, 1-s\}$ betragen kann. Über die drei Peakerkennungen hinweg sind hier fünf verschiedene Beobachtungen von Fehlklassifikationen betroffen. Für die manuelle sowie die automatische Peakerkennung ohne Alignierung verschlechtert sich die Klassifikation dadurch insgesamt, bei der automatischen Peakerkennung mit Alignierung wird jedoch in Kombination mit Standardisierung oder Standardisierung$^+$ nur noch eine Beobachtung fehlklassifiziert und dies noch etwas seltener als bei feststehendem Schwellenwert von 0.5 (vgl. Abbildung B.10, im Anhang auf Seite 176).

Insgesamt ist das Klassifikationsproblem zu einfach, um sichere Schlüsse auf Vor- und Nachteile der einzelnen Varianten zu ziehen. Es könnte sich bei den Unterschieden auch um zufällige Schwankungen handeln, da die Unterschiede im Wesentlichen auf der Klassifikation nur einer oder sehr weniger Beobachtungen beruhen.

Die Variablenwichtigkeit gibt an, welche Bedeutung die einzelnen Peaks für die Klassifikation haben. Die jeweils drei Peaks mit den höchsten (über die Kreuzvalidierungswiederholungen medianen) Werten für die Variablenwichtigkeit sind für die drei Peakerkennungsmethoden in Kombination mit Standardisierung bzw. Standardisierung$^+$ in Abbildung 6.24 abgebildet. Für die beiden Skalierungen sind die drei wichtigsten Peaks jeweils dieselben, dargestellt sind hier die Werte nach Anwendung von Standardisierung$^+$. Von links nach rechts nimmt jeweils die Variablenwichtigkeit ab. Teilweise reicht bereits eine einzige Variable aus, um die Messungen mit und ohne Saft zu unterscheiden. Die Variablen, die eine höhere Variablenwichtigkeit aufweisen (links) trennen die Beobachtungen optisch besser als die Variablen mit geringerer Variablenwichtigkeit (rechts).

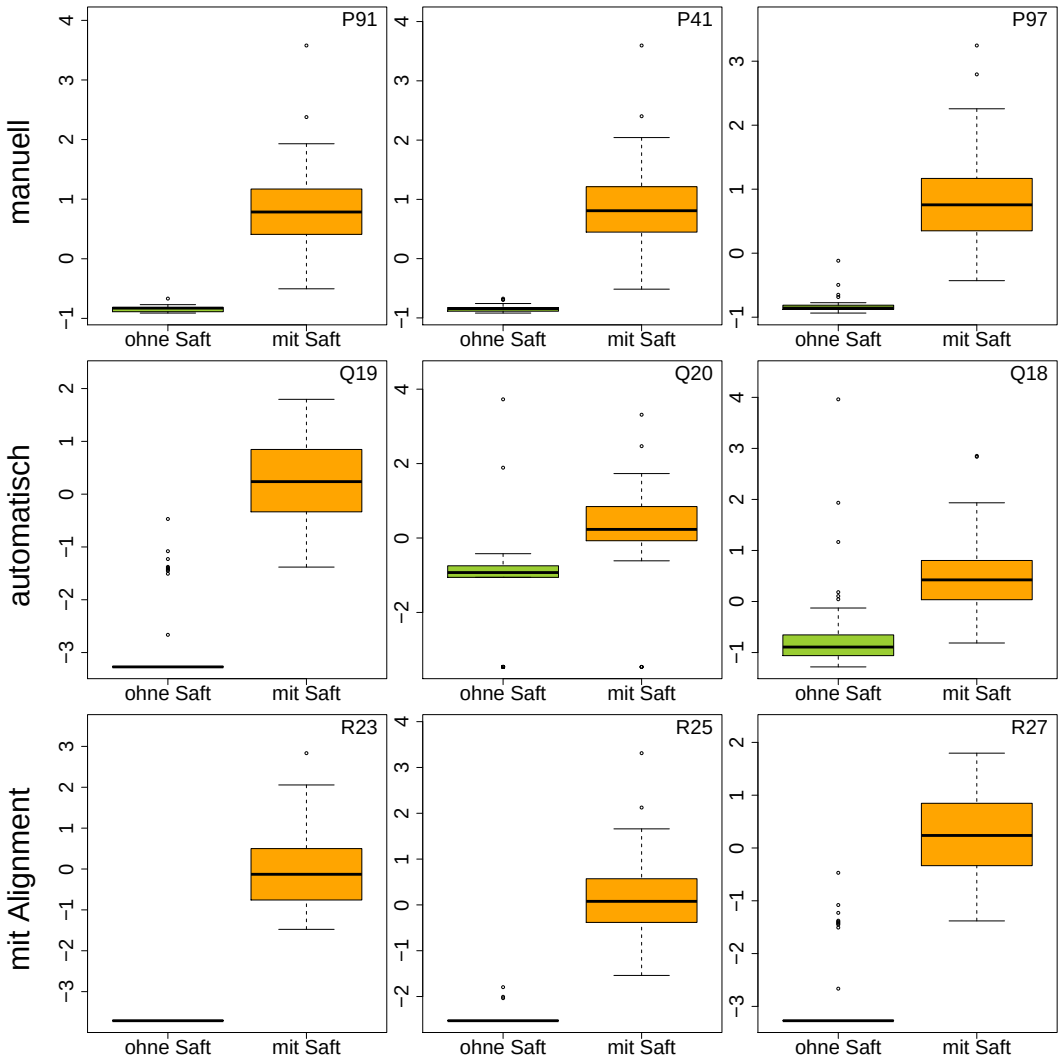


Abbildung 6.24: Boxplots der jeweils drei Variablen mit den höchsten Variablenwichtigkeiten beim Klassifikationsproblem ohne/mit **Saft** für die Datensätze der drei Peakerkennungen mit Standardisierung$^+$ (die drei wichtigsten Variablen unterschieden sich nicht bei Standardisierung und Standardisierung$^+$). Von links nach rechts nimmt die Variablenwichtigkeit ab.

Um beurteilen zu können, ob für die Klassifikation bei den verschiedenen Peakerkennungsmethoden ähnliche Bereiche auf den Rohmessungen relevant sind, sind in Abbildung 6.25 die Peakpositionen der jeweils fünf wichtigsten Peaks (bei Anwendung der Standardisierung$^+$) auf die gemittelten Rohmessungen als Kreuze eingezeichnet. Diese fünf Peaks kamen bei allen Skalierungen am häufigsten unter den wichtigsten fünf vor. Die übrigen gefundenen Peaks im Bildausschnitt weisen niedrigere Variablenwichtigkeiten auf und sind als Kreise dargestellt. Jeweils links sind nur die Rohmessungen der Personen, die keinen Saft getrunken haben, gemittelt, rechts die Messungen der Personen nach dem Saftkonsum. Bei der manuellen Peakerkennung liegen zweimal zwei der fünf wichtigsten Peaks recht dicht beieinander (P41/P91 und P31/P117). Bei der automatischen Peakerkennung ohne Alignierung ist an einer dieser Stellen nur ein Peak gefunden (Q19, ebenfalls in den Top 5), bei der anderen Stelle sind ebenfalls zwei Peaks gefunden, von denen jedoch nur eines unter den Top 5 ist (Q17). Bei der automatischen Peakerkennung mit Alignierung ist an beiden Stellen jeweils nur ein Peak gefunden worden. Beide Peaks (R25 und R27) gehören zu denen mit den fünf höchsten Wichtigkeits-Werten. Diese drei Bereiche wurden also in allen drei Peakerkennungsmethoden als relevant für die Saft-Klassifikation identifiziert. Da die manuelle Peakerkennung in zwei Bereichen zwei relevante statt nur einem Peak detektiert, unterscheiden sich die automatischen Peakerkennungen von der manuellen Peakerkennung, indem sie jeweils zwei zusätzliche Bereiche auf der Rohmessung als relevant identifizieren. Insgesamt stimmen die relevanten Bereiche der verschiedenen Peakerkennungsmethoden jedoch ungefähr überein.

Anhand der gemittelten Rohmessungen ist deutlich zu sehen, wo Unterschiede zwischen den Messungen mit und ohne Saft bestehen. Im Bereich der Peaks P97 (beziehungsweise Q22, R23) ist in den Messungen ohne Saft kein Peak und bei den Messungen mit Saft ein sehr deutlicher Peak zu sehen. Für Peak P91/P41 (beziehungsweise Q17, R25) sowie P117/P31 (beziehungsweise Q19, R27) ist in den Messungen ohne Saft nur ein sehr kleiner Peak zu sehen. Bei R28, dem zusätzlichen Bereich der automatischen Peakerkennung mit Alignierung, ist bei den Messungen ohne Saft in der gemittelten Darstellung kein deutlicher Peak zu sehen, möglicherweise überlappt dieser mit dem direkt darunter liegenden Peak R24. Dieser ist bei den Messungen mit Saft deutlicher von R28 getrennt und nimmt im Mittel höhere Werte an, was sich durch eine leuchtendere Färbung des Peaks zeigt.

Die Rohmessungen der automatischen Peakerkennung mit Alignierung (untere Bilder) wurden auch hier auf die Skala der Rohmessungen von Gerät A verschoben, bevor die Rohmessungen gemittelt wurden. Aus diesem Grund sehen die mittleren Rohmessungen in diesem Fall leicht verändert aus und veranschaulichen die Korrektur des Geräte-Effekts. An den beiden Stellen,

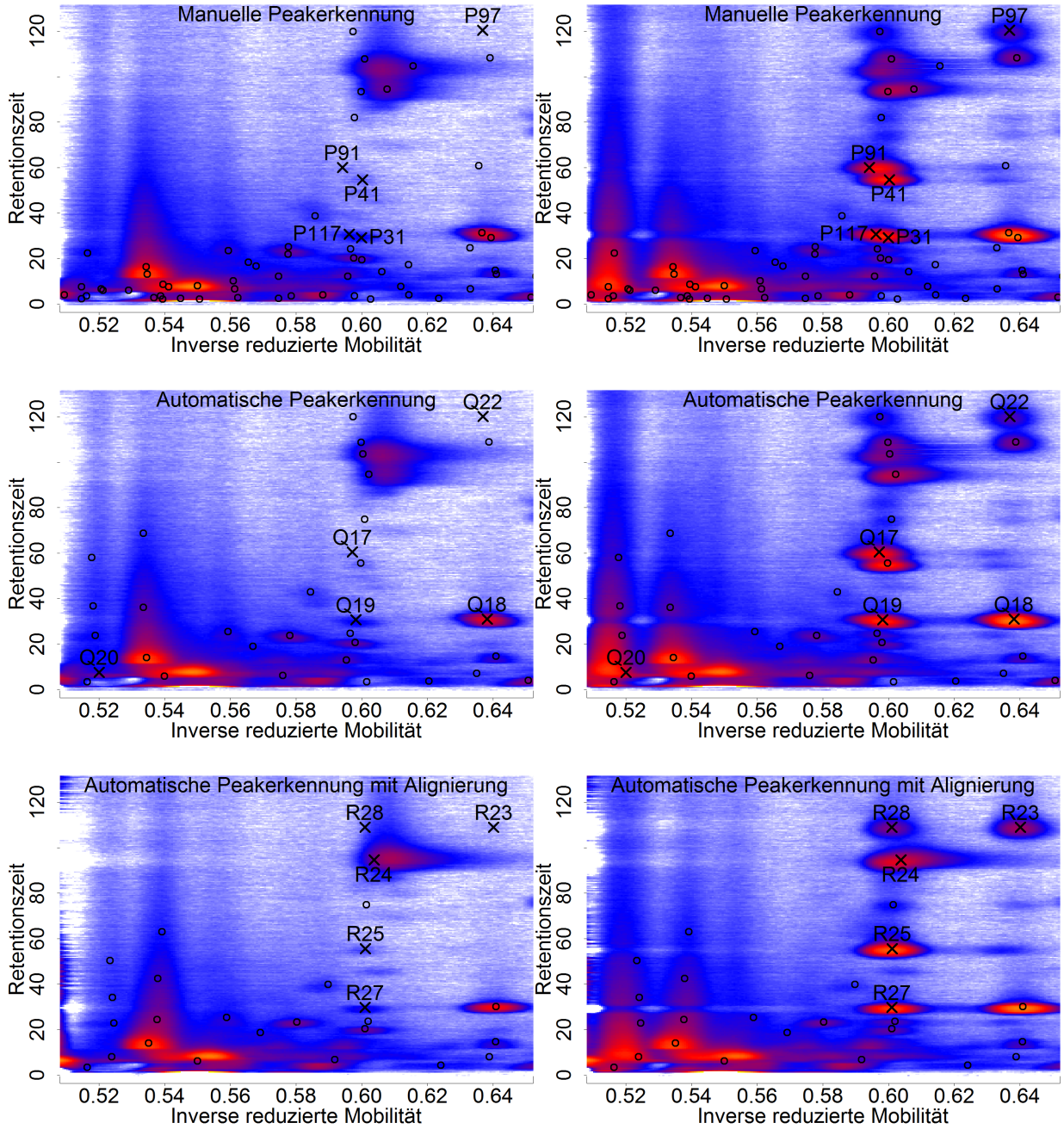


Abbildung 6.25: Ausschnitt der gemittelten Rohmessungen (für die automatische Peakerkennung mit Alignierung wurden zusätzlich die Achsen der Rohmessungen von Gerät B auf die von Gerät A aligniert, bevor die Messungen gemittelt wurden). Jeweils links sind die Messungen **ohne Saft**, jeweils rechts die Atemluftmessungen **mit Saft** gemittelt worden. Mit einem Kreuz markiert sind die Positionen der jeweils fünf Peaks mit der höchsten Variablenwichtigkeit im Saft-Klassifikationsproblem für die drei Peakerkennungsmethoden in Kombination mit Standardisierung$^+$. Die übrigen gefundenen Peaks mit niedrigeren Variablenwichtigkeits-Werten sind als Kreise dargestellt.

an denen die manuelle Peakerkennung (und teilweise auch die automatische Peakerkennung ohne die Alignierung) zwei Peaks detektieren, sieht es optisch bei der automatischen Peakerkennung nur noch wie ein Peak aus. Es liegt hier also nahe, dass die manuelle Peakerkennung und die automatische Peakerkennung ohne Alignierung fälschlicherweise zwei getrennte Peaks ausmachen. Die Rohmessungen auf den beiden Geräten sind in diesem Fall so stark verschoben, dass der Peak für den Experten der manuellen Auswertung für zwei getrennte Peaks gehalten wurde. Ebenso entstanden bei der automatischen Peakerkennung ohne Alignierung zwei getrennte Consensus Peaks. Die Verschiebung der Single Peaks vor dem Cluster-Schritt hingegen bewirkte, dass bei der automatischen Peakerkennung mit Alignierung nur noch ein Consensus Peak entstand. Eine händische Betrachtung einer Stichprobe der Rohmessungen ergab tatsächlich, dass an dem entsprechenden Bereich in Saft-Messungen immer nur ein Peak auffindbar ist, jeweils an verschobenen Positionen der beiden Geräte.

Entsprechend dieser Logik müssten diese detektierten Peaks, die jeweils nur auf einem Gerät gefunden werden, aber nahe beieinander liegen, jeweils für ein Gerät hohe Werte für Beobachtungen mit Saft und niedrige für alle Messungen des anderen Geräts aufweisen. Für den Peak Q17 der automatischen Peakerkennung ohne Alignierung bestätigt sich dies in Abbildung B.14 im Anhang auf Seite 180. Nur auf Gerät B wurde dieser Metabolit entdeckt und gleichzeitig ist er nur für die Saft-Messungen detektiert worden. Zusätzlich ist der Peak Q43 dargestellt, welcher in Abbildung 6.25 direkt rechts unterhalb von Q17 liegt. Für diesen ist es genau umgekehrt. Dieser Metabolit wird ausschließlich auf Gerät A gefunden und dort ebenfalls fast nur bei Messungen nach Saftkonsum.

Für die manuelle Peakerkennung hingegen kann dieser Effekt so nicht beobachtet werden. Im Gegenteil sind die Werte für die Saft-Messungen nicht jeweils auf einem Gerät hoch und auf dem anderen niedrig, sondern hoch positiv korreliert. In Abbildung 6.26 sind die beiden nahe beieinander liegenden Peaks der beiden Peakpaare gegeneinander aufgetragen. Die Werte sind hier unskaliert dargestellt, um zu zeigen, dass dieses Problem bereits ohne Vorverarbeitung besteht. Für das Peakpaar P31/P117 liegen die Punkte fast exakt auf der Winkelhalbierenden, für das Peakpaar P41/P91 nehmen die Werte von P91 auf Gerät A häufig etwas niedrigere Werte an als P41. Die Intensitätswerte sind nicht nur hoch korreliert, sie sind für das Peakpaar P31/P117 in 50% und für P41/P91 in 17% der Fälle sogar exakt gleich. Messungen mit Saft nehmen unabhängig vom Gerät im Allgemeinen höhere Werte an als Messungen ohne Saft. Dies passt nicht zu den vorigen Überlegungen und auch nicht zu der Beobachtung, dass in den Rohmessungen nur eines dieser beiden Peaks sichtbar ist, aber nicht beide.

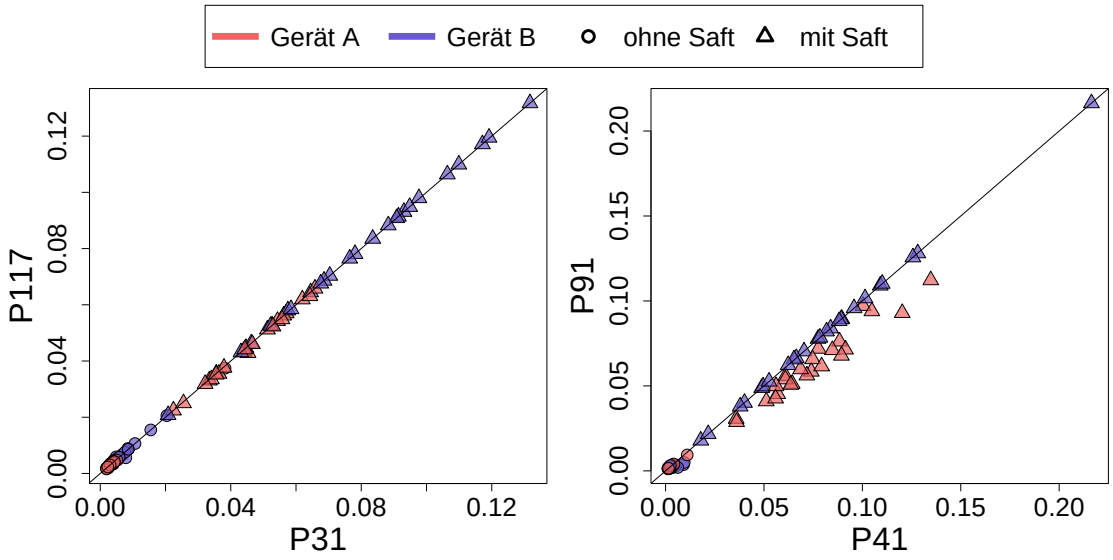


Abbildung 6.26: Korrelationen jeweils zweier nahe beieinander liegender Peaks bei der manuellen Peakerkennung. Dargestellt sind die Werte ohne Skalierung.

Die Ursache für dieses Phänomen liegt in der manuellen Peakerkennung begründet. Nach der (semi-)manuellen Bestimmung von Peakpositionen wird um jedes Peakzentrum ein Rechteck gebildet. Anschließend wird in jeder einzelnen Rohmessung das Maximum aus dem entsprechenden Rechteck dem Peak zugeordnet, unabhängig davon, ob ein Peak sichtbar ist. Die entsprechenden Rechtecke sind in Abbildung 6.27 zusätzlich zu den Peakpositionen dargestellt. Für die Peaks P31 und P117 überlappen sich die zugehörigen Rechtecke im Zentrum des Peaks. Es ist also plausibel, dass in den beiden Rechtecken häufig das gleiche Maximum gefunden und als Peakintensität festgesetzt wird. Bei den Peaks P41 und P91 überlappen sich die Rechtecke zwar nicht, dennoch decken die Rechtecke teilweise Bereiche beider Peaks ab. Liegt der Peak also beispielsweise in einer Messung an der Position von P41, so ist an der Position von P91 in Wahrheit folglich kein Peak vorhanden. Wird der Wert für diesen Peak jedoch über das Rechteck bestimmt, so ragt ein Teil des Peaks von P41 in den Rechteck-Bereich von P91 hinein, sodass dem Peak P91 trotzdem ein recht hoher Wert zugeordnet wird.

Insgesamt kann festgehalten werden, dass die manuelle Peakerkennung in solchen Fällen überlappender Rechtecke nicht in der Lage ist, die entsprechenden Peaks auseinanderzuhalten. Dies wäre auch eine Erklärung dafür, dass die Geräteunterschiede nicht so stark vorhanden sind wie bei den automatischen Peaks. Anstatt von zwei Peaks, die jeweils nur auf einem Gerät festgestellt werden, werden hier zwei Peaks mit redundanter Information gebildet. In dem Fall, dass es sich um Geräteunterschiede handelt, ist dies eher vorteilhaft, allerdings

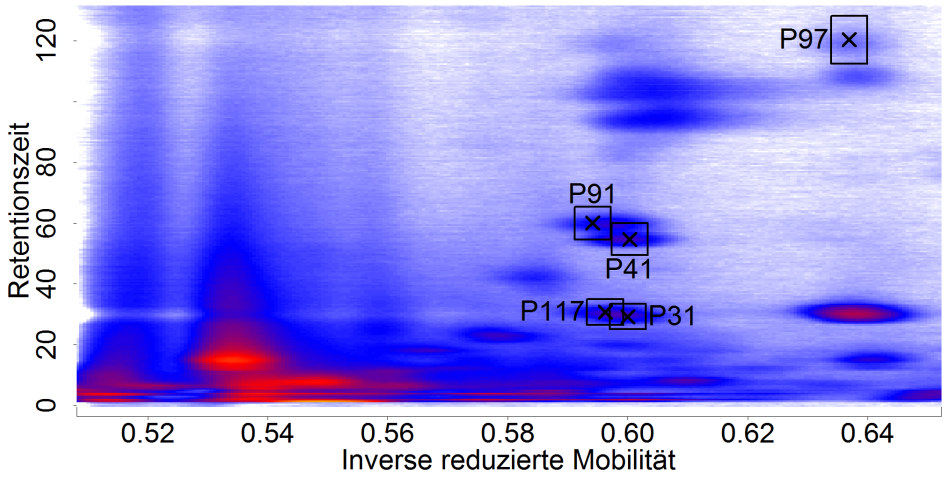


Abbildung 6.27: Ausschnitt der gemittelten Rohmessung mit den Peakpositionen der Peaks mit den fünf höchsten Variablenwichtigkeits-Werten der manuellen Peakerkennung. Zusätzlich sind die Rechtecke eingezeichnet, die für die Bestimmung des Intensitätswerts eines Peaks verwendet wurden.

zeugt dieses Phänomen von einer mangelnden Trennschärfe der manuellen Peakerkennung, die auch dazu führen kann, dass andere nahe beieinander liegende Peaks nicht voneinander unterschieden werden können.

6.6.2 Gerät

Als nächste Zielvariable wird das Gerät betrachtet. Hier ist es wünschenswert, dass das Gerät beispielsweise nach Alignierung und einer Standardisierung nicht mehr oder zumindest deutlich schlechter klassifizierbar ist, da diese Schritte unternommen wurden, um den Geräte-Effekt zu reduzieren.

Die Ergebnisse der Klassifikation bei Anwendung des Schwellenwerts 0.5 sind in Tabelle 6.7 dargestellt. Der gewünschte Effekt, dass das Gerät sich nicht oder nur schlecht klassifizieren lässt, ist für keine Kombination aus Peakerkennung und Standardisierung erzielt. Die Klassifikation ist bei einem mittleren AUC von 1 in den meisten Fällen perfekt möglich. Die Ergebnisse verändern sich bei Anwendung der Prävalenz-Methode praktisch nicht, sie sind im Anhang in Tabelle B.2 auf Seite 181 enthalten.

In Kapitel 6.5 wurde bereits erläutert, dass für einige Kombinationen auch nach Alignierung und Skalierung noch signifikante Unterschiede vorliegen. Diese Unterschiede können sich auch in der Klassifikation niederschlagen. In Abbildung 6.28 sind für die drei Peakerkennungs-

Tabelle 6.7: Gemittelte Gütemaßzahlen der kreuzvalidierten Klassifikationsergebnisse für die Zielvariable **Gerät** bei Verwendung des **Schwellenwerts 0.5**. Die Klassifikation wurde für die unskalierten Daten und für zwei Skalierungen durchgeführt, Standardisierung und Standardisierung mit separater Behandlung der Nullen (= Standardisierung$^+$).

Skalierung	AUC	ACC	TPR	TNR	PPV	NPV
Manuelle Peakerkennung						
ohne	1	1.000	1.000	1.000	1.000	1.000
Standard.	1	0.994	0.989	0.998	0.998	0.990
Standard.$^+$	1	0.993	0.987	0.999	0.999	0.987
Automatische Peakerkennung						
ohne	1	1.000	1.000	1.000	1.000	1.000
Standard.	1	1.000	1.000	1.000	1.000	1.000
Standard.$^+$	1	1.000	1.000	1.000	1.000	1.000
Automatische Peakerkennung mit Alignierung						
ohne	1	1.000	1.000	1.000	1.000	1.000
Standard.	1	1.000	1.000	1.000	1.000	1.000
Standard.$^+$	1	1.000	0.999	1.000	1.000	0.999

methoden und jeweils Standardisierung und Standardisierung$^+$ die Verteilungen der drei Peaks mit der höchsten Variablenwichtigkeit dargestellt. Da sich die beiden Skalierungen bei der manuellen Peakerkennung nicht unterscheiden, ist für diese nur eine Zeile enthalten. Bei einigen Grafiken sind für eine verbesserte Darstellung Ausreißer nach oben nicht eingezeichnet, ihre Anzahl ist neben einem Sternsymbol ablesbar. Für alle abgebildeten Peaks sind deutliche Unterschiede zwischen den Geräten erkennbar. Bei der manuellen Peakerkennung sind die Werte für Gerät B im Mittel niedriger als die Werte auf Gerät A. Bei den automatischen Peakerkennungen gibt es große Unterschiede in Abhängigkeit von der verwendeten Skalierung. Wird die Standardisierung auf Mittelwert 0 und Varianz 1 verwendet, so sind die Werte für fast alle Beobachtungen vor der Standardisierung auf beiden Geräten genau Null gewesen. Dadurch, dass dabei einzelne Werte ungleich Null vorkamen (teilweise sind das die nicht dargestellten Ausreißer) und je Gerät skaliert wurde, wurden die vorher gleichen Werte auf unterschiedliche verschoben. Dies entspricht den Erkenntnissen aus Kapitel 6.4. Auf diese Weise wurden neue Unterschiede zwischen den Geräten in den Datensatz eingeführt. Die erweiterte Standardisierung$^+$ zeigt dieses Verhalten bei den automatischen Methoden nicht. Hier sind fünf der sechs Peaks mit den hohen Werten für die Variablenwichtigkeit fast ausschließlich auf einem Gerät gefunden worden.

Abbildung 6.29 zeigt jeweils die Positionen der fünf Peaks mit den höchsten Variablenwichtigkeiten der drei Peakerkennungsmethoden bei Anwendung von Skalierung$^+$. Diese fünf Peaks kamen bei allen Skalierungen am häufigsten unter den wichtigsten fünf vor. Die übrigen gefundenen Peaks im Bildausschnitt weisen niedrigere Variablenwichtigkeiten auf und sind als Kreise dargestellt. Jeweils links sind nur die Atemluftrohmessungen von Gerät A gemittelt worden, rechts die Messungen von Gerät B. Dabei fallen tatsächlich Unterschiede zwischen den Geräten auf.

Bei der manuellen Peakerkennung werden, wie schon im vorigen Unterkapitel beobachtet, durch Verschiebungen der Rohmessungen auf beiden Geräten Peaks nahe beieinander liegende Peaks detektiert, die nur auf einem der Geräte zu sehen sind (P74/P111 auf jeweils einem Gerät sowie P87 und P90, die jeweils nur bei Gerät B zu sehen sind). Zusätzlich ist P57 unter den fünf Peaks mit der höchsten Variablenwichtigkeit, obwohl in den gemittelten Rohmessungen kein Peak zu sehen ist. Dies kann am angewendeten Filter (spaltenweises Abziehen des 20%-Quantils) liegen, der in der Darstellung eventuell vorhandene Unterschiede entfernt haben könnte.

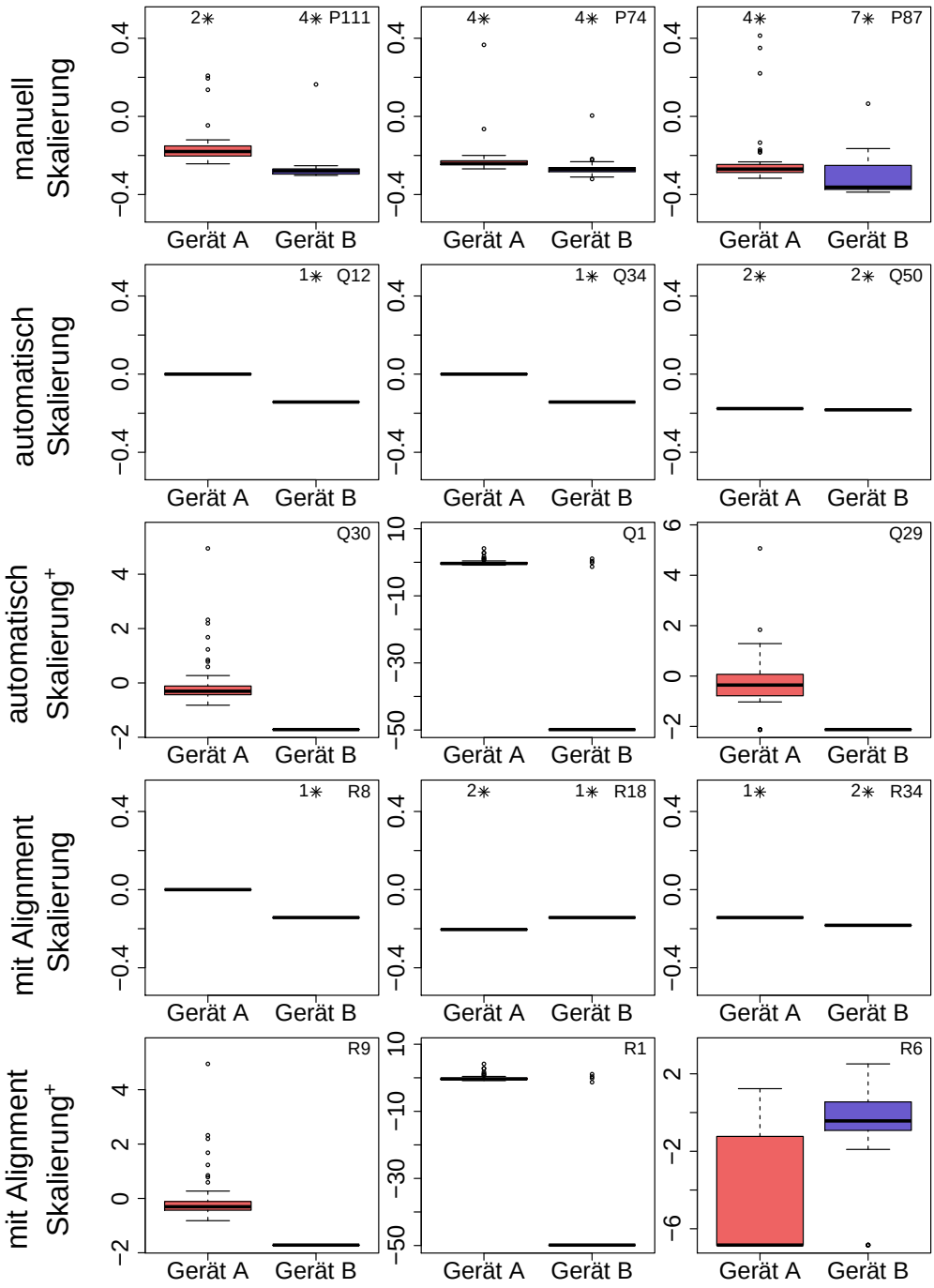


Abbildung 6.28: Boxplots der jeweils drei Variablen mit den höchsten Variablenwichtigkeiten beim Klassifikationsproblem **Gerät** für die Datensätze der drei Peakerkennungen mit Standardisierung und Standardisierung$^+$ (bei der manuellen Peakerkennung unterschieden sich diese Skalierungen nicht). Zur besseren Darstellbarkeit sind in einigen Grafiken Ausreißer nach oben nicht eingezeichnet. Sie sind mit dem Stern-Symbol gekennzeichnet. Die nebenstehende Zahl gibt die Anzahl der nicht dargestellten Ausreißer an.

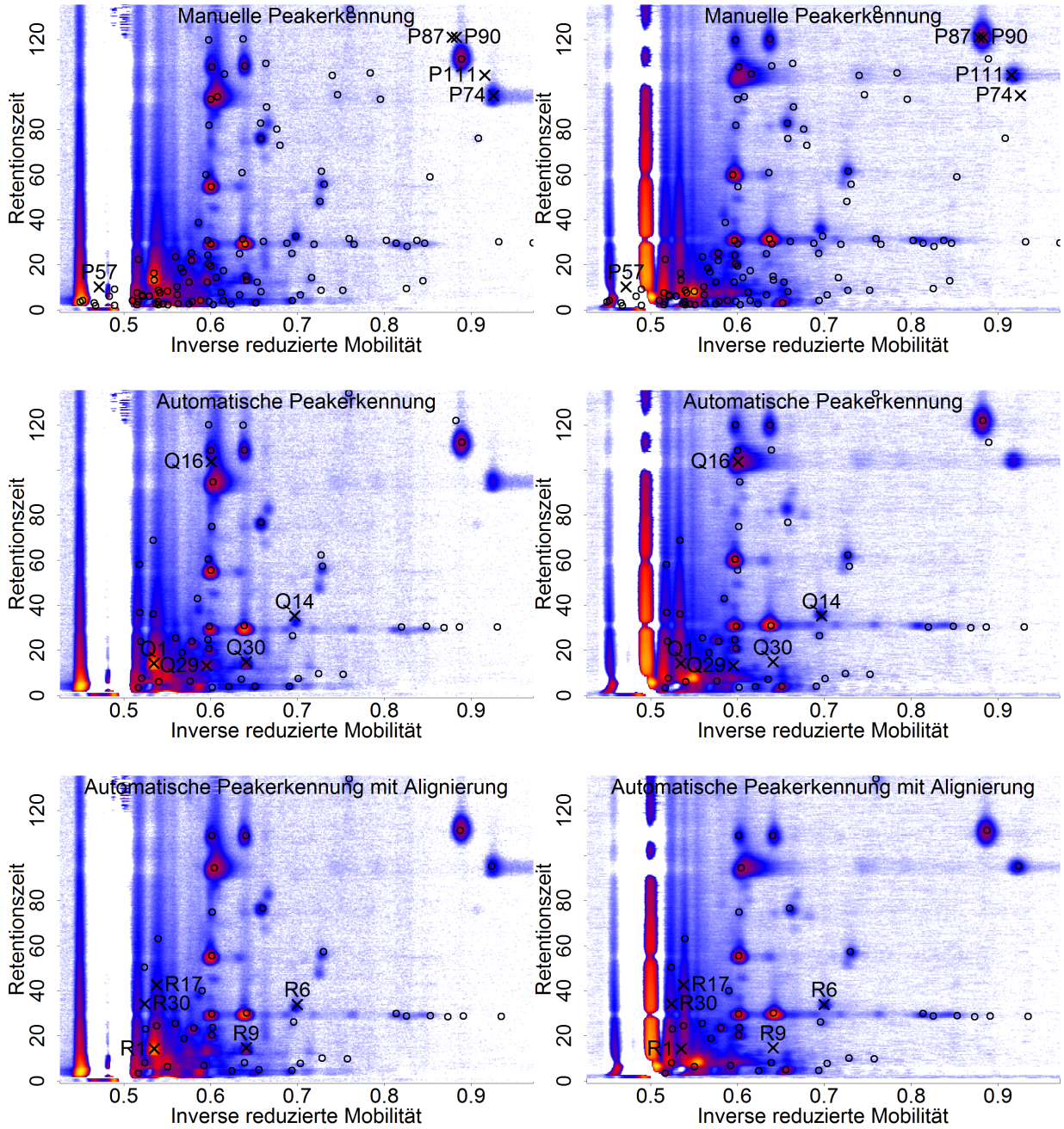


Abbildung 6.29: Ausschnitt der gemittelten Rohmessungen (für die automatische Peakerkennung mit Alignierung wurden zusätzlich die Achsen der Rohmessungen von Gerät B auf die von Gerät A aligniert, bevor die Messungen gemittelt wurden). Jeweils links sind die Atemluftmessungen von **Gerät A**, jeweils rechts die Messungen von **Gerät B** gemittelt worden. Mit einem Kreuz markiert sind die Positionen der jeweils fünf Peaks mit der höchsten Variablenwichtigkeit im Geräte-Klassifikationsproblem für die drei Peakerkennungsmethoden in Kombination mit Standardisierung$^+$. Die übrigen gefundenen Peaks mit niedrigeren Variablenwichtigkeits-Werten sind als Kreise dargestellt.

Bei den automatischen Peakerkennungen sind andere Bereiche der Rohmessungen relevant für den Geräteunterschied. Bei der automatischen Peakerkennung ohne Alignierung tritt ebenfalls ein Peak (Q16) auf, der durch die Verschiebung nur auf einem Gerät (Gerät A) zu sehen ist. Bei der automatischen Peakerkennung mit Alignierung ist dies nicht mehr der Fall. Die Positionen Q30/R9, welche den gleichen Peak beschreiben, sind nur im linken Bild, also bei Gerät A, zu sehen. Hierbei könnte es sich also um einen gerätespezifischen Peak handeln. An den Positionen von Q1/R1 liegt ebenfalls bei Gerät A ein sehr großer Peak vor. Bei Gerät B ist dieser deutlich schmaler und liegt nach der Alignierung (Bild unten rechts) auch etwas weiter rechts. Bei den Positionen Q14/R6 ist auf beiden Geräten deutlich ein Peak zu sehen. Für Gerät B waren für R6 jedoch auch viele Werte Null (vgl. Abbildung 6.28), sodass der Peak möglicherweise bei einigen Rohmessungen übersehen wurde oder nicht in allen enthalten ist. Optisch ist für diese Peaks nicht zu erkennen, dass tatsächlich Unterschiede zwischen den Geräten vorliegen. Die Peaks R17 und R30 beschreiben Peaks, die vor allem bei Gerät B sichtbar sind. Dort sieht es allerdings so aus, als wären diese Peaks möglicherweise keine neuen Peaks sondern Teile jeweils eines über viele Retentionszeiten langgestreckten Peaks. Dieser ist bei einer Retentionszeit von etwa 30 „eingeschnürt“, also scheinbar unterbrochen, sodass die Peakerkennungsmethoden zwei getrennte Peaks ausmachen könnten. Die Einschnürungen werden durch große Peaks bei der entsprechenden Retentionszeit hervorgerufen und sind auch beim RIP beobachtbar (beispielsweise auch bei RT$\approx$ 30 oder RT$\approx$ 55 sichtbar). Dieses Phänomen wurde auch von Egorov u. a. (2014) beschrieben. Diese Einschnürungen erschweren zusätzlich die Peakerkennung.

Insgesamt zeigen diese Beobachtungen, dass die Geräteunterschiede sowohl durch Defizite in der Peakerkennung als auch durch tatsächliche Unterschiede zwischen den Geräten verursacht werden können.

Die Differenzen der einzelnen Beobachtungen der Klassifikationswahrscheinlichkeiten für die korrekte Klasse zum Schwellenwert 0.5 sind in den Abbildungen B.15 bis B.17 im Anhang auf den Seiten 182 bis 184 dargestellt. Für die manuelle Peakerkennung (Abbildung B.15) ist ersichtlich, dass die Differenzen zum Schwellenwert bei den skalierten Werten deutlich kleiner werden, die Klassifikation also einer größeren Unsicherheit unterliegt. Allerdings kommt es nur bei drei Beobachtungen in sehr wenigen Kreuzvalidierungswiederholungen zu Fehlklassifizierungen. Hierbei ist zu beachten, dass Standardisierung und Standardisierung$^+$ identische Datensätze sind und die Unterschiede zwischen den beiden Grafiken (Mitte und unten) ausschließlich auf die Variabilität des Random Forests zurückzuführen sind. Dass es dabei nur zu minimalen Unterschieden kommt, verdeutlicht die Stabilität des Random Forests. Für die

Tabelle 6.8: Gemittelte Gütemaßzahlen der kreuzvalidierten Klassifikationsergebnisse für die Zielvariable **Geschlecht** bei Verwendung des **Schwellenwerts 0.5**. Die Klassifikation wurde für die unskalierten Daten und für zwei Skalierungen durchgeführt, Standardisierung und Standardisierung mit separater Behandlung der Nullen (= Standardisierung$^+$).

Skalierung	AUC	ACC	TPR	TNR	PPV	NPV
Manuelle Peakerkennung						
ohne	0.402	0.427	0.492	0.353	0.461	0.379
Standard.	0.426	0.462	0.557	0.355	0.494	0.413
Standard.$^+$	0.435	0.471	0.557	0.375	0.500	0.429
Automatische Peakerkennung						
ohne	0.534	0.500	0.541	0.454	0.529	0.466
Standard.	0.483	0.471	0.526	0.410	0.502	0.432
Standard.$^+$	0.525	0.487	0.545	0.423	0.516	0.452
Automatische Peakerkennung mit Alignierung						
ohne	0.571	0.529	0.599	0.451	0.553	0.498
Standard.	0.509	0.503	0.572	0.424	0.530	0.465
Standard.$^+$	0.585	0.549	0.592	0.501	0.574	0.521

beiden automatischen Peakerkennungsmethoden (Abbildungen B.16 und B.17) sind die Differenzen zum Schwellenwert ohne Skalierung und mit der einfachen Standardisierung sehr groß, sodass es keine Fehlklassifikationen gibt. Bei der Anwendung von Standardisierung$^+$ sinken auch hier die Differenzen, es kommt allerdings lediglich bei der Peakerkennung mit Alignierung bei einer einzigen Beobachtung zu zwei Fehlklassifizierungen. Da sich die Ergebnisse für den Schwellenwert nach Prävalenzmethode nicht stark von denen mit festem Schwellenwert 0.5 unterscheiden, werden die entsprechenden Abbildungen hier nicht aufgeführt.

6.6.3 Geschlecht

Für die Zielvariable Geschlecht sind zuvor keine univariat signifikanten Metaboliten gefunden worden. Es ist jedoch möglich, dass ein multivariates Klassifikationsverfahren dennoch eine sinnvolle Klassifikationsregel aufstellen kann. Die Gütemaßzahlen für diese Problemstellung unter Verwendung des Schwellenwertes 0.5 finden sich in Tabelle 6.8. Die Klassifikation ist

Tabelle 6.9: Gemittelte Gütemaßzahlen der kreuzvalidierten Klassifikationsergebnisse für die Zielvariable **Rauchen** bei Verwendung des **Schwellenwerts 0.5**. Die Klassifikation wurde für die unskalierten Daten und für zwei Skalierungen durchgeführt, Standardisierung und Standardisierung mit separater Behandlung der Nullen (= Standardisierung$^+$).

Skalierung	AUC	ACC	TPR	TNR	PPV	NPV
Manuelle Peakerkennung						
ohne	0.501	0.520	0.502	0.524	0.191	0.825
Standard.	0.532	0.534	0.509	0.539	0.199	0.831
Standard.$^+$	0.541	0.538	0.517	0.543	0.204	0.833
Automatische Peakerkennung						
ohne	0.542	0.522	0.544	0.516	0.202	0.835
Standard.	0.529	0.523	0.528	0.522	0.199	0.831
Standard.$^+$	0.534	0.524	0.523	0.524	0.198	0.830
Automatische Peakerkennung mit Alignierung						
ohne	0.567	0.530	0.592	0.516	0.216	0.849
Standard.	0.596	0.552	0.607	0.540	0.230	0.859
Standard.$^+$	0.620	0.578	0.600	0.573	0.242	0.864

unabhängig von der Peakerkennungsmethode und unabhängig von der verwendeten Skalierung nicht erfolgreich. Für die manuelle Peakerkennung liegen die AUC-Werte sogar unterhalb von 0.5, für die anderen schwanken sie um 0.5, was für eine zufällige Klassifizierungsregel spricht. Für die automatische Peakerkennung mit Alignierung und Standardisierung$^+$ wird ein AUC von fast 0.59 erreicht, allerdings liegt die Accuracy nur bei etwa 0.55. Dieses Ergebnis kann durch zufällige Schwankungen entstanden sein. Die Verwendung des Schwellenwerts mit der Prävalenzmethode führt zu keinen großen Änderungen. Die zugehörige Tabelle B.3 ist im Anhang auf Seite 185 aufgeführt.

6.6.4 Rauchen

Für die Zielvariable Rauchen wurden Anpassungen am Aufbau der Kreuzvalidierung vorgenommen. Diese wurden zu Beginn dieses Kapitels erläutert. Die Klassifikationsergebnisse unter Verwendung des Schwellenwertes 0.5 sind in Tabelle 6.9 dargestellt. Die Klassifikation

des Raucherstatus ist für die verschiedenen Peakerkennungsmethoden und Skalierungen nicht erfolgreich. Die mittleren AUC-Werte liegen etwa bei 0.5, für die automatische Peakerkennung mit Alignierung und Standardisierung$^+$ wird ein AUC von 0.62 erzielt. PPV und NPV entsprechen jeweils ungefähr der relativen Häufigkeiten der positiven beziehungsweise negativen Klasse. Dieses Ergebnis wird dadurch erzielt, dass in den jeweiligen Trainingsdatensätzen beide Klassen gleich stark vertreten waren und bei einer nicht erfolgreichen Klassifikation die Label etwa entsprechend ihrer relativen Häufigkeiten im Trainingsdatensatz vergeben werden. Hier werden also beide Label ungefähr mit Wahrscheinlichkeit 0.5 an die Beobachtungen vergeben. Daher entsprechen auch Sensitivität, Spezifität und Accuracy im Testdatensatz in etwa 0.5. Der positive prädiktive Wert ist der Anteil der positiv klassifizierten Beobachtungen, die in Wahrheit der positiven Klasse angehören. Sei p der Anteil positiver Beobachtungen im Testdatensatz und n die Gesamtanzahl der Beobachtungen. Dann werden im Schnitt $n \cdot p \cdot 0.5$ Beobachtungen richtigerweise der positiven Klasse zugeordnet, insgesamt werden etwa $n \cdot 0.5$ Beobachtungen der positiven Klasse zugeordnet. Als PPV ergibt sich dann $\frac{n \cdot p \cdot 0.5}{n \cdot 0.5} = p$ und entspricht damit der relativen Häufigkeit der Beobachtungen im Testdatensatz. Analog gilt dies für den NPV. Die hier beobachteten Werte für PPV und NPV ergeben sich also aufgrund der Häufigkeitszusammensetzung der Testdatensätze. Die Ergebnisse bei Verwendung der Schwellenwerte nach der Prävalenzmethode unterscheiden sich kaum von den Ergebnissen bei festem Schwellenwert 0.5, da der mittlere bestimmte Schwellenwert dabei ebenfalls sehr nahe an 0.5 liegt und zusätzlich die Klassengröße im Trainingsdatensatz bereits durch das Undersampling angeglichen wurde. Die Ergebnisse sind in Tabelle B.4 im Anhang auf Seite 186 dargestellt.

Der AUC-Wert von 0.62 ist sehr niedrig, könnte jedoch ein Hinweis auf Variablen mit zumindest geringem Informationsgehalt sein. Da in der Gruppe der Rauchenden auch die (Ex-)Rauchenden enthalten sind, muss ohnehin davon ausgegangen werden, dass das Klassifikationsproblem dadurch erschwert ist und möglicherweise nur die Subgruppe der tatsächlich noch aktiv Rauchenden klassifizierbar ist. Um die Ursache des leicht über 0.5 liegenden AUC-Werts zu ergründen, sind die drei wichtigsten Variablen der Klassifikation für die Kombination der automatischen Peakerkennung mit Alignierung und mit Skalierung$^+$ in Abbildung 6.30 dargestellt. Es sind bei allen drei Peaks keine großen Unterschiede zwischen den beiden Gruppen zu sehen, es sind stets kleine Mengen an Beobachtungen, die sich von übrigen Beobachtungen trennen (bei R4 einige hohe Werte für die Nichtrauchenden, bei R21 einige hohe und niedrige Werte für die Nichtrauchenden, bei R38 ein paar mittige Beobachtungen der (Ex-)Rauchenden). Insgesamt sehen diese Abweichungen eher zufällig aus, beziehungsweise sind die Auffälligkeiten inhaltlich unintuitiv. Dass für Nichtrauchende

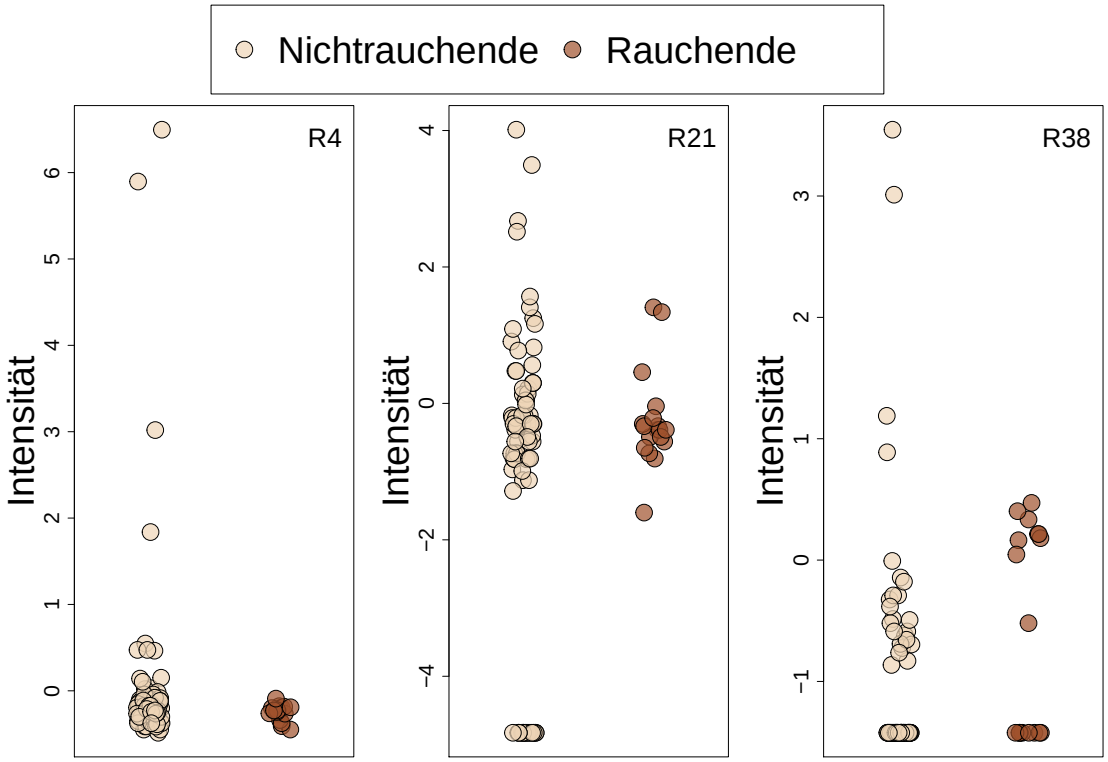


Abbildung 6.30: Beobachtungen der drei Variablen mit den höchsten Variablenwichtigkeiten beim Klassifikationsproblem Rauchen für die automatische Peakerkennung mit Alignierung und Standardisierung^{+}.

extremere Werte angenommen werden und für (Ex-)Rauchende niedrige oder gar mittlere, erscheint unplausibel. Intuitiv wäre davon auszugehen, dass bei den Rauchenden ein Stoff zusätzlich in der Atemluft enthalten ist und somit höhere Werte für relevante Metaboliten erzielt werden. Theoretisch wäre es auch denkbar, dass bei (Ex-)Rauchenden ein metabolischer Prozess inhibitiert wird, sodass niedrigere Werte als bei den Nichtrauchenden auftreten, mittlere Werte hingegen erscheinen biologisch zumindest nicht intuitiv. Da deutlich mehr Nichtrauchende als (Ex-)Rauchende enthalten sind, ist es eher plausibel, dass die Extremwerte häufiger von diesen Beobachtungen angenommen werden. Da die Effekte insgesamt auch nur sehr gering ausgeprägt sind, ist hier nicht davon auszugehen, dass den dargestellten Metaboliten eine Bedeutung bei der Unterscheidung von Nichtrauchenden und (Ex-)Rauchenden beizumessen ist.

6.7 Auswirkung der Vernachlässigung des Geräte-Effekts

Im folgenden Abschnitt wird demonstriert, wie sich die Vernachlässigung des Geräte-Effekts auf die Klassifikation auswirken kann. Zu diesem Zweck werden die Daten jeweils eines Gerätes als Trainingsdatensatz und die des anderen Gerätes als Testdatensatz verwendet. Dies ist ein realistisches Szenario, wenn an zwei unterschiedlichen Standorten jeweils ein Gerät vorhanden ist und zunächst an einem Standort Daten erhoben und ein Klassifikationsmodell aufgestellt werden und zu einem späteren Zeitpunkt neue Daten des anderen Standorts übermittelt werden, welche anschließend mit Hilfe des Modells ausgewertet werden sollen.

Wird der Geräte-Effekt vollkommen außer Acht gelassen, so kann die Auswertung von den aus dieser Arbeit vorgestellten Methoden nur mit der manuellen Peakerkennung und der automatischen Peakerkennung ohne Alignierung und jeweils ohne Skalierung angewendet werden. Der Einfachheit halber werden die Daten beider Peakerkennungen so verwendet, wie sie in den vorigen Abschnitten gebildet wurden, das heißt, die Peakerkennung basiert jeweils auf allen Daten und wird nicht erneut nur für die Daten eines Gerätes durchgeführt. Die Ergebnisse der Klassifikation für den festen Schwellenwert 0.5 sind in Tabelle 6.10 dargestellt. Im Unterschied zu den Ergebnissen zuvor wurde hier keine Kreuzvalidierung durchgeführt, da in dieser Situation genau ein Trainings- und ein Testdatensatz vorliegen. Zudem ist in diesen Ergebnissen der AUC-Wert nicht enthalten, da der Schwellenwert in der konkreten Anwendung auf dem Testdatensatz nicht mehr variiert/optimiert werden darf sondern fix gewählt oder auf dem Trainingsdatensatz bestimmt werden muss (vgl. Tabelle 6.11, für welche

Tabelle 6.10: Gütemaßzahlen der Klassifikationsergebnisse für die Zielvariable **Saft**, wenn auf einem Gerät trainiert und auf dem anderen Gerät vorhergesagt wird. Der verwendete **Schwellenwert** ist **0.5**. Die Klassifikation wurde für die unskalierten Daten und für zwei Skalierungen durchgeführt, Standardisierung und Standardisierung mit separater Behandlung der Nullen (= Standardisierung$^+$). Die Kombinationen, welche den Geräte-Effekt vollständig vernachlässigen, sind grau unterlegt.

Skalierung	ACC	TPR	TNR	PPV	NPV
Trainingsgerät: A					
manuell					
ohne Skalierung	0.980	1.000	0.958	0.962	1.000
Standardisierung	0.959	0.920	1.000	1.000	0.923
Standardisierung$^+$	0.959	0.920	1.000	1.000	0.923
automatisch					
ohne Skalierung	0.490	0.000	1.000	-	0.490
Standardisierung	0.510	1.000	0.000	0.510	-
Standardisierung$^+$	0.490	0.000	1.000	-	0.490
mit Alignierung					
ohne Skalierung	0.959	0.920	1.000	1.000	0.923
Standardisierung	0.959	0.920	1.000	1.000	0.923
Standardisierung$^+$	0.959	0.920	1.000	1.000	0.923
Trainingsgerät: B					
manuell					
ohne Skalierung	0.980	0.958	1.000	1.000	0.962
Standardisierung	0.980	1.000	0.960	0.960	1.000
Standardisierung$^+$	0.980	1.000	0.960	0.960	1.000
automatisch					
ohne Skalierung	0.510	0.000	1.000	-	0.510
Standardisierung	0.490	1.000	0.000	0.490	-
Standardisierung$^+$	0.510	0.000	1.000	-	0.510
mit Alignierung					
ohne Skalierung	1.000	1.000	1.000	1.000	1.000
Standardisierung	1.000	1.000	1.000	1.000	1.000
Standardisierung$^+$	1.000	1.000	1.000	1.000	1.000

der Schwellenwert auf dem Trainingsdatensatz mit der Prävalenzmethode bestimmt wurde). In den beiden Tabellen sind die Kombinationen aus Peakerkennungsmethode und Skalierung, welche den Geräte-Effekt vernachlässigen, grau hinterlegt.

Für den fixen Schwellenwert (Tabelle 6.10) gelingt die Klassifikation des Saft-Problems für die manuelle Peakerkennung, unabhängig davon, ob auf Gerät A oder auf Gerät B trainiert wurde, beinahe perfekt. Die Accuracy liegt fast bei 1. Lediglich die TNR, knapp unterhalb von 1, weist darauf hin, dass wenige Beobachtungen der Saft-Gruppe zugeordnet werden, obwohl kein Saft getrunken wurde. Die automatische Peakerkennung ohne Skalierung hingegen schlägt komplett fehl, die Accuracy beträgt ungefähr 0.5, eine TPR von 0 und eine TNR von 1 lassen erkennen, dass in diesem Fall alle Beobachtungen der negativen Klasse (ohne Saft) zugeordnet wurden. Aus diesem Grund kann der PPV nicht berechnet werden, der NPV liegt nahe bei 0.5, da nur die Hälfte der negativ klassifizierten Beobachtungen tatsächlich negativ ist. Auch dies ist unabhängig davon, auf welchem Gerät trainiert wurde. Die Anwendung von Standardisierungen bringt für die automatische Peakerkennung ohne Alignierung keinen Vorteil, auch für diese Fälle ist die Klassifikationsregel nicht sinnvoll. Die Standardisierung führt dazu, dass im Gegensatz zur Variante ohne Skalierung oder Standardisierung$^+$ alle Beobachtungen der positiven Klasse zugeordnet werden, sodass sich die entsprechenden Maßzahlen umkehren. Die Klassifikation der Daten der manuellen Peakerkennung verbessern sich durch die Standardisierungen nicht. Die automatische Peakerkennung mit Alignierung erzielt unabhängig von der verwendeten Skalierung, ähnlich wie die manuelle Peakerkennung, fast perfekte Klassifikationsergebnisse.

Die Maßzahlen, welche sich durch die Anwendung der Prävalenzmethode ergeben, sind in Tabelle 6.11 dargestellt. Für die automatische Peakerkennung mit Alignierung ergeben sich keinerlei Änderungen, für die manuelle Peakerkennung mit Standardisierung oder Standardisierung$^+$ verschlechtert sich die Klassifikation leicht (sinkende TPR bei gleichbleibender FPR) gegenüber der Klassifikation mit fixem Schwellenwert 0.5. Für die automatische Peakerkennung ohne Alignierung bleiben die Ergebnisse unverändert, wenn Gerät A das Trainingsgerät ist. Ist hingegen Gerät B das Trainingsgerät, so verbessern sich die Ergebnisse für die Variante ohne Skalierung und mit Standardisierung$^+$, nicht hingegen für die gewöhnliche Standardisierung. Dabei werden Accuracy-Werte von etwa 0.78 (ohne Skalierung) bzw. 0.88 (bei Standardisierung$^+$) erreicht. Die gewählten Schwellenwerte liegen dabei mit jeweils 0.375 deutlich niedriger als 0.5.

Tabelle 6.11: Gütemaßzahlen der Klassifikationsergebnisse für die Zielvariable **Saft**, wenn auf einem Gerät trainiert und auf dem anderen Gerät vorhergesagt wird. Der verwendete Schwellenwert wurde mit der **Prävalenz-Methode** bestimmt. Die Klassifikation wurde für die unskalierten Daten und für zwei Skalierungen durchgeführt, Standardisierung und Standardisierung mit separater Behandlung der Nullen (= Standardisierung$^+$). Die Kombinationen, welche den Geräte-Effekt vollständig vernachlässigen, sind grau unterlegt.

Skalierung	ACC	TPR	TNR	PPV	NPV	Schwellenwert
Trainingsgerät: A						
manuell						
ohne Skalierung	0.980	1.000	0.958	0.962	1.000	0.579
Standardisierung	0.939	0.880	1.000	1.000	0.889	0.633
Standardisierung$^+$	0.939	0.880	1.000	1.000	0.889	0.628
automatisch						
ohne Skalierung	0.490	0.000	1.000	-	0.490	0.554
Standardisierung	0.510	1.000	0.000	0.510	-	0.520
Standardisierung$^+$	0.490	0.000	1.000	-	0.490	0.550
mit Alignierung						
ohne Skalierung	0.959	0.920	1.000	1.000	0.923	0.540
Standardisierung	0.959	0.920	1.000	1.000	0.923	0.547
Standardisierung$^+$	0.959	0.920	1.000	1.000	0.923	0.559
Trainingsgerät: B						
manuell						
ohne Skalierung	1.000	1.000	1.000	1.000	1.000	0.407
Standardisierung	0.980	1.000	0.960	0.960	1.000	0.406
Standardisierung$^+$	0.980	1.000	0.960	0.960	1.000	0.413
automatisch						
ohne Skalierung	0.776	0.542	1.000	1.000	0.694	0.375
Standardisierung	0.490	1.000	0.000	0.490	-	0.338
Standardisierung$^+$	0.878	0.750	1.000	1.000	0.806	0.375
mit Alignierung						
ohne Skalierung	1.000	1.000	1.000	1.000	1.000	0.376
Standardisierung	1.000	1.000	1.000	1.000	1.000	0.339
Standardisierung$^+$	1.000	1.000	1.000	1.000	1.000	0.394

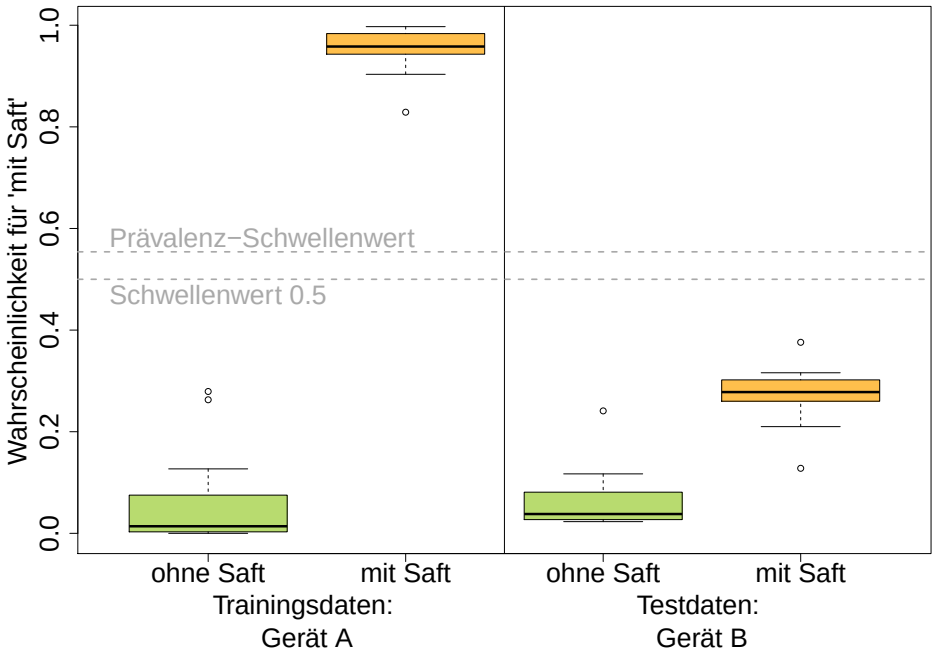


Abbildung 6.31: Wahrscheinlichkeiten für die positive Klasse (mit Saft), wenn das Klassifikationsmodell ohne Berücksichtigung des Geräte-Effekts nur auf einem Gerät (Gerät A) trainiert wird und auf die Beobachtungen des anderen Geräts (Gerät B) angewendet wird. Die Wahrscheinlichkeiten sind getrennt nach wahrer Klasse dargestellt, jeweils für den Trainingsdatensatz (links) und den Testdatensatz (rechts).

Die Ursache für die Fehlklassifikationen der automatischen Peakerkennung ohne Alignierung wird im Folgenden am Beispiel der Daten ohne Skalierung näher erläutert. Die prognostizierten Klassifikationswahrscheinlichkeiten für die Beobachtungen der beiden Klassen sind für das Modell, das auf den Werten von Gerät A trainiert wurde, in Abbildung 6.31 dargestellt. Links sind die auf den Trainingsdaten vorhergesagten Wahrscheinlichkeiten dargestellt, rechts die auf den Testdaten vorhergesagten Wahrscheinlichkeiten. Die beiden verwendeten Schwellenwerte sind durch horizontale Linien markiert. Es ist deutlich zu erkennen, dass die Klassifikation auf den Trainingsdaten perfekt gelingt. Aus diesem Grund erzielen auf den Trainingsdaten viele mögliche Schwellenwerte perfekte Resultate. Der fixe Wert 0.5, sowie der mittig zwischen den realisierten Wahrscheinlichkeiten liegende Schwellenwert nach der Prävalenz-Methode, erzielen dieses Ergebnis gleichermaßen. Auf den Testdaten hingegen liegen alle realisierten Wahrscheinlichkeiten unterhalb dieser beiden Schwellenwerte, sodass sich die Maßzahlen wie in den zuvor beschriebenen Tabellen ergeben. Interessanterweise sind die Beobachtungen des Testdatensatzes anhand ihrer realisierten Wahrscheinlichkeiten theoretisch ebenfalls recht gut trennbar, jedoch müsste der Schwellenwert hierfür deutlich

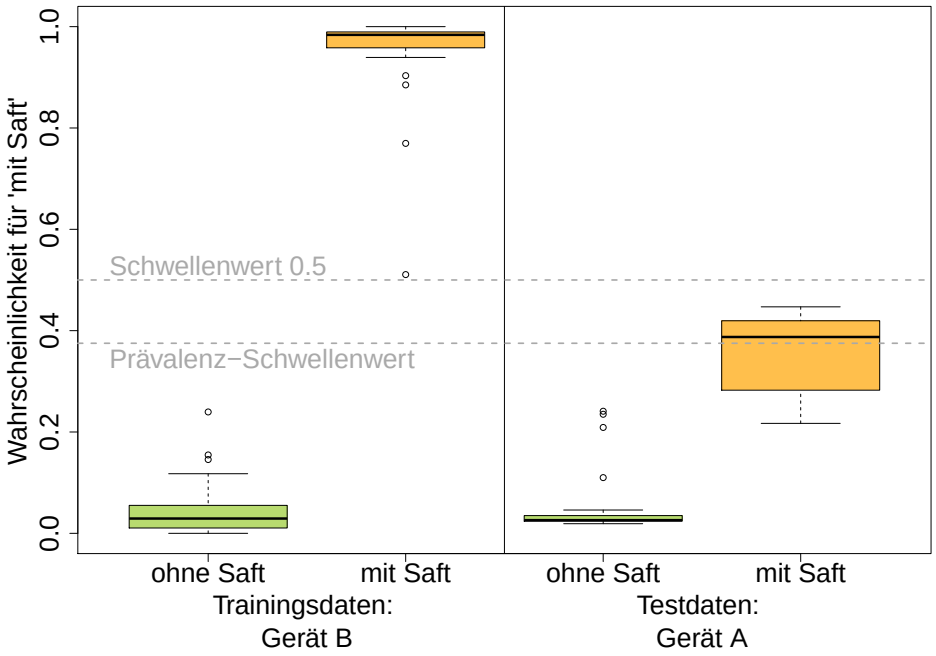


Abbildung 6.32: Wahrscheinlichkeiten für die positive Klasse (mit Saft), wenn das Klassifikationsmodell ohne Berücksichtigung des Geräte-Effekts nur auf einem Gerät (Gerät B) trainiert wird und auf die Beobachtungen des anderen Geräts (Gerät A) angewendet wird. Die Wahrscheinlichkeiten sind getrennt nach wahrer Klasse dargestellt, jeweils für den Trainingsdatensatz (links) und den Testdatensatz (rechts).

niedriger gewählt werden. Ein derartiger Schwellenwert war anhand der Trainingsdaten jedoch nicht plausibel. Dieses Ergebnis zeigt deutlich, dass die Testdaten wesentliche Informationen über das Klassifikationsproblem enthalten, diese mit Hilfe des Modells auf den Trainingsdaten jedoch nicht angemessen genutzt werden können.

Analog sind die entsprechenden Wahrscheinlichkeiten, wenn Gerät B als Trainingsgerät verwendet wird, in Abbildung 6.32 dargestellt. Auf den Trainingsdaten zeigt sich ein vergleichbares Bild, die Klassifikation erfolgt auch hier bei perfekter Trennung. Der fixe Schwellenwert 0.5 führt fast zu einer Fehlklassifikation auf den Trainingsdaten, wohingegen der Prävalenz-Schwellenwert auch hier per Definition mittig zwischen den extremsten realisierten Wahrscheinlichkeiten der beiden Gruppen liegt. Dies erscheint plausibel, könnte jedoch auch eine Überanpassung darstellen, da der Schwellenwert sich durch eine sehr niedrige Wahrscheinlichkeit der Saft-Klasse stark nach unten verschiebt. Wenn die Wahrscheinlichkeiten der Saft-Klasse generell stärker streuen als die der Klasse ohne Saft, so wäre dieser nach unten verschobene Schwellenwert sinnvoll, handelt es sich hierbei jedoch um einen Ausreißer,

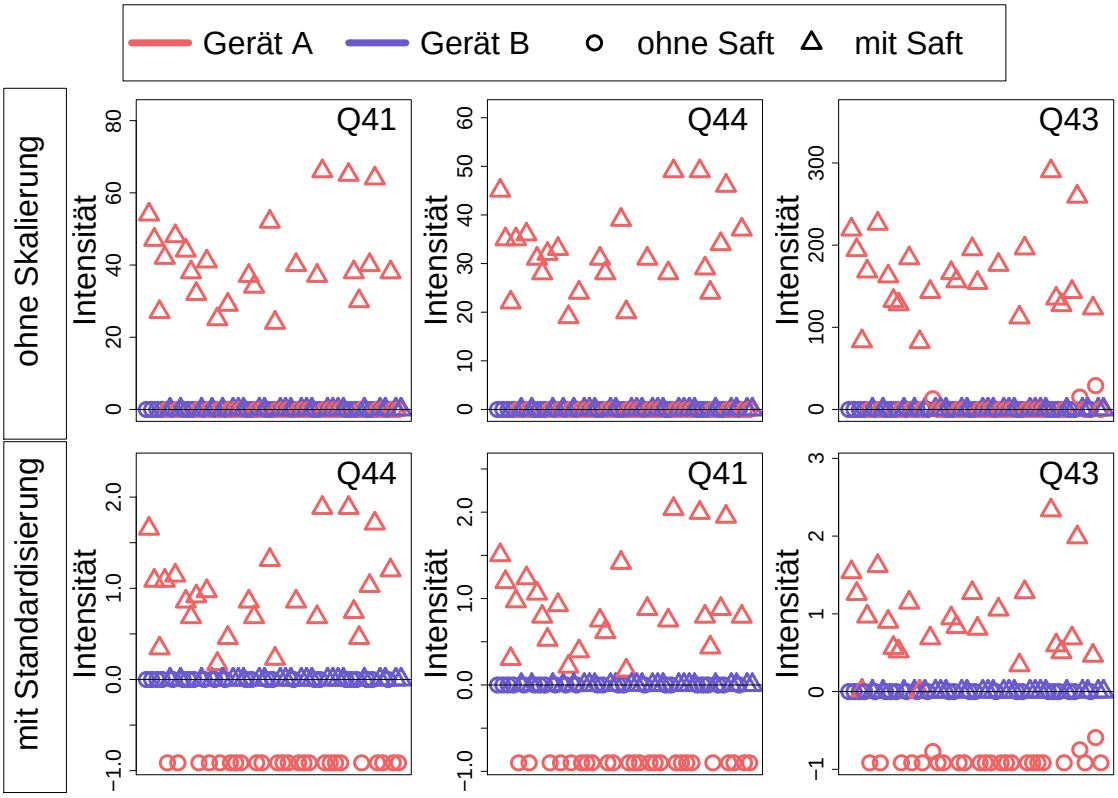


Abbildung 6.33: Beobachtungen der drei Variablen mit den höchsten Variablenwichtigkeiten beim Klassifikationsproblem Saft für die automatische Peakerkennung ohne Alignierung und ohne Skalierung (oben) sowie mit Standardisierung (unten), wenn nur auf Gerät A trainiert wurde.

so wäre die Wahl des festen Schwellenwerts sinnvoller. Diese Frage kann hier aufgrund des geringen Stichprobenumfangs nicht abschließend beantwortet werden. Der etwas niedrigere Schwellenwert führt dazu, dass auf dem Testdatensatz einige Beobachtungen der Saft-Klasse korrekt klassifiziert werden und sich die Maßzahlen in der zuvor beschriebenen Tabelle entsprechend verbessern. Dies kann jedoch Zufall (durch die Überanpassung) sein und ist kein sicheres Zeichen für eine Überlegenheit der Prävalenz-Methode in diesem Fall, insbesondere da dieser Effekt bei Vertauschung der Geräte (Abbildung 6.31) nicht zu beobachten war.

Die Ursache für die hohen Fehlklassifikationsraten im Fall der automatischen Peakerkennung ohne Alignierung und die umgekehrten Maßzahlen, wenn die Daten ohne Skalierung und bei Anwendung der Standardisierung verwendet werden, werden im Folgenden noch weiter untersucht. Zu diesem Zweck sind in Abbildung 6.33 beispielhaft für den Fall, dass nur auf Gerät A trainiert wurde, die jeweils drei wichtigsten Peaks für die unskalierten Daten und für die standardisierten Daten dargestellt. Die drei wichtigsten Peaks sind jeweils die

gleichen (lediglich die Reihenfolge von R41 und R44 ist vertauscht). Es ist zu sehen, dass die ausgewählten Variablen jeweils ausschließlich auf Gerät A vorkommen und auf diesem die beiden Gruppen ohne/mit Saft unterscheiden. Im Fall ohne Skalierung (obere Zeile) sprechen, betrachtet man nur die Werte von Gerät A, hohe Werte für ein Auftreten von Saft. Da die Werte von Gerät B alle Null sind, werden diese Beobachtungen im Testdatensatz folglich der negativen Klasse zugeordnet. Aus diesem Grund liegt die TPR in den Tabellen 6.10 und 6.11 bei 0, die TNR hingegen bei 1. Wird die Standardisierung angewendet (Abbildung 6.33 untere Zeile), so verschieben sich durch die getrennte Standardisierung der beiden Geräte die Werte der Saft-Messungen von Gerät A näher an die Null-Werte von Gerät B. Wird in den Klassifikationsbäumen ein Split für eine Variable festgelegt, so wird dieser bei einer perfekt möglichen Klassifikation mittig zwischen den positiven und den negativen Beobachtungen platziert. Dieser Wert wird also eher unterhalb der Beobachtungen von Gerät B liegen, sodass die Beobachtungen für Gerät B in die positive Klasse eingeordnet werden. Aus diesem Grund ergeben sich in den zugehörigen Tabellen in diesem Fall die umgekehrten Maßzahlen (TPR = 1, TNR = 0).

Insgesamt verdeutlichen diese Ergebnisse, dass der Geräte-Effekt besonders für die automatische Peakerkennung eine wichtige Rolle spielt und der Klassifikationserfolg davon abhängen kann, dass dieser Effekt in die Analyse einbezogen wird. Dabei ist die Standardisierung nicht ausreichend, da entscheidende Peaks (hier zur Safterkennung) nur auf einem der Geräte gefunden werden beziehungsweise getrennt für beide Geräte. Dies ist in Abbildung 6.34 verdeutlicht. Dargestellt sind links die Positionen der drei Peaks, welche bei Training nur auf Gerät A die höchsten Variablenwichtigkeits-Werte aufweisen. Für jedes der drei Peaks gibt es ein entsprechendes Äquivalent, welches hauptsächlich auf Gerät B gefunden wurde (jeweils der blaue Punkt oberhalb links). Rechts sind zum Vergleich die Peakpositionen nach der Alignierung der Peakpositionen dargestellt sowie die drei Peaks hervorgehoben, welche in Kapitel 6.6.1 kreuzvalidiert die höchsten Variablenwichtigkeiten für die Saft-Klassifikation aufwiesen. Zwei der drei Peaks stimmen in den Positionen überein. Durch den Alignierungs-Schritt kann die Information beider Geräte sinnvoll genutzt werden.

Insgesamt wurde gezeigt, dass der Alignierungs-Schritt unerlässlich ist, um die automatische Peakerkennung geräteübergreifend nutzbar zu machen. Die Problematik der verschiedenen Geräte steht an dieser Stelle auch stellvertretend für andere mögliche Störgrößen, welche unbeachtet oder unbekannt die Ergebnisse verfälschen können.

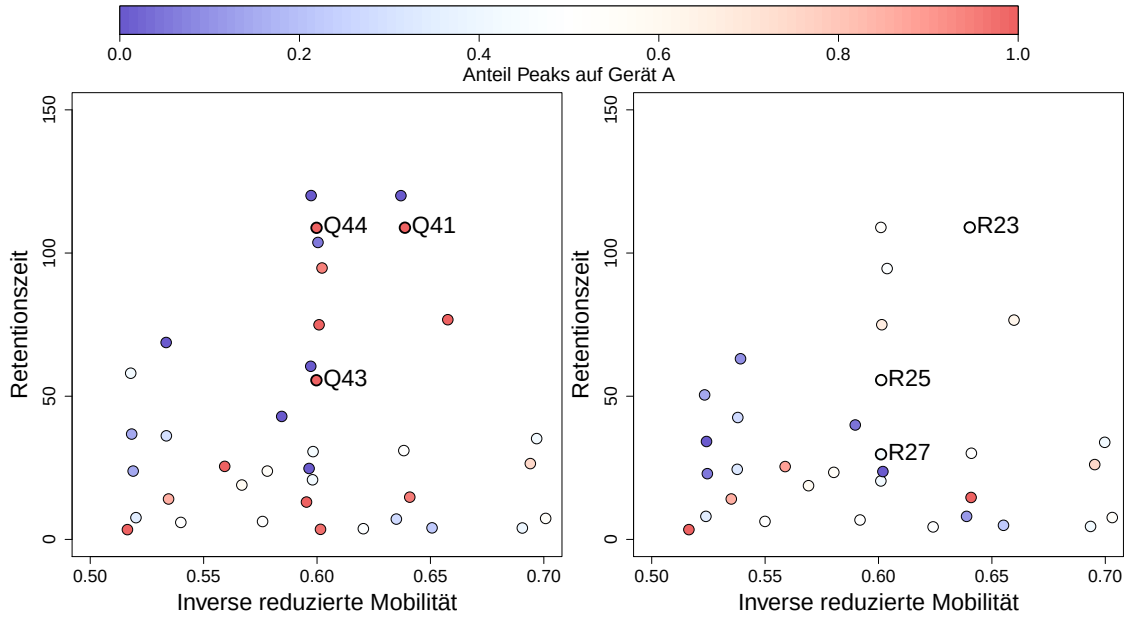


Abbildung 6.34: Peakpositionen der drei Variablen mit den höchsten Variablenwichtigkeiten beim Klassifikationsproblem Saft für die automatische Peakerkennung ohne Alignierung sowie ohne Skalierung, wenn nur auf Gerät A trainiert wurde (links) und mit Alignierung sowie ohne Skalierung, entsprechend der kreuzvalidierten Ergebnisse aus Kapitel 6.6.1 (rechts). Die Farben indizieren die Anteile auf beiden Geräten (rot: Peak wurde hauptsächlich auf Gerät A gefunden, blau: Peak wurde hauptsächlich auf Gerät B gefunden, weiß: ausgeglichene Verteilung).

7 Diskussion und Ausblick

Diese Arbeit befasst sich mit zwei Themenblöcken aus der Auswertung von MCC-IMS-Messungen. Im ersten Block wurde der Gesamtanalyseprozess von den Rohmessungen bis zur Lösung eines Klassifikationsproblems untersucht. Ziel war es, automatische Methoden für die Peakerkennung zu finden, welche in einem nachgeschalteten Klassifikationsproblem mindestens genauso gut abschneiden wie der (semi-)manuelle Goldstandard (Auswertung der Rohmessungen mit der Software VisualNow). Gleichzeitig war die geeignete Wahl eines Klassifikationsalgorithmus notwendig. In Hinblick auf die Klassifikationsgüte wurden automatische Algorithmenkombinationen (beispielsweise SGLTR Peakauswahl kombiniert mit DBSCAN Peakclustern) gefunden, welche diese Anforderungen erfüllen. Dieser Erfolg garantiert jedoch noch keine optimale Peakerkennung. In Kapitel 5.6 wird deutlich, dass auch die empfohlene automatische Peakerkennung nicht alle mit bloßem Auge sichtbaren Peaks in den Rohmessungen detektiert. Dem Anspruch, dass die automatische Peakerkennung jeden sichtbaren Peak detektieren muss, wird die Methode nicht gerecht. Um eine unmittelbare Bewertung der Qualität der Single Peaks vornehmen zu können, müssten alle Rohmessungen aufwändig einzeln untersucht werden, worauf in dieser Arbeit verzichtet wurde. Es wäre jedoch denkbar, für jede Rohmessung manuell festzulegen, wo ein Peak vorliegt und in welcher Umgebung dieser Peak hätte annotiert werden müssen. Dabei müsste dann auch festgelegt werden, wie mit mehreren detektierten Peaks in der Umgebung umzugehen ist. Dann wäre eine Einteilung in korrekt annotierte Peaks, fehlende Peaks und falsch positive Peaks möglich. Der zweite Schritt, das Clustern der Peaks, ist noch schwieriger zu beurteilen, da auch für das bloße Auge nicht eindeutig ersichtlich ist, welche Peaks aus den verschiedenen Messungen tatsächlich durch den gleichen Analyt in der Atemluft erzeugt werden. Dass gleichzeitig auch die manuelle Peakerkennung nicht ohne Fehler ist und beispielsweise nahe beieinander liegende Peaks in ihren Intensitäten manchmal kaum auseinanderzuhalten sind, wurde in Kapitel 6.6.1 festgestellt.

Beim Versuch, den Geräte-Effekt durch eine Skalierung zu reduzieren, waren bei der automatischen Peakerkennung die vielen Werte, die Null sind, problematisch (Kapitel 6.4). Die Verteilungen der Variablen sind damit nicht mehr stetig. Dies könnte auch zum Teil durch

Messungenauigkeiten bedingt sein, also durch nicht messbare Konzentrationen der Stoffe oder durch Fehler bei der Peakerkennung. Bei der (semi-)manuellen Peakerkennung kommen Werte, die exakt Null sind, praktisch nicht vor, da für jeden Peak die Signalintensität an der Stelle ausgewertet wird, an der ein (Consensus) Peak liegt, unabhängig davon, ob in dieser Rohmessung optisch ein Peak sichtbar ist oder nicht. Bei der automatischen Peakerkennung hingegen wird nur dann ein Wert ungleich Null eingetragen, wenn in dieser Messung auch ein Peak an dieser Stelle detektiert wird. Dabei kann es passieren, dass Peaks übersehen werden. Um diese mögliche Fehlerquelle zu reduzieren, könnten alle Rohmessungen nach der Bildung der Consensus Peaks erneut betrachtet werden. An den Stellen, an denen ein Consensus Peak vorliegt, jedoch kein Single Peak annotiert wurde (das heißt, der Wert für den Consensus Peak ist Null), kann die Signalintensität erneut untersucht werden. Ist die Signalintensität an dieser Stelle sehr hoch, könnte nachträglich noch ein Wert für diese Rohmessung eingetragen werden. Um zu vermeiden, dass dabei zu viel Rauschen in den Datensatz aufgenommen wird, könnte ein höherer Schwellenwert gewählt werden als sonst. Dieses Vorgehen ist allerdings nicht möglich, wenn die Peaks „online“ detektiert werden, da dann die Rohmessung nicht gespeichert wird.

Für die Gerätealignierung wurde hier ein Referenzgemisch verwendet. Dieses Vorgehen hat den Vorteil, dass die Transformation leicht durchgeführt werden kann. Das Referenzgemisch sollte idealerweise zeitnah mit den übrigen Messungen durchgeführt werden (hier lagen nur ältere Messungen des Referenzgemisches vor). Der Nachteil des Referenzgemischs ist, dass es nur einen kleinen Teil der möglichen Peakpositionen abdeckt. Denkbar wäre auch, ein Verfahren zu entwickeln, das nur die Rohmessungen selbst verwendet und dabei Unterschiede zwischen den Geräten untersucht werden. Sind alle Rohmessungen randomisiert auf den Geräten gemessen worden (messen beide Geräte im Mittel also das gleiche), könnten beispielsweise die Rohmessungen beider Geräte separat gemittelt und verglichen werden. Es wäre also beispielsweise möglich, die lineare Transformation der Retentions- und IRM-Werte zu finden, welche ein Abstandsmaß zwischen den beiden Messungen minimiert. Dabei muss jedoch beachtet werden, dass die Messungen sehr groß sind und viele Bereiche, an denen nur Rauschen vorliegt, nicht relevant für mögliche Verschiebungen sind. Gleichzeitig dürfen Peaks, die beispielsweise spezifisch für ein Gerät sind und somit keinen Gegenpart auf dem anderen Gerät besitzen, nicht stark ins Gewicht fallen. Zudem müssten Nebenbedingungen aufgestellt werden, welche das Ausmaß der möglichen Verschiebung sinnvoll eingrenzen. Einerseits wäre diese Herangehensweise komplex und nicht mehr so leicht nachvollziehbar, andererseits könnte auf diese Weise gegebenenfalls auf die Messung eines Referenzgemischs verzichtet werden.

Der Random Forest hat sich in dieser Arbeit als geeignetes Klassifikationsverfahren bewährt. In der Zukunft könnte überprüft werden, ob die Klassifikation durch weitere Optimierungen noch verbessert werden kann (andere Implementierungen, anderes Unreinheitsmaß, Optimierung des Parameters, wie viele Variablen in jedem Split zur Verfügung stehen etc.). Darüber hinaus kann die Laufzeit eventuell reduziert werden, wenn die Anzahl der verwendeten Bäume nicht festgelegt wird sondern nur so viele Bäume gebildet werden, bis ein Konvergenzkriterium erreicht ist. Beispielsweise kann nach Hastie u. a. (2009, S. 592f.) das Training beendet werden, wenn sich der Out-Of-Bag-Fehler (dieser Fehler wird berechnet, indem für jede Beobachtung nur die Bäume gemittelt werden, für die diese Beobachtung nicht verwendet wurde, weil sie nicht in der Bootstrap-Stichprobe enthalten war) stabilisiert. Durch eine vorgeschaltete Variablenselektion könnte sich das Ergebnis ebenfalls verbessern. Besonders bei der automatischen Peakerkennung gab es viele Variablen, die nur für sehr wenige Beobachtungen Werte ungleich Null aufwiesen (siehe Abbildung 6.11). Diese könnten vorab entfernt werden.

In Bezug auf die Untersuchung von Einflussfaktoren auf die Atemluft konnten für die Faktoren Geschlecht und Rauchen keine Unterschiede zwischen den teilnehmenden Personen festgestellt werden. Da nur sehr wenige Personen rauchten und diese daher mit Ex-Rauchenden zusammengefasst wurden, ist das Ergebnis zum Rauchen jedoch nicht sehr aussagekräftig. Um diesen Faktor angemessen bewerten zu können, hätten mehr aktiv Rauchende untersucht werden müssen. Für den Faktor Geschlecht überrascht das Ergebnis jedoch, da für andere Technologien der Atemluftforschung in der Literatur bereits häufig Unterschiede festgestellt werden konnten (vergleiche Kapitel 2.3). Möglicherweise wurden auch hier nicht genug Personen in die Studie einbezogen oder die Unterschiede sind mit der MCC-IMS aktuell nicht messbar. Vor der Adjustierung der p-Werte gab es vereinzelt signifikante Metaboliten bezüglich eines Lageunterschieds zwischen den Geschlechtern, welche noch explorativ betrachtet werden könnten. In der Literatur als relevant identifizierte Stoffe könnten gezielt in den MCC-IMS-Messungen gesucht werden, wenn ihre Peakpositionen bekannt sind. Auf diese Weise könnte überprüft werden, ob diese Stoffe möglicherweise durch keine der Peakerkennungen detektiert wurden oder ob diese Stoffe hier keine Auffälligkeiten bezüglich der Geschlechter aufweisen.

Ein weiterer in der Literatur häufig aufgeführter Einflussfaktor wurde in dieser Arbeit nicht berücksichtigt. Die Raumluft wurde zwar bei jeder Atemluftmessung ebenfalls analysiert (nur bei den Daten zur Analyse der Einflussfaktoren), jedoch wurde die Atemluft nicht bezüglich der Raumluft adjustiert. Für die Korrektur der Atemluft hinsichtlich der Raumluft gibt es bisher kein Standardverfahren (Risby, 2008). In einigen Studien wird die Raumluft von der Atemluft abgezogen (beispielsweise in Kischkel u. a. (2010)). Diese Methode wird jedoch häufig kritisiert,

da der Gasaustausch in der Lunge deutlich komplexer abläuft (Miekisch u. a., 2004; Kushch u. a., 2008). Als Beispiel führen Kushch u. a. (2008) an, dass der Kohlenstoffdioxidgehalt in der ausgeatmeten Luft stets 4% beträgt, unabhängig davon, wie hoch die Konzentration in der eingeatmeten Luft ist (0%, 1% oder 2%). Eine andere Möglichkeit, Fehlschlüsse aus der Atemluft aufgrund einer hohen Konzentration in der Raumluft zu vermeiden, ist, Stoffe aus dem Datensatz zu entfernen, bei denen die Raumluftkonzentration einen bestimmten Prozentsatz (5-50%) der Konzentration der Atemluft übersteigt (Risby, 2008; Kushch u. a., 2008; Beauchamp, 2011). Pleil u. a. (2013) verdeutlichen die Komplexität der Problematik, indem sie beispielhaft für einen Stoff ein Modell aufstellen, welches die Aufnahme, Verteilung, Verstoffwechselung und Ausscheidung eines Stoffes im Körper beschreibt. Sie schließen, dass die Kinetik der Stoffe die Interpretation der gemessenen Daten erschwert, da sie für jeden Stoff verschieden ist und auch nur dann konstant ist, wenn sich zwischen der Atemluft der untersuchten Person und der Raumluft ein Gleichgewicht eingestellt hat. Wie lange es dauert, bis ein solches Gleichgewicht erreicht ist, ist jedoch unklar (Risby, 2008). Da bei den hier vorliegenden Daten (zur Analyse der Einflussfaktoren) die Geräte stets im gleichen Raum standen und die Messungen nur in einem Zeitraum von vier Tagen stattfanden, ist davon auszugehen, dass sich die Raumluft innerhalb der Messungen nicht sehr stark unterscheidet. Dennoch sollten zukünftig Maßnahmen geprüft werden, um die Atemluftmessungen um die Raumluftbestandteile zu bereinigen. Um den Effekt der Raumluft zu untersuchen, könnten zusätzlich Studien mit Messungen der gleichen Personen an verschiedenen Orten durchgeführt werden.

8 Zusammenfassung

Die Peakerkennung bei der Auswertung von MCC-IMS-Rohmessungen erfordert aktuell eine manuelle Begutachtung der Messungen. Um diesen Goldstandard durch automatische Verfahren ersetzen zu können, wurden in dieser Arbeit mehrere Algorithmen getestet. Die automatischen Verfahren setzen sich aus zwei Schritten, der „Peakauswahl“ und dem „Peakclustern“ zusammen. Während die Peakauswahl für jede Rohmessung einzeln untersucht, an welchen Positionen Peaks vorliegen, werden diese Peaks aus mehreren Rohmessungen beim Peakclustern einander zugeordnet. Dann liegt eine Liste von Peaks vor, die den gesamten Datensatz repräsentiert. Insgesamt werden in dieser Arbeit 25 Kombinationen aus Peakauswahl- und Peakcluster-Verfahren gebildet und mit der manuellen Peakerkennung verglichen. Da es in der Atemluftanalyse häufig das Ziel ist, kranke und gesunde Personen voneinander zu unterscheiden, wurden die Methoden auf drei verschiedene Datensätze angewendet, welche alle ein derartiges Zwei-Klassen-Problem beinhalten. Da die Zusammensetzung der Atemluft im Voraus nicht bekannt ist, ist die Güte der Peakerkennung nicht einfach zu beurteilen. In dieser Arbeit galt ein automatisches Verfahren zur Peakerkennung als erfolgreich, wenn die Klassifikation der Datensätze mindestens ebenso gut gelang wie bei der (semi-)manuellen Peakerkennung. Damit dieses Ergebnis nicht vom Klassifikationsverfahren abhängig ist, und um gleichzeitig einen geeigneten Klassifikationsalgorithmus zu finden, wurden verschiedene Klassifikationsalgorithmen getestet. Dabei schnitt der Random Forest bezüglich des AUC-Wertes deutlich besser ab als die übrigen Algorithmen in der Auswahl. Für den Schritt der Peakerkennung erzielten SGLTR (Savitzky-Golay Laplace-operator filter thresholding regions) und LM (Local Maxima) die besten Ergebnisse in Kombination mit DBSCAN (Density-Based Spatial Clustering of Applications with Noise) oder EM (Expectation Maximization Clustering) Peakclustern. Die Klassifikationsgüte ist für die derart automatisch ausgewerteten Daten nicht schlechter als die (semi-)manuelle Peakerkennung. Für die weitere Anwendung in dieser Arbeit wurde die Kombination aus SGLTR Peakauswahl, DBSCAN Peakclustern und Random Forest Klassifikation verwendet.

Der zweite Abschnitt dieser Arbeit beschäftigte sich mit Einflussfaktoren auf die Atemluft bei MCC-IMS-Messungen. Während für andere Technologien bereits Einflussfaktoren identifiziert wurden, fehlen derartige Studien für MCC-IMS-Messungen bisher. Variablen wie das Geschlecht oder der Raucherstatus von Personen werden in Untersuchungen an erkrankten Menschen bisher kaum einbezogen. Es wurden Messungen an 49 Personen (ohne spezielle Erkrankungen) durchgeführt, um die Effekte des Geschlechts, des Raucherstatus und des verwendeten Gerätes zu untersuchen. Außerdem wurde die Atemluft jeder Person zweimal gemessen, einmal vor und einmal nach dem Konsum eines Glases Orangensaft. Diese Beeinflussung der Atemluft kann ebenfalls als Störgröße interpretiert werden, wurde hier jedoch hauptsächlich als Surrogat eines Krankheitsstatus verwendet. Auf diese Weise konnte an gesunden Personen ein Klassifikationsproblem erzeugt werden, welches normalerweise in der Unterscheidung von gesunden und kranken Personen besteht. Der Effekt des Geräts wurde untersucht, indem die beiden Messungen einer Person auf zwei unterschiedlichen Geräten durchgeführt wurden. Dabei wurde darauf geachtet, dass der Saft-Status und das Gerät zufällig kombiniert werden und nicht das gleiche Gerät alle Messungen mit Saft maß.

Zwischen den Geschlechtern zeigten sich weder in univariaten Tests noch in der Klassifikation Unterschiede. Dieses Ergebnis steht nicht im Einklang mit Ergebnissen für andere Technologien der Atemluftauswertung. In zukünftigen Untersuchungen könnte versucht werden, die Peakpositionen der dort nachgewiesenen VOCs (flüchtige organische Verbindungen) für die MCC-IMS zu bestimmen und dort gezielt nach den entsprechenden Peaks zu suchen. Auf diese Weise könnte herausgefunden werden, ob diese Stoffe in den MCC-IMS-Messungen nicht detektiert werden oder ob die Unterschiede für diese Variablen in den vorliegenden Daten nicht (oder weniger stark ausgeprägt) vorliegen.

Da die untersuchten Personen nicht gezielt nach ihrem Raucherstatus ausgewählt wurden, lagen nur Daten sehr weniger aktiv rauchender Personen vor. Diese wurden daraufhin mit den Daten der Ex-Rauchenden zusammengefasst. Es ergaben sich dann keine Unterschiede zwischen den (Ex-)Rauchenden und den Nichtrauchenden. Dies ist jedoch aufgrund des sehr kleinen Stichprobenumfangs und der wahrscheinlichen Verringerung des Effekts des Rauchens durch Personen, die schon lange nicht mehr rauchen, nicht sehr aussagekräftig. Eine gezielte Untersuchung des Effekts für MCC-IMS-Messungen in der Zukunft ist noch notwendig.

Zwischen den beiden Geräten zeigten sich große Unterschiede. Diese lassen sich zum Teil darauf zurückführen, dass die Peakpositionen der Messungen für beide Geräte leicht verschieden sind. Insbesondere bei der automatischen Peakerkennung führt dies dazu, dass beim Schritt des Peakclusterns zwei Peaks gebildet werden (je eines pro Gerät), obwohl in Wahrheit

nur eine Komponente in der Atemluft vorliegt. Mit Hilfe eines Referenzgemischs, dessen Zusammensetzung bekannt ist, kann dieses Phänomen reduziert werden. Wird das Gemisch auf beiden Geräten gemessen, so können die unterschiedlichen Peakpositionen direkt verglichen werden. Durch lineare Regressionen in die beiden Dimensionen der Rohmessungen können die Peakpositionen, die nach der Peakauswahl vorliegen, für ein Gerät so verschoben werden, dass sie denen des anderen Geräts entsprechen. Wird anschließend das Peakclustern durchgeführt, so ergeben sich deutlich plausiblere Ergebnisse. Die Messwerte separat für die beiden Geräte zu skalieren, führte zwar teilweise zu plausibleren Werten, jedoch waren die Geräte in der Klassifikation weiterhin deutlich voneinander unterscheidbar, sodass das Gerät als Störfaktor nicht ausgeschlossen werden konnte. Für die automatische Peakerkennung wurde demonstriert, dass die Klassifikation scheitern kann, wenn ein auf nur einem Gerät trainiertes Modell ohne Berücksichtigung des Geräte-Effekts auf Daten eines anderen Gerätes angewendet wird. Aus diesem Grund müssen Daten von verschiedenen Geräten sehr sorgfältig ausgewertet werden. Auch in der Planung von Experimenten sollte darauf geachtet werden, dass die Zuordnung der Untersuchungsobjekte randomisiert auf die Geräte erfolgt.

Den Wahrscheinlichkeitsschwellenwert in der Klassifikation durch die Prävalenzmethode zu wählen, anstatt ihn auf den fixen Wert 0.5 zu setzen, brachte keine Vorteile bezüglich der AUC-Werte.

Anhang

A Algorithmen im Gesamtanalyseprozess

Optimierte Parameter in der Klassifikation

Die Klassifikation in Kapitel 5 wurde in R (R Core Team, 2016) mit Version 3.2.2 durchgeführt. Die Parameter-Optimierung wurde innerhalb einer geschachtelten Kreuzvalidierung mit einer einfachen Gittersuche durchgeführt. Die dabei verwendeten Parameter (sowie die sich von den Standard-Parametereinstellungen unterscheidenden), sind im Folgenden aufgelistet. Für die Gittersuche wurde häufig eine große Bandbreite des Parameterraums untersucht, da keine Vorinformationen über diese Parameter vorliegen. Dafür wurden die Werte auf einem exponentiellen Gitter 2^x verwendet, wobei x die 10 äquidistanten Werte zwischen -15 und 15 annimmt. Diese Parameteroptionen werden im Folgenden als „exponentielles Gitter“ bezeichnet.

Support Vector Machine (SVM)

- **R-Paket:** e1071 (Meyer u. a., 2014)
- **Parameter:**
  - `cost`: Optimierung auf dem exponentiellen Gitter
  - `kernel`: „linear“ (für die lineare SVM) und „radial“ für die RBF-SVM
  - `gamma`: Optimierung auf dem exponentiellen Gitter (nur für die RBF-SVM)

k-Nächste-Nachbarn

- **R-Paket:** kknn (Schliep und Hechenbichler, 2014)
- **Parameter:**

– `k`: Optimierung auf den Werten $\{1, 2, ..., 10\}$

Klassifikationsbaum

- **R-Paket:** rpart (Therneau u. a., 2015)
- **Parameter:**
  - `minbucket`: Optimierung auf den Werten $\{1, 2, ..., 5\}$

Gradienten Basiertes Boosting

- **R-Paket:** gbm (Ridgeway, 2013)
- **Parameter:**
  - `shrinkage`: Optimierung auf dem exponentiellen Gitter
  - `n.minobsinnode`: Optimierung auf den Werten $\{1, 2, ...5\}$
  - `interaction.depth`: Optimierung auf den Werten $\{1, 2, 3\}$

Random Forest

- **R-Paket:** randomForest (Liaw und Wiener, 2002)
- keine Parameteroptimierung

B Analyse von Störgrößen

Tabellen und Abbildungen

Tabelle B.1: Zeitlicher Ablauf der Pilotstudie.

Uhrzeit	Testperson	Beeinflussung
08:15	Person 1	keine
08:55	Person 1	Orangensaft
09:35	Person 1	Schokolade
10:15	Person 2	keine
10:55	Person 2	Orangensaft
11:34	Person 2	Schokolade
12:16	Person 3	keine
12:55	Person 3	Orangensaft
13:35	Person 3	Schokolade
14:34	Person 1	keine
15:19	Person 2	keine
16:00	Person 3	keine

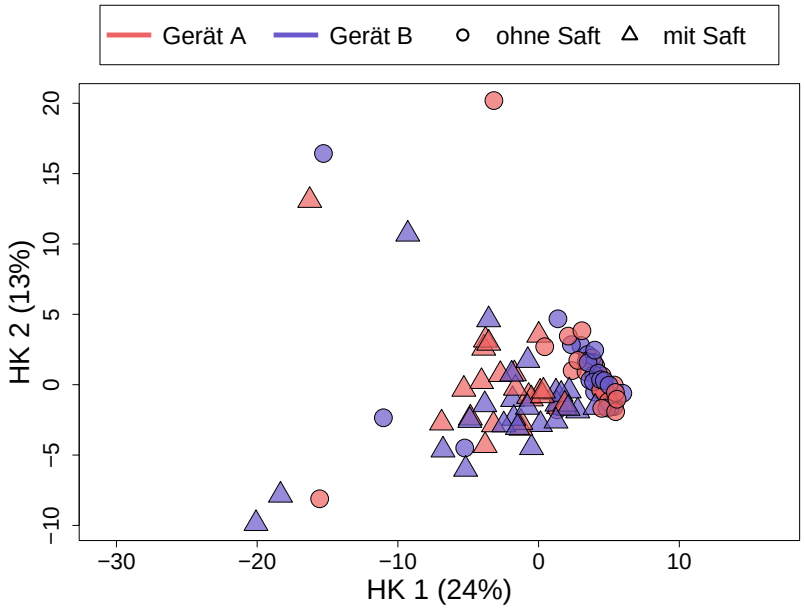


Abbildung B.1: Die ersten beiden Hauptkomponenten der metabolomischen Atemluftdaten der **manuellen** Peakerkennung mit Skalierung auf *Median 0 und MAD 1* (Nullen werden theoretisch separat verschoben, was auf die manuellen Daten allerdings keinen Einfluss hat, da keine Nullen vorkommen). Die Farben markieren, auf welchem Gerät die jeweilige Messung durchgeführt wurde (rot: Gerät A, blau: Gerät B), die Symbole, ob es sich um eine Messung mit oder ohne zuvor konsumierten Orangensaft handelt (Kreis: ohne Orangensaft, Dreieck: mit Orangensaft).

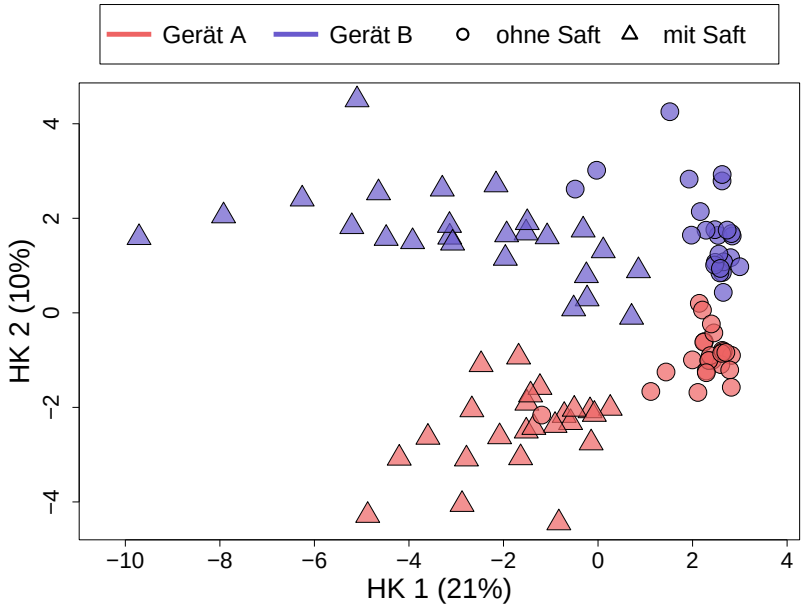


Abbildung B.2: Die ersten beiden Hauptkomponenten der metabolomischen Atemluftdaten der automatischen Peakerkennung **mit Alignierung** und Skalierung auf *Median 0 und MAD 1, wobei die Nullen separat verschoben werden.* Die Farben markieren, auf welchem Gerät die jeweilige Messung durchgeführt wurde (rot: Gerät A, blau: Gerät B), die Symbole, ob es sich um eine Messung mit oder ohne zuvor konsumierten Orangensaft handelt (Kreis: ohne Orangensaft, Dreieck: mit Orangensaft).

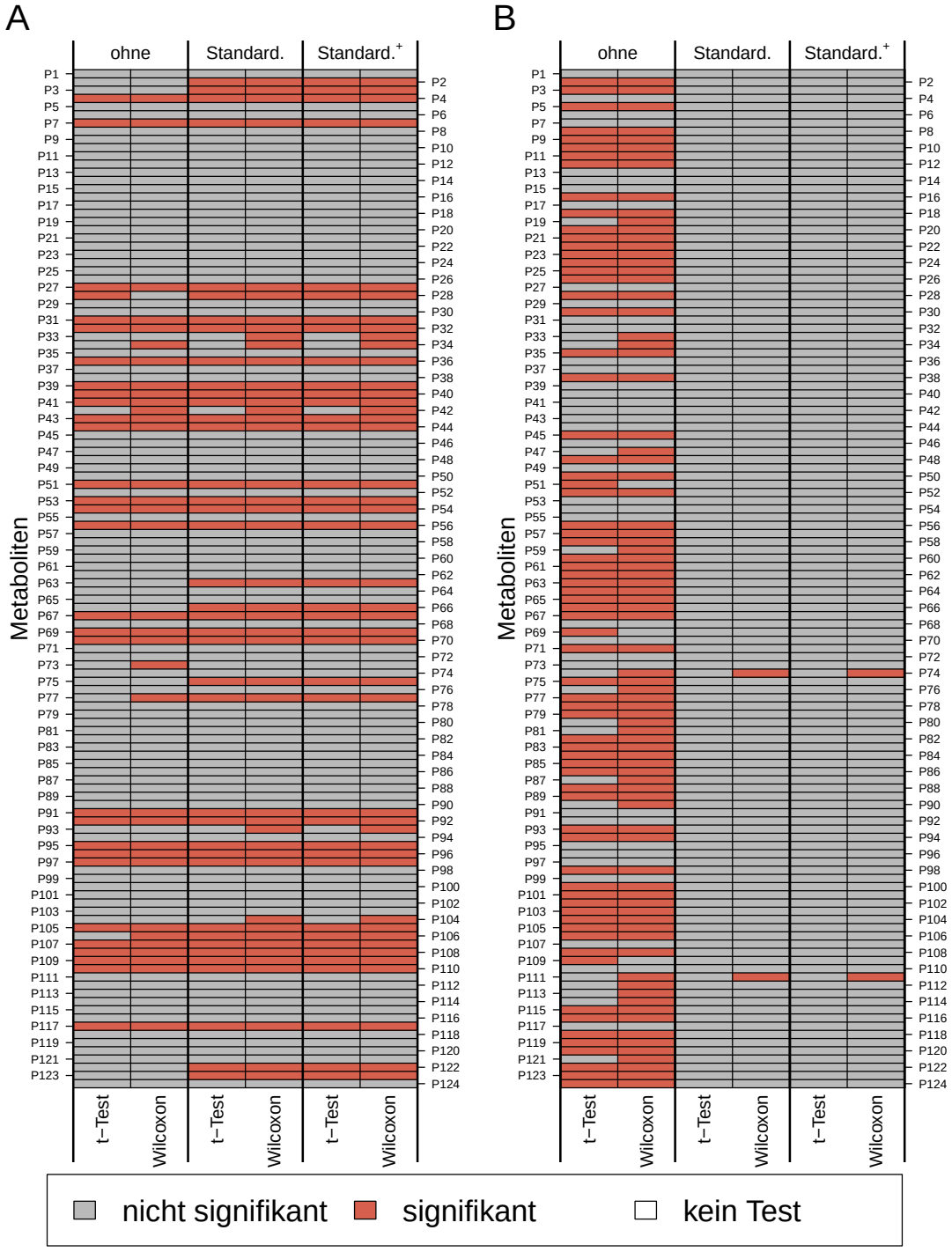


Abbildung B.3: Ergebnisse der univariaten gepaarten Signifikanztests auf Lageunterschiede (adjustiert) der **manuellen** Daten, jeweils für die Daten ohne Skalierung, mit Standardisierung und mit Standardisierung$^+$ sowie für den t-Test und den Wilcoxon Test. A: Unterschiede zwischen ohne/mit Saft. B: Unterschiede zwischen Gerät A/B.

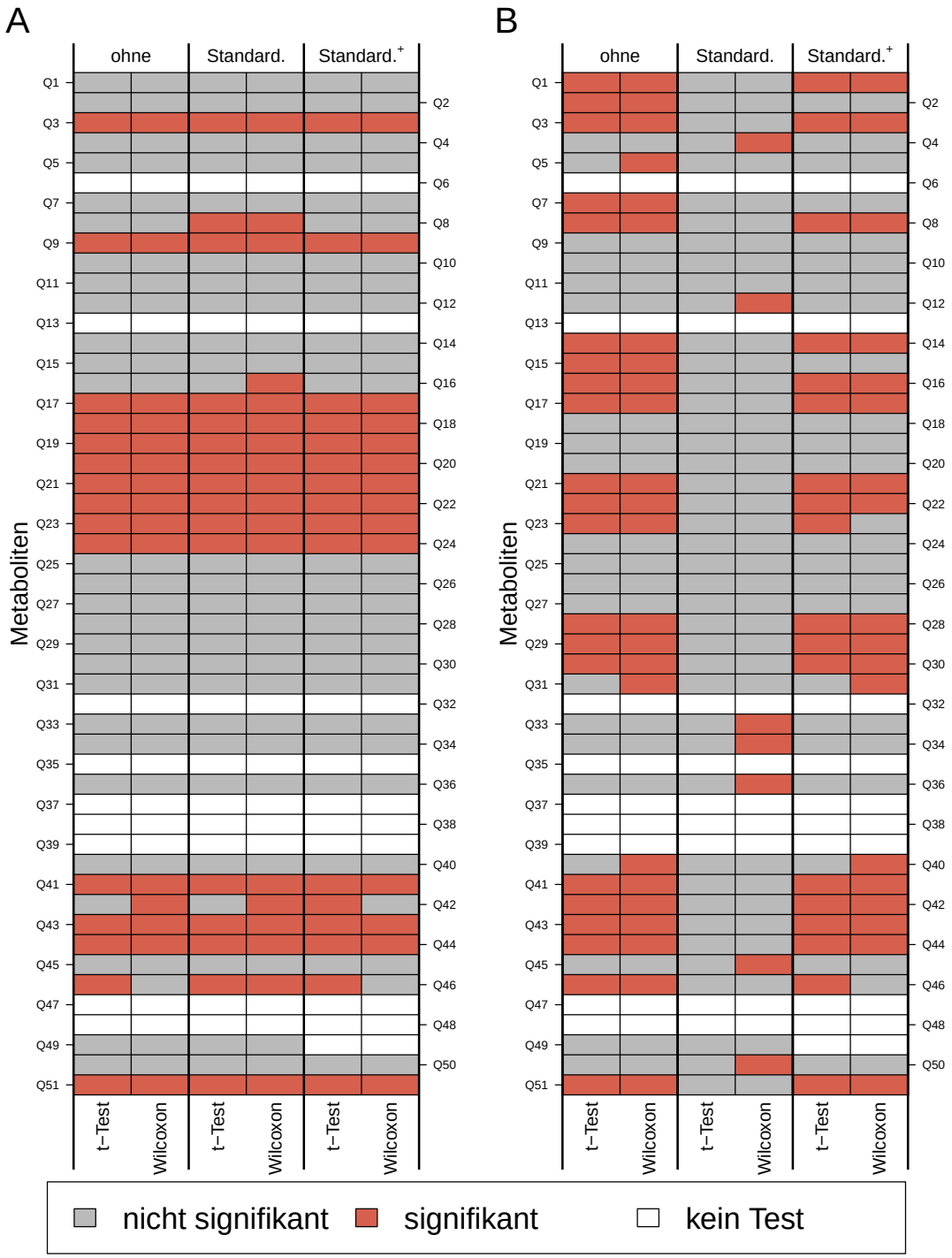


Abbildung B.4: Ergebnisse der univariaten gepaarten Signifikanztests auf Lageunterschiede (adjustiert) der Daten **ohne Alignierung**, jeweils für die Daten ohne Skalierung, mit Standardisierung und mit Standardisierung$^+$ sowie für den *t*-Test und den Wilcoxon Test. A: Unterschiede zwischen ohne/mit Saft. B: Unterschiede zwischen Gerät A/B.

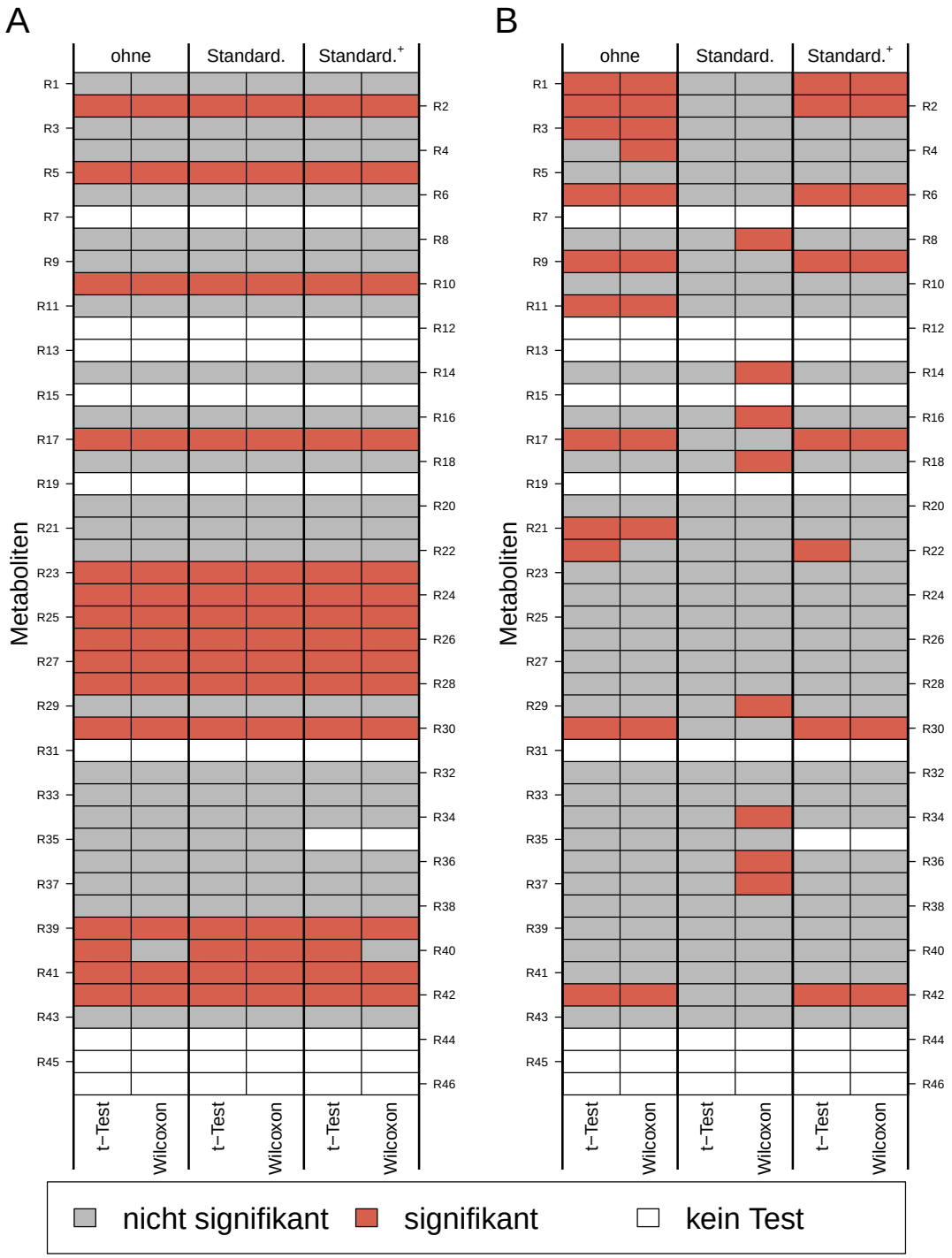


Abbildung B.5: Ergebnisse der univariaten gepaarten Signifikanztests auf Lageunterschiede (adjustiert) der Daten **mit Alignierung**, jeweils für die Daten ohne Skalierung, mit Standardisierung und mit Standardisierung$^+$ sowie für den *t*-Test und den Wilcoxon Test. A: Unterschiede zwischen ohne/mit Saft. B: Unterschiede zwischen Gerät A/B.

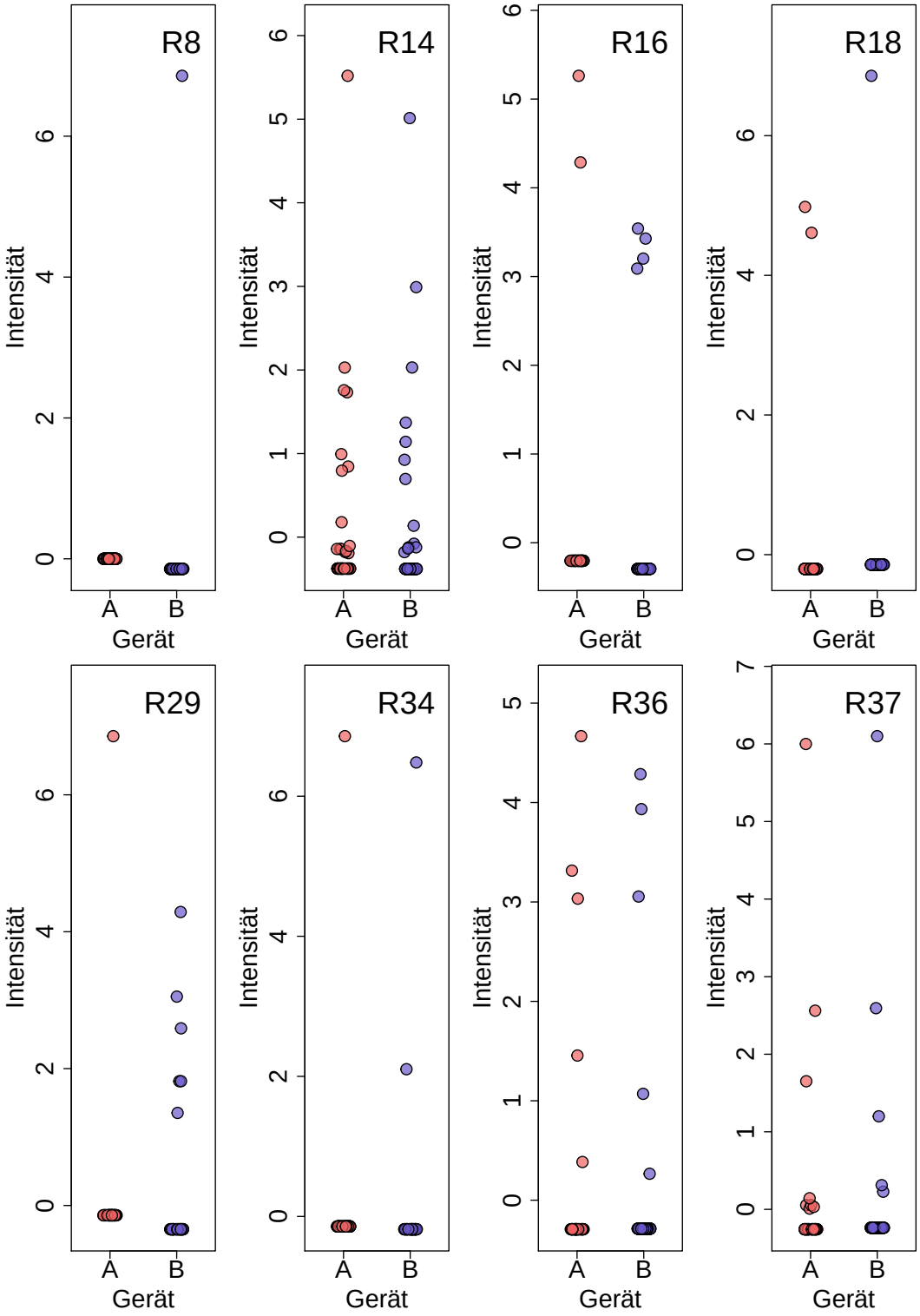


Abbildung B.6: Verteilung der Metaboliten des **alignierten**, *standardisierten* Datensatzes, die für Geräteunterschiede unter Anwendung des Wilcoxon-Tests univariat (adjustiert) signifikant sind.

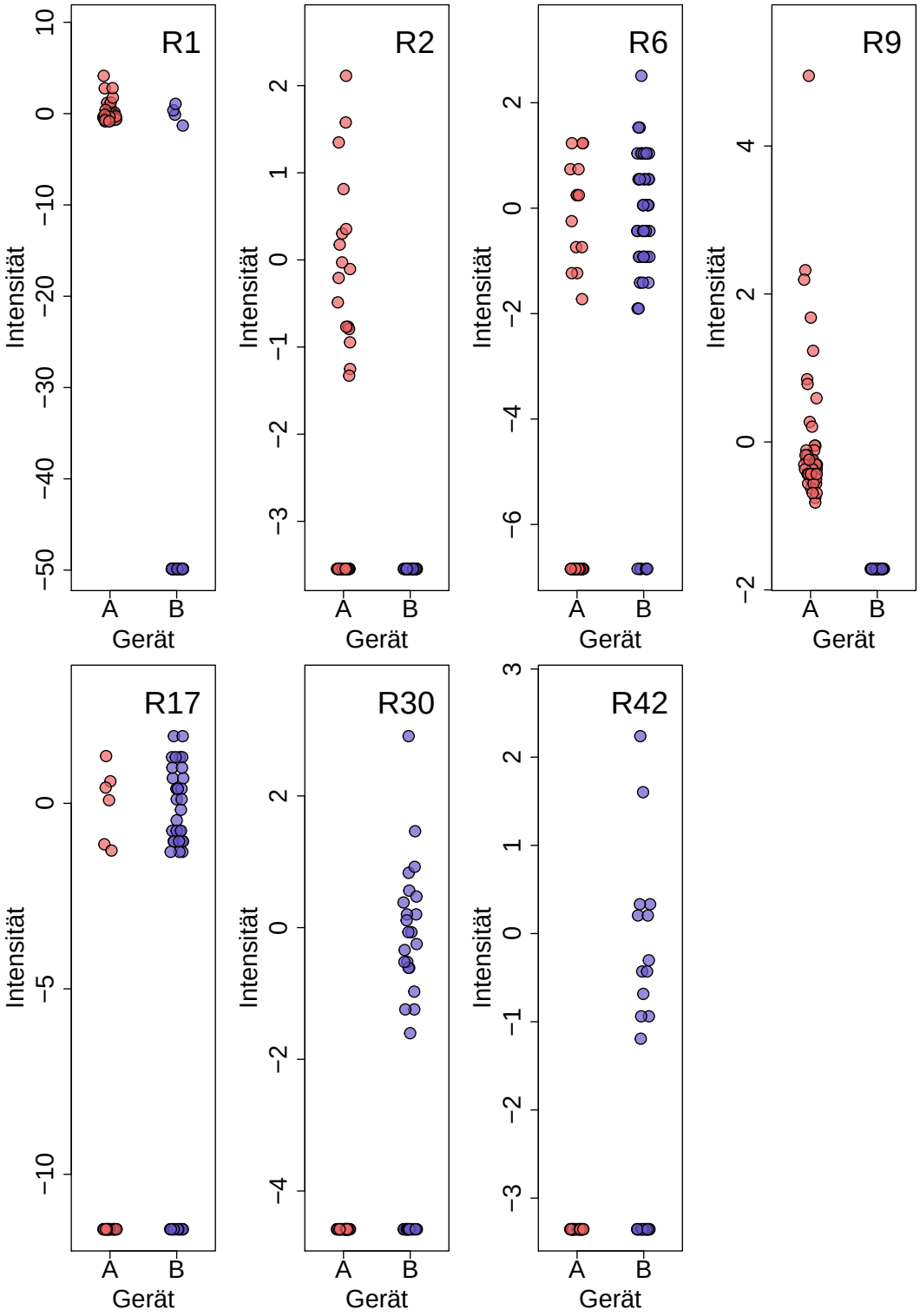


Abbildung B.7: Verteilung der Metaboliten des **alignierten** Datensatzes mit der *erweiterten Standardisierung*, die für Geräteunterschiede unter Anwendung des Wilcoxon-Tests univariat (adjustiert) signifikant sind.

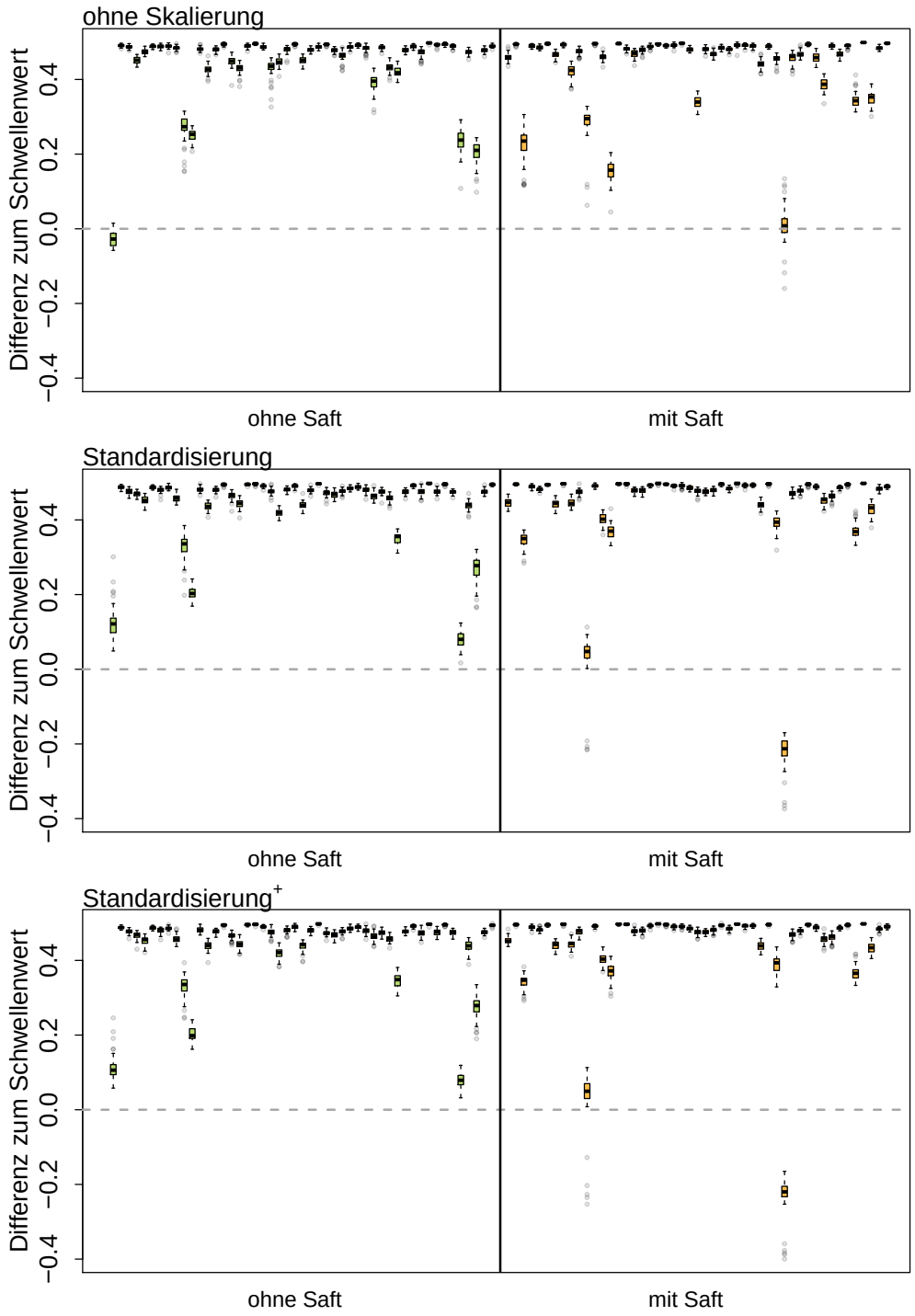


Abbildung B.8: Differenz der Wahrscheinlichkeit für die wahre Klasse zum **Schwellenwert** 0.5 für alle Beobachtungen der Kreuzvalidierungswiederholungen. Betrachtet werden die Daten der **manuellen** Peakerkennung, jeweils ohne Skalierung, mit Standardisierung und mit Standardisierung$^+$ im **Saft**-Klassifikationsproblem.

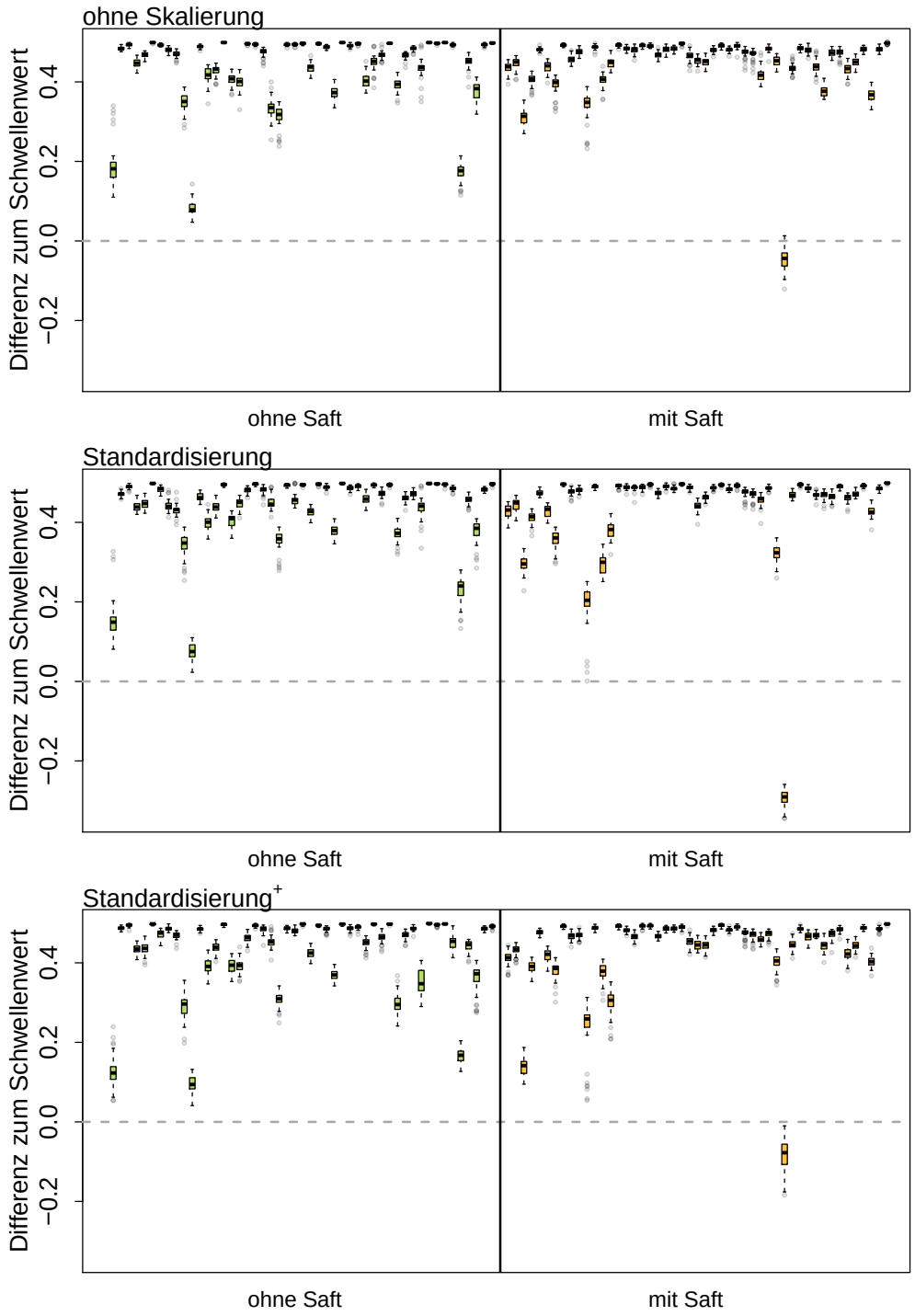


Abbildung B.9: Differenz der Wahrscheinlichkeit für die wahre Klasse zum **Schwellenwert** 0.5 für alle Beobachtungen der Kreuzvalidierungswiederholungen. Betrachtet werden die Daten der automatischen Peakerkennung **ohne Alignierung**, jeweils ohne Skalierung, mit Standardisierung und mit Standardisierung$^+$ im **Saft**-Klassifikationsproblem.

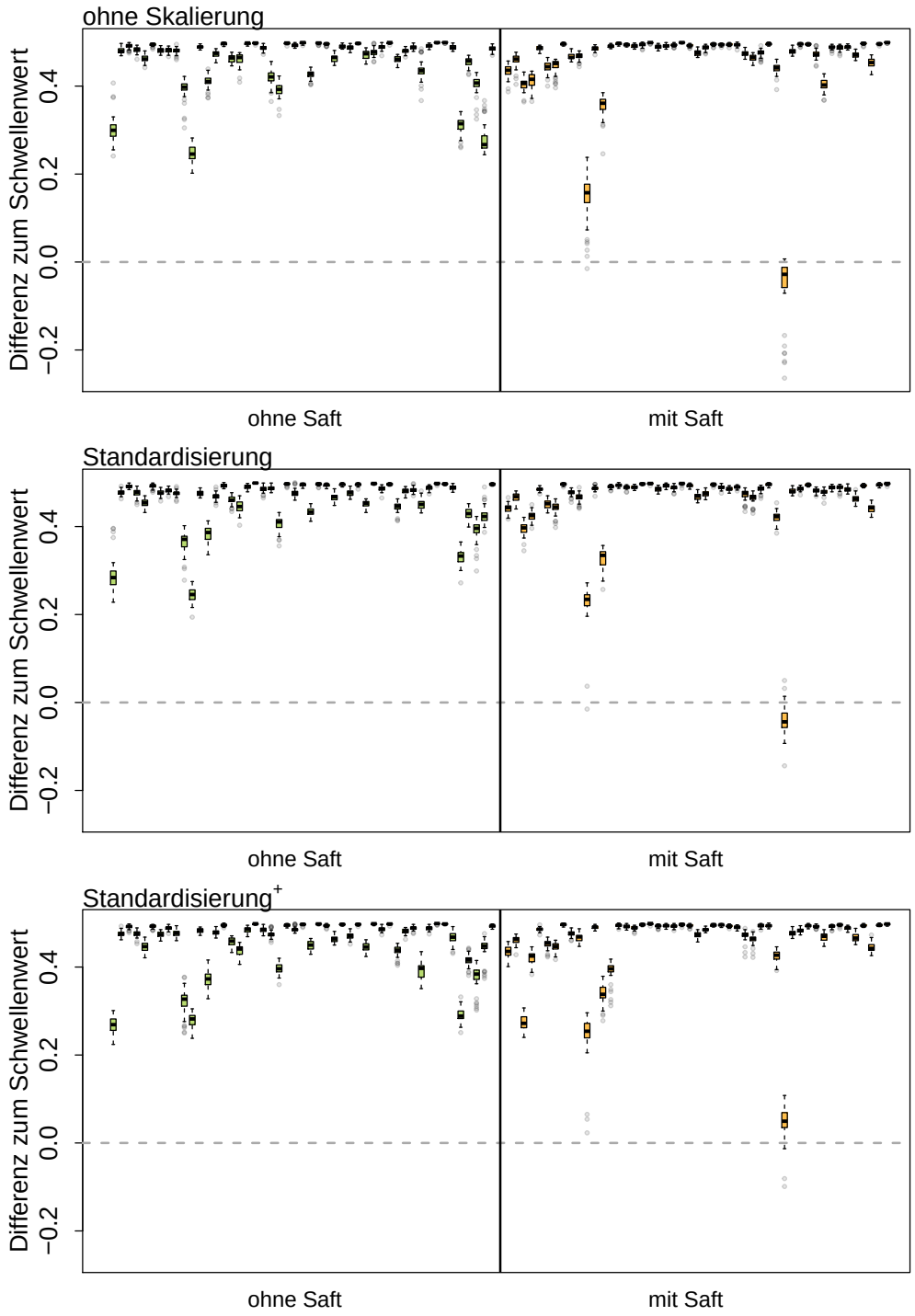


Abbildung B.10: Differenz der Wahrscheinlichkeit für die wahre Klasse zum **Schwellenwert** 0.5 für alle Beobachtungen der Kreuzvalidierungswiederholungen. Betrachtet werden die Daten der automatischen Peakerkennung **mit Alignierung**, jeweils ohne Skalierung, mit Standardisierung und mit Standardisierung$^+$ im **Saft**-Klassifikationsproblem.

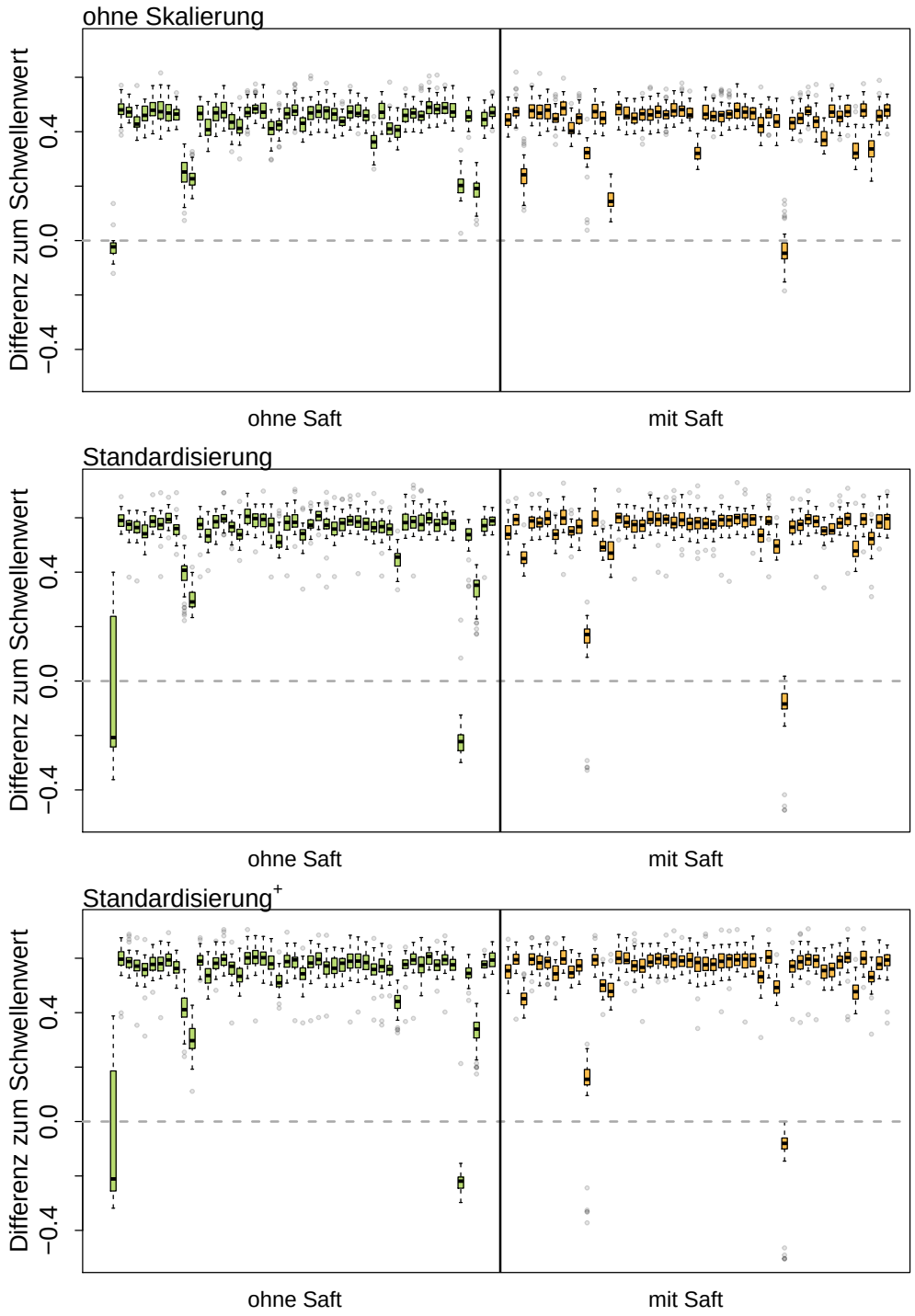


Abbildung B.11: Differenz der Wahrscheinlichkeit für die wahre Klasse zum Schwellenwert nach der **Prävalenzmethode** für alle Beobachtungen der Kreuzvalidierungswiederholungen. Betrachtet werden die Daten der **manuellen** Peakerkennung, jeweils ohne Skalierung, mit Standardisierung und mit Standardisierung$^+$ im **Saft**-Klassifikationsproblem.

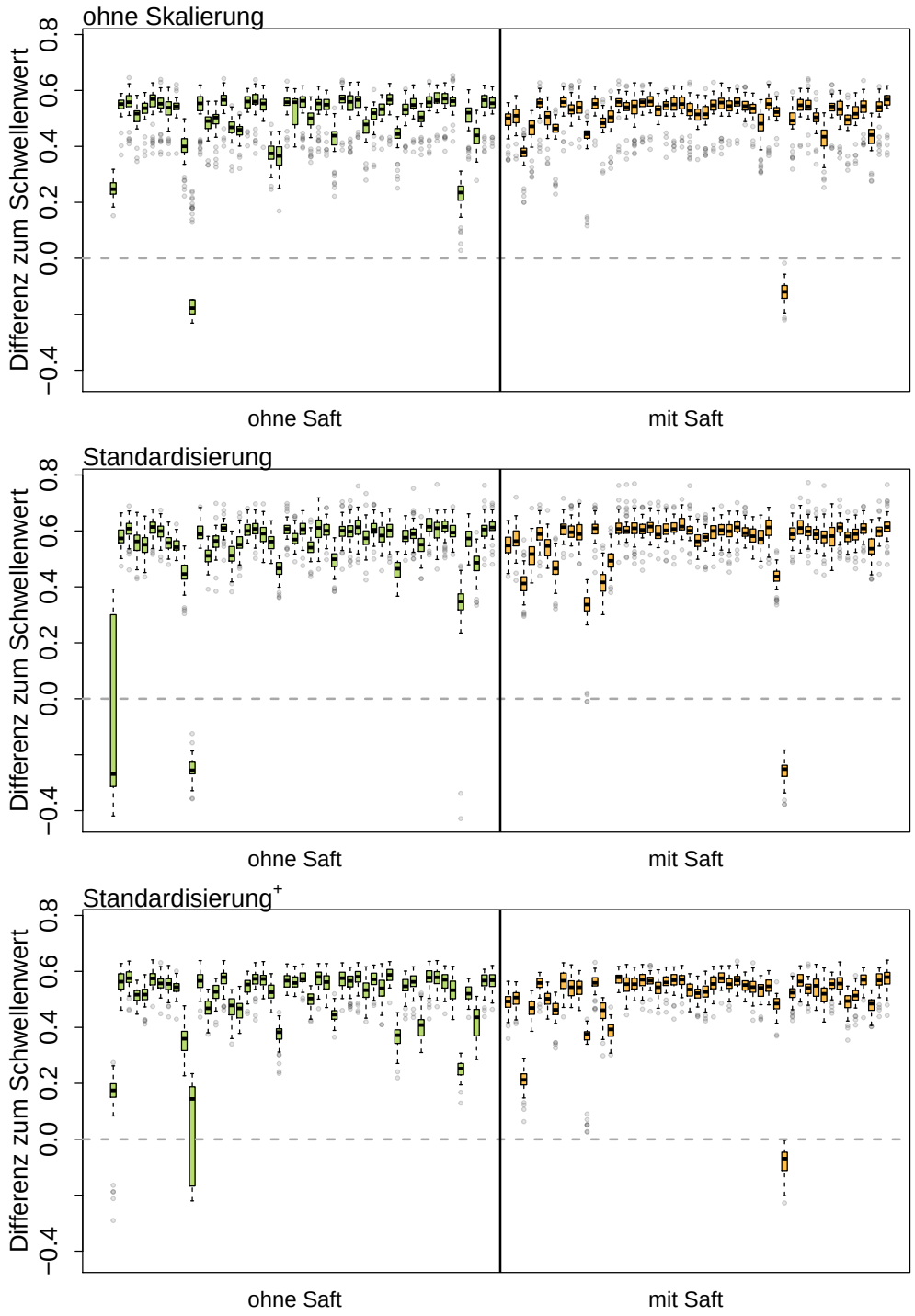


Abbildung B.12: Differenz der Wahrscheinlichkeit für die wahre Klasse zum Schwellenwert nach der **Prävalenzmethode** für alle Beobachtungen der Kreuzvalidierungswiederholungen. Betrachtet werden die Daten der automatischen Peakerkennung **ohne Alignierung**, jeweils ohne Skalierung, mit Standardisierung und mit Standardisierung$^+$ im **Saft**-Klassifikationsproblem.

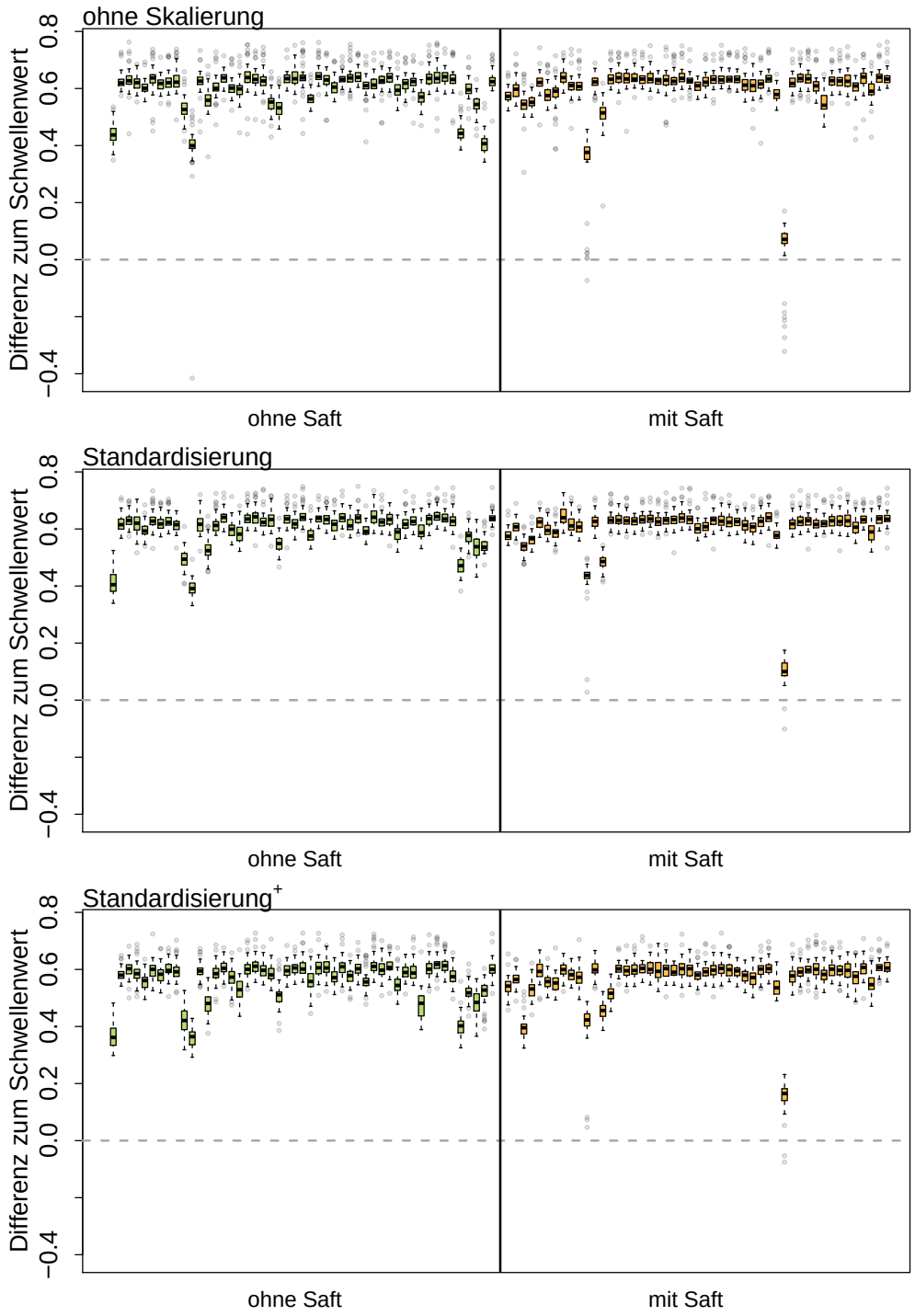


Abbildung B.13: Differenz der Wahrscheinlichkeit für die wahre Klasse zum Schwellenwert nach der **Prävalenzmethode** für alle Beobachtungen der Kreuzvalidierungswiederholungen. Betrachtet werden die Daten der automatischen Peakerkennung **mit Alignierung**, jeweils ohne Skalierung, mit Standardisierung und mit Standardisierung$^+$ im **Saft**-Klassifikationsproblem.

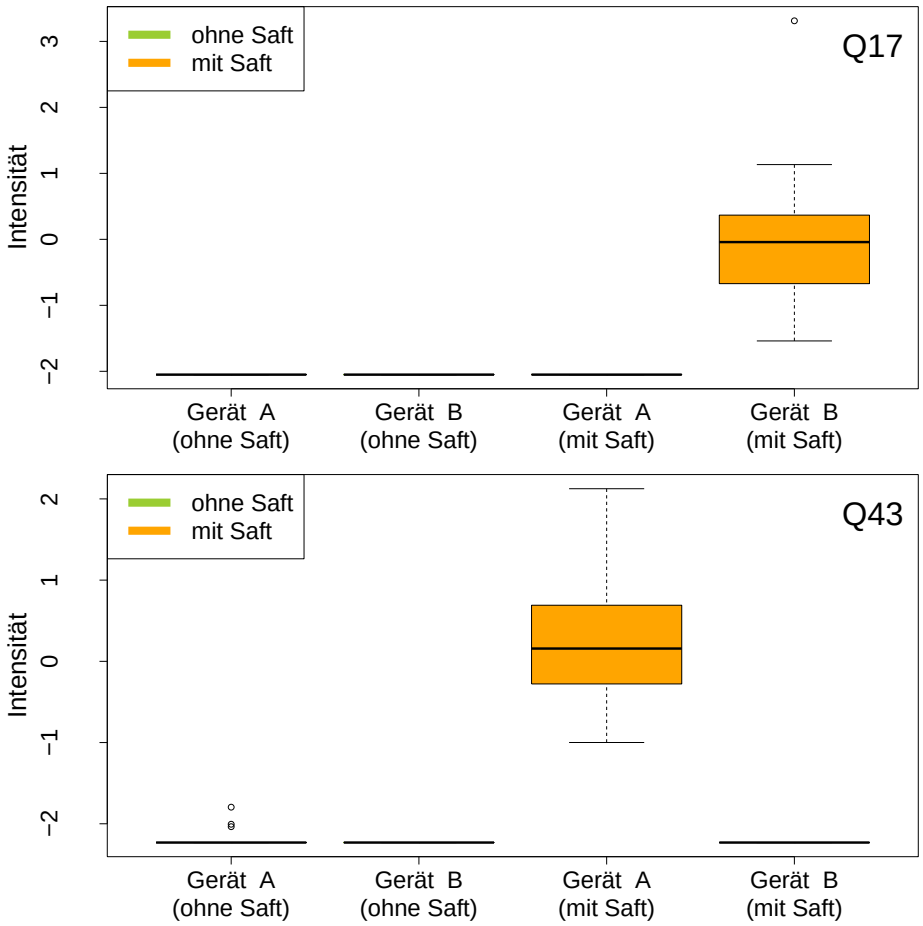


Abbildung B.14: Verteilung der Peaks P17 und P43 der automatischen Peakerkennung ohne Alignierung mit Standardisierung^{+} getrennt nach Gerät und Saftkonsum.

Tabelle B.2: Gemittelte Gütemaßzahlen der kreuzvalidierten Klassifikationsergebnisse für die Zielvariable **Gerät**, sowie das arithmetische Mittel der verwendeten Schwellenwerte s (entsprechend der **Prävalenz-Methode**). Die Klassifikation wurde für die unskalierten Daten und für zwei Skalierungen durchgeführt, Standardisierung und Standardisierung mit separater Behandlung der Nullen (= Standardisierung$^+$).

Skalierung	AUC	ACC	TPR	TNR	PPV	NPV	s
Manuelle Peakerkennung							
ohne	1	1.000	1.000	1.000	1.000	1.000	0.544
Standard.	1	0.989	0.991	0.986	0.986	0.992	0.475
Standard.$^+$	1	0.989	0.989	0.989	0.990	0.989	0.476
Automatische Peakerkennung							
ohne	1	1.000	1.000	1.000	1.000	1.000	0.444
Standard.	1	1.000	1.000	1.000	1.000	1.000	0.389
Standard.$^+$	1	1.000	1.000	1.000	1.000	1.000	0.480
Automatische Peakerkennung mit Alignierung							
ohne	1	1.000	1.000	1.000	1.000	1.000	0.466
Standard.	1	1.000	1.000	1.000	1.000	1.000	0.469
Standard.$^+$	1	1.000	1.000	1.000	1.000	1.000	0.413

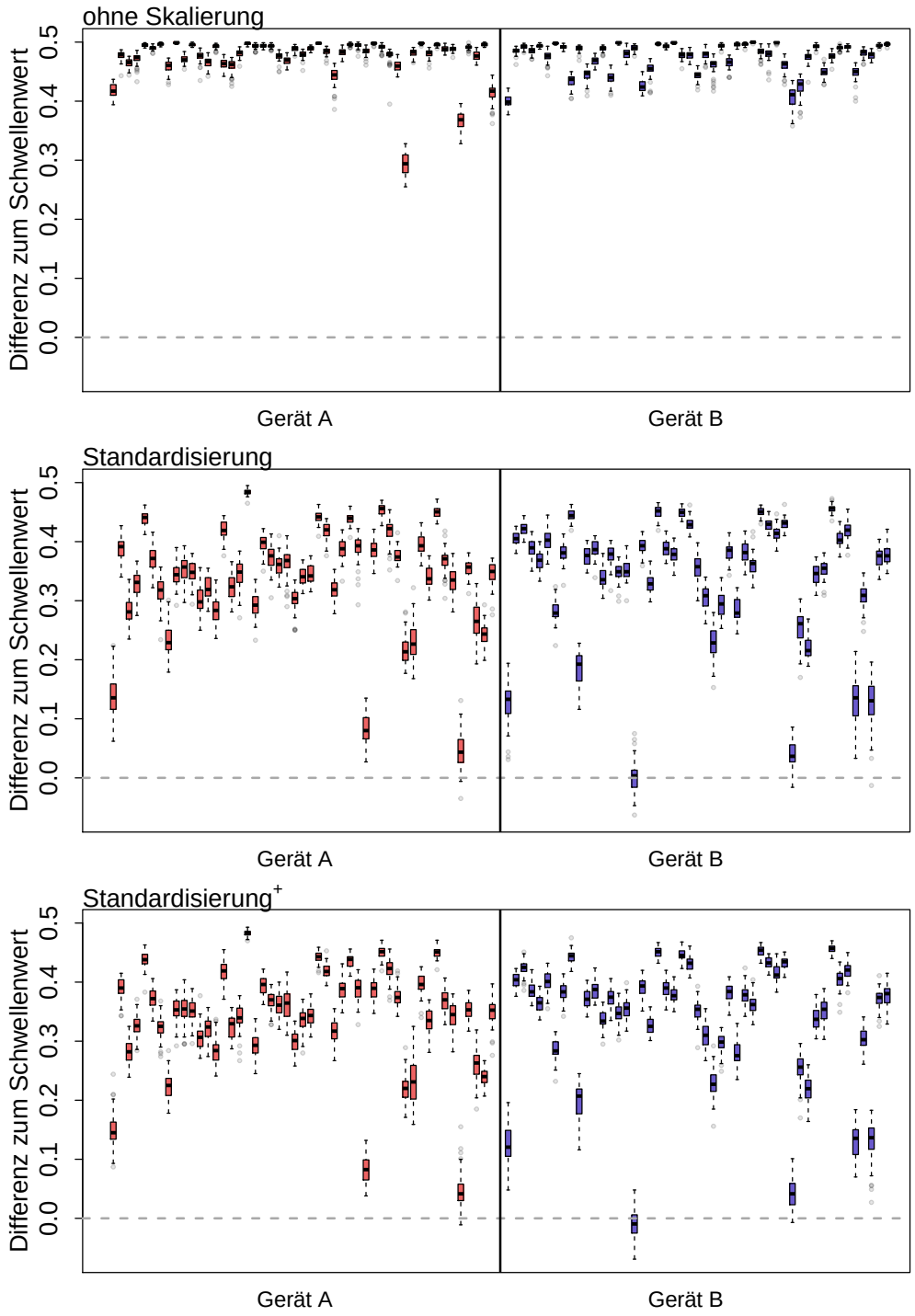


Abbildung B.15: Differenz der Wahrscheinlichkeit für die wahre Klasse zum **Schwellenwert 0.5** für alle Beobachtungen der Kreuzvalidierungswiederholungen. Betrachtet werden die Daten der **manuellen** Peakerkennung, jeweils ohne Skalierung, mit Standardisierung und mit Standardisierung$^+$ im **Geräte**-Klassifikationsproblem.

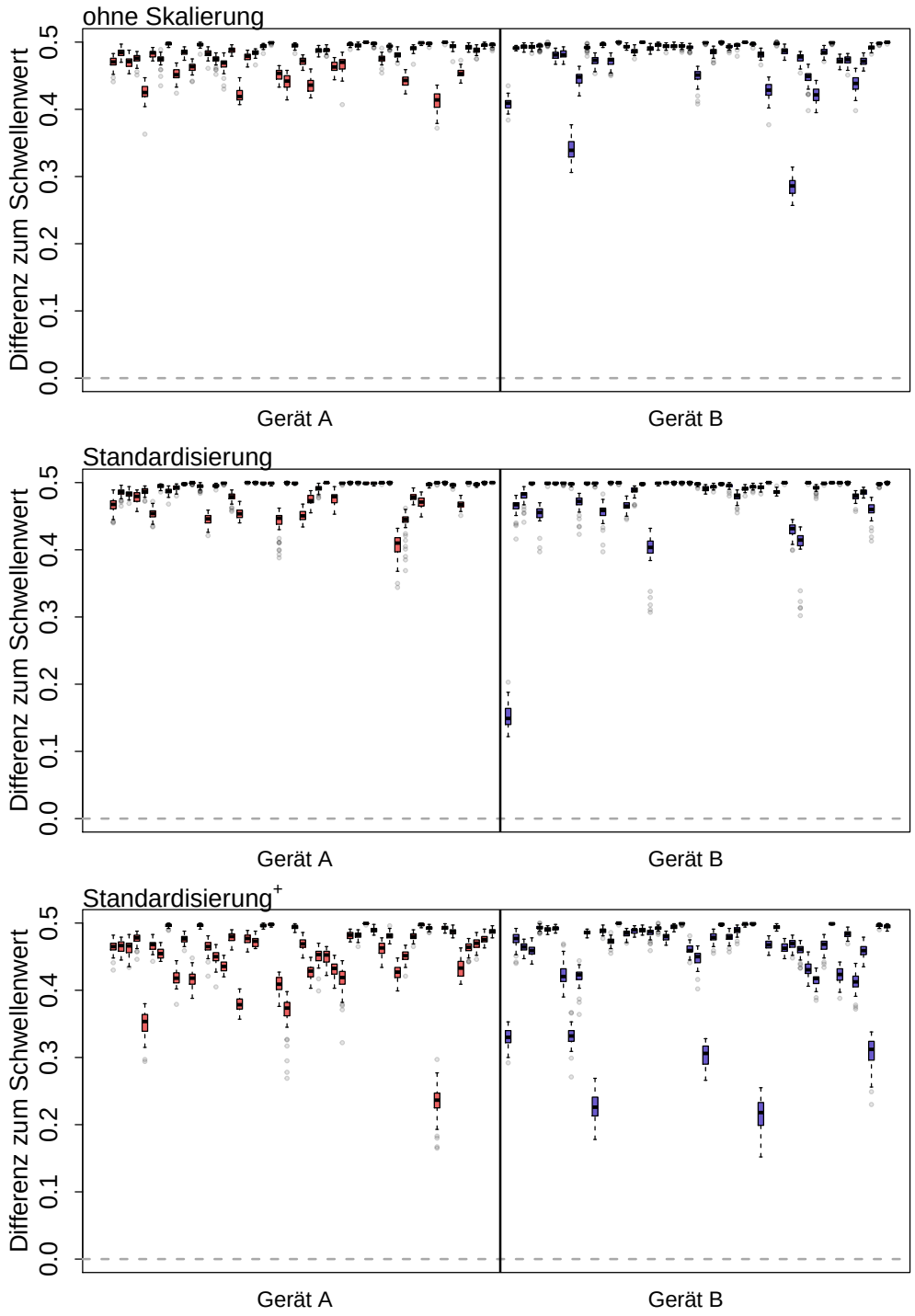


Abbildung B.16: Differenz der Wahrscheinlichkeit für die wahre Klasse zum **Schwellenwert 0.5** für alle Beobachtungen der Kreuzvalidierungswiederholungen. Betrachtet werden die Daten der automatischen Peakerkennung **ohne Alignierung**, jeweils ohne Skalierung, mit Standardisierung und mit Standardisierung$^+$ im **Geräte**-Klassifikationsproblem.

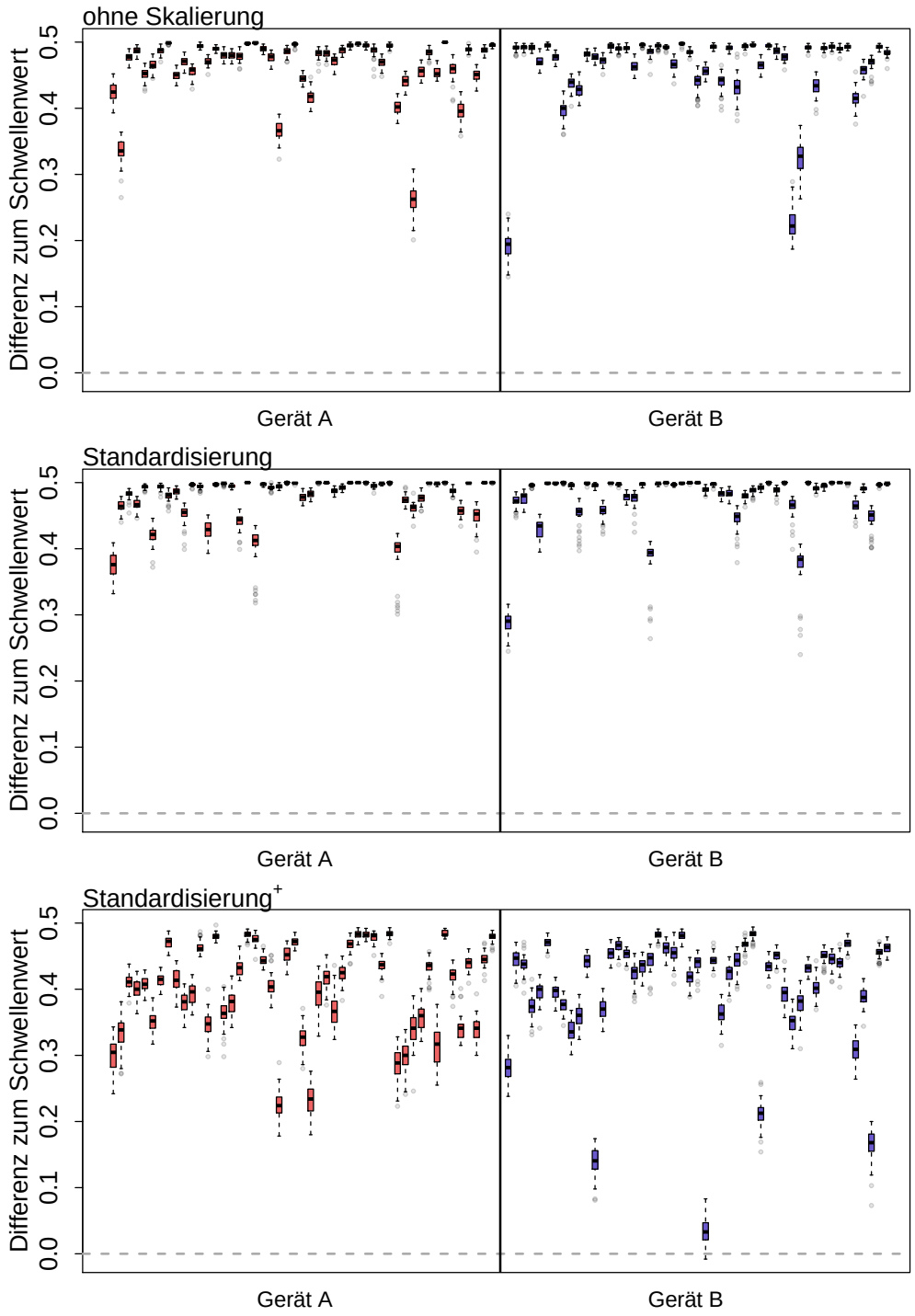


Abbildung B.17: Differenz der Wahrscheinlichkeit für die wahre Klasse zum **Schwellenwert 0.5** für alle Beobachtungen der Kreuzvalidierungswiederholungen. Betrachtet werden die Daten der automatischen Peakerkennung **mit Alignierung**, jeweils ohne Skalierung, mit Standardisierung und mit Standardisierung$^+$ im **Geräte**-Klassifikationsproblem.

Tabelle B.3: Gemittelte Gütemaßzahlen der kreuzvalidierten Klassifikationsergebnisse für die Zielvariable **Geschlecht**, sowie das arithmetische Mittel der verwendeten Schwellenwerte s (entsprechend der **Prävalenz-Methode**). Die Klassifikation wurde für die unskalierten Daten und für zwei Skalierungen durchgeführt, Standardisierung und Standardisierung mit separater Behandlung der Nullen (= Standardisierung$^+$).

Skalierung	AUC	ACC	TPR	TNR	PPV	NPV	s
Manuelle Peakerkennung							
ohne	0.402	0.428	0.450	0.403	0.459	0.392	0.510
Standard.	0.426	0.456	0.495	0.412	0.488	0.419	0.514
Standard.$^+$	0.435	0.456	0.481	0.428	0.486	0.422	0.515
Automatische Peakerkennung							
ohne	0.534	0.499	0.522	0.473	0.528	0.467	0.508
Standard.	0.483	0.473	0.502	0.441	0.504	0.438	0.514
Standard.$^+$	0.525	0.492	0.515	0.466	0.522	0.459	0.514
Automatische Peakerkennung mit Alignierung							
ohne	0.571	0.527	0.561	0.489	0.555	0.496	0.515
Standard.	0.509	0.501	0.533	0.464	0.530	0.466	0.515
Standard.$^+$	0.585	0.546	0.562	0.528	0.574	0.517	0.510

Tabelle B.4: Gemittelte Gütemaßzahlen der kreuzvalidierten Klassifikationsergebnisse für die Zielvariable **Rauchen**, sowie das arithmetische Mittel der verwendeten Schwellenwerte s (entsprechend der **Prävalenz-Methode**). Die Klassifikation wurde für die unskalierten Daten und für zwei Skalierungen durchgeführt, Standardisierung und Standardisierung mit separater Behandlung der Nullen (= Standardisierung$^+$).

Skalierung	AUC	ACC	TPR	TNR	PPV	NPV	s
Manuelle Peakerkennung							
ohne	0.501	0.516	0.507	0.518	0.191	0.825	0.497
Standard.	0.532	0.525	0.516	0.527	0.197	0.829	0.497
Standard.$^+$	0.541	0.532	0.524	0.534	0.203	0.833	0.497
Automatische Peakerkennung							
ohne	0.542	0.525	0.521	0.526	0.199	0.830	0.506
Standard.	0.529	0.527	0.529	0.526	0.200	0.833	0.501
Standard.$^+$	0.534	0.524	0.517	0.526	0.197	0.828	0.501
Automatische Peakerkennung mit Alignierung							
ohne	0.567	0.542	0.551	0.540	0.213	0.843	0.508
Standard.	0.596	0.559	0.582	0.554	0.227	0.855	0.507
Standard.$^+$	0.620	0.578	0.589	0.576	0.240	0.862	0.504

Fragebogen

Fragebogen Nr. _____

Bitte füllen Sie den Fragebogen vollständig aus.

PERSÖNLICHE ANGABEN

1. Geschlecht:

☐ männlich ☐ weiblich

2. Alter: _____

ERNÄHRUNG

3. Wann haben Sie Ihre letzte vollwertige Mahlzeit eingenommen?

☐ < 30 min
☐ 30 – 60 Minuten
☐ 1 – 3 Stunden
☐ 3 – 6 Stunden
☐ 6 – 12 Stunden
☐ > 12 Stunden

4. Wann haben Sie das letzte Mal etwas gegessen (auch Obst, Joghurt,…)?

☐ < 30 min
☐ 30 – 60 Minuten
☐ 1 – 3 Stunden
☐ 3 – 6 Stunden
☐ 6 – 12 Stunden
☐ > 12 Stunden

5. Haben Sie in den letzten 24 Stunden Zitrusfrüchte konsumiert (z.B. Mandarine, Zitrone, Orangensaft,...)?

☐ Ja

5.a Wann haben Sie das letzte Mal Zitrusfrüchte konsumiert?

- ☐ < 30 min
- ☐ 30 – 60 Minuten
- ☐ 1 – 3 Stunden
- ☐ 3 – 6 Stunden
- ☐ 6 – 12 Stunden
- ☐ > 12 Stunden

☐ Nein

6. Welche der folgenden Getränke haben Sie innerhalb der letzten Stunde zu sich genommen? (Mehrfachauswahl möglich)

☐	Leitungswasser	☐	Abgefülltes Wasser
☐	Saft	☐	Limonade/Cola
☐	Tee	☐	kaffeehaltige Getränke
☐	Kakao	☐	alkoholische Getränke
☐	Sonstige und zwar:	☐	keine

_____

RAUCHERSTATUS

7. Rauchen Sie zurzeit Zigaretten (keine E-Zigaretten) – wenn auch nur gelegentlich?

☐ Ja

7.a Wie viel rauchen Sie zurzeit gewöhnlich?

Anzahl Zigaretten pro Tag: _____

Falls Sie regelmäßig* rauchen:

7.b Wann haben Sie angefangen, regelmäßig* zu rauchen?

Im Alter von _____ Jahren bzw. im Jahr _____

7.c Wann haben Sie die letzte Zigarette geraucht?

Vor...

- ☐ < 30 min
- ☐ 30 – 60 Minuten
- ☐ 1 – 3 Stunden
- ☐ 3 – 6 Stunden
- ☐ 6 – 12 Stunden
- ☐ > 12 Stunden

☐ Nein, ich habe früher regelmäßig* geraucht, aber jetzt nicht mehr

7.d Wie viel haben Sie früher gewöhnlich geraucht?

Anzahl Zigaretten pro Tag: _____

7.e Wann haben Sie angefangen, regelmäßig* zu rauchen?

Im Alter von _____ Jahren bzw. im Jahr _____

7.f Wann haben Sie aufgehört, regelmäßig* zu rauchen?

Im Alter von _____ Jahren bzw. im Jahr _____

☐ Nein, ich habe noch nie regelmäßig* geraucht

** Unter regelmäßig verstehen wir:*
*1 Zigarette pro Tag oder mindestens 5 Zigaretten pro Woche **oder** mindestens 1 Packung Zigaretten pro Monat für mindestens 6 Monate*

8. Rauchen Sie zurzeit (auch) andere Tabakwaren als Zigaretten oder E-Zigaretten?

☐ Ja

8.a Welche Tabakwaren/E-Zigaretten rauchen Sie zurzeit - wenn auch nur gelegentlich? (Mehrfachnennungen möglich)

☐	E-Zigarette	☐	Zigarillo
☐	Zigarre	☐	Pfeife
☐	Wasserpfeife/Shisha	☐	Sonstige und zwar:

_____

☐ Nein

Ihre Kommentare

Haben Sie Kommentare zum Fragebogen?

_____

_____

_____

_____

Vielen Dank für Ihre Teilnahme!

Einwilligungserklärung

Information und Einwilligungserklärung
zur Teilnahme an der Beobachtungsstudie

Studie zur Ermittlung von Einflussfaktoren auf die Zusammensetzung der Atemluft

Sehr geehrte Teilnehmerin, sehr geehrter Teilnehmer!

Wir laden Sie ein an der oben genannten Beobachtungsstudie teilzunehmen.

Ihre Teilnahme an dieser Studie erfolgt freiwillig. Sie können jederzeit ohne Angabe von Gründen aus der Studie ausscheiden.

1. Was ist der Zweck dieser Studie?

Der Zweck dieser Beobachtungsstudie ist die Identifizierung von Einflussfaktoren auf die menschliche Atemluft, z.B. Geschlecht, Rauchverhalten und Ernährung.

2. Wie läuft die Beobachtungsstudie ab?

Diese Studie wird in den Räumlichkeiten der B&S Analytik GmbH durchgeführt.

Sie werden gebeten, einen Fragebogen auszufüllen. Darüber hinaus bitten wir Sie, zweimal Ihre Atemluft mit Hilfe von MCC-IMS-Geräten messen zu lassen. Vor der zweiten Messung soll ein Glas Orangensaft getrunken werden.

3. Worin liegt der Nutzen einer Teilnahme an der Beobachtungsstudie?

Es ist nicht zu erwarten, dass Sie aus Ihrer Teilnahme an dieser Studie gesundheitlichen Nutzen ziehen werden, aber möglicherweise werden künftig Patienten von den Ergebnissen profitieren.

4. Gibt es Risiken, Beschwerden und Begleiterscheinungen?

Nein.

5. In welcher Weise werden die im Rahmen dieser Beobachtungsstudie gesammelten Daten verwendet?

Ihr Name wird ausschließlich für den Zweck Ihrer Einwilligung mit Hilfe dieses Formulars erhoben und in elektronischer Form nicht mit ihren Messdaten oder Ihrem Fragebogen verbunden.

Die Weitergabe der Daten erfolgt ausschließlich zu wissenschaftlichen Zwecken und Sie werden ausnahmslos nicht namentlich genannt. Auch in etwaigen wissenschaftlichen Veröffentlichungen der Daten dieser Studie werden Sie nicht namentlich genannt.

6. Einwilligungserklärung

Name des Probanden in Druckbuchstaben: .....

Geb.Datum: ..... Code: .....

Ich habe dieses Informationsblatt gelesen und verstanden. Alle meine Fragen wurden beantwortet und ich habe zurzeit keine weiteren Fragen mehr.

Mit meiner persönlich datierten Unterschrift gebe ich hiermit freiwillig mein Einverständnis, dass meine Daten gespeichert und ohne direkten Personenbezug für wissenschaftliche Zwecke verwendet und weitergegeben werden dürfen.

Ich weiß, dass ich diese Zustimmungen jederzeit und ohne Angabe von Gründen widerrufen kann.

.....
(Datum und Unterschrift des Probanden)

Tabellenverzeichnis

Abbildungsverzeichnis

Literaturverzeichnis

Bader, S., Urfer, W. und Baumbach, J. I. (2006). „Reduction of ion mobility spectrometry data by clustering characteristic peak structures“. In: *J. Chemom.* 20, Seiten 128–135.

Baumbach, J. I. (2009). „Ion mobility spectrometry coupled with multi-capillary columns for metabolic profiling of human breath“. In: *J. Breath Res.* 3.3, 034001.

Beauchamp, J. (2011). „Inhaled today, not gone tomorrow: pharmacokinetics and environmental exposure of volatiles in exhaled breath“. In: *J. Breath Res.* 5.3, 037103.

Blanchet, L., Smolinska, A., Baranska, A., Tigchelaar, E., Swertz, M., Zhernakova, A., Dallinga, J. W., Wijmenga, C. und van Schooten, F. J. (2017). „Factors that influence the volatile organic compound content in human breath“. In: *J. Breath Res.* 11.1.

Bödeker, B., Vautz, W. und Baumbach, J. I. (2008). „Visualisation of MCC/IMS-data“. In: *Int. J. Ion Mobil. Spec.* 11.1–4, Seiten 77–81.

Breiman, L. (2001). „Random Forests“. In: *Machine Learning* 45.1, Seiten 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A. und Stone, C. J. (1984). *Classification and regression trees*. Belmont: Wadsworth.

Cao, W. und Duan, Y. (2006). „Breath Analysis: Potential for Clinical Diagnosis and Exposure Assessment“. In: *Clinical Chemistry* 52.5, Seiten 800–811.

Cheng, Z. J., Warwick, G., Yates, D. H. und Thomas, P. S. (2009). „An electronic nose in the discrimination of breath from smokers and non-smokers: a model for toxin exposure“. In: *J. Breath Res.* 3, 036003.

Cumeras, R., Schneider, T., Favrod, P., Figueras, E., Gràcia, I., Maddula, S. und Baumbach, J. I. (2012). „Stability and alignment of MCC/IMS devices“. In: *Int. J. Ion Mobil. Spec.* 15.1.

D'Addario, M., Kopczynski, D., Baumbach, J. I. und Rahmann, S. (2014). „A modular computational framework for automated peak extraction from ion mobility spectra“. In: *BMC Bioinformatics* 15.25.

Dempster, A. P., Laird, N. M. und Rubin, D. B. (1977). „Maximum Likelihood from Incomplete Data via the EM Algorithm“. In: *Journal of the Royal Statistical Society* 39.1, Seiten 1–38.

Di Francesco, F., Fuoco, R., Trivella, M. G. und Ceccarini, A. (2005). „Breath analysis: trends in techniques and clinical applications“. In: *Microchem. J.* 79.1–2.

Egorov, A., König, A., Köppen, M., Kühn, H., Kullack, I., Kuthe, E., Mitkovska, S., Niehage R. Pawelko, A., Sträßer, M. und Striewe, C. (2014). *Ressourcenbeschränkte Analyse von Ionenmobilitätsspektren mit dem Raspberry Pi.* Technical Report. Faculty of computer science, TU Dortmund.

Ernstgård, L., Sjögren, B., Warholm, M. und Johanson, G. (2003). „Sex differences in the toxicokinetics of inhaled solvent vapors in humans 2. 2-Propanol“. In: *Toxicol. Appl. Pharmacol.* 193, Seiten 158–167.

Ester, M., Kriegel, H.-P., Sander, J. und Xu, X. (1996). „A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise“. In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining.* Herausgegeben von E. Simoudis, J. Han und U. Fayyad. AAAI Press, Seiten 226–231.

Fink, T., Baumbach, J. I. und Kreuer, S. (2014). „Ion mobility spectrometry in breath research“. In: *J. Breath Res.* 8.2, 027104.

Freeman, E. A. und Moisen, G. G. (2008). „A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa“. In: *Ecological Modelling* 217.1–2, Seiten 48–58.

Hastie, T., Tibshirani, R. und Friedman, J. (2009). *The Elements of Statistical Learning.* 2. Auflage. New York: Springer.

Hauschild, A.-C., Baumbach, J. I. und Baumbach, J. (2012). „Integrated statistical learning of metabolic ion mobility spectrometry profiles for pulmonary disease identification“. In: *Genet. Mol. Res.* 11.3, Seiten 2733–2744.

Hauschild, A.-C., Kopczynski, D., D'Addario, M., Baumbach, J. I., Rahmann, S. und Baumbach, J. (2013). „Peak Detection Method Evaluation for Ion Mobility Spectrometry by Using Machine Learning Approaches“. In: *Metabolites* 3, Seiten 277–293.

Hauschild, A.-C., Schneider, T., Pauling, J., Rupp, K., Jang, M., Baumbach, J. I. und Baumbach, J. (2012). „Computational Methods for Metabolomic Data Analysis of Ion Mobility Spectrometry Data — Reviewing the State of the Art“. In: *Metabolites* 2, Seiten 733–755.

Holm, S. (1979). „A Simple Sequentially Rejective Multiple Test Procedure“. In: *Scand. J. Stat.* 6.2, Seiten 65–70.

Horsch, S., Kopczynski, D., Baumbach, J. I., Rahnenführer, J. und Rahmann, S. (2015). „From raw ion mobility measurements to disease classification: a comparison of analysis processes“. In: *PeerJ PrePrints* 3, e1591.

Horsch, S., Kopczynski, D., Kuthe, E., Baumbach, J. I., Rahnenführer, J. und Rahmann, S. (2017). „A detailed comparison of analysis processes for MCC-IMS data in disease classification — Automated methods can replace manual peak annotations“. In: *PLOS ONE* 12.9, e0184321.

Horsch, S., Rahnenführer, J. und Baumbach, J. I. (2019). „Statistical analysis of MCC-IMS data for two group comparisons— an exemplary study on two devices“. In: *J. Breath Res.* 13, 036011.

Jordan, A., Hansel, A., Holzinger, R. und Lindinger, W. (1995). „Acetonitrile and benzene in the breath of smokers and non-smokers investigated by proton transfer reaction mass spectrometry (PTR-MS)“. In: *Int. J. Mass Spectrom. Ion Process,* 148.1–2, Seiten L1–L3.

Kim, K.-H., Jahan, S. A. und Kabir, E. (2012). „A review of breath analysis for diagnosis of human health“. In: *Trends Anal. Chem.* 33, Seiten 1–8.

Kischkel, S., Miekisch, W., Sawacki, A., Straker, E. M., Trefz, P., Amann, A. und Schubert, J. K. (2010). „Breath biomarkers for lung cancer detection and assessment of smoking related effects - confounding variables, influence of normalization and statistical algorithms“. In: *Clin. Chim. Acta* 411.21–22, Seiten 1637–1644.

Kopczynski, D. (2017). „Resource-Constrained Analysis of Ion Mobility Spectrometry Data“. Dissertation. TU Dortmund.

Kopczynski, D. und Rahmann, S. (2014a). „An Online Peak Extraction Algorithm for Ion Mobility Spectrometry Data“. In: *Algorithms in Bioinformatics*. Herausgegeben von D. Brown und B. Morgenstern. Band 8701, Seiten 232–246.

– (2014b). *Using the Expectation Maximization Algorithm with Heterogeneous Mixture Components for the Analysis of Spectrometry Data*. arXiv: `1405.5501 [cs.OH]`.

Kushch, I., Schwarz, K., Schwentner, L., Baumann, B., Dzien, A., Schmid, A., Unterkofler, K., Gastl, G., Španel, P., Smith, D. und Amann, A. (2008). „Journal of Breath Research Compounds enhanced in a mass spectrometric profile of smokers' exhaled breath versus non-smokers as determined in a pilot study using PTR-MS“. In: *J. Breath Res.* 2.2, Seite 026002.

Latza, U., Hoffmann, W., Terschüren, C., Chang-Claude, J., Kreuzer, M., Schaffrath Rosario, A., Kropp, S., Stang, A., Ahrens, W. und Lampert, T. (2005). *Erhebung, Quantifizierung und Analyse der Rauchexposition in epidemiologischen Studien*. Technischer Bericht. Robert Koch Institut.

Liaw, A. und Wiener, M. (2002). „Classification and Regression by randomForest“. In: *R News* 2.3, Seiten 18–22.

Manel, S., Williams, H. C. und Ormerod, S. J. (2001). „Evaluating presence–absence models in ecology: the need to account for prevalence“. In: *Journal of Applied Ecology* 38, Seiten 921–931.

McWilliams, A., Beigi, P., Srinidhi, A., Lam, S. und MacAulay, C. E. (2015). „Sex and Smoking Status Effects on the Early Detection of Early Lung Cancer in High-Risk Smokers Using an Electronic Nose“. In: *IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING* 62.8, Seiten 2044–2054.

Meyer, David, Dimitriadou, Evgenia, Hornik, Kurt, Weingessel, Andreas und Leisch, Friedrich (2014). *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*. R package version 1.6-4.

Miekisch, W., Schubert, J. und Noeldge-Schomburg, G. F. E. (2004). „Diagnostic potential of breath analysis—focus on volatile organic compounds“. In: *Clinica Chimica Acta* 347.1–2, Seiten 25–39.

Morgan, J. N. und Sonquist, J. A. (1963). „Problems in the Analysis of Survey Data, and a Proposal“. In: *Journal of the American Statistical Association* 58.302, Seiten 415–434.

Needleman, S. B. und Wunsch, C. D. (1970). „A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins“. In: *J. Mol. Biol.* 48.3, Seiten 443–453.

Pleil, J. D., Stiegel, M. A. und Risby, T. H. (2013). „Clinical breath analysis: discriminating between human endogenous compounds and exogenous (environmental) chemical confounders“. In: *J. Breath Res.* 7.1, Seite 017107.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Rabis, T., Sommerwerck, U., Anhenn, O., Darwiche, K., Freitag, L., Teschler, H., Bödeker, B., Maddula, S. und Baumbach, J. I. (2011). „Detection of infectious agents in the airways by ion mobility spectrometry of exhaled breath“. In: *Int. J. Ion Mobil. Spec.* 14.4, Seiten 187–195.

Rahmann, S., Wittkop, T., Baumbach, J., Martin, M., Truß, A. und Böcker, S. (2007). „Exact and Heuristic Algorithms for Weighted Cluster Editing“. In: *Computational Systems Bioinformatics Conference*. Band 6, Seiten 391–401.

Ridgeway, G. (2013). *gbm: Generalized Boosted Regression Models*. R package version 2.1.

– (2019). *Generalized Boosted Models: A guide to the gbm package*. URL: `https://cran.r-project.org/web/packages/gbm/index.html`.

Risby, T. H. (2008). „Critical issues for breath analysis“. In: *J. Breath Res.* 2.3, Seite 030302.

Schliep, K. und Hechenbichler, K. (2014). *kknn: Weighted k-Nearest Neighbors*. R package version 1.2-5.

Silverman, B. W. und Jones, M. C. (1989). „E. Fix and J.L. Hodges (1951): An Important Contribution to Nonparametric Discriminant Analysis and Density Estimation: Commentary on Fix and Hodges (1951)“. In: *International Statistical Review* 57.3, Seiten 233–247.

Therneau, T., Atkinson, B. und Ripley, B. (2015). *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1-9.

Vautz, W., Nolte, J., Fobbe, R. und Baumbach, J. I. (2009). „Breath analysis-performance and potential of ion mobility spectrometry“. In: *J. Breath Res* 3.3, 036004.

Wehberg, S., Sauerbrei, W. und Schumacher, M. (2007). „Diagnosestudien: Wertigkeit der Sonographie bei der Differenzierung von gut- und bösartigen Brusttumoren bei Patientinnen mit klinischen Symptomen“. In: *Methodik klinischer Studien*. Berlin, Heidelberg: Springer, Seiten 323–330.

Westhoff, M., Litterst, P., Maddula, S., Bödeker, B., Rahmann, S., Davies, A. N. und Baumbach, J. I. (2010). „Differentiation of chronic obstructive pulmonary disease (COPD) including lung cancer from healthy control group by breath analysis using ion mobility spectrometry“. In: *Int. J. Ion Mobil. Spec.* 13.3–4, Seiten 131–139.