Einsatz von Verfahren der Künstlichen Intelligenz insbesondere des Maschinellen Lernens in der amtlichen Statistik

Abschlussbericht der Ludwig-Maximilians-Universität München
in Zusammenarbeit mit dem Munich Center for Machine Learning zum
Kooperationsprojekt mit Destatis

Bernd Bischl,
Thomas Augustin, Andreas Bender, Ludwig Bothmann,
Florian Karl, Anne-Laure Boulesteix, Roman Hornung, Thomas Fetzer,
Julian Rodemann, Christoph Kern, Frauke Kreuter

Berichtszeitraum: 20.12.2021 – 30.06.2024

Inhaltsverzeichnis

Management Fassung

Maschinelles Lernen (ML) gewinnt auch in der amtlichen Statistik zunehmend an Bedeutung. Aufbauend auf dem internen Bericht "Proof of Concept Machine Learning" (Beck, Dumpert und Feuerhake, 2018) wurden in diesem gemeinsamen Projekt von Destatis mit der Ludwig-Maximilians-Universität (LMU) München und in Zusammenarbeit mit dem Munich Center for Machine Learning (MCML), Fraunhofer Institut und der Universität Mannheim weitere Aspekte des Einsatzes von ML in der amtlichen Statistik untersucht und vertieft. Diese fokussieren sich auf für die amtliche Statistik und ihre verantwortungsvolle Rolle beim Umgang mit Daten und dem Einsatz von ML zentrale Aspekte: Schätzung des Generalisierungsfehlers (Abschnitt 1), Quantifizierung von Unsicherheit (Abschnitte 1.3 und 4), Interpretierbarkeit (Abschnitt 2), Einsatz von ML bei stichprobenbasiert erhobenen Daten (Abschnitt 3), technische Aspekte beim konkreten Einsatz von ML in der Praxis und insbesondere Destatis (Abschnitt 5), Fairness (Abschnitt 6) und Beleuchtung von ethischen und rechtlichen Aspekten beim Einsatz von ML (Abschnitt 7).

1. Der Generalisierungsfehler (GE) ist der Fehler, den ein ML Model bei der Prädiktion auf neuen Daten macht. Die Schätzung dieses Fehlers basiert in der Regel auf "Resampling"-basierten Verfahren wie der bekannten Kreuzvalidierung. Die realistische Abschätzung des GE sowie die Quantifizierung seiner Unsicherheit, z.B. in Form von Konfidenzintervallen, sind somit essentielle Bestandteile des verantwortungsvollen Einsatzes von ML in der Praxis. In Abschnitt 1 wurden in diesem Zusammenhang zwei Beiträge erarbeitet:

   (a) Der erste, Abschnitt 1.2 beschäftigt sich mit der Schätzung des GE in komplexen Situationen, in denen der Einsatz von Standardverfahren zur Schätzung des GE zu Verzerrungen (Bias) führen würde. Insbesondere behandelt werden geclusterte Daten, räumliche Daten, Daten aus komplexen Stichproben, Daten mit Konzeptdrift sowie Daten mit hierarchisch strukturierten Outcomes.

   (b) Der zweite, Abschnitt 1.3, fasst den aktuellen Erkentnissstand der, aufgrund von Abhängigkeiten der resamplingbasierten Trainings- und Testdaten nicht trivialen Schätzung der Unsicherheit des GE zusammen und vergleicht die Ansätze in einer umfangreichen empirischen Studie.

2. Neben der Schätzung des GE ist auch die Interpretierbarkeit von Verfahren des Maschinellen Lernens ein zentraler Aspekt, um Nachvollziehbarkeit und Vertrauen in die von ML beeinflussten Entscheidungen (Prädiktionen) zu erhöhen. Abschnitt 2 fasst die Ziele des Interpretierbaren Maschinellen Lernens (IML) zusammen und gibt eine Übersicht über die im Rahmen dieses Projekts entstandenen Beiträge.

   (a) Abschnitt 2.3.1 beschreibt eine Zusammenfassung und einen Leitfaden zu sogenannten "Feature Importance" Verfahren, welche auf beliebige ML Modelle angewandt werden können, um zu bestimmen, welche Variablen für das Modell von besonderer Wichtigkeit waren.

   (b) Ein weiterer Beitrag beschäftigt sich mit sogenannten kontrafaktischen Erklärungen, die verwendet werden können um zu verstehen, wie eine spezifische Prädiktion zustande gekommen ist bzw. was sich ändern müsste, um die Prädiktion/Entscheidung zu ändern (Abschnitt 2.3.2).

   (c) Letztlich wurde in diesem Bereich auch eine neue Software entwickelt, die dem Anwender eine nützliche Übersicht über verschiedene Aspekte des geschätzten ML Modells Auskunft gibt (Abschnitt 2.4).

3. In Abschnitt 3 wird der Einsatz von ML bei komplexen Stichprobendesigns untersucht. Dies ist im Kontext der amtlichen Statistik besonders wichtig, da viele Daten nicht als unabhängige und identische Ziehungen aus der Grundpopulation betrachtet werden können. Neben einer ausführlichen Übersicht über die relevanten Grundlagen und Konzepte wird hier insbesondere detailliert eine Korrektur beim Einsatz von Regressionsbäumen sowie Random Forests vorgestellt.

4. In Abschnitt 4 wird die Schätzung von Unsicherheit beim Einsatz von ML wieder aufgegriffen und ausführlich dargestellt, wie mengenwertige Verfahren zu ihrer Quantifizierung beitragen können. Insbesondere wurden hier drei Beiträge erarbeitet:

   (a) Abschnitt 4.3.1 beschreibt, wie die Unsicherheit bezüglich der Modelspezifikation im Kontext der Bayesianischen Optimierung deren Performanz beeinflusst. Darauf aufbauend wird eine robuste Variante der Bayesianischen Optimierung vorgeschlagen, die auf einer mengenwertige Spezifikation des Modells beruht.

   (b) In Abschnitt 4.3.2 wird verdeutlicht, wie wichtig eine verlässliche Unsicherheitsquantifizierung im Bereich des halb-überwachten Lernens ist, bei dem Modelle anhand gelabelter und ungelabelter Daten trainiert werde. Basierend auf Ideen aus der Bayesianischen und robusten Statistik wird mittels mengenwertiger Verfahren eine Methode zur verlässlichen Auswahl pseudo-gelabelter Daten im halb-überwachten Lernen entwickelt.

   (c) Eine Übertragung und Erweiterung aktueller Entwicklungen aus der theoretischen Entscheidungstheorie auf die Wahl von Algorithmen ist Gegenstand von Abschnitt 4.3.3. Dort wird ein neues Konzept stochastischer Dominanz genutzt, um Algorithmen anhand von Benchmarkstudien statistisch gesichert (partiell) zu ordnen.

5. In Abschnitt 5 wird Machine Learning Operations (MLOps) im Kontext der amtlichen Statistik untersucht. MLOps vereint Tools und Best Practices, um Deployment, Betrieb, Überwachung und Wartung von ML-Modellen zu vereinfachen und bessere Reproduzierbarkeit von Ergebnissen zu ermöglichen. Nach einer Untersuchung des Status quo bzgl. ML und MLOps bei Destatis sowie einer Anforderungsanalyse wurde eine detaillierte Handlungsempfehlung erarbeitet, die insbesondere eine MLOps-Architektur für die speziellen Anforderungen der amtlichen Statistik vorschlägt. Abschnitt 5 ist eine kurze Zusammenfassung der wichtigsten Erkenntnisse aus dem Abschlussbericht (siehe Karl, Kaminwar und Frechen, 2024) zu diesem Thema.

6. Der Abschnitt Fairness 6 reflektiert zunächst die Beziehungen der amtlichen Statistik zu datenbasierten Entscheidungen, insbesonderen zu automatisierten Entscheidungssystemen (ADM). Eine wichtige Rolle kommt zudem der Perspektive zu, wofür Machine Learning in der amtlichen Statistik eigentlich hauptsächlich verwendet wird: weniger im Sinne einer klassischen Datenanalyse, sondern vor allem in der Datenerhebung, Datenaufbereitung und bei sonstigen Schritten der Genese und Produktion von Daten und Datenprodukten.

   Ausgehend von dieser Basis wird der Bezug hergestellt zu den Qualitätsdimensionen, die die amtliche Statistik prägen. Insbesondere wird hierbei auf dem QF4SA Framework von Yung u. a. (2022) aufgebaut, welcher die Dimensionen Interpretierbarkeit, Genauigkeit, Robustheit,

Reproduzierbarkeit, Kosteneffizienz und Zeiteffizienz umfasst. Das Arbeitspaket untersucht einerseits das Wechselspiel einer jeder dieser Dimensionen mit dem Konzept Fairness, argumentiert aber auch, weshalb Fairness als eigenständige (Qualitäts-)Dimension etabliert werden sollte. In der einschlägigen Literatur existieren verschiedene Fairness-Konzepte, die (großteils) so zusammengefasst werden können, dass die (Prädiktionen der) ML Modelle gleich gut funktionieren sollen für verschiedene gesellschaftliche Gruppen (Alter, Geschlecht, Herkunft, etc.).

Eine wichtige Schlussfolgerung des Arbeitspakets ist, dass die (frühzeitige und kontinuierliche) Berücksichtigung von Fairness in der amtlichen Statistik eine echte Bereicherung darstellen kann – nicht zuletzt dadurch, dass so frühzeitig spätere (Fairness-)Probleme erkannt werden, sodass nicht Ressourcen in ML Modelle gesteckt werden, die sich letztendlich nicht als praxistauglich herausstellen werden.

7. Ziel des Abschnitts 7 war es, die rechtlichen und ethischen Auswirkungen des Einsatzes von Machine Learning in der amtlichen Statistik zu beleuchten. Dazu wurde untersucht, inwieweit das Datenschutzrecht sowie die geplante europäische KI-Verordnung den Einsatz von Machine Learning beeinflussen, um zu überprüfen, ob sich für im Rahmen einer Erhebung befragte Personen eine Verschlechterung mit Hinblick auf den Schutz ihrer persönlichen Daten ergibt. Der erste Teil des Abschnitts analysiert daher, inwieweit eine Anwendbarkeit des Datenschutzrechts durch den Einsatz moderner Datenverarbeitungsverfahren auch bei statistischen Daten gegeben ist und welche Folgen sich daraus ergeben. Im zweiten Teil wird untersucht welche Verpflichtungen sich für statistische Ämter aus der KI-Verordnung ergeben werden.

1 Schätzung des Generalisierungsfehlers und seiner Unsicherheit

1.1 Einführung

Die Validierung von prädiktiven Machine Learning (ML)-Modellen ist ein sehr wichtiger Teil der Entwicklung solcher Modelle. Ohne Schätzung der zu erwartenden Fehlerrate eines ML-Modells kann es nicht zur Anwendung empfohlen werden. Dies ist besonders wichtig in Anwendungen der amtlichen Statistik wo häufig hohe Ansprüche an die Qualität solcher Modelle gestellt werden müssen oder es wichtig ist, genau über die Performanz solcher Modelle Bescheid zu wissen. Der Generalisierungsfehler (GE) von prädiktiven Modellen wird in der Regel über Resamplingverfahren wie Kreuzvalidierung (KV) geschätzt; ausgenommen hiervon sind Fälle in denen die verfügbaren Datensätze sehr groß sind, wo ein einfacher Split in Trainings- und Testdaten genügt.

Während das Grundprinzip von Resamplingverfahren simpel ist, ist ihre Anwendung und ihre Interpretation in der Praxis mit hohen methodischen Herausforderungen verbunden. Dieses Arbeitspaket widmete sich diesen methodischen Schwierigkeiten und gliedert sich in zwei Teile, deren Hintergrund und Motivation im Folgenden einleitend beschrieben werden.

Der erste Teil beschäftigte sich mit der Schätzung des GEs in komplexen Datensituationen. In der amtlichen Statistik liegen häufig komplexe Datensituationen vor. Dazu zählen räumliche Daten, geclusterte Daten oder Stichproben deren Beobachtungen mit ungleichen Zugwahrscheinlichkeiten aus der Population gezogen wurden. Es ist bekannt, dass in solchen Datensituationen konventionelle Resamplingverfahren häufig verzerrte, insbesondere optimistisch verzerrte, GE-Schätzer liefern. Allerdings scheint dieses Problem in der Praxis häufig vernachlässigt zu werden. Außerdem existieren nicht für alle Situationen Leitfäden, wie der GE in komplexen Datensituationen (weitestgehend) unverzerrt geschätzt werden kann. Für manche Datensituationen, wie etwa räumliche Daten, existieren dazu ausreichende Ergebnisse aus der Literatur, während es für andere Datensituationen noch Erkenntnislücken gibt. Vor diesem Hintergrund hatten wir im ersten Teil dieses Arbeitspaketes zwei Ziele. Das erste Ziel war, einen möglichst umfassenden Überblick über den aktuellen Erkenntnisstand aus der Literatur zu geben. Das zweite Ziel war, Erkenntnislücken mittels Simulationsstudien zu füllen. In beiden Fällen beschränken wir uns dabei auf fünf komplexe Datensituationen die in der amtlichen Statistik von besonderer Relevanz sind.

Der zweite Teil befasste sich mit der Konstruktion von Konfidenzintervallen (KI) für den GE, wobei wir uns im Gegensatz zum ersten Teil auf unabhängig und identisch verteilte (i.i.d.) Daten beschränkten. Da es in der amtlichen Statistik von großer Bedeutung ist, verlässliche Maßzahlen zu berichten, ist es wichtig, die Unsicherheit von GE-Schätzern zu quantifizieren, um zu vermeiden, falsche Rückschlüsse über die Performanz von ML-Modellen zu ziehen. Klassischerweise wird die Unsicherheit von Schätzern mit KIs quantifiziert. Es existiert in der Literatur eine Vielzahl von Schätzern für die Varianz von resamplingbasierten Fehlerschätzern und Methoden zur Konstruktion von KIs für den GE. Allerdings hat sich bisher keine dieser Methoden etabliert und es nicht klar, wie gut diese Methoden funktionieren. Des weiteren sind die theoretischen Eigenschaften dieser Methoden in vielen Fällen nicht ausreichend erforscht. Unserer Vermutung nach ist ein wichtiger Grund hierfür, dass die Komplexität des Problems noch nicht ausreichend verstanden wird. Daher verfolgten wir in diesem zweiten Teil zwei Ziele. Das erste Ziel war, das Problem möglichst allgemeinverständlich zu beschreiben. Das zweite Ziel war, die Performanz einiger dieser Methoden in einer neutralen Vergleichsstudie unter Verwendung verschiedener existierender Simulationsdesigns zu vergleichen mit dem Ziel, Hinweise darauf zu bekommen, welche dieser Methoden zu empfehlen sind und welche

vermieden werden sollten. In den nächsten beiden Abschnitten geben wir einen kurzen Überblick über die Ergebnisse der beiden Teilprojekte.

1.2 Schätzung des Generalisierungsfehler in komplexen Datensituationen

Die folgenden komplexen Datensituationen wurden betrachtet: geclusterte Daten, räumliche Daten, ungleiche Stichprobenwahrscheinlichkeiten, Konzeptdrift und hierarchisch strukturierte Outcomes. Im Folgenden geben wir einen Überblick über die wichtigsten Erkenntnisse aus Hornung u. a. (2023) über die GE-Schätzung in diesen Datensituationen.

Geclusterte Daten In der Gestaltung von Umfragen und Volkszählungen werden Daten häufig in Gruppen geclustert, zum Beispiel in Haushalte. Beobachtungen innerhalb dieser Gruppen sind typischerweise ähnlicher zueinander als zu Beobachtungen aus anderen Gruppen. Bei der GE-Schätzung muss die Clusterstruktur berücksichtigt werden, indem Beobachtungen clusterweise den Folds in der KV zugewiesen werden, was Clusterüberschneidungen zwischen Trainings- und Testdaten verhindert. Unsere Simulationsergebnisse zeigen, dass das Ignorieren der Clusterstruktur zu einem leichten, aber nicht zu vernachlässigenden Optimismus führen kann. Dieser Effekt kann sich verstärken, wenn Merkmale innerhalb von Clustern identische Werte annehmen, eine Situation, die in der amtlichen Statistik häufig auftritt.

Räumliche Daten Jede Art von Daten, die geografische Informationen enthält, kann als räumliche Daten betrachtet werden. Zum Beispiel kann Satellitenbildgebung verwendet werden, um die Landnutzung in verschiedenen Gebieten zu bestimmen. Hier sind Regionen, die einander nahe sind, in Landnutzung und anderen Merkmalen einander ähnlicher als entfernte Gebiete, was zu räumlichen Korrelationen zwischen einander nahen Beobachtungen führt. Bei der GE-Schätzung von räumlichen Vorhersagemodellen ist es oft entscheidend, eine geeignete räumliche Trennung zwischen Trainings- und Testdaten sicherzustellen. Dies wird durch verschiedene Verfahren erreicht, die gemeinsam als räumliche KV bezeichnet werden. Die Wahl und genaue Konfiguration dieser Verfahren hängt von der Datenstruktur und dem spezifischen Anwendungsszenario des ML-Modells ab. Wird das Modell innerhalb des Beobachtungsraums verwendet und sind die Trainingsdaten gleichmäßig im Beobachtungsraum verteilt, könnte eine standardmäßige KV ohne räumliche Trennung ausreichen.

Ungleiche Ziehungswahrscheinlichkeiten In der amtlichen Statistik und anderen Bereichen wie der Ökologie werden aus praktischen und prinzipiellen Gründen häufig Stichprobenverfahren verwendet, die von i.i.d. Sampling abweichen. Hier werden die Beobachtungen mit unterschiedlichen Wahrscheinlichkeiten aus der Population in die Stichprobe gezogen. Wie wir analytisch und durch eine Simulationstudie gezeigt haben, können solche ungleichen Ziehungswahrscheinlichkeiten, wenn sie nicht berücksichtigt werden, zu einer Verzerrung bei der GE-Schätzung führen. Wie wir ebenfalls zeigen, kann diese Verzerrung jedoch mittels eines auf dem Horvitz-Thompson-Theorem basierenden Schätzer vermieden werden, wenn die Stichprobenwahrscheinlichkeiten bekannt sind, selbst wenn das ML-Modell fehlspezifiziert ist.

Konzeptdrift Veränderungen in der Verteilung von Datenquellen im Laufe der Zeit, die als Konzeptdrift bekannt sind, können die Vorhersageleistung von ML-Modellen beeinträchtigen, wenn ihnen nicht angemessen begegnet wird. Das Erkennen von Konzeptdrift und die Umsetzung geeigneter Maßnahmen sind daher von entscheidender Bedeutung. Für diese Aufgaben stehen zahlreiche

Methoden zur Verfügung. Allerdings ist die vorhandene Literatur zur Schätzung von GE unter Konzeptdrift relativ spärlich. Prequential Validation, auch als Zeitreihenkreuzvalidierung bekannt, ist die vorherrschende Methode zur Schätzung des GE im Kontext von Konzeptdrift. Unsere Simulationsergebnisse zeigen jedoch, dass sowohl die Prequential Validation als auch die konventionelle KV überoptimistische GE-Schätzungen liefern können. Im Gegensatz dazu scheint die Out-of-Sample-Validation, bei der nur die jüngsten Beobachtungen für die GE-Schätzung und frühere Beobachtungen für das Training verwendet werden, weitgehend unverzerrte Schätzungen zu liefern.

Hierarchisch strukturierte Outcomes In einigen Klassifikationsproblemen gehören die Beobachtungen nicht zu einzelnen Klassen, sondern zu einer Hierarchie von Klassen, wobei jede Klasse innerhalb einer breiteren Kategorie angesiedelt ist. Beispielsweise könnte Klasse 1.3.2 eine Unterklasse innerhalb von Klasse 1.3 sein, die wiederum eine Unterklasse unter der breiteren Kategorie von Klasse 1 ist. In der amtlichen Statistik sind Klassifikationssysteme häufig hierarchisch. Ein Beispiel ist das hierarchische Berufsklassifikationsschema ISCO-08 der Internationalen Arbeitsorganisation (ILO). Die hierarchische Struktur solcher Schemata liefert wertvolle Informationen in statistischen Analysen, die angemessen berücksichtigt werden müssen.

Es existiert eine Vielzahl von Performanzmaßen für hierarchische Klassifikationsprobleme, jedes mit eigenen Vorteilen und Einschränkungen. Diese Performanzmaße sind darauf ausgelegt zu berücksichtigen, dass die Schwere einer Fehlklassifikation durch das Ausmaß bestimmt wird, in dem die vorhergesagte Klasse von der wahren Klasse abweicht.

Während es also bereits viel Literatur zu Performanzmaßen gibt, scheint bisher noch nicht untersucht worden zu sein, welche Resampling-Techniken am besten für hierarchische Klassifikationsprobleme geeignet sind. Wir haben in einer Simulationsstudie unsere Vermutung untersucht, dass stratifizierte KV im Vergleich zur Standard-KV für hierarchische Klassifikationsprobleme zu einer geringeren Verzerrung und Varianz führt. Während beide Methoden bei kleinen Datensätzen eine gewisse Verzerrung zeigten, zeigte die stratifizierte KV, anders als die Standard-KV, für größere Stichproben eine vernachlässigbare Verzerrung. Hierarchische Klassifikationsprobleme erfordern typischerweise relativ große Datensätze. Diese Eigenschaft macht die stratifizierte KV zu einer zuverlässigeren Methode zur Performanzschätzung in solchen Kontexten. Daher empfehlen wir die Verwendung von stratifizierter KV zur Performanzschätzung von hierarchischen Klassifikationsmodellen. Unsere Simulationsstudie zeigte keine bemerkenswerten Unterschiede in der Varianz zwischen stratifizierter und Standard-KV.

Im Rahmen der Simulationsstudie für hierarchisch strukturierte Outcomes wurde ein R-Paket, "hierclass", entwickelt. Dieses Paket beinhaltet den im Rahmen der Studie verwendeten Learner "Top-Down-Klassifikation mit Random Forests als lokale Multi-Class-Klassifikatoren". Unseres Wissens nach ist das derzeit der einzige in R öffentliche verfügbare Learner für hierarchische Klassifikation. Zusätzlich sind in "hierclass" verschiedene Performanzmaße für hierarchische Klassifikationsprobleme integriert. Sowohl der Learner als auch die Performanzmaße sind in das "mlr3"-Ökosystem eingebettet, einer Sammlung von R-Paketen für maschinelles Lernen in R. Das Paket ist auf GitHub öffentlich verfügbar: https://github.com/RomanHornung/hierclass.

Ebenfalls öffentlich verfügbar ist der Code für alle in diesem Teilprojekt durchgeführten Simulationen, zu finden unter: https://github.com/RomanHornung/PPerfEstComplex.

1.3 Konstruktion von Konfidenzintervallen für den Generalisierungsfehler

1.3.1 Konzeptionelle Grundlagen

Die Erstellung geeigneter KIs für den GE stellt durch die Überlappung der Trainings- und Testdatensätze in Resamplingverfahren eine Herausforderung dar, was sich auch daran zeigt, dass asymptotische Garantien für entsprechende Varianzschätzer in der Literatur selten sind. Um die Komplexität des Problems klar darzustellen und so insbesondere hoffentlich methodischen Wissenschaftlern bei der Entwicklung neuer Ansätze zu erleichtern stellten wir in Schulz-Kümpel u. a. (2024) die konzeptionellen Grundlagen des Problems dar.

Diese konzeptionellen Grundlagen zur Schätzung des GEs umfassen die Zieldefinition, die Komplexität des Resampling, die Quellen der Unsicherheit und die theoretische Validität. Der GE wird oft als "Risiko" oder "erwartetes Risiko" bezeichnet, wobei diese Begriffe in der Literatur nicht einheitlich verwendet werden.

Ziel der Inferenz: Der GE dient als Maß für die durchschnittliche Vorhersagegenauigkeit eines Modells, basierend auf Daten, die aus demselben Prozess wie die Trainingsdaten stammen. Dabei wird zwischen dem "Risiko" (Performance eines spezifischen Modells) und dem "erwarteten Risiko" (durchschnittliche Performance eines Schätzverfahrens) unterschieden.

Rolle des Resampling: Resampling-Verfahren sind essenziell, um (weitestgehend) unverzerrte Schätzungen des GE zu erhalten. Sie teilen die Daten wiederholt in Trainings- und Testsets, wodurch Abhängigkeiten zwischen den beobachteten Verlusten entstehen, die die Inferenz erschweren. Das Resampling-Verfahren, das am ehesten dem in der Praxis betrachteten Prädiktionszenario entspricht ist Holdout-Resampling, bei dem die Daten einmalig in Trainings- und Testdaten unterteilt werden. Aber selbst für dieses einfache Verfahren bringt die Konstruktion von KIs theoretische Herausforderungen mit sich.

Quellen der Unsicherheit: Die Unsicherheit bei der GE-Schätzung kann in "Validierungsunsicherheit" und "Trainingsunsicherheit" unterteilt werden. Erstere entsteht dadurch, dass die geschätzten Modelle immer nur auf einer endlichen Menge von Testdaten evaluiert werden, und letztere dadurch, dass das gelernte Modell von den zufällig gezogenen Trainingsdaten abhängt und das Schätzverfahren sogar selbst stochastisch sein kann.

Theoretische Validität: Asymptotische Exaktheit, also die Garantie, dass das KI den wahren GE mit zunehmender Stichprobengröße mit der zuvor festgelegten Überdeckungswahrscheinlichkeit einschließt, ist selten nachweisbar. Die meisten KIs basieren auf heuristischen Anpassungen und nur wenige KIs sind für spezifische Arten von GE-Schätzungen theoretisch validiert. Da in der Literatur zu KIs für den GE häufig zufällige Parameter (z.B. das Risiko) betrachtet wurden, schlagen wir eine Definition eines "Coverage Intervals" als Generalisierung von KIs vor, in der der interessierende Parameter zufällig sein kann.

Zusammenfassend ist die Konstruktion von KIs für den GE eine komplexe Aufgabe, die präzise Definitionen des Ziels, eine sorgfältige Auswahl des Resampling-Verfahrens und ein tiefes Verständnis der Unsicherheitsquellen erfordert. Die theoretische Fundierung dieser Methoden ist entscheidend, um ihre Zuverlässigkeit in der Praxis zu gewährleisten.

1.3.2 Zusammenfassung existierender Methoden

Die existierenden Methoden zur Konstruktion von KIs für den GE basieren auf verschiedenen Resampling-Verfahren. Diese Methoden umfassen unter anderem Holdout, wiederholtes Subsampling, KV, Leave-One-Out KV, wiederholte KV, genestete KV und verschiedene Bootstrapping-Ansätze. Jedes Verfahren hat spezifische Eigenschaften bezüglich der Datentrennung und der Wiederholung der Schätzprozesse, die zu unterschiedlichen Abhängigkeiten und Varianzen in den Verlustbeobachtungen führen.

Wir betrachteten folgende Inferenzmethoden:

- Standard-Single-Split-Schätzung: Diese Methode nutzt eine einmalige Aufteilung in Trainings- und Testdaten, um Punktschätzungen und KIs für den GE zu berechnen, wobei sie, wie von uns bewiesen, asymptotisch exakte Abdeckungen liefert.
- Replace-One KV: Austern und Zhou, 2020 berechnen asymptotisch exakte KIs durch wiederholtes berechnen einfacher KVs, bei denen einzelne Beobachtungen ausgetauscht werden. Wir konnten hier einen Fehler in der Herleitung finden und haben die Formel für das Konfidenzintervall korrigiert.
- Heuristic Repeated KV: Diese von uns konzipierte Abwandlung von Replace-One KV verwendet wiederholte statt einfache KV.
- Test Error: Bayle u. a., 2020 verwendet K-fache KV, um sowohl Punkt- als auch KI-Schätzungen für eine von ihnen definierte Größe, namens "Testfehler" zu liefern, wobei sie einen Nachweis der asymptotischen Exaktheit der KI-Schätzung in Bezug auf diese Größe erbringen.
- Corrected Resampled t-test: Diese Methode verwendet wiederholtes Subsampling um GE-Schätzer und Standardfehler-Schätzer zu erhalten. Hierbei wird eine heuristische Anpassung verwendet um für die Überlappung in den Trainings- und Testdaten zu korrigieren (Nadeau und Bengio, 2003).
- Konservativer z-Test: Die Varianz wird durch gepaartes Subsampling geschätzt (Nadeau und Bengio, 2003).
- 5×2-KV: Die 5×2-KV-Methode von Dietterich, 1998 konstruiert KIs mittels wiederholter 2-facher KV. Sie wurde ursprünglich für Modellvergleiche entwickelt, kann aber auch zur Einzelmodellbewertung eingesetzt werden.
- Nested KV: Bates, Hastie und Tibshirani, 2021 verwenden eine spezielle verschachtelte KV zur Schätzung des Standardfehlers.
- Out-of-Bag: Diese Methode nutzt Bootstrap-Resampling zur Schätzung des GEs und des Standardfehlers (Efron und Tibshirani, 1997).
- 0.632+ Bootstrap: Diese Methode kombiniert Bootstrap-Resampling mit In-Sample-Fehlerbewertungen zur Schätzung des GEs und des Standardfehlers (Efron und Tibshirani, 1997).
- Bootstrap Case CV Percentile: Die Bootstrap Case CV Percentile-Methode von Jiang, Varma und Simon, 2008 kombiniert Bootstrap-Resampling mit einer Variante von Leave-One-Out KV, um KIs zu konstruieren, optional mit einer Bias-Korrektur.

- Two-stage Bootstrap: Hierbei werden Bootstrap-Schätzer innerhalb eines äußeren Bootstraps verwendet und anhand empirischer Quantile die Konfidenzintervalle geschätzt (Noma u. a., 2021).
- Location-shifted Bootstrap: Dieses Verfahren schätzt die Breite des Konfidenzintervalls durch eine Kombination aus Insample-Schätzer und Bootstrapping, sowie einer Biaskorrektur (Noma u. a., 2021).

1.3.3 Empirische Vergleichsstudie

Die Motivation für diese Vergleichsstudie lag in der Vielzahl von Resampling-Techniken und möglichen Ansätzen zur Varianzschätzung begründet, die fortwährend zu neuen Vorschlägen für Methoden zur Ableitung von KIs für den GE führt. Da derzeit wenige theoretische Ergebnisse existieren, die eine allgemeine Analyse des asymptotischen Verhaltens von KIs über Resampling-Einstellungen hinweg ermöglichen, ist eine gründliche empirische Untersuchung dieser Methoden von größter Bedeutung.

Unsere Studie (Schulz-Kümpel u. a., 2024) liefert mehrere wichtige Beiträge zum Verständnis der Komplexität der resampling-basierten Inferenz über den GE und etabliert eine Grundlage für die Bewertung zukünftiger Methoden, indem sie die Vergleichsmetriken transparent darstellt und alle Daten sowie Code öffentlich zugänglich macht. Dadurch soll die Forschungsgemeinschaft ermutigt werden, das Problem weiter zu erkunden und ein tieferes Verständnis zu entwickeln. Der Code für die Vergleichsstudie ist auf GitHub öffentlich zugänglich: `https://github.com/slds-lmu/paper_2023_ci_for_ge`.

Für den Vergleich der oben beschriebenen Inferenzverfahren wurden 18 verschiedene Datengenerierungsverfahren, vier ML-Modelle und verschiedene Verlustfunktionen verwendet. Die Verfahren wurden hinsichtlich der Überdeckungsfrequenz und Breite der generierten Konfidenzintervalle, sowie des notwendigen Rechenaufwands bewertet. Die Überdeckungsfrequenz wurde bezüglich des Risikos, erwarteten Risikos und - falls vorhanden - der entsprechenden Proxygröße berechnet,.

Es wurden zahlreiche Datensätze generiert, um die Abdeckungsfrequenz der KIs zu schätzen. Dabei wurde das (erwartet) Risiko auf großen Validierungsdatensätzen approximiert. In bestimmten Fällen wurde auch Proxygrößen berechnet, wenn die KIs asymptotisch exakt in Bezug auf diese Größen waren.

In der empirischen Vergleichsstudie konnten wir fünf Methoden identifizieren, die eine gute durchschnittliche Überdeckungswahrscheinlichkeit aufweisen: Nested KV, Conservative z-Test, Standard-Single-Split, Test Error und Corrected Resampled t-Test. Diese unterscheiden sich teilweise stark in ihrer Kostenintensität, wobei insbesondere Nested KV in der von den Autoren empfohlenen Einstellung (ca. 5000 Resampling-Iterationen) deutlich teurer ist als der Rest. Das Conservative z-Test-Verfahren ist mit 315 Iterationen zwar schon deutlich günstiger und weist eine ähnliche durchschnittliche Abdeckungswahrscheinlichkeit auf, führt aber zu breiteren Konfidenzintervallen. Beide Verfahren sind tendenziell eher konservativ. Die anderen drei Inferenzmethoden sind deutlich günstiger, dafür aber auch liberaler. Der Corrected Resampled t-Test weist tendenziell eine bessere Überdeckungswahrscheinlichkeit als das Test Error-Verfahren auf. Der Standard-Single-Split Schätzer hat zwar eine gute Überdeckungswahrscheinlichkeit, führt allerdings - aufgrund der geringeren Anzahl an Testbeobachtungen - zu breiteren Intervallen.

Die Analyse des Einflusses der Parameter der Inferenzverfahren auf deren Güte ist derzeit noch nicht abgeschlossen, wird aber Aufschluss darüber geben, wie beispielsweise die Anzahl der Wiederholungen eines Inferenzverfahrens die Überdeckungswahrscheinlichkeit und Breite beeinflusst.

Neben dem Vergleich der Inferenzmethoden lieferte das Experiment auch Aufschluss über methodenunabhängige Herausforderungen bei der Konstruktion von Konfidenzintervallen. Beispielsweise haben wir beobachtet, dass das lineare Modell auf bestimmten datengenerierenden Prozessen sehr instabil war und dadurch zu sehr schlechten Überdeckungswahrscheinlichkeiten für alle Verfahren führte. Auch können starke Ausreißer - vor allem für Regressionsprobleme - zu Schwierigkeiten und in Einzelfällen extrem breiten Konfidenzintervallen führen, selbst wenn das Inferenzverfahren auf dem datengenerierenden Prozess an sich eine gute Überdeckungswahrscheinlichkeit aufweist. Die Wahl von robusteren Verlustfunktionen kann hierbei zumindest zu einem gewissen Grad abhilfe schaffen.

1.4 Fazit und Ausblick

In diesen beiden Teilprojekten wurden zum einen GE-Schätzer für komplexe Datensituationen und zum anderen Inferenzverfahren für den GE im i.i.d-Fall behandelt.

Es liegt nahe, in einem Folgeprojekt Inferenzverfahren für den GE in komplexen Datensituationen zu untersuchen. Dies wäre von besonderer praktischer Bedeutung, da die Daten in Anwendungen der amtlichen Statistik häufig komplexe Strukturen aufweisen und unklar ist, inwieweit die für den i.i.d.-Fall entwickelten Inferenzverfahren für den GE auf solche Datensituationen übertragbar sind. Die Erweiterung auf weitere Stichprobendesigns (siehe Arbeitspaket 1.4) wäre ebenfalls sinnvoll.

Weitere interessante Forschungsaktivitäten könnten darin bestehen, Inferenzverfahren für komplexere Lernverfahren inklusive Hyperparameter-Tuning zu vergleichen oder den Einfluss des Resampling-Verfahrens auf das Hyperparameter-Tuning näher zu untersuchen.

Ein nächster Schritt wird es außerdem sein, die gut funktionierenden Inferenzmethoden in dem mlr3-Ökosystem verfügbar zu machen, sodass unsere wissenschaftlichen Erkenntnisse leichter den Weg in die Praxis finden.

2 Interpretierbares Maschinelles Lernen

2.1 Einführung

Maschinelles Lernen (ML) hat ein enormes Potenzial, Entscheidungsprozesse aufgrund seiner Vorhersageleistung zu unterstützen. Leider sind diese ML-Modelle oft Black Boxes und zu komplex, um von Menschen verstanden zu werden, beispielsweise ein Random Forest, der aus mehreren Entscheidungsbäumen besteht, oder ein tiefes neuronales Netzwerk mit mehreren Schichten von Neuronen. Das Fehlen von Erklärungen kann ein Hindernis für die Anwendung von ML-Modellen sein, insbesondere in kritischen Bereichen wie der amtlichen Statistik, in denen Vorhersagen nachteilige Auswirkungen auf bestimmte Gruppen haben können. In den letzten zehn Jahren hat sich ein ganzes Forschungsfeld um die Interpretation von ML-Modellen entwickelt, bekannt als interpretierbares maschinelles Lernen (IML) oder erklärbare künstliche Intelligenz (XAI). Das Feld kann grob in zwei Forschungsbereiche unterteilt werden: Der erste befasst sich mit der Entwicklung von (leistungsstarken) interpretierbaren Modellen, z.B. durch die Destillation eines neuronalen Netzwerks in einen einfachen Entscheidungsbaum (Frosst und Hinton, 2017); der zweite befasst sich mit der Entwicklung von Methoden zur nachträglichen Interpretation komplexer Modelle, also nach der Modellanpassung. In diesem Abschnitt behandeln wir das letztgenannte Feld der nachträglichen Interpretationsmethoden. Wir fassen deren Zwecke zusammen (Abschnitt 2.2) und geben einen Überblick über Methoden (Abschnitt 2.3). Wir heben auch einige der neuesten Fortschritte innerhalb dieses Projekts hervor, die insbesondere in der amtlichen Statistik angewendet werden können (Abschnitte 2.3.1-2.4.1), und diskutieren offene Forschungsfragen (Abschnitt 2.5).

2.2 Ziele der Interpretation

Die Interpretierbarkeit von ML-Modellen ist aus mehreren Perspektiven wichtig. Erstens können Interpretationsmethoden helfen, *globale* Einblicke in ein Modell zu gewinnen. Benutzer:innen können etwas darüber lernen, welche Merkmale eine Vorhersage am meisten beeinflussen und wie diese Merkmale die Vorhersage im Durchschnitt beeinflussen. Interpretationsmethoden können auch helfen, *individuelle* Entscheidungen zu verstehen und zu kontrollieren. Das Recht auf eine Erklärung für individuelle algorithmische Entscheidungen ist auch in der Datenschutz-Grundverordnung (Erwägungsgrund 71, DSGVO)[1] verankert.

> In jedem Fall sollte eine solche Verarbeitung mit angemessenen Garantien verbunden sein, einschließlich der spezifischen Unterrichtung der betroffenen Person und des Anspruchs auf direktes Eingreifen einer Person, auf Darlegung des eigenen Standpunkts, auf Erläuterung der nach einer entsprechenden Bewertung getroffenen Entscheidung sowie des Rechts auf Anfechtung der Entscheidung.

Diese Einblicke können die Grundlage für die Identifizierung von Fehlern im Modell oder in den Daten sein. Daher können sie auch bei der Modellprüfung helfen, für die die gewonnenen Einblicke mit dem Fachwissen verglichen werden müssen. Die Ungenauigkeiten können dann für ein nachfolgendes Modell korrigiert werden. Dies ist besonders relevant, wenn das Modell diskriminierendes Verhalten aufdeckt – beispielsweise wenn Menschen aufgrund ihres Geschlechts oder ihrer ethnischen Zugehörigkeit ein höheres Risiko vorhergesagt wird, einen Kredit nicht zurückzuzahlen. Dies führt zu ethischen Fragen, siehe z.B. Bothmann, Peters und Bischl (2024) für einen Ansatz zur Übersetzung philosophischer

[1] Siehe https://dsgvo-gesetz.de/erwaegungsgruende/nr-71/

Fragen in ein umsetzbares Rahmenwerk für fairnessbewusstes maschinelles Lernen und Caton und Haas (2024) für eine Übersicht über die aktuelle Forschung zur algorithmischen Fairness, siehe auch Kapitel 6. Der folgende Abschnitt gibt einen Überblick über nachträgliche Interpretationsmethoden, die auf ein einzelnes oder eine Teilmenge der Interpretationsziele abzielen.

2.3 Überblick über nachträgliche Interpretationsmethoden

Nachträgliche Interpretationsmethoden können in Klassen unterteilt werden, basierend auf bestimmten Eigenschaften dieser Methoden. Die folgende Differenzierung basiert auf Molnar (2022) und Molnar u. a. (2022).

Erstens können wir zwischen modellspezifischen und modellagnostischen Methoden unterscheiden. Modellspezifische Methoden sind nur für bestimmte ML-Modelle anwendbar, z.B. die Gini-Importance-Methode von Breiman (2001a) für Random Forests oder Saliency Maps für neuronale Bildklassifizierer von Simonyan, Vedaldi und Zisserman (2013). Diese Methoden nutzen die Modellstruktur, d.h. die Baumstruktur von Random Forests oder den Zugriff auf die Gradienten des neuronalen Netzwerks. Daher können sie nicht auf andere Modelltypen angewendet werden. Im Gegensatz dazu nutzen modellagnostische Methoden die Modellstruktur nicht aus und können auf jeden Modelltyp angewendet werden.[2] Modellagnostische Methoden sind besonders nützlich, wenn verschiedene Modelle miteinander verglichen werden sollen, um zu bewerten, ob das Modell dem eigenen Fachwissen entspricht.[3] Für den Rest dieses Kapitels konzentrieren wir uns nur auf modellagnostische Methoden.

Eine zweite Differenzierung von Methoden basiert darauf, ob die Methoden tatsächlich darauf abzielen, das Modellverhalten im Allgemeinen zu erklären – diese Methoden werden als global bezeichnet – oder ob die Methoden darauf abzielen, das lokale Verhalten des Modells für einen einzelnen Datenpunkt von Interesse und dessen nahe Umgebung zu erklären – diese Methoden werden als lokal bezeichnet.

Drittens können wir Methoden in Feature Effect- und Feature Importance-Methoden unterteilen. Das Ziel von Feature Effect-Methoden ist es, die Richtung und Größe einer Änderung in den Ergebnissen aufgrund von Änderungen der Features zu bewerten. Feature Importance-Methoden bewerten den Beitrag eines Features zur Modellleistung (z.B. durch eine Verlustfunktion) oder zur Varianz der Vorhersagefunktion.

Die folgenden zwei Unterabschnitte konzentrieren sich auf zwei Unterklassen von Feature Importance- und Feature Effect-Methoden: Verlustbasierte Feature Importance-Methoden (Abschnitt 2.3.1) und Kontrafaktische Erklärungen (Abschnitt 2.3.2).

2.3.1 Spotlight: Verlustbasierte Feature Importance

Dieser Abschnitt bietet einen kurzen Überblick über globale, verlustbasierte, modellagnostische Feature Importance-Methoden. Er basiert stark auf der in diesem Projekt erstellten Arbeit von Ewald u. a. (2024), einem umfassenden Leitfaden, der darauf abzielt, Forscher:innen die Werkzeuge

[2] Dies bedeutet nicht unbedingt, dass Methoden auf binäre, mehrklassige oder Regressionsmodelle angewendet werden können. Zum Beispiel sind kontrafaktische Erklärungsmethoden oft auf binäre Klassifikatoren zugeschnitten, jedoch spielt es keine Rolle, ob das Modell ein Random Forest, ein lineares Modell oder ein neuronales Netzwerk ist.

[3] Zum Beispiel, wenn eine Expertin in einem bestimmten Bereich sicher ist, dass ein Merkmal X einen großen Einfluss auf ein Ergebnis Y hat, würde es sie beunruhigen, wenn das Merkmal X nicht das wichtigste Merkmal für die gegebenen Vorhersagemodelle ist.

an die Hand zu geben, um fundierte Entscheidungen bei der Auswahl der geeignetsten Methode für ihre spezifischen Analysebedürfnisse zu treffen.

Feature Importance-Methoden dienen als Brücke zwischen den komplexen Vorhersagen, die von ML-Modellen generiert werden, und der Notwendigkeit interpretierbarer Einblicke in den zugrunde liegenden DGP. Indem sie den Beitrag jedes Merkmals zum Vorhersageprozess quantifizieren, bieten diese Methoden wertvolle Einblicke in die relative Bedeutung verschiedener Merkmale bei der Bestimmung der Modellergebnisse. Dieses Verständnis ist nicht nur für die Modellinterpretation, sondern auch für die Hypothesengenerierung, die Merkmalselektion und domänenspezifische Einblicke unerlässlich.

Ewald u. a. (2024) unterscheiden zwischen drei Klassen von Feature Importance-Methoden: univariate Perturbationen, Marginalisierung und Modellanpassung. Im Folgenden werden drei der bekanntesten verlustbasierten Feature Importance-Methoden kurz vorgestellt und ihre Unterschiede diskutiert. Für weitere Methoden und detaillierte Einblicke verweisen wir auf Ewald u. a. (2024).

Permutation Feature Importance (PFI) (Breiman, 2001a; Fisher, Rudin und Dominici, 2019) ist eine univariate Perturbations-FI-Methode. Um die PFI für ein Merkmal von Interesse (feature of interest, FOI) X_j zu berechnen, werden die entsprechenden Beobachtungen permutiert, sodass die Abhängigkeit zwischen dem FOI und dem Ziel sowie zwischen dem FOI und allen anderen Merkmalen aufgehoben wird. Die Diskrepanz zwischen dem erwarteten Verlust des Modells unter Verwendung des permutierten Merkmals und dem Modell mit dem ursprünglichen Merkmal ergibt die PFI für das Merkmal X_j. Ewald u. a. (2024) geben an, dass PFI verwendet werden kann, um bedingungslose Feature Importance zu bewerten. PFI erfordert jedoch Annahmen über die Unabhängigkeit der Merkmale, die in realen Datensätzen typischerweise unrealistisch sind. Dies kann ihre praktische Anwendbarkeit trotz ihrer theoretischen Fundierung einschränken.

Die nächsten beiden Methoden sind besonders geeignet, wenn bedingte Feature Importances von Interesse sind, d.h. die Bedeutung eines Merkmals bedingt durch die anderen verbleibenden Merkmale.

Conditional Feature Importance (CFI) (Strobl u. a., 2008) ist ebenfalls eine Methode, die auf univariater Permutation basiert. Sie ist der PFI ähnlich, der einzige Unterschied besteht darin, dass die Permutation der Beobachtungen des Merkmals von Interesse so durchgeführt wird, dass die Abhängigkeiten zwischen den anderen Merkmalen erhalten bleiben, während die Beziehung zwischen dem Ziel und X_j aufgehoben wird. CFI erfordert genaue Modelle der univariaten bedingten Verteilung, die möglicherweise nicht immer verfügbar oder komplex zu ermitteln sind.

Im Gegensatz zu PFI und CFI ist **Leave-One-Covariate-Out (LOCO)** (Lei u. a., 2018) eine Methode zur Modellanpassung der FI. Wie der Name schon sagt, bestimmt LOCO die Bedeutung eines Merkmals, indem es aus den Daten entfernt und das Modell ohne es neu angepasst wird. Die Diskrepanz im Risiko des Modells ohne das Merkmal und dem vollständigen Modell quantifiziert das LOCO-Importance-Maß. Aufgrund mehrerer Modellanpassungen ist die Methode rechnerisch aufwändig.

Insgesamt erfordert die Auswahl der geeignetsten Feature Importance-Methode eine sorgfältige Abwägung mehrerer Faktoren. Forscher:innen müssen die Natur der Daten, die Komplexität des Modells und die spezifischen Fragen, die sie beantworten möchten, berücksichtigen. Während jede FI-Technik ihre Stärken hat, hängt ihre Praktikabilität von der spezifischen Anwendung und den verfügbaren Rechenressourcen ab. Praktiker sollten Methoden wählen, die basierend auf ihren spezifischen Bedürfnissen eine Balance zwischen Genauigkeit, Annahmen und Rechenanforderungen bieten.

Um die praktische Anwendung dieser Methoden zu veranschaulichen, verwenden Ewald u. a.

(2024) den bekannten „Bike Sharing"-Datensatz. Der Datensatz umfasst 731 Beobachtungen und 12 Merkmale im Zusammenhang mit Wetter, Temperatur, Windgeschwindigkeit, Jahreszeit und Wochentag. Durch die Anwendung von sowohl PFI als auch LOCO auf diesen Datensatz zeigen sie, wie unterschiedliche Methoden zu unterschiedlichen Ergebnissen führen können und diskutieren die Implikationen dieser Unterschiede für das Verständnis des datengenerierenden Prozesses.

Die Schätzung der Unsicherheit ist ein kritischer Aspekt der Interpretation von Feature Importance-Maßen. Ewald u. a. (2024) diskutieren verschiedene Techniken zur Schätzung der Unsicherheit dieser Maße, wie Resampling-Methoden und statistische Inferenztechniken. Sie betonen die Bedeutung der Verwendung unabhängiger Testdaten, um verzerrte Schätzungen zu vermeiden, und heben bewährte Verfahren zur Sicherstellung zuverlässiger Ergebnisse hervor.

Abschließend identifiziert das Papier mehrere offene Herausforderungen und Bereiche für zukünftige Forschung, darunter: (i) Entwicklung von Methoden, die die Feature Importance bei komplexen Interaktionen zwischen Merkmalen genau schätzen können, (ii) Erstellung von Benchmarks und empirischen Studien zum Vergleich der Leistung verschiedener Feature Importance-Methoden in verschiedenen Szenarien, (iii) Untersuchung der kausalen Zusammenhänge zwischen Merkmalen und Zielvariable, um über bloße Assoziationen hinaus die zugrunde liegenden Mechanismen zu verstehen.

Durch die Bewältigung dieser Herausforderungen kann die zukünftige Forschung die Zuverlässigkeit und Anwendbarkeit von Feature Importance-Methoden verbessern und sie sowohl für wissenschaftliche Inferenz als auch für praktische Anwendungen nützlicher machen.

2.3.2 Spotlight: Kontrafaktische Erklärungen

Kontrafaktische Erklärungen (CFEs) und Semi-Faktische Erklärungen (SFEs) sind lokale Interpretationsmethoden, die darauf abzielen, das Verhalten eines Modells für individuelle Beobachtungen zu erklären (Doshi-Velez und Kim, 2017). CFEs heben minimale Änderungen der Merkmale hervor, die erforderlich sind, um eine Vorhersage zu ändern, während SFEs die maximalen Änderungen zeigen, die erforderlich sind, um eine Vorhersage gleich zu halten. Zum Beispiel könnte eine CFE in einem Kreditantrags-Szenario vorschlagen, dass die Beantragung eines niedrigeren Kreditbetrags zur Genehmigung geführt hätte. Umgekehrt könnte eine SFE darauf hinweisen, dass der Antrag selbst bei einem geringfügig niedrigeren Betrag abgelehnt worden wäre. Zu verstehen, warum ein bestimmtes Ereignis eingetreten ist, beinhaltet die Identifizierung seiner Ursachen, ein Konzept, das im kontrafaktischen Denken verwurzelt ist. Wie Hume (1748) und Lewis (1973) vorgeschlagen haben, geht es dabei darum, zu überlegen, was passiert wäre, wenn die Umstände anders gewesen wären.

Dieser Abschnitt beschreibt eine spezifische Methode der SFEs, nämlich „Interpretierbare Regionale Deskriptoren". Diese Methode wurde von uns in Dandl u. a. (2023) vorgeschlagen, und wir verweisen für Details der Methode auf die vollständige Publikation.

Interpretierbare Regionale Deskriptoren (IRDs) stellen eine neuartige Technik zur Generierung lokaler, modellagnostischer Interpretationen dar. Diese Methode verwendet Hyperboxen, um zu umreißen, wie sich Merkmalswerte einer Beobachtung ändern können, ohne deren Vorhersage zu beeinflussen. Dieser Ansatz erleichtert das Verständnis der Robustheit von Vorhersagen und bietet SFEs, die den Bereich der Merkmalswerte angeben, die die Vorhersage konstant halten. Solche Einblicke sind sowohl für Modellentwickler:innen als auch für Personen, die von den Entscheidungen der ML-Modelle betroffen sind, von Wert.

Konzept und Motivation IRDs befassen sich mit der Frage, wie viel ein Merkmalswert variiert werden kann, während die gleiche Vorhersage beibehalten wird. Sie erzeugen eine Reihe von „Selbst-

wenn"-Erklärungen, die helfen, eine Entscheidung zu rechtfertigen, indem sie zeigen, dass auch bei unterschiedlichen Merkmalswerten das Ergebnis unverändert bleiben würde. Zum Beispiel könnte ein IRD in einem Kreditrisikobewertungsszenario darauf hinweisen, dass selbst wenn die Ersparnisse des Antragstellers moderat statt gering wären, das vorhergesagte Risikoniveau moderat bleiben würde und damit eine Begründung für die Entscheidung liefern.

Methodik Der Prozess der Generierung von IRDs umfasst die Formulierung eines Optimierungsproblems, bei dem das Ziel darin besteht, die größte Hyperbox um einen Interessenspunkt x' zu finden. Diese Hyperbox muss Merkmalswerte enthalten, die Vorhersagen innerhalb einer benutzerdefinierten Nähe-Region, Y', ergeben. Die Optimierung zielt darauf ab, die Abdeckung der Hyperbox zu maximieren und gleichzeitig eine hohe Präzision zu gewährleisten, was bedeutet, dass alle Punkte innerhalb der Hyperbox Vorhersagen innerhalb des angegebenen Bereichs ergeben sollten.

Erstellung von IRDs Der Prozess der Erstellung eines IRD umfasst mehrere Schritte:

1. **Einschränkung des Suchraums**: Der initiale Suchraum wird auf die größte lokale Hyperbox um den Interessenspunkt eingeschränkt. Für numerische Merkmale bedeutet dies, die Merkmalswerte auf einem Gitter zu variieren, bis die Vorhersage außerhalb des gewünschten Bereichs fällt. Für kategoriale Merkmale werden alle Kategorien eingeschlossen, die weiterhin Vorhersagen innerhalb des gewünschten Bereichs ergeben.
2. **Auswahl des Datensatzes**: Der Datensatz, der zur Erstellung des IRD verwendet wird, kann entweder die Trainingsdaten oder Daten sein, die gleichmäßig den interessierenden Merkmalsraum abdecken. Die Auswahl beeinflusst die empirischen Maße der Abdeckung und Präzision.
3. **Initialisierung**: Abhängig vom Ansatz kann die initiale Hyperbox die größte lokale Box sein, die alle Datenpunkte abdeckt, oder die kleinstmögliche Box, die den Interessenspunkt enthält.
4. **Optimierung**: Die Grenzen der Hyperbox werden iterativ angepasst, um die Abdeckung zu maximieren und gleichzeitig die Präzision zu erhalten. Top-down-Methoden verkleinern die größte Box, während Bottom-up-Methoden die kleinste Box vergrößern, wobei immer gewährleistet sein muss, dass der Interessenspunkt innerhalb der Hyperbox bleibt.
5. **Nachbearbeitung**: Zur Verfeinerung der Hyperbox-Grenzen werden zusätzliche Datenpunkte gesammelt und die Box angepasst, um Präzision und Abdeckung zu verbessern. Dieser Schritt hilft, Bereiche in der Hyperbox zu adressieren, die möglicherweise suboptimale Abdeckung oder Präzision aufweisen.

Durch das Befolgen dieser Schritte bieten IRDs eine umfassende und interpretierbare Methode, um die Stabilität von Vorhersagen komplexer ML-Modelle zu verstehen. Sie sind besonders nützlich, um Entscheidungen zu rechtfertigen und Merkmale zu identifizieren, die lokal keinen Einfluss auf die Vorhersage haben.

2.4 Modelbewertung basierend auf Interpretationsmethoden

2.4.1 Spotlight: Open Source Software in R

Das in diesem Projekt entwickelte R-Paket `mlr3summary` von Dandl u. a. (2024) ist darauf ausgelegt, prägnante und interpretierbare Zusammenfassungen für maschinelle Lernmodelle zu erstellen.

Inspiriert von der `summary`-Funktion für generalisierte lineare Modelle (GLMs) in R, erweitert `mlr3summary` seine Funktionalität, um modellagnostisch zu sein und somit einheitliche Zusammenfassungen sowohl für parametrische als auch für nicht-parametrische maschinelle Lernmodelle bereitzustellen (Dandl u. a., 2024). Dieses Paket ermöglicht Informationen zu Datensatzmerkmalen, Modellleistung, Komplexität, geschätzter Merkmalswichtigkeit, Merkmalswirkungen und Fairnessmetriken, die alle mithilfe von Resampling-Strategien für unverzerrte Leistungsschätzungen bewertet werden.

Da maschinelles Lernen in verschiedenen Bereichen immer mehr in Entscheidungsprozesse integriert wird, ist die Notwendigkeit interpretierbarer Modelle von größter Bedeutung. Traditionelle Zusammenfassungsfunktionen in R, wie die für GLMs, bieten Einblicke in Modellparameter und -anpassung, sind jedoch auf spezifische Modelltypen beschränkt und verallgemeinern nicht auf komplexere maschinelle Lernmodelle. Das `mlr3summary`-Paket schließt diese Lücke, indem es eine standardisierte Diagnoseausgabe für eine Vielzahl von Modellen bietet und so den Vergleich und die Auswahl von Modellen erleichtert.

`mlr3summary` umfasst mehrere Schlüsseleigenschaften zur Verbesserung der Interpretierbarkeit und Bewertung von maschinellen Lernmodellen:

- **Modellagnostische Zusammenfassungen**: Bietet ein konsistentes Zusammenfassungsformat für verschiedene Modelltypen, einschließlich sowohl linearer als auch komplexer Modelle wie Random Forests und Gradient Boosted Trees.
- **Resampling-basierte Bewertung**: Verwendet Techniken wie Cross-Validation, um unverzerrte Schätzungen der Modellleistung und Merkmalswichtigkeit bereitzustellen.
- **Detaillierte Metriken**: Beinhaltet Leistungsmetriken (z.B. AUC, F1-Score), Komplexitätsmaße des Modells (z.B. Sparsity, Interaktionsstärke), Merkmalswichtigkeit (z.B. Partial Dependence Plots, Permutation Feature Importance) und Fairnessbewertungen.
- **Anpassbare Ausgabe**: Benutzer:innen können die Zusammenfassungsausgabe mit der `summary_control`-Funktion anpassen, um sie an spezifische Bedürfnisse und Präferenzen anzupassen.

Die Kernfunktion von `mlr3summary` ist die S3-basierte Zusammenfassungsfunktion für `mlr3`-Learner-Objekte. Der typische Workflow umfasst das Initialisieren einer Aufgabe, das Auswählen eines Lerners, das Trainieren des Modells und die Anwendung einer Resampling-Strategie zur Bewertung des Modells. Die Zusammenfassungsfunktion bietet dann einen umfassenden Überblick über das Modell und seine Leistung, einschließlich Abschnitten zu allgemeinen Modellinformationen, Residuen, Leistungsmetriken, Modellkomplexität, Merkmalswichtigkeit und Wirkungsdiagrammen. Jeder Abschnitt wird aus den Resampling-Ergebnissen abgeleitet, um eine unverzerrte Bewertung zu gewährleisten.

Die Zusammenfassungsausgabe kann mithilfe der `summary_control`-Funktion an spezifische Bedürfnisse angepasst werden. Benutzer:innen können angeben, welche Metriken enthalten sein sollen, die Anzahl der anzuzeigenden wichtigen Merkmale und welche Abschnitte ausgeblendet werden sollen. Zusätzlich können Fairnessmetriken durch Angabe eines geschützten Attributs integriert werden.

Das Paket enthält Funktionen zur effizienten Handhabung großer Datensätze und komplexer Modelle. Eine Simulationsstudie zeigte, dass die Laufzeit zwar mit der Anzahl der Merkmale und

Beobachtungen zunimmt, die Parallelisierung über Resampling-Iterationen jedoch die Leistung erheblich verbessert.

`mlr3summary` bietet ein robustes Werkzeug zur Zusammenfassung von maschinellen Lernmodellen auf konsistente und interpretierbare Weise. Zukünftige Entwicklungen könnten die erweiterte Unterstützung zusätzlicher Interpretationsmethoden, verbesserte Visualisierungsmöglichkeiten und die Integration mit anderen Modellauswertungstools umfassen.

Weitere Informationen und Zugriff auf das Paket finden Sie im GitHub-Repository unter https://github.com/mlr-org/mlr3summary und auf der CRAN-Seite unter https://cran.r-project.org/package=mlr3summary.

2.5 Diskussion

Die Arbeit von Ewald u. a. (2024), Dandl u. a. (2023) und Dandl u. a. (2024), die oben vorgestellt wurde, hebt mehrere Fortschritte im Bereich des interpretierbaren maschinellen Lernens hervor. Jeder Beitrag verbessert die Möglichkeiten, maschinelle Lerntechniken in der amtlichen Statistik einzusetzen: Während Ewald u. a. (2024) beschreibt, wie man zwischen verschiedenen Methoden zur Merkmalswichtigkeit wählt, zeigen Dandl u. a. (2023), wie man interpretierbare semi-faktische Erklärungen erstellt – indem man „Selbst-wenn"-Fragen beantwortet –, und Dandl u. a. (2024) bieten ein benutzerfreundliches Werkzeug für Praktiker, um Black-Box-Modelle des maschinellen Lernens zusammenzufassen und dadurch die Möglichkeiten zur Modelldiagnose und Modellauswahl zu verbessern. Zusätzlich zu fortgeschrittenen Interpretationsmethoden sind jedoch ethische Fragen in der amtlichen Statistik von Bedeutung. Wie von Bothmann, Dandl und Schomaker (2023) beschrieben und auf die Frage der „fairen"Mietpreise von Bothmann und Peters (2024) angewendet, sollten Methoden zur Berücksichtigung historischer Verzerrungen in realen Daten unter Verwendung von Kausalitätsinferenzen in Zukunft mehr in den Fokus rücken – insbesondere in der amtlichen Statistik.

3 Maschinelles Lernen bei komplexen Stichprobendesigns

3.1 Ausgangssituation, Fragestellungen und Überblick

Ein fundamentales Hindernis bei der leistungsfähigen, direkten Anwendung von Verfahren des Maschinellen Lernens in der Primär- wie Sekundäranalyse von Datensätzen der amtlichen Statistik könnte in der Komplexität des Erhebungsdesign liegen. Während die gängigen Verfahren zum maschinellen Lernen praktisch ausnahmslos und selbstverständlich von unabhängig und identisch verteilten Daten ausgehen, sind die wenigsten amtlichen Erhebungen einfache Zufallsauswahlen im stichprobentheoretischen Sinne. Vielmehr ist das Stichprobendesign typischerweise komplex; nicht alle Einheiten besitzen dieselbe Wahrscheinlichkeit, in die Stichprobe zu gelangen. Gängige Beispiele sind Clusterstichproben, geschichtete Stichproben oder größenproportionale Ziehungen, wobei in der Praxis oft auch Kombinationen dieser Designs verwendet werden.

Im Kontext von klassischen Regressionsanalysen ist das Problem des Umgangs mit komplexen Stichproben intensiv untersucht. Survey-Statistik und Methodik bieten klare Handlungsempfehlungen, wann und wie Stichprobegewichte in der Analyse explizit zu berücksichtigen sind (siehe etwa die Überblicke in (Lumley und Scott, 2017; Pfeffermann, 2009)). Gänzlich anders ist die Situation im Kontext von maschinellen Lernverfahren. Hier gibt es bisher nur verstreute Einzelergebnisse, die noch kein Gesamtbild gestatten(Toth und Eltinge, 2011; Nahorniak u. a., 2015; MacNell u. a., 2023, siehe auch Zadrozny, 2004 im Zusammenhang mit dem Selektionsbias).

Aufgabe des Arbeitspaketes war es deshalb, systematisch das Verhalten von Regressionsbäumen unter komplexen Stichprobendesigns zu untersuchen. Sogenannte CARTs (Klassifikations- (für kategoriale Zielvariablen) und Regressionsbäume (für metrische Zielvariable)) sind im Sinne der Breimanschen Unterscheidung der Modellierungskulturen (Breiman, 2001b) prototypische Beispiele algorithmischer Lernverfahren. Sie besitzen – wohl auch wegen ihrer besonders anschaulichen Interpretierbarkeit – große Verbreitung; hinzu kommt, dass Klasssifikations- und Regressionsbäume als zentraler Baustein weiterführender Verfahren dienen, wie insbesondere die als besonders prädiktionsstark geltenden Random Forests.

Im Zentrum standen dabei die folgenden Fragestellungen, die auch im Folgenden die Gliederung dieses Kapitels bestimmen werden:

- Muss davon ausgegangen werden, dass eine Nichtberücksichtigung des Stichprobendesigns zu deutlichen Verzerrungen bei der Bestimmung von Regressionsbäumen führen kann?

Da relativ schnell festgestellt werden konnte, dass die Antwort zu dieser Frage affirmativ ist, ergaben sich natürlich darauf aufbauende Fragenstellungen zur Systematisierung der Effekte und zur Ausdehnung der Ergebnisse:

- Lassen sich Situationen charakterisieren, in denen mit besonders starken Effekten bei der Nichtberücksichtgung des komplexen Stichprobendesigns zu rechnen ist?
- Gibt es die Möglichkeit einer Korrektur? Wie lassen sich Regressionsbäume bilden, die das Stichprobendesign angemessen berücksichtigen?
- Was lässt sich über Random Forest bei komplexen Stichprobendesign sagen?
- Was lernt man aus den Ergebnissen für andere Verfahren und Fragestellungen?

Zur Beantwortung der Fragestellungen wurde die konkrete analytische Bestimmung von Regressionsbäumen aufgebrochen und untersucht. Dabei zeigte sich, dass der wesentliche Konstruktionsschritt

als Varianzschätzungsproblem reinterpretiert werden kann. Dies erlaubt es, Ergebnisse aus der klassischen Survey-Statistik zur unverzerrten Varianzschätzung bei komplexen Stichprobendesigns heranzuziehen. Dadurch konnte das Ausmaß der Verzerrung bei verschiedenen Designs charakterisiert werden und entsprechende Korrekturen vorgeschlagen werden, die sich als erfolgreich erwiesen.

3.2 Zur Erinnerung: Grundlegende Konzepte komplexer Stichprobendesigns, Konstruktion von Regressionsbäumen

Zur Erinnerung I: Komplexe Stichprobendesigns

Im Folgenden sei jedes Stichprobendesign[4] als *komplex* bezeichnet, bei dem nicht jede Einheit dieselbe Wahrscheinlichkeit besitzt, in die Stichprobe zu gelangen. Bekanntermaßen entstehen solche Designs unmittelbar, wenn das gewählte Stichprobendesign Schichtungen nach bestimmten Merkmalen oder geclusterte Bestandteile enthält. Dies kann geplant geschehen, wenn Schichtungseffekte auf die Varianz ausgenutzt werden sollen, oder auch sich implizit ergeben, wenn keine Urliste der Gesamtpopulation vorliegt, sondern nur übergeordnete Mengen von Einheiten erfasst werden können. Ein anderes wichtiges Beispiel, auf das auch hier wiederholt zurückgegriffen wird, ist die *größenproportionale Ziehung*, auch als *PPS-(Proportional Per Size) Stichprobe* bezeichnet. Hier wird mithilfe von Auswahlwahrscheinichkeiten, die proportional zu einem mit dem interessierenden Merkmal korrelierenden Hilfsmerkmal sind, gezogen. Dies stellt sicher, dass in dem Bereich der Werte, der meist inhaltlich besonders interessant, aber oft auch sehr volatil ist, mit überdurchschnittlich vielen Beobachtungen, und somit mit einer größeren Genauigkeit, gerechnet werden darf. Ein praktisches Beispiel für ein solches Untersuchungsmerkmal könnte der Nettogewinn von Unternehmen sein; als Hilfsmerkmale könnten der Umsatz oder auch der Gewinn bei einer früheren Totalerhebung dienen.

Wichtig für die Formalisierung unten stehender Überlegungen sind folgende Notationen und die darauf aufbauenden gängigsten Schätzer: Für eine endliche Gesamtheit $\mathcal{U}$ von Einheiten $i = 1, \ldots, N$ sei die (als echt positiv vorausgesetzte) *Inklusionswahrscheinlichkeit erster Ordnung*, also die Wahrscheinlichkeit, dass i in einer konkret gezogenen Stichprobe vom Umfang n ist, mit π_i bezeichnet; gelegentlich benötigt wird auch die Inklusionswahrscheinlichkeit zweiter Ordnung π_{ij}, dass beide Elemente eines Paares $i, j, i \neq j$ von Einheiten gleichzeitig in der Stichprobe sind.[5] Bei der einfachen Zufallsauswahl vom Umfang n ist die Inklusionswahrscheinlichkeit π_i für jede Einheit i gleich, nämlich n/N. Bei der größenproportionalen Ziehung auf der Basis eines Hilfsmerkmals A mit Ausprägungen a_i setzt man, sofern $\pi \leq 1$ sichergestellt ist

$$\pi_i = n \frac{a_i}{\sum_{j=1}^{N} a_j} \,. \tag{1}$$

Bei der Schätzung des Populationsmittelwerts der Variable Y aus den Werten $y_1, \ldots, y_n$ einer Stichprobe sind die Inklusionswahrscheinlichkeiten als inverse Gewichte

$$w_i = 1/\pi_i \tag{2}$$

zu berücksichtigen, um systematische Verzerrungen zu vermeiden. Verwendet werden meist der Horvitz-Thompson-Schätzer

$$\bar{y}_{HT} = \frac{1}{N} \sum_{i=1}^{n} y_i w_i \,, \tag{3}$$

[4] Siehe etwa Lohr, 2021 für eine einführende Monographie.

[5] Wir gehen nachfolgend davon aus, dass die Inklusionswahrscheinichkeiten bekannt (oder hinreichend genau ermittelt worden) sind.

oder alternativ ein auf Hajek zurückgehender Vorschlag (Särndal, Swensson und Wretman, 2003)

$$\hat{\bar{y}}_{HJ} = \frac{\sum_{i=1}^{n} y_i w_i}{\sum_{i=1}^{n} w_i} \,. \tag{4}$$

Letzterer ist auch für Situationen geeignet, in denen der Populationsumfang N nicht bekannt ist. Der Schätzer ist nur approximativ erwartungstreu; in vielen Anwendungen als attraktiv gilt aber, dass durch den gewichtungsabhängigen Nenner der starke Einfluss extrem großer Gewichte etwas kompensiert wird.

Generell kann gerade bei der größenproportionalen Auswahl die Varianz substantiell reduziert werden (z.B. Valliant, Dever und Kreuter, 2018): Extreme Beobachtungen werden häufiger von der Stichprobe erfasst, gehen aber nur mit niedrigerem Gewicht in den Schätzer ein; unter fiktiven Wiederholungen der Zufallsauswahl und des Schätzvorgangs würden die resultierenden Schätzwerte homogener und damit ihre Varianz geringer.

Erinnerung II: Regressionsbäume

Klassifikations- und Regressionsbäume gelten als prototypisches Beispiel algorithmischer Lernprozesse. Ihr Ziel ist es, den Raum der potentiellen Werte der Feature-Variablen in optimaler Weise so in verallgemeinerte Quader einzuteilen, dass diese in sich jeweils möglichst homogen sind bezüglich der Zielvariablen, so dass man dann aus ihnen eine spezifische Prädiktion ableiten kann. Betrachtet man wie hier vorwiegend metrische Zielvariablen, also Regressionsbäume, so dient bei einem neu beobachteten Feature-Vektor der Mittelwert der Werte aller Beobachtungen der Zielvariable, die in den entsprechenden Quader fallen, als Prädiktion. Wesentlicher Vorteil dieser Quaderstruktur ist ihre Nichtlinearität in den Feature-Variablen, so dass sich auch komplexe Interaktionen ergeben können, ohne dass man wie bei der klassischen Regression ihre Gestalt fest vorgeben müsste.

Charakteristisch für (Klassifikations- und) Regressionsbäume ist nun, dass nicht in einem einzigem Schritt direkt die homogenste Aufteilung gesucht wird, sondern rekursiv gearbeitet wird (*rekursives Partitionieren*). Dabei wird in jedem Schritt bis zur Erfüllung eines Stopkriteriums eine informationsoptimale Teilung in Richtung einer einzigen Variablen (*optimale Splitvariable*) an einem *optimalen Splitpunkt* vorgenommen und dann die entstehenden beiden "Datensätze" links und rechts des Splitpunkts getrennt weiter verarbeitet. Konkret wird also folgender Algorithmus verwendet:

- Wiederhole bis Stop
  1. Iteriere durch alle Variablen $x_1, \ldots, x_q$ und für jede Variable durch alle möglichen Splitpunkte s_j im Wertebreich $\mathcal{X}_j$ der Variable x_j !
  2. Für jeden Splitpunkt s_j bestimme die Mittelwertsprädiktion $(\widehat{\mu_L}, \widehat{\mu_R})$ aus den links und rechts gebildeten Datensätzen!
  3. Bestimme über alle Variablen und alle Splitpunkte den "informations-optimalen" Splitpunkt s_j^* !
  4. Teile die Daten anhand des optimalen Splitpunkts s_q^* und wiederhole das Vorgehen für die neu entstandenen Datensätze!

Wie in Abbildung 1 grob skizziert, produziert dieser Algorithmus eine Baumstruktur mit Entscheidungsknoten und nicht mehr zu verfeinernden Blättern: Bei den Entscheidungsknoten (dargestellt durch Rechtecke) werden die Beobachtungen nach rechts und links aufgeteilt; bei den Blättern

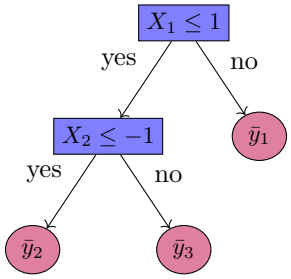

Abbildung 1: Skizze der Baumstruktur eines Regressionsbaums.

(durch Kreise repräsentiert) wird prädiziert, wobei als Prädiktion der Mittelwert aus den Werten der Zielvariable aller Beobachtungen herangezogen, die diesen Knoten erreicht haben.

Diese sehr anschaulich interpretierbare Baumstruktur hat sicher auch entscheidend zur Beliebtheit von Klassifikations- und Regressionsbäumen in der Praxis beigetragen. Ein wichtiger Nachteil der Baumstruktur, gerade im hier betrachteten Kontext, ist allerdings, dass sich natürlicherweise eine "unglückliche Auswahl" eines Splits auf alle nachfolgenden Bestandteile des jeweiligen Astes auswirkt. So führt etwa, wie sich nachfolgend (vgl. Abschnitt 3.8) auch im Beispiel zeigen wird, eine durch das Ignorieren des Stichprobendesigns hervorgerufene falsche Wahl der ersten Splitvariable zu einem stark unterschiedlichen Baum.

Ein weiterer Nachteil von Klassifikationsbäumen ist ihre Tendenz zum "Overfitting", also die Gefahr zufällig entstandene Strukturmuster als Bestandteil der eigentlichen Struktur zu interpretieren. Diese Gefahr hat zu einer Reihe von Weiterentwicklungen geführt. Am prominentesten sind *Random Forests*, bei denen verschiedene Einzelbäume aggregiert werden, die jeweils durch zufällige Auswahlen der Datenpunkte und der zur Erzeugung von Splits zur Verfügung stehenden Variablen generiert werden; sie bringen anschaulich gesprochen, verschiedene Aspekte des Datensatzes zur Geltung bringen. In der Tat wird dadurch meist die Prädiktionsgüte deutlich gesteigert, allerdings auf Kosten einer direkten Interpretierbarkeit.

Die nachfolgend berichteten Ergebnisse fußen in weiten Teilen auf der aus dem Projekt hervorgegangenen Artikelpublikation (Nalenz, Rodemann und Augustin, 2024), wobei aus darstellerischen Gründen die Notation teilweise abgewandelt wird.

3.3 Potentielle Verzerrung des naiven Vorgehens

Erste Voruntersuchungen zeigten unmittelbar, dass in der Tat zu vermuten ist, dass eine Vernachlässigung des Stichprobendesign deutliche Auswirkungen auf die Ermittlung von Regressions- und Klassifikationsbäumen hat. Sofort erkennt man: die aus Simulationen etwa unter einem PPS-Design gewonnenene Daten und analoge i.i.d. Daten sind so unterschiedlich, dass sie zu typischerweise zu unterschiedlichen Bäumen führen. Dies bestätigt sich auch an einem zur Illustration herangezogenen Datensatz über Wohnungsverkäufe in Seoul (vgl. Abschnitt 3.8).

Bezeichnet man – in Anlehnung an die Messfehlertheorie – alle Methoden und Verfahrensschritte, die das komplexe Stichprobendesign ignorieren, als "naiv", so zeigt sich in der Tat auch in Simulationen, dass die der naiven Auswahl zugrundeliegende Kriteriumsfunktion von Splitvariablen und Splittpunkten deutlich verzerrt ist (vgl. die blau gestrichelte Kurve in Abbildung 3). Wichtig ist

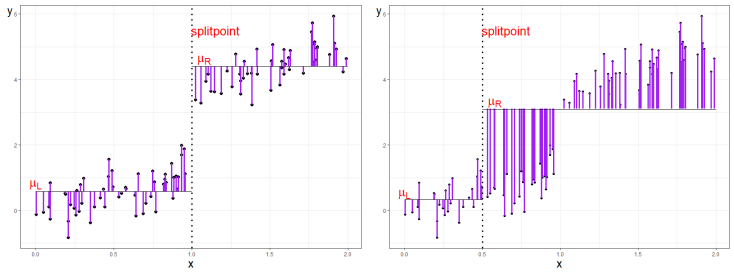

Abbildung 2: Beispiele für potentielle Splitpunkte: links: bester Splitpunkt, rechts: schlechter Kandidat

dabei auch festzuhalten, dass es sich um prinzipielle Verzerrungen handelt, die auch nicht durch einen beliebig wachsenden Umfang der Trainingsdaten gelindert werden können; der Bias der naiven Schätzung verschwindet auch nicht asymptotisch.

3.4 Technisches Hauptargument: Impurity-Reduktion als Varianzreduktion

Die weitere Analyse und die Entwicklung von Korrekturverfahren hat sich mit den Kriterien der Auswahl von Splitvariablen und Splitpunkt auseinanderzusetzen. Die restlichen Schritte, mit denen ein Regressionsbaum erzeugt wird, sind vom Stichprobendesign unabhängig.

Als optimaler Splittpunkt gilt derjenige Punkt, bei dem die Aufteilung den größten Informationsgewinn bietet in dem Sinne, dass die Impurity (Unreinheit) maximal reduziert wird (vergleiche zur Illustration Abbildung 2). Üblicherweise wird die Impurity gemessen über eine quadratische Verlustfunktion, die die Unterschiede zwischen den einzelnen Beobachtungen und der jeweiligen Prognose bestimmt. Bei der Mittelwertsprognose ergeben sich hier also einfach die einzelnen Summanden des MSEs. Als zugehörige Risikofunktion erhält man den *erwarteten Informationsgewinn* $\Delta(s)$ eines Aufteilens im potententiellen Splitpunkt s

$$\Delta(s) = \sigma_B^2 - p_L\sigma_L^2 - p_R\sigma_R^2\,. \tag{5}$$

Dabei bezeichnet σ_B^2 die Varianz ohne Split, während σ_L^2, σ_R^2 für die Varianz im linken bzw. rechten Teil stehen. Analog sind p_L und p_R die Wahrscheinlichkeiten, dass eine Beobachtung dem linken bzw. dem rechten Teil angehört. Hierdurch ist das Lernen eines Klassifikationsbaum auf das Arbeiten mit Wahrscheinlichkeiten und Varianzen, bzw. Anteilen und empirischen Varianzen, zurückgeführt. Dies ist insbesondere hilfreich, da für diese Größen zahlreiche Ergebnisse aus der allgemeinen Survey-Statistik für komplexe Stichproben zur Verfügung stehen; ihre Übertragung erlaubt eine Beurteilung der erwarteten Verzerrung und die Entwicklung korrigierter Verfahren.

3.5 Einfluss des Stichprobendesigns

Der Einfluss (der Vernachlässigung des) Stichprobendesigns lässt sich damit auf die Frage nach der Verzerrung der naiven Varianzschätzung zurückspielen. Hierzu liegen geeignete Ergebnisse der klassischen Survey-Statistik vor. So zeigten Courbois und Urquhart, 2004, dass unter einem

komplexen Design für die Verzerrung des naiven Varianzzschätzers $\widehat{\sigma^2}_{Naive}$ – unter der Verwendung der stichprobentheoretischen Konvention, die Werte der Zielgröße aus der Grundgesamtheit mit $Y_i, i = 1, \ldots, N$ zu bezeichnen– gilt:

$$\mathbb{E}(\sigma^2 - \widehat{\sigma^2}_{Naive}) = \frac{1}{n}\sum_{i=1}^{N} Y_i^2 \underbrace{\left(\pi_i - \frac{n}{N}\right)}_{\Delta_i} - \frac{1}{n(n-1)}\sum_{j=1}^{N}\sum_{j\neq i} Y_i Y_j \underbrace{\left(\pi_{ij} - \frac{n(n-1)}{N(N-1)}\right)}_{\Delta_{ij}} . \tag{6}$$

Die Ausdrücke Δ_i und Δ_{ij} stellen genau die Abweichungen zu den Standard-Inklusionswahrscheinlichkten einer einfachen Zufallsauswahl ohne Zurücklegen dar, wobei sich diese Abweichungen über alle Einheiten hinweg kompensieren:

$$\sum_{i=1}^{N} \Delta_i = 0 \qquad \text{and} \qquad \sum_{i\neq j}^{N} \Delta_{ij} = 0\,.$$

Wichtig für die Frage nach der Verzerrung ist insbesondere der erste Summand. In ihn geht entscheidend die Korrelation zwischen Y^2 und Δ ein; Abweichungen von der einfachen Zufallsauswahl sind ceteris paribus umso verzerrender, je stärker sie mit dem Quadrat der Zielvariable korrelieren. Dies macht sofort deutlich, dass bei größenproportionaler Auswahl immer mit einem beträchtlichen Bias komplexer Struktur bei der Varianzschätzung zu rechnen ist, der sich dann auch in einer verzerrten Lage des Optimums und damit in einer falschen Wahl von Splitpunkten und -variablen niederschlagen wird. Bei einer geschichteten Stichprobe etwa kommt es darauf an, wie stark das Schichtungsmerkmal mit der Zielgröße zusammenhängt; nur bei mit Y^2 unkorrelierten Inklusionswahrscheinlichkeiten verschwindet diese Verzerrung.

3.6 Korrigierte Regressionsbäume

Diese Überlegungen zeigen, dass es als nächster Schritt in der Tat notwendig ist, nach einer korrigierten Schätzung der wahren Kriteriumsfunktion, also des erwarteten Informationsgewinn zu suchen. Ähnlich wie in Gleichung (4) wird man dabei auf einen Schätzer der Hajek-Form zurückgreifen, bei dem die Zahl der Beobachtungen nicht apriorisch bekannt sein muss.

Behandelt man die Mittelwertschätzer als fest und bekannt, vernachlässigt also ihre Variabilität, so lässt sich ein auf Lumley, 2004 zurückgehender Plug-in-Schätzer vom Hajek-Typ als korrigierter Varianzschäzter heranziehen:

$$\hat{\sigma}^2_{HJ} = \frac{\sum_{i=1}^{n} w_i (y_i - \hat{\hat{y}}_{HJ})^2}{\sum_{i=1}^{n} w_i}\,. \tag{7}$$

Will man die Varianzkorrekturen in die zentrale Formel (5) einbringen, so fehlen noch die Schätzungen der zu erwartenden Populationsanteile links und rechts vom Split. Auch hier ist auf eine entsprechende Hajek-Version der Anteilsschätzung unter Gewichten zurückzugreifen.

Insgesamt ergibt sich also als neue Zielfunktion die folgende, um das komplexe Stichprobendesign korrigierte, Schätzung des erwarteten Informationsgewinns (Nalenz, Rodemann und Augustin, 2024, p. 3384)

$$\Delta_{HJ}(s) = \widehat{\sigma^2}_{HJ} - \widehat{p_L}_{HJ}\,\widehat{\sigma^2_L}_{HJ} - \widehat{p_R}_{HJ}\,\widehat{\sigma^2_R}_{HJ}\,, \tag{8}$$

wobei, mit $\mathcal{L}$ und $\mathcal{R}$ als Indexmengen der nach links bzw. rechts fallenden Einheiten, $\widehat{p_L}_{HJ} = \sum_{i\in\mathcal{L}}^{n} w_i / (\sum_{i\in\mathcal{L}}^{n} w_i + \sum_{j\in\mathcal{R}}^{n} w_j)$ und $\widehat{p_R}_{HJ} = \sum_{i\in\mathcal{R}}^{n} w_i / (\sum_{i\in\mathcal{L}}^{n} w_i + \sum_{j\in\mathcal{R}}^{n} w_j)$ die angesprochenen

Hajek-Schätzer der links/rechts-Anteile sind und $\widehat{\sigma^2_{LHJ}}$ bzw. $\widehat{\sigma^2_{RHJ}}$ die Hájek-Varianzschätzer im linken bzw. rechten Subknoten sind.

Die klassische Stichprobentheorie bietet auch noch eine alternative Variante an, die Varianz zu schätzen, die auf Chaudhuri, 1978 zurückgeht und explizit von den Inklusionswahrscheinlichkeiten 2. Ordnung Gebrauch macht. Wäre neben π_{ij} auch N bekannt, so lieferte

$$\widehat{\sigma^2_*} = \frac{1}{2N^2} \sum_{i \neq j}^{n} \frac{(y_i - y_j)^2}{\pi_{ij}} \,. \tag{9}$$

einen erwartungstreuen Schätzer für die Populationsvarianz. Um diesen Schätzer auf die unbekannte Populationsgröße in den Blättern zu adaptieren, bietet es sich an, wie beim Hajek-Schätzer vorzugehen und das jeweilige N über $\sum_{i=1}^{n} w_i$ zu schätzen. Ersetzt man in (8) die einzelnen Hajek-Schätzer durch ihre Gegenstücke im Sinne von (9), so erhält man eine alternative Form einer Korrektur um das Stichprobendesign.

3.7 Simulationsergebnisse

Die Korrekturansätze werden anhand von Simulationsdaten in Abbildung 3 zusammen mit der Kurve des naiven Schätzers und der wahren Zielfunktion dargestellt. Wie oben bereits angesprochen ist die naive Kriteriumsfunktion nicht nur in Höhe sondern insbesondere in der Lage verzerrt, während die Korrekturverfahren allesamt ähnliche Ergebnisse aufweisen, die erfreulicherweise nahe an der Populationskurve liegen. Mit wachsendem Stichprobenumfang n nivellieren sich die Unterschiede zwischen den korrigierten Ansätzen weiter, während der beträchtliche Bias des naiven Vorgehens erhalten bleibt.

3.8 Illustratives Datenbeispiel:

Die hergeleiteten und an der Simulation überprüften Ergebnisse wurden in Nalenz, Rodemann und Augustin, 2024, Section 6 auch an einem Datenbeispiel zu Verkäufen von Wohneigentum in Seoul illustriert. Der grundlegende Datensatz besteht aus 2.65 Millionen Häusern und Wohnungen, die zwischen 2005 and 2023 verkauft wurden; ausgewertet wird eine selbstgezogenen PPS-Stichprobe vom Umfang 1000. Zielvariable ist der Preis in 10 Millionen Won[6].); als Feature-Variablen dienen die Wohnungsgröße in Quadratmeter (sqm), die Grundstücksgröße in Quadratmeter ($lsqm$), das Baujahr ($YearBuilt$) und das Verkaufsjahr ($YrSold$). Während der naive Baum eine deutlich andere Struktur mit einem ersten Split in der Grundstücksgröße produziert, sind sich der Populationsbaum und der Hajek-korrigierte Baum deutlich ähnlicher, mit sogar einem identischen Spilt in der ersten Variable.

3.9 Korrigierte Random Forests

Will man Random Forests an das komplexe Stichprobendesign anpassen, bieten sich zwei Möglichkeiten an:

[6] 1 Euro entspricht etwas weniger als 1.500 Won.https://www.bundesbank.de/dynamic/action/de/statistiken/zeitreihen-datenbanken/zeitreihen-datenbank/723452/723452?tsId=BBEX3.D.KRW.EUR.BB.AC.000&dateSelect=2024; 28.6.2024

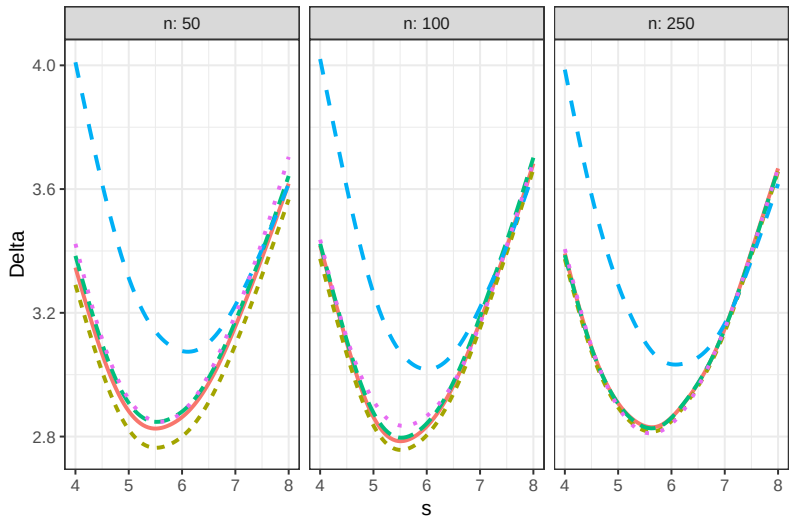

Abbildung 3: Exemplarische Simulation zur Darstellung der Korrektur des Information Gains gemäß(5): wahres $\Delta(s)$ in der Population (violett, gepunktet), naive Schätzung (blau, gestrichelt), $\widehat{\Delta}_{HJ}(s)$ (grün, vgl. (8) und die Korrektur unter Verwendung von $\widehat{\sigma_*^2}$ mit bekanntem bzw. geschätztem N (durchgehende rote bzw. gestrichelte gelbe Linie); Graphik aus Nalenz, Rodemann und Augustin, 2024, p. 3386

- **Hajek-korrigierte Forests**: Die erste Variante schließt unmittelbar an die bisherigen Überlegungen an, indem Sie einen korrigierten Wald aus lauter korrigierten einzelnen Bäumen konstruiert. Konkret entsteht auf natürliche Weise der nachfolgende "Hajek-korrigierte Forest" als Ensemble aus Hajek-korrigierten Bäumen.
- **Bootstrap-korrigierte Wälder**: Hier macht man sich die Tatsache zu nutze, dass in der Produktion des Waldes ja Zufallsstichproben eine entscheidende Rolle spielen; die einzelnen Bäume werden aus Bootstrap-Stichproben der PPS-Stichprobe generiert. Intuitiv mag es einen Versuch wert sein, die PPS-Struktur zu kompensieren, indem man die Beobachtungen mit großer Inklusionswahrscheinlichkeit nach unten gewichtet und umgekehrt. Konkret wurde hier mit den inversen Inkluionswahrscheinlichkeiten gearbeitet. (Interessanterweise lässt sich dieses Vorgehen im Paket Wright und Ziegler, 2017 unmittelbar umsetzen, das ganz allgemein im Bootstrap-Prozess Fallgewichte zulässt.)

Beide Verfahren wurden intensiver untersucht und verglichen. Sie erweisen sich in der Tat als leistungsfähige Korrekturen; beim Hajek-korrigierten Forest ist es bei sehr schiefen Verteilungen wichtig, rechtzeitig zu prunen. Wie die verschiedenen Simulationsszenarien in Nalenz, Rodemann und Augustin, 2024, Section 5 deutlich machen, ist die Güte des naiven Verfahrens hingegen sehr situationsabhängig; der MSE schwankt zwischen vergleichbarer Güte bis hin zu absolut inakzeptablen Größenordnungen. Eine große Rolle spielt dabei, wie schief die bedingte Verteilung der Zielvariable und in wieweit sehr große Werte noch einer Struktur zuordenbar sind oder als reines Rauschen zu interpretieren sind.

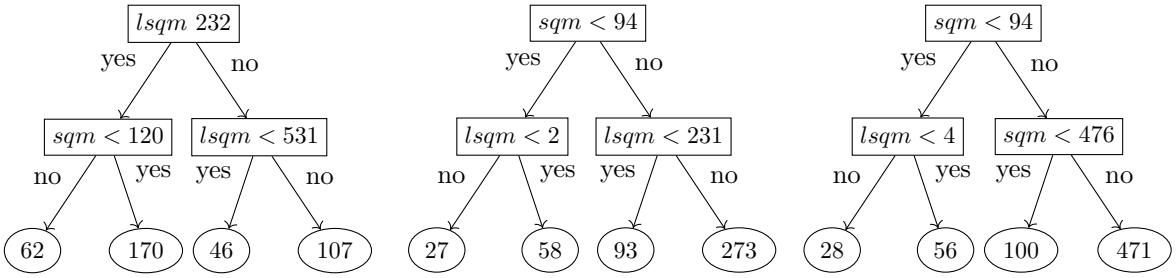

Abbildung 4: Datenbeispiel zu Wohnungsverkäufen in Seoul: naiver Baum (links), Baum auf der Gesamtpopulation der Daten (Mitte), Hájek-korrigierter Baum (rechts); Graphiken unter anderer Anordnung aus Nalenz, Rodemann und Augustin, 2024, Section 6.

	$n = 100$	$n = 500$	$n = 1000$
Naive	1.69	1.77	1.59
Hájek forests	1.11	0.98	0.96
Regularized[7] Hájek forests	1.10	1.12	1.20
Weighted bootstrap	0.93	0.88	0.90

Tabelle 1: Relativer MSE der verschiedenen Random-Forest Methoden unter einem PPS-Design im Vergleich zu einem Random Forest auf Daten einer einfachen Zufallsauswahl; Tabelle aus Nalenz, Rodemann und Augustin, 2024, p. 3394

Die Methoden wurden auch auf den illustrierenden Datensatz zu den Wohnungsverkäufen in Seoul angewendet. Wiedergegeben seien hier kurz die Ergebnisse für den relativen MSE im Vergleich zu einer Random-Forest Analyse direkt mit Daten, die aus einer reinen Zufallsauswahl aus dem Datensatz stammen (siehe Tabelle 1 und Nalenz, Rodemann und Augustin, 2024, p. 3394; für Partial Dependence Plots und permutationsorientierte Variablenwichtigkeit siehe ebd., Abschnitt 6.1).

3.10 Was lernt man aus den Ergebnissen für verwandte Situationen?

Adaptive Designs im Machine Learning

Generell bemerkenswert ist an dem eben besprochenen Beispiel auch, dass die Korrekturverfahren sogar eine größere Genauigkeit erzielen können als eine direkte Analyse auf einer einfachen Zufallsstichprobe. Komplexe Auswahlverfahren können – sofern dies bei der Analyse der Daten geeignet berücksichtigt wird – für maschinelle Lernverfahren besonders informativ sein. Dies stellt einen unmittelbaren Bezug her zu all jenen Verfahren des maschinellen Lernens, die auf einer adaptiven Wahl der weiter zu prozessierenden Datenpunkte beruhen, wie zum Beispiel die Bayesianische Optimierung oder das sog. semi-überwachte Lernen. Projektergebnisse zu diesen Methoden werden in Kapitel 4 dargestellt, da sie in besonderer Weise auf mengenwertigen Analyseparadigmata aufbauen.

Generalisierungsfehler unter komplexen Stichproben

Eine der Nichtstandardsituationen in der Studie zum Generalisierungsfehler (Hornung u. a. (2023), siehe auch Abschnitt 1.2 zu Arbeitspket 1.2 in diesem Bericht) waren komplexe Stichprobendesigns, vorwiegend unter einem PPS-Design. Dabei wurde analytisch gezeigt, dass der naive Schätzer des Generalisierungsfehlers verzerrt sein kann, und eine Horvitz-Thompson Korrektur hergeleitet. Eine Simulationsstudie bekräftigte die Vermutung, dass der Fehler etwa von Random-Forests beträchtlich sein kann und auch bei wachsendem Stichprobenumfang erhalten bleibt; ähnliches gilt auch für lineare Modelle, sofern diese fehlspezifiert sind, Die vorgeschlagene Horvitz-Thompson Korrektur erweist sich in der Tat als zielführend; ähnliches gilt für eine von Hodel (2024) in ihrer Seminararbeit untersuchte analoge Hajek-Korrektur.

Software und Dokumentation

Der Hauptartikel (Nalenz, Rodemann und Augustin, 2024) ist in dem renomierten Journal *Machine Learning* frei zugänglich publiziert. Die Implementierungen aller vorgestellten Methoden sowie Skripte zur Reproduktion der präsentierten Experimente sind auf Github unter `https://github.com/maltenlz/ComplexTreesAndForests` zur freien Verwendung verfügbar. Interessanterweise sind die in Nalenz, Rodemann und Augustin, 2024 theoretisch hergeleiteten Korrekturverfahren bereits zum Teil in gängigen Softwarepaketen implementiert, wie etwa in den `R`-Paketen `rpart` von Therneau und Atkinson, 2022 und `ranger` von Wright und Ziegler, 2017. Dies spricht für deren praktische Relevanz einerseits und ermöglicht andererseits eine einfach Integration in bestehende Implementierungen, die auf einem der Pakete basieren.

4 Mengenwertige Verfahren zur Unsicherheitsquantifizierung

4.1 Unsicherheit im Maschinellen Lernen

Unsicherheitsquantifizierung (UQ) im maschinellen Lernen gewinnt für die amtliche Statistik an Bedeutung (Molladavoudi und Yung, 2023). UQ ermöglicht es, die Zuverlässigkeit von Vorhersagen zu bewerten und Risiken besser einzuschätzen. Traditionell liefern viele Machine-Learning-Modelle Punktvorhersagen, ohne anzugeben, wie sicher diese sind. Doch gerade in der amtlichen Statistik, wo Entscheidungen weitreichende gesellschaftliche Auswirkungen haben können, ist es besonders wichtig zu wissen, wie vertrauenswürdig eine Vorhersage ist.

Durch die Quantifizierung von Unsicherheit können statistische Ämter besser informierte und fundiertere Entscheidungen treffen. Insgesamt trägt also die Unsicherheitsquantifizierung dazu bei, die Transparenz und Verlässlichkeit von Machine-Learning-Modellen in der amtlichen Statistik zu erhöhen.

Konkret wird dies bei der Einhaltung diverser Qualitätsstandards (siehe insbesondere die Entwicklung des Quality Framework for Statistical Algorithms: Yung u. a., 2022). Diese sehen vor, dass die Veröffentlichung statistischer Schätzungen stets mit Informationen zu deren Unsicherheit einhergeht (Molladavoudi und Yung, 2023). In der klassischen Statistik sind UQ-Methoden seit langem integraler Bestandteil statistischer Modelle. Ein in der Survey-Statistik beliebtes Beispiel ist das Verhältnis einer Schätzung zu ihrem Standardfehler anzugeben. Eine solche direkte, generische Möglichkeit der Unsicherheitsquantifizierung gibt es bei Methoden des Maschinellen Lernens nur bedingt. Dies hat die Entwicklung neuer UQ-Methoden in den vergangenen Jahren angeregt. Abdar u. a., 2021; Hüllermeier und Waegeman, 2021; Gruber u. a., 2023 bieten einen Überblick über das wachsende Forschungsfeld der UQ im Machine Learning.

4.2 Imprecise Probabilities und Mengenwertige Verfahren

Grundsätzlich wird in der Literatur zur Unsicherheitsquantifizierung im maschinellen Lernen zwischen epistemischer (reduzierbarer) und aleatorischer (nicht reduzierbarer) Unsicherheit unterschieden. Letztere ist dem Zufallsprozess inhärent und entzieht sich daher meist der expliziten Einflussnahme durch Akteure wie statistische Ämter. Epistemische Unsicherheit entsteht im Kontext des maschinellen Lernens hingehen unter anderem durch (zumeist probabilistische) Modellierungsentscheidungen. Beispiele hierfür sind die Annahme einer bestimmten Verteilungsklasse oder die Verwendung einer Verlustfunktion.

Als ein vielversprechender Forschungszweig zur Quantifizierung dieser Art von Unsicherheit haben sich mengenwertige Verfahren herauskristallisiert, siehe etwa Hüllermeier, Destercke und Shaker, 2022; Sale, Caprio und Hüllermeier, 2023; Caprio u. a., 2023. Diese gründen zu einem beachtlichen Teil auf der Theorie der *Imprecise Probabilities* (Walley, 1991; Augustin u. a., 2014). Die grobe Idee hierbei ist, dass sich Akteure – wie etwa statistische Ämter – nicht auf die Annahme einer einzigen Verteilung oder einer einzigen Verlustfunktion festlegen müssen. Stattdessen bildet eine *Menge* von Verteilungen oder Verlustfunktionen die Unsicherheit der Akteurin darüber ab. Im Falle von Verteilungsannahmen (oder dazu korrespondierenden Verlustfunktionen) entspricht eine solche Menge einer Menge von Wahrscheinlichkeitsmaßen, in diesem Kontext als *Credal Sets* bezeichnet. Die Theorie der *Imprecise Probabilities* bietet reichhaltige Möglichkeiten, anhand solcher *Credal Sets* auf robuste Art und Weise statistische Schlüsse oder Prädiktionen im Sinne des maschinellen Lernes aus Stichproben abzuleiten. Die Grundidee ist denkbar simpel: Die Prädiktionen basieren nun nicht mehr auf einem *einzigen* ML-Model (und dem damit korrespondierenden Wahrscheinlichkeitsmaß),

sondern auf einer Menge solcher Modelle. Die Vorteile liegen auf der Hand. Die Prädiktionen bleiben valide, auch wenn man das *korrekte* Model nicht genau kennt, es aber Teil der spezifizierten Menge an Modellen ist. Darüber hinaus lässt sich die epistemische Unsicherheit in den Prädiktionen nun explizit mit der Modelwahl in Verbindung bringen. Man erhält beispielsweise ein Intervall für reellwertige Prädiktionen, das als die durch die Modellwahl verursachte Variation in den Prädiktionen interpretiert werden kann.

4.3 Mengenwertige ML-Verfahren

Im Folgenden zeigen wir anhand dreier für die amtliche Statistik relevanter Beispiele, wie mengenwertige Verfahren dabei helfen, Methoden des Maschinellen Lernens verlässlicher und effizienter zu machen. Die Beispiele beruhen jeweils auf Forschungsarbeiten, die als, teilweise mit AP 1.4 (vgl. hier Kapitel 3) eng verwobener, PI-Eigenbeitrag im Rahmen dieses Kooperationsprojekts entstanden sind.

4.3.1 Spotlight: Robuste Bayesianische Optimierung

Bayesianische Optimierung (BO) ist eine Methode zur Optimierung unbekannter Funktionen, die teuer zu evaluieren sind, wie z.B. in der Hyperparameteroptimierung von Machine-Learning-Modellen. Sie basiert auf dem Bayes'schen Theorem und verwendet ein probabilistisches Modell, typischerweise einen Gauß-Prozess (GP), um die zu optimierende Funktion zu lernen. Durch die Kombination von Vorhersagen und Unsicherheitsabschätzungen wird versucht, mit möglichst wenigen (weil teuren) Funktionsauswertungen einen möglichst optimalen Funktionswert zu finden. Der Prozess besteht aus einer iterativen Schleife, in der das Modell im Lichte neuer Funktionswerte aktualisiert und eine sogenannte Akquisitionsfunktion maximiert wird, um den nächsten Evaluierungspunkt zu bestimmen. Dies ermöglicht es, die Anzahl der teuren Funktionsauswertungen so gering wie nötig zu halten und dennoch ein optimales Ergebnis zu erzielen.

Aufgrund ihrer allgemeinen Formulierung ist Bayesianische Optimierung auf eine große Bandbreite von Optimierungsproblemen anwendbar. BO ist jedoch besonders nützlich in Bereichen, in denen jede Auswertung der Zielfunktion zeitaufwendig oder kostspielig ist. Statistisch betrachtet ist die BO eine Methode des experimentellen Designs und somit eng verwandt mit der für die Survey-Statistik hochrelevanten Stichprobentheorie. Vor diesem Hintergrund liegt eine Verwendung Bayesianischer Optimierung in der amtlichen Statistik auf der Hand.

Um Bayesianische Optimierung allerdings tatsächlich für statistische Ämter zugänglich zu machen, sollte sie ihren Nimbus als effiziente, aber unzugängliche *Black Box* verlieren. Dies erfordert einerseits Methoden des interpretierbaren Maschinellen Lernens, siehe Kapitel 2.2. Beispiele umfassen die Interpretation des zugrundeliegenden Modells (Moosbauer u. a., 2021), der Akquisitionsfunktion (Rodemann u. a., 2024) oder der erreichten Optima (Chakraborty, Seifert und Wirth, 2024). Auf der anderen Seite ist jedoch auch eine Quantifizierung relevanter Unsicherheiten innerhalb der BO unerlässlich. Nur so können die mittels BO gefundenen Lösungen als verlässlich eingestuft und ihre Unsicherheit im Sinne einschlägiger Disseminationsstrategien und Qualitätsanforderungen der Ämter kommuniziert werden.

Die eben angesprochenen Methoden der epistemischen Unsicherheitsquantifizierung mittels mengenwertiger Verfahren bieten sich hierfür an, um mit der *Imprecise Bayesian Optimization* (Rodemann und Augustin, 2024) eine robustifizierte Variante Bayesianischer Optimierung vorzuschlagen: Prior-Robust Bayesian Optimization (PROBO), welche darauf abzielt, die Schwächen klassischer BO-Methoden zu überwinden, die oft empfindlich auf fehlerhafte Modellannahmen

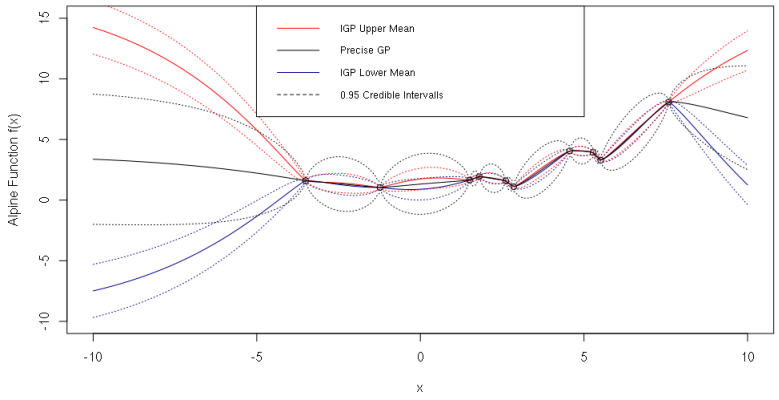

Abbildung 5: Illustration mengenwertiger Gauß-Prozesse: Obere Grenze (rot) und untere Grenze (blau) der GP-Posteriori-Schätzung, inklusive Posteriori-Varianzschätzung (gestrichelt). Vergleiche dazu die klassische präzise Gauß-Prozess-Schätzung (schwarz). Abbildung aus Rodemann und Augustin, 2024, Seite 22.

reagieren. PROBO erlaubt eine Vielzahl möglicher GP-Spezifikationen, um robuster gegenüber Modellunsicherheiten zu sein. Dazu verwendet sie die sogenannte *impräzisen Gauß-Prozesse* (IGP) (Mangili, 2016), welche aus einer Menge von Posteriori-Schätzungen der Funktionen bestehen, aus denen eine obere und untere Grenze abgeleitet werden kann, siehe Abbildung 5. Diese oberen und unteren Grenzen können wie folgt interpretiert werden: Die obere (untere) Grenze ist die Schätzung, die entstanden wäre, hätte man da die maximale (minimale) GP-Spezifikation gewählt. Man muss sich also in der Anwendung nicht mehr auf die Korrektheit einer einzelnen Spezifikation verlassen, was die Ergebnisse verlässlicher macht. Ein wesentlicher Bestandteil von PROBO ist die Generalized Lower Confidence Bound (GLCB), eine neue Akquisitionsfunktion, die speziell entwickelt wurde, um Unsicherheiten in den Mittelwertparametern zu berücksichtigen.

Motiviert wird PROBO mittels einer klassischen Bayesianischen Sensitivitätsanalyse, siehe auch Rodemann und Augustin, 2022. Dabei untersuchen die Autoren die Auswirkungen der Spezifikation einzelner Bestandteile des GP-Modells auf die Konvergenz von BO. Sie stellen fest, dass die Mittelwertparameter des GP die größte Auswirkung auf die Konvergenz haben. Bei falscher Spezifikation dieser Parameter steigen die kumulativen Regret-Werte linear an, während sie bei korrekter Spezifikation sublinear bleiben. Das bedeutet, dass die Leistung der Optimierung stark von den Annahmen über den GP abhängt. Wenn die Annahmen falsch sind, kann dies zu ineffizienten Optimierungen und schlechteren Ergebnissen führen.

In einer detaillierten Simulationsstudie zeigen die Autoren, dass PROBO schneller konvergiert als klassische BO-Methoden. Diese Ergebnisse werden durch Tests an einem realen Optimierungsproblem in der Materialwissenschaft, nämlich der Optimierung der Graphenproduktion, untermauert. Graphen ist ein Material mit herausragenden Eigenschaften, dessen Produktionsoptimierung jedoch komplex und kostspielig ist. PROBO erweist sich dabei als leistungsfähiger und robuster gegenüber Unsicherheiten in den Modellannahmen.

Die Arbeit betont die Bedeutung von Modellunsicherheiten und zeigt, dass deren Berücksichtigung zu besseren und robusteren Optimierungsergebnissen führen kann. Die Autoren beschreiben ausführlich, dass klassische BO-Methoden empfindlich auf falsche Annahmen über die Mittelwertparameter reagieren. Dies führt oft zu einer suboptimalen Performance und linearen kumulativen Regret-Werten. Durch die Verwendung ungenauer GP-Modelle als Surrogate kann PROBO diese Schwächen überwinden und robustere Ergebnisse erzielen.

Als Gründe für die bessere Performanz von PROBO führen Rodemann und Augustin, 2024 an, dass die Flexibilität des Optimierungspfades zur Erfassung globaler Optima erhöht werden kann, indem die Annahmen über die probabilistischen Elemente durch Verllagmeinerte Wahrscheinlichkeiten (Imprecise Probabilities) gelockert werden. Anlehnend an das berühmte Zitat von Charles Manski (Manski, 2003, page 1): „Die Glaubwürdigkeit der Inferenz nimmt mit der Stärke der aufrechterhaltenen Annahmen ab,"wird für das Beispiel BO gezeigt, dass eine Lockerung der Annahmen die Modellierungskapazität der Optimierer und damit ihre Leistung erhöhen kann, was auf ein „Gesetz der abnehmenden Flexibilität"hinweist:

> *„Die explorative Flexibilität der Bayesian Optimization nimmt mit der Stärke der aufrechterhaltenen probabilistischen Annahmen ab."*

Die Allgemeinheit der IP-Modelle erlaubt offensichtlich mehr Flexibilität der BO durch eine zusätzliche explorative Dimension im gut verstandenen Exploration-Exploitation-Abwägung. Dies steht im Einklang mit jüngsten Überlegungen von Hüllermeier und Waegeman, 2021, die eine Aufteilung der reduzierbaren (epistemischen) Unsicherheit in Modellierungsunsicherheit und Approximationsunsicherheit vorschlagen, wobei letztere sich auf die klassische statistische Schätzunsicherheit bezieht. Durch die Exploration der Domäne der zu optimierenden Funktion zielt die klassische Bayesian Optimization auf die Reduktion dieser letzteren Approximationsunsicherheit. Durch die explizite Berücksichtigung der Modellierungsunsicherheit mittels eines Prior-Near-Ignorance-Modells aus IP exploriert PROBO darüber hinaus, um diese zweite Art der reduzierbaren Unsicherheit zu verringern.

Die im Rahmen des Kooperationsprojekts geförderten Forschungsarbeit *Imprecise Bayesian Optimization* (Rodemann und Augustin, 2024) erscheint in Kürze im renommierten Fachjournal *Knowledge-based Systems*. Die dazugehörige Software ist unter `https://github.com/rodemann/imprecise-bayesian-optimization` frei verfügbar. Neben einer ausführlichen Dokumentation enthält das Repository unter anderem eine Reihe von illustrativen Anwendungsbeispielen sowie Code, um die Benchmarking-Experimente zu reproduzieren. Auch Visualisierungen sind verfügbar.

4.3.2 Spotlight: Pseudo-Label Auswahl im Halb-Überwachten Lernen

Halb-überwachtes Lernen (*Semi-supervised Learning, SSL*) hat das Potenzial, die Leistungsfähigkeit von Machine-Learning-Modellen zu erhöhen, indem es sowohl gelabelte (klassifizierte) als auch ungelabelte (unklassifizierte) Daten nutzt. Im Gegensatz zum rein überwachten Lernen, das ausschließlich auf gelabelte Daten angewiesen ist, kombiniert semi-supervised Learning die Stärken von überwachten und unüberwachten Lernmethoden. Dies ist besonders vorteilhaft, da das Labeln von Daten oft zeitaufwendig und teuer ist, während ungelabelte Daten in großen Mengen verfügbar sind.

Durch die Einbindung ungelabelter Daten kann das Modell eine bessere allgemeine Struktur der Daten erfassen und somit genauere Vorhersagen treffen. In den letzten Jahren hat sich dieses Forschungsfeld stark entwickelt, mit Methoden wie dem Co-Training, dem Self-Training und der Nutzung von generativen Modellen. Der Einsatz von semi-supervised Learning stellt somit eine

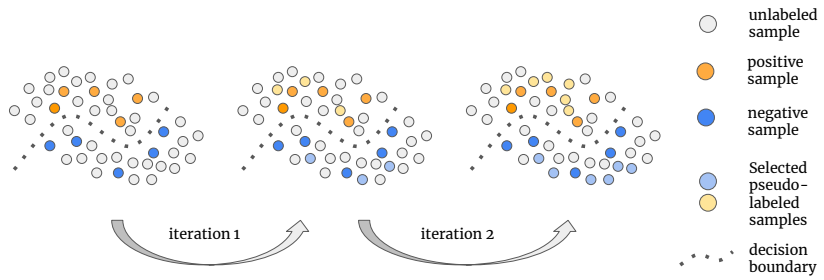

Abbildung 6: Illustration des *self-trainings* im binären Fall Nach einer ersten Modellanpassung an die gelabelten Daten (dunkelblaue und dunkle orangefarbene Punkte), werden Teile der ungelabelten Daten (grau) selbst gelabelt (*pseudo-labels*, hellblaue und hellorange) und für eine erneute Modellanpassung auf diesem erweiterten (Pseudo-)gelabeleten Datensatz verwendet. Das iterative Verfahren wird fortgesetzt, bis ein Stopp-Kriterium erfüllt ist. Abbildung aus Goschenhofer, 2023, Seite 16.

vielversprechende Möglichkeit dar, die Effizienz und Genauigkeit von Machine-Learning-Modellen erheblich zu steigern. Diese Methoden werden bisher insbesondere in Bereichen wie der Bild- und Spracherkennung angewendet, wo sie erhebliche Verbesserungen in der Modellleistung bewirken können; eine Anwendung etwa auf die automatisierte Auswertung offener Fragen oder im Rahmen des Editing processes erscheint attraktiv.

Ein besonders populärer Ansatz im halb-überwachten Lernen ist das sogenannte *self-training*, bei dem das Model sich gewissermaßen selbst trainiert, indem es die ungelabelten Daten einfach selbst labelt und damit trainiert wird. Fundamental dabei ist natürlich die Frage nach einer fundierten Entscheidungsbasis, welchen dieser selbst-generierten Labels man vertrauen kann. Dies war der Ausgangspunkt zu den Arbeiten *Approximately Bayes-Optimal Pseudo-Label Selection* (Rodemann u. a., 2023c) und *In All Likelihoods: Robust Selection of Pseudo-Labeled Data* (Rodemann u. a., 2023d), die beide im Rahmen des Kooperationsprojekt entstanden sind. Beide Arbeiten verdeutlichen, wie wichtig hierfür Methoden der Unsicherheitsquantifizierung sind: Nur wenn bekannt ist, wie hoch die Unsicherheit der prädizierten Label ist, kann man die relativ sicheren auswählen. Insbesondere die letztere Arbeit verdeutlicht, wie fruchtbar mengenwertige Verfahren zur Unsicherheitsquantifzierung aus dem Bereich der *Imprecise Probabilities* hierfür sind.

- **Problemstellung:** Ein prinzipielles Problem beim selbsttrainierenden Ansatz ist, dass er stark von den anfänglichen Modellergebnissen abhängt. Wenn das anfängliche Modell *overfitted*, das heißt, wenn es an die Trainingsdaten überangepasst ist und nicht gut auf neue Daten generalisiert, können falsche Pseudo-Labels erzeugt werden. Diese fehlerhaften Pseudo-Labels werden dem Modell hinzugefügt, was zu einem sogenannten *confirmation bias* führt. Dieser Bias verstärkt die anfänglichen Fehler des Modells und kann die Gesamtleistung erheblich beeinträchtigen. Dieses Problem wird durch die Unsicherheit in den Modellparametern und

den zugrunde liegenden Daten weiter verschärft.

- **Lösung:** Das Papier *Approximately Bayes-Optimal Pseudo-Label Selection* (Rodemann u. a., 2023c) stellt einen neuen Ansatz namens *Bayesian Pseudo-Label Selection (BPLS)* vor, um diese Probleme zu lösen. BPLS nutzt bayesianische Methoden, um Pseudo-Labels auszuwählen. Der Kern von BPLS ist ein Kriterium, das auf einer analytischen Approximation der posterioriprädiktiven Verteilung der Pseudo-Stichproben basiert. Möglich wird der Bayesianische Ansatz durch eine Einbettung der Pseudo-Label Auswahl in die Entscheidungstheorie, siehe auch Rodemann, 2023. Nachdem die Auswahl der Pseudo-Label als Entscheidungsproblem formalisiert wird, erscheinen mehrere bekannte Kriterien in einem neuen Licht. Das optimistische *superset learning* (Rodemann, Kreiss und Hüllermeier, 2022; Hüllermeier, 2014) etwa ergibt sich als Spezialfall des max-max-Kriteriums. Anstatt sich auf eine einzelne Schätzung der Modellparameter zu verlassen, berücksichtigt BPLS die Unsicherheit in diesen Parametern. Wie in der klassischen Bayesianischen Statistik wird über mehrere Modelle gemittelt, sodass pseudo-gelabelte Daten ausgewählt werden, die nicht nur im Lichte des gefitteten Models, sondern im Lichte eines gewichteten Durchschnitts aller möglicher Modelle plausibel erscheinen. Dadurch wird die Robustheit erhöht. Durch die Einbeziehung der Unsicherheit in die Parameter wird BPLS stabiler und widerstandsfähiger gegen Überanpassung. Dies führt zu einer zuverlässigeren Auswahl von Pseudo-Labels, die die Gesamtleistung des Modells verbessern können. Trotz seiner Komplexität ist BPLS rechnerisch effizient. Es verwendet Näherungen wie die Methode von Laplace und das Gaußsche Integral, um die Berechnungen zu vereinfachen, ohne die Genauigkeit wesentlich zu beeinträchtigen, siehe Rodemann u. a., 2023a, Kapitel 3. Diese Näherungen ermöglichen eine effiziente Implementierung von BPLS, die in realen Szenarien angewendet werden kann, ohne erhebliche Rechenressourcen zu erfordern.

- **Empirische Evaluation:** BPLS wurde anhand von simulierten und realen Daten getestet, siehe Rodemann u. a., 2023a, Kapitel 4. Es zeigte sich, dass BPLS herkömmlichen Methoden zur Pseudo-Label-Auswahl überlegen ist, besonders bei hochdimensionalen Daten, die anfällig für *overfitting* sind. Die Tests umfassten verschiedene Datensätze und Szenarien, die die Robustheit und Effizienz von BPLS unter Beweis stellten. Die Ergebnisse zeigen, dass BPLS in der Tat nicht nur die Genauigkeit der Pseudo-Labels verbessert, sondern auch die Gesamtleistung des Modells in verschiedenen Anwendungsfällen erhöht.

Die Einbettung der Auswahl pseudo-gelabelter Daten in Rodemann u. a., 2023a in die Entscheidungstheorie ermöglicht sodann einige robuste Erweiterungen, die darauf abzielen, verschiedene zusätzliche Unsicherheitsquellen zu berücksichtigen. Diese Erweiterungen sind Gegenstand des zweiten Papiers *In All Likelihoods: Robust Selection of Pseudo-Labeled Data* (Rodemann u. a., 2023d). Ein wichtiges Werkzeug für die robusten Erweiterungen sind die oben skizzierten *Credal Sets*, also Mengen von Wahrscheinlichkeitsmaßen, mit den die Unsicherheit bezüglich der Verteilungsannahme beziehungsweise der Modellwahl dargestellt werden kann.

Konkret ermöglicht die Formalisierung der Pseudo-Label Auswahl als Entscheidungsproblem die Einführung einer mehrdimensionalen Nutzenfunktion, die darauf abzielt, pseudo-gelabelte Daten auszuwählen, die vorteilhaft unter einer Reihe unterschiedlicher Szenarien sind. Diese Nutzenfunktion berücksichtigt verschiedene Unsicherheitsquellen wie die Modellwahl, etwaige Fehlerakkumulation und *Covariate Shift*, also eine Veränderung der marginalen Verteilung der Kovariablen. Darüber hinaus schlagen Rodemann u. a., 2023b vor, die generalisierte Bayesianische α-cut-Regel für Credal-Sets zu verwenden, wenn die Unsicherheitsquelle unbekannt ist, siehe unten.

- **Berücksichtigung mehrerer Modellklassen:** Eine offensichtliche und allgegenwärtige Quelle der Ungenauigkeit ist die Modellwahl: Welche Verteilungsannahme (und entsprechendes Modell) sollte bei der Berechnung der Nutzenfunktion berücksichtigt werden? Bisher wurde dafür das Modell verwendet, das wir zur Vorhersage der Pseudo-Labels verwendet haben. Dies ist jedoch keineswegs zwingend erforderlich. Statt sich nur auf eine einzige Modellklasse zu verlassen, berücksichtigt die robuste PLS die Wahrscheinlichkeit, dass die pseudo-gelabelten Daten unter verschiedenen Modellklassen auftreten. Dies führt zu einem multi-objektiven Entscheidungsproblem, das durch eine gewichtete Summe der Wahrscheinlichkeiten der Modelle gelöst wird. Es wird also auf zwei Eben gemittelt: Zum einen wird im Bayesianischen Sinne für ein Modell über alle möglichen Parameter (dieses einen Modells) gemittelt. Zum anderen wird danach noch zusätzlich über die so entstandenen Mittelwerte mehrere Modellklassen gemittelt. Diese Herangehensweise erhöht die Robustheit der Pseudo-Label-Auswahl, da sie nicht von einem einzigen Modell abhängt. Durch die Berücksichtigung mehrerer Modelle kann das System auch in Fällen, in denen ein Modell stark abweicht, weiterhin korrekte Entscheidungen treffen. Dies ist besonders nützlich in komplexen und hochdimensionalen Datensätzen, bei denen die Modellwahl einen großen Einfluss auf die Ergebnisse haben kann. Die technischen Details sind in Rodemann u. a., 2023b, Kapitel 3.1 zu finden.

- **Fehlerakkumulation:** Die wiederholte Verwendung von pseudo-beschrifteten Daten als wahre Labels kann zu einer Akkumulation von Fehlern führen. Die vorgeschlagene Methode berücksichtigt nicht nur die vorhergesagten Labels, sondern auch alle hypothetischen Labels und weist diesen eine Gewichtung zu. Indem alle möglichen Labels berücksichtigt werden, kann das Modell eine umfassendere Sicht auf die Daten entwickeln und potenzielle Fehlerquellen besser identifizieren und minimieren. Dies reduziert die Gefahr, dass sich Fehler im Laufe des Trainings verstärken und die Gesamtgenauigkeit des Modells negativ beeinflussen. Die technischen Details sind in Rodemann u. a., 2023b, Kapitel 3.2 zu finden.

- **Covariate Shift:** Durch die Aufnahme selbst-gelabelter Daten verändert sich die Verteilung der Trainingsdaten. Zu beachten ist, dass die Aufnahme nicht zufällig, sondern durch das Selektionskriterium erfolgt. Die i.i.d.-Annahme kann also nicht aufrecht erhalten werden. Eine solche, selbst-induzierte Veränderungen in der Datenverteilung über die Zeit können die Unsicherheitsquantifizierung des Modells beeinträchtigen. Die vorgeschlagene robuste Erweiterung berücksichtigt diese Verschiebung. Die Idee ist, dass nur solche pseudo-gelabelte Daten ausgewählt werden, die im Lichte der verschobenen und der initialen, unverzerrten Verteilung plausibel erscheinen, also einen hohen Nutzenfunktionswert haben. Durch die Anpassungsfähigkeit an unterschiedliche Verteilungen kann das Modell seine Genauigkeit und Zuverlässigkeit über verschiedene Datensätze und Zeiträume hinweg aufrechterhalten. Diese Flexibilität ist besonders wichtig in dynamischen Umgebungen, in denen die Datenverteilung häufig wechselt und starre Modelle schnell an Leistungsfähigkeit verlieren können. Die technischen Details sind in Rodemann u. a., 2023b, Kapitel 3.3 zu finden.

- **Generische Robustifizierung:** Rodemann u. a., 2023b schlagen darüber hinaus vor, die generalisierte Bayesianische α-cut-Regel für Credal-Sets zu verwenden, wenn keine näheren Informationen über diese Unsicherheiten vorliegen, siehe auch Dietrich, Rodemann und Jansen, 2024. Die Idee ist, wieder eine Menge Bayesianischer Priori-Verteilungen zu spezifizieren und alle im Licht der Beobachtungen zu Posteriori-Verteilungen aufzudatieren. Dann wählt man die pseudo-gelabeleten Daten aus, die den höchsten erwarteten Nutzen (Likelihood) gehabt hätten,

wenn wir die Priori so festgelegt hätten, dass sie der (möglicherweise überanpassten) Likelihood des Modells am meisten widersprochen hätte. Diese generische robuste Erweiterungen zeigt im Vergleich zu traditionellen Methoden bessere Ergebnisse und erhöhen die Zuverlässigkeit und Leistung des Modells. Die technischen Details sind in Rodemann u. a., 2023b, Kapitel 4 sowie in Dietrich, Rodemann und Jansen, 2024 zu finden.

Die auch im Rahmen des Kooperationsprojekts entstandene Arbei *In All Likelihoods: Robust Selection of Pseudo-Labeled Data* (Rodemann u. a., 2023d) ist in den Proceedings of Machine Learning Research (PMLR) zur Konferenz *International Symposium on Imprecise Probabilities Theory and Application (ISIPTA)* erschienen. Die dazugehörige Software ist unter `https://github.com/rodemann/robust-pls` frei verfügbar. Neben einer ausführlichen Dokumentation enthält das Repository unter anderem eine Reihe illustrativer Anwendungsbeispiele sowie Code, um die Benchmarking-Experimente zu reproduzieren. Die generische Robustifizierung wurde in (Dietrich, Rodemann und Jansen, 2024) implementiert und ausführlich getestet. Die dazugehörige Implementierung ist unter `https://github.com/Stefan-Maximilian-Dietrich/reliable-pls` ebenfalls frei verfügbar.

4.3.3 Spotlight: Entscheidungstheoretische Präferenzsysteme und Algorithmenwahl

Abschließend soll noch über einige weitere Arbeiten aus den Eigenleistungen im Rahmen dieser Kooperation berichtet werden. Sie übertragen und erweitern aktuelle Entwicklungen aus der theoretischen Entscheidungstheorie auf die Wahl von Algorithmen. Vereinfacht gesprochen wird ein neues Konzept (mengenwertiger, verallgemeinerter) stochastischer Dominanz genutzt, um Algorithmen anhand von Benchmarkstudien, denen multidimensionale Gütekriterien und mehrere Datensätze zugrunde liegen, statistisch gesichert (partiell) zu ordnen.

Ausgangssituation: die dreifache Problematik beim Algorithmenvergleich

Typischerweise sind viele verschiedene Algorithmen prinzipiell für eine bestimmte Problemstellung geeignet. Eine wichtige Entscheidungshilfe bei der Frage, welche Algorithmen mit welcher Variante letztendlich in die nähere Wahl kommen, bieten Benchmarkstudien mit Datensätzen ähnlicher Struktur. Dort werden verschiedene Verfahren bezüglich verschiedener Kriterien anhand verschiedener Datensätze verglichen. Tabelle 2 zeigt den prinzipiellen Aufbau einer solchen Benchmarkstudie bei einem Klassifikationsproblem; dabei seien

- $\mathcal{C}$ eine Menge von Klassifikationsverfahren,
- $\mathcal{D}$ eine Menge von Datensätzen,
- $\phi_1, \ldots, \phi_n : \mathcal{C} \times \mathcal{D} \to \mathbb{R}$ verschiedene kardinale oder ordninale Kriterien, die Güte der Klassifkation zu beurteilen.

Eine allgemein überzeugende direkte Entscheidung für einen der Algorithmen aufgrund der Ergebnisse scheitert üblicherweise an der Mulitplizität der Vergleichsdimensionen (Kriterien und Datensätze). Typischerweise sind selbst bei einem festen Datensatz manche Algorithmen hinsichtlich eines Kriteriums besser aber schlechter bezüglich anderer; ferner ändert sich das Ergebnis des Vergleichs bei der Betrachtung eines anderen Datensatzes. Erschwerend kommt als dritter Problemkreis im Sinne von Demšar, 2006 hinzu, dass es sich ja bei den betrachteten Datensätzen um eine Zufallsauswahl aus einem potentiellen Universum von Datensätzen handelt, so dass sichergestellt sein muss, dass beobachtete Unterschiede auch tatsächlich signifikant sind.

data sets / classifier	D_1	$\dots$	D_s
C_1	$\begin{pmatrix} \phi_1(C_1, D_1) \\ \vdots \\ \phi_n(C_1, D_1) \end{pmatrix}$	$\dots$	$\begin{pmatrix} \phi_1(C_1, D_s) \\ \vdots \\ \phi_n(C_1, D_s) \end{pmatrix}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
C_q	$\begin{pmatrix} \phi_1(C_q, D_1) \\ \vdots \\ \phi_n(C_q, D_1) \end{pmatrix}$	$\dots$	$\begin{pmatrix} \phi_1(C_q, D_s) \\ \vdots \\ \phi_n(C_q, D_s) \end{pmatrix}$

Tabelle 2: Schematischer Aufbau einer Benchmarkstudie, aus Jansen u. a., 2023a, p. 3

Gängige Methoden, mit der Multiplizität umzugehen, bewegen sich zwischen zwei Extrempolen,
a) der radikalen Informationsreduktion bis zur eindeutigen finalen Entscheidung für eine lineare Ordnung auf den Algorithmen
b) der vorsichtigen Verarbeitung der Situation aus der Perspektive der Pareto-Optimalität
Vorgehensweise a) versucht durch verschiedene Methoden der Indexbildung (Weiterverabeitung von Rangsstatistiken, gewichtete Mittelung der Werte oder Aggregation paarweiser Vergleiche) eine lineare Ordnung unter den Algorithmen zu produzieren. Eine solche Vorgehensweise unterliegt zwangsläufig einer Willkür; die Existenz einer allgemein akzeptierten Präferenzaggretation in diesem Kontext stünde auch im Widerspruch zu Unmöglichkeitsresultaten der Präferenzaggregations-Theorie (vgl. Jansen u. a., 2023a). Vorgehensweise b) ist extrem zurückhaltend und wenig diskriminierend; es werden nur Algorithmen ausgeschlossen, die von einem anderen in jeder Hinsicht überboten werden.

Präferenzsysteme

Zur Motivation unseres Vorgehens ist ein Blick in die Entscheidungstheorie hilfreich. Dort werden seit den 1960iger Jahren mengenwertige Zustandswahrscheinlichkeiten als bedeutendes Paradigma gesehen, um mit komplexer Unsicherheit über die Umwelt umzugehen. Potenziert werden die Möglichkeiten konzeptkonformer Anwendungen, wenn zudem mit mengenwertigen Nutzenfunktionen gearbeitet wird, um komplexe Präferenzstrukturen über der Konsequenzenmenge, und letztendlich der Aktionenmenge, ausdrücken zu können. Fundamental für das folgende sind Präferenzsysteme im Sinne von Jansen, Schollmeyer und Augustin, 2018. Sie ermöglichen es, ordinale und metrische Präferenzinformation strukturtreu aufzugreifen, durch mengenwertige Nutzenfunktionen zu repräsentieren und so auch mit Situationen mit variierendem Skalenniveau geeignet umzugehen (Jansen u. a., 2023b). Betrachtet werden partielle Ordnungen, die auch Unvergleichbarkeit zulassen. Die Ordinalität findet sich in partiellen Ordnungen der Form "a wird gegenüber b bevorzugt", wieder; fürdie Kardinalität werden partielle Ordnungen der in Relation stehen Paare elizitiert ("Die Präferenz von a gegenüber b ist stärker als die Präferenz von c gegenüber d.") Mit diesen allgemein skalierten Zufallsvariablen lässt sich ein neues leistungsfähiges Konzept Verallgemeinerter Stochastischer Dominanz aufbauen.

Algorithmenwahl mit Verallgemeinerter Stochastischer Dominanz

In Jansen u. a., 2023a werden diese Konzepte im Kontext der Algorithmenwahl weiterentwickelt. Entscheidender Gedanke ist, dass über die multiplen – teils ordinalen, teils kardinalen – Gütekriterien ein aussagekräftiges Präferenzsystem aufgebaut werden kann. Die beobachteten Gütewerte eines Algorithmus werden als Realisationen eines präfenzsystemwertigen Zufallselements gesehen, so dass die Frage des Algorithmenvergleichs über die Verallgemeinerte Stochastische Dominanz beantwortet werden kann. In diesem Zusammenhang ist es wichtig, dass auch geeignete Permutationstests entwickelt werden konnten, die Auskunft geben, ob auf den konkret beobachteten Datensätze festgestellte Unterschiede auch statistisch signifikant sind.

Das Konzept der Verallgemeinerten Stochastischen Dominanz erlaubt also die Konstruktion einer statistisch wie messtheoretisch fundierten partiellen Ordnung der Algorithmen, die die Performance in den Benchmarkuntersuchungen strukturtreu aussagekräftig abbilden.

Jansen u. a., 2024 entwickeln weitergehend das Konzept einer GSD-Front (GSD für Generalized Stochastic Dominanz) als einer informationseffizienten Erweiterung der Pareto-Front, wobei es bei den aufbauenden statistischen Test- und Schätzverfahren wiederum gelingt, der statistischen Unsicherheit geeignet Rechnung zu tragen. Die Tests lassen sich zudem gegenüber der Annahme, dass die Datensätze über einen reinen i.i.d.-Prozess ausgewählt wurden, robustifizieren.

Softwate und Dokumentation: Die dieses Teilgebiet leitenden Arbeiten sind frei zugänglich und in renomierten Publikationsorganen (Journal of Machine Learning Research: Jansen u. a., 2023a und Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence: Jansen u. a., 2023b) publiziert bzw. aktuell in Begutachtung (Jansen u. a., 2024). Die aktuellsten Implementierungen aller entwickelten Methoden sowie Skripte zur Reproduktion von Experimenten sind unter `https://anonymous.4open.science/r/Statistical-Multicriteria-Benchmarking-via-the-GSD-Front-3BF3/README.md` beziehungsweise unter `https://github.com/hannahblo/Robust_GSD_Tests` frei verfügbar.

5 Machine Learning Operations und Reproduzierbarkeit in der Amtlichen Statistik

Das statistische Bundesamt (Destatis) setzt zunehmend auf den Einsatz von Machine Learning (ML) und etliche Methoden haben sich für eine Vielzahl von Anwendungen in Bereichen wie Nowcasting und Natural Language Processing als nützlich für die amtliche Statistik erwiesen. Durch die zunehmende Wichtigkeit von ML in Arbeitsabläufen bei Destatis und die natürlich gewachsenen Strukturen arbeiten viele Teams parallel an ML-Anwendungen und der Aufwand für Entwicklung, Einsatz und Wartung der Modelle nimmt zu. Weil entsprechende Tools und Prozesse nicht eingesetzt werden, ist darüber hinaus die Reproduzierbarkeit von wichtigen Ergebnissen nur mit Einschränkungen gegeben, was nicht den Standards der amtlichen Statistik entspricht. Aus diesem Grund wurden im Rahmen des Projektes der aktuelle Stand innerhalb von Destatis - speziell des Referats "Künstliche Intelligenz, Big Data" - bei den Themen ML und Machine Learning Operations (MLOps) untersucht, Anforderungen erhoben und anschließend ein Lösungsvorschlag für eine MLOps-Architektur erarbeitet. Als Projektergebnis dieses Arbeitspakets wurde ein ausführlicher Abschlussbericht erstellt (siehe Karl, Kaminwar und Frechen, 2024), der von Steffen Moritz[8] bereitgestellt werden kann.

5.1 Aktuelle Situation und Anforderungsanalyse

Um die derzeitige Situation und bestehenden Anforderungen an MLOps besser zu verstehen, wurden im Rahmen des Projekts diverse Interviews mit relevanten Expert:innen und Anwender:innen durchgeführt. Schwerpunkte in diesen Gesprächen waren die ML-Anwendungen innerhalb von Destatis, bestehende sowie geplante Infrastruktur und Tools, Arbeitsabläufe über den gesamten ML-Lebenszyklus und Sicherheit. Aus diesen Informationen abgeleitet wurde eine MLOps-Reife bestimmt. Aufbauend auf einer Stakeholderanalyse, in der vor allem die drei Rollen Data Scientist, Fachstatistiker:in und IT-Verantwortliche:r identifiziert wurden, wurden in Abstimmung mit Destatis Anforderungen, die diese Schlüsselrollen an eine MLOps-Architektur haben, erhoben. Die identifizierten Pflichtanforderungen sind in einer Tabelle des Abschlussberichts (siehe Karl, Kaminwar und Frechen, 2024) übersichtlich aufgelistet. Zusätzlich zu den Anforderungen wurden Nebenbedingungen identifiziert, nach denen sich eine empfohlene Architektur richten muss. Die wichtigsten Nebenbedingungen sind Kompatibilität mit Python und R (technische Nebenbedingung), eine Möglichkeit die Architektur auf eigenen Servern zu verwalten (organisatorische Nebenbedingung), Open-Source Status (organisatorische Nebenbedingung) und Umgang mit Text-, Zeitreihen- und tabellarischen Daten (Nebenbedingung bezüglich der Daten). Die ausführlichen Nebenbedingungen finden sich ebenfalls in einer Tabelle des Abschlussberichts (siehe Karl, Kaminwar und Frechen, 2024) wieder.

5.2 Lösungsstrategie

Als Lösung wird die in Abbildung 7 visualisierte MLOps-Architektur empfohlen.

Empfohlene Architektur Die Architektur setzt auf eine einheitliche Datenspeicherung durch das relativ schlanke und durch seinen objektorientierten Speicher für ML-Anwendungen geeignete Tool *minIO*. Durch die Einführung von Datenversionskontrolle durch das Tool *DVC* wird dabei aufbauend

[8] steffen.moritz@destatis.de

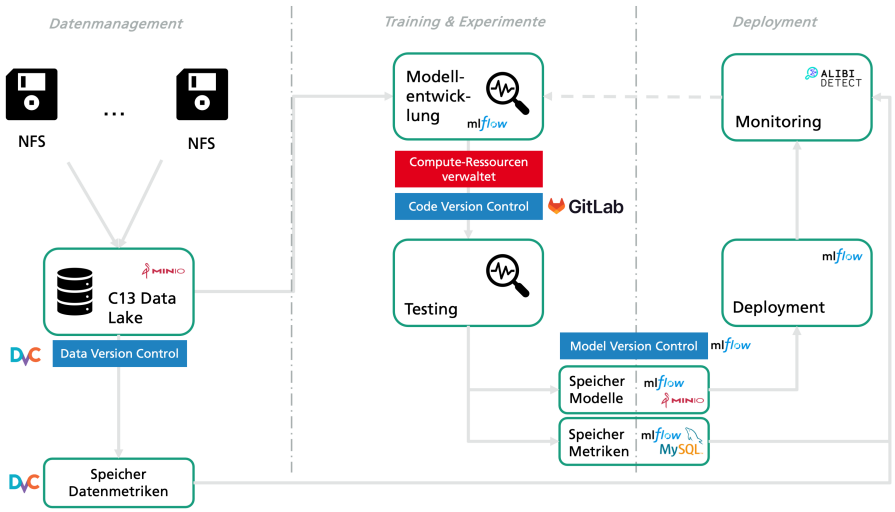

Abbildung 7: Empfohlene MLOps-Architektur.

auf der Datenspeicherung eine erste Voraussetzung für Reproduzierbarkeit von Experimenten und Ergebnissen geschaffen. *DVC* ist eine der führenden Open-Source Lösungen für Datenversionskontrolle, ist geeignet für alle in der amtlichen Statistik relevanten Datentypen und setzt auf eine einfache Handhabung. Gleichzeitig werden mit verschiedenen Versionen von Datensätzen auch Metadaten und Datenmetriken gespeichert, die Reproduzierbarkeit und Transparenz begünstigen und darüber hinaus hilfreich in späteren Schritten des ML-Lebenszyklus sein können (z.B. Überwachung von Modellen im Produktivbetrieb).

Die Experimente und Modellentwicklung werden in der vorgeschlagenen Architektur maßgeblich durch *MLflow* unterstützt. Experiment-Tracking, also das Festhalten von Ergebnissen und zugehörigen Konfigurationen im Sinne besserer Zusammenarbeit sowie Reproduzierbarkeit, ist eine zentrale Komponente der Architektur. *MLflow* liefert die in den Anforderungen als notwendig für die amtliche Statistik identifizierte Funktionalität in diesem Bereich, bietet im Vergleich zu einigen Alternativen außerdem eine umfangreiche Integration mit Python und R und schafft durch seinen Status Flexibilität für die Zukunft, da es in vielen Frameworks und Cloud-Lösungen Beachtung findet. Aus diesem Grund wird *MLflow* in der empfohlenen Architektur auch für das Deployment eingesetzt. Für das Deployment gibt es durchaus Open-Source Lösungen mit vergleichbarer Funktionalität, die ebenfalls viele Anforderungen erfüllen, sie sind aber oft auf bestimmte ML-Frameworks und meistens für die Programmiersprache Python optimiert. Dass *MLflow* drei Lösungskomponenten - neben Experiment-Tracking und Deployment auch den Modellspeicher - abdeckt, macht es für diese Architekturempfehlung besonders wertvoll, da so kaum Schnittstellen entstehen und die Wartung deutlich vereinfacht wird. Sowohl *MLflow* als auch *minIO* bieten die Möglichkeit der Zugriffsverwaltung, so dass der Zugriff auf Daten, Modelle und Ergebnisse den Anforderungen gemäß beschränkt werden kann.

Eine letzte Komponente, das Monitoring von ML-Modellen im Produktivbetrieb, wird durch *Alibi Detect* abgedeckt. Ein hoher Grad an Automatisierung des Monitorings ist für die amtliche

Statistik aufgrund der Datensituation - neue Daten werden meist erst nach Monaten erhoben - nicht zielführend. *Alibi Detect* begünstigt die einfache Erstellung von Skripten, die Modelle im Produktivbetrieb überwachen, und empfiehlt sich besonders durch die Möglichkeit individuelle Tests zu integrieren. Es soll an dieser Stelle darauf hingewiesen werden, dass *Alibi Detect* von den empfohlenen Tools am wenigsten "in die Architektur integriert ist". Es bedient lediglich die Komponente der Modellüberwachung und sollten sich die Anforderungen an diese Komponente in Zukunft ändern (z.B. hinsichtlich Nutzerfreundlichkeit für Fachstatistiker:innen), kann ein neues Tool verhältnismäßig einfach integriert werden. Eine durchaus realistische Alternative ist die Entscheidung für ein Monitoring-Tool je nach Anforderungen auf Projektebene. Ein Vergleich der MLOps-Reife nach Einführung der Architektur im Vergleich zur aktuellen MLOps-Reife ist in Tabelle 3 visualisiert.

	Level 0: Kein MLOps, kein DevOps	Level 1: DevOps, aber kein MLOps	Level 2: Automat. Training	Level 3 Automat. Deployment von Modellen	Level 4: Automat., umfängliches MLOps
Dokumentation & Standardisierung		X	X		
Zusammenarbeit	X		X		
Reproduzierbarkeit		X			X
Expperimente	X			X	
Monitoring			X	X	
Deployment		X	X		
Integration von Applikationen	X		X		

Tabelle 3: Aktuelle MLOps Reife innerhalb von Destatis (schwarz) gegenüber der MLOps-Reife mit der empfohlenen Architektur (rot). Eine Erläuterung der Stufen findet sich im Anhang von Karl, Kaminwar und Frechen, 2024.

Erweiterung der empfohlenen Architektur Die empfohlene Architektur - erarbeitet um die Pflichtanforderungen zu erfüllen - erfüllt bereits einen Großteil der wünschenswerten Anforderungen; eine umfassende Erweiterung ist deswegen wenig sinnvoll. Als mögliche Ergänzungen werden Datenvalidierung sowie eine Automatisierung des Deployments empfohlen und in einem entsprechenden Kapitel des Abschlussberichts (siehe Karl, Kaminwar und Frechen, 2024) ausgeführt.

Empfehlung für einen ersten Schritt Falls eine Einführung der empfohlenen Architektur schrittweise erfolgen soll, bietet sich als erster Schritt eine Integration von *MLflow* in die Arbeitsabläufe von Destatis an. Experiment-Tracking und ein Modellspeicher sind essentiell für die notwendige Reproduzierbarkeit und können bereits von einzelnen Personen oder Arbeitsgruppen

produktiv genutzt werden. Außerdem kann *MLflow* durchaus sinnvoll eingesetzt werden, wenn nicht die gesamte Funktionalität genutzt wird: So kann z.B. Experiment-Tracking durchaus unabhängig von Deployment und dem Modellspeicher in der Phase der Modellentwicklung verwendet werden. Nicht zuletzt schafft *MLflow* Zukunftssicherheit durch seine Verbreitung und Flexibilität und könnte im Falle eines anschließenden Umstiegs auf eine Cloud-Lösung wahrscheinlich weiterhin genutzt werden. Eine Erweiterung auf die empfohlene Architektur ist anschließend dann schrittweise, z.B. durch die Einführung von *DVC*, möglich.

5.3 Funktionsabgleich mit Cloudera

Cloudera wird derzeit intern für einen Einsatz in den Bereichen ML und Data Science diskutiert und könnte auch für das Referats "Künstliche Intelligenz, Big Data" von Interesse sein. Insbesondere zeigten die durchgeführten Interviews, dass Cloudera innerhalb der amtlichen Statistik bereits erfolgreich zum Einsatz kommt. Ein auf der Anforderungsanalyse aufbauender Funktionsabgleich der empfohlenen MLOps-Architektur mit *Cloudera* belegt, dass auch die Cloud-Lösung einen Großteil der relevanten Funktionalität bietet. Abstriche müssen dabei vor allem im Bereich der Datenversionskontrolle und eingeschränkter Funktionalität in Bezug auf die Programmiersprache R (Experiment-Tracking, Modellspeicher) gemacht werden. Ein Vorteil von *Cloudera* sind die - je nach Version - integrierten und nutzerfreundlichen Module zur Visualisierung und Datenanalyse. Eine ausführliche Abhandlung zu *Cloudera* findet sich im Abschlussbericht (siehe Karl, Kaminwar und Frechen, 2024).

5.4 MLOps Best Practices

Etliche der erhobenen Anforderungen setzen sich nicht mit Technik und Funktionalität auseinander, sondern betreffen eher interne Arbeitsstandards. Sie richten sich vor allem an Programmierungs-Richtlinien, Dokumentation und Weiterbildung. Um auch diese Anforderungen zu berücksichtigen, werden zusätzliche generelle ML(Ops) Best Practices empfohlen, die in die Arbeit bei Destatis integriert werden können. Es handelt sich dabei um Best Practices wie z.B. die Verwendung von Vorlagen für ML-Projekte (Programmierungs-Richtlinien), Dokumentation von Ergebnissen oder Datensätzen und ein einheitliches Wissensmanagement (Weiterbildung). Eine ausführliche Beschreibung ist im Abschlussbericht (siehe Karl, Kaminwar und Frechen, 2024) zu finden.

Als Ergänzung zum Abschlussbericht wurde ein Demonstrator entwickelt, der die Architektur anhand eines beispielhaften ML-Projekts greifbarer machen soll. Dafür wurde ein *Git*-Repository mit umfangreicher Dokumentation erstellt, das die Installation aller verwendeten Tools ermöglicht, und mit dem sich exemplarisch einige typische Arbeitsabläufe durchführen lassen. Der Zugang zum Demonstrator sowie eine Vorstellung in Form eines Videos können ebenfalls von Steffen Moritz[9] bereitgestellt werden.

[9] steffen.moritz@destatis.de

6 Fairness und Bias Auditing

6.1 Einführung

Das Arbeitspaket zu "Fairness in ML" beschäftigt sich mit der Verknüpfung der Konzeption, Identifikation und Quantifizierung von (Un)fairness in Machine Learning (ML) in Anwendungen im Wechselspiel mit der spezifischen Nutzung von ML in der amtlichen Statistik. Ausgangspunkt war die Beobachtung, dass die zunehmende Nutzung von Verfahren des maschinellen Lernens (insbesondere) im öffentlichen Sektor nicht nur mit technischen, sondern auch sozial-gesellschaftlichen Herausforderungen einhergeht. Diese Perspektive speist sich insbesondere aus der Nutzung von ML als zentralem Bestandteil von automatisierten Entscheidungssystemen (automated decision-making; ADM), welche das Risiko tragen, gleichsam auch historische Diskriminierung gegenüber sozialen Gruppen zu automatisieren. Vor diesem Hintergrund war eine zentrale Aufgabe des Arbeitspakets die Zusammenführung von Perspektiven aus der "fair ML" Literatur und deren Übersetzung in den Kontext der amtlichen Statistik. Dies erfolgte zentral durch die Verknüpfung mit und Ergänzung des "Quality Framework for Statistical Algorithms" (QF4SA; Yung u. a. 2022) um Fairnessaspekte in Schenk und Kern, 2024 (siehe auch Unterabschnitt 6.3.1). Weiterführende aktuelle Arbeiten des Arbeitspakets beschäftigen sich mit dem "bias auditing" von Trainingsdaten, bei dem mithilfe von Metriken aus der Surveyforschung (Schouten, Cobben und Bethlehem, 2009) vorausschauende Indikatoren für Fairnessprobleme *vor* dem eigentlichen Modelltraining eingeführt werden.

Die gestiegene Bedeutsamkeit von Fairness im Kontext von ML-basierten Vorhersagen und ADM ergibt sich aus unserer Sicht insbesondere aus zwei Aspekten: Verschiebungen in der Entscheidungsverantwortung und in der immensen technischen Skalierbarkeit. 1) Da Modellvorhersagen bei der Entscheidungsfindung in vielen Bereichen immer wichtiger werden, verlieren menschliche Entscheider dort zunehmend ihre Ermessensfreiheit bei der Entscheidungsfindung. Dies ist insofern von Bedeutung, als dass man von anderen Menschen (z.B. von Sachbearbeitern in Jobcentern) erwartet, dass sie bei ihren Entscheidungen auf Gerechtigkeit und Fairness (Kuppler u. a., 2022) Rücksicht nehmen (Bogdan u. a., 2023). Wenn der Mensch zunehmend in der Entscheidungsfindung auf Vorhersagen zurückgreift oder Prozesse gänzlich automatisiert werden, stellt sich die zentrale Frage, wie Fairnessaspekte in algorithmisch-gestützten Entscheidungen berücksichtigt werden können. 2) Der Einsatz eines einzigen ML-Modells kann großflächig Vorhersagen liefern, die vormals von einer Vielzahl von menschlichen Entscheidern getroffen wurden. Diese Skalierung erhebt Bedeutung und Auswirkungen der Fairness dieses einzelnen ML-Modells weit über die Bedeutung der Fairness im Entscheidungsverhalten von einer dieser vielen Personen. Gleichsam lohnt sich die Investition von Ressourcen zum Trainieren eines Modells insbesondere für solche groß angelegten Einsätze. Organisationen und öffentliche Behörden, die automatisierte Entscheidungen in zentralen Kontexten einsetzen, wie etwa die Arbeitsagentur eines Landes, können dabei auf vielfältige Daten zum Modelltraining zurückgreifen, unter anderem auch auf Datenprodukte der amtlichen Statistik. Beide Entwicklungen in Kombination unterstreichen die Relevanz von Fairnessaspekten im Kontext der Datenproduktion amtlicher Statistik.

Vor diesem Hintergrund stellen wir in Abschnitt 6.2 eine kurze Zusammenfassung der Forschung zu Fairness in Machine Learning vor (Unterabschnitt 6.2.1), in Verknüpfung mit ML Anwendungen in der amtlichen Statistik (Unterabschnitt 6.2.2). Abschnitt 6.3 diskutiert die Verknüpfung des QF4SA mit Fairnessaspekten (Unterabschnitt 6.3.1) und argumentiert darüber hinausgehend für Fairness als eine eigene Qualitätsdimension (Unterabschnitt 6.3.2). Wir schließen mit einer Zusammenfassung und einem Ausblick auf aktuelle und zukünftige Forschung (Abschnitt 6.4).

6.2 Faire Vorhersagen und algorithmische Entscheidungssysteme

Nach vieldiskutierten Anwendungen von Machine Learning in verschiedenen sensiblen Bereichen (z.B. Justiz, Gesichtserkennung, Profiling von Arbeitssuchenden, siehe Angwin, Mattu und Kirchner 2016; Buolamwini und Gebru 2018; Allhutter u. a. 2020), haben Arbeiten zu "Fairness in Machine Learning" ein vielschichtiges und multidisziplinäres Forschungsfeld initiiert, welches sich primär mit den sozialen Auswirkungen algorithmischer Entscheidungsfindung (ADM) befasst. Die Forschung zu Fairness in Machine Learning ("fair ML"; siehe Mehrabi u. a. 2021; Mitchell u. a. 2021; Makhlouf, Zhioua und Palamidessi 2021; Caton und Haas 2024) konzentriert sich in der Regel auf Vorhersagemodelle als Teil größerer soziotechnischer Systeme, die den Zugang zu Positionen regeln oder Ressourcen zuweisen. Der Anwendungsbereich von fair ML geht jedoch über solche Kontexte hinaus und umfasst auch die Auswirkungen des Einsatzes von ML in anderen Zusammenhängen (Rodolfa, Saleiro und Ghani, 2020). Dies betrifft aus Sicht dieses Arbeitspakets insbesondere auch Prozesse der Datenproduktion in der amtlichen Statistik, auf die sich fair ML Konzepte anwenden und erweitern lassen.

6.2.1 Fairnesskonzepte und -metriken

Ein zentrales Konzept in der fair ML Literatur ist der Begriff der *geschützten Attribute*. Geschützte Attribute sind inhärente oder zugeschriebene Eigenschaften von Personen (wie ethnische Herkunft, Geschlecht, Alter oder Religion), für die sie nicht verantwortlich gemacht werden können (oder sollten), die aber dennoch aufgrund von Vorurteilen und Diskriminierung die Grundlage für eine unterschiedliche Behandlung von Personen in der realen Welt sein können. Im engeren Sinne können geschützte Attribute auf der Grundlage von Antidiskriminierungsgesetzen definiert werden (Simson, Fabris und Kern, 2024; Mehrabi u. a., 2021), aber die endgültige Auswahl von Attributen, der in einer bestimmten Anwendung berücksichtigt werden sollten, ist in der Regel kontextspezifisch.

Ein übergreifendes Verständnis von Fairness in Machine Learning folgt dem Prinzip, Ergebnisse oder Praktiken zu verhindern, die negative Auswirkungen auf Mitglieder geschützter Gruppen haben (Barocas und Selbst, 2016). Es gibt verschiedene Wege, über welche dieses Prinzip in der Praxis des maschinellen Lernens verletzt werden kann – der zentralste Punkt ist (verschiedene Arten von) *Bias in Daten* (Mehrabi u. a., 2021). Historische Biases können in jeglichen Daten vorhanden sein, die aus sozialen Prozessen resultieren: Administrative Arbeitsmarktdaten erfassen Diskriminierungsprozesse auf dem Arbeitsmarkt mit, so wie Bildungshistorien auch herkunftsspezifisch unterschiedliche Chancen im Bildungssystem widerspiegeln. Historische Biases können leicht von ML-Modellen gelernt und übernommen werden, wenn Daten, die soziale Prozesse widerspiegeln, für das Modelltraining verwendet werden. Das Modelltraining kann jedoch auch durch Messfehler beeinflusst werden. So kann etwa die in den Daten beobachtete abhängige Variable ein verzerrter Proxy für das tatsächliche Outcome sein, so dass sich soziale Biases im Schritt der Modellspezifikation in das Modell einschleichen (Obermeyer u. a., 2019). Schließlich bezieht sich Repräsentationsbias auf Verzerrungen in der Zusammensetzung der Trainingsdaten. Solche Defizite können sich auf die (unzureichende) Repräsentation bestimmter sozialer Subgruppen beziehen oder auf die Übereinstimmung zwischen den für das Modelltraining verfügbaren Daten und der letztendlichen Zielpopulation. Neben diesen Datenproblemen kann es sogenannte "feedback loops" geben: Wenn (verzerrte) Vorhersagen reale Outcomes beeinflussen, können diese wiederum Verzerrungen in den zeitlich folgenden Trainingsdaten herbeiführen oder verstärken und so Verzerrungen in den Vorhersagen verstetigen (Perdomo u. a., 2020).

Vor dem Hintergrund der verschiedenen Arten von Datenbiases wurde in der fair ML Literatur eine Reihe von *Fairness-Konzepten* vorgeschlagen, die verschiedene Vorstellungen von Fairness

formalisieren und oft entsprechende *Fairness-Metriken* implizieren. Fairness-Konzepte fokussieren typischerweise auf binäre Klassifikationsprobleme und werden auf Gruppen-, Subgruppen- oder individual-Ebene formuliert. Konzepte der "Group fairness" vergleichen Mitglieder geschützter und nicht-geschützter Gruppen hinsichtlich verschiedener vorhersagebasierter Maße. Bei einem gegebenen geschützten Attribut A und einer Vorhersage $\hat{Y}$ fordern unabhängigkeitsbasierte Fairness-Konzepte, dass die Vorhersagen unabhängig von der Gruppenmitgliedschaft sein sollen: $\hat{Y} \perp A$. Das "separation" Konzept berücksichtigt zusätzlich das beobachtete Ergebnis Y und fordert Unabhängigkeit gegeben dem wahren Outcome: $\hat{Y} \perp A \mid Y$. Das Konzept der "sufficiency" hingegen verlangt $Y \perp A \mid \hat{Y}$ (Barocas, Hardt und Narayanan, 2023; Makhlouf, Zhioua und Palamidessi, 2021). Wenngleich gruppenbasierte Fairness-Metriken in der Praxis einfach zu berechnen sind, ist ein zentrales Ergebnis der fair ML Literatur, dass sich die dahinterstehenden Konzepte in der Regel ausschließen: Abgesehen von Spezialfällen kann ein Vorhersagemodell nicht gleichzeitig Unabhängigkeit, "separation" und "sufficiency" erfüllen (Chouldechova, 2016).

Neben Konzepten auf Gruppenebene zielt Subgruppenfairness darauf ab, stärkere Fairnessgarantien zu erreichen, indem Fairnessanforderungen für komplexe Sets von Subgruppen formuliert werden, die sowohl über geschützte als auch nicht-geschützte Attribute definiert werden können (Hebert-Johnson u. a., 2018; Kim, Ghorbani und Zou, 2019; Kearns u. a., 2018). Schließlich formulieren Konzepte der individuellen Fairness Anforderungen auf der Individualebene, z.B. in Hinblick auf Distanzen zwischen Individuen im Verhältnis zu Distanzen in den Vorhersagen (konkret: ähnliche Individuen sollen ähnliche Vorhersagen erhalten; Dwork u. a. 2012) oder unter Einbeziehung von kausalen Argumenten (Kilbertus u. a., 2017).

6.2.2 Verknüpfung zu Machine Learning in der amtlichen Statistik

Die klassischen Anwendungsfälle, die in der fair ML Literatur diskutiert werden, beinhalten Machine Learning Modelle als Teil eines automatisierten Entscheidungssystems (ADM), in dem die ML-basierte Vorhersage hilft, Personen gemäß ihres (statistischen) Risikos (unter nicht-Behandlung) zu priorisieren. Wenngleich die amtliche Statistik selbst nicht unmittelbar solche datengesteuerten Entscheidungen durchführt, ist der Blick auf algorithmische Entscheidungsfindung aus zwei Gründen relevant. 1) Die Datenprodukte der amtlichen Statistik können, selbst oder ggf. nach Verknüpfung mit anwenderseitig vorhandenen Daten, von nachgelagerten Nutzern verwendet werden, um Modelle für ADM Anwendungen zu trainieren. Da solche Modelle nur so gut sind wie die Daten, mit denen sie trainiert werden – und die Dokumentation, die für diese Daten verfügbar ist –, spielt die amtliche Statistik eine entscheidende (vorgelagerte) Rolle bei der Vermeidung von Fairnessproblemen. 2) Entscheidungsfindung und insbesondere ADM sind von zentraler Bedeutung in der fair ML Literatur, und daher ist dieser Hintergrund notwendig, um die spezifische Perspektive der Literatur und der dort vorgeschlagenen Fairnesskonzepte und -techniken zu verstehen.

Bei der Analyse von ADM Systemen ist es hilfreich, zunächst zwischen dem Vorhersageschritt (Training des ML-Modells) und dem Entscheidungsschritt (Verteilung einer Resource auf Basis der Vorhersage) zu unterscheiden (Kuppler u. a., 2022; Scantamburlo, Baumann und Heitz, 2024). Ein faires prädiktives ML-Modell (Kuppler u. a., 2022) oder ein faires Gesamtsystem (d. h. beide Schritte zusammen; Scantamburlo, Baumann und Heitz 2024) tragen in diesem Kontext zu gerechten Entscheidungen bei.

In der Literatur wird der Vorhersageschritt häufig aus der Perspektive eines Datenanalysten betrachtet, der Daten erhält, die er dann, möglicherweise nach einer Datenaufbereitung, in das Modelltraining einspeist. Wir schlagen vor, diesen Blick zu erweitern: Alles vor dem Entschei-

dungsschritt sollte als Teil des Vorhersageschritts betrachtet werden. Somit tragen nicht nur das letztendliche Modelltraining, sondern alle Schritte entlang der Datenverarbeitung zur (Un-)Fairness des endgültigen Systems bei, vom Design der Datenerhebung über ihre Durchführung bis hin zu allen Schritten, die während der Verarbeitung an den Daten vorgenommen werden. Insbesondere in der amtlichen Statistik ist dies ein mehrstufiger Prozess, wie er z. B. im Rahmen von Total Survey Error (TSE)-Frameworks (siehe Groves u. a. 2009; Amaya, Biemer und Kinyon 2020; West u. a. 2023) beschrieben wird. Entlang dieser Schritte können sich Fairnessfehler summieren, sodass es wichtig ist, die kumulativen Auswirkungen auf Fairness entlang der Datenverarbeitungskette zu berücksichtigen. Zudem können Fairnessfehler mehr als nur additiv sein: Entscheidungen, die z.B. im Rahmen der Datenerhebung getroffen werden können, somit eine bedeutsame Hebelwirkung entwickeln. Jeder Schritt in der Erstellung eines Datenprodukts liefert Input für den nächsten Schritt. Obwohl die amtliche Statistik möglicherweise nicht an den allerletzten Schritten (insb. dem Entscheidungsschritt) eines ADM Systems beteiligt ist, liefert sie entscheidenden Input, wodurch die Fairness ihrer Datenprodukte (und Datenproduktion) von entscheidender Bedeutung ist.

6.3 Fairness als ein Qualitätskriterium für ML in der amtlichen Statistik

Vor dem Hintergrund der Bedeutsamkeit von Fairnessaspekten für die Anwendung von Machine Learning auch in der amtlichen Statistik stellt sich die Frage nach deren systematischen Einbeziehung in bestehende Prozesse der Datenproduktion. In Schenk und Kern, 2024 schlagen wir eine Erweiterung des "Quality Framework for Statistical Algorithms" (QF4SA; Yung u. a. 2022), mit welcher Fairnessüberlegungen in bestehende Qualitätsrichtlinien eingebettet werden können. Wir argumentieren, dass die Abbildung von Fairnessüberlegungen auf das QF4SA die aktuellen Qualitätsdimensionen bereichert, bestehende Anforderungen schärft und damit eine umfassende Evaluation potenzieller sozialer Biases von ML-Modellen in der amtlichen Statistik ermöglicht. Eine Neubewertung von Qualitätskonzepten aus der Fairnessperspektive kann zudem auf blinde Flecken hinweisen und zusätzliche Kriterien einführen, deren Schwerpunkt auf den verschiedenen nachgelagerten Verwendungen von Datenprodukten der amtlichen Statistik liegt.

6.3.1 Fairness in Interaktion mit bestehenden Kriterien

Interpretierbarkeit (interpretability) Interpretierbarkeit (siehe auch Kapitel 2) und Fairness können als eng miteinander verflochtene Prozesse in der gesamten ML-Pipeline betrachtet werden (siehe auch S. 14). In der *Modellentwicklung* können Methoden zur Interpretation von ML-Modellen (IML) helfen, zu verstehen, ob und wie ein Modell gesellschaftliche Vorurteile gelernt hat. Zu diesem Zweck kann als erster Schritte die Rolle und Bedeutung von geschützten oder sensiblen Attributen untersucht werden, um u.a. zu verstehen, ob "legitime" Merkmale für soziale Gruppen auf unterschiedliche Weise verwendet werden. In der Praxis kann die Zusammenführung von Modellinterpretation und Fairness über die Verwendung geschützter Attribute als Gruppierungsvariablen in der Anwendung von IML-Techniken umgesetzt werden. In der *Modellimplementierung* kann der (wahrgenommene) Grad der Interpretierbarkeit eines Modells die Fairnesswahrnehmung der späteren Nutzer des Algorithmus und das Vertrauen in die Ergebnisse des Modells beeinflussen. Wenn IML-Methoden im Kontext von Modellentwicklung oder -implementierung angewendet werden, ist ein weiterer Aspekt das Ausmaß, in dem die Genauigkeit der IML-Methoden ("fidelity") zwischen Gruppen variiert: Wenn die Erklärungen die Entscheidungen der Modelle nicht über den gesamten Merkmalsraum hinweg gleichermaßen korrekt widerspiegeln können, können Schlussfolgerungen,

die über die Funktionsweise der Modelle gezogen werden, in den verschiedenen Untergruppen unterschiedlich genau bzw. korrekt sein.

Genauigkeit (accuracy) Fairness interagiert mit (Vorhersage-)Genauigkeit in mehreren Aspekten des Datenproduktionsprozesses. Die *Aggregation von Daten* ist eine der Kernaufgaben der amtlichen Statistik: sowohl im Hinblick auf die Datenanalyse, bei der es sich um deskriptive Statistiken handeln kann, als auch im Hinblick auf die Erstellung (aggregierter) Daten, die dann öffentlich zur Verfügung gestellt werden. Machine Learning kann nützlich sein, wenn ein Parameter von Interesse nicht in allen Subpopulationen identisch ist. Insbesondere wenn viele Subpopulationen untersucht werden, wie dies bei intersektionaler (d.h. Subgruppen-)Fairness der Fall ist, kann ML helfen, Heterogenität zu entdecken. Wenn interpretierbare Heterogenität das Ziel ist, können decision trees für univariate Statistiken oder für komplexere Analysen causal trees (Athey und Imbens, 2016) geeignet sein, um einen Parameter und dessen Heterogenität akkurat abzubilden. Wir schlagen daher vor, solche Methodiken zu verwenden, um die Ergebnisse der Datenanalyse fairer zu berichten: Subgruppen, bei denen der Parameter mehr als einen vorab festgelegten Wert oder Anteil vom globalen Durchschnitt abweicht, sollten identifiziert und zusammen mit dem globalen Durchschnitt berichtet werden. Für fehlerbasierte Fairnesskonzepte kann die selbe Methodik verwendet werden, um Subpopulationen zu finden, die überdurchschnittliche Vorhersagefehler aufweisen. Somit können Gruppen gefunden werden, bei denen der Datenproduktionsprozess im Vergleich zu anderen Gruppen oder zu einem absoluten Schwellenwert schlechter abschneidet. Im Kontext des *überwachten Lernens* (supervised ML) kann Genauigkeit als Qualitätsdimension insofern um Fairnessaspekte erweitert werden, als dass genaue Vorhersagen nicht nur insgesamt, sondern auch für Subgruppen verlangt werden, die durch geschützte Attribute oder andere relevante Merkmale definiert sein können (Hebert-Johnson u. a., 2018; Kim, Ghorbani und Zou, 2019).

Robustheit (robustness) Mangelnde Robustheit und Stabilität kann auf verschiedene Weise mit Fairnessaspekten in Verbindung gebracht werden. Auf organisatorischer Ebene kann "'model decay" oder "model drift" (d. h. eine mit der Zeit nachlassende Vorhersagegenauigkeit) ein Grund für (möglicherweise selektive) Skepsis von Modellnutzern gegenüber algorithmischen Lösungen sein (Choi u. a., 2022). In nachgelagerten Anwendungen kann sich Modelldrift auf unterschiedliche Weise auf verschiedene Teile der Zielpopulation auswirken. Das heißt, es können unterschiedliche Fehler zwischen sozialen Subgruppen auftreten oder aufgrund von Änderungen in den Daten, auf die das Modell angewendet wird, verstärkt werden. Es kann auch schwieriger sein, Modelldrift zu erkennen, der hauptsächlich oder zuerst in (kleinen) geschützten Gruppen auftritt. Außerdem kann eine spezifische Art von Drift, nämlich das Auftauchen neuer Kategorien in einem kategorialen Merkmal, unmittelbar mit der Erkennung von geschützten Gruppen in Zusammenhang stehen. Wir schlagen aus der Fairnessperspektive daher vor, dass Gütemaße auch auf Subgruppenebene in der Modellimplementierung über die Zeit überwacht werden sollten. Dies gibt nicht nur Aufschluss darüber, wann Subgruppenfehler einen vorab festgelegten Schwellenwert überschreiten, sondern auch über mögliche Ursachen und Gegenmaßnahmen. Wir argumentieren außerdem, dass eine sorgfältige Überwachung auch dann erforderlich ist, wenn Modelle regelmäßig neu trainiert werden, da im Laufe der Zeit möglicherweise neue Verzerrungen (in den Daten) auftreten können.

Reproduzierbarkeit (reproducibility) Aus der Fairnessperspektive wirft (unzureichende) Reproduzierbarkeit Fragen danach auf, wie stark sich Designentscheidungen in der Entwicklung des Machine Learning Modells auf dessen Outputs auswirken, und zwar nicht nur insgesamt, sondern

auch separat für sensible Subgruppen. Fairnessrelevante Entscheidungspunkte können nicht nur das Training des Modells selbst umfassen (z. B. die Wahl des Modelltyps und der Hyperparameter), sondern auch subtilere Aspekte wie implizite Entscheidungen in der Datenaufbereitung. So kann gezeigt werden, das Entscheidungen hinsichtlich der Aufteilung der Daten in Trainings- und Testset, die Festlegung des Klassifizierungsschwellenwertes sowie das Vorgehen zum imputieren fehlender Werte Fairnessmaße auf unterschiedliche Weise beeinflussen können (Caton, Malisetty und Haas, 2022; Simson, Pfisterer und Kern, 2024). In der Praxis können die Auswirkungen von Nichtreproduzierbarkeit wiederum durch die Strukturierung der Modellevaluation nach geschützten Attributen bewertet werden, gepaart mit Modellergebnissen welche unter unterschiedlichen Designentscheidungen zu erwarten wären. Eine starke Bedeutsamkeit von einzelnen Designentscheidungen ist wieder besonders dann besorgniserregend, wenn die Modellergebnisse nachgelagert weiterverwendet werden, z.B. als Input für weitere Analysen.

Kosteneffizienz (cost effectiveness) Die Kosten für die Einführung von Machine Learning Verfahren in der amtlichen Statistik gehen über Aspekte wie technische Ausstattung und Weiterbildungen hinaus. Aus Fairnessperspektive betonen wir Qualitätssicherung und -kontrolle als kritische Komponenten, nicht nur als Mittel zur Evaluation von ML-Modellen im Hinblick auf z. B. Schwankungen in der (Subgruppen-)Performanz, sondern auch als Sicherheitsmaßnahme: Menschen können (müssen) die Vorhersagen bzw. Outputs der Modelle überschreiben, wenn deren Unsicherheit einen vorab festgelegten Schwellenwert überschreitet (Bhatt u. a., 2020). Die Einführung einer "reject option" bei Machine Learning Modellen, d.h. die Weiterleitung schwieriger Fälle zur manuellen Klassifizierung, kann Fairnessmaße verbessern (Kaiser, Kern und Rügamer, 2022), geht aber per Definition mit zusätzlichen Kosten für manuelle Arbeit einher. Die Bewertung der Notwendigkeit und des Ausmaßes menschlicher Aufsicht sollte daher in die Kosten-Nutzen-Analyse von ML-Anwendungen in der amtlichen Statistik einbezogen werden. Kurzum: Fairness kann nicht (vollständig) automatisiert werden (Weerts u. a., 2023).

Zeiteffizienz (timeliness) Die Berücksichtigung von Fairness bei der Diskussion und Bewertung von Qualitätsdimensionen sollte nicht als zusätzliche (zeitliche) Belastung wahrgenommen werden. Wie wir in den bisherigen Abschnitten zu veranschaulichen versucht haben, können Fairnessaspekte in der Praxis in bestehende Prozesse integriert evaluiert und als zusätzliche Absicherung betrachtet werden, um sicherzustellen, dass die durch ML-basierte Automatisierung möglicherweise verbesserte Effizienz nicht auf Kosten von Fairnessproblemen in der Modellimplementierung geht. Wir argumentieren, dass die bestehenden Qualitätsdimensionen des QF4SA-Frameworks jeweils von der Fairness-Perspektive profitieren, da sie die gewissenhafte Kontrolle von Algorithmen bereichert, indem sie die entscheidende Rolle (sozialer) Subgruppen hervorhebt.

6.3.2 Was fehlt?

Fairness Betrachtungen spielen auch eine Rolle jenseits der eben diskutierten Wechselwirkungen mit bereits etablierten Qualitätskriterien der amtlichen Statistik wie sie z.B. im QF4SA aufgeführt sind.

Ein zentraler Mechanismus, der Fairness-Probleme erzeugt, ist, wenn der funktionale Zusammenhang zwischen Input- und Outputvariablen variiert zwischen verschiedenen Gruppen A, diese Modellheterogenität aber im Modelltraining unerkannt bleibt. Selbst sehr flexible ML-Modellklassen sind nicht dagegen gewappnet, wenn es einfach zu wenige Fälle aus einer geschützten Gruppe A gibt (entweder in einem absoluten Sinne zu wenig oder relativ zu der zu erkennenden Komplexität des

funktionalen Zusammenhangs). Die Gründe für eine solche Unterrepräsentation können sowohl in der Datenerhebung liegen (zu Representationsfehlern wie coverage, sampling, und nonresponse errors im TSE Framework siehe Groves u. a. 2009) als auch während der Datenaufbereitung entstehen. Zu letzterem sind z.B. Methoden des *unüberwachten Lernens* (unsupervised ML) zu nennen: So können Angehörige von kleinen Minderheiten schon rein aufgrund ihrer Seltenheit fälschlich als Ausreißer identifiziert und automatisiert aus den Daten entfernt werden. Ebenso können bei der Verknüpfung von verschiedenen Datenquellen (record linkage) vermehrt Individuen aus bestimmten Gruppen herausfallen: das für die Verknüpfung nötige Erkennen desselben Individuums in zwei Datensätzen, welches häufig die Namen der Individuen einbezieht, kann häufiger scheitern für Personen mit nicht-lokalen Namen, z.B. aufgrund von verschiedenen Transkriptionsmöglichkeiten oder vermehrten Schreibfehlern.

Als Datenproduzentin kann die amtliche Statistik in der fortlaufenden Qualitätskontrolle beobachten, ob Angehörige von Gruppen A überhäufig aus den Daten entfernt wurden oder ob ihre absolute Zahl für spätere Datenanalysen zu gering wird. Daneben gibt es idealerweise Goldstandard-Daten, die den wahren Anteil von Gruppen A in der Zielpopulation akkurat widerspiegeln, sodass eingeschätzt werden kann, ob und, wenn ja, welche Gruppen in den Daten unterrepräsentiert sind. Wie bereits oben ist auch hier unsere Empfehlung, mit den (ggf. anwendungsspezifischen) typischen Fairness-relevanten Gruppen zu beginnen und dies zu flankieren durch die automatisierte Suche nach betroffenen (Sub-)Gruppen.

Ohne rechtliche Vorgaben (bzw. zusätzlich zu diesen) bzgl. der nötigen Mindestrepräsentation von Gruppen in Daten sehen wir hier die Notwendigkeit, organisationsinterne Schwellen und Kriterien zu diskutieren, die aber durchaus zwischen Anwendungsfällen variieren können. Spezifisch sehen wir hier die statistischen Konsequenzen in weiteren Datenanalysen und Datenanalyseschritten als eine Quelle für Kriterien: Wie viele Einheiten einer Gruppe sind nötig, um Modellheterogenität von einem inhaltlich relevanten Ausmaß wahrscheinlich erkennen zu können? Wie viele Einheiten sind nötig, um Unsicherheiten (in der Schätzung sowohl von interessierenden Parametern als auch von Fairness-Metriken) innerhalb eines tolerierbaren Rahmens zu halten? Power Analysen und andere statistische, häufig simulationsbasierte Verfahren können bei der Quantifizierung dieser Aspekte helfen.

Die amtliche Statistik kann dabei nicht alle möglichen Verwendungen ihrer Daten(produkte) antizipieren. Jedoch sollten, nach Möglichkeit, zukünftige Anwender in die Lage versetzt werden, einschätzen zu können, welche Daten(produkte) in welchem Maße welchen (Fairness-)Ansprüchen gerecht werden können. Metadaten und "report cards" mit Fairness-relevanten Informationen können dazu ein wichtiger Beitrag sein.

6.4 Diskussion und Ausblick

Die zunehmende Nutzung multipler Datenquellen zum Trainieren von ML-Modellen im ADM Kontext in Kombination mit dem kumulativen Effekt von Fehlern und Verzerrungen in Daten entlang der ML-Pipeline unterstreichen die Bedeutung von Fairness als ein Qualitätskriterium in der Datenproduktion der amtlichen Statistik. Faire Datenprodukte und deren transparente Dokumentation legen den Ausgangspunkt für faire datengestützte Entscheidungssysteme. Wir argumentieren weiterhin, dass sich Fairnessaspekte in bestehende Qualitätskontrollen und -prozesse in der amtlichen Statistik einbinden lassen und letztlich auch für die bereits bestehenden Qualitätsdimensionen einen Mehrwert bieten. Insbesondere soll die kontinuierliche Integration von Fairnessüberlegungen und -evaluation entlang der ML Pipeline auch frühzeitig verhindern, dass Ressourcen (Zeit, Arbeitsstunden, Energie

und computationale Ressourcen) für die Weiterarbeit an solchen ML-Modellen verwendet werden, deren (Fairness-)Performance sich im Nachhinein als zu ungenügend erweisen wird, um diese Modelle dann auch tatsächlich anzuwenden. Die kontinuerliche Fairness-Evaluation hilft also dabei, (Fairness- und anderweitige Qualitäts)Probleme möglichst dort in der ML-Modell Entwicklungskette zu entdecken, wo sie auftreten. Der explizite Fokus auf soziale Gruppen in der Modellevaluation ist für zentrale Aspekte wie (Vorhersage)genauigkeit und Robustheit gleichsam förderlich. In der Praxis kann hierbei auf ein breites Toolkit von Metriken und Verfahren der fair ML Literatur zurückgegriffen werden, welche in "Fairness Auditing" Softwarepaketen bereits implementiert sind (Pfisterer, Siyi und Lang, 2024; Pfisterer u. a., 2021).

Die Fairnessperspektive stärkt explizit die Bedeutsamkeit von Daten. Erstens ist die Datenqualität wohl der größte Faktor, der zur (Un)Fairness von ML-Modellen beiträgt. Zweitens sind Datenproduzenten wie die amtliche Statistik dadurch in einer einzigartigen Position, da sie viel Ermessensspielraum und zugleich Expertise bei der Gestaltung, Erhebung, Aufbereitung und Produktion von (bedeutsamen) Daten haben. Dies ist eine dringend notwendige Erweiterung der Perspektive und wir sind dankbar für die Kooperation mit der amtlichen Statistik, die klar aufzeigt, dass Machine Learning i.A. und fair ML im Besonderen nicht nur im Sinne von "Datenanalyse" gedacht werden sollte, sondern verstärkt auch die Genese von Daten(produkten) in den Blick nehmen muss.

Vor diesem Hintergrund können Fairnessüberlegungen auch als Chancen verstanden werden: Die amtliche Statistik kann neue Datenprodukte mit Blick auf soziale Gruppen erstellen, bestehende Produkte speziell für diese Gruppen verbessern und faire ML-Modelle nutzen, um (Sub-)Gruppen zu finden, die durch ein bestimmtes Datenprodukt tatsächlich benachteiligt sind. Die neuen Chancen betreffen gleichermaßen bestehende Datenprodukte: Daten der amtlichen Statistik können als "benchmark" Daten für Fairnessevaluationen genutzt werden, sowohl im Vergleich zu anderen Daten als auch für ML-Modelle selbst. Die Bedeutsamkeit rigoros hochwertiger Datenprodukte, und somit die Relevanz von Methoden der amtlichen Statistik für die fair ML Forschung, kann somit nur unterstrichen werden.

7 Rechtliche Aspekte

7.1 Relevanz und Aufgabe der amtlichen Statistik in Deutschland

Die Notwendigkeit und Aufgabe von amtlicher Statistik in Deutschland sind stark durch die Rechtsprechung des Bundesverfassungsgerichts geprägt. Vorgaben zur Durchführung von Statistiken sind ferner durch das Bundesstatistikgesetz, welches an mancher Stelle auch als Grundgesetz der Statistik bezeichnet wird[10], sowie die einzelnen Landesstatistikgesetze gegeben. Die Gesetzgebungskompetenz des Bundes zur Durchführung von Statistiken auf Bundesebene ergibt sich aus Art. 73 Abs. 1 Nr. 11 GG.

Die amtliche Statistik bildet für jede moderne und leistungsfähige Gesellschaft eine notwendige Infrastruktur[11]. Aus rechtlicher Sicht führt in Deutschland für die dabei unerlässliche Datenverarbeitung kein Weg am wegweisenden Volkszählungsurteil des Bundesverfassungsgerichts[12] von 1983 vorbei, das sich mit dem Eingriff in Grundrechte durch statistische Erhebungen auseinandersetzte. Dieses Urteil stellt einen Meilenstein für die Rechtsprechung zur Statistik dar, da sich das BVerfG sehr detailliert mit den Grundsätzen der statistischen Erhebung auseinandersetzte. Dies war der Vielzahl an Verfassungsbeschwerden geschuldet, die auch einer allgemeinen Angst geschuldet waren, dass die Anzahl an automatisierter Datenverarbeitung von staatlicher Stelle zugenommen hatte (Simitis, 2000). In dem Urteil konstatierte das Gericht, dass eine am Sozialstaatsprinzip orientierte staatliche Politik ökonomische, soziale und ökologische Veränderungen nicht einfach als unabänderliches Schicksal hinnehmen darf[13]. Als Handlungsgrundlage für Politik liefert die amtliche Statistik als Informationsinstrument und -quelle die notwendige informationelle Entscheidungsgrundlage. Sie stellt eine informationelle Basis sowohl für staatliches Handeln als auch für individuelle Meinungsbildung dar, denn Statistiken „ermöglichen es prinzipiell jedem, gesellschaftliche, wirtschaftliche und ökologische Phänomene zu beobachten und zu beurteilen“ (Radermacher, 2017).

Dabei gilt jedoch der Grundsatz, dass die statistische Erhebung zu einer Datenwiedergabe in strukturierter und anonymer Form führen muss[14]. Die amtliche Statistik liefert Ergebnisse über Massenerscheinungen und dient gerade nicht der Darstellung von personen- oder institutionsbezogenen Informationen. Folglich sind auch die Entscheidungen, die auf Grundlage dieser Darstellungen erfolgen, nicht an den Einzelfall gerichtet, sondern adressieren gesamtgesellschaftliche Beobachtungen und Probleme[15]. Methodik und Erhebungsform richten sich nach dem aktuellen Stand der Methodendiskussion, es sind also moderne Methoden zur Datenverarbeitung zu verwenden und neue Erkenntnisquellen auszuschöpfen[16].

Das BVerfG setzte sich in seinem Urteil aber auch mit allgemeinen Grundsätzen der Datenerhebung seitens staatlicher Stellen auseinander. So stellte es fest, dass bei einer Datenerhebung für statistische Zwecke Besonderheiten für die Zweckbindung greifen, da es in der Natur der Statistik liegt, dass die Daten nach ihrer Aufbereitung für verschieden Zwecke verwendet werden[17]. Diese Lockerung der Zweckbindung hat das BVerfG in seiner weiteren Rechtsprechung zum Zensus 2011 bestätigt[18]. Das

[10] Etwa hier: https://www.destatis.de/DE/Ueber-uns/Geschichte/HistorischesDossier/_inhalt.html (Stand 18.04.24).
[11] BT-Drs. 10/5345, S. 13.
[12] BVerfGE 65, 1.
[13] BVerfGE 65, 1 (47).
[14] BT-Drs. 10/5345, S. 13; BVerfGE 65, 1 (53f.).
[15] *Kühling/Schmid* in: Kühling, § 1 BstatG Rn. 7; BVerfGE 65, 1 (64)(Kühling, 2023).
[16] Vgl. BVerfGE 65, 1 (55f.); BVerfGE 150, 1 (110).
[17] BVerfGE 65, 1 (47).
[18] BVerfGE 150, 1 (108 Rn. 223).

Volkszählungsurteil ist damit das Grundsatzurteil im Bereich des deutschen Datenschutzrechts[19].
Das BVerfG meint mit Statistik die methodische Erhebung, Sammlung, Darstellung und Auswertung von Daten und Fakten für staatliche Zwecke[20]. Auch innere Tatsachen und Vorgänge sowie politische Wertungen und Meinungen können Teil davon sein, solange sie einen dienenden Charakter haben[21]. Sinn und Zweck der amtlichen Statistik ist die Erhebung von Tatsachenmaterial und nicht das Bewirken politischer Handlungen oder Aktionen[22]. Sie ist dabei auf die abstrakt-generelle Darstellung allgemeiner Entwicklungen und Phänomene beschränkt[23].
In § 1 BStatG werden die Aufgaben der Bundesstatistik erläutert, die Norm orientiert sich dabei an Vorgaben aus dem Volkszählungsurteil des Bundesverfassungsgerichts. Unter Statistik ist demnach Statistik für Bundeszwecke zu verstehen. Satz 1 normiert es als Aufgabe der Bundesstatistik „laufend Daten über Massenerscheinungen zu erheben, zu sammeln, aufzubereiten, darzustellen und zu analysieren.“ In der Wahrung ihrer Aufgaben ist die amtliche deutsche Statistik nach § 1 Satz 2 BStatG den Grundsätzen der Neutralität, der Objektivität und der fachlichen Unabhängigkeit unterworfen[24]. Neutralität verlangt dabei, dass keinen einzelnen fremden Interessen Vorzug gegeben werden darf. Objektivität fordert die Entwicklung und Verbreitung von Bundesstatistiken auf systematische, zuverlässige und unvoreingenommene Weise. Fachliche Unabhängigkeit gilt vor allem im Hinblick auf die Verfahren, Definitionen, Methoden und Quellen sowie den Zeitpunkt und Inhalt aller Verbreitungsformen[25]. Diese Unabhängigkeit wird jedoch insoweit aufgehoben, dass der Gesetzgeber Vorgaben hinsichtlich der Methode geben kann, zu deren Vorgabe er bereits durch die Verfassung gezwungen ist[26].
Wichtig für eine moderne Statistik ist insbesondere Satz 3, der zu einem Einsatz sachgerechter Methoden und Informationstechniken beim Datengewinn aufruft. Damit wird konkretisiert, wie die Gewinnung der zugrundeliegenden Daten erfolgt. Die Bundesstatistik hat dabei wissenschaftliche Erkenntnisse zu verwenden, hierunter fallen sämtliche Resultate eines Verfahrens, das nach Inhalt und Form als ernsthafter Versuch zur Ermittlung der Wahrheit anzusehen ist[27]. Zudem sind auch die jeweils sachgerechten Methoden einzusetzen. Dies ist im Einzelfall abzuwägen, wobei zum einen ein möglichst hoher Grad an Genauigkeit[28] gefordert wird und andererseits eine hinreichend genaue Ermittlung notwendig ist[29]. Genauigkeit, Erforderlichkeit und Ressourcenaufwand sind dabei abzuwägen[30].

7.2 Relevanz des Datenschutzes insbesondere bei statistischen Daten

7.2.1 Datenschutz durch das Grundgesetz

Bereits vor dem Volkszählungsurteil hatte das Bundesverfassungsgericht bejaht, dass dem einzelnen Bürger durch das allgemeine Persönlichkeitsrecht ein Schutzrecht vor Eingriffen in den Bereich

[19] Vgl. *Kühling/Raab* in: Kühling/Buchner, DSGVO/BDSG, Einführung, Rn. 122m (Kühling und Buchner, 2024).
[20] BVerfGE 150, 1 (79 Rn. 144).
[21] BVerfGE, 8, 104 (111).
[22] BVerfGE, 8, 104 (111).
[23] *Kühling/Schmid* in: Kühling, § 1 BstatG Rn. 7.
[24] Diese Grundsätze sind denen aus Art. 2 der VO (EG) Nr. 223/2009 ähnlich. Die Verordnung erweitert die Grundsätze zudem noch um die der Zuverlässigkeit, der Geheimhaltung sowie der Kostenwirksamkeit.
[25] *Kühling/Schmid*, in: Kühling, § 1 BStatG, Rn. 19ff.
[26] *Kühling/Schmid* in: Kühling, § 1 BStatG, Rn. 19ff.
[27] BVerfGE 150, 1 (110, 113f.).
[28] BVerfGE 150, 1 (109, 157); BVerfGE 65, 1 (50).
[29] BVerfGE 150, 1 (111, 138, 153).
[30] BVerfGE 150, 1 (133).

privater Lebensführung zusteht. Dazu unterschied das Gericht zwischen verschiedenen Sphären der privaten Lebensführung: Der Sozial-, der Privat- und der Intimsphäre. Das Rechtfertigungsbedürfnis ist bei Eingriffen umso höher, je tiefer der Eingriff in das Privatleben des Betroffenen reicht[31].
Das Grundrecht auf informationelle Selbstbestimmung, welches das Gericht im Volkszählungsurteil aus dem allgemeinen Persönlichkeitsrecht aus Art. 2 Abs. 1 und Art. 1 Abs. 1 des Grundgesetzes ableitete, schützt vor staatlicher Informationsverarbeitung(Schantz, Wolff u. a., 2017). Es überlässt dem Einzelnen die Entscheidung, was über ihn bekannt sein soll[32]. Ein Grundgedanke dieses Urteils war es, dass die Bevölkerungsstatistik möglichst eingriffsarm für die einzelnen Bürger sein muss. Das BVerfG definierte den Begriff des Datums vor dem Hintergrund der automatisierten Datenverarbeitung sehr weit, indem es festhielt, dass es kein an sich belangloses Datum mehr gebe[33]. Seit der Entscheidung kommt dem Datenschutz in Deutschland Verfassungsrang zu[34]. Das Grundrecht schützt aber nicht schrankenlos vor Eingriffen, da individuelle Realität auch teilweise ein Abbild des allgemeinen sozialen Gefüges darstelle und daher der Allgemeinheit diene[35], sodass Eingriffe in das Grundrecht zulässig sein können. Das Bedürfnis einer belastbaren Entscheidungsgrundlage des Staates für sein Handeln dient dabei als Gegenbelang, der einen Eingriff in die informationelle Selbstbestimmung rechtfertigen kann. Solche Eingriffe müssen durch verhältnismäßige Maßnahmen aber eine möglichst geringe Belastungstiefe haben.
Der amtlichen Statistik ist es laut Bundesverfassungsgericht jedoch untersagt, umfassende Persönlichkeitsprofile von Personen zu erstellen[36]. Bestünde diese Möglichkeit, könnten sich Einschüchterungseffekte einstellen, weil Bürger nicht mehr wissen, „wer was wann und bei welcher Gelegenheit über sie weiß“[37], was sie in der Ausübung weiterer Freiheitsrechte einschränken könnte. Durch die Vermeidung von abschreckenden Effekten wird gerade die Handlungs- und Mitwirkungsfähigkeit eines freiheitlich-demokratischen Gemeinwesens garantiert[38]. Statistische Geheimhaltung kann daher als „Fundament der amtlichen Statistik“[39] bezeichnet werden.
Zudem ist insbesondere die Nutzung von statistischen Daten für den Verwaltungsvollzug nicht mit dem Grundrecht auf informationelle Selbstbestimmung vereinbar (sogenanntes Rückspielverbot)[40] (Thiel und Puth, 2023). Aufgabe der amtlichen Statistik ist es vielmehr, Massenphänomene abzubilden und statistisch darzustellen, die keinen Personenbezug mehr aufweisen. Die abstrakt-generellen Darstellung dient gerade nicht dem Verwaltungsvollzug, da dieser auf eine Adressierung der einzelnen Person ausgelegt ist.
Als einfachgesetzlicher Ausfluss des oben genannten Grundrechts auf informationelle Selbstbestimmung fordert § 16 BStatG daher die Geheimhaltung von statistischen Einzeldaten für die amtliche Bundesstatistik[41]. Die Geheimhaltungspflicht stellt ein Gegengewicht dar zur Auskunftspflicht der Befragten und der Tatsache, dass sich die genauen Zwecke der statistischen Erhebung im Vorhinein nicht abschließend eingrenzen lassen[42].
Für Unternehmen besteht durch andere Grundrechte, ein Schutz vor der Preisgabe von Daten

[31] *Lang* in: BeckOK Grundgesetz, Art. 2, Rn. 75ff (Epping und Hillgruber, 2024).
[32] BVerfGE 65, 1 (43).
[33] BVerfGE 65, 1 (43).
[34] *Gusy/Eichenhofer* in: BeckOK Datenschutzrecht, § 1 BDSG Rn. 17 (Wolff, Brink und Ungern-Sternberg, 2024).
[35] BVerfGE 65, 1 (43f.).
[36] BVerfGE 65, 1 (53).
[37] BVerfGE 65, 1 (42f.); BVerfGE 115, 166 (188).
[38] BVerfGE 65, 1 (43).
[39] Dorer/Mainusch/Tubies, § 16 BStatG Rn. 1 (Dorer, Mainusch und Tubies, 1988).
[40] BVerfGE 65, 1 (51f); BVerfGE 150, 1 (109f.).
[41] *Kühling/Sauerborn*, in: Kühling, § 16 BStatG Rn. 1.
[42] BVerfGE 65, 1 (48).

gegenüber statistischen Ämtern. So kann eine Mitwirkungspflicht Unternehmen in ihrem Grundrecht auf Berufsfreiheit aus Art. 12 Abs. 1 GG beschränken. Eine Berufung der Unternehmen auf dieses Grundrecht ist nach Art. 19 Abs. 3 GG bei einer wesensgemäßen Ausübung vergleichbar mit der einer natürlichen Person möglich[43]. Ebenso ist ein Eingriff in das Grundrecht der Eigentumsfreiheit aus Art. 14 GG denkbar[44]. Auch diese bedürfen zur Rechtfertigung eines Eingriffs eines hohen Standards bei der Geheimhaltung.

2.2. Sekundärrechtlicher Datenschutz durch die DSGVO

Da Statistik zumindest bei der Erhebung und Verarbeitung personenbezogene Daten erfassen kann (beispielsweise beim Zensus), ist der Anwendungsbereich der DSGVO eröffnet. Diese etabliert einen einheitlichen europäischen Rechtsrahmen, der das europäische Grundrecht auf Datenschutz aus Art. 8 GrCH, welches sich auch in Art. 16 AEUV wiederfindet, durch Rechte der betroffenen Personen sowie Verpflichtungen derjenigen, die die Daten verarbeitet, sekundärrechtlich konkretisiert. Die DSGVO betrifft die automatisierte ebenso wie die manuelle Verarbeitung personenbezogener Daten. Für Bereiche der amtlichen Statistik sieht die DSGVO bei der Wahrung von Interessen der betroffenen Personen spezifische Privilegierungen vor, deren Umsetzung durch nationale Öffnungsklauseln teilweise den Mitgliedsstaaten überlassen ist. Statistik im Sinne der Verordnung ist als der methodische Umgang mit empirischen Daten zu verstehen[45].
So fordert Art. 89 Abs. 1 DSGVO geeignete Garantien beim Umgang mit personenbezogenen Daten für statistische Zwecke, die die Rechte und Freiheiten der betroffenen Person wahren. Unter diesen Garantien[46] ist die weitgehende Anonymisierung und hilfsweise Pseudonymisierung von personenbezogenen Daten zu verstehen[47]. Insbesondere Art. 89 Abs. 2 DSGVO erlaubt dafür im Gegenzug auch die Einschränkung von Betroffenenrechten, sofern diese die statistischen Zwecke erheblich beeinträchtigen oder unmöglich machen. Dies ist insbesondere denkbar durch einen übermäßigen Gebrauch dieser Rechte, der den betroffenen verarbeitenden Stellen einen hohen Verwaltungsaufwand einbringt oder aber deren Datengrundlage entzieht[48]. Ob die spezifischen Zwecke tatsächlich eingeschränkt oder gar unmöglich gemacht werden, hängt von einer Prognose im Einzelfall ab[49].
Eine weitere Ausnahme betrifft die Zweckbindung der Datenverarbeitung. Art. 5 Abs. 1 lit. b DSGVO sieht im Grundsatz die Verwendung von personenbezogenen Daten für festgelegte, eindeutige und legitime Zwecke vor. Ist eine Weiterverarbeitung beabsichtigt, erfordert dies gewöhnlich einen Kompatibilitätstest nach Art. 6 Abs. 4 DSGVO. Eine Weiterverarbeitung zu statistischen Zwecken wird jedoch – die Garantien des Art. 89 DSGVO vorausgesetzt – als kompatibel angesehen, sofern die Zwecke nicht auch mit anonymen Daten erreicht werden könnten[50].
Wenn die DSGVO von Daten für statistische Zwecke spricht, ist damit gemeint, dass die Daten bei Veröffentlichung keinen Personenbezug mehr aufweisen (vgl. ErwG 162 S. 3–5). Somit greift auch nur dann die Privilegierung. Ein Personenbezug, der bei der Erhebung und der Verarbeitung noch vorliegen kann, wird dabei durch die Trennung der Hilfsmerkmale[51] für interne Zwecke –

[43] BVerfGE 50, 290 (363).
[44] Vgl. hierzu Kühling, Einleitung BStatG, Rn. 59.
[45] *Buchner/Tinnefeld* in: Kühling/Buchner, Art. 89 DS-GVO Rn. 15.
[46] Zum Umfang und Ausmaß dieser Garantien vgl. *Weichert*, ZD 2020, 18 (22).
[47] *Pauly* in: Paal/Pauly, Art. 89 DS-GVO Rn. 13.
[48] *Pauly* in: Paal/Pauly, Art. 89 DS-GVO Rn. 14.
[49] *Pauly* in: Paal/Pauly, Art. 89 DS-GVO Rn. 14.
[50] *Buchner/Tinnefeld* in: Kühling/Buchner, Art. 89 DSGVO Rn. 21; ErwG 159 der DSGVO.
[51] Hilfsmerkmale dienen der technischen Durchführung von Statistiken und ermöglichen eine Identifizierung einzelner Personen, die vor allem der anfänglichen Prüfung der erhobenen Daten dienen und eine Rückfrage ermöglichen.

auf nationaler Ebene im Rahmen von § 12 BStatG – aufgehoben[52](Weichert, 2020). Ist dieser bei aggregierten statistischen Ergebnissen noch vorhanden, handelt es sich nicht mehr um privilegierte statistische Zwecke[53].
Der Datenschutz in der Statistik dient somit dem Erhalt dieser Privilegierung gegenüber den Regelungen der DSGVO, um eine effiziente, funktionierende Statistik zu gewährleisten. Daneben wäre die Funktion der Statistik auch dadurch eingeschränkt, dass bei einer niedrigen Geheimhaltung eine Partizipation der Befragten nicht gleichermaßen akkurat ausfallen würde und mit Falschangaben zu rechnen wäre(Couper u. a., 2008). Kommt es nicht zu einer strikten Geheimhaltung, könnte dies zu einer „schwindenden Kooperationsbereitschaft“ seitens der befragten Bürger führen[54].
Die Geheimhaltung von statistischen Einzelangaben ist auch Teil des Code of Practice für europäische Statistiken[55]. Dieser fordert in Art. 5 den Datenschutz und die Geheimhaltung von Einzeldaten durch Maßnahmen seitens der statistischen Ämter, gesetzliche Verpflichtung sowie durch Selbstverpflichtung der Mitarbeitenden der statistischen Ämter. Bei diesem Kodex handelt es sich zwar nicht um einen verbindlichen Rechtsakt, er wird jedoch in Art. 1, 2 Abs. 1 UAbs. 2 und 11 der Verordnung (EG) 223/2009, die auch als statistische Rahmenverordnung bezeichnet wird[56], erwähnt[57]. Im Statistikrecht kommt ihm daher also erhebliche Bedeutung zu.

7.3 Vorteile von Machine Learning für die amtliche Statistik

Den Wesenskern des Machine Learning (ML) fasst der Technologiejournalist Thomas Ramge so zusammen: „Bei Maschinellem Lernen erkennen Computersysteme Muster in Beispielen und können ihre ‚Erkenntnisse‘ auf andere Beispiele übertragen. So lernen sie, aus Daten immer genauere Schlüsse zu ziehen und Entscheidungen abzuleiten.“(Ramge, 2018)
In der amtlichen Statistik ergeben sich verschiedene Möglichkeiten Machine Learning in die Datenverarbeitung einzubinden. Ein Einsatz bietet sich insbesondere für Prozesse der Datenaufbereitung an.
Der Einsatz von ML in der amtlichen Statistik ist beispielsweise bei Klassifikationsverfahren gut denkbar, die andernfalls nicht oder nur zeitintensiv händisch durchführbar wären[58](Dumpert, 2021). Hierzu zählt die Kodierung erhobener Mikrodaten[59], wobei Daten einem amtlichen systematischen numerischen Code zugeordnet werden. Dies geschieht durch Texterkennung der Erhebungsdaten, die anschließend über einen Algorithmus zugeordnet werden. Bei den Erhebungsvariablen kann es sich

Hilfsmerkmale sind beispielsweise Name, Nachname oder Anschrift. Vgl. Dorer/Mainusch/Tubies, § 10 BStatG Rn. 4.

[52]Buchner/Tinnefeld in: Kühling/Buchner, Art. 89 DS-GVO Rn. 15.

[53]*Buchner/Tinnefeld* in: Kühling/Buchner, Art. 89 DS-GVO Rn. 15, 15a.

[54]Vgl. BVerfGE 65, 1 (50); Vgl. auch BT-Drs. I/982, S. 20 zu § 10 des Entwurfs des Volkszählungsgesetzes; Vgl. Dorer/Mainusch/Tubies, § 16 BStatG Rn. 4.

[55]Amt für Veröffentlichungen der Europäischen Union (2018): Verhaltenskodex für europäische Statistiken für die nationalen statistischen Ämter und Eurostat (statistisches Amt der EU), angenommen vom Ausschuss für das Europäische Statistische System am 16. November 2017; abzurufen unter: https://ec.europa.eu/eurostat/documents/4031688/8971242/ (Stand: 09.02.2024).

[56]*Kingreen* in: Callies/Ruffert, Art. 338 AEUV Rn. 3 (Callies und Ruffert, 2022).

[57]*Kühling*, Einleitung BStatG, Rn. 77.

[58]https://www.destatis.de/DE/Service/Hauptstadtkommunikation/Zukunft/maschinelles-lernen.html (Stand: 09.02.2024).

[59]Mikrodaten, auch Einzelangaben genannt, sind Angaben über persönliche und sachliche Verhältnisse eines Merkmalsträgers, z. B. über eine bestimmte Person (deren Beruf, gesundheitliche Verhältnisse, Noten etc.). Einzelangaben finden sich in der amtlichen Statistik beispielsweise für Personen, Haushalte oder Unternehmen. Die Einzelangaben enthalten die maximale Informationsmenge einer Statistik und sind damit sozusagen der Rohstoff des Statistikers (https://www.destatis.de/DE/Service/Statistik-Campus/ESC/mikrodaten.html).

um verschiedene Eigenschaften wie beispielsweise das Alter, Gehalt, eine Berufsbeschreibung oder die Beschreibung einer Verletzung handeln[60]. Ein Beispiel hierfür ist die Zuordnung von Berufen nach einem Occupational Classification System[61].
Die Plausibilisierung bietet eine weitere Einsatzmöglichkeit. Dieser Prozess hat das Ziel, ungewöhnliche Werte in Daten zu erkennen und diese zu kennzeichnen. Dabei können auch Regeln für die Plausibilisierung erkannt werden, die vorher auf Intuition der zuständigen Mitarbeitenden gebaut haben; zudem besteht die Möglichkeit, versteckte Strukturen in Daten zu finden[62]. Dieser Prozess lässt sich durch den Einsatz von ML beschleunigen[63]. Eng damit verknüpft ist die Imputation, bei der (vorher erkannte) fehlende oder falsche Werte ergänzt beziehungsweise ersetzt werden(Preising, Lange und Dumpert, 2021).
Bei der Betrachtung der rechtlichen Perspektiven des Einsatzes von ML in der amtlichen Statistik sind die Nutzungsmöglichkeiten und Vorteile relevant, die die Technologie bietet. Einen ersten Anhaltspunkt für die Notwendigkeit einer Ausweitung der Methoden in der amtlichen Statistik ist die Tatsache, dass das Bundesverfassungsgericht im Volkszählungsurteil zur kontinuierlichen Methodendiskussion ermahnt hat und die Methodenentwicklung dahingehend zu „beobachten, ob sie grundrechtsschonendere Verfahren ermöglicht“[64]. Dies hat den Zweck, abzuschätzen, ob der technische Fortschritt Erhebungsmethoden hervorbringt oder bereits hervorgebracht hat, die die Eingriffstiefe in das Grundrecht auf informationelle Selbstbestimmung verringern[65]. So können (mittlerweile unübliche) Vollerhebungen zu einer Preisgabe von Daten führen, die nicht Teil der eigentlichen Erhebung sind[66]. Auch im Qualitätshandbuch der amtlichen Statistik findet sich dementsprechend der Grundsatz, dass die Methodik statistischer Prozesse dem „Stand der wissenschaftlichen Forschung entsprechen“ muss(Statistische Ämter des Bundes und der Länder, 2021; Dumpert, 2021). Gegebenenfalls lässt sich durch den Einsatz von ML die Plausibilisierung so gestalten, dass weniger Rückfragen bei den Befragten notwendig sind, was ebenfalls eine geringere Eingriffstiefe erreichen würde.
Ein weiteres Argument ist, dass ein effektiverer Einsatz von personellen Ressourcen zu erwarten ist, wenn die Mitarbeitenden der statistischen Ämter durch ML unterstützt würden(Dumpert, 2021). Damit wäre dem Allgemeininteresse an einer kosteneffizienten Verwaltung Rechnung getragen. Gegebenenfalls kommt es dadurch auch zu einem höheren Datenschutzniveau, weil weniger Personen Kontakt mit den erhobenen Daten haben.

[60] United Nations Economic Commission for Europe – High Level Group for the Modernization of Official Statistics (UNECE – HLG-MOS) Machine Learning Project – Classification and Coding Theme Report, p. 3; abzurufen unter: https://statswiki.unece.org/display/ML/WP1+-+Theme+1+Coding+and+Classification+Report (Stand: 14.03.2024).
[61] UNECE – HLG-MOS Machine Learning Project - Classification and Coding Theme Report; Hierbei handelt es sich um Klassifikationssystem für die Berufszugehörigkeit von Personen.
[62] UNECE HLG-MOS Machine Learning Project Theme – Report of the Editing Imputation group; abzurufen unter: https://statswiki.unece.org/display/ML/WP1+-+Theme+2+Edit+and+Imputation+Report(14.03.2024).
[63] UNECE – HLG-MOS Machine Learning Project – Work Package 1 - Executive Summary; abzurufen unter: https://statswiki.unece.org/display/ML/WP1+-+Executive+Summary (14.03.2024).
[64] BVerfGE 150, 1 (133); vgl. auch BVerfGE 65, 1 (55f.).
[65] BVerfGE 150, 1 (110, 113f.).
[66] BVerfGE 150, 1 (134).

7.4 Gefahren für den Datenschutz durch den Einsatz von Machine Learning?

Der Einsatz von Machine Learning könnte aber auch Gefahren für den Datenschutz nach sich ziehen. Die Vergangenheit hat in mehreren Fällen gezeigt, dass eine Identifizierung von Personen anhand weniger Merkmale möglich sein kann. Die amerikanische PhD-Studentin Latanya Sweeney etwa bewies Anfang der 2000er Jahre, dass ein Großteil der Bevölkerung der USA allein anhand von drei scheinbar unpersönlichen Merkmalen eindeutig zugeordnet werden konnte. Sie demonstrierte dies eindrücklich, indem sie Gesundheitsdaten einer Versicherung mit den günstig zu erwerbenden Daten einer Voter Registration Liste abglich. Die Zuordnung der Daten ermöglichte die eindeutige Klassifikation von Einzelpersonen, unter ihnen der Gouverneur von Massachusetts(Sweeney, 2002).
Ein weiterer Fall handelt von einer Veröffentlichung von Filmbewertungen durch Netflix mit dem Ziel, die Präferenzen von Nutzern basierend auf deren Verlauf und Bewertungen zu bestimmen. Arvind Narayanan and Vitaly Shmatikov von der University of Texas zeigten, dass diese Daten ausreichten, um mindestens einen Teil der Nutzer eindeutig zu identifizieren, sofern man sie mit weiteren Daten darüber verknüpfte, welche Filme von den Personen bereits geschaut wurden(Narayanan und Shmatikov, 2008).
Hieran wird deutlich, dass anhand weniger Daten mit vermeintlich geringem Aussagegehalt ein konkreter Personenbezug hergestellt werden kann. Wie oben dargelegt, weist ML die Fähigkeit auf, Daten in großen Mengen zu verknüpfen und Muster in diesen zu erkennen. Eine solche Möglichkeit zur besseren Datenverknüpfung könnte aber ein Risiko für die Sicherheit von Daten darstellen und statistische Daten auf individueller Ebene zuordenbar machen.
Der technische Fortschritt und die damit einhergehende Datenverarbeitung wird teilweise als Rückschritt für den Datenschutz angesehen. An mancher Stelle geht man sogar davon aus, dass Anonymität nicht mehr möglich sei(Boehme-Neßler, 2016; Sarunski, 2016). Ob diese Aussage mit derartiger Schärfe zutrifft, kann hier nicht abschließend erläutert werden. Sicher ist jedoch, dass sich die Mittel zur Wiederherstellung eines Personenbezugs verbessert haben und die Grenze dessen, was noch als anonym gilt durch technischen Fortschritt verrückt wurde[67].
Relevant ist daher die Frage, ob solche Fälle eines Verlustes der Anonymität auch durch den Einsatz von moderner Datenverarbeitung in Form von ML in der amtlichen Statistik auftreten könnten.
Die zuvor genannten Prozesse der Kodierung, Plausibilisierung und Imputation von Erhebungsdaten dienen der Datenaufbereitung, das heißt sie erfolgen noch vor Trennung und Löschung der Hilfsmerkmale, sodass ohnehin noch ein Personenbezug vorliegt[68]. Darin unterscheiden sich die Methoden insoweit nicht von der ihnen analogen Pendants. Auch dabei werden die Hilfsmerkmale erst getrennt und gegebenenfalls gelöscht, wenn die Plausibilisierung der Werte erfolgt ist. Die Möglichkeit, die Daten schnell den Hilfsmerkmalen zuzuordnen, um etwaige Rückfragen zu ermöglichen, muss bis zu deren Plausibilitätsprüfung gegeben sein[69]. Insoweit ist also keine Schlechterstellung durch moderne, ML-gestützte Datenverarbeitung ersichtlich.
Denkbar wäre es jedoch, dass ein ML-Modell, das zur Prüfung der Daten verwendet wird, Daten speichert, die es ermöglichen, einen Personenbezug zu einem späteren Zeitpunkt wieder herzustellen(Song, Ristenpart und Shmatikov, 2017; Veale, Binns und Edwards, 2018). Um die gelöschten Hilfsmerkmale nach Löschung nicht wieder über den ML-Algorithmus herstellen zu können, wäre es daher ratsam, dass dieser lediglich mit den Erhebungsmerkmalen in Berührung kommt. Um

[67] *Ernst* in: Paal/Pauly, Art. 4 DSGVO, Rn. 50; Klar/Kühling in: Kühling/Buchner, Art. 4 DSGVO, Rn. 22.
[68] *Kühling/Schmid* in: Kühling, § 12 BStatG Rn. 15; BT-Drs. 10/5348, 18.
[69] Dorer/Mainusch/Tubies, § 12 BStatG Rn. 3.

ein hohes Datenschutzniveau zu gewährleisten, wäre auch ein Training mit bereits anonymen oder synthetischen Daten zu befürworten.
Gegebenenfalls ist aber entweder das Training oder die Benutzung eines ML-Modells nicht möglich, ohne dass Hilfsmerkmale verwendet werden. In diesem Fall wäre eine spätere Löschung der Modelldaten angebracht, wenn es zur Trennung und Löschung von Hilfsmerkmalen kommt, um nach erfolgreicher Plausibilisierung oder Kodierung keine Rekonstruierung über das Modell mehr zu ermöglichen. Dies liegt allein schon daran, dass § 12 BStatG eine irreversible Löschung der (Hilfs-)Daten fordert[70]. Darüber hinaus wäre eine Wiederherstellung des Personenbezugs auch durch § 21 BStatG ausgeschlossen, der die Reidentifizierung verbietet.
Demnach handelt es sich nach der Aufhebung des direkten Personenbezugs durch Trennung und Löschung nicht mehr um identifizierte Daten im Sinne des Art. 4 Nr. 1 DSGVO. Solche Daten liegen nur vor, wenn die Identität unmittelbar aus der Information selbst folgt[71]. Es wäre jedoch weiterhin denkbar, dass die Analyse mit einem ML-Algorithmus ein Muster offenbart, das die Zuordnung zu einer einzelnen Person wieder möglich macht(Boehme-Neßler, 2016). Insbesondere über die Verknüpfung mehrerer Datensätze, könnten sich einzigartige Merkmalskombinationen ergeben.
Daher ist also zu erläutern, ob es sich bei den Daten, die den statistischen Ämtern vorliegen, um personenbezogene Daten im Sinne des Art. 4 Nr. 1 DSGVO handelt, weil die dahinterstehende Person identifizierbar ist. Eine Person ist demnach dann identifizierbar, „wenn sie direkt oder indirekt, insbesondere mittels Zuordnung zu einer Kennung wie einem Namen, zu einer Kennnummer, zu Standortdaten, zu einer Online-Kennung oder zu einem oder mehreren besonderen Merkmalen, die Ausdruck der physischen, physiologischen, genetischen, psychischen, wirtschaftlichen, kulturellen oder sozialen Identität dieser natürlichen Person sind, identifiziert werden kann.“
Welches Wissen und welche Mittel zu berücksichtigen sind, um zu beurteilen, ob ein Personenbezug hergestellt und damit von personenbezogenen Daten ausgegangen werden kann, ist umstritten[72]. ErwG 26 der DSGVO ordnet an, dass bei der Frage, ob eine Person identifizierbar ist, alle Mittel zu berücksichtigen, die von dem Verantwortlichen oder einer anderen Person nach allgemeinem Ermessen wahrscheinlich genutzt werden, um die natürliche Person direkt oder indirekt zu identifizieren. Bei der Frage, ob Mittel nach allgemeinem Ermessen genutzt werden, sollten alle objektiven Faktoren, wie die Kosten der Identifizierung und der dafür erforderliche Zeitaufwand, herangezogen werden. Dabei sind die zum Zeitpunkt verfügbaren Technologien und technologischen Entwicklungen zu berücksichtigen. Es ist demnach eine Risikoanalyse vorzunehmen, die berücksichtigt, wie wahrscheinlich die Identifizierung der betroffenen Person ist(Nink und Pohle, 2015).
Hinsichtlich der Frage, ob auch Wissen anderer Personen berücksichtigt werden kann, wird teilweise ein absolutes Verständnis angenommen, nach welchem Informationen als personenbezogen angesehen werden, wenn der Verantwortliche oder eine beliebige dritte Person über das Wissen verfügen, um den Personenbezug herzustellen[73]. Nach dem relativen Verständnis des Personenbezugs hingegen sind für die Frage, ob mit Informationen der Personenbezug hergestellt werden kann, nur solche Mittel zu berücksichtigen, die nach allgemeinem Ermessen wahrscheinlich genutzt werden und dem Verantwortlichen tatsächlich im konkreten Einzelfall zur Verfügung stehen[74]. Diese Auffassung wird

[70] *Kühling/Schmid* in: Kühling, § 12 BStatG Rn. 11.
[71] EuGH, Urt. v. 19.10.2016 - C-582/14, ECLI:EU:C:2016:779 Rn. 38 – *Breyer*.
[72] *Karg* in: Simitis/Hornung/Spiecker gen. Döhmann, Datenschutzrecht, Art. 4 Nr. 1 Rn. 58 (Simitis, Hornung und Spiecker gen. Döhmann, 2019).
[73] *Karg* in: Simitis/Hornung/Spiecker gen. Döhmann, Datenschutzrecht, Art. 4 Nr. 1 Rn. 58; Vgl. *Klar/Kühling* in: Kühling/Buchner, Art. 4 DSGVO Rn. 25.
[74] *Karg* in: Simitis/Hornung/Spiecker gen. Döhmann, Datenschutzrecht, Art. 4 Nr. 1 Rn. 59.

auch vom Europäischen Gerichtshof (EuGH)[75] sowie in jüngerer Rechtsprechung vom Gericht der Europäischen Union (EuG)[76] geteilt[77]. Daher wäre auch bei einer technischen Möglichkeit nicht zwingend von einem Personenbezug auszugehen, solange keine relative Zuordenbarkeit bejaht werden kann[78]. Der Einsatz von rechtswidrigen Mitteln ist jedenfalls im Einklang mit der Rechtsprechung des EuGH von vornherein nicht relevant[79]. Damit ist die Personenbeziehbarkeit von Daten bei der amtlichen Statistik zu verneinen, da hier durch § 21 BStatG ein Verbot zur Reidentifizierung existiert.
Wenngleich also eine abstrakte Gefahr durch den Einsatz von intelligenter Datenverarbeitung für die Anonymität ausgeht, so relativiert sich diese bei näherer Betrachtung. Die Nutzung von ML beschränkt sich auf den Bereich innerhalb der amtlichen Statistik. Sie beeinflusst also nicht die statistische Geheimhaltung nach außen bei Veröffentlichung der Daten. Eine Reidentifizierung ist auch beim Einsatz moderner Datentechnik untersagt. Daher ist beim Training der Modelle darauf Rücksicht zu nehmen, dass diese keine personenbezogenen Daten speichern. Die strengen gesetzlichen Anforderungen der Geheimhaltung für die Mitarbeitenden der amtlichen Statistik schließen zudem eine Reidentifizierung durch diese aus. Insgesamt ist also davon auszugehen, dass dadurch geeignete Garantien im Sinne des Art. 89 DSGVO auch beim Einsatz von ML gewährleistet sind.

7.5 Auswirkungen der KI-Verordnung auf ML in der amtlichen Statistik

7.5.1 Allgemeines

Ein weiterer regulatorischer Ansatz, der ML betrifft, ist die geplante KI-Verordnung der EU. Wenngleich sie noch nicht verabschiedet wurde, handelt es sich dabei nach einer Einigung von Rat und Parlament im Dezember 2023[80] um reine Formsache. Es wird mit einer Veröffentlichung im ersten Halbjahr dieses Jahres gerechnet[81]. Über den geeinigten Verordnungstext wurde im Parlament im März abgestimmt, sodass mit einer Annahme im Rahmen des Berichtigungsverfahren vor Ende der Wahlperiode im Sommer zu rechnen ist[82]. Zum Erlass des Rechtsaktes bedarf es zudem noch der Zustimmung des Rates. Der ursprüngliche Vorschlag der EU-Kommission hat durch die Stellungnahmen von und verschiedene Änderungen erfahren. Es ist davon auszugehen, dass es sich bei dem Text, über den das Parlament im März abgestimmt hat, jedoch um den finalen Verordnungstext

[75] EuGH, Urt. v. 19.10.2016 - C-582/14, ECLI:EU:C:2016:779 Rn. 31ff. – *Breyer*.
[76] EuG 26.4.2023 – T-557/20, ECLI:EU:T:2023:219 Rn. 94ff. - *SRB v EDPS*.
[77] Das EuG geht nicht von personenbezogenen Daten aus, wenn dem Wortlaut der DSGVO nach pseudonyme Daten vorliegen, sofern der Verantwortliche aber nicht über den Schlüssel verfügt, um die Pseudonymisierung aufzuheben.
[78] EuGH, Urt. v. 19.10.2016 - C-582/14, ECLI:EU:C:2016:779 Rn. 46ff. – *Breyer*.
[79] *Karg* in: Simitis/Hornung/Spiecker gen. Döhmann, Datenschutzrecht, Art. 4 Nr. 1 Rn. 64; *Klar/Kühling* in: Kühling/Buchner, Art. 4 DSGVO Rn. 28; EuGH, Urt. v. 19.10.2016 - C-582/14, ECLI:EU:C:2016:779 Rn. 46ff. – *Breyer*.
[80] Rat der EU, Pressemitteilung vom 9. Dezember 2023 (https://www.consilium.europa.eu/de/press/press-releases/2023/12/09/artificial-intelligence-act-council-and-parliament-strike-a-deal-on-the-first-worldwide-rules-for-ai/).
[81] https://rsw.beck.de/aktuell/daily/magazin/detail/ki-verordnung-ante-portas?bifo=port (abgerufen am: 29.04.2024); *Heller*, Wie mächtig dürfen Maschinen sein?, FAS vom 18.02.2024, S. 1.
[82] https://www.europarl.europa.eu/news/de/press-room/20240308IPR19015/gesetz-uber-kunstliche-intelligenz-parlament-verabschiedet-wegweisende-regeln (Stand: 23.04.2024).

handelt[83]. Nach einigen Anpassungsvorschlägen durch Rat[84] und Europaparlament[85] unterscheidet sich dieser in einigen Punkten deutlich vom ursprünglichen Kommissionsvorschlag[86].
Eine – voraussichtlich finale – Version der Verordnung ist auf der Seite des europäischen Parlaments abzurufen[87]. Die Verordnung hat es zum Ziel, KI-Praktiken, die mit fundamentalen Werten der EU im Konflikt stehen, einzugrenzen.
Einerseits wird sie als notwendige Regulierungsmaßnahme eines (zu) schnell wachsenden Technologiebereichs gesehen, um die Risiken, die diese Technologie birgt, einzuhegen[88]. Anderseits wird sie jedoch auch als Hemmnis für Innovation betrachtet und man befürchtet, dass eine mögliche Überregulierung zur Abwanderung von Technologieunternehmen führt. Als Verordnung ist sie – ab Inkrafttreten – unmittelbar in den einzelnen Mitgliedsstaaten anwendbar und müsste nicht erst in nationales Recht übersetzt werden.
Die Verordnung untergliedert sich sekundärrechtstypisch in Erwägungsgründe (ErwG), die bei der Auslegung und dem Verständnis des Normtextes helfen, sowie rechtlich verbindliche Normen. Ziel der Verordnung ist es, einheitliche Regeln für die Entwicklung, Verwendung und Vermarktung von KI im Einklang mit den Werten der Union zu etablieren[89]. Gleichzeitig soll sie Schutz vor Gefahren – materieller und immaterieller Art – bieten, die künstliche Intelligenz hervorrufen kann[90].
Der persönliche Anwendungsbereich nach Art. 2 umfasst primär die Anbieter von KI-Systemen, aber auch Betreiber und Einführer. Das statistische Bundesamt wäre ausgehend davon, dass ein ML-Algorithmus selbst und für eigene Zwecke entwickelt wird, als Anbieter eines KI-Systems zu sehen und würde auch räumlich unter die Verordnung fallen, da ein geplantes KI-System innerhalb der EU genutzt würde.
Art. 3 Nr. 1 definiert zuvorderst den Begriff des KI-Systems. Während der Kommissionsentwurf hierbei noch mit einem Verweis auf Annex I arbeitete, definieren sowohl Rat als auch EU-Parlament ein KI-System abschließend: Die Definition soll eine Abgrenzung von klassischer Software ermöglichen[91]. Entscheidend ist, dass ein KI-System maschinenbasiert ist, mit einem gewissen Grad an Autonomie agieren kann und die Fähigkeit besitzt, Informationen basierend auf deren Eingabewerten herzuleiten, die das Umfeld beeinflussen können. Bei Maschinellem Lernen handelt es sich um einen Teilbereich der künstlichen Intelligenz[92]. Eingesetzte ML-Modelle in der amtlichen Statistik würden also generell der KI-Verordnung unterfallen.
Der Entwurf der Kommission sah noch einen sehr weiten Begriff für künstliche Intelligenz vor, unter den auch teilweise normale Software gefallen wäre(Bomhard und Merkle, 2022). Das Verhandlungsergebnis arbeitet nun aber mit einem weniger weitreichenden Begriff, da auch Anhang I der Ursprünglichen Version gestrichen wurde. In der neuen Begriffsdefinition wurde klargestellt, dass Künstliche Intelligenz die Fähigkeit hat, zu lernen bzw. sich anzupassen. Damit handelt es

[83] https://www.europarl.europa.eu/news/de/press-room/20240308IPR19015/gesetz-uber-kunstliche-intelligenz-parlament-verabschiedet-wegweisende-regeln (abgerufen am 16.04.2024).
[84] Abzurufen unter: https://www.consilium.europa.eu/de/press/press-releases/2022/12/06/artificial-intelligence-act-council-calls-for-promoting-safe-ai-that-respects-fundamental-rights/ (Stand: 29.04.2024).
[85] P9_TA(2023)0236 abzurufen unter: https://www.europarl.europa.eu/doceo/document/TA-9-2023-06-14_EN.html#sdocta6 (Stand: 29.04.2024).
[86] COM(2021) 206 final.
[87] https://www.europarl.europa.eu/doceo/document/TA-9-2024-03-13-TOC_DE.html (Stand: 19.04.2024).
[88] Jannis Brühl, SZ Online, v. 09.12.2023; Holzki, Handelsblatt v. 11.12.2023, S. 14.
[89] ErwG 1.
[90] ErwG 4 und 5.
[91] ErwG 12.
[92] Baum in: Leupold/Wiebe/Glossner, IT-Recht, Teil 9.1 Rn. 8.

sich weiterhin um eine weite Definition des Begriffs. Die Gefahr, dass es sich dabei um einen „Software-Act“(Hacker und Berz, 2023) handelt, ist aber nicht mehr vorhanden.

7.5.2 Verbotene KI

Die Verordnung verfolgt einen risikobasierten Ansatz[93] und legt KI-Systemen stärkere Beschränkungen auf, je stärker deren Nutzung Grundrechte gefährdet[94]. Dementsprechend unterscheidet die Verordnung vier Risikolevel: Inakzeptables Risiko, hohes Risko, mittleres Risiko und niedriges Risiko[95](Ebert und Spiecker gen. Döhmann, 2021). Je höher das Risikolevel beim Einsatz eines KI-Systems, desto tiefgehender ist die Regulierung. Mit welchem Risiko ML für statistische Zwecke zu bewerten ist, soll im Folgenden verdeutlicht werden. Ausgegangen wird dabei von den oben erwähnten Methoden der Kodierung, Plausibilisierung sowie Imputation. Dass es sich bei diesen Methoden um verbotene KI-Praktiken nach Art. 5 handelt, lässt sich zweifelsfrei ausschließen. Solche Praktiken sind insbesondere Methoden der kognitiven Verhaltenssteuerung, soziales Scoring, biometrische Identifizierung und Kategorisierung oder biometrische Echtzeit-Fernidentifizierung. All diese Anwendungsfelder entsprechen jedoch nicht den Anwendungsfällen von ML für statistische Zwecke.

7.5.3 Hochrisiko-KI

Der zweiten Risikoklasse (Hochrisiko) unterfallen Systeme, die ein hohes Gefährdungspotential für Grundrechte haben[96]. Als „Hauptadressat“ der Verordnung unterwirft die Verordnung deren Anbieter zahlreichen Einschränkungen nach den Art. 8 ff. Derartige KI-Systeme sind zum einen solche, die in Produkten verwendet werden, die unter Produktsicherheitsvorschriften der EU fallen, also etwa Luftfahrt, Fahrzeuge, medizinische Geräte und Spielzeug(Bomhard und Siglmüller, 2024). Zum anderen sind es Systeme, die in spezifische Bereiche fallen, welche in Annex III der Verordnung genannt werden[97].
Bei den Anwendungsfällen des ML in der Statistik handelt es sich um Unterstützungsmethoden der menschlichen Auswertung und Aufbereitung von statistischen Erhebungsdaten. Es ist nicht ersichtlich, inwieweit sie unter einen Fall dieser Hochrisiko-KI zu subsumieren wäre. Seit dem Trilogverfahren im Gesetzgebungsverfahren zur KI-Verordnung haben zudem generelle Ausnahmen für Hochrisiko KI-Systeme Einzug in die KI-Verordnung gefunden, wenn die Anwendung kein erhebliches Risiko für bestimmte Rechtsgüter darstellt(Bomhard und Siglmüller, 2024). Explizit ausgenommen sind demnach erstens KI-Systeme, die einen sehr engen Aufgabenbereich innerhalb eines Prozesses haben, zweitens Systeme, die das Ziel haben, ein vorher rein menschlich durchgeführtes Ergebnis zu verbessern oder drittens solche, die Entscheidungsmuster oder Abweichungen davon erkennen sollen, aber nicht dazu bestimmt sind, menschliche Entscheidungen zu beeinflussen oder zu ersetzen.
Unabhängig davon, dass die erwähnten Einsatzbereiche von ML in der amtlichen Statistik nur

[93] ErwG 26.
[94] ErwG 5.
[95] COM(2021) 206 final, S. 15. Wobei zwischen den beiden unteren Risikokategorien nicht trennscharf abgegrenzt wird.
[96] ErwG 46.
[97] Hierunter fallen vor allem Systeme zum Einsatz in den folgenden Bereichen: Biometrische Identifizierung und Kategorisierung, Verwaltung und Betrieb kritischer Infrastrukturen, allgemeine und berufliche Bildung, Beschäftigung, Personalmanagement und Zugang zur Selbstständigkeit, Zugänglichkeit und Inanspruchnahme grundlegender privater und öffentlicher Dienste und Leistungen, Strafverfolgung, Mirgration, Asyl und Grenzkontrolle, Rechtspflege und demokratische Prozesse.

schwer als Hochrisiko-KI zu klassifizieren wären, ließe sich argumentieren, dass Klassifikation oder Plausibilisierung zur Datenaufbereitung ohnehin nur ein enges Anwendungsfeld darstellen oder der Verbesserung von vorher rein menschlich durchgeführten Verfahren dienen, was eine Ausnahme rechtfertigen würde.

7.5.4 Sonstige Pflichten

Die Einhaltung ethischer Grundsätze ist in den Erwägungsgründen[98] der Verordnung für die Entwickler von KI-Systemen vorgesehen(Bomhard und Siglmüller, 2024). Den Erwägungsgründen kommt allerdings keine rechtliche Verbindlichkeit zu[99]. Das Einhalten ethischer Grundsätze wurde auch in die Selbstverpflichtungen des Art. 95 aufgenommen. Mit ethischen Grundsätzen im Sinne dieser Verordnung sind dabei die Grundsätze gemeint, die von der hochrangigen Expertengruppe für künstliche Intelligenz (High Level Expert Group, HLEG) 2019 aufgestellt wurden[100]. Diese Grundsätze umfassen menschliches Handeln und Aufsicht, technische Robustheit und Sicherheit, Datenschutz und Datenverwaltung, Transparenz, Vielfalt, Nichtdiskriminierung und Fairness, gesellschaftliches und ökologisches Wohlergehen und Verantwortlichkeit. Zwischen diesen Grundsätzen den Grundsätzen aus dem europäischen Statistikkodex existieren teilweise bereits Überschneidungen, etwa beim Datenschutz und der Transparenz. Ebenso lassen sich Parallelen zu den Standards eines Quality Framework for Statistical Machine Learning erkennen(Yung u. a., 2022), insbesondere bei den Punkten Erklärbarkeit (Explainibility) und Transparenz. Ein neuer Punkt, der seit dem Parlamentsvorschlag in die Verordnung aufgenommen wurde, ist die Anforderung an KI-Kompetenz für Mitarbeitende und Dritte, die mit dem KI-System arbeiten bzw. zu dessen Betrieb eingesetzt werden[101]. KI-Kompetenz bezeichnet dabei die Kenntnis über die Chancen und Risiken sowie möglichen Schäden, die die Nutzung von AI nach sich zieht, um informierte Entscheidungen beim Umgang mit einem KI-System treffen zu können[102]. Daneben sollen das Amt für künstliche Intelligenz (Art. 3 Nr. 47, ErwG) und die Mitgliedsstaaten die Aufstellung von Verhaltenskodizes in Bezug auf die Anwendung freiwilliger Standards im Bereich der künstlichen Intelligenz fördern und erleichtern (Art. 95). Dies soll einen Anreiz zur Einführung solcher Selbstverpflichtungen darstellen, welche die Regelungen von Kapitel III Abschnitt 2 der Verordnung umfassen, die auch für Hochrisiko-KI gelten (Art. 8ff.). Die Selbstverpflichtungen durch Verhaltenskodizes sollen nach Art. 95 Abs. 2 dabei klar messbare Ziele haben wie:

- Elemente der Ethik-Leitlinien für eine vertrauenswürdige KI,
- Beurteilung und Minimierung der Auswirkungen von KI-Systemen in Bezug auf deren Energieverbrauch bei der Entwicklung, dem Training und der Benutzung,
- die Unterstützung von Maßnahmen, die der KI-Kompetenz dienen,
- inklusive Gestaltung von AI-Systemen, unter anderem durch vielseitige Entwicklungsteams,
- Gefahrenprävention beim Einsatz von KI-System, insbesondere diskriminierungsfreie Ausgestaltung von KI-Systemen und Verhinderung von negativen Auswirkungen auf schutzbedürftige Personen.

[98] ErwG 27.
[99] EuGH Urt. v. 19.6.2014 Rs. C-345/13, ECLI:EU:C:2014:2013 Rn. 31 – *Karen Millen Fashions*.
[100] Ethics Guidelines for Trustworthy AI. Abzurufen unter: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (15.03.2024).
[101] ErwG 14a bzw. Art. 4a Abs. 1 d) des Parlamentsvorschlag.
[102] Vgl. Art. 3, Nr. 56.

Es handelt sich hierbei um reine Selbstverpflichtungen, ein Anreiz zur Implementierung könnte dabei im gesteigerten Vertrauen der Personen liegen, deren Daten durch ein KI-System genutzt werden. Daneben könnten sich Reputationsvorteile für das statistische Bundesamt ergeben und gegebenenfalls Überschneidungen mit bereits bestehenden Complianceanforderungen oder Qualitätsstandards ausgenutzt werden (Linardatos, 2023).

Insgesamt lässt sich daher festhalten, dass ML in der amtlichen Statistik der KI-Verordnung unterfallen dürfte. Bei den geplanten Anwendungen handelt es sich aber um KI-Methoden mit einem geringen Risiko. Sie sind daher nicht durch die Verordnung verboten und auch die hohen Anforderungen, die die Verordnung an Hochrisiko-KI stellt, treffen sie nicht. Daher sind die Anforderungen an das statistische Bundesamt als Anbieter eines KI-Systems vor allem Schulungen der zuständigen Mitarbeitenden im Bereich der KI-Kompetenz. Daneben ist das Etablieren eines Verhaltenskodex, der sich an den Ethik-Guidelines der HLEG orientiert, in Erwägung zu ziehen, wenngleich es sich hierbei nicht um eine Rechtspflicht handelt. Dabei ergeben sich teilweise bereits Überschneidungen mit anderen Qualitätsstandards der amtlichen Statistik.

7.6 Fazit und Ausblick

Das geltende Datenschutzrecht sowie die zu erwartende KI-Verordnung sollten zu keinen starken Beschränkungen für dein Einsatz von ML bei der Datenaufbereitung in der amtlichen Statistik führen. Aus datenschutzrechtlicher Sicht erscheint die Nutzung von anonymen Daten sinnvoll, wenn ML-Systeme trainiert werden, um einen Eingriff in das Recht auf informationelle Selbstbestimmung möglichst niedrig zu halten. Das Verbot der Reidentifizierung des Bundesstatistikgesetzes stellt zudem aber eine geeignete Garantie im Sinne der DSGVO dar. Gleichzeitig ist zu hoffen, dass durch die intelligente Datenverarbeitung weniger Rückfragen bei den Befragten einer statistischen Erhebung notwendig sind. Dadurch verringert sich die Eingriffstiefe in das Recht auf informationelle Selbstbestimmung. Hinsichtlich der KI-Verordnung ergeben sich niedrige Auflagen für statistische Ämter, sofern sie ML zur Datenaufbereitung verwenden. Schulungen im Bereich der KI-Kompetenz sind bereits ohne die KI-Verordnung voraussichtlich notwendig. Verpflichtungen aus dem Soft Law der Verordnung dürften für statistische Ämter zudem zu Synergieeffekten führen, da sich hier Überschneidungen mit anderen Qualitätsstandards aufzeigen.

Literatur

Abdar, M., F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya u. a. (2021). „A review of uncertainty quantification in deep learning: Techniques, applications and challenges“. In: *Information Fusion* 76, S. 243–297.

Allhutter, D., F. Cech, F. Fischer, G. Grill und A. Mager (2020). „Algorithmic Profiling of Job Seekers in Austria: How Austerity Politics Are Made Effective“. In: *Frontiers in Big Data* 3, S. 1–17. DOI: 10.3389/fdata.2020.00005.

Amaya, A., P. P. Biemer und D. Kinyon (2020). „Total Error in a Big Data World: Adapting the TSE Framework to Big Data“. In: *Journal of Survey Statistics and Methodology* 8.1, S. 89–119.

Angwin, J., S. Mattu und L. Kirchner (23. Mai 2016). *Machine Bias*. ProPublica. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Athey, S. und G. Imbens (2016). „Recursive partitioning for heterogeneous causal effects“. In: *Proceedings of the National Academy of Sciences* 113.27, S. 7353–7360. DOI: 10.1073/pnas.1510489113.

Augustin, T., F. P. Coolen, G. De Cooman und M. C. Troffaes, Hrsg. (2014). *Introduction to imprecise probabilities*. Bd. 591. John Wiley & Sons.

Austern, M. und W. Zhou (2020). *Asymptotics of cross-validation*. arXiv: 2001.11111.

Barocas, S., M. Hardt und A. Narayanan (2023). *Fairness and Machine Learning: Limitations and Opportunities*. Cambridge, MA: MIT Press. URL: www.fairmlbook.org.

Barocas, S. und A. D. Selbst (2016). „Big Data's Disparate Impact“. In: *California Law Review* 104.3, S. 671–732.

Bates, S., T. Hastie und R. Tibshirani (2021). *Cross-Validation: What Does It Estimate and How Well Does It Do It?* arXiv: 2104.00673.

Bayle, P., A. Bayle, L. Janson und L. Mackey (2020). *Cross-validation Confidence Intervals for Test Error*. arXiv: 2007.12671.

Beck, M., F. Dumpert und J. Feuerhake (2018). *Proof of Concept Machine Learning*. Abschlussbericht. Wiesbaden: Statistisches Bundesamt (Destatis).

Bhatt, U., Y. Zhang, J. Antorán, Q. V. Liao, P. Sattigeri, R. Fogliato, G. G. Melançon, R. Krishnan, J. Stanley, O. Tickoo, L. Nachman, R. Chunara, A. Weller und A. Xiang (2020). *Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty*. arXiv: 2011.07586.

Boehme-Neßler, V. (2016). „Das Ende der Anonymität – Wie Big Data das Datenschutzrecht verändert“. In: *Datenschutz und Datensicherheit*, S. 419–423.

Bogdan, P. C., F. Dolcos, M. Moore, I. Kuznietsov, S. A. Culpepper und S. Dolcos (2023). „Social Expectations are Primarily Rooted in Reciprocity: An Investigation of Fairness, Cooperation, and Trustworthiness“. In: *Cognitive Science* 47.8, e13326. DOI: 10.1111/cogs.13326.

Bomhard, D. und M. Merkle (2022). „Europäische KI-Verordnung – Der aktuelle Kommissionsentwurf und praktische Auswirkungen“. In: *Recht Digital (RDi)*, S. 276–283.

Bomhard, D. und J. Siglmüller (2024). „AI Act – das Trilogergebnis“. In: *Recht Digital (RDi)*, S. 45–96.

Bothmann, L., S. Dandl und M. Schomaker (2023). „Causal Fair Machine Learning via Rank-Preserving Interventional Distributions“. en. In: *Proceedings of the 1st Workshop on Fairness and Bias in AI co-located with 26th European Conference on Artificial Intelligence (ECAI 2023)*. CEUR Workshop Proceedings. URL: https://ceur-ws.org/Vol-3523/.

Bothmann, L. und K. Peters (2024). „Fairness als Qualitätskriterium im Maschinellen Lernen – Rekonstruktion des philosophischen Konzepts und Implikationen für die Nutzung außergesetzlicher

Merkmale bei qualifizierten Mietspiegeln". In: *AStA Wirtschafts- und Sozialstatistisches Archiv*. To appear.

Bothmann, L., K. Peters und B. Bischl (2024). *What is fairness? On the role of protected attributes and fictitious worlds*. arXiv: 2205.09622.

Breiman, L. (2001a). „Random Forests". In: *Machine Learning* 45.1, S. 5–32. DOI: 10.1023/a:1010933404324.

— (2001b). „Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author)". In: *Statistical Science* 16.3, S. 199–231.

Buolamwini, J. und T. Gebru (2018). „Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification". In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. Hrsg. von S. A. Friedler und C. Wilson. Bd. 81. Proceedings of Machine Learning Research. PMLR, S. 77–91. URL: https://proceedings.mlr.press/v81/buolamwini18a.html.

Callies, C. und M. Ruffert, Hrsg. (2022). *EUV/AEUV, Kommentar*. 6. Aufl. München: C.H.Beck.

Caprio, M., S. Dutta, K. J. Jang, V. Lin, R. Ivanov, O. Sokolsky und I. Lee (2023). *Imprecise Bayesian neural networks*. arXiv: 2302.09656.

Caton, S. und C. Haas (2024). „Fairness in Machine Learning: A Survey". In: *ACM Computing Surveys* 56.7, S. 1–38. DOI: 10.1145/3616865.

Caton, S., S. Malisetty und C. Haas (19. Sep. 2022). „Impact of Imputation Strategies on Fairness in Machine Learning". In: *Journal of Artificial Intelligence Research* 74. DOI: 10.1613/jair.1.13197.

Chakraborty, T., C. Seifert und C. Wirth (2024). *Explainable Bayesian Optimization*. arXiv: 2401.13334.

Chaudhuri, A. (1978). „On estimating the variance of a finite population". In: *Metrika* 25.1, S. 65–76.

Choi, I., A. del Monaco, E. Law, S. Davies, J. Karanka, A. Baily, R. Piela, T. Turpeinen, A. Mharzi, S. Rastan, K. Flak und S. Jentoft (2022). *ML Model Monitoring and Re-training in Statistical Organisations*. ONS-UNECE Machine Learning Group 2022, Theme Group - Model Retraining, v2, available at https://statswiki.unece.org/display/ML/Machine+Learning+Group+2022.

Chouldechova, A. (Okt. 2016). *Fair prediction with disparate impact: A study of bias in recidivism prediction instruments*. arXiv: 1610.07524.

Couper, M. P., E. Singer, F. G. Conrad und R. M. Groves (2008). „Risk of disclosure, perceptions of risk, and concerns about privacy and confidentiality as factors in survey participation". In: *Journal of Official Statistics* 24.2, S. 255–275.

Courbois, J.-Y. P. und N. S. Urquhart (2004). „Comparison of survey estimates of the finite population variance". In: *Journal of Agricultural, Biological, and Environmental Statistics* 9.2, S. 236–251.

Dandl, S., M. Becker, B. Bischl, G. Casalicchio und L. Bothmann (2024). *mlr3summary: Concise and interpretable summaries for machine learning models*. Accepted at 2nd World Conference on eXplainable Artificial Intelligence 2024 (Demo Track). arXiv: 2404.16899.

Dandl, S., G. Casalicchio, B. Bischl und L. Bothmann (2023). „Interpretable Regional Descriptors: Hyperbox-Based Local Explanations". In: *Machine Learning and Knowledge Discovery in Databases: Research Track*. Hrsg. von D. Koutra, C. Plant, M. Gomez Rodriguez, E. Baralis und F. Bonchi. Cham: Springer Nature Switzerland, S. 479–495.

Demšar, J. (2006). „Statistical Comparisons of Classifiers over Multiple Data Sets". In: *Journal of Machine Learning Research* 7.1, S. 1–30.

Dietrich, S., J. Rodemann und C. Jansen (2024). *Semi-Supervised Learning guided by the Generalized Bayes Rule under Soft Revision*. Accepted at the 11th International Conference on Soft Methods in Probability and Statistics (SMPS). arXiv: 2405.15294.

Dietterich, T. G. (1998). „Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms“. In: *Neural Computation* 10.7, S. 1895–1923. DOI: 10.1162/089976698300017197.

Dorer, P., H. Mainusch und H. Tubies, Hrsg. (1988). *Bundestatistikgesetz - Kommentar*. 1. Aufl. München: C.H.Beck.

Doshi-Velez, F. und B. Kim (2017). *Towards A Rigorous Science of Interpretable Machine Learning*. arXiv: 1702.08608v2.

Dumpert, F. (2021). „Machine Learning in der amtlichen Statistik–Ergebnisse und Bewertung eines internationalen Projekts“. In: *WISTA–Wirtschaft und Statistik* 73.4, S. 53–63.

Dwork, C., M. Hardt, T. Pitassi, O. Reingold und R. Zemel (2012). „Fairness through awareness“. In: *ITCS '12: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. Cambridge, MA: ACM Press, S. 214–226. DOI: 10.1145/2090236.2090255.

Ebert, A. und I. Spiecker gen. Döhmann (2021). „Der Kommissionsentwurf für eine KI-Verordnung der EU“. In: *Neue Zeitschrift für Verwaltungsrecht*, S. 1188–1193.

Efron, B. und R. Tibshirani (1997). „Improvements on Cross-Validation: The .632+ Bootstrap Method“. In: *Journal of the American Statistical Association* 92.438, S. 548–560.

Epping, V. und C. Hillgruber, Hrsg. (2024). *BeckOK Grundgesetz - Kommentar*. 57. Ed. München: C.H.Beck.

Ewald, F. K., L. Bothmann, M. N. Wright, B. Bischl, G. Casalicchio und G. König (2024). *A Guide to Feature Importance Methods for Scientific Inference*. arXiv: 2404.12862.

Fisher, A., C. Rudin und F. Dominici (2019). „All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously“. In: *Journal of Machine Learning Research* 20.177.

Frosst, N. und G. Hinton (2017). *Distilling a Neural Network Into a Soft Decision Tree*. arXiv: 1711.09784.

Goschenhofer, J. (2023). „Reducing the effort for data annotation: contributions to weakly supervised deep learning“. Diss. München, Ludwig-Maximilians-Universität, 2023.

Groves, R. M., F. J. Fowler Jr, M. P. Couper, J. M. Lepkowski, E. Singer und R. Tourangeau (2009). *Survey Methodology*. 2. Auflage. Hoboken, NJ: John Wiley & Sons.

Gruber, C., P. O. Schenk, M. Schierholz, F. Kreuter und G. Kauermann (2023). *Sources of Uncertainty in Machine Learning – A Statisticians' View*. arXiv: 2305.16703.

Hacker, P. und A. Berz (2023). „Der AI Act der Europäischen Union – Überblick, Kritik und Ausblick“. In: *Zeitschrift für Rechtspolitik*, S. 226–229.

Hebert-Johnson, U., M. Kim, O. Reingold und G. Rothblum (2018). „Multicalibration: Calibration for the (Computationally-Identifiable) Masses“. In: *Proceedings of the 35th International Conference on Machine Learning*. Hrsg. von J. Dy und A. Krause. Bd. 80. Proceedings of Machine Learning Research. PMLR, S. 1939–1948. URL: https://proceedings.mlr.press/v80/hebert-johnson18a.html.

Hodel, M. (2024). „A simulation study based on: "Evaluating Machine Learning Models in Non-Standard Settings: An Overview and New Findings" by Hornung et al.“ Seminar Paper for 'Machine Learning in Official Statistics — Methodological Perspectives and Challenges' by F. Dumpert and T. Augustin. LMU Munich.

Hornung, R., M. Nalenz, L. Schneider, A. Bender, L. Bothmann, B. Bischl, T. Augustin und A.-L. Boulesteix (2023). *Evaluating machine learning models in non-standard settings: An overview and new findings*. arXiv: 2310.15108.

Hüllermeier, E. (2014). „Learning from imprecise and fuzzy observations: Data disambiguation through generalized loss minimization". In: *International Journal of Approximate Reasoning* 55.7, S. 1519–1534.

Hüllermeier, E., S. Destercke und M. H. Shaker (2022). „Quantification of credal uncertainty in machine learning: A critical analysis and empirical comparison". In: *Uncertainty in Artificial Intelligence*. PMLR, S. 548–557.

Hüllermeier, E. und W. Waegeman (2021). „Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods". In: *Machine Learning* 110.3, S. 457–506.

Hume, D. (1748). *An Enquiry Concerning Human Understanding*.

Jansen, C., G. Schollmeyer und T. Augustin (2018). „Concepts for decision making under severe uncertainty with partial ordinal and partial cardinal preferences". In: *International Journal of Approximate Reasoning* 98, S. 112–131.

Jansen, C., M. Nalenz, G. Schollmeyer und T. Augustin (2023a). „Statistical Comparisons of Classifiers by Generalized Stochastic Dominance". In: *Journal of Machine Learning Research* 24, S. 1–37.

Jansen, C., G. Schollmeyer, H. Blocher, J. Rodemann und T. Augustin (2023b). „Robust statistical comparison of random variables with locally varying scale of measurement". In: *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*. Hrsg. von R. J. Evans und I. Shpitser. Bd. 216. Proceedings of Machine Learning Research. PMLR, S. 941–952.

Jansen, C., G. Schollmeyer, J. Rodemann, H. Blocher und T. Augustin (2024). *Statistical Multicriteria Benchmarking via the GSD-Front*. arXiv: 2406.03924 [stat.ML]. URL: https://arxiv.org/abs/2406.03924.

Jiang, W., S. Varma und R. Simon (2008). „Calculating confidence intervals for prediction error in microarray classification using resampling". In: *Statistical Applications in Genetics and Molecular Biology* 7.1.

Kaiser, P., C. Kern und D. Rügamer (2022). *Uncertainty-aware predictive modeling for fair data-driven decisions*. arXiv: 2211.02730.

Karl, F., S. R. Kaminwar und H. Frechen (2024). *Beschreibung einer MLOps-Architektur: Abschlussbericht und Empfehlung für das Statistische Bundesamt*. Techn. Ber. Fraunhofer IIS, Supply Chain Services.

Kearns, M., S. Neel, A. Roth und Z. S. Wu (2018). „Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness". In: *Proceedings of the 35th International Conference on Machine Learning*. Hrsg. von J. Dy und A. Krause. Bd. 80. Proceedings of Machine Learning Research. PMLR, S. 2564–2572. URL: https://proceedings.mlr.press/v80/kearns18a.html.

Kilbertus, N., M. Rojas-Carulla, G. Parascandolo, M. Hardt, D. Janzing und B. Schölkopf (2017). „Avoiding Discrimination through Causal Reasoning". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Red Hook, NY: Curran Associates Inc., S. 656–666. URL: https://proceedings.neurips.cc/paper_files/paper/2017/hash/f5f8590cd58a54e94377e6ae2eded4d9-Abstract.html.

Kim, M. P., A. Ghorbani und J. Zou (2019). „Multiaccuracy: Black-Box Post-Processing for Fairness in Classification". In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '19. New York, NY: Association for Computing Machinery, S. 247–254. DOI: 10.1145/3306618.3314287.

Kühling, J., Hrsg. (2023). *Bundestatistikgesetz - Kommentar*. 1. Aufl. München: C.H.Beck.

Kühling, J. und B. Buchner, Hrsg. (2024). *Datenschutzgrundverordnung/BDSG - Kommentar*. 4. Aufl. München: C.H.Beck.

Kuppler, M., C. Kern, R. Bach und F. Kreuter (2022). „From fair predictions to just decisions? Conceptualizing algorithmic fairness and distributive justice in the context of data-driven decision-making". In: *Frontiers in Sociology* 7. DOI: 10.3389/fsoc.2022.883999.

Lei, J., M. G'Sell, A. Rinaldo, R. J. Tibshirani und L. Wasserman (2018). „Distribution-Free Predictive Inference for Regression". In: *Journal of the American Statistical Association* 113.523, S. 1094–1111. DOI: 10.1080/01621459.2017.1307116.

Lewis, D. K. (1973). *Counterfactuals*. Malden, MA: Blackwell.

Linardatos, D. (2023). „§ 7 – Qualitätskontrolle, Korrekturmechanismen und Code of Conduct". In: *Die neue Verordnung der EU zur Künstlichen Intelligenz*. Hilgendorf, Eric und Roth-Isigkeit, David, S. 125–143.

Lohr, S. L. (2021). *Sampling: Design and Analysis*. 3rd. Boca Raton, FL, USA: CRC Press.

Lumley, T. (2004). „Analysis of complex survey samples". In: *Journal of Statistical Software* 9, S. 1–19.

Lumley, T. und A. Scott (2017). „Fitting Regression Models to Survey Data". In: *Statistical Science* 32.2, S. 265–278.

MacNell, N., L. Feinstein, J. Wilkerson, P. M. Salo, S. A. Molsberry, M. B. Fessler, P. S. Thorne, A. A. Motsinger-Reif und D. C. Zeldin (2023). „Implementing machine learning methods with complex survey data: Lessons learned on the impacts of accounting sampling weights in gradient boosting". In: *PLoS One* 18.1, e0280387.

Makhlouf, K., S. Zhioua und C. Palamidessi (2021). „Machine learning fairness notions: Bridging the gap with real-world applications". In: *Information Processing & Management* 58.5. arXiv preprint previously titled On the Applicability of ML Fairness Notions. DOI: 10.1016/j.ipm.2021.102642.

Mangili, F. (2016). „A prior near-ignorance Gaussian process model for nonparametric regression". In: *International Journal of Approximate Reasoning* 78, S. 153–171.

Manski, C. F. (2003). *Partial identification of probability distributions*. Springer Science & Business Media.

Mehrabi, N., F. Morstatter, N. Saxena, K. Lerman und A. Galstyan (2021). „A Survey on Bias and Fairness in Machine Learning". In: *ACM Computing Surveys* 54.6. DOI: 10.1145/3457607.

Mitchell, S., E. Potash, S. Barocas, A. D'Amour und K. Lum (2021). „Algorithmic Fairness: Choices, Assumptions, and Definitions". In: *Annual Review of Statistics and Its Application* 8.1, S. 141–163. DOI: 10.1146/annurev-statistics-042720-125902.

Molladavoudi, S. und W. Yung (2023). „Exploring quality dimensions in trustworthy Machine Learning in the context of official statistics: model explainability and uncertainty quantification". In: *AStA Wirtschafts-und Sozialstatistisches Archiv* 17.3, S. 223–252.

Molnar, C. (2022). *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. 2. Aufl. URL: https://christophm.github.io/interpretable-ml-book.

Molnar, C., G. König, J. Herbinger, T. Freiesleben, S. Dandl, C. A. Scholbeck, G. Casalicchio, M. Grosse-Wentrup und B. Bischl (2022). „General Pitfalls of Model-Agnostic Interpretation Methods for Machine Learning Models". In: *xxAI - Beyond Explainable AI: International Workshop*. Hrsg. von A. Holzinger und et al. Cham: Springer International Publishing, S. 39–68. DOI: 10.1007/978-3-031-04083-2_4.

Moosbauer, J., J. Herbinger, G. Casalicchio, M. Lindauer und B. Bischl (2021). „Explaining hyperparameter optimization via partial dependence plots“. In: *Advances in Neural Information Processing Systems* 34, S. 2280–2291.

Nadeau, C. und Y. Bengio (1. Sep. 2003). „Inference for the Generalization Error“. In: *Machine Learning* 52.3, S. 239–281. DOI: 10.1023/A:1024068626366.

Nahorniak, M., D. P. Larsen, C. Volk und C. E. Jordan (2015). „Using inverse probability bootstrap sampling to eliminate sample induced bias in model based analysis of unequal probability samples“. In: *PLoS One* 10.6, e0131765.

Nalenz, M., J. Rodemann und T. Augustin (2024). „Learning De-Biased Regression Trees and Forests from Complex Samples“. In: *Machine Learning* 113.6, S. 3379–3398.

Narayanan, A. und V. Shmatikov (2008). „Robust de-anonymization of large sparse datasets“. In: *2008 IEEE Symposium on Security and Privacy (sp 2008)*. IEEE, S. 111–125.

Nink, J. und J. Pohle (2015). „Die Bestimmbarkeit des Personenbezugs–Von der IP-Adresse zum Anwendungsbereich der Datenschutzgesetze“. In: *Zeitschrift für IT-Recht und Digitalisierung (MMR)* 18.9, S. 563–567.

Noma, H., T. Shinozaki, K. Iba, S. Teramukai und T. A. Furukawa (2021). „Confidence intervals of prediction accuracy measures for multivariable prediction models based on the bootstrap-based optimism correction methods“. In: *Statistics in Medicine* 40.26, S. 5691–5701.

Obermeyer, Z., B. Powers, C. Vogeli und S. Mullainathan (2019). „Dissecting racial bias in an algorithm used to manage the health of populations“. In: *Science* 366.6464, S. 447–453. DOI: 10.1126/science.aax2342.

Perdomo, J. C., T. Zrnic, C. Mendler-Dünner und M. Hardt (2020). *Performative Prediction*. arXiv: 2002.06673.

Pfeffermann, D., Hrsg. (2009). *Handbook of Statistics, Volume 29B; Sample Surveys: Inference and Analysis*. Amsterdam: Elsevier.

Pfisterer, F., C. Kern, S. Dandl, M. Sun, M. P. Kim und B. Bischl (2021). „mcboost: Multi-Calibration Boosting for R“. In: *Journal of Open Source Software* 6.64, S. 3453. DOI: 10.21105/joss.03453.

Pfisterer, F., W. Siyi und M. Lang (2024). *mlr3fairness: Fairness Auditing and Debiasing for 'mlr3'*. R package version 0.3.2. URL: https://mlr3fairness.mlr-org.com.

Preising, M., K. Lange und F. Dumpert (2021). „Imputation zur maschinellen Behandlung fehlender und unplausibler Werte in der amtlichen Statistik“. In: *WISTA-Wirtschaft und Statistik* 73.5, S. 40–52.

Radermacher, W. J. (2017). „Governance in der amtlichen Statistik“. In: *AStA Wirtschafts-und Sozialstatistisches Archiv* 11.2, S. 65–81.

Ramge, T. (2. Feb. 2018). *Mensch fragt, Maschine antwortet*. URL: https://www.bpb.de/shop/zeitschriften/apuz/263680/mensch-fragt-maschine-antwortet/.

Rodemann, J. (2023). „Pseudo Label Selection is a Decision Problem“. In: *Proc. of the 46th German Conference on Artificial Intelligence*. Springer.

Rodemann, J. und T. Augustin (2024). „Imprecise Bayesian Optimization“. In: *Knowledge-based Systemes*. forthcoming.

Rodemann, J., J. Goschenhofer, E. Dorigatti, T. Nagler und T. Augustin (2023a). *Bayesian PLS! Approximate Bayes Optimal Pseudo-Label Selection (PLS)*. URL: https://arxiv.org/abs/2302.08883.

Rodemann, J., C. Jansen, G. Schollmeyer und T. Augustin (2023b). *In all LikelihoodS: How to Reliably Select Pseudo-Labeled Data for Self-Training in Semi-Supervised Learning*. arXiv: 2303.01117. (in Begutachtung).

Rodemann, J., D. Kreiss und T. Hüllermeier E. Augustin (2022). „Levelwise Data Disambiguation by Cautious Superset Classification“. In: *Scalable Uncertainty Management: Proceedings of SUM 2022*. Hrsg. von F. Dupin de Saint-Cyr, M. Öztürk-Escoffier und N. Potyka. Lecture Notes in Computer Science, vol 13562. Cham: Springer, S. 263–276.

Rodemann, J. und T. Augustin (2022). „Accounting for Gaussian process imprecision in Bayesian optimization“. In: *International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making*. Springer, S. 92–104.

Rodemann, J., F. Croppi, P. Arens, Y. Sale, J. Herbinger, B. Bischl, E. Hüllermeier, T. Augustin, C. J. Walsh und G. Casalicchio (2024). *Explaining Bayesian Optimization by Shapley Values Facilitates Human-AI Collaboration*. arXiv: 2403.04629.

Rodemann, J., J. Goschenhofer, E. Dorigatti, T. Nagler und T. Augustin (2023c). „Approximately Bayes-Optimal Pseudo-Label Selection“. In: *39th Conference on Uncertainty in Artificial Intelligence (UAI): Pittsburgh, USA*. Hrsg. von R. J. Evans und I. Shpitser. Proceedings of Machine Learning Research, Bd. 216, S. 1762–1773.

Rodemann, J., C. Jansen, G. Schollmeyer und T. Augustin (2023d). „In all likelihoods: robust selection of pseudo-labeled data“. In: *13th International Symposium on Imprecise Probability: Theories and Applications (ISIPTA): Oviedo, Spain*. Hrsg. von E. Miranda, I. Montes, E. Quaeghebeur und B. Vantaggi. Proceedings of Machine Learning Research, Bd. 215, S. 412–425.

Rodolfa, K. T., P. Saleiro und R. Ghani (2020). „Bias and Fairness“. In: *Big Data and Social Science*. Hrsg. von I. Foster, R. Ghani, R. S. Jarmin, F. Kreuter und J. Lane. 2. Auflage. Boca Raton, FL: CRC Press. Kap. 7. URL: https://textbook.coleridgeinitiative.org.

Sale, Y., M. Caprio und E. Hüllermeier (2023). „Is the volume of a credal set a good measure for epistemic uncertainty?“ In: *Uncertainty in Artificial Intelligence*. PMLR, S. 1795–1804.

Särndal, C.-E., B. Swensson und J. Wretman (2003). *Model assisted survey sampling*. New York, NY, USA: Springer.

Sarunski, M. (2016). „Big Data–Ende der Anonymität? Fragen aus Sicht der Datenschutzaufsichtsbehörde Mecklenburg-Vorpommern“. In: *Datenschutz und Datensicherheit-DuD* 40, S. 424–427.

Scantamburlo, T., J. Baumann und C. Heitz (2024). „On Prediction-Modelers and Decision-Makers: Why Fairness Requires More Than a Fair Prediction Model“. In: *AI & Society*. DOI: 10.1007/s00146-024-01886-3.

Schantz, P., H. A. Wolff u. a. (2017). *Das neue Datenschutzrecht*. München: C.H.Beck.

Schenk, P. O. und C. Kern (2024). *Connecting Algorithmic Fairness to Quality Dimensions in Machine Learning in Official Statistics and Survey Production*. (unter Begutachtung in AStA Wirtschafts- und Sozialstatistisches Archiv). arXiv: 2402.09328.

Schouten, B., F. Cobben und J. Bethlehem (2009). „Indicators for the representativeness of survey response“. In: *Survey Methodology* 35.1, S. 101–113. URL: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2009001/article/10887-eng.pdf.

Schulz-Kümpel, H., S. Fischer, A. Bender, B. Bischl, A.-L. Boulesteix, T. Nagler, L. Schneider, A. Stöcker und R. Hornung (2024). *Constructing confidence intervals for "the" Generalization Error*. (in Vorbereitung).

Simitis, S. (2000). „Das Volkszählungsurteil oder der lange Weg zur Informationsaskese-(BVerfGE 65, 1)“. In: *Kritische Vierteljahresschrift für Gesetzgebung und Rechtswissenschaft (KritV)* 83.3/4, S. 359–375.

Simitis, S., G. Hornung und I. Spiecker gen. Döhmann (2019). *Datenschutzrecht - Kommentar*. 1. Aufl. München: C.H.Beck.

Simonyan, K., A. Vedaldi und A. Zisserman (2013). *Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps*. arXiv: 1312.6034.

Simson, J., A. Fabris und C. Kern (2024). „Lazy Data Practices Harm Fairness Research“. In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '24. New York, NY, USA: Association for Computing Machinery, S. 642–659. DOI: 10.1145/3630106.3658931.

Simson, J., F. Pfisterer und C. Kern (2024). „One Model Many Scores: Using Multiverse Analysis to Prevent Fairness Hacking and Evaluate the Influence of Model Design Decisions“. In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '24. New York, NY, USA: Association for Computing Machinery, S. 1305–1320. DOI: 10.1145/3630106.3658974.

Song, C., T. Ristenpart und V. Shmatikov (2017). „Machine learning models that remember too much“. In: *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, S. 587–601.

Statistische Ämter des Bundes und der Länder (2021). *Qualitätshandbuch der Statistischen Ämter des Bundes und der Länder*. Statistische Ämter des Bundes und der Länder.

Strobl, C., A.-L. Boulesteix, T. Kneib, T. Augustin und A. Zeileis (2008). „Conditional variable importance for random forests“. In: *BMC Bioinformatics* 9.1. DOI: 10.1186/1471-2105-9-307.

Sweeney, L. (2002). „k-anonymity: a model for protecting privacy“. In: *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems* 10.5, S. 557–570. DOI: 10.1142/S0218488502001648.

Therneau, T. und B. Atkinson (2022). *rpart: Recursive partitioning and regression trees*. R package version 4.1.19. URL: https://CRAN.R-project.org/package=rpart.

Thiel, G. und M.-C. Puth (2023). „Der Zensus der Zukunft: Registerzensus – Die Leitlinien des Bundesverfassungsgerichts als Maßstab des Registerzensus“. In: *Neue Zeitschrift für Verwaltungsrecht*, S. 305–308.

Toth, D. und J. L. Eltinge (2011). „Building consistent regression trees from complex sample data“. In: *Journal of the American Statistical Association* 106.496, S. 1626–1636.

Valliant, R., J. A. Dever und F. Kreuter (2018). *Practical tools for designing and weighting survey samples*. Cham: Springer.

Veale, M., R. Binns und L. Edwards (2018). „Algorithms that remember: model inversion attacks and data protection law“. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376.2133, S. 20180083. DOI: 10.1098/rsta.2018.0083.

Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities*. London: Chapman and Hall.

Weerts, H., F. Pfisterer, M. Feurer, K. Eggensperger, E. Bergman, N. Awad, J. Vanschoren, M. Pechenizkiy, B. Bischl und F. Hutter (2023). *Can Fairness be Automated? Guidelines and Opportunities for Fairness-aware AutoML*. arXiv: 2303.08485.

Weichert, T. (2020). „Die Forschungsprivilegierung in der DS-GVO. Gesetzlicher Änderungsbedarf bei der Verarbeitung personenbezogener Daten für Forschungszwecke“. In: *Zeitschrift für Datenschutz* 10.1, S. 18–23.

West, B. T., J. Wagner, J. Kim und T. D. Buskirk (2023). *The Total Data Quality Framework*. https://www.coursera.org/specializations/total-data-quality. (Besucht am 13. 03. 2023).

Wolff, H.-A., S. Brink und A. v. Ungern-Sternberg, Hrsg. (2024). *BeckOK Grundgesetz - Kommentar*. 47. Ed. München: C.H.Beck.

Wright, M. N. und A. Ziegler (2017). „ranger: A fast implementation of random forests for high dimensional data in C++ and R“. In: *Journal of Statistical Software* 77, S. 1–17.

Yung, W., S.-M. Tam, B. Buelens, H. Chipman, F. Dumpert, G. Ascari, F. Rocci, J. Burger und I. Choi (2022). „A Quality Framework for Statistical Algorithms“. In: *Statistical Journal of the IAOS* 38.1. Preprint at https://statswiki.unece.org/download/attachments/285216420/QF4SA_2020_Final.pdf, S. 291–308.

Zadrozny, B. (2004). „Learning and evaluating classifiers under sample selection bias“. In: *Proceedings of the Twenty-first International Conference on Machine Learning*, S. 114.