Saidani, Younes et al.

Article — Published Version

Qualitätsdimensionen maschinellen Lernens in der amtlichen Statistik

AStA Wirtschafts- und Sozialstatistisches Archiv

Provided in Cooperation with:
Springer Nature

Suggested Citation: Saidani, Younes et al. (2023) : Qualitätsdimensionen maschinellen Lernens in der amtlichen Statistik, AStA Wirtschafts- und Sozialstatistisches Archiv, ISSN 1863-8163, Springer, Berlin, Heidelberg, Vol. 17, Iss. 3, pp. 253-303, https://doi.org/10.1007/s11943-023-00329-7

This Version is available at:
https://hdl.handle.net/10419/308978

AStA Wirtschafts- und Sozialstatistisches Archiv (2023) 17:253–303
https://doi.org/10.1007/s11943-023-00329-7

ORIGINALVERÖFFENTLICHUNG

Qualitätsdimensionen maschinellen Lernens in der amtlichen Statistik

Younes Saidani · Florian Dumpert · Christian Borgs · Alexander Brand · Andreas Nickl · Alexandra Rittmann · Johannes Rohde · Christian Salwiczek · Nina Storfinger · Selina Straub

Eingegangen: 3. April 2023 / Angenommen: 4. Oktober 2023 / Online publiziert: 17. November 2023

Zusammenfassung Die amtliche Statistik zeichnet sich durch ihren gesetzlich auferlegten Fokus auf die Qualität ihrer Veröffentlichungen aus. Dabei folgt sie den europäischen Qualitätsrahmenwerken, die auf nationaler Ebene in Form von Qualitätshandbüchern konkretisiert und operationalisiert werden, sich jedoch bis dato hinsichtlich Ausgestaltung und Interpretation an den Anforderungen der „klassischen" Statistikproduktion orientieren. Der zunehmende Einsatz maschineller Lernverfahren (ML) in der amtlichen Statistik muss daher zur Erfüllung des Qualitätsanspruchs durch ein spezifisches, darauf zugeschnittenes Qualitätsrahmenwerk begleitet werden. Das vorliegende Papier leistet einen Beitrag zur Erarbeitung eines solchen Qualitätsrahmenwerks für den Einsatz von ML in der amtlichen Statistik, indem es (1) durch den Vergleich mit bestehenden Qualitätsgrundsätzen des Verhaltenskodex für Europäische Statistiken relevante Qualitätsdimensionen für ML identifiziert und (2) diese unter Berücksichtigung der besonderen methodischen Gegebenheiten von ML ausarbeitet. Dabei (2a) ergänzt es bestehende Vorschläge durch den Aspekt der

Anmerkung zur Sortierung der Autorenliste Der Erst- und Zweitautor werden nach Beitrag, die weiteren Mitautorinnen und -autoren dagegen in alphabetischer Reihenfolge aufgeführt.

✉ Younes Saidani · Florian Dumpert
Statistisches Bundesamt, Gustav-Stresemann-Ring 11, 65189 Wiesbaden, Deutschland
E-Mail: younes.saidani@destatis.de

Christian Borgs · Johannes Rohde
Information und Technik Nordrhein-Westfalen, Mauerstraße 51, 40476 Düsseldorf, Deutschland

Alexander Brand · Andreas Nickl · Nina Storfinger · Selina Straub
Bayerisches Landesamt für Statistik, Nürnberger Straße 95, 90762 Fürth, Deutschland

Alexandra Rittmann
Statistisches Landesamt Sachsen-Anhalt, Merseburger Straße 2, 06110 Halle/Saale, Deutschland

Christian Salwiczek
Statistisches Amt für Hamburg und Schleswig-Holstein, Steckelhörn 12, 20457 Hamburg, Deutschland

Robustheit, (2b) stellt Bezug zu den Querschnittsthemen Machine Learning Operations (MLOps) und Fairness her und (2c) schlägt vor, wie die Qualitätssicherung der einzelnen Dimensionen in der Praxis der amtlichen Statistik ausgestaltet werden kann. Diese Arbeit liefert die konzeptionelle Grundlage, um Qualitätsindikatoren für ML-Verfahren formell in die Instrumente des Qualitätsmanagements im Statistischen Verbund zu überführen und damit langfristig den hohen Qualitätsstandard amtlicher Statistik auch bei Nutzung neuer Verfahren zu sichern.

Schlüsselwörter Qualität · Maschinelles Lernen · Amtliche Statistik · Erklärbarkeit · Interpretierbarkeit · Robustheit · Stabilität

JEL C80 · C52 · C18

Quality Dimensions of Machine Learning in Official Statistics

Abstract Official statistics distinguishes itself through the legally stipulated requirement to ensure the quality of its publications. To this end, it adheres to European quality frameworks, which are operationalised at the national level in the form of quality manuals. Hitherto, these have been designed and interpreted with the requirements of "classical" statistical production processes in mind. Thus, in order to ensure continued adherence to quality standards, a tailored quality framework must be developed to accompany the increasing use of machine learning (ML) methods in official statistics. This paper makes three contributions to the development of such a quality framework for the use of ML in official statistics: (1) It identifies relevant quality dimensions for ML by analysing the quality principles contained in the European Statistics Code of Practice and (2) fleshes them out in light of the methodological peculiarities of ML. Unlike previous works, (2a) robustness is proposed as a stand-alone quality dimension, (2b) machine learning operations (MLOps) and fairness are discussed as two cross-cutting issues with relevance to most quality dimensions, and (2c) suggestions are made how quality assurance can be conducted in practice for each quality dimension. This work provides the conceptual groundwork for embedding ML quality indicators in the quality management systems used by official statistics for assessment and reporting, thus ensuring that the quality standard of official statistics continues to be met when new statistical procedures are used.

Keywords Quality · Machine Learning · Official Statistics · Explainability · Interpretability · Robustness · Stability

1 Einleitung

Die amtliche Statistik genießt gegenüber anderen Statistikanbietern das besondere Privileg, dass nicht der Markt über ihren Bestand entscheidet, sondern dass sie von Gesetzes wegen laufend Daten über Massenerscheinungen zu erheben, zu sammeln, aufzubereiten, darzustellen und zu analysieren hat. Dieses Privileg erfordert

im Gegenzug aber auch einen besonderen Qualitätsstandard – zum einen, um dem eigenen Qualitätsanspruch gerecht zu werden, und zum anderen, um ihrer bedeutenden Rolle als verlässliche Datenlieferantin in einer Demokratie erfüllen zu können. Schließlich „höhlt schlechte Qualität sehr, sehr schnell das Vertrauen [in die amtliche Statistik] aus“[1], wie Walter Radermacher, der ehemalige Präsident des Statistischen Bundesamtes und der europäischen Statistikbehörde Eurostat, 2022 im Rahmen seiner Antrittsvorlesung als Honorarprofessor an der Ludwig-Maximilians-Universität München bemerkte. Die Qualität statistischer Daten hat darum in der amtlichen Statistik seit jeher eine große Bedeutung. Basierend auf europäischen Rahmenwerken – dem Quality Assurance Framework (European Statistical System 2019) und dem Verhaltenskodex für Europäische Statistiken bzw. European Statistics Code of Practice (Europäische Kommission und Eurostat 2018)[2] – entwickelten die Statistischen Ämter des Bundes und der Länder daher ein Qualitätshandbuch, das Leitlinien für den institutionellen Rahmen, die statistischen Prozesse und die statistischen Produkte enthält (Statistische Ämter des Bundes und der Länder 2021).

Dabei orientiert sich das Qualitätshandbuch hinsichtlich Interpretation und Ausarbeitung der Leitlinien zwar an den Anforderungen und Herausforderungen der „klassischen“ Statistikproduktion, verlangt jedoch explizit, dass die amtliche Statistik methodisch mit der Zeit geht, um ihren Qualitätsstandards gerecht zu werden: So sollen „[d]ie statistischen Prozesse zur Erhebung, Aufbereitung und Verbreitung von Statistiken [...] internationalen Standards und Leitlinien in vollem Umfang genügen und zugleich dem aktuellen Stand der wissenschaftlichen Forschung entsprechen. Dies gilt sowohl für die eingesetzte Methodik als auch für die angewendeten statistischen Verfahren“ (ebd., S. 19). Die deutsche amtliche Statistik beschäftigt sich folgerichtig immer wieder mit Fragen der Qualität in verschiedenen Kontexten, zuletzt beispielsweise mit Bezug zu zusammengeführten Daten (Tümmler 2020), zur Nutzung von Mobilfunkdaten (Saidani et al. 2022), zum Zensus (Tümmler und Meinke 2019; Meinke und Hentschke 2022) oder zur Beobachtung und Analyse spezieller Unternehmensgruppen (Ahlborn et al. 2021). International befassten sich in jüngerer Zeit de Waal et al. (2019) und Gootzen et al. (2023) mit der Qualität beim Zusammenführen von Informationen aus verschiedenen Datenquellen, Nguyen und Hogue (2019) mit automatisierten Qualitätskontroll- und Qualitätssicherungssystemen und Reister (2023) mit der Übertragbarkeit von statistikbezogenen Qualitätsrahmenwerken auf Fragen der Datenqualität.

Methoden des maschinellen Lernens (ML)[3] sind in den letzten Jahren von der Zukunftstechnologie zum Industriestandard herangereift und bieten u. a. neuartige Möglichkeiten im Umgang mit großen Datenmengen. Folglich wurden ML-Methoden auch in der amtlichen Statistik aktiv aufgegriffen und pilotiert (Dumpert und Beck 2017; Beck et al. 2018), sodass inzwischen erfolgreiche Beispiele für die produktive Nutzung in der deutschen amtlichen Statistikproduktion u. a. bei Imputa-

[1] Englische Zitate sind zur besseren Lesbarkeit durchgängig ins Deutsche übersetzt.

[2] Zur historischen Entwicklung siehe Kopsch et al. (2006) sowie Klumpen und Schäfer (2012). Zum Code of Practice gibt es bereits Ergänzungsvorschläge, etwa Sæbø und Holmberg (2019).

[3] Die Begriffe „ML“, „maschinelle Lernverfahren“, „Methoden des maschinellen Lernens“, „ML-Algorithmen“ u. ä. werden in diesem Artikel synonym verwendet.

tions- und Schätzverfahren zu verzeichnen sind (z. B. Preising et al. 2021; Levagin et al. 2022; Dumpert 2023). Auch auf europäischer und internationaler Ebene werden ML-Verfahren verstärkt in der Statistikproduktion eingesetzt. Zur Bewältigung methodischer, technischer und regulatorischer Herausforderungen wurde darüber hinaus im Rahmen des Machine-Learning-Projektes der UNECE HLG-MOS (Julien 2020; UNECE 2021) ein Austausch zwischen nationalen Statistikämtern etabliert.

Da sich der Einsatz von maschinellem Lernen in Teilaspekten erheblich von bisherigen Herangehensweisen unterscheidet, können bestehende Qualitätsrahmenwerke der amtlichen Statistik nicht ohne weitere Spezifizierung auf diese – für sie neuen – Verfahren angewandt werden. Um in der amtlichen Statistik langfristig etabliert werden zu können, muss daher der Einsatz von ML-Verfahren durch ein spezifisches, darauf zugeschnittenes Qualitätsrahmenwerk begleitet werden (Julien 2020; Dumpert 2021). Ein solches Rahmenwerk erfüllt mehrere Funktionen: Zum einen verdeutlicht es, inwiefern die Anwendung maschineller Lernverfahren in den verschiedenen Schritten des Geschäftsprozessmodells Amtliche Statistik (Blumöhr et al. 2017) die Qualität der statistischen Produkte beeinflussen kann. Dies ist besonders für die statistischen Fachbereiche relevant, die die Gesamtverantwortung für die Qualität ihres Statistikproduktes tragen. Zum anderen erhalten Statistikproduzenten konkrete Qualitätsrichtlinien, die bei der Entwicklung und dem Einsatz von ML-Algorithmen berücksichtigt werden müssen. Zuletzt schaffen klare Qualitätsanforderungen auch Transparenz für die Nutzenden und stärken das Vertrauen in die amtliche Statistik.

Das vorliegende Papier soll einen Beitrag zur Erarbeitung eines Qualitätsbegriffs für den Einsatz von Methoden des maschinellen Lernens in der amtlichen Statistik leisten. Als Startpunkt werden die bestehenden Qualitätsrahmenwerke genutzt, um Qualitätsdimensionen zu identifizieren, die bei der Nutzung maschineller Lernverfahren tangiert werden oder Implikation für ihren Einsatz haben. Dieser Beitrag zeigt – nach bestem Wissen der Autorinnen und Autoren – erstmalig detailliert den Bezug zwischen den Qualitätsprinzipien amtlicher Statistik und den Besonderheiten der Arbeit mit maschinellen Lernverfahren auf, was nach Ansicht der Autorinnen und Autoren eine essenzielle Grundlage für die Akzeptanz solcher Methoden in der amtlichen Statistik und damit ihre langfristige Nutzung ist. Im Hauptteil des Papiers werden die herausgearbeiteten Dimensionen dann unter Berücksichtigung der besonderen methodischen Gegebenheiten von ML konkretisiert und mit Prozessindikatoren versehen. Abgeschlossen wird der Aufsatz durch eine kurze Diskussion übergreifender Themen in Bezug auf den Einsatz maschinellen Lernens in der amtlichen Statistik – Machine Learning Operations (MLOps) und Fairness.

2 Qualitätsprinzipien der amtlichen Statistik

Die Qualitätsprinzipien der amtlichen Statistik leiten sich – für Deutschland als Mitglied des Europäischen Statistischen Systems (ESS) – primär aus den europäischen Rahmenwerken ab. An erster Stelle steht hier der Verhaltenskodex für Europäische Statistiken (Europäische Kommission und Eurostat 2018), dessen Einhaltung für die nationalen Statistischen Ämter der 27 EU-Mitgliedsstaaten sowie für das

Statistische Amt der Europäischen Union (Eurostat) im Rahmen einer Selbstverpflichtung bindend ist (ebd., S. 4). Der Kodex soll gewährleisten, dass die amtliche Statistikproduktion frei von politischer Einflussnahme und nach anerkannten wissenschaftlichen Verfahren durchgeführt wird. Er enthält 16 Qualitätsgrundsätze sowie eine Reihe von Indikatoren zur Qualitätsmessung und -überprüfung. Der Verhaltenskodex wird ergänzt und spezifiziert durch das Quality Assurance Framework (European Statistical System 2019). Dieses enthält eine Sammlung von detaillierten Methoden, Werkzeugen und vorbildlichen Praktiken (Best Practices), die unter Berücksichtigung der nationalen Gegebenheiten eigenverantwortlich zur Umsetzung des Verhaltenskodex genutzt werden können. Auf Grundlage dieser beiden Rahmenwerke wurde in Deutschland durch die Statistischen Ämter des Bundes und der Länder ein Qualitätshandbuch ausgearbeitet, das eine Anleitung zur konkreten Umsetzung der Qualitätsanforderungen des Verhaltenskodex für Europäische Statistiken in der deutschen amtlichen Statistik liefert (Statistische Ämter des Bundes und der Länder 2021).

Maßstab und Ausgangspunkt aller Überlegungen bezüglich eines Qualitätsbegriffs für maschinelles Lernen in der amtlichen Statistik müssen somit die im Verhaltenskodex und im Qualitätshandbuch verankerten Grundsätze sein.[4] Von den drei Ebenen, für die Richtlinien formuliert werden – das institutionelle Umfeld, die statistischen Prozesse und die statistischen Produkte – sind vornehmlich die letzten beiden von Interesse, da die Grundsätze zu institutionellen und organisatorischen Faktoren durch den Einsatz von maschinellen Lernverfahren nicht berührt werden.

Die im Verhaltenskodex festgelegten Grundsätze für die statistischen Prozesse definieren europäische Standards, Leitlinien und Praktiken für das Management, die Effizienz und die Innovation der Prozesse. Darüber hinaus wirken sie darauf hin, die Glaubwürdigkeit der amtlichen Statistik zu erhalten und zu stärken. Die Grundsätze für die statistischen Produkte nehmen vor allem den Nutzerbedarf in den Blick. Sie sollen sicherstellen, dass die Statistiken den Bedarf der europäischen Institutionen, Regierungen, Forschungseinrichtungen und Unternehmen sowie der Öffentlichkeit im Allgemeinen decken. Einen Überblick über alle neun Grundsätze für statistische Prozesse und Produkte bietet Tab. 1.

Maschinelle Lernverfahren, die im Rahmen des statistischen Produktionsprozesses neu eingesetzt werden, bestehende Prozesse ergänzen oder bisherige Arbeiten vollständig ersetzen, müssen als Mindestanforderung die Qualitätsanforderungen an die statistischen Prozesse und Produkte weiterhin erfüllen. Tragen sie darüber hinaus auch noch zu einer Verbesserung von Prozessen oder Produkten im Sinne des Verhaltenskodex bei, ist der Mehrwert von maschinellem Lernen offensichtlich. Im Folgenden werden die einzelnen Grundsätze und die damit verbundenen Indikatoren im Detail vorgestellt und im Hinblick auf die Relevanz für maschinelles Lernen bewertet.

[4] Da die Grundsätze und ihre Indikatoren in beiden Dokumenten quasi identisch enthalten sind, werden der Verhaltenskodex und das Qualitätshandbuch im Folgenden austauschbar verwendet.

Tab. 1 Grundsätze für die statistischen Prozesse und Produkte (Statistische Ämter des Bundes und der Länder 2021)

Grundsätze für die statistischen Prozesse	
7. Solide Methodik	Qualitativ hochwertige Statistiken basieren auf einer soliden Methodik. Diese erfordert geeignete Instrumente und Verfahren sowie ein entsprechendes Know-how
8. Geeignete statistische Verfahren	Geeignete statistische Verfahren – von der Erhebung bis zur Validierung der Daten – bilden die Grundlage für qualitativ hochwertige Statistiken
9. Vermeidung einer übermäßigen Belastung der Auskunftgebenden	Der Beantwortungsaufwand steht in einem angemessenen Verhältnis zum Bedarf der Nutzerinnen und Nutzer und ist für die Auskunftgebenden (Respondenten) nicht übermäßig hoch. Die statistischen Stellen überwachen den Beantwortungsaufwand und legen Ziele für dessen schrittweise Verringerung fest
10. Wirtschaftlichkeit	Ressourcen werden effektiv eingesetzt
Grundsätze für die statistischen Produkte	
11. Relevanz	Die europäischen Statistiken entsprechen dem Bedarf der Nutzerinnen und Nutzer
12. Genauigkeit und Zuverlässigkeit	Die europäischen Statistiken spiegeln die Realität genau und zuverlässig wieder
13. Aktualität und Pünktlichkeit	Die europäischen Statistiken sind aktuell und werden pünktlich veröffentlicht
14. Kohärenz und Vergleichbarkeit	Die europäischen Statistiken sind untereinander und im Zeitablauf konsistent und zwischen Regionen und Ländern vergleichbar; es ist möglich, miteinander in Beziehung stehende Daten aus unterschiedlichen Quellen zu kombinieren und gemeinsam zu verwenden
15. Zugänglichkeit und Klarheit	Die europäischen Statistiken werden klar und verständlich präsentiert, in geeigneter und benutzerfreundlicher Weise veröffentlicht und sind zusammen mit einschlägigen Metadaten und Erläuterungen entsprechend dem Grundsatz der Unparteilichkeit verfügbar und zugänglich

2.1 Qualitätsgrundsätze für statistische Prozesse

Der Verhaltenskodex definiert in vier Grundsätzen und 27 Indikatoren die Leitlinien der Qualitätssicherung für die statistischen Prozesse (vgl. Tab. 13 im Anhang). Im Folgenden werden die Grundsätze und die dazugehörigen Indikatoren dahingehend betrachtet, ob sie durch den Einsatz von maschinellem Lernen tangiert werden und ob sich daraus Qualitätsanforderungen an den Einsatz von maschinellen Lernverfahren ableiten lassen.

Grundsatz 7 „Solide Methodik“ wird in sieben Indikatoren untergliedert, wovon alle einen Bezug zum Einsatz von maschinellen Lernverfahren haben. Der erste Aspekt bezieht sich darauf, dass die verwendeten Methoden i. d. R. den internationalen Empfehlungen entsprechen müssen (Indikator 7.1). Abweichungen von diesem Vorgehen müssen daher auch beim Einsatz von ML in Qualitäts- und Methodenberichten erläutert werden. Des Weiteren wird in den Indikatoren beschrieben, dass Standardkonzepte, -definitionen und -klassifikationen einheitlich verwendet werden sowie dass diese Klassifikationssysteme auf nationaler und europäischer Ebene übereinstimmen müssen (7.2 und 7.4). Da maschinelle Lernverfahren auch bei Klassifikationen unterstützend eingesetzt werden können, sollten sie sich den internationalen Systemen anpassen, sodass beispielsweise Trainings-, Validierungs- und Testdaten-

sätze ausgetauscht werden könnten. Darüber hinaus können bei der regelmäßigen Aktualisierung des Unternehmensregisters (7.3) ML-Verfahren unterstützend eingesetzt werden, etwa um fehlende Werte zu imputieren. Die weiteren Indikatoren halten fest, dass für die Erstellung von qualitativ hochwertigen Statistiken und für die Sicherstellung einer soliden Methodik Absolventen inhaltlich einschlägiger Studiengänge eingestellt werden (7.5) sowie dass die Mitarbeitenden kontinuierlich weitergebildet werden müssen (7.6). Darüber hinaus sollten zur steten Verbesserung der Methodik konkrete Maßnahmen, wie etwa eine Zusammenarbeit mit der Wissenschaft, durchgeführt werden (7.7). Wissenserwerb und Wissenstransfer sind gerade im Umgang mit dem sich rapide weiterentwickelnden Gebiet des maschinellen Lernens entscheidende Bausteine. Zwar halten maschinelle Lernverfahren mittlerweile Einzug in universitäre Lehrpläne, daraus folgt jedoch nicht unmittelbar, dass die erforderlichen Kenntnisse ohne aktive Bemühungen mittelfristig bei einer ausreichenden Anzahl an Mitarbeitenden in allen Statistischen Ämtern vorhanden sein werden. Nur durch die gezielte Anwerbung von Fachkräften sowie die Fortbildung der Mitarbeitenden (etwa als Teil der Gemeinsamen Fortbildungen der Statistischen Ämter des Bundes und der Länder) können ausreichende Kompetenzen zu maschinellem Lernen in Statistikbehörden geschaffen werden. Um den Austausch mit der Wissenschaft zu fördern, muss bereits in der amtlichen Statistik aufgebautes Erfahrungswissen im Rahmen von wissenschaftlichen Veranstaltungen (z. B. universitäre Workshops) geteilt und diskutiert werden.

Grundsatz 8 „Geeignete statistische Verfahren“ wird im Verhaltenskodex in neun Indikatoren untergliedert, wovon drei in direktem Zusammenhang mit Methoden des maschinellen Lernens stehen: Erstens sollen die Erhebungs- und Stichprobenpläne, die auf Grundlage der gesetzlichen Vorgaben festgelegt werden, „auf soliden Grundlagen“ basieren (Indikator 8.3). Die Methoden und Verfahren werden z. B. in den Fachgremien für die einzelnen Statistiken mit den Fachreferentinnen und Fachreferenten der Länder konkretisiert, abgestimmt und regelmäßig überprüft, Stichproben werden regelmäßig neu gezogen, zwischen zwei Stichprobenziehungen werden Stichproben ggf. um Neuzugangsstichproben ergänzt und der Einsatz von Schwellenwerten (z. B. bzgl. Umsatz oder Anzahl an Beschäftigten), unterhalb derer nicht befragt wird, wird bei Bedarf überprüft. Maschinelle Lernverfahren finden bei Schätzverfahren Anwendung und können sich auf die Festlegung von Erhebungs- und Stichprobenplänen auswirken. Wie bei klassischen statistischen Verfahren muss das jeweils zuständige Fachgremium für die einzelnen Statistiken auch bei maschinellen Lernverfahren die Rechtskonformität der verwendeten Methoden und Verfahren sicherstellen, da mittels ML-Algorithmen geschätzte (und somit unter Umständen vom wahren Wert abweichende) Werte potenziell Konsequenzen für die betreffenden administrativen Einheiten nach sich ziehen können. Der zweite relevante Indikator betrifft die Methoden und Verfahren der Datengewinnung, -eingabe und -kodierung (8.4). Diese werden in den Fachgremien des Statistischen Verbunds[5] konkretisiert, abgestimmt und regelmäßig überprüft. Insbesondere in den Phasen 4 (Daten gewinnen) und 5 (Daten aufbereiten) des Geschäftsprozessmodells Amtliche Statistik (GMAS) bietet maschinelles Lernen – etwa durch automatisierte

[5] Den Statistischen Verbund bilden die Statistischen Ämter des Bundes und der Länder.

Kodierung – großes Potenzial für die Erhöhung der Datenqualität und langfristig für die Einsparung von Personalkosten. Drittens sollen für das Editieren und die Imputation „geeignete Verfahren“ eingesetzt, überprüft und regelmäßig überarbeitet werden (8.5). Während beim Editieren Fehler in den Daten identifiziert und korrigiert werden, wird die Imputation zur Ersetzung fehlender Werte genutzt. In der Praxis werden dafür Verfahren zur Plausibilisierung und zur Imputation entwickelt, getestet und schließlich in den statistischen Produktionsprozess implementiert. Die Konzeption, Diskussion von Ergebnissen und Verbesserung von Plausibilitätsprüfungen und Imputationen werden im Rahmen der Fachgremien für die einzelnen Statistiken thematisiert. Prozesse der Datenaufbereitung – und damit auch das Editieren und Imputieren – sind geeignete Anwendungsfelder für maschinelles Lernen. Um zur Qualitätserhaltung beizutragen, müssen ML-Algorithmen adäquat eingesetzt, kontrolliert und falls erforderlich regelmäßig angepasst werden.

Grundsatz 9 „Vermeidung einer übermäßigen Belastung der Auskunftgebenden“ beinhaltet zwei konkrete Anknüpfungspunkte für maschinelle Lernverfahren: So soll die Belastung der Auskunftgebenden „auf das absolut erforderliche Maß“ begrenzt werden (Indikator 9.1), unter anderem durch Maßnahmen, die „die Verknüpfung von Datenquellen“ ermöglichen (9.6). In der Praxis geht der Durchführung von Statistiken meist ein Gesetzgebungsprozess voraus.[6] Dieser berücksichtigt die Begründung für eine Erhebung sowie die zu erhebenden Merkmale und legt die sachliche und räumliche Gliederungstiefe fest. Zusammen mit dem entsprechenden Kostenmodell wägt der Gesetzgeber so das Verhältnis zwischen Informationsbedarf und Belastung der Auskunftgebenden ab. Damit dieses Verhältnis ausgeglichen ist, muss der Umfang der zu beantwortenden Fragen auf das notwendige Maß minimiert werden. Durch die Verknüpfung verschiedener Datenquellen (z. B. Webscraping in der Preisstatistik, Zuschätzung von Variablen der Verdienststatistik aus dem Mikrozensus) und durch die Extraktion bereits vorhandener Informationen – klassische Anwendungsfelder von maschinellem Lernen – ist es prinzipiell möglich, den Beantwortungsaufwand bei primärstatistischen Erhebungen zu reduzieren. Dabei ist jedoch zu beachten, dass auch das Zusammenführen von Datenquellen den Bestimmungen des Bundesstatistikgesetzes unterliegt (vgl. § 13a und § 6 Abs. 5 BStatG); Daten aus frei zugänglichen Quellen dürfen etwa grundsätzlich mit amtlichen Daten verknüpft werden.

Grundsatz 10 „Wirtschaftlichkeit“ ist in vier Indikatoren unterteilt, von denen zwei für den Einsatz von maschinellem Lernen relevant sind. Erstens liefert die Anweisung, das Produktivitätspotenzial der Informations- und Kommunikationstechnologien (IKT) auszuschöpfen (Indikator 10.2), ein Mandat für die Prüfung des Einsatzes von ML-Verfahren und darauf basierender IKT in der Statistikbearbeitung. Zweitens ist auch eine zunehmende Standardisierung in der Entwicklung von ML-Lösungen einerseits sowie in ihrer Anwendung andererseits anzustreben (10.4), wo-

[6] Dies gilt mit Ausnahme von „Erhebungen für besondere Zwecke“ nach § 7 BStatG, die zur Erfüllung eines kurzfristigen Datenbedarfs (Absatz 1) oder zur Klärung wissenschaftlich-methodischer Fragestellungen (Absatz 2) durchgeführt werden können, jedoch ohne Auskunftspflicht und mit maximal 20.000 Befragten.

bei hier jedoch methodische Besonderheiten einzelner maschineller Lernverfahren berücksichtigt werden müssen.

2.2 Qualitätsgrundsätze für statistische Produkte

Der Verhaltenskodex beschreibt fünf Grundsätze für die Qualität statistischer Produkte (vgl. Tab. 14 im Anhang). Während die eben diskutierten Qualitätsgrundsätze für statistische Prozesse die Art und Weise der Statistikproduktion thematisierten, erlauben jene eine Beurteilung und Messung der Qualität von Produkten der amtlichen Statistik. Im Folgenden wird untersucht, inwiefern sich der Einsatz von Methoden des maschinellen Lernens auf die einzelnen Grundsätze und ihre Indikatoren auswirkt.

Grundsatz 11 „Relevanz“ steht in keinem direkten Zusammenhang mit dem Einsatz von maschinellem Lernen bei der Statistikproduktion.

Die Indikatoren des Grundsatzes 12 „Genauigkeit und Zuverlässigkeit“ beschreiben mit der Basisdatenevaluation und -validierung (Indikator 12.1) sowie der Messung des Stichprobenfehlers und des Nicht-Stichprobenfehlers (12.2) zwei Aspekte, die beim Einsatz von maschinellen Lernverfahren wesentlich sind. Modelle des maschinellen Lernens werden vor ihrem Einsatz mit den verfügbaren Einzeldaten (auch: Basisdaten) trainiert, um ein möglichst genaues Modell mit hoher Vorhersagequalität zu entwickeln. Je besser die Basisdaten den interessierenden Sachverhalt repräsentieren, desto besser fällt auch die Vorhersagequalität der Modelle aus. Beim Einsatz von maschinellen Lernverfahren ist daher besonderes Augenmerk auf die Evaluation der verwendeten Basisdaten zu legen, um die Genauigkeit der Ergebnisse sicherzustellen. Ebenfalls ist standardmäßig die Güte des Modells mit geeigneten Kennzahlen (z. B. Accuracy, Precision) zu beschreiben und zu dokumentieren.

Grundsatz 13 „Aktualität und Pünktlichkeit“ thematisiert in fünf Indikatoren den Veröffentlichungszeitpunkt und die Periodizität der Statistiken. In diesem Zusammenhang sind für maschinelles Lernen besonders zwei Indikatoren relevant. Beim Einsatz von Verfahren des maschinellen Lernens muss die Aktualität von Datenveröffentlichungen und -lieferungen z. B. an Eurostat grundsätzlich weiterhin „europäische und andere internationale Veröffentlichungsstandards“ erfüllen (Indikator 13.1). Darüber hinaus kann die Statistikerstellung jedoch auch beschleunigt werden, da ML-Algorithmen manuelle Arbeiten ersetzen, bestimmte Prozessschritte automatisieren und dadurch qualitätsverbessernd wirken können. Außerdem können sie oft ermöglichen, dass zu einem früheren Zeitpunkt „vorläufige Ergebnisse von akzeptabler Gesamtgenauigkeit“ veröffentlicht werden (13.5). Beispielsweise nutzen Salgado et al. (2023) ein Gradient Boosting-Verfahren, um noch nicht eingegangene Erhebungsdaten zu imputieren und so eine zunehmend genaue Frühschätzung für den spanischen Industrieumsatzindex zu ermöglichen.

Grundsatz 14 „Kohärenz und Vergleichbarkeit“ mit seinen fünf Indikatoren tangiert Methoden des maschinellen Lernens in zweierlei Hinsicht: Einerseits müssen Statistiken in sich kohärent und vergleichbar sein (Indikator 14.1). Andererseits müssen sie auch über die Zeit hinweg vergleichbar bleiben – vor allem dann, wenn es zu methodischen Änderungen kommt (14.2). Abweichungen müssen dokumentiert und den Nutzerinnen und Nutzern zugänglich gemacht werden. Sollten Methoden

des maschinellen Lernens in einem der Produktionsschritte ergänzend oder substituierend eingesetzt werden, müssen statistikübergreifende Beziehungen unbedingt erhalten bleiben, um die Qualität der Statistiken nicht zu gefährden. Darüber hinaus sollten derartige methodische Änderungen in Aufsätzen oder Methodenpapieren transparent kommuniziert werden.

Grundsatz 15 „Zugänglichkeit und Klarheit“ und dessen sechs Indikatoren behandeln den Datenzugang und die Beschreibung der Statistiken, welche jeweils möglichst nutzerfreundlich gestaltet werden sollten. Für den Einsatz maschineller Lernverfahren sind zwei Indikatoren relevant: Den Nutzenden müssen alle nötigen Informationen zur Verfügung gestellt werden, die für „eine korrekte Interpretation und aussagekräftige Vergleiche“ benötigt werden (Indikator 15.1). Zudem müssen grundlegende Informationen zu den Methoden präsentiert werden. Der Einsatz von ML-Algorithmen muss daher in Qualitätsberichten und Methodenaufsätzen kommuniziert werden; auch die entsprechenden Metadaten (15.5) müssen methodische Informationen enthalten.

2.3 Zwischenfazit

Amtliche Statistik ist ein komplexes Projekt, dessen Abläufe aus Gründen der Vergleichbarkeit und Qualitätssicherung hochgradig strukturiert sind. Der klassische Statistikproduktionsprozess von der Erhebung bis zur Veröffentlichung nutzt daher etablierte Werkzeuge und bewährte Schnittstellen. Diesen Prozess beschreiben die Qualitätsrahmenwerke der amtlichen Statistik – der Verhaltenskodex für Europäische Statistiken (Europäische Kommission und Eurostat 2018) und das Qualitätshandbuch der Statistischen Ämter des Bundes und der Länder (Statistische Ämter des Bundes und der Länder 2021) – und auf diesem Abstraktionsniveau formulieren sie Qualitätsgrundsätze und -indikatoren, welche durchaus Anhaltspunkte für die Erarbeitung eines Qualitätsbegriffs für maschinelles Lernen bieten: So enthalten die meisten Qualitätsgrundsätze Indikatoren, die durch den Einsatz von maschinellen Lernverfahren tangiert werden oder Auswirkungen auf deren Nutzung haben. Gleichzeitig drängt sich aber die Frage auf, ob diese Anhaltspunkte die Besonderheiten bei der Nutzung von ML vollumfänglich abbilden können. Dieser Frage soll nun mit Blick auf die Literatur nachgegangen werden, bevor Qualitätsdimensionen für den Einsatz maschineller Lernverfahren formuliert werden.

3 Qualitätsdimensionen maschinellen Lernens

Maschinelles Lernen bezeichnet eine Sammlung von Methoden, die sich gegenüber traditionellen statistischen Methoden durch den Verwendungszweck abgrenzen: Während bei der „klassischen“, forschungsorientierten Statistik unter anderem Hypothesentests im Vordergrund stehen, zielen ML-Algorithmen darauf ab, bestmöglich Eigenschaften neuer Beobachtungen vorherzusagen – in der amtlichen Statistik ebenso wie in anderen Teilen der Verwaltung, in der Industrie oder der Finanzwirtschaft oftmals mit dem Ziel der Prozessautomatisierung. Verfahren des maschinellen Lernens, wie beispielsweise baumbasierte Verfahren (einschließlich Random Forests

und Boosting-Ansätze), Support Vector Machines oder neuronale Netze, kommen in der amtlichen Statistik hauptsächlich (aber nicht ausschließlich) in Schritten der Datenaufbereitung zum Einsatz. Sie übernehmen dort Aufgaben in den Bereichen Klassifikation und Codierung, Fehlererkennung und Fehlerkorrektur sowie Imputation fehlender Werte.

Um dem Anspruch amtlicher Statistik gerecht zu werden, müssen auch maschinelle Lernverfahren ihren Qualitätsstandard erfüllen. Voraussetzung hierfür ist die Entwicklung eines konkreten Qualitätsbegriffs für ML in der amtlichen Statistik.[7] Einen ersten Beitrag hierzu liefern de Broe et al. (2021): Sie betrachten die Möglichkeiten und Herausforderungen, die sich durch die Nutzung neuer Daten und neuer Methoden allgemein in der amtlichen Statistik ergeben, und leiten daraus mehrere Dimensionen von Qualität ab – hinsichtlich der Forschung, Methodik, Daten, Prozesse und Ergebnisse. Im Vergleich mit den Qualitätsrahmenwerken der amtlichen Statistik resümieren sie, dass „die methodischen Qualitätsaspekte von statistischem Lernen und Big Data eindeutig neue Elemente enthalten […]. Diese Elemente können als Erweiterung der bestehenden Qualitätsdimensionen verstanden werden anstatt als komplett neue Dimensionen“ (ebd., S. 357). In Verbindung mit der Analyse des Verhaltenskodex hinsichtlich ML aus dem vorigen Kapitel ergibt sich also folgendes Bild: Zum einen sind die amtlichen Qualitätsgrundsätze an vielen Stellen vom Einsatz von ML-Verfahren betroffen und liefern Ansatzpunkte für eine Qualitätsbewertung derselben. Zum anderen lassen sich die methodischen Eigenheiten von ML auch relativ nahtlos unter die bestehenden Qualitätsgrundsätze subsummieren. Jedoch sind die bestehenden Indikatoren nicht ausreichend, um die Qualitätsaspekte beim Einsatz von ML vollumfänglich zu beschreiben; die eher generische Ausrichtung der Grundsätze für statistische Prozesse ist zu abstrakt und die Grundsätze für statistische Produkte (die sich allesamt auf statistische Veröffentlichungen beziehen) zu mittelbar. Aus diesem Grund müssen die relevanten Grundsätze konkreter auf ML bezogen werden, um eine umfassende und angemessene Qualitätsmessung zu ermöglichen.

Dieser Aufgabe nahm sich eine Gruppe von nationalen Expertinnen und Experten aus Mitgliedsstaaten der UNECE und Australien an und entwickelte ein Quality Framework for Statistical Algorithms (Yung et al. 2022), welches viele der von de Broe et al. (2021) herausgearbeiteten Qualitätsaspekte aufgreift, hinsichtlich der Besonderheiten statistischer Algorithmen in der Statistikproduktion – insbesondere ihre Rolle bei der Erstellung von Zwischenprodukten – schärft und in fünf Qualitätsdimensionen wie folgt zusammenfasst:

[7] Neben der amtlichen Statistik befassen sich auch weitere Akteure mit Fragen der Qualität oder den wünschenswerten Eigenschaften von ML-Verfahren. Beispielsweise betrachten die Bundesanstalt für Finanzdienstleistungsaufsicht (BaFin) und die Deutsche Bundesbank den Einsatz von ML im Kontext ihrer Aufsichtsfunktion und fokussieren sich dabei auf Aspekte wie Stabilität, Entwicklung vs. Betrieb, Erklärbarkeit und Adaptivität beim Einsatz von ML im Versicherungs- und Bankenbereich (BaFin und Deutsche Bundesbank 2021, 2022). Das Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme (IAIS) legt einen Leitfaden zur Gestaltung vertrauenswürdiger künstlicher Intelligenz (KI-Prüfkatalog) vor, der die Vertrauenswürdigkeitsdimensionen Fairness, Autonomie und Kontrolle, Transparenz, Verlässlichkeit, Sicherheit und Datenschutz beleuchtet (Poretschkin et al. 2021).

- ***Genauigkeit*** ist der Grad eines statistischen Outputs, zu dem dieser in der Lage ist, das zu messende Phänomen korrekt zu beschreiben, d. h. den Abstand zwischen Schätzung und wahrem Wert zu minimieren.
- ***Erklärbarkeit*** ist die Eigenschaft zu verstehen, welche Zusammenhänge der Algorithmus nutzt, um Vorhersagen zu treffen, also den (gegebenenfalls nur lokalen) Zusammenhang zwischen Eingabe- und Ausgabevariablen darlegen zu können.
- ***Reproduzierbarkeit*** ist die Fähigkeit, bei Nutzung der gleichen Daten und des gleichen Algorithmus identische Ergebnisse zu erzielen. Darüber hinaus kann darunter auch verstanden werden, dass die Anwendung des gleichen Algorithmus auf verschiedene Zufallsstichproben aus ein und derselben Grundgesamtheit regelmäßig zu im Wesentlichen gleichen Ergebnissen führen soll.
- ***Aktualität und Pünktlichkeit*** ist die Fähigkeit, den Algorithmus innerhalb der geforderten Zeit konzipieren, trainieren und anwenden zu können sowie aktuelle Ergebnisse zu veröffentlichen.
- ***Wirtschaftlichkeit*** ist das Verhältnis der Ausprägung der anderen Qualitätsdimensionen zu den Kosten der Umsetzung. Diese Definition hat den Vorteil, dass sie verschiedene Methoden mit unterschiedlicher Genauigkeit, Schnelligkeit oder Erklärbarkeit vergleichbar macht.

Nicht berücksichtigt wird hier eine Fragestellung, die bei maschinellem Lernen von besonderer Bedeutung ist, sich aber keinem der fünf genannten Begriffe zuordnen lässt: Wie verlässlich ist ein einmal trainiertes Modell im Produktivbetrieb, wenn es mit „unbekannten" Störeinflüssen konfrontiert wird? Nachdem ein Modell etwa bestimmte Zusammenhänge aus den Daten erlernt hat, können sich diese Beziehungen im Zeitverlauf verändern. Um dann keine verzerrten Ergebnisse zu erhalten, müssen bestimmte Vorkehrungen getroffen werden, etwa könnten ML-Modelle regelmäßig oder bei der Überschreitung definierter Schwellenwerte auf Grundlage aktueller Daten neu trainiert werden. Den Umgang mit diesem und verwandten Phänomenen beschreibt der Begriff der ***Robustheit***, der für eine adäquate Qualitätsbewertung von maschinellen Lernverfahren unerlässlich und daher in der obigen Liste zu ergänzen ist. Insgesamt schlägt dieser Aufsatz also folgende sechs Grundsätze zur Qualitätsmessung beim Einsatz maschineller Lernverfahrenen in der amtlichen Statistik vor, sortiert nach Bezugspunkt vom Produkt zum Prozess:

1. Genauigkeit
2. Robustheit
3. Erklärbarkeit
4. Reproduzierbarkeit
5. Aktualität und Pünktlichkeit
6. Wirtschaftlichkeit

Mit Rückbezug auf die Analyse des Verhaltenskodex im letzten Abschnitt wird deutlich, dass mit dieser Auswahl an Dimensionen zur Qualitätsmessung alle für den Einsatz von ML relevanten Aspekte aus dem Verhaltenskodex aufgegriffen bzw. konkretisiert werden können, auch wenn diese zunächst zu mittelbar oder zu generisch erschienen: Die wesentlichen Themen der Grundsätze solide Methodik, geeignete statische Verfahren und Vermeidung einer übermäßigen Belastung der Auskunftge-

benden werden aufgrund ihrer Kleinteiligkeit verschiedenen Dimension zugeordnet (v. a. Genauigkeit, Robustheit, Erklärbarkeit und Reproduzierbarkeit). Die Grundsätze Genauigkeit und Zuverlässigkeit werden in den zwei Dimensionen Genauigkeit sowie Robustheit konkretisiert. Aspekte des Grundsatzes Kohärenz und Vergleichbarkeit werden ebenso bei der Robustheit aufgegriffen; für ML relevante Indikatoren des Grundsatzes Zugänglichkeit und Klarheit finden sich bei den Dimensionen Reproduzierbarkeit und Erklärbarkeit wieder. Zuletzt werden die Grundsätze Aktualität und Pünktlichkeit sowie Wirtschaftlichkeit jeweils in gleichnamigen Dimensionen spezifiziert.

Das vorliegende Papier leistet zur Entwicklung eines amtlichen Qualitätsbegriffs für maschinelles Lernen[8] drei Beiträge: Erstens ergänzt es die in Yung et al. (2022) vorgeschlagenen Qualitätsdimensionen um den Aspekt der Robustheit eines ML-Verfahrens. Zweitens arbeitet es die einzelnen Qualitätsdimensionen stärker aus, schärft ihre Definitionen und stellt Bezug zu den Querschnittsthemen Machine Learning Operations (MLOps) und Fairness her. Drittens macht es einen Vorschlag, wie die Qualitätssicherung der einzelnen Dimensionen in der Praxis ausgestaltet werden kann, indem es Indikatoren formuliert. Damit schafft es für die deutsche amtliche Statistik die notwendigen Grundlagen, um der Forderung von de Broe et al. (2021), „diese Aspekte [der Qualität maschinellen Lernens] in die aus dem Verhaltenskodex abgeleiteten Qualitätshandbücher sowie die Qualitätsrichtlinien der nationalen Statistischen Ämter einzugliedern“ (ebd., S. 357), nachzukommen.

Im Folgenden werden die vorgeschlagenen Dimensionen und die sich daraus ableitbaren Anforderungen an ML-Verfahren ausführlich definiert. Anschließend wird aufgezeigt, welche Besonderheiten maschineller Lernverfahren jeweils explizit berücksichtigt werden müssen. Zuletzt werden Indikatoren abgeleitet, mittels derer gemessen werden kann, inwieweit die definierten Qualitätsanforderungen im Einzelfall erfüllt wurden.

3.1 Genauigkeit

Die europäischen Qualitätsprinzipien stellen explizite Ansprüche an die Genauigkeit und Zuverlässigkeit von amtlichen statistischen Prozessen und Produkten. Damit handelt es sich bei der Genauigkeit um eine notwendige Qualitätsdimension, die bei allen Statistiken zwingend erfüllt sein muss. Das betrifft zwar vorrangig die zu veröffentlichenden Zahlen, doch auch die Genauigkeit von Zwischenergebnissen ist dafür mindestens förderlich. Im Kontext der amtlichen Statistik ist Genauigkeit definiert als der (a) Grad, zu dem ein Algorithmus in der Lage ist, (b) das zu messende Phänomen (c) korrekt zu messen und zu beschreiben (Yung et al. 2022). Diese Definition soll im Folgenden anhand der drei hervorgehobenen Formulierungen erläutert werden:

[8] Der hier entwickelte Qualitätsbegriff ist in Teilen auch für andere statistische Algorithmen relevant und ist immer dann anwendbar, wenn automatisiert oder datenbasiert gearbeitet wird. Im Detail konzentriert er sich jedoch auf die Anforderungen und Eigenheiten von ML.

a. Genauigkeit ist nicht als absolutes Kriterium definiert, sondern kann graduell erfüllt werden. Welcher Grad „ausreichend“ ist, um die Qualitätsanforderungen zufriedenstellend zu erfüllen, kann nicht pauschal für alle Anwendungsfälle festgelegt werden, sondern ist Abwägungssache. Mag bei der Zuschätzung eines weniger bedeutsamen Erhebungsmerkmals ein relativ geringer Grenzwert für die Vorhersagegenauigkeit ausreichend sein, so erfordert etwa die Klassifizierung von Wirtschaftszweigen ein deutlich höheres Maß an Genauigkeit. Bei dieser Abwägung muss bedacht werden, dass auch bestehende (nicht-ML, oft menschliche) statistische Prozesse keine perfekte Genauigkeit bieten (siehe 3.1.3 Qualität der Trainingsdaten). Wenn die Genauigkeit des bestehenden Verfahrens relativ gering ist, so ist die Messlatte für alternative Methoden niedriger und eine ML-Methode erfüllt *ceteris paribus* die Qualitätsdimension „Genauigkeit“ zu einem höheren Grad. Darüber hinaus erlaubt die Definition von Genauigkeit als graduell erfüllbarer Kennwert die gegenseitige Abwägung verschiedener Qualitätsdimensionen. So kann eine geringere Genauigkeit in der Praxis u. U. durch geringere Implementierungskosten aufgewogen werden. Zuletzt ist anzumerken, dass ein hohes Maß an Kontextwissen notwendig ist, um zu definieren, welcher Grad der Genauigkeit ausreichend ist. Während rein methodische Untersuchungen nur notwendige Gütemaße bzw. Kennzahlen liefern, obliegt die Entscheidung schlussendlich dem zuständigen Fachbereich, der das notwendige Kontextwissen der jeweiligen Statistik besitzt.
b. Die Evaluation der Genauigkeit ist abhängig von dem zu messenden Phänomen. Im Kontext des überwachten maschinellen Lernens wird dabei grob zwischen zwei Klassen von Phänomenen unterschieden: Wird die Gruppenzugehörigkeit einer Beobachtung betrachtet, so handelt es sich um ein Klassifikationsproblem, für die Schätzung der Ausprägung einer stetigen Variablen nutzt man Regressionsverfahren. Welche Gütemaße angegeben werden können, unterscheidet sich maßgeblich je nach Klasse des Problems. Darüber hinaus ist es wichtig, bei der Evaluation der Genauigkeit das zu messende Phänomen korrekt abzubilden und zur Schätzung der Gütemaße realistische Testszenarien zu nutzen: Liegen etwa geschichtete Stichproben oder Cluster vor, so ist dies entsprechend zu berücksichtigen. Auch eine eventuell vorhandene zeitliche Struktur in den Daten muss beim Testen korrekt simuliert werden, um Data Leakage zu vermeiden.
c. Genauigkeit wird anhand der Distanz zwischen Schätzung und wahrem Wert gemessen. Diese Distanz kann je nach Anwendungsfall (abhängig etwa von den Konsequenzen einer Fehlklassifikation) anders interpretiert und anhand unterschiedlicher Gütemaße gemessen werden (siehe folgender Abschnitt). Dies ist in der Praxis von hoher Bedeutung, da verschiedene Gütemaße oftmals im Konflikt miteinander stehen und die Performanz verschiedener Algorithmen sich hinsichtlich bestimmter Metriken durchaus unterscheiden kann. Methodische Untersuchungen sollten daher ein Spektrum an verfügbaren Gütemaßen nutzen, um verschiedene Algorithmen auf ihre Genauigkeit hin zu untersuchen. Ausgestattet mit einer Erklärung der prinzipiellen Bedeutung der Gütemaße kann der zuständige statistische Fachbereich, der über das notwendige Kontextwissen verfügt, eine Abwägung zwischen verschiedenen Korrektheitsbegriffen vornehmen und die letztendliche Entscheidung treffen.

Tab. 2 Konfusionsmatrix bei Klassifikationsproblemen

Klassifikation | Wahrer Wert	Positiv	Negativ
Positiv	True positive (*TP*)	False Positive (*FP*)
Negativ	False negative (*FN*)	True Negative (*TN*)

Zusammenfassend unterscheidet sich der vorausgesetzte Grad der Genauigkeit je nach Anwendungsfall. Verschiedene Fragestellungen erlauben die Nutzung verschiedener Methoden, die wiederum anhand unterschiedlicher Gütemaße evaluiert werden. Der statistische Fachbereich definiert, welche „Art" der Genauigkeit besonders wichtig ist, wägt ab, welcher Grad ausreichend ist und trifft schließlich eine Auswahl zwischen denjenigen getesteten ML-Algorithmen, die den definierten Qualitätsanforderungen genügen.

3.1.1 Messung

Die Messung der Genauigkeit von ML-Algorithmen erfolgt durch die Evaluation verschiedener Gütemaße. Die meisten Gütemaße für Klassifikationsalgorithmen basieren auf dem Vergleich einer Klassifikation eines Algorithmus mit den „wahren" Werten, wie in Tab. 2 dargestellt.

Ähnlich wie bei anderen statistischen Anwendungsgebieten bieten ML-Methoden eine Vielzahl an Gütemaßen, die die Frage nach der „Korrektheit" des Modells aus unterschiedlichen Blickwinkeln beantworten. Klassifikationsalgorithmen werden beispielsweise häufig anhand folgender Gütemaße evaluiert:

- Accuracy: Anteil der richtig klassifizierten Einheiten an allen zu klassifizierenden Einheiten; $\frac{TP+TN}{TP+FP+FN+TN}$
- Sensitivität (auch: Recall): Wie viele Einheiten werden korrekterweise positiv klassifiziert? $\frac{TP}{TP+FN}$
- Spezifität: Wie viele Einheiten werden korrekterweise negativ klassifiziert? $\frac{TN}{TN+FP}$
- Relevanz (auch: Precision): Wie hoch ist der Anteil der korrekterweise positiv klassifizierten Einheiten an allen positiv klassifizierten Einheiten? $\frac{TP}{TP+FP}$
- Segreganz: Wie hoch ist der Anteil der korrekterweise negativ klassifizierten Einheiten an allen negativ klassifizierten Einheiten? $\frac{TN}{TN+FN}$
- Cohens Kappa: Bewertung der zufallskorrigierten Übereinstimmung von Testdatenklassen und Vorhersagen. Gibt es über den Zufall hinaus eine Vorhersagekraft? $\frac{2\times(TP\times TN-FN\times FP)}{(TP+FP)\times(FP+TN)+(TP+FN)\times(FN+TN)}$

Dazu gibt es weitere Gütemaße, die sich aus den genannten ableiten (z. B. Balanced Accuracy, F-Maß, Intersection over Union). Bei Regressionsproblemen können u. a. folgende Gütemaße betrachtet werden:

- R^2: Anteil der durch das Modell erklärten Varianz an der gesamten Varianz, sowie Adjusted R^2: angepasst unter Berücksichtigung der genutzten Freiheitsgrade, sodass zusätzliche unabhängige Variablen nicht mehr zwangsläufig zu einem höheren Wert führen

- Mean Absolute Error (MAE): Maß für den durchschnittlichen Schätzfehler (d. h. die durchschnittliche Abweichung der geschätzten Werte von den tatsächlichen)
- Mean Absolute Percentage Error (MAPE) oder Symmetric Mean Absolute Percentage Error (SMAPE): Maß für den relativen Schätzfehler
- Root Mean Squared Error (RMSE): Maß für die Standardabweichung des Schätzfehlers

Aufgrund der Vielzahl der verfügbaren Gütemaße ist eine Abwägung der verschiedenen Optionen notwendig: So ist bei der automatisierten Identifikation von Ausreißern mit nachgeschaltetem manuellen Review beispielsweise eine besonders hohe Sensitivität (Recall) wichtig, damit keine potenziellen Ausreißer übersehen werden. Eine höhere Relevanz (Precision) könnte aber aus wirtschaftlichen Gesichtspunkten wünschenswert sein, um den manuellen Prüfaufwand in Grenzen zu halten. Die Entscheidung, nach welchem (ggf. zusammengesetzten, gewichteten) Gütemaß schließlich die Genauigkeit gemessen werden soll, leitet sich somit aus dem Kontext des Anwendungsfalls ab und muss folglich von der Fachstatistik getroffen werden.

Nach der Auswahl der Gütemaße werden akzeptable Grenzwerte festgelegt, die den Grad der Erfüllung dieser Qualitätsdimension bestimmen. Dafür kann es hilfreich sein, ein alternatives Modell zu definieren, um den Mehrwert des ML-Verfahrens besser einschätzen zu können („Benchmarking“). Entweder kann es sich hierbei um das bestehende Verfahren handeln, das durch ein ML-Modell ersetzt wird – wenn dessen Genauigkeit denn beurteilbar bzw. dokumentiert ist – oder um ein Null-Modell. Schlussendlich leiten sich die konkreten Anforderungen an die Genauigkeit jedoch wieder aus dem fachstatistischen Kontextwissen ab: Sind etwa bei der Arzneimitteltestung ob der Konsequenzen einer Fehlklassifikation anspruchsvolle Grenzwerte unerlässlich, so genügt beispielsweise bei der Umsatzschätzung von Unternehmen ein geringerer Grad an Genauigkeit.

3.1.2 Stichprobenbedingte Unsicherheit

Wie in der statistischen Produktion üblich und mit Blick darauf, die zukünftige Performanz des ML-Verfahrens auf neuen Daten gut zu schätzen (vgl. Abschn. 3.2 in Bischl et al. 2023), sollte die Evaluation der Genauigkeit nicht nur durch Punktschätzer erfolgen, sondern auch durch die Angabe ihrer geschätzten Varianz. Oft sind Konfidenzintervalle modell- und kennzahlenagnostisch bestimmbar, etwa mittels der Wilson-Methode (Wilson 1927). Die Bestimmung verfeinerter Konfidenzintervalle ist jedoch abhängig vom betrachteten Gütemaß und der verwendeten Methodik. Bei komplexeren Kenngrößen, deren Varianz nicht ohne weiteres analytisch ableitbar oder kompliziert zu berechnen ist, können approximative statistische Schätzungen oder Resampling-Verfahren zur empirischen Varianzschätzung genutzt werden (Bruch 2015). Die Schätzung des Generalisierungsfehlers von ML-Verfahren in komplexen Stichprobendesigns ist zudem aktueller Forschungsgegenstand.

3.1.3 Qualität der Trainingsdaten

Zuletzt sollte beachtet werden, dass die beschriebenen Gütemaße lediglich den Modellfehler wiedergeben und die Trainings-, Validierungs- und Testdaten[9] als gegeben annehmen. Für eine Evaluation der Genauigkeit einer Statistik sollte jedoch auch die Qualität dieser Datengrundlage identifiziert und eingeschätzt werden. Yung et al. (2018), Puts und Daas (2021) und Puts et al. (2022) beobachten etwa, dass Trainingsdaten in der Praxis systematisch verzerrt, fehlerbehaftet oder unvollständig sein sowie zu wenige Datenpunkte enthalten können. Die Annahme, dass manuell durch Fachexperten klassifizierte Trainingsdaten fehlerfrei sind, ist ohne Einschränkungen nicht haltbar. Beispielsweise ist bekannt, dass bei Klassifikationsaufgaben menschliche Schätzer eine gewisse Fehlervarianz aufweisen. Insbesondere stimmen die manuellen Klassifikationsergebnisse mehrerer Expertinnen und Experten annähernd nie perfekt überein („inter-rater reliability“) und weichen desto stärker voneinander ab, je komplexer die Klassifikationsaufgabe ist (Kraff et al. 2020). Somit ist es möglich, dass Abweichungen eines ML-Verfahrens von der manuellen Klassifikation keinen Fehler im Modell darstellen, sondern auf Fehler in den Trainingsdaten zurückgehen, die vom Modell „entdeckt“ wurden. Im deutschsprachigen Raum sind nur wenige Beispiele aus der amtlichen Statistik bekannt, in denen Fragen der „inter-rater reliability“ untersucht und quantifiziert wurden. Jedoch lässt sich auch qualitativ einschätzen, ob und wie weit erstellte Trainingsdatensätze hinsichtlich des Grades der Korrektheit vom Ideal abweichen. Wenn in einem Anwendungsfall etwa keine Klassifikationsrichtlinien vorhanden sind und primär die manuelle Einschätzung der Sachbearbeiterinnen und Sachbereiter entscheidend ist, kann erwartet werden, dass auch ein Algorithmus Schwierigkeiten haben wird, Strukturen aus den Daten auf Basis der manuellen Klassifikationen zu erlernen. Besteht in den Trainingsdaten sogar ein Bias – etwa weil diese auf einer nicht-repräsentativen Stichprobe (z. B. Online-Befragungen) basieren – so wird dieser bei Nutzung eines maschinellen Lernverfahrens perpetuiert.

Wie diese Fehlerquelle eingeschätzt und quantifiziert werden kann, ist aktueller Forschungsgegenstand. Idealerweise sollte eine mehrfache Klassifikation der Trainingsdaten erfolgen, um die Fehlerrate der menschlichen Schätzer abschätzen und damit einen realistischen Zielwert für die Genauigkeit des ML-Algorithmus setzen zu können – dies ist in der Praxis aber aus Kosten- oder Zeitgründen selten umsetzbar. In der amtlichen Statistik etwa beschäftigte sich die Machine Learning Group 2022 von ONS[10] und UNECE u. a. mit dem Thema „Quality of Training Data“ und untersuchte „Fragestellungen mit Bezug zur Klassifizierung durch Menschen sowie Stichprobenverfahren, mit dem Ziel, repräsentative Trainingsdatensätze zu generieren“[11] (Puts et al. 2022). Ein von der Gruppe verfolgter Forschungsansatz ist, das „Total Survey Error“-Modell (Biemer 2010) zu nutzen; jedoch ist die Schätzung

[9] Im Folgenden wird aus stilistischen Gründen auf die Nennung aller drei Begriffe verzichtet; gemeint sind auch bei Erwähnung eines Begriffs i. d. R. immer Trainings-, Validierungs- und Testdaten.

[10] Das Office for National Statistics (ONS) ist das Statistische Amt des Vereinigten Königreichs.

[11] Originalwortlaut: „[with the aim of exploring] issues related to human annotation process and sampling methods to obtain representative training sets“.

Tab. 3 Indikatoren für die Evaluation der Genauigkeit maschineller Lernmodelle

Qualitätsindikatoren: Genauigkeit	
1	Statistische Methodik[a] und ML Best Practices wurden berücksichtigt (z. B. Anwendung geeigneter Resampling-Verfahren sowie Vermeidung von Data Leakage, Overfitting, Bias durch hochkorrelierte Features, Verzerrungen durch Ausreißer, abschließendes Training auf möglichst allen zur Verfügung stehenden Daten[b] ...)
2	Die aus fachstatistischer Sicht relevanten Gütemaße wurden ermittelt, d. h. es wurde festgelegt, welche Gütemaße betrachtet und bzgl. welcher Gütemaße die zu testenden ML-Verfahren optimiert werden sollen
3	Für alle aus fachstatistischer Sicht relevanten Gütemaße wurden Punktschätzer und Konfidenzintervalle angegeben
4	Gütemaße wurden für alle aus fachstatistischer Sicht relevanten Subgruppen angegeben; systematische Vorhersagefehler wurden auf ihre Relevanz hin überprüft (siehe 3.7.1 Fairness)

[a] Selbstverständlich dürfen auch bei ML grundlegende statistische Prinzipien nicht vernachlässigt oder gar außer Acht gelassen werden (Friedrich et al. 2022). Wenn in den Trainingsdaten beispielsweise Multikollinearität auftritt und dadurch Variablen selektiert werden, die keinen kausalen Zusammenhang mit den abhängigen Variablen aufweisen, kann dies auch in ML-Modellen zu Verzerrungen (im Sinne eines Bias) führen. Overfitting liegt vor, wenn ein Modell zu sehr an die Trainingsdaten angeglichen wird – überangepasste Modelle scheitern dann bei der Anwendung auf Testdaten (oder, wenn an Testdaten optimiert wird, an der Realität).

[b] In der Trainings- und Testphase wird üblicherweise nur mit einem Teil der Daten das Modell trainiert, während der verbleibende Teil zum Testen zurückgehalten wird. Wenn nach erfolgter Modellauswahl das ML-Verfahren final trainiert wird, so werden dafür alle verfügbaren Daten genutzt. Die Daten dürfen zum Zwecke des Trainings auch angepasst werden (z. B. durch Up- oder Downsampling, s. oben); hier ist alles erlaubt, was das Training verbessert. Im Gegensatz dazu soll in der (vorangegangenen) Testphase die Messung der Genauigkeit, d. h. die Schätzung der späteren Performanz des ML-Verfahrens, anhand von Daten erfolgen, die den zukünftigen Daten, auf welche das ML-Verfahren angewandt werden soll, in Inhalt und Struktur hinreichend ähnlich sind.

der einzelnen Fehlerkomponenten äußert herausfordernd und es existieren bis dato keine ausgearbeiteten Richtlinien zur Umsetzung eines solchen Modells.

3.1.4 Genauigkeit in der Praxis

Tab. 3 stellt zusammenfassend dar, welche Schritte bei der Evaluation der Qualitätsdimension „Genauigkeit“ in der Praxis zwingend erfolgen sollten.

3.2 Robustheit

Die europäischen Qualitätsprinzipien der amtlichen Statistik enthalten Anweisungen zur Nutzung von „geeigneten statistischen Verfahren“ und einer „soliden Methodik“. Zudem wird „Zuverlässigkeit“ als Grundsatz für die statistische Produktion genannt. Um zu berücksichtigen, wie verlässlich ein einmal trainiertes Modell im Produktivbetrieb sein wird, wird hier die „Robustheit“ eines ML-Verfahrens – obwohl in den Rahmenwerken nicht explizit als Qualitätsdimension genannt – als weitere Qualitätsanforderungen an die amtliche Statistik betrachtet. Hierbei handelt es sich um eine natürliche Bedingung für den Einsatz von maschinellem Lernen und anderen statistischen Verfahren im statistischen Produktionsprozess: Fehlt Robustheit, so ist die Verlässlichkeit der Verfahren nicht gegeben (Meyer und Alsabah 2022).

In der statistischen Fachliteratur ist der Qualitätsbegriff der Robustheit seit Tukey (1959), Huber (1964) und Hampel (1968) zwar fest etabliert, aber gleichzeitig nicht eindeutig definiert. Grundsätzlich bezeichnet er die Eigenschaft einer statistischen Inferenzmethode, „stabil" gegen Abweichungen der Modellannahmen, insbesondere mit Blick auf Verteilungsannahmen, zu sein (Hampel et al. 1986, S. 8). Yu und Kumbier (2020) nennen Stabilität als eines von drei Prinzipien von „wahrheitsgetreuer Data Science" und definieren sie wie folgt: „Während des Modellierens misst Stabilität, wie eine Zielgröße (‚data result') sich verändert, wenn die Daten und/oder das Modell verändert werden." Da Stabilität notwendig ist, um in der Praxis bei wiederholter Anwendung eines Modells auf verschiedene Zufallsstichproben der Grundgesamtheit verlässliche Ergebnisse zu erhalten, ist zur Erfüllung der europäischen Qualitätsprinzipien die Berücksichtigung einer Qualitätsdimension zur „Robustheit" unabdingbar. Im Kontext des maschinellen Lernens ist dies besonders relevant, da es besondere Vorteile in der automatisierten (d. h. auch wiederholten, skalierbaren) Verarbeitung großer (d. h. manuell schwer validierbarer) Datenmengen bietet. Robustheit trifft dann konkret eine Aussage darüber, ob und wie lange ein einmal trainiertes und evaluiertes Modell auch dann noch gute Ergebnisse liefert, wenn sich etwa die Grundgesamtheit verändert (z. B. über die Zeit), das Hyperparametertuning suboptimal ist, eine andere Stichprobe aus der identischen Grundgesamtheit verwendet wird oder die Annahmen für die Gültigkeit des Modells nicht voll erfüllt sind. In Formeln (stilisiert) ausgedrückt, wird von einem robusten maschinellen Lernverfahren erwartet, dass es die Veränderung der Zielgröße A beschränkt:

$$||A(U_1) - A(U_0)|| \leq ||U_1 - U_0||$$

wobei U_1 und U_0 die Umstände sind, unter denen die interessierende Größe A bestimmt (oder geschätzt) wurde. Häufig wird dabei ergänzt, dass diese Ungleichung nur für kleine Unterschiede in den Umständen erfüllt sein muss, dass also keine globale, stetige Abhängigkeit der Zielgröße von den Umständen gelten muss.

Diesem Verständnis folgend besteht auch eine Beziehung zwischen Robustheit und zwei anderen Qualitätsdimensionen: So kann ***Reproduzierbarkeit*** im weiten Sinne auch die Replizierbarkeit beinhalten – also die Eigenschaft eines Modells, trotz kleinerer Abweichungen (Perturbationen) in den Daten und der Methodik das ungefähr gleiche Ergebnis zu liefern (siehe 3.4 Reproduzierbarkeit). Die Qualitätsdimension der „Robustheit" beschreibt diesen Aspekt jedoch unmittelbar, während ein robustes Modell lediglich – wenngleich notwendigerweise – einen Beitrag zur Reproduzierbarkeit leistet. Auch die Qualitätsdimension der ***Genauigkeit*** ist mit der Robustheit verwandt. So beschreiben Konfidenzintervalle von Gütemaßen, wie oben beschrieben, wie genau das Verfahren Gütemaß-Punktschätzungen bestimmt. Dadurch leiten sich auch Aussagen über die Robustheit der Punktschätzungen gegenüber stichprobenbedingten Datenperturbationen ab. Da es sich hierbei aber primär aber um eine Meta-Aussage über die „Genauigkeit der geschätzten Genauigkeit" handelt, wird dieser Aspekt der Genauigkeitsdimension zugeordnet.

Weiterhin ist wichtig zu verstehen, was Gegenstand der Betrachtung Robustheit als Qualitätsdimension ist: Welche Zielgröße soll hinsichtlich der o. g. Aspekte

robust sein? Die interessierende Größe[12] *A* könnte beispielsweise für konkrete, datenpunktbezogene Schätzungen (bei Klassifikation und Regression) stehen, beispielsweise für die Zuordnung von statistischen Einheiten zu Klassen eines Klassifikationsschemas.[13] Auch könnte *A* Modellparameter wie beispielsweise die Koeffizienten bei Lasso beschreiben; in diesem Fall stünde die robuste Schätzung des Einflusses einer erklärenden Variablen auf die Schätzung im Vordergrund. Für die Anwendung von ML in der amtlichen Statistik erscheint dieser Gegenstand von Robustheit allerdings regelmäßig weniger relevant, da i. d. R. die Prädiktion und selten die Bedeutung einzelner Parameter im Fokus steht. Ist das primäre Ziel eher, ein im Mittel brauchbares Modell zu trainieren, sollte das Augenmerk auf die Robustheit der Genauigkeitsmaße des betrachteten ML-Verfahrens (siehe 3.1 Genauigkeit) gerichtet werden. Zuletzt kann die Robustheit sich auch auf (mittels ML-Verfahren bestimmte) Aggregate beziehen, wie beispielsweise die Umsätze je Wirtschaftszweig oder das Außenhandelsvolumen je Unternehmensgrößenklasse. In diesem Fall wäre etwa bei einer Klassifikation eine nicht-robuste Zuordnung von statistischen Einheiten zu den Klassen weniger kritisch, solange die Schätzung der Gesamtumsätze je Klasse robust ist.

Im statistischen Produktionsprozess ist Robustheit bezüglich einer solchen „prozessabwärts" zu bestimmenden Zielgröße wie des Gesamtumsatzes (im Sinne der „Verlässlichkeit") regelmäßig das eigentliche Ziel, um den Qualitätsanforderungen an die statistischen Produkte (siehe 2.1 Qualitätsgrundsätze für statistische Prozesse) genüge zu tun. Gleichzeitig ist Robustheit auf dieser Ebene aufgrund der vielen Zwischenschritte zwischen dem Einsatz des ML-Verfahrens und der Bestimmung der Zielgröße am schwierigsten zu untersuchen: Um die Robustheit eines ML-Verfahrens in Bezug auf bestimmte Aggregate zu bestimmen, müsste die Abhängigkeit der prozessabwärts zu bestimmenden Zielgröße modellierbar sein oder automatisiert berechnet werden können. In diesem Fall wäre es möglich, unter Nutzung von Resampling-Methoden zu analysieren, wie sich künstliche Störungen in einem Teilschritt auf das statistische Endprodukt und die daraus abgeleiteten Aggregate auswirken.[14] Jedoch ist in der amtlichen Statistik eine automatisierte Verschaltung

[12] Der Singular soll nicht ausdrücken, dass es sich bei *A* um eine Zahl handeln muss. Auch ein Vektor, eine Matrix o. ä. sind denkbar. Durch Einsatz von Normen, (gewichteten) Summen von Einträgen, Extremwertbetrachtungen usw. ist eine Reduktion auf einen Skalar immer möglich.

[13] Klassifikationsschemata sind in der amtlichen Statistik weit verbreitet und in der Regel auch international abgestimmt. Beispiele sind die „Statistische Systematik der Wirtschaftszweige in der Europäischen Gemeinschaft" oder die „Classification of Individual Consumption by Purpose" (COICOP). Anwendungsfälle für Klassifikation in der amtlichen Statistik können aber auch binär sein, beispielsweise um die Einschlägigkeit von Berichtseinheiten für die vorliegende Statistik einzuschätzen (vgl. Feuerhake und Dumpert 2016).

[14] Beispielsweise erlaubt eine an die multiple Imputation angelehnte Vorgehensweise, die mehrere Datensätze aus einer Verteilung generiert, prinzipiell die Schätzung von Unsicherheiten im statistischen Endprodukt. Um das zu ermöglichen, müssten weitere Aufbereitungsschritte aber für alle multipel imputierten Datensätze erfolgen, was in der Praxis aufgrund der mangelnden automatisierten Verschaltung und des hohen Personalaufwands einer manuellen Implementierung nicht möglich ist. Stattdessen wird derzeit einer dieser Datensätze ausgewählt, als neue „Ground Truth" definiert und für die weiteren Schritte der Statistikproduktion verwendet (Preising et al. 2021).

Tab. 4 Aspekte der Robustheit von ML-Verfahren

Datenperturbationen
Robustheit hinsichtlich alternativer Ziehungen aus der Grundgesamtheit mit gleichbleibenden Verteilungseigenschaften (Stichprobenvariabilität)
Robustheit hinsichtlich vorgelagerter Schritte (z. B. Transformationen, Plausibilisierungen etc.)
Robustheit hinsichtlich nicht-stichprobenbedingter Fehlerquellen (z. B. Fehler in den Trainingsdaten)
Modellperturbationen
Robustheit hinsichtlich alternativer Ziehungen aus einer anderen Grundgesamtheit (i. S. v. anderen Verteilungseigenschaften) oder bewusst manipulierter Stichproben
Robustheit hinsichtlich alternativer Ziehungen aus der prinzipiell gleichen Grundgesamtheit, jedoch zu einem anderen Zeitpunkt als die ursprüngliche Stichprobe (Concept Drift)
Robustheit hinsichtlich Perturbation von Modellparametern (sowohl von optimierten Parametern als auch von Hyperparametern)

der für die verschiedenen GMAS-Phasen[15] genutzten Werkzeuge derzeit i. d. R. nicht gegeben; zudem würde ein solches Resampling-Verfahren hohe Anforderungen an die Rechenleistung stellen.

Bei Modellen von üblichem Komplexitätsgrad ist es somit nicht ohne Weiteres möglich, eine Aussage darüber zu treffen, wie die Unsicherheiten der einzelnen Schätzer interagieren und welche Eigenschaften die veröffentlichte Statistik hat. Folglich lässt sich festhalten: Auch wenn die Bestimmung der Robustheit hinsichtlich einer GMAS-prozessabwärts zu bestimmenden Zielgröße idealerweise das Objekt der Robustheitsuntersuchungen sein sollte, so ist dies aktuell unter praktischen Gesichtspunkten kaum umsetzbar.[16] Ersatzweise sollten Robustheitsuntersuchungen die Robustheit der Modellgüte (gemessen an ausgewählten Gütemaßen) betrachten. Darüber hinaus sollte bei Bedarf stets die Robustheit von einigen konkreten, datenpunktbezogenen Schätzungen (z. B. von Unternehmen mit besonders hohem Umsatz, hoher Beschäftigtenzahl, ...) sichergestellt werden.

Bis hierhin wurde das Konzept der Robustheit allgemein definiert und von anderen Qualitätsdimensionen abgegrenzt. Konkret kann Robustheit im Hinblick auf verschiedene Aspekte bewertet werden (siehe teilweise auch Yu und Kumbier 2020). Diese sind in Tab. 4 zusammengefasst und werden im Folgenden im Detail beschrieben, sowie in Hinblick auf ihre Messung diskutiert.

[15] Vgl. hierzu das Geschäftsprozessmodell Amtliche Statistik (Blumöhr et al. 2017) bzw. das GSBPM (UNECE 2019). Der Einsatz von ML erfolgt häufig – wenngleich nicht ausschließlich – in den Phasen 5.2 bis 5.4 dieser Prozessmodelle.

[16] Auch wenn etwa eine quantitative Evaluation des Gesamtfehlers eines veröffentlichten statistischen Aggregats nicht möglich ist, so ist zumindest eine qualitative Einschätzung angemessen. Dies ist besonders relevant, wenn Verarbeitungsprozesse teils außer Haus geschehen – etwa im Falle von nicht-traditionellen, sogenannten Neuen Digitalen Daten. Mobilfunkdaten beispielsweise entstehen als Resultat einer komplexen Kette von Bereinigungsschritten, die mit einer Vielzahl von potenziellen Ungenauigkeiten und Fehlern einhergehen (Saidani et al. 2022). Selbst aus einer sehr hohen Modellgenauigkeit kann man somit nicht unbedingt schließen, dass die entstandenen statistischen Produkte das Phänomen von Interesse unverzerrt und mit einer hohen Genauigkeit beschreiben.

3.2.1 Datenperturbation

Ein robustes Modell sollte nicht primär Aussagen über den Trainingsdatensatz treffen, sondern auch mit anderen Stichproben ähnliche oder ähnlich gute Ergebnisse liefern wie während des Trainings. Somit ist es zur Erlangung von Robustheit erforderlich, Overfitting zu vermeiden: Gelernt werden sollen nicht alle Spezifika der vorliegenden Daten, sondern generalisierbare Eigenschaften der Stichprobe. Um zu verhindern, dass bestimmte Beobachtungen die Ergebnisse übermäßig beeinflussen, kann in der Praxis getestet werden, ob das Entfernen weniger Beobachtungen einen Effekt hat (Yu 2013, S. 6). Eine allgemeinere Form der Datenperturbation ist das Resampling; darunter fallen etwa Methoden wie Jackknife, Sub-Sampling und Bootstrapping.

Neben zufälligen Perturbationen sind je nach Anwendungsfall auch gezielte (verteilungserhaltende) Manipulationen sinnvoll, um die Robustheit eines Modells zu testen: Soll beispielsweise eine Wirtschaftszweigklassifikation von Unternehmen erfolgen, kann durch das zufällige Austauschen von Labels einer bestimmten Hierarchiestufe im Testdatensatz getestet werden, ob die hierarchische Struktur der Daten erfolgreich gelernt wurde. Ein robustes Modell sollte mit dem veränderten Datensatz bei der Klassifikation von höheren Hierarchiestufen eine ähnliche Genauigkeit erreichen.

Die Robustheit gegenüber Datenperturbationen sollte in der Praxis getestet und dokumentiert werden. Allgemeine Regeln und Prinzipien zu formulieren fällt ob der Breite an genutzten Modellen schwer. Zwar schlägt bereits Andrews (1986) vor, die Robustheit eines Regressionsmodells als „ein Maß dafür, wie groß der Effekt einer einzelnen Beobachtung in der Stichprobe auf den realisierten Wert des Schätzers [ist]“ (ebd., S. 1207) zu betrachten. Die Frage, welche Gütemaße hierfür im Kontext maschineller Lernverfahren genutzt werden können, ist aber nicht geklärt und aktueller Forschungsgegenstand: Yu und Kumbier (2020) etwa „entwickeln Prozeduren, die – aufbauend auf dem Konzept des PCS [‚predictability, computability, and stability‘], insbesondere auf PCS-Perturbationsintervallen und PCS-Hypothesentests – relativ zur Problemformulierung, Datenaufbereitung, Modellierungsentscheidungen und Interpretationen eine Aussage über die Stabilität eines ‚data results‘ treffen“ (ebd., S. 1). Liu et al. (2022) diskutieren, wie Stabilität (im Sinne der Robustheit gegenüber kleinen Datenperturbationen) im Falle von Unsupervised Clustering operationalisiert und gemessen werden kann. Zu neuronalen Netzwerken im Besonderen existiert eine umfangreiche und wachsende Literatur unter dem Schlagwort „Adversial Robustness“, die sich mit der Verlässlichkeit von solchen Modellen im Umgang mit absichtlich manipulierten Eingangsdaten beschäftigt. Die Gütemaße, die zur Messung der Anfälligkeit entwickelt wurden (oft: die minimale Perturbation der Eingangsdaten, die notwendig ist, um eine Fehlklassifikation zu erreichen), sind prinzipiell auch im Kontext der amtlichen Statistik anwendbar.

Zuletzt sollte neben der praktischen Performanz auch auf die theoretischen Eigenschaften der genutzten Modelle geachtet werden. Beispielsweise ist nach Yu (2013, S. 8) „Lasso-Regression kombiniert mit Kreuzvalidierung nicht stabil gegenüber Bootstrapping- oder Subsampling-Perturbationen, wenn die Prädiktorvariablen untereinander korreliert sind“. Darauf aufbauend entwickeln Lim und Yu (2016)

eine alternative Kreuzvalidierungsmethodik. Grundsätzlich sollte daher gelten, wo möglich robuste Schätzverfahren oder Methoden, deren Robustheitseigenschaften theoretisch belegt sind, zu verwenden.

In der Praxis hängt die Testung der Robustheit gegenüber Datenperturbationen von der Art des Modells ab. Programmierer von ML-Algorithmen in der amtlichen Statistik sollten daher Entwicklungen in der Fachliteratur verfolgen und sich nach den aktuellen wissenschaftlichen Standards richten.[17]

3.2.2 Modellperturbation: Verletzung von Verteilungsannahmen

Statistische Modelle garantieren wünschenswerte Eigenschaften wie Erwartungstreue, Konsistenz und Effizienz nur unter der Voraussetzung, dass die Eingangsdaten bestimmte Annahmen hinsichtlich ihrer Verteilung erfüllen. Was passiert, wenn Annahmen des Modells nicht erfüllt sind? Zwar mag man naiv hoffen, dass Abweichungen in der Praxis von geringer Konsequenz sind – in den Worten von J. W. Tukey „hatte man, als man über Abweichungen vom idealen Modell hinwegsah, die unterschwellige Hoffnung, dass diese Abweichungen nichts ausmachen würden; dass statistische Verfahren, die im Falle des richtigen Modells optimal waren, auch mit einem halbwegs richtigen Modell halbwegs optimal sein würden. Leider stellte sich heraus, dass diese Hoffnung oft komplett falsch ist; selbst geringe Abweichungen haben oft einen deutlich größeren Effekt als die meisten Statistiker erwarten würden“ (zitiert nach Hampel et al. 1986).

Daher ist es unabdingbar, im Einzelfall zu evaluieren, wie groß der Einfluss unerfüllter Modellannahmen ist. Wird die Robustheit des Modells gegenüber einer Verletzung dieser Annahmen getestet, ist von Modellperturbation die Rede. Hier wird – anders als bei der Datenperturbation – bewusst die Verteilung der Eingangsdaten manipuliert, um zu beobachten, wie stark das Modell von den bestimmten Verteilungseigenschaften der Trainings-, Validierungs- und Testdaten abhängt. Ein klassisches Beispiel ist die Frage, wie das Modell mit Daten aus endlastigen Verteilungen (wie den t- oder Pareto-Verteilungen) umgeht. Auch asymmetrische Ausreißer – ein häufiges Phänomen in der amtlichen Statistik – können je nach genutzter Methode einen großen Einfluss auf die Genauigkeit des Modells haben. Auch Feature-Permutation kann in diesem Zuge betrieben werden, indem man im Datensatz spaltenweise Werte verändert.

In der Praxis ist es wichtig, dass die getroffenen Annahmen an die Eingangsdaten bekannt sind und dokumentiert werden. Darauf aufbauend sollte eine Verletzung dieser Annahmen simuliert werden, um die Robustheit des Modells gegenüber Modellperturbationen festzustellen. Auf dieser Grundlage können dann etwa Parameter festgelegt werden, die im laufenden Betrieb überwacht werden und bei der Überschreitung von Grenzwerten manuellen Review nötig machen.

[17] Exemplarisch für eine Zusammenfassung des wissenschaftlichen Standes sei im Fall von Support Vector Machines auf Köhler und Christmann (2022) und die dort zitierten Arbeiten verwiesen.

3.2.3 Modellperturbation: Concept Drift

Das Konzept des Concept Drifts bezieht sich ebenfalls auf Situationen, in denen die Verteilung der aktuellen Eingangsdaten von der Verteilung der Trainingsdaten abweicht – entweder in Bezug auf die erklärenden Variablen (Feature Drift), die zu erklärende Variable (Target Drift) oder den funktionalen Zusammenhang zwischen der zu erklärenden und den erklärenden Variablen (Posterior Drift). Während sich die im vorigen Abschnitt diskutierte Modellperturbation auf die Eigenschaften einer statischen Verteilung konzentriert, beschreibt Concept Drift das Phänomen, dass sich im Laufe der Zeit die Verteilung der Grundgesamtheit verschieben kann. Dies ist in Anwendungsfällen der amtlichen Statistik besonders relevant, da produktive Prozesse mit maschinellen Lernmodellen grundsätzlich auf eine lange, mehrjährige Laufzeit angelegt sind. Aus diesem Grund wurde diese Fragestellung im Rahmen des internationalen ONS-UNECE-ML-2022-Projekts zur Verbesserung der amtlichen Statistik vertieft behandelt (Choi et al. 2022).

Je nach Erhebungsinfrastruktur ist es denkbar, dass Concept Drift eine direkte Auswirkung auf die Repräsentativität der Stichprobe hat: Korreliert die Variable, deren Verteilung sich verschiebt, mit der Auswahl- oder Antwortwahrscheinlichkeit bei einer Umfrage, dann unterscheiden sich die erfassten Daten im Laufe der Zeit möglicherweise systematisch von der Grundgesamtheit. In diesem Fall sind weitläufige Anpassungen des gesamten Erhebungs- und Datenverarbeitungsprozesses notwendig, insbesondere auch ein Neutrainieren des ML-Modells (anhand eines ggf. neu zu erstellenden Trainingsdatensatzes, vgl. 3.1.3 Qualität der Trainingsdaten). Dieses Problem kann zwar grundsätzlich bei allen Produkten der amtlichen Statistik auftreten. Da maschinelles Lernen oft mit Prozessautomatisierung einhergeht, erfordert der Einsatz von ML-Algorithmen aber besondere Vorsicht und das Treffen von Maßnahmen zur automatisierten Detektion von Concept Drift. Veränderungen in der Verteilung können prinzipiell über statistische Distanzmaße[18] bestimmt werden, die bei Überschreitung bestimmter Schwellenwerte einen Anpassungsbedarf signalisieren (Goldenberg und Webb 2019; Meertens et al. 2022). Je nach Datentyp (z. B. Textdaten) ist ein solcher traditioneller Ansatz jedoch nicht immer praktikabel. Alternativ kann ein zweites ML-Modell implementiert werden, das fortwährend versucht, vorhandene von neu eingehenden Beobachtungen zu unterscheiden. Sobald der neue Zustand der Population zunehmend zuverlässig vom früheren unterschieden werden kann, ist dies ein Hinweis auf Drift (mindestens in den erklärenden Variablen) und ein Neutrainieren des primären Modells ist möglicherweise angebracht. Auch ein auf Basis von stichprobenartigen Prüfungen festgestellter Abfall der Performanz des ML-Verfahrens kann ein Hinweis auf Drift (in allen drei oben genannten Ausprägungen) sein (Choi et al. 2022, S. 8f.).

Selbst wenn Concept Drift keine Auswirkung auf die Repräsentativität der Stichprobe hat, kann dieser Konsequenzen für die Genauigkeit von maschinellen Lernmo-

[18] Es gibt eine große Zahl an denkbaren Distanzmaßen. Choi et al. (2022, S. 9) sowie Thurow et al. (2021, S. 19ff.) nennen im Kontext von Fragestellungen der amtlichen Statistik einige Beispiele. Die Maße unterscheiden sich bezüglich ihrer Voraussetzungen, ihrer Aussagen und ihrer Empfindlichkeit hinsichtlich spezieller Abweichungen in den Verteilungen.

dellen haben: Verschiebt sich etwa bei einem Klassifizierungsmodell, dessen Parameter einmal für den Trainingsdatensatz optimiert wurden, im Laufe der Zeit die zugrundliegende Verteilung der Eingangsdaten, so ist meistens[19] eine Verschlechterung der Genauigkeit des Modells zu erwarten. Im schlimmsten Fall führen stabile Parameter in diesem Fall zu einem Bias, sodass die Erwartungstreue des Modells nicht mehr gegeben ist. Grundsätzlich kann auch mit dieser Situation umgegangen werden, indem das Modell neutrainiert wird oder indem Methoden zur Bias-Korrektur genutzt werden. Da ersteres oft kostspielig ist, werden in der Praxis oft – sofern möglich – letztere genutzt.[20] Alternativ wird nicht mit allen Daten neutrainiert, oder ältere Beobachtungen werden geringer gewichtet. Die optimale Methodik ist aktueller Forschungsgegenstand.[21]

Anders als bei anderen Aspekten der Robustheit ist das Ziel bei Concept Drift nicht die Verhinderung; da es sich um ein Phänomen handelt, das durch externe Prozesse entsteht, ist dies auch gar nicht möglich. Stattdessen gilt es zu vermeiden, dass Concept Drift unbemerkt geschieht und das implementierte Modell obsolet macht – sei es durch Prozesse, die Drift erkennen und ausweisen können, oder durch Modelle, die durch Bias-Korrektur mit Drift umgehen können. Ein regelmäßiges Neutrainieren zum Umgang mit unentdeckten Drifts erscheint letztlich dennoch geboten.

3.2.4 Modellperturbation: Hyperparameter

Ein besonderes Merkmal von Methoden des maschinellen Lernens ist die Anzahl und Bedeutung von Hyperparametern und ihres Tunings (Bartz et al. 2023; Bischl et al. 2023). Für einen qualitätsgesicherten Einsatz von ML in der amtlichen Statistik ist ein angemessenes Verständnis für die Wirkung der Hyperparameter und ein entsprechendes Tuning notwendige Voraussetzung (Dumpert und Schmidt 2023). So haben die Eingangsdaten nicht nur einen direkten Effekt auf das Ergebnis durch die im Modell definierten Zusammenhänge, sondern sie können auch Einfluss auf das Verfahren selbst haben, wenn einzelne Hyperparameter bei jedem Durchlauf neu optimiert werden. Ein Modell ist folglich hinsichtlich seiner Hyperparameter robust, wenn generierte Ergebnisse oder Genauigkeitsmaße bei leichten Perturbationen der Hyperparameter stabil bleiben. „Leichte Perturbationen“ meint dabei die Variation von Hyperparametern in der Nähe einer optimalen Hyperparameterkonstellation oder auf Basis von fortgeschrittenen Tuningverfahren (wie beispielsweise solchen, die den Zusammenhang zwischen Hyperparameterkonstellationen und Performanz

[19] Drift kann auch in einer Art und Weise geschehen kann, die die Validität des Modells nicht beeinträchtigt: „Wenn beispielsweise die Verteilung einer Feature-Variablen X sich in einer Art und Weise verändert, sodass [die Zusammenhänge im] Modell ihre Gültigkeit behalten, dann sinkt auch die Modellgüte nicht.“ (Choi et al. 2022, S. 5)

[20] Die schlechtere Wirtschaftlichkeit durch häufiges Neutrainieren ist in der Praxis jedoch abzuwägen gegen andere Qualitätsdimensionen, wie z. B. die Relevanz der veröffentlichten Ergebnisse.

[21] Beispielsweise diskutieren Meertens et al. (2022) die Performanz zweier Korrekturalgorithmen bei einer Verschiebung der A-Priori-Wahrscheinlichkeiten („prior probability shift“).

des Modells erlernen und abbilden).[22] Dies ist umso wichtiger, je weniger fachliche Begründung für eine bestimmte Hyperparameterwahl existiert: Aus Sicht der Fachstatistik relevante Abweichungen der Zielgröße durch leichte Änderungen von quasi-zufälligen Werten (wie dem Startwert bei einem Gradient Descent oder dem für die Generierung von Zufallszahlen genutzten Seed) sind aus Sicht der Robustheit inakzeptabel (Yu und Kumbier 2020, S. 6). Bei Hyperparametern mit inhaltlicher Bedeutung kann eine Sensitivitätsanalyse (Saltelli et al. 2007) und die Visualisierung des Tunings (Bartz et al. 2023, Kap. 8–10) ebenso aufschlussreich sein, auch wenn hier fachliche Begründungen eine spezifische Parameter-Wahl rechtfertigen können.

3.2.5 Auswahl robuster Verfahren

Darüber hinaus sollte bereits bei der Modellformulierung darauf geachtet werden, robuste Modelle zu erstellen, beispielsweise durch Auswahl von Verfahren mit theoretisch garantierten Robustheitseigenschaften[23] oder indem die Anzahl an genutzten Variablen möglichst reduziert wird. Aufgrund der multivariaten Komplexität des Optimierungsproblems in maschinellen Lernmodellen ist der geschätzte Einfluss einer erklärenden Variablen i. d. R. von allen anderen abhängig. So ist der effektivste Weg, um unbeabsichtigte Effekte auf das Modell durch einzelne Variablen zu minimieren, Variablen mit geringem zusätzlichem Wert für das Modell zu streichen. Sculley et al. (2015) empfehlen etwa, „ungenutzte Abhängigkeiten [d. h. Variablen, die nur einen geringen Beitrag zur Modellgüte leisten] mittels umfangreicher Leave-One-Out-Analysen zu ermitteln. Diese sollten regelmäßig durchgeführt werden, um überflüssige Features zu identifizieren und zu entfernen." (ebd., S. 3).

Modelle, die hinsichtlich ihrer Hyperparameter nicht robust sind, bieten wenig Vertrauen in die Fähigkeit des Modelles, das Signal in den Daten von zufälligen Effekten zu trennen. Ist eine Modellspezifikation nicht stabil, kann nicht davon ausgegangen werden, dass der Algorithmus allgemein genug ist, um von den Trainingsdaten ausgehend generalisierbar zu sein und mit anderen Daten gleich gute Ergebnisse zu erzielen.

3.2.6 Robustheit in der Praxis

Um in der amtlichen Statistik robuste maschinelle Lernmodelle zu erstellen, sind alle oben genannten Robustheitsaspekte zu berücksichtigen. In der Praxis kann man für einen bestimmten Anwendungsfall von einem „hinreichend robusten Verfahren" sprechen, wenn bei der Ausführung bestimmter Tests keine fachstatistisch relevante Änderung der Zielgröße auftritt. Tab. 5 enthält eine Checkliste solcher Tests, die Quellen mangelnder Stabilität aufdecken können und so das Vertrauen in die Stabili-

[22] Nicht gemeint ist der Vergleich der Ergebnisse basierend auf verschiedenen, mehr oder weniger willkürlich ausgewählten Hyperparameterkonstellationen (wie sie bei Grid Search oder Random Search auftreten) – denn dass sich dabei in der Regel keine Robustheit einstellt, ist zu erwarten.

[23] Die „Robustifizierung" von Schätzern (für Klassifikations- und Regressionsprobleme) und Hypothesentests ist eine wichtige Teildisziplin der robusten Statistik. Aus Sicht der amtlichen Statistik ist neben den theoretischen Eigenschaften solcher robuster Verfahren entscheidend, dass die zugehörigen Algorithmen auch ausgereift und effizient implementiert (z. B. in R) vorliegen.

Tab. 5 Indikatoren für die Evaluation der Robustheit maschineller Lernmodelle

Qualitätsindikatoren: Robustheit	
1	Zielgrößen, die Gegenstand der Robustheit sein sollen, wurden definiert
2	Bei der Vorauswahl zu untersuchender ML-Verfahren wurden robuste Verfahren berücksichtigt
3	Die Auswirkungen von Datenperturbationen auf die Zielgröße in Form eines geeigneten Resampling-Verfahrens wurden untersucht
4	Die Auswirkungen von Datenperturbationen in Form der bewussten Manipulation oder Hinzufügung grob falscher Datenpunkte wurden untersucht
5	Die Auswirkungen von Modellperturbationen in Form einer Verletzung der Annahmen an das Modell wurden untersucht
6	Ein Verfahren zur Detektion von Concept Drift wurde implementiert
7	Die Auswirkungen von Modellperturbationen in Form verschiedener Hyperparameterauswahlen wurden untersucht

tät des Modells stärken. Ob das eingesetzte ML-Verfahren im Einzelfall ausreichend robust für den Einsatz im Produktionsprozess der amtlichen Statistik ist, ist eine fachstatistische Entscheidung.

3.3 Erklärbarkeit

Wenn ein statistischer Algorithmus Aufgaben übernehmen soll, die vormals von Menschen erledigt wurden, dann muss er in der Regel zuerst einmal das Vertrauen der Nutzenden gewinnen, um den Verzicht auf umfassende menschliche Kontrolle und die Übertragung menschlicher Urheberschaft auf die Maschine zu rechtfertigen: Verhält sich das Modell hinsichtlich der Fehlerrate „ähnlich“ wie ein Mensch? Besitzt es die beim Menschen stark ausgeprägte Kapazität, auf geänderte Umstände zu reagieren? Liefert es über die reinen Vorhersagen hinaus auch ein Verständnis der tieferen Zusammenhänge und Kausalitäten in den Daten? Nutzt das Modell Muster zur Vorhersage, die aus ethischer Sicht problematisch sind? Diese berechtigten Fragen können teilweise entkräftet werden, wenn ein ML-Modell gewisse wünschenswerte Eigenschaften aufweist, die unter dem breiten Begriff der „Erklärbarkeit“ (auch: Transparenz, Interpretierbarkeit) zusammengefasst werden. Auf EU-Ebene definieren die Ethics Guidelines for Trustworthy AI (AI HLEG 2019), das White Paper on AI (European Commission 2020) und das AI Framework (European Parliament 2020) die Erklärbarkeit als zentralen Wert bei der Etablierung von KI. Was ist jedoch konkret unter dieser Qualitätsdimension zu verstehen? Nach Lipton (2018) ist zwischen zwei Erklärbarkeitsbegriffen zu unterscheiden:

a. Erklärbarkeit im Sinne der ***Transparenz der Methodik***, auf verschiedenen Ebenen: Simulierbarkeit („simulatability“) erlaubt es Menschen, „das Modell in seiner Gänze mental zu erfassen“[24] (ebd., S. 13). Zerlegbarkeit („decomposability“) ist gegeben, wenn jeder Bestandteil des Modells – insbesondere die Parameter, aber auch etwaige generierte Features – eine intuitive Bedeutung aufweisen. Al-

[24] Originalwortlaut: „In the strictest sense, a model might be called transparent if a person can contemplate the entire model at once.“

gorithmische Transparenz („algorithmic transparency“) beschreibt das Vorhandensein wohldefinierter theoretischer Eigenschaften wie Optimalitätsbedingungen und Konvergenzverhalten.

b. Erklärbarkeit im Sinne der ***Post-hoc-Interpretierbarkeit der Ergebnisse***. Diese liegt vor, wenn es möglich ist nachzuvollziehen, welche Zusammenhänge der Algorithmus nutzt, um Vorhersagen zu treffen oder Kategorien zuzuweisen. Dies kann etwa durch automatisch generierte textliche Erklärungen, durch Visualisierungen oder anhand von ausgewählten Beispielen geschehen.

Für die amtliche Statistik ist der erste Begriff nur teilweise von Bedeutung: Um Vertrauen in maschinelle Lernverfahren zu schaffen, ist es durchaus essenziell, dem europäischen Qualitätsprinzip „Zugänglichkeit und Klarheit“ folgend die Methodik transparent darzustellen, genutzte Verfahren detailliert zu dokumentieren und die grobe Funktionsweise des Modells auch Nicht-Experten begreifbar machen zu können. Jedoch muss für ein tieferes Verständnis – wie bei jeder Fachmethode – ein gewisses Maß an (statistischem und mathematischem) Wissen und Erfahrung vorausgesetzt werden. Simulierbarkeit ist selbst bei einem einfachen, linearen Modell nicht gegeben, wenn es viele Variablen oder Interaktionsterme enthält (ebd., S. 13). Zerlegbarkeit ist wichtig, wenn ein Modell bei Entscheidungen unterstützen oder tiefergehende Einsichten liefern soll, aber bei Anwendungsfällen in der amtlichen Statistik mit Einbindung in automatisierte Abläufe weniger relevant. Algorithmische Transparenz ist theoretisch wünschenswert, aber in der Praxis angesichts des erwiesenen Erfolges von ML-Modellen bei Prädiktionsaufgaben nicht wesentlich. Nichtsdestotrotz ist es möglich, dass Fachbereiche im Einzelfall begründete Anforderungen an die Transparenz der ML-Methodik stellen.

Die Interpretierbarkeit der Ergebnisse ist beim Einsatz maschineller Lernverfahren in der amtlichen Statistik der primär relevante Erklärbarkeitsbegriff. So begründet der Deutsche Ethikrat in seiner Stellungnahme zu den Herausforderungen Künstlicher Intelligenz den Bedarf nach Erklärbarkeit durch eine „technisch bedingte Opazität“ (Deutscher Ethikrat 2023, S. 142) bei bestimmten Klassen maschineller Lernmodelle. Diese läge vor, wenn „die algorithmischen Strategien, die im Laufe des Trainings zur Bewältigung der jeweiligen Aufgaben entwickelt werden [...] selbst für geschultes Personal, das den Code vollständig einsehen kann, nicht auf Anhieb nachvollziehbar [sind]“ (ebd., S. 72). Beispielsweise können umfangreiche Sprachmodelle im Bereich des Natural Language Processing[25] sehr genaue Ergebnisse generieren (Devlin et al. 2019), die sich aber praktisch nicht mehr nachvollziehen oder erklären lassen (Kovaleva et al. 2019).[26] Die konkreten Anforderungen an Interpretierbarkeit seien „in Abhängigkeit von den jeweiligen Zielen, die mit Transparenz und Nachvollziehbarkeit verfolgt werden, nach den Personen, die Informationen erhalten, und nach dem jeweiligen Anwendungskontext zu konkretisieren.“ (Deutscher

[25] Natural Language Processing (NLP) beschreibt Techniken und Methoden zur maschinellen Verarbeitung natürlicher Sprache.

[26] Vor allem bei Anwendungen mit Bezug zu NLP zeigt sich daher ein deutlicher Zielkonflikt zwischen der Genauigkeit und der Erklärbarkeit eines Algorithmus, der in der Praxis durch eine Abwägung und Priorisierung seitens des statistischen Fachbereichs aufgelöst werden muss.

Ethikrat 2023, S. 285). Im Anwendungskontext der amtlichen Statistik sind hier drei mögliche Zielgruppen denkbar, für die Erklärbarkeit unterschiedliche Aufgaben erfüllt: ML-Entwickler bzw. Data Scientists, statistische Fachbereiche und Externe (beispielsweise im Falle einer rechtlichen Verpflichtung).

Die in der amtlichen Statistik tätigen ***Data Scientists*** sollten nachvollziehen, wie der ML-Algorithmus aus den Eingabevariablen Vorhersagen oder Klassifikationen generiert, um einschätzen zu können, wie verlässlich und wie generalisierbar das entwickelte Modell ist. Dafür sollte betrachtet werden, welche Variablen einen großen Einfluss auf das Ergebnis haben („variable importance“) und wie das Ergebnis konkret von den einzelnen Variablen abhängt. Bei bestimmten, einfachen Modellen ist dies aus dem Modell an sich ablesbar. Wenn Variablen allerdings mehrfach vorkommen (z. B. linear und quadratisch) oder wenn Interaktionen (insb. höherer Ordnung) von Variablen auftreten, können Variableneffekte nicht mehr ohne Weiteres abgeleitet werden. Andere Verfahren, insbesondere neuronale Netze, lernen nicht einmal explizit Variablenzusammenhänge und weisen demnach auch keine festen, globalen Beziehungen zwischen Eingabe- und Ausgabevariablen auf. In Anwendungsfällen, in denen aufgrund großer Unterschiede in den übrigen Qualitätsdimensionen nicht auf interpretierbare Modelle zurückgegriffen werden kann, können alternativ modell-agnostische Methoden des Explainable AI bzw. Interpretable ML genutzt werden: Global Sensitivity Analysis (GSA) erlaubt etwa, Veränderungen in der Zielvariable auf einzelne erklärende Variablen zurückzuführen (Kuhnt und Kalka 2022). Local Interpretable Model-Agnostic Explanations (LIME) schätzen dagegen lokale Zusammenhänge zwischen Variablen (Ribeiro et al. 2016). SHapley Additive exPlanations (SHAP) liefern neben dem globalen Einfluss von Variablen auch Erklärungen für die Ausprägung der Zielvariable bei jeder einzelnen Beobachtung, erlauben die Analyse von nicht-linearen Zusammenhängen und Interaktion und stellen einfach nutzbare Visualisierungswerkzeuge bereit (Lundberg und Lee 2017). Die dafür genutzten Shapley-Werte können auch für klassische statistische Inferenz und die Kommunikation der Ergebnisse in Form von Regressionstabellen genutzt werden (Joseph 2022). Darüber hinaus existiert eine Vielzahl weiterer Methoden. Solche Verfahren gewinnen seit wenigen Jahren an Bedeutung und Verbreitung, sodass bereits umfassende Literatur zu dem Thema existiert (z. B. Kamath und Liu 2021; Molnar 2022). Die Anwendung solcher Methoden setzt jedoch ein gutes Verständnis ihrer Limitationen voraus: Etwa sind sie nur in bestimmten Kontexten und auf gut generalisierbare Modelle anwendbar. Werden Abhängigkeiten und Interaktionen zwischen Variablen, Varianzmaße der Schätzwerte und die besonderen Anforderungen im hochdimensionalen Anwendungsfall nicht berücksichtigt, läuft man Gefahr, falsche Schlüsse zu ziehen (Molnar et al. 2022). Zudem können Methoden wie LIME und SHAP auch getäuscht werden, indem das Modell bewusst so verändert wird, dass es auf perturbierten Beobachtungen, die mit hoher Wahrscheinlichkeit nicht der Stichprobe entstammen, einen tatsächlich im Modell vorhandenen Bias korrigiert und so in seiner Gesamtheit unverzerrt wirkt (Slack et al. 2020). Die Auswahl und korrekte Anwendung von geeigneten Explainable AI Methoden ist also durchaus anspruchsvoll.

Auch im Austausch mit ***statistischen Fachbereichen*** ist Erklärbarkeit wichtig, um das entwickelte Modell zu kontrollieren und zu verbessern: So steigt das Vertrauen

in die Fähigkeit des Algorithmus, im Einsatz mit neuen Daten genaue Ergebnisse zu liefern, wenn er Muster erlernt, die fachstatistisch plausibel erscheinen und er beobachtete Zusammenhänge in der Realität approximativ wiedergibt.

Zuletzt ist es denkbar, dass auf Schätzungen basierende, statistische Ergebnisse aufgrund daran geknüpfter ***Rechtsfolgen*** extern gerechtfertigt werden müssen. Im Kontext der amtlichen Statistik ist dies sicherlich überwiegend unwahrscheinlich, sofern man den aktuellen Entwurf des europäischen AI-Acts zugrunde legt. Hier wird nämlich eine Unterteilung von KI-Systemen, in drei Kategorien vorgenommen: jene, „die ein i) unannehmbares Risiko, ii) ein hohes Risiko und iii) ein geringes oder minimales Risiko darstellen" (Europäische Kommission 2021, S. 15). Hochrisiko-KI-Systeme sind dabei solche, die Grundrechte gefährden und das Gebot der Nichtdiskriminierung missachten. Sie müssen „so konzipiert und entwickelt [werden], dass ihr Betrieb hinreichend transparent ist, damit die Nutzerinnen und Nutzer die Ergebnisse des Systems angemessen interpretieren und verwenden können" (ebd., S. 57) – also ein gewisses Maß an Erklärbarkeit aufweisen. Nach aktuellem Stand der europäischen Verordnungsentwürfe und der darauf basierenden juristischen Bewertungen könnte bei der Verwendung maschineller Lernmodelle etwa für die statistischen Register (wie das Statistische Unternehmensregister[27]), den Mikrozensus sowie die Steuerstatistiken eine Klassifizierung als Hochrisiko-KI erfolgen und folglich ein höheres Maß an Erklärbarkeit für Betroffene notwendig machen. Eine abschließende Bewertung – auch von juristischen Experten – wird letztlich erst nach der Verabschiedung der europäischen Verordnung möglich sein. Es ist beispielsweise möglich, dass im weiteren Prozess noch eine Ausnahmeregelung für die amtliche Statistik analog zum Forschungssektor getroffen wird oder bei der Verwendung öffentlicher Verwaltungsdaten geringere Erklärbarkeitsanforderungen gestellt werden.

3.3.1 Erklärbarkeit in der Praxis

Beim Einsatz in der amtlichen Statistik hängen die Anforderungen an die Erklärbarkeit von ML-Modellen wie beschrieben vom Anwendungsfall und der Zielgruppe ab. So stellt der Deutsche Ethikrat in seiner Stellungnahme zu KI fest: „Während es in manchen Bereichen sinnvoll oder sogar notwendig sein kann, ein Höchstmaß an Erklärbarkeit der jeweiligen Resultate anzustreben (Explainable AI), wofür gegebenenfalls ein hoher technischer und finanzieller Aufwand erforderlich ist, dürfte es in anderen Bereichen ausreichen sicherzustellen, dass die Personen, die diese Systeme anwenden, deren Resultate stets einer eigenen Plausibilitätsprüfung unterziehen, um den Gefahren eines ungerechtfertigten blinden Vertrauens in die Technik (Automation Bias) zu entgehen" (Deutscher Ethikrat 2023, S. 142). Gleichzeitig wird jedoch auch ein höherer Standard für die öffentliche Verwaltung empfohlen: „Aufgrund ihrer Grundrechtsbindung sind an staatliche Einrichtungen bei der Entwicklung und Nutzung algorithmischer Systeme hohe Anforderungen in Bezug auf Transparenz

[27] Weitere Informationen zum Unternehmensregister sind unter https://www.verwaltungsdaten-informationsplattform.de/SharedDocs/Register/Statistisches_Unternehmensregister.html zu finden. Stand: 09.02.2023.

Tab. 6 Indikatoren für die Evaluation der Erklärbarkeit maschineller Lernmodelle

Qualitätsindikatoren: Erklärbarkeit	
1	Es wurde bestimmt, auf welcher Ebene Erklärbarkeit gegeben sein soll – insbesondere, ob sich aus der Rechtslage für die gegebene Statistik zusätzliche Anforderungen ergeben und ob der Fachbereich zur Evaluation des Modells Zerlegbarkeit, Algorithmische Transparenz oder Interpretierbarkeit der Ergebnisse voraussetzt
2	Nach der Vorauswahl geeigneter ML-Verfahren wurden diese hinsichtlich ihrer Erklärbarkeit (mglw. in Verbindung mit Methoden des Interpretable ML) bewertet
3	Nur solche ML-Verfahren wurden eingesetzt, die die in 1. definierten Anforderungen hinsichtlich Erklärbarkeit erfüllen
4	Bei (annähernd) gleicher Ausprägung der anderen Qualitätsdimensionen wurde das erklärbarere Modell verwendet
5	Post-hoc-Analysen wurden – sofern erforderlich – bereitgestellt

und Nachvollziehbarkeit zu stellen, um den Schutz vor Diskriminierung zu gewährleisten sowie Begründungspflichten erfüllen zu können“ (Deutscher Ethikrat 2023, S. 42).

Zusammenfassend ist ein Mindestmaß an Erklärbarkeit notwendig, um die Entwicklung (seitens Data Scientists) und Akzeptanz (seitens der Fachbereiche) guter Modelle zu fördern, die sinnvolle Vorhersagen machen und deren Einsatz in der amtlichen Statistik entsprechend zweckmäßig ist. Darüber hinaus können sich im Einzelfall striktere Anforderungen für einzelne Statistiken aus der geltenden Rechtslage ergeben. Tab. 6 stellt dar, welche Schritte bei der Evaluation der Qualitätsdimension „Erklärbarkeit“ in der Praxis zwingend erfolgen sollten.

3.4 Reproduzierbarkeit

Die Reproduzierbarkeit von Forschungsergebnissen gehört zu den grundlegenden Qualitätskriterien in der Wissenschaft (Munafò et al. 2017; Peng 2011). Sie erhöht die Glaubwürdigkeit wissenschaftlicher Arbeiten, indem sie die unabhängige Überprüfung der Ergebnisse durch Dritte ermöglicht. Für die Prozesse und Produkte der Statistischen Ämter des Bundes und der Länder stellt die Reproduzierbarkeit von Ergebnissen ebenfalls eine wichtige Eigenschaft dar. Zum einen müssen die Produkte der amtlichen Statistik internationalen Standards und Leitlinien in vollem Umfang genügen (siehe 2.2 Qualitätsgrundsätze für statistische Produkte). Darüber hinaus müssen der Öffentlichkeit im Rahmen der Qualitätsberichte Informationen zu den verwendeten Methoden und Definitionen sowie zur Qualität der Produkte zur Verfügung gestellt werden. Reproduzierbarkeit schafft zudem Rechtssicherheit, da sie eine langfristige Nachvollziehbarkeit und erneute Durchführung von Prozessen ermöglicht.

Dies sichert das Vertrauen in die amtliche Statistik. Für die im Rahmen der Produkterstellung verwendeten Algorithmen gilt dies ebenso, d. h. diese müssen einen Grad der Nachvollziehbarkeit erreichen, der sich in der Rigidität an Kriterien der unabhängigen Wissenschaft orientiert (Radermacher 2022; Statistischer Beirat 2010).

Die Minimalanforderung an Reproduzierbarkeit ist, dass Ergebnisse mit derselben Datengrundlage und denselben Codes (bzw. der Beschreibung der Reihenfolge und Auswahl der Schritte bei Nutzung einer grafischen Oberfläche) wiederholt generierbar sein müssen. Darüber hinaus sind jedoch auch andere Definitionen möglich; hierbei kann zwischen drei Arten der Reproduzierbarkeit unterschieden werden (Yung et al. 2022):

a. Methodische Reproduzierbarkeit: die Fähigkeit, die experimentellen und rechnerischen Verfahren mit denselben Daten und Werkzeugen so genau wie möglich zu wiederholen und dieselben Ergebnisse zu erhalten.
b. Inferenzielle Reproduzierbarkeit: das Treffen von Wissensaussagen ähnlicher Stärke aus einer Studienreplikation oder Re-Analyse. Dies ist nicht identisch mit der Reproduzierbarkeit von Ergebnissen, da nicht alle Forscher aus denselben Ergebnissen dieselben Schlussfolgerungen ziehen.
c. Ergebnisbezogene Reproduzierbarkeit (Replizierbarkeit): die Erzielung übereinstimmender Ergebnisse in einer unabhängigen Studie (d. h. mit neuen Daten) unter Nutzung derselben Methoden.

Es stellt sich jedoch die Frage, ob in der praktischen Anwendung alle drei Dimensionen der Reproduzierbarkeit einzuhalten sind. Für die amtliche Statistik und ihre konkreten Anwendungsfälle ist die Betrachtung der methodischen Reproduzierbarkeit notwendig, um dem Qualitätsanspruch der amtlichen Statistik angemessen Rechnung zu tragen. Dies basiert darauf, dass amtliche Statistik aufgrund des gesetzlichen Auftrages zur Bereitstellung bestimmter Statistiken einerseits keine Beliebigkeit in der Fallauswahl hat und anderseits das bloße Beobachten von Tendenzen keine adäquate Entscheidungsgrundlage bilden kann.

Um diese sicherzustellen, müssen gewisse Anforderungen an Daten, Code und Umgebung erfüllt werden, die im Folgenden ausgeführt werden sollen. Dabei sind die genannten Kriterien als Mindeststandards zu verstehen, welche immer, also unabhängig vom Anwendungsfall, erfüllt sein müssen und von denen nur in sehr spezifischen, ausreichend geprüften Fällen abgewichen werden sollte.

3.4.1 Daten

Prinzipiell muss die Gesamtheit der genutzten Daten vorliegen und für den reproduzierenden Nutzenden zugänglich sein. Diese sind in der amtlichen Statistik die ausführenden Personen in den Ämtern, da konträr zur wissenschaftlichen Praxis keine externe Reproduktion möglich ist. Abweichend von der üblichen Praxis des Archivierens in der amtlichen Statistik findet man im wissenschaftlichen Betrieb zum Beispiel den Verweis auf Daten aus offiziellen Veröffentlichungen mit entsprechender Pfadangabe, aber auch DOIs (Digital Object Identifier) für Datenpublikationen. Für einige Datenpools (besonders Social-Media-Daten) ergeben sich hier Einschränkungen aufgrund rechtlicher Regelungen wie zum Beispiel AGBs (Kinder-Kurlanda et al. 2017).

Tab. 7 Anforderungen der Reproduzierbarkeit an die Daten

Konzept	Kriterium
Vollständigkeit	Alle in der zu reproduzierenden Analyse genutzten Daten müssen verfügbar sein
Adäquate Zugriffsmöglichkeiten	Der Zugang zu den Daten muss so einfach wie möglich, jedoch unter Berücksichtigung aller notwendigen rechtlichen Vorgaben gestaltet werden
Zeitliche Invarianz	Daten müssen unabhängig vom Zugriffszeitpunkt identisch sein
Definition des Umfangs und der Erfassung	Der Umfang der Daten muss exakt definiert und dokumentiert sein und es muss eine klare Beschreibung aller Informationen sowie der ursprünglichen Basisdaten vorliegen

Diese Problematik sollte zwar bei der Evaluierung neuer, experimenteller Datenquellen[28] mitgedacht werden, ist in der Praxis der amtlichen Statistik bislang aber von untergeordneter Bedeutung. Da die traditionellen Datenquellen der amtlichen Statistik primärstatistische Erhebungen und sekundäre Verwaltungs- und Registerdaten (Hartmann und Lengerer 2014; Signorelli et al. 2022) sind, stellt sich das Problem der Reproduzierbarkeit der Datengrundlage hier etwas anders dar. Insbesondere unterliegen die Roh- bzw. Einzeldaten und etwaige methodische Merkmale aufgrund ihrer besonderen rechtlichen Stellung besonderen Zugriffsbestimmungen (BStatG § 16 Geheimhaltung, Rothe 2015), sodass die Daten ausschließlich durch die Statistischen Ämter des Bundes und der Länder eingesehen und genutzt werden können. Dementsprechend ist eine Reproduzierbarkeit durch Externe nicht möglich; es besteht keine Möglichkeit, die Datengrundlage für Nutzende außerhalb der amtlichen Statistik zur Verfügung zu stellen. In der amtlichen Statistik liegt der Fokus stattdessen darauf, dass Veränderungen, die intern an der Datengrundlage vorgenommen werden, begründet und gut dokumentiert sein müssen. Die Reproduzierbarkeit ist immer dann gefährdet, wenn Fachbereiche im Rahmen von Plausibilitätsprüfungen Änderungen am Datensatz vornehmen und dies nicht dokumentieren, sodass die ursprünglichen, nicht-plausibilisierten Daten nicht mehr rekonstruiert werden können. Sollten, wie in der amtlichen Statistik üblich,[29] Revisionen vorgenommen werden, so sollte dabei folgendes beachtet werden, um die Reproduzierbarkeit beim Einsatz von maschinellen Lernverfahren zu gewährleisten: Erstens sollten Änderungen an den Datensätzen nur vorgenommen werden, wenn unbedingt notwendig. Zweitens sollten begründete Änderungen unbedingt dokumentiert werden. Drittens sollte eine Archivierung durchgeführt werden, sodass alte Versionen vor den Revisionen weiterhin zugänglich bleiben.

Aus den generellen Anforderungen an die Verwaltung großer Datenbestände und den spezifischen Strukturen der amtlichen Statistik ergeben sich spezifische An-

[28] Siehe hierzu auch einige Anwendungen auf den EXSTAT-Portalen des Statistischen Bundesamts [https://www.destatis.de/DE/Service/EXSTAT/_inhalt.html] und des Verbunds [https://www.statistikportal.de/de/experimentelle-statistiken].

[29] Siehe hierzu auch „Umgang mit Fehlern“ [https://www.statistikportal.de/de/umgang-mit-fehlern] und die „Allgemeine Revisionspolitik der Statistischen Ämter des Bundes und der Länder“ [https://www.destatis.de/DE/Methoden/Qualitaet/allgem-revisionspolitik.pdf] sowie „Revisionen“ [https://www.destatis.de/DE/Methoden/Revisionen/_inhalt.html].

forderungen in Bezug auf die Reproduzierbarkeit, die in Tab. 7 zusammengefasst sind.

3.4.2 Programmcode und Umgebung

Zusätzlich zu den Anforderungen an die Datengrundlage bestehen ebenso Anforderungen an den Programmcode, der das Modell definiert und seine Ausführung steuert, sowie an die technische Umgebung, in der er ausgeführt wird. Letztere ist hierbei besonders kritisch, da Software ständig aktualisiert wird und alter Code mit neuen Programmversionen nicht zwangsläufig ausführbar ist. In vielen Fällen kann auf Containerlösungen zurückgegriffen werden, die den Code zusammen mit der benötigten Software einfriert und für zukünftige Nutzung abspeichert.[30]

Die Nutzung von Open-Source-Software bietet zudem weitere Vorteile in Bezug auf die Reproduzierbarkeit: So können frei verfügbare Programme auch von intern Berechtigten ausgeführt werden, die keinen Zugriff auf proprietäre Software hätten. Dabei ist jedoch zu beachten, dass – ähnlich wie bei den Daten – eine externe Reproduktion nicht möglich ist. Lediglich auf interner Infrastruktur (wenn auch nicht notwendigerweise auf demselben Rechner) ist bei gleicher Ausstattung wie beim Primärdurchlauf ein Reproduktionsversuch durchführbar.

Zudem erlaubt Open-Source-Software direkte Einblicke in die innersten Strukturen der genutzten Funktionen und Objekte. Allerdings ergeben sich durch diese offene Struktur auch Einschränkungen: Änderungen an noch nicht ausgereiften Funktionen können häufiger vorkommen und die Wahl einer richtigen Paketversion kann eine relevante Voraussetzung für die korrekte Nutzung des Codes sein. Insofern besteht hier die verstärkte Notwendigkeit, auf Versionierung zu achten.

Hinsichtlich des Programmiercodes ist sicherzustellen, dass dieser auch für die Reproduktion verfügbar ist. Immer noch stellt dies bei vielen Veröffentlichungen nicht die Regel dar – mit großen Unterschieden zwischen den Fachdisziplinen (Baker 2016) – ist jedoch die Grundlage für reproduzierbare Analysen. Auch wenn die amtliche Statistik der in der Wissenschaft üblichen Veröffentlichungspraxis aus rechtlichen Gründen nicht nachkommen kann, ist zumindest die interne Archivierung des Codes notwendig für Reproduzierbarkeit in der Zukunft. Wenn Algorithmen mit zufälligen Komponenten verwendet werden, ist der genutzte Seed im Code explizit festzulegen oder zu speichern, um bei der Reproduktion identische Ergebnisse zu ermöglichen. Weiterhin ist beim Programmieren auf eine ausreichende und verständliche Dokumentation der Schritte zu achten – sei es durch Kommentare im Code oder durch adäquate Begleitdokumente – um die Codestruktur nachvollziehen zu können und etwaige Probleme lösen zu können. Hierbei ist die Anwendung von (internen) Styleguides zu empfehlen. Zusammen mit einer angemessenen Dokumentation kann die Orientierung an z. B. Konventionen bei der Namensgebung von

[30] Hierbei ist jedoch zu beachten, dass auch Containerlösungen nach mehreren Jahren aufgrund technischer Weiterentwicklungen oder sicherheitstechnischer Einschränkungen potenziell nicht mehr verwendet werden können. Selbst wenn die technischen Voraussetzungen weiterhin gegeben sind, so ist es möglich, dass z. B. ausschließlich andere Programmiersprachen genutzt werden oder Kenntnisse verloren gehen. Hier ist es Aufgabe der Ämter, eine möglichst langfristige Sicherheit zu schaffen.

Tab. 8 Anforderungen der Reproduzierbarkeit an Code und Umgebung

Konzept	Kriterium
Vollständigkeit	Der gesamte genutzte Code muss verfügbar sein; genutzte Seeds müssen explizit gesetzt oder gespeichert werden
Adäquate Zugriffsmöglichkeiten	Der Zugang zu den Codes muss so einfach wie möglich, jedoch unter Berücksichtigung aller notwendigen Sicherheitsstufen gestaltet werden
Kommentierung, Begleitdokumentation	Eine ausreichende Kommentierung der Codes muss vorhanden sein
Versionierung	Versionierungen der Codes und deren Dokumentation müssen vorliegen
Spezifische Eigenheiten von Implementierungen	Sollten spezifische Funktionen/Klassen besondere Eigenheiten ausweisen, ist dies ebenfalls zu dokumentieren oder zu referenzieren
Gleiche Software	Es muss ein Zugang zur gleichen Software mit der gleichen Version wie in der zu reproduzierenden Auswertung bestehen
Computing Infrastruktur	Es muss Zugang zu einer gleichwertigen (optimalerweise der identischen) Infrastruktur für die Berechnungen bestehen

Objekten, Funktionen und Verzeichnissen dabei helfen, eine effiziente Überprüfung der einzelnen Schritte zu ermöglichen.

Tab. 8 zeigt eine Zusammenfassung der genannten Anforderungen an Code und Umgebung.

3.4.3 Besonderheiten maschineller Lernverfahren

Bei der Anwendung von maschinellen Lernverfahren ist darüber hinaus eine klare Beschreibung des Algorithmus bereitzustellen, die auch die genutzten Hyperparameter und Gründe für die Auswahl bestimmter Werte beinhaltet. Zudem sollte klar beschrieben werden, wie Trainings-, Test- und Validierungsdatensätze generiert wurden und wie das Tuning durchgeführt wurde. Bei der Modellauswahl an sich ist zu beachten, dass komplexere Methoden schwieriger erfolgreich reproduziert werden können (Hu et al. 2021; Sani et al. 2018). Werden unübliche, experimentelle Methoden eingesetzt, deren Interpretation nicht offensichtlich ist, so sollten Ergebnisse ausführlich erklärt werden.

Werden die oben ausgeführten Standards bei der Entwicklung von maschinellen Lernmodellen beachtet, so können sie in der amtlichen Statistik zu einer verbesserten Reproduzierbarkeit beitragen: Zum einen kann ihr Einsatz vormals manuelle Klassifikationsaufgaben stärker strukturieren, indem Kategorisierungsregeln expliziert gemacht werden. So produziert ein ML-Algorithmus bei gleichem Input immer den exakt gleichen Output. Unabhängig von der Genauigkeit des Modells können im Vergleich zur manuellen Kodierung stabilere Fehler erwartet werden, was eine Untersuchung der Ursachen vereinfachen kann: Wenn das ML-Verfahren falsch liegt, dann liegt es zumindest „konsistent falsch“. Gleichzeitig wird dadurch auch deutlich, welch große Bedeutung qualitativ hochwertige Trainingsdaten innehaben: Sonst werden inkorrekte Muster mitgelernt und zukünftig auf alle Vorhersagen angewandt.

Tab. 9 Indikatoren für die Evaluation der Reproduzierbarkeit maschineller Lernmodelle

Qualitätsindikatoren: Reproduzierbarkeit	
1	Alle in der zu reproduzierenden Analyse genutzten Daten wurden wiederauffindbar abgelegt
2	Der zukünftige Zugang zu den Daten wurde geklärt und dokumentiert sowie über entsprechende Berechtigungen (unter Berücksichtigung aller notwendigen rechtlichen Vorgaben) ermöglicht
3	Es wurde sichergestellt, dass die Daten unabhängig vom Zugriffszeitpunkt identisch vorliegen, d. h. unverändert an einem bekannten Ort archiviert wurden
4	Der Umfang der Daten wurde exakt definiert und dokumentiert, z. B. durch ein Codebook; eine Beschreibung der Informationen und Variablen in den ursprünglichen Basisdaten liegt vor; potenziell zusätzlich nötige Informationen wurden ebenfalls verschriftlicht
5	Alle in der zu reproduzierenden Analyse genutzten Skripte wurden wiederauffindbar abgelegt
6	Der zukünftige Zugang zu den Skripten wurde geklärt und dokumentiert sowie über entsprechende Berechtigungen (unter Berücksichtigung aller notwendigen Sicherheitsstufen) ermöglicht
7	Alle Codeblöcke wurden ausreichend kommentiert, um die Schritte in Bezug auf Input, Output und Modifizierung durch den Code nachvollziehen zu können
8	Informationen zu den genutzten Paketen, Modulen oder Bibliotheken wurden hinterlegt; Versionsentscheidungen – falls von der neuesten abweichend – wurden begründet
9	Sollten spezifische Funktionen oder Klassen besondere Eigenheiten ausweisen, wurde dies in ausreichendem Maße schriftlich dokumentiert
10	Die genutzte Software wurde schriftlich dokumentiert und der zukünftige Zugriff wurde geregelt

3.4.4 Reproduzierbarkeit in der Praxis

Zusammengefasst lassen sich aus den oben herausgearbeiteten Konzepten und Kriterien eine Reihe an Indikatoren ableiten, die in Tab. 9 dargestellt sind. Dabei fällt auf, dass bereits das Nichterfüllen eines einzigen Aspektes zum Scheitern einer Reproduktion führen kann. Fehlen etwa die Daten, ist eine Reproduktion sofort ausgeschlossen. Insofern ist die folgende Liste als holistischer Ansatz zu verstehen, bei dem keine selektive Vorauswahl getroffen werden kann.

3.5 Aktualität und Pünktlichkeit

Aktualität und Pünktlichkeit gehören zu den im Verhaltenskodex benannten Anforderungen an die Qualität statistischer Produkte. Grundsätzlich fordert diese Qualitätsdimension die pünktliche Veröffentlichung statistischer Produkte sowie die Berücksichtigung von Nutzerbedarfen hinsichtlich der Aktualität der dargestellten Ergebnisse in den Veröffentlichungsprodukten.

In Bezug auf den Einsatz maschineller Lernverfahren sind grundsätzlich die identischen Indikatoren und Anforderungen gültig, die der Verhaltenskodex generell an den Qualitätsgrundsatz „Aktualität und Pünktlichkeit“ stellt (vgl. 2.2 Qualitätsgrundsätze für statistische Produkte). Hiernach gelten Statistiken als pünktlich veröffentlicht, wenn die im Veröffentlichungskalender definierten Publikationszeitpunkte eingehalten werden. Zur Messung der Aktualität einer Veröffentlichung wird zumeist die Zeit in Tagen zwischen der Datenlieferung bzw. -erhebung und der Veröffentlichung der Ergebnisse betrachtet.

Mit dem Einsatz von maschinellen Lernverfahren ist grundsätzlich auch die Hoffnung verbunden, bestehende Prozesse z. B. mit Blick auf eine rechtzeitige Veröffent-

lichung oder eine höhere Aktualität von statistischen Produkten zu verbessern bzw. effizienter zu gestalten. Aus diesem Grund sind einige Aspekte beim Einsatz maschineller Lernverfahren besonders zu berücksichtigen.

3.5.1 Pünktlichkeit

Hinsichtlich der Sicherstellung rechtzeitiger Veröffentlichungen gelten in der Praxis für Verfahren des maschinellen Lernens und konventionelle Methoden der Statistikproduktion ähnliche Heuristiken. Jedoch ist vor allem die Konzeptionsphase für den Einsatz maschineller Lernverfahren oftmals mit zeitlichen Aufwänden verbunden, die bei der Terminierung des Endprodukts berücksichtigt werden müssen. Beim Einsatz von ML-Algorithmen muss insbesondere die häufig langwierige Sammlung und Aufbereitung von Trainings-, Validierungs- und Testdaten eingeplant werden, wobei in der Regel viele Stellen zu involvieren sind. Zumeist muss auch die Fachseite unbearbeitete Rohdaten oder Informationen aus dem Aufbereitungsprozess liefern. Zudem muss für eventuelle Verzögerungen bei der Datenbeschaffung ein ausreichender Puffer sowie für die Konzeption und das Training der Algorithmen ausreichend Zeit eingeplant werden. Eine Fertigstellung und Lieferung von Endprodukten ist demnach mit deutlichem Vorlauf zu planen.

Eine Messung der Pünktlichkeit kann wie bei konventionellen statistischen Verfahren durch einen Abgleich der geplanten und tatsächlichen Veröffentlichungs- bzw. Lieferzeitpunkte des Endprodukts erfolgen.

3.5.2 Aktualität

Hinsichtlich der Aktualität sind insgesamt nur wenige Änderungen und Besonderheiten von ML-Anwendungen gegenüber normalen Verfahren herauszustellen. Im Produktivbetrieb sollte durch den Einsatz von maschinellen Lernverfahren (bei gleichbleibender Genauigkeit) grundsätzlich eine Verbesserung der Aktualität zu beobachten sein, wenn langwierige oder aufwändige manuelle Prozesse durch den Einsatz von maschinellen Lernverfahren ersetzt bzw. beschleunigt werden können. Dies gilt insbesondere für Anwendungsfälle, die ohne den Einsatz maschineller Lernverfahren ein hohes Maß an Personalressourcen erfordern. Ein Beispiel stellt die automatische Generierung von Wirtschaftszweig-Schlüsseln aus Freitexten dar, deren Klassifikation ohne maschinelle Lernverfahren mit viel Aufwand manuell durchgeführt werden müsste. Auch kann der Einsatz eines ML-Verfahrens dazu beitragen, dass vorläufige Ergebnisse von akzeptabler Gesamtgenauigkeit früher als bisher veröffentlicht werden können.

Mögliche Verbesserungen der Aktualität können durch einen Vergleich der benötigten Zeit von Datenerhebung bzw. Datenlieferung bis zur Veröffentlichung vor und nach der Implementation eines ML-Verfahrens messbar gemacht werden.

3.5.3 Aktualität und Pünktlichkeit in der Praxis

Tab. 10 fasst zusammen, welche Schritte bei der Nutzung maschineller Lernverfahren zur Sicherstellung und Messung der Aktualität und Pünktlichkeit notwendig sind.

Tab. 10 Indikatoren für die Evaluation der Aktualität und Pünktlichkeit maschineller Lernmodelle

Pünktlichkeit	
1	Die Fachseite wurde rechtzeitig eingebunden
2	Es wurde hinreichend Zeit für die Datenbeschaffung und Datenaufbereitung eingeplant
3	Es wurde hinreichend Zeit für die Konzeption sowie die Auswahl und die Testung möglicher ML-Verfahren eingeplant
Aktualität	
1	Die Durchlaufzeit bei Nutzung des ML-Verfahrens im Produktivbetrieb wurde gemessen

3.6 Wirtschaftlichkeit

Der Verhaltenskodex führt zum Grundsatz der Wirtschaftlichkeit unter anderem aus, dass zur Gewährleistung der Wirtschaftlichkeit das Produktivitätspotenzial von Informations- und Kommunikationstechnologie (IKT) soweit wie möglich ausgeschöpft werden soll und dass Effizienz und Wirksamkeit über standardisierte Lösungen realisiert werden sollen. Die Verwendung von maschinellen Lernverfahren in der Statistikproduktion zahlt also insoweit auf die Wirtschaftlichkeit ein, als technologische Entwicklungen zur Verbesserung der Produktivität genutzt werden und der langfristige Aufwand durch die Entwicklung standardisierter Prozesse gesenkt wird. Im „Quality Framework for Statistical Algorithms“ (Yung et al. 2022) wird Wirtschaftlichkeit im Sinne einer Kosteneffizienz aufgefasst. In der Praxis ist der „effektive Einsatz von Ressourcen“ im Sinne der Wirksamkeit zu verstehen, also als Verhältnis der Ausprägung der anderen Qualitätsdimensionen zu den Kosten der Umsetzung. Als Maß wird das Verhältnis von der Ergebnisqualität zu den angefallenen Kosten vorgeschlagen.

Der Nutzen von Methoden des maschinellen Lernens ist abhängig vom Einsatzgebiet. ML-Algorithmen können etwa bei gleichbleibender Ergebnisqualität manuelle Personalaufwände reduzieren. Aber auch in anderen Bereichen sind Verbesserungen denkbar: So sind standardisierte Methoden üblicherweise weniger anfällig gegenüber Flüchtigkeitsfehlern und erlauben in bestimmten Fällen eine deutlich schnellere Statistikproduktion. Mittels maschineller Lernverfahren lassen sich gegebenenfalls auch Auswertungen realisieren, die mit aktuellen Prozessen nicht umsetzbar sind – etwa eine umfassende, hochfrequente Plausibilisierung und damit eine monatliche Veröffentlichung einer bis dato nur halbjährlich produzierten Statistik. In der Praxis ist jedoch zu beachten, dass die Erwartungen an eine ML-Implementierung klar definiert werden und realistisch sein müssen. So ist in den wenigsten Fällen eine gleichzeitige Verbesserung der Genauigkeit und Schnelligkeit bei gleichzeitigem reduzierten Ressourceneinsatz möglich. Um im Nachhinein negative Überraschungen zu vermeiden, wird deshalb empfohlen, vorab die Nutzungserwartungen systematisch z. B. über eine Anforderungsanalyse zu erfassen. Auf dieser Basis lässt sich dann abschätzen, ob die Anforderungen methodisch, technisch und wirtschaftlich mit maschinellen Lernmethoden erfüllt werden können. Anschließend sollten möglichst alle Anforderungen in messbare Nutzenindikatoren überführt werden, die im Nachhinein evaluiert werden können.

Beim Aufwand der Einführung von Verfahren des maschinellen Lernens ist zwischen einmaligen und laufenden Aufwänden zu unterscheiden. Nicht immer ist diese

Tab. 11 Aufwände bei der Nutzung von maschinellen Lernverfahren

Einmalige Aufwände für Implementierung, Umstellung oder Neukonzeptionierung
Entwicklung von ML-Verfahren, Konzepten, Modellen und Algorithmen
Erstellung oder Beschaffung von Trainings-, Validierungs- und Testdaten
Anschaffungskosten für Hard- und Software sowie Kosten für externe Beratung
Einstellung und Bindung von qualifizierten Fachkräften (z. B. Data Scientists) und Qualifizierung von Beschäftigten
Überführung in standardisierte Verfahren und Entwicklung von Standardwerkzeugen
Fortlaufende Aufwände bei der Nutzung von ML-Verfahren in der Statistikerstellung
Aufwände für die Bereinigung und Vorbereitung von Trainings-, Validierungs- und Testdaten sowie für das Labeln von Daten bei Klassifikationen
Studium und Prüfung von (neu entwickelten) Technologien und Verfahren im Bereich ML inkl. Wissensmanagement
Wartung, Pflege und Aktualisierung von Verfahren, Modellen und Algorithmen

Unterscheidung trennscharf – dennoch ist es in der Praxis wichtig, sich bewusst zu sein, welche Aufwände bei der Nutzung von maschinellen Lernverfahren regelmäßig auftreten und welche eher einmalig bei der Implementierung, Umstellung oder Neukonzeptionierung entstehen. Einige solcher Aufwände werden in Tab. 11 aufgelistet.

Die Aufwands- bzw. Kostentreiber sollten in quantifizierbare Aufwandsindikatoren überführt werden. Auf Basis der Nutzen- und Aufwandsindikatoren wird es möglich, quantitative Ziele abzuleiten sowie im Rahmen einer Wirtschaftlichkeitsprognose (etwa nach Kazmierski und Ritsert 2010) systematisch feststellen, wie groß der erwartete Aufwand (zeitlich, personell, prozessual) im Vergleich zum erwarteten Nutzen sein wird. Insgesamt wird dabei aber grundsätzlich auf die Verfahren zur Gewährleistung der Wirtschaftlichkeit zurückgegriffen, die im Statistischen Verbund bereits verwendet werden.[31]

3.6.1 Wirtschaftlichkeit in der Praxis

In der Praxis besteht bei der Arbeit mit ML-Algorithmen häufig die Herausforderung, dass der potenzielle Nutzen mit den zu Projektbeginn zur Verfügung stehenden Informationen schwer abschätzbar ist. Darüber hinaus ist der Umsetzungsaufwand stark abhängig vom Ausgangszustand, insbesondere der Qualität der vorhandenen Datenbasis (Trainings-, Validierungs- und Testdaten). Empfohlen wird deshalb ein zweigeteiltes Vorgehen: In einem ersten Schritt sollte ein „Proof of Concept“ (PoC) durchgeführt werden. Wenn sich dadurch herausstellt, dass die Anforderungen und Wünsche mit ML-Methoden nicht oder nur mit relativ großem Aufwand erfüllt werden können, kann oder sollte von einer weiteren Bearbeitung abgesehen werden. Sind die Ergebnisse des PoC in Verbindung mit der Wirtschaftlichkeitsprognose

[31] Um einen effizienten Ressourceneinsatz zu erreichen, werden in den Statistischen Ämtern des Bundes und der Länder sowohl der personelle und als auch der finanzielle Ressourceneinsatz unter Zuhilfenahme der jeweils eingesetzten Rechnungslegungsinstrumente (Haushaltsbewirtschaftungssysteme, Personal- und Stellenbewirtschaftungssysteme, Kosten- und Leistungsrechnungen usw.) überwacht.

Tab. 12 Indikatoren für die Evaluation der Wirtschaftlichkeit maschineller Lernmodelle

Qualitätsindikatoren: Wirtschaftlichkeit	
1	Eine Anforderungsanalyse wurde zu Beginn des Projekts durchgeführt. Ziele und Erwartungen an das ML-Verfahren wurden systematisch erfasst
2	Im Rahmen eines Proof of Concept wurde die vorhandene Datenbasis geprüft und die Machbarkeit des Vorhabens eingeschätzt
3	Eine Wirtschaftlichkeitsprognose wurde durchgeführt. Erwartete Aufwände (zeitlich, personell, prozessual; siehe Tab. 11) wurden ermittelt und mit dem erwarteten Nutzen verglichen[a]

[a] Ist das Vorhaben etwa, die Plausibilisierung von Erhebungsdaten mittels ML zu beschleunigen, so kann nach einer Machbarkeitsstudie konkret folgende Kennzahl berechnet werden, um die Wirtschaftlichkeit des Vorhabens quantitativ zu bestimmen: Veränderungen der Durchlaufzeit ΔDLZ relativ zu den Gesamtkosten der ML-Umsetzung, wobei $\Delta DLZ = DLZ_{\text{mitML}} - DLZ_{\text{ohneML}}$

hingegen vielversprechend, kann im Weiteren mit einer vertieften Bearbeitung oder direkt mit der Implementierung begonnen werden. Durch ein solches Vorgehen wird außerdem sichergestellt, dass zum Zeitpunkt der Aufwandsschätzung der Implementierung bereits ein gutes Verständnis der Chancen und Herausforderungen gegeben ist. Tab. 12 fasst die beschriebene Vorgehensweise zusammen.

3.7 Ergänzende Aspekte

Abschließend seien zwei übergreifende Themen erwähnt, die zwar keine eigenständigen Qualitätsdimensionen darstellen, aber in engem Bezug zu den sechs genannten Dimensionen stehen: Zum einen kann im Rahmen der Evaluierung der Genauigkeit auch untersucht werden, ob ein ML-Modell „faire“ Ergebnisse für interessierende Subgruppen liefert; auch Erklärbarkeit erlaubt die Analyse und Bewertung der Modellstrukturen unter Fairness-Gesichtspunkten. Zum anderen sind für die zeit- und kosteneffiziente Umsetzung der Qualitätsanforderungen (insbesondere der Reproduzierbarkeit) in der Praxis bestimmte technische Voraussetzungen nötig, die unter dem Begriff MLOps zusammengefasst werden.

3.7.1 Fairness

Die Fairness von statistischen Verfahren bezieht sich auf die Auswirkungen, die algorithmische Entscheidungen oder Klassifikationen auf befragte Individuen oder administrative Einheiten haben können. Im Sinne der Gerechtigkeit ist „bei der Beurteilung von KI zu berücksichtigen, für wen eine Anwendung jeweils Chancen oder Risiken [...] mit sich bringt“ (Deutscher Ethikrat 2023, S. 22) – es sollte also vermieden werden, dass bestimmte Gruppen infolge der Anwendung maschineller Lernverfahren in relevanter Weise anders behandelt werden. Im Kontext der amtlichen Statistik sind solche Auswirkungen i. d. R. mittelbar, also etwa durch politische Entscheidungen auf Grundlage der veröffentlichten Daten. Neben allgemeinen Erwägungen wie der Qualität der Trainingsdaten – ein maschinelles Lernmodell kann unabhängig von der Schätzgenauigkeit falsche Zusammenhänge erlernen, wenn bereits die Trainingsdaten strukturelle Verzerrungen enthalten (Mehrabi et al. 2022; Schwemmer et al. 2020) – kann die ***Genauigkeit*** eines ML-Verfahrens Implikationen für die Fairness haben, wenn statistische Aggregate für bestimmte Subgruppen

systematisch über- oder unterschätzt werden. Dies ist insbesondere relevant für Subgruppen, die in den Daten weniger stark vertreten sind oder tendenziell am Rande der Verteilung liegen: André und Meslin vom französischen nationalen Statistikinstitut entwickeln beispielsweise ein maschinelles Lernmodell zur Schätzung des Immobilienvermögens von Haushalten auf Basis eines umfangreichen Datensatzes französischer Immobilien (André und Meslin 2021) und stellen fest, dass konventionelle ML-Algorithmen den Marktwert von Luxusobjekten – die zwar nur einen geringen Anteil der Beobachtungen ausmachen, aber für die Schätzung der Vermögensverteilung von hohem Interesse sind – systematisch unterschätzen. Um dieses Problem zu umgehen, wird ein Hilfsalgorithmus zur Klassifizierung von Luxusimmobilien verwendet, deren Marktwert dann mit einem angepassten Schätzmodell geschätzt wird, welches im Training wertvolle Objekte deutlich stärker gewichtet.

Allgemein gibt es nach Branco et al. (2017) verschiedene Strategien, um die Modellgenauigkeit für kleine Subgruppen zu erhöhen und so Verzerrungen zu vermeiden: Up- und Downsampling-Methoden oder hybride Formen wie SMOTE (Synthetic Minority Over-sampling Technique, Chawla et al. 2002) und ROSE (Random Over-Sampling Examples, Menardi und Torelli 2014; Lunardon et al. 2014) können währen des Preprocessings genutzt werden, um Datensätze mit einer ausgewogeneren Verteilung zu erhalten. Active Learning verbessert das Downsampling, indem mittels Expertenfeedback besonders informative Beobachtungen ausgewählt werden. Alternativ zu Resampling-Verfahren können (wie im Beispiel oben) Gewichte vergeben werden, die die Kosten einer Fehlklassifikation für bestimmte Subgruppen erhöhen. Auch existieren mittlerweile angepasste ML-Modelle, die eigens für die Verwendung mit unbalancierten Daten optimiert sind.

In der Praxis muss der Einsatz der genannten Methoden nicht zwingend notwendig sein, wenn auch einfache Modelle eine zufriedenstellende Schätzgenauigkeit für Subgruppen liefern. Dies kann evaluiert werden, indem bei dem trainierten Modell im Rahmen der Qualitätssicherung berechnet wird, ob einfache Aggregate für interessierende Subgruppen[32] in relevantem Ausmaß unter- oder überschätzt werden.

Auch die Qualitätsdimension der ***Erklärbarkeit*** hat Berührpunkte mit der Fairness: Selbst wenn Aggregate für verschiedene Subgruppen korrekt geschätzt werden, so ermöglicht ein Verständnis des Effekts der Subgruppenausprägungen auf die Zielvariable das Ziehen wertvoller Schlussfolgerungen über die Funktionsweise und die Verlässlichkeit des Modells. Zu Zwecken der Prädiktion ist es a priori nicht ausgeschlossen, Muster in den Daten zu verwenden, die auf bestehende Ungleichheiten und Stereotypen hindeuten, wenn sie de facto eine große Vorhersagekraft haben – schließlich ist das Ziel der amtlichen Statistik, bestehende Strukturen und Gegebenheit abseits politischer Vorstellungen möglichst wahrheitsgetreu abzubilden.[33] Nur auf einer soliden Datengrundlage können Ist-Zustände objektiv evaluiert und Strate-

[32] Subgruppen können etwa aufgrund ihrer politischen Bedeutung oder aufgrund ihrer Wichtigkeit in den Daten relevant sein. Letztere können algorithmisch identifiziert werden; die Entwicklung von Standardmethoden ist aktueller Forschungsgegenstand, siehe z. B. Bothmann et al. (2022).

[33] Siehe exemplarisch für die Bundesstatistik § 1 Satz 2 des Gesetzes über die Statistik für Bundeszwecke (Bundesstatistikgesetz – BStatG): „Für [die Bundesstatistik] gelten die Grundsätze der Neutralität, Objektivität und fachlichen Unabhängigkeit.“.

gien zum Erreichen von Soll-Zuständen formuliert werden. Jedoch dürfen Effekte, die ohne Anspruch auf eine wissenschaftlich korrekte Modellierung realer Beziehungen geschätzt wurden, nicht ohne Weiteres als kausale Zusammenhänge interpretiert werden.[34] In jedem Fall erlaubt ein interpretierbares Modell erst, zu evaluieren, ob ein bestimmter erlernter Zusammenhang im Einzelfall das Vertrauen in die Fähigkeit des Modells, mit ungesehenen Daten umzugehen, infrage stellt.

3.7.2 MLOps

Zur bestmöglichen Erfüllung der Qualitätsdimensionen ist es unabdingbar, dass bestimmte standardisierte Prozesse der Datenverarbeitung und des Datenmanagements etabliert werden: So würde die Evaluation der ***Genauigkeit*** durch Werkzeuge erleichtert werden, die beim Modelltraining standardmäßig relevante Gütemaße und ihre Varianz ausgeben. Die Sicherstellung der ***Robustheit*** und ***Erklärbarkeit*** kann im Einzelfall enorm aufwendig sein, wenn nicht bestimmte Test- und Auswertungsroutinen bereits programmatisch definiert und einfach ausführbar sind. Auch ***Aktualität*** und ***Wirtschaftlichkeit*** würden durch Standardprozesse begünstigt werden, so sie denn einmal entwickelt wurden. Die Qualitätsdimension, für die standardisierte Prozesse am relevantesten sind, ist ohne Frage die ***Reproduzierbarkeit***: Um zu erreichen, dass Daten, Codes und Umgebungen reproduzierbar sind, müssen bestimmte Vorgehensweisen und technische Werkzeuge einheitlich etabliert und genutzt werden. Praktiken und Prozesse, die darauf abzielen, Modelle für maschinelles Lernen zuverlässig und effizient zu entwickeln, produktiv bereitzustellen, zu verwalten, zu überwachen und zu warten, werden unter dem Schlagwort „Machine Learning Operations“ (MLOps) zusammengefasst (Shankar et al. 2022; Kreuzberger et al. 2022). Insofern solche Prozesse, Instrumente und Praktiken notwendig sind, um die in den Qualitätsrahmenwerken der amtlichen Statistik vorgegebenen Qualitätsdimensionen zu erfüllen, ist MLOps ebenfalls eine Grundvoraussetzung für maschinelles Lernen in der amtlichen Statistik. Folglich ist es essenziell, dass Aspekte des MLOps bei der Entwicklung von IT-Plattformen und Datenmanagementsystemen berücksichtigt und von Anfang an mitgedacht werden. Hierzu wurden auf internationaler Ebene im Rahmen des ONS-UNECE-ML-2022-Projekts bereits Anforderungen an MLOps-Systeme formuliert und mögliche Systemarchitekturen diskutiert (Engdahl et al. 2022).

4 Fazit

Um maschinelle Lernverfahren langfristig in die amtliche Statistik zu integrieren, bedarf es klarer Qualitätskriterien, die dem Anspruch amtlicher Statistik genügen und Vertrauen sichern. Dieser Aufsatz konnte durch eine systematische Analyse der Grundsätze und Indikatoren des Verhaltenskodex nachweisen, dass das Qualitätsverständnis der amtlichen Statistik breit genug ist, um auch auf neue Methoden wie maschinelles Lernen anwendbar zu sein. So lassen sich alle für den Einsatz von ML

[34] Es existieren jedoch auch Ansätze zur kausalen Interpretation von Zusammenhängen, die mittels ML-Modellen festgestellt wurden, siehe etwa Wager und Athey (2018) oder Freiesleben et al. (2022).

relevanten Aspekte unter die bestehenden Qualitätsgrundsätze subsummieren. Auf Grundlage bestehender Literatur wurden sechs Qualitätsdimensionen für den Einsatz von maschinellen Lernverfahren in der amtlichen Statistik aus dem europäischen Verhaltenskodex und den Qualitätsanforderungen an statistische Algorithmen abgeleitet: Genauigkeit, Robustheit, Erklärbarkeit, Reproduzierbarkeit, Aktualität und Pünktlichkeit sowie Wirtschaftlichkeit. Daraus wurden Anforderungen an ML-Algorithmen definiert und konkrete Prozessindikatoren zur Sicherstellung der Qualität in der Praxis abgeleitet.

Dabei wurde deutlich, dass maschinelles Lernen in vielerlei Hinsicht zur Verbesserung der Qualität amtlicher Statistikprodukte beitragen kann. Auch wurde argumentiert, dass zur Implementierung einer qualitätsgesicherten ML-Entwicklungsarbeit standardisierte Prozesse der Datenverarbeitung und des Datenmanagements etabliert werden müssen, die als MLOps bezeichnet werden. Abseits technischer Überlegungen hat maschinelles Lernen potenzielle Auswirkungen auf die Fairness statistischer Produkte, wenn Resultate für Subgruppen systematisch über- oder unterschätzt werden; dies ist in der Praxis im Rahmen der Genauigkeits- und Erklärbarkeitsuntersuchungen zu berücksichtigen. Zuletzt zeigte sich, dass qualitativ hochwertige ML-Algorithmen auf menschliche Unterstützung und Überwachung angewiesen sind: Werden Anpassungen an einer Statistik vorgenommen, indem etwa neue Merkmale erhoben werden, so ist i. d. R. eine händische Anpassung des Modells durch ML-Entwickler notwendig. Fälle, die der Algorithmus nicht mit ausreichender Genauigkeit vorhersagen oder klassifizieren kann, müssen weiterhin von der Fachstatistik überprüft werden. Viele potenzielle Fehlerquellen können durch stichprobenartige Kontrollen des Algorithmus und seiner Ergebnisse erkannt werden – etwa um Concept Drift zu erkennen und zu entscheiden, wann ein Neutrainieren erforderlich ist. Um Ad-hoc Entscheidungen zu minimieren, sollten Kriterien für menschlichen Review möglichst schon während der Entwicklung festgelegt werden.

Werden maschinelle Lernverfahren in der Praxis eingesetzt, so können fast nie alle Qualitätsdimensionen perfekt erfüllt werden. In Konfliktfällen ist es möglich und notwendig, verschiedene Dimensionen gegeneinander abzuwägen: So kann eine geringere Erklärbarkeit bei bestimmten Anwendungsfällen möglicherweise in Kauf genommen werden, wenn dadurch eine deutlich höhere Genauigkeit erreicht werden kann. Ebenfalls kann bei einem Modell, das nur zur Deckung eines kurzfristigen, einmalig auftretenden Bedarfs ausgeführt werden soll (etwa vor der Einführung einer neuen Software), eine geringere Robustheit zugunsten einer schnellen Implementierung toleriert werden. Dies ist dann jeweils eine fachstatistische oder amtspolitische Entscheidung. Kurz gesagt: „Ein Verständnis [der Notwendigkeit] der Abwägung der Qualitätskriterien und das Anstellen fundierter Überlegungen hierzu sind notwendig, um ein gutes Gesamtergebnis zu erreichen. So sieht professionelle statistische Arbeit in der Praxis aus“[35] (Sæbø und Holmberg 2019, S. 177).

Es verbleiben einige offene methodische Fragen, die aktueller Forschungsgegenstand in der Wissenschaft sind und daher zum derzeitigen Stand nicht abschließend beantwortet werden können. Wie kann der Generalisierungsfehler von ML-Verfah-

[35] Originalwortlaut: „Comprehension of and reflection on balancing quality criteria is necessary to achieve a good total result. This is statistical professionalism in practice.“

ren bei komplexen Stichprobendesigns geschätzt werden (3.1.2 Stichprobenbedingte Unsicherheit)? Wie kann die Qualität der Trainingsdaten quantifiziert werden (3.1.3 Qualität der Trainingsdaten)? Mittels welcher Gütemaße kann die Robustheit eines Modells gegen Datenperturbationen gemessen werden (3.2.1 Datenperturbation)? Wie können relevante Subgruppen algorithmisch und datengetrieben identifiziert werden (3.7.1 Fairness)? Weiterhin gibt es für viele Problemstellungen zwar Lösungsvorschläge, aber noch keine Best-Practice Methoden, die effizient implementiert in ausgereiften Softwarepaketen bereitstehen. Etwa mangelt es bis dato an einer optimalen Methodik für das Neutrainieren von Modellen bei Concept Drift (3.2.3 Modellperturbation: Concept Drift), Standardprozeduren für die Evaluation und Sicherstellung der Robustheit (3.2.6 Robustheit in der Praxis), Standardmethoden zur Verbesserung der Interpretierbarkeit (3.3 Erklärbarkeit) oder „robustifizierten" ML-Schätzern und Hypothesentests (3.2.5 Auswahl robuster Verfahren). Dadurch zeigte sich, wie wichtig die kontinuierliche Aus- und Weiterbildung zu dem aktiven und sich rasant entwickelnden Forschungsfeld des maschinellen Lernens für die Nutzung zeitgemäßer Methoden ist.

Die hier vorliegende Ausarbeitung eines Qualitätsbegriffs für maschinelles Lernen in der amtlichen Statistik schafft lediglich eine konzeptionelle Grundlage. Darauf aufbauend ist es einerseits notwendig, durch Gremienbeschlüsse eine Verbindlichkeit bei der Qualitätssicherung zu erwirken und Qualitätsindikatoren so weit wie möglich in bestehende Qualitätsmanagementsysteme der deutschen amtlichen Statistik zu integrieren.[36] Andererseits besteht mittelfristig ein großer Bedarf nach der Entwicklung von Standardroutinen im Sinne des MLOps, sodass Einzellösungen durch standardisierte Systeme abgelöst werden, durch die eine effiziente Umsetzung qualitätsgesicherter maschineller Lernverfahren möglich wird.

5 Anhang

Tab. 13 Grundsätze und Indikatoren für die statistischen Prozesse (Statistische Ämter des Bundes und der Länder 2021)

Grundsatz 7: Solide Methodik	
Indikator 7.1	Der für europäische Statistiken verwendete allgemeine methodische Rahmen trägt europäischen und anderen internationalen Standards, Leitlinien und vorbildlichen Praktiken Rechnung
Indikator 7.2	Es gibt Verfahren, die gewährleisten, dass Standardkonzepte, -definitionen und -klassifikationen in der gesamten statistischen Stelle einheitlich verwendet werden
Indikator 7.3	Um eine hohe Qualität zu gewährleisten, werden das Unternehmensregister und die Erhebungsgrundlagen für Bevölkerungserhebungen regelmäßig evaluiert und sofern erforderlich angepasst
Indikator 7.4	Zwischen den nationalen und den europäischen Klassifikationssystemen besteht eine enge Übereinstimmung
Indikator 7.5	Es werden Absolventen der einschlägigen Studiengänge eingestellt

[36] Bestehende Instrumente des Qualitätsmanagements in der deutschen amtlichen Statistik sind etwa die Qualitätsrichtlinien (QRL), die Qualitätsdatenblätter im Verbund (QuiV) und die Qualitätsberichte einzelner Statistiken. Insbesondere die QRLs und die Qualitätsberichte bieten sich für die Qualitätsmessung und transparente Kommunikation etwaig genutzter ML-Verfahren an.

Tab. 13 (Fortsetzung)

Indikator 7.6	Die statistischen Stellen verfolgen eine Politik der kontinuierlichen beruflichen Weiterbildung ihrer Mitarbeiterinnen und Mitarbeiter
Indikator 7.7	Zur Verbesserung der Methodik sowie der Wirksamkeit angewandter Methoden und, sofern möglich, zur Förderung besserer Instrumente werden Maßnahmen in Zusammenarbeit mit der Wissenschaft durchgeführt
Grundsatz 8: Geeignete statistische Verfahren	
Indikator 8.1	Falls europäische Statistiken auf Verwaltungsdaten basieren, werden die für administrative Zwecke verwendeten Definitionen und Konzepte den Erfordernissen der Statistik soweit wie möglich angepasst
Indikator 8.2	Die Fragebogen für statistische Erhebungen werden vor der Erhebung der Daten systematisch getestet
Indikator 8.3	Die Erhebungspläne sowie die Stichprobenziehung und Schätzverfahren basieren auf soliden Grundlagen und werden regelmäßig überprüft und sofern erforderlich überarbeitet
Indikator 8.4	Die Datengewinnung sowie die Eingabe und Kodierung der Daten werden regelmäßig kontrolliert und sofern erforderlich angepasst
Indikator 8.5	Für das Editieren und Imputationen werden geeignete Verfahren eingesetzt, die regelmäßig überprüft und sofern erforderlich überarbeitet oder aktualisiert werden
Indikator 8.6	Revisionen erfolgen nach standardisierten, bewährten und transparenten Verfahren
Indikator 8.7	Die statistischen Stellen sind an der Gestaltung von Verwaltungsdaten beteiligt, um deren Eignung für statistische Zwecke zu erhöhen
Indikator 8.8	Es werden Vereinbarungen mit den Eignern von Verwaltungsdaten getroffen, in denen die gemeinsame Verpflichtung zur Nutzung dieser Daten für statistische Zwecke bekräftigt wird
Indikator 8.9	Die statistischen Stellen arbeiten mit den Eignern von Verwaltungsdaten zusammen, um die Datenqualität zu gewährleisten
Grundsatz 9: Vermeidung einer übermäßigen Belastung der Auskunftgebenden	
Indikator 9.1	Der Bedarf an Angaben für europäische Statistiken wird in Bezug auf Umfang und Gliederungstiefe auf das absolut erforderliche Maß begrenzt
Indikator 9.2	Der Beantwortungsaufwand wird so gleichmäßig wie möglich auf die Erhebungspopulationen verteilt
Indikator 9.3	Die von den Unternehmen verlangten Angaben werden soweit möglich direkt aus deren Buchhaltung entnommen, und im Interesse der leichteren Übermittlung dieser Angaben werden möglichst elektronische Hilfsmittel eingesetzt
Indikator 9.4	Administrative Datenquellen werden – wann immer möglich – herangezogen, um doppelte Datenanforderungen zu vermeiden
Indikator 9.5	Innerhalb der statistischen Stellen erfolgt generell eine gemeinsame Datennutzung, um eine Vervielfachung der Erhebungen zu vermeiden
Indikator 9.6	Die statistischen Stellen fördern Maßnahmen, die die Verknüpfung von Datenquellen ermöglichen, um den Beantwortungsaufwand zu reduzieren
Grundsatz 10: Wirtschaftlichkeit	
Indikator 10.1	Durch interne und unabhängige externe Maßnahmen wird der Ressourceneinsatz der statistischen Stelle überwacht
Indikator 10.2	Das Produktivitätspotenzial der Informations- und Kommunikationstechnologie wird bei der Datenerhebung, -verarbeitung und -verbreitung soweit als möglich ausgeschöpft
Indikator 10.3	Zur Vergrößerung des statistischen Potenzials von Verwaltungsdaten und zur Begrenzung des Zurückgreifens auf direkte Erhebungen werden proaktive Anstrengungen unternommen
Indikator 10.4	Zur Steigerung der Effizienz und Wirksamkeit fördern und realisieren die statistischen Stellen standardisierte Lösungen

Indikatoren mit Relevanz für den Einsatz von maschinellen Lernverfahren sind fett hervorgehoben.

Tab. 14 Grundsätze für die statistischen Produkte (Statistische Ämter des Bundes und der Länder 2021)

Grundsatz 11: Relevanz	
Indikator 11.1	Es gibt Verfahren zur Konsultation der Nutzerinnen und Nutzer, zur Überwachung der Relevanz bestehender Statistiken und des Ausmaßes, in dem sie den Bedarf der Nutzerinnen und Nutzer tatsächlich decken sowie zur Einbeziehung des neu entstehenden Bedarfs und der neu entstehenden Prioritäten der Nutzerinnen und Nutzer
Indikator 11.2	Prioritäre Anforderungen werden erfüllt und im Arbeitsprogramm abgebildet
Indikator 11.3	Die Zufriedenheit der Nutzerinnen und Nutzer wird regelmäßig überprüft und systematisch verfolgt
Grundsatz 12: Genauigkeit und Zuverlässigkeit	
Indikator 12.1	Die Basisdaten, die vorläufigen Ergebnisse und die statistischen Produkte werden regelmäßig evaluiert und validiert
Indikator 12.2	Stichprobenfehler und Nicht-Stichprobenfehler werden gemessen und systematisch gemäß den europäischen Standards dokumentiert
Indikator 12.3	Zur Verbesserung statistischer Prozesse werden Datenrevisionen regelmäßig analysiert
Grundsatz 13: Aktualität und Pünktlichkeit	
Indikator 13.1	Die Aktualität erfüllt europäische und andere internationale Veröffentlichungsstandards
Indikator 13.2	Für die Veröffentlichung der Statistiken wird ein täglicher Standardzeitpunkt bekanntgegeben
Indikator 13.3	Die Periodizität der Statistiken trägt dem Nutzerbedarf weitest möglich Rechnung
Indikator 13.4	Abweichungen vom Veröffentlichungskalender werden vorab bekanntgegeben und erläutert, und ein neuer Veröffentlichungszeitpunkt wird festgesetzt
Indikator 13.5	Vorläufige Ergebnisse von akzeptabler Gesamtgenauigkeit können veröffentlicht werden, wenn dies für nützlich erachtet wird
Grundsatz 14: Kohärenz und Vergleichbarkeit	
Indikator 14.1	Die Statistiken sind in sich kohärent und konsistent (d. h. die rechnerischen und buchungstechnischen Identitätsbeziehungen bleiben gewahrt)
Indikator 14.2	Die Statistiken sind über einen ausreichenden Zeitraum betrachtet vergleichbar
Indikator 14.3	Die Erstellung der Statistiken erfolgt auf der Grundlage von einheitlichen Standards in Bezug auf den Geltungsbereich, die Definitionen, die Einheiten und die Klassifikationen, die für die verschiedenen Erhebungen und Quellen gelten
Indikator 14.4	Die Statistiken aus den verschiedenen Quellen und von verschiedener Periodizität werden verglichen und miteinander in Einklang gebracht
Indikator 14.5	Die Vergleichbarkeit der Daten verschiedener Länder wird innerhalb des Europäischen Statistischen Systems durch regelmäßige Kontakte zwischen dem Europäischen Statistischen System und anderen statistischen Systemen gewährleistet, methodische Untersuchungen werden in enger Zusammenarbeit zwischen den Mitgliedstaaten und Eurostat durchgeführt
Grundsatz 15: Zugänglichkeit und Klarheit	
Indikator 15.1	Die Statistiken und die entsprechenden Metadaten werden in einer Weise präsentiert und archiviert, die eine korrekte Interpretation und aussagekräftige Vergleiche erleichtert
Indikator 15.2	Die Verbreitung erfolgt mit Hilfe moderner Informations- und Kommunikationstechnologie sowie, falls angemessen, durch gedruckte Veröffentlichungen
Indikator 15.3	Maßgeschneiderte Analysen werden, wenn dies möglich ist, bereitgestellt, und die Öffentlichkeit wird davon in Kenntnis gesetzt
Indikator 15.4	Der Zugang zu Mikrodaten ist zu Forschungszwecken gestattet und unterliegt besonderen Regeln oder Vorschriften
Indikator 15.5	Die Metadaten sind im Einklang mit standardisierten Metadaten-Systemen dokumentiert
Indikator 15.6	Die Nutzerinnen und Nutzer werden fortlaufend über die Methodik der statistischen Prozesse, einschließlich der Verwendung von Verwaltungsdaten, informiert

Indikatoren mit Relevanz für den Einsatz von maschinellen Lernverfahren sind fett hervorgehoben.

Funding Open Access funding enabled and organized by Projekt DEAL.

Interessenkonflikt Die Autorinnen und Autoren geben an, dass kein Interessenkonflikt besteht.

Literatur

Ahlborn M, Draken F, Schulz V (2021) Qualitätssicherung in der amtlichen Statistik: Large Cases Unit. Wista – Wirtschaft Stat (2):31–40 (www.destatis.de/DE/Methoden/WISTA-Wirtschaft-und-Statistik/2021/02/qualitaetssicherung_022021.html)

André M, Meslin O (2021) Housing wealth concentration and redistributive impact of property tax: evidence from a database on French households' housing wealth, S 2021–2004 (www.insee.fr/en/statistiques/5893230)

Andrews DWK (1986) Stability comparison of estimators. Econometrica 54(5):1207. https://doi.org/10.2307/1912329

Baker M (2016) 1,500 scientists lift the lid on reproducibility. Nature 533(7604):452–454. https://doi.org/10.1038/533452a

Bartz E, Bartz-Beielstein T, Zaefferer M, Mersmann O (Hrsg) (2023) Hyperparameter tuning for machine and deep learning with R. Springer, Singapore

Beck M, Dumpert F, Feuerhake J (2018) Machine learning in official statistics https://doi.org/10.48550/arXiv.1812.10422

Biemer PP (2010) Total survey error: design, implementation, and evaluation. Public Opin Q 74(5):817–848. https://doi.org/10.1093/poq/nfq058

Bischl B, Binder M, Lang M, Pielok T, Richter J, Coors S, Thomas J, Ullmann T, Becker M, Boulesteix A-L, Deng D, Lindauer M (2023) Hyperparameter optimization: foundations, algorithms, best practices, and open challenges. WIREs Data Min Knowl Discov 13(2):1–43. https://doi.org/10.1002/widm.1484

Blumöhr T, Teichmann C, Noack A (2017) Standardisierung der Prozesse: 14 Jahre AG SteP. Wista – Wirtschaft Stat (5):58–75 (www.destatis.de/DE/Methoden/WISTA-Wirtschaft-und-Statistik/2017/05/standardisierung-prozesse-052017.html)

Bothmann L, Peters K, Bischl B (2022) What is fairness? Implications for fairML https://doi.org/10.48550/arXiv.2205.09622

Branco P, Torgo L, Ribeiro RP (2017) A survey of predictive modeling on imbalanced domains. ACM Comput Surv 49(2):1–50. https://doi.org/10.1145/2907070

de Broe S, Struijs P, Daas P, van Delden A, Burger J, van den Brakel J, ten Bosch O, Zeelenberg K, Ypma W (2021) Updating the paradigm of official statistics: New quality criteria for integrating new data and methods in official statistics. Stat J IAOS 37(1):343–360. https://doi.org/10.3233/SJI-200711

Bruch C (2015) Varianzschätzung unter Imputation und bei komplexen Stichprobendesigns. Dissertation. Universität Trier, Trier.

Bundesanstalt für Finanzdienstleistungsaufsicht (BaFin), Deutsche Bundesbank (2022) Maschinelles Lernen in Risikomodellen – Charakteristika und aufsichtliche Schwerpunkte. Antworten auf das Konsultationspapier. www.bafin.de/SharedDocs/Downloads/DE/Konsultation/2021/dl_kon_11_21_Ergebnisse_machinelles_Lernen_Risikomodelle.html

Bundesanstalt für Finanzdienstleistungsaufsicht (BaFin), Deutsche Bundesbank (2021) Maschinelles Lernen in Risikomodellen – Charakteristika und aufsichtliche Schwerpunkte. Konsultationspapier (11/2021) (www.bafin.de/SharedDocs/Downloads/DE/Konsultation/2021/dl_kon_11_21_Diskussionspapier.html)

Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357. https://doi.org/10.1613/jair.953

Choi I, Del Monaco A, Law E, Davies S, Karanka J, Baily A, Piela R, Turpeinen T, Mharzi A, Rastan S, Flak K, Jentoft S (2022) ML model monitoring and re-training in statistical organisations (statswiki.unece.org/display/ML/Machine+Learning+Group+2022)

Deutscher Ethikrat (2023) Mensch und Maschine – Herausforderungen durch Künstliche Intelligenz. Stellungnahme. Vorabversion vom 20. März 2023. www.ethikrat.org/fileadmin/Publikationen/Stellungnahmen/deutsch/stellungnahme-mensch-und-maschine.pdf

Devlin J, Chang M-W, Lee K, Toutanova K (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Burstein J, Doran C, Solorio T (Hrsg) Proceedings of the 2019 Conference of the North, 4171–4186. Association for Computational Linguistics, Stroudsburg

Dumpert F (2021) Machine Learning in der amtlichen Statistik – Ergebnisse und Bewertung eines internationalen Projekts. WISTA Wirtsch Stat (4):53–63 (www.destatis.de/DE/Methoden/WISTA-Wirtschaft-und-Statistik/2021/04/machine-learning-042021.pdf)

Dumpert F (2023) Machine learning in German official statistics. In: Snijkers G, Bavdaž M, Bender S, Jones J, MacFeely S, Sakshaug JW, Thompson KJ, van Delden A (Hrsg) Advances in business statistics, methods and data collection. Wiley, S 537–560

Dumpert F, Beck M (2017) Einsatz von Machine-Learning-Verfahren in amtlichen Unternehmensstatistiken. AStA Wirtsch Sozialstat Arch 11(2):83–106. https://doi.org/10.1007/s11943-017-0208-6

Dumpert F, Schmidt E (2023) Hyperparameter Tuning in German Official Statistics. In: Bartz E, Bartz-Beielstein T, Zaefferer M, Mersmann O (Hrsg) Hyperparameter Tuning for Machine and Deep Learning with R. Springer, Singapore, S 177–185

Engdahl J, Choi I, Deeben E, Karanka J, Karlsson A, Meszaros M, Pocknee J, Holroyd P, Baily A (2022) Building an ML ecosystem in statistical organisations. statswiki.unece.org/display/ML/Machine+Learning+Group+2022

Europäische Kommission (2021) Vorschlag für eine VERORDNUNG DES EUROPÄISCHEN PARLAMENTS UND DES RATES zur Festlegung harmonisierter Vorschriften für Künstliche Intelligenz (Gesetz über Künstliche Intelligenz) und zur Änderung bestimmter Rechtsakte der Union. COM(2021) 206 final

Europäische Kommission, Eurostat (2018) Verhaltenskodex für Europäische Statistiken. Für die nationalen statistischen Ämter und Eurostat (statistisches Amt der EU). Amt für Veröffentlichungen der Europäischen Union, Luxemburg

European Commission (2020) On Artificial Intelligence—A European approach to excellence and trust. White Paper (COM(2020) 65 final). eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0065

European Commission High-Level Expert Group on Artificial Intelligence (AI HLEG) (2019) Ethics Guidelines for Trustworthy AI. strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

European Parliament (2020) European Parliament resolution of 20 October 2020 with recommendations to the Commission on a framework of ethical aspects of artificial intelligence, robotics and related technologies (2020/2012(INL)). Framework of ethical aspects of artificial intelligence, robotics and related

European Statistical System (2019) Quality assurance framework of the European statistical system (version 2.0). ec.europa.eu/eurostat/documents/64157/4392716/ESS-QAF-V2.0-final.pdf

Feuerhake J, Dumpert F (2016) Erkennung nicht relevanter Unternehmen in den Handwerksstatistiken. WISTA Wirtsch Stat (2):79–94 (www.destatis.de/DE/Methoden/WISTA-Wirtschaft-und-Statistik/2016/02/nichtrelevante-unternehmen-handwerk-022016.html)

Freiesleben T, König G, Molnar C, Tejero-Cantero A (2022) Scientific inference with interpretable machine learning: analyzing models to learn about real-world phenomena https://doi.org/10.48550/arXiv.2206.05487

Friedrich S, Antes G, Behr S, Binder H, Brannath W, Dumpert F, Ickstadt K, Kestler HA, Lederer J, Leitgöb H, Pauly M, Steland A, Wilhelm A, Friede T (2022) Is there a role for statistics in artificial intelligence? Adv Data Anal Classif 16(4):823–846. https://doi.org/10.1007/s11634-021-00455-6

Goldenberg I, Webb GI (2019) Survey of distance measures for quantifying concept drift and shift in numeric data. Knowl Inf Syst 60(2):591–615. https://doi.org/10.1007/s10115-018-1257-z

Gootzen YA, Daas PJ, van Delden A (2023) Quality framework for combining survey, administrative and big data for official statistics. Stat J IAOS 39(2):439–446. https://doi.org/10.3233/SJI-220110
Hampel FR (1968) Contributions to the theory of robust estimation. Ph.D. Thesis. University of California, Berkeley.
Hampel FR, Ronchetti EM, Rousseeuw PJ, Stahel WA (1986) Robust statistics. The approach based on influence functions. Wiley, New York
Hartmann PH, Lengerer A (2014) Verwaltungsdaten und Daten der amtlichen Statistik. In: Baur N, Blasius J (Hrsg) Handbuch Methoden der empirischen Sozialforschung. Springer, Wiesbaden, S 907–914
Hu X, Chu L, Pei J, Liu W, Bian J (2021) Model complexity of deep learning: a survey. Knowl Inf Syst 63(10):2585–2619. https://doi.org/10.1007/s10115-021-01605-0
Huber PJ (1964) Robust estimation of a location parameter. Ann Math Stat 35(1):73–101. https://doi.org/10.1214/aoms/1177703732
Joseph A (2022) Parametric inference with universal function approximators Bd. 784. https://doi.org/10.2139/ssrn.3351091
Julien C (2020) UNECE—HLG-MOS Machine Learning Project. Project report. statswiki.unece.org/display/ML/Machine+Learning+Project+Report
Kamath U, Liu J (2021) Explainable artificial intelligence: an introduction to interpretable machine learning. Springer, Cham
Kazmierski U, Ritsert R (2010) Zur Methodik von Wirtschaftlichkeitsuntersuchungen. In: Barthel C, Lorei C (Hrsg) Empirische Forschungsmethoden. Eine praxisorientierte Einführung für die Bachelor- und Masterstudiengänge der Polizei. Verl. für Polizeiwiss. Lorei, Frankfurt, M., S 161–188
Kinder-Kurlanda K, Weller K, Zenk-Möltgen W, Pfeffer J, Morstatter F (2017) Archiving information from geotagged tweets to promote reproducibility and comparability in social media research. Big Data Soc 4(2):205395171773633. https://doi.org/10.1177/2053951717736336
Klumpen D, Schäfer D (2012) Der Verhaltenskodex für europäische Statistiken (Code of Practice) in überarbeiteter Fassung 2011. WISTA Wirtsch Stat: 1035–1047 (www.destatis.de/DE/Methoden/WISTA-Wirtschaft-und-Statistik/2012/12/verhaltenskodex-2011-122012.html)
Köhler H, Christmann A (2022) Total Stability of SVMs and Localized SVMs. J Mach Learn Res 23(100):1–41
Kopsch G, Köhler S, Körner T (2006) Der Verhaltenskodex Europäische Statistiken (Code of Practice). WISTA Wirtsch Stat (8):793–804 (www.destatis.de/DE/Methoden/WISTA-Wirtschaft-und-Statistik/2006/08/verhaltenskodex-europaeische-statistiken-082006.pdf)
Kovaleva O, Romanov A, Rogers A, Rumshisky A (2019) Revealing the Dark Secrets of BERT. In: Inui K, Jiang J, Ng V, Wan X (Hrsg) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics, Stroudsburg, S 4364–4373
Kraff NJ, Wurm M, Taubenbock H (2020) Uncertainties of human perception in visual image interpretation in complex urban environments. IEEE J Sel Top Appl Earth Obs Remote Sens 13:4229–4241. https://doi.org/10.1109/JSTARS.2020.3011543
Kreuzberger D, Kühl N, Hirschl S (2022) Machine learning operations (MLops): overview, definition, and architecture https://doi.org/10.48550/arXiv.2205.02302
Kuhnt S, Kalka A (2022) Global sensitivity analysis for the interpretation of machine learning algorithms. In: Steland A, Tsui K-L (Hrsg) Artificial intelligence, big data and data science in statistics. Springer, Cham, S 155–169
Levagin B, Lange K, Walprecht S, Gerls F, Kühnhenrich D (2022) Vereinfachtes Verfahren zur interaktiven Schätzung des Erfüllungsaufwands mittels maschinellen Lernens. WISTA Wirtsch Stat (3) (www.destatis.de/DE/Methoden/WISTA-Wirtschaft-und-Statistik/2022/03/vereinfachtes-verfahren-erfuellungsaufwand-032022.pdf)
Lim C, Yu B (2016) Estimation Stability With Cross-Validation (ESCV). J Comput Graph Stat 25(2):464–492. https://doi.org/10.1080/10618600.2015.1020159
Lipton ZC (2018) The mythos of model Interpretability. Queue. https://doi.org/10.1145/3236386.3241340
Liu T, Yu H, Blair RH (2022) Stability estimation for unsupervised clustering: a review. WIREs Comput Stat 14(6):1–18. https://doi.org/10.1002/wics.1575
Lunardon N, Menardi G, Torelli N (2014) ROSE: a package for binary Imbalanced learning. R J 6(1):79–89. https://doi.org/10.32614/RJ-2014-008
Lundberg SM, Lee S-I (2017) A unified approach to interpreting model predictions. In: Proceedings of the 31st international conference on neural information processing systems. Curran Associates, Red Hook, S 4768–4777

Meertens QA, Diks C, van den Herik HJ, Takes FW (2022) Improving the output quality of official statistics based on machine learning algorithms. J Off Stat 38(2):485–508. https://doi.org/10.2478/jos-2022-0023
Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A (2022) A survey on bias and fairness in machine learning. ACM Comput Surv 54(6):1–35. https://doi.org/10.1145/3457607
Meinke I, Hentschke J (2022) Kern-Qualitätskennzahlen im Zensus 2022. Eine zensusspezifische Ausgestaltung der Qualitätsdatenblätter im Verbund. WISTA Wirtsch Stat (3):25–38 (www.destatis.de/DE/Methoden/WISTA-Wirtschaft-und-Statistik/2022/03/kern-qualitaetskennzahlen-032022.html)
Menardi G, Torelli N (2014) Training and assessing classification rules with imbalanced data. Data Min Knowl Disc 28(1):92–122. https://doi.org/10.1007/s10618-012-0295-5
Meyer C, Alsabah N (2022) Herausforderung „Verlässliche KI". Behörden Spieg 38(IV):40
Molnar C (2022) Interpretable machine learning. A guide for making black box models explainable. Christoph Molnar, Munich.
Molnar C, König G, Herbinger J, Freiesleben T, Dandl S, Scholbeck CA, Casalicchio G, Grosse-Wentrup M, Bischl B (2022) General pitfalls of model-agnostic interpretation methods for machine learning models. In: Holzinger A, Goebel R, Fong R, Moon T, Müller K-R, Samek W (Hrsg) xxAI—Beyond Explainable AI. Springer, Cham, S 39–68
Munafò MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, Du Sert NP, Simonsohn U, Wagenmakers E-J, Ware JJ, Ioannidis JPA (2017) A manifesto for reproducible science. Nat Hum Behav 1:21. https://doi.org/10.1038/s41562-016-0021
Nguyen JD, Hogue CR (2019) Automatically generated quality control tables and quality improvement programs1. Stat J IAOS 35(2):193–200. https://doi.org/10.3233/SJI-180461
Peng RD (2011) Reproducible research in computational science. Science 334(6060):1226–1227. https://doi.org/10.1126/science.1213847
Poretschkin M, Schmitz A, Akila M, Adilova L, Becker D, Cremers AB, Hecker D, Houben S, Mock M, Rosenzweig J, Sicking J, Schulz E, Voß A, Wrobel S (2021) Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz (KI-Prüfkatalog). www.iais.fraunhofer.de/de/forschung/kuenstliche-intelligenz/ki-pruefkatalog.html
Preising M, Lange K, Dumpert F (2021) Imputation zur maschinellen Behandlung fehlender und unplausibler Werte in der amtlichen Statistik. WISTA Wirtsch Stat (5) (www.destatis.de/DE/Methoden/WISTA-Wirtschaft-und-Statistik/2021/05/imputation-maschinelle-behandlung-052021.pdf)
Puts M, Daas P (2021) Machine learning from the perspective of official statistic. Surv Stat 84:12–17
Puts M, da Silva A, Di Consiglio L, Choi I, Salgado D, Clarke C, Jones S, Baily A (2022) Quality of training data. statswiki.unece.org/display/ML/Machine+Learning+Group+2022
Radermacher WJ (2022) Statistical awareness promoting a data culture. Stat J IAOS 38(2):453–461. https://doi.org/10.3233/SJI-220956
Reister M (2023) Assuring quality in the new data ecosystem: mind the gap between data and statistics! Stat J IAOS 39(2):421–430. https://doi.org/10.3233/SJI-230008
Ribeiro M, Singh S, Guestrin C (2016) "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: DeNero J, Finlayson M, Reddy S (Hrsg) Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations. Association for Computational Linguistics, Stroudsburg, S 97–101
Rothe P (2015) Statistische Geheimhaltung – der Schutz vertraulicher Daten in der amtlichen Statistik. Teil 1: Rechtliche und methodische Grundlagen. Bayern Zahl (5):294–303 (www.statistischebibliothek.de/mir/receive/BYMonografie_mods_00000049)
Saidani Y, Bohnensteffen S, Hadam S (2022) Qualität von Mobillfunkdaten – Projekterfahrungen und Anwendungsfälle aus der amtlichen Statistik. WISTA Wirtsch Stat (5):55–67 (www.destatis.de/DE/Methoden/WISTA-Wirtschaft-und-Statistik/2022/05/qualitaet-mobilfunkdaten-052022.html)
Salgado D, Barragán S, Rosa-Pèrez E (2023) Timeliness and accuracy with machine learning algorithms: early estimates of the industrial turnover index. unece.org/statistics/documents/2023/05/ml2023s1spainsalgadopaperpdf
Saltelli A, Ratto M, Andres T, Campolongo F, Cariboni J, Gatelli D, Saisana M, Tarantola S (2007) Global sensitivity analysis. The primer. Wiley
Sani HM, Lei C, Neagu D (2018) Computational complexity analysis of decision tree algorithms. In: Bramer M, Petridis M (Hrsg) Artificial intelligence XXXV. Springer, Cham, S 191–197
Schwemmer C, Knight C, Bello-Pardo ED, Oklobdzija S, Schoonvelde M, Lockhart JW (2020) Diagnosing gender bias in image recognition systems. Socius. https://doi.org/10.1177/2378023120967171
Sculley D, Holt G, Golovin D, Davydov E, Phillips T, Ebner D, Chaudhary V, Young M, Crespo J-F, Dennison D (2015) Hidden technical debt in machine learning systems. In: Cortes C, Lawrence N,

Lee D, Sugiyama M, Garnett R (Hrsg) Advances in neural information processing systems. Curran Associates,
Shankar S, Garcia R, Hellerstein JM, Parameswaran AG (2022) Operationalizing machine learning: an interview study https://doi.org/10.48550/arXiv.2209.09125
Signorelli S, Fontana M, Gabrielli L, Vespe M (2022) Challenges and opportunities of computational social science for official statistics https://doi.org/10.48550/arXiv.2207.13508
Slack D, Hilgard S, Jia E, Singh S, Lakkaraju H (2020) Fooling LIME and SHAP. In: Markham A, Powles J, Walsh T, Washington AL (Hrsg) Proceedings of the AAAI/ACM conference on AI, ethics, and society. ACM, New York, S 180–186
Statistische Ämter des Bundes und der Länder (2021) Qualitätshandbuch der Statistischen Ämter des Bundes und der Länder (Version 1.21). www.destatis.de/DE/Methoden/Qualitaet/qualitaetshandbuch.pdf
Statistischer Beirat (2010) Eckpunkte zur Weiterentwicklung der amtlichen Statistik in der 17. Legislaturperiode. bdi.eu/media/themenfelder/industriepolitik/downloads/201002_Eckpunkte-Weiterentwicklung-der-amtlichen-Statistik.pdf
Sæbø HV, Holmberg A (2019) Beyond code of practice: new quality challenges in official statistics. Stat J IAOS 35(2):171–178. https://doi.org/10.3233/SJI-180463
Thurow M, Dumpert F, Ramosaj B, Pauly M (2021) Goodness (of fit) of imputation accuracy: the goodimpact analysis https://doi.org/10.48550/arXiv.2101.07532
Tukey JW (1959) A survey of sampling from contaminated distributions. Princeton University Press, Princeton, New Jersey
Tümmler T (2020) Qualität bei zusammengeführten Daten. In: Klumpe B, Schröder J, Zwick M (Hrsg) Qualität bei zusammengeführten Daten. Springer, Wiesbaden, S 81–95
Tümmler T, Meinke I (2019) Aufbau des Qualitätsmanagements für den Zensus 2021. WISTA Wirtsch Stat: 59–73 (www.destatis.de/DE/Methoden/WISTA-Wirtschaft-und-Statistik/2019/07/aufbau-qualitaetsmanagement-zensus-2021-072019.html)
United Nations Economic Commission for Europe (UNECE) (2019) Generic statistical business process model (GSBPM). statswiki.unece.org/display/GSBPM/GSBPM+v5.1 (Erstellt: 01.2019)
United Nations Economic Commission for Europe (UNECE) (2021) Machine learning for official statistics. unece.org/statistics/publications/machine-learning-official-statistics
de Waal T, van Delden A, Scholtus S (2019) Quality measures for multisource statistics. Stat J IAOS 35(2):179–192. https://doi.org/10.3233/SJI-180468
Wager S, Athey S (2018) Estimation and inference of heterogeneous treatment effects using random forests. J Am Stat Assoc 113(523):1228–1242. https://doi.org/10.1080/01621459.2017.1319839
Wilson EB (1927) Probable inference, the law of succession, and statistical inference. J Am Stat Assoc 22(158):209–212. https://doi.org/10.1080/01621459.1927.10502953
Yu B (2013) Stability. Bernoulli 19(4):1484–1500. https://doi.org/10.3150/13-BEJSP14
Yu B, Kumbier K (2020) Veridical data science. Proc Natl Acad Sci U S A 117(8):3920–3929. https://doi.org/10.1073/pnas.1901326117
Yung W, Karkimaa J, Scannapieco M, Barcarolli G, Zardetto D, Sanchez JAR, Braaksma B, Buelens B, Burger J (2018) The use of machine learning in official statistics. statswiki.unece.org/download/attachments/120128748/The%20use%20of%20machine%20learning%20in%20official%20statistics.pdf
Yung W, Tam S-M, Buelens B, Chipman H, Dumpert F, Ascari G, Rocci F, Burger J, Choi I (2022) A quality framework for statistical algorithms. Stat J IAOS 38(1):291–308. https://doi.org/10.3233/SJI-210875