Transparenz und Güte der Ergebnisse von Wertermittlungen – Teil 3: Resampling-Verfahren, Bootstrapping und die Ergebnisgenauigkeit in der Wertermittlung

Transparency and Quality of Valuation Results – Part 3: Resampling Methods, Bootstrapping, and the Accuracy of Results in Valuation

Peter Ache | Christian Müller-Kett

Zusammenfassung
Unter Berücksichtigung der in den Teilen 1 und 2 der Folge zu Transparenz und Güte von Wertermittlungen gezogenen Schlussfolgerungen werden in diesem Betrag beide Begriffe tiefergehend erläutert und anhand datenwissenschaftlicher Verfahren wie Kreuzvalidierung und Bootstrapping zur Überprüfung der Genauigkeit von Bewertungsmodellen diskutiert. Durch die Quantifizierung von Verzerrungsfreiheit und Präzision werden die Anforderungen der ImmoWertV unterstützt und es wird die Nachvollziehbarkeit der Ergebnisse für Anwender und Entscheidungsträger verbessert.

Schlüsselwörter: Immobilienwertermittlung, Ergebnisgenauigkeit, Modellperformance, Künstliche Intelligenz, Resampling, Kreuzvalidierung, Bootstrapping

Summary
Taking into account the conclusions drawn in parts 1 and 2 of the series on transparency and quality of the results of valuation, this contribution provides a deeper examination of both concepts and discusses them by means of data-scientific methods such as cross-validation and bootstrapping for determining the accuracy of valuation models. By quantifying bias-freeness and predictive accuracy, the requirements of the ImmoWertV are supported and the comprehensibility of the results for users and decision-makers is enhanced.

Keywords: *real estate valuation, model accuracy, model performance, artificial intelligence, resampling, cross-validation, bootstrapping*

1 Einleitung

Die Immobilienwertermittlung in Deutschland befindet sich inmitten eines tiefgreifenden Wandels: Die zunehmende Verfügbarkeit digitaler Kaufpreisdaten, die rechtliche Verankerung statistischer Verfahren in der Immobilienwertermittlungsverordnung (vom 14. Juli 2021, Immo WertV 2021, BGBl. I S. 2805) und deren Muster-Anwendungshinweisen (ImmoWertA) vom 20. September 2023 sowie der mittlerweile zwingend vorauszusetzende Einzug von Künstlicher Intelligenz (KI) und maschinellem Lernen in die eher konservative und vergleichsweise streng regulierte Umgebung der klassischen Wertermittlung von Immobilien erfordern ein – nahezu disruptives – und zügiges Umdenken. Dies gilt nicht nur für Deutschland, sondern wird intensiv auch auf internationaler Ebene diskutiert (Dimopoulos et al. 2024).

Ein zentrales Element dieses Wandels ist die Forderung nach einer stärker evidenzbasierten Wertermittlung, wie sie in Ache (2025a) grundlegend motiviert wird. Hier geht es vorrangig um die Anforderungen von Immobilienwertermittlungen auf der Grundlage empirischer Daten, die aus dem realen Markt stammen, sowie tatsächlich gezahlter Kaufpreise und Informationen zu den Umständen der jeweiligen Transaktion. Die bisherigen traditionellen, eher theoretischen Methoden sollten zunehmend in den Hintergrund treten und den Weg für verzerrungsfreie, präzise und damit vornehmlich datenbasierte Wertermittlungsansätze frei machen.

In der in dieser Reihe geführten Diskussion zur Güte evidenzbasierter Immobilienbewertungen stellen die Konzepte der Modellperformance und der Ergebnisgenauigkeit die zentralen Kriterien der Bewertung von Ergebnissen bei der Ermittlung von Verkehrswerten, insbesondere aber auch bei der Ermittlung der für die Wertermittlung erforderlichen Daten dar. Die Modellperformance beschreibt dabei die Fähigkeit eines Schätzverfahrens, unter unterschiedlichen Stichprobenbedingungen konsistente, robuste, wiederholbare und nicht systematisch verzerrte Ergebnisse zu generieren. Im Gegensatz dazu bezieht sich die Ergebnisgenauigkeit auf die Abweichung der Modellschätzungen von den wahren Werten in Bezug auf das tatsächliche Marktgeschehen. Hierbei lassen sich weiter die Unverzerrtheit und Präzision der Schätzwerte unterscheiden. Ein wesentliches Instrument zur Beurteilung der Ergebnisse dieser Konzepte ist die Bestimmung geeigneter Metriken, welche die Unsicherheit der Schätzwerte in probabilistischer Form

  DOI 10.12902/zfv-0534-2025

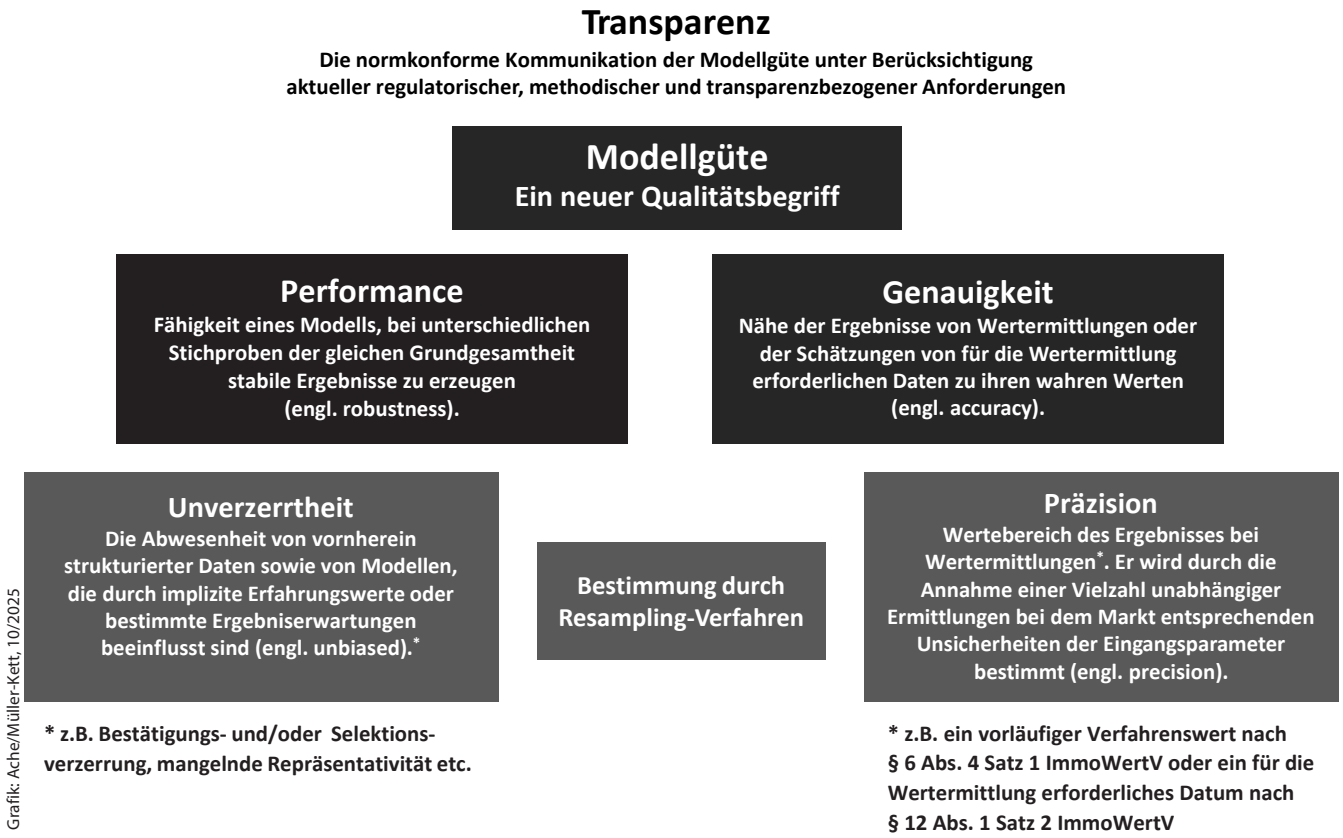

Abb. 1: Konzept der Modellgüte bei der Immobilienwertermittlung

abbilden. Die explizite Quantifizierung dieser Unsicherheit ist insbesondere vor dem Hintergrund steigender Anforderungen an Transparenz und Reproduzierbarkeit in der Wertermittlung von essenzieller Bedeutung.

Die Ermittlung und Kommunikation der Modellgüte ist dabei nicht nur eine methodische Option, sondern ein direktes Resultat unter anderem der aktuellen Anforderungen nach § 12 ImmoWertV, wonach geeignete statistische Verfahren verpflichtend zur Anwendung kommen sollen, wenn für die Wertermittlung erforderliche Daten ermittelt werden. Gleiches gilt auch für die Ermittlung von Verkehrswerten, geht es auch hier darum, die Aussagefähigkeit von Verfahrenswerten nach § 6 ImmoWertV zu beurteilen, um auf dieser Grundlage den Verkehrswert als »Punktwert« abzuleiten. Es genügt daher nicht mehr, lediglich einen Verfahrenswert zu ermitteln, vielmehr ist es erforderlich, dessen Güte zu quantifizieren (vgl. Abb. 1). Neben der ikonischen Beschreibung von Unverzerrtheit und Präzision am Beispiel von Karten mit unterschiedlichen Maßstäben (Ache 2025b) kann auch die in Anlehnung an die Darstellungen in Neumann (2009, S. 17 f.) verwendete Zielscheibenanalogie zur Klarstellung beitragen (Abb. 2). Dabei ist das Zentrum der konzentrischen Kreise als der wahre Wert einer Wertermittlung im weitesten Sinne zu verstehen; die Abweichungen der Wertermittlungsergebnisse können negativ oder positiv sein.

Abb. 2 verdeutlicht, dass Modelle trotz hoher Präzision die wahren Werte der Grundgesamtheit verfehlen können (Teilbild II). Modellverzerrungen (eng. biases) sind schwer zu identifizieren, weil die wahren Werte unbekannt sind und lediglich Modelle zur Schätzung vorliegen. Eine einzelne Kennzahl, die eine Verzerrung verlässlich »aufdeckt«, existiert nicht. Verzerrungen sind vielmehr Folge der Modellstruktur und ihrer Annahmen (eng. model bias) sowie möglicher Verzerrungen in der Datengrundlage (Stichprobenverzerrung, eng. sampling bias). Die Nicht- oder falsche Berücksichtigung von Lage- oder Nachbarschaftseinflüssen (z. B. Straße trennt Wohnqualitätszonen) kann ebenso zu Modellverzerrungen führen, wie die Nichtbeachtung von Annahmen z. B. bei der linearen Regressionsanalyse. Die unreflektierte Herausnahme von Beobachtungen als

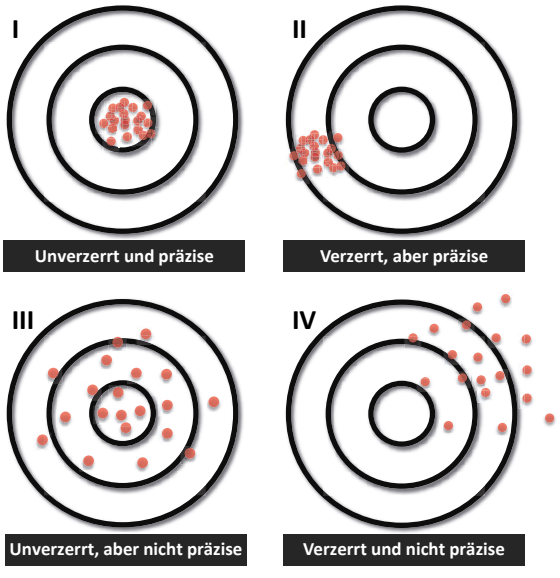

Abb. 2: Unverzerrtheit und Präzision

»Ausreißer« oder die vorab auf ein Ergebnis gerichtete Selektion von Kaufpreisinformationen aus Preisdatenbanken kann zu Stichprobenverzerrungen führen, die das Ergebnis stark von seinem wahren Wert abweichen lässt.

So ist die präzise Bestimmung der Repräsentativität einer Stichprobe von zentraler Bedeutung, insbesondere im Hinblick auf die Beurteilung der Leistungsfähigkeit statistischer Modelle (Ache 2025b). George Gallup (1901–1984), einem der maßgeblichen Begründer der modernen Meinungsforschung, wird hierfür die bekannte Analogie der Suppenschüssel zugeschrieben: Um den Geschmack einer Suppe zu beurteilen, sei es nicht erforderlich, den gesamten Inhalt zu verzehren; vielmehr genüge eine kleine Kostprobe, sofern die Suppe zuvor gründlich durchmischt wurde. Übertragen auf die Stichprobentheorie bedeutet dies, dass jedem Element der Grundgesamtheit die gleiche Auswahlwahrscheinlichkeit zukommen muss, damit die gezogene Stichprobe als repräsentativ gelten kann.

In diesem Sinne dienen repräsentative Zufallsstichproben dazu, systematische Verzerrungen, Voreingenommenheiten und andere Formen der Stichprobenverzerrung zu minimieren. Davon abzugrenzen ist die Modellverzerrung: Auch bei einer methodisch korrekt gezogenen Stichprobe kann ein Modell fehlerhafte Ergebnisse produzieren, wenn es auf unzulässigen Vereinfachungen beruht oder die theoretischen Voraussetzungen für seine Anwendung nicht erfüllt sind. Ein klassisches Beispiel ist die Verwendung linearer Regressionsmodelle zur Abbildung nichtlinearer Zusammenhänge und die Nichteinhaltung der Anwendungsvoraussetzungen.

Gleichwohl sind hohe Fehlermetriken in Trainings- und Testdaten ein starkes Indiz für Bias. Ergänzend bieten Streudiagramme der prädizierten gegenüber den beobachteten Werten ein wirksames Diagnoseinstrument (vgl. Ache 2025b, S. 259, Abb. 1).

Bei hoher, jedoch unverzerrter Streuung (Abb. 2, Teilbild III) ist der Schätzer erwartungstreu und trifft den wahren Wert im Mittel, die Präzision der Einzelprognosen bleibt jedoch gering. Diese geringe Präzision lässt sich mit Metriken wie RMSE, RMdSE, R^2 und MdAPE gut messen. Dabei ist nach den Erfahrungen in der Wertermittlung die Annahme nicht symmetrisch verteilter Residuen naheliegender als eine symmetrische Verteilung. Bei der Anwendung von Machine-Learning-Methoden liegt ja – im Gegensatz zu linearen Regressionsmodellen – oft keine modellbezogene Grundannahme der symmetrischen Verteilung der Residuen vor; daher liegt es nahe, den Medianwert der Residuen als für die Beurteilung der Präzision geeigneteren Lagewert zu verwenden. Zudem entspricht der arithmetische Mittelwert bei symmetrischer Verteilung ohnehin dem Medianwert. Sollen allerdings explizit unsymmetrische Abweichungen überrepräsentiert werden, bieten sich Fehlermaße an, welche anfällig für Ausreißer sind, wie z. B. der RMSE und der Vergleich mit dem RMdSE, um solche Auffälligkeiten sichtbar zu machen.

Angesichts der heute verfügbaren Software, hoher Rechenkapazitäten und des vermehrten Einsatzes automatisierter Modelle (eng. Automated Valuation Models, AVM), Machine-Learning-Methoden sowie algorithmisch erzeugter Massenbewertungen, etwa im steuerlichen Kontext, gewinnt die Veröffentlichung der Modellgüte stark an Bedeutung. Die systematische Integration geeigneter Gütemaße ist daher keine fakultative Ergänzung, sondern eine zwingende Voraussetzung für eine risikoadäquate Modellierung von Immobilienmärkten und die Ermittlung von Immobilienwerten insgesamt. Dabei ist die – eher historisch insinuierte – Annahme von normalverteilten Beobachtungen auf dem Immobilienmarkt regelmäßig nicht zutreffend. Nur durch die transparente Kommunikation geeigneter Metriken können Bewertungsrisiken offengelegt und in nachgelagerte Finanzierungs-, Investitions- oder Steuerungsentscheidungen adäquat einbezogen werden.

Die Vernachlässigung einer fundierten Einschätzung der Modellgüte in der Immobilienbewertung birgt erhebliche Risiken auf mikro- wie makroökonomischer Ebene. Es besteht die Gefahr, dass Immobilienmarktinformationen mit trügerischer Präzision oder gar mit unerkannten Verzerrungen (vgl. Abb. 2, Teilbild II) interpretiert werden. Dies kann gravierende Fehlentscheidungen in Finanzierungs-, Investitions- oder Besteuerungskontexten nach sich ziehen. Auch bei der Ableitung von Liegenschaftszinssätzen, Bodenrichtwerten oder Sach- und Vergleichsfaktoren können unzureichend validierte Modelle strukturelle Verzerrungen enthalten, die mangels transparenter Angaben zur Modellgüte unbeachtet bleiben. Solche »falschen Modelle« bergen das Potenzial systematischer Fehlbewertungen, die kumulativ zu massiven Fehlallokationen langfristig gebundenen Kapitals führen und letztlich das Vertrauen in die Bewertungsinstitutionen unterminieren.

Im vorliegenden Zusammenhang ist festzuhalten, dass die Veröffentlichung von für die Wertermittlung erforderlichen Daten, die auf »falschen« oder nicht normkonformen Modellansätzen beruhen, nicht durch den Grundsatz der fachlichen Unabhängigkeit der Gutachterausschüsse gedeckt ist. Ein solches Vorgehen ist vielmehr dem Bereich der Dienstaufsicht zuzuordnen, da einschlägige gesetzliche und untergesetzliche Vorgaben – insbesondere die in der ImmoWertV normierten Anforderungen an Datenqualität und Modelltransparenz – verletzt werden. Die rechtliche Verortung ist damit eindeutig und in die Systematik der ImmoWertV und des Baugesetzbuches (BauGB) eingebettet.

Dies gilt in Wertermittlungsverfahren sowohl für die Ermittlung der vorläufigen Verfahrenswerte (§ 6 Abs. 3 Nr. 1 ImmoWertV), die nach Bewertung ihrer Aussagekraft in den jeweiligen Verkehrswert überführt werden (§ 6 Abs. 4 ImmoWertV), als auch für die Bestimmung sonstiger für die Wertermittlung erforderlicher Daten gemäß § 12 Abs. 1 Satz 2 ImmoWertV. Bisher wurde in der Immobilienwertermittlung das Konzept des systematischen Trainierens und Testens von Modellen nur eingeschränkt, meistens aber gar nicht berücksichtigt. Angesichts der wachsenden Anforderungen an Transparenz und Nachvollziehbarkeit rückt es jedoch zunehmend in den Fokus.

Eine naive, bislang aber oft praktizierte Vorgehensweise besteht darin, ein Modell mit sämtlichen verfügbaren Daten der Stichprobe zu schätzen und anschließend seine Güte mit denselben Daten zu prüfen. Dabei bleibt jedoch unklar, in welchem Maße das Modell tatsächlich präzise und unverzerrt ist, wenn es auf bislang unbekannte Phänomene der Grundgesamtheit angewandt wird. Aus diesem Grund dürfen Modelle nicht mit allen verfügbaren Daten trainiert und validiert werden, sondern ausschließlich mit einem ausgewählten Teil der Stichprobe (engl. resamples), wobei zurückgehaltene Daten als Testdaten zur Einschätzung der Modellgüte verwendet werden.

2 Resampling-Verfahren

Als Resampling-Verfahren werden im Allgemeinen Methoden bezeichnet, die wiederholtes Ziehen eines Anteils aus vorhandenen Daten nutzen, um die Performance und die Genauigkeit von Modellergebnissen zu beurteilen. Zu den klassischen Ansätzen zählen das Bootstrapping (Efron und Tibshirani 1986, 1993) und die Kreuzvalidierung (Hastie et al. 2009). Gerade in der Immobilienwertermittlung, in der häufig heterogene und begrenzte Datensätze vorliegen, ermöglichen diese Methoden eine zuverlässige Einschätzung der Modellgüte. Resampling-Verfahren benötigen keine theoretischen Annahmen über die Verteilungen von Variablen (z. B. sind Kaufpreisdaten selten normalverteilt) und eignen sich, wenn klassische Voraussetzungen wie z. B. die Normalverteilung der zu untersuchenden Daten fraglich sind.

Ähnliche Grundprinzipien nutzen auch andere Resampling-Verfahren, die aus einer beobachteten Stichprobe viele neue simulierte Stichproben erzeugen, um so Rückschlüsse auf die unbekannte Grundgesamtheit und damit auf einen unbekannten Prozess der Datengenerierung ziehen zu können.

Ein älteres Verfahren, entwickelt in den 1950er Jahren, ist die **Jackknife-Methode** (z. B. Tukey 1958). Auch hier werden Metriken ermittelt, die es erlauben – wie beim Bootstrapping – Unsicherheiten von statistischen Schätzungen zu quantifizieren (Streuung, Verzerrung, Konfidenzintervall). Dies geschieht, indem aus einem bestehenden Datensatz eine Vielzahl von Teildatensätzen erzeugt werden, in denen systematisch jeweils eine Beobachtung entfernt wird. Bei z. B. 80 Beobachtungen zu Wohnungspreisen je m^2 würden 80 Stichproben simuliert, bei denen systematisch jeweils eine Beobachtung fehlt und der jeweilige Medianwert aus diesen 79 Beobachtungen ermittelt wird. Somit entstehen insgesamt 80 Medianwerte, aus denen dann die erforderlichen Streuungsstatistiken ermittelt werden können. Dieses Verfahren ist jedoch gegenüber dem Bootstrapping-Verfahren weniger effizient.

Der **Permutationstest**, eine der ältesten Resampling-Methoden, wurde von Ronald A. Fisher in den 1930er Jahren entwickelt (Fisher 1935). Ziel dieses Verfahrens ist es, die Nullhypothese zu überprüfen, nach der eine Behandlung oder ein Merkmal keinen Einfluss auf die Zielgröße hat. In der Immobilienwertermittlung lässt sich so beispielsweise untersuchen, ob Photovoltaikanlagen einen signifikanten Einfluss auf Immobilienpreise ausüben. Dazu werden die Preise von Objekten mit und ohne Photovoltaikanlage wiederholt zufällig permutiert, um eine Verteilung der Preisunterschiede unter Annahme der Nullhypothese zu erzeugen. Auf diese Weise kann geprüft werden, ob der beobachtete Preiseffekt statistisch signifikant ist. Permutationsverfahren eignen sich zudem zur Bewertung des Einflusses kategorialer und quantitativer Merkmale (»Feature Importance«) auf die Zielgröße, etwa indem die Auswirkung der Wohnfläche auf den Preis durch wiederholtes Permutieren und Modellberechnungen quantifiziert wird (Fisher et al. 2019).

Im erweiterten Vorgehen können diese Methoden zur Merkmalsbewertung mit dem Bootstrapping-Verfahren kombiniert werden, z. B. indem beurteilt wird, welche Variablen in mehreren Bootstrapping-Durchgängen am häufigsten in den besten Modellen vorkommen. Dazu soll hier lediglich auf Hastie et al. (2009, S. 658, 681 ff.), Guyon und Elisseeff (2003), James et al. (2013, Kap. 6) und Yates et al. (2023) verwiesen werden.

Die **Monte-Carlo-Simulation** (z. B. Carsey und Harden 2013) nutzt ein anderes Anwendungsprinzip. Hier liegen dem Grunde nach keine Beobachtungen vor. Der Prozess der Bildung einer theoretischen – oder besser simulierten – Grundgesamtheit ist durch den Anwender vollständig kontrolliert und verantwortet. Auch hier werden durch wiederholtes Ziehen von sehr vielen Stichproben aus einer kontrollierten Wahrscheinlichkeitsverteilung der interessierenden Variablen Effekte und Metriken der Performance und Genauigkeit von komplexen mathematischen Prozessen sichtbar gemacht. Diese Simulationsmethode wird eher im Zusammenhang mit klassischen Verfahren wie der Ertrags- und Sachwertmethode oder bei deduktiven Berechnungsverfahren zum Tragen kommen (Haak 2008, Long 2017). Es kommt hier in erster Linie darauf an, die Wahrscheinlichkeitsverteilungen der in der Wertermittlungsmethode einzusetzenden Daten (z. B. Liegenschaftszinssatz, Restnutzungsdauern etc.) möglichst nahe an dem tatsächlichen Immobilienmarkt zu bestimmen.

Diese Techniken ermöglichen eine gezielte Bewertung der Modellgüte und erhöhen so die Transparenz der Wertermittlung, vor allem angesichts frei verfügbarer und weit entwickelter Statistikprogramme. Dazu ist jedoch fundiertes statistisches Wissen für eine sinnvolle Anwendung unerlässlich.

Im Folgenden sollen die Verwendung der Kreuzvalidierung und des Bootstrapping-Verfahrens für Anwendungen in der Immobilienwertermittlung näher diskutiert werden.

3 Quantifizierung der Verzerrung durch Kreuzvalidierung

Zur Schätzung der Vorhersagekraft eines Modells auf bislang unbekannte Daten sowie zur Identifikation möglicher Überanpassungen (eng. overfitting) eignet sich insbesondere die Kreuzvalidierung (eng. cross validation). In diesem Zusammenhang sprechen Yates et al. (2023) von »prediktiver Performance«, die durch eine Verlustfunktion (eng. loss function) numerisch quantifiziert und als »Score« geschätzt wird. Dabei wird gemessen, wie gut die Vorhersagen von Modellen mit den tatsächlichen Beobachtungen übereinstimmen. Bei Regressionsproblematiken entspricht dies der Kleinste-Quadrate-Regression. Besonders im Kontext der Machine-Learning-Algorithmen wird die Kreuzvalidierung allgemeinhin als geeignete Methode zur Abschätzung der Güte eines Modells angesehen. Von den verschiedenen Ausprägungen dieser Methode gilt die *k*-fache Kreuzvalidierung – teilweise auch als V-fold Cross Validation bezeichnet – als die etablierteste und am weitesten verbreitete Variante.

Die Kreuzvalidierung wird im vorliegenden Kontext zur Modellschätzung innerhalb des Trainingsdatensatzes eingesetzt (Abb. 3). Ihr grundlegendes Schema ähnelt dem der klassischen Resampling-Verfahren.

Zunächst wird die verfügbare Stichprobe nach dem Zufallsprinzip in einen Trainings- und einen Testdatensatz unterteilt, häufig im Verhältnis 80:20, 75:25 oder 90:10. Einige Anwendungsanleitungen geben an dieser Stelle pauschale Zahlenverhältnisse vor, die sich auf keine mathematische Argumentation stützen. Vielmehr sollte das tatsächlich gewählte Verhältnis dem Anwendungsfall und den zur Verfügung stehenden Daten entsprechen. Der Testdatensatz bleibt vollständig unberührt von der Modellbildung und dient ausschließlich der abschließenden Überprüfung. Die Modellierung erfolgt ausschließlich auf Grundlage des Trainingsdatensatzes. Für die Kreuzvalidierung wird dieser wiederum in *k* gleich große, zufällig gebildete Teilmengen zerlegt, wobei jeweils *k*–1 Teilmengen zur Schätzung und die verbleibende Teilmenge zur Validierung genutzt werden. In diesem Zusammenhang ist darauf hinzuweisen, dass bei – wie in der Immobilienwertermittlung zurzeit noch oft der Fall – kleinen Datensätzen eine Kreuzvalidierung ebenso wenig sinnvoll ist, wie die Entwicklung von komplexen Modellen im Allgemeinen. Datensätze mit sehr wenigen Immobilienpreisen eignen sich schon allein aufgrund des allgemeinen Rauschens in den betreffenden Teilmärkten nicht, um geeignete Modelle zu entwickeln. Auch an dieser Stelle kann keine pauschale Mindestanzahl genannt werden, da sich die jeweiligen Anwendungsfälle, Datenstrukturen, lokalen Besonderheiten etc. unterscheiden. Als Anhaltspunkte zur Einschätzung der Eignung einer Datengrundlage zur Modellbildung können aber z. B. die Anzahl der statistischen Freiheitsgrade und die statistische Teststärke dienen (siehe z. B. Dorey 2011, Serdar et al. 2021).

Ein Beispiel verdeutlicht das Vorgehen: Liegt ein Trainingsdatensatz mit 80 Beobachtungen vor und wird eine 5-fache Kreuzvalidierung durchgeführt, so wird der Datensatz in fünf Teilmengen zu je 16 Beobachtungen aufgeteilt (Abb. 4).

In den einzelnen Folds werden jeweils Subtrainings- und Validierungsdatensätze gebildet, sodass im Beispiel mit einer gleichen Methode (z. B. Random Forest) fünf unterschiedliche Modelle entstehen. Die Unterschiede der prädizierten Parameter resultieren aus den jeweils variierenden Kombinationen der 64 Trainingsbeobachtungen und der entsprechenden Validierungsdatensätze. Die Beurteilung der Modellgüte erfolgt auf Grundlage geeigneter

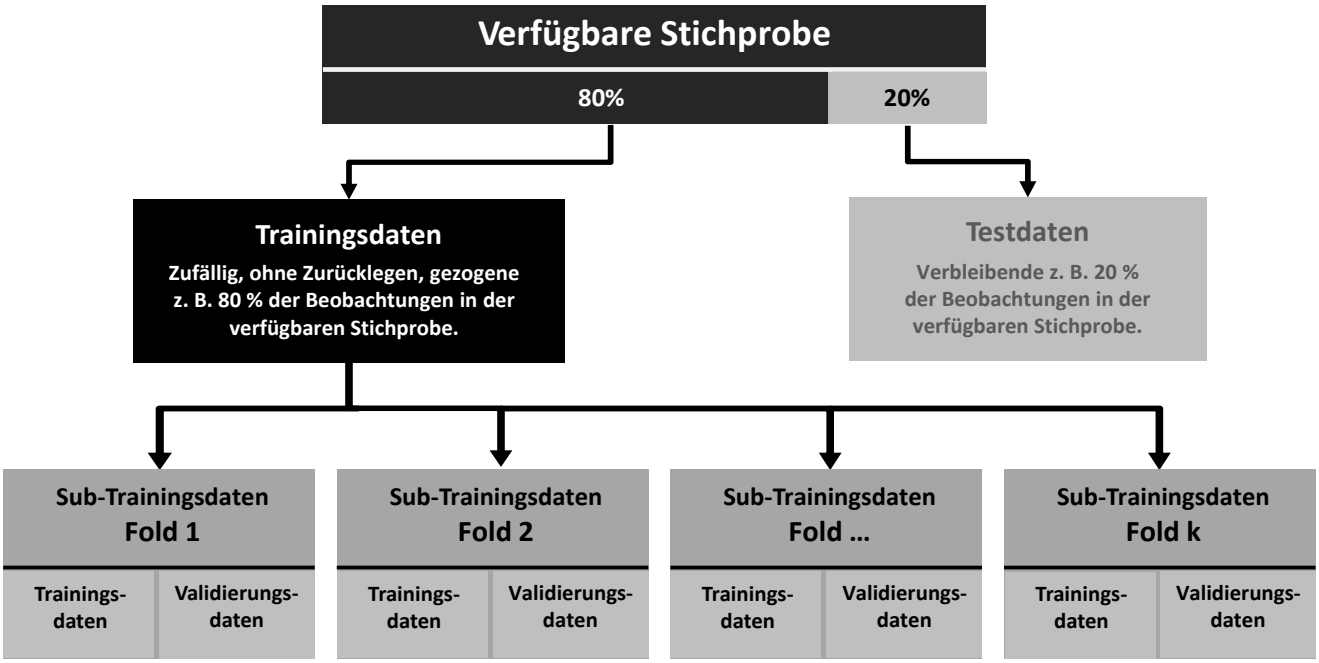

Abb. 3: Grundprinzip der Aufteilung einer verfügbaren Stichprobe

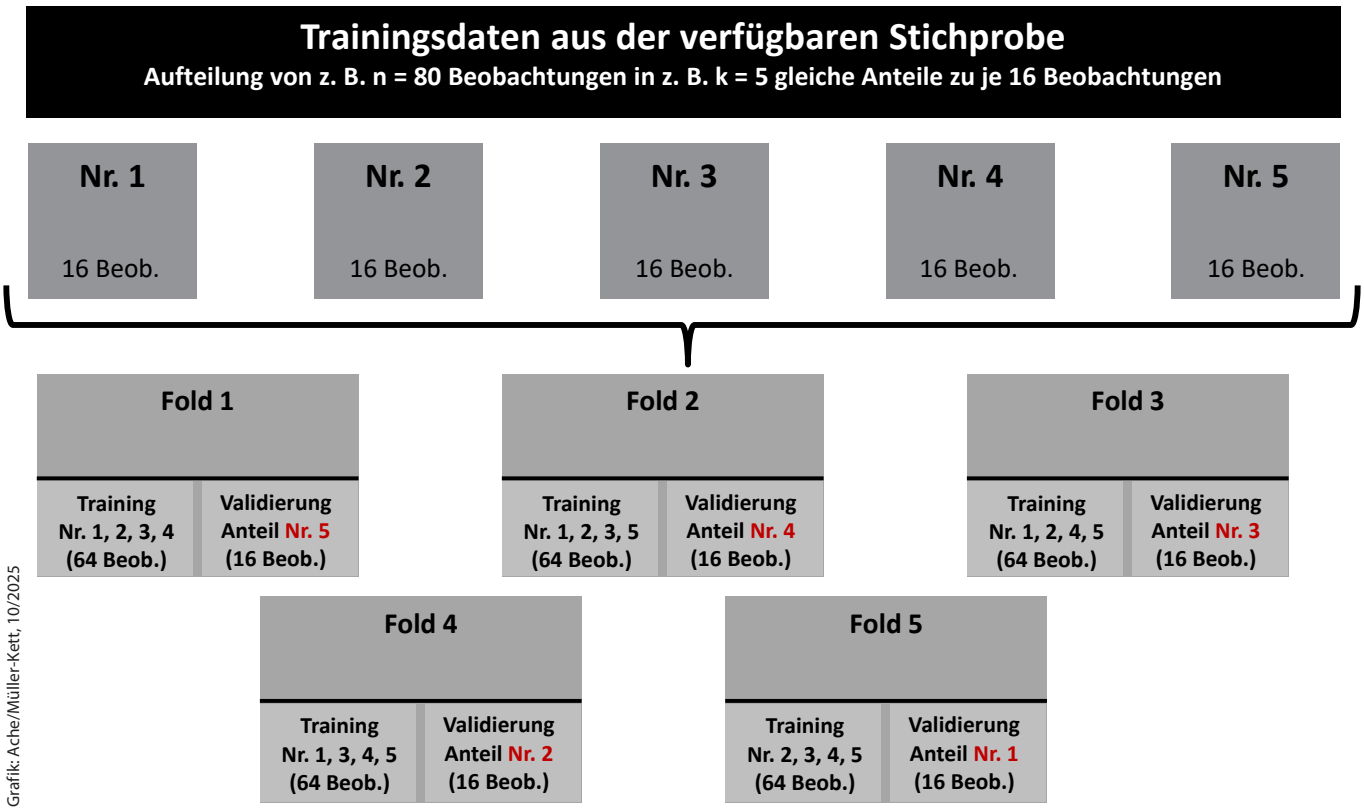

Abb. 4.: Grundprinzip der Bildung von Folds für die Kreuzvalidierung (Beispiel)

Metriken, die den Vergleich zwischen den tatsächlichen Werten und den auf den Validierungsdaten geschätzten Werten innerhalb der Folds abbilden.

Im Gegensatz zu einem naiven Ansatz, bei dem alle Daten für das Training verwendet würden, erlaubt die Aufteilung in Trainings- und Validierungsdaten innerhalb des eigentlichen Trainings eine realistische Einschätzung der Modellgüte für bislang unbekannte Beobachtungen. Dieser Grundsatz gilt methodenübergreifend, von der linearen Regression über Random Forests (z. B. Anzahl der Bäume, Baumtiefe) bis hin zu neuronalen Netzen (z. B. Anzahl verborgener Neuronen, Regularisierung, Lernrate) und Gradient Boosting (z. B. Iterationszahl, Lernrate, maximale Tiefe).

Zur Beurteilung also z. B. eines Trainingsdatensatzes von 80 Wohnflächenpreisen, bei denen ein Zusammenhang des Preises mit dem Baujahr der Objekte und der Lagequalität zur Prädiktion von Preisen genutzt werden soll, ergeben sich die mittleren Werte für Fehlermetriken (Ache 2025b) aus den fünf Folds z. B. wie in Tab. 1 dargestellt:

Ziel ist es, Hyperparameter, wie etwa die Anzahl und Tiefe der Bäume, so zu wählen, dass das Modell z. B. ein möglichst niedriges RMdSE und ein hohes R^2 erreicht. So lässt sich systematisch beurteilen, wie die Parameterauswahl die Vorhersagequalität beeinflusst und ob der gewählte Suchraum angemessen ist.

Die tatsächliche Verzerrung des finalen Modells wird anhand eines zuvor ungenutzten Testdatensatzes bewertet. Das trainierte Modell sagt beispielsweise Quadratmeterpreise für diese dem Modell unbekannten Kauffälle voraus.

Tab. 1: Beispielhafte Metriken zur Beurteilung der Modell-Verzerrung

Metrik	Abkürzung	Wert im Beispiel
Root Median Squared Error	RMdSE	± 342 €/m² (durchschnittliche Abweichung vom tatsächlichen Wert für die Wohnflächenpreise)
Root Mean Squared Error	RMSE	± 412 €/m² (durchschnittliche Abweichung vom tatsächlichen Wert für die Wohnflächenpreise)
Median Absolute Percentage Error	MdAPE	± 10 % (mittlere prozentuale Abweichung von den tatsächlichen Werten)
Mean Absolute Proportional Error	MAPE	± 8 % (mittlere absolute prozentuale Abweichung von den tatsächlichen Werten)
Bestimmtheitsmaß in %	R^2	81 % (erklärte Varianz des Modells für die Wohnflächenpreise)

Anschließend werden die Modellvorhersagen mit den tatsächlichen Preisen mittels Fehlermetriken wie RMdSE und R^2 verglichen. Ein niedriger Test-RMdSE und ein hohes R^2 deuten z. B. auf eine gute Generalisierungsfähigkeit hin. Dieses Vorgehen kann somit eingesetzt werden, um eine Überanpassung des Modells an die Trainingsdaten zu quantifizieren.

Das Prinzip der Kreuzvalidierung kann auch auf die Testphase ausgeweitet werden. Dabei wird der gesamte Prozess, von der Trennung der Trainings- und Testdaten bis hin zur *k*-fachen Aufteilung in Subtrainings- und Validierungsdaten, *p*-fach durchgeführt. So dient in jedem Durchlauf ein anderer Teil der verfügbaren Daten als Testdaten und wird dem Modell vorenthalten. Dieses Vorgehen wird als geschachtelte Kreuzvalidierung bezeichnet und die finalen Vorhersagen lassen sich für den gesamten Datensatz als Mittelwerte der einzelnen Iterationen bilden. Auf diese Weise wird eine wiederholte Fehlerberechnung vollzogen, deren Mittelwert ein Maß für die Modellgüte als Kombination aus Unverzerrtheit und Präzision darstellt. In diesem Fall ist die Rede von einem Fold-Fehler, für den sich auch die Streuung über alle Durchläufe berechnen lässt, was in gewisser Weise eine Einschätzung der Präzision erlaubt. Allerdings sollte beachtet werden, dass diese Fehlerstreuung die Variabilität der bereits gemittelten Fehler beschreibt und nicht direkt die Modellvariabilität. Alternativ können die Fehler über alle Durchläufe gesammelt und gemittelt werden, was als Global-Fehler bezeichnet wird und zur genauen und gezielten Einschätzung der Gesamtmodellgüte verwendet werden kann. Das so gewonnene Fehlermaß ähnelt damit in gewisser Weise dem Out-of-Bag-Fehler eines Random-Forest-Modells (s. unten und vgl. Hastie et al. 2009, S. 592). Eine Kreuzvalidierung kann also sowohl zur Einschätzung der Unverzerrtheit als auch zur Validierung der modellinternen Parameter und Hyperparameter genutzt werden. Während Ersteres häufig in der Wertermittlung vorrangig von Belang ist, ist Letzteres fundamentales Vorgehen des maschinellen Lernens. In der Methodik, Interpretation und Bedeutung für die Modellgüte bleibt das Verfahren der Kreuzvalidierung allerdings gleich, weshalb an dieser Stelle nicht nach Anwendungsgebiet getrennte Begriffe eingeführt werden sollten.

Ein Hauptnachteil dieser Dreiteilung der verfügbaren Daten besteht allerdings in der Tatsache, dass das Modell mit weniger Daten trainiert wird, als in der ursprünglichen Stichprobe zur Verfügung stehen, Informationen der Testdaten also zu einem gewissen Grad ungenutzt bleiben. Dieses Problem stellt sich als besonders herausfordernd für viele Anwendungen in der Wertermittlung dar, weil hier oftmals kleine Stichprobengrößen vorliegen. Dies kann dazu führen, dass bei der *k*-fachen Kreuzvalidierung die Metriken systematisch zu hoch ausfallen und die Verzerrung eines so trainierten Modells schlechter erscheint, als sie tatsächlich ist (Browne 2000, Yates et al. 2023).

Eine solche Überschätzung des Fehlers kann allerdings quantitativ erfasst und korrigiert werden und wird als Bias-Corrected-Kreuzvalidierung bezeichnet (s. Yates et al. 2023). Wird in jeder Iteration jeweils nur ein Datenpunkt dem Training zur Berechnung des Fehlers vorenthalten, ist die Rede von einer Leave-One-Out-Kreuzvalidierung, bei der die Überschätzung des berechneten Fehlers vernachlässigbar ist und keine Anpassung erfolgen muss.

Ein zweiter Nachteil der Kreuzvalidierung ist ein um den Faktor *k* vervielfältigtes Modelltraining, wobei sich bei einer geschachtelten Kreuzvalidierung die Anzahl der zu trainierenden Modelle weiter vervielfältigt. Angesichts langer und ressourcenintensiver Berechnungszeiten für komplexe Algorithmen kann eine Kreuzvalidierung an die Grenzen der praktischen Anwendung stoßen, insbesondere wenn die Berechnungen auf lokalen Maschinen erfolgen sollen. Einen entscheidenden Faktor stellt dabei die Einteilung des Datensatzes in *k* Anteile dar. Einerseits erhöht sich mit steigenden Werten die Berechnungszeit. Andererseits sollte dieser Wert so hoch wie möglich gewählt werden, um möglichst verlässliche Aussagen über die Streuung der berechneten Fehler zu erhalten und um die systematische Überschätzung dieses Fehlers zu minimieren. In der Praxis wird häufig eine Leave-One-Out-Kreuzvalidierung oder $k \geq 10$ empfohlen, sofern die Rechenressourcen dies erlauben. Andernfalls sollte eine Bias-Corrected-Kreuzvalidierung durchgeführt werden (Yates et al. 2023).

In diesem Zusammenhang seien Strategien zur Aufteilung der Daten in Trainings-, Validierungs- und Testdaten beleuchtet (Yates et al. 2023). Wird in jeder Iteration nicht ein anderer Anteil der Daten zur Validierung genutzt, sondern findet jeweils eine zufällige Auswahl statt, wird dies als Leave-d-Out-Kreuzvalidierung bezeichnet. Bei diesem Vorgehen kann es zu Überschneidungen der in den jeweiligen Iterationen ausgewählten Validierungsdaten kommen, was in gewisser Weise dem Bootstrapping-Verfahren ähnelt (s. unten). Sind bestimmte Gruppen von Daten von Belang, z. B. unterschiedliche Teilmärkte, sollte sichergestellt sein, dass jede dieser Gruppen angemessen in jeder Datenpartition vertreten ist, ein Vorgehen, welches als stratifizierte Kreuzvalidierung bekannt ist. Alternativ oder in Ergänzung sollten die Verteilungen der jeweiligen Datenpartitionen verglichen werden (z. B. Wahrscheinlichkeitsdichteverteilung der Validierungsdaten im Vergleich zu den Trainingsdaten), ein Vorgehen, das sich in der Praxis bewährt hat. Beziehen sich diese Gruppen nicht auf attributive Klassen, sondern auf geografische Gebiete, ist die Rede von einer räumlich blockierten Kreuzvalidierung. Jede dieser Kreuzvalidierungsstrategien lässt sich programmatisch in Analyseprozesse implementieren.

Ein in der Praxis zu berücksichtigender Aspekt der Kreuzvalidierung ist, dass dieses Verfahren primär zur realistischen Schätzung der Verzerrung eines Modells dient, nicht jedoch für inferenzstatistische Schlussfolgerungen auf Modellebenen eingesetzt werden sollte (siehe z. B. Efron und Tibshirani 1993). So ist es statistisch nicht zulässig, mittlere Werte von Regressionskoeffizienten aus den einzelnen Iterationen zu bilden, um daraus z. B. einen Umrechnungskoeffizienten abzuleiten. Der Grund liegt darin,

dass in jeder Iteration der Kreuzvalidierung auf einer anderen Teilmenge der Ausgangsdaten ein neuer Parametersatz geschätzt wird, der als Gesamtheit für diesen Durchgang zwar korrekt ist, allerdings nicht, wenn einzelne Koeffizienten extrahiert und mit anderen verrechnet werden. Diese Teilmengen werden für die jeweiligen Folds systematisch aufgeteilt, was dazu führt, dass eine einzelne Beobachtung per Definition nicht mit der gleichen Wahrscheinlichkeit Teil der Trainingsdaten ist. Dies führt dazu, dass die von Durchlauf zu Durchlauf gebildeten Koeffizienten systematisch und nicht zufällig variieren. Hier besteht ein grundlegender Unterschied zum Bootstrapping-Verfahren, welches »Resamples« generiert, in denen jeder Datenpunkt prinzipiell mit der gleichen Wahrscheinlichkeit vertreten ist und jede Bootstrap-Stichprobe als Alternativstichprobe der Grundgesamtheit interpretiert werden kann. Mittlere Werte z. B. der Regressionskoeffizienten über die Iterationen einer Kreuzvalidierung ergeben damit keinen konsistenten, inferenzstatistisch sinnvollen Schätzer für den zugrundeliegenden »wahren« Parameter in der Grundgesamtheit.

Ebenso kann die Kreuzvalidierung nicht zur Berechnung von Konfidenzintervallen in Bezug auf die Präzision eines Modells eingesetzt werden (mit Ausnahme der Streuungsmaße über die einzelnen Folds). Die Kreuzvalidierung bietet also keinen Rückhalt für Inferenz, da die verschiedenen Trainingsdaten jeweils strukturell variieren und nicht als unabhängige Größen aus einer Grundgesamtheit aufzufassen sind.

Sollen zur Einschätzung der Präzision von Ergebnissen Konfidenzintervalle für Modellparameter bestimmt werden, empfiehlt sich das Bootstrapping-Verfahren. Hier bleibt die Datenbasis strukturell stabil und die wiederholten Ziehungen und Modellierungen aus dem Gesamtdatensatz liefern fundierte Aussagen zur Schwankungsbreite und Vertrauenswürdigkeit einzelner Parameter.

4 Quantifizierung der Präzision durch Bootstrapping

Die bisherigen Ausführungen haben gezeigt, dass die Betrachtung der Modellverzerrung – etwa durch Kreuzvalidierung – zur Quantifizierung von Unsicherheiten als zentrale Anforderung den gestiegenen Ansprüchen an Transparenz und Nachvollziehbarkeit gerecht werden kann. Allerdings eignet sich dieses Verfahren nicht umfänglich für eine Bewertung der Modellgenauigkeit. Vor diesem Hintergrund rückt das Bootstrapping als Resampling-Verfahren in den Fokus, da es eine probabilistische Einschätzung der Präzision von Schätzwerten ermöglicht und somit die Aussagekraft datenbasierter Bewertungsmodelle erheblich stärkt.

4.1 Bootstrapping – Das Verfahren im Allgemeinen

Wie bereits dargestellt, umfasst das Konzept der Modellgüte die komplementären Dimensionen Verzerrungsfreiheit und Präzision in Bezug auf die wahren Werte der Grundgesamtheit eines Teilmarktes. Für die Immobilienwertermittlung bedeutet dies, dass alle Objekte eines relevanten Teilmarktes zur Grundgesamtheit zählen, unabhängig davon, ob sie in einer Kaufpreissammlung oder anderen Datenbanken erfasst sind. Dies bedeutet in der Praxis, dass – entgegen mitunter anderer Auffassungen – auch die bislang nicht veräußerten Immobilien als Teil der Grundgesamtheit zu betrachten sind und nicht nur diejenigen, die in z. B. den Kaufpreissammlungen der Gutachterausschüsse als Verkäufe registriert sind (Ache 2025a). Da die vollständige Erhebung aller relevanten Marktdaten praktisch unmöglich ist, werden Stichproben herangezogen, um Rückschlüsse auf die Grundgesamtheit zu ziehen. Der aus der Stichprobe berechnete mittlere Wert weicht dabei in der Regel vom theoretischen Erwartungswert und auch von dem wahren Wert ab.

Das Bootstrapping-Verfahren ermöglicht es, durch wiederholtes Ziehen von Stichproben mit Zurücklegen die Schwankungsbreite und Unsicherheit von Schätzparametern wie Mittelwerten oder Regressionskoeffizienten empirisch zu bestimmen. So lassen sich beispielsweise Vertrauensintervalle für Liegenschaftszinssätze oder andere Bewertungsparameter ableiten, die eine probabilistische Aussage über deren wahre Werte in der Grundgesamtheit erlauben.

Das Vertrauensintervall ist hier wie folgt zu interpretieren: Liegt z. B. ein 95 %-Vertrauensintervall für Liegenschaftszinssätze bei großen Mehrfamilienhäusern in einem Wertebereich zwischen 1,8 % und 2,5 %, dann kann von folgendem Verhalten der Liegenschaftszinssätze in der Grundgesamtheit aller veräußerbaren Mehrfamilienhäuser der Region XY ausgegangen werden:

Liegen 100 unterschiedliche, unverzerrte Stichproben der genannten Grundgesamtheit vor, so befinden sich die Erwartungswerte (z. B. Medianwerte) der Liegenschaftszinssätze in 95 Fällen im Bereich zwischen 1,8 % und 2,5 %. Das bedeutet, dass der Liegenschaftszinssatz eines beliebigen Objekts mit einer Wahrscheinlichkeit von 95 % in dieser Spanne liegt. Wird das Vertrauensintervall beispielsweise auf 65 % reduziert, liegt der Wert nur in 65 von 100 Fällen innerhalb der entsprechend engeren Spanne, während er in 35 Fällen außerhalb liegt. Eine solche Verengung suggeriert zwar eine höhere Genauigkeit, ist für die praktische Anwendung jedoch wenig hilfreich.

Erwartungswerte können z. B. der Medianwert oder auch die Regressionskoeffizienten von Regressionsfunktionen oder Ergebnisse aus anderen Methoden sein.

Es ist in diesem Zusammenhang darauf hinzuweisen, dass die Begriffe »Vertrauensintervall« und »Prognoseintervall« mitunter verwechselt werden. Das Vertrauensintervall beschreibt den Bereich, in dem der wahre Mittelwert mit einer bestimmten Wahrscheinlichkeit liegt, während das Prognoseintervall den Bereich angibt, in dem ein zukünftiger Einzelwert mit einer bestimmten Wahrscheinlichkeit zu erwarten ist. Für die Immobilienwertermittlung ist diese Unterscheidung essenziell, da das Prognoseintervall stets breiter ist und sowohl die Unsicher-

Tab. 2: Allgemeine Grundlage des Bootstrapping, Basisstichprobe

Basisstichprobe		
Wohnflächenpreise in Euro/m²		
	lfd. Nr.	Wohnflächenpreis
1	1	1.500
2	2	1.625
3	3	1.750
4	4	1.875
5	5	2.000
6	6	2.125
7	7	2.250
8	8	2.375
9..18		
19	19	3.750

heit des Mittelwerts als auch die Streuung der Einzelwerte berücksichtigt. Mit Hilfe des Bootstrap-Verfahrens können sowohl das Vertrauens- als auch das Prognoseintervall berechnet werden, auch bei unsymmetrischen Verteilungen.

Beim Bootstrapping entstehen durch Ziehen mit Zurücklegen aus einer Originalstichprobe – mit z. B. Wohnflächenpreisen von 1.500 €/m² bis 3.750 €/m²– neue Zahlenreihen, wobei jede Reihe ebenso viele Elemente enthält wie die Ausgangs- bzw. Basisstichprobe (Tab. 2).

Dabei treten einige Beobachtungen zwar mehrfach und andere gar nicht auf, dennoch ist dieses Verfahren dann besonders vorteilhaft, wenn eher kleine Stichproben vorliegen und wenn die Verteilung der Zielgröße in der Grundgesamtheit nicht bekannt oder unsymmetrisch ist, wie es bei Kaufpreissammlungen häufig der Fall ist. In Tab. 3 sind beispielhaft die Ergebnisse von fünf Ziehungen dargestellt.

Es ist erkennbar, dass z. B. in der Stichprobe »Bootstrap 2« der Wert 1.875 €/m² zweimal vorkommt, obwohl in der Basisstichprobe dieser Wert nur einmal anfällt. Um zu hinreichend aussagekräftigen Informationen über die Grundgesamtheit zu kommen, sollte eine erheblich höhere Zahl von solcherart simulierten Stichproben erzeugt werden, als in diesem Beispiel dargestellt.

Aus einem solchen Datensatz können dann Statistiken der jeweiligen simulierten Stichproben (engl. bootstraps), wie z. B. Medianwerte, arithmetische Mittelwerte oder Standardabweichungen, ermittelt werden (Tab. 4).

Aus den Statistiken der Vielzahl von Bootstrapping-Stichproben ergeben sich dann Schätzungen für die Intervalle der Parameter (z. B. Median, arithmetischer Mittelwert, Standardabweichung) in der Grundgesamtheit. In diesem einfach gefassten Beispiel kann also davon ausgegangen werden, dass die Grundgesamtheit die in Tab. 5 dargestellten Statistiken aufweist.

Tab. 3: Beispielhafte Darstellung von fünf durch Bootstrapping simulierte Stichproben

Bootstrapping mit Basisstichprobe					
5 Bootstraps mit 19 Stichprobenwerten					
	Bootstrap 1	Bootstrap 2	Bootstrap 3	Bootstrap 4	Bootstrap 5
1	2.750	1.875	3.250	3.375	3.250
2	2.375	2.625	3.500	1.500	2.750
3	1.625	1.875	2.250	3.125	1.875
4	2.375	2.750	2.250	1.500	3.375
5	1.500	2.250	3.625	1.875	2.625
6	3.750	3.625	2.625	2.000	2.000
7..18					
19	2.250	2.375	3.375	3.500	3.125

Tab. 4: Statistiken aus einzelnen Bootstrapping-Datensätzen

Bootstrapping, Wohnflächenpreise				
500 Bootstraps – Statistiken				
	id	Median	arithm. Mittel	Standardabweichung
1	Bootstrap001	3.000	2.855	818
2	Bootstrap002	3.000	2.816	611
3	Bootstrap003	2.750	2.691	733
4	Bootstrap004	2.625	2.697	657
5	Bootstrap005	2.750	2.678	693
6	Bootstrap006	2.500	2.513	785
7	Bootstrap007	2.625	2.711	689
8	Bootstrap008	2.625	2.579	670
9..499				
500	Bootstrap500	3.250	2.947	742

Tab. 5: Erwartungswerte und Konfidenzintervalle für Bootstrapping-Datensätzen

Statistiken der Wohnflächenpreise*	Erwartungswert	95 %-Konfidenzintervall
Medianwert	2.625 €/m²	2.125 – 3.125 €/m²
arithmetischer Mittelwert	2.618 €/m²	2.325 – 2.928 €/m²
Standardabweichung	680 €/m²	528 – 819 €/m²

* n = 500 simulierte Beobachtungen für die jeweiligen Statistiken

Der Median von 2.625 €/m² liegt innerhalb des 95 %-Konfidenzintervalls des arithmetischen Mittels (2.325–2.928 €/m²), und umgekehrt fällt der Mittelwert von 2.618 €/m² in das 95 %-Konfidenzintervall des Medians (2.125–3.125 €/m²). Damit wird gut erkennbar, dass sich Median und arithmetischer Mittelwert in dieser Grundgesamtheit nicht signifikant unterscheiden. Mittelwert und Median der ursprünglichen Stichprobe entsprechen den aus den Bootstrapping ermittelten Erwartungswerten.

Das Beispiel verdeutlicht einen für die Wertermittlung relevanten Vorteil des Bootstrappings. Gerade bei geringer Datenverfügbarkeit, wie sie in der Immobilienwertermittlung häufig vorkommt, eignet sich Bootstrapping gut, da es auch mit kleinen Stichproben zuverlässige Analysen ermöglicht (für Lösungsansätze siehe auch Horvath et al. 2021). Die Methode kann nicht nur zur Schätzung einfacher statistischer Parameter, sondern auch zur Evaluierung komplexer Modelle wie multipler räumlicher Regressionen, Random Forests oder Neuronaler Netze eingesetzt werden, indem simulierte Stichproben zur Beurteilung der Modellparameter, etwa der Regressionskoeffizienten, verwendet werden.

4.2 Bootstrapping bei der Random-Forest-Methode

Bei der Random-Forest-Methode oder anderen Machine-Learning-Ansätzen (s. Bishop 2006) gibt es keine Koeffizienten im Sinne von Regressionsanalysen. Hier werden stattdessen andere modellinterne Parameter geschätzt, welche das Verhalten des Modells spezifizieren. Ein Random Forest besteht z. B. aus einer Vielzahl einzelner Entscheidungsbäume, deren interne Struktur simpel und intuitiv ist (Breiman 2001; Hastie et al. 2009, Kap. 15). Diese gliedert sich in die Wurzel, verschiedene Entscheidungsknoten und schließlich die Blätter des Baumes. Die »Baumwurzel« stellt den Startpunkt für eine Modellvorhersage dar. Die Knoten sortieren Datenpunkte jeweils nach einer einfachen Regel in zwei Unterkategorien, in einem fiktiven Beispiel könnte die Wohnfläche weniger als 120 m² betragen oder nicht. Im nächsten Schritt könnte z. B. ein weiterer Knoten die Datenpunkte auf der Entscheidungsbasis Baujahr nach 1970 oder nicht einer von zwei Unterkategorien zuordnen. Als Datenpunkte (eng. samples) werden in diesem Zusammenhang die Beobachtungen bezeichnet, die in der Immobilienwertermittlung häufig Kauffälle darstellen, die in einem tabellarischen Datensatz als Zeilen modelliert werden. Die Blätter, denen die Datenpunkte über diese einfachen dichotomen Entscheidungen zugewiesen werden, stellen schließlich die Modellvorhersage dar.

Bestehen diese Vorhersagen aus qualitativen Variablen (also distinkten Klassen wie z. B. »einfache Lage«, »mittlere Lage«, »gute Lage«), ist die Rede von Entscheidungsbäumen, während sogenannte Regressionsbäume numerische Vorhersagen liefern (z. B. der vorhergesagte Verkehrswert). Im Folgenden werden beide Ansätze vereinfacht als »Entscheidungsbäume« (eng. decision trees) bezeichnet. Die Reihenfolge der Knotenlogik und deren Schwellenwerte (z. B. Wohnfläche < 120 m², Baujahr > 1970, etc.) verkörpern die modell-internen Parameter und deren Schätzung findet als Teil des Random-Forest-Algorithmus mit Hilfe von Bootstrapping statt.

Die ausgewählten Bootstrap-Datenpunkte werden als »In-Bag-Samples« bezeichnet, die übrigen, für diese Iteration nicht verwendeten Datenpunkte als »Out-of-Bag-Samples«, oder kurz »OOB-Samples«. Im Random-Forest-Algorithmus findet anschließend eine weitere Randomisierung, also eine zufällige Auswahl der Einflussgrößen, statt. Dabei werden zufällig Einflussgrößen für den zu trainierenden Entscheidungsbaum entfernt, z. B. die Wohnfläche oder das Baujahr, sodass ein niedrig-dimensionierter Datenraum entsteht. Die Methodik überschneidet sich an dieser Stelle mit weiterführenden Techniken zur Merkmalsauswahl und Merkmalswichtigkeit (eng. feature importance), die hier nicht im Detail diskutiert werden sollen. Durch diese doppelte Randomisierung (Datenpunkte und Merkmale) entsteht ein »Random Subspace« und, da diese Schritte für jeden zu trainierenden Entscheidungsbaum durchgeführt werden, eine Anzahl recht unterschiedlicher Entscheidungsbäume – ein sogenanntes baumbasiertes Ensemble-Modell – mit distinkten Parametern, eine genügend variable Stichprobe vorausgesetzt.

Dieses Vorgehen erlaubt auf zweifache Weise eine Einschätzung der Modellgüte im Sinne der Verzerrungsfreiheit und Präzision. Zum einen liefert ein Random-Forest-Modell zunächst nicht eine, sondern – basierend auf den einzelnen Entscheidungsbäumen – viele Vorhersagen. Die finale Vorhersage ergibt sich aus dem häufigsten Wert oder dem Mittelwert der individuellen Vorhersagen. Auf diese Weise kann die Streuung der individuellen Vorhersagen Aufschluss über die Präzision geben. Liegen z. B. die vorhergesagten Quadratmeterpreise der individuellen Entscheidungsbäume weit auseinander, äußert sich dies in einer hohen Vorhersagestreuung (z. B. einer 1,5-fachen Standardabweichung oder eines 95 %-Vertrauensintervalls) und zeigt an, dass das Modell zwar Vorhersagen liefert, diese allerdings nicht belastbar sind, also eine schlechte Präzision aufweisen. Umgekehrt lässt eine niedrige Vorhersagestreuung zwischen den individuellen Entscheidungsbäumen auf präzise Vorhersagen schließen.

Dies veranschaulicht ein einfaches Beispiel (Tab. 6). Es wurden exemplarisch zwei einfache Random-Forest-Modelle mit jeweils 500 Entscheidungsbäumen zur Vorhersage von Quadratmeterpreisen trainiert. Die jeweiligen Modellparameter sowie die Struktur der für beide Modelle gleichen Ausgangsdaten sind an dieser Stelle nicht von Belang. Vielmehr soll folgendes Phänomen veranschaulicht werden: Obwohl zwei Modelle exakt gleiche Vorhersagen liefern können, können diese Vorhersagen unterschiedlich präzise sein, wobei Bootstrapping zur Quantifizierung dieser Präzisionsunterschiede verwendet werden kann. Im betrachteten Beispiel ist der arithmetische Mittelwert für einen einzelnen unbekannten Kaufpreis über die 500 Baumvorhersagen bei beiden fiktiven Modellen

Tab. 6: Ermittlung des Konfidenzintervalls zur Schätzung der Präzision eines Random-Forest-Modells basierend auf der Vorhersagevariabilität der einzelnen Entscheidungsbäume

Konfidenzintervalle durch Bootstrapping in Random-Forest-Modellen		
500 Entscheidungsbäume pro Modell		
	Random-Forest-Modell 1	Random-Forest-Modell 2
Baum-id	mittlerer vorhergesagter Quadratmeterpreis in €	mittlerer vorhergesagter Quadratmeterpreis in €
1	4.919	2.431
2	1.720	2.783
3	1.222	2.587
4	3.324	2.374
5…499	…	…
500	4.569	2.629
arithmetisches Mittel	**2.630**	**2.630**
Median	2.506	2.618
Standardabweichung	644	175
1,5-fache Standardabweichung	966	262
untere Grenze des Konfidenzintervalls	1.664	2.368
obere Grenze des Konfidenzintervalls	3.596	2.892
1,5-fache Standardabweichung als Konfidenzintervall	**1.664 – 3.596**	**2.368 – 2.892**
0.25-Quantil	2.150	2.618
0.975-Quantil	4.182	2.639
95 %-Interquantilbereich als Konfidenzintervall	**2.150 – 4.182**	**2.618 – 2.639**

identisch (2.630 €/m²), während ein niedrigerer Medianwert (Modell 1: 2.506 €/m², Modell 2: 2.618 €/m²) auf – wie auf dem Immobilienmarkt üblich – eine rechtsschiefe Verteilung der Werte für die Variable »Quadratmeterpreis« hindeutet.

Das Konfidenzintervall lässt sich z. B. basierend auf der Standardabweichung der individuellen Entscheidungsbaumvorhersagen (Modell 1: 1.664–3.596 €/m², Modell 2: 2.368–2.892 €/m²) oder deren 95 %-Vertrauensintervall (Modell 1: 2.150–4.182 €/m², Modell 2: 2.618–2.639 €/m²) berechnen. Dieses Beispiel verdeutlicht erneut die Notwendigkeit, neben der reinen Vorhersage zusätzlich das Vertrauensintervall anzugeben. In diesem Beispiel ist das Modell 2 deutlich präziser als Modell 1.

Eine weitere Möglichkeit der Ermittlung von Vertrauensintervallen ist der Vergleich zwischen den Schätzwerten aus dem Modell ($\hat{y}$) für die OOB-Datenpunkte und den wahren Werten aus der Grundgesamtheit (y), ein Vergleich, der auch zur Schätzung der Verzerrung angestellt werden kann. Dieser Vergleich ist legitim, da dem Modell diese Datenpunkte zum Zeitpunkt des Trainings in der jeweiligen Iteration nicht bekannt waren, wodurch dieser Ansatz der Kreuzvalidierung ähnelt (s. oben).

Auch hier kann die Streuung dieser Fehler zur Berechnung des Konfidenzintervalls herangezogen werden, wie das folgende Beispiel veranschaulicht (Tab. 7). In diesem Beispiel wurden erneut zwei Random-Forest-Modelle mit jeweils 500 Entscheidungsbäumen trainiert und auch hier spielen die jeweiligen Modellspezifika keine Rolle für die folgende Betrachtung. Für jeden Baum standen 1.000 Kauffälle zur Verfügung, wovon einzelne Datenpunkte für die jeweiligen Entscheidungsbäume nicht in das Training eingeflossen sind, die sogenannten »OOB-Samples« (s. Ausführungen zur Bildung des »Random Subspaces« am Kapitelanfang). Für jeden dieser Kauffälle wurde von den jeweiligen Entscheidungsbäumen, in deren Training dieser Fall nicht eingeflossen ist, eine Modellvorhersage berechnet und anschließend der Median über alle Bäume gebildet (OOB-Vorhersage). Der OOB-Fehler ($|\varepsilon|$) ergibt sich aus der absoluten Differenz zwischen dem vorhergesagten und dem tatsächlichen Quadratmeterpreis.

Der Mehrwert einer Betrachtung dieser OOB-Fehler zeigt sich z. B. darin, dass beide Modelle die gleichen Vorhersagen liefern, wenn diese auf dem arithmetischen Mittel über alle Entscheidungsbäume basieren (In-Bag-Modellvorhersage). Eine Betrachtung der OOB-Fehler offenbart

Tab. 7: Ermittlung des Konfidenzintervalls zur Schätzung der Präzision eines Random-Forest-Modells basierend auf Out-of-Bag-Fehlern

Konfidenzintervalle durch Out-of-Bag-Fehler in Random-Forest-Modellen													
500 Entscheidungsbäume und 1000 Kauffälle													
		Random-Forest-Modell 1						Random-Forest-Modell 2					
	Sample-id →	1	2	3	4	5…999	1000	1	2	3	4	5…999	1000
Baum-id	bekannter Quadratmeterpreis in € (y)	**1.506**	**3.106**	**1.067**	**4.638**	**…**	**1.531**	**1.506**	**3.106**	**1.067**	**4.638**	**…**	**1.531**
1	vorhergesagter Quadratmeterpreis in € (ŷ)	2.515	*kein OOB*	*kein OOB*	3.114	…	1.113	*kein OOB*	3.019	*kein OOB*	4.628	…	2.046
2		*kein OOB*	*kein OOB*	1.288	4.730	…	*kein OOB*	1.564	*kein OOB*	*kein OOB*	4.575	…	1.448
3…499		…	…	…	…	…	…	…	…	…	…	…	…
500		*kein OOB*	3.005	*kein OOB*	5.941	…	*kein OOB*	1.504	*kein OOB*	*kein OOB*	4.603	…	1.507
In-Bag-Modellvorhersage		1.265	3.512	1.115	4.203	...	2.043	1.265	3.512	1.115	4.203	...	2.043
OOB-Vorhersage (Median)		**1.505**	**3.171**	**756**	**3.726**	**…**	**2.546**	**1.564**	**2.964**	**949**	**4.280**	**…**	**1.576**
OOB-Fehler (|ε|)		1	65	311	912	…	1.015	57	142	118	358	…	45
OOB-0.25-Quantil		327						123					
OOB-0.975-Quantil		1.030						607					
OOB-95 %-Interquantilbereich als Konfidenzintervall		**327 – 1.030**						**123 – 607**					

allerdings, dass Modell 2 präziser ist als Modell 1. So zeigt das 95 %-Konfidenzintervall eine deutlich niedrigere Fehler-Streuung für Modell 2 (123–607 €/m²), verglichen mit Modell 1 (327–1.030 €/m²). Zugleich zeigt Modell 2 weniger Verzerrung, was sich in niedrigeren mittleren Fehlern äußert.

In beiden Beispielen verdeutlicht sich die Notwendigkeit der berechneten Konfidenzintervalle zur Steigerung der Transparenz der verwendeten Wertermittlungsmodelle.

5 Fazit

Die Integration datenwissenschaftlicher Methoden in die Immobilienwertermittlung eröffnet neue Möglichkeiten, die Genauigkeit und Nachvollziehbarkeit von Bewertungsmodellen substanziell zu verbessern. Die konsequente Anwendung von Verfahren wie Kreuzvalidierung und Bootstrapping ermöglicht es, Unsicherheiten nicht nur zu erkennen, sondern auch quantitativ zu erfassen und transparent zu kommunizieren. Dies ist nicht nur ein methodischer Fortschritt, sondern entspricht auch den gestiegenen Anforderungen der ImmoWertV an die Dokumentation und Nachvollziehbarkeit von Bewertungsprozessen.

Die Modellgüte kann dabei in Unverzerrtheit und Präzision differenziert werden. In diesem Beitrag wurde exemplarisch verdeutlicht, wie eine Kreuzvalidierung zur Quantifizierung der Verzerrung dienen kann. Dazu werden dem Modell iterativ Daten vorenthalten, für die nach Abschluss des Modelltrainings Vorhersagen generiert werden. Eine Einschätzung der Modellverzerrung erfolgt dann aus den gemittelten Differenzen zwischen vorhergesagten und tatsächlichen Werten. Außerdem wurde veranschaulicht, dass die Kreuzvalidierung, insbesondere die geschachtelte Kreuzvalidierung, durch eine Quantifizierung der Fold-Variabilität in Ansätzen zur Einschätzung der Präzision dienen kann.

Stehen weniger die Modellvorhersagen als vielmehr ein durch das Modell ausgedrücktes Wirkungsprinzip im Zentrum einer Analyse (wenn z. B. Umrechnungskoeffizienten berechnet werden), kann stattdessen das Bootstrapping-Verfahren eingesetzt werden. Wie dargelegt wurde, eignet sich dieses Verfahren für derart inferenzstatistische Belange. Weiterhin kann das Bootstrapping primär zur Quantifizierung der Präzision (z. B. Variabilität über einzelne »Resamples«) genutzt werden. Allerdings lässt auch das Bootstrapping-Verfahren in Ansätzen eine Einschätzung der Verzerrung zu, wenn z. B. die »Out-of-Bag«-Fehler berechnet werden.

Folglich lässt sich die Unverzerrtheit primär (aber nicht ausschließlich) durch Kreuzvalidierung und die Präzision primär aber ebenfalls nicht ausschließlich durch Bootstrapping-Methoden quantifizieren. Diese Einschätzung der Modellgüte ist dabei nicht bloß eine akademische Differenzierung, sondern hat unmittelbare praktische Relevanz: Sie erlaubt es, die Aussagekraft und die Grenzen von Modellergebnissen klar zu benennen und so die Risikokommunikation gegenüber Auftraggebern, Gerichten und anderen Stakeholdern zu verbessern. Gerade in einem zunehmend datengetriebenen Umfeld wird die Fähigkeit, Unsicherheiten explizit auszuweisen, zu einem Qualitätsmerkmal der Wertermittlung.

Gleichzeitig zeigt sich, dass die Quantifizierung der Modellgüte nicht als Selbstzweck verstanden werden darf. Vielmehr ist ihre Wirksamkeit an den spezifischen Anforderungen der Wertermittlung zu messen – insbesondere an der Transparenz, Nachvollziehbarkeit und rechtlichen Absicherung der Ergebnisse. Nur wenn diese Aspekte in der praktischen Umsetzung berücksichtigt werden, kann der Mehrwert datenbasierter Ansätze voll ausgeschöpft werden.

Insgesamt bietet die Verbindung klassischer Bewertungsprinzipien mit modernen Methoden zur Einschätzung der Modellgüte die Chance, die Wertermittlung auf ein neues Qualitätsniveau zu heben – vorausgesetzt, die Methoden werden kritisch reflektiert, sachgerecht eingesetzt und ihre Ergebnisse verantwortungsvoll kommuniziert.

Literatur

Ache, P. (2025a): Transparenz und Güte der Ergebnisse von Wertermittlungen – Teil 1: Grundüberlegungen für eine moderne Wertermittlung. In: zfv – Zeitschrift für Geodäsie, Geoinformation und Landmanagement, Heft 2/2025, 150. Jg., 179–186. DOI: 10.12902/zfv-0503-2024.

Ache, P. (2025b): Transparenz und Güte der Ergebnisse von Wertermittlungen – Teil 2: Modellbildung und Immobilienwertermittlung. In: zfv – Zeitschrift für Geodäsie, Geoinformation und Landmanagement, Heft 4/2025, 150. Jg., 252–261. DOI: 10.12902/zfv-0511-2025.

Bishop, C. M. (2006): Pattern Recognition and Machine Learning. Springer.

Breiman, L. (2001): Random forests. In: Machine Learning, Vol. 45, 5–32. DOI: 10.1023/A:1010933404324, letzter Zugriff 10/2025.

Browne, M. W. (2000): Cross-Validation Methods. In: Journal of Mathematical Psychology, Vol. 44, Iss. 1, 108–132. DOI: 10.1006/jmps.1999.1279, letzter Zugriff 10/2025.

Carsey, T. M., Harden, J. J. (2013): Monte Carlo Simulation and Resampling Methods for Social Science. Sage Publications.

Dimopoulos, T., Renigier-Biłozor, M., Ache, P., Janowski, A. (2024): The Future of Real Estate Valuation: Connecting Automated Valuation Models (AVMs) and Artificial Intelligence (AI) for Ethical and Scalable Solutions. In: Brazil´s Economic Perspectives and Progress – Valuation 20 Conference Proceedings 2024, 77–90. https://journals.aarvf.org/_files/ugd/9b866a_1e07b9ea06a2433797a672b5bd5f8ed8.pdf, letzter Zugriff 10/2025.

Dorey, F. J. (2011): In Brief: Statistics in Brief: Statistical Power: What Is It and When Should It Be Used? In: Clinical Orthopaedics and Related Research, Vol. 469, Iss. 2, 619–620. DOI: 10.1007/s11999-010-1435-0.

Efron, B., Tibshirani, R. (1986): Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. In: Statistical Science, Vol. 1, No. 1, 54–75. DOI: 10.1214/ss/1177013815, letzter Zugriff 10/2025.

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

Fisher, A., Rudin, C., Dominici, F. (2019): All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. In: Journal of Machine Learning Research, 20 (177), 1–81.

Fisher, R. A. (1935): The Design of Experiments. Hafner Press.

Guyon, I., Elisseeff, A. (2003): An Introduction to Variable and Feature Selection. In: Journal of Machine Learning Research, Vol. 3, 1157–1182.

Haack, B. (2008): Sensitivitätsanalyse zur Verkehrswertermittlung von Grundstücken. Dissertation im Fachbereich Geowissenschaften/Geographie – Wirtschaftsgeographie der Rheinischen Friedrich-Wilhelms-Universität Bonn, GRIN Verlag.

Hastie, T., Tibshirani, R., Friedman, J. (2009): The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd Edition, Springer.

Horvath, S., Soot, M., Zaddach, S., Neuner, H., Weitkamp, A. (2021): Deriving adequate sample sizes for ANN-based modelling of real estate valuation tasks by complexity analysis. In: Land Use Policy, Vol. 107, 105475. DOI: 10.1016/j.landusepol.2021.105475, letzter Zugriff 10/2025.

James, G., Witten, D., Hastie, T., Tibshirani, R. (2013): An Introduction to Statistical Learning with Applications in R. Springer.

Long, H. (2017): Monte-Carlo-Simulation, Immobilienbewertung. Masterarbeit, Hochschule Anhalt. https://opendata.uni-halle.de/bitstream/1981185920/12369/1/Masterarbeit%20Haochen%20Long.pdf, letzter Zugriff 10/2025.

Neumann, I. (2009): Zur Modellierung eines erweiterten Unsicherheitshaushaltes in Parameterschätzung und Hypothesentests. Dissertation. Wissenschaftliche Arbeiten der Fachrichtung Geodäsie und Geoinformatik der Leibniz Universität Hannover, Nr. 277, zugleich Deutsche Geodätische Kommission bei der Bayerischen Akademie der Wissenschaften, Reihe C, Nr. 634. https://dgk.badw.de/fileadmin/user_upload/Files/DGK/docs/c-634.pdf, letzter Zugriff 10/2025.

Serdar, C. C., Cihan, M., Yücel, D., Serdar, M. A. (2021): Sample size, power and effect size revisited: simplified and practical approaches in pre-clinical, clinical and laboratory studies. In: Biochemia Medica 2021, Vol. 31, Iss. 1, 1–27. DOI: 10.11613/BM.2021.010502, letzter Zugriff 10/2025.

Tukey, J. W. (1958): Bias and Confidence in Not-Quite Large Sample. In: Annals of Mathematical Statistics, Vol. 29, Iss. 2, 614–623.

Yates, L. A., Aandahl, Z., Richards, S. A., Brook, B. W. (2023). Cross validation for model selection: A review with examples from ecology. In: Ecological Monographs, Vol. 93, e1557. DOI: 10.1002/ecm.1557, letzter Zugriff 10/2025.

Kontakt

Dipl.-Ing. Peter Ache
FIG – International Federation of Surveyors
Vorsitzender der FIG-Commission 9 »Valuation and the Management of Real Estate«
peter.ache.fig@achemail.de

Prof. Dr. Christian Müller-Kett
IU Internationale Hochschule
Professor für Data Science
christian.mueller-kett@iu.org

Dieser Beitrag ist auch digital verfügbar unter www.geodaesie.info.

