Transparenz und Güte der Ergebnisse von Wertermittlungen – Teil 2: Modellbildung und Immobilienwertermittlung

Transparency and Quality of Valuation Results – Part 2: Modelbuilding and Real Estate Valuation

Peter Ache

Zusammenfassung

Mit dem zunehmenden Einsatz KI-gestützter Methoden gewinnt die Transparenz und Qualität von Immobilienbewertungen an Bedeutung. Insbesondere bei machine learning-Verfahren ist es entscheidend, die zugrunde liegenden Modelle sowie die Genauigkeit der Ergebnisse sorgfältig zu validieren. Der Beitrag beleuchtet konkret die Kernbegriffe der Modell-Performance und Ergebnisgenauigkeit. Während Modell-Performance die Fähigkeit eines Modells beschreibt, konsistente und unverzerrte Ergebnisse zu liefern, misst die Ergebnisgenauigkeit die Präzision der Schätzwerte. Eine klare Trennung dieser Begriffe und ihre verständliche Darstellung ist unerlässlich, da sie entscheidend für die Akzeptanz der Ergebnisse moderner Bewertungsverfahren in einer zunehmend KI-dominierten Ära sind. Die Kernbotschaft ist, dass auch auf dem Sektor der Immobilienwertermittlung die guten Grundsätze der Geodäsie, nämlich der Darstellung von Zuverlässigkeit und Genauigkeit von Messungen, wieder mehr Beachtung finden müssen, als es bislang der Fall ist.

Vor dem Hintergrund der aktuellen Entwicklungen muss zudem mittelfristig davon ausgegangen werden, dass klassische Wertermittlungs- und Analysemethoden an ihre Grenzen stoßen. Auch deshalb muss sich die moderne Immobilienbewertung verstärkt mit Methoden wie beispielsweise Random-Forest und neuronalen Netzen auseinandersetzen. Modelle werden trainiert und getestet, um Überanpassung zu vermeiden und durch Konfidenzintervalle die Genauigkeit der Vorhersagen abzusichern. In diesem Zusammenhang ermöglicht das Bootstrap-Verfahren – wie im Artikel erläutert – belastbare Schätzungen, selbst bei nicht normalverteilten Daten.

Schlüsselwörter: Immobilienwertermittlung, Ergebnisgenauigkeit, Modell-Performance, Künstliche Intelligenz, Bootstrap-Verfahren, Transparenz, Immobilienmarkt

Summary

With the increasing digitization and the use of AI-supported methods, transparency and quality in real estate valuation are becoming increasingly important. In particular, for machine-learning approaches, it is crucial to carefully validate the underlying models and the accuracy of results. This article examines concrete key concepts for assessing model performance and result accuracy. While model performance describes a model's ability to produce consistent and unbiased results, result accuracy measures the precision of estimated values. A clear distinction between these terms and their comprehensible presentation is essential, as they play a decisive role in the acceptance of modern valuation methods in an increasingly AI-driven era. The key message of the article is that, even in the field of real estate valuation, the fundamental principles of geodesy – namely the representation of reliability and accuracy in measurements – must receive greater attention than has been the case so far.

Given current challenges, it must be assumed that traditional valuation and analysis methods will soon reach their limits. Consequently, modern real estate valuation must increasingly engage with methods such as e.g. Random Forest and Neural Networks. Models are trained and tested to prevent overfitting and to ensure prediction accuracy through confidence intervals. In this context, the Bootstrap method – as explained in this article – enables robust estimations, even when data are not normally distributed.

Keywords: *real estate valuation, accuracy of results, model performance, artificial intelligence, bootstrap method, transparency, real estate market*

1 Einleitung

Die Bedeutung von Immobilien für die wirtschaftliche, ökologische und soziale Entwicklung wird häufig unterschätzt, obwohl sie eine zentrale Rolle für das Bruttoinlandsprodukt und den sozialen Frieden spielen. Ein unzureichendes Wohnraumangebot führt zu steigenden Preisen und Mieten, was soziale Ungleichheiten und Unsicherheiten verstärkt. Hinzu kommt, dass zukünftig ein erheblicher Fachkräftemangel im Immobiliensektor zu erwarten ist, und von einem Wandel der erforderlichen Kompetenzen in Bereichen wie Wertermittlung, Projektentwicklung, Handel, Finanzierung, Besteuerung und Management begleitet sein wird. Parallel dazu nehmen Digitalisierung und der Einsatz von KI deutlich zu, deren ethische Implikationen berücksichtigt werden müssen (UNESCO 2021).

Gleichzeitig eröffnen maschinelles Lernen und automatisierte Bewertungsmodelle neue Möglichkeiten, präzisere und transparentere Ergebnisse zu erzielen. Diese Ansätze lösen sich von traditionellen Methoden wie der linearen

  DOI 10.12902/zfv-0511-2025

Regressionsanalyse und den ökonomisch geprägten, theoretischen Verfahren der Wertermittlung. Besonders bei Massenbewertungen, etwa für die Besteuerung, die Beleihung oder die Bewertung von Klimaschutzmaßnahmen, stoßen klassische Methoden zunehmend an ihre Grenzen.

Der Beitrag baut auf dem ersten Teil der Reihe »Transparenz und Güte der Ergebnisse von Wertermittlungen« (Ache 2025) auf und widmet sich der vertiefenden Definition und Abgrenzung der Begriffe »Modell-Performance« und »Ergebnisgenauigkeit«. Neben einer Erläuterung der theoretischen Grundlagen und praktischen Anwendungen wird die spezifische Bedeutung statistischer Intervalle zur Darstellung der Ergebnisgenauigkeit in der Immobilienbewertung diskutiert.

Das Ziel ist es, ein vertieftes Verständnis für diese Konzepte zu schaffen und ihre zentrale Rolle in modernen Bewertungsverfahren zu verdeutlichen. Ein zukünftiger Beitrag der Reihe wird den Schwerpunkt auf Resampling-Techniken legen, insbesondere das Bootstrap-Verfahren und die Monte-Carlo-Simulation, um deren Potenzial zur Steigerung der Modellgüte und Ergebnisgenauigkeit exemplarisch darzustellen.

2 Modell-Performance und Ergebnisgenauigkeit

Mit Bezug auf den ersten Teil der Reihe sollen die Begriffe »Modell-Performance« und »Ergebnisgenauigkeit« in diesem Beitrag detaillierter untersucht und in den Gesamtzusammenhang des Analyseprozesses gebracht werden. Die Beziehung von Modell-Performance und Ergebnisgenauigkeit lässt sich mit dem Maßstab einer Karte vergleichen:

Eine Karte im Maßstab 1:1.000.000 bildet die Realität unverzerrt und korrekt ab, doch die Genauigkeit der Messungen ist begrenzt. So entspricht eine Messgenauigkeit von ±0,1 mm auf der Karte einem realen Wert von ±100 m. Bei einer Karte im Maßstab 1:100 beträgt die reale Messgenauigkeit dagegen ±1 cm. Die Modell-Performance beschreibt also die Fähigkeit eines Modells, unverzerrte Ergebnisse zu liefern, die auch bei variierenden Stichproben konsistent die Realität widerspiegeln. Ein Modell gilt als performant, wenn es systematisch und unabhängig von äußeren Einflüssen richtige Resultate erzeugt. Die Modell-Performance kann also mit einer unverzerrten und die Wirklichkeit widerspiegelnden Karte verglichen werden, unabhängig von ihrem Maßstab.

Die Ergebnisgenauigkeit bezieht sich demgegenüber nicht auf die grundsätzliche Korrektheit oder Performance des Modells (in diesem Fall der »Karte«), sondern vielmehr auf den Maßstab, der die Präzision der Vorhersagen bestimmt. Sie beschreibt die Präzision der von einem Modell gelieferten Ergebnisse in Bezug auf die realen Werte und gibt an, mit welcher Sicherheit die Schätzungen innerhalb eines definierten Konfidenzintervalls liegen. Dabei spiegelt die Ergebnisgenauigkeit nicht die grundsätzliche Unverzerrtheit des Modells wider, sondern dessen Fähigkeit, Vorhersagen mit einer spezifischen Präzision zu treffen, die den Anforderungen der jeweiligen Anwendung entspricht.

Es wird zunehmend bedeutsam, transparente Metriken zur Bewertung dieser Genauigkeit zu entwickeln, wie es in evidenzbasierten Ansätzen der modernen Immobilienwertermittlung gefordert wird. Dabei steht die Modell-Performance im Kontext einer zunehmenden Digitalisierung und der steigenden Bedeutung datengetriebener Verfahren, wie sie in Bezug auf die aktuellen Entwicklungen der Immobilienbewertung hervorgehoben werden. Deshalb ist es entscheidend, die Begriffe Modell-Performance und Ergebnisgenauigkeit klar zu trennen, um die Validierung und Beurteilung von Modellen fundiert und transparent zu gestalten. Eine getrennte Betrachtung dieser Aspekte, insbesondere vor dem Hintergrund der Anforderungen an Transparenz und Modellgüte, wie sie von Ache (2025) betont werden, erleichtert nicht nur die Weiterentwicklung, sondern auch die Vergleichbarkeit von Modellen. Zusammengefasst können die Begriffe wie folgt ausdifferenziert werden:

Modell-Performance:

- **Erklärbarkeit** ist die Fähigkeit eines Modells, seine Entscheidungsprozesse nachvollziehbar und verständlich darzustellen (Datenbasis, fehlende Werte, Effekte der Einflussvariablen etc.).
- **Robustheit** ist die Fähigkeit eines Modells, auf andere Daten des gleichen Teilmarktes immer gleich zu reagieren.
- **Fairness** ist die Unverzerrtheit (Unvoreingenommenheit) des Modells im Hinblick auf die starke Reduktion von systematischen Verzerrungen, die dazu führen, dass z. B. bestimmte Gruppen von Immobilien, wie etwa diejenigen in sehr schlechten Lagen, unter- oder überschätzt werden (Clemmensen und Kjærsgaard 2023).

Ergebnisgenauigkeit:

- **Genauigkeit** misst die durchschnittliche Abweichung zwischen den vorhergesagten und den tatsächlichen Werten und gibt an, wie nah die Schätzwerte voraussichtlich an der Wirklichkeit liegen (ähnlich auch Schwieger und Zhang 2019, S. 14 f.).
- **Präzision** zeigt den Anteil der Vorhersagen, die innerhalb einer von vorneherein festgelegten Grenze von den tatsächlichen Werten abweichen. Sie lässt eine Aussage darüber zu, wie konsistent ein Modell bestimmte Werte innerhalb einer Fehlermarge trifft.
- **Vertrauensintervall des Ergebnisses** qualifiziert die Unsicherheit einer konkreten Schätzung mit einem üblicherweise verwendeten Vertrauensbereich von 95 %, der erkennbar macht, dass der tatsächliche Wert in 95 von 100 Fällen in einer bestimmten Spannweite liegen wird (Brydges 2019).

Die zur Beurteilung von Performance und Ergebnisgenauigkeit nützlichen Metriken werden weiter unten dargestellt.

3 Immobilienwertermittlung und Modellbildung

Der Modellbildungsprozess in der Immobilienwertermittlung bildet das methodische Fundament für die Entwicklung präziser, transparenter und nachvollziehbarer Wertermittlungen sowie für die Erhebung und Analyse der hierfür erforderlichen Daten. Er umfasst eine klar strukturierte Abfolge von Schritten, die mit der Datenerhebung und -aufbereitung beginnen, über die Auswahl und Anwendung geeigneter statistischer Methoden verlaufen und in der Validierung, Interpretation und Darstellung der Ergebnisse münden, einschließlich der sachverständigen Entscheidung über den finalen Wert und ggf. der Bewertung seiner Genauigkeit.

Eine zentrale Bedeutung kommt der Qualität und Repräsentativität der verwendeten Daten zu, da Verzerrungen oder unzureichende Datengrundlagen die Performance der Modelle und die Genauigkeit der Ergebnisse erheblich beeinträchtigen können. Der Modellbildungsprozess ist dabei dynamisch und iterativ angelegt, wodurch eine kontinuierliche Anpassung und Optimierung der Modelle durch regelmäßiges Training und die dabei erfolgende Integration neuer Daten ermöglicht wird.

Im Gegensatz zu traditionellen Bewertungsverfahren, die oft auf subjektiven Einschätzungen der in die Methode eingehenden Daten durch Experten beruhen, zeichnet sich der moderne Modellbildungsprozess durch Transparenz, Reproduzierbarkeit und evidenzbasierte Methoden aus. Ziel ist es, statistisch basierte Modelle zu entwickeln, die nicht nur präzise Vorhersagen liefern, sondern auch unter variierenden Marktbedingungen und bei unterschiedlichen Stichproben robust und zuverlässig bleiben. Dabei kommt es darauf an, die Performance des Modells und die Genauigkeit der Ergebnisse so zu optimieren, dass die Wirklichkeit, entsprechend der jeweiligen Anforderung an die Bewertungsaufgabe, so genau wie nötig abgebildet wird.

Der Modellbildungsprozess inklusive der Entscheidung über einen zu erwartenden Wert umfasst dem Grunde nach acht Schritte (siehe Kasten).

MODELLBILDUNG IN 8 SCHRITTEN

1. Sammlung von Daten über Immobilien und von Preisen zu veräußerten Immobilien (Rohdaten),
2. Überprüfung der zu verwendenden Daten auf Repräsentativität sowie Verstehen der Grundgesamtheit, fehlender Werte, von Zusammenhängen und Verteilungen (Datenexploration),
3. Entscheidung für zu verwendende Methoden und die entsprechende Vorverarbeitung der Daten (Preprocessing),
4. Trainieren von Modellen mit einer oder mehreren Methoden,
5. Validierung der Performance der Modelle,
6. Ermittlung der Genauigkeit der Ergebnisse aus den Modellen,
7. Interpretation der Modelle sowie weitere Iterationen der Schritte 4 bis 6 und
8. Entscheidung des Sachverständigen über den Verkehrswert oder ein anderes für die Wertermittlung erforderliches Datum, wie z. B. Bodenrichtwerte, Liegenschaftszinssätze oder Umrechnungskoeffizienten, unter Einbeziehung der Analyseergebnisse und der persönlichen Expertise.

4 Rohdaten, Kaufpreissammlungen der Gutachterausschüsse

Die Erstellung und Führung von Datenbanken zu Immobilien stellt angesichts der Komplexität des Immobilienmarktes eine besondere und kontinuierlich wachsende Herausforderung dar. Die Qualität dieser Datensammlungen bildet die Grundlage für die Zuverlässigkeit und Präzision von Immobilienwertermittlungen. Ein treffender Vergleich hierzu ist der bekannte Ausdruck aus den Anfängen der Computertechnologie der 1950er und 1960er Jahre: »Garbage In, Garbage Out«. Für die Immobilienwertermittlung und -marktbeobachtung bedeutet dies, dass ungenaue, lückenhafte oder verzerrte Daten zwangsläufig zu fehlerhaften und unpräzisen Ergebnissen führen.

Besonders problematisch ist dies vor dem Hintergrund, dass nur ein Bruchteil der Bestandsimmobilien tatsächlich auf dem Markt gehandelt wird. Dadurch steht lediglich eine eingeschränkte Stichprobe der Grundgesamtheit zur Verfügung, was die Relevanz hochwertiger und repräsentativer Datensammlungen für eine evidenzbasierte Wertermittlung im modernen Zeitalter unterstreicht. Ziel ist es daher, möglichst umfangreiche und qualitativ hochwertige Daten zu Kaufpreisen und den jeweiligen Entstehungsumständen zu sammeln und einer breiten Anwendung zugänglich zu machen.

Gemäß § 193 Abs. 5 BauGB obliegt den Gutachterausschüssen diese Verpflichtung. Sie haben eine entsprechend geeignete Kaufpreisammlung zu führen. Ziel dieser Sammlung ist es, eine verlässliche Datenbasis für die Ermittlung immobilienbezogener Werte bereitzustellen. Die Einführung der Kaufpreissammlung geht auf § 143 des Bundesbaugesetzes (BBauG vom 23. Juni 1960, BGBl. I S. 341) zurück, wodurch Kaufpreise bereits seit den frühen 1960er Jahren (in den alten Bundesländern) systematisch erfasst werden.

Im Kontext datengetriebener Modelle und der Entwicklung KI-gestützter Methoden zur Modellbildung stellt die Kaufpreissammlung eine unverzichtbare Grundlage dar. Für die Führung und Qualität dieser Sammlungen sind gemäß § 199 Abs. 2 Nr. 4 BauGB die Länder verantwortlich. Sie sind ermächtigt, durch Rechtsverordnungen die Führung und Auswertung der Kaufpreissammlungen zu regeln. Die Umsetzung unterliegt nach den sogenannten »Gutachterausschussverordnungen« (§ 199 Abs. 2 BauGB) der Rechtsaufsicht der zuständigen Landesministerien, die

für die Einhaltung der Grundsätze zur Qualität und Nutzbarkeit der Daten verantwortlich sind.

Die operative Umsetzung obliegt den Gutachterausschüssen als Behörden der Länder. Nach § 192 Abs. 4 BauGB wird diese Aufgabe üblicherweise von den Geschäftsstellen der Gutachterausschüsse ausgeführt. Diese sind in der Regel bei bereits existierenden Behörden der Länder oder Kommunen, zumeist den Vermessungs- und Katasterbehörden, angesiedelt. Dadurch sind die Gutachterausschüsse ein wesentlicher Kern zur Sicherstellung einer konsistenten und qualitativ hochwertigen Datenerfassung als Basis für alle weiteren Schritte zur Verbesserung der Transparenz auf dem Immobilienmarkt, zur Datenbereitstellung, Marktbeobachtung und Ermittlung von Immobilienwerten. Vor dem Hintergrund der aktuellen und zukünftigen Herausforderungen ist hier an vielen Stellen derzeit durchaus Anpassungs- und Modernisierungsbedarf zu konstatieren.

5 Repräsentativität der Stichprobe und Grundgesamtheit

Bei der Datenanalyse im Kontext der Immobilienwertermittlung kommt es in der Regel darauf an, aus einer Stichprobe Informationen über das Verhalten der Grundgesamtheit zu bekommen. Essenziell ist damit, dass die Stichprobe die Grundgesamtheit, der Aufgabe entsprechend, hinreichend repräsentiert. Der Datensatz z. B. aus einer »Auskunft aus der Kaufpreissammlung« ist dahingehend zu überprüfen, ob jede Immobilie einer Grundgesamtheit – unabhängig davon, ob in der Kaufpreissammlung als Kauffall registriert oder nicht – die gleiche Chance gehabt hätte, als Verkauf in diese Kaufpreissammlung zu gelangen oder nicht. Die Stichprobe muss sowohl den räumlichen als auch den sachlichen Teilmarkt hinreichend repräsentieren. Die Verteilung der Kaufpreise aus der Stichprobe muss also sowohl über den geografischen Raum (räumlicher Teilmarkt) als auch über die Charakteristika der Immobilien (sachlicher Teilmarkt) aus der für die Wertermittlungsaufgabe interessierenden Grundgesamtheit unverzerrt sein. Dabei kommt es darauf an, die Grundgesamtheit für den betreffenden sachlichen und räumlichen Teilmarkt sorgfältig und klar zu definieren.

Der Begriff »repräsentative Stichprobe« ist ein sehr vielschichtiger Begriff, der, ebenso wie der Begriff »Transparenz«, gerne und manchmal sogar leichtfertig verwendet wird, jedoch nicht einheitlich definiert ist. In der Literatur ist eine allgemein akzeptierte Definition dessen, was eine repräsentative Stichprobe ausmacht und wie Repräsentativität gemessen werden kann, schwer zu finden. Daher werden eine Reihe von unterschiedlichen Methoden (Bootstrapping, Kreuzvalidierung, Modell-Performance-Messungen etc.) vorgeschlagen, die einen Eindruck darüber vermitteln können, inwieweit eine bestehende Stichprobe für die Grundgesamtheit repräsentativ ist. Insgesamt überwiegt jedoch der Eindruck, dass es noch an Forschungen zu messbaren Konzepten für Repräsentativität und zu den Anforderungen an Messgrößen, die für die Anwendung in Machine-Learning-Umgebungen geeignet sind, fehlt (Clemmensen und Kjærsgaard 2023).

Ein Mittel, um einen Eindruck von der Repräsentativität einer Stichprobe zu erlangen, ist, die Stichprobe in zwei zufällig aufgeteilte Datensätze als Trainings- und Testdatensätze zu splitten und dabei darauf zu achten, dass die

Tab. 1: Hinweise auf die Repräsentativität einer Stichprobe

Grundstücksmerkmale und Repräsentativität der Stichprobe … geringe Abweichungen sind ein Hinweis auf gute Repräsenatitvität der Stichprobe					
	Trainingsdaten (n = 1238)		Testdaten (n = 312)		Test/Train
Merkmal	vorhandene Werte	Medianwert	vorhandene Werte	Medianwert	Abweichung
Kaufpreis (€)	100 %	500.000	100 %	505.500	1,1 %
Rechtswert (m)	100 %	569.326	100 %	569.311	–0,0 %
Hochwert (m)	100 %	5.835.026	100 %	5.834.512	–0,0 %
Entfernung vom Zentrum (m)	100 %	67.172	100 %	67.126	–0,1 %
Bodenrichtwert	100 %	470	100 %	460	–2,1 %
Kaufzeitpunkt (Jahre)*	100 %	–2,68	100 %	–2,74	2,1 %
Grundstücksfläche (m^2)	100 %	587	100 %	568	–3,3 %
Wohnfläche (m^2)	34 %	147	40 %	151	2,7 %
Baujahr	46 %	1962	53 %	1965	0,2 %
Ausstattungsstandard**	2 %	3	3 %	3	0,0 %

* Wertermittlungsstichtag (2025-01-01) bis Kaufdatum (in Jahren)

** Ausstattungsstandard von 1 (sehr einfach), 3 (mittel) bis 7 (stark gehoben)

Beobachtungen nach bestimmten Kriterien gleichmäßig in beiden Datensätzen auftauchen (Stratifikation).

Das Vorgehen kann an dem Beispieldatensatz (Tab. 1) dargestellt werden. Angenommen, in der Stadt »A« soll der Wert je m² Wohnfläche für Ein- und Zweifamilienhäuser ermittelt werden. Dazu liegen aus den Jahren 2020 bis 2024 insgesamt 1.550 Kaufpreise und einige Informationen zu den Objekten vor, wie z. B. Wohnfläche, Baujahr, Grundstücksfläche und Ausstattungsstandard. Diese Stichprobe wird in zwei Substichproben so aufgeteilt, dass eine Stichprobe nunmehr 80 % der Beobachtungen (Trainingsdatensatz) enthält und die andere 20 % der Beobachtungen (Testdatensatz) umfasst. Die Aufteilung der jeweiligen Zeilen der Stichprobe ist zufällig. Danach können die Medianwerte der Ausprägungen der Objektmerkmale jeweils für die unterschiedlichen Stichproben gebildet werden (siehe Tab. 1).

Die Abweichungen der Testdaten von den Trainingsdaten liegen in nahezu allen Fällen unterhalb von ± 5 %. Auch die Anteile der fehlenden Werte bei der Wohnfläche und dem Baujahr sind sehr ähnlich. Dies ist ein deutlicher Hinweis darauf, dass der vorliegende Datensatz die Grundgesamtheit über den räumlichen und sachlichen Teilmarkt gut repräsentiert. Der Vergleich der Medianwerte der Entfernung zum Zentrum z. B. zeigt, dass keine räumliche Verzerrung vorliegt. Auch könnten die Verteilungen der Daten zu Wohn- und Grundstücksfläche, Baujahr, Kaufpreis etc. verglichen werden, um zu überprüfen, ob die sachlichen Merkmale ähnlich sind und somit auch hier keine Verzerrung vorliegt. Über diesen Ansatz hinaus können auch andere Verfahren wie Bootstrapping oder Kreuzvalidierung hilfreich sein, um Erkenntnisse zur Frage der Repräsentativität der Stichprobe zu erlangen.

Über die Hinweise zur Repräsentativität hinaus ergeben sich aus Tab. 1 auch Informationen über den Umfang fehlender Werte. So fehlen z. B. Informationen über das Baujahr in mehr als 50 % der Fälle; über die ausgesprochen wertrelevante Größe der Wohnfläche bei den Ein- und Zweifamilienhäusern fehlen sogar 64 % der Informationen und was den Ausstattungsstandard betrifft, ist die Anzahl der vorhandenen Angaben mit 2 % aller Fälle verschwindend gering. Demgegenüber ist in allen Fällen die Grundstücksfläche bekannt.

Eine Überlegung für den folgenden Schritt des Preprocessing könnte daher sein, einen Zusammenhang zwischen der Wohnfläche und dem Baujahr mit z. B. der Grundstücksfläche und anderen Variablen (z. B. Koordinaten) zu überprüfen. Liegt hier ein signifikanter Zusammenhang vor und ist eine valide Schätzung der Wohnfläche aus anderen Daten möglich, so können die fehlenden Werte in der Stichprobe durch prädizierte Werte ersetzt (imputiert) werden. Dies soll in dem hier vorgestellten Beispiel mit Hilfe des Random-Forest-Verfahrens erfolgen. Damit gehen keine Informationen verloren und es kann die Zielgröße »Wohnflächenpreis je m²« auf alle Datensätze verwendet werden.

Aufgrund der sehr schwachen Datenlage der Werte für den Ausstattungsstandard ist es eher unwahrscheinlich, z. B. einen evidenz-basierten Zusammenhang zwischen dem Baujahr und dem Ausstattungsstandard zu ermitteln, sodass die fehlenden Werte nicht imputiert werden können. Diese Variable würde in den Analysen daher nicht berücksichtigt werden können.

Durch die Medianwerte der qualitativen Variablen ist im Zuge der Datenexploration gut erkennbar, dass es sich z. B. in dieser räumlichen und sachlichen Grundgesamtheit (dem sog. Immobilienteilmarkt) um eher ältere Ein- und Zweifamilienhäuser, im Mittel des Jahrgangs um 1960, handeln dürfte, deren Wohnfläche sich um etwa 140 bis 150 m² bewegt und deren Grundstücksfläche etwa 550 bis 650 m² betragen dürfte.

Um Zusammenhänge der Grundstücksmerkmale zu ermitteln, ist die Ermittlung von Korrelationen zwischen den Variablen erforderlich. Der Korrelationskoeffizient misst die Stärke und die Richtung des Zusammenhangs zweier Variablen (z. B. Kaufpreis und Wohnfläche). Dabei ist zu beachten, dass es unterschiedliche Ansätze gibt, je nach Art der Variablen (qualitativ oder quantitativ) und der Art des Zusammenhangs (linear oder nicht linear). Die Ansätze werden z. B. als Korrelation nach »Pearson«, »Spearman« oder »Kendal« bezeichnet.

Der Zusammenhang des Kaufpreises mit den quantitativen Variablen (Wohnfläche, Grundstücksfläche und Baujahr) sowie der Zusammenhang dieser Variablen untereinander können in einer Korrelationsmatrix dargestellt werden (Tab. 2).

Aus der Matrix für die Korrelationen aus dem Trainingsdatensatz ist ersichtlich, dass der Kaufpreis je m² Wohnfläche (in Euro) mit einer Reihe von Variablen Korrelationen aufweist. In der Regel werden die Korrelationskoeffizienten dann mit »r« bei Korrelationen nach Pearson, »ρ (rho)« bei Korrelationen nach Spearman und »τ (tau)« bei Korrelationen nach Kendall bezeichnet (Fahrmeir et al. 2011).

Insbesondere stellt sich aber auch die Frage, wie die Ausprägungen der Korrelationskoeffizienten zu interpretieren sind. Die grundlegenden Hinweise hierzu geben die folgenden Kategorien wieder (Cohen 1988):

- schwache Korrelation: $r \approx 0{,}1$
- mittlere Korrelation: $r \approx 0{,}3$
- starke Korrelation: $r \approx 0{,}5$

Diese Angaben allerdings sind Heuristiken, also Faustregeln, die je nach Anwendungsgebiet abweichende Werte haben können. Für den Bereich der Immobilienwertermittlung können aufgrund der allgemeinen Erfahrungen die Werte als nutzbar eingestuft werden.

Insgesamt ist jedoch auch hier zu beachten, dass die Korrelationskoeffizienten instabil sein können, da die Voraussetzungen monotoner Zusammenhänge nicht unbedingt gegeben sind. Bei z. B. der Preisentwicklung im Zeitverlauf (»Jahre.seitWE_stichtag«) kann davon ausgegangen werden, dass die Kaufpreise noch bis zum Jahr 2021/2022 gestiegen, danach jedoch gefallen sind; hier ist demnach kein monotoner Zusammenhang zu erwarten. Der nach Spearman ausgewiesene schwache Korrelationskoeffizient

Tab. 2: Korrelationskoeffizienten nach Spearman

Korrelationen, Trainingsdaten Korrelationskoeffizienten nach Spearman (ρ) – Voraussetzung sind monotone Zusammenhänge										
	Rechtswert	Hochwert	Entfernung vom Zentrum	Bodenrichtwert	Jahre seit WE-Stichtag	Grundstücksfläche	Wohnfläche	Baujahr	Kaufpreis m² Wohnfläche	Kaufpreis
Kaufpreis m² Wohnfläche	**0.20**	**0.02**	**0.19**	**0.38**	**–0.11**	**0.01**	**0.00**	**0.33**	–	**0.77**
Kaufpreis	0.11	–0.02	0.14	0.36	–0.10	0.20	0.58	0.31	0.77	–
Rechtswert	–	0.07	0.74	0.29	0.01	–0.13	–0.10	0.25	0.20	0.11
Hochwert	0.07	–	–0.58	0.17	–0.07	0.08	0.00	0.02	0.02	–0.02
Entfernung vom Zentrum	0.74	–0.58	–	0.16	0.03	–0.16	–0.07	0.17	0.19	0.14
Bodenrichtwert	0.29	0.17	0.16	–	0.39	0.13	0.13	–0.06	0.38	0.36
Jahre seit WE-Stichtag	0.01	–0.07	0.03	0.39	–	0.10	–0.01	–0.10	–0.11	–0.10
Grundstücksfläche	–0.13	0.08	–0.16	0.13	0.10	–	0.36	–0.39	0.01	0.20
Wohnfläche	–0.10	0.00	–0.07	0.13	–0.01	0.36	–	0.02	0.00	0.58
Baujahr	0.25	0.02	0.17	–0.06	–0.10	–0.39	0.02	–	0.33	0.31

von –0,11 dürfte daher nicht die tatsächlichen Zusammenhänge darstellen.

Es zeigt sich aber auch, dass es regelmäßig Multikollinearität gibt. Es bestehen Zusammenhänge zwischen den Einflussgrößen. Dieser Umstand ist ein Hinweis darauf, dass z. B. die Methode der klassischen Regressionsanalyse nicht besonders geeignet sein dürfte. Insofern kann es angezeigt sein, alternative Methoden zur Modellbildung zu wählen.

Insgesamt zeigt dieser Schritt, wie es gelingen kann, auf der Grundlage einer Stichprobe auf ihre Repräsentativität, die Zusammenhänge der Merkmale und schließlich auf eine anzuwendende Methode zu kommen. Wichtig ist in diesem Zusammenhang, dass zukünftig eine Automatisierung der Prüfvorgänge zu Machine-Learning-Prozessen und damit zur Anwendung Künstlicher Intelligenz führen muss. Bei einigen Methoden (z. B. Random-Forest) sind detaillierte Prüfvorgänge nicht erforderlich, wenn der Workflow der Analyse dies im Modellierungsprozess mit abfängt.

6 Methoden und Preprocessing

Für die Entwicklung von Modellen zur Abbildung des Immobilienmarktes in sachlichen und räumlichen Teilmärkten ist bislang die Anwendung linearer Regressionsanalysen die klassische und am weitesten verbreitete Methode. Im internationalen und in wirtschaftswissenschaftlichen Kontexten wird diese Methode oft als »hedonisches Preis-Modell (hedonic price model, HPM)« bezeichnet. Allerdings zeigen neuere Forschungen, dass die mit dieser Methode entwickelten Modelle durchaus nachteilig sein können, weil gerade im Immobilienmarkt die für klassische Regressionsmodelle geltenden Annahmen, wie z. B. das nicht Vorhandensein von Multikollinearität, lineare Beziehungen zwischen den Ziel- und Einflussgrößen oder schwer zu beurteilende geografische Attribute (Lee et al. 2024), nicht gelten.

Im Zusammenhang mit der Nutzung Künstlicher Intelligenz und hier der Verwendung von »Machine-Learning-Prozessen« kommen daher zunehmend auch andere Methoden zum Tragen. Es entwickelt sich der Grundsatz, dass nicht pre-analytische statistische Zusammenhänge gelten müssen, sondern diese Zusammenhänge durch die Analyse von Daten extrahiert und im Zuge der Modellbildung weiterverarbeitet werden. Zu diesen Ansätzen gehören z. B. die »Random-Forest-Methode« oder – als »Deep-Learning-Ansatz« – die Verwendung von Neuronalen Netzwerken (Neural Networks, NNs) bzw. Convolutional Neural Networks (CNNs). CNNs eignen sich besonders zur Verarbeitung von Bildinformationen, um visuelle Merkmale, wie Nachbarschaftseigenschaften, besondere Gebäude- oder Grundstücksmerkmale oder auch Innenansichten zur Ausstattung von Wohnungen, zu verarbeiten und daraus Preiseffekte abzuleiten.

Nach der aktuellen Literatur wird besonders die Random-Forest-Methode als eine mittlerweile etablierte und leistungsfähige Methode zur Immobilienwertermittlung erwähnt, aber auch die Verwendung von Neuronalen Netzen kommt zunehmend in den Blick. Im Vergleich zu traditionellen Regressionsmodellen zeigt sich hier eine erheblich größere Verwendungsbreite insbesondere deshalb, weil ihre Anwendung in Märkten mit den üblichen nichtlinearen Zusammenhängen von Grundstücksmerkmalen und deren Multikollinearität robuster und flexibler ist (Yazdani und Raissi 2023).

Tab. 3: Metriken zur Messung der Modell-Performance

n	*Anzahl der Beobachtungen* (= Kauffälle) der Stichprobe (Trainingsdaten, Testdaten).	n = Anzahl der Beobachtungen von y
R^2.*perc*	*Bestimmtheitsmaß* R^2 oder auch Coefficient of Determination (CODE): Bildet ab, wie viel Prozent der ursprünglichen Varianz durch das Modell erklärt wird. Bestehen erhebliche Unterschiede zwischen dem trainierten Modell und den Testdaten, dann ist zu vermuten, dass das Modell überangepasst ist.	$R^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$ $\hat{y}_i$ = Schätzwert für jede Beobachtung $\bar{y}$ = Mittelwert aller Beobachtungen
MdAPE	*Median Absolute Percentage Error:* Beschreibt die Vorhersagegenauigkeit eines Modells als prozentuale Abweichungen der prädizierten Werte von den tatsächlich beobachteten Werten. Ein *MdAPE* von 15 % bedeutet, dass diese Abweichung im Mittel bei ± 15 % liegt. Es können die folgenden Faustzahlen für die Interpretation verwendet werden: ■ 0 bis 10 % : hohe Genauigkeit ■ 10 bis 20 % : gute Genauigkeit ■ 20 bis 50 % : noch angemessene Genauigkeit ■ > 50 % : nicht mehr angemessene Genauigkeit Der *MdAPE* ist bei kleineren Stichproben ($n < 30$) unsicher.	$MdAPE = median\left(abs\left(\frac{\epsilon_i}{y_i}\right)\right)\times 100$ $\epsilon_i = y_i - \hat{y}_i$
PPE	*Percent Predicted Error:* Bildet ab, wie viel Prozent der Schätzwerte innerhalb eines bestimmten Fehlerbereichs (z. B. ± 15 %) der tatsächlichen Werte liegen.	$PPE = \frac{n_{in}}{n}\times 100$ Fehlerbereich = $y_i \pm 15\%$ n_{in} = Anzahl im Fehlerbereich
RMdSE	*Root Median Square Error:* Bildet die Abweichungen der Vorhersagen von den Schätzwerten in absoluten Zahlen ab. Zu beachten ist jedoch, dass es zu einer Überanpassung kommen kann, wenn das Modell mit dem geringsten *RMdSE* gewählt wird.	$RMdSE = \sqrt{median(\epsilon^2)}$

7 Trainieren und Testen, Modell-Performance und Ergebnisgenauigkeit

Bei der Anwendung von Machine-Learning-Methoden zur Nutzung von Künstlicher Intelligenz bei der Wertermittlung von Immobilien ist es Standard, die Modelle zu trainieren und in mehreren Iterationen das für die Aufgabe geeignetste Modell abzuleiten. Dabei ist in den Blick zu nehmen, dass nicht immer jenes Modell das »beste« ist, welches z. B. das höchste Bestimmtheitsmaß (R^2) aufweist oder die geringste Varianz. In Wahrheit kommt es darauf an, dass das finale Model auf alle Daten, die der interessierenden Grundgesamtheit angehören, die gleiche Güte aufweist wie auf den Trainingsdaten. Dieses Vorgehen stellt zunächst die Ermittlung der Performance eines Modells dar. Hier geht es darum, festzustellen, wie robust das Modell auf andere Daten der Grundgesamtheit reagiert. Bleibt die Performance gleich oder sind die Metriken für die Messung der Performance z. B. auf die Testdaten erheblich abweichend? Ist Letzteres der Fall, liegt es nahe, dass das Modell z. B. überbestimmt ist und auf den Trainingsdaten eine Genauigkeit insinuiert, die es tatsächlich auf andere Stichproben (hier die Testdaten) nicht aufweist.

In Anlehnung an neuere Literatur, die im Kontext der Entwicklung von Machine-Learning-Modellen bzw. der Entwicklung des Automatisierten-Valuation-Modells veröffentlicht ist (z. B. Dimopoulos und Bakas 2019; Lee et al. 2024), werden die wesentlichen Metriken zum Vergleich von Trainings- und Testdatensatz in Tab. 3 vorgeschlagen, um insbesondere die Leistungsfähigkeit von Modellen zu ermitteln und ggf. auch verschiedene Modelle zu vergleichen. Insgesamt ist festzustellen, dass diese Ansätze geeignet sind, die Transparenz und Validität bei der Immobilienwertermittlung insoweit zu gewährleisten, als die Leistungsfähigkeit der Modelle anhand von quantitativen Metriken dargestellt werden kann (Ache 2025).

Beispielhaft sollen hier die Metriken aus einem Random-Forest-Modell (RF-Modell) dargestellt werden,

Tab. 4: Metriken zur Messung der Modell-Performance aus dem Trainingsdatensatz für den Wert von Ein- und Zweifamilienhäusern

Performance-Metriken … starke Unterschiede = ungünstige Modell-Performance			
	Werte		Relation
Metrik	Train	Test	Test/Train
n	1.238	312	25.2 %
R^2.*perc*	95.4 %	33.9 %	35.5 %
MdAPE	5.2 %	15.5 %	298.1 %
*PPE**	88.5 %	48.7 %	55.0 %
RMdSE	182	572	313.9 %

* Fehlerbereich der Beobachtungen = ± 15 %

welches mit den Trainingsdaten der Tab. 1 entwickelt worden ist und sodann auf die entsprechenden Testdaten angewendet wurde (Tab. 4). Das Modell resultiert aus der Anwendung der Random-Forest-Methode, bei der es sich um ein ensemblebasiertes Verfahren aus dem Umfeld der »machine leaning methods« handelt. Die Methode beruht auf einer Vielzahl zufällig ausgewählter Stichproben der Trainingsdaten (Bootstraping) und einer zufälligen Auswahl von Einflussvariablen. Diese Kombinationen ergeben eine Vielzahl von unabhängigen Entscheidungsbäumen, das RF-Modell. Die Aggregation der Vorhersagen dieser Bäume ergibt dann die jeweiligen Schätzungen der Zielgröße.

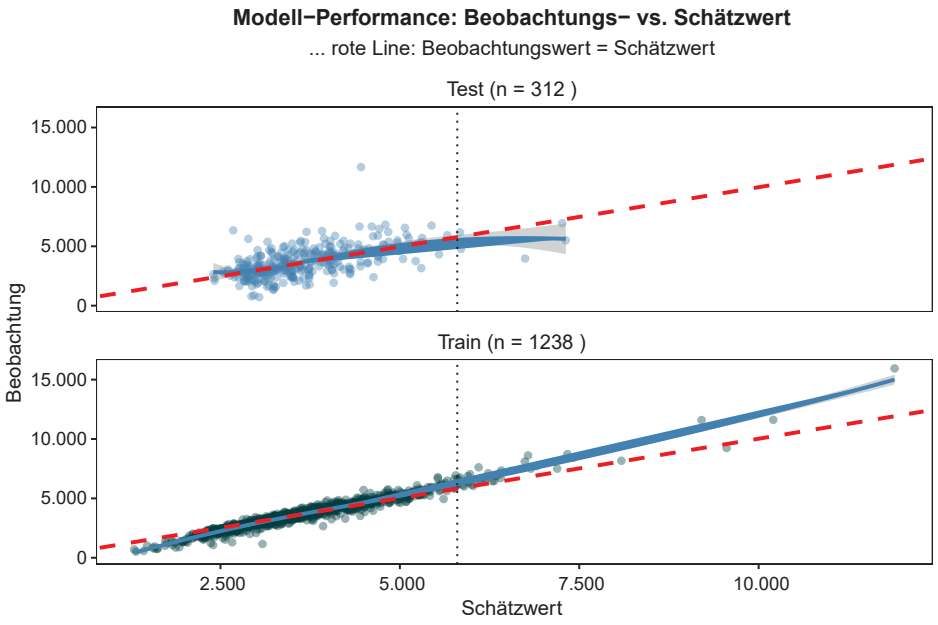

Abb. 1: Modell-Performance: Beobachtungs- vs. Schätzwert

Entsprechend der Beschreibung der Metriken in Tab. 3 zeigt die zusammenfassende Tab. 4 nach dem Vergleich der Metriken von Trainings- und Testdatensatz allerdings auch, dass das trainierte Modell mit einem Bestimmtheitsmaß R^2 = 95,6 % und einem um den Faktor 3,1 höherem RMdSE im Testdatensatz überangepasst ist und damit zunächst eine eher ungenügende Performance aufweist. Der Anteil der innerhalb eines 15 %-Fehlerbereiches liegenden Beobachtungen (*PPE*) beträgt bei den Trainingsdaten ca. 88 % und bei den Testdaten ca. 50 %.

Diese teilweise erheblichen Abweichungen verdeutlichen eine zentrale Schwäche bestimmter Modellierungsansätze, nämlich das Risiko, dass ein Modell die Trainingsdaten nahezu auswendig lernt (Überanpassung oder Overfitting). Dieses Phänomen führt dazu, dass das Modell im Trainingsdatensatz präzise Vorhersagen liefert, während seine tatsächliche Leistungsfähigkeit auf unbekannten Daten eingeschränkt bleibt.

Allerdings ermöglichen die Daten der Tab. 4 keine Aussage darüber, wie der Zusammenhang der Schätzwerte mit den tatsächlichen Werten ist, ob also z. B. ab einer bestimmten Preisklasse eine systematische Unter- oder Überschätzung erfolgt, die das Modell nicht abbilden kann. Dies kann aus der Gegenüberstellung von Schätzwerten und Beobachtungen in einem Streudiagramm für den Test- und Trainingsdatensatz leicht erfolgen (Abb. 1). Dabei muss das besondere Augenmerk auf dem Testdatensatz liegen.

Die tatsächlichen Beobachtungen sind in Abb. 1 in gestreuten Punkten dargestellt, wobei die rote Linie (gestrichelt) die vollständige Übereinstimmung von Beobachtungswert und Schätzwert wiedergibt. Bei einem Schätzwert in Höhe von 5.000 Euro ergäbe sich nach dieser Linie der Beobachtungswert eben auch zu 5.000 Euro. Die blaue Linie hingegen ist die geglättete mittlere Linie (Locally Weighted Scatterplot Smoothing, LOWES) des Verhaltens der Beobachtungen; sie stellt den Trend des tatsächlichen Zusammenhangs zwischen Beobachtungs- und Schätzwerten dar. Der graue Bereich um diese Linie bezeichnet den Konfidenzbereich dieser Linie und es wird – besonders in den Testdaten – die Unsicherheit bei höheren Preisen erkennbar.

In dem Beispiel liegt das Bestimmtheitsmaß $R^2 \approx 35$ % und auch der Wert *PPE* ≈ 50 % für die Testdaten in einem Wertebereich, bei dem auf dem Immobilienmarkt von einer nicht besonders guten, jedoch nicht unüblichen Genauigkeit der Schätzungen ausgegangen werden kann. Auch wird gut erkennbar, dass ab einem Wert von ca. 6.000 €/m² Wohnfläche das Modell eher zur Überschätzung der Werte führt.

8 Konfidenzintervalle und Bootstrapping-Verfahren

In der Immobilienbewertung wird zunehmend eine stärkere Transparenz der Ergebnisse gefordert, insbesondere vor dem Hintergrund der wachsenden Menge digitaler Daten, gestiegener Rechenkapazitäten sowie gesellschaftlicher, umwelt- und sicherheitspolitischer Herausforderungen. Vor allem international gibt es konkrete Bestrebungen, die Transparenz auf dem Immobilienmarkt zu definieren und zu verbessern (Wiejak-Roy et al. 2024). So macht auch die weltweit agierende FIG (International Federation of Surveyors) deutlich, dass mangelnde Transparenz auf dem Immobilienmarkt zu Wertverlusten für die Eigentümer führt und ohne hinreichende Transparenz den aktuellen Herausforderungen nur schwer begegnet werden kann (Ache et al. 2024). Transparenz ist demnach nicht nur ein wesentliches Qualitätskriterium, sondern auch die Voraussetzung für Validität und Akzeptanz von Wertermittlungsergebnissen, insbesondere bei KI-gestützten Methoden. Die Genauigkeit des ermittelten Wertes spielt eine primäre Rolle für

die praktische Verwertbarkeit der Ergebnisse. Die Güte einer Wertermittlung ergibt sich damit einerseits aus der Performance des Modells und andererseits aus der daraus resultierenden Genauigkeit der Verfahrenswerte (§ 6 ImmoWertV) oder der für die Wertermittlung erforderlichen Daten (§ 12 ImmoWertV).

Die Angabe eines Konfidenzintervalls um den ermittelten Verfahrenswert oder die sonstigen für die Wertermittlung erforderlichen Daten ist eine erforderliche Information zur Genauigkeitsbeurteilung. Bei der Angabe eines Konfidenzintervalls hat sich das 95 %-Konfidenzintervall als weit akzeptierter Standard im Kontext der Immobilienwertermittlung etabliert. Ein höheres Konfidenzniveau (z. B. 99 %) führt zu einem breiteren Intervall und damit zu größerer Sicherheit, ein engeres Intervall (z. B. 65 %) birgt dagegen ein höheres Risiko, dass der wahre Wert außerhalb des Intervalls liegt. Gelegentlich werden engere Vertrauensintervalle auch verwendet, um eine höhere Genauigkeit der Ergebnisse zu insinuieren, auf längere Sicht verfängt eine solche Angabe jedoch in der Regel nicht. Es ist zudem entscheidend, die Art des angegebenen Intervalls zu kennen. Grundsätzlich wird zwischen Prognose- und Konfidenzintervallen unterschieden. Prognoseintervalle geben plausible Bereiche für Einzelwerte innerhalb einer Grundgesamtheit an, während Konfidenzintervalle verwendet werden, um statistische Parameter (z. B. Mittelwerte) und deren Unsicherheiten zu quantifizieren. Konfidenzintervalle werden jedoch häufig fälschlicherweise mit Prognoseintervallen verwechselt (Spence und Stanley 2016). In der Immobilienwertermittlung geht es, insbesondere bei der Ermittlung von für die Wertermittlung erforderlichen Daten, um die Ermittlung von allgemeingültigen Erwartungswerten; aus diesem Grund ist die Verwendung des Konfidenzintervalls angezeigt.

Die Ermittlung von Konfidenzintervallen kann durch verschiedene statistische Verfahren erfolgen, deren Auswahl maßgeblich von der Art und Größe der Stichprobe sowie den zugrunde liegenden Modellbildungsannahmen abhängt. Insbesondere erweist sich das Bootstrap-Verfahren als vorteilhaft, wenn die Annahme einer Normalverteilung der abhängigen Variable nicht erfüllt ist oder Unsicherheiten hinsichtlich anderer theoretischer Voraussetzungen bestehen. Auch hier gilt allerdings, dass dieses Verfahren bei kleineren Stichproben ($n < 30$) nicht geeignet ist. Die Ursache ist, dass Bootstrap-Intervalle auf Stichprobenwiederholungen aus der ursprünglichen Stichprobe beruhen, was bei kleinen Stichproben eine nicht ausreichende Variation erzeugt und die daraus resultierenden Konfidenzintervalle die Unsicherheit der Schätzer zu gering bewerten (Dalitz 2018). Aufgrund der Einschränkungen bei der Anwendung dieses Verfahrens sind Varianten des Grundansatzes entwickelt worden, die z. B. Nachteile kleinerer Stichproben oder Verzerrungen bei nicht-normalverteilten Daten auffangen (Percentile Bootstrap und Bias-Corrected and Acceleratred Bootstrap, BCa). Das Percentile-Verfahren ist die derzeit am weitesten verbreitete Methode zur Ermittlung von Konfidenzintervallen mit Hilfe des Bootstrap-Verfahrens. Es werden bestimmte Quantile (i. d. R. das 0,025- und 0,975-Quantile für ein 95 %-Intervall) genutzt, um die Grenzen des Intervalls zu bestimmen. Es handelt sich bei dem Bootstrap-Verfahren um ein nichtparametrisches Verfahren, welches keine besonderen Annahmen über die Verteilung der Daten macht (Justus et al. 2024). Dies kommt der Verteilung von Daten auf dem Immobilienmarkt sehr entgegen und es erübrigen sich damit ebenfalls Transformationen von z. B. Kaufpreisen, sodass die Verteilung der Daten zumindest annähernd einer Normalverteilung entspricht.

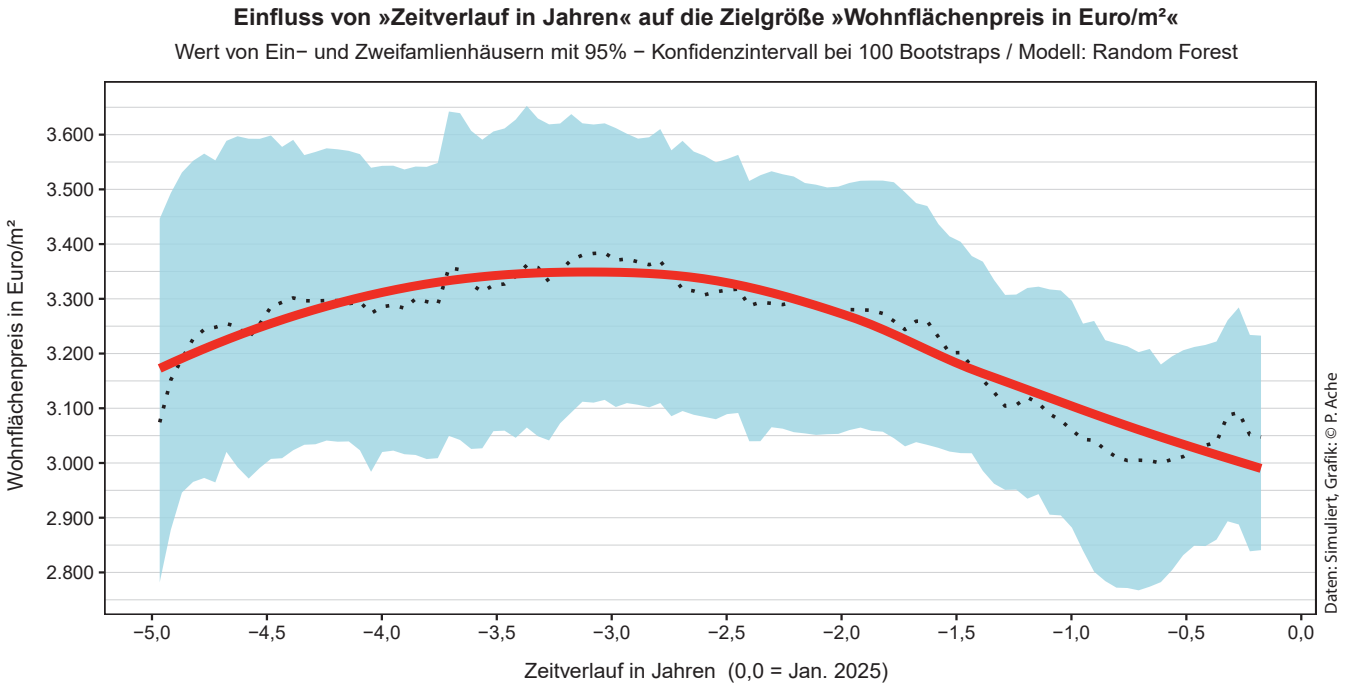

Abb. 2: Preiseffekt des Zeitablaufes in Jahren ab Januar 2025 mit 95 %-Konfidenzintervall aus Bootstrap-Schätzungen

Um die Genauigkeit der Schätzwerte für die interessierende Variable – in diesem Beispiel der Kaufpreis pro m² Wohnfläche – realistisch zu bewerten, wird zunächst ein Datensatz simuliert, der die Medianwerte aller Einflussgrößen enthält. Zusätzlich wird eine Sequenz variierter Werte für die Variable erstellt, deren Effekt isoliert untersucht werden soll. Bei einer Sequenz von 100 Zeilen und 100 Bootstraps entstehen 100 unterschiedliche Stichproben, die jeweils 100 Zeilen umfassen. Auf diese Stichproben wird das zuvor ermittelte Modell angewendet, sodass für jede Zeile der Sequenz 100 Schätzwerte generiert werden. Aus diesen 100 Schätzungen pro Zeile lassen sich das 0,025- und 0,975-Quantil berechnen, die anschließend grafisch dargestellt werden. Dies führt zur Visualisierung eines Konfidenzbandes, das die Genauigkeit der Schätzung für den Einfluss der interessierenden Variable veranschaulicht (Abb. 2) – also der Visualisierung der Ergebnisgenauigkeit aus dem RF-Modell. Auf dieser Grundlage können die Umrechnungskoeffizienten und deren Konfidenzbereiche durch einfache Transformation bestimmt werden.

Dabei stellt die rote Linie in Abb. 2 den Trend des Zusammenhangs zwischen dem Zeitverlauf in Jahren und dem Vergleichsfaktor »Wohnflächenpreis in €/m²« aus dem RF-Modell dar (LOWES). Die gestrichelte schwarze Linie zeigt den detaillierten Verlauf der Schätzwerte aus dem RF-Modell. Erkennbar ist, dass zum Zeitpunkt »Januar.2025 – 5,0 = Januar.2020« ein Schätzwert von etwa 3.080 €/m² ermittelt wird, der im Verlauf des ersten Quartals stark anzusteigen scheint; möglich ist, dass hier bereits ein Preiseffekt aus dem Beginn der Corona-Pandemie Anfang des Jahres 2020 vorliegt. Da dies angesichts des breiten Konfidenzintervalls jedoch eher einer Spekulation zugeordnet werden sollte, führt dazu, dass in Abb. 2 die Trendlinie präsenter dargestellt werden muss als die Einzelschätzungen.

9 Fazit

Zusammenfassend lässt sich festhalten, dass Transparenz und Qualität in der Immobilienwertermittlung durch den Einsatz KI-gestützter Methoden zunehmend an Bedeutung gewinnen. Die klare Unterscheidung zwischen Modell-Performance und Ergebnisgenauigkeit ist essenziell für die Validität der Bewertungen. Eine verlässliche Wertermittlung erfordert repräsentative Daten, den Einsatz geeigneter Modelle sowie die Berücksichtigung von Konfidenzintervallen, um präzise und nachvollziehbare Ergebnisse zu gewährleisten. Moderne statistische Verfahren wie das Bootstrap-Verfahren verbessern die Genauigkeit der Schätzwerte, insbesondere im Kontext der fortschreitenden Digitalisierung. Es ist daher ein breiter fachlicher Diskurs erforderlich, um neue statistische Methoden jenseits der klassischen Regressionsanalyse weiterzuentwickeln und so die Transparenz und Güte von Wertermittlungen nachhaltig zu verbessern.

Literatur

Ache, P. (2025): 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., Wiejak-Roy, G., Kavanagh, J., Korinke, E. K., Reydon, B. (2024): FIG Position Paper: Viewpoint on Transparency in Real Estate Markets. www.fig.net/resources/monthly_articles/2024/November_2024/FIG-Viewpoint_Transparency.pdf, letzter Zugriff 2/2025.

Brydges, C. R. (2019): Effect Size Interpretation, Sample Size Calculation, and Statistical Power in Gerontology. In: PsyArXiv. DOI: 10.31234/osf.io/u2jbm.

Clemmensen, L. H., Kjærsgaard, R. D. (2023): Data Representativity for Machine Learning and AI Systems. In: arXiv. DOI: 10.48550/arXiv.2203.04706.

Cohen, J. (1988): Statistical Power Analysis for the Behavioral Sciences. 2nd edition, L. Erlbaum Associates. www.utstat.toronto.edu/~brunner/oldclass/378f16/readings/CohenPower.pdf, letzter Zugriff 2/2025.

Dalitz, C. (2018): Construction of Confidence Intervals. In: arXiv. DOI: 10.48550/arXiv.1807.03582.

Dimopoulos, T., Bakas, N. (2019): Sensitivity Analysis of Machine Learning Models for the Mass Appraisal of Real Estate. Case Study of Residential Units in Nicosia, Cyprus. In: Remote Sensing, Vol. 11, Iss. 24, 3047. DOI: 10.3390/rs11243047.

Fahrmeir, L., Künstler, R., Pigeot, I., Tutz, G. (2011): Statistik – Der Weg zur Datenanalyse. 7. Auflage, Springer.

Justus, V. L., Rodrigues, V. B., Sousa, A. R. dos S. (2024): Bootstrap confidence intervals: A comparative simulation study. In: arXiv. DOI: 10.48550/ARXIV.2404.12967.

Lee, H., Han, H., Pettit, C., Gao, Q., Shi, V. (2024): Machine learning approach to residential valuation: a convolutional neural network model for geographic variation. In: The Annals of Regional Science, Vol. 72, Iss. 2, 579–599. DOI: 10.1007/s00168-023-01212-7.

Renigier, M. (2008): Residuals Analysis for Constructing »More Real« Property Value. In: Kauko, T., d'Amato, M. (eds.): Mass Appraisal Methods: An International Perspective for Property Valuers. Wiley, 148–163. DOI: 10.1002/9781444301021.ch7.

Schwieger, V., Zhang, L. (2019): Qualität in der Ingenieurgeodäsie – Begriff und Modellierung. In: DVW e.V. (Hrsg.): Qualitätssicherung geodätischer Mess- und Auswerteverfahren 2019. DVW-Schriftenreihe, Band 96, Augsburg, 9–30.

Spence, J. R., Stanley, D. J. (2016): Prediction Interval: What to Expect When You're Expecting … A Replication. In: PLOS ONE, Vol. 11, Iss. 9, e0162874. DOI: 10.1371/journal.pone.0162874.

UNESCO (2021): UNESCO-Empfehlung zur Ethik der Künstlichen Intelligenz. www.unesco.de/dokumente-und-hintergruende/publikationen/detail/unesco-empfehlung-zur-ethik-der-kuenstlichen-intelligenz/, letzter Zugriff 2/2025.

Wiejak-Roy, G., Reydon, B., Ache, P., Neubrand, E., James, K. (2024): Global inquiry into transparency in the Real Estate Markets. In: 30th Annual European Real Estate Society Conference, Sopot and Gdańsk, Poland. DOI: 10.15396/eres2024-159.

Yazdani, M., Raissi, M. (2023): Real Estate Property Valuation using Self-Supervised Vision Transformers. In: arXiv. DOI: 10.48550/arXiv.2302.00117.

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