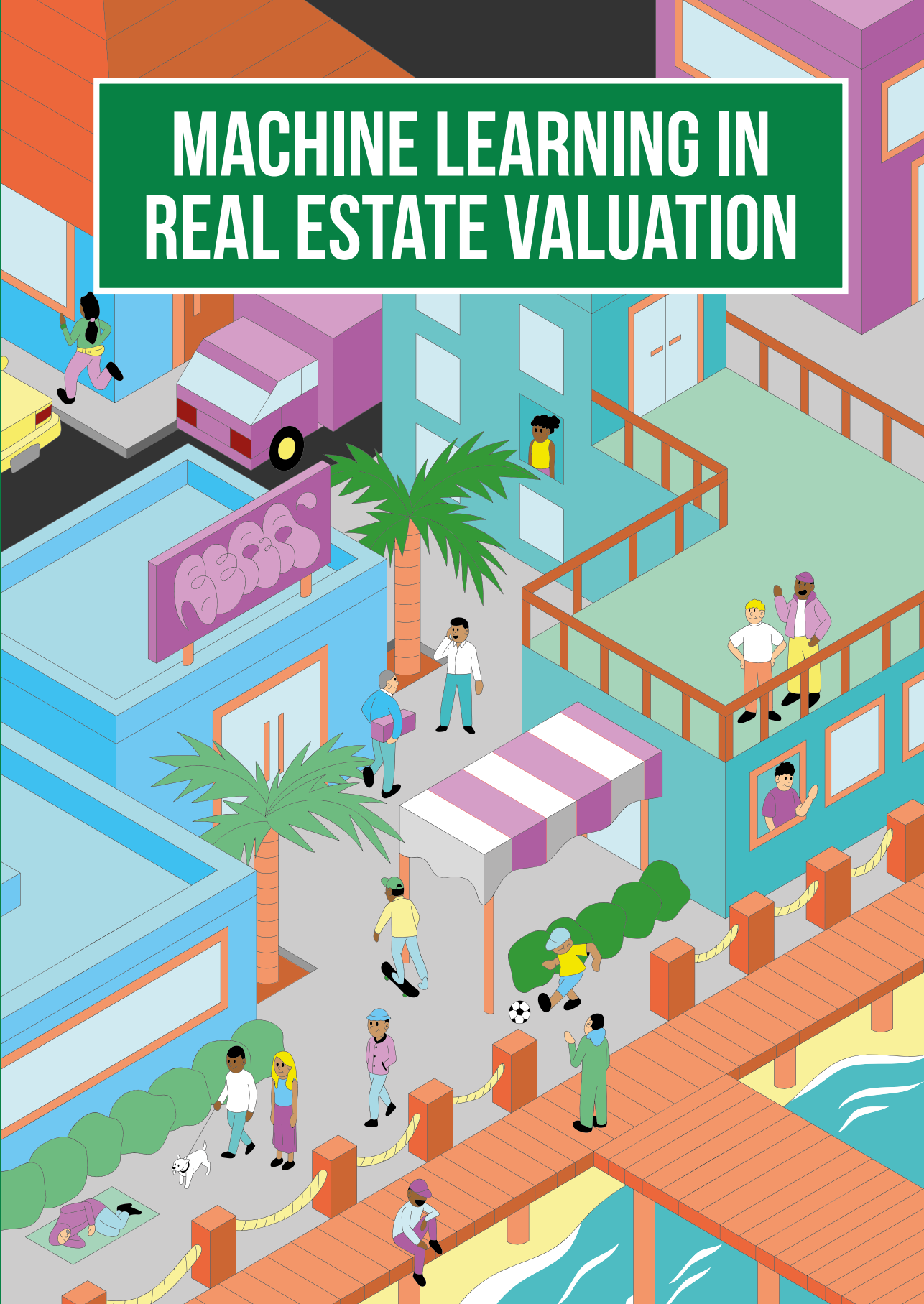
MACHINE LEARNING IN REAL ESTATE VALUATION



Propositions

Felipe Dutra Calainho

22nd of October 2026

1. In real estate valuation, machine learning is not just a way to predict prices more accurately, but a way to measure markets: through indices, uncertainty estimates, and distributional diagnostics.

2. Machine learning predicts commercial real estate prices more accurately than linear models, but that edge shrinks as data becomes scarce, and in thin samples it can disappear entirely.

3. Out-of-time prediction errors offer a principled way to build price indices with any regression model, machine learning included.

4. A good price index must be stable over time, not merely accurate at the transaction level.

5. In automated valuation models, coverage depends mainly on the interval method, while the tightness of the interval depends mainly on the regression model.

6. An automated valuation model should report the uncertainty around its predictions, not just the predictions themselves.

7. Conformal prediction is a practical way to quantify uncertainty without distributional assumptions, but it becomes unreliable when samples are small or exchangeability breaks down.

8. Mean and median indices hide how different market segments move. Monitoring the market well requires looking at the whole price distribution.

9. Housing price segments do not move together; some lead and others follow, and the pattern may vary across markets.

Propositions belonging to the dissertation entitled

"Machine Learning in Real Estate Valuation"

Machine Learning in Real Estate Valuation

Felipe Dutra Calainho

The template used to format this PhD dissertation has been released under the permissive AGPL license and can be found online from the author's github repository, at github.com/lhoangan/template-uva-thesis, which originates from its lead author, Jordan Suchow, at github.com/suchow/Dissertate.

Copyright © 2026 by Felipe Dutra Calainho
Cover Ilustration by: Guilherme Abrão Da Silva, @beraverso.
The printing of this thesis was self-funded.
This thesis is licensed under a Creative Commons "Attribution 4.0 International" license.

ISBN 978-94-6534-636-6

Machine Learning in Real Estate Valuation

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. P. P. C. C. Verbeek
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op weekdag 22 oktober 2026, te 16:00 uur

door

Felipe Dutra Calainho

geboren te Brasília

Promotiecommissie

Promotor:	prof. dr. Marc Francke	Universiteit van Amsterdam
Co-promotor:	dr. Alexander van de Minne	Cornell University
Overige leden:	prof. dr. Thies Lindenthal	University of Cambridge
	prof. dr. Jan Rouwendal	Vrije Universiteit Amsterdam
	prof. dr. Marc Salomon	Universiteit van Amsterdam
	dr. Torsten Jochem	Universiteit van Amsterdam
	dr. Martijn Dröes	Universiteit van Amsterdam

Faculteit Economie en Bedrijfskunde

The work described in this thesis has been carried out within the Finance Group of the University of Amsterdam. The research is supported by Ortec Finance.

Contents

List of Figures 9

List of Tables 13

Acronyms 17

1 General introduction 1
1.1 Summary of the included papers and doctoral contribution 3

2 Using Machine Learning to Build Real Estate Price Indices 7
2.1 Introduction 9
2.2 Methodology 11
2.3 Data 16
2.4 Results 17
2.5 Conclusion 29
Appendix 2.A1 New York Indices and Errors per Property Type 35
Appendix 2.A2 Welch's *t*-Test Results 37

3 Quantifying Uncertainty in Real Estate Price Predictions 41
3.1 Introduction 43
3.2 Literature review 45
3.3 Methodology 48
3.4 Data 58
3.5 Results 60
3.6 Conclusion 68
Appendix 3.A1 Additional Prediction Interval Performance 71

4 Lead-lag Relationships in Quantile Index Percentiles 73
4.1 Introduction 75
4.2 Methodology 79
4.3 Data 91
4.4 Results 95
4.5 Conclusion 108
Appendix 4.A1 Smooth Quantile Indices Summary Statistics 111
Appendix 4.A2 Lag-weighted Lasso VAR 114
Appendix 4.A3 Generalized Impulse Response Function 118
Appendix 4.A4 Granger Causality and Reduced VAR Additional Results 122

5 Conclusion 125
5.1 Synthesis of key findings 126
5.2 Contributions 127
5.3 Implications for research and practice 128

5.4 Limitations 128
5.5 Directions for future research 129
5.6 Closing remarks 130

Bibliography 131

Summary 137

Nederlandse samenvatting 145

Publications 153

Acknowledgements 155

List of Figures

2.1 **New York Chained Paasche Price Indices**. *OLS* = Ordinary Least Squares; *XGBT* = Extreme Gradient Boosting Tree; *SVR* = Support Vector Regression; *avNNet* = Neural Networks Using Model Averaging. 19

2.2 **New York Chained Paasche Price Returns**. *OLS* = Ordinary Least Squares; *XGBT* = Extreme Gradient Boosting Tree; *SVR* = Support Vector Regression; *avNNet* = Neural Networks Using Model Averaging. 20

2.3 **New York Chained Paasche Indices RMSEs**. *OLS* = Ordinary Least Squares; *XGBT* = Extreme Gradient Boosting Tree; *SVR* = Support Vector Regression; *avNNet* = Neural Networks Using Model Averaging; *RMSE* = Root Mean Squared Error. 21

2.4 **Returns for Stress Tests**. All the indices returns are similar until the 10% level. At the 5% level, indices returns exhibit a higher volatility, detaching from the other percentage levels in years 2001, 2004 and 2011. Numbers refer to the percentage of the sample size being used: 50, 25, 10 and 5 percent. When omitted the full sample has been used. 24

2.5 **RMSEs for Stress Tests**. The Root Mean Squared Error (RMSE) for the Machine Learning (ML) algorithms is noticeably more sensitive to data availability as it is higher then Ordinary Least Squares (OLS). Starting from the 10% level, the RMSE of the ML algorithms increases, decoupling from the previous levels. At the 5% level all ML algorithms display a higher RMSE than the OLS benchmark. Numbers refer to the percentage of the sample size being used: 50, 25, 10 and 5 percent. When omitted the full sample has been used. 25

2.6 **RMSEs per Window Size.** Each regression algorithm behaves differently, depending on the training window size. The best performing algorithms are Averaged Neural Networks (avNNet) with a window size of 2 years and Extreme Gradient Boosting Tree (XGBT) with a window size of 7 years; *RMSE* = Root Mean Squared Error. 26

2.7 **Cumulative difference between single and double imputation**. All ML algorithms exhibit a difference between single and double imputation, Support Vector Regression (SVR) and Cubist have the highest cumulative difference; *Index difference* = Single minus double imputation difference. 27

2.8 **New York Chained Paasche Price Indices with Double Imputation**. Starting from year 2013 all the indices built using the ML regression algorithms are below the benchmark, different from the single imputation approach, where SVR and Cubist had all their values above the benchmark. 28

2.A1.1 **New York Chained Paasche Price Indices per Property Type**. *OLS* = Ordinary Least Squares; *XGBT* = Extreme Gradient Boosting Tree; *SVR* = Support Vector Regression; *avNNet* = Neural Networks Using Model Averaging. 35

2.A1.2 **New York Apartment Chained Paasche Indices per Property Type RMSEs**. *OLS* = Ordinary Least Squares; *XGBT* = Extreme Gradient Boosting Tree; *SVR* = Support Vector Regression; *avNNet* = Neural Networks Using Model Averaging; *RMSE* = Root Mean Squared Error. . 36

3.1 **Probability Density Estimation of Prediction Interval Width for Three Different Prediction Interval Methods** ($\alpha = 0.05$). *OLS* = Ordinary Least Squares; *XGBT* = Extreme Gradient Boosting Tree; *NCP* = Normalized Conformal Prediction; *DP* = Direct Prediction; *DLUP* = Direct Lower and Upper Prediction. Number of Breaks = 30 63

3.2 **Probability Density Estimation of the Upper Minus Lower Bound Difference ($U - L$) for the DLUP Prediction Interval Method** ($\alpha = 0.05$). *OLS* = Ordinary Least Squares; *XGBT* = Extreme Gradient Boosting Tree; *DLUP* = Direct Lower and Upper Prediction. Number of Breaks = 20 64

3.3 **Amsterdam Map of Mean Prediction Interval Width (MPIW) per Neighborhood for Three Different Prediction Interval Methods** ($\alpha = 0.05$). *OLS* = Ordinary Least Squares; *XGBT* = Extreme Gradient Boosting Tree; *NCP* = Normalized Conformal Prediction; *DP* = Direct Prediction; *DLUP* = Direct Lower and Upper Prediction. 65

3.4 **Rolling Window Coverage Results for the Four Different Prediction Interval Methods** ($\alpha = 0.05$). *OLS* = Ordinary Least Squares; *XGBT* = Extreme Gradient Boosting Tree; *CCP* = Cross Conformal Prediction; *NCP* = Normalized Conformal Prediction; *DP* = Direct Prediction; *DLUP* = Direct Lower and Upper Prediction. 66

3.5 **Mean Prediction Interval Width (MPIW) of the Rolling Window Results for the Four Different Prediction Interval Methods** $\alpha = 0.05$. *OLS* = Ordinary Least Squares; *XGBT* = Extreme Gradient Boosting Tree; *CCP* = Cross Conformal Prediction; *NCP* = Normalized Conformal Prediction; *DP* = Direct Prediction; *DLUP* = Direct Lower and Upper Prediction. 68

3.A1.1 **Rolling Window Coverage Results for the Four Different Prediction Interval Methods** ($\alpha = 0.01$). *OLS* = Ordinary Least Squares; *XGBT* = Extreme Gradient Boosting Tree; *CCP* = Cross Conformal Prediction; *NCP* = Normalized Conformal Prediction; *DP* = Direct Prediction; *DLUP* = Direct Lower and Upper Prediction. 71

3.A1.2 **Mean Prediction Interval Width (MPIW) of the Rolling Window Results for the Four Different Prediction Interval Methods** $\alpha = 0.01$. *OLS* = Ordinary Least Squares; *XGBT* = Extreme Gradient Boosting Tree; *CCP* = Cross Conformal Prediction; *NCP* = Normalized Conformal Prediction; *DP* = Direct Prediction; *DLUP* = Direct Lower and Upper Prediction. 72

4.1 **Warsaw RIF Quantile Index Returns.** *RIF* = Recentered Influence Function. 95
4.2 **Warsaw Quantile Indices.** *RIF* = Recentered Influence Function. . . . 96
4.5 **Generalized Impulse Response Function** (GRIF). One Standard Deviation shock in the 60th Quantile. 104
4.6 **Generalized Impulse Response Function** (GRIF). One Standard Deviation shock in the 10th Quantile. 105
4.A3.1 **Generalized Impulse Response Function** (GRIF). One Standard Deviation shock in the 45th Quantile. 120
4.A3.2 **Generalized Impulse Response Function** (GRIF). One Standard Deviation shock in the 75th Quantile. 121

List of Tables

2.1 **New York sample summary statistics**. *Avg. Price* = Average Price in US Dollars; *SD. Price* = Price Standard Deviation in US Dollars; *Avg. Area* = Average Area in Square Feet; *SD. Area* = Area Standard Deviation in Square Feet; *Obs.* = Number of Observations. 18

2.2 **Categorical variables summary**. *Obs.* = Number of Observations. . . . 18

2.3 **Chained Paasche performance summary**. *OLS* = Ordinary Least Squares; *XGBT* = Extreme Gradient Boosting Tree; *SVR* = Support Vector Regression; *avNNet* = Neural Networks Using Model Averaging; *RMSE* = Root Mean Squared Error; *Volatility* = Index Return Standard Deviation; *ACF(1)* = Index Return First-Order Autocorrelation. 22

2.4 **Summary statistics of the volatility from the 30 samples with 50%, 25%, 10% and 5% of the available observations**. Index return volatility is, on average, lower for OLS when compared to ML algorithms. Also, only at the 5% level a significant increase in index return volatility is noticeable for all regression algorithms; *Min.* = Minimum volatility; *1st Qu.* = First Quartile volatility; *3rd Qu.* = Third Quartile volatility; *Max.* = Maximum volatility; *SD.* = Standard Deviation of index volatility. 31

2.5 **Optimal window performance summary**. ML algorithms display a lower RMSE then the OLS benchmark. The lowest RMSE pertains to avNNet with a window size of 2 years; *RMSE* = Root Mean Squared Error; *ACF(1)* = Index Return First-Order Autocorrelation; *Volatility* = Index Return Standard Deviation; *OWS* = Optimal Window Size in Years. 32

2.6 **Expanding window performance summary**. Expanding window performed worst than the optimal rolling window or using only one year as in the chained index methodology. XGBT regression algorithm has the best performance for the expanding window methodology; *RMSE* = Root Mean Squared Error; *ACF(1)* = Index Return First-Order Autocorrelation; *Volatility* = Index Return Standard Deviation. 32

2.7 **Difference between true and fitted log price means**. The increase in variance of the fitted values for the ML regression algorithms comes at the cost of having some estimation bias. The OLS benchmark has no estimation bias. 32

2.8 **Difference between true and fitted log price variance**. All the ML regression algorithms present a fitted variance closer to the true variance when compared to the OLS benchmark. 33

2.A2.1 **OLS t-test Results.** *LBCI* = Lower Bound of the Confidence Interval; *UBCI* = Upper Bound of the Confidence Interval. 37

2.A2.2 **XGBT t-test Results** *LBCI* = Lower Bound of the Confidence Interval; *UBCI* = Upper Bound of the Confidence Interval. 38

2.A2.3 **SVR t-test Results** *LBCI* = Lower Bound of the Confidence Interval; *UBCI* = Upper Bound of the Confidence Interval. 38

2.A2.4 **avNNet t-test Results** *LBCI* = Lower Bound of the Confidence Interval; *UBCI* = Upper Bound of the Confidence Interval. 39

2.A2.5 **Cubist t-test Results** *LBCI* = Lower Bound of the Confidence Interval; *UBCI* = Upper Bound of the Confidence Interval. 39

3.1 **Amsterdam sample summary statistics**. *Avg. Price* = Average Price in Euros; *SD. Price* = Price Standard Deviation in Euros;*Avg. Area* = Average Area in Square Meters; *SD. Area* = Area Standard Deviation in Square Meters; *Obs.* = Number of Observations. 59

3.2 **Regression Out-of-Sample Residual Metrics**. *OLS* = Ordinary Least Squares; *XGBT* = Extreme Gradient Boosting Tree; *RMSE* = Root Mean Squared Error; *MAPE* = Mean Absolute Percentage Error. 60

3.3 **Results for the Four Different Prediction Interval Methods** ($\alpha = 0.05$). *CCP* = Cross Conformal Prediction; *NCP* = Normalized Conformal Prediction; *DP* = Direct Prediction; *DLUP* = Direct Lower and Upper Prediction; *OLS* = Ordinary Least Squares; *XGBT* = Extreme Gradient Boosting Tree; *Algo.* = Algorithm; *MPIW* = Mean of the Prediction Interval Width; *SDPIW* = Standard Deviation of the Prediction Interval Width; *PI* = Price Interval; *MWIS* = Mean Winkler Interval Score. 62

3.A1.1 **Results for the Four Different Prediction Interval Methods** ($\alpha = 0.01$). *CCP* = Cross Conformal Prediction; *NCP* = Normalized Conformal Prediction; *DP* = Direct Prediction; *DLUP* = Direct Lower and Upper Prediction; *OLS* = Ordinary Least Squares; *XGBT* = Extreme Gradient Boosting Tree; *Algo.* = Algorithm; *MPIW* = Mean of the Prediction Interval Width; *SDPIW* = Standard Deviation of the Prediction Interval Width; *PI* = Price Interval; *MWIS* = Mean Winkler Interval Score. 72

4.1 **Summary statistics before and after data cleaning.** 92

4.2 **Variable definitions.** 93

4.4 **LWL-VAR model performance summary** (squared-error loss, one-step-ahead horizon $h = 1$). *Lag* denotes the VAR lag order; λ^* is the penalty parameter selected by rolling cross-validation on a grid of 50 candidate values; and *Frac. active* is the proportion of non-zero VAR coefficients. *In-sample* and *OOS* report, respectively, the in-sample and out-of-sample mean squared forecast errors (MSFE) of the penalized LWL-VAR. *Mean*, *AIC*, *BIC*, and *RW* give the out-of-sample MSFE of, respectively, a conditional mean forecast, an unpenalized VAR with lag order chosen by AIC, an unpenalized VAR with lag order chosen by BIC, and a random-walk benchmark. All models are estimated on $T = 71$ quarterly observations of quantile returns; λ^* is chosen using 24 rolling cross-validation origins ($t = 23, \dots, 47$), and all out-of-sample MSFEs are computed from $N_{\text{OOS}} = 24$ one-step-ahead forecasts ($t = 48, \dots, 71$). 99

4.5 **System-wide Granger-causality tests in the five-variable VAR(2).** . . 106

4.6 **Pairwise Granger causality tests between smoothed quantile returns (VAR(2)).** 107

4.A2.1 **Lag-weighted Lasso VAR (order $p = 2$) coefficient estimates for smoothed RIF-based quantile-return indices.** The model is $\mathbf{y}_t = \mathbf{c} + \Phi^{(1)}\mathbf{y}_{t-1} + \Phi^{(2)}\mathbf{y}_{t-2} + \mathbf{u}_t$. Panel A reports $\Phi^{(1)}$ (lag $t-1$) and Panel B reports $\Phi^{(2)}$ (lag $t-2$); the intercept is repeated in both panels for reference. 115
4.A2.2 **Lag-weighted Lasso VAR (order $p = 1$) coefficient estimates for smoothed RIF-based quantile-return indices.** The model is $\mathbf{y}_t = \mathbf{c} + \Phi^{(1)}\mathbf{y}_{t-1} + \mathbf{u}_t$. Rows denote response quantiles at time t; columns denote lagged quantile returns. Entries equal to 0 indicate coefficients shrunk to (approximately) zero by the lag-weighted Lasso. 117
4.A4.1 **VAR(2) estimates for equation in $q_{0.10}$ (smoothed returns).** 122
4.A4.2 **VAR(2) estimates for equation in $q_{0.45}$ (smoothed returns).** 122
4.A4.3 **VAR(2) estimates for equation in $q_{0.60}$ (smoothed returns).** 123
4.A4.4 **VAR(2) estimates for equation in $q_{0.75}$ (smoothed returns).** 123
4.A4.5 **VAR(2) estimates for equation in $q_{0.80}$ (smoothed returns).** 124
4.A4.6 **Residual covariance matrix of the VAR(2) in five quantile returns.** . . 124
4.A4.7 **Residual correlation matrix of the VAR(2) in five quantile returns.** . . 124

Acronyms

ACF1 first-order autocorrelation.

AIC Akaike Information Criterion.

AVM Automated Valuation Model.

avNNet Averaged Neural Networks.

BIC Baysian Information Criterion.

CCP Cross Conformal Prediction.

CP Chained Paasche Index.

CV Cross-Validation.

DLUP Direct Lower and Upper Prediction.

DP Direct Prediction.

EW Expanding Window.

GIE Generalized Inverse Estimator.

GIRF Generalized Impulse Response Function.

GLM Generalized Linear Models.

LOOCV Leave One Out Cross-Validation.

LWL-VAR Lag-weighted Lasso vector autoregression.

MAPE Mean Absolute Percentage Error.

ML Machine Learning.

MPIW Mean Prediction Interval Width.

MRPIW Mean Relative Prediction Interval Width.

MSFE mean squared forecast error.

MWIS Mean Winkler Interval Score.

NCP Normalized Conformal Prediction.

NVM Dutch Brokerage Association.

OLS Ordinary Least Squares.

OOS Out-of-Sample.

OWS Optimal Window Size.

PI Prediction Interval.

RCA Real Capital Analytics.

RIF Recentered Influence Function.

RMSE Root Mean Squared Error.

RW Rolling Window.

$\mathbf{RW_{walk}}$ Random Walk.

SDPIW Standard Deviation of the Prediction Interval Width.

SVM Support Vector Machine.

SVR Support Vector Regression.

VAR vector autoregression.

XGBT Extreme Gradient Boosting Tree.

"Vaak moet er iets gebeuren voordat er iets gebeurt."

Johan Cruyff

1

General introduction

THE BUILT ENVIRONMENT is simultaneously a physical system (e.g. buildings, infrastructure, land) and an information system (e.g. contracts, appraisals, permits, blueprints, transactions, sensors). Over the past decades, the "information layer" has expanded rapidly: more events are recorded digitally, more attributes are observed at finer spatial scales, and more data products are created by combining administrative records, market listings, geospatial layers, and proprietary sources. In parallel, advances in computation and statistical learning have broadened the set of models that can be applied to these data, ranging from classical hedonic frameworks to modern machine learning and, more recently, generative Artificial Intelligence.

In real estate finance and economics, improvements in measurement matter because valuation underpins a wide set of economic and institutional functions. Prices and valuations shape capital allocation, inform investment performance assessment, and support risk management in settings where real estate serves as collateral and as a long-duration asset held by households and institutions. The increasing availability of data and computational tools has therefore stimulated renewed interest in applying modern predictive methods to property valuation problems. In this setting, it is natural to view Machine Learning (ML) as part of the same methodological toolbox as Ordinary Least Squares (OLS), Generalized Linear Models (GLM), or other probabilistic models: a collection of instruments that map heterogeneous information into estimates used for decisions, each with its own strengths and limitations. ML algorithms can offer improved predictive accuracy in complex, non-linear settings, but typically provide weaker interpretability

and less transparent links to causal mechanisms (Hastie et al., 2009; Mullainathan and Spiess, 2017).

1

The adoption of these methods also broadens the set of data that can be used in property valuation, including high-dimensional tabular information as well as text and images that have historically been difficult to incorporate into empirical work. At the same time, expanding the modeling frontier does not remove the need to address more fundamental questions about what constitutes informative, comparable, and decision-relevant valuation output. Three questions are particularly salient. First, markets and portfolios require coherent summaries of price change over time, yet index construction remains challenging in the presence of heterogeneous assets, changing composition, and thin markets (Balk et al., 2013; Geltner et al., 2017). Second, operational valuation systems have tended to emphasize point predictions, while the quantification of prediction uncertainty remains less standardized, despite its centrality for risk-sensitive applications. Third, market narratives often rely on average measures, even though segmentation implies that different parts of the price distribution may behave differently, and potentially convey distinct information about market dynamics.

A further complication is that these property valuation outputs are often produced under imperfect information. Real estate data remain fragmented across jurisdictions, market platforms, and private vendors, frequently collected under non-uniform definitions and affected by missingness, duplication, and measurement error. Considerable resources are therefore devoted to data integration and validation, and even sophisticated market participants typically face trade-offs between coverage, timeliness, and consistency. In such a setting, methodological work that improves the robustness of market indicators, and clarifies what can and cannot be inferred from valuation models, is valuable for both research and practice. This thesis contributes to that goal by developing and evaluating approaches that remain informative under heterogeneity, thin markets, and imperfect data, with particular attention to how predictive models can be translated into market- and decision-relevant measurement.

Against this background, this thesis investigates how Automated Valuation Models (AVMs) and modern predictive models can be used to improve real estate valuation and market measurement from three complementary perspectives. Two chapters adopt a primarily cross-sectional perspective, focusing on how predictive models can be used to construct market-level measures (price indices) and property-level measures (value predictions with uncertainty). A third chapter adopts a time-series perspective from a cross-sectional data foundation, studying whether distributional market measures (quantile price indices) move together or exhibit systematic lead-lag relationships that

reveal price segmentation in housing market dynamics. While these perspectives differ in unit of analysis and temporal focus, they are united by a common theme: AVMs are not only predictive tools, but also measurement tools, and their value depends on the quality and interpretability of the market information they produce.

More concretely, this thesis asks how ML-based valuation models can be used (i) to construct real estate price indices in a principled, model-agnostic way; (ii) to quantify price prediction uncertainty by studying how predictive models interact with prediction-interval methodologies and identifying combinations that achieve reliable coverage without becoming uninformatively wide; and (iii) to assess whether housing price quantiles move together over time or whether certain segments systematically lead others, thereby offering a distributional view of market dynamics. Together, these questions connect the three included papers and frame the thesis contribution as an empirical and methodological investigation of ML applications in real estate valuation across indices, uncertainty quantification, and distributional time-series dynamics.

THESIS STRUCTURE. This thesis is written as a collection of three research papers. Chapter 1 introduces the broader context and motivates the research agenda; it also provides concise summaries of the three papers and clarifies the doctoral contribution in co-authored work. Chapter 2 presents the first paper on machine-learning-based price index construction. Chapter 3 presents the second paper on uncertainty quantification for AVMs via model-agnostic prediction intervals. Chapter 4 presents the third paper on lead-lag relationships across quantile price index percentiles. Chapter 5 concludes by synthesizing the findings across the three papers and outlining implications for research and practice.

1.1 Summary of the included papers and doctoral contribution

1.1.1 Chapter 2: A Machine Learning Approach to Price Indices

Motivation and goal. Real estate price indices are a foundational measurement tool for investment analysis, risk management, and market monitoring. Classical approaches are typically developed around linear or semi-parametric models and can be restrictive when the price function is highly non-linear. Building on the hedonic tradition (Court, 1939; Lancaster, 1966; Malpezzi, 2002), this paper proposes a *model-agnostic* framework that allows machine learning models to be used in index construction.

Methodological contribution. The key innovation is to use *out-of-time prediction errors* at the transaction level to measure price change, enabling the construction of indices while

1

flexibly choosing the underlying predictive model (e.g., linear regression, tree ensembles, support vector regression, neural networks). The framework is designed to separate (i) the predictive model used to estimate prices from (ii) the index aggregation logic, so that non-linear learners can be incorporated without redesigning the index methodology.

Empirical setting and main findings. Using 29,998 commercial real estate transactions in New York (2000–2019), the results show that machine learning models improve prediction accuracy relative to linear benchmarks. However, the analysis also shows that flexible learners can be less stable in smaller samples and may exhibit estimation bias, highlighting a practical governance trade-off: performance gains can come with higher sensitivity to the calibration data and to sampling variation. The paper therefore motivates bias-reduction and stability checks as core components of ML-based index production (Calainho et al., 2022).

1.1.2 Chapter 3: Quantifying Uncertainty in Real Estate Price Predictions

Motivation and goal. AVMs are widely used in property tax, mortgage underwriting, and investment analysis, yet point predictions alone are often insufficient when decisions and contracts are written on individual properties and valuation errors have asymmetric consequences. Chapter 3 therefore studies prediction intervals as practical, property-level summaries of price prediction uncertainty that can complement standard AVM outputs and support risk-sensitive decision rules (Bellotti, 2017; Lim and Bellotti, 2021; Pollestad and Oust, 2025; Shao et al., 2015).

Methodological contribution. The chapter provides a transparent, like-for-like comparison of commonly used, model-agnostic prediction-interval constructions across distinct model classes. Specifically, it evaluates four Prediction Interval (PI) methods—Cross Conformal Prediction (CCP), Normalized Conformal Prediction (NCP), Direct Prediction (DP), and Direct Lower–Upper Prediction (DLUP)—paired with a conventional hedonic OLS benchmark and two high-performing machine-learning regression engines (XGBoost and Cubist). By holding the empirical setup, feature set, and evaluation metrics fixed, the analysis isolates how interval properties (coverage, width, heterogeneity, and asymmetry) change when the same interval logic is combined with different prediction engines, clarifying the sources of variation that matter in applied AVM uncertainty reporting.

Empirical setting and main findings. Using 86,838 residential transactions in Amsterdam (2001–2021), results show a clear division of labor: the regression engine primarily determines PI *efficiency* and *heterogeneity* (average width and its dispersion across properties), whereas the interval construction plays a larger role for attaining nominal *validity*

(coverage). In the main split-sample exercise, Cubist combined with DLUP offers the most favorable width–coverage trade-off. A rolling-window design highlights that coverage can fall below nominal levels when annual training samples are small and exchangeability conditions underlying conformal calibration are strained. Overall, the findings support treating uncertainty quantification as a first-class component of AVMs, and they clarify what practitioners can expect from standard PI methods across model classes in an operational valuation setting.

1.1.3 Chapter 4: Lead–lag Relationships in Quantile Index Percentiles

Motivation and goal. Market summaries based on mean or median indices can conceal heterogeneous dynamics across the price distribution. This paper asks whether different parts of the housing price distribution move together or whether some quantiles systematically lead others, providing a segmentation-based perspective on housing market dynamics.

Methodological contribution. The paper constructs quantile price indices (10th to 90th percentiles) using Recentered Influence Function (RIF) regressions for unconditional quantiles (Firpo et al., 2009), applies frequency conversion/smoothing via a generalized inverse estimator, and studies the joint dynamics in a lag-weighted Lasso vector autoregression (VAR) complemented by generalized impulse responses and a low-dimensional VAR with Granger-causality tests (Machado and Mata, 2005; Nicholson et al., 2020; Park and Sakaori, 2013). The framework is designed to be portable across markets and suitable for identifying leading segments under high dimensionality.

Empirical setting and main findings. Using 175,070 apartment transactions in Warsaw (2006–2024), the quantile indices form a persistent "fan" with widening dispersion, indicating that cheaper and more expensive segments drift apart over time. Dynamic analysis shows that quantile returns do not move in lockstep: lower-middle and upper-middle segments, especially the 10th and 60th percentiles, contain disproportionate predictive content for the rest of the distribution, while the upper tail behaves mainly as a follower. The results primarily contribute a segmentation insight that can inform market monitoring; while the quarterly frequency limits tactical timing applications, the approach can be used as a distributional "thermometer" for public and private stakeholders.

"We look at the present through a rear-view mirror.
We march backwards into the future."

Marshall McLuhan

2

Using Machine Learning to Build Real Estate Price Indices

Authors: Felipe D. Calainho, Alex M. van de Minne, Marc K. Francke

Original title: A Machine Learning Approach to Price Indices: Applications in Commercial Real Estate

Published in: Journal of Real Estate Finance and Economics (Volume: 68, Pages: 624–653, April 2024)

https://doi.org/10.1007/s11146-022-09893-1

Abstract

This article presents a model agnostic methodology for producing property price indices. The motivation to develop this methodology is to include non-linear and non-parametric models, such as Machine Learning (ML), in the pool of algorithms to produce price indices. The key innovation is the use of individual out-of-time prediction errors to measure price changes. The data used in this study consist of 29,998 commercial real estate transactions in New York, in the period 2000–2019. The results indicate that the prediction accuracy is higher for the ML models compared to linear models. On the other hand, ML algorithms depend more on the data used for calibration; they produce less stable results when applied to small samples and may exhibit estimation bias. Hence, measures to reduce or eliminate bias need to be implemented, taking into consideration the bias and variance trade-off.

2.1 Introduction

HAVING RELIABLE real estate price indices is pivotal for several reasons. Firstly, index volatility is an important input in determining the cost of capital of real estate (Geltner et al., 2017). This is helpful for investors, underwriters and policymakers to determine interest rates. Secondly, it allows one to (re-)appraise real estate portfolios mark-to-market, by multiplying (historic) book values by the corresponding index returns (Francke and Van de Minne, 2022). This is of obvious interest for investors, but also for owner-occupiers in determining their household wealth. This can also help appraise commercial and residential property values for property tax purposes. Finally, real estate indexing would allow for derivative trading (Deng and Quigley, 2008), especially in the absence of index revisions. However, producing real estate indices is a nontrivial task as real estate is a heterogeneous asset class that transacts infrequently (Francke and Van de Minne, 2021a).

In the real estate literature, several methods haven been proposed to compile transaction price indices. Hedonic pricing models and repeat sales are two of the most popular ones. The hedonic pricing model is the "original" method of price index construction (Malpezzi, 2002), and will be the focus of this paper. The hedonic pricing model has its origins in appraising farmland values (Haas, 1922a,b; Wallace, 1926), and has been used for making constant quality indices in the automobile industry (Court, 1939). Later examples of literature using hedonic pricing models include Griliches (1961), Lancaster (1966) and Rosen (1974). The hedonic pricing model assumes that the price of a commodity is composed by aggregating the individual contributions of each of its characteristics.[1]

Classic hedonic approaches employ linear models, estimated by OLS, or its generalizations for index construction. These linear algorithms have the convenience of the β parameter, that elicits a linear relationship, in units, between the dependent and independent variables. Nevertheless, the dependence on estimating a parameter vector β constrains the use of a wide family of other regression algorithms, such as ML approaches, that are non-linear or non-parametric. This paper proposes a (hedonic) framework which admits the usage of parametric as well as non-parametric/non-linear models to construct price indices.

The proposed method can be viewed as a model that predicts the price of a property ($\hat{Y}$) sold in the current period (t) as if it had been sold in the previous period ($t-1$) The difference between the estimated price ($\hat{Y}_{t-1}$) and actual price (Y_t) at a property level can be

[1]The repeat sales approach, initially proposed by Bailey et al. (1963), is a method that utilizes the relative price difference of a real estate property that was sold twice or more in a given time interval as the basis for index construction.

decomposed into a market price change and some noise. More specifically, we first estimate some (linear/non-linear/non-parametric) model using only transactions in (or up to and including) period $t-1$. In a second step, we predict the values of all properties sold in period t, using the estimated model in the first step. In the final step, we take the out-of-sample prediction errors for transactions in period t, and compute the average change in order to compute the price index change. To find the returns in subsequent periods, we simply redo these three steps for the remaining time periods $(t, t+1, \ldots, T-2, T-1)$. This methodology is called the chained "Paasche" price index (Geltner et al., 2017). Note that this three step framework is agnostic as to which model and estimation technique is used in the first step, and it can include, but is not limited to, ML techniques.

One of the many reasons to have an price index methodology that incorporates other (regression) algorithms such as ML is that these permit a more flexible relationship between variables when compared to linear models (Varian, 2014). This flexibility increases our expectations of better approximating the unknown complex data generating process that underlies real estate property prices. In other words, it might improve our ability to find the "true" price change between periods that is less affected by noise.

Following this line of thought, this article proposes a new methodology that generalizes the chained hedonic approach in a way that any ML regression algorithm can be used to compile property price indices. Hence, this paper contributes to the existing literature for building non-linear price indices, such as McMillen and Dombrow (2001), and it connects the fields of number theory, econometrics, and ML. In fact, to the best of our knowledge, this paper is the first to estimate price *indices* using ML approaches[2]. Our proposed framework therefore opens up the field of chained hedonic pricing models for numerous new possibilities.

The data used in this study have been provided by Real Capital Analytics (RCA) and consist of 29,998 individual transactions from commercial real estate properties in New York metropolitan area, in the period from 2000 to 2019. We construct an aggregate price index and indices per property type.

In general, the results show that using individual predictions of real estate transaction prices is a viable solution for building price indices. Another finding is that Out-of-Time prediction accuracy is higher for the ML algorithms when compared to OLS. However, ML algorithms are more dependent on the data used for estimating the models, and have less stability when applied to smaller data sets. Additionally, the bias and variance trade-off

[2] However, it should be noted that research in the field of real estate price *levels*, or appraised values, is quite established by now. See for example: Do and Grudnitski (1992); Evans et al. (1992); Kok et al. (2017); McGreal et al. (1998); Nghiep and Al (2001); Peterson and Flanagan (2009); Tay and Ho (1992); Wong et al. (2002); Worzala et al. (1995).

from the ML algorithms has an important role in this methodology, as bias affects the index being estimated. A solution for regression algorithms that exhibit estimation bias is the use of the double imputation method.

We perform a stress test to determine sensitivity to sample size and to examine how much leverage the sample size has on the proposed methodology. This test is performed by sampling 50%, 25%, 10%, and 5% of the available transactions 30 times for each percentage level, totalling 120 samples. The number of repetitions was chosen to ensure that the hyperparameters would be estimated 30 times.

Additionally, we allow the model's predicted property prices in period t to be estimated on *multiple* periods up to and including period $t-1$. Two approaches for determining the optimal window size are tested: Rolling Window (RW) with windows of 2 through 8 years and Expanding Window (EW). These tests determine the relation between model complexity (additional observations at the cost of additional year variables) and model accuracy for each of the algorithms used.

Chained index results for the full sample show that all ML algorithms have lower RMSE and higher R^2 than OLS. Regarding volatility and first-order autocorrelation (ACF1), ML algorithms have similar results to OLS; none of the indices exhibit extreme volatility or a negative ACF1. The trade-off between having more observations by using expanding or rolling windows versus the cost of including time fixed effects, affects the algorithms differently. Another finding is that while performing the stress test, ML algorithms have higher errors compared to OLS when data is restricted. This is more evident at the 5% data restriction level, which suggests a higher data dependence for the ML algorithms.

2.2 Methodology

This paper presents a single imputation Chained Paasche approach towards building real estate price indices (Balk et al., 2013). Linear and non-linear models are used for imputing the value of a real estate property in a different time period. The linear model, estimated by ordinary least squares (OLS), is used as a benchmark. For each period, hyperparameter tuning, model estimation and price prediction are performed. Therefore, there is a model estimated or trained using each algorithm for each period.

Hyperparameter tuning is performed to minimize the risk of overfitting the data, consequently lowering the out-of-sample generalization power. To best capture the price dynamics and to avoid overfitting, an optimization of a vector of hyperparameters λ is required. Therefore, a five-fold cross-validation with random search is performed to estimate the best set of hyperparameters for each regression algorithm. This optimization is

explained in depth further in Subsection labeled "Training and bias-variance trade-off".

2

2.2.1 Chained Index

The chained hedonic approach is a series of (numeric) indices for a long sequence of periods. It is obtained by connecting price changes covering shorter time intervals. This approach controls for changes in the pool of real estate properties available in period t by specifying a "representative property". This allows the indexing process to explicitly relate property price change either to changes in the implicit prices of its characteristics or to general market conditions and omitted longitudinal variables (Geltner et al., 2017).

According to Geltner et al. (2017), the representative property can be specified relatively freely within the bounds of the sample characteristics. As a result, the chained hedonic index is a more powerful approach for exploring asset price dynamics. In this study, the representative property is the average real estate property in period t.

The chained hedonic index effectively requires separate, purely cross-sectional regression algorithms to be run on each single index period (t), hence, constituting an out-of-sample estimation. The index number formula can be defined so as to establish a Laspeyres, Paasche or Fisher index. In this research, only the Paasche index is applied.

The hedonic price modelling approach is susceptible to omitted variable bias or model misspecification as it is not possible to control for and observe every real estate characteristic. Therefore, the resulting price indices are constant quality with respect to the observed characteristics.[3]

The Chained Paasche Index (CP) is calculated each period by taking log sale prices Y_t and characteristics X_t in period t for training and then predicting log prices of properties transacted in period $t+1$ as if they had been transacted in period t, $\hat{Y}^t_{t+1}|(X_{t+1}, \mathscr{F}_t)$, where the superscript in $\hat{Y}^t_{t+1}$ denotes the price level date t, and the subscript denotes the transacted properties in period $t+1$, and $\mathscr{F}_t$ is the training set, given by (Y_t, X_t). All transactions in period t are used for training the model, and all transactions in period $t+1$ for out-of-sample predictions with price level date t. Note that $\hat{Y}^t_{t+1}|(X_{t+1}, \mathscr{F}_t)$ does not use time fixed effects, as the data used for estimation (training) is for a single period, t. Thus, within the training data set, time is a constant.

The average difference between observed log sale prices in period $t+1$ and its corresponding predictions in period t is an estimate of the log price change between period t

[3] Omitted variable bias arising from the lack of quality controls can be partially mitigated using other data sources such as text. See for example, Nowak and Smith (2020).

and $t+1$, denoted by δ_{t+1},

$$\hat{\delta}_{t+1} = \frac{1}{n_{t+1}} \sum_{i \in S_{t+1}} \left(Y_{i,t+1} - \hat{Y}^{t}_{i,t+1} | (X_{t+1}, \mathscr{F}_t) \right), \tag{2.1}$$

where S_{t+1} denotes the set of transactions in period $t+1$, and n_{t+1} its corresponding number.

The Paasche Single Imputation index can be calculated as

$$\mathrm{CP}_t = \mathrm{CP}_{t-1} \times \mathbf{exp}(\hat{\delta}_{t+1}). \tag{2.2}$$

As Balk et al. (2013) point out, an alternative to the single imputation approach is to use the predicted (fitted) values $\hat{Y}^{t+1}_{t+1} | (X_{t+1}, \mathscr{F}_t + 1)$ instead of observed prices Y_{t+1}. This approach is known as double imputation. The estimated log price change between period t and $t+1$ is then given by

$$\hat{\delta}_{t+1} = \frac{1}{n_{t+1}} \sum_{i \in S_{t+1}} \left(\hat{Y}^{t+1}_{i,t+1} | \mathscr{F}_{t+1} - \hat{Y}^{t}_{i} | (X_{t+1}, \mathscr{F}_t) \right). \tag{2.3}$$

It can be argued that double imputation is a better way to construct indices, as biases derived from omitted variables in the model would at least partially offset each other (Balk et al., 2013). For the present study, both single and double imputation have been investigated. For simplicity, this paper focuses on the single imputation approach. Brief results of double imputation are presented in the Subsection labeled "Double and single imputation".

Additionally, we allow the model to predict property prices in period t to be estimated on *multiple* periods up to and including period $t-1$. We test two variants of the CP model: (1) Rolling Window (RW) and (2) Expanding Window (EW) models. Both the rolling and expanding window variants have the benefit that they use more observations to train the model. For each additional period in the training window, a time dummy variable is included in the model, increasing the model complexity (dimensionality). The window approach has the objective of analyzing the trade-off between model complexity and accuracy.

RW models are estimated based on all observations in periods $t-1$ and t (a 2-period window), so $\mathscr{F}_t = (y_t, X_t), (y_{t-1}, X_{t-1})$, and then predict the prices of properties sold in period $t+1$ as if they had been sold in t. EW models are estimated based on all observations available from the first period in the time series (2000) until period t and then predict the prices of real estate sold in period $t+1$ as if they had been sold in period t. In this approach, the extension of the training window varies (expands) over time, so $\mathscr{F}_t = (y_t, X_t), (y_{t-1}, X_{t-1}), \ldots, (y_1, X_1)$.

2.2.2 Machine Learning Algorithms and comparison metrics

There are several studies in the real estate literature that implement machine learning algorithms. For example, Kok et al. (2017) use machine learning to construct an AVM. Most of these studies benchmark traditional hedonic pricing models to Artificial Neural Networks (Do and Grudnitski, 1992; Evans et al., 1992; McGreal et al., 1998; Nghiep and Al, 2001; Peterson and Flanagan, 2009; Tay and Ho, 1992; Wong et al., 2002; Worzala et al., 1995).

As there are countless ML regression algorithms that allow us to predict the value of next period's transaction prices $\hat{Y}^t_{t+1}|(X_{t+1}, \mathscr{F}_t)$, a preselection of algorithms is necessary for practicality. The preselection phase consisted in applying several ML algorithms e.g. Lasso (Tibshirani, 1996), Random Forest (Breiman, 2001), and k-Nearest Neighbours (Altman, 1992) to the same trial dataset and selecting the four with the lowest RMSE. The algorithms which had the best performance in the preselection phase were SVR (Drucker et al., 1997), XGBT (Chen and Guestrin, 2016), avNNet (Ripley, 1996) and Cubist (Quinlan et al., 1992). These algorithms are among the most popular in the ML domain and are applied in a variety of fields.

The linear model estimated by OLS is selected as our benchmark for two reasons. First, it is the most popular regression model. Second, it is the norm for chained indices (Balk et al., 2013). SVR, XGBT, avNNet and Cubist are non-linear ML regression algorithms. Specifically, SVR is a variation of Support Vector Machine (SVM) for regression, XGBT and avNNet are algorithms that can be used for both regression and classification, while Cubist is an algorithm used exclusively for regression.

The estimated indices are assessed on two criteria: Regression model accuracy and index quality. RMSE and (adjusted) R^2 are among the most popular metrics for measuring the performance of regression algorithms. According to Steurer et al. (2021), these measures are used by several researchers for evaluating regression algorithms in the AVM field. Hence, these are the selected metrics for assessing the regression algorithms that form the indices.

For assessing the quality of price indices, Guo et al. (2014) argue that volatility and first-order autocorrelation (ACF1) of the index returns are more appropriate measures. Volatility is defined as the standard deviation of the index returns. The authors argue that noise (cross-sectional random error) in the index adds to volatility, beyond and in addition to the true volatility. Thus, noise generates excess or erroneous volatility in the index. This excess volatility reduces the ACF1 to the point that a pure noise index would yield an ACF1 of -0.5. Hence, extreme volatility and a negative ACF1 are indicative of poor index quality. Finally, the four factors established to compare the estimated price indices are: out-of-sample RMSE, R^2, volatility and the ACF1 of the index returns.

2.2.3 Training and bias-variance trade-off

The training of a ML algorithm is marked by the trade-off between bias (central tendency) and variance (deviations from the central tendency) of the predicted variable (Geman et al., 1992; Webb, 2000). The objective is to have the least biased model with the highest generalization power. Hyperparameters are settings used to control how flexible the resulting model should be when fitting the training data. Refinements in ML implementation emphasize stable out-of-sample performance to explicitly guard against overfitting (Gu et al., 2020). Choosing the best set of hyperparameters minimizes the occurrence of overfitting and ensures out-of-sample generalization.

Cross-Validation (CV) is a sampling technique used during training for hyperparameter optimization, which results in a nearly unbiased estimator of the generalization error given a finite sample (Scholköpf and Smola, 2002). When employing k-fold CV, the lowest bias can be achieved by setting k to n, where n equals the number of observations in the training data set. This type of CV is also known as Leave One Out Cross-Validation (LOOCV), and can be performed at the cost of increasing the error variance (Hastie et al., 2009; Smola and Scholköpf, 2004). Conversely, setting k to 10 or 5 yields the best compromise between bias and variance (Breiman and Spector, 1992; Kohavi et al., 1995).

Considering the trade-off between bias and variance, in this paper we opt for performing a 5-fold CV with hyperparameter values selected via random search. This choice favors the balance between bias and variance. As a result, it cannot be asserted that all performed forward predictions are mean unbiased. Note, however, that the proposed methodology allows the choice of hyperparameter tuning to warrant either less bias or less variance.

Consider $\mathcal{G}_{\mathcal{X}}$ as a natural distribution with parameters Θ and $\mathcal{X}^{(train)}$ as an i.i.d. sample set drawn from $\mathcal{G}_{\mathcal{X}}$. The learning algorithm $\mathcal{A}$ operates as a function that maps the data set $\mathcal{X}^{(train)}$ to a function f which minimizes some expected loss $\mathcal{L}(x; f)$. Bergstra and Bengio (2012) argue that the ultimate goal of a standard learning algorithm $\mathcal{A}$ is to find the function f. A learning algorithm usually produces f through the optimization of a training criterion in connection with the set of parameters Θ, *e.g.* mean and variance of $\mathcal{G}_{\mathcal{X}}$. For the present work, the selected training criterion is RMSE.

The learning algorithm has additional features called *hyperparameters* (λ) which control the shape of function f. The actual learning algorithm is the one obtained after selecting λ, which can be denoted $\mathcal{A}$, where $f = \mathcal{A}(X^{(train)})$ for training set $X^{(train)}$. Identifying the best set of values for the hyperparameters λ is called *hyperparameter optimization*. The objective is to choose λ that minimizes generalization error, hence

$$\mathbb{E}_{x \sim \mathcal{G}_x}[\mathcal{L}(x; \mathcal{A}_\lambda(\mathcal{X}^{(train)})\,].$$

Essentially, there are no efficient algorithms to perform the optimization:

$$(\hat{Y}_{t+1}^{t} \mid X_{t+1}, \mathscr{F}_t, \hat{\lambda}_t) \tag{2.4}$$

Additionally, it is not possible to evaluate the expectation over the unknown natural distribution $\mathcal{G}_x$. Hence, cross-validation is employed to estimate the expectation of $\mathcal{G}_x$. Cross-validation replaces the expectation with the mean over a *validation set* $\mathcal{X}^{(train)}$, whose elements are drawn i.i.d $x \sim \mathcal{G}_x$.

$$\lambda^{(*)} = \underset{\lambda \in \Lambda}{\textbf{argmin}}\, \mathbb{E}_{x \sim \mathcal{G}_x}[\mathcal{L}(x; \mathcal{A}_\lambda(\mathcal{X}^{(train)})] \tag{2.5}$$

$$\approx \underset{\lambda \in \Lambda}{\textbf{argmin}} \underset{x \in \mathcal{X}^{(valid)}}{mean} \mathcal{L}(x; \mathcal{A}_\lambda(\mathcal{X}^{(train)}) \equiv \underset{\lambda \in \Lambda}{\textbf{argmin}}\, \Psi(\lambda) \tag{2.6}$$

$$\approx \underset{\lambda \in \{\lambda^{(1)} \ldots \lambda^{(S)}\}}{\textbf{argmin}}\, \Psi(\lambda) \equiv \hat{\lambda} \tag{2.7}$$

Equation (2.6) expresses the hyperparameter optimization problem in terms of a *hyperparameter response function*, Ψ. Hyperparameter optimization is the minimization of $\Psi(\lambda)$ over $\lambda \in \Lambda$. Different data sets, tasks, and learning algorithm families yield different sets Λ and functions Ψ. As the information about Ψ and Λ are scarce, the current dominant strategy for finding a good λ is to choose some number (S) of trial points $\{\lambda^{(1)}, ..., \lambda^{(S)}\}$, to evaluate $\Psi(\lambda)$ for each one, and return the $\lambda^{(i)}$ that works the best as $\hat{\lambda}$. This strategy is represented by Equation (2.7).

The trial points $\{\lambda^{(1)}, ..., \lambda^{(S)}\}$ are randomly selected, hence the name random search. Bergstra and Bengio (2012) suggest that a random search is better than other methods, such as a grid search, because it has a higher chance of finding the global optima. The ML regression algorithms explored in this research have different numbers of hyperparameters.

2.3 Data

The data used in this study consist of individual commercial real estate transactions in New York over the period 2000–2019. It is divided in five regions, four property types and four building periods. For each property, the data contain the price in USD and the area in squared feet. The data are provided by Real Capital Analytics (RCA), a well-established commercial real estate transaction dataset in the United States. Founded in 2000, they collect over 90% of all transactions of "investable" real estate by now.

Observations in the bottom and top 1% of the distribution $\textbf{log}(\text{price})/\,\textbf{log}(\text{area})$ have

been removed. The main reason for doing so is that data entry errors and outliers tend to be concentrated at the extremes of the distribution (Steurer et al., 2021). After performing the previous step and removing any observation with missing values from the original 31,727 observations, the resulting data set is composed of 29,998 observations.

The yearly summary by transaction date is available in Table 2.1. A few observations can be made from this table. First of all, the average square footage of properties sold decreased substantially in the first four years. This is mostly explained by the fact that the data collecting process improved in that period.[4] This is also reflected in the relatively low number of observations in the same period. Secondly, note that the Great Financial Crisis (GFC, 2007 – 2009) is clearly visible as well. The average transaction price was $11M during the trough of the GFC, considerably lower compared to the $27M long run average. The number of observations also fell almost by half between 2008 and 2009, falling from 1,354 to 719.

A categorical variable for building period has been created by splitting the data into quartiles of the building date. Information about the categorical variables is available in Table 2.2. Note that over 70% of the transactions happen in just Manhattan and the NYC Boroughs (Brooklyn, Queens, Bronx and Staten Island). Likewise, over 40% of the transactions are income producing apartment complexes.

2.4 Results

This section first presents chained index estimates, followed by stress tests that analyze the sensitivity of the results to the sample size. The Subsection labeled "Optimal Window Size" contrasts estimation results using rolling and expanding windows. Finally, the bias-variance trade-off confirms that double imputation is better for regression algorithms that suffer from estimation biases.

2.4.1 Chained index results

Figure 2.1 shows the yearly CP indices. It shows that SVR (Figure 2.1a) and Cubist (Figure 2.1d) indices have higher cumulative price changes compared to OLS. Also, XGBT (Figure 2.1b) and avNNet (Figure 2.1c) have almost identical results as OLS for most of the time, but are slightly lower in the last five years.

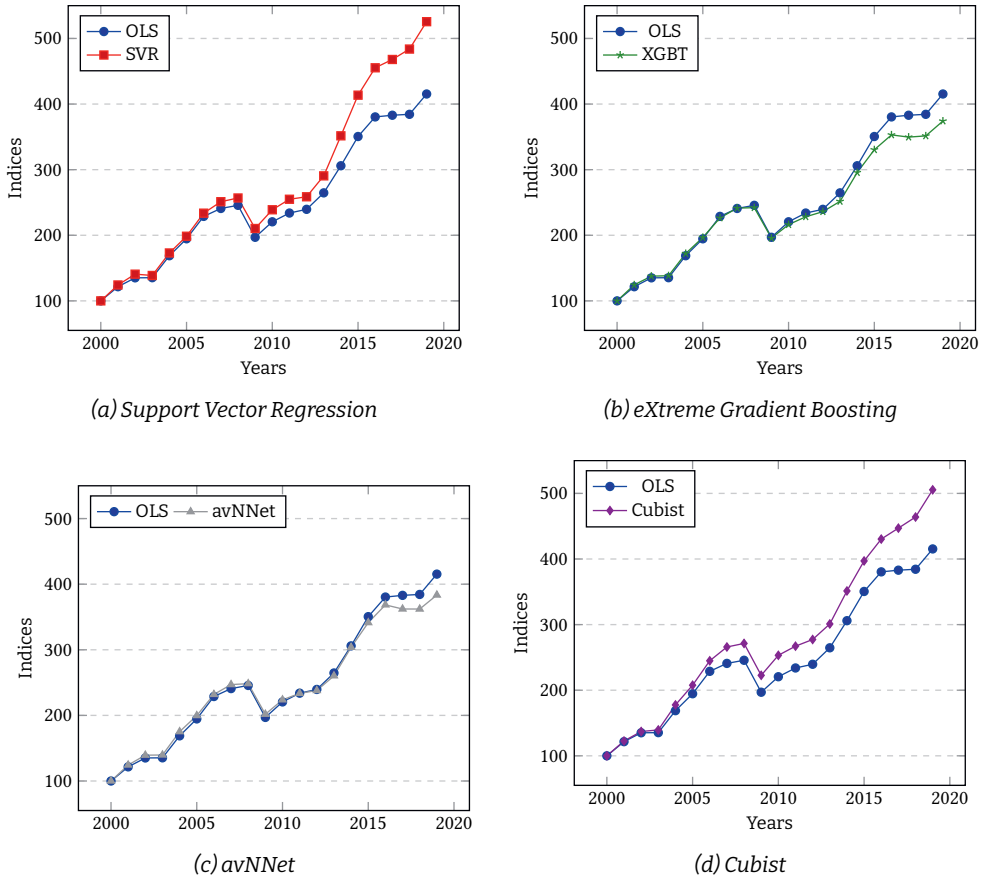
[4] Before 2005, RCA used a $5M lower boundary threshold for collecting data. They lowered this number to $2.5M in 2005. Even though they then retroactively collected transactions of properties sold for less than $5M for the period 2000–2004, it is probably still incomplete.

Year	Avg. Price	SD. Price	Avg. Area	SD. Area	Obs.
2000	37,478,974	96,366,158	178,061	323,727	313
2001	27,056,215	81,077,075	132,114	236,834	387
2002	30,258,411	83,920,959	141,908	258,308	450
2003	27,545,576	102,639,934	130,694	260,487	640
2004	23,705,006	60,436,108	113,049	211,754	1,077
2005	21,373,296	74,473,445	84,128	184,572	1,895
2006	26,404,120	159,581,210	79,109	274,189	2,164
2007	26,954,778	109,581,309	72,508	161,258	2,227
2008	22,357,677	112,259,746	63,023	134,594	1,354
2009	11,472,711	42,536,163	57,141	135,668	719
2010	25,275,415	99,605,492	87,926	200,933	892
2011	37,932,610	140,323,829	96,011	239,206	1,149
2012	23,290,705	81,586,946	62,811	140,172	1,907
2013	25,782,133	126,978,651	66,192	157,284	2,311
2014	29,147,148	127,380,169	68,031	166,633	2,429
2015	33,435,428	173,464,626	68,909	254,577	2,681
2016	35,058,311	158,851,647	74,161	200,809	2,205
2017	26,996,847	121,011,114	72,363	175,786	1,927
2018	29,107,151	130,042,150	63,662	144,370	2,082
2019	27,169,438	99,302,021	66,360	131,411	1,189

Table 2.1: New York sample summary statistics. *Avg. Price* = Average Price in US Dollars; *SD. Price* = Price Standard Deviation in US Dollars; *Avg. Area* = Average Area in Square Feet; *SD. Area* = Area Standard Deviation in Square Feet; *Obs.* = Number of Observations.

Regions	Obs.	Property Type	Obs.	Building Period	Obs.
Long Island	2,014	Apartment	12,265	1 - 1732 to 1920	9,313
Manhattan	10,465	Industrial	4,848	2 - 1921 to 1931	7,057
New Jersey	4,418	Office	6,196	3 - 1932 to 1970	6,174
NYC Boroughs	11,035	Retail	6,689	4 - 1971 to 2019	7,454
Stamford	969				
Westchester	1,097				
Total			29,998		

Table 2.2: Categorical variables summary. *Obs.* = Number of Observations.



(a) Support Vector Regression

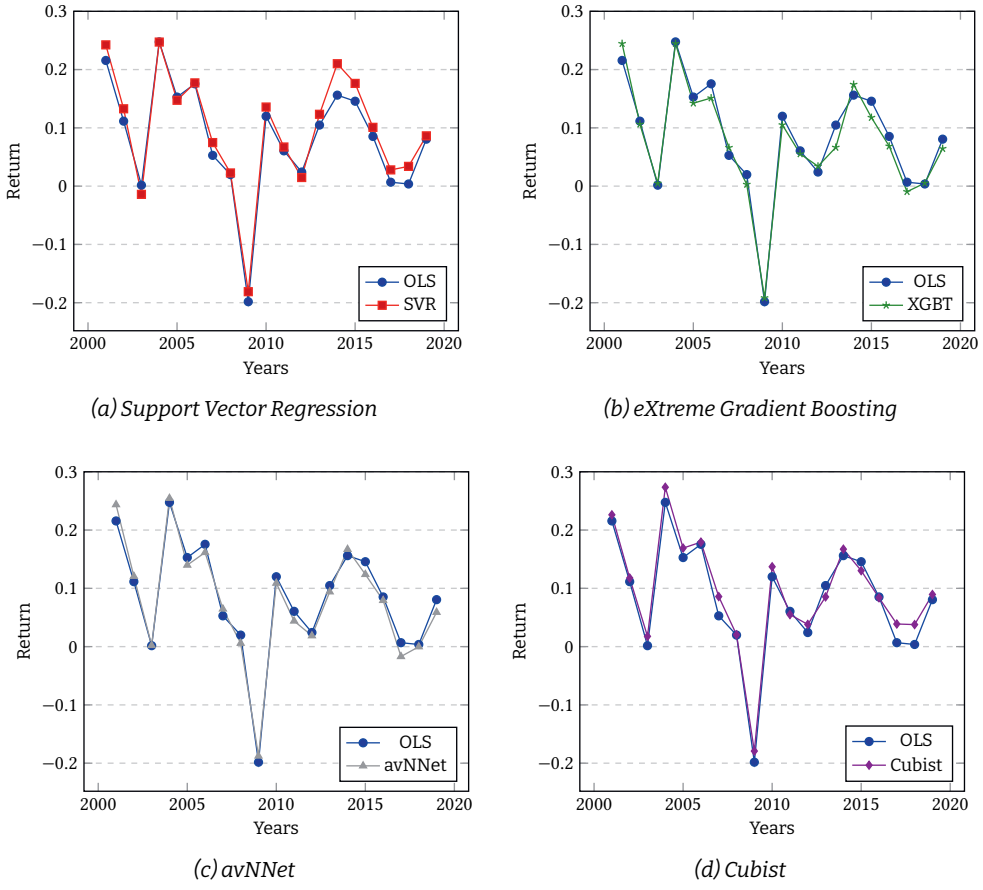
(b) eXtreme Gradient Boosting

(c) avNNet

(d) Cubist

Figure 2.1: New York Chained Paasche Price Indices. *OLS* = Ordinary Least Squares; *XGBT* = Extreme Gradient Boosting Tree; *SVR* = Support Vector Regression; *avNNet* = Neural Networks Using Model Averaging.

Figure 2.2 shows the yearly CP index returns. Despite the fact that the indices are different in levels (see Figure 2.1), these differences is not that evident when looking at the index returns (first differences). Among all the regression algorithms, the most notable differences in the returns are displayed by SVR (Figure 2.2a) and Cubist (Figure 2.2d), both of which yield overall higher returns than OLS.



(a) Support Vector Regression

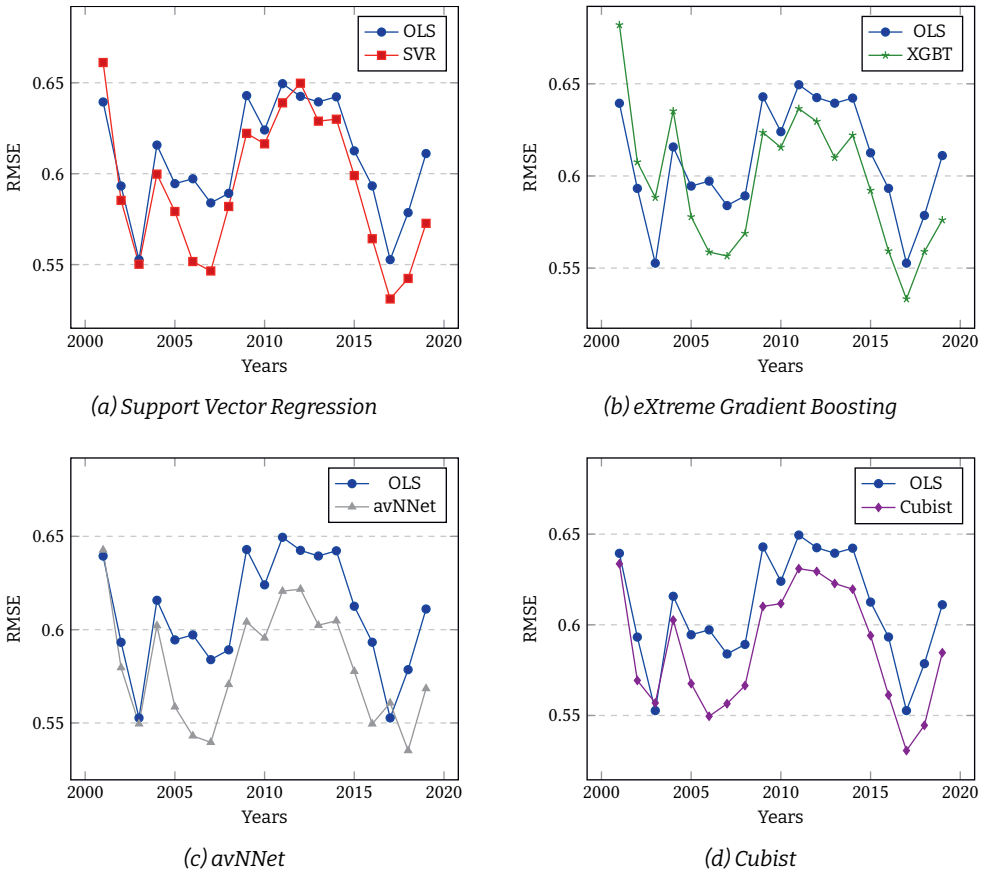
(b) eXtreme Gradient Boosting

(c) avNNet

(d) Cubist

Figure 2.2: New York Chained Paasche Price Returns. *OLS* = Ordinary Least Squares; *XGBT* = Extreme Gradient Boosting Tree; *SVR* = Support Vector Regression; *avNNet* = Neural Networks Using Model Averaging.

Figure 2.3 and Table 2.3 show that overall, the non-linear models have a lower RMSE and a higher R^2 when compared to OLS. This is expected, as non-linear models are more flexible and usually yield a better fit to the data, especially when the data is complex and sparse as is the case for real estate. Table 2.3 shows that for all the regression algorithms, the index returns have a non-negative first-order autocorrelation and do not exhibit extreme volatility.



(a) Support Vector Regression

(b) eXtreme Gradient Boosting

(c) avNNet

(d) Cubist

Figure 2.3: New York Chained Paasche Indices RMSEs. *OLS* = Ordinary Least Squares; *XGBT* = Extreme Gradient Boosting Tree; *SVR* = Support Vector Regression; *avNNet* = Neural Networks Using Model Averaging; *RMSE* = Root Mean Squared Error.

Since the commercial real estate market in New York is distinct for each property type, an index per property type has been calculated. Appendix Figure 2.A1.1a exhibits the index for apartments, Appendix Figure 2.A1.1b for retail, Appendix Figure 2.A1.1c for offices and Appendix Figure 2.A1.1d for industrial. Appendix Figures 2.A1.2a, 2.A1.2b, 2.A1.2c and 2.A1.2d present the respective RMSE values for the underlying algorithms.

	RMSE	Residual Mean	Rsquared	Volatility	ACF(1)
OLS	0.6081	0.0749	0.7473	0.0991	0.1360
XGBT	0.5964	0.0694	0.7588	0.0992	0.1693
SVR	0.5922	0.0873	0.7610	0.1020	0.1169
avNNet	0.5803	0.0707	0.7709	0.1012	0.1719
Cubist	0.5864	0.0853	0.7657	0.0968	0.1317

Table 2.3: Chained Paasche performance summary. *OLS* = Ordinary Least Squares; *XGBT* = Extreme Gradient Boosting Tree; *SVR* = Support Vector Regression; *avNNet* = Neural Networks Using Model Averaging; *RMSE* = Root Mean Squared Error; *Volatility* = Index Return Standard Deviation; *ACF(1)* = Index Return First-Order Autocorrelation.

As the regression algorithms used to estimate the property type indices were trained on a restricted data set, the RMSE for the ML algorithms are higher when compared to the indices with the full sample. Apartment is the most prevalent property type. Therefore, the ML algorithms show a lower RMSE than the benchmark, with the only exception being SVR. XGBT and avNNet produce an index almost identical to the one generated by the benchmark, while SVR and Cubist have much higher values.

Industrial properties are the type with the fewest transactions. The RMSE for this category is around the same level for all the algorithms used, with ML being below the benchmark in some years and above in others. The variation of the RMSE values is higher for the ML algorithms when compared to the benchmark. For this property type, XGBT has a very similar index to the benchmark, whereas SVR and Cubist have much higher index values. avNNet is the only regression algorithm that exhibits an index with a lower value than the benchmark.

Office and Retail have similar numbers of transactions and the RMSE of the ML algorithms is usually higher than the benchmark. SVR, XGBT and avNNet generated indices with values lower than the benchmark. Cubist produces an index that has lower values than the benchmark up until 2015, but after that the index values are higher.

The restriction per property type imposed in the training sample limits the performance of the ML algorithms. This is more evident after examining the RMSE plots, which show no conclusive improvement when compared to the benchmark.

2.4.2 Stress Test

In order to check model and index stability, a stress test has been performed by sampling 50%, 25%, 10% and 5% of the available data, 30 times for each percentage level, so in total 120 samples. An index has been generated for each sample.

Figure 2.4 shows the average index return for each percentage level. Interestingly, all indices exhibit a higher volatility at the 5% level, detaching from the other percentage levels

in years 2001, 2004 and 2011. Table 2.4 presents summary statistics of the volatility (standard deviation of the index returns) per percentage level. All ML algorithms have a similar behaviour, displaying almost no noticeable change in the index until the 5% mark. Only at the 5% level is there a significant index return change, with the index return exhibiting a higher volatility when compared to the full sample. It is possible to corroborate the results in Figure 2.4 by examining the mean volatility of the 30 samples in Table 2.4, where, only at the 5% level, a significant increase in volatility of the returns is noticeable for all regression algorithms.

Figure 2.5 presents the average RMSE over all stress tests performed. The RMSE for the ML algorithms is noticeably more sensitive to data availability. Nevertheless, for several percentage levels, the RMSE for the ML algorithms is lower than the OLS benchmark. Starting from the 10% level, the RMSE of the ML algorithms increases, decoupling from the previous levels. At the 5% level all ML algorithms display a higher RMSE than the OLS benchmark.

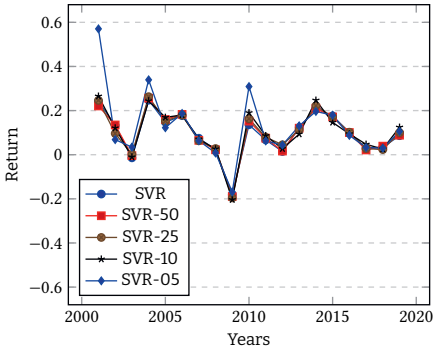
Table 2.4 provides insight in how "stable" the index returns remain when using less observations (see Francke and Van de Minne, 2017, who employ a similar stress test). On average, the mean of the volatility of the returns for OLS is lower than the ML algorithms for all suppressed percentage levels, except for the 5% level where XGBT and Cubist have a lower average return volatility. A common trend among all regression algorithms is that the volatility of the returns increases as more data are suppressed. The standard deviation of the return volatility also increases when more data are removed. Only at the 50% and 10% levels is the standard deviation of the return volatility from the benchmark lower than the ML algorithms.

2.4.3 Optimal Window Size

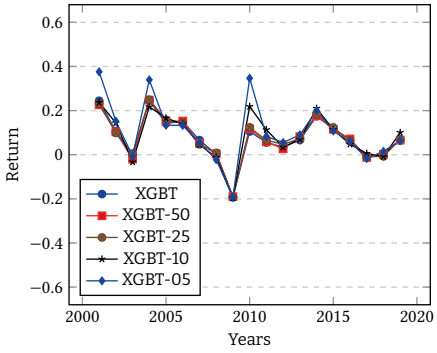
This subsection presents model results from both rolling (RW) and expanding windows (EW) samples. The main motivation for investigating alternative samples is to account for the trade-off between model complexity and out-of-sample accuracy.

A larger time window size is associated with more observations and more year control variables, thus adding more dimensions and complexity to the models. Up to an Optimal Window Size (OWS), the addition of an extra year is expected to have an overall benefit of increasing the out-of-sample accuracy, as the model fits the data better and becomes more generalizable. The balance between model complexity and out-of-sample accuracy can be explored by testing different window sizes and measuring the impact in terms of out-of-sample RMSE.

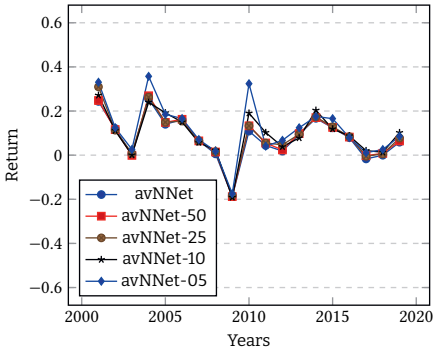
In our search for the optimal window size, we test windows of 2 to 8 years. Window sizes



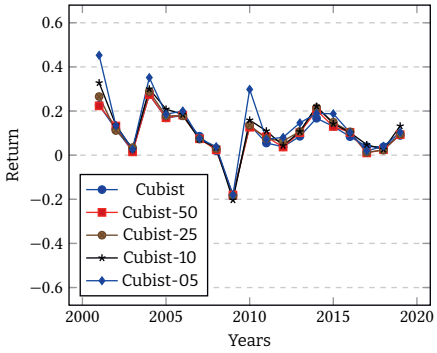
(a) Support Vector Regression



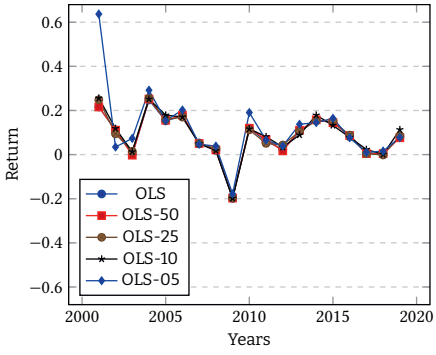
(b) eXtreme Gradient Boosting



(c) avNNet

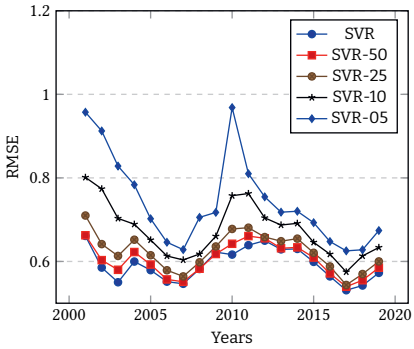


(d) Cubist

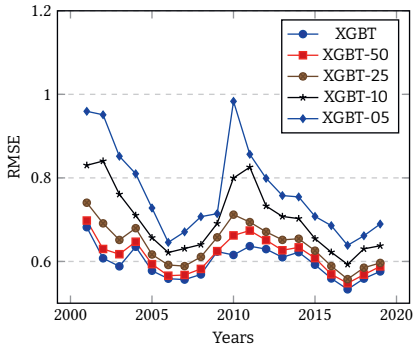


(e) OLS

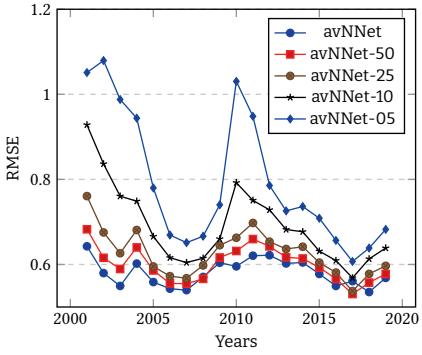
Figure 2.4: Returns for Stress Tests. All the indices returns are similar until the 10% level. At the 5% level, indices returns exhibit a higher volatility, detaching from the other percentage levels in years 2001, 2004 and 2011. Numbers refer to the percentage of the sample size being used: 50, 25, 10 and 5 percent. When omitted the full sample has been used.



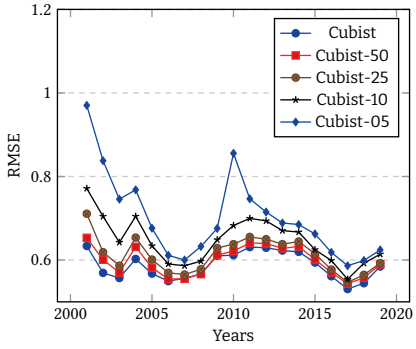
(a) Support Vector Regression



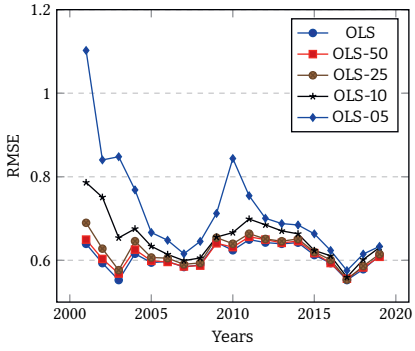
(b) eXtreme Gradient Boosting



(c) avNNet



(d) Cubist



(e) OLS

Figure 2.5: RMSEs for Stress Tests. The RMSE for the ML algorithms is noticeably more sensitive to data availability as it is higher then OLS. Starting from the 10% level, the RMSE of the ML algorithms increases, decoupling from the previous levels. At the 5% level all ML algorithms display a higher RMSE than the OLS benchmark. Numbers refer to the percentage of the sample size being used: 50, 25, 10 and 5 percent. When omitted the full sample has been used.

larger than 8 years would hamper this exercise by being too limiting on the data availability for training. The test data used for the different window sizes must be identical to allow for a fair comparison of results. Therefore, for each window size m, the regression algorithms have been trained on data in the years $t - m + 1$ to t and predicted for year $t + 1$, where $t + 1 = 2008, \ldots, 2019$.

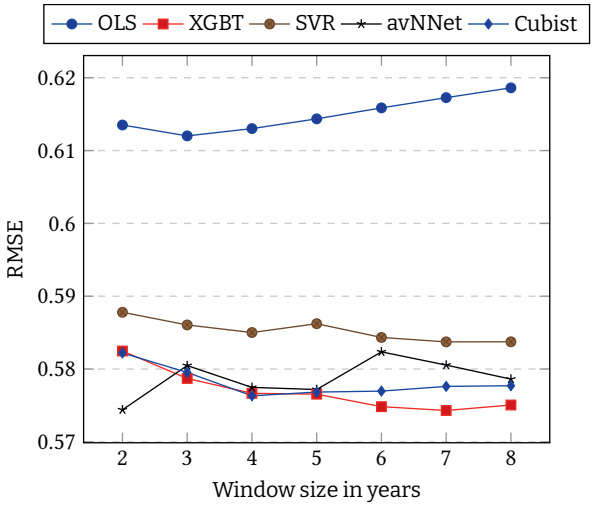


Figure 2.6: RMSEs per Window Size. Each regression algorithm behaves differently, depending on the training window size. The best performing algorithms are avNNet with a window size of 2 years and XGBT with a window size of 7 years; *RMSE* = Root Mean Squared Error.

Figure 2.6 presents out-of-sample RMSEs for the different algorithms and window sizes. The figure shows that each regression algorithm behaves differently, depending on the training window size. This shows that the out-of-sample model performance and corresponding indices are quite sensitive to the window length. Table 2.5 provides model and index performance statistics for the optimal window size for each algorithm. Table 2.6 presents the results for the EW approach, which has a lower performance when compared to RW. The results for RW and EW suggest that the optimal window size is bigger than two years and smaller than using all years as in EW.The best performing algorithms are avNNet with a window size of 2 years and XGBT with a window size of 7 years.

2.4.4 Double and single imputation

When comparing the results obtained using single and double imputation (Figure 2.7) one can notice a significant cumulative difference of single minus double imputation for the ML algorithms. SVR and Cubist have the highest cumulative difference. Conversely, avNNet shows a moderate cumulative difference, while XGBT is the only ML algorithm that

exhibits a negative difference between indices using the true or fitted values. OLS shows no cumulative difference because the fitted mean is equal to the true mean by definition (when a constant is included).

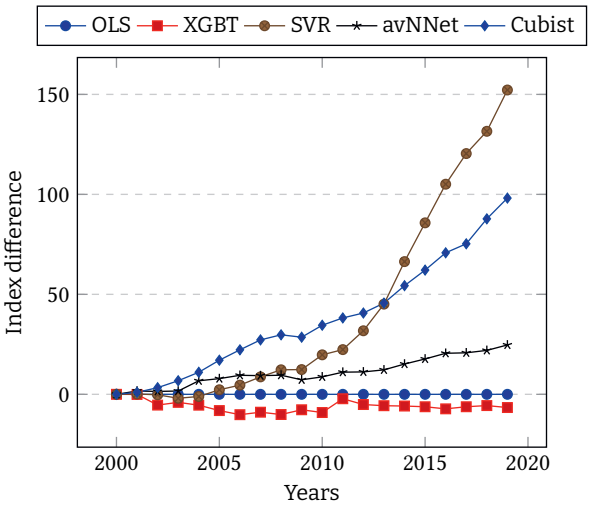


Figure 2.7: Cumulative difference between single and double imputation. All ML algorithms exhibit a difference between single and double imputation, SVR and Cubist have the highest cumulative difference; *Index difference* = Single minus double imputation difference.

As presented in the Subsection labeled "Training and bias-variance trade-off ", there is a trade-off between bias and variance that should be taken into consideration when selecting and evaluating ML algorithms. Algorithms with higher fitted variance have a tendency to overfit, which implies that the mean of the fitted values will be equal or virtually equal to the mean of the true (observed) values as the bias is close to zero.

Table 2.7 shows the difference between true mean log price and fitted mean log price. All ML algorithms display values different from zero, which indicates bias. On the other hand, Table 2.8 shows that all ML algorithms have a lower difference between true and fitted variance. All ML algorithms have a higher fitted variance when compared to the benchmark. Hence, Tables 2.7 and 2.8 display and quantify the bias-variance trade-off for this application.

The double imputation approach is preferred for the ML algorithms as the bias in the fitted value $(\hat{Y}_{t+1}^{t+1}|\mathscr{F}_{t+1})$ and the Out-of-Time prediction $(\hat{Y}^{t}|(X_{t+1}, \mathscr{F}_{t}))$ will partially offset each other. Hill and Melser (2008) also suggest double imputation for similar reasons. For the purpose of price index construction, it is necessary to be aware of the estimation bias, as this can be transferred into the index. A biased model can potentially produce an index that over or under estimates long-term price trends, especially over long time periods, as

the index is built using prior index values and the bias will accumulate.

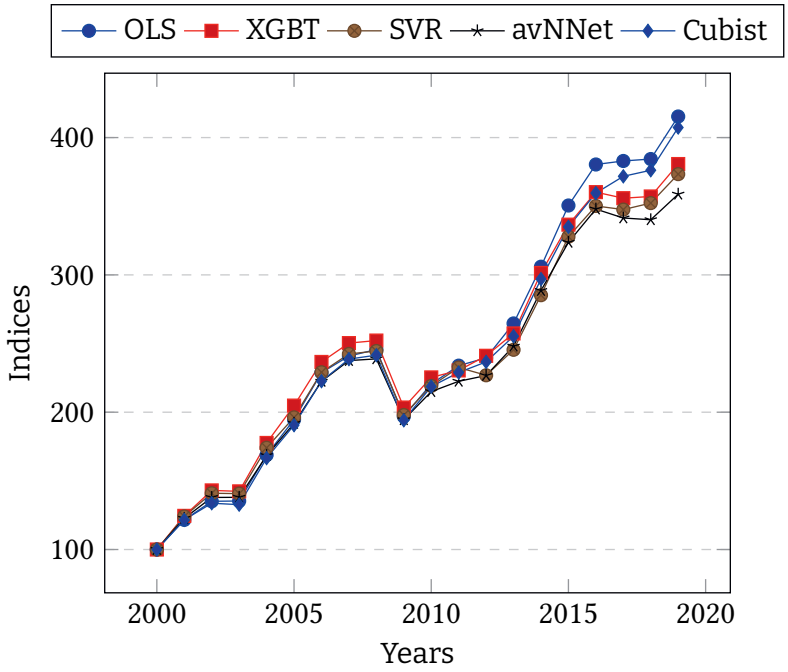


Figure 2.8: New York Chained Paasche Price Indices with Double Imputation. Starting from year 2013 all the indices built using the ML regression algorithms are below the benchmark, different from the single imputation approach, where SVR and Cubist had all their values above the benchmark.

Figure 2.8 shows the price indices for the New York commercial real estate market using double imputation. Notice that after the year 2013 all the indices built using the ML regression algorithms are below the benchmark, different from the single imputation approach, where SVR and Cubist had all their values above the benchmark.

Additionally, one of the central assumptions of this paper is that the difference of the means is the periodic price change. With a biased estimator this difference will be composed of the periodic price change plus bias. Hence, when using the proposed methodology, the bias-variance trade-off should be taken into consideration as well as the measures to attenuate or eliminate the estimation bias. As mentioned in the Subsection labeled "Training and bias-variance trade-off ", in the context of this paper, increasing the number of folds during training might be a solution to mitigate estimation bias during training.

As a further check, Welch's t-tests were performed in all actual/fitted price pairs to test the hypothesis that the two populations have equal mean. The test results, such as t-values, p-values and confidence intervals can be seen in the Appendix labeled "Welch's t-Test Results". Note that the algorithms which exhibit a higher difference between single and dou-

ble imputation indices are the ones with lower p-values on average. However, with a 95% confidence interval, none of the tests rejected the null hypothesis. This result indicates the likelihood that the two populations have equal means.

2.5 Conclusion

This paper presents a model-agnostic approach based on Out-of-Time individual transaction predictions to build price indices for commercial real estate using a variety of non-linear machine learning (ML) algorithms. The key innovation is the use of prediction error to measure time trends. The results obtained support the viability of using ML for constructing price indices. Overall, the non-linear ML algorithms yielded higher accuracy and lower volatility with non-negative first-order autocorrelation of index returns.

The comparison between single and double imputation shows that some of the ML algorithms have display estimation bias. Using the proposed methodology for index construction requires attention to the bias and variance trade-off. The findings also highlight the importance of the hyperparameter selection phase in minimizing the introduction of bias while keeping the out-of-sample generalization power. Regression algorithms that exhibit estimation bias could use the double imputation method as a straightforward way to reduce the bias problem.

The stress tests show that linear models (OLS) generate overall more stable indices when few training data are available than the non-linear ML regression algorithms used in this paper, as linear models are less dependent on the number of observations. Also, looking at the index volatility in the stress test, OLS has lower values, on average, than the ML algorithms. Additionally, the variations of the loss function across the tests with 50, 25, 10 an 5 percent of the data are higher for the ML algorithms, especially at the 5% level. The RMSE from the property type indices corroborate the idea that ML regression algorithms are more dependent on sample size. These indices were generated using regression algorithms estimated using only the property type sample. Data sets composed of property types like apartment that contain more observations produce lower RMSE when compared to the other types with fewer observations, such as industrial real estate.

The analysis of the optimal window size for the Rolling Window (RW) approach demonstrates that the magnitude of the optimal window varies greatly across the different algorithms; the window sizes range from 2 to 8 years. The single year window corresponds to the Chained Paasche index (CP) and is ruled out, as the RMSE is higher than both the RW and the Expanding Window (EW).

Considering all the tests performed in this study, it is possible to conclude that in cases

where more observations are available, even at the cost of adding more dimensions (controls or features), ML algorithms tend to produce better results than OLS. Cases where few observations or characteristics are accessible favor OLS, as it performs better in restricted data sets when compared to ML.

		50%	25%	10%	5%
OLS	Min.	0.0885	0.0915	0.0970	0.1422
	1st Qu.	0.1009	0.1025	0.1312	0.1947
	Median	0.1054	0.1106	0.1495	0.2251
	Mean	0.1046	0.1138	0.1466	0.2471
	3rd Qu.	0.1079	0.1179	0.1536	0.2623
	Max.	0.1239	0.1679	0.1836	0.5313
	SD.	0.0071	0.0171	0.0220	0.0975
XGBT	Min.	0.0900	0.0887	0.1171	0.1403
	1st Qu.	0.0975	0.1060	0.1387	0.1871
	Median	0.1066	0.1120	0.1486	0.2187
	Mean	0.1053	0.1144	0.1581	0.2258
	3rd Qu.	0.1102	0.1190	0.1696	0.2417
	Max.	0.1224	0.1574	0.2369	0.4328
	SD.	0.0089	0.0139	0.0316	0.0629
SVR	Min.	0.0925	0.0930	0.0973	0.1324
	1st Qu.	0.1023	0.1079	0.1396	0.2038
	Median	0.1077	0.1198	0.1610	0.2211
	Mean	0.1081	0.1194	0.1583	0.2621
	3rd Qu.	0.1128	0.1287	0.1729	0.2681
	Max.	0.127	0.1610	0.2112	1.0164
	SD.	0.0082	0.0147	0.0275	0.1546
avNNet	Min.	0.0856	0.0898	0.1010	0.1466
	1st Qu.	0.1015	0.1114	0.1331	0.2086
	Median	0.1064	0.1179	0.1539	0.2468
	Mean	0.1079	0.1218	0.1544	0.253
	3rd Qu.	0.1109	0.1295	0.1672	0.2896
	Max.	0.1283	0.1894	0.2401	0.3852
	SD.	0.0103	0.0198	0.0308	0.0573
Cubist	Min.	0.0846	0.0903	0.1180	0.1430
	1st Qu.	0.1014	0.1080	0.1342	0.1819
	Median	0.1057	0.1167	0.1534	0.2261
	Mean	0.1055	0.1174	0.1573	0.2281
	3rd Qu.	0.1120	0.1255	0.1846	0.2642
	Max.	0.1206	0.1445	0.2024	0.3789
	SD.	0.0089	0.0132	0.0259	0.0576

Table 2.4: Summary statistics of the volatility from the 30 samples with 50%, 25%, 10% and 5% of the available observations. Index return volatility is, on average, lower for OLS when compared to ML algorithms. Also, only at the 5% level a significant increase in index return volatility is noticeable for all regression algorithms; *Min.* = Minimum volatility; *1st Qu.* = First Quartile volatility; *3rd Qu.* = Third Quartile volatility; *Max.* = Maximum volatility; *SD.* = Standard Deviation of index volatility.

	RMSE	Residual Mean	R^2	Volatility	ACF(1)	OWS
OLS	0.6120	0.0432	0.7283	0.0913	0.2203	3
XGBT	0.5751	0.0454	0.7600	0.0901	0.1678	7
SVR	0.5837	0.0567	0.7526	0.1024	0.2062	7
avNNet	0.5744	0.0413	0.7605	0.0892	0.1584	2
Cubist	0.5763	0.0651	0.7590	0.0951	0.3031	4

Table 2.5: Optimal window performance summary. ML algorithms display a lower RMSE then the OLS benchmark. The lowest RMSE pertains to avNNet with a window size of 2 years; *RMSE* = Root Mean Squared Error; *ACF(1)* = Index Return First-Order Autocorrelation; *Volatility* = Index Return Standard Deviation; *OWS* = Optimal Window Size in Years.

	RMSE	Residual Mean	R^2	Volatility	ACF(1)
OLS	0.6186	0.0414	0.7226	0.0914	0.2092
XGBT	0.5761	0.0376	0.7588	0.0898	0.2405
SVR	0.5852	0.0593	0.7513	0.1010	0.2110
avNNet	0.5791	0.0416	0.7560	0.0914	0.1633
Cubist	0.5773	0.0594	0.7584	0.0926	0.3182

Table 2.6: Expanding window performance summary. Expanding window performed worst than the optimal rolling window or using only one year as in the chained index methodology. XGBT regression algorithm has the best performance for the expanding window methodology; *RMSE* = Root Mean Squared Error; *ACF(1)* = Index Return First-Order Autocorrelation; *Volatility* = Index Return Standard Deviation.

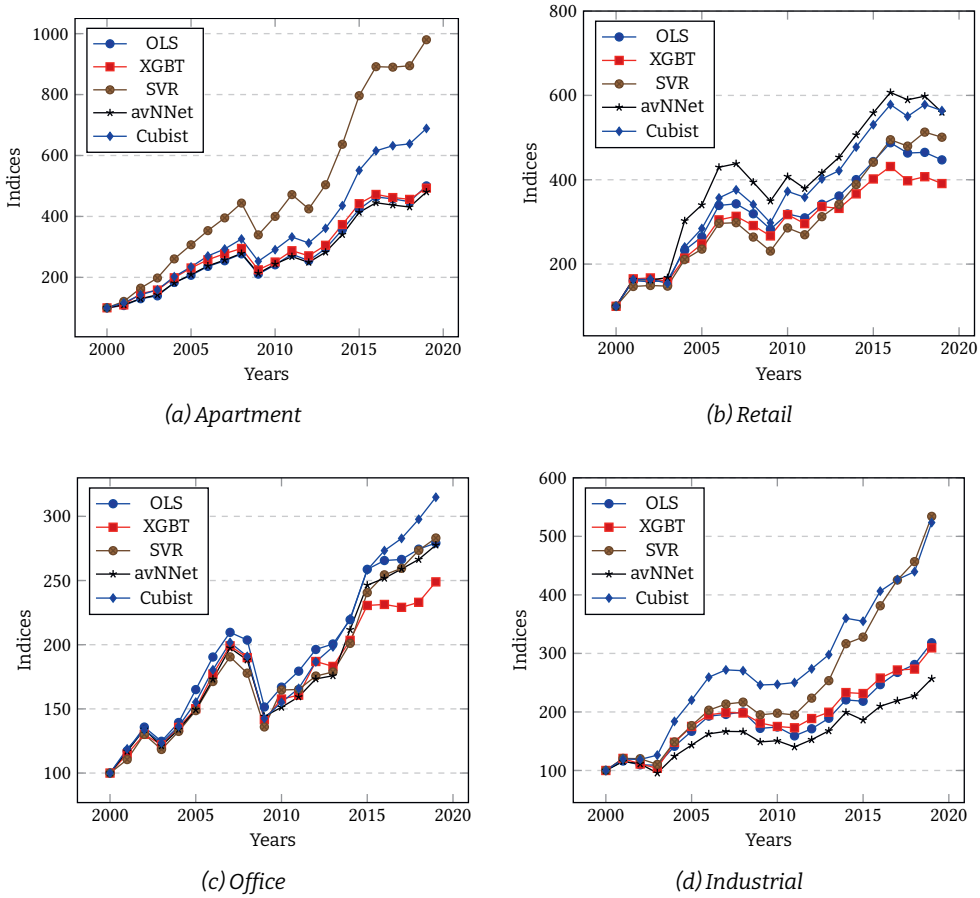
Year	OLS	XGBT	SVR	avNNet	Cubist
2001	0.0000	-0.0056	0.0014	0.0134	0.0081
2002	0.0000	-0.0410	-0.0032	-0.0029	0.0165
2003	0.0000	0.0017	-0.0123	0.0013	0.0254
2004	0.0000	0.0026	0.0078	0.0276	0.0141
2005	0.0000	0.0023	0.0178	0.0007	0.0217
2006	0.0000	-0.0013	0.0083	0.0017	0.0095
2007	0.0000	-0.0011	0.0155	-0.0033	0.0127
2008	0.0000	-0.0017	0.0136	0.0007	0.0082
2009	0.0000	-0.0006	0.0117	-0.0026	0.0211
2010	0.0000	0.0037	0.0257	0.0032	0.0094
2011	0.0000	-0.0021	0.0055	0.0087	0.0077
2012	0.0000	-0.0002	0.0393	-0.0003	0.0042
2013	0.0000	0.0006	0.0376	-0.0001	0.0058
2014	0.0000	-0.0001	0.0405	0.0030	0.0037
2015	0.0000	0.0000	0.0232	0.0020	0.0021
2016	0.0000	0.0006	0.0300	0.0041	0.0095
2017	0.0000	-0.0015	0.0350	0.0017	0.0047
2018	0.0000	0.0002	0.0198	0.0036	0.0253
2019	0.0000	-0.0054	0.0246	0.0040	0.0062

Table 2.7: Difference between true and fitted log price means. The increase in variance of the fitted values for the ML regression algorithms comes at the cost of having some estimation bias. The OLS benchmark has no estimation bias.

Year	OLS	XGBT	SVR	avNNet	Cubist
2001	0.3302	0.3768	0.2315	0.3271	0.3162
2002	0.3127	0.2053	0.3234	0.2919	0.2467
2003	0.2954	0.3239	0.2296	0.2711	0.2838
2004	0.3136	0.2865	0.2683	0.5195	0.2838
2005	0.3242	0.2838	0.2382	0.2833	0.2740
2006	0.3206	0.2801	0.1976	0.2747	0.2151
2007	0.3333	0.2633	0.2489	0.2936	0.2638
2008	0.3252	0.2648	0.2773	0.6765	0.2581
2009	0.3323	0.3042	0.3230	0.2880	0.3039
2010	0.3566	0.3263	0.3533	0.3736	0.2768
2011	0.3885	0.3550	0.3367	0.3377	0.3387
2012	0.3922	0.4243	0.3393	0.3630	0.3120
2013	0.3903	0.3517	0.3345	0.3558	0.3095
2014	0.3812	0.3640	0.3025	0.3151	0.2741
2015	0.3466	0.3034	0.2993	0.3414	0.2864
2016	0.3349	0.3574	0.2761	0.3023	0.2441
2017	0.2988	0.3472	0.2528	0.2737	0.2147
2018	0.3218	0.3190	0.2458	0.2693	0.2740
2019	0.3570	0.2638	0.2645	0.3040	0.2175

Table 2.8: Difference between true and fitted log price variance. All the ML regression algorithms present a fitted variance closer to the true variance when compared to the OLS benchmark.

2.A1 New York Indices and Errors per Property Type



(a) Apartment

(b) Retail

(c) Office

(d) Industrial

Figure 2.A1.1: New York Chained Paasche Price Indices per Property Type. *OLS* = Ordinary Least Squares; *XGBT* = Extreme Gradient Boosting Tree; *SVR* = Support Vector Regression; *avNNet* = Neural Networks Using Model Averaging.

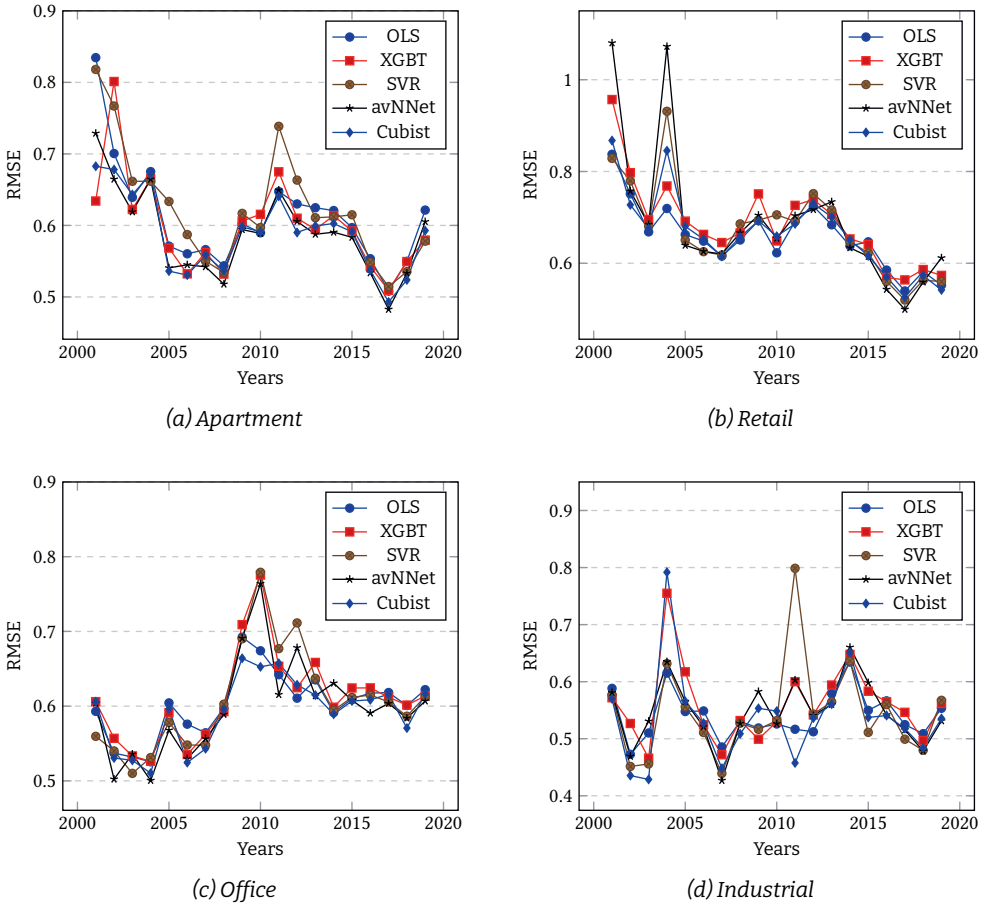


Figure 2.A1.2: New York Apartment Chained Paasche Indices per Property Type RMSEs. *OLS* = Ordinary Least Squares; *XGBT* = Extreme Gradient Boosting Tree; *SVR* = Support Vector Regression; *avNNet* = Neural Networks Using Model Averaging; *RMSE* = Root Mean Squared Error.

2.A2 Welch's *t*-Test Results

Sample Year	t-value	p-value	LBCI	UBCI
2000	-0.0358	0.9715	-0.2644	0.2549
2001	0.0149	0.9881	-0.1908	0.1937
2002	0.0156	0.9876	-0.1839	0.1868
2003	0.0168	0.9866	-0.1454	0.1480
2004	-0.0148	0.9882	-0.1059	0.1043
2005	-0.0177	0.9859	-0.0708	0.0695
2006	0.0002	0.9999	-0.0650	0.0650
2007	0.0070	0.9944	-0.0660	0.0665
2008	-0.0245	0.9805	-0.0851	0.0830
2009	0.0072	0.9943	-0.0918	0.0924
2010	-0.0088	0.9930	-0.1054	0.1045
2011	0.0023	0.9982	-0.1067	0.1070
2012	0.0115	0.9908	-0.0698	0.0706
2013	0.0072	0.9943	-0.0630	0.0634
2014	-0.0176	0.9859	-0.0614	0.0603
2015	-0.0122	0.9902	-0.0610	0.0602
2016	-0.0233	0.9814	-0.0676	0.0660
2017	-0.0080	0.9936	-0.0690	0.0685
2018	-0.0100	0.9920	-0.0678	0.0671

Table 2.A2.1: OLS t-test Results. *LBCI* = Lower Bound of the Confidence Interval; *UBCI* = Upper Bound of the Confidence Interval.

Sample Year	t-value	p-value	LBCI	UBCI
2000	-0.0245	0.9804	-0.2626	0.2561
2001	0.0393	0.9687	-0.1852	0.1928
2002	0.0635	0.9494	-0.1801	0.1921
2003	-0.0741	0.9409	-0.1501	0.1392
2004	0.0197	0.9843	-0.1044	0.1065
2005	-0.1352	0.8925	-0.0754	0.0656
2006	-0.0252	0.9799	-0.0664	0.0647
2007	0.0414	0.9670	-0.0654	0.0682
2008	-0.0408	0.9675	-0.0843	0.0809
2009	-0.0691	0.9449	-0.0948	0.0884
2010	-0.0137	0.9891	-0.1062	0.1047
2011	0.0460	0.9633	-0.1045	0.1095
2012	0.0875	0.9303	-0.0664	0.0726
2013	0.1258	0.8999	-0.0594	0.0675
2014	0.0106	0.9916	-0.0601	0.0607
2015	-0.0582	0.9536	-0.0616	0.0580
2016	-0.0702	0.9440	-0.0687	0.0640
2017	0.0184	0.9853	-0.0681	0.0694
2018	-0.0635	0.9494	-0.0699	0.0655

Table 2.A2.2: XGBT t-test Results *LBCI* = Lower Bound of the Confidence Interval; *UBCI* = Upper Bound of the Confidence Interval.

Sample Year	t-value	p-value	LBCI	UBCI
2000	-0.1645	0.8694	-0.2836	0.2397
2001	-0.1158	0.9078	-0.2037	0.1810
2002	0.1279	0.8982	-0.1724	0.1964
2003	0.0403	0.9679	-0.1430	0.1490
2004	-0.2014	0.8404	-0.1141	0.0928
2005	-0.6683	0.5040	-0.0951	0.0468
2006	-0.2633	0.7924	-0.0754	0.0575
2007	-0.4853	0.6275	-0.0840	0.0507
2008	-0.1754	0.8608	-0.0924	0.0772
2009	-0.2705	0.7868	-0.1045	0.0791
2010	-0.6156	0.5382	-0.1386	0.0724
2011	0.1647	0.8692	-0.0995	0.1177
2012	-0.8191	0.4128	-0.1010	0.0415
2013	-0.9747	0.3297	-0.0957	0.0321
2014	-1.2794	0.2008	-0.1019	0.0214
2015	-0.6992	0.4845	-0.0832	0.0394
2016	-0.9726	0.3308	-0.1009	0.0340
2017	-0.8629	0.3882	-0.0999	0.0388
2018	-0.6286	0.5296	-0.0903	0.0465

Table 2.A2.3: SVR t-test Results *LBCI* = Lower Bound of the Confidence Interval; *UBCI* = Upper Bound of the Confidence Interval.

Sample Year	t-value	p-value	LBCI	UBCI
2000	-0.0792	0.9369	-0.2716	0.2505
2001	0.0012	0.9990	-0.1925	0.1928
2002	0.0436	0.9652	-0.1817	0.1900
2003	-0.0210	0.9832	-0.1483	0.1452
2004	-0.0475	0.9621	-0.1079	0.1028
2005	-0.0156	0.9876	-0.0712	0.0701
2006	-0.1705	0.8646	-0.0714	0.0600
2007	-0.0448	0.9643	-0.0682	0.0652
2008	-0.1927	0.8472	-0.0930	0.0764
2009	-0.2213	0.8249	-0.1040	0.0829
2010	-0.0046	0.9963	-0.1054	0.1049
2011	-0.0099	0.9921	-0.1081	0.1070
2012	-0.0583	0.9535	-0.0729	0.0687
2013	0.0572	0.9544	-0.0619	0.0657
2014	0.0511	0.9592	-0.0598	0.0630
2015	0.1346	0.8929	-0.0569	0.0652
2016	-0.1186	0.9056	-0.0715	0.0633
2017	0.0599	0.9522	-0.0670	0.0712
2018	-0.1053	0.9162	-0.0716	0.0643

Table 2.A2.4: avNNet t-test Results *LBCI* = Lower Bound of the Confidence Interval; *UBCI* = Upper Bound of the Confidence Interval.

Sample Year	t-value	p-value	LBCI	UBCI
2000	-0.0069	0.9945	-0.2632	0.2613
2001	-0.1327	0.8944	-0.2057	0.1796
2002	-0.1619	0.8715	-0.2015	0.1708
2003	-0.3393	0.7344	-0.1723	0.1215
2004	-0.3779	0.7055	-0.1256	0.0850
2005	-0.6837	0.4942	-0.0958	0.0463
2006	-0.2806	0.7790	-0.0757	0.0567
2007	-0.5355	0.5924	-0.0856	0.0488
2008	-0.1580	0.8745	-0.0916	0.0779
2009	-0.4876	0.6259	-0.1158	0.0697
2010	-0.2805	0.7791	-0.1215	0.0911
2011	-0.2065	0.8364	-0.1187	0.0961
2012	-0.0990	0.9212	-0.0748	0.0676
2013	-0.0371	0.9704	-0.0654	0.0630
2014	-0.0699	0.9443	-0.0643	0.0599
2015	-0.1221	0.9028	-0.0652	0.0575
2016	-1.1266	0.2600	-0.1057	0.0286
2017	-0.9633	0.3354	-0.1031	0.0352
2018	-0.6509	0.5152	-0.0906	0.0454

Table 2.A2.5: Cubist t-test Results *LBCI* = Lower Bound of the Confidence Interval; *UBCI* = Upper Bound of the Confidence Interval.

"Il n'est pas certain que tout soit incertain."

Blaise Pascal

3

Quantifying Uncertainty in Real Estate Price Predictions: Methodological and Algorithmic Synergies

Author: Felipe D. Calainho

Original title: Quantifying Uncertainty in Real Estate Price Predictions: Methodological and Algorithmic Synergies

Abstract

Automated valuation models (AVMs) are widely used in mortgage underwriting and investment analysis, yet point predictions are often insufficient when payoffs and contracts are written on individual properties. This paper compares model-agnostic prediction-interval (PI) constructions as practical measures of property-level price prediction uncertainty. Using 86,838 Amsterdam housing transactions (2001–2021), we evaluate four PI methods—cross conformal prediction, normalized conformal prediction, direct prediction, and direct lower–upper prediction—paired with a hedonic OLS benchmark and two machine-learning regressors (XGBoost and Cubist). Regression choice mainly affects interval efficiency (average width and heterogeneity), whereas PI construction matters more for achieving nominal coverage. In split-sample tests, Cubist with direct lower–upper prediction offers the best width–coverage trade-off. A rolling-window design shows coverage can fall below nominal levels when annual training samples are small and conformal exchangeability is strained. Overall, the results clarify what practitioners can expect from common PI methods across model classes in an applied AVM setting.

3.1 Introduction

Accurate property valuation is central to real estate transactions, mortgage underwriting, and investment decisions. AVMs provide a scalable and objective way to estimate property values from observed characteristics (Lim and Bellotti, 2021). Yet point predictions alone are often insufficient for decision-making because many payoffs and constraints are written on *individual dwellings* rather than on an index, making valuation errors economically consequential. In mortgage and collateral contexts, for example, downside realizations can have asymmetric costs, and contract features can amplify the relevance of house-specific valuation errors (Bellotti, 2017; Shao et al., 2015). This motivates uncertainty quantification as a core output of modern AVMs.

Developing reliable AVMs is challenging due to the heterogeneity of housing, limited turnover, and incomplete measurement of property and locational attributes (Francke and Van de Minne, 2021b). These features generate several sources of prediction error—including measurement error, functional-form misspecification, omitted variables (unobserved heterogeneity), and transaction noise—that translate into uncertainty around predicted prices. While these statistical sources are distinct from broader macroeconomic or demand-driven uncertainty, they interact with them in practice, especially when uncertainty is consumed *ex-ante* to guide choices before outcomes are realized (Lin and Liu, 2008; Pollestad and Oust, 2025).

A useful way to connect statistical prediction uncertainty to real estate economics is through the concept of idiosyncratic housing risk. Asset-level risk in housing is often operationalized as residual dispersion after conditioning on observable characteristics, reflecting property-specific variation that is difficult to diversify and varies across segments and locations (Giacoletti, 2021; Peng and Thibodeau, 2017; Sommervoll and de Haan, 2014). Prediction intervals can be interpreted as an operational, property-level version of this residual-dispersion idea: they convert the residual distribution (or conditional residual distribution) from an AVM into a calibrated range of plausible transaction prices for a given property. This interpretation clarifies why prediction intervals are not merely technical add-ons. They provide a decision-relevant summary of the heterogeneity in valuation uncertainty across properties, which can matter for underwriting margins, risk controls, and screening rules (Pollestad and Oust, 2025).

Among the ways to report uncertainty, PIs are particularly attractive because they map naturally into decisions: at a chosen confidence level, they describe a range within which the realized transaction price is expected to fall. Interval width can be interpreted as a measure of valuation uncertainty, and asymmetry can be informative when downside and

upside risks are economically different. Despite the growing use of uncertainty quantification in real estate (Krause et al., 2020), there remains limited consensus on which practical interval-construction methods work well across different model classes and evaluation setups.

Recent studies have advanced the use of conformal prediction and related approaches for housing valuation uncertainty, including empirical applications and simulation-based comparisons (Francke et al., 2024; Hjort et al., 2023; Pollestad et al., 2024). This paper is positioned as an incremental contribution in this emerging literature. The paper focuses on a transparent, apples-to-apples comparison of commonly used, model-agnostic interval constructions across a conventional hedonic benchmark and two high-performing machine-learning regression algorithms, using the same dataset and evaluation metrics throughout. The emphasis is on clarifying how interval properties (coverage, width, heterogeneity, and asymmetry) change when the same interval logic is paired with different prediction engines, and on interpreting these differences in economically meaningful terms.

Concretely, the study evaluates four model-agnostic PI methodologies—Cross Conformal Prediction (CCP), Normalized Conformal Prediction (NCP), Direct Prediction (DP), and Direct Lower and Upper Prediction (DLUP)—combined with three regression models: OLS, XGBT, and Cubist. These regression algorithms were selected because they delivered the best predictive performance in preliminary benchmarking against additional candidates (including SVR and avNNet), while OLS serves as a conventional hedonic baseline widely used in practice. The empirical setting is a dataset of 86,838 residential transactions in Amsterdam over 2001–2021. The results show that the choice of regression model primarily affects interval efficiency (average width and heterogeneity), while the choice of interval methodology matters more for achieving nominal coverage. Across the evaluated combinations, Cubist paired with DLUP yields the most favorable trade-off between interval width and coverage in the main split-sample exercise, and rolling-window results illustrate how coverage can fall below nominal levels when annual training samples are small and when the exchangeability conditions underlying conformal calibration are strained.

By focusing on comparability across model classes and on decision-relevant interval properties, the paper contributes to ongoing efforts to make uncertainty outputs from AVMs interpretable and usable in applied real estate settings. The goal is to clarify what practitioners and researchers can (and cannot) expect from commonly used prediction-interval constructions when implemented with different regression engines, and how these outputs relate to economically meaningful notions of property-level uncertainty.

3.2 Literature review

Real estate prices are uncertain for multiple reasons, and the mechanisms that generate uncertainty differ in their economic interpretation. In this paper, *uncertainty* refers to limited predictability of future transaction prices (or cash flows) given available information, whereas *risk* refers to the exposure of households, lenders, or investors to that uncertainty through payoffs, constraints, and contract terms. These concepts are closely related in housing because most decisions and contracts are written on individual properties rather than on a market index. As a result, property-level prediction errors and their dispersion can have direct economic consequences even when aggregate price movements are well understood. This literature review places the paper's focus on *statistical uncertainty in AVM predictions* alongside related strands on housing-market uncertainty and asset-level risk, and it motivates why interval-based summaries of prediction error are a useful bridge from econometric residual variation to decision-relevant measures of property-specific price dispersion.

3.2.1 Economic Uncertainty: Market Volatility and Its Implications

Economic uncertainty, often proxied by volatility in housing and land prices, is commonly studied through real options and cyclical market dynamics. In real options settings, price volatility affects the timing of tenure choice and investment because housing is costly to trade and investments are partly irreversible. For example, Wang et al. (2020) model house price uncertainty as a stochastic process and show how volatility influences rental outcomes through timing incentives. Related work extends option-based reasoning to development and land-use decisions, emphasizing that volatility can delay or accelerate investment when adjustment is costly and irreversible (Clapp et al., 2013; Lin, 2024). More broadly, research on real estate cycles links uncertainty to feedback between demand, supply responses, and construction dynamics (Wheaton, 1999). These strands highlight macroeconomic and market-wide drivers of price uncertainty, which are conceptually distinct from (but not independent of) uncertainty at the level of individual transactions.

3.2.2 Demand Uncertainty: Imperfect Information and Search Frictions

A second source of uncertainty arises from imperfect information and search frictions, which affect listing strategies, time on market, and ultimately transaction prices. Sellers face uncertainty about buyers' valuations and the distribution of offers, and this can lead to listing price revisions and strategic timing (Knight, 2002). In matching and search frame-

works, uncertainty is generated endogenously through vacancy, turnover, and heterogeneous arrival rates of buyers and sellers, affecting price formation and market liquidity (Wheaton, 1990). These mechanisms emphasize that even with stable fundamentals, frictions and imperfect information can generate meaningful uncertainty around realized transaction prices.

3.2.3 Historical and Additional Sources of Uncertainty

Long-run evidence provides context for how uncertainty and price dynamics evolve over time. For instance, Eichholtz and Geltner (2002) document centuries of location value movements in Amsterdam, illustrating that long-run trends and regime shifts can shape the predictability of real estate outcomes. At a more structural level, small changes in turnover or vacancy can generate disproportionate price movements, contributing to volatility in transaction outcomes even when long-run anchors exist (Wheaton, 1990). A complementary perspective emphasizes thin markets, heterogeneous preferences, and market power. In this view, idiosyncratic demand and limited market depth can amplify unpredictability in rents and vacancy dynamics and can create persistent wedges between private and socially optimal outcomes (Arnott, 1989). These channels are important because they foreshadow a core practical problem for valuation: uncertainty is not only market-wide, but also property-specific, and it can be higher in markets where trading is thin or information is poor.

3.2.4 Idiosyncratic Risk and the Economic Meaning of Property-Level Prediction Uncertainty

The discussion above motivates why uncertainty is economically relevant, but it does not yet explain why a *property-level* uncertainty measure is meaningful in a price-uncertainty paper. This link is naturally made through the literature on idiosyncratic (asset-level) housing risk. In empirical housing research, idiosyncratic risk is often defined as the component of price variation that remains after conditioning on observable characteristics—that is, the dispersion of hedonic residuals or comparable measures of house-specific pricing errors (Giacoletti, 2021; Peng and Thibodeau, 2017; Sagi, 2021; Sommervoll and de Haan, 2014). This residual dispersion captures heterogeneity in outcomes that is difficult to hedge or diversify for many market participants, especially when trading is infrequent and portfolios are concentrated.

Prediction intervals from AVMs can be interpreted as an operational, property-level counterpart to this idea. Whereas the idiosyncratic-risk literature uses hedonic residual

dispersion to summarize asset-level uncertainty at the market or segment level, prediction intervals translate (conditional) residual behavior into a calibrated range of plausible prices for an individual property. This mapping is economically relevant because many contracts and decisions depend on the distribution of potential outcomes for *a particular dwelling*. Shao et al. (2015), for example, emphasizes that pricing errors matter precisely because household balance sheets and mortgage-related payoffs are defined at the level of individual houses (rather than an index), so the tails of the house-level outcome distribution can be pivotal in assessing downside exposure.

This perspective also clarifies why *ex-ante* uncertainty is distinct from (and can exceed) *ex-post* realized variation. When trading time is uncertain, measures constructed from realized transactions can understate the uncertainty faced *before* sale because the relevant exposure includes both price risk and timing risk (Lin and Liu, 2008). In applied decision contexts, the relevant object is therefore often an ex-ante assessment of uncertainty around future transaction outcomes. Recent work in real estate investment decision support similarly motivates uncertainty measures as decision inputs rather than as purely diagnostic model outputs (Pollestad and Oust, 2025). In this paper, prediction intervals are interpreted in that ex-ante spirit: they summarize the uncertainty a user faces when acting today on an AVM valuation, acknowledging that the realized transaction price is observed only later and under market conditions that may differ from the calibration sample.

Taken together, the idiosyncratic-risk literature provides a coherent economic interpretation for property-level prediction uncertainty. Wide (or asymmetric) prediction intervals can be read as indicators of greater exposure to house-specific uncertainty, consistent with the notion that residual dispersion is an economically meaningful component of housing risk (Giacoletti, 2021; Peng and Thibodeau, 2017; Sommervoll and de Haan, 2014). This interpretation is particularly relevant when uncertainty is used to set underwriting margins, screening rules, or investment criteria that are sensitive to downside realizations or asymmetric loss functions (Pollestad and Oust, 2025; Shao et al., 2015).

3.2.5 Statistical Uncertainty in AVMs: Methods for Quantifying Prediction Uncertainty

Against this broader backdrop, the present paper focuses on *statistical* uncertainty in AVMs: uncertainty arising from data limitations and model imperfections that affect the reliability of property-level price predictions. As summarized by Francke and Van de Minne (2021b), key contributors include measurement error in observed characteristics, functional-form misspecification, omitted-variable bias due to unobserved heterogeneity, and transaction noise related to bargaining and market frictions. These sources

matter because AVMs are typically trained on historical transactions and must generalize to new properties and changing market conditions.

A growing literature proposes and evaluates practical methods to quantify this statistical uncertainty. Pollestad et al. (2024) apply conformal prediction ideas to construct prediction intervals designed to achieve reliable coverage, highlighting that conventional approaches can understate uncertainty. Hjort et al. (2023) propose locally weighted conformal prediction to account for spatial heterogeneity, showing that neighborhood-level dynamics can meaningfully affect interval calibration. Bastos and Paquette (2024) emphasize that data incompleteness and market shifts influence prediction errors, motivating careful interpretation of uncertainty measures across regimes. Krause et al. (2020) distinguish error-based and model-based uncertainty measures and argue for clearer taxonomy in applied work. Finally, Francke et al. (2024) provide simulation-based evidence on how well different interval construction strategies capture uncertainty, helping to clarify strengths and limitations of commonly used, model-agnostic procedures.

The present paper contributes to this stream by emphasizing a comparative framing across model classes: it evaluates how a common set of model-agnostic prediction-interval methods behave when paired with different regression engines (including a conventional hedonic benchmark), using the same empirical setting and performance criteria throughout. This comparative emphasis is intended to clarify which aspects of interval performance are driven primarily by the interval construction itself versus by the underlying prediction model that generates the residual structure on which most model-agnostic methods rely.

3.3 Methodology

This section presents the empirical framework used to construct and compare PI for residential AVMs. The paper evaluates how widely used, *model-agnostic* PI procedures behave when paired with different regression algorithms that are standard in applied valuation. This comparative emphasis is motivated by the fact that most PI methods are built from the residual structure produced by a first-stage point predictor; consequently, the regression model can affect interval *efficiency* and size *heterogeneity*.

The analysis combines four model-agnostic PI methodologies with three regression models: a hedonic OLS benchmark and two machine-learning algorithms, XGBT (Chen and Guestrin, 2016) and Cubist (Quinlan et al., 1992). OLS is included because linear hedonic specifications remain a common reference point in both research and practice, and because it provides a transparent baseline against which more flexible learners

can be interpreted. The two machine-learning models reflect a pragmatic model set: in preliminary screening runs (not reported), these algorithms delivered the strongest predictive performance relative to alternatives considered by the author (including SVR and avNNet). The goal is therefore not to claim that the selected models span the space of AVMs, but to compare PI behavior across a conventional parametric benchmark and two high-performing, flexible predictors within a fixed empirical setting.

The four PI methodologies are model-agnostic in the sense that they do not require the regression model to be trained under a particular likelihood or distributional assumption. Instead, they use out-of-sample residuals (or transformations of residuals) from the first-stage predictor during a calibration step to construct an interval around a point prediction. This structure makes the methods portable across regression engines and matches common implementation in applied AVMs: one first produces a point valuation and then adds an uncertainty layer based on predictive errors.

In this study, all regression algorithms are trained to minimize RMSE. The PI procedures, however, may implicitly correspond to different loss functions (e.g., absolute-residual quantiles versus separate lower/upper quantiles). Importantly, a mismatch between the loss function used to fit the point predictor and the loss function implicit in the PI construction does not, by itself, invalidate the PI methods. Rather, the regression model shapes the residual distribution available for calibration, which can affect the attainable interval width, the degree of asymmetry, and the extent to which uncertainty varies across properties. Put differently, the PI method primarily targets *validity* (coverage), while the regression model largely governs *efficiency* and the economic shape of uncertainty at the property level.

Following standard desiderata for PIs, the empirical comparison centers on validity and efficiency, complemented by measures that are particularly relevant in valuation settings: (i) *asymmetry*, which is informative when under- versus overvaluation have different economic consequences (e.g., underwriting margins or downside-risk screening), and (ii) *heterogeneity*, which captures whether uncertainty varies across neighborhoods and property types rather than being imposed uniformly. The PI methods are evaluated under two complementary frameworks described in Subsection 3.3.2. The first uses a conventional random split to characterize average out-of-sample performance in a pooled sample with time controls. The second uses a rolling-window design to assess stability over time when models are trained on a single year and tested on the subsequent year, reflecting a more operational forecasting setting. Throughout, the empirical setup, feature set, models, and reported results are held fixed so that differences in PI behavior can be attributed to the interaction between PI construction and regression engine rather than to changes in the

data or specification.

While validity is the primary requirement—a $(1 - \alpha)$ interval should contain the realized transaction price with frequency close to $1 - \alpha$ in repeated samples—efficiency governs how *informative* that uncertainty statement is once coverage is fixed. Conditional on achieving the desired coverage level, narrower intervals are preferred because they provide a tighter set of plausible prices and thus sharper guidance for valuation decisions; by contrast, excessively wide intervals can be "valid" yet economically unhelpful if they accommodate outcomes that are too far apart to support underwriting, pricing, or risk screening. This highlights an important limit-case intuition. In the extreme, the narrowest conceivable interval is a zero-width interval at the point prediction. However, a point prediction cannot deliver nontrivial coverage for a continuous outcome unless the conditional distribution is essentially degenerate, which is implausible in housing markets given transaction noise, unobserved heterogeneity, and bargaining frictions. The relevant objective is therefore not to make intervals as narrow as possible in an absolute sense, but to make them as narrow as possible *subject to* meeting the coverage target. This rationale underlies reporting coverage together with mean width and proper scoring rules (such as the Winkler score), which jointly penalize miscalibration (intervals that are too narrow and undercover) and lack of informativeness (intervals that are overly wide despite meeting nominal coverage).

3.3.1 Prediction Interval Methods

The four model-agnostic methodologies share the same objective which is specified in Equation (3.1a), where $\mathbb{P}$ is the probability that the price y_{n+1} of a new observation $n + 1$ should be within the lower (l) and upper (u) bound of $PI(x_{n+1})$, with coverage probability $1 - \alpha$. The characteristics of the real estate are represented by x_{n+1}.

$$\mathbb{P}(y_{n+1} \in PI(x_{n+1})) = \tag{3.1a}$$

$$\mathbb{P}([l_{n+1} < y_{n+1} < u_{n+1}]) \geq 1 - \alpha \tag{3.1b}$$

The PI methodologies share a common two-step structure, though they differ significantly in their approach to the second step. In the first step, a regression algorithm is used to generate predictions and corresponding residuals outside the sample, which serve as the basis for constructing the PIs. These residuals are then used to estimate the uncertainty in the predictions.

In the second step, the methodologies diverge in their strategies for creating the PIs.

Some approaches employ a secondary regression algorithm to predict the lower and upper bounds of the interval. Others rely on quantile regression, which estimates specific quantiles of the outcome distribution to define the bounds. Alternatively, certain methodologies estimate a quantile of the residuals distributions. The outputs of the second step are then used to calculate the PI.

As one starts exploring the uncertainty quantification methodologies it becomes evident that there are a multitude of possible approaches for creating PIs. An innumerable amount of combinations of steps that could result in new PI methodologies to be further defined and investigated.

As discussed previously, validity and efficiency are the primary performance indicators for PI methodologies, but symmetry and heterogeneity also play a significant role in evaluating their overall performance. These criteria are the benchmarks for assessing how well a PI methodology addresses uncertainty in real estate price predictions.

Validity reflects whether the PI contains the true price of a new observation with probability $1 - \alpha$, where α represents the pre-specified significance level. This confidence level is chosen before estimating the PIs, ensuring that the intervals meet the desired coverage probability. In essence, validity ensures that the methodology produces reliable intervals that are appropriately calibrated to capture uncertainty. Efficiency, on the other hand, measures the average width of the PIs. The width of a PI serves as a proxy for the level of uncertainty in the price prediction: wider intervals indicate greater uncertainty, while narrower intervals suggest more precise predictions. Understanding this relationship is crucial because overly wide intervals may limit the practical utility of predictions, while overly narrow intervals risk being invalid.

Beyond validity and efficiency, the symmetry of the PI provides additional insights. Symmetry refers to the balance between the lower and upper bounds of the interval. Asymmetric PIs are particularly informative in real estate applications, as they address the distinct uncertainties associated with underpricing and overpricing. When the lower or upper bound is significantly larger than its counterpart, the interval reflects these uncertainties unevenly, offering a more nuanced perspective than a perfectly symmetric PI, which distributes uncertainty equally between the bounds.

Finally, heterogeneity in PI width plays an important role in understanding how uncertainty is distributed across different data points. Some methodologies, such as CCP, produce intervals of uniform width regardless of the characteristics x_{n+1}. In contrast, other methodologies allow the width of the PI to vary based on specific features of the data, such as location, property size, or market conditions. This variability provides a more tailored measure of uncertainty, as it reflects the heterogeneity inherent in real estate markets.

Consequently, the degree of width heterogeneity is a key factor in evaluating how well a PI methodology captures uncertainty across diverse real estate observations.

3.3.1.1 Cross Conformal Prediction

Conformal Prediction is a system for creating valid[1] PIs (Shafer and Vovk, 2008; Vovk, 2015; Vovk et al., 2005). CCP is a derivation of conformal prediction in which the entire training set is used for calibration. In CCP, the training undergoes a k-fold CV. This procedure provides out-of-sample residuals for all the observations in the training set, this being one of its main advantages. Equation (3.2b) defines the out-of-sample residuals R_i as the absolute difference between true price y_i and predicted price $\hat{\mu}(x_i)$, this measure is also known as the nonconformity measure in the conformal prediction literature. $\mathcal{I}_1$ represents the training fold of the k-fold CV, while $\mathcal{I}_2$ represents the test fold. The predicted price can come from any generic algorithm $\mathcal{A}$, hence this methodology is model-agnostic.

$$\hat{\mu}(x) \leftarrow \mathcal{A}\left(\{(x_i, y_i) : i \in \mathcal{I}_1\}\right) \tag{3.2a}$$

$$R_i = |y_i - \hat{\mu}(x_i)|, \quad i \in \mathcal{I}_2 \tag{3.2b}$$

The subsequent step is to estimate ϵ by ranking the out-of-sample residuals R_i in ascending order and selecting the value at the desired quantile $1 - \alpha$, as shown in Equation (3.3a). The quantile Q represents the value below which a given percentage of the data falls. In this case, it determines the residual magnitude that captures $1 - \alpha$ of the distribution. Then ϵ is subtracted/added to the individual prediction $\hat{\mu}(x_{n+1})$ forming the lower/upper bound of the PI. This procedure is defined in Equation (3.3b), where it is possible to note that the PI are homogeneous. In other words, all intervals have the same size, independent of the real estate characteristics x_{n+1}.

$$\epsilon = Q_{1-\alpha}(R) \tag{3.3a}$$

$$PI(x_{n+1}) = [\hat{\mu}(x_{n+1}) - \epsilon, \hat{\mu}(x_{n+1}) + \epsilon] \tag{3.3b}$$

Even though simple, CCP is capable of guaranteeing the desired coverage. Nevertheless, it will likely have a reduced efficiency, as some observations could have smaller intervals than estimated (Kato et al., 2023). Finally, the homogeneity of the PI might not be suitable for all applications, specially when you want to measure prediction uncertainty at the individual level.

[1] Valid in this context means that the PI contains the true value of new observations given a desired confidence level $1 - \alpha$.

3.3.1.2 Normalized Conformal Prediction

As in subsection 3.3.1.1, NCP also defines the out-of-sample residuals as the absolute difference between real and predicted price. The following step consists on training a regression algorithm $\mathscr{A}_r$ in which the residuals R are regressed on real estate characteristics x. The main takeaway of this step is conditioning uncertainty (residuals) on real estate characteristics. This allows the PI to vary across the different real estates being transacted. Therefore, this methodology produces heterogeneous PI.

$$\hat{\mu}_R(x) \leftarrow \mathscr{A}_r\left(\{(x_i, R_i) \,:\, i \in \mathscr{I}_2\}\right) \tag{3.4}$$

To ensure the desired coverage $1-\alpha$, we first train the regression algorithm $\mathscr{A}_r$, and then estimate ϵ. In this methodology, ϵ is defined as the quantile $Q_{1-\alpha}$ of the element-wise division between the out-of-sample residuals R_i and the fitted residuals $\hat{\mu}_r(x_i)$, as expressed in Equation (3.5):

$$\epsilon = Q_{1-\alpha}\left(R \oslash \hat{\mu}_R(x)\right). \tag{3.5}$$

Here, $\oslash$ represents the Hadamard (element-wise) division, meaning each residual R_i is divided by its corresponding fitted residual $\hat{\mu}_R(x_i)$. This normalization step ensures that the prediction intervals scale according to the difficulty of each prediction. Specifically, ϵ guarantees coverage (consistency), while the normalizing predicted residual $\hat{\mu}_R(x)$ ensures optimal interval size (efficiency). Kato et al. (2023) explain that using this PI methodology makes the length of the PI proportional to the predicted accuracy of the underlying regression algorithm for a new example. This improves the efficiency of conformal predictions by assigning larger PIs to more challenging examples and smaller ones to those that are easier to predict. In this context, the difficulty is determined by the accuracy of the underlying regression algorithm.

$$PI\left(x_{n+1}\right) = \left[\hat{\mu}\left(x_{n+1}\right) - \epsilon \times \hat{\mu}_r\left(x_{n+1}\right), \hat{\mu}\left(x_{n+1}\right) + \epsilon \times \hat{\mu}_r\left(x_{n+1}\right)\right], \tag{3.6}$$

Equation (3.6) demonstrates how the PI of the NCP methodology is calculated. The product of ϵ and the predicted residual $\hat{\mu}_r\left(x_{n+1}\right)$ is subtracted/added to the property predicted price in order to obtain the lower/upper bound of the PI. This procedure generates a symmetric PI, as the lower and upper bounds are the same size.

3.3.1.3 Direct Prediction

The DP methodology employed in this work is derived from Francke et al. (2024). This methodology uses the same nonconformity measure as the previous PI methodologies, specified in Equation (3.2b). After estimating the out-of-sample residuals, and given the confidence level $1 - \alpha$, a quantile regression algorithm $\hat{q}_{1-\alpha}$ is trained, where the residuals R are regressed on real estate characteristics x. Different from the previous methods, the predicted price $\hat{\mu}(x_{n+1})$ is added to the real estate characteristics matrix (x_{n+1}) before training the quantile regression algorithm. Hence, x^*_{n+1} stands for the real estate characteristics matrix that had the point prediction $\hat{\mu}(x_{n+1})$ added to it.

In Francke et al. (2024) the authors use a Light Gradient-Boosting Machine with a pinball loss function as the quantile regression algorithm for predicting the residuals. The current work, different from the aforementioned authors, uses a traditional linear quantile regression model following Koenker and Bassett Jr (1978) and Koenker and d'Orey (1987). This choice facilitates interpretation because the estimated coefficients can be accompanied by standard asymptotic inference under familiar regularity conditions. They provide a parsimonious description of how observed covariates are associated with shifts in the relevant conditional quantiles of the residual distribution. At the same time, the interpretation is inherently conditional: quantile regression coefficients capture partial associations at a given conditional quantile, holding other regressors fixed, and they should not be read as marginal effects or structural parameters. With these qualifications in mind, the linear specification remains informative as an empirical diagnostic of which characteristics are most strongly associated with higher predicted residual quantiles, and therefore with wider prediction intervals, compared with more flexible black-box quantile learners for which attributing uncertainty to specific covariates is less straightforward.

The trade off between using more traditional algorithms versus the ML can be summarized as efficiency vs interpretability. The ML algorithms with quantile loss functions might generate more efficient (narrower) intervals when compared to the traditional quantile regression methods, assuming the same confidence level $1 - \alpha$. This higher efficiency comes with the cost of loosing the interpretablity of the results.

$$PI(x^*_{n+1}) = \left[\hat{\mu}(x_{n+1}) - \hat{q}_{1-\alpha}\left(x^*_{n+1}\right), \hat{\mu}(x_{n+1}) + \hat{q}_{1-\alpha}\left(x^*_{n+1}\right)\right]. \tag{3.7}$$

Equation (3.7) demonstrates how the PI is built for this methodology. The predicted quantile residual $\hat{q}_{1-\alpha}\left(x^*_{n+1}\right)$ is subtracted/added to the predicted price $\hat{\mu}(x_{n+1})$ forming the lower/upper bound of the PI.

3.3.1.4 Direct Lower and Upper Prediction

The DLUP methodology is also from Francke et al. (2024) and is the only methodology in this work with a different nonconformity measure. In this case, the out-of-sample residuals are defined as the real price y_i minus the predicted price $\hat{\mu}(x_i)$, as in Equation (3.8). This allows the distribution of the residuals to have negative values. The idea behind this choice is to train two regression quantile algorithms, each predicting the lower/upper bound. The lower/upper bound algorithm will have a confidence level of $\alpha/2$ and $1-\alpha/2$ respectively. The lower/upper bound algorithm will predict the negative/positive side of the residuals distribution.

$$R_i = y_i - \hat{\mu}(x_i), \quad i \in \mathcal{I}_2, \tag{3.8}$$

All PI methodologies that were previously shown have symmetric PI, centered at the price prediction $\hat{\mu}(x_i)$, by design. As a result of having two algorithms, one for the lower and the other for the upper bound, the PI will be asymmetric. In other words, this allows the PI to better quantify a real estate price that might have more lower or upper bound uncertainty. Larger lower/upper bound means that there is a higher uncertainty in over/under pricing the real estate in question. This information can be crucial to investors and developers that want to distinguish a lemon from a peach real estate.

The quantile regression algorithms for estimating the lower ($q_{\alpha/2}$) and upper ($q_{1-\alpha/2}$) bound of the residuals R_i are trained using the real estate characteristic matrix with the appended price prediction x_i^*. The prediction of the lower and upper bound of the residuals R_i are added to the price prediction $\hat{\mu}(x_i)$ as exemplified in Equation (3.9). Because the lower bound model produces negative values, the lower bound value is added to the price prediction, different from the previous methods, where the lower bound value is subtracted from the price prediction.

$$PI(x_{n+1}^*) = \left[\hat{\mu}(x_{n+1}) + \hat{q}_{\alpha/2}\left(x_{n+1}^*\right), \hat{\mu}(x_{n+1}) + \hat{q}_{1-\alpha/2}\left(x_{n+1}^*\right)\right]. \tag{3.9}$$

3.3.2 Performance Evaluation

As expressed by Vovk (2012), validity and efficiency are the two fundamental desiderata for prediction intervals. The best-performing algorithm or methodology would exhibit both high efficiency and validity, resulting in the smallest PI achievable given the desired confidence level. A confidence predictor is considered valid if the miscoverage rate (i.e., the fraction of true values falling outside the prediction interval) does not exceed α at the chosen confidence level $1-\alpha$. A predictor is deemed efficient if the prediction intervals it

generates are as small as possible considering the limitations of the underlying regression algorithm[2].

Efficiency is measured as the mean difference between the upper bound u_i and the lower bound l_i of the PI. In this study, the prices are transformed into log-prices. As a consequence of using the log-prices, the interval width will be equal to a percentage of the real estate price, hence, the interval width is proportional no the property price. The Mean Prediction Interval Width (MPIW) efficiency metric is defined in Equation (3.10).

In the alternative case where price is used as the dependent variable, it is important to use the Mean Relative Prediction Interval Width (MRPIW) (Equation (3.11)), which is a similar measure. The main difference between MPIW and MRPIW is that MRPIW divides the interval width $u_i - l_i$ by the prediction $\hat{y}_i$. This adjustment makes the interval width relative to the property price, ensuring that cheaper properties do not consistently have narrower intervals compared to expensive properties. In this study, MPIW and MRPIW yield nearly identical values, differing only at the millesimal level. When the model uses log-price as the dependent variable, MPIW is used, while MRPIW is used when price is the dependent variable or after transformation of log price predictions back to currency price.

$$\text{MPIW} = \frac{1}{N}\sum_{i=1}^{N}(u_i - l_i). \tag{3.10}$$

$$\text{MRPIW} = \frac{1}{N}\sum_{i=1}^{N}\frac{(u_i - l_i)}{\hat{y}_i}. \tag{3.11}$$

This work also looks at Standard Deviation of the Prediction Interval Width (SDPIW) as a way to measure how heterogeneous the PIs are. For example, the CCP methodology will have the SDPIW equal to zero since all the PIs have the same size. Therefore, the closer SDPIW is to zero, the more homogeneous the model is. The heterogeneity metric SDPIW is defined in Equation (3.12).

$$\text{SDPIW} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(u_i - l_i)^2 - \left(\frac{1}{N}\sum_{i=1}^{N}(u_i - l_i)\right)^2} \tag{3.12}$$

The metric for coverage is equal to the number of out-of-sample prices that fall within the predicted interval. When the true price of the real estate is within the PI it counts as 1, if not, zero. The mean of this count is equal to the coverage, as described in Equation (3.13). The coverage rate number can range from 0 to 1, but it is required to be equal to the desired confidence level $1 - \alpha$. Hence, it is best to the algorithm to have the coverage rate closer or

[2] A smaller prediction interval is deemed more efficient than a larger one.

equal to the desired confidence level.

As mentioned previously, this work also examines how many of the true prices fall under and over the PI. The under/over PI measure can help us better understand if the algorithm is more prone to over/under estimating the true price of the real estate. If the proportion of true prices are under/over the PI, this suggests that the regression algorithm is over/under pricing the real estate.

$$\text{Coverage} = \frac{1}{N}\sum_{i-1}^{N} c_i \quad \text{where} \quad \begin{cases} c_i = 1 & \text{if } y_i \in [l_i, u_i] \\ c_i = 0 & \text{if } y_i \notin [l_i, u_i] \end{cases}. \tag{3.13}$$

This work uses an additional measure called Mean Winkler Interval Score (MWIS) (Winkler, 1972). It evaluates the PIs by combining the efficiency and coverage metric. When the true price is within the PI, only the size of the interval is counted as the Winkler Score. If the true price lies outside of the PI, there is a penalty term added to the PI size that is being counted as the Winkler Score. This penalty term is equal to $2/\alpha$ times the distance between true price and upper or lower bound. This procedure is demonstrated in Equation (3.14). The mean value of these individual Winkler Scores is the metric used to evaluate the algorithms.

$$\text{MWIS} = \frac{1}{N}\sum_{i-1}^{N} W_i \quad \text{where} \quad \begin{cases} W_i = (u_i - l_i) + \frac{2}{\alpha}(l_i - y_i) & \text{if } y_i < l_i, \\ W_i = (u_i - l_i) & \text{if } l_i \leq y_i \leq u_i, \\ W_i = (u_i - l_i) + \frac{2}{\alpha}(y_i - u_i) & \text{if } y_i > u_i, \end{cases} \tag{3.14}$$

The performance evaluation of the PIs will take into account these four metrics under two different frameworks. The first framework is a simple split, where 70% of the data is randomly sampled to be used as a training set. The other 30% of the data is used as a test set to check out-of-sample performance of the PI methodologies and their respective regression algorithms. In this first framework, year dummy variables are used to control the time variations, and the best hyperparameters of the ML algorithms are estimated using a 10 fold CV in the training set.

The second framework is a rolling window approach that is used to check the robustness of the methodologies over time. In this approach, the data is divided in transactions per year, the regression algorithms are trained yearly, the following year is used as a test set for out-of-sample evaluation of performance. As the training sample is only one year, there are no control variables for time. Prices in the training set are updated to current price levels by using the returns from a price index that is calculated using the double imputation

methodology from Calainho et al. (2022).

The performance metrics displayed for the underlying regression algorithms used in the out-of-sample log-price prediction are RMSE, Mean Absolute Percentage Error (MAPE), the residual mean, and Rsquared. These metrics have been widely used in the real estate literature as well as in econometrics and machine learning. The accuracy of these algorithms play a particular role in this study as the PI are generated based on the distribution of the residuals. Hence, models with lower errors tend to generate more efficient/narrower PIs.

3.3.2.1 Rolling Window

A rolling window approach is employed to assess the robustness of the methodologies over time, spanning the period from 2001 to 2020. In this approach, the data is divided into yearly transaction sets, with machine learning algorithms trained on the data for each year t, and the subsequent year $t + 1$ is used as the test set for out-of-sample evaluation of efficiency and consistency. Additionally, the machine learning algorithms are hypertuned annually to ensure optimal performance in each year's specific market conditions.

Since the training sample for each year consists of only one year of data, no time control variables are included. To account for time-related differences, prices in the training set are adjusted to current levels using returns derived from a price index calculated via the double imputation methodology, as described by Calainho et al. (2022). The double imputation approach is considered a more reliable method for constructing price indices, as it helps to mitigate biases resulting from omitted variables in the model, with these biases at least partially offsetting each other (Balk et al., 2013). Only the training set is subject to this price index correction, the test set remained unaltered.

The same regression algorithm used to predict real estate prices in the AVM is also employed to generate the price index. In other words, the AVM algorithm corrects prices over time using an index that is calculated with the same corresponding algorithm. For example, if the AVM utilizes XGBT, the price index used for time-correction is also generated using XGBT.

3.4 Data

The data used in this study was provided by the Dutch Brokerage Association (NVM). The data encompass 75% of all housing transactions within the city of Amsterdam and includes many property characteristics (De Wit et al., 2013; van Dijk and Francke, 2018). Observations in the bottom and top 1% of the log(price/area) distribution have been removed as,

according to Steurer et al. (2021), data entry errors and outliers tend to be concentrated at the extremes of the distribution. This leaves this study with a total of 86,838 transactions between 2001 and 2021. The training set has 60,787 observations and the test set 26,050.

The log prices are predicted based on the following property characteristics, size in square meters and dummy variables for the presence of a garden, construction period (1500-1905, 1906-1930, 1945-1959, 1960-1970, 1971-1980, 1981-1990, 1991-2000, after 2001), year of transaction, zipcode, and property type (terraced, back-to-back, corner, semi-detached, detached, ground floor, split level apartment, upper floor split level, other apartment).[3]

Year	Avg. Price	SD. Price	Avg. Area	SD. Area	Obs
2001	254,702	157,227	99	47	2,897
2002	256,079	163,097	98	47	3,397
2003	249,484	164,023	97	46	3,649
2004	246,572	158,758	94	48	4,407
2005	259,499	184,965	94	50	5,113
2006	274,153	187,336	92	47	5,311
2007	288,261	203,024	91	49	5,358
2008	291,341	177,914	88	41	4,919
2009	283,597	173,752	88	42	4,141
2010	293,067	208,911	89	49	4,213
2011	288,177	208,042	88	50	3,777
2012	264,687	191,635	87	46	3,698
2013	262,802	202,549	86	47	3,746
2014	271,180	178,690	85	42	5,677
2015	294,570	197,525	87	44	6,445
2016	335,721	208,618	91	43	5,654
2017	379,945	198,235	97	41	4,026
2018	428,918	211,144	103	43	2,780
2019	433,987	190,216	103	39	3,213
2020	460,087	196,345	106	39	2,983
2021	491,997	237,751	110	45	1,434

Table 3.1: Amsterdam sample summary statistics. *Avg. Price* = Average Price in Euros; *SD. Price* = Price Standard Deviation in Euros;*Avg. Area* = Average Area in Square Meters; *SD. Area* = Area Standard Deviation in Square Meters; *Obs.* = Number of Observations.

The yearly summary by transaction date is available in Table 3.1. From this table we can notice that the number of transactions varies across the years. The average number of transactions over the 21 years in the sample is equal to 2897 per year. The average area exhibits a pattern, it decreases until 2014. In 2014 the average floor space starts to increase, and in 2018 it surpasses the average floor space level from 2001.

[3] Some properties are freehold whereas others are leasehold, this is corrected for by including a dummy-variable in the case the property is leasehold (van Dijk, 2024)

3.5 Results

This section first presents the accuracy of the real estate log-price prediction algorithms, followed by the performance of the PI methodologies on the 70/30 split. An analysis of the interval width distribution and a map showing the variability of this metric throughout the neighborhoods of Amsterdam is also available. Finally, the performance of the PI methodologies on the rolling window approach is presented. This showcases, on average and for the duration of the sample, which neighborhood has a higher or lower price prediction uncertainty. All results presented in the corpus of this paper use $\alpha = 0.05$, while the results in Appendix labeled "Additional Prediction Interval Performance" correspond to $\alpha = 0.01$.

3.5.1 Log-Price Prediction Algorithms

The accuracy performance of the real estate log-price prediction is shown in Table 3.2. The regression algorithm with the lowest RMSE and MAPE is Cubist followed by XGBT. OLS and Cubist have a slightly negative residual mean, while XGBT has a positive residual mean. This results are calculated based on the results obtained in the test set, hence, out-of-sample. With these results, it can be expected that the algorithm with the lowest RMSE will also have the most efficient PI, as the methodologies PI are based on the residuals of these regression algorithms. Hence, Cubist is expected to have the narrowest MPIW.

	RMSE	MAPE	Residual Mean	R squared
OLS	0.1960	0.1866	-0.0020	0.8572
XGBT	0.1462	0.1113	-0.0007	0.9205
Cubist	**0.1344**	**0.1008**	-0.0016	**0.9328**

Table 3.2: Regression Out-of-Sample Residual Metrics. *OLS* = Ordinary Least Squares; *XGBT* = Extreme Gradient Boosting Tree; *RMSE* = Root Mean Squared Error; *MAPE* = Mean Absolute Percentage Error.

3.5.2 Prediction Interval Performance

The performance of the PI methodologies and its underlying regression algorithms can be see in Table 3.3. The PI methodology and regression algorithm with the lowest MPIW (0.5068) is DP with Cubist. The coverage of this combination is equal to 0.9487, slightly less than the desired confidence level of 0.95. The methodology and algorithm with the highest coverage is CCP with XGBT equal to 0.9507. Having the coverage value higher then the confidence level is not suitable as it means that the intervals width could have been, on average, smaller to reach the desired confidence level. That is, the additional coverage can be seen as a loss in efficiency. When taking into account the MWIS measure, which combines

efficiency and coverage metrics, the best combination o PI methodology and regression algorithm is DLUP with Cubist, being the lowest value with 0.6573.

As mentioned previously, the SDPIW measure tells how homogeneous the interval widths are. As explained in subsection 3.3.1.1, the CCP methodology produces homogeneous[4] intervals. This is the reason why the SDPIW values for all the regression algorithms in this methodology are equal to zero. When observing the SDPIW value range for all three methodologies that have values different from zero, the least efficient underlying regression algorithm has the largest SDPIW. Regression algorithms with higher efficiency have lower SDPIW. The lowest MWIS score, which is the DLUP with Cubist combination, has also the lowest SDPIW. Hence, we can conclude from these results that there is no explicit link between efficiency/coverage and SDPIW.

Regarding the Under and Over PI measures, NCP with OLS was one of the combinations that showed the highest difference between the two measures, 0.0339 Under and 0.0185 Over. This means that this combination is more likely overpricing the real estate predictions. The most stable combination is DLUP with XGBT, this means that the Under and Over value are the closest to being equal, in this case it was 0.0248 and 0.0254. When comparing the Under and Over PI measures of the two most similar PI methodologies, namely DP and DLUP. DLUP exhibits a lower difference between the two measures when compared to the DP methodology. This suggests that the asymmetric intervals that DLUP generates mitigates Under or Over pricing.

All regressions algorithms using PI methodologies with symmetric intervals have some difference between the Under and Over PI measure. The lowest difference for all the regression algorithms was DLUP methodology, which is the only methodology that generates asymmetric PI. This suggests that the asymmetric interval counter-acts the difference between Under and Over PI measure. Another point that further corroborates to this idea is the MWIS measure. This measure has a penalty term that is equal to the distance from the upper or lower bound to the true price. Hence, if you allow the PI to be asymmetric, the distance between the PI bounds and the true price will be shortened. This is assuming that the true price is closer to the largest bound of this asymmetric PI. Thus, the penalty term will be lower for PI methodologies that have their PI bounds closer to the true price, resulting in a lower MWIS, which DLUP has the lowest among all PI methodologies and for all regression algorithms.

[4] Homogeneous in this case means that the intervals have identical width.

Method	Algo.	MPIW	SDPIW	Coverage	Under PI	Over PI	MWIS
CCP	OLS	0.7019	0.0000	0.9464	0.0236	0.0300	1.0049
	XGBT	0.5741	0.0000	0.9507	0.0262	0.0231	0.7661
	Cubist	**0.5286**	0.0000	**0.9506**	0.0279	0.0215	**0.7186**
NCP	OLS	0.6975	0.2773	0.9476	0.0339	0.0185	0.8846
	XGBT	0.5494	0.1310	**0.9499**	0.0294	0.0207	0.7105
	Cubist	**0.5069**	0.1357	0.9481	0.0291	0.0228	**0.6639**
DP	OLS	0.6662	0.2035	0.9463	0.0315	0.0222	0.8654
	XGBT	0.5512	0.1339	**0.9498**	0.0291	0.0212	0.7099
	Cubist	**0.5068**	0.1312	0.9487	0.0285	0.0228	**0.6644**
DLUP	OLS	0.6600	0.2480	0.9463	0.0274	0.0263	0.8486
	XGBT	0.5451	0.1256	**0.9498**	0.0248	0.0254	0.6953
	Cubist	**0.5074**	0.1300	0.9491	0.0260	0.0250	**0.6573**

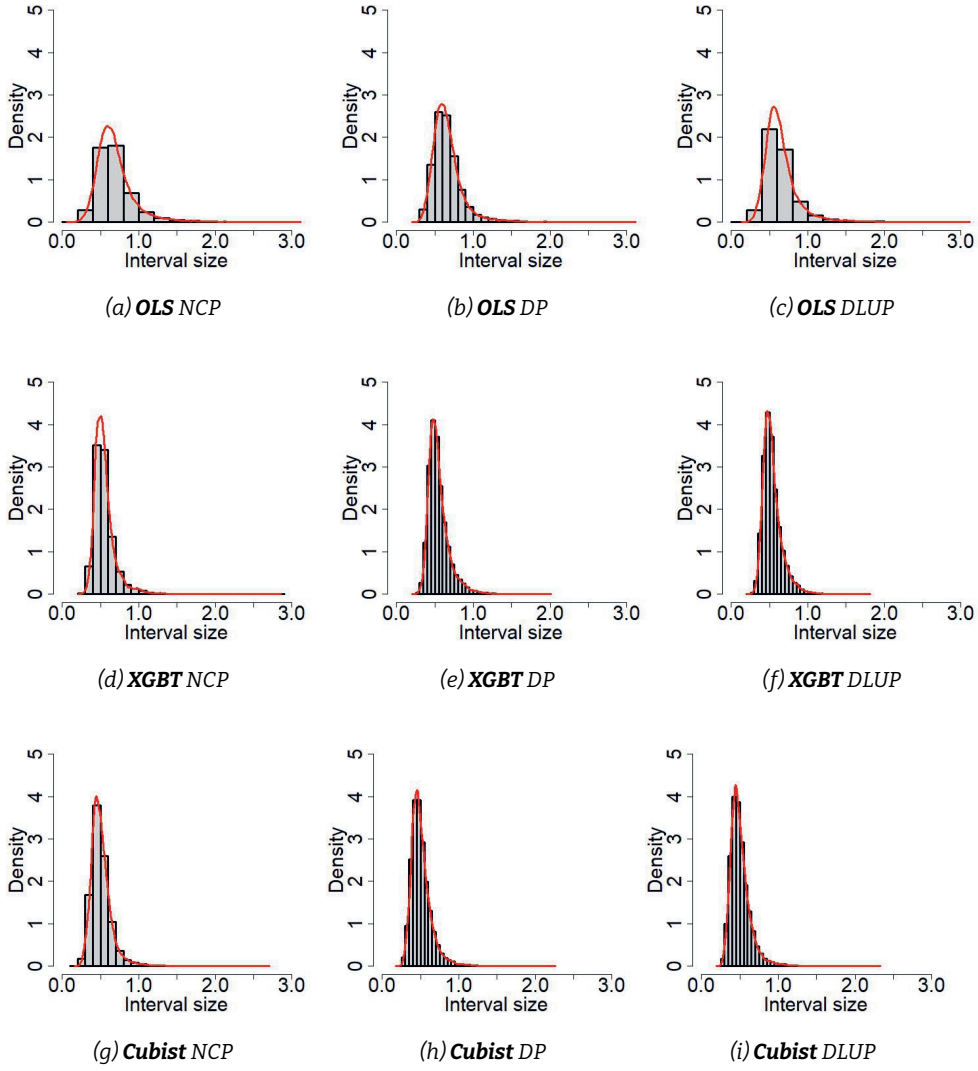
Table 3.3: Results for the Four Different Prediction Interval Methods ($\alpha = 0.05$). *CCP* = Cross Conformal Prediction; *NCP* = Normalized Conformal Prediction; *DP* = Direct Prediction; *DLUP* = Direct Lower and Upper Prediction; *OLS* = Ordinary Least Squares; *XGBT* = Extreme Gradient Boosting Tree; *Algo.* = Algorithm; *MPIW* = Mean of the Prediction Interval Width; *SDPIW* = Standard Deviation of the Prediction Interval Width; *PI* = Price Interval; *MWIS* = Mean Winkler Interval Score.

3.5.2.1 Prediction Interval Width Distribution Analysis

Figure 3.1 shows the histogram and probability density estimation of the prediction interval width for the all the algorithms with the three PI methodologies that do not have a fixed interval width. All distributions in Figure 3.1 are positively skewed. The interval width distribution results are within expectations when taking into account the SDPIW values displayed in Table 3.3. OLS results obtained a higher SDPIW, this translates into flatter distributions with longer tails when compared to the ML algorithms.

The DLUP methodology with the XGBT algorithm, produced the shortest tail with a value bellow 2. It is worth mentioning that the interval width distribution for all of the OLS regression results had a long right-hand tail that ended in the interval width value of 6. This is also the reason why the bins for all the OLS results are wider than the other regression algorithms. The distribution figure was cropped to an interval width of 3, with the number of breaks set to 30 for visualization purposes.

As a way to better understand the outcomes of the asymmetric intervals that the DLUP methodology generates, is important to measure the different levels and frequencies of asymmetry. This study considers the lower/upper bound width as the distance between the lower/upper bound l_i/u_i and the predicted price $\hat{y}_i$. Hence, Figure 3.2 shows the histogram of the difference between the upper and lower bound width. Another important point noted is that for all of the observations where the true price falls within the interval predicted, the true price is always within the largest bound, be it the lower or upper bound.



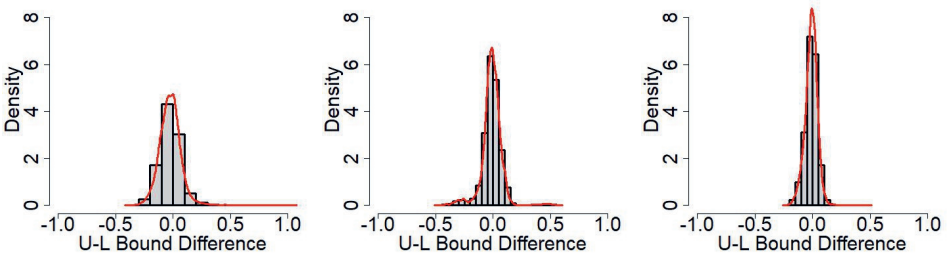
(a) ***OLS*** *NCP* *(b)* ***OLS*** *DP* *(c)* ***OLS*** *DLUP*

(d) ***XGBT*** *NCP* *(e)* ***XGBT*** *DP* *(f)* ***XGBT*** *DLUP*

(g) ***Cubist*** *NCP* *(h)* ***Cubist*** *DP* *(i)* ***Cubist*** *DLUP*

Figure 3.1: Probability Density Estimation of Prediction Interval Width for Three Different Prediction Interval Methods ($\alpha = 0.05$). *OLS* = Ordinary Least Squares; *XGBT* = Extreme Gradient Boosting Tree; *NCP* = Normalized Conformal Prediction; *DP* = Direct Prediction; *DLUP* = Direct Lower and Upper Prediction. Number of Breaks = 30

Figure 3.2 shows that the results differ significantly by regression algorithm. OLS has a very long right-hand tail, which means that the upper bound width is larger. This suggests that the OLS regression algorithm occasionally significantly under prices real estate. The XGBT regression algorithms has the result that appears to be the most symmetric, this suggest no consistent under or over pricing. Whereas the Cubist algorithm has a slightly longer right-hand tail, meaning it occasionally under prices real estate.



(a) ***OLS*** *NCP* *(b)* ***XGBT*** *DP* *(c)* ***Cubist*** *DLUP*

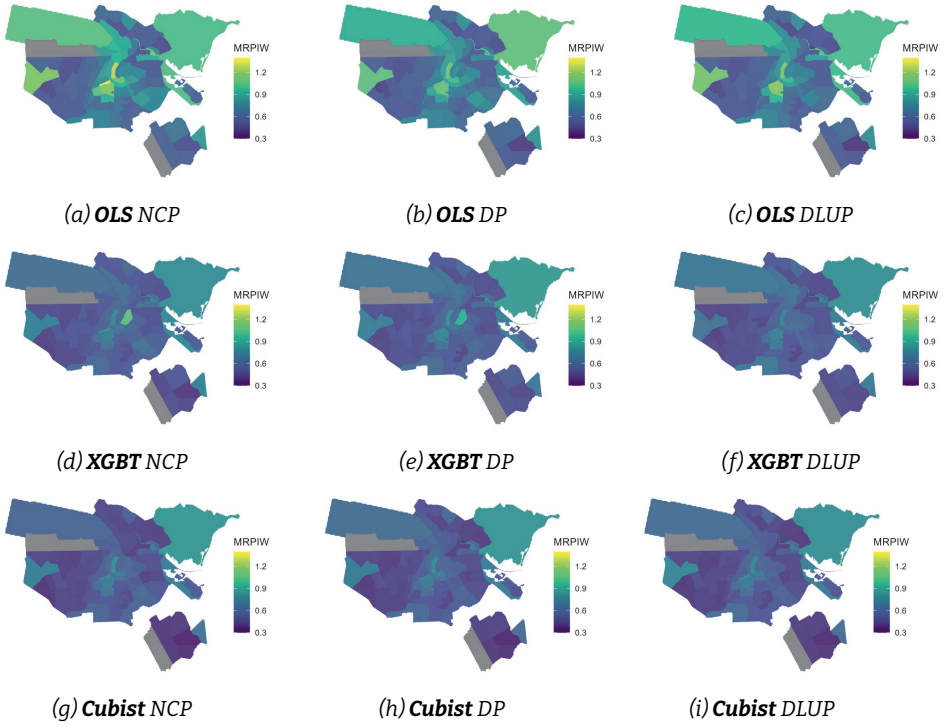
Figure 3.2: Probability Density Estimation of the Upper Minus Lower Bound Difference ($U - L$) for the DLUP Prediction Interval Method ($\alpha = 0.05$). *OLS* = Ordinary Least Squares; *XGBT* = Extreme Gradient Boosting Tree; *DLUP* = Direct Lower and Upper Prediction. Number of Breaks = 20

3.5.2.2 Amsterdam Mean Prediction Intervals Width Map

Figure 3.3 shows the MPIW per neighborhood of Amsterdam. These are calculated on the intervals predicted for the test set of the 70/30 split. It contains transactions from all years in the sample. Hence, Figure 3.3 represents the average MPIW over the years for each neighborhood. The results available are for the three different methodologies with non-zero SDPIW, and their underlying regression algorithm.

Observe in Figure 3.3 that the three methodologies are able to distinguish the price prediction uncertainty per neighborhood. This location dependence feature is important to have as real estate price is location dependent (Bourassa et al., 2007). Is important to note that, there is a higher variation in MPIW across the different regression algorithms when compared to the different methodologies. Hence, the regression algorithm has a higher impact on efficiency when compared to the PI methodology.

Another important point to note is that, there are several neighborhoods that have the same analogous higher PI size, across different methodologies and regression algorithms. Therefore, for most of the neighborhoods, all the PI methodologies and regression algorithms agree on which neighborhoods have a higher price prediction uncertainty relative



(a) ***OLS*** *NCP* (b) ***OLS*** *DP* (c) ***OLS*** *DLUP*

(d) ***XGBT*** *NCP* (e) ***XGBT*** *DP* (f) ***XGBT*** *DLUP*

(g) ***Cubist*** *NCP* (h) ***Cubist*** *DP* (i) ***Cubist*** *DLUP*

Figure 3.3: Amsterdam Map of Mean Prediction Interval Width (MPIW) per Neighborhood for Three Different Prediction Interval Methods ($\alpha = 0.05$). *OLS* = Ordinary Least Squares; *XGBT* = Extreme Gradient Boosting Tree; *NCP* = Normalized Conformal Prediction; *DP* = Direct Prediction; *DLUP* = Direct Lower and Upper Prediction.

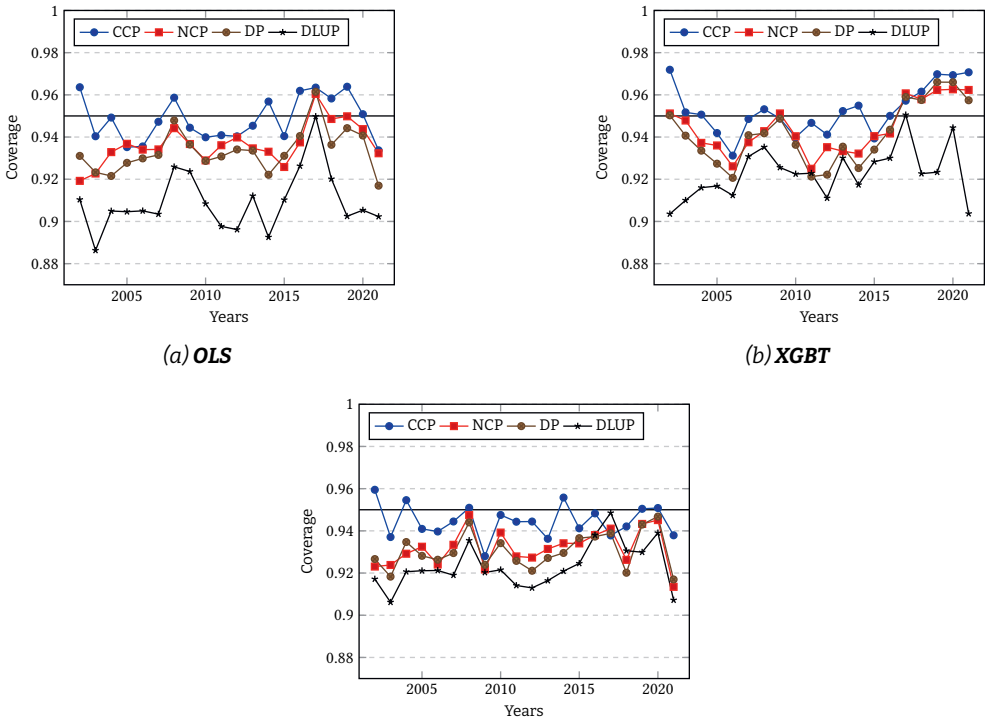
to the other neighborhoods. This is an indication of a common factor that is being capture by all the methodologies and their respective regression algorithms. This argument will be further developed in Section 3.5.3.2.

3.5.3 Rolling Window Results

3.5.3.1 Coverage

Figure 3.4 displays the coverage of the four methodologies for each of the three underlying regression algorithms over the 2001-2020 period. Figure 3.4 shows that CCP is the only methodology that manages, on average, to stay close to the desired confidence level of 0.95. NCP is the second methodology that manages to stay closer to the desired confidence level. The other methodologies stay, on average, bellow the desired confidence level. DLUP is the worst methodology when taking this metric into account, in its worst case it is more than 5% bellow the desired confidence level.

There are two main reasons that might explain this decrease in the coverage metric



(a) ***OLS*** *(b)* ***XGBT*** *(c)* ***Cubist***

Figure 3.4: Rolling Window Coverage Results for the Four Different Prediction Interval Methods ($\alpha = 0.05$). *OLS* = Ordinary Least Squares; *XGBT* = Extreme Gradient Boosting Tree; *CCP* = Cross Conformal Prediction; *NCP* = Normalized Conformal Prediction; *DP* = Direct Prediction; *DLUP* = Direct Lower and Upper Prediction.

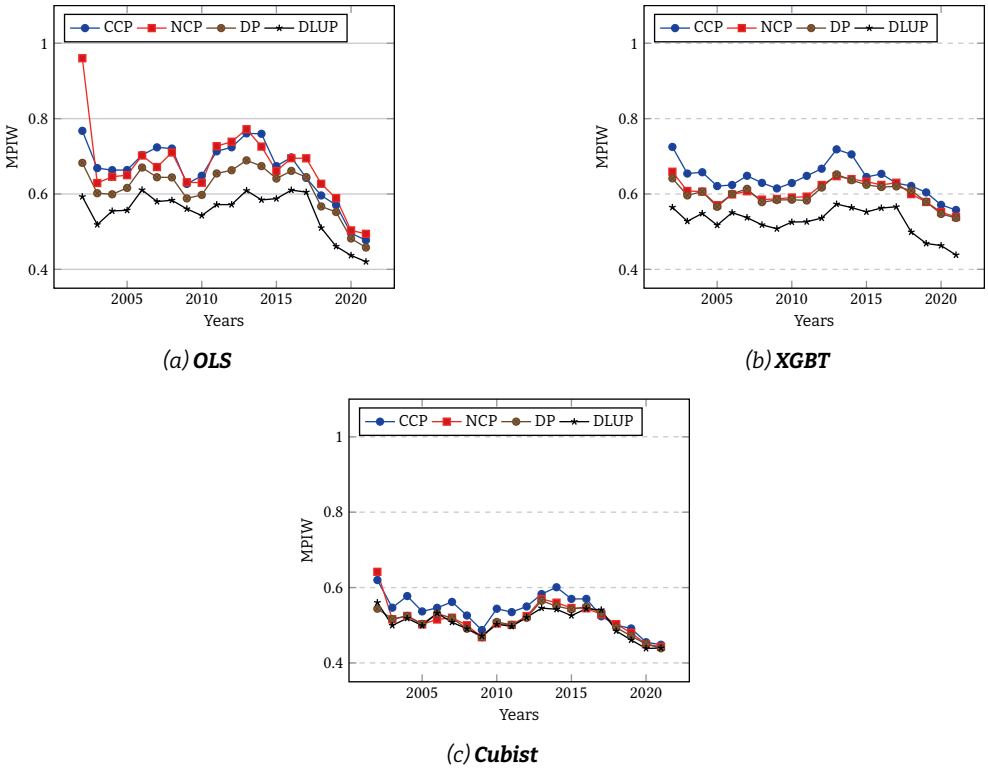
when compared to the 70/30 split, it can be one of the two or even a combination of the two. The first explanation for this phenomena is the fact that the training set is comprised of only one year. This drastically limits the variability of property characteristics in the sample, meaning that the sample is too small to significantly represent all the different real estate that exists in that market, there are on average 4197 observations per year. In other words, the observations exchangeability requirement is being pushed to its limits. The second reason behind the coverage decrease is that the prices are being updated to present value by multiplying the sales price by the return of the double-imputation price index methodology from Calainho et al. (2022). This procedure might be adding some noise to the results, as the index return is the average price movement of the market. Hence, not all real estate price segments follow the market average price movement. This distortion plus some index pricing estimation error might be causing the models to under-perform in the coverage metric. However, all the coverage results in the rolling window section seems to be flat with no trends and to vary within a fixed level that is slightly bellow the

desired confidence level.

3.5.3.2 Mean Prediction Interval Width

Figure 3.5 shows how the MPIW, of the four methodologies for each of the three underlying regression algorithms, varies across the years. This metric over the years uncovers several interesting points. The Cubist regression algorithm is not sensitive to the PI methodology, meaning that the MPIW for all methodologies using this algorithm were very similar, with almost no variation from one another. It is worth noting that, DLUP generates the narrowest MPIW for most of the years, for all regression algorithms. But it is also the method with the lowest coverage. Hence, even though the methodology is efficient by providing narrow intervals, it might not guarantee the desired coverage level. NCP and DP are the methodologies with the best efficiency-coverage trade-off, they manage to stay close to the desired confidence level while providing narrow intervals.

Since the MPIW is directly derived from the residuals of the underlying regression algorithm, its size depends on the sources of uncertainty that influence these residuals. According to Francke and Van de Minne (2021b), these sources may include error in variables, which refers to measurement inaccuracies in property characteristics, specification error, which arises from incorrect functional form assumptions that misrepresent relationships between variables, omitted variable bias, which results from unobserved heterogeneity due to missing property or locational characteristics, and transaction noise, which stems from market frictions such as bargaining power and asymmetric information. Each of these components contributes to the residual variability measured by the MPIW. Figure 3.5 shows that all MPIWs have a similar pattern for every year in the sample. From 2015 to 2022 all MPIWs goes down to its lowest level. A possible explanation to this downward trend is an increase in data quality over the last years in the sample, This would decrease the measurement inaccuracies in property characteristics and consequently the error in variable component.



(a) ***OLS*** *(b)* ***XGBT*** *(c)* ***Cubist***

Figure 3.5: Mean Prediction Interval Width (MPIW) of the Rolling Window Results for the Four Different Prediction Interval Methods $\alpha = 0.05$. *OLS* = Ordinary Least Squares; *XGBT* = Extreme Gradient Boosting Tree; *CCP* = Cross Conformal Prediction; *NCP* = Normalized Conformal Prediction; *DP* = Direct Prediction; *DLUP* = Direct Lower and Upper Prediction.

3.6 Conclusion

This paper studies how commonly used, model-agnostic prediction-interval constructions behave when paired with different regression engines in residential AVMs. Using the same data, evaluation metrics, and empirical designs throughout, the analysis compares four interval methodologies—CCP, NCP, DP, and DLUP—implemented on three regression models: a conventional hedonic OLS baseline and two high-performing machine-learning predictors, XGBT and Cubist. The central empirical message is that the regression algorithm largely governs interval *efficiency*, whereas the interval methodology has a stronger influence on achieving nominal *validity* (coverage). This division of roles is useful for applied work because it clarifies which design choice is primarily responsible for tightness versus calibration.

Interpreting these results through the lens of property-level uncertainty helps connect

the statistical exercise to economic use cases. Prediction intervals translate the (conditional) residual distribution of an AVM into an operational, dwelling-specific range of plausible prices. In that sense, interval width provides a decision-relevant analogue to residual dispersion measures used to quantify idiosyncratic housing risk, while interval asymmetry is informative when under- and overvaluation have different consequences (for example in underwriting margins or downside-risk screening). The empirical findings indicate that heterogeneous interval methods can capture meaningful variation in uncertainty across properties and neighborhoods, but they also reinforce the trade-off that narrower intervals are only valuable to the extent that coverage is maintained at the chosen confidence level.

Among the evaluated combinations, Cubist yields the strongest point-prediction performance and, correspondingly, tends to produce the narrowest intervals across methods. In the split-sample evaluation, the most favorable width–coverage trade-off is obtained by Cubist combined with DLUP when performance is summarized by the Winkler score, reflecting the usefulness of allowing asymmetric bounds when residual risk is not symmetric. At the same time, methods can differ materially in their calibration: designs that deliver efficiency gains may do so at the cost of undercoverage, particularly in more demanding forecasting setups.

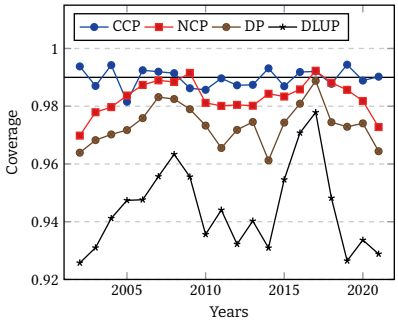
The rolling-window analysis, which trains on a single year and tests on the subsequent year, provides a complementary view of operational performance. Coverage is broadly stable over time but is, on average, below the nominal level for several method–model combinations. Two features of the empirical design plausibly contribute to this pattern. First, annual training samples can be too small to span the full heterogeneity of the next year's transactions, so the exchangeability conditions that underlie conformal calibration are strained. Second, in the rolling-window framework the training sample is updated using a mean-based price index, which can introduce additional noise when different market segments deviate from the market-wide average. These considerations do not overturn the main comparisons, but they bound the interpretation and underscore that interval calibration is most reliable when the calibration data are representative of the forecast environment.

Overall, the paper contributes an incremental, model-class-spanning comparison that clarifies what practitioners and researchers can expect from widely used prediction-interval methodologies when implemented with different regression engines. The results support a practical takeaway: interval construction and regression choice should be treated as jointly important design decisions, with the former primarily affecting calibration and the latter primarily affecting tightness and heterogeneity. Future work may

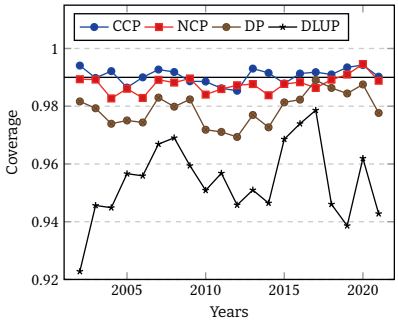
extend these ideas toward richer probabilistic outputs, such as conformal predictive distributions (Vovk et al., 2017), which can support threshold-based decision rules by allowing users to query tail probabilities directly. More broadly, improving the interpretability and decision relevance of uncertainty outputs remains central to deploying AVMs in valuation, underwriting, and investment settings where uncertainty is consumed *ex-ante* and where asset-level errors are economically consequential.

Chapter appendix

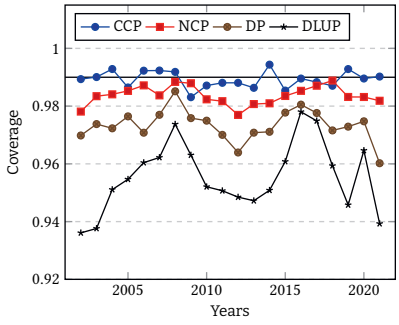
3.A1 Additional Prediction Interval Performance



(a) ***OLS***



(b) ***XGBT***

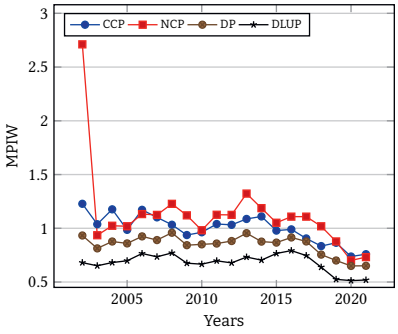


(c) ***Cubist***

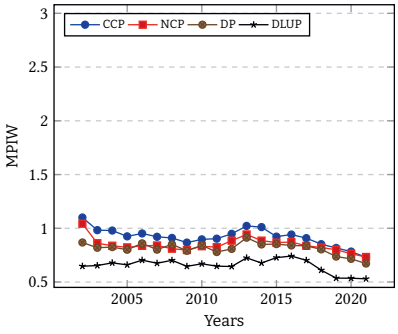
Figure 3.A1.1: Rolling Window Coverage Results for the Four Different Prediction Interval Methods ($\alpha = 0.01$). *OLS* = Ordinary Least Squares; *XGBT* = Extreme Gradient Boosting Tree; *CCP* = Cross Conformal Prediction; *NCP* = Normalized Conformal Prediction; *DP* = Direct Prediction; *DLUP* = Direct Lower and Upper Prediction.

Method	Algo.	MPIW	SDPIW	Coverage	Under PI	Over PI	MWIS
CCP	OLS	1.0424	0.0000	0.9895	0.0072	0.0033	1.7047
	XGBT	0.8526	0.0000	0.9896	0.0068	0.0036	1.1463
	Cubist	**0.8013**	0.0000	0.9897	0.0074	0.0028	**1.0798**
NCP	OLS	1.0421	0.4143	0.9898	0.0084	0.0018	1.2924
	XGBT	0.7776	**0.1854**	0.9897	0.0072	0.0031	1.0099
	Cubist	**0.7362**	0.1971	0.9895	0.0068	0.0037	**0.9412**
DP	OLS	0.9411	0.3014	**0.9891**	0.0073	0.0036	1.2124
	XGBT	0.7856	0.2193	0.9891	0.0077	0.0033	0.9866
	Cubist	**0.7353**	**0.2001**	0.9887	0.0071	0.0042	**0.9372**
DLUP	OLS	0.9158	0.3244	0.9882	0.0061	0.0057	1.1563
	XGBT	0.7635	0.1860	0.9889	0.0060	0.0051	0.9375
	Cubist	**0.7218**	**0.1845**	0.9888	0.0059	0.0054	**0.9145**

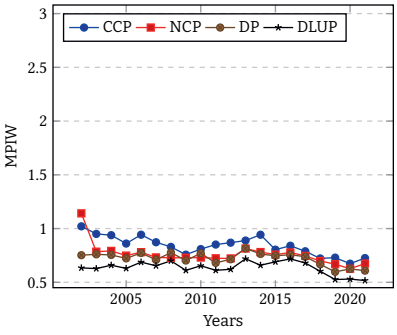
Table 3.A1.1: Results for the Four Different Prediction Interval Methods ($\alpha = 0.01$). *CCP* = Cross Conformal Prediction; *NCP* = Normalized Conformal Prediction; *DP* = Direct Prediction; *DLUP* = Direct Lower and Upper Prediction; *OLS* = Ordinary Least Squares; *XGBT* = Extreme Gradient Boosting Tree; *Algo.* = Algorithm; *MPIW* = Mean of the Prediction Interval Width; *SDPIW* = Standard Deviation of the Prediction Interval Width; *PI* = Price Interval; *MWIS* = Mean Winkler Interval Score.



(a) ***OLS***



(b) ***XGBT***



(c) ***Cubist***

Figure 3.A1.2: Mean Prediction Interval Width (MPIW) of the Rolling Window Results for the Four Different Prediction Interval Methods $\alpha = 0.01$. *OLS* = Ordinary Least Squares; *XGBT* = Extreme Gradient Boosting Tree; *CCP* = Cross Conformal Prediction; *NCP* = Normalized Conformal Prediction; *DP* = Direct Prediction; *DLUP* = Direct Lower and Upper Prediction.

"Confía en el tiempo, que suele dar dulces salidas a muchas amargas dificultades."

Miguel de Cervantes

4

Lead-lag Relationships in Quantile Index Percentiles

Authors: Felipe D. Calainho, Alex M. van de Minne, Radoslaw Trojanek, Marc K. Francke

Abstract

This paper examines whether different parts of the housing price distribution move together or whether some quantiles systematically lead others. Using 175,070 residential real estate transactions in Warsaw (2006–2024), we construct recentered influence function (RIF)–based quantile price indices (10th to 90th percentiles), smooth their returns using a generalized inverse estimator, and study their joint dynamics in a lag-weighted Lasso VAR complemented by generalized impulse responses and a low-dimensional VAR with Granger-causality tests. The quantile indices form a persistent fan with widening dispersion, showing that cheaper and more expensive segments drift apart over time. The dynamic analysis reveals that quantile returns do not move in lockstep: lower-middle and upper-middle segments, especially the 10th and 60th percentiles, contain disproportionate predictive content for the rest of the distribution, while the upper tail behaves mainly as a follower. The framework is portable and can be used to identify leading segments in other housing markets.

4.1 Introduction

Understanding the overall health and trajectory of housing markets is important for several obvious reasons. First, given that a large portion of many households' wealth accumulation is locked in their own homes, flagging dangers (and opportunities) early can help policymakers soften any macroeconomic spillover effects. For example, too much aggregate debt (compared to home values) can result in a decline in household mobility (Genesove and Mayer, 2001; Liebersohn and Rothstein, 2025). Likewise, understanding the trajectory of (local) housing markets allows banks to set interest rates in a fairer way, especially since mortgage interest rates are typically fixed for an extended period. Also, the (local) housing market is simply a good indicator of the overall health of the (local) economy. It is therefore no surprise that extant literature has taken an interest in finding time series that describe—or even predict—home values. Typical factors that influence house prices and are mentioned in existing literature include: construction costs, new supply, income/unemployment, interest rates/credit conditions, and population variables. (See, Ambrose et al., 2013; Englund and Ioannides, 1997; Fitzgerald et al., 2020; Harter-Dreiman, 2004, among many others.)

In this paper, the interest likewise lies in such *differences in timing* across the housing market, but viewed through a novel lens. More specifically, the analysis asks whether changes in some price quantiles tend to occur earlier than in others. A burgeoning literature has shown that price segments do not move in unison (Geltner and Minne, 2017; McMillen, 2008). We wish to exploit such differences in timing over segments in our search for a leading indication of the *overall* housing market. This approach has several benefits over using other covariates as leading indicators. For one, the choice of variables can be rather ad hoc, and certain variables might not be available everywhere, or for the same time period or frequency. In our proposed setup, one only need the underlying transaction data, which is typically (publicly) available through deeds or cadasters, and is needed to produce a house price index in the first place. A priori, it is not known which quantile will be leading; this is primarily an empirical question. In fact, this can be different for any given market. However, the next step is to outline possible drivers or channels through which certain segments *could* be leading.[1]

Extant literature has established that lower-income households respond the most to changes in interest rates, see Bhutta and Ringo (2021). Where higher income households might have a "cushion" to ride out a market lull, buyers of smaller price point properties will be immediately impacted in their borrowing capacity, and, therefore, prices will

[1] It should be noted that we do not investigate the potential drivers of the price segments themselves directly. We leave that to future research.

respond more quickly. A similar argument can be made for other economic variables like (aggregate) unemployment, general credit conditions and income. Similarly, Cvijanović and Spaenjers (2021) show that wealthy individuals have a lower marginal utility of wealth and relatively high opportunity costs of time, leading them to bargain less intensely compared to lower-income households. Changing circumstances might, therefore, play a more prominent role for lower-income households during the bargaining process. Median and average prices are typically the most widely reported by news media and governmental agencies. As such, there is a saliency argument to be made, where property buyers only find out about changes in market circumstances after media attention about the mean. As far as we are aware, there is no literature on this topic for housing markets. However, in general finance literature, there is plenty of research on the impact of media attentiveness on stock values. See, for example, Huberman and Regev (2001), Tetlock (2007, 2011), and Barber et al. (2022) for a few influential papers in this space.

For some or all of these reasons, certain price points might move first after some event. Obviously, it is never as clear-cut as indicated here. Wealthier individuals also care about interest rates and might be better or more sophisticated negotiators, for example. It should also be noted that we are not claiming that any potential findings indicate arbitrage opportunities in the market (Geltner and Minne, 2017). The argument is that the housing market is *segmented*, in the sense that lower-income or wealth households do not compete for large price point properties, and wealthier households might not be interested in smaller homes. This segmentation is causing differences in risk and return (and timings) over price points.

Methodology. To test this conjecture, the present paper takes the next step in this evolving literature by integrating RIF-based quantile estimation with time-series modeling. Specifically, it introduces two methodological extensions to the framework of McMillen and Shimizu (2021). First, it applies the Generalized Inverse Estimator (GIE) of Penrose (1955) to smooth the time series of quantile indices. GIE has been used in the literature to denoise repeat-sales indices (Bokhari and Geltner, 2012); here it is adapted to quantile indices, allowing stable index construction even in the presence of market sparsity and volatility. Second, it employs a Lag-weighted Lasso vector autoregression (LWL-VAR) model (Park and Sakaori, 2013) as a sparse variable-selection device for dynamic links across quantile returns, and then uses these selected connections to guide conventional VAR estimations for a subset of quantiles, on which standard Granger-causality tests are conducted. This two-step procedure allows the lead–lag patterns suggested by the penalized LWL-VAR to be corroborated by formal hypothesis testing in low-dimensional conventional VAR systems.

Data. We estimate our proposed methodology using micro-level transaction data for Warsaw in Poland, spanning the period from 2006 to 2024. For 175 thousand transactions, we observe: sales prices, time of sale, locational, distance to the nearest park and subway, distance to the city center, size and age of the building, and other variables, like architectural style, and whether or not there is storage and parking included. Poland is an interesting laboratory, as it has been rocked by multiple challenges that overlap with our analyzed time period, which allows us to find which quantile was leading. Other than the usual suspects—the GFC and COVID—Poland eventually had problems with many mortgages being originated in Swiss Francs during the mid-2000s, and welcomed and sheltered 2 million Ukrainian refugees at one point in the wake of the Russian invasion in 2022 (compared to a Polish population of less than 38 million). Many of the refugees have returned or moved to another country by now (mostly Germany). The impact on the housing market was unexpected and large.

Results. The empirical results suggest that Warsaw house prices are systematically segmented: cheaper and more expensive parts of the market remain at different levels and drift further apart over time. The quantile indices constructed with the RIF method spread out as the sample progresses, and the smoothed versions preserve this broad pattern while reducing high-frequency volatility in the series. In the LWL-VAR framework, smoothed index returns deliver lower forecast errors than regular index returns. This is partly mechanical: smoothing reduces high-frequency noise that is intrinsically unpredictable, so any reasonable forecasting model will perform better on the smoothed series. The contribution of the analysis is therefore not that smoothing improves forecastability per se, but that once a cleaner signal is obtained, only a sparse subset of quantiles appears to contain systematic predictive information for the rest of the distribution. Therefore, the resulting coefficient matrices for the LWL-VAR model highlight a small set of dominant cross-quantile links. In particular, the 10th and 60th percentiles display systematically non-zero coefficients toward neighbouring quantiles at short horizons of one quarter, while mid-quantiles around the 45th and 75th percentiles play a much more passive role. Generalized impulse response functions based on the preferred two-lag specification corroborate this pattern, with shocks to the 10th and 60th percentiles generating strong, system-wide but transitory reactions, and shocks to the 45th and 75th percentiles producing weaker and less distinctive responses, with large confidence bands.

Robustness and Additional Findings. Several checks support the stability of these findings. Summary statistics and unit-root tests confirm that the smoothed indices behave as highly persistent, non-stationary series, whereas their returns are stationary with only moderate autocorrelation, providing an appropriate input for VAR analysis.

Correlation matrices show that smoothing mainly removes idiosyncratic noise while leaving the strong common trend and high cross-quantile co-movement intact. Across alternative LWL-VAR specifications, different lag lengths and nominal versus real returns, the same qualitative structure re-emerges, with a prominent 60th-percentile column and a weaker but visible role for the 10th percentile, while models based on RIF returns (not smoothed) consistently underperform and produce coefficient patterns with many small, scattered effects and no clear cross-quantile structure. As an additional check, an unrestricted VAR(2) in a five-quantile subset (10th, 45th, 60th, 75th, 80th) shows that system-wide Granger causality tests reject the null that any single series fails to forecast the remaining four, indicating a richly interconnected system; within this VAR, the 60th percentile delivers the strongest Granger statistic, and pairwise tests reveal bidirectional predictability between the 10th and 45th percentiles and between the 10th and 60th percentiles, together with predominantly one-directional links from the 10th and 60th percentiles toward higher quantiles and a more "receiving" role for the 80th percentile. The 45th percentile appears more active in this unpenalized, low-dimensional VAR than in the high-dimensional LWL-VAR setting, which is consistent with its predictive content being highly collinear with that of the 60th percentile and therefore absorbed by the latter when shrinkage is imposed. Taken together, the LWL-VAR results, the unrestricted VAR, and the Granger tests all point to a tightly connected system in which the lower-middle and upper-middle segments, especially the 10th and 60th percentiles, contain disproportionate predictive content for the rest of the distribution. In addition, the RIF quantile indices document that, over the 2006–2024 period, lower quantiles appreciate somewhat faster than upper quantiles, indicating a gradual tightening of the lower part of the Warsaw price distribution.

Contribution. The contribution of the study is both substantive and methodological. Substantively, the results demonstrate that distributional dynamics within the housing market contain information about future price movements that is not visible from aggregate indices alone, and that, in the Warsaw case, an entry-level (10th percentile) and an upper-middle (60th percentile) segment tend to move one quarter ahead of the rest of the distribution. Methodologically, the study assembles RIF-based quantile indices, GIE smoothing, LWL-VAR estimation, and generalized impulse responses into a unified framework that can be implemented using only transaction-level data. This framework extends existing quantile index work from static description to dynamic lead–lag analysis under high dimensionality, while explicitly accommodating changes in the composition of the housing stock. Applied to Warsaw, it identifies the 10th and 60th percentiles as leading segments, but its main value lies in providing a portable, data-driven procedure

that can be used to uncover whichever quantiles are leading in other markets and periods.

Roadmap. The remainder of the paper is structured as follows. Section 2.2 outlines the methodology, including the construction of the different index specifications and the LWL-VAR model. Section 2.3 describes the dataset and preprocessing steps. Section 2.4 presents the empirical findings and interprets the dynamic relationships between quantiles. Section 2.5 concludes with a discussion of policy and research implications.

4.2 Methodology

4.2.1 Motivation and Conceptual Framework

The construction of residential property price indices traditionally relies on mean-based methods, such as hedonic pricing or repeat sales models. While these approaches offer a concise summary of average price trends, they conceal rich heterogeneity in how different market segments evolve over time. As McMillen (2008) first argued, aggregate indices can obscure asymmetric movements within the distribution, for instance, when upper-tier properties appreciate rapidly while lower-tier prices stagnate. Such phenomena are critical for understanding distributional dynamics, price dispersion, and market segmentation. Consequently, a growing strand of literature has turned toward estimating and analyzing quantile-based price indices (see, for example, Machado and Mata, 2005; McMillen and Shimizu, 2021; McMillen, 2008).

This literature builds on earlier work on wage and outcome decompositions such as the Oaxaca–Blinder framework (Blinder, 1973; Oaxaca, 1973), which focuses on mean differences between groups. While powerful, mean-based decompositions cannot characterize changes across the entire outcome distribution. The quantile decomposition of Machado and Mata (2005) was a key step forward, allowing researchers to construct counterfactual distributions and quantile indices using conditional quantile regression. However, it remains simulation-intensive and does not yield a simple additive decomposition of unconditional [2] quantiles.

A further methodological development came with the Recentered Influence Function (RIF) regression of Firpo et al. (2009). The RIF expresses an unconditional statistic as the av-

[2] In this context, "unconditional" refers to quantiles of the marginal distribution of transaction prices, $Q_\tau(Y)$, computed without conditioning on any particular bundle of dwelling characteristics. By contrast, "conditional" quantiles $Q_{Y|X}(\tau)$ describe the distribution of prices for dwellings with given characteristics (for example, a 60 m^2 flat in a particular district). Because quantiles do not satisfy a law-of-iterated-expectations or "tower" property, $Q_\tau(Y)$ cannot in general be recovered by averaging or aggregating conditional quantiles, and changes in the composition of X over time can move conditional measures even when the underlying marginal distribution is unchanged. Unconditional quantile indices are therefore designed to track shifts in the overall market-wide price distribution rather than shifts driven purely by who happens to transact in a given period.

erage of observation-specific pseudo-outcomes: for each observation, the influence function records how much the statistic would change if that observation were given slightly more weight, and these recentered values are then regressed on covariates to show how shifts in covariates move the unconditional statistic without the need to simulate full counterfactual distributions. McMillen and Shimizu (2021) apply this framework to real estate, constructing quantile housing price indices that describe how the unconditional distribution of prices evolves over time. The present paper follows that line of work and constructs RIF-based quantile indices from hedonic pricing models.[3]

Interpreting quantile indices as "price segments" is useful but requires some care. Price segments in practice are often conceptualized as discrete bins (e.g., "lower tier", "middle tier", "upper tier"), whereas quantiles are points on a continuous distribution. A high quantile (say, the 90th percentile) will predominantly reflect upper-tier properties, but it may also include the very best units of a middle-tier segment or the lower end of an ultra-luxury segment. Conversely, what market participants consider the "upper tier" in one period may shift in relative terms as the distribution of prices and demand changes. In other words, quantiles summarize relative positions in the price distribution, not fixed product categories. Throughout this paper, its therefore used the language of "price segments" as a convenient shorthand for regions of the price distribution, while recognizing that the mapping between quantiles and market segments is imperfect and time-varying.

Evidence from hedonic pricing models reinforces this interpretation of quantile indices as proxies for price segments. Goodman and Thibodeau (1998) show that within a metropolitan area, dwellings with different combinations of structural and locational characteristics trade in distinct submarkets, each with its own structure of implicit prices for housing attributes. In their setting, units of lower quality or in less desirable locations occupy a different part of the hedonic function than high-quality, well-located dwellings, so that movements in one submarket need not be mirrored one-for-one in another. This perspective emphasizes that the citywide price distribution is an aggregation of several partially segmented markets rather than a single homogeneous one. Against this background, the quantile indices in the present study can be viewed as proxies of these underlying submarkets: lower quantiles are populated disproportionately by low-quality or peripheral units, while higher quantiles reflect higher-quality, better-located stock.

Within this conceptual framework, the methodological choices are guided by a single goal: to assess whether certain parts of the price distribution move before others. To do

[3] Earlier quantile index applications often rely on conditional quantile regressions, which characterize the distribution of prices given covariates. RIF-based methods instead approximate how covariates shift the unconditional quantiles of the price distribution, see Firpo et al. (2009) and McMillen and Shimizu (2021) for a detailed discussion. For brevity, the measures are referred to as "quantile indices" throughout.

so, the analysis proceeds in a sequence of steps. First, it constructs a panel of RIF-based quantile price indices from transaction-level data, following the unconditional approach of Firpo et al. (2009) and McMillen and Shimizu (2021). Second, because quarterly quantile indices can be noisy and exhibit a low signal-to-noise ratio, their log returns are smoothed using the Generalized Inverse Estimator (GIE) of Penrose (1955). This step acts as a frequency-conversion and denoising procedure, yielding more stable series for time-series analysis. Third, the dynamic interaction between quantile returns is modeled using a Lag-weighted Lasso vector autoregression (LWL-VAR) in the spirit of Park and Sakaori (2013), which addresses the high dimensionality of the setting (17 indices and multiple lags) and embeds the economic intuition that recent information is more informative than distant history. Fourth, generalized impulse response functions in the sense of Pesaran and Shin (1998) are computed from the LWL-VAR to trace how shocks to selected quantiles propagate through the entire distribution over time in an order-invariant way. Fifth, as a conventional, low-dimensional complement to this penalized framework, unrestricted VAR(2) models are estimated for a selected subset of quantiles and standard Wald-type Granger non-causality tests are carried out, both system-wide and pairwise, to verify that the leading segments suggested by the LWL-VAR also emerge in an unpenalized VAR. The quantile indices, their smoothing, the LWL-VAR, the Generalized Impulse Response Functions (GIRFs), and the auxiliary VAR-based Granger tests thus work together to uncover and validate the dynamic structure across the housing price distribution.

4.2.2 RIF-based Quantile Index

Let i denote transactions and t denote quarters. This study works with log transaction prices

$$y_i^t = \log P_i^t,$$

and a vector of property and location characteristics X_i^t (structural attributes, locational dummies, distances to amenities, building age, etc.). For a set of quantile levels $\tau \in \mathcal{T}$ (covering lower-, middle-, and upper-priced parts of the market), our target object is the *unconditional* τ-quantile of y_i^t in each quarter t, i.e. the τ-th percentile of the overall price distribution in that period, rather than the quantile conditional on a particular dwelling type.

4.2.2.1 Influence functions and recentered influence functions

The starting point is the notion of an *influence function* from robust statistics.[4] Given any distributional statistic $\nu(F)$ (for example a mean, a quantile, or a Gini coefficient), the influence function $\mathbf{IF}(y;\nu)$ answers the following question:

If we put a tiny bit more probability mass on the value y, how much would the statistic ν change?

For the τ-quantile q_τ of Y, the influence function takes a simple form:

$$\mathbf{IF}(y;q_\tau) \;=\; \frac{\tau - \mathbf{1}\{y \le q_\tau\}}{f_Y(q_\tau)},$$

where $f_Y(q_\tau)$ is the density of Y at q_τ and $\mathbf{1}\{\cdot\}$ is an indicator function. Observations below the quantile ($y \le q_\tau$) receive a negative value, observations above receive a positive value, and the magnitude reflects how "locally sparse" the distribution is around q_τ (through $f_Y(q_\tau)$). Intuitively, $\mathbf{IF}(y;q_\tau)$ assigns each sale a score that measures how strongly that particular transaction would pull the τ-th quantile up or down if it became slightly more common. Because the average of the influence function is zero, it is convenient to "recenter" it around the statistic of interest:

$$\mathbf{RIF}(y;q_\tau) \;\equiv\; q_\tau + \mathbf{IF}(y;q_\tau).$$

By construction it holds that $\mathbb{E}\big[\mathbf{RIF}(Y;q_\tau)\big] \;=\; q_\tau$, where $\mathbb{E}\big[\,\cdot\,\big]$ denotes the expectation. So the recentered influence function behaves like a dependent variable whose unconditional mean is exactly the quantile we care about. This is the key trick in Firpo et al. (2009). It turns a nonlinear object (a quantile) into the expectation of a transformed outcome, which can then be studied with familiar regression tools.

In our panel notation, the unconditional τ-quantile of y_i^t in quarter t is denoted q_τ^t, and the corresponding recentered influence function for observation (i,t) is

$$\mathbf{RIF}(y_i^t;q_\tau^t) = q_\tau^t + \mathbf{IF}(y_i^t;q_\tau^t).$$

4.2.2.2 RIF–regressions

Let

$$m_\tau^t(x) \;\equiv\; \mathbb{E}\big[\mathbf{RIF}(Y_i^t;q_\tau^t) \mid X_i^t = x\big],$$

[4] See Firpo et al. (2009) for a detailed introduction in the context of unconditional quantile regressions.

denote the (unknown) conditional expectation of the recentered influence function given characteristics $X_{it} = x$ in quarter t. Using the law of iterated expectations, it is possible to write the unconditional quantile as

$$q_\tau^t \;=\; \mathbb{E}\big[\mathbf{RIF}(Y_i^t; q_\tau^t)\big] \;=\; \mathbb{E}\big[m_\tau^t(X_i^t)\big].$$

In words: if we knew how the RIF varies with X_{it}, then averaging this relationship over the distribution of characteristics would give us back the unconditional τ-quantile.

The *RIF–regression* approach approximates $m_\tau^t(x)$ by a linear function of x:

$$\mathbf{RIF}(y_i^t; q_\tau^t) \;=\; (X_i^t)^\top \beta_\tau^t + \varepsilon_i^t. \tag{4.1}$$

estimated by ordinary least squares for each quarter t and quantile τ. The coefficient vector β_τ^t summarizes how differences in property and location characteristics are associated with shifts in the *unconditional* τ-quantile of log prices, rather than with conditional quantiles for a fixed dwelling type. In the terminology of Firpo et al. (2009), these coefficients approximate "unconditional quantile partial effects": the effect of shifting the distribution of a covariate on the marginal quantile q_τ^t.

Operationally, for each (t, τ) it proceed as follows:

1. Estimate the unconditional τ-quantile q_τ^t of y_i^t using the empirical distribution of prices in quarter t.
2. Estimate the density $f_{Y^t}(q_\tau^t)$ at the quantile (using a kernel estimator).
3. Compute $\mathbf{RIF}(y_i^t; q_\tau^t)$ for all transactions in quarter t.
4. Run the linear regression (4.1) of the recentered influence function on X_i^t to obtain $\hat{\beta}_\tau^t$.

4.2.2.3 From RIF–regressions to quantile indices

To obtain a time series of quantile indices, we follow the chain-linking strategy of McMillen and Shimizu (2021). For each quarter t and quantile τ, we use the estimated coefficients $\hat{\beta}_\tau^t$ together with the empirical distribution of characteristics in that quarter, X_t, to obtain a prediction of the unconditional τ-quantile,

$$\hat{q}_\tau(X^t, \hat{\beta}_\tau^t),$$

where $\hat{q}_\tau(\cdot)$ denotes the prediction implied by the RIF–regression averaged over the observed X_i^t in quarter t. Because y_i^t are log prices, differences in these predicted quantiles can be interpreted as log price changes at quantile τ.

The log change in the τ-quantile between quarter t and $t+1$ is

$$\Delta q_\tau^{(t\to t+1)} = \hat{q}_\tau(X^{t+1}, \hat{\beta}_\tau^{t+1}) - \hat{q}_\tau(X^t, \hat{\beta}_\tau^t), \tag{4.2}$$

4

and starting from a base-period index normalized to 100, the (not smoothed) RIF-based quantile index is obtained by cumulative chain-linking:

$$\mathbf{Index}_\tau^{(t)} = 100 \cdot \exp\left(\sum_{s=1}^{t} \Delta q_\tau^{(s-1\to s)}\right). \tag{4.3}$$

By construction, $\mathbf{Index}_\tau^{(t)}$ tracks the evolution of the unconditional τ-quantile[5] of the log price distribution, adjusting for changes in the composition of transacting dwellings in each quarter. These indices can be interpreted as proxies for the evolution of different price segments of the market, subject to the conceptual caveats discussed above. Finally, note that the period-by-period log return of the τ-quantile index is simply

$$r_\tau^t \equiv \log \mathbf{Index}_\tau^{(t)} - \log \mathbf{Index}_\tau^{(t-1)} = \Delta q_\tau^{(t-1\to t)}.$$

and these log returns are the main input to the smoothing and VAR analysis that follows.

4.2.3 Smoothed Quantile Index

Quarterly quantile indices, especially in the tails of the distribution, may display substantial short-run volatility. This volatility can arise from limited numbers of transactions within a quarter, changing sample composition, and idiosyncratic transaction shocks. In turn, excessive short-run variation lowers the signal-to-noise ratio of the estimated quantile returns and may obscure the dynamic relationships that the subsequent vector autoregression (VAR) analysis is intended to capture. To attenuate this problem, the Generalized Inverse Estimator (GIE) of Penrose (1955) is applied to the RIF-based quantile returns, following the logic of frequency-conversion and alignment procedures used in the repeat-sales literature by Bokhari and Geltner (2012).

The key idea is that the raw quantile returns are not treated as the final object of interest, but as noisy measurements of an underlying latent sequence of lower-frequency returns.

[5] In this study the quantiles estimated are 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50, 0.55, 0.60, 0.65, 0.70, 0.75, 0.80, 0.85, and 0.90. In total there are 17 quantile indices being estimated.

In the present setting, the procedure begins by constructing several overlapping quantile index series for each quantile $\tau \in \mathcal{T}$. With quarterly data, four separate aligned index series are estimated. Each of these series starts in a different quarter of the calendar year and then proceeds in annual steps over observations occupying the same within-year position. Hence, one series links first quarters across years, another links second quarters, and so forth. Because these aligned series are shifted relative to one another, they contain complementary information about the same latent quarterly return process.

This alignment step is useful because each shifted series provides a partial and noisy view of the underlying quarterly dynamics. Consider the quarterly case. One aligned series summarizes changes from the first quarter of one year to the first quarter of the next year, another from the second quarter of one year to the second quarter of the next year, and similarly for the remaining quarters. Each observed annual change therefore aggregates several adjacent latent quarterly returns. The GIE is then used to reconcile these overlapping annual changes and recover the quarterly return sequence that best fits all aligned series simultaneously in a least-squares sense.

Formally, for a given quantile τ, let $\mathbf{y}_\tau$ denote the stacked vector of observed aligned returns obtained from all shifted index series. Each element of $\mathbf{y}_\tau$ is an observed lower-frequency return that can be written as the sum of several consecutive latent quarterly returns. Let $\mathbf{w}_\tau$ denote the latent quarterly return vector to be recovered. The relationship between observed aligned returns and latent quarterly returns can then be expressed as

$$\mathbf{y}_\tau = \mathbf{X}\mathbf{w}_\tau + \boldsymbol{\varepsilon}_\tau, \tag{4.4}$$

where $\mathbf{X}$ is an incidence matrix indicating which latent quarterly returns enter each observed aligned return, and $\boldsymbol{\varepsilon}_\tau$ collects measurement noise and approximation error. In the quarterly case, each row of $\mathbf{X}$ contains four consecutive ones, reflecting the fact that an annual return spans four quarters. The rows are shifted by one quarter across observations, so that the full system links together all overlapping aligned returns.

The GIE estimate of the latent return sequence is then obtained as

$$\hat{\mathbf{w}}_\tau = (\mathbf{X}'\mathbf{X})^{+}\mathbf{X}'\mathbf{y}_\tau, \tag{4.5}$$

where $(\mathbf{X}'\mathbf{X})^{+}$ denotes the Moore–Penrose pseudoinverse. The pseudoinverse is required because the alignment system is generally not square and may be rank-deficient. Intuitively, Equation (4.5) selects the quarterly return path that comes closest to reproducing all observed aligned returns at once. Rather than relying on a single noisy index path, the procedure pools information across all shifted alignments and extracts the common latent

component that is most consistent with the data.

This interpretation is important. The method does not merely smooth the series mechanically. Instead, it imposes a coherent temporal structure: observed aligned returns are assumed to be linear combinations of adjacent latent quarterly returns, and the estimator recovers the latent sequence that best rationalizes those overlapping observations. As a result, high-frequency idiosyncratic noise that is specific to one alignment tends to be averaged out, while persistent movements that appear consistently across alignments are retained. The resulting series is therefore smoother, but it remains grounded in the return structure implied by the estimated quantile indices.

For each quantile $\tau \in \mathcal{T}$, the estimated latent returns $\{\hat{w}_{\tau,t}\}_{t=1}^{T}$ are then cumulated to obtain the smoothed quantile index,

$$\widetilde{\text{Index}}_{\tau}^{(t)} = 100 \cdot \exp\left(\sum_{s=1}^{t} \hat{w}_{\tau,s}\right). \tag{4.6}$$

This transformed series preserves the medium- and long-run trajectory of the original quantile index while reducing erratic quarter-to-quarter fluctuations. In the subsequent time-series analysis, the VAR is estimated using the smoothed log returns $\hat{w}_{\tau,t}$ rather than the raw returns $r_{\tau,t}$, thereby improving the signal-to-noise ratio without discarding economically meaningful variation across quantiles.

The procedure implemented for each quantile can be summarized as follows:

1. Estimate several aligned quantile index series, each beginning in a different quarter of the year.

2. Convert each aligned index into a sequence of aligned log returns.

3. Stack all aligned returns into a vector $\mathbf{y}_{\tau}$ for quantile τ.

4. Construct an incidence matrix $\mathbf{X}$ that maps each observed aligned return to the set of latent quarterly returns it spans.

5. Apply the Moore–Penrose pseudoinverse to obtain the least-squares estimate of the latent quarterly return vector according to Equation (4.5).

6. Cumulate the estimated latent returns to obtain the smoothed quantile index in Equation (4.6).

4.2.4 Dynamic Interaction and Predictive Structure: LWL-VAR

Our main interest lies in whether movements in some quantiles of the price distribution systematically precede movements in others. To study these dynamic interactions, this paper models the joint evolution of quantile returns using a Lag-weighted Lasso vector autoregression (LWL-VAR) as proposed by Park and Sakaori (2013).

Let $\mathscr{T}$ contain K quantile levels (in our application, $K = 17$), and define the K-dimensional vector of smoothed log returns at time t as

$$R^t = \left(\hat{w}^t_{\tau_1}, \hat{w}^t_{\tau_2}, \ldots, \hat{w}^t_{\tau_K}\right)'.$$

Because the underlying indices are non-stationary (they exhibit persistent trends), modeling them in levels would risk spurious regression results. Working with log returns R^t instead removes these unit roots and yields series that are approximately covariance-stationary, which is the standard assumption for VAR analysis.[6]

This paper considers a VAR(p) representation for the K-dimensional return vector:

$$R^t = c + \sum_{l=1}^{p} \Phi^{(l)} R^{t-l} + \varepsilon^t, \tag{4.7}$$

where c is a vector of intercepts, $\Phi^{(l)}$ are $K \times K$ coefficient matrices, and ε_t is a K-dimensional error term. With $K = 17$ and a moderate lag order p, the unrestricted VAR quickly becomes high-dimensional: there are $K^2 \cdot p$ slope coefficients, which can easily exceed what is reliably estimable given the available quarterly sample size.

To address this, we estimate Equation (4.7) using a lag-weighted LASSO penalty (Park and Sakaori, 2013):

$$\min_{\{c,\Phi^{(1)},\ldots,\Phi^{(p)}\}} \left\{ \frac{1}{2} \sum_{t=p+1}^{T} \left\| R^t - c - \sum_{l=1}^{p} \Phi^{(l)} R^{t-l} \right\|_2^2 + \lambda \sum_{l=1}^{p} \omega_l \left\| \Phi^{(l)} \right\|_1 \right\}, \tag{4.8}$$

where $\|\cdot\|_1$ is the element-wise ℓ_1 norm, $\|\cdot\|_2$ is the Euclidean norm, $\lambda > 0$ is a global penalty parameter, and $\{\omega_l\}_{l=1}^{p}$ are non-negative lag weights. The lag weights encode the idea that more distant lags are less informative and should be shrunk more aggressively; for instance, one can choose ω_l to be increasing in l or to follow a tapered (decaying) structure. Economically, this reflects the expectation that recent price movements contain more relevant information for current pricing decisions than events many years in the past,

[6] Formal unit-root tests on the index levels support the presence of stochastic trends; results are available in Section 2.4.

particularly in segmented and informationally imperfect housing markets.

In practice, this work implements Equation (4.8) using the "Tapered" specification of the `BigVAR` package (Nicholson et al., 2017), which corresponds exactly to the lag-weighted LASSO structure in Equation (4.8). Throughout the paper, this study refers to the resulting specification simply as the LWL-VAR. The tuning parameter λ is selected through an expanding-window, one-step-ahead forecasting exercise. Starting from an initial estimation window, the sample is recursively expanded over time. At each time t in the validation period, the model is re-estimated using all observations up to $t-1$, and a one-step-ahead forecast $\hat{R}^{t|t-1}_{(\lambda)}$ is produced for each candidate value of λ. The corresponding forecast errors $e^t_{(\lambda)} = R^t - \hat{R}^{t|t-1}_{(\lambda)}$ are collected and summarized by the mean squared forecast error (MSFE),

$$\text{MSFE}(\lambda) = \frac{1}{KN} \sum_{t \in \mathcal{T}_{\text{val}}} \left\| R^t - \hat{R}^{t|t-1}_{(\lambda)} \right\|_2^2, \tag{4.9}$$

where K is the number of series and N is the number of one-step-ahead forecasts in the validation window $\mathcal{T}_{\text{val}}$. The algorithm chooses the value of λ that minimizes this Out-of-Sample (OOS) MSFE and then re-estimates the LWL-VAR on the full sample using this optimal penalty.

In the results, the associated in-sample loss is reported as In-sample, which is the average of squared residuals for the penalized VAR computed over the estimation window. The OOS corresponds to the average MSFE of the penalized LWL-VAR forecasts over the validation period as in Equation (4.9). For comparison, this study also compute several benchmark losses over the same validation window: (i) the conditional mean OOS loss (Mean) is obtained by replacing the penalized VAR forecast with a predictor that always forecasts the sample mean of each series; (ii) the Akaike Information Criterion (AIC) and (iii) Baysian Information Criterion (BIC) OOS losses are computed using forecasts from unpenalized VAR models whose lag order is chosen by minimizing the AIC and BIC, respectively; and the (iv) Random Walk (RW_{walk}) OOS loss evaluates a naive benchmark in which the one-step-ahead forecast for each series equals its last observed value.[7] These metrics document that the chosen penalized specification achieves a low MSFE relative to conventional benchmarks, and the regularization level is calibrated to improve OOS predictive performance rather than in-sample fit alone.

Formally, let T denote the total number of quarterly observations for the return series. Following Nicholson et al. (2017), the sample is split into an initial estimation block, a rolling cross-validation block, and a final evaluation block. Define $T_1 = \lfloor T/3 \rfloor$ as the start of the cross-validation window and $T_2 = \lfloor 2T/3 \rfloor$ as the start of the true out-of-sample evaluation

[7] See Nicholson et al. (2017) and the package documentation for details on these performance metrics.

window. For each candidate penalty λ_i on the grid, the algorithm runs a rolling expanding-window scheme: at each origin $t = T_1, \dots, T_2 - 1$ the VAR is estimated on all observations up to t, a one-step-ahead forecast for y_{t+1} is generated, and the resulting squared forecast errors are averaged over the $N_{\mathsf{CV}} = T_2 - T_1$ cross-validation forecasts. The value λ^* minimizes this cross-validation MSFE. Given λ^*, the model is then used to produce genuinely out-of-sample one-step-ahead forecasts over the final block $t = T_2 + 1, \dots, T$, yielding $N_{\mathsf{OOS}} = T - T_2$ observations on which the reported OOS losses in Table 4.4 are computed. In our application, $T = 71$ so that $T_1 = 23, T_2 = 47, N_{\mathsf{CV}} = 24$, and $N_{\mathsf{OOS}} = 24$.

The LWL-VAR yields a sparse set of estimated cross-quantile lag coefficients. Non-zero off-diagonal elements in $\Phi^{(l)}$ at short lags indicate that returns at quantile τ_j help forecast returns at quantile τ_k beyond the information contained in the latter's own history. This paper interprets such patterns as evidence of lead–lag relationships between different parts of the price distribution. In Section 4.2.5, This study complements these coefficient-based results with GIRF that trace the dynamic response of each quantile to shocks in selected parts of the distribution.

4.2.5 Generalized Impulse Response Function

To further interrogate and visually validate the dynamic structure implied by the LWL-VAR, This work computes Generalized Impulse Response Functions (GIRFs) in the sense of Pesaran and Shin (1998). More information on their calculation including confidence bounds is provided in Appendix Section 4.A3. Unlike orthogonalized impulse responses based on Cholesky decompositions, GIRFs do not require orthogonalizing the shocks and are invariant to the ordering of variables in the VAR. In our context, GIRFs provide an order-robust way of tracing how an innovation to one quantile of the price distribution propagates to all other quantiles over time.

It is important to clarify how this impulse–response framework should be interpreted in a quantile setting. In macroeconomic applications, an impulse is often thought of as a shock to a clearly defined variable such as GDP or unemployment, for which it is at least conceptually plausible to imagine a disturbance hitting that series. By contrast, a price quantile is not a single object but a summary of an entire segment of the cross-sectional distribution, and price segments themselves are tightly linked to location and quality tiers rather than mapping one-to-one into quantiles. Even if one neighbourhood or quality tier were affected by a local event (for example, a flood or fire that disproportionately impacts dwellings in a particular price range), any resulting adjustment in prices would almost inevitably spill over to neighbouring quantiles. Thus, one should not interpret our GIRFs as describing a literal experiment in which only a single quantile is shocked while all oth-

ers are untouched. Instead, a generalized impulse here is best viewed as a stylized disturbance that is *concentrated* in a given quantile equation within the estimated LWL-VAR system, while allowing for contemporaneous comovement through the full residual covariance matrix.

Within this interpretation, the generalized impulse responses serve as a robustness check and a complementary visualization of the LWL-VAR results. Because the GIRFs are order-invariant and derived from the same estimated lag-weighted VAR, they provide an internally consistent way to examine whether the quantiles that appear as dynamically important in the coefficient matrices and Granger-causality tests also generate economically plausible and statistically meaningful paths of adjustment throughout the price distribution when subjected to a standardized innovation. This paper therefore uses the GIRF analysis not to identify structural "shocks" to specific market segments, but to summarize in a transparent way how the estimated multivariate dynamics redistribute a localized disturbance across different parts of the housing price distribution over time.

4.2.6 Low-Dimensional VAR and Granger-Causality Tests

As a robustness check on the dynamic structure uncovered by the LWL-VAR and GIRFs, the analysis also estimates a conventional low-dimensional VAR and performs standard Granger-causality tests. To keep the system tractable and the results easy to interpret, the VAR is restricted to five smoothed quarterly log-return series, chosen to represent different parts of the price distribution (lower, middle, and upper segments). Stacking these returns into a vector Y^t, the benchmark specification is an unrestricted VAR with two quarterly lags and a constant term,

$$Y^t = c + \Phi_1 Y^{t-1} + \Phi_2 Y^{t-2} + u^t, \tag{4.10}$$

where c is a vector of intercepts, Φ_1 and Φ_2 are coefficient matrices, and u^t is a vector of innovations. A two-lag structure is a standard choice for quarterly data: it captures short-run dynamics without introducing an excessive number of parameters. The VAR(2) in (4.10) is estimated equation-by-equation using ordinary least squares, and stability is verified by checking that all roots of the characteristic polynomial lie strictly inside the unit circle.

Granger causality is then examined at two levels, following standard practice in multivariate time-series analysis. First, *system-wide* tests ask whether the past of a given quantile return improves the joint prediction of the remaining four series. Formally, for each candidate series, a block-exogeneity null is tested that sets all coefficients on its lagged values in the other equations equal to zero. This restriction is evaluated with a Wald F-test

in the sense of Granger (1969); rejection at conventional significance levels indicates that the history of that quantile contains predictive information for at least one of the other segments, over and above their own lags. Instantaneous-causality tests based on the residual covariance matrix provide a complementary check for contemporaneous dependence that is not captured by lagged terms.

Second, *pairwise* Granger-causality tests are conducted for each ordered pair of the five quantiles. For a given pair $(Y^t_{(i)}, Y^t_{(j)})$, a bivariate VAR(2) is estimated, and two null hypotheses are tested: (i) past values of $Y^t_{(i)}$ do not help predict $Y^t_{(j)}$ once past values of $Y^t_{(j)}$ are controlled for, and (ii) the reverse direction. Each null corresponds to a joint restriction on the relevant lag coefficients and is assessed via an F-test. Rejection in one direction but not the other points to a unidirectional predictive relation between those two segments, while rejection in both directions suggests bidirectional predictability.

The five-variable VAR, the system-wide Granger tests, and the pairwise Granger-causality analysis thus provide a standard, low-dimensional benchmark. They allow the lead–lag patterns suggested by the sparse LWL-VAR and GIRFs to be cross-checked in an unpenalized setting that is familiar in the time-series literature, thereby strengthening the overall interpretation of which parts of the price distribution tend to move earlier.

4.3 Data

This section describes the dataset and all variables used in the following hedonic price estimation. The entire data set covers 180,898 residential real estate transactions in Warsaw, the capital and largest city of Poland, spanning from the second quarter of 2006 to the fourth quarter of 2024. After applying data cleaning criteria explained below, 175,070 unique observations remain, each of which records the sale of a single apartment. The raw data is provided by the official Polish Property Price Register. Only transactions in the secondary market are considered, where full private ownership rights in multi-family residential buildings are sold. Multi-family dwellings are the most common form of housing in Warsaw.

In order to link the database with administrative register, additional variables including building height and year of construction were merged using the official Polish Cadastre. In the case of incomplete cadastral data, building technology was coded from Google Maps Street View images in a manual data collection process. All property addresses were converted into longitude and latitude coordinates by the Google Maps API to compute distance variables to urban amenities. The data collection and merging procedure is presented in Trojanek and Huderek-Glapska (2018).

Distance variables are based on geographic coordinates. In particular, distance from a property to the closest public park was computed using vector layers from Warsaw City Hall WMS server. Similarly, distance to the closest subway station was measured using official subway locations. Distance from the property to the city center is based on longitude and latitude of the geographic centroid of the housing unit. The Palace of Culture and Science was used as a proxy of the city center.

After merging all relevant data sources, several observations are dropped from the raw transaction database to improve comparability. First, all non-market transactions were dropped from the sample, which includes debt-enforcement sales, intra-family transfers, and inheritances. Second, transactions with multiple dwellings were also excluded. Third, to focus on multi-family housing, properties located in single-family houses were dropped from the sample. Fourth, to remove extreme observations, housing units smaller than 20 m^2 or larger than 300 m^2 were also excluded, after all these four procedures we are left with 180,898 observations.

Finally, hedonic regression diagnostics were used to identify outliers. Cook's distance was computed for each observation using the commonly used $4/(n-k)$ cutoff, where n is sample size and k is the number of regressors. Observations with large values of Cook's distance have a large influence on the estimated hedonic price regression. This process was repeated separately for each year of the sample to avoid the potential problem of extreme house price changes that happened during the observation period. After removing observations based on Cook's distance, 175,070 observations are left. Table 4.1 summarizes the raw data and cleaned sample.

Variable	Raw	Cleaned
Price per m^2 (PLN)	9,923	9,946
Dwelling area (m^2)	53.5	53.4
Building age (years)	34.0	34.8
Observations	180,898	175,070
Retention rate (%)	100.0	96.8
Sample period: 2006:Q2–2024:Q4 (75 quarters)		

Table 4.1: Summary statistics before and after data cleaning.

Cleaning the data reduces the sample size by about 3.2% or 5,828 observations. Importantly, mean price per square meter increases by less than 0.25%, from 9,923 PLN to 9,946 PLN, suggesting that removal of outliers does not have a material impact on the sample. The mean dwelling area is virtually unchanged at 53.4 m^2 after the cleaning. However, the mean building age increases from 34.0 years to 34.8 years, which suggests that the

cleaning procedure was dropping particularly young observations. The final cleaned sample consists of 175,070 georeferenced transactions with multi-family apartments sold in Warsaw.

4.3.1 Variable Specification

Table 4.2 shows the complete list of variables that enter into the hedonic price models. It includes temporal, spatial, and structural dwelling characteristics. Temporal heterogeneity is captured by quarterly dummy variables that span the entire sample period. Spatial heterogeneity is controlled for by district fixed effects, where all 18 administrative districts of Warsaw[8] are included. Structural characteristics of the housing stock include a mix of continuous (dwelling area, building age, distance to urban amenities) and categorical variables (construction technology, parking and storage availability).

Variable	Type	Description
Temporal		
Quarter	Dummy	Quarter ($q_1, \ldots, q_{75}$) for 2006:Q2–2024:Q4
Spatial		
District	Dummy	District fixed effects for Warsaw's 18 administrative districts
Park	Continuous	Distance to nearest public park (meters)
Subway	Continuous	Distance to nearest subway station (meters)
City center	Continuous	Distance to city center (meters)
Structural: continuous		
Area	Continuous	Usable floor area (m^2)
Age	Continuous	Building age at transaction (years)
Structural: binary		
Technology	Dummy	= 1 if traditional construction, 0 if prefabricated
Storage	Dummy	= 1 if basement/storage unit included
Parking	Dummy	= 1 if garage/parking space included

Table 4.2: Variable definitions.

Some of these variables require additional explanation. The construction technology variable differentiates between prefabricated (mainly large-panel system buildings) and traditional buildings constructed from brick. Many of these prefabricated buildings, mostly built during the socialist era, command lower prices as they are considered to be of lower quality and to have limited architectural appeal. Age of the building was calculated as the difference between the year of sale and the year of construction.

[8] Warsaw has a total of 1.861.599 inhabitants and 517.81 km^2 according to Statistics Poland (GUS) and Statistical Office in Warsaw. This represents a density of 3.595 people per km^2.

Spatial variables reflect accessibility to urban amenities. Park distance indicates proximity to green spaces, which are an important urban amenity, particularly in dense city centers. Subway distance is a measure of accessibility to Warsaw's metro system. Distance to the city center is a proxy for accessibility to urban economic opportunities and agglomeration economies in general.

To avoid the problem of many small and sporadically observed districts, a simple frequency-based aggregation rule was used to combine small districts. For each district and quarter, the number of transactions was tabulated. Because the RIF-based hedonic pricing models are estimated separately for each quarter, we require that a district have at least five sales in that quarter to be modelled separately; district–quarter cells with fewer than five observations are recoded into an artificial "rest" category. This aggregation is applied when estimating district fixed effects in the hedonic regression models.

4.4 Results

4.4.1 Index Return Discussion

Figure 4.1 presents the RIF-based quantile index returns for Warsaw together with their smoothed counterparts obtained via the GIE. In the left panel, the index returns are highly volatile, especially in the first years of the sample, with several quantiles experiencing changes in excess of ±10% per quarter. Even later in the sample, the lines are jagged and quantile-specific spikes are frequent. The mean return is itself noisy and it is difficult to distinguish systematic cross-quantile patterns from idiosyncratic variation.

The right panel of Figure 4.1 displays smoothed index returns. Here, the early spikes vanish, most observations lie within a band of roughly ±5%, and all quantile series form a tight bundle. The co-movement across quantiles is visually much clearer, and any cross-sectional differences are small and persistent rather than irregular. In other words, smoothing shifts the focus from short-lived, quantile-specific shocks toward the common cyclical component that is relevant for dynamic modelling. Working with smoothed index returns therefore offers a cleaner basis for identifying lead–lag relations in the LWL-VAR.

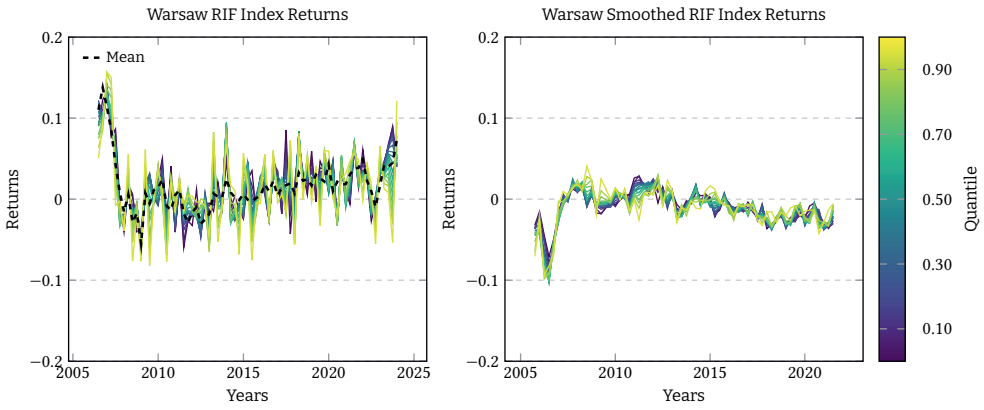


Figure 4.1: Warsaw RIF Quantile Index Returns. *RIF* = Recentered Influence Function.

Figure 4.2 displays the corresponding quantile indices. The left panel shows the indices (not smoothed), where a mean index is also plotted as a reference. All series are normalized to 100 at the start of the sample. The quantile fan exhibits a well-ordered, upward-sloping pattern: lower quantiles remain above the mean and upper quantiles bellow it throughout the sample. The spread between the tails widens notably in the later years, indicating that lower quantiles appreciate faster than upper quantiles. The mean index tracks

the central part of the distribution, but the divergence between, say, the 10th and 90th percentiles reveals a non-trivial increase in dispersion that would be masked by looking only at central tendencies (mean index).

The right panel of Figure 4.2 shows the smoothed indices. The overall narrative remains the same: a strong boom in the mid-2000s, a slowdown around the Global Financial Crisis, followed by a sustained price increase. However, high-frequency fluctuations and short-lived crossings between neighboring quantiles largely disappear. The quantile curves become smoother and more parallel, and their relative ordering is much clearer.

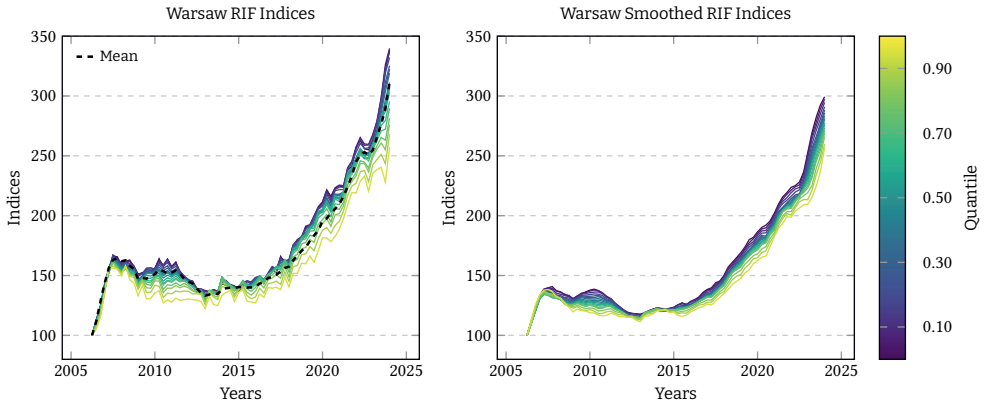


Figure 4.2: Warsaw Quantile Indices. *RIF* = Recentered Influence Function.

Table 4.3 reports summary statistics for the smoothed index returns. Average quarterly growth rates are between 1.3 and 1.5 percent across all quantiles, indicating that the long-run appreciation is broadly shared throughout the distribution. Return volatilities are of similar magnitude across quantiles, with only a mild increase toward the upper tail, which is consistent with the tightly clustered lines in the right panel of Figure 4.1. First-order autocorrelations range from about 0.61 to 0.78, so returns remain persistent but much less so than the levels. Most importantly, the Augmented Dickey–Fuller statistics are strongly negative and the associated p-values are below 0.01 for every quantile. The unit-root null hypothesis is therefore decisively rejected, confirming that the smoothed index return series are covariance-stationary and suitable for VAR analysis.

Table 4.A1.1 provides summary statistics for the smoothed quantile indices. The sample means are very similar across quantiles, reflecting the strong common trend already visible in Figure 4.2. The variances are largest at the lower quantiles and gradually decline toward the upper tail, which is consistent with somewhat more variation over time in the lower end of the distribution. The first-order autocorrelation coefficients are all around 0.91, indicating a high degree of persistence in levels. Augmented Dickey–Fuller statis-

Series	Mean	Std. dev.	ACF(1)	ADF stat.	ADF p-value
q: 0.10	0.015	0.026	0.748	-6.196	< 0.01
q: 0.15	0.015	0.026	0.781	-6.205	< 0.01
q: 0.20	0.015	0.026	0.760	-5.715	< 0.01
q: 0.25	0.015	0.025	0.765	-6.115	< 0.01
q: 0.30	0.015	0.025	0.729	-6.085	< 0.01
q: 0.35	0.015	0.025	0.718	-6.409	< 0.01
q: 0.40	0.015	0.025	0.729	-6.617	< 0.01
q: 0.45	0.015	0.025	0.721	-7.047	< 0.01
q: 0.50	0.015	0.025	0.709	-7.417	< 0.01
q: 0.55	0.014	0.025	0.730	-7.722	< 0.01
q: 0.60	0.014	0.025	0.744	-7.365	< 0.01
q: 0.65	0.014	0.025	0.725	-7.670	< 0.01
q: 0.70	0.014	0.025	0.696	-7.420	< 0.01
q: 0.75	0.014	0.025	0.670	-7.503	< 0.01
q: 0.80	0.014	0.026	0.663	-7.240	< 0.01
q: 0.85	0.013	0.026	0.610	-7.473	< 0.01
q: 0.90	0.013	0.028	0.614	-5.461	< 0.01

Table 4.3: Smoothed Index Log Returns: Summary Statistics. Means and standard deviations are for quarterly log returns of the smoothed RIF-based quantile indices. ACF(1) denotes the first-order autocorrelation. ADF stat. and ADF p-value refer to Augmented Dickey–Fuller tests for a unit root. *Notes:* All series have $N = 71$ quarterly observations; the ADF regressions include 2 lags.

tics are positive with p-values above 0.99 for all quantiles, so the unit-root null cannot be rejected. Together, these statistics confirm that the smoothed indices behave like non-stationary, highly persistent series, which motivates working with log returns in the time-series analysis.

Appendix Figures 4.A1.1 and 4.A1.2 provide correlations between the quantile index returns, for the regular RIF-based quantile index returns and the smoothed ones, respectively. In Appendix Figure 4.A1.1, correlations of RIF-based quantile returns are positive and exhibit a clear distance effect: neighbouring quantiles typically correlate in the 0.85–0.95 range, while the extremes can fall to around 0.65. This is consistent with the noisy, quantile-specific spikes in the left panel of Figure 4.1. By contrast, Appendix Figure 4.A1.2 shows that correlations between smoothed index returns are uniformly high, mostly above 0.90 even for distant quantiles, and the gradient across the matrix flattens considerably. This mirrors the tightly clustered lines in the right panel of Figure 4.1 and the stationary, moderately persistent behaviour documented in Table 4.3.

The correlations between the quantile indices, and their respective returns, are provided in the appendix. Appendix Figure 4.A1.3 shows that the RIF-based quantile indices are already extremely tightly comoving: pairwise correlations between the mean index and each quantile, and across quantiles themselves, are all above 0.98. This confirms that all parts of the distribution share a dominant common trend, consistent with the fan-like

pattern in the left panel of Figure 4.2 and with the high persistence and non-stationarity documented in Appendix Table 4.A1.1. After smoothing, the correlations in Appendix Figure 4.A1.4 move even closer to one, indicating that the GIE has mainly removed small, idiosyncratic wiggles rather than altering the underlying co-movement.

4.4.2 Lead-lag Analysis

In this study, model performance is evaluated along several complementary dimensions. The key measure is the OOS loss, which reports the mean squared one-step-ahead forecast error of the penalized VAR over the validation window. Lower values indicate better forecasting accuracy. This OOS loss is compared to a set of simple benchmarks. The "Mean" column corresponds to a conditional mean forecast that always predicts the sample mean of each series; the "AIC" and "BIC" columns report losses from unpenalised VARs whose lag order is selected by minimising the Akaike and Bayesian information criteria, respectively; and the "RW_{walk}" column corresponds to a random-walk benchmark that forecasts each series by its most recently observed value. A penalised specification is regarded as useful only if its OOS loss is no worse than, and ideally clearly below, these benchmark losses. The "In-sample" column reports the average squared residuals over the estimation window and mainly reflects within-sample fit, while the "Frac. active" column records the proportion of VAR coefficients that remain non-zero after penalization, providing a summary of model sparsity. Together, these metrics allow an assessment of whether a given specification achieves both competitive forecast performance and a parsimonious, interpretable coefficient structure.

Table 4.4 summarizes the EW cross-validation performance of the different LWL-VAR specifications considered. All entries report one-step-ahead MSFE computed under the same rolling forecasting scheme and penalty selection procedure described in Section 4.2. The RIF model, which uses RIF-based index returns, performs clearly worst: its out-of-sample MSFE (0.02130) is an order of magnitude larger than that of any smoothed specification and even larger than the simple benchmarks (Mean, AIC, BIC, RW_{walk}). This confirms that feeding the noisy, index returns into the VAR renders the system essentially unforecastable and underscores the need for GIE smoothing.

Among the three nominal "Smooth" models, the differences in out-of-sample MSFE are relatively small. The four-lag specification attains the lowest loss (0.00335), followed by the one-lag model (0.00346) and the two-lag model (0.00383). However, the four-lag model also has by far the highest fraction of active coefficients (0.51), and also the highest number of total coefficients (1156). More than half of all potential coefficients survive shrinkage, but many of them are only slightly different from zero, so the predictive signal is dispersed

Model	Lag	λ^*	Frac. active	In-sample	OOS	Mean	AIC	BIC	RW
Smooth	2	0.0006177	0.1987	0.00270	0.00383	0.01010	0.00583	0.01010	0.00399
RIF	2	0.0007305	0.4092	0.02260	0.02130	0.01950	0.02690	0.01950	0.03700
Smooth	1	0.0009226	0.3248	0.00259	0.00346	0.00909	0.00456	0.00909	0.00399
Smooth	4	0.0001637	0.5109	0.00254	0.00335	0.01130	0.00794	0.01010	0.00399
Real Smooth	2	0.0006377	0.2060	0.00396	0.00533	0.01090	0.00981	0.01090	0.00619

Table 4.4: LWL-VAR model performance summary (squared-error loss, one-step-ahead horizon $h = 1$). *Lag* denotes the VAR lag order; λ^* is the penalty parameter selected by rolling cross-validation on a grid of 50 candidate values; and *Frac. active* is the proportion of non-zero VAR coefficients. *In-sample* and *OOS* report, respectively, the in-sample and out-of-sample mean squared forecast errors (MSFE) of the penalized LWL-VAR. *Mean*, *AIC*, *BIC*, and *RW* give the out-of-sample MSFE of, respectively, a conditional mean forecast, an unpenalized VAR with lag order chosen by AIC, an unpenalized VAR with lag order chosen by BIC, and a random-walk benchmark. All models are estimated on $T = 71$ quarterly observations of quantile returns; λ^* is chosen using 24 rolling cross-validation origins ($t = 23, \dots, 47$), and all out-of-sample MSFEs are computed from $N_{\text{OOS}} = 24$ one-step-ahead forecasts ($t = 48, \dots, 71$).

over a large number of small effects. By contrast, the two-lag smoothed model has the lowest fraction of active coefficients (about 0.20), meaning that roughly 80% of all possible coefficients are shrunk exactly to zero. The remaining coefficients are larger in magnitude and concentrated in a small set of cross-quantile links. In other words, the two-lag model achieves a competitive MSFE while encoding the dynamics in a much sparser and more interpretable way, with a few strong connections rather than many weak ones. The "Real Smooth" specification, which uses real rather than nominal smoothed returns, also outperforms the benchmarks but exhibits a higher out-of-sample MSFE (0.00533), suggesting that deflating by inflation does not enhance short-run predictive content for our purposes.

In light of this trade-off between fit and parsimony, this study treats the smoothed two-lag LWL-VAR as our baseline specification for the GIRF part. It delivers substantial improvements over the not smoothed index returns and benchmark models (Conditional Mean, AIC, BIC, and RW_{walk}), while concentrating the predictive structure in a relatively small number of economically meaningful coefficients instead of relying on a dense web of marginal effects.

All coefficient plots in Figures 4.3 and 4.4, plus the Appendix Figures 4.A2.1, 4.A2.2, and 4.A2.3 share the same layout and colour scale. Each figure shows the lag-weighted VAR coefficient matrices for the 17 quantile-return series. Rows correspond to the equations (the quantiles whose returns are being predicted), columns to the regressors (lagged returns of each quantile), and both rows and columns are ordered from the 10th to the 90th percentile. Separate panels display the different lag matrices $\Phi^{(1)}, \Phi^{(2)}, \dots$, so moving from left to right corresponds to moving further back in time (one-quarter, two-quarter lags, etc.). Darker cells indicate coefficients with larger absolute value, while nearly white cells correspond to coefficients that have been shrunk to (almost) zero by the lag-weighted Lasso.

A vertical band of darker cells therefore signals that a given quantile helps forecast many others, whereas a horizontal band would signal that a given quantile is strongly influenced by many others. Interpreting the figures in this way makes it immediately clear whether the system behaves like a "single factor" (everything moving together) or whether different parts of the distribution interact in a more segmented fashion.

In the baseline specification with smoothed nominal returns and two lags (Figure 4.3), the first-lag matrix $\Phi^{(1)}$ shows a pronounced vertical band around the 60th-percentile column and a weaker band around the 10th percentile. These columns contain non-zero coefficients in many rows, meaning that past movements in these quantiles enter a large share of the forecasting equations across the distribution. Columns associated with the median and upper quantiles are much lighter and have relatively few active cells, so their lagged returns are selected far less often as predictors. The second-lag matrix $\Phi^{(2)}$ is considerably sparser, with only a few faint cells, suggesting that once we condition on the first lag of the smoothed indices, additional lagged information contributes relatively little to the fitted dynamics. The corresponding coefficient values (with quantile-labelled rows and columns) are reported in Appendix Table 4.A2.1.

Reducing the lag order to one (Figure 4.4) compresses this structure into a single matrix. The overall picture is very similar: the mid-distribution column associated with the 60th percentile remains the darkest and most pervasive, and the lower-tail column around the 10th percentile again shows up as a secondary source of predictive content. Because the entire dynamics must now be captured in $\Phi^{(1)}$, more cells are lightly shaded, indicating that the regularization spreads a comparable total amount of signal over a larger number of small coefficients, but the qualitative cross-quantile pattern remains unchanged. For completeness, the full set of estimated coefficients for this $p = 1$ specification is provided in Appendix Table 4.A2.2.

When the lag order is increased to four (Figure 4.A2.3), the first-lag matrix still looks close to the baseline two-lag case, with visible bands around the 10th and 60th percentiles. The additional lag matrices $\Phi^{(2)}$–$\Phi^{(4)}$ display many more light cells scattered across rows and columns, reflecting the extra degrees of freedom: the model is allowed to use a larger set of small lagged effects, and the penalty responds by activating more coefficients with modest magnitude. The fact that these later-lag panels are pale and diffuse, compared with the sharper structure in $\Phi^{(1)}$, suggests that the economically relevant dependencies are predominantly short-run, while further lags mostly pick up residual noise.

Estimating the same two-lag model on smoothed *real* returns (Figure 4.A2.1) leads to an almost identical configuration. The strongest vertical band again lies in the middle of the distribution, with a weaker but discernible column in the lower tail, and the second-lag

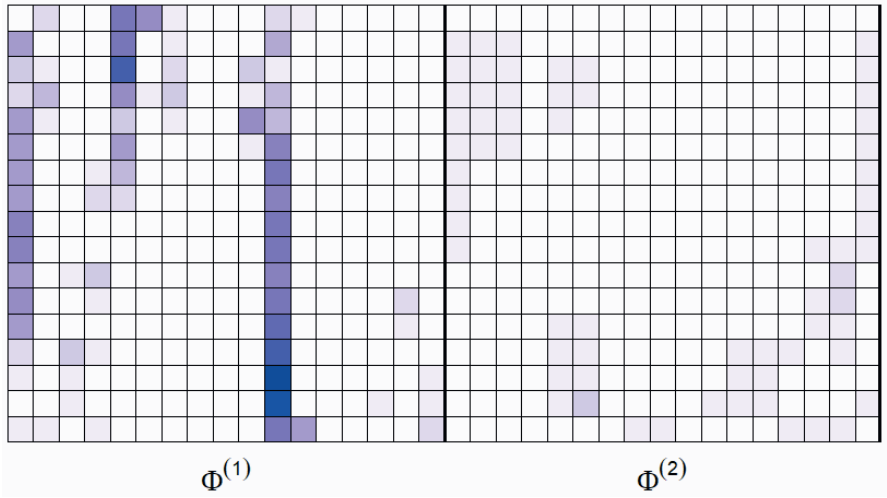


Figure 4.3: Lag-weighted Lasso VAR (order $p = 2$) for smoothed RIF-based quantile-return indices. Panels report the lag coefficient matrices $\Phi^{(1)}$ (lag $t-1$) and $\Phi^{(2)}$ (lag $t-2$), where rows are the predicted quantile returns at time t and columns are lagged quantile returns. Darker cells indicate larger $|\Phi_{ij}^{(n)}|$ (lighter cells ≈ 0).

matrix remains relatively sparse. The colours are slightly lighter overall—consistent with somewhat smaller coefficients in real terms—but the pattern of which quantiles transmit information to which others is remarkably robust to the nominal/real distinction.

A different picture emerges once smoothing is switched off. In the model based on the RIF-based returns (Figure 4.A2.2), the coefficient mass is much less concentrated: colored cells appear in many more positions across both $\Phi^{(1)}$ and $\Phi^{(2)}$, and the contrast between a few dominant columns and the rest of the matrix is far weaker. Several mid-range quantiles (such as those around the 0.30 and 0.45 levels) show noticeable activity, but without a clear, coherent band structure. Visually, the heatmap looks much "busier", consistent with the idea that the model is now trying to track high-frequency movements and idiosyncratic noise, in line with the weaker forecasting performance reported for this specification in Table 4.4.

Taken together, these plots show that the quantile indices do *not* move in lockstep. Across all specifications with smoothed returns—nominal or real, and with different lag orders—the matrices display a structured pattern of vertical bands concentrated in particular parts of the distribution, rather than a single uniform block. Some quantiles systematically exert more influence on others, while some are mainly followers, which is exactly what is expect in a segmented housing market where different price tiers adjust at different speeds. The specification using index returns (not smoothed), by contrast, produces a much more diffuse pattern, reinforcing the conclusion that smoothing the

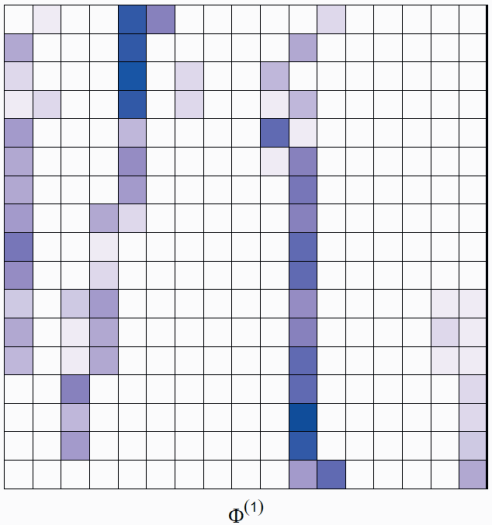


Figure 4.4: Lag-weighted Lasso VAR (order $p = 1$) for smoothed RIF-based quantile-return indices. The panel reports $\Phi^{(1)}$, the coefficient matrix on $\mathbf{y}_{t-1}$ (lag $t-1$); rows are equations and columns are lagged regressors. Darker cells indicate larger $|\Phi^{(1)}_{ij}|$ (lighter cells ≈ 0).

indices and using a parsimonious lag structure helps reveal this segmentation in a sparse and interpretable way.

4.4.3 Generalized Impulse Response Function Results

Figures 4.5 and 4.6 report generalized impulse response functions in the sense of Pesaran and Shin (1998), based on the preferred two-lag LWL-VAR estimated on smoothed RIF returns. Each figure considers a one-standard-deviation innovation in a particular quantile and traces the response of all seventeen quantile returns over a twelve-quarter horizon, together with 95% bootstrap confidence bands. Across all four shocks the responses are positive on impact, decay monotonically toward zero within roughly three to four years, and rarely cross zero within the first few quarters. This pattern is consistent with the VAR results, which suggested that shocks propagate rapidly through the system but have a fairly short half-life.

The GIRFs also corroborates the leading role of the lower-middle and middle-upper segments highlighted by the LWL-VAR coefficient matrices. When the 10th or 60th quantile is shocked (Figures 4.5 and 4.6), all other quantiles react contemporaneously and significantly, with the strongest responses typically observed in nearby quantiles and in the central part of the distribution. The impact is largest in the first one or two quarters and then gradually fades, indicating that innovations in these segments transmit quickly to the rest

of the market. This is exactly what would be expected if the 0.10 and 0.60 quantiles play a leading role in the dynamic structure: they not only respond to aggregate conditions but also help shape the subsequent evolution of other parts of the distribution.

By contrast, shocks to quantiles whose LWL-VAR coefficients are essentially zero, such as the 0.45 and 0.75 quantiles (Appendix Figures 4.A3.1 and 4.A3.2), do not generate disproportionate or highly persistent reactions in other quantiles. The GIRFs remain positive and quickly revert to zero, and the confidence bands widen rapidly after a few quarters. Because generalized impulse responses allow shocks to propagate through the contemporaneous covariance matrix of residuals, some response is still observed even when a given quantile does not Granger-cause the others directly. However, the magnitude and persistence of these responses are comparable across quantiles and do not single out the 0.45 or 0.75 quantiles as dynamically special. Taken together, the GIRF analysis therefore acts as a robustness check on the LWL-VAR findings: it confirms that innovations in the 10th and especially the 60th quantile have system-wide repercussions, whereas shocks originating in quantiles with zero LWL-VAR coefficients do not appear to drive the rest of the distribution in a systematic way.

In our Warsaw application, the joint evidence from the LWL-VAR and GIRFs thus points to the 10th and 60th percentiles as segments that tend to move first, but this should be interpreted as context-specific rather than universal. Over the sample period of this study, the 10th percentile can reasonably be viewed as an entry-level segment populated by more credit-constrained households, while the 60th percentile corresponds to an upper-middle segment in which trading is relatively frequent and price discovery is rapid. Within this particular market and time frame, it is plausible that macro–financial shocks first show up where households sit closest to their borrowing constraints and where transactions are most informative, and that the rest of the distribution adjusts with a lag.

A key implication of our analysis, however, is not that policymakers or market participants should always monitor the 10th and 60th percentiles per se, but rather that the identity of any "leading" quantile is an empirical question that will generally depend on local market structure and the period under consideration. More broadly, the analysis sets out a framework—based on RIF quantile indices, GIE smoothing, LWL-VAR estimation, and GIRF—for systematically uncovering lead–lag relationships across the price distribution. Applied to Warsaw in 2006–2024, this framework highlights the 10th and 60th percentiles as leading segments. In other markets or periods, the same procedure may highlight different quantiles. The key message is that the framework provides a disciplined way to identify them rather than prescribing a fixed set of "universal" leading indicators.

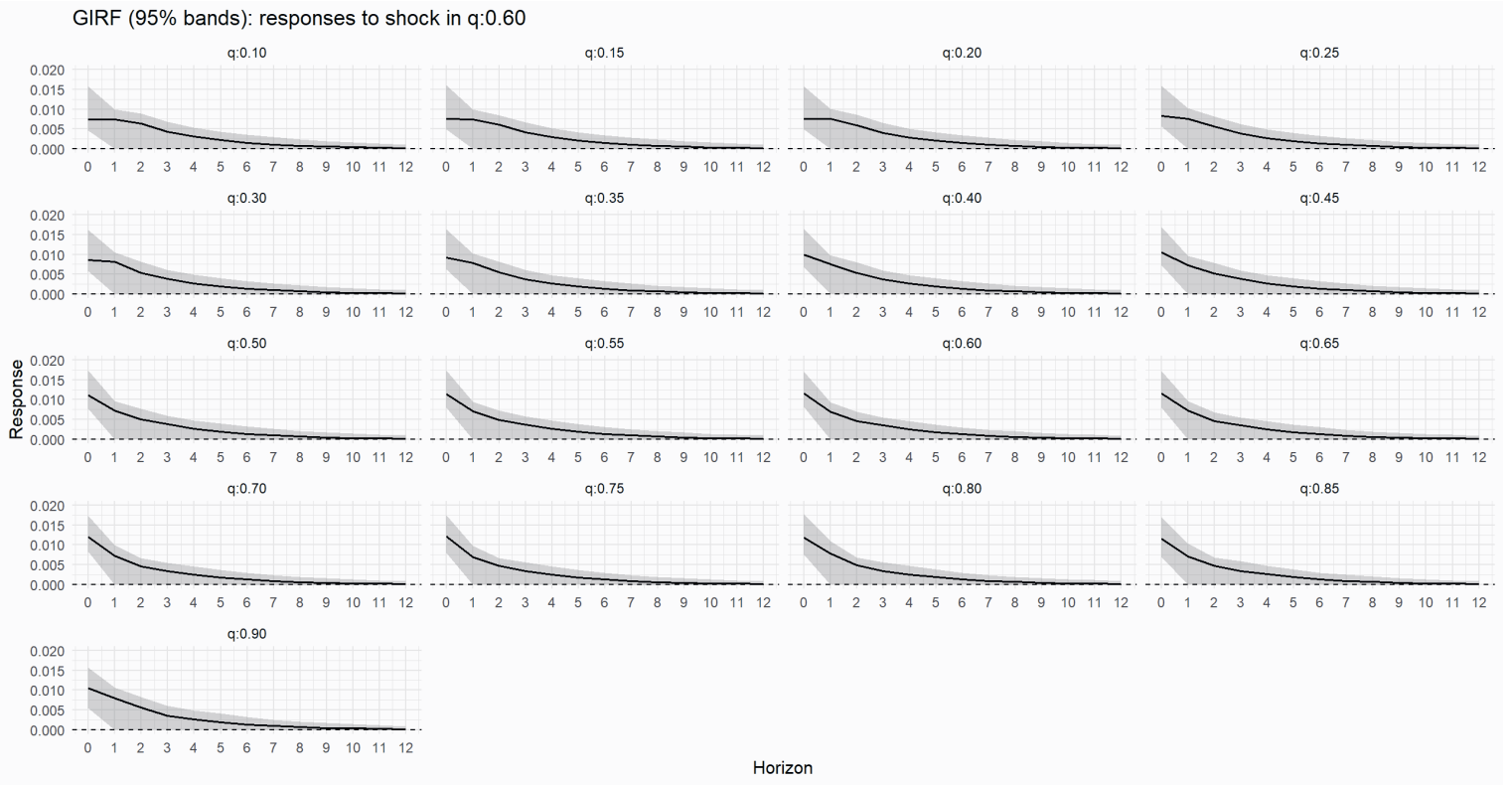


Figure 4.5: Generalized Impulse Response Function (GRIF). One Standard Deviation shock in the 60th Quantile.

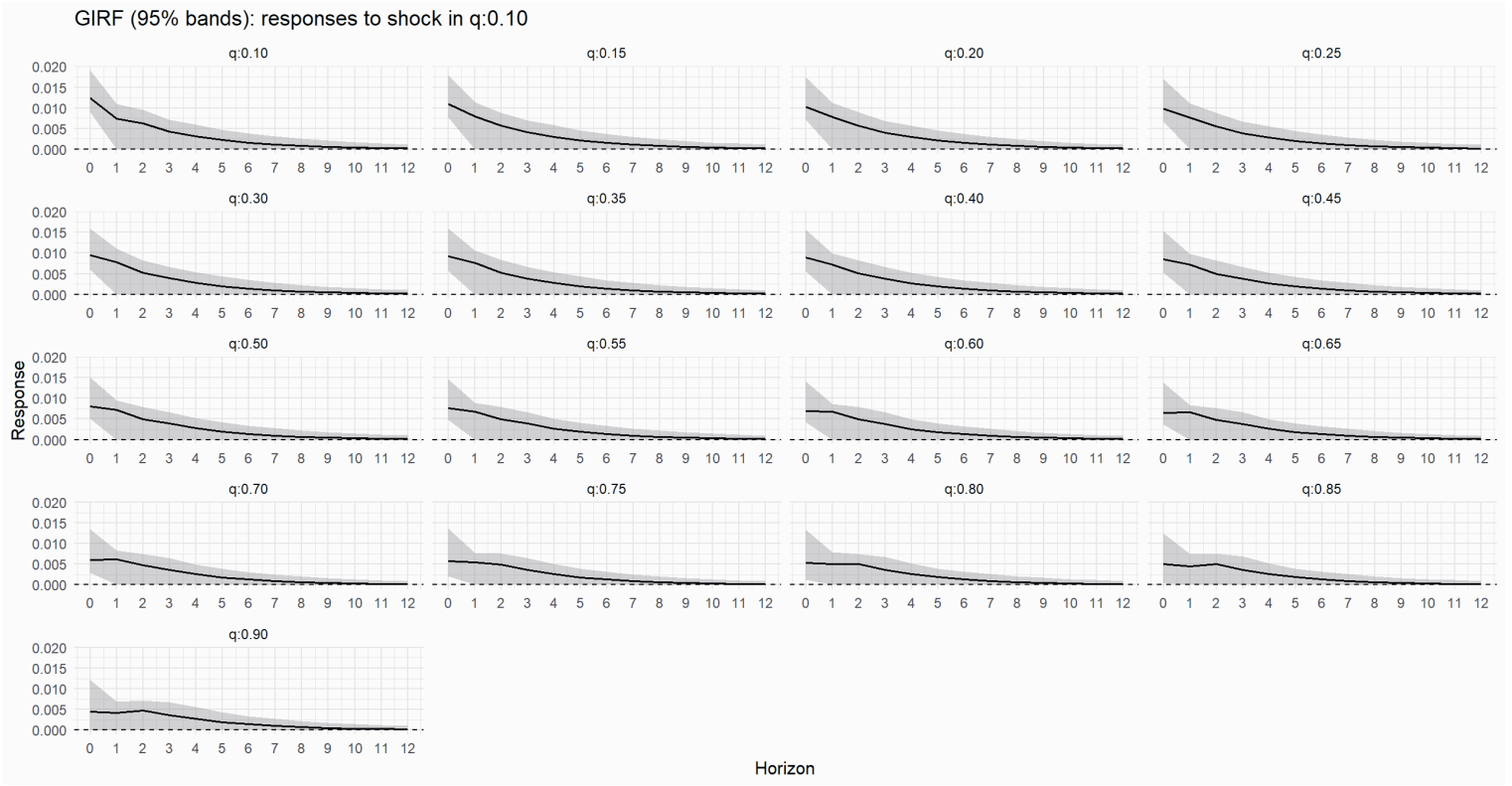


Figure 4.6: Generalized Impulse Response Function (GRIF). One Standard Deviation shock in the 10th Quantile.

4.4.4 Reduced VAR and Granger Causality Tests

As a complement to the high-dimensional LWL-VAR, the study also estimates an unrestricted VAR(2) for a reduced system of five smoothed quantile returns (0.10, 0.45, 0.60, 0.75, 0.80). These quantiles are chosen to span lower, middle, and upper parts of the distribution and to reflect the heterogeneous roles suggested by the LWL-VAR coefficient maps: $q_{0.10}$ and $q_{0.60}$ are among the most frequently active quantiles, $q_{0.45}$ and $q_{0.75}$ have all coefficients shrunk to zero, and $q_{0.80}$ shows the largest estimated response to $q_{0.60}$ at lag 1. This lower-dimensional VAR is easier to interpret equation by equation and allows us to check whether the dynamic patterns highlighted by the penalized model also appear in a conventional, unpenalized setting. The system-wide Granger-causality tests in Table 4.5 reject, at the 5% level, the null that any of the five quantiles has no predictive content for the remaining four: lagged values of each selected quantile improve the prediction of at least some others beyond their own past, indicating a tightly interconnected but asymmetric set of dynamics across price segments.

Cause	Affected variables	F-statistic	df_1, df_2	p-value
$q_{0.10}$	$q_{0.45}, q_{0.60}, q_{0.75}, q_{0.80}$	2.492	8, 290	0.0125
$q_{0.45}$	$q_{0.10}, q_{0.60}, q_{0.75}, q_{0.80}$	5.379	8, 290	2.5×10^{-6}
$q_{0.60}$	$q_{0.10}, q_{0.45}, q_{0.75}, q_{0.80}$	5.096	8, 290	6.0×10^{-6}
$q_{0.75}$	$q_{0.10}, q_{0.45}, q_{0.60}, q_{0.80}$	3.210	8, 290	0.0016
$q_{0.80}$	$q_{0.10}, q_{0.45}, q_{0.60}, q_{0.75}$	3.668	8, 290	4.3×10^{-4}

Table 4.5: System-wide Granger-causality tests in the five-variable VAR(2).

The individual VAR equations in Appendix Tables 4.A4.1, 4.A4.2, 4.A4.3, 4.A4.4, and 4.A4.5 provide more detail on how this predictability varies across quantiles. For the 10th percentile, no single lagged coefficient is individually significant at conventional levels, although the regression as a whole has a relatively high R^2, indicating that several small effects combine to explain a sizeable share of its variance. In contrast, the equation for $q_{0.45}$ features statistically significant effects from $q_{0.60,t-1}$ and $q_{0.75,t-1}$ (one-quarter lags) and from $q_{0.10,t-2}$ (two-quarter lag), suggesting that past movements in both lower and somewhat higher quantiles help predict the evolution of the central part of the distribution. The $q_{0.60}$ equation includes a significant effect of $q_{0.45,t-2}$, while the $q_{0.75}$ and $q_{0.80}$ equations both load significantly on $q_{0.60,t-1}$ and $q_{0.45,t-2}$. Thus, middle quantiles (around 0.45 and 0.60) play an important role in driving the dynamics of upper quantiles, whereas the top quantile rarely feeds back into the lower part of the system. The residual covariance and correlation matrices in Tables 4.A4.6–4.A4.7 show strong positive contemporaneous correlation among all five series, especially between adjacent quantiles, but

also a gradual increase in residual variance towards the upper quantiles, indicating that higher-priced segments are somewhat more volatile.

The pairwise Granger-causality tests in Table 4.6 further clarify the nature of these cross-quantile links. For most pairs, this study finds asymmetric relations: in eight out of ten pairs, only one direction is significant at the 5% level. The main exception is the lower part of the distribution: for the pairs $(q_{0.10}, q_{0.45})$ and $(q_{0.10}, q_{0.60})$, Granger causality is bidirectional, indicating a tight two-way interaction between the lower and middle quantiles. For other combinations, the predictive direction typically runs from lower or middle quantiles towards higher ones (e.g. $q_{0.45} \rightarrow q_{0.75}, q_{0.80}$ and $q_{0.60} \rightarrow q_{0.75}, q_{0.80}$), while we never reject the null that the top quantile $q_{0.80}$ fails to Granger-cause any of the others. Together with the equation-by-equation results, this pattern suggests that shocks tend to propagate upward through the distribution, with the upper tail behaving more as a responder than as an initiator of movements.

Pair	Direction	F-stat	p-value	Decision (5%)
q10, q45	q45 → q10	7.60	0.0011**	reject H_0
	q10 → q45	4.69	0.0125*	reject H_0
q10, q60	q60 → q10	3.52	0.0355*	reject H_0
	q10 → q60	4.46	0.0154*	reject H_0
q10, q75	q75 → q10	0.67	0.5151	do not reject
	q10 → q75	4.33	0.0173*	reject H_0
q10, q80	q80 → q10	0.50	0.6074	do not reject
	q10 → q80	3.46	0.0373*	reject H_0
q45, q60	q60 → q45	2.51	0.0895.	do not reject
	q45 → q60	5.24	0.0078**	reject H_0
q45, q75	q75 → q45	0.21	0.8133	do not reject
	q45 → q75	6.45	0.0028**	reject H_0
q45, q80	q80 → q45	0.64	0.5281	do not reject
	q45 → q80	5.40	0.0068**	reject H_0
q60, q75	q75 → q60	2.12	0.1285	do not reject
	q60 → q75	7.98	0.0008***	reject H_0
q60, q80	q80 → q60	1.96	0.1499	do not reject
	q60 → q80	9.82	0.0002***	reject H_0
q75, q80	q80 → q75	1.74	0.1834	do not reject
	q75 → q80	4.77	0.0117*	reject H_0

Table 4.6: Pairwise Granger causality tests between smoothed quantile returns (VAR(2)).

Notes: Each row reports an F-test for the null hypothesis H_0 that the lagged values of the "from" quantile do not Granger-cause the "to" quantile in a bivariate VAR(2) with a constant. "Reject H_0" indicates that this null is rejected at the 5% level, i.e. that the lags of the "from" quantile add statistically significant predictive power for the "to" quantile beyond its own lags. Significance codes: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$, $^{.}p < 0.15$.

Overall, the reduced five-variable VAR points to a rich but structured pattern of dynamic dependence across price segments. Quantile returns are strongly contemporaneously correlated, yet their past values contain additional information that is not symmetric across the distribution. These findings dovetail with the LWL–VAR heatmaps and GIRFs: both approaches indicate that price changes do not occur uniformly across the distribution, but instead ripple from some segments to others over time, reinforcing the view that housing markets exhibit persistent segmentation and staggered adjustment across quantiles.

4.5 Conclusion

This paper set out to examine whether different parts of the housing price distribution move together or whether some quantiles systematically lead others. Rather than relying on external macro–financial covariates, the analysis investigates whether the dynamics of the distribution itself contain information about the future path of the market. To address this question, the study proposes and implements a distributional time-series framework that combines Recentered Influence Function (RIF)-based quantile indices, Generalized Inverse Estimator (GIE) smoothing, a Lag-weighted Lasso vector autoregression (LWL-VAR), Generalized Impulse Response Functions (GIRFs), and a reduced-form vector autoregression (VAR) with Granger-causality tests.

Using micro-level data for Warsaw over 2006–2024, the analysis first constructs a panel of RIF-based quantile indices spanning the 10th to the 90th percentiles. The indices display a clear, well-ordered fan with widening dispersion in later years: lower quantiles remain above the mean and upper quantiles below it, with the lower tail appreciating somewhat faster. Summary statistics and unit-root tests show that these index levels are highly persistent and non-stationary, while their (smoothed) returns are stationary with moderate serial correlation. Applying the GIE to quantile log-differences eliminates high-frequency noise and stabilizes the relative ordering of the quantiles without altering the underlying common trend. The smoothed indices and their returns therefore provide a coherent and internally consistent description of how the entire price distribution has evolved.

On this basis, the joint dynamics of smoothed quantile returns are modeled using a LWL-VAR. Out-of-sample forecast comparisons show that models based on index returns (not smoothed) perform poorly, whereas specifications using smoothed returns substantially reduce forecast errors relative to standard benchmarks. Among the smoothed models, a two-lag LWL-VAR strikes a balance between fit and parsimony: roughly four-fifths of all potential coefficients are shrunk to zero, and the remaining non-zero

coefficients are economically interpretable. Coefficient heatmaps reveal structured heterogeneity rather than a uniform block: across one-, two-, and four-lag specifications in nominal and real terms, a small set of columns—most consistently around the 10th and 60th percentiles—contains many active coefficients, while others (notably around the 45th and 75th percentiles) are largely shrunk to zero. By contrast, the model estimated on index returns (not smoothed) produces a more diffuse and noisy pattern. Taken together, these results indicate that quantile returns do not move in lockstep: predictive links are concentrated in particular parts of the distribution, while other segments behave mainly as followers.

Generalized impulse response functions provide an additional, order-robust perspective on these dynamics. Generalized impulses concentrated in the 10th and 60th percentiles generate system-wide responses: all other quantiles react on impact, with the largest and most persistent effects occurring in nearby quantiles and in the central part of the distribution, and responses decaying smoothly over a horizon of roughly three to four years. In contrast, impulses in quantiles whose LWL-VAR coefficients are essentially zero (such as the 45th and 75th percentiles) produce smaller and less distinctive responses once contemporaneous covariance effects are taken into account. Interpreted as stylized disturbances within the estimated system rather than literal shocks to isolated market segments, the GIRF analysis corroborates the coefficient-based evidence: some parts of the distribution occupy more central positions in the propagation of disturbances, while others adjust mainly in response.

A conventional unrestricted VAR(2) in a reduced five-quantile system provides a complementary, low-dimensional check on these conclusions. System-wide Granger-causality tests reject the null that any of the selected quantiles fails to help forecast the remaining four, and pairwise tests reveal predominantly asymmetric relations: lower and middle quantiles tend to Granger-cause higher ones more often than the reverse, while the upper-most quantile behaves mainly as a receiver of information. Mid-quantiles such as the 45th percentile appear more active in this unpenalized setting than in the high-dimensional LWL-VAR, which is consistent with their predictive content being highly collinear with that of neighbouring quantiles and therefore absorbed by those neighbours when shrinkage is imposed. Overall, the reduced VAR confirms a tightly interconnected but segmented system, with movements tending to ripple through the distribution rather than all quantiles adjusting simultaneously.

Two types of implications follow. Substantively, the results for Warsaw over 2006–2024 suggest that monitoring distributional measures beyond the mean can provide earlier and richer information about market conditions. In a context where the 10th percentile proxies

an entry-level segment and the 60th percentile an active upper-middle tier, it is plausible that macro–financial disturbances are reflected relatively quickly where households are most constrained and where trading is most informative, and that the remainder of the distribution adjusts with a lag of one quarter. For lenders, regulators, and local policymakers, such segment-specific dynamics matter for assessing vulnerabilities and designing macroprudential tools, and they underscore that standard mean-based price indices provide only a partial view of market conditions: while headline indices track the central tendency of prices, the quantile indices reveal how different segments of the distribution can diverge and adjust at different speeds.

More broadly, the main contribution of this study lies in the empirical framework itself. The combination of RIF quantile indices, GIE smoothing, LWL-VAR estimation, generalized impulse responses, and auxiliary VAR-based Granger tests provides a structured way to uncover lead–lag relationships across the price distribution using only transaction data. Applied to Warsaw, the framework highlights a lower-middle (10th percentile) and an upper-middle (60th percentile) segment as particularly influential in the estimated dynamics, but there is no reason to expect these particular quantiles to be universally informative. In other cities or periods, with different credit institutions, supply constraints, or demand shocks, the same procedure may point to different parts of the distribution. The key message is not which specific quantile is "the" leader, but that distributional dynamics can be systematically analyzed with readily available data, revealing segmentation and staggered adjustment that are invisible in aggregate indices alone.

Chapter appendix

4.A1 Smooth Quantile Indices Summary Statistics

Series	Mean	Std. dev.	ACF(1)	ADF stat.	ADF p-value
q: 0.10	157.2	46.9	0.919	0.479	> 0.99
q: 0.15	156.5	46.0	0.917	0.941	> 0.99
q: 0.20	155.5	45.2	0.916	1.335	> 0.99
q: 0.25	154.0	44.4	0.917	1.386	> 0.99
q: 0.30	152.7	43.5	0.915	1.374	> 0.99
q: 0.35	152.0	43.0	0.915	1.168	> 0.99
q: 0.40	151.4	42.5	0.916	0.945	> 0.99
q: 0.45	150.7	42.2	0.916	0.873	> 0.99
q: 0.50	150.5	42.0	0.917	0.846	> 0.99
q: 0.55	149.8	41.4	0.917	0.988	> 0.99
q: 0.60	149.6	41.1	0.917	1.465	> 0.99
q: 0.65	148.9	40.5	0.917	1.359	> 0.99
q: 0.70	148.3	39.9	0.916	1.536	> 0.99
q: 0.75	148.1	39.3	0.916	1.489	> 0.99
q: 0.80	146.2	38.3	0.913	1.451	> 0.99
q: 0.85	144.4	37.1	0.914	1.185	> 0.99
q: 0.90	142.5	35.8	0.909	1.105	> 0.99

Table 4.A1.1: Smoothed Index Levels: Summary Statistics. Means and standard deviations are for the levels of the smoothed RIF-based quantile indices (base = 100 in the first quarter). ACF(1) denotes the first-order autocorrelation. ADF stat. and ADF p-value refer to Augmented Dickey–Fuller tests for a unit root. *Notes:* All series have $N = 72$ quarterly observations; the ADF regressions include 2 lags.

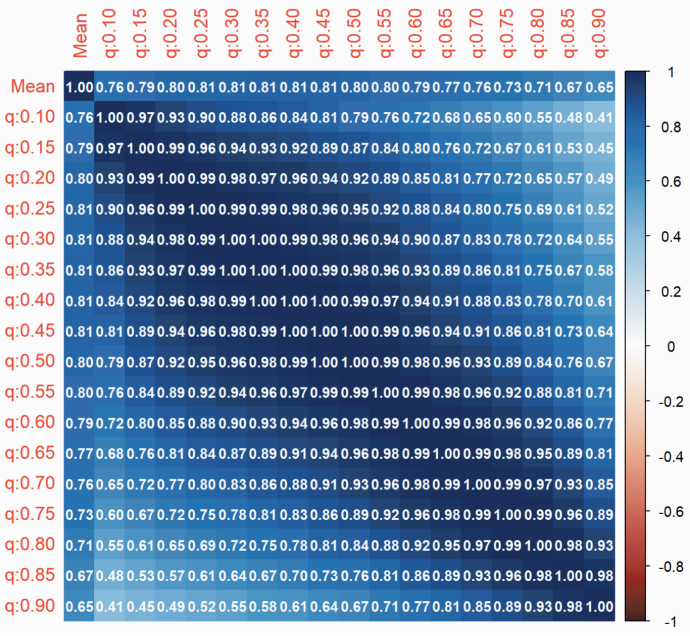


Figure 4.A1.1: Correlation Matrix of the RIF-based Quantile Index Log Return (not smoothed).

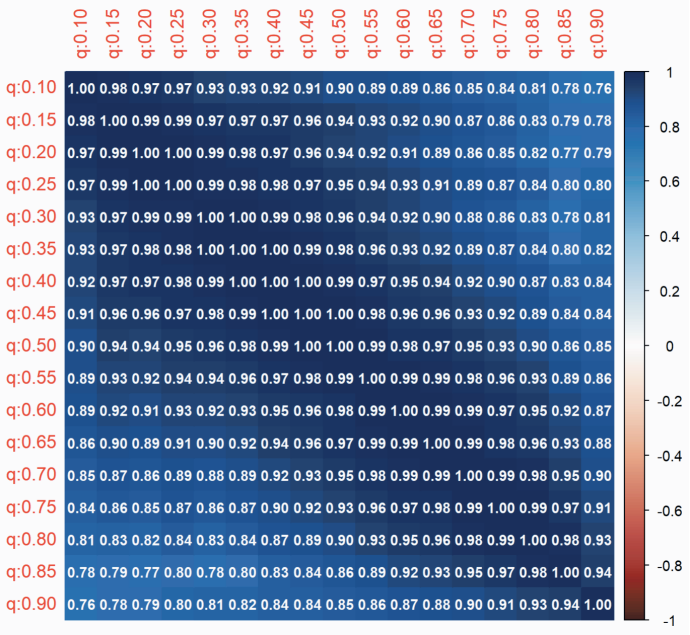


Figure 4.A1.2: Correlation Matrix of the Smoothed RIF-based Quantile Index Log Return.

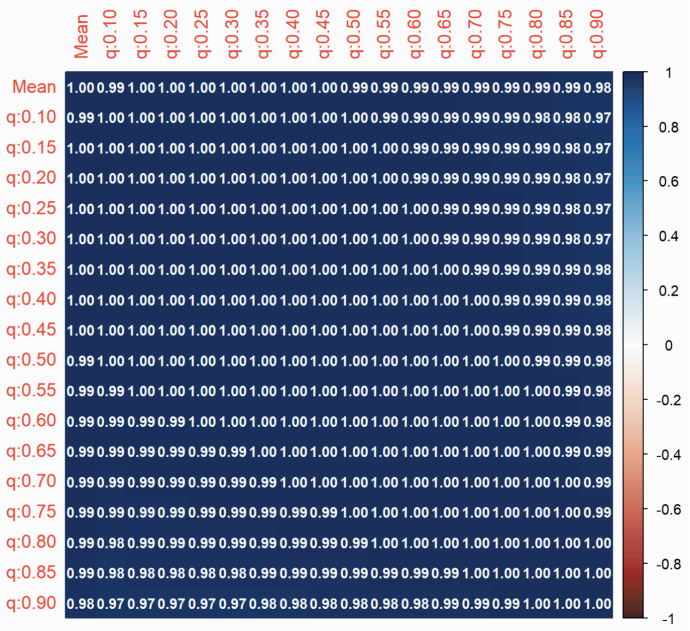


Figure 4.A1.3: Correlation Matrix of the RIF-based Quantile Indices (not smoothed).

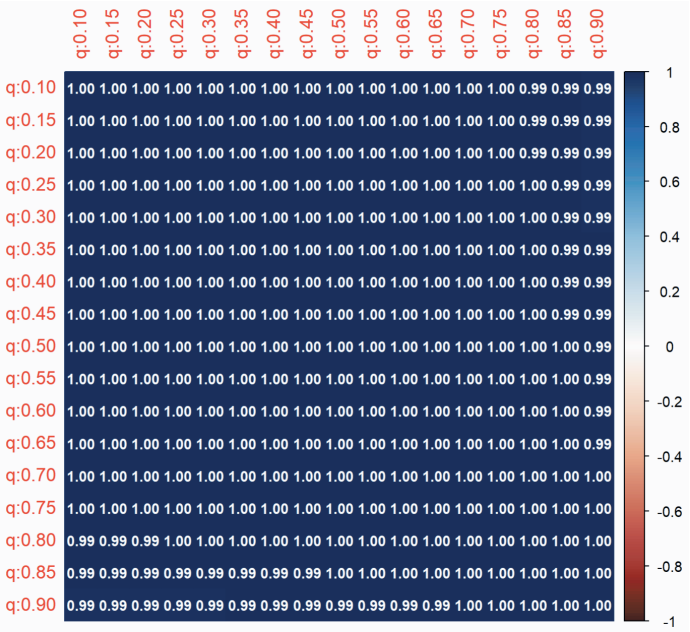


Figure 4.A1.4: Correlation Matrix of the Smoothed RIF-based Quantile Indices.

4.A2 Lag-weighted Lasso VAR

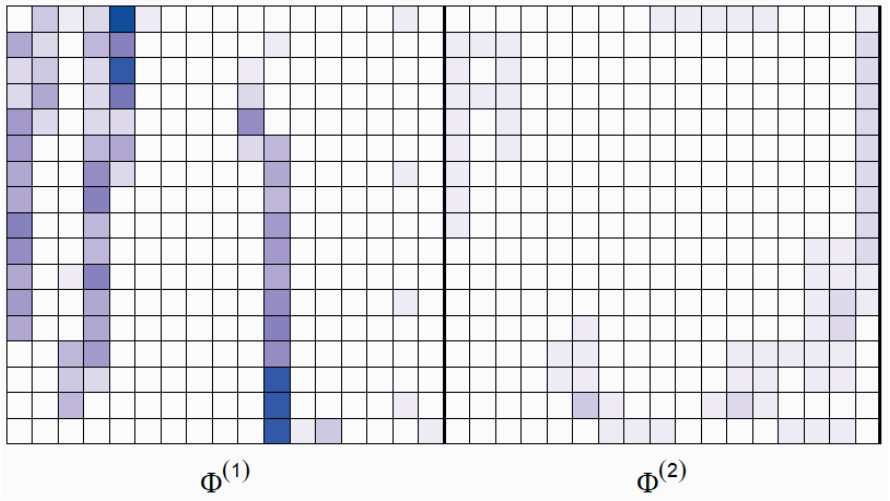


Figure 4.A2.1: Lag-weighted Lasso VAR with Smoothed RIF-based Quantile Index Real Returns.

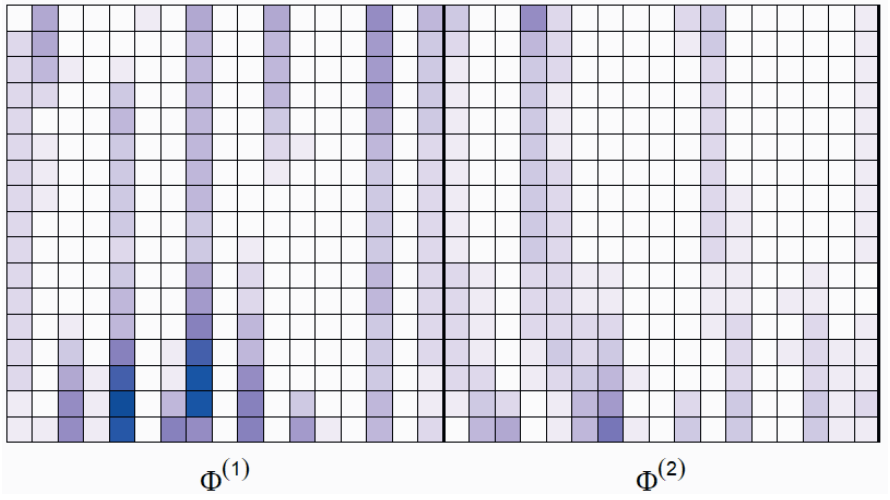


Figure 4.A2.2: Lag-weighted Lasso VAR with RIF-based Quantile Index Returns (no smoothing).

Panel A: $\Phi^{(1)}$ (lag $t-1$)

Response	Intercept	q:0.10	q:0.15	q:0.20	q:0.25	q:0.30	q:0.35	q:0.40	q:0.45	q:0.50	q:0.55	q:0.60	q:0.65	q:0.70	q:0.75	q:0.80	q:0.85	q:0.90
q:0.10	0.0023	0	0.0484	0	0	0.3745	0.2954	0.0128	0	0	0	0.0569	0.0206	0	0	0	0	0
q:0.15	0.0031	0.2389	0	0	0	0.3801	0	0.0090	0	0	0	0.1999	0	0	0	0	0	0
q:0.20	0.0033	0.1155	0.0221	0	0	0.4682	0	0.0696	0	0	0.1146	0.0394	0	0	0	0	0	0
q:0.25	0.0036	0.0646	0.1523	0	0	0.2902	0.0098	0.1089	0	0	0.0255	0.1754	0	0	0	0	0	0
q:0.30	0.0040	0.2586	0.0244	0	0	0.1146	0	0.0047	0	0	0.2987	0.1455	0	0	0	0	0	0
q:0.35	0.0039	0.2426	0	0	0	0.2377	0	0	0	0	0.0120	0.3399	0	0	0	0	0	0
q:0.40	0.0036	0.2505	0	0	0.0027	0.1591	0	0	0	0	0	0.3748	0	0	0	0	0	0
q:0.45	0.0034	0.2715	0	0	0.0787	0.0801	0	0	0	0	0	0.3334	0	0	0	0	0	0
q:0.50	0.0034	0.3659	0	0	0	0	0	0	0	0	0	0.3886	0	0	0	0	0	0
q:0.55	0.0032	0.3259	0	0	0	0	0	0	0	0	0	0.4059	0	0	0	0	0	0
q:0.60	0.0036	0.2506	0	0.0012	0.1151	0	0	0	0	0	0	0.3646	0	0	0	0	0	0
q:0.65	0.0036	0.3029	0	0	0.0046	0	0	0	0	0	0	0.3823	0	0	0	0	0.0513	0
q:0.70	0.0032	0.2380	0	0	0	0	0	0	0	0	0	0.4583	0	0	0	0	0.0121	0
q:0.75	0.0025	0.0575	0	0.1189	0.0355	0	0	0	0	0	0	0.4618	0	0	0	0	0	0
q:0.80	0.0020	0.0067	0	0.0435	0	0	0	0	0	0	0	0.6467	0	0	0	0	0	0.0022
q:0.85	0.0009	0	0	0.0125	0	0	0	0	0	0	0	0.6364	0	0	0	-0.0251	-0.0001	0.0024
q:0.90	0.0004	-0.0261	-0.0251	0	-0.0013	0	0	0	0	0	0	0.3931	0.2704	0	0	0	0	0.0672

Panel B: $\Phi^{(2)}$ (lag $t-2$)

Response	Intercept	q:0.10	q:0.15	q:0.20	q:0.25	q:0.30	q:0.35	q:0.40	q:0.45	q:0.50	q:0.55	q:0.60	q:0.65	q:0.70	q:0.75	q:0.80	q:0.85	q:0.90
q:0.10	0.0023	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
q:0.15	0.0031	-0.0189	-0.0167	-0.0064	0	0	0	0	0	0	0	0	0	0	0	0	0	-0.0010
q:0.20	0.0033	-0.0129	-0.0109	-0.0121	0	-0.0044	-0.0028	0	0	0	0	0	0	0	0	0	0	-0.0158
q:0.25	0.0036	-0.0174	-0.0163	-0.0140	0	-0.0018	-0.0036	0	0	0	0	0	0	0	0	0	0	-0.0197
q:0.30	0.0040	-0.0250	-0.0140	-0.0225	0	-0.0048	-0.0004	0	0	0	0	0	0	0	0	0	0	-0.0313
q:0.35	0.0039	-0.0404	-0.0064	-0.0178	0	0	0	0	0	0	0	0	0	0	0	0	0	-0.0284
q:0.40	0.0036	-0.0433	0	-0.0003	0	0	0	0	0	0	0	0	0	0	0	0	0	-0.0187
q:0.45	0.0034	-0.0395	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-0.0050
q:0.50	0.0034	-0.0366	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-0.0061
q:0.55	0.0032	-0.0015	0	0	0	0	0	0	0	0	0	0	0	0	0	-0.0021	-0.0302	-0.0019
q:0.60	0.0036	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-0.0587	0
q:0.65	0.0036	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-0.0097	-0.0527	0
q:0.70	0.0032	0	0	0	0	0.0033	0.0097	0	0	0	0	0	0	0	0	-0.0073	-0.0432	0
q:0.75	0.0025	0	0	0	0	0.0199	0.0208	0	0	0	0	0	-0.0081	-0.0115	-0.0064	0	-0.0128	0
q:0.80	0.0020	0	0	0	0.0004	0.0265	0.0287	0	0	0	0	0	-0.0289	-0.0184	0	0	0	0
q:0.85	0.0009	0	0	0	0	0.0325	0.1073	0.0007	0	0	0	-0.0248	-0.0461	-0.0228	0	0	0	0.0137
q:0.90	0.0004	0	0	0	0	0	0	0	0.0107	0.0013	0	0	0	0	0.0104	0.0185	0.0177	0

Table 4.A2.1: Lag-weighted Lasso VAR (order $p = 2$) coefficient estimates for smoothed RIF-based quantile-return indices. The model is $\mathbf{y}_t = \mathbf{c} + \Phi^{(1)}\mathbf{y}_{t-1} + \Phi^{(2)}\mathbf{y}_{t-2} + \mathbf{u}_t$. Panel A reports $\Phi^{(1)}$ (lag $t-1$) and Panel B reports $\Phi^{(2)}$ (lag $t-2$); the intercept is repeated in both panels for reference.

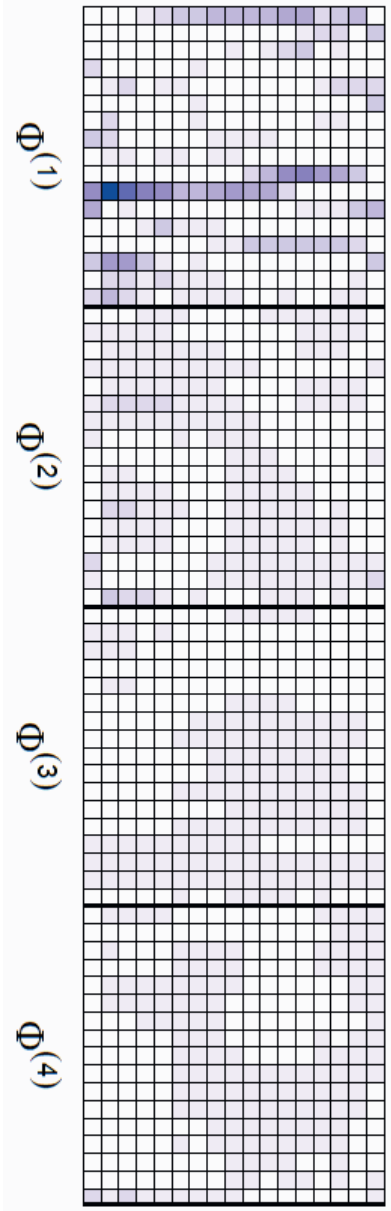


Figure 4.A2.3: Lag-weighted Lasso VAR with Smoothed RIF-based Quantile Index Returns.

		$\Phi^{(1)}$ (lag $t-1$)															
Response	Intercept	q:0.10	q:0.15	q:0.20	q:0.25	q:0.30	q:0.35	q:0.40	q:0.45	q:0.50	q:0.55	q:0.60	q:0.65	q:0.70	q:0.75	q:0.80	q:0.85	q:0.90
q:0.10	0.0021	0	0.0098	0	0	0.4709	0.2999	0	0	0	0	0	0.0593	0	0	0	0	0
q:0.15	0.0022	0.1841	0	0	0	0.4650	0	0	0	0	0	0.1679	0	0	0	0	0	0
q:0.20	0.0023	0.0670	0	0	0	0.5425	0	0.0523	0	0	0.1331	0	0	0	0	0	0	0
q:0.25	0.0024	0.0399	0.0557	0	0	0.4592	0	0.0571	0	0	0.0231	0.1472	0	0	0	0	0	0
q:0.30	0.0025	0.2112	0	0	0	0.1512	0	0	0	0	0.3902	0.0121	0	0	0	0	0	0
q:0.35	0.0025	0.1936	0	0	0	0.2643	0	0	0	0	0.0072	0.2904	0	0	0	0	0	0
q:0.40	0.0025	0.1963	0	0	0	0.2284	0	0	0	0	0	0.3276	0	0	0	0	0	0
q:0.45	0.0026	0.2035	0	0	0.1798	0.0638	0	0	0	0	0	0.2906	0	0	0	0	0	0
q:0.50	0.0026	0.3353	0	0	0.0286	0	0	0	0	0	0	0.3661	0	0	0	0	0	0
q:0.55	0.0026	0.2768	0	0	0.0714	0	0	0	0	0	0	0.3820	0	0	0	0	0	0
q:0.60	0.0025	0.0953	0	0.1012	0.2373	0	0	0	0	0	0	0.2808	0	0	0	0	0.0218	0.0010
q:0.65	0.0025	0.1891	0	0.0120	0.1628	0	0	0	0	0	0	0.2972	0	0	0	0	0.0715	0.0015
q:0.70	0.0024	0.1239	0	0.0205	0.1666	0	0	0	0	0	0	0.3792	0	0	0	0	0.0198	0.0262
q:0.75	0.0022	0	0	0.2850	0	0	0	0	0	0	0	0.3954	0	0	0	0	0	0.0602
q:0.80	0.0017	0	0	0.1437	0	0	0	0	0	0	0	0.5629	0	0	0	0	0	0.0744
q:0.85	0.0012	0	0	0.2096	0	0	0	0	0	0	0	0.4666	0	0	0	0	0	0.1165
q:0.90	0.0010	0	0	0	0	0	0	0	0	0	0	0.2366	0.3783	0	0	0	0	0.1783

Table 4.A2.2: Lag-weighted Lasso VAR (order $p = 1$) coefficient estimates for smoothed RIF-based quantile-return indices. The model is $\mathbf{y}_t = \mathbf{c} + \Phi^{(1)}\mathbf{y}_{t-1} + \mathbf{u}_t$. Rows denote response quantiles at time t; columns denote lagged quantile returns. Entries equal to 0 indicate coefficients shrunk to (approximately) zero by the lag-weighted Lasso.

4.A3 Generalized Impulse Response Function

Consider the estimated LWL-VAR in Equation (4.7), written in terms of the reduced-form innovations ε^t, with covariance matrix $\Sigma = \mathbb{E}[\varepsilon_t \varepsilon_t']$. Under standard assumptions, this VAR admits a moving-average representation

$$R^t = \sum_{h=0}^{\infty} \Psi_h \varepsilon^{t-h},$$

where $\Psi_0 = I_K$ and Ψ_h are $K \times K$ matrices of dynamic multipliers. Following Pesaran and Shin (1998), the generalized impulse response at horizon h of the system to a one-standard-deviation shock to component j is defined as

$$\text{GIRF}_j(h) = \mathbb{E}\big[R^{t+h} \,\big|\, \varepsilon_j^t = \sigma_j, \mathcal{I}_{t-1}\big] - \mathbb{E}\big[R^{t+h} \,\big|\, \mathcal{I}_{t-1}\big],$$

where $\sigma_j = \sqrt{\Sigma_{jj}}$ is the standard deviation of the j-th innovation and $\mathcal{I}^{t-1}$ denotes the information set up to time $t-1$. In a linear Gaussian VAR this expression simplifies to

$$\text{GIRF}_j(h) = \Psi_h \Sigma e_j / \sigma_j.$$

where e_j is the j-th canonical basis vector. Each column of the $K \times (H+1)$ matrix $\big[\text{GIRF}_j(0), \ldots, \text{GIRF}_j(H)\big]$ thus traces the expected path of all quantile returns following a standardized shock to quantile j, for horizons $h = 0, \ldots, H$.

Starting from the trained model object and the original matrix of quantile returns Y (with T rows and $K = 17$ columns), This work first reconstruct the lagged design in companion form. The estimated coefficient matrix is partitioned into an intercept and lag coefficient blocks $(A_1, \ldots, A_p)$, and stack these into the companion matrix A of dimension $Kp \times Kp$. A selection matrix J of dimension $K \times Kp$ is used to extract the first K states from the companion system. The reduced-form residuals from the fitted LWL-VAR are collected and used to estimate the innovation covariance matrix Σ. Dynamic multipliers are then obtained as

$$\Psi_h = J A^h J', \quad h = 0, \ldots, H,$$

where H is the maximum horizon of interest (in our application, $H = 12$ quarters). For each shock variable j and each horizon h, we compute $\text{GIRF}_j(h)$ using the Pesaran–Shin formula above, normalizing the shock by its own standard deviation σ_j so that all responses are expressed in terms of a one-standard-deviation innovation.

To quantify uncertainty around these impulse responses, we construct bootstrap confi-

dence bands using a residual-based moving-block bootstrap that preserves temporal dependence. The fitted residuals $\{\hat{\varepsilon}_t\}$ from the original LWL-VAR are segmented into overlapping blocks of fixed length. For each bootstrap replication, blocks are resampled with replacement in circular fashion to generate a synthetic sequence of residuals. Using the estimated LWL-VAR parameters as the data-generating mechanism, we recursively generate a synthetic series of returns R_t^* by feeding the bootstrap residuals through the VAR. On each bootstrap sample, we re-estimate a LWL-VAR with the same penalization structure and recompute the full set of generalized impulse responses. Repeating this procedure B times yields an empirical distribution of GIRFs for each response quantile, horizon, and shocked quantile.

From the collection of bootstrap GIRFs we extract point wise percentile bands. For each response quantile τ_k and horizon h, and for a given shocked quantile j, we take the $\alpha/2$ and $(1-\alpha/2)$ empirical quantiles of the B bootstrap values of $\mathrm{GIRF}_{k,j}(h)$ to obtain, for example, 95% confidence intervals. These bands are then plotted together with the point GIRF obtained from the original LWL-VAR estimates. In practice, we repeat this procedure for shocks to a small set of economically relevant quantiles (in our case, $\tau = 0.10, 0.45, 0.60$, and 0.75), which allows us to examine whether shocks originating in the lower, middle, or upper parts of the distribution generate responses consistent with the leading–lagging structure suggested by the LWL-VAR coefficients.

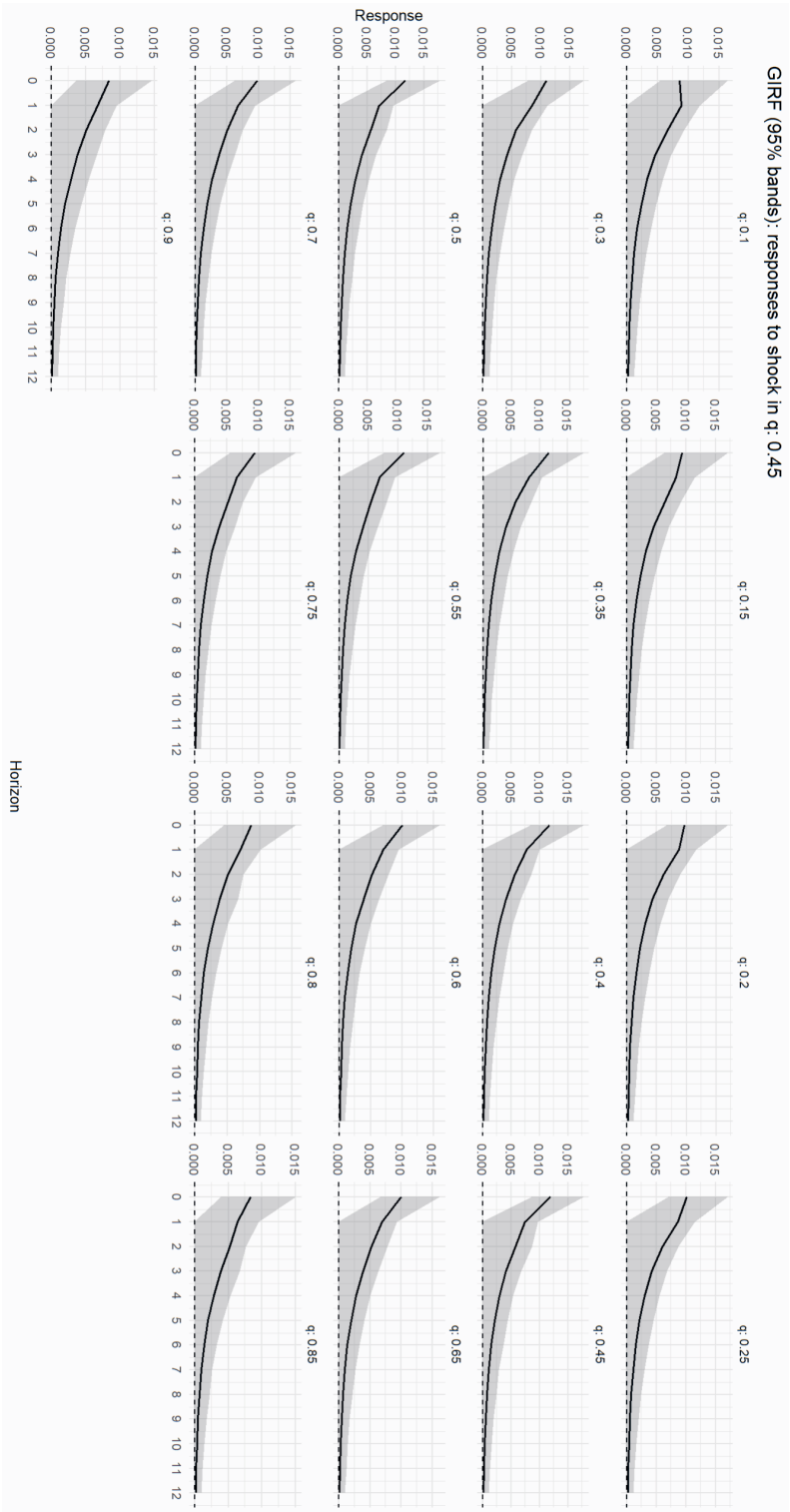


Figure 4.A3.1: Generalized Impulse Response Function (GRIF). One Standard Deviation shock in the 45th Quantile.

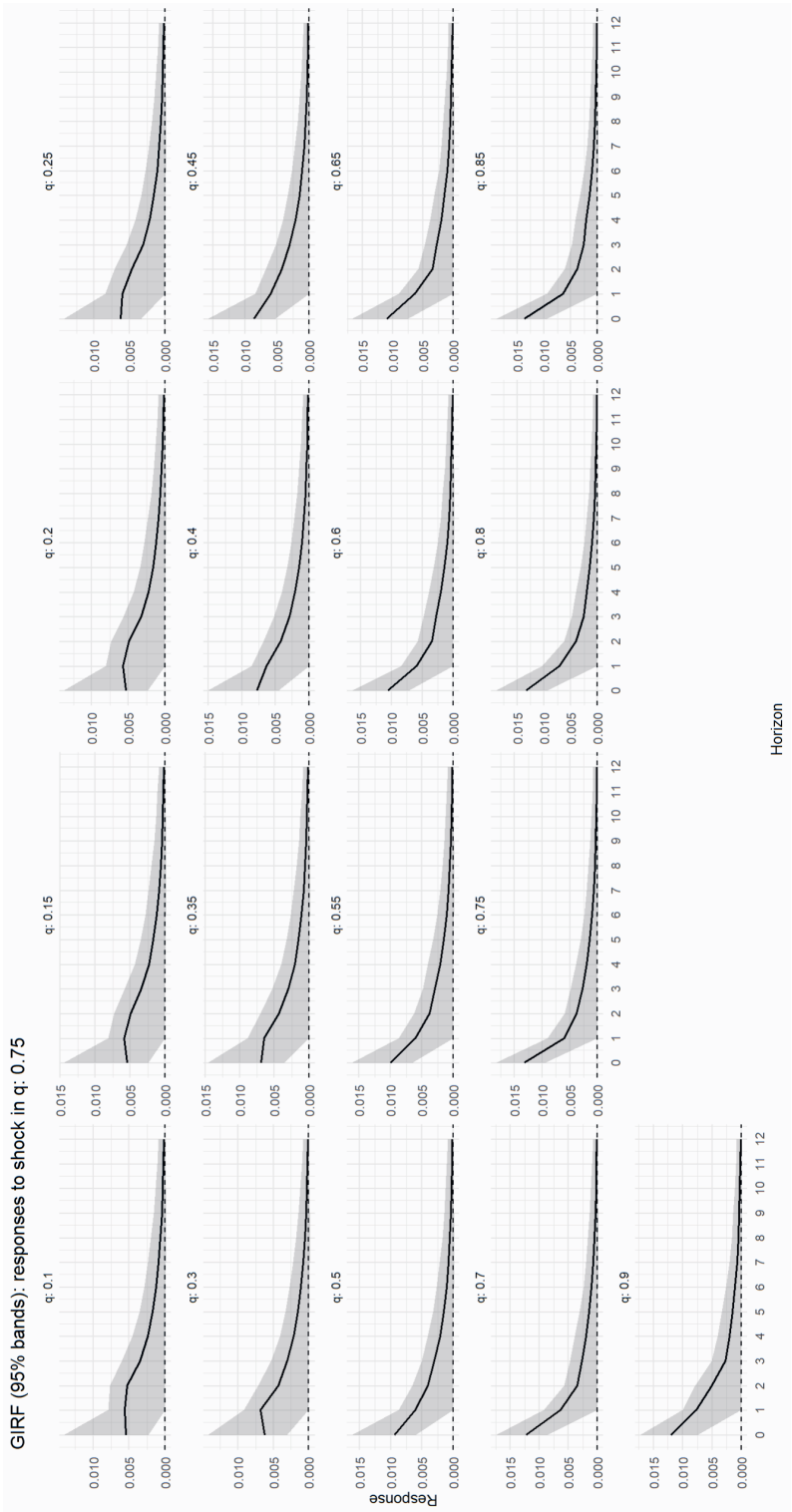


Figure 4.A3.2: Generalized Impulse Response Function (GRIF). One Standard Deviation shock in the 75th Quantile.

4.A4 Granger Causality and Reduced VAR Additional Results

Regressor	Coefficient	Std. error	t-stat	p-value	Sig.
$q_{0.10,t-1}$	0.1701	0.1987	0.856	0.3960	
$q_{0.45,t-1}$	0.6432	0.4352	1.478	0.1450	
$q_{0.60,t-1}$	0.6550	0.6126	1.069	0.2890	
$q_{0.75,t-1}$	-0.7403	0.8701	-0.851	0.3980	
$q_{0.80,t-1}$	0.1746	0.7166	0.244	0.8080	
$q_{0.10,t-2}$	-0.1083	0.1675	-0.646	0.5210	
$q_{0.45,t-2}$	0.2490	0.3916	0.636	0.5270	
$q_{0.60,t-2}$	-0.4624	0.6474	-0.714	0.4780	
$q_{0.75,t-2}$	-0.2066	0.9348	-0.221	0.8260	
$q_{0.80,t-2}$	0.4266	0.6710	0.636	0.5270	
Constant	0.0025	0.0019	1.302	0.1980	

Table 4.A4.1: VAR(2) estimates for equation in $q_{0.10}$ (smoothed returns).

Residual s.e. = 0.0131 (58 df); $R^2 = 0.7014$, adjusted $R^2 = 0.6499$; overall $F(10, 58) = 13.62$, $p = 6.29 \times 10^{-12}$.
Sig. codes: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.10$.

Regressor	Coefficient	Std. error	t-stat	p-value	Sig.
$q_{0.10,t-1}$	0.2760	0.1773	1.557	0.1249	
$q_{0.45,t-1}$	-0.1718	0.3883	-0.442	0.6599	
$q_{0.60,t-1}$	1.7676	0.5466	3.234	0.0020	**
$q_{0.75,t-1}$	-2.0084	0.7764	-2.587	0.0122	*
$q_{0.80,t-1}$	1.0102	0.6394	1.580	0.1196	
$q_{0.10,t-2}$	-0.3138	0.1495	-2.099	0.0401	*
$q_{0.45,t-2}$	0.3647	0.3494	1.044	0.3010	
$q_{0.60,t-2}$	-0.0613	0.5776	-0.106	0.9158	
$q_{0.75,t-2}$	-0.2462	0.8340	-0.295	0.7689	
$q_{0.80,t-2}$	0.0961	0.5987	0.160	0.8731	
Constant	0.0029	0.0017	1.707	0.0933	·

Table 4.A4.2: VAR(2) estimates for equation in $q_{0.45}$ (smoothed returns).

Residual s.e. = 0.0117 (58 df); $R^2 = 0.7214$, adjusted $R^2 = 0.6733$; overall $F(10, 58) = 15.02$, $p = 9.4 \times 10^{-13}$.
Sig. codes: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.10$.

Regressor	Coefficient	Std. error	t-stat	p-value	Sig.
$q_{0.10,t-1}$	0.1302	0.1781	0.731	0.4679	
$q_{0.45,t-1}$	0.2213	0.3902	0.567	0.5727	
$q_{0.60,t-1}$	0.8780	0.5492	1.599	0.1153	
$q_{0.75,t-1}$	-1.0958	0.7801	-1.405	0.1655	
$q_{0.80,t-1}$	0.7010	0.6425	1.091	0.2797	
$q_{0.10,t-2}$	-0.1942	0.1502	-1.293	0.2011	
$q_{0.45,t-2}$	0.7986	0.3511	2.275	0.0266	*
$q_{0.60,t-2}$	-0.7165	0.5804	-1.235	0.2220	
$q_{0.75,t-2}$	-0.1112	0.8381	-0.133	0.8949	
$q_{0.80,t-2}$	0.0685	0.6016	0.114	0.9098	
Constant	0.0032	0.0017	1.867	0.0669	·

Table 4.A4.3: VAR(2) estimates for equation in $q_{0.60}$ (smoothed returns).

Residual s.e. = 0.0117 (58 df); $R^2 = 0.6918$, adjusted $R^2 = 0.6387$; overall $F(10, 58) = 13.02$, $p = 1.49 \times 10^{-11}$.
Sig. codes: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.10$.

Regressor	Coefficient	Std. error	t-stat	p-value	Sig.
$q_{0.10,t-1}$	-0.0649	0.1957	-0.332	0.7412	
$q_{0.45,t-1}$	0.0663	0.4286	0.155	0.8777	
$q_{0.60,t-1}$	1.3224	0.6032	2.192	0.0324	*
$q_{0.75,t-1}$	-1.4310	0.8568	-1.670	0.1003	
$q_{0.80,t-1}$	0.8955	0.7057	1.269	0.2095	
$q_{0.10,t-2}$	-0.1865	0.1650	-1.130	0.2630	
$q_{0.45,t-2}$	1.1555	0.3856	2.997	0.0040	**
$q_{0.60,t-2}$	-0.9544	0.6375	-1.497	0.1398	
$q_{0.75,t-2}$	-1.1206	0.9205	-1.217	0.2284	
$q_{0.80,t-2}$	0.9938	0.6607	1.504	0.1380	
Constant	0.0031	0.0019	1.664	0.1015	

Table 4.A4.4: VAR(2) estimates for equation in $q_{0.75}$ (smoothed returns).

Residual s.e. = 0.0129 (58 df); $R^2 = 0.6532$, adjusted $R^2 = 0.5935$; overall $F(10, 58) = 10.93$, $p = 3.67 \times 10^{-10}$.
Sig. codes: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.10$.

Regressor	Coefficient	Std. error	t-stat	p-value	Sig.
$q_{0.10,t-1}$	-0.0918	0.2082	-0.441	0.6610	
$q_{0.45,t-1}$	-0.1728	0.4560	-0.379	0.7061	
$q_{0.60,t-1}$	1.5076	0.6418	2.349	0.0223	*
$q_{0.75,t-1}$	-0.8744	0.9117	-0.959	0.3415	
$q_{0.80,t-1}$	0.4304	0.7508	0.573	0.5687	
$q_{0.10,t-2}$	-0.1330	0.1755	-0.758	0.4516	
$q_{0.45,t-2}$	0.9231	0.4103	2.250	0.0283	*
$q_{0.60,t-2}$	-0.8118	0.6783	-1.197	0.2362	
$q_{0.75,t-2}$	-1.0029	0.9794	-1.024	0.3101	
$q_{0.80,t-2}$	0.9390	0.7030	1.336	0.1869	
Constant	0.0025	0.0020	1.238	0.2208	

Table 4.A4.5: VAR(2) estimates for equation in $q_{0.80}$ (smoothed returns).

Residual s.e. = 0.0137 (58 df); $R^2 = 0.6370$, adjusted $R^2 = 0.5745$; overall $F(10, 58) = 10.18$, $p = 1.25 \times 10^{-9}$.
Sig. codes: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.10$.

	$q_{0.10}$	$q_{0.45}$	$q_{0.60}$	$q_{0.75}$	$q_{0.80}$
$q_{0.10}$	0.0001715	0.0001130	0914	0683	0610
$q_{0.45}$	0.0001130	0.0001365	0.0001250	0.0001131	0.0001061
$q_{0.60}$	0914	0.0001250	0.0001379	0.0001395	0.0001394
$q_{0.75}$	0683	0.0001131	0.0001395	0.0001663	0.0001737
$q_{0.80}$	0610	0.0001061	0.0001394	0.0001737	0.0001883

Table 4.A4.6: Residual covariance matrix of the VAR(2) in five quantile returns.

	$q_{0.10}$	$q_{0.45}$	$q_{0.60}$	$q_{0.75}$	$q_{0.80}$
$q_{0.10}$	1.0000	0.7387	0.5945	0.4041	0.3392
$q_{0.45}$	0.7387	1.0000	0.9112	0.7507	0.6617
$q_{0.60}$	0.5945	0.9112	1.0000	0.9213	0.8652
$q_{0.75}$	0.4041	0.7507	0.9213	1.0000	0.9818
$q_{0.80}$	0.3392	0.6617	0.8652	0.9818	1.0000

Table 4.A4.7: Residual correlation matrix of the VAR(2) in five quantile returns.

"Tenha até pesadelos, se necessário for. Mas sonhe."

Patrícia Galvão

5

Conclusion

This thesis examines how automated valuation models (AVMs) and modern machine learning (ML) methods can be used to strengthen real estate valuation and the market measures derived from it. The motivation is grounded in the role that valuation plays in real estate finance and economics. Valuation outputs enter a wide range of institutional settings, including mortgage lending, portfolio monitoring, investment performance assessment, and public-sector appraisal and taxation. As richer data and more flexible predictive methods become available, ML increasingly complements established econometric approaches such as OLS and GLMs. In this setting, the central question is not whether flexible models can be estimated, but whether they can be translated into valuation outputs that remain comparable over time, informative about uncertainty, and useful for understanding market heterogeneity.

The thesis addresses this agenda through three connected questions. It asks how ML-based valuation models can be used (i) to construct real estate price indices in a principled, model-agnostic way, (ii) to quantify price prediction uncertainty by studying how predictive models interact with prediction-interval methodologies and identifying combinations that achieve reliable coverage without becoming uninformatively wide, and (iii) to assess whether housing price quantiles move together over time or whether certain segments systematically lead others, thereby offering a distributional view of market dynamics.

5.1 Synthesis of key findings

Chapter 2 focuses on market-level measurement through price indices. It proposes a model-agnostic framework in which indices can be constructed using out-of-time transaction predictions produced by alternative predictive models. Empirically, flexible ML methods improves transaction-level predictive performance relative to linear benchmarks in the commercial real estate application, and the framework shows how such models can be embedded in an index construction pipeline without redefining the index methodology for each algorithm. The chapter also documents an important practical regularity: gains from flexibility depend on data conditions. When samples are restricted, simpler models can exhibit greater stability, and design choices such as tuning and imputation can influence index behaviour. These findings reinforce a familiar insight from hedonic and regression-based index construction: quality adjustment and comparability require disciplined design, regardless of the predictive model used (Bailey et al., 1963; Hill, 2011; Rosen, 1974).

Chapter 3 addresses property-level valuation by studying how to attach statistically meaningful uncertainty to AVM point predictions. The empirical results shows that prediction-interval performance is jointly determined by the interval methodology and the regression engine, but in different ways. Across the evaluated combinations, the interval method primarily governs empirical coverage, while the regression model influences interval width. This separation clarifies why uncertainty quantification should be evaluated on two criteria simultaneously: validity, in the sense of achieving the intended coverage rate, and efficiency, in the sense of remaining sufficiently narrow to be informative. The chapter further identifies combinations that balance these criteria in a rolling, out-of-sample setting, consistent with the view that conformal-style approaches provide a practical route to distribution-free uncertainty quantification in prediction problems (Shafer and Vovk, 2008; Vovk et al., 2005).

Chapter 4 shifts from cross-sectional valuation to distributional time-series dynamics. It constructs quantile price indices and models their joint evolution to examine whether movements in the housing price distribution are synchronized or segmented. The results supports a distributional interpretation of market dynamics. Quantile indices displays persistent dispersion, and lead–lag analysis suggests that predictive relationships concentrate in particular segments of the distribution rather than being evenly shared across quantiles. The main contribution is therefore a segmentation result, operationalized through a replicable pipeline that links distributional measurement to dynamic dependence, building on unconditional quantile methods and decomposition tools used in

applied microeconometrics (Firpo et al., 2009; Machado and Mata, 2005).

Taken together, the three chapters provide evidence that ML methods can be used to improve valuation practice. Price indices, prediction intervals, and distributional dynamics represent distinct, but complementary, ways of converting transaction data into market information.

5.2 Contributions

The thesis makes three contributions.

First, it contributes to the price index literature by providing and testing a model-agnostic way to connect modern predictive models to index construction using an out-of-time prediction design. This extends the practical scope of regression-based and hedonic index ideas by allowing the prediction step to be chosen flexibly while keeping the index logic interpretable and comparable (Bailey et al., 1963; Hill, 2011; Rosen, 1974).

Second, it contributes to the applied AVM literature by clarifying, through a systematic comparison, how prediction-interval methodology and regression model choice interact in practice. The results show that interval validity and interval efficiency respond to different components of the modelling pipeline, which provides an actionable way to design and evaluate uncertainty quantification in AVMs. This contribution aligns with, and offers real-estate-specific evidence for, the broader conformal prediction literature that motivates model-agnostic uncertainty quantification under weak distributional assumptions (Shafer and Vovk, 2008; Vovk et al., 2005).

Third, it contributes to work on housing market heterogeneity by developing a framework that treats quantile indices as a multivariate time series and uses sparse dynamic modelling to identify lead–lag relationships across the distribution. This provides an empirical route to studying segmentation as a dynamic phenomenon, complementing established distributional methods used to analyze heterogeneous effects and counterfactual distributions (Firpo et al., 2009; Machado and Mata, 2005).

A unifying interpretation is that AVMs produce measurement objects that are subsequently compared across time and space. The scientific and practical value of these objects depends on whether they are empirically defensible under heterogeneity, thin markets, and imperfect data.

5.3 Implications for research and practice

The thesis has two broad implications.

The first concerns evaluation standards. In many applied settings, valuation systems are assessed primarily through point prediction accuracy. The results in this thesis indicate that such an evaluation is incomplete when the valuation output is used to monitor, benchmark or make risk-sensitive decisions. Index construction requires attention to stability and comparability over time. Uncertainty quantification requires explicit assessment of empirical coverage alongside interval width. Distributional monitoring requires tools that can detect heterogeneous dynamics that average measures can mask. A common implication is that valuation systems should be evaluated against the specific measurement task they are intended to support, rather than against a single generic accuracy metric.

The second concerns the interpretation of market information produced by predictive models. The chapters collectively show that richer modelling capacity does not eliminate long-standing measurement challenges in real estate. Instead, it changes where those challenges appear. Flexible models can improve fit, but they also make design choices, sample restrictions, and distributional assumptions more consequential for the resulting market indicators. For research, this points to the importance of methodological work that links predictive modelling to interpretable measurement objects. For practice, it suggests that the adoption of ML is most defensible when accompanied by transparent validation strategies and monitoring protocols aligned with the output of interest, whether an index, an interval, or a distributional diagnostic.

5.4 Limitations

Several limitations qualify the interpretation of the results.

First, each chapter relies on a specific market and data environment. The index application uses a commercial real estate transaction sample in New York; the uncertainty analysis uses residential transactions in Amsterdam; and the distributional dynamics study uses apartment transactions in Warsaw. Transferability of empirical magnitudes therefore requires care, particularly across markets with different liquidity, institutional arrangements, and measurement conventions.

Second, the index framework is evaluated using an out-of-time design that aligns with the intended use of indices, but index behaviour can remain sensitive to sample selection, feature availability, and modelling choices, especially in thin segments. This reflects a gen-

eral limitation of index construction in heterogeneous markets and motivates additional robustness work when indices are produced operationally.

Third, Chapter 3 focuses on prediction intervals rather than full predictive distributions. While intervals capture an important aspect of uncertainty, some applications require probability statements at specific thresholds or tail-focused risk measures. This boundary of scope motivates one of the future research directions below.

Fourth, in Chapter 4 the results provide evidence of systematic dependence across quantile indices, but they do not identify structural mechanisms driving the observed leading patterns.

5.5 Directions for future research

Three extensions follow directly from the thesis.

First, index construction can be strengthened by incorporating richer information on quality. Text and images offer a promising way to capture property characteristics that are difficult to encode in structured variables. In index settings, such information is valuable primarily because it can improve quality adjustment and comparability over time, not simply because it can raise point prediction accuracy. A related research direction is to evaluate how these modalities affect index stability when market composition changes, and whether they reduce sensitivity to thin-market segments.

Second, uncertainty quantification can be extended from intervals to full predictive distributions. Recent developments in conformal methodology allow the construction of prediction distributions for individual observations, making it possible to answer probability queries that align naturally with risk management questions, rather than restricting attention to interval bounds. This would also facilitate the evaluation of tail behaviour and density-based diagnostics within AVMs.

Third, the lead–lag evidence in Chapter 4 motivates work on mechanisms. The quantiles' dynamic results document asynchronous movements across segments of the price distribution, but they do not identify why particular segments appear to lead others. Future research could connect distributional lead–lag patterns to candidate drivers such as shifts in credit conditions, changes in buyer composition, differential supply responses across market segments, or policy and institutional frictions that affect submarkets unevenly. Empirically, this could be pursued by augmenting the quantile VAR framework with macro-financial indicators, credit and lending variables, construction and permit measures, or local labour market conditions, and by testing whether lead–lag relationships strengthen or weaken around identifiable regime changes. A complementary

approach is to move beyond reduced-form dependence by estimating structural or semi-structural models that allow the propagation of shocks to differ across segments, thereby providing an explanation for the observed asynchronous adjustment in the distribution.

More broadly, these directions connect to the longer-run agenda of building coherent digital representations of the built environment. As more heterogeneous data sources become available, the research frontier will increasingly involve turning those sources into measurement objects that remain comparable, interpretable, and useful for economic analysis and decision-making. The results in this thesis represent incremental steps toward that agenda by focusing on three foundational outputs: indices, uncertainty statements, and distributional dynamics.

5.6 Closing remarks

This thesis argues that the most consequential role of ML in real estate valuation is not simply to improve prediction accuracy, but to strengthen the informational content of valuation outputs that are used for benchmarking, risk assessment, and market interpretation. By developing model-agnostic approaches to price indexing, by identifying how uncertainty methods and predictive models interact in practice, and by providing a distributional perspective on housing market dynamics, the thesis contributes tools and evidence that help bring greater clarity to real estate measurement. As data and algorithms continue to evolve, progress in real estate analytics will ultimately be judged by whether it produces market information that is credible, interpretable, and useful for decisions. The results presented here provide concrete steps in that direction.

Bibliography

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185.

Ambrose, B. W., Eichholtz, P., and Lindenthal, T. (2013). House prices and fundamentals: 355 years of evidence. *Journal of Money, Credit and Banking*, 45(2-3):477–491.

Arnott, R. (1989). Housing vacancies, thin markets, and idiosyncratic tastes. *Journal of Real Estate Finance and Economics*, 2(1):5–30.

Bailey, M. J., Muth, R. F., and Nourse, H. O. (1963). A regression method for real estate price index construction. *Journal of the American Statistical Association*, 58(304):933–942.

Balk, B., Diewert, W. E., Fenwick, D., Prud'homme, M., and de Haan, J. (2013). Handbook on residential property prices indices (RPPIs). Technical report, Eurostat.

Barber, B. M., Huang, X., Odean, T., and Schwarz, C. (2022). Attention-induced trading and returns: Evidence from Robinhood users. *The Journal of Finance*, 77(6):3141–3190.

Bastos, J. A. and Paquette, J. (2024). On the uncertainty of real estate price predictions. *Journal of Property Research*, pages 1–19.

Bellotti, A. (2017). Reliable region predictions for automated valuation models. *Annals of Mathematics and Artificial Intelligence*, 81(1):71–84.

Bergstra, J. and Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(Feb):281–305.

Bhutta, N. and Ringo, D. (2021). The effect of interest rates on home buying: Evidence from a shock to mortgage insurance premiums. *Journal of Monetary Economics*, 118:195–211.

Blinder, A. S. (1973). Wage discrimination: reduced form and structural estimates. *Journal of Human resources*, pages 436–455.

Bokhari, S. and Geltner, D. (2012). Estimating real estate price movements for high frequency tradable indexes in a scarce data environment. *The Journal of Real Estate Finance and Economics*, 45(2):522–543.

Bourassa, S. C., Cantoni, E., and Hoesli, M. (2007). Spatial dependence, housing submarkets, and house price prediction. *The Journal of Real Estate Finance and Economics*, 35:143–160.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Breiman, L. and Spector, P. (1992). Submodel selection and evaluation in regression. the x-random case. *International Statistical Review/Revue Internationale de Statistique*, pages 291–319.

Calainho, F. D., van de Minne, A. M., and Francke, M. K. (2022). A machine learning approach to price indices: Applications in commercial real estate. *The Journal of Real Estate Finance and Economics*, pages 1–30.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM.

Clapp, J. M., Eichholtz, P., and Lindenthal, T. (2013). Real option value over a housing market cycle. *Regional Science and Urban Economics*, 43(6):862–874.

Court, A. (1939). *Hedonic price indexes with automotive examples*.

Cvijanović, D. and Spaenjers, C. (2021). "we'll always have paris": Out-of-country buyers in the housing market. *Management Science*, 67(7):4120–4138.

De Wit, E. R., Englund, P., and Francke, M. K. (2013). Price and transaction volume in the dutch housing market. *Regional Science and Urban Economics*, 43(2):220–241.

Deng, Y. and Quigley, J. M. (2008). Index revision, house price risk, and the market for house price derivatives. *The Journal of Real Estate Finance and Economics*, 37(3):191–209.

Do, A. Q. and Grudnitski, G. (1992). A neural network approach to residential property appraisal. *The Real Estate Appraiser*, 58(3):38–45.

Drucker, H., Burges, C. J., Kaufman, L., Smola, A. J., and Vapnik, V. (1997). Support vector regression machines. In *Advances in neural information processing systems*, pages 155–161.

Eichholtz, P. and Geltner, D. (2002). *Four centuries of location value: implications for real estate capital gain in central places*. MIT Center for Real Estate.

Englund, P. and Ioannides, Y. M. (1997). House price dynamics: an international empirical perspective. *Journal of Housing Economics*, 6(2):119–136.

Evans, A., James, H., and Collins, A. (1992). *Artificial Neural Networks: An Application to Residential Valuation in the UK*. University of Portsmouth, Department of Economics.

Firpo, S., Fortin, N. M., and Lemieux, T. (2009). Unconditional quantile regressions. *Econometrica*, 77(3):953–973.

Fitzgerald, M., Hansen, D. J., McIntosh, W., and Slade, B. A. (2020). Urban land: price indices, performance, and leading indicators. *The Journal of Real Estate Finance and Economics*, 60(3):396–419.

Francke, M., Kadhum, A., and Kroon, D. (2024). Creating model-agnostic prediction intervals. *Available at SSRN 4872625*.

Francke, M. and Van de Minne, A. (2021a). Modeling unobserved heterogeneity in hedonic price models. *Real Estate Economics*, 49(4):1315–1339.

Francke, M. and Van de Minne, A. (2021b). Modeling unobserved heterogeneity in hedonic price models. *Real Estate Economics*, 49(4):1315–1339.

Francke, M. and Van de Minne, A. (2022). Daily appraisal of commercial real estate a new mixed frequency approach. *Real Estate Economics*, 50(5):1257–1281.

Francke, M. K. and Van de Minne, A. (2017). The hierarchical repeat sales model for granular commercial real estate and residential price indices. *The Journal of Real Estate Finance and Economics*, 55(4):511–532.

Geltner, D., Francke, M., Shimizu, C., Fenwick, D., and Baran, D. (2017). Commercial property price indicators: sources, methods and issues. Technical report, Eurostat.

Geltner, D. M. and Minne, A. v. d. (2017). Do different price points exhibit different investment risk and return in commercial real estate?

Geman, S., Bienenstock, E., and Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1–58.

Genesove, D. and Mayer, C. (2001). Loss aversion and seller behavior: Evidence from the housing market. *The Quarterly Journal of Economics*, 116(4):1233–1260.

Giacoletti, M. (2021). Idiosyncratic risk in housing markets. *The Review of Financial Studies*, 34(8):3695–3741.

Goodman, A. C. and Thibodeau, T. G. (1998). Housing market segmentation. *Journal of housing economics*, 7(2):121–143.

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438.

Griliches, Z. (1961). Hedonic price indexes for automobiles: An econometric of quality change. In *The price statistics of the federal goverment*, pages 173–196. NBER.

Gu, S., Kelly, B., and Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5):2223–2273.

Guo, X., Zheng, S., Geltner, D., and Liu, H. (2014). A new approach for constructing home price indices: The pseudo repeat sales model and its application in china. *Journal of Housing Economics*, 25:20–38.

Haas, G, C. (1922a). *Sale prices as a basis for farmland appraisal*, volume 9. University Farm.

Haas, G. C. (1922b). A statistical analysis of farm sales in blue earth county, minnesota, as a basis for farm land appraisal. Technical report.

Harter-Dreiman, M. (2004). Drawing inferences about housing supply elasticity from house price responses to income shocks. *Journal of Urban Economics*, 55(2):316–337.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Hill, R. (2011). Hedonic price indexes for housing. *OECD Statistics Working Paper Series*.

Hill, R. J. and Melser, D. (2008). Hedonic imputation and the price index problem: an application to housing. *Economic Inquiry*, 46(4):593–609.

Hjort, A., Hermansen, G. H., Pensar, J., and Williams, J. P. (2023). Uncertainty quantification in automated valuation models with locally weighted conformal prediction. *arXiv preprint arXiv:2312.06531*.

Huberman, G. and Regev, T. (2001). Contagious speculation and a cure for cancer: A nonevent that made stock prices soar. *The Journal of Finance*, 56(1):387–396.

Kato, Y., Tax, D. M., and Loog, M. (2023). A review of nonconformity measures for conformal prediction in regression. *Conformal and Probabilistic Prediction with Applications*, pages 369–383.

Knight, J. R. (2002). Listing price, time on market, and ultimate selling price: Causes and effects of listing price changes. *Real estate economics*, 30(2):213–237.

Koenker, R. and Bassett Jr, G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50.

Koenker, R. and d'Orey, V. (1987). Algorithm as 229: Computing regression quantiles. *Applied statistics*, 46(1):383–393.

Kohavi, R. et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada.

Kok, N., Koponen, E.-L., and Martínez-Barbosa, C. A. (2017). Big data in real estate? From manual appraisal to automated valuation. *The Journal of Portfolio Management*, 43(6):202–211.

Krause, A., Martin, A., and Fix, M. (2020). Uncertainty in automated valuation models: Error-based versus model-based approaches. *Journal of Property Research*, 37(4):308–339.

Lancaster, K. J. (1966). A new approach to consumer theory. *Journal of Political Economy*, 74(2):132–157.

Liebersohn, J. and Rothstein, J. (2025). Household mobility and mortgage rate lock. *Journal of Financial Economics*, 164:103973.

Lim, Z. and Bellotti, A. (2021). Normalized nonconformity measures for automated valuation models. *Expert Systems with Applications*, 180:115165.

Lin, D. (2024). From vacancy to verdancy? the impact of land options on the timing of vacant lot investment. *Real Estate Economics*, 52(5):1263–1307.

Lin, Z. and Liu, Y. (2008). Real estate returns and risk with heterogeneous investors. *Real Estate Economics*, 36(4):753–776.

Machado, J. A. and Mata, J. (2005). Counterfactual decomposition of changes in wage distributions using quantile regression. *Journal of applied Econometrics*, 20(4):445–465.

Malpezzi, S. (2002). Hedonic pricing models: a selective and applied review. *Housing economics and public policy*, pages 67–89.

McGreal, S., Adair, A., McBurney, D., and Patterson, D. (1998). Neural networks: the prediction of residential values. *Journal of Property Valuation and Investment*, 16(1):57–70.

McMillen, D. and Shimizu, C. (2021). Decompositions of house price distributions over time: The rise and fall of tokyo house prices. *Real Estate Economics*, 49(4):1290–1314.

McMillen, D. P. (2008). Changes in the distribution of house prices over time: Structural characteristics, neighborhood, or coefficients? *Journal of Urban Economics*, 64(3):573–589.

McMillen, D. P. and Dombrow, J. (2001). A flexible fourier approach to repeat sales price indexes. *Real Estate Economics*, 29(2):207–225.

Mullainathan, S. and Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106.

Nghiep, N. and Al, C. (2001). Predicting housing value: A comparison of multiple regression analysis and artificial neural networks. *Journal of Real Estate Research*, 22(3):313–336.

Nicholson, W., Matteson, D., and Bien, J. (2017). Bigvar: Tools for modeling sparse high-dimensional multivariate time series.

Nicholson, W. B., Wilms, I., Bien, J., and Matteson, D. S. (2020). High dimensional forecasting via interpretable vector autoregression. *The Journal of Machine Learning Research*, 21(1):6690–6741.

Nowak, A. D. and Smith, P. S. (2020). Quality-adjusted house price indexes. *American Economic Review: Insights*, 2(3):339–56.

Oaxaca, R. (1973). Male-female wage differentials in urban labor markets. *International economic review*, pages 693–709.

Park, H. and Sakaori, F. (2013). Lag weighted lasso for time series model. *Computational Statistics*, 28:493–504.

Peng, L. and Thibodeau, T. G. (2017). Idiosyncratic risk of house prices: Evidence from 26 million home sales. *Real Estate Economics*, 45(2):340–375.

Penrose, R. (1955). A generalized inverse for matrices. *Proceedings of the Cambridge Philosophical Society*, 51(3):406–413.

Pesaran, H. H. and Shin, Y. (1998). Generalized impulse response analysis in linear multivariate models. *Economics letters*, 58(1):17–29.

Peterson, S. and Flanagan, A. (2009). Neural network hedonic pricing models in mass real estate appraisal. *Journal of Real Estate Research*, 31(2):147–164.

Pollestad, A. J., Næss, A. B., and Oust, A. (2024). Towards a better uncertainty quantification in automated valuation models. *The Journal of Real Estate Finance and Economics*, pages 1–37.

Pollestad, A. J. and Oust, A. (2025). Harnessing uncertainty: a new approach to real estate investment decision support. *Quantitative Finance*, pages 1–18.

Quinlan, J. R. et al. (1992). Learning with continuous classes. In *5th Australian joint conference on artificial intelligence*, volume 92, pages 343–348. World Scientific.

Ripley, B. D. (1996). *Pattern recognition and neural networks*. Cambridge University Press.

Rosen, S. (1974). Hedonic prices and implicit markets: product differentiation in pure competition. *Journal of Political Economy*, 82(1):34–55.

Sagi, J. S. (2021). Asset-level risk and return in real estate investments. *The Review of Financial Studies*, 34(8):3647–3694. Published online 2020-10-27.

Scholköpf, B. and Smola, A. J. (2002). *Learning with Kernels*. The MIT Press.

Shafer, G. and Vovk, V. (2008). A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3).

Shao, A. W., Hanewald, K., and Sherris, M. (2015). Reverse mortgage pricing and risk analysis allowing for idiosyncratic house price risk and longevity risk. *Insurance: Mathematics and Economics*, 63:76–90.

Smola, A. J. and Scholköpf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222.

Sommervoll, D. E. and de Haan, J. (2014). Homes and castles: Should we care about idiosyncratic risk? *Land Economics*, 90(4):700–716. DOI not consistently listed in standard metadata; add if your source provides it.

Steurer, M., Hill, R. J., and Pfeifer, N. (2021). Metrics for evaluating the performance of machine learning based automated valuation models. *Journal of Property Research*, 38(2):99–129.

Tay, D. P. and Ho, D. K. (1992). Artificial intelligence and the mass appraisal of residential apartments. *Journal of Property Valuation and Investment*, 10(2):525–540.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3):1139–1168.

Tetlock, P. C. (2011). All the news that's fit to reprint: Do investors react to stale information? *The Review of Financial Studies*, 24(5):1481–1512.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Trojanek, R. and Huderek-Glapska, S. (2018). Measuring the noise cost of aviation: the association between the Limited Use Area around Warsaw Chopin Airport and property values. *Journal of Air Transport Management*, 67:103–114.

van Dijk, D. W. (2024). Local constant-quality housing market liquidity indices. *Regional Science and Urban Economics*, 106:103997.

van Dijk, D. W. and Francke, M. K. (2018). Internet search behavior, liquidity and prices in the housing market. *Real Estate Economics*, 46(2):368–403.

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28(2):3–28.

Vovk, V. (2012). Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pages 475–490. PMLR.

Vovk, V. (2015). Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence*, 74:9–28.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic learning in a random world*, volume 29. Springer.

Vovk, V., Shen, J., Manokhin, V., and Xie, M.-g. (2017). Nonparametric predictive distributions based on conformal prediction. In *Conformal and probabilistic prediction and applications*, pages 82–102. PMLR.

Wallace, H. A. (1926). Comparative farmland values in iowa. *Journal of Land and Public Utility Economics*, Oct.:385–392.

Wang, H., Yu, F., and Zhou, Y. (2020). Property investment and rental rate under housing price uncertainty: a real options approach. *Real Estate Economics*, 48(2):633–665.

Webb, G. I. (2000). Multiboosting: A technique for combining boosting and wagging. *Machine learning*, 40(2):159–196.

Wheaton, W. C. (1990). Vacancy, search, and prices in a housing market matching model. *Journal of political Economy*, 98(6):1270–1292.

Wheaton, W. C. (1999). Real estate 'cycles': Some fundamentals. *Real Estate Economics*, 27(2):209–230.

Winkler, R. L. (1972). A decision-theoretic approach to interval estimation. *Journal of the American Statistical Association*, 67(337):187–191.

Wong, K., So, A. T., and Hung, Y. (2002). Neural network vs. hedonic price model: Appraisal of high-density condominiums. In *Real estate valuation theory*, pages 181–198. Springer.

Worzala, E., Lenk, M., and Silva, A. (1995). An exploration of neural networks and its application to real estate valuation. *Journal of Real Estate Research*, 10(2):185–201.

Summary

Machine Learning in Real Estate Valuation

Real estate markets are characterized by extreme asset heterogeneity, infrequent trading, and data that are often incomplete or measured with error. These features make valuation difficult, but they also make valuation outputs unusually consequential: appraisals and automated valuation models (AVMs) influence mortgage credit decisions, collateral management, portfolio monitoring, investment performance measurement, taxation, and market narratives. Over the past decades, the "information layer" of the built environment has expanded rapidly as transaction registers, listings, geospatial layers, remote-sensing products, and administrative records have become more widely available in digital form. In parallel, advances in computation and statistical learning have broadened the set of predictive models that can be applied to these data. Yet the core challenge remains: the value of a valuation system is ultimately judged not only by whether it can produce accurate point predictions, but by whether it can produce market information that is comparable over time, informative about uncertainty, and capable of revealing heterogeneous dynamics across segments.

This thesis advances that measurement perspective on machine learning (ML) in real estate valuation. The overarching argument is that AVMs are not merely predictive tools; they are measurement devices that transform heterogeneous micro data into objects used for economic interpretation and decision-making. The thesis therefore examines how modern predictive methods can be integrated into three valuation outputs that recur across research and practice. The first is market-level price indices, which summarize price change through time while adjusting for changing quality and composition. The second is property-level uncertainty statements, which translate historical prediction errors into calibrated information about the range of plausible prices for a given dwelling. The third is distributional market monitoring, which shifts attention from a single mean or median index to the joint evolution of multiple quantile indices and asks whether

some segments of the price distribution systematically lead others. Together, these perspectives connect a cross-sectional valuation problem (predicting prices) to three distinct but complementary measurement tasks (indexing, uncertainty quantification, and distributional dynamics).

The thesis is organized as three research articles preceded by a general introduction and followed by a synthesis. Chapter 2 develops and evaluates a model-agnostic framework for price index construction that allows the prediction step to be carried out with either econometric benchmarks or flexible ML models, while keeping the index aggregation logic interpretable and comparable. Chapter 3 studies prediction intervals (PIs) as practical, model-agnostic measures of uncertainty around AVM point predictions and clarifies how interval behavior depends jointly on the PI methodology and the underlying regression algorithm. Chapter 4 constructs quantile price indices for the housing market and treats them as a multivariate time series, using sparse dynamic modeling to investigate segmentation and lead–lag relationships across the price distribution. Chapter 5 synthesizes the results and draws implications for how valuation systems should be evaluated when their outputs serve measurement and risk-sensitive purposes.

Chapter 2 focuses on the long-standing challenge of constructing price indices in markets where assets are not directly comparable over time. Classical index approaches—such as hedonic regressions, repeat-sales methods, and hybrid designs—typically embed strong parametric structure. The chapter proposes an alternative perspective that separates two components: the predictive model used to estimate counterfactual prices and the index logic used to aggregate those predictions into a time series. The methodological core is to use out-of-time prediction errors at the transaction level to measure price change. Each transaction is valued using a model trained on data that exclude the transaction's time period, and the resulting prediction error is interpreted as a period-specific price movement conditional on observed characteristics. These transaction-level signals are then aggregated using a chained index formula, yielding a market index that can be constructed with any regression model capable of producing point predictions.

The out-of-time setup operationalizes the counterfactual that defines the chained index. In each period t, the regression model is estimated using only the cross-section of transactions observed in t, so it captures the period-t pricing relationship between observed characteristics and log sale prices. The model is then applied to the next period's transaction set, S_{t+1}, to predict what the properties sold in $t+1$ would have transacted for under period-t market conditions, i.e., $\widehat{Y}^{t}_{t+1} \mid (X_{t+1}, \mathscr{F}_t)$. The period log return is obtained as the average gap between observed log prices in $t+1$ and these period-t counterfactual valuations, and chaining these returns yields the Chained Paasche index.

This construction follows the logic of chained hedonic indices but makes the "pricing step" explicitly modular: any regression engine that maps characteristics to prices can be used to generate the counterfactual valuations, which makes the approach model-agnostic and enables like-for-like comparisons across linear and non-linear predictors within the same index formula.

The empirical application uses Real Capital Analytics data covering 29,998 commercial real estate transactions in the New York metropolitan area from 2000 to 2019 and estimates both an aggregate index and indices by property type. The benchmark is an OLS hedonic model, compared against several non-linear ML regressors (XGBT, SVR, avNNet, and Cubist), with hyperparameters tuned period-by-period. In the full-sample chained setting, the ML models achieve higher out-of-time accuracy than OLS (lower RMSE and higher R^2), while the resulting index returns remain well-behaved: volatilities are comparable and first-order autocorrelation is non-negative across algorithms. Differences are more visible in index levels than in returns, with SVR and Cubist producing the largest deviations from the benchmark. A key practical qualification concerns data dependence. When the methodology is stress-tested by repeatedly subsampling the available transactions (50%, 25%, 10%, and 5% of the data), ML performance deteriorates more sharply than OLS as sample sizes shrink, and this becomes most evident under the strongest restriction, consistent with greater sensitivity of flexible learners to limited information. The property-type indices reinforce this point: segments with more observations (e.g., apartments) are more supportive of ML gains, whereas thin segments display higher variation and weaker separation from the benchmark. Finally, the paper highlights that some ML algorithms exhibit non-zero differences between fitted and observed mean log prices, i.e., estimation bias, which matters in a single-imputation chained setting because the index return is constructed from mean differences. Comparing single and double imputation illustrates how this bias can affect long-run index levels and motivates double imputation as a straightforward mitigation when bias is present.

Chapter 3 treats uncertainty as an explicit output of residential AVMs and asks how reliably different prediction-interval (PI) constructions translate point-prediction errors into calibrated ranges for transaction prices. The chapter's contribution is to isolate, within a single empirical design, how much PI performance is driven by the interval methodology itself versus the underlying regression algorithm. Prediction intervals are evaluated not only by whether they attain the nominal coverage rate, but also by how informative they are in practice—their average width, how strongly width varies across dwellings and neighborhoods, and whether the resulting bounds are symmetric or skewed.

Methodologically, the chapter implements four model-agnostic PI approaches—Cross

Conformal Prediction (CCP), Normalized Conformal Prediction (NCP), Direct Prediction (DP), and Direct Lower–Upper Prediction (DLUP)—on three regression algorithms: a hedonic OLS benchmark and two ML (XGBoost and Cubist). The methods differ in the way residual information is converted into bounds: CCP produces constant-width intervals; NCP scales interval length using a conditional scale model to obtain heterogeneous widths; DP targets an upper residual quantile and yields symmetric intervals by construction; and DLUP predicts lower and upper residual quantiles separately, allowing asymmetric bounds. Performance is assessed using coverage and mean interval width, complemented by dispersion of widths, under/over-interval diagnostics, and the Winkler score, under both a pooled split-sample evaluation and a rolling one-year-ahead design.

Empirically, using 86,838 Amsterdam transactions (2001–2021), the results show a robust separation between calibration and tightness. Across specifications, the regression engine mainly determines how narrow and heterogeneous the intervals can be, while the PI methodology is more influential for attaining nominal coverage. In the pooled split-sample setting, Cubist paired with DLUP yields the strongest trade-off between coverage and width in the main results, reflecting the value of allowing asymmetric bounds when residual risk is not symmetric; CCP attains coverage closest to nominal but at the cost of wider, homogeneous intervals. The rolling-window exercise is more demanding: coverage is broadly stable but several combinations under-cover, with CCP remaining the most conservative and DLUP among the tightest yet most prone to undercoverage when trained on a single year. Overall, the chapter shows that selecting an uncertainty layer cannot be separated from selecting a point predictor: interval calibration is largely methodological, but interval informativeness is largely model-driven, so both should be evaluated jointly under the intended deployment regime.

Chapter 4 studies distributional market monitoring by asking whether changes in some parts of the housing price distribution tend to precede changes in others. The motivation is that mean- or median-based indices summarize a dominant common trend but can obscure segmentation: different price tiers may adjust at different speeds, and these timing differences may contain information about how market conditions propagate across the distribution. The chapter therefore treats lead–lag relations across quantile indices as an empirical question and investigates whether the distribution's own internal dynamics—using only transaction data—reveal systematic predictive structure.

Methodologically, the chapter integrates distributional price-index estimation with high-dimensional multivariate time-series methods. It first constructs quarterly quantile price indices from the 10th to the 90th percentiles using Recentered Influence Function (RIF) regressions, which target unconditional quantiles and therefore track movements

in the marginal price distribution rather than conditional quantiles for a fixed dwelling type. Because quarterly quantile returns can be noisy, their log returns are smoothed using the Generalized Inverse Estimator (GIE), adapted as a denoising step to recover a cleaner latent return signal for each quantile. The joint dynamics of the resulting panel of (smoothed) quantile returns are then modeled with a lag-weighted Lasso VAR (LWL-VAR), which imposes sparsity in a 17-series system while shrinking distant lags more strongly than recent ones; the regularization level is selected by an expanding-window forecasting procedure that targets out-of-sample mean squared forecast error. Dynamic propagation is summarized with generalized impulse response functions (GIRFs), which provide order-invariant response profiles, and the high-dimensional results are cross-checked in a conventional low-dimensional VAR with standard Granger-causality tests for a representative subset of quantiles.

The empirical application uses 175,070 residential real estate transactions in Warsaw from 2006 to 2024. The estimated quantile indices display a persistent "fan": lower quantiles remain above higher quantiles throughout the sample and the dispersion widens over time, with lower quantiles appreciating somewhat faster than upper quantiles. After smoothing, the return series are stationary and suitable for VAR analysis, and the forecasting exercise shows that smoothing is not cosmetic: models estimated on unsmoothed quantile returns perform poorly out of sample, whereas specifications using smoothed returns markedly reduce forecast errors relative to standard benchmarks. Within the preferred smoothed specifications, the LWL-VAR coefficient maps exhibit a stable pattern of concentrated predictive links, with a pronounced role for the 60th percentile and a weaker but persistent role for the 10th percentile; higher quantiles, especially the upper tail, behave more as receivers than initiators of movements. The GIRFs corroborate this asymmetry: impulses concentrated in the 10th or 60th percentiles generate stronger system-wide responses than impulses in quantiles whose LWL-VAR coefficients are largely shrunk to zero (such as the 45th or 75th percentiles). A reduced VAR and Granger-causality tests provide a complementary confirmation in a familiar setting, with evidence that predictive relations run predominantly from lower and middle segments toward higher quantiles and that the uppermost segment contributes comparatively little predictive content. The chapter's substantive message is thus that quantile returns do not move in lockstep even when the underlying indices share a strong common trend, and that distributional dynamics can be used to identify which parts of the market tend to move first—with the specific "leading" quantiles (here, the 10th and 60th percentiles) best interpreted as context-dependent rather than universal.

Across the three papers, a unifying theme is that more flexible prediction does not elim-

inate traditional measurement challenges in real estate; instead, it changes where those challenges appear and how they should be managed. In index construction, stronger predictive models can improve fit, but they can also increase sensitivity to sample composition and to design choices that are easy to overlook when focusing only on average accuracy. In uncertainty quantification, the validity of interval statements is not guaranteed by a high-performing point predictor, and reliable uncertainty reporting requires explicit evaluation of coverage and width under realistic out-of-sample designs. In distributional monitoring, the use of multiple quantile indices reveals segmentation and staggered adjustment that aggregate measures can mask, but it also requires careful treatment of noise and high dimensionality to extract interpretable dynamics.

The thesis makes three contributions. First, it extends the practical scope of regression-based and hedonic index ideas by providing a model-agnostic framework that connects modern predictive models to index construction through an out-of-time prediction design. This clarifies how ML can be used in indexing without redefining the index methodology for each algorithm and highlights stability as a central criterion alongside accuracy. Second, it provides real-estate-specific evidence on model-agnostic uncertainty quantification by showing how PI methodology and regression model choice interact in practice and by documenting a consistent separation between validity and efficiency. This offers an actionable way for practitioners to design uncertainty layers for AVMs and to anticipate how uncertainty measures behave across model classes and data regimes. Third, it contributes to the literature on housing market heterogeneity by treating quantile indices as a multivariate time series and applying sparse dynamic modeling to uncover lead–lag relationships across the distribution. This extends distributional index work from static description to dynamic segmentation analysis under high dimensionality.

Two implications follow. First, evaluation should be matched to the valuation output being produced. Point-prediction accuracy is informative about the regression component of an AVM, but it does not by itself establish that a price index is stable and comparable over time, that prediction intervals attain the intended coverage while remaining informative, or that distributional monitoring captures segmentation rather than only a common trend. Each output therefore requires output-specific diagnostics: stability and sensitivity checks for indices, joint assessment of coverage and width (including heterogeneity) for prediction intervals, and dependence measures that are able to represent dynamic links across quantiles. Second, the results underscore that modelling choices and data constraints transmit directly to the reported measurement objects. In operational settings, this motivates clearly documented modelling protocols and routine robustness checks that assess the outputs directly, such as index paths, prediction interval coverage

and width, and the stability of estimated dependence patterns, rather than relying on a single summary metric for the underlying predictor.

The limitations and extensions are chapter-specific. A general limitation is that the three empirical applications use different markets and data environments, so the numerical magnitudes should not be treated as directly transferable across settings with different liquidity, institutions, and measurement conventions. Beyond this general point, Chapter 2 is most affected by data availability and sample composition. In thin segments or under restricted samples, index behaviour can become sensitive to feature availability, imputation, and estimation choices, which motivates further work on robustness when such indices are produced operationally and on whether richer measures of quality, including information extracted from text and images, can improve comparability through time. Chapter 3 is limited by its focus on prediction intervals rather than full predictive distributions. This leaves open extensions that deliver probability statements at thresholds and tail-focused risk measures, and that allow calibration to be assessed beyond coverage and width. Chapter 4 documents systematic dependence across quantile indices, but the evidence remains reduced-form. A natural next step is to connect the observed timing patterns to candidate mechanisms, for example credit conditions, buyer composition, and differential supply responses, and to test whether these relations change around identifiable regime shifts.

Looking ahead, the thesis fits into a measurement agenda in which valuation links heterogeneous data to economic interpretation and institutional use. As data environments become richer, the key requirement is that valuation outputs remain comparable over time and open to empirical scrutiny, not merely computable at scale. The thesis therefore supports a simple conclusion: progress in machine learning matters for real estate economics when it improves the reliability and interpretability of the measures used to monitor markets and make decisions, not when it only reduces average prediction error.

Nederlandse samenvatting

Machine Learning in Real Estate Valuation

Vastgoedmarkten worden gekenmerkt door extreme heterogeniteit van objecten, infrequente verkopn en gegevens die vaak onvolledig zijn of met meetfouten zijn belast. Deze kenmerken maken het waarderen lastig, maar zorgen er ook voor dat waarderingsuitkomsten uitzonderlijk belangrijk zijn: taxaties en geautomatiseerde waarderingsmodellen (automated valuation models, AVMs) beïnvloeden beslissingen over hypotheekverstrekking, onderpandbeheer, portefeuillesurveillance, het meten van beleggingsprestaties, belastingheffing en marktpercepties. In de afgelopen decennia is de "informatielaag" van de gebouwde omgeving snel gegroeid doordat transactieregisters, woning advertenties, georuimtelijke lagen, remote-sensingproducten en administratieve bestanden in digitale vorm breder beschikbaar zijn geworden. Tegelijkertijd hebben vooruitgang in rekenkracht en statistische leermethoden de verzameling voorspellende modellen vergroot die op deze data kan worden toegepast. De kernuitdaging blijft echter bestaan: de waarde van een waarderingssysteem wordt uiteindelijk niet alleen beoordeeld op het vermogen om nauwkeurige puntvoorspellingen te leveren, maar ook op het vermogen om marktinformatie te produceren die door de tijd vergelijkbaar is, inzicht geeft in onzekerheid en heterogene dynamiek tussen segmenten kan blootleggen.

Dit proefschrift ontwikkelt die meetkundige invalshoek op machine learning (ML) in vastgoedwaardering. Het overkoepelende uitgangspunt is dat AVMs niet louter voorspellende instrumenten zijn, maar meetinstrumenten die heterogene micro-data omzetten in objecten die worden gebruikt voor economische interpretatie en besluitvorming. Het proefschrift onderzoekt daarom hoe moderne voorspellingsmethoden kunnen worden geïntegreerd in drie waarderingsuitkomsten die terugkeren in onderzoek en praktijk. De eerste betreft marktbrede prijsindices, die prijsveranderingen door de tijd samenvatten en tegelijk corrigeren voor verandering in kwaliteit en samenstelling. De tweede betreft onzekerheidsuitspraken op objectniveau, die historische voorspellingsfouten vertalen

naar gekalibreerde informatie over de bandbreedte van plausibele prijzen voor een specifieke woning. De derde betreft distributionele marktmonitoring, die de aandacht verlegt van één gemiddelde of mediane index naar de gezamenlijke ontwikkeling van meerdere kwantielindices en onderzoekt of sommige segmenten van de prijsverdeling systematisch andere segmenten leiden. Samen verbinden deze perspectieven een cross-sectioneel waarderingsprobleem (prijzen voorspellen) met drie onderscheiden maar complementaire meetopgaven (indexering, onzekerheidskwantificatie en distributionele dynamiek).

Het proefschrift bestaat uit drie onderzoeksartikelen, voorafgegaan door een algemene inleiding en afgesloten met een synthese. Hoofdstuk 2 ontwikkelt en evalueert een model-agnostisch raamwerk voor de constructie van prijsindices waarin de voorspellingsstap kan worden uitgevoerd met zowel econometrische benchmarks als flexibele ML-modellen, terwijl de aggregatielogica van de index interpreteerbaar en vergelijkbaar blijft. Hoofdstuk 3 bestudeert voorspellingsintervallen (prediction intervals, PIs) als praktische, model-agnostische onzekerheidsmaten rondom AVM-puntvoorspellingen en verduidelijkt hoe intervalgedrag gezamenlijk wordt bepaald door de PI-methodologie en het onderliggende regressie-algoritme. Hoofdstuk 4 construeert kwantielprijsindices voor de woningmarkt en behandelt deze als een multivariate tijdreeks, waarbij spaarzame dynamische modellering wordt gebruikt om segmentatie en lead–lag-relaties over de prijsverdeling te onderzoeken. Hoofdstuk 5 vat de resultaten samen en trekt conclusies voor de evaluatie van waarderingssystemen wanneer hun uitkomsten een meetfunctie hebben en in risicogevoelige toepassingen worden gebruikt.

Hoofdstuk 2 richt zich op de klassieke uitdaging om prijsindices te construeren in markten waarin objecten door de tijd niet direct vergelijkbaar zijn. Klassieke indexbenaderingen, zoals hedonische regressies, herhaalde verkopen methoden en hybride ontwerpen, bevatten doorgaans een sterke parametrische structuur. Dit hoofdstuk stelt een alternatief perspectief voor dat twee componenten onderscheidt: het voorspellende model waarmee contrafeitelijke prijzen worden geschat en de indexlogica waarmee die voorspellingen worden geaggregeerd tot een tijdreeks. De methodologische kern is het gebruik van out-of-time voorspellingsfouten op transactieniveau om prijsverandering te meten. Elke transactie wordt gewaardeerd met een model dat is getraind op data die de periode van de transactie uitsluiten, en de resulterende voorspellingsfout wordt geïnterpreteerd als een periodieke prijsbeweging conditioneel op de waargenomen kenmerken. Deze transactiesignalen worden vervolgens geaggregeerd met een chained indexformule, wat resulteert in een marktindex die kan worden berekend met elk regressiemodel dat puntvoorspellingen kan leveren.

De out-of-time opzet operationaliseert het contrafeitelijke gedachte-experiment dat de ketting (chained) index definieert. In elke periode t wordt het regressiemodel geschat met uitsluitend de cross-sectie van transacties die in t worden waargenomen, zodat het de periode-t prijsrelatie tussen waargenomen kenmerken en log-verkoopprijzen vastlegt. Het model wordt vervolgens toegepast op de verzamling van transacties in de volgende periode, S_{t+1}, om te voorspellen voor welke prijs de in $t+1$ verkochte objecten onder periode-t marktomstandigheden zouden zijn verhandeld, dat wil zeggen $\widehat{Y}^{t}_{t+1} \mid (X_{t+1}, \mathscr{F}_t)$. Het periodieke log-rendement wordt verkregen als het gemiddelde verschil tussen waargenomen log-prijzen in $t+1$ en deze periode-t contrafeitelijke waarderingen, en door deze rendementen te verbinden ontstaat de Chained Paasche-index. Deze constructie volgt de logica van chained hedonische indices, maar maakt de "pricing step" expliciet modulair: elk regressie-algoritme dat kenmerken aan prijzen koppelt kan worden gebruikt om contrafeitelijke waarderingen te genereren, wat de aanpak model-agnostisch maakt en directe vergelijkingen tussen lineaire en niet-lineaire voorspellers binnen dezelfde indexformule mogelijk maakt.

De empirische toepassing gebruikt Real Capital Analytics-data met 29,998 commerciële vastgoedtransacties in het grootstedelijk gebied New York over 2000–2019 en schat zowel een geaggregeerde index als indices naar objecttype. De benchmark is een hedonisch OLS-model, vergeleken met meerdere niet-lineaire ML-regressors (XGBT, SVR, avNNet en Cubist), waarbij hyperparameters per periode worden getuned. In de full-sample chained setting behalen de ML-modellen een hogere out-of-time nauwkeurigheid dan OLS (lagere RMSE en hogere R^2), terwijl de resulterende indexrendementen zich ordelijk gedragen: volatiliteiten zijn vergelijkbaar en de eerste-orde autocorrelatie is niet-negatief over de algoritmen heen. Verschillen zijn zichtbaarder in indexniveaus dan in rendementen, waarbij SVR en Cubist de grootste afwijkingen ten opzichte van de benchmark laten zien. Een belangrijke praktische nuancering betreft data-afhankelijkheid. Wanneer de methodologie wordt onderworpen aan stresstesten door herhaaldelijk de beschikbare transacties te subsamplen (50%, 25%, 10% en 5% van de data), verslechtert de ML-prestatie sneller dan die van OLS naarmate de steekproef krimpt, en dit wordt het meest evident bij de sterkste restrictie, in overeenstemming met een grotere gevoeligheid van flexibele leeralgoritmen voor beperkte informatie. De indices naar objecttype ondersteunen dit punt: segmenten met meer observaties (bijv. appartementen) bieden meer ruimte voor ML-verbeteringen, terwijl dunne segmenten grotere variatie en minder duidelijke scheiding van de benchmark vertonen. Tot slot benadrukt het artikel dat sommige ML-algoritmen een niet-nul verschil vertonen tussen gefitte en waargenomen gemiddelde log-prijzen, dat wil zeggen schattingsbias, wat rele-

vant is in een single-imputation chained setting omdat het indexrendement is gebaseerd op gemiddelde verschillen. De vergelijking tussen single en double imputation illustreert hoe deze bias de indexniveaus op langere termijn kan beïnvloeden en motiveert double imputation als een eenvoudige mitigatie wanneer bias aanwezig is.

Hoofdstuk 3 behandelt onzekerheid als een expliciete uitkomst van residentiële AVMs en onderzoekt hoe betrouwbaar verschillende constructies van voorspellingsintervallen puntvoorspellingsfouten vertalen naar gekalibreerde bandbreedtes voor transactieprijzen. De bijdrage van het hoofdstuk is het isoleren, binnen één empirisch ontwerp, in welke mate PI-prestaties worden gedreven door de intervalmethodologie zelf versus het onderliggende regressie-algoritme. Voorspellingsintervallen worden niet alleen beoordeeld op het behalen van de nominale dekkingsgraad, maar ook op hun informatiewaarde in de praktijk: de gemiddelde breedte, de mate waarin breedte varieert tussen woningen en buurten, en de vraag of de grenzen symmetrisch dan wel scheef zijn.

Methodologisch implementeert het hoofdstuk vier verschillende model-agnostische PI-benaderingen – Cross Conformal Prediction (CCP), Normalized Conformal Prediction (NCP), Direct Prediction (DP) en Direct Lower–Upper Prediction (DLUP) – op drie regressie-algoritmen: een hedonische OLS-benchmark en twee ML-modellen (XGBoost en Cubist). De methoden verschillen in de manier waarop residu-informatie wordt omgezet in grenzen: CCP produceert intervallen met constante breedte; NCP schaalt de intervalbreedte met een conditioneel schaalmodel om heteroskedastische breedtes te verkrijgen; DP richt zich op een bovenkwantiel van de residuen en levert per constructie symmetrische intervallen; en DLUP voorspelt onder- en bovengrenzen als afzonderlijke residukwantielen, waardoor asymmetrische intervallen mogelijk worden. De prestatie wordt beoordeeld op basis van dekking en gemiddelde intervalbreedte, aangevuld met de spreiding van intervalbreedtes, under/over-intervaldiagnostiek en de Winkler-score, zowel onder een gepoolde split-sample evaluatie als onder een rollend one-year-ahead ontwerp.

Empirisch, met 86,838 transacties in Amsterdam (2001–2021), laten de resultaten een robuuste scheiding zien tussen calibratie en intervalbreedte. Over specificaties heen bepaalt het regressie-algoritme vooral hoe smal en heterogeen de intervallen kunnen zijn, terwijl de PI-methodologie meer invloed heeft op het behalen van de nominale dekking. In de gepoolde split-sample setting levert Cubist gecombineerd met DLUP in de hoofdresultaten de sterkste afruil tussen dekking en breedte, wat de waarde weerspiegelt van asymmetrische grenzen wanneer residueel risico niet symmetrisch is; CCP benadert de nominale dekking het best, maar ten koste van bredere, homogene intervallen. De rolling-window oefening is veeleisender: de dekking is in grote lijnen stabiel, maar

meerdere combinaties leiden tot onderdekking, waarbij CCP het meest conservatief blijft en DLUP vaak tot de smalste intervallen behoort maar ook het meest gevoelig is voor onderdekking wanneer slechts één jaar als trainingsperiode wordt gebruikt. In het geheel laat het hoofdstuk zien dat de keuze van een onzekerheidsniveau niet los kan worden gezien van de keuze van een puntvoorspeller: Interval kalibratie is grotendeels methodologisch, maar interval informativiteit is grotendeels modelgestuurd, dus beide moeten gezamenlijk worden geëvalueerd onder het beoogde implementatie regime.

Hoofdstuk 4 bestudeert distributionele marktmonitoring door te onderzoeken of veranderingen in bepaalde segmenten van de woningprijsverdeling veranderingen in andere segmenten voor gaan. De motivatie is dat indices gebaseerd op het gemiddelde of de mediaan een dominante gemeenschappelijke trend samenvatten, maar segmentatie kunnen verhullen; verschillende prijsklassen kunnen met verschillende snelheid aanpassen, en deze timingverschillen kunnen informatie bevatten over hoe marktomstandigheden zich over de verdeling verspreiden. Het hoofdstuk behandelt lead–lag-relaties tussen kwantielindices als een empirische vraag en onderzoekt of de interne dynamiek van de verdeling, uitsluitend op basis van transactiedata, een systema-tische voorspellende structuur blootlegt.

Methodologisch integreert het hoofdstuk distributionele prijsindexschatting met hoog-dimensionale multivariate tijdreeksmethoden. Eerst worden kwartaalgewijze kwantielprijsindices van het 10e tot en met het 90e percentiel geconstrueerd met Recentered Influence Function (RIF)-regressies, die zich richten op onconditionele kwantielen en daarom bewegingen in de marginale prijsverdeling volgen in plaats van conditionele kwantielen voor een vast type woning. Omdat kwartaalgewijze kwantielrendementen ruis kunnen bevatten, worden deze in logartime-vorm gesmoothd met de Generalized Inverse Estimator (GIE), als filter om een schoner rendementsignaal per kwantiel te verkrijgen. De gezamenlijke dynamiek van het resulterende panel van (gesmoothde) kwantielrendementen wordt vervolgens gemodelleerd met een lag-weighted Lasso VAR (LWL-VAR), die sparsity (spaarzaamheid) oplegt in een systeem van 17 reeksen en verre lags (vertragingen) sterker krimpt (shrinks) dan recente lags; het regularisatieniveau wordt gekozen via een expanding-window forecastingprocedure die out-of-sample mean squared forecast error minimaliseert. Dynamische propagatie wordt samengevat met generalized impulse response functions (GIRFs), die orde-invariante responsprofielen leveren, en de hoog-dimensionale resultaten worden getoetst met een conventionele laag-dimensionale VAR met standaard Granger-causaliteitstesten voor een representatieve deelverzameling van kwantielen.

De empirische toepassing gebruikt 175,070 woningverkopen in Warschau van 2006

tot 2024. De geschatte kwantielindices vertonen een persistent "waaierpatroon": lagere kwantielen blijven gedurende de steekproef boven hogere kwantielen en de spreiding neemt door de tijd toe, waarbij lagere kwantielen iets sneller stijgen dan hogere kwantielen. Na smoothing zijn de rendementsreeksen stationair en geschikt voor een VAR-analyse, en de voorspellingsoefening laat zien dat smoothing niet louter cosmetisch is: modellen geschat op niet-gesmoothde kwantielrendementen presteren slecht out-of-sample, terwijl specificaties met gesmoothde rendementen de forecast errors duidelijk verlagen ten opzichte van standaardbenchmarks. Binnen de voorkeursspecificaties laten de LWL-VAR-coëfficiëntenkaarten een stabiel patroon zien van geconcentreerde voorspellende verbanden, met een uitgesproken rol voor het 60e percentiel en een zwakkere maar persistente rol voor het 10e percentiel; hogere kwantielen, met name de bovenstaart, gedragen zich meer als ontvangers dan als initiators van bewegingen. De GIRFs bevestigen deze asymmetrie: schokken geconcentreerd in het 10e of 60e percentiel genereren sterkere systeembrede reacties dan schokken in kwantielen waarvan de LWL-VAR-coëfficiënten grotendeels naar nul worden gekrompen (zoals het 45e of 75e percentiel). Een reduced VAR en Granger-causaliteitstesten bieden een aanvullende bevestiging in een bekend kader, met aanwijzingen dat voorspellende relaties voornamelijk lopen van lagere en middensegmenten naar hogere kwantielen en dat het hoogste segment relatief weinig voorspellend is. De inhoudelijke boodschap van het hoofdstuk is daarmee dat kwantielrendementen niet synchroon bewegen, zelfs wanneer de onderliggende indices een sterke gemeenschappelijke trend delen, en dat distributionele dynamiek kan worden gebruikt om te identificeren welke delen van de markt de neiging hebben als eerste te bewegen, waarbij de specifieke "leidende" kwantielen (hier het 10e en 60e percentiel) het best contextafhankelijk kunnen worden geïnterpreteerd.

In alle drie de artikelen komt als rode draad naar voren dat flexibelere voorspellingen de traditionele meetproblemen in de vastgoedsector niet wegnemen, maar wel veranderen waar die problemen zich voordoen en hoe ze moeten worden aangepakt. Bij indexconstructie kunnen sterkere voorspellende modellen de fit verbeteren, maar zij kunnen ook de gevoeligheid vergroten voor steekproefsamenstelling en voor ontwerpkeuzes die gemakkelijk uit het oog worden verloren wanneer de aandacht vooral uitgaat naar gemiddelde nauwkeurigheidsmaten. Bij onzekerheidskwantificatie volgt de geldigheid van intervaluitspraken niet automatisch uit een goed presterende puntvoorspeller, en betrouwbare onzekerheidsrapportage vereist expliciete evaluatie van dekking en breedte onder realistische out-of-sample ontwerpen. Bij distributionele monitoring maakt het gebruik van meerdere kwantielindices segmentatie en gespreide aanpassing zichtbaar die geaggregeerde maten kunnen verhullen, maar dit vereist ook zorgvuldige behandeling

van ruis en hoge dimensionaliteit om interpretabele dynamiek te verkrijgen.

Het proefschrift levert drie bijdragen. Ten eerste verbreedt het de praktische reikwijdte van regressie-gebaseerde en hedonische indexideeën door een model-agnostisch raamwerk te bieden dat moderne voorspellende modellen via een out-of-time ontwerp aan indexconstructie koppelt. Dit verduidelijkt hoe ML in indexering kan worden toegepast zonder de indexmethodologie voor elk algoritme te herdefiniëren en benadrukt stabiliteit als een centraal criterium naast nauwkeurigheid. Ten tweede levert het vastgoedspecifiek bewijs over model-agnostische onzekerheidskwantificatie door te laten zien hoe PI-methodologie en regressiekeuze in de praktijk op elkaar inwerken en door een consistente scheiding tussen validiteit en efficiëntie te documenteren. Dit biedt een handelbaar kader voor het ontwerpen van onzekerheidslagen voor AVMs en voor het anticiperen op het gedrag van onzekerheidsmaten over modelklassen en dat regimes heen. Ten derde draagt het bij aan de literatuur over heterogeniteit in woningmarkten door kwantielindices als een multivariate tijdreeks te behandelen en spaarzame dynamische modellering toe te passen om lead–lag-relaties over de verdeling bloot te leggen. Dit breidt distributioneel indexonderzoek uit van statische beschrijving naar dynamische segmentatieanalyse onder hoge dimensionaliteit.

Twee implicaties volgen. Ten eerste moet evaluatie worden afgestemd op het product van waardering. Puntvoorspellingsnauwkeurigheid is informatief over de regressiecomponent van een AVM, maar zij toont op zichzelf niet aan dat een prijsindex stabiel en door de tijd vergelijkbaar is, dat voorspellingsintervallen de beoogde dekking behalen en tegelijk informatief blijven, of dat distributionele monitoring segmentatie vastlegt in plaats van uitsluitend een gemeenschappelijke trend. Elke uitkomst vereist daarom eigen diagnostiek: stabiliteits- en gevoeligheidscontroles voor indices, gezamenlijke beoordeling van dekking en breedte (inclusief heterogeniteit) voor voorspellingsintervallen, en afhankelijkheidsmaten die dynamische verbanden tussen kwantielen kunnen representeren. Ten tweede onderstrepen de resultaten dat modelkeuzes en databeperkingen direct doorwerken in de gerapporteerde meetobjecten. In operationele toepassingen motiveert dit helder gedocumenteerde modelleringsprotocollen en routinematige robuustheidscontroles die de uitkomsten zelf beoordelen, zoals indexpaden, dekking en breedte van voorspellingsintervallen en de stabiliteit van geschatte afhankelijkheidspatronen, in plaats van te vertrouwen op één samenvattende maat voor de onderliggende voorspeller.

De beperkingen en uitbreidingen zijn hoofdstukspecifiek. Een algemene beperking is dat de drie empirische toepassingen verschillende markten en data-omgevingen gebruiken, zodat de numerieke grootheden niet zonder meer als direct overdraagbaar kunnen worden beschouwd tussen settings met verschillende liquiditeit, instituties en

meetconventies. Los van dit algemene punt wordt Hoofdstuk 2 het sterkst beïnvloed door databeschikbaarheid en steekproefsamenstelling. In dunne segmenten of onder beperkte steekproeven kan indexgedrag gevoelig worden voor feature-beschikbaarheid, imputatie en schattingskeuzes, wat verder onderzoek motiveert naar robuustheid wanneer dergelijke indices operationeel worden geproduceerd en naar de vraag of rijkere kwaliteitsmaten, waaronder informatie uit tekst en beelden, de vergelijkbaarheid door de tijd kunnen verbeteren. Hoofdstuk 3 wordt begrensd door de focus op voorspellingsintervallen in plaats van volledige voorspellingsverdelingen. Dit laat uitbreidingen open die kansuitspraken op drempels en staartgerichte risicomaten leveren en die calibratie kunnen beoordelen voorbij dekking en breedte. Hoofdstuk 4 documenteert systematische afhankelijkheid tussen kwantielindices, maar het bewijs blijft reduced-form. Een logische volgende stap is de geobserveerde timingpatronen te verbinden met mogelijke mechanismen, bijvoorbeeld kredietcondities, koperssamenstelling en differentiële aanbodreacties, en te testen of deze relaties veranderen rond identificeerbare regimeschokken.

Vooruitkijkend past het proefschrift in een meetagenda waarin waardering heterogene data verbindt met economische interpretatie en institutioneel gebruik. Naarmate data-omgevingen rijker worden, is het essentieel dat waarderingsuitkomsten door de tijd vergelijkbaar blijven en empirisch getoetst kunnen worden, en niet alleen op schaal berekend kunnen worden. Het proefschrift ondersteunt daarom een eenvoudige conclusie: vooruitgang in machine learning is relevant voor de vastgoedeconomie wanneer zij de betrouwbaarheid en interpretabiliteit van de maatstaven verbetert waarmee markten worden gevolgd en beslissingen worden genomen, en niet wanneer zij uitsluitend de gemiddelde voorspellingsfout verlaagt.

Publications

Journal publications

Felipe D. Calainho, Alex M. van de Minne, Marc K. Francke, "A machine learning approach to price indices: applications in commercial real estate.", *Journal of Real Estate Finance and Economics*, vol. 68, no. 4, pp. 624–653, 2024.

Poster sessions

Felipe D. Calainho, "Quantifying Uncertainty in Real Estate Price Predictions:Methodological and Algorithmic Synergies," in *American Real Estate and Urban Economics Association–Allied Social Science Associations Conference*, Philadelphia (US), 2026.

Working papers in conference

Felipe D, Calainho, Alex M. van de Minne, Radoslaw Trojanek, Marc K. Francke, "Lead-lag Relationships in Quantile Index Percentiles," in *30th European Real Estate Society Annual Conference*, Gdansk (PL), 2024.

Author contributions

Chapter 2. "Using Machine Learning to Build Real Estate Price Indices", was co-authored with my promoter Marc Francke and co-promoter Alex van de Minne, with me as lead author. I developed the indexing methodology, implemented the empirical framework, and conducted the full set of model estimations and analyses. Alex van de Minne proposed the research question and contributed to the writing of the introduction; he also provided the Real Capital Analytics (RCA) transaction data used in the empirical application. Marc Francke provided detailed substantive feedback throughout the project and contributed to revising, editing, structuring, and formatting the manuscript.

Chapter 3. "Quantifying Uncertainty in Real Estate Price Predictions: Methodological and Algorithmic Synergies", is single-authored. The empirical analysis draws on residential transaction data provided by NVM, made available with the support of Ortec Finance. I designed and implemented the full methodological comparison, carried out all empirical analyses, and wrote the manuscript. Marc Francke provided critical feedback and guidance during the development of the chapter.

Chapter 4. "Lead–lag Relationships in Quantile Index Percentiles", was co-authored with my promoter Marc Francke, co-promoter Alex van de Minne, and Prof. Radoslaw Trojanek, with me as lead author. I led the empirical implementation, including the quantile index construction pipeline, the dynamic modelling and lead–lag analyses, and the drafting of the manuscript. Alex van de Minne proposed the research question, contributed to the writing of the introduction, and provided initial code for the index construction. Prof. Radoslaw Trojanek was responsible for data collection and cleaning and wrote the data section of the chapter. Marc Francke provided extensive feedback and contributed to reviewing, challenging, editing, structuring, and formatting the manuscript.

"The language of friendship is not words but meanings."

Henry David Thoreau

Acknowledgements

Few experiences reshape a person the way a PhD does. This one was no exception. It brought unexpected changes, took me to different corners of the world, and placed me alongside people I deeply respect and will not forget. A PhD transforms those who go through it, sometimes quietly, sometimes not. There were stretches of solitude and moments where the finish line felt genuinely distant, but doubt was never part of it. The topic kept me genuinely engaged throughout, artificial intelligence and real estate are fields where there is always more to understand, and the research environment I found myself in made the harder periods easier to endure. Looking back, I recognize how much I have changed since the beginning of this process, and how much of that change I owe to the people and institutions that accompanied me along the way.

I would like to thank Ortec Finance for sponsoring this research and for the consistent support and feedback provided throughout. I am also grateful to the Dutch Broker Association (NVM), Real Capital Analytics (RCA/MSCI), and the Polish Property Price Register for supplying the data that made this work possible. I want to acknowledge the MIT Center for Real Estate and the University of Connecticut for hosting me as a visiting researcher, experiences that broadened my perspective and enriched this work in ways that are difficult to fully quantify. Finally, I want to thank the European Real Estate Society and the American Real Estate and Urban Economics Association for the opportunity to present and discuss this research with peers across the field.

The foundation of this thesis rests on two people. Marc Francke brought patience, rigor, and an intellectual standard that raised the quality of everything produced under his guidance. He is a source of knowledge and inspiration in the truest sense, and a mentor whose influence will extend well beyond this thesis. Alex van de Minne is one of the most creative minds I have had the pleasure of working with. He opened his home to me during my visit to UConn, and while he would be the first to say he did not do much, the truth is that he consistently showed up, invested his time, and made things happen. To me, he has been something like an older brother from another mother, and I mean that in the fullest sense.

Among the members of my committee and the wider academic community that shaped this work, I want to acknowledge several people with particular warmth. Peter van Gool carries a rare passion for this field, one that is contagious, and that he generously extends to those who come after him. David Geltner taught me what it means to be a professional researcher: how to carry oneself, how to deliver ideas with precision, and how to bridge the world of practice with the demands of academic inquiry. The breadth of his thinking, spanning engineering, architecture, and real estate finance and economics, left a lasting impression. John Clapp contributed countless insights throughout the development of this thesis, always available, always generous with his time. He is the kind of scholar who invests genuinely in the next generation of researchers. Thies Lindenthal offered important comments and reflections, and his enthusiasm for the use of machine learning and artificial intelligence in real estate finance and economics made him a natural and valued interlocutor.

The years of this PhD were also years of friendship, and I want to name those friendships honestly. Dorinth van Dijk was my first friend in this field, the person who showed me around and set the tone for what was to come. From beers on the steps of a Boston house to a rollercoaster at Six Flags to what I will only refer to as the Cannoli incident, our shared memories are ones I return to with real affection. Yumei Wang was a constant and generous presence in the office, a great friend with whom I shared many lunches and many hours of work, and it was an honor to stand beside her as her paranymph. Willem Vlaming was one of the great pleasures of daily life at the university. Our conversations and the occasional beer after work were always something to look forward to, and watching him go from the person I first met to a husband was a privilege I am glad I did not miss. Ujjwal Sharma taught me a great deal, in particular about new technologies, and I still clearly remember him demonstrating GPT in room M4.04 back when the chat interface did not yet exist. Those conversations were always engaging and almost always ended in laughter. Robert Lee is one of the funniest and most genuinely lovable people I encountered during this period, and attending his wedding was one of the highlights of these years. Jaqueline Trinta arrived near the end of the journey and immediately understood the potential of this work. She was the first person outside my research group to truly believe in it and in the possibility of building something meaningful for the real estate industry. That belief meant more than she may know.

I also want to thank the broader group of PhD students from the finance department and room M40.4. Special mention to Adam, Yingjie, Remo, Andy, Lukas, Andre, Arjan, Thu, Yasmine, and Xinyi, and to everyone else whose name I may have failed to record here but whose company I valued.

To my parents: thank you for everything that came before this. This thesis is, in some ways, the materialization of a long-held dream, and it would not exist without the years of effort and sacrifice that preceded it. What you built made this possible.

And finally, Marina. Only you know the full weight of what these years contained. The sleepless nights, the setbacks, the distance, the doubt. I have said it before and I mean it completely: without you, I would not have left Brazil. I would not be here. During this PhD you gave me the two most important things in the world, Lucas and Daniel. Raising two children abroad, far from family, is not a small thing, and you faced every challenge that came with it with a strength and determination that I have always admired. This work belongs to both of us.

Draegd en Verdraagd[1]

[1]While going through our family history, my parents and I discovered the Lott family crest bearing the sentence above. Written in Early Modern Dutch, it translates to the classic maxim "Bear and Forbear" (or "Carry and Endure"). This stoic phrase delivers a two fold piece of wisdom: Draegd urges one to carry responsibilities and life's burdens with fortitude, while Verdraagd commands the patience, tolerance, and resilience required to peacefully endure hardships.

