

Nowcasting Transaction-Based House Price Indices Using Web-Scraped Listings and MIDAS Regression

Radoslaw Trojanek¹ Luke Hartigan² Norbert Pfeifer³ Miriam Steurer³

¹Poznan University of Economics and Business

²University of Sydney

³University of Graz

18 February 2026



THE UNIVERSITY OF SYDNEY

Acknowledgement: Hartigan partially funded by DP230100959

Motivation

Motivation

Real-time monitoring of housing markets remains a challenge for statistical agencies

Motivation

Real-time monitoring of housing markets remains a challenge for statistical agencies

Transaction-based price indices are considered most reliable. However, agencies cannot compile these on time because of delays in getting transaction data

Motivation

Real-time monitoring of housing markets remains a challenge for statistical agencies

Transaction-based price indices are considered most reliable. However, agencies cannot compile these on time because of delays in getting transaction data

Investigate if online real-estate list-price data can be used to reduce the time gap and improve house price measurement

Contribution

Contribution

Make two contributions:

Contribution

Make two contributions:

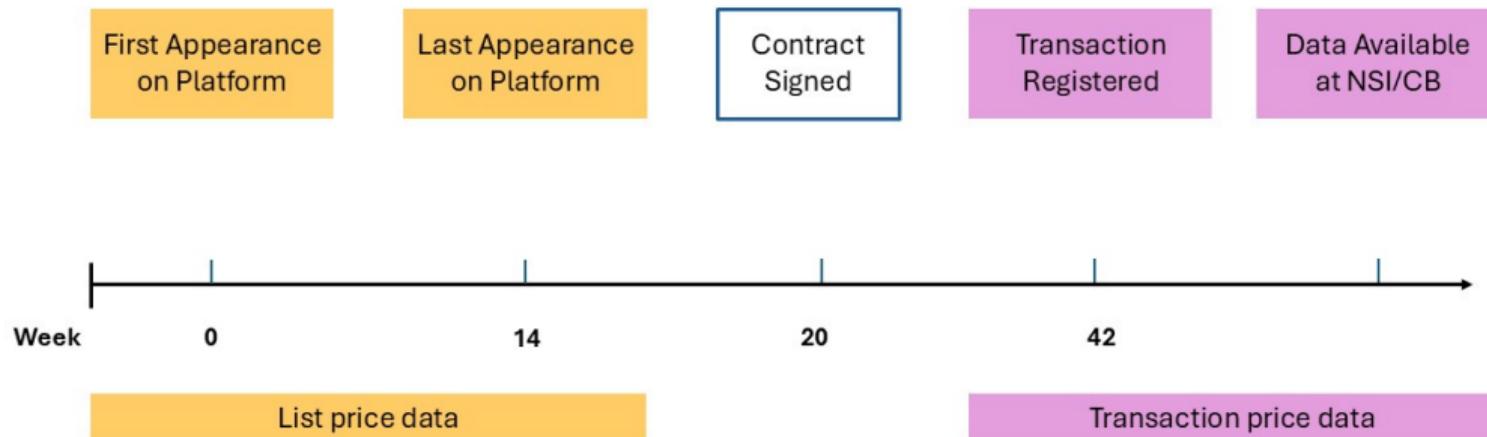
- 1) Compile hedonic price indices for both list and transaction data using 16 years of micro-level list and transaction data for Warsaw and Poznan

Contribution

Make two contributions:

- 1) Compile hedonic price indices for both list and transaction data using 16 years of micro-level list and transaction data for Warsaw and Poznan
- 2) Use MIDAS regression to nowcast property prices and show timely list-price indices can provide early and reliable signals of transaction-price dynamics

Timeline of the transaction process for Warsaw



Micro-level dataset for two Polish cities

Micro-level dataset for two Polish cities

Use a unique property database with 16 years of micro-level transaction and list price data for two major Polish cities

Micro-level dataset for two Polish cities

Use a unique property database with 16 years of micro-level transaction and list price data for two major Polish cities

Transaction data obtained from a range of official sources. Each record provides detailed property characteristics

Micro-level dataset for two Polish cities

Use a unique property database with 16 years of micro-level transaction and list price data for two major Polish cities

Transaction data obtained from a range of official sources. Each record provides detailed property characteristics

Listing data collected through web scraping from major real estate portals. Retain only final observed price per property to align closer with transaction prices and reduce bias

Micro-level dataset for two Polish cities

Use a unique property database with 16 years of micro-level transaction and list price data for two major Polish cities

Transaction data obtained from a range of official sources. Each record provides detailed property characteristics

Listing data collected through web scraping from major real estate portals. Retain only final observed price per property to align closer with transaction prices and reduce bias

Without adjustment this would result in over-representation of expensive or atypical properties in the list-price dataset

Summary of micro-dataset before and after cleaning

	Listings		Transactions	
	Raw	Cleaned	Raw	Cleaned
Warsaw				
Mean price (per m ²)	10,675	10,446	9,564	9,832
Mean area (m ²)	62	60	54	53
Mean age (years)	31	32	34	34
Observations	1,674,796	760,273	162,015	154,729
Poznan				
Mean price (per m ²)	6,404	6,935	6,239	6,264
Mean area (m ²)	58	57	51	51
Mean age (years)	34	36	41	41
Observations	338,164	133,026	50,891	44,384

Constructing hedonic price indices

Constructing hedonic price indices

Hedonic methods preferred approach among statistical agencies for constructing quality-adjusted residential property price indices

Constructing hedonic price indices

Hedonic methods preferred approach among statistical agencies for constructing quality-adjusted residential property price indices

Use hedonic rolling-time-dummy (RTD) method to compile residential property price indices for Warsaw and Poznan

Constructing hedonic price indices

Hedonic methods preferred approach among statistical agencies for constructing quality-adjusted residential property price indices

Use hedonic rolling-time-dummy (RTD) method to compile residential property price indices for Warsaw and Poznan

RTD hedonic models incorporate property characteristics and time dummies for consecutive period with a fixed window length (both one year)

Constructing hedonic price indices

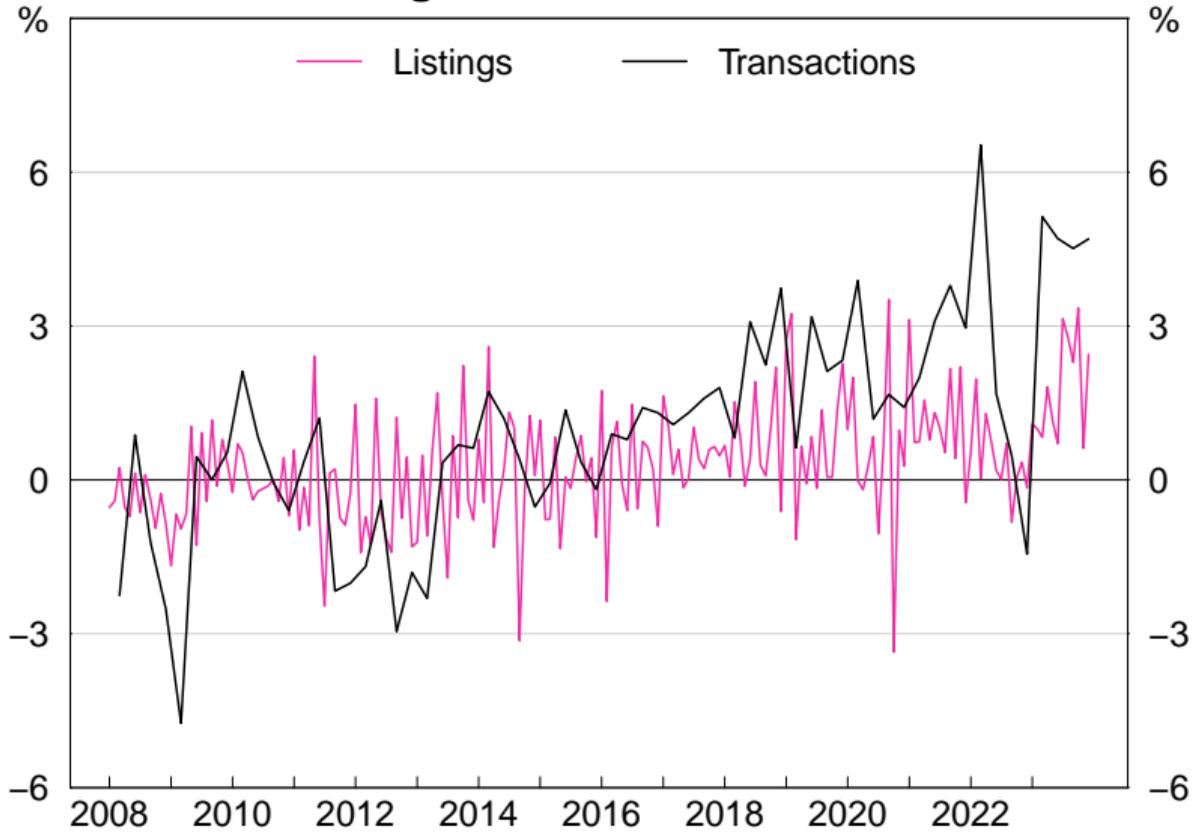
Hedonic methods preferred approach among statistical agencies for constructing quality-adjusted residential property price indices

Use hedonic rolling-time-dummy (RTD) method to compile residential property price indices for Warsaw and Poznan

RTD hedonic models incorporate property characteristics and time dummies for consecutive period with a fixed window length (both one year)

When new data is available, window is rolled forward and hedonic model is re-estimated

Warsaw – Listings and Transactions Price Growth



Modelling mixed frequency time series

Modelling mixed frequency time series

Traditionally employ temporal aggregation

Modelling mixed frequency time series

Traditionally employ temporal aggregation

- e.g. create quarterly series by taking three month average or last month in quarter

Modelling mixed frequency time series

Traditionally employ temporal aggregation

- e.g. create quarterly series by taking three month average or last month in quarter

Leads to potential loss of high frequency information

Modelling mixed frequency time series

Traditionally employ temporal aggregation

- e.g. create quarterly series by taking three month average or last month in quarter

Leads to potential loss of high frequency information

Alternative is **MI**xed **DA**ta **S**ampling (MIDAS) regression

Modelling mixed frequency time series

Traditionally employ temporal aggregation

- e.g. create quarterly series by taking three month average or last month in quarter

Leads to potential loss of high frequency information

Alternative is **MI**xed **DA**ta **S**ampling (MIDAS) regression

- Exploits *all* high frequency information in highly parsimonious way

Modelling mixed frequency time series

Traditionally employ temporal aggregation

- e.g. create quarterly series by taking three month average or last month in quarter

Leads to potential loss of high frequency information

Alternative is **MI**xed **DA**ta **S**ampling (MIDAS) regression

- Exploits *all* high frequency information in highly parsimonious way

State-space models also feasible, but MIDAS models shown to perform as well in forecasting and much easier to implement

MIDAS regression introduction

MIDAS regression introduction

Allows dependent and explanatory variables of time series regressions to be sampled at different frequencies:

MIDAS regression introduction

Allows dependent and explanatory variables of time series regressions to be sampled at different frequencies:

$$y_t = \beta_0 + \alpha(L)y_t + \beta_1 \mathcal{W}(L^{1/m}; \theta) x_t^m + \varepsilon_t \quad (1)$$

where $\alpha(L)$ is a lag polynomial, $\mathcal{W}(L^{1/m}; \theta) = \sum_{k=0}^K W(k; \theta) L^{1/m}$ and $L^{1/m}$ is a lag operator such that $L^{1/m} x_t^m = x_{t-1/m}^m$ with m indicating the higher sampling frequency

MIDAS regression introduction

Allows dependent and explanatory variables of time series regressions to be sampled at different frequencies:

$$y_t = \beta_0 + \alpha(L)y_t + \beta_1 \mathcal{W}(L^{1/m}; \theta) x_t^m + \varepsilon_t \quad (1)$$

where $\alpha(L)$ is a lag polynomial, $\mathcal{W}(L^{1/m}; \theta) = \sum_{k=0}^K W(k; \theta) L^{1/m}$ and $L^{1/m}$ is a lag operator such that $L^{1/m} x_t^m = x_{t-1/m}^m$ with m indicating the higher sampling frequency

Number of lags can be significant (i.e. monthly y_t and daily x_t) so \mathcal{W} represents a set of weights as a function of a low dimensional vector of j parameters θ ($j \ll K$)

MIDAS regression introduction

Allows dependent and explanatory variables of time series regressions to be sampled at different frequencies:

$$y_t = \beta_0 + \alpha(L)y_t + \beta_1 \mathcal{W}(L^{1/m}; \theta) x_t^m + \varepsilon_t \quad (1)$$

where $\alpha(L)$ is a lag polynomial, $\mathcal{W}(L^{1/m}; \theta) = \sum_{k=0}^K W(k; \theta) L^{1/m}$ and $L^{1/m}$ is a lag operator such that $L^{1/m} x_t^m = x_{t-1/m}^m$ with m indicating the higher sampling frequency

Number of lags can be significant (i.e. monthly y_t and daily x_t) so \mathcal{W} represents a set of weights as a function of a low dimensional vector of j parameters θ ($j \ll K$)

Specifications for \mathcal{W} include: Normalised Exponential Almon and Normalised Beta

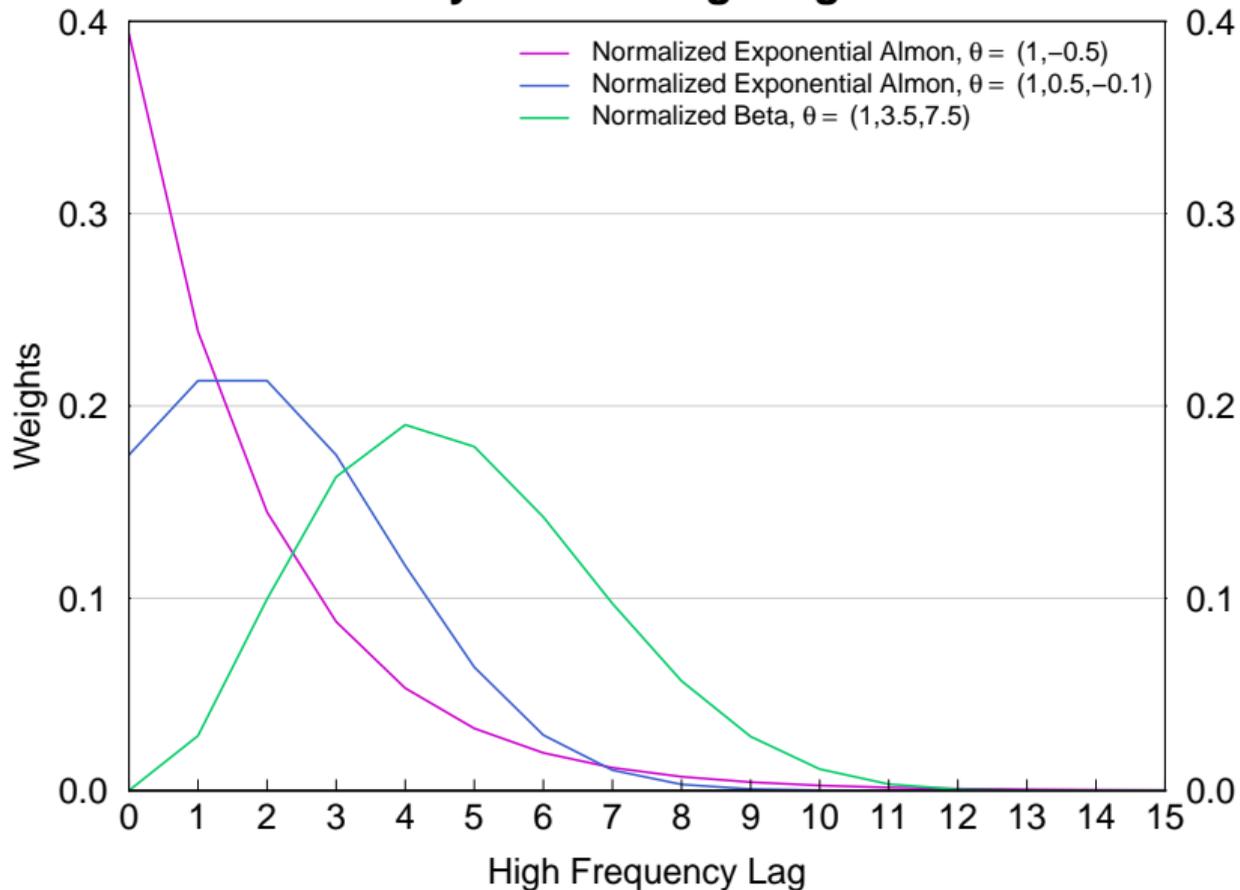
Data Frequency Alignment

Example: Simple MIDAS model assuming only monthly data in the current quarter has explanatory power (i.e. $K = 3$):

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix} = \begin{bmatrix} 1 & x_3 & x_2 & x_1 \\ 1 & x_6 & x_5 & x_4 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{3T} & x_{3T-1} & x_{3T-2} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_T \end{bmatrix}$$

Note: if y_t is of length T then x_t must be of length $m \times T$

MIDAS Polynomial Weighting Functions



Specifying MIDAS models

Specifying MIDAS models

Two questions about specifying MIDAS regression models:

Specifying MIDAS models

Two questions about specifying MIDAS regression models:

- 1) Suitable functional constraint (i.e. \mathcal{W})

Specifying MIDAS models

Two questions about specifying MIDAS regression models:

- 1) Suitable functional constraint (i.e. \mathcal{W})
- 2) Appropriate maximum lag order (i.e. K)

Specifying MIDAS models

Two questions about specifying MIDAS regression models:

- 1) Suitable functional constraint (i.e. \mathcal{W})
- 2) Appropriate maximum lag order (i.e. K)

Can address both issues together using an information criterion, such as BIC, to select best model in terms of parameter restriction and lag order

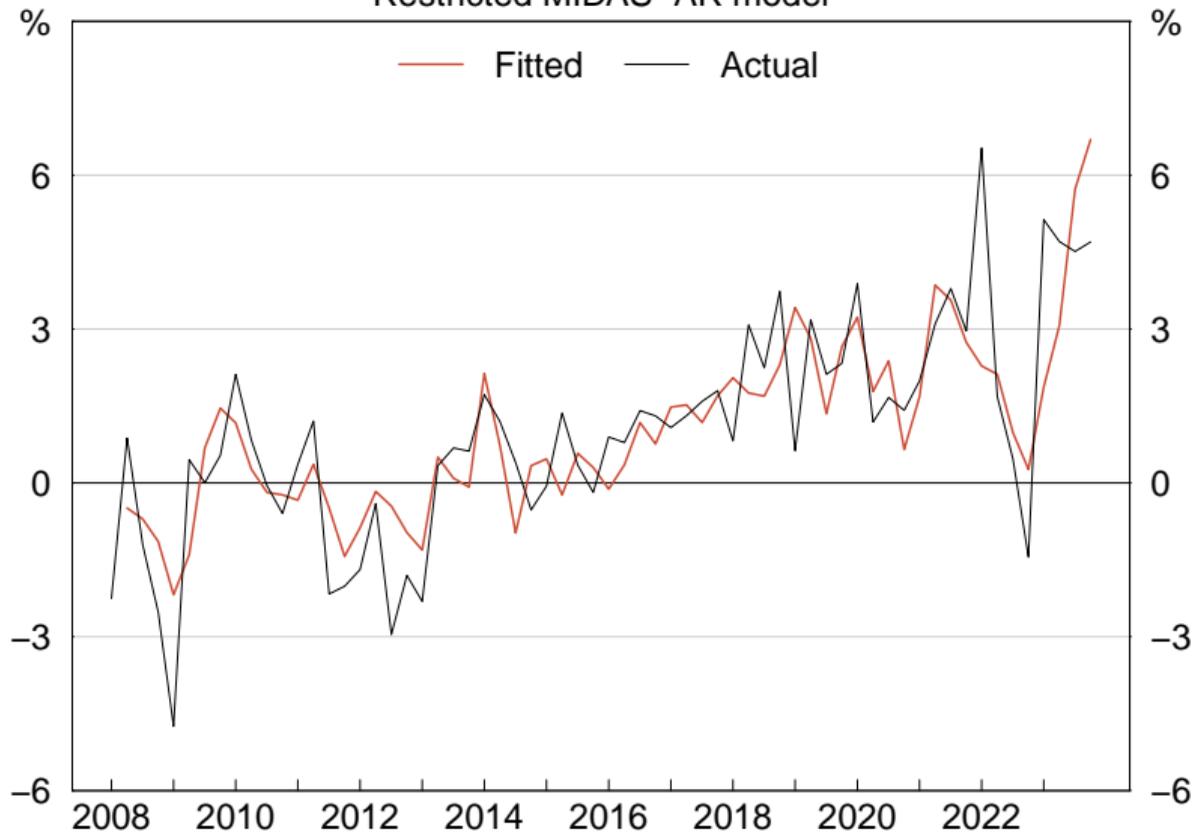
Warsaw: MIDAS model comparisons

Lag K	Normalised Exponential Almon				Normalised Beta		Unrestricted MIDAS	
	$j = 2$		$j = 3$		$j = 3$			
	BIC	p -value	BIC	p -value	BIC	p -value	BIC	p -value
0:2	236.21	0.54	239.90	0.00	252.33	0.00	239.90	–
0:3	228.76	0.71	232.14	0.93	250.62	0.00	236.28	–
0:4	231.93	0.28	232.13	0.66	243.60	0.00	239.82	–
0:5	<u>225.89</u>	<u>0.25</u>	227.78	0.27	241.91	0.00	238.03	–

Notes: p -value is for test of null hypothesis that the restrictions on the MIDAS regression coefficients implied by the polynomial weighting function are valid. Failure to reject the null implies the functional restrictions are supported by the data. Bold values denote best model per lag. A bold and underline value denote best overall model.

Warsaw – Transacted Prices and Fitted Prices

Restricted MIDAS-AR model



Nowcasting transaction prices using list prices

Nowcasting transaction prices using list prices

Only consider current quarter predictions of quarterly transaction price growth

Nowcasting transaction prices using list prices

Only consider current quarter predictions of quarterly transaction price growth

For each quarter can produce **four predictions** of current quarter transaction price growth as new monthly list-price series are produced:

Nowcasting transaction prices using list prices

Only consider current quarter predictions of quarterly transaction price growth

For each quarter can produce **four predictions** of current quarter transaction price growth as new monthly list-price series are produced:

- 1) **FC** in $t - 2/3$ (i.e. month 1 in quarter t) includes data up to $t - 1$ (last month of previous quarter)

Nowcasting transaction prices using list prices

Only consider current quarter predictions of quarterly transaction price growth

For each quarter can produce **four predictions** of current quarter transaction price growth as new monthly list-price series are produced:

- 1) **FC** in $t - 2/3$ (i.e. month 1 in quarter t) includes data up to $t - 1$ (last month of previous quarter)
- 2) **M1** in $t - 1/3$ (month 2) includes data up to $t - 2/3$

Nowcasting transaction prices using list prices

Only consider current quarter predictions of quarterly transaction price growth

For each quarter can produce **four predictions** of current quarter transaction price growth as new monthly list-price series are produced:

- 1) **FC** in $t - 2/3$ (i.e. month 1 in quarter t) includes data up to $t - 1$ (last month of previous quarter)
- 2) **M1** in $t - 1/3$ (month 2) includes data up to $t - 2/3$
- 3) **M2** in t (month 3) includes data up to $t - 1/3$

Nowcasting transaction prices using list prices

Only consider current quarter predictions of quarterly transaction price growth

For each quarter can produce **four predictions** of current quarter transaction price growth as new monthly list-price series are produced:

- 1) **FC** in $t - 2/3$ (i.e. month 1 in quarter t) includes data up to $t - 1$ (last month of previous quarter)
- 2) **M1** in $t - 1/3$ (month 2) includes data up to $t - 2/3$
- 3) **M2** in t (month 3) includes data up to $t - 1/3$
- 4) **M3** in $t + 1/3$ (month 1 in quarter $t + 1$) includes data up to t

Nowcasting transaction prices using list prices

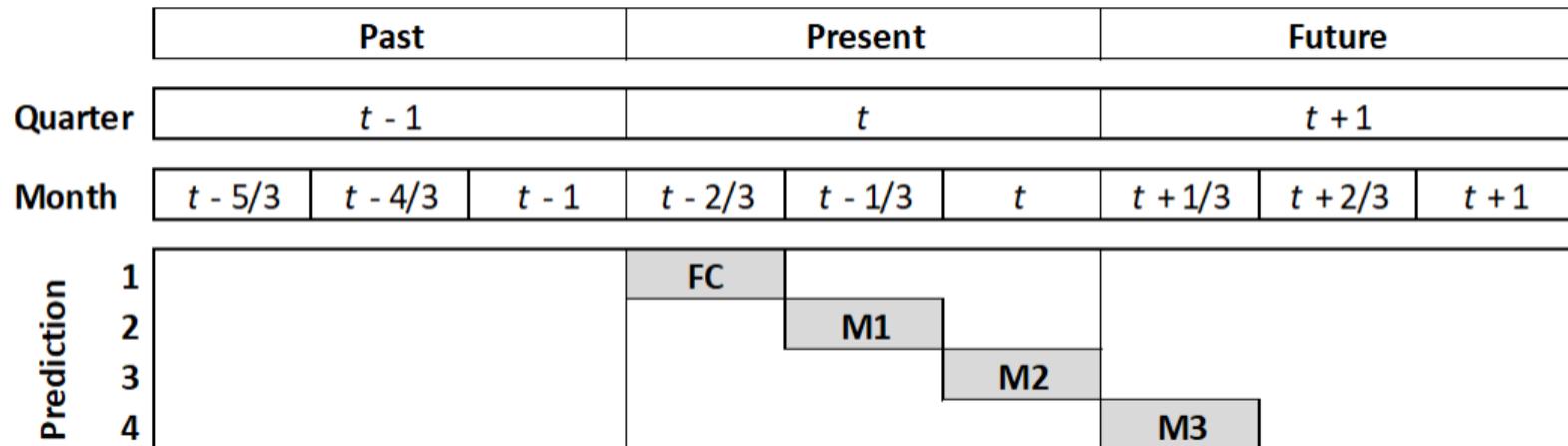
Only consider current quarter predictions of quarterly transaction price growth

For each quarter can produce **four predictions** of current quarter transaction price growth as new monthly list-price series are produced:

- 1) **FC** in $t - 2/3$ (i.e. month 1 in quarter t) includes data up to $t - 1$ (last month of previous quarter)
- 2) **M1** in $t - 1/3$ (month 2) includes data up to $t - 2/3$
- 3) **M2** in t (month 3) includes data up to $t - 1/3$
- 4) **M3** in $t + 1/3$ (month 1 in quarter $t + 1$) includes data up to t

R-MIDAS models differ in number of coefficients to estimate for each prediction but include a lag of y_t and use the Normalised Exponential Almon weighting function

Timeline of nowcasting transaction-price growth



Out-of-sample prediction exercise

Out-of-sample prediction exercise

Compare four versions of R-MIDAS model plus a 'quarter-average' version (QA) against traditional ARMA models

Out-of-sample prediction exercise

Compare four versions of R-MIDAS model plus a 'quarter-average' version (QA) against traditional ARMA models

Goal: Predict current quarter transaction price growth

Out-of-sample prediction exercise

Compare four versions of R-MIDAS model plus a 'quarter-average' version (QA) against traditional ARMA models

Goal: Predict current quarter transaction price growth

- Rolling window (i.e. fixed 8-year window) and recursive estimation (i.e. expanding window)

Out-of-sample prediction exercise

Compare four versions of R-MIDAS model plus a 'quarter-average' version (QA) against traditional ARMA models

Goal: Predict current quarter transaction price growth

- Rolling window (i.e. fixed 8-year window) and recursive estimation (i.e. expanding window)
- Training sample: 2008:Q1–2015:Q4

Out-of-sample prediction exercise

Compare four versions of R-MIDAS model plus a 'quarter-average' version (QA) against traditional ARMA models

Goal: Predict current quarter transaction price growth

- Rolling window (i.e. fixed 8-year window) and recursive estimation (i.e. expanding window)
- Training sample: 2008:Q1–2015:Q4
- Evaluation sample: 2016:Q1–2023:Q4

Comparing prediction accuracy

Comparing prediction accuracy

Test statistical significance of predictive accuracy of competing models via Model Confidence Set (MCS) procedure

Comparing prediction accuracy

Test statistical significance of predictive accuracy of competing models via Model Confidence Set (MCS) procedure

MCS identifies models whose predictive accuracy is statistically indistinguishable at a given level of significance and specified loss criterion

Comparing prediction accuracy

Test statistical significance of predictive accuracy of competing models via Model Confidence Set (MCS) procedure

MCS identifies models whose predictive accuracy is statistically indistinguishable at a given level of significance and specified loss criterion

Provides a statistically grounded way to identify set of best models from a candidate pool, rather than picking one by many pairwise tests

Comparing prediction accuracy

Test statistical significance of predictive accuracy of competing models via Model Confidence Set (MCS) procedure

MCS identifies models whose predictive accuracy is statistically indistinguishable at a given level of significance and specified loss criterion

Provides a statistically grounded way to identify set of best models from a candidate pool, rather than picking one by many pairwise tests

Use range statistic (T_R), with $\alpha = 10\%$, and compute p -value via a stationary bootstrap (bloc length 8 quarters and 5,000 replications)

Results: Warsaw – rolling window

	AR(1)	MA(1)	AR(2)	ARMA(1,1)	FC	M1	M2	M3	QA
Past 3-years									
RMSE	2.51	2.47	2.72	2.68	2.33	2.19	2.75	2.76	2.68
(90%) \mathcal{M}						*			
p -value	0.00	0.04	0.00	0.00	0.04	1.00	0.00	0.00	0.00
Past 5-years									
RMSE	2.18	2.18	2.27	2.25	2.13	1.90	2.26	2.27	2.19
(90%) \mathcal{M}						*			
p -value	0.05	0.05	0.05	0.05	0.05	1.00	0.05	0.05	0.05
Full sample									
RMSE	1.87	1.96	1.92	1.90	1.82	1.57	1.95	1.91	1.89
(90%) \mathcal{M}						*			
p -value	0.01	0.01	0.01	0.01	0.01	1.00	0.01	0.01	0.01

Notes: Rolling window estimation begins in 2016:Q1 with window length of 32 quarters. Full sample: 2016:Q1–2023:Q1. Bold values denote best model(s) for each horizon

Results: Warsaw – recursive

	AR(1)	MA(1)	AR(2)	ARMA(1,1)	FC	M1	M2	M3	QA
Past 3-years									
RMSE	2.60	2.86	2.63	2.64	2.50	2.22	2.00	2.00	1.99
(90%) \mathcal{M}							*	*	*
p -value	0.01	0.00	0.01	0.01	0.01	0.01	0.21	0.21	1.00
Past 5-years									
RMSE	2.25	2.50	2.22	2.22	2.16	1.98	2.12	1.80	1.80
(90%) \mathcal{M}	*		*	*	*	*	*	*	*
p -value	0.14	0.00	0.14	0.14	0.14	0.19	0.14	0.76	1.00
Full sample									
RMSE	1.96	2.22	1.92	1.92	1.82	1.62	1.74	1.55	1.51
(90%) \mathcal{M}			*	*	*	*	*	*	*
p -value	0.03	0.00	0.26	0.27	0.27	0.48	0.31	0.57	1.00

Notes: Recursive estimation begins in 2016:Q1 with initial sample length of 32 quarters. Full sample: 2016:Q1–2023:Q4. Bold values denote best model(s) for each horizon

Conclusion

Conclusion

Transaction-based indices regarded as most reliable indicator of housing market trends but recorded with a delay creating a challenge for policy makers

Conclusion

Transaction-based indices regarded as most reliable indicator of housing market trends but recorded with a delay creating a challenge for policy makers

Drawing on 16 years of micro-level data from Warsaw and Poznan construct quality-adjusted list and transaction price indices using hedonic rolling time dummy method

Conclusion

Transaction-based indices regarded as most reliable indicator of housing market trends but recorded with a delay creating a challenge for policy makers

Drawing on 16 years of micro-level data from Warsaw and Poznan construct quality-adjusted list and transaction price indices using hedonic rolling time dummy method

Employ MIDAS regression to nowcast quarterly transaction-price growth using monthly list-price indices and achieve more accurate predictions than traditional ARMA models

Conclusion

Transaction-based indices regarded as most reliable indicator of housing market trends but recorded with a delay creating a challenge for policy makers

Drawing on 16 years of micro-level data from Warsaw and Poznan construct quality-adjusted list and transaction price indices using hedonic rolling time dummy method

Employ MIDAS regression to nowcast quarterly transaction-price growth using monthly list-price indices and achieve more accurate predictions than traditional ARMA models

Results confirm list-price indices provide a timely indication to future movements in transaction-price indices

Spares

Constructing hedonic price indices – details

Assuming the first period in the window is t , the semi-log hedonic model is:

$$\ln(p_{\tau n}) = \sum_{c=1}^C \beta_c z_{\tau n} + \sum_{s=t+1}^{t+m} \delta_s d_{\tau sn} + \varepsilon_{\tau n} \quad (2)$$

where $p_{\tau n}$ is price of property n in period τ , $z_{\tau n}$ are the property characteristics while $d_{\tau sn}$ is a dummy variable equal to 1 when $\tau = s$ and 0 otherwise

From (2) obtain the price index as:

$$\frac{P_{t+m}}{P_{t+m-1}} = \frac{\exp(\hat{\delta}_{t+m}^t)}{\exp(\hat{\delta}_{t+m-1}^t)} \quad (3)$$

where P_{t+m-1} and P_{t+m} denote the level of the price index in periods $t+m-1$ and $t+m$ while $\hat{\delta}_{t+m-1}^t$ and $\hat{\delta}_{t+m}^t$ are the estimated coefficients of the last two time dummies

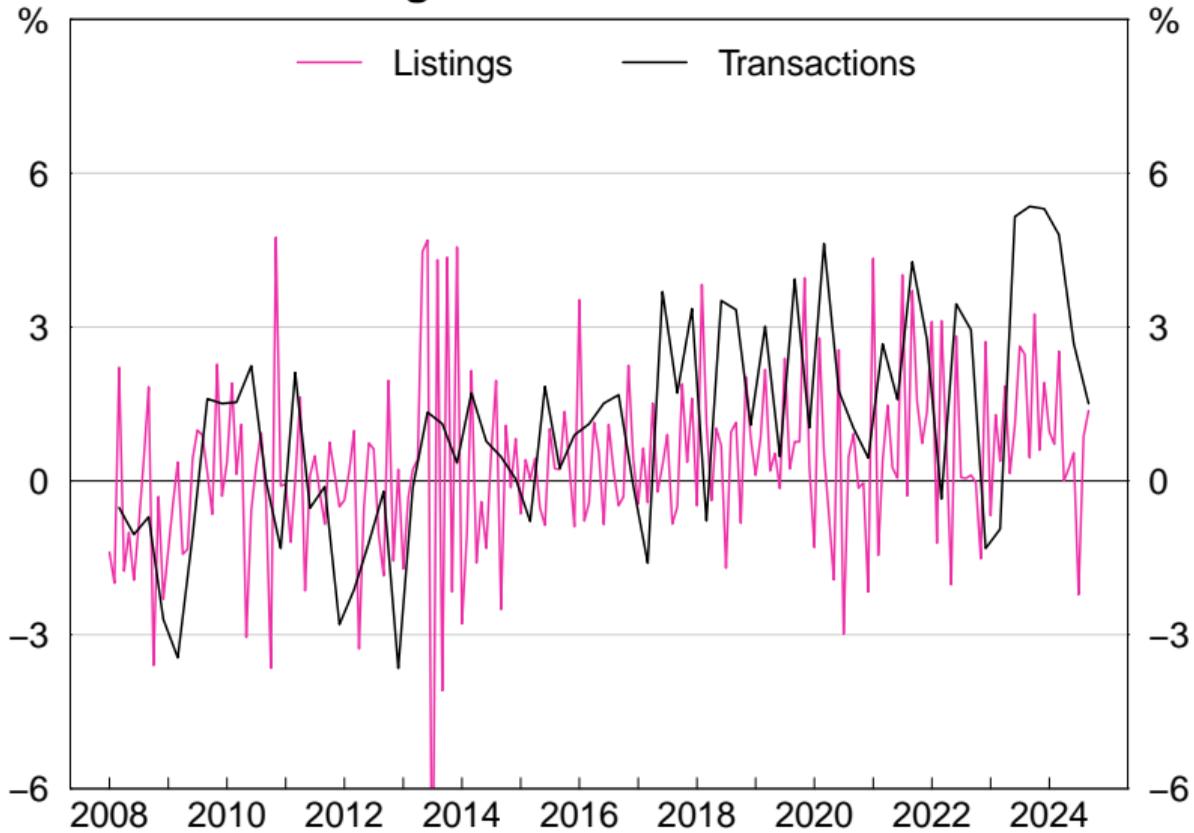
Constructing hedonic price indices – details

Constructing hedonic price indices – details

Overall price index starting in $t = 1$ is calculated by chaining all the pairwise price indices together as follows:

$$\frac{P_{t+m}}{P_1} = \prod_{\tau=1}^{t-m} \left[\frac{\exp(\hat{\delta}_{\tau+m}^{\tau})}{\exp(\hat{\delta}_{\tau+m-1}^{\tau})} \right] \quad (4)$$

Poznan – Listings and Transactions Price Growth



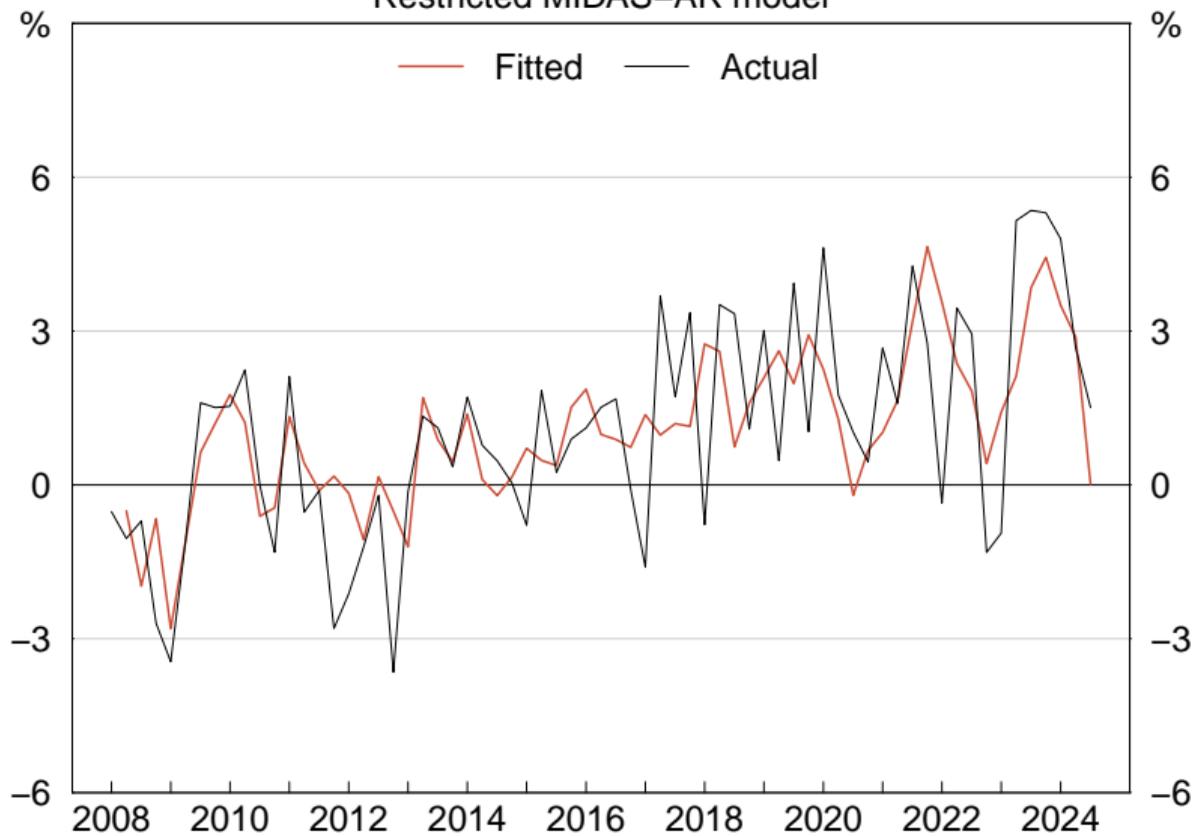
Poznan: MIDAS model comparisons

Lag K	Normalised Exponential Almon				Normalised Beta		Unrestricted MIDAS	
	$j = 2$		$j = 3$		$j = 3$			
	BIC	p -value	BIC	p -value	BIC	p -value	BIC	p -value
0:2	285.50	0.87	289.66	0.00	293.32	0.00	289.66	–
0:3	277.10	0.54	280.35	0.54	282.53	0.11	284.28	–
0:4	277.10	0.01	292.16	0.00	275.13	0.00	276.86	–
0:5	274.11	0.00	268.68	0.64	269.18	0.34	280.04	–

Note: p -value is for test of null hypothesis that the restrictions on the MIDAS regression coefficients implied by the polynomial weighting function are valid. Failure to reject the null implies the functional restrictions are supported by the data. Bold values denote best model per lag. A bold and underline value denote best overall model.

Poznan – Transacted Prices and Fitted Prices

Restricted MIDAS-AR model



Results: Poznan – rolling window

	AR(1)	MA(1)	AR(2)	ARMA(1,1)	FC	M1	M2	M3	QA
Past 3-years									
RMSE	2.61	2.63	2.84	2.69	2.63	2.43	2.52	2.74	2.68
(90%) \mathcal{M}	*	*				*	*		*
p -value	0.21	0.27	0.01	0.03	0.09	1.00	0.27	0.09	0.27
Past 5-years									
RMSE	2.33	2.35	2.42	2.34	2.30	2.09	2.24	2.50	2.34
(90%) \mathcal{M}						*			
p -value	0.05	0.03	0.03	0.03	0.03	1.00	0.05	0.00	0.05
Full sample									
RMSE	2.33	2.34	2.34	2.27	2.29	2.20	2.33	2.47	2.33
(90%) \mathcal{M}	*	*	*	*	*	*	*		*
p -value	0.55	0.24	0.38	0.55	0.55	1.00	0.24	0.04	0.24

Notes: Rolling window estimation using the extended dataset begins in 2016:Q3 with window length of 34 quarters. Full sample: 2016:Q3–2024:Q3. Bold values denote best model(s) for each horizon

Results: Poznan – recursive

	AR(1)	MA(1)	AR(2)	ARMA(1,1)	FC	M1	M2	M3	QA
Past 3-years									
RMSE	2.45	2.54	2.67	2.58	2.28	2.13	2.20	2.03	2.08
(90%) \mathcal{M}	*	*		*	*	*	*	*	*
p -value	0.29	0.29	0.05	0.16	0.29	0.64	0.46	1.00	0.64
Past 5-years									
RMSE	2.28	2.38	2.34	2.29	2.11	1.88	2.05	1.89	1.89
(90%) \mathcal{M}			*	*	*	*	*	*	*
p -value	0.05	0.06	0.16	0.16	0.16	1.00	0.16	0.90	0.90
Full sample									
RMSE	2.32	2.40	2.32	2.29	2.19	1.98	2.21	2.09	2.03
(90%) \mathcal{M}			*	*	*	*	*	*	*
p -value	0.03	0.02	0.15	0.15	0.15	1.00	0.15	0.56	0.56

Notes: Recursive estimation using the extended dataset begins in 2016:Q3 with initial sample length of 34 quarters. Full sample: 2016:Q3–2024:Q3. Bold values denote best model(s) for each horizon