# Forecasting the Growth rate in Residential housing across Danish Municapalities

Asger Ingemann Rasmussen

# Introduction

## General intro

Real estate has long been on average the asset that accounts for the largest percent of individual Danes net wealth, accounting for roughly 54% in 2021.[1] Fueled by *"The Danish Mortgage Model"*, seen as one of most effective and secure models for financing residential properties in the world, amounting to approximately 402bn EUR, making it the largest mortgage bond market in Europe. Of the adult population 58% are home owners at one time in their life, understanding and predicting growth in the house market is not just important for the individual Dane, but the overall economy.

## Research question

The goal of this paper is to forecast real growth in house prices for municipalities in Denmark. The primary motivation for this study is the National Bankens paper from 2021 *"Housing Market Robustness Should be Strengthened"*, highlighting the importance of forecasting the growth in house prices. The paper highlights how the Covid-19 pandemic lead to rapid demand increases for housing, despite the general economic downturn resulting in large increases in property prices across Denmark. Specifically the urban regions such as Copenhagen and Aarhus experienced significant growth in both property prices, and lending activity. This highlights the regional disparities observed between urban and rural municipalities, in the case of property prices. In general these rapid price increases outpaces the growth rate of the income level for the individual Dane, raising concerns for the stability of the housing market, as households take on larger level of debt to finance their purchase of property alongside an increased speculative behavior by investors and individuals at the prospect of future growth in the property market. National Banken highlights these factors, as the driving force for the increased the possibility of a strong market correction in the property market.

## How?

Specifically this study seeks to forecast growth in housing prices using a simple Autoregressive (AR) model benchmark alongside an Autoregressive distributed (ARDL) lag model using Economic variables motivated by historical evidence, against a modern machine learning approach using the framework of XGBoost, which has been one of the best performing machine learning methods for forecasting house prices. Sharma, Harsora, and Ogunleye [9]

This study does not seek to forecast the property market as a whole, but selectively the residential property market of single family homes. The property market is characterized by large heterogeneity, as properties are valued based on location and its physical properties which differs significantly both between regions and within regions. Because of the infrequent nature of property sales it is not feasible to obtain prices on individual properties, instead this study focuses on the aggregate *sqm* prices on a municipality level. Denmark consits of 98 different municipalities, over the last 12 years 68 municipalities have seen an increase in their population while 30 have seen a decrease, the general trend is that municipalities within the capital area has seen the largest growth in population, while municipalities in Jutland and Fyn has experienced smaller or negative growth. A working paper by the National Bank in 2017 with the goal of understanding the regional model for the danish housing market, found that regional fundamentals are the dominant determinant for real estate prices in the long run, where as the short term is mostly explained by the "ripple effect" where a increase in one region impacts prices in neighboring regions, the goal of this paper is to forecast the long-term growth in house prices, therefor cross

[1] Statistics Denmark – StatBank.dk/bef1

regional impact is omitted, and the focus will be on disparities in the regional fundamentals between municipalities.

## Data

The data is primarily sourced by statbanken.dk, managed by Statistics Denmark a government agency responsible for compiling official statics in Denmark. Other sources consist of Nationalbanken, Kommunale nøgletal, boligstatistiken. Data has been transformed into a singular dataset by the researcher containing 23 variables spanning from 1993 - 2024.

## What other people did

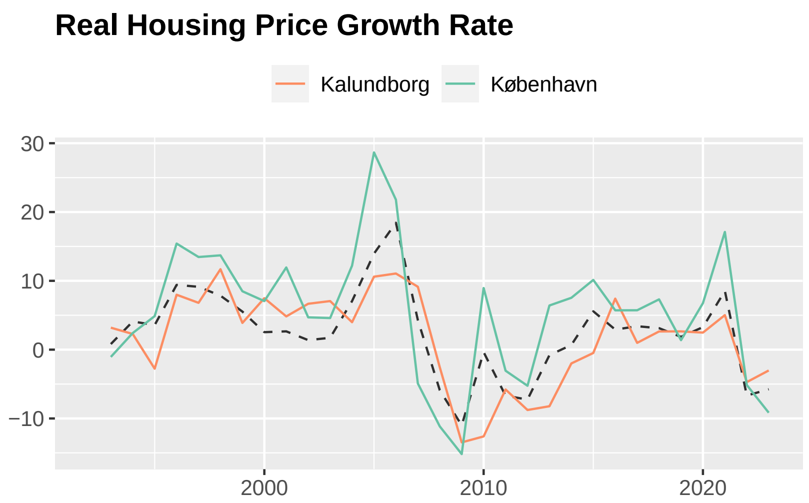
Past literature has found compelling evidence that economic variables are associated with growth in house prices. Historic papers such as (1) rent prices, construction cost, population and housing starts are significant for explaining excess returns in real estate. Abraham and Hendershott [1] and (**dipasqualewheaton**) find similar findings. Historicaly models have been split into structural and non structural models, a quick overview of notable papers on the topic of forecasting house prices.

1. **dipasqualewheaton** work within the framework of the dynamic Gordon growth treating real estate as a purely financial asset akin to the stock market. They make use of a bayesian VAR model to forecast the rent-price ratio and housing premium, finding substantial evidence for their respective predictability.
2. (1) forecast real estate prices and excess returns, make use of a structural model looking at theoretical economic variables and their effect on prices, their simple model has laid much of the groundwork for future work.
3. (2) expands on (Case, Shiller) structural model by forecasting across US states, making use of an ARDL framework alongside combination forecast. The ARDL model presented in this paper is build up on the model presented in this paper.

# Data set

## House Prices

Regional house prices are obtained from “Finans Danmark”, consist of quarterly observations for 1992Q1:2023Q4 prices are collected from the Real-ViewTM database, which purpose is to document market price from realized sales of properties on the open market by professional brokers. Prices are calculated as *sqm* for the individual property, for the portal district and the municipality as a whole, using a summation of the individual properties *sqm* price divided by the number of sales in that region. Prices are calculated four times a year to ensure the reliability and robustness of the reported data. The nomial *sqm* prices are then converted into real terms, using the (CPI) deflator from Danmark statistik, with 2015 serving as the base year. The computed quarterly growth rates are calculated as the differences in the log levels of real housing prices for each region. The quarterly real housing price growth rates are plotted in Figure 1.

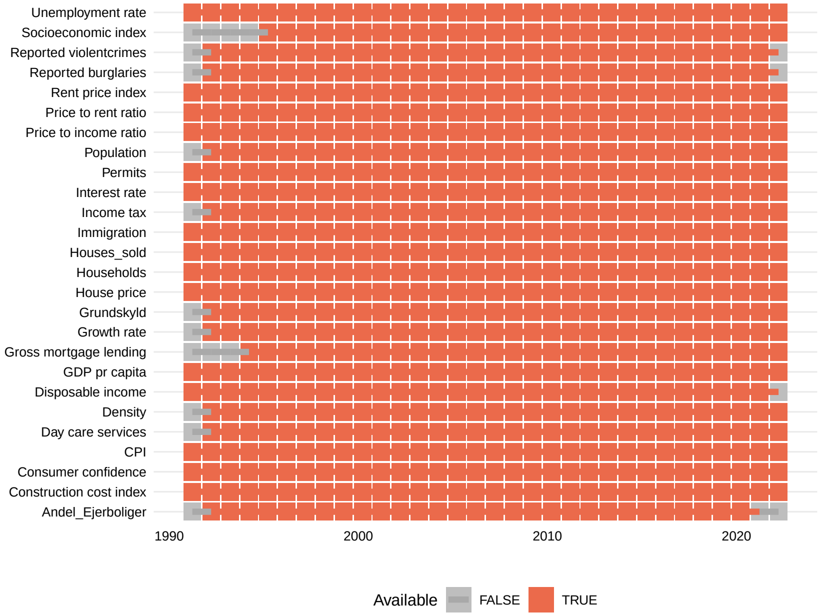


The blacked dashed line indicates the average real growth rates for house prices in all municipalities, this seems to indicate that housing price changes in locations are subject to similar market forces and economic conditions, while still maintaining heterogeneity between them.

## Variables

For our structural forecast I consider 19 potential predictors of real housing price growth for each municipality, these predictors have been transformed by differentiating and/or deflating them by the (CPI) when deemed appropriate in an effort to make them stationary. The predictors are split up into national level economic variables or municipality level economic variables.

Figure 1: Data availability for municipalities



Out of the nineteen potential predictors nine of them are **national level** economic variables:

- Real Interest rate (*Nationalbank offical lending rate by end of month, discounted by CPI and summarized quarterly, diff*)
- Consumer confidence (*Statisk Denmark monthly survey comprising the perceived economic situation, summarized quarterly, diff)*
- Construction cost index *(Statisk Denmark construction cost index for residential buildings total, quarterly, diff)*
- Employment rate *(FRED Employment rate aged 15-64 all persons for Denmark, yearly, diff)*
- GDP Pr. capita *(Statisk Denmark in 1000 DKK current prices, yearly, diff)*
- Real Gross Lending *(Nationalbank Gross lending by mortgage banks for all property types deflated by CPI in bn DKK, quarterly, diff)*
- Price to Income ratio *(OECD Analytical house prices indicators for Denmark, seasonal adjusted, quarterly, diff)*
- Price to Rent ratio *(OECD Analytical house prices indicators for Denmark, seasonal adjusted, quarterly, diff)*
- Rent price index *(OECD Analytical house prices indicators for Denmark, seasonal adjusted, quarterly, diff)*

Real interest rate serves as an indication for cost of borrowing, as the long term and short term mortgage rents follow it closely, this has been deflated by CPI. Consumer confidence is an indication of the general economic perspective in the country, and include measures of intention of pruchasing or building a home. Including real gross lending they serve as way to predict the demand for housing. The construction cost index purpose is reflect the cost of housing construction in Denmark, used primarily by construction organizations and housing developers, alongside Price to Rent ratio they serve as an indication for the supply side of housing in Denmark. Employment rate and GDP Pr. Capita is general macro economic indicators for the danish economy.

The remaining ten variables are **municipality level** economic variables:

- Disposable income *(Statisk Denmark disposable income for municipalities, log, yearly, diff)*
- Land tax *(Municipality keynumbers tax rate on land, yearly)*
- Income tax rate *(Municipality keynumbers income tax rate for municipality, yearly)*
- Immigration *(Municipality keynumbers number of immigrants log, yearly)*
- Population *(Municipality keynumbers population log, yearly, diff)*
- Reported violent crimes *(Municipality keynumbers population, yearly)*
- Reported burglaries *(Municipality keynumbers population, yearly)*
- Density *(Municipality keynumbers population, yearly)*
- Permits *(Denmark Statistik new housing permits, yearly)*

- Socioeconomic index *(Municipality keynumbers population log, yearly, diff)*

Numerical disposable income which has been deflated the CPI to be Real Disposable income and population approximate demand for housing within municipalities. To account for supply of housing I make use of permits for new residential housing within region. Land tax, income tax rate, density, socioeconomic index and reported crimes accounts for the general economic structure within municipalities to account for the heterogeneity between them. Figure 2 shows data availability for all listed variables for a single municipality, two variables raise concerns namely socioeconomic index and employment rate.

# Method

## Expanding window

The dataset on real growth in house prices spans from 1992Q1 to 2023Q4 amounting to 124 individual observations, a fairly modest number of observations which spans across three decades.

Assessing the accuracy of our models require us to split our data into an initial training period for training our model, and a testing period to test the accuracy of our model, on data is has not seen. For this purpose I will make use of a expanding window, because of the long run dynamics of prices on real estate and the modest number of observation present, this is favorable over a rolling window, as it allows for the full data to be used accounting for long term trends in the housing market. A expanding window works by choosing an initial training period, for this paper 20 was chosen accounting for five years of data, after training the model on the initial period a one step ahead forecast is produced. Expanding the window by the next observation a new model is trained, and a new one step ahead forecast is produced, this patterns continuous until the last observation is reached, in our case this would be 2023Q4. This method is done for all produced models to assure the best model is chosen.

## RMSE

Historically the RMSFE has been used to evaluate proposed model for forecasting house prices, this paper will use a more general metric namely RMSE. RMSE is well known and popular choice for its symmetrical weighting scheme and intrepretable results. The reported RMSE is the average RMSE for all individual municipalities, this might not guarantee that the best model for each municipality is found, but rather the model that on average is best at forecasting for all individual municipality.

## ARIMA

ARIMA (AutoRegressive Integrated Moving Average) is a widely popular model for forecasting time series, gaining widespread adoption through the seminal work of Box and Jenkins. ARIMA models are made up of two components.

$$y_t = \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q}.$$

The first, an AutoRegressive (AR) component $y_t = \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p}$ , employs lagged iterations of the dependent variable, leveraging historical data to address and account for serial correlation in the stochastic process. The Moving Average (MA) component $y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q}$, incorporates past error terms to predict the variable of interest. ARIMA models can be extended to include seasonal components such as trend and cyclical behavior.

## ARDL

Autoregressive distributed lag model (ARDL), are a popular model choice in the framework of analyzing economic growth, such as Foreign direct investments Belloumi [2], immigration Morley [5] and Real housing price growth Rapach and Strauss [8]. Like ARIMA models it is encompasses two components, an Autoregressive component such as the one mentioned above in the ARIMA model, and a Distributed lag (DL) component. Traditionaly a distributed lag model takes the form.

$$y_t = \sum_{j=-\infty}^{\infty} \beta_j x_{t-j} + \varepsilon_t$$

Where $y_t$ is our *response* time series, and $x_t$ is our *input* time series, using lag values of our *input* time series we seek to analyze or predict our *response* time series. Peng [6] This dynamic relationship between our variables, are often modeled using vector autoregreesive (VAR) and vector error-correction (VEC) models, which allows for contemporaneous feedback from a response variable to other variables. Kripfganz and Schneider [4] ARDL is a single equation model and therefor has a uni directional relationship between the *response* and *input*, meaning that the *response* variable does not contemporaneously affect our *input* variables like in the VAR & VEC models.

When modeling two distinct time series the concept of a spurious regression and co-integration plays an important role, spurious regression refers to the phenomenon where two non stationary time series display a statistical significant relationship even when no theoretical or substantial relation exist in reality. GRANGER and NEWBOLD [3] This statistical significant relationship is due to the some common underlying trend, this can be circumvented by making our time series stationary with the notion of differentiating or testing for co-integreated relationships between our time series. A co-integrated relationship refers to the case where the statistical properties of a linear combination of non stationary time series is stationary *The cointegrated VAR model : methodology and applications - Royal Danish Library* [10], in the ARDL framework co-integration is tested using a *"bounds test"* propsed by Pesaran, Shin, and Smith [7].

Following the example of David E. Rapach & Jack K. Strauss for forecasting Real growth in house prices, an individual ARDL model based on a single predictor is given as

$$y^h_{t+h} = \alpha + \sum_{j=0}^{q_1-1} \beta_j \Delta y_{t-j} + \sum_{j=0}^{q_2-1} \gamma_j x_{t-j} + \epsilon^h_{t+h}$$

Where $y^h_{t+h}$ is the growth of real housing prices from time $t$ to $t+h$. $\Delta y_{t-j}$ for period $h$ and the included lags of the growth rate in real housing prices, denoted by $j$. $x_{t-j}$ is a given predictor out of our 21 available predictors with its lags denoted by $j$, $\alpha$ is our constant and $\epsilon^h_{t+h}$ is an error term. The amount of lags to be included are chosen by the Bayesian information criteria (BIC), with the minimum value of $q_1 = 0$ and $q_2 = 1$ to ensure that our potential predictor is included. For each municipality an ARDL model is run for each individual explanatory in a single time period in our expanding window, all of these ARDL models are then combined using a combination forecast, the resulting value is the foretasted value $y^h_{t+h}$ for that municipality in that time period.

### Combination forecast

The discount mean square forecast error (DMSFE), was the best performing forecast combination in the paper by David E. Rapach & Jack K. Strauss, and will therefor be implemented in this case. The weights are given by

$$w_{i,t} = m_{i,t}^{-1} \sum_{j=1}^{n} m_{j,t}^{-1} \quad (i = 1, \ldots, n)$$

where

$$m_{i,t} = \sum_{s=-h}^{h-\theta} \theta^{-h-s} (y_{s+h} - \hat{y}_{i,s+h|s})^2$$

## XGBOOST

XGBoost, an acronym for eXtreme Gradient Boosting, marks a significant evolution in the domain of machine learning, presenting an advanced and highly efficient implementation of gradient boosting algorithms. At its core, XGBoost is engineered to optimize both computational speed and model performance, utilizing a sophisticated framework that intelligently handles large-scale data. The algorithm's prowess lies in its ability to conduct parallel tree construction, a method that significantly accelerates the learning process without compromising accuracy. Furthermore, XGBoost incorporates a novel regularization technique, which enhances the model's capability to generalize by mitigating overfitting through the penalization of complex models. This dual emphasis on speed and regularization underpins XGBoost's superiority in handling a vast spectrum of predictive modeling tasks, ranging from regression and classification to ranking. Its versatility and robustness have catapulted XGBoost to the forefront of predictive analytics, making it a preferred choice among data scientists and practitioners for tackling challenging datasets and competitions. Through its exemplary performance across numerous benchmarks and real-world applications, XGBoost has firmly established itself as a cornerstone technology in the predictive modeling landscape.

# Forecast

## ARIMA

To understand the choices of specification for the ARIMA model, I present results for a single municipality in the hopes of making it easier to follow. The first step as with most forecasts is to visualize our time series, for this we will be focusing on the municipality of Copenhagen. The goal of visualizing our time series, is to identify any clear trend or seasonality within the stochastic process.

Figure 2: Real growth rate in house prices for Copenhagen

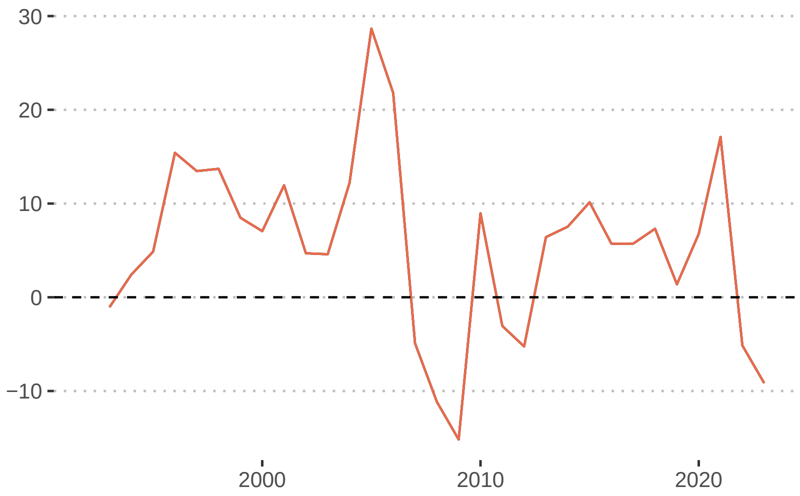


Figure 3, represent the real growth rate in house prices for the municipality of Copenhagen, the stochastic process does not visually seem to exhibit any clear long term trend or seasonality. To confirm this initial conclusion a ADF test is run containing 8 lags corresponding to 2 years, all p-values are below the threshold of 0.05 indicating that our time series is stationary.

It is not clear from the PACF and ACF which model should be chosen. For efficiency I will make use of Rob J. Hyndman fable package, containing the ARIMA function which searches through the model space to identify the best ARIMA model for each individual time series, the choice of model is based on the Information criteria BIC. The information criteria BIC was chosen over AIC and AICc as it has been showed to favor less complex models. This procedure is done over the expanding window detailed above in section “Method, Expanding Window”.

This initial search produced 73 unique model specification across over 10000 model specification. Out of these 73 unique models, the 3 most common was chosen as candidates for the ARIMA model, the chosen models are.

ARIMA(0,0,1) w/ mean, ARIMA(0,0,1) and ARIMA(1,0,0). To choose between these three models, I will run an expanding window including all three models, and make one-step ahead forecast for all periods in the testing period, the model with the lowest RMSE will be chosen as our ARIMA model.

Based on table 3, the model which achieves the lowest RMSE is an ARIMA(0,0,1) with mean, this leads me to conclude that the for comprising against ARDL and XGBoost this model will be chosen.

## ARDL

## XBGOOST

# Conclusion

# Appendix

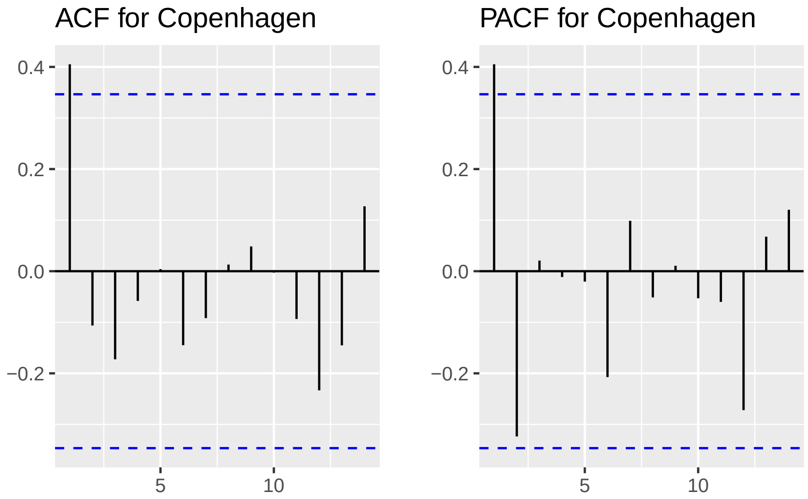
```
#| label: Appendix 1 ACF and PACF for Copenhagen

ACF <- Ejendomspriser |>
  filter(Kommune == "København") |>
  select(Growth_rate) |>
  feasts::ACF(Growth_rate) |>
  autoplot() +
  labs(
    y = "",
    x = "",
    title = "ACF for Copenhagen"
  )
```

```
Adding missing grouping variables: `Kommune`
```

```
PACF <- Ejendomspriser |>
  filter(Kommune == "København") |>
  feasts::PACF(Growth_rate) |>
  autoplot() +
  labs(
    y = "",
    x = "",
    title = "PACF for Copenhagen"
  )

grid.arrange(ACF, PACF, ncol=2)
```



# Bibliography

[1] Jesse M. Abraham and Patric H. Hendershott. “Bubbles in Metropolitan Housing Markets”. In: *Journal of Housing Research* 7.2 (1996). Publisher: American Real Estate Society, pp. 191–207. URL: http://www.jstor.org/stable/24832859.

[2] Mounir Belloumi. “The relationship between trade, FDI and economic growth in Tunisia: An application of the autoregressive distributed lag model”. In: *Economic Systems*. Symposium: Performance of Financial Markets 38.2 (June 1, 2014), pp. 269–287. DOI: 10.1016/j.ecosys.2013.09.002. URL: https://www.sciencedirect.com/science/article/pii/S093936251400003X.

[3] C. W. J. GRANGER and PAUL NEWBOLD. “CHAPTER SIX - FORECASTING FROM REGRESSION MODELS”. In: ed. by C. W. J. GRANGER and PAUL NEWBOLD. Second Edition. DOI: https://doi.org/10.1016/B978-0-12-295183-1.50012-1. Academic Press, 1986, pp. 187–215. DOI: https://doi.org/10.1016/B978-0-12-295183-1.50012-1. URL: https://www.sciencedirect.com/science/article/pii/B9780122951831500121.

[4] Sebastian Kripfganz and Daniel C. Schneider. “ardl: Estimating autoregressive distributed lag and equilibrium correction models”. In: *The Stata Journal* 23.4 (Dec. 1, 2023). Publisher: SAGE Publications, pp. 983–1019. DOI: 10.1177/1536867X231212434. URL: https://doi.org/10.1177/1536867X231212434.

[5] Bruce Morley. “Causality between economic growth and immigration: An ARDL bounds testing approach”. In: *Economics Letters* 90.1 (2006), pp. 72–76. DOI: https://doi.org/10.1016/j.econlet.2005.07.008. URL: https://www.sciencedirect.com/science/article/pii/S0165176505002594.

[6] Roger D. Peng. *4.3 Distributed Lag Models | A Very Short Course on Time Series Analysis*. URL: https://github.com/rdpeng/timeseriesbook.

[7] M. Hashem Pesaran, Yongcheol Shin, and Richard J. Smith. “Bounds Testing Approaches to the Analysis of Level Relationships”. In: *Journal of Applied Econometrics* 16.3 (2001). Publisher: Wiley, pp. 289–326. URL: http://www.jstor.org/stable/2678547.

[8] David E. Rapach and Jack K. Strauss. “Differences in housing price forecastability across US states”. In: *Forecasting Returns and Risk in Financial Markets using Linear and Nonlinear Models* 25.2 (Apr. 1, 2009), pp. 351–372. DOI: 10.1016/j.ijforecast.2009.01.009. URL: https://www.sciencedirect.com/science/article/pii/S0169207009000119.

[9] Hemlata Sharma, Hitesh Harsora, and Bayode Ogunleye. “An Optimal House Price Prediction Algorithm: XGBoost”. In: *Analytics* 3.1 (Mar. 2024). Number: 1 Publisher: Multidisciplinary Digital Publishing Institute, pp. 30–45. DOI: 10.3390/analytics3010003. URL: https://www.mdpi.com/2813-2203/3/1/3.

[10] *The cointegrated VAR model : methodology and applications - Royal Danish Library*. URL: https://soeg.kb.dk/discovery/fulldisplay?docid=alma99123019170705763&context=L&vid=45KBDK_KGL:KGL&lang=en&search_scope=MyInst_and_CI&adaptor=Local%20Search%20Engine&tab=Everything&query=any,contains,Cointegrated%20VAR%20Model&sortby=date_d&facet=frbrgroupid,include,9045911366128049316&offset=0.