A Work Project, presented as part of the requirements for the Award of a Master's degree in Economics from the Nova School of Business and Economics.

Residential Real Estate Price Forecasting in Portugal

Sepehr Zarrabi

Work project carried out under the supervision of:

Francesco Franco

23/09/2022

Abstract

This paper employs three panel data and seven machine learning methods, including linear and nonlinear models, to perform accurate predictions of house prices for fifty-one parishes in six municipalities of Portugal. To construct the predictive models, nine time series economic factors and two non-time series features are applied as explanatory variables. Finally, the neighboring parish's lagged house prices per square meter data is added as a predictor to increase the forecasting accuracies. The utilized models are Artificial Neural Network, eXtream Gradient Boosting, Linear regression, Lasso and Ridge regression, Bayesian regression, Polynomial regression, Pooled OLS, Panel OLS, and First Difference OLS.

Keywords: Forecasting, Machine Learning, Econometrics, panel data, neural networks, gradient boosting

This work used infrastructure and resources funded by Fundação para a Ciência e a Tecnologia (UID/ECO/00124/2013, UID/ECO/00124/2019 and Social Sciences DataLab, Project 22209), POR Lisboa (LISBOA-01-0145-FEDER-007722 and Social Sciences DataLab, Project 22209) and POR Norte (Social Sciences DataLab, Project 22209).

1. Introduction

One of the most important assets for each family is their real estate property, particularly houses. In recent years, certain unfortunate occurrences have happened that undoubtedly impacted the real estate market and housing prices in countries all over the world, including Portugal. COVID-19 and Russian-Ukraine war in 2022 are two major incidents that affected the global economy. The supply shortage of oil, natural gas, and food due to the Russian-Ukraine war led to a steady increase in inflation in most nations. According to analysts, European countries will notice a higher inflation rate and a disruption in their supply chains (Mbah and Wasum, 2022). Furthermore, concerns about safety issues in European countries may influence the investors' behavior in the real estate market (Mbah and Wasum, 2022).

COVID-19, like previous pandemics, causes exogeneity and brings difficulties in measuring its influence on the real estate industry. In addition, the investigation is more complicated by a shortage of available data, especially for low-frequency time series data in real estate. Due to the pandemic, the number of commercial and residential property transactions falls, individuals quit their flats in urban regions, and households have financial difficulties paying their mortgages (Balemi et al., 2021). Hence, an accurate understanding is essential for different sectors, such as national banks, financial organizations, private investors, and most importantly, governments, to be ready for any possible future waves (Balemi et al., 2021). The uncertainties caused by the Russian-Ukrain war and coronavirus make a proper prediction for residential real estate prices valuable.

Moreover, Idealista, a famous Portuguese real estate advertising webpage, presents various articles which show that during the pandemic, the Portuguese real estate market attracted domestic and foreign investors for construction development due to the lack of supply and increase in prices for this market (Idealista.pt, 2022). The rise in Portuguese housing prices was

higher than the average growth for EU and Eurozone countries (Idealista.pt, 2021). According to the International Monetary Fund's (IMF) research, the expansion in the real estate market during Covid was influenced by several causes. The decrease applied by the government in the interest rate on mortgages to prevent unemployment and the adverse pandemic effects on the household's income, and developing concepts such as remote working, led to a decrease in consumption and increased savings. Due to individuals spending the majority of their time indoors, bigger homes with more significant outside space and facilities like swimming pools and gardens became more appealing. The demand increased as a result of these causes. On the other hand, the construction market slowed due to a lack of materials and labor. As a result, the real estate market's supply shrank. This mismatch in the housing market between demand and supply generated a price boom, prompting governments to seek remedies (Idealista.pt, 2021; IMF.org, 2021).

To find a proper forecasting model that performs accurate results for Portuguese residential real estate prices, linear and nonlinear approaches using econometrics panel data models and machine learning methods are applied in this paper. The utilized time series data is quarterly from 2016:Q1 to 2021:Q1 for the median value per square meter of dwellings sales as a target variable. Additionally, eleven time series and non-time series explanatory variables are employed to improve the forecasting models. In general, previous research on Portuguese real estate price prediction acquired data at high geographical aggregate levels; hence the parish (Freguesia) geographical level is used for this study. Fifty-one parishes from six municipalities, including Porto, Almada, Lisbon, Amadora, Cascais, and Oeiras, are selected.

The remainder of the paper is structured as follows. Section 2 contains the related literature about the topic, section 3 describes the utilized data and models, section 4 presents the results and model comparison, and the conclusions are presented in Section 5.

2. Literature Review

For a variety of reasons, residential real estate price forecasting is tricky. The major challenge is the limited amount of time series data that is currently accessible. It is challenging to design and test models since house price indices are commonly generated at monthly and quarterly frequencies, which restricts the duration of calculated indices. Milunovich performs different linear and nonlinear models to forecast the log real house prices in Australia with seven other predictors, including the consumer price index and unemployment rate. According to his findings, linear models perform better for predicting the first quarter of the future. For eight-quarter prediction, however, nonlinear methods appear to be more appropriate. Moreover, to deal with nonlinear relationships in the data, he recommends machine learning and deep learning algorithms (Milunovich, 2020).

Nonlinear behaviors in the house market data are amplified because of the high transaction costs in this market (Muellbauer and Murphy, 1997). Therefore, we can use several tools from machine learning to outline the various sorts of nonlinear relationships in the data. Varian (2014) encourages economists to learn and use new machine learning methods, and he introduces some tools from machine learning for data analysis. In econometrics, data analysis falls into four categories: prediction, summarization, estimation, and hypothesis testing. The main focus of machine learning is prediction. The related field of data mining is also concerned with summarization, particularly finding exciting patterns in the data. The most popular approach in the case of summarization is regression analysis (Varian, 2014).

Siwicki (2021) uses different methods such as linear regression, artificial neural network (ANN), random forest, extreme gradient boosting (XGBoost), and spatial error model with KNN weight matrix to predict housing prices in Warsaw and compare their forecasting accuracy. He mentions linear regression as an essential statistical tool and neural network as

the most popular machine learning model in real estate price analysis. Finally, Siwicki recommends the extreme gradient boosting model (XGBoost), which can help to deal with real-world problems. The evaluation results show that XGBoost and random forest regressions perform the best predictions among all methods (Siwicki, 2021).

Sang-Hyang Lee and his colleagues (2021) perform a forecasting model for land prices in South Korea. They study the correlations between land prices and thirteen micro and macro variables with python and R. As a result of the analysis by R, interest rate and the number of crimes are two variables with significant correlation among the four significantly correlated variables. From the python analysis, the real estate tax charge is one of the independent variables with a significant correlation (Lee et al., 2021).

The socioeconomic makeup of communities can shift significantly due to crime (Tita et al., 2006). Crime victimization close to one's house increases the likelihood that people change their residence neighborhood (Dugan, 1999). Frischtak and Mandel (2012) remark that if crime declines, low-priced homes will likely respond more positively, and price inequality may drop (Frischtak and Mandel, 2012). According to other authors, earnings frequently play a crucial part in forecasting real estate prices (Favilukis et al., 2017; Malpezzi, 1999).

Since 1989, modern homes have been constructed on smaller plots following the density maximization idea. As a result, green spaces in neighborhoods are usually restricted and might be seen as a club good that is only accessible to residents. According to their findings, which are consistent with earlier studies, being close to a park positively correlates with apartment pricing. An apartment's price rises by 2.8% to 3.1% on average if an urban green space (park) is within 100 meters. The impact of park access on home values is more significant for newer units than those constructed before 1989. These are the findings of Trojanek and his colleagues about the impact of green areas on Warsaw's house prices (Trojanek et al., 2018).

One recent Portuguese real estate price prediction research is from Samadani and Costa (2021). They used economic and machine learning methods to analyze house prices at the aggregate level of the country with four independent variables: crime rate, selected usage (the form of processing waste as a proxy to measure the comfort of the house sold), purchasing power, and tax rate (IMI and IMT per capita). Lasso regression and linear regression had the highest accuracy on the test subset between their machine learning models (Samadani and Costa, 2021).

According to these publications, econometrics and machine learning methods, including linear and nonlinear models, are employed in this study to forecast residential real estate prices using more current data. This paper includes parish geographical aggregate levels instead of country levels to be more practical.

3. Methodology

3.1. Data

There was a rise in the value of construction sales until 2010. The overall value then dropped but has subsequently recovered since 2013. Prior to the middle of 2012, the average annual decline in housing prices since the financial crisis started was roughly 3%. Furthermore, Portugal's housing market has seen substantial changes since 2013, including a sharp increase in real estate transactions and a rise in home prices due to a limited supply of real estate and a massive increase in demand. (Samadani and Costa, 2021; Del Giudice et al., 2020; Nicola et al., 2020). Regarding these findings, the sample from more recent data is used to exclude the fluctuations before 2013. The utilized data is quarterly time series from the first quarter of 2016 until the last quarter of 2021.

The Median value per square meter of dwellings sales is the dependent variable in the mentioned time horizon and at the geographic aggregate level of the parish (Freguesia) for fifty-one parishes in six municipalities in Portugal, including Porto, Almada, Lisbon, Amadora, Cascais, and Oeiras. To construct the prediction models, time series data for nine economic aspects are obtained as features such as consumer price index (CPI), construction cost index (CCI), the interest rate on housing loans, unemployment rate, the number of live births, crime rate, the municipal property tax (Imposto Municipal sobre Imóveis (IMI)), The Municipal Tax on Property Transfers (Imposto Municipal sobre as Transmissões Onerosas de Imóveis (IMT)), and average earnings. The geographic levels of these datasets are divided into parishes, municipalities, NUTS II, and the country. In addition, the density of subway stations for parishes with subway and the density of parks are two non-time series independent variables. These two fixed features were extracted from Google Maps as the number of subway stations and the number of parks for each parish. Then, the areas per square kilometer of parishes are used as a proxy for the constructed area in each parish to perform the subway station's and park's density. Moreover, the municipalities' areas per square kilometer have been considered to normalize the aggregate IMI and IMT tax and the number of live births. For the machine learning models, the lagged housing price per square meter of the most correlated neighbor parish is included as an extra explanatory variable to improve the predictive models. The parish being studied shares the same border with this neighboring parish.

The primary source of data is INE (Instituto Nacional de Estatística) Portugal national institute for statistics. The source for the crime rate data is the PORDATA database.

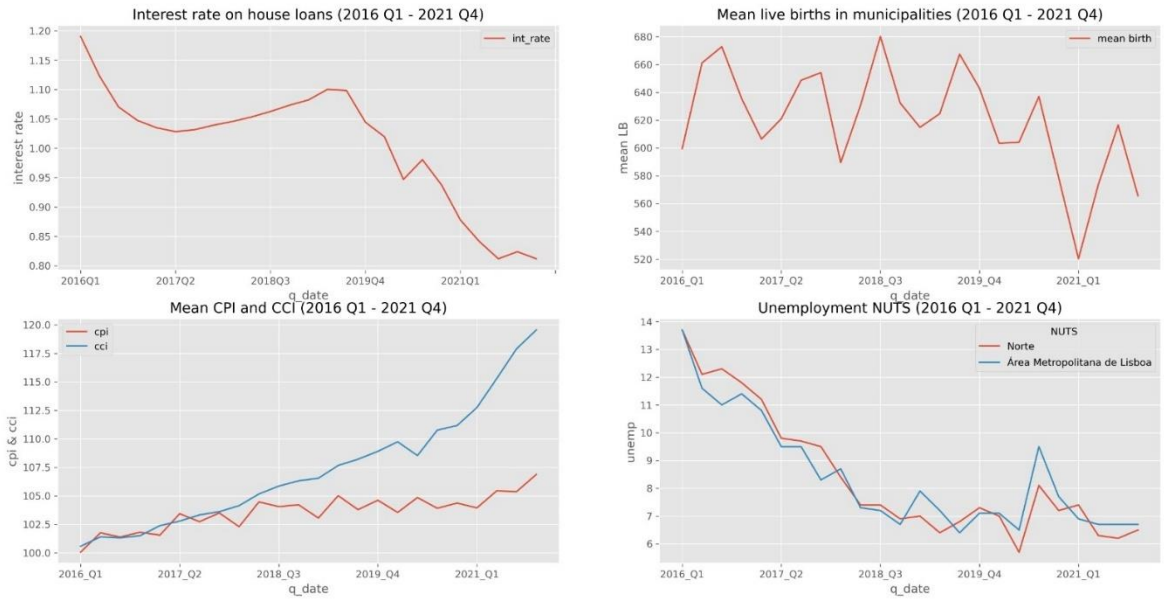

Figure 1 *Time series data (2016 Q1-2021 Q4) plots for seven of the predictors.*

Train and Test split

While making predictions, the objective is to provide accurate forecasts on the subset of data that is not included in training the models. It is simple to build a model that performs admirably for in-sample data but is inaccurate for out-of-sample datasets, which refers to the "overfitting" problem. It is common practice to split data into different subsets for training and testing. The training subset is used to estimate the model, while the test subset is used to evaluate the model's performance. The acceptable model is a model with high accuracy in both train and test subsets (Varian, 2014).

According to Varian's (2014) recommendations, each model is trained once on the in-sample period 2016:Q1-2020:Q4, then forecasts are generated across the out-of-sample range 2021:Q1-2021:Q4 (80%-20%). As a result, no models are trained using the test subset.

Variable Selection

For the panel data analysis, after fitting each model, the statistically significant variables at 0.05 are chosen for prediction by checking their p-values. The five most correlated variables are picked to fit the models and make predictions for machine learning models by running a correlation test between the dependent variable and time series explanatory variables. Each Pearson correlation coefficient's p-value is evaluated. Another correlation test is run to select one parish from the neighbor parishes with common borders to include its lagged price as an additional predictor. Predictor variables differ for each parish in machine learning approaches. Due to the missing values in data for average earnings, it dropped from explanatory variables.

3.2. Models

This literature uses three panel data approaches and seven machine learning methods to anticipate the house value per square meter for chosen parishes. The panel data estimators include Pooled OLS, Panel OLS, and First Difference OLS. The machine learning models contain linear and nonlinear methods, including Linear regression, Lasso and Ridge regression, Bayesian regression, Polynomial regression, Artificial Neural Network (ANN), and eXtream Gradient Boosting (XGBoost). The data is quarterly from 2016:Q1 to 2021:Q4 for fifty-one parishes in six municipalities in Portugal. Finally, the models are compared through their accuracy score, the coefficient of determination (R-squared), and the standard error of the estimate. R-squared is a statistical metric that indicates how near the data is to the estimated regression model from 0 to 100% (it can also have negative values). The standard error shows how different are the estimated values from the actual values. The equations for R-squared and the standard error of the estimate are shown below.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \quad \text{(Eq.1)}$$

N is the observation number, $\hat{y}_i$ is the fitted value, and $\bar{y}$ is the mean of y_i.

$$SE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}} \quad \text{(Eq.2)}$$

I) Panel Data Models

Panel data, often known as longitudinal data, are time series observations of a set of individuals. Usually, statements in panel data include a minimum of two dimensions, including cross-sectional and time series dimensions denoted by i and t. The general equation for panel data is shown below. (Hsiao, 2007).

$$y_{it} = \alpha_i + \beta'_{it} x_{it} + e_{it} \quad \text{(Eq.3)}$$

The three panel data models used in this paper are Pooled OLS, Panel OLS, and First Difference OLS. Pooled OLS, or the constant coefficient model, has constant coefficients for both the slopes and the intercepts. It is a simple panel data method that pools the whole data and performs simple OLS regression. Even though typically entity or time effects are present, sometimes they might not be statistically significant (check equation 5) (Yaffee, 2003). Panel OLS is the same as Pooled OLS but with time or entity effects, or both; the general formula is shown in equation 6. Examples of entity effects are park and subway station density, which are constant in the time but different for each parish. Examples of time effects are the time series variables in the geographical aggregate of the country, such as the interest rate on housing loans.

$$y_{it} = \beta'_{its} x_{its} + e_{it} \quad \text{(Eq.4)}$$

In this equation, s refers to the twelve exogenous variables, including time series and non-time series characters.

$$y_{it} = \alpha_{is} + \gamma_{ts} + \beta'_{its}x_{its} + e_{it} \quad \text{(Eq.5)}$$

The first-difference (FD) estimator is a panel data model which is employed to deal with the issue of omitted variables. It takes the first difference of variables in the equation and then works the same as The Fixed Effect model. Sometimes, it may be more effective than the typical fixed effects estimator.

$$\Delta y_{it} = \Delta\gamma_{ts} + \beta'_{its}\Delta x_{its} + \Delta e_{it} \quad \text{(Eq.6)}$$

The linearmodels library from python has functions for the panel data methods.

I) Machine Learning Models

This section briefly introduces nonlinear models, such as artificial neural network and XGBoost, and linear models, including Linear, Lasso, Ridge, Bayesian, and Polynomial regression.

Artificial Neural Network (ANN)

One of the most prevalent machine learning methods utilized in real estate market analysis research is the artificial neural network. The neural network concept is based on how the biological human brain organizes data. It is a collection of linked neurons that perform a series of changes on input and construct its perspective of the input layer nodes through hidden layers as an output. Therefore, a neural network model contains at least three layers: input, hidden (one or more), and output. In the layers, each neuron is assigned a connection power to other neurons in the next layer, denoted as the weight, representing the impact of one neuron

on another. Because of its many layers, the neural network can handle issues with both linear and nonlinear interactions (Siwicki, 2021; Bency et al., 2017; Limsombunchai, 2004; Selim, 2009). Each hidden layer has the output h_i for neuron i as below.

$$h_i = \sigma\left(\sum_{j=1}^{N} W_{ij} x_j + T_i^{hid}\right) \quad \text{(Eq.7)}$$

In this equation, σ is the activation or transfer function that contains W_{ij} and x_j as the weights and information from the input layer nodes, and T_i^{hid} as the hidden layer neurons' threshold terms. The number of input neurons is denoted by N. The activation function's main objective is to limit the value of the primary nodes so that diverse neurons do not confuse the neural network model (Wang, 2003). The Multilayer Perceptron (MLP) is the most basic neural network model employed from the Scikit-learn python's library in this study.

eXtreme Gradient Boosting (XGBoost)

Additionally to the artificial neural network, decision tree models are suitable for dealing with nonlinear data behaviors. Moreover, trees handle missing data and work well for large datasets. Still, their major disadvantage is that they are prone to overfitting and give a weak outcome for forecasting the future. Adding randomness to the dataset is one technique to handle the overfitting issue. Three common strategies for adding randomness are Boosting, Bagging, and Bootstrap. Boosting is a technique of repetitive estimations in which the misclassified observations gain weights among each iteration and then take an average (or vote) between the repeated estimations. Boosting may be used in almost any classification or regression model and has a tendency to increase an estimator's predictive ability (Varian, 2014). Between Boosting models, data scientists frequently use XGBoost, a scalable tree boosting technique, to get more accurate outcomes on various problems, which can handle real-world issues with limited resources (Chen and Guestrin, 2016). Python has the xgboost library for the extreme gradient boosting method.

Linear Regression

The primary purpose of linear regression is to find the best coefficients for variables that reduce the residual sum of squares (RSS) between the known objects in the dataset and the forecasted ones as much as possible. Generally, linear models are popular among statisticians, and this popularity is for even before the invention of computers (Siwicki, 2021; Conway et al., 2008; Osland, 2010; Wilhelmsson, 2002; Zhang et al., 2015). This paper's first linear regression model is the Ordinary least squares linear regression imported from the Scikit-learn python's library. The general OLS equation is shown as follows.

$$y_i = \sum_{j=1}^{P} \beta_j X_{ij} + e_i \quad \text{(Eq.8)}$$

In the equation, X_{ij} is the observation of predictor j at period i, and β_j is the coefficient of that predictor. P stands for the number of independent variables.

The RSS, also known as the cost function, is a measure of how different the actual and predicted values are in the chosen linear model.

$$RSS = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \quad \text{(Eq.9)}$$

$$RSS = \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{P} \beta_j X_{ij}\right)^2 \quad \text{(Eq.10)}$$

Where $\hat{y}_i$ stands for the fitted value of y_i.

Lasso and Ridge Regression

Lasso (Least Absolute Shrinkage and Selection Operator) and Ridge regressions are two modifications of the linear regression model that are used to overcome the potential overfitting and multi-colinearity in a multiple linear regression model. The central equation to find the coefficients of regressors is the same as linear regression, but with adding a shrinkage through a penalty term known as α. The range of α is from zero to infinity. The zero α makes the model

similar to simple linear regression (OLS), and the higher α increases the penalty effect. The difference between Lasso and Ridge is how they imply this penalty effect into the coefficients in the equation. They are also known as L1 regularization for Lasso and L2 regularization for Ridge (Milunovich, 2020).

$$L_1 = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 + \alpha \sum_{j=1}^{P} |\beta_j| \quad \text{(Eq.11)}$$

$$L_2 = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 + \alpha \sum_{j=1}^{P} \beta_j^2 \quad \text{(Eq.12)}$$

In this study, α is set to 0.5 for Lasso and Ridge regression. Scikit-learn has functions for both Lasso and Ridge regression.

Bayesian Regression

Bayes' theorem describes the relationship between the past and the subsequent beliefs and updates the initial ideas regarding the model parameters. The Bayesian method is a comprehensive probability model that captures uncertainty in the output value conditional on unknown coefficients and prior uncertainty about the coefficients. The purpose is to find the updated posterior probability distribution for the model parameters with the input variables and outcomes. The normal (Gaussian) distribution is assumed in Bayesian Linear Regression.

$$p(\beta|y,X) \propto p(y|\beta,X)p(\beta) \quad \text{(Eq.13)}$$

The assumption is that the set of explanatory variables, X, is not related to β, then $p(\beta|X)$ is equal to $p(\beta)$ (Muth et al., 2018).

In this paper, the utilized Bayesian model from the Scikit-learn library is Bayesian Ridge Regression as a type of Bayesian regression, which fits a probabilistic model to maximize the posterior probability.

Polynomial Regression

The last regression model in this study is Polynomial Regression, a special case of multiple linear regression. Additionally to the first two machine learning methods, ANN and XGBoost, Polynomial regression is a practical method to deal with nonlinearity in the data. The general equation of polynomial regression is shown as follows, where k, the order of the equation, is known as the degree of polynomial (Ostertagová, 2012).

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \cdots + \beta_k x_i^k + e_i \qquad \text{(Eq.14)}$$

In this paper, the chosen value for the degree of polynomial is 2. To use python for polynomial regression, Pipeline, PolynomialFeatures, and LinearRegression are imported functions from the Scikit-learn library.

4. Results

This section displays the results of the methods stated above and evaluates and interprets each method's accuracy to identify the best-fitted model for residential real estate price forecasting in Portugal with the mentioned dataset. The accuracy of each model is calculated by R-squared and the standard error of the estimate. R-squared is employed on both the train and the test subsets not only for model comparison but also to determine whether or not each model has the overfitting problem. On the other hand, the standard error is applied just for the test subset.

The panel data and machine learning models are fitted on the dataset with 1224 observations for all parishes, then done once again only for the parishes with subway stations, including 576 observations. Moreover, in order to improve the prediction accuracy by employing the neighbor parish's lagged house prices, machine learning models are used once more on the time series data for each parish separately with 24 rows and five most correlated predictors in

addition to the neighborhood lagged prices. These predictors are different for each parish, and non-time series explanatory variables are not included.

Starting with panel data models, Pooled OLS is the most accurate method with the lowest standard error of the estimate and greatest R-square in the test and train subsets among the three utilized models. The overall accuracy decreases by adding extra time and entity effects to the simple Pooled OLS model and changing it to the Panel OLS. If requested, the Panel OLS estimator is able to remove fully absorbed variables. The dropped variables are CCI and interest rate in this case. The First Difference OLS eliminates the fixed variables over time; hence, this model cannot use subway station and park densities.

Figure 2 represents the summary of Pooled OLS estimator and coefficients' information for the whole dataset. According to figure 2, the Model's R-squared score is 0.9546, and the p-value is zero. The consumer price index and unemployment rate are not statistically significant for the significance level of 0.05. The obtained Pooled OLS equation is as follows.

$$P_{it} = 271.03 * PARK_i + 28.46 * CCI_t + (-648.82) * IntRate_t + \quad \text{(Eq.15)}$$
$$(-2.61) * BIRTH_{it} + (-14.20) * CRIME_{it} + (-0.87) * IMI_{it} + 1.31 *$$
$$IMT_{it} + e_{it}$$

The park density, CCI, and IMT are the variables that positively impact the price per square meter. For instance, increasing one unit of the park density in a parish raises the price per square meter of dwellings by 271.03 euros in that parish. In contrast, the interest rate on housing loans, number of live births, crime rate, and IMI negatively affect the target variable.

PooledOLS Estimation Summary

Dep. Variable:	price	R-squared:	0.9546
Estimator:	PooledOLS	R-squared (Between):	0.9657
No. Observations:	1224	R-squared (Within):	0.7265
Date:	Mon, Sep 05 2022	R-squared (Overall):	0.9546
Time:	17:18:10	Log-likelihood	-9394.8
Cov. Estimator:	Unadjusted		
		F-statistic:	2836.6
Entities:	51	P-value	0.0000
Avg Obs:	24.000	Distribution:	F(9,1215)
Min Obs:	24.000		
Max Obs:	24.000	F-statistic (robust):	2836.6
		P-value	0.0000
Time periods:	24	Distribution:	F(9,1215)
Avg Obs:	51.000		
Min Obs:	51.000		
Max Obs:	51.000		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
cpi	-3.3848	9.7925	-0.3457	0.7297	-22.597	15.827
cci	28.463	7.2886	3.9051	0.0001	14.163	42.762
birth	-2.6134	0.4671	-5.5944	0.0000	-3.5299	-1.6969
crime_rate	-14.199	1.7044	-8.3308	0.0000	-17.543	-10.855
imi	-0.8664	0.1454	-5.9579	0.0000	-1.1516	-0.5811
imt	1.3079	0.0491	26.639	0.0000	1.2115	1.4042
unemp	-5.2121	10.344	-0.5039	0.6144	-25.506	15.082
interest_rate	-648.82	291.67	-2.2245	0.0263	-1221.1	-76.584
park_density	271.03	28.859	9.3918	0.0000	214.42	327.65

Figure 2 *Pooled OLS estimator for all parishes.*

The following figure (Figure 3) displays Pooled OLS results for parishes with subway stations, including 24 parishes from Lisbon, Porto, and Amadora. The estimator's R-squared score is 0.9609, and the p-value is zero. The construction cost index, consumer price index, and interest rate are insignificant variables for a 0.05 p-value as the significance level.

PooledOLS Estimation Summary

Dep. Variable:	price	R-squared:	0.9609
Estimator:	PooledOLS	R-squared (Between):	0.9725
No. Observations:	576	R-squared (Within):	0.7191
Date:	Mon, Sep 05 2022	R-squared (Overall):	0.9609
Time:	19:48:28	Log-likelihood	-4442.2
Cov. Estimator:	Unadjusted		
		F-statistic:	1391.1
Entities:	24	P-value	0.0000
Avg Obs:	24.000	Distribution:	F(10,566)
Min Obs:	24.000		
Max Obs:	24.000	F-statistic (robust):	1391.1
		P-value	0.0000
Time periods:	24	Distribution:	F(10,566)
Avg Obs:	24.000		
Min Obs:	24.000		
Max Obs:	24.000		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
cpi	-1.4457	15.891	-0.0910	0.9275	-32.658	29.766
cci	-2.1159	11.531	-0.1835	0.8545	-24.766	20.534
birth	-8.0872	0.9116	-8.8710	0.0000	-9.8778	-6.2965
crime_rate	-28.316	3.4899	-8.1137	0.0000	-35.171	-21.461
imi	3.1002	0.4995	6.2062	0.0000	2.1190	4.0814
imt	0.7774	0.1106	7.0267	0.0000	0.5601	0.9947
unemp	-95.428	20.033	-4.7634	0.0000	-134.78	-56.079
interest_rate	484.91	480.00	1.0102	0.3128	-457.88	1427.7
subway_density	524.50	46.699	11.231	0.0000	432.77	616.22
park_density	278.55	39.828	6.9938	0.0000	200.32	356.78

Figure 3 *Pooled OLS estimator for parishes with subway stations.*

Following is the Pooled OLS equation for the mentioned parishes.

$$P_{it} = 524.50 * SUBWAY_i + 278.55 * PARK_i + (-95.43) * \quad \text{(Eq.16)}$$

$$UNEMP_{nt} + (-8.09) * BIRTH_{it} + (-28.32) * CRIME_{it} + 3.10 *$$

$$IMI_{it} + 0.78 * IMT_{it} + e_{it}$$

The parameter n for the unemployment rate stands for the NUTS II geographical level, which contains "Norte" or "Área Metropolitana de Lisboa".

Adding one unit to the subway station density in a parish increases the house prices per square meter by 524.50 euros. The variables with positive effects are the subway station density, park density, IMI, and IMT. The unemployment rate, the number of live births, and the crime rate are all negatively impacting factors.

The accuracy scores for the test subset after predicting with Pooled OLS are presented in Table 1.

Table 1 *Pooled OLS prediction accuracies*

Measure	All parishes	Subway parishes
R-squared	0.633	0.682
SE	35.42	51.27

The Panel OLS has an acceptable R-squared and p-value for the model but performs poorly in the prediction (Figures 4 and 5 in the appendix). The estimator's R-squared for the First Difference OLS is low (Figure 6 in the appendix).

Table 2 *Machine learning models' accuracies for all parishes.*

Model	Train subset R-squared	Test subset R-squared	Test subset SE
Neural Network	0.606	0.396	37.81
XGBoost	0.817	0.626	33.81
Linear Regression	0.702	0.152	37.47
Lasso Regression	0.701	0.203	37.44
Ridge Regression	0.702	0.176	37.48
Bayesian Regression	0.701	0.248	37.39
Polynomial Regression	0.780	-4.415	67.25

Table 2 represents the accuracy results for machine learning models with the same explanatory variable as the Pooled OLS model, which shows the potential nonlinear behaviors in the data that make predictions from linear-based models inaccurate. High accuracy in the training subset and low accuracy in the test subset is the output of the overfitting models. XGBoost is the only accurate predictive model among the other six methods. The negative R-squared for Polynomial regression indicates that the fitted line performs more poorly than the horizontal line in the test subset.

Table 3 *Machine learning models' accuracies for parishes with subway stations.*

Model	Train subset R-squared	Test subset R-squared	Test subset SE
Neural Network	0.614	0.381	66.43
XGBoost	0.919	0.805	39.80
Linear Regression	0.769	-0.752	58.31
Lasso Regression	0.769	-0.724	58.32
Ridge Regression	0.769	-0.739	58.32
Bayesian Regression	0.769	-0.707	58.41
Polynomial Regression	0.891	-1.296	145.45

According to table 3, adding the subway station density as a predictor worsens the accuracies of linear-based models in the test subset. Compared to table 2, XGBoost has better R-squared scores in training and test subsets but a worse standard error of the estimate. The overfitting problem still exists in the neural network model.

The two prediction models with the best output for the data without the subway station density are Pooled OLS and XGBoost. XGBoost performs better in parishes with subway stations.

Finally, to add the neighbor parish's lagged housing price per square meter, the machine learning models are employed on the time series datasets for each parish. The following tables show the accuracy results for the first parish of each municipality.

Table 4 *R-squared of the test subsets for the first parishes of Porto, Almada, Amadora, Cascais, Lisbon, and Oeiras.*

Parish	ANN	XGB	LR	Lasso	Ridge	Bayesian	Polynomial
Bonfim	0.947	0.973	0.928	0.929	0.929	0.947	0.834
Costa da Caparica	0.970	0.986	0.985	0.984	0.984	0.969	0.923
Alfragide	0.987	0.958	0.979	0.979	0.979	0.989	0.966
Alcabideche	0.975	0.851	0.981	0.989	0.987	0.985	0.981
Ajuda	0.987	0.914	0.974	0.977	0.978	0.969	0.712
Barcarena	0.902	0.908	0.932	0.944	0.860	0.806	0.681

The small test subset might cause a high R-squared score since there is a shortage of out-of-sample values.[1] Increasing the share of the test subset helps to overcome this issue; however, as there are only 24 time periods of data, decreasing the training subset negatively affects the training and prediction accuracies.

According to Table 5 in the following, the standard error scores improve. Since models behave differently for each parish's time series data, each parish should be analyzed independently for the model comparison. Generally, The polynomial model with order 2 has the weakest output among other models. In some parishes, XGBoost tends to overfit due to the shortage of data.

[1] Note that the R-squared scores for the training subset are more significant than the test subset. Increasing the test subset's percentage is mainly used when the test R-squared is large and is greater than the train R-squared (Table 6 in the appendix).

One technique to find better prediction outputs is to average over the used models (Varian, 2014).

Table 5 The standard error of the estimate for the first parishes of Porto, Almada, Amadora, Cascais, Lisbon, and Oeiras.

Parish	ANN	XGB	LR	Lasso	Ridge	Bayesian	Polynomial
Bonfim	23.06	24.15	20.02	21.15	21.19	20.73	56
Costa da Caparica	25.54	15.79	15.43	16.69	16.12	26.08	41.43
Alfragide	14.44	27.4	18.52	18.45	18.31	14.01	24.78
Alcabideche	21.22	52.72	20	14.51	16.12	17.53	20
Ajuda	30.02	67.82	41	39	38.05	42.78	104.71
Barcarena	39.58	37.01	32.01	29.59	47.12	54.17	66.45

5. Conclusions

The primary purpose of this research is to develop prediction models that provide reliable results for residential real estate future prices per square meter in Portugal. Regarding this issue, the data from 2016:Q1 to 2021:Q2 at the geographical aggregate level of the parish is utilized for employing econometrics and machine learning models. The data is split into training and test subset to analyze the model's strength in estimating out-of-sample data and capture the potential overfitting problem. R-squared and standard error of the estimate are used for the model evaluation. The results represent XGBoost and Pooled OLS as the most accurate methods on the test subset among the machine learning and panel data approaches. Moreover, the data for subway station density is added as an explanatory variable to analyze the models'

outcome. XGBoost performs best in dealing with the nonlinearity in the data, while the neural network is overfitted.

The other aim of this study is to improve the machine learning models by adding the lagged house prices per square meter of the most correlated neighbor parish with a common border as another predictor. The machine learning models are applied to the time series data for each parish separately, and the outcomes show high accuracies in training and test subsets, which the short test subset might cause.

One possible suggestion for future research is to use the more recent time series data with a lower time frame. Using web scraping techniques through real estate advertising websites is a way to find this data (Siwicki, 2021).

References

- Balemi, Nadia, Roland Füss, and Alois Weigand. "COVID-19's impact on real estate markets: review and outlook." Financial Markets and Portfolio Management 35, no. 4 (2021): 495-513.

- Bency, Archith J., Swati Rallapalli, Raghu K. Ganti, Mudhakar Srivatsa, and B. S. Manjunath. "Beyond spatial auto-regressive models: Predicting housing prices with satellite imagery." In 2017 IEEE winter conference on applications of computer vision (WACV), pp. 320-329. IEEE, 2017.

- Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785-794. 2016.

- Del Giudice, Vincenzo, Pierfrancesco De Paola, and Francesco Paolo Del Giudice. "COVID-19 infects real estate markets: Short and mid-run effects on housing prices in Campania region (Italy)." Social sciences 9, no. 7 (2020): 114.

- Dugan, Laura. "The effect of criminal victimization on a household's moving decision." Criminology 37, no. 4 (1999): 903-930.

- Favilukis, Jack, Sydney C. Ludvigson, and Stijn Van Nieuwerburgh. "The macroeconomic effects of housing wealth, housing finance, and limited risk sharing in general equilibrium." Journal of Political Economy 125, no. 1 (2017): 140-223.- Frischtak, C. and Mandel, B.R., 2012. Crime, house prices, and inequality: The effect of UPPs in Rio. Available at SSRN 1995795.

- Hsiao, Cheng. "Panel data analysis—advantages and challenges." Test 16, no. 1 (2007): 1-22.

- Idealista.pt. **“**Real estate development will be one of the most promising sectors in 2022 for Portugal.” (2022). https://www.idealista.pt/en/news/financial-advice-in-portugal/2022/02/07/4824-real-estate-development-will-be-one-of-the-most-promising-sectors-in-2022-for

- Idealista.pt. “Property prices in Portugal have soared in the last decade.” (2021). https://www.idealista.pt/en/news/property-for-sale-in-portugal/2021/01/21/845-property-prices-in-portugal-have-soared-in-the-last-decade

- Idealista.pt. “House prices in Portugal are rising faster than wages.” (2021). https://www.idealista.pt/en/news/financial-advice-in-portugal/2021/12/07/4753-house-prices-in-portugal-are-rising-faster-than-wages-why

- IMF.org. “Housing Prices Continue to Soar in Many Countries Around the World.” (2021) https://blogs.imf.org/2021/10/18/housing-prices-continue-to-soar-in-many-countries-around-the-world/

- Lee, Sang-Hyang, Jae-Hwan Kim, and Jun-Ho Huh. "Land Price Forecasting Research by Macro and Micro Factors and Real Estate Market Utilization Plan Research by Landscape Factors: Big Data Analysis Approach." Symmetry 13, no. 4 (2021): 616.

- Limsombunchao, Visit. "House price prediction: hedonic price model vs. artificial neural network." New Zealand Agricultural and Resource Economics Society Conference, 25-26 June 2004. Blenheim, New Zealand: New Zealand Agricultural and Resource Economics Society, 2004.

- Malpezzi, Stephen. "A simple error correction model of house prices." Journal of housing economics 8, no. 1 (1999): 27-62.

- Mbah, Ruth Endam, and Divine Forcha Wasum. "Russian-Ukraine 2022 War: A review of the economic impact of Russian-Ukraine crisis on the USA, UK, Canada, and Europe." Advances in Social Sciences Research Journal 9, no. 3 (2022): 144-153.

- McDonald, Gary C. "Ridge regression." Wiley Interdisciplinary Reviews: Computational Statistics 1, no. 1 (2009): 93-100.

- Milunovich, George. "Forecasting Australia's real house price index: A comparison of time series and machine learning methods." Journal of Forecasting 39, no. 7 (2020): 1098-1118.

- Muellbauer, John, and Anthony Murphy. "Booms and busts in the UK housing market." The Economic Journal 107, no. 445 (1997): 1701-1727.

- Muth, Chelsea, Zita Oravecz, and Jonah Gabry. "User-friendly Bayesian regression modeling: A tutorial with rstanarm and shinystan." Quantitative Methods for Psychology 14, no. 2 (2018): 99-119.

- Nicola, Maria, Zaid Alsafi, Catrin Sohrabi, Ahmed Kerwan, Ahmed Al-Jabir, Christos Iosifidis, Maliha Agha, and Riaz Agha. "The socio-economic implications of the coronavirus pandemic (COVID-19): A review." International journal of surgery 78 (2020): 185-193.

- Ostertagová, Eva. "Modelling using polynomial regression." Procedia Engineering 48 (2012): 500-506.

- Samadani, Sanam, and Carlos J. Costa. "Forecasting real estate prices in Portugal: A data science approach." In 2021 16th Iberian Conference on Information Systems and Technologies (CISTI), pp. 1-6. IEEE, 2021.

- Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

- Selim, Hasan. "Determinants of house prices in Turkey: Hedonic regression versus artificial neural network." Expert systems with Applications 36, no. 2 (2009): 2843-2852.

- Siwicki, Dawid. The Application of Machine Learning Algorithms for Spatial Analysis: Predicting of Real Estate Prices in Warsaw. University of Warsaw, Faculty of Economic Sciences, 2021.

- Tita, George E., Tricia L. Petras, and Robert T. Greenbaum. "Crime and residential choice: a neighborhood level analysis of the impact of crime on housing prices." Journal of quantitative criminology 22, no. 4 (2006): 299-317.

- Trojanek, Radoslaw, Michal Gluszak, and Justyna Tanas. "The effect of urban green spaces on house prices in Warsaw." International Journal of Strategic Property Management 22, no. 5 (2018): 358-371.

- Varian, Hal R. "Big data: New tricks for econometrics." Journal of Economic Perspectives 28, no. 2 (2014): 3-28.

- Wang, Sun-Chong. "Artificial neural network." In Interdisciplinary computing in java programming, pp. 81-100. Springer, Boston, MA, 2003.

- Yaffee, Robert. "A primer for panel data analysis." Connect: Information Technology at NYU 8, no. 3 (2003): 1-11.

Appendix

PanelOLS Estimation Summary

Dep. Variable:	price	R-squared:	0.6307
Estimator:	PanelOLS	R-squared (Between):	-2.0637
No. Observations:	1224	R-squared (Within):	-2.5199
Date:	Mon, Sep 05 2022	R-squared (Overall):	-2.0850
Time:	20:15:38	Log-likelihood	-9340.6
Cov. Estimator:	Unadjusted		
		F-statistic:	291.12
Entities:	51	P-value	0.0000
Avg Obs:	24.000	Distribution:	F(7,1193)
Min Obs:	24.000		
Max Obs:	24.000	F-statistic (robust):	291.12
		P-value	0.0000
Time periods:	24	Distribution:	F(7,1193)
Avg Obs:	51.000		
Min Obs:	51.000		
Max Obs:	51.000		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
cpi	25.957	65.756	0.3948	0.6931	-103.05	154.97
birth	-2.5540	0.4556	-5.6057	0.0000	-3.4479	-1.6601
crime_rate	-14.588	2.2225	-6.5639	0.0000	-18.949	-10.228
imi	-0.9879	0.1566	-6.3080	0.0000	-1.2952	-0.6807
imt	1.3605	0.0696	19.539	0.0000	1.2239	1.4971
unemp	418.75	69.612	6.0156	0.0000	282.18	555.33
park_density	270.28	28.535	9.4720	0.0000	214.30	326.27

F-test for Poolability: 7.1914
P-value: 0.0000
Distribution: F(23,1193)

Included effects: Time

Figure 4 *Panel OLS estimator for all parishes.*

PanelOLS Estimation Summary

Dep. Variable:	price	R-squared:	0.7103
Estimator:	PanelOLS	R-squared (Between):	-347.04
No. Observations:	576	R-squared (Within):	0.3232
Date:	Mon, Sep 05 2022	R-squared (Overall):	-331.14
Time:	20:14:26	Log-likelihood	-4394.8
Cov. Estimator:	Unadjusted		
		F-statistic:	166.75
Entities:	24	P-value	0.0000
Avg Obs:	24.000	Distribution:	F(8,544)
Min Obs:	24.000		
Max Obs:	24.000	F-statistic (robust):	166.75
		P-value	0.0000
Time periods:	24	Distribution:	F(8,544)
Avg Obs:	24.000		
Min Obs:	24.000		
Max Obs:	24.000		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
cpi	464.30	118.46	3.9194	0.0001	231.60	696.99
birth	-8.9778	0.8641	-10.390	0.0000	-10.675	-7.2804
crime_rate	-19.012	6.0650	-3.1347	0.0018	-30.926	-7.0981
imi	3.4344	0.6018	5.7071	0.0000	2.2523	4.6165
imt	0.4170	0.1914	2.1781	0.0298	0.0409	0.7930
unemp	76.542	104.04	0.7357	0.4622	-127.82	280.91
subway_density	576.70	44.344	13.005	0.0000	489.59	663.80
park_density	236.81	37.788	6.2667	0.0000	162.58	311.04

F-test for Poolability: 4.0647
P-value: 0.0000
Distribution: F(23,544)

Included effects: Time

Figure 5 *Panel OLS estimator for parishes with subway stations.*

FirstDifferenceOLS Estimation Summary

Dep. Variable:	price	R-squared:	0.1892
Estimator:	FirstDifferenceOLS	R-squared (Between):	0.1550
No. Observations:	1173	R-squared (Within):	0.5707
Date:	Mon, Sep 05 2022	R-squared (Overall):	0.1743
Time:	20:16:44	Log-likelihood	-7121.6
Cov. Estimator:	Unadjusted		
		F-statistic:	33.977
Entities:	51	P-value	0.0000
Avg Obs:	24.000	Distribution:	F(8,1165)
Min Obs:	24.000		
Max Obs:	24.000	F-statistic (robust):	33.977
		P-value	0.0000
Time periods:	24	Distribution:	F(8,1165)
Avg Obs:	51.000		
Min Obs:	51.000		
Max Obs:	51.000		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
cpi	-5.4878	3.5595	-1.5417	0.1234	-12.472	1.4960
cci	38.731	2.8691	13.499	0.0000	33.102	44.361
birth	-0.2497	0.9916	-0.2518	0.8012	-2.1953	1.6958
crime_rate	6.5120	1.0617	6.1338	0.0000	4.4290	8.5949
imi	0.6345	0.2059	3.0821	0.0021	0.2306	1.0385
imt	0.0534	0.0277	1.9242	0.0546	-0.0010	0.1078
unemp	-25.990	3.9714	-6.5444	0.0000	-33.782	-18.199
interest_rate	-31.006	109.57	-0.2830	0.7772	-245.98	183.97

Figure 6 *First Difference OLS estimator for all parishes.*

Table 6 *R-squared of the training subsets for the first parishes of Porto, Almada, Amadora, Cascais, Lisbon, and Oeiras.*

Parish	ANN	XGB	LR	Lasso	Ridge	Bayesian	Polynomial
Bonfim	0.964	0.998	0.966	0.966	0.966	0.949	0.997
Costa da Caparica	0.986	0.999	0.993	0.993	0.993	0.985	0.998
Alfragide	0.996	0.997	0.996	0.996	0.996	0.992	0.998
Alcabideche	0.977	0.981	0.989	0.989	0.989	0.969	0.982
Ajuda	0.987	0.998	0.988	0.988	0.988	0.975	0.997
Barcarena	0.968	0.999	0.986	0.984	0.958	0.937	0.992

Model Comparison

	Pooled	Panel	FD
Dep. Variable	price	price	price
Estimator	PooledOLS	PanelOLS	FirstDifferenceOLS
No. Observations	1224	1224	1173
Cov. Est.	Unadjusted	Unadjusted	Unadjusted
R-squared	0.9546	0.6307	0.1892
R-Squared (Within)	0.7265	-2.5199	0.5707
R-Squared (Between)	0.9657	-2.0637	0.1550
R-Squared (Overall)	0.9546	-2.0850	0.1743
F-statistic	2836.6	291.12	33.977
P-value (F-stat)	0.0000	0.0000	0.0000
cpi	-3.3848	25.957	-5.4878
	(-0.3457)	(0.3948)	(-1.5417)
cci	28.463		38.731
	(3.9051)		(13.499)
birth	-2.6134	-2.5540	-0.2497
	(-5.5944)	(-5.6057)	(-0.2518)
crime_rate	-14.199	-14.588	6.5120
	(-8.3308)	(-6.5639)	(6.1338)
imi	-0.8664	-0.9879	0.6345
	(-5.9579)	(-6.3080)	(3.0821)
imt	1.3079	1.3605	0.0534
	(26.639)	(19.539)	(1.9242)
unemp	-5.2121	418.75	-25.990
	(-0.5039)	(6.0156)	(-6.5444)
interest_rate	-648.82		-31.006
	(-2.2245)		(-0.2830)
park_density	271.03	270.28	
	(9.3918)	(9.4720)	
Effects		Time	

T-stats reported in parentheses

Figure 7 *Panel Data estimators' comparison for all parishes.*

Model Comparison

	Pooled Subway	Panel Subway
Dep. Variable	price	price
Estimator	PooledOLS	PanelOLS
No. Observations	576	576
Cov. Est.	Unadjusted	Unadjusted
R-squared	0.9609	0.7103
R-Squared (Within)	0.7191	0.3232
R-Squared (Between)	0.9725	-347.04
R-Squared (Overall)	0.9609	-331.14
F-statistic	1391.1	166.75
P-value (F-stat)	0.0000	0.0000
cpi	-1.4457	464.30
	(-0.0910)	(3.9194)
cci	-2.1159	
	(-0.1835)	
birth	-8.0872	-8.9778
	(-8.8710)	(-10.390)
crime_rate	-28.316	-19.012
	(-8.1137)	(-3.1347)
imi	3.1002	3.4344
	(6.2062)	(5.7071)
imt	0.7774	0.4170
	(7.0267)	(2.1781)
unemp	-95.428	76.542
	(-4.7634)	(0.7357)
interest_rate	484.91	
	(1.0102)	
subway_density	524.50	576.70
	(11.231)	(13.005)
park_density	278.55	236.81
	(6.9938)	(6.2667)
Effects		Time

T-stats reported in parentheses

Figure 8 *Panel Data estimators' comparison for subway parishes.*