Analysis & Prediction of Real Estate House Pricing for Yerevan, Armenia

Author: Davit Nazlukhanyan
BS in Data Science
American University of Armenia

Author: Awadis Shikoyan
BS in Data Science
American University of Armenia

Supervisor: Pakrad Balabanian
MSC, Masters in GIS
Lund University

Abstract

This study is aimed to predict real estate prices in Yerevan using three regression models, trained and tested using ArcGIS, namely, Generalized Linear Regression (GLR), Geographically Weighted Regression (GWR), and Forest-based Classification and Regression (FBCR). We aim to answer the following research question:

- What is the best spatial data science technique for predicting real estate prices in Yerevan, Armenia, and how can it be visualized and validated to inform decision-making for property buyers, sellers, and urban planners?

The project involves exploring various techniques and tools in ArcGIS Pro & Python to analyze real estate data in Yerevan. The first stage of the project includes using IQR for outlier detection, analyzing the correlation matrix of variables, and assessing the Global Moran's Spatial Autocorrelation in order to understand our data better. The spatiotemporal distribution of house sales is visualized using the 3D tools in ArcGIS to identify patterns and trends in the data. The analysis also includes spatiotemporal data science techniques, such as hotspot analysis, Space-Time-Cube patterns for sale data, and evaluation of model performance based on standardized residuals, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2). The results showed that the FBCR model outperformed the GLR and GWR models in terms of predictive accuracy, as demonstrated by the lowest MAE and RMSE and the highest R^2. In order to have a pipeline that properly answers our research question, we used the data of real estate houses that were not sold yet to examine how well each of these models can be generalized for new data. We also try to optimize the quality of the models by analyzing the spatially varying relationships between our independent variables, constructing a baseline approach to analyze the residuals, and from there, choosing the most optimum/significant number of variables to include in each model. We conclude this project by approving that the prediction by the FBCR and GWR model comes much closer to our actual distribution prices as both models are far better at capturing the spatial heterogeneity of the data.

Keywords— real estate, price prediction model, ArcGIS, GLR, GWR, FBCR, Yerevan, spatiotemporal, spatial analytics

Table of Content

‘

I. Introduction

The real estate industry is a great indicator of the economic wellness of a country/city; thus, it is essential to be able to identify the right patterns and make predictions regarding prices of real estate, which would help buyers, sellers and investors to make informed decisions.

Over the years, many systems and techniques have been developed that would allow anyone who can leverage historical data regarding real estate prices to be able to predict it using techniques such as regression analysis, machine learning, and spatial analysis. This research aims to concentrate on two main aspects:

1. Identifying the best spatiotemporal data science technique to predict real estate hotspots and patterns in Yerevan, Armenia
2. Building, analyzing, validating, and visualizing results of the regression models that we have used in order to predict real estate prices.

Before we provide an explanation of what data we are using for the research, we would like to acknowledge GeoVibe for providing us with the data for the research and talk about the geographical location where the data is set for.

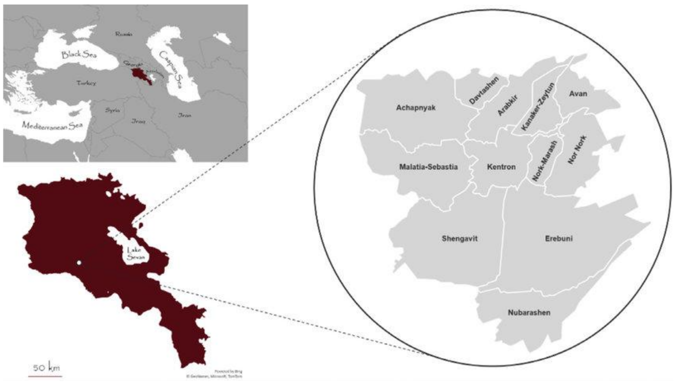

Fig. 1. The map of Armenia and the 12 districts of Yerevan[1]

In Fig.1. we can see the map of Armenia, which is the only city this project is interested in analyzing and studying the real estate market for. It is important to give some context of the region to understand better how we should approach this analysis, its longevity, and its relevance to the present and future.

The first important note to make here is that while machine learning and visualizations can give insight and help us with analysis, they can't predict the future for certain. Thus, this study intends to analyze the market drivers for real estate house pricing within the context of the last year of developments that have taken place in Yerevan, Armenia. The real estate market is always changing, and certain data can't be accounted for. Something that we haven't taken into account for this research is the metric of how the geopolitical situation of Armenia impacts the real estate market. For example, Armenia is in constant threat of engaging in war, and these scenarios are unlikely to predict and cannot be measured for certain what effects they would have on the real estate market. Overall, it is important to approach this paper within the current context and setting of the real estate market in Yerevan and analyze the data we can extract and predict as data scientists [6, 7].

The analysis and implementation of the approaches have been done by Python, Jupyter Notebooks, and ArcGIS. We have used Python for most of the data processing, cleaning, and analysis of the results from the models. ArcGIS was used in order to have 2D and 3D visualizations using the preprocessed data, building the models, and validating them on new points of data.

First, we begin by describing the data used in section II. It includes section II.A that covers the data cleaning process; detailed steps are taken to ensure data quality and reliability, and section II.B where we do data exploration for gaining insights, including relationships between variables. Section III focuses on the methods and results obtained; in this section, we use three models: Generalized Linear Regression(GLR), Geographically Weighted Regression(GWR), and Forest-Based Classification and Regression(FBCR). Section III.D is for evaluating and comparing the models' performance on a new dataset. Finally, in section IV we conclude by summarizing the key findings of this study.

II. Data

As part of our research on the Yerevan real estate market, we have data that has been scraped from three prominent real estate websites - list.am, estate.am, and real-estate.am. The GeoVibe team has been utilizing a script that has been extracting information such as id, price, square meters, height, and other relevant details from each website. This is the main data we use for this research in order to gain and provide valuable insights into the Yerevan real estate market, including trends, prices, and demand throughout time.

There are two main databases that the script populates

- Yerevan actual - current active house listing in the websites
- Yerevan historical - house price changes from Yerevan actual data, and listings that were removed from the websites

Arc-GIS was used in order to map the addresses to a specific longitude and latitude coordinates they belong to. By having such data, we wrote a script that provides more context and features that we will explore later for our models and visualizations.

We used Google Maps API to generate the fields

- The walking distance (in meters) from each location to their closest Metro
- The neighborhood they belong to.

We used the Nominatim package to generate the field

- Which district each house is settled in.

Thus, that leaves us with Yerevan actual containing houses that are not sold, and Yerevan historical containing houses that have been sold and houses that have experienced a change in price over time in the raw dataset that was provided to us by GeoVibe.

[1] Armenia Map

The timeframe of the data is starting from August 30th, 2022, up to March 13th, 2023, and the raw data provided to us by GeoVibe contained 729,498 rows of historical data and 51,893 rows of active listings. We performed data cleaning and preprocessing, including handling missing values, outliers, and duplicates, to ensure the data was suitable for analysis. However, we should note that the biggest assumption for the sold houses data is that a house was considered 'sold' when the listing was taken down. As a result, a few houses appeared as duplicates in Yerevan Historical. We try to handle this issue by removing the duplicate id’s which we explain further in the data cleaning section.

Overall, the collected data provides a comprehensive view of the Yerevan real estate market, and with the added features to the table that we were able to generate, it will allow us to further examine the spatial characteristics that have an effect on the price and location of our data.

A. Data Cleaning

The first important step was to ensure our data was as accurate, reliable, and suitable for analysis as possible. We performed quite a few data cleaning and preprocessing techniques on both of the datasets.

To begin with, we filter out any null values from the price column. This was necessary because the price is a critical component of our analysis, and any missing values could have affected our results. We also remove any duplicated IDs from Yerevan Historical, which arise from our assumption that a house is considered 'sold' when the listing is taken down. We keep only the most recent entry based on the sold_date column, which helps ensure that our data accurately reflects the sale of each house. As we will document further in our Methods and Results section (III), we have used built-in models in Arc-GIS, which handle missing and null values for each of our independent variables by completely leaving out that row in the model[2].

Outlier detection is a critical step in any data analysis project, as outliers can skew the results and lead to inaccurate conclusions. In our project, we conduct IQR outlier detection on the price and square_meters columns for each district in both Yerevan Historical and Yerevan Actual datasets [4]. However, we also check for outliers in every numeric variable in the dataset. We know outliers can arise due to human errors when the property is listed on the website. Therefore, we carefully examine each variable in our datasets for outliers. After this initial examination, we identify price and square meters as the variables with the highest potential for outliers and therefore focus our outlier detection efforts on these variables.

To identify outliers, we calculate the IQR (Interquartile Range) for each district in both Yerevan Historical and Yerevan Actual datasets. By using the IQR measure, we will identify the statistical dispersion lies in the data by taking the difference between the 75th (Quantile 3) and 25th (Quantile 1) percentile of the data.

$$IQR = Q_3 - Q_1 \tag{1}$$

$$Lower\ bound = Q_1 - 1.5 \times IQR \tag{2}$$

$$Upper\ bound = Q_3 + 1.5 \times IQR \tag{3}$$

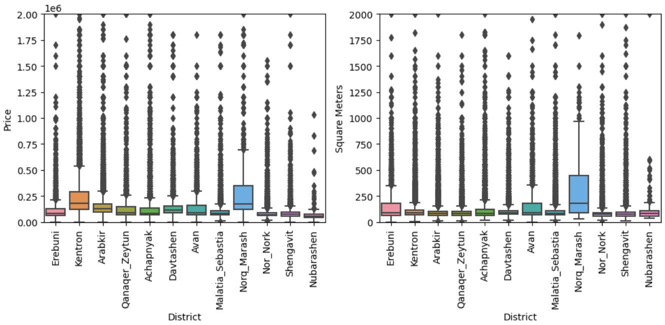

Fig. 2. The graph illustrates the boxplot of real estate prices and square meters in different districts of Yerevan. The boxes show the interquartile range (IQR) of the data, with the median value indicated by a horizontal line. Outliers beyond 1.5 times the IQR are shown as individual points.

Any point that lies outside of the whiskers is considered an outlier and subsequently is removed from the data

It is worth noting that the bottom whisker of each boxplot is very close to zero, indicating the presence of a significant number of low-priced or small-sized properties. To address this issue, we apply an IQR outlier detection method and remove extreme values from the dataset. Specifically, we remove any price or square meter values that fall below the lower bound defined as the first quartile minus 1.5 times the IQR as outlined by formula (2), however from the graph, we can observe the lower whiskers of the boxplot are sometimes very low so we set a ‘default’ value of \$10,000 and $50m^2$ for price and square_meters in case the lower bounds fall any lower than those values.

Finally, we perform data type conversions and data standardization to ensure consistency across our datasets. We convert some columns from string to numerical data types, standardize the format of some columns to ensure consistency across the datasets and remove unnecessary columns.

[2] Documentation Regarding Missing Values

B. Data Exploration

Let's have a closer look at the attributes of our data and scatterplot matrix and explore the relationships between them.

TABLE I. Description of Every Attribute in Dataset

Field Name	**Description**	**Data Type**
backup_status	Status of houses (sold or not_sold)	String
id	Unique id (categorized by website)	Numeric
price	Price of house	Float
rooms	Number of rooms	Integer
square_meters	Living space size	Float
address	Address of house	String
sold_date	Date of sale	Date
furniture	Is there furniture (1: yes, 0: no)	Boolean
renovation	Categorical variable for renovation of house	String
price_per_meter	Price per meter of house	Float
floor	Floor number of house	Integer
building_floor	Number of floors of building	Integer
height	Height of house	Float
bathroom	Number of bathrooms	Integer
download_date	Date the data is scraped from websites	Datetime
x	Longitude of house	Float
y	Latitude of house	Float
closest_metro	Closest metro to house	String
walking_distance_to_metro(m)	Walking distance to nearest metro	Integer
district	District house is located at	String
latitude_jittered	Small deviation from latitude of house	Float
longitude_jittered	Small deviation from longitude of house	Float
neighborhood	Neighborhood house is located	String

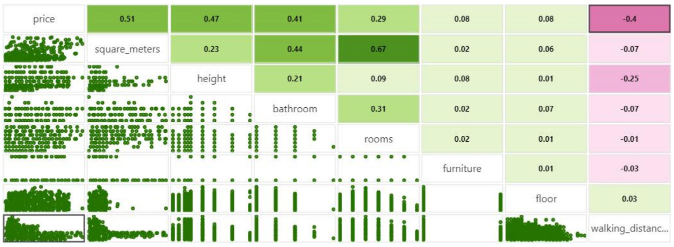

Fig. 3. Sold houses Correlation Matrix and Pearson's r value. A value of Pearson's r close to 1 or -1 indicates a strong linear relationship, whereas values close to 0 indicate a weak linear relationship.

In Fig. 3, we can see that price has a strong positive linear relationship with square_meters, being 0.57. This implies that as square_meters increases, the price also increases and vice versa. A negative r value indicates an increase in one variable and a decrease in another, as we can see from the price, walking_distance_to_metro(m) pair. Pearson's r value also shows some weak linear relationship between furniture and price (0.08) and floor and price pairs (0.08).

From the above analysis, we find that square_meters has the strongest correlation to our target variable(price). Other variables that show a strong relationship with each other can cause problems if we put them in the same linear model with square_meters, as they could show multicollinearity, meaning both independent variables are correlated, and will result in less reliable statistical inferences.

Thus we can say that running a Linear Regression model between two or more predictor variables' prices has to be evaluated so that it will not result in redundant information to the model, which may lead to unstable and unreliable estimates of the coefficients. Multicollinearity can also have an effect on the coefficients of the model variables hence interfering with the power of statistical tests. In Fig. 3, we can see a high correlation between square_meters and rooms (0.67), which can cause a case of multicollinearity if both of the variables are included in the model. This is something we examine further in the methods section.

Now that we have aggregated the sales to detect patterns for sale, a question can arise in the form of if these patterns are actually statistically significant or if they have been caused by pure randomness of the variance. That is why there is a need to understand the spatial correlation between each of these houses to understand if we have spatially similar houses cluttered next to each other.

For this, we use a statistical measure in ArcGIS called Spatial AutoCorrelation (Global Moran's)[3]. Global Moran's is a statistic that quantifies the degree to which similar values of a variable are clustered together in space. In a way, it measures a degree of similarity between each observation and its neighbors; then, it compares the original layout of points to a version of the data if they were randomly distributed in the given space. Global Moran's will take two things into account when measuring spatial autocorrelation.

[3] Spatial Autocorrelation (Global Moran's I)

1. Feature Locations - which refer to the longitude and latitude coordinates of the houses in our data
2. Feature Values - which refer to all of the attributes that are non-categorical in the dataset

Global Moran’s I uses inferential statistics to construct the following null hypothesis.

H_0: *values associated with features are randomly distributed*

The formula to calculate the Global Moran Index is the following[4]

$$I = \frac{n}{S_0} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j} z_i z_j}{\sum_{i=1}^{n} z_i^2} \qquad (4)$$

$$S_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j} \qquad (5)$$

Where:

- n - number of spatial units
- z_i and z_j - deviation of the attributes feature from it’s mean between units i and j
- $w_{i,j}$- measure of the spatial proximity between units i and j
- S_0 - the aggregate of the spatial weights calculated by formula (5)

Then the z-score in order to check the significance of the result is calculated by:

$$z = \frac{I - E[I]}{\sqrt{V[I]}} \qquad (6)$$

Now that we know the math behind Global Moran’s I we can test it on our data using ArcGISs’ Spatial Autocorrelation (Global Moran’s I) tool.

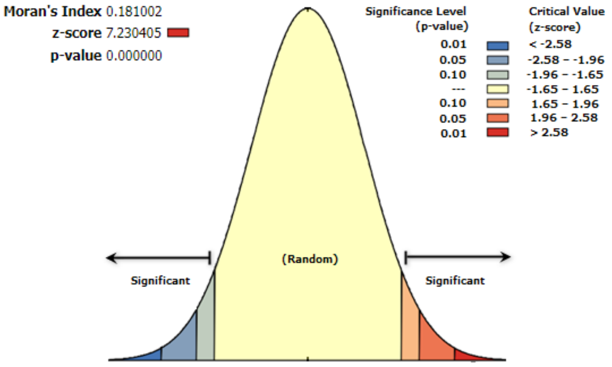

Fig. 4. Global Moran’s Spatial Autocorrelation

From Fig. 4, we can see Moran's Index value of 0.181 which suggests that there is a moderate level of clustering in the data. This implies that houses with similar attribute values tend to be located near each other in space. The z-score of 7.2 tell us that the result is significant and that we should reject the null value that the values associated with features are randomly distributed. The p-value of 0 further supports this, as it is less than the typical level of significance (0.05) used in hypothesis testing. Therefore, we can reject the null hypothesis and conclude that there is a statistically significant positive spatial autocorrelation present in the real estate house sales data.

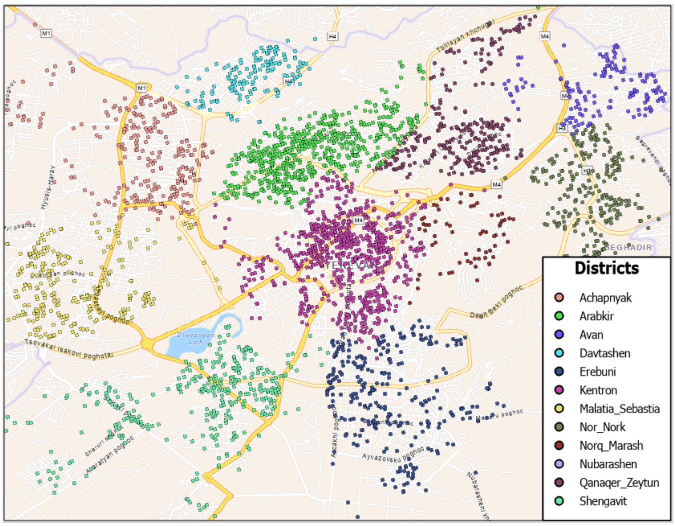

Fig. 5. Sold houses grouped by their district

In Fig. 5, we can see a spatial distribution of all the points of the sold houses data, which are grouped by the district they belong to. This graph can’t tell us much about which of the regions are considered hotspots (except for making an obvious assumption that the center of the city would be a popular hotspot for buying a new house). To help decide where to open your business in terms of a hotspot, we have analyzed sold houses data and determined the most popular neighborhoods in Yerevan. To gain more meaningful insight, we have created a grid of hexagon bins that covers Yerevan and use this grid to aggregate sold houses. Then, we symbolize the result layer to determine which areas have the most sales.

Additionally, we can make this analysis spatiotemporal in terms of incorporating the metric of time with it as well. The spatiotemporal analysis involves analyzing data that has a temporal component (i.e., time) in addition to the spatial component (i.e., location). With that in mind, we visualize Space Time hexagons in 3D using ArcGIS to identify trends and patterns in housing demand across different regions. Additionally, it can help us detect clusters of high or low house sales, which will indicate areas of high or low housing demand or property values.

[4] How Spatial Autocorrelation (Global Moran's I) works

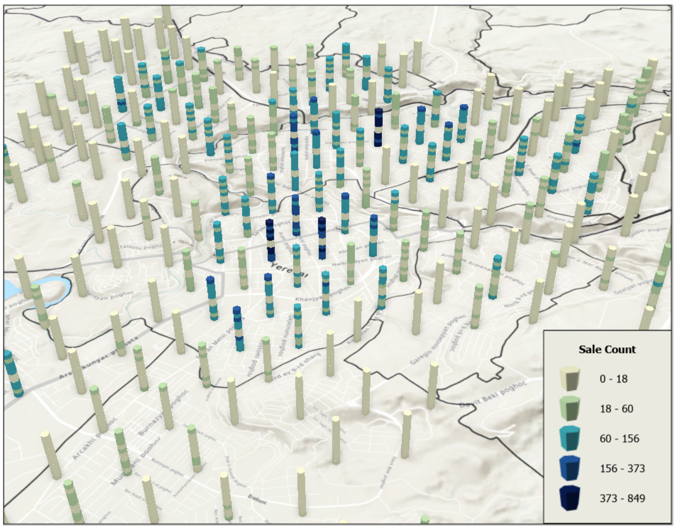

Fig. 6. Spatiotemporal distribution of house sales in Yerevan using Visualize Space Time Cube in 3D tool in ArcGIS[5]. The sale count interval for each class is calculated by minimizing the sum of the square of the number of elements in each class.[6]

For Fig. 6, the map for Yerevan is divided into hexagons, each occupying $800m^2$ of the area and the time interval being accounted for on a tri-weekly basis. With this information, we can detect spatiotemporal trends and patterns in the district of Kentron and Arabkir. This data can lead us to make informed decisions about real estate investments, urban planning, and business investments in the regions where we can see a constant pattern of growth or maintenance of that intensivity for the market desire throughout time. The fact that most of the sales across time have been in the center of the city and the areas close to the center suggests that these areas are more economically active compared to other parts of the city. This further validates our approach of trying to train and test our models on spatial characteristics is the right approach so far. For example, we observe that sales are concentrated in certain areas of the city, with each hexagon representing different levels of intensity or patterns for sale; that is why it is important to run our models based on different spatial areas for the city of Yerevan. To further analyze the spatiotemporal patterns in the wholesale sales data, we can apply hot spot and cold spot analysis. This analysis can help us identify areas with statistically significant high values (hotspots) or low values (coldspots) of wholesale sales and can provide insights into areas of the city that are more economically active compared to others [2].

We can achieve the following results by running the 3D Space-Time Cubes to be triggered by the hot and cold spot display theme[7].

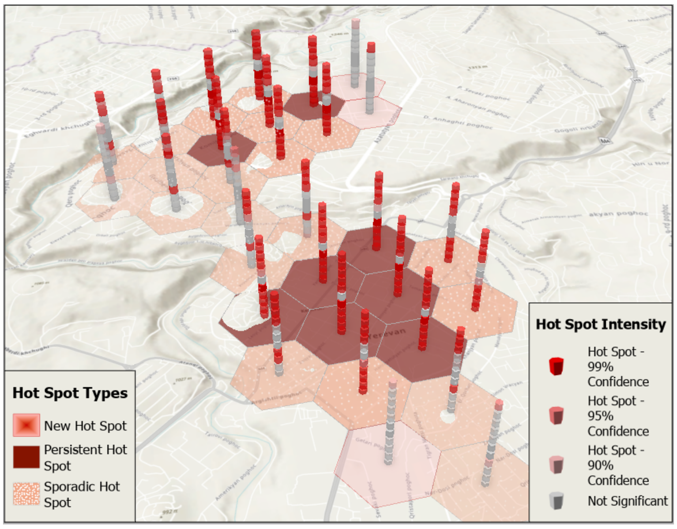

Fig. 7. Spatiotemporal hot spot detection. The legend on the right represents the confidence level for the hotspots that were detected, and it is based on the z-score and p-value of each feature, which in this case is the sale of the houses[8]. The legend on the left indicates the hot spot analysis for our hexagons[9]. The hexagons that did not show any significance were left out of the graph so that we can only the ones that showed some level of significance throughout time.

As the analysis for Fig. 6 and Fig 7. further supports the claim that hotspots of house sales are concentrated in certain neighborhoods or areas of the city, which are Kentron and Arabkir. In areas considered new hot spots, only the most recent month (the uppermost hexagon bin on the column) is considered a hot spot. Sporadic hot spots alternate between being hot spots and not being hot spots. In the center of Yerevan, areas are hot spots during every interval, making them persistent hot spots. In the case of the new hotspots, these regions could be interesting points for new business owners and real estate agents to review and invest their time and resources in that area. These are areas that have newly constructed buildings which would create the need for businesses such as supermarkets to be opened in that area. However, from the image, we can see three main spots for this, which can also imply that the urban planning developments for the past year in Armenia have not significantly impacted how the real estate market is developing. This is especially in the case of continually seeing the exact center of Yerevan as a hotspot.

[5] Visualize Space Time Cube in 3D
[6] Geometric Interval
[7] Hot and cold spot results
[8] Hot Spot Analysis (Getis-Ord Gi*)
[9] Emerging Hot Spot Analysis

III. Methods & Results

In this section, we further explain the regression models that we have used and the results that we obtained for each of them. Our main goal is to develop accurate models that could predict house prices by taking into account spatial and temporal trends in the data. To achieve this, we use several modeling techniques, including

1. **Generalized Linear Regression (GLR)**: Covered in detail in section III.A
2. **Geographically Weighted Regression (GWR)**: Covered in detail in section III.C
3. **Forest-Based Classification & Regression (FBCR)**: Covered in detail in section III.D

All of these models will be run using ArcGIS geoprocessing tool solution, and in case any of the independent variables in the model has missing values, the model just leaves them out of the analysis, as mentioned previously. Before we explain how each model would work and the results we obtain for them, we consider two approaches to gain the best possible results from our models.

1. **Baseline approach**: Developing a baseline model for each that would be representative of the test results and how they would need to be optimized for the final product
2. **Optimized approach**: Developing an optimal model based on the following
   a. Checking for normality of the residuals for the predicted price
   b. Checking for multicollinearity between the independent variables used to predict the price
   c. Independent variable percentage importance to the models
   d. Clustering the city into zones (regions) that represent similar traits so they can be grouped together, and models will be fit based on each group/zone.

One of the important metrics that we use to evaluate model performance is the Coefficient of Determination (R^2), which shows how well the model fits the data and its prediction for future observations. R^2 ranges from 0 to 1, with 1 indicating perfect fit and 0 no fit at all.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2} \qquad (7)$$

- n - the number of observations
- y_i - the actual value of the dependent variable
- $\hat{y}_i$ - the predicted value of the dependent variable
- $\bar{y}$ - the mean of actual values of the dependent variable

A. Generalized Linear Regression

Generalized Linear Regression (GLR) is a statistical method native to ArcGIS that models the relationship between a dependent variable and one or more independent variables. GLR can be a more flexible form of a traditional linear regression model, which can be utilized to predict variables that are not continuous or unbounded. However, as we have predicted a continuous variable for the price, it ultimately functions just like a linear regression model[10] that you would come across in a Python package.

For developing the baseline approach of the model, we look at the variables that show a high correlation with price. From Fig. 3, we can see that includes square_meters, bathrooms, height, walking_distance_to_metro(m). However, we need to also be aware of cases of multicollinearity that could exist, thus looking at the same figure, we can see that all of these independent variables have low relation to each other, and reviewing the VIF scores that we obtained for them, we can conclude that these variables are good predictors to include in the baseline model as of now.

A.A.1 Baseline GLR

Next, we further explore property characteristics and sale prices of houses using exploratory regression. With the ArcGIS Spatial Statistics geoprocessing tool, we create a Generalized Linear Regression model. As we predict, a continuous variable(price) Gaussian model is chosen. One of the outputs of this tool is a standardized residual map (Fig. 8).

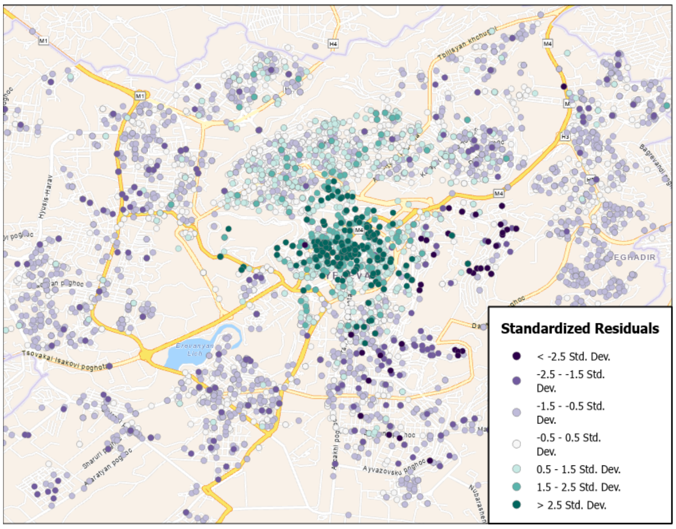

Fig. 8. Baseline GLR price overestimation and underestimation. Dark green and dark purple are indicators of a large mismatch between the predicted price and the actual price of houses.

The map clearly shows that the model underestimates houses in the central area of the city (Kentron district), whereas areas around it show both strong and weak fluctuations. There is a need to detect spatial autocorrelation present in the data, meaning that houses in close proximity to each other may

[10] Generalized Linear Regression

have similar prices, which we further discuss III.C & III.D. But for now, we can see that the house price predictions in the center of the city are almost all 2.5 standard deviations above the actual price of the house listing. This obviously resulted in most of the underestimations being in the regions outside of the city center.

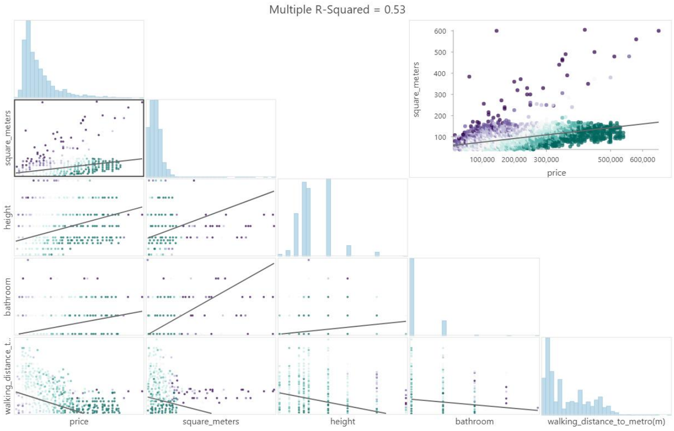

Fig. 9. On the top-right, we have a baseline GLR scatterplot for the actual price and square meter area of the house, along with a linear regression line fit to the data. The darker colors here represent the overestimations and underestimations. The left half of the figure represents the scatterplot between each and every variable, but in this case, price and its interaction with the other variables are more important to look at.

To analyze the data points GLR model also provides a chart of the same data points. Ideally, all points should be close to the line, as the closer the points are to the line, the stronger the relationship is between variables. Green points in Fig. 8 and Fig. 9 tell that actual house prices are higher than the ones predicted by the model, whereas purple indicates that actual prices are below the predicted ones. The GLR model provides R^2 value of 0.53, meaning only 53% of the dependent variable(price) can be explained by the independent variable (square_meters, height, bathroom, walking_distance_to_metro) in this model. A reason why predicting house prices in the city center is more difficult is that the city center tends to have a higher concentration of unique properties, such as historical buildings or commercial spaces that have been converted into residential properties, and as we do not track the conditions of our buildings, none of our models would take such a variable into account.

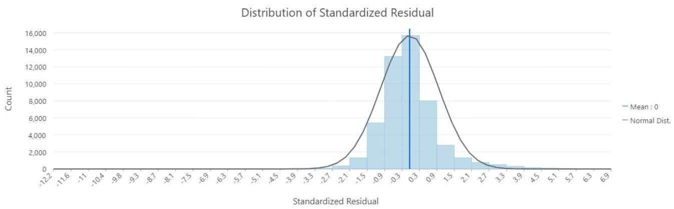

Fig. 10. Standardized residuals of baseline GLR based on the difference between the actual price of the house and the predicted price of the house, divided by the estimated standard deviation of the residuals.

As the map illustrates in Fig. 8, even though most of the strong underestimations are concentrated in the city center, the majority of house data points in all other districts are mostly light purple, indicating a slight deviation from the correct prediction. The histogram confirms this prediction as we can see the highest bin includes negative values as it starts from -0.3, and the second highest bin in Fig. 10 includes the interval between -0.9 to -0.3 of standardized residual.

To further explain why we don't obtain optimum results, especially for the city center, which is subject to more fluctuations in market conditions, and the model doesn't take into account any spatial features to train it. These factors can be difficult to predict and may require more nuanced modeling techniques. However, one way we can improve this same model is by trying to run the same GLR with the context of the zone each house belongs to, thus adding a spatial element to the model. We tried three different ways of dividing each house into the zone it belongs.

1. Neighborhood: grouping each house by the neighborhood they belong to (total 80 neighborhoods)
2. District: grouping each house by the District they belong to (total 12 districts)
3. Multivariate Clustering[11]: a geostatistical tool in ArcGIS used to group similar geographic areas or spatial units together based on multiple variables or attributes (a total of 4 clusters)

With this, we experimented running the model with a sub-zone of each zone and ended up getting the best results for the zone that was grouped by each neighborhood. Review Appendix 4 to review improvement over baseline approach.

A.A.2 Neighborhood GLR

We then perform GLR in every neighborhood. For this, we use the ModelBuilder in the ArcGIS[12] tool and specify the iterator model as GLR. Overall model quality for each of the neighborhoods is greater than the baseline result; only 19% of neighborhoods performed worse.

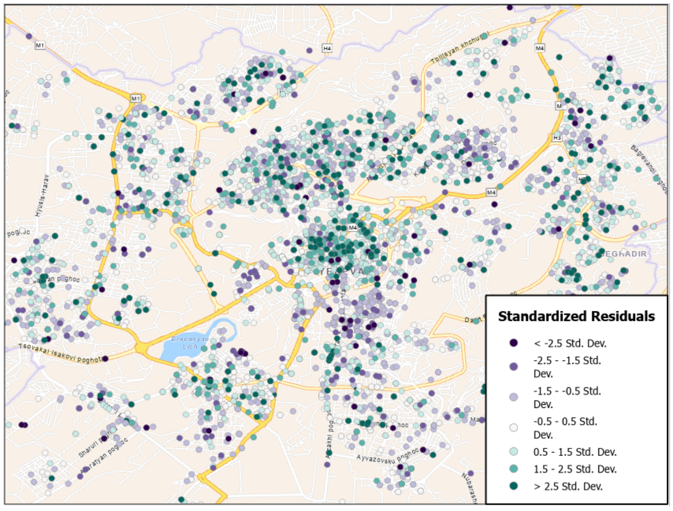

Fig. 11. Neighborhood GLR price overestimation and underestimation

[11] Multivariate Clustering

[12] Model Builder

Considering both standardized residuals from Fig. 10 and the outputs of the comparison of the two models, we can claim that there is a significant improvement in the house price prediction for the neighborhood GLR model. Fig. 11 shows us that although there is more variation of the darker colors in the regions, the higher R^2 of 0.66 and better distribution of the residuals implies that this model works better for each segmented neighborhood compared to the approach of the one with the baseline. This makes sense as the neighborhood GLR will try to fit the model on a subset of the 80 neighborhoods compared to the one with the baseline approach. Additionally, the residuals for the baseline are far higher than the neighborhood approach. Having the data from the residuals and the predicted price, we calculate the mean squared error (MSE) for the prediction.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 \quad (8)$$

n - the number of observations in a dataset

Y_i - the actual value of target variable(price)

$\widehat{Y}_i$ - the predicted value of target variable(price)

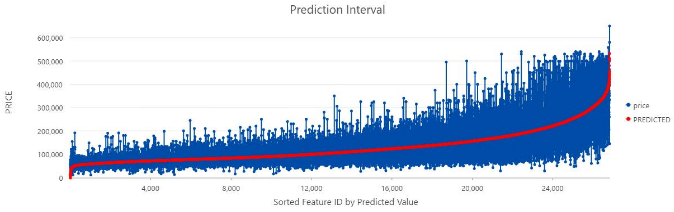

Fig. 12. Neighborhood GLR price prediction interval

Fig. 12 shows the uncertainty bounds of the prediction, with the red line being the actual prediction and the blue lines representing how far the prediction deviated away from the actual price. The x-axis line represents an ‘id’ that increases incrementally based on how high or low the predicted price is. We can see the bounds formed by the actual price sporadically start to widen for homes priced at more than $150,000. This trend is due to the lower sample size for the more expensive homes. For homes more expensive than $150,0000, the bounds become bigger and more frequent, as there are even fewer samples in this price range. This plot is quite useful in showing the uncertainty relating to the predictions for the training sample.

B. Spatially Varying Relationships

Next, we use Geographically Weighted Linear Regression (GWR) and Forest-Based Classification and Regression to model house prices. GWR’s main idea is that the relationship between the dependent variable and the independent variables can vary across different locations, and therefore a single global regression model may not be appropriate. Instead, GWR estimates a local regression model for each location that captures the local spatial relationships between a dependent variable and independent variables.

Before running the model, it is a best practice to run the Local Bivariate Relationships tool in ArcGIS, which is an entropy-based approach to discover spatial relationships to check whether statistically significant spatial relationships exist between the variables. If a strong relationship exists between variables when randomizing data, entropy does not increase; however, if there is no significant relationship when randomizing data, entropy increases[13].

Running a local bivariate relationship model before GWR can be useful for several reasons, such as

1. A simple way to visualize the spatial patterns of the relationship between the variables, which can help to identify areas with strong/weak relationships.

2. It helps identify potential outliers or influential observations that may need to be handled in the GWR model.

In our case, this is run to check what type of relationship we have between price and the two highest correlation values we observe in Fig. 3. (square_meter, height) to check their relationship.

B.A.1 Local Bivariate Relationship Analysis Between Price & Square Meter Area of the House

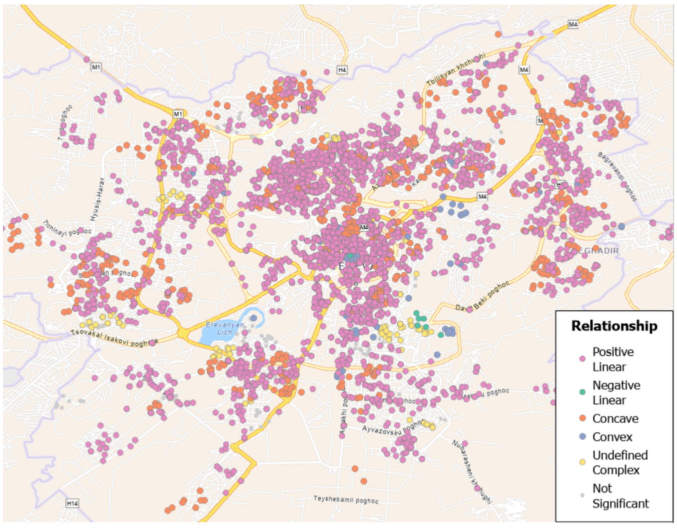

Fig. 13. Local bivariate relationship between price and square_meters

From Fig. 13, we see a good amount of data points with positive linear correlation for price and square_meter. Sometimes, data points overlap, making one of the relationships look dominant, whereas, in reality, that is not true. Fortunately, ArcGIS also provides a section with more detailed information on results.

[13] Local Bivariate Relationships

TABLE II. Categorical Summary of Relationship Between Price and Square Meter Area of the House

Description	# of features	% of features
Positive Linear	24310	46.34
Negative Linear	0	0.00
Concave	26824	51.13
Convex	1307	2.49
Undefined Complex	8	0.02
Not Significant	9	0.02
Total	52458	100

From Table II we can see that in reality, Concave Relationship is more governing than the Positive Linear Relationship as indicated in Fig. 13, with 51.13 % and 46.34 % respectively.

In the GWR model, the regression equation is estimated separately for each data point, allowing for the relationship between the variables to vary based on local conditions and factors. This means that GWR can work with both Linear and Polynomial(Concave) relationships and capture complex associations.

TABLE III. FDR Detection of False Positives

Description	# significant	% significant
Without FDR	52450	99.98
With FDR	52449	99.98

On this page in ArcGIS, you can also find a False Discovery Rate(FDR) detection table, which measures the proportion of false positives among all significant results. The results in Table III show that without controlling for FDR number of significant data points is 52450, which corresponds to 99.98% of overall data points, whereas with FDR controlling, the number decreased by one point to 52449. Indicating that one relationship is identified as a false positive and was not actually significant. This small difference shows that the overall results of the local bivariate relationship model were not considerably influenced by the control for FDR and that the relationships are likely to be robust and reliable.

In comparison to the abovementioned pair, Local Bivariate Analysis is also run on the Price(dependent variable) and Height(explanatory variable) pair.

B.A.2 Local Bivariate Relationship Analysis Between Price & Height of the House

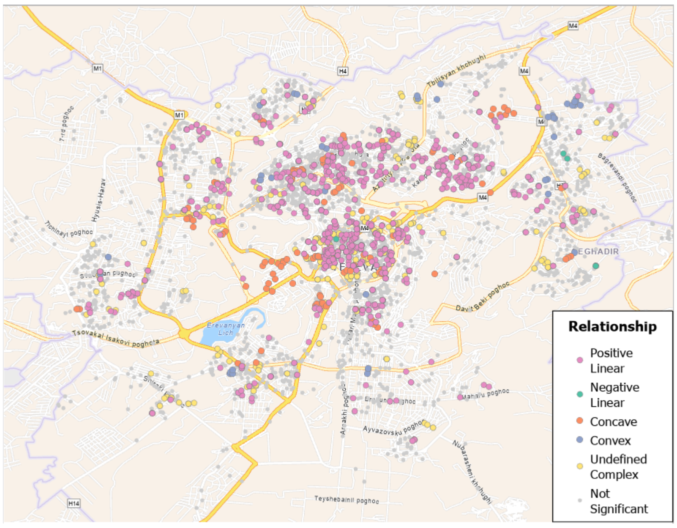

Fig. 14. Local bivariate relationship between price and height

The Local Bivariate Analysis mostly shows no significant points, which tells that the model did not find a statistically significant relationship between the two variables. From Fig. 14, we acknowledge the fact that in Yerevan city, most of the house prices are not directly related to their height. Kentron and Arabkir districts majorly include the points which are in a Positive Linear Relationship. And the following table, Table IV, demonstrates the ratio between the type of relationships, excluding the possibility of inaccurate representation of data in Fig. 14.

TABLE IV. Categorical Summary of Relationship Between Price and Height of the House

Description	# of features	% of features
Positive Linear	7085	14.15
Negative Linear	22	0.04
Concave	2032	4.06
Convex	556	1.11
Undefined Complex	1425	2.85
Not Significant	38964	77.80
Total	50084	100

Table IV provides analyzed data of 50,084 data, majority of which found not significant (77.8%), indicating little or no evidence of a relationship in those areas. Among the significant ones, the positive linear correlation is the second with 14.15%. This shows that as the height of the house increases, the price also tends to increase, in contrast to the negative correlation, which is merely 0.04%, where a height increase leads to a price decrease.

TABLE V. FDR Detection of False Positives

Description	# significant	% significant
Without FDR	21562	43.05
With FDR	11120	22.20

As opposed to the previous Local Bivariate Relationship Analyses, this one makes use of False Discovery Rate detection a lot more. According to Table V, the analysis without FDR detected a total of 21562 significant results, whereas FDR is controlled, the number decreases to 11120, suggesting that a huge number of significant results were, in fact, false positives.

C. Geographically Weighted Regression

We learned from analyzing the results from the baseline GLR and neighborhood GLR that having the baseline assumption of the relationship between the dependent and independent variables being constant across the entire study area is not ideal and certainly does not lead to the best result for the model. We can definitely state that the effect of independent variables on the dependent variable may be different in different locations. For that, ArcGIS has a model called Geographically Weighted Regression (GWR) that takes into account the spatial variation in the relationship between the dependent and independent variables[14].

GWR enables this by having a kernel function used to define a neighborhood around each data point. The kernel function determines the weight of each neighboring data point based on its distance from the focal data point. The neighbors option specifies the number of neighboring data points to use in the local regression calculation. When using GWR, one of the most important parameters to take into account is the neighborhood(bandwidth), as it controls the degree of smoothing in the model. We determine the possible neighbors option by using Golden search[15], which is a parameter technique to find the optimal value of the bandwidth in GWR within ArcGIS Pro. After optimizing the bandwidth, GWR can more easily deal with underfitting and overfitting. And Golden Search does so by selecting different values for bandwidth parameters and comparing the results of the GWR model based on the Akaike Information Criterion(AIC), Bayesian Information Criterion(BIC), or Cross-validation(CV). In our case, AICc is used, which is a variant of AIC and stands for Akaike Information Criterion corrected. It includes a correction factor for small sample sizes helping to reduce the overfitting and inaccurate model selection that AIC can introduce with small sample sizes.

$$AICc \ = \ \frac{AIC + (2k * (k+1))}{(n-k-1)} \qquad (9)$$

- k - the number of model parameters
- n - the sample size

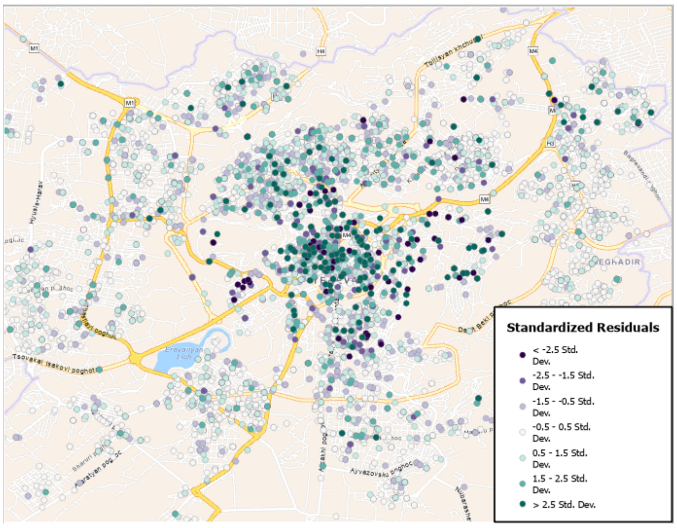

Fig. 15. GWR price overestimation and underestimation

Similar to GLR model, the GWR also underestimates the houses in the Kentron district, yet reduces the number of underestimation and overestimation for house prices in other districts surrounding it significantly(Malatia-Sebastia, Nor-Nork, Shengavit, Davtashen, Nubarashen, etc.). Compared to the model that is run with GLR by neighborhoods, GWR with Golden search makes most of its overestimations and underestimations in the center of the city, and it manages to perform better in terms of its residuals by obtaining a R^2 value of 0.75 compared to the 0.66 for GLR.

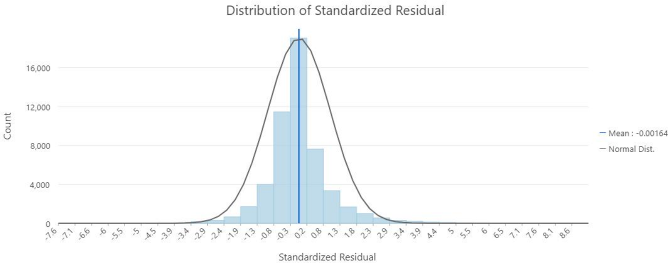

Fig. 16. Standardized residuals of GWR based on the difference between the actual price of the house and the predicted price of the house, divided by the estimated standard deviation of the residuals

Most of the standardized residuals are close to 0, indicating fewer overestimations and underestimations(fewer darker colors), and the model is generally well-fitted to the data compared to the GLR model in Fig. 10. Additionally, the bars that are next to the 0 lack significantly from the density. This can suggest that the GWR model has better spatial prediction capabilities than the GLR model, as the spatial dependence in the data is taken into account in the GWR model.

[14] How Geographically Weighted Regression (GWR) works

[15] How Golden Search Works in GWR

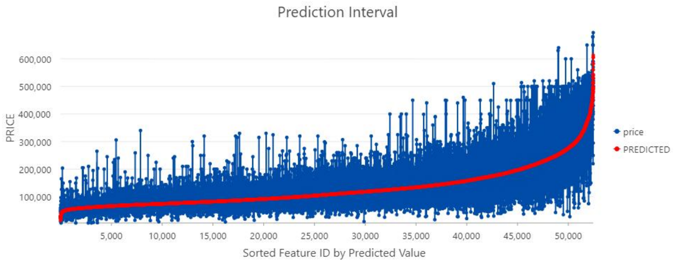

Fig. 17. GWR price prediction interval

Fig. 17 shows the uncertainty bounds of the prediction, with the red line being the actual prediction and the blue lines representing how far the prediction deviated away from the actual price. The x-axis line represents an ‘id’ that increases incrementally based on how high or low the predicted price is. Compared to the neighborhood GLR model, the GWR model makes sporadic overestimations and underestimations throughout the graph. This is most likely due to the fact that regardless of what the Golden search identified as the best possible number of neighbors to take that are similar to each other, there is still big variability across those numbers of neighbors. Thus, we can say that Yerevan is a city where you would encounter a lot of price variability in houses from the same exact building in some cases, even though they share quite similar attributes with each other. Additionally, as GWR dismisses any slight chance of multicollinearity, we are limited to running the model only in the context of the square_meters value. As we observed from section III.B and Fig. 3, height and square_meters would present multicollinearity with each other thus, we leave the height component out of the context from the model.

D. Forest-based Classification & Regression

We have a dataset containing a lot of independent variables, we want to incorporate them into a single regression model. The FBCR[16] model is not affected by multicollinearity and can work with spatial and non-spatial variables. It creates models and generates predictions based on the random forest algorithm, which is a supervised machine-learning method developed by Leo Breiman and Adele Cutler. An ensemble of many decision trees is created based on explanatory variables, and each tree is trained using a random subset of house data and explanatory variables. As individual trees on their own are prone to overfitting, the model uses the entire forest to generate final predictions, solving this issue. By default, FBCR takes out 10% of training data for validation, meaning once the data is trained, you can check how well it predicts the validation data. Another method we use to check the quality of predictions is Out of Bag(OOB) errors and a Variable importance chart.

D.A.1 Baseline FBCR

We have the possibility to run the model over 20 different variations and then examine the results for each. When the FBCR model is finished running, it will examine the iteration that obtained the best residuals and R^2 use that to train the 10% of data that it left out for testing and validation. We can further examine the R^2 values that the model obtained over each iteration of the run.

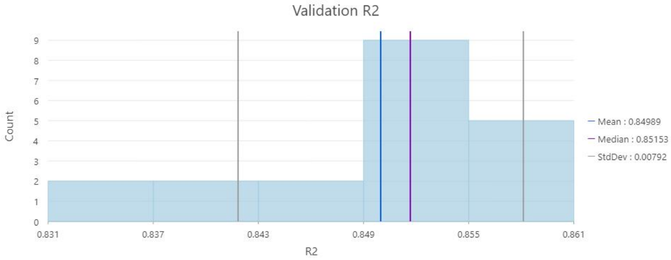

Fig. 18. Validation of R^2 based on the 20 times the baseline FBCR model runs.

The standard deviation of R^2 is relatively small (0.00792), meaning R^2values are consistent and include a little variation between observations. Mean and median values are close to each other, approximately 0.085, which indicates that R^2 values are almost symmetrically distributed. We can also observe that the highest bar is also around the mean and the median, resulting in a good performance of prediction. Overall the model seems to be stable, as the validation interval is between 0.831 and 0.861 on the 20 runs. (Fig. 18)

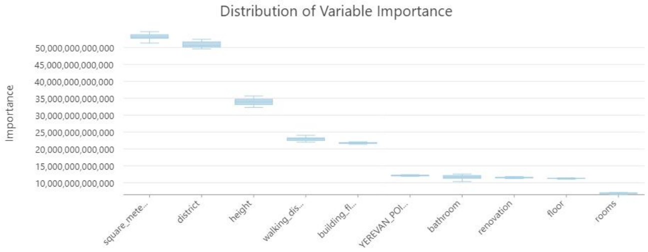

Fig. 19. Importance of each variable based on the 20 runs for the baseline FBCR model validation

As mentioned above, one of the other factors that affect the prediction quality is the importance of variables. As we can see from Fig. 19, square meters, district, height, walking distance, building floor, and Yerevan POIs have the highest importance. It should be noted that the value of importance shows the number of tree splits based on a variable, indicating an impact of a variable on the final result.

D.A.2 Reduced FBCR

One way to improve the model is to reduce the FBCR variables to only include the influential ones. Since removing the non-significant ones reduces the chance of randomly selecting them for a particular tree and hinder the important ones to provide better results.

[16] Forest-based Classification and Regression

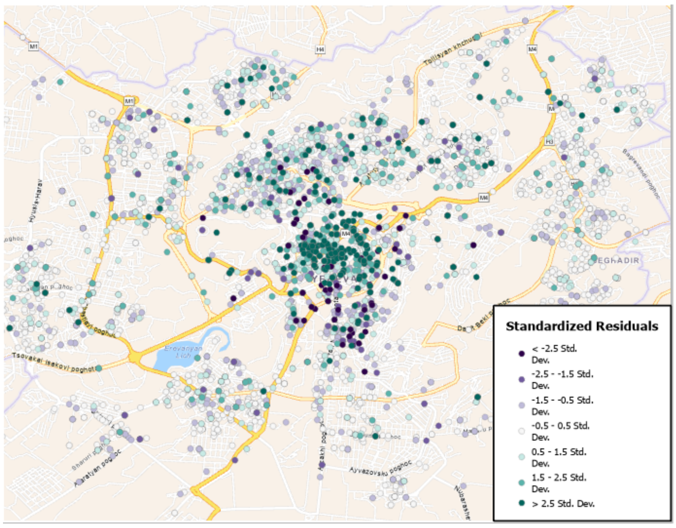

Fig. 20. Reduced FBCR price overestimation and underestimation

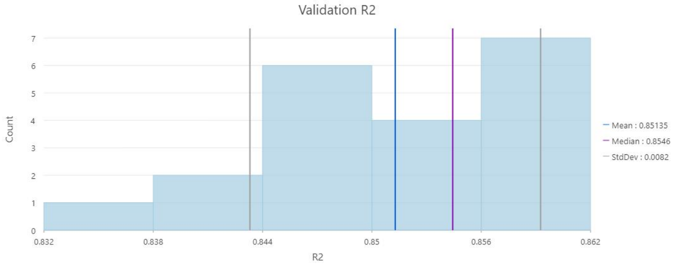

Fig. 21. Validation of R^2 based on the 20 times the reduced FBCR model runs

From the histogram, we observe the mean being 0.85135 and median 0.8546 pointing to a slightly left-skewed distribution, with the majority of distributions falling upper part of the histogram, implying the better performance of $R^2$for overall model. As the standard deviation is approximately 0, we can say that most of the outputs are relatively close to the mean. Fig. 21

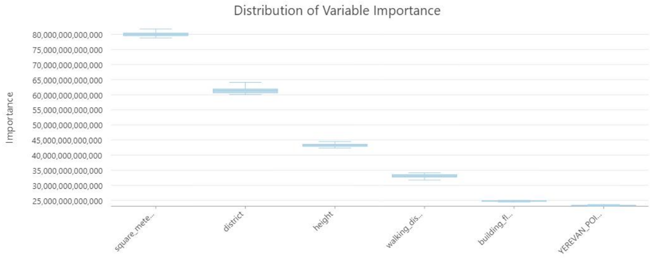

Fig. 22. Importance of each variable based on the 20 runs for the reduced FBCR model validation

If we look at the values from Fig. 19, we can see that all the values that are mentioned in the Fig. 22 have smaller values since, as mentioned earlier, the features that are not as important can hold back the others. By comparing graphs, it is clear that variables have significantly increased:

- square meters by $26.000.000.000.000$
- district by $10.000.000.000.000$
- height by $10.000.000.000.000$
- walking distance by $11.000.000.000.000$
- building floors by $3.000.000.000.000$
- Yerevan POIs by $12.000.000.000.000$

TABLE VI. FBCR Model Out of Bag Errors

Number of Trees	500	1000
MSE	1199739661.12	1183562173.8
% of Variation Explained	81.174	81.428

Table VI shows Out of Bag(OOB) Errors, including:

- Mean Squared Error(**MSE**): A metric that calculates the average squared difference between actual and predicted values in regression models. It is commonly used to evaluate the accuracy of regression models, with a lower value indicating better performance. Review formula (7)
- % of Variation Explained: A measure indicating the percentage of variation in the dependent variable(price) explained by the independent variables(square meters, height, etc.).

Using 1000 decision trees for the FBCR model resulted in a lower MSE value and a slightly higher percentage of variation explained compared to using 500 trees. This suggests that using more decision trees improved the performance of the model. However, to make sure that the model is not overfitting the training data, we also need to analyze the validation dataset (the excluded 10% of training data).

TABLE VII. FBCR Diagnostics

Data	**R-Squared**	**p-value**	**Standard Error**
Training	0.940	0.000	0.001
Validation	0.819	0.000	0.005

For both Training and Validation subsets, R^2s perform better than GLR and GWR models. The remaining two diagnostic measures: P-values and standard errors of both datasets, are closer to 0, showing regression coefficients are significant at a high level of confidence, and the model is highly accurate in estimating coefficients of datasets. As the scores for validation data are not substantially lower than training data, we can confirm that the model is generalizable, meaning it can predict unknown data points with high accuracy.

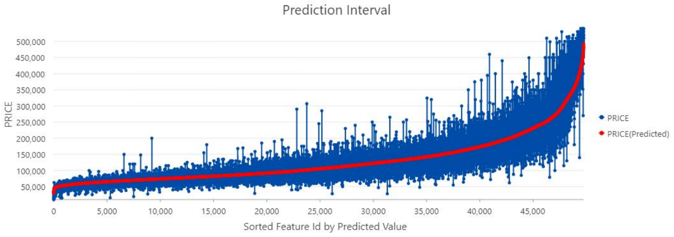

Fig. 23. Reduced FBCR price prediction interval

Compared to the previous models in Fig. 23, we can see that FBCR makes the best possible predictions out of all of them, especially for the lower values of the prices where the sample of the data is higher. It only starts to increase as the value of the price also increases, which can be a result of now having enough data for houses with higher prices.

For the FBCR model, it is crucial to investigate the spatial distribution of uncertainty. It will calculate a 90 percent prediction interval for each predicted value. During the process, it uses P95 and P05 parameters which represent the upper and lower bounds for prediction. This means for new observations; you have 90 percent confidence that they will fall within a specific range, given the same explanatory variables.[17]

The spatial distribution of uncertainty can help us understand the areas that are less reliable for the model and more prone to abrupt changes so that informed decisions are made for further analyses.

The uncertainty level is calculated with the following formula:

$$Uncertainty \ = \ \frac{P95 - P05}{P50}$$

- P95 (95th percentile) - Upper bound for prediction interval
- P05 (5th percentile) - Lower bound for prediction interval
- P50 (50th percentile) - Median for prediction interval

This provides a measure of the relative spread of data around the median, which is not affected by specific units of measurements. It is used for our spatial data to track the uncertainty of different areas and neighborhoods.

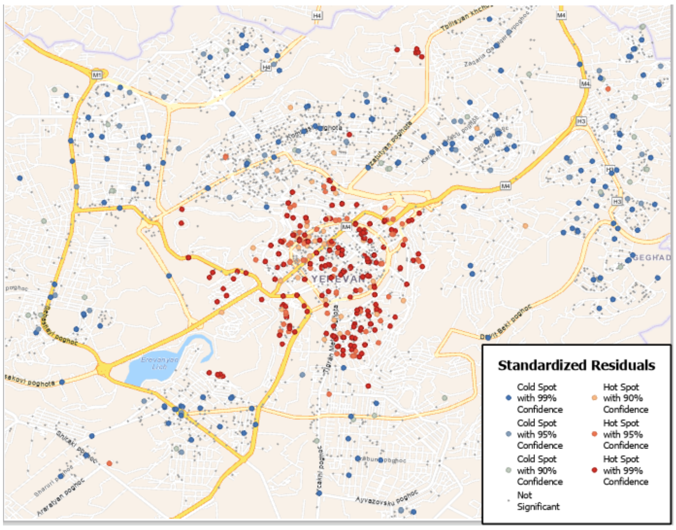
YEREVAN
GEGHAD
Standardized Residuals
Cold Spot with 99% Confidence
Cold Spot with 95% Confidence
Cold Spot with 90% Confidence
Not Significant
Hot Spot with 90% Confidence
Hot Spot with 95% Confidence
Hot Spot with 99% Confidence

Fig. 24. Spatial distribution for the uncertainty of predicted prices for sold houses using the FBCR model

Simply said cold spots refer to areas where the actual price is significantly higher than the predicted ones, and hot spots refer to predicted prices being higher than actual house prices in that area.

And the map shows 3 different levels (99%, 95%, 90%) of certainty in statistical analyses. (i.e., 99% probability suggests the prediction has a high probability of being correct, given the data and model used). As we are analyzing the uncertainty, Fig. 24 shows that uncertainty of sale price predictions tends to be more prone to underestimation for the areas in the center of the city, while neighborhoods outside of the center are more likely to be underestimated. This could be a result of the price range difference in the center of the city compared to regions/neighborhoods outside of it. Additionally, when trying to compare this figure to Fig. 7, we can see that the hotspots for both of these figures align together approximately for the same area, which further suggests to us that although there much more data points for Kentron, the price range for each of the houses in Kentron varies much more, thus making it much harder to have accurate predictions for the data in Kentron, regardless of how many important variables we take into account as independent variables. Although, the FBCR model predicts this quite well on the 10% of the data it takes out for testing (R^2=0.819).

Overall, we reached our goal of training models based on the dataset that we have and created a model that should be applicable to new real estate sale data points in Yerevan. We will use each model to train new data points, those data points being all the houses that were up for sale in Yerevan, Armenia, as of March 13th, 2023. We have obtained the following R^2 for each of our trained models…

- GLR by neighborhood: 0.67
- GWR with Golden Search: 0.75
- FBCR: 0.92

Although we see that the neighborhood GLR model performs worse than GWR, and we have mentioned that both are modified version of how a Linear Regression model would work, but trained on a subset of similarly grouped data, there could be a case that GWR model may have overfit the data based on the different independent variables that were passed so it is important to test the neighborhood GLR model performance as well on the validation data, to see if it would possibly perform better on the new data points.

E. *Comparison of validation data for each model*

A key component of evaluating the performance of any model is to analyze:

1. **Training Data Performance**: When we have a model that performs well on the training data, it can lead us to believe that it has learned the patterns and relationships in the data. However, this can also indicate overfitting, where the model has learned the noise in the training data and does not generalize well to new data. This is why we need the following…
2. **Validation data performance**: This helps to ensure that the model is able to generalize well to unseen data. A model that performs well on the validation data suggests that it has learned the underlying patterns and relationships in the data and can make accurate predictions on new data.

[17] For example, if the model returns $50,000 as the prediction, $40,000 as the lower bound, and $60,000 as the upper bound small changes may affect the model to fluctuate between $40,000 and $60,000 during the prediction.

3. **Metrics**: To quantify the performance of the model, it's important to choose appropriate metrics that capture the goals and requirements of the problem at hand. We would need to analyze these metrics for the training data and the validation data. Some of these metrics include residuals analysis (R^2), and error analysis (MAE, RMSE).

With that being said, now that we have covered all aspects of training data for our models, we have to consider what data we are taking for the purpose of validating how well each of our trained models are performing. We have done the analysis of the training data on real estate houses that have been sold from August 2022 to March 2023 and have chosen to perform the validation for our models on the real estate houses that were listed up for sale as of March 13, 2023. This decision allows us to evaluate the following

- **Representative of the same distribution**: Using unsold houses as validation data means that the validation data is representative of the same distribution as the sold houses. This ensures that the validation data has similar characteristics and features to the data used to train the models.
- **Real-world application**: The ultimate goal of the models is to make predictions on houses that have not yet been sold. Using unsold houses as validation data provides a more realistic assessment of the model's performance in a real-world setting.
- **Potential to identify areas for model improvement**: Evaluating the models on unsold houses may also help to identify areas for model improvement. If the models perform poorly on the unsold houses, it suggests that there may be areas where the models are not capturing the underlying patterns and relationships in the data.

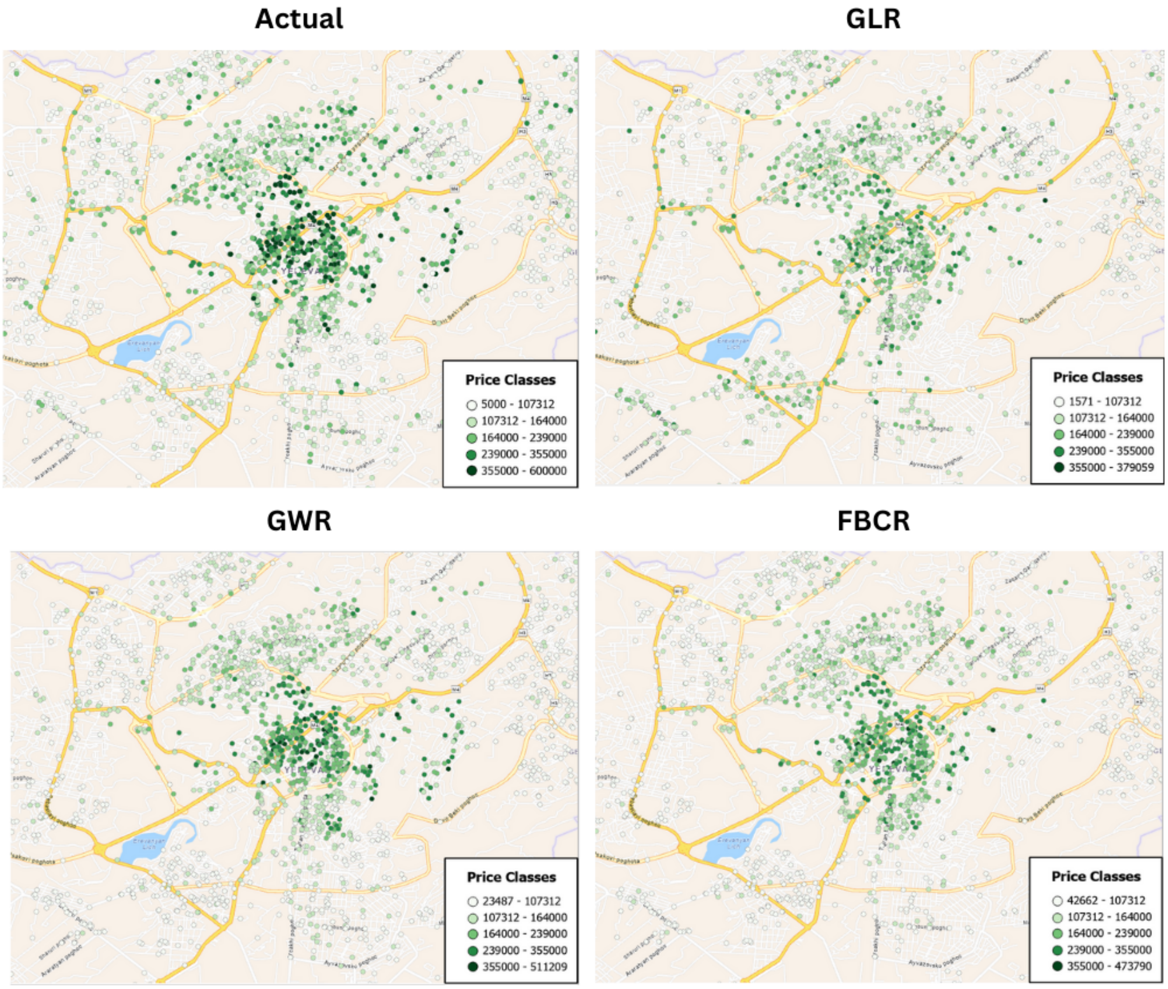

Fig. 25. Class distribution of the predicted prices for each model across the map of Yerevan, compared to the class distribution of the actual price the new homes are being sold for. The price classes for each of the models are adjusted from how they were originally defined by ArcGIS[18] to correspond to the interval for the Actual price. We can see from these graphs that GLR most definitely performed the worst for the new data points that were higher in price. This could be a result of the distribution of the higher prices not being enough to accommodate for shifting the slope of the prediction to start increasing faster. Also, as a result of having houses with significantly low values (starting from 5000) and having values that are distributed towards lower prices of houses Fig 26. the GLR model can become a better predictor of the houses with lower values. From the map, we can see that the FBCR and GWR models performed fairly better, as there is more variety of darker colors throughout the neighborhoods in Yerevan.

[18] Geometric Intervals set by ArcGIS

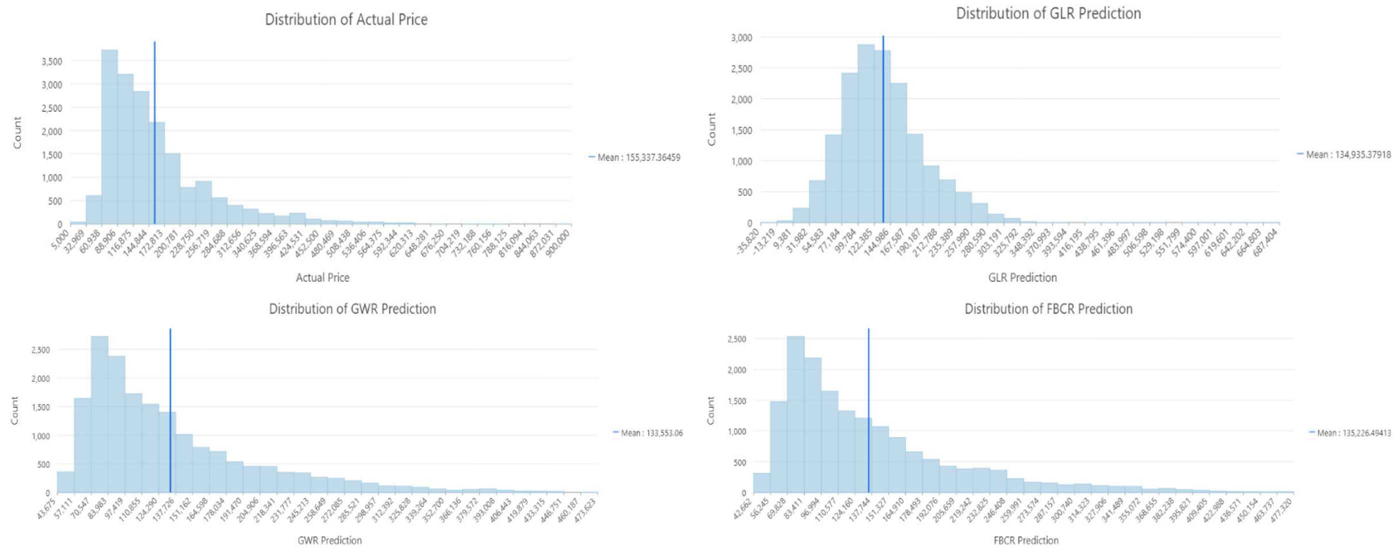

Fig. 26. Distribution of the Predicted prices by each model and the actual price distribution. Although the GLR model is grouped so that the validation is done based on each neighborhood, the distribution of the predicted prices is still somewhat normally distributed. This may suggest that the GLR being trained and tested ultimately gave somewhat significant improvements to our results; however, when compared to the actual price distribution, it’s still pretty far off. Regarding the prediction by the FBCR and GWR model, it comes much closer to our actual distribution prices as both models are commonly used for real estate price predictions because they can capture the spatial heterogeneity of the data.

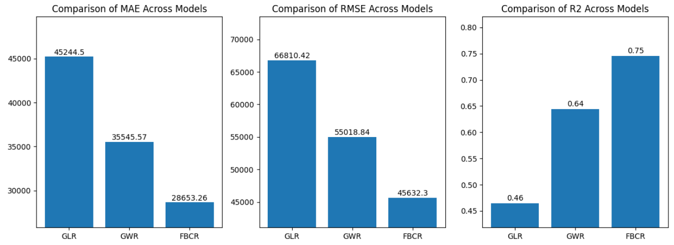

Fig. 27. Validation data performance (MAE, MSE, and R^2) diagnostics for the models GLR, GWR, and FBCR

Firstly, let's understand what information each of the metrics provides [3]:

Note: The letters used in the following two formulas have the same concepts.

- n - the number of observations
- y_i - the actual value of the dependent variable
- $\hat{y}_i$ - the predicted value of the dependent variable
- $\bar{y}$ - the mean of actual values of the dependent variable

Mean Absolute Error(**MAE**) measures the absolute magnitude of the errors. It is calculated by taking the absolute difference between each predicted and actual value.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \qquad (10)$$

Root Mean Squared Error(**RMSE**) is similar to MSE, but as a result of having a squared term in the formula, it makes much more sensitive to larger errors compared to MSE. This means larger errors have a greater effect on the RMSE value than smaller ones.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (11)$$

The coefficient of Determination (R^2) formula can be found in (7)

From the first and second bar charts, we can see the comparison of MAE and RMSE across three different models. As it shows the error rate, the lower the value, the better the prediction of a model. FBCR beats the GLR and GWR models significantly, by approximately 16600 and 6900 points for the MAE score and by 21000 and 9300 points for the RMSE score, respectively, indicating that it has the best predictive power among the three models. And the highest R^2 is again achieved by the FBCR; in other words, it provides the best overall fit in predicting real estate prices.

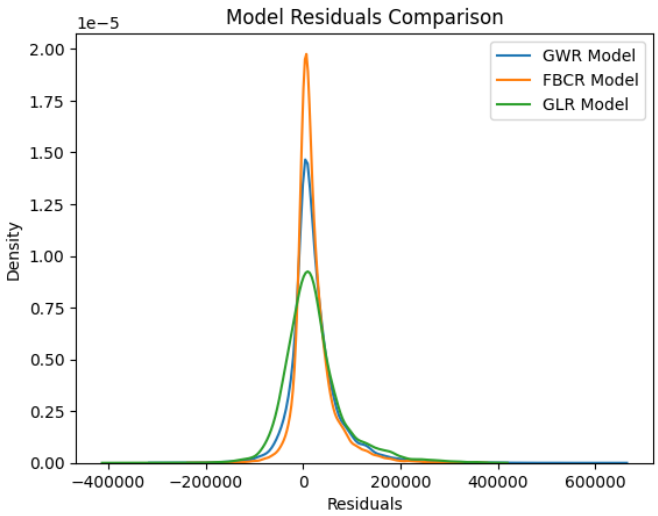

Fig. 28. Validation data residuals from predicted price and the actual price for every model (GWR, FBCR, GLR).

This density plot is an estimate of the distribution of residuals. It is a smoothed version of the histogram and provides insights into the shape of the residual distribution, which is calculated by subtracting the actual price from the model's predicted price. We can see that all of these models have a symmetric distribution on relative to the 0 on the x-axis line. However, this data further back up the fact that the FBCR model works the best for this data, as its higher peak around 0 indicates that the model has fewer prediction errors, as the residuals are mostly concentrated near zero, indicating that the actual values and predicted values are very close to each other.

At this point, we have the reasoning to conclude the following about each model.

- **FBCR**: the FBCR model has the lowest errors for both testing and training, and it is less prone to overfitting the data
- **GWR**: the GWR model performs quite well and does not overfit the data. Also, it further suggests to us as GWR models are designed to account for spatial autocorrelation (Fig 4), real estate prices in Yerevan do not have the tendency of having house sale observations to be more similar to each other than to distant observations.
- **GLR**: the GLR model performs the worst and assumes that the price distributions would be normally distributed, and although we tried to add a spatial measure of training and testing the model by grouping it by the neighborhood in Yerevan, it only slightly improved the results from the baseline approach. This could lead us to claim that including neighborhood as an independent variable only slightly increase performance.

IV. Conclusion & Discussion

In this paper, we used data regarding sold and not sold real estate houses in Yerevan, Armenia, to provide an analysis of how the urban planning developments in Yerevan actually reflect on the real estate market. Additionally, to provide a data-driven solution as to what models are best fit to predict the trajectory of the real estate pricing market and which variables play a key role in becoming a valuable predictor for the pricing of houses. This could eventually become a pipeline that is continually updated to understand how the market would develop and predict the house pricing for any new upcoming project for the city. The following points are the main steps this project looked at:

- Conduct some data processing and cleaning in order to avoid human errors that were scraped from websites and have data spatial correlation range that is close to each other. Meaningful spatial elements to the data itself are also added in order to evaluate whether spatial characteristics of the houses have any effect eventually on our models.
- Evaluating how the urban planning developments for the city have impacted the real estate market. The results from here could be improved and better analyzed when the data has been scraping for a longer time.
- Building and testing models locally from ArcGIS for the data that we have preprocessed. Our initial approach involved us constructing two ways of evaluating the importance of our variables and the performance of our models. The baseline approach is used to establish a benchmark for each model. The idea is to create a simple model that would represent the typical performance of the model without any optimization or fine-tuning. This approach was useful in providing us with a starting point to compare the optimized models. Then eventually, based on the results achieved from the baseline approach, we fine-tuned our models.
- The eventual results of the models should be easily applicable to any new data points that are added to the dataset. This project eventually acts as a way of predicting the real estate prices in Yerevan.

We have trained and evaluated three different models for real estate house price prediction in Yerevan, including Forest-based classification & Regression, Geographically weighted regression, and Generalized linear regression. The FBCR model has the lowest errors for both testing and training, and it is less prone to overfitting the data (Fig 29, 30). The GWR model performs quite well and does not overfit the data. Moreover, it accounts for spatial autocorrelation, suggesting that house sale observations in Yerevan do not have the tendency to be more similar to each other than to distant observations. In contrast, the GLR model performs the worst and assumes that the price distributions would be normally distributed. Although we tried to add spatial measures of training and testing the model by grouping it by the neighborhoods in Yerevan, it only slightly improved the results from the baseline approach. This suggests that if the house belongs to a certain neighborhood, its price will generally be slightly impacted.

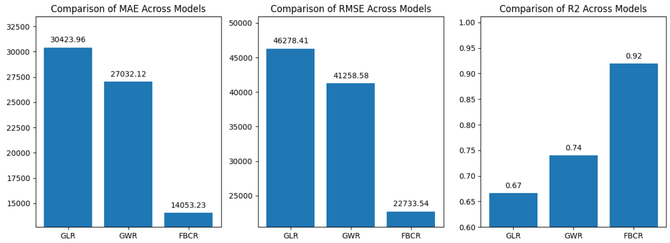

Fig. 29. Trained performance (MAE, MSE, and R^2) diagnostics for the models GLR, GWR, and FBCR

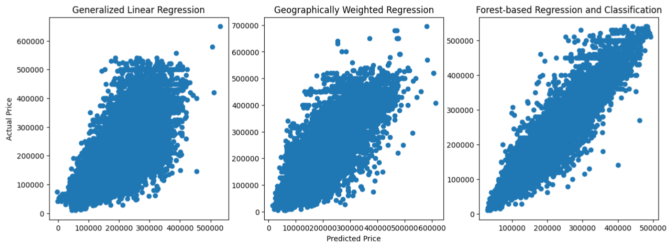

Fig. 30. Scatter plots to visualize the relationship between actual and predicted house prices for the models GLR, GWR, and FBCR

We acknowledge the limitations of our research. One limitation is that the data is scraped from online websites, and a house is considered sold when the listing for it is removed from the website. This approach is quite flawed and may not accurately represent the exact count of sales for the houses. Furthermore, we did not consider other factors that could affect the real estate house prices, such as the age of the house. These limitations could be addressed in future research by using more reliable data sources that would require some form of contact in the Yerevan municipality. Future research could also explore the inclusion of other features that were not considered in this study, such as the condition of the building, the quality of the local schools, and crime rates.

Additionally, the generalizability of our results for the future or even for the past could be questionable as Armenia is a geopolitically unstable country. The geopolitical state of a country can greatly affect its economy, which subsequently includes the real estate market. However, future research for this project can explore the impact of geopolitics on the real estate market, but this would require gathering past data and collecting data that would span a much larger timeframe.

In summary, based on our evaluation, we recommend using the FBCR and GWR models for real estate house price prediction in Yerevan. These models perform well and have unique strengths that could be useful for different purposes. However, any future readers of this project should be aware of the limitations of the data and the constraints of the time period this was evaluated.

Acknowledgments

We would like to thank Sergey and Karen from GeoVibe for their continued guidance and support while using ArcGIS.

References

[1] Winky K.O. Ho, Bo-Sin Tang & Siu Wai Wong (2021) Predicting property prices with machine learning algorithms, Journal of Property Research, 38:1, 48-70, DOI: 10.1080/09599916.2020.1832558

[2] Gale, H., Roy, S.S. Optimization of United States Residential Real Estate Investment through Geospatial Analysis and Market Timing. Appl. Spatial Analysis 16, 315–328 (2023). https://doi.org/10.1007/s12061-022-09475-x

[3] Al-Hamadin, Rashed & al-sit, walid. (2020). Real Estate Market Data Analysis and Prediction Based on Minor Advertisements Data and Locations' Geo-codes. International Journal of Advanced Trends in Computer Science and Engineering. 9. 4077 – 4089. https://doi.org/10.30534/ijatcse/2020/235932020

[4] Quang Truong, Minh Nguyen, Hy Dang, Bo Mei, Housing Price Prediction via Improved Machine Learning Techniques, Procedia Computer Science, Volume 174, (2020), Pages 433-442, ISSN 1877-0509. https://doi.org/10.1016/j.procs.2020.06.111

[5] Ali Soltani, Mohammad Heydari, Fatemeh Aghaei, Christopher James Pettit. Housing price prediction incorporating spatio-temporal dependency into machine learning algorithms, Cities, Volume 131, 2022. https://doi.org/10.1016/j.cities.2022.103941

[6] Real Estate Market in Armenia. (2019b). amcham.am. https://amcham.am/wp-content/uploads/2019/07/GT-_-real-estate-market-analysis.pdf

[7] Current Trends in Armenia's Real Estate Market. (2021, March 17). evnreport.com. https://evnreport.com/economy/current-trends-in-armenia-s-real-estate-market/

Appendix 1: Results for Trained Models

Model	R-Squared	RMSE	MAE
GLR	0.58	52320	34715
GWR	0.74	41258	27032
FBCR	0.92	22733	14053

Appendix 2: Variable Importance for Reduced FBCR Model

Variable	%
square_meters	29
district	23
height	17
walking_distance_to_metro(m)	12
building_floors	11
Yerevan_POI	8

Appendix 3: Reduced FBCR Model Diagnostics

Number of Trees	1000
Leaf Size	5
Tree Depth Range	29-42
Mean Tree Depth	34
% of Training Available per Tree	100
# of Randomly Sampled Variables	2
% of Training Data Excluded for Validation	10

Appendix 4: GLR Neighborhood Improvement Over Baseline

Neighborhood	Improvement over baseline (%)	Mean residual (baseline)	Mean residual (neighborhood)
11th_Quarter	92,08902	38570,33	10902,7
15th_Quartier	64,00756	38570,33	21068,54
1st_Block	67,92505	38570,33	20840,01
1st_Quarter	71,93066	38570,33	20122,36
2nd_Block	65,51693	38570,33	23645,59
2nd_Quartier	86,22116	38570,33	15848,8
3rd_Block	64,05524	38570,33	23251,83
4th_Block	52,8121	38570,33	26576,03
4th_Quartier	76,59514	38570,33	19964,69
5th_Quartier	60,10945	38570,33	25906,58
7th_Quartier	77,62196	38570,33	18997,57
8th_Quarter	44,13512	38570,33	25947,04
A1	77,33754	38570,33	17446,58
A2	10,43955	38570,33	26918
A3	91,83717	38570,33	11993,63
Ajapniak	69,49882	38570,33	19868,06
Anastasavan	63,16979	38570,33	24895,83
Antarayin	-151,609	38570,33	65813,69
Arabkir	43,29172	38570,33	29989,09
Araratian	76,18703	38570,33	17685,16
Avan	67,46129	38570,33	20257,74
Avan_Arinj	71,63191	38570,33	18841,08
Aygedzor	29,96829	38570,33	36273,36
Aygestan	30,04744	38570,33	33212,93
B1	87,92049	38570,33	13346,22
B2	70,59616	38570,33	15721,46
B3	27,90365	38570,33	31049,85
Bryusov	67,02019	38570,33	22177,36
Charents	74,6321	38570,33	17856,5

Davtashen_Block	64,1786	38570,33	23701,21
Duryan	68,74745	38570,33	22740,87
Erebuni	74,60352	38570,33	17825,79
Haghtanak	-86,1709	38570,33	48013,73
Isahakyan	-2,67116	38570,33	44388,56
Jrvezh	41,99088	38570,33	27655,42
Kanaker	72,55857	38570,33	18205,52
Kanaker_Zeytun	61,19213	38570,33	22427,55
Kentron	-35,5707	38570,33	45023,7
Kharberd	83,6634	38570,33	16435,44
Koghb	-3,39836	38570,33	32061,01
Kond	16,11766	38570,33	33991,39
Kuchak	74,46303	38570,33	17674,16
Lukashin	79,33145	38570,33	16920,41
Malatia_Sebastia	29,78976	38570,33	29997,97
Mayak	53,57939	38570,33	23597,77
Narekatsi	71,47037	38570,33	19488,6
Nazarbekian	66,98091	38570,33	20213,08
Nerkin_Shengavit	75,21718	38570,33	18150,58
Nor_Arabkir	51,3856	38570,33	27718,9
Nor_Aresh	70,90094	38570,33	20718,91
Nor_Butania	70,76962	38570,33	20986,33
Nor_Kilikia	-166,836	38570,33	60861,61
Nor_Malatia	70,23709	38570,33	22373,36
Nor_Nork	76,02717	38570,33	17670,39
Nor_Nork_1st_Microdistrict	73,51499	38570,33	17211,06
Nor_Nork_2nd_Microdistrict	79,79412	38570,33	16807,92
Nor_Nork_3rd_Microdistrict	81,32644	38570,33	15892,45
Nor_Nork_4th_Microdistrict	75,51518	38570,33	18283,56
Nor_Nork_5th_Microdistrict	75,87511	38570,33	17763,47
Nor_Nork_6th_Microdistrict	83,98224	38570,33	15289,99
Nor_Nork_7th_Microdistrict	75,37909	38570,33	17163,84
Nor_Nork_8th_Microdistrict	87,48067	38570,33	12958,42
Nor_Nork_9th_Microdistrict	49,36268	38570,33	25415,02
Nor_Sebastia	72,4969	38570,33	19056,71
Nor_Zeytun	59,35677	38570,33	24313,81
Norashen	64,94245	38570,33	22914,47
Nork	28,65779	38570,33	33018,36
Nork_Marash	-5,51233	38570,33	39210,41
Nubarashen	81,12124	38570,33	19014,05
Old_Yerevan	-29,9464	38570,33	46485,78
Sari_Tagh	90,01355	38570,33	12921,01
Sayat_Nova	47,04134	38570,33	22400,96
Shahumyan	63,12139	38570,33	23031,65
Shengavit	63,30689	38570,33	22223,26
Tumanyan	79,14691	38570,33	17891,15
Vardashen	65,40448	38570,33	22299,55
Varuzhan	71,24246	38570,33	16744,25
Verin_Charbakh	68,8852	38570,33	16915,21
Verin_Shengavit	69,59201	38570,33	20642,96
Zoravar_Andranik	73,2	38570,33	17654,22