Pricing Factors of Student Housing around UC Santa Barbara: A Spatial and Non-Spatial Machine Learning Analysis

Jinze Li, Yixuan Wang, Weitian Xie[1]

[1]Department of Statistics and Applied Probability, UC Santa Barbara

May 2021

Abstract

Place holder

Contents

Student Housing, Rent Prices, Geographically Weighted Regression, Elastic Net, Random forest, Neural Network

1 Introduction

COVID-19 has exposed many issues in the economy. Something that is deeply connected to students and University-based researchers – like us – is student housing issues. The student housing calamity unfolding at UCSB in Fall 2021 epitomized the inability of the student housing market near college towns to adapt to market stress. UCSB even had to subsidize student housing to relieve the shortages by contracting with local hotels. However, UCSB's student housing issues have plagued the community for years, which the 2020 Grad Students Strike best exemplifies (Brian Osgood). In fact, due to the inelasticity of supply due to lengthy planning and construction time, high cost of entry due to California legislations such as proposition 13, and lack of substitute housing due to the Santa Barbara South Coast region's relative geographical isolation, the UCSB student housing rental market is more valuable to short term – such as a pandemic – and medium-term – such as increase enrollment – market stress.

Therefore, identifying students' housing preferences is crucial for UCSB's future expansion and sustained growth as a prestigious academic institution. The demand for such knowledge is made even more urgent with UCSB's controversial windowless "Mega-Dorm" proposal that sparked national debate (Maria Cramer; Dennis Mcfadden).

Therefore, the purpose of this study is to identify influential factors on students' housing rental prices to deduce students' rental preferences using online proprietary and open-source databases and a purpose-built campus-wide online survey in the case study submarket of UCSB student residential rentals. We will compare multiple spatial and non-spatial machine learning methods using the different data courses. In the end, we hope to turn our research into real-world software that would help students find their price range and ideal property based on their preferences.

2 Literature Review

Our research approach is primarily inspired by and developed upon the methods in "Influence Factors and Regression Model of Urban Housing Prices based on Internet Open Access Data" published in 2018 with the leading scholar Hao Wu[1]. Investigating apartment sales price and consumer house-buyer preference, their research employed hedonic linear regression, geographically weighted regression model (GWR), and Artificial Neural Network which are all commonly used to investigate housing prices and their influencing factors.

The hedonic linear regression uses “the regression analysis to build a regression model of housing prices and various influence factors, and then discusses the effect of each factor on a house price” [1]. However, the hedonic linear regression overlooks the auto correlation among influencing factors, and this problem can be solved by including in the GWR model which “parameters and their weights can be adjusted locally in accordance with their spatial locations” [1]. Plainly, GWR tries to fit different (linear) regression functions at different geographic locations based on the dependent and feature variables near that location.

Wu’s paper also introduced what kinds of data are relevant to such study and what kind of source are used to collect them. The types of variables they used are: housing price-related data, point of interest data, location-based service data, and urban planning and internet map data. Most of their data are sourced from publicly available information from internet and telecommunication companies – such as property listing website, dianping.com website, online maps, and mobile phone usage data – which inspired us to search for available data from their American counterpart. Together, Wu used 10 explanatory variables. Their best linear model – GWR – has $R^2 = 0.359$ while their best non-linear model – ANN combined with GWR– has $R^2 = 0.725$.

Some improvements can be made to improve the prediction performance and depth of analysis. Namely, the relatively small number of explanatory variables left many variations in the response unaccounted for and may be plagued by confounding issues; moreover, the specific explanatory variables they chose or defined may seem somewhat arbitrary and lack rigor. To address this, we perform non-linear transformation on some variables, define similar variables in more nuanced and rigorous manner such distance as measured in travel time rather than euclidean distance, expand the total number of variables to 34, and use dimension reduction and variable selection techniques to identify the relevant variables and control model variance. In the end, our model produced better goodness of fit in terms of both R^2 and $AICc$ and allows us to analyze more factors contributing to housing prices. Many additional variables we used are also relevant to the unique geography and economy of our area of study.

3 Data Collection and Processing

External data are appended in R or extracted in ArcGIS. All spatial coordinates are projected from Longitudes and Latitudes to Northing and Eastings to enable accurate calculation of Euclidean distance and direction.

3.1 Study Area

The study area is Santa Barbara South Coast Region, where University of California, Santa Barbara (UCSB) is located, with an approximation of 448,000 people live and an area of 9,810 square kilometer. [2] In the studied area, Isla Vista is the small beach town near UCSB, where most of the houses are rented to UCSB students and faculties. Goleta provides daily necessities for UCSB students, through its access to marketplace, hospitals, restaurants and gyms. Other than Isla Vista and Goleta, Santa Barbara Downtown offers more spacious houses for UCSB students with easy access to makers, hospitals, restaurants and gyms.

3.2 Online Internal Housing Characteristic Data

Reassemble the data by bedroom, only keep data with per SF data, How we decide which variables to use: some observations may not have certain data such as year built; preliminary model suggest year built is useless so we removed it.

Dimension Reduction on Internal Amenities

Data are collected from the Costar group, one of the largest commercial real estate data providers. It owns many residential listing sites, including Apartments.com, Apartamentos.com, ApartmentFinder.com, ApartmentHomeLiving.com. The data contains 250 properties that most of the aforementioned information for multifamily properties that are around twenty minutes' driving distance from campus – we deem this area, which stretches from west of Goleta to Summerland, to be most relevant in our study. For data processing, we first added the travel time and distance via multiple transit methods to the UCSB campus for each property. This is achieved in R through a function we defined that would append travel time and distance via walking, biking, public transit, and driving using Google Cloud's Distance Matrix API.

Then, we mutated the data frame to contain different boolean variables of different types of amenities, such as gyms, pools, air conditioning, indoor-gathering space, outdoor gathering space, electric car charging station, safety-related features, and on-site serves. The reason for this step is that these features are originally combined into one column and some properties may state similar amenities differently (such as having a Lounge or a Clubhouse).

3.3 Point of Interest Data

Point-of-interest(POI) data are locations near each property that may influence the tenant's preference, which includes restaurants, shops, gyms, and parks in our study. The four types of POI are chosen because of their close connection to student's daily life or are found to be important in previous research[3, 4]. We used Google Cloud's Places API to access the rich POI data from Google Maps – or, as Anish Nahar of Harvard Business School calls it, "the most expansive data machine"[5]. For each category, a few relevant "place types", defined here, are requested from said API [6]. Note that each of the four POI categories encompasses a broader set of respective locations subscribed to a more general and inclusive definition. For example, the restaurant includes cafes and bars; stores include plazas, malls, and supermarkets; parks include tourist attractions, museums, campsites, and beaches.

However, since Google tries to prevent mass data extraction, we had to employ some web scraping tricks such as adding random time breaks in our code and performing extraction using multiple center locations. After removing duplicates, we are able to mass 477 restaurants, 295 stores, 51 gyms, and 120 parks, all with geo-coordinates and rating numbers attached.

These data points are fed into a weighted kernel density function – a kernel that measures both geographical proximity and density of POI – to be extracted as explanatory variables. See 4.1 for details on kernel density estimation in ArcMap. Traditional wisdom tells us that the closer we are to these businesses, the more convenient the location is, so the kernel density of POIs should reflect the preference for convenience with respect to the particular types of business or locations. Moreover, this also correlates with the proximity to business centers without artificially defining where they are.

Since kernel density estimates require the researcher to input the "search radius" parameter – which governs the rate of decrease of the kernel with respect to Euclidean distance from the residence to the POI – we selected three interpretable radii of 0.5 miles, 1 mile, and 3 miles to feed into the variable selection scheme. One can interpret them as POI within walking distance, biking and transit distance, or driving distance respectively.

The weights used in weighted kernel density are the total number of ratings on Google Maps, which should correlate with the frequency of visits and influence of a particular business. This agrees with empirical results as large retailers – such as Paseo Nuevo Mall, Costco Wholesale, Best Buy, and La Cumbre Plaza – and local staple restaurants – such as In-N-Out Burgers, Boathouse, Brophy Bros, Free Bird – received the most ratings. However, the issue with this approach is that we cannot account for the popularity particularly in the student population, but we still believe it's a decent estimation. Note, for parks and restaurants, we applied Box-Cox transformation on the total number of ratings to address extremely large outliers; this normalizing transformation is also motivated by the fact that for these two types of POI, the availability of more choices should be more preferable than just having one highly popular location.

In short, compared with finding the Euclidean distance to POI, weighted kernel density estimation encodes proximity, density, and quality while allowing for flexible interpretations.

3.4 Transit Accessibility

Transit Accessibility data includes transit and driving time to UCSB campus, bus stops, bike paths, road density, and distance to highway.

Commute time is imperative to housing choice[7]. To investigate student's preference for commute time, we employ Google Cloud's Distance Matrix API through the 'gmapsdistance' R package to estimate the public transit time and driving time to UCSB library from each housing location on a typical morning (Citation, data source). A function is coded in R to automate this process given any data frames.

This method takes into account the paths, average speed of each link, traffic congestion, transit stop location, transit schedules, and other factors which offers a more nuanced estimation than Euclidean distance. Note, for residents extremely close to campus, transit time will be equivalent to walking time. The accuracy compared to other platforms and numerical implementation is beyond the scope of this paper, but the prediction is realistic from empirical experience. Reference Bahman Lahoorpoor and David M. Levinson's paper for more details[8].

The only form of public transit in the study area is buses managed by Santa Barbara Metropolitan Transit District (MTD)(Citation). The General Transit Feed Specification files (GTFS), a standard data specification widely used by public transit providers and mapping services, are obtained from the Santa Barbara MTD website[9]. The data are manually processed in R to obtain the total number of times bus stops at each station during a typical week. Similar to POI data, the weekly stoping count undergoes a similar Box-Cox transformation to become weights in kernel density estimation, which reflects the density of bus stops in the research area. Two estimations with different search radii – 0.5 miles and 1 mile – are generated; this is motivated by the Department of Transportation's research stating that most people are willing to walk 0.25-0.5 miles to a transit stop [10]. Additionally, the Euclidian Distance between residence and nearest bus station and its inverse is also calculated to be plugged into variable selection.

Bike path shapefile data are provided by the Santa Barbara Bike Collection's ArcGIS server.[11] An (unweighted) kernel density with a search radius of 0.5 miles is generated to capture the density of official bike paths; additionally, the distance to the nearest bike path and its inverse is also calculated.

Road and highway shapefile data are provided by "hkeswanihs" on ArcGIS.com, which he extracted from U.S. Census Bureau's Master Address File and Topologically Integrated Geographic Encoding and Referencing[12]. Similarly, an (unweighted) kernel density with a search radius of 0.5 miles is generated to capture the density of roads while the distance to highways and its inverse is computed.

3.5 Additional Geographical Data

Additional data we incorporate into our model includes population density, coastline, flood risk, and noise level.

Facebook's Population High Resolution Population Density Map is one of the highest quality publicly available population data with a resolution of 30 meters! Moreover the data is stratified by demographics. In short, the data is generated using computer vision machine learning on satellite images, building type identification, and census data; a detailed breakdown of its methodology is in the paper from Tobias G. Tiecke et al[13].

In particular, we used the data of the entire population as well as youth aged 15 to 24[14]. The latter correlates well with college students, as the highest density in this age group are in the college town Isla Vista, dorms, and in Santa Barbara City College. Since the resolution is too fine, Kriging interpolation is performed in ArcMap.

Coast line data are credited to United States Geological Survey's "The National Assessment of Shoreline Change"[15]. The distance of each residence to the nearest coast line along with its inverse are determined. Noise and flood risk are produced by Santa Barbara County government[16]. Since, the shapefile indicates the noise level if it is over 60dB, so we assume the rest of the area has a baseline noise level of 50dB. Note, the noise data suggest that the noise courses are airplane departure routes, train tracks, highways, and major roads. Therefore, the data seems to only capture transportation noise. The flood shapefile is a boolean type indicating where flood risk is prevalent.

One type of data lacking in this research is crime rate. Although we see the importance, we are not able to trackdown such public data. However, since the south coast region is generally safe, it becomes less important in our case.

3.6 Survey Data

Survey Design Data Cleaning Geo Coding Time Discounting

3.7 Variable Definition

Variable definition of Costar data is shown in Table 1.

4 Methodology

In this study, two non-spatial linear hedonic models –elastic net (EN) and partial least squares (PLS) regression – are fitted. The variable selection or dimension reduction results from elastic net and PLS are used as a starting point to construct a reduced model to be used in spatial machine learning model – geographically weighted regression(GWR). Variable selection is necessary for GWR due to the inherent higher model variance and numerical issues with sparse data. Taken together, the three highly interoperable linear models are used for analysis and interpretation to identify the factors influencing student housing prices and infer UCSB student's housing preference.

Lastly, two nonlinear or non-parametric models are used to validate our data construction process and above linear models. Random forest (RF) and deep neural network (DNN) for their high prediction performance[17]. In particular, RF can help to identify if nonlinear transformation is needed on explanatory variables and DNN provides an insight into the information carried in the explanatory variables we constructed.

This section will provide a summary of the models or methods used in this study.

Number	Variable Name	Variable Type	Description
1	Effective_Rent_SF	Numeric	Rent price per square feet
2	Style	Categorical	Style of the house
3	Number_Of_Units	Numeric	Number of units in the house
4	Transit_Time	Numeric	Transit time to UCSB
5	Drive_Time	Numeric	Drive time to UCSB
6	Bed_Count	Numeric	Number of beds in the house
7	Size	Numeric	Size of the house
8	Amenities_AC	Boolean	Air conditioning
9	Amenities_Safety	Boolean	Safety measures (eg. gated or doorman)
10	Amenities_Pool	Boolean	Pool
11	Amenities_Entertainment	Boolean	Entertainment equipment
12	Amenities_Gym	Boolean	Gym
13	Amenities_EV	Boolean	Electric vehicles
14	Amenities_Service	Boolean	Front desk service
15	Distance_Beach	Numeric	Distance to beach
16	Bike_800	Numeric	Distance to bike path (Radius = 800 in ArcGIS)
17	Bus_800	Numeric	Distance to bus station (Radius = 800 in ArcGIS)
18	Bus_1600	Numeric	Distance to bus station (Radius = 1600 in ArcGIS)
19	Store_800	Numeric	Distance to grocery store (Radius = 800 in ArcGIS)
20	Store_1600	Numeric	Distance to grocery store (Radius = 1600 in ArcGIS)
21	Store_4800	Numeric	Distance to grocery store (Radius = 4800 in ArcGIS)
22	Rest_800	Numeric	Distance to restaurant (Radius = 800 in ArcGIS)
23	Rest_1600	Numeric	Distance to restaurant (Radius = 1600 in ArcGIS)
24	Rest_4800	Numeric	Distance to restaurant (Radius = 4800 in ArcGIS)
25	Gym_800	Numeric	Distance to gym (Radius = 800 in ArcGIS)
26	Gym_1600	Numeric	Distance to gym (Radius = 1600 in ArcGIS)
27	Gym_4800	Numeric	Distance to gym (Radius = 4800 in ArcGIS)

Table 1: Variable Definition for Costar Data Analysis.

4.1 Kernel Density Estimation

4.2 Principal Component Analysis

4.3 Partial Least Square

The PLS regression model is chosen for variable selection as it deals with collinearity problem which is common in multivariate regression. Multiple literature highlighted the importance of variable selection in the existence of multi-dimensional variable space. (Mehmood et.al, 2012; Peres and Fogliatto, 2018) Variable selection serve the purpose of improve estimation performance and predictability of the model. To that end, Partial Least Square Regression (PLS) has proven to be a commonly used method in multivariate data analysis. (Mehmood et.al, 2012) PLS is known for its ability to deal with large number of explanatory variables, highly likely due to collinearity among them. (Afanador et.al, 2014) From its algorithm, PLS is able to ignore the effect of the variable space that are spanned by irrelevant, noisy variables.

PLS is argued to obtain three advantages in multivariate modeling towards response variables. (Willaby et.al, 2015) The first advantage is claimed to be the use of ordinary least squares manages to minimize the unexplained variance for all explanatory variables in a multivariable regression model. Secondly, PLS algorithm ensures the ease of modeling formative constructs, as the explanatory variables are antecedent which offers straightforward interpretation from items to construct. (Willaby et.al, 2015) The thirst advantage is the fact that PLS algorithm does not require large sample size which only limited by the single largest regression equation.

The PLS has now developed into three main categories: filter-, wrapper-, and embedded methods. In this paper, we mainly use PLS regression which we assume a linear relationship between a set of independent variables $X_{(n,p)}$ and a single response variable $y_{(n,1)}$ through the equation $y = \alpha + X\beta + \epsilon$. [18] In the equation, unknown regression parameters are defined to be α and β, with error term ϵ. [19]From our Costar data, $n = 281$ and $p = 26$. Also we assume that $A(A \leq p)$ is the number of relevant components under analysis, then follows the algorithm that repeats for $a = 1, 2, ..., A$:

1. Calculate the loading weights through $w_a = X'_{a-1} y_{a-1}$, where $X_0 = X - 1\bar{x}'$ and $y_0 = y - 1\bar{y}$. In this equation, the weights are the directions in the space with the span of X_{a-1} of maximum covariance with y_{a-1}. This step extracts the weights of X_{a-1} from $X_{(n,p)}$, the entire set of independent variables from our data where $n = 281$ and $p = 26$.

2. Calculate the score vector t_a through $t_a = X_{a-1} w_a$.

3. Calculate the loading weights of X, p_a, through the regression of X_{a-1} on the score vectorz $p_a = X'_{a-1} \frac{t_a}{t'_a t_a}$. Similarly with the loading weights of y $q_a = y'_{a-1} \frac{t_a}{t'_a t_a}$.

4. Subtract t_a

$$\begin{matrix} X_a = X_{a-1} - t_a p_a \\ y_a = y_{a-1} - t_a q_a \end{matrix}.$$

From the algorithm, the loadings of X is matrix $P = [p_1, p_2, ..., p_A]$, the loadings of Y is matrix $Q = [q_1, q_2, ..., q_A]$. The regression estimators are $\hat{\beta} = W\left(P'W\right)^{-1}$ and $\hat{\alpha} = \bar{y} - \bar{x}\hat{\beta}$, where the matrices W,P,T are defined by $W = [w_1, w_2, ..., w_A]$,$T = [t_1, t_2, ..., t_A]$.

4.4 Regularization Regression

xx

4.5 Geographic Weighted Regression

Geographically weighted regression (GWR) is a spatial analysis technique. Briefly, it fits separate proximity weighted OLS at every location in the dataset to address, which incorporates the dependent and explanatory variables of locations falling within the bandwidth of each target location(Citation, How GWR works). Formally it is defined as[20]:

$$\hat{y}_j = \beta_{j,0} + \beta_{j,1}x_1 + \beta_{j,2}x_2 + ... + \beta_{j,p}x_p + \epsilon_j \tag{1}$$

With the coefficient at each location index j:

$$\vec{\beta_j} = (X^T W_j X)^{-1} X^T W_j \vec{y} \tag{2}$$

In this study, we use the Gaussian weighting kernel[21] subject to some bandwidth. Thus, diagonal weight matrix's i's diagonal entry is given by:

$$W_j(i) \propto \begin{cases} e^{-r_{j,i}^2} & if\ r_{j,i} \leq bandwidth \\ 0 & otherwise \end{cases} \tag{3}$$

where $r_{j,i}$ is the geographical distance between j to i. Notice, this is near identical to OLS except for the weight matrix $W(i)$ which is specific to each location j such that observations near j have a greater influence on the model coefficients.

Therefore, compared to non-spatial hedonic models, GWR makes the modeling more sophisticated by allowing the relationships between the independent and dependent variables to vary by locality[22]. It also accounts for spatial lag of variables – that is geographically nearer observations are more correlated or "related" than further observations[23]. This autocorrelation arose from Tobler's first law of geography, which states that "everything is related to everything else, but near things are more related than distant things"(Citation, https://doi.org/10.1093/acrefore/9780190264079.013.325)/. Sometimes, this autocorrelation can be explained by explanatory variables not included in the dataset(Citation, https://www.publichealth.columbia.edu/research/population-health-methods/geographically-weighted-regression). Explicitly, spatial lag is manifested in autocorrelation of regression coefficients with respect to geographical proximity.

In this study, we implement GWR in ArcMap to utilize the vast array of shapefile and geographical visualization toolkits. Because GWR fits different regressions at different locations, it leads to greater model flexibility and thus high model variance that is sensitive to multicollinearity – especially when observations are sparse. To address this, robust variable selection that takes multicollinearity into account are used and suitable neighborhood bandwidth are identified using the Akaike information criterion (AIC). Note, AIC are chosen over cross validation due to the already limited sample size. The above measure turns out to be necessary for GWR's to be fitted numerically as well.

The advantage of GWR is in its ability to conduct more sophisticated inference and interpretation rather than prediction performance(Citation, https://link.springer.com/chapter/10.1007/978-3-642-02664-5$_5$; $https : //link.springer.com/chapter/10.1007/978-3-642-03647-7_22).ThereforeGWRisprimarilyuse$

4.6 Random Forest

Random Forest is a traditional and common machine learning algorithm which combines the output of multiple decision trees to reach a single result. It us flexible and easy to use, which have fueled its adoption, and it can solve both classification and regression problems.These questions make up the decision nodes in the tree, acting as a means to split the data. Each question helps an individual to arrive at a final decision, which would be denoted by the leaf node. Observations that fit the criteria will follow the "Yes" branch and those that don't will follow the alternate path. Decision trees seek to find the best split to subset the data, and they are typically trained through the Classification and Regression Tree (CART) algorithm. Metrics, such as Gini impurity, information gain, or mean square error (MSE), can be used to evaluate the quality of the split. While decision trees are common supervised learning algorithms, they can be prone to problems, such as bias and over fitting. However, when multiple decision trees form an ensemble in the random forest algorithm, they predict more accurate results, particularly when the individual trees are uncorrelated with each other.

Discuss how it contribute to identify variable that does not have a linear relationship.

4.7 Neural Network

4.8 Additional Methods Considered

Knock-off filter, Boosting, Step-wise,

5 Results

5.1 Descriptive Statistics and Unsupervised Machine Learning

5.1.1 Descriptive Statistics for CoStar Data Analysis

Figure 1: Histogram of Rent Price Around UCSB Per Square Feet

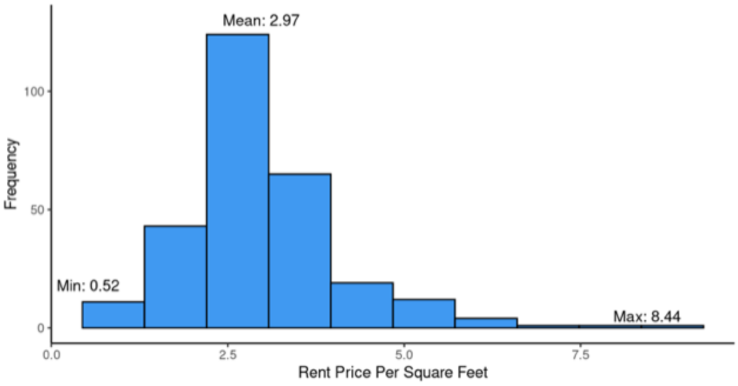


Figure 1 shows the general distribution of Rent Price Around UCSB per square feet, according to the survey result. From the histogram, the mean of the rent is 2.97 per square feet, with minimum $0.52 and maximum $8.44 per square feet. From the figure, the majority of the houses of respondents are around $2.5.

In terms of housing conditions for the survey respondents, 204 out of 277 houses are of low-rise building; 67 of 277 are houses with gardens and 6 are mid-rise buildings, as shown in figure 2. From figure 4, 253 of total respondents have air conditions in their houses, with 28 do not. In addition, 253 answered to have houses with gates or doormen, as in figure 5.

In terms of housing conditions for the survey respondents, 204 out of 277 houses are of low-rise building; 67 of 277 are houses with gardens and 6 are mid-rise buildings, as shown in figure 2. From

Figure 2: Proportion of Different Types of Houses

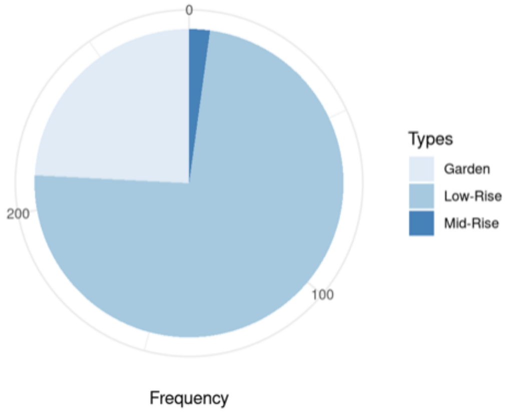


Figure 3: a (left) - Proportion of Air Conditioner; b (right) - Proportion of Gated Houses

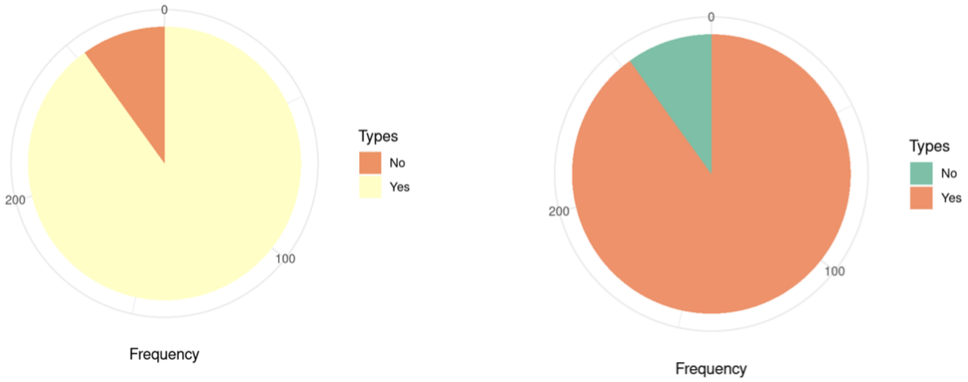


figure 4, 253 of total respondents have air conditions in their houses, with 28 do not. In addition, 253 answered to have houses with gates or doormen, as in figure 5.

The scatter plot, figure 3, does not show an evident relationship between distance to beach and rent price near UCSB, opposed to economic hypothesis. Referenced from Kennedy's article, it is a reasonable hypothesis that location is important in explaining a rent price; specifically in this case, we assume distance to beach mainly determines a housing location which explains a rent price. From the figure, most of the survey data points at the place with close distance to beach (0.00 - 0.01)and low rent price. Moreover, with 209 of rent price are in the same interval (2 - 4), they are also evenly distributed in different intervals, (0.01-0.02, 0.02-0.03, 0.03-0.04) for distance from houses to beach.

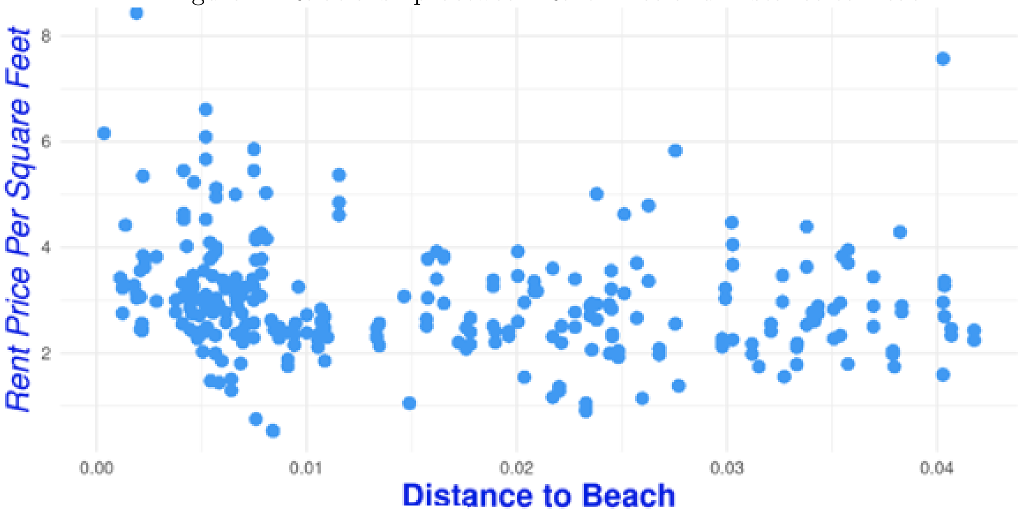
5.1.2 Descriptive Statistics for Survey Data Analysis

5.1.3 Clustering and Principal Component Analysis

5.2 Co-Linearity

Co-Linearity issues underlines many of the decisions in our research.

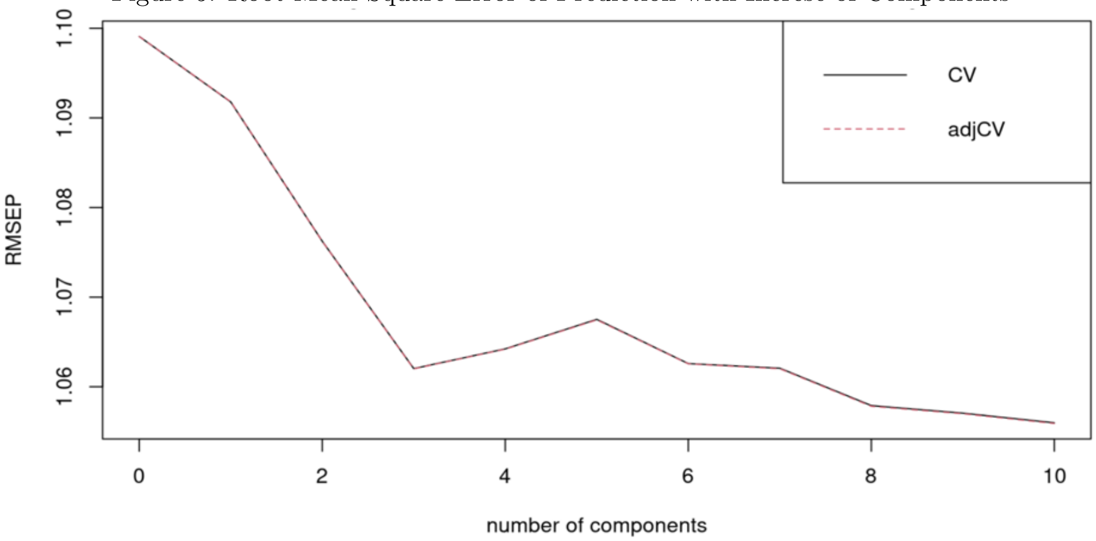
Figure 4: Relationship between Rent Price and Distance to Beach



Variable Name	Minimum	Median	Mean	Maximum
Transit Time	439	2865	2563	5199
Drive Time	356	711	692	1031
Distance to Beach	0.0004	0.0110	0.0162	0.0418
Distance to Bike Path	0.00014	0.00115	0.00150	0.00824

Table 2: Minimum, Median, Mean and Maximum values for variables Transit Time, Drive Time, Distance to Beach and Distance to Bike Path.

Figure 5: Root Mean Square Error of Prediction with Increse of Components



5.3 Model Results

5.3.1 Variable Selection - Lasso

5.3.2 Variable Selection - Partial Least Square

Table 3 shows the proportion of Explained Variance with increase of components. The increase in the number of components raises the ability of PLS model of explaining the variable of the y variable. With 1 component, the PLS model explains 83.12% of the variance of the y variable; with 2 components, the proportion increases to 97.91% of the variance. This in turn reflects the usefulness of applying PLS in variable selection process.

Components	Variance Explained
1	83.122
2	97.913
3	98.821
4	99.484
5	99.71
6	99.82
7	99.94
8	99.97
9	99.99
10	100

Table 3: Proportion of Explained Variance with increase of components.

As shown in Figure 6, the RMSEP (Root Mean Square Error of Prediction) drastically decrease with the increase in number of components being calculated by the algorithm. This is indicted by two cross-validation (CV) estimate - CV, as the ordinary CV estimate, and adjCV, as the bias-corrected CV estimate. The LOO CV do not have much difference in this paper. Dropping from

1.09 at 1 component to 1.063 at 3 components, the RMSEP of the response variable reaches its lowest at the 10 components. This suggest that 10 components is the best number of components for the PLS model.

Rank	Variable	Importance
1	Restaurant (Radius = 1600)	8.538945E-08
2	Park (Radius = 800)	6.948683E-08
3	Restaurant (Radius = 4800)	6.612163E-08
4	Bus (Radius = 1600)	6.025469E-08
5	Gym (Radius = 4800)	5.805304E-08
6	Restaurant (Radius = 800)	5.604880E-08
7	Gym (Radius = 800)	5.205775E-08
8	Bus (Radius = 800)	4.954927E-08
9	Gym (Radius = 1600)	4.595372E-08
10	Store (Radius = 4800)	3.890560E-08

Table 4: PLS Result for Variable Importance.

The PLS model results is shown in Table 3, showing the variables with top ten influencing abilities. The variable importance measure is the generalized cross-validation (gcv) estimate for each variable. This measure accumulates the reduction in the gcv statistic when more predictors are added. According to this measure, the top three explanatory variables are distance to restaurant (radius = 1600), distance to park (radius = 800) and distance to restaurant (radius = 4800). Accordingly, they have the importance measure of $8.538945E-08$, $6.948683E-08$, and $6.612163E-08$, the three largest among all variables. In other words, the reduction of prediction error when adding the variables of distance to restaurant (radius = 1600), distance to park (radius = 800), and distance to restaurant (radius = 4800). Looking at the result of the PLS model, distance to restaurant, park, bus and gym generally have the greatest explaining power to the response variable.

5.3.3 Geographic Weighted Regression

6 Discussion

Discuss how which VC method is better for GWR.

6.1 Comparison with Random Forest and Neural Network

6.2 Error Analysis

Discuss spatial clustering of residual(Moran's I) https://desktop.arcgis.com/en/arcmap/latest/tools/spatial-statistics-toolbox/cluster-and-outlier-analysis-anselin-local-moran-s.htm

6.3 Policy Advice

7 Conclusion

References

[1] Hao Wu, Hongzan Jiao, Yang Yu, Zhigang Li, Zhenghong Peng, Lingbo Liu, and Zheng Zeng. Influence factors and regression model of urban housing prices based on internet open access data. *Sustainability*, 10(5):1676, 2018.

[2] Santa barbara, california, Jun 2022.

[3] Manuel A. Zambrano-Monserrate, María Alejandra Ruano, Cristina Yoong-Parraga, and Carlos A. Silva. Urban green spaces and housing prices in developing countries: A two-stage quantile spatial regression analysis. *Forest Policy and Economics*, 125:102420, Apr 2021.

[4] Hao Wu, Hongzan Jiao, Yang Yu, Zhigang Li, Zhenghong Peng, Lingbo Liu, and Zheng Zeng. Influence factors and regression model of urban housing prices based on internet open access data. *Sustainability*, 10(5):1676, 2018.

[5] Anish Nahar. Google maps - the most expansive data machine.

[6] Place types, google developer.

[7] Evelyn Blumenberg and Fariba Siddiq. Commute distance and jobs-housing fit. *Transportation*, 2022.

[8] Bahman Lahoorpoor and David M. Levinson. The transit travel time machine: Comparing three different tools for travel time estimation, Nov 2019.

[9] Developer resources.

[10] Pedestrians and transit - safety: Federal highway administration.

[11] Story map series.

[12] Santa barbara county roads.

[13] Tobias G. Tiecke, Xianming Liu, Amy Zhang, Andreas Gros, Nan Li, Gregory Yetman, Talip Kilic, Siobhan Murray, Brian Blankespoor, Espen B. Prydz, and et al. Mapping the world population one building at a time. 2017.

[14] Home.

[15] Cheryl Hapke and David Reid. The national assessment of shoreline change:.

[16]

[17] Trevor Hastie, Jerome Friedman, and Robert Tisbshirani. *The elements of Statistical Learning: Data Mining, Inference, and prediction.* Springer, 2017.

[18] Tahir Mehmood, Solve Sæbø, and Kristian Hovde Liland. Comparison of variable selection methods in partial least squares regression. *Journal of Chemometrics*, 34(6), 2020.

[19] Il-Gyo Chong and Chi-Hyuck Jun. Performance of some variable selection methods when multicollinearity is present. *Chemometrics and Intelligent Laboratory Systems*, 78(1-2):103–112, 2005.

[20] Chris Brunsdon, A. Stewart Fotheringham, and Martin E. Charlton. Geographically weighted regression: A method for exploring spatial nonstationarity. *Geographical Analysis*, 28(4):281–298, 2010.

[21] Carol Rogers. The gaussian kernel.

[22] Mailman School of Public Health Columbia University. Geographically weighted regression.

[23] UrbanSpatial. Urbanspatial/public-policy-analytics-landing: All the data for the public policy analytics book by ken steif can be downloaded from this repo.