Violet Lingenfelter

Geoff Boeing

Big Data for Cities, PPUA 5262

14 April 2019

Final Paper: Luxurious Revere

Introduction

This paper will discuss luxury in Revere, MA., as calculated from tax assessor data. "Luxury" by definition is "the state of great comfort and extravagant living". The concept of luxury is often used by realtors to market certain living arrangements, i.e. "luxury apartments" or a "luxury lifestyle" as being the norm of a neighborhood. This concept of luxury has implications of class, and the concept of luxury apartments date back to the beginning of urban renewal projects. Because of these implications, I decided to create a latent construct to model luxury in the town of Revere, MA to see what I could find.

Methodology

To get at the luxury of homes in Revere from tax assessor data, I chose to aggregate several manifest variables. These variables include yard size of a parcel and the number of rooms. Other variables I would have liked to include but could not find a meaningful way to quantify include style of home and year built. I will discuss the implications of excluding these manifest variables in the discussion section of this paper.

To calculate yard size, I converted lot size from acres to square feet, and then subtracted the building size from the lot size to get the metric. I then calculated the yard size z-score for

each parcel as a way of normalizing the data. This metric is indicative of luxury because it shows potential for private greenspace, which is a hallmark of single home luxury.

To get a metric for number of rooms, I took the existing number of rooms value and normalized it by calculating the z-score for each entry. Again, this was to normalize the data and make it comparable to other variables. The idea behind including this metric is that the more rooms in a single family home, the more luxurious that home will feel. Take for example comparing the perceived luxury of a studio apartment compared to that of a three bedroom home. The home with more rooms, particularly if some of the rooms are for recreation as opposed to living space (an office or studio in a spare room vs. having a child living in the spare room), will feel more luxurious.

In order to make these categories comparable, I chose to find the z-score for each manifest variable. I chose z-scores because that metric is robust to underlying distribution and scales the data so they are comparable to each other. Once I assigned values for each variable and calculated the z-scores, I added these values together to calculate a 'luxury' score. This luxury score therefore is a composite z-score of yard-size, which I calculated, and number of rooms, which was given in the data. I specifically chose to exclude any data about the value or last prices of the parcels, because I am hoping that the development of a “luxury” construct could be used as a predictor for these variables.

I chose to examine how well this latent construct worked for our data by creating and examining a correlation matrix and making a linear model to predict for luxury using variables that were not included in the creation of the construct. For the linear model, I chose to include in the array of explanatory variables total value, residential area, number of units, and last sale price

of the parcel. Total value and last sale price were included because more expensive properties tend to be perceived as luxurious. Residential area was included because more spacious living areas are generally perceived as nicer to live in.

It is also important to note that this analysis looks only at single family homes in Revere, as found by their use code in the data. This means that in theory, the number of units should be one. From examining the data, it was clear to me that there were inconsistencies in the use code that the parcels were filed under and the number of units each parcel contained. This may be due to "granny flats", or a part of a home that was renovated to be self contained and suitable for an individual to live in. I would have liked to include this in the analysis but did not because it the number of units in theory should have been one for each single family home. Total number of units should be included because it is a marker of how many families are living on a property, and at least in the US, multiple families living in the same home is generally considered a negative condition.

Results

	BLDG_VAL	LAND_VAL	TOTAL_VAL	LS_PRICE	UNITS	RES_AREA	NUM_ROOMS	LUXURY
r.BLDG_VAL	1.000	0.347	0.919	0.218	0.064	0.793	0.485	0.468
r.LAND_VAL	0.347	1.000	0.688	0.073	0.023	0.306	0.184	0.324
r.TOTAL_VAL	0.919	0.688	1.000	0.198	0.059	0.744	0.455	0.502
r.LS_PRICE	0.218	0.073	0.198	1.000	0.013	0.127	0.094	0.067
r.UNITS	0.064	0.023	0.059	0.013	1.000	0.076	0.160	0.107
r.RES_AREA	0.793	0.306	0.744	0.127	0.076	1.000	0.479	0.420
r.NUM_ROOMS	0.485	0.184	0.455	0.094	0.160	0.479	1.000	0.717
r.LUXURY	0.468	0.324	0.502	0.067	0.107	0.420	0.717	1.000

Figure 1: Table of our correlation matrix including building value, land value, total value, list price, number of units, residential area, number of rooms, and our latent construct, luxury.

Fig. 1 shows the correlation matrix calculated with these variables. From this we can see that the variable most highly correlated with luxury is the number of rooms in the residence on a parcel, which makes sense considering a derivation of that metric was included as a manifest variable for our latent construct. Very lowly correlated with the construct for luxury is last list price. This perhaps indicates that the way the construct was made is missing some key pieces of what makes parcels luxurious, assuming that people will pay for luxurious properties. Also of note is that last list price is not strongly correlated with any of the variables included in our matrix. This is of note because it illustrates what seems to be a disconnect in how properties are assessed versus how they sell on market. This should be investigated in a separate analysis.

```
Call:
lm(formula = LUXURY ~ TOTAL_VAL + RES_AREA + LS_PRICE, data = luxury)

Residuals:
    Min      1Q  Median      3Q     Max
-5.2420 -0.7418 -0.1039  0.5990 17.6112

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.137e+00  8.678e-02 -36.146  < 2e-16 ***
TOTAL_VAL    9.930e-06  4.486e-07  22.135  < 2e-16 ***
RES_AREA     2.324e-04  4.308e-05   5.396 7.16e-08 ***
LS_PRICE    -2.999e-07  1.313e-07  -2.283   0.0225 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.237 on 4488 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.2579, Adjusted R-squared:  0.2574
F-statistic:   520 on 3 and 4488 DF,  p-value: < 2.2e-16
```

Figure 2: The results of our regression model

Fig. 2 shows the output of our regression model. All of the variables included in this model were significant, but our R-squared value was a meager 0.25, indicating that only 25% of the variation in our construct for luxury can be explained by variation in our model. By examining our P-values, we can see that the explanatory variables we included were worthwhile, as they were all significant at the $\alpha=0.05$ level. Our F statistic has a $p < 2.2\times10^{-16}$, allowing us to reject the null hypothesis that there is not significant linear interaction between our explanatory variables and our response variables.

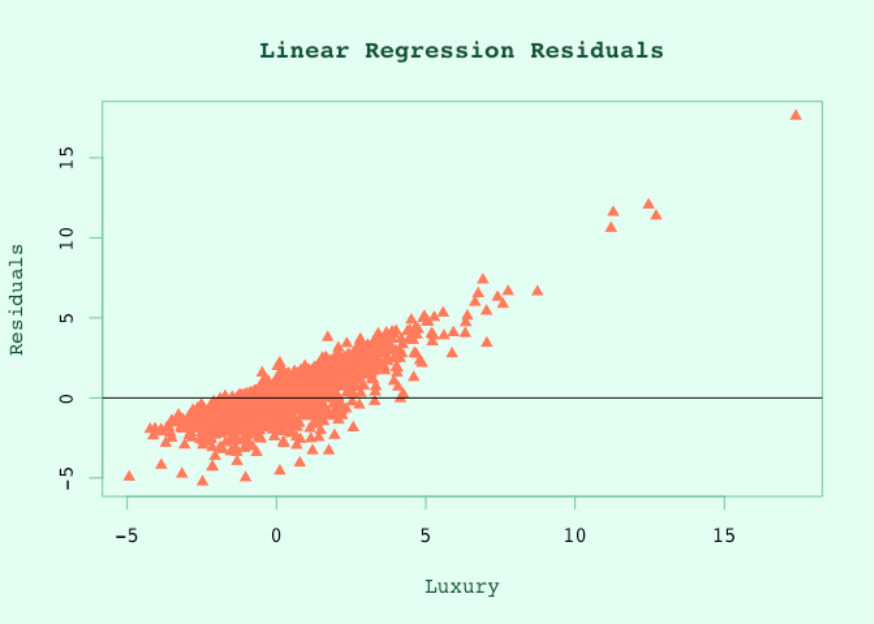

Figure 3: Residuals plotted against the luxury construct (observed variable in our model).

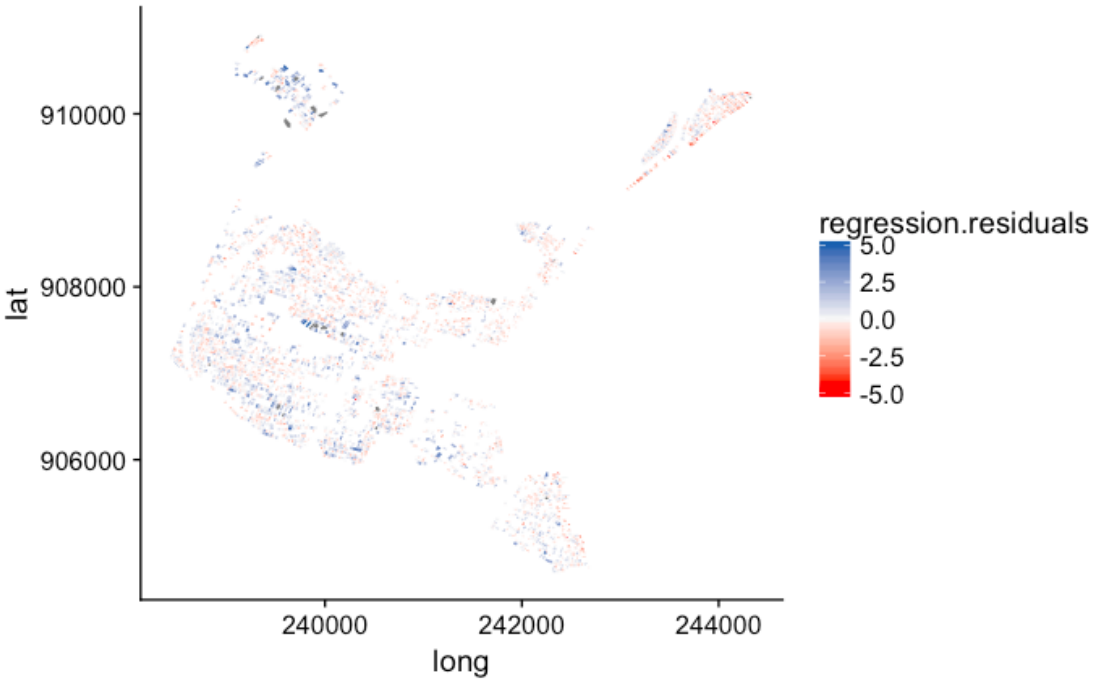

Figure 4: Geographical representation of the residuals for our model

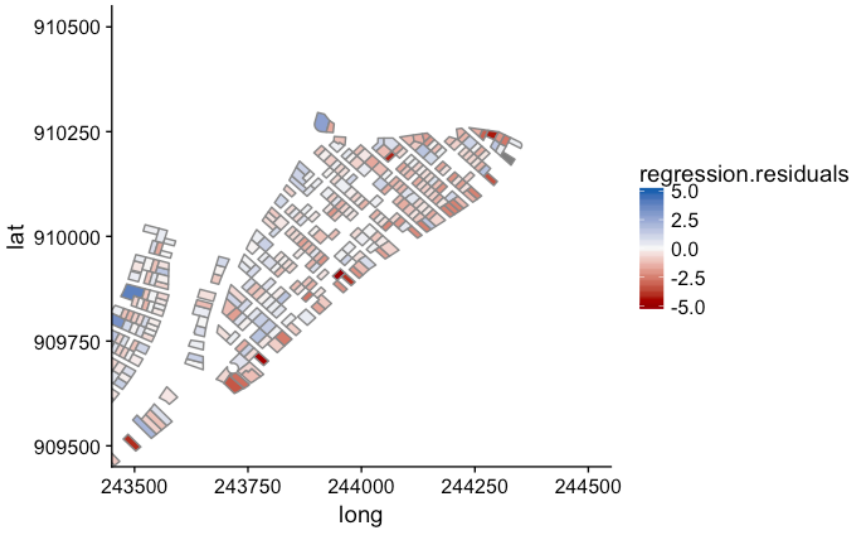

Figure 5: Figure 4 zoomed into show detail at the north east tip of the city.

Fig. 3 and Fig. 4 allow us to take a further look at the residuals of our model. Fig. 3 is a residual plot and from this plot we can see that our model is not a good one. The residuals are

clearly non-random and increase as our construct for luxury increases. This indicates heteroscedasticity, and means that our model was probably not the best.

Fig.4 shows the geography of the residuals. This is useful because it allows us to look for geographic clustering of our residuals, and determine if we should be using a geographic model. Looking at the geographic clustering of residuals allows us to see if perhaps the location of the parcel is affecting how well our model fits. Figure 5 shows a zoomed in portion of the map of the residuals, with the focus set to the north eastern corner of the city. This area has a cluster of low residual values. This means that our model is over estimating the luxury of properties in this geographic cluster. This indicates that perhaps a model that takes geographic location into account is necessary.

Discussion

This analysis shows that the model we made for luxury is flawed. The data was heteroscedastic, making it a bad candidate for the type of modelling we used. The data would benefit from being fitted to a model that takes geography into account.

Jumping off points from this analysis include making this model of luxury more dependent on the neighborhood in which the parcel lies. This could include measuring distance from each parcel to the nearest public and private school, distance to business districts, to the coastline, or to cultural hubs such as churches or community centers. This model could also benefit from the inclusion of categorical variables such as the style of the home, the neighborhood that the home is in, or whether the home is a rental or owned by its inhabitant. Combining tax assessor data with data from the census may also prove useful.

Further investigation should be put into determining whether this model would benefit from being spatially weighted. Given the way luxury works in neighborhoods, it is logical to assume that nearness to other luxurious parcels enhances the luxury of any given parcel.