Neighborhood clustering around Boston

2015 crimes committed and property values

Raaid Arshad and Michael Clawar

Abstract:

Valuing properties and understanding the driving forces behind their value is difficult to determine in cities. To do so with an appropriate amount of resolution is even more challenging. We attempt to find a spatial relationship between property values with a fitting amount of granularity. Our approach is to use a k-nearest neighbors valuation model for Boston neighborhoods. We found certain “pockets of value” across the downtown area of Boston, such as Beacon Hill and waterfront properties.. Our results demonstrate how significantly a property’s location within a particular area of the city influences its value and that we are able to explain approximately two-thirds of the variability in our data.

1. Introduction

Valuing properties across a city is a difficult, if not impossible, task that requires expertise in the area. Because property values can vary significantly from block to block, individuals unfamiliar with both local and national real estate trends may struggle to understand the driving forces behind valuation. In our paper, we propose a relatively simple and computationally efficient method to identify "pockets of value" across Boston, MA, which can be used in further research as a way of modeling neighborhood- and geographic-specific impacts, instead of property-level impacts. Because medium- to large-scale real estate developers and government policymakers are likely to be more interested in generalized local housing trends rather than, e.g., a home with a bay window sells for $300 more, our approach of k-nearest neighbors valuation has potential value in neighborhood analysis.

Using data from the City of Boston's 2015 property assessments, we train a *k*-nearest neighbors regression on a training set, and are able to explain nearly two-thirds of the variability in the test set. As a baseline, we believe this is a productive start. Our model reflects the real estate mantra of "location, location, location" that drives pricing, as the neighborhood is the major driver of property values.

2. Methodology

Our data set consists of 27,679 observations from the 2015 property assessment by the City of Boston. We split these into 100 training and test set using a rough 80/20 rule. Each observation has an associated latitude, longitude, and assessed property value (in USD per square foot).

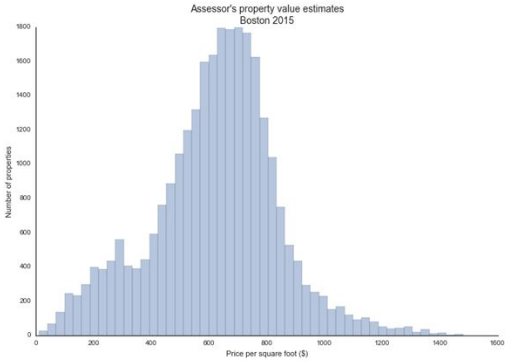

Figure 1: Distribution of property values ($/sqft)

Metric	Price per square feet ($)
25th percentile	$499
Median	$642
75th percentile	$755
Mean	$623

Table 1: Summary statistics for property values

Our model attempts to identify and find a spatial relationship between property values. k-nearest neighbors is a simplest model for nonparametric spatial relationships, and we define the k-nearest neighbors as the k-closest properties using the cartesian distance of latitude and longitude, or

$$\mathbf{k}\text{-nn}_{\mathrm{j}} = \text{argmin}_{i \in I} \sqrt{(lat_i - lat_j)^2 + (lon_i - lon_j)^2}$$

For a small area, cartesian distance is a reasonable approximation. However, in spatial analysis with distances larger than blocks in downtown Boston, other spherical approximations would be more appropriate.

We then estimate an optimal number of neighbors to use in the k-nearest neighbors spatial regression to balance fit and smoothing. The number of neighbors in possible models range from 1 to 30.

The modeling flow to select the number of neighbors is given by:

```
for $test_i$, $train_i$ in test sets, train sets:
    for k in range(1, 31):
        fit k-nn model on $test_i$
        predict $train_i$ using k-nn model
        estimate score as $R^2$ of k-nn predictions
```

Then select the **k** which maximizes the mean R^2 of the training set predictions.

3. Results

We find an optimal **k**-number of neighbors to be 5, which gives an average R^2 of roughly 0.64 (*Figure 2*). This means that about two-thirds of the variability in prices in Boston can be explained by the nearest five properties

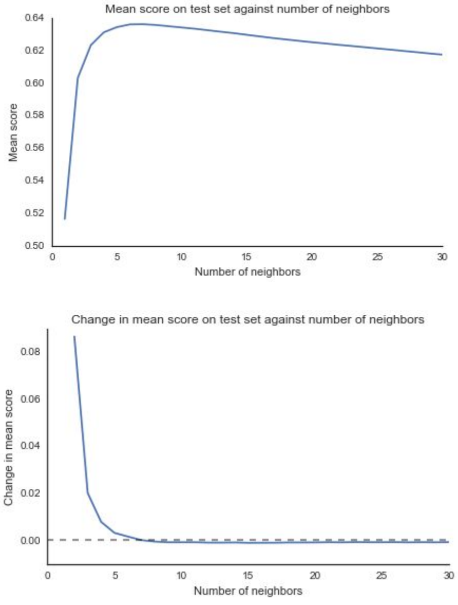

Figure 2: Model scores for varying number of neighbors

Predictions are generated on a grid over the *lat-lon* bounds of the data set and mapped using **mplleaflet**, **Leaflet**, and the CartoDB positron basemap (*Figure 3*).

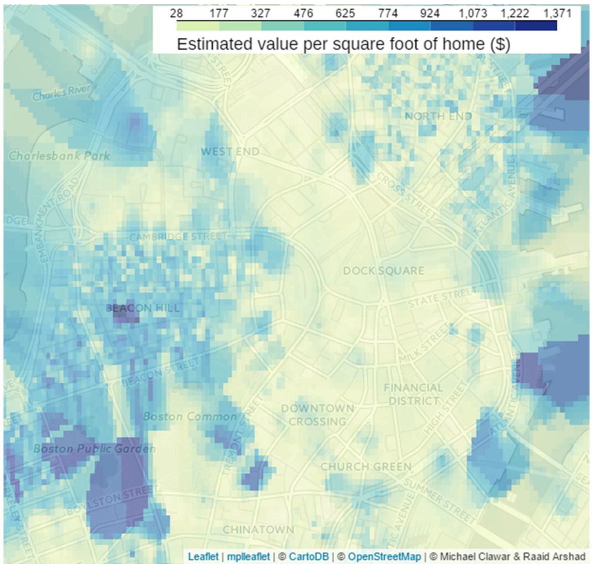

Figure 3: Model predictions for home prices in the area of downtown Boston

Beacon Hill is most obviously an expensive area, while Chinatown and East Boston are relatively less expensive. In general, it matches expectations that waterfront properties would be relatively more expensive per square foot than other properties in Boston. Pockets in the North End are also more expensive, as a mix between somewhat more upscale housing and commercial properties.

The center of the downtown area, composed of mostly commercial buildings, appears to be less expensive per square foot than the residential areas surrounding it.

In addition, the waterfront properties in East Boston are relatively more expensive, as are pockets in the North End. Our model also predicts the Boston Public Garden to be relatively expensive, despite having no property in the area. This approach may be a helpful path to pursue for developers or government in valuing public or undeveloped land.

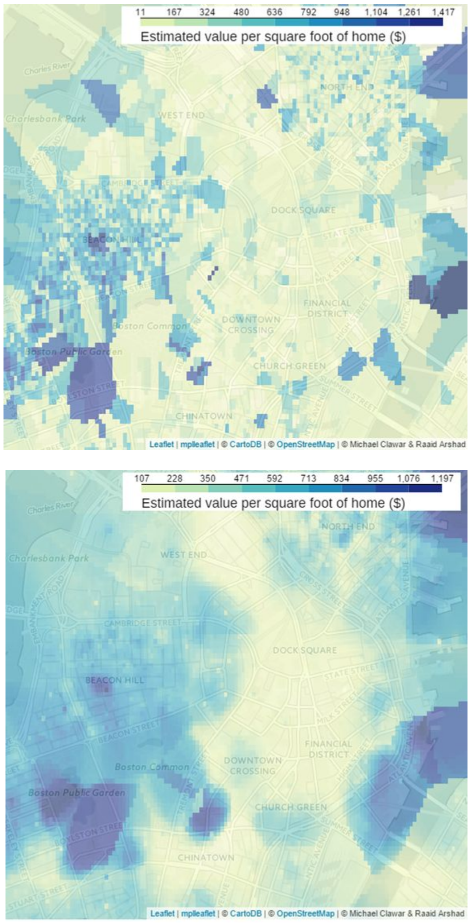

Figure 4: Model predictions with k=1, 30

Figure 4 shows predictions from the model with *k*=1 (top) and *k*=30 (bottom). With only 1 neighbor used to predict the assessed property value, the model overfits, and is subject to variability from the nearest neighbor, with only a 0.52 R^2. Predictions for the Beacon Hill area in particular are very noisy, suggesting that the variability in individual property values is relatively large in the neighborhood.

With 30 neighbors, the model tends to oversmooth, and may be more useful for identifying larger, general neighborhoods, rather than small pockets of value. For example, we can easily identify waterfront properties as a whole to be more expensive, the North End to have a cluster of more expensive housing, and Beacon Hill to still be the most expensive properties in the downtown area.

4. Conclusion

Our data reveals clusters of valuable properties in the downtown Boston area, concentrated in the Beacon Hill and waterfront neighborhoods. Using a 5-nearest neighbors regression, we explain over roughly two-thirds of the variability in property values (per square foot) in downtown Boston.

We also predict higher property values within the Boston Public Gardens (around $650 to $1,000 per square foot), if construction were permitted to be built in the Gardens.

Future work could focus on extending the regression to use other factors beyond geographic closeness, as well as extending and testing the predictions on new buildings. This may prove useful in valuing land and property bids, and provide simple, fairly accurate estimation of baseline property values. In general, this model could also help individuals make estimates of their own property's assessed value and prepare taxes accordingly.

Packages used

Data management packages:

- Pandas
- NumPy
- Scikit-learn
- PyMongo

Visualization packages:

- Matplotlib
- Mplleaflet

Data sources

- Boston Public Schools
  http://data.cityofboston.gov/resource/e29s-ympv
- Crime Incidents
  http://data.cityofboston.gov/resource/7cdf-6fgx
- Hospital Locations
  http://data.cityofboston.gov/resource/46f7-2snz
- Property Assessment
  http://data.cityofboston.gov/resource/yv8c-t43q