NYC_analysis

Maxime Rischard

October 5, 2016

Contents

The New York City school district is itself divided into sub-districts. From now on, by "district", I will be talking about these sub-districts. Residents are guaranteed a slot in a school within their district, but there is some system whereby a child can apply to attend schools outside of the district, which might possibly dampen the treatment effect.

Our goal is to detect discontinuities in the price of housing across the school district boundaries.

1 Dataset

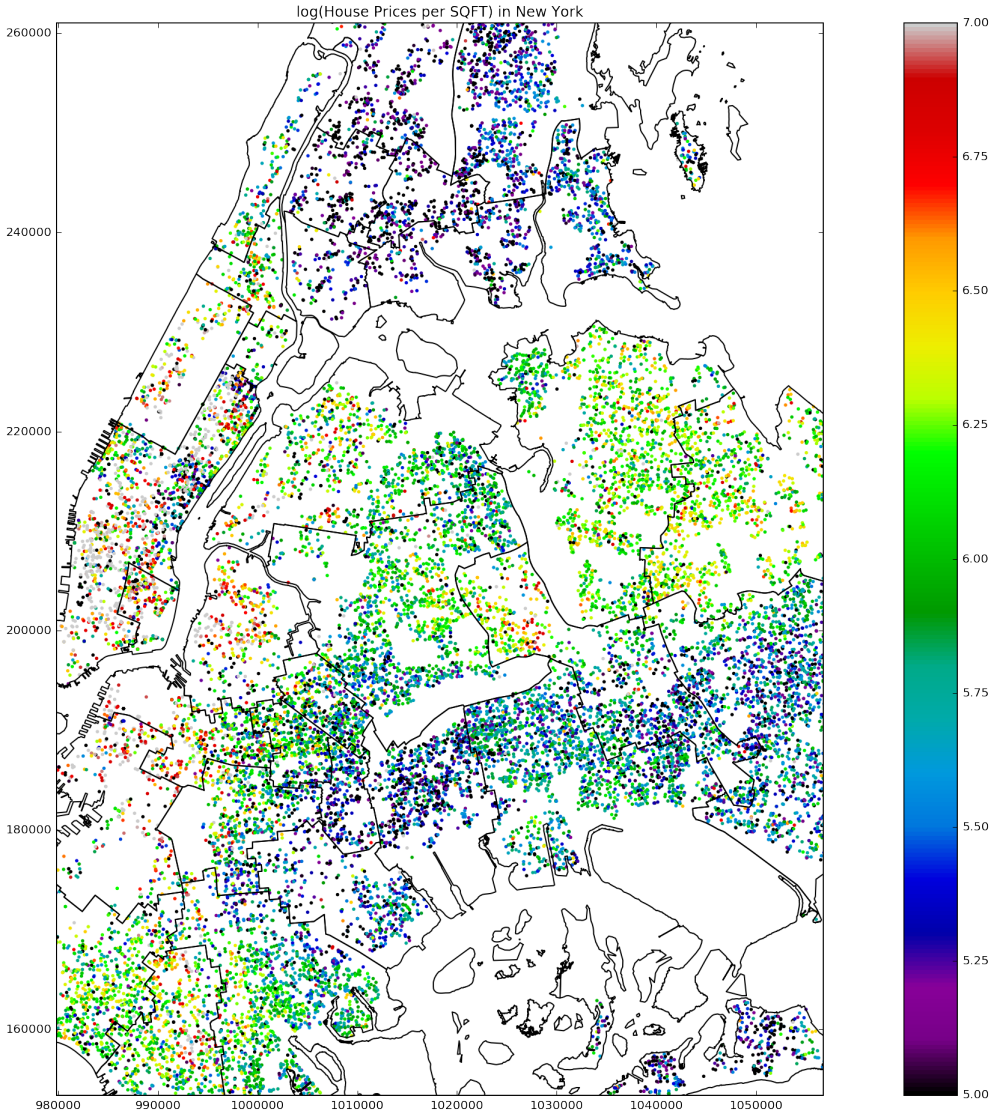
New York publishes a year's worth of house sales on a rolling basis. Annoyingly, this means that data that is older than a year disappears, and I've not been able to find this data online. But New York is a density, so even just a year's data is quite a large dataset.

I also have some covariates, though somewhat less rich than I found in Tucson. What I have is square footage, tax class and building class. I've geocoded the address of each sale to get a latitude and longitude, and projected the coordinates onto a Euclidean plane in order to be able to calculate distances.

1.1 Data cleaning

I remove the following sales:

- any sale with missing data in the sale price, square footage, property covariates, geographical coordinates (due to failed geocoding)
- sales outside of any NYC school district
- properties smaller than 100 sq ft
- outliers in the price per sqft, which I defined as sales outside of $3 < Y_i < 8$


log(House Prices per SQFT) in New York
260000
240000
220000
200000
180000
160000
980000
990000
1000000
1010000
1020000
1030000
1040000
1050000
7.00
6.75
6.50
6.25
6.00
5.75
5.50
5.25
5.00

sales map

I remove tiny properties because I feel like they might behave differently than the rest of the market. Removing outliers is possibly contentious, but it makes sense to me that some sales between friends or family members could have a very low Y_i that isn't a true representation of the market price, and that some high prices could be equally disconnected from the market.

This leaves 23727 out of 56815 sales records in NYC, mostly because of properties that don't have a reported gross square footage.

2 Model

I decided to define the outcome to be the log price per square foot. The model is a Gaussian Process in the spatial covariates on top of a ridge regression on the property covariates (building and tax class). Within a school district we could write the model as [suggestions for clearer notation welcome]:

$$Y_i = \log\left(\frac{\texttt{SalePrice}_i}{\texttt{SQFT}_i}\right) = \beta_0 + \beta_{1\texttt{TaxClass}[i]} + \beta_{2\texttt{BuildingClass}[i]} + f(\mathbf{x}_i) + \epsilon_i \tag{1}$$

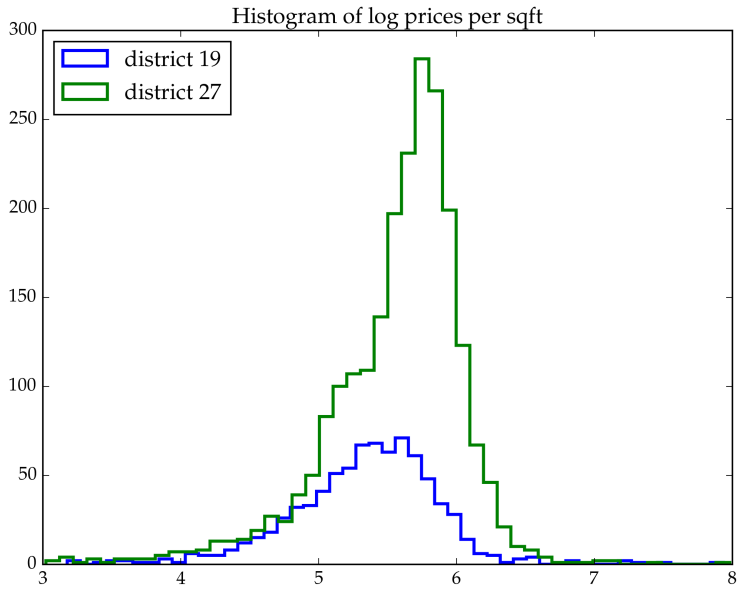
$$\epsilon_i \sim \mathcal{N}\left(0, \sigma_y^2\right) \tag{2}$$

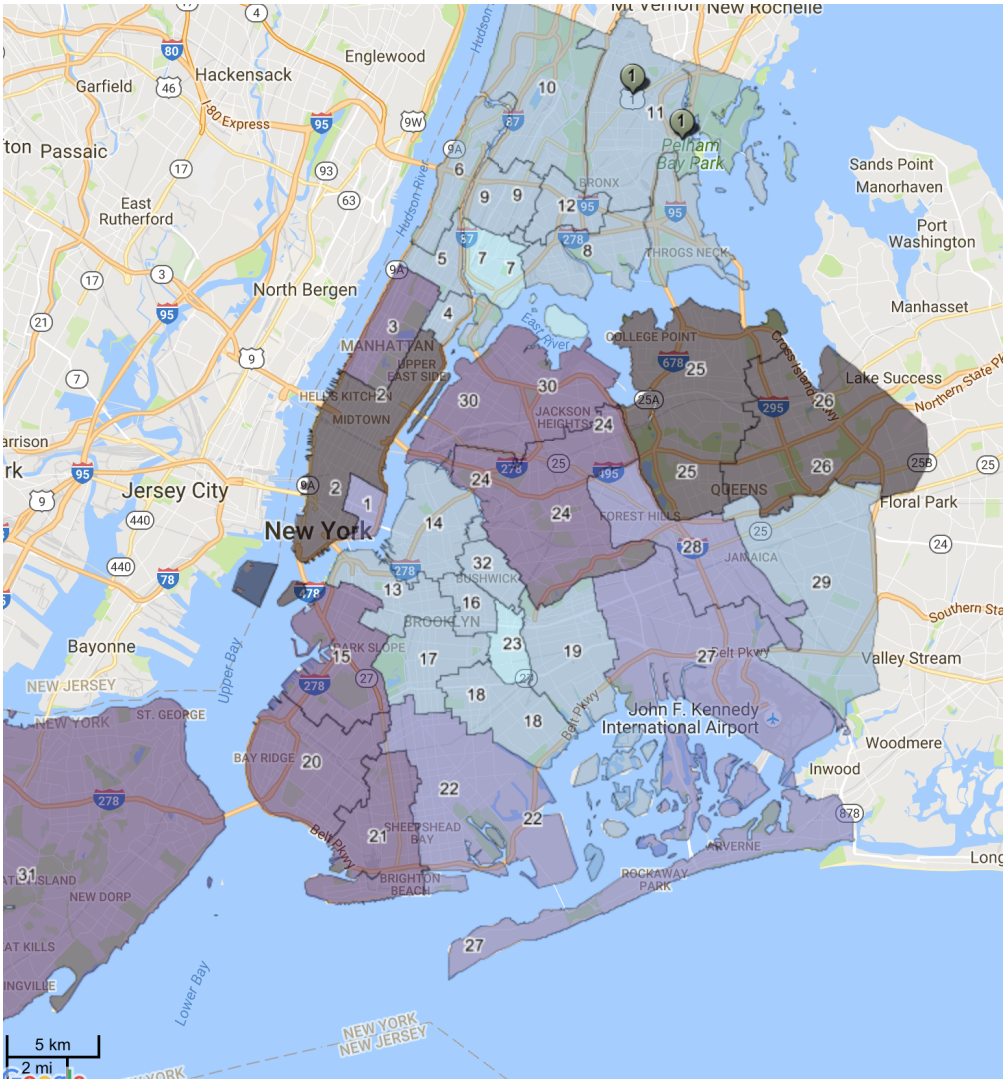
$$\beta_{1j}, \beta_{2j} \sim \mathcal{N}\left(0, \sigma_\beta^2\right) \tag{3}$$

$$f(\mathbf{x}_i) \sim \mathcal{GP}\left(0, k(\mathbf{x}, \mathbf{x}')\right) \tag{4}$$

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left\{-(\mathbf{x} - \mathbf{x}')^\intercal(\mathbf{x} - \mathbf{x}')/2\,\ell^2\right\} \tag{5}$$

A visual inspection of the house sales map above led me to focus on the boundary between districts 19 and 27. I found a map online of average maths performance in each school district, which shows that districts 19 and 27 are quite different. It's important to note that the boundary between the two districts is also part of the boundary between Brooklyn and Queens, so we won't be able to attribute a causal effect solely to the difference in school districts. A histogram of Y in both districts also shows that marginally the house prices are very different.




Englewood
Hackensack
Garfield
Passaic
East Rutherford
North Bergen
Jersey City
New York
Bayonne
MANHATTAN
UPPER EAST SIDE
HELL'S KITCHEN
MIDTOWN
JACKSON HEIGHTS
COLLEGE POINT
QUEENS
FOREST HILLS
JAMAICA
BUSHWICK
BROOKLYN
PARK SLOPE
BAY RIDGE
SHEEPSHEAD BAY
BRIGHTON BEACH
ROCKAWAY PARK
THROGS NECK
BRONX
Pelham Bay Park
Sands Point
Manorhaven
Port Washington
Manhasset
Lake Success
Floral Park
Valley Stream
Woodmere
Inwood
John F. Kennedy International Airport
Hudson River
East River
Upper Bay
Lower Bay
Belt Pkwy
Cross Island Pkwy
Northern State Pkwy
Southern State Pkwy
NEW JERSEY
NEW YORK
ST. GEORGE
NEW DORP
5 km
2 mi

districts

3 Optimize hyperparameters using district 27

I start by optimizing the hyperparameters β_0, σ_β, σ_f, ℓ and σ_y using only data for district 27 (the larger of the two, with 2249 sales. Optimizing within a single district makes it computationally quicker and easier to implement, and ensures that there is no interference from the treatment effect.

The optimization takes a very reasonable 20-30 seconds.

```
Results of Optimization Algorithm
 * Algorithm: Conjugate Gradient
 * Starting Point: [-1.6094379124341003,5.560919115673984, ...]
 * Minimizer: [-0.8487823942160716,5.5608570663288095, ...]
 * Minimum: 1.379086e+03
 * Iterations: 6
 * Convergence: true
   * |x - x'| < 1.0e-04: false
   * |f(x) - f(x')| / |f(x)| < 1.0e-03: true
   * |g(x)| < 1.0e-08: false
   * Reached Maximum Number of Iterations: false
 * Objective Function Calls: 32
 * Gradient Calls: 25
```

The fitted hyperparameters are:

Parameter	Fitted Value
σ_y	0.4279
σ_f	0.1716
σ_β	0.4360
ℓ (feet)	3425.2545

4 Get posterior mean $\hat{\beta}$

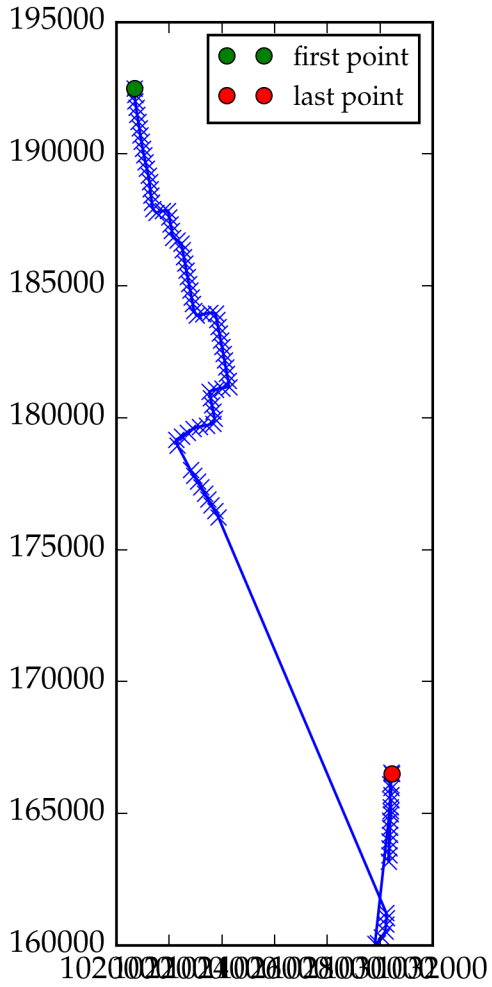
With these hyperparameters in hand, I extract the posterior mean of the linear regression parameters β_{1j} and β_{2j}, again only using district 27 data. This makes our lives easier, and I've convinced myself in other examples that it makes very little difference.

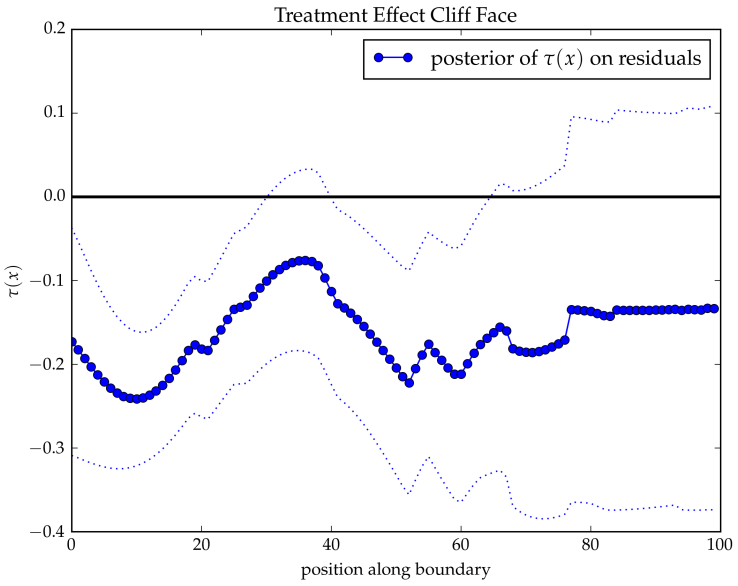
Obtaining the posterior means is pretty quick, about 0.5 seconds.

5 Fit GPs to residuals

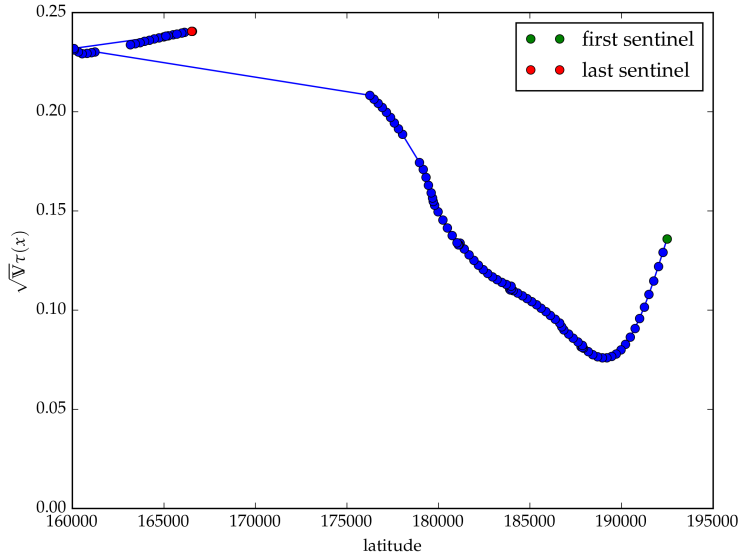
I then go back to the other school districts, obtain their residuals from the linear regression, and fit $\mathcal{GP}$'s to the residuals (using the hyperparameter values fitted on district 27), with constant mean set to the mean of the residuals. I then look at the boundary between districts 19 and 27, using our machinery to obtain a treatment effect cliff face, and an inverse-variance weighted average.

For context, here is a plot of my sentinel points between the two districts. The districts are on the coastline, so some of the boundary runs in the water, and then there's some sentinels on an island in Jamaica Bay. In the treatment effect cliff face that follows, we see a corresponding discontinuous jump in $\tau(x)$ and its posterior variance.


first point
last point
195000
190000
185000
180000
175000
170000
165000
160000


Treatment Effect Cliff Face
posterior of $\tau(x)$ on residuals
$\tau(x)$
position along boundary
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
0
20
40
60
80
100

The following plot just shows the posterior standard deviation against the latitude of the sentinel points (the Northmost sentinel point is on the right of this plot, so it's horizontally flipped from the cliff face plot above). We see that the standard deviation is lowest in the most densely populated part of the boundary, and very high on the island.



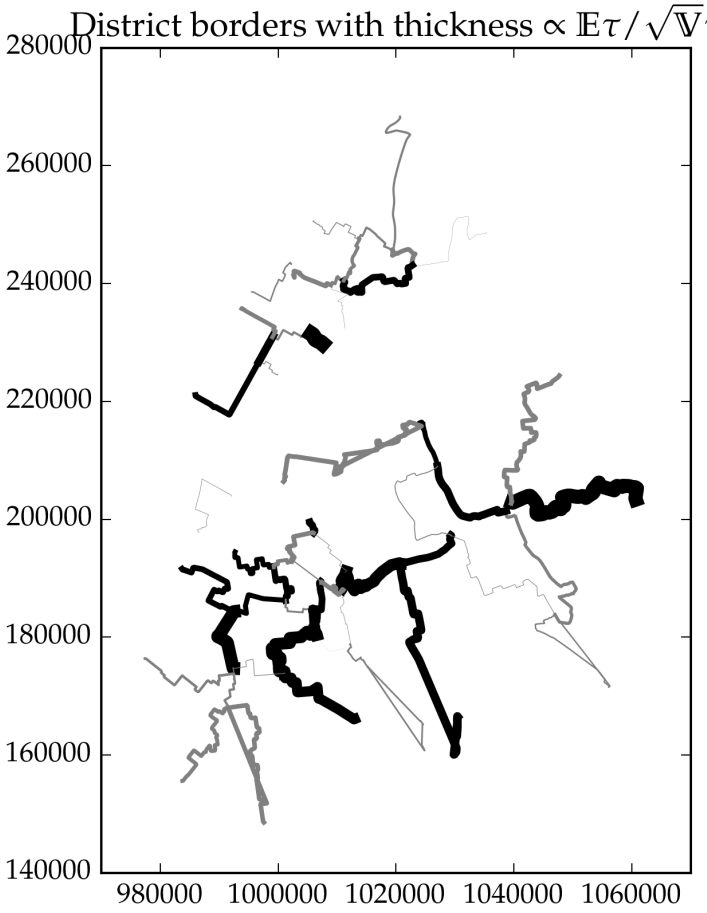
The inverse-variance treatment effect is very strong, we get:

$$\tau \mid Y \sim \mathcal{N}(\mu = -0.174, \sigma = 0.049) \tag{6}$$

$$\mathbb{P}(\tau > 0 \mid Y) = 0.022\% \tag{7}$$

6 Pairwise treatment effect

Beyond districts 19 and 27, I now look at every pair of contiguous NYC school districts, and compute the inverse-variance treatment effect. I then draw a map of all the district boundaries with the thickness of the boundaries drawn proportionally to the effect size $\mathbb{E}\,\tau/\sqrt{\mathbb{V}\,\tau}$. Boundaries with τ at least two standard deviations away from 0 are shown in black.



7 Next steps

- Fit hyperparameters on more districts than just the 27th
- Decide if this is a sufficiently convincing example to be used in the paper
- Revisit once we've implemented Bayesian inference on hyperparameters