CS281: Advanced Machine Learning

Predicting Intersection-Specific Accidents in NYC

Irene Chen and Brian Zhang

HARVARD School of Engineering and Applied Sciences

Introduction

Given an unseen intersection in New York City, our goal is to predict how many accidents occur there over a given time period.

Crash analysis has been performed on other cities (London, Philadelphia), but the lack of data has prevented similar analysis on NYC. Accurate and efficient prediction of accidents would allow city officials to allocation resources like additional signage, medical help, or increased surveillance.

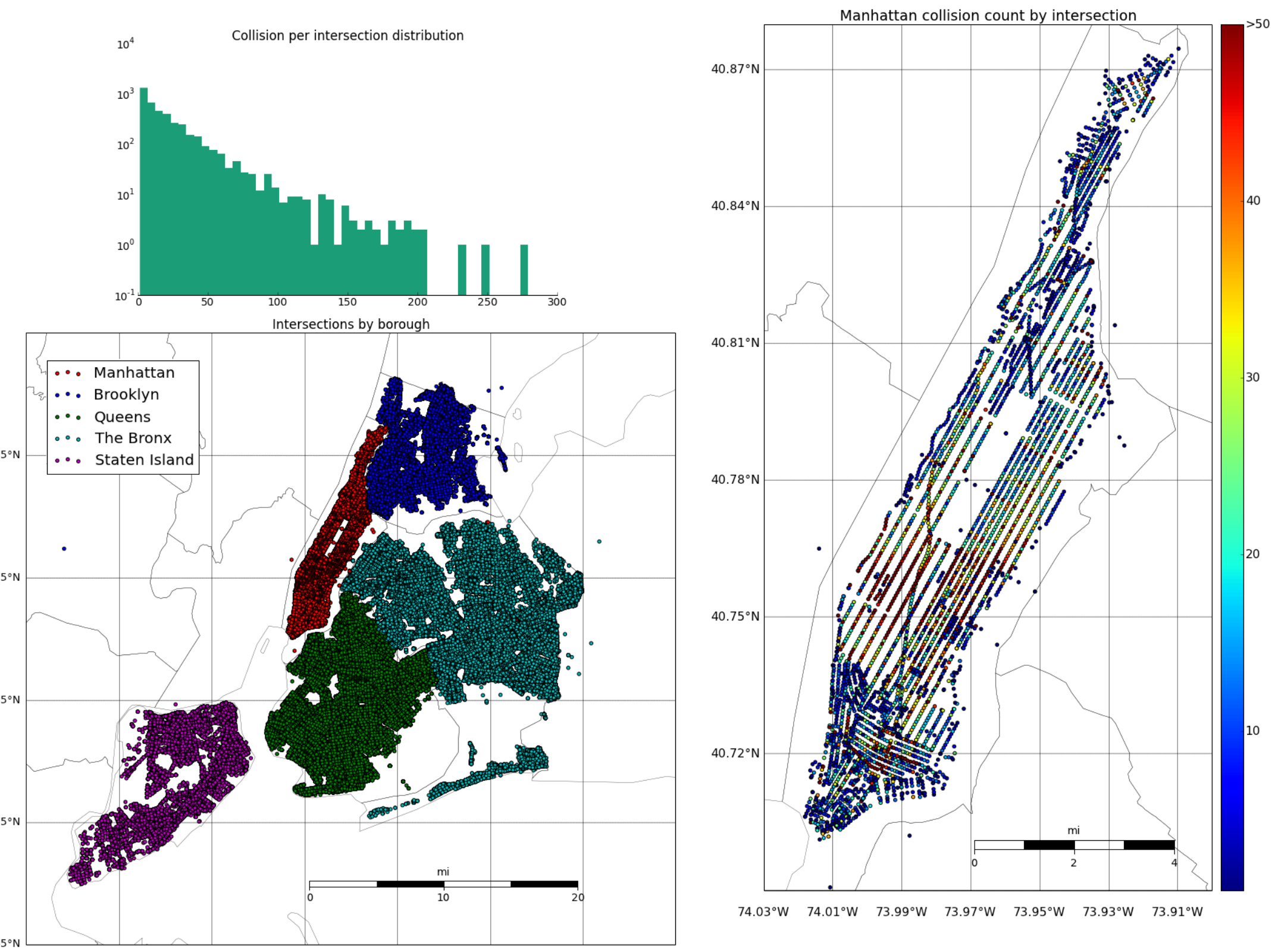

Figure 1: (Upper Left) Distribution of collisions per intersection. (Lower Left) The five boroughs of New York City as covered by our crash data. (Right) The crash density of Manahattan.

Dataset

Our crash dataset covers the five boroughs of Manhattan and gives the number of intersection-specific collisions for over 250,000 collisions in NYC between Aug 2011 and Sept 2013.

The data was recently released by the New York City Police Department. Because the NYPD has historically been resistant to releasing NYC crash data since it could be manipulated by people who "want to make a point of some sort," the data has recently been made available through an open source project. [1] [3]

After examining the data, we chose to focus on Manhattan since the mean collisions per intersection was the highest of the five boroughs and because Manhattan is spatially contiguous. Additionally, we decided to aggregate the data over the entire time period.

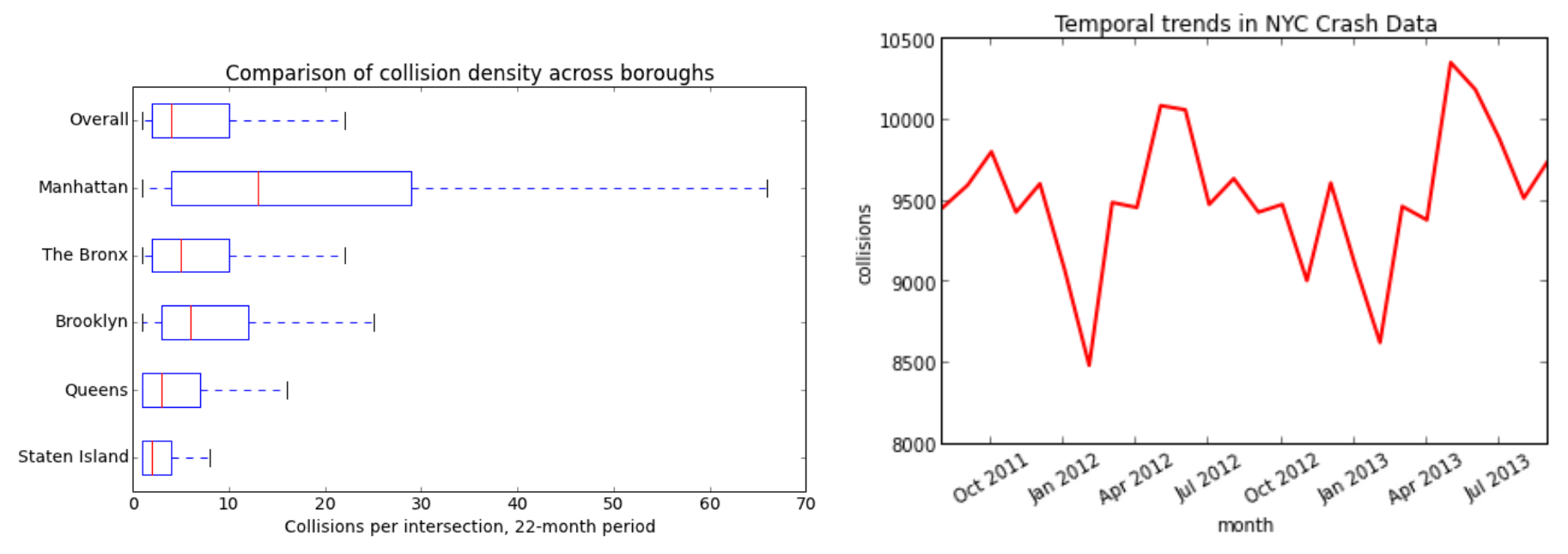

Figure 2: (Left) Box-and-whiskers plot of crash distribution according to borough.(Right) Temporal changes in collisions over the entire period of the dataset.

Performance Baselines

To determine performance of each model, we examined the negative log likelihood per data point NLL/N over N intersections assuming a Poisson distribution of collisions per intersection.

$$NLL/N = -\frac{1}{N} \cdot \sum_{i=1}^{N} \log\left(Poisson(k_i, \lambda_i)\right)$$

where at intersection i, λ_i is the predicted value of collisions and k_i is the actual value of collisions. We used the negative log likelihood per data point to make the values more manageable and interpretable.

Our two baselines were guessing the average of the seen data (using the average of the training data with 5-fold cross validation) and guessing perfectly, which we compare against our K-nearest neighbors approach in Figure 3.

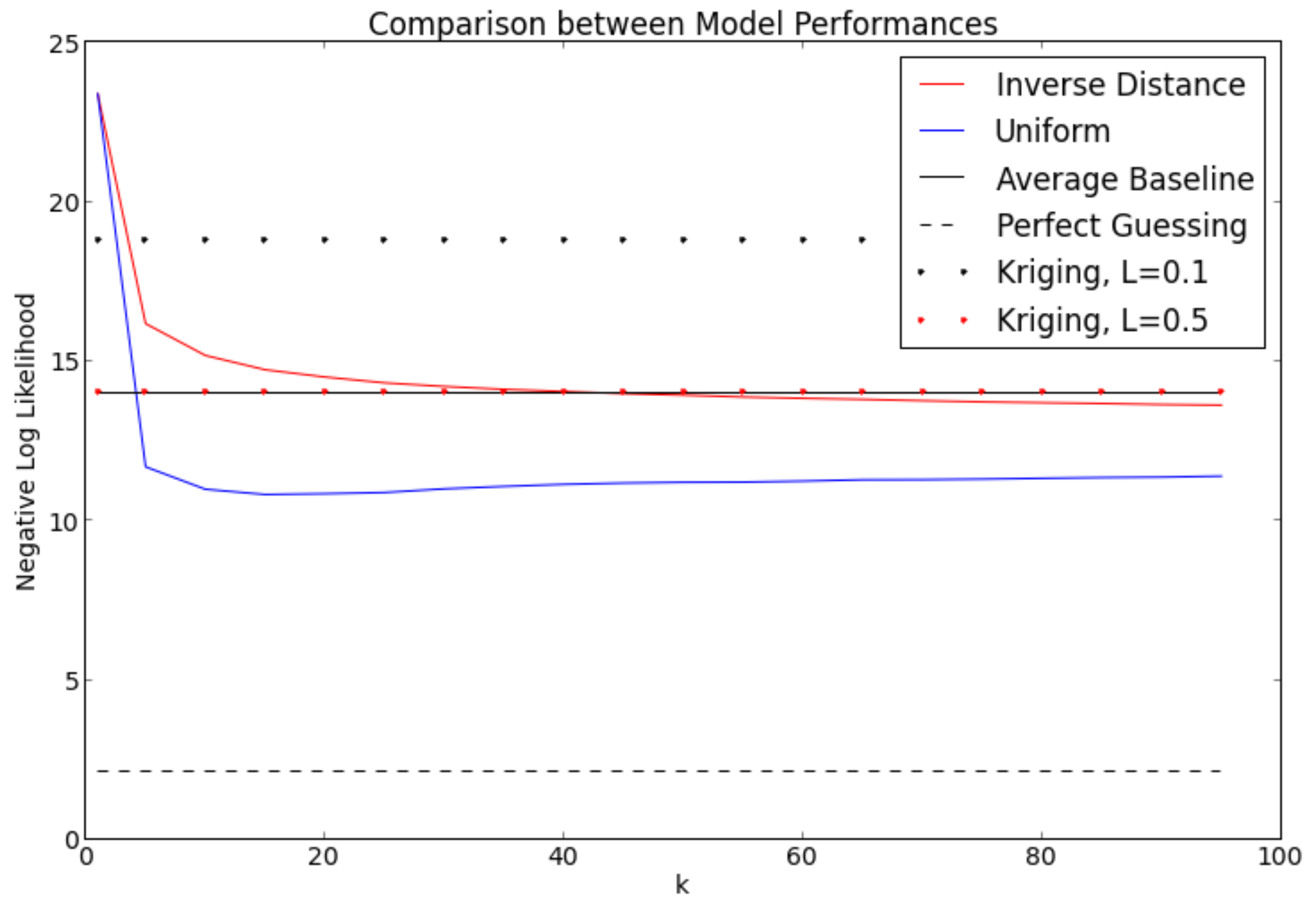

Figure 3: Performance of K-nearest neighbors for inverse distance and uniform weighting, kriging with two length scales, the average baseline, and perfect guessing.

K-Nearest Neighbors

Additionally, we used the K-nearest neighbors to predict the count for an unseen intersection.

For each new data point, we computed the Euclidean distance to ever previously seen intersection, ordered them in increasing distance, and used either the uniform average or inverse distance weighted average to predict the number of collisions.

As before, we calculated the negative log likelihood per data point, and the comparison with our baseline is shown in Figure 3.

Kriging

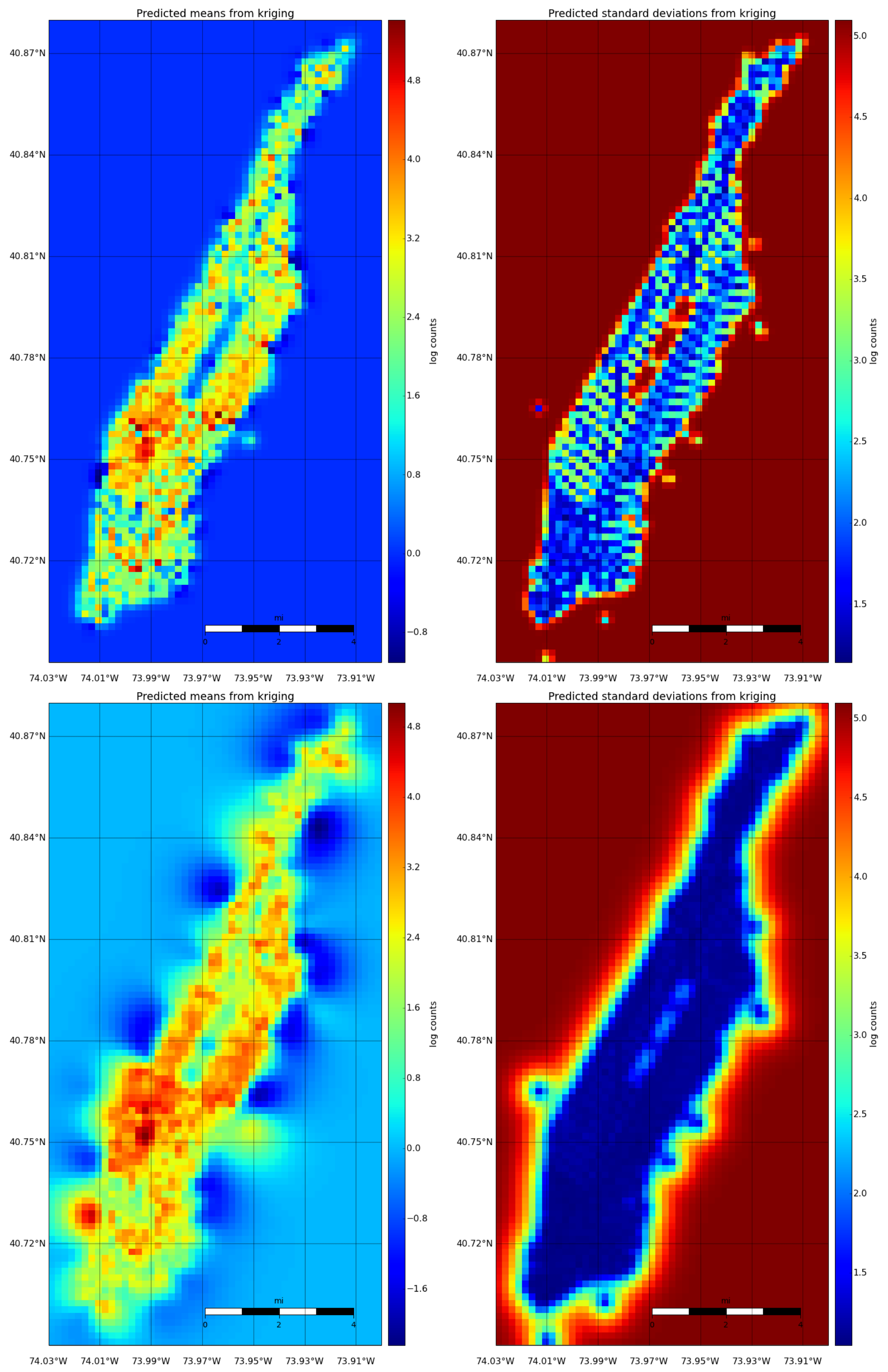
Predicted means from kriging
Predicted standard deviations from kriging
log counts

Figure 4: Predicted means (left) and standard deviations (right) using kriging over a mesh grid of Manhattan. Two length scales are shown: $l = 0.1$ miles (top row) and $l = 0.5$ miles (bottom row).

Kriging is the spatial statistics term for a 2D Gaussian process. We used a Matern kernel with $\nu = 3/2$ and fit it to log-count data. If we allow for noisy observations, then the kernel function can be written as

$$\kappa_y(x_p, x_q) = \sigma_f^2 \mathrm{Matern}_{3/2}(x_p, x_q, l) + \sigma_y^2 \delta_{pq}$$

Here σ_f corresponds to the variance of the function, σ_y is the noise in our observations, and l is the relevant length scale. For our analysis, we fixed $\sigma_f = 5, \sigma_y = 1$, and varied the length scale. Results are shown in Figure 3 and 4.

Further Work

Possible extensions to be explored include:

- Fitting of GP parameters: Currently, K-nearest neighbors with uniform weighting still performs better than both kriging examples. We could consider a more intelligent selection of the model parameters to improve performance.
- Seasonal data: Because our data also includes the per-month collision breakdown over the given time period, exploration into the temporal trends could help make our prediction model more accurate.
- Different boroughs: Although we focused mainly on Manhattan in our analysis, we could easily adapt our model for another borough, tuning parameters as needed.

References

[1] Aaron, B. NYPD's Lax Crash Investigations May Violate State Law. In *Streetsblog.org*, 15 Feb 2013.

[2] Diggle, Peter and Ribeiro, Paulo Justiniano, *Model-based Geostatistics*, Springer, 2007.

[3] Krauss, J, NYC Crash Data Band-Aid, `http://nypd.openscrape.com/`.

[4] Murphy, Kevin P, *Machine Learning: A Probabilistic Perspective*, MIT Press, 2012.

[5] Rasmussen, Carl and Williams, C. K. I. *Gaussian Processes for Machine Learning*, MIT Press, 2006.