Predicting Intersection-Specific Accidents Using New York City Crash Data

Irene Chen
Applied Math
Harvard University
Cambridge, MA 02138
iychen@college.harvard.edu

Brian Zhang
Physics
Harvard University
Cambridge, MA 02138
brianzhang@college.harvard.edu

Abstract

In the context of crash data, probabilistic models allow city officials to better understand crash trends and implement preventative actions. In this paper, we examine a dataset of New York City crash reports over the past three years labeled by intersection. Given an unseen intersection, we seek to predict the number of collisions in a given time period, evaluating our results using cross-validation assuming a Poisson likelihood. Building from a simple baseline of predicting the average, we tested a K-nearest neighbor (KNN) regressor with uniform and inverse distance weighting and the Gaussian process method of kriging. Our prediction results ranked KNN with uniform weighting first on both the Manhattan and Brooklyn datasets. However, our Gaussian process model is more readily interpretable and gives information about the distribution of crashes through both a visual representation and optimal hyperparameter values.

1 Introduction

Probabilistic models are fundamental tools for machine learning, providing a coherent framework for finding structure in real-world data. In the context of crash data, probabilistic models allow city officials to better understand crash trends and to implement preventive actions. Researchers in London have explored spatial dependence and uncorrelated hetereogeniety [1]. Environmental engineerings have also examined spatial distributions of fatal and injury crashes in Pennsylvania [2].

Although analysis has been done at the county level, there has been a lack of crash data analysis performed at the road level. As Aguero-Valverde and Jovanis note,

> Given the existence of spatial correlation, at least for injury models, it is expected that spatial correlation plays a more important role at smaller spatial scales. FB [full-Bayes] models with spatial correlation may be even more useful at road section level where units are smaller and closer and therefore, the probability of spatial correlation is higher. Using FB models with spatial correlation is a natural next step in research work.[2]

We propose to conduct similar crash data analysis on an intersection-specific basis, using crash data from the New York City Police Department. Historically, the New York City Police have been resistant to releasing New York City crash data. As one senior attorney from the New York Polic Department explained, they are reluctant to release the data since it could be manipulated by people who want "to make a point of some sort." [3] Fortunately an open source project has recently scraped the deliberately obfuscated New York Police PDF files and has made the data public. [4]

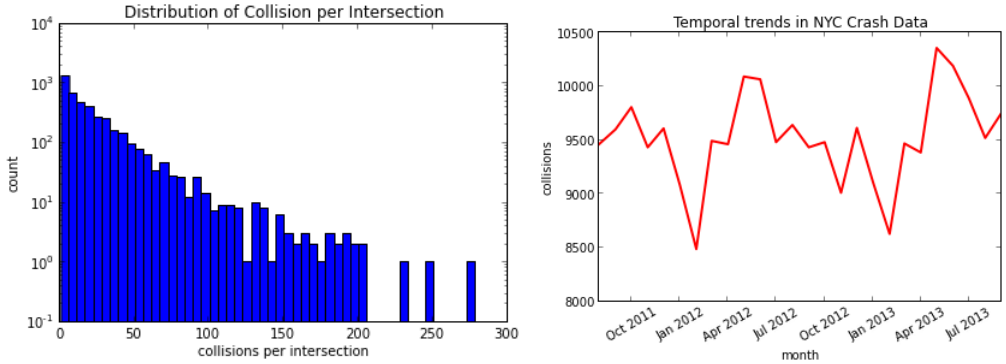


Figure 1: (Left) Histogram of collisions per intersection on a log-scale. (Right) Temporal changes in New York City crash data over the 25-month period from August 2011 to September 2013.

Although existing analysis on this dataset has been limited to visualization projects ([5]), we aim to build a probabilistic model over the dataset that is capable of predicting the collision count at unseen intersections. A predictive model would allow city officials to infer traffic trends across the city at by only measuring data from a few intersections. Specifically, we use Gaussian processes, also known as kriging in the spatial statistics literature, to model spatial trends in collisions. The model fit of a Gaussian process can be evaluated using the marginal likelihood, and its predictive performance can be evaluated using cross-validation on a test set. By comparing with other techniques for spatial prediction such as K-nearest neighbors, we can determine the best predictive technique for this type of data.

The body of our paper is organized as follows. In the next section, we will describe the specifics of our dataset and explain our exploratory data analysis. In Section 3, we describe our metrics and baselines upon which the rest of our analysis will rest, specifically our assumption of a Poisson likelihood. In Section 4, we describe our K-nearest neighbors model in detail and present our findings. One limitation of K-nearest neighbors is that the predictions are not inherently smooth or interpretable. Hence, in Section 5, we build a probablistic predictive model using Gausian processes to model spatial trends in collisions. Finally, in Section 6, we present our model performances on two borough datasets and discuss potential explanations. We conclude in Section 7 with additional research directions for New York City crash data analysis.

2 Exploratory Data Analysis

We focused our analysis on the dataset of collisions scraped from the PDFs from the New York City Police Department. The dataset itself covers the five boroughs of New York City and consists of monthly collision counts at intersections, spanning August 2011 to September 2013 and listing 249,462 monthly totals.

The dataset itself consists of 39,040 unique intersections and 364,330 total collisions. The distribution of the collisions per intersection, as seen in Figure 1, is skewed heavily towards smaller values with an overall mean of 9.32 and standard deviation of 15.41. In addition to longitude, latitude, month, borough, and the number of collisions, each datum also includes additional features such as the two streets which label the intersection, the number of people injured and killed, and the types of vehicles involved. For our model, we examined only longitude, latitude, month, borough, and number of collisions.

Note that the data seems to exhibit some seasonal changes—as shown in Figure 1. However, we chose to aggregate the data over the 25-month period in order to obtain larger collision counts, which should decrease the relative noise at each intersection.

Each borough has a slightly different distribution, as seen in Figure 2. We treated each borough as an individual dataset since geographic and political factors may influence the spatial distribution of collisions in each borough. Within our season-aggregated data, we narrowed our problem to

Borough	Intersections
Manhattan	4,191
The Bronx	5,701
Brooklyn	10,432
Queens	14,247
Staten Island	4,511

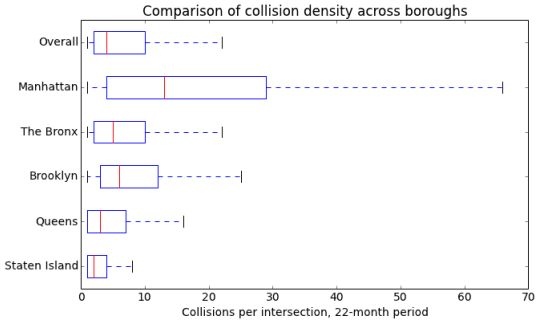
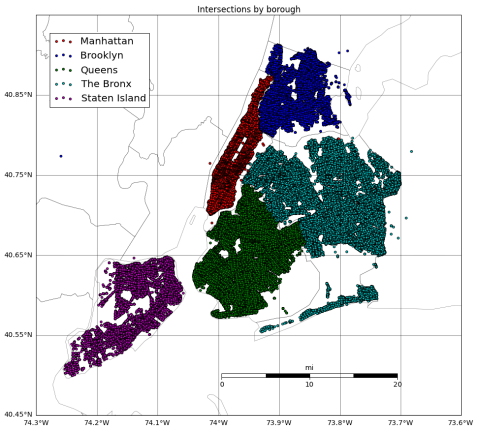


Figure 2: (Left) Count of distinct intersections for each of the five boroughs of New York City. (Right) Box-and-whiskers plot of collisions per intersection over a 25-month period.



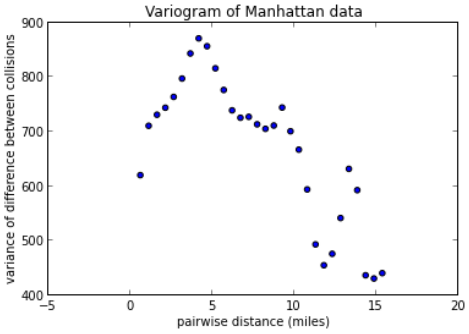


Figure 3: (Left) Scatterplot of intersections in dataset with boroughs as labeled by New York City Police Department. (Right) Variogram of Manhattan data relating pairwise distances and variance of the difference of collisions.

Manhattan and Brooklyn because the two boroughs have the highest collisions per intersection. Manhattan is more densely populated than Brooklyn, while Brooklyn has more intersections overall, which may yield interesting results.

The dataset itself contains a few biases. First, intersections are only included for a particular month if they had reported collisions. Therefore, the minimum number of collision counts is 1 rather than 0. By visual inspection of the data (Figure 3), we see that there are some points that fall far outside of the five boroughs. Moreover, a few intersections appear to be mislabeled as the incorrect borough. Although we were able to spot these outliers visually, we allowed them to remain in the dataset because they were relatively few in number.

Our proposed predictive models rely on the underlying intuition that close intersections have close collision count numbers. To test this relationship, we plotted the empirical variogram for Manhattan data, as modeled after [8] (Figure 3). For all pairs of intersections i, j, we computed the variance of difference between the collision counts $0.5(x_i - x_j)^2$ and plotted against pairwise distance d_{ij} and displayed a binned average over distance. As we hoped, we see that at small distances, the empirical variance increases as the distance between intersections increases. For larger distances, we see the reverse effect, potentially explained by the fact that Manhattan is a long island with low collision counts at either extreme.

By focusing on the two boroughs with the highest rate of collisions (Manhattan and Brooklyn), we will leverage this relationship between pairwise distance and similarity of collision counts to build a spatial predictive model.

3 Metrics and Baselines

Our ultimate goal was to develop the best predictive techniques for city crash data. To compare across different methods, we partitioned our datasets for the two boroughs into training and test sets in a 3 to 1 size ratio. We used the training set to fit the parameters of the model, then evaluated on the test set.

To measure performance, we calculated the negative log likelihood assuming a Poisson distribution of collisions per intersection ([6]). If at intersection i, the actual count of collisions was x_i and we predicted λ_i, then our average negative log probability per data point would be

$$NLL/N = -\frac{1}{N} \cdot \sum_{i=1}^{N} \log\left(\frac{e^{-\lambda_i}\lambda_i^{x_i}}{x_i!}\right).$$

By taking a derivative, one can show that this quantity is minimized by setting $\lambda_i = x_i$ for each i. This is the value obtained by perfect guessing. Since this cannot be achieved in practice, our simplest baseline is to guess the average of the training data, $\lambda = (\sum_i x_i)/n$, for each of the test points. We call this the average guessing baseline. Both of these values are plotted in Figure 7 in comparison with our performances from K-nearest neighbors and kriging.

4 K-nearest Neighbors

K-nearest neighbors is a non-parametric method for regression. In the context of crash data analysis, we use K-nearest neighbors to predict the count of collisions for an unseen intersection. For each new data point, we computed the Euclidean distance from the new data point to every 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. Uniform average weighting simply averages the closest K points while inverse distance weighted average weights each point by $1/d$ where d is the distance between the two intersections.

As before, we calculated the negative log likelihood per data point, and we can see the K-nearest neighbors performance for varying number of neighbors K in Figure 4 for both Manhattan and Brooklyn. Our performance was found by performing 10-fold cross validation on our training data.

We see that for both Manhattan and Brooklyn, the uniform K-nearest neighbors model performed better than the inverse distance model. Furthermore, Brooklyn has a lower negative log likelihood than Manhattan across the board, potentially because Manhattan has a higher variance in collisions per intersection and a higher mean (see Figure 2).

We found that our model selected the optimal number of K neighbors to be 15 for Manhatan and 30 for Brooklyn, again supporting the claim that there is more variance in Manhattan than Brooklyn. Using these tuned parameters, our calculated negative log likelihood per data point on our held-out test data was then 10.36 and 6.96 for Manhattan and Brooklyn, respectively. These are plotted in Figure 7.

5 Kriging

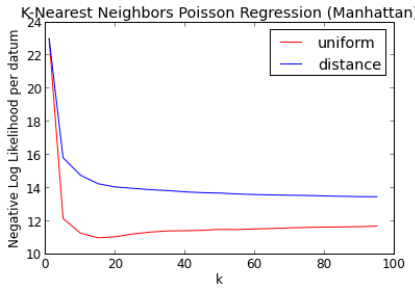
Kriging is the spatial statistics term for modeling spatial behavior using a Gaussian process. Following the example in Murphy 15.3.3 [7], we fit a model to log-count data because we assume collisions at each intersection are Poisson distributed. We also decided on the Matern kernel with $\nu = 3/2$:

$$\text{Matern}_{3/2}(r, l) = \left(1 + \frac{\sqrt{3}r}{l}\right)\exp\left(-\frac{\sqrt{3}r}{l}\right)$$

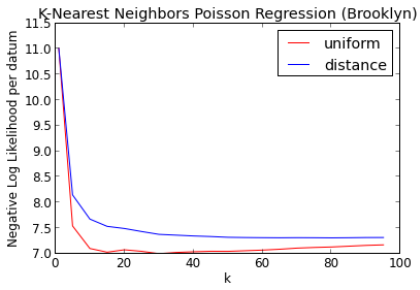
This kernel leads to GPs which are 1-time mean-square differentiable and is frequently used in the kriging literature [8]. The covariance between the values at two points $\mathbf{x}_p, \mathbf{x}_q$ is then

$$k_y(\mathbf{x}_p, \mathbf{x}_q) = \sigma_f^2 \text{Matern}_{3/2}(||\mathbf{x}_p - \mathbf{x}_q||^2, l) + \sigma_n^2 \delta_{pq}$$

where our hyperparameters are the length-scale l, signal variance σ_f^2, and noise variance σ_n^2. Although our locations were given in terms of latitude and longitude coordinates, we wrote a simple function to compute the relative distance in miles.



(a) Manhattan



(b) Brooklyn

Figure 4: K-nearest neighbors results for uniform and inverse distance weighting on data from the two New York City boroughs with the most collisions per intersection.

Given a choice of hyperparameters, prediction using Gaussian processes is done by computing kernel matrices for the test and train data and then performing a few matrix operations (see Rasmussen Chapter 2 [9]). To set the hyperparameters, we used gradient descent to optimize the log marginal likelihood,

$$\log p(\mathbf{y}|X) = -\frac{1}{2}\mathbf{y}^\top K^{-1}\mathbf{y} - \frac{1}{2}\log|K| - \frac{n}{2}\log 2\pi,$$

where K is the kernel matrix of the training data, and depends on l, σ_f^2, and σ_n^2. The gradient of log marginal likelihood is given by

$$\frac{\partial}{\partial\theta_j}\log p(\mathbf{y}|X,\theta) = \frac{1}{2}\text{tr}\left((\alpha\alpha^\top - K^{-1})\frac{\partial K}{\partial\theta_j}\right)$$

where $\alpha = K^{-1}\mathbf{y}$.

We first applied kriging to Manhattan, with results shown in Figure 5. The optimal hyperparameter values found using gradient descent were $(L, \sigma_f^2, \sigma_n^2) = (1.95, 8.66, 1.34)$, and the resulting mesh grid prediction is shown at the bottom right panel. We also plot similar grids for Gaussian models at two other length scales. Note that when L is small (0.1 or 0.5 miles), the Gaussian model can represent fine structure but has a lower marginal likelihood. By contrast, $L = 1.95$ miles seems to give a smoother variation over Manhattan which still represents some trends of interest. In all three Gaussian model plots, we can observe a peak in collisions near New York Penn Station as well as a lull in activity in the middle of Central Park. Using the optimal hyperparameter values (lower right), we obtained a test set prediction performance of 13.02.

For Brooklyn, we used hyperparameter values of $(L, \sigma_f^2, \sigma_n^2) = (5.5, 8, 1.06)$. Because the dataset was larger, we ran into computation time limits in trying to optimize the marginal likelihood of the whole dataset. Hence, these were the optimal values obtained through optimizing the marginal likelihood of the first 1,000 data points (out of 10,432). The prediction grid is shown in Figure 6, and the performance on the test set was 8.36.

6 Discussion

The results from kriging on the Manhattan and Brooklyn data sets are compared with the average baseline and K-nearest neighbors in Figure 7. Contrary to our expectations, the optimal KNN regressor performed better than kriging in both regions. Kriging still outperformed the average baseline in both regions, and outperformed KNN with an inverse distance weighting for the Manhattan dataset.

We hypothesize reasons for the superior performance of uniform KNN over inverse distance weighting KNN and kriging. Because of the high density of intersections in city data, a good predictor should average over many nearby intersections and not just look in the immediate vicinity. In inverse distance weighting, the closest intersections dominate the KNN average, which could be a reason why uniform KNN performs better. We hypothesize that kriging would perform better when the data is more sparse, so it has a chance to learn more interesting interpolations between data

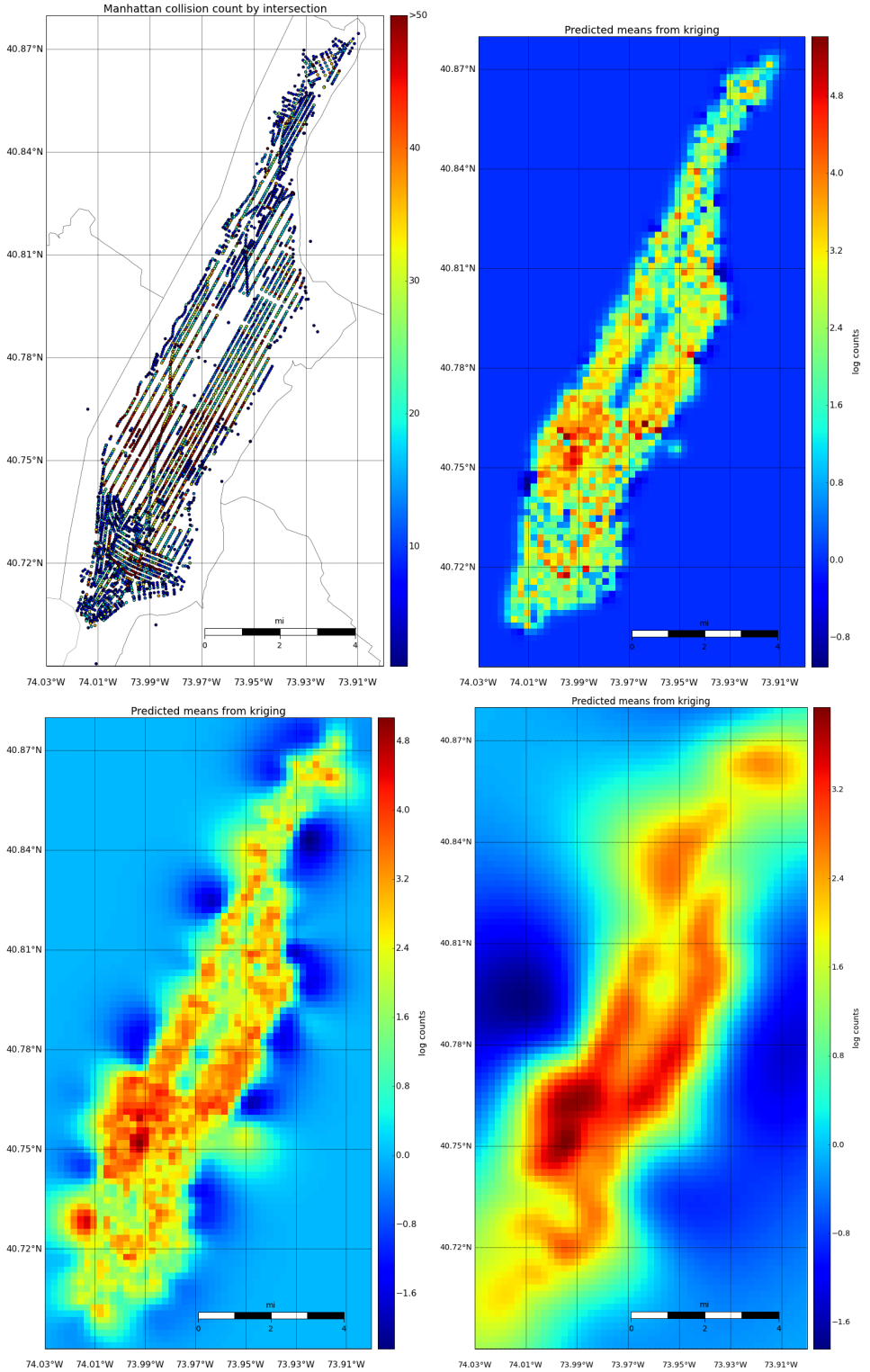


Figure 5: Scatter plot of Manhattan intersection data, along with kriging predictions over a mesh grid for different parameter values: $(L, \sigma_f^2, \sigma_n^2) = (0.1, 25, 1), (0.5, 25, 1)$, and $(1.95, 8.66, 1.34)$ as we go from top right to bottom left to bottom right. Log marginal likelihoods are -7028, -5390, and -5052. L has units of miles and σ_f and σ_n are fit to log collision counts. The bottom right panel optimizes the log marginal likelihood, with parameters found by gradient descent.

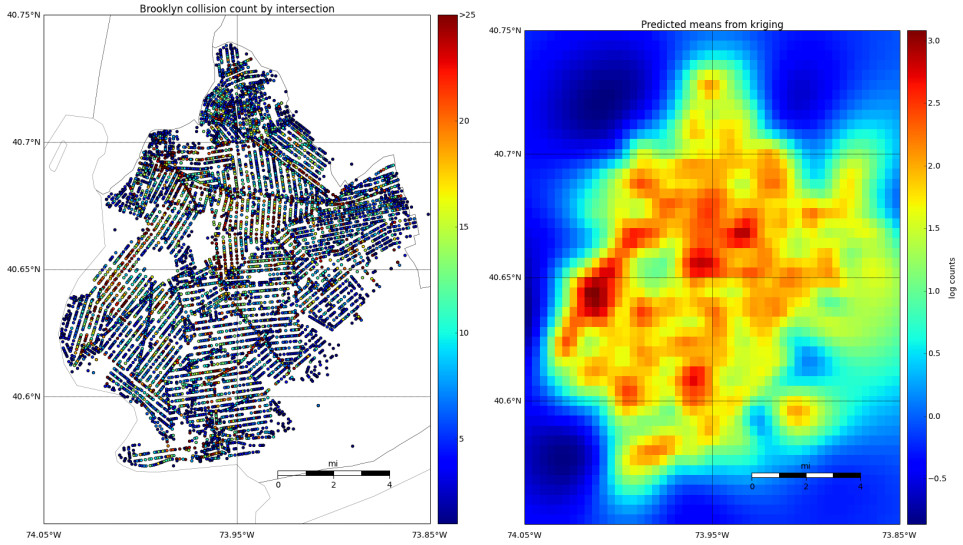


Figure 6: Scatter plot of Brooklyn intersection data, along with kriging predictions over a mesh grid for parameters $(L, \sigma_f^2, \sigma_n^2) = (5.5, 8, 1.06)$. Because this dataset was larger, the parameter values were found by optimizing the marginal likelihood of only a subset of the training data.

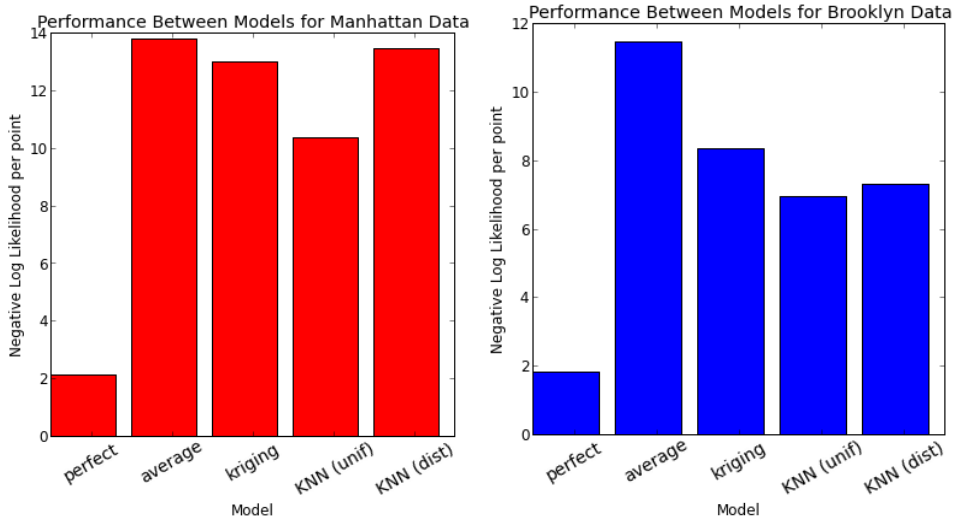


Figure 7: Model Performance on Manhattan (left) and Brooklyn (right) crash data. Other than the unrealistic perfect guessing baseline, K-nearest neighbors using a uniform weighting consistently performed the best, with kriging performing the second best. Note that performance values were found using tuned parameters on withheld test data.

points. Perhaps if we did a similar analysis in the suburbs, kriging would do a better job than KNN. Kriging could also be hurt by some of the spatial outliers in our data, which KNN would be more likely to get correct since it looks at the average of data points.

Since uniform KNN performed better than kriging on prediction, does that mean kriging is a useless model for city data? We do not think so, and offer a few reasons why. As a probabilistic model, kriging gives us access to the language of Bayesian statistics, which a non-parametric KNN regressor would not. For instance, the marginal likelihood is well-understood to represent the fit of a model. Gaussian models can offer additionally functionality which a KNN regressor would not. For instance, kriging can not only predict means at certain locations, but can also output variances in those predictions and infer correlations between counts at different locations.

The predicted means from kriging vary smoothly across the region, as opposed to a KNN predictor which jumps discontinuously when points get swapped in and out of the nearest neighbor set. This yields a nice visualization of the data which could be useful for city officials or other parties interested in intuitively understanding the spatial variation of collisions. For instance, we see in Figures 5 and 6 that a Gaussian model helps to identify hot spots in traffic collisions for both Manhattan and Brooklyn.

The optimal hyperparameters from kriging can yield information about the data as well. For Manhattan, these values were $(L, \sigma_f^2, \sigma_n^2) = (1.95, 8.66, 1.34)$ and for Brooklyn, they were $(5.5, 8, 1.06)$. We see that the optimal values of σ_f^2 and σ_n^2 were in a similar ballpark. However, the optimal L was 1.95 miles for Manhattan and 5.5 miles for Brooklyn. It appears that as the area of our region grows, a Gaussian model fits better if smoothed over a larger length scale. We can also see a slight drop in σ_f^2 as we go from Manhattan to Brooklyn, possibly because Brooklyn has fewer collisions per intersection on average. These values from kriging are more intuitive than interpreting the value of K in a KNN regressor, which can be more difficult to decipher.

7 Conclusion

The introduction of probabilistic models yields new possibilities for crash data analysis. To explore these possibilities, we formalized and implemented Gaussian processes to model spatial trends in collisions. By comparing our kriging results with other spatial predictive techniques, we found that the K-nearest neighbors with uniform distance weighting outperformed our kriging model.

To build on our current prediction methods, we conclude with other possible applications for crash data analysis in New York City. One natural extension of our analysis would be to incorporate seasonal trends in our prediction. As noted by Figure 1, there seem to be strong seasonal trends in our data with lows in February and highs in May over the roughly two year span. Although this variance may partially be explained by having a fewer number of days in February, more research can still be done to extend our prediction methods to foresee number of collisions per intersection per month.

Another extension would be to examine the other three boroughs that we excluded from our analysis: Staten Island, The Bronx, and Queens. Just as we interpreted the parameters and hyperparameters for Manhattan and Brooklyn, deeper analysis on the remaining three boroughs might provide more insights into why our KNN model with uniform weighting outperformed kriging in the two boroughs analyzed.

One last extension to which our crash data analysis predictive model can be applied arises when we consider the features of our dataset that we excluded in Section 2. The the New York City Police Department collision dataset includes not only location and monthly collision totals, but also the number of people involved, the vehicles involved, the number of people killed, and the number of people injured. Crash data analysis with these additional features might provide additional insight into the spatial crash distribution in New York City.

References

[1] QUDDUS, MOHAMMED A., 2008. Modelling area-wide count outcomes with spatial correlation and heterogeneity: An analysis of London crash data. In *Accidental Analysis and Prevention*,

Vol. 40, 1486-1497.

[2] J. AGUERO-VALVERDE, P.P. JOVANIS, Spatial analysis of fatal and injury crashes in Pennsylvania. In *Accident Analysis and Prevention*, Vol. 38, 2006, 618-625.

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

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

[5] KRAUSS, J. NYC Crash Mapper. `http://nyc.crashmapper.com/`.

[6] L. FRISTRM, J. IFVER, S. INGEBRIGTSEN, R. KULMALA, L.K. THOMSEN, Measuring the contribution of randomness, exposure, weather, and daylight to the variation in road accident counts. In *Accidental Analysis and Prevention*, Vol. 27, 1995, 1-20.

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

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

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