UNIVERSITY OF MICHIGAN

STATS406 Computational Methods in Statistics and Data Science

Exploring the Relationship between Green Coverage and Crime Rate in American Cities

Project Group 8
Xixiao Pan
Ziyang Xiong
Jiawen Liu

December 13, 2023

Contents

1 Introduction

Urbanization, marked by the relentless expansion of cities, has become one of the defining global trends of the 21st century. Within this urban landscape, urban green space (UGS), loosely defined as any type of public or private plant-cover environment within each city[1], has risen to prominence as a pivotal element. In recent years, urban green space has transcended its role as mere landscaping and has emerged as a symbol of sustainable urban planning and design. Many people hold the opinion that it offers a host of benefits, from providing sanctuary to city dwellers seeking a breath of fresh air to contributing to the overall well-being of urban communities. And since the ability of urban green space to improve mental health and strengthen social cohesion [2] have already been approved. More and more people are curious about whether UGS can serve as a deterrent to crime rate.

Unfortunately, although the relationship between UGS and crime rate has been a subject of both intrigue and debate among academics, urban planners, the studying results on whether UGS can reduce crime risk are often mixed. There are competing theories stating that specific types of UGS can potentially aggravate the crime which has become a main concern of many communities and police departments[3].

Is this concern justifiable? In this paper, our objective is to assess whether a correlation exists between urban green space and crime rates. To ensure a nuanced exploration, we introduced an additional variable, the income level of different cities, to mitigate potential confounding effects driven by economic factors. To the end, we applied two independent variables related to green coverage and income, paired with one dependent variable tied to crime rates. The analysis is carried out through the use of permutation tests and Markov Chain Monte Carlo (MCMC) under both a uniform null hypothesis and a uniform alternative hypothesis. And we used the non-parametric technique, local regression, looking for the relationship between variables.

This research is of paramount importance as it has the potential to provide valuable insights for urban planners and policymakers. By figuring out whether the impact of parks and green spaces on crime rates exist or not, we can better allocate public resources, with a specific emphasis on enhancing community safety and improving the well-being of urban residents. And this study holds particular significance for underserved communities in the United States, where the benefits of green spaces may be transformative.

The following report contains five main sections: Data, Method, Simulation, Analysis, and Discussion. In the data section, we'll introduce where we get the data and how we pre-process it at first. Followed by a brief introduction to permutation tests and Markov Chain Monte Carlo and Metropolis-Hastings Algorithms and local regression. And the reasons and advantages of choosing these methods will also be explained in the Method section. In the following section, Simulation, we'll compare the effectiveness of the permutation tests and correlation test under different situations and try to choose an appropriate prior distribution for the MCMC method. And in the Analysis and Discussion section, we'll apply real data we gathered to the methods we mentioned before and discuss how these

results achieve our experimental goals, and what the room for improvement is.

2 Data

To find the relationship between UGS and the crime rate, we focus on the green afforestation and crime rate in different cities in the USA. Also, the income of different cities is added to reduce possible effects on crime rate. We found three datasets to support our intention. The first dataset is the ParkServe dataset from Trust for Public Land. To represent the green degree of a city, we extracted two labels of data from the raw dataset——Acreage of Parks per 1000 residents by city and Parkland as Percent of Adjusted City Area. We found the green park data of 100 different cities from 2013 to 2020. We used RStudio to help us extract and splice the data.

The second dataset is the crime rate of different cities. We found the data from Federal Bureau of Investigation Crime Data Explorer. However, the website only provides tables of crime numbers of one city during 2013 to 2020. Therefore, we randomly selected 24 cities from the 100 cities in green park dataset, where 8 cities are highly-populated cities, 8 ones are medium and 8 ones are low according to the classification on ParkServe website. Then, we manually downloaded the crime numbers reported by these 24 city's police departments from 2013 to 2020 and the population from the ParkServe website. And use RStudio to calculate crime numbers/population to obtain the crime rate. Finally, we get the crime rate of 24 cities in the USA from 2013 to 2020.

The third dataset is the income of different cities from 2013 to 2020. The dataset is from GDP by Industry — U.S. Bureau of Economic Analysis (BEA).

In this paper, we denote the crime rate as CrimeRate, the Acreage of Parks per 1000 residents as AcrePerPeople, the Parkland as Percent of Adjusted City Area as ParksPercent and the income as Income.

To investigate the distribution of different variables, we found they all follow non normal distributions. Here, we displayed two plots for simplicity. The QQ-plots of normal distribution for Crime Rate and Acres Per People in Figure 1, and Figure 2 show that they deviate much from the normal distribution.

3 Method

3.1 Permutation Test

To explore whether the relationship between urban green space (in this study, we use AcresPerPeople in data to reflect it) and crime rate exist or not, we use the permutation test first.

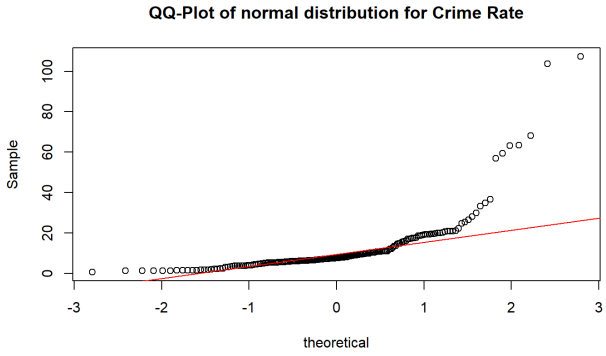

Figure 1: QQ-plots of normal distribution for Crime Rate, where the red line represents a normal distribution.

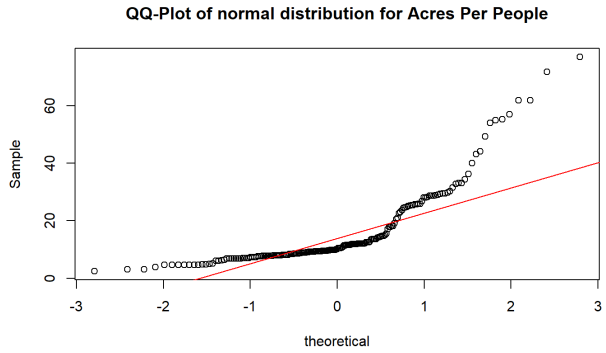

Figure 2: QQ-plots of normal distribution for Acres Per People, where the red line represents a normal distribution.

Null hypothesis (H0): There is no relationship between crime rate and AcresPerPeople.

Alternative hypothesis (H1): There is a negative correlation between the AcresPerPeople and crime rate.

Compared to traditional parametric tests, such as T-test, permutation tests are a non-parametric method that doesn't require the assumptions of normal distribution and homogeneity of variance. According to the visualization of the data in the last section, it's obvious that the distribution of the crime rate and AcresPerPeople did not conform well to a normal distribution. Moreover, considering the small size of our sample, it is highly susceptible to outliers. The permutation test, known for its insensitivity to outliers and sample heterogeneity, often yields more robust results. Furthermore, the outcomes of the permutation test are generally straightforward to interpret and understand, making them more applicable to practical problem analysis, which underpins our choice of this method for the study.

In our study, we conducted a permutation test by randomly shuffling the independent variable AcresPerPeople while holding the dependent variable CrimeRate steady. This approach disrupts any existing link between the variables under the null hypothesis of no association. For each of the 1000 permutations, we recalculated the correlation coefficient between the shuffled AcresPerPeople and the fixed 'crime rate'. The p-value was then

calculated as the proportion of permutations where the recalculated statistic was as extreme or more than the original. A low p-value would strongly indicate a rejection of the null hypothesis, suggesting there may exist a relationship between two variables.

To further explore the relationship existing or not, we employed bootstrap methods to construct 95% confidence intervals of the correlation coefficient and p-value. If the range of confidence intervals of correlation coefficients is always under 0. It may imply that if a relationship exists, two variables should be negatively correlated. And if the confidence intervals of p-value are always less than 0.05, we may have enough evidence to reject the null hypothesis. And if confidence intervals can be significantly small but varied a lot, this may suggest that our conclusion of rejecting H0 in the results of the permutation test is less accurate, because the randomness of p-value is so strong that we just happen to be able to reject H0.

3.2 MCMC

Except for the permutation test, we used another method, Markov Chain Monte Carlo and Metropolis-Hastings algorithms.

Null hypothesis (H0): There is no relationship between crime rate and AcresPerPeople.

Alternative hypothesis (H1): There is a negative correlation between the AcresPerPeople and crime rate.

First, we use Bayesian Analysis to specify a statistical model describing the relationship between crimerate and AcresPerPeople. We construct a linear model

$$y = \beta_0 + \beta_1 x + \sigma \tag{1}$$

where y is CrimeRate, x is AcresPerPeople, σ is the error term, β_0 and β_1 are the estimated parameters. The prior distributions of β_0, β_1, and σ are unknown , which are explored in Simulation part. The log-likelihood is applied for numerical stability and computational efficiency. By applying Bayes' theorem, the posterior distributions of the parameters can be obtained, which updates the beliefs about the parameters. Then, we apply Metropolis-Hastings as a MCMC method to generate a sequence of random samples. We set the initial parameters to be 0, 0, and 0 because there is no particular knowledge of the starting values. The initial posterior probability is calculated. The proposal distribution is chosen with a normal distribution with mean 0 and standard deviation 0.1 to achieve numerical stability. Then, the new parameter values are proposed and accepted or rejected based on the posterior probabilities. The process is iterated for 10000 times and the samples converge to the true posterior distribution. To reduce the effects of initialization effects, we set the burn-in period to be 3000 to discard the first 3000 results. Finally, we analyzed the 7000 posterior probabilities by observing the 95% confidence interval, the mean value and the mean value of the distribution that β_1 is less than 0, which means the possibility to reject H0.

Furthermore, in case of randomness, we used Monte Carlo simulation to increase the accuracy of the results. We repeat the previous MCMC process for 100 times to obtain the 95% confidence interval that β_1 does not include 0 and the mean value of all means of β_1 values. It reduces randomness and provides more reliable results.

3.3 Local Regression (LOWESS)

In the context of exploring the relationship between urban greenery levels and crime rates in American cities, we employed local regression, also known as Locally Weighted Scatterplot Smoothing, as the primary method for data analysis. This approach is chosen for its ability to flexibly model potential nonlinear relationships without the need for presetting a parametric form. We will detail its computational aspects and tie it back to relevant statistical theory in this part.

Local regression is a non-parametric regression technique used for modeling the relationship between variables. Unlike parametric methods, local regression does not assume a predetermined form, such as linear or polynomial, for the relationship, making it particularly suited for exploring complex and potentially nonlinear relationships between variables. Mathematically, the model at each point Xi is given by:

$$\hat{y_i} = \beta_0 + \beta_1(x_i - x) \tag{2}$$

where $\hat{y_i}$ is the predicted value for the dependent variable at point x_i , and 0 , 1 are coefficients estimated by minimizing the weighted least squares:

$$\sum_{j=1}^{n} w_j(x_i)(y_j - \beta_0 - \beta_1(x_j - x))^2 \tag{3}$$

Here, $w_j(x_j)$ is the weight assigned to each point x_j, typically determined by a kernel function that assigns higher weights to points closer to x_i. The choice of kernel function is critical. Commonly used kernels include the Gaussian and Tricubic kernels. For instance, the tricube weight function for a point x_j relative to x_i is:

$$w_i(x_i) = (1 - (\frac{|x_j - x_i|}{h})^3)^3 \tag{4}$$

, where h is the bandwidth.

This function provides weights that decrease as points are farther from xi, up to a distance of h, beyond which the weight is zero.

In LOWESS, the focus is placed on each observation in the dataset. Around each point, a neighborhood is considered, within which a simple model (typically linear) is fitted using weighted least squares. The prediction for each point is based on nearby points within this window, with weights determined by their proximity to the center point.

In our study, local regression is utilized to explore the relationship between greenery levels (as measured by per capita green space and park percentage of city area) and crime rates. This method allows us to investigate whether there is a trend or pattern between two variables, without the concern of fitting these relationships into a predefined mathematical form.

There are three primary assumptions for local regression.

1. Bandwidth (Smoothing Parameter): too small bandwidth might lead to overfitting, while too large can obscure important local features.

2. Data Density: Assumes a sufficient number of data points around each observation to effectively perform local fitting.

3. Weighting Scheme: Weights are typically assigned using kernels such as tricubic or Gaussian.

Connecting to statistical theory, the bandwidth selection illustrates the bias-variance trade-off. A smaller bandwidth can capture more local variation at the risk of leading to high variance and overfitting. Conversely, a larger bandwidth provides a smoother estimate but may introduce bias by oversmoothing. This non-parametric method contrasts to parametric methods, where the functional form of the relationship between variables must be specified a prior.

4 Simulations

4.1 Permutation Test

Based on the QQ-plot analysis in the data section, it is evident that the data used in this study does not follow a normal distribution. In the following, we aim to demonstrate that correlation tests perform well when data follows a normal distribution, while permutation tests are more efficient in situations where the data does not conform to a normal distribution.

To investigate this, we conducted two controlled trials. In the first trial, we used the Cholesky decomposition method to generate two groups of data from a normal distribution, each consisting of 100 data points, with their covariance close to 1. In the second trial, we randomly sampled 100 data points from an exponential distribution and added random small noise while maintaining a high correlation between these two sets of 100 data points.

The null hypothesis for both controlled trials states that the correlation between the two data samples is equal to 0. Given that we already know the underlying data relationships, our goal is to confidently reject this null hypothesis. In other words, the more significant the p-value obtained, the stronger the evidence supporting the effectiveness of the chosen method.

After applying both the permutation test and correlation test in these two trials, we obtained the following results in Figure 3 : In the first trial, both tests yielded p-values of zero, making it impossible to determine which method is superior, but we can find that they both work well. However, in the second trial, the permutation test produced much more significant p-values compared to the correlation test. Consequently, we can conclude that the permutation test is better suited for non-normally distributed data.

```
$cor_test_p_value
[1] 0

$perm_test_p_value
[1] 0

$cor_test_p_value
[1] 0.0001982076

$perm_test_p_value
[1] 0
```

Figure 3: R output of Permutation Test simulation results.

4.2 MCMC

To construct MCMC, Bayesian statistics is applied to construct a linear model. As the prior distributions of beta0, beta1 and sigma are unknown, the choice of the proper prior distribution greatly affects the results. Because there is little prior knowledge about the estimated parameters, we simply assume that they follow the normal distribution and the mean should be 0. However, for the standard deviation of the normal distribution, we simulate different values of standard deviation to observe the changes of the estimated β_1. Following the process in Method, we calculated the mean value of the distribution that β_1 is less than 0, which means the possibility to reject the null hypothesis to present the result. Then, we set the standard deviation to be 2, 4, ... and 20, and calculated the result as shown in Figure 4. We found that as the standard deviation increases, the probability of negative β_1 first increases and then floats in a small range. Because the parameter is flexible, we believe that the standard deviation should be sufficiently large to accommodate variability. Also, when the standard deviation is larger than 8, the probability tends to stabilize. We assume the stable y-value is close to the practical value. Thus, we select the turning point where $sd = 10$ for the following analysis.

4.3 LOWESS

To illustrate how local regression works under various conditions, including when its assumptions hold and when they do not, I perform several simulations. These simulations help to understand the operating characteristics of local regression, such as bias, mean

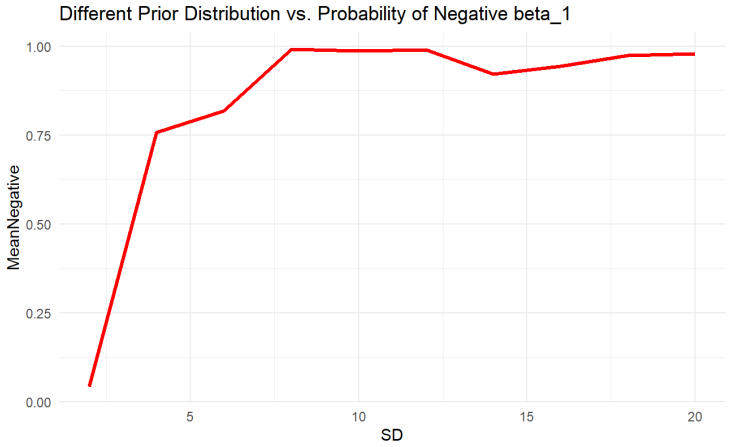

Figure 4: The relationship of different prior distribution and probability of negative β_1.

squared error (MSE), and prediction error. We will also compare the computational efficiency of different settings within the local regression framework.

We will now carry out the simulations and analyze the results for each scenario. In Scenario 1, we simulate data that follows a nonlinear relationship, which is ideal for local regression. In Scenario2, the data follow a linear relationship, testing how local regression performs when its nonlinearity assumption is violated. And Scenario 3 introduces outliers to see how robust local regression is against them.

Scenario	Bias	MSE	Computation Time (seconds)
Nonlinear Relationship	0.00089	0.03265	0.0001
Linear Relationship	0.00105	0.02956	0.0001
Nonlinear with Outliers	0.5674	3.494	0.00008

Figure 5: Simulation Results.

In Figure 5, the local regression approach demonstrated remarkable precision in the first scenario where we explore a non-linear relationship. The bias of mere 0.00089 indicates an exceptional alignment with the actual data, suggesting that when the method's assumptions are perfectly met, it captures complex, nonlinear patterns with minimal systematic error. This is further corroborated by the mean squared error (MSE) of 0.03265, which attests to the model's ability to produce predictions that closely mirror the true values. Furthermore, the computation time of just 0.0001 seconds highlights the method's computational efficiency, a critical factor in the era of big data and real-time analytics.

Transitioning to the second scenario, which involves a linear relationship, we find that local regression maintains a commendable level of accuracy. The slightly higher bias of 0.00105, while still low, subtly indicates the method's primary orientation towards nonlinear relationships. However, an MSE of 0.02956 is particularly noteworthy compared to the first scenario. It suggests that local regression, despite its non-parametric nature, is versatile enough to adapt to scenarios that deviate from its fundamental assumption of nonlinearity. This adaptability is a significant advantage in practical applications, where

the exact nature of data relationships may not always be clear. The consistent computation time in this scenario further reinforces the method's applicability across a diverse array of datasets.

The third scenario, which introduced outliers into a nonlinear relationship, presented a more challenging situation. The method exhibited a notable increase in bias to 0.5674, a clear indication of its sensitivity to outliers. This substantial deviation from the true values can be particularly problematic in real-world datasets, where outliers are not uncommon. The dramatic rise in MSE to 3.494 underscores reflects a considerable decrease in predictive accuracy due to the outliers. However, the computation time was slightly faster in this scenario, though this minor gain is overshadowed by the significant impact on bias and MSE.

These simulations paint a comprehensive picture of local regression as a highly effective tool for modeling and understanding complex relationships, especially those of a nonlinear nature. Its efficiency and adaptability, even in linear contexts, are commendable. However, its vulnerability to outliers necessitates caution, particularly in datasets where such anomalies are prevalent. This analysis suggests that while Local Regression is a powerful analytical tool, its application should be accompanied by careful data preprocessing, especially in the presence of outliers, to ensure the robustness and reliability of its predictions.

5 Analysis

5.1 Permutation Test

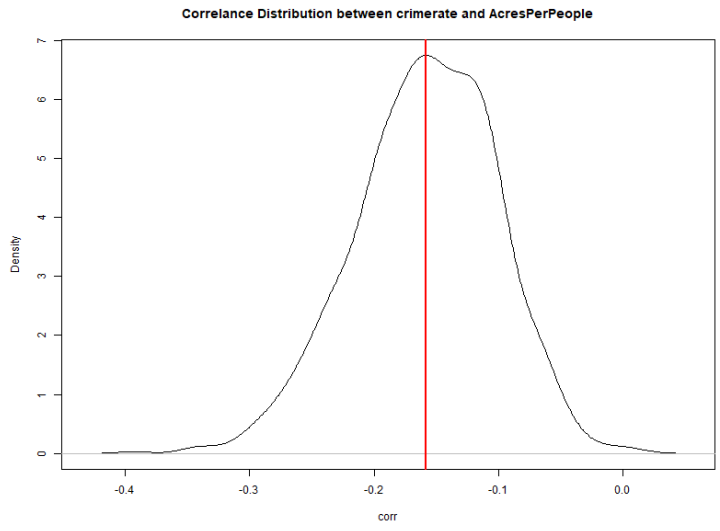

Figure 6: Correlation distribution between CrimeRate and AcresPerPeople.

To investigate the relationship between crime rate and green space in different cities, first we implemented the permutation test to calculate the correlation between CrimeRate and

AcresPerPeople, ParksPercent and Income relatively. After randomly permutating AcresPerPeople for 1000 times, the observed correlation and the p-value of three variables were obtained. For AcresPerPeople, the observed correlation is -0.153 and the p-value is 0.031; for ParksPercent, the observed correlation is -0.018 and the p-value is 0.804; for Income, the observed correlation is 0.076 and the p-value is 0.292. The observed correlation between CrimeRate and AcresPerPeople is the smallest, which means it is most negatively correlated with CrimeRate. The other variables only have the abstract correlation less than 0.1, which means the correlations are not significant. To further explore AcresPerPeople, bootstrap was used to repeatedly calculate the correlation for 1000 times. The 95% bootstrap confidence interval is (-0.2785, -0.0572), and the distribution of the correlation is shown in Figure 6. Additionally, the permutation test p_value of AcresPerPeople is less than 0.05, which provides evidence to reject the null hypothesis. From Figure 6, the correlation after bootstrapping distribution is stable.

Then, we implemented the bootstrap method to calculate the $p-value$ of permutation for 500 times. The permutation is also repeated 500 times. The bootstrap distributions of $p-values$ for AcresPerPeople and ParksPercent are shown in Figure 4 and 5 respectively. The 95% bootstrap confidence interval is (0.0000, 0.4168) and (0.1582, 0.9659) as shown in Figure 7 and Figure 8.

```
p-value of AcresPerPeople is:  0.031
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot.ci(boot.out = boot_results2, conf = 0.95, type = "perc")

Intervals :
Level     Percentile
95%   ( 0.0000,  0.4168 )
Calculations and Intervals on Original Scale
```

Figure 7: R output of bootstrap for AcresPerPeople.

```
p-value of AcresPerPeople is:  0.796
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot.ci(boot.out = boot_results3, conf = 0.95, type = "perc")

Intervals :
Level     Percentile
95%   ( 0.1041,  0.9840 )
Calculations and Intervals on Original Scale
```

Figure 8: R output of bootstrap for ParksPercent.

From Figure 9 and 10, the main distribution of p-value is less than 0.05 for AcresPerPeople and higher than 0.05 for ParksPercent. It provides evidence that AcresPerPeople is related to crimerate. Thus, we reject the null hypothesis at the 0.05 significance level for the variable AcresPerPeople.

5.2 MCMC

In the simulation part, we select the proper prior distribution for beta0, beta1, and sigma. Then, we followed the process mentioned above to calculate the mean value of the distri-

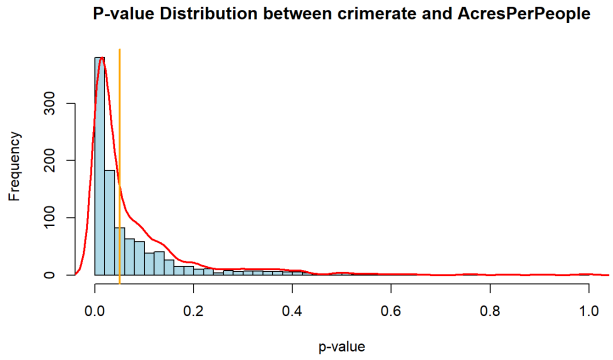

Figure 9: P-value Distribution between CrimeRate and AcresPerPeople

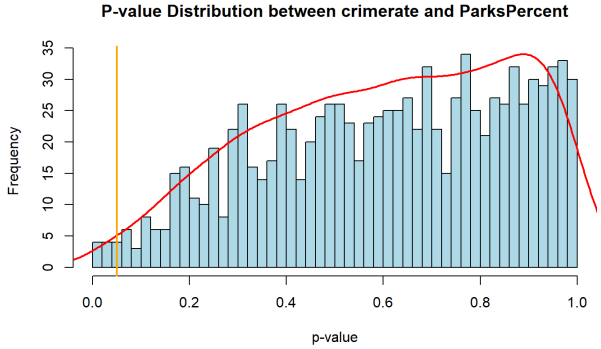

Figure 10: P-value Distribution between CrimeRate and ParksPercent

bution that β_1 is less than 0. To investigate in more aspects, we also calculated the 95% confidence interval of β_1 and the mean value of the β_1 distribution. When we set the seed to be 10, we get the following result in Figure 11. The frequency distribution of β_1 is generated in Figure 12. We observed that beta1 is less than 0 for 99.1% situations, and the 95% confidence interval does not include 0. It means that it is significant to reject the null hypothesis.

```{r}
beta1_CI1 <- quantile(beta1_posterior, probs = c(0.025, 0.975))
print(beta1_CI1)
mean(beta1_posterior)
mean(beta1_posterior < 0)
```

```
      2.5%      97.5%
-0.3147986 -0.0309686
[1] -0.1741222
[1] 0.9914286
```

Figure 11: Results of Probability of negative upper bound and mean $beta_1$

However, in case of randomness, we used Monte Carlo simulation to increase the accuracy of the results. After repeating the previous MCMC process for 100 times, we got the probability that the 95% confidence interval does not include 0 and the mean value of all means of β_1 values in Figure 8. However, according to the Figure 13, the probability is only 0.24,

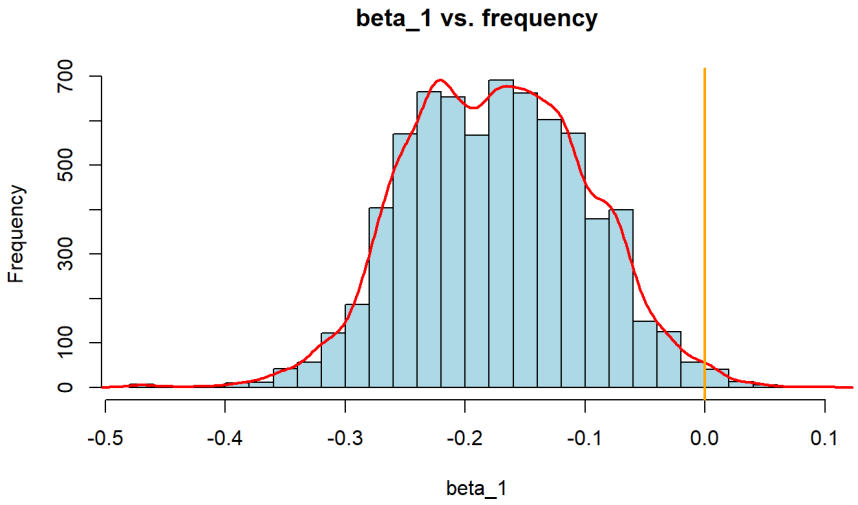

Figure 12: Frequency distribution of β_1

which means it is not significant to reject the null hypothesis. In the previous condition, we are able to prove that AcresPerPeople is relevant to CrimeRate due to randomness, which is not reliable. Nevertheless, the mean is still less than zero. It means that AcresPerPeople is only possible to be negatively correlated with CrimeRate. In conclusion, we cannot predict whether AcresPerPeople is linearly dependent on CrimeRate, but AcresPerPeople has another negative relationship with CrimeRate.

```
Probability of 97.5% CI upper bound < 0: 0.24
Mean of beta1_posterior means: -0.1163944
```

Figure 13: Results of MCMC after Monte Carlo simulation

5.3 LOWESS

We conducted an analysis on two indicators of greenness to see if they were significantly associated with crime rates. The left graph illustrates how crime rates change as green space per capita changes. It was observed that as the per capita green space area increases, the crime rate shows a fluctuating trend. However, there is no obvious linear relationship. In Figure 14, the right graph shows the trend of crime rate as a function of park area percentage. Likewise, crime rates show some variability but still no clear linear pattern.

Although local regressions reveal certain trends, these results indicate that there is no direct linear relationship between urban greenness (measured as green area per capita or percentage of park area) and crime rates. This suggests that other factors may influence crime rates, or that the relationship between greenness levels and crime rates may be more complex than a simple linear association. Further research may be needed to consider additional variables and potential interactions.

Next, we employed a multivariate local regression model for a more comprehensive analysis.

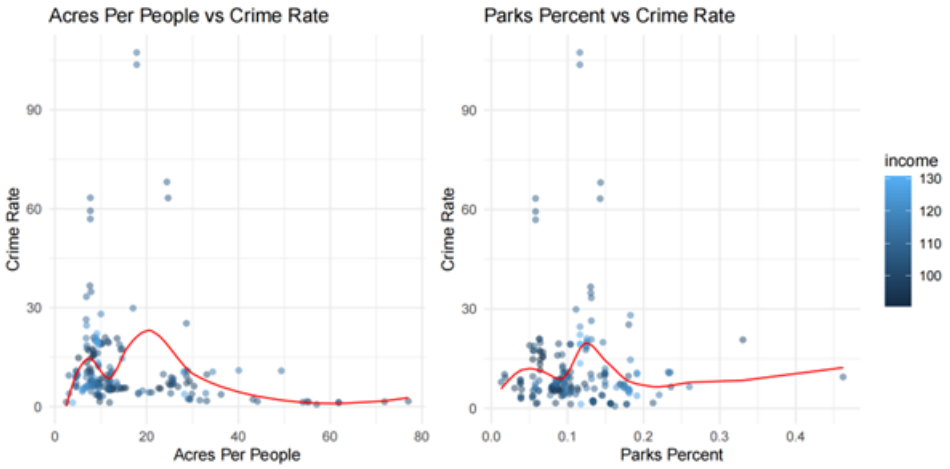

Figure 14: Relationship between green space per capita (acres per person) / Percentage of City Area Covered by Parks (Parks Percent) and crime rate.

This model simultaneously considers additional factors that might impact crime rates, such as income level. In multivariate local regression, it's important to note that as the number of variables in the model increases, so does the demand on the data. This is because effectively estimating relationships among multiple variables requires a sufficient number of data points to cover the higher-dimensional space.

Our analysis also takes into account factors like income levels. From Figure 15, the red dots represent the predictions of the multivariate local regression. The relationship between per capita green space and crime rates remains unclear. On the right, the red dots denote the model's predictions. This graph also indicates an unclear relationship between the percentage of park area and crime rates.

Even when other variables are considered in the multivariate local regression model, the relationship between greenery levels (measured either by per capita green space or the percentage of park area) and crime rates is still not explicit. This shows that there is no significant relationship between urban greenness and crime rates.

6 Conclusion & Discussion

The primary objective of this paper is to employ permutation tests, the Markov Chain Monte Carlo (MCMC) method, and local regression to investigate the potential relationship between green spaces and crime rates. The null hypothesis posits that "There is no relationship between the crime rate and AcresPerPeople." To achieve this, we collected data on crime rates, population numbers, Acreage of Parks per 1000 residents (used to represent green spaces), and GDP values from various cities across the United States. After thorough data preprocessing, we retained eight years of relevant data for 24 cities.

As shown in Figure 8, the results of the permutation test following bootstrap analysis

```
## Estimation type: Local Regression
##
## Call:
## locfit(formula = y ~ lp(AcresPerPeople, ParksPercent, income),
##     data = data)

## Number of data points:  191
## Independent variables:  AcresPerPeople ParksPercent income
## Evaluation structure: Rectangular Tree
## Number of evaluation points:  62
## Degree of fit:  2
## Fitted Degrees of Freedom:  24.895
```

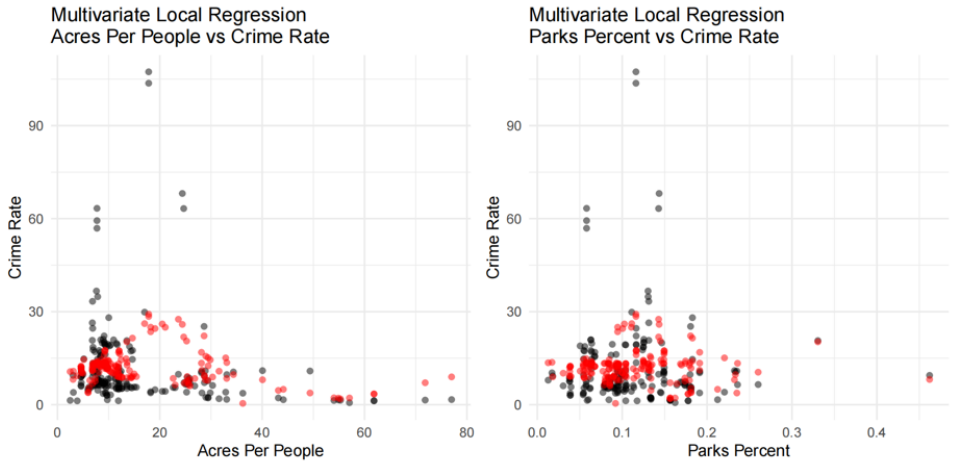

Figure 15: Relationship Between Per Capita Green Space (Acres Per People) / Percentage of City Area Covered by Parks (Parks Percent) and Crime Rate.

reveal that most p-values are less than 0.05. However, it's worth noting that the upper limit of the confidence interval is considerably larger. This raises some uncertainty about rejecting the null hypothesis, as the rejection might be due to chance.

Moving on to the results obtained using the MCMC method, we observed that with the selection of an appropriate prior distribution, the coefficient of AcresPerPeople has a likelihood of more than 99% of being negative, and its mean is also negative. It's important to mention that the chosen standard deviation for the prior distribution is relatively large. This may be attributed to the coefficient's lack of stability, and it suggests that the formula we developed for the crime rate concerning AcresPerPeople may not be very precise. However, the consistent appearance of negative coefficients could indicate a negative relationship between green spaces and crime rates.

Local regression result enhanced above discussions. Both local regression and multivariate local regression indicates that there is no explicit relationship between urban greenness and crime rates. Even though we take the income level factors into account, the relationship is still unclear.

Fortunately, we have gathered enough evidence to confidently reject the null hypothesis, suggesting a negative correlation between green spaces and crime rates. However, re-

grettably, we are unable to provide a consistent mathematical expression that perfectly captures this relationship in the data.

In conclusion, our findings can offer valuable insights to government authorities. When it comes to investing in crime reduction strategies, allocating resources towards urban greenery initiatives is a worthy consideration.

However, many improvements exist for further research. In general, our sample is not large enough, and if we can collect data from more cities, the results may be more rigorous. Besides, in order to obtain a more accurate coefficient for the greening space, we may need to add more potential variables affecting the crime rate, including Population density, poverty rate, average level of education and so on.

7 Contribution

All three members discussed together and decided the topic of this research. All three members conducted data cleaning and exploratory data analysis together. Ziyang Xiong and Xixiao Pan completed all parts about permutation test and MCMC methods together, and Jiawen Liu completed all parts about local regression method.

References

[1] Z. Zhu, H. Pei, B. S. Schamp, et al., "Land cover and plant diversity in tropical coastal urban Haikou, China," Urban Forestry Urban Greening, Vol. 44, 2019, p. 12. [Online]. Available: https://doi.org/10.1016/j.ufug.2019.126395. Accessed on: Nov. 30, 2023

[2] S. S. Ogletree, L. R. Larson, et al., "Urban greenspace linked to lower crime risk across 301 major U.S. cities," Cities, 2022. [Online]. Available: https://doi.org/10.1016/j.cities.2022.103949. Accessed on: Nov. 30, 2023

[3] W. Huang and G. Lin, "The relationship between urban green space and social health of individuals: A scoping review," Urban Forestry Urban Greening, Vol. 85, 2023, p. 127969. [Online]. Available: https://doi.org/10.1016/j.ufug.2023.127969. Accessed on: Nov. 30, 2023