ASTRO 410 — Final Project

Using two-point correlation function to analysis US population clustering

Haoyu Wang

May 1, 2021

Introduction

Clustering usually refers to the tendency for subjects of the same type to be distributed in a non-random way, thus creating one or more feature distances between them. In the area of Machine Learning, people often use clustering when studying galaxies, population, and the distribution of mountain ranges. By analyzing clustering, one can often figure out the probability of finding the same type of research subject over a certain distance. Further, we can simulate the distribution of research objects in unknown areas through our known databases. However, there may be some bias in clustering, which could be caused by a variety of factors, so these factors should also be taken into account in the process of computing, and I will explain in detail the factors for population clustering bias later in this report.

US population distribution

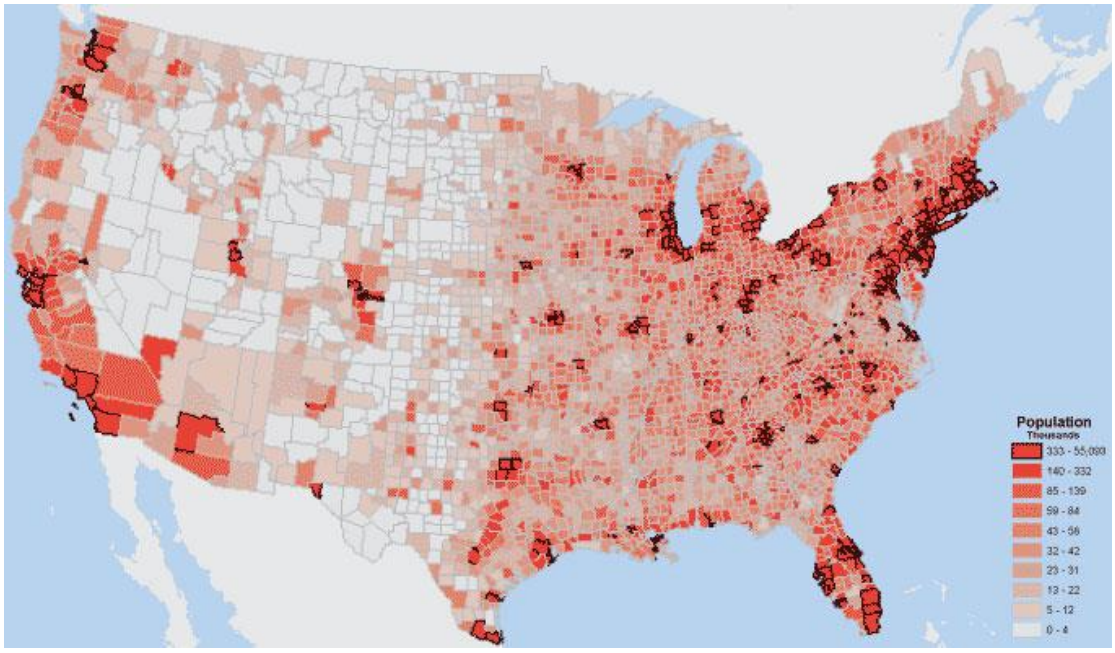

Figure 1

The quantification of clustering is a powerful process that can lead to physical insights about the objects in question. For example, from figure 1, we can clearly see that the distribution of population in the United States is not uniform, and the phenomenon of population clustering appears in some places, such as, California, Florida, and some Northwestern states. This kind of distribution intrigues me, and it is also the subject of this paper.

In addition to studying population clustering, many astronomers are making pretty good headway in the area of galaxy clustering. For example, MAssive Cluster Survey (MACS) discovered the majority of the targets studied in depth by the Hubble Frontier Fields initiative and, more recently, the extended MAssive Cluster Survey (eMACS) focuses on yet more distant systems at

redshifts beyond z=0.5. By studying such clustering, we can better understand the formation and operation of galaxies.

Methods & Models

For this project, I used two-point correlation function to evaluate population clustering, and it describes the excess probability of finding two objects separated by certain distance. Before start computing the function, I collected a census of 1,000 cities and towns in the United States. Since there are nested for loops in my program, (although I have written most nested for loops as comprehension, it still takes very long time to run thousands of data), I divided them into five different groups by setting different population range to ease the burden on my program, as shown in table 1.

Group	Population
1	Below 50,000
2	50,000 to 80,000
3	80,000 to 100,000
4	100,000 to 200,000
5	Above 200,000

Table 1

Next, I created sets of random points to correspond to each population data group, and made the random sets have the same amount of data with corresponding datasets. Meanwhile, the random longitude set was limited between -176 to -65 degrees, and random latitude set was limited between 17 to 71 degrees. These limited was set after finding the maximum and minimum values for longitude and latitude respectively in my dataset.
So with these data sets, I introduced the pairwise separation measure,$\Phi_\theta(x,y)$. This function determines if a given pair of points has angular separation between θ and $\theta + \Delta$. In this project, I set θ_{min} to be 4.5° , and θ_{max} to be 175°.

$$\Phi_\theta(x,y) = \begin{cases} 1 & if\ \ \theta < \theta_{sep} \leq \theta + \Delta \\ 0 & otherwise \end{cases}$$

With this formula, we can determine the total number of pairs of points with angular separations for the possible values θ_{sep} of using a double sum. And the angular separation between two points on a sphere is given by the following equation, where (x1,y1) and (x2,y2) are the coordinates (longitude, latitude) of the points, and they are transformed into a coordinate between $[0, 2\pi]$ and $[0, \pi]$.

$$\cos(\theta_{sep}) = cosy_1 cosy_2 + siny_1 siny_2 \cos(x_1 - x_2)$$

In order to make the computation less expensive, I modified the angular separation equation using the cosine difference formula.

$$\cos(\theta_{sep}) = cosy_1 cosy_2 + siny_1 siny_2(\cos x_1 cosx_2 + sinx_1 sinx_2)$$

In order to get a pair of points, we have three different choices: 1. Choose both points from dataset; 2. Choose one point from dataset, and the other one from random set; 3. Choose both points from random set. And a point is not allowed to be its own pair in all three cases. So we need to calculate the double sum for each cases, and formula are shown below. where D is the set of data points and R is the set of random points.

$$P_{DD} = \sum_{x\in D}\sum_{y\in D} \Phi_\theta(x,y)\,; \qquad P_{DR} = \sum_{x\in D}\sum_{y\in R} \Phi_\theta(x,y)\,; \qquad P_{RR} = \sum_{x\in R}\sum_{y\in R} \Phi_\theta(x,y)$$

After that, we normalize the three different pairwise sums by dividing by the number of pairs we calculated. Where N is the number of points in each dataset.

$$DD = \frac{P_{DD}}{N(N-1)}; \qquad DR = \frac{P_{DR}}{N^2}; \qquad RR = \frac{P_{RR}}{N(N-1)}$$

Finally, plugging the numbers we just calculated to the correlation function (Landy & Szalay), which is showed below, we can get the correlation at the specific value θ_{sep}.

$$\omega(\theta) = \frac{DD(\theta) - 2DR(\theta) + RR(\theta)}{RR(\theta)}$$

Another important quantity of clustering is bias. In this project, US population clustering bias can be caused by the distribution of population in the individual state. So the equation of bias can be written as the square root of the angular correlation function of the population in question to that of population of individual states. As I mentioned earlier, the clustering appears in Florida and California. So I collect the population data for these two states, and calculate their angular correlation function. Since there are both positive and negative values in the angular correlation functions, I decided to calculate bias square, b^2.

$$b^2 = \frac{\omega_{US}}{\omega_{FL}} \quad or \quad \frac{\omega_{US}}{\omega_{CA}}$$

Results

When I was calculating the angular correlation function for each population group, I selected 15 angles evenly spaced between $4.5°$and $175°$. In order to get the error, I run the program 10 times for each angle in each group, and present the standard deviation as the error. After that, I used SciPy's cubic spline function to create smooth curves between points. And it turns out most of the images have curves that look like power law $\xi(\theta) \sim \left(\frac{\theta}{\theta_0}\right)^{-\gamma}$, the value of γ varies from group to group. The errors in some graphs may be caused by the instability of randomly generated data, and the inaccurate geographical location of cities may also lead to the imperfect curve generated. Here are the final plots for each population group, and the clustering bias.

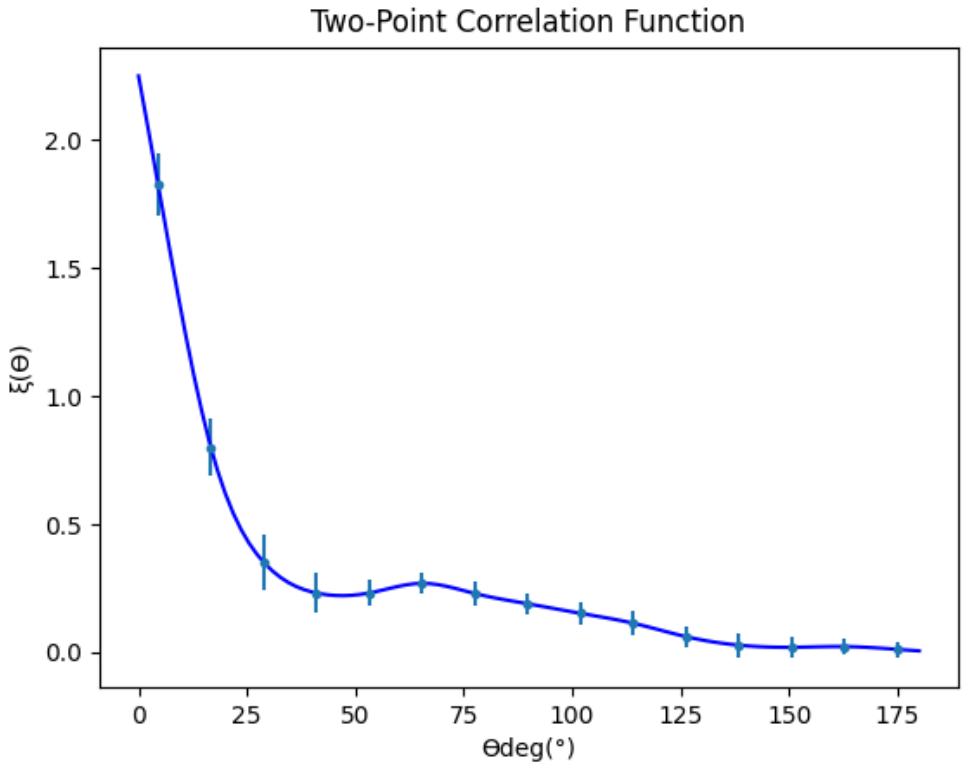

Figure 2

Here, I plot the angular correlation function of population between 50,000 and 80,000, and there is a small bump around $70°$.

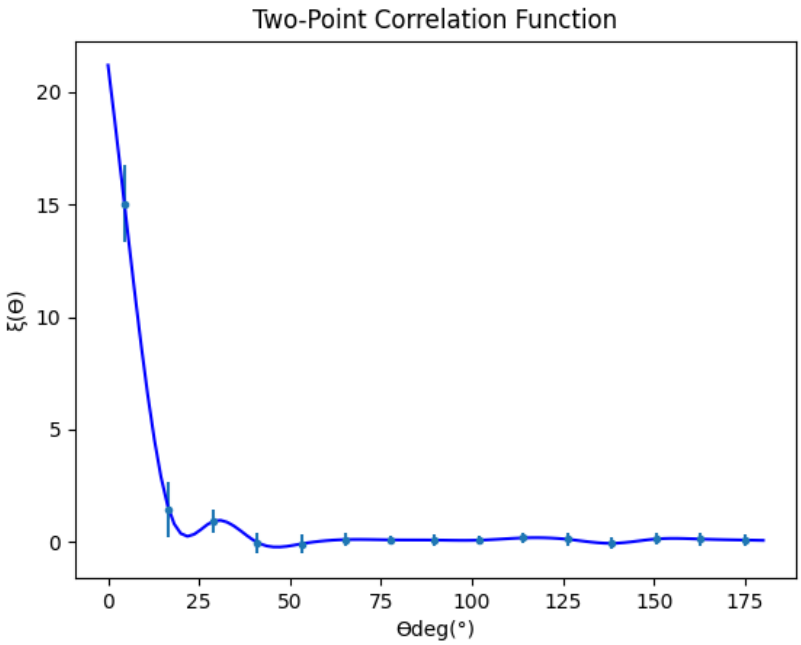

Figure 3

Here, I plot the angular correlation function of population between 80,000 and 100,000, we notice a high degree of clustering at low angular separations, and there is a small bump around $30°$.

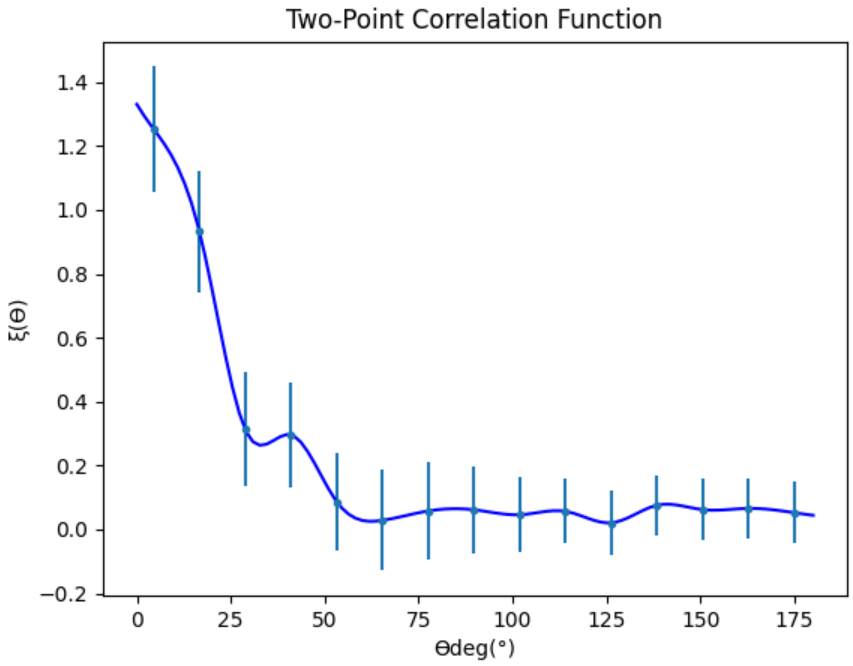

Figure 4

Here, I plot the angular correlation function of population between 100,000 and 200,000, and there are several bumps in the graph, the conspicuous one is a around $35°$.

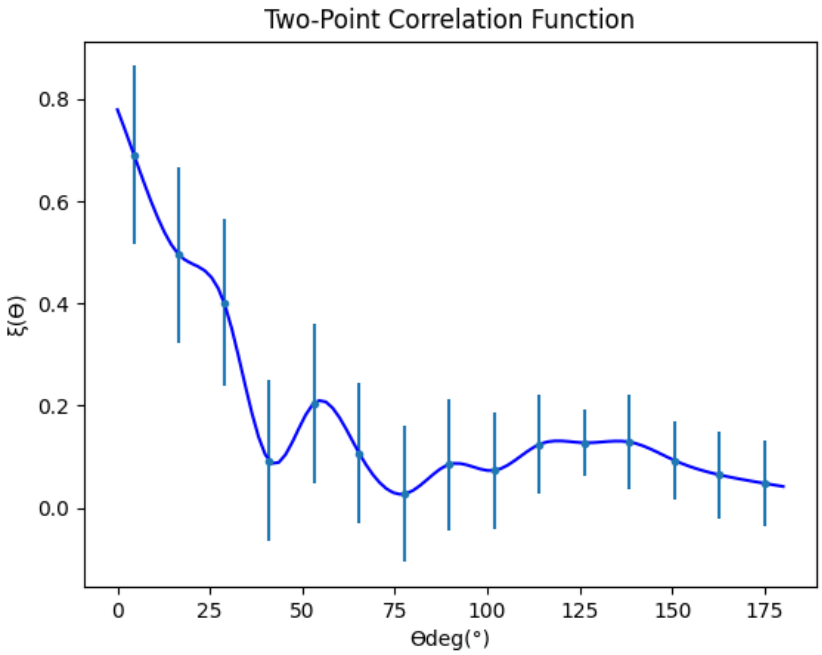

Figure 5

Here, I plot the angular correlation function of population above 200,000, and there are several bumps in the graph, the conspicuous one is around $55°$. The small number of cities with a population of more than 200,000 also contributes to the picture's imperfection.

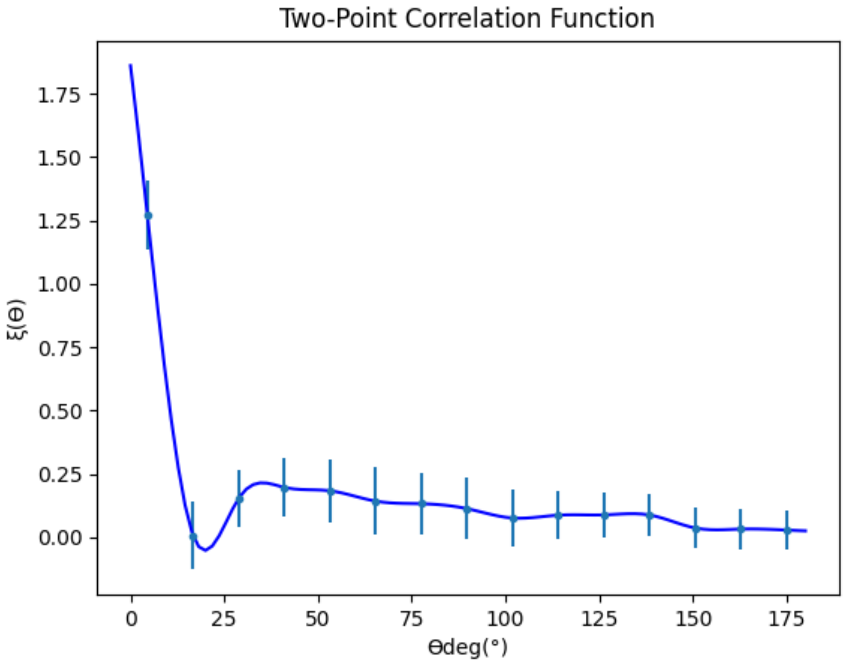

Figure 6

Here, I randomly selected 257 cities from the group of population below 50,000 and plot the angular correlation function. There are several bumps in the graph, but none of them are conspicuous.

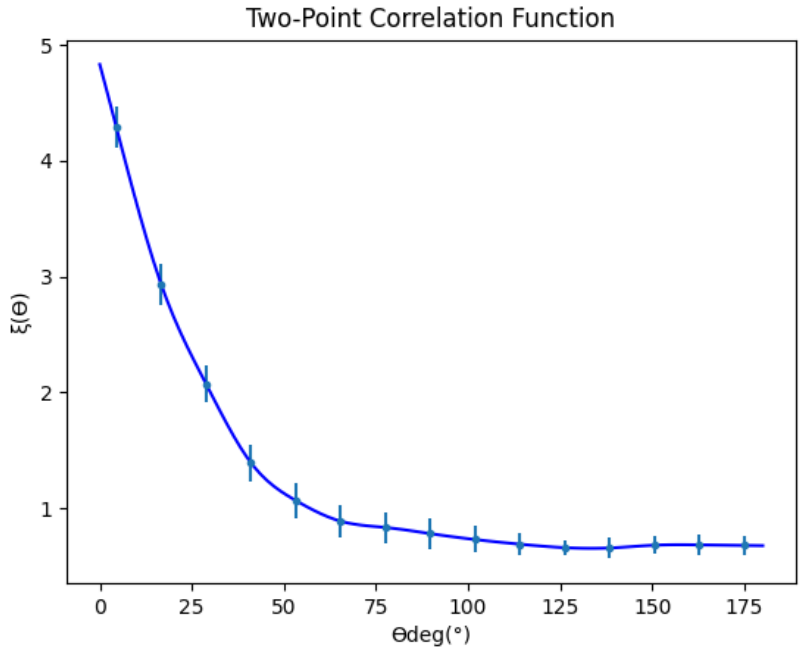

Figure 7

Here, I plot the angular correlation function of California State population, and there are no conspicuous bumps in the graph.

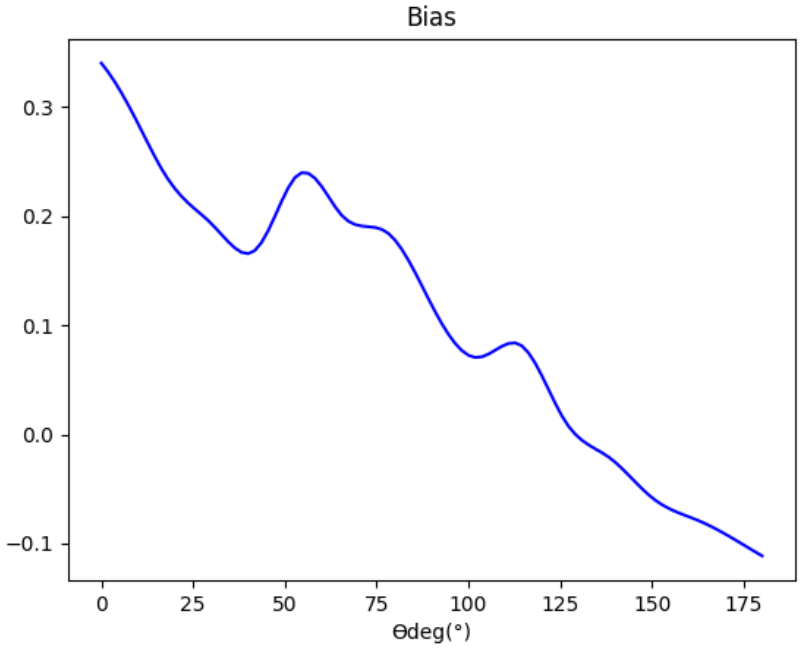

Figure 8

Here I plot the bias of the population between 50,000 and 80,000 with respect

to California State population.

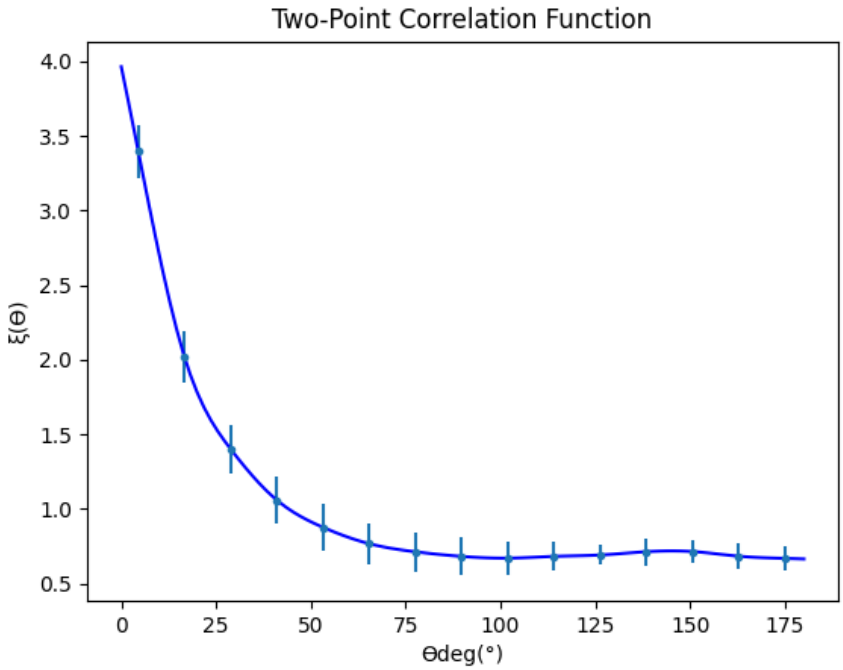

Figure 9

Here, I plot the angular correlation function of Florida population, and there is a little bump around $150°$ in the graph.

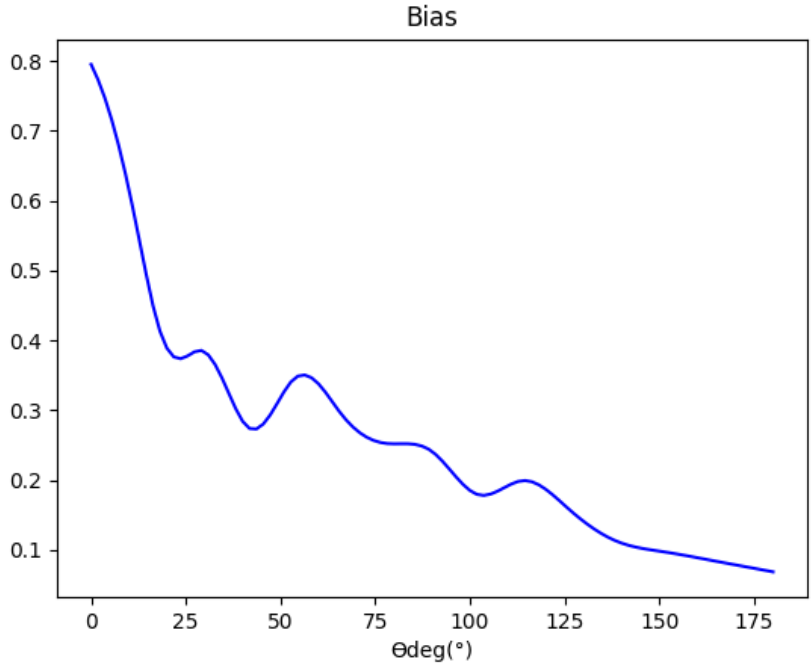

Figure 10

Here I plot the bias of the population between 50,000 and 80,000 with respect

to Florida State population.

Conclusions

From these plots of angular correlation function, we can clearly see that population tend to be highly clustered at small angular separations, and drops dramatically with angular separation increase, like a power law. Meanwhile I implemented correlation function for the states which have large population, California and Florida, to test clustering bias.

The biggest problem the project has encountered so far is the limitation of computation. Even though I've collected population from all over the country, and got a very good database, I can't fit them all into my program. Though I rewrite nested for loops as comprehensive, my program can only run no more than 500 data at a time. So when it comes to the group which includes huge amount of cities, such as group of population below 50,000, I have to randomly select data from my database. Doing so, however, can lead to inaccurate results. So using a supercomputing facility, such as, CyberLamp or ROAR, would be a great improvement for this project.

In addition, calculating power spectra by employing the Fourier Transformation on the correlation functions would be another way to improve the data evaluation. The real test of population clustering would come from simulating the formation and evolution of cities and towns from the beginning with various initial conditions such as marketing, employment, and weathering. I believe with these conditions in place, the results will be more authentic and reliable.

Reference

Galaxy Clustering. (2017). Institute for Astronomy, University of Hawaii. https://www.ifa.hawaii.edu/research/Galaxy_Clustering.shtml

Springel, V. (2017, December 22). First results from the IllustrisTNG simulations: matter and galaxy clustering. Monthly Notices of the Royal Astronomical Society,. https://academic.oup.com/mnras/article/475/1/676/4772886

Selection Effects on the Two-Point Correlation Function of Voronoi Distributions of Galaxies. (2004). Jonathan Heiner. https://www.astro.rug.nl/opleidingsinstituut/reports/Astro_Ms_2004_JHeiner.pdf

Martin, K. (2000, May). A Comparison of Estimators for the Two-Point Correlation Function. NASA/ADS. https://ui.adsabs.harvard.edu/abs/2000ApJ...535L..13K/abstract