# Pubs vs Starbucks': Understanding Spatial Point Process Analysis Methods

**Author: Anup De**

## Abstract

This project aims to model spatial point process data and determine to what extent analysis can be garnered from known methods of spatial point process model distribution. It also seeks to apply these methods to simple datasets that have interpretability in social and economic analysis. Because of the difficulties in understanding spatial point process without temporal data, studying this field, especially when using a simpler dataset, is of importance in the field of point process data. To approach the problem, this research explored homogenous Poisson spatial distributions, and Papangelou distributions, along with applying Matern's Hard Core Process and its ramifications in analysis. These processes are applied to datasets showing the geographic locations of pubs and Starbucks' café restaurants in the city of London. It was found that the certain aspects of the Homogenous Process can describe Starbucks locations while data features prevent much analysis on the pubs and effective comparisons between models.

## Introduction

With limited data arises limited possible analysis. In the data-driven age however, it is important that statistical analysis can still be done (or at least better understood the restrictions) on purely spatial datasets. These datasets are more common than spatiotemporal data, as any geographic tool (i.e. Google Maps), can serve as an easy data collection option. To make clear these limitations in interpretability and applied processes, it is perhaps best to use the simplest data possible. That is why using locations of pubs and Starbucks' locations throughout London were chosen to represent the spatial data.

Furthermore, there are a number of practical applications for useful spatial analysis. Modern geographic tools could easily implement a feature that presents the likelihood of a certain locations being present in an area. It could aid businesses (such as pub owners and Starbucks' franchises) to open or close locations depending on how spatial distributions currently appear.[1] This analysis could even reverse engineer locations and population centers by understanding the locations of some known process, such as Starbucks' cafes, which are likely nearer to city centers. Finally, based upon analysis of many economic or social locations, information can be given that shows the most profitable or most useful geographic locations for newer locations. For example, given the locations and spatial

[1] This technically does imply some history feature, as new stores are opening upon knowledge of other existing stores, but practically the information of where stores currently are is trivial (as it does not matter when they opened). This application of this analysis still does not require data that presents history of all current stores.

analysis of tourist centers, competitor hotels, and restaurants in Brussels, a new hotel chain could find most optimal regions to open hotels.[2] Use of hard-core processes could be useful for modeling new hotels against competitor hotels, while a Papangelou distribution might aid in understanding how existing tourist locations or restaurants influence where hotels optimally appear.

Any of these possible endeavors would be aided by this analysis, as it will present the limitations that must be held, and also show how current methods of statistical point process understanding can be used to begin each of those practical applications. It is also of note that this analysis provides insights into how a Papangelou and Matern process could be applied to practical data along with interpretations and feasibility. This is important in reasoning how useful these processes are how they can be advanced theoretically do be better suited for interpretations.

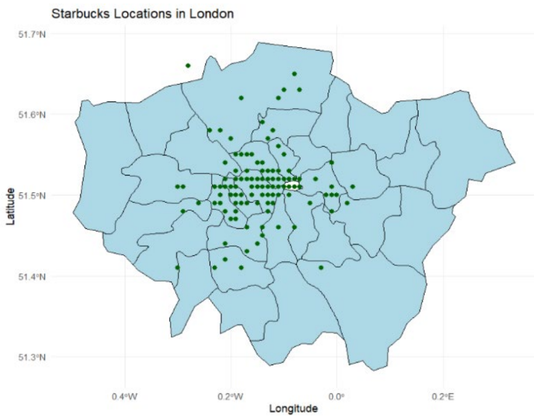
## Exploratory Data Analysis

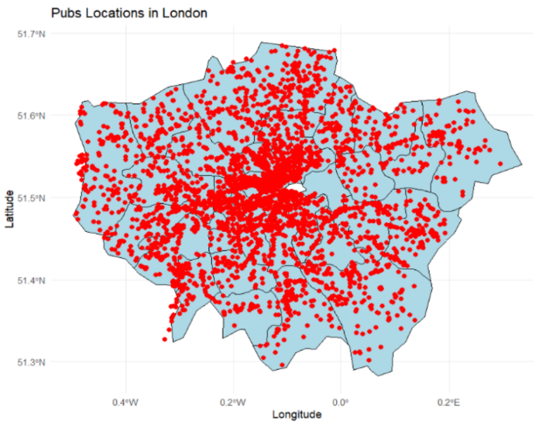
This project data included information about the 901 Starbucks' locations across Great Britian. Relevant to this analysis is the given cities as well as the franchise's longitude and latitude. In order to further limit the data, the project restricted the data to locations within London. This was done so that the data analysis could be easily reviewed and tested for logical interpretability. Furthermore, it is likely that the distribution of Starbucks' across the island of Great Britian is vastly different than that of the distribution variation across London.[3] Also the relevant use-cases for this analysis largely operate on a city level and therefore it is reasonable to limit the dataset to the 193 locations within London. This could obviously limit the generalization level of this analysis, but as limiting the data made the research more feasible, the choice was taken to perform this data cleaning. Performing the same data cleaning techniques on the pubs dataset shows 4,273 pubs within London along with relevant longitude and latitudes of each datapoint. London was nicely chosen to its relatively quadrilateral boundaries, which made visuals over a grid much easier than performing visuals over a field representing the area of Great Britian, for instance.

By using shapefiles, the project could also visualize the Starbucks' and pub locations within London as a view of London's administrative districts. While these districts are not relevant to later analysis, it does enhance the data visual and could be used in further research.

[2] Again there is some history relationship here, in that this assumes this hotel chain has no current hotels. When it comes to data collection however, this assumption is trivial. But because this does not require analysis based upon the history of either those tourist locations, competitor hotels, or restaurants, this can be classified as purely spatial.

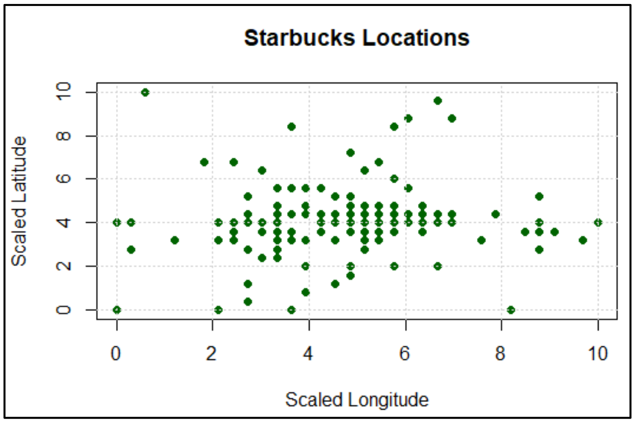
[3] Including the entirety of Great Britain would likely find extreme centrality around London and this would make practical analysis hard (there is not much need in concluding that there are more Starbucks' in London then elsewhere). Ideal data would have widespread datapoints that does not cause major centrality issues.

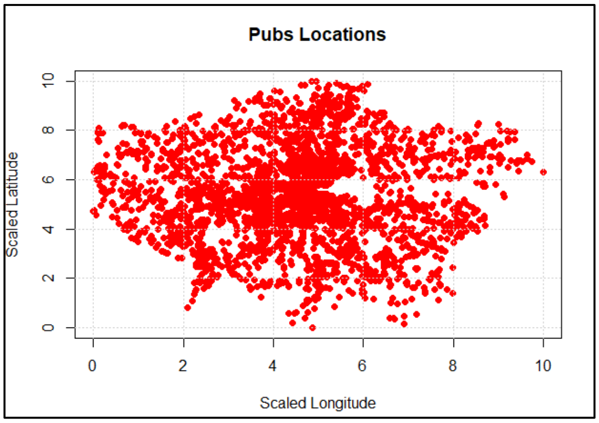




Since the geographical data was presented in the form of longitude and latitude, it would make sense to scale both to ranges of [0,10]. This removes any dependence between longitude and latitude and is a form of map projection onto a 2-dimensional level, a minor modification that makes for simpler interpretation. It also largely removes the London administrative districts from the map. Below are the distributions of Starbucks' and pubs on these scaled grids for longitude and latitude.

## Methods





### *Homogenous Poisson Process (Number of Locations per Circle of Radius 1 Around Each Location)*

The first analysis performed was to plot the number of neighbors around each location. Neighbors would be defined as number of locations that are within 1 unit radius of the center location.[4] With this information about data, the number of locations within 1 unit radius could be set against a Weibull distribution, with a lambda rate parameter set as the average number of locations within 1 unit radius. This was performed on both the Starbucks' dataset and the pubs dataset and was evaluated with a Q-Q plot and a Mean Squared Error of the Weibull predictions. With the Weibull distribution now present for the entirety of the field, likelihood of the data being present at the locations it was found at

[4] One unit represents units along the scaled longitude and latitude metrics.

were calculated. Since we will compare this model to other models that use the same dataset (either Starbucks or pubs), we can simply add the scaled lambda (scaling such that the field sums to 1) for each point and make a fair comparison between models.

### *Homogenous Poisson Process (Distance Till Nearest Starbucks/Pubs)*

The second analysis model applied to each dataset was the closest distance to another location. With this information, a new lambda rate parameter could be set as the average of these closest distances and used in the Weibull distribution. Once again evaluations of these models were done with a Q-Q plot and Mean Squared Error.

### *Papangelou Process (Number of Close Neighbors to Starbucks/Pub Locations)*

The Papangelou Model that this research employed, attempted to find the distribution of Starbucks/pub locations based upon the distances to all other locations. Therefore the Lambda Papangelou value for each data point, *i,* is:

$$= \beta + e^{-\alpha * \sum_{j=1}^{n} dist(i,j)}$$

We can then interpret the Papangelou value as a form of centrality for each datapoint. This is because the Papangelou value for each datapoint is a form of an inverse of the pairwise distances, meaning locations that have all other locations close to it, will have a higher Lambda Papangelou. We can then compare this with another measure of centrality, which is the existing number of close neighbors a value has. Once more Q-Q plots and Mean Squared Error estimates are produced for this modeling. Finally visuals regarding the strength of Lambda Papangelou can also be constructed, both for the actual data points and for the field of interest.

### *Papangelou Process (Distance Till Nearest Starbucks/Pubs)*

Another way to evaluate the Papangelou process is to derive it in terms of the distance to the closest location. It will still use the distance to all other locations from a specified point, but this new Lambda Papangelou will solve for the distance until 1 neighbor as shown to the right:

$$\text{Expected Number of Points} = \pi r^2 \lambda(x)$$

$$1 = \pi r^2 \lambda(x)$$

$$r = \frac{1}{\sqrt{\pi \lambda(x)}}$$

This is a projected distance to the closest neighbor which can again be modeled across the entire field of interest or evaluated against the actual closest neighbors of each point. Q-Q Plots, Margins of Error, and likelihood probabilities can also be found similarly to the Homogenous Poisson Models.

### *Matern Process*

The Matern Process seeks to find the number of points that can be accepted under a specific hard-core parameter. Once the datapoints are ordered, points are accepted if all points existing within the already accepted datapoints have distances larger than the parameter. Largely, the ordering of the numbers matters extremely high, and without a history process, this is a difficult task to garner interpretability. The visuals created from this can show how many and what points can be accepted under various parameters.

### Unreasonable Assumptions

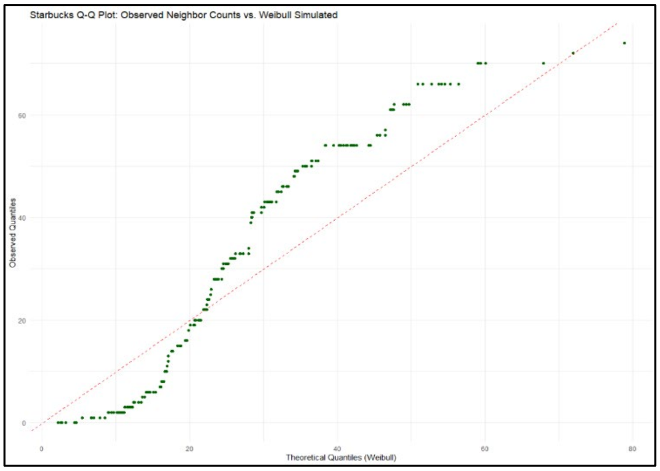
The metrics that are being compared against this Papangelou model are the Number of Locations within a Circle Radius 1 and a Distance Till Nearest Location. Since the dataset was isolated to look within London, these metrics are slightly misleading without the correct assumptions. For instance, a datapoint near the edge of London will only have a few Starbucks / pubs in London within radius 1 and a certain distance to the nearest other location in London. It is actually likely that there are more locations within this radius that are not in London and that the nearest distance is shorter, but to a location that is outside the boundaries of London. Since both the model and the metric use these potentially flawed numbers, the comparisons may be valid, but the interpretation must come with an assumption that there are no Starbucks/pubs outside of London. Future studies can still focus on London and counter this fact by still counting the number of Starbucks/pubs that are outside London, but only computing for field spaces within London.

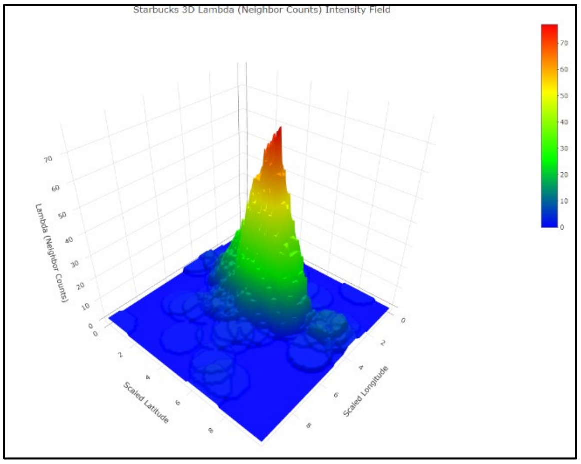
## Results

The results will be shown per model. Within each model, analysis of the Starbucks dataset will be followed by analysis of the Pubs dataset. Modes will be separated by their headers.

### Homogenous Poisson Process (Number of Loc. per Circle of Radius 1 Around Each Loc.)

Shown in the below figures are the Q-Q Plot of observed counts of Starbucks locations within radius 1 in dark green against a Weibull simulated theoretical counts. Also shown is a 3-D plot of this Poisson model over the entire field of London.

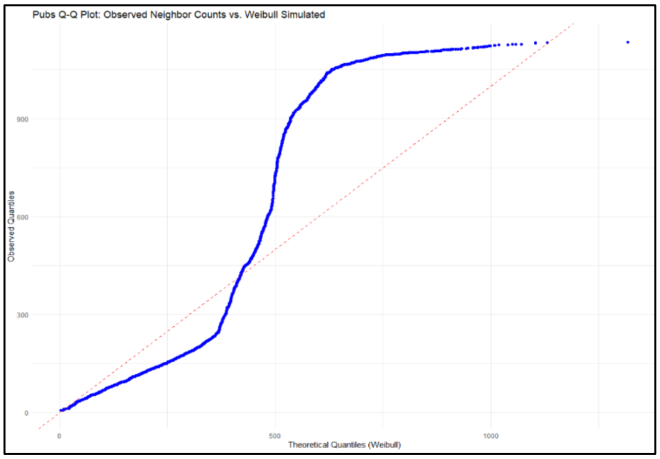


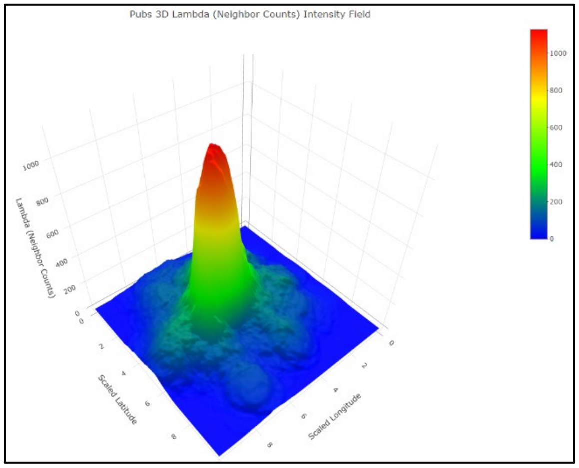


The quantiles for the observed entities follow closely to those of those of the Weibull distribution on the scale of all the datapoints. Localizing analysis to specific regions of points, would show that the lowest observed quantiles (those with the fewest nearby locations), are expected to have more neighbors under Weibull. This is noted in the obvious centrality of the data, which a Weibull distribution cannot note.

This yields a Mean Square Error of 89.1 and as per the formula outlined in methods, the likelihood probability is 0.0005656.

Shown in the below figures are the Q-Q Plot of observed counts of Pubs locations within radius 1 in dark green against a Weibull simulated theoretical counts. Also shown is a 3-D plot of this Poisson model over the entire field of London.



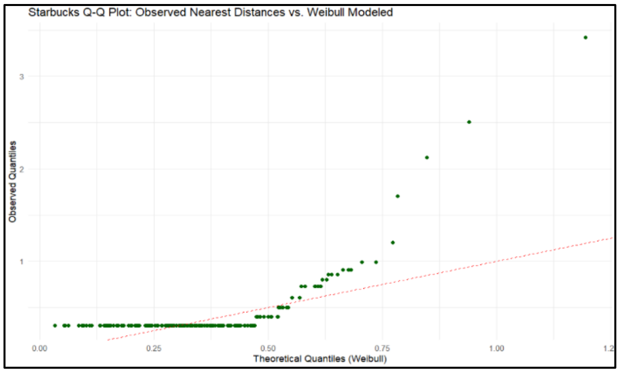


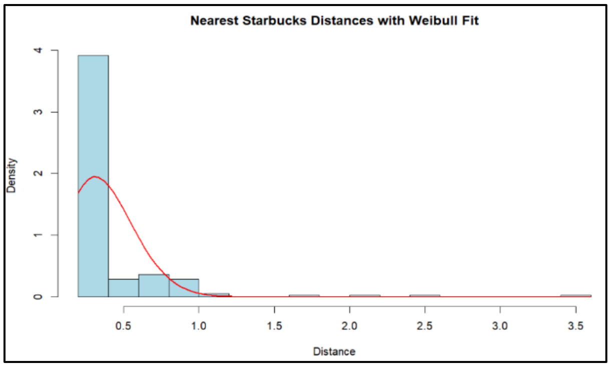
The quantiles follow closely to the expected Weibull distribution for the smaller quantiles but quickly deviate. More pubs fall in the highest quantile level of nearest neighbors than expected. About half feature themselves in these higher quantiles.

This yields a Mean Square Error of 37749.42 and as per the formula outlined in methods, the likelihood probability is 0.0003268.

### *Homogenous Poisson Process (Distance Till Nearest Starbucks/Pubs)*

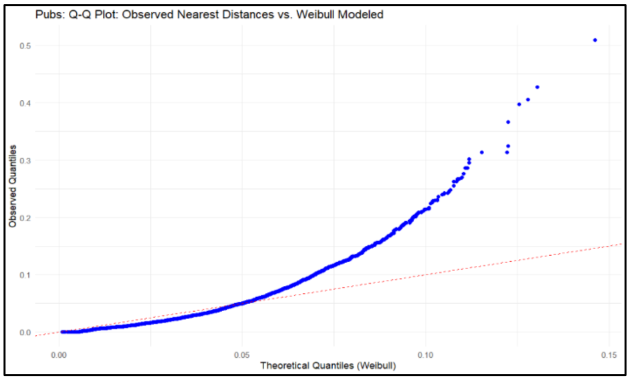
Below is the Q-Q Plot for the Homogenous Poisson Model when considering the lambda rate parameter as the distance to the nearest other Starbucks. When considering this model against the proposed Weibull model, the following Q-Q plot and histogram are created. The histogram appears with the true data values in blue and the proposed distribution in red.

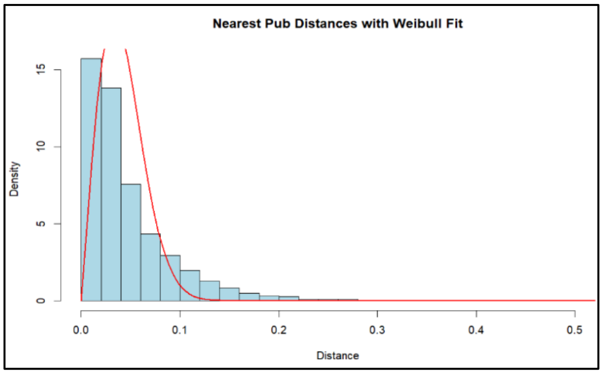




The Q-Q Plot gives insights into the dataset and shows that several datapoints have the minimal distance to nearest location showing their centrality. The fact that locations outside of London were not included means that the data breaks the Weibull distribution as some datapoints on the edges of London have an artificially high distances till the nearest location. The Mean Squared Error of this distribution's predictions is 0.068.

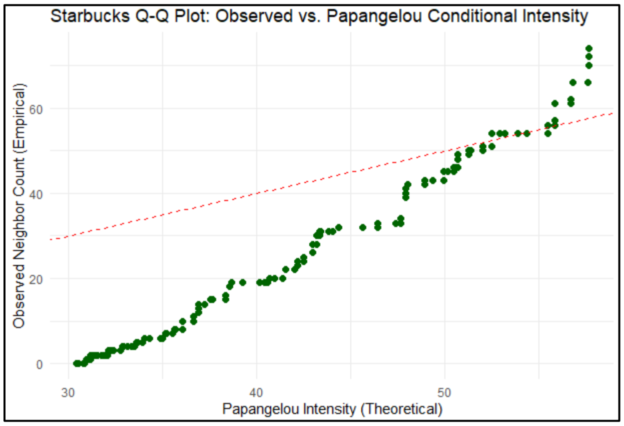
For the pubs shown below, the dataset (which has more points) features a smoother version of the Starbucks data, again showing that the points with a farther closest distance fail to fit the standard Weibull distribution. Pubs maintains less centrality however, as shown in the histogram, which shows nearest distances maintained at several bins rather than just the first. The Mean Squared Error is 0.00075.

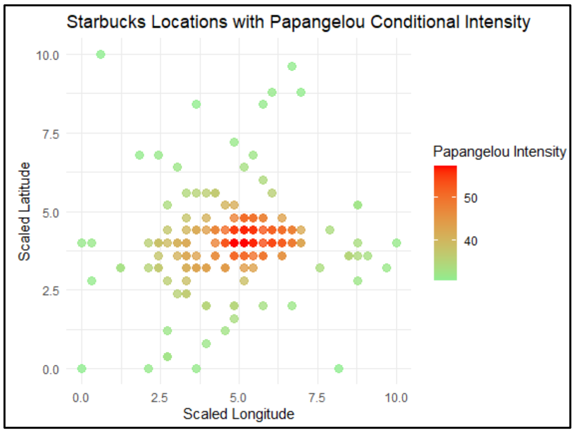




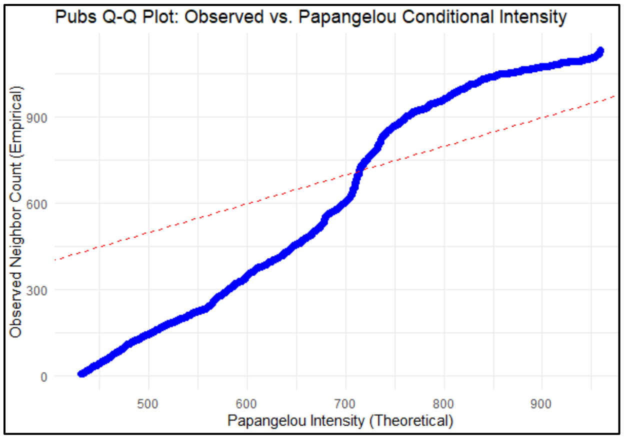
### *Papangelou Process (Number of Close Neighbors to Starbucks/Pub Locations)*

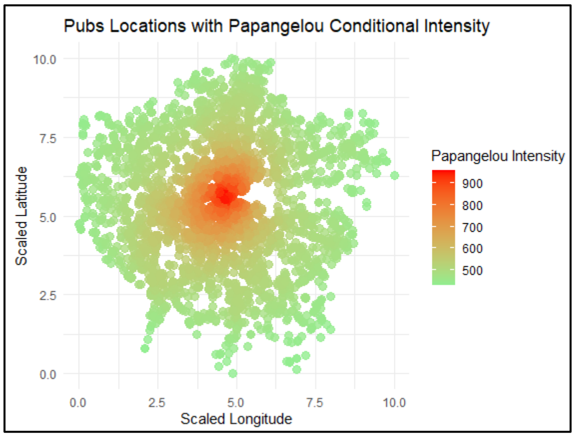
Shown above are the Q-Q plots for the Papangelou Process that was built for modeling the number of neighbors within a 1-unit radius of each individual location. Along with that is the colored map showing the intensity found from this Papangelou process for each datapoint. For Starbucks datapoints, the Papangelou conditional intensity that is describing the number of neighbors within a 1 radius circle does capture the growth towards the center of London but perhaps not the extent of this growth, as the true datapoints follow the strongly linear growth but towards the center (where locations with the highest number of neighbors within 1 radius would be. The Mean Squared Error of this distribution's predictions is 364.13.





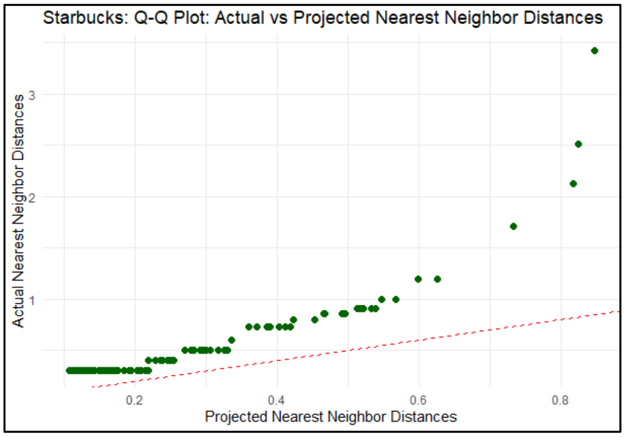
Shown next in figures below are the Q-Q plots and colored maps for the Papangelou Distribution relevant to the number of neighbors in the nearby region of 1-unit radius. These datapoints reach a threshold of about 900+ locations much faster than the Papangelou models expects and also much faster than Starbucks' locations are able to. Due to the more widespread nature, the Papangelou model holds closer to the true datapoints but still cannot reflect the rate of increase of intensity as the model moves closer to its center. The Mean Squared Error is 83763.17.

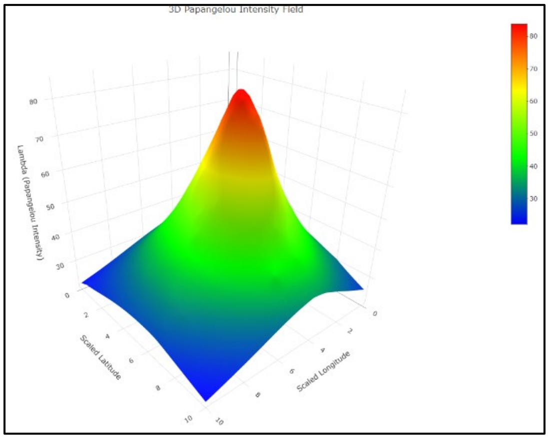




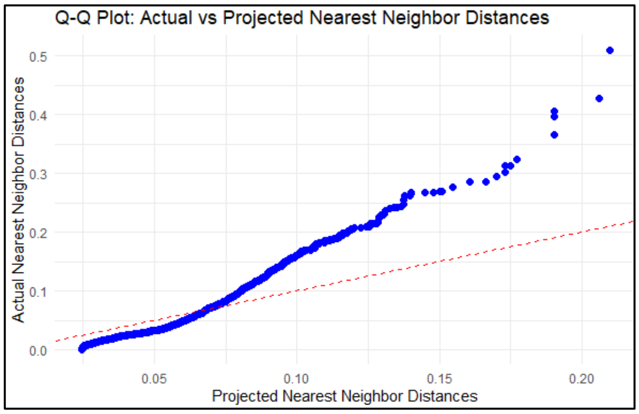
### *Papangelou Process (Distance Till Nearest Starbucks/Pubs)*

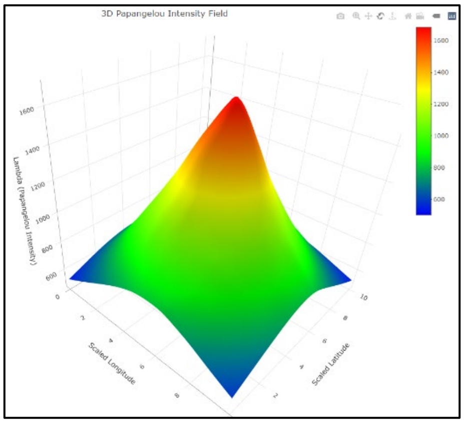
The Papangelou Process that determines the distance until the next location reflects similar problems as earlier analysis has shown as represented in the below figures showing Q-Q Plots and 3D models across the space of interest. Once again, the bigger distances in the dataset (closer to the edges of London) are not captured well. The general distribution of the Starbucks closer to the center however are captured better as they nominally increase, the further they are from central London. The Mean Squared Error now is 0.107 with a likelihood of 0.0015.





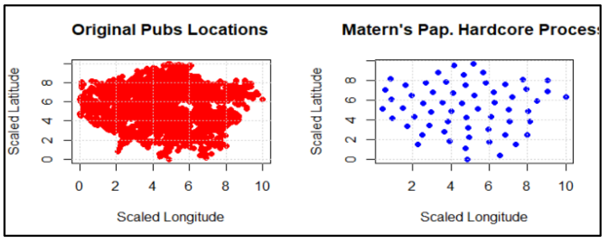
For the pubs shown below, the Q-Q Plot features less representation from the predicted Papangelou distribution. The nearest neighbor for each point is thought to be much closer than actuality, due to the more spread data (more pubs having longer distances, rather than close cafes and very far cafes). Pubs across the space mean that the peak likelihood is smaller and features more spread than the Starbucks. The Mean Squared Error is 0.0006 and the likelihood is 0.00013.

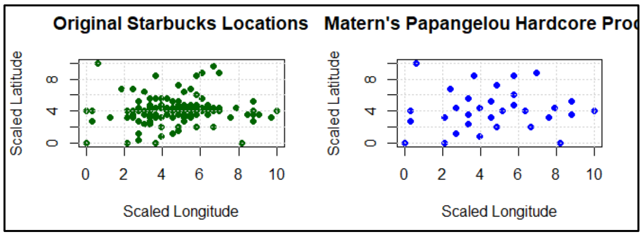




## Matern Process

Finally the Matern Process that was done on each dataset was simply done to show what levels of hard-core refraction could be applied to each dataset safely. Model building was not applicable because there was little practical reason to remove datapoints and imagine different distributions. Below are the Matern Process' for each Starbucks and Pubs. On the left of each image will be the actual distribution and then the Matern Process applied for a hard-core parameter of 1.





## Conclusion

Neither the Homogenous Poisson nor the Papangelou methods fit the data particularly well. Analyzing the metrics for distance to the nearest location was particularly ineffective because of the model's assumptions that there would be no hard edge to the data as there was in this case. Instead the metric for number of locations within distance one served as more robust to this data flaw. The level of centrality of the data was also particularly important, as the more widespread pubs featured faster deviations from the model's expected metrics, as the pubs become more closer and seemingly reached a threshold maximum number of nearby pubs (referring to the first homogenous Poisson process on page 6).

To build upon this research, certain assumptions and models can be adjusted. Data from the outskirts of London can perhaps still feed into the metrics of distance to, and amount of, neighbors but do not have to be analyzed themselves. This would fix the edge problem. It would be interesting to see if a dynamic Matern's Process could be built into another process such that locations farther from the city center could be subject to more refractory behavior. Finally, a better likelihood representation should also be built to describe and compare these processes. This analysis used a sum of probabilities which could only then allow comparisons across models of the same dataset rather than the log of a product. The current method shown of model evaluation is quite poor in its interpretability and ability to provide effective comparisons. The assumption of no locations outside of the boundaries of the data and the evaluation methods are the biggest drawbacks to this assessment and prevent any serious model fitting to be performed and judged.

## Bibliography

Preda, Gabriel, *"Beer or Coffee in London - Tough Choice? No More!" Kaggle.com*, Kaggle, 19 Dec. 2018, www.kaggle.com/code/gpreda/beer-or-coffee-in-london-tough-choice-no-more/report

### *Datasets*

Starbucks, *"Starbucks Locations Worldwide." www.kaggle.com*, www.kaggle.com/datasets/starbucks/store-locations

Preda, Gabriel. *"GADM Data for UK." Kaggle.com*, 2018, www.kaggle.com/datasets/gpreda/gadm-data-for-uk

Tatman, Rachael. *"Every Pub in England." Kaggle.com*, 2017, www.kaggle.com/datasets/rtatman/every-pub-in-england