Starbucks in China – A Spatial Analysis with Socioeconomic & Demographic Data

Yuwei (Johnny) Meng

21 Dec 2024

Project GitHub Repository

Project Presentation

1 Introduction

As globalization continues, the presence of multinational corporations becomes increasingly prominent, among which Starbucks is one of the most successful companies that expand their business to many countries, including China. As of 31 December 2019, there are 4166 Starbucks stores in Mainland China, spanning over 168 cities and occupying a large portion of China's coffee market. Still, the number is growing today.

To understand the success of Starbucks in China, a great starting point is to analyze the spatial patterns in Starbucks stores in China. Therefore, given a dataset consisting of locations of Starbucks stores in Mainland China, this project aims to use techniques in spatial statistics to discover spatial patterns within the data. Socioeconomic and demographic factors such as GDP and population are included in the spatial models as covariates to improve model performance.

In addition, Shanghai, one of the most developed cities in China, has a remarkably large number of Starbucks stores compared to other cities. Hence, this study also delves into Shanghai and focuses on finding spatial patterns in locations of Starbucks stores in Shanghai.

Ultimately, this project aims to achieve the following objectives:

1. **Areal Analysis** – On a province scale, is there an autocorrelation in the number of Starbucks stores? Can we construct spatial models that include GDP and population data to model the number of Starbucks stores in a province?
2. **Point Pattern Analysis** – Can we build point process models to study the distribution of Starbucks stores in Shanghai?

By addressing the research questions above, we can paint a holistic picture of Starbucks' development in China.

2 Data

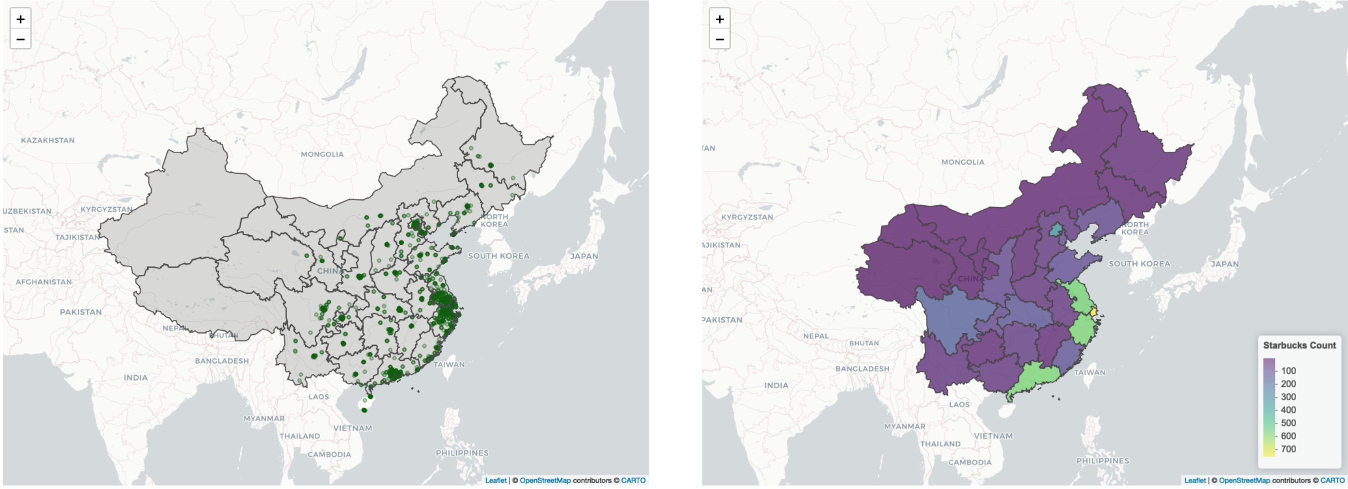
This study joins multiple datasets. A dataset containing locations of Starbucks stores in Mainland China, obtained from Kaggle, is the main subject of this project. This dataset gives the latitudes and longitudes of the 4166 Starbucks stores in Mainland China along with some features of each store, such as the business hours and artwork status. However, these features are of no interest to the study and thus were abandoned for the remaining analysis.

Regarding the spatial analysis, the shapefiles of China on the district scale, the city scale, and the province scale for conducting spatial computations are provided in a cited GitHub repository. For the analysis, I removed Hong

Kong, Macau, and Taiwan from the shapefiles because the Starbucks dataset does not contain store information in these regions. Furthermore, Hainan is an island province that does not spatially touch any other province and so I removed Hainan from the shapefiles as well. Overall, there are 28 provinces remaining for the analysis.

By merging the Starbucks dataset and the shapefile on the province scale, we can visualize the locations of Starbucks stores in Mainland China on a map, shown in *Figure 1a* below. From the map, we observe that most Starbucks stores are clustered in the eastern part of China. Particularly, Shanghai and its two neighboring provinces, Zhejiang and Jiangsu, seem to have a large cluster of Starbucks stores.

Upon conducting a spatial join, I grouped the Starbucks stores by province and counted the number of stores in each province. The grouped data is shown on the map in *Figure 1b*. From the map, we see that Shanghai, Zhejiang, Jiangsu, and Guangdong have the most Starbucks stores. Note that Xinjiang and Tibet, two provinces in western China, have no Starbucks stores.



(a) Starbucks Locations in Mainland China

(b) Number of Starbucks Stores by Province

Figure 1: Maps of Starbucks Stores in China

As mentioned in the introduction, GDP and population were incorporated into the analysis. The GDP dataset was pulled from Kaggle, originally containing GDP information by province from 1992 to 2020. Since the Starbucks dataset contains the store locations by the end of 2019, I only kept the GDP information in 2019 and removed the rest. *Table 1* below shows the 5 provinces with the highest GDP and the total GDP. Additionally, the population information was extracted from the National Bureau of Statistics of China regarding the 2020 population census. *Table 2* below shows the 5 provinces with the greatest population and the total population of China. Note that Guangdong, Jiangsu, and Henan had the greatest populations and meanwhile the highest GDPs.

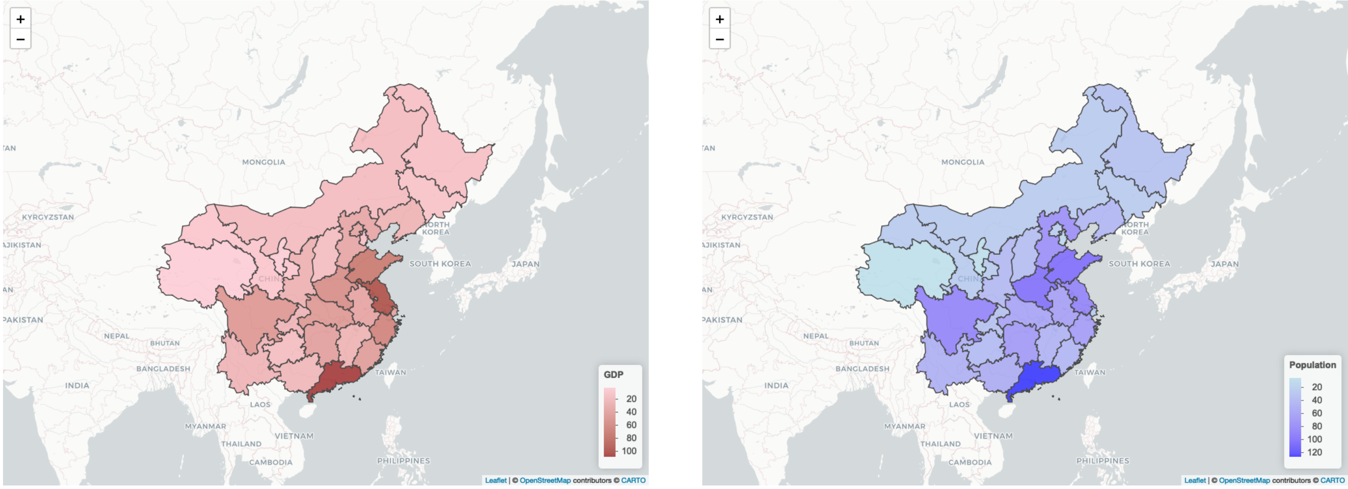
Province	**GDP (100 Billion CNY)**
Guangdong	108
Jiangsu	98.7
Shandong	70.5
Zhejiang	62.5
Henan	53.7
Total	**961.7**

Table 1: Top 5 Provinces in GDP

Province	**Population (Million)**
Guangdong	126
Shandong	102
Henan	99.4
Jiangsu	84.7
Sichuan	83.7
Total	**1411.8**

Table 2: Top 5 Provinces in Population

Figure 2 below contains the maps of GDP and population by province.



(a) GDP (100 Billion CNY) by Province

(b) Population (Million) by Province

Figure 2: Maps of GDP and Population in China

3 Methodology, Results & Discussion

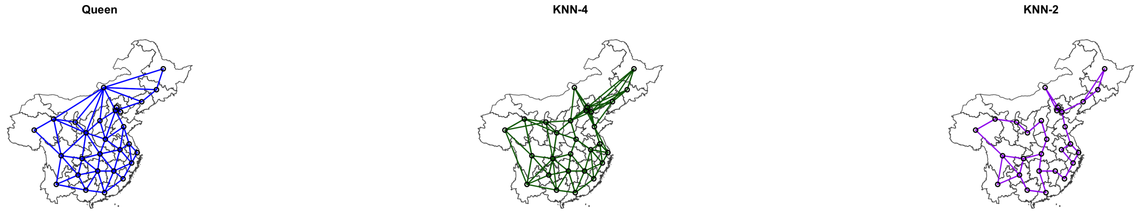
Since there are two parts to the overarching research question, this section will present the two parts separately.

3.1 Areal Analysis

> **Research Question** – On a province scale, is there an autocorrelation in the number of Starbucks stores? Can we construct spatial models that include GDP and population data to model the number of Starbucks stores in a province?

3.1.1 Neighborhood Matrices

To conduct an areal analysis and answer this research question, I first built several neighborhood matrices using the queen method and k-nearest neighbors. For kNN, I tried $k = 4$ and $k = 2$ and analyzed their performance.



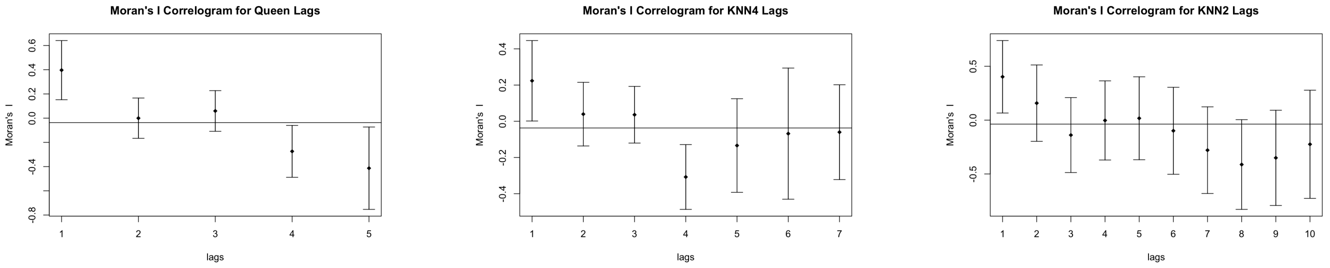
(a) Queen Neighborhood Matrix

(b) kNN-4 Neighborhood Matrix

(c) kNN-2 Neighborhood Matrix

Figure 3: Neighborhood Matrices

Nevertheless, from *Figure 3*, we notice that virtually all provinces in China are irregular in shape. Specifically, Inner Mongolia, the province in northern China with a very wide shape, spatially touches many other provinces and so the queen neighborhood matrix shows many connections with neighboring provinces for Inner Mongolia. Beijing, on the other hand, is surrounded by Hebei and Tianjin and so the queen neighborhood matrix shows only two connections for Beijing. Furthermore, Gansu, a province in northwestern China with a long and irregular shape, does not even have the centroid of the province in its territory. Thus, the practicality of kNN for computing neighbors might be questionable. To address these potential issues, I plotted the Moran's I correlograms for the three neighborhood matrices in *Figure 4* below, all using the row-standardized version, and assessed their usefulness.



(a) Queen Lags

(b) kNN-4 Lags

(c) kNN-2 Lags

Figure 4: Correlograms

In *Figure 4*, note that all three neighborhood matrices have a significant positive Moran's I only at lag 1. The kNN-4 neighborhood matrix has a significant negative Moran's I at lag 4, and the queen one has a significant negative Moran's I at lags 4 and 5. Conversely, the kNN-2 neighborhood matrix shows a more gradual and smooth change in Moran's I.

Considering that the three correlograms above all showed a significant positive Moran's I at lag 1, I also conducted some hypothesis tests to verify this result. *Table 3* below contains the test statistics and p-values for the two-sided Moran's I hypothesis tests on the three neighborhood matrices with a row-standardized weights list.

Neighborhood Matrix	**Moran's I Statistic**	**p-value**
Queen	0.397	0.0004
kNN-4	0.224	0.019
kNN-2	0.402	0.009

Table 3: Moran's I Tests under Normality

From *Table 3*, we observe that the tests for all three neighborhood matrices returned a statistically significant positive Moran's I statistic. This means that there is a positive autocorrelation in the number of Starbucks stores by province, or neighboring provinces have similar numbers of Starbucks stores. This result can be visualized in *Figure 1b* since Shanghai is surrounded by Zhejiang and Jiangsu which have a cluster of Starbucks stores, while provinces in northwestern China all have low numbers of Starbucks stores.

Since I could only plot the queen matrix up to lag 5 and the kNN-4 matrix up to lag 7 because there are only 28 provinces in the dataset, I decided to use the kNN-2 neighborhood matrix for the remaining analysis, considering its smooth change in Moran's I over the 10 lags and that its Moran's I statistic was the highest among the three neighborhood matrices.

3.1.2 Linear Models

The preliminary analysis of the neighborhood matrices above indicates that there is a positive autocorrelation in the number of Starbucks stores by province in China and so we should build autoregressive models to capture this pattern. By incorporating the GDP and population data as covariates, I first fitted some linear models to predict the number of Starbucks stores by province, which are summarized in *Table 4* below.

Model Specification	**Adj.** R^2	**AIC**
GDP Only	0.446	366.263
Population Only	0.067	380.845
GDP + Population	**0.715**	**348.508**
GDP × Population	0.711	349.822

Table 4: Fitted Linear Models

Coefficient	**Estimate**	p**-value**
Intercept	60.411	0.155
GDP	12.628	4.08×10^{-8}
Population	-7.090	3.17×10^{-5}

Table 5: Best Linear Model Coefficient Estimates

Noted in *Table 4*, the model with the highest adjusted R^2 and the lowest AIC value is specified by GDP + Population. The coefficient estimates for this model are summarized in *Table 5* above. Notice that both GDP and population are statistically significant predictors in the model. The positive GDP coefficient estimate indicates that a province with high GDP tends to also have a large number of Starbucks stores. On the contrary, population has an inverse correlation with the number of Starbucks stores in the province. I suspect that this relationship might be confounded by Shanghai because Shanghai has low population but many Starbucks stores. Overall, this model explains 71.5% of the variation in the number of Starbucks stores by province in China and thus it is a generally acceptable model.

With this linear model, I examined the spatial autocorrelation in the residuals by conducting a Moran's I hypothesis test. Interestingly, the test returns a Moran's I statistic of 0.009 and a p-value of **0.783**. This result indicates that there is no autocorrelation in the residuals after accounting for GDP and population, which suggests that GDP and population are capable of modeling the number of Starbucks stores and there is technically no autocorrelation in the number of Starbucks stores in the first place. This conclusion refutes the Moran's I tests given in *Table 3*.

To conclude this section of linear models, I was curious of whether there is any spatial autocorrelation in GDP or population so that the linear model above actually has spatial information in fitting the number of Starbucks stores. Therefore, I conducted two-sided Moran's I hypothesis tests on GDP and population using the kNN-2 and the queen row-standardized weights matrices. The results are summarized in *Table 6* below.

Covariate	**Moran's** I **Statistic**	p**-value**
GDP kNN-2	0.193	0.171
Population kNN-2	0.055	0.584
GDP Queen	0.258	0.016
Population Queen	0.166	0.097

Table 6: Moran's I Tests under Normality for Covariates

From the table, only the test on GDP with the queen matrix returned a p-value of 0.01 which is statistically significant, while the other tests had non-significant results. The evidence for this significant result is not very strong either. Therefore, we can conclude that there is only moderate evidence for a positive autocorrelation in GDP and no evidence for autocorrelation in population.

3.1.3 Areal Models

Building upon linear models, I fitted several spatial models for the number of Starbucks stores in China, incorporating GDP and population as covariates, and examined their performance compared to the linear models. *Table 7* below is a summary of the fitted areal models.

Model	**Estimate for λ**	**p-value for λ**	**Estimate for ρ**	**p-value for ρ**	**AIC**
SAR Error	0.029	0.931	NA	NA	350.5
SAR Lag	NA	NA	0.183	0.136	348.28
SAR Lag-Error	-0.032	0.897	0.184	0.171	350.27
CAR	0.028	0.952	NA	NA	350.5

Table 7: Fitted Areal Models

Below is a description for each of the areal models:

1. SAR Error – This model assumes that there is an autocorrelation in the residuals or errors after fitting the linear model with the covariates. Specifically, the model is specified by

$$Y = X\beta + \varepsilon,$$

where

$$\varepsilon = \lambda W \varepsilon + \nu.$$

Here λ measures the strength and direction of the spatial autocorrelation in the errors. From *Table 7*, we see that the fitted SAR Error model estimated $\lambda = 0.029$ with a p-value of 0.931, which is extremely non-significant. Therefore, we conclude that there is no spatial autocorrelation in the errors.

2. SAR Lag – Instead of assuming autocorrelation in the errors, this model assumes that there is a spatial component in the response variable Y. Mathematically, the model is specified by

$$Y = X\beta + \rho WY + \varepsilon,$$

where ρ is the spatial autoregressive parameter in Y. From *Table 7*, the fitted model estimated $\rho = 0.183$ with a p-value of 0.136, which is also non-significant. Hence, we conclude that there is no spatial component in the number of Starbucks stores by province in China.

3. SAR Lag-Error – Combining the previous two models, this model assumes that there is an autocorrelation in the errors and a spatial component in the response variable at the same time. Hence the model is represented by

$$Y = X\beta + \rho WY + \varepsilon,$$

where

$$\varepsilon = \lambda W \varepsilon + \nu.$$

From *Table 7*, we notice that neither λ nor ρ is statistically significant in the model, corroborating the results above.

4. CAR – In the SAR models, we assume that the numbers of Starbucks stores in different provinces form a joint distribution. In the CAR model, we assume that there is spatial dependence of a province on all other provinces. This dependence is represented mathematically by

$$Y_i | Y_{-i} \sim \mathcal{N}\left(\sum_{j \in N(i)} \lambda_{ij} Y_j, \sigma_i^2 \right),$$

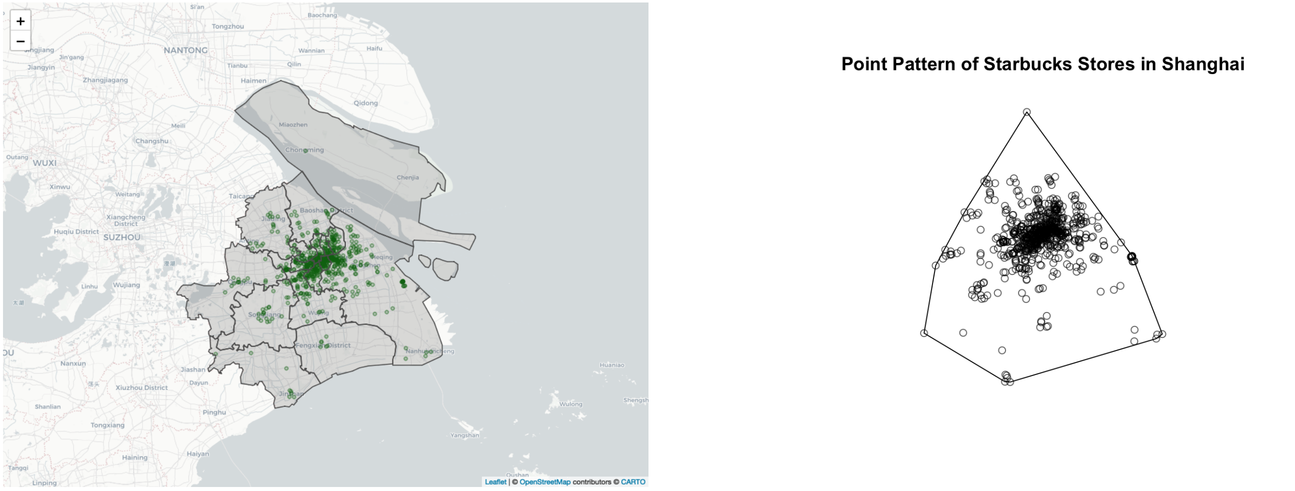
where λ_{ij} is the autoregressive term that measures the extent to which Y_i depends on its neighbors. In the fitted CAR model, the estimate for λ is 0.028 with a p-value of 0.952, which is non-significant. Hence, we conclude that there is no spatial dependence in the number of Starbucks stores in a province on other provinces.

In the previous subsection on linear models, I mentioned that GDP and population well capture the patterns in the number of Starbucks stores by province. The analysis in this subsection confirms this finding that any spatial model yields non-significant spatial results. By looking at the AIC values in *Table 7*, we can also observe that the areal models perform comparably or worse than the best linear model. Therefore, we can confidently conclude that there is no spatial autocorrelation in the number of Starbucks stores by province in China. GDP and population, or more generally, the development of a province, largely determine the number of Starbucks stores in that province.

3.2 Point Pattern Analysis

> **Research Question** – Can we build point process models to study the distribution of Starbucks stores in Shanghai?

In this section, we focus on Shanghai specifically. As per the Starbucks dataset, there are 754 Starbucks stores in total in Shanghai by the end of 2019. *Figure 5a* below is a map of the Starbucks stores in Shanghai.


Point Pattern of Starbucks Stores in Shanghai
Leaflet | © OpenStreetMap contributors © CARTO

(a) Map of Starbucks Stores in Shanghai (b) Point Pattern of Starbucks Stores in Shanghai

Figure 5: Starbucks Stores in Shanghai

The map shows that most Starbucks stores are clustered in the central part of Shanghai. The farther from the city center, the fewer Starbucks stores. *Figure 5b* on the right shows the enclosing window of the Starbucks stores as point pattern and confirms that the stores are clustered at the center of the window.

3.2.1 Testing for Complete Spatial Randomness (CSR)

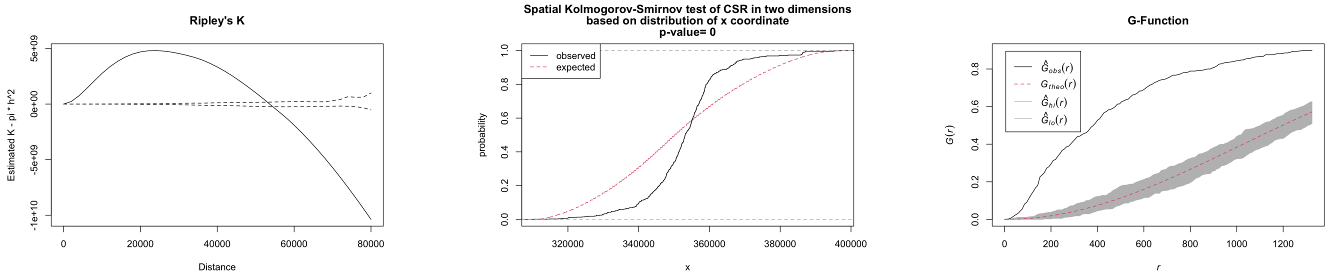
To study point patterns, the exploratory step is to test whether the points are distributed in a completely random fashion across the study area. *Figure 6* below shows three types of exploratory CSR testing methods. A brief description of each method is written below:

1. Ripley's K – Ripley's K tests for CSR by considering the expected number of points within a distance r of a point and normalizing the count by the intensity. Mathematically,

$$K(r) = \frac{1}{\lambda}\mathbb{E}[N(r)].$$

From *Figure 6a*, we notice that the empirical Ripley's K computed from the Starbucks data does not fall into the theoretical envelope at all, indicating that there is very unlikely CSR in the locations of Starbucks stores in Shanghai.

2. Spatial K-S Test – The K-S test assesses the goodness-of-fit of how well the distribution of the point pattern of Starbucks stores in Shanghai fits into a theoretical distribution of CSR. Methodologically, we can plot the theoretical



(a) Ripley's K (b) Spatial Kolmogorov-Smirnov Test (c) G-Function

Figure 6: Tests for CSR

cumulative distribution function (CDF) and the empirical distribution function (EDF) together, like in *Figure 6b*. Then the K-S test statistic D is computed as

$$D = \max |F_{\text{obs}}(z) - F_{\text{exp}}(z)|.$$

Here, the K-S test returned a p-value of 0, indicating that there is unlikely CSR in the locations of Starbucks stores in Shanghai.

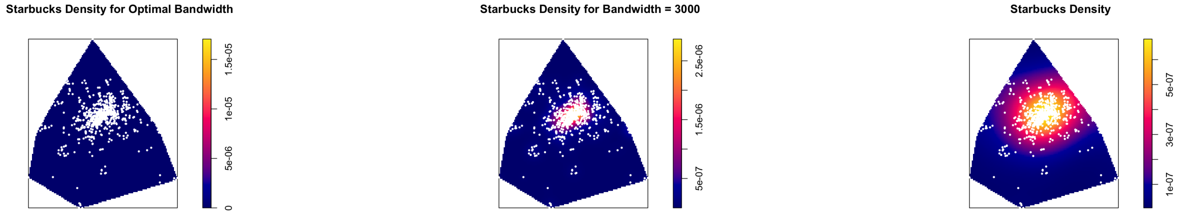
3. G-Function – The G-function measures the cumulative distribution function for the probability that the nearest neighbor distance is less than r. Namely,

$$G(r) = \mathbb{P}(D_i \leq r).$$

In *Figure 6c*, note that the empirical G-function is above the theoretical envelope for CSR. This result suggests that there is likely a clustering pattern in the locations of Starbucks stores in Shanghai.

3.2.2 Kernel Density & Dirichlet Tesselation

In the previous subsection, the three CSR tests all came to the conclusion that there is unlikely CSR in the data. An alternative way to visualize the distribution of Starbucks stores in Shanghai is to examine the density function of the distribution. In *Figure 7* below, three density functions are plotted using different bandwidths and methods.


Starbucks Density for Optimal Bandwidth
1.5e-05
1e-05
5e-06
0
Starbucks Density for Bandwidth = 3000
2.5e-06
1.5e-06
5e-07
Starbucks Density
5e-07
3e-07
1e-07

(a) Density with Optimal Bandwidth (b) Density with Bandwidth = 3000 (c) Default Density

Figure 7: Density Functions

In *Figure 7a*, the density function was plotted using the `bw.diggle` function in R. This function uses cross-validation to find the optimal bandwidth for the density. In this case, the optimal bandwidth is estimated to be 340.062, indicating that there is spatial variation at a scale of approximately 340 meters, which is a localized pattern. This result is seen on the plot that there is no global spatial pattern in the distribution of Starbucks stores and thus this bandwidth is not a good estimate.

Subsequently, I tried a bandwidth of 3000 and plotted the density in *Figure 7b*. In this plot, we start to see high density values centered at the cluster of Starbucks stores. As we get farther from the center, we see fewer Starbucks stores and thus would expect that density decreases.

Lastly, *Figure 7c* was plotted using the `density` function in R without any parameter. In this figure, we observe a larger region of high density at the center of the cluster. Similar as above, the farther from the cluster, the smaller the density.

In addition to density, *Figure 8* below shows the Dirichlet tesselation for the Starbucks stores. Agreeing with the density functions above, the tesselation creates very small regions at the cluster of Starbucks stores at the center of the window, and large regions as we get farther from the cluster. This confirms that there is a cluster of Starbucks stores at the center, and fewer Starbucks stores in the outer regions of Shanghai.

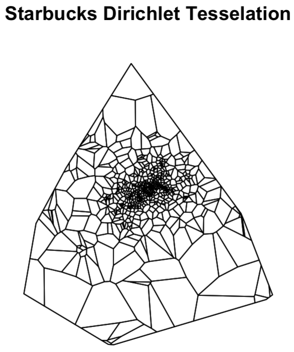


Figure 8: Dirichlet Tesselation of Starbucks Stores

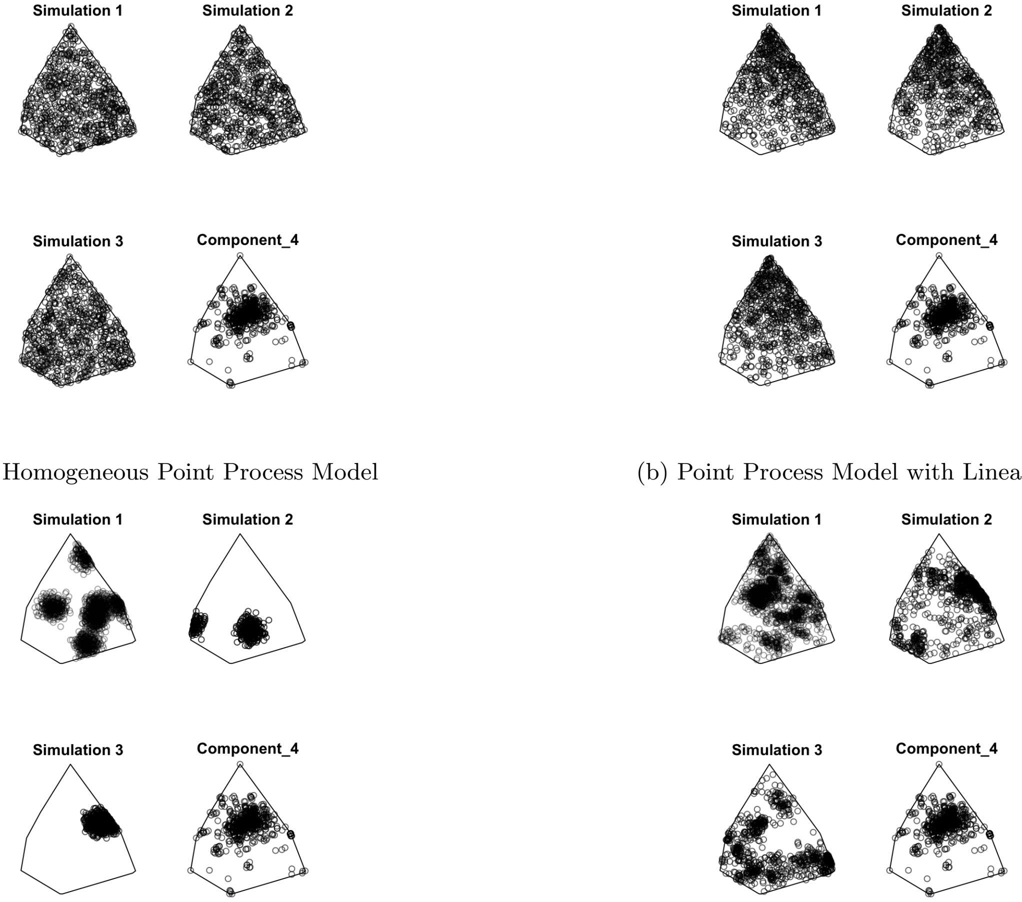
3.2.3 Point Process Models

To fully understand the distribution of Starbucks stores in Shanghai, we need to build point process models to capture the point pattern. For this study, I fitted 4 point process models and using each model, I simulated 3 point processes in the study area to examine the extent to which the model fits the point process of Starbucks stores. *Figure 9* below shows the simulations of point processes after fitting the models. Here is the description and analysis of each point process model:

1. Homogeneous Point Process Model – This model assumes that the training point process is a homogeneous Poisson point process and thus the only parameter the model fits is the intensity parameter of the Poisson distribution. In this case, the point process model estimates that the intensity $\lambda = 1.540 \times 10^{-7}$. However, in the previous subsections, we already established that the point process of Starbucks stores in Shanghai is not a homogeneous process. Therefore, this model does not fit the point process well. We also reach this conclusion in *Figure 9a* that there is obviously no clustering happening in the simulations.

2. Point Process Model with Linear Trend – Instead of a homogeneous Poisson point process, this model assumes that there is a linear trend in the point process of the data. For this data, the model estimates that $\log(x)$ is associated with a coefficient of 3.254 and $\log(y)$ is associated with a coefficient of 89.217. Both coefficient estimates are statistically significant. This means that the farther to the north or the east, we would expect to see more Starbucks stores. Nevertheless, from *Figure 9b* we also decide that this model does not capture the pattern well because this linear trend is not definitive in the original dataset.

3. Cluster Process Models – *Figures 9c* and *9d* demonstrate simulations using two cluster models, fitted with



(a) Homogeneous Point Process Model

(b) Point Process Model with Linear Trend

(c) Thomas Cluster Process Model

(d) Log-Gaussian Cox Process (LGCP) Model

Figure 9: Point Process Models Simulations

different methods. The advantage of these models is that they assume a clustering process in the data, which is the case in the Starbucks data. Among these two models, I believe the LGCP model fits the Starbucks point process better than the Thomas model. For the Thomas model in *Figure 9c*, we see no Starbucks stores outside of the clusters defined by the model, which is not the case in the original data. Contrarily, the LGCP model correctly captures the clustering nature of Starbucks stores while allowing for non-clustered Starbucks stores to be present.

4 Conclusion

In short, this project consists of two types of spatial analysis – areal analysis and point pattern analysis – in order to discover and model the spatial patterns in the locations of Starbucks stores in Mainland China.

In the areal analysis, we were interested in finding an autocorrelation in the number of Starbucks stores by province, that is, whether neighboring provinces have similar numbers of Starbucks stores. Exploratory Moran's I hypothesis tests revealed that there is a significant positive autocorrelation in the number of Starbucks stores and so we proceeded with fitting areal models. However, we discovered that GDP and population of a province are capable of explaining 71.5% of the variability in the number of Starbucks stores, leaving the residuals spatially uncorrelated. This result was further confirmed by the statistically non-significant estimates of the spatial parameters in the spatial models. Overall, we concluded that GDP and population of a province greatly determine the number of Starbucks stores in a province.

In the point pattern analysis, we were curious of finding point patterns in the locations of Starbucks stores in Shanghai. Basic tests of CSR returned that the locations of Starbucks stores are not randomly distributed, and so we fitted kernel density estimations, Dirichlet tesselation, and point process models to capture the spatial patterns. Conclusively, we determined that there is a cluster of Starbucks stores at the center of Shanghai. As we move farther from the city center, there are fewer Starbucks stores.

4.1 Limitations

There are several limitations to this project. Firstly, as I read in and examined the dataset containing the locations of Starbucks stores in areas that I am familiar with, I figured out that the latitudes and longitudes of the stores given in the dataset are not perfectly accurate. This inaccuracy might cause potential issues to the spatial computations, which can lead to unexpected outcomes. Therefore, a systematic review and correction of the latitudes and longitudes should be carried out.

Secondly, the point process analysis was only conducted using Starbucks stores in Shanghai specifically. Many other cities, like Beijing and Guangzhou, also have many Starbucks stores and thus are worth delving into. Future studies should focus on more cities so that a more complete picture of Starbucks' development in China can be painted.

5 Acknowledgements

I would like to express my sincere gratitude to Professor Meredith Franklin in the Department of Statistical Sciences at the University of Toronto for her valuable suggestions and comments towards completing this project.

6 References

- Starbucks Stores in China Mainland – Kaggle
- China's GDP in Province – Kaggle
- CTA Map – GitHub
- Communiqué of the Seventh National Population Census (No. 3) – National Bureau of Statistics of China