Starbucks: Spatial Analysis and Classification of the Different Scale

Jaceline Preval

Abstract

In this project, I analyzed Starbucks locations in the US. Based on some key variables, can I predict the scale of numbers of Starbucks in a zip code and are the locations of those stores random. Two datasets were collected for this project. The first dataset was collected from Kaggle and include every Starbucks store locations as of February 2017 [1]. The second data was collected from the United States Census Bureau that includes more than X samples [2]. I narrowed down the scope to Texas and approach the problem with two methodologies. First, I performed a chi-square test to test complete spatial randomness. Second, I implemented a logistic regression to predict the different scale of number of stores per zip code in Texas.

I. Introduction

1.1 Background

When the first Starbucks store opened in 1971 in Seattle historic Pike Place Market, the vision for the future was to offer the finest fresh roasted whole bean coffees. Now with more than 21,00 stores in 54 countries, Starbucks is the premier roaster and retailer of specialty coffee in the world. As a coffee lover, I wanted to predict how many Starbucks locations will be in my zip code if I moved to Texas.

Before going further, I plotted a heat map to display the number Starbucks stores in the US. The difference among the states could be due to population density, income, social and economic class. In the following study, I focused on Texas *(n=983)* even though California *(n=2,759)* has the most number of study since it the largest state [Fig.1].

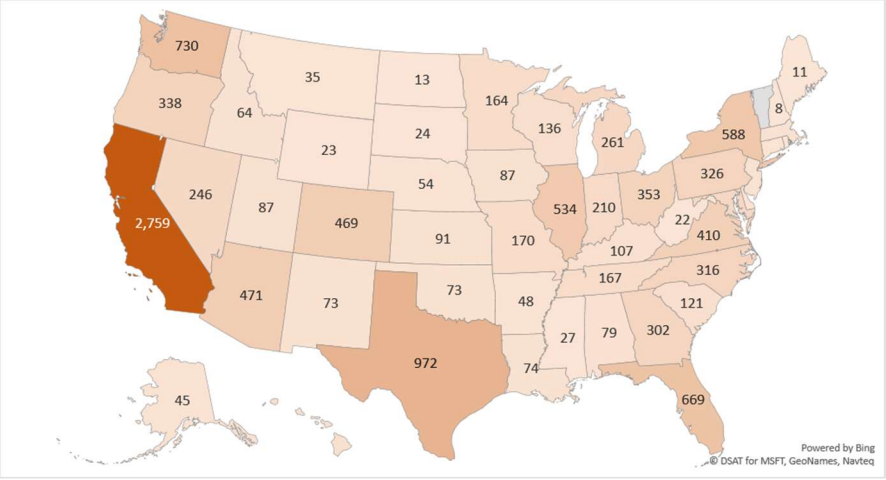


Fig.1 The US map shows the number of Starbucks stores *(Darkest: 2,759 stores; lightest: 8 stores)*

1.2 Pre-Processing

The first dataset from Kaggle contained:

1. **Brand:** The name of each brand under the umbrella of Starbucks i.e. Starbucks, Teavana, Evolution Fresh.
2. **Ownership Type:** The type of ownership of each store i.e. Company Owned, Licensed, Joint Venture or Franchise.
3. **City**
4. **State**
5. **Zip code**
6. **Longitude**
7. **Latitude**

I filtered this data set to only reflect Starbucks Brand in Texas. I also created two new variables (X, Y) which represent the cartesian coordinates of the longitude and latitude. The second dataset included those fields after I manipulated the data:

1. **Zip Code**
2. **Median Household Income:** Median household income at the zip code level for 2010.
3. **Percent of Number of Household with Income less than or equal $50,000:** Based on income and number of household at the zip code level
4. **Percent of Number of Household with Income more than $50,000:** Based on income and number of household at the zip code level
5. **Population:** The total population number in the zip code in 2010.
6. **Price Increase of Median House Prices between 2000 and 2010**

The two datasets were then joined by zip code to establish two new datasets:

A. ***Dataset A*** *(n=783)* **-** This data set included: City, Median Household Income, Percent of household with income 50k or less, Percent of household with more 50k, Population, Median House Price Increase. I also added a new field which categorized the numbers of Starbucks stores in the zip code as Low: 0-1; Medium: 2-4; High: > 4.

B. ***Dataset B*** *(n=422)* - The second dataset only contained the x and y cartesian coordinates for each store locations in Texas.

1.3 Exploratory Data Analysis

With Texas having 983 Starbucks, I mapped all the locations to get a sense of where they were all located.

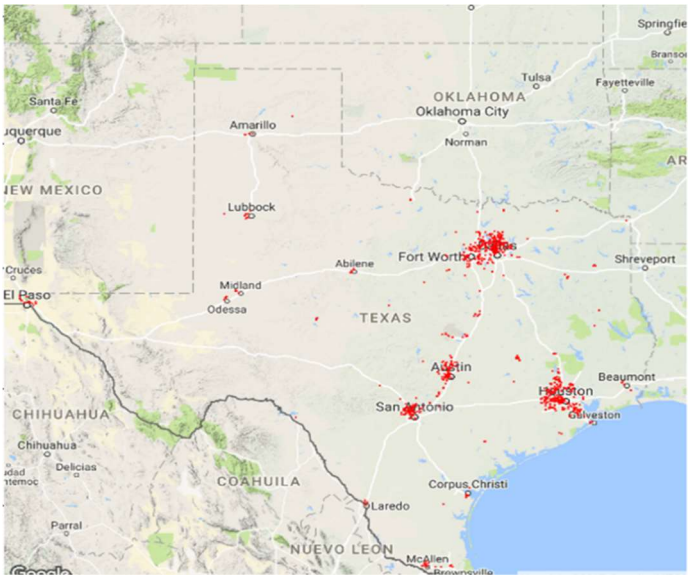


Fig.2 Map showing in what cities are the most numbers of Starbucks stores.

Looking at the map, it was evident that the number of stores were concentrated in Dallas and Houston. The top 5 cities by number of stores are Houston: 151, San Antonio: 75, Dallas: 74, Austin: 58 and Fort Worth: 29. I then tried to understand and analyze which of those variables were driving the numbers of stores in those top cities. First, I averaged the features in my dataset by cities and looked to see that which features sorting descending will have most of my top 5 cities. This strategy failed to show meaningful insights. Since multiple zip codes had the same city, I took the maximum value for each of the cities in my dataset. Surprisingly looking at the maximum percentage of household with income less than or equal to 50,000, Dallas, San Antonio and Houston were in my top 5 results with 93.0%, 91.9% and 91.8% respectively. However, when I looked at the maximum value of the increase median house price between 2000 and 2010, my top 5 results were Dallas, Houston. Austin, Bellaire and San Antonio. This reaffirmed my belief that the housing market was driving the numbers of Starbucks stores. As shown in Fig.3, the average median house price was the highest where the number of store was also the highest.

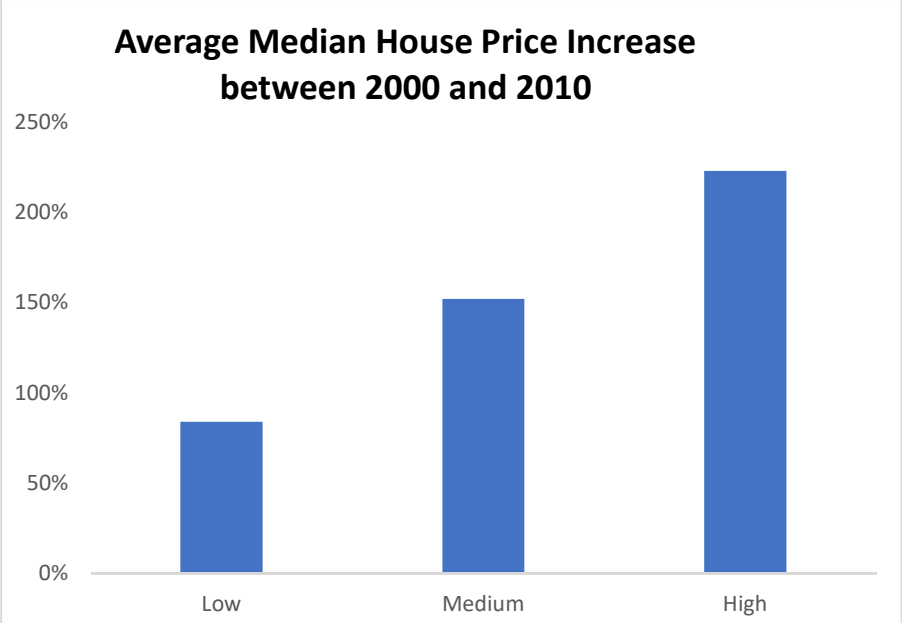


Fig.3 Graph showing as the higher the increase in price, the higher the total number of stores.

II. Complete Spatial Randomness

2.1 Method

Complete Spatial Randomness [4] is a standard model and states that the events follow a homogeneous Poisson Process over the study region. In this model, point pattern is the number of events occurring in arbitrary sub-regions or areas, A, of the whole study region R.

Spatial point process is defined by:

$$\{Y(A), A \subseteq R\} \text{ Where}$$

where *Y(A)* is the number of events occurring in the area A.

A hypothesis of complete spatial randomness for a spatial point pattern $\{Y(A), A \subseteq R\}$ asserts that:

- The number of events in any planar region with area A follows a Poisson distribution with mean λA.

$$f_{Y(A)}(y) = \frac{(\lambda A)^y}{y!} e^{-\lambda}$$

- Given n events in A, the events are an independent random sample from a uniform distribution on A

- implies constant intensity – no first order effects
- implies no spatial interaction

This means that any event has an equal probability of occurring at any position in R. Furthermore, the position of any event is independent of the position of any other, i.e. events do not interact with one another. Therefore, by simulating n events from such a process by enclosing R in a rectangle, i.e. generating events with x coordinates from a uniform distribution on (x_1, x_2) and y coordinates from a uniform distribution on (y_1, y_2) the observed pattern of points can be compared with the simulated ones based on CSR. i.e. CSR represents a baseline hypothesis against which to assess whether observed patterns are regular, clustered or random.

We can conduct statistical tests for significant patterns in our data, with this hypothesis:

- Ho: events exhibit complete spatial randomness (CSR)
- Ha: events are spatially clustered or dispersed

There are at least three approaches to testing the CSR hypothesis: the quadrat method, the nearest-neighbor method, and the method of K-functions. However, for this I will only use the quadrat method. Since the CSR Hypothesis also implies that each of the cell counts, $N_i = N(C_i), i = 1, \dots, k,$ is independent, it follows that $(N_i: i = 1, \dots, k)$ must be an independent random sample from this Poisson distribution. Hence the simplest test of this hypothesis is to use the Pearson $\mathcal{X}^2$ goodness-of-fit test. Hence the observed value of N_i is denoted by n_i , then the chi-square statistic

$$\mathcal{X}^2 = \sum_{i=1}^{m} \frac{(n_i - \bar{n})^2}{\bar{n}} = (m-1)\frac{s^2}{\bar{n}}$$

where $s^2 = \frac{1}{(m-1)} \sum_{i=1}^{m} (n_i - \bar{n})^2$ is the sample variance.

2.2 Results & Interpretation

I used R to implement a quadrat method to test for CRS. I first plotted the x and y cartesian coordinates from *dataset B.* At first glance, Fig.4 showed that there was concentration of Starbucks stores located in the middle. Let's apply the test to see if this was indeed true.

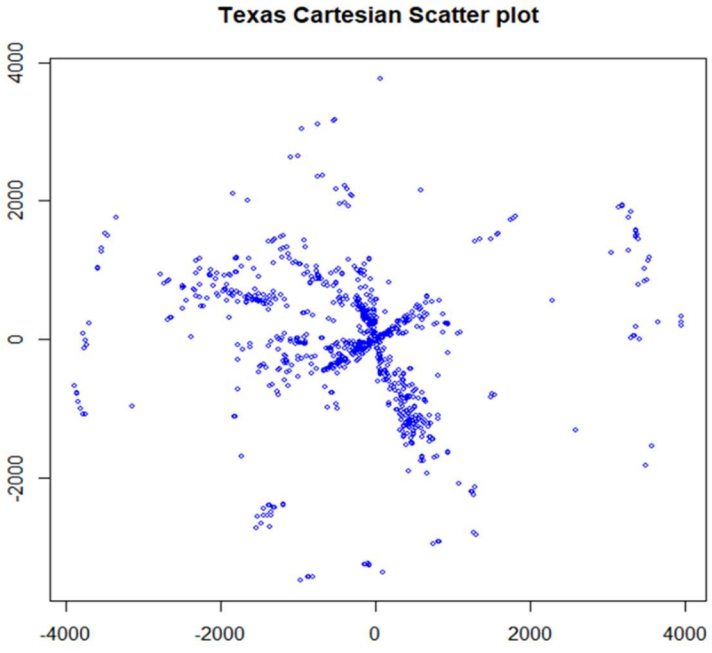


Fig.4 Scatterplot of the cartesian coordinates in Texas.

For the first attempt, I divided the plot into 2 by 2 quadrants.

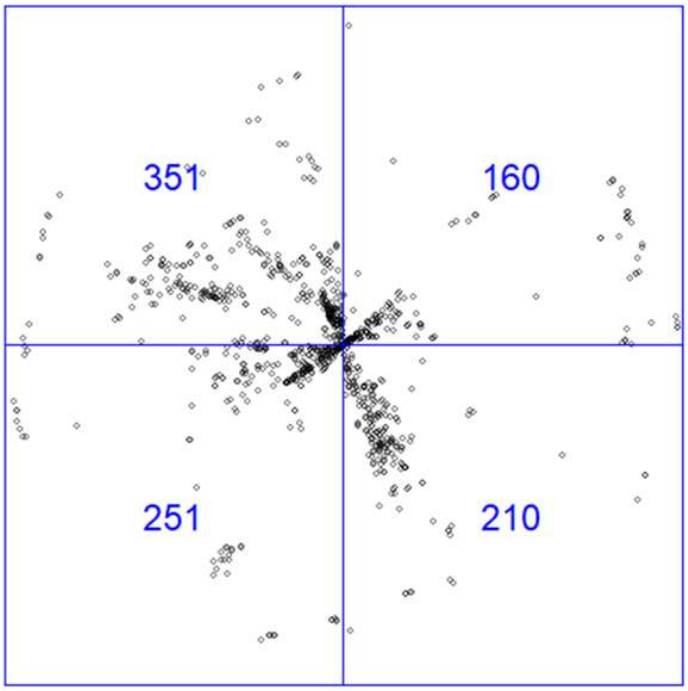
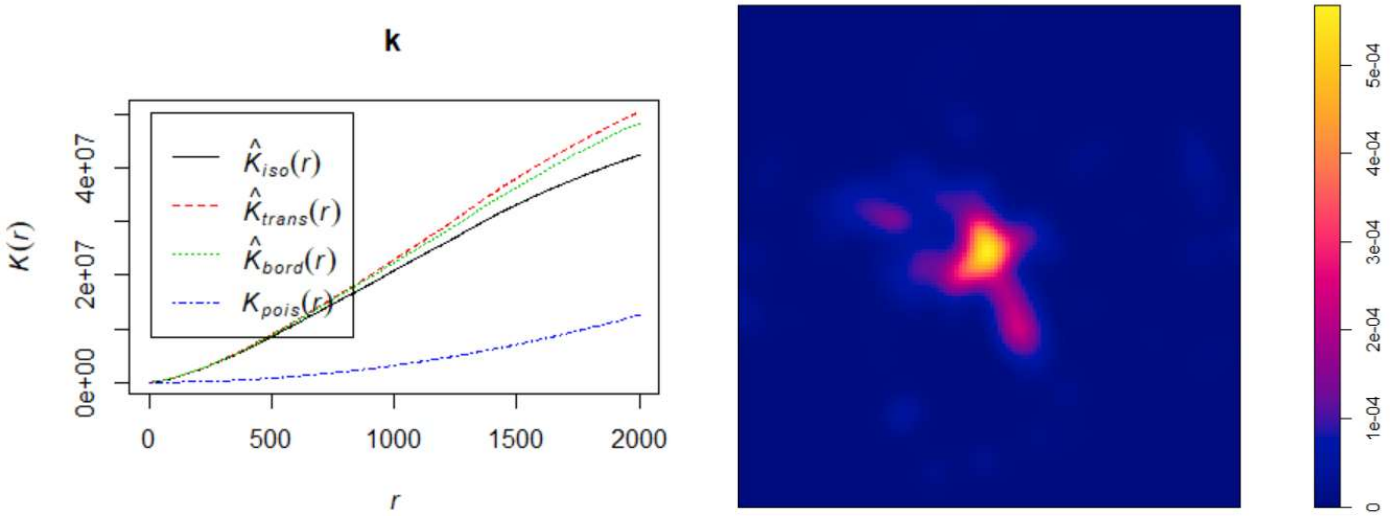


Fig.5 Quadrant count of Starbucks locations in Texas.

After performing the chi-square test in R using the quadrant counts, I got $\wp$ value of $2.20e^{-16}$ with a degree of freedom of 3 and $\mathcal{X}^2$ equals to 81.095. Since the $\wp$ value is less the significance level, we cannot accept the null hypothesis. Thus, we conclude this data set was not generated under CSR.

Since I have concluded that the locations are not random, let's plot the kernel smoothing and second order measure.



III. Multinomial Logit Model

3.1 Method

Consider a random variable Y_i that may take one of several discrete values, which we index 1, 2, . . ., J. Let π_{ij}= Pr{Y_i = j} denote the probability that the i-th response falls in the j-th category. Assuming that the response categories are mutually exclusive and exhaustive, we have

$\sum_{j=1}^{J} \pi_{ij} = 1$ for each i, i.e. the probabilities add up to one for each individual, and we have only $J - 1$ parameters. The probability distribution of the counts Y_{ij} given the total n_i is given by the multinomial distribution

$$\Pr\{Y_{i1} = y_{i1}, \ldots, Y_{iJ} = y_{iJ}\} = \binom{n_i}{y_{i1}, \ldots, y_{iJ}} \pi_{i1}^{yi1} \ldots \pi_{iJ}^{yiJ}$$

In the multinomial logit model, we assume that the log-odds of each response follow a linear model

$$\eta_{ij} = log \frac{\pi_{ij}}{\pi_{iJ}} = \alpha_j + x_i' \beta_j$$

where α_j is a constant and β_j is a vector of regression coefficient for $j = 1, 2, \ldots, J - 1$.

3.2 Results

Based on household income, demographics and housing market information, I want to predict the scale of Starbucks in a zip code. I added an L2 regularization parameter and tuned it using cross validation. After training a multinomial in python, I have analyzed my results to see if I was successful. I have created a confusion matrix to better see my results. My accuracy is 90.99%.

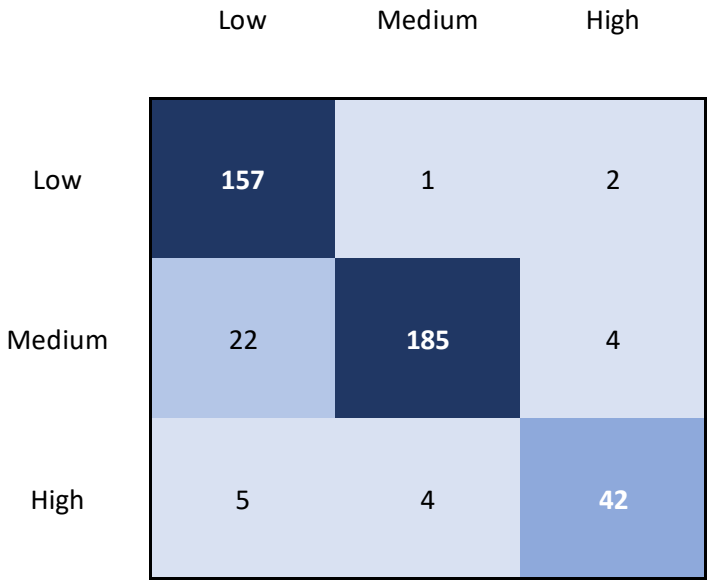


Fig. 6 Confusion Matrix based on multinomial logit model

References

[1] https://www.kaggle.com/starbucks/store-locations

[2] https://www.census.gov

[3] https://www.thoughtco.com/us-states-by-area-1435125

[4] Grabarnik, P., and S. N. Chiu. "Goodness-of-fit test for complete spatial randomness against mixtures of regular and clustered spatial point processes." Biometrika 89.2 (2002): 411-421.