GEOG 318 Winter 2017 Assignment 2

Spatial Distribution Analysis

THE HOME DEPOT

ONTARIO, CANADA

Understanding site and surroundings of the Home Depot stores.

Jiachen Wei 20695056

University of Waterloo

Wednesday, 1st February 2017

I. Introduction

The Home Depot requested an analysis of the spatial distribution of their stores across Ontario, considering the site and situation characteristics around the current store locations. Home Depot provided a shape file of the locations of their existing stores in Ontario, and nearest distance calculations from Geospatial Modelling Environment (GME). This report summarizes the point pattern statistics, quadrat counts and probability distributions, among maps and other deliverables generated by RStudio with assistance from ArcGIS and Excel.

II. Results

First, the point pattern statistics are summarized using R (Table 1). The mean coordinate (X and Y) of the Home Depot stores in Ontario is (1371645.99, 11959567.33), which is in the south-eastern coner of Ontario (the red cross in Figure 1). The squared residuals indicate that distances from the mean coordinate reach up to hundreds of kilometers, and that the points are closer vertically (along Y axis) than horizontally.

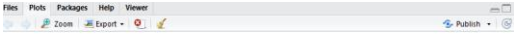


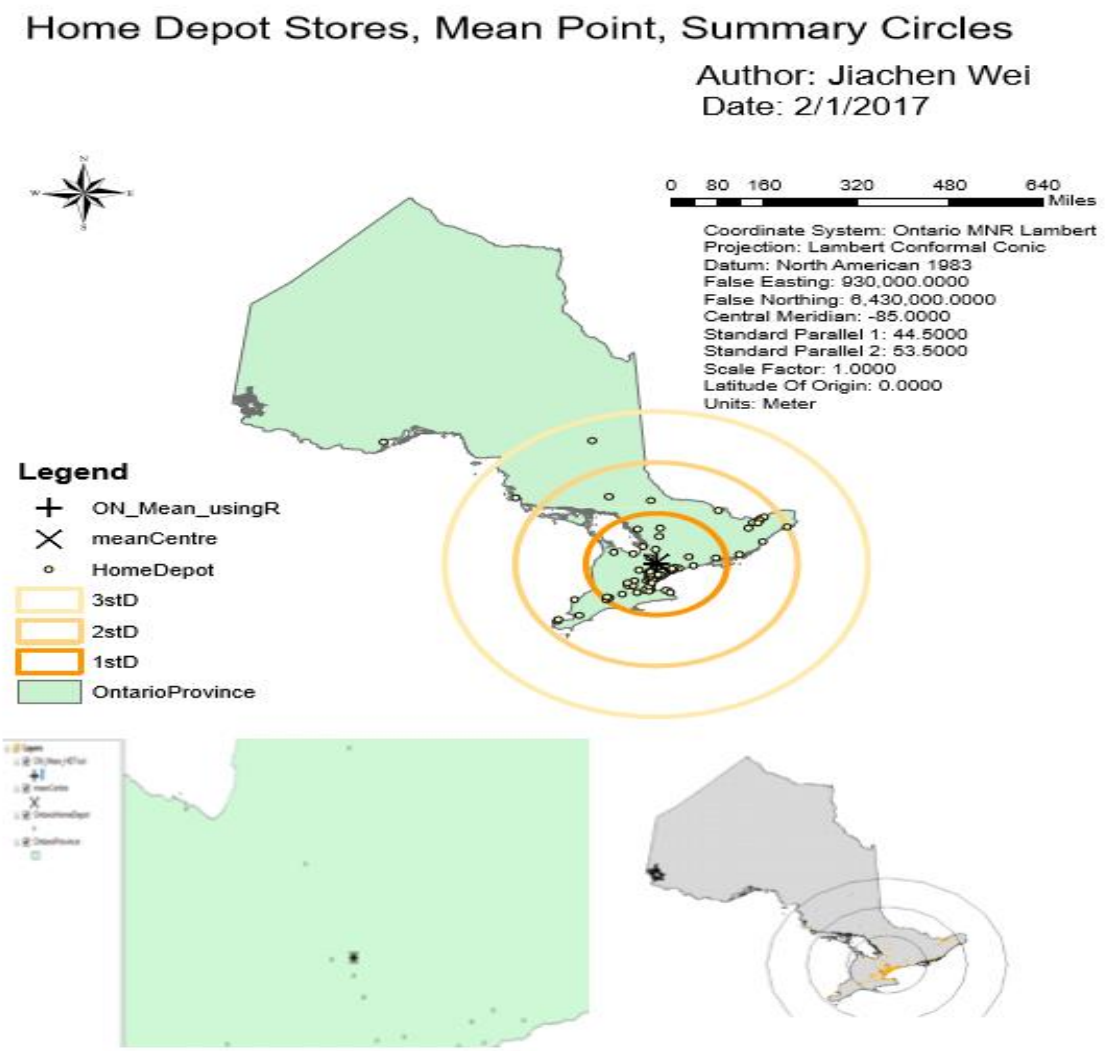
Table 1: Point Pattern Summary Statistics

Mean X	1371645.99
Mean Y	11959567.33
Sum of Squared X Residuals	2.052145e+12
Sum of Squared Y Residuals	1.350904e+12
Standard Distance	195541.59

Figure 1 Home Depot Stores in Ontario

The standard distance (stD) is related to the concept of standard diviation in basic statistics. We plot in R circles with radius set as 1 stD, 2 stD, and 3stD; the result shows how likely the stores are to fall in different areas of Ontario. To verify the result in R, we visualize the mean point and the three circles for in ArcGIS; the results match (Figure 2). Unfortunately, there was no way to zoom in and out in RStudio. Table 2 shows the probability of stores falling in each circle, and the corresponding standard deviation values.

Figure 2 Result Map



	DISTANCE (M)	# OF FEATURES	% OF ALL FEATURES	STANDARD DEVIATION VALUES %
1 STANDARD DISTANCE	195541.59	68	76.4	68
2 STANDARD DISTANCE	391083.18	85	95.51	95
3 STANDARD DISTANCE	586624.77	88	98.88	99.7
TOTAL	N/A	89	100	N/A

Table 2: Variance from the mean information

Next, we perform quadrat count, and compare the derived variance-to-mean value (VMR) to the Poisson distribution. We create 50km*50km quadrats from minimum X to maximum X, and minimum Y to maximum Y in RStudio; we erase the quadrats out of Ontario by creating a subset in RStudio and poltting it (Figure 3).

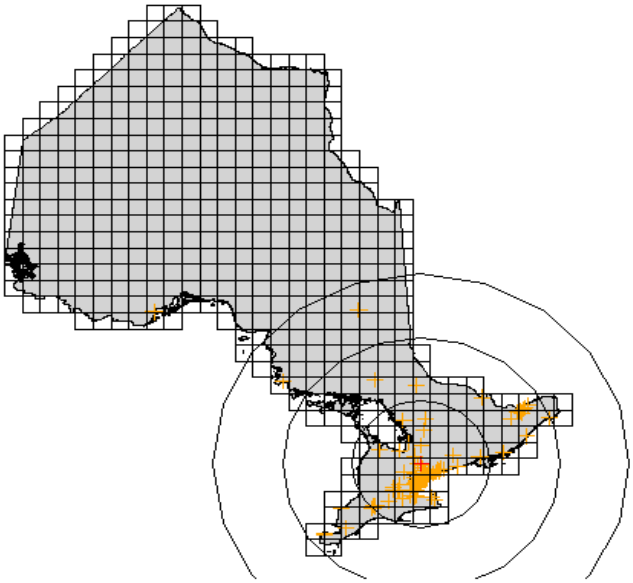


Figure 3 Quadrats

The quadrat count allows us to analyze the locations of Home Depot stores based on VMR. For Poisson distribution, the VMR value is 1.0. In other words, the VMR is expected to be 1.0 if the events occur with a fixed interval of space (or time), or randomly over space. We calculate VMR through dividing the variance $s^2 = \frac{\sum X(K-\mu)^2}{(m-1)}$ by the mean number of events in one quadrat (μ). A VMR greater than 1.0 indicates a tendency towards dispersion, while the VMR for our data (0.563450419) suggests a degree of clustering.

Table 3 Quadrat Count and Calculation of the Variance for Store Locations

No. of events, K	No. of quadrats, X	μ	K - μ	$(K-\mu)^2$	$X(K-\mu)^2$
0	447	0	0	0	0
1	23	0.043478261	0.956521739	0.914933837	21.04347826
2	8	0.25	1.75	3.0625	24.5
3	2	1.5	1.5	2.25	4.5
4	1	4	0	0	0
6	1	6	0	0	0
10	1	10	0	0	0
24	1	24	0	0	0
Totals	484	N/A	N/A	N/A	50.04347826

Finally, we employed distance-based approaches. We use the nearest distance result from Geospatial Modelling Environment (GME). In order to plot the G function (or the refined nearest neighbor), we first generate the cumulative number of events in Excel. We only need one value for events at the same distance, and the automated way to do this is adding a module from Microsoft Visual Basic for Applications. Due to limited time, we highlight the clums with duplicate values and manually delete to unwanted records in

the numCumuEvent column from 1 to 89. G(d) indicates the how likely a store is located within the corresponding distance (in ascending order) to the mean point. It is calculated by distance divided by the numCumuEvent value, and the result is shown in Figure 4. Over 84% of the points are between 0 to 40000 meters away from their nearest neighbors, so we plotted another graph for these points to see the G function in detail. The rapid increase from 5000m to 10000m suggests that a large number of stores are at this distance range to their nearest neighbors.

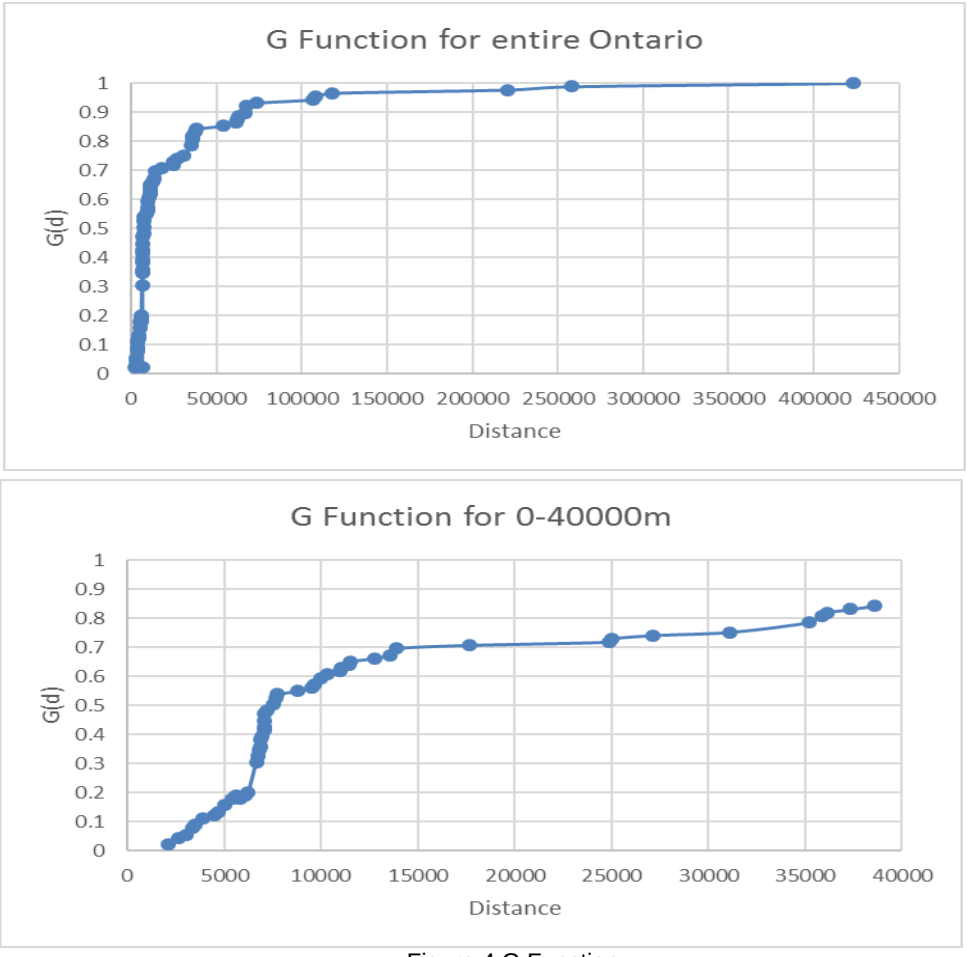
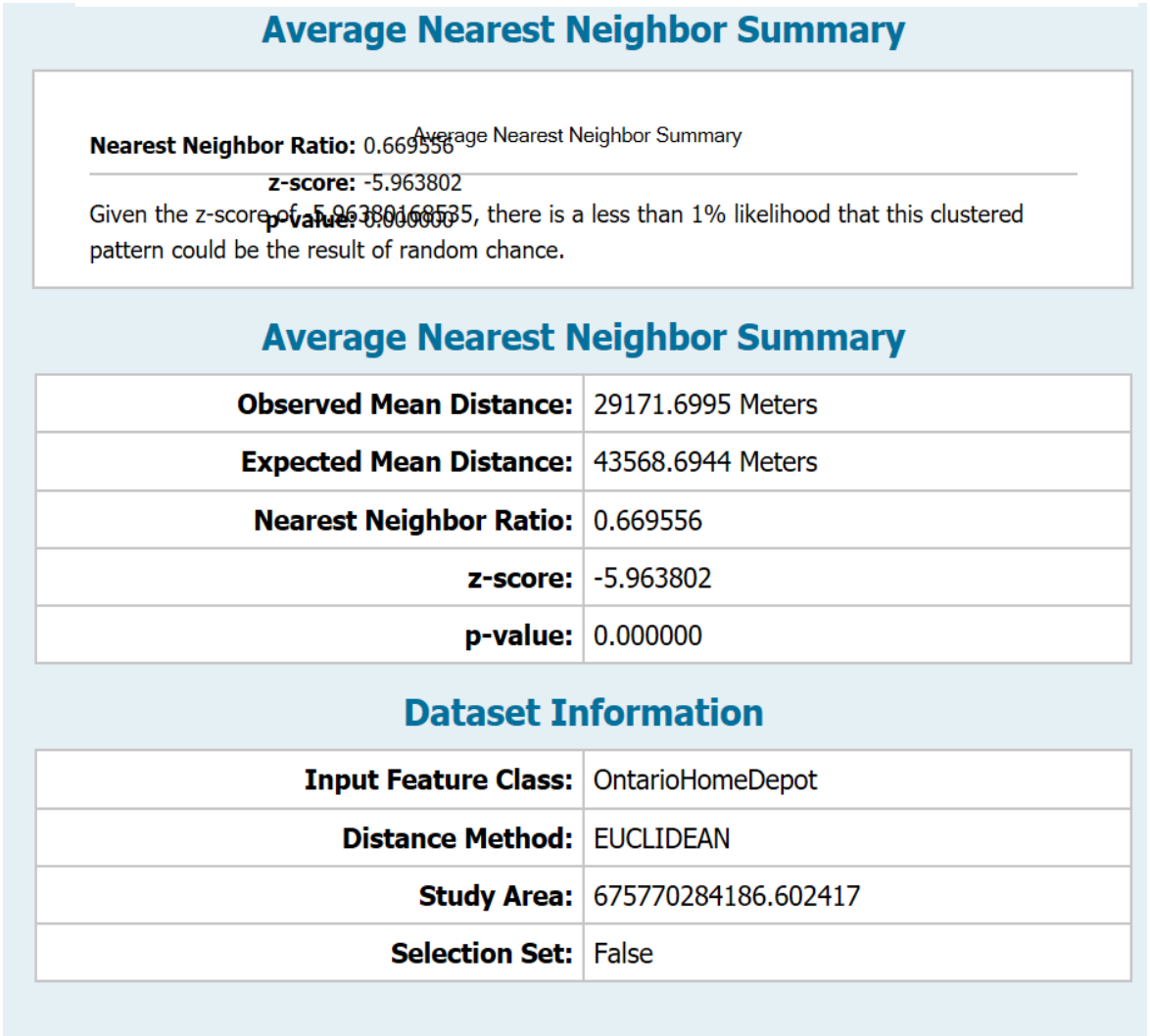


Figure 4 G Function

To examine G function in ArcGIS, we joined the Excel file as an attribute table to the OntarioHomeDepot layer, and performed Average Nearest Neighbor Tool. Figure 5 shows the report of the results. Overall, the pattern is 1% likely to be random. The Nearest Neighbor Ratio (0.669556 < 1.0) suggests a tendency of clustering. The z-score (-5.96308< -3.58) and the p-value (0.0 0< 0.001) indicate 99% confident level.

Figure 5 ArcGIS Report



Average Nearest Neighbor Summary

Average Nearest Neighbor Summary

Nearest Neighbor Ratio: 0.669556

z-score: -5.963802

p-value: 0.000000

Given the z-score of -5.963801685535, there is a less than 1% likelihood that this clustered pattern could be the result of random chance.

Average Nearest Neighbor Summary

Observed Mean Distance:	29171.6995 Meters
Expected Mean Distance:	43568.6944 Meters
Nearest Neighbor Ratio:	0.669556
z-score:	-5.963802
p-value:	0.000000

Dataset Information

Input Feature Class:	OntarioHomeDepot
Distance Method:	EUCLIDEAN
Study Area:	675770284186.602417
Selection Set:	False

III. Response to the Questions

Question 1: Explain why the proportion of features at each standard distance do not align with the proportion of variance from the mean typically found in aspatial data?

Basic statistics assumes that events happen randomly to a degree, while spatial events occur with autocorrelation, which means they are more likely to be closer together. In Figure 2, Home Depot stores tend be fall in eastern Ontario. This "clustering effect" explains why in Table 2, 76.40% stores fall in the circle with the radius of standard distance from the mean point, significantly more likely than how many events (68%) are within the standard deviation from the mean value. However, since statistical events happen more randomly, more (99.7%) events happen within three standard deviation from the mean, than the stores (98.88%) that are located within three standard distance from the mean point. In other words, spatial events are more likely than the statistical events to fall out of three standard deviation.

Question 2: What would our spatial summary statistics tell us if we had additional data that described all of the Rona stores in Ontario? In other words, how could we compare these two point patterns based on the above summary spatial statistics and what could the results tell you?

We could compare the mean locations of Home Depot and Rona based on the coordinates, compared the X and Y residuals to see whether the locations are closer along X or in Y direction, and plot the summary circles to see if Rona has more stores closer to the central location than Home Depot does. By comparing with Home Depot's

counterpart retailers, we can examine the locations of other retail companies and learn from their strategies (for example, to identify a potential market or avoid competition), or see the bigger picture of the retail industry in Ontario.

Question 3: How would your results change if you separated Ontario into Northern and Southern Ontario? Be as explicit as possible and use the knowledge you have gained from the lecture, text, and the methods above.

Question 4: How would the G function change if it was calculated separately for Northern and Southern Ontario?

If we split the data into north and south by the mean Y value, and did the analysis with two sets of objects in R, we would see discrepant results because almost all the Home Depot stores are located in south Ontario. Obviously the mean location and residuals would change based on the point coordinates. The standard distance for southern Ontario would be smaller than that for the entire Ontario, while the standard distance in the north would be much greater because the northern stores are so far and few between.

G function for the south would rise faster at closer distance because the points are so closely together, while for the south G function might be much smoother (as the X values spread out) inflecting from each of the few points. In other words, the south would look more like the graph for 0-40000m in Figure 4, while the north world look like polylines connecting a small number of points rising from lower left to upper right.

Question 5: What information does the average nearest neighbour provide and when is this calculation useful?

Question 6: Test for significant difference between the expected and observed quadrat distributions for the Home Depot data, similar to the text and lecture 3. Can you accept or reject the null hypothesis that the data is represented by an IRP?

The Average Nearest Neighbor Tool in ArcGIS measures the distances between each centroid and its nearest centroid. We can compare the result of a certain dataset to that of a random distribution to observe if the data are clustered or spread-out. In the ArcGIS Average Nearest Neighbor Tool report, a Nearest Neighbor Ratio below 1.0 suggests a tendency of clustering, while a ratio above 1.0 suggests dispersion.

This calculation is useful for businesses to evaluate competition and opportunities. In addition to commercial purposes, this tool may also allow us to study the distribution or temporal changes of any type of land use in the interest of preserving of species and habitats, facilitating public services, optimizing networks for storage and transport, and so on. It also enables comparing data to a control distribution.

Many statistical tests begin with a null hypothesis, which for pattern analysis assumes the features or values are generated by independent random process (IRP), or interchangeably the complete spatial randomness (CSR). The z-score and p-value returned in ArcGIS tell us whether to reject the null hypothesis. The z-score is standard deviation; the p-value is a probability indicating how likely the pattern is created by a random process. They both are associated with the standard normal distribution, the areas of which indicate condifence levels, or how confident we are to reject the null hypothesis. Such test for difference between the expected and observed pattern can be interpreted into spatial preferences or restrictions.

In our case, the confidence level is 99%, which is the safest or most prudent to reject the null hypothesis. For less conservative confidence levels (90% and 95%), ArcGIS offers FDR Correction to help us accept or reject null hypothesis.