CSE 511- PROJECT PHASE 2 REPORT – GROUP 8

Sachin Araballi **Sandhya Balaji** **Pooja Dilip Bihani**

Kiran Shanthappa **Radhika Kulkarni** **Arushi Gaur**

1. INTRODUCTION

In Phase 1 of the project, we implemented 2 functions - ST_Contains and ST_Within. These were used to perform 4 spatial queries - Range Query, Range Join Query, Distance Query, Distance Join Query. In Phase 2 of the project, we implemented two functions to perform Hot Spot Analysis - Hot Zone Analysis and Hot Cell Analysis. In Hot Zone Analysis, we are calculating the hotness of each locality which is represented by a rectangle. In Hot Cell Analysis, we are calculating the Getis-Ord statistic to identify the hot spots.

2. SYSTEM REQUIREMENTS

Ubuntu 20.04 Virtual machine, OpenJDK 8, OpenJRE 8, sbt 1.5.5, Scala 2.11.12, spark-2.4.7-bin-hadoop2.7.

3. TASK IMPLEMENTATION

3.1. Hot Zone Analysis:

In this task we are analyzing the hotness of the rectangle by using the number of points inside the rectangle as our comparison criteria. Hotness is directly proportional to the number of points inside the rectangle i.e. the more the points, the more is the hotness. We started by loading the files and creating the data frame. ST_Contains is used to determine points inside the rectangle. If the points of the diagonals of rectangles are (x1, y1) and (x2, y2). If the x coordinate of the point falls between x1 and x2, and y coordinate falls between y1 and y2, then the point is inside the rectangle and it will count towards the hotness.

Input: pointPath: path of the csv file containing points, rectangePath: path of the csv file containing the diagonal endpoints of the rectangle.

Output: File containing the rectangles and their hotness. (sample output in Fig. 1)

3.2. Hot Cell Analysis:

The goal of this task is to use Apache Spark to apply spatial statistics to spatio-temporal large data in order to discover statistically significant spatial hot spots. The use of spatial statistics to discover statistically significant clusters or outliers in spatial data is a well-known analytic technique used by GIS experts. When making decisions, spatial statistics are essential – for example, if we are 95 percent certain that the levels observed in this location surpass a legal threshold, we must respond by doing so. The Getis-Ord statistic is a widely used statistic for identifying statistically significant clusters (also referred to as Hot Spot Analysis). It provides a z-score and p-values, allowing users to see where characteristics with high or low values are spatially grouped. It's worth noting that this statistic can be utilized in both spatial and spatio-temporal domains; however, we'll be focusing on the spatio-temporal Getis-Ord statistic in this competition.

Input: From January 2009 through June 2015, a collection of New York City Yellow Cab cab trip records was collected. To reduce some of the noisy error data, the source data can be clipped to an envelope containing the five boroughs of New York City.

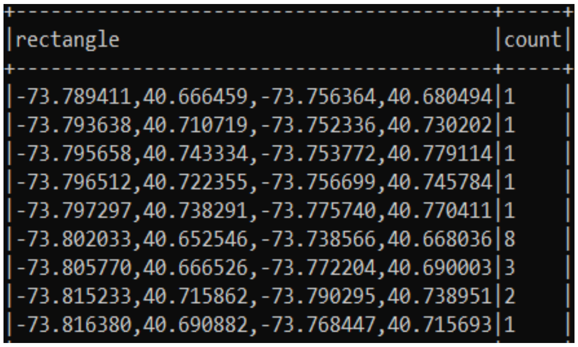
Output: A list of the fifty most significant hotspot cells in time and space (sample output Fig 2) as identified using the Getis-Ord Gi* statistics.

$$G_i^* = \frac{\sum_{j=1}^{n} w_{i,j} x_j - \bar{X} \sum_{j=1}^{n} w_{i,j}}{S \sqrt{\frac{\left[n \sum_{j=1}^{n} w_{i,j}^2 - \left(\sum_{j=1}^{n} w_{i,j}\right)^2\right]}{n-1}}}$$

$$\bar{X} = \frac{\sum_{j=1}^{n} x_j}{n}$$

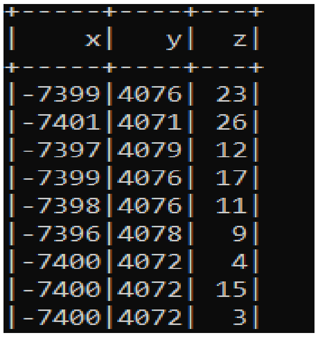
$$S = \sqrt{\frac{\sum_{j=1}^{n} x_j^2}{n} - (\bar{X})^2}$$

where xj is the attribute value for cell j, wij is the spatial weight between cell i and j, n is equal to the total number of cells.



```
+---------------------------------------+-----+
|rectangle                              |count|
+---------------------------------------+-----+
|-73.789411,40.666459,-73.756364,40.680494|1    |
|-73.793638,40.710719,-73.752336,40.730202|1    |
|-73.795658,40.743334,-73.753772,40.779114|1    |
|-73.796512,40.722355,-73.756699,40.745784|1    |
|-73.797297,40.738291,-73.775740,40.770411|1    |
|-73.802033,40.652546,-73.738566,40.668036|8    |
|-73.805770,40.666526,-73.772204,40.690003|3    |
|-73.815233,40.715862,-73.790295,40.738951|2    |
|-73.816380,40.690882,-73.768447,40.715693|1    |
```

Fig 1. Output of Hot Zone Analysis



```
+-----+----+---+
|    x|   y|  z|
+-----+----+---+
|-7399|4076| 23|
|-7401|4071| 26|
|-7397|4079| 12|
|-7399|4076| 17|
|-7398|4076| 11|
|-7396|4078|  9|
|-7400|4072|  4|
|-7400|4072| 15|
|-7400|4072|  3|
```

Fig 2. Output of Hot Cell Analysis

4. TASK DISTRIBUTION

Name	Tasks
Sachin Araballi - 1219530884 saraball@asu.edu	Study of Hotspot analysis, Code setup, Implementation of HotCell analysis
Kiran Shanthappa - 1222119418 khshanth@asu.edu	Study of Hotspot analysis, Code setup, Implementation of HotCell analysis
Sandhya Balaji - 1222169507 sbalaj17@asu.edu	Study of Hotspot analysis, Code setup, Implementation of HotCell analysis
Radhika Kulkarni - 1222165776 rkulka16@asu.edu	Study of Hotspot analysis, Code setup, Implementation of HotZone analysis
Pooja Dilip Bihani - 1219789363 pbihani@asu.edu	Study of Hotspot analysis, Code setup, Implementation of HotZone analysis
Arushi Gaur - 1219396022 agaur17@asu.edu	Study of Hotspot analysis, Code setup, Implementation of HotZone analysis

5. REFERENCES

1. https://github.com/steveloughran/winutils/tree/master/hadoop-2.7.1/bin
2. https://spark.apache.org/docs/latest/
3. https://github.com/jiayuasu/CSE512-Project-Phase2-Template
4. https://github.com/sbt/sbt/releases/download/v1.3.13/sbt-1.3.13.msi
5. http://sigspatial2016.sigspatial.org/giscup2016/problem
6. http://sigspatial2016.sigspatial.org/giscup2016/submit