Leveraging Data Mining Methods in Geospatial Contexts

Dr. William J. Carter *[1], Matthew R. Davis[2] & Emma L. Greenfield[3]
***[1]Guest Faculty, Department of Geography, University of Melbourne, Melbourne, Australia**
[2]M.Tech Geoinformatics Student, Department of Geography, University of Toronto, Toronto, Canada
[3]Research Scholar, Department of Geography, University of Birmingham, Birmingham, UK

ABSTRACT

Geospatial technology involves a combination of data capturing and analyzing devices that tend to generate myriad amounts of data day in and day out in a systematic manner if instructed to do so which puts in front of us a serious issue of vetting out the irrelevant and culling out objects of interest from a maze of collections that can really aid in decision making to solve issues of concern. Data mining, in general addresses this predicament by providing apt statistical tools that carry out an in depth search for obscured correlations among data variables which is the same in case of geospatial data mining with the exception of giving specific emphasis on locational identities that can be either descriptive or quantitative in nature.

Keywords: *Conventional statistical models, Spatial Database, Data Mining Algorithms, Spatial Statistics, Cluster identification.*

I. INTRODUCTION

Data Mining, as the name implies is a diligent and elaborate search for signs or symptoms that establish or sort of prove that there exists a tacit underlying relationship between two or more variables which by doing so helps in forecasting or estimatingunknown variables from their related known counterparts using rudimentary algorithms. With this being said, there are two algorithms that are prevalent among statisticians in implementing data mining techniques namely the Hierarchical clustering and regression algorithms. Though they are not of much relevance in geospatial context, understanding how they work is of paramount importance if one needs to get hold of geospatial data mining concepts.

- **Hierarchical Clustering**

In keeping with it's name, data variables taken for n number of samples or data points are compared with each other so that slightly varying ones are grouped together thereby eventually creating clusters of characteristically similar data points. In order to better understand this let's look at an example, consider the brightness value for Visible, Near infrared (NIR) and Shortwave Infrared (SWIR) bands at two sampled locations namely X and Y. For the purpose of mathematical simplicity the values are scaled down to a maximum of 5.

Table 1

	X	Y
Visible	3	3
NIR	1	5
SWIR	2	4

The variables of these two points are compared with each other to identify the most similar valued pair which is later merged to form a cluster. Though it is glaringly obvious that Visible and SWIR bands look identical for both the samples, there are certain ways to determine the degree of similarity with Euclidean distance being the most commonly used. It's been calculated as follows,
Eucl. Distance = ((Difference between two bands for X)^2 + (Difference between two bands for Y)^2)½which on applying Visible and SWIR band values will giveD = ((3-2)^2 + (3-4)^2)½ = 1.414

The value D gives the diagonal distance between the taken variables, in other words Visible and SWIR values are 1.414 units apart on a given measurement space. On calculating D for different band combinations it's been found

that Visible and SWIR are very less apart as opposed to others which make them more likely to get merged into a single cluster thereby giving a strong sense that the impending cluster may indicate Waterbody spectral class.

- **Regression Algorithm**

One of the most commonly used algorithms in business intelligence and scientific studies where a variable's value is predicted based on the value of another, while the former is called dependent (Y) the latter is called an independent variable (X).Though there are umpteen variants of it, Linear regression holds good for a pair of variables varying simultaneously say for instance Water availability and Plant productivity parameters that are linearly related to each other with plants(Y) depending on water (X). Based on this an unknown productivity value for a given water availability data can be estimated by plotting yesteryear values as points on a 2 D scatter plot thereby trying to fit a (Regression) line that best fits in it.

II. METHODOLOGY

As obvious as it gets, Data mining is of utmost importance in each and every technological aspect with the Geospatial front being no exception which is quite evident from the examples cited. Due to a constrain in storage of dimensions, querying capabilities and incorporation of attributes pertaining to space, conventional databases were rendered incapacitated thereby giving rise to what are known as Spatial databases. Certain questions have been formulated and justified so as to what makes them different and so on as follows.

- Why should we prefer Spatial database to conventional ones?
- What makes Spatial statistics so special?
- What sort of algorithms do Spatial data mining tools adhere to?

III. ANALYSIS

- Spatial Databases

While data mining refers to mining algorithms adopted by ordinary databases, spatial data mining are the ones done by spatial databases. Though databases like Oracle and Postgre SQL were enriched with a pool of analytical functions to gain insights they fall short in incorporating geometrical data types that include point (0 Dimensional), Line (1 Dimensional) and Polygon (3 Dimensional) objects that has prompted them to develop Oracle Spatial and PostGIS respectively. With this marking a stark difference, Spatial Indexing, which is a lucid arrangement of spatial and non spatial attributes for effective storage, spatial querying and analysis has what made the gap look more pronounced.

Goehashing, just like Geocoding is one of the common ways of indexing spatial data where a combination of letters and numbers are assigned for each entity that serves as a unique identifierwhich an analyst can use in his/her spatial queries, further concisely storing large entities in a crisp manner using cryptic codes . This cryptic letter-number combo can be deciphered into latitude and longitude values using a base 32 encodingwhere decimal to binary conversion takes place. For instance, a geohash index value of **bcd14** is converted into binary bits as per base 32 standards as follows,

Table 2

Decimal	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
Base 32	0	1	2	3	4	5	6	7	8	9	b	c	d	e	f	g	h	j	k	m	n	p	q	r	s	t	u	v	w	x	y	z

Substituting appropriate number for each index value and converting them to binary number will give a series of bits which eventually has to be interpreted in the form of latitude and longitude coordinate pair.
b = 10 = 01010, c = 11 = 01011, d = 12 = 01100, 1 = 1 = 00001, 4 = 4 = 00100

Bits 01010 01011 01100 00001 00100 are segregated into X and Y values by culling out even numbered values starting from left to right as Longitude and odd numbered as Latitude. Further, the intervals are assigned based on bit values, that is the entity is assumed to be in the upper interval range (0 to 90 degrees) if it's value is 1 and lower interval range (-90 to 0 degrees) if it is 0 and so is the case for longitude(0 to 180 deg for 1, -180 to 0 deg for 0), the procedure is iterated until all the bits of an index are done away with. From another perspective, since the index

values are encoded based on latitude and longitudinal positions, objects that are nearby tend to have the same prefix which can be really helpful for an end user in terms of rudimentary interpretation.

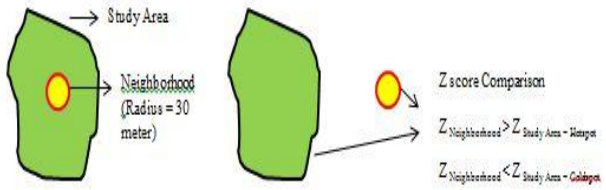
- Importance of Spatial Statistics

Multitudes of spatial statistical tools are on offer of which two of them were discussed to emphasize their significance and usefulness over on paper statistical tools.

✓ *Hotspot Analysis:*

As opposed to basic Choropleth maps which are nothing but different shades of a colour whose intensity varies based on magnitude of the taken variable, hotspot analysis produces decisive hot and cold spots in an area of study by performing Getis ord gi* statistical test. Since delving further into the mathematical concepts of Getis ord gi* is not the scope of this paper, it's nature of work is concentrated. Simply put, this tool is a test of spatial randomness, that is it determines what is the percentage of probability that a cluster is randomly achieved. Higher the probability is , the more likely a cluster is randomly formed and vice versa. This in turn is achieved by carving a feature of interest along with it's neighboring features that fall inside a user definable radius out of the study area and comparing their Z score values. If the Z score of carved out subset is **significantly** higher than the study area's Z score it belongs to a hotspot and vice versa.

Figure 1



Z score Comparison

$Z_{Neighborhood} > Z_{Study\ Area} = Hotspot$

$Z_{Neighborhood} < Z_{Study\ Area} = Coldspot$

Though Neighborhood sizes can be defined using different methods, the one used here (Fixed Distance band) is predominantly used as it allows the analyst to exercise his/her's sole discretion. Z score is a statistical parameter that takes into account a population's mean and standard deviation and is calculated as follows,

$$Z = \frac{\text{Value of a Feature} - \text{Population Mean}}{\text{Population Standard Deviation}}$$

Z scores obtained for all the features are averaged out for neighborhood and study area which are later compared to delineate hot and cold spots, ones that **do not vary significantly** are termed Insignificant or randomly formed clusters.

✓ *Spatial Autocorrelation*

It is a test for feature similarity that is carried out using Moran's I index thereby declaring whether the dataset of interest is clustered, random or dispersed based on the obtained index value. In a crux, it calculates the mean value of the entire dataset and compares how much each feature's value within a specified neighborhood deviates from the calculated mean. These deviation values are multiplied with each other to get a final cross product of the neighborhood, if the cross product value is positive so will be the index value and area of interest is deemed clustered, if cross product is negative it is deemed dispersed and if positive and negative cross products balance each other out, the neighborhood is deemed random and agrees with the null hypothesis that there exists a complete spatial randomness in the dataset.

- Algorithms of Interest

Now that some of the statistical tools were discussed, a brief note on some of the clustering algorithms were portrayed.

✓ *DBSCAN*

Density Based Clustering for Applications with Noise (DBSCAN) gives utmost importance to how densely data points are clustered in a measurement space by stressing on two important parameters namely Epsilon (ε), which is

an analyst specified radius for drawing a circle from a data point's center and Min Points refers to the minimum number of specified points that need to be present inside a drawn circle of radius Epsilon. If a point with a circle of radius Epsilon satisfies the Min points condition, it is denoted as Core point, that is a point situated in a highly dense region. In DBSCAN a cluster is bound to formed as long as the points in ti are connected only through Core points. If a point is deemed Core, it looks within

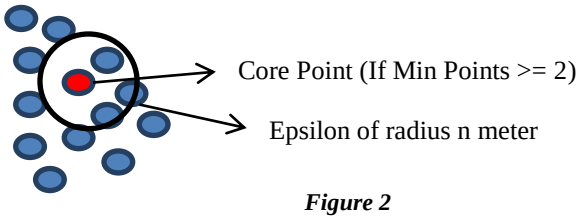


Figure 2

it's neighborhood for fellow Core points thereby including neighborhood points of the newly assigned core points in it and thus gradually increasing it's Cluster size until it breaks off due to shortage of Min points within an Epsilon.

- ✓ *K – Means Clustering*

K - Means Clustering approach takes an input from the user in the form of number of clusters to be assigned which on feeding will randomly assign cluster centers on an n dimensional measurement space. Points/Pixels in a dataset are assigned to a cluster in such a way that the distance between the former and a cluster's centroid is as minimum as possible. This process of allocation is carried on until all the data points are ensconced within anyone of the specified number of clusters. Now the filled up clusters' new centroids are calculated which on doing so will alter their positions to some extent and the whole procedure is repeated iteratively until there is no significant difference in a cluster centroid's position before and after the iteration is observed.

- ✓ *ISODATA Clustering*

Iterative Self Organizing Data Analysis Technique (ISODATA) is a slight variant of K – Means which allows the number of clusters to be modified in the likes of merging, splitting and deleting based on user specified rules after undergoing each iteration. Some noteworthy rules are two clusters would merge if their mean distance is less than a specified threshold and a cluster can split up into two if it's standard deviation is greater than the maximum specified value. As it is the case with K – Means, ISODATA algorithm carries on until it is done with the specified number of iterations or if it fails to find any discernible change in cluster statistics.

IV. CONCLUSION

Data mining concepts in general and it's necessity in geospatial contexts were discussed from a bird' eye view in a brief manner with some core technical know-how being touched upon. As mentioned earlier, Spatial data mining is punching above it's weight globally amidst some developing countries still trying to get a grasp of geospatial technical perks due to the abundant availability of data generated irrespective of it's relevancy. Statistics and data mining always go hand in hand and it's high time that sophisticated stat algorithms are integrated in geospatial tools offered by both commercial and open source software if decision making solutions have to be made.

REFERENCES

1. *Melissa Rudy,"Article-Data Mining"Techopedia, (2014)*
2. *E-Book,"Spatial Data Mining: Association and Clustering", Indian Agricultural Statistics Research Institute (IASR), (2017)*
3. *David Smiley,"Article-Spatial Search with Geohashes", Lucidworks, , (2010)*
4. *E-Tutorial-"How Spatial Autocorrelation Works?" ArcGIS Pro–Analyzing PatternToolset Concepts, (2018)*
5. *Video Source, "Spatial Data Mining I:Essentials of Cluster Analysis" ESRI Events, (2017)*
6. *Video Source,"Machine Learning #75 Density Based Clustering", Xoviabcs, (2017)*
7. *Video Source,"Z Score Introduction", Khan Academy, (2018)*