GEO SPATIAL DOMAIN IN BIG DATA: A COLLECTIVE STUDY USING OSM

[1]**Hunny Yadav,** ***Research Scholar,***
BMU University, Rohtak,
(Haryana)
m.hunnyyadavgmail.com

[2]**Prof.Dr.Brij Mohan Goel**
BMU, Asthal Bohar, Rohtak
Brijmohan.vce@gmail.com

Abstract

The “Geo Spatial big data” has a collective term. It is very difficult, complicated and complex to analyze the Diverse, enormous, invariable semi-structured or unstructured digital material by using typical DBMS tools and methodology. This data is also known as semi-structured data. Using “Geo Spatial big data techniques” the main goals are: providing best decision making, by increasing efficiency to reduce cost, the presence of real time assistance and spatial link analyses. The research methodology depends upon the reviewed research. If we talk about the trends of Geo Spatial big data analytics application- First is conceptual method and in this method we will see about data processing, overlay analysis, the prediction of changes in land use, land analysis on Global Scale. In the field of Data Mining, the second focus pertains to the study of human mobility and disaster management. The third area of interest involves knowledge representation, including problem-solving, Geo-Spatial knowledge representation, and the exploration of big data.

Keywords: Geo Spatial Big Data, Management, Geo Spatial Technology, Big Spatial Data (BSD), Location’s Base Services (LBS).

1. Introduction

The epoch of "Big Data" is rapidly approaching, poised to significantly transform our approach to environmental management. The term "Big Data" denotes data sets of such immense size that they surpass the capacities of conventional data management techniques in current use. "Big data" encompasses the vast data sets that can be mathematically analyzed uncover patterns, trend and correlations, especially those related to human behavior and social interactions. Geographical data that express three V's and surpasses the capabilities of existing computer systems could be identified as "Big" based on these attributes. Big Spatial Data (BSD) within the various area such as satellite images, the IOT’s, the Location’s Base Services (LBS) and climate-simulation demonstrates a strong alignment with the characteristics mentioned above, highlighting the need for customized systems, methodologies, and algorithms from the very beginning. Before the advent of big data, there was already a surge in organizations incorporating BSD.

The subsequent are numerous sources of spatial big data: Data acquired by sensors through the internet of things (IOT) represents one source. Another source is the application of Landsat and other remote sensing techniques for land assessment worldwide, climate ontology, as well as the accumulation of diverse data through web services and navigation (A Trajectory).

1.1 Big Data

Big data arises from diverse sources of significant magnitude (Volume), displaying a range of data structures (Variety), encompassing structured (in 2D form), unstructured (text, images, video, etc.), or semi-structured data (emails, files with HTML, XML extensions), and is produced at a remarkably rapid pace (Velocity). These elements collectively constitute the "3V's characteristics of Big Data." Subsequently, two additional 'V's, specifically Value and Veracity, are introduced in alignment with the criteria. Value and Veracity are the terms denoting these aspects. The union of these five dimensions of big data is known as the "5Vs of Big Data," which also represents a concept.

1.2 Geo Spatial technology

All the many forms of technology required for the collection, archiving, and organization of geographical data are together referred to as "Geo Spatial technology". It makes use of the satellite technology that allowed for the mapping of Earth's surface and the performance of scientific studies on it. The phrase "Geo Spatial technology" describes a collection of interconnected technologies, Global positioning systems (GPS), geographic information systems (GIS), remote sensing, and some other technologies.

1.2.1 Geo Spatial technology and Python

Python is a widely used programming language that is well-suited for working with geographic data. It can handle vector and raster data, which are the two common ways that Geo Spatial data are stored. Because of this, Python is a great option for anybody handling geographic data. Applications like as Fiona and Geoplanids may be utilized to perform a variety of operations on vector data. Raster data may be handled in many ways with a programmer such as arrays.

1.3 Data Cleaning

"Data cleaning" is a useful procedure that is a component of the larger field of "data management." Data cleaning is the act of evaluating all of the information that is currently kept in a database to either update or eliminate information that is redundant, inaccurate, missing, or unnecessary in any other manner. The purpose of this study is to update the database with the most current version of the data. The process of data cleaning is finding a way to optimize the correctness of the dataset without necessarily altering the data that is already there, as opposed to just removing out-dated information to create room for new data.

1.4 Objectives of the Study

1. To investigate the challenges associated with spatial big data analytics applications.
2. To investigate big data computing for applications related to geography.

1.5 Review Literature

Benguigui L. and Czamanski D. (2014), In order to determine the regions that surround the Geographic features (like points, lines, and polygons) a buffer analysis was performed. The procedure consisted of constructing a buffer around preexisting geographic features and then identifying or selecting characteristics depending on whether or not they were located inside the buffer's border or outside of its boundaries. Several of the research articles that were reviewed included the use of the buffer tool.

Bennett J. (2010), A collection of spatially specified alternatives, such as parcels of land, and a set of assessment criteria, represented as map layers, were used in the GIS-based multi-criteria evaluation methods. Despite the reality that the large number of decision rules has proposed in the MCE (Multi Criteria Evaluation) literature, the procedures for making decisions took into account a number of variables in order to determine the extent to which each site was appropriate for the allocation that was being evaluated.

Bolin Centre Database (2018), A Geographic Information System enables the data to be represented with real world spatial connections or topology using appropriate data types such as polygon, network, etc. A geographic information system is sometimes referred to as a GIS or a Geographic database. It makes it easier to store spatial and non-spatial data so that it may be analyzed using the spatial analysis tools that are very helpful to planners. The second characteristic that makes it fascinating is that it can provide geographical data and findings in the form of a map, which facilitates quick comprehension, accurate issue diagnosis, and creative problem solving.

According to Brelsford C, Martin T, Hand J, Bettencourt L.M.A. (2015), an electric utility places a preponderant amount of importance on its distribution system because of two factors: the system's close proximity to the end consumer and its expensive investment cost. Although a significant amount of work is put into the production of electricity, the transmission and distribution of that electricity should not be neglected (Igbokwe & Emengini 2005).

According to Brockmann D., Hufnagel L. and Geisel T. (2016), the electrical distribution system includes a spatial component; using GIS, it is possible to see the system while it is being put out on the ground. You just need to glance at the map and click on a certain feature in the network for all of the pertinent information to be presented. According to Harini and Santhosh (2006), this results in a more flexible knowledge of the network and hence a more expedient approach to problem solving.

In Cattani C. and Ciancio A. (2016), their research, Rahmati and colleagues wanted to determine how well statistical models, particularly the weights and frequency ratios of the evidence models, could predict the likelihood of flooding in the Golestan Province of Iran. (Rahmati et al. 2016). (Rahmati et al. 2016). In order to achieve this goal, they developed a flood inventory that included 144 flood sites and used 9 flood causative elements from the spatial database. Some of these factors included land use, distance from rivers, and other similar characteristics.

Chen Y. (2019), carried out research that resulted in the development of a river flood susceptibility mapping in response to an exceptional flood event that occurred in Brisbane in 2011. For the purpose of mapping flood-prone zones in the region, we applied the different machine learning models: one is a decision trees and second is the support vector machine. In research conducted, the Decision Tree model performed better than the Support Vector Machine model and attained a high level of accuracy of 88.47%.

1.5 Methodology

Because this study is conducted in a big-data environment and requires geo-computation of large amounts of data, the overall conclusions heavily depend on the methodology being used. A few of the technical requirements for handling and simulating the VGI datasets are discussed here. More specifically, the datasets are primarily sourced from Bright kite, Gowalla, Twitter, and Open Street Map (OSM) environments. The minimum number of people in each dataset is tens of thousands, as are the number of geographic characteristics. When data is acquired from the bottom up, it means that every single user is contacted. The method of data processing will be emphasized, along with how to construct the data structure, remove unnecessary portions, and extract important information, as the framework implies. This will also cover a data organization.

1.6 STUDY I: The Heterogeneity of OSM Data and Community

In order to explore the basic scaling property of geographic space, this study examined the global OSM historical record, involving human actions. Approximately 2.7 billion unique contributions, one million users, and two billion geographic features can all be stored in this XML file format database. Because of this, the dataset used in this study is quite large, spanning eight years and 692 gigabytes when uncompressed. All of the OSM objects that were part of the history dump yielded pertinent historical and attribute data for our research. Subsequently, head tail break and detection of Power-Law were used in this study. From the user viewpoints the global OSM database describes the Scaling Trend. This was carried out in an effort to identify the best scaling technique.

1.7 STUDY II: A Socio-Geographic Perspective on Human Activities

Study focuses on the relation of social & spatial human aspect activity on the platform of social media. For more accuracy this uses the scaling analysis of socio-geographic networks and human mobility behaviour to search for a relationship between check-in locations and social relationships. The data came from the now-defunct American location-based social networking apps Bright Kite and Gowalla. We used a TIN-based clustering approach to construct natural cities based on user locations, which we first obtained as the first or most recent check-in location for each user. We were able to successfully finish the work as a result. Every social media platform, the research found three different kinds of socio-geographic networks: people to people network (P-P), location to location network (L-L) and city to city network(C-C). Up to

10's of thousands of nodes and 10's of millions of linkages are found in the socio geographic network and as on the bases of social-media data from about 50K user of having check-ins locations are 6 million. The complex aspects of network design are separated and illustrated. The networks were used to do a correlation study between the spatial and social aspects of human activity.

1.8 STUDY III: A Smooth Curve as a Fractal under the Third Definition

Prior study has shown that the Geographic space exhibits a fractal or scaling configuration across different scales. This study focuses on examining the fractal properties of diverse geographical features.

The utilization of this novel concept in the context of a simple polyline is exemplified in **Figure 1**. The polyline consists of a total of ten angles, all of which are calculated recursively. Moreover, the proportion of significantly more minor angles to major ones repeating twice (htindex = 3) is articulated as under: the x1 **+ x2 + x3 > x4 + x5 + x6 + x7 + x8 + x9 + x10** and **x1 > x2 + x3**, thereby indicating the fractal nature of the polyline. As per third stipulation, nearly all continuous curves fulfill the conditions to be categorized as fractals, given that the recurrence of scaling a considerably larger quantity of small angles than large angles transpires multiple times. To secure this assertion, the publication presents interconnected analyses of four diverse forms of smoothie curves: semicircle, semi ellipse, logarithmic spirals, and the shoreline of the United Kingdom (following smoothing processes).

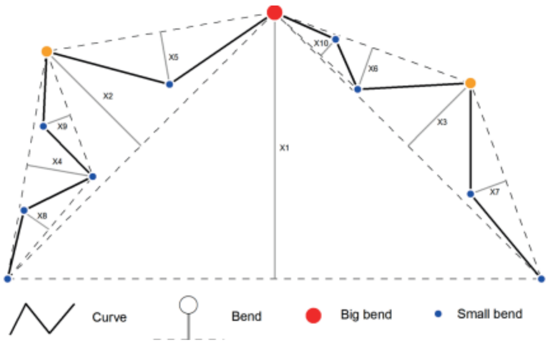

Figure: 1

The illustration of the new definition of fractal

1.9 STUDY IV: Why Topology Matters in Predicting Human Activities

The present study explores the reasons behind the importance of spatial topology, in particular the topological connection between natural streets on cities, for estimating the actions of people. The study includes a thorough comparison between the two types of representations and assessed how well geometric and topological representations captured or predicted human behaviours. In this study, natural streets and axial lines are the focus of attention for the topological representations, while segment-based models are the focus of attention for the geometric representations at the city level. The results of this study show that segment-analysis methods are essentially geometric, which makes them unsuitable for forecasting human behaviour. These techniques focus on variables that hardly ever exhibit scaling properties, including segment lengths and the turning angle of two intersecting segments. It is, however, possible to study the underlying scaling of a far greater number of streets with fewer connections than those with more established linkages to topological models.

We have a series of experiments using **Location Data from A week worth of tweets and the streets of London to give evidence in support of this argument.** These experiments were based on the related concepts of axial line and axial line segments (sometimes called line segments for short) and natural-streets and natural-street-segments (NSS) (also called short street segments). Different power-law fitting metrics also revealed that the streets connectivity values distribution & line segments did not hold significant scaling features, in contrast to the distributions of connection values for the line i.e. axial line & natural-streets. Therefore, we found that the connection value distributions for the line i.e. axial line & natural-streets have remarkable scaling features. Subsequent investigation demonstrated that street connectedness's scaling feature could be helpful in predicting human activity.

1.10 STUDY V: Least Community as a Homogeneous Group in Complex Networks

The existence of communities dispersed throughout a complex network has a substantial impact on its topological structure. Due to the possible variation in the link distribution between different nodes within a complex network, there is a heightened likelihood of community presence within the network. Despite the diverse relationships among them, fractals can emerge from these communities within a complex network. The objective of this research is to identify the scaling & characteristics of the community arrangement within a complicated network, with the aim of extending the uses of the 3rd description of fractals and laws of scaling to network domain.

This approach enables researchers to accurately delineate the heterogeneity under investigation. A comparable random graph representation is identified for each real-world network, with equivalent numbers of nodes and connections.

1.11 STUDY VI: Spatial Distribution of City Tweets and Their Densities

In this study The UrbanSpaceStructures (USS) their effects on the human activity’s spatial distribution (SD) are examined. Here the study describes the SD of the quantity of **GeoTaggedTweets (GTT)** & city level density across blocks of street, it also construct the natural city portion. This study employed the Open Street Map street network as its geographic unit, which was totally created from the bottom up, in contrast to previous studies that relied on T-D (Top-down) Geographic units like governmental area boundaries.

The Government imposed Top-down units and it may be out dated and subjective. In contrast, Bottom-up units are more dynamic, objective, and have exceptionally fine spatial-temporal resolution. Government governed an order and in this order Top-down units is random.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper © 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 12.Iss 01, 2023

Therefore, the goal of this research is to examine, within the context of geographic big data, the various forms that cities might take and how these forms affect human activities.

Figure: 2

The spatial distributions of tweet numbers and densities in London

1.12 STUDY VII: How Complex is a Fractal

The breaks between the head and tail, also referred to as the htindex induced by a fractal, form the foundation for the third definition of a fractal. The ht_index spans from 1 to ∞ and serves to evaluate the fractal nature of a geographical element (with a higher ht_index indicating a greater level of fractality in the geographical feature). An illustration follows to demonstrate the computation of the ht_index for a dataset having 10 values based on Zipf's Law (Zipf 1949): [1, 1/2, 1/3..., 1/10]. The intention of this illustration is to ensure the self-sufficiency of this section. By utilizing head and tail breaks, two distinct averages are derived. The initial average of 0.28 divides the dataset into a head comprising "1, 1/2, 1/3" and a tail comprising "1/4, 1/5, 1/6, 1/7, 1/8, 1/9, 1/10," while the subsequent average of 0.61 further separates the head in to a fresh head containing "1" and a new fresh tail containing "1/2, 1/3." Consequently, the ht_index value assigned to this dataset is 3. In addition, during the execution of the head/tail breaks procedure, it is possible to assign the ht_index value to each individual element within the series. For example, the numerical value "1" corresponds to the ht_index of 3, while "1/2, 1/3" is associated with the ht_index of 2, and all other values receive the ht_index of 1. Consequently, the hierarchical organization of data series provides insights into the fractal nature of each specific value. Nevertheless, the ht_index lacks the sensitivity required to detect subtle variations that may arise during the transition between different data series. Even with the inclusion of "1/11, 1/12, 1/13, 1/14, and 1/15" in the original data series, the ht_index value remains unchanged for the updated series. Evidently, there exists a sensitivity issue pertaining to the ht_index metric.

1.13 Results and Discussion

This covers the results of all seven studies and talks about the unique contributions each one made to science. The result focuses on the Scaling analysis of topological order of Geo Spatial big data in order to gain a deeper understanding of the Geographic space's scaling structure and how it affects human activities.

1.14 Conclusion

Geometrically speaking, big data exhibits a fractal structure that Euclidean geometry struggles to get completely. A statistical perspective on big data indicates that it typically has a heavy-tailed distribution, meaning that the data cannot be described by a single, well-functioning mean value. This is why it is necessary to develop a completely new theoretical framework, grounded in Paretian Statistics and fractal geometry, for Geo Spatial analysis the big data stage. The new paradigm upon which they are founded, big data analytics is endowed with data-intensive computing approaches.

REFERENCE

1. Brockmann D., Hufnagel L. and Geisel T. (2016), The scaling laws of human travel, Nature, 439, 462–465.
2. Cattani C. and Ciancio A. (2016), On the fractal distribution of primes and prime-indexed primes by the binary image, Physica A, 460, 222–229.
3. Chen Y. (2019), A new model of urban population density indicating latent fractal structure, International Journal of Urban Sustainable Development, 1(1–2), 89–110.
4. Chen Y. (2021), Modelling fractal structure of city-size distributions using correlation functions, PLOS ONE, 6(9): e24791. doi:10.1371/journal.pone.0024791.
5. Chen Y. (2015), Power-law distributions based on exponential distributions: Latent scaling, spurious Zipf's law, and fractal rabbits, Fractals, 23(2): 1550009.
6. Christaller W. (2023), Central Places in Southern Germany, Englewood Cliffs, NJ: Prentice Hall.
7. Cho E., Myers S. A., and Leskovec J. (2021), Friendship and mobility: user movement in location–based social networks, In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, U.S.A, 1082–1090.
8. Clark C. (2021), Urban population densities, Journal of the Royal Statistical Society: Series A (General), 114(4), 490–496.
9. Clauset A., Shalizi C. R., and Newman M. E. J. (2019), Power-law distributions in empirical data, SIAM Review, 51, 661–703.
10. Cohen R. and Havlin S. (2019), Complex Networks: Structure, Robustness and Function, Cambridge University Press: Cambridge.
11. Corbett J. P. (2019), Topological Principles in Cartography, Technical study 48, U.S. Dept. of Commerce, Bureau of the Census: Washington D. C.

12. Cranshaw J., Schwartz R., Hong J. I., and Sadeh N. (2022), The livehoods project: Utilizing social media to understand the dynamics of a city, Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media, 58–65.
13. Diakoulakli D., Mavrotas G. and Papayannakis L. (2015), Determining objective weights in multiple criteria problems: The critic method, Computers and Operations Research 22(7), 763–770.
14. Dubin R. (2018), Estimation of regression coefficients in the presence of spatially autocorrelated error terms, The Review of Economics and Statistics, 70(3), 466–474.
15. Edmonds J. (2015), Paths, trees, and flowers, Canadian Journal of Mathematics, 17, 449–467.
16. Edenhofer M. J. and Herring J. R. (2010), A mathematical framework for the definition of topological relationships, Proceedings of the Fourth International Symposium on Spatial Data Handling, International Geographical Union, Zurich 1990, 803–813.
17. Erdős P. and Rényi A. (2019), On random graphs I, Publicationes Mathematicae, 6, 290–297.