Buffer Analysis and Modified K-Means Clustering for Geo-Spatial Amenities on Gujarat state

Shikhar Brajesh, Karan Katyal, Manoj Pandya, Bharat Chaudhary, Hitesh Bodar, Paru Thakkar, Leenal Patel, Bhoomi Patel, Krunal Patel

Abstract—increasing digitization of demographic data provides us with an excellent opportunity to combine it with available geo-spatial data about facilities like schools to derive meaningful conclusions from the data and formulate better plans for betterment of these facilities. Various data mining techniques can be studied and extended to use with geo-spatial data.In this effort; we have used the concept of spatial buffers along with a modified K-Means clustering algorithm for knowledge discovery from the available data on assets like schools in Gujarat State.

Index Terms— Buffer Analysis, Geographic Information Systems, Geo-Server, K-Means clustering algorithm, Multi-point Buffer, OpenLayers, Point Buffer, Postgres, PostGIS

1 Introduction

THERE is great variety of data being generated continously in the world today.in this world. Computers have become cheaper, more powerful and easily accessible to common masses. There is e-commerce data, environmental data, data from social media and also extremely useful geo-spatial data.

Geo-spatial data refers to information that identifies the geographic location and characteristics of natural or constructed features and boundaries on the earth. It is typically represented using points, lines, polygons and other complex geographic features. It includes both original and interpreted data collected at stored at enormous speeds using remote sensing satellites, aerial surveys, telescopes and scientific simulations.

Buffer analysis is among the various techniques to analyse the geo-spatial data. It refers to a technique of data retrieval which is useful in analyzing information about a region surrounding a specific location of interest. The location of interest can be a single point, a linetring joining a set of points or even an irregular polygon. The structure of these different type of shapes is defined through a geometry data type supported by various spatial data base management systems and software. The other parameter necessary to define a buffer is the buffer radius. According to the buffer radius specified, a region is selected in the spatial domain around the given geometry data.

A point buffer refers to a section of area taken around a selected point on map and all the data points that come under this area are listed. In our statistical analysis we would be using multiple point buffers along with various set operations to filter out the data to be processed. After data retrieval step, we will use a modified version of the popular K-Means Spatial Clustering algorithm. We would also be merging data points to make our analysis less computer intensive.

In Section 2, we would be discussing existing literature and terminology related to buffer analysis, various clustering algorithms with their extensions for spatial data and basic devices to be used in our analysis. Section 3 would comprise of implementation methodology and system design followed by Section 4 about results, performance evaluation and statistical anlysis. The paper would then conclude with a suitable conclusion in Section 5.

- *Shikhar Brajesh & Karan Katyal are currently pursuing undergraduate degree program in Computer Science in Birla Institute of Technology and Science, Pilani, India, PH-+91-9772064361. E-mail: shikharbrajesh9@gmail.com*
- *Manoj Pandya, Paru Thakkar, Krunal Patel, Leena Patel, Bhoomi Patel, Bharat Chaudhary & Hitesh Bodar are affiliated with Bhaskaracharya Institute for Space Applications & Geo-Informatics (BISAG), E-mail: mjpandya@yahoo.com*

2 Background

2.1 Rationale

Geographic Information System (GIS) is a computer-based information system used to digitally represent and analyze the geospatial data or geographic data.
GIS can be thought of as a system that provides spatial data entry, management, retrieval, analysis, and visualization functions. The implementation of a GIS is often driven by jurisdictional (such as a city), purpose, or application requirements.

2.2 Components of a typical GIS

A typical GIS implementation consists of the following functional parts:

Spatial database that can provide random access to large data sets, query processing that understands spatial relationships, and transactional integrity during concurrent editing. Examples include Oracle Spatial, SQL Server Spatial and PostGIS/PostgreSQL.

Desktop software that can provide direct editing and visualisation of data in the database. For data management, quality control, and ad hoc reporting. Examples include ArcGIS, MapInfo, QGIS, uDig.

Cartographic map renderer reads spatial data from the database, applies styling rules and outputs map images. Examples include ArcIMS, ArcGIS Server, MapServer, MapGuide and GeoServer.

Application server that can provide a programming framework for custom applications. Examples include ArcGIS

Server, GeoServer and MapGuide.

Map tile server that can store pre-rendered image tiles and serve them up quickly to make maps refresh faster. Examples include ArcGIS Server, MapGuide, Tilecache and GeoWebCache.

Web map component that can provide a map component inside a web browser. Examples include Google Maps, OpenLayers.

For more detailed information refer [1].

2.3 Partitional Clustering Algorithms

Partitional clustering methods [6] determine a partition for dividing a group of points into different clusters, such that the points in a cluster are more similar to each other than to points in different clusters. These methods start with some arbitrary-initial clusters and iteratively reallocate points into clusters until a stopping criterion is met. They tend to find clusters with hyperspherical shapes. Examples of a partitional clustering algorithm is k-means algorithm which clusters data based upon their mean value.

3 System Design

The application is developed to be a highly interactive and user-friendly. There are three types of data needed for doing our analysis:

1. Spatial data of village centres and boundary in Gujarat
2. Spatial data of different types of school.
3. Demographic data of the various schools.

The process begins by identifying data to be analysed by creating point buffer(s) and performing set operations as specified by user to select the data points on which we need to perform clustering operations. After this step, we perform K-Means clustering algorithm which has been modified to cluster spatial data in K clusters determined according to the maximum buffer distance specified by the user, which is shown as below.

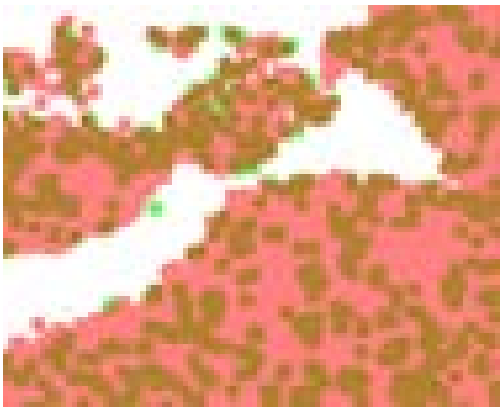


Figure 1: User Interface with school centres in red and green points as village centres

3.1 Creating point buffer

For taking input from the user, we use a web portal based on GeoServer ver. 2.4.2. The code is written in JAVA using JavaServer Pages technology. The platform used for coding is NetBeans IDE and JavaScript is used to do the client side programming. PostGIS is used as spatial database management system.

OpenLayers is a JavaScript library which can be used to display and customize maps to be displayed on client side. GeoServer along with OpenLayers library was used to display the map of Gujarat state to the user and input coordinates through mouse clicks on the map. Before clicking the user specifies the buffer radius through a menu panel. The user can also specify if he wishes to perform set operations on multiple buffers to obtain the desired data points. Ex. If a user wants points within a radius of 3 km from A and 7 km from B, A and B will be centres of respective point buffers and we will take an intersection of the regions. Similiarly, union, difference and complement operations can be performed.

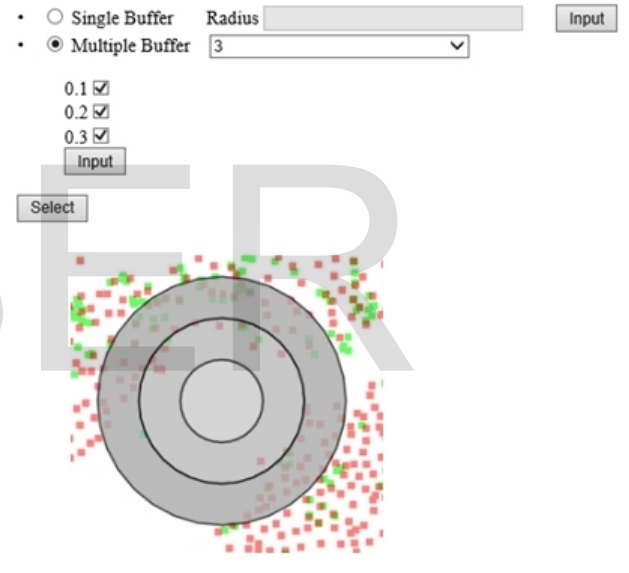


Figure 2: Multiple point buffer with radius 0.1, 0.2 and 0.3 degrees (1 degree is approx. 110 km)

3.2 Buffer based K-Means clustering

For a conventional K-Means clustering algorithm, the user has to determine the most suitable K value heuristically. More information on determining K values in [5]. However the modified K Means algorithm that we have used takes buffer distance d as a parameter. This is the distance specified by the user as *'the maximum possible euclidean distance that a school centre can have from the cluster's centre'*.

The algorithm works as follows:

1. Select the output data points available after point/multipoint buffer operations.
2. Move from top to bottom in the region and store the N points in a separate database.
3. Take centroid of all the cluster data points.
4. Take initial value of K to be N/factor where factor = N if

all points are within user's specified buffer distance d.

5. If not, then factor is incremented by a suitable number fc. This number is set depending on the value of N and computation time required by the user. Higher the value of fc, better the computation time.

6. After increasing value of K, we select K points from the sorted N points that we have so that the K points are as far away from each other as possible.

7. A loop then runs to assign each of the N points to one of the K clusters for which the points' distance from cluster centroid is minimum. This distance should also be less than d. If not then the loop breaks, K value is incremented and process continues from step 6.

8. When all the N points are clustered with a suitable K, such that the points are within d distance from their respective cluster centroids, the clustering is complete.

The clusters finally formed are independent of any administrative boundaries and the statistical analysis is thus dependent on actual aerial distances between schools.

Fig 3: School centres (top left), School cluster centroids (top right), both superimposed. Subset of gujart state is the geographical extent of this research study area.

3.3 Data derivation

Using PostgreSQL commands along with PostGIS extension commands, we derive various data like: number of students in individual class, number of teachers in primary and secondary sections, the facilities available like laboratories, computer centre, playground, child health care centres.

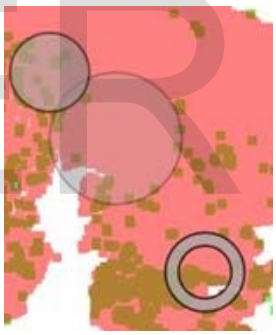
These data are then processed cluster wise to find attributes like number of students per teacher in various primary, upper primary, secondary and higher secondary classes. We find where there is a shortage of teachers and wher there is an excess. Same goes for various other facilities which are looked up using the given school data. The village boundary data and the demographic data of the schools in the region is used to determine the enrollment percentage of the children of various ages within the region. The funding on school facilities can be better utilized if we know what is being spent on which spatial region and by integrating and treating schooling facilities near each other as a cluster.

3.4 Data Merging

Based upon our analysis done in the above steps, we have spatial clusters of school centres near each other. If a school is within the village boundary, lies within the buffer distance from a nearby school and has less number of students/insufficient facilities, it can be merged with the nearby school which has more students/better facilities. After the merging has been done, the databases need to be altered accordingly and the steps are repeated again.

4 Performance Evaluation, Result and Analysis

The result of buffer operations is displayed using Geo-Server as soon as the parameters are received from the user.



INTERSECTION

UNION
503: MEMAR PRIMARY SCHOOL
504: BHAMSARA PRIMARY SCHOOL
505: HASAN NAGAR PRIMARY SCHOOL
506: SANKOD PRIMARY SCHOOL
507: DURAGI PRIMARY SCHOOL
508: MENI PRIMARY SCHOOL
509: DEVATHAL PRIMARY SCHOOL
510: NAVA DEVATHAL PRIMARY SCHOOL
511: AMIPURA PRIMARY SCHOOL
512: PANAR PRIMARY SCHOOL
536: SARANDI PRIMARY SCHOOL
537: LILAPUR PRIMARY SCHOOL
539: VALTHERA PRIMARY SCHOOL
541: LAT KHANPUR PRIMARY SCHOOL
542: JALALPUR GODHNESHWAR
543: NANA UBHADA PRIMARY SCHOOL
553: RAMPURA PRIMARY SCHOOL
554: LILAPUR PRIMARY SCHOOL
555: KAYALA PRIMARY SCHOOL
2512: SHREE PADHAR ASHRAMSHALA
2535: KGBV MERUPAR
2566: JETHLAJ PRI. SCHOOL
2870: VINAY SCIENCE SCHOOL
2871: SHREE SWAMINARAYAN VIDHYAMANDIR
2872: SHREE SWAMINARAYAN GURUKUL INT SCHOOL
2879: SHREE GOVERMENT HIGHER SECONDRY SCHOOL
2880: SAMARTH VIDHYALAY
2883: SHISHU SANSKAR PRIMARY SCHOOL

DIFFERENCE
503: MEMAR PRIMARY SCHOOL
504: BHAMSARA PRIMARY SCHOOL
505: HASAN NAGAR PRIMARY SCHOOL
506: SANKOD PRIMARY SCHOOL
507: DURAGI PRIMARY SCHOOL
508: MENI PRIMARY SCHOOL
509: DEVATHAL PRIMARY SCHOOL
510: NAVA DEVATHAL PRIMARY SCHOOL
511: AMIPURA PRIMARY SCHOOL
512: PANAR PRIMARY SCHOOL
536: SARANDI PRIMARY SCHOOL

Fig 4: Selection of data using set operations on multiple buffers. Intersection is empty as seen in the figure.

The clustering is performed to generate a detailed statistical report. A report formed by processing 697 records of Gandhinagar district schools gives the final K clusters in a processing

time of

2 min, factor =50, final K=300
5 min, factor =30, final K=310
12 min, factor =10, final K=300
Initial K was kept at 100 and buffer distance was 5km.

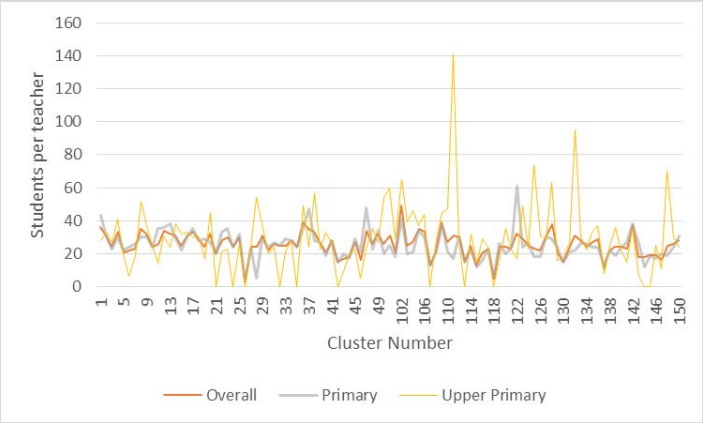


Fig. 5: Students per teacher in clusters 1-50 and 100-150. Sharp jump at 111th cluster shows more teachers required in upper primary at that area

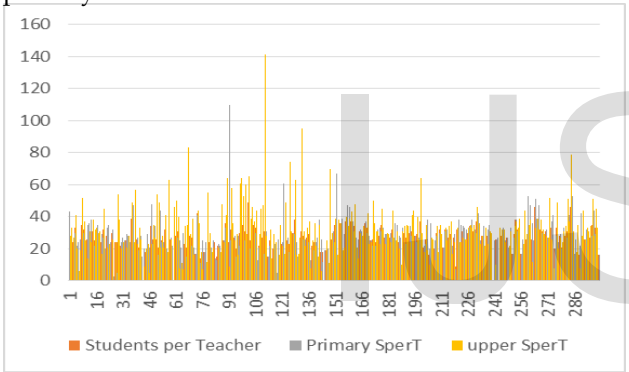


Fig. 6: Students per teacher shown for all clusters. The clusters requiring attention are the ones with huge peaks and trenches.

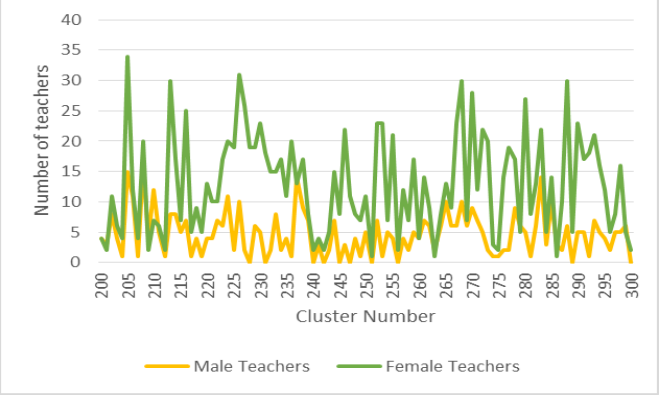


Fig. 7: Male teachers and female teachers are projected in the range of clusters 200-300.

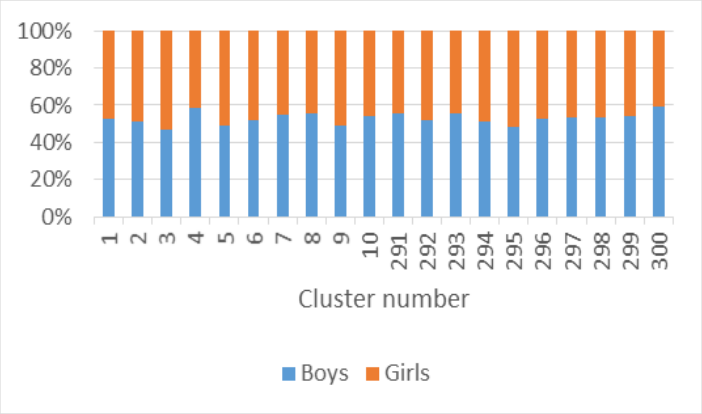


Fig. 8: Percentage of girls vs. boys is projected and shown in first 10 and last 10 clusters

5 Conclusion

With advancement of geospatial data collection and storage technology, geospatial algorithms and data analysis techniques will play a vital role in policy formation and execution. The increased rate of computerization of government departments and digitalization of data presents us an excellent opportunity to develop automated analysis techniques. This paper contributes to these efforts by providing an implementation using the concept of buffers with clustering algorithms. The research can be further extended in this field by optimizing the existential algorithms by using better data structures and filtering noisy data.

Acknowledgment

The authors wish to thank T.P. Singh, Director, Bhaskaracharya Institute of Space Applications and Geo-informatics and BITS, Pilani for their valuable support and inputs in the research work.

References

[1] D.J. Macguire and J. Dangermond, "The Functionality of GIS", url: http://www.wiley.com/legacy/wileychi/gis/Volume1/BB1v1_ch21.pdf

[2] Fernando L. Bacao, "Geospatial Data Mining", ISEGI, New University of Lisbon

[3] J. Han, M. Kamber and A.K.H. Tung, "Spatial Clustering Methods in Data Mining: A Survey," School of Computing Science, Simon Fraser University, Burnaby, Canada

[4] Dr.Chandra.E and Anuradha V. P., "A Survey on Clustering Algorithms for Data in Spatial Database Management Systems," *International Journal of Computer Applications (0975 – 8887)*, Volume 24– No.9, June 2011

[5] D T Pham, S S Dimov and C D Nguyen, " Selection of K in K-means clustering," Manufacturing Engineering Centre, Cardiff University, Cardiff, UK.

[6] Xin Wang and Jing Wang, "Using Clustering Methods in Geospatial Information System," Department of Geomatics Engineering, Schulich School of Engineering, University of Calgary, Calgary, Alberta.