

Stream Data Clustering for Development of Real Time Disease Outbreak Detection System

Thesis submitted to the Andhra University, Visakhapatnam in partial fulfilment of the requirement for the award of *Master of Technology in Remote Sensing and GIS*



Submitted By:

Vineet Kumar

M.Tech (Remote Sensing and GIS)
Geoinformatics Department

Supervised By:

Koti Shiva Reddy

Scientist SD

Geoinformatics Department



**Indian Institute of Remote Sensing, ISRO,
Dept. of Space, Govt. of India Dehradun – 248001
Uttarakhand, India**

June, 2015

CERTIFICATE

This is to certify that the project entitled “**Stream Data Clustering for Development of Real Time Disease Outbreak Detection System**” is a bonafide record of work carried out by Mr. Vineet Kumar. The report has been submitted in partial fulfilment of requirement for the award of Master of Technology in Remote Sensing and GIS in Geoinformatics Department, conducted at Indian Institute of Remote Sensing, Dehradun, during August 18, 2013 to August 16, 2015. The work has been carried out under the supervision of Mr. Koti Shiva Reddy, Scientist/Engineer- ‘SD’, Geoinformatics Department.

Mr. Koti Shiva Reddy
Scientist/Engineer 'SD',
Geoinformatics Department,
Indian Institute of Remote Sensing,
Dehradun

Dr. Sameer Saran
Head,
Geoinformatics Department,
Indian Institute of Remote Sensing,
Dehradun

Dr. S.P.S. Kushwaha
Dean (Academics),
Indian Institute of Remote Sensing,
Dehradun

DISCLAIMER

This document describes work undertaken as part of a program of study at the Indian Institute of Remote Sensing of Indian Space Research Organization, Department of Space, Government of India. All views and opinions expressed therein remain the sole responsibility of the author, and do not necessarily represent those of the Institute.

ACKNOWLEDGEMENTS

Foremost, I would like to express my sincere gratitude to my supervisor, Mr. Koti Shiva Reddy, for his never ending support and motivation during the research phase. He knew when to motivate and when to scold me for getting the best results from the research. His kindness and patience have always encouraged me to ask my doubts and errors many times during the research phase.

I would also like to take this opportunity to thank Dr Senthil Kumar, Director, Indian Institute of Remote Sensing, Dehradun for giving me this opportunity to pursue the course. I also thank Ms. Shefali Aggarwal (Course Director) and Dr S.P.S. Kushwaha (Dean, Academics) for providing the infrastructure and environment to carry out the research work. I take this opportunity for expressing my gratitude towards Dr Sameer Saran, Head, GID for his support and valuable suggestions. I especially thank the staff and faculty at IIRS for their kind support.

I am extremely grateful to the Government of Uttarakhand for providing me the disease data to carry out my Research. I especially thank ISDP unit, Dehradun, officers and staff for their support and knowledge sharing.

I express my heartfelt gratitude to all my classmates to make my stay in IIRS very special and memorable. Special thanks to Rajkumar Sir (my room partner), Kuldeep Sir, Sukant, Rigved, Akshat, Varun, Rohit and Manohar. I especially thank Kavisha, Amol and Raunak with whom sitting next to in lab (dexter lab) made my project work special.

And who, my I am grateful to the almighty for every opportunity and success. Last but most importantly, I owe this to my family and specially my father, who have been greatest source of inspiration and hope in my life. They have always been my strength and their love motivates me to do more.

ABSTRACT

Public health has become a great concern in recent years for developing countries like India. During the recent time, the spreading of deadly disease like Ebola and H1N1(bird flu) has shown the importance for the implementations of outbreak detection systems in order to have the minimum casualties and also the further spreading of diseases should be tracked and controlled by timely implementation of control and preventive measures. In this study we developed a GUI based system based on comparison of space and time complexity analysis of different stream data clustering algorithms. These algorithms are compared with the results obtained from a standard clustering tool SaTScan. After comparison we found the DBSCAN (Density based spatial clustering of applications with noise) is better in time and space complexity than SaTScan which is implemented using python scripting language.

To generate the continuous data, DHIS has been implemented on the local sever and connected to the Postgres database. DHIS results has also been displayed and analyzed for various diseases. The GUI of the system has been developed using PyQt module of python with web view to display the clusters over the base layers. WMS of Google and BHUVAN has been used as base layers. Clusters have been displayed with the help of Geoserver.

Keywords: DHIS, DBSCAN, Stream data clustering, SaTScan.

Table of Contents

List of Figures	4
List Of Tables	6
1. INTRODUCTION	7
1.1 Background	7
1.2 Motivation and Problem Statement.....	9
1.3 Research Objectives.....	9
1.4 Research Questions	10
1.5 Thesis Structure	10
2. Literature Review.....	11
2.1 Disease Outbreak Detection System	11
2.1.1 District Health Information System (DHIS)	11
2.1.2 Real-time Outbreak and Disease Surveillance System (RODS).....	12
2.2 Integrated Disease Surveillance Program (IDSP).....	13
2.3 Clustering.....	14
2.3.2 Scanning Window	16
2.4 SaTScan	16
2.4.1 Monte Carlo Simulation.....	17
2.4.2 Likelihood P-Ratio test	17
2.4.3 Algorithm for SaTScan	18
2.5 Distribution Models	18
2.5.1 Discrete Poisson Model	18
2.5.2 Normal Model.....	18
2.5.3 Ordinal Model.....	19
2.5.4 Exponential Model.....	20
2.5.5 Bernoulli Model.....	20
2.5.6 Multinomial.....	20
2.5.7 Space-Time Permutation Model	21
2.6 Retrospective Analysis.....	21
2.7 Prospective Analysis	21
2.8 Visualization	21

2.9	Stream Data Clustering	22
2.9.1	Clustering Algorithms and Space –Time Analysis	22
3.	Study Area and Datasets Used	26
3.1	Demographic Details of Study Area	26
3.2	Datasets Used.....	27
4.	Methodology	30
4.1	Flow Diagram	30
4.2	Algorithms Analysis	30
4.3	Algorithm Implementation.....	31
4.4	Database Creation	33
4.5	DHIS Implementation for Continuous Data Source	34
4.5.1	Organization Unit Architecture	35
4.5.2	Data Element Architecture.....	36
4.5.3	Datasets Architecture	37
4.5.4	Data Quality	38
4.5.5	Indicators.....	38
4.5.6	Reports	38
4.5.7	GIS Visualizer.....	39
4.6	Methodology To Display Clusters using WMS	39
4.6.1	Google Map Service	39
4.6.2	Bhuvan WMS.....	40
4.6.3	Geosever	41
5.	RESULTS AND DISCUSSIONS	42
5.1	SaTScan Output Analysis	42
5.1.1	Input Parameters Analysis	43
5.1.2	Cluster File.....	44
5.1.3	Data Models	45
5.2	Evaluation of different stream data clustering algorithms	46
5.3	Database Creation through DHIS Implementation	48
5.3.1	Data Entry Form	48
5.3.2	PostGIS Database Connection	49

5.3.3	GIS Visualization.....	51
5.4	Real Time Disease Outbreak Detection System	54
5.4.1	Algorithm Implementation.....	54
5.4.2	Output Comparison	55
5.5	GUI design of disease outbreak detection system.....	56
5.5.1	Displaying Clusters Using Google Maps.....	57
5.5.2	Displaying Clusters Using Geosever	60
5.5.3	Displaying Cluster Using BHUVAN.....	60
6.	CONCLUSIONS AND RECOMMENDATIONS	62
6.1	Conclusions.....	62
6.2	Recommendations.....	63
	References.....	64

List of Figures

Figure 2-1: DHIS surveillance system	12
Figure 2-2: IDSP visualization system	14
Figure 2-3: Normal Distribution Curve	19
Figure 2-4: Ordinal Distribution Curve	19
Figure 2-5: Bernoulli Distribution Curve.....	20
Figure 2-6: Cluster visualization over Google Earth	22
Figure 2-7:CURE Algorithm Processing	23
Figure 3-1:Map showing location of study area	26
Figure 3-2:S-from reporting format	27
Figure 3-3:P-from reporting format	28
Figure 3-4:L-from reporting format.....	29
Figure 4-1:Methodology	30
Figure 4-2:SatScan Methodology	31
Figure 4-3:DBSCAN Flow Chart	32
Figure 4-4:Data Collection Methodology	33
Figure 4-5: Database Design Methodology	34
Figure 4-6:Dashboard User Interface Of DHIS	35
Figure 4-7:Organization Unit Architecture.....	36
Figure 4-8:Data Element Architecture.....	37
Figure 4-9:Datasets Architecture	38
Figure 4-10:GIS Visualizer.....	39
Figure 4-11: Google WMS methodology	40
Figure 4-12:BHUVAN WMS methodology	41
Figure 4-13:Geoserver Publishing methodology	41
Figure 5-1:SaTScan Output of Clusters using Covariates	42
Figure 5-2:Clustering using only one disease data	43
Figure 5-3:Input Parameters File	44
Figure 5-4:Output Cluster File.....	45
Figure 5-5:Data Entry Form.....	49
Figure 5-6:Postgres Database	50
Figure 5-7:GIS Visualization	51
Figure 5-8:GIS Visualization of fever data.....	52
Figure 5-9:GIS Visualization of cough data	53
Figure 5-10:GIS Visualization of pneumonia data	53
Figure 5-11:Algorithm Processing.....	54
Figure 5-12:DBSCAN output	55
Figure 5-13:SaTScan Output cluster.....	55
Figure 5-14:Interface of Developed System	57
Figure 5-15:Clustering Using Google Map	58

Figure 5-16:Clustering Using Google Map at zoom level 4..... 59
Figure 5-17:Clustering Using Google Map at zoom level 59
Figure 5-18:Displaying Clusters Using Geosever 60
Figure 5-19:Displaying Clusters Using BHUVAN 61

List Of Tables

3-1 Block Details	27
5-1:Data model analysis.....	46
5-2:Evaluation of different stream data clustering algorithms.....	47
5-3:Space-time complexity analysis	48
5-4:Table of Comparison	56

1. INTRODUCTION

1.1 Background

Public health weather for a developed nation or a developing nation has become a great concern. During the recent time, the spreading of deadly disease like Ebola and H1N1(bird flu) has shown the importance for the implementations of outbreak detection systems in order to have the minimum casualties and also the further spreading of diseases should be tracked and controlled by timely implementation of control and preventive measures. The best existing way to do the surveillance is spatial clustering of the disease data having the geo-locations of the effected person i.e. the latitude and longitude of the location, type of the disease and population of the area. These parameters are generally fetched from the daily medical reports from the hospitals, emergency response systems and like ambulance dispatch calls etc.

The distribution of infectious disease cases in an area involves various social and demographic factors. These include human population density and their living conditions, sewage and waste management systems ,housing type and location, water supply, land use and irrigation systems, availability and use of various vector borne disease control programs, access to health care units like district hospitals, CHC's and PHC's, and general environmental conditions like pollution. There can be meteorological factors like temperature, humidity, and rainfall patterns that can influence transmission intensity of infectious diseases. Recent cases of diseases like Ebola and swine flu has shown the effect of climate change on virus transmission, swine flu virus intensity became low as the temperature increased.

The Intergovernmental Panel on Climate Change(IPCC) in its 2007 report published that climate change may influence the increase in the areas under risk by infectious diseases such as malaria and may increase the intensity of other vector borne diseases, putting more people at risk. One of the most epidemic diseases that spread widely and quickly is malaria. More than 400-500 million cases of malaria and about 1 million malaria-related deaths occur globally each year. Global revival of malaria has been caused by various factors like high population growth, unavailability of drugs and health infrastructure deterioration. Changes in temperature, rainfall, humidity of different areas may have different effect on people as they have different immunity levels.

As in country like India where a lot of variations are there socially as well as demographically, a system is needed to map the disease cases distribution so to have the pre- knowledge of any disease outbreak in an area. The best way to have the outbreak detection is clustering(Charu C. Aggarwal & Yu, 2008)(Guha S, Meyerson A., 2003). A disease cluster is basically the repeatedly occurring of same disease cases clubbed together

with in a specific time period and geographic area. These cases are reported from their family members and local health units. Some recent disease clusters reported includes Swine Flu(H1N1) case from states like Maharashtra ,Andhra Pradesh and Rajasthan. Another, very recent disease cluster, was the 2003 outbreak of a respiratory illness, later identified as severe acute respiratory syndrome (SARS), caused by a previously unrecognized virus.

Cluster detection has become an important part of the agenda for health system specialist and various public health organizations; the identification of areas having high and low risk is the basic aim of authorities to develop the strategies during the disease outbreak. There are several challenges in investigating the disease clusters as the basic factors like population density and origin of cases may vary time. People who are infected may travel to other places and can infect those who come into the contacts of such. Health system specialist must make sure that the cluster which is under investigating involves one disease (not many). Epidemiologists must make sure that that a suspected exposure of the infected person could have actually triggered disease outbreak, which could be the most casual cause of any disease outbreak happen. To show a disease outbreak in an area the number of infected cases must be greater than the usual or normal cases reported during the same period and same conditions. Then they must ensure that whether the cluster have occurred by chance. Currently, there are different cluster detection techniques used, the most popular are circular window scanning and rectangular window scanning of the studied region. The circular window scanning use the centroid based approach in which it select one case as the centroid and scan the nearby cases in a circular manner, however, when the areas have heterogeneity in populations sizes, scan window methods may lead to inaccurate results. In order to perform cluster detection on heterogeneously populated areas, developed methods are not based on scanning windows but instead on standard mortality ratios (SMR) using irregular spatial aggregation (ISA). Its extension, i.e. irregular spatial aggregation with covariates (ISAC), includes covariates with residuals from Poisson regression.

Clustering algorithms like K-MEAN, BIRCH, COBWEB(C.C. Aggarwal, 2002) are being used in currently implemented disease outbreak detection systems. These algorithms work on data gathered prior to analysis. The basic difference between algorithms is how they process the input data, based on which they are categorized into two types clustering algorithms partitioning and hierarchical algorithms. Partitioning algorithms create the partition of a database D of n objects into a set of k clusters. k is an input parameter for these algorithms, that requires some domain knowledge which unfortunately is not available for many applications. The partitioning algorithm generally starts with an initial partition of D database and then uses an iterative method to optimize an objective function. Each cluster is represented by the gravity center of the cluster (k-means algorithms) or by one of the objects of the cluster located near its center (k-medoid algorithms)(Charu C. Aggarwal, 2003). Consequently, partitioning algorithms follow a two-step procedure. First, it determines k clusters to minimize the objective function. Second, assign each object to the

cluster whose is nearest to the considered object. The second step implies that a partition is equivalent to a voronoi diagram and each cluster is contained in one of the voronoi cells.

An algorithm called CLARANS (Nagesh, Goil, & Choudhary, 2001)(Clustering Large Applications based on RANdomized Search) has also been developed which is an improved k-medoid method. Compared to k-medoid algorithms used, CLARANS is more effective and more efficient. An experimental evaluation indicates that CLARANS runs efficiently on databases of thousands of objects. CLARANS assumes that all objects to be clustered can remain in main memory at the same time which does not hold for large databases. Furthermore, the run time of CLARANS is prohibitive on large databases. Therefore, Ester, present several focusing techniques which address both of these problems by focusing the clustering process on the relevant parts of the database. First, the focus is small enough to be memory resident and second, the run time of CLARANS on the objects of the focus is significantly less than its run time on the whole database.

1.2 Motivation and Problem Statement

Disease outbreak detection systems using clustering algorithms like k-mean, CluStream etc. does not support for continuous evolving data like real time data. Outbreak detected from these systems comes generally with some duration of time gap which create concern, when there is a huge outbreak spreading within a little span of time. To have better and effective response toward controlling epidemic disease, is to have a real time disease outbreak detection system which can perform clustering of the real time stream data directly coming from the ground sources like ASHA workers, village level community health centers(CHC) and district hospitals.

To do the real time stream data processing, algorithms needs number of resources like high performance computing and much important an online source for stream data, which is difficult to generate as in developing countries till now much of the work is not automated. The data collected until it reaches the last level of management who do the clustering on the basis of their experience is in hard copy format or some word document format. So, to have the knowledge of disease outbreak, evaluation of algorithms having optimum space and time complexity is needed and a system to generate stream data is to be developed to have real time stream data clustering

1.3 Research Objectives

- Stream Data Clustering for Development of Real Time Disease Outbreak Detection System.

Sub Objectives

1. Evaluate the different stream data clustering algorithms and comparative analysis from SatScan.
2. Design and implementation of health data database.
3. Development of real time disease outbreak detection system.
4. Design of web based interface for real time disease outbreak.

1.4 Research Questions

1. What are the parameters for comparison of different algorithm results using health benchmark data?
2. How to overcome the limitations of existing disease outbreak detection system to make it real time detection system?
3. What will be the database constraints in designing stream database?
4. What will be the user interface design requirement of disease outbreak detection system?

1.5 Thesis Structure

The whole thesis is divided into six chapters. The first chapter provides general introduction about the research work including problem statement, research objectives and questions. The second chapter deals with literature review in which the related works with respect to the research work are presented. In the third chapter information about the chosen study area and datasets has been used is given. The fourth chapter gives description about the methodology used. Fifth chapter presents the findings of this research work and a detailed discussion on the results obtained. In the sixth chapter a conclusion is drawn based on the results obtained along with some recommendations.

2. Literature Review

2.1 Disease Outbreak Detection System

Disease outbreak detection has been there in the decision making level from the past few decades on the basis of experts knowledge and analysis experience. But because of the recent advance development in the field of technology like high performance hardware, software for processing and visualization has made the detection automated which leads to the development of disease outbreak detection systems. DHIS, RODS are some of the examples of existing systems which generate daily or weekly reports on disease outbreak in an area depending upon the severity of the disease. These systems used different stream data processing clustering algorithms and have different ways to store data streams to analyze trends. Outbreak detection systems are fully automated system, from the time they are connected to a data source it start processing without any human intervention. There are various free available tools which can perform clustering like SaTScan which is discussed later in the thesis, having implementation of different data models and facilities of prospective and retrospective analysis of disease data available from an area(Kulldorff, Heffernan, Hartman, Assunção, & Mostashari, 2005).

2.1.1 District Health Information System (DHIS)

The District Health Information Software (DHIS) is the centralized database management information system for Health Facilities, which provides users the option of data entry, analysis and visualization over GIS based interface. Different versions of DHIS have been developed by different organizations and are being used in many countries. The first version named Hospital Management Information System - HMIS was developed way back in early 1990's with an MS Access based database system and was used in public health department. Later version was developed by AZM, funded by JICA completely based on open source technologies. Current version is developed as web based application for online system by Eycon Pvt. Ltd. and funded by TRF. One of the mostly use DHIS named system is DHIS2 developed by Health Information System Programme (HISP). DHIS2 has overcome all the limitations from previous versions. It is more flexible, high performance, open source health management information system and data warehouse. It was primarily implemented only in the three district of South African city Cape Town, but later with the expansion of HISP network DHIS was implemented in nearly half of the African continent covering almost 300- 400 million population. Some of the countries in Asia too have implemented it. In India state like Himachal Pradesh, Kerala, Maharashtra and some others are using it where they have linked it to the data collected by village level health workers through some mobile based application or data is coded into it with the help of pdf or excel based direct entry.

DHIS covers all aspects from capture of case data or patient data, location from where the data is collected, population of respective locations, facilities like district hospitals, PHC's

Stream Data Clustering for Development of Real Time Disease Outbreak Detection System

and CHC's name and locations, diseases under surveillance, indicators for diseases. It support data collection at any level, that is it provide data entry forms at different administrative hierarchal boundaries with any frequency of data entry hourly, daily or weekly. It also provides customized outputs in the form of reports, charts, graphs and thematic map based results.

Continuous efforts are being made by several institutes to make it more advanced. Department of Informatics at the University of Oslo act as the core development center for it, supported by The Norwegian Research Council, NORAD, The University of Oslo, and The Norwegian Centre for International Cooperation in Education.

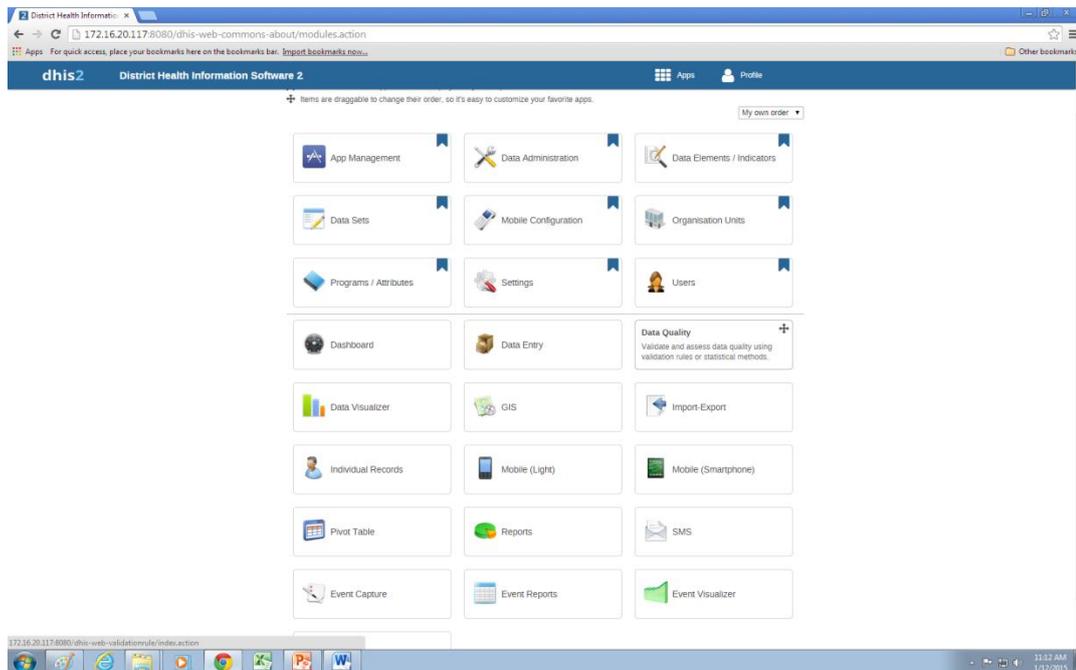


Figure 2-1: DHIS surveillance system

2.1.2 Real-time Outbreak and Disease Surveillance System (RODS)

The Real-time Outbreak and Disease Surveillance System (RODS) is an online web based public health disease outbreak detection and surveillance system developed by the RODS Laboratory situated within the Center for Biomedical Informatics of the University of Pittsburgh. RODS work on the principle that while in common network, systems can share data, the data collected exist in various computer systems in clinical and other settings and display them for public health departments through a secure web-based user interface. The

major advantage of RODS is that it uses the data collected from routine data entry in various departments of hospitals or other health institutions like how many patients are admitted, what vaccination is given to them or with what disease they are suffering from, this make possible to not enter data into the system again and again. RODS monitor the health of the community for disease that have high frequency of occurrence like fever, cough, malaria, and also that which occurs rarely by examining these data. RODS provide temporal and spatial data analysis tools, automatic outbreak detection algorithms that enable public health specialists to detect the presence of a disease outbreak, and support the characterization of that outbreak.

2.2 Integrated Disease Surveillance Program (IDSP)

IDSP is a central public health department funded program started in 2004 used to detect disease outbreaks in all parts of India. Under this program an IDSP unit has to be set up in all the districts of states. The data is collected weekly from various health institutions like primary health centers, community health centers, district hospitals whether public or private and medical colleges. The data are being collected in 'S' forms used by ASHA workers which is syndromic, 'P' forms from the doctors that is probable and 'L' forms from the laboratory. 'S' form basically have the symptoms case data like cold and cough or fever. 'P' forms have the case data of the disease notified from the doctors in patients. The P form data may be positive or negative cases as they are no verified from any laboratory. 'L' form data is the actual number of positive cased of any diseases tested in that local area laboratory. The data collected from all centers collected at district IDSP unit through mails/hard copy from where it is uploaded onto the IDSP portal through web based interface. The weekly data is analyzed by SSU/DSU experts to detect and control an outbreak.

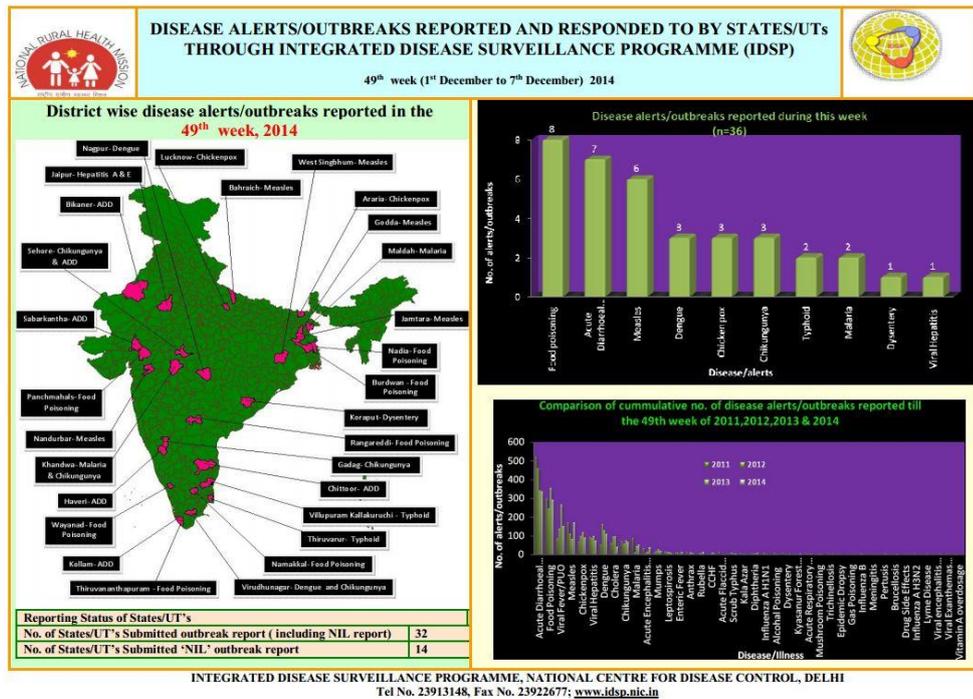


Figure 2-2: IDSP visualization system

2.3 Clustering

Clustering(Charu C. Aggarwal, 2009) is defined as the grouping of objects having some similarity in nature. To have a cluster we should have some parameters defined on the basis of which clusters can be defined like distance between points, points or objects behaviors, there characteristics properties, there space and time variation or there statistical properties.

Disease clustering is defined as the detection of clusters of disease spreading in an area, so to minimize the consequences like high mortality, increase number of infected people and others. With the help of clustering we can do the following:

- Perform the prospective surveillance for early detection of disease outbreak.
- Geographical surveillance of disease to detect the space-time or spatial cluster.
- To do the statistical analysis of clusters.
- To detect whether the disease is continuously or randomly spreaded over space and time.

2.3.1. Clustering Algorithms

2.3.1.1 K-Mean

The k-mean algorithm divides the set of N samples into K number of clusters. Each having a mean of the samples in the clusters. The mean is described as the centroid of the cluster.(Bifet, Holmes, Pfahringer, Kirkby, & Gavalda, 2009);

Steps:

- Select K number of initial centroids (clusters).
- For each point, find its closest centroid and assign it to that one, repeat it for all the points. This will give k number of clusters.
- Recompute the centroids of every cluster until centroid do not change.

2.3.1.2 D-Stream

D-Stream is most widely used stream clustering algorithm (Amini A. et. Al, 2014) and uses an innovative method that bind a decay point to every point density. The decay factor automatically forms the clusters by putting more emphasis on new data points without neglecting old data. Like other clustering algorithms D-Stream does not require prior declarations of number of clusters k. D-stream can found clusters of any arbitrary shape and is more efficient than other algorithms on time and space factors (Yixin C, Li Tu, 2007).

Steps of Algorithm:

1. Initialize an empty hash table grid list $T_c=0$.
2. While data stream is active read records $a=(a_1, a_2, a_3, \dots, x_n)$.
3. Determine the density grid g that contains x .
4. If g is not in grid list than insert it into it and update the characteristics vector of g .
5. If $tc == gap$ then
6. Call initial clustering
7. end if
8. if $tc \bmod gap == 0$ then
9. detect and remove spreaded grids from grid list;
10. call adjust clustering(grid list);
11. end if
12. $tc = tc + 1$;
13. end while
14. end procedure

2.3.1.3 COBWEB

COBWEB is an incremental system for hierarchical conceptual clustering. COBWEB incrementally organizes points into a classification tree. Each node in a classification tree represents a class (concept) and is labeled by a probabilistic concept that summarizes the attribute-value distributions of objects classified under the node. This classification tree can be used to predict missing attributes or the class of a new object.

There are four basic steps COBWEB performing building the classification tree. Which operation is selected depends on the category utility of the classification achieved by applying it. The operations are:

- Merging Two Nodes-Merging two nodes means replacing them by a node whose children is the union of the original nodes' sets of children and which summarizes the attribute-value distributions of all objects classified under them.
- Splitting a node-A node is split by replacing it with its children.
- Inserting a new node-A node is created corresponding to the object being inserted into the tree.
- Passing an object down the hierarchy-Effectively calling the COBWEB algorithm on the object and the subtree rooted in the node (Charu C. Aggarwal, 2009).

2.3.2 Scanning Window

To do the clustering a circular or rectangular scan window(Guha & Munagala, 2009) over the point is used to detect the clusters. The points in the required geographical area are discrete or continuously distributed. In discrete scan statistics the geographical location of the recorded data are non-random and are fixed by the users whereas for continuous distribution the data points are random and are self-generated. For two or more dimensional feature space there are square, circular, rectangle and triangular scanning windows used which shows different dimensions of the data. For example circular scanning window is a cylindrical shape that is to be fitted over the points to show it as the cluster. When is shown in 2-Dimensional system like map on the geographical user interface it appears to be circular. The diameter of the circular window gives the total area covered whereas the height of the cylinder gives us the time dimension of the data.

2.4 SaTScan

SaTScan(Kleinman, Abrams, Kulldorff, & Platt, 2005) is a GUI based tool to analyse spatial, temporal and spacio-temporal data using spatial, temporal and spacio-temporal scan statistics. With SaTScan we can perform the following:

- Perform geographical disease surveillance, to detect spatial and space-time disease clusters.

- Test whether a disease is randomly distributed over space, over time or over space and time.
- Evaluate the quantitative behavior of disease cluster alarms.
- Perform retrospective and prospective analysis for the early warning for disease outbreaks.

SaTScan data types and models and its methodology can be used for both continuous as well as discrete scan statistics. For discrete scan statistics the geographical location where cases has been recorded are user-fixed positions and are non-random in nature, such as houses, schools, hospitals etc. or they can be center locations representing large area such as population weighted centroids of district, state and country. For continuous scan statistics the geographical locations of the cases are not fixed by users rather they can appear anywhere in the area under study.

2.4.1 Monte Carlo Simulation

When a process is repeated continuously over same data points n number of times and test statistics is calculated for each random replications as well as for real datasets, and if the later is among 2 percent higher, then the test is significance at 0.02 level and so on. This is called montecalro hypothesis testing. For any number of Monte Carlo simulations performed, the test is always biased, giving correct significance level that is no prediction and no estimation is done. The number of replications affect the power of test in such a way that more number of simulations give more higher power (Huang, Kulldorff, & Gregorio, 2007).

2.4.2 Likelihood P-Ratio test

Likelihood P-Ratio test gives the p-value that suggest where to put a point that is if there are two clusters nearby a point, than in which of them the point should be merged. During the scan statistics, an alternative hypothesis is that, there may be some elevated risk within the window as compared to outside feature space.

The likelihood ratio function for a specific window is given as

$$D = -2 \ln \left(\frac{\text{likelihood for null model}}{\text{likelihood for alternative model}} \right) \\ = -2 \ln(\text{likelihood for null model}) + 2 \ln(\text{likelihood for alternative model})$$

For the Bernoulli model the likelihood function is $\ln N$, where N is total number of cases and control combined together.

2.4.3 Algorithm for SaTScan

1. Pick one grid point. Calculate the distance for all population points from selected one and sort those in increasing order. After sorting store the population points in an array.
2. Repeat step 1 for all other grid point.
3. Pick a grid point and create a circle taking point as center and continuous increase the radius.
4. For each point entered into the circle, update the event number n and measure $\mu(W)$ of the circular area (W).
5. Repeat step 3 and 4 for each grid point and report the largest likelihood based on μ and $\mu(W)$ pairs.
6. Repeat step 3 to 5 for each monte-carlo simulation (Kleinman et al., 2005).

2.5 Distribution Models

2.5.1 Discrete Poisson Model

It is a distribution related to the probabilities of events which are extremely rare, but which have a large number of independent opportunities for occurrence. The number of persons born blind per year in a large city and the number of deaths by horse kick in an army corps are some examples of discrete Poisson distribution model (Kulldorff et al., 2005).

2.5.2 Normal Model

The normal (or Gaussian) distribution is a very commonly occurring continuous probability distribution—a function that tells the probability that any real observation will fall between any two real limits or real numbers, as the curve approaches zero on either side. A normal distribution is (Kulldorff, Huang, & Konty, 2009)

$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The parameter μ in this definition is the mean or expectation of the distribution (and also its median and mode). The parameter σ is its standard deviation; its variance is therefore σ^2 . A random variable with a Gaussian distribution is said to be normally distributed and is called a normal deviate.

If $\mu = 0$ and $\sigma = 1$, the distribution is called the standard normal distribution or the unit normal distribution, and a random variable with that distribution is a standard normal deviate.

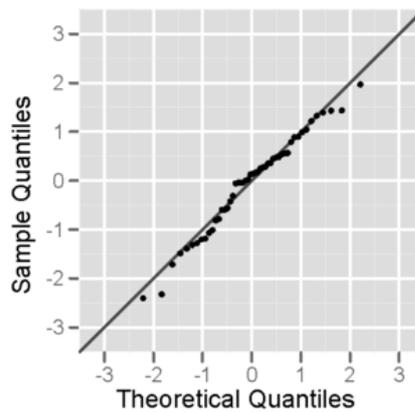


Figure 2-3: Normal Distribution Curve

2.5.3 Ordinal Model

The ordinal logistic model considers a set of dichotomies, one for each possible cut-off of the response categories into two sets, of “high” and “low” responses.

Data: $(Y_i, X_{1i}, \dots, X_{ki})$ for observations $i = 1, \dots, n$, where Y is a response variable with C ordered categories $j = 1, \dots, C$, and probabilities π

$$\pi_j = P(Y = j)$$

X_1, \dots, X_k are k explanatory variables (Kleinman et al., 2005)

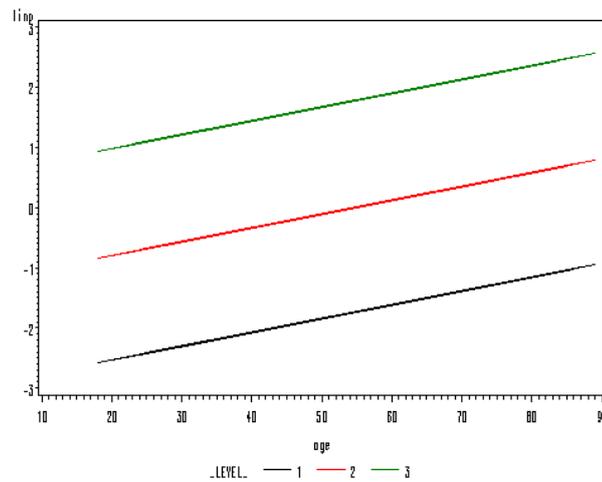


Figure 2-4: Ordinal Distribution Curve

2.5.4 Exponential Model

The exponential distribution is the probability distribution that describes the time between events in a poisson process, i.e. a process in which events occur continuously and independently at a constant average rate. It is the continuous analogue of the geometric distribution, and it has the key property of being memory less (Kulldorff et al., 2005).

2.5.5 Bernoulli Model

This distribution best describes all situations where a "trial" is made resulting in either "success" or "failure," such as when tossing a coin, or when modeling the success or failure of a surgical procedure. The Bernoulli distribution (Jung, Kulldorff, & Richard, 2010) is defined as:

$$f(x) = p^x (1-p)^{1-x}, \quad \text{for } x = 0, 1,$$

where :

P is the probability that a particular event (e.g., success) will occur.

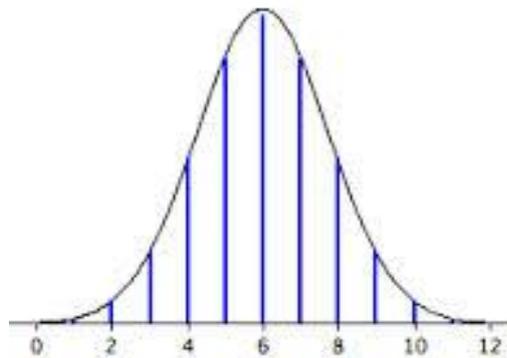


Figure 2-5: Bernoulli Distribution Curve

2.5.6 Multinomial

In a multinomial distribution, the analog of the Bernoulli distribution is the categorical distribution, where each trial results in exactly one of some fixed finite number k possible outcomes, with probabilities p_1, \dots, p_k (so that $p_i \geq 0$ for $i = 1, \dots, k$ and $\sum p_i = 1$), and there are n independent trials. Then if the random variables X_i indicate the number of times outcome number i is observed over the n trials, the vector $X = (X_1, \dots, X_k)$ follows a multinomial distribution with parameters n and p , where $p = (p_1, \dots, p_k)$ (Jung et al., 2010).

2.5.7 Space-Time Permutation Model

The space–time permutation scan statistic utilizes thousands or millions of overlapping cylinders to define the scanning window, each being a possible candidate for an outbreak. The circular base represents the geographical area of the potential outbreak(Tango, Takahashi, & Kohriyama, 2011).

2.6 Retrospective Analysis

A retrospective cohort (Charu C. Aggarwal et al., 2003) study deals with the detection of outbreaks that has already occurred in past. It is generally a post occurrence analysis. In it we have number of cases occurred in an area on some particular time interval, which shows the outbreak already occurred and have the ground truth values associated with it. Retrospective analysis is good when we want to look up about the comparison between outbreaks conditions happened in two or more successive or random years. The main advantages this study is, it is conducted on a small scale, and requires less time to complete. It is also useful to address diseases which occur rarely and have low incidences of occurring.

2.7 Prospective Analysis

Prospective analysis(Jones, Liberatore, Fernandez, & Gerber, 2006) is prediction based analysis. In this type of analysis we have some case data already occurred and we want to predict how it will change in future time. As we are doing the prediction we can only have purely temporal and spacio- temporal analysis. For example if there is an outbreak occurred in an area which is increasing with time and space and we want to predict how much will it spread more, we do the prospective analysis Prospective analysis is influenced by nay external factors like climatic factors temperature, precipitation, rainfall and human factors like population density, pollution, migrants population. Prospective study is conducted on a large scale and it does not give proper analysis for incidences diseases.

2.8 Visualization

Visualization is the most important step in data analysis and results presentation. With the help of visualization it make possible to have relationship between raw and processed data and also to have better understanding of pattern discovered. Various tools are available to visualize the clusters like QGIS, BHUVAN which are open source and in commercial ones like ArcGIS, Mapbox. We can also visualize the clusters by publishing them through Geosever.

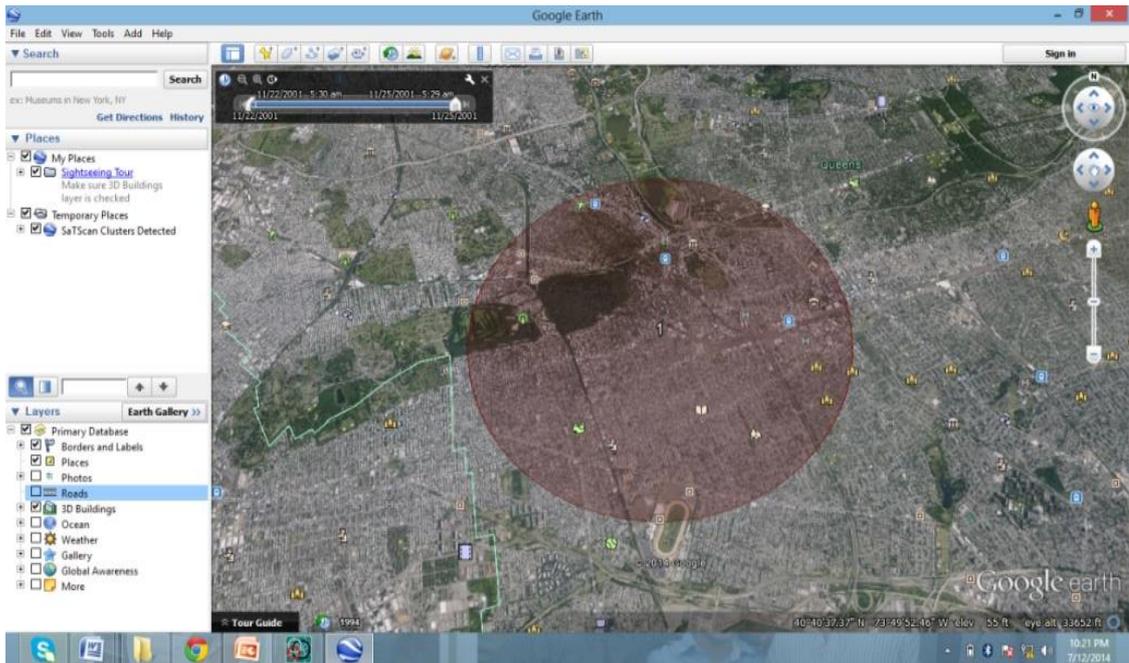


Figure 2-6: Cluster visualization over Google Earth

2.9 Stream Data Clustering

In recent years, clustering has been evolved extensively and studied a great with new algorithms came into existence(Charu C. Aggarwal & Yu, 2008).These algorithms makes grouping of data points based on some similarities into a cluster which are different from the one, in another cluster. At the same time, data stream concept has also been evolved, which van have periodic information about the person localization, therefore can be useful in many aspects of the real world. Combining these two gives us the clustering data streams, which is rather a difficult task (Cao, Liang, Bai, Zhao, & Dang, 2010) as traditional clustering algorithms are inefficient to work with data stream. In recent years only one pass algorithms has been developed which are efficient to only limited data but when data evolved largely over time the quality of the algorithm get diminished. Therefore to get better cluster over time window exploring clustering data stream algorithm is needed(Guha S, Meyerson A, Mishra N. (2003).

2.9.1 Clustering Algorithms and Space –Time Analysis

2.9.1.1 CURE

CURE algorithm is same as that to hierarchical clustering approach, the only difference is rather than taking every point in the cluster it takes only sample points variants as cluster representatives.

Steps:

1. Select required target of sample number “c”.
2. Now select “c” sample points which are well scattered from the cluster.
3. The selected sample points are bound tightly toward the centroid in fraction of time α where $0 \leq \alpha \leq 1$.
4. These points are used as cluster representatives and are used in calculation for d_{min} cluster merging approach.
5. After completion of each merging, c sample points from original cluster representation will be selected to represent new cluster.
6. Process of merging will continue until target k clusters are formed(Chen & Liu, 2009).

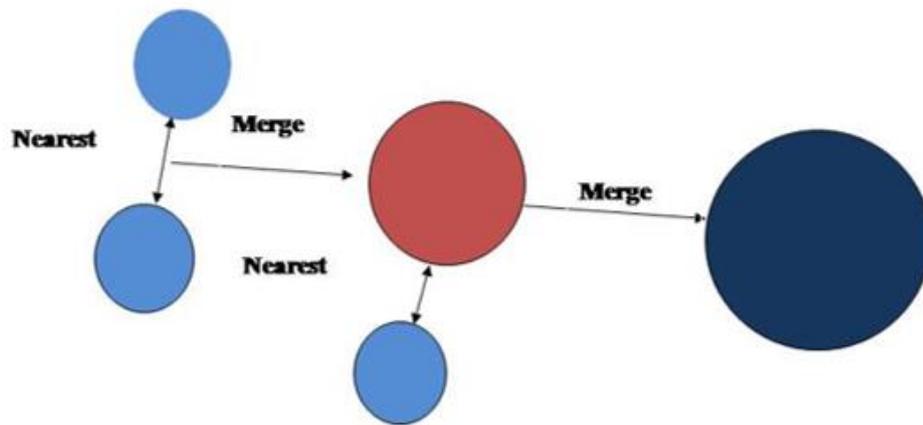


Figure 2-7: CURE Algorithm Processing

Analysis

- The worst-case time complexity is $O(n^2 \log n)$
- The space complexity is $O(n)$ due to the use of k-d tree and heap.

2.9.1.2 DBSCAN

Density-Based Spatial Clustering of Applications with Noise (DBSCAN)(Cormode & McGregor, 2008) is the most widely used density based algorithm giving non-linear shape structure based on point density. Density reachability and density connectivity are two concepts which govern this algorithm.

Density Reachability - A point "a" is said to be density reachable from a point "b" if the point "a" is within distance ϵ from point "b" and "b" has sufficient number of points in its neighborhood who are at a distance ϵ from it.

Density Connectivity - Two points "r" and "s" are said to be density connected if there exist a third point "t" which is within a distance ϵ from points r and s and have sufficient number of points in its neighborhood within ϵ distance. For ex- if "a" is a neighbor of b and b is the neighbor of c and c is the neighbor of d, then by density connectivity a is the neighbor of d.

Algorithmic steps for DBSCAN clustering

Let $P = \{p_1, p_2, p_3, \dots, p_n\}$ be the set of data points, ϵ (eps) is the distance and minPts (minimum number of points required to form a cluster).

- 1) Start with an random point that has not been visited.
- 2) Find the neighborhood of this point which are within the ϵ distance from it.
- 3) If there are sufficient points around the neighborhood of the selected point then the clustering will start and the point will be marked as visited otherwise the point will be marked as noise.
- 4) If a point is already part of a cluster than same procedure will be repeated from step 2 until all points in the cluster are determined.
- 5) Every new point visited will lead to a cluster formation or noise.
- 6) Process completes when all the points are marked as visited.

Analysis:

- $O(n)$ Space Complexity where n in number of data points.
- Using KD Trees the overall Time Complexity reduces to $O(n * \log n)$ from $O(n^2)$.

Advantages

- 1) Does not require setting of any target cluster number.
- 2) Able to identify points and noise.
- 3) Able to generate indefinite size and indefinite shape clusters.

Disadvantages

- 1) Algorithm fails in case of non-homogeneous density clusters.
- 2) Do not work for multi-dimensional datasets.

2.9.1.3 STREAM

STREAM algorithm use partitioning methodology for sample points and then do the clustering using bottom-up approach similar to that of CURE (Forestiero, Pizzuti, & Spezzano, 2009). However CURE generate the clusters of non-homogeneous type where STREAM generate the provably good clusters.

Algorithm Small-Space(S)

1. Select a sample space(S) and divide it into n disjoint pieces $p_1, p_2, p_3, \dots, p_n$.
2. For each point, find center $O(k)$ in P_i and assign in P_i to closest center.
3. Let P^* be the total $O(nk)$ centers obtained where each center is weighted by the total number of points assigned to it.
4. Cluster P^* to generate k clusters.

3. Study Area and Datasets Used

3.1 Demographic Details of Study Area

The study area Dehradun falls under the Uttarakhand state of India. It is the capital city lies between 30°18'56.52" North latitude and 78°21'30.96" East latitude and covers an area of about 3088 km². According to 2011 census it has a total population of 1,696,694. The area is divided into six health blocks namely Chakrata, Raipur, Kalsi, Doiwala, Sahaspur and Vikasnagar. The total numbers of health facilities in the district are about 240 situated at district level, block level and at village level. The study is carried out by the data collected at reporting units, which act as data collection centers for 5-6 villages at rural areas and 1-2 wards at urban areas.

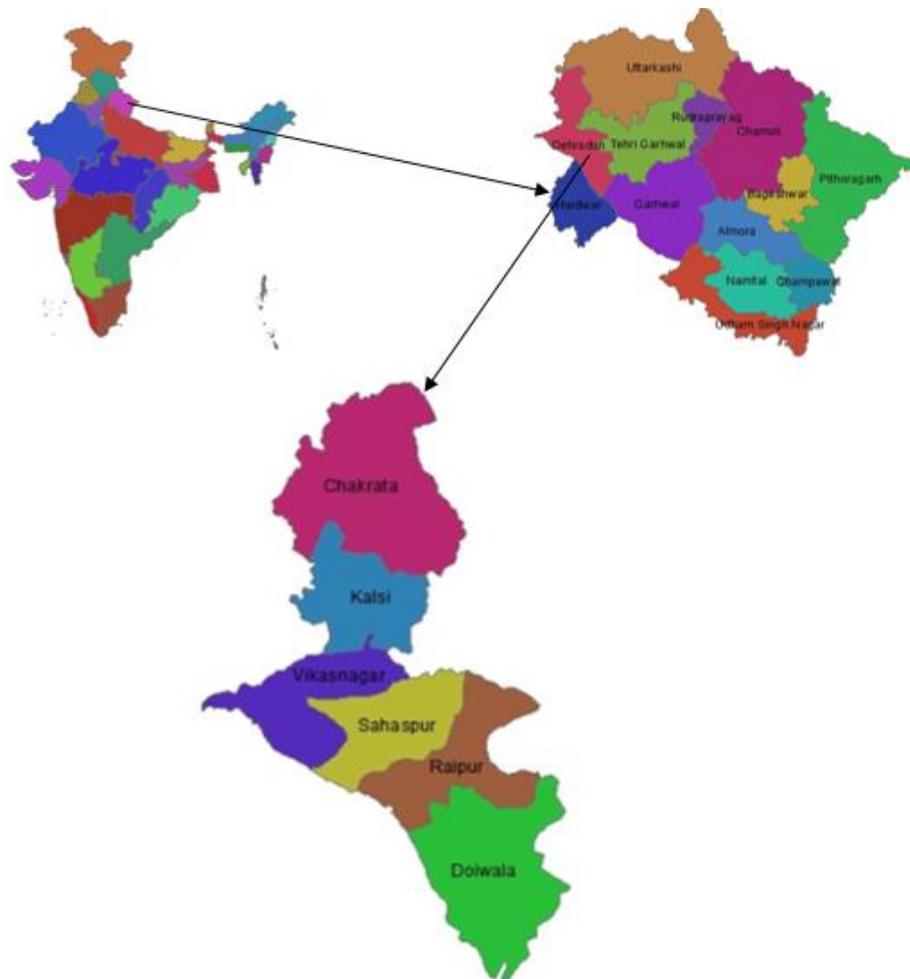


Figure 3-1:Map showing location of study area

3-1 Block Details

Block name	Population	Number of Reporting Units		
		S-form	P-form	L-form
Chakrata	83500	12	11	0
Kalsi	47329	20	18	2
Doiwala	260000	9	16	6
Raipur	578420	27	28	20
Sahaspur	409587	19	17	10
Vikasnagar	317482	16	7	5

3.2 Datasets Used

S-form – S-form dataset is the basic symptoms data collected by health workers like ASHA and ANM at village level. The data is collected on daily basis but the reporting to the IDSP unit is done weekly. The major symptoms checked in this form are severity of fever, cough, dehydration, acute jaundice cases and acute flaccid paralysis.

The screenshot shows an Excel spreadsheet titled "Statement 4:- Block Wise Weekly Surveillance Report on the basis of District : Dehradun" for the period 10/11/2014 to 16/11/2014. The report is organized into columns for various symptoms and reporting units. The columns include: S.No., BLOCK, Name of Reporting Unit, Fever less than 7 days (subdivided into Only Fever, With Rash, With Bleeding, With Daze or Semiconsciousness/Unconsciousness, and Fever more than 7 days), Cough with or without fever (subdivided into <3 Weeks and >3 weeks), Loose Watery Stools < 2 weeks (subdivided into With some / much dehydration, With no dehydration, and With blood in stool), Jaundice cases of less than 4 weeks (cases of Acute Jaundice), and Acute Flaccid Paralysis (cases in less than 4 years of age (cases of Acute Flaccid Paralysis)).

S.No.	BLOCK	Name of Reporting Unit	Fever					Cough with or without fever		Loose Watery Stools < 2 weeks			Jaundice cases of less than 4 weeks (cases of Acute Jaundice)	Acute Flaccid Paralysis (cases in less than 4 years of age (cases of Acute Flaccid Paralysis))
			Fever less than 7 days					<3 Weeks	>3 weeks	With some / much dehydration	With no dehydration	With blood in stool		
			Only Fever	With Rash	With Bleeding	With Daze or Semiconsciousness/Unconsciousness	Fever more than 7 days							
1	CHAKRATA	Viala	2								1			
2		Barontha	2					1			2			
3		Dasau	0											
4		Kwaru	2					1			2			
5		Kwansi	3					2			1			
6		Sawra	2					2			2			
7		Mairawana	3					1			3			
8		Buraswa	3					1			3			
9		Jahdi	0								1			
10		Mohana	3					2			1			
11	Tyuni													
12	Manthat													
13	Doiwala		10											
14	DOIWALA	Bhaniawala	3					2			2			
15		Raiwala	3											
16		Baleswala	2								2			
17		Chedderwala	0					1						

Figure 3-2:S-from reporting format

Stream Data Clustering for Development of Real Time Disease Outbreak Detection System

L-form - L-form dataset is the disease cases referred from doctors and tested at laboratory located at the block level. The form contains the total number of samples tested and number of cases found positive. The laboratory testing is done for total of 18 diseases, major like typhoid, malaria, Dengue, HCV, VDRL, sputum, hepatitis etc.

SI No		BLOCK	Name of RU's	Dengue/DHF/DSS		Chikungunya		JE		Meningococcal Meningitis		Typhoid Fever		Diphtheria		Cholera		Shigella Dysentery		Viral Hepatitis A		Viral Hepatitis E	
				No. of samples test	No. found Positive	No. of samples test	No. found Positive	No. of samples test	No. found Positive	No. of samples test	No. found Positive	No. of samples test	No. found Positive	No. of samples test	No. found Positive	No. of samples test	No. found Positive	No. of samples test	No. found Positive	No. of samples test	No. found Positive	No. of samples test	
1	2	Chakrata	CHC Chakrata									15	2										
3	4	Doiwala	CHC Doiwala									50	9										
5	6		SPS Rishikesh	4	0							8	2										
7	8		Nirmal Ashram Hospital	4	0																		
9	10		HIHT, Jollygrant	32	5				4	0	128	25					2	0	4	2	4		
11	12	Kalsi	Dabral Pathology	1	1							21	1										
13	14		Prabhu Pathology Centre									5	0										
15	16		PHC Kalsi										9	3									
17	18		CHC Sahiya										13	3									
19	20		Doon Hospital	225	13							195	14										
21	22		Coronation Hospital									59	21										
23	24		Doon Female Hospital									12	4										

Figure 3-4:L-form reporting format

4. Methodology

4.1 Flow Diagram

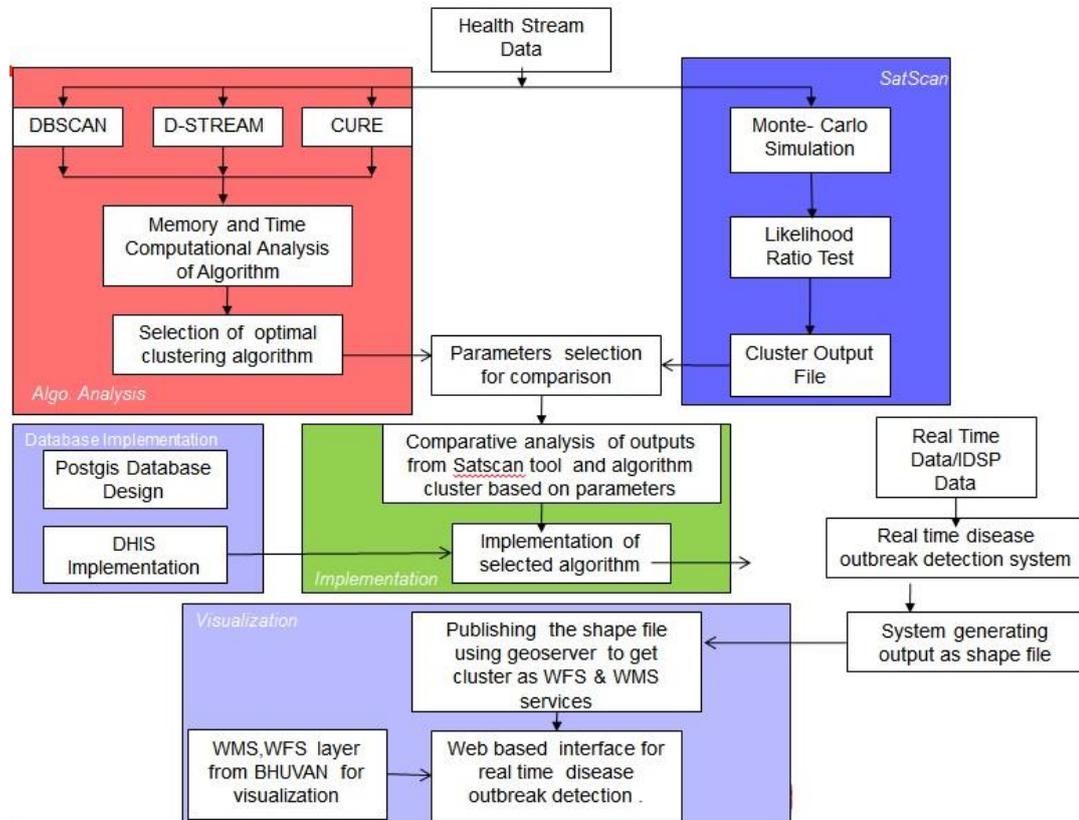


Figure 4-1:Methodology

The overall methodology for the project implementation is divided into four parts namely algorithms analysis, algorithm implementation, database creation and visualization. Each part covers one objective and corresponding research question. The flow diagram shows the required operations to be performed to reach the execution of each phase of project.

4.2 Algorithms Analysis

Algorithms analysis means the theoretical and practical, space and time complexity analysis of various stream data clustering algorithms with the help of specific tool and by literature review. The tool used for cluster algorithm analysis is SaTScan.

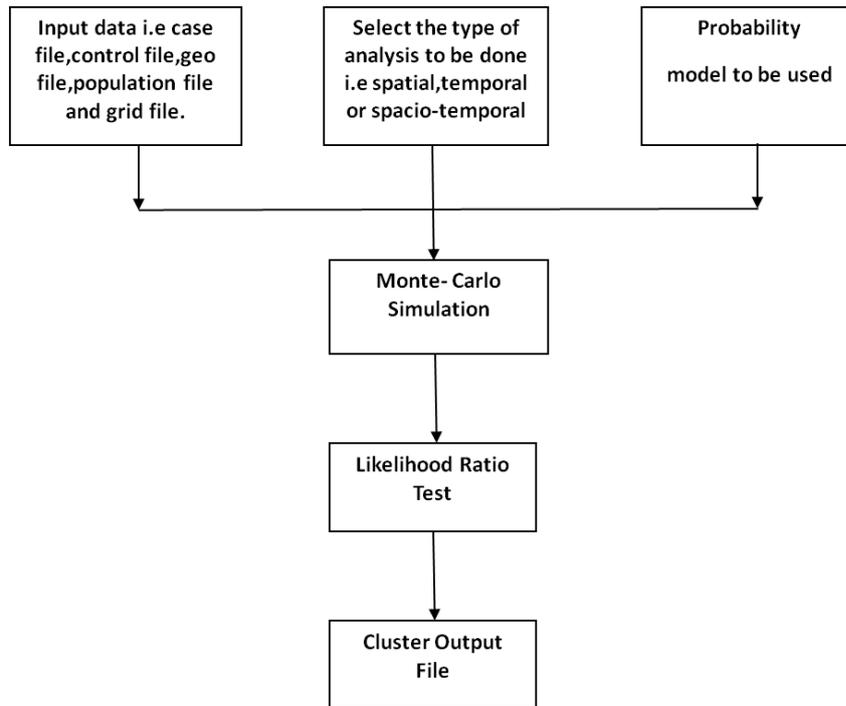


Figure 4-2:SatScan Methodology

The flow diagram for cluster algorithms analysis requires the health data having the geolocations of the area where disease cases are found, number of cases in an area total population of the area and the cases type. The required file in given as input then the type of analysis is given that is spatial, temporal or spatio-temporal is to be performed. The third part is to choose the specific data model for the analysis from Bernoulli, Multinomial, Ordinal, Exponential, Space-time permutation, Discrete Poisson and Normal. After giving all the inputs the Monte-Carlo simulation will start. For each monte-carlo simulation likelihood test ratio is calculated and clusters are formed after all the simulation gets completed.

The required files generated from the cluster formation give space that is the memory used and time taken by algorithms to generate the clusters. This space-time complexity analysis gives us the optimum algorithm that is to be implemented for real time disease outbreak detection system.

Detailed comparison between the SaTScan and the implemented algorithm DBSCAN has been discussed in the result chapter.

4.3 Algorithm Implementation

From the space time complexity analysis of different stream data clustering algorithms we found Density-Based Spatial Clustering of Applications with Noise (DBSCAN) most

optimum to implement with $O(n)$ space complexity using KD Trees the overall time complexity reduces to $O(n * \log n)$ from $O(n^2)$.

Algorithmic steps for DBSCAN clustering

Let $P = \{p_1, p_2, p_3, \dots, p_n\}$ be the set of data points, ϵ (eps) is the distance and minPts (minimum number of points required to form a cluster).

- 1) Start with an random point that has not been visited.
- 2) Find the neighborhood of this point which are within the ϵ distance from it.
- 3) If there are sufficient points around the neighborhood of the selected point then the clustering will start and the point will be marked as visited otherwise the point will be marked as noise.
- 4) If a point is already part of a cluster than same procedure will be repeated from step 2 until all points in the cluster are determined.
- 5) Every new point visited will lead to a cluster formation or noise.
- 6) Process completes when all the points are marked as visited.

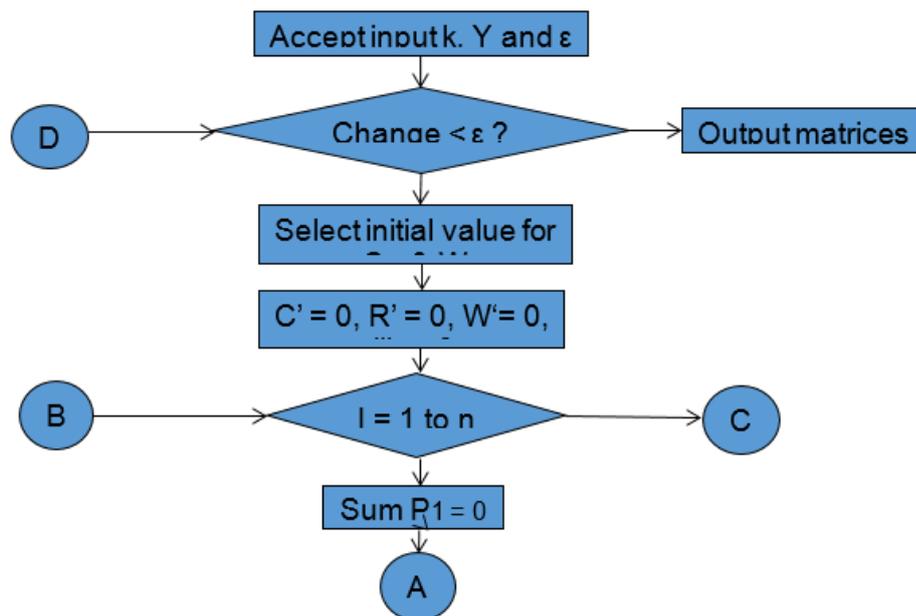


Figure 4-3:DBSCAN Flow Chart

Complexity Analysis:

1. Time complexity of DBSCSAN is $O(n)$.
2. Space complexity analysis of the algorithm is $O(n * \log n)$.

Where n = number data points

4.4 Database Creation

The methodology to create a continuous database starts with how we are generating the data and what the various sources of data. The health data collected comes from three different sources. The S-form data means the data collected by the ASHA workers in the villages from household reports. The data then collected at one reporting unit under which defined number of villages comes. The P-form data comes from the doctors present at various reporting units and L-form data is collected from the laboratory reports. The laboratory data gives us the actual number of persons suffering from the disease.

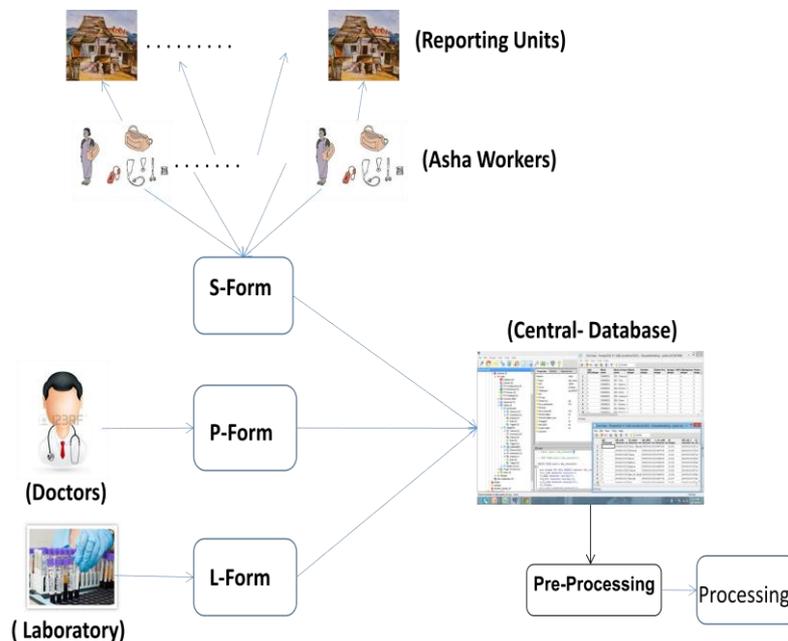


Figure 4-4:Data Collection Methodology

The database design methodology shows the normalization of data collected in s-form, p-form and L-form. The data than collected is entered into postgresql database and the spatial reference table is made for the shape file or base layer visualization.

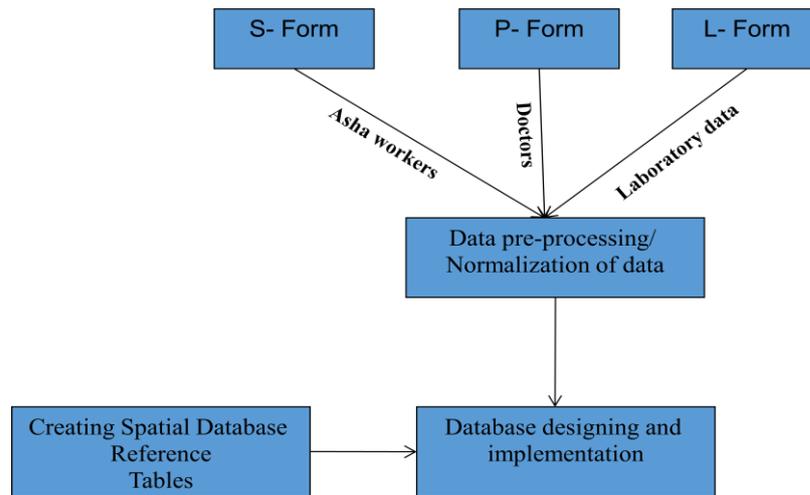


Figure 4-5: Database Design Methodology

4.5 DHIS Implementation for Continuous Data Source

DHIS is a desktop and also web based software package for collection, analysis, and visualization of case data collected for integrated health information management activities. The software package is a fully functional tool rather than only a data collection mechanism. It is having a user friendly and customized user interface that gives user the advantage to design the GIS based health information systems without any programming. Latest versions of DHIS are completely web based software packages developed with free and open source Java framework.

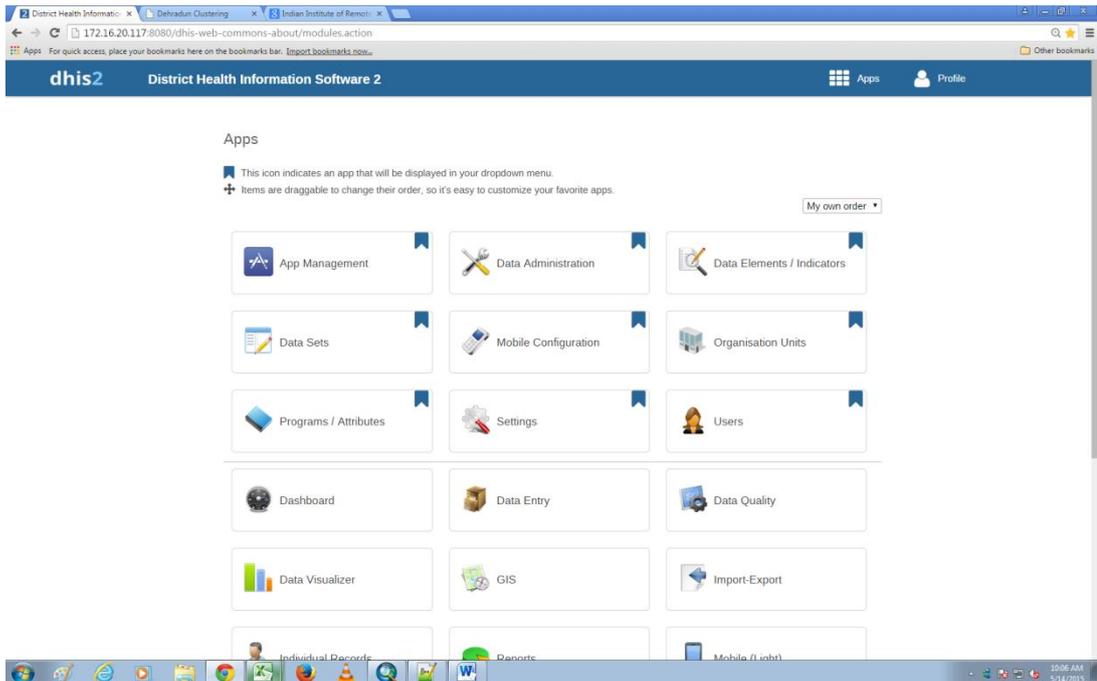


Figure 4-6:Dashboard User Interface Of DHIS

4.5.1 Organization Unit Architecture

The organizational units defines the health facilities like district hospitals, community health centers etc and also the administrative boundaries of the areas for which the data collection and analysis has to be done. Organization units follow a hierarchy that defines under which organization unit comes and so on, for example the hierarchy starting from state than under state there will be districts than blocks, that reporting facility (CHC or PHC). We can visualize organization units on DHIS GIS after converting shape files to gml with the help of ogr library.

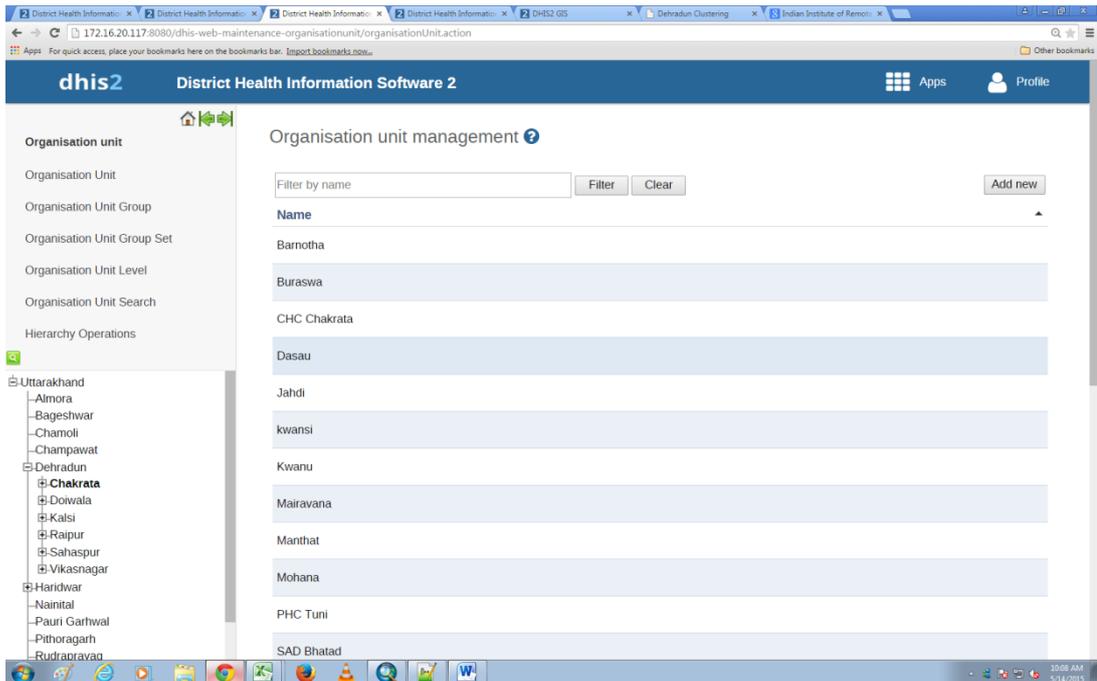


Figure 4-7: Organization Unit Architecture

4.5.2 Data Element Architecture

The Data Element can be called as the most important part of DHIS 2 database. Data elements represent what is to be collected and analyzed like malaria cases, fever or dengue etc. Different data sets represent different data elements. S-form data sets represents data elements showing basic symptoms of the disease like fever, dehydration, diarrhea etc., p-form represents data elements as disease names like chicken pox, cholera, malaria, dengue etc and L-form represents data elements in two parts positive or negative that is whether a cases found in lab testing in in positive or negative. We can have any number of data elements in data sets.

Stream Data Clustering for Development of Real Time Disease Outbreak Detection System

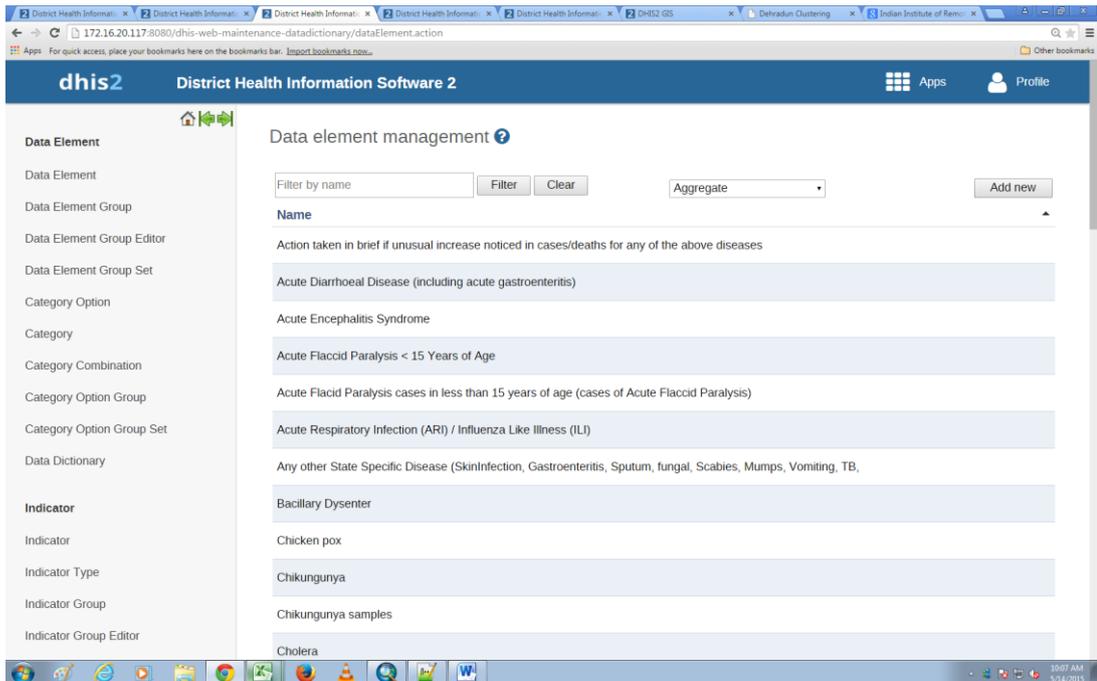


Figure 4-8:Data Element Architecture

4.5.3 Datasets Architecture

All data entry in DHIS is done through the forms created in datasets having the required fields for data elements. Data entered manually or through mobile interface directly goes to these datasets. Every data entry form in data sets is assigned to some organization unit. Many organization units can have same data entry forms. A dataset has a time period which shows how the data is collected, it can be daily, weekly, monthly, and quarterly, six-monthly, or yearly. These constraints are defined by the user as per the requirements. Once the data set has been assigned to an organization unit the dataset form will appear in the Data Entry component of DHIS.

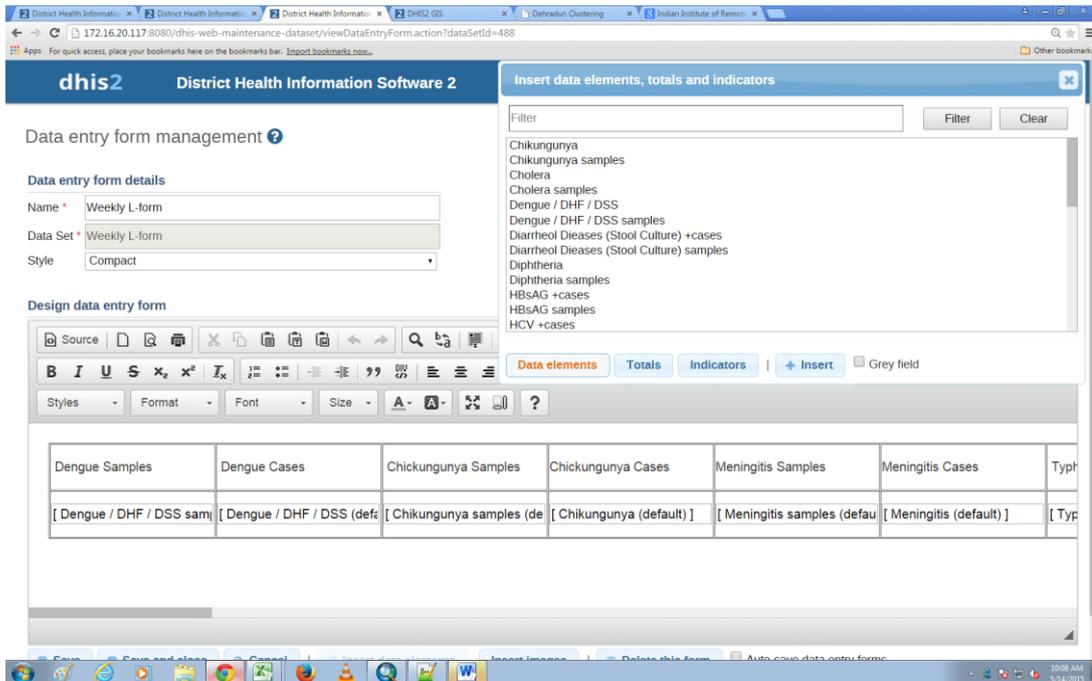


Figure 4-9: Datasets Architecture

4.5.4 Data Quality

Once the data entry form has been created and data is entered into the database, there comes a point of data quality. A data quality mechanism is set to check redundancy, incompleteness and feasibility. Rules are made to check to safeguard quality of data.

4.5.5 Indicators

Indicators are the most important part of analysis, however data elements give the numbers of infected people indicators shows the results. They are values based on ratios determined from numerators and denominators like children malnourished per thousand or sex ratios (1:100,1000,10000). Some of the common used indicators are ANC checking's and neonatal deaths per thousands.

4.5.6 Reports

DHIS gives the functionality of preparing reports of database analysis. It has features of preparing pie-charts, simple graphs, line graphs, bar graphs, scatter plots and the maps. It give the flexibility of having multiple datasets and multiple time interval data reports presentation. We can have monthly, quarterly, half yearly or any time interval reports generated form DHIS.

4.5.7 GIS Visualizer

DHIS have the integrated GIS module with it. We can display areas to be under study as polygons and health facilities under the area as point. Polygons shape files first needed to be converted into xml format by ogr library and then loaded into it. Points features can be directly loaded by entering the latitude and longitude of the location. We can also have buffers around the points to shows how much area is covered by the facilities.

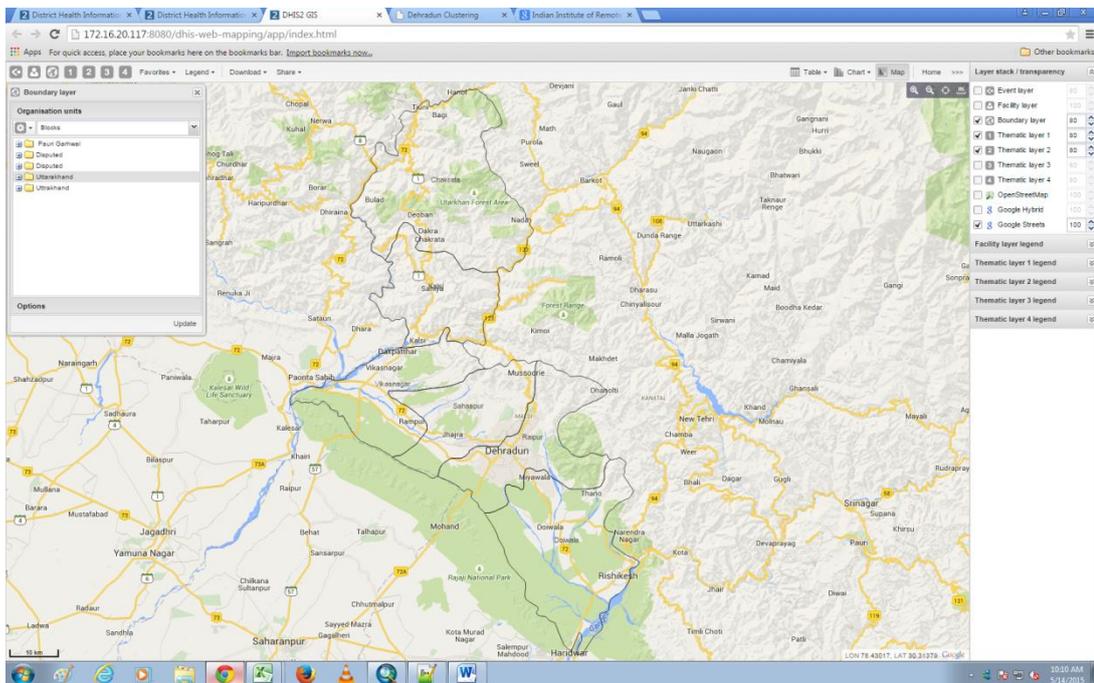


Figure 4-10:GIS Visualizer

4.6 Methodology To Display Clusters using WMS

4.6.1 Google Map Service

To use google map service ,googlewms layer has been used as base layer. Google WMS uses the Google API architecture. On clicking the check box button on GUI Wms will be displayed on the web view with the cluster points case data on it. As base layer we can also use google open street map. Fig. hows how the wmslaer of google has beed used..the client connect to the googleapi engine with the help of html, css and javascript code.

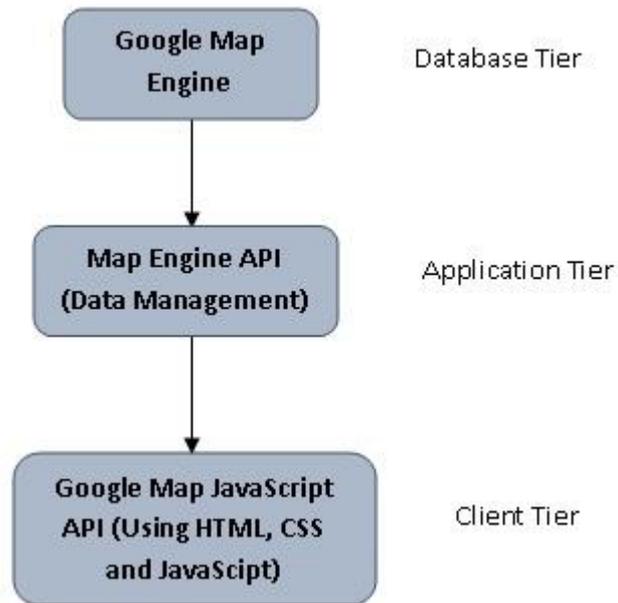


Figure 4-11: Google WMS methodology

4.6.2 Bhuvan WMS

Bhuvanwms service can be used by getting the url of the layer which we want to display as base layer. Its layer display code contains the extent of the layer, format as well as the EPSG code.

```
URL - "http://bhuvan5.nrsc.gov.in/bhuvan/gwc/service/wmts/"  
version - 1.0.0  
layer - "vector:UK_LULC50K_0506"  
matrixSet- "EPSG:4326"  
tileFullExtent :77.575,28.715,81.043,31.467,  
format: "image/png"
```

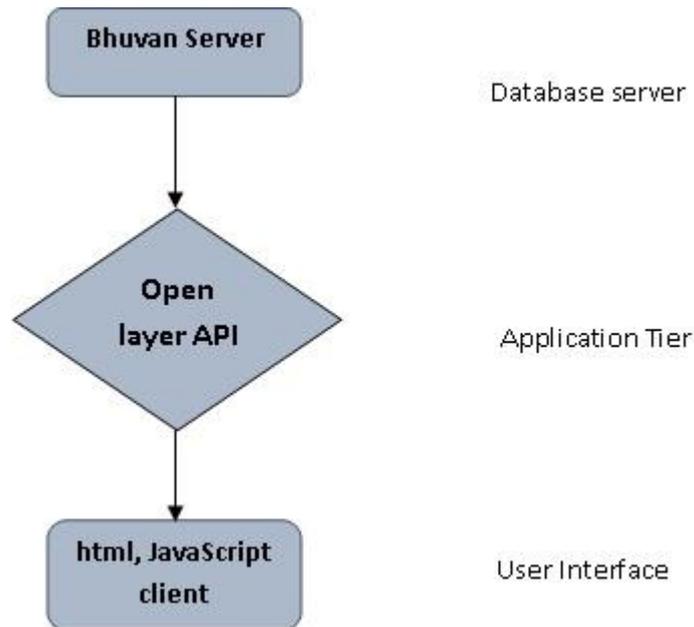


Figure 4-12: BHUVAN WMS methodology

4.6.3 Geosever

Geoserver architecture include the backend data source as PostGIS database containing the shape files to be published in the geoserver. The PostGIS is connected to the application server of geoserver .

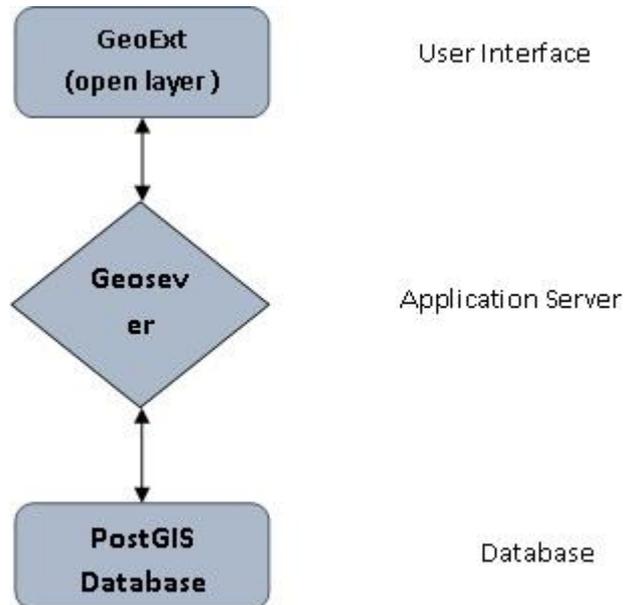


Figure 4-13: Geoserver Publishing methodology

5. RESULTS AND DISCUSSIONS

5.1 SaTScan Output Analysis

Satscan output are the kml and shape files which can be viewed on any gis viewer. Here we used mapbox to display them as it have the multi layer architecture and can have base layers of google maps, satellite images or open layer maps. Data used for the scan statistics have the latitude longitude positions, number of cases occurred in that area and the population of theta area.

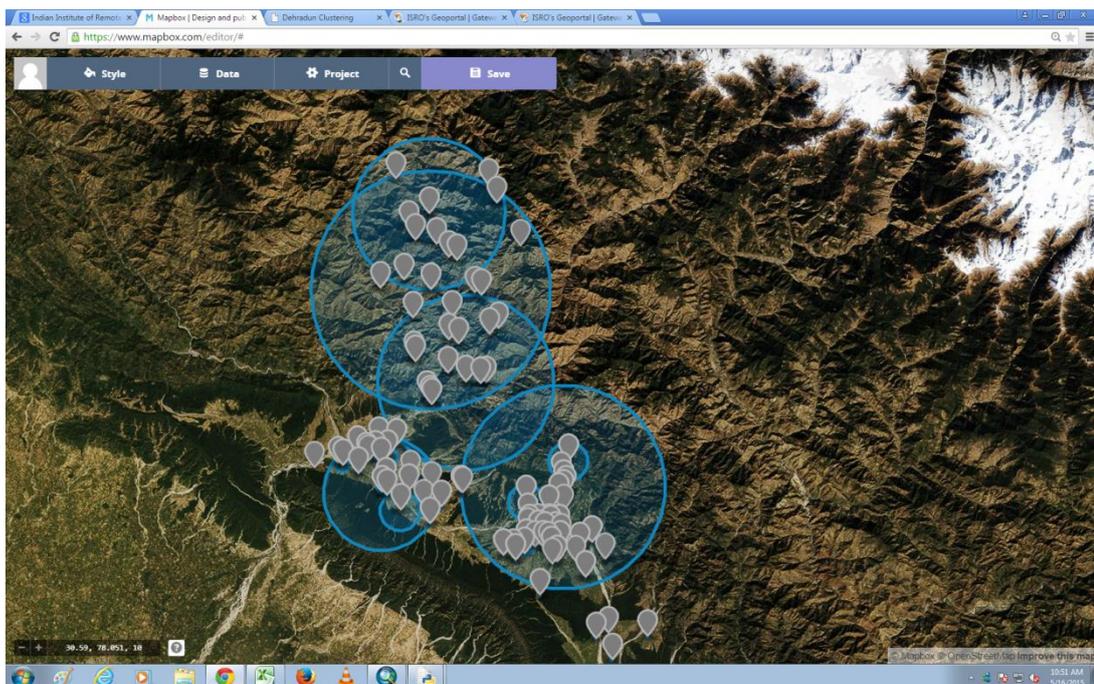


Figure 5-1:SaTScan Output of Clusters using Covariates

We can also get the clusters of more than one disease at same time by taking one as cases and other as covariates. Covariates are the secondary cases that we are taking during clustering to show multiple disease cluster at once. We can have number of covariates depending on the case data we have. For example for P-Form we can take malaria as major case disease and other like chicken pox, dengue, diphtheria, measles etc as covariates.

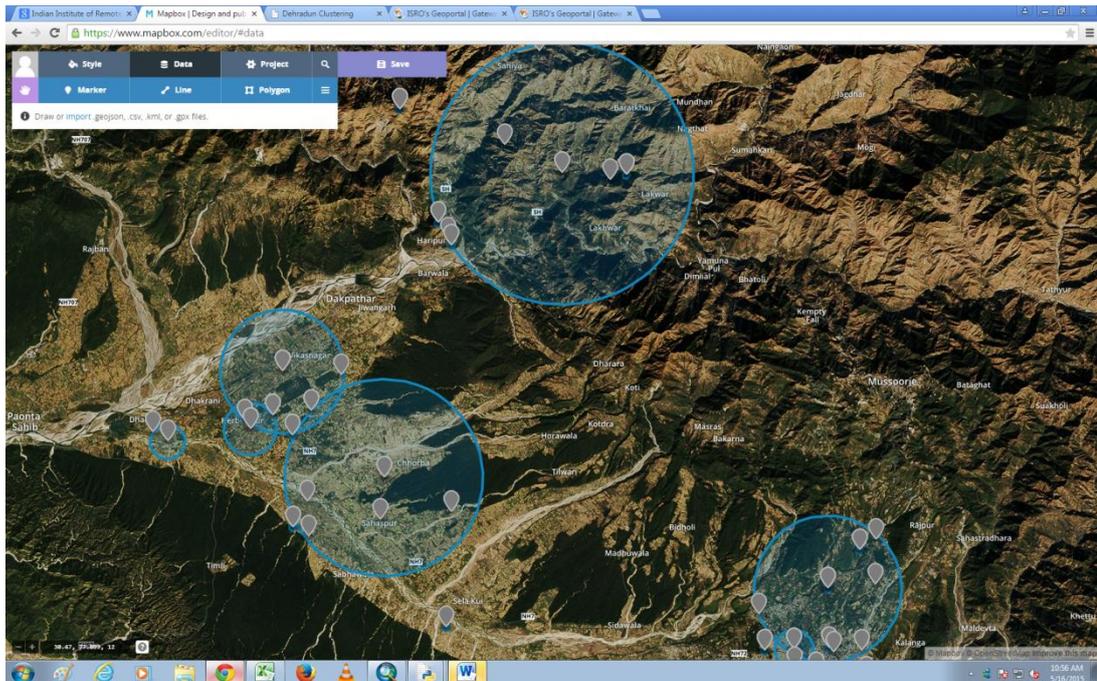
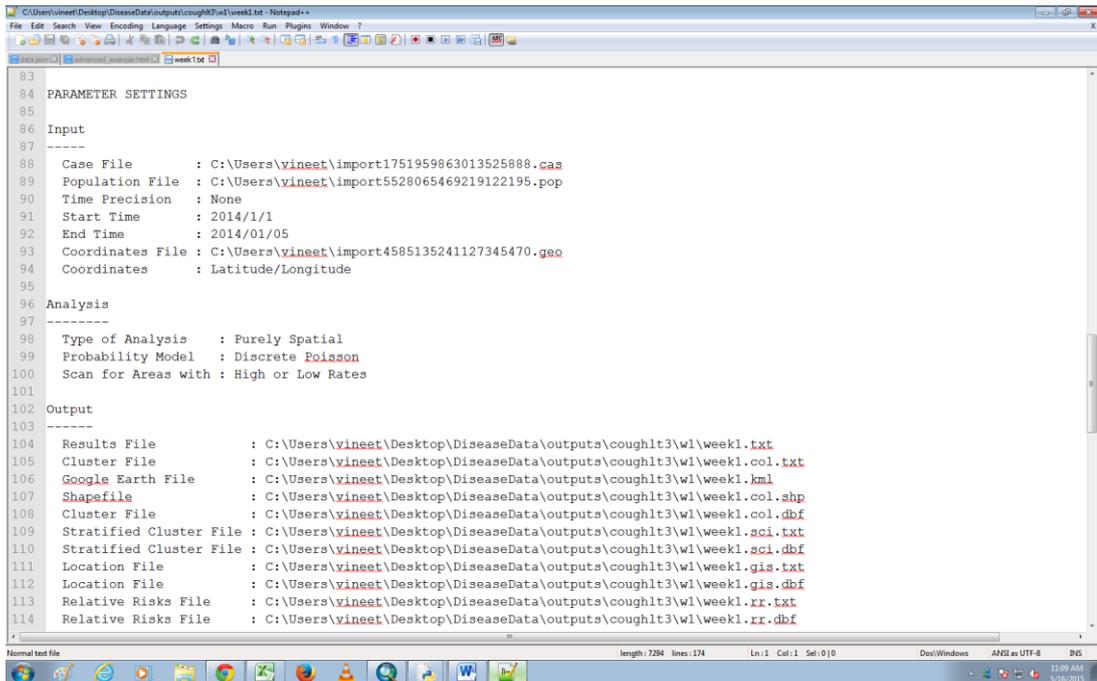


Figure 5-2: Clustering using only one disease data

5.1.1 Input Parameters Analysis

The inputs given to the cluster formation are the case files containing the case information, the population file of locations, the coordinates of the locations and the start and end time of the data collected. The type of analysis of the data is done in purely spatial, temporal or space-time in retrospective analysis where as in prospective analysis only temporal and space time is allowed. The models used are Poisson, Bernoulli, multinomial, ordinal, exponential for discrete scan statistics and for continuous scan statistics Poisson is used.

Stream Data Clustering for Development of Real Time Disease Outbreak Detection System



```
83
84 PARAMETER SETTINGS
85
86 Input
87 -----
88 Case File      : C:\Users\vineet\import1751959863013525888.cas
89 Population File : C:\Users\vineet\import5528065469219122195.pop
90 Time Precision  : None
91 Start Time     : 2014/1/1
92 End Time       : 2014/01/05
93 Coordinates File : C:\Users\vineet\import4585135241127345470.geo
94 Coordinates     : Latitude/Longitude
95
96 Analysis
97 -----
98 Type of Analysis : Purely Spatial
99 Probability Model : Discrete Poisson
100 Scan for Areas with : High or Low Rates
101
102 Output
103 -----
104 Results File      : C:\Users\vineet\Desktop\DiseaseData\outputs\coughlt3\w1\week1.txt
105 Cluster File     : C:\Users\vineet\Desktop\DiseaseData\outputs\coughlt3\w1\week1.col.txt
106 Google Earth File : C:\Users\vineet\Desktop\DiseaseData\outputs\coughlt3\w1\week1.kml
107 Shapefile       : C:\Users\vineet\Desktop\DiseaseData\outputs\coughlt3\w1\week1.col.shp
108 Cluster File     : C:\Users\vineet\Desktop\DiseaseData\outputs\coughlt3\w1\week1.col.dbf
109 Stratified Cluster File : C:\Users\vineet\Desktop\DiseaseData\outputs\coughlt3\w1\week1.sci.txt
110 Stratified Cluster File : C:\Users\vineet\Desktop\DiseaseData\outputs\coughlt3\w1\week1.sci.dbf
111 Location File    : C:\Users\vineet\Desktop\DiseaseData\outputs\coughlt3\w1\week1.gis.txt
112 Location File    : C:\Users\vineet\Desktop\DiseaseData\outputs\coughlt3\w1\week1.gis.dbf
113 Relative Risks File : C:\Users\vineet\Desktop\DiseaseData\outputs\coughlt3\w1\week1.rr.txt
114 Relative Risks File : C:\Users\vineet\Desktop\DiseaseData\outputs\coughlt3\w1\week1.rr.dbf
```

Figure 5-3:Input Parameters File

5.1.2 Cluster File

The output generated from input parameters in the cluster file. The cluster file can be a shape file or the kml file for google earth. The cluster file will give us the number of cluster formed, locations in each cluster with number of cases in them, cluster population size, relative risk value, log likelihood ration and p-value. The shape file can be visualize over any gis based software like mapbox, QGIS or ArcGIS.

Stream Data Clustering for Development of Real Time Disease Outbreak Detection System

```
14 SUMMARY OF DATA
15
16 Study period.....: 2014/1/1 to 2014/01/05
17 Number of locations.....: 103
18 Total population.....: 1696693
19 Total number of cases.....: 87
20 Annual cases / 100000.....: 374.6
21
22
23 CLUSTERS DETECTED
24
25 1.Location IDs included.: Juddo, Kamla, Khatar, Pipaya, Samalta, Sureu, PHCKalsi, ByasBhoor
26 Overlap with clusters.: No Overlap
27 Coordinates / radius..: (30.551770 N, 77.904160 E) / 6.75 km
28 Gini Cluster.....: Yes
29 Population.....: 19353
30 Number of cases.....: 29
31 Expected cases.....: 0.99
32 Annual cases / 100000.: 10946.1
33 Observed / expected...: 29.22
34 Relative risk.....: 43.34
35 Log likelihood ratio..: 75.022713
36 P-value.....: < 0.000000000000000001
37
38 2.Location IDs included.: Rajawala, Telpura, Selaqui, Keinchiwala, Rampur, Sherpur, Sahaspur,
39 Charba, Sabhawala, HariyawalaKhurd, Kaulagarh, Majri, Jassowala,
40 Mehuwala, UHPBallupur, Gaziawala, Badripur, Pashchimwala, UHCMajra,
41 Jannipur, Kanwali, AsanBag, UHCKishanNagar, Majra, SewalaKala,
42 UHPKhubura, Garhi, UHPPatelNagar, MCHHerbetpur
43
44 Overlap with clusters.: No Overlap
45 Coordinates / radius..: (30.390940 N, 77.896240 E) / 14.95 km
46 Gini Cluster.....: Yes
47 Population.....: 607783
```

Figure 5-4: Output Cluster File

5.1.3 Data Models

The case data collected can be of any type like 0/1 that is whether the person is affected or not, categorical data like cancer data giving stages of cancer. After processing different types of data using various analyses we found which model to be used for what type of data. Bernoulli model is specially used for data values 0/1 that it is only considering the positive and negative values. Multinomial model is used when we have categorical data having no relationships with each other like the case data of children of different age groups. For disease case data like cancer data of state 1 stage 2 and stage 3 we use ordinal model as there is relationship between every stage that is category.

5-1:Data model analysis

Data Type Used	Analysis Performed	Probability Model Used
Data in form of 0/1 that is cases and non-cases	Purely Spatial	Bernoulli
Categorical Data do not having any relationship	Purely Spatial	Multinomial Model
Categorical Data having relationship	Purely Spatial	Ordinal Model
When only case information is present	Space-Time	Space-Time permutation Model
Continues data with positive and negative outcomes	Purely Spatial	Normal Model
Survival time data	Space-Time	Exponential Model
Covariates data	Space-Time	Discrete Poission Model

5.2 Evaluation of different stream data clustering algorithms

To develop the system an algorithms is to be implemented to process the case data. To implement it an evaluation of various stream clustering algorithms has been done theoretically. During analysis different parameters has been taken under considerations like how many parameters the algorithms is taking as inputs, what type of data structure it is using to store the data to be get processed in RAM and what is the output format that we getting to visualize. All this parameters will give the space and time complexity of the algorithm. Fig is showing the comparison between different stream clustering algorithms

5-2:Evaluation of different stream data clustering algorithms

S.No.	Algorithm	Cluster Algorithm	Quality Metrics	Data Structure	Cluster Shape	No. Of Parameters
1.	BIRCH	k-mean	Purity	feature vector	hyper-sphere	5
2.	CluStream	k-mean	Rand index	feature vector	hyper-sphere	9
3.	ClusTree	k-mean/DBSCAN	Purity	feature vector	arbitrary	3
4.	D-Stream	DBSCAN	SSQ	grid	arbitrary	5
5.	DenStream	DBSCAN	Purity	feature vector	arbitrary	4
6.	DGClus	k-mean	Purity	grid	hyper-sphere	5
7.	ODAC	Hierarchal Clustering	Purity	correlation matrix	hyper-sphere	3
8.	Scalable k-mean	k-mean	Purity	feature vector	hyper-sphere	5
9.	Single pass k-mean	k-mean	Purity	feature vector	hyper-sphere	2
10.	Stream	k-mean	Purity	prototype array	hyper-sphere	3
11.	Stream LSearch	k-mean	Purity	prototype array	hyper-sphere	2
12.	StreamKM++	k-mean	Purity	corset tree	hyper-sphere	3
13.	SW Clustering	k-mean	Purity	prototype array	hyper-sphere	5

After doing the theoretical analysis the comparison has been done with the time and space complexity of algorithms and SatScan algorithm. The comparison has been done with three algorithms namely K-mean, DBSCAN and CURE. After comparative analysis DBSCAN is found to be the optimal one .

5-3:Space-time complexity analysis

Complexity	SaTScan Tool	Complexity of different stream clustering algorithms		
		D-Stream	DBSCAN	CURE
Space-Complexity	$O(RGM)$	g	$O(n * \log n)$	$O(n^2 \log n)$
		$O(1) + O(g)$	$O(n)$	$O(n)$
Time-Complexity	$O(n * \log n)$			

Where RGM is referred to Representative Graph method,
 n= number of data points taken,
 d= distance to neighborhood
 k= number of clusters defines
 g= number of grids in the grid list

5.3 Database Creation through DHIS Implementation

5.3.1 Data Entry Form

Data entry in DHIS 2 is done through data entry forms. Data entry forms are linked to data sets that have the structure of the respective form. From the left hand side panel of the page we have to select the organization unit for which the data is to be entered then we have to select the data set that is the S, P or L form in our respect and the last the date for which the data is entered. This gives the format to enter data .

The screenshot shows the DHIS2 Data Entry interface. On the left is a tree view of the organizational structure under Uttarakhand, with Barnotha selected. The main area is titled 'Data Entry' and shows the following details:

- Organisation Unit: Barnotha
- Data Set: Weekly S-Form
- Period: W46 - 2014-11-10 - 2014-11-16

Below these fields is a row of five data entry boxes with values 5, 7, 8, 1, and 5. At the bottom of the form are buttons for 'Complete', 'Incomplete', and 'Run validation'. A status bar at the bottom indicates 'Completed by: admin at: 2015-03-09 See details'.

Figure 5-5:Data Entry Form

We can enter data also through pdf or excel files, in which we have to check for the name convention or through specific codes generated by DHIS for each organization unit, data sets, data elements or

5.3.2 PostGIS Database Connection

The data entry done into DHIS goes to the postgres database connected to it at the time of DHIS war file deployment configuration. We can make the PostGIS connection to the DHIS with the PgAdmin interface

	id	pk integer	code	created	lastupdated	name	startdate	enddate	description	formname	valuetype	numberof	texttype	domaintype	aggregation	date
		character varying(1)	character varying(1)	timestamp with time zone	timestamp with time zone	character varying(200)	character varying(50)	text	character varying(100)	character varying(100)	character varying(100)	number	character varying(100)	character varying(100)	character varying(100)	timestamp with time zone
1	203	app01010af		2015-02-16	2015-03-04	Jambose cases of less than 4 weeks	Jambose				int	number		AGGREGATE	sum	15
2	204	XGJZampW9		2015-02-16	2015-03-04	Acute Flaccid Paralysis cases in less Acute					int	number		AGGREGATE	sum	15
3	205	SS0h0w1tm		2015-02-16	2015-03-04	Unusual Symptoms Leading to Death	Unusual Symptoms Leading to Death				int	number		AGGREGATE	sum	15
4	230	wzd0m01uc		2015-02-20	2015-03-04	Fever less than 7 days(Only Fever)	Fever less than 7 days(Only Fever)				int	number		AGGREGATE	sum	15
5	231	Gdz0w070q		2015-02-20	2015-03-04	Fever less than 7 days(With rash)	Fever less than 7 days(With rash)				int	number		AGGREGATE	sum	15
6	232	8080Z0y00T		2015-02-20	2015-03-04	Fever less than 7 days(With Bleeding)	Fever less than 7 days(With Bleeding)				int	number		AGGREGATE	sum	15
7	233	10190T000UF		2015-02-20	2015-03-04	Fever less than 7 days(With Daze or	With Daze or Semi-consciousness/ Uncons				int	number		AGGREGATE	sum	15
8	234	RD00000000		2015-02-20	2015-03-04	Fever more than 7 days	Fever more than 7 days				int	number		AGGREGATE	sum	15
9	235	ab0300000Q		2015-02-20	2015-03-04	Cough with or without fever(<3 week)	Cough with or without fever(<3 week)				int	number		AGGREGATE	sum	15
10	236	1w7000P0013		2015-02-20	2015-03-04	Cough with or without fever(>3 week)	Cough with or without fever(>3 week)				int	number		AGGREGATE	sum	15
11	237	7000000000		2015-02-20	2015-03-04	Loose Watery Stools < 2 weeks(With	Loose watery stools / much dehydration				int	number		AGGREGATE	sum	15
12	238	0Y71000000E		2015-02-20	2015-03-04	Loose Watery Stools < 2 weeks(With	Loose watery stools / much dehydration				int	number		AGGREGATE	sum	15
13	239	00000000000		2015-02-20	2015-03-04	Loose Watery Stools < 2 weeks(With	Loose Watery Stools < 2 weeks(With				int	number		AGGREGATE	sum	15
14	240	00000000000		2015-02-20	2015-03-04	Outcome (In case of death, please use	Outcome (In case of death, please use				int	number		AGGREGATE	sum	15
15	461	00100000000		2015-03-16	2015-03-17	Malaria	Malaria				int	int		AGGREGATE	sum	15
16	462	00000000000		2015-03-16	2015-03-16	Measles	Measles				int	number		AGGREGATE	sum	15
17	463	00000000000		2015-03-16	2015-03-16	Chicken pox	Chicken pox				int	number		AGGREGATE	sum	15
18	464	00000000000		2015-03-16	2015-03-16	Dengue / DHF / DSS	Dengue / DHF / DSS				int	number		AGGREGATE	sum	15
19	465	00000000000		2015-03-16	2015-03-16	Chikungunya	Chikungunya				int	number		AGGREGATE	sum	15
20	466	00000000000		2015-03-16	2015-03-16	Measles	Measles				int	number		AGGREGATE	sum	15
21	467	00000000000		2015-03-16	2015-03-16	Acute Encephalitis Syndrome	Acute Encephalitis Syndrome				int	number		AGGREGATE	sum	15
22	468	00000000000		2015-03-16	2015-03-16	Enteric Fever	Enteric Fever				int	number		AGGREGATE	sum	15
23	469	00000000000		2015-03-16	2015-03-16	Fever of Unknown Origin (FUO)	Fever of Unknown Origin (FUO)				int	number		AGGREGATE	sum	15

Figure 5-6:Postgres Database

5.3.3 GIS Visualization

GIS tab on DHIS dashboard will link us to the gis interface. GIS interface have the panel on right hand side to select the layers to be displayed as base layer like googleopen street and google map and the thematic layers. On left hand side a bar is given having link to have boundary layer , facility layer.

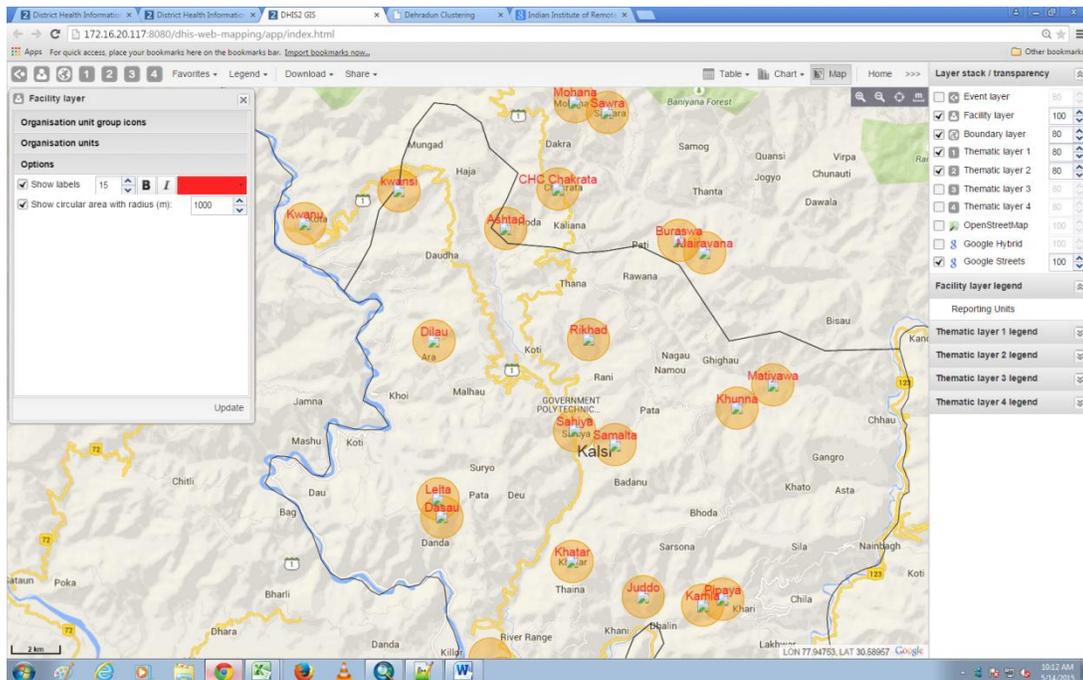


Figure 5-7:GIS Visualization

The data that is entered through the data entry form or through any mobile interface medium, can be visualize using thematic layers in DHIS GIS viewer. The thematic layer will have the parameters like, whether data elements or indicators are used as input, date and period of the data collection, name of the organization units of which data is collected and to be visualize and legend of the map. In fig. 10 data of fever less than 7 days has been visualized taking input as data elements.

Stream Data Clustering for Development of Real Time Disease Outbreak Detection System

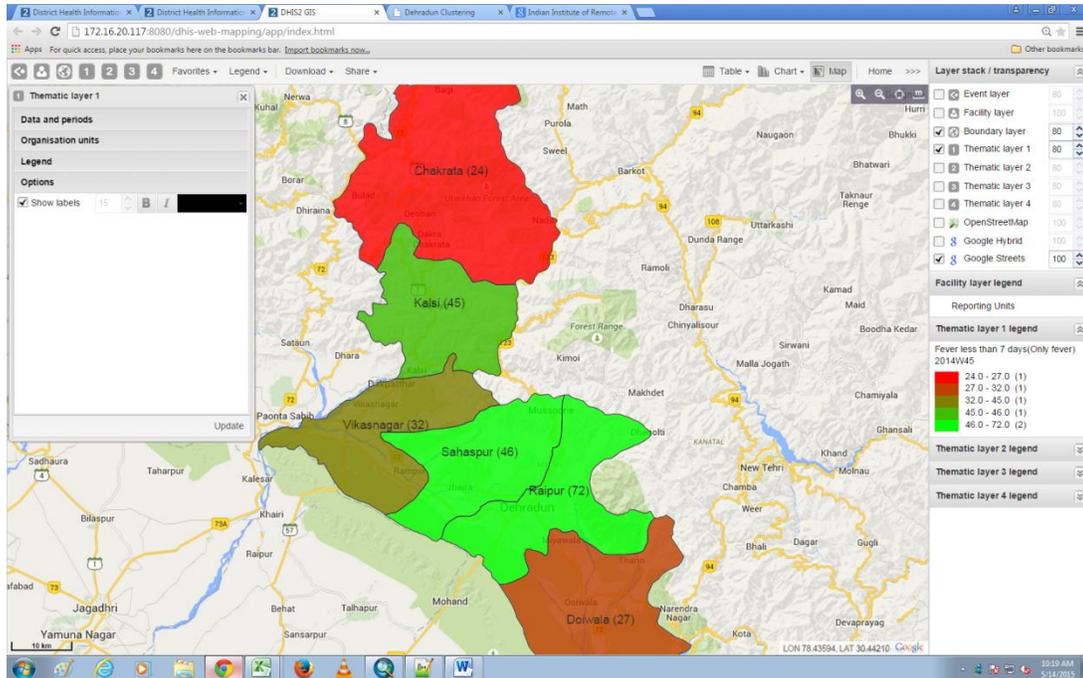


Figure 5-8:GIS Visualization of fever data

We can have multiple thematic layer visualization on one DHIS GIS visualizer. Fig 11 and 12 shows the data visualization of cough and pneumonia. We can have base layer change by clicking on the layer stock/transparency panel on the right hand side panel on the GIS page.

Stream Data Clustering for Development of Real Time Disease Outbreak Detection System

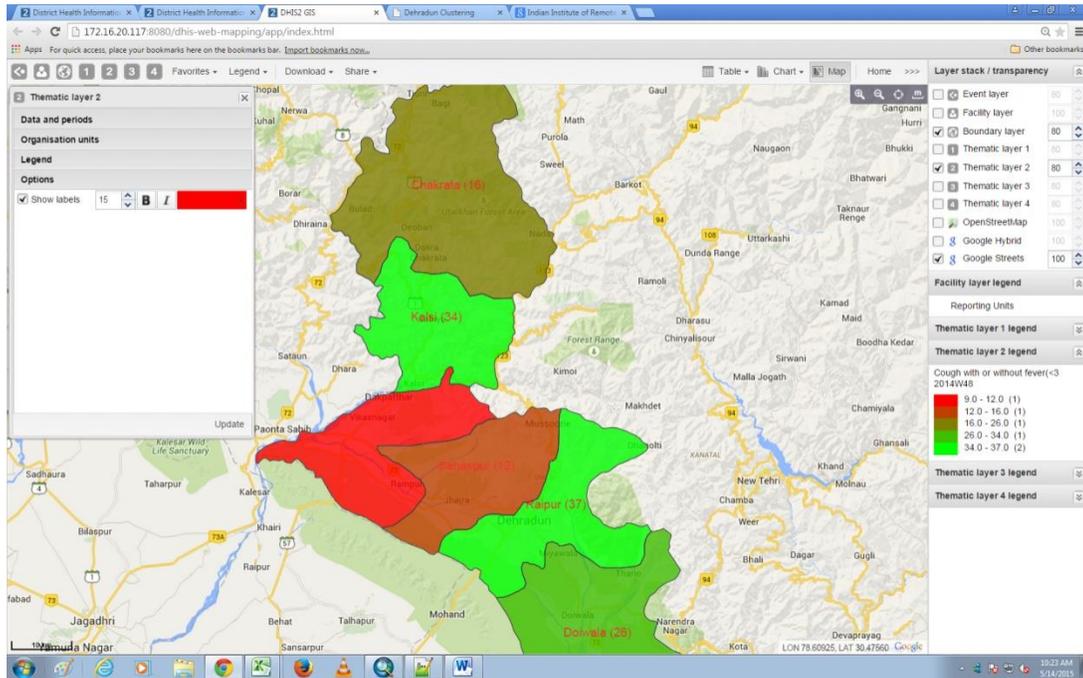


Figure 5-9:GIS Visualization of cough data

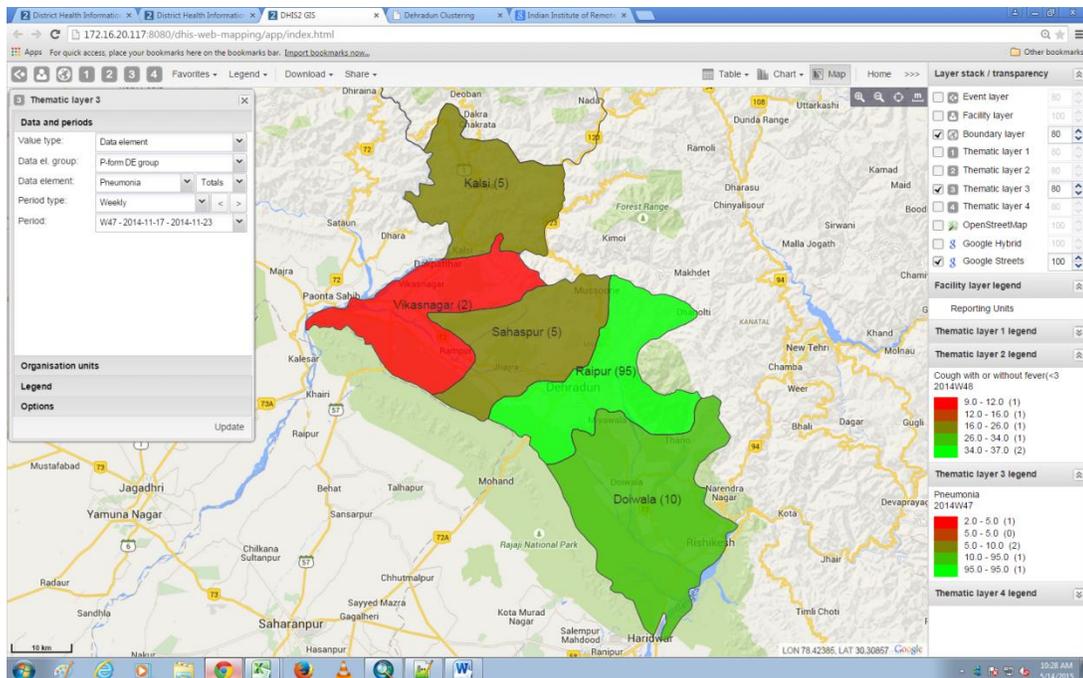
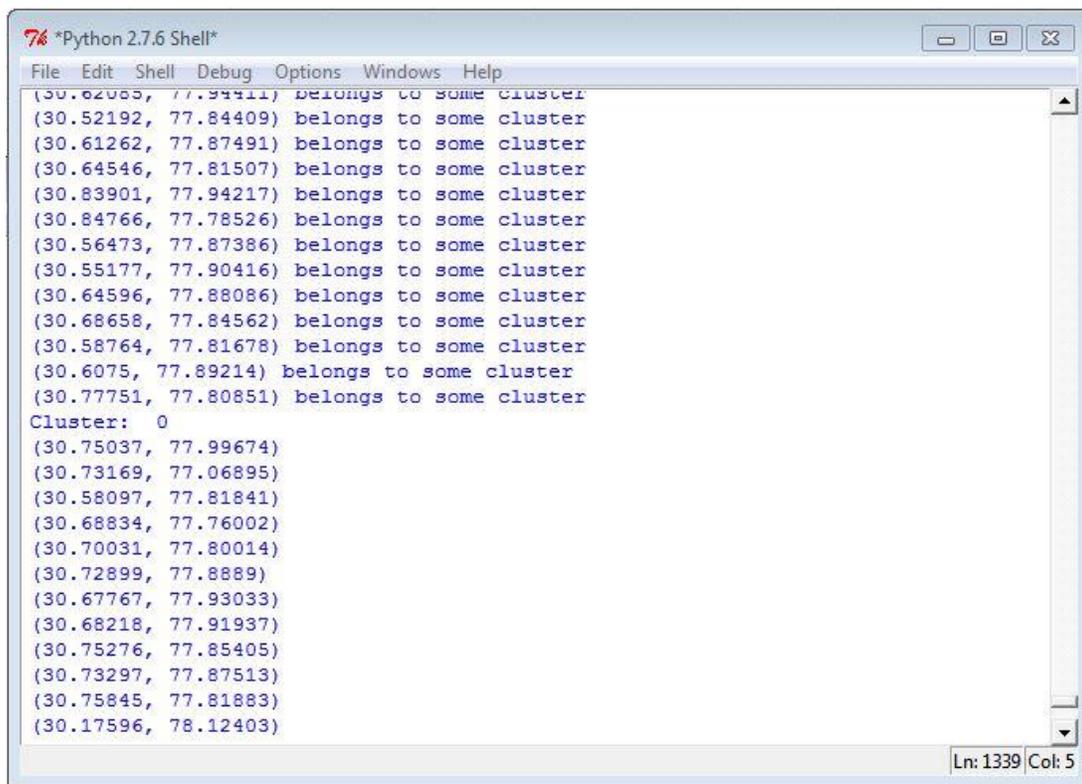


Figure 5-10:GIS Visualization of pneumonia data

5.4 Real Time Disease Outbreak Detection System

5.4.1 Algorithm Implementation

The algorithm used is density based spatial clustering algorithm (DBSCAN). We have to pass the latitude and longitude of the locations where the cases occurred individually like if there are two cases occurred on same location, two different entry will be counted as the time stamp of two will be different. As it connected to the stream source it will select on centroid and check for the distance of the random points coming from the centroid. It will allocate points in the clusters as per the locations. Numbers of points in any clusters differ depending on the occurrence of the points.



```
*Python 2.7.6 Shell*
File Edit Shell Debug Options Windows Help
(30.62065, 77.94411) belongs to some cluster
(30.52192, 77.84409) belongs to some cluster
(30.61262, 77.87491) belongs to some cluster
(30.64546, 77.81507) belongs to some cluster
(30.83901, 77.94217) belongs to some cluster
(30.84766, 77.78526) belongs to some cluster
(30.56473, 77.87386) belongs to some cluster
(30.55177, 77.90416) belongs to some cluster
(30.64596, 77.88086) belongs to some cluster
(30.68658, 77.84562) belongs to some cluster
(30.58764, 77.81678) belongs to some cluster
(30.6075, 77.89214) belongs to some cluster
(30.77751, 77.80851) belongs to some cluster
Cluster: 0
(30.75037, 77.99674)
(30.73169, 77.06895)
(30.58097, 77.81841)
(30.68834, 77.76002)
(30.70031, 77.80014)
(30.72899, 77.88889)
(30.67767, 77.93033)
(30.68218, 77.91937)
(30.75276, 77.85405)
(30.73297, 77.87513)
(30.75845, 77.81883)
(30.17596, 78.12403)
Ln: 1339 Col: 5
```

Figure 5-11:Algorithm Processing

The algorithm run for n number of Monte Carlo simulations, n defines the number of case in an area or the latitude and longitude of the disease occurrence.

5.4.2 Output Comparison

SaTScan clusters and clusters generated from algorithm

The figure shows the difference between the clusters generated form DBSCAN algorithm clusters output and the SaTScan clusters. The clusters generated from the algorithm are random points clusters displayed on the window while the SaTScan clusters are the circular window cluster. The optimal time to run the algorithm comes out to be approximately 0.25 seconds where as the SaTScan process will completes in approximately 1 second.

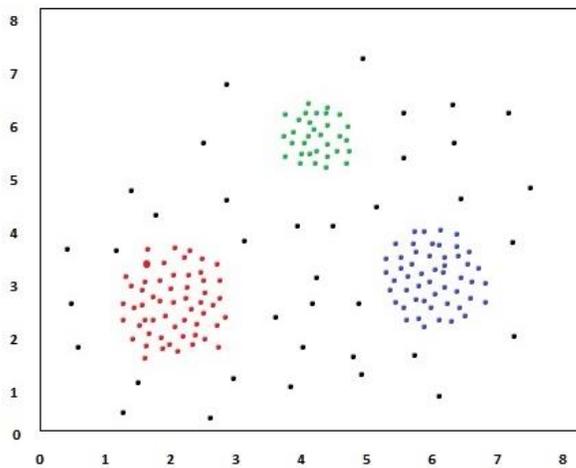


Figure 5-12:DBSCAN output

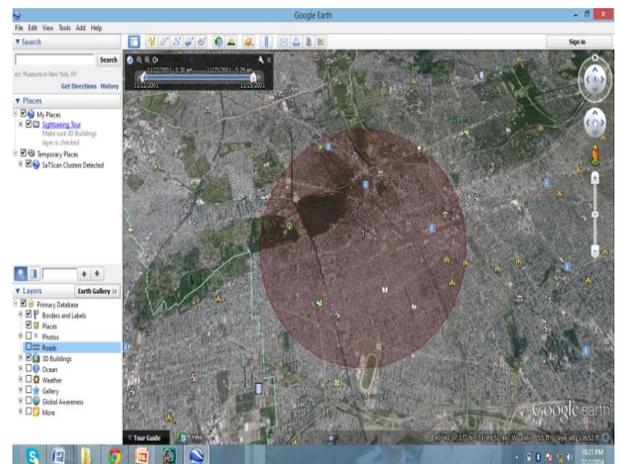


Figure 5-13:SaTScan Output cluster

The table of comparison shows that the DBSCAN runs faster than SaTScan algorithm and have better time and space complexity. The number of parameters taken by DBSCAN are also less as compared to SaTScan and other clustering algorithms.

5-4:Table of Comparison

Algorithm	Running Time	Complexity		Input parameters	Cluster Shape
		Time	Space		
DBSCAN	0.25 sec approx.	$O(n)$	$O(n * \log n)$	ϵ - distance and minPoints	Arbitrary Shape and Size
SaTScan Algorithm	2 sec approx.	$O(n * \log n)$	$O(RGM)$	Case file, pop file, geo file	Circular

5.5 GUI design of disease outbreak detection system

To design the interface of the system PyQt module of python has been used. Fig—shows the interface design of the developed system, PyQt combo box containing the entry fields for district, disease vector and database file is used, in output combo box, layers like BHUVAN, Geosever and Google map on which the clusters can be displayed has been given. We can also get the cluster file on clicking the download option in shape file format. To display the clusters on layers web view is needed. In web view layer will be displayed using the url of the source of where the layer has been published.

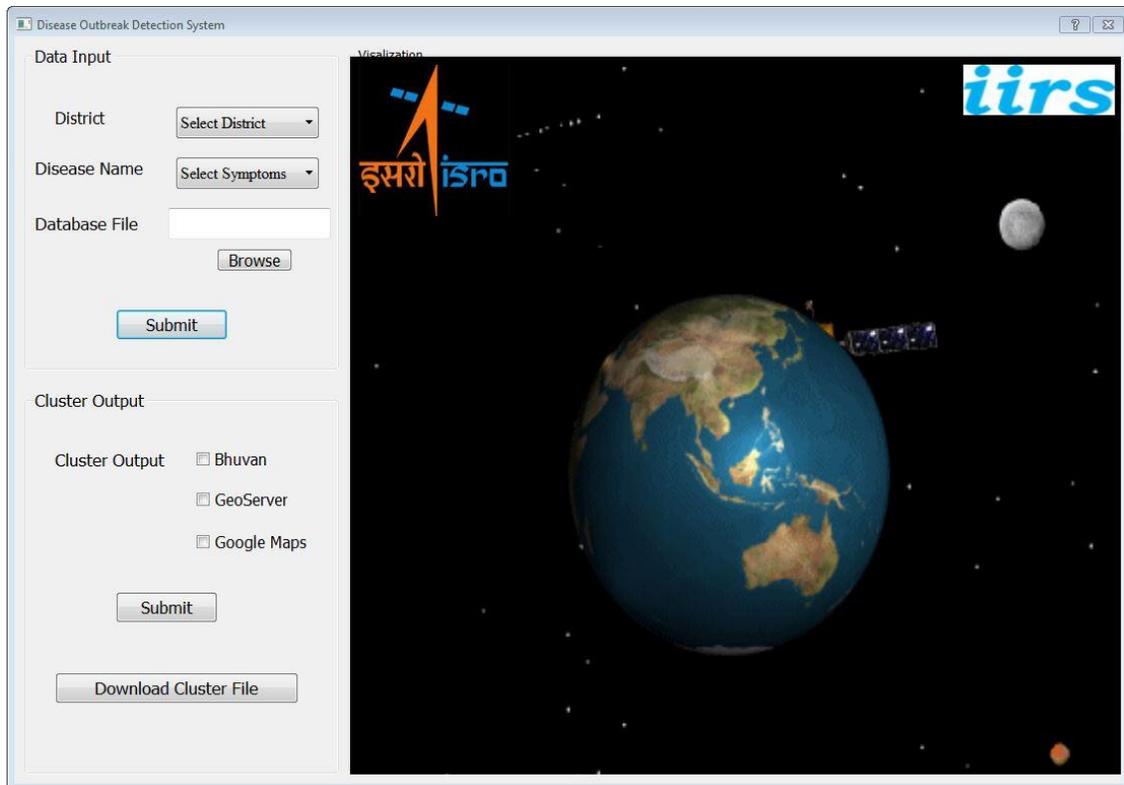


Figure 5-14:Interface of Developed System

5.5.1 Displaying Clusters Using Google Maps

To perform clustering using Google map select a district and disease vector as input parameters. Clustering need the basic parameters namely latitude and longitude of the location where the cases has been occurred, number of cases occurred and the time of the reporting of that case. These parameters have been parsed using the Jason file format. Jason is the specific file format .Processing of Jason file using the clustering code will map the points on the Google map layer. These points are displayed as individual's points showing the lat longs of the cases on the layer. Figure 12 showing the points of the Google map layer as individual points.

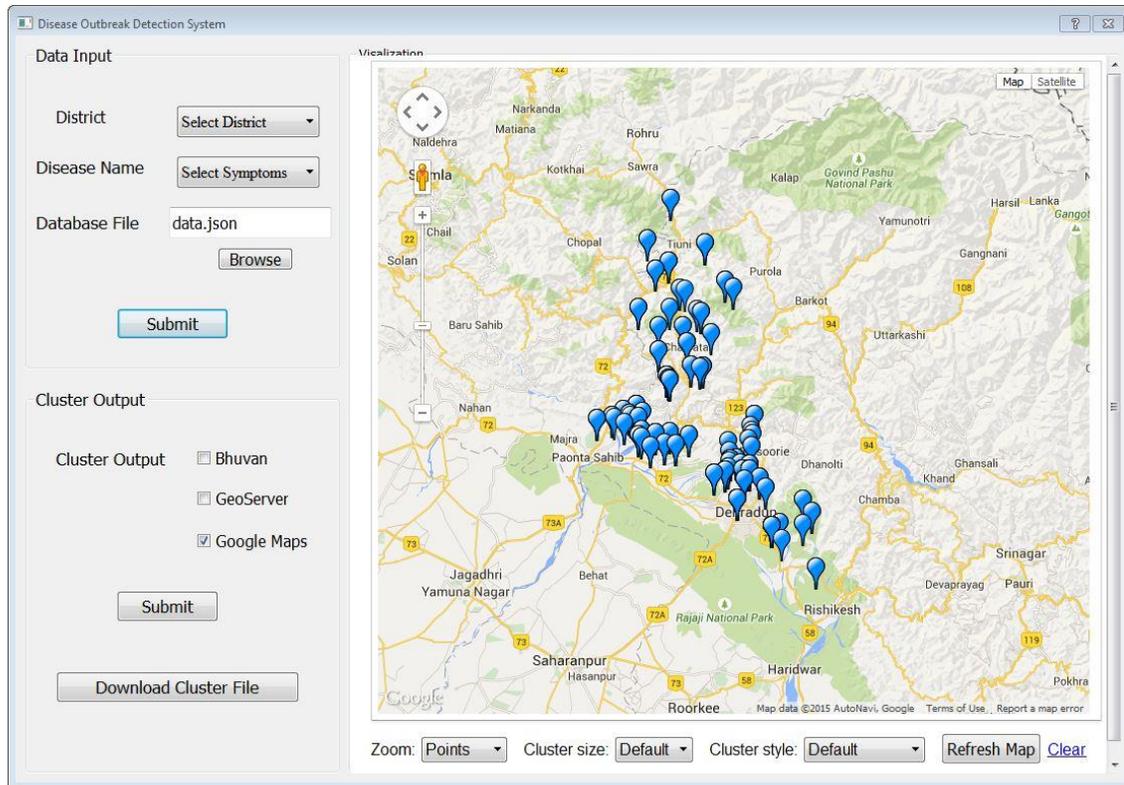


Figure 5-15: Clustering Using Google Map

After displaying the points on the Google map layer, clustering process will start, on zoom out the points will start combining together and cluster will be formed of different colors. Color coding will show the intensity of the cluster. Fig 13 shows the cluster formed using the data.json file format. Here, blue color will show a low intensity cluster having points club together less than ten. Yellow color will show the cluster of points less than hundred.

On further zoom out more number of lat long points will combine together and the cluster size will start increasing. Fig 14 shows the clusters having number of cases more than hundred at some location covering specific area. We can also vary the size of the clusters by selecting the minimum number of points that to be required to form a cluster by drop down field of cluster size given in the bottom of the layer.

Stream Data Clustering for Development of Real Time Disease Outbreak Detection System

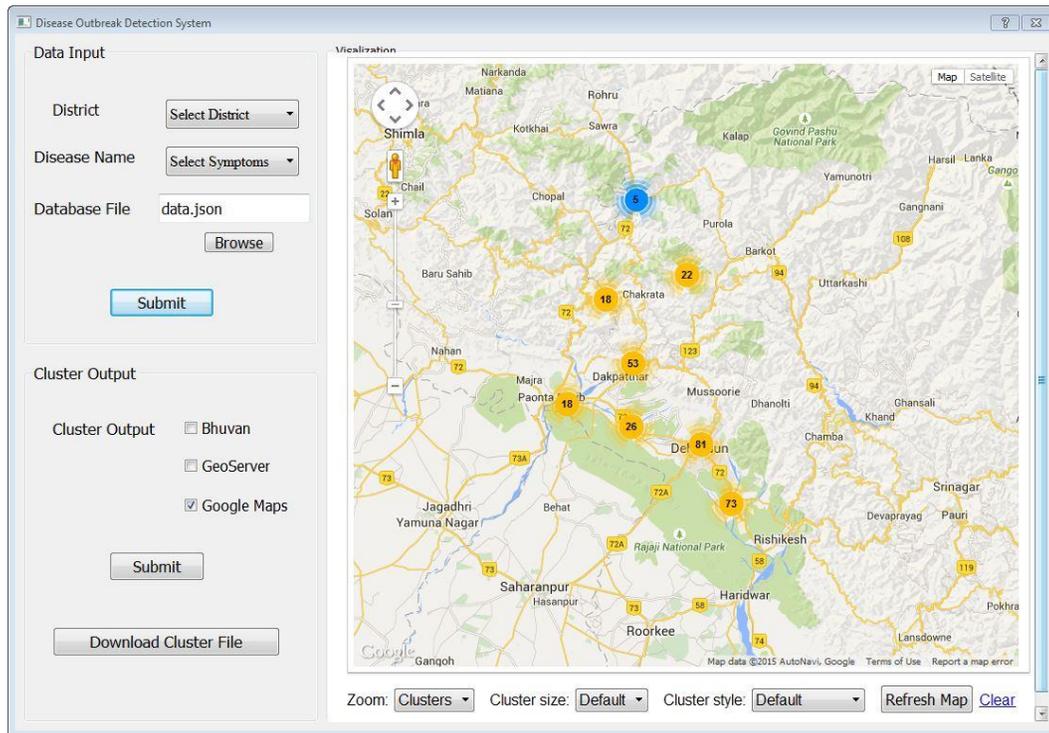


Figure 5-16: Clustering Using Google Map at zoom level 4

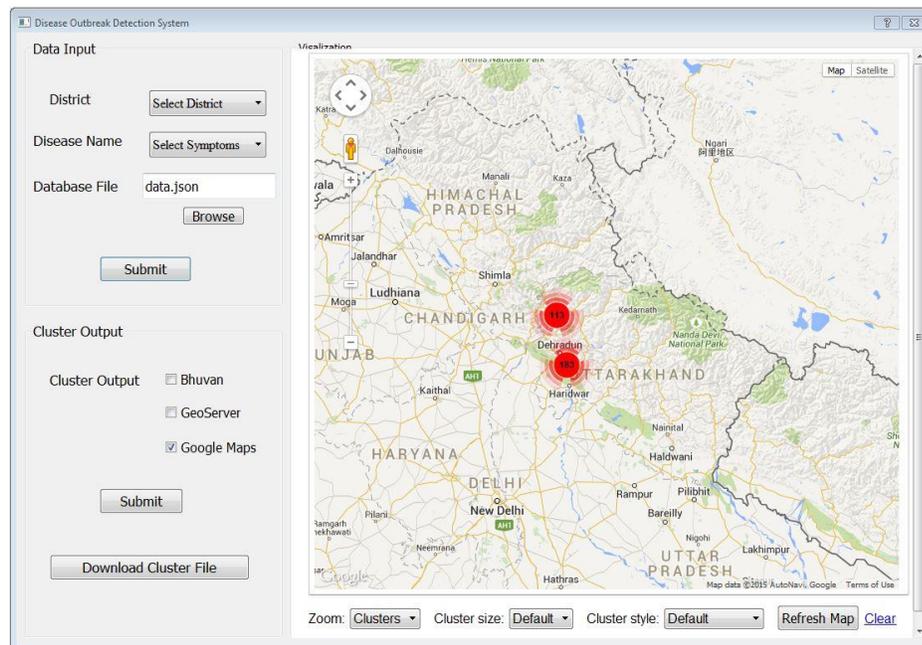


Figure 5-17: Clustering Using Google Map at zoom level

5.5.2 Displaying Clusters Using Geosever

We can also display disease clusters using geosever. After the processing of lat long cases data by the clustering algorithm the output generated will be the shape files of number of cases in an area depending on the high or the low rate of occurrence. The each cluster shape file will contain the cluster id,name of the centroid of the cluster areas covered by that cluster, radius of the cluster, start and end date of the data collected, test-statistics and p-value.

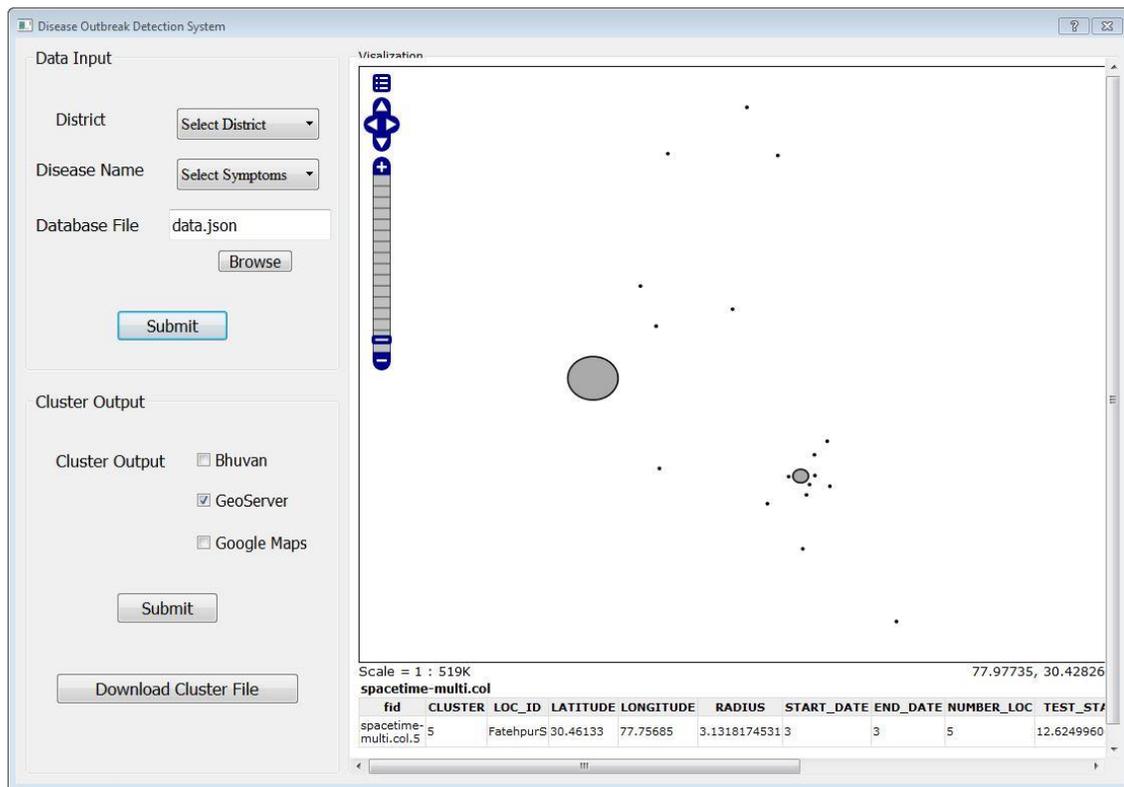


Figure 5-18:Displaying Clusters Using Geosever

5.5.3 Displaying Cluster Using BHUVAN

BHUVAN in the ISRO portal used to display wms and wcs services on the web based environment. 2D and 3D versions of the bhuvan has been developed. We can used the bunvanwms services like that of google and can display clusters on that.

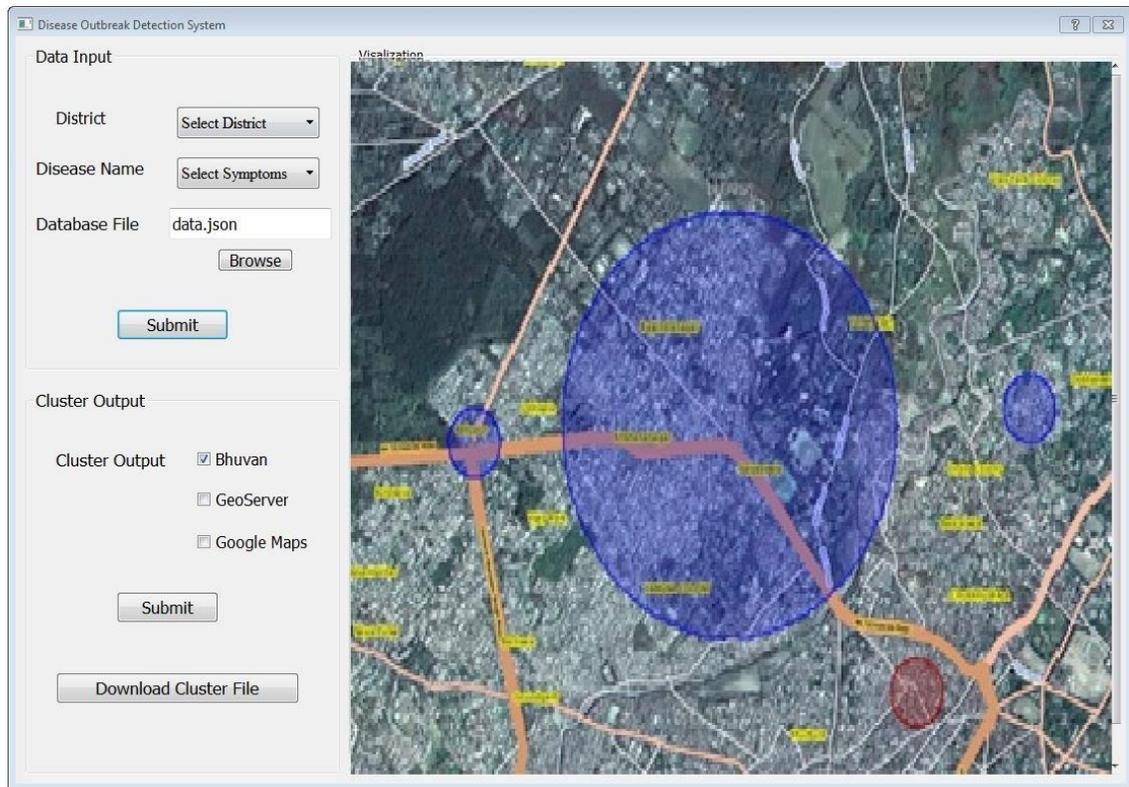


Figure 5-19: Displaying Clusters Using BHUVAN

6. CONCLUSIONS AND RECOMMENDATIONS

6.1 Conclusions

What are the parameters for comparison of different algorithm results using health benchmark data?

By analysis of different stream data clustering algorithms, parameters like cluster shape, number of input parameters required, space and time complexity has been found to be used for the comparative study. These parameters govern the whole algorithms processing. Number of input parameters shows the maximum data fields that are to be given as input like population, date and time, lat-long of the area etc. whereas the space and time complexity gives the total memory and time required by the algorithm for complete data processing.

How to overcome the limitations of existing disease outbreak detection system to make it real time detection system?

Existing systems are human experts based non-automated or semi-automated systems having weekly or monthly reporting data formats. To make an existing system into real time a continuous data source has been created by implementing DHIS which gives .xml data format source that can be used as a data link having case information of diseases, thier locations and time period of the data entry. The DHIS data base can also be accessed by PgAdmin through a connection to the DHIS.

What will be the database constraints in designing stream database?

During the implementation of database through DHIS which fields are necessary for data entry is to be selected so that to have fast and non-redundant database.

What will be the user interface design requirement of disease outbreak detection system?

A simple user friendly interface using PyQt has been design taking district, disease and a database connection as the input parameters and the output base layer on which the cluster is to be shown is given in output. A web view panel has been given on the right hand side of the system to visualize output.

6.2 Recommendations

- Some other algorithms such as D-Stream, Den-Stream should also be tried to do the clustering.
- Various other modes to enter the data into the database can be developed like entry through app, or message based interface.
- In place of using WMS services, cluster on the 3D layers should be tried to display. To display 3D, dem data of strm or cartosat can be used. With 3D, we can get the clusters displayed on both side of hills or mountains which can help in analysing the trends in cases occurrence. 3D display is also useful to have analysis for the diseases cases that are affected by height factor.
- Rectangular window option is also there to display the cluster.

References

- Aggarwal, C. C. (2002). An intuitive framework for understanding changes in evolving data streams. Proceedings 18th International Conference on Data Engineering. doi:10.1109/ICDE.2002.994715
- Aggarwal, C. C. (2003). A framework for diagnosing changes in evolving data streams. Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data - SIGMOD '03, 575. doi:10.1145/872824.872826
- Aggarwal, C. C., & Yu, P. S. (2008). A framework for clustering uncertain data streams. In Proceedings - International Conference on Data Engineering (pp. 150–159). doi:10.1109/ICDE.2008.4497423
- Aggarwal, C. C. (2009). A framework for clustering massive-domain data streams. In Proceedings - International Conference on Data Engineering (pp. 102–113). doi:10.1109/ICDE.2009.13
- Amini A, Wah TY, Saboohi H. On density-based data streams clustering algorithms: A survey. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 29(1): 116–141 Jan. 2014. DOI 10.1007/s11390-013-1416-3
- Bifet, A., Holmes, G., Pfahringer, B., Kirkby, R., & Gavalda, R. (2009). New Ensemble Methods For Evolving Data Streams. In KDD '09 Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 139–148). doi:10.1145/1557019.1557041
- Cao, F., Liang, J., Bai, L., Zhao, X., & Dang, C. (2010). A framework for clustering categorical time-evolving data. IEEE Transactions on Fuzzy Systems, 18, 872– 882. doi:10.1109/TFUZZ.2010.2050891
- Chen, Yixin., & Li, Tu. (2009). “Density-based clustering for real-time stream data.” Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining.
- Chen, K., & Liu, L. (2009). HE-Tree: A framework for detecting changes in clustering structure for categorical data streams. VLDB Journal, 18, 1241– 1260. doi:10.1007/s00778-009-0134-5

- Cormode, G., & McGregor, A. (2008). Approximation algorithms for clustering uncertain data. Proceedings of the Twenty-Seventh ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, 191–200. doi:10.1145/1376916.1376944
- Forestiero, A., Pizzuti, C., & Spezzano, G. (2009). FlockStream: A bio-inspired algorithm for clustering evolving data streams. In Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI (pp. 1–8). doi:10.1109/ICTAI.2009.60
- Forestiero, A., Pizzuti, C., & Spezzano, G. (2011). A single pass algorithm for clustering evolving data streams based on swarm intelligence. Data Mining and Knowledge Discovery. doi:10.1007/s10618-011-0242-x
- Guha, S., & Munagala, K. (2009). Exceeding expectations and clustering uncertain data. Proceedings of the Twenty-Eighth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems - PODS '09, 269. doi:10.1145/1559795.1559836
- Guha S, Meyerson A. (2003) Clustering data streams: theory and practice. IEEE TKDE Special Issue Cluster 3(2):37–46
- Guha S, Meyerson A, Mishra N. (2003) Clustering data streams: theory and practice. IEEE Trans Knowledge Data Eng 15(3):505–528
- Guha S, Harb B (2005) Wavelet synopsis for data streams: minimizing non-euclidean error. In: Proceeding of the 11th ACM SIGKDD international conference on knowledge discovery in data mining. pp 88–97
- Guha S, Mishra N, Motwani R et al (2000) Clustering data streams. In: Proceedings of the 41st annual symposium on foundations of computer science. pp 359–366
- Huang, L., Kulldorff, M., & Gregorio, D. (2007). A spatial scan statistic for survival data. Biometrics, 63(1), 109–118. doi:10.1111/j.1541-0420.2006.00661.x
- Jones, R. C., Liberatore, M., Fernandez, J. R., & Gerber, S. I. (2006). Use of a Prospective Space-Time Scan Statistic to Prioritize Shigellosis Case Investigations in an Urban Jurisdiction. Public Health Reports, 121, 133–139. doi:10.1080/08985620500531865
- Jung, I., Kulldorff, M., & Richard, O. J. (2010). A spatial scan statistic for multinomial data. Statistics in Medicine, 29(18), 1910–8. doi:10.1002/sim.3951

- Kleinman, K. P., Abrams, A. M., Kulldorff, M., & Platt, R. (2005). A model- adjusted space-time scan statistic with an application to syndromic surveillance. *Epidemiology and Infection*, 133, 409–419. doi:10.1017/S0950268804003528
- Kulldorff, M., Heffernan, R., Hartman, J., Assunção, R., & Mostashari, F. (2005). A space-time permutation scan statistic for disease outbreak detection. *PLoS Medicine*, 2(3), e59. doi:10.1371/journal.pmed.0020059
- Kulldorff, M., Huang, L., & Konty, K. (2009). A scan statistic for continuous data based on the normal probability model. *International Journal of Health Geographics*, 8, 58. doi:10.1186/1476-072X-8-58
- Nagesh, H., Goil, S., & Choudhary, A. (2001). Adaptive grids for clustering massive data sets. *Proceedings of the 1 St SIAM ICDM, Chicago, IL*, 1–17.
- Tango, T., Takahashi, K., & Kohriyama, K. (2011). A Space-Time Scan Statistic for Detecting Emerging Outbreaks. *Biometrics*, 67, 106–115. doi:10.1111/j.1541-0420.2010.01412