Geographical Visualization Approach to Perceive Spatial Scan Statistics: An Analysis of Dengue Fever Outbreaks in Delhi

Student Name: Shuchi Mala

IIIT-D-MTech-CS-DE-11-032
June 18, 2013

Indraprastha Institute of Information Technology
New Delhi

Thesis Committee
Dr. Raja Sengupta (Chair)
Dr. Rahul Dev Garg
Dr. Vinayak Naik

Submitted in partial fulfillment of the requirements
for the Degree of M.Tech. in Computer Science,
with specialization in Data Engineering

This research was partially funded by Geoinformatics group at Indraprastha Institute of Information Technology, Delhi.

Keywords:Disease Surveillance, clusters, statistically significant, outbreaks, visualization, dengue fever, p-value.

Certificate

This is to certify that the thesis titled **"Geographical Visualization Approach to Perceive Spatial Scan Statistics: An Analysis of Dengue Fever Outbreaks in Delhi "** submitted by Shuchi Mala for the partial fulfilment of the requirements for the degree of Master of Technology in Computer Science & Engineering(Data Engineering Specialization) is a record of the bona fide work carried out by her under my guidance and supervision in the Geoinformatics group at Indraprastha Institute of Information Technology, Delhi. This work has not been submitted anywhere else for the reward of any other degree.

Professor Raja Sengupta
Indraprastha Institute of Information Technology, New Delhi

Abstract

In India, there is a strong need of a nation-wide disease surveillance system. As of now there are very few surveillance systems in India to detect disease outbreaks. IDSP (Integrated Disease Surveillance Project) was launched by Government of India with assistance of World Bank to detect and respond to disease outbreaks quickly. Still efforts are needed to strengthen the disease surveillance and response system for early detection of disease outbreaks. The strongest pillar of an accurate disease surveillance system is data related to cases and various risk factors. After data collection, the next important step is transformation of the collected data into meaningful information. Precise statistical methods are then required to analyse the information at hand. Disease outbreaks are detected using statistical analysis tools but for effective disease control a visualization approach is required. Without appropriate visualization it is very difficult to interpret the results of analysis. In the work presented here, a statistical analysis is performed to detect space-time disease clusters and then the developed visualization approach is used to visualize the disease outbreaks. SaTScan software is integrated with the visualization approach to detect location of disease clusters and to test whether the detected clusters are statistically significant. Without the developed visualization approach users will have to run SaTScan software for each disease per data source. Hence, the presented work provides an extremely efficient and accurate technique for early detection of disease outbreak in the region covered by the surveillance system.

Acknowledgments

First and foremost, I would like to thank my supervisor *Dr. Raja Sengupta* for his valuable guidance and advice. He inspired me greatly to pursue my work in this research area. Without his guidance and persistent help this dissertation would not have been possible. I would also like to thank my committee members *Dr. Rahul Dev Garg* and *Dr. Vinayak Naik* for agreeing to be in my committee.

I would like to thank *Dr Ashok Rawat*, *DHO*,*Civil Lines Zone*, *Municipal Corporation of Delhi* and *Dr. Sanjay Sinha*, *Senior Doctor*, *Civil Lines Zone*, *Municipal Corporation of Delhi* for providing me with the vital data on dengue fever to perform analysis. I also highly appreciate their invaluable feedback on my research work. I am grateful to my friend and collegue *Radhika Tayal* and *Shiva Reddy Koti* for their constant support and help.

Finally, an honourable mention goes to my family; for their understanding, endless love and wishes for the successful completion of this project.

This thesis is dedicated to my loving father and my sister who have always been there for me whenever I needed encouragement, motivation and strength to continue this research work.

Contents

List of Figures

List of Tables

Chapter 1

Introduction and Literature review

1.1 Introduction

Disease surveillance is a continuous process of collecting information, as well as organizing, analysing and interpreting the information collected via this process, information about public health can be generated and utilized to take effective actions. In geographical disease surveillance the three components include selection of a (a) disease, (b) a geographical area and (c) a fixed time period for mapping disease rates. The outbreaks of infectious diseases can be discovered by collecting timely information of cases in space and time and analysing this information to detect changes in disease occurrence in an area. This can facilitate pre-emptive actions to be taken by public health officials.
Outbreaks for diseases can be discovered by monitoring space-time trends of disease occurrences which can highlight changing patterns in risk and help to identify new risk factors. However surveillance datasets are mostly large in size.Therefore, the availability and performance of software capable of analysing space-time disease surveillance data using spatial statistical methods on a continuous basis is essential for practical surveillance. Once statistical methods have been utilized to detect changes in the underlying disease process. Geographic information system (GIS) can be used to visualize the spatial variations in disease risk.
Many nations of the world maintain disease database for various diseases like cancer, dengue fever, malaria any many more. Many national and local health organisations continuously collect information about occurrences of different infectious diseases. Number of cases is typically added to a disease database daily, monthly or yearly, with the duration depending on the type of disease and the limit of database system. An early detection of emerging geographical outbreaks due to the presence of suddenly occurring risk factor is important for the health of people. Therefore, there is the need for a disease outbreak surveillance system which can predict and observe the outbreaks and discover the factors contributing to such occurrences.

1.1.1 Traditional Disease Surveillance

A traditional disease surveillance mechanism involves reporting of diseases confirmed by labs to local or national organizations of health.This does not normally allow for early detection of new outbreaks. New surveillance systems have been developed to find out where the outbreaks will occur and what will be the geographical sizes of these outbreaks. In these systems clustering is done which tells about number of cases of disease within study area. Disease clustering is very useful in identifying outbreaks. Disease clustering can be classified as temporal cluster-

ing, spatial clustering or space-time clustering. Temporal clustering observes whether cases are located close to each other in time, spatial clustering observes whether cases are located close to each other in space and space-time clustering observes whether cases are close in space as well as close in time. To examine whether clustering is real and significant many tests have been proposed for various purposes. Besag and Newell (1991) classified these tests into general tests and focused tests. General tests intend to examining of whether clustering occurs over the study region. Focused tests include assessing the clustering around a pre-defined location. Kulldorff (1998) classified general tests into Global clustering tests and cluster detection tests. Global clustering tests evaluating whether cases are located close to each other no matter when and where they occur. Cluster detection tests are developed for detecting localized clusters and evaluating their significance.

In 2000 Wakefield separated methods into four groups: (a) traditional methods,(b) distance or adjacency methods,(c) moving windows methods and (d) risk surface estimation methods. Traditional methods detects over dispersion in a really aggregated data. The tests done in this method are global tests and therefore do not provide an indication of geographical location but detect the presence or absence of clustering over the whole study region. Examples of such test include Pearsons chi-squared statistic, the Put-off and Whittinghills method. Distance or adjacency methods comprise global tests that go through the spatial dependence in the data set. The techniques included in this method are autocorrelation statistics, Whittemores method, Tangos method, and K-functions. Moving windows have been developed to determine whether the number of cases within a window exceeds the expected number of cases occurring by chance. A window can be a circular shape in which cases of disease occurrence can be observed. The window moves systematically throughout the study region. These methods include local tests with the ability to detect spatial locations of disease outbreaks. The techniques used in this method are Openshaws method, Besag and Newells method, scan statistics, and Cuzick and Edwardss method. Scan methods are usually recommended when there is sparse data. Risk surface estimation method includes kernel estimation, generalized additive models, and geo-statistical methods. In these techniques, the stress is less on hypothesis testing and more on estimation of the underlying residual risk surface. These methods help to analyse in depth the nature of the clusters since they not just statistically identify clusters, but produces continuous surfaces of risk across the whole study region. The limitations of risk surfaces is a less well-developed statistical understanding that requires the use of specialized software. These methods are computationally expensive and more difficult to implement.

1.1.2 Statistical methods for disease cluster detection

For detection of disease cluster various statistical methods have also been developed. One well-known method is Openshaws Geographical Analysis Machine (GAM).GAM is a cluster detection approach which works by examining a large number of overlapping circles at a variety of scales and assesses the statistical probability of the number of events occurring by chance. The drawbacks of this method is that it has a multiple testing problems where there are large number of test and original version was heavily computer intensive. FlexScan is a free software which was developed to analyse spatial count data using the flexible spatial scan statistic and circular spatial scan statistic. The current version includes a spatial scan statistic with a restricted likelihood ratio. FLeXScan is quite similar to SaTScan, but the current version of FleXScan is still restricted specifically to spatial analyses, ignoring the temporal component. Another software Splancs was developed by Rowlingson and Diggle (1993) for spatial and space-time point pattern analysis. Finally Kulldroff together with Information Management Services Inc. developed

SaTScan software that can perform geographical surveillance of disease, detect clusters and test whether these clusters are statistically significant or not. It can also perform time-periodic disease surveillance for early detection of disease outbreaks. Unlike techniques such as Openshaws GAM, SaTScan does not take into account the problem of multiple testing and reports the significance of each reported cluster.

In the analysis of cluster detection test there can be issues concerning data, the scale of analysis, correction for covariates and the underlying background population. Covariate in data is the single most important problem affecting cluster study. Any spatial or temporal variation of covariates like gender, age, ethnicity, diet, smoking behaviour or population density can worsen the real disease patterns. For example, people of similar ethnic origin traditionally tend to live close together, although in todays world this is decreasing due to increased population migration. As we know some diseases are inherited, one would expect to observe spatial clusters of genetic diseases. Therefore, examining clusters for such diseases requires evidence for clustering over background population, after adjustments made for the genetic covariates. Methods like SaTScan should be adopted for performing adjustments in covariates. Another case can be related to the background population at risk. Correction of spatial variations in the population at risk is an important part of spatial epidemiological research because any observed pattern of health events needs to be adjusted by the background population distribution. The method implied in point-pattern analysis is to define a suitable set of controls from the at-risk population. By comparing the spatial arrangement of the observed cases with that of the controls, a relative risk estimate can be produced that has adjusted for the underlying non uniformity. The control set can be defined as representation of the whole underlying population distribution then an issue can arise due to inaccuracies introduced here may generate false results in the analysis. A different method can be used for this purpose which makes use of cartograms to correct for the non-uniformity in the background population data. If disease data is required to be transformed by the geography of background population density or by some means of population at risk, one would produce a distorted space map with the disease events still in their relative positions. The adjustment for variations in the background population density with aggregated count data can be achieved by using figures obtained from a population census.

1.1.3 Visualization of Detected Disease Clusters

Visualization of spatial distribution of the disease over a defined area helps user to analyse clusters easily and efficiently, and users can also detect unusual patterns of disease outbreaks. To study geographical variations of disease risk, the locations of cases are mostly proxies for residential addresses such as pin codes. If individual addresses are available, it is easy to plot the locations of cases on the map. In case Cartesian coordinates (GPS readings) are available, these can also be used. The most common type of map for visualization is Choropleth maps for spatial distribution of disease in the well-defined geographical area using health indicators as occurrence, incidence rate and mortality rate. Choropleth maps usually uses colour or pattern combinations to show different levels of disease risk associated with each geographical area. These geographical areas are small areas usually defined for administrative purposes, such as counties, zones, wards, colonies, villages, towns and cities. For visualization various geographical information systems can be used like QGIS which is a cross-platform free and open source desktop geographic information systems application.

1.2 Research Motivation

In this thesis report, the development of a user interface for SaTScan is presented, which is designed to provide an easier way to run SaTScan multiple times and add graphical output for analysing results obtained by this software. This standalone package comprises of following steps:

i It takes data sets or files of population, cases information, geographical coordinates of each location and optional controls for each location as an input in a simple prescribed format and generates text files in SaTScan format. .

ii It allows the user to choose SaTScan analysis options. Thirdly, it reads the results from SaTScan and creates geographical outputs, based on a separate map boundary file.

These steps are combined in one step, which allows the user to enter data sets which are in the form of text files acceptable by SaTScan, run the software, and generates output in graphical form. The frontend was being developed using PyQt4, and QGIS, with Python used as the interfacing programming language.

Effectiveness of the user interface is analysed by a case study of dengue fever cases in New Delhi from 2010-2012. Work presented in this report is motivated by the following facts:

i In India, health information exists at various levels, forms and systems. There is a wide variety of data that is collected by number of agencies mainly government both at the central and state level through routine data collection and also periodic sample surveys. While various initiatives meant to strengthen health information systems is underway, challenges continue in terms of reliability, relevance, timeliness, harmonization as well as quality of data.[1]

ii GIS can visualize complex spatio-temporal events and can help to analyse disease data on geographic maps consisting of several layers of information.[2]

iii Using thoughtful visualization aids will help public health officials, epidemiologists and researchers more efficiently identify and trace outbreaks of potential public health significance.[3]

iv For the case study data regarding dengue fever is collected as Dengue is the worlds most rapidly spreading mosquito-borne viral disease and is taking a far bigger human toll than was believed to be the case. India is the country with the worlds highest dengue burden, with about 34 per cent of all such cases occurring here. [4]

1.3 Research Aim

The aim of this research work is to promote the use of statistical analysis to detect disease outbreaks and to gain information regarding the outbreaks by visualizing them. With the help

[1] http://www.whoindia.org/en/section2_1522.html

[2] http://www.medscape.com/viewarticle/579612

[3] AMIA 2005 Symposium Proceedings Page - 967

[4] http://www.thehindu.com/sci-tech/health/policy-and-issues/india-leads-the-world-in-dengue-burden-nature/article4592098.ece

of work done in this thesis, proactive actions can be taken to prevent disease outbreaks. It is also helpful in identifying the hot zone of an epidemic. On the basis of information gathered from statistical analysis and visualization the overall quality of health of the nation can be improved. This application is particularly useful to health Officials who do not have knowledge of GIS software.

1.4 SaTScan Theory and Applications:

Kulldorff (1997) described a statistical method for the detection of multi-dimensional point process using spatial scan statistics. It uses variable window size and a baseline process as an inhomogeneous Poisson process or Bernoulli Process. Scanning window can be any predefined shape and is modelled on a geographical space G with a measure u. Monte Carlo sampling is done in which a regular or irregular grid of centroids covering the whole study region is created and then an infinite number of circles around each centroid are created, with the radius anywhere from zero up to a maximum so that at most 50 percent of the population is included. Actual and expected number of cases inside and outside the circle is obtained and Likelihood Function is calculated. By using Monte Carlo Simulation random replicas of the data set are generated under the null-hypothesis of no cluster. Likelihood function value is ranked with the maximum likelihood ratio from the Monte Carlo replications. These ranks are called p-values. For a cluster to be statistically significant its p-value should be less than one. The cluster with the smallest p-value is the most likely cluster that has occurred not by alone.

Kulldorff et.al. (2005) proposed a prospective space-time permutation scan statistic for the early detection of disease outbreaks by using only case numbers without any need for population at risk data. This method can be used prospectively to regularly scan a geographical region for outbreaks at any location and of any size. It monitors one-day or muti-days for any location and size in order to quickly detect a rapidly rising outbreak and still has the power to detect a slowly emerging outbreak by combining information from multiple days. The New York City Emergency Department Syndrome surveillance system is described as an example. In space-time permutation scan statistic method the input data is daily case counts for zip-code areas. The scanning window is in cylindrical shape with location as a base and time as a height .For each cylinder, the expected number of cases is calculated. Also, for a particular cylinder the expected number of cases and generalized likelihood ratio are calculated. Random replicas of the data set are conditioned on the marginal and by permutations of pairs of spatial locations and times. Maximum likelihood ratio and random data sets using Monte Carlo hypothesis testing are compared. Rank of the maximum generalized likelihood ratio from the real dataset is calculated. The radius of the circle is increased from zero to some maximum value which is to be defined by the user. The height of the cylinder represents the number of days with the requirement that the last day is always included with a variable number of preceding days up to some maximum limit defined by the user. In the paper, historical data for Diarrhoea Surveillance are considered. They detected citywide outbreak which started locally. The advantage of this method is ease of use. It requires case data and is capable to handling missing data.

One-dimensional scan statistic has been used for purely temporal disease surveillance and a spatial scan statistic has been used for purely geographical surveillance Kulldorff (2001).In this paper, a time periodic geographical disease surveillance system is proposed which is based on a space-time scan statistic. The statistical inference is adjusted for the multiple testing issue arising from multiple geographical regions and sizes of disease clusters. The proposed method can

detect spatial clusters irrespective of any predefined geographical boundaries by making a collection of close locations into the same cluster. The purpose of this method is to detect only those clusters that are still _"alive"_, which means that excess risk is still present during the last time period for which data was made available. It can detect clusters with moderately excess risk as well as recently emerged clusters with a high excess risk. The statistical inference adjusts for the many possible time periods for which analysis is done. Multiple testing is required for repeated time period analyses; so the surveillance system is developed such that it adjusts for large number of analyses that have already been performed. The method proposed in this paper is quite different from traditional space-time disease clustering methods. The traditional methods are retrospective in nature which is designed to test whether a disease is randomly distributed over space and time for a fixed geographical region during a fixed time period, whereas the proposed method is prospective in nature, with repeated time periodic analyses. In this paper Thyroid cancer incidence data for men in New Mexico was used to describe time periodic geographical disease surveillance system. All analyses were adjusted for age by using indirect standardization.

A two- dimensional scan statistic was also proposed to study purely spatial disease clusters Kulldorff et al., (1998). The space-time scan statistic is defined by a cylindrical window with a circular geographical base and with height corresponding to time. The base is present on one of the several possible centroids located throughout the study region with the radius continuously changing. The height indicates any possible time interval of less than or equal to half the study period. The window is then moved in both space and time so that for each possible geographic location and size, it also travels each possible time interval. For every cylinder,the number of cases inside and outside the cylinder is observed and thus indicatces the population at risk and relevant cofounders. The likelihood ratio is calculated for each cylinder. The cylinder with maximum likelihood ratio that has more cases than expected, is denoted as most likely cluster. Monte Carlo simulation is used to find out the significance of the reported clusters. The adjustments in confounding variables are made by calculating the expected number of cases in each area and time period through indirect standardization and conditioning on the total number of cases observed. If a temporal trend is observed then this trend can be adjusted by multiplying the covariates i.e., the expected number of cases for each geographical area and time period by the overall rate during that particular year. This method was applied to study brain cancer cases in Los Alamos. For each case county of residence, year of diagnosis, age in 5-years interval, race and gender was known. Incidence rates are adjusted for race, age and gender. Most of the cases of brain cancer in Los Alamos were by chance and hence were not statistically significant.

Kulldorff et. al. (1997) investigated whether the high breast cancer mortality was distributed evenly over the Northeast. They collected demographic data and age-specific breast cancer mortality rates for women from 244 counties in 11 north eastern states and for the District of Columbia for 1988-1992. In the paper, they used spatial scan statistic to obtain clusters of cases without specifying their size or location ahead of time. To discover the presence of disease clusters and to identify their approximate location, they used spatial scan statistic. In the paper they made an assumption that the number of deaths in each county follow Poisson distribution. Their method unfolds the increased breast cancer mortality on Long Island to be statistically significant and suggested that the increase was extended down to parts of New Jersey and Philadelphia. The proposed method is useful when encountering new cases of disease clusters.

In jankowska et al. (2008), a procedure named AMOEBA is proposed which is based on Multidi-

rectional Optimum Ecotope-Based Algorithm developed to identify hot and cold spots in mapped data by examining the spatial association of a mapped unit to surrounding units. AMOEBA is not designed to assign all units to clusters but to assign those units which are statistically significant. They illustrated their procedure by using data from the 2000 Ghana Census to identify social neighbourhoods in Accra, Ghana. The purpose of the algorithm is to detect a cluster from a selected seed location by calculating Gi* values of all locations near the seed. If the neighbour increases the Gi* value then it is included in the cluster. SaTScan limits cluster size by population while Tango and Takahashis 2005 allows limiting number of units in a cluster. These two identical features were introduced in AMOEBA and those are: (1) The maximum number of observations within a cluster is limited. (2) Allows a user to place a threshold on a variable of choice (e.g., area or population) where the cluster cannot surpass the sum of the selected variable.

Piarroux et al. (2011) present an analysis of cholera incidences in Haiti where spatial and temporal clusters were identified and analysed. They used regression model and discovered that cholera more severely affected communities in the coastal plains along the Artibonite River. They calculated cholera incidences by using population numbers mapped together with environmental settings by using ARCGIS. To investigate for space-time clustering, they analysed the daily case numbers in each Haitian community using SaTscan software. Their work suggested that contamination of Artibonite and one of its tributaries downstream from a military camp caused this epidemic. In the paper, they concluded that management of water and sewage is very important to prevent cholera spread.

In Odoi (2004), spatial scan statistic is used to identify spatial clusters and geographical correlation analysis which is used to discover associations of giardiasis rates with fertilizer application on agricultural land and livestock density. “Giardia lamblia” is a very common human intestinal parasite found in Canada with an estimation of 4-10% prevalence but infection rates vary by geographical area and sub-regions of high or low infection rates are thought to exist. The major routes of infection are the water-borne transmission. The statistical and geographical analysis described in the paper has four steps. First step is detection and identification of giardiasis clusters. SaTScan software was used to test for the presence of giardiasis spatial clusters and to identify their approximate location. Second step is Cartographic and GIS manipulations. All GIS manipulations and cartographic visualization is done in ArcView GIS. Mapping of geographical distributions of agricultural and land-use factors including livestock densities and the percentage of agricultural land on which fertilizer is applied is done. To calculate the critical intervals for livestock density and fertilizer Jenk’s optimization classification method is used .Third step is Geographical correlation analyses. At the Census Consolidated Sub-divisions spatial level data on livestock density and fertilizer application on agricultural land are made available. This geographical unit is used to correlate the land-use variables (Which are livestock density and fertilizer application on agricultural land) and giardiasis rates.

1.4.1 Issues with SaTScan

There are two issues related to SaTScan software itself. Those are : (1) Lack of cartographic support for interpreting the detected geographical clusters. (2) Outcomes being sensitive to parameter choices related to cluster scaling. The software does not directly provide any visualization support.
Chen et al. (2008) suggested that the geovisual analytics method make it efficient for users

to understand the SaTScan results. Authors have illustrated geovisual analytics approach in a case study analysis of cervical cancer mortality in the U.S. They have analysed the cervical cancer mortality data for the counties of United State between 2000 and 2004. For all the counties Standardized Mortality Ratio and reliability scores are visualized to identify stable and homogeneous clusters. The results of their analysis are compared with the results produced by other independent techniques including the Empirical Bayes Smoothing and Kafadar spatial smoother methods. The proposed geovisual analytics approach is implemented in Java-based Visual Inquiry Toolkit.

Further, running SaTScan frequently through its graphical user interface can be unmanageable and the output can be difficult to visualize. Abrams and Kleinman (2007) have developed and described a package called as SMAC within which SatScan software is used. This package makes the output more explanatory. The SMAC package comprises of four SAS macros which are developed to resolve difficulties in running SaTScan through its GUI. The first macro creates the input files in SatScan format, the second macro allows the user to choose the SatScan options and the third macro reads in the SatScan output and combines it with boundary files to create a summary of the most likely cluster, its statistical information as well as a list of the locations within the cluster. The boundary files are not created within the package and are obtained from another source. The fourth macro in the SMAC package calls all the macros one by one to create the input files, run SaTScan, and create the graphical output all in one step.

Chapter 2

Space Time Permutation Scan Statistic

2.1 Space - Time Permutation Model

2.1.1 Introduction

Space-time data can be analysed by retrospective and prospective analysis. Retrospective analysis is done only once for a predefined geographical region and time period while the prospective analysis is done to detect early disease outbreaks by repeating analysis every day, week, month or year and only alive clusters are detected. The retrospective analysis can be carried out using space-time permutation model which is used when only case data with information about the spatial location and time for each case is available with no information about controls or background population at risk. The number of observed cases in a cluster is compared with the expected number of cases as if the spatial and temporal locations of all cases were independent of each other which mean that there is no space-time interaction. This model is used to detect a cluster in a geographical area during a specific time period and the geographical area has a higher proportion of its cases in that time period compared to the rest of the areas. The space-time permutation model automatically adjusts for both purely spatial and purely temporal clusters therefore there are no purely temporal or purely spatial versions of this model.

2.1.2 Reasons of bias in results

There is a possibility that detected space-time permutation clusters are due to an increased risk of disease or due to different geographical population distribution at different time periods e.g., population in some areas grows faster than in others, so it is favourable to choose total study period to be less than a year. If the study period is more than a year then it is advised to be very careful when using this method. If the geographical population increases or decreases faster in some areas than in others then there is a risk for population shift bias which will produce biased p-values when the study period is longer than a few years. The bias occurs because space-time permutation model cannot distinguish an increase in geographical population due to a local population increase versus an increase in the disease risk. If an increase or decrease in the geographical population is same across the study region then it will not lead to biased results.
The space time permutation model can also give biased results due to missing data. It is

highly likely that some data could be missing for cases which were not reported to the hospital due to lack of awareness about treatment of the disease. Spatial uncertainty due to lack of or the error in knowledge about a case geographic position that is latitude or longitude may lead to uncertainty about the spatial relationship among its neighbours. For example an error in a patient's residential address may introduce spatial uncertainty about the location where the patient lives and this error will further bias any relationship between the patient's health information and specific environmental conditions

2.1.3 Adjustments for missing data in the model

In space-time permutation model it is quite complex to adjust for missing data. In the analysis file day-of-week should be added first as a covariate. When a location or time period is missing, then remove all data for the days of the week for which any data is missing then it is necessary to remove all data for the days of the week for which data is missing for that particular location. The similar method can be used with any other categorization of the data but the categorizations must be in any time-periodic unit that appears several times and is evenly spread out over the study period. Two more methods can be used to deal with missing data in the space-time permutation model and those are: (a) For selected region remove all data if some data are missing for that region. (b) Remove all data for a specific time period for dates on which there is missing data in any region. The latter is very useful in prospective surveillance when data is missing in the beginning of the study period to prevent removal of current data that are critical for the early detection of disease outbreaks.

In [3] a method for space-time permutation analysis is proposed which has been used in the present analysis. To detect most likely clusters by using space-time permutation model two types of data should be available: (a) case data (b) geographical region data. Case data should have the information about residential pin codes for each observed cases or the hospital address where the case was reported, dates for each observed cases and number of cases that occurred on a particular date and on a particular geographical region. The latitude and longitude geographical coordinates for each region with a study area should be available. Adjustments should be made for missing data by using any of the above discussed methods.

2.2 Algorithm of Space-Time Permutation Method

The algorithm for the Space-time permutation scan statistic can be described is as follows:

1. The input data is daily case counts for pin-code areas.
2. Let C_{zd} is the observed number of cases in pin code area z during day d.
3. The space-time analysis is done by using cylindrical window with a circular geographic base and with height corresponding to time.This cylindrical window is moved in both space and time, such that for each possible geographical region and size it also visits each possible time period. Thus it covers the whole study area.
4. For each cylinder, calculate the expected number of cases that is μ_{zd} as:

$$(\sum_z C_{zd} \sum_d C_{zd})/ \text{ c}$$

Where C is the total number of observed cases defined as:

$$\mathrm{C} = \sum_z \sum_d C_{zd}$$

5. For a particular cylinder A, the expected number of cases are: $\sum_{zd}$

$$\mu_{zd} = \sum_{(z,d)\in A} \mu_{zd}$$

6. For each cylinder calculate generalized likelihood ratio, that is T_A(if $C_A > \mu_A$) as:

$$T_A = \left\{ \begin{array}{r} \frac{C_A}{\mu_A}^{C_A} \frac{C - C_A}{C - \mu_A}^{[C-C_A]} \\ 1, otherwise \end{array} \right\} \tag{2.1}$$

This is the observed number of cases divided by the expected number of cases to the power of the observed number of cases inside the cylinder multiplied by the observed number of cases divided by the expected to the power of the observed outside the cylinder.

7. R = $max_A\, T_A$, among the cylinders evaluated, the cylinder with the maximum generalized likelihood ratio constitutes the spacetime cluster of cases that is least likely to be occur purely by chance. Thus, it is the primary cluster for a true outbreak.

8. Generate random replicas of the data set by permutations of pairs of the spatial locations and time periods.

9. P-value is calculated by comparing R and random data sets using Monte Carlo hypothesis testing:

$$\frac{R}{(1 + totalnumberofreplicas)} \tag{2.2}$$

The most likely cluster is calculated for each simulated dataset in exactly the same manner as for the real data. To examine statistical significance Monte Carlo hypothesis testing is used. P-value is calculated by comparing the maximum generalized likelihood ratio calculated from 999 simulated datasets to the maximum generalized likelihood ratio calculated from real data. If the p-value is less than one then the cluster is considered as statistically significant.

The radius of the circle is increased from zero to the maximum value that is predefined and the height of the cylinder represents the number of days, weeks, months or year with the requirement that the last day, week, month, year is always included together with the variable number of preceding days, weeks, months, years (Up to a maximum predefined value).

2.3 Limitation of Space-Time permutation model

Space-Time permutation model has some limitations i.e., as the method is highly sensitive to missing or incomplete data. To prevent false alarms while using space-time permutation model systematic data quality checks and the analytical adjustments should be done before using this method. If population at risk data is available, then spacetime analysis with Poisson model utilizes this extra information to perform better than the spacetime permutation scan statistic.

But if population at risk data is of poor quality or is not available, then the spacetime permutation scan statistic should be used. Since the spacetime permutation scan statistic makes adjusts for purely temporal clusters it can therefore detect outbreaks if they have started locally but if they occur simultaneously in the whole city then it will not detect. It is important to understand that the geographical boundary of the detected outbreak is not certainly the same as the boundary of the true outbreak.

All detected outbreaks are approximately circular because circles are used as the base for the scanning cylinder. It is possible to have scanning window of other shapes (e.g., elliptical) but the circular scanning window is the one used most often to detect outbreak areas.
To promote the use of this method worldwide, the method has been integrated into freely available software called SaTScan (Software for the Spatial and Space-Time Statistic).

Chapter 3

SaTScan-Software for the Spatial and Space-Time

3.1 SaTScan Software

SaTScan is a free software designed for geographical surveillance of disease to detect spatial, temporal or space-time disease clusters and to examine whether these clusters are statistically significant or not missing. It is used to perform discrete as well as continuous scan statistics. For discrete scan statistics the geographical locations where cases are observed are not random, i.e., is the location of each case is prefixed by the user. For continuous scan statistics the locations of the observations are random and can occur anywhere within a study area defined by the user. The type of analysis in SatScan can broadly be divided into two categories retrospective analysis and prospective analysis. Retrospective analysis is performed when past data is analysed to detect clusters of disease outbreaks. There are four types of retrospective analysis: Purely spatial, purely temporal, space-time and spatial variation in temporal trends. Prospective analysis is performed when clusters for early detection of disease outbreak is needed. There are only two types of prospective analysis possible: Purely temporal and space-time analysis. To perform the analysis a probability model is required. SaTScan provides seven probability models to perform discrete scan statistics: Poisson model, Bernoulli model, Multinomial model, Ordinal model, Normal model, Exponential and Space-Time permutation. To perform continuous scan statistics SaTScan uses a continuous Poisson probability model.

3.1.1 Data Requirements

The input data needs to be stored as in various files. In SaTScan five input files are required, but not all file is mandatory and a case file is needed for all probability models except for the continuous Poisson model. If Poisson probability model is used then population file is mandatory and if a Bernoulli probability model is used then control file is mandatory. For space-time, temporal analysis it is essential to have a time stamp related to each case. Grid file is always an optional file. The input file should be in SaTScan ASCII format or in dBase, comma delimited or space delimited files. If Poisson and space-time permutation model is used then it is possible to adjust for multiple categorical covariates by including them in case and population files. Covariates can be adjusted for using data sets in Bernoulli, ordinal or exponential models. As shown in Figure 3.1,the user interface of the software has three tabs: input tab, analysis tab and output tab. Input tab has various sections to accept input files. The required input files

are of following types:

1. Case file: The case file contains the information about observed cases of a disease within a study area. It includes following information:

    - Location id: It should be a numerical value or a string of characters.
    - Number of Cases: It is the number of observations or individuals with the disease for the specified location, time and covariates. It should be the total number of observations in the locations independent of the values of their continuous attribute for ordinal, multinomial, normal and exponential models.
    - Date/Time: It should be in years, months or days or in a generic format. The format should follow specified time precision format. In the case file all the cases with their specified times must fall within the study period which is specified on the Input Tab by the user.
    - Attribute: It is a variable that describes some characteristics of the case. It is needed for multinomial, ordinal, exponential and normal models. These may be covariates when discrete Poisson or space-time permutation model is used, category when multinomial or ordinal models is used, survival time or censored when exponential model is used, a continuous variable or weight value when normal model is used.
    - Censored: It is a variable with value 0 or 1 where 1=censored and 0=uncensored for exponential model.
    - Weight: It is required if covariates are used in normal model, even if all observations have the same variance.
    - Covariates: It is needed when discrete Poisson, space-time permutation and normal models are used for analysis. There can be any number of categorical covariates which can be specified as either numbers or characters. Covariates for the normal model is only included if weights are also present.

2. Control File: Control file is needed when Bernoulli Model is used. Multiple lines can be used for different controls with the same location, time and attributes. SaTScan add them automatically. It should contain the following information:

    - Location id: It should be a numerical value or a string of characters. Blank spaces are not allowed.
    - Controls: It should be the number of controls for the specified location and time.
    - Time: It should be in years, months or days. The time specified for each control should be within the study period. It should be in the same format as in the case file.

3. Population File: Population file is needed when Discrete Poisson model is used. Population file provides the information about the background population at risk. It can be actual population count or it could be covariate adjusted expected counts from a statistical regression model. It should contain the following information:

    - Location id: It should be a numerical value or string of characters. Blank spaces are not allowed.
    - Time: The time for a specific population size refers to. It should be in years, months or days, If the population time is unknown but identical for all population numbers, then a dummy year must be given, the choice should not affect the result.

- Population: For a particular location, year and covariate combination population size is given. It can be in decimal number to reflect a population size at risk rather than an actual number of people.
- Covariates: There can be any number of categorical covariates, each is expressed by a different column separated by blank spaces. It should be a numerical value or string of characters. It must be same as in the case file.

4. Coordinates File:The geographic coordinates for each Location ID is provided by a coordinate file. Each geographical location is presented by each line in the file. Area-based information may be aggregated and represented by one single geographical point location. Coordinates must be specified in any of the two formats which are standard Cartesian coordinate system or in latitude and longitude.

    - Cartesian Coordinates: It is the regular planar x,y-coordinate system .These may be specified in any number of dimensions. If Cartesian coordinates are used, then coordinates file should contain location id and coordinates which must be specified in the same units and there is no upper limit on the number of dimensions.
    - Latitude and Longitude: It should be in decimal number of degrees. Latitude represents the north or south distance from the equator and locations south of the equator should be presented as negative numbers. Longitude represents the east or west distance from the Prime Meridian and locations west of the Prime Meridian should be presented as negative numbers. If latitude or longitude coordinates are used, the coordinates file should contain the information about Location ID , latitude and longitude.

5. Grid File:The optional grid file defines the centroids of the circles used by the scan statistic. If no grid file is specified, the coordinates given in the coordinates file are used for this purpose.

3.1.2 Type of Analysis

Figure 3.2, shows the two types of analysis possible in SaTScan i.e., Retrospective analysis and Prospective analysis. Retrospective analysis is used when cluster are detected on the basis of past reported cases of disease. The types of retrospective analysis which can be performed in SaTScan are as follows:

1. Purely Spatial: This type of analysis is performed when the time of occurrence of cases is not considered only geographical location of the case is considered.

2. Purely Temporal: This type of analysis is performed when geographical locations of the cases are not considered but the time of occurrence of cases are considered.

3. Space-Time: This type of analysis is performed when fixed geographical locations of the cases within the fixed study area as well as time of the cases are considered.

4. Spatial Variation in Temporal Trends: This type of analysis is performed when spatial variations are required to be evaluated in temporal trends. This type of analysis can only be performed with discrete Poisson probability model. Prospective analysis is used for the early detection of disease outbreaks and this is done when analysis is repeated every day, week, month or year. The types of prospective analysis which can be performed in SaTScan are as follows:

5. Purely Temporal: This type of analysis is repeatedly performed when daily, weekly, monthly or yearly times of the cases are considered.

6. Space-Time: This type of analysis is repeatedly performed when geographical locations as well as times of the cases are considered.

3.1.3 Probability Model

As shown in Figure 3.2, there are eight different probability models that can be used to perform the analysis which are as follows:

1. Discrete Poisson Model: It should be used when the background population is required to be considered.

2. Bernoulli Model: It should be used when the data set has the information about number of individuals who may or may not have a disease in the form of 0 or 1.Those who have the disease are cases and should be listed in the case file and those who are without the disease are controls and should be listed in the control file. The Bernoulli model is a special case of the multinomial model or ordinal model when there only two cofounding variables are considered.

3. Space-Time Permutation Model: It should be used when only case data is present and when an individual wants to adjust for purely spatial and purely temporal clusters.

4. Multinomial Model: It is used when individuals belong to one or more than one categories and when there is no ordinal relationship between them.

5. Ordinal Model: It is used when individuals belong to one or more than one categories and when there exist an ordinal relationship between those categories such as small, medium and large.

6. Exponential Model: It is used for survival time data to search for spatial or temporal clusters of short or long survival. The survival time is a positive continuous variable.

7. Normal Model: It is used for continuous data. Observations may be either positive or negative.

8. Continuous Poisson Model: It should be used when observations are distributed randomly with constant intensity according to a homogeneous Poisson process over a predefined study area.

3.1.4 Scan for High or Low Rates

In SaTScan it is possible to perform analysis to detect clusters with high rates only or low rates only or simultaneously for areas with either high or low rates. When continuous Poisson model is used then it is only possible to scan for high rates. In case of spatial variation in temporal trends scan statistic only clusters with a trend that is higher or lower than a trend outside the cluster is reported. It looks for the clusters with high trend only or low trend only or simultaneously for both.

3.1.5 Time Aggregation

Time Aggregation is used to reduce computation time by aggregating case times into time intervals. The units in which the length of the time intervals are years, months, days or generic. The units of the time intervals cannot be more precise than the time precision specified in the software.

3.1.6 Outputs of SaTScan

As shown in Figure 3.3, Output tab has an option to choose a drive where user wants to save the output file. The output file generated in SaTScan is a text based output file in ASCII format and five different optional output files in column format which can be generated in either ASCII or dBase format. Output files are:

1. Standard Results File (*.out.*): This file is automatically shown when all the calculations are completed like in Figure 3.4 . It has following information:
    - Summary Of Data: This is used to check that the input data files contain the correct number of cases, locations etc.
    - Total population: Average population during the study period this is used when discrete Poisson model is used.
    - Annual rate per 100,000: This is evaluated taking leap years into account and is based on the average length of a year of 365.2425 when discrete Poisson model is used.
    - Variance: This is calculated for all observations in the data assuming common mean when normal model.
    - Most Likely Cluster: Summary information about the most likely cluster; it is the cluster that is least likely to be due to chance.
    - Radius: When latitude and longitude are used in coordinate file the radius of the circle is given in kilometres. When Cartesian coordinates are used in coordinate file then the radius of the circle is given exactly in the same units as given in the coordinates file.
    - Population: It represents the average population in the geographical area of the cluster.
    - Relative Risk: It is calculated as the ratio of estimated risk within the cluster to the estimated risk outside the cluster.
2. Cluster Information File (*.col.*): In the cluster information file, the information about each cluster is represented on one line. The exact columns included in the file depend on the chosen analysis but it can be verified by comparing it with the standard results file.
3. Stratified Cluster Information File (*.sci.*): In the stratified cluster information file, In each data set for each cluster there is one line for each ordinal or multinomial category. For each data set combination there is one column each for the observed number of cases, the expected number of cases observed divided by expected and the relative risk.
4. Location Information File (*.gis.*): It a special output file which is created to describe various clusters that can be integrated into a geographical information system (GIS). This file includes Columns with information about Location ID, Cluster Number, p-Value of Cluster, Observed Cases in Cluster, Expected Cases in Cluster, ratio of Observed and

Expected cases in Cluster, Observed Cases in Location, Expected Cases in Location and ratio of Observed and Expected cases in Location.

5. Risk Estimates for Each Location File (*.rr.*): The risk estimated for each location is represented in risk estimates for each location file. This file includes columns which information about Location id, Observed Cases, Expected Cases, ratio of Observed and Expected cases and Relative Risk.

6. Simulated Log Likelihood Ratios File (*.llr.*):The log likelihood ratio test statistics is calculated from random data sets which are not provided as part of the standard output. This is presented in an optional simulated log likelihood ratio file when needed.

3.1.7 Advanced Features

SaTScan analysis is performed by using three basic tabs which are input, analysis and output tabs. But SaTScan also provides various additional features which are available as advanced features. Each tab has an advanced option. According to the analysis adjustments are made like if user wants to use a specific method to calculate p-value then this can be changed in inference tab in advanced analysis feature as shown in Figure 3.5 . There are three options except default option from which user can make a choice. In the similar manner, there are various additional features present in the software for the users who wants to make other choices then default settings

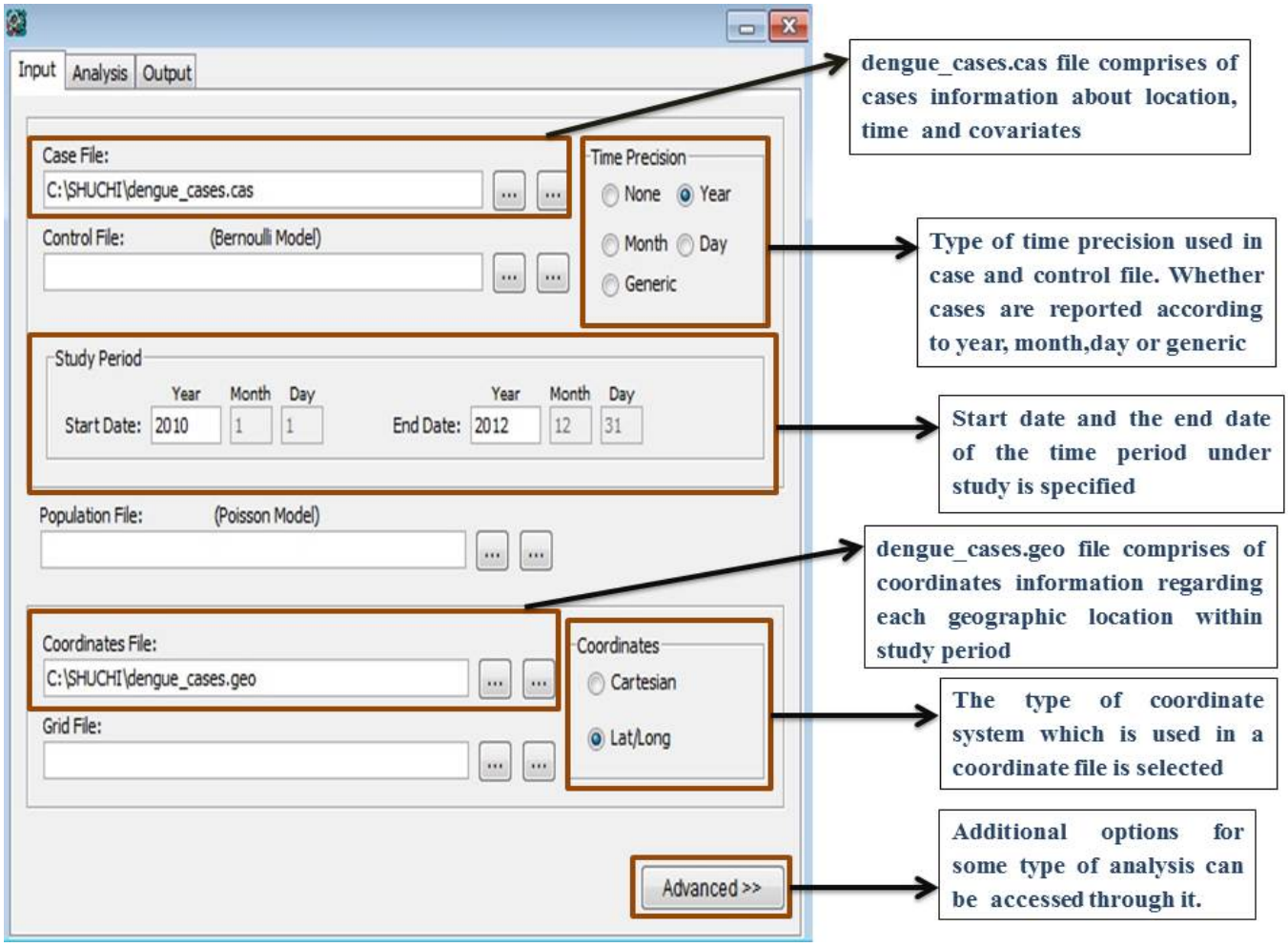


Figure 3.1: Input Tab

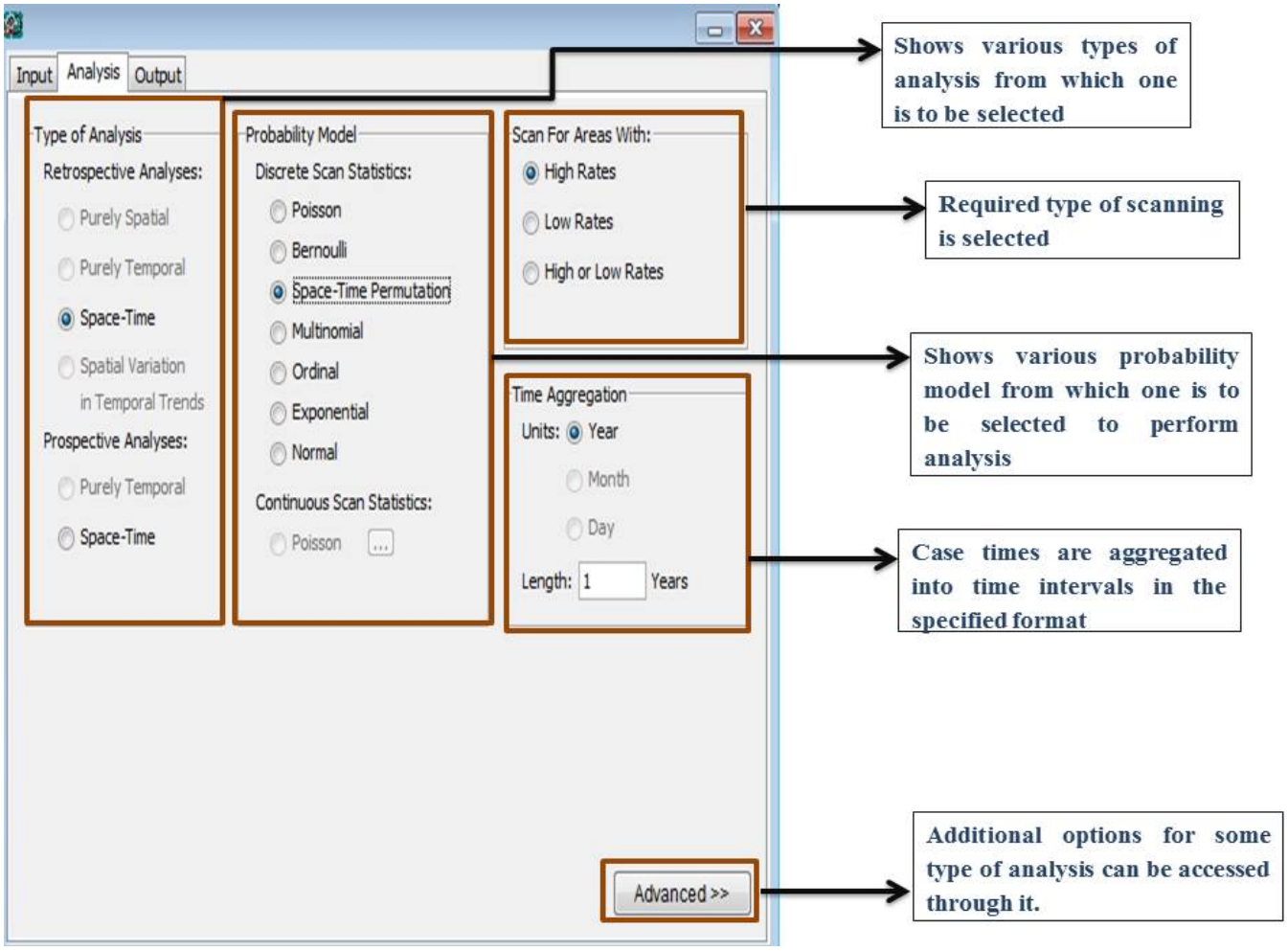


Figure 3.2: Analysis Tab

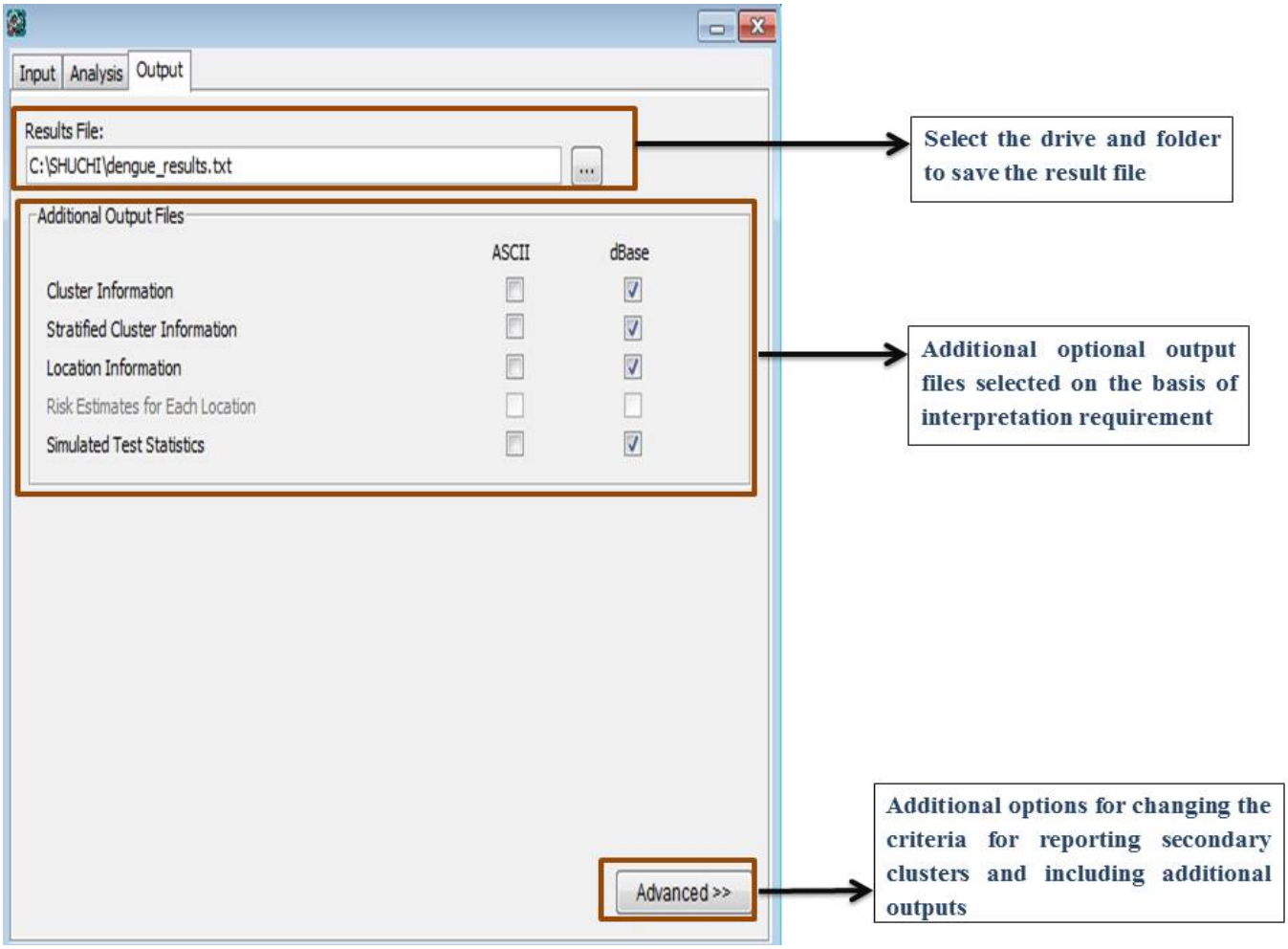


Figure 3.3: Output Tab

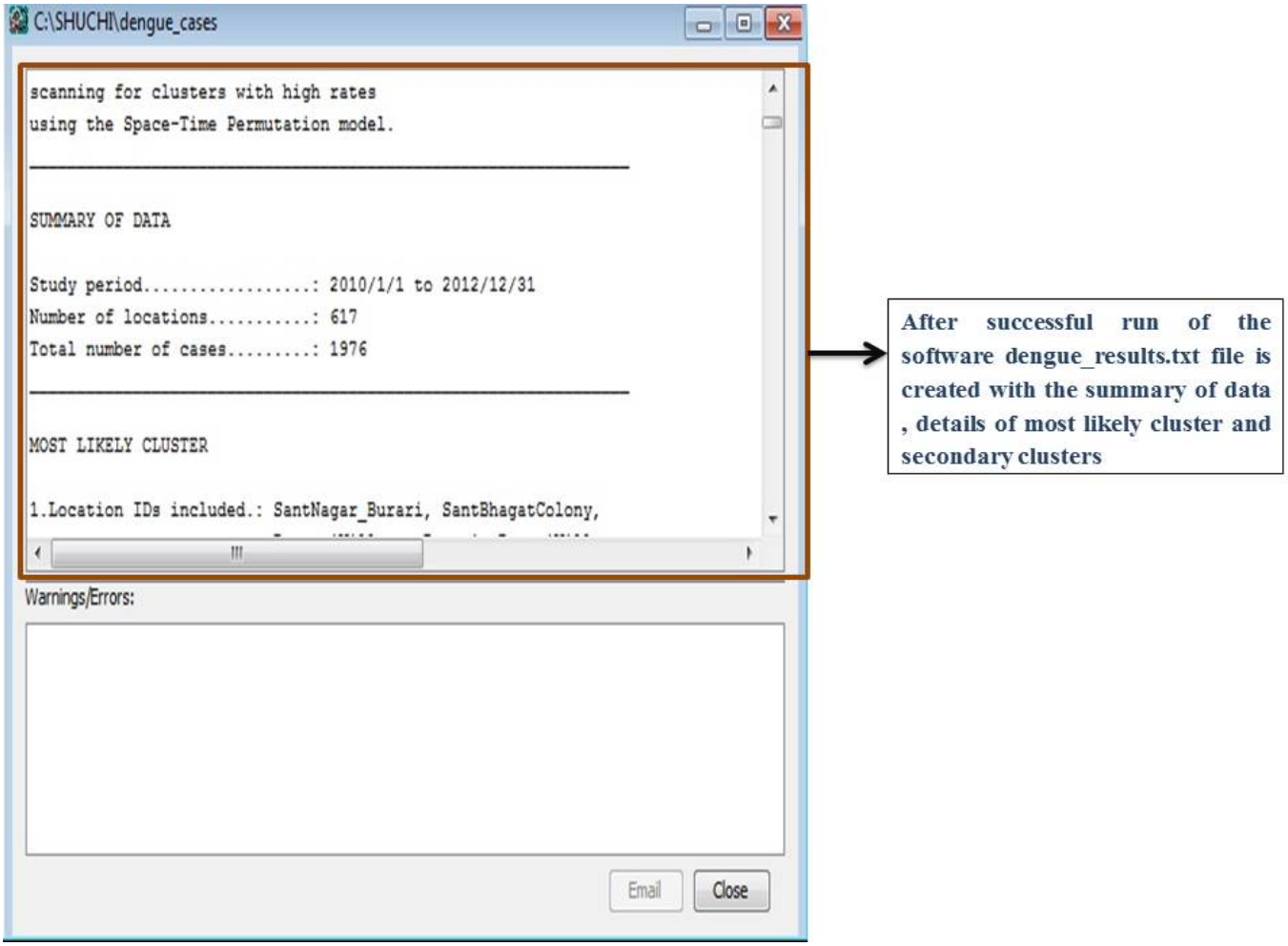


Figure 3.4: Result File

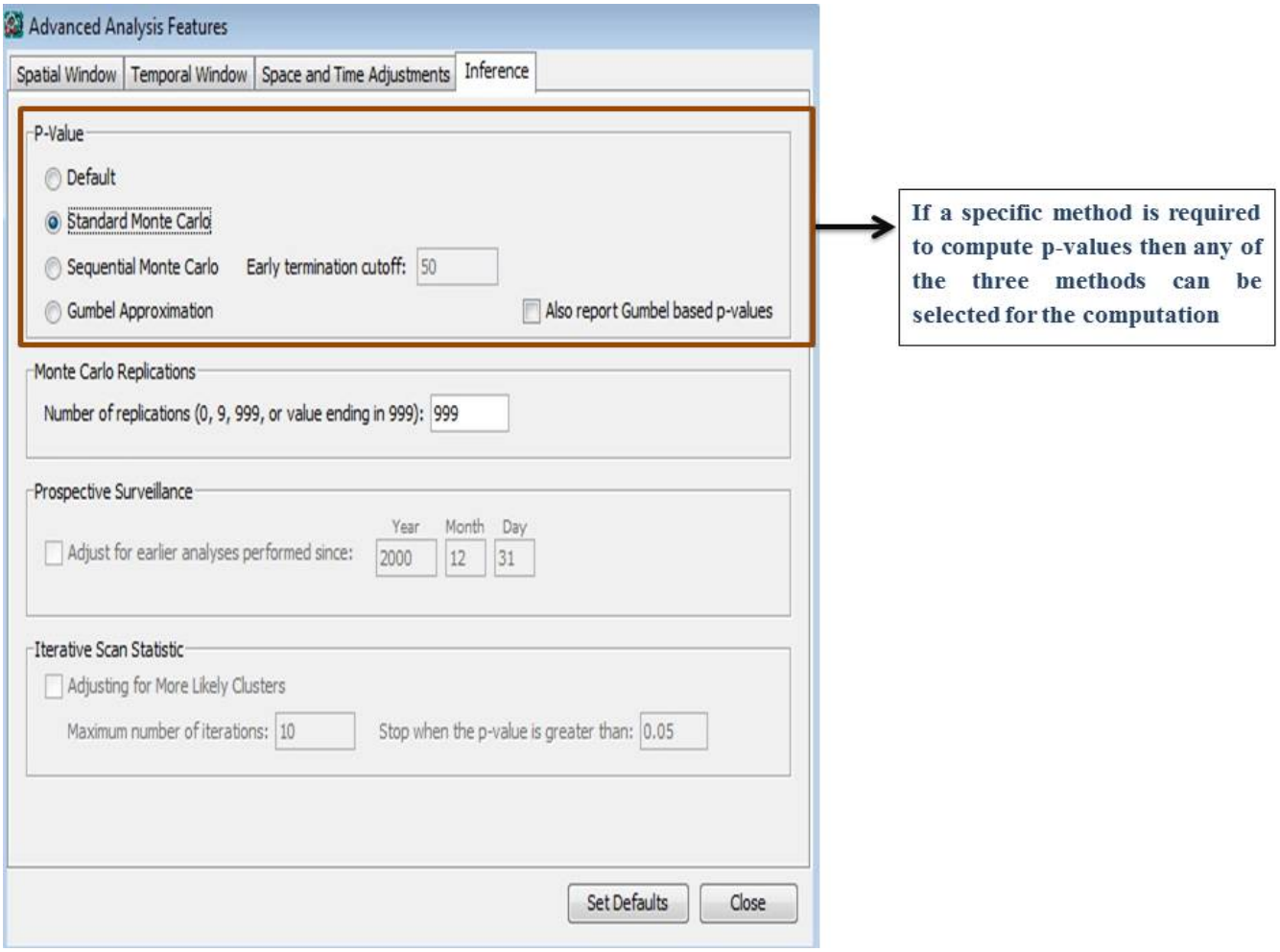


Figure 3.5: Advanced Analysis features

Chapter 4

Results, Discussion and Conclusion

4.1 Need for visualization

In the field of public health, spatial clustering analysis and subsequent geoprocessing of clustering results is the most efficient yet technically comprehendible way. SaTScan software[12], offers many advantages. It is robust, computationally efficient, has flexibility of options, corrects for multiple comparisons, adjusts for heterogeneous population densities among the different areas in the study, detects and identifies the location of the clusters without prior specification of their suspected location or size thereby overcoming pre-selection bias, and allows for adjustment for covariates. However, one of the drawbacks of SaTScan is that it does not have a visualization system for presenting the results. In this regard GIS has sophisticated mechanisms to visualize data. In addition to being able to asses disease cases with a general categorical definition of "place", GIS systems provide the means to analyze spatial/temporal relationships between sets of variables, allowed users to identify spatial patterns in data, and provided the means to integrate databases on the basis of geography. But it is a time-consuming process and one has to learn how to work with GIS packages.

4.2 Geographical Visualization Approach

Here, a standalone package is developed which provides a single integrated map that combines the geographical location of diseases and clusters to enhance the understanding of SaTScan results. It does not require a knowledge of GIS packages. GIS support is provided by QGIS library which provides many spatial algorithms and native GIS functions. This library is accessed through PyQGIS by Python bindings which provides simpler programming environment. For developing GUI, PyQt4, Qt4-devel, Qt4-doc and Qt4-libs have been used (PyQt is a Python binding of the cross-platform GUI toolkit Qt)

The relationships between software components of the developed standalone application termed "Visual Interpretation of Statistical Analysis" **(VISA)** is shown in Figure 4.1 is shown in Figure4.1

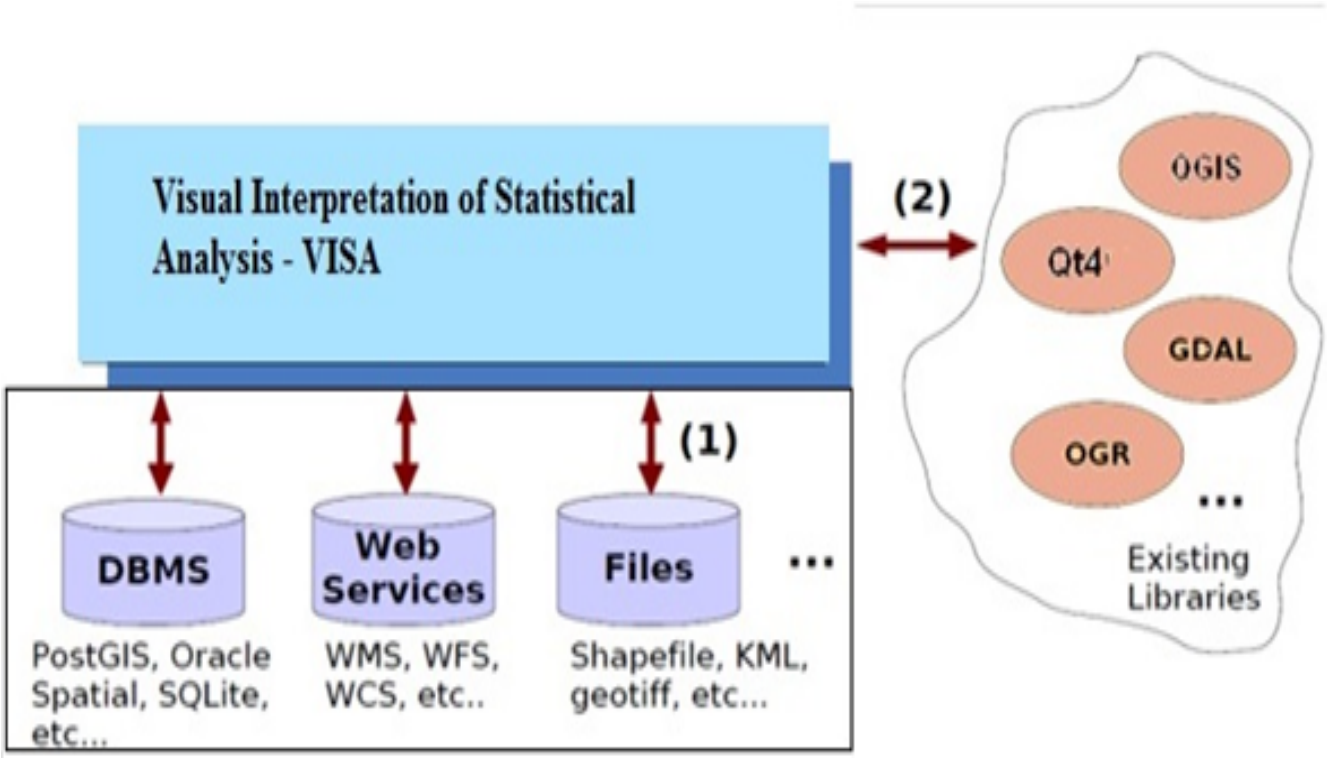


Figure 4.1: Structure of Standalone Application VISA

As shown in Figure 4.1 and4.2 are developed in the proposed standalone package.

4.3 Development of GUI

The packages PyQt4, Qt4-devel, Qt4-doc and Qt4-libs which are development tools, libraries and documentation are used for developing GUI for the standalone package VISA . The Qt Designer is used which allows creation and modification of a form in WYSIWYG (WYSIWYG implies a user interface that allows a user to view something quite similar to the end result while the document is being created.) without editing the code and can be converted into the code in Python. Once this code is converted, any modification needed can be done in the Python program editor.
Figure 4.2 shows a sample form that was developed using the methods described in Appendix-A

4.4 Development of Application

After creating the form python code was written for developing a standalone application. Some statements of the program are given in parts along with the explainations.:

The first line indicates the encoding of the source text and the

-- coding: utf-8 -* -
based on C + + Tutorial (C) 2006 Tim Sutton
Porting to Python: Martin Dobias, with the participation of Gary Sherman, changes to

FOSS4G2007
Adapted and modified by Shuchi Mala
#
Imports modules Qt, QGIS and the form created in section 5.2.

```
from PyQt4 import QtCore, QtGui
from qgis.core import *


from qgis.gui import *
from mainwindow_ui import U_MainWindow
```

A new class SubWindow0 is created

```
class SubWindow0( QtGui.QMainWindow, Ui_MainWindow ):
def __init__( self ):
QtGui.QMainWindow.__init__( self )
```

```
# Requires Qt4 to initialize the UI
self.setupUi( self )
```

Set the title of the window, create a map with background colour.

```
# Set the window title
self.setWindowTitle( "Delhi Map with Zones" )



# create a map
self.canvas = QgsMapCanvas()
# Change the background color card
self.canvas.setCanvasColor( QtGui.QColor( 200, 200, 255 ) )
self.canvas.enableAntiAliasing( True )
self.canvas.show()
```

Arrange the components of widgets in the application window

```
# The widgets in a box
self.layout = QtGui.QVBoxLayout( self.frame )
self.layout.addWidget( self.canvas )
```

Next, define actions for each signal and connect them with appropriate methods.

```
# Specify the action for each of the instruments and connect them with appropriate methods
self.connect( self.mpActionAddLayer, QtCore.SIGNAL( "triggered()" ), self.addLayer );
self.connect( self.mpActionZoomIn, QtCore.SIGNAL( "triggered()"), self.zoomIn );
self.connect( self.mpActionZoomOut, QtCore.SIGNAL( "triggered()" ), self.zoomOut );
self.connect( self.mpActionPan, QtCore.SIGNAL( "triggered()" ), self.pan );
self.connect( self.mpActionZoomFullExtent, QtCore.SIGNAL( "triggered()" ), self.zoomFull );
```

Create toolbars wich will add to the actions.
# New Toolbar
self.toolbar = self.addToolBar(“Map”);
# Adds an action to the toolbar
self.toolbar.addAction(self.mpActionAddLayer);
self.toolbar.addAction(self.mpActionZoomIn);
self.toolbar.addAction(self.mpActionZoomOut);
self.toolbar.addAction(self.mpActionPan);
self.toolbar.addAction(self.mpActionZoomFullExtent);

Create the tools.
self.toolPan = QgsMapToolPan(self.canvas)
self.toolPan.setAction(self.mpActionPan)
self.toolZoomIn = QgsMapToolZoom(self.canvas, False) # false = reduce
self.toolZoomIn.setAction(self.mpActionZoomIn)
self.toolZoomOut = QgsMapToolZoom(self.canvas, True) # true = increase
self.toolZoomOut.setAction(self.mpActionZoomOut)

Form procedure for the tools.
# Set the tool to increase
def zoomIn(self):
self.canvas.setMapTool(self.toolZoomIn)

# Set the tool to reduce
def zoomOut(self):
self.canvas.setMapTool(self.toolZoomOut)

# Set the tool to move
def pan(self):
self.canvas.setMapTool(self.toolPan)

# Set the instrument display card is fully
def zoomFull(self):
self.canvas.zoomFullExtent()

To load the layer following statements are used.
# Add OGR-layer
def addLayer(self):
# add a (hardcoded) layer and zoom to its extent
self.color = QtGui.QColor(“white”)
self.update()
# The layer must be in the same folder as the application file
# Layer name in the code
layerPath1 = “district boundary.shp”
layerName1 = “district boundary”
layerProvider = “ogr”

```
# Create a layer
layer1 = QgsVectorLayer( layerPath1, layerName1, layerProvider )
if not layer1.isValid():
print “Layer failed to load!”
return

# Add layer to the registry
QgsMapLayerRegistry.instance().addMapLayer( layer1 );

# Set equal to the coverage map coverage layer
self.canvas.setExtent( layer1.extent() )
# specify a set of
c1 = QgsMapCanvasLayer( layer1 )
layers = [ c1]
self.canvas.setLayerSet( layers )
```

To load the layer following section of the program is used:

```
wnd0 = SubWindow0()
wnd0.show()
wnd0.exec_()
```

If the layer is the map of Delhi with zone, the above program will give the result as shown in Figure 4.2 which contains tool boxes for display Map, ZoomIn, ZoomOut, Pan and zoomfull toolboxes.

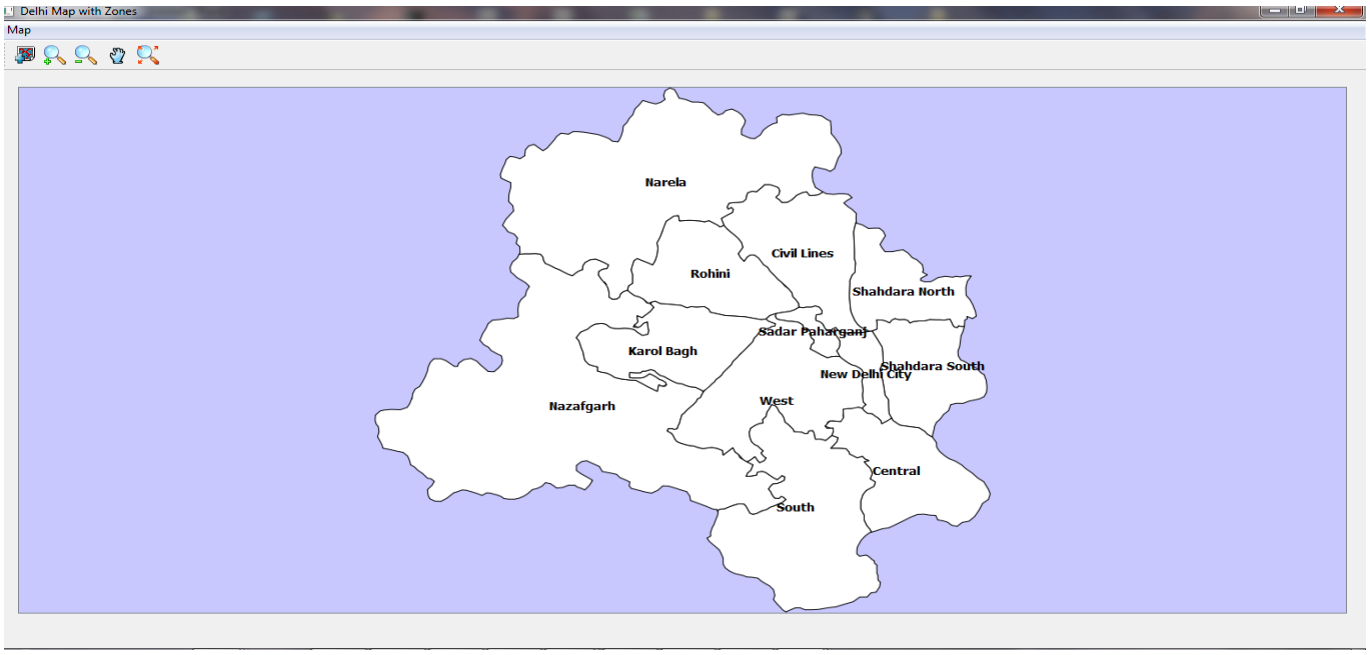


Figure 4.2: Screenshot Of The District Boundary Map

The above mentioned procedure is used for loading other layers.

4.5 Data Preparation and Execution

The data preparation methodology for the application can be described as follows:

1. In application folder there is a text file yeardata.txt which is modified according to the study period. If Dengue fever analysis is done for three years (2010-2012) separated by comma.This is a requirement of SaTScan data and hence cannot be incorporated into VISA.

2. For analysis in SaTScan two main files are needed. If either is unavailable then it is not possible to run SaTScan. The two main files are: case file and coordinate file. Case file contains information about number of cases, time at which case was reported and covariates like age, gender etc. Coordinate file contains name of location and its longitude and latitude or Cartesian coordinates. These two files are stored in a folder. Now run VISA application, which will show the following menu with buttons of disease cases maps and disease density maps for those years which were entered before running it as shown in Figure 4.3

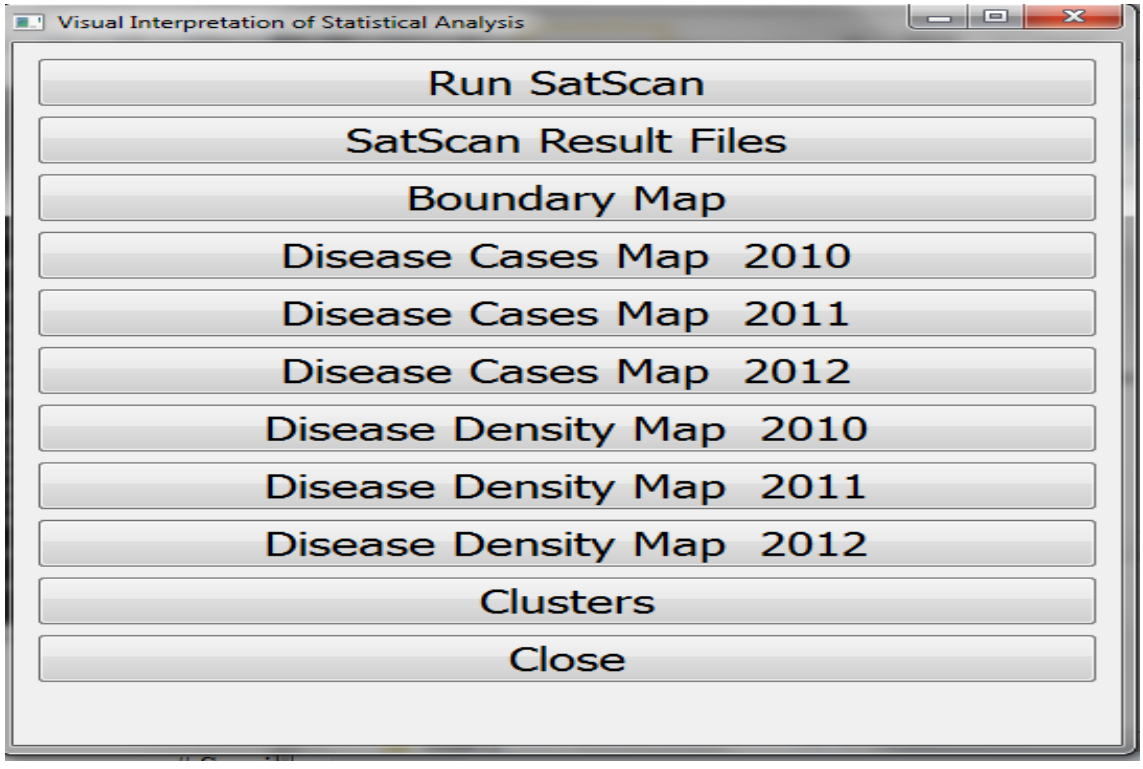


Figure 4.3: Main Menu

- When Main Menu appears click Run SaTScan button .
- Now SaTScan menu will be displayed.
- Select new session.
- In the Input menu, case file and coordinate files are selected

- Study period is entered.
- In the Analysis menu, type of analysis and probability model is selected. Under Advanced, a Circle with radius is selected and value of radius is entered.
- In Output menu name of result file is entered and additional output files are marked. Now the run button is pressed.
- Results are stored in four output files. For example in the presented work analysis is done on dengue fever outbreaks from 2010 to 2012 in Delhi, Consequently the four output files generated are: dengue_cases.sci.txt, dengue_cases.llr.txt, dengue_cases.gis.txt and dengue_cases.col.txt. The output files are stored in the same folder.

3. When SaTScan Result Files is clicked, in which SaTScan result with extension col.txt is selected .

4. On clicking Boundary Map, shape file of state or county or country map is imported in the above program as layer and displayed. Since here the analysis of dengue cases in Delhi is done then Delhi Map with Zones is displayed as shown in Figure 4.2

5. On clicking Disease Cases Map 2010, first of all, function AppLatLonCase.py is called which extract cases of 2010 and append longitude and latitude values for each case. A separate file case_lt2010.csv is generated. This csv file is read in the program and the layer showing cases which occured in the year is 2010 displayed as shown in Figure 4.4

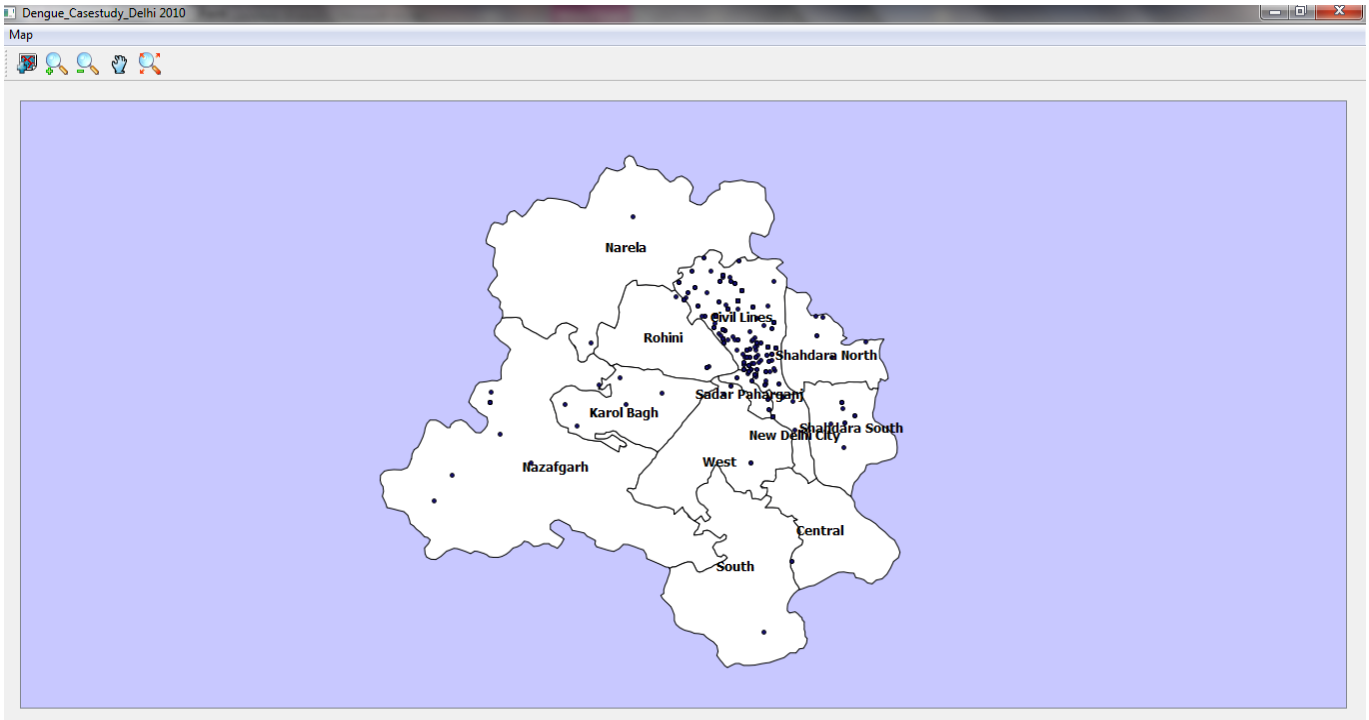


Figure 4.4: Dengue cases in Delhi in the year 2010

6. On clicking Disease Cases Map 2011 and Disease Cases Map 2012, the operations mentioned in step 5) are repeated for cases occurring in the years 2011 and 2012 as shown in Figure 4.5 and Figure 4.6 respectively

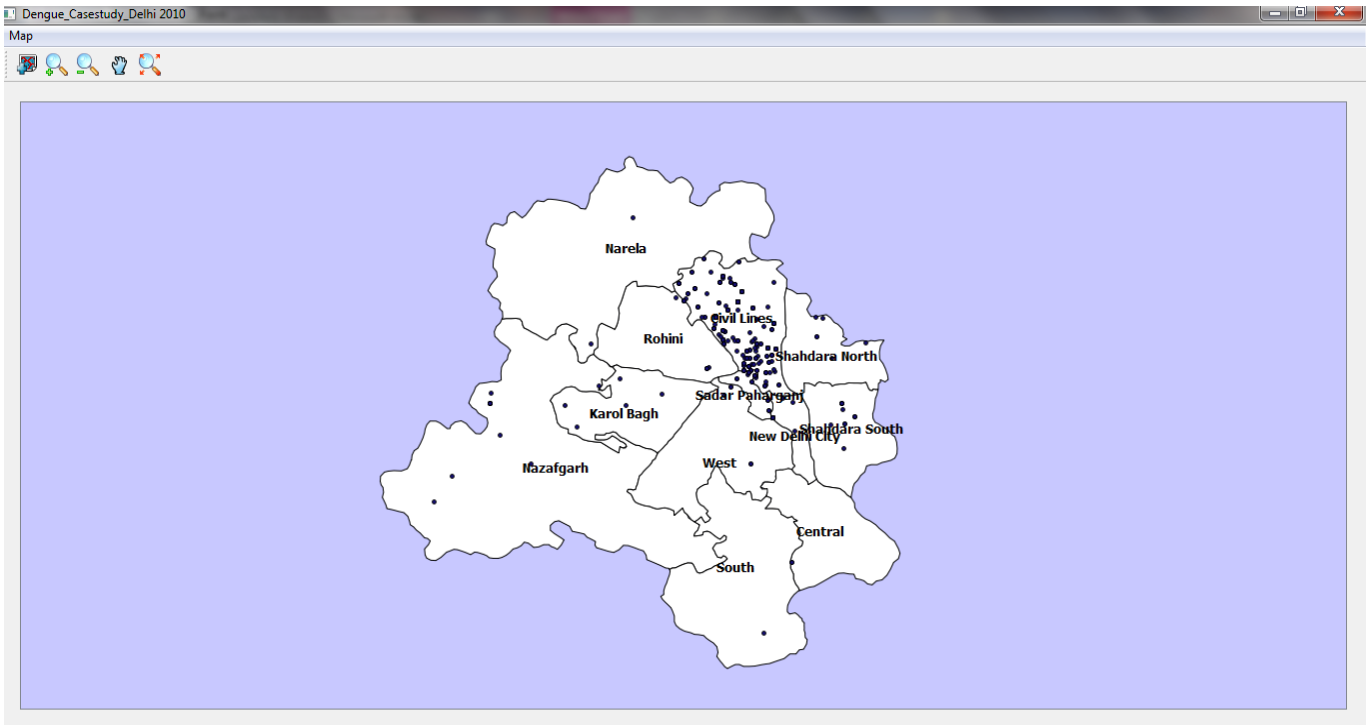


Figure 4.5: Dengue cases in Delhi in the year 2011

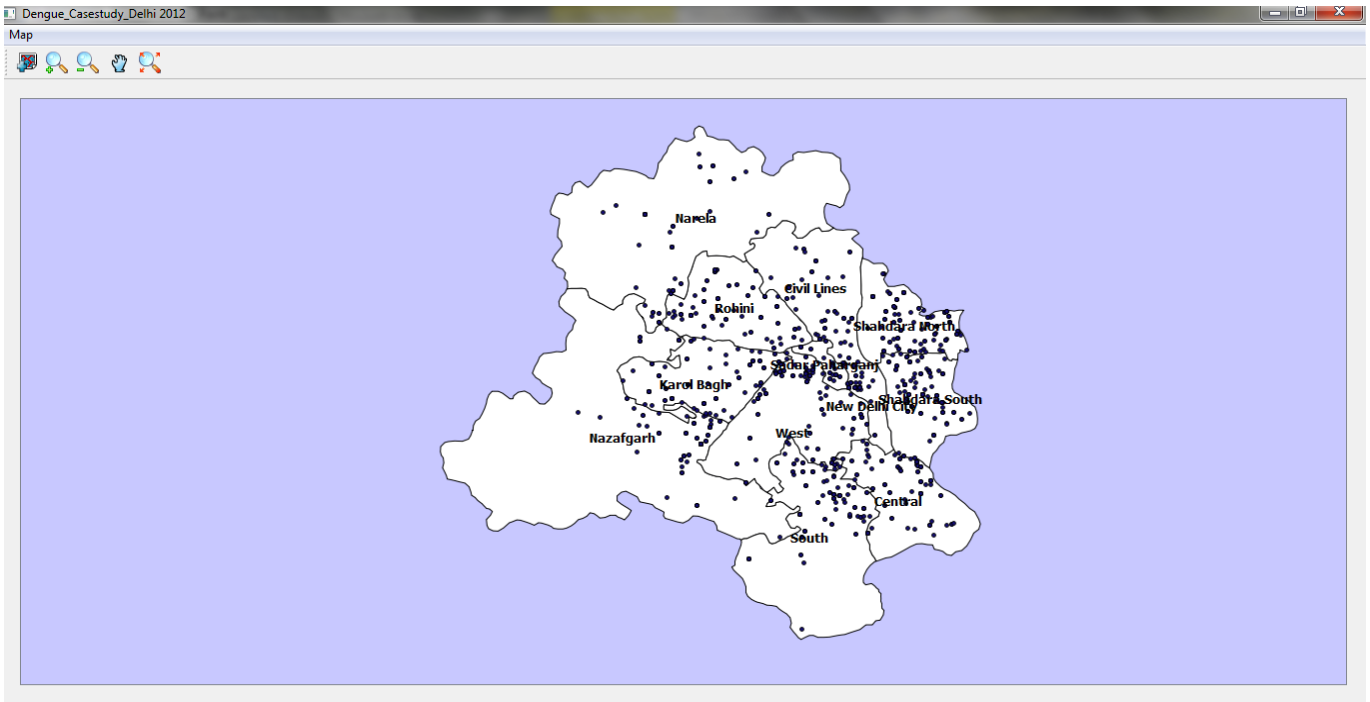


Figure 4.6: Dengue cases in Delhi in the year 2012

7. On selecting Disease Density Map 2010, all the dengue cases in the year 2010 in Delhi are

considered. A program computepoint.py is called to calculate number of dengue cases in each zone, A shape file is created and displayed which is shown in Figure 4.7 . Number of cases which occurred in the year 2010 in each zone is also displayed.

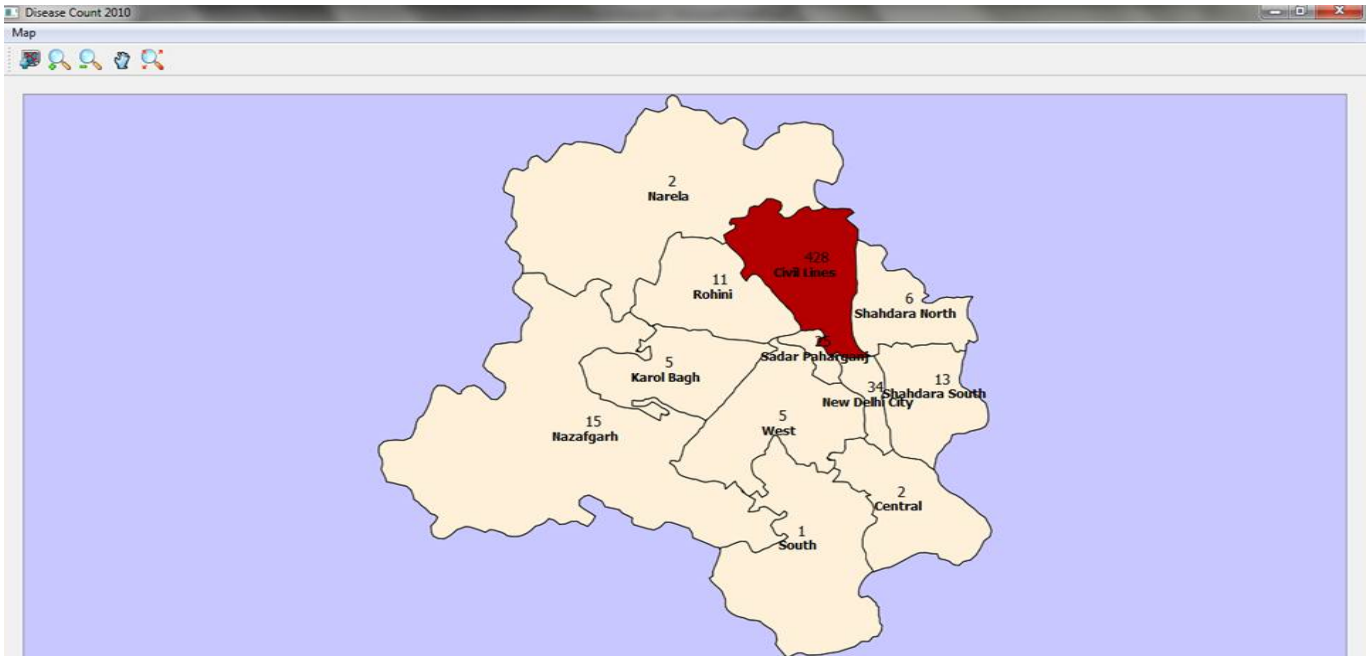


Figure 4.7: Dengue cases in 2010 is shown by graduated color

i On clicking Disease Density Map 2011 and Disease Density Map 2012, the operations mentioned in step 7) are repeated for cases occurring in the years 2011 and 2012 as shown in Figure 4.8 and Figure 4.9 respectively.

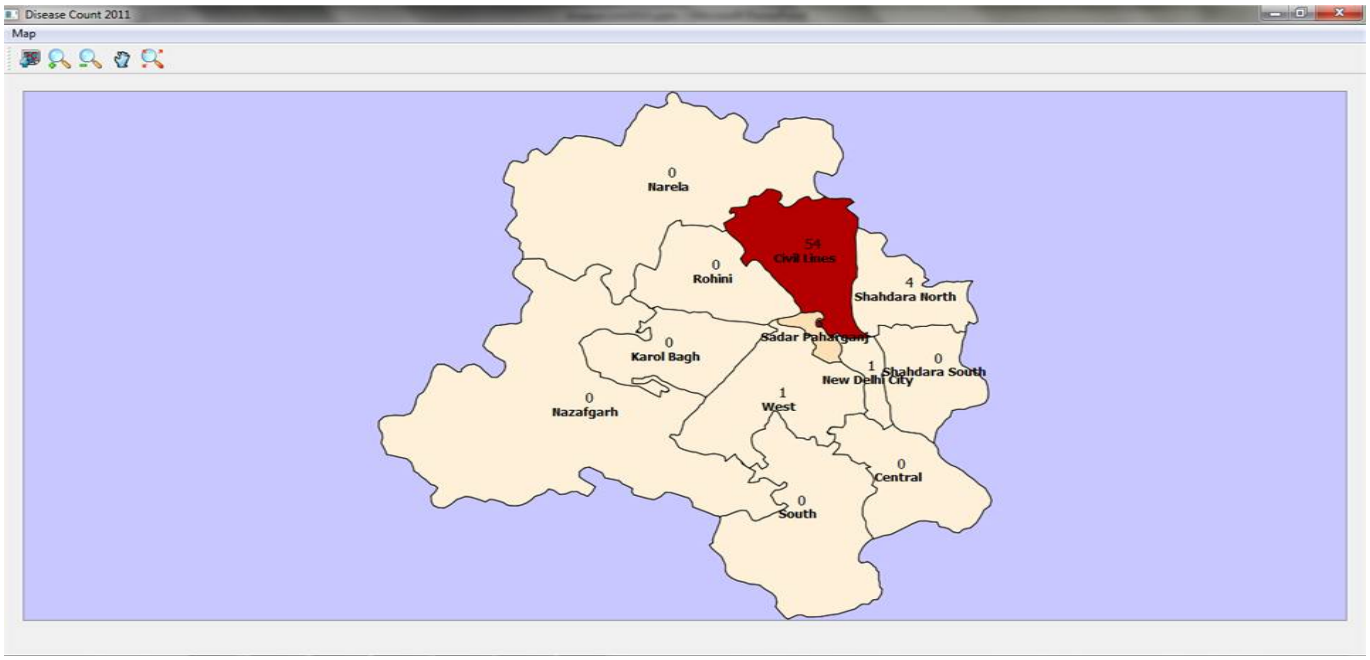


Figure 4.8: Dengue cases in 2011 is shown by graduated color

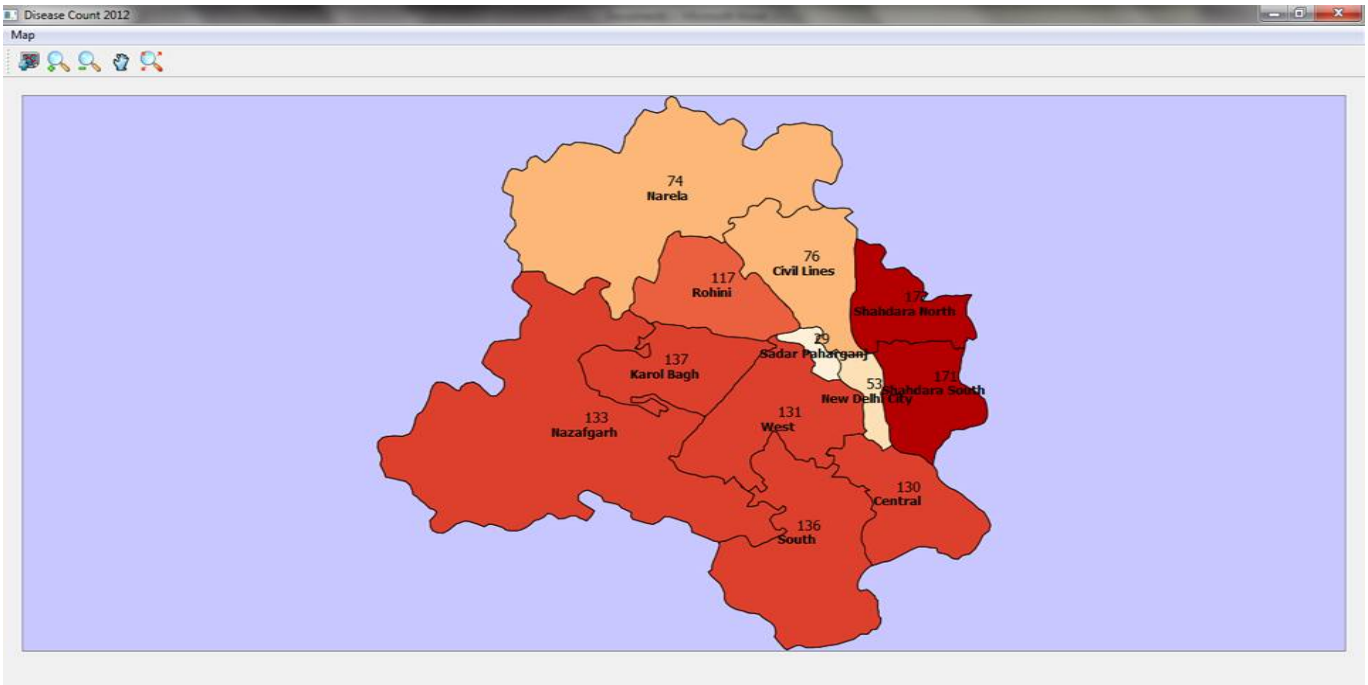


Figure 4.9: Dengue cases in 2012 is shown by graduated color

ii On clicking close, the application is terminated.
In the developed application, several free softwares have been used. SaTScan software which is freely available has been called in the application. The developed package is a user-friendly tool with automated GIS functionalities utilizing the core and gui libraries of QGIS.

iii When Cluster button is clicked, windowSub3.py is called. This program reads the SaTScan result file dengue_cases.col.txt and forms a file cluster.csv which contains all cluster having p-value less than a specified value. This csv file is read and a layer is created. Following statements are used for forming a buffer of a specified radius
outputFilename = "cluster.shp"
bufferLength = 0.0166
polygonSides=20
provider = layer4.dataProvider()
fields = provider.fields()
writer = QgsVectorFileWriter(outputFilename, "CP1250", fields, QGis.WKBPolygon,
provider.crs(), "ESRI Shapefile")
inFeat = QgsFeature()
outFeat = QgsFeature()
inGeom = QgsGeometry()
provider.select(provider.attributeIndexes())
while provider.nextFeature(inFeat):
point = inFeat.geometry().asPoint()
inGeom = inFeat.geometry()
outFeat.setGeometry(QgsGeometry.fromPolygon(
QgsPoint(point[0]+np.sin(angle)*bufferLength, point[1]+np.cos(angle)*bufferLength)
for angle in np.linspace(0,2*np.pi,polygonSides, endpoint=False)]]))
outFeat.setAttributeMap(inFeat.attributeMap())

```
writer.addFeature( outFeat )
del writer
newlayer = QgsVectorLayer(outputFilename, "Polygons", "ogr")
newlayer.loadNamedStyle('newlayer.qml')
```

The created layers are displayed which is shown in Figure 4.10 .

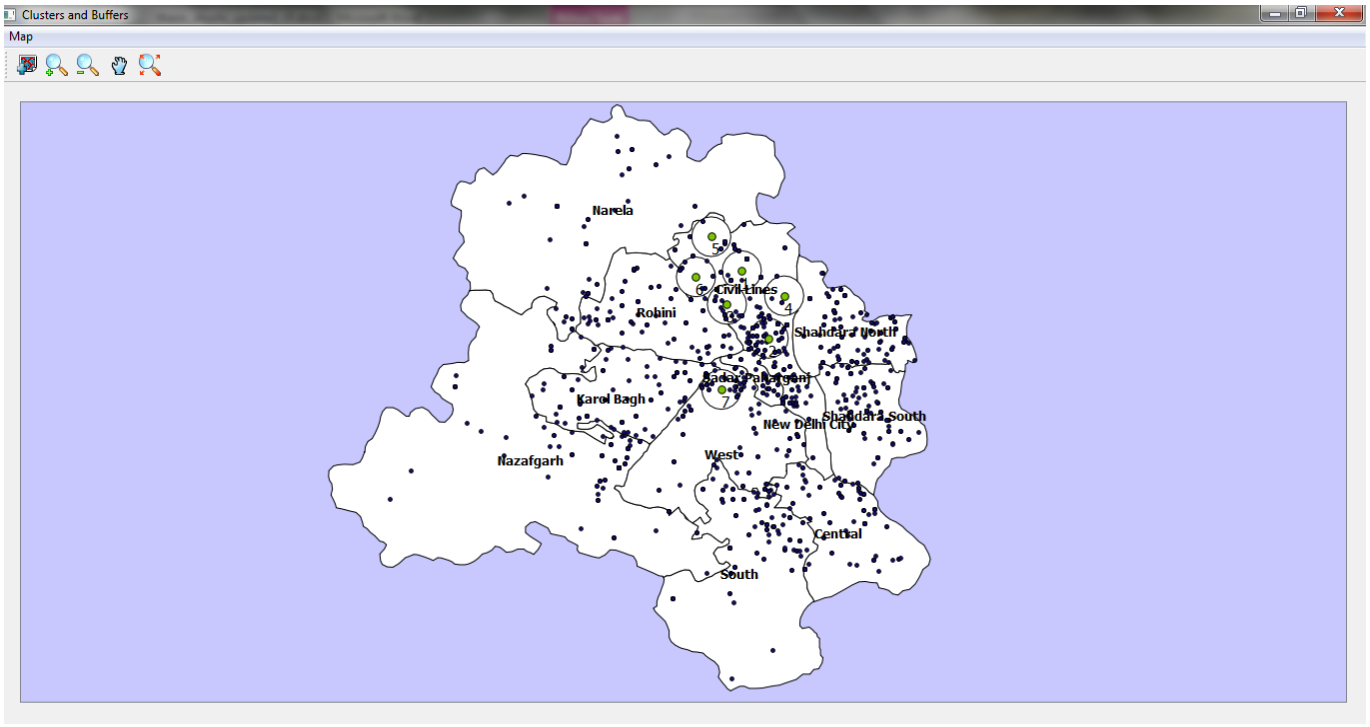


Figure 4.10: Buffer zones of 2 km radius around the clusters

4.6 Workflow of the implementation of developed package:

The brief steps which are followed by the developed package are as follows:

a) The program shuchi_main.py is executed as a main program and windowSub0.py, AppLatLonCase.py, windowSub4.py,windowSub2.py,windowSub3.py,mainwindow_ui.py and these programs are called in it. The executable file of SaTScan is called in the program.

b) mainwindow-ui.py is the program which generates the main menu.

c) windowsub0.py has a class SubWindow0 which creates canvas when init function is called and other functions are zoomin, zoomOut, pan, addLayer functions that is responsible to display shapefile of a map.

d) AppLatLonCase.py is a program which reads all the cases from the file and append latitude and longitude corresponding to each case.

e) windowsub4.py has a class subWindow4 which creates canvas with the same functions as described in step c) and generate maps with disease cases using ApplatLonCase.py. According to the study period all the disease cases maps are generated by this program.

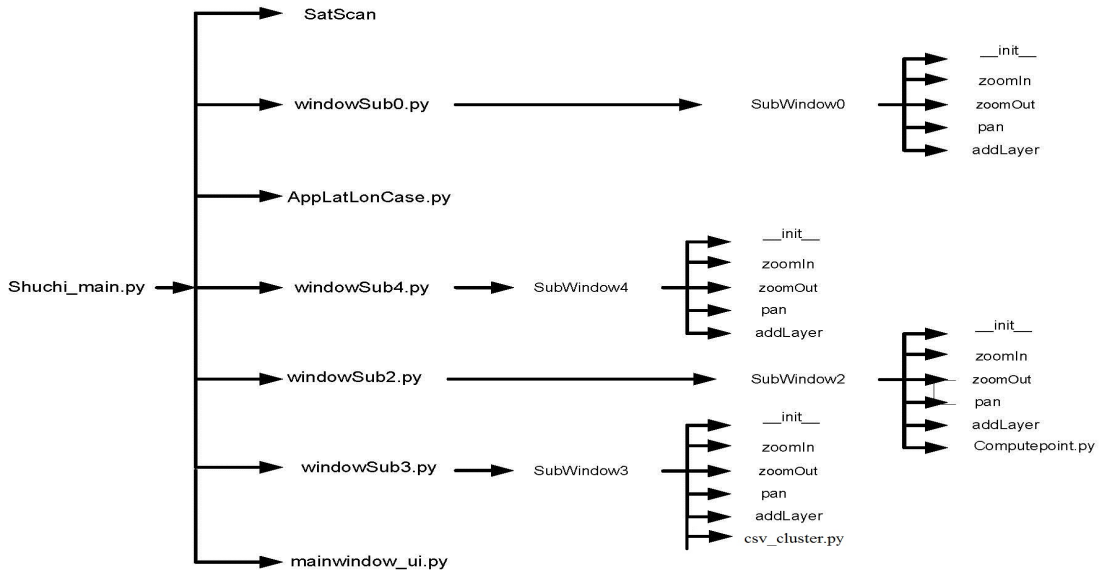


Figure 4.11: Workflow of implementation of the package

f) Windowsub2.py has a class subWindow2 which creates canvas with the same functions as described in step c) and generate maps with total number of cases reported per district by using Computepoint.py program. According to the study period all the disease cases maps are generated by this program.

g) Windowsub3.py has a class subWindow3 which creates canvas with the same functions as described in step c) and generate a map with statistically significant clusters by using csv_cluster.py. This map shows all the clusters with p-value less than equal to 0.5 along with the cases reported in all the years according to the entered study period.

4.7 A Case Study- analysis of past three years Dengue fever outbreaks in Delhi

In this section an analysis of dengue fever outbreaks in Delhi for past three years is presented. In India 28,000 dengue cases were reported in 2010 [1] and there is a significant rise in the incidence of dengue from 18,860 cases in the year 2011 to 49,606 cases in the year 2012 and 1,700 dengue cases were reported from Delhi due to the disease last year [2]. If disease surveillance was done nation-wide for early detection of disease outbreaks then effective actions would have been taken and outbreaks would have been controlled. To perform disease surveillance statistical methods

[1] http://www.ndtv.com/article/india/india-sees-highest-ever-35-000-dengue-cases-in-2012-297948

[2] http://articles.timesofindia.indiatimes.com/2013-02-27/delhi/37329822_1_denguecasesdenguefeverapex-referrallaboratories

to detect disease clusters are required. It is equally important to have effective visualization approach. The presented work covers both the aspects of disease surveillance along with the surveillance results.

4.7.1 Data Collection

To perform any type of analysis the most importance requirement is data. The analysis cannot be performed without collection of data. Data for analysis was collected from DHO Civil Lines Zone, Municipal Corporation of Delhi, Health Department, Dr. Ashok Rawat. The data covered details on dengue cases in Delhi for the years 2010, 2011 and 2012. Dr. Ashok Rawat was very interest in the results of analysis as it would assist his department in preventing outbreak of dengue cases. The details were taken from the Epidemiological Investigation form.

4.7.2 Data Pre-processing

The provided data is transformed in the format that is required to perform statistical analysis. In year 2010, dengue cases reported were 547, in year 2011, 66 cases were reported and in the year 2012, dengue cases reported were 1363 in Civil lines Zone, Municipal Corporation of Delhi. Between 2010 and 2012 total number of cases reported are 1976 in Civil Lines Zone. The data is transformed into a comma separated value file which contains information about the observed cases Location name, case count, date of reported case, age of the case and gender. Location name is represented as a string of characters, case count is represented as a numerical value, date is represented in a time format which is supported by the SaTScan , age is represented as a numerical value and gender is represented as a character. In this file age and gender are the covariates. Another comma separated value file is created for the coordinate information about the geographical location where case is occurred. In this file information about location name, latitude and longitude of the location of the case is stored. Location name is represented as a string of characters and latitude and longitude are represented as a decimal number of degrees.

4.7.3 Map Digitization

In [15], the process of map digitization in QGIS is described. Georeferencing process is used to assign real-world coordinates to each pixel of the raster. In the presented work, scanned map of Delhi is digitized by obtaining coordinates from the markings on the map image itself. Using these GCPs (Ground Control Points), the image is warped and it is made to fit within the chosen coordinate system.

4.7.4 Statistical Analysis

To detect clusters of disease outbreaks statistical analysis is done. SaTScan software is used to perform statistical analysis. Satscan is integrated in the developed standalone package. To begin analysis, in the GUI of the presented application click on “Run SaTScan” button. On clicking the button, SaTScan window opens. In the input tab, import the created comma separated value file which contains the details about observed cases under case file option. This file is saved as dengue_cases.cas . The extension supported by SaTScan for case file is .cas. Under coordinate file option import the other comma separated file which contains latitude and longitude information about each location of the observed case. This file is saved as dengue_cases.geo . The extension supported by SaTScan for coordinate file is .geo. The study period is from

2010/1/1 to 2012/12/31 with time precision to be year and coordinates lat/long. In the analysis space-time analysis is selected and space time permutation model is used because dengue fever has a relation with environmental variables as many cases are observed in months with warm and humid weather. Hence with geographical location time is also an important parameter to perform analysis. Adjustment for the maximum spatial cluster size is set at 2 kilometres. In The output tab result file dengue_cases.txt is saved and SaTScan software is executed. Some additional output files are created which are mentioned earlier. The work flow is described in the Figure 4.12

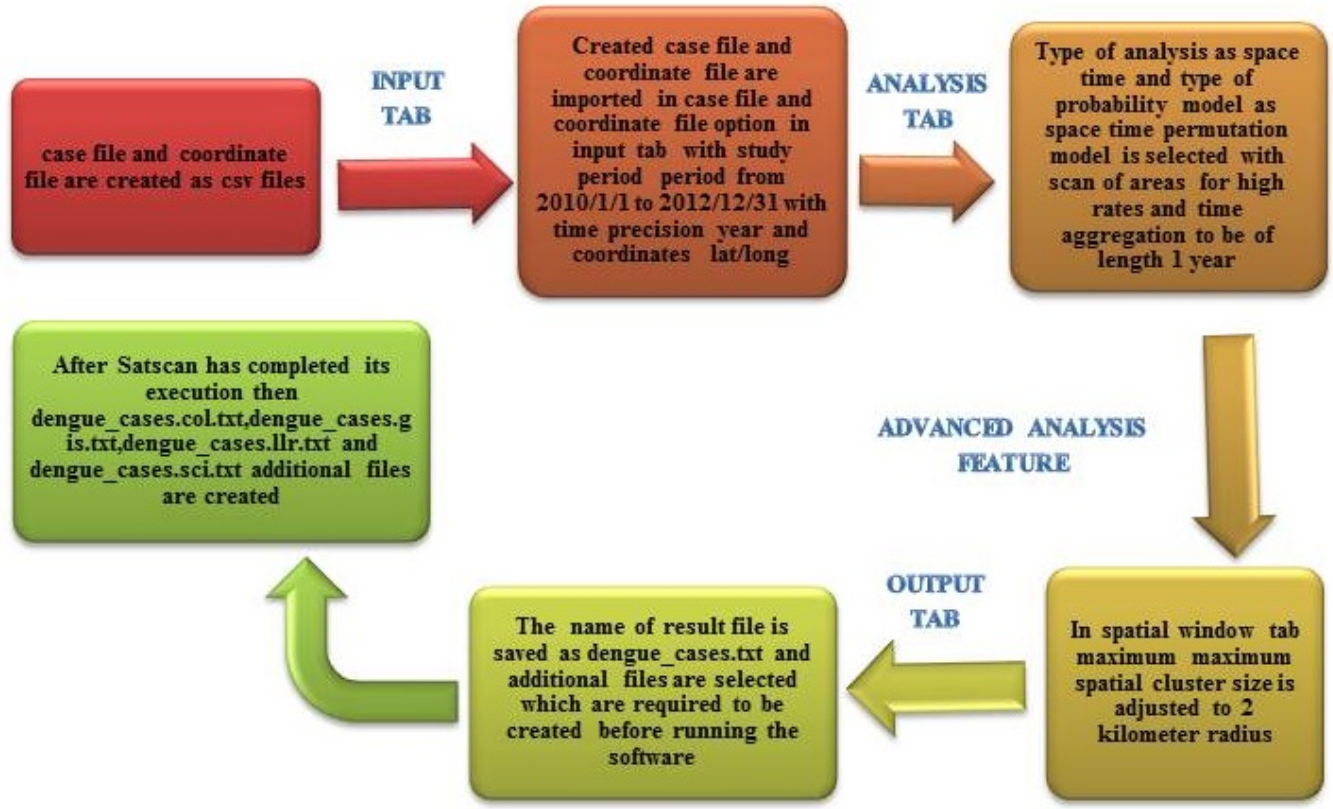


Figure 4.12: Workflow of SaTScan

4.7.5 Visualization results generated by VISA package

Delhi region is divided into 12 zones which are Narela, Rohini, Civil Lines, Shahdara North, Shahdara South, New Delhi City, Karol Bagh, Nazafgarh, central, west , Sadar Paharganj and south. The map of district boundary of Delhi is created by digitizing Delhi Map. On clicking Delhi Map button, the zone boundary map is generated as shown in Figure 4.2 . On Clicking Disease Cases Map 2010 button, dengue fever cases spread in year 2010 map are generated as shown in Figure 4.4 .High number of cases occurred in Sant Nagar(64), Jahangirpuri(47), Timarpuri(24) and Burari Village(22). On Clicking Disease Cases Map 2011 button, dengue fever cases reported year 2011 map are generated as shown in Figure 4.4 Majority of cases occurred in Jahangirpuri(12), Sant Nagar(5), Malkaganj(4) . On Clicking Disease Cases Map 2012 button, dengue fever spread in year 2012 map is generated as shown in Figure 4.5 High number of cases occurred in Rajnagar-II (34), Mandawali(26) and Sangam Vihar(23).There are 8 locations in Delhi where cases were reported in 2010, 2011 as well as 2012 as Table 4.1

Table 4.1: locations with number of cases in 2010-2012

Location	**2010(cases count)**	**2011(cases count)**	**2012 (cases count)**
Timarpur	24	2	5
Sant Nagar	62	5	3
Nehru Vihar	2	3	3
MajnuKaTilla	19	6	2
KingswayCamp	10	1	1
Kamla Nagar	7	1	1
Jahangirpuri	47	12	12
Azadpuri	16	3	1

Table 4.2: Details of detected cluster

Cluster Number	**Central Location ID**	**Latitude**	**Long - itude**	**Cluster Start Date**	**Cluster End Date**	**Log Likelihood Ratio**	**P-value of Cluster**	**Observed Cases**	**Expected cases**	**Observed /Expected**
1	SantNagar_Burari	28.74544	77.190212	01-01-2010	31-12-2010	37.490136	1.00E-17	119	48.85	2.45
2	DelhiUniversityi	2.87E+01	77.213	01-01-2010	31-12-2010	30.182756	1.11E-16	142	69.66	2.04
3	Majlisparki	2.87E+01	77.177633	01-01-2010	31-12-2010	24.49099	2.7E-13	100	45.94	2.18
4	WaziraabadVillage	2.87E+01	77.226145	01-01-2010	31-12-2010	8.08196	0.0022	22	8.13	2.71
5	KadiVillage	2.88E+01	77.164402	01-01-2010	31-12-2010	8.03457	0.0023	24	9.35	2.57
6	Badli	2.87E+01	77.151392	01-01-2011	31-12-2010	5.921857	0.042	6	0.97	6.2
7	EastPatelNag-ar	2.86E+01	77.172961	01-01-2012	31-12-2010	3.848719	0.548	65	45.35	1.43

On Clicking "No. of cases Map" button Choropleth map is generated to show density of dengue cases with graduated colors. On the basis of zone, most cases are reported from Civil Lines Zone, Shahdara North Zone and Shahdara South Zone as shown in Figure 4.6 On clicking Clusters button, clusters and buffer map are generated as shown in Figure 4.7 . There are seven detected clusters out of which one is most likely cluster and six are secondary clusters as shown in Table 4.2

4.7.6 Interpretation of detected clusters

Total number of locations where cases were reported are 617 and total number of cases reported are 1976. The shape file of the clusters generated and saved by the package is easily imported as a KML file in Google Earth so that interpretations can be made. A team of doctors of health department, Civil lines zone, Municipal Corporation of Delhi helped in confirming the interpretation of the detected clusters. The following are the detected clusters:

1. The locations included in the most likely cluster are SantNagar_Burari, SantBhagatColony, Burari, BurariVillage, Mukandpur, RadahVihar, BengaliColony, Bhalswa and KaushikEnclave with number of cases observed to be 119 and expected number of cases 48.16 with test statistic to be 38.097 approximately, highest among detected clusters. It is statistically significant because its P-value is smaller than 0.000000000000000010 which shows that its occurrence is not by chance. The cluster on Google Earth is shown in Figure 4.13

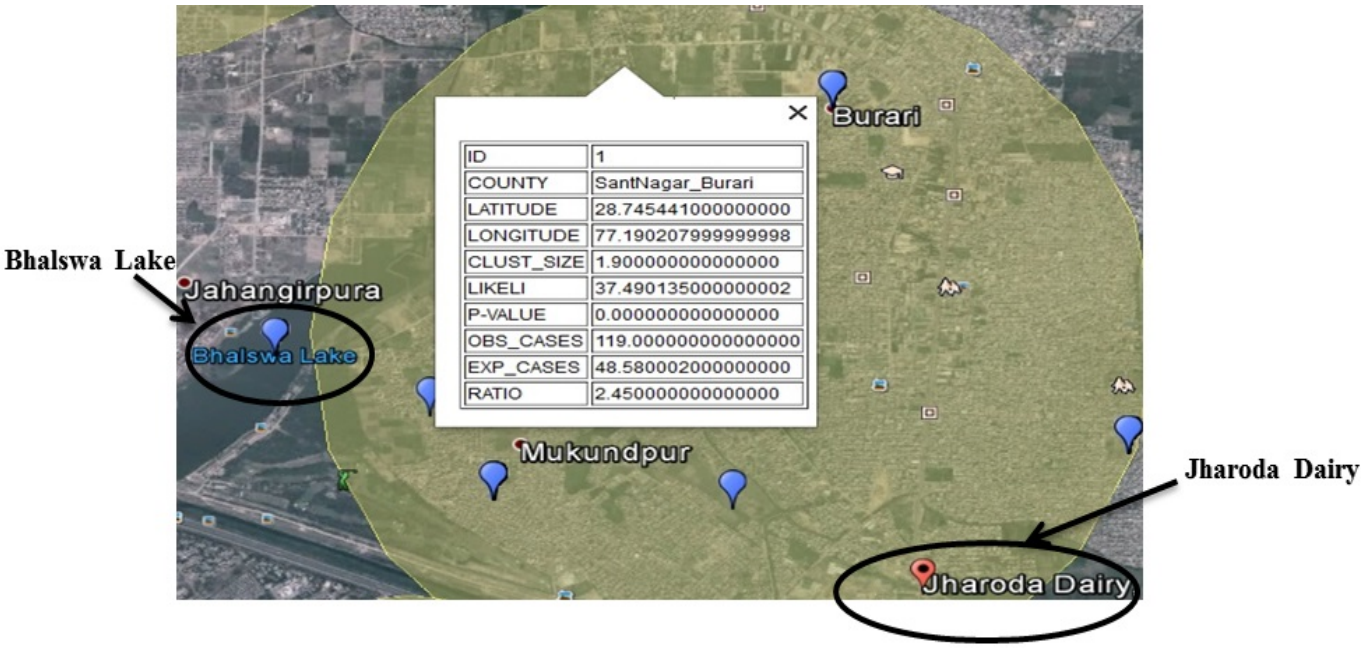


Figure 4.13: Screenshot of Mostly Cluster on Google Earth

The reasons for having high number of outbreaks in the regions related to this cluster are as follows:

- Burari regionhas Jharoda Dairy where cattle are raised. Mostly, dairies use earthen pots which are not cleaned frequently and results in stagnant water, helping in breeding of mosquitoes. [3]

[3]http://health.india.com/news/dengue-breeding-spots-more-than-households-report/

- Bhalswa region has Bhalswa lake and there is a high probability that the water present in the lake is stagnant.
- Massive construction sites with some unregulated construction sites.
- Demographic in this cluster is lower middle class and migrant population.
- During rainfall Yamuna bank is flooded and there is no place to drain water from these places.

2. The locations included in the second cluster are Delhi University, St.Stephen College, Patel Chest, Miranda House, Ramjas College, Mall Road, Old Vidhan Sabha, Vidha Sabha, Mourice Nagar, Vijaynagar, Khyber Pass, HansRaj College, Hakikat Nagar, Old Chandrawal, Kamla Nagar, Rajpur Road, Old Gupta Colony, New Chandrawal, UnderHill Road, IP College Campus, Timarpur, Outram Lane, Aruna Nagar, Civil Line,Outram Line, Kalyan Vihar, Malkaganj, Rajpur, Kingsway Camp, Ghanta Ghar, Gur Mandi, GTB Hospital Campus, RamKishore Marg, CCColony, Kabirbasti, Ishwar Colony, SinghSabhaRoad, IndraVihar, RanaPratapBagh, MukherjiNagar, WestMukherjiNagar, ShaktiNagar and MajnuKaTilla with number of cases observed as 142 and expected number of cases 68.82 with test statistic to be 31.074300 approximately, second highest among detected clusters. It is statistically significant because its P-value is 0.00000000000000011 which shows that its occurrence is not by chance. The cluster on Google Earth is shown in Figure 4.14

Figure 4.14: Screenshot of second detected cluster on Google Earth

The reasons for having high number of outbreaks in the regions related to this cluster are as follows

- Most of the areas are institutional and government office area. These areas are vulnerable due to lack of awareness.
- There is no sense of ownership because of multiplicity of authorities.
- There are large open areas near Rajpur Road where rain water gets collected
- Large student population and lack awareness about prevention and control of disease.

3. The locations included in the third cluster are Majlispark, AdarshNagar, NewsabziMandi, KewalPark, BunglowRoad, Azadpur, ModelTown-III, MCDColony, SaraiPipalThala, MahendraPark, LalBagh, ModelTown-II, MahendraEnclave, ModelTown, Jahangirpuriwith number of cases observed are 100 and expected number of cases 45.94 with test statistic to be 24.479 approximately, highest among detected clusters. It is statistically significant because its P-value is 0.00000000000047 which shows that its occurrence is not by chance. The cluster on Google Earth is shown in Figure4.15

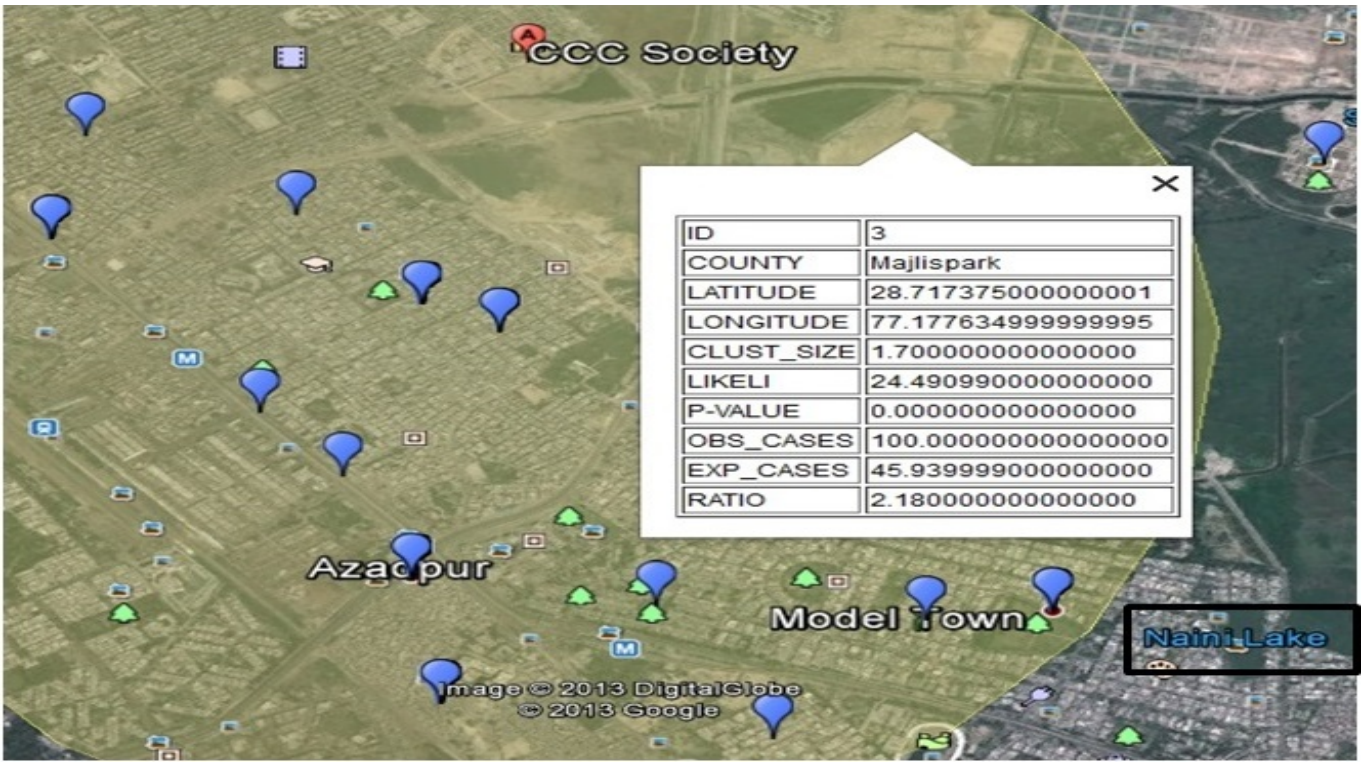


Figure 4.15: Screenshot of third detected cluster on Google Earth

The reason for outbreaks in regions related to this cluster is that the areas are largely residential localities with middle income group residents. In these areas houses are relatively smaller in size and hence care can be easily taken but storage of drinking water in unhygienic conditions can help in breeding of mosquitoes.

4. The locations included in the fourth cluster are WazirabadVillage, GopalPur,GandhiVihar, JharodaMajra with number of cases observed to be 22 and expected number of cases 7.92 with test statistic to be 8.448150 approximately. It is statistically significant because its P-value is .0014 which shows that its occurrence is not by chance. The cluster on Google Earth is shown in Figure 4.16

   The reasons for outbreak in regions related to this cluster are as follows:

   - The areas in the cluster are present in the flood plains of Yamuna.
   - It was main centre of the disease but a lot of steps were taken to control spread of disease.
   - Not densely population areas so control activities are quite difficult.

5. The locations included in the fifth cluster are KadiVillage, Nathupura, NathuColony, West-NathuColony, SwaroopVihar,Mukhmelpur, NangliPoonawith number of cases observed to

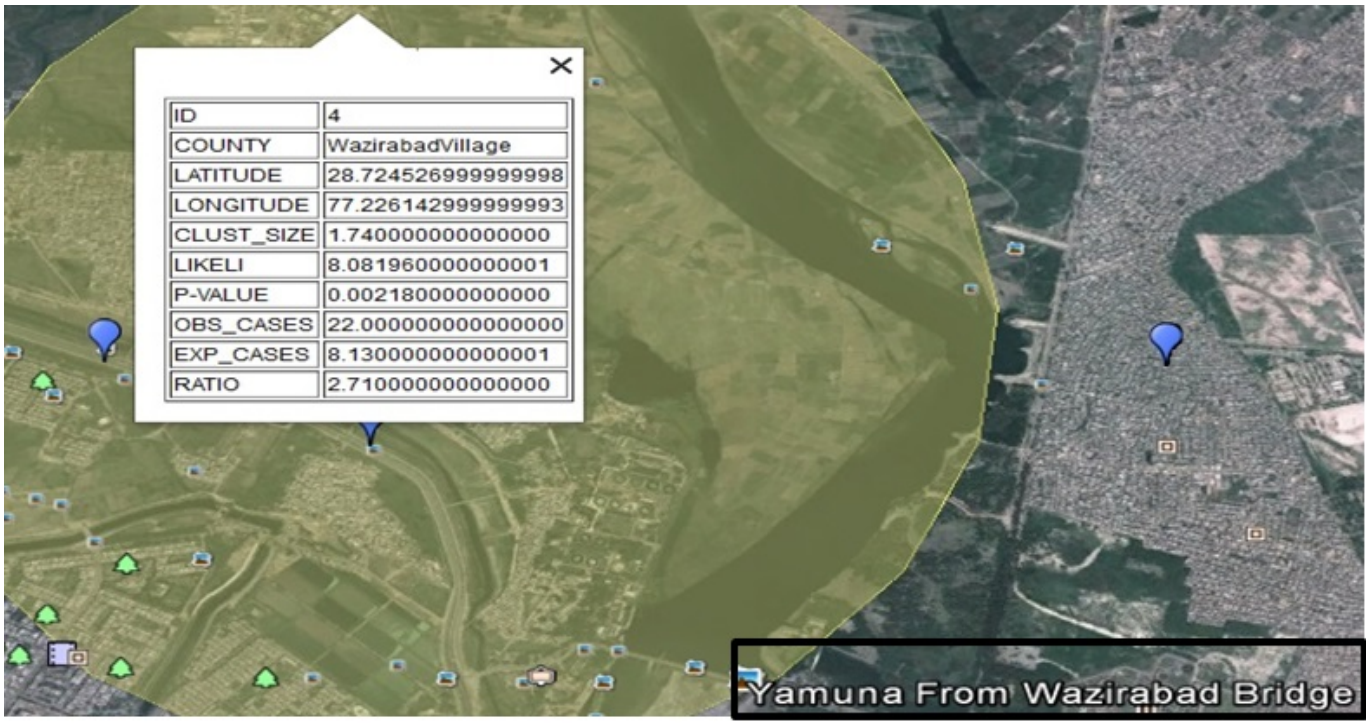


Figure 4.16: Screenshot of fourth detected cluster on Google Earth

be 24 and expected number of cases 9.35 with test statistic to be 8.033746 approximately. It is statistically significant because its P-value is 0.0025 which shows that its occurrence is not by chance. The cluster on Google Earth is shown in Figure 4.17

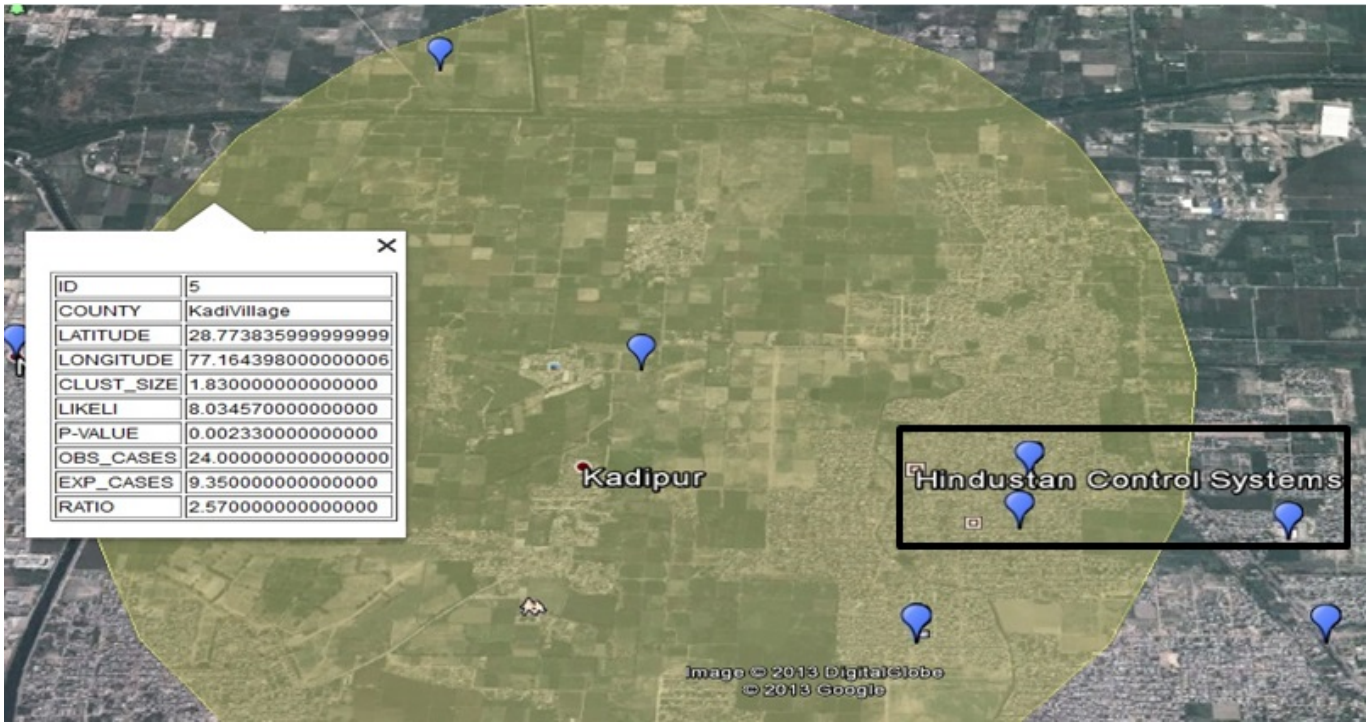


Figure 4.17: Screenshot of fifth detected cluster on Google Earth

The reasons for outbreak in regions related to this cluster are as follows:

- Rural areas and a lot of agriculture activities are done but less stagnant water.
- Industrial activities are growing here.
- Under process of urbanization.

6. The locations included in the sixth cluster are Badli, BadliVillage, SanjayColony, Jahangirpuri_IndustrialArea, GTKarnalRoad, BhalswaDairy, YadavNagarwith number of cases observed are 6 and expected number of cases 0.97 with test statistic to be 5.92 approx. It is statistically significant because its P-value is 0.043 which shows that its occurrence is not by chance. The cluster on Google Earth is shown in Figure 4.18

Figure 4.18: Screenshot of sixth detected cluster on Google Earth

The reasons for outbreak in region related to this cluster are as follows:

- Large number of industries including household industries. These industries are authorised industries.
- Hub of large labour population including migrant population
- Catered by private health facilities.

7. The locations included in the seventh cluster are EastPatelNagar, PrasadNagar, PatelNagar, RanjitNagar, UGFRanjitNagar, ChannaMarket, PunjabiBasti, AnandParbat, SatNagar, TankRoad, WestPatelNagar, ShankarRoad, NewRajenderNagar, OldRajinderNagar, DevNagar, WestExtensionArea_KarolBagh, RegarPura, NewRanjeetNagar, Shadipur, BaljitNagar, AryaSamajRoad, NehruNagar, KarolBagh with number of cases observed to be 65 and expected number of cases 45.35 with test statistic to be 3.848 approx. It is statistically significant because its P-value is 0.522 which shows that its occurrence is not by chance. The cluster on Google Earth is shown in Figure 4.19

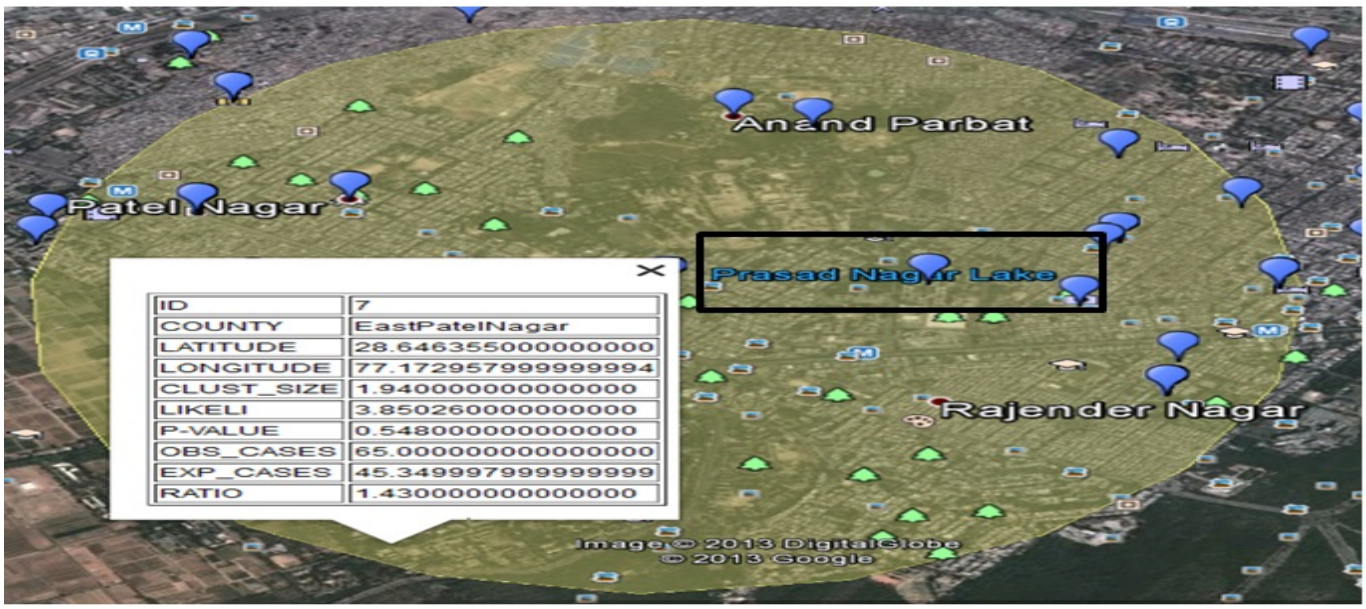
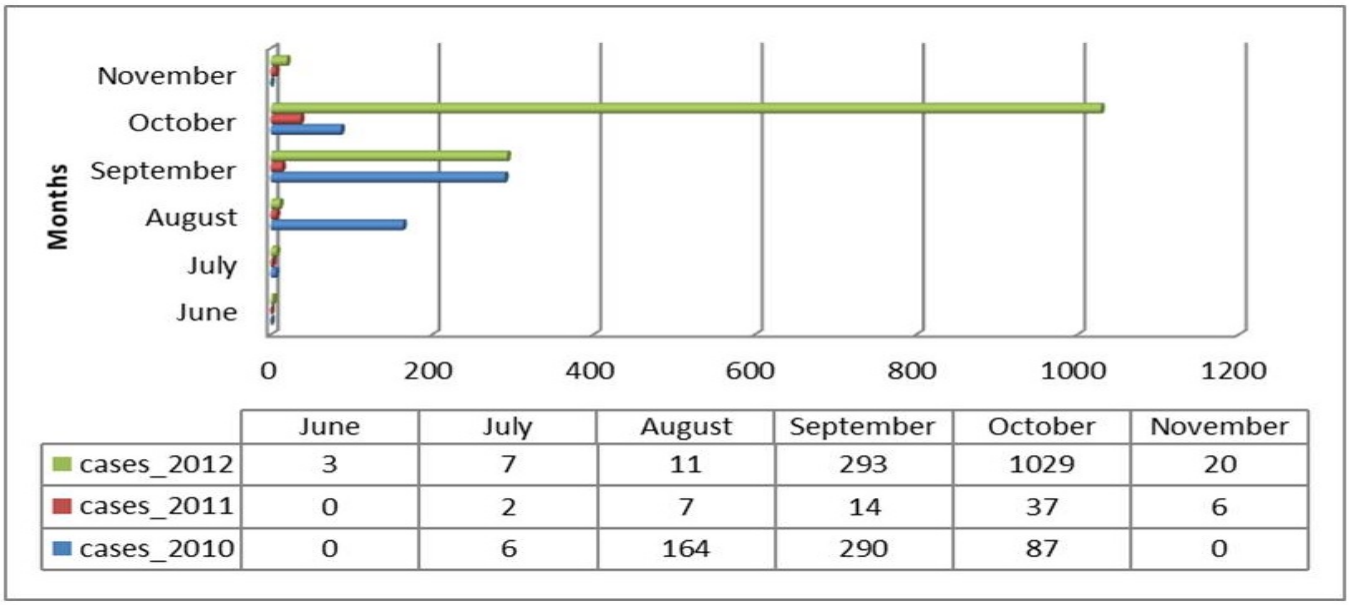


Figure 4.19: Screenshot of seventh cluster on Google Earth

The reasons for outbreak in regions related to this cluster are as follows:

- There are slum areas like Ranjit Nagar, Baljit Nagar as well as some commercialized areas and Prasad Nagar Lake.
- They were unauthorized but now these areas are regulated areas with proper facilities of drainage and water supply.

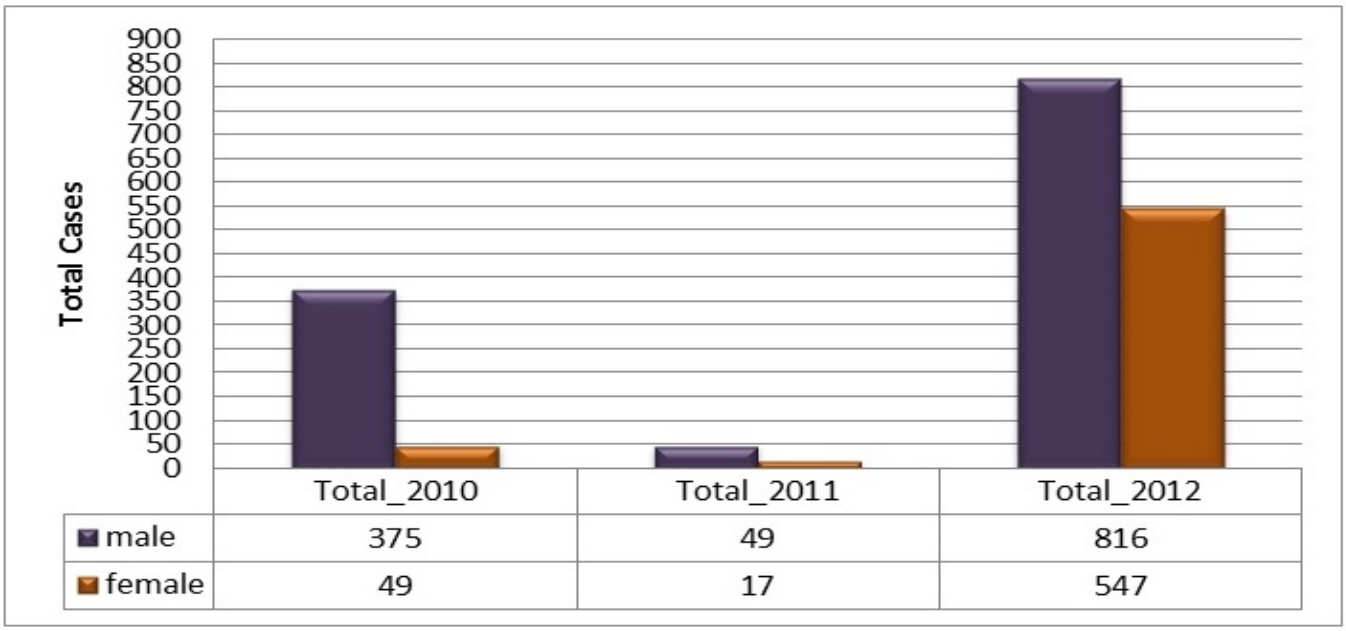
In months of July and August breeding of mosquitoes takes place and in months of September and October they become adults and thus large number of cases are observed in these two months as shown in Figure 4.20



	June	July	August	September	October	November
cases_2012	3	7	11	293	1029	20
cases_2011	0	2	7	14	37	6
cases_2010	0	6	164	290	87	0

Figure 4.20: Number of cases on the basis of months

On the basis of gender, infected male population was very high in 2010 and 2011 in comparison to female population but in 2012 there was a high increase in the number of infected females as shown in Figure 4.21



	Total_2010	Total_2011	Total_2012
male	375	49	816
female	49	17	547

Figure 4.21: Number of cases on the basis of gender

On the basis of age group in the year 2010, most cases occurred in 16-20 years of age, in the year 2011 most of the cases occurred in the 11-15 years of age and in the year 2012 most of cases occurred in the 11-15 years and 20-25 years of age groups as shown in Figure 4.22

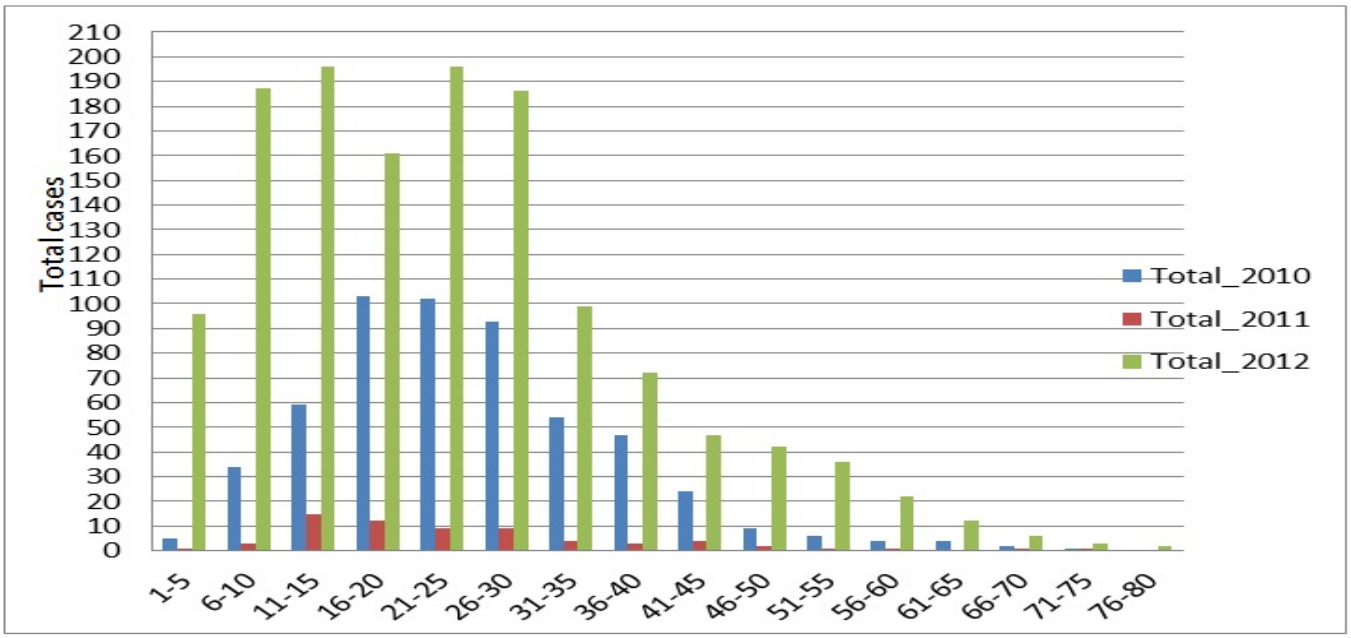


Figure 4.22: Number of cases on the basis of age groups

4.8 Conclusion

The Geographical visualization approach developed in this report facilitates the interpretation of space-time cluster detection methods by providing an efficient representation of the results of statistical analysis in geographical space. The standalone package is developed by using python, PyQGIS, QGIS, PyQT4, Qt Designer and SaTScan. To perform statistical analysis data collection is the main concern. Data on dengue fever from 2010-2012 is collected from Civil lines zone, Municipal corporation of Delhi. Proposed method is used to analyse the results of statistical analysis of dengue fever.

Space-time analysis is performed by using space-time permutation model. The disease clusters are detected with the cluster radius as 2 kilometres. With the proposed visualization method maps are generated.These maps show spreading of cases, density of cases within each district or county and statistically significant clusters. The team of doctors in Municipal Corporation of Delhi found these results very informative; efficient and accurate and the interpretations related to the clusters were very similar to their interpretations.

Chapter 5

Future Work

The future work of the presented work includes development of three more modules in the developed package. First proposed module will include a statistical interface for space-time analysis instead of integration with SaTScan. Second proposal is an addition of module to facilitate the use of web-services to make developed package available on the web. Third suggestion is have a module which will enable the user to save analysed data in a database. This section will be very helpful for those health officials who are doing daily surveillance of disease outbreaks. With the facility of a database user can save data whenever required and can perform analysis at chosen time intervals such as on weekly basis or based on months or years.

Bibliography

[1] Renaud Piarroux , Robert Barrais, Benot Faucher, Rachel Haus, Martine Piarroux, Jean Gaudart, Roc Magloire, and Didier Raoult, *UNDERSTANDING THE CHOLERA EPIDEMIC, HAITI.* Emerging Infectious Diseases, 2011, Vol.17, No.7, pp. 1161-1167.

[2] Martin Kuldorff, *A SPATIAL SCAN STATISTIC* Commun Statist Theory Meth, 1997, Vol.26, No.6, pp. 1481-1495.

[3] Martin Kulldorff, Richard Heffernan, Jessica Hartman, Renato Assunc, Farzad Mostashari, *A SPACETIME PERMUTATION SCAN STATISTIC FOR DISEASE OUTBREAK DETECTION*, plosjournals, 2005, Vol.2, No.3, pp. 216-224.

[4] Martin Kulldorff, *PROSPECTIVE TIME PERIODIC GEOGRAPHICAL DISEASE SURVEILLANCE USING A SCAN STATISTIC* Royal Statistical Society, 2001, Vol. 164, Part 1, pp. 61-72.

[5] Martin Kulldorf, William F. Athas, Eric J.Feuer, Barry A. Miller and Charles R.Key, *EVALUATING CLUSTER ALARMS: A SPACE-TIME SCAN STATISTIC AND BRAIN CANCER IN LOS ALAMOS, NEW MEXICO* American Journal of Public Health, 1998, Vol.88, No.9, pp. 1377-1380.

[6] Martin Kulldorff, Eric J. Feuer, Barry A. Miller and Laurence S. Freedman, *BREAST CANCER CLUSTERS IN THE NORTHEAST UNITED STATES: A GEOGRAPHIC ANALYSIS*, American Journal of Epidemiology, 1997, Vol. 146, No. 2, pp. 161-170.

[7] Marta Jankowska, Jared Aldstadt, Arthur Getis, John Weeks, Grant Fraley, *AN AMOEBA PROCEDURE FOR VISUALIZING CLUSTERS*, GIScience 2008, Park City, UT

[8] Toshiro Tango and Kunihiko Takahashi,*A FLEXIBLY SHAPED SPATIAL SCAN STATISTIC FOR DETECTING CLUSTERS* International Journal of Health Geographics, 2005, Vol.4, No.11, pp 1-15.

[9] Jin Chen, Robert E Roth, Adam T Naito, Eugene J Lengerich and Alan M MacEachren, *GEOVISUAL ANALYTICS TO ENHANCE SPATIAL SCAN STATISTIC INTERPRETATION:AN ANALYSIS OF U.S. CERVICAL CANCER MORTALITY* International Journal of Health Geographics, 2008, Vol.7, No.57, pp. 1-18.

[10] Agricola Odoi, S Wayne Martin, Pascal Michel, Dean Middleton, John Holt and Jeff Wilson, *INVESTIGATION OF CLUSTERS OF GIARDIASIS USING GIS AND A SPATIAL SCAN STATISTIC* International Journal of Health Geographics, 2004, Vol.3, No.11, pp 1-11.

[11] Allyson M Abrams and Ken P Kleinman, *A SATSCAN MACRO ACCESSORY FOR CARTOGRAPHY (SMAC) PACKAGE IMPLEMENTED WITH SAS SOFTWARE*, International Journal of Health Geographics, 2007, Vol. 6, No.6, pp. 1-8.

[12] http://www.satscan.org/cgi/bin/satscan/register.pl/SaTScan_Users_Guide.pdf?todo=process_userguide_dow

[13] http://www.riverbankcomputing.com/software/pyqt

[14] http://www.qgis.org/pyqgis-cookbook/

[15] http://qgis.spatialthoughts.com/2012/02/tutorial-georeferencing-topo-sheets.html

[16] Martin Kulldorff, Lan Huang , Linda Pickle and Luiz Duczmal *AN ELLIPTIC SPATIAL SCAN STATISTIC*, Wiley InterScience,2006, Vol. 25, pp. 39293943.

[17] Sheehan TJ, DeChello LM. *A space-time analysis of the proportion of late stage breast cancer in Massachusetts, 1988 to 1997*, International Journal of Health Geographics, 2005,Vol. 4, No.15

Chapter 6

Appendix A

The steps involved in the development of GUI are:

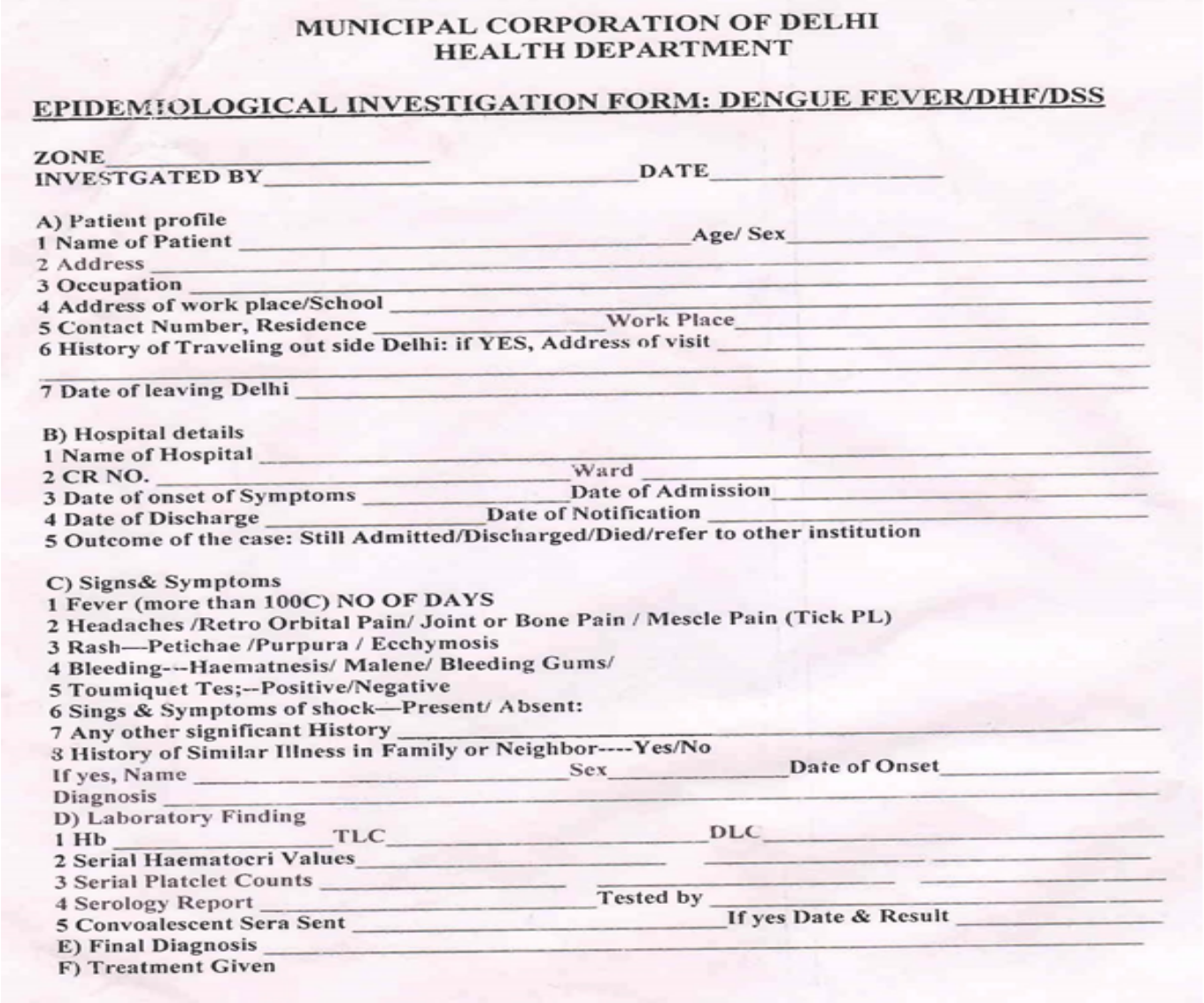
1. Create a directory for developing the application and storing any changes to it.
2. Run Qt Designer (by default C:\OSGeo4W\bin\designer.exe)
3. A "New Form" dialogue will appear. If it does not, choose from the menu "New Form".
4. From the list "templates / forms" choose "Main Window" and click on "Create".
5. Drag widget "Frame" located in the group "Containers" into the window. Click the mouse anywhere in the form but outside the container that we just added. Press the button "Lay Out in a Grid"
6. Save the created form as mainwindow.ui in the folder created in step 1.
7. Select the inscription Type Here and enter the name of the item "Map" and press enter. Repeat this step for other items such as ZoomIn, ZoomOut, Pan and zoomfull toolboxes.
8. Create a new prefix by clicking "Add Prefix". Now add here five icons for the tools.
9. When creating a menu for Qt Designer automatically item is given a name through "Action Editor" (Editor commands). For each menu item create command. These command can be named as mpActionZoomIn, mpActionZoomOut, mpActionPan, mpActionAddLayer and mpActionZoomFullExtent.
10. Assign each action to the related icon by double clicking on the the action and choosing the icon.
11. Save the changes
12. Exit from Qt Designer.
13. Now compile the form using the PyQt interface compiler:
    Pyuic4 - o mainwindow_ui.py mainwindow.ui

    This generate Python code for the mainwindow GUI.

Resource file resources.qrc can be compiled by:
Pyrcc4 -0 resource_rc.py resources.qrc

Chapter 7

Appendix B



MUNICIPAL CORPORATION OF DELHI
HEALTH DEPARTMENT

EPIDEMIOLOGICAL INVESTIGATION FORM: DENGUE FEVER/DHF/DSS

ZONE_____
INVESTGATED BY_____DATE_____

A) Patient profile
1 Name of Patient _____Age/ Sex_____
2 Address _____
3 Occupation _____
4 Address of work place/School _____
5 Contact Number, Residence _____Work Place_____
6 History of Traveling out side Delhi: if YES, Address of visit _____
_____
7 Date of leaving Delhi _____

B) Hospital details
1 Name of Hospital _____
2 CR NO. _____Ward _____
3 Date of onset of Symptoms _____Date of Admission_____
4 Date of Discharge _____Date of Notification _____
5 Outcome of the case: Still Admitted/Discharged/Died/refer to other institution

C) Signs& Symptoms
1 Fever (more than 100C) NO OF DAYS
2 Headaches /Retro Orbital Pain/ Joint or Bone Pain / Mescle Pain (Tick PL)
3 Rash—Petichae /Purpura / Ecchymosis
4 Bleeding---Haematnesis/ Malene/ Bleeding Gums/
5 Toumiquet Tes;--Positive/Negative
6 Sings & Symptoms of shock—Present/ Absent:
7 Any other significant History _____
8 History of Similar Illness in Family or Neighbor----Yes/No
If yes, Name _____Sex_____Date of Onset_____
Diagnosis _____
D) Laboratory Finding
1 Hb _____TLC_____DLC_____
2 Serial Haematocri Values_____ _____ _____
3 Serial Platclet Counts _____ _____ _____
4 Serology Report _____Tested by _____
5 Convoalescent Sera Sent _____If yes Date & Result _____
E) Final Diagnosis _____
F) Treatment Given

1 PATIENT'S LOCALITY **Type of locality: Authorized /Un authorized/Slums**

. Total house checked _____ Breeding Detected _____
. Cooler Checked _____ Cooler _____
. OT Checked _____ OT _____
. Containers Checked _____ Containers _____
. If any Other _____ If any Other _____

2 WORK PLACE

. Total house/ Office/ Shops checked _____ Breeding Detected _____
. Cooler Checked _____ Cooler _____
. OT Checked _____ OT _____
. Containers Checked _____ Containers _____
. If any Other _____ If any Other _____

3 House Sprayed _____ Fogging Done _____
4 No of Prosecutions _____ No of Legal Notice Issued _____
5 Results of Ovitrap if placed in / Around Patient's House _____
6 Any breeding detected on Routine DBC in the area of Patient's residence / Work place before notification of the case (previous Weeks) _____
7 Identification of Source of infection _____
8 If concerned DHO (pertaining of Work place) informed Yes /No
9 If CMO / DHO of he neighboring State/ District informed Yes/ No

H) IEC Activities

1 Meeting with local RWAs conducted:- Yes/ No
2 If Yes Name of RWA _____ Date _____
3 Hand bills distributed _____ Poster's _____ Sticcers _____ others _____
4 Area covered by Munadi _____

Dy. Health Officer Zonal Epidemiologist AMO/SMI

Dy.MHO (Mal&OVBD)

Copy to

MHO

Addl. MHO