


# **Clustering and Time Series Prediction for Spatio-Temporal Geographic Dataset**

**Nirma University**  
**Institute of Technology**  
**Certificate**

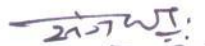
This is to certify that the thesis entitled **Clustering and Time Series Prediction for Spatio-Temporal Geographic Dataset** has been prepared by **Kedar Prasad Agrawal (11EXTPHDE71)** under my supervision and guidance. The thesis is his own original work completed after careful research and investigation. The work of the thesis is of the standard expected of a candidate for Ph.D. Programme in **Computer Science and Engineering** and I recommend that it be sent for evaluation.

Date: Jan 28, 2016


  
(Dr. Sanjay Garg)  
Signature of the Guide

---

Forwarded Through:

  
\_\_\_\_\_  
(i) Name and signature of the  
Head of the Department (if any)

  
\_\_\_\_\_  
(ii) Name and signature of the Dean Faculty of Technology and Engineering

  
30/1/2016 (Dr. M. GHATE)  
\_\_\_\_\_  
(iii) Name and signature of the Dean Faculty of Doctoral Studies and Research

To:  
Executive Registrar  
Nirma University

## **Abstract**

Owing to the generation of petabytes of data (may be of type classical, spatial, temporal or hybrid) on daily basis from different sources, work is required to be carried out such that these voluminous amount of data can be utilized meaningfully using relevant data mining tasks. When it is required to deal with Spatio-Temporal dataset, data mining related tasks becomes more challenging specially in case of obtaining arbitrary shaped clusters of good quality and reliable forecasting. Based on reliable forecasting, some anticipatory action like Land Usage, availability of good and healthy crops or no crops, good rains, flood or detecting drought areas etc. can be taken which is beneficial to masses. In clustering, issues like detection of arbitrary shaped clusters, handling high dimensional data, independence from order of data input, interpretability, ability to deal with nested clusters, scalability etc. and while forecasting, issues like handling non-stationarity of time series, non-linear domain, selection and tuning of parameters of existing or newly developed technique(s) needs to be addressed with utmost care.

Spatio-Temporal Data Mining (STDM) is a process of the extraction of implicit knowledge, spatial and temporal relationships, or other patterns not explicitly stored in spatio-temporal databases. As data is growing not only from static view point, but they also evolve spatially and temporally which is dynamic in nature that is the reason why this field is now becoming very important field of research. In addition Spatio-Temporal (ST) -Data tends to be highly auto-correlated, because of which assumptions which are taken in Gaussian distribution models fails, as in Gaussian Distribution, an assumption of independence is taken into consideration, which is not the case with ST Data. Vital issues in spatio temporal clustering technique for Earth observation data is to obtain good quality arbitrarily shaped clusters and its validation. The presented research work addresses these issues and presents their solutions. In order to achieve said objective, an attempt has been made to develop a clustering algorithm named as “Spatio-Temporal - Ordering Points to Identify Clustering Structure (ST-OPTICS)” which is modified version of existing density based technique OPTICS.

Experimental work carried out is analyzed and found that quality of clusters obtained and run time efficiency are much better than existing technique i.e. ST DBSCAN. An attempt has been made to

hybridize the results generated by ST-OPTICS with agglomerative approach to improve the visualization and the interpretation of obtained clusters. Validations of the obtained results have also been performed by visualization and various performance indices. Results shows performance improvement of ST-OPTICS clustering technique.

In order to improve the accuracy of prediction, fusion of statistical and machine learning models have been done. Statistical model like Integration of Auto Regressive (AR) and Moving Average (MA) is capable to handle non-stationary time series but it can deal with only single time series. While machine learning approach (i.e. Support Vector Regression (SVR)) can handle dependency among different time series along with non-linear separable domains, however it cannot incorporate the past behavior of time-series. This led to combine these two approaches for improving accuracy of time series prediction, where focus has been given on minimization of forecast error using residuals, which helps to take appropriate action for near future. Keeping in view objective, hybridization of Auto Regressive Integrated Moving Average (ARIMA) with SVR models has been done. In order to reduce number of area wise models and reduction in time complexity for tuning different parameters, emphasis has been laid down on handling issues related to scalability by taking suitable representative samples from each sub-areas. Results obtained shows that the performance of proposed hybrid model is better than individual models.

**Nirma University  
Institute of Technology**

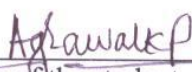
**Declaration**


I, **Kedar Prasad Agrawal**, registered as Research Scholar, bearing Registration No. **11EXTPHDE71** for Doctoral Programme under the Faculty of **Technology & Engineering** of Nirma University do hereby declare that I have completed the course work, pre-synopsis seminar and my research work as prescribed under R. Ph.D. 3.5.

I do hereby declare that the thesis submitted is original and is the outcome of the independent investigations / research carried out by me and contains no plagiarism. The research is leading to the discovery of new facts / techniques / correlation of scientific facts already known. (Please tick whichever is applicable). This work has not been submitted to any other University or Body in quest of a degree, diploma or any other kind of academic award.

I do hereby further declare that the text, diagrams or any other material taken from other sources (including but not limited to books, journals and web) have been acknowledged, referred and cited to the best of my knowledge and understanding.

Date: Jan 28, 2016

  
\_\_\_\_\_  
Signature of the student

  
(Guide)  
(Dr. Sanjay Gang)

## Acknowledgement

It gives me immense pleasure to present this research work entitled “**Clustering and Time Series Prediction for Spatio-Temporal Geographic Dataset**”, to the **Almighty** for being present in all my endeavors.

I would like to sincerely thank **Dr. Sanjay Garg**, my Supervisor and Head of Department, Computer Science and Engineering, Institute of Technology, Nirma University, Ahmedabad for his continuous motivation, guidance and support throughout my Doctoral studies.

My sincere thanks are also due, to the reviewers **Dr. Asim Banerjee**, Professor, DAIICT and **Dr. Tanish Zaveri**, Sr. Associate Professor, Nirma University who had provided their valuable feedback for improvement during my Research Progress Committee meetings. I take this opportunity to thank **Dr. P.N. Tekwani**, Director, Institute of Technology, Nirma University for providing infrastructural facilities on campus.

I render my sincere thanks to **Mr. Shashikant Sharma**, Senior Scientist, SAC, ISRO, Ahmedabad for providing data and support time to time.

I render my thanks to **Mr. Pinkal Patel**, Ph.D. scholar for providing his support in writing publication related work time to time.

I shall be failing in my duties if I do not render my sincere thanks to my wife **Mrs. SUSHMA** and my loving daughter **Ms. ANMOL**, for their lots of sacrifices, continuous support and patience towards this long journey of research work.

I am indebted to my loving **Parents** and **Almighty** without their support and blessings this journey would have not been completed.

I am also thankful to all faculty members, staff members and students of our department for providing all types of resources and support on time directly or indirectly.

At last but not least I sincerely thank to one and all who have supported me directly or indirectly to achieve this bigger objective.

**Kedar Prasad Agrawal**

**(11EXTPHDE71)**

# Contents

Certificate	i
Declaration	ii
Publication Related To Thesis	iii
Abstract	iv
Acknowledgement	vi
List of Tables	ix
List of Figures	x
Abbreviation	xii
Chapter 1 : Introduction	1
1.1 Motivation	2
1.2 Objective of the work	2
1.3 Scope of the work	2
1.4 Major Contributions	3
1.5 Organization of the Thesis	4
Chapter 2 : Preliminaries	5
2.1 Basic Terminologies	5
2.2 Spatio-Temporal Clustering Technique	12
2.3 Time Series Prediction Technique	17
Chapter 3 : Literature Survey	19
3.1 Spatio-Temporal Clustering Technique	19
3.1.1 Related work	19
3.1.2 Issues and Challenges	27
3.2: Time Series Prediction	28
3.2.1 Related work	28
3.2.2 Issues and Challenges	30
Chapter 4 : Proposed Spatio-Temporal Clustering Technique	33
4.1 : ST-OPTICS: Clustering Technique	34
4.2: Experimental Analysis	45
4.2.1 Experimental Setup	45
4.2.2 Data Specification	46
4.2.3 Performance Parameters and Validation Indices	46
4.2.4 Results	46
4.3 Discussions	58

Chapter 5 : Proposed Time Series Prediction Technique	61
5.1 SARIMA – Statistical Prediction Technique	62
5.2 RBF – SVR : Machine Learning based Prediction Technique	63
5.3 Hybrid Model for Prediction	65
5.4 Experimental Analysis	68
5.4.1 Experimental Setup	68
5.4.2 Data Specification	69
5.4.3 Evaluation of Performance Parameters	69
5.4.4 Results	70
5.5 Scalability Enhancements	75
5.6 Discussions	78
Chapter 6 : Summary, Conclusions and Future Scope	79
References	81



## List of Tables

Table 2.1	A sample data set of Events Data.....	8
Table 3.1	Categories of Clustering Algorithms – A Comparison .....	21
Table 3.2	Density and Grid Based Clustering Algorithms – A Comparison.....	24
Table 3.3	Shortcomings of AR, MA, ARMA, ARIMA, SARIMA models.....	30
Table 3.4	Shortcoming of RBF-SVR model.....	30
Table 4.1	Selected Validation Indices with their Significance and Applicabilities.....	39
Table 4.2	Unselected Validation Indices with their Significance and Reasons .....	40
Table 4.3	Cluster validation indices using k-means and DBSCAN on classical datasets..	44
Table 4.4	Dataset Specification.....	46
Table 4.5	Comparison - ST-OPTICS and ST-DBSCAN for Gujarat state.....	46
Table 4.6	Interpretation – Results obtained.....	49
Table 4.7	Comparison table showing results using ST-OPTICS and ST-DBSCAN with hybrid approach.....	50
Table 4.8	Results - Validation Indices using a) ST-OPTICS and b) ST-DBSCAN ..... (Gujarat State)	55
Table 4.9	Cluster validation indices’s results of ST-OPTICS and ST-DBSCAN for ..... the Himachal Pradesh and Arunachal Pradesh	55
Table 4.10	Cluster validation indices’s results of ST-OPTICS and ST-DBSCAN for ..... the Chhattisgarh and Karnataka	56
Table 5.1	Dataset Specification.....	67
Table 5.2	Performance comparison for Kharif season.....	72
Table 5.3	Performance comparison for Rabi season.....	73
Table 5.4	Performance comparison on 7 districts of Gujarat for Kharif season.....	74
Table 5.5	Performance of hybrid approach on 7 districts of Gujarat for Rabi season.....	75

## List of Figures

Figure 2.1	Spatio – Temporal Data Types .....	7
Figure 2.2	Basic Terms1: (a) Border, Core and Noise objects (b) Density- Reachability (c) Density-Connected .....	12
Figure 2.3	Basic Terms2: Core-distance and reachability-distances.....	14
Figure 4.1	Proposed Approach (ST-OPTICS). .....	34
Figure 4.2	Classical Datasets: (a) Jain Dataset (b) Spiral Dataset and (c) Compound .....	42
Figure 4.3	Results of (a) k-means (b) DBSCAN on Jain dataset.....	43
Figure 4.4	Results of (a) k-means (b) DBSCAN on Spiral dataset.....	43
Figure 4.5	Results of (a) k-means (b) DBSCAN on Compound dataset.....	43
Figure 4.6	Clustered Maps (a) Forest survey map (b) Clusters given by ST-DBSCAN (c) Clusters given by ST-OPTICS.....	47
Figure 4.7	Hybrid Approach.....	47
Figure 4.8	Comparison of Result for validation-Hybrid approach (a) Forest survey map (b) Result of ST-DBSCAN (c) Result of ST-OPTICS.....	49
Figure 4.9	Clustered maps of Arunachal Pradesh, Chhattisgarh, Karnataka and Himachal Pradesh for forest survey map shown in (a),(d),(g) and (j), results of ST- DBSCAN shown in (b), (e),(h) and (k) and results of ST-OPTICS shown in (c),(f),(i) and(l) .....	54
Figure 4.10	Performance Comparison (Average Runtime) of ST-OPTICS and ST-BSCAN	56
Figure 4.11	Performance Comparison of ST-OPTICS and ST-DBSCAN with different dataset size.	59
Figure 5.1	Flow chart of hybrid approach.....	63
Figure 5.2	Time Series of NDVI from 2002 to 2010.....	67
Figure 5.3	ACF and PACF plots for NDVI time series.....	68
Figure 5.4	ACF and PACF plots for NDVI time series with seasonal difference(D=1).....	68
Figure 5.5	Actual NDVI Time Series (2002-2010) and Prediction only for 2010 using ARIMA model.....	69
Figure 5.6	Actual and Predicted NDVI Time Series for the year 2010 using ARIMA model.....	69
Figure 5.7	Pearson's Correlation Coefficient between NDVI and rainfall.....	70

Figure 5.8	Actual and predicted NDVI Time Series for the year 2010 using RBF-SVR model.....	71
Figure 5.9	Prediction of NDVI for Kharif season by SARIMA, RBF-SVR and Hybrid approach for the year 2010.....	72
Figure 5.10	Prediction of NDVI for Rabi season by SARIMA, RBF-SVR and Hybrid approach for the year 2010.....	72

# Chapter 1

## Introduction

Data Mining (alternatively known as Knowledge Discovery for Databases-KDD) is a major banner under which attempt have been made to perform various tasks such as clustering, classification, prediction, summarization, aggregation, outlier detection etc. Basically the whole crux of the data mining is to extract hidden pattern present in available dataset(s), its evaluation and interpretation of the extracted pattern(s) and using it the way we want (depending upon the domain of interest under consideration). Petabytes of data are being generated on daily basis from different sources viz satellites, mobile communication, super bazar, traffic, meteorological department etc. some useful work is required to be carried out such that these voluminous data can be utilized meaningfully i.e. one should be able to extract hidden knowledge from it. Data mining tasks are performed on domains relevant data taken from Data Marts or Data Warehouse where cleaned, transformed data are kept in an integrated way. Lot of research have already been done on classification and clustering in different areas for utilizing these enormously generated data effectively. Clustering is more challenging task compare to classification due to absence of class labels in dataset and in real life there are many instances where class labels are not present especially when obtained data are from satellites

Classical datasets are often discrete. Classical techniques focus on global pattern. Generally, classical dataset obeys Gaussian distribution where the assumption of independence is true, while in case of Spatio-Temporal Data Mining, data samples are highly self-correlated i.e. they are auto-correlated, they are embedded in continuous space instead of discrete space and at the same time focus is on the local pattern instead of global.

All geographic phenomena evolve over time both spatially and temporally, so it is felt that if some tasks related to clustering and prediction on spatial, temporal and spatio-temporal data are done, it may be useful in many ways for example Land Utilization, Urban and Rural planning by Governmental agencies, Identification of earth quake prone areas etc.

## **1.1 Motivation**

Owing to the generation of petabytes of data (may be classical, spatial, temporal or hybrid type) on daily basis from different sources, work is required to be done such that these voluminous data can be utilized meaningfully i.e. one should be able to extract hidden knowledge from it. Lot of research has already been done on classification and clustering in different areas for utilizing generated data effectively. Clustering is more challenging task compare to classification due to absence of class labels in data set and in real life there are many instances where class labels are absent especially when obtained data are from satellites. All geographic phenomena evolve over time both spatially and temporally. It has been felt that if some tasks related to clustering on spatial, temporal and spatio-temporal data and prediction about the future based on historical data is carried out, it may be useful to the masses in many ways viz. Land Utilization, Urban and Rural planning, Identification of earth quake prone areas etc by Governmental agencies,.

## **1.2 Objectives**

Objectives of research work areas follows

- A) To develop new Clustering Technique with special emphasis on Spatio-Temporal Geographic Dataset.
- B) To develop hybridized Time Series Prediction Technique for Spatio-Temporal Geographic Dataset.

## **1.3 Scope of the Work**

Scope of research work includes

In order to utilize voluminous data meaningfully one should be able to extract hidden knowledge from it. Literature survey revealed that lot of research has already been done on classification and clustering in different areas. As all geographic phenomena evolve over time both spatially and

temporally, these tasks becomes more challenging. In this context amount of work done found to be less compared to work done only on classical/non-ST-dataset. More focus is required to be laid on this aspect as the data is to be seen not only from static viewpoint, but also from the viewpoint of time and space which is dynamic in nature. Keeping in view these, an assumption has been made, that there exist technique(s), which deals with spatio-temporal dataset and find out shortcomings present in it and propose a solution to overcome these shortcoming(s). Assumption has also been made about the dataset, is that the available ST-Dataset is not continuously varying both spatially and temporally frequently, but emphasis has been laid down on handling voluminous (scalability issue), high dimensional, spatial and time varied dataset, handling nested clusters and accordingly attempt has been made to develop proposed spatio-temporal clustering technique. Moreover while developing time series prediction technique, an assumption of solid mathematical background about machine learning especially Support Vector Machine (SVM), Support Vector Regression (SVR) and statistical approaches (ARIMA) do exists. For validation purpose also from among already existing performance indices, few of them are appropriately selected, which are used for validating the results obtained using ST-Dataset.

## **1.4 Major Contributions**

Major contribution towards present research works are i) Able to develop new Clustering Technique with special emphasis on Spatio-Temporal Geographic Dataset which is able to handle high dimensional data and able handle arbitrary shaped and nested clusters. ii) As clustering can also be used to detect outliers, ST-SNN clustering technique has been developed to find spatio-temporal outliers which is discussed in section 4.2.

In order to meet second objective, a new set of thoughts have been incorporated to develop a hybrid model which is the hybridization of statistical model ARIMA and machine learning model RBF-SVR which is capable to improve the quality of prediction which is clearly indicated in section 5.4.4. Care has been taken to make the newly developed hybrid model for time series prediction, scalable, which helps in reducing the time required to tune the parameters area-wise.

## **1.5 Organization of the Thesis**

For the ease of readers, this section reveals the organization of the present work which is as follows:

Chapter 2: It deals with preliminaries of data mining and its related tasks in general and about clustering, prediction and also basic terminologies in order to understand the proposed clustering and time series techniques in particular.

Chapter 3: Literature survey plays vital role in any research in order to know historical and current development, issues and challenges which can be addressed to overcome the problems present in current systems.

Chapter 4: The first objective of the present research work is discussed here in detail including design and developmental aspects of spatio-temporal clustering technique, it first covers other researcher's contribution in the field, followed by study and experimentation on existing clustering techniques and modification in the existing technique to fulfill the objectives. Emphasis has been laid down on the validation of the obtained results.

Chapter 5: The strategy which has been followed in chapter 4 is also used for prediction of time series. Moreover, owing to the very large number of area-wise models and time required for tuning the parameters, attempt has been made in the same chapter for making the model, scalable. Finally, Chapter 6 deals with summary, conclusion and some suggestion for enhancing the current work in future.

Thesis is ended with cited work which are referred during entire period of research work.

# Chapter 2

## Preliminaries

In order to understand and work presented in forthcoming chapters, this chapter deals with basic terminologies followed by spatio-temporal clustering and time series prediction technique.

### 2.1 Basic Terminologies

#### 2.1.1 Data Mining(Han, Kamber, and Pei)

Enormous amount of data of the order of 1 TB and more are being generated on daily basis. Data Mining is a term referred to as extracting or “mining” hidden patterns/knowledge from large amounts of data, so that these data can be utilized meaningfully and appropriate decision can be taken to perform certain tasks can be taken. Sometimes it is also popularly known as “Knowledge Data Mining, Knowledge Discovery from Data, or simple KDD. This knowledge discovery process consists of following steps of iterative sequence:

1. Data cleaning (to remove noise and inconsistent data)
2. Data integration (where multiple data sources may be combined)
3. Data selection (where data relevant to the analysis task are retrieved from the database)
4. Data transformation (where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance)
5. Data mining (an essential process where intelligent methods are applied in order to extract data patterns)
6. Pattern evaluation (to identify the truly interesting patterns representing knowledge based on some interestingness measures)
7. Knowledge presentation (where visualization and knowledge representation techniques are used to present the mined knowledge to the user)

The knowledge extracted can be used for wide varieties of applications. Major application areas of data mining include land utilization, town planning, market analysis, fraud detection, retention



of customers, production control and science exploration etc. to name a few. Major tasks which may be performed under the banner of Data Mining are Clustering, Classification, Prediction, Summarization, Aggregation, Outlier Detection etc depending upon the need of application domain. Latest research trend in data mining are developing scalable clustering, classification and prediction techniques capable of handling large disk-resident data.

### **2.1.2 Spatial Data Mining(SDM)(Kisilevich et al.)**

Spatial Data Mining have attributes of spatial objects such as latitude, longitude and shape which are highly dependent on location, often influenced by neighboring objects and are stationary in nature in general. Non-spatial attributes are used to characterize non-spatial features of objects (such as name, population, and unemployment rate for a city, NDVI, rainfall etc ). Spatial Data Mining task includes(Kisilevich et al.):Spatial outlier detection,Co-location pattern discovery, Spatial classification, Regression modeling,Spatial clustering, and Spatial hotspot analysis.

### **2.1.3 Temporal Data Mining(Kisilevich et al.)**

Time period attached to the data expresses when it was valid or stored in the database. The ultimate goal of temporal data mining is to discover hidden relations between sequences and subsequences of events.

### **2.1.4 Spatio-Temporal Data Mining(Kisilevich et al.)**

Major tasks of Spatio-Temporal Data Mining includes "The extraction of implicit knowledge, spatial and temporal relationships, or other patterns not explicitly stored in spatiotemporal databases."

In Spatio-temporal clustering basically there are three components: Spatial (i.e. location/where ?), Temporal (i.e. time/when ?) and Objects (i.e. what ?) are used.

Accordingly following three kinds of queries are possible:

**When + Where->What:** Describe the objects or set of objects that are present at a given location or set of locations at a given time or set of times.

**When + What->Where:** Describe the location or set of locations occupied by a given object or set of objects at a given time or set of times.

**Where + What->When:** Describe the times or set of times that a given object or set of objects occupied a given location or set of locations.

### 2.1.4.1 Classification of Spatio-Temporal Data Types (Kisilevich et al.)

Fig 2.1 represents various Spatio-Temporal data types viz. ST events, Geo-referenced variable, Geo-referenced time series, Moving points, Trajectories etc.

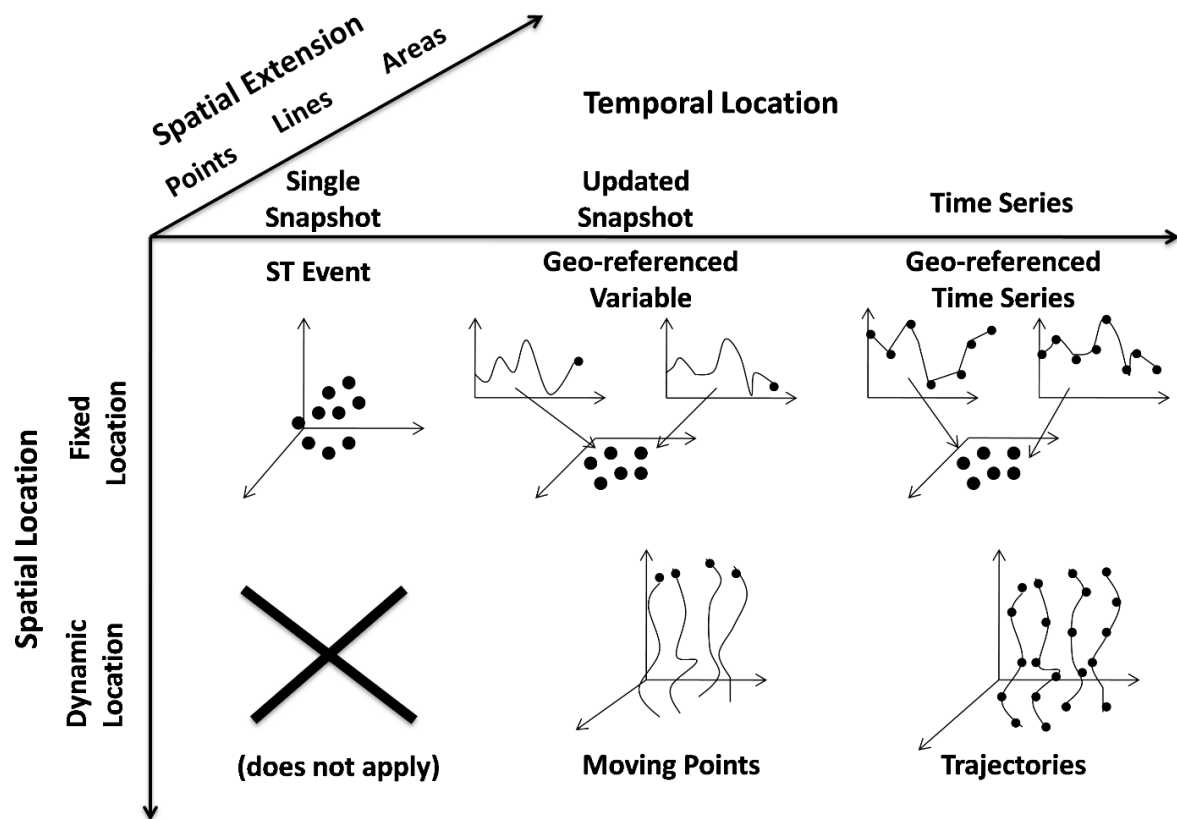


Figure 2.1: Spatio – Temporal Data Types

#### 2.1.4.1.1 ST events

Each event is usually associated with the location where it was recorded and the corresponding timestamp as shown in table 2.1. Both the spatial and the temporal information associated with the events are static, since no movement or any other kind of evolution is possible, for example in case of Earth tremors captured by sensors or geo-referenced records of an epidemic.

Table 2.1: A sample data set of Events Data

Longitude	Latitude	Time
X1	Y1	1
.	.	.
.	.	.
X4	Y4	1
.	.	.
.	.	.
.	.	.
X8	Y8	3
.	.	.

#### 2.1.4.1.2 Geo-referenced variable

"Evolution in time of some phenomena in a fixed location". For example each data items in this case can be a weather station location and corresponding temperature value at the different time sequences.

#### 2.1.4.1.3 Geo-referenced time series

A time series is a sequence of data that represent recorded values of a phenomenon over time means stored the whole history of the evolving object.

#### 2.1.4.1.4 Moving points

When the spatial location of the data object is time-changing, we are dealing with moving objects. e.g. real -time monitoring of vehicles for security applications, and no trace of the past locations is kept.

#### **2.1.4.1.5 Trajectories**

When the whole history of a moving object is stored and available for analysis, the sequence of spatial locations visited by the object, together with the time-stamps of such visits, forms what is called a trajectory.

#### **2.1.5 Clustering (Han, Kamber, and Pei)**

Clustering is the process of forming groups of data into clusters, such that objects within the same cluster have high similarity in comparison to one another but they are very dissimilar to objects in other clusters. These similarities/dissimilarities are assessed based on the attribute values describing the objects, very often, distance measures are used. In a simplistic way it can be said that intra-cluster distances are minimized and inter-cluster distances are maximized. It is a difficult/challenging process compared to classification because of absence of class labels of each object, especially in large databases. Because of absence of class labels, clustering is a form of learning by observation, rather than learning by examples.

Cluster analysis is widely used in numerous applications, including market research, pattern recognition, data analysis, and image processing. For example in business, clustering helps marketers discover distinct groups in their customer bases and characterizes customer groups/clusters based on their purchasing patterns. In biology, it is used to derive plant and animal taxonomies, categorize genes with similar functionality. Clustering also helps in the identification of areas of similar Land Use, using Earth Observation Database, In identification of groups of houses in a city according to house type, value, and geographic location, identification of groups of automobile insurance policy holders with a high average claim cost clustering methods are very useful.

Clustering is also called data segmentation in some applications because clustering partitions large data sets into groups according to their similarity. In machine learning jargon, clustering is an example of unsupervised learning. Tools based on k-means, k-medoids etc. have also been built

into many statistical analysis software packages or systems, such as S-Plus, SPSS, SAS, WEKA, R Package/Language.

### **2.1.6 Classification (Han, Kamber, and Pei)**

Unlike clustering, classification is an example of supervised learning, is a form of learning by examples, rather than learning by observation. In classification, first a model/function which describes and distinguishes data classes or concepts is constructed, which is then used to predict the class of objects whose class label is unknown. For the purpose of constructing classification models, methods available are naïveBayesian classification, support vector machines, k-nearest neighbor classification, to name a few. There is thin line demarcation between classification and prediction, classification predicts categorical (discrete, unordered) labels, whereas prediction models continuous-valued functions, which is used to predict unavailable or missing numerical value rather than class labels.

### **2.1.7 Prediction(Han, Kamber, and Pei)**

Prediction models continuous-valued functions, which is used to predict unavailable or missing numerical value rather than class labels which is categorical in nature ( like “yes”, “no”, “safe”, “risky” ) in case of classification. The term prediction may be referred for both numeric prediction and class label prediction.

Besides other methods, regression analysis is a statistical methodology that is often used for numeric prediction. In prediction, data distribution trends may also be identified which is based on the available data. Relevance analysis, which is an attempt to determine attributes that do not participate in or contribute to the classification or prediction process is performed before classification and prediction, such attributes are then excluded, this process is called dimensionality reduction which helps in improving runtime complexity.

### **2.1.8 Summarization (Han, Kamber, and Pei)**

Summarization gives the subsets of data with related simple description Basic analytical foundation for data preprocessing is provided by summarization of data. Mean, Mode and Median,

are for measuring the central tendency of data, and range, quartiles, interquartile range, variance, and standard deviation are the terms used for measuring the dispersion of data, they provide the basic statistical measures for data summarization. For graphical representations histograms, box plots (Min, 1st quartile, Median, 3rd quartile and Max), scatter plots etc provide the facility for visual inspection of the data and are thus useful for data preprocessing and mining. It is also called as generalization or characterization. It extracts the representative information about the database.

### **2.1.9 Aggregation (Han, Kamber, and Pei)**

Many a times we need to collect data from heterogeneous sources, so preprocessing steps (cleaning, selection, integration, transformation etc) are required to be performed. For the purpose of proper interpretation, and taking decision, it requires consolidation i.e. aggregation, summarization which results in clean, integrated and high quality data.

### **2.1.10 Outlier Detection (Han, Kamber, and Pei)**

Existence of data objects which do not comply with the general behavior or model of the data, such data objects, which are heavily different from or inconsistent with the remaining set of data, are named as outliers. They can be there in the dataset because of measurement or execution error. For instance, the display of a person's age as 350, may be caused by a program default setting of an unrecorded age or may be due to data entry operator's mistakes. They can also present due to inherent data variability. Many algorithms (data mining) attempt to minimize the effect of outliers or eliminate them all together. However many a times outliers is helpful too, for example outliers can help in identifying or detecting fraudulent activities, so it can be said that outlier detection and analysis is very interesting data mining task which is referred to as outlier mining. It has wider application base viz fraud detection (unusual usage of credit cards or telecommunication), customized marketing (for identifying the purchasing behavior of customers), or in medical analysis (for finding unusual responses to various medical treatments).

## 2.2 Spatio-Temporal Clustering Technique

Some basic concepts and terms required to understand the proposed clustering algorithm are given below:

### Definition 1 (Neighbourhood)(Birant and Kut):

It is determined by a distance function (e.g., Manhattan Distance, Euclidean Distance) for two points  $p$  and  $q$ , denoted by  $\text{distance}(p,q)$ .

### Definition 2 ( $\epsilon$ -neighbourhood)(Birant and Kut):

The  $\epsilon$ -neighbourhood of a point  $p$ , denoted by  $N_\epsilon(p)$  is defined by  $\{q \in D \mid \text{dist}(p,q) \leq \epsilon\}$ .

### Definition 3 (Core object)(Birant and Kut):

A core object refers to such point that its neighbourhood of a given radius ( $\epsilon$ ) has to contain at least a minimum number (MinPts) of other points (figure 2.2a).

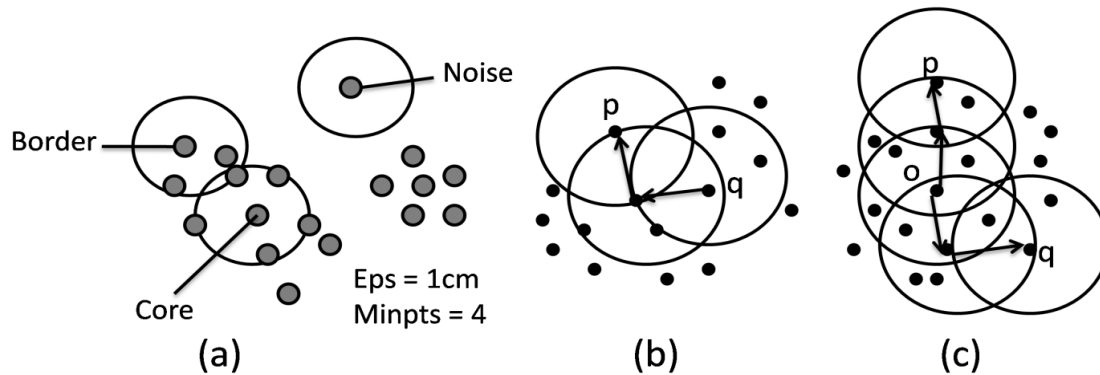


Figure 2.2: Basic Terms1: (a) Border, Core and Noise objects (b) Density-Reachability of  $p$  from  $q$  (c)  $p$  and  $q$  are Density-Connected via object  $o$ .

### Definition 4 (Directly density-reachable)(Birant and Kut):

An object  $p$  is directly density-reachable from the object  $q$  if  $p$  is within the  $\epsilon$ -neighbourhood of  $q$ , and  $q$  is a core object.

### Definition 5 (Density-reachable)(Birant and Kut):

An object  $p$  is density-reachable from the object  $q$  with respect to  $\epsilon$  and MinPts if there is a chain of objects  $p_1, \dots, p_n$ ,  $p_1 = q$  and  $p_n = p$  such that  $p_{i+1}$  is directly density-reachable from  $p_i$  with respect to  $\epsilon$  and MinPts, for  $1 \leq i \leq n$ ,  $p_i$  belongs to  $D$  (figure2.2b).

**Definition 6 (Border object)(Birant and Kut):**

An object is a border object if it is not a core object but density-reachable from another core object (figure2.2a).

**Definition 7 (Density-connected)(Birant and Kut):**

An object  $p$  is density-connected to object  $q$  with respect to  $\epsilon$  and MinPts if there is an object  $o$  belongs to  $D$  such that both  $p$  and  $q$  are density-reachable from  $o$  with respect to  $\epsilon$  and MinPts (figure2.2c).

**Definition 8 (Density-based cluster)(Birant and Kut):**

A cluster  $C$  is a non-empty subset of  $D$  satisfying the following “maximality” and “connectivity” requirements:

- (1)  $\forall p, q$ : if  $q \in C$  and  $p$  is density-reachable from  $q$  with respect to  $\epsilon$  and MinPts, then  $p$  belongs to  $C$ .
- (2)  $\forall p, q \in C$ :  $p$  is density-connected to  $q$  with respect to  $\epsilon$  and MinPts.

**Definition 9 (core-distance of an object  $p$ )(Ankerst et al.) :**

Let  $p$  be an object from a database  $D$ , let  $\epsilon$  be a distance value, let  $N_\epsilon(p)$  be the  $\epsilon$ -neighbourhood of  $p$ , let MinPts be a natural number and let MinPts-distance( $p$ ) be the distance from  $p$  to its farthest MinPts' neighbour. Then, the core-distance of  $p$  defined as (e.g. Eq. (1)):

$$\text{core-distance}_{\epsilon, \text{Minpts}}(p) = \begin{cases} \text{UNDEFINED, if } |N_\epsilon(p)| < \text{MinPts} \\ \text{MinPts-distance}(p), \text{ otherwise} \end{cases} \dots\dots\dots (1)$$

The core-distance of an object  $p$  is simply the smallest distance( $\epsilon'$ ) between  $p$  and an object in its  $\epsilon$ -neighbourhood such that  $p$  would be a core object with respect to  $\epsilon'$  if this neighbour is contained in  $N_\epsilon(p)$ . Otherwise, the core-distance is UNDEFINED.



**Definition 10 (reachability-distance objectpi w.r.t. object o)(Ankerst et al.) :**

Let p and o be objects from a database D, let  $N_\epsilon(o)$  be the  $\epsilon$ -neighbourhood of o and let MinPts be a natural number. Then, the reachability-distance of p with respect to o is defined as (e.g. Eq. (2)):

$$\text{reachability-distance}_{\epsilon, \text{MinPts}}(p, o) = \begin{cases} (\text{UNDEFINED}, \text{if } |N_\epsilon(o)| < \text{MinPts} \\ \max(\text{core-distance}(o), \text{distance}(o, p)), \text{otherwise} \end{cases} \dots\dots\dots(2)$$

Intuitively, the reachability-distance of an object p with respect to another object o is the smallest distance such that p is directly density-reachable from o if o is a core object. In this case, the reachability-distance cannot be smaller than the core-distance of o because for smaller distances no object is directly density-reachable from o. Otherwise, if o is not a core object, even at the generating distance  $\epsilon$ , the reachability-distance of p with respect to o is UNDEFINED. The reachability-distance of an object p depends on the core object with respect to which it is calculated. figure2.3 illustrates the notions of core-distance and reachability-distance.

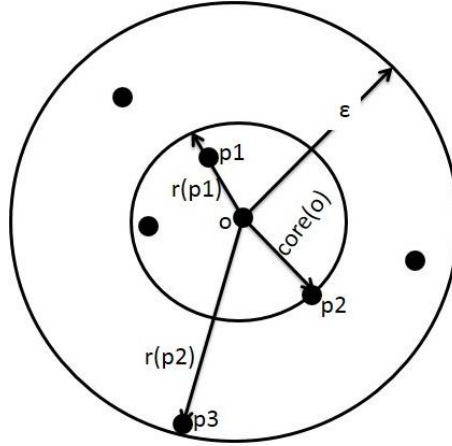


Figure 2.3: Basic Terms2: Core-distance ( $r(o, p2)$ ), reachability-distances ( $r(p1,o)$ ,  $r(p3,o)$ ) for  $\text{MinPts}=4$ .

**Definition 11 (Spatial Radius  $\epsilon_1$ )(Birant and Kut):**

It measures the closeness of two objects/points that are geographically located. For example, there are two objects P and Q and their locations are given by of latitude and longitude such as  $P(x_1, y_1)$  and  $Q(x_2, y_2)$ . The Spatial radius (also called as spatial distance) can be defined as

$$\epsilon_1 = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

**Definition 12 (Non-Spatial Radius  $\varepsilon_2$ )(Birant and Kut):**

It measures the similarity between non-spatial values. For example, there are two objects P and Q and their non-spatial values are day-time-temperature and night-time-temperature such as  $P(dt_1, dt_2)$  and  $Q(nt_1, nt_2)$ . The non-spatial radius (also called as non-spatial distance) can be defined as

$$\varepsilon_2 = \sqrt{(dt_1 - nt_1)^2 + (dt_2 - nt_2)^2}$$

To take into consideration of temporal dimension, spatio-temporal data (year wise) is first merged by retaining the temporal neighbours and their corresponding spatial and non-spatial values. Two objects are temporal neighbours if the values of these objects are observed in consecutive time units, for example back to back days in that year or in that day in continuous years.

Spatio-Temporal Data Mining (STDM) is a process of the extraction of implicit knowledge, spatial and temporal relationships, or other patterns not explicitly stored in spatio-temporal databases. As data is growing not only from static view point, but they also evolve spatially and temporally which is dynamic in nature that is the reason why this field is now becoming very important field of research. In addition, Spatio-Temporal (ST) -Data tends to be highly auto-correlated, because of which assumptions which are taken in Gaussian distribution models fails. Vital issues in spatio-temporal clustering technique for Earth observation data is to obtain good quality arbitrarily shaped clusters and its validation.

Spatio-Temporal Data (Kisilevich et al.) can be explored in term of spatial dimension, temporal dimension and non-spatial measures. Spatial dimension is related about the location for fixed object on the earth, for example latitude and longitude are describing the location of object. Temporal dimension is described about time period attached to the data which expresses when it was valid or stored in the database. Non-spatial measures (also called attribute or characteristic data) are used to characterize non-spatial features of objects (such as name, population, and unemployment rate for a city, normalized difference vegetation index, temperature, rainfall etc).

### **Major requirements of Spatio Temporal Clustering technique :**

- Ability to handle high dimensional data
- Ability to deal with spatial, non-spatial and temporal attributes
- Independent of input data order
- Good Interpretability and usability
- Ability to deal with nested clusters
- Should be Scalable

In real life, most of clustering techniques are very sensitive to their input parameters and results obtained from them are also different and there are no predefined structures of clusters and hence it is difficult to identify:

- i) Correct clustering techniques/algorithms for a dataset,
- ii) Correct clustering structure for different parameters of same clustering algorithm and
- iii) The number of correct clusters in a dataset.

## **2.3 Time Series Prediction Technique(Han, Kamber, and Pei) (Montgomery, Jennings, and Kulahci)**

A time-series consists of sequences of values or events obtained over repeated measurements of time, typically measured at equal time intervals (e.g., hourly, daily, weekly). It is utilized in many applications, such as stock market analysis, economic and sales forecasting, budgetary analysis, utility studies, inventory studies, yield projections, workload projections, process and quality control, observation of natural phenomena (such as atmosphere, temperature, wind, earthquake), scientific and engineering experiments, and medical treatments to name a few. A time-series database is also a sequence database, vice versa is not true, as sequence database consists of sequences of ordered events, with or without concrete notions of time e.g. web page traversal sequences and customer shopping transaction sequences are sequence data, but they may not be time-series data. The amount of time-series data is increasing rapidly, in the order of gigabytes per day (in stock trading), per hour or even or even per minute (such as space programs – receiving data from different satellites, which are varying spatially and temporally as well). Challenging tasks are how we can find correlation relationships, to find similar or regular patterns, trends, bursts (such as sudden sharp changes), and outliers, with fast or even on-line real-time response within time-series data? It can be done with a special emphasis on trend analysis and similarity search.

**Trend analysis consists of the following four major components for characterizing time-series data:**

**Trend/Long-Term Movements:** These indicate the general direction in which a times eries graph is moving over a long interval of time.

**Cyclic Movements:** It refers to the cycles, which may or may not be periodic i.e. the long-term oscillations about a trend line or curve. These cycles need not necessarily follow exactly similar patterns after equal intervals of time.

**Seasonal Movements:** These are systematic or calendar related, e.g. events that recur annually, such as the sudden increase in sales of chocolates and flowers before Valentine's Day or of department store items before Christmas, increase in water consumption in summer due to warm weather etc. In seasonal movements, movements are the identical or nearly identical patterns that a time series appears to follow during corresponding months of successive years.

**Irregular/Random Movements:** These characterizes the sporadic/infrequent motion of time series due to random or chance events, such as labor disputes, floods, or announced personnel changes within companies.

# **Chapter 3**

## **Literature Survey**

Literature survey is very strong component of any research work, as it reveals what research have already been done by different researchers across the globe and what are the issues which still needs to be addressed.

### **3.1 Spatio-Temporal Clustering Technique**

#### **3.1.1 Related Work**

This section summarizes the work related to clustering carried out by various researchers in the field. Basic principle behind clustering (Han, Kamber, and Pei) is the process of putting the objects together in groups based upon certain similarity measures with the aim of minimizing intra-cluster distance among the objects and maximizing inter-cluster distance. Clustering techniques can be categorized as Partitional, Hierarchical, Density Based, Grid Based and Model based Clustering (Kisilevich et al.)(Kolatch).

Partitioning based clustering: Technique possesses the characteristics of offering spherical shaped clusters, similar in size, it does not give natural clusters, but reallocation of an object from one cluster to another is taken care which improves clustering quality. Moreover number of cluster to be obtained is required to be given apriori. Example algorithms are PAM (Partitioning Around Medoids) (Kaufman and Rousseeuw), CLARANS (Ng and Han), K-MEANS and its variants (Baboo and Tajudin) (S. and R. C.).

Hierarchical based clustering: It offers spherical or convex shaped clusters, no reallocation of the object from one cluster to another, it offers natural clusters, and terminal condition has to be specified. Example algorithms are BIRCH (Zhang, Ramakrishnan, and Livny), CURE (Guha, Rastogi, and Shim) etc. Lee and Crawford (Lee and Crawford) have proposed multistage method using hierarchical clustering with a bayesian for unsupervised image classification and provided accurate classification of image with smooth spatial patterns in region merging approach.

Density based clustering: The properties of this kind of technique include detection of arbitrary shaped and different sizes clusters, without taking no. of clusters as input parameter but it is very sensitive to other input parameters like radius (spatial and non-spatial), minimum points, threshold. Example algorithms include DBSCAN (Ester et al.), ST-DBSCAN (Birant and Kut), OPTICS (Ankerst et al.) , DENCLUE (Hinneburg and Keim). A comparative study of three density based clustering algorithms that are DENCLUE, DBCLASD and DBSCAN have been done by (Nagpal and Mann). Result analysis helps in finding the appropriate density based clustering algorithm in variant situations.

Grid based clustering (Ilango and Mohan): It gives natural cluster of any shape and size (arbitrary), independent of number of objects, fast processing time, efficiency better than density based cluster are the few important characteristics of grid based clustering algorithm, such as STING (Wang, Yang, and Muntz), CLIQUE (Agrawal et al.), Wave Cluster (Barnathan)(Chaovalit et al.)(Sheikholeslami, Chatterjee, and Zhang).

Table 3.1 gives the features of different categories of clustering algorithms and examples algorithms of each category.

Where different symbols in table 3.1 indicates :

- n is number of objects.
- k is number of clusters.
- s is the size of sample.
- D is active dataset.
- t is number of iteration.
- h is highest dimensionality.
- g is total no. of Grid cells.

Table 3.1: Categories of Clustering Algorithms – A Comparison

Sr No.	Methods	Features	Algorithms and Complexity	References
1.	Partitioning based clustering	<ul style="list-style-type: none"> <li>- Spherical Shape.</li> <li>- Similar in Size.</li> <li>- Gives no Natural Clusters.</li> <li>- Re-allocation of object from one cluster to other to improve clustering quality.</li> </ul>	<ul style="list-style-type: none"> <li>- K-means (<math>O(nkt)</math>).</li> <li>- K-medoids (<math>O(k(n-k)^2)</math>).</li> <li>- CLARA (<math>O(ks^2 + k(n-k))</math>).</li> <li>- CLARANS (<math>O(kn^2)</math>).</li> <li>- EM (<math>O(n^2)</math>).</li> </ul>	(Baboo and Tajudin) (Han, Kamber, and Tung) (Kolatch)
2.	Hierarchical based clustering	<ul style="list-style-type: none"> <li>- No Reallocation of object from one cluster to other.</li> <li>- Gives Natural Clusters.</li> <li>- Terminal condition has to be specified.</li> <li>- Spherical or convex in shape.</li> </ul>	<ul style="list-style-type: none"> <li>- CHAMELEON (<math>O(n^2)</math>).</li> <li>- CURE (<math>O(n^2(\log(n)))</math>).</li> <li>- BIRCH (<math>O(n)</math>).</li> </ul>	(Han, Kamber, and Tung) (Kisilevich et al.) (Kolatch)
3.	Density based clustering	<ul style="list-style-type: none"> <li>- Gives Natural Cluster.</li> <li>- Any shape and size.</li> <li>- Index based data structure- to reduce time complexity.</li> <li>- Sensitive to input parameter.</li> </ul>	<ul style="list-style-type: none"> <li>- DBSCAN (<math>O(n \log n)</math>).</li> <li>- OPTICS (<math>O(n \log n)</math>).</li> <li>- DENCLUE (<math>O(D(\log(D)))</math>).</li> <li>- DBCLASD (<math>O(3n^2)</math>).</li> </ul>	(Birant and Kut) (Han, Kamber, and Tung) (Kolatch) (Nagpal and Mann)
4.	Grid based clustering	<ul style="list-style-type: none"> <li>- Gives Natural Clusters.</li> <li>- Any shape and size.</li> <li>- Independent of number of objects.</li> <li>- Fast processing time.</li> <li>- Efficiency better than Density based cluster.</li> </ul>	<ul style="list-style-type: none"> <li>- STING (<math>O(ng)</math>).</li> <li>- WAVE Cluster (<math>O(n)</math>).</li> <li>- CLIQUE (<math>O(k^h + nh)</math>).</li> </ul>	(Han, Kamber, and Tung) (Ilango and Mohan) (Kolatch) (Sheikhol-eslami, Chatterjee, and Zhang)



Keeping in view major requirements of Spatio-temporal Clustering as mentioned in section 2.2, from the comparison (table 3.1) it is clear that Density Based and Grid Based Clustering algorithms suites to our requirements as they are able to identify arbitrary shaped clusters, unlike Partitional and Hierarchical based categories of algorithms which are able to identify only spherical shaped clusters.

Now as next step in the literature survey, an attempt has been made to dig into Density Based and Grid Based Clustering algorithms.

Clustering method DBSCAN (Density Based Spatial Clustering of Applications with Noise) is proposed by(Ester et al.) for obtaining arbitrary shaped clusters and compared with CLARANS (Clustering Large Applications based on RANdomized Search) and found that average run time complexity of DBSCAN is  $O(n \cdot \log n)$  which is better than CLARANS i.e.  $O(kn^2)$ .

ST-DBSCAN algorithm (Birant and Kut), an extension of DBSCAN algorithm which focuses on obtaining arbitrary shaped cluster with special emphasis on spatio-temporal data having attributes like sea surface temperature, sea surface height residual, significant wave height and wind speed values of four seas. The average run time complexity of ST-DBSCAN is similar to DBSCAN algorithm.

Another approach for obtaining good quality clusters which is also taking care of nesting of clusters (unlike ST-DBSCAN) besides arbitrary shaped clusters called OPTICS (Ordering Points to Identify Clustering Structure) proposed by (Ankerst et al.) , which actually does not produce clusters explicitly but gives ordering to points which represents its density based clustering structure, which is helpful in detecting nested level of clustering.

(Kisilevich et al.) first discusses on the classification of different types of spatio-temporal data: ST events, Geo-referenced variables, Moving objects and Trajectories and focuses on trajectory. Moreover work shows different approaches of spatio-temporal clustering and finally presents several scenarios in different application domain such as movement, cellular networks and environmental studies.

The work presented by (Han, Kamber, and Tung), focuses on spatial clustering in different application such as identification of similar area of Land Usages in Earth observation data or in merging regions with similar weather pattern etc. Moreover, authors have also discussed the categories of clustering and presented their algorithms. (J., J., and P.) in 2013 has presented temporal trends in the Indian seasons and its variability at spatial resolution.

(Shekhar et al.) explored the different tasks of spatial data mining (i.e. spatial outlier detection, co-location patterns, spatial classification and regression models, spatial clustering etc.) and presented different methods to extract patterns from spatial information.

The differences among classical, spatial and spatio-temporal data have described by (Rashid and Hossain), moreover authors have highlighted the different challenging issues in spatio-temporal data mining.

(Wang, Wang, and Li) have proposed two spatio-temporal clustering methods (i.e. ST-GRID and ST-DBSCAN) based on neighborhood searching strategy to obtain clusters using Geo-databases. ST-GRID searches between neighborhood cells while ST-DBSCAN searches the neighborhood of points. Moreover, ST-GRID needs additional disk space to store the grid structure, because of which it has lower space efficiency than ST-DBSCAN and clustering precision of ST-GRID is a little lower than ST-DBSCAN. Because of these shortcomings of ST-GRID, ST-DBSCAN has been selected for comparing the results of proposed technique.

(Morgan et al.) presented land cover classification based on hyper spectral data and validation done through ground truth data.

Table 3.2 gives the Comparative study of different Density and Grid Based clustering algorithms and examples algorithms of each category.

Table 3.2: Density and Grid Based Clustering Algorithms – A Comparison

Sr. No	Algorithm	RunTime Efficiency (Avg) where n - Number of objects k - Number of clusters. D- Active portion of dataset. h - Highest dimensionality.	Handle High Dimensionality	Handle Irregular shape	Insensitive to Noise	Independent of Data input order	Input Parameters	Support ST Data?
1	DBSCAN(Ester et al.)(Han, Kamber, and Tung)(Kolatch)(Nagpal and Mann)	$O(n \log n)$	YES	Not completely	YES	NO	Radius, Minimum no. of points	NO
2	ST-DBSCAN(Birant and Kut)(Han, Kamber, and Tung)(Kolatch)(Nagpal and Mann)	$O(n \log n)$	YES	YES	YES	NO	Spatial radius, Non-spatial radius, Minimum no. of points, Threshold	YES
3	OPTICS (Ankerst et al.) (Han, Kamber, and Tung)(Kolatch)(Nagpal and Mann)	$O(n \log n)$	YES	YES	YES	YES	Radius, Minimum no. of points	NO
4	DENCLUE (Hinneburg and Keim)(Han, Kamber, and Tung)(Kolatch)(Nagpal and Mann)	$O(D \log D)$	Some what	YES	YES	YES	Threshold for the influence of data points, Threshold for significance of density-attractor	NO
5	DBCLASD(Han, Kamber, and Tung)(Kolatch)(Nagpal and Mann)	$O(3n^2)$	NO	Better than DBSCAN	YES	NO	No input parameters	NO
6	WAVE Cluster(Han, Kamber, and Tung)(Ilango and Mohan)(Lee and Crawford)(Sheikholeslami, Chatterjee, and Zhang)	$O(n)$	Some what	YES	YES	YES	Wavelet, the no. of grid cells for each dimensions, the no. of application of wavelet transform	NO
7	CLIQUE (Agrawal et al.)(Han, Kamber, and Tung)(Ilango and Mohan)(Kolatch)	$O(k^h + nh)$	YES	Minimal	Partially	YES	No. of intervals, Density Threshold	NO
8	STING (Han, Kamber, and Tung) (Ilango and Mohan)(Kolatch) (Wang, Yang, and Muntz)	$O(k)$	NO	Cluster are approx	YES	YES	No. of objects in a cell	NO

From table 3.2, it is obvious that ST-DBSCAN algorithm fulfills objectives of research work as mentioned in section 1.2 and that is why it has been chosen for experimentation purpose.

### **Related Work: Validation Indices**

Cluster validation indices plays important role in validating the obtained clusters which has been categorized as follows (Halkidi, Batistakis, and Vazirgiannis)(Theodoridis and Koutroumbas): 1. External Index: Measure similarity of clusters against known class labels. e.g. Entropy, F-measures and Rand statistic etc. 2. Internal Index: Measure the goodness of clusters without any external information. e.g. using Sum of Squared Error (SSE) method, others (inter, intra cluster distances) etc. 3. Relative Index: Compare two different clustering structure using either external or internal indices measures.

(Weingessel, Dimitriadou, and Dolnicar) have used cluster validation indices for determining the number of clusters in high dimensional binary datasets using k-means and hard competitive learning algorithms. It is based on voting criteria, in which it is required to find out ‘number of clusters’ in particular dataset, they considered maximum choice obtained from all indices.

(Milligan) and (Cooper) have studied and experimented many validation indices also called them “Stopping criteria” as these indices were helpful to hierarchical clustering algorithm to terminate. In the study (G. et al.), cluster validation indices are used for a quantitative evaluation of clustering results using DB and Huberts indices and checked their reliability.

Performance/reliability of indices are varying depending on clustering method, data structure and clustering objective suggested by (Shim, Chung, and Choi) and found best six indices (i.e. the Calinsky and Harabasz, Ray and Turi, Davis and Bouldin, and G(+)) among sixteen cluster validation indices that supports k-means algorithm.

(Bernard) discussed and implemented 27 internal cluster validation indices developed for R-package named “clusterCrit”. This package also includes some external cluster validation indices.

New sum of squares based validation index (WB) for homogeneous data based on independent variables is proposed and also done the effective comparison to some other commonly used indices proposed by (Zhao, Xu, and Fränti) using k-means algorithm.

More recently, (Arbelaitz et al.) compared cluster validation indices using three different algorithms: k-means, Ward and average-linkage. Their studies showed that the experimental factors, noise and cluster overlap have greatest impact on cluster validation indices.

A new tool CVAP (Cluster Validity Analysis Platform) has been developed by (Wang, Wang, and Peng) to evaluate clustering results and to validate clusters.

Relative criteria was used to validate clusters and have performed experimentation with RS, RMSSTD, DB, SD and S\_Dbw indices using k-means and CURE clustering algorithm by (Halkidi, Batistakis, and Vazirgiannis), but not effective for arbitrary shaped clusters.

(Kovács, Legány and Babos) have put their sincere effort to validate clusters using relative and internal criteria, attempt is made to validate arbitrary shaped clusters. For experimentation they used Dunn, DB, SD, S\_Dbw indices using two different clustering algorithm k-means and DBSCAN and on three different datasets.

Based on literature survey also revealed that lot of work have been done to validate obtained clusters using internal validation indices. All of them used artificial and synthetic datasets using non-hierarchical (i.e. k-means) and hierarchical clustering algorithms (i.e. agglomerative and divisive). Very less amount of work has been done to validate densed and arbitrary shaped clusters formed using spatio-temporal data.

### **Mathematical Background for Validation Indices(Bernard):**

There are total of 27 internal cluster validation indices and also they have been used for comparison using relative criteria. All of them work on following mathematical background.

Let us define dataset  $M$  as a set of  $N$  observation and  $d$  dimension i.e.  $N \times d$  Where  $M = \{M_1, M_2, \dots, M_N\}$ .

The dataset is assumed to be partitioned in  $K$  clusters or groups. The coordinates of  $M_i$  are the

coefficients of the  $i^{\text{th}}$  row of dataset  $M$ . The set of observation belongs to cluster  $C^k$  or  $C^{(k)}$  is denoted by  $I_k$ .

Let us denote by  $G^{(k)}$  the centre of the observations ( $n_k$ ) in the cluster  $C^{(k)}$  and by  $G$  the centre of all the observations.

$$G^{(k)} = \frac{1}{n_k} \sum_{i \in I_k} M_i$$

$$G = \frac{1}{N} \sum_{i=1}^N M_i$$

The within-cluster dispersion, noted  $WGSS^{(k)}$ , is the trace of scatter matrix and define as

$$WGSS^{(k)} \text{ or } Tr(WG^{(k)}) = \sum_{i \in I_k} ||M_i^{(k)} - G^{(k)}||^2$$

Finally the within-cluster sum of squares WGSS is the sum of the within-cluster dispersion for all the clusters:

$$WGSS \text{ or } Tr(WG) = \sum_{k=0}^K WGSS^{(k)}$$

The between-cluster dispersion, noted BGSS, is define as

$$BGSS \text{ or } Tr(BG) = \sum_{k=1}^K n_k ||G^{(k)} - G||^2$$

And TSS (total sum of squared) = WGSS + BGSS.

The basic idea for some indices is based on pairs of points in which one can try to distinguish the pairs of points belonging to the same and different cluster. The total number of pairs of distinct points in the cluster  $C^{(k)}$  are  $n_k(n_k-1)/2$  which are not depends on order of points. The total number of such pairs NW is defined as

$$NW = \sum_{k=1}^K \frac{n_k(n_k-1)}{2}$$

The total number of pairs of distinct points in the dataset is defined as

$$NT = \frac{N(N-1)}{2}$$

Let us  $NT = NW + NB$  where NB is the number of pairs constituted of points which do not belong to the same cluster and it is defined as

$$NB = \sum_{k < k'} n_k n_{k'}$$

Let us denote by IB the set of the NB pairs of between cluster indices and IW the set of the NW pairs of within cluster indices.

### 3.1.2 Issues and Challenges

Following are the issues and challenges related to clustering : Discovery of clusters with arbitrary shape, ability to handle high dimensional data, ability to deal with spatial, non-spatial and temporal attributes, independent of input data order, interpretability and usability, ability to deal with nested clusters, scalability.

### **3.2 Time Series Prediction(Han, Kamber, and Pei)(Montgomery, Jennings, and Kulahci)**

In order to take appropriate action well in advance, for example proper growth of vegetation in near future with the help of its prediction having correlation with rainfall. It also helps in analysis of drought area, we can also differentiate bare soil from grass or forest, we can detect plants under stress, and differentiate between crops and crop stages. Prediction of an accurate NDVI with the help of rainfall means a lot from economic perspective too. But as said earlier, analysis of data which are varying spatially and temporally is a challenging tasks.

Many statistical models are available like AR, MA, ARMA, ARIMA etc. for prediction purpose. All these models are for stationary time series (where mean remains constant) except ARIMA which can also forecast for non-stationary time series(where mean is variant). But all these statistical models work for a single time-series only.

The SVR creates “Prediction-maker” system which is trained from historical time series and attempts to predict new value for near future based on dependency, like prediction of NDVI time series from rainfall data, as NDVI depends on rainfall. The basic idea about SVR as summarized by (Smola and Scholkopf) is to map the input data into high dimensional feature space through a nonlinear mapping function, and to solve a linear regression problem in this feature space. Different kinds of kernel functions are available in SVR like Linear, Polynomial, RBF (Radial Basis Function) etc. The usability of SVR increases if we use any one of the kernel function listed above, but at the same time it gives rise to new problem of scalability as parameters have to be tuned for optimal performance.

### 3.2.1 Related Work

NDVI can be used effectively for crop yield forecasting as shown by (Mkhabela et al.) and similarly (Zhang, Lei, and Yan) shows that along with other factors like precipitation and temperature data, it can be used for comparison of regression models like OLS (Ordinary Least Square) and Spatial Autoregressive method for crop yield prediction. The above work enables one to predict crop yields accurately if prediction of NDVI is accurate.

(Iwasaki) have put his sincere efforts get NDVI prediction models using Multiple Linear Regression over precipitation and temperature data.

(Omuto et al.) Using have attempted to perform tasks like detection of human-induced loss of vegetation cover using NDVI and rainfall data, where authors proposed a mixed-effects method which considers the influence of vegetation types whereas the global model does not consider the influence of vegetation types in the modeling process. Mixed-effects method had a higher accuracy in modeling NDVI-rainfall relationship than the global model.

(Foody) uses OLS and GWR (Geographically Weighted Regression) for NDVI-rainfall modeling and find out that GWR gives better results compared to OLS. (Kumar and Rao) has contributed to find the relationship between NDVI and rainfall considering other factors like air temperature, soil moisture adequacy and ENSO (El Nino Southern Oscillation) indices particularly in India using linear regression.

Machine learning algorithm, SVR has been used for classification of crops using NDVI time series as shown in (Niazmardi et al.) where they have proposed modified ML (Maximum likelihood) algorithm that uses SVR algorithm to estimate the probability of each class.



(Deng, Shen, and Tian) has explored the applicability of time series analysis for stock trend forecast and present Self projecting Time Series Forecasting (STSF). (Liu, Yao, and Zhu) has nicely attempted hybrid approach combining SVM and Improved Quantum-behaved Particle Swarm Optimization (IQPSO) for image classification.

(Zhu and Xu) have presented their work on prediction algorithm based on data mining especially for multimedia objects in next generation Digital Earth. Statistical Techniques to identify stable relationship in the monitored data which characterises normal operation and can help in detecting anomalies is presented by (Munawar and Ward).

### **3.2.2 Issues and Challenges**

Issues related to prediction of time series having different measures are :

- Handling non-stationarity of time series,
- Handling high dimensionality of dataset,
- Handling non-linear domain,
- Selection and tuning of parameters in case of ACF and PACF plots and Kernel function (Radial Basis Function).

Statistical model like Integration of Auto Regressive (AR) and Moving Average (MA) i.e. ARIMA model is capable to handle non-stationary time series but it can deal with only single time series. While machine learning approach (i.e. Support Vector Regression (SVR)) can handle dependency among different time series along with non-linear separable domains, however it cannot incorporate the past behavior of time-series. Summary of these shortcomings are given in table 3.3 and table 3.4.

Table 3.3: Shortcomings of AR, MA, ARMA, ARIMA, SARIMA models

Property	AR	MA	ARMA	ARIMA	SARIMA
1. Support Non-Stationary Time Series	NO	NO	NO	YES	YES
2. Support Seasonal Time Series	NO	NO	NO	NO	YES
2. Support Multi Measures Time Series	NO	NO	NO	NO	NO

Table 3.4: Shortcoming of RBF-SVR model

Property	RBF-SVR
1. Model Perform well- when no correlation exists	NO
2. Parameter tuning	Difficult

Keeping in view challenges and shortcomings of different models independently, a new model has been developed by hybridizing statistical model ARIMA and machine learning model RBF-SVR considering residuals to predict time series of NDVI using rainfall time series data as explained in section 5.1.

Moreover, in order to reduce number of area wise models and reduction in time complexity for tuning different parameters, emphasis has been laid down on handling issues related to scalability by taking suitable representative samples from each sub-areas.

## Chapter 4

### Proposed Spatio-Temporal Clustering Technique

Spatio-Temporal Data (Kisilevich et al.) can be explored in terms of spatial dimension, temporal dimension and non-spatial measures. Spatial dimension is related about the location for fixed object on the earth, for example latitude and longitude are describing the location of object. Temporal dimension is described about time period attached to the data which expresses when it was valid or stored in the database. Non-spatial measures (also called attribute or characteristic data) are used to characterize non-spatial features of objects (such as name, population, and unemployment rate for a city, normalized difference vegetation index, temperature, rainfall etc). Major requirements of Spatio Temporal Clustering technique are :Discovery of clusters with arbitrary shape, Ability to handle high dimensional data, Ability to deal with spatial, non-spatial and temporal attributes, Independent of input data order, Good Interpretability and usability, Ability to deal with nested clusters moreover it should be scalable too.

From section 3.1.1 and table 3.1 and table 3.2 it is very clear that DBSCAN, ST-DBSCAN and OPTICS suites to objective of the research work, initial experimentation have been performed using these algorithms. Sincere efforts have been put to propose new technique named as Spatio-Temporal – Ordering Points to Identify Cluster Structure (ST-OPTICS) which is modified version of OPTICS to emphasize on ST-Dataset, which has been further hybridized with agglomerative approach for better visualization and interpretation of results obtained. Main motivation behind selection of OPTICS is that “it is able to handle with nested clusters” (Ankerst et al.) , but does not support spatio-temporal data, care has been taken in proposed technique for the same.

## 4.1ST-OPTICS: Clustering Technique

OPTICS (Ankerst et al.) method is started by exploring the dataset and enumerating all the objects. For each object  $p$  it checks if the core object conditions are satisfied and in the positive case, starts to enlarge the potential cluster by checking the condition for all neighbours of  $p$ . If the object  $p$  is not a core object, the scanning process continues with the next unvisited object in dataset. The results were discussed in a reachability plot (Ankerst et al.), the objects are represented on the horizontal axis in the order of visiting them and the vertical dimension represents their reachability distances. Intuitively, the reachability distance of an object  $p_i$  corresponds to the minimum distance from the set of its predecessors  $p_j$ ,  $0 < j < i$ . As a consequence, a high value of the reachability distance roughly means a large distance from the other objects, which indicates that the object is in a sparse area. The actual clusters may be determined by defining a reachability distance threshold and grouping together the consecutive items that are below the chosen threshold in the plot. The result of the OPTICS algorithm is insensitive to the original order of the objects in the dataset, however this technique do not take care of spatio-temporal data.

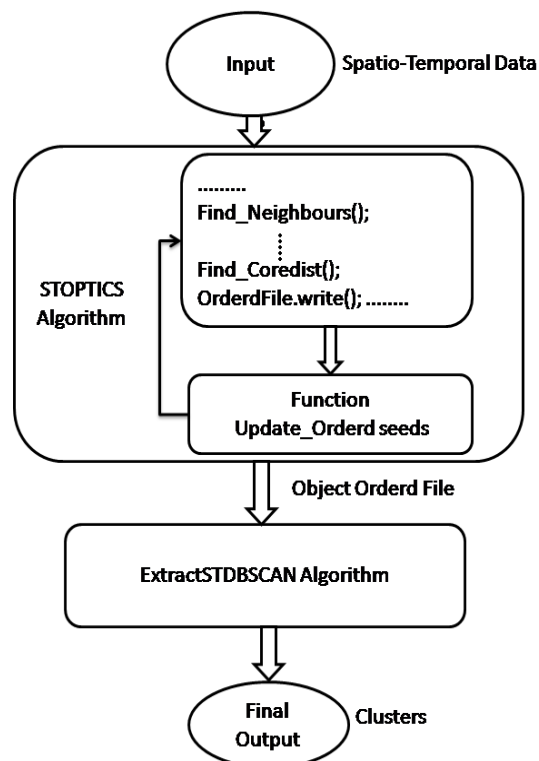


Figure 4.1:Proposed Approach (ST-OPTICS).

Logical flow of proposed technique ST-OPTICS has been shown in Figure4.1, the results obtained from this technique are compared with the result of ST-DBSCAN but as DBSCAN and OPTICS are not able to deal with spatio-temporal database, their result are not comparable.

#### 4.1.1 The Description of the Algorithm

While OPTICS algorithm creates an ordering of a database based on core distances and reachability distances and uses three parameters ( $\epsilon$ , MinPts, OrderedFile). Our algorithm ST-OPTICS creates an ordering of Spatio-Temporal (ST) database based on core and reachability distances but they are calculated from the two radius ( $\epsilon_1$ ,  $\epsilon_2$ ) and also uses one additional parameter CoRememberFile.  $\epsilon_1$  and  $\epsilon_2$  are used to handle ST data and processing time is reduced due to  $\epsilon_1$ . CoRememberFile is very useful to automatically calculate the values of new  $\epsilon'$  i.e. min\_RD and max\_CD for clustering. The heuristic (Ester et al.) is used to determine the values of radius ( $\epsilon_1$ ,  $\epsilon_2$ ) and threshold, moreover in order to improve accuracy while finding these values, work has also been done to obtain these values automatically from k-distance graph rather than observing manually. Euclidean distance measure is used for distance calculations.  $\epsilon_1$  is used to calculate spatial distances and it also helps to reduce computational time by restricting search spatial area.  $\epsilon_2$  is used to calculate non-spatial distances. figure 4.1 shows the logical flow of proposed algorithm ST-OPTICS.

Proposed algorithm (*Algorithm 4.1*) starts with the first object ( $ob_1$ ) in database D and checks if  $ob_1$  is not yet processed then goes further, after that it finds the neighbours with respect to  $\epsilon_1$  and  $\epsilon_2$ . If  $ob_1$  is core object (please refer definition 3 in section 3), it is remembered by CoRememberFile which is used at the time of clustering to calculate parameters min\_RD, max\_CD. Set the reachability-distance (RD - a vector) of  $ob_1$  to UNDEFINED, determines its core-distance (CD - a vector) and writes object to OrderedFile. If CD of  $ob_1$  is defined, it goes for processing of neighbours. A call of function **Update\_Orderseeds (neighbours, object, CD)** as shown in *Algorithm 4.2* returns the RD of neighbours and Seeds (it contain the neighbours (objects) in sorted order by their reachability distance which means according to closeness to core object) and

manages Seeds-list by i) Inserting objects in a Seeds which are not yet present in priority-queue of Seeds. ii) Objects which are already in the queue (Seeds)

are moved further to the top of the queue if their new\_RD is smaller than their previous RD. iii) Sorting objects in the Seeds according to reachability-distance<sup>1</sup>.

### **Algorithm ST-OPTICS(D, $\epsilon_1$ , $\epsilon_2$ , MinPts, OrderdFile, CoRememberFile)**

// Input:

// D={ob<sub>1</sub>,ob<sub>2</sub>,.....,ob<sub>n</sub>} set of points(objects)  
 //  $\epsilon_1$ = Spatial distance value.  
 //  $\epsilon_2$ = Non-spatial distance value.  
 // Minpts= Minimum number of points required in radiuses ( $\epsilon_1$ ,  $\epsilon_2$ )

// Output:

// OrderdFile= It contains Objects in Order.  
 // CoRememberFile= It contains only Core object that satisfies condition of MinPts.

For all objects do

  If ob<sub>i</sub>.Processed != TRUE Then  
    N=neighbours(ob<sub>i</sub>,  $\epsilon_1$ ,  $\epsilon_2$ ) //Find neighbours of object.  
    ob<sub>i</sub>.processed = TRUE // Set object as processed.  
    InsertToCoRememberFile(ob<sub>i</sub>, Minpts, N) // Remember in CoRememberFile.  
    ob<sub>i</sub>.CD = Find Core distance of ob<sub>i</sub>//Find Core Distance(CD).  
    ob<sub>i</sub>.RD = UNDEFINED  
    InsertToOrderFile(ob<sub>i</sub>) //Write in Order File.  
    If ob<sub>i</sub>.CD != UNDEFINED Then  
     CallUpdate\_Orderseeds(N, ob<sub>i</sub> , ob<sub>i</sub>.CD) // **call of function.**  
     Repeat Seeds until not empty  
       // Seeds={S<sub>1</sub>,S<sub>2</sub>,.....,S<sub>m</sub>} set of points(objects)  
       S<sub>i</sub>.N = neighbours(S<sub>i</sub>,  $\epsilon_1$ ,  $\epsilon_2$ )  
       S<sub>i</sub>.processed = TRUE  
       InsertToCoRememberFile(S<sub>i</sub>, Minpts, S<sub>i</sub>.N)  
       S<sub>i</sub>.CD = FindCoredistance of S<sub>i</sub>  
       S<sub>i</sub>.RD = UNDEFINED  
       InsertToOrderFile(S<sub>i</sub>)  
       If S<sub>i</sub>.CD != UNDEFINED Then  
         CallUpdate\_Orderseeds(N, S<sub>i</sub> , S<sub>i</sub>.CD) // **call of function.**  
       End If  
     End Repeat  
 End If

---

<sup>1</sup>K.P. Agrawal, Sanjay Garg, Shashikant Sharma and Pinkal Patel, Development and Validation of OPTICS Based Spatio -Temporal Clustering Technique, Information Sciences (2016), Elsevier, Thompson Reuters Impact Factor 2015 : 3.364, H index :114,

```

        End If
    End For
    Find the max_CD and min_RD from the CoRememberFile.

```

Algorithm 4.1: ST-OPTICS algorithm.

Process each object in Seeds until it is not empty following the step given. First, find the  $\epsilon$  - neighbourhood of object which is available from Seeds, its CD, mark processed as TRUE and also check condition for CoRememberFile. If current Object ( $S_i$ ) is a core object, then proceed for further candidates/objects to call Update\_Orderseeds function. At the end, calculate min\_RDs and max\_CDs from CoRememberFile (min\_RD is the minimum reachability distance from the number of objects in CoRememberFile and similarly max\_CD is the maximum core-distance).

**Function Update\_Orderseeds (neighbours, Centerobject, CD)**

```

// For ordering structure, this function deals with both spatial and temporal dimensions.
// neighbours={p1,p2,...,pk}
For all neighbours do
    If pi.Processed != TRUE Then
        pi.New_RD = Find RD of pi// Find the Reachability Distance (RD).
        If pi.RD = UNDEFINED Then
            pi.RD = pi.New_RD
            InsertToSeeds(pi) // append the object pi at Seeds.
        Else
            If pi.New_RD < pi.RD Then
                RD. pi = pi.New_RD
                DecreaseToSeeds(pi) // put the object pi at top of Seeds.
            End If
        End If
    End If
End For
Sort the Seeds file according to reachability distance.

```

Algorithm 4.2: Update\_Orderseeds function.

For clustering, **Extract\_STDBSCAN(OrderdFile, min\_RD, max\_CD, threshold)** algorithm is used as shown in *Algorithm 4.3*. Initially, it assigns ClusterID as Noise and set zero value in Cluster\_Avg() (Note: the value of Cluster\_Avg() for particular cluster is varying according to objects added to them). Further, it is first checked whether the RD of current object are larger than the min\_RD (for both spatial and non-spatial) and the difference between the average value of the cluster and current object value is larger than the threshold (Birant and Kut), If condition is false, same ClusterID is assigned to object else check for the CD of current object whether they are

smaller or equal to the max\_CD (for both spatial and non-spatial) then assign new ClusterID otherwise mark it as NOISE.

#### **Extract\_STDBSCAN (OrderFile, min\_RD, max\_CD, threshold)**

```
// OrderFile = { od1, od2, ..., odn } set of points(objects)
Set ClusterID as NOISE;
For all objects do
    Select object from OrderFile.
    If odi.RD > min_RD and |Cluster_Avg() – odi.value| > threshold Then
        If odi.CD < max_CD Then
            Assigns new ClusterID to object.
        Else
            Set Object as NOISE.
        End If
    Else
        Assigns old ClusterID to object.
    End If
End For
```

Algorithm 4.3: Extract\_STDBSCAN algorithm.

Note: i) The reachability distance of the first object in the OrderdFile is always UNDEFIND and it is assumed that UNDEFIND to be much greater than any defined distance. ii) The variables (CD, RD, min\_RD, max\_RD) are arrays containing both spatial and non-spatial values.

Based on working of proposed technique, it has been felt intuitively that the technique is capable to handle issues mentioned in section 3.1.2, and experimental result in section 4.3.4 also indicates the same.

#### **Validation of Spatio-Temporal Clustering Algorithm**

In order to validate results obtained from developed technique, following approaches have been used:

1. Theoretical Approach
2. Practical Approach

##### **Theoretical Approach:**



While applying first approach, theoretical principle based on mathematical background following indices are studied and in the second approach, experimentations are performed on data which generates dense and arbitrary shaped clusters having different sizes (where in representative point of each cluster is never fixed(Kovács, Legány and Babos)). After theoretical study, indices which are selected for validation are given in Table 4.1 with remarks of their applicability and indices which are not applicable for spatio-temporal clustering are given in table 4.2 with remarks/justification saying that why they not suitable (please note that remarks are objective and extensive rather than an exhaustive).

In first column of table 4.1 and 4.2, (↑) indicates that larger value of index provided by algorithm means it generates good quality of clusters compared with other algorithms and in case of (↓) it is vice - versa.

**Table 4.1: Selected Validation Indices with their Significance and Applicabilities**

<b>Index</b>	<b>Significance of Index</b>	<b>Remarks on Applicability</b>
Ball-Hall (↑)	It works on finding mean dispersion within each cluster irrespective shape and size. Works on representative point	Able to work on arbitrary shaped cluster which is need of ST-data.
Det_Ratio(↓)	It is ratio of total dispersion and within cluster dispersion	Minimization rule of this index suggest that this is not suitable for spherical shaped cluster, however for arbitrary shaped cluster it seems to be suitable, and also verified after performing experimentation.
Dunn (↑)	It is ratio of minimal distance between points of different clusters and the maximal distance within cluster. It is very sensitive to noise	It does not depend on representative points, good for arbitrary shape cluster and verified experimentally.
Gamma (↑)	It is the index of correlation between two vectors of data with the same size. One vector is the set of distances between pairs of points and the second vector is a binary.	It does not depend on representative points, good for well separability between clusters and verified experimentally.
G + (↓)	It is based on Gamma. It measures separation based on distance between two points those do not belonging to the same cluster and this distance is greater than the distance between two points belonging to the same cluster.	It does not depend on representative points, good for well separability between clusters and verified experimentally.
Generalized Dunn's Index (GDI) (↑)	It's generalization of dunn index, ratio of minimum of between cluster and maximum of within cluster distances. It has 18 variations. These indices use different definition for cluster distance and cluster diameter.	Based on theoretical principle, it can be concluded that inter-cluster distance play vital role in well separability of clusters obtained, literature (Bezdek and Pal) reveals that two variations GDI31 and GDI51 are most reliable measure for inter cluster distance and also have been verified experimentally.

Ksq_DetW (↑)	It computes cohesion by within cluster-dispersion only.	As it does not employ calculation of between cluster dispersion and in case of arbitrary shaped cluster and ST-Data, as such no parameters like distances among clusters are same so this index is independent of between cluster dispersion.
Log_Det_Ratio (↓)	It is logarithm of Det_Ratio.	Minimization of this index suggest that it is not suitable for spherical shaped cluster, however for arbitrary shaped cluster it seems to be suitable, also verified after performing experimentation, however in few cases authors got invalid index value, and it is analysed that when variability among the data is very less, then it is feasible.
Point Biserial(↑)	It is a correlation measure between continuous variable and binary variable.	It does not depend on representative points, Good for well separability between clusters and verified experimentally.
Tau (↑)	It is the ratio where numerator part is same as Gamma index to the total number of pair of points.	It does not depend on representative points, good for well separability between clusters and verified experimentally.
Trace_W (↑)	It finds the best division of cluster by analysis of variance.	It does not depend on representative points.

Table 4.2: Unselected Validation Indices with their Significance and Reasons

Index	Significance of Index	Remarks on Non-applicability
BanfeldRaftery (↓)	It is the weighted sum of the logarithms of traces of variance-covariance matrix of each cluster. It computes index based on mean of squared distances between points in cluster.	Sum of squares criterion is most appropriate when clusters all have the same dimensions (Theodoridis and Koutroumbas), Works on representative point
C- index (↓)	This index is measures of between-cluster isolation and within-cluster coherence.	It is data dependent and it is not straight forward to compute maximum or minimum within group sum of square in non-hierarchical clustering (Symons).
CalinskiHarabasz (↑)	It is the ratio where the cohesion is estimated based on the instances from the points in a cluster to its centroid.	It depends on representative points of clusters which is not fixed in arbitrary shaped clustered.
Davies Bouldin (↓)	It estimates the cohesion based on the distance from the points in a cluster to its centroid and the separation based on the distance between centroids.	It depends on representative points of clusters which is not fixed in arbitrary shaped clustered.
Log_SS_Ratio (↓)	It is logarithm of ratio of between cluster dispersion and within cluster dispersion.	Distances among the clusters are not uniform/same in ST-Data, it depends on representative points of clusters which is not fixed in arbitrary shaped clusters.
McClain Rao (↓)	It is minimized the ratio of average dispersion of within cluster and between clusters.	Distances among the clusters are not uniform/same in ST-Data, it depends on representative points of clusters which is not fixed in arbitrary shaped clustered.
PBM (↑)	The PBM index is based on distance between the points and their centre in clusters and the distances between the centres of clusters.	Distances among the clusters are not uniform/same in ST-Data, it depends on representative points of clusters which is not fixed in arbitrary shaped clustered.

Ray Turi (↓)	It is ratio of within group dispersion and minimum of the squared distances between all the cluster centres.	It is used to identify well compact clusters which are available at unique/same distances but in case of ST-data, is not possible.
Ratkowsky Lance (↑)	It is the mean of ratio of between- -cluster dispersion and total dispersion for each dimension of data.	As in case ST-Data, the distances among the clusters and size of clusters is not uniform/same.
Scott Symons (↓)	It works weighted sum of the logarithms of the determinants of the variance-covariance matrix of each clusters.	As logarithms of the determinants of the variance-covariance matrix of few clusters may be zero/close to zero for densed ST-data.
SD(SD_Scat & SD_Dis) (↓)	It is summation of two quantities i.e. average scattering for clusters and the total separation between clusters.	It is not able to handle arbitrary shaped clusters (Halkidi and Vazirgiannis).
S_Dbw (↓)	It is sum of the mean dispersion in the clusters and of the between cluster density.	It is not able to handle arbitrary shaped clusters (Halkidi and Vazirgiannis).
Silhouette (↑)	Used for graphical display where each cluster is represented by a silhouette, which is based on the comparison of its tightness and separation.	Used for graphical display, suitable for partitioning techniques, so worth using this index for spherical shape clusters only and not for arbitrary shaped clusters which can be detected by Density based algorithm(Rousseeuw).

Table 4.2 Continued

Index	Significance of Index	Remarks on Non-applicability
Trace_WiB (↑)	It is ratio of between cluster dispersion and within cluster dispersion.	Distances among the clusters are not uniform/same in ST-Data, it depends on representative points of clusters which is not fixed in arbitrary shaped clustered.
WemmertGancarski (↑)	It is described using quotients of distances between the points and the centres of all the clusters.	It depends on representative points of clusters which is not fixed in arbitrary shaped clustered.
XieBeni (↓)	The XieBeni index is based on concept of compactness and separation. It is ratio of within group dispersion and minimum of the squared distances between the points in the all the cluster.	Distances among the clusters are not uniform/same in ST-Data.

### Practical Approach (KP, Sanjay and Pinkal):

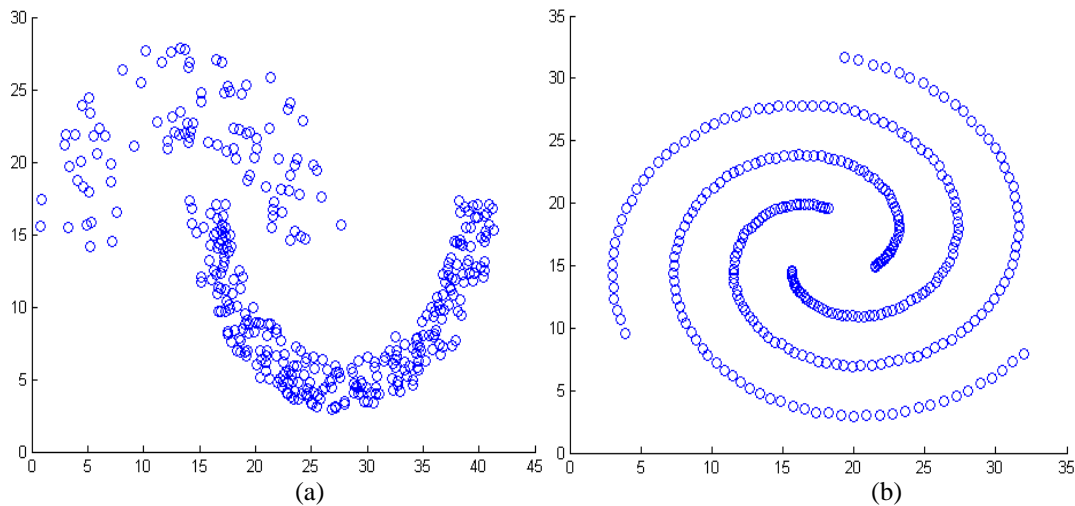
To prove theoretical approach, experimentation has been performed on classical dataset (i.e. having densed, sparse and arbitrary shaped clusters) with all internal indices described in table 4.1 and table 4.2. This enables to use the same set of selected validation indices to use them for validating clustering algorithms using ST dataset which is one of the objective.

Steps for performing practical approach:

- 1) Select classical datasets having dense, sparse and arbitrary shape in nature.
- 2) Select two or more clustering algorithms in which one do not identify correct clustering structure while other clustering algorithms can.
- 3) Select indices that supports correct clustering algorithm.
- 4) Compare the indices obtained from practical approach with theoretical approach.
- 5) Select appropriate indices which identify correct clustering structure for all given classical dataset(s).

For performing experimentation, three different classical datasets i.e. Jain, Spiral and Compound(Clustering Datasets) have been used as shown in figure 4.2 having following nature:

1. Jain Dataset with 02 cluster and its nature (dense and any shape) is shown in figure 4.2(a).
2. Spiral Dataset with 03 cluster and its nature (any shape) is shown in figure 4.2(b).
3. Compound Dataset with 06 cluster and its nature (dense, sparse, ring and any shape with outliers) is shown in figure 4.2(c).



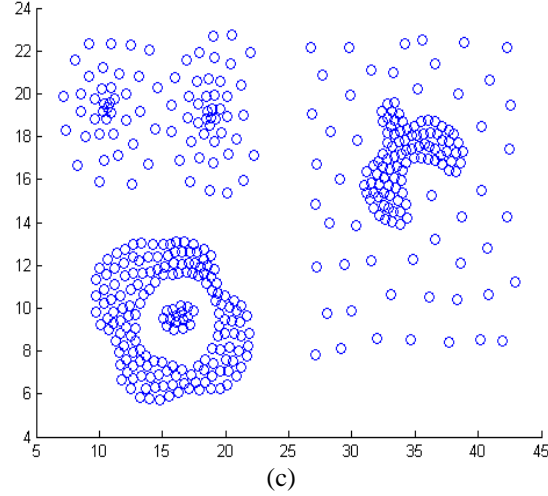


Figure 4.2: Classical Datasets: (a) Jain Dataset (b) Spiral Dataset and (c) Compound Dataset

## Results and Discussion

For classical dataset, following clustering algorithms are used:

1. k-means (Partitional Algorithm)
2. DBSCAN (Density based algorithm)

Reason for selecting these two algorithms is that, as it is known that k-means algorithm is able to detect only spherical shaped clusters while DBSCAN algorithm is capable to detect arbitrary shaped clusters also besides spherical shaped clusters. This enables one to compare that which of validation indices are performing well for only spherical shaped clusters and which of them are doing well for arbitrary shaped clusters, as major objective is to validate clusters having arbitrary shape as focus is given on ST data.

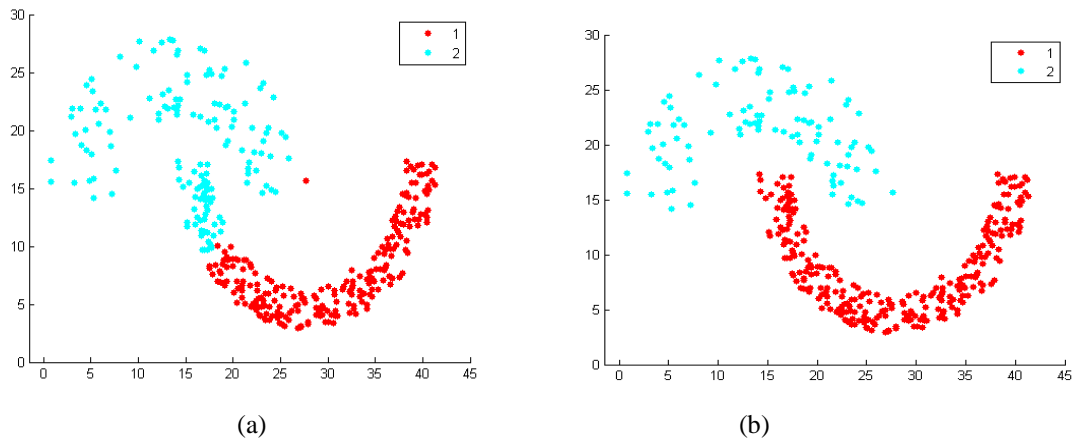


Figure 4.3: Results of (a) k-means (b) DBSCAN on Jain dataset

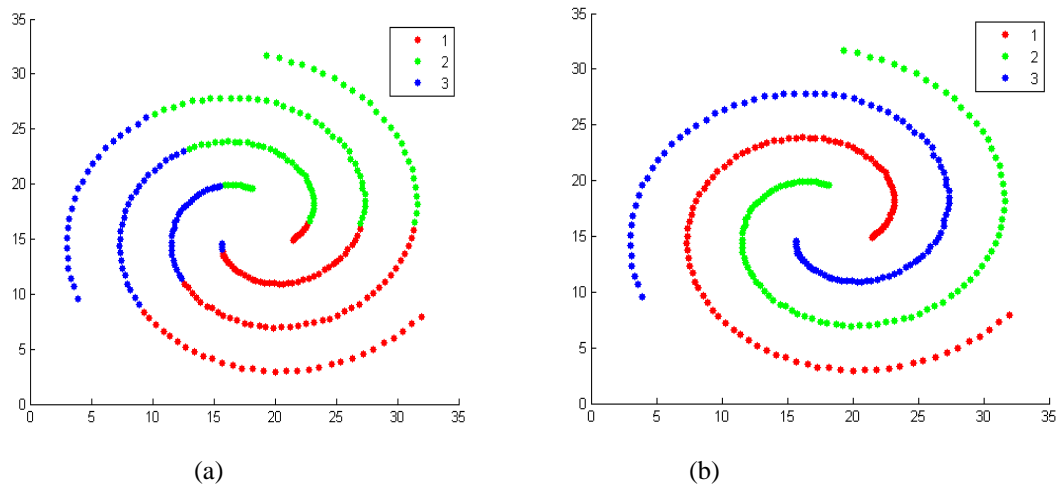


Figure 4.4:Results of (a) k-means (b) DBSCAN on Spiral dataset

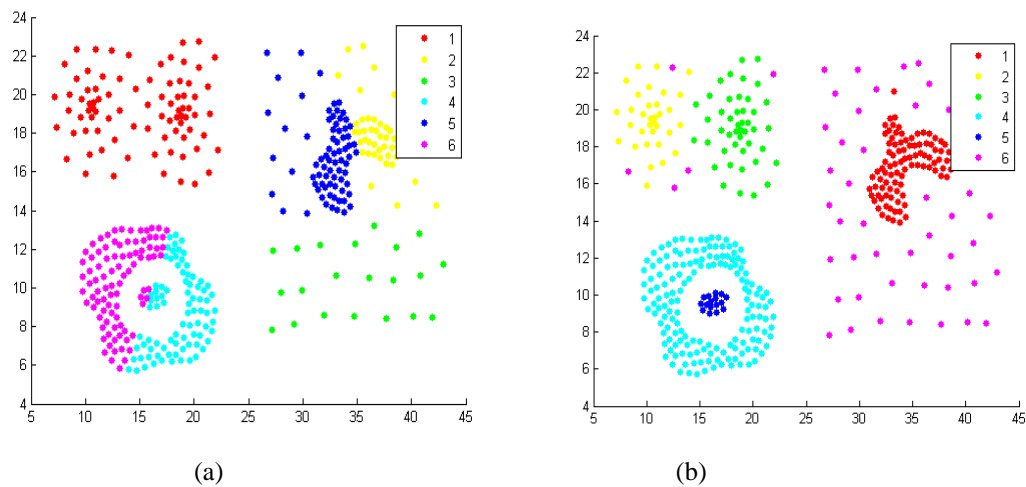


Figure 4.5:Results of (a) k-means (b) DBSCAN on Compound dataset

Results of both these algorithms for Jain, Spiral and Compound datasets have been shown in figure 4.3, 4.4 and 4.5 respectively. From figures 4.3, 4.4, 4.5,it is clear that the results (clustering structure) of all three datasets using DBSCAN algorithm are much better than k-means algorithm.

Table 4.3:Cluster validation indices using k-means and DBSCAN on classical datasets<sup>2</sup>

Indices	Rules	Jain		Spiral		Compound	
		DBSCAN	K-means	DBSCAN	K-means	DBSCAN	K-means
ball_hall	max	<b>74.803333</b>	59.8308923	<b>96.53504591</b>	39.38572141	<b>20.1677008</b>	9.133920716
banfeld_raftery	min	1631.7452	1522.939249	1425.656936	1146.05003	982.782114	941.1428588
c_index	min	0.1807141	0.119262481	0.458272443	0.136240372	0.06438973	0.049867022

<sup>2</sup>K. P. Agrawal, Sanjay Garg and Pinkal Patel, *Performance Measures for Densded and Arbitrary Shaped Clusters*, International Journal of Computer Science & Communication, Volume 6(2), pp. 338-350, Serial Publication, 2015, ISSN: 0973-7391.

calinski_harabasz	max	279.48312	503.4754674	5.797852185	238.3129233	361.906202	679.0463101
davies_bouldin	min	0.8996695	0.78307242	5.948939631	0.894538706	4.56622113	0.34420882
<b>det_ratio</b>	min	<b>3.3460281</b>	4.082517028	<b>1.076389026</b>	6.444023922	<b>21.3647099</b>	51.85237085
<b>Dunn</b>	max	<b>0.0924237</b>	0.018699307	<b>0.141071797</b>	0.006667315	<b>0.04310513</b>	0.023997109
<b>Gamma</b>	max	<b>-0.598760</b>	-0.734569132	<b>-0.01336378</b>	-0.693398654	<b>-0.84863883</b>	-0.8798154
<b>g_plus</b>	min	<b>0.3788764</b>	0.433648422	<b>0.224517184</b>	0.375394933	0.34545416	0.335967732
<b>gdi11</b>	max	<b>0.0924237</b>	0.018699307	<b>0.141071797</b>	0.006667315	<b>0.04310513</b>	0.023997109
<b>gdi12</b>	max	<b>0.4459971</b>	0.093651276	<b>0.584360131</b>	0.035059924	<b>0.26333017</b>	0.165633773
<b>gdi13</b>	max	<b>0.1447294</b>	0.032165271	<b>0.19145759</b>	0.012062476	<b>33</b>	0.059715173
gdi21	max	1.4893557	1.607545413	1.106082483	1.35576615	0.22369928	0.503744322
gdi22	max	7.1869919	8.051029809	4.581713126	7.129265458	1.36658381	3.47696349
gdi23	max	2.3322325	2.765189793	1.50113553	2.452845937	0.47747084	1.253533449
gdi31	max	0.7372479	0.783392929	0.494695097	0.692602063	0.11459354	0.24898925
gdi32	max	3.5576427	3.923447369	2.049169979	3.642032194	0.70005441	1.718583206
gdi33	max	1.1544815	1.347538995	0.671382467	1.253052495	0.24459208	0.619592797
gdi41	max	0.6501577	0.713822242	0.121020417	0.610918185	0.01683587	0.211982271
gdi42	max	3.1373824	3.575018227	0.501301522	3.212499377	0.10285071	1.463152208
gdi43	max	1.0181040	1.227868254	0.164244575	1.105270395	0.03593502	0.527503449
<b>gdi51</b>	max	<b>0.3053413</b>	0.278857359	<b>0.359959904</b>	0.273222817	0.04296875	0.099740607
<b>gdi52</b>	max	<b>1.4734464</b>	1.396594391	<b>1.491057895</b>	1.43673597	0.26249705	0.688433464
gdi53	max	0.4781443	0.479671377	0.48852469	0.494313476	0.09171387	0.24819771
<b>ksq_detw</b>	max	<b>56398492</b>	462241649.1	<b>2032325883</b>	339473177.6	<b>511682683</b>	210828394.5
<b>log_det_ratio</b>	min	<b>450.49970</b>	524.7042164	<b>22.96692673</b>	581.3037914	<b>1221.63446</b>	1575.411863
<b>log_ss_ratio</b>	min	<b>-0.283260</b>	0.305332924	<b>-3.28270656</b>	0.433390518	<b>1.52701337</b>	2.156317629
mcclain_rao	min	0.5546708	0.476216739	0.957905994	0.528246482	0.31594955	0.289991739
Pbm	max	135.99496	196.5497618	1.290347306	56.3356262	123.49936	202.3043751
<b>point_biserial</b>	max	<b>-4.351642</b>	-5.175361703	<b>-0.256728239</b>	-3.364320447	<b>-4.69923163</b>	-4.74318189
ray_turi	min	0.2554356	0.183635578	9.747806492	0.234527926	59.8961278	0.732054837
ratkowsky_lance	max	0.4932935	0.517825311	0.10974467	0.449354775	0.36192205	0.373282454
scott_symons	min	2574.3994	2302.89743	2405.707588	1828.565898	1340.5641	1316.453629
sd_scatt	min	0.5821945	0.412999618	0.972330935	0.401221253	0.16003072	0.067195479
sd_dis	min	0.1129814	0.111070935	0.500501531	0.120898548	3.6382463	0.606202327
s_dbw	min	1.6638272	0.59332748	3.981600292	3.323213794	<b>2.35446996</b>	4.560024802
Silhouette	max	0.42473838	0.493174045	0.001217834	0.360534171	0.33772354	0.424481508
<b>Tau</b>	max	<b>-0.412215</b>	-0.519422468	<b>-0.008895828</b>	-0.461702874	<b>-0.51881178</b>	-0.52601785
<b>trace_w</b>	max	<b>29856.328</b>	22208.78484	<b>30109.35035</b>	12286.92822	<b>8330.50162</b>	4843.470606
trace_wib	max	2.3460281	3.082517028	0.074988889	3.081063342	8.48908363	17.01666466
wemmert_gancarski	max	0.415748	0.601653787	0	0.488620804	0.14250205	0.607735727
xie_beni	min	12.640154	267.5999017	7.173709574	1969.059009	9.13717732	57.12481918

Results of all 27 indices are shown in table 4.3 for all three classical datasets with k-means and DBSCAN algorithms. As we know that DBSCAN algorithm is more suitable compared to k-means for obtaining dense and arbitrary shaped clusters which have been shown in figure 4.3, figure 4.4 and figure 4.5. Column 2 i.e. ‘Rules’ in table 4.3 shows the rules (“max” or “min”) which suggests that results given by two chosen algorithms are compared and whichever algorithm is giving larger value is considered better than the other algorithm if rule is ‘max’ and otherwise if rule is

‘min’. Therefore, the indices giving better results for DBSCAN (i.e. which are suitable) have been marked in bold as shown in table 4.3, this enable us to select set of good indices. Indices which supports dense, sparse, ring and arbitrary shaped clustering structure given by DBSCAN are as follows (i.e. total 12): Ball-Hall, Det\_Ratio, Dunn, Gamma, G\_plus, GDI (gdi11, 12, 13, 51, 52), Ksq\_DetW, Log\_Det\_Ratio, Log\_SS\_Ratio, Point Biserial, Tau and Trace\_W.

Based on the results obtained, these indices have been used for performing experimentation using ST-Dataset for validation purpose (as shown in table 4.7 in section 4.2.4).

## 4.2 Experimental Analysis

In order to achieve research objective i.e. “To develop new Clustering Technique with special emphasis on Spatio-Temporal Geographic Dataset”. Following steps have been taken into considerations :

- 1) From literature survey as given in section 3.1.1 and table 3.1 and table 3.2, after required pre-processing/cleaning of dataset, ST-DBSCAN (Birant and Kut) algorithm is applied, results are analyzed.
- 2) It is followed by application of developed ST-OPTICS algorithm on same dataset, results are obtained, analyzed and compared with the results obtained from step 1. Results obtained from newly developed technique are appealing (section 4.3.4)
- 3) Finally, result are validated (section 4.3.4)

### 4.2.1 Experimental Setup

The "R" language/environment and Quantum Geographic Information System (QGIS) are the tool being open source have been used for experimentation purpose for obtaining the results, as They supports all existing clustering algorithms, provides facilities for coding our own algorithm i.e. ST-OPTICS and visualization of clusters obtained respectively.

### 4.2.2. Data Specifications

Dataset for experimentation is provided by Space Application Centre (SAC, ISRO) Ahmedabad.

Table 4.4: Dataset Specification.

Satellite	:	MODIS (Moderate Resolution Imaging Sepctroradiometer)
Measure	:	NDVI (Normalized Difference Vegetation Index) (16 days composite)



Area	:	Entire Country (India)
Grid Size	:	5*5 km
Total No of Grids	:	1,30,307 (includes different states and their districts)

Before experimentation, cleaning of data have been taken care of (i.e. values which are negative or zero continuously at many locations). In dataset (specification given in table 4.4) which has been used for experimentation having latitude and longitude are spatial, days, months & year are temporal and NDVI are the non-spatial dimensions. Dataset has the columns: Grid code for each grid, Latitude and Longitude of the grid, State, City and 23 columns for NDVI values, where each column contains 16 days composite NDVI values.

### 4.2.3 Performance Parameters and Validation Indices

In order to measure the performance of clustering algorithm, care has been taken to improve the run time efficiency, quality of clusters obtained and handling high dimensional ST Data along handling nesting of clusters. Moreover, different validation indices have been used for validating arbitrary shaped and densed clusters with special emphasis to ST data. This has been discussed in section 4.1 in detail.

### 4.2.4 Results

Initially, data have been chosen related to Gujarat state which is containing 7028 grids for obtaining clusters (with  $\epsilon_1=0.2022$  and  $\epsilon_2=0.6561$  calculated from k-dist graph) using both the algorithms ST-OPTICS and ST-DBSCAN. The results are shown in table 4.5 and its geographical visualization is also shown in figure 4.6.

Table 4.5: Comparison - ST-OPTICS and ST-DBSCAN for Gujarat state.

State	Algorithm	Number of Cluster	Time in Sec
Gujarat	ST-OPTICS	1751	60.38
	ST-DBSCAN	280	1797.03

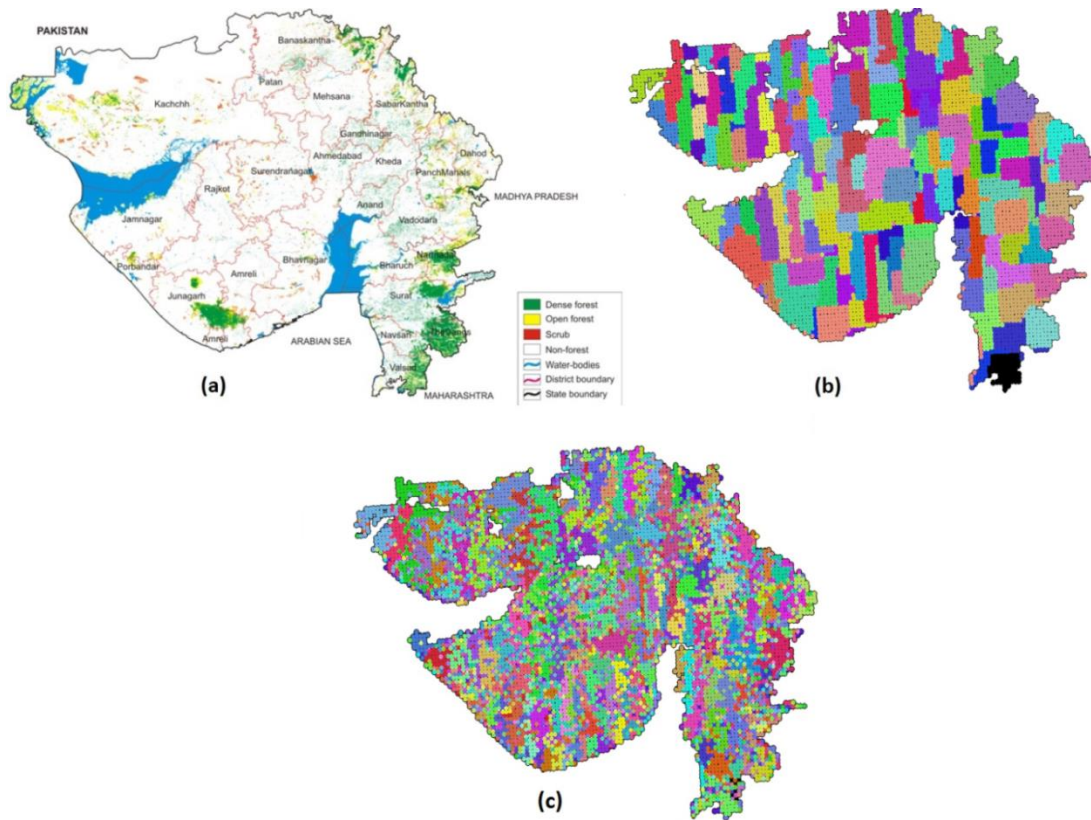


Figure 4.6: Clustered Maps (a) Forest survey map (b) Clusters given by ST-DBSCAN (c) Clusters given by ST-OPTICS.

Different colors in figure 4.6 shows different clusters. In case of clustered map given by ST-OPTICS algorithm (figure 4.6), owing to very large no. of clusters obtained due to clustering at micro level, it is very difficult to interpret them properly. Consequently, sincere attempt has been made to explore the possibility of merging the clusters using hybrid approach (figure 4.7), where it is found that the performance of ST-OPTICS is better than ST-DBSCAN.

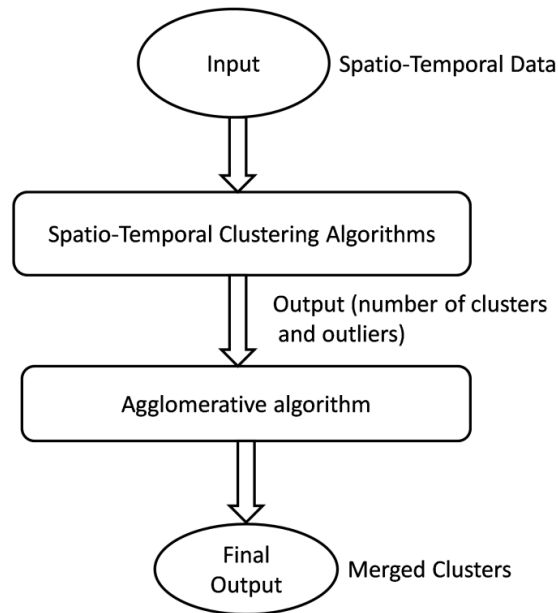


Figure 4.7: Hybrid Approach.

Flow chart for hybrid approach has been shown in figure 4.7 which is the hybridization of two categories of clustering algorithms i.e. Density and Hierarchical based for getting merged cluster so that better analysis, visualization and interpretation of obtained clusters can be done effectively. Results generated by both the spatio-temporal clustering algorithms (i.e. ST-OPTICS and ST-DBSCAN) are given as input to Agglomerative algorithm where the required no. of natural clusters were obtained from the guidelines about crops available from the website of agriculture department of Gujarat (Gujarat State Seeds Corporation Limited) (Ministry of Statistics and Implementation) and final results obtained are shown in figure 4.8.

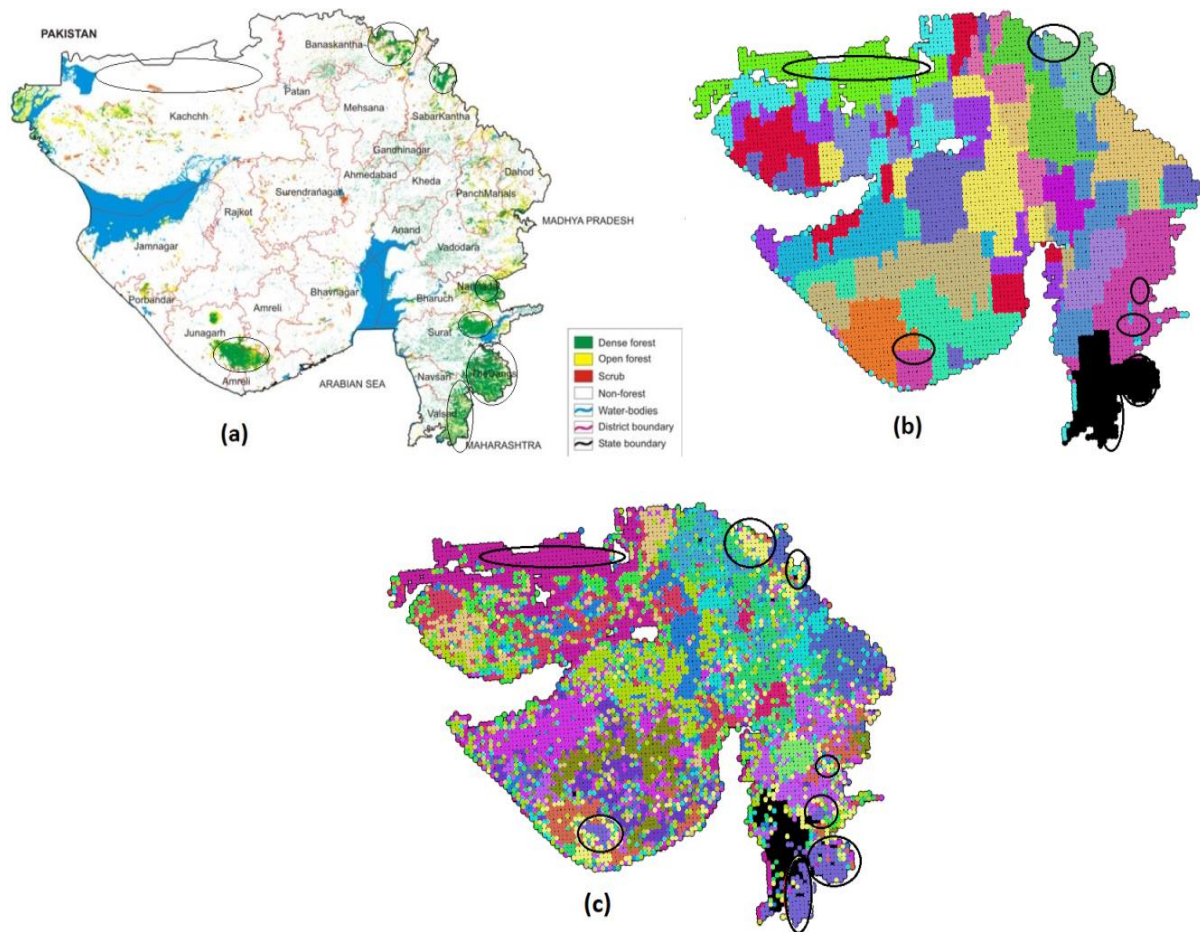


Figure 4.8: Comparison of Result for validation-Hybrid approach (a) Forest survey map (b) Result of ST-DBSCAN (c) Result of ST-OPTICS.

### Interpretation of Results:

The interpretation of results obtained (figure 4.8) is given in table 4.6 below.

Table 4.6: Interpretation – Results obtained

Figure	Title	Interpretation	Remarks
figure4.8 (a)	Standard map of Gujarat state given by FSI for comparison purpose.	Clusters about presence of dense vegetation are encircled. Empty ellipse indicates no vegetation.	Clusters shown in blue color are the water bodies.

figure 4.8 (b)	Clusters given by ST-DBSCAN.	i) Cluster shown in black colour cover the areas like ‘The Dangs’, Valsad and ‘Navsari’ and others areas also in vicinity, not matching with figure 4.8 (a). ii) Similary Area like ‘Gir Forest of Junagarh’ covering three areas of vicinities, not matching with figure 4.8 (a). iii) Similar observations are visible for other encircled areas.	Not providing good quality clusters.
figure 4.8 (c)	Clusters given by ST-OPTICS.	i) Cluster shown in black colour cover the areas like ‘The Dangs’, ‘Valsad’ only and put in same cluster, which match exactly with figure 4.8 (a). ii) Cluster given cover only Area like ‘Gir Forest of Junagarh’, which match exactly with figure 4.8 (a). iii) Similar observations are visible for other encircled areas.	Providing good quality clusters.

For Validation of obtained clusters, following two approaches have been used:

1. By Visualization
2. By Performance Indices

## Validation By Visualization:

Validation of the obtained clusters (figure 4.8b and figure 4.8c) have been performed with map provided on Forest Survey of India’s [FSI, an organization under the Ministry of Environment & Forests, Government of India] website (Forest Survey of India) and the information about major crops in corresponding areas are taken from (Gujarat State Seeds Corporation Limited) (Ministry of Statistics and Implementation) and it is found that result given by ST-OPTICS algorithm is better.

For further verification and validation, similar experimentations for different states (Himachal Pradesh, Arunachal Pradesh, Chhattisgarh, and Karnataka) from the different corner of country ‘India’ are performed. The results are shown in table 4.7 and also geographic visualization of these results (figure 4.9) are analyzed and validated. For different states, different areas such as in

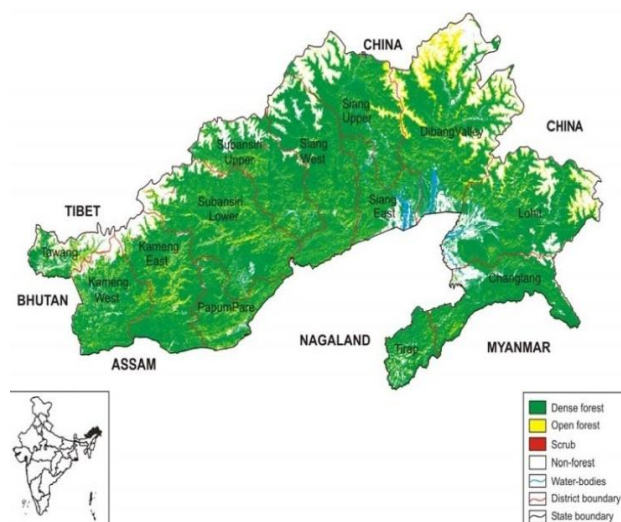
Arunachal Pradesh (areas like Lohit, Dibangvalley, Tawang, Chandlang and Siang East etc), in Chhattisgarh (areas like Surguja, Jashpur, Mahasamund, Baster etc.), in Karnataka (areas like Bidar, Raichur, Udupi, Dakshin Kannada, Chamrajnagar etc.) and in Himachal Pradesh (areas like Lahul & Spiti, Kinnaur, Una, Hamirpur, Mandi etc.) are analyzed for NO, LOW and HIGH vegetation and it is obvious that ST-OPTICS is giving better cluster formation compared to ST-DBSCAN.

Table 4.7: Comparison table showing results using ST-OPTICS and ST-DBSCAN with hybrid approach

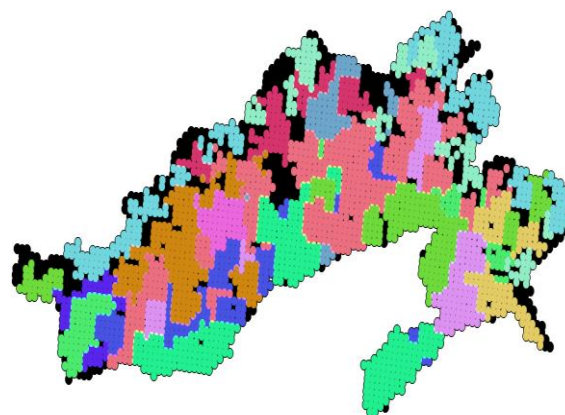
Sr. No	State	Number of Grids	Algorithm	Number of Cluster without using Hybrid approach	*Number of Cluster using Hybrid approach	Time (in sec) without using Hybrid approach	Time (in sec) taken by Agglomerative algorithm	Total Time in sec
1	Gujarat	7028	ST-OPTICS	1751	20	60.38	1.77	62.15
			ST-DBSCAN	280		1797.0	0.28	1797.3
2	Himachal Pradesh	2186	ST-OPTICS	531	10	8.53	0.47	9.0
			ST-DBSCAN	167		131.10	0.14	131.24
3	Arunachal Pradesh	3176	ST-OPTICS	1263	15	13.95	1.18	15.13
			ST-DBSCAN	196		106.41	0.19	106.60
4	Chhattisgarh	5212	ST-OPTICS	1473	10	35.59	1.43	37.02
			ST-DBSCAN	341		801.83	0.33	802.16
5	Karnataka	7581	ST-OPTICS	2394	15	65.13	2.53	67.66
			ST-DBSCAN	506		1031.3	0.5	1031.8

\*Indicates major crops in particular area have been taken from FSI and agriculture department of concerned state.

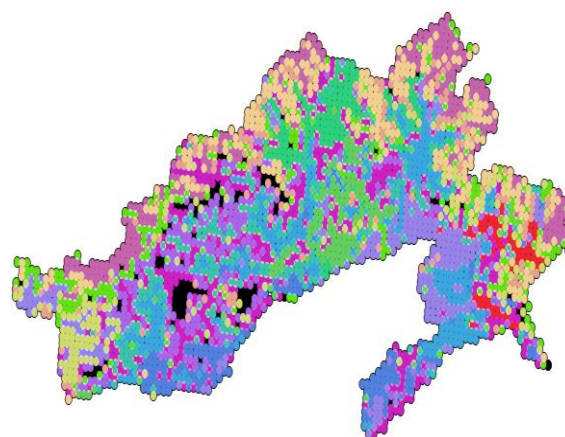
Clustered maps of Arunachal Pradesh, Chhattisgarh, Karnataka and Himachal Pradesh for forest survey map are given in fig 4.9 (a),(d),(g) and (j), results of ST-DBSCAN shown in figure(b), (e),(h) and (k) and results of ST-OPTICS shown in figure(c),(f),(i) and (l).



(a)



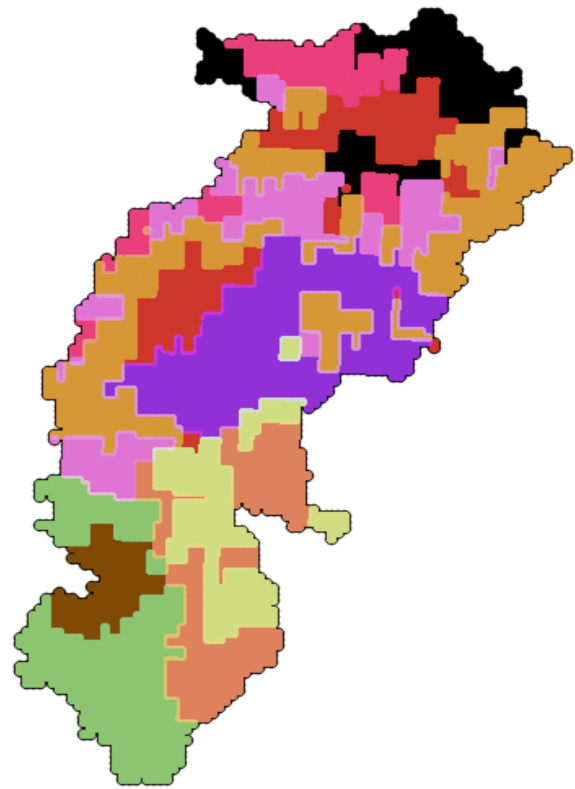
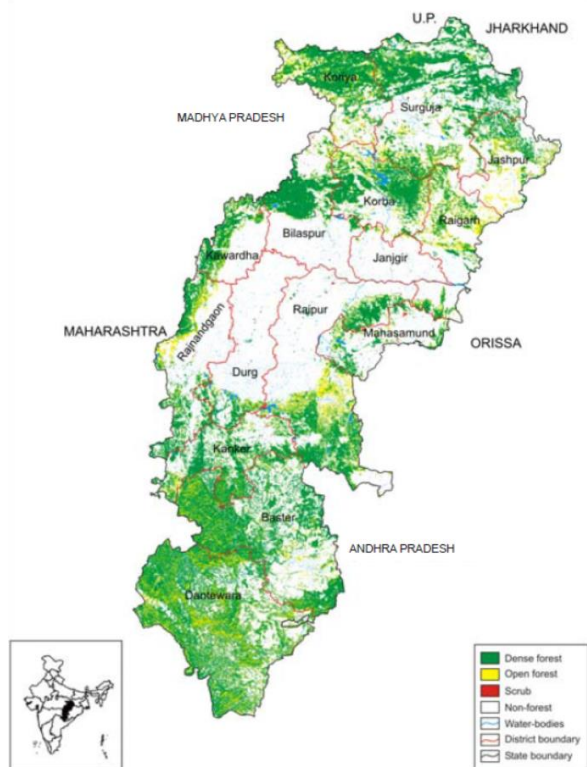
(b)



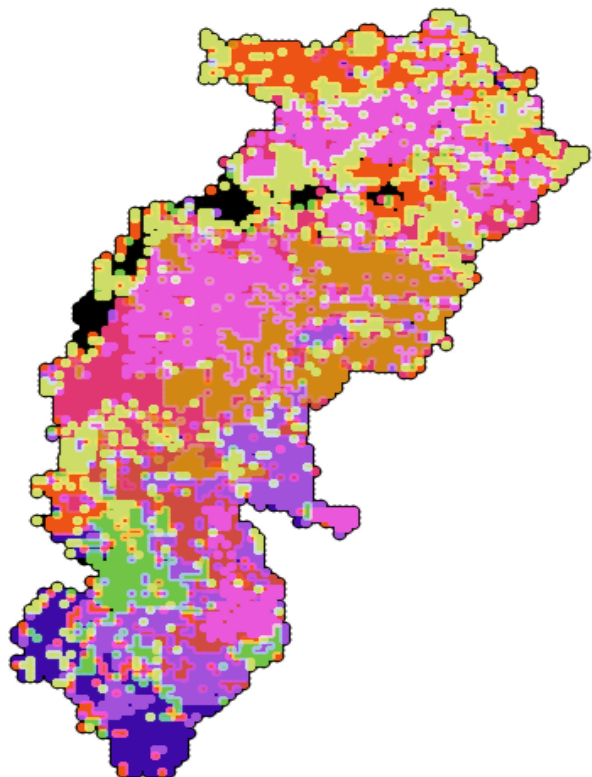
(c)

Figure 4.9 Continued.





(e)

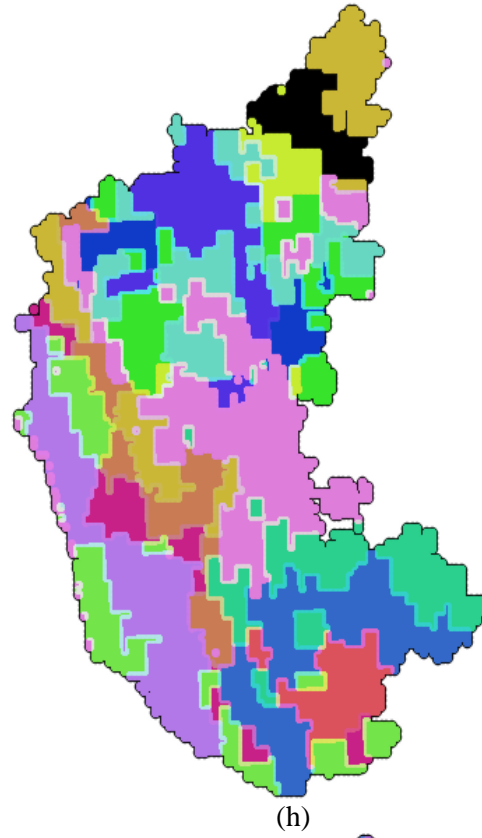
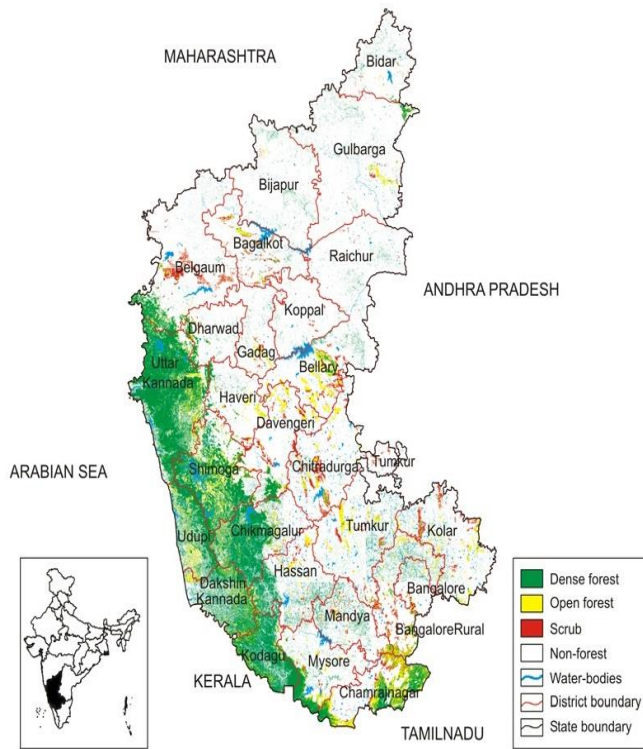




(d)

(f)

Figure 4.9 Continued.



(g)

(i)

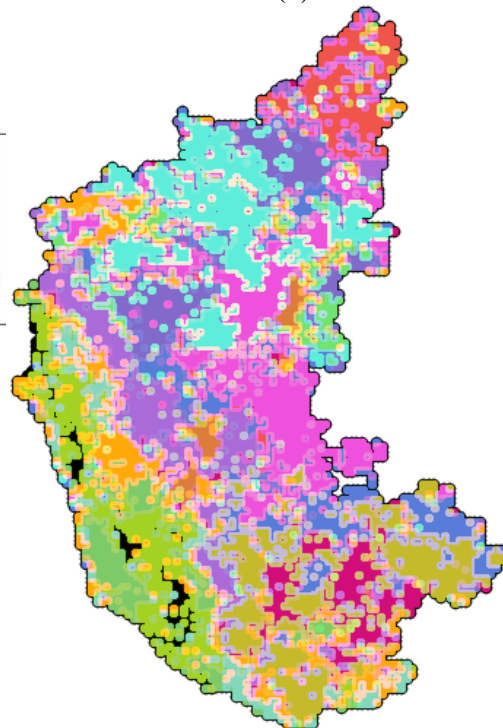


Figure 4.9 Continued.

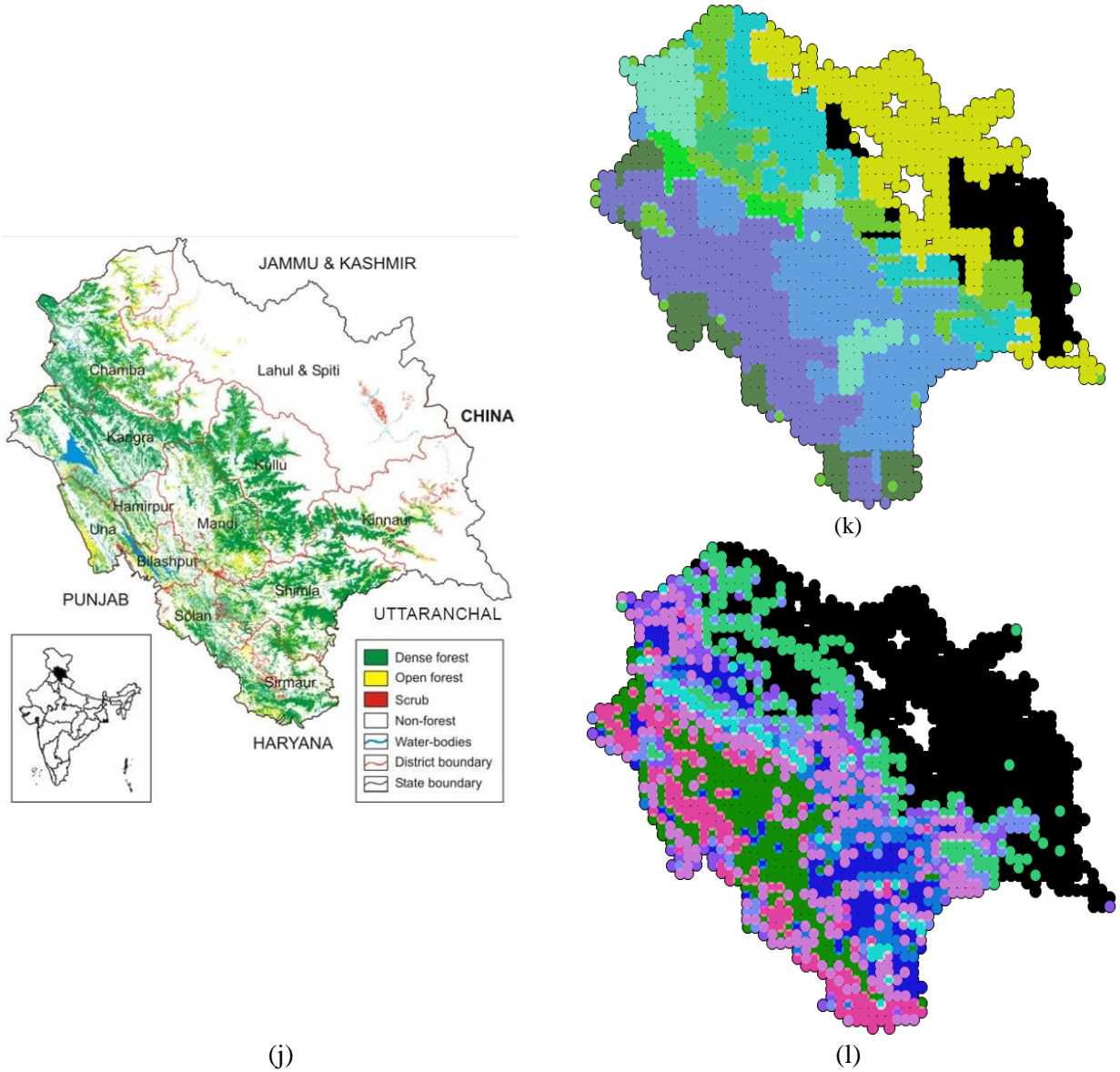


Figure 4.9: Clustered maps of Arunachal Pradesh, Chhattisgarh, Karnataka and Himachal Pradesh for forest survey map shown in (a),(d),(g) and (j), results of ST-DBSCAN shown in (b), (e),(h) and (k) and results of ST-OPTICS shown in (c),(f),(i) and (l).

## Validation By Performance Indices:

Results of all 12 indices (i.e. selected through theoretical and practical approaches as given in table 4.1 and table 4.2) are shown in table 4.8 for Gujarat dataset with ST-OPTICS and ST-DBSCAN algorithms. Indices which supports to results of ST-OPTICS are shown in bold letter in table 4.8.

Most of these indices supported to results of ST-OPTICS compared to the results of ST-DBSCAN except ksq\_detw. The results in table 4.8 clearly convey that the ST-OPTICS algorithm is performing much better than ST-DBSCAN algorithm.

Table 4.8: Results - Validation Indices using a) ST-OPTICS and b) ST-DBSCAN (Gujarat State).

Indices	Indices-Value	
	STOPTICS	STDBSCAN
Ball-Hall (↑)	<b>1.825358</b>	0.94273431
Det_Ratio(↓)	<b>120.5571</b>	1675.54768
Dunn (↑)	<b>0.007791</b>	0.00746244
Gamma (↑)	<b>-0.40856</b>	-0.63646775
G + (↓)	<b>0.081162</b>	0.08599348
GDI31 (↑)	<b>0.168413</b>	0.1607403
GDI51 (↑)	<b>0.075723</b>	0.06623729
Ksq_DetW (↑)	-1.23E+23	<b>-8.82E+21</b>
Log_Det_Ratio (↓)	<b>33679.05</b>	52175.1366
Point Biserial (↑)	<b>-0.18916</b>	-0.28280074
Tau (↑)	<b>-0.1387</b>	-0.20633383
Trace_W (↑)	<b>13770.22</b>	7761.81158

Further verification and validation of spatio-temporal clustering technique by selected indices, same technique has been also applied on different states (i.e. Himachal Pradesh, Arunachal Pradesh, Chhattisgarh and Karnataka) (table 4.9 and table 4.10) and got good results i.e. selected indices are proper and conveys that ST-OPTICS is performing well compared to existing one.

Table 4.9: Cluster validation indices' results of ST-OPTICS and ST-DBSCAN for the Himachal Pradesh and Arunachal Pradesh

Indices	Indices - Value for different States			
	Himachal Pradesh		Arunachal Pradesh	
	STOPTICS	STDBSCAN	STOPTICS	STDBSCAN
Ball-Hall (↑)	<b>0.925945</b>	0.7430249	<b>2.162726</b>	1.6145338
Det_Ratio(↓)	-369.525	<b>-616.1915</b>	<b>184.3184</b>	199.76345
Dunn (↑)	<b>0.014378</b>	0.0109474	<b>0.015736</b>	0.0122538
Gamma (↑)	<b>-0.70101</b>	-0.727297	<b>-0.36166</b>	-0.405789
G + (↓)	0.23715	<b>0.2012307</b>	<b>0.096006</b>	0.1238663
GDI31 (↑)	<b>0.258908</b>	0.242937	<b>0.199267</b>	0.1715524
GDI51 (↑)	<b>0.15014</b>	0.1220003	<b>0.089935</b>	0.0839707
Ksq_DetW (↑)	<b>3.11E+17</b>	1.864E+17	<b>2.66E+35</b>	2.45E+35
Log_Det_Ratio (↓)	NaN	NaN	<b>16547.26</b>	16802.509

Point Biserial ( $\uparrow$ )	-0.4296	<b>-0.407758</b>	<b>-0.1756</b>	-0.216393
Tau ( $\uparrow$ )	-0.37017	<b>-0.351067</b>	-0.13581	<b>-0.170346</b>
Trace_W ( $\uparrow$ )	<b>1733.426</b>	1485.1937	<b>7316.808</b>	6319.0357

Table 4.10: Cluster validation indices' results of ST-OPTICS and ST-DBSCAN for the Chhattisgarh and Karnataka

Indices	Indices - Value for different States			
	Chhattisgarh		Karnataka	
	STOPTICS	STDBSCAN	STOPTICS	STDBSCAN
Ball-Hall ( $\uparrow$ )	<b>1.97964</b>	0.9912552	<b>2.044328</b>	1.2313262
Det_Ratio( $\downarrow$ )	<b>-32.1562</b>	284.17864	<b>121.4647</b>	590.85344
Dunn ( $\uparrow$ )	0.010005	0.0145338	0.008056	0.0114849
Gamma ( $\uparrow$ )	<b>-0.32447</b>	-0.58228	<b>-0.42892</b>	-0.639471
G + ( $\downarrow$ )	<b>0.139277</b>	0.1640892	<b>0.108234</b>	0.123474
GDI31 ( $\uparrow$ )	0.234173	<b>0.246309</b>	<b>0.201978</b>	0.170098
GDI51 ( $\uparrow$ )	0.123677	0.135355	<b>0.112779</b>	0.0765183
Ksq_DetW ( $\uparrow$ )	<b>6.95E+23</b>	-7.86E+22	-9.24E+31	-1.90E+31
Log_Det_Ratio ( $\downarrow$ )	NaN	29440.081	<b>36381.15</b>	48372.285
Point Biserial ( $\uparrow$ )	<b>-0.2127</b>	-0.381948	<b>-0.24254</b>	-0.360621
Tau ( $\uparrow$ )	<b>-0.1488</b>	-0.265183	<b>-0.16694</b>	-0.248183
Trace_W ( $\uparrow$ )	<b>11430.14</b>	5719.3419	<b>15975.37</b>	9672.041

## Run Time Performance:

The Run time efficiency of ST-OPTICS and its comparison with ST-DBSCAN algorithm are shown in Table 4.6 and corresponding histogram has been shown in figure 4.10.

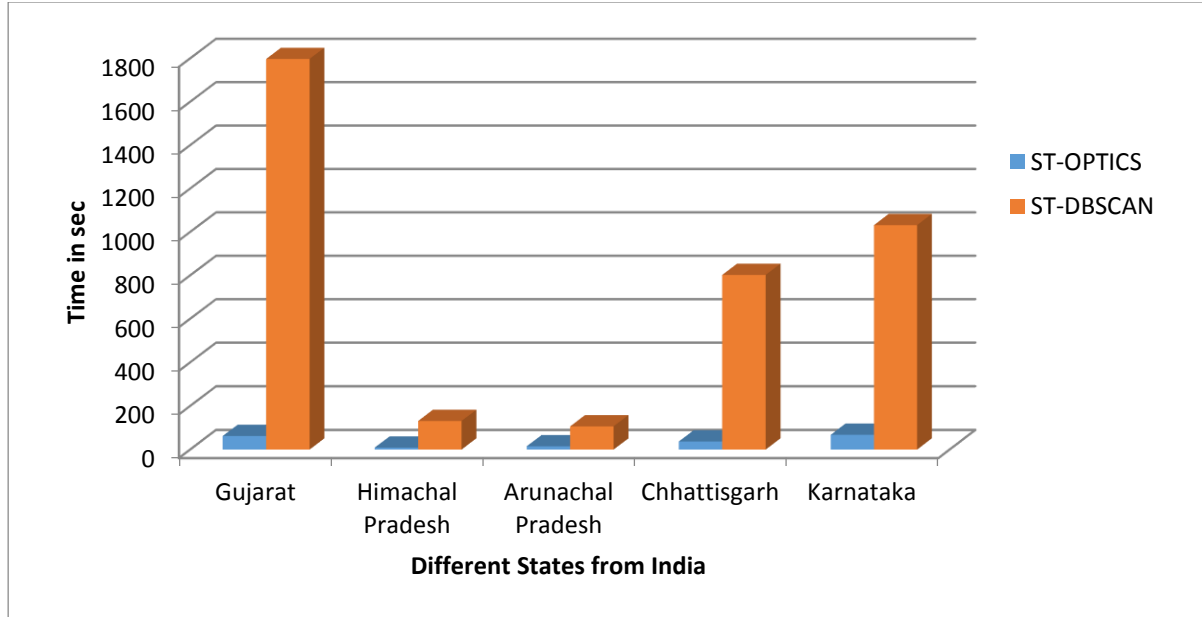


Figure 4.10: Performance Comparison (Average Runtime) of ST-OPTICS and ST-DBSCAN.

Proposed technique ST-OPTICS gives similar average runtime complexity as that of OPTICS algorithm i.e.  $O(n \cdot \log n)$  (Ankerst et al.), where  $n$  is the number of objects in the dataset. Modifications made in existing OPTICS algorithm do not affect the runtime complexity of proposed approach; however ST-OPTICS deals with spatio-temporal data while OPTICS cannot. Moreover, average runtime complexity of ST-DBSCAN algorithm is  $O(n \cdot \log n)$  (Birant and Kut). Both algorithms (i.e. ST-DBSCAN and ST-OPTICS) have same asymptotic time complexity but, while perform computation, ST-DBSCAN took more time compared to ST-OPTICS algorithm for every case as shown in table 4.6. Analysis reveals two reasons behind this: first, in ST-DBSCAN algorithm, it always checks the condition (no. of neighbours  $\geq$  Minpts) after finding the neighbours and if condition is not true, it repeats main loop, so inner computations also increases. In ST-OPTICS algorithm, it doesn't checks this condition so iteration in main loop reduces. Secondly, the condition  $(|Cluster\_Avg() - o.Value| \leq \Delta\epsilon)$  that is computed in inner loop of ST-DBSCAN algorithm takes more time, however in ST-OPTICS before testing of condition, objects are already ordered.

Outlier detection is very important functionality of data mining, it has enormous applications (discussed in section 2.1.10). Looking to very wider application base of this data mining tasks, parallel work has been carried out to detect outliers on same ST dataset using Shared Nearest

Neighbor (SNN)(Jarvis and Patrick) (Ertoz, Steinbach, and Kumar) clustering approach, but this existing technique is not able to handle ST data, attempt has been made to modify this existing approach to suite to the requirement. This newly developed technique named as Spatio-Temporal Shared Nearest Neighbor (ST-SNN) is capable to handle high dimensional spatio-temporal data having different densities and sizes and also capable to identify arbitrary shaped cluster.

### **ST-SNN Approach to detect ST-Outliers**

It uses three step approach to detect spatio-temporal outliers.

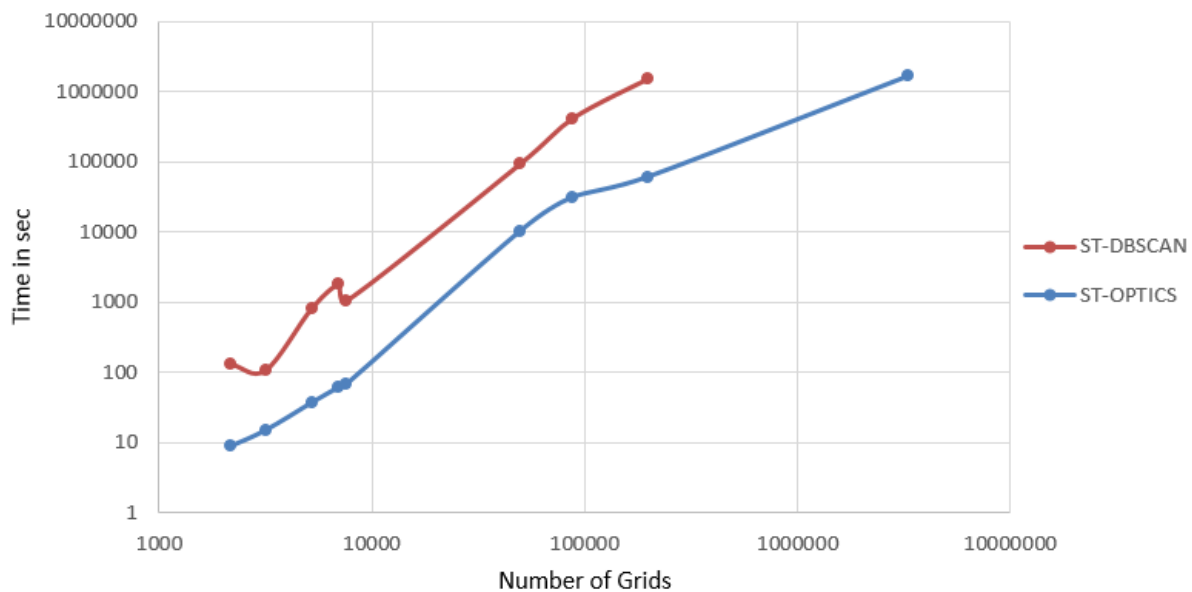
- i) In the first step of outlier detection, clustering is performed on the spatio-temporal dataset using modified clustering approach. First finds nearest neighbors of each data points, secondly it finds the shared nearest neighbor similarity between pair of points in terms of how many nearest neighbors the two points share. Using this similarity measure, algorithm identifies core points and build clusters around the core points.
- ii) In the second step of outlier detection, spatial outliers are identified.
- iii) Finally, in the third step, to find presence of outliers in our dataset, identified spatial outliers are compared with temporal neighbors.

In nutshell, steps which is performed in modified algorithm are: Clustering, Identifying Spatial Outliers, Identifying Temporal Outliers and finally Spatio-temporal outliers are detected.

## **4.3 Discussion**

In the presented work above, an enhanced spatio-temporal clustering technique namely ST-OPTICS has been developed, performed experimentations and validated the cluster obtained. Developed technique is capable to generate spatio-temporal clusters and it can address the issues like i) Handling spatio-temporal data. ii) As technique does not depend on dimensions of data, so it can be concluded that it is ready to handle n-dimensions. iii) Independence of ordering of observation in database can be observed by working principle of proposed technique i.e. it first identifies order of observation before clustering and hence it results into improvement in the run time efficiency of the developed algorithm which can seen from fig 4.10.iv)The technique is also capable to discover nested (or micro level) clusters with arbitrary shape which is achieved using input parameters (min\_RD, max\_CD, threshold) in Extract\_STDBSCAN algorithm 4.3.v) Scalable.

Section 4.2.2, table 4.7 shows the comparison of total time taken by proposed ST-OPTICS and ST-DBSCAN algorithms where it reveals that proposed technique takes drastically less time compared to existing ST-DBSCAN algorithm, for same number of grids. As for as scalability issue is concerned, very large dataset is required, which have been obtained from SAC, ISRO, belongs to Proba-V satellite (Fourth satellite in the European Space Agency's PROBA series having 3304683 grids i.e. no. of observations, here V stands for Vegetation). This dataset has been pre-processed and made suitable for this application by SAC, ISRO. Experimentation results are plotted as given below, which clearly proves that developed approach ‘ST-OPTICS’ supports scalability too and outperforms existing ST-DBSCAN algorithm in the context of time taken by newly developed method ‘ST-OPTICS’ for same number of grids.



**Figure 4.11.** Performance Comparison of ST-OPTICS and ST-DBSCAN with different dataset size.

Note: For last dataset size (having 3304683 grids) experimentation could not be performed for ST-DBSCAN, as it took too much longer time. (Please refer fig. 4.11)

Owing to generation of micro level clusters (very large in numbers but good quality clusters) difficulty is faced in visualization, analysis and better interpretation, which led to hybridization of developed spatio-temporal clustering techniques with agglomerative approach. This hybridization helped a lot in interpreting the results.

To validate the obtained results, two approaches are used i.e. first by visualization and and secondly by performance indices. For evaluating the quality of clusters, various performance indices are studied with their theoretical principles and performed experimentation on theoretically

selected indices first on classical data to finally select performance indices and then selected 12 validation indices are applied on ST-DBSCAN and ST-OPTICS which generates clusters using ST Data, which are dense and arbitrary shaped in nature. Experimental results as given in section 4.2.4 clearly reveal that the ST-OPTICS is performing much better when compared with existing ST-DBSCAN algorithm in the context of quality of clusters and average run time efficiency.



## **Chapter 5**

### **Proposed Time Series Prediction Technique**

Statistical model like Integration of Auto Regressive (AR) and Moving Average (MA) i.e. ARIMA is capable to handle non-stationary time series but it can deal with only single time series. While machine learning approach (i.e. Support Vector Regression (SVR)) can handle dependency among different time series along with non-linear separable domains, however it cannot incorporate the past behavior of time-series. This led to combine these two approaches for improving accuracy of time series prediction, where focus has been given on minimization of forecast error using residuals, which helps to take appropriate action for near future. Focus is given on fusion of statistical and machine learning models for improving the accuracy of prediction. Keeping in view our objective, hybridization of Auto Regressive Integrated Moving Average (ARIMA) with SVR models has been done. Moreover, in order to reduce number of area wise models and reduction in time complexity for tuning different parameters, emphasis has been laid down on handling issues related to scalability by taking suitable representative samples from each sub-areas.



## 5.1 SARIMA: Statistical Prediction Technique

ARIMA(Montgomery, Jennings, and Kulahci) is basically a combination of two models: AR model and MA model. Individually, AR and MA models stationary time series. A time series whose mean is constant and gives no trend over time can be defined as stationary time series. But, if a trend in time series exists, it may result into non-stationarity which can be dealt by differencing the time series. In ARIMA, the letter "I" ensures that the series is transformed into stationary time series. In ARIMA the forecast is assumed to be a linear combination of past values and past errors. The form of ARIMA(p, d, q) model is written as follows:

$$\Phi(B)(1 - B)^d Y_t = \Theta(B)a_t \quad \dots\dots\dots (1)$$

where

- $\Phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p$  represents the autoregressive operator of non-negative order  $p$ .
- $\Theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q$  represents the moving average operator of non-negative order  $q$ .
- $Y_t$  is the obtained series.
- $B$  is the backshift operator.
- $d$  is order of difference to make series stationary.
- $a_t$  denotes the random errors (white noise) which are assumed to be independently and identically distributed with the mean of zero and a constant variance.

Now, if a seasonal component is there in series, then this seasonality factor can be taken care of by Seasonal-ARIMA which can be modeled as –

SARIMA( $p, d, q$ )x( $P, D, Q$ ) $_s$  :

$$\Phi(B)\phi(B^s)(1 - B)^d(1 - B^s)^D Y_t = \Theta(B)\theta(B^s)a_t \quad \dots\dots\dots (2)$$

where

- $s$  is the seasonal length.
- $\phi(B^s) = 1 - \phi_1 B^s - \phi_2 B^{2s} - \dots - \phi_P B^{Ps}$  represents the seasonal autoregressive operator of non-negative order  $P$ .

- $\theta(B^s) = 1 - \theta_1 B^s - \theta_2 B^{2s} - \dots - \theta_Q B^{Qs}$  represents the seasonal moving average operator of non-negative order  $Q$ .
- $D$  represents the number of seasonal differences.
- Rest of the terms are same as defined in equation (1).

In order to estimate the parameters  $p$ ,  $q$ ,  $P$  and  $Q$  of SARIMA, first it is required to determine  $d$  and  $D$  which are the orders of differencing needed to make the series stationary. With the help of Auto Correlation Function (ACF) and Partial Auto Correlation Function (PACF) plots of differenced series, we can evaluate the suitable values for  $p$ ,  $q$ ,  $P$  and  $Q$ .

## 5.2 RBF-SVR: Machine Learning Based Prediction Technique

The Support Vector Machine (SVM) was proposed by (Vapnik). It is firmly grounded in the framework of statistical learning theory or VC theory which has been developed over the last three to four decades by (Cortes and Vapnik). It works on the principle of structural risk minimization where an upper bound on generalization error is minimized rather than usual empirical error. SVM is used for classification. For regression and time series prediction it has been modified to get SVR. SVR can be formulated as follows:

$$y = \omega \varphi(x) + b \dots \dots \dots (3)$$

where

- $\omega$  is a vector and  $b$  is some scalar.
- $\varphi(x)$  is called the feature vector, which is nonlinear mapping from the input space  $x$  into feature space.

It is required to define a function such that it has at most  $\varepsilon$  deviation from the actually obtained targets  $d_i$  for all  $N$  training data-set. This means that the error less than  $\varepsilon$  is not penalized. Such function is defined as  $\varepsilon$ -insensitive loss function:

$$L_\varepsilon(d_i, y_i) = \begin{cases} |d_i - y_i| - \varepsilon & ; |d_i - y_i| > \varepsilon \\ 0 & ; otherwise \end{cases} \dots \dots \dots (4)$$

The coefficients  $\omega$  and  $b$  are estimated by solving the following convex optimization problem:

$$\min \frac{1}{2} \|\omega\|^2 + C \frac{1}{N} \sum_{i=1}^N L_{\varepsilon}(d_i, y_i) \dots\dots\dots(5)$$

The first term in the above equation defines the flatness of the function. C evaluates the trade-off between the empirical risk and the flatness of the model. Using the positive slack variables  $\xi$  and  $\xi^*$ , which represent the distance of the  $\varepsilon$  tube boundaries to the actual values, equation (5) can be transformed as follows:

$$\min \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^N (\xi_i + \xi_i^*) \dots\dots\dots(6)$$

Subject to:

$$d_i - \omega \varphi(x_i) - b \leq \varepsilon + \xi_i$$

$$\omega \varphi(x_i) + b - d_i \leq \varepsilon + \xi_i^*$$

$$\xi_i, \xi_i^* \geq 0$$

where  $i = 1, 2, \dots, N$

Finally, on introducing the Lagrangian multipliers  $\alpha_i, \alpha_i^*$  and by applying Karush-Kuhn-Tucker (KKT) conditions, we obtain the dual problem of equation (6) which is maximization problem as follows:

$$\max \sum_{i=1}^N d_i (\alpha_i - \alpha_i^*) - \varepsilon \sum_{i=1}^N (\alpha_i^* + \alpha_i) - \frac{1}{2} \sum_{i,j=1}^N (\alpha_i - \alpha_i^*) (\alpha_j - \alpha_j^*) K(x_i, x_j) \dots\dots(7)$$

Subject to:

$$\sum_{i=1}^N (\alpha_i^* - \alpha_i) = 0 \text{ and } \alpha_i, \alpha_i^* \in [0, C]$$

$$\text{Also } \alpha_i \cdot \alpha_i^* = 0$$

Lagrangian multipliers are calculated and the optimal weight vector is expressed as follows:

$$\omega^* = \sum_{i=1}^N (\alpha_i - \alpha_i^*) \varphi(x_i) \dots\dots\dots(8)$$

Hence, the solution is expressed as:

$$y = \sum_{i=1}^N (\alpha_i - \alpha_i^*) K(x_i, x) + b \dots\dots\dots(9)$$

where  $K(x_i, x)$  is called the kernel function.

Value of kernel function is equal to the inner product of two vectors  $x_i$  and  $x$  in the feature space  $\varphi(x_i)$  and  $\varphi(x)$  such that  $K(x_i, x_j) = \varphi(x_i) \cdot \varphi(x_j)$  satisfies the Mercer's condition. Kernel functions can be linear, polynomial, Gaussian etc. In our study, Gaussian kernel function is used which is specified as follows:

$$K(x_i, x_j) = \exp(-\gamma ||x_i - x_j||^2) \dots\dots\dots (10)$$

### 5.3 Hybrid Model for Prediction<sup>1</sup>

Initially ARIMA and SVR model are used individually to appreciate the pros and cons of these models independently, it is followed by Hybridization of ARIMA and SVR.

ARIMA model uses only single time series for modeling, it does not incorporate or check the dependency of other factors. Also, it deals with the linear domain only. While SVR is best suitable for non-linear domain problems and it incorporates other factors which have correlation among them, so by applying the hybrid approach, the dependency problem of ARIMA can be tackled using SVR and also, the forecast error can be minimized. Detailed description of hybridization can be understood by figure 5.1. Experimental results are given in section 5.2.

---

<sup>1</sup>K. P. Agrawal, Sanjay Garg, Shashikant Sharma, Pinkal Patel and Ayush Bhatnagar, Fusion of Statistical and Machine Learning Approaches for Time Series Prediction using Earth Observation Data, International Journal of Computational Science and Engineering, Inderscience. (Accepted on July 21, 2015, available in Forthcoming articles), H index :12.

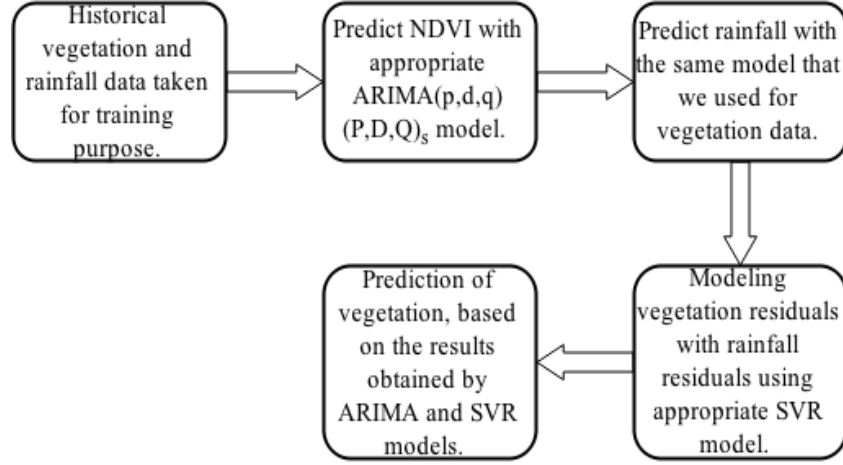


Figure 5.1: Flow chart of hybrid approach

The data is available from 2002-2010 from which 2002-2009 has been used for training purpose and the data of 2010 has been used for testing purpose. As shown in figure 5.1, prediction models of NDVI and rainfall are made through ARIMA with appropriate parameters (i.e. (p, d, q)x(P,D,Q)s ) using ACF and PACF plots.

Now after obtaining the values from ACF and PACF plots, we obtained model of SARIMA (0,0,0)x(1,1,0)23 for NDVI. Substituting these values (i.e. p, d, q, P, D, Q and s) in model equation 2 given in section 5.1, we have the following :

- $\Phi(B) = 1 - \phi_0 B^0 = 1$
- $\Theta(B) = 1 - \theta_0 B^0 = 1$
- $\phi(B^{23}) = 1 - \phi_1 B^{23}$
- $\theta(B^{23}) = 1 - \theta_0 B^{0*23} = 1$

which results into the forecasting equation 11 that is obtained as follows:

$$(1 - \phi_1 B^{23})(1 - B^{23})Y_t = a_t$$

$$(1 - B^{23} - \phi_1 B^{23} + \phi_1 B^{46})Y_t = a_t$$

$$Y_t - Y_{t-23} - \phi_1 Y_{t-23} + \phi_1 Y_{t-46} = a_t$$

$$\boxed{\hat{Y}_t = Y_{t-23} + \phi_1 Y_{t-23} - \phi_1 Y_{t-46} + a_t} \dots\dots\dots(11)$$

From residuals of NDVI and rainfall for training data, the prediction model of NDVI residuals is obtained using rainfall residuals with the help of RBF(Radial Basis Function)-

SVR by taking best parameters for RBF kernel function. This model as described is used to predict the NDVI residuals for testing data.

For example, the RBF-SVR ( $C, \varepsilon, \gamma$ ) model can be defined as:

$$y = \omega\varphi(x) + b$$

i.e.  $Dependentvariable = \omega\varphi(Independentvariable) + b$

where,

Dependent variable is the NDVI residuals.

Independent variable is the Rainfall residuals.

The  $\varphi(Independentvariable)$  solved by using RBF kernel function.

In generic sense dependent and independent variables can be other also.

RBF kernel function is as mentioned in equation 10 in section 5.2 where  $\gamma$  has been obtained optimally in this case. Optimal  $\omega$  can be obtained using Lagrangian multipliers which in turn can be obtained using  $C$  and  $\varepsilon$  (i.e. described in equation 7 and 8 in section 5.2).

Finally, we have predicted NDVI values for testing year (i.e. 2010) by taking either summation or subtraction of predicted values of NDVI given by ARIMA and residual values of NDVI given by RBF-SVR, which has given good results.

Stepwise procedure for performing prediction (as shown in the flow chart of figure5.1) is mentioned below:

- 1) First of all identify the best lag, which is a point where Pearson's Correlation coefficient between NDVI and Rainfall is maximum (say it is 'n').

- 2) From NDVI time series 'n' number of observations from beginning are removed and same number of observations(i.e. 'n') are removed from the end of rainfall time series (as it takes some time delay/lag for vegetation to cultivate after rainfall).



- 3) Using ARIMA model, predict NDVI for 2002-10, say predicted and actual data for the year 2002-09 as " $Y_{02-09}^*$ " and " $Y_{02-09}$ " and say predicted and actual data for the year 2010 as " $Y_{10}^*$ " and " $Y_{10}$ ".
- 4) Using ARIMA model, predict the rainfall for 2002-10, say predicted and actual data for the year 2002-09 as " $X_{02-09}^*$ " and " $X_{02-09}$ " and say predicted and actual data for the year 2010 as " $X_{10}^*$ " and " $X_{10}$ ".
- 5) Obtain the NDVI residual for the year 2002-09 as:  $Y_{resi.02-09} = Y_{02-09} - Y_{02-09}^*$
- 6) Obtain the rainfall residual for the year 2002-09 as:  $X_{resi.02-09} = X_{02-09} - X_{02-09}^*$
- 7) Apply SVR-RBF model on the residuals of NDVI and rainfall from 2002-09, which will provide us  $\omega$  and  $b$  as mentioned above.
- 8) Obtain the rainfall residual for the year 2010  $X_{resi.10} = X_{10} - X_{10}^*$  (provided by ARIMA).
- 9) Using parameters  $\omega$  and  $b$  provided by SVR-RBF model (in step 7), to predict NDVI residuals ( $Y_{resi.10}$ ) for 2010 with the help of rainfall residual ( $X_{resi.10}$ ) of the year 2010.
- 10) Final prediction of NDVI for 2010 = predicted NDVI for 2010 ( $Y_{10}^*$ ) + predicted NDVI residuals for 2010 ( $Y_{resi.10}$ ).

Hence, this hybrid approach not only minimizes the prediction error of ARIMA but also incorporates the dependency of other factors.

## 5.4 Experimental Analysis

### 5.4.1 Experimental Setup

The "R" language/environment is the tool being open source has been used for experimentation purpose for obtaining the resultsof prediction, as it supports all statistical models and machine learning algorithms, and for coding our own algorithm.

Figure 5.2 shows NDVI time series from 2002-2010 for Ahmedabad district where the data available from year 2002-2009 has been used for training purpose and the data of year 2010 has been used for the testing purpose. The time series shown in figure5.2 possess two characteristics:

- i) Non-Stationarity
- ii) Seasonality.

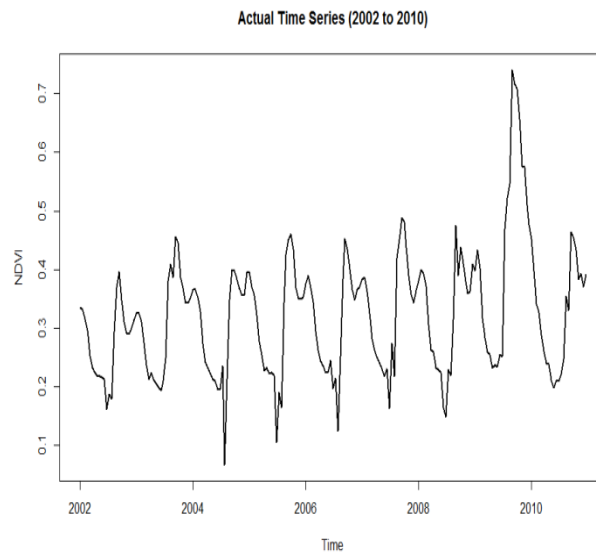


Figure 5.2: Time Series of NDVI from 2002 to 2010

### 5.4.2 Data Specification

Table 5.1:Dataset Specification

Satellite / Sensor	<ul style="list-style-type: none"><li>•MODIS (Moderate Resolution Imaging Spectroradiometer)</li><li>•NOAA-CPC (National Oceanic and Atmospheric Administration-Climate Prediction Center)</li></ul>
Measure	<ul style="list-style-type: none"><li>•NDVI (Normalized Difference Vegetation Index) (16 days composite)</li><li>• Rainfall (Daily basis but converted into 16 days composite)</li></ul>
Area	Entire Country (India)
Grid Size	5*5 km <sup>2</sup>
Total No of Grids	1,30,307
Period	2002-2010

In dataset whose specification is given in table5.1, each grid (specified by location having its latitude and longitude) have 23 NDVI and rainfall observations that are taken at an interval of 16 days for the complete year starting from January to December. Before experimentation, proper care has been taken for cleaning the dataset. For experimentation purpose, dataset is divided into two seasons namely "Kharif" (May to October) and "Rabi" (November to April)which are the two major crop season of India.

### 5.4.3 Evaluation of Performance Parameters

Root Mean Square Error (RMSE) and Coefficient of Determination (R-square) are used as performance indices in order to evaluate the adaptability of different approaches.

### 5.4.4 Results

#### Prediction using ARIMA model

In order to model using ARIMA we need to analyze the ACF and PACF plots of a time series. So, figure5.3 shows the ACF and PACF plots of the time series for available data and from graphs it's clear that given time series is non-stationary (where mean is variant) and seasonal (where we found cut offs at lag 1, 23, 46 and so on).

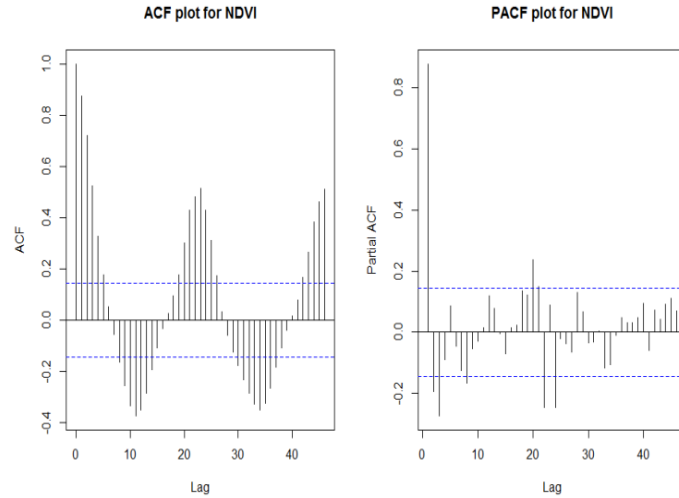


Figure 5.3: ACF and PACF plots for NDVI time series

figure5.4 shows ACF and PACF plots with seasonal differencing( $D=1$ ). Based on the observation of ACF and PACF plots for non-seasonal and seasonal behavior,  $ARIMA(0,0,0)(1,1,0)_{23}$  model has been selected.

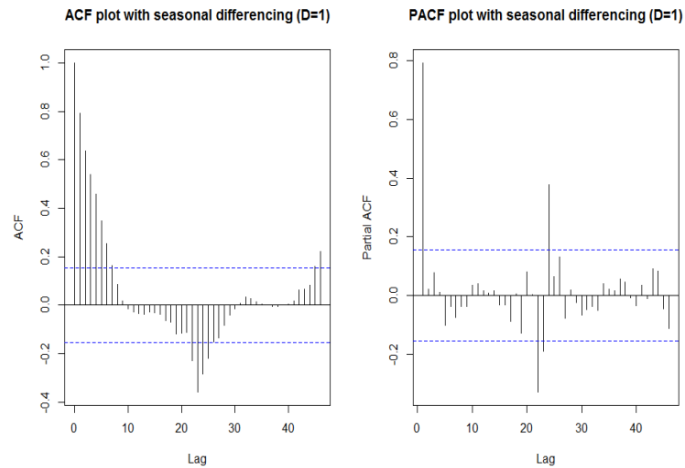


Figure 5.4: ACF and PACF plots for NDVI time series with seasonal difference( $D=1$ ).

So  $ARIMA(0,0,0)(1,1,0)_{23}$  model is applied on the data of Ahmedabad district available from the year 2002-2009 which has been taken for training purpose and found that the RMSE value for the test data of year 2010 is 6.86%. Figure 5.5 shows the actual time series from year 2002-2010 of NDVI and predicted time series by ARIMA model for the year 2010.

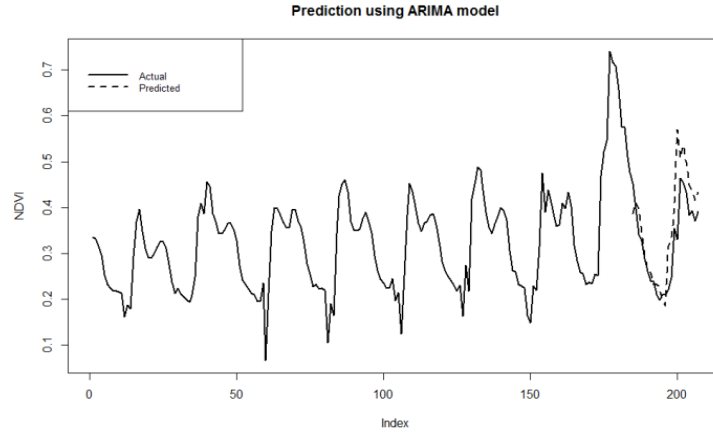


Figure5.5: Actual NDVI Time Series (2002-2010) and Prediction only for 2010 using ARIMA model

Note that in figure 5.5, Index indicates the number of readings for non-spatial data i.e. we have 23 readings of NDVI in a year for a particular district, so total readings for 9 years would be 207. figure5.6 is the expanded view of the figure 5.5 over the year 2010 to show the original time series and predicted value by seasonal ARIMA more effectively.

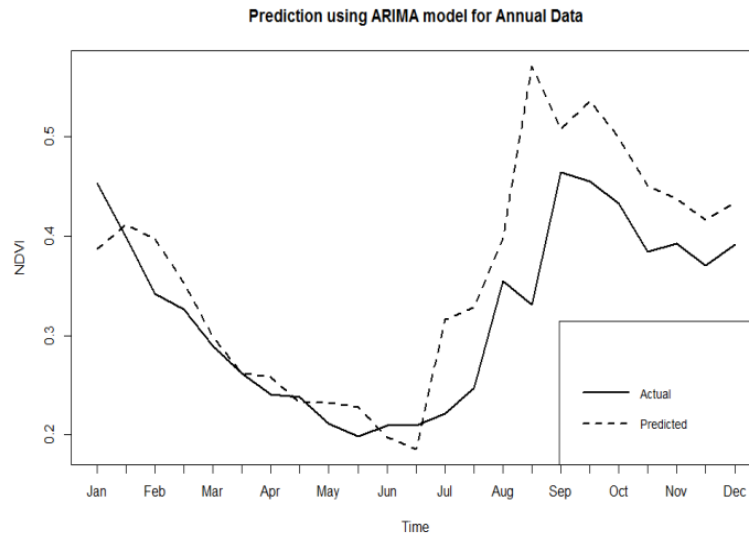


Figure 5.6: Actual and Predicted NDVI Time Series for the year 2010 using ARIMA model

### Prediction using SVR model

In SVR algorithm, important step is to decide appropriate kernel function from the available functions (i.e. Linear, Polynomial, Gaussian/RBF etc). In the present study, the Radial Basis

Function has been adapted based on its robustness (Chen and Zhang), excellent learning capability (Kim et al.) and less numeric difficulties (Munoz-Mari, Bruzzone, and Camps-Vails).

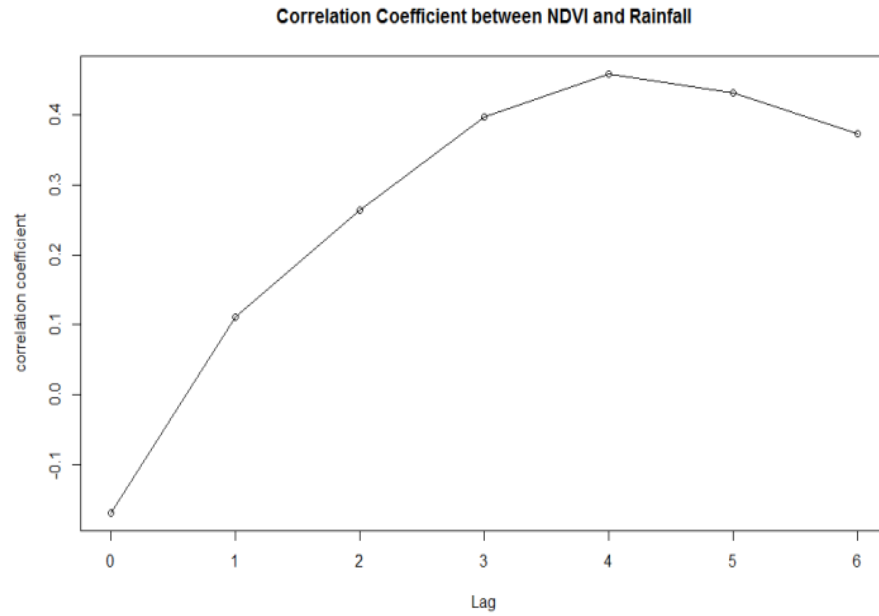


Figure 5.7: Pearson's Correlation Coefficient between NDVI and rainfall

Before applying Gaussian/RBF-SVR, it is necessary to detect the lag between rainfall and NDVI as vegetation is not expected immediately after rainfall. Using Pearson's correlation coefficient, the best lag found is 4 as shown in figure 5.7 which means the 4<sup>th</sup> observation of NDVI is highly correlated with 1<sup>st</sup> observation of rainfall. The range constraints of the RBF-SVR parameters are set as  $C$  [0.1,2000],  $\epsilon$  [0.001,1] and  $\gamma$  [0.01,4]. The three parameters  $C$ ,  $\epsilon$  and  $\gamma$  are then adjusted in a way to keep error minimum.

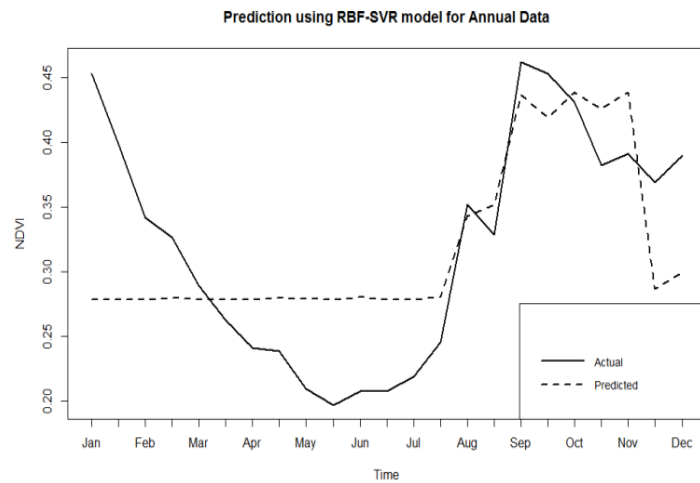


Figure 5.8: Actual and predicted NDVI Time Series for the year 2010 using RBF-SVR model.

Prediction of NDVI from rainfall using RBF-SVR is shown in figure 5.8. It is clear that the prediction is not accurate during initial periods from November to June as there is no correlation between NDVI and rainfall during these months but it is predicting accurately during the period from July to October due to correlation existence between them or we can say due to rainfall in these months.

**Conclusion drawn from the results of ARIMA and RBF-SVR model:**

ARIMA (results are shown in figure 5.6) and RBF-SVR (results are shown in figure 5.8) models which are used to predict NDVI values for the year 2010 are performing differently in the months from July to October and from November to June respectively. Hence, we are proposing hybrid approach that is the combination of seasonal ARIMA and RBF-SVR for better prediction than the individual.

**Prediction using Hybridization of ARIMA and RBF-SVR model:**

Before applying the hybrid approach, the data is divided into two major crop seasons of India i.e. "Kharif" (May to October) and "Rabi" (November to April). Also we have seen through experimentation that there is no correlation between NDVI and rainfall from November to June, which more or less falls under the Rabi season. So division of the year becomes more helpful. Results of hybrid model for both Kharif and Rabi season are summarized in the tables 5.2 and 5.3 respectively, also figure 5.9 and 5.10 respectively show the visualization.

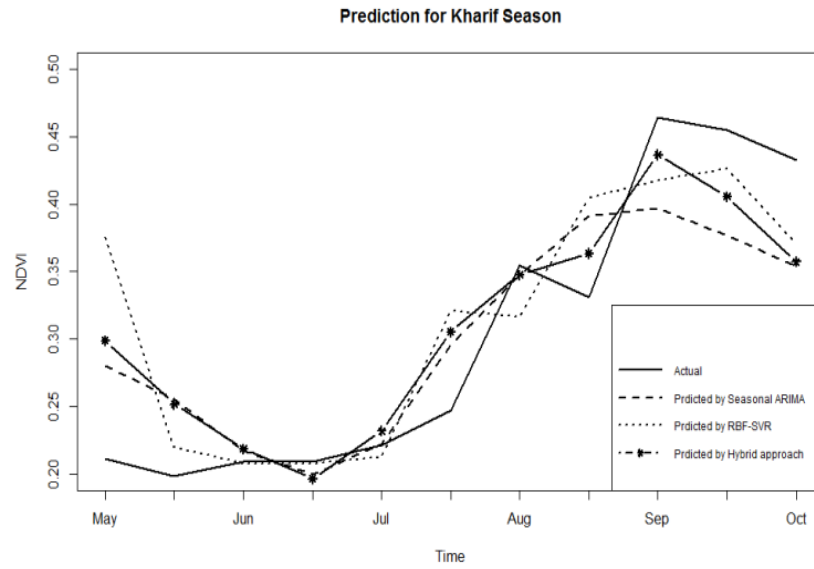


Figure 5.9: Prediction of NDVI for Kharif season by Seasonal ARIMA, RBF-SVR and Hybrid approach for the year 2010.

Table 5.2: Performance comparison for Kharif season

Model	RMSE(%)	R-square(%)
ARIMA	5.30	75.04
RBF-SVR	6.51	39.50
Hybrid approach	4.68	76.97

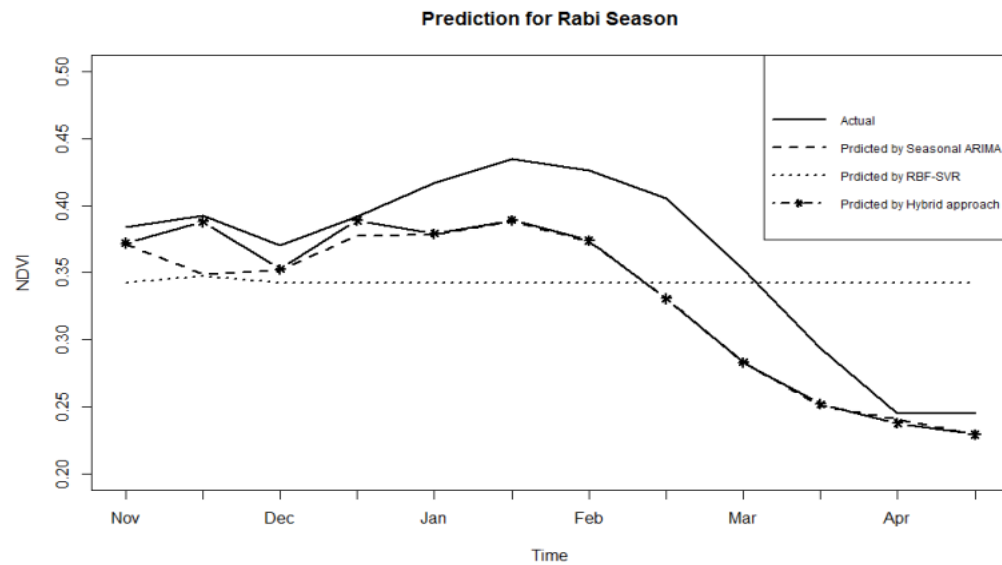


Figure 5.10: Prediction of NDVI for Rabi season by Seasonal ARIMA, RBF-SVR and Hybrid approach for the year 2010



Table 5.3: Performance comparison for Rabi season

Model	RMSE(%)	R-square(%)
ARIMA	4.25	90.12
RBF-SVR	6.66	2.03
Hybrid approach	4.00	93.51

### Performance Improvement

It is clear that in Ahmedabad district, the hybrid approach outperforms the individual approaches. For Kharif and Rabi season, results given by hybrid model are better than RBF-SVR and ARIMA. Though with a little gain in RMSE but this inspired us to do the same on a big scale which is shown in section 5.5 and the results are quite satisfactory.

## 5.5 Scalability Enhancements

It has been shown that the hybrid approach is better than the individual approaches for Ahmedabad district, but we cannot make a separate model for each and every district of the state and consequently of entire country to prove the same thing, as the number of models will increase drastically. So we have made sincere efforts for making our proposed approach scalable by making a group of adjacent districts (in our case we have considered 7 districts which are in the vicinity of Ahmedabad district called “central zone”). This enables one to reduce number of models at the state level and the same notion can be extended at the country level. In order to take into account the effect of different locations, we took a representative sample which includes 20% of data from each of the 7 districts. This line of action also helps in estimating parameters (total 9) of hybrid approach which is very time consuming task if done individually for each district. This attempt improves the time complexity of making scalable model to a greater extent.

Attempt has been made to obtain the optimal results by setting various parameters of ARIMA and RBF-SVR. After tuning the parameters of ARIMA, we obtained (22,1,0)(0,1,3) for Kharif and (16,2,0)(0,1,0) for Rabi season as the best set and similarly for RBF-SVR tuning has been made

over  $C$  [0.1,2000],  $\varepsilon$  [0.001,1] and  $\gamma$  [0.01,4] ranges. After deployment of our proposed hybrid model with tuned parameters over the representative sample, the results obtained on all the 7 districts are shown in table 5.4 and table 5.5 for both seasons i.e. Kharif and Rabi respectively. Moreover they also contain the number of grids in each district, RMSE and  $R^2$  for individual models i.e. ARIMA and SVR-RBF along with performance of hybrid approach, in order to show its comparison. It is found that their RMSE is well within 2-5% and also  $R^2$  (R-square) for each district is above 80%. This shows that hybrid approach is scalable too. Also from tables it is clear that the hybrid approach improves over individual ARIMA and RBF-SVR though small gain in few districts but in majority we can see its success.

A small observation to note is that in Kharif season, the value of  $R^2$  for RBF-SVR is more compared to Hybrid approach in most of the districts. We know that the value of  $R^2$  represents the percentage of variance in NDVI explained by model and predictor variable(rainfall). Since RBF-SVR solely depends on rain for prediction of NDVI, more  $R^2$  seems legit, while in Hybrid not only rain is a factor but NDVI also incorporates its past behavior. So here variance in NDVI is not only explained by rainfall but by itself too. And for the support of this justification if we see Rabi season then here Hybrid is better in each district in terms of both RMSE and  $R^2$ . During this season rainfall is less, so explaining variance in NDVI using solely rainfall won't suffice. Thus  $R^2$  of Hybrid approach is better here compared to RBF-SVR.

Table 5.4: Performance comparison on 7 districts of Gujarat for Kharif season

District	Number of Grid	Models	RMSE (%)	$R^2$ (%)
Ahmedabad	338	ARIMA	4.41	78.45
		RBF-SVR	7.12	95.34
		HYBRID Approach	4.31	80.80
Anand	149	ARIMA	5.15	95.38
		RBF-SVR	5.95	96.42
		HYBRID Approach	4.85	95.40
Bhavnagar	433	ARIMA	4.85	90.28
		RBF-SVR	9.29	93.88
		HYBRID Approach	4.27	90.71
Gandhinagar	24	ARIMA	6.67	83.66
		RBF-SVR	8.27	93.99
		HYBRID Approach	4.70	99.00

Table 5.4 continued

District	Number of Grid	Models	RMSE (%)	R <sup>2</sup> (%)
Kheda	128	ARIMA	4.36	90.18
		RBF-SVR	7.14	96.00
		HYBRID Approach	4.34	89.89
Mahesana	205	ARIMA	4.44	94.66
		RBF-SVR	7.89	94.47
		HYBRID Approach	4.20	94.89
Surendranagar	407	ARIMA	4.32	81.51
		RBF-SVR	8.07	93.41
		HYBRID Approach	3.98	84.56

Table 5.5 Performance of hybrid approach on 7 districts of Gujarat for Rabi season

District	Number of Grid	Models	RMSE (%)	R <sup>2</sup> (%)
Ahmedabad	338	ARIMA	3.55	86.33
		RBF-SVR	12.57	56.16
		HYBRID Approach	3.50	86.31
Anand	149	ARIMA	2.33	86.01
		RBF-SVR	6.72	85.51
		HYBRID Approach	2.31	86.04
Bhavnagar	433	ARIMA	5.35	93.91
		RBF-SVR	7.22	88.08
		HYBRID Approach	4.63	93.99
Gandhinagar	24	ARIMA	2.43	95.83
		RBF-SVR	10.60	72.11
		HYBRID Approach	2.39	95.95
Kheda	128	ARIMA	2.62	93.56
		RBF-SVR	7.44	83.29
		HYBRID Approach	2.58	94.12
Mahesana	205	ARIMA	4.13	95.46
		RBF-SVR	12.46	61.75
		HYBRID Approach	4.05	95.50
Surendranagar	407	ARIMA	5.14	88.02

		RBF-SVR	5.98	81.03
		HYBRID Approach	4.87	87.77

## 5.6 Discussion

A time-series consists of sequences of values or events obtained over repeated measurements of time, typically measured at equal time intervals (e.g., hourly, daily, weekly). It is utilized in many applications (section 2.3). From among many statistical models like AR, MA, ARMA, ARIMA etc. which are available for prediction purpose, except ARIMA, all other models are for stationary time series (where mean remains constant) consequently ARIMA model has been chosen which can also forecast for non-stationary time series (where mean is variant). But all these statistical models work for a single time-series measure only. To take into consideration the effect of seasonal (Kharif and Rabi) changes SARIMA is finally chosen.

As SVR creates system which is trained from historical time series and attempts to predict new value for near future based on dependency (like prediction of NDVI time series does depends on rainfall data), SVR model has been chosen. Basic functionality of SVR is to map the input data into high dimensional feature space through a nonlinear mapping function, and to solve a linear regression problem in this feature space. Various kernel functions are available in SVR like Linear, Polynomial, RBF (Radial Basis Function) etc. From among this set of kernel functions, literature survey reveals that RBF kernel performs better, so RBF-SVR model is finally chosen. Owing to some restriction/limitation of these model when used independently, hybridization of SARIMA with RBF-SVR is done in order to improve the quality of time series prediction. Time lag between rainfall and cultivation of vegetation is also taken into consideration for finding correlation between the two measure i.e. rainfall and vegetation.

In order to reduce number of area wise model, adjacent sub-areas are merged into single model taking representative samples into consideration, which also helped in saving time which is normally consumed in turning parameters of both these models for getting optimal performance.

## **Chapter 6**

### **Summary, Conclusions and Future Scope**

#### **Summary**

Presented research work is summarized as, to start with an enhanced Spatio-Temporal clustering technique namely ST-OPTICS has been developed which is based on existing algorithm called OPTICS. Next in order to predict time series, hybrid model is designed and developed which is combining the features of statistical (SARIMA) and machine learning (RBF-SVR) model where scalability issue is also taken into consideration and addressed successfully.

#### **6.1 Conclusions**

Based on literature survey, issues and challenges which are prevailing as mentioned in section 3.1.2 have been kept in view while designing and developing new clustering technique ST-OPTICS and they have been addressed successfully. Moreover experimental work is said to be incomplete without validation of results obtained, so validation have also been performed and results are compared with existing ST-DBSCAN algorithm and observed that ST-OPTICS is performing much better. Thus conclusively it can be said that first objective (section 1.2) of my research work is achieved successfully.

As it is known that there exists a natural dependency between vegetation and rainfall. So in order to predict time series and to find out above said dependency, a new approach, which is hybridization of ARIMA and RBF-SVR is developed and when applied, has produced very good results. The models are defined for both the crop seasons Kharif and Rabi which are two major crop seasons of India, independently. This hybrid approach is also scalable with respect to area in

order to save time which is due to reduction in number of models and hence reduction in timing for parameter tuning. The delay of rainfall effect on vegetation is also taken into consideration. Thus, if we have rainfall data, we can easily forecast vegetation of future depending on the lag between them with a very high accuracy, and hence second objective (section 1.2) of my research work is also achieved successfully.

## **6.2 Future Scope**

Research is never ending process, no existing system is complete, so there are few suggestions which can be taken into consideration in future research such as incremental management of clusters on updates of database, application of parallel computing and usage of spatial indexing data structure to improve performance, processing over distributed architectural platform etc.

Owing to the variability of climatic conditions, the other parameters like air temperature, soil moisture adequacy etc. may be taken into consideration for further enhancement in the results. Moreover this work can be extended to make generic model which can support different domains too i.e. domains other than vegetation field.

## **Chapter 6**

### **Summary, Conclusions and Future Scope**

#### **Summary**

Presented research work is summarized as, to start with an enhanced Spatio-Temporal clustering technique namely ST-OPTICS has been developed which is based on existing algorithm called OPTICS. Next in order to predict time series, hybrid model is designed and developed which is combining the features of statistical (SARIMA) and machine learning (RBF-SVR) model where scalability issue is also taken into consideration and addressed successfully.

#### **6.1 Conclusions**

Based on literature survey, issues and challenges which are prevailing as mentioned in section 3.1.2 have been kept in view while designing and developing new clustering technique ST-OPTICS and they have been addressed successfully. Moreover experimental work is said to be incomplete without validation of results obtained, so validation have also been performed and results are compared with existing ST-DBSCAN algorithm and observed that ST-OPTICS is performing much better. Thus conclusively it can be said that first objective (section 1.2) of my research work is achieved successfully.

As it is known that there exists a natural dependency between vegetation and rainfall. So in order to predict time series and to find out above said dependency, a new approach, which is hybridization of ARIMA and RBS-SVR is developed and when applied, has produced very good results. The models are defined for both the crop seasons Kharif and Rabi which are two major crop seasons of India, independently. This hybrid approach is also scalable with respect to area in order to save time which is due to reduction in number of models and hence reduction in timing for parameter tuning. The delay of rainfall effect on vegetation is also taken into consideration. Thus, if we have rainfall data, we can easily forecast vegetation of future depending on the lag between them with a very high accuracy, and hence second objective (section 1.2) of my research work is also achieved successfully.

## **6.2 Future Scope**

Research is never ending process, no existing system is complete, so there are few suggestions which can be taken into consideration in future research such as incremental management of clusters on updates of database, application of parallel computing and usage of spatial indexing data structure to improve performance, processing over distributed architectural platform etc.

Owing to the variability of climatic conditions, the other parameters like air temperature, soil moisture adequacy etc. may be taken into consideration for further enhancement in the results. Moreover this work can be extended to make generic model which can support different domains too i.e. domains other than vegetation field.



## References

- Agrawal, Rakesh et al. "Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications." *SIGMOD '98: Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data*. Ed. Laura Haas et al. ACM Press, 1998. 94–105.
- Ankerst, Mihael et al. "OPTICS: Ordering Points To Identify the Clustering Structure." *SIGMOD 1999, Proceedings ACM SIGMOD International Conference on Management of Data, June 1-3, 1999, Philadelphia, Pennsylvania, USA*. Ed. Alex Delis, Christos Faloutsos, and Shahram Ghandeharizadeh. ACM Press, 1999. 49–60.
- Arbelaitz, Olatz et al. "An Extensive Comparative Study of Cluster Validity Indices." *Pattern Recognition* 46.1 (2013): 243–256. Web. 8 Dec. 2014.
- Baboo, S Santhosh, and K Tajudin. "Clustering Centroid Finding Algorithm (CCFA) Using Spatial Temporal Data Mining Concept." *2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering* (2013): 30–36.
- Barnathan, Michael. "Mammographic Segmentation Using WaveCluster." *Algorithms* 5.3 (2012): 318–329.
- Bernard, Desgraupes. "clusterCrit: Clustering Indices." 2013.
- Bezdek, James C, and Nikhil R Pal. "Some New Indexes of Cluster Validity." *IEEE Transactions on Systems, Man, and Cybernetics, Part B* 28.3 (1998): 301–315.
- Birant, Derya, and Alp Kut. "ST-DBSCAN: An Algorithm for Clustering Spatial–temporal Data." *Data & Knowledge Engineering* 60.1 (2007): 208–221. Web. 23 May 2013.
- Chaovalit, Pimwadee et al. "Discrete Wavelet Transform-Based Time Series Analysis and Mining." *ACM Computing Surveys* 43.2 (2011): 1–37. Web. 26 Aug. 2013.
- Chen, Songcan, and Daoqiang Zhang. "Robust Image Segmentation Using FCM with Spatial Constraints Based on New Kernel-Induced Distance Measure." *IEEE Transactions on Systems, Man, and Cybernetics, Part B* 34.4 (2004): 1907–1916.
- Cooper, Glenn W Milligan; Martha C. "An Examination of Procedures for Determining the Number of Clusters in a Data Set." *Psychometrika* 50.2 (1985): n. pag.
- Cortes, Corinna, and Vladimir Vapnik. "Support-Vector Networks." *Mach. Learn.* 20.3 (1995): 273–297.
- Deng, Ke, Hong Shen, and Hui Tian. "Self Projecting Time Series Forecast: An Online Stock Trend Forecast System." *IJCSE* 2.1/2 (2006): 46–56.
- Ertoz, Levent, Michael Steinbach, and Vipin Kumar. "Finding Clusters of Different Sizes, Shapes, and Densities in Noisy, High Dimensional Data." *Proceedings of the Third SIAM International Conference on Data Mining (SDM 2003)*. Ed. Daniel Barbara and

- Chandrika Kamath. Vol. 112. Society for Industrial and Applied Mathematics, 2003. Proceedings in Applied Mathematics.
- Ester, Martin et al. "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise." *KDD* (1996): n. pag. Web. 6 May 2014.
- Foody, G. "Geographical Weighting as a Further Refinement to Regression Modeling: An Example Focused on the NDVI-Rainfall Relationship." *Remote Sensing of Environment* 88.3 (2003): 193–283.
- Forest Survey of India, Ministry of Environment Forest, Government of India. "State of Forest Report." 2014. <<http://fsi.nic.in/details.php?pgID=qu 4>>.
- G., Erika Johana Salazar et al. "A Cluster Validity Index for Comparing Non-Hierarchical Clustering Methods." 2002: n. pag.
- Guha, Sudipto, Rajeev Rastogi, and Kyuseok Shim. "CURE: An Efficient Clustering Algorithm for Large Databases." *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data*. New York, NY, USA: ACM, 1998. 73–84. Web. SIGMOD '98.
- Gujarat State Seeds Corporation Limited, Guj info Petro Limited. "Crop Lists." 2014. <<http://www.gurabini.com/cropList.aspx?id=1>>.
- Jarvis, R. and Patrick, E. A., "Clustering using a similarity measure based on shared near neighbour's," *Computers, IEEE Transactions on*, vol. C-22, pp. 1025-1034, Nov1973.
- Halkidi, Maria, Yannis Batistakis, and Michalis Vazirgiannis. "Clustering Validity Checking Methods: Part II." *SIGMOD Record* 31.3 (2002): 19–27.
- Halkidi, Maria, Yannis Batistakis, and Michalis Vazirgiannis. "On Clustering Validation Techniques." *Journal of Intelligent Information Systems* 17 (2001): 107–145.
- Halkidi, Maria, and Michalis Vazirgiannis. "Clustering Validity Assessment: Finding the Optimal Partitioning of a Data Set." *ICDM*. Ed. Nick Cercone, Tsau Young Lin, and Xindong Wu. IEEE Computer Society, 2001. 187–194.
- Han, J, M Kamber, and J Pei. *Data Mining: Concepts and Techniques*. Elsevier Science, 2011. The Morgan Kaufmann Series in Data Management Systems.
- Han, Jiawei, Micheline Kamber, and Anthony K H Tung. "Spatial Clustering Methods in Data Mining: A Survey." *Geographic Data Mining and Knowledge Discovery, Research Monographs in GIS*. Ed. Harvey J Miller and Jiawei Han. Taylor and Francis, 2001.
- Hinneburg, Alexander, and Daniel A Keim. "An Efficient Approach to Clustering in Large Multimedia Databases with Noise." *KDD*. Ed. Rakesh Agrawal, Paul E Stolorz, and Gregory Piatetsky-Shapiro. AAAI Press, 1998. 58–65.

- Ilango, M R, and V Mohan. "A Survey of Grid Based Clustering Algorithms." *International Journal of Engineering Science and ...* 2.8 (2010): 3441–3446.
- Iwasaki, H. "NDVI Prediction over Mongolian Grassland Using GSMaP Precipitation Data and JRA-25/JCDAS Temperature Data." *Journal of Arid Environments* 73.4-5 (2009): 557–562. Web. 7 Oct. 2014.
- J., Duncan, Dash J., and Atkinson P. "Analysing Temporal Trends in the Indian Summer Monsoon and Its Variability at a Fine Spatial Resolution." *Climatic Change* 117.1 (2013): 119–131.
- Kaufman, L, and Peter J Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley, 1990.
- Kim, Dae-Won et al. "Rapid and Brief Communication: Evaluation of the Performance of Clustering Algorithms in Kernel-Induced Feature Space." *Pattern Recogn.* 38.4 (2005): 607–611.
- Kisilevich, Slava et al. "Spatio-Temporal Clustering." *Data Mining and Knowledge Discovery Handbook*. Ed. Oded Maimon and Lior Rokach. Springer US, 2010. 855–874.
- Kolatch, Erica. "Clustering Algorithms for Spatial Databases: A Survey." *PDF is available on the Web* (2001): 1–22.
- Kovács, Ferenc, Csaba Legány, and Attila Babos. "Cluster Validity Measurement Techniques." *6th international symposium of Hungarian ...* (2005): n. pag. Web. 20 Dec. 2014.
- KP Agrawal, Sanjay Garg, and Pinkal Patel. "Performance Measures for Densed and Arbitrary Shaped Clusters." *IJCSE* 6.2 (2015): 338–350.
- Kumar, T V, and K K Rao. "Studies on Spatial Pattern of NDVI over India and Its Relationship with Rainfall, Air Temperature, Soil Moisture Adequacy and ENSO." *Geofizika* 30.1 (2013): 1–18.
- Lee, Sanghoon, and Melba M Crawford. "Unsupervised Multistage Image Classification Using Hierarchical Clustering with a Bayesian Similarity Measure." *IEEE Transactions on Image Processing* 14.3 (2005): 312–320.
- Liu, Yiming, Min Yao, and Rong Zhu. "A Novel Hybrid Model for Image Classification." *IJCSE* 6.1/2 (2011): 96–104.
- Milligan, Glenn W. "A Monte Carlo Study of Thirty Internal Criterion Measures for Cluster Analysis." *Psychometrika* 46.2 (1981): n. pag.
- Ministry of Statistics and Programme Implementation. "Agriculture Survey." 2014. <<http://mospi.nic.in/Mospi New/site/inner.aspx?status=3&menu id=57>>.

- Mkhabela, M.S. et al. "Crop Yield Forecasting on the Canadian Prairies Using MODIS NDVI Data." *Agricultural and Forest Meteorology* 151.3 (2011): 385–393. Web. 1 Sept. 2014.
- Montgomery, Douglas C, Cheryl L Jennings, and Murat Kulahci. *Introduction to Time Series Analysis and Forecasting*. 1st ed. Wiley-Interscience, 2008.
- Morgan, Joseph T et al. "Adaptive Feature Spaces for Land Cover Classification with Limited Ground Truth Data." *Multiple Classifier Systems*. Ed. Fabio Roli and Josef Kittler. Vol. 2364. Springer, 2002. 189–200. Lecture Notes in Computer Science.
- Munawar, Mohammad Ahmad, and Paul A S Ward. "Leveraging Many Simple Statistical Models to Adaptively Monitor Software Systems." *IJHPCN* 7.1 (2011): 29–39.
- Munoz-Mari, Jordi, Lorenzo Bruzzone, and Gustavo Camps-Valls. "A Support Vector Domain Description Approach to Supervised Classification of Remote Sensing Images." *IEEE T. Geoscience and Remote Sensing* 45.8 (2007): 2683–2692.
- Nagpal, PB, and PA Mann. "Comparative Study of Density Based Clustering Algorithms." *International ...* 27.11 (2011): 44–47.
- Ng, Raymond T, and Jiawei Han. "Efficient and Effective Clustering Methods for Spatial Data Mining." N.p., 1994. 144–155.
- Niazmardi, Saeid, and J Shang. "A New Classification Method Based on the Support Vector Regression of NDVI Time Series for Agricultural Crop Mapping." *Agro-Geoinformatics (Agro-Geoinformatics), 2013 Second International Conference on*. 2013. 361-364 Web. 23 Dec. 2014.
- Omuto, C T et al. "Mixed-Effects Modeling of Time Series NDVI-Rainfall Relationship for Detecting Human-Induced Loss of Vegetation Cover in Dry Lands." *Journal of Arid Environments* 74.11 (2010): 1552–1563.
- Rashid, ANMB, and MA Hossain. "Challenging Issues of Spatio-Temporal Data Mining." *Computer Engineering and Intelligent ...* 3.4 (2012): 55–64. Web. 10 Sept. 2013.
- Rousseeuw, Peter. "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis." *Journal of Computational and Applied Mathematics* 20.1 (1987): 53–65.
- S., Garg, and Jain R. C. "A Heuristic Based Variation of K-Mean Clustering Algorithm for Dealing with Outlier." *IJCSE* 4.3 (2006): 56–60.
- S., Garg, and Jain R. C. "Variations of K-Mean Algorithm: A Study for High-Dimensional Large Data Sets." *Information Technology* 5.6 (2006): 1132–1135.
- Sheikholeslami, Gholamhosein, Surojit Chatterjee, and Aidong Zhang. "WaveCluster: A Wavelet-Based Clustering Approach for Spatial Data in Very Large Databases." *The*

- VLDB Journal *The International Journal on Very Large Data Bases* 8.3-4 (2000): 289–304.
- Shekhar, Shashi et al. “Identifying Patterns in Spatial Information: A Survey of Methods.” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1.3 (2011): 193–214. 14 Aug. 2013.
- Shim, Yosung, Jiwon Chung, and In-Chan Choi. “A Comparison Study of Cluster Validity Indices Using a Nonhierarchical Clustering Algorithm.” *Computational Intelligence for Modelling, Control and Automation, 2005 and International Conference on Intelligent Agents, Web Technologies and Internet Commerce, International Conference on*. Vol. 1. N.p., 2005. 199–204.
- Smola, Alex J, and Bernhard Schölkopf. “A Tutorial on Support Vector Regression.” *Statistics and computing* 14.3 (2004): 199–222.
- Speech and Image Processing Unit, School of Computing University of Eastern Finland. “Clustering Datasets.” 2014. Web. 20 Mar. 2002.
- Symons, M J. “Clustering Criteria and Multivariate Normal Mixtures.” *Biometrics* 37.1 (1981): pp. 35–43.
- Theodoridis, Sergios, and Konstantinos Koutroumbas. *Pattern Recognition, Third Edition*. Orlando, FL, USA: Academic Press, Inc., 2006.
- Vapnik, V N. *The Nature of Statistical Learning Theory*. New York: Springer, 1995.
- Wang, Kaijun, Baijie Wang, and Liuqing Peng. “CVAP: Validation for Cluster Analyses.” *Data Science Journal* 8.May (2009): 88–93.
- Wang, Min, Aiping Wang, and Anbo Li. “Mining Spatial-Temporal Clusters from Geo-Databases.” 1 (2006): 263–270.
- Wang, Wei, Jiong Yang, and Richard R Muntz. “STING: A Statistical Information Grid Approach to Spatial Data Mining.” *Proceedings of the 23rd International Conference on Very Large Data Bases*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1997. 186–195. VLDB ’97.
- Weingessel, A, E Dimitriadou, and S Dolnicar. *An Examination of Indexes for Determining the Number of Clusters in Binary Data Sets*. N.p., 1999.
- Zhang, Li, Liping Lei, and Dongmei Yan. “Comparison of Two Regression Models for Predicting Crop Yield.” *IGARSS. IEEE*, 2010. 1521–1524.
- Zhang, Tian, Raghu Ramakrishnan, and Miron Livny. “BIRCH: An Efficient Data Clustering Method for Very Large Databases.” *SIGMOD ’96: Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data*. Ed. Jennifer Widom. ACM Press, 1996. 103–114.

Zhao, Qinpei, Mantao Xu, and Pasi Fränti. "Sum-of-Squares Based Cluster Validity Index and Significance Analysis." *Adaptive and Natural Computing Algorithms* (2009): 313–322. Web. 20 Dec. 2014.

Zhu, Li, and Shengyong Xu. "Prediction Algorithm Based on Web Mining for Multimedia Objects in next-Generation Digital Earth." *IJES* 7.1 (2015): 79–87.