AMITY UNIVERSITY

UTTAR PRADESH

Major Project Report

On

WATER UTILITY MANAGEMENT USING DATA ANALYTICS AND GEOSPATIAL ANALYTICS

Submitted to
Amity University Uttar Pradesh

In partial fulfilment of the requirements for the award of the degree of
Bachelor of Technology
In
Computer Science and Engineering
By

ANKITA GUPTA
A2305214086

Under the guidance of

Mr. ABHISHEK SINGHAL
Dy. Head of Department,
Department of Computer Science and Engineering
And
DR. A. SAI SABITHA
Head of Department,
Department of Information Technology

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
AMITY SCHOOL OF ENGINEERING AND TECHNOLOGY
AMITY UNIVERSITY UTTAR PRADESH
NOIDA (U.P.)
May 2018

DECLARATION

I, **Ankita Gupta**, student of B.Tech (CSE) hereby declare that the project titled **Water Utility Management Using Data Analytics And Geospatial Analytics,** which is submitted by me to Department of Computer Science and Engineering, Amity School of Engineering and Technology, Amity University Uttar Pradesh, Noida, in partial fulfilment of requirement for the award of the degree of Bachelor of Technology in Computer Science and Engineering, has not been previously formed the basis for the award of any degree, diploma or other similar title or recognition.

Date: 30 April 2018

Place: Noida

Ankita Gupta

DECLARATION FORM
(HEALTH, SAFETY & PLAGIARISM)

I, **Ankita Gupta**, student of Bachelor of Technology, Computer Science and Engineering, Enrolment Number - A2305214086, batch 2014-2018, Department of Computer Science and Engineering, Amity School of Engineering and Technology, Amity University, Uttar Pradesh, Noida, hereby declare that I have carefully gone through the project guidelines including policy on health and safety, policy on plagiarism.

Date: 30 April 2018 Student Signature

Place: Noida

CERTIFICATE OF ORIGINALITY

On the basis of declaration submitted by **Ankita Gupta**, student of B. Tech CSE-2X, I hereby certify that the project titled "WATER UTILITY MANAGEMENT USING DATA ANALYTICS" which is submitted to Department of Computer Science and Engineering, Amity School of Engineering and Technology, Amity University Uttar Pradesh, Noida, in partial fulfilment of the requirement for the award of the degree of Bachelor of Technology in CSE, is an original contribution with existing knowledge and faithful record of work carried out by her under my guidance and supervision.

To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Date: 30 April 2018

Place: Noida

MR. ABHISHEK SINGHAL
Department of Computer Science and Engineering
Amity School of Engineering and Technology
Amity University Uttar Pradesh, Noida

DR. A. SAI SABITHA
Department of Information Technology
Amity School of Engineering and Technology
Amity University Uttar Pradesh, Noida

ACKNOWLEDGEMENT

I take this opportunity to express my profound sense of gratitude and respect to all those who helped me throughout this project.

This report acknowledges to the intense driving and technical competence of the every individual that has contributed to it. It would have been almost impossible to complete this project without the support of these people. I extend warm regards and gratitude to **Prof. (Dr.) Abhay Bansal,** Head, Dept. of CSE, Joint Head, ASET, Director, International Collaboration for Engineering and Technology, **Mr. Abhishek Singhal,** Deputy Head, Dept. of CSE, for his constant guidance throughout the course of work, **Dr. A Sai Sabitha**, Head of department**,** Department of Information Technology, for her imparting knowledge and mentorship in all aspects. They shared their valuable time from their busy schedule to guide me and provide their active and sincere support for my activities.

I wish to thank **Mr. V. S Gupta, D.J.B** for his profound insight into the problem domain of the project and providing industrial perspective and data for goal achievement and practical deductions and results.

This report is authentic record of individual work which is accomplished by the sincere and active support by all the faculty of my college. I have tried my best to summarize this report.

ANKITA GUPTA
B. Tech CSE-2(X)
Amity School of Engineering and Technology,
Noida

ABSTRACT

Over the years the utility-customer relationship has rapidly transformed with vast number of factors affecting consumption including environmental factors like geography, weather, population, migration etc. as well as social and political factors like economy, Household sizes, government policies, tariffs etc. Amid this, customer satisfaction has become the utmost concern and key performance metric for utilities. A result of this, utility has enlightened to turn towards data-driven and information-centric and decision-making models for understanding demand-supply patterns, and overall sustainable management of utilities also.

The research work focuses on implementing a data driven approach towards understanding the water utility management, finding loopholes in demand supply patterns and customer base analysis using data analytics methods and geospatial approach. The first half focuses research work aims at understanding the water utility consumption in metropolitan city of parts of City of Delhi, using data analytics that aims at understanding per household consumption of consumers and segmentation of customers based on customer profiling. The second half of the research work focuses on geospatial analysis using visual GUI based tool like ArcGIS. The customer profiling is carried out by mapping consumption based data and analytical observations on DSSDI map in order to capture the density based distributions and variability of consumers across various geographical regions.

This study can overcome the problems of mismanagement of revenue and identification of potential consumers of water utility and focus on sustainability of water resources through policy enhancements and implementations.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF PUBLICATIONS

1) Ankita Gupta, Abhishek Singhal, A. Sai Sabitha, Tanupriya Chaudhary, “Water Utility Management using Data Analytics and Geospatial Analytics”.

Related Works

1) Ankita Gupta, Gaurav Raj, Vikram Singh, “Traversal based Ordering of Arc-Node Topological dataset of sewer network”, 2nd International Conference on Computational Intelligence and Informatics (ICCI-2017), Springer 2017.

CHAPTER I
INTRODUCTION

1.1 Introduction

For over 5 decades, Delhi water utility has been meeting the needs of potable water for the National Capital Territory of Delhi. The population of Delhi has seen a start rise over the past few decades with the figures rising up to 180 lakhs, and an additional floating population of about 8 to 10 lakhs. With the help of systematic planning and implementation process, more than 82% population of Delhi has been provided access to piped water supplied, through a network of more than 12000 Kms. of water main. By optimizing all the resources the water production has been augmented upto 878 MGD per day. The used and raw water is obtained from various sources like the river Bhakra Storage, Upper Ganga Canal, Yamuna and Ground Water.

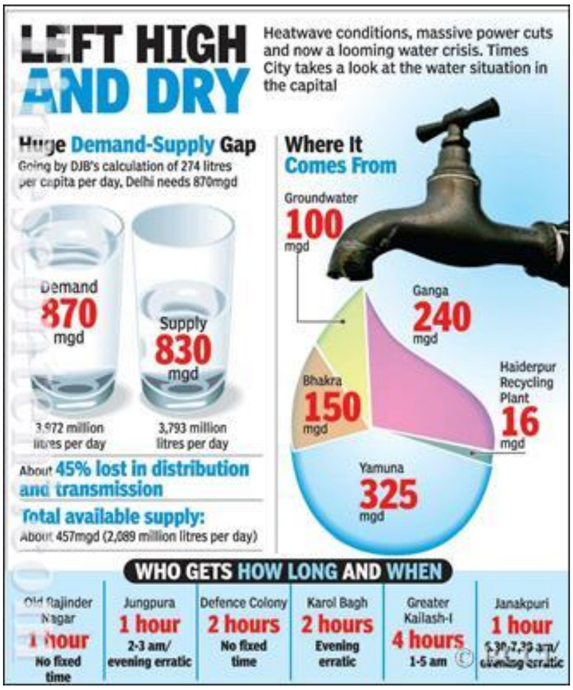


Figure 1.1 Water Scarcity News Piece

In spite of such massive statistics, the management and efficiency of water utility reaching each household continues to be below par with the demand needs of the population. Seasonal shortages of water supply, unbilled customer base due to mismanagement, loss of revenue due to hidden customer base leading to excessive exploitation of resource etc. are some of the problems that plague the water industry and continue to affect the population needs.

The use of analytics plays a crucial part in improving and sustaining a utility’s all-important customer connection, along with its business performance. One of the biggest drivers at play for the ability to provide more customized, individual service is the customer focused and more personal and effective relationship with each one of its customers. The best picture of this customer based orientation towards understanding of utility can be captured by user centric data and information generated from customer base.

Considering the rising worldwide concern for the scarcity and sustainability of natural resources, an efficient management of resources is need of the hour. With information technology bringing revolutionary developments in the past few decades, the problem of resource management can be tackled and understood better through the use of Data mining. A large amount of data is generated from the supply and consumption of water by the users of a defined population over a fixed period of time. This data is able to capture the seasonal variants and trends of data consumption in relation various other factors that influence the consumption of resources.

The purpose of this research project aims at understanding the water resource consumption factors and patters by understanding the various other influencing factors like customer demographics, geographical topology, climate and weather changes, population demographics, manner of land use patterns etc. With this knowledge, a behavioural understanding can be established which can help target better policies for the government in sustainability of natural water resources.

The research work aims at understanding the water utility consumption in metropolitan city using data analytics techniques considering various sources of data pertaining to water utility and monthly metering schemes, geographic and climate topography of the area, customer demographics etc. in different areas of NCT of Delhi.

Along with the social cause of mismanagement and scarcity of water utility, the developed models help the government in accessing the revenue model of the Utility management Bodies. The problem of Hidden customers, arrears, potential customers can be easily targeted through a data mining based approach to the problem.

1.2 Motivation

This study draws its motivation from the limitations of other alternatives available for the purpose some of which are summarised as below:

1. *Hardware Implementation*: For equitable distribution of water supply and rationalisation can be achieved by implementation of SCADA systems, consisting of field instrumentation such as pressure sensors, flow meters, and actuator valves which are installed at various strategic locations in the water supply networks, starting from source to the supply end. This type of system is very cost intensive and require considerable technical knowhow and skills which are normally not easily available in typical government utility. The system also needs extensive efforts towards day-to-day maintenance and operations.

2. *GIS and Spatial Analysis*: Another approach of understanding population based analyses of utility management is through spatial projection of population and consumption based attributes through Geospatial information systems. To keep the system functional, alive, and useful to all incremental system improvements and its linkage to master plan action, the GIS becomes a critical resource. The institutional and functional aspects of maintaining and keeping the database current pose interesting challenges that are necessary to address for a progressive water utility to operate, maintain, and manage its functions. The mapping of resources and population over geographical locations gives clearer idea of availability and supply of water utility however it does not target the quantitative aspect of optimization and rationalisation problem for resource. Hence pure GIS based systems without analytics can be limited in scope of the current problems.

A hybrid approach of Geographic information systems with data analytics can provide the qualitative and quantitate analysis of revenue management issues and sustainable management of resource.

1.3 Objective

To propose a system that analyses the problem related to water utility and help understand the customer base better. Following are the key objective of the project-

1. Segmentation of customers and categorising them under different categories based on consumption.
2. Identification of Yearly Usage patterns.
3. Identification of Potential Customers for Extension of customer base.
4. Identification of Excessive Utility Exploiting customers.

The final research outcomes will provide a quantitative and qualitative analysis of the water utility of a zone of Delhi region. It will provide the insight into how the customer base can be identified with categories of similar consumption trends and factors affecting their behaviour. The study further focuses on the water consumption pattern throughout the year for different customer groups.

This analysis can be implemented through unsupervised learning algorithms and techniques. Some of the algorithms that can be used are as follows:

1. Clustering of datasets in order to capture consumption, region based trends.

2. Time series analysis provides insight into the seasonal and annual trends of water consumption that can be further used to predict usage spikes in near future and reason with anomalous behaviour of consumption.

3. Correlation analysis can be used to determine trends and patterns of consumption with respect to seasonally variant factors like temperature, precipitation, humidity etc.

1.4 Limitations

1) Inadequate Data: The dataset collected can be inadequate accounting to lot of factors. The unauthorized localities and slum areas are difficult to be field surveyed and project an erroneous and approximate estimation of the data quantitate data collected. The data collected further is unreliable for accurate estimation and

prediction analysis since the dynamics such as building units, area, and land use pattern as subjective to changes as per individual inhabitants without any legal notices. Thus prediction patterns for such areas can be erroneous.

2) Insufficient Geographic mapping: The geographic mapping of the field surveyed data cannot be 100% mapped to the digital GIS based maps. Demographic profile of Delhi city is very complex as the city has developed over the years in a haphazard manner. Therefore the property address system is not standardised and cannot it does not support the standard geocoding systems for the mapping of streets and house addresses. This essentially requires physical surveys of each and every property to capture latitude and longitude and map the same with corresponding billing address and customer details.

3) Incomplete Billing: The water utility department being a public service utility does not accounts for various public welfare schemes by providing subsidised water to a section of consumers. Further some schemes also include providing unmetered water connection on lump sum monthly charges. Under these circumstances the quantity of water supplied in not 100% metered and exact estimation are difficult to ascertain.

4) Incoherent view of geographic Data in analytical software: The data collected from the various regions of the area are geographic in nature comprising of latitude and longitudes. Such features cannot be taken into account or visualised using pure data analytical tools like R etc. Some area based patterns may be missed out considering the pure ordinal and categorical nature of data.

5) Faulty Metering and hardware: The data for the consumption of water units per household may have discrepancies due to illegal tampering of metering system by the residents or excessive consumption by illegal residents or third parties. This account for hidden customers. Also many households have built-in hand pumps and access to ground water reserve which does not accounts in the portable water distribution through the utility. Therefore the actual consumption patterns cannot be governed through the water Utility analysis.

1.5 Project Timeline

Progress Phase wise	Proposed Start	Proposed Finish
Literature Survey and Data Collection	10-07-2017	12-09-2017
Learning algorithms and application	27-09-2017	20-10-2017
Implemented algorithms and running comparative analysis	12-01-2018	15-03-2018
Project Report	15-03-2018	06-04-2018

Table 1.5.1. Project Timeline

1.6 Risk Assessment

Risk No.	Risk Description	Probability (1-3)	Impact (1-3)	Risk Score	Remedy/Mitigation/Fall-Back	Action By
1	Developer/ analyst not devoting enough time for project due to day to day activities	3	3	9	To ensure that student accord highest priority to the project	Student
2	Scope Creep	3	3	9	A sign-off on Business Requirements Strict adherence to Change Control Procedures	Student and Mentor
3	Delay in Reviews	2	3	6	Project plan to highlight milestones, deliverables and sign-off expected date clearly	Student and Mentor
4	Delay in Providing clean data	2	2	4	Student needs to take up Data cleansing activity on foremost and high priority	Student
5	Insufficient geographic mapping	3	2	6	Selection of Subset narrow training dataset and	Student

Risk No.	Risk Description	Probability (1-3)	Impact (1-3)	Risk Score	Remedy/Mitigation/Fall-Back	Action By
					prototyping for conceptual understanding	
6	Denial of Service because of viruses /Trojans/ worms	1	3	3	The proposed anti-virus solution would be deployed and will be continuously kept updated for looking and knowing the new virus signatures so that an attack from new virus is prevented from creeping in.	Student

Table 1.6.1 Risk Assessment

1.7 Report Structure

This project report is organised in seven chapters.

Chapter 1: This Chapter deals with the introduction to the problem statement, motivation, objective and scope of the work done.

Chapter 2: This chapter provides the background study of fields in which work is done, literature survey explaining the various research related to this work. Current practices for data mining and their relevance in the proposed research work.

Chapter 3: This chapter accounts the methodology and implemented model for analysis, pre-processing and modelling of data, related tools and software involved.

Chapter 4: This chapter focuses on the major results and outcomes of experimentation, laying emphasis and reasoning of obtained results and discussions.

Chapter 5: This chapter illustrates the conclusion and future prospects of implementation of proposed research methodology into commercial software and billing, revenue based systems.

CHAPTER II

LITERATURE REVIEW

2.1 Research Reviews

S. No	Publication	Author/Year	Key Concepts and Implementations
1.	Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks	Geoffrey K. F. Tso, Kelvin K.W. Yau (Elsevier, 2005)	Focused on prediction of electricity utility consumption and comparative study of 3 predictive models i.e. Regression analysis decision tree, neural networks. Model selection is decided by the square root of average squared error.
2.	Measuring and comparing the efficiency of water utility companies: A data development analysis approach	Giulia Romano, Andrea Guerrini (Elsevier, 2011)	Focuses on finding the efficiency of the water utility establishments in different regions on the basis of the size of the organisation and clustering them to determine maximum efficiency organisations.
3.	Exploring patterns in water consumption by clustering	Chrysi Laspidoua, Elpiniki Papageorgiou, Konstantinos Kokkinosb, Sambit Sahud, Arpit Guptae, Leandros Tassiulasf (Elsevier 2015)	Focuses on implementing of automatic Clustering of consumers using SOM Matlab Tools to identify clusters of different users and group them on the basis of mixed or natural consumers, i.e. residential or commercial consumers. The case study was divided into 2 subgroups, i.e. pure household and the consumption patterns based on the occupancy levels. Second case study was based on the mixed consumers that includes residential and commercial.
4.	A Statistical Analysis of Water Utility Operating Data for 1965 and 1970	Harris F. Seidel (Journal (American Water Works Association), Vol. 70)	The research paper gave a comprehensive statistical view of the water production units per volume groups to the consumption including revenue, allowances, and water rates, water utility trends, operation and maintenance costs etc. and related consumption trends over the year 1965-70.

5.	Water Utility Operation Data:An Analysis	Harris F. Seidel Thea, (Journal, American Water Works Association)	The research was purely statistical and included of means, medians, mode, quintiles and quartile distributions of water consumption over the year for various consumer groups and customers. It was one of the earliest attempts at understanding energy and utility using data.
6.	Clustering Time-Series Energy Data from Smart Meters	Alexander Lavin, Diego Klabjan, (Elsevier, 2016)	The approach used by the author here was to cluster the time series based consumption data from meter readings for electricity was clustered into various groups based on calendar months and customer profiles.
7.	Usage Analysis for Smart Meter Management	Hongfei Li, Dongping Fang, Shilpa Mahatma, Arun Hampapur, (IEEE 2011)	The research focused on flood of usage data obtained from meter in real-time that uses data analytics and optimization tool to support smart meter management. Various approach for meter anomaly detection, usage demand forecasting and association analysis for water utility was statistically proven including identification of malfunctioning meters, optimizing water supply in the future and understanding factors that associate and drive meter failures and water demand.
8.	On Building a Big Data Analysis System for California Drought	Pengcheng Zhang, Jerry Gao, A. G. Thomas, K. P. Alagupackiam, K. Mannava, P. I. Bosco, and Sen Chiao, (IEEE 2017)	The research proposes a drought forecasting approach that is based upon the PDSI index and focuses on the moisture level anomaly using climatic factor precipitation value and the previous month's PDSI index.
9.	Estimation of residential water demand: a state-of-the-art review	Fernando Arbués, Mar´ıa Ángeles Garc´ıa-Valiñas, Roberto Mart´ınez-Espiñeira, (2003)	This paper surveys the issues in the residential water demand. Different tariff types and their objectives are realised. Then, the main contributions in residential water demand estimation are reviewed, with specific attention to variables, models, data set, and the most common econometric constrains.

10.	Utilities and Big Data: Using Analytics for Increased Customer Satisfaction	Oracle (2013)	The Oracle published White paper for analytics services for Water utility management uses data mining techniques for understanding utilities and providing consumer specific and personalised services for management of bills and consumption.
11.	China’s water sustainability in the 21st century: a climate informed water risk assessment covering multi-sector water demands	X. Chen, D. Naresh, L. Upmanu, Z. Hao, L. Dong, Q. Ju, J. Wang, and S. Wang, (Copernicus Publications on behalf of the European Geosciences Union)	Paper proposes spatial approach to understanding the problem of water scarcity and identification of high risk zones that are vulnerable to drought under the effect of climate/water changes and water scarcity in China.
12.	Consumption Analytics and Forecasting Engine	Giorgos Giannopoulos, Sophia Karagiorgou, Yannis Kouvaras, Michalis Alexakis, Pantelis Chronis, (DAIAD Research project, E.C.)	The research deliverable report implemented a scalable utility forecasting engine, that implemented FML-KNN classification technique to classify users based on consumption by system trained using the similar dataset.
13.	Managing Water Utilities With Geographic Information Systems: The Case Of The City Of Tampa, Florida		The Thesis presented Geographic Information System (GIS) applications that is developed for the water utilities of the city of Tampa, Florida. The application proses tools and activities that helps monitor the city's water and wastewater infrastructure development, planning and maintenance using geo-database.
14.	SMAS: A Smart Meter Data Analytics System	Xiufeng Liu, Lukasz Golab and Ihab F. Ilyas, (IEEE 2011)	The author proposed and implemented SMAS inside a relational database management system with the help of open source tools like PostgreSQL and the MADLib machine learning toolkit.

Table 2.1.1 Research Review

2.2 Data Mining Methods

Data mining is the study of data and extraction of coherent and useful patterns, relations and information that can be used for facilitating businesses, researches and industries by understanding the market, customers, and commodities better. There are various data mining methods and algorithms that are used for such analysis, for example, segmentation, classification, prediction, association, time series analysis, and statistical analysis.

2.2.1 Clustering

Clustering is a technique that identifies patterns of similarities existing between the data sets and their interrelationship between them and groups them coherently in order to capture maximum similarity trends within the subgroups or clusters. It helps us to understand the hidden similarities and dissimilarities within the data and derive patterns for understanding the data in coherent manner. Some of the widely used algorithms for clustering are as follows:

1. **Density Based Clustering:** The density based clustering identifies the dense and parse regions within the spatial domain of data. It segregated the data by forming clusters around the high density regions.
2. **Centroid Based Clustering:** In centroid-based clustering, clusters are designed around central vector, which are not primarily needed to be a member of the data set. The algorithm finds k clusters and assigns each member of the dataset to the locality of these clusters such that the intra-cluster distance is minimised and inter-cluster distance is maximized.
3. **Distributed Clustering:** The original problem is decomposed into sub-problems that are processed in a distributed fashion.
4. **Hierarchical Method:** It is also known as 'nesting clustering', where the data is clustered within the bigger clusters to form a hierarchy of clusters.

Within these types of clustering approaches there are many types of algorithms proposed to implement these clustering. Some of the popularly used algorithms are the Fuzzy C means method, K-means algorithm, K-Medoids clustering, Hierarchical clustering and others alike, which are all used widely for deriving patterns and clustering. Here in this research we use K-Means clustering analysis to find patterns and groups within the customer base to identify their consumption patterns that vary with respect to seasonal influences.

2.2.1.1 K-means Algorithm

The k-means algorithm uses a dataset with n data entries. It clusters them into K clusters such that k<=n. The goal is to identify k seed points or mean points in the data set and associate every other data entry to either of these seed points so that every entry forms a part of a cluster or a partition. Various indices are used to check the optimal configuration of clusters some of which are Davis Bouldin Index, Silhouette index etc. The algorithm uses an iterative refinement technique wherein a set number of steps are iterated until the optimal solution is obtained or finite cohesive clusters are achieved. It is also referred to as Lloyd's algorithm.

Given an initial set of *k* seed points or means $m_1^{(1)},\ldots,m_k^{(1)}$, the algorithm executes by alternating the two steps as given below:

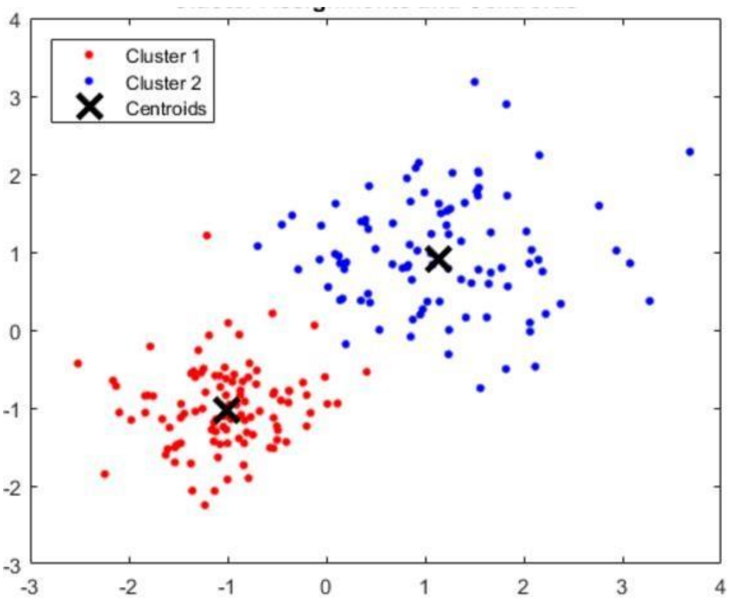


Figure 2.1 Cluster assignment and Centroid

1) **Assignment step**: Every point is assigned to one of the cluster based on the Euclidian distance between the sed points and the given point. The point is assigned the cluster that has the least distance of its seed point from the data point.

$$S_i^{(t)} = \{x_p : \| x_p - m_i^{(t)} \|^2 \leq \| x_p - m_j^{(t)} \|^2 \ \forall j, 1 \leq j \leq k\},$$

where each x_p is assigned to exactly one $S_i^{(t)}$, based on the Euclidean distance in spite of the fact that it can assigned to two of clusters.

2) **Update step**: The new means are calculated within the new clusters identified in the previous step. The algorithm converges at the point when the points in data do not change their associations with the final clusters. Whereas the algorithm do not guarantees the optimal solution.

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j$$

Here we use Davis Boudin index to find the most suitable value that defines the value of k. It is decided by checking the value of Davis Bouldin index as the smallest value tells the cohesiveness of the points within the clusters. Therefore the lesser the value of DB index is, better the value of K for clustering orientation is. The K-Means algorithm executes via following pseudo code.

```
Let $a_i … a_k$ be randomly selected inputs from $x_i … x_k$
 Repeat
            for i=1…n do
                  $y_{ij}$ = 1 if j = argmin$_j$ $\|xi(j) - cj\|^2$
                  $y_{ij}$ = 0 otherwise
            end for
            for j=1…k do
                  $n_j = \sum_{i=1}^{n} yij$
                  $a_j = \frac{1}{N} \sum_{i=1}^{n} yij\ xi$
            end for
        until convergence
End procedure
Return a1 … ak
```

The following expression is used during analysis of k means:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \|xi(j) - cj\|^2$$

J=Objective function, n= number of cases, c_j = the centroid for cluster j, x= case number, k= total number of clusters.

2.2.2 Correlation Analysis

Correlation is used to test relationships between quantitative variables or categorical variables. In other words, it’s a measure of how things are related. The study of how variables are correlated is called correlation analysis.

A correlation coefficient is a way to put a value to the relationship. Correlation assigns a value between -1 to 1 for any two variable set of values. A 0 value signifies that there is no set pattern or relationship between the quantities. A -1 signifies a perfect negative correlation between the quantities while a value of 1 signifies a perfect positive correlation. The negative or positive nature of their relationship can be understood by the graph of the two variables.

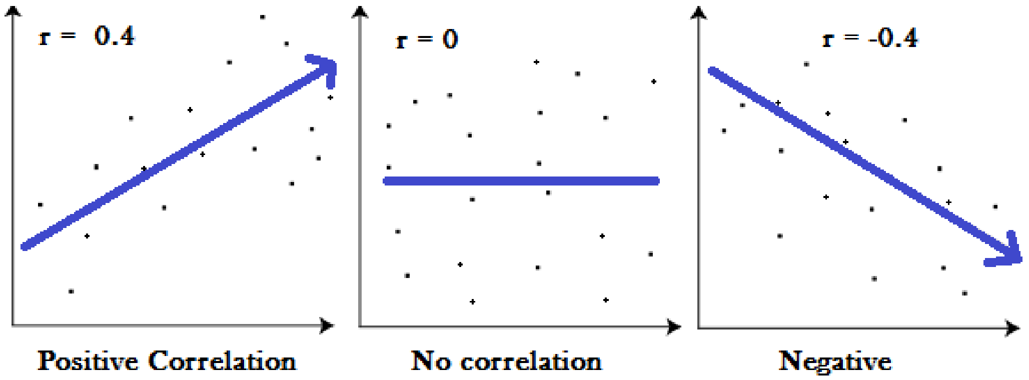


Figure. 2.2. Correlation Analysis.

The Pearson's correlation coefficient when applied, the coefficient value is represented by the letter *r*. For, two distinct dataset A={x_1,...,x_n} containing *n* values and dataset B = {y_1,...,y_n} containing *n* values, the formula for *r* is:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

Where, n = sample size of population;

x_i, y_i = are the samples indexed as i;

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean for x_i and similarly for y.

The most commonly used correlation coefficient is the Pearson Correlation Coefficient. It is used for testing linear relationships existing between data. Various different correlation coefficients are used based on the nature of data and the their use. For example, Kruskal lambda coefficient, Goodman coefficient are used, and are quite common coefficient.

2.3. Geospatial analysis

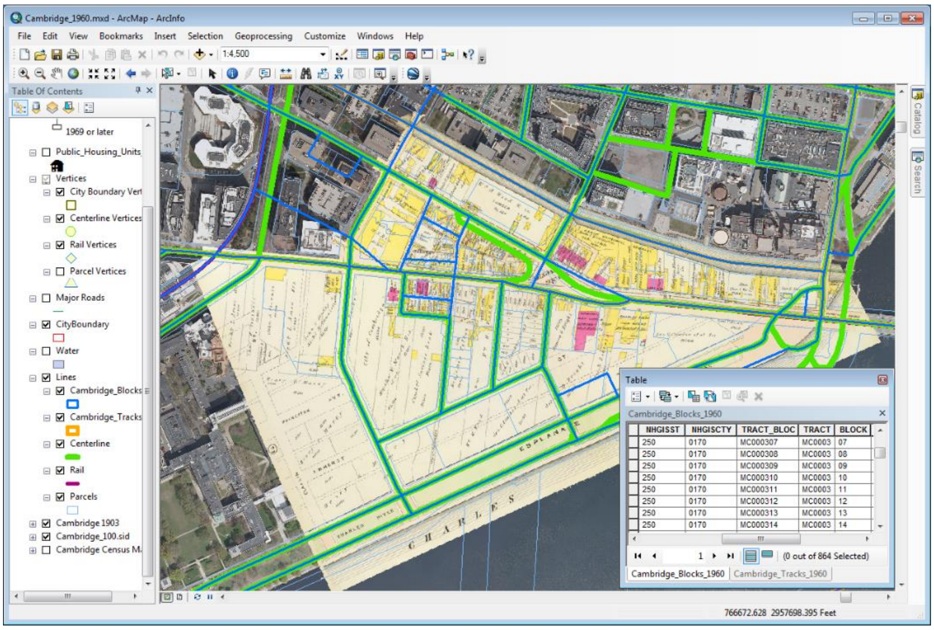
Geospatial analysis, or spatial analysis, is an approach towards geographical or spatial data that applies statistical analysis and analytical data techniques. This analysis employees soft wares that are capable of rendering spatial maps and graphs that process spatial data, It also applies analytical methods on geographic datasets and terrestrial datasets, and also includes the application of geographic information systems (GIS) and Geomatics.

Apart from the 2D and 3D mapping analysis Geospatial analysis includes the following functions:

1) **Surface analysis:** Surface analysis deals with analysis of physical surface properties like visibility and aspect, gradient along with analysing fields that pertains to surface-like data.
2) **Network analysis**: The network analysis understands and models the properties of manmade or natural networks with the aim of understanding the flow behaviour around and in such networks. GIS-based network analysis is being used to address a wide variety of problems that cater to route selection and facility location, problems that involve flows of transportation and hydrology
3) **Geo-visualization**: Geo-visualisation refers to creation and manipulation of diagrams, charts, images, maps, and other 3D views with their linked tabular datasets. Different visual tools like static and rotating, spatio-temporal visualisations, surface representation, animations etc. are packed into such GIS packages.

2.3.1 Types of GIS Analysis

1. **Spatial measurements:** The spatial distance between two points in a spatial plane is made easy to be calculated and analysed through areas, polygons, points and line analysis. Identification of areas, overlaps, and other complex renditions are simplified.
2. **Information Retrieval:** The GIS tools allows extraction of information of any element on the map using point and click feature. The information of area, features, attributes, objects, etc. can be easily accessed in this way. The GIS systems have embedded DBMS features that allows easy extraction and retrieval of data or specify information on point click.
3. **Searches by geography**: The records of data are features in a map database. The GIS spatial retrieval generates maps, that allows for visual searching of information and it highlights the end result. For instance in order to generate a report; the spatial equivalent will produce a finished map that finds the locate features. Spatial equivalents of the DBMS queries result in locating sets of features, or building new GIS layers. The Combination of spatial queries and attribute based queries build complex GIS operations.
4. **The query interface:** The query interface helps manage the data through SQL based querying system. An SQL based standard interface is used for interacting with data stored in RDBMS.
5. **Spatial overlay:** The spatial relationships are built using spatial overlay that overlays layers of features over one spatial plane, and maps all the features together that share a common area. The new spatial plane results in new relationships between features.
6. **Boundary analysis:** Boundary analysis, is the process of districting that is used to define regions within the spatial based on some criterion. The defined area can be of specific demographic characterises. The GIS software interactively defines the boundaries that can further calculate aggregate values and population statistics within the defined new geographic boundaries.
7. **Neighborhood Operations:** The neighborhood features can calculate various parameters surrounding a defined region. For example for a boundary region defined, the neighborhood operations can identify the maximum, minimum, average or median values of a feature within 5 unit distance of the boundary.



Cambridge_Blocks_1960

NHGISST	NHGISCTY	TRACT_BLOC	TRACT	BLOCK
250	0170	MC000307	MC0003	07
250	0170	MC000308	MC0003	08
250	0170	MC000309	MC0003	09
250	0170	MC000310	MC0003	10
250	0170	MC000311	MC0003	11
250	0170	MC000312	MC0003	12
250	0170	MC000313	MC0003	13
250	0170	MC000314	MC0003	14

1 (0 out of 864 Selected)

Cambridge_Blocks_1960 Cambridge_Tracks_1960

766672.628 2957698.395 Feet

Fig. 2.3. ArcGIS analysis- Spatial Overlay

2.4. DSSDI Geodatabase Delhi

Delhi State Spatial Data Infrastructure (DSSDI) Project created a GIS Base Map of Delhi at the scale of 1:2000 that mapped all building parcels and household units. The feature table of this entire Map has been exported for data analytics. The base map provides spatial features like building units, roads, railway lines, streets etc. through the use of overlay functions an aggregate visualisation of a geographic region is provided. The process of creation of this Geodatabase has been given below through a diagram.

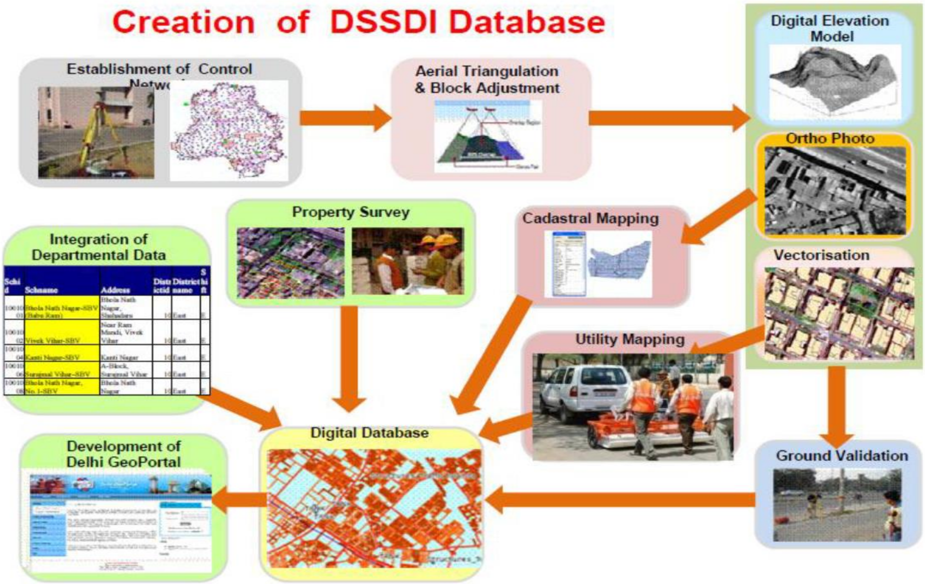

Creation of DSSDI Database
Establishment of Control Network
Aerial Triangulation & Block Adjustment
Digital Elevation Model
Ortho Photo
Vectorisation
Cadastral Mapping
Property Survey
Integration of Departmental Data
Utility Mapping
Digital Database
Development of Delhi GeoPortal
Ground Validation

Figure.2.4. Creation of DSSDI Database

CHAPTER III
DESIGN AND METHODOLOGY

3.1. Methodology

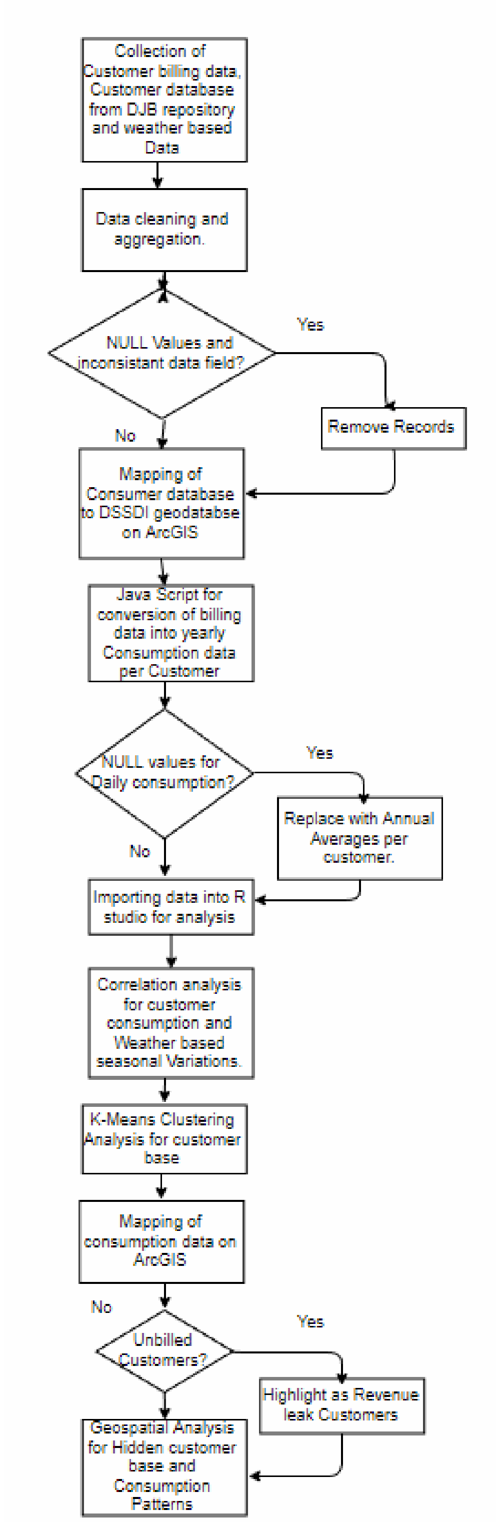

Collection of Customer billing data, Customer database from DJB repository and weather based Data
Data cleaning and aggregation.
NULL Values and inconsistant data field?
Yes
Remove Records
No
Mapping of Consumer database to DSSDI geodatabse on ArcGIS
Java Script for conversion of billing data into yearly Consumption data per Customer
NULL values for Daily consumption?
Yes
Replace with Annual Averages per customer.
No
Importing data into R studio for analysis
Correlation analysis for customer consumption and Weather based seasonal Variations.
K-Means Clustering Analysis for customer base
Mapping of consumption data on ArcGIS
No
Unbilled Customers?
Yes
Highlight as Revenue leak Customers
Geospatial Analysis for Hidden customer base and Consumption Patterns

Figure 3.1. Adopted Methodology

The proposed methodology has been explained as below.

1) Obtaining the Data

The data has been obtained from Delhi Jal Board. The data collected through monthly consumption and billing cycles has been stored on publically servers at NIC Delhi. The DSSDI Base Map has been accessed through on ArcGIS software, access granted through Delhi Jal Board. Data corresponding to weather i.e. Max temperature, Average Temperature and Minimum Temperature and Precipitation levels are collected on daily basis. The master consumer table is as below.

Attribute Name	**Description**
Zone	The entire NCT of Delhi has been divided into 4 major zones i.e. east, west north and south. It is further followed by an area based division of each zone.
Location	Location defines the locality within the defined Zone
MRCODE	Meter reading code is the code assigned to each locality depending upon the metering activity and related area division
AREACODE	Area Code is unique code assigned to each billable area.
K No. CATERGORY	Connection has been divided into different categories on the basis of urban/rural locality
K No.	Unique number assigned to every water meter connection allotted to the household.
K. No. NAME	The name of the resident.
K. No. ADDRESS	Address location of the resident
OldWCNumber	Old Water Connection Number, an earlier used scheme for connection identification.

Table 3.1.1. Master Consumer Table

The dataset for Consumers Billing Cycles. Every cycles generates an excel file for a specific zone area. There are 6 billing cycles in a year.

Attribute Name	**Description**
Zone	The entire NCT of Delhi has been divided into 4 major zones i.e. east, west north and south. It is further followed by an area based division of each zone.
Location	Location defines the locality within the defined Zone
MRCODE	Meter reading code is the code assigned to each locality depending upon the metering activity and related area division
AREACODE	Area Code is unique code assigned to

K No.	Unique number assigned to every water meter connection allotted to the household.
Current Reading	Current reading value (in Kilo Litres)
Previous Reading	Previous Reading Value (in Kilo Litres)
Current Billing Date	Date of latest Billing Reading
Previous Billing Date	Date for previous Billing Reading
Net Bill Amount	Net amount for the consumption

Table 3.1.2. Consumer Billing Cycles

The dataset for daily weather conditions is as follows:

Attribute Name	**Description**
Minimum Temp	Minimum Temperature on a given day in Celsius
Maximum Temp	Maximum Temperature on a given day in Celsius
Average Temp	Average Temperature on a given day in Celsius
Precipitation	Precipitation level in (millilitre) on a given day

Table 3.1.3. Daily weather Forecast

2) **Data Cleaning**

Data is cleaned for inconsistent Connections and spatially unmapped entries. The NULL values are replaced with yearly consumption averages. The yearly data is normalised using a java script. The billing records are converted to customer wise daily consumption. The inconsistent and missing fields are removed for elimination of error in analysis.

3) **Consumer Mapping**

The consumer mapping is carried out using the technique of geotagging. The attributes of DSSDI Base Map are joined with the dataset to convert multiple attributes into single attribute by the name KNO, that identifies an individual billed customer across the entire Delhi region by a unique 8-10 digit Number. This activity is carried out using Excel toolkit. The attributes Zone, Location, MRCODE, AREACODE, OldWCNumber are Inner Joined with the consumer Billing data attributes namely, Zone, Location, MRCODE, AREACODE and KNO. The KNO are new coding scheme that simplifies Data Mapping. The Mapped KNO consist of an individual data layer that can be viewed over the existing Base Map.

4) Normalisation of Data

A java code is used to generate the billing records of every customer into yearly consumption. The pseudocode is given as below.

```
Step 1: Read KNO from Master database.
Step 2: Identify the billing record for identified customer.
Step 3: Calculate the average unit of consumption between the previous and current billing date by calculating the duration of cycle in days and total consumption as per the billing record.
Step 4: Populate the master Consumption table with average values as calculated from previous step.
Step 5: Repeat step 2 to 4 for all records.
Step 6: Repeat step 1 to 5 for all Consumers.
```

5) Importing Data

Data is converted to excel formats and imported into R using data frames.

6) Correlation Analysis

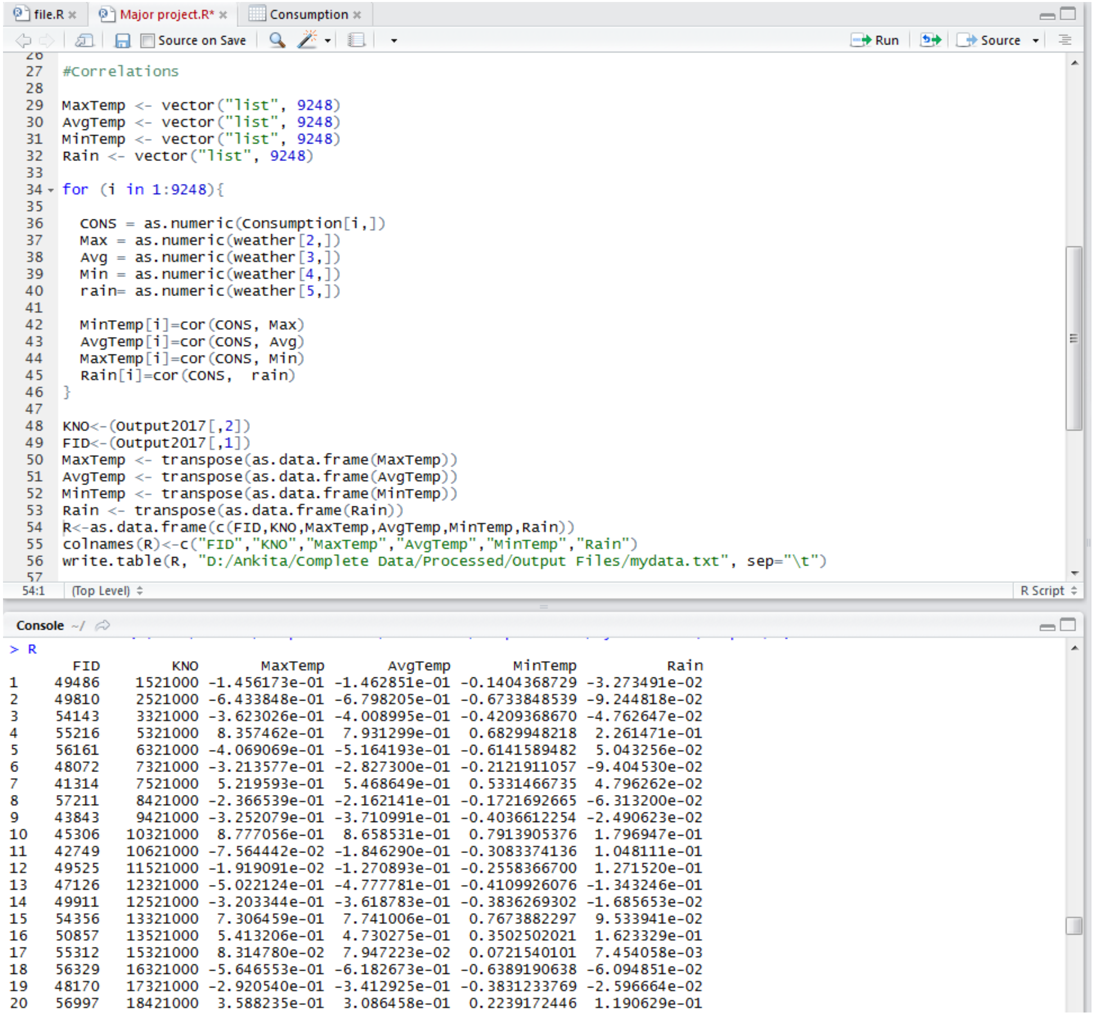
The correlation analysis is carried out, where daily consumption for users is correlated with the Daily temperature and precipitation values in order to find the correlation values of every user towards weather conditions. Correlation is found using Pearson’s correlation coefficient to determine how strongly the values are correlated with seasonal variations. See Fig.

7) K means Clustering Algorithm

K-means clustering algorithm was then run on the data for clustering and analysis. The optimal value of k was found by considering the smallest Davies Bouldin Index value. The clustering results are exported back into the excel file that is further joined with the feature table in ArcGIS for visualising results.

8) Geospatial Analysis

The cluster results are visualised over the DSSDI Base Map for spatial analytics. Hidden customers are identified. Cluster based consumption pattern are also analysed.



```r
#Correlations

MaxTemp <- vector("list", 9248)
AvgTemp <- vector("list", 9248)
MinTemp <- vector("list", 9248)
Rain <- vector("list", 9248)

for (i in 1:9248){

  CONS = as.numeric(Consumption[i,])
  Max = as.numeric(weather[2,])
  Avg = as.numeric(weather[3,])
  Min = as.numeric(weather[4,])
  rain= as.numeric(weather[5,])

  MinTemp[i]=cor(CONS, Max)
  AvgTemp[i]=cor(CONS, Avg)
  MaxTemp[i]=cor(CONS, Min)
  Rain[i]=cor(CONS,  rain)
}

KNO<-(Output2017[,2])
FID<-(Output2017[,1])
MaxTemp <- transpose(as.data.frame(MaxTemp))
AvgTemp <- transpose(as.data.frame(AvgTemp))
MinTemp <- transpose(as.data.frame(MinTemp))
Rain <- transpose(as.data.frame(Rain))
R<-as.data.frame(c(FID,KNO,MaxTemp,AvgTemp,MinTemp,Rain))
colnames(R)<-c("FID","KNO","MaxTemp","AvgTemp","MinTemp","Rain")
write.table(R, "D:/Ankita/Complete Data/Processed/Output Files/mydata.txt", sep="\t")
```

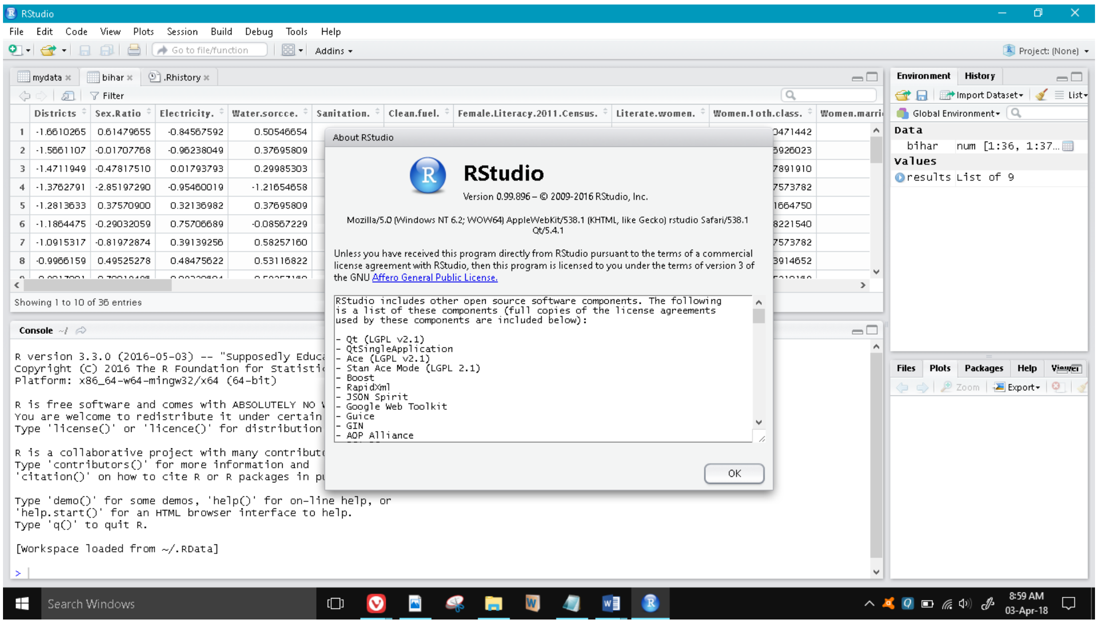
```
> R
      FID        KNO       MaxTemp       AvgTemp       MinTemp          Rain
1   49486    1521000 -1.456173e-01 -1.462851e-01 -0.1404368729 -3.273491e-02
2   49810    2521000 -6.433848e-01 -6.798205e-01 -0.6733848539 -9.244818e-02
3   54143    3321000 -3.623026e-01 -4.008995e-01 -0.4209368670 -4.762647e-02
4   55216    5321000  8.357462e-01  7.931299e-01  0.6829948218  2.261471e-01
5   56161    6321000 -4.069069e-01 -5.164193e-01 -0.6141589482  5.043256e-02
6   48072    7321000 -3.213577e-01 -2.827300e-01 -0.2121911057 -9.404530e-02
7   41314    7521000  5.219593e-01  5.468649e-01  0.5331466735  4.796262e-02
8   57211    8421000 -2.366539e-01 -2.162141e-01 -0.1721692665 -6.313200e-02
9   43843    9421000 -3.252079e-01 -3.710991e-01 -0.4036612254 -2.490623e-02
10  45306   10321000  8.777056e-01  8.658531e-01  0.7913905376  1.796947e-01
11  42749   10621000 -7.564442e-02 -1.846290e-01 -0.3083374136  1.048111e-01
12  49525   11521000 -1.919091e-02 -1.270893e-01 -0.2558366700  1.271520e-01
13  47126   12321000 -5.022124e-01 -4.777781e-01 -0.4109926076 -1.343246e-01
14  49911   12521000 -3.203344e-01 -3.618783e-01 -0.3836269302 -1.685653e-02
15  54356   13321000  7.306459e-01  7.741006e-01  0.7673882297  9.533941e-02
16  50857   13521000  5.413206e-01  4.730275e-01  0.3502502021  1.623329e-01
17  55312   15321000  8.314780e-02  7.947223e-02  0.0721540101  7.454058e-03
18  56329   16321000 -5.646553e-01 -6.182673e-01 -0.6389190638 -6.094851e-02
19  48170   17321000 -2.920540e-01 -3.412925e-01 -0.3831233769 -2.596664e-02
20  56997   18421000  3.588235e-01  3.086458e-01  0.2239172446  1.190629e-01
```

Fig. 3.2 Correlation of User Consumption with seasonal variables.

3.2 Tools and Technologies

3.2.1 R Studio

R Studio is a free and open-source integrated development environment (IDE) for R, a programming language for statistical computing and graphics. R Studio was founded by JJ Allaire. R Studio is written in the C++ programming language and uses the QT framework for its graphical user interface



	Districts	Sex.Ratio	Electricity.	Water.sorcce.	Sanitation.	Clean.fuel.	Female.Literacy.2011.Census.	Literate.women.	Women.1oth.class.	Women.marri
1	-1.6610265	0.61479655	-0.84567592	0.50546654					0471442	
2	-1.5661107	-0.01707768	-0.96238049	0.37695809					5926023	
3	-1.4711949	-0.47817510	0.01793793	0.29985303					7891910	
4	-1.3762791	-2.85197290	-0.95460019	-1.21654658					7573782	
5	-1.2813633	0.37570900	0.32136982	0.37695809					1664750	
6	-1.1864475	-0.29032059	0.75706689	-0.08567229					8221540	
7	-1.0915317	-0.81972874	0.39139256	0.58257160					7573782	
8	-0.9966159	0.49525278	0.48475622	0.53116822					3914652	

Showing 1 to 10 of 36 entries

RStudio

Version 0.99.896 – © 2009-2016 RStudio, Inc.

Mozilla/5.0 (Windows NT 6.2; WOW64) AppleWebKit/538.1 (KHTML, like Gecko) rstudio Safari/538.1 Qt/5.4.1

Unless you have received this program directly from RStudio pursuant to the terms of a commercial license agreement with RStudio, then this program is licensed to you under the terms of version 3 of the GNU Affero General Public License.

```
RStudio includes other open source software components. The following
is a list of these components (full copies of the license agreements
used by these components are included below):

- Qt (LGPL v2.1)
- QtSingleApplication
- Ace (LGPL v2.1)
- Stan Ace Mode (LGPL 2.1)
- Boost
- RapidXml
- JSON Spirit
- Google Web Toolkit
- Guice
- GIN
- AOP Alliance
```

```
R version 3.3.0 (2016-05-03) -- "Supposedly Educa
Copyright (C) 2016 The R Foundation for Statisti
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO W
You are welcome to redistribute it under certain
Type 'license()' or 'licence()' for distribution

R is a collaborative project with many contribut
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in pu

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Workspace loaded from ~/.RData]

>
```

```
Data
  bihar     num [1:36, 1:37...
Values
  results   List of 9
```

Fig 3.3: R Studio

The R language is extremely efficient for churning out large quantity of data and requires relevant representation and illustration of analysis. R Studio is being used for industry and research experts for performing data mining and analytics processes. R studio makes the large quantity of user consumption based data processing easier and quicker hence it is ideally the best choice for this project.

3.2.2 ArcGIS Software

ArcGIS is a geographic information system (GIS) for working with maps and geographic information. Arc GIS software is primarily used for generation of maps, analysis of data as features over map, discovery of geographic patterns and information for a range of applications and databases.

The system provides an infrastructure for making maps and geographic information available throughout an organization, across a community, and openly on the Web.

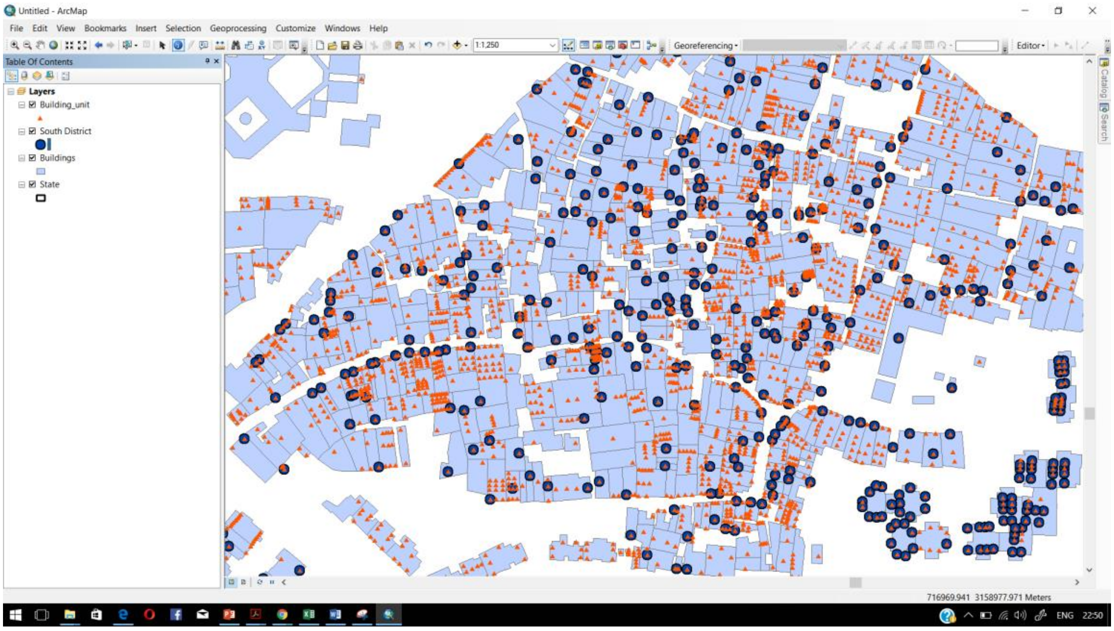


Figure 3.4 ArcGIS 10.5

3.3 System Specifications

- Operating System: Windows 10
- Processors: Intel core i5
- RAM: 8 GB DDR3
- R Studio version 0.99.896
- ESRI ArcMap 10.5
- Eclipse Oxygen

CHAPTER IV
EXPERIMENTATION AND RESULTS

4.1. Clustering Analysis

The K-Means Clustering yeileds 3 optimal clusters for the data set of customers compimisng of following parameters.

Attribute Name	Description
K No.	Unique number assigned to every water meter connection allotted to the household.
MinTempCoeff	Correlation Coefficient of yearly consumption with Minimum Temperature
AvgTempCoeff	Correlation Coefficient of Yearly consumption with Average Temperature
MaxTempCoeff	Correlation Coefficient of Yearly consumption with Maximum Temperature
D1, D2,…D365	Attributes that store the daily consumption of a customer across 365 days of the year. Each day is numbered through D1, D2, D3…D365.

Table 4.1.1 Clustering Dataset

The clustering results are dipicted as follows. Customers on X axis are plotted against the variation of Correlation Coefficients.

4.1.1. Optimal Value of K

An optimal value of Davis Bouldi index is identified in order to find the optimal number of clusters K configuration. The clustering algorithm was executed multiple times with different number of seed points or K values and the Davis Bouldin Index was noted for every configuration. Davies Bouldin index is a measure of how well the clustering has been executed over the given dataset. As the DB value is closer to 0 (negative or positive), more optimal is the clustering outcome from the dataset. The optimal value of K can be found by considering the highest Davies Bouldin Index which is found using the equation as below:

$$\text{DB} = \qquad \text{DB} = \frac{1}{N}\sum_{i=1}^{N} Di$$

Here, the number of clusters is given by N, Di is cluster distance ratio i.e. distance between points and centroid in that cluster, for all the clusters in the data set. The observed values of DB index for different values of K are tabulated as below.

OPTIMAL DB INDEX	-0.565	**-0.561**	-0.651	-0.686	-0.910
K	K=2	**K=3**	K=4	K=5	K=10

Table 4.1.2 Optimal DB Index

Further the 3 cluster orientation is subjectively justified as the customers follow 3 different patterns of usage, where they may be strongly positive correlated with the seasonal variations of weather, strongly negatively correlated with the seasonal variations in weather or slighly or not correlated with the seasonal variations at all. These classifications can be furtehr studies in detail using geospatial analytics further.

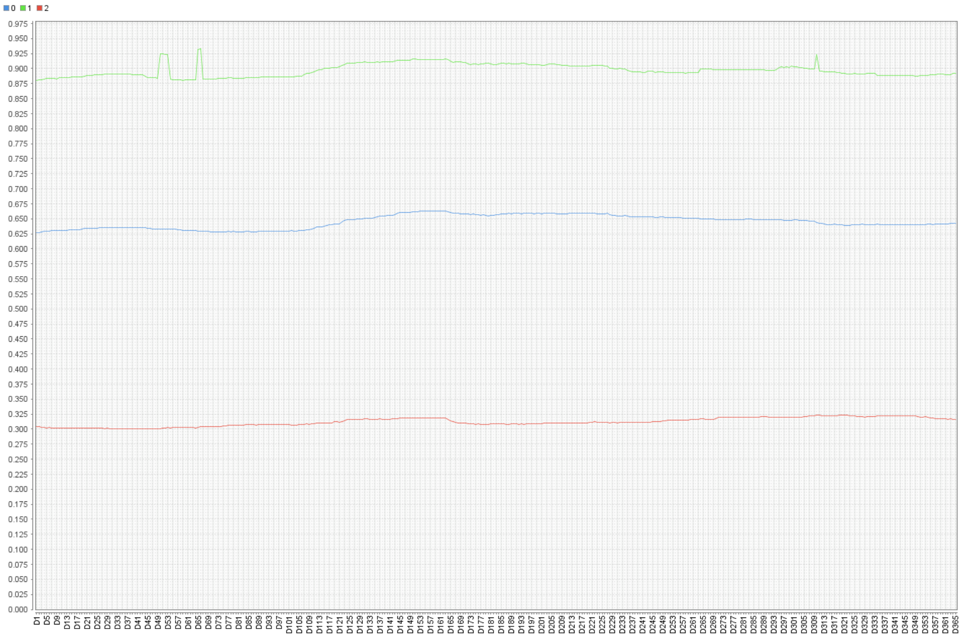


Figure. 4.1. Cluster wise Daily Consumption v/s Coefficients of correlation pattern

The clustering analysis is carried out and the line graph for different clusters against daily consumption and coefficient values has been depicted as below. The graph follows as clear distinction between the clustered customers that fall under clearly defined beahviours of correlation patterns with seasonal changes.

4.1.2. Effect of daily Temperatures.

The clustering shows that the influence of daily minimum temperatures are prominent on the clustering orientation. Cluster 0 shows a slight negative or positive correlation of consumption to the minimum temperature. The Coefficient of correlation ranges from -0.35 to 0.4. Majority of the consumers falling into this cluster are independent of the seasonal demand surges and continue to consume an average pattern of units annually.

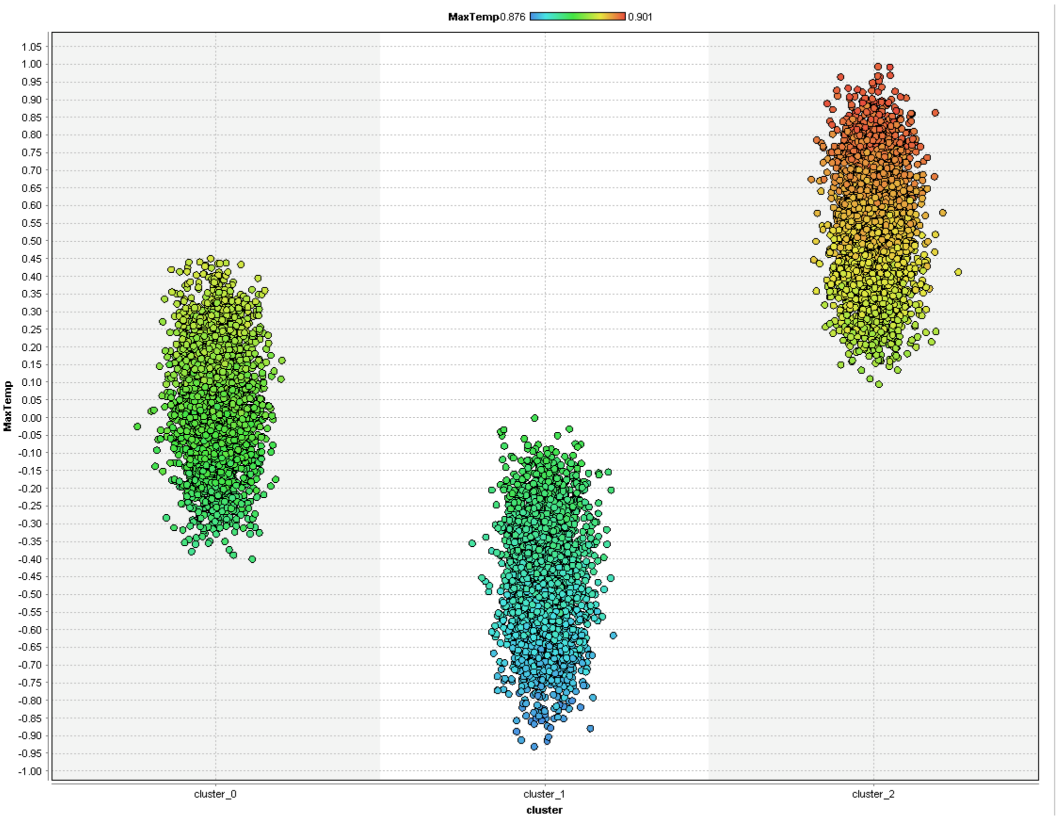


Figure 4.2. Cluster of Customers plotted agaianst the coefficient of correlation with Maximum Temperature

The cluster 1 have consumers that have highly negative correlation pattern with the seasonal variations of temperature. As the temperatures surges their demand pattern decreases which can be due to various reasons and choices like use of electrical cooling devices instead of water consuming coolers. Also another prominent reason could be extension of consumption through multiple connections thus reducing the overall consumption per connection. The analysis per connection can be highlighted through the geospatial analysis.

The cluster 2 have highly positive correlations to the temperature. These consumers are highly affected by the rising and falling temperatures. Their primary demand for water consumption lies for cooling and hydration during summers. A surge in temperatures linearly increases the demand for water, while as temperatures drop their demand is also reduced. Such customers fall in high potential household based customers. They form the base for potential revenue generation.

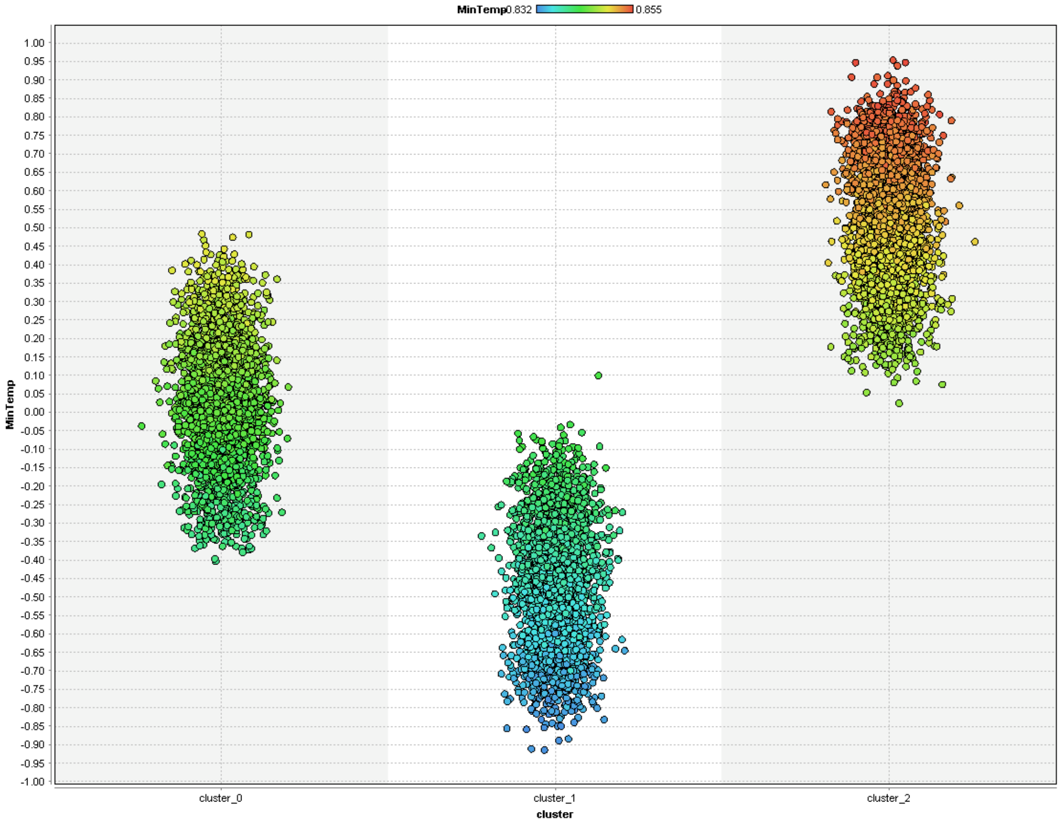


Figure 4.3. Cluster of Costumers plotted agaianst the coefficient of correlation with Minimum Temperature.

The pattern for temperatures across the clusters remain similar in trend for minimum, average as well as maximum temperatures. Except for the consumers show high correlation values in case for maximum temperature. The clusters spread out evenly in case of minimum temperature as different consumers show different level of affinity towards water consumption based on personal needs.

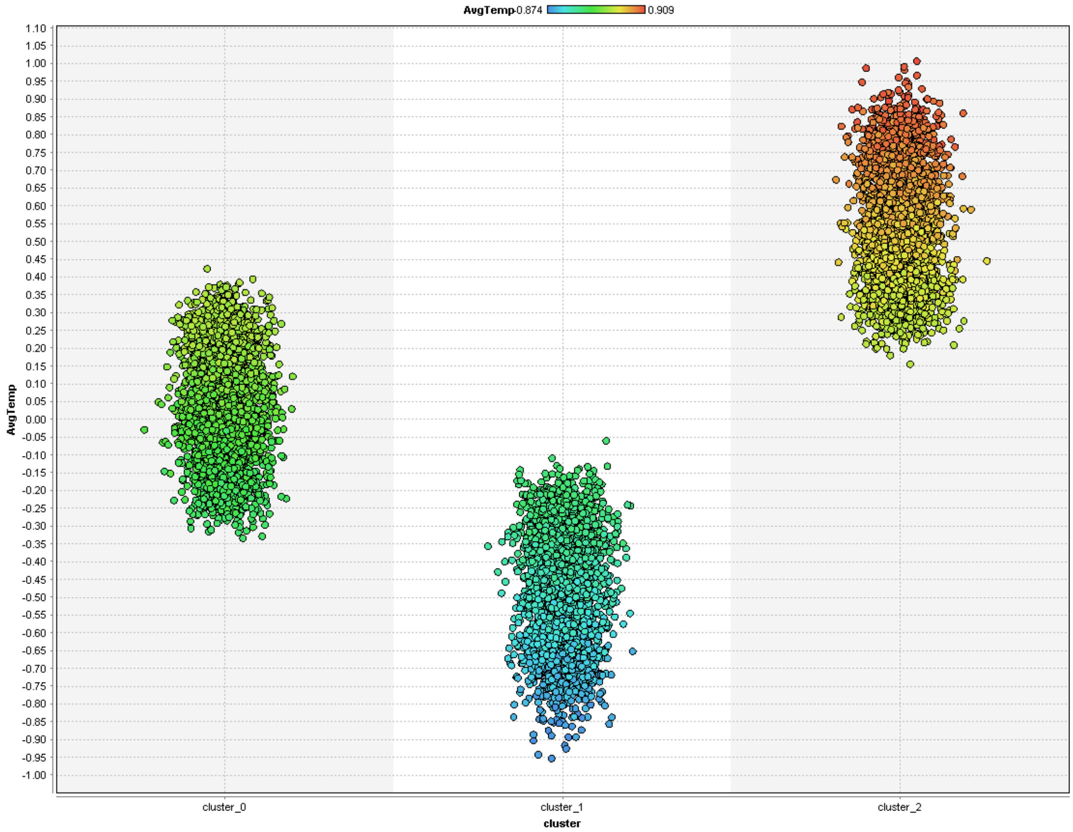


Figure 4.4. Cluster of Customers pllotted agaianst the coefficient of correlation with Average Temperature.

4.1.3 Effect of Precipitation levels.

The effect of precipitation follows the trend of temperatures to slightly significant level. The cluster 0 consumers do not show major consumption dependence on the level of precipitation, ranging the coefficient of relation from -0.2 to 0.2. The cluster 1 shows slight negative correlation patterns towards precipitation levels. This may hint towards rain water harvesting and switching to natural fresh water alternatives instead of piped supply. The cluster 2 show slight positive correlation trend to the precipitation levels. Tis trend can be analysed as the commercial consumers with some dependence on precipitation level for their consumption.

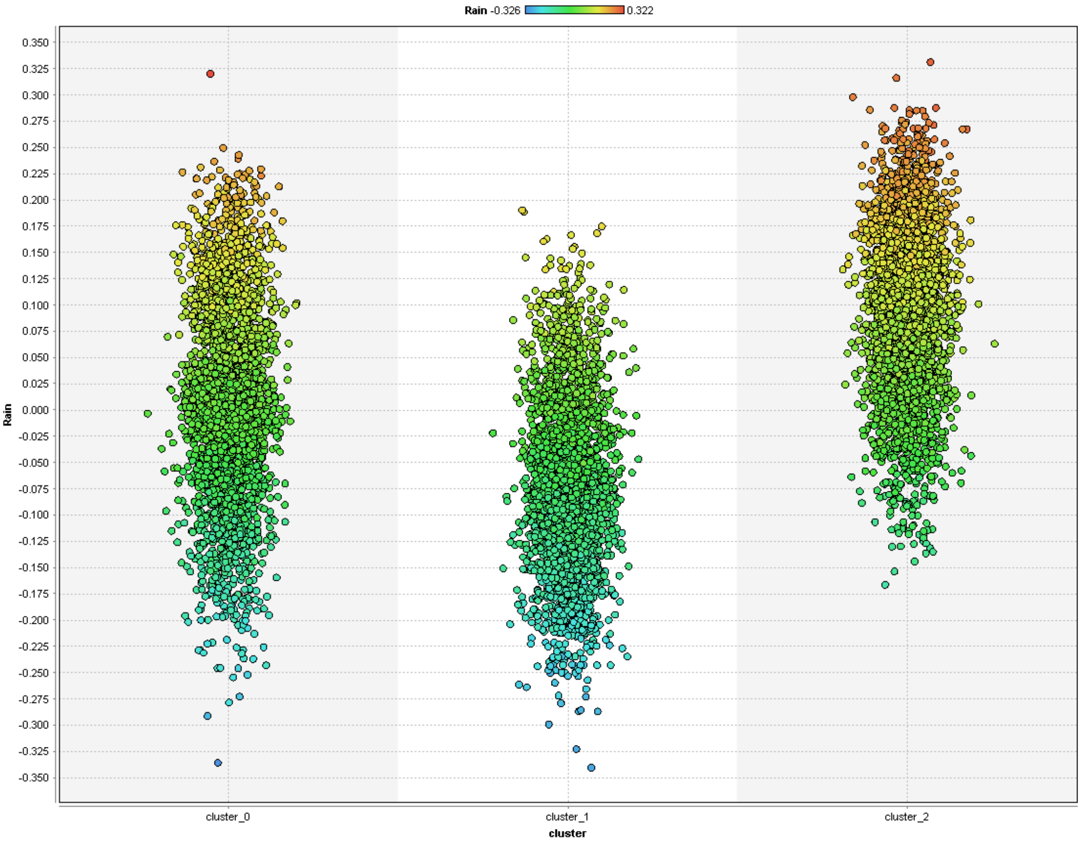


Figure 4.5. Cluster of Customers plotted against the coefficient of correlation with precipitation level.

Cluster	Coefficient of Correlation	Discussion
Cluster 0	Invariant or near to 0	The consumer demand is not influenced by seasonal weather variations and the customers are least affected by the seasonal demand surges.
Cluster 1	Strong Negative	These consumers are affluent consumers who are not dependent on water based cooling solutions such as water coolers, sprinklers etc. They migrate towards use of Electrical cooling systems or temporaryly migrate to other geographies.
Cluster 2	Strong Positive	The consumer demand show high dependence towards the weather variations. They are most affected by the seasonal demand surges. These customers are to be targeted towards water conservation schemes and techniques.

Table 4.1.3 Cluster Analysis.

4.2. Geospatial Analysis.

The geospatial analysis involves mapping of 10,000 customers on the Base Map, That involves 2165 buildings in the R.K. Puram area of Delhi. The number of building units or number of independent families risiding per building is denoted by building unit that varies from 1 in small households to over 20 in commercial residential areas.

4.2.1. Identification of Potential customer base and Exploiting customer base

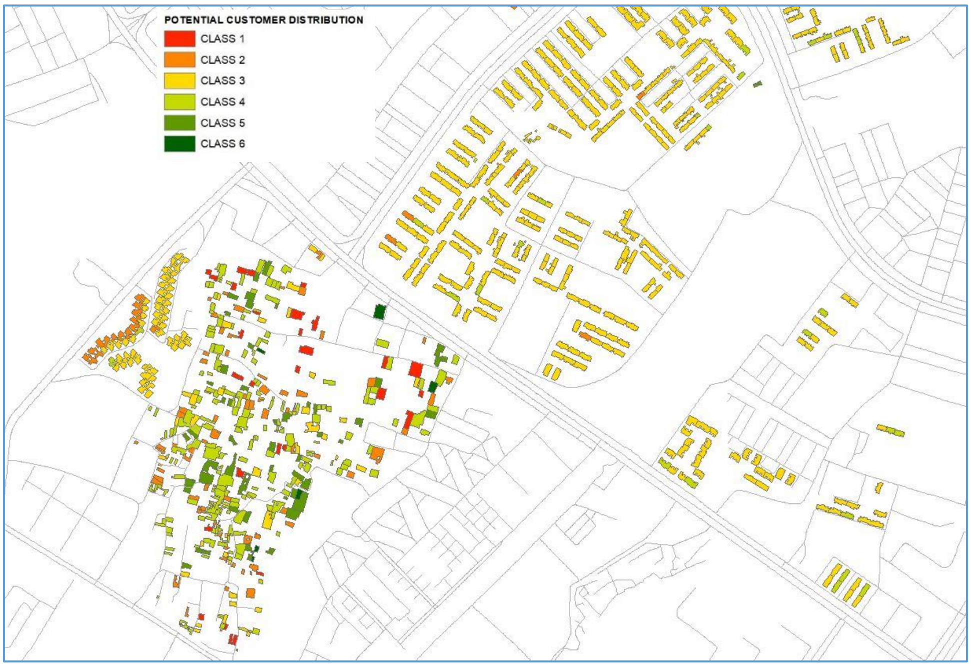
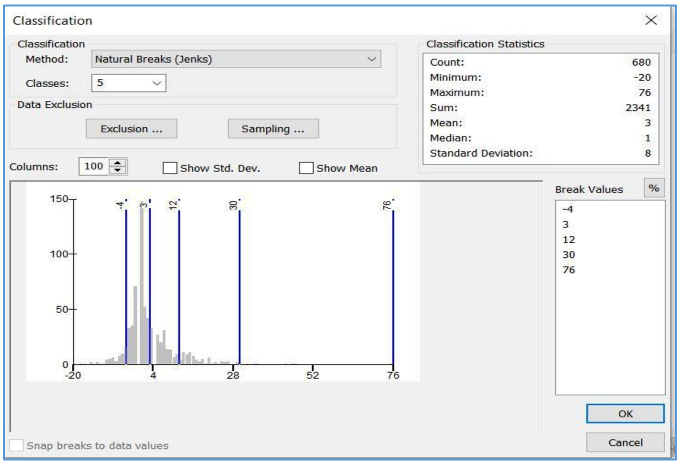


Figure. 4.6. Geospatial Classification of Customer base

The building units layer is projected over the consumer connection mapping. This gives potential customers and the consumer over exploitation. The difference of building unit and number of connection within per building is identified. The values thus obtained are classified using classification techniques using natural breaks (Jenks).



Classification

Classification

Method: Natural Breaks (Jenks)

Classes: 5

Data Exclusion

Columns: 100

□ Show Std. Dev.

□ Show Mean

Classification Statistics

Count:	680
Minimum:	-20
Maximum:	76
Sum:	2341
Mean:	3
Median:	1
Standard Deviation:	8

Break Values

-4
3
12
30
76

□ Snap breaks to data values

Figure 4.7. Classification using natural breaks (Jenks)

The classes are colour coded and projected on the BaseMap. The summarised clasification is as follows.

Class	Range	Building Count	Potential Number of Customers	Monthly Average Consumption (Kilo litres)	Interpretation
1	-20 – (-5)	46	NA	21.285	These categories denote extremely high number of connections per building units. This result in exploitation of water beyod the avergage family requirements. This further indicates loss of revenue due to distribution of per consumer consumption across multiple unethically gained connections.
2	-4 –(-1)	155	NA	20.390	These consumers are small households with justified 1:1 proportion of connections to building unit. There are 0-3 deviations for number of unethically gained connections.
3	0	1485	0	17.554	This class signifies ideal consumption scenario with 0 potential for expansion.

4	1-12	399	1418	19.89	These building units show a significant gap of an average of 8 extra building units than the number of connections acquired. This shows possibility for new customer base and increased potential revenues.
5	13-30	73	1313	23.02	These building units have accessively hight gaps in the number of connections and building units. These require improvised planning schemes for commercial planning.
6	31-76	7	314	19.588	These denote extreme hidden revenue situation with 50 or more hidden connections per building establishment.

Table 4.2.1 Summarising Potential and Exploiting Customer base.

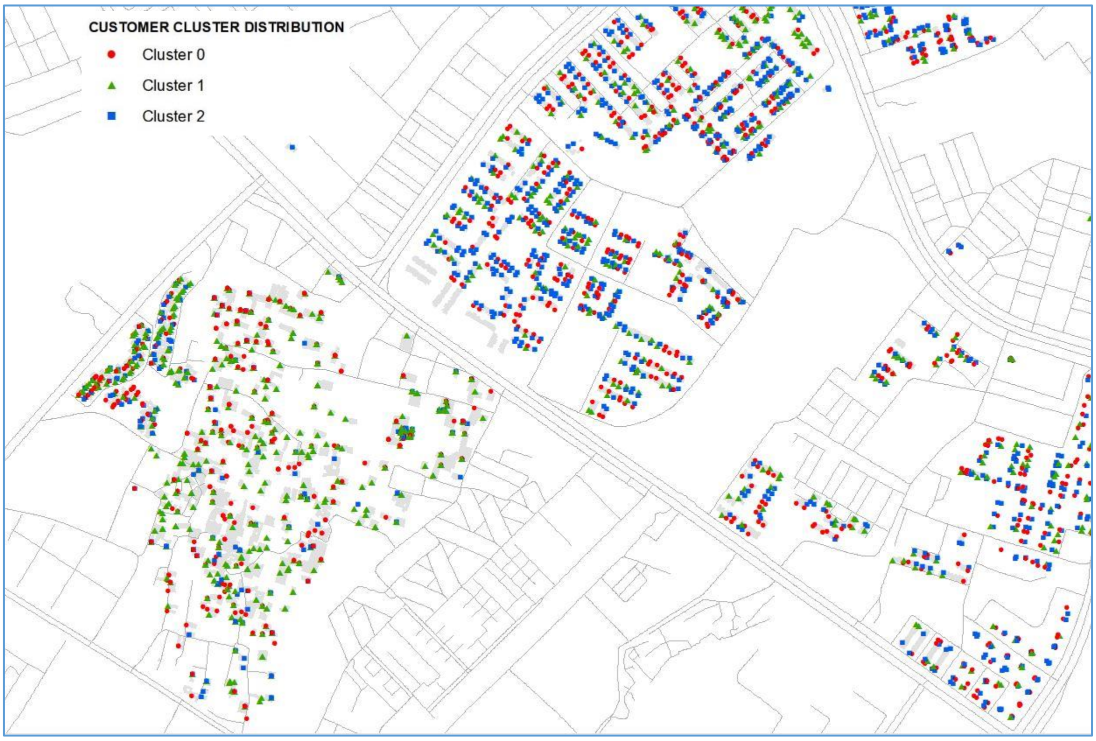


Figure 4.8. Distribution of Consumer Clusters

The distribution of clusters across the geography in figure 4.9. reveal patterns that show that for specific areas the consumer cluster 1 is prominent. As analysed from the basemap these areas are unplanned residential areas. The other part of the geography shows planned city and cluster distribution of customers from cluster 2. The distribution of consumers of cluster 0 is uniform across the entire geography.

Overlapping of Analytical clusters with geospatial classification results give significant findings as follows.

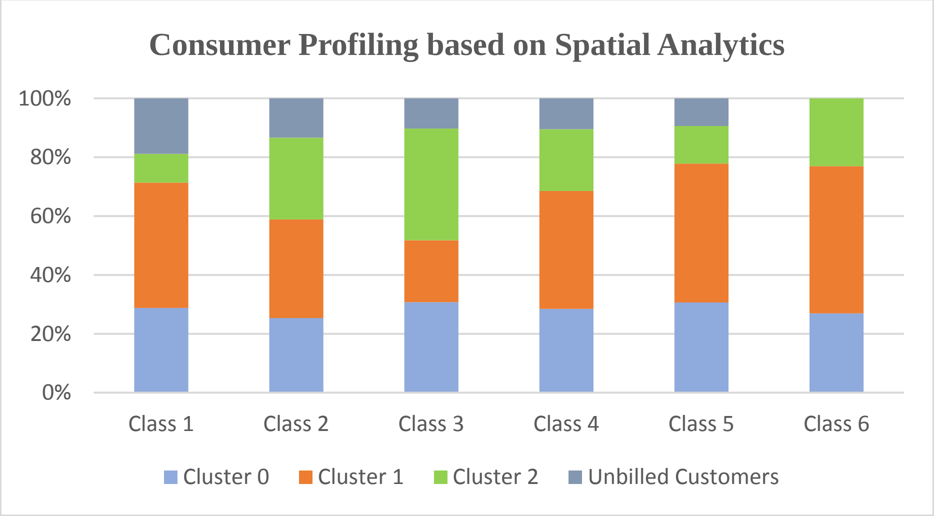


Figure 4.9. Consumer Profiling based on Spatial Analysis.

Potential Class	Cluster 0	Cluster 1	Cluster 2	Unbilled Consumers
1	197	291	67	129
2	220	290	241	116
3	2037	1393	2512	680
4	425	598	313	157
5	65	100	27	20
6	7	13	6	0

Table 4.2.2 Consumer Profiling distribution based on Spatial Analysis.

1) The class 1 and 2 consumers overexploiting the resource through excessive number of connections do not face seasonal variations and demand surges for water. The consumption correlation for majority of these consumers follows negative trend where seasonal spikes in demand do not affect these consumers. They rely on multiple connections to fulfill their demand. Also this results in loss of revenue due latency of high consumers amidst low resential consumers. These customers are exploiters in nature. This customer base also highlights low billing efficincy and posibility of increased revenues.

2) The class 3 consumers form the base of ideal consumers that have 1:1 connection to family Ratio. These consumers show demand dependence on seasonal variation with positive trends. These areas face acute shortage of supply during summers and consumption is directly correlated to the coefficient of correlation of weather conditions. These households can be targetted for water harvesting and conservation schemes.
3) The class 4 and 5 consuming groups of consumers show high negative correlation towards the seasonal changes and consumption. These users depict potential for high expanse in customer base and their primary sources depend on alternate sources like groundwater or portable water.

4.2.2. Identification of Customer Billing Efficiency

The hidden customer base is identified as the consumers residing in a speciafic geography and are not billed for the utility services that they consume. The consmers are identified from the annual billing data and are mapped on basemap from the master table of consumers.

The billig efficiency is defines as the percentage of total customers that are billed on regular basis. The unbilled customers are either charged on lumpsum basis or are not contributing anything to the revenue. The billing efficiency for the current customer base works out as follows:

Total Customers	10436
Billed Customers	9248
Unbilled Customers	1190
Billing Efficiency	**88.60%**

Table 4.2.3 Billing Efficiency

As observed from the spatial distribution the unplanned region have low billing efficiency as compared to planned residential areas. This requires efficient measures to be taken in order to comensate for the revenue loss in rural/unplanned area.

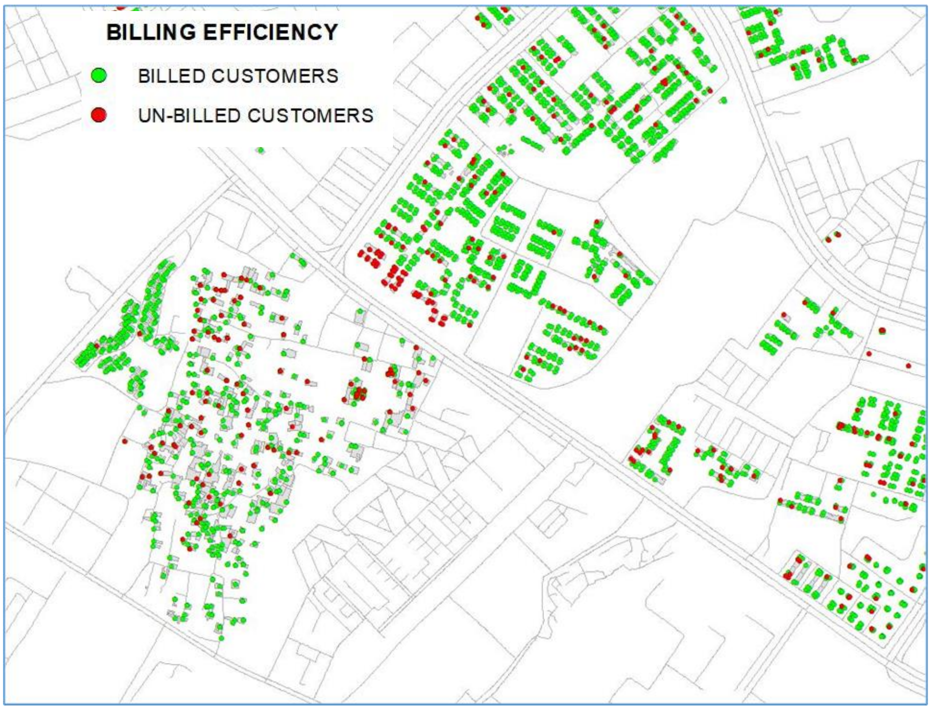


Figure 4.10. Distribution of Billed/ Unbilled customers.

CHAPTER V
DISCUSSION

5.1. Proposed Policy Implementations

The water is a scarce resourse and is depleting day by day, requiring utilities to constantly evolve a new policy measures aiming at conservation of water by its customers at the same time enhancing its revenue to maintain sustainability of its operations.

Customer profiling provies an efficient ground for formulating new policies targetting specific customer groups to achieve tradeoff between demand management and revenue enhancement.

The tarrif structure follows the principle of “use more-pay more” and is based on telescopic model where the rate for higher slabs of consumption increases telescopically. However this system poses a challenge where customers tend to gain by taking mutiple water connections thereby reducing consumption per connection and moving towards lower slabs resultsing in revenue loss in the utilities.

1) The potential areas where such customers exist in the system have been identified and categorized as class 1 and class 2. Further these customer groups are also overexploiting and potentially misusing the water use. Therefore the policy can be formulated to enforce regulations on these customer groups to surrender multiple connections. Alternatively water supply can be restricted to these areas based on designated per capita water use.

2) The potential classes 4 and 5 consists of those areas where alternate sources of water such as groundwater are being used and there is scope for increasing cutomer base by adding new customer to the existing network of Customers.

   Policey for extending services by providing new water connections in these target groups can be framed which will result in enhanced revenue for the utility.

CHAPTER VI
CONCLUSION & FUTURE PROSPECTS

6.1. Conclusion

The purpose of data collection and its statistical analysis is to reach (closest to) the truth. While there might be many software available for basic data analysis, each one has its own attributes and pros and cons. Analysis of the big data set as a whole has the limitation of having extremes in the data, which produces fallacious results on analysis, thus deviating from the truth. The technique of clustering of a big data set into smaller groups based on common attributes and their subsequent analysis helps overcome this limitation. In the present study clustering enabled us to understand the real correlations seasonal variations and its corresponding effect on the consumption and demand of water utility in the state of Delhi. This would help the policy makers understand their customer base better and target policies for revenue according to the user profiling. This can improve the demand supply relations and equity of reach of utility to each and every household.

Also through this research, the geospatial analysis concludes the extent of mismanagement of water connections and utility across the city. It helps to summarise and overlap different visual data to aggregate the various parameters and provide practical industry oriented results for identification of cause of exploitation.

6.2 Future Prospects

The project further expands the scope of work to GIS Spatial data mining techniques. A parallel processing and big data solution can also be foreseen as an extension of work. The expected results and outcome of the analysis will be documented in the form of an analytical report that will be useful for establishing analytical understanding of demand-supply-consumption patterns and amend and refine policies for water utility consumption and revenue generation.

REFERENCES

[1] Geoffrey K. F. Tso, Kelvin K.W. Yau, Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks, Elsevier, 2005

[2] Giulia Romano, Andrea Guerrini, "Measuring and comparing the efficiency of water utility companies: A data development analysis approach", Elsevier, 2011

[3] Chrysi Laspidoua, Elpiniki Papageorgiou, Konstantinos Kokkinosb, Sambit Sahud, Arpit Guptae, Leandros Tassiulasf, "Exploring patterns in water consumption by clustering", Elsevier 2015

[4] Harris F. Seidel, "A Statistical Analysis of Water Utility Operating Data for 1965 and 1970" (Journal (American Water Works Association), Vol. 70)

[5] Harris F. Seidel Thea, "Water Utility Operation Data: An Analysis", Journal, American Water Works Association

[6] Alexander Lavin, Diego Klabjan, "Clustering Time-Series Energy Data from Smart Meters", Elsevier, 2016

[7] Hongfei Li, Dongping Fang, Shilpa Mahatma, Arun Hampapur, "Usage Analysis for Smart Meter Management",IEEE- 2011.

[8] Pengcheng Zhang, Jerry Gao, A. G. Thomas, K. P. Alagupackiam, K. Mannava, P. I. Bosco, and Sen Chiao, "On Building a Big Data Analysis System for California Drought" , IEEE 2017

[9] Fernando Arbués, Mar´ıa Ángeles Garc´ıa-Valiñas, "Estimation of residential water demand: a state-of-the-art review", Roberto Mart´ınez-Espiñeira, (2003)

[10] Utilities and Big Data: Using Analytics for Increased Customer Satisfaction, Retrieved from www.oracle.com on 23rd October 2013.

[11] X. Chen, D. Naresh, L. Upmanu, Z. Hao, L. Dong, Q. Ju, J. Wang, and S. Wang, "China's water sustainability in the 21st century: a climate informed water risk assessment covering multi-sector water demands" , (Copernicus Publications on behalf of the European Geosciences Union)

[12] Giorgos Giannopoulos, Sophia Karagiorgou, Yannis Kouvaras, Michalis Alexakis, Pantelis Chronis, "Consumption Analytics and Forecasting Engine", DAIAD Research project, E.C.

[13] Ramiro Vega, Managing Water Utilities With Geographic Information Systems: The Case Of The City Of Tampa, Florida, (Student thesis, 2009)

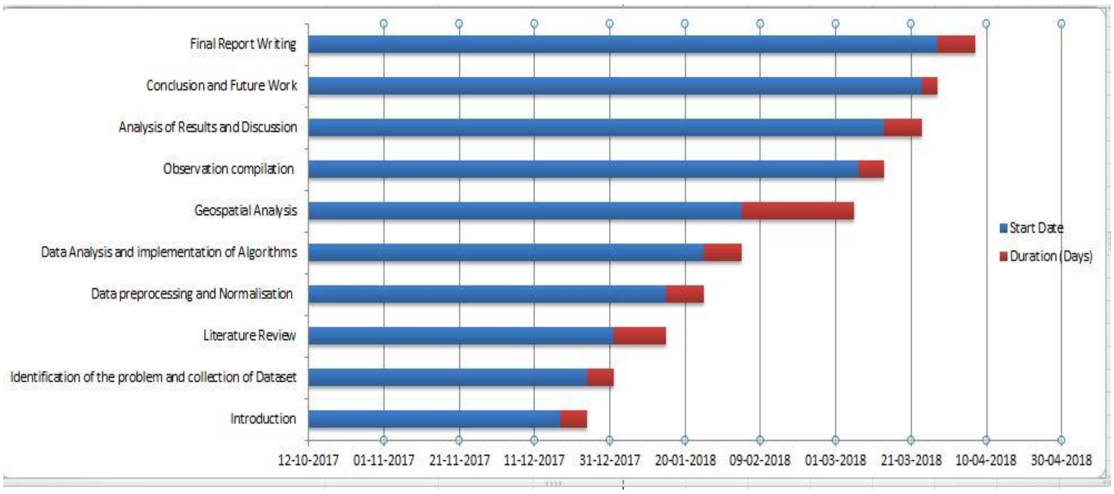
[14] Xiufeng Liu, Lukasz Golab and Ihab F. Ilyas, "SMAS: A Smart Meter Data Analytics System", IEEE 2011

[15] Asnashari, A. and Shahrour, I., 2007. Geostatistical analysis of water mains failure: A case study from Iran. Water Asset Management International, 3(3), 8–13.

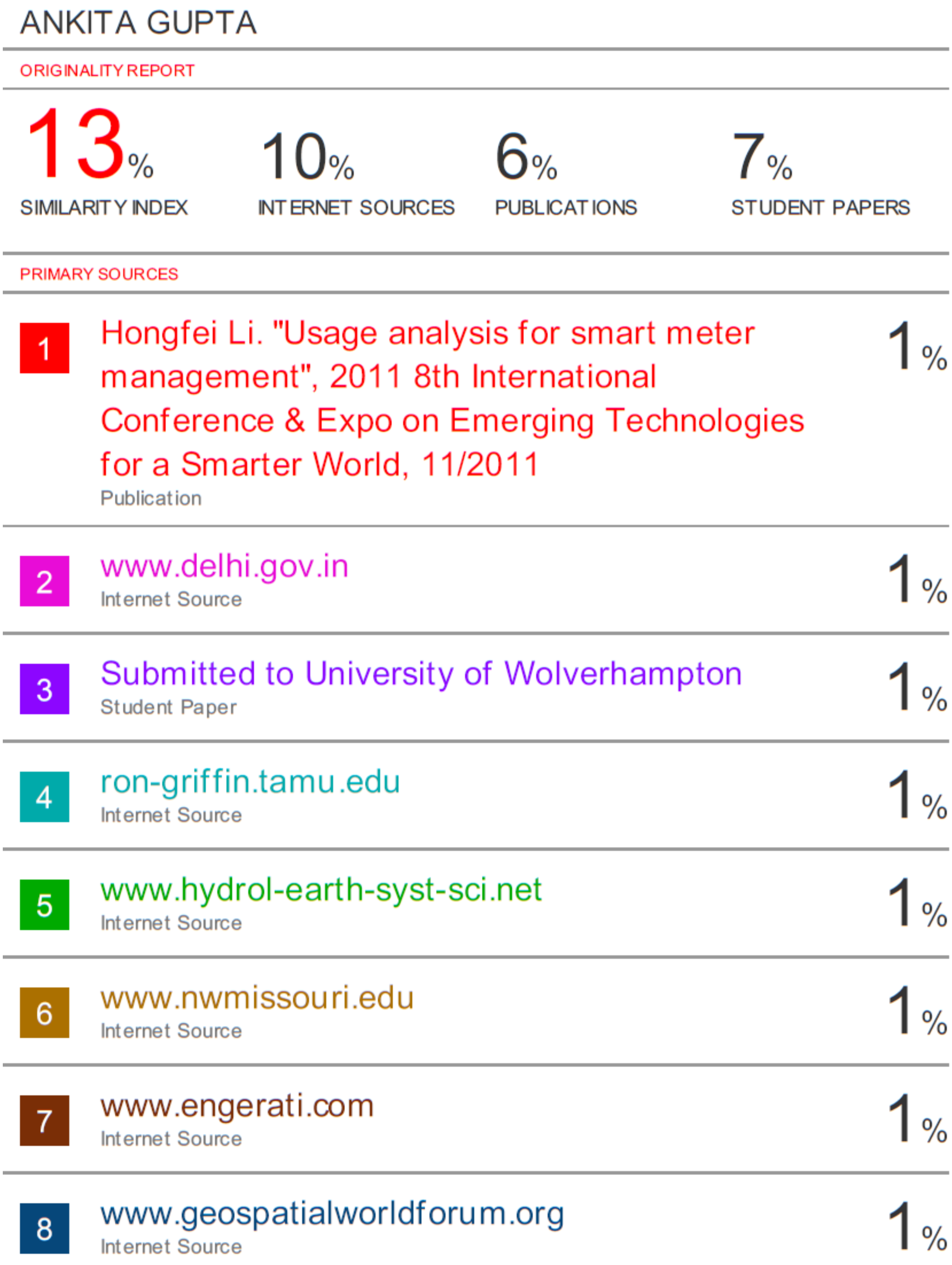
APPENDIX

(A) GANTT. CHART

Activity	Duration (Days)	Start Date	Finish Date
Introduction	7	18-12-2017	24-12-2017
Identification of the problem and collection of Dataset	7	25-12-2017	31-12-2017
Literature Review	14	01-01-2018	14-01-2018
Data pre-processing and Normalisation	10	15-01-2018	24-01-2018
Data Analysis and implementation of Algorithms	10	25-01-2018	03-02-2018
Geospatial Analysis	30	04-02-2018	06-03-2018
Observation compilation	7	07-03-2018	13-03-2018
Analysis of Results and Discussion	10	14-03-2018	23-03-2018
Conclusion and Future Work	4	24-03-2018	27-03-2018
Final Report Writing	10	28-03-2018	06-04-2018



(B) Plagiarism Report



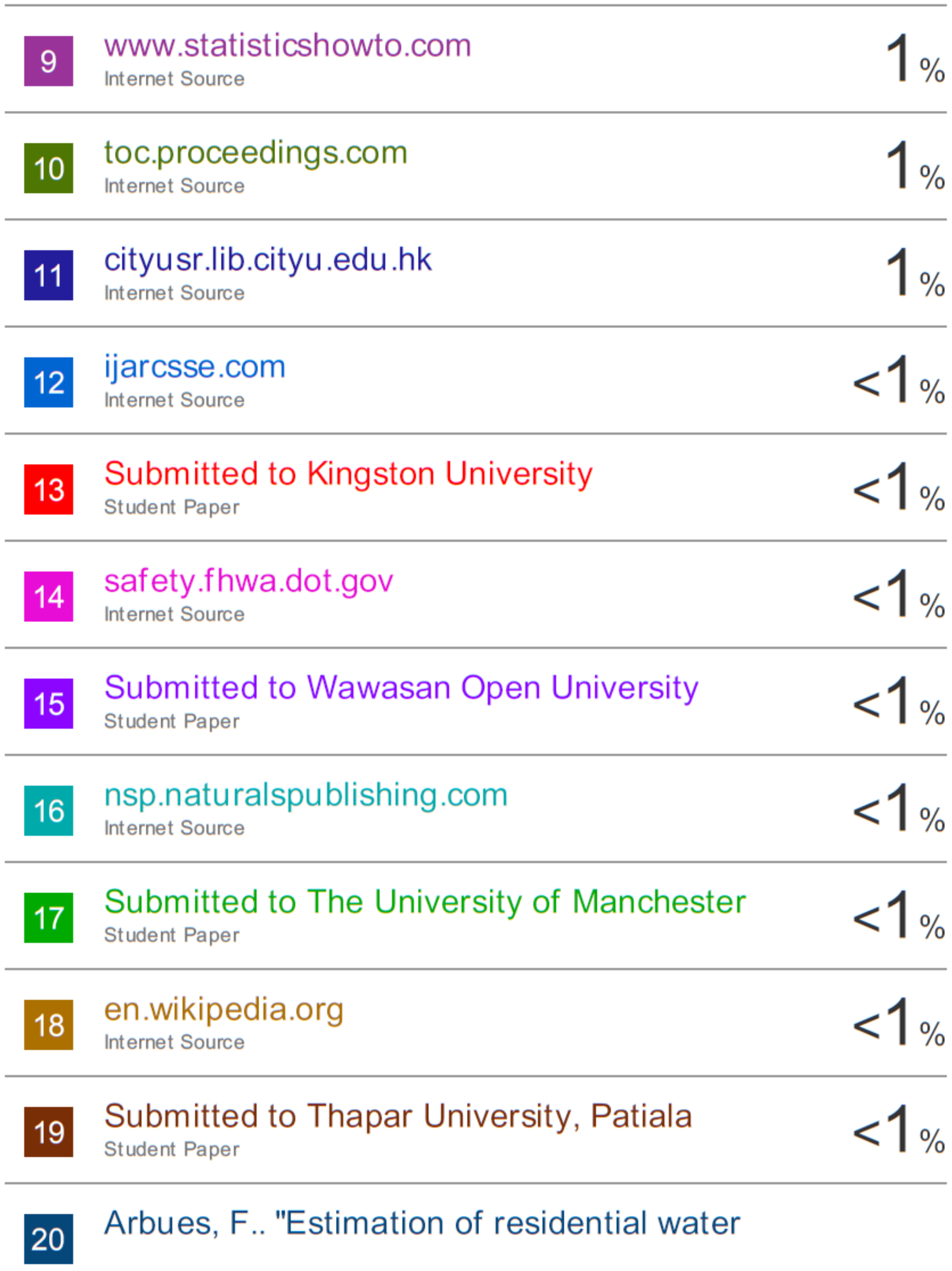
ANKITA GUPTA

ORIGINALITY REPORT

13%	10%	6%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Hongfei Li. "Usage analysis for smart meter management", 2011 8th International Conference & Expo on Emerging Technologies for a Smarter World, 11/2011 Publication	1%
2	www.delhi.gov.in Internet Source	1%
3	Submitted to University of Wolverhampton Student Paper	1%
4	ron-griffin.tamu.edu Internet Source	1%
5	www.hydrol-earth-syst-sci.net Internet Source	1%
6	www.nwmissouri.edu Internet Source	1%
7	www.engerati.com Internet Source	1%
8	www.geospatialworldforum.org Internet Source	1%



9	www.statisticshowto.com Internet Source	1 %
10	toc.proceedings.com Internet Source	1 %
11	cityusr.lib.cityu.edu.hk Internet Source	1 %
12	ijarcsse.com Internet Source	<1 %
13	Submitted to Kingston University Student Paper	<1 %
14	safety.fhwa.dot.gov Internet Source	<1 %
15	Submitted to Wawasan Open University Student Paper	<1 %
16	nsp.naturalspublishing.com Internet Source	<1 %
17	Submitted to The University of Manchester Student Paper	<1 %
18	en.wikipedia.org Internet Source	<1 %
19	Submitted to Thapar University, Patiala Student Paper	<1 %
20	Arbues, F.. "Estimation of residential water	

COMMENT BY EXTERNAL EXAMINER

(Page to be added at the end of the Project Report)

Name of the External Examiner:

Signature: