JADAVPUR UNIVERSITY
Department of Computer Science and Engineering
Kolkata,West Bengal,India
2022

Matrix profile based analysis of repeated patterns of pollutants

by
Animesh Adhikari
REGN. NO. -154148 OF 2020-2021
EXAM ROLL NO. -M4CSE22024

under the supervision of

Dr. Sarbani Roy
Professor

*Thesis submitted in partial fulfillment of requirements
for the degree of*

MASTER OF COMPUTER SCIENCE OF ENGINEERING
OF
JADAVPUR UNIVERSITY

June 23, 2022

Certificate from the Supervisor

This is to certify that the work embodied in this thesis entitled **"Matrix profile based analysis of repeated patterns of pollutants"** has been satisfactorily completed by **Animesh Adhikari** (Registration Number 154148 of 2020-21; Class Roll No. 002010502024; Examination Roll No. M4CSE22024). It is a bona-fide piece of work carried out under my supervision and guidance at Jadavpur University, Kolkata for partial fulfilment of the requirements for the awarding of the **Master of Engineering in Computer Science and Engineering** degree of the Department of Computer Science and Engineering, Faculty of Engineering and Technology, Jadavpur University, during the academic year 2021-22.

Dr. Sarbani Roy,
Professor,
Department of Computer Science and Engineering,
Jadavpur University.
(Supervisor)

Forwarded By:

Prof. Anupam Sinha,
Head,
Department of Computer Science and Engineering,
Jadavpur University.

Prof. Chandan Majumdar,
DEAN,
Faculty of Engineering & Technology,
Jadavpur University.

Department of Computer Science and Engineering
Faculty of Engineering And Technology
Jadavpur University, Kolkata - 700 032

Certificate of Approval

This is to certify that the thesis entitled **"Matrix profile based analysis of repeated patterns of pollutants"** is a bona-fide record of work carried out by **Animesh Adhikari** (Registration Number 154148 of 2020-21; Class Roll No. 002010502024; Examination Roll No. M4CSE22024) in partial fulfilment of the requirements for the award of the degree of **Master of Engineering in Computer Science and Engineering** in the **Department of Computer Science and Engineering, Jadavpur University**, during the period of June 2021 to June 2022. It is understood that by this approval, the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein but approve the thesis only for the purpose of which it has been submitted.

Examiners:

(Signature of The Examiner)

(Signature of The Supervisor)

Department of Computer Science and Engineering
Faculty of Engineering And Technology
Jadavpur University, Kolkata - 700 032

Declaration of Originality and Compliance of Academic Ethics

I hereby declare that the thesis entitled **"Matrix profile based analysis of repeated patterns of pollutants"** contains literature survey and original research work by the undersigned candidate, as a part of his degree of **Master of Engineering in Computer Science and Technology** in the **Department of Computer Science and Engineering, Jadavpur University**. All information have been obtained and presented in accordance with academic rules and ethical conduct.

I also declare that, as required by these rules and conduct, I have fully cited and referenced all materials and results that are not original to this work.

Name: Animesh Adhikari

Examination Roll No.: M4CSE22024

Registration No.: 154148 of 2020-21

Thesis Title: Matrix profile based analysis of repeated patterns of pollutants

Signature of the Candidate:

ACKNOWLEDGEMENT

I am pleased to express my gratitude and regards towards my Project Guide Dr. Sarbani Roy, Assistant Professor, Department of Computer Science and Engineering, Jadavpur University, without whose valuable guidance, inspiration and attention towards me, pursuing my project would have been impossible.
For his recommendations and persistent support, I would like to convey my sincere gratitude to Asif Iqbal Middya, Research Fellow at Jadavpur University in Kolkata.

Last but not the least, I express my regards towards my friends and family for bearing with me and for being a source of constant motivation during the entire term of the work.

Animesh Adhikari
MCSE Final Year
Exam Roll No. - M4CSE22024
Regn. No. - 154148 of 2020-21
Department of Computer Science and Engineering,
Jadavpur University.

JADAVPUR UNIVERSITY

Abstract

Faculty of Engineering and Technology
Computer Science and Engineering

Master of Engineering

Matrix profile based analysis of repeated patterns of pollutants

by Animesh Adhikari

Air pollution is one of the most pressing challenges nowadays. Almost every country has been contaminated by the environment. The majority of air pollution in the environment is caused by the consumption of energy for certain purposes. There is a significant risk of toxic gas production. Air pollution has a variety of negative effects on people's health. In humans, it causes a wide range of cutaneous and respiratory issues. Air pollution causes asthma, bronchitis, and a range of other illnesses. Furthermore, it hastens the ageing of the lungs, lowers lung function, and destroys the cells of the respiratory system. In this work, we are trying to find the pollution level and behaviour of pollutants in different stations in Delhi by finding repeated patterns on a daily, weekly, and monthly basis using a matrix profile. Several pollutants are used to discover repeated patterns such as $PM_{2.5}$, BP, Ozone, RH, etc. and for each pollutant, there are thirty-two stations. The z-normalize method and without the z-normalize method are used to collect results after finding patterns from stations and then cluster them based on euclidean distance between them.

Keywords: *matrix profile, time series data, time series pattern, air pollution, multi graniular analysis*

Contents

List of Figures

List of Tables

List of Abbreviations

MP	**M**atrix **P**rofile
MASS	**M**ueen's **A**lgorithm **For** **S**imilarity **S**earch
GIS	**G**eographic **I**nformation **S**ystem
API	**A**pplication **P**rogramming **Interface**
OSM	**O**pen **S**treet **M**ap
STAMP	**S**calable **T**ime series **A**nytime **M**atrix **P**rofile.
STOMP	**S**calable **T**ime **S**eries **O**rdered-search **M**atrix **P**rofile
GPU	**G**raphics **P**rocessing **U**nit
FET	**F**ast **F**ourier **T**ransform
CSV	**C**omma **S**eparated **V**alues
CPCB	**C**entral **P**ollution **C**ontrod **B**oard
LTS	**L**ong **T**erm **S**upport
NaN	**N**ot a **N**umber
MDL	**M**inimum **D**escription **L**ength
LAMP	**L**earned **A**pproximate **M**atrix **P**rofile

List of Symbols

T	Time series data
m	Granularity
k	Number of cluster
P	Set of Pollutant
S	Set of stations
A	Pattern of different length
$G_{S_1}^{P_1}$	Time series data of P_1 pollutant of S_1 stations
T^P	Time series of pollutant P
T^S1	Station S1 time series data
GL^S	Geolocation of stations
PD^S	Population density of area of stations S
LU^D	Landuse locations in Delhi
d[i, j]	Distance matrix

Chapter 1

Introduction

1.1 Overview

Pollutant monitoring is one of the most important tasks in recent days. Recently global warming increases extensively because of different types of pollution caused by the burning of fossil fuels, automobiles, agriculture activities, factories and industries, and domestic sources. In our work, we mainly focus on Air Pollution. In our environment, several pollutants exist that cause Air pollution. Here we try to find the characteristics of these pollutants and relate their behaviour by finding a pattern of different length of pollutant values in different stations. With the help of this experiment, we monitor the pollutants of different stations. Suppose an area has ten different pollutants and we try to find which pollutant has similar behaviour or which are different and pollution level of pollutant in different stations is high or low.

We collect the data of different pollutants of different stations from the government of India website. The dataset that we get from the website is time series data of almost 400 days' Air pollution data and data of pollutants that are taken at an interval of one hour. And from the time series data we try to find the repeated pattern of three different granularity(daily, monthly and weekly). There are several machine learning, deep learning and statistical methods which discover repeated patterns from time series data. In our work a new technology, matrix profile has been used.

1.2 Motivation

These days, one of the most serious issues is air pollution. Almost all countries have become victims of environmental contamination. This is the process of introducing pollutants into the air, which is harmful to human life and the environment as a whole. The majority of air pollution occurs in the environment as a result of the use of energy for particular activities. Whatever the reason for burning the fossil, it releases some unexpected and chemical elements into your environment. Reports points out that burning fossil fuels, such as coal, releases dangerous and stifling chemical compounds into the atmosphere. Apart from that, there is a considerable risk of harmful gas production. People's health is harmed by air pollution in different ways. It is the cause of a variety of cutaneous and respiratory problems in humans. It also causes cardiac problems. Asthma, bronchitis, and a variety of other ailments are caused by air pollution.Furthermore, it accelerates lung ageing, reduces lung function, and damages respiratory system cells. In this work we analyze different pollutants behaviour and pollution level in different cities. And also analyze which pollutant is harmful of an area. After analyzing that we try to reduce the uses those things that causes pollution level high. This way if we analyze the pollution level of different area and maintain the pollution level then we may control air pollution significantly.

1.3 Objective

The main objective of this work is to find the repeated pattern from the Air pollutant dataset collected from different pollutant monitoring stations in Delhi. To discover repeated patterns we segregate stations in four different ways such as landuse-based, geolocation-based, population density based, and randomly. Specifically, it is required to identify patterns of different pollutants concentration in geographically distributed stations' time series data. The z-normalization method is used to normalize the pollutant values after finding the repeated pattern. Matrix profile is utilized to find repeated patterns with three granularity . For segregating the stations k-means algorithm has been used based on geolocation as well as population density. QGIS API has been used for segregating stations based on landuse. After finding the pattern, they are clustered based on the distance between them.

1.4 Contribution

We use statistical data of Delhi to build our model which can discover repeated pattern of a particular pollutant in different areas in Delhi. In this work we try find repeated pattern of $PM_{2.5}, CO, NO_2, SO_2$ in several areas in Delhi with three different granularity that is daily, weekly and monthly basis. Matrix profile is one of the powerful tool to find repeated pattern from set of time series data with different granularity. Two main algorithm are used to find repeated pattern one is 'OSTINATO' and other is 'MASS'.In this study we also graphically shows that similar behaviour of pollutants in different stations. We grouping stations in four different ways like geolocations based that is consider all the stations as points in 2d plane the classify them using k-means algorithm, landuse based means using QGIS Api we fall stations in different landuse, randomly select stations and segregate stations on the basis of population density of particular area, here we use k-means too to segregate stations.

1.5 Organization

The thesis is organized as follows: chapter 2 describe some existing work related to motif discovery using matrix profile and some statistical method. In chapter 3 we have discussed the problem statement mathematically and graphically present workflow of the thesis and discuss each part. In this chapter we have also discussed how air pollution data is collected and described the model that is used, parameters that can be tuned and the algorithms those helps to find repeated pattern. Chapter 4 discusses the experimental setup in which we built our model and results that we get. Finally chapter 5 concludes the thesis and future scope of our work.

Chapter 2

Literature Review

In this chapter we discuss the details of existing works on matrix profile based motif discovery. Individuals can monitor a variety of indicators related to their personal and professional activity. Sensor technology is improving all the time, and the number of sensors is growing, as seen in finance and seismic investigations, for example. This produces vast amounts of complex data, usually in the form of time series, which makes knowledge discovery difficult.With such large and complicated collections of data, rapid and effective similarity search is essential for various data mining applications like as Shapelets, Motif Discovery, Classification, and Clustering.In a recent paper [1] STAMP, an anytime algorithm is introduced, which uses a quick similarity search approach to locate exact pair-motifs of a particular length.As a subroutine, it uses the standard Fast Fourier Transform (FFT) algorithm to leverage the overlap between subsequences under z-normalized Euclidean distance.They proposed storing the information acquired from the similarity join in a Matrix Profile data structure(MP).

Several related works can be found in [2] [3] [4].In [2] STOMP(Scalable Time series Ordered-search Matrix Profile) algorithm is given.It is shown how the scalability of exact motif identification may be improved by utilising GPU hardware and adapting the previously published STAMP algorithm.In [3],an algorithm is provided that can handle any length of unlabeled time series and make intuitive and informative visualisations.A scalable and understandable technique is presented for extracting such subsequences based on the Minimum Description Length (MDL).The usefulness of this method is illustrated on cardiology, human activity, audio, seismic, and electrical power consumption time series.Time series patterns can be of various lengths in real-world industrial settings, the class annotations may only belong to a general part of the data, may contain errors, and the class distribution is often extremely skewed.SDTS, a scalable method has been introduced in [4] that can learn in such difficult situations.

In [5] the reasons behind the issues of time series motif discovery is explained, and a novel and general framework is introduced to address them.This framework allows a user to produce vectors using only a few lines of code in a scripting language. This framework is also simple and adaptable enough to support domains and constraints that have yet to be explored. The multidimensional matrix profile, similar the original matrix profile [1][2], can be generated using a variety of algorithms and used in a variety of time series data miming applications with some modification and/or postprocessing. The multidimensional matrix profile is the motivation for the mSTAMP-based motif finding approach,introduced in [6]. In [7] time series chains, a novel time series data mining primitive is introduced. They've demonstrated that chains may be discovered fast and reliably from noisy and complex information to deliver important insights.

In the time series domain, unsupervised semantic segmentation is a well-studied problem. Current methodologies have a number of flaws such as most approaches necessitate a large number of parameters to be set/learned, and so may have difficulty generalising to novel scenarios, most methods implicitly assume that all data can be segmented, which causes problems when that assumption is incorrect; that have limited the use of time series semantic segmentation beyond the academic contexts. To solve these challenges, [8] provided a domain-agnostic method with only one parameter that can handle high-rate data streaming. In this context, they puts the algorithm to the test on the world's largest and most diversified collection of time series datasets.One of the most useful primitives in time series data mining is the finding of time series motifs. Its utility for exploratory data mining, summarization, visualisation, segmentation, classification, clustering, and rule development has been demonstrated by researchers. Despite the well-documented prevalence of missing data in scientific, industrial, and medical datasets, there is still no technique that allows the finding of time series motifs in the presence of missing data, despite substantial investigation. [9] presents a technique for motif discovery in the presence of missing data in this paper. It has been formally demonstrated that this method is admissible by demonstrating that it produces no false negatives. It is further shown that with a minor constant factor time/space overhead, their method can "piggyback" on the fastest known motif finding method.

The scope of motif discovery has substantially extended in recent years as a result of algorithmic improvements. We argue that there is an unquenchable need for more motifs to be accommodated. [10] introduces SCRIMP++, an O(n^2) time and anywhere algorithm that combines the greatest aspects of STOMP and STAMP. [11] have developed a new primitive called top-k time series snippets that can discover snippets in enormous datasets even when they're corrupted by noise, dropouts, or a wandering baseline, among other things.Usually Euclidean Distance, or DTW, is used in several time series data mining algorithms. We believe that these distance metrics are not as robust as the general public believes. In [12], the authors present MPdist, a novel distance measure. The proposed distance metric is far more reliable than existing distance measures. It can deal with data that has missing values or erroneous areas.In a relatable paper [13] that by combining numerous fresh ideas with a novel scalable framework and cloud deployment to commercial GPU clusters, we can expand the motif's discovery envelope.

In recent years, the data mining community has broadly agreed that many time series analytics concerns pertain to detecting and then reasoning about repeated structure in time series. A concept of time series consensus motifs as well as a scalable approach for finding them in huge data sets is presented in [14].Time series semantic motifs, a generalisation of classic time series motifs, have been introduced in [15]. They've also demonstrated that semantic motifs are far more expressive than classic motifs, allowing us to search big, complicated datasets for recurring structure that would otherwise be impossible to find using present methods. In another work[16] LAMP, a flexible and generic framework for approximating Matrix Profile values in the face of fast-moving streams is introduced. Because the Matrix Profile is at the heart of many time series algorithms for classification, motif discovery, anomaly detection, segmentation, and other tasks, LAMP enables higher-level algorithms to be used in real-time on fast moving streams.

Analysts are often asked to identify enormous amounts of time series data in areas as disparate as entomology and sports medicine. This can be done automatically in rare situations using a classification system. Complex, noisy, and polymorphic data, on the other hand, can defeat state-of-the-art classifiers in many domains while easily yielding to human examination and annotation.This is particularly true if the person has access to more data and past annotations. This tagging effort could be a major stumbling block for scientific development. In [17], we present an algorithm that significantly minimises the amount of human effort necessary. This interactive algorithm organises subsequences and asks the user to designate a group's prototype, which is then applied to all group members.Time series motif discovery was reduced to a single parameter, the length of time series motifs we expect (or hope) to find, using the matrix profile. In many circumstances, this is an appropriate restriction because the user can set this parameter using out-of-band data or domain expertise. Pan Matrix Profile is introduced in [18], a new data structure that contains closest neighbour information for all subsequences of any length. The first totally parameter-free motif discovery technique in the literature is made possible by this data format.

Anomaly detection in time series is a perennially important research topic. In fact, in the booming age of IoT, it is a work that has become increasingly crucial.While the literature contains hundreds of anomaly detection approaches, one definition, time series discords, has emerged as a competitive and popular option for practitioners. Subsequences of a time series that are maximum far away from their nearest neighbours are called time series discords.Unlike many parameterladen methods, discords require only a single parameter to be set by the user: the subsequence length.MERLIN[19] is an algorithm that can efficiently and efficiently find discords of all lengths in massive time series archives. Being familiar with time series we have come across the concepts of discord and motifs.Time series motifs refer to two subsequences that are unusually close together, whereas time series discords refer to subsequences that are far apart. However, we claim that it is occasionally beneficial to reason about a subsequence's proximity to specific data while also considering its distance from other data. In [20], we offer the Contrast Profile, a new fundamental that allows us to efficiently compute such a specification in a principled manner.

TABLE 2.1: Existing work related to the time series motif discovery using matrix Profile

Reference	**Year**	**Contribution**
Chin-Chia Michael Yeh et al[1]	2016	A Unifying View that Includes Motifs, Discords and Shapelets.
Chin-Chia Michael Yeh et al [3]	2016	Visualization of Salient Subsequences in Massive Time Series.
Chin-Chia Michael Yeh et al[4]	2017	Using Weakly Labeled Time Series to Predict Outcomes.
Kaveh Kamgar et al[14]	2019	Finding concensus motif from time series data.
Shima Imani et al[17],	2019	Semantic motif discovery from time series data.
Zachary Zimmerman et al[16]	2019	Time Series Mining in the Face of Fast Moving Streams using a Learned Approximate Matrix Profile.
Frank MadridFinding et al[18]	2019	Finding and Visualizing Time Series Motifs of All Lengths using the Matrix Profile

Chapter 3

Methodology

This chapter is basically graphically present and describes the overall process of the thesis work. The problem statement of the work is presented mathematcally and repeated pattern in different time series data with different granularity. This chapter also describes how pollutant data is collected and how to preprocess the dataset to fit into the model of finding the repeated pattern. It also describes different ways to segregate the stations.

3.1 Problem Statement

The dataset that is collected from the central government website is the Air Pollution dataset of different stations in different cities of Delhi, which has several stations for a particular pollutant. There are a set $P = \{P_1, P_2, ...P_m\}$ of twenty-three pollutants and a set $S = \{S_1, S_2, ..., S_n\}$ of thirty-two stations. Here m is number of pollutants and n is the number of stations. In the dataset there is one year pollution record and we try to find the pattern of different stations of pollutants of three different granularities such as $w_1, w_2, w_3 = m$ that is daily(24-hour window), weekly(7x24) and monthly(30x24) basis. Four methods are used to find repeated patterns and those are population density, landuse, geolocation and randomly. Suppose T be the time series data $T = T_1, T_2, T_3, ...T_l$ of length l. Patterns of length 24, 168 and 720 of time series data T can be represented as $A_{24} = T_i, T_{i+1}, ...T_{i+24}$, $A_{168} = T_j, T_j + 1, ...T_{j+168}$ and $A_{720} = T_k, T_k + 1, ..., T_{k+720}$. Matrix profile is a technique that helps to discover repeated structure from time series data of several stations of a pollutant.

Let $G_{S_1}^{P_1}$ be the time series data of P_1 pollutant and S_1 station and try to find repeated pattern for a particular pollutant in four randomly selected stations. For example randomly select P_2 pollutant and S_3, S_6, S_9, S_{11} be the randomly selected stations. Repeated pattern of the stations of particular pollutant of length 24 and initial stating points of these stations are a, b, c and d can be represented as $S_{3a}, S_{3a+1}, ...,$ $S_{3a+24}, S_{6b}, S_{6b+1}, ..., S_{6b+24}, S_{9d}, S_{9d+1}, ..., S_{9d+24}$ and $S_{11c}, S_{11c+1}, ..., S_{11c+24}$

For the remaining two granularity repeated patterns of stations can represent the same way. For population density-based finding repeated patterns k-means algorithm is applied to classify the stations. For finding land-use base and geolocation-based repeated patterns use QGIS API to classify stations.

For example, graphically present the repeated structure of four stations of pollutant instance CO in whole time series data and also shows the only repeated pattern of the stations.

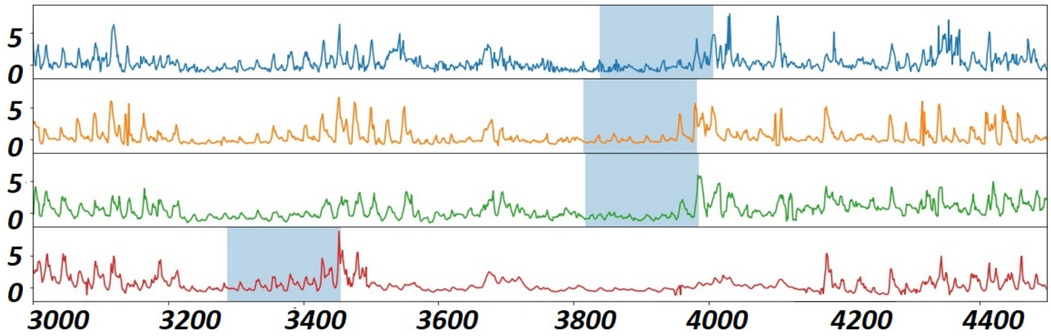


FIGURE 3.1: Similar patterns found in four different stations of CO pollutant

Fig.3.1 shows the repeated structure of CO pollutant and the length of the repeated structure is 168 that is found the repeated structure weakly basis and the figure shows some portion of the stations' time series data. The highlighted part in the figure represent repeated structure among the four stations and it also shows the location of every pattern. And all the station values are z-normalized. Here the z-normalized method is used for visualizing the pattern clearly.

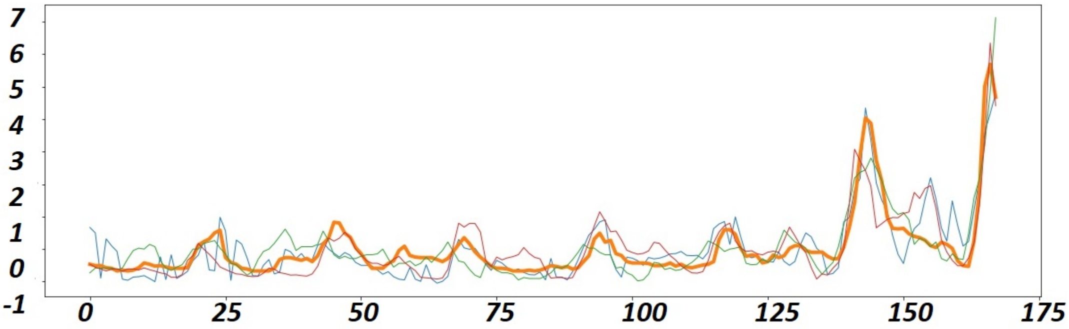


FIGURE 3.2: Optimal pattern and repeated pattern of CO pollutant in four different stations

Fig.3.2 reveals the repeated structure of randomly selected four stations. The bold-shaped pattern shows the optimal pattern among the four patterns. In the above figure, some parts of the pattern are similar and some are different. That means for a particular pollutant air pollution levels of some stations are the same and some are different.

3.2 A Workflow

This section shows the overall architecture of the work and discusses what is a task of every module. It also shows the output of one module flows to another module. The below figure shows the diagram of the overall work. It can be divided into eight modules and the modules are Data collection, Data pre-processing, preparing data for individual models, four different models for segregation of stations, and model for collecting results.

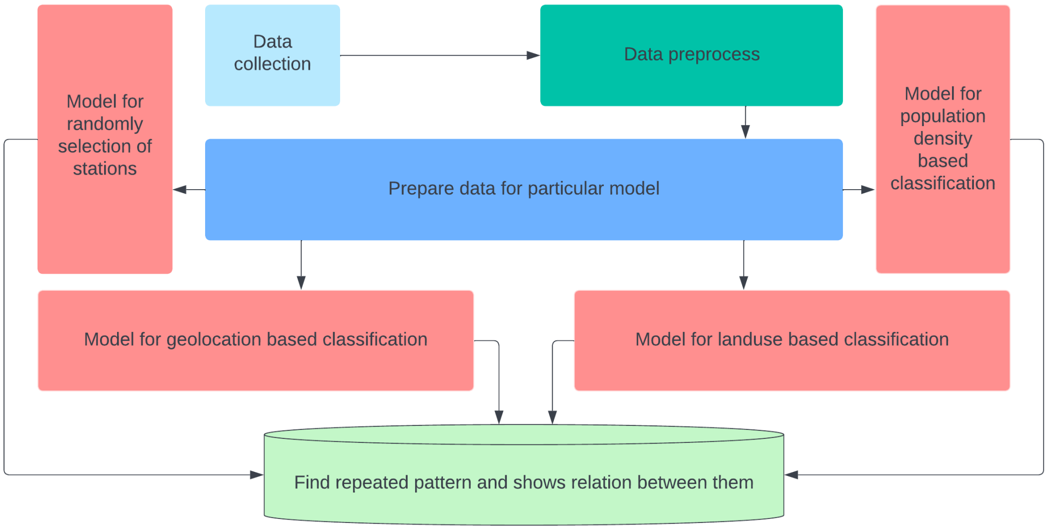

Data collection
Data preprocess
Model for randomly selection of stations
Prepare data for particular model
Model for population density based classification
Model for geolocation based classification
Model for landuse based classification
Find repeated pattern and shows relation between them

FIGURE 3.3: Overall Architecture of the Work

3.2.1 Module Description

Data Collection :

In the data collection phase of finding repeated structures collect data from the central government website for pollution monitoring of stations. From there we collect the air pollution dataset in CSV format.

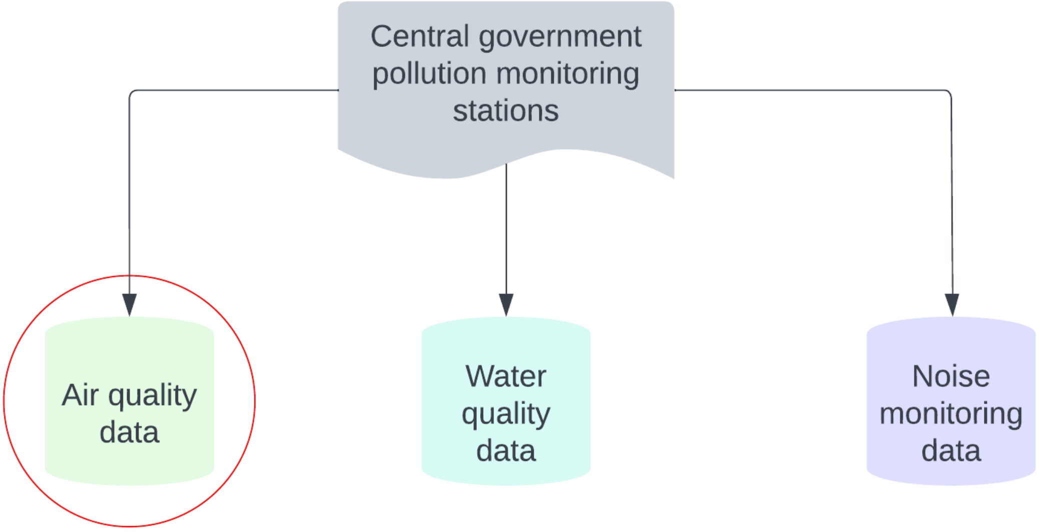


FIGURE 3.4: Collect Air Pollution Dataset from CPCP

From the above Fig.3.4 we can visualize the Indian Government set up a pollution monitoring station in different states in India which records several pollutants instances data for particular types of pollution. In this work, discover repeated structures from the Air Pollution dataset and pollutant instances are NH_3, NO_2, $PM_{2.5}$, SO_2, etc.

Data Pre-processing :

Pre-processing stage is used to eliminate the unnecessary information from the raw dataset that is not useful to find repeated structures between the stations of pollutant instances. The unnecessary information like the logo of the Central Pollution Control Board(CPCB), stations that have particular pollutants, Pollutant name, duration of collection pollution data, dataset collection date and time, and null values are removed from the dataset.

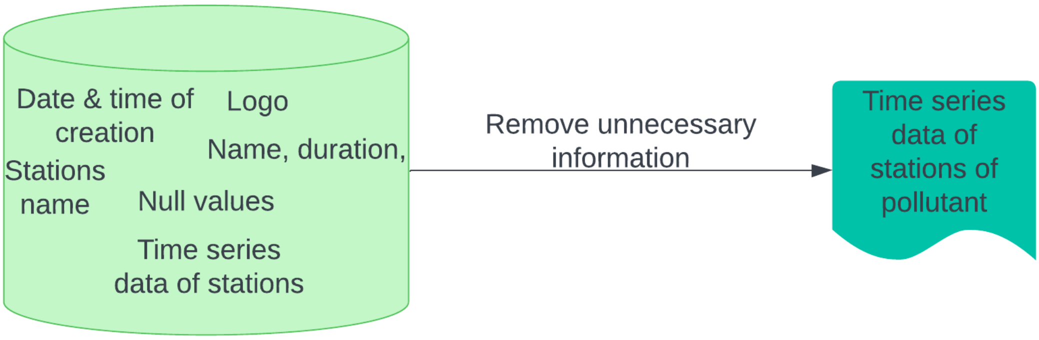

Date & time of creation
Logo
Name, duration,
Stations name
Null values
Time series data of stations
Remove unnecessary information
Time series data of stations of pollutant

FIGURE 3.5: Dataset after removing Unnecessary Information

Prepare Dataset to Fit into the Model :

In this phase, we abbreviate stations with code which helps to segregate stations with a particular model and easy to represent the relationship between them after finding the repeated pattern. Here we normalized the population density of the location of stations using z-normalize and remove null values. Table 3.1 shows that pollution monitoring stations of Delhi are abbreviated with codes and geolocations also stores of each station from Central Pollution Control. Board(CPCB)[1]. With the help of these information, stations are classified based on nearest location of the stations.

[1] https://app.cpcbccr.com/ccr//caaqm-dashboard-all/caaqm-landing

TABLE 3.1: Stations name, stations location and stations code

LATITUDE	LONGITUDE	STATION NAME	STATION CODE
28.6811736	77.3025234	IHBAS-CPCB	S1
28.5710274	77.0719006	Dwarka-S8-DPCC	S2
28.531346	77.190156	Sri-Aurobindo-Marg-DPCC	S3
28.7500499	77.1112615	DTU-CPCB	S4
28.530785	77.271255	Okhla-Phase-2-DPCC	S5
28.815329	77.15301	Alipur-DPCC	S6
28.4706914	77.1099364	Aya-Nagar-IMD	S7
28.710508	77.249485	Sonia-Vihar-DPCC	S8
28.695381	77.181665	Ashok-Vihar-DPCC	S9
28.628624	77.24106	ITO-CPCB	S10
28.58261	77.234238	JNS-DPCC	S11
28.684678	77.076574	Mundka-DPCC	S12
28.639645	77.146262	Pusa-DPCC	S13
28.6514781	77.1473105	Shadipur-CPCB	S14
28.672342	77.31526	Vivek-Vihar-DPCC	S15
28.636429	77.201067	Mandir-Marg-DPCC	S16
28.56789	77.250515	Nehru-Nagar-DPCC	S17
28.699793	77.165453	Wazirpur-DPCC	S18
28.5512005	77.2735737	CRRI-Mathura-Road	S19
28.563262	77.186937	RK-Puram-DPCC	S20
28.674045	77.131023	Punjabi-Bagh-DPCC	S21
28.4995128	77.2662248	KSSR-DPCC	S22
28.73282	77.170633	Jahangirpuri-DPCC	S23
28.7762	77.051074	Bawana-DPCC	S24
28.822836	77.101981	Narela-CPCB	S25
28.623763	77.287209	Patparganj-DPCC	S26
28.732528	77.11992	Rohini-DPCC	S27
28.647622	77.315809	Anand-Vihar-DPCC	S28
28.60909	77.0325413	NSIT-Dwarka-CPCB	S29
28.612569	77.2346097	MDCNS-DPCC	S30
28.570173	76.933762	Najafgarh-DPCC	S31
28.655935	77.294904	East-Arjun-Nagar-CPCB	S32

Randomly Classify Stations :

In the random segregation of stations, initially, five stations are selected randomly from one to two thirty-two (number of stations) then the next five stations are selected randomly except for previously selected stations. This process is repeated two more times to segregate all the stations. This way we cluster all the stations into four clusters according to the pollutant instances. Some station has only two to four station that's why ignored those stations. The pollutant is namely Eth-Benzene, Mp-Xylene, O, Temp, and VWS. And remaining all the stations we find repeated patterns of pollutants.

Landuse based Classification :

Based on human use of land or area it is called landuse. There are several landuse like industrial areas, commercial, agriculture, economic etc. Landuse-based classification stations means area of station belongs which particular landuse and to find this we use open street map API. From this API we generate CSV file for different landuse and each file contains the locations of area which belongs to particular landuse. From previously created a file which contains the locations of each stations and calculate the distance between each stations with the landuse location that we get from QGIS API. The distance is calculated using GeoPy python library and it is measure the distance using geodesic method in GeoPy library. Geodesic method mainly measure the shortest distance between two coordinate in the globe. We segregate stations by taking first six stations of each landuse that we get from the API. Table 3.2 defined that stations are grouped based on landuse. There are nine different landuse and for each landuse several stations exists.

TABLE 3.2: Landuse based segregation of stations

Landuse	Stations
Industrial	[S14, S13, S26, S18, S19, S5]
Historical	[S11, S30, S10, S3, S22, S20],[S17, S16, S5, S19, S31, S8]
Water_Tower	[S17, S1, S27, S14, S18, S16],[S13, S15, S4, S3, S9, S10]
Residential	[S32, S9, S17, S26, S21, S19],[S20, S30, S13, S14, S3, S2]
Commercial	[S5, S30, S20, S16, S19, S3]
Sports_facilities	[S23, S17, S32, S11, S30, S10]
Education	[S16, S5, S17, S27, S29, S26]
Garden	[S15, S3, S28, S11, S17, S16]
Wood	[S20, S5, S19, S12, S3, S22],[S24, S25, S27, S6, S29, S7]

Latitude and Longitude Based Classification :

In geolocation based segregation of stations, first remove those rows that has NaN values. After that k-means classification algorithm helps to segregate stations. To segregate stations we use five cluster and k-means++ method to initialize the each clusters centre. After k-means clustering method applied on the locations of the stations, all the stations are labeled with particular cluster number and then we use some brute force method to segregate the stations into different clusters.

Population Density Based Classification :

We collect population and population density of Delhi from GeoIQ.[2] website and only use population density to segregate the stations shown in Table3.3. Firstly we normalize the values of population density using z-normalized method. K-means clustering algorithm is applied on the normalized data of population density and five clusters are used to segregate the stations. After labeling the stations with cluster number using k-means algorithm a brute force method is applied to segregate stations into different clusters.

[2] https://geoiq.io/places/Vivek-Vihar/qWVYIcIzZG

TABLE 3.3: Population density of each stations area

Station Name	Population Density/km2^2
IHBAS-CPCB	20931
Dwarka-S8-DPCC	11588
Sri-Aurobindo-Marg-DPCC	14370
DTU-CPCB	8565
Okhla-Phase-2-DPCC	33659
Alipur-DPCC	6369
Aya-Nagar-IMD	5263
Sonia-Vihar-DPCC	5662
Ashok-Vihar-DPCC	18909
ITO-CPCB	44718
JNS-DPCC	12757
Mundka-DPCC	10275
Pusa-DPCC	4718
Shadipur-CPCB	23942
Vivek-Vihar-DPCC	28261
Mandir-Marg-DPCC	17368
Nehru-Nagar-DPCC	17602
Wazirpur-DPCC	29299
CRRI-Mathura-Road	18942
RK-Puram-DPCC	14620
Punjabi-Bagh-DPCC	20746
KSSR-DPCC	6803
Jahangirpuri-DPCC	17652
Bawana-DPCC	6660
Narela-CPCB	5728
Patparganj-DPCC	22088
Rohini-DPCC	5900
Anand-Vihar-DPCC	19518
NSIT-Dwarka-CPCB	24838
MDCNS-DPCC	8693
Najafgarh-DPCC	11531
East-Arjun-Nagar-CPCB	24168

Find Repeated Structure and Represent Relation Between Stations :

In this phase we clearly describe how repeated patterns are generated. Three different granularity(Daily, Weakly, Monthly) are used to analyze the pollutant instant in different station properly. The algorithm that we have used to discover repeated pattern is 'OSTINATO'[14]. This algorithm is used to find best radius, time series id and start index of optimal pattern from more than one time series data. Here 'radius' term is used because all the subsequence in all time series of length 'm' in m dimensional space. Th algorithm finds shortest distance from a subsequence to all other subsequence then it gets the best radius, time series id and subsequence id. After finding the index of optimal subsequence then using MASS[1] (Mueen's Algorithm for Similarity Search) algorithm calculate shortest distance from all subsequence in other stations time series data. The MASS algorithm is basically calculates distance using euclidean distance between subsequences. After finding the index of repeated pattern in other time series data we can use z-normalized method to normalized data points in the pattern to show the pattern properly among the pattern. Dendrogram is technology which shows the relation among the stations using MASS algorithm.

Below figure shows how to find repeated pattern from set of time series data. OSTINATO[14] ans MASS [1]algorithm are use to find the repeated pattern and another method is used to normalize the data is z-normalized.

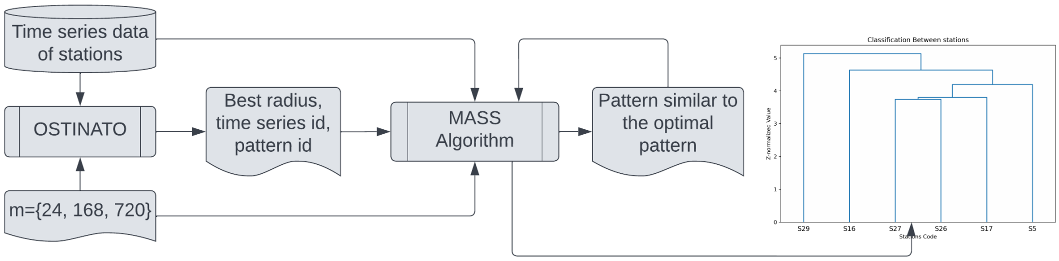

Time series data of stations
OSTINATO
m={24, 168, 720}
Best radius, time series id, pattern id
MASS Algorithm
Pattern similar to the optimal pattern
Classification Between stations
Z-normalized Value
0
1
2
3
4
5
S29
S16
S27
S26
S17
S5
Stations Code

FIGURE 3.6: Graphically present how repeated pattern a discovered.

3.3 Pollution Data Collection

This section is basically presents how to collect dataset, from where we collect dataset and what are the tools and technologies are used to collect the dataset. The dataset that are used to find repeated pattern is Air Pollutant Dataset from Central Pollution Control Board(CPCB)[3] of India. Indian government built pollution monitoring station in different cities in different states such as Delhi, West Bengal, Punjab, Assam, Gujarat, Odisha etc in India. For our work we collect the dataset of state Delhi. There are total forty one pollution monitoring stations out of which thirty two stations are taken to find repeated structure in Air Pollution Dataset. Air pollution dataset of Delhi state has twenty three pollutant instances such as Benzene, BP, CO, $PM_{2.5}$, etc. A particular area in Delhi, pollution level of some pollutants are high and some are low and some are zero. So, for a particular pollutant not all the thirty two stations are included. Pollution level of each is recorded in one hour interval and total record is almost 400 days. Here we use three granularity that is 24 hour window(daily basis), 24x7 hour window(weakly basis) and 24x30 hour window(monthly basis) to find the repeated structure in different stations of pollutants. Random, population density, geolocaton and landuse are for types of methods which is used for classify the stations. QGIS is a geographic information system software which helps to detect landuse for cities in a state. To get landuse information we use QuickOSM API, which provides Open Street Map information and if we enter the state name and which landuse information that we want then it returns the locations of this particular landuse in the state.

3.4 Proposed Approach

This section basically presents the description of the model, parameters that helps to build the model and after tuning these parameters we gets the better results, and preprocessing steps which preprocess the dataset that suitable for the model and the algorithm which help to find the repeated pattern among the time series data. In this work we use k-means clustering and some brute force methods method which helps to classify the stations of pollutants. In k-means clustering method number of

[3] https://app.cpcbccr.com/ccr//caaqm-dashboard-all/caaqm-landing

cluster that we take is 5 and k-means++ method is to select cluster centre in a way that convergence rapidly. In randomly classification of stations we use geodesic method which find the distance between coordinate value of two areas.

Parameters that are used for the experiment discussed below

1. **Randomly select stations :** Here we use random method for selecting four random stations number and we consider these four stations as the start station number and from that station number we select consecutive four stations.

2. **Find distance between coordinate :** Every location in the globe has coordinate value(i.e. latitude and longitude), we take from the website of GeoIQ[4]. Using geodesic method we calculate the distance between two coordinate values which is to find the shortest distance between all the stations.

3. **Number of clusters :** This parameter is used to classify the thirty two station into few groups. For our work we set this parameter as five that means we group these stations into five groups.

4. **Length of the patterns :** From the Air pollution dataset of several stations we use three different granularity that is daily(24 hour), weekly(24x7 hour) and monthly(24x30) to analyze the level of pollution properly of a pollutant in different stations.

Now, we discuss algorithms which helps to discover repeated pattern of several stations of pollutant

Algorithm 1 DiscoveringRepeatedPattern

```
Input: Pollutant time series data of stations ${T^P = T^{s1}, T^{s2}, ..., T^{sp}}$, abbreviate stations with code ${S = s^1, s^2, ...., s^{32}}$, geolocation of each pollution monitoring stations $GL^S = LL^{s1}, LL^{s2}, ...., LL^{s32}}$, population density of station's area $PD^S = PD^{s1}, PD^{s2}, ..., PD^{S32}}$, landuse locations in Delhi $LU^D = LU^D_1, LU^D_2, ..., LU^D_z}$, granularity $M = m_1, m_2, m_3}$.
Output: Repeated pattern of stations of Pollutants ${ T^{s1}_i, T^{s1}_i + 1, ...., T^{s1}_{i+m}}$, ${ T^{s2}_j, T^{s2}_j + 1, ...., T^{s2}_{j+m}}$, ....., ${ T^{sq}_t, T^{sq}_t + 1, ...., T^{sq}_{t+m}}$
if Is randomly segregation then
    $statData \leftarrow$ RandClassSta($T^P, S$)
else if Is population density based segregation then
    $statData \leftarrow$ PopDenClassSta($T^P, S, PD^S$)
else if Is Geolocation based segregation then
    $statData \leftarrow$ GeolClassSta($T^P, S, GL^S$)
else if Is landuse based segregation then
    $statData \leftarrow$ LanduseClassSta($T^P, S, LU^D$)
end if
for $m \leftarrow M$ do
    $radius, TsIdx, SubseqIdx \leftarrow$ OSTINATO($statData, m$)
    $optPatt \leftarrow stat_data.values[Subseq_idx : Subseq_idx + m]$
    for $station \leftarrow statData.stat_name$ do
        $strtIdxOfPatt \leftarrow$ MASS($optPatt, statData[station]$)
        $statSubseq[station] \leftarrow stat_data[station][strtIdxOfPatt : strtIdxOfPatt + m]$
    end for
end for
return statSubseq
```

The algorithm shows the steps of finding repeated pattern of different stations of pollutant instance but inside the algorithm there is four different subroutines and two major algorithms. The main objective of these four subroutines are segregation of stations in four different ways and objective of one algorithm find optimal pattern from set of time series data and other is to find distance matrix from optimal pattern

[4] https://geoiq.io/places/Vivek-Vihar/qWVYIcIzZG

that we found to all other stations time series data.
Distance matrix that we define above which stores the distance from a subsequence to all other sub sequence in the time series[1]. So here distance matrix basically stores the distance from optimal pattern that we found from OSTINATO[14] algorithm to all other time series data of different stations of pollutants.

Chapter 4

Results and Analysis

This chapter describes the experimental setup in which configuration of the machine where our model is run. It mentions the language used to find repeated pattern and describes the libraries that have been used for this model and also the machine learning model used. This chapter also elaborates several ways that we can use to segregate the stations and analyze among the patterns of different stations with the help of normalization as well as without normalization.

4.1 Experimental Setup

The experiment for finding the repeated pattern in different time series data is done on Ubuntu 20.04.4 LTS version with 11th Gen Intel® Core™ i5-1135G7 @ 2.40GHz × 8 CPU with 8GB memory machine. We have used Python language and libraries like Numpy, Pandas, stumpy, matrixprofile, matplotlib, scipy, sklearn . Numpy[1] is useful when we apply complex mathematical function, linear algebra operation fourier transformation etc to the time series data. Matplotlib[2] library is used to visualize experimental results and input data graphically. Matrixprofile[3] and stumpy[4] are usually used to discover optimal pattern, repeated pattern, normalize time series data among the time series data. From scipy[5] library we use two methods: linkage and dendrogram to show the relation and similar pattern from different stations of particular pollutant. And from sklearn[6] we have used k-means classification to classify the stations to find repeated pattern of a pollutant.

4.2 Results

This section presents the results of discovering the repeated pattern of pollutants in different stations which we obtain from the Air pollutant dataset of Delhi. Tools, technology, and methodology for finding repeated patterns are discussed in previous sections. Space and Time complexity depend on the algorithm, dataset, and granularity(i.e. length of the pattern). Of all the repeated patterns that we have found from several pollutants here, we choose only $PM_{2.5}, CO, NO_2$ and SO_2. The results that we get from the air pollution dataset include z-normalized as well as without z-normalized pollutant values.

[1] https://numpy.org/

[2] https://matplotlib.org/

[3] https://matrixprofile.org/

[4] https://stumpy.readthedocs.io/en/latest/index.html

[5] https://scipy.org/

[6] https://scikit-learn.org/stable/

The algorithms that are used to find the repeated pattern is discussed above. Here we discuss the methodology for clustering the stations graphically as well as mathematically. The input of this clustering method is either a one-dimensional condensed matrix or a two-dimensional observation vector[7]. Two clusters form a new cluster by calculating the distance between two existing clusters C_1 and C_2. There are several methods for calculating the distance between two clusters such as single, complete, average, weighted, centroid, median, and ward. Suppose in cluster X has N original observations $X_1, X_2, ..., X_N$ and in cluster Y has M original objects $Y_1, Y_2, ..., Y_M$. Now X can be created by combining two clusters C_1 and C_2 and any existing clusters that are not in X are the cluster Y. The distance between i and j cluster is stored in a matrix called the distance matrix and every iteration keeps track of the distance matrix. Table 4.1 shows the ways in which we can calculate the distance between two clusters, the mathematical expression and the names of the algorithms. The algorithm starts with a forest of clusters that are yet to be categorized hierarchically. Two clusters are removed from the forest only when they are combined into a single cluster, for example, C_1 and C_2 be two clusters that combine to form cluster X then C_1 and C_2 can be removed from the forest, and X is to be added to the forest. The algorithm halts only when a single cluster remains and it is the root of the forest. Let d[i, j] be the distance matrix that stores the distance between two clusters and it is updated after every iteration of the algorithm. There must be more than one minimum distance between clusters when minimum distance pair is chosen.
Among the different methods of finding minimum distance, 'Single' is the minimum spanning tree-based algorithm and 'complete', 'average', 'weighted', and 'ward' are the nearest neighbour-based algorithms. The time complexity of all the algorithms is O(N^2). And the remaining two methods 'centroid', and 'median' are euclidean distance based.

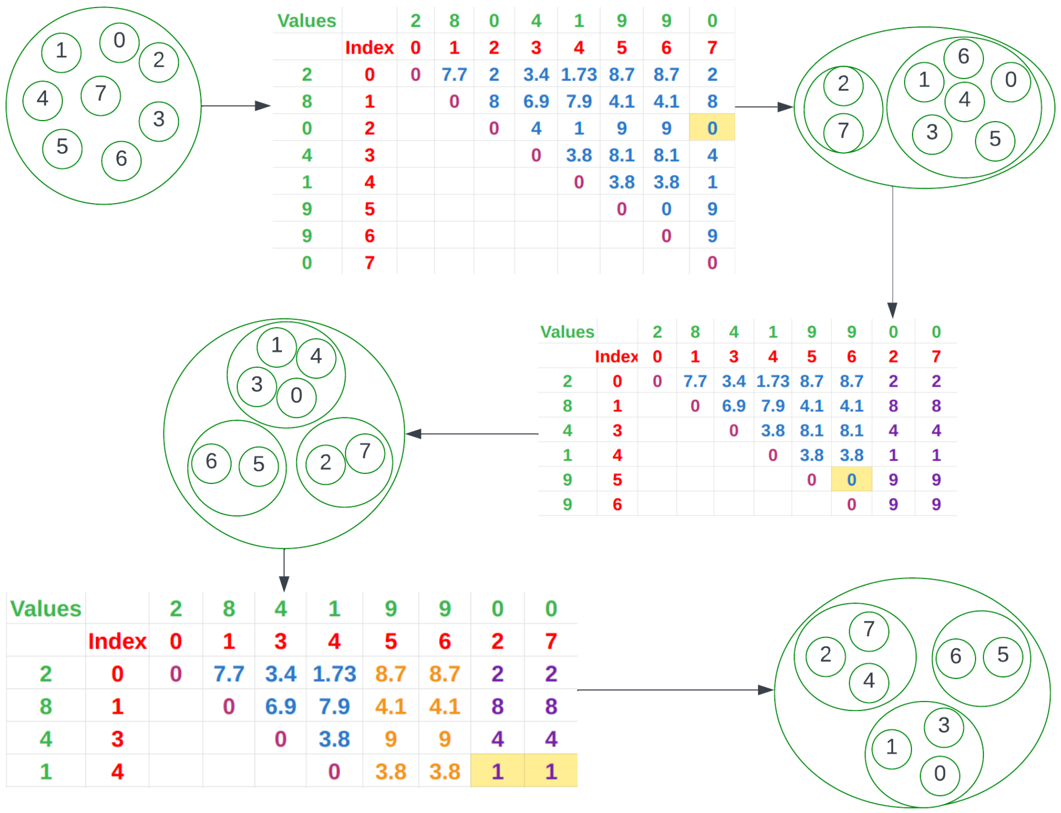
[7] https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.linkage.html

TABLE 4.1: Distance Calculating Methods

Method Name	**Mathematical Expression**	**Algorithm Name**	**Description**
Single	d(X, Y) = min(dist(X[i], Y[j]))	Nearest Point Algorithm	This method calculate the distance between all the cluster in the forest and take the minimum distance.
Complete	d(X, Y) = max(dist(X[i], Y[j]))	Farthest Point Algorithm or Voor Hees Algorithm	It takes maximum distance from all the distance which is calculated between all the clusters in the forest.
Average	d(X, Y) = $$\sum_{ij} \frac{(d(X[i], Y[j])}{(\lvert X\rvert * \lvert Y\rvert)}$$	UPGMA Algorithm	It calculates distance between cluster X and cluster Y then add all the distance values together. Finally divided by number of clusters in X and number of clusters in Y.
Weighted	d(X, Y) = $(d(C_1, X) + d(C_2, Y))/2$	WPGMA Algorithm	y using cluster s and t cluster X is formed and Y is the remaining cluster in the forest.
Centroid	$d(C_1, C_2) = \lVert c_{C1} - c_{C2} \rVert_2$	UPGMC Algorithm	Here cluster is formed by calculating distance between centroid of the clusters. From original objects of C_1 and C_2 new centroid is computed.
Median	Same as method centroid	WPGMC Algorithm	Like centroid method median of cluster is assign. The average of centroids C_1 and C_2 gives the new centroid X when two clusters C_1 and C_2 are joined to form a new cluster X.
Ward	d(X,Y)=$\sqrt{S+T-U}$ $Where\ S = \frac{\lvert Y\rvert+\lvert C1\rvert}{Z} d(Y, C1)^2, T = \frac{\lvert Y\rvert+\lvert C2\rvert}{Z} d(Y, C_2)^2 and\quad U = \frac{\lvert Y\rvert}{Z} d(C_1, C_2)^2$	Incremental Algorithm	The entries of the distance matrix is computed by the formula given in column 3. Cluster C_1 and C_2 joined together to make cluster X.

We have already discussed theoretically the working paradigm of clustering and the mathematical expression of methods to find the distance between two clusters. Now we illustrate an example. We consider a forest of eight different clusters and will discuss how the distance matrix is created and how it is updated at every iteration. We will also discuss the method that can classify the forest from a single cluster to more than one cluster according to the distance between clusters.
The overall process has been depicted in three figures Fig. 4.1 and Fig. 4.2 and we show how the distance matrix is created and updated and clusters are classified based on distance between clusters. And in Fig. 4.3 we present clusters using dendrogram.

A distance matrix created from the clusters is an upper triangular matrix. There are several methods for finding distance between clusters that we discussed in Table 4.1, here we use a single method which mainly uses Euclidean distance. In the distance matrix, green coloured part is the value of the clusters and the red coloured part is the indices of the clusters. From the distance matrix we take the minimum value and make a cluster with reference to the minimum value. In Fig. 4.1 matrix in the upper left corner minimum value is zero and its reference points are (2,7) and (5,6). We take (2,7) reference point and create a cluster and these two reference points are removed from the forest. And the values of these reference points are removed from the distance matrix. The next minimum value is also zero and its reference point is (5, 6) and creates a separate cluster with these reference points. And it is removed from the forest and as well from the distance matrix.



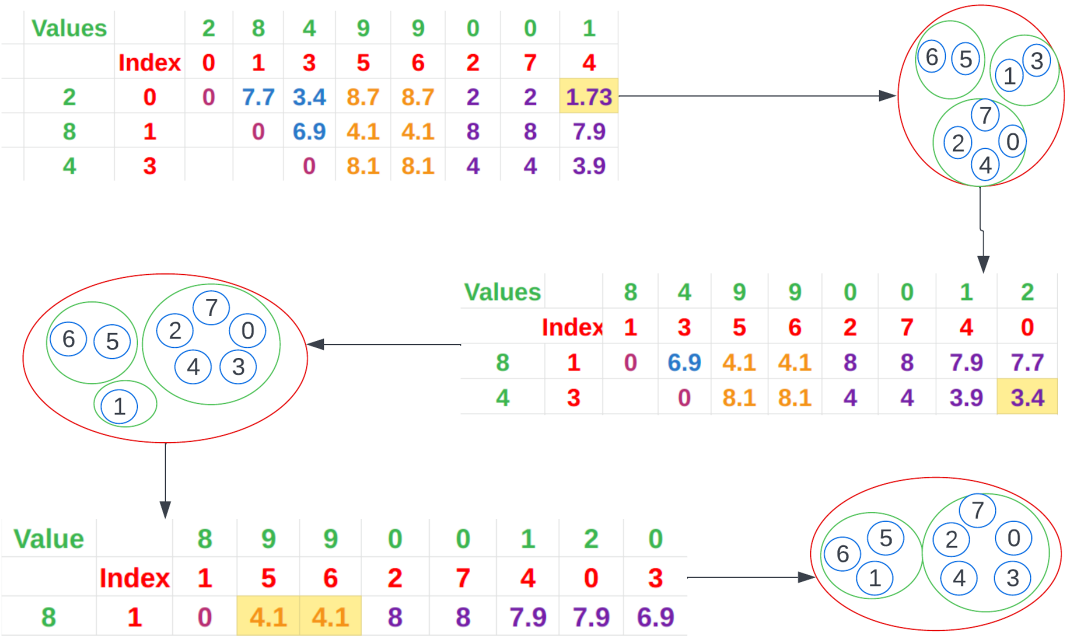
Values		2	8	0	4	1	9	9	0
	Index	0	1	2	3	4	5	6	7
2	0	0	7.7	2	3.4	1.73	8.7	8.7	2
8	1		0	8	6.9	7.9	4.1	4.1	8
0	2			0	4	1	9	9	0
4	3				0	3.8	8.1	8.1	4
1	4					0	3.8	3.8	1
9	5						0	0	9
9	6							0	9
0	7								0

Values		2	8	4	1	9	9	0	0
	Index	0	1	3	4	5	6	2	7
2	0	0	7.7	3.4	1.73	8.7	8.7	2	2
8	1		0	6.9	7.9	4.1	4.1	8	8
4	3			0	3.8	8.1	8.1	4	4
1	4				0	3.8	3.8	1	1
9	5					0	0	9	9
9	6						0	9	9

Values		2	8	4	1	9	9	0	0
	Index	0	1	3	4	5	6	2	7
2	0	0	7.7	3.4	1.73	8.7	8.7	2	2
8	1		0	6.9	7.9	4.1	4.1	8	8
4	3			0	3.8	9	9	4	4
1	4				0	3.8	3.8	1	1

FIGURE 4.1: Procedure of clustering stations after finding repeated Pattern-1

Now the distance between cluster four and cluster with the reference to point(2, 7) is minimum i.e. one. Hence cluster four is added to the cluster (2, 7) to form a cluster (2, 7, 4) and 4 is removed from the main forest as well as the distance matrix. In Fig. 4.2 distance between cluster 0 and all the clusters in the forest (2, 7, 4) is minimum so it is added to it and removed from the main forest and from a distance matrix.



Values		2	8	4	9	9	0	0	1
	Index	0	1	3	5	6	2	7	4
2	0	0	7.7	3.4	8.7	8.7	2	2	1.73
8	1		0	6.9	4.1	4.1	8	8	7.9
4	3			0	8.1	8.1	4	4	3.9

Values		8	4	9	9	0	0	1	2
	Index	1	3	5	6	2	7	4	0
8	1	0	6.9	4.1	4.1	8	8	7.9	7.7
4	3		0	8.1	8.1	4	4	3.9	3.4

Value		8	9	9	0	0	1	2	0
	Index	1	5	6	2	7	4	0	3
8	1	0	4.1	4.1	8	8	7.9	7.9	6.9

FIGURE 4.2: Procedure of clustering stations after finding repeated Pattern-2

In the above figure, the next cluster to be added to any of the clusters which is created is cluster 3. And its distance is minimum from the sub-cluster (2, 7, 4, 0) so it is added to it and removed from the forest. Finally only one cluster is remaining in the main forest which is 0 and its distance is minimum from cluster (5,6). So, cluster 0 is added to the sub-cluster of the main forest (5, 6) as (5, 6, 0). Hence two separate cluster is formed from the main forest that is cluster (5,6,0) and (2,7,4,1,3) and from these two clusters we join both of them and make root of the tree.

Now in the Fig. 4.3 we have shown above discussed clustering from main forest to sub forest on the basis of distance between them using dendrogram.

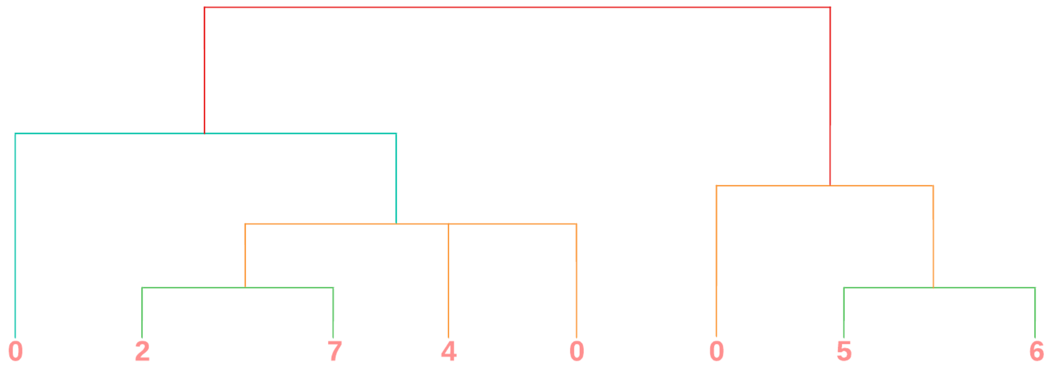


FIGURE 4.3: Procedure of clustering stations after finding repeated Pattern-3

Below we present results graphically of each pollutant with three different granularity and discuss them briefly.

4.2.1 Randomly classified stations :

Here stations are classified randomly. Four initial stations are selected in a random manner then the remaining stations are selected consecutively. The below sections represents the results of each pollutant in different stations with both normalized and non-normalized value of pollutants and present results with three granularity i.e. daily(24 hours), weekly(24x7 hours) and monthly(24x30 hours) basis. In these figures left sided part represents the time series subsequence or motif or pattern of time series data of different stations of pollutants and the right-sided part represents grouping of stations using the linkage and dendrogram method.

Normalized Results :

Here we show repeated patterns of $PM_{2.5}$ pollutant of randomly selected stations of Delhi state of three different granularity. In the Fig.4.4 left-hand side represents the repeated pattern of randomly selected pollution monitoring stations like IHBAS-CPCB, Dwarka-S8-DPCC, Sri-Aurobindo-Marg-DPCC, DTU-CPCB, Okhla-Phase-2-DPCC of $PM_{2.5}$ pollutant and x-axis represent the length of repeated pattern and y-axis represent the normalized value of pollutant in stations. In patterns that we have found, some parts of the pattern are overlapped with others which means behaviour of the pollutant in these stations at particular times are same. In the beginning, patterns are totally different but as time exceeds all the pattern looks the same, and after some time patterns look like zig-zag. The pattern that we found from different positions of a set of time series data we observe that at the beginning the value of the pollutant is high to these respective stations which means that the pollution level of this pollutant is high and when time increases its level is decreased.

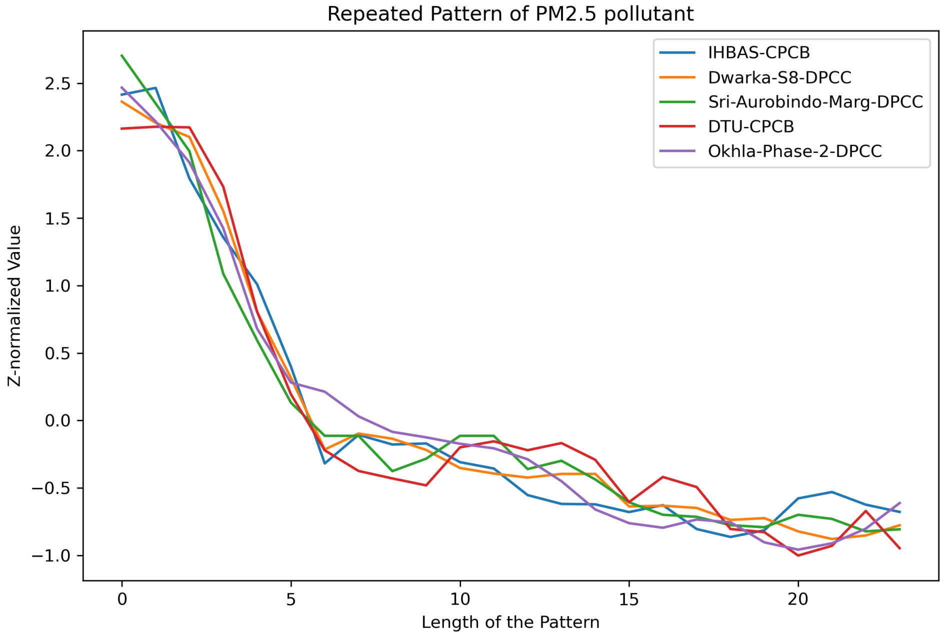


FIGURE 4.4: Repeated pattern of $PM_{2.5}$ pollutant of length 24 of different stations

In Fig.4.5, we group the stations according to the distance between stations and use euclidean distance to calculate the distance between them. In this case distance between S2(Dwarka-S8-DPCC) and S5(Okhla-Phase-2-DPCC) is minimum. So, we make a group with these two stations and the name of the group is S_{25}. After that distance from S1(IHBAS-CPCB) to group S_{25} is minimum and group name is S_{251}, distance from S4(DTU-CPCB) to the last created group is minimum. So, make a group S_{2514}, and the remaining station is grouped with the lastly created group i.e. S_{25413}. This way we can group the stations and represent this grouping using a dendrogram. In the dendrogram of the stations, the x-axis represents the station's code and the y-axis represents the distance between the stations.

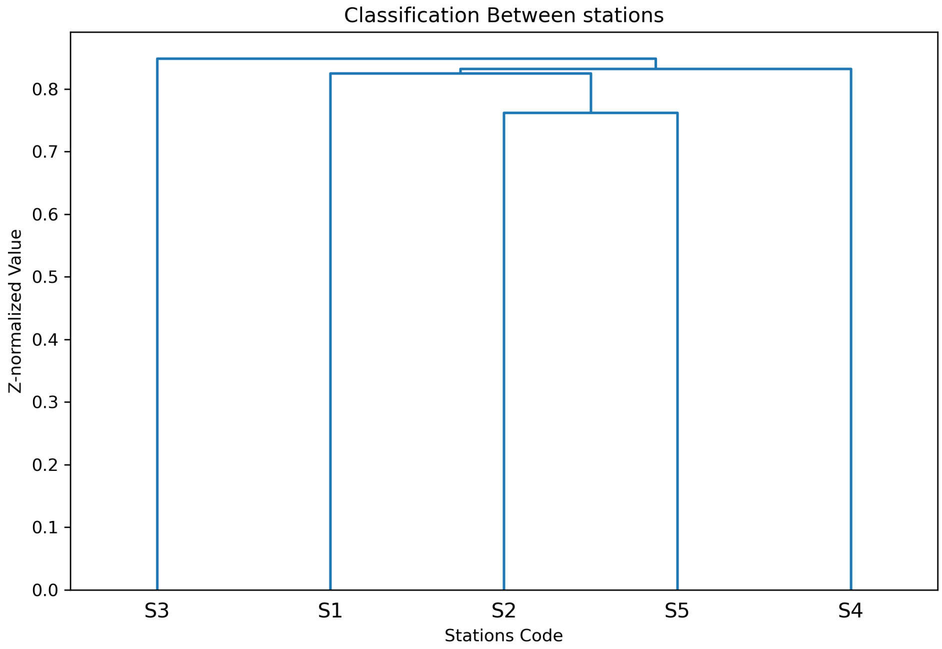


FIGURE 4.5: Clustering among the stations

Fig.4.6 represents repeated patterns that we get weekly basis(i.e. 24x7=168 hours) and classification between them. Repeated patterns of randomly selected stations of all the patterns are overlapped to each other from beginning to end of the pattern. Here we have also seen that the pattern that we discovered from different locations of a collection of time series data is that the value of the pollutant is high at the beginning for these respective stations, indicating that the pollution level of this pollutant is high, and as time passes, the level of pollution reduces.

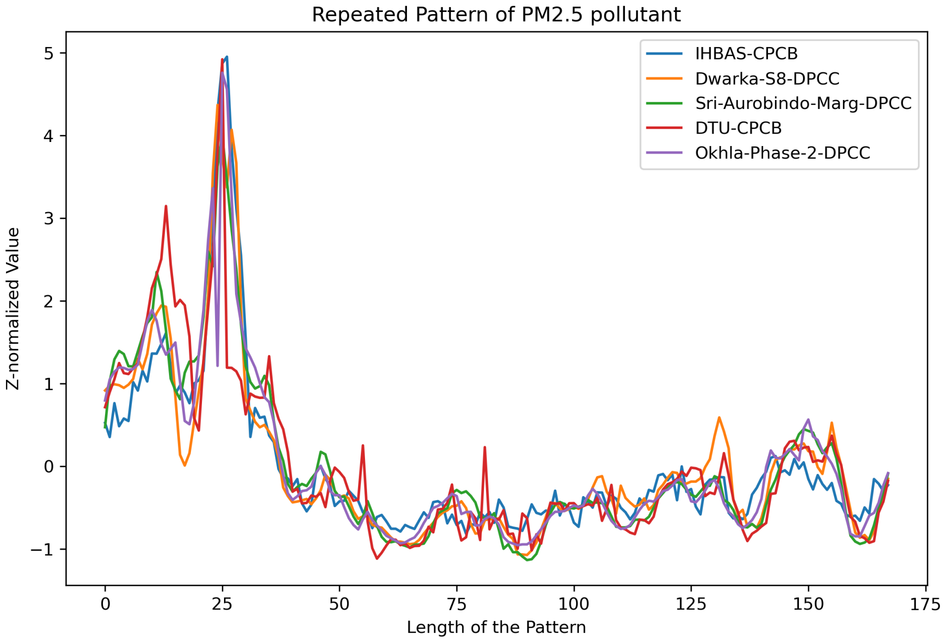


FIGURE 4.6: Repeated pattern of $PM_{2.5}$ pollutant of length 168 of different stations

Here the stations are clusters by calculating the distance between every station that is selected randomly in Fig.4.7. The distance between station S_3(Sri-Aurobindo-Marg-DPCC) and S5(Okhla-Phase-2-DPCC) is minimum so we make a class with them as S_{35}. Now among the remaining stations, we get S_2 as the minimum distance from the newly created class. So, the class that we created is S_{352}. Hence the stations that we have to make class them are S_1 and S_4. From these two stations, the linkage[8] method finds the minimum distance from S_1(IHBAS-CPCB) to S_{352} and S_1 is added to the existing class and the newly formed class is S_{3521}. Finally, one station S_4 is added to the class S_{3521} and the final class is S_{35214}.

[8] https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.linkage.html

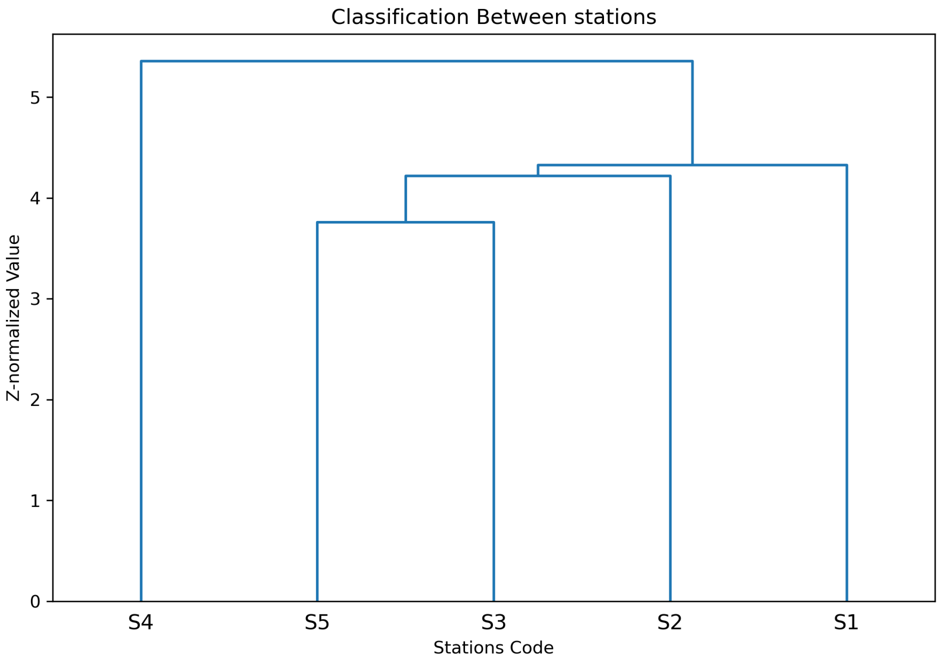


FIGURE 4.7: Clustering among stations

In the above, we discussed the repeated pattern of $PM_{2.5}$ pollutant of randomly selected five pollutant monitoring stations with 24-hour and 24x7 hour windows. Now in Fig.4.8 we found repeated pattern of 24x30 hour window. The patterns we have found in a particular position of time series data of five different stations of $PM_{2.5}$ pollutant, we can observe that the values of the pollutant in different stations at the beginning of patterns are low but when the time goes away the level of the pollutants are increasing accordingly. And at the end of the pattern, the level of pollution of $PM_{2.5}$ pollutant of Okhla-Phase-2-DPCC station is very high in comparison to other stations' pollution levels. All the values of the pollutant are z-normalized so we get a negative value at some point in time. We also observed that all the patterns overlapped with each other in most of the places of the pattern. From that observation, we can conclude that the behavior of the pollutant in that places where pollution levels of different stations meet each other is the same.

In the Fig.4.9 reveals how they grouped to each other and according to the behavior of the pollutant of different areas are grouped. The pollution level of station S_3(Sri-Aurobindo-Marg-DPCC) and S_5(Okhla-Phase-2-DPCC) are similar and the distance between values of stations are minimum. So we can make a cluster with S_3 and S_5 as S_{35}. From the remaining three stations, from S_4(DTU-CPCB) to cluster S_{35} is minimum and make cluster S_{354}. Now from S_2(Dwarka-S8-DPCC) to cluster S_{354} is minimum and make a new cluster S_{3542} and then remaining station S_1 is added to the cluster to form new cluster S_{35421}.

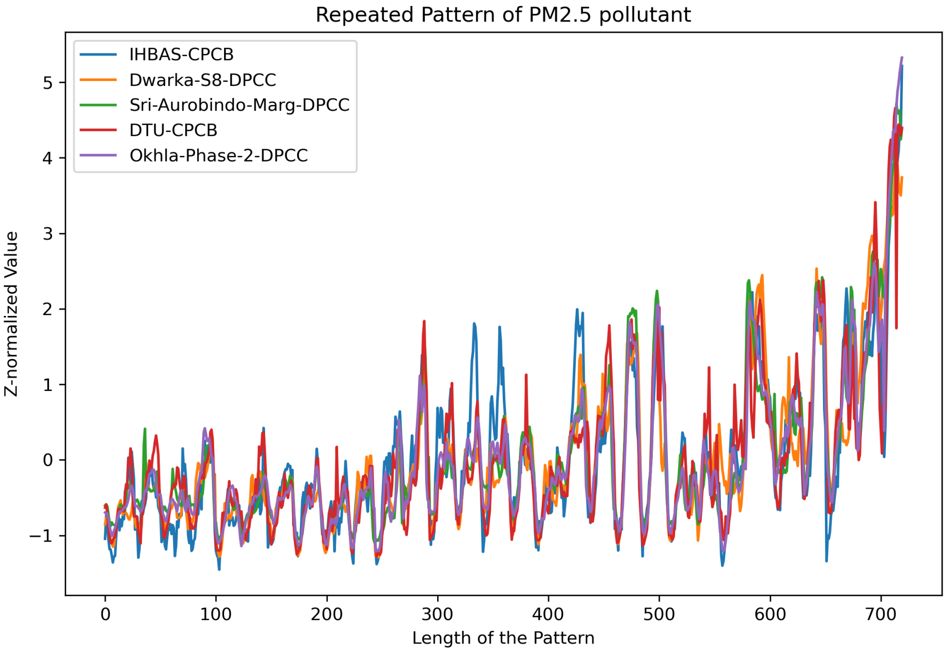


FIGURE 4.8: Repeated pattern of $PM_{2.5}$ pollutant of length 720 of different stations

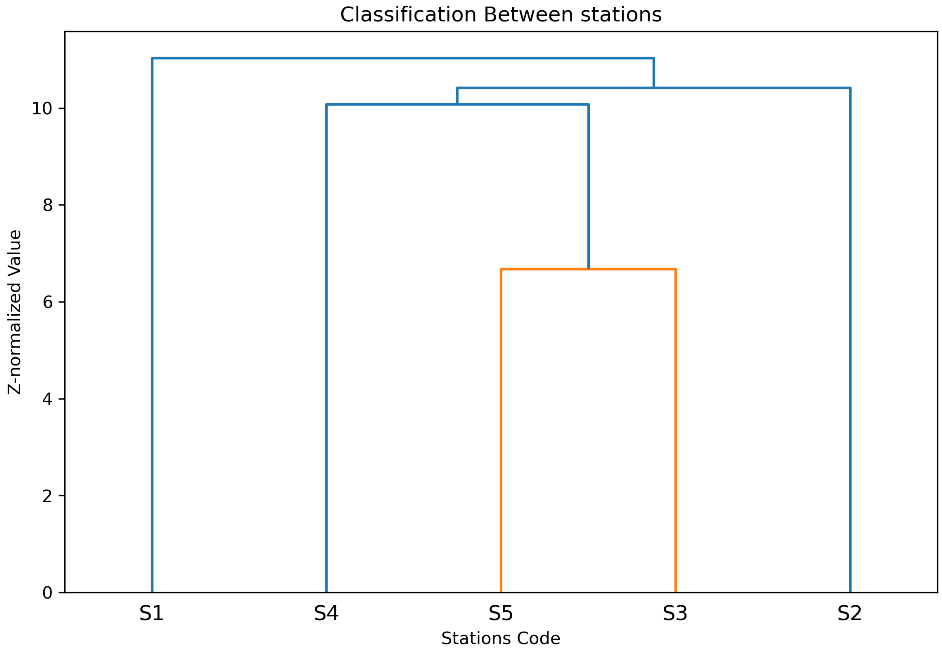


FIGURE 4.9: Clusters among stations

In the above section we have discussed the repeated pattern of $PM_{2.5}$ pollutant of three different granularity that is 24, 168 and 720. We found the repeated patterns of other pollutants like CO, NO, NH_3, Ozone etc. of these three granularities to randomly classify stations. Similarly we can discuss the results of other pollutants as we have discussed the results of $PM_{2.5}$.

Non-Normalized Results :

Above discussed repeated patterns are based on the z-normalized method applied to the pollutant values in different stations. In this section, we discuss repeated patterns without normalizing the values of the pollutant in different stations. In Fig.4.10 we found the repeated pattern of length 24 of $PM_{2.5}$ pollutant of five randomly selected stations. The pollution level of Sri-Aurobindo-Marg-DPCC station is low in comparison with the other four stations and the pollution level of IHBAS-CPCB station is high at the beginning of the day but after some time its value decreases. The pollution level in the other three stations is also high at the beginning of the day but decreases after some time.

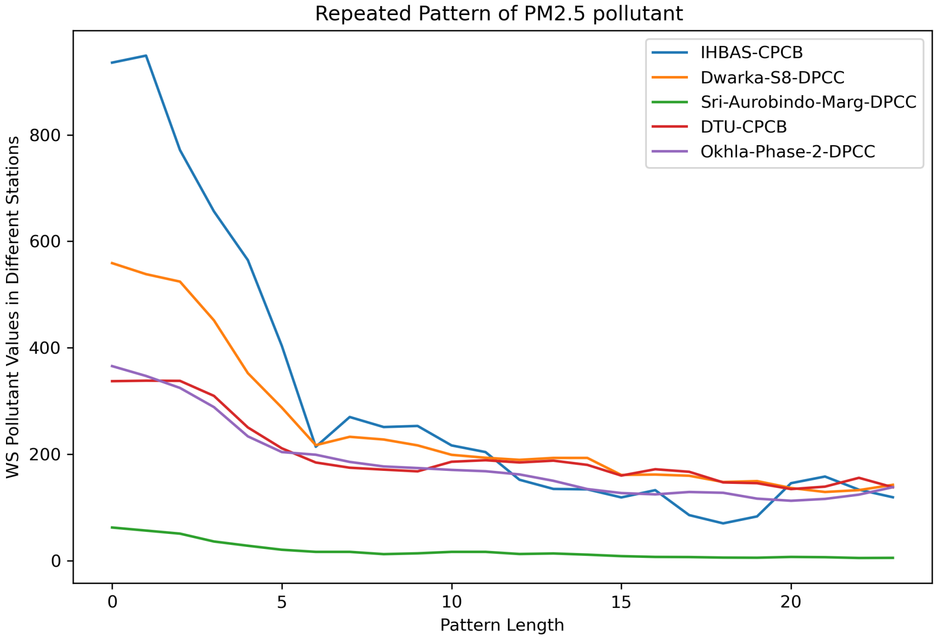


FIGURE 4.10: Repeated pattern of $PM_{2.5}$ pollutant of length 24 of different stations

In the Fig.4.11distance between stations, S_2 and S_5 is minimum. Hence a cluster is created with them as S_{25} and among the remaining three other stations S_1 has the minimum distance from the cluster S_{251}. So, station S_1 is added to it. Next station S_4 and then station S_3 is added to this cluster and formed S_{25143}.

In Fig.4.12 we have seen that patterns are similar to the pattern that we have found after applying the z-normalization method to the pollutant values of different stations. All the patterns are overlapped with each other and look similar. Hence we can say that behaviour is the same in a different station. And pollution level at the beginning of the week is high but with the passing of each day pollution level decreases.

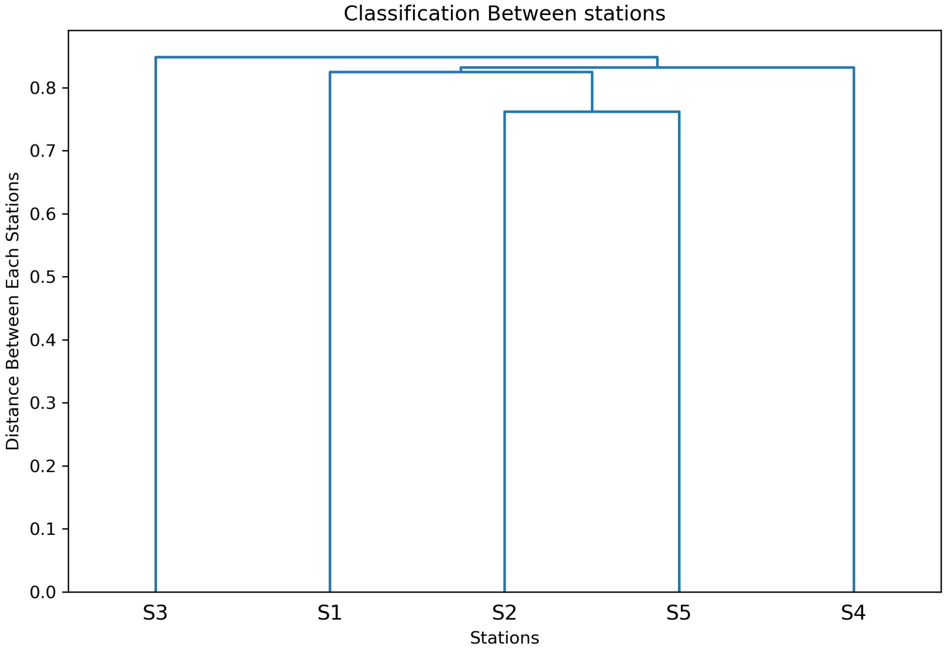


FIGURE 4.11: Clusters among stations

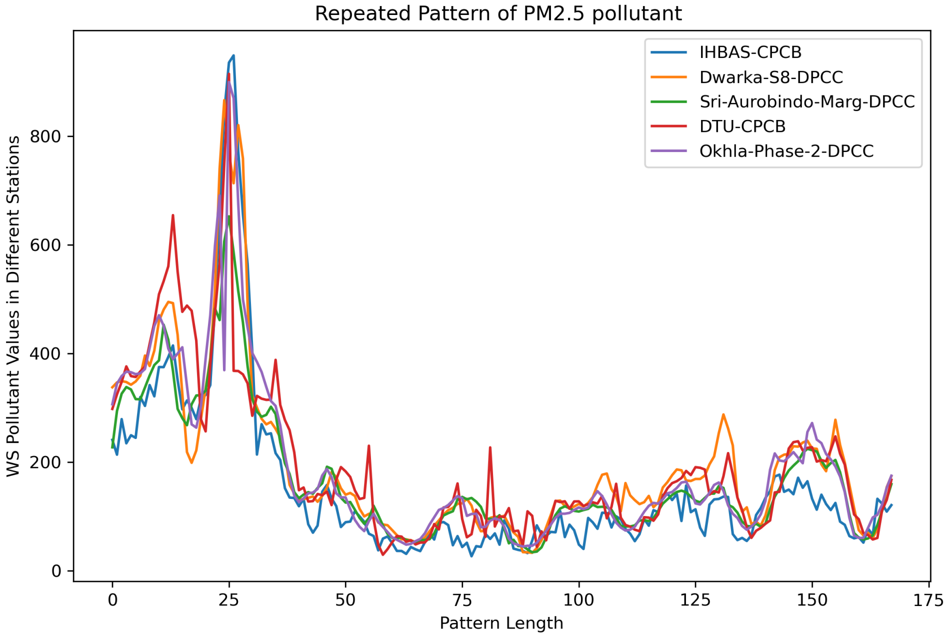


FIGURE 4.12: Repeated pattern of $PM_{2.5}$ pollutant of length 168 of different stations

The cluster is formed by taking station S_5 and station S_3 as they have the minimum distance among all other stations and form S_{35} is shown in Fig.4.13. Then from S_2 to the newly formed cluster is minimum hence it is added and form cluster S_{352} and station S_1 and S_4 is added to this cluster and form cluster S_{35214}.

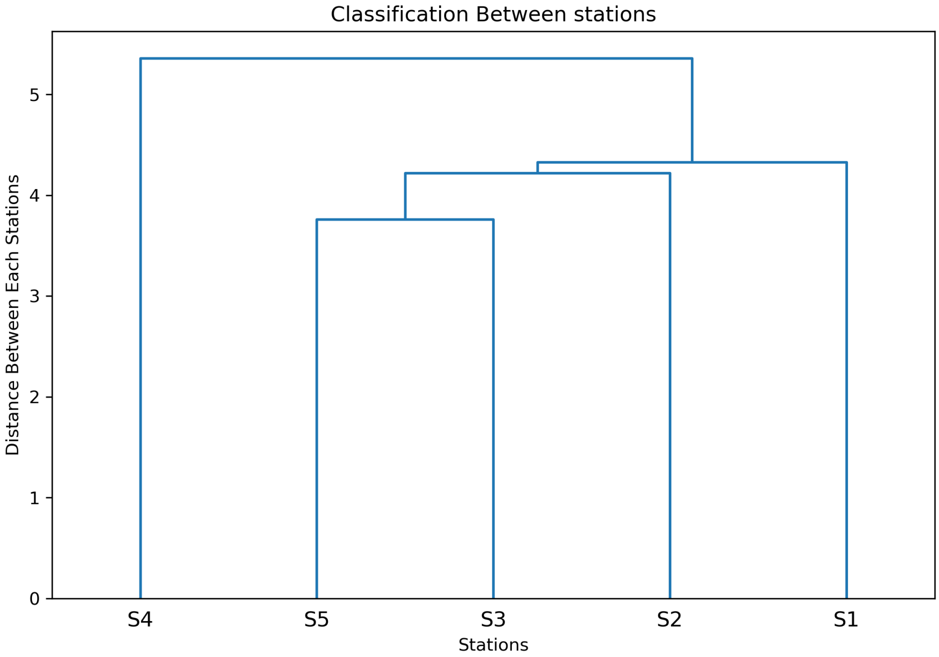


FIGURE 4.13: Cluster among stations

Lastly, we have discovered the repeated pattern of length 720 of PM2.5 of five different stations is shown in Fig.4.14. In the figure, we have seen that the pollution level at the beginning of the month is low, and at the end of the month pollution level increases. The patterns that we have found look like zig-zag and all the patterns are overlapped to each other in most the places.
Stations S_3 and S_5 form a cluster in the diagram above because their minimum distance is S_{35}, and S_4 has the shortest distance from the newly formed cluster of all the stations. Then S_2 and S_1 are added, and when we have a single cluster we construct S_{35421} at the end.

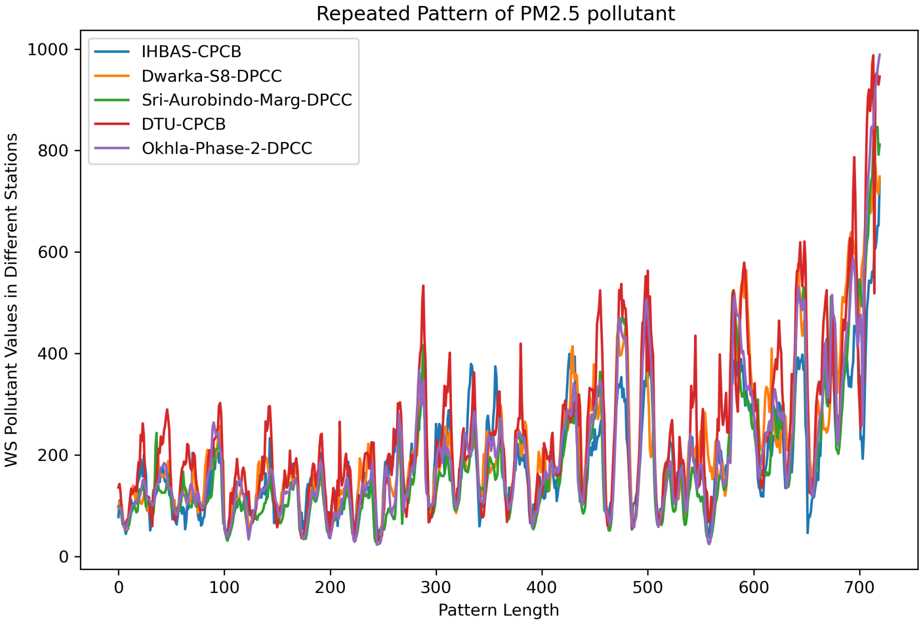


FIGURE 4.14: Repeated pattern of $PM_{2.5}$ pollutant of length 720 of different stations

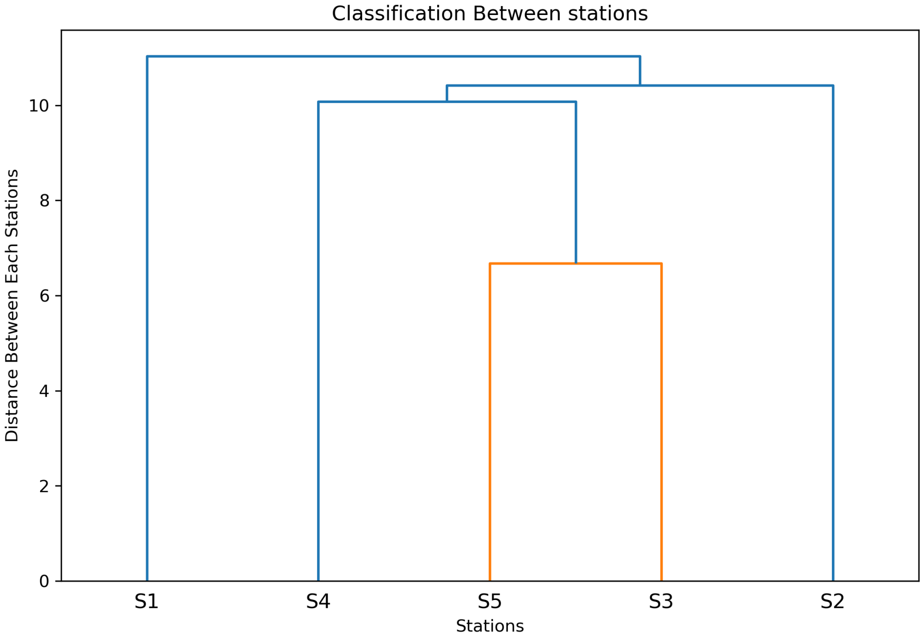


FIGURE 4.15: Cluster among stations

4.2.2 Population density based classification of stations :

We collect the population density of locations of Delhi from GeoIQ[9] where different pollution monitoring stations and population density is measured using per Km^2. The population density of Delhi is in the year 2020. Population density of Delhi

[9] https://geoiq.io/places/Vivek-Vihar/qWVYIcIzZG

in different areas is different and after analyzing it we observe that the population density of some areas are very dense and some less.

Normalized Results :

After finding the patterns from different stations of pollutant we normalize values of pollutant using z-normalized process.

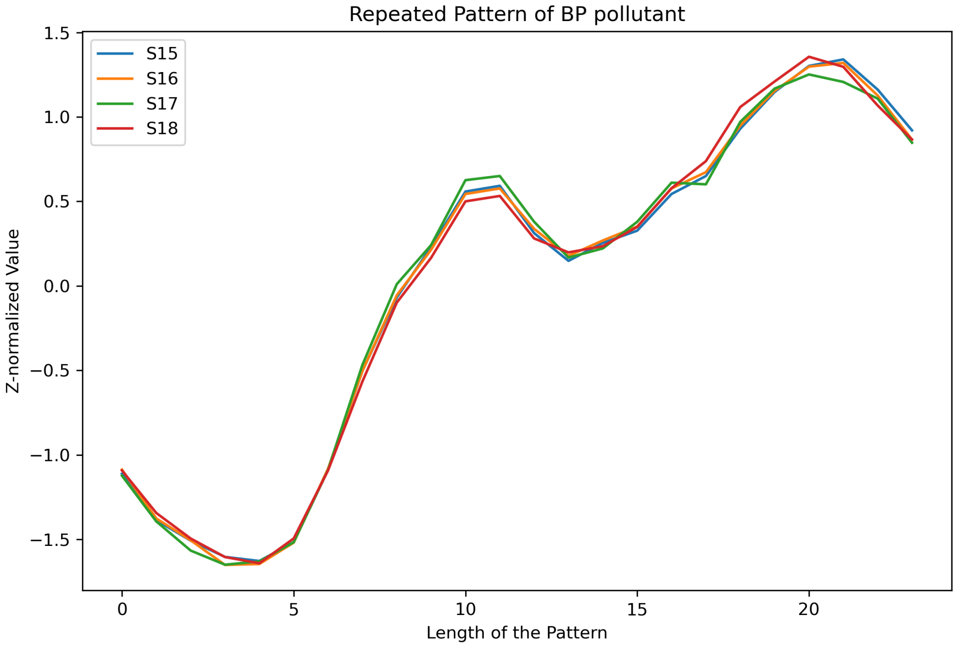


FIGURE 4.16: Repeated pattern of BP pollutant of length 24 of different stations

In the above Fig.4.16 the pattern we have found has a length of 24(i.e. daily basis) of stations 'Vivek-Vihar-DPCC', 'Mandir-Marg-DPCC', 'Nehru-Nagar-DPCC', 'Wazirpur-DPCC' which are situated in Delhi. From this figure, we observed that the patterns are exactly the same. The population density of these stations are nearly same and from these pattern we can say that the pollution level and the behaviour of BP pollutant is the same. And at the beginning of the pattern we can see that pollution level is low but when time increases the level of pollution is increases accordingly in these stations. All the station values of the pollutant are positive and negative because we have normalized using the z-normalization method.

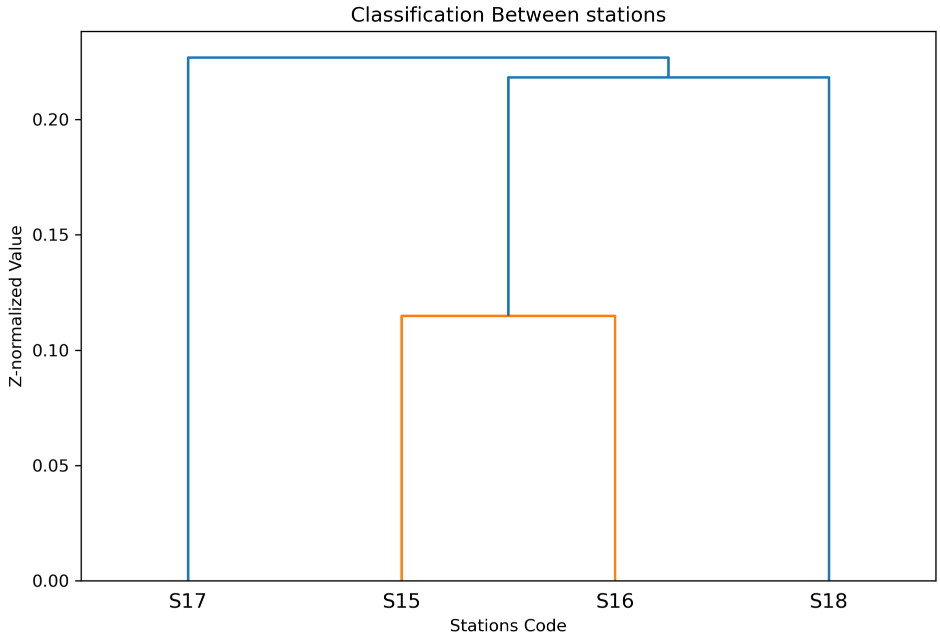


FIGURE 4.17: Clustering among stations

Classification of stations based on how patterns are similar is shown in Fig.4.17. Pollutant values of stations S_{16}(Vivek-Vihar-DPCC) and S_{15}(Shadipur-CPCB) are almost the same so we cluster them together. Next pollutant values of station S_{18}(Nehru-Nagar-DPCC) are almost similar to the newly created cluster so it is added to it. The remaining station S_{17}(Mandir-Marg-DPCC) is added to this cluster. The methodology of cluster formation has been discussed in the previous section.

In the below Fig. 4.18 patterns are found on weakly basis that is the length of the pattern 168. Patterns of BP pollutants in four different stations are overlapped to each other and patter are almost similar. So, from that observation, we can say that the behaviour of the pollutant of 'Vivek-Vihar-DPCC', 'Mandir-Marg-DPCC', 'Nehru-Nagar-DPCC', 'Wazirpur-DPCC' stations are the same. Initially, values of the pollutant in stations are low but at some point in time, it increases so we can say that the pollution level increases with the passing of each day level of pollution increases.

Fig.4.19 showing the values of BP pollutants of station S_{16}(Vivek-Vihar-DPCC) and S_{15}(Shadipur-CPCB) are similar to make a cluster with them and how they are similar is found by calculating the Euclidean distance to every station and take the minimum value. Next S_{17} is added to this cluster because its distance is minimum from the new form cluster and then only S_{18} is remaining and is added to this.

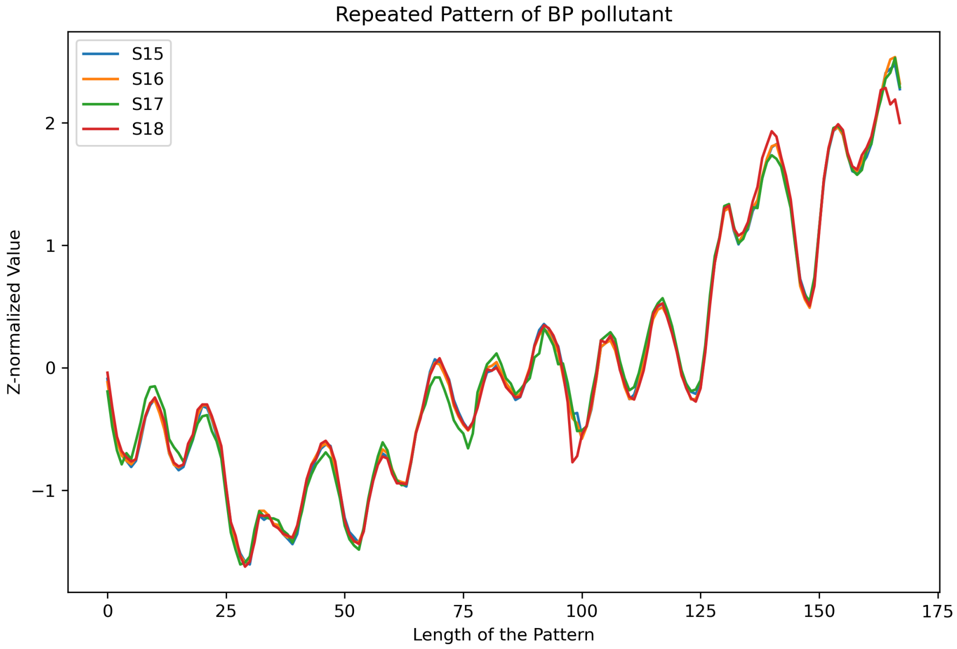


FIGURE 4.18: Repeated pattern of BP pollutant of length 168 of different stations

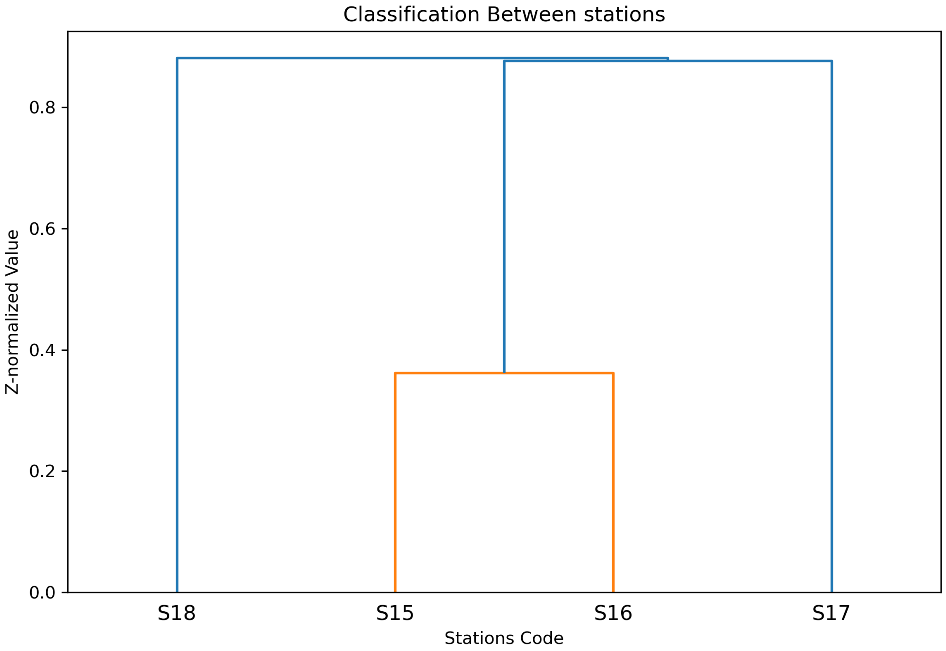


FIGURE 4.19: Clustering among stations

Now the patterns we found are the monthly basis data that is the pattern length is 720 of BP pollutant of four stations. The patterns are overlapped with each other at most of the points of other pollutants and some points are different in Fig.4.20. So, we can conclude that the behaviour of BP pollutants in these four stations is the same. The level of pollutant is initially low at the beginning of the month but day by day the level of the pollutant increases.

Clustering of stations shown in Fig.4.21 here four stations are clusters based on euclidean distance using a dendrogram. The distance between stations S_{15} and S_{17}

is minimum so make a cluster with them and the distance between stations S_{16} and S_{18} is the same as a previously created cluster. Hence these two stations are added to the previously created cluster.

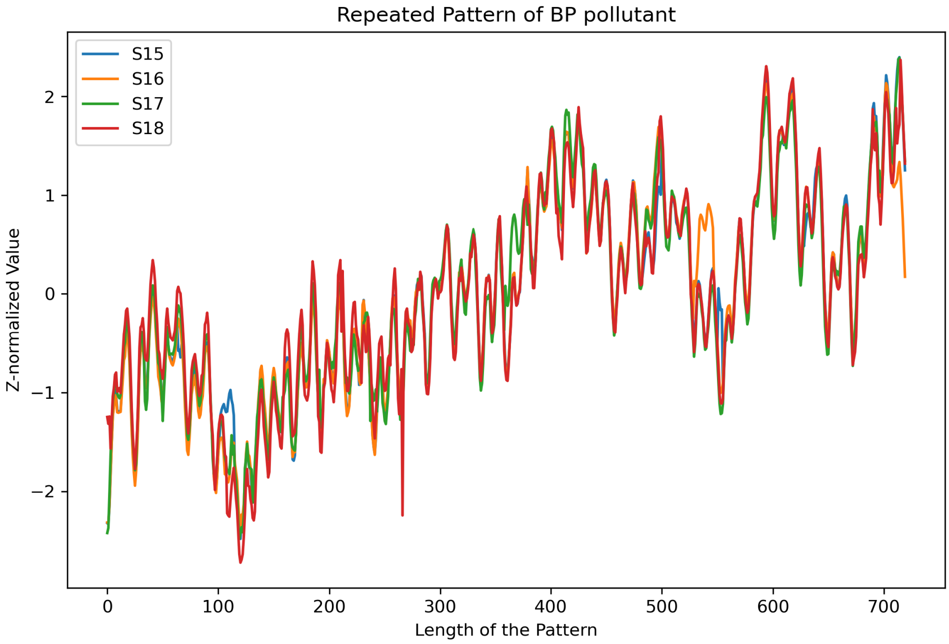


FIGURE 4.20: Repeated pattern of BP pollutant of length 720 of different stations

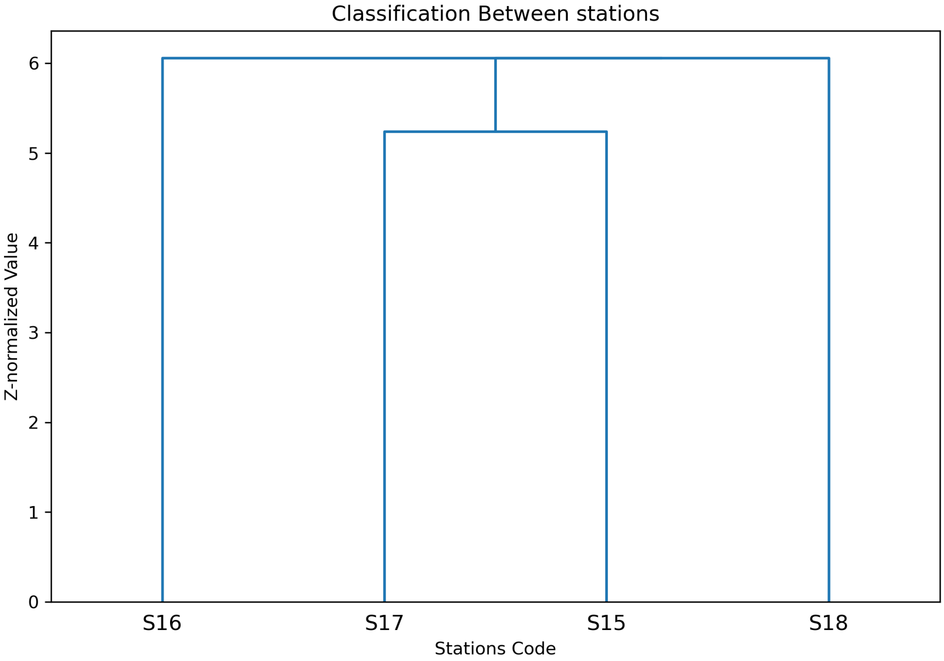


FIGURE 4.21: Clustering among stations

The repeating pattern of BP pollutant of three different granularities, 24, 168, and 720, is discussed above. We have discovered repeated patterns of other pollutants such as CO, NO, NH_3, Ozone, and others at this three granularity, and we presented the results of BP in the same way we discussed the results of other pollutants.

Non-Normalized Results :

Segregation of stations using population density of the area of pollution monitoring stations have been discussed in the previous section and also discussed the patterns that are created using the z-normalized method. Now we have to discuss the repeated pattern of BP pollutants without the normalization method. In the Fig. 4.22 four patterns of length 24 are generated from pollutant values of four different stations. We can observe that the pollution level of BP pollutant in station S_{16} is low at the beginning of the day but after some time it increases rapidly and pollution level of station S_{18} is high in comparison with other stations.

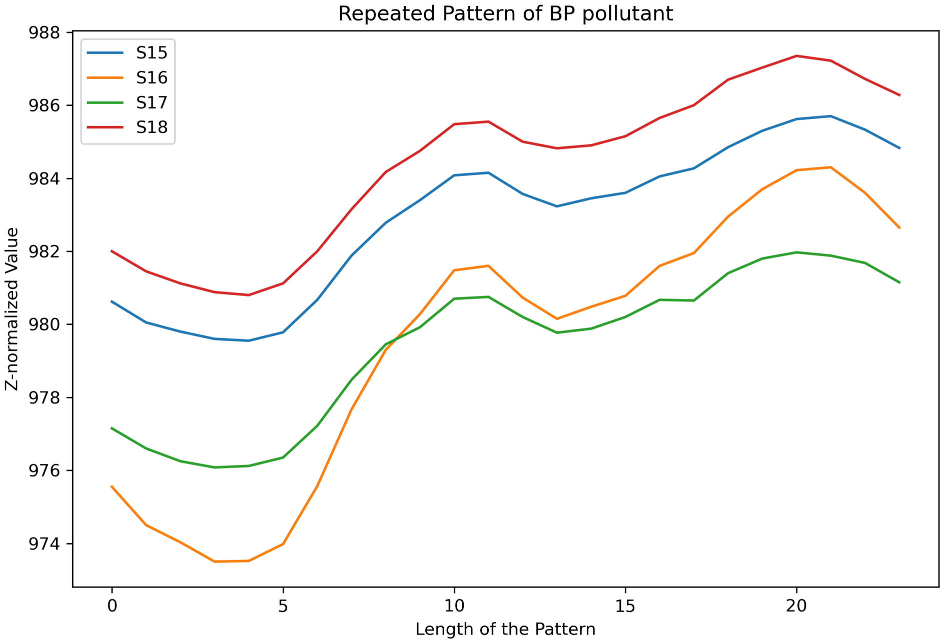


FIGURE 4.22: Repeated pattern of BP pollutant of length 24 of different stations

Here stations S_{15} and S_{16} are combined with each other to form a cluster on the basis of distance. Then among the stations S_{17} and S_{18}, the distance from S_{18} to the newly formed cluster is minimum so it is added to this. And finally, S_{17} is added to this cluster as only one station is left.

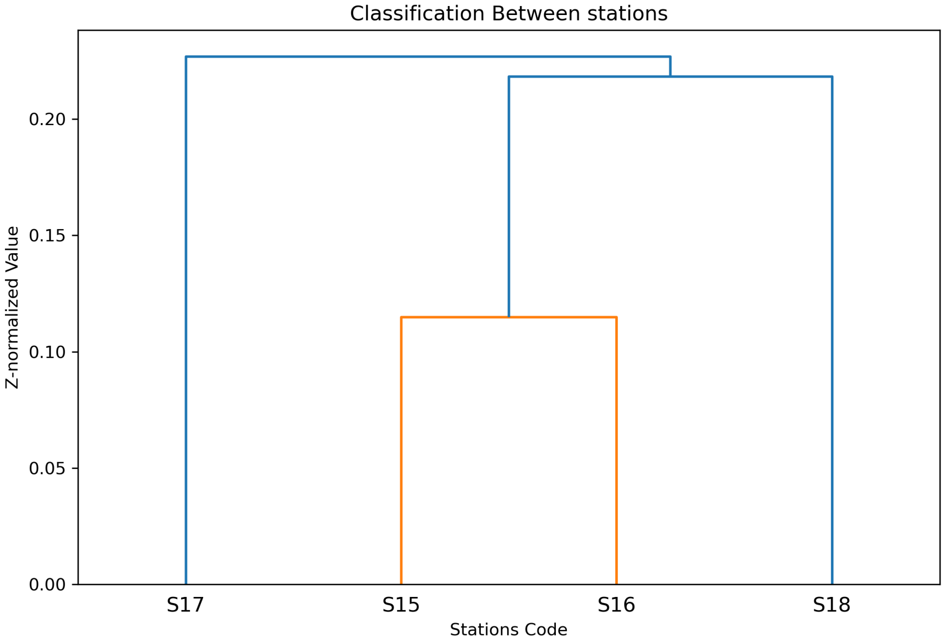


FIGURE 4.23: Clustering among stations

Here in Fig.4.24, we have seen that the pattern of length 168 is discovered from the BP pollutant of Vivek-Vihar-DPCC, Mandir-Marg-DPCC, Nehru-Nagar-DPCC, and Wazirpur-DPCC station. The pollution level of both granularity 24 and 168 is the same (i.e., high) for station S_{18}. In comparison to a repeated pattern of length 24, all the patterns are overlapped to each other but here stations are non-overlapped to each other.

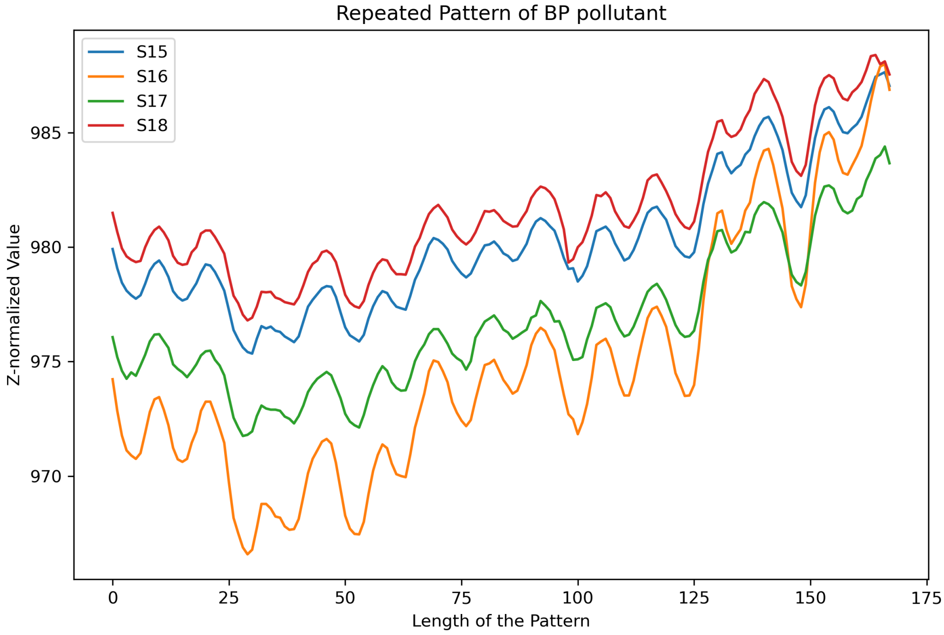


FIGURE 4.24: Repeated pattern of BP pollutant of length 168 of different stations

On the basis of Euclidean distance station S_{15}, and station S_{16} create a cluster, and then station S_{17} is added to this and forms a cluster with these three stations.

Then station S_{18} is added to this. and the final cluster is formed.

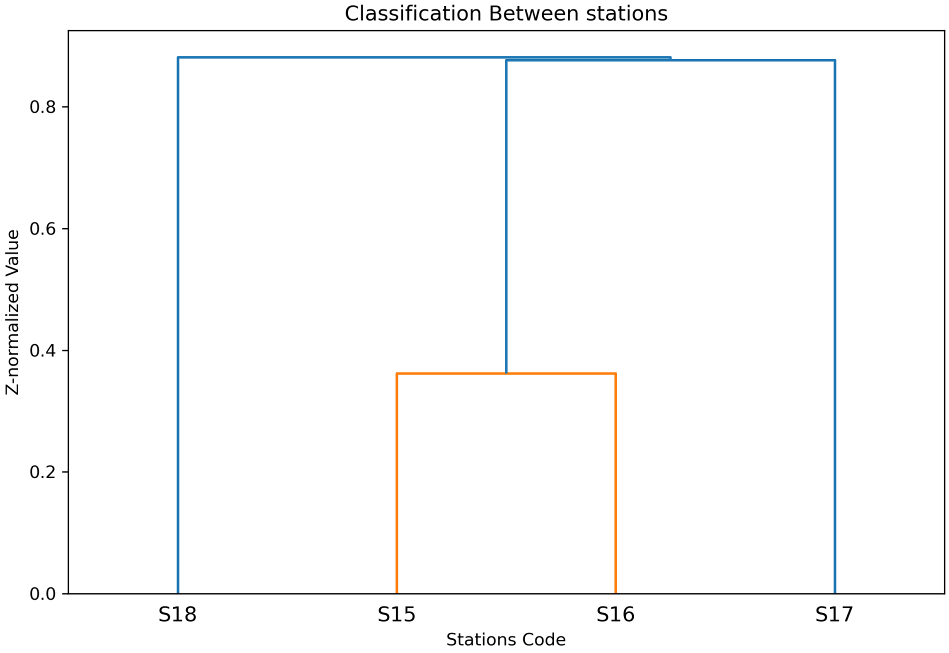


FIGURE 4.25: Clustering among stations

For the pattern of length 720 in Fig.4.26, all the patterns look similar but the values are not the same in a different station. Here also pollution level is high for station S_{18} and low for station S_{16} and pollution level rest of the stations are between these two stations.

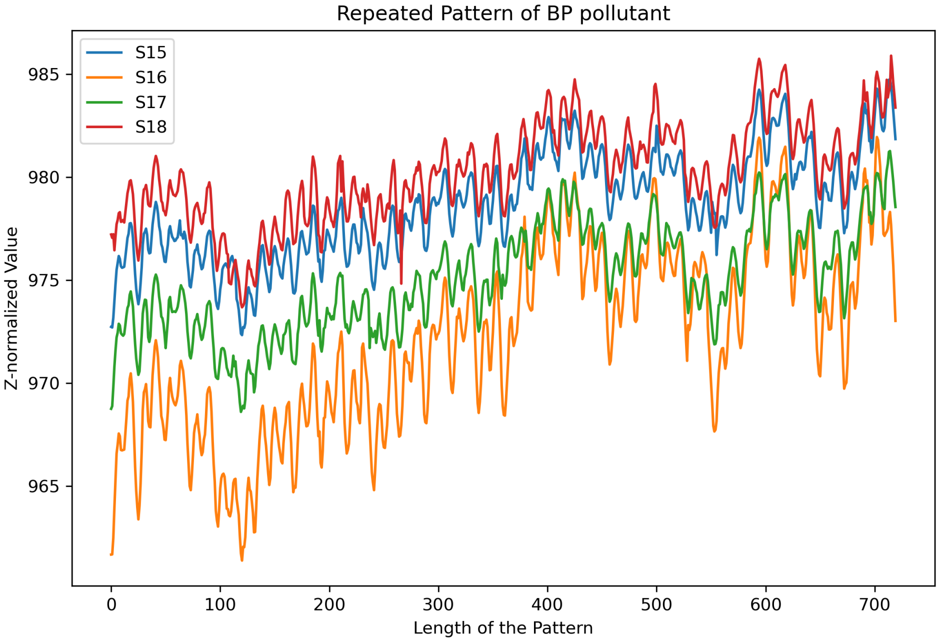


FIGURE 4.26: Repeated pattern of BP pollutant of length 720 of different stations

Now stations S_{15} and S_{17} are grouped together to form a cluster as they have minimum distance pair among the other stations' pairs. Then stations S_{16} and S_{18} are added together to form a single cluster.

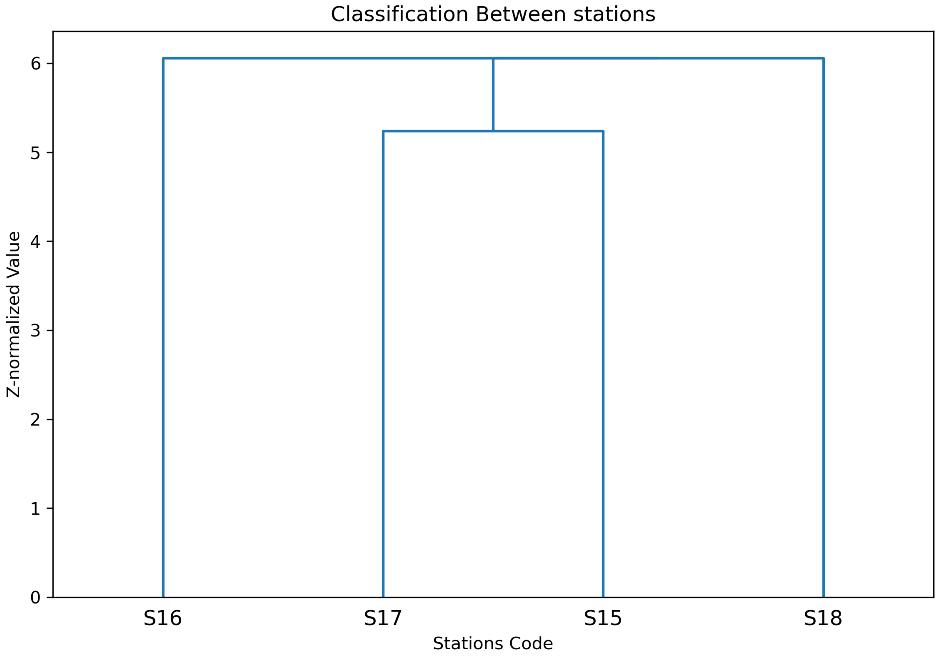


FIGURE 4.27: Clustering among stations

In this section we have discussed elaborately how repeated pattern of BP pollutant discovered based on population density of three different granularity. The same way we can elaborate repeated patterns of stations of other pollutants

4.2.3 Geolocation based classification of stations :

Geolocation is the coordinate(latitude and longitude) of an area in the globe. Each pollution monitoring station has a coordinate value and from these coordinate values we segregate the stations and the k-means clustering algorithm is used. Using this machine learning model we cluster thirty-two stations into five different clusters based on how close stations belong. Here we discuss only one cluster among five clusters of stations 'Dwarka-S8-DPCC', 'Aya-Nagar-IMD', 'Najafgarh-DPCC' in Delhi in three different granularity.

Normalized Results :

In this section, experimental results are collected using the z-normalized method of every subsequence of time series data of pollutants in different stations in Delhi. In Fig.4.28 we found the repeated pattern of Ozone of pollutant monitoring stations 'Dwarka-S8-DPCC', 'Aya-Nagar-IMD', 'Najafgarh-DPCC' in Delhi daily that is the 24-hour window. Patterns that we found of pollutant Ozone in three stations are almost similar from start to the end of the pattern but only a few parts of the pattern are different and most of the parts are overlapped to each other. So, we can say that the behavior of these patterns in three different stations is the same. All the patterns look like English Alphabet **A** and the pollution level of Ozone at the beginning of the day is very low but as time passes its value increases. When the time is around 4 PM the level of the pollutant reaches a peak but after 4 PM pollution level decreases and some values of the pollutant are negative because we apply the z-normalize method to the pollutant values of the stations. Here we discuss repeated pattern of Ozone of length 24, 168 and 720.

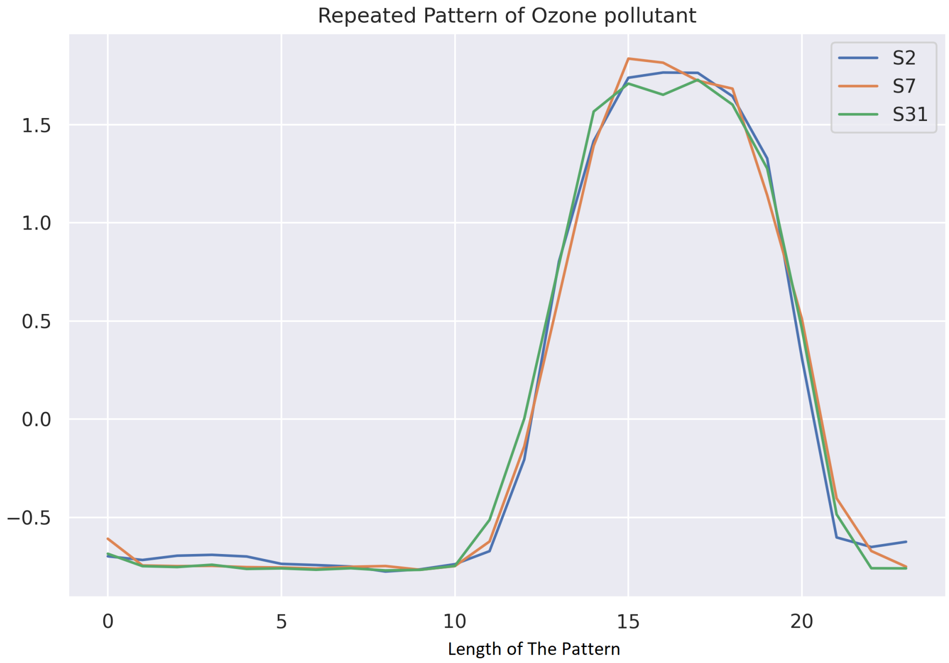


FIGURE 4.28: Repeated pattern of Ozone of length 24 of different stations

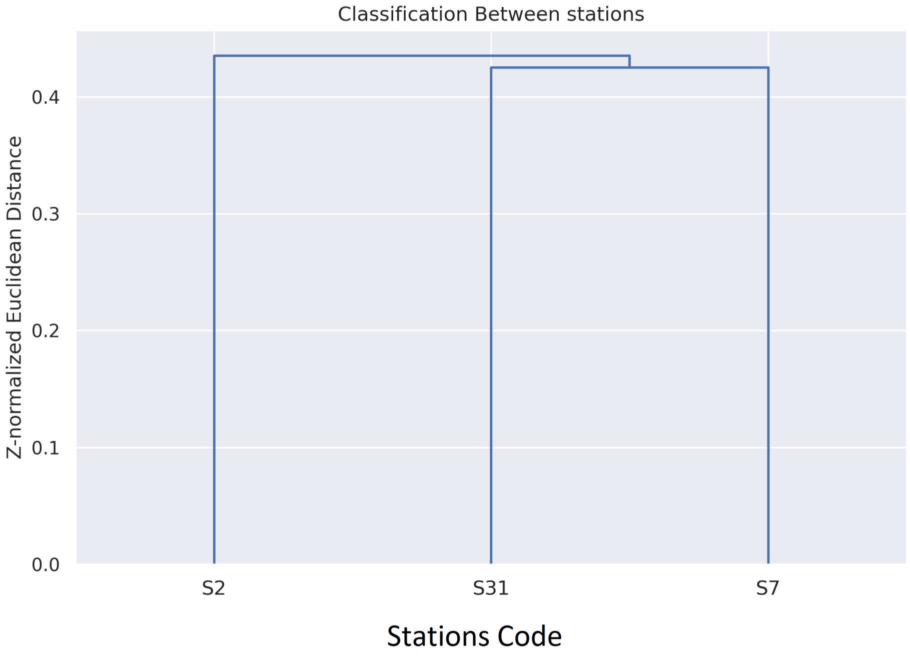


FIGURE 4.29: Clustering among stations

In the above Fig.4.29 shows that the distance between stations S_7(Aya-Nagar-IMD) and S_{31}(Najafgarh-DPCC) is minimum so make a cluster with them and one

station is remaining so it is added to the newly formed cluster. And the z-normalized distance between stations varies from zero to one.
The behavior of repeated pattern of these three stations of length 168(weekly) in a zig-zag manner in Fig. 4.30. That is pollution level on the first day is from high to low and pollution level on the next day is from low to high and so on. Most of the places of the pattern are overlapped and all the patterns look the same. So, from here we can conclude that the behavior of this pollutant in these stations is the same.

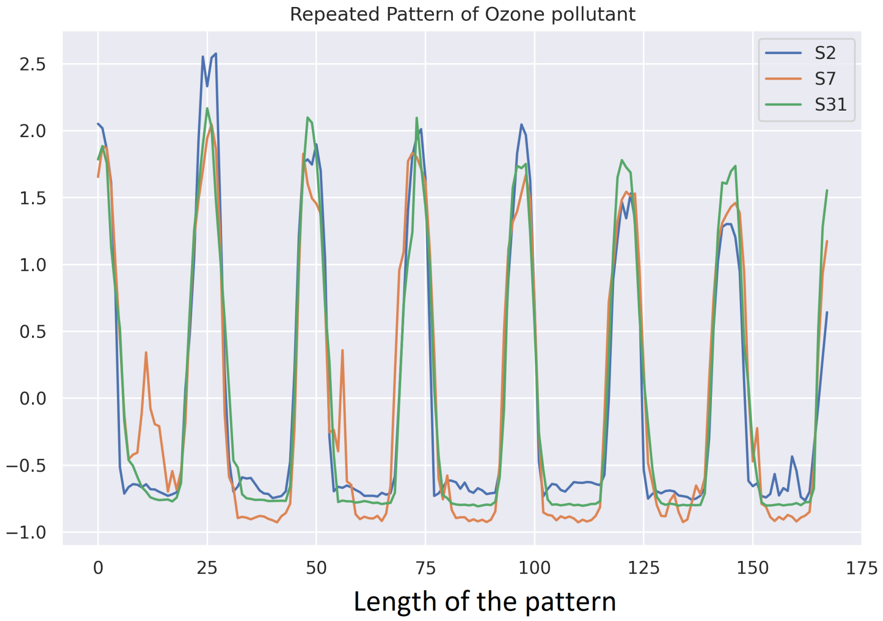


FIGURE 4.30: Repeated pattern of Ozone pollutant of length 168 of different stations

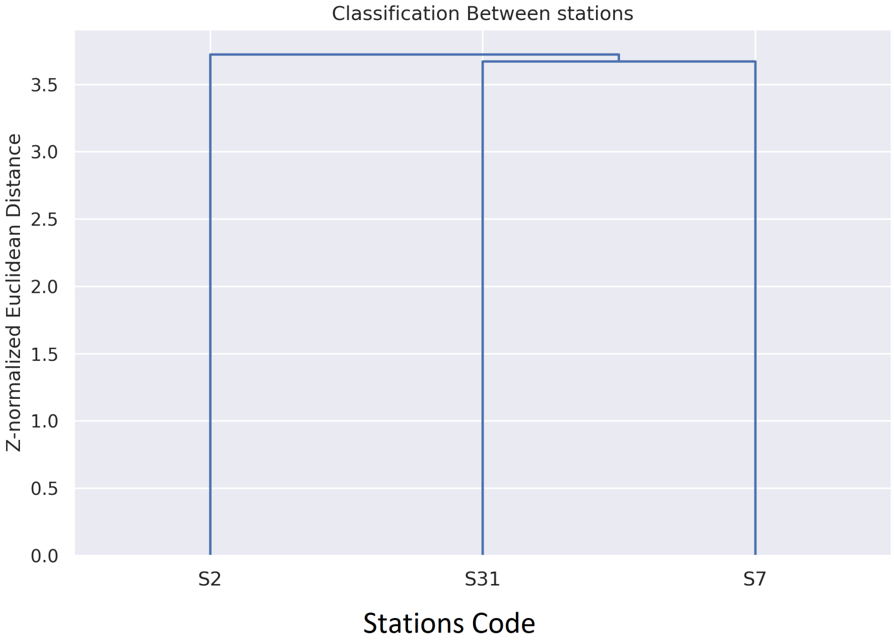


FIGURE 4.31: Clustering among stations

Here clustering is done the same way as discussed in the previous part of this section but the only difference is that distance between stations varies from 0 to 4 and distance is a normalized value shown in Fig.4.31. Stations S_7 and S_{31} are close to each other on the basis distance of the pollutant of these two stations and make a cluster with them and the distance from S_2 is minimum then it is added to the newly formed cluster.

We have already discussed the repeated pattern of pollutants of different stations with two different granularity 24-hour and 168-hour windows. Now we discuss the repeated pattern of Ozone in three different stations with a 720-hour window shown in Fig.4.32. This pattern also looks like a zig-zag manner and vertical parts of the patterns are overlapped in the upper and lower parts, some patterns are overlapped some are not. So, the behaviour of the patterns in these stations are the same, and the pollution level of each day of the month is almost the same but in a few portions of the pattern, the pollution level is low.

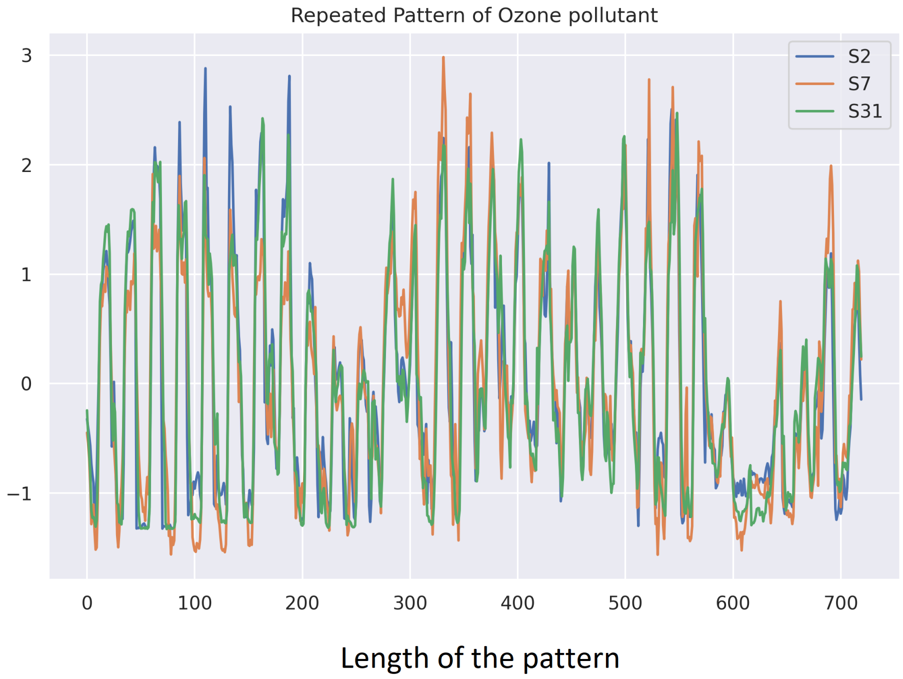


FIGURE 4.32: Repeated pattern of Ozone pollutant of length 720 of different stations

In the below figure, the dendrogram shows the clustering of stations graphically. Station S_7 and station S_{31}'s pollutant values are almost similar that's why a cluster is formed between these two stations next remaining station is added to this cluster and the distance of these pollutant values of stations is between 0 to 11. In this figure x-axis labeled as stations code and y-axis labeled as distance between clusters.

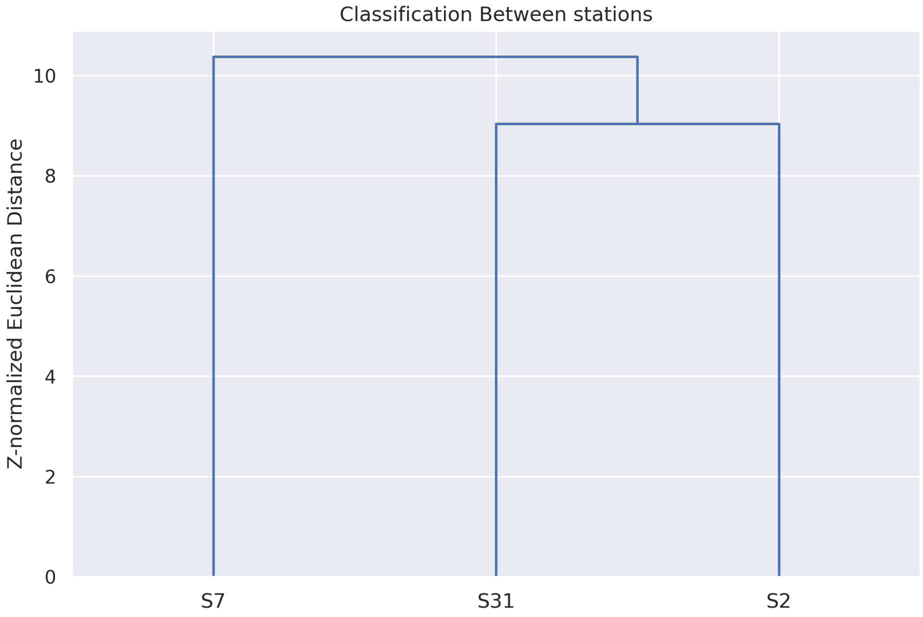


FIGURE 4.33: Clustering among stations

In the same way, we can discuss other pollutant instances such as CO, NH3, Benzene, etc., and how clusters are formed based on the pollutant value of different stations of different granularity.

Non-Normalized Results :

In this section, we discuss all the patterns of Ozone in different stations of geolocation-based segregation of stations without normalizing the pollutant values. In Fig.4.34 we have seen that at the beginning of the day pollution level of ’Dwarka-S8-DPCC’, ’Aya-Nagar-IMD’ and ’Najafgarh-DPCC’ station is the same and low but when time passes only pollution level of station ’Najafgarh-DPCC’ increases significantly but pollution level of other two stations increases but not very much. In comparison with the normalized pattern, the pollution level of all the stations is increasing significantly but in the non-normalized pattern, only one station’s pollution level increases significantly.

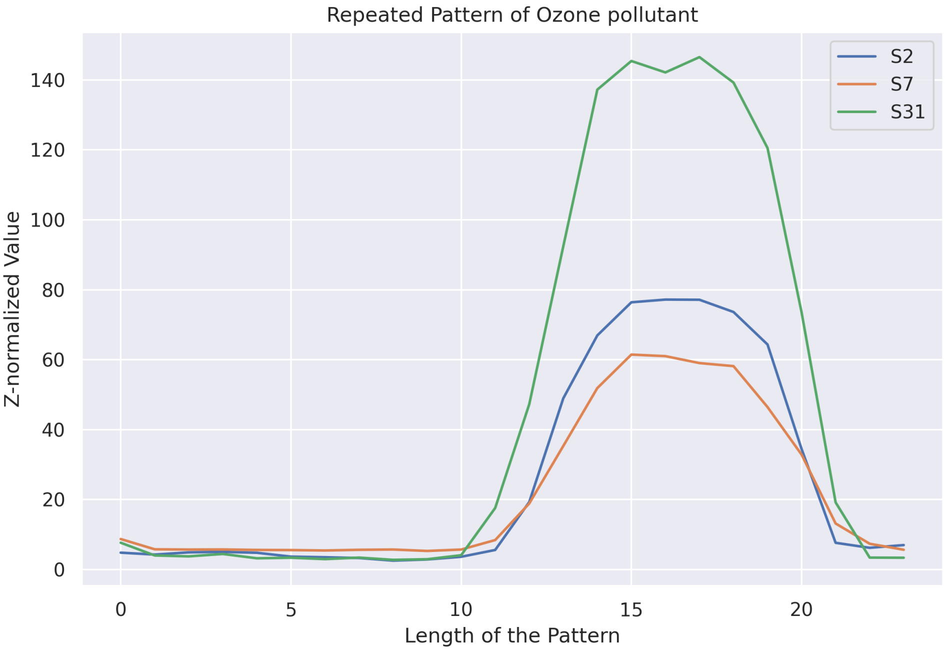


FIGURE 4.34: Repeated pattern of Ozone pollutant of length 24 of different stations

Station S_7 and S_{31} form a cluster as they have the minimum distance among all the stations. Then station S_2 is added to the cluster because only one station is left and clustering is done as there is only one cluster remaining.

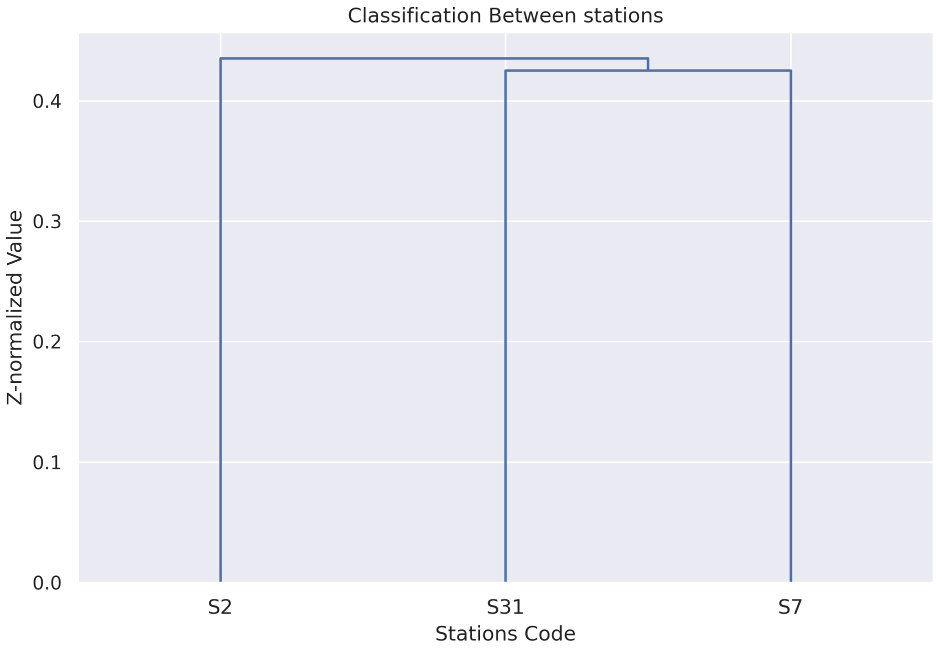


FIGURE 4.35: Clustering among stations

In the above, we have discussed the repeated pattern of length 24. Now we analyze the repeated pattern of length 168 that is a weekly basis. In this figure, we have seen that the pollution level in every station is not certain it varies every moment of the day. Here also pollution level of station S_{31} is increased remarkably. But some time of the week pollution level is very low in all the stations. The pollution level of the other two stations is not increasing enormously.

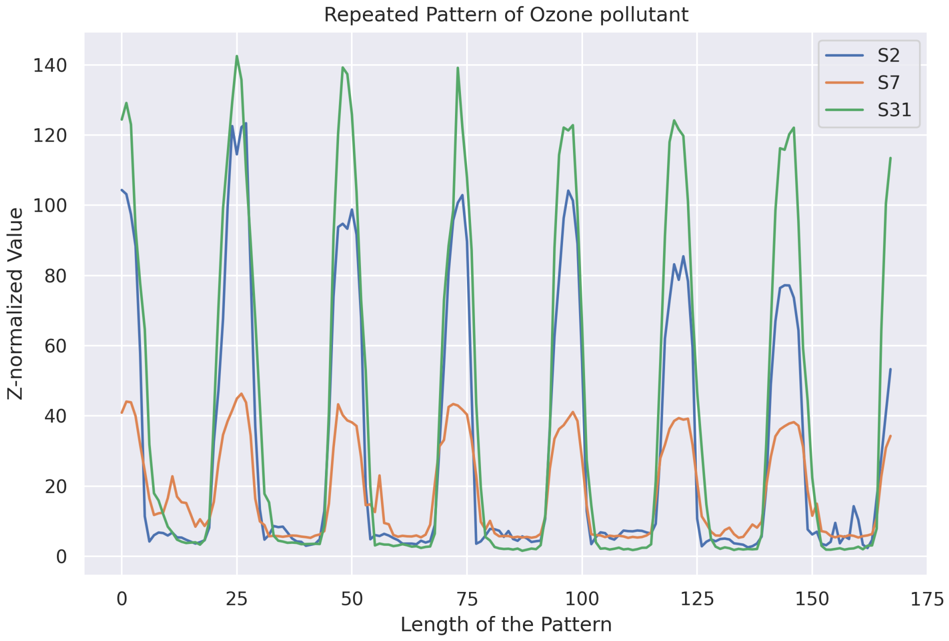


FIGURE 4.36: Repeated pattern of Ozone pollutant of length 168 of different stations

Station S_{31} and S_7 merge to form a cluster in Fig.4.37 based on the distance between every pair of stations. Then S_2 is added to this cluster.

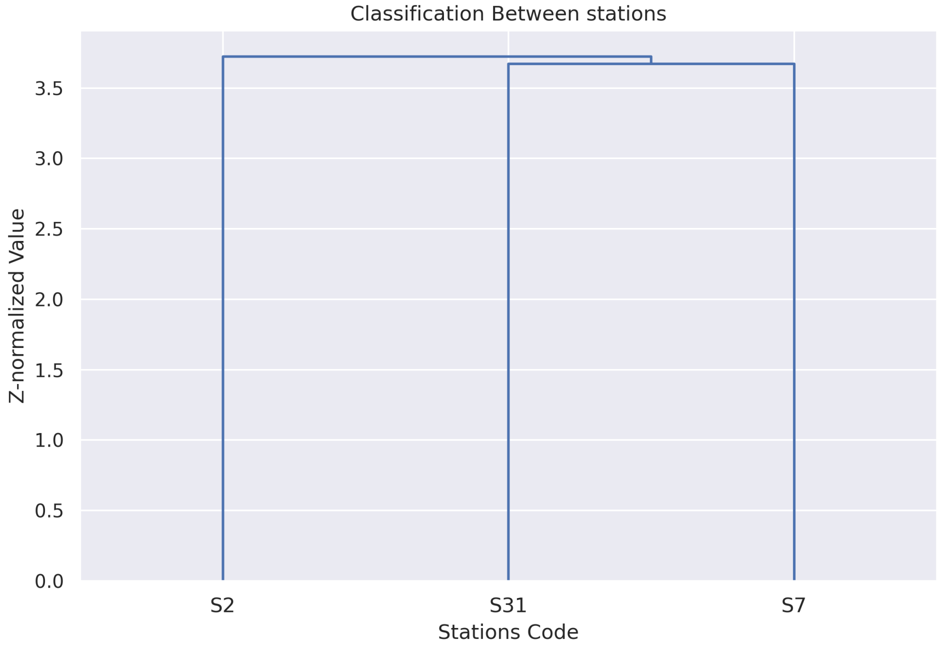


FIGURE 4.37: Clustering among stations

Now in Fig.4.38 we have seen that the behaviour of the pattern in all the stations is very uncertain that is pollution level is low on some days of the month and some days is high. The pollution level of station S_7 is low and S_2 is high.

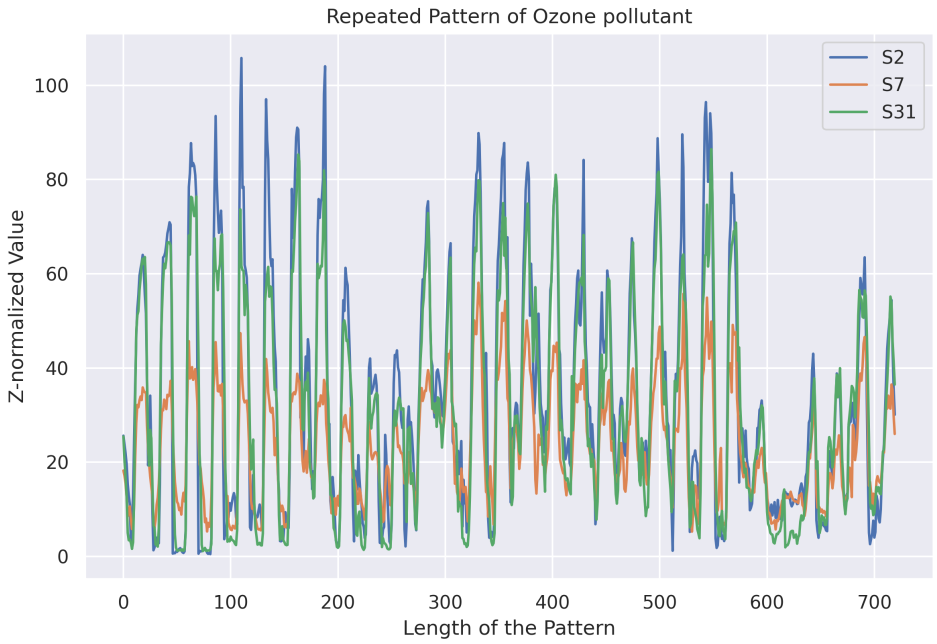


FIGURE 4.38: Repeated pattern of Ozone pollutant of length 720 of different stations

Based on the distance between every pair of stations, stations S_{31} and S_2 combine to form a cluster. Because between station S_{31} and S_2 distance is minimum. After that, station S_7 is added to the cluster as only one station remaining and its distance from the newly created cluster is minimum.

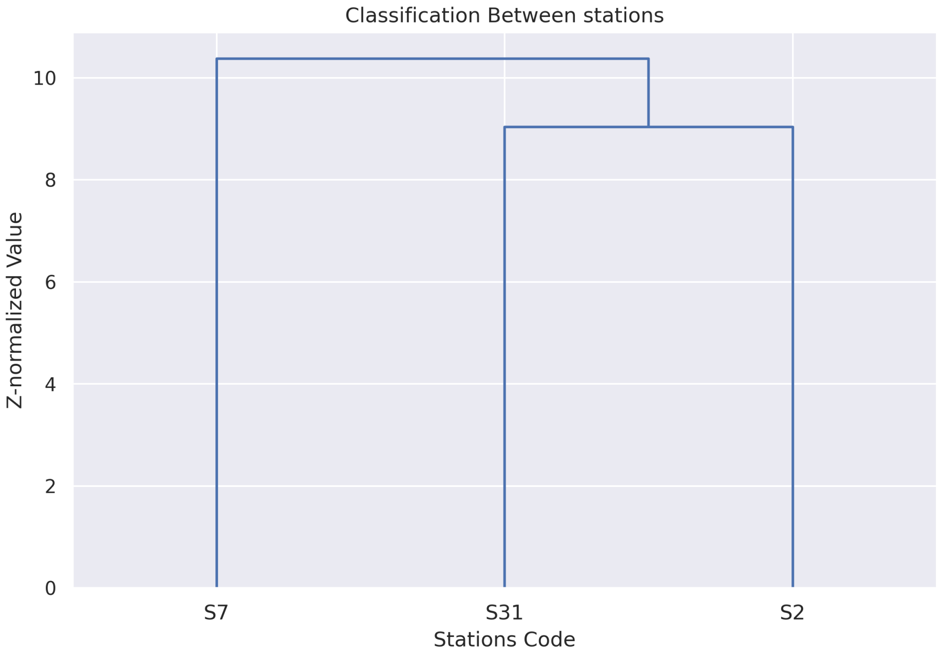


FIGURE 4.39: Cluster among stations

4.2.4 Landuse based classification of stations :

In landuse based classification of stations, each station belongs to a particular landuse; the categories of landuse being industrial, historical, commercial, residential, etc. . We took nine different landuse of Delhi and put a minimum of five stations in each landuse on basis of the calculating distance between geolocation of landuse and station. In this section, we discuss repeated patterns of RH pollutants on a daily, weekly, and monthly basis.

Normalized Results :

For all values of the pollutant in different stations, we normalize values using the z-normalized method. In Fig.4.40 we have seen that the pattern we have found is of length 24 and all the pattern of the four stations looks the same. At the beginning of the day station S_9 and S_{26} are overlapped to each other but when time passes in a day pollutant values of these two stations are non-overlapped to each other. So, the behaviour of the pollutant in these two stations at beginning of the day is the same but after some time they are different. And S_{17} and S_{32} are slightly different from other patterns. So the behaviour of the pollutant is different from others. The pollution level is high at the beginning but after 2 PM pollution level decreases.

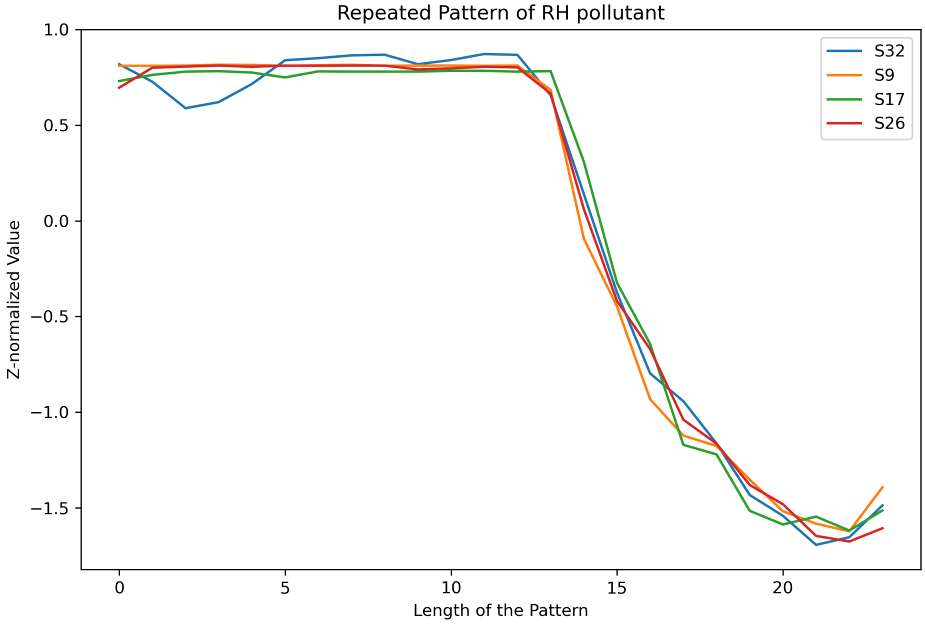


FIGURE 4.40: Repeated pattern of RH pollutant of length 24 of different stations

From the below figure, we can say that the distance between stations S_9, S_{17}, and S_{16} is the same, as well as the minimum, and from station S_{32} distance, is large. So, create a cluster with these three stations, and the S_{32} is added to the cluster.

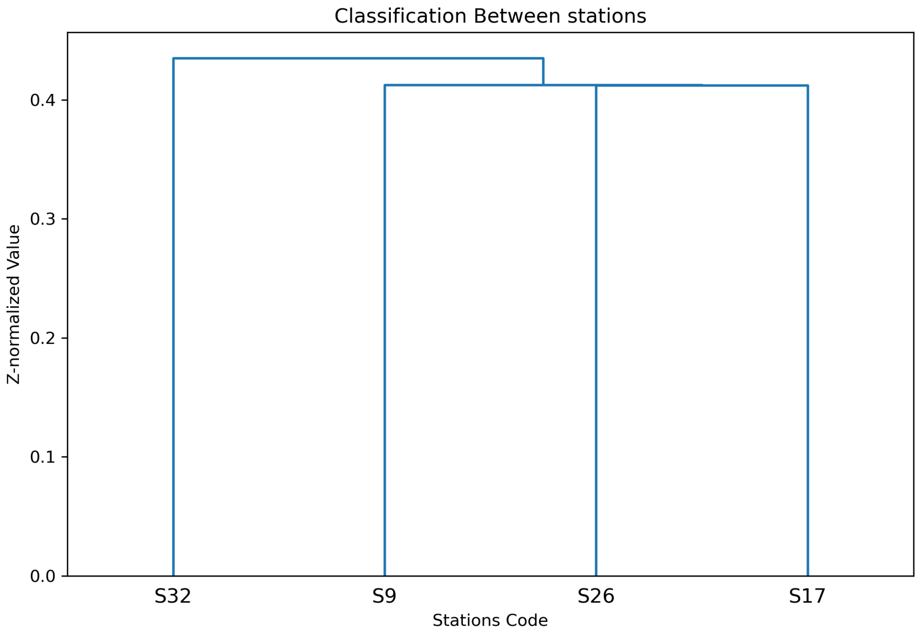


FIGURE 4.41: Cluster among stations

Now patterns that are found from RH pollutant in Fig.4.40 of length 168 (i.e. weekly) are almost similar and look like a zig-zag manner. All the vertical parts of the patterns are similar but the upper portion of the pattern is dissimilar. Hence we can say that when the pollution level is low and medium the reaction of patterns is same but when the pollution level is high the reaction is different.

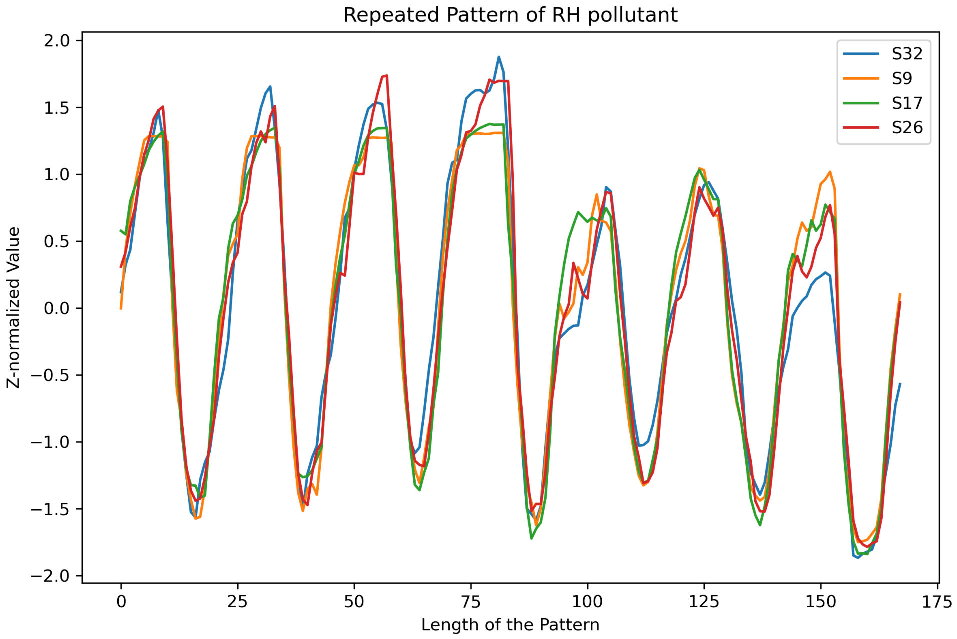


FIGURE 4.42: Repeated pattern of RH pollutant of length 168 of different stations

In the above figure, a cluster of four stations created based on the Euclidean distance between them. The minimum distance between S_9 and S_{17}, so a cluster is formed with these two stations, and the distance is almost 2.5. Then, the distance between station S_{26} and S_{32} is minimum, and a separate cluster is created with them distance is more or less 3, and a new cluster is formed to combine the earlier created cluster.

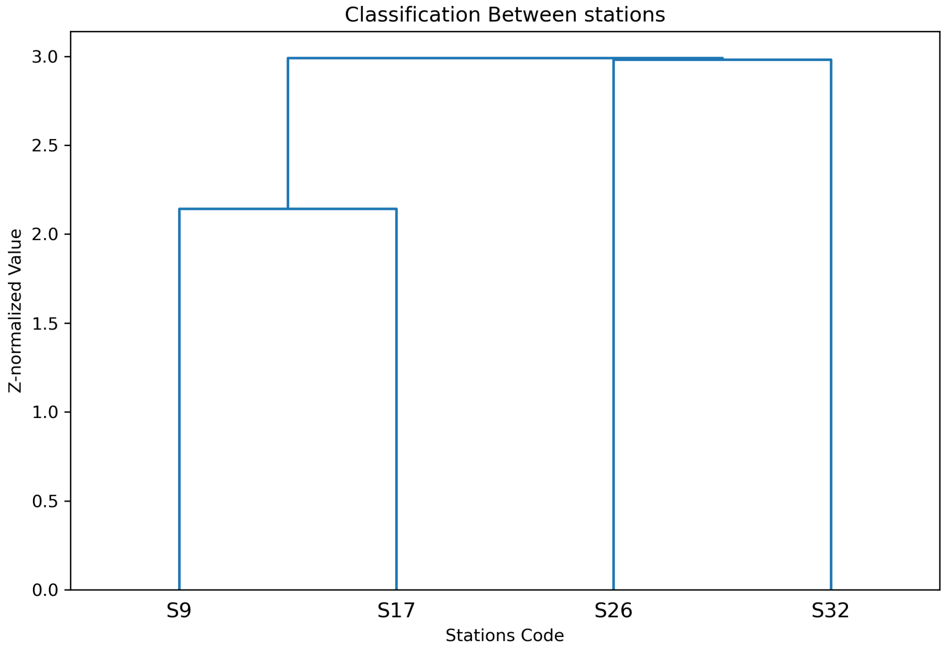


FIGURE 4.43: Cluster among stations

Now we discuss the pattern of length 720(i.e. monthly basis) of pollutant instance RH. In Fig.4.44 left-hand sided part is the repeated pattern of four different stations of RH pollutant and the right-hand side part shows classification between them. These patterns behave the same as the pattern of length 168 where vertical parts are the same but the upper and lower region is different and the level of pollution is almost the same every day of the month.

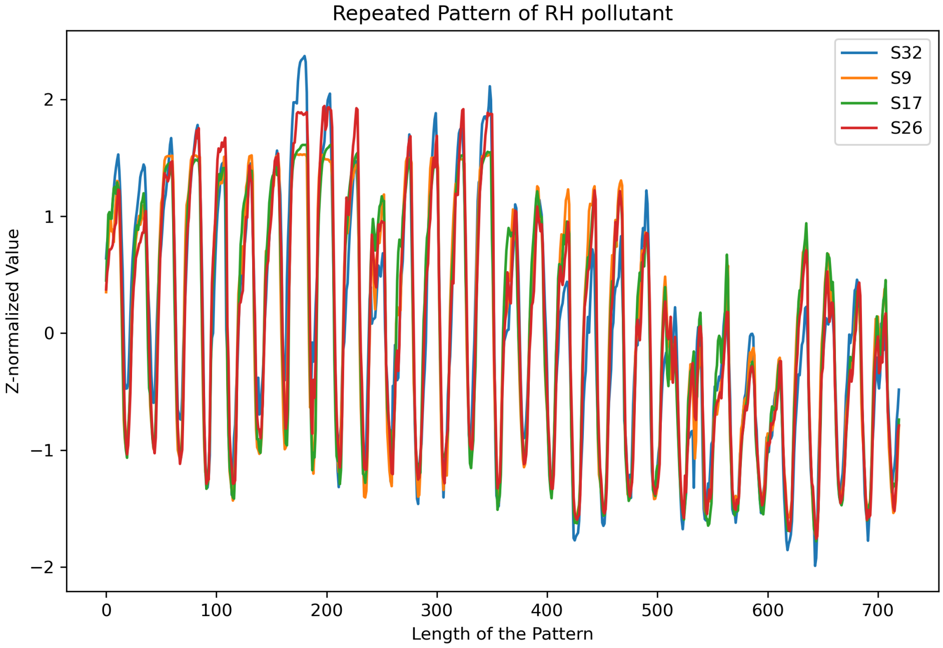


FIGURE 4.44: Repeated pattern of RH pollutant of length 720 of different stations

The distance between stations S_9 and S_{17} is minimum so we make a cluster with them that is shown in the above figure. Then the distance between S_{26} to the newly formed cluster is minimum hence S_{26} is added to new cluster . And finally, the remaining station S_{32} is added to this cluster.

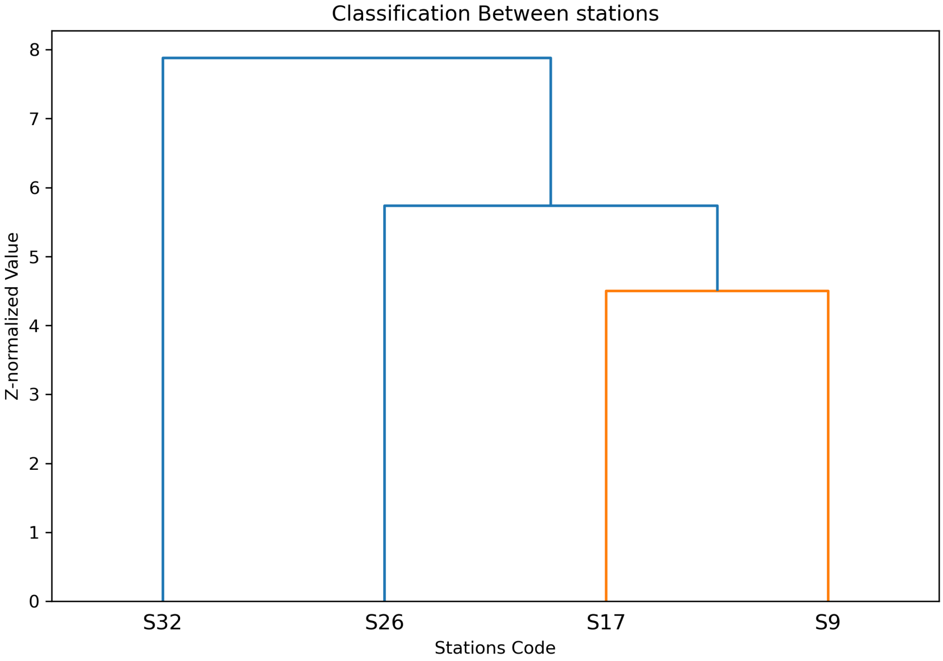


FIGURE 4.45: Cluster among stations

The way we have discussed the repeated pattern of station and classification between them of RH pollutant of several stations, the same way we can discuss other pollutants repeated pattern

Non-Normalized Results :

Previously we have shown that all the results of the repeated pattern are normalized with z-normalization method. In this section we discuss the patterns without normalization. All the patterns looks similar in Fig.4.46 but only difference is pollutant values varies significantly in different stations in comparison to pattern with normalization. The pollution level is high in station S_{17}(Nehru-Nagar-DPCC) and population density is low in station S_{32}(East-Arjun-Nagar-CPCB). All the pattern values of RH pollutant are actual values but in normalized patterns values are normalized. The repeated pattern is of length 24.

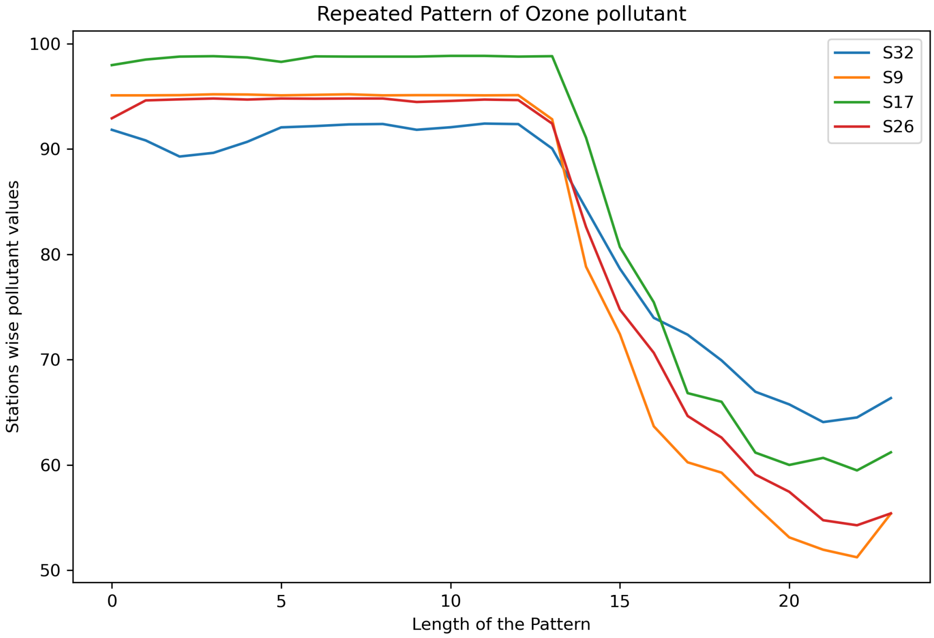


FIGURE 4.46: Repeated pattern of RH pollutant of length 24 of different stations

And the classification of stations is also similar to the normalized pattern and a cluster is formed with the same stations. The distance is calculated between every possible combinations of stations and distance is calculated all the values of pattern of one station pollutant values to other station. Here station S_{17}, S_9 and S_{26} are combined as they have the minimum distance and distance is almost 4.25. Then the distance from station S_{32} to newly formed cluster is almost 4.5 and it is the only one station remaining. So, it is added to the cluster and form a single cluster with these three stations.

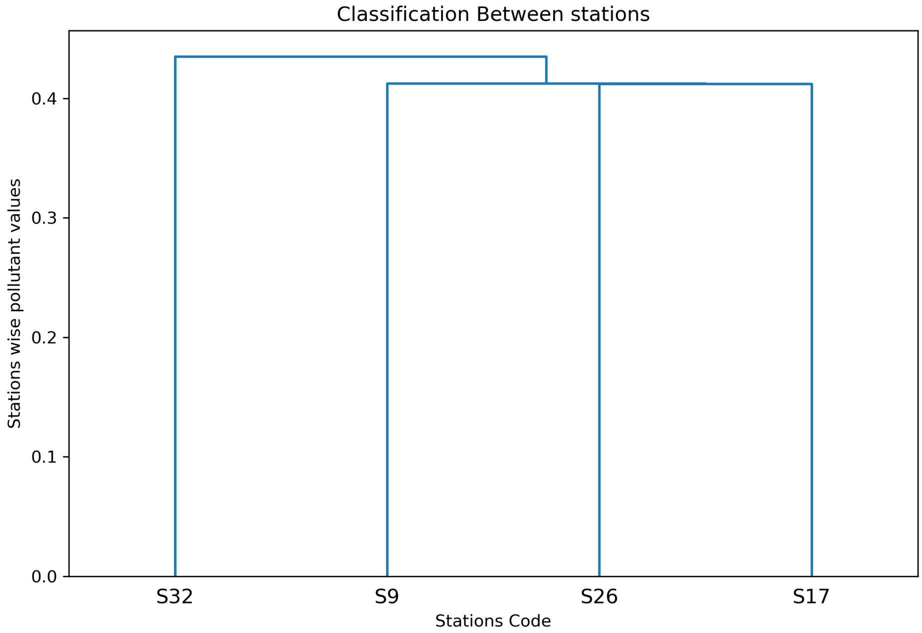


FIGURE 4.47: Cluster among stations

Now for repeated patterns of length 168 in Fig.4.49 only a few parts are similar but most of the parts are different. The behavior of RH pollutants in station Patparganj-DPCC and Ashok-Vihar-DPCC is similar but in Nehru-Nagar-DPCC and East-Arjun-Nagar-CPCB is dissimilar. The pollution level varies every hour of the day.

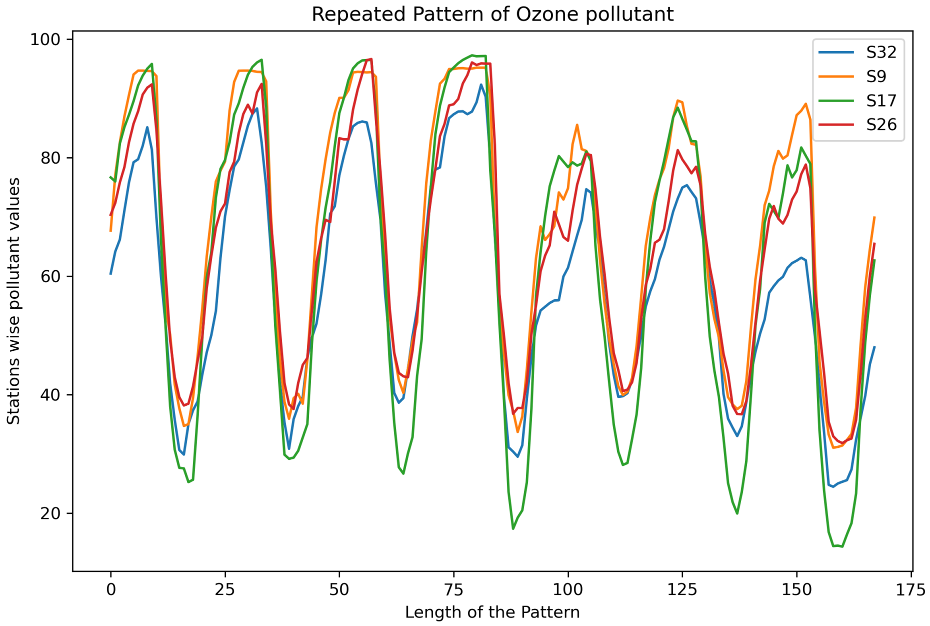


FIGURE 4.48: Repeated pattern of RH pollutant of length 168 of different stations

Classification of stations of RH is done in a similar way as in the previous section. Station Ashok-Vihar-DPCC and Nehru-Nagar-DPCC make a cluster because the distance between them is minimum and Patparganj-DPCC and East-Arjun-Nagar-CPCB create a cluster as they have minimum distance. Finally, combining these two clusters a single cluster is formed with four stations.

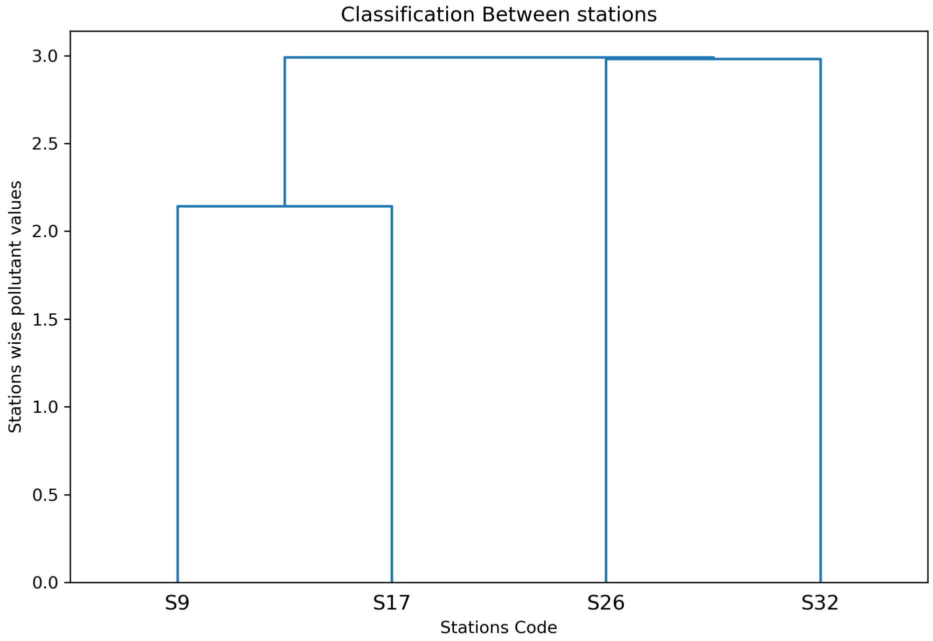


FIGURE 4.49: Cluster among stations

Now for pattern length 720, stations S_9 and S_{26} are almost similar but stations S_{17} and S_{32} are different. The pollution level of station S_{17} is low at some point in the day in comparison with other stations' pollution levels. And at the beginning of the month, the pollution level in all the stations is slightly higher than on the remaining days of the month.

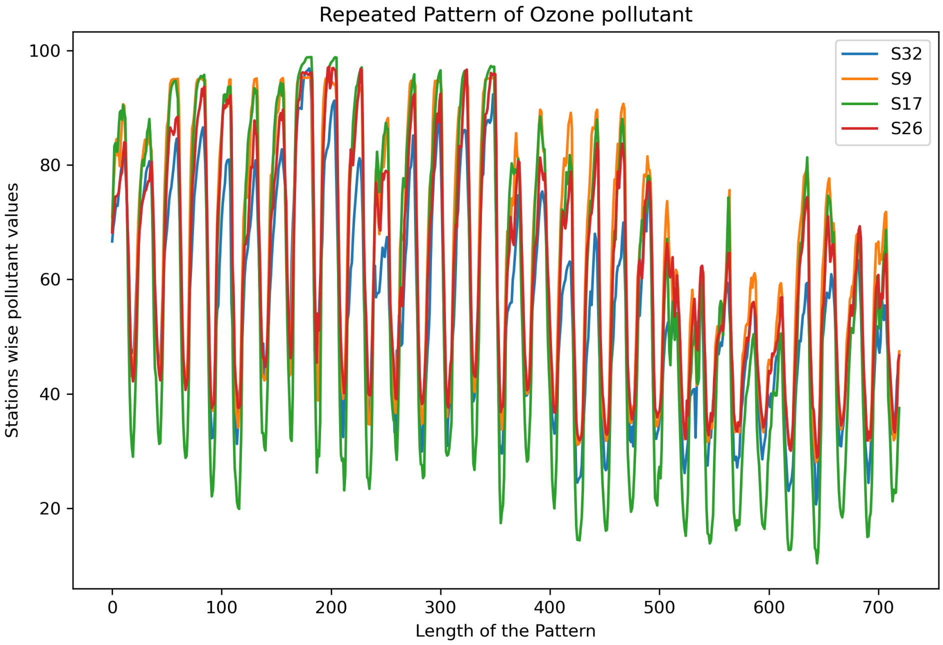


FIGURE 4.50: Repeated pattern of RH pollutant of length 720 of different stations

First stations S_9 and S_{17} combine to form a cluster then S_{26} is added to the newly formed cluster and finally, S_{32} is added on the distance between them.

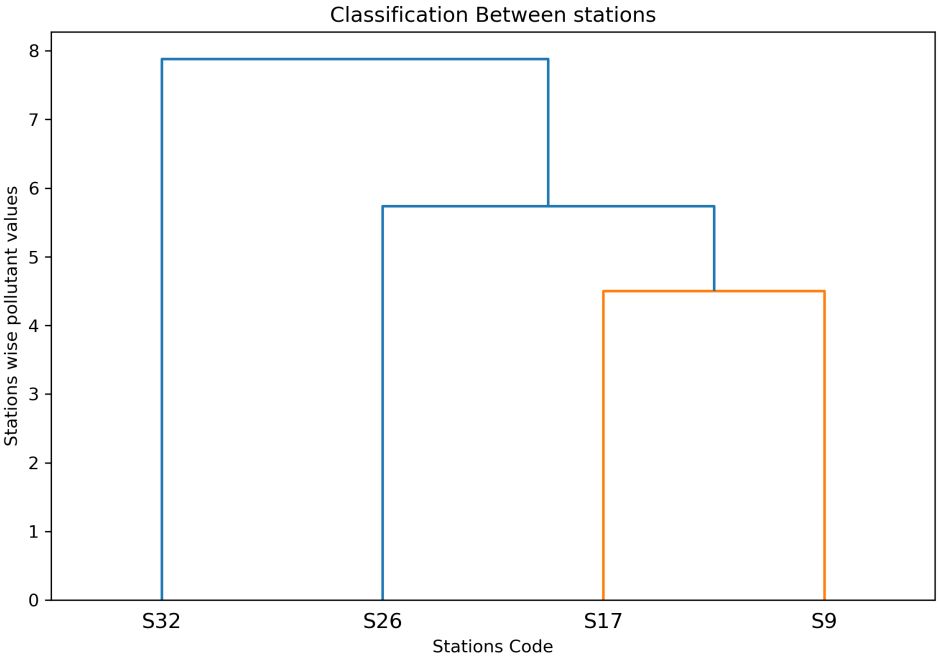


FIGURE 4.51: Cluster among stations

All other pollutants can be described similarly as discussed above for RH pollutants. And the distance between them varies from three to eight.

Chapter 5

Conclusion and Future Work

We have discussed the conclusion in this chapter, which will provide an overview of our work and its prospects. In the future, we can explore how we can utilize more segregation methods to separate stations for better visualization of repeating patterns, as well as how we can improve our model parameter for more accurate results.

5.1 Conclusion

In this thesis work, we discovered the repeated pattern with three different granularity that is daily, weekly and monthly basis using matrix profile. For finding repeated patterns we collect an air pollution dataset of different pollution monitoring stations in Delhi and we took one-year data on air pollution. In this work, we classify the stations in different ways such as randomly, geolocation based, landuse based, and population density based. Machine learning models are used to sort out stations. The complexity of our work depends on the length of the time series data and the size of the pattern. We have discussed previously repeated pattern of $PM_{2.5}$, Ozone, BP and RH with three different granularities of four different segregation methods of stations. The way we have elaborated on these pollutants same way we can elaborate on other pollutants.

5.2 Future Work

Ideas for the future include identifying repeating patterns from other time series dataset such as human activity, animal activity, App usages activity, temperature over a year, another pollution dataset, and so on, and the length of the patterns varies depending on the dataset. We will strive to use various categorization methodologies in the future to assist us classify stations differently. After identifying repeated patterns in time series data from various pollution stations, various approaches for grouping stations have been developed. We solely utilised one approach in this project, but we can use more methods in the future.

Bibliography

[1] C.-C. M. Yeh, Y. Zhu, L. Ulanova, N. Begum, Y. Ding, H. A. Dau, D. F. Silva, A. Mueen, and E. Keogh, "Matrix profile i: All pairs similarity joins for time series: A unifying view that includes motifs, discords and shapelets," in *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pp. 1317–1322, 2016.

[2] Y. Zhu, Z. Zimmerman, N. S. Senobari, C.-C. M. Yeh, G. Funning, A. Mueen, P. Brisk, and E. Keogh, "Matrix profile ii: Exploiting a novel algorithm and gpus to break the one hundred million barrier for time series motifs and joins," in *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pp. 739–748, 2016.

[3] C.-C. M. Yeh, H. Van Herle, and E. Keogh, "Matrix profile iii: The matrix profile allows visualization of salient subsequences in massive time series," in *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pp. 579–588, 2016.

[4] C.-C. M. Yeh, N. Kavantzas, and E. Keogh, "Matrix profile iv: Using weakly labeled time series to predict outcomes," *Proc. VLDB Endow.*, vol. 10, p. 1802–1812, aug 2017.

[5] A. Dau and E. Keogh, "Matrix profile v: A generic technique to incorporate domain knowledge into motif discovery," pp. 125–134, 08 2017.

[6] C.-C. M. Yeh, N. Kavantzas, and E. Keogh, "Matrix profile vi: Meaningful multidimensional motif discovery," pp. 565–574, 11 2017.

[7] Y. Zhu, M. Imamura, D. Nikovski, and E. Keogh, "Matrix profile vii: Time series chains: A new primitive for time series data mining (best student paper award)," in *2017 IEEE International Conference on Data Mining (ICDM)*, pp. 695–704, 2017.

[8] S. Gharghabi, Y. Ding, C.-C. M. Yeh, K. Kamgar, L. Ulanova, and E. Keogh, "Matrix profile viii: Domain agnostic online semantic segmentation at superhuman performance levels," in *2017 IEEE International Conference on Data Mining (ICDM)*, pp. 117–126, 2017.

[9] Y. Zhu, A. Mueen, and E. Keogh, "Admissible time series motif discovery with missing data," 2018.

[10] Y. Zhu, C.-C. M. Yeh, Z. Zimmerman, K. Kamgar, and E. Keogh, "Matrix profile xi: Scrimp++: Time series motif discovery at interactive speeds," in *2018 IEEE International Conference on Data Mining (ICDM)*, pp. 837–846, 2018.

[11] S. Imani, F. Madrid, W. Ding, S. Crouter, and E. Keogh, "Matrix profile xiii: Time series snippets: A new primitive for time series data mining," in *2018 IEEE International Conference on Big Knowledge (ICBK)*, pp. 382–389, 2018.

[12] S. Gharghabi, S. Imani, A. Bagnall, A. Darvishzadeh, and E. Keogh, "An ultra-fast time series distance measure to allow data mining in more complex real-world deployments," *Data Mining and Knowledge Discovery*, vol. 34, 07 2020.

[13] Z. Zimmerman, K. Kamgar, N. S. Senobari, B. Crites, G. Funning, P. Brisk, and E. Keogh, "Matrix profile xiv: Scaling time series motif discovery with gpus to break a quintillion pairwise comparisons a day and beyond," in *Proceedings of the ACM Symposium on Cloud Computing*, SoCC '19, (New York, NY, USA), p. 74–86, Association for Computing Machinery, 2019.

[14] K. Kamgar, S. Gharghabi, and E. Keogh, "Matrix profile xv: Exploiting time series consensus motifs to find structure in time series sets," in *2019 IEEE International Conference on Data Mining (ICDM)*, pp. 1156–1161, 2019.

[15] S. Imani and E. Keogh, "Matrix profile xix: Time series semantic motifs: A new primitive for finding higher-level structure in time series," in *2019 IEEE International Conference on Data Mining (ICDM)*, pp. 329–338, 2019.

[16] Z. Zimmerman, N. Shakibay Senobari, G. Funning, E. Papalexakis, S. Oymak, P. Brisk, and E. Keogh, "Matrix profile xviii: Time series mining in the face of fast moving streams using a learned approximate matrix profile," in *2019 IEEE International Conference on Data Mining (ICDM)*, pp. 936–945, 2019.

[17] F. Madrid, Q. Chesnais, K. E. Mauck, S. Singh, and E. J. Keogh, "Efficient and effective labeling of massive entomological datasets," 2019.

[18] F. Madrid, S. Imani, R. Mercer, Z. Zimmerman, N. Shakibay Senobari, and E. Keogh, "Matrix profile xx: Finding and visualizing time series motifs of all lengths using the matrix profile," pp. 175–182, 11 2019.

[19] T. Nakamura, M. Imamura, R. Mercer, and E. Keogh, "Merlin: Parameter-free discovery of arbitrary length anomalies in massive time series archives," in *2020 IEEE International Conference on Data Mining (ICDM)*, pp. 1190–1195, 2020.

[20] R. Mercer, S. Alaee, A. Abdoli, S. Singh, A. Murillo, and E. Keogh, "Matrix profile xxiii: Contrast profile: A novel time series primitive that allows real world classification," in *2021 IEEE International Conference on Data Mining (ICDM)*, pp. 1240–1245, 2021.