Time Series Motif and Pattern Mining To Analyse Air Pollution Data

A project report

submitted in partial fulfillment of the requirements for the Degree of

Master of Computer Application

of

Department of Computer Science and Engineering of

Jadavpur University

by

Roma Mandal

Registration No.: 149905 of 2019-2020

Examination Roll No.: MCA226044

Under the guidance of

Dr. Sarbani Roy

Professor

Department of Computer Science and Engineering

Jadavpur University, Kolkata-700032

India

2022

Faculty of Engineering And Technology

Jadavpur University

Certificate of Recommendation

This is to certify that the project report entitled “Time Series Motif and Pattern Mining To Analyse Air Pollution Data” has been carried out by **Roma Mandal** (University Registration No.: 149905 of 2019-2020, Examination Roll No.: MCA226044) under my guidance and supervision and be accepted in partial fulfillment of the requirement for the Degree of Master of Computer Application of the Department of Computer Science and Engineering, Jadavpur University. The research results presented in the project report have not been included in any other paper submitted for the award of any degree in any other University or Institute.

_____

Prof. Sarbani Roy (Project Supervisor)
Department of Computer Science and Engineering
Jadavpur University, Kolkata-32

Countersigned

_____

Prof. Anupam Sinha
Head, Department of Computer Science and Engineering,
Jadavpur University, Kolkata-32

_____

Prof. Chandan Mazumdar
Dean, Faculty of Engineering and technology,
Jadavpur University, Kolkata-32

Faculty of Engineering And Technology

Jadavpur University

Certificate of Approval

This is to certify that the project report entitled “Time Series Motif and Pattern Mining To Analyse Air Pollution Data” is a bona-fide record of work carried out by Roma Mandal in partial fulfillment of the requirements for the award of the degree of Master of Computer Application in the Department of Computer Science and Engineering, Jadavpur University during the period of January 2022 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 project report only for the purpose for which it has been submitted.

_____

Signature of Examiner 1

Date:

_____

Signature of the Examiner 2

Date:

Faculty of Engineering And Technology

Jadavpur University

Declaration of originality and compliance of academic ethics

I hereby declare that this project report entitled “Time Series Motif and Pattern Mining To Analyse Air Pollution Data” contains a literature survey and original research work by the undersigned candidate, as part of her degree of Master of Computer Application. All information has 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 referenced all materials and results that are not original to this work.

Name: Roma Mandal

Registration No.: 149905 of 2019-2020

Exam Roll No.: MCA226044

Project Title: Time Series Motif and Pattern Mining To Analyse Air Pollution Data

_____

Signature with date

Jadavpur University

Abstract

Faculty of Engineering and Technology, Jadavpur University
Computer Science and Engineering

Master of Computer Application

Time Series Motif and Pattern mining To Analyse Air Pollution Data

by Roma Mandal

Air pollution is a global issue in our contemporary civilization, and it is the primary cause of impending global warming. The main cause of air pollution is the growing economic condition and the high number of chemical industries in urban and suburban areas. Although a rise in the number of automobiles and changing lifestyle are also major contributors to air pollution. Respiratory disease, heart disease, irritation of the eyes, breathlessness, etc. are some of the dangerous effects of air pollution. Because of the significant growth of air pollutants, government agencies have set up pollution monitoring stations throughout the country using sensors, and those sensors generate a large amount of data and that data may be visualized as time series data for pollution pattern identification to better understand variability in pollution levels. In the above-mentioned context, the focus of this project is to collect real world air pollution data from government setup pollution monitoring stations of Delhi and find temporal motif (i.e., pattern) in the pollutants' time series. Specifically, our study focuses on identifying consensus motif for various subsequence lengths. An existing motif discovery algorithm such as ostinato, MASS, etc. are employed to achieve the project goal.

Keywords: *Time series, Matrix Profile, Motif, Pollutants, Air Pollution, Consensus Motif*

Acknowledgements

The writing of the project report as well as the related work has been a long journey with input from many individuals, right from the first day till the development of the final project. I would like to express my deepest gratitude to my supervisor, Prof. Sarbani Roy, Professor, department of computer science and engineering, Jadavpur University for giving me the opportunity to do research and providing invaluable guidance throughout this work. Her dynamism, vision, sincerity and motivation have deeply inspired me. She has taught me the methodology to carry out the work and to present the works as clearly as possible. It was a great privilege and honor to work and study under her guidance. I would like to express my sincere, heartfelt gratitude to Asif Iqbal Middya, research scholar, department of computer science and engineering, Jadavpur University for his patience, guidance, suggestions and moral support in times of need. I am also particularly thankful to Prof. Anupam Sinha, Head of the department of computer science and engineering, Jadavpur University for allowing us to carry out research in the department.

I would also like to thank all the faculty members of the department of computer science and engineering of Jadavpur University for their continuous support. This project report would not have been completed without the inspiration and support of my family and friends, and a number of wonderful individuals including my batchmates of Master of Computer Application in Jadavpur University — my thanks and appreciation to all of them for being part of this journey and making this project report possible.

_____

Roma Mandal

Examination Roll No.: MCA226044

Registration No.: 149905 of 2019-2020

Class Roll No.: 001910503045

CONTENTS

List of Tables

List of Figures

List of Symbols

Symbol	Description
T	Time series data
m	Subsequence length
$T^{a}_{i,m}$	It is the subsequence of T with starting index i, and m as the length of the index
$T^{b}_{j,m}$	It is the subsequence of T with starting index j, and m as the length of the index
$Q_{i,j}$	Dot product between $T^{a}_{i,m}$ and $T^{b}_{j,m}$
σ	Standard deviation
μ	Mean

In Dedication to my family for supporting me all the way!

Chapter 1

Introduction

In today's world data play an important role in day-to-day life and if those data are collected properly and used for time series data analysis then it can give various results that may be beneficial in some or the other way. In recent times, pollution has become a source of concern due to continuous industrial progress and a huge increase in the urban population. Carbon monoxide, nitrogen dioxide, ozone, and fine particulate matter (PM10, PM2.5) are more prevalent in densely populated cities. And as a result of continuously increasing pollution, pollution monitoring has become a crucial element of modern living. Continuous monitoring is necessary and because of this many government agencies have set up pollution monitoring stations throughout the country using sensors, and those sensors generate a large amount of data and that data may be utilized as time series data for pollution pattern identification for better understanding pollution variability.

Time series data are set of real-valued observations taken on equal time intervals; the time intervals can be chosen as per the application domain. There are a lot of application domains that use time series data for finding solutions to different kinds of problems. Some of the domains that are proposed in literatures are seismology, motion capture animal behavior detection, music similarities [1, 3, 4, 5, 10], etc. Time series motifs are repeated sub-sequences or segments embedded in a long time series data. Motifs are nothing but useful patterns used in retrieving information from time series data by using different data mining techniques. Motif discovery recently has started receiving significant attention beyond the data mining community. In recent decades, time series motif analysis has gained much popularity as it offers solutions to most of the time series problems including semantic segmentation, anomaly detection, etc.

The occurrence of motifs in time series data is not only due to chance rather they usually contain underlying information about the system and provide valuable insights about the problem being investigated. Time series motif discovery can be done on different types of data sets such as one-dimensional time series data or multidimensional time series data etc.

For finding motifs in time series data, Matrix Profile became a generic data tool to solve a host of time series data mining problems. Matrix Profile is a data structure that stores the nearest neighbor information for every sub-sequence of particular time series. The distance measures for the matrix profile can be calculated using different techniques. Some of the techniques proposed in literatures are Euclidean distance, MPdist, DTW (Dynamic time warping) [36, 2, 14], etc. The main aim of motif discovery is to find unknown patterns in a time series data without any prior information about the type of data being used. Below Fig. 1.1. shows an example of time series motif discovery [https://bit.ly/3O9Lfbl]. Fig. 1.1(a) shows the time series motif in red color. A, B, and C are the repeating sub-sequences or patterns which are almost similar to each other and we can also see the location of each repeating sub-sequences. Fig. 1.1(b) shows the closeup look of these repeating patterns and, we can see that they are almost similar to each other with a very minute difference.

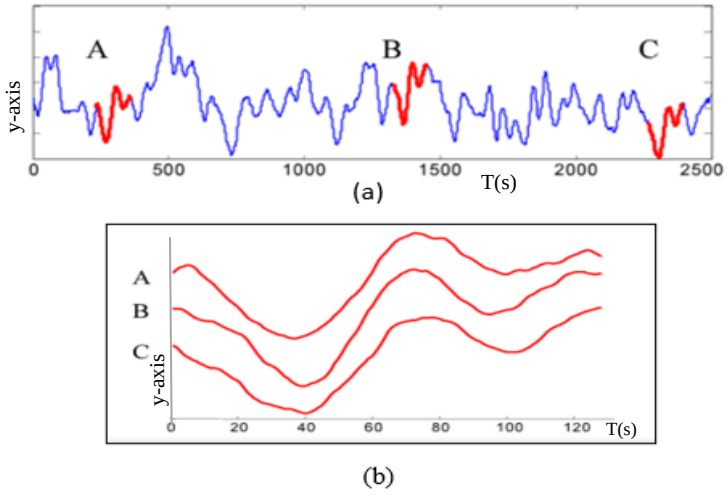


Fig. 1.1 An example of time series motif (a) time series motif containing three sub-sequences of similar type at three different locations (b) closeup look of the three sub-sequences and how similar they are to each other [https://bit.ly/3O9Lfbl].

Time series motifs can be of different lengths and types. Some of the common types of time series motifs are one-dimensional time series motifs, multi-dimensional time series motifs, semantic motifs, consensus motifs, etc. [9, 8,14, 21, 25, 26]. In a one-dimension or single dimension time series motif, the discovery of conserved patterns is done for a single time series data, the repeated patterns are searched in the same time series but in the case of a multi-dimension time series motif the discovery for finding conserved patterns is done for a set of time series data, similar patterns are searched in every time series present in that set. Semantic motifs are a different kind of motif, they are not like the usual time series motif. Semantic motifs are repeated patterns in a time series data, suppose there is a similar conserved pattern consisting of two parts in a single time series data, each of length n, separated by some interval, that type of conserved semantic pattern is known as a semantic motif. A "consensus motif" is a time series motif for discovering conserved patterns that is common to all of the time series in a set. Matrix profile can be used to detect conserved patterns both within a single time series data by using self-join and across two or more time series using AB-join. These conserved patterns are referred to as time series consensus motifs. Kamgar et al. [25], proposed a novel algorithm to find a time series consensus motif named ostinato. In this work, we will explore more about the consensus motif in detail.

Since the topic is very vast, it is impossible to cover every aspect of it. Therefore, this work will focus on time series motif discovery and pattern mining to analyze air pollution data. We have tried to find consensus motifs for the time series pollution dataset of Delhi. The dataset contains pollution levels of 21 different stations and 8 pollutants (CO (carbon monoxide), NO2 (nitrogen dioxide), O3 (ozone), PM2.5, PM10, Sr (strontium) WS (wind speed), RH (Relative humidity)). The main objectives of this study can be categorized in the following two aspects: (i) Single station multiple pollutants for finding out the dependency of every pollutant and how they vary relative to each other across the stations (ii) single pollutant multiple stations for finding how the pollutant vary and have what type on the impact on the cities.

The rest of the project report is organized as follows: In Chapter 2, we present background and related works. Chapter 3 describes motif discovery techniques. In Chapter 4, we describe our experiment of finding similar patterns in air pollution time series data. Chapter 5 gives a conclusion and future work.

Chapter 2

Background and Related Work

Motif discovery for time series was first proposed in 2003 [37] but it foreshadows the classic study of motifs by computing all-pairs similarity for time series [38], and it has created a flurry of research effort since then. The use of motifs to solve issues in disciplines is as diverse as hemodynamics, animal behavior [17], music processing [1, 4], neurology [7, 16], and entomology [7] has been some of the key trends. Extensions and generalizations of the original work particularly attempt to increase scalability, have been another key research topic.

There is little research on repeating pattern discovery in real-valued series that we are aware of. Techniques like Matrix Profile [3] try to find the best-matched subsequence pair rather than the most common one when it comes to motif finding. If one has access to the matrix profile, it has been shown in most recent literatures [2, 3] that one can easily compute all top-k motifs (for any k) range motifs (for arbitrary ranges), and several other important time series primitives [10].

2.1 Definitions and Notation

Definition 1 (*Time Series*) A time series $T \in \mathbb{R}$ is a contagious sequence of real-valued numbers $t \in \mathbb{R} : T = [t_1, t_2, t_3. \ldots, t_n]$ where n is the length of series T.

We will discuss about the local, not global properties of time series in this work, local region of time series is known a subsequence.

Definition 2 (*Subsequence*) Given a time series T of length n, a subsequence $T_{i,m}$ of T is a continuous subset of length m < n of contiguous positions from T, that is, $T_{i,m} = \{t_i, t_{i+1}, t_{i+2}, \ldots, t_{i+m-1}\}$ where $1 < i < n - m + 1$.

The particular local property we are interested in in this work is finding time series motifs. With the help of distance profile, we can take a subsequence and compute its distance to all subsequences in the same time series.

Definition 3 (*Time Series Motif*) The most identical subsequence pair of a time series is called a time series motif. $T_{a,m}$ and $T_{b,m}$ is the time series motif pair iff $dist(T_{a,m}, T_{b,m}) \leq dist(T_{i,m}, T_{j,m}) \; \forall \; i,j \in [1, 2, \dots, n - m + 1]$ where n is the length of time series T, m is the length of the subsequence, $a \neq b$ and $i \neq j$, and $dist$ is a function that computes the z-normalized Euclidean distance between the input subsequences [2, 3, 9, 37].

In an ordered array called a distance profile, we store the distance between a subsequence of a time series and all other subsequences from the same time series.

Definition 4 (*Distance Profile*) Distance profile (DP for short) of a subsequence $T^a_{i,m}$ is a vector that stores the z-normalized Euclidean distances between $T^a_{i,m}$ and each subsequence of length m in a time series $T^b_{j,m}$. The z-normalized Euclidean distance can be calculated as:

$$\text{dist}\,(T^a_{i,m}, T^b_{j,m}) = \sqrt{2m\left(1 - \frac{Q_{i,j} - m\mu_i\mu_j}{m\sigma_i\sigma_j}\right)}$$

Here, m is the subsequence length, $Q_{i,j}$ is the dot product between $T^a_{i,m}$ and $T^b_{j,m}$, μ_i and μ_j are the mean of $T^a_{i,m}$ and $T^b_{j,m}$, σ_i and σ_j are the standard deviation of $T^a_{i,m}$ and $T^b_{j,m}$. Formally, $D_i = [d_{i,1}, d_{i,2}, \dots, d_{i,n-m+1}]$, where $d_{i,j}$ $(1 \leq i, j \leq n\text{-}m+1)$ is the distance between $T_{i,m}$ and $T_{j,m}$. The distance profile can be computed efficiently by using a convolution-based method such as MASS [https://bit.ly/3aJ8c5K].

Recent work shows that the matrix profile [3][4], a data structure that contains nearest neighbour information for every subsequence in a time series, it solves a variety of problems in time series data mining, including motif discovery. The matrix profile is the most efficient means of precisely finding time series patterns.

Definition 5 (*Matrix Profile*) A matrix profile MP [3] of time series T is a vector of the Euclidean distances between each subsequence $T_{i,m}$ and its nearest neighbor in time series T. Formally, MP = [min(D_1), min(D_2),…, min(D_{n-m+1})], where D_i ($1 \leq i \leq n-m+1$) is the distance profile Di of time series T.

We call this vector matrix profile because one way to compute it would be to compute the full distance matrix of all pairs of subsequences in time series T, and then evaluate the minimum value of each column. Below fig. 2.1. [3] shows the relationship between the distance profile DP and matrix profile MP.

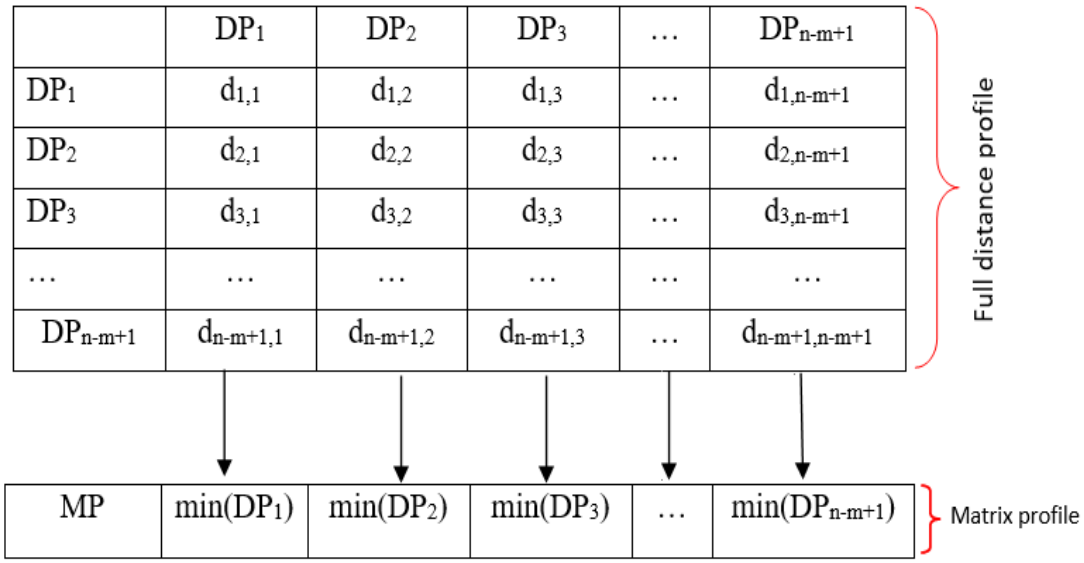


Fig. 2.1. Relationship between the distance profile, the matrix profile and the full distance matrix [3].

In the above picture we can see that the Distance matrix is symmetric. The Euclidean distance between subsequence $T_{i,m}$ and its nearest neighbour in time series T is shown by the i^{th} element in the matric profile MP. The minimum (off diagonal elements should only be considered since the diagonal elements are 0 as they are trivial matches) values of each column of the distance matrix are recorded in the matrix profile and the location of the minimum value within each column is saved in the matrix profile index.

Definition 6 (*Trivial Match*) Given a subsequence S_i and its nearest neighbor S_j from a same time series S, a trivial match occurs when i = j. In distance profile its value is 0 or

approximately equal to 0. Suppose we are searching for a sub-sequence of length 10 in a time series and let we find its best match at location 30 but the second best match will probably be at location 29 or 31but those are trivial matches since they have minor to no difference. To avoid such unimportant matches, an exclusion zone is created around the best match, followed by the search for the second best match.. We avoid such matches by ignoring an "exclusion zone" of length m/4 before and after the location of the query [3]. .Below Fig. 2.2. [https://bit.ly/3HCcyZ2] Shows a time series T in which a query Q is being searched, the 1st location where the query Q is located is around 50 then according to the above definition locations around 50 are considered to be an exclusion zone and is shown by a purple patch and at location 90 2nd best match is found which is a non-trivial one.

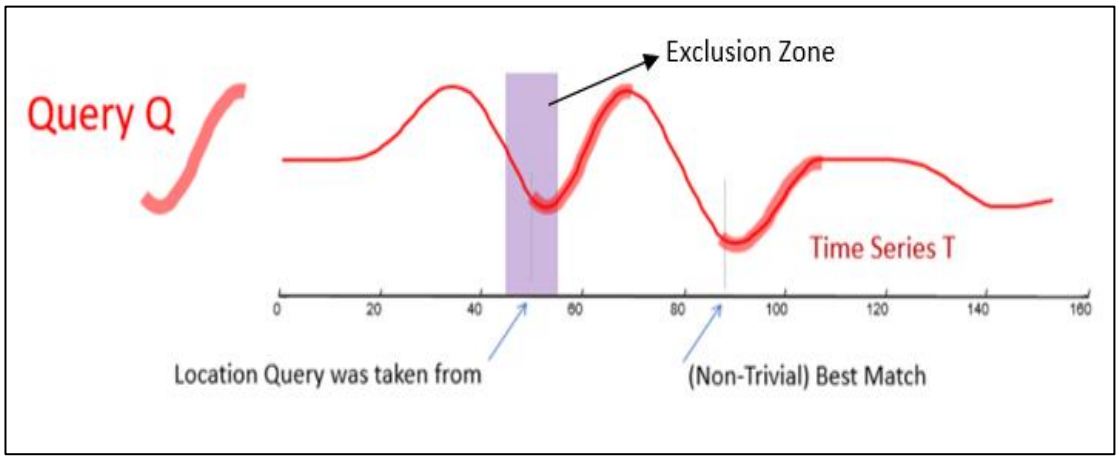


Fig. 2.2. A graphical representation of exclusion zone in a time series data. [https://bit.ly/3HCcyZ2]

Definition 7 (*Matrix Profile Index*) A matrix profile index I of time series T is a vector of integers: $I= [I_1, I_2, \dots I_{n+m-1}]$, where $I_i=j$ if $d_{i,j} = \min(D_i)$, where $D_i = [d_{i,1}, d_{i,2},\dots, d_{i,n-m+1}]$.

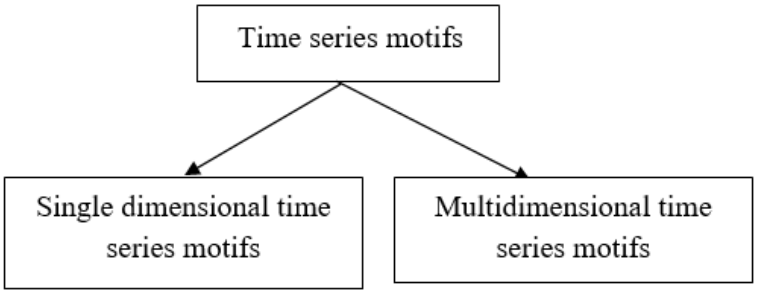
We can quickly find the nearest neighbour of query $T_{i,m}$ by accessing the ith element in the matrix profile index by storing the adjacent information in this manner.

We have summarised the important takeaways from the preceding section because it was a bit dense. To annotate a time series A with the distance to and location of all its

subsequences' nearest neighbours in itself or another time series B, we can use time series for basically finding two things: the matrix profile and the matrix profile index. These two data items explicitly provide the answers to the motif discovery tasks in time series data mining [15, 18]. Furthermore, as we'll see later, the matrix profile and matrix profile index as primitives can be used to perform a variety of additional types of analytics.

2.2 Types of time series motif

Basically, time series motifs are divided into two parts:



- **Single dimensional time series motifs [9, 18, 21]:** Conserved patterns found within 1-dimensional time series T is known as single dimensional time series motif. These can be computed with the help of matrix profile.
- **Multidimensional time series motifs [9, 14]:** A multidimensional time series $\boldsymbol{T} \in \mathbb{R}^{dxn}$ is a set of co-evolving time series $T^{(i)} \in \mathbb{R}^{n}$: $\boldsymbol{T} = [T^{(1)}, T^{(2)}, \ldots, T^{(n)}]^{T}$ where d is the dimensionality of $\boldsymbol{T}$ and n is the length of the time series $\boldsymbol{T}$.

Some more types of time series motifs are:

- **Missing value time series motifs [30]**: A missing value time series $\bar{T}$is a sequence of missing value time series that are either real-valued numbers or NaNs, $\overline{t1}$: $T = \overline{t1}, \overline{t2}, \ldots, \bar{t}$n, where n is the length of the time series $\bar{T}$. We assume T is the real time series of $\bar{T}$, before the missing values were produced by some process, let us assume if the sensors had been working properly, we would have gotten T instead of $\bar{T}$.

Works related to missing values time series motifs has been discussed in the paper [30]. In this paper some methods have also been introduced to find such missing value so that they can overcome the false negative outcomes.

- **Semantic motifs [26]:** Semantic motifs are repeated patterns in a time series data. The starting location of the semantic motif is the same as the starting position of its prefix. A semantic motif has three parts: prefix, don't-care, and suffix. The prefixes and suffixes are almost similar to each other. The don't-care regions are random and are ignored. Moreover, we allow don't-care regions to be of different lengths ranging from 0 to r, where r is a user-defined value. Formally, a pattern meeting all these requirements is a time series semantic motif [26]. Imani et al. [26] proposed the definition for semantic motif pair as: Ta,m and Tb,m is said to be the semantic motif pair iff (a, b)~argmini, j {dist(Ti,m, Tj,m) + dist(Tk,m , Tℓ,m)} where $1 \leq i \leq n - 2(r + m) + 1$, $i + r \leq k \leq n - (r+ 2m) + 1$, $k + m \leq j \leq n - (r + m) + 1$, $j + m + r \leq l \leq n - m + 1$ $\forall$ i, j, k, l $\in$ [1,2,… , n − m + 1]. Where n is the length of the time series, m is the length of the semantic motif, $i \neq j$, $k \neq \ell$, and the dist function computes the z-normalized Euclidean distance between the input subsequences. r is the maximum length of the don't-care region. The first term i.e., dist(Ti,m, Tj,m) is the distance between prefixes, and the second term i.e., dist(Tk,m , Tℓ,m) is the distance between suffixes concerning the ordered triple (prefix, don't-care, suffix).

- **Consensus motifs [25]:** It is the subsequence from one of k time series T 1...T k that has the smallest radius of any subsequence occurring in any of the time series T 1...T k. The radius r of a subsequence Tij,m of time series T i with respect to a sequence of time series T 1…Tk is the maximum distance between Tij,m and its nearest neighbor in each of T 1…Tk . Each subsequence in each of the k time series is encircled by an m-ball with a minimum distance that contains at least one subsequence from each of the remaining (k-1) time series if all subsequences in the set of k time series are imagined as points in m-dimensional space. Then this distance is known as the radius. Fig. 2.3, shows the visual representation of radius [25]. It shows the subsequence of three time series A, B, and C that exists as a point in m-dimensional space.

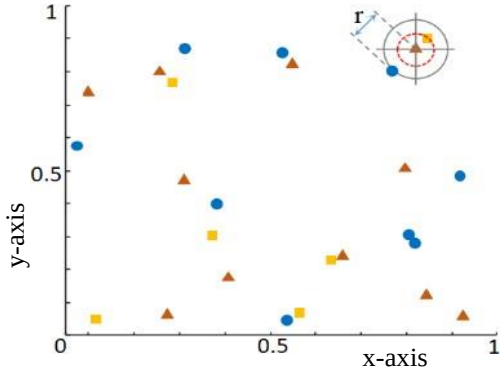


Fig. 2.3. Shows a visual representation of radius r from the subsequences of three time series, A, B, and C exist as points in a m-dimensional space [25].

In order to find a consensus motif, we must determine which subsequence from the k time series produces the smallest radius in order to obtain a consensus motif asdefined. Fig. 2.4 shows an example of consensus motif discovery [25]. We start by concatenating all 3-time series into a single time series T, with null markers in between to indicate where one time series ends and another begins. The length of this long time series may be denoted by the letter N. This long time series can be used to generate a distance matrix D, which contains the pairwise distance between every z-normalized subsequences of length m in T. By sweeping across all columns of the pairwise distance matrix (visualized by the red line) and finding the minimum value ineach of the 3 regions. The radius of the relevant subsequence is determined to be the largest of these 3 column-wise values, while the radius of the consensus motif is foundto be the smallest of all such r

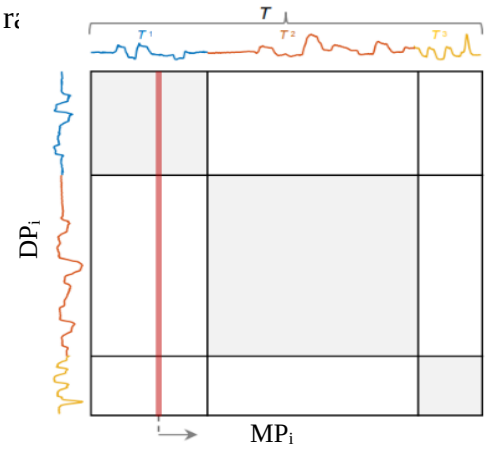


Fig.2.4. An example of the consensus motif discovery under a time series T made from concatenating three time series T1, T2, T3. As the red line sliding across all the columns of D, we calculate the minimum distances of three regions. The maximum value among the 3 column wise values is the radius of the corresponding subsequence, and the smallest among them will be the radius of the consensus motif [25].

2.3 Acronyms Used in This Domain

This Section contains the acronyms used in this domain along with their meaning and description and references.

Table 2.1 List of acronyms

Acronym	Meaning	Description	Reference
SiMPle	**S**imilarity **M**atrix **P**rofiLE	It is a vector which stores the Euclidean distance of A to its nearest neighbor in B and the position of its nearest neighbor in B	[1]
STAMP	**S**calable **T**ime **S**eries **A**nytime **M**atrix **P**rofile	STAMP is an algorithm which computes time series subsequence joins with an efficient anytime algorithm	[2], [3], [20]
VALMOD	**V**ariable **L**ength **M**otif **D**iscovery	It is an exact and scalable motif discovery algorithm that efficiently finds all motifs in a given range of lengths.	[12]
LAMP	**L**earned **A**pproximate **M**atrix **P**rofile	It is a model which enables constant time approximation of the MP (matrix profile) value given a newly arriving time series.	[27]
MIR	**M**usic **I**nformation **R**etrieval	Music information retrieval is the interdisciplinary science of retrieving some information from music for different purposes such as originality and authenticity of the music etc.	[1]
DTW	**D**ynamic **T**ime **W**arping	It is an algorithm used to dynamically compare time series data when the time series indices between comparison data points do not sync up perfectly.	[36]
MDS	**M**ultidimensional **S**caling	Multidimensional Scaling is a family of statistical methods that focus on creating mappings of items based on distance.	[4]
EPG	**E**lectrical **P**enetration **G**raph	EPG is a technology widely used by entomologists to study how different kinds of insects feed on plants.	[7]
MASS	**M**ueen's **A**lgorithm for **S**imilarity **S**earch	MASS is an algorithm to create Distance Profile of a query to a long time series.	[18]

2.4 Related work

In this section, we have summarized existing literature on motif discovery. Silva et al. [1], proposed a novel approach to assess similarities between audio recordings for MIR (Music Information Retrieval) tasks by using a self-similarity-matrix and named the representation as SiMPle (Similarity Matrix Profile). Yeh et al. [4] have also worked in similar areas. A new performance of a previously recorded material is referred to as a "cover song." A cover song, for example, can refer to a live performance, a remix, or a musical interpretation in a different genre. Most of the works are for automatic cover identification such as Copyright management, collection organizing, etc.

Yeh et al. [2, 4], Zhu et al. [13], Gharghabi et al. [14], Imani et al. [15], Linardi et al. [29] all have worked on finding similarities in electric power consumption. Most of them have used the time series data of power consumption for freezer and how they vary during a year for each season change. By knowing the similarities of power consumption in each season better power distribution and maintenance techniques can be applied to the areas to overcome electricity problems.

Yeh et al. [2], Dau et al. [8], Zimmerman et al. [27], Zhu et al. [30] discussed about repeated pattern (i.e., motif) discovery is a fundamental tool in seismology, which allows discovery of foreshocks, triggered earthquakes, swarms, volcanic activity, and induced seismicity. Conserved patterns help to detect earthquakes as, at that time there will be a spike in the amplitude and by continuous monitoring early precautions may be taken. However, it can fail to detect smaller or more distant earthquakes, whose average waveform amplitudes are close to the noise floor [27]. So more sensitive detection method should be used but that will require a lot of cost for installing system for sensitive detection. To overcome this situation Zimmerman et al. [27] proposed a method LAMP (Learned Approximate Matrix Profile) which has a potential solution for sensitive, rapid and inexpensive real-time seismic event detection.

Imani et al. 2019 [26] proposed an exact algorithm for finding semantic motifs named Semantic Motif Finder. In this they have used a human speech/ recorded audio of the poem “The Raven” by Edgar Allen Poe and converted that to a time series data and have

tried to find semantic motifs present in that dataset. However, they deliberately used a poor quality audio recording to demonstrate that Semantic Motif Finder algorithm works with challenging data. They have also tried to find out about animal behavior by using semantic motifs, for this they have used a seal behavior dataset that contains motion recordings of seventy-two seals, belonging to four species. This dataset contains data from a wearable accelerometer mounted on the seal's back. They have particularly chosen this dataset as the data is complex and challenging.

Zhu et al. [3,5,10,12] , Imani et al. [26] have tried to analyse behaviours of penguin and seal by recording their movements. This will help us to study about their behaviour by finding similar patterns. Yeh et al. [4] , Linardi et al. [29], Nakamura et al. [32] all have worked in finding similarities in ECG datasets , allowing us to distinguish between different ECG reports they have also tried to find out human activity behaviours by recording their breathing, activities throughout the day or by measuring their pedestrian count etc. Finding motifs in all these datasets helps to know about different kind of similarities and dissimilarities found.

Neuroscience and hemodynamics are another major research domain in time series motif discovery. Zhu et al. [7, 10, 16], Gharghabi et al. [11] have searched on Syncope, the loss of consciousness caused by a fall in blood pressure. They have also experimented with the Parkinson Disease (PD) dataset they do not claim any medical significance in this rather they simply want to show that the Matrix Profile and a few lines of code can allow you to quickly test ideas that may be fruitful.

Entomology has been a popular domain in time series motif discovery. Yeh et al. [7] , they have experimented on insects that feed by ingesting plant fluids cause devastating damage to agriculture worldwide, primarily by transmitting pathogens of plants. The behavior of the Asian citrus psyllid has been used in their experiments. The three-minute and three-hour samples demonstrate how the behavior is structure-suggestive but intrusive and complex. However, they do not want to make any claims of entomological relevance. But the output from such monitoring is often evaluated manually, which is a

time-consuming and laborious operation. So, motif discovery can be a used for finding such results.

Table2.2 Summary of related works in existing literature

Topic	Description	References
Music/Audio Similarity	Includes works for finding similarities in audio and music recordings, copyright/ originality identification etc.	[1], [4], [6], [9]
Speech recognition Analysis	Includes survey for finding semantic motifs present in human audio.	[26]
Electric power demand	Includes surveys related to knowing the power consumption pattern.	[2], [4], [9], [13], [14], [15], [29]
Seismology	Includes detection and monitoring of seismic waveforms to detect earthquakes, aftershocks etc.	[3], [4], [5], [8], [27], [30]
Animal behaviour	Surveys that contain motion recording of penguin, seal to know about their behaviour.	[3], [5], [10], [17], [26], [34]
Heart beat analysis	Survey related to monitoring ECG reports	[4], [29], [32], [34]
Human activity analysis	Finding similarity in behaviours by recoding pedestrian counts, tracking activity for all day, etc.	[4], [9], [15], [27], [31], [35]
Neuroscience	Experiments related to Parkinson's disease	[7],[16]
Hemodynamic	Experiments related to fall in blood pressure levels, loss of consciousness etc.	[7], [10], [11]
Entomology	Experiments related to insects EPG for finding out their feeding trend on plants.	[7], [24], [27], [34]

2.5 Timeline

This section presents the temporal evolution of the mile-stones works that contributed to shaping the Time series motif discovery in real-valued datasets. Fig. 3.1 shows graphically the time of appearance and groups the research works into nine categories:

- Music/Audio Similarities and Speech Recognition Analysis
- Electrical Power Demand
- Seismology
- Animal Behavior Analysis
- Heart Beat Analysis
- Human Activity Analysis
- Neuroscience
- Hemodynamics
- Entomology

To uncover the time-evolution, we have created this timeline. We can see that works related to seismology are being carried out since 2016 and has a lot of potential in terms of finding foreshocks, aftershocks, triggered earthquakes, and volcanic activity, all of which can be reduced to finding these recurrent patterns. Music /audio similarity analysis was first done in 2016 by Silva et al. [1] and since then many works have been done in this domain till 2019. But in last few years works related to this domain is not seen. Electrical power consumption has been another key domain for research in time series motif analysis. Similarly works related to hemodynamics, entomology, neuroscience all these are gaining interest in recent times. Although motif identification in time series data has been performed extensively in existing literature it is not well investigated for air pollution time series. Hence this project focuses on exploring motif identification for various air pollutants' time series data.

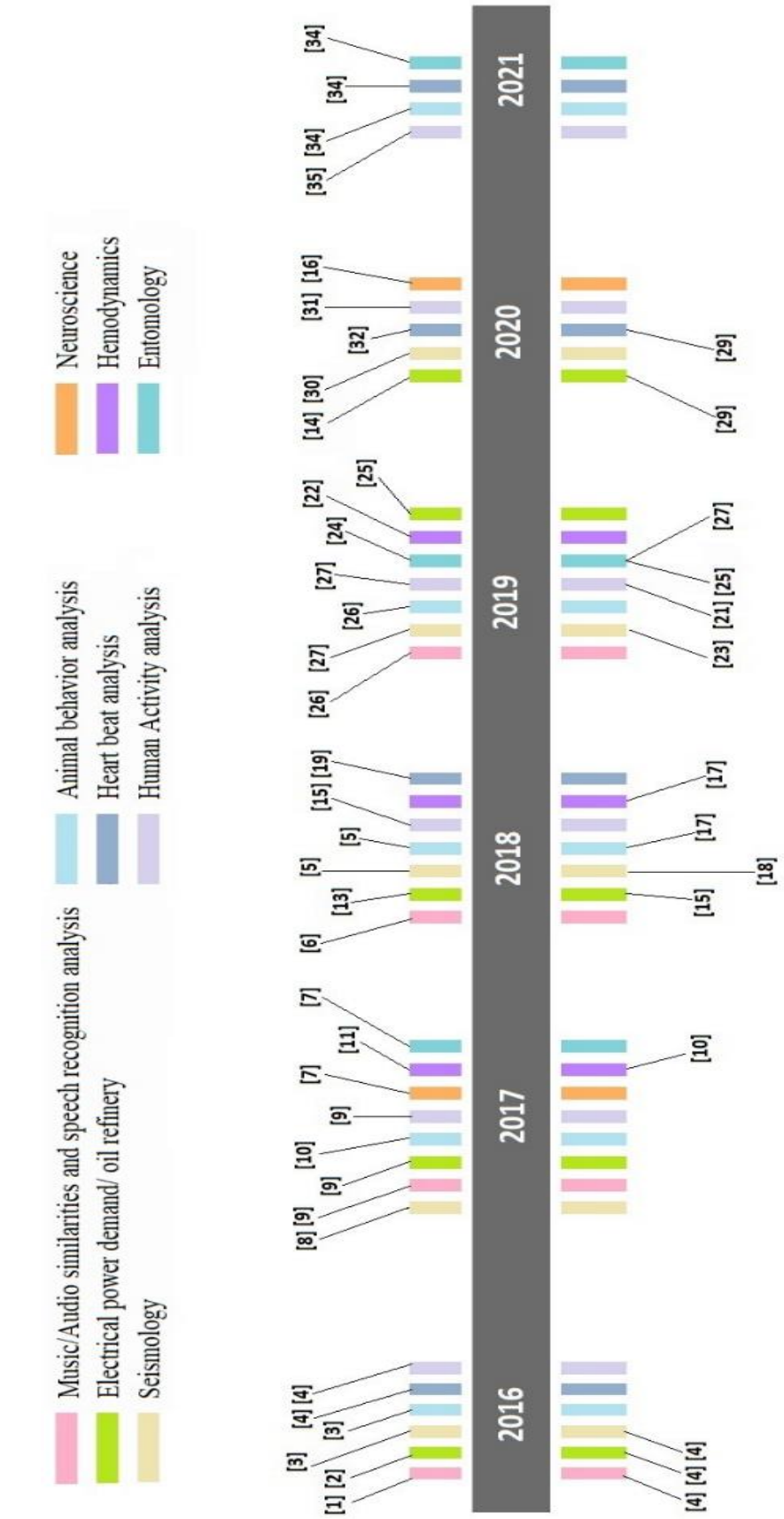


Fig. 2.5. Time series motif analysis timeline. It presents the main works that have contributed to the evolution of time series motif and pattern mining, divided between music/audio similarities and speech recognition analysis, Electrical power demand, seismology, animal behavior analysis, heart beat analysis, human activity analysis, neuroscience, hemodynamics, entomology.

Chapter 3

Motif Discovery Technique

Motif discovery techniques has been divided into 2 types; fixed length motif discovery and variable length motif discovery. The motif discovery methods use time series data and the data may be one-dimensional or multi-dimensional. Since their introduction in 2003, motif discovery techniques have been improved to reduce their time complexity and increase efficiency in large data sets. Most of the motif discovery algorithms are based on two common steps: Similarity measure and time series representation.

3.1 Representation Techniques

Numerous representation techniques for time series are introduced in literature to support time series motif discovery or pattern mining. But in this paper, we will basically focus on the representation technique i.e., Z-normalization. In Z-normalization, the input is a vector representation converted to an output vector with mean approximately zero and standard deviation close to one. This technique allows to focus on the similarities that are structural and not amplitude driven. The formula to calculate the z-normalized Euclidean distance D[i] between two time series subsequence Q and $T_{i,m}$ using their dot product, QT[i] is ([https://www.cs.ucr.edu/~eamonn/MatrixProfile.html] visit this page for detailed derivation):

$$D[i] = \sqrt{2m\left(1 - \frac{QT[i] - m\mu_Q M_T[i]}{m\sigma_Q \sum T[i]}\right)}$$

where m is the sub-sequence length, μ_Q is the mean of Q, $M_T[i]$ is the mean of $T_{i,m}$, σ_Q is the standard deviation of Q, and $\Sigma T[i]$ is the standard deviation of $T_{i,m.}$

3.2 Similarity Measures

The similarity of data in time series can be represented by constructing a suitable distance function for time series data. The fundamental aspect in motif finding algorithms is the similarity measure. Euclidean distance[1, 2, 3, 4, 5, 7, 8 etc.] dynamic temporal warping (DTW) [36]are some common similarity methods used for motif finding. In time series motif finding, Euclidean distance (ED) is one of the most improved distance measurements and in most of the literatures z-normalized Euclidean distance is used for finding motifs. However, in some contexts, Euclidean distance is not necessarily the most appropriate distance measure. DTW is commonly used to deal with local time-axis distortions. By enabling the time axis to be warped, this distance metric can match different regions of a time series. The shortest warping path in a distance matrix determines the best alignment. A warping path W is a set of consecutive matrix indices that define a time series mapping. The ideal path among numerous alternative warping paths is the one that minimizes the global warping cost. Dynamic programming with temporal complexity $O(n^2)$ can be used to calculate DTW. However, in our work we will apply a novel Euclidean distance similarity search technique called MASS introduced by Yeh et al. [2] for time series data motif discovery. The technique doesn't simply report the distance to the query's closest neighbour; it also returns the distance to each subsequence. It calculates the distance profile, in particular. By using the FFT to calculate the dot products between the query and all subsequences of the time series, the approach takes just O(n log n) time.

Chapter 4

Evaluation and Results

In this Section we have provided an experimental setup which describe processor details, size physical memory and types of libraries, packages , dataset description section, implementation details and discussion about result and analysis based on those results.

4.1 Experimental Setup

We implement our program in Jupyter Notebook (6.4.5), and execute the experiments in a machine running a Windows 10, 64bit operating system, x64-based processor and equipped with the following hardware: 11th Gen Intel(R) Core(TM) i5-1135G7 CPU @ of 2.40 GHz (8 GB RAM).

4.2 Libraries/Packages and Datasets Used

This section includes a table listing all of the libraries, packages, and datasets utilised in our experiment, along with their names, descriptions, and sources.

Table 4.1 Libraries, Packages and datasets used in the experiment

	Names	**Description**	**Source**
Libraries/ Packages	Stumpy	STUMPY is a powerful and scalable Python library for modern time series analysis that efficiently computes a matrix profile for solving host of time series problems.	[https://bit.ly /3xyS2DJ]
	matplotlib.pyplot	It is a state-based interface to matplotlib. It provides an implicit, MATLAB-like, way of plotting. It also opens figures on your screen, and acts as the figure GUI manager.	[https://bit.ly /3QrGdbg]

	Pandas	Pandas is a python library that is used for working with relational or labelled data. It provides various data structures and operations for manipulating numerical data and time series data.	[https://pandas.pydata.org/docs/]
	Numpy	It is a Python library that is used for scientific computation	[https://numpy.org/doc/stable/]
	Time	It is a python module that provides various time related functions.	[https://bit.ly/2ztZGEP]
	scipy.cluster.hierarchy	This module provides functions for hierarchical clustering. Its features include generating hierarchical clusters from distance matrices, calculating statistics on clusters, cutting linkages to generate flat clusters, and visualizing clusters with dendrograms.	[https://bit.ly/3N15F4t]
	matplotlib.patches	It is a python class used to patch a plot in different shapes. A patch is a 2D artist with a face color and an edge color.	[https://bit.ly/3NWmX3X]
	stumpy.ostinato	It is a function used to find the z-normalized consensus motif of multiple time series.	[https://bit.ly/3QtHXAU]
	stumpy.core.mass	It is a function used to compute the distance profile using the MASS algorithm.	[https://bit.ly/3QtHXAU]
Datasets	CO (Carbon Monoxide)	These are 8 different pollutants that causes air pollution, 3 of them are gas concentrations (CO, NO_2, O_3), other 3 of them are particulate matter (PM2.5, PM10, Sr) and rest are meteorological factors (RH, WS). All the datasets contain pollution level	[https://cpcb.nic.in/]
	NO_2 (Nitrogen Dioxide)		
	O_3 (Ozone)		
	PM2.5 (Particulate Matter 2.5, here 2.5 means that the diameter of the particle is <= 2.5 mm)		

	PM10 (Particulate Matter 10, here 10 means that the diameter of the particle is <= 10 mm)	for 21 different stations of Delhi. All the datasets have hourly logged data for around a year and a month i.e., 396 days. There are a total of 9504 tuples per pollutants. The total size of the dataset of single pollutant is 9504*21≈199584.
	RH (Relative Humidity)	
	Sr (Strontium)	
	WS (Wind Speed)	

4.3 Dataset Description

We perform our experiments on pollution time series dataset of 8 pollutants from different stations of Delhi. We collected our air pollution raw data from Central Pollution Control Board (CPBP) (https://www.india.gov.in/official-website-central-pollution-control-board). Dataset consists of gas concentration, particulate air pollutant and meteorological factors which help in analyzing air pollution. The first three of them are gas concentration which includes CO (carbon monoxide), NO_2 (nitrogen dioxide), O_3 (ozone); rest of them are particulate air pollutant includes PM2.5, PM10, Sr (strontium); and meteorological factors include WS (wind speed), RH (Relative humidity). There are 21 different stations in all the datasets. And concentration of all pollutants is taken at every regular interval of time of 60 minutes, i.e., every dataset contains hourly logged pollutant concentration and covers data for around a year and a month, i.e.,396 days. There are a total of 9504 tuples per pollutants. The total size of the data set of single pollutant is 9504*21≈199584

4.4 Implementation

In this work we have tried to implement the algorithm named Ostinato in our pollution time series dataset for finding consensus motif and analyzing the pollution levels. The algorithm was proposed by Kamgar et.al in their work [25]. Firstly, we have imported the suitable packages of python that we'll need to load the datasets, analyze those datasets, and plot the data. After that, we have loaded the dataset by reading the .csv files and storing the values of the datasets in an array and after that we have calculated

the exponential average for the values stored in the array. Exponential average is calculated so that we can avoid any sudden peak rise or fall in the dataset values. After that we have plotted the individual time series so that we can visualize the data and is shown in Fig. 4.1

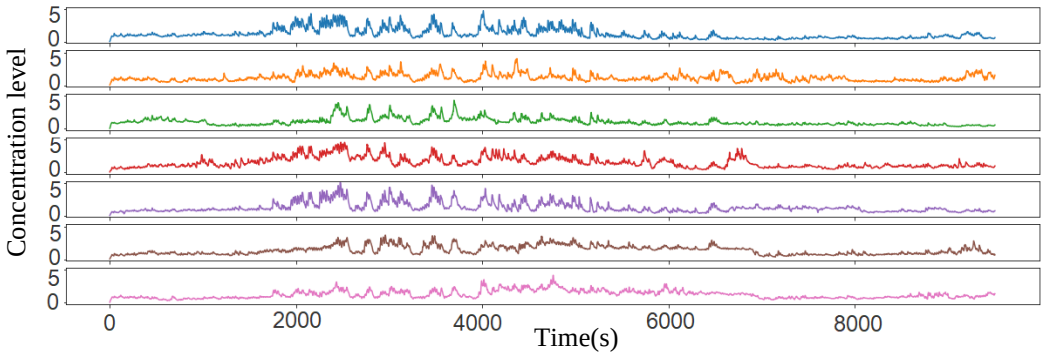
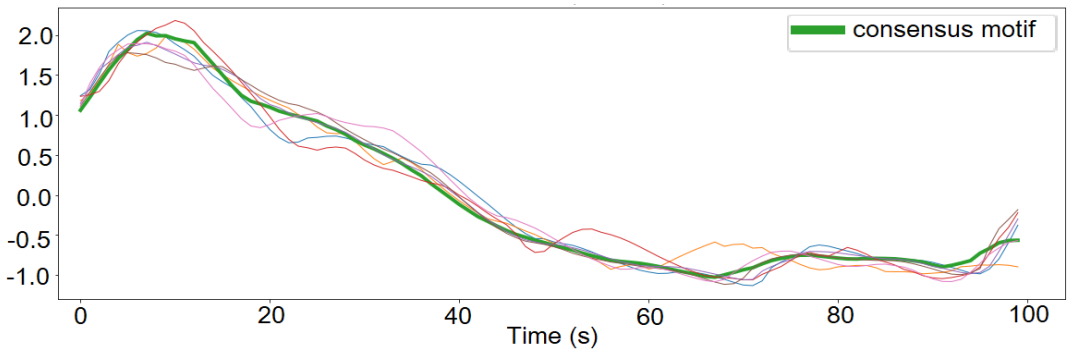


Fig. 4.1 Shows visual representation of the dataset of a pollutant from 7 different stations

After that we have tried to find the consensus motif among all the datasets by using the ostinato algorithm by passing the dataset values and the subsequence length after that we have plotted the z-normalized consensus motifs. After that we have recorded the radius and time taken for execution in finding the motif so that we can compare how the execution time varies. And we have also recorded the starting index of the consensus motif and in which time series it is present which can be seen in Fig. 4.2




```
Found Best Radius 1.62 in time series 2 starting at subsequence index location 2781.
Time: 1.5523
```

Fig. 4.2. The radius, index of the time series, and starting point of the subsequence of consensus motif in a set of 7 time series. for subsequence length 100.

After that we have plotted the consensus motif region in the graph plotted for individual datasets so that we can visually see where the consensus motifs are present and is shown in Fig. 4.3.

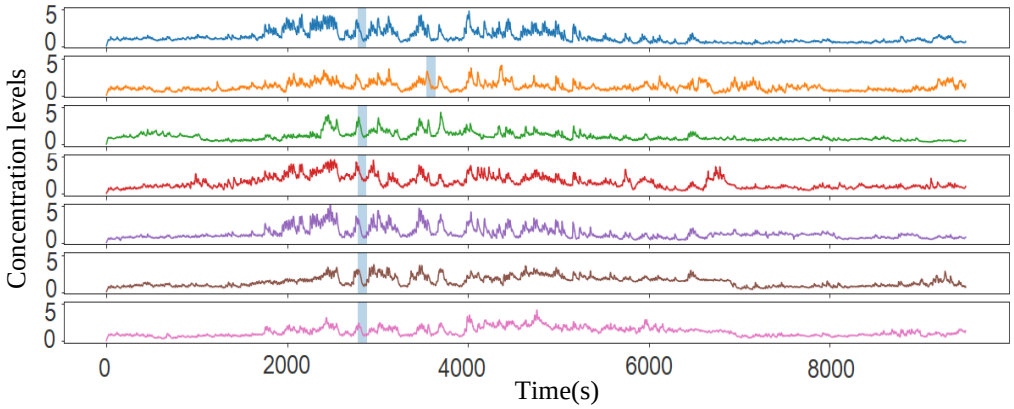


Fig 4.3. Shows the exact location of the subsequences of the consensus motif

Description of Ostinato Function:

List of Parameters: (i) a list (ii) subsequence length or window size. Here, Ts is the list that contains list of pollutant datasets. It is the list of time series for which nearest neighbour subsequence closest to the query subsequence should be found. The second parameter is m which is subsequence length or window size. The function returns: radius of the consensus motif , index of the time series in which it is occurring and the starting location of the subsequence index. Compare subsequences with the same radius and return the most central motif (i.e., having the smallest average nearest neighbour radii). In Ts, the best radius is found. The radius is the shortest distance a subsequence must travel to embrace at least one nearest neighbour subsequence from all other time series. The best radius in Ts is the smallest of all the radii. Multiple subsequences with the same optimal radius may exist in some data sets. Only one of them is found by the greedy Ostinato algorithm, and it may not be the most important motif. The subsequence with the least mean distance to nearest neighbours in all other time series is the most central motif among the subsequences with the best radius.

4.5 Results and Analysis

We have tried to identify similar pattern in pollution time series dataset. Basically, in this work we have found out consensus motifs in different sets of time series data. This work has been done in 2 ways: (i) Single station multiple pollutants for finding out the dependency of every pollutant and how they vary relative to each other across the station. (ii) Single pollutant multiple stations for finding how the pollutant vary and have what type on impact on the cities.

4.5.1 Single Station Multiple Pollutant

In this we have taken datasets of CO, NO_2, O_3, PM2.5, PM10, Sr, and WS. All the datasets are sorted in a way so that every dataset have same stations in all their respective columns. After that we have tried to find out consensus motifs for all the 21 stations for a set of 7 pollutants time series datasets mentioned above. We have done this experiment for different subsequence length i.e., 100, 200, 300, and 400. Fig. 4.4. shows the time series data visualization for the station Dwarka showing CO concentration in blue color, NO2 concentration in Orange, O3 in green, PM2.5 in red, PM10 in purple, WS in maroon and Sr in pink.

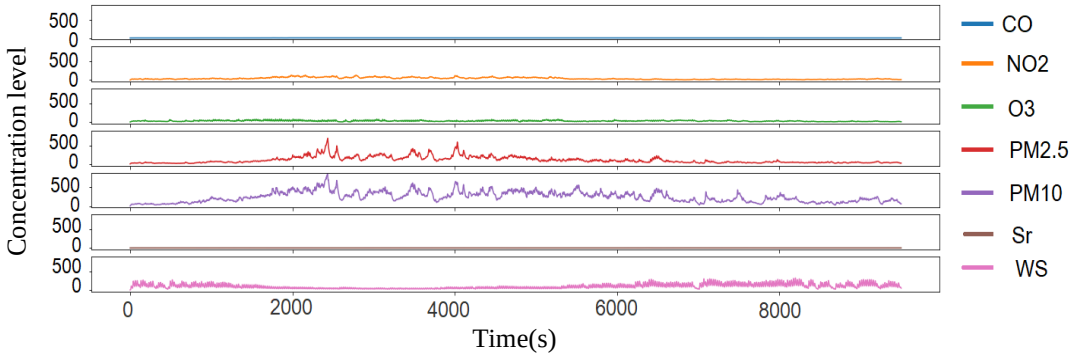


Fig. 4.4. Every pollutant's time series data visualization of Dwarka.

We'll pick four stations at random and discuss about the radius of the consensus motif, the time it takes to execute for different subsequence lengths, and so on. In Table 4.2. we can see that execution time for subsequence length is 100 is smaller the subsequence

length 200 and so on. Basically, we can say that execution time may depend on the subsequence length.

Table 4.2. Experimental results of execution time for finding consensus motif for a set of 7 pollutants of four stations.

Execution Time / Station Name	Subsequence length (100)	Subsequence length (200)	Subsequence length (300)	Subsequence length (400)
Dwarka-S8-DPCC	6.0719	12.8843	13.8924	22.6659
Okhla-Phase-2-DPCC	5.8498	16.0474	25.9133	32.3754
Sonia-Vihar-DPCC	4.014	15.0978	15.8049	16.0829
Mundka-DPCC	3.2533	4.6694	6.934	8.2225

In Table 4.3 we can see that the radius for subsequence length 100 does not vary much for all the four station they lie between 2.15 to 2.74. But , for subsequence length 200 they lie between 5.64 to 6.4, for subsequence length 300 they lie between 8.29 to 9.45, for subsequence length 400 they lie between 10.67 to 12.72, as the subsequence length increases variation in the radius also increases.

Table 4.3. Experimental results of consensus motif radius for a set of 7 pollutants of four stations.

Radius / Station Name	Subsequence length (100)	Subsequence length (200)	Subsequence length (300)	Subsequence length (400)
Dwarka-S8-DPCC	2.61	6.05	9.37	12.32
Okhla-Phase-2-DPCC	2.74	5.82	9.36	12.37
Sonia-Vihar-DPCC	2.15	5.56	8.29	10.67
Mundka-DPCC	2.63	6.4	9.45	12.72

In Fig 4.5, we can see that set of z-normalized time series consensus motifs of subsequence length 100 for the station Dwarka, Okhla Phase2, Sonia Vihar and Mundka. We can see in the figure one time series is shown with a thicker and bolder line because

it is the most central or seed time series. We can also notice that the variations in Okhla Phase 2 and Mundka are practically identical; this may be because both of these areas are industrial and have comparable trends in air pollution. However, the graph differs for each station, demonstrating that the consensus motifs are not random.

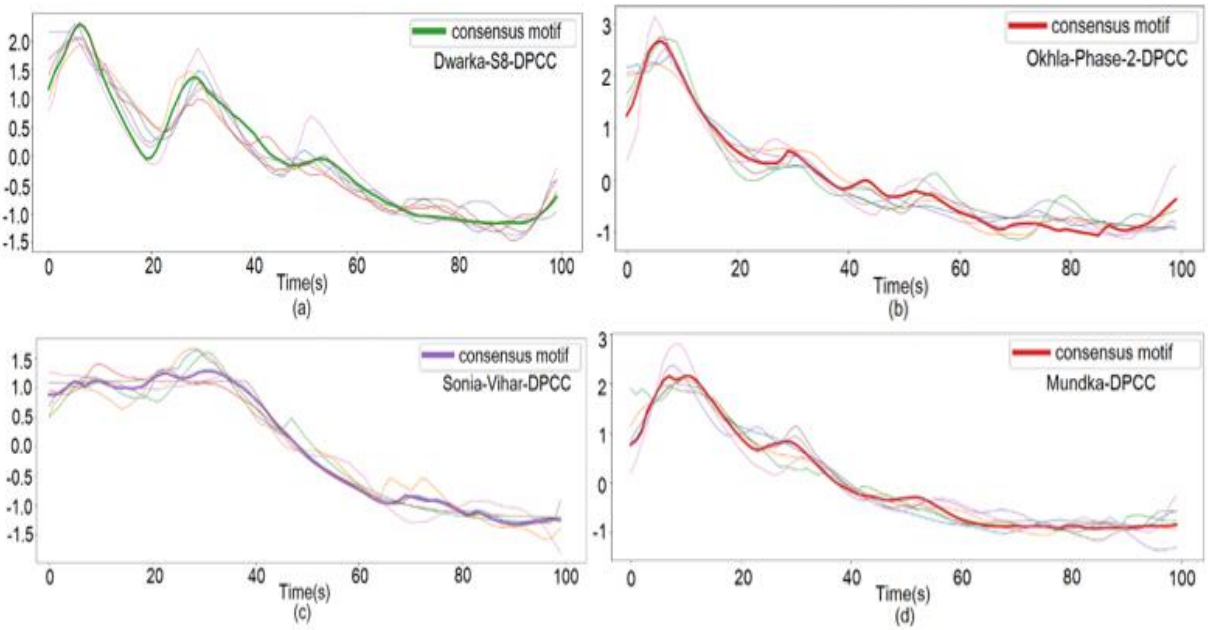


Fig. 4.5. Z-normalized consensus motif for subsequence length 100 of 4 stations: (a) Dwarka-S8-DPCC, (b) Okhla-phase-2-DPCC, (c) Sonia-Vihar-DPCC, (d) Mundka-DPCC

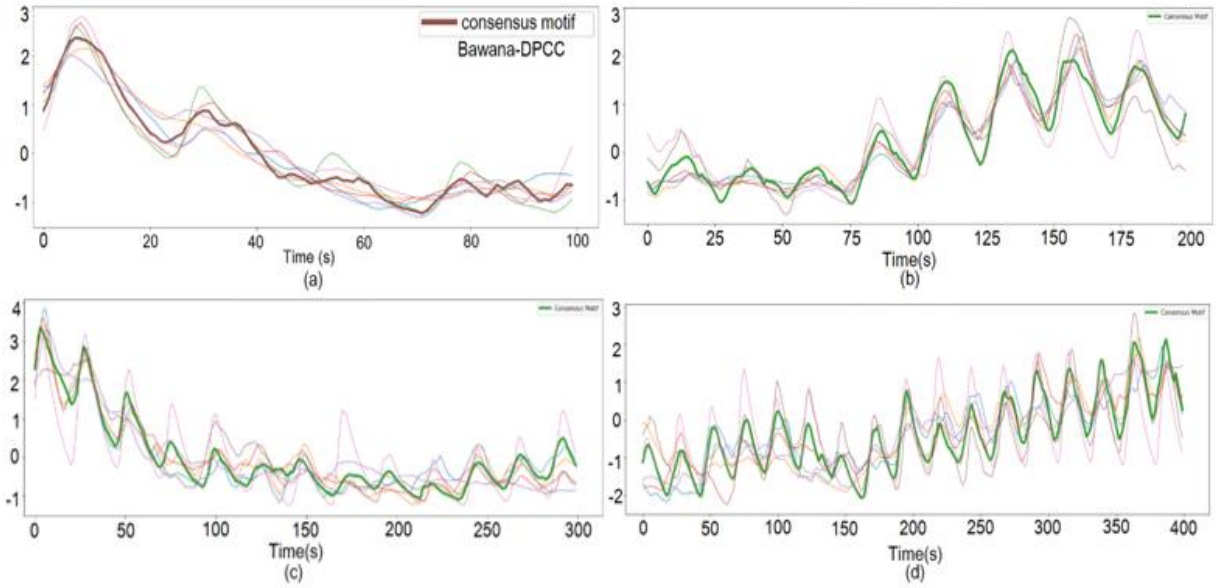


Fig 4.6. Z-normalized consensus motif of Bawana for different subsequence length: (a) 100, (b) 200, (c) 300, (d) 400

In the above figure i.e., Fig 4.6 , we can see consensus motifs of the station Bawana for subsequence length 100 in Fig 4.6 (a), subsequence length 200 in Fig 4.6 (b), subsequence length 300 in Fig 4.6 (c), subsequence length 400 in Fig 4.6 (d), and all of them have different patterns.

Fig.4.7. Comparison of execution time for all 21 stations based on different subsequence length for finding consensus motif in a set of 7 pollutants.

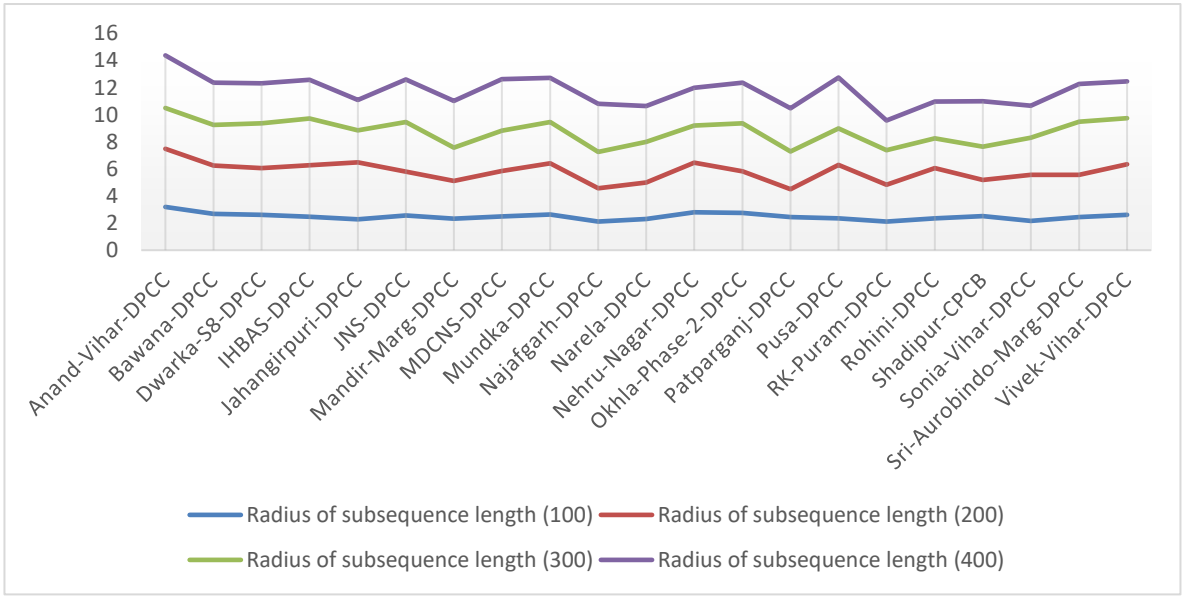


Fig.4.8. Comparison of radius of consensus motif in a set of 7 pollutants for all 21 stations based on different subsequence length.

4.5.2 Single Pollutant Multiple Stations

In this we have created datasets by station names and every station's dataset will contain pollutants concentration of CO, NO_2, O_3, PM2.5, PM10, RH, Sr, and WS in each column. So, that while reading the .csv files every data set's 1st column will be read first and will contain only the 1st pollutants concentration i.e., CO and so on. In this we have particularly taken only 7 stations i.e., Dwarka, Okhla Phase2, Mundka, Rohini, Sonia Vihar, and Sri Aurobindo Marg. We have done this experiment for different subsequence length i.e., 100, 200, 300, and 400.

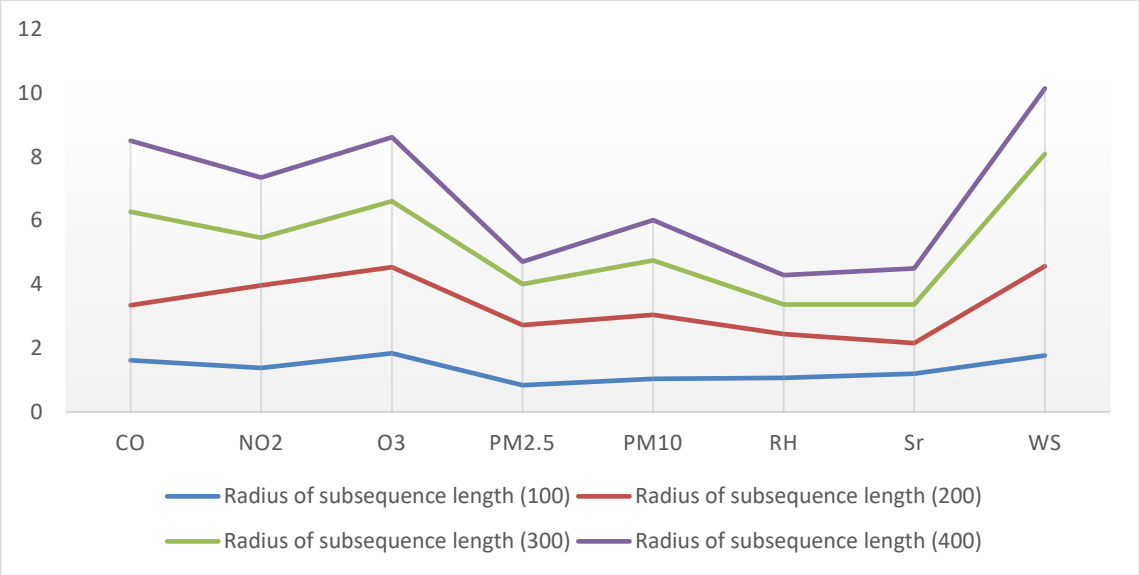


Fig.4.9. Comparison of radius of consensus motif in time series data of 8 different pollutants in group of 7 different stations.

In Fig. 4.9, we can see that radius is increasing as the subsequence length increase i.e., radius is directly proportional to subsequence length because for longer window size the search in done for a large area so, automatically the radius will increase. We can also see that the variation in radius for subsequence length 400 is the most, fluctuations can be seen very clearly but in on the other hand variation for radius of length 100 is minimal , the blue line is almost a straight line and from the Table 4.4., also we can see that the radius ranges from 0.84 to 1.84 but for the maximum subsequence length it ranges from 4.29 to 10.14. This may happen due to the pollutants not showing similar

trend in their concentration value for a longer period of time and that is why the variation is too much.

Table 4.4. Experimental results of consensus motif radius of 8 pollutants for a group of 7 stations

Radius / Pollutant	Subsequence length (100)	Subsequence length (200)	Subsequence length (300)	Subsequence length (400)
CO	1.62	3.35	6.28	8.5
NO2	1.38	3.97	5.46	7.35
O3	1.84	4.54	6.61	8.61
PM2.5	0.84	2.72	4.01	4.71
PM10	1.04	3.05	4.75	6.01
RH	1.07	2.44	3.37	4.29
Sr	1.20	2.16	3.37	4.5
WS	1.77	4.57	8.09	10.14

In Fig.4.10, we can see that the execution time does not does not depends on the subsequence length as it was for the above experiment where we have calculated the radius for a single station have multiple pollutant rather in this case it is varying based on some underlying factor.

Fig.4.10 Comparison of execution time for all pollutants based on different subsequence length for finding consensus motif in group of 7 different stations.

Table 4.5. Experimental results of execution time for finding consensus motif of 8 pollutants for a group of 7 stations

Execution Time / Pollutant	Subsequence length (100)	Subsequence length (200)	Subsequence length (300)	Subsequence length (400)
CO	1.5523	2.1342	3.902	5.7712
NO2	1.339	3.0965	2.6023	2.9233
O3	2.051	3.8655	4.7598	6.3505
PM2.5	1.2402	2.7553	5.8967	6.6645
PM10	1.3145	3.1948	5.6933	9.0229
RH	2.0047	4.2208	6.4822	7.3731
Sr	11.2149	5.438	7.8041	11.975
WS	1.0734	3.3135	8.2319	9.3081

In Fig. 4.11, we can see consensus motifs of the pollutant CO for subsequence length 100 in (a), subsequence length 200 in (b), subsequence length 300 in (c), subsequence length 400 in (d), and all of them have different patterns.

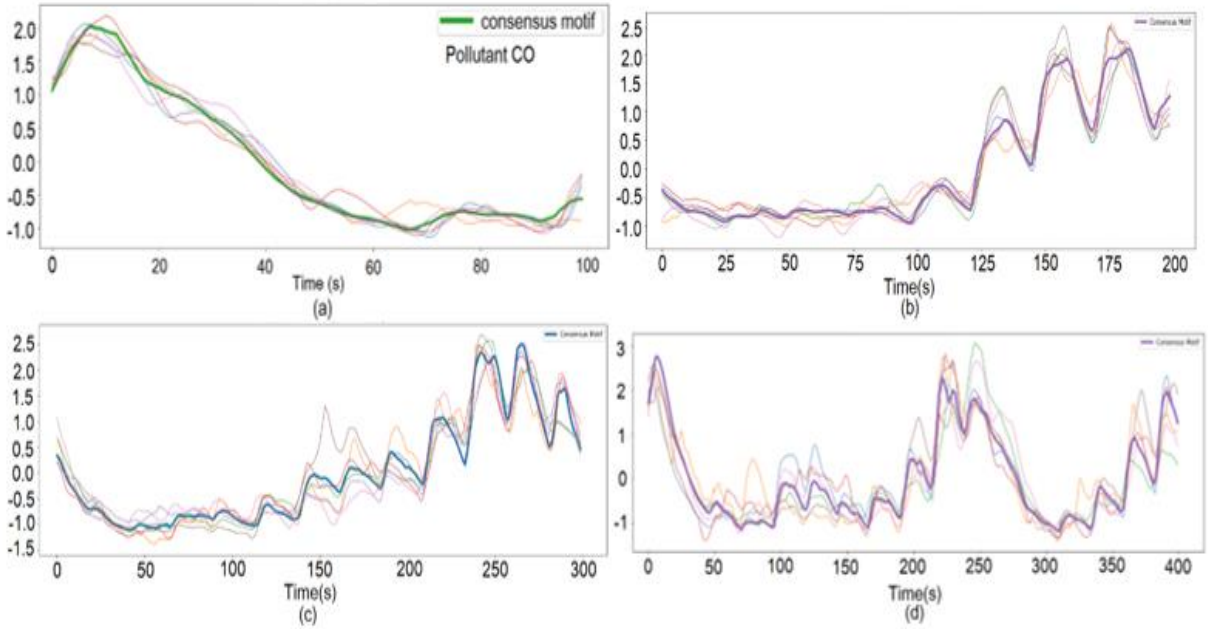


Fig. 4.11 Z-normalized consensus motif of the pollutant CO for different subsequence length: (a) 100, (b) 200, (c) 300, (d) 400

Chapter 5

Conclusion and Future Work

In this Chapter, we discussed conclusion which will give an overall glimpse of our work and future scope. The future scope section will provide how motif discovery can be more useful for accurately knowing the patterns of air pollutants.

5.1 Conclusion

In this work we have found consensus motif for a set of air pollutants time series data, we have chosen air pollution data for 21 different stations of Delhi and have used data from roughly around a year and a month, i.e., 396 days. We attempted to discover consensus motifs for various subsequence lengths, and also logged their execution times in order to determine the time taken for finding the radius of the consensus motif, as well as the starting point of the consensus motif subsequence so that we could plot them in the raw data representation easily. It is advantageous to improve the prediction of air pollution particles in urban and sub-urban areas and it will help further to prevent air pollution in metropolitan areas.

5.2 Future Work

The future scope of this work can include the application of finding similar patterns in different pollutants based on some other techniques to find time series motifs for more accurately predicting which pollutant is having high concentration at which point of time and according to those data, necessary precautions can be taken for reducing their effects on the environment and human health. Moreover, this air pollution-based pattern identification will assist us in determining the factors that aggravated pollutants concentrations at what times, as well as whether there is any similarity in their trend of causing air pollution, and will further assist us in preventing air pollution in highly polluted urban and suburban areas.

Bibliography

[1] Silva, D.F., Yeh, C.C.M., Batista, G.E. and Keogh, E.J., 2016, August. SiMPle: Assessing Music Similarity Using Subsequences Joins. In *ISMIR* (pp. 23-29).

[2] Yeh, C.C.M., Zhu, Y., Ulanova, L., Begum, N., Ding, Y., Dau, H.A., Silva, D.F., Mueen, A. and Keogh, E., 2016, December. 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). Ieee.

[3] Zhu, Y., Zimmerman, Z., Senobari, N.S., Yeh, C.C.M., Funning, G., Mueen, A., Brisk, P. and Keogh, E., 2016, December. 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). IEEE.

[4] Yeh, C.C.M., Van Herle, H. and Keogh, E., 2016, December. 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). IEEE.

[5] Zhu, Y., Zimmerman, Z., Shakibay Senobari, N., Yeh, C.C.M., Funning, G., Mueen, A., Brisk, P. and Keogh, E., 2018. Exploiting a novel algorithm and GPUs to break the ten quadrillion pairwise comparisons barrier for time series motifs and joins. *Knowledge and Information Systems*, *54*(1), pp.203-236.

[6] Silva, D.F., Yeh, C.C.M., Zhu, Y., Batista, G.E. and Keogh, E., 2018. Fast similarity matrix profile for music analysis and exploration. *IEEE Transactions on Multimedia*, *21*(1), pp.29-38.

[7] Yeh, C.C.M., Kavantzas, N. and Keogh, E., 2017. Matrix profile IV: using weakly labeled time series to predict outcomes. *Proceedings of the VLDB Endowment*, *10*(12), pp.1802-1812.

[8] Dau, H.A. and Keogh, E., 2017, August. Matrix profile v: A generic technique to incorporate domain knowledge into motif discovery. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 125-134).

[9] Yeh, C.C.M., Kavantzas, N. and Keogh, E., 2017, November. Matrix profile VI: Meaningful multidimensional motif discovery. In *2017 IEEE international conference on data mining (ICDM)* (pp. 565-574). IEEE.

[10] Zhu, Y., Imamura, M., Nikovski, D. and Keogh, E., 2017, November. 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). IEEE.

[11] Gharghabi, S., Ding, Y., Yeh, C.C.M., Kamgar, K., Ulanova, L. and Keogh, E., 2017, November. Matrix profile VIII: domain agnostic online semantic segmentation at superhuman performance levels. In *2017 IEEE international conference on data mining (ICDM)* (pp. 117-126). IEEE.

[12] Linardi, M., Zhu, Y., Palpanas, T. and Keogh, E., 2018, May. Matrix profile X: VALMOD-scalable discovery of variable-length motifs in data series. In *Proceedings of the 2018 International Conference on Management of Data* (pp. 1053-1066).

[13] Zhu, Y., Yeh, C.C.M., Zimmerman, Z., Kamgar, K. and Keogh, E., 2018, November. Matrix profile XI: SCRIMP++: time series motif discovery at interactive speeds. In *2018 IEEE International Conference on Data Mining (ICDM)* (pp. 837-846). IEEE.

[14] Gharghabi, S., Imani, S., Bagnall, A., Darvishzadeh, A. and Keogh, E., 2020. An ultra-fast time series distance measure to allow data mining in more complex real-world deployments. *Data Mining and Knowledge Discovery*, *34*(4), pp.1104-1135.

[15] Imani, S., Madrid, F., Ding, W., Crouter, S. and Keogh, E., 2018, November. 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). IEEE.

[16] Zhu, Y., Gharghabi, S., Silva, D.F., Dau, H.A., Yeh, C.C.M., Shakibay Senobari, N., Almaslukh, A., Kamgar, K., Zimmerman, Z., Funning, G. and Mueen, A., 2020. The Swiss army knife of time series data mining: ten useful things you can do with the matrix profile and ten lines of code. *Data Mining and Knowledge Discovery*, *34*(4), pp.949-979.

[17] Zhu, Y., Imamura, M., Nikovski, D. and Keogh, E.J., 2018, July. Time Series Chains: A Novel Tool for Time Series Data Mining. In *IJCAI* (pp. 5414-5418).

[18] Yeh, C.C.M., Zhu, Y., Ulanova, L., Begum, N., Ding, Y., Dau, H.A., Zimmerman, Z., Silva, D.F., Mueen, A. and Keogh, E., 2018. Time series joins, motifs, discords and shapelets: a unifying view that exploits the matrix profile. *Data Mining and Knowledge Discovery*, *32*(1), pp.83-123.

[19] Linardi, M., Zhu, Y., Palpanas, T. and Keogh, E., 2018, May. VALMOD: A suite for easy and exact detection of variable length motifs in data series. In *Proceedings of the 2018 International Conference on Management of Data* (pp. 1757-1760).

[20] Shi, J., Yu, N., Keogh, E., Chen, H.K. and Yamashita, K., 2019, November. Discovering and labeling power system events in synchrophasor data with matrix profile. In *2019 IEEE Sustainable Power and Energy Conference (iSPEC)* (pp. 1827-1832). IEEE.

[21] Gharghabi, S., Yeh, C.C.M., Ding, Y., Ding, W., Hibbing, P., LaMunion, S., Kaplan, A., Crouter, S.E. and Keogh, E., 2019. Domain agnostic online semantic segmentation for multi-dimensional time series. *Data mining and knowledge discovery*, *33*(1), pp.96-130.

[22] Zhu, Y., Imamura, M., Nikovski, D. and Keogh, E., 2019. Introducing time series chains: a new primitive for time series data mining. *Knowledge and Information Systems*, *60*(2), pp.1135-1161.

[23] Zimmerman, Z., Kamgar, K., Senobari, N.S., Crites, B., Funning, G., Brisk, P. and Keogh, E., 2019, November. 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* (pp. 74-86).

[24] Madrid, F., Singh, S., Chesnais, Q., Mauck, K. and Keogh, E., 2019, October. Matrix profile xvi: efficient and effective labeling of massive time series archives. In *2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 463-472). IEEE.

[25] Kamgar, K., Gharghabi, S. and Keogh, E., 2019, November. 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). IEEE.

[26] Imani, S. and Keogh, E., 2019, November. 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). IEEE.

[27] Zimmerman, Z., Senobari, N.S., Funning, G., Papalexakis, E., Oymak, S., Brisk, P. and Keogh, E., 2019, November. 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). IEEE.

[28] Madrid, F., Imani, S., Mercer, R., Zimmerman, Z., Shakibay, N. and Keogh, E., 2019, November. Matrix profile xx: Finding and visualizing time series motifs of all lengths using the matrix profile. In *2019 IEEE International Conference on Big Knowledge (ICBK)* (pp. 175-182). IEEE.

[29] Linardi, M., Zhu, Y., Palpanas, T. and Keogh, E., 2020. Matrix profile goes MAD: variable-length motif and discord discovery in data series. *Data Mining and Knowledge Discovery*, *34*(4), pp.1022-1071.

[30] Zhu, Y., Mueen, A. and Keogh, E., 2018. Admissible time series motif discovery with missing data. *arXiv preprint arXiv:1802.05472*.

[31] Zhu, Y., Yeh, C.C.M., Zimmerman, Z. and Keogh, E., 2020, April. Matrix profile xvii: Indexing the matrix profile to allow arbitrary range queries. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)* (pp. 1846-1849). IEEE.

[32] Nakamura, T., Imamura, M., Mercer, R. and Keogh, E., 2020, November. 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). IEEE.

[33] Alaee, S., Kamgar, K. and Keogh, E., 2020, November. Matrix profile XXII: exact discovery of time series motifs under DTW. In *2020 IEEE International Conference on Data Mining (ICDM)* (pp. 900-905). IEEE.

[34] Mercer, R., Alaee, S., Abdoli, A., Singh, S., Murillo, A. and Keogh, E., 2021, December. 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). IEEE.

[35] Alaee, S., Mercer, R., Kamgar, K. and Keogh, E., 2021. Time series motifs discovery under DTW allows more robust discovery of conserved structure. *Data Mining and Knowledge Discovery*, *35*(3), pp.863-910.

[36] Agrawal, R., Faloutsos, C. and Swami, A., 1993, October. Efficient similarity search in sequence databases. In *International conference on foundations of data organization and algorithms* (pp. 69-84). Springer, Berlin, Heidelberg.

[37] Chiu, B., Keogh, E. and Lonardi, S., 2003, August. Probabilistic discovery of time series motifs. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 493-498).