Original research article

Intelligent Decision Technologies
1–21

DOI: 10.1177/18724981251318529
journals.sagepub.com/home/idt
Sage | IOS Press

Discovering new relationships using multipoles through correlation and threshold analysis

Pratistha Gaur[1], Pavan Kumar Vadla[1], P Sudarshan[1], Prrathyush Kedeelaya[1] and C Narendra Babu[1]

Abstract

Uncovering novel relationships in time series data is crucial for understanding complex systems. Traditional methods, such as linear regression and frequency domain analysis, often fail to capture the complex, non-linear interactions present in time series data. Multipoles represent a new class of linear relationships among multiple time series, where each series significantly contributes to the overall linear dependence. In this paper, we demonstrate the utility of multipoles in climate science and neuroscience, revealing meaningful patterns and interactions. The proposed approach transforms time series data into a multipole representation, decomposing it into multipolar components to detect intricate dependencies typically obscured by conventional techniques. This work uses scalable algorithms for efficient multipole transformation and analysis, validated through experiments on real-world datasets. Experimental results using climate data and blood pressure data show the significance in detecting hidden patterns and understanding underlying correlations helping us map relationships between the relevant parameters. In climatic time series, our method uncovers weather dynamics with threshold values, while in healthcare, it reveals complex interactions between physiological parameters demonstrating graphical results. This paper contributes a novel perspective on time series analysis, enriching the analytical toolbox for researchers and practitioners, and offers a promising direction for exploring complex data relationships in various domains, including climate science and neuroscience.

Keywords

multipoles, bipoles, time series analysis, correlation networks, Pearson correlation, ARIMA model, logarithmic spiral

Received: 17 July 2024; accepted: 22 January 2025

1 Introduction and related works

Time series data, consisting of sequences of data points recorded at successive time intervals, is pivotal across diverse fields such as finance, healthcare, environmental monitoring, and engineering. The primary goal of time series analysis is to understand the underlying structure of the data, identify patterns, and make forecasts. This analysis can reveal crucial insights, including trends, seasonal variations, and irregular patterns, which are essential for informed decision-making. This paper defines a novel class of linear relationships involving more than two time series- multipoles We outline a framework for identifying and analysing multipoles and demonstrate their application in scientific domains.

Researchers have explored its applications in various fields and have come up with significant results.[1] Presents a novel approach for identifying strongly correlated pairs of variables by leveraging a support-based upper bound of Pearson's correlation coefficient, with a focus on efficiency, possibly with applications in data mining and knowledge discovery. A graph-based approach[2] to identify teleconnections in climate data, which could have significant implications for

[1]Department of Computer Science and Engineering, MS Ramaiah University of Applied Sciences, Bengaluru, India

Corresponding author:
C Narendra Babu, Department of Computer Science and Engineering, MS Ramaiah University of Applied Sciences, Bengaluru, India.
Email: narendrababu.c@gmail.com

understanding global climate dynamics and its impacts on regional weather patterns. The authors[3] propose a method for discovering dynamic dipoles in climate data, which could lead to insights into complex climate phenomena and their temporal variations. The research work[4] highlights a method for sparse inverse covariance estimation using the graphical lasso, which has wide-ranging applications in various domains, particularly in analysing high-dimensional data with complex dependencies between variables. In,[5]a method for variable selection in high-dimensional datasets using the Lasso method is outlined. This approach has wide-ranging applications in fields where datasets often contain a large number of variables relative to the number of observations. In,[6] climate dynamics is explored by investigating the community structure and dynamics within climate networks. This research helps identify patterns of interconnectedness between different climate variables or regions, leading to insights into the complex interactions that drive climate variability and change. The authors[7] propose a method in the field of climate science and data mining by discovering dynamic dipoles in climate data. By identifying pairs of climate phenomena with opposing behaviours that change over time, this research enhances our understanding of complex climate dynamics and their temporal variations. Time series analysis using tripoles is introduced in[8] which outlines an approach for discovering a new class of relationships within temporal data. This research expands the analytical capabilities for understanding complex temporal dynamics and patterns. In field of machine learning and climate science[9] introduces the Sparse Group Lasso method and providing theoretical insights into its consistency properties. Additionally, it demonstrates the practical relevance of the method through applications in climate science, highlighting its potential for variable selection and regularization in high-dimensional climate datasets. The work in[10] proposes a method for classifying short time series data with missing values using Multivariate Functional Linear Discriminant Analysis (MFLDA). The research[11] introduces the concepts of tripoles as a new class of relationships within temporal data. By extending traditional pairwise relationships to include interactions between three variables, this research expands the analytical capabilities for understanding complex temporal dynamics and patterns.[12] Presents temporal causal modelling by a methodological approach that leverages graphical Granger methods for analysing causal relationships among time series variables. A methodological approach in[13] is tailored to the task of discovering groups of time series data with similar behaviour within short intervals of time, addressing the limitations of traditional clustering techniques in handling temporal data effectively. The research work[14] insinuates a novel approach for outlier detection in multivariate time series data using independent component analysis, offering a methodological alternative to traditional outlier detection methods. A methodology for detecting changes in time series data is introduced in[15] that considers contextual information, providing a more nuanced and accurate understanding of how changes manifest within different contexts. In the field of probabilistic approach for discovering motifs in time series data,[16] offers a methodological alternative to traditional motif discovery methods that can better handle noisy or uncertain data. Eichler's paper[17] contributes to the understanding of causal inference in the context of multiple time series data, highlighting both the principles underlying such inference and the challenges involved in its practical application. Fu's paper[18] provides a comprehensive overview of time series data mining, covering foundational concepts, analytical techniques, practical applications, and challenges, making it a valuable resource for researchers, practitioners, and students in the field. Keogh, Chu, Hart, and Pazzani's research in[19] contributes to the field of time series analysis by providing a comprehensive survey of existing segmentation techniques and presenting a novel approach that addresses the limitations of current methods, offering potential improvements in segmenting time series data. Paper[20] proposes an efficient algorithm, Hot SAX, for detecting the most unusual subsequence within time series data, offering a methodological advancement that addresses the limitations of existing techniques in terms of efficiency and scalability. Multivariate Time Series Analysis and Applications serves as a comprehensive resource for researchers, practitioners, and students interested in understanding and applying multivariate time series analysis techniques in diverse fields.

In this context, this work introduces a multipole-based framework for time series analysis, aiming to capture complex, non-linear relationships while maintaining interpretability, demonstrating improved predictive accuracy, and uncovering hidden patterns in various datasets in comparison to the contemporary works in existence.

To assess the performance of this method compared to traditional time series analysis techniques, a comparative evaluation was conducted. While conventional methods like PCA or linear regression capture simple linear dependencies, the proposed multipole-based approach excels at uncovering complex, non-linear interactions, particularly in multidimensional datasets. However, it requires more computational resources and can be sensitive to threshold selection, making it less robust in some scenarios. Future work could further benchmark this method across diverse datasets to clarify its strengths in handling intricate dependencies.

The rest of the paper is organized as follows. Section 2 highlights the concepts of Tripoles and Multipoles. The proposed method is outlined in section 3. Case studies with the application of multipoles are outlined in section 4. Results and discussion are presented in section 4. The paper concludes in section 5.

2 Multipoles and its relationships in time series data

The multipole-based framework offers several key advantages. It serves as a robust tool for capturing non-linear dependencies and improving the interpretability of results. This approach is both theoretically innovative and practically valuable, with wide applications in domains where time series data is crucial. The research is driven by the potential of multipole theory to tackle key challenges in the field. This work highlights the multipole-based framework's application, including its theoretical basis, algorithmic implementation, and empirical validation. Through experiments on real-world datasets, the framework's effectiveness in identifying new relationships in time series data is demonstrated. Extensive experiments validate the methodology, showcasing its applicability in fields such as healthcare and environmental science. This section provides a thorough overview of the core concepts and existing models in time series analysis.

2.1 Overview of time series analysis

Time series analysis involves statistical methods and techniques for examining data points collected or recorded at specific time intervals.

2.2 Bipoles, tripoles and multipoles

- Bipole: It refers to a relationship between two time series where their correlation, either positive or negative, is significant. Dipoles, a specific type of bipole, often represent pairs of distant geographical regions whose climate anomaly time series are negatively correlated.
- Tripole: It involves three time series T0, T1, and T2. The sum of the time series T1 + T2, exhibits a much stronger correlation with T0 than either T1 or T2 individually.
- Multipole: These extend the concept of tripoles to more than three time series. They involve complex relationships where the combined behaviour of multiple time series provides a better understanding of another time series.

2.3 Pearson correlation

The Pearson Correlation Coefficient (r) measures the linear relationship between two continuous variables.

$$(r) = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{\sqrt{\left[n\sum x^2 - \left(\sum x\right)^2\right]\left[n\sum y^2 - \left(\sum y\right)^2\right]}} \tag{1}$$

where
n = number of pairs of score s
$\sum xy$ = sum of the product of paired scores
$\sum x$ = sum of x score s
$\sum y$ = sum of y scores
$\sum x^2$ = sum of squared x scores
$\sum y^2$ = sum of squared y scores

2.4 Logarithmic spiral

A logarithmic spiral, also known as an equiangular spiral, is a self-similar curve which appears frequently in nature. The defining characteristic of a logarithmic spiral is that the angle between the tangent to the curve and the radial line from the origin is constant.

In polar coordinates (r, θ), the logarithmic spiral can be expressed with the following equation: $r = ae^{b\theta}$ (2) Where r = the radial distance from the origin, θ = the angle in radians, a = a positive real constant that determines the size of the spiral, b = a real constant that determines the rate of growth of the spiral. If $b > 0$, the spiral winds outward as θ increases; if $b < 0$, the spiral winds inward.

2.5 Naïve approach

The naive approach computes statistical correlations between time series data and forms a baseline for evaluating techniques like multipole transformation.

Naive Logarithmic Spiral Algorithm:
Step 1 Initialization of parameters:

a : Scale facto r
b : Growth rate
$\theta_{\min}$: Starting angl e
$\theta_{\max}$: Ending angle
$\Delta\theta$: Step size for angle increment .

Step 2 Point Generation:

Initialize an empty list point to store Cartesian coordinates.
For each θ in the range $[\theta_{\min}, \theta_{\max}]$ with step $\Delta\theta$:
Compute radial distance r using equation $r = a \cdot e^{b \cdot \theta}$.
Convert polar coordinates to Cartesian coordinates:

$$x = r \cdot \cos(\theta) \tag{3}$$

$$y = r \cdot \sin(\theta) \tag{4}$$

Append (x, y) to points.

Step 3 Output:

Return the list points containing the generated Cartesian coordinates.

2.6 Case study 1: climate data and geography of cities

This section focuses on the geography and climate of three major cities: Bengaluru, Mumbai and Hyderabad.

- Bengaluru: Located in southern India, the state of Karnataka, with coordinates 12.9716° N, 77.5946° E. Its terrain is mostly flat to slightly undulating. The city experiences a tropical savanna climate with three distinct seasons: moderate summers (March to May) with average high temperatures around 34°C (93°F); heavy monsoon rains (June to September) bringing approximately 970 mm of annual precipitation; and mild, pleasant winters (October to February) with average highs of 27°C (81°F) and lows around 15°C (59°F).
- Mumbai (Bombay): Located in western India along the Arabian Sea coast, the state of Maharashtra, is positioned at 19.0760° N, 72.8777° E. The city features a mix of coastal plains and low-lying hills. It has a tropical wet and dry climate with hot and humid summers (March to May), temperatures ranging from 25°C (77°F) to 35°C (95°F); extremely heavy monsoon rainfall (June to September) totalling about 2200 mm annually, often causing flooding; and mild, relatively dry winters (October to February) with temperatures between 18°C (64°F) and 30°C (86°F).
- Hyderabad: Situated in south-central India, the state of Telangana, with coordinates 17.3850° N, 78.4867° E. The city is characterized by rocky terrain and hilly regions. It experiences tropical wet and dry climate with very hot summers (March to May), where temperatures often exceeds 40°C (104°F); moderate to heavy monsoon rains (June to September) bringing around 800 mm of annual precipitation; and mild, comfortable winters (October to February) with temperatures ranging from 13°C to 29°C.

2.7 Case study 2: physiological data (blood pressure)

Blood pressure is the force exerted by circulating blood on the walls of blood vessels. It is measured using two readings-systolic pressure (the pressure in the arteries when the heart beats and fills them with blood) and diastolic pressure (the pressure in the arteries when the heart rests between beats).

The normal blood pressure range is systolic less than 120 mmHg and diastolic less than 80 mmHg. The factors that influence blood pressure are – age, diet, weight, physical activity, stress and genes (family history).

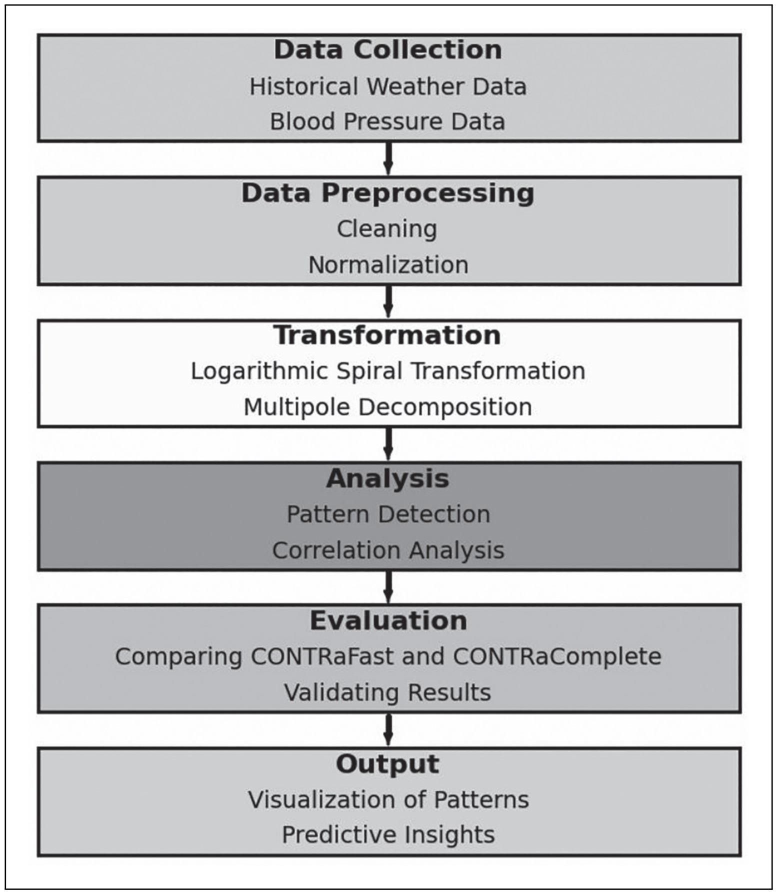

Figure 1. Workflow for uncovering novel relationships in time series data using multipole transformation.

3 Discovering relationships in time series data using multipoles with case studies

The project takes an advanced approach to tackle the challenges of classifying complex, multidimensional climate and health datasets. By utilizing correlation and threshold analysis techniques, it captures intricate relationships within the data. This approach enhances relational accuracy by accounting for both the intrinsic structure and temporal patterns, creating a solid framework for managing diverse and high-dimensional datasets. Additionally, it thoroughly evaluates the proposed methods, confirming their effectiveness and practical importance. This framework represents a promising direction for exploring complex data relationships by merging principles from statistical analysis and time-series correlation, with applications across climate science, neuroscience, epidemiology, and public health monitoring.

Figure 1 presents a block diagram of the data processing pipeline, which begins with data collection and proceeds through preprocessing, transformation, analysis, and evaluation, ultimately uncovering novel relationships in time series data and affirming the significance of multipole transformation.

Data preprocessing is essential for effectively analyzing complex climate and health datasets. The preprocessing pipeline begins with data cleaning, addressing missing values and outliers to ensure data integrity. We then apply logarithmic transformations to stabilize variance and normalize distributions, followed by normalization to standardize variable scales. Temporal alignment is performed to synchronize the time series data across variables. Feature extraction is employed to capture key patterns within the datasets. Subsequently, correlation and threshold analysis, including bipole and tripole relationships, are conducted to uncover significant interactions between variables. Finally, data aggregation and visualization techniques are applied to summarize and present the processed data, setting the stage for in-depth analysis.

3.1 Bipole, tripole and multipole relationship

- Bipoles: It is instrumental in unravelling climate intricacies in urban hubs like Bengaluru, Bombay, and Hyderabad. Specifically, on crucial relationships, such as temperature and rainfall patterns in Bengaluru, aiding in drought and flood assessments. In Bombay, they uncover links between sea surface temperatures and monsoon activity, enhancing seasonal rainfall predictions. Similarly, in Hyderabad, bipoles elucidate the interplay between humidity and visibility, vital for managing heat waves and supporting public health initiatives. Leveraging bipoles allows for a nuanced understanding of climate dynamics, facilitating tailored resilience strategies tailored to each city's unique challenges.

- Tripoles: It serves as a sophisticated tool for delving into climate nuances across Bengaluru, Bombay, and Hyderabad. Particularly, they offer insights into essential correlations, such as temperature-precipitation dynamics in Bengaluru, crucial for understanding monsoon behaviours. In Bombay, tripoles reveal intricate connections between the density of clouds, visibility, and humidity, aiding in the prediction of extreme weather events. Meanwhile, in Hyderabad, they uncover relationships between temperature fluctuations and agricultural productivity, guiding strategies for drought resilience. Utilizing tripoles provides a holistic perspective on climate factors, enabling informed decision-making for urban sustainability.
- Multipoles: It provides a holistic framework for comprehensively understanding climate complexities in urban settings like Bengaluru, Bombay, and Hyderabad. They unveil interconnected dynamics, including wind direction, wind pattern, cloud and humidity offering insights crucial for sustainable development. In Bengaluru, multipoles illuminate interactions between temperature, rainfall, and humidity, informing water resource management strategies. In Bombay, they decipher patterns in sea surface temperatures, atmospheric pressure, and precipitation, enhancing disaster preparedness efforts. Similarly, in Hyderabad, multipoles reveal relationships between wind patterns, and wind direction guiding urban heat island mitigation strategies. Leveraging multipoles enables policymakers to develop resilient cities amidst evolving environmental challenges.

3.2 Bipole analysis using Pearson correlation

A bipole involves two time series variables, T_0 and T_1, and focuses on their correlation.

Let the correlation between T_0 and T_1 be denoted as $\mathrm{corr}(T_0, T_1) = a$

$$\sigma_\Delta = a \tag{5}$$

Here, the bipole strength is simply the correlation coefficient between the two variables.

3.3 Tripole analysis using Pearson correlation

A tripole involves three time series variables, T_0, T_1, and T_2. The relationships among these variables are examined through their pairwise correlations.

Let the pairwise correlation strengths be denoted as

$$\mathrm{corr}(T_0, T_1) = a, \mathrm{corr}(T_1, T_2) = b \text{ and } \mathrm{corr}(T_0, T_2) = c$$

In a tripole $\Delta = (T_0 : T_1, T_2)$, the jump δ_Δ of the tripole is given by:

$$\delta_\Delta = \frac{(a+b)^2}{2(1+c)} - a^2 \tag{6}$$

Using the formulae of Pearson correlation for the sum of x

$$\sigma_\Delta = \mathrm{corr}(T_0, T_{1+2}) = \frac{\mathrm{cov}(T_0, T_1) + \mathrm{cov}(T_0, T_2)}{\sqrt{\mathrm{var}(T_0) + \mathrm{var}(T_1) + 2\mathrm{cov}(T_1, T_2)}} \tag{7}$$

3.4 Quadpole analysis using Pearson correlation

A quadrupole involves four time series variables, T_0, T_1, T_2, and T_3. It extends the idea of tripoles to include an additional variable and examines the correlations among all pairs.

Let the pairwise correlation strengths be denoted as $\mathrm{corr}(T_0, T_1) = a, \mathrm{corr}(T_1, T_2) = b$, $\mathrm{corr}(T_2, T_3) = d$, and $\mathrm{corr}(T_0, T_3) = e$. Additionally, include $\mathrm{corr}(T_0, T_2) = c$ and $\mathrm{corr}(T_1, T_3) = f$

Proposition: The jump δ_Δ of the quadrupole can be derived similarly to the tripole, considering the additional complexity of the relationships among four variables.

$$\delta_\Delta = \frac{(a+b+d)^2}{3(1+c+f)} - a^2 - d^2 \tag{8}$$

Using the principles of Pearson correlation for multiple variables, the quadrupole strength can be expressed as:

$$\sigma_\Delta = \mathrm{corr}(T_0, T_{1+2+3}) = \frac{\sum \mathrm{cov}(T_0, T_i)}{\sqrt{\sum \mathrm{var}(T_i) + 2\sum \mathrm{cov}(T_i, T_j)}} \tag{9}$$

where the sums are taken over all relevant variables and their pairwise covariances.

3.5 Logarithmic spiral

The logarithmic spiral exhibits properties that make it suitable for capturing the self-similar and fractal-like structures found in time series data. Its non-linear nature helps to model intricate dependencies more accurately. It can efficiently handle large datasets. Its inherent scalability ensures that the analysis remains computationally feasible, even with extensive time series data. By transforming data into a logarithmic spiral representation, hidden patterns and correlations can be uncovered which are not apparent through conventional methods.

Naive Logarithmic Spiral Algorithm
Step 1 Initializing Parameters:
a : Scale factor, : Growth rate, $\theta_{\min}$: Starting angle and $\theta_{\max}$: Ending angle .
$\Delta\theta$: Step size for angle increment.
Step 2 Initialize an Empty List:
Create an empty list of points to store the (x, y) coordinates.
Step 3 Iterate Over the Range of Angles:
Start with $\theta = \theta_{\min}$.
Step 4 For Each θ from $\theta_{\min}$ to $\theta_{\max}$ with step size $\Delta\theta$:
• Compute the Radial Distance r:$r = a \cdot e^{b\cdot\theta}$ (10)
• Convert Polar Coordinates to Cartesian Coordinates:

$$x = r \cdot \cos(\theta) \tag{11}$$

$$y = r \cdot \sin(\theta) \tag{12}$$

Step 5 Store the Coordinates:
Append the point (x, y) to the list points.
Step 6 Increment the Angle:

$$\theta = \theta + \Delta\theta \tag{13}$$

Step 7 Return the List of Points:
Return the list points containing all the generated (x, y) coordinates where r is the radial distance, a is a scaling factor, b is a growth factor, and θ is the angle.
We use the correlation values as the growth factor b to generate the logarithmic spiral for each type of pole.

3.6 CONTRaComplete

This employs exhaustive algorithms to achieve high precision, making it ideal for the proposed work where accuracy is paramount, despite the higher computational cost.

Input: $\mathcal{D}$, Parameters: β, τ
Output: Non - redundant and complete set of tripoles, C.
Initializing $C \leftarrow \phi$
Select lowest value of $s_\delta \in [0, 1] s.t. \frac{s_\beta^2(1+s_\beta)}{(1-s_\beta)} \geq \beta$
$L \leftarrow$ finding candidate super pairs($\mathcal{D}, s_\beta$)
for each super pair $(T_A, T_B) \in L$
$P \leftarrow$ finding tripoles for super pair $(\mathcal{D}, (T_A, T_B), \beta)$
$C \leftarrow C \cup P$
end for
$C \leftarrow$ getting non - redundant tripoles (C, τ)
return C

3.7 CONTRaFast

Designed for rapid analysis, CONTRaFast utilizes approximations and heuristics to reduce computational overhead.

Here, is the algorithm for CONTRaFast

Input: $\mathcal{D}$,Parameters: s_β
Output: Set $Lofcandidatesuperpairs$
Initializing $L \leftarrow \phi$
if negative tripoles are to be only found then
$E \leftarrow$ All pairs of time series with correlation below 0 and magnitude $\geq$ s$_\beta$
else
$E \leftarrow$ All pairs of time series with correlation magnitude $\geq s_\beta$
end if
return L

3.8 Applications of multipoles through case studies

This section presents two case studies where the application of multipoles significantly enhances the analysis and yields superior results, offering deeper insights and more accurate interpretations.

3.8.1 Application on climate data. Bangalore, Hyderabad, and Mumbai are three major cities in India with distinct geographical locations and climates, leading to notable differences in their weather patterns and overall environmental characteristics.

3.8.1.1 Relationship based on climate and geography.

- Bangalore and Hyderabad: Similar inland locations on the Deccan Plateau result in moderate temperatures, distinct seasons, and reliance on monsoon for rainfall.
- Mumbai: Coastal location leads to higher humidity and more pronounced monsoon effects, contrasting with Bangalore and Hyderabad.

3.8.1.2 Relationship among parameters. Bangalore, Hyderabad, and Mumbai exhibit distinct weather patterns and relationships between various weather parameters.

Temperature:

- Bangalore: Mild climate due to high elevation, rarely reaching extreme temperatures.
- Hyderabad: Hot and dry, with summer temperatures often exceeding 40°C (104°F).

- Mumbai: Coastal climate with stable temperatures from warm to hot.

Humidity:

- Bangalore: Moderate humidity, especially during the monsoon.
- Hyderabad: High humidity levels during the monsoon months.
- Mumbai: High humidity throughout the year, particularly during the monsoon.

Precipitation:

- Bangalore: Significant rainfall during the monsoon and occasional showers year-round.
- Hyderabad: Monsoon-influenced rainfall pattern from June to September.
- Mumbai: Heavy monsoon rainfall contributing to lush greenery.

Wind Patterns:

- Bangalore: Moderate winds influenced by geography and surrounding hills.
- Hyderabad: Gusty winds, especially during summer thunderstorms.
- Mumbai: Strong coastal winds, particularly during the monsoon season.

Temperature and Humidity Relationship:

- Bangalore: Stable relationship due to moderate climate.
- Hyderabad: Strong correlation between temperature and humidity during hot and humid periods.
- Mumbai: High humidity during the monsoon affects temperature perception and comfort levels.

3.8.1.3 Regional wise correlation. To calculate and compare monthly correlations between variables across different cities.

df: Data frame containing the dataset with columns for date-time, variable A, variable B, and variable C.
M: Set of months $1, 2, \dots, 12$.
Y: The specific year for which data is being analyzed (e.g., 2019).
Variable A for humidity, B for visibility and C for cloud cover.
$\rho\varepsilon$ Pearson correlation coefficient .
Step 1: Data Initialization:
• Load the dataset into a data frame df.
• Define the set of months $M - 1, 2, \dots, 12$.
• Define the year Y.
Step 2: Data Preparation:
• Convert the 'date_time' column in df to a DateTime format.
• Initializing an empty list C to store the correlation values for each month.
Step 3: Monthly Data Filtering and Validation:
For each month $\in M$:
• Filter Data: Create a new DataFrame, df_m, containing rows from df where the year of the date_time column equals Y and the month equals m.
• Check Data Validity: Ensure df_m is not empty and contains sufficient data points in columns A, B, and C.
Step 4: Correlation Calculation:
For each valid month m:
• Extract Columns:

$$A_m - row[A] \| \ row \ \in \mathrm{df}_m$$
$$B_m - \mathrm{row}[B] \| \ row \ \in \mathrm{df}_m$$
$$C_m - row[C] \| \ row \ \in \mathrm{df}_m$$

• Calculate Correlations:

Compute (A_m, B_m) : Pearson correlation between A_m and B_m.
Compute (A_m, C_m) : Pearson correlation between A_m and C_m.
Step 6: Calculate Difference:

$$\Delta\rho_m - \rho(A_m, B_m) - \rho(A_m, C_m) \tag{13}$$

Step 7: Results

$$C \leftarrow C \cup \{\Delta\rho_m\} \tag{14}$$

This comprehensive analysis of climate data across Bangalore, Hyderabad, and Mumbai highlights how geographical differences influence weather patterns, providing valuable insights into regional climate variations and trends.

3.8.2 Application on blood pressure data.

3.8.2.1 Relationship among parameters. The model calculates correlations between age and BMI for males and females across different age intervals and then plots these correlations along with logarithmic spiral curves based on the correlation values.

The logarithmic spiral function generates points on a logarithmic spiral based on the input parameters t (angle) and *correlation* using the formula:

$$\boldsymbol{x = a \cdot e^{b \cdot t} \cdot \cos(t)} \tag{15}$$

$$\boldsymbol{y = a \cdot e^{b \cdot t} \cdot \sin(t)} \tag{16}$$

Where a is a scale factor controlling the size of the spiral, and b is derived from the correlation value.

Here's the mathematical algorithm for the programme:

Data Loading and Processing

- The programme loads data from a CSV file ('data.csv') containing age, sex, and BMI information.
- It defines age intervals (width of 10 years each) for analysis.

Correlation Calculation

- It calculates the correlation between age and BMI for males and females within each age interval.
- For each interval, it filters the data by age and sex (male or female) and computes the Pearson correlation coefficient.

This algorithm integrates data processing, and correlation analysis using logarithmic spirals to represent correlations between age and BMI for different genders and age groups.

4 Results and discussion

This section highlights the results of the relationships established between the different parameters in the form of two case studies

4.1 Dataset description

The dataset is suitable for a variety of training tasks of different difficulty levels, allowing instructors to match the complexity to their audience. It is compact enough to be quickly trained on a standard laptop, making it useful for hands-on sessions.

4.1.1 Climate dataset. The World Weather Online (WWO) API is a service that provides comprehensive weather data for various locations worldwide. It offers current conditions, forecasts, and historical data, making it a reliable source for accessing extensive weather information. The wwo_hist package is a Python tool that simplifies the retrieval of historical weather data using the WWO API. It automates API requests and organizes the returned data into a user-friendly format, making it ideal for researchers and developers needing extensive historical weather data for analysis.

- Time Frame:

Year: 2009
Frequency: Hourly data, providing a detailed view of weather conditions over ten years

- Cities Covered

The dataset includes historical weather data for several Indian cities, focusing on Bengaluru, Mumbai, and Hyderabad.

- Attributes and Columns

The dataset contains several key weather parameters for each hour, including:

- Date and Time: The specific hour of the observation.
- Temperature (TemperatureC): in degrees Celsius.
- Humidity (Humidity%): Relative humidity as a percentage.
- Wind (WindSpeedKmph and WindDirDegree): Wind speed in kilometres per hour and Wind direction in degrees.
- Precipitation (PrecipMM): in millimetres.
- Pressure (PressureMB): Atmospheric pressure in millibars.
- Visibility (VisibilityKM): in kilometres.
- Cloud Cover (CloudCover%): Percentage of cloud cover.
- Heat Index (HeatIndexC): Perceived temperature considering humidity.
- Dew Point (DewPointC): Temperature at which dew forms.
- Weather Description (WeatherDesc): Textual description of weather conditions.

4.1.2 Blood pressure data. The blood pressure dataset, provided by Pavan Bodanki on Kaggle, originates from the National Health and Nutrition Examination Survey (NHANES) conducted by the Centers for Disease Control and Prevention (CDC). NHANES combines clinical measurements and self-reported information. Blood pressure readings, including systolic and diastolic pressures, are taken using standard medical equipment like sphygmomanometers, ensuring accuracy and consistency.

4.2 Discussion

This study aims to analyze the relationships between climate variables and blood pressure (BP) factors using a multipoles approach. By examining the interplay between environmental conditions and BP, the research seeks to identify potential correlations and dependencies that could inform public health interventions and climate adaptation strategies. The selected climate variables include wind speed, wind direction, humidity, visibility, and cloud cover. For BP factors, the study considers Body Mass Index (BMI), salt intake, alcohol consumption, and BP abnormalities.

Climate Data and Blood Pressure Analysis Using Bipoles, Tripoles, and Multipoles

- Bipole Analysis: Bipoles are utilized to explore pairwise correlations between climate variables and BP factors. For example, the relationship between humidity and systolic BP can be analyzed to understand how atmospheric moisture affects blood pressure levels.
- Tripole Analysis: Tripoles extend the analysis to three variables, providing insights into more complex interactions. Analyzing the interaction between wind speed, humidity, and diastolic BP can help identify how combined climatic conditions influence blood pressure.
- Multipole Analysis: Multipoles involve examining multiple variables simultaneously to capture the intricate dependencies between climate conditions and BP factors. This comprehensive approach can reveal how a combination of wind speed, wind direction, humidity, visibility, and cloud cover collectively impact BP and related health outcomes. The multipole approach to analyzing climate data and blood pressure provides a nuanced understanding of the relationships between environmental conditions and BP levels.

By leveraging bipoles, tripoles, and multipoles, researchers can uncover complex interactions and dependencies that inform both public health and climate adaptation strategies. This methodology is particularly valuable for identifying the multifaceted influences on BP, thereby aiding in the development of comprehensive and targeted health interventions.

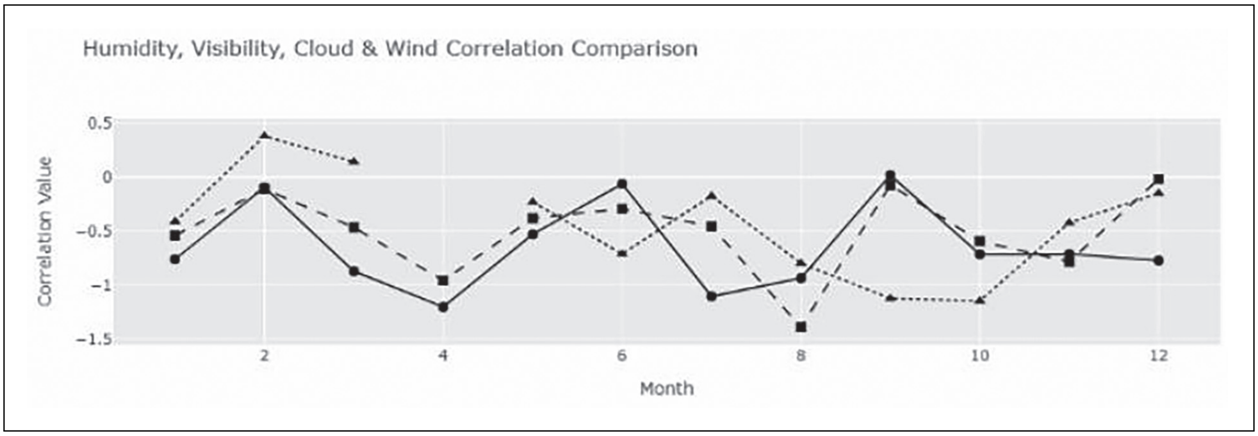

Figure 2. Humidity and visibility correlation analysis.

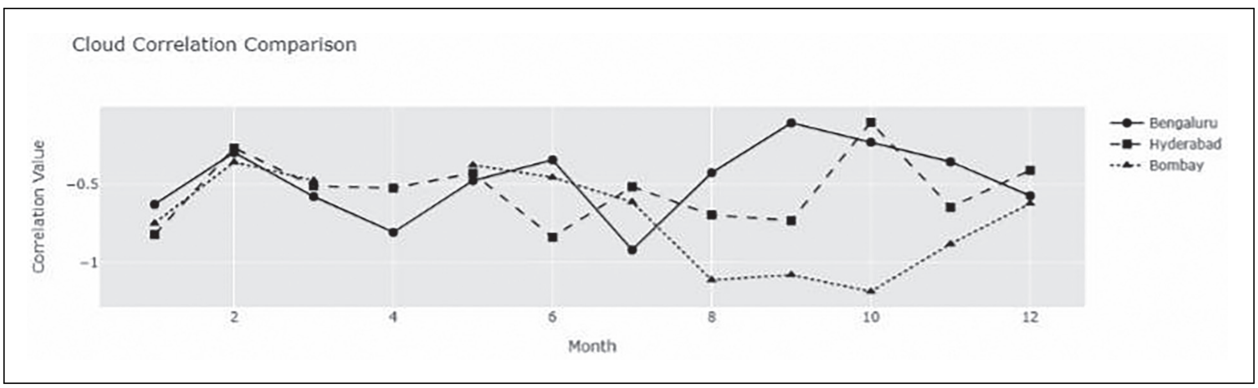

Figure 3. Cloud correlation analysis based on regions.

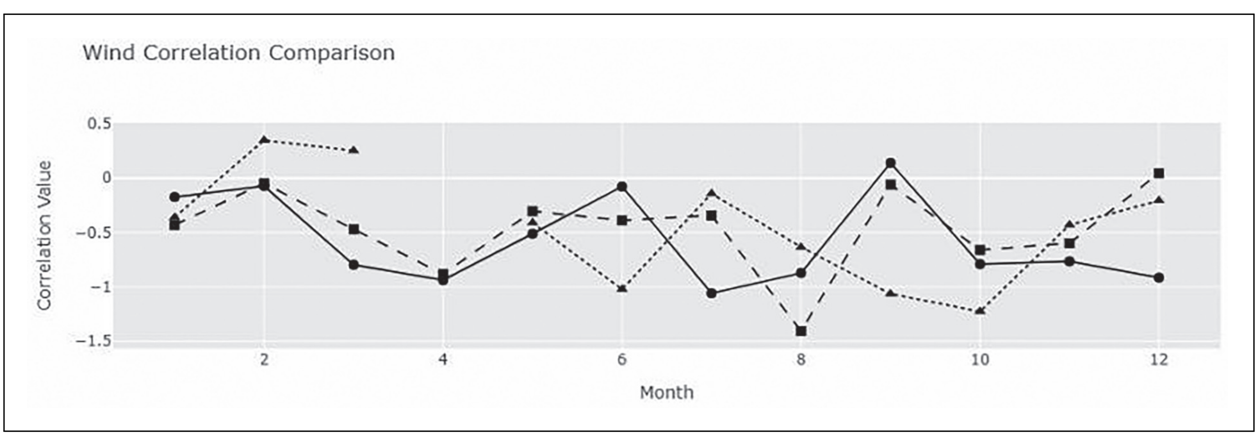

Figure 4. Wind correlation analysis based on regions.

4.3 Output and results

The following figures depict the observations drawn to investigate the relationship between weather parameters and blood pressure levels in individuals residing in Indian cities. To establish an understanding of how the variations in temperature, humidity, and other weather factors may impact blood pressure readings.

Case Study I: Climate Data

The following figures depict the observations drawn from monthly data analysis conducted on:

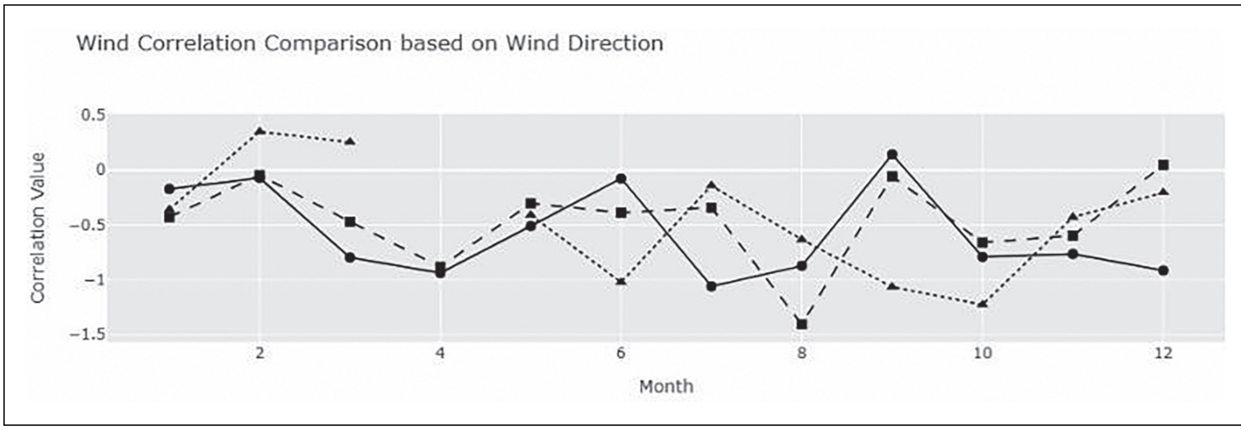

Figure 5. Wind correlation analysis based on regions.

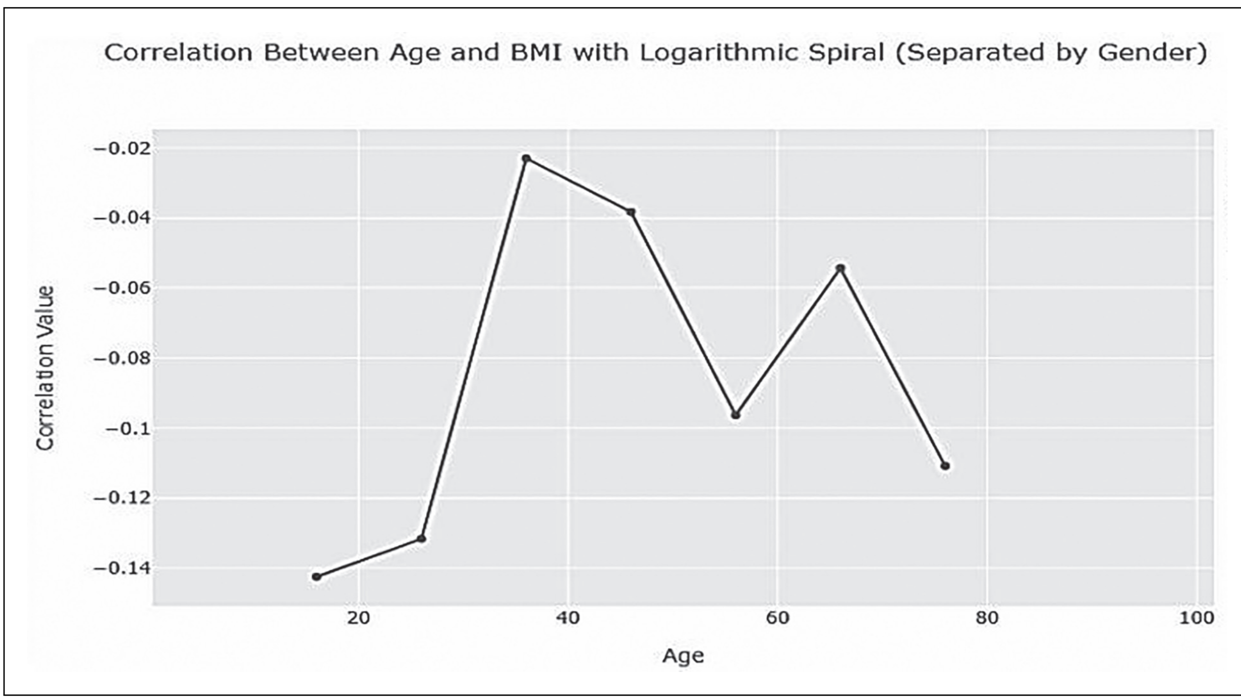

Figure 6. Blood pressure correlation analysis on BMI –men.

Bangalore(), Bombay() and Hyderabad().

Figure 1 brings out a clear indication that as humidity levels rise, visibility tends to decrease. Figure 2 illustrates the city's cloud cover variation during different seasons, influencing visibility conditions. In summer, Hyderabad and Mumbai have high humidity and cloud coverage, reducing visibility, while Bangalore's lower humidity results in better visibility despite occasional thunderstorms. In winter, Hyderabad and Mumbai experience lower humidity and improved visibility. Bangalore maintains good visibility with moderate cloud coverage, occasionally impacted by morning mist or fog. Overall, winter offers clearer skies and better visibility in these cities compared to summer.

Figure 3 explains that in summer, Hyderabad and Mumbai experience higher wind speeds that lead to cloud formation and reduced visibility, while Bangalore has lighter winds with occasional gusts. In winter, Hyderabad's calmer winds result in clearer skies and better visibility, Mumbai's coastal winds cause occasional haze or fog, and Bangalore's light to moderate winds disperse clouds, improving visibility.

In Figure 4 it is observed that in summer, Hyderabad and Mumbai experience higher wind speeds and humidity, reducing visibility, with wind direction influencing cloud formation. Bangalore has lighter winds but occasional gusts affect

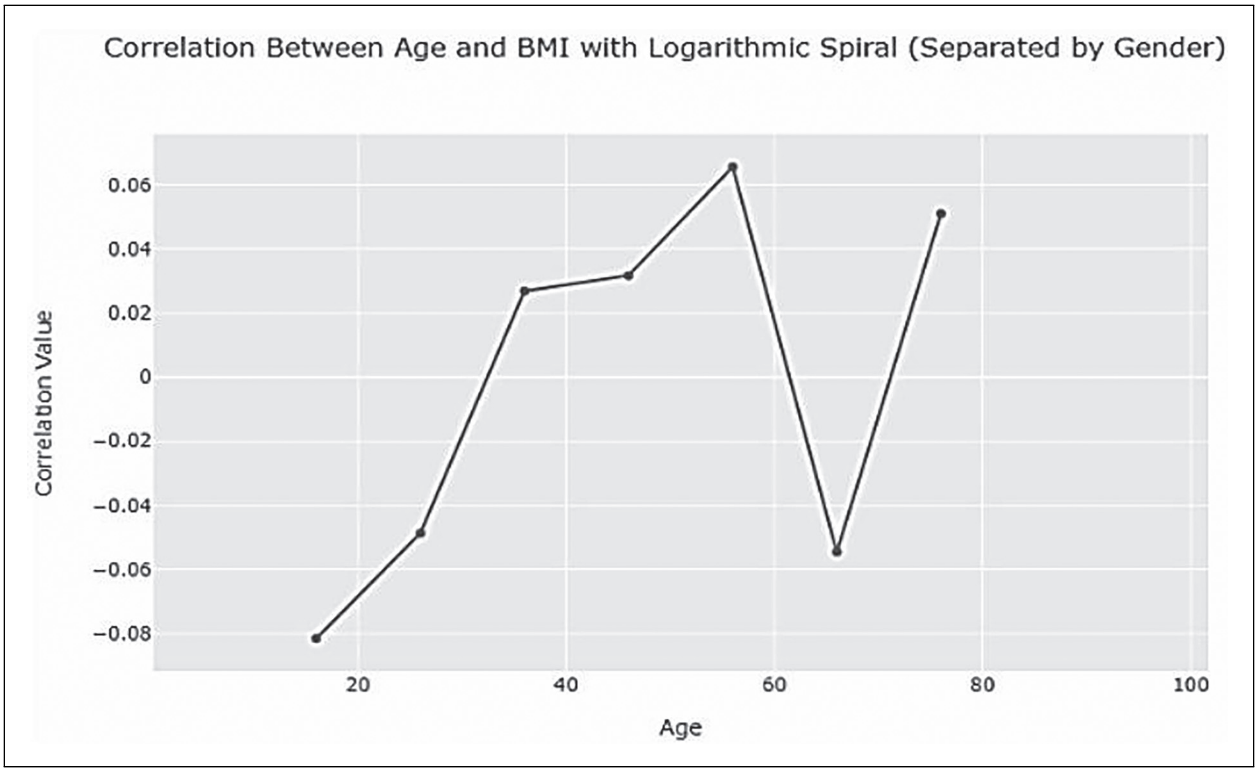

Figure 7. Blood pressure correlation analysis on BMI- female.

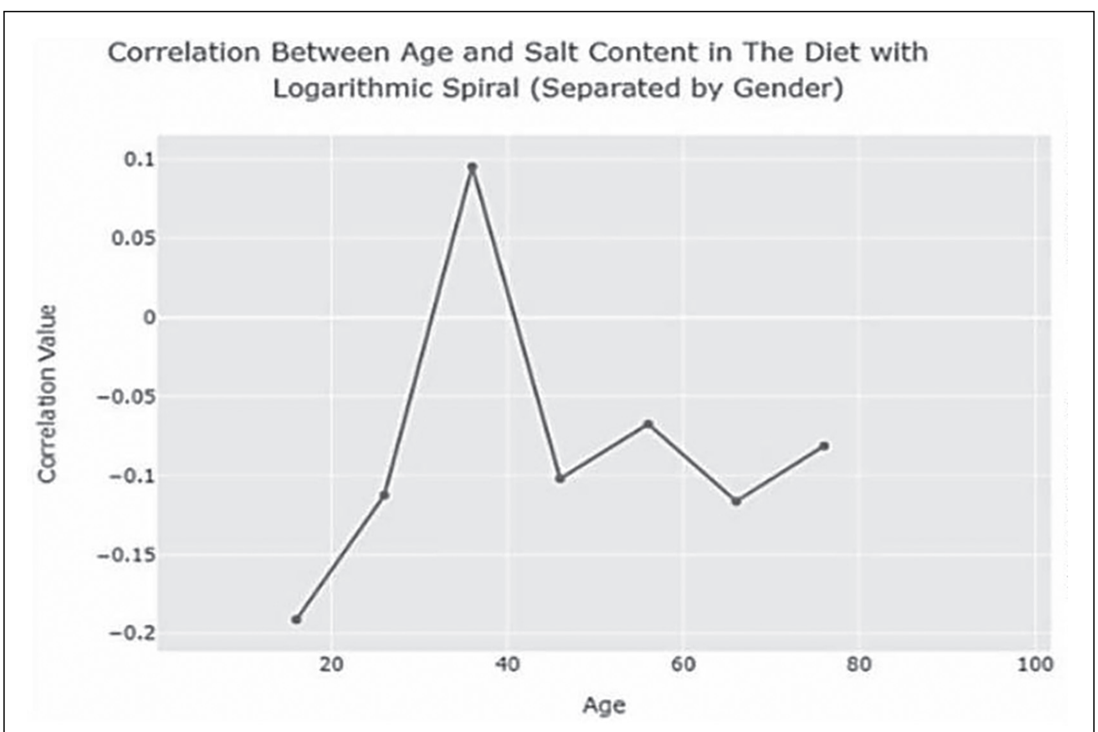

Figure 8. Blood pressure correlation analysis on salt content- men.

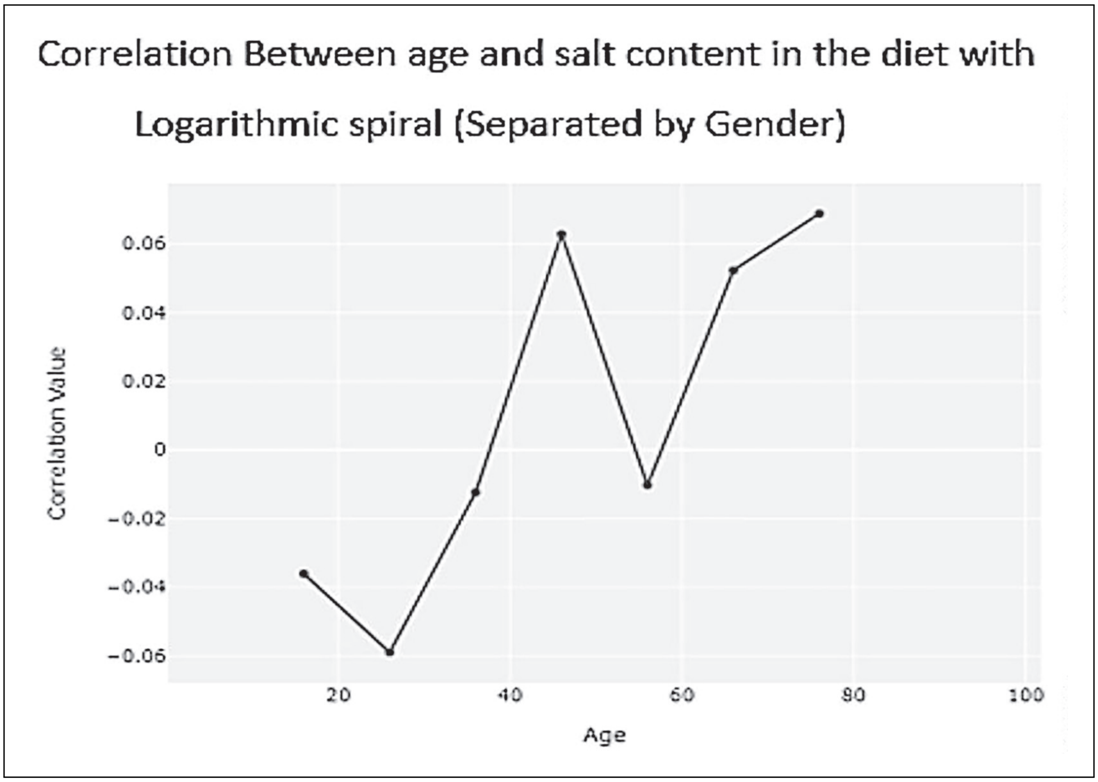

Figure 9. Blood pressure correlation analysis on salt content (female).

visibility. In winter, calmer winds and northerly directions in Hyderabad and Bangalore improve visibility, while Mumbai's moderate coastal winds can cause occasional haze or fog. Understanding wind patterns and speeds helps interpret weather conditions and visibility trends in each city across seasons.

Figure 5 exemplifies that in summer, Hyderabad and Mumbai experience higher wind speeds and humidity, affecting visibility due to cloud formation, while Bangalore's lighter winds occasionally impact local weather. In winter, Hyderabad and Bangalore have calmer winds and clearer skies, improving visibility, whereas Mumbai's coastal winds can cause occasional haze or fog. Wind direction and speed significantly influence weather conditions and visibility trends in these cities throughout the year.

Case Study II: Blood Pressure Data

The following figures depict the observations of how changes in the various health-related parameters affect blood pressure readings and overall health.

In Figure 6 and Figure 7, Correlation with BMI male and Female (▬). The graph illustrates the relationship between age and BMI for males and females separately. Each line represents the average BMI for different age groups across a range of ages.

Figure 8 and Figure 9 correlation with salt content in the diet, Male and Female (▬). It provides insights into how salt content in diet varies across different age groups. we observe either a steady increase, a plateau, or even a decrease in salt intake with age. Higher values indicate higher salt consumption, potentially increasing the risk of health issues related to high salt intake.

The graphs in Figures 10 and 11 show the Correlation with Alcohol consumption per day, Male and Female (▬), the relationship between gender and alcohol consumption per day. there may be variability in alcohol consumption within each gender category. Some individuals may consume significantly more or less alcohol than the average for their gender group.

Figure 12 and Figure 13 illustrate the Correlation with Blood Pressure Abnormality, Male and Female(▬), the relationship between gender and the prevalence of blood abnormalities. Some individuals may have a higher or lower prevalence of blood abnormalities than the average for their gender group. Higher values indicate a higher prevalence of blood abnormalities.

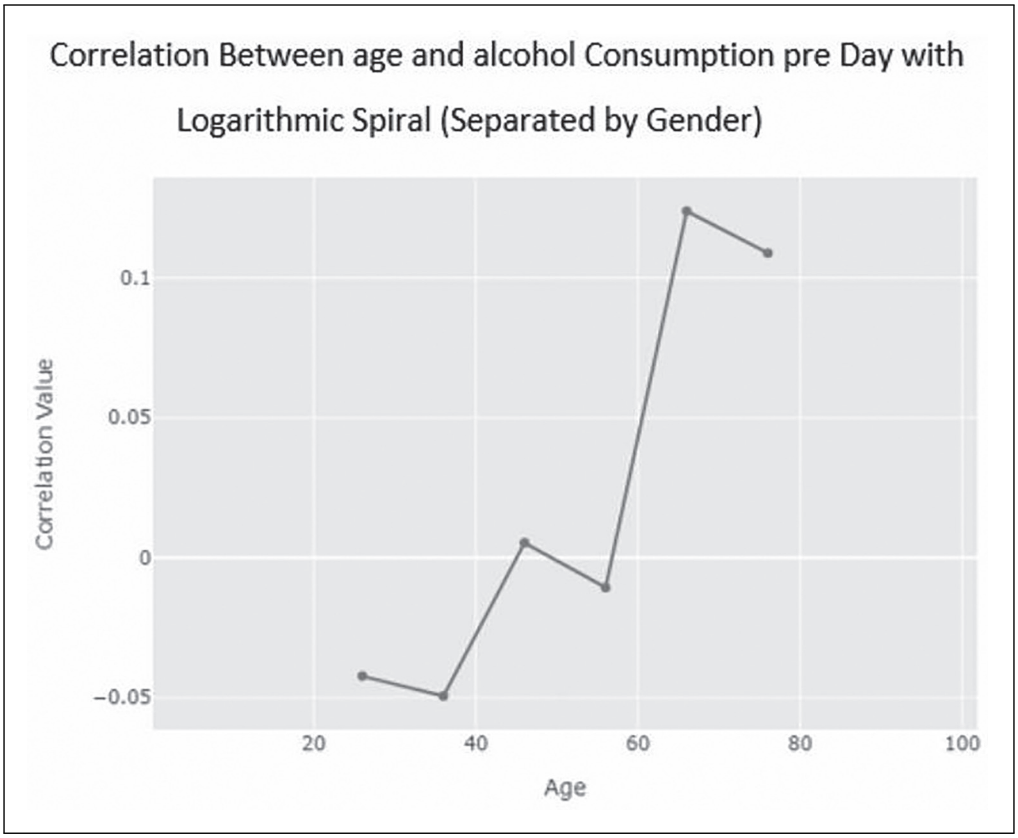

Figure 10. Correlation analysis on alcohol consumption /day (men).

Understanding these differences is crucial for identifying age-specific dietary patterns, gender-specific patterns of alcohol use, and blood abnormalities to develop targeted interventions to promote healthier eating habits, reduce the risk of health issues associated with excessive salt intake promote responsible drinking behaviours, reduce the risk of alcohol-related harm and to prevent, diagnose, and manage blood-related health conditions in both males and females. Figure 14 shows the comparison of the computational efficiency of CONTRaFast(■) with CONTRaComplete(■).

Based on the observations CONTRaFast takes less amount of computational time, compared to CONTRaComplete.

In Figure 15, we see CONTRa (Fast and Complete) for climate data correlation Analysis with the values obtained as results.

In Figure 16, the impact of the threshold for climate data correlation analysis with the values is obtained as a result.

While correlation values fluctuate with varying threshold values, this variation is expected and does not undermine the model's robustness. The overall correlation patterns remain consistent across different thresholds (K values of −0.5, −0.4, −0.3) for all cities (Bangalore, Hyderabad, and Bombay), indicating that the method effectively uncovers underlying weather dynamics. These fluctuations simply reflect the model's natural sensitivity to parameter changes without indicating significant shifts in detected patterns. Thus, the reliability of the method in detecting weather dynamics is evident, allowing domain experts to confidently apply the approach, knowing that threshold variations do not critically impact the core findings. In Figure 17 the time complexity on the threshold values of Bengaluru, Hyderabad and Bombay can be observed with variations with respect to computational time calculated in seconds.

Cost of CONTRa for finding multipoles in Climate Dataset: Time complexity remains unaffected by the change of the threshold k. Table 1, Table 2 and Table 3 show the time complexity (seconds) and overall correlation against every value of k, for every city i.e., Bengaluru, Bombay, and Hyderabad.

5 Conclusion

The conclusion of this project highlights the significant findings from the correlation analysis of climate data and blood pressure data, illustrating the intricate relationships between various environmental and lifestyle factors and their respective impacts. In climate data analysis, the study uncovered critical links between humidity, visibility, cloud cover, wind

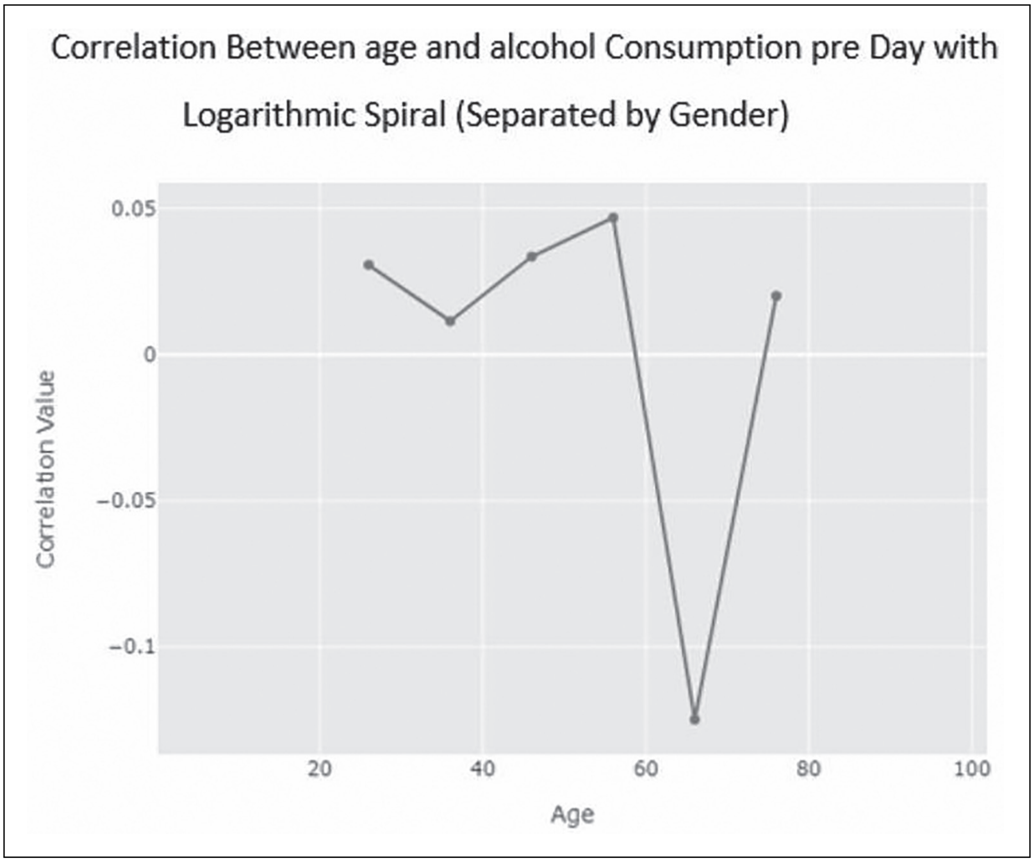

Figure 11. Correlation analysis based on alcohol consumption/day (female).

Table 1. Contrafast time complexity for Bengaluru.

K	Time Complexity (in seconds)	Overall Correlation
−0.5	0.064122915	−0.012798239
−0.4	0.064029932	−0.043849621
−0.3	0.064013958	−0.074901004
−0.2	0.063063145	−0.105952386
−0.1	0.06301403	−0.137003769
0	0.063014269	−0.168055152
0.1	0.064116001	−0.199106534
0.2	0.064014435	−0.230157917
0.3	0.064015865	−0.261209299
0.4	0.063014507	−0.292260682
0.5	0.063015223	−0.323312065

speed, and regional differences, which are essential for understanding weather patterns and assessing climate change impacts. The CONTRa algorithm proved efficient in analysing these complex datasets, offering valuable insights into environmental influences on local weather conditions. Similarly, the analysis of blood pressure data revealed substantial correlations between age, gender, BMI, dietary habits, and blood pressure levels, emphasizing the multifaceted nature of cardiovascular health. The study identified critical factors like salt intake and alcohol consumption that significantly impact blood pressure, underscoring the need for targeted dietary and lifestyle interventions to prevent hypertension and related health issues. These findings provide a comprehensive understanding of the demographic and lifestyle influences on cardiovascular health, offering a foundation for preventive strategies and health interventions. The insights gained can help policymakers and health practitioners address the combined impacts of climate and lifestyle factors on blood pressure, ultimately promoting better health outcomes and resilience to climatic changes.

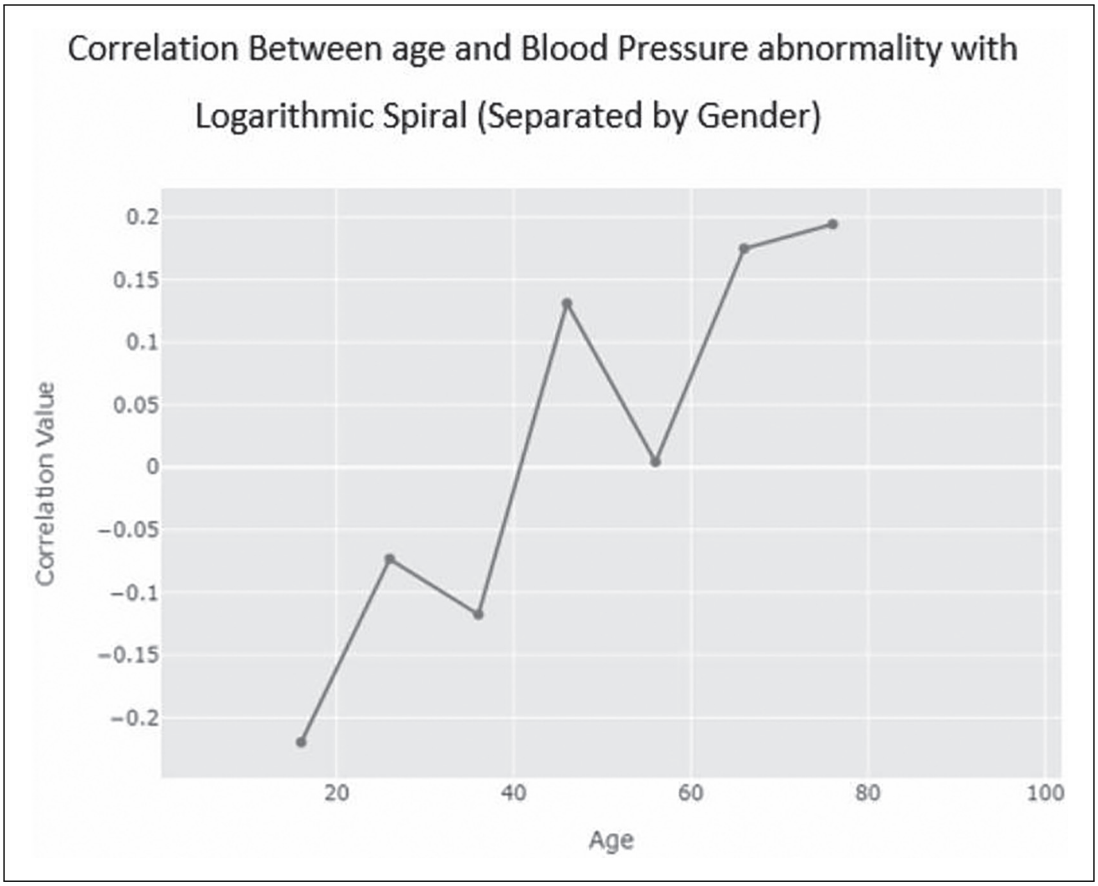

Figure 12. Blood pressure correlation analysis on BP abnormality – men.

Table 2. Contrafast time complexity for Hyderabad.

K	Time Complexity (seconds)	Overall Correlation
−0.5	0.064017057	−0.051291277
−0.4	0.064032793	−0.083940636
−0.3	0.06401515	−0.116589995
−0.2	0.063050747	−0.149239354
−0.1	0.062014341	−0.181888713
0	0.065016747	−0.214538073
0.1	0.063014984	−0.247187432
0.2	0.063014269	−0.279836791
0.3	0.076015472	−0.31248615
0.4	0.06301403	−0.345135509
0.5	0.063014507	−0.377784868

Domain-Specific Explanation for Multipolar Components:

Climate Science: When analyzing the interaction between wind direction, cloud cover, and humidity, we observe that changes in wind patterns directly influence cloud formation and humidity levels. For example, increased wind speeds from specific directions can bring moisture-laden air, leading to higher humidity and denser cloud cover. These multipolar relationships help predict extreme weather conditions such as storms or low-visibility periods, making them actionable for meteorologists and environmental planners who rely on these insights to improve forecasting accuracy and disaster preparedness.

Healthcare (Blood Pressure): In the context of blood pressure analysis, the interaction between age, BMI, and salt intake shows how these factors collectively impact cardiovascular health. Older individuals with higher BMI and greater

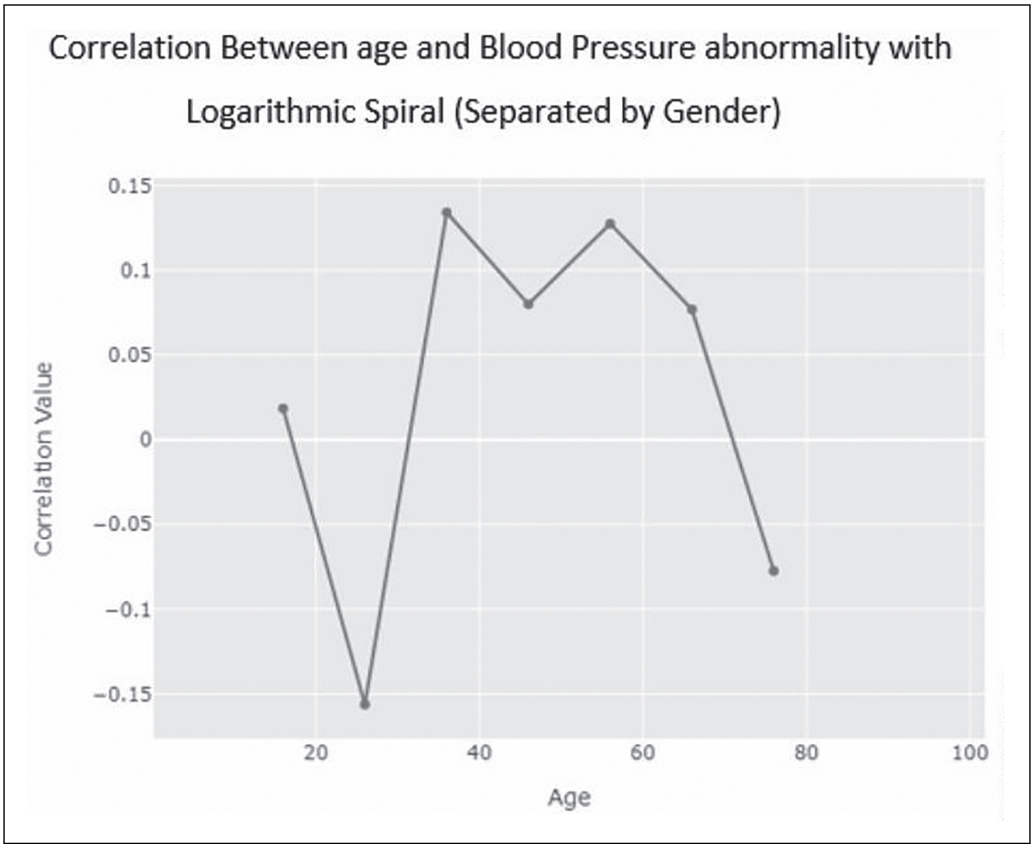

Figure 13. Blood pressure correlation analysis on BP abnormality – men.

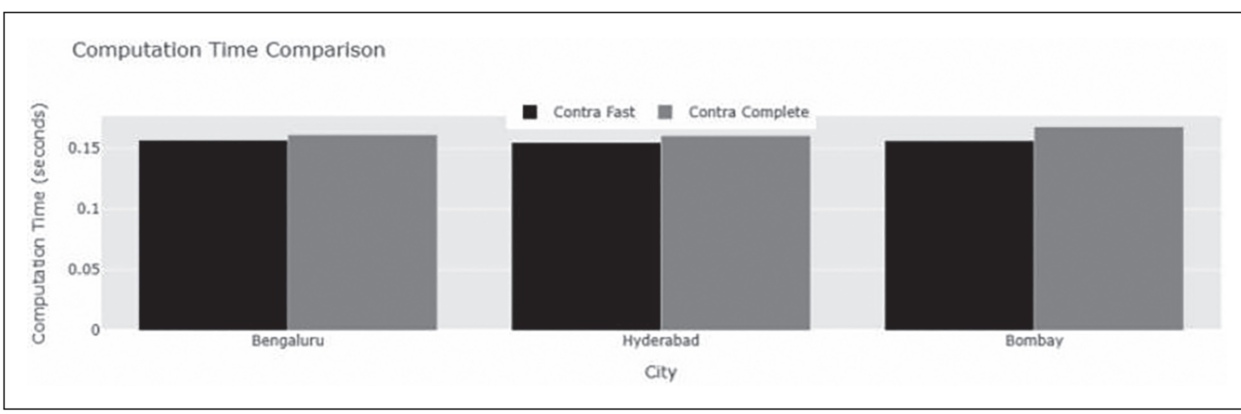

Figure 14. CONTRa complete and CONTRa fast time complexity based on clime data.

salt intake are at increased risk of hypertension. By understanding these multipolar relationships, healthcare professionals can better assess patient risk profiles and develop targeted interventions, such as dietary recommendations or lifestyle modifications, to prevent and manage high blood pressure more effectively.

Influence of the factors like demographic, environmental and lifestyle related characteristics on blood pressure are not considered under the present study. These investigations can become a part of the further work. Integrating climate data with health datasets could reveal patterns in disease prevalence and effectiveness of public health interventions. Leveraging advanced machine learning algorithms and interdisciplinary approaches will enhance the ability to extract meaningful patterns from large datasets, driving innovation in precision medicine and public health, ultimately improving population health and well-being.

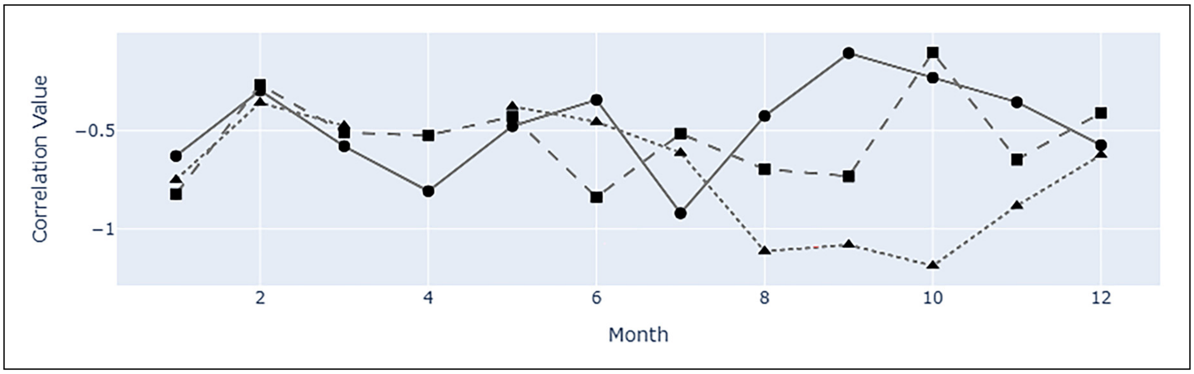

Figure 15. Impact of CONTRa (fast and complete) for climate data correlation analysis.

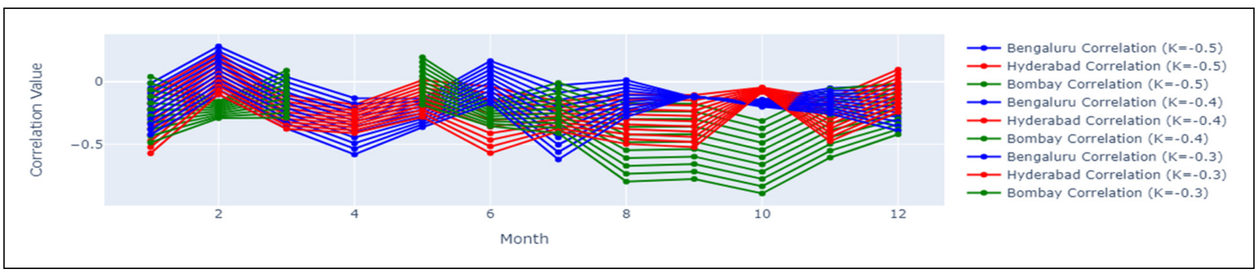

Figure 16. Impact of threshold correlation for climate data correlation analysis.

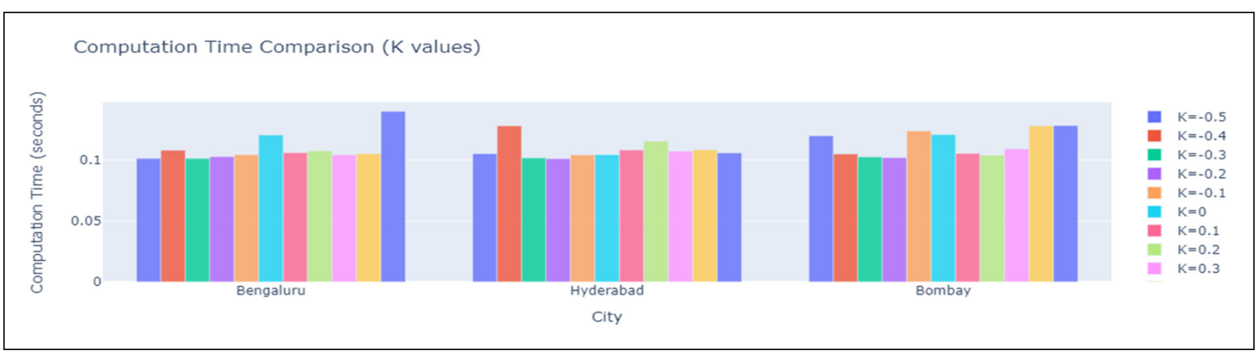

Figure 17. Time complexity on a threshold value.

Table 3. Contrafast time complexity for Bombay.

K	Time Complexity (seconds)	Overall Correlation
−0.5	0.063055038	−0.066363134
−0.4	0.064014435	−0.109925986
−0.3	0.062088251	−0.153488837
−0.2	0.062023878	−0.197051689
−0.1	0.063014269	−0.240614541
0	0.06301403	−0.284177393
0.1	0.063014269	−0.327740244
0.2	0.064014673	−0.371303096
0.3	0.063014269	−0.414865948
0.4	0.063014746	−0.4584288
0.5	0.063106537	−0.501991651

ORCID iD

Pratistha Gaur https://orcid.org/0009-0005-2126-0471
Pavan Kumar Vadla https://orcid.org/0009-0002-9861-2218
P Sudarshan https://orcid.org/0009-0002-9793-7643
Prrathyush Kedeelaya https://orcid.org/0009-0008-6353-7178
C Narendra Babu https://orcid.org/0000-0002-2358-3443

Statements and declarations

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

1. Xiong H, Shekhar S, Tan P-N, et al. Exploiting a support-based upper bound of Pearson's correlation coefficient for efficiently identifying strongly correlated pairs. In: *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004, pp.334–343. ACM.
2. Kawale J, Liess S, Kumar A, et al. A graph-based approach to find teleconnections in climate data. *Statist Anal Data Mining, ASA Data Sci J* 2013; 6(3): 158–179.
3. Kawale J, Steinbach M and Kumar V. Data Guided Discovery of Dynamic Climate Dipoles. *CIDU* 2011.
4. Dallakyan A, Rakheon K and Mohsen P. Time series graphical lasso and sparse VAR estimation. *CSDA* 2022; 176: 107557.
5. Cui Caihao and Wang Dianhui. High dimensional data regression using Lasso model and neural networks with random weights. *Information Sciences* 2016; 372: 505–517.
6. Cui C and Dianhui W. High dimensional data regression using Lasso model and neural networks with random weights. *Information Sciences* 2016; 372: 505–517.
7. Zuckerberg B, Strong C, LaMontagne JM, et al. Climate dipoles as continental drivers of plant and animal populations. *TREE* 2020; 35(5): 440–453.
8. Kawale J, Michael S and Vipin K. Discovering dynamic dipoles in climate data. In Proceedings of the 2011 SIAM international conference on data mining. Society for Industrial and Applied Mathematics, 2011.
9. Chatterjee S, Steinhaeuser K, Banerjee A, et al. Sparse group lasso: Consistency and climate applications. Proceedings of the 2012 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2012.
10. Bordoloi R, Réda C, Trautmann O, et al. Multivariate Functional Linear Discriminant Analysis for the Classification of Short Time Series with Missing Data. arXiv preprint arXiv:2402.13103 (2024).
11. Agrawal Saurabh, et al. Tripoles: A new class of relationships in time series data. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2017.
12. Arnold Andrew, Liu Yan and Abe Naoki. Temporal causal modeling with graphical granger methods. Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. 2007.
13. Atluri G, Steinbach M, Lim KO, et al. Discovering groups of time series with similar behavior in multiple small intervals of time. Proceedings of the 2014 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2014.
14. Baragona Roberto and Battaglia Francesco. Outliers detection in multivariate time series by independent component analysis. *Neural computation* 2007; 19(7): 1962–1984.
15. Wei William WS. *Multivariate time series analysis and applications*. John Wiley & Sons, 2019.
16. Chiu B, Eamonn K and Stefano L. Probabilistic discovery of time series motifs. Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. 2003.
17. Eichler M. Causal inference with multiple time series: principles and problems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 2013; 371(1997): 20110613.
18. Fu Tak-chung. A review on time series data mining. *Engineering Applications of Artificial Intelligence* 2011; 24(1): 164–181.
19. Keogh E, Chu S, Hart D, et al. Segmenting time series: A survey and novel approach. *Data Mining in Time Series Databases* 2004; 1–21.
20. Keogh E, Jessica L and Hot Ada F. sax: Efficiently finding the most unusual time series subsequence. Fifth IEEE International Conference on Data Mining (ICDM'05). Ieee, 2005.