
**Electronic Journal of Applied Statistical Analysis**
**EJASA, Electron. J. App. Stat. Anal.**
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v17n3p764

# Compositional Principal Component Analysis on Air Concentrations data from Kodungaiyur region of Chennai city

By

15 December 2024

*Electronic Journal of Applied Statistical Analysis*
Vol. 17, Issue 03, December 2024, 764-780
DOI: 10.1285/i20705948v17n3p764

# Compositional Principal Component Analysis on Air Concentrations data from Kodungaiyur region of Chennai city

Kalvinithi D[a], M R Sindhumol[b], Aswany B Lal[c], and Menini T. *[d]

[a] *University of Madras, Chennai ,*
[b] *University of Madras, Chennai ,*
[c] *Mahatma Gandhi University, Kottayam,*
[d] *University of Naples, L'Orientale ,*

15 December 2024

This study applies Compositional Principal Component Analysis (CoDA PCA) to analyze air pollutant concentrations in the Kodungaiyur region of Chennai, focusing on seven key pollutants: PM10, PM2.5, NO2, SO2, NH3, O3, and CO. Compositional transformations—Centered Log-Ratio (CLR), Isometric Log-Ratio (ILR), and Additive Log-Ratio (ALR)—are employed to address the inherent constraints of compositional data, ensuring the sum of the proportions remains constant. The PCA results reveal that the first two principal components explain approximately 76% of the variance, providing valuable insights into the relationships between pollutants and their potential sources. The CLR, ILR, and ALR biplot highlight different pollutant groupings, suggesting distinct emission sources and environmental impacts. This study underscores the importance of using compositional data techniques in air quality research and offers a detailed understanding of pollutant dynamics in Chennai, India, aiding in more targeted pollution control strategies.

**keywords:** Air pollutants concentrations, compositional data, log-ratios, principal component analysis

*Corresponding authors: menini@unior.it


ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

# 1 Introduction

Air pollution poses a significant threat to both human health and the environment. Addressing air pollution as a science-policy initiative is crucial for mitigating its adverse effects on health. Concerns surrounding nitrogen pollution, particularly nitrogen oxides in the air and ammonia nitrogen in water, highlight the need for effective bioremediation strategies using microalgae. Additionally, studies on air quality and thermal comfort in kitchen environments demonstrate the use of computational fluid dynamics to assess and improve air pollution levels.

Changing climate conditions may impact future air pollution control strategies, emphasizing the importance of considering emission reduction measures within the context of climate change. A novel approach to air filtration involves using polypropylene nonwoven materials loaded with cerium-doped manganese oxides submicron particles for ozone decomposition. Strategies for mitigating air pollution can also include solar chimneys, which have the potential to eliminate pollutants, particularly particulate matter, on a large scale.

Recent advancements in using metal-organic framework-derived nanostructures for the removal of volatile organic compounds address the industrially released gases and anthropogenic particles contributing to air pollution. Environmental sustainability challenges posed by the COVID-19 pandemic necessitate the development of suitable implementation strategies to ensure sustainability during this period. Moreover, creating a toxic-free environment is essential for protecting human health and sustaining the health of the planet for future generations.

In conclusion, there is an urgent need to address air pollution as a threat to human health and the environment through innovative bioremediation strategies, advanced air quality monitoring techniques, and sustainable pollution control measures. The literature underscores the interconnectedness of air pollution with climate change, industrial activities, and global health crises, highlighting the importance of collaborative efforts to mitigate the impacts on both human populations and ecosystems.

Principal component analysis (PCA) has been widely used to analyze air pollution and its associated factors. Probabilistic PCA has been utilized to aggregate PAH concentrations in residential locations during wildfires. Investigations into the relationship between technological progression and ambient air pollution have led to the development of separate indicators for both aspects. PCA-GLM analysis has explored the most significant air pollutants and gestational periods in relation to preterm birth risk in rural populations.

The effectiveness of air pollution control policies emphasizes the importance of integrating evaluation models and deep learning forecast methods. Assessments of air quality improvement strategies consider various factors, including meteorology, pollutant emissions, and energy structure. The impact of lockdown measures on air quality has been observed, revealing fluctuations in pollutant concentrations during different phases. Changes in air pollution levels during the pandemic underscore the importance of monitoring air quality during such events.

In a different context, spatial data analysis using moss biomonitoring showcases diverse

applications of environmental monitoring techniques. Studies conducted in critically polluted industrial areas employ multivariate statistical methods to assess air quality and propose strategies for future sustainability.

### 1.1 Compositional nature of air pollutants concentration

Jarauta et al (2016) explains that the concentrations of air pollutants are inherently classified as compositional data, meaning that they are represented as proportions or parts of a whole in a designated air volume or mass (Gallo et al., 2019). This characteristic is crucial for understanding the interplay between different pollutants and their impacts on air quality. The compositional nature of air pollution data signifies that the presence of each pollutant is not independent; rather, the concentration of one pollutant affects and is affected by the concentrations of others. This interconnectedness is essential for accurate modeling and analysis of air quality and pollution dynamics. Monitoring and interpreting pollution data as compositional data poses challenges, as traditional statistical methods may not be suitable. Specialized techniques are required to account for the interdependencies among various pollutant concentrations while ensuring accurate assessment of air quality. Also the detrimental effects of air pollution are widely acknowledged; effectively combating this issue necessitates a deep understanding of the specific pollutants at play, their sources, and their evolving trends. This knowledge gap underscores the critical need for robust statistical research on air pollutant data. Compositional Principal Component Analysis (CoDA PCA) holds the competency to critically analyze compositional data, a type of data where the relative proportions of components are more important than their absolute values. This makes it particularly well-suited for fields like air quality research, where understanding the interplay between various pollutants is crucial. Unlike traditional PCA, CoDA PCA doesn't treat each pollutant concentration independently (Gallo e Simonacci, 2013). Instead, it considers the closed-sum nature of the data, where all pollutant measurements add up to a constant value (often 100%). This allows CoDA PCA to identify critical combinations of pollutants that contribute the most to air quality variations. By critically analyzing these relationships, researchers can move beyond studying individual pollutant concentrations and gain a deeper understanding of the complex dynamics at play. This enriched knowledge can then be used to develop more targeted and effective strategies for air pollution control. The paper is organized into three sections, in which second section deals with the methodology adopted for the analysis, the third section focuses on analysis of the data and interpretation of the findings and the last section marks the discussion and concluding remarks.

## 2 Methodology

### 2.1 About the Study location and data set

Kodungaiyur, situated in Chennai, Tamil Nadu, is significant for air quality studies due to its urban classification, industrial activities, and waste management practices. These

factors contribute to complex air pollution dynamics, making it a vital location for monitoring air quality. As shown in figure 1, Kodungaiyur is home to one of Chennai's major municipal solid waste dumping grounds. The area has been utilized for waste disposal since 1986, with a current dumping rate of approximately 1,800 tons per day. From the studies conducted by Karthikeyan et al. (2011) and Vijiyalakshmi et al. (2017), it can be understood that, the open burning of waste at this site releases a variety of pollutants, including particulate matter (PM), volatile organic compounds (VOCs), and the hazardous gases like sulfur dioxide (SO2) and nitrogen oxides (NOx) into the atmosphere. The report prepared by Chennai Metropolitan Water Supply and Sewerage Board, 2016 highlights the region's proximity to industrial zones and heavy traffic contributes to elevated levels of air pollutants. Emissions from vehicles, construction activities, and industrial operations are primary sources of air pollution in Kodungaiyur, exacerbating the existing environmental issues. Research by Mariselvam, et al (2019) indicates that the air quality in Kodungaiyur has direct health impacts on the local population. Studies have shown a correlation between high levels of PM2.5 and increased incidences of respiratory and eye illness among residents. This underscores the importance of monitoring air quality as a means to protect public health. The Central Pollution Control Board (CPCB) monitors air quality in Kodungaiyur, providing crucial data that can be used to analyze trends in pollution levels over time. Sangeetha P (2022) shows that this information is vital for policymakers and researchers aiming to identify pollution sources and evaluate the effectiveness of air quality management strategies in Chennai. Furthermore, Suresh et al (2023) highlights the upsurge of PM10 exceedance as time elapses, hence making the region more prone to particulate matter emissions. For the following parameters, PM2.5 (particulate matter less than 2.5 micrometers in diameter), PM10 (particulate matter less than 10 micrometers in diameter), NO (Nitric oxide), NO2 (Nitrogen dioxide), NOx (Oxides of Nitrogen), NH3 (ammonia), SO2 (sulfur dioxide), CO (carbon monoxide), Ozone, Benzene and Toluene, the 24-hour average measurements were taken over a period from January 1, 2024 to June 10, 2024.

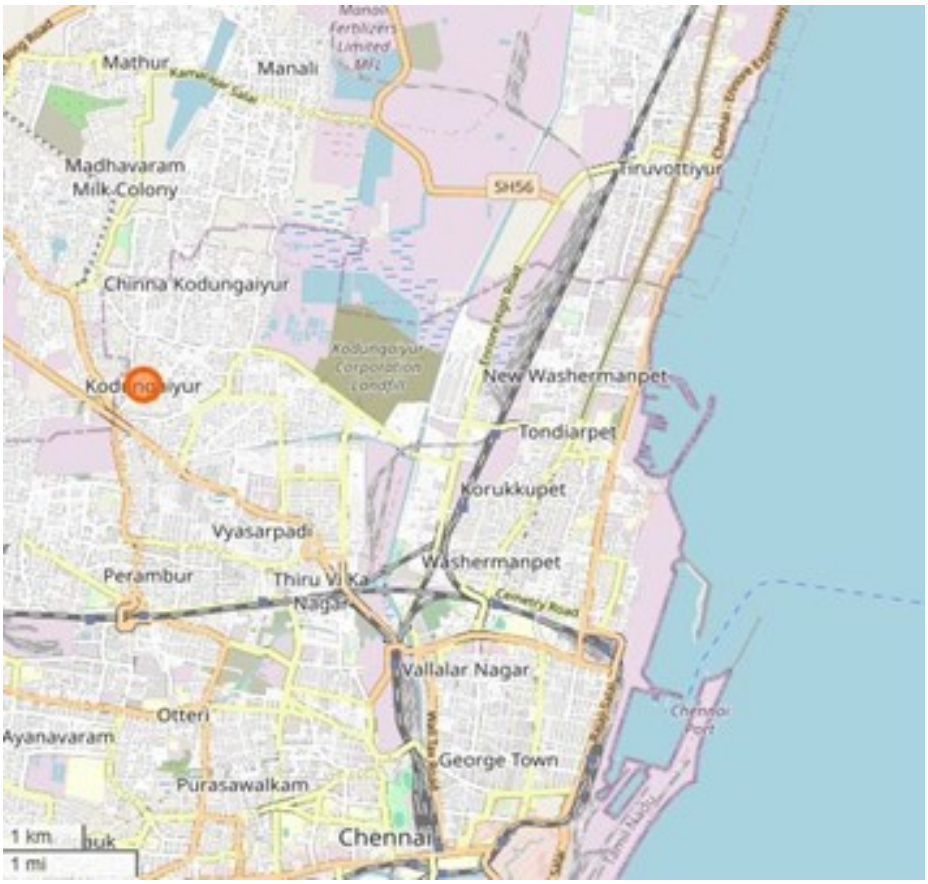


Figure 1: Geographical Location of Kodungaiyur

## 2.2 Principal Component Analysis

Principal Component Analysis (Trendafilov et al., 2021) is a feature reduction technique used to simplify datasets with many features by transforming them into a new set of uncorrelated variables called principal components (PCs). PCA reduces the number of features while retaining essential information by selecting the top PCs based on the variance they explain. For example, if the first three PCs explain 98% of the total variance, only these three are retained, reducing dimensionality without significant information loss. PCA is particularly useful when original variables are strongly correlated, enhancing model efficiency and performance. The major purpose of standardizing continuous variables prior to performing Principal Component study (PCA) is to ensure that each variable contributes equally to the study. PCA is sensitive to the variances of the starting variables, therefore without standardization; variables with greater ranges may dominate the analysis, resulting in biased conclusions and also standardization handles the variables measured with different units uniformly. Standardization gives each variable a mean of zero and a standard deviation of one, putting them on the same scale. This enables PCA to precisely identify the principle components that capture the most variance in the data, resulting in more trustworthy and interpretable findings.

## 2.3 PCA on Compositional Data:

Compositional Principal Component Analysis (PCA)(Aitchison, 1986) is a statistical technique that aims to analyze and reduce the dimensionality of compositional data. This methodology is vital for exploring relationships among components that are part of a whole, which is particularly important in fields where data are constrained to sum to a constant, such as chemical compositions or environmental data. Since compositional data are constrained to a simplex, traditional PCA methods are not appropriate as they do not account for the specific geometry of the data. To address this, compositional PCA employs specific transformations, such as the Centered Log-Ratio (CLR), Additive Log-Ratio transformation (ALR), and Isometric Log-Ratio (ILR) transformation, to project the data into a Euclidean space where standard PCA can be applied. The first step in PCA is transforming the data into a suitable form.

### 2.3.1 Centered Log-Ratio (CLR) Transformation:

The CLR is a technique for analyzing compositional data that addresses the constraint that variables reflect proportions of a whole and must sum to a constant total. The transformation converts compositional data from a simplex space to a real Euclidean space, allowing regular statistical techniques to be more efficiently used. The method involves taking the natural logarithm of component proportions to a reference component and centering them on their mean. Given a *D-dimensional vector*, say $X = (x_1, x_2, \ldots, x_D)$ where $x_i > 0$ for $i = 1, 2, \ldots, D$ and $\sum_{i=1}^{D} x_i = 1$ the CLR transfor-

mation of $X$ is given by:

$$\text{clr}(X) = \ln\left(\frac{x_i}{\bar{x}}\right) \quad \text{where } \bar{x} = \left(\prod_{i=1}^{D} x_i\right)^{\frac{1}{D}} \tag{1}$$

where $x_i$ represents the components of the compositional data and $(D)$ is the number of components.CLR transformation provides a way to analyze data with respect to the geometric mean, making it suitable for identifying relative changes in components (Egozcue et al., 2003).

### 2.3.2 Isometric Log-Ratio (ILR) Transformation::

The ILR transforms compositional data into a multivariate observation $z = (z_1, \ldots, z_{D-1})$ in the Euclidean space. This transformation is useful for retaining the distances between data, which is critical for statistical analysis. The ILR transformation for a compositional vector $X = (x_1, x_2, \ldots, x_D)$ is given by:

$$z_j = \sqrt{\frac{i}{i+1}} \ln\left(\frac{\prod_{j=1}^{i} x_j}{x_i + 1}\right), \quad i = 1, \ldots, D-1 \tag{2}$$

This transformation reduces the number of dimensions in the data while keeping important information about the component relationships. When considering only two parts, $(x_1, x_2)$, the equation simplifies to:

$$z = \sqrt{\frac{1}{2}} \ln \frac{x_1}{x_2} \tag{3}$$

In this case, the ILR variable z is univariate but contains information about the relationship between$(x_1, x_2)$ by means of their logarithmic ratio.

Using the basis vectors obtained from the SBP (Sequential Binary Partition), transform the CLR coordinates into ILR coordinates. If $V_1, V_2, \ldots, V_{D-1}$ are theorthonormal basis vectors, the ILR coordinates are:

$$z_j = \text{clr}(X) \cdot V_j \quad \text{for } j = 1, 2, \ldots, D-1 \tag{4}$$

where $\cdot$ denotes the dot product.

The ILR transformation effectively reduces the dimensionality of the compositional data from $D$ to $D-1$, making it possible to apply standard statistical methods to the transformed data.

### 2.3.3 Additive Log-Ratio (ALR) Transformation:

The ALR transformation uses one component as a reference and computes the natural log-ratios of other components relative to this reference to get a closed composition.

$$\text{alr}(x_i) = \ln\left(\frac{x_i}{x_D}\right) \quad \text{for } i = 1, 2, \ldots, D-1 \tag{5}$$

This means that each component $x_i$ of the composition is divided by the lastcomponent $x_D$ and then the logarithm is taken. The ALR transformation is not an isometric transformation. This means that it does not retain the geometric qualities of the original data, such as distances, angles, and forms. As a result, the ALR transformation should never be used to compute distances, angles, or forms because the results would be distorted and incomprehensible in the context of the original data. Following the transformation of the compositional data, PCA can proceed as normal. The principal components and matching eigenvalues of the modified data are then determined by the PCA by calculating the covariance matrix of the data. The directions of highest variability in the modified compositional data space are represented by these primary components.

## 3 Analysis and Findings

The CoDA-PCA is applied on the transformed data (from three transformations) and the findings from the analysis are interpreted in this section.

Table 1: Variance ratio explained by each Principal Component obtained from all three log-ratio transformations

| PCs | CLR | ILR | ALR |
|---|---|---|---|
| PC1 | 0.5353 | 0.5353 | 0.5904 |
| PC2 | 0.2282 | 0.2282 | 0.1787 |
| PC3 | 0.1093 | 0.1093 | 0.1423 |
| PC4 | 0.0757 | 0.0757 | 0.0503 |
| PC5 | 0.0186 | 0.0186 | 0.0125 |
| PC6 | 0.0128 | 0.0128 | 0.0099 |
| PC7 | 0.0106 | 0.0106 | 0.0070 |
| PC8 | 0.0065 | 0.0065 | 0.0050 |
| PC9 | 0.0030 | 0.0030 | 0.0040 |
| PC10 | 0.0001 | 0.0001 | 0.0000 |
| PC11 | 0.0000 | - | - |

This table summarizes the variance ratio explained by each Principal Component (PC) for three different log-ratio transformations: Centered Log-Ratio (CLR), Isometric Log-Ratio (ILR), and Additive Log-Ratio (ALR).The other observations from the table as follows; Consistency Between CLR and ILR - The variance ratios for CLR and ILR transformations are identical across all principal components, which suggest that these two transformations yield the same PCA results in this context. This similarity might

occur because both transformations respect the geometric properties of compositional data and preserve the relative information between components. ALR Transformation - The variance ratios for ALR differ slightly from those of CLR and ILR, particularly for the first three principal components (PC1, PC2, and PC3). This difference indicates that the ALR transformation, which uses a reference component, captures variance differently, potentially emphasizing certain components over others. For example, PC1 explains more variance in ALR (0.5904) compared to CLR and ILR (0.5353), suggesting that the first principal component in the ALR-transformed data captures more of the total variance. However, PC2 and PC3 explain less variance under ALR, which indicates that the variability is more concentrated in PC1 after the ALR transformation. Dominance of the First Few PCs - Across all three transformations, the first two principal components (PC1 and PC2) explain a large proportion of the total variance, with PC1 alone explaining over 50% of the variance in all cases. This suggests that the majority of the information in the compositional data can be captured by just the first two principal components. Moving further to higher-order components (PC4 and beyond), the explained variance decreases sharply, indicating that these components contribute only marginally to explaining the total variance in the data. The results suggest that PCA on compositional data, regardless of the log-ratio transformation used, reveals a strong underlying structure where the first few principal components capture most of the variance. The slight differences in variance explained between ALR and the other transformations highlight that the choice of transformation can influence the distribution of explained variance among the principal components. This means that the first two or three principal components are likely sufficient for representing the underlying structure of the compositional data.

This table shows the loadings of the first two principal components (PC1 and PC2) obtained from PCA performed on data transformed using three different log-ratio methods: CLR (Centered Log-Ratio), ILR (Isometric Log-Ratio), and ALR (Additive Log-Ratio). The loadings represent how much each original variable contributes to each principal component. CLR Transformation – Ozone has the highest positive loading (0.8427), indicating that it strongly contributes to PC1. This suggests that variations in ozone levels are a significant part of the overall variance captured by PC1.SO2 (-0.3288) and PM10 (-0.2599) have notable negative loadings, meaning they inversely contribute to PC1. This could indicate an inverse relationship between these pollutants and the overall structure of the data as represented by PC1. SO2 (0.8519) strongly contributes to PC2, indicating that variations in SO2 levels drive a significant portion of the variance captured by PC2.Benzene (-0.2911) and Toluene (-0.3436) have significant negative loadings, suggesting an inverse relationship with PC2. ILR Transformation – CO (-0.8547) has the highest negative loading, suggesting it significantly drives the variance captured by PC1 but in an inverse direction. PM10 (-0.2433) also contributes negatively, though less strongly.PM2.5 (0.0442) contributes positively, though weakly, indicating it has a minor positive relationship with PC1.NH3 (-0.8484) has a strong negative loading, indicating it heavily influences the variance captured by PC2.Ozone (0.3431) and Benzene (0.3604) contribute positively, suggesting that higher levels of these pollutants are associated with the variance captured by PC2. ALR Transformation – CO (-0.7766) and PM10 (-0.3043)

Table 2: First two principal components obtained from all three log-ratio transformations

| **Log-ratio Transformations** | **CLR** | | **ILR** | | **ALR** | |
|---|---|---|---|---|---|---|
| **PCs** | **PC1** | **PC2** | **PC1** | **PC2** | **PC1** | **PC2** |
| **$PM_{2.5}$** | -0.1973 | -0.0614 | 0.0442 | 0.0409 | 0.0601 | -0.0436 |
| **$PM_{10}$** | -0.2599 | -0.1193 | -0.2433 | 0.0256 | -0.3043 | 0.1924 |
| **NO** | 0.0694 | -0.1216 | -0.1465 | -0.0469 | -0.2542 | 0.1345 |
| **$NO_2$** | 0.0399 | -0.0466 | -0.1273 | -0.0124 | -0.2766 | 0.1559 |
| **$NO_x$** | 0.0553 | -0.0733 | -0.1232 | -0.1094 | -0.2499 | 0.0522 |
| **$NH_3$** | 0.0764 | 0.0354 | 0.2711 | -0.8484 | -0.0567 | 0.8309 |
| **$SO_2$** | -0.3288 | -0.8519 | -0.1049 | 0.0652 | -0.2460 | 0.1412 |
| **CO** | 0.0343 | -0.0033 | -0.8547 | -0.1085 | -0.7766 | -0.3622 |
| **Ozone** | 0.8427 | 0.1728 | 0.2331 | 0.3431 | -0.0766 | 0.1792 |
| **Benzene** | -0.2088 | -0.2911 | 0.1293 | 0.3604 | -0.1659 | 0.2057 |
| **Toluene** | -0.1232 | -0.3436 | - | - | - | - |

are significant negative contributors, indicating they strongly influence PC1 but in an inverse direction.PM2.5 (0.0601) has a positive, albeit weak, contribution, meaning it has a slight positive relationship with PC1.NH3 (0.8309) contributes strongly to PC2, suggesting it is a key driver of the variance captured by this component.CO (-0.3622) contributes negatively, indicating it has an inverse relationship with PC2. Across all transformations, certain pollutants like NH3 and CO appear to consistently have strong loadings, albeit in different directions depending on the transformation and principal component. This suggests that these pollutants are key drivers of the data's variance, regardless of the specific log-ratio transformation used. The choice of transformation (CLR, ILR, and ALR) influences which variables are emphasized in the principal components. For example, CO is a major contributor to PC1 in ILR and ALR but not in CLR, where Ozone is more prominent. This implies that the interpretation of the principal components and their environmental or health implications can vary depending on the transformation applied to the data. PC1 generally represents a linear combination where a few key pollutants such as Ozone, CO, SO2 dominate, suggesting that these components capture the most substantial variation in the data. PC2 often highlights a different set of pollutants e.g., SO2 in CLR, NH3 in ILR/ALR, which may indicate different patterns or sources of pollution. A scree plot can significantly enhance the interpretation of the tables provided by visually demonstrating how much of the total variance in the data is explained by each principal component (PC).

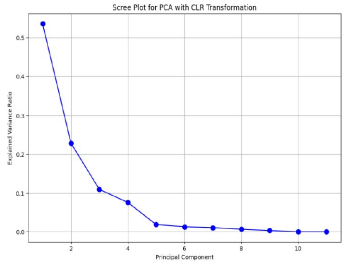


Figure 2: Scree plot – CLR Transformation

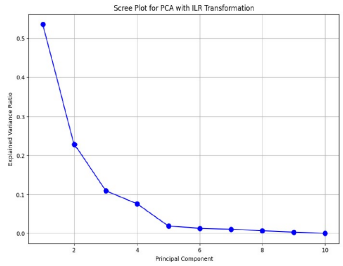


Figure 3: Scree plot – ILR Transformation

The scree plots in Figures 2, 3 and 4 which all exhibit similar patterns, visually confirm that the variance drops significantly after the first two PCs. The "elbow" point in each plot supports the conclusion that the first two PCs capture the majority of the variance, consistent with the high variance ratios shown in Table 1. The elbow method, observed in the scree plots, consistently indicates that retaining the first two PCs is a reasonable choice because they capture the majority of the variance. This aligns with the quantitative results in Table 1, where the cumulative variance explained by the first two PCs is substantial across all transformations. Since the scree plots suggest that only the first two PCs are necessary, this supports focusing on the loadings of these

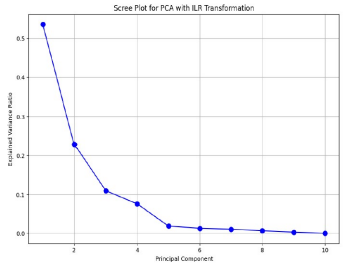


Figure 4: Scree Plot – ALR Transformation

PCs in Table 2. The loadings in Table 2 for PC1 and PC2 are thus the most critical for interpretation, as they represent the variables contributing to the most significant dimensions of variation in the data. Although the ALR transformation shows a slightly higher variance explained by PC1, the scree plot still confirms that the general pattern is similar to CLR and ILR, suggesting that the transformation choice doesn't drastically alter the conclusion about the number of important PCs. A biplot becomes essential following the PCA results as it provides a comprehensive visual representation of the data, allowing for the simultaneous interpretation of both the principal components and the original variables.

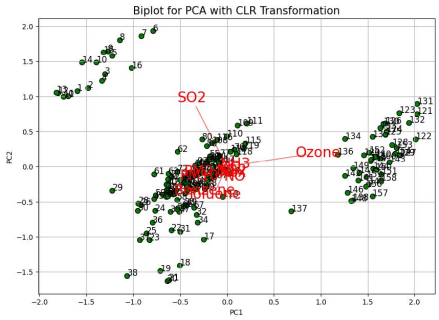


Figure 5: Biplot – CLR Transformation

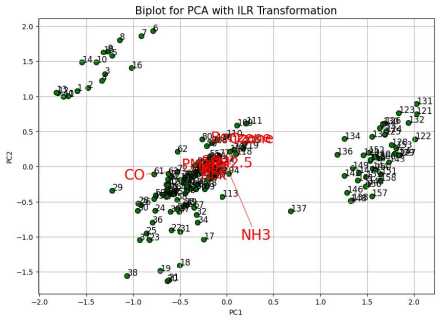


Figure 6: Biplot – ILR Transformation

Figure 5 provides a visual representation through biplot for PCA with CLR transformation. The arrow for Ozone (O3) points strongly in the direction of PC1 with a high positive loading (0.8427) and a moderate negative loading on PC2 (-0.1728). This indicates that Ozone strongly contributes to the variance captured by PC1 and plays a significant role in separating the data along this component. The arrow for SO2 points in the negative direction for both PC1 (-0.3288) and PC2 (-0.8519), indicating it has a strong influence, particularly along PC2. Its placement suggests that SO2 contributes heavily to the variance captured by PC2 and is inversely related to the pollutants that load positively on this component. Both PM10 and Toluene have moderate positive

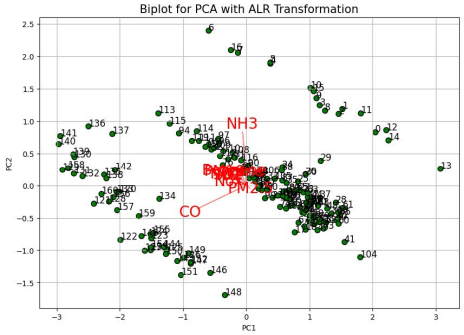


Figure 7: Biplot – ALR Transformation

loadings on PC2 and moderate negative loadings on PC1. Their arrows will point toward the quadrant where PC2 is positive, indicating that these pollutants contribute to PC2, but in a manner that is opposite to Ozone along PC1. These pollutants have negative loadings on PC1 and positive loadings on PC2, which suggests that they contribute more to PC2 but are inversely related to the pollutants that load positively on PC1. PM2.5, PM10, and Benzene are grouped relatively close together with negative loadings on PC1 and positive loadings on PC2. This proximity suggests that these pollutants are related and might be influenced by similar sources or processes in the environment. The variables NO, NO2, and NOx have low but positive loadings on both PC1 and PC2, indicating that they are somewhat related but do not dominate the variation captured by either principal component. Their clustering close to the origin suggests they do not strongly influence the separation of data along either PC1 or PC2. The biplot likely shows a clear separation between O3 and SO2 due to their strong but opposite loadings on PC1 and PC2. This suggests that these pollutants are negatively correlated and represent different aspects of the data variability. Clusters of observations near certain arrows would suggest that those pollutants are dominant in those observations. For example, a cluster near the PM2.5, PM10, and Benzene arrows would indicate that these pollutants are jointly prevalent in those observations. Biplot for PCA with ILR transformation is showed in figure 6. The arrow for CO points strongly in the positive direction of PC1 (-0.8547) and slightly negative in PC2 (-0.1085). This indicates that CO is the most influential pollutant along the first principal component (PC1), contributing significantly to the variance captured by PC1. The arrow for NH3 points strongly in the negative direction for both PC1 (-0.2711) and particularly for PC2 (-0.8484). This suggests that NH3 has a strong influence on the second principal component (PC2) and negatively correlates with variables that load positively on PC2.Both O3 and Benzene have positive loadings on PC2 (0.3431 and 0.3604, respectively) but negative loadings on PC1 (-0.2331 and -0.1293, respectively). This suggests that these pollutants contribute to PC2 and are inversely related to the variables contributing to PC1, particularly CO. The arrow for PM10 points moderately in the positive direction of PC1 (0.2533) and slightly on PC2 (0.0256). This indicates that PM10 is a contributor to PC1 and has a

mild influence on the variance captured by PC2. The pollutants NO, NO2, and NOx have positive loadings on PC1 but slightly negative or very small loadings on PC2. This indicates that they contribute to PC1 but have little impact on PC2. NO, NO2, NOx, and PM10 are grouped closely with positive loadings on PC1, suggesting they are related and may share similar sources or effects in the environment. Their proximity indicates that they jointly influence the variance captured by PC1.O3 and Benzene are grouped with positive loadings on PC2 and negative loadings on PC1, indicating they are related and may act together in influencing the variance captured by PC2. Their direction suggests an inverse relationship with pollutants like CO, which dominate PC1.CO and NH3 are likely plotted in opposite quadrants due to their contrasting loadings and CO with strong positive influence on PC1 and NH3 with strong negative influence on PC2. This suggests that these pollutants are inversely related, perhaps indicating different sources or environmental conditions affecting them. Observations near the CO arrow would indicate high CO levels, while those near NH3 would indicate high NH3 levels. Clusters of observations near the O3 and Benzene arrows would suggest higher levels of these pollutants. Observations that cluster near NO, NO2, NOx, and PM10 arrows would indicate that these pollutants are prevalent together in those data points. Figure 7 shows the biplot of ALR transformation. CO (-0.7765) has the largest negative loading on PC1, indicating that it is the strongest driver of variation along this component. This suggests that CO is highly influential in differentiating samples along PC1, possibly due to its distinct sources or high concentration levels compared to other pollutants. The pollutants PM10 (-0.3042), NO (-0.2542), NO2 (-0.2766), NOx (-0.2499), and SO2 (-0.2459) have moderate negative loadings on PC1, indicating they are somewhat aligned with CO in driving the variation along this axis. These pollutants likely come from similar sources or processes, such as combustion or traffic emissions. PM2.5 (0.0600) and Ozone (-0.0766) have smaller loadings, meaning they have less influence on PC1. Ozone has a small negative loading, while PM2.5 has a small positive loading, suggesting they may not be strongly associated with the dominant source(s) of pollution represented by PC1 and seems to represent pollution from combustion-related sources, primarily dominated by CO, and to a lesser extent, PM10, NOx, and SO2. Since many of these pollutants are linked to vehicular or industrial emissions, PC1 likely captures the variation due to these sources. NH3 (0.8309) has a very large positive loading on PC2, making it the dominant pollutant driving variation along this component. This indicates that PC2 is likely associated with sources specific to NH3, such as agricultural activities, waste management, or certain industrial processes. Benzene (0.2057), PM10 (0.1924), and NO2 (0.1559) also contribute to PC2, although to a lesser extent than NH3. The positive loadings suggest that they may share some secondary sources with NH3 or reflect different processes that contribute to pollution variation captured by PC2. CO (-0.3621) has a negative loading on PC2, indicating it is inversely related to NH3. This suggests that areas or time periods with high NH3 concentrations may have lower CO levels, pointing to different sources for these pollutants. PC2 is primarily driven by NH3, likely representing pollution from agricultural or waste sources. The positive loadings of pollutants like PM10 and Benzene might indicate that they are associated with industrial or urban activities, but on a smaller scale compared to NH3. The negative contribution of CO suggests that PC2

distinguishes between NH3-driven pollution and CO-related emissions. PC1 likely represents pollution driven by combustion and vehicular emissions, with CO, PM10, NOx, and SO2 contributing strongly. PC2 differentiates pollution from agricultural or waste sources, with NH3 being the primary driver. The findings from the study highlights the need for targeted pollution control measures, particularly focusing on major pollutants like ozone (O3), NH3, and SO2, which have significant impacts as shown in both CLR and ALR analyses. Policy efforts should aim to reduce O3 by addressing emissions from industrial activities and vehicles, as it is a secondary pollutant formed from precursors like NOx and VOCs. NH3 and SO2, which are influenced by agricultural and industrial sources respectively, require stricter emission standards and the promotion of cleaner technologies to mitigate their levels. In addition, addressing source-specific pollution is crucial. CO, identified as a dominant factor in the ILR analysis, can be managed by improving vehicle emissions standards, promoting cleaner fuels, and enhancing industrial practices to reduce incomplete combustion. Since NOx plays a dual role by contributing to O3 formation and overall pollution, strategies should focus on reducing emissions from transportation and industry through advanced combustion technologies and enhanced regulatory controls. Pollution source identification is also critical, with PM2.5, PM10, and benzene clustering in the CLR analysis, suggesting a shared source, likely from construction, industrial activities, and vehicular emissions. Comprehensive air quality monitoring is necessary to identify and regulate these sources effectively. To strengthen air quality management, establishing a network of air quality monitoring stations across Chennai is recommended, with real-time data systems to provide immediate feedback and enable timely interventions during pollution spikes. Public awareness must be a key component of these efforts. Educating the public on the health impacts of pollutants like O3, NH3, and SO2, and encouraging behavioral changes, such as using public transport or reducing vehicle idling, can help minimize individual contributions to air pollution. Finally, cross-sector collaboration is essential for policy integration. A coordinated approach across transportation, industry, and agriculture is necessary to ensure that efforts in one sector do not adversely affect others. Strengthening regulatory frameworks and incentivizing cleaner technologies will further support compliance with air quality standards and foster long-term improvements in air quality.

## 4 Conclusion

In this study, CoDA PCA applied to air concentration data from the Kodungaiyur region of Chennai, using three different log-ratio transformations: Centered Log-Ratio (CLR), Isometric Log-Ratio (ILR), and Additive Log-Ratio (ALR). The analysis demonstrates that the first two principal components account for approximately 76% of the variance in the data, indicating a robust reduction in dimensionality while retaining significant information. The CLR transformed biplot reveals a clear structure in the data, with distinct relationships between pollutants. The negative correlation between O3 and pollutants like SO2 and Toluene suggests that different environmental processes or sources are influencing these pollutants. Additionally, the clustering of PM2.5, PM10, and Benzene

highlights their potential common sources or similar environmental impacts, underscoring their importance in the data's variability. The ILR transformed biplot emphasizes CO as the primary contributor to PC1 and NH3 as a significant factor for PC2. This bifurcation indicates that CO and NH3 are associated with distinct aspects of pollution, with CO dominating the variance captured by PC1 and NH3 significantly influencing PC2. The proximity of NO, NO2, NOx, and PM10 suggests they are interrelated and may originate from similar sources. The ALR transformed biplot further confirms O3 as a major pollutant affecting both principal components, particularly in the negative direction. NH3 and SO2 are identified as key contributors to PC2, suggesting their substantial role in the data's variability. The clustering of NO, NO2, NOx, and CO indicates they share related environmental sources or effects, predominantly influencing PC1. The inverse relationship between O3 and pollutants like NH3 and SO2 suggests different underlying sources or impacts on the environment. In conclusion, long-term strategic planning for sustainable development must prioritize air quality by integrating it into urban planning, designing cities to minimize pollution and enhance green spaces. Addressing the impacts of climate change on air quality is also essential, with adaptation measures to mitigate future changes in pollution patterns. Furthermore, support for research and development is vital to drive technological advances in pollution control and monitoring, while investing in advanced data analysis tools will provide deeper insights into pollution sources and trends, ultimately improving air quality management and fostering innovative solutions

## Acknowledgement

This research is funded by RUSA 2.0, Ministry of Human Resource Development, Government of India.

## References

Aitchison, J. (1986). *The statistical analysis of compositional data*. Chapman & Hall Ltd.

Pawlowsky-Glahn, V. e Egozcue, J. J. (2001). Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment*, 15(5), 384–398.

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., e Barcelo-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3), 279–300.

Aitchison, J. e Greenacre, M. (2002). Biplots of compositional data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 51(4), 375–392.

Egozcue, J. J. e Pawlowsky-Glahn, V. (2005). Groups of parts and their balances in compositional data analysis. *Mathematical Geology*, 37(7), 795–828.

Hron, K., Engle, M., Filzmoser, P., e Fišerová, E. (2021). Weighted symmetric pivot

coordinates for compositional data with geochemical applications. *Mathematical Geosciences*, 53(4), 655–674.

Gallo, M. (2012). CoDa in three-way arrays and relative sample spaces. *Electronic Journal of Applied Statistical Analysis*, 5(3), 400-405.

Gallo, M. e Simonacci, V. (2013). A procedure for the three-mode analysis of compositions. *Electronic Journal of Applied Statistical Analysis*, 6(2), 202-210.

Gallo, M., Simonacci, V., e Di Palma, M. A. (2019). An integrated algorithm for three-way compositional data. *Quality & Quantity*, 53, 2353-2370.

Trendafilov, N., et al. (2021). Principal component analysis (PCA). In *Multivariate Data Analysis on Matrix Manifolds: (with Manopt)* (pp. 89-139).

Trendafilov, N., Gallo, M., e Tierney, R. J. (2023). PCA and other dimensionality-reduction techniques. In *International Encyclopedia of Education (Fourth Edition)* (pp. 590-599). Elsevier.

Jarauta-Bragulat, E., Hervada-Sala, C., & Egozcue, J. J. (2016). Air Quality Index revisited from a compositional point of view. *Mathematical Geosciences*, 48, 581–593.