

# Análise Estatística do Impacto de Diferentes Fatores na Qualidade do Ar em Países da Ásia Meridional

Vinícius Lôbo

Departamento de Engenharia de Teleinformática  
Universidade Federal do Ceará  
Fortaleza, Brasil  
vinicius.alp@alu.ufc.br

Mateus R. Gomes

Departamento de Engenharia de Teleinformática  
Universidade Federal do Ceará  
Fortaleza, Brasil  
matribg04@alu.ufc.br

Gabriel Melo

Departamento de Engenharia de Teleinformática  
Universidade Federal do Ceará  
Fortaleza, Brasil  
gabriel.abm@alu.ufc.br

Filipe Caetano

Departamento de Engenharia de Teleinformática  
Universidade Federal do Ceará  
Fortaleza, Brasil  
filipecaetano@alu.ufc.br

**Abstract**—A baixa qualidade do ar pode ocasionar irritação nos olhos, tosse, dificuldade para respirar e até mesmo o agravamento de doenças respiratórias ou crônicas. Diante disso, este trabalho busca compreender o impacto de diferentes fatores na determinação da qualidade do ar por meio de uma análise exploratória de dados de diversas regiões de países da Ásia Meridional, a região mais populosa do mundo. Foram empregadas medidas estatísticas descritivas e técnicas de visualização de dados, analisadas tanto de forma incondicional quanto condicional à qualidade do ar. Além disso, investigamos as correlações entre os diferentes poluentes e aplicamos estratégias de normalização e detecção de outliers para o adequado tratamento dos dados. Por fim, realizamos uma análise de componentes principais, cuja representação bidimensional dos dois primeiros componentes permitiu evidenciar padrões e agrupamentos associados aos diferentes níveis de qualidade do ar. Os resultados obtidos indicam que determinados poluentes exercem influência mais significativa sobre a qualidade do ar, evidenciando a importância de abordagens estatísticas multivariadas para a compreensão dos fenômenos ambientais subjacentes. Os códigos base e experimentos reproduutíveis estão disponíveis publicamente em: <https://github.com/gabrielmelo7/Homeworks-de-ICA>

**Index Terms**—Qualidade do ar, poluição, análise de dados, estatística.

## I. INTRODUÇÃO

A qualidade do ar é o grau de pureza do ar que é respirado em um ambiente, é medido pela concentração de gases tóxicos e material particulado fino ou grosso. A alta concentração desses materiais no ar afeta diretamente a saúde humana, podendo causar danos agudos como irritação respiratória e danos crônicos tais quais doenças respiratórias e cardíacas, câncer de pulmão, infecção respiratória em crianças e bronquite em adultos [1]. Evitar a concentração de poluentes em ambientes urbanos se torna, portanto, uma prioridade quando se visa a saúde global.

Ademais, o Sul da Ásia surge como uma região de estudo crítico. Países como Bangladesh, Índia, Nepal e Paquistão estão entre os mais poluídos do mundo. Segundo o Relatório

Mundial de Qualidade do Ar de 2020, 37 das 40 cidades mais poluídas do globo estão localizadas no Sul da Ásia. A poluição nesta região possui causas complexas e multivariadas, influenciadas por muitos fatores, como emissões industriais e de veículos.

A análise exploratória dos dados consiste no uso de técnicas de estatística para descrever, caracterizar e poder inferir sobre um conjunto de dados. O foco da análise exploratória dos dados é simplificar o entendimento de um conjunto de dados com diferentes formas de visualização, priorizando o entendimento por meio da percepção, não da cognição [2]. Nesse contexto, tais estratégias têm se mostrado eficientes para identificar certos tipos de poluentes e fatores que impactam diretamente na qualidade do ar, como demonstrado em [3] e [4].

Este trabalho propõe uma análise exploratória e estatística de um conjunto de dados sobre a qualidade do ar em países do Sul da Ásia, com o objetivo de investigar a estrutura do *dataset* e avaliar o impacto de suas variáveis para a variabilidade da poluição. A contribuição deste estudo reside na metodologia aplicada para extrair os padrões mais significativos, demonstrando a relevância dos principais vetores da poluição na região.

## II. MATERIAIS E MÉTODOS

O dataset selecionado para análise pode ser encontrado em <https://www.kaggle.com/datasets/mujtabamatin/air-quality-and-pollution-assessment>. O conjunto captura fatores ambientais e demográficos críticos que influenciam os níveis de poluição e contém 5.000 observações e 10 atributos, dos quais 9 correspondem a variáveis preditoras e 1 representa a variável alvo (classe) associada a cada amostra. As variáveis analisadas são apresentadas na tabela I.

TABLE I  
COLUNAS DO DATASET E SEUS TIPOS

Coluna	Tipo
Temperature	float64
Humidity	float64
PM2.5	float64
PM10	float64
NO2	float64
SO2	float64
CO	float64
Proximity to Industrial Areas	float64
Population Density	int64
Air Quality	object

Onde "Air Quality" é o rótulo dos dados e possui quatro categorias:

- *Good*: Ar com baixos níveis de poluição. 2000 ocorrências no dataset.
- *Moderate*: Qualidade de ar aceitável com alguns poluentes presentes. 1500 ocorrências no dataset.
- *Poor*: Poluição notável que pode causar riscos de saúde para grupos sensíveis. 1000 ocorrências no dataset.
- *Hazardous*: Ar altamente poluído que causa sérios riscos de saúde para a população. 500 ocorrências no dataset.

#### A. Análise monovariada

Neste estudo foram utilizadas as métricas da média, desvio padrão e assimetria para analisar os dados de qualidade do ar.

A média é definida como o valor central de uma distribuição e o desvio padrão é uma medida de dispersão em torno do valor central. Já a assimetria é a medida de obliquidade de uma distribuição em relação à sua média. Para calcular a média, o desvio padrão e a assimetria foram utilizadas as equações 1, 2, 3.

$$\mu = \frac{1}{N} \sum_{i=1}^n a_i \quad (1)$$

$$\sigma = \sqrt{\frac{\sum (a_i - \mu)^2}{N}} \quad (2)$$

$$g_1 = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\left[ \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right]^{3/2}} \quad (3)$$

$$G_1 = \frac{\sqrt{n(n-1)}}{n-2} g_1 \quad (4)$$

Sendo a equação 4 a equação da assimetria ajustada.

Metodologicamente, esta análise foi conduzida em duas etapas. Primeiramente, as métricas foram aplicadas de forma agregada a todo o conjunto de dados, visando caracterizar as propriedades estatísticas globais de cada variável e a identificação de padrões relevantes, como distribuições assimétricas. Em seguida, a análise foi segmentada por classe, permitindo uma investigação condicional para comparar a tendência central, a dispersão e a forma da distribuição entre os diferentes grupos de rótulos. Nas duas etapas foram traçados, para melhor entendimento, os histogramas e na primeira etapa, um diagrama de caixa das distribuições.

#### B. Análise bivariada

A análise bivariada do dataset permite melhor entendimento das correlações entre os pares de preditores. Para fazer a análise levamos em consideração o cálculo do coeficiente de correlação entre os pares de preditores, que mede o grau de correlação linear entre duas variáveis. A equação utilizada para calcular o coeficiente pode ser vista em 5.

$$r = \frac{\sum (a_i - \mu_a)(b_i - \mu_b)}{\sqrt{\sum (a_i - \mu_a)^2(b_i - \mu_b)^2}} \quad (5)$$

Onde  $a_i$  e  $b_i$  são os valores e  $\mu_a$  e  $\mu_b$  são as médias das diferentes distribuições.

#### C. Métodos de transformação dos dados

Uma etapa fundamental que antecede a análise estatística e a aplicação de métodos de redução de dimensionalidade é o pré-processamento dos dados. A importância deste tratamento reside em dois fatores principais: a análise de componentes principais é sensível à escala e muitas variáveis ambientais apresentam distribuição assimétrica.

Com o objetivo de normalizar as escalas, corrigir assimetrias e mitigar o impacto de valores extremos, as seguintes técnicas de transformação e padronização foram consideradas neste estudo:

- Transformação Z-score [5]
- Transformação de Box-Cox [6]
- Transformação de Yeo-Johnson [7]
- Transformação Spatial Sign [8]

#### D. Análise Multivariada Incondicional

A realização de uma análise multivariada incondicional nos permitiu visualizar e entender melhor como as *features* afetam o comportamento e a variabilidade do conjunto de dados. A partir de análises preliminares, como a matriz de correlação obtida com base na análise bivariada, foi possível constatar que há *features* que são altamente correlacionadas, sendo assim, apresentam comportamentos redundantes. Dessa forma, nos foi possível realizar uma redução de dimensionalidade para analisar de forma mais efetiva os dados.

Partindo dessa premissa, o método PCA foi utilizado para fazer uma análise **não supervisionada** nos dados, transformando o nosso conjunto de dados original em um novo conjunto de dados com dimensionalidade reduzida e com componentes ortogonais (linearmente independentes).

O PCA [9] é um método de diminuição de dimensionalidade que tem como objetivo maximizar a variância dos dados em seus componentes. O PCA realiza isso ao criar uma nova dimensão na qual os novos eixos são combinações lineares dos eixos originais que representam as *features* de forma a criar vetores ortogonais entre si. Sendo assim, o componente principal 1 deve concentrar a maior variância dos dados, seguido pelo 2 e assim por diante até chegarmos à dimensão dos preditores do conjunto de dados.

O PCA se utiliza da matriz de covariância do *dataset*. Contudo, para resultados mais apropriados e acurados, os dados foram pré-processados, utilizando técnicas descritas anteriormente, de tal forma a estarem com obliquidade baixa e normalizados.

Sendo assim, para a realização do PCA foi necessário seguir uma série de passos [10].

- 1) Determinar os autovalores e os autovetores da matriz de covariância dos dados normalizados e centralizados.
- 2) Análise dos Principal Components (PC's) encontrados. Obs: Nessa análise a visualização foi realizada a partir de um *Scree plot*.
- 3) Escolha dos PC's que melhor explicam a distribuição.
- 4) Determinar como as *features* afetam cada PC.
- 5) Cálculo dos Scores a partir do conjunto de dados normalizado.
- 6) Realizar o *Scatter plot* dos scores nos eixos escolhidos.

#### E. Detecção de outliers

A fim de detectar e tratar possíveis *outliers*, foi utilizada uma abordagem baseada na distância de mahalanobis (MD) [11]. Tal distância é uma métrica que calcula a distância entre o ponto e a distribuição. Ela funciona de maneira bastante eficaz em dados multivariados, pois utiliza a matriz de covariância das variáveis para calcular a distância entre os pontos de dados e o centro (6). Isso significa que a MD detecta outliers com base no padrão de distribuição dos pontos de dados, ao contrário da distância euclidiana. A diferença pode ser vista na Fig. 1.

$$D^2 = (x - \mu)^T \cdot C^{-1} \cdot (x - \mu) \quad (6)$$

Onde:

- $D^2$ : é o quadrado da distância de Mahalanobis.
- $x$ : é o vetor da observação (linha no conjunto de dados).
- $\mu$ : é o vetor de valores médios das variáveis (média de cada coluna).
- $C^{-1}$ : é a matriz de covariância inversa das variáveis.

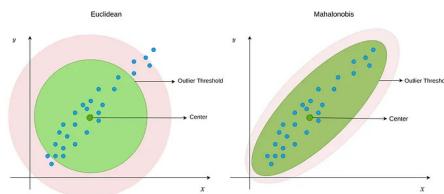


Fig. 1. Distância Euclidiana vs Distância de Mahalanobis

Sob a premissa de que os dados seguem uma distribuição normal multivariada, o quadrado da MD segue uma distribuição  $\chi^2$ . Desta forma, um ponto é classificado como *outlier* se seu valor  $D^2$  excede um limiar crítico determinado pela distribuição  $\chi^2$  para um nível de significância pré-definido e graus de liberdade iguais ao número de variáveis analisadas.

Na análise em questão, a detecção de *outliers* foi realizada em dois estágios: primeiramente, sobre o conjunto de dados original (porém normalizado) e, em seguida, após a aplicação do PCA.

### III. RESULTADOS

#### A. Análise monovariada

O Cálculo da média, desvio padrão e assimetria utilizando 1, 2 e 4 respectivamente resultou nos valores expostos na tabela II, onde foi visto que no geral os dados que mais variam

TABLE II  
MÉTRICAS DAS COLUNAS DO DATASET

Colunas	Média	Desvio Padrão	Assimetria
Temperature	30.03	6.72	0.752
Humidity	70.06	15.86	0.28
PM2.5	20.14	24.55	2.89
PM10	30.22	27.35	2.53
NO2	26.41	8.90	0.639
SO2	10.02	6.75	1.17
CO	1.50	0.546	0.879
Proximity to Industrial Areas	8.43	3.61	0.47
Population Density	497.4	152.7	0.204

são a Densidade Populacional, o PM10 e o PM2.5. Além disso, pode-se afirmar que os dados em geral são assimétricos, pois a maioria dos valores apresentaram considerável valor de assimetria. Todos os parâmetros apresentaram assimetria positiva, o que significa que seus valores extremos são os maiores que a média e que a distribuição se concentra nos valores menores. Além disso, pode ser dito que a maioria das variáveis apresenta valor baixo de desvio padrão, exceto a Densidade Populacional, que apresenta o maior valor. Podemos visualizar os histogramas dos valores na Fig. 2.

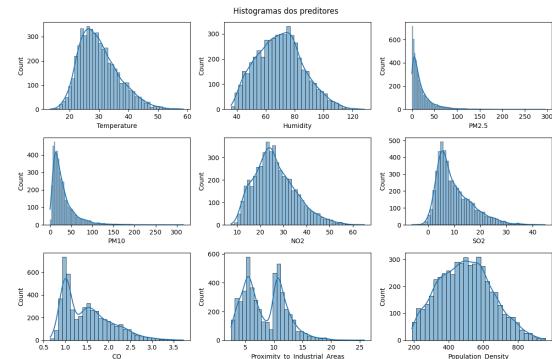


Fig. 2. Histogramas dos preditores

Onde foi visto que, de fato, a distribuição da maioria dos parâmetros se concentrou à esquerda do gráfico. Isso pode ser visualizado de forma mais acentuada nos gráficos de PM2.5, PM10, SO2 e NO2. Analisando os diagramas de caixa dos preditores na Fig. 3, vê-se mais uma vez que as distribuições dos preditores se concentraram nos valores menores, pois os quartis e as medianas estão à esquerda do diagrama, nos

valores baixos, e que seus valores extremos são os maiores que a média.

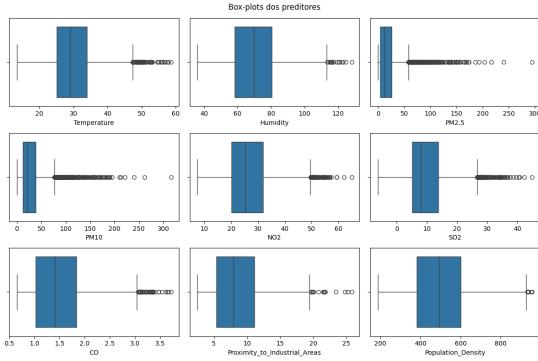


Fig. 3. Diagramas de caixa dos preditores

Com base na análise segmentada pelas classes, foi possível aferir como evidenciado pelas tabelas III e IV (tabelas referentes à melhor e à pior qualidade do ar respectivamente) que à medida que a qualidade piora, a distribuição dos dados se “alarga”, como indicado pelo aumento do desvio padrão. Isso significa que as classes rotuladas como piores não são apenas mais poluídas na média, mas também significativamente mais instáveis e imprevisíveis. As métricas para as demais classes, assim como os plots dos diagramas de caixa, podem ser encontradas no repositório do trabalho.

TABLE III  
MÉTRICAS PARA A CLASSE “GOOD”

Colunas	Média	Desvio Padrão	Assimetria
Temperature	24.94	3.29	-0.047
Humidity	60.021	11.62	0.0014
PM <sub>2.5</sub>	9.91	9.70	1.93
PM <sub>10</sub>	14.98	9.93	1.86
NO <sub>2</sub>	19.44	4.69	-0.038
SO <sub>2</sub>	5.035	2.057	-0.027
CO	0.99	0.101	-0.052
Proximity_to_Industrial_Areas	11.98	2.021	1.98
Population_Density	398.94	115.54	0.031

TABLE IV  
MÉTRICAS PARA A CLASSE “HAZARDOUS”

Colunas	Média	Desvio Padrão	Assimetria
Temperature	40.34	6.47	-0.0064
Humidity	89.47	14.081	0.073
PM <sub>2.5</sub>	41.92	41.34	1.93
PM <sub>10</sub>	61.50	41.90	1.83
NO <sub>2</sub>	40.59	7.73	0.066
SO <sub>2</sub>	20.02	7.99	0.087
CO	2.49	0.40	0.27
Proximity_to_Industrial_Areas	4.59	2.16	2.27
Population_Density	696.01	120.09	0.018

A análise dos histogramas da Fig. 4 ajuda a visualizar o que os números das tabelas indicam. Fica evidente que, à medida que a qualidade do ar piora, os dados se tornam muito mais “espalhados”, confirmando a instabilidade nas classes de ar piores.

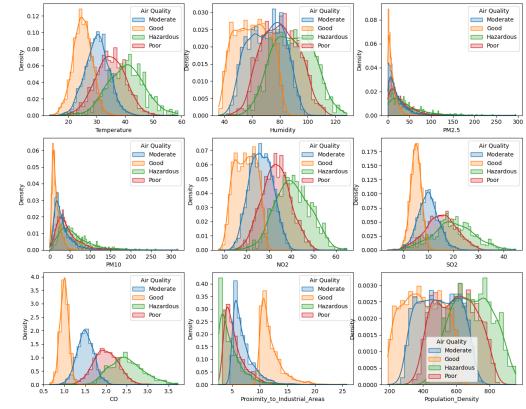


Fig. 4. Histogramas dos preditores com base nas classes

Além disso, os gráficos mostram que diferentes poluentes se comportaram de maneiras distintas. Variáveis como a temperatura e gases (como o NO<sub>2</sub> e o CO) apresentaram gráficos mais simétricos. Nesses casos, a principal diferença entre as classes é que a média de poluição simplesmente se desloca para valores mais altos.

### B. Análise bivariada

Cada coeficiente de correlação calculado foi adicionado a uma matriz de correlação que pode ser vista na Fig. 5, onde pudemos ver uma forte correlação linear entre PM2.5 e PM10, além de uma correlação moderada entre CO e NO<sub>2</sub>, CO e SO<sub>2</sub> e NO<sub>2</sub> e SO<sub>2</sub>. Pudemos ver também que a temperatura tem considerável correlação com NO<sub>2</sub>, SO<sub>2</sub> e CO. Foi possível observar também que há forte correlação negativa entre a proximidade a áreas industriais e o CO e moderada entre a proximidade a áreas industriais e a temperatura, NO<sub>2</sub> e SO<sub>2</sub>.

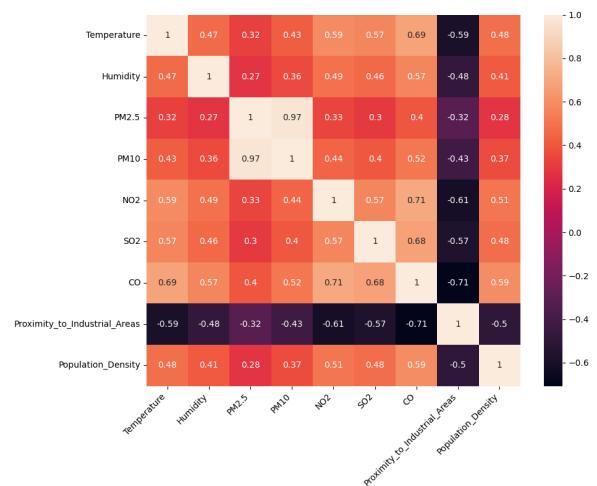


Fig. 5. Matriz de correlação entre os preditores

As relações entre os pares podem ser vistas em formato de gráfico de dispersão na Fig.6, onde é possível constatar visualmente a correlação linear entre *PM2.5* e *PM10* previamente mencionada, as relações entre *CO*, *NO2* e *SO2* e também as relações entre temperatura e *NO2*, *SO2* e *CO* além das correlações negativas entre proximidade a áreas industriais e *CO*, *NO2*, *SO2* e temperatura. As classificações com cores também permitiram ver que altas concentrações de *CO*, *SO2* e *NO2* e alta temperatura são mais comuns em áreas cuja qualidade do ar é considerada danosa à saúde.

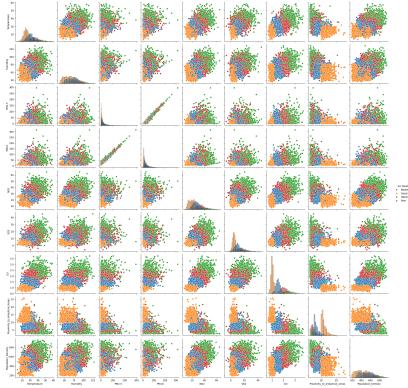


Fig. 6. Gráficos de dispersão dos pares de preditores

### C. Efeito das transformações dos dados

Conforme estabelecido, uma análise visual preliminar do conjunto de dados brutos confirmou a presença de assimetrias em diversas variáveis, além de escalas de magnitude incompatíveis. A Fig. 2 ilustra o estado original dos dados.

Para mitigar esses problemas, foi aplicada a sequência de pré-processamento composta pela transformação Yeo-Johnson, seguida pela padronização Z-score.

Com os dados devidamente transformados para corrigir a assimetria e padronizados para normalizar a escala, o dataset foi considerado apto para as etapas subsequentes. Os efeitos das transformações nos dados podem ser vistos na Fig. 7.

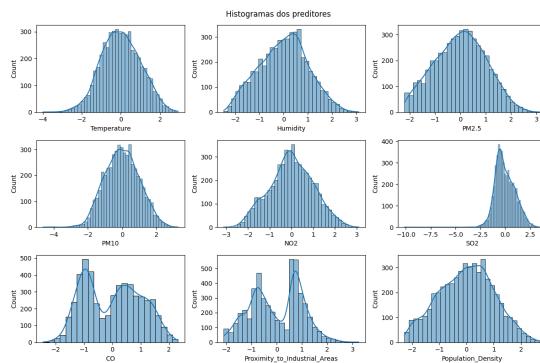


Fig. 7. Histogramas dos preditores transformados

### D. Análise PCA

A partir do que foi discutido na sessão II-D acerca de PCA e dos passos descritos em II-D, podemos perceber pela Fig.8 que os dois maiores PC's possuem uma contribuição total de 70% de variância acumulada. Logo, utilizamos o PC1 e o PC2 que já estão ordenados de forma decrescente, de maior contribuição para a menor.

A contribuição de cada *feature* para os PC's correspondentes pode ser visualizada por meio dos *loadings* na representação dos dados em menor dimensionalidade, como pode ser visto na Fig.9. É possível aferir que as variáveis *Proximity\_to\_Industrial\_Areas* e *CO* possui uma forte contribuição para a PC1, já para a PC2, os preditores *PM2.5* e *PM10* se destacaram.

Os gráficos contendo as contribuições absolutas de maneira mais detalhada estão presentes no repositório do projeto.

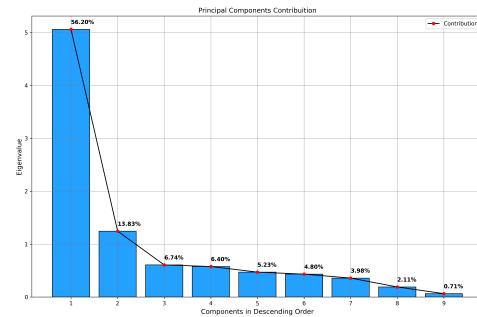


Fig. 8. Scree plot da contribuição de cada PC

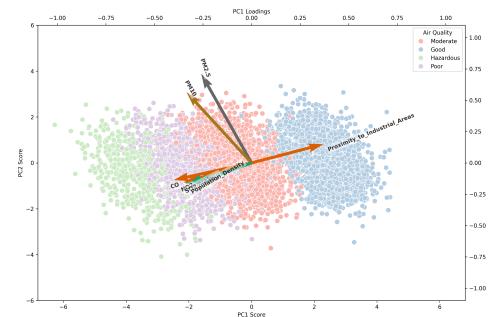


Fig. 9. Biplot contendo as direções de cada *feature* e a distribuição

### E. Outliers

No conjunto de dados normalizado, utilizando um nível de significância de 1% e com 9 graus de liberdade na distribuição  $\chi^2$ , foram encontrados 140 *outliers* em 5000 amostras. Este nível de significância indica que o critério de corte rejeita aproximadamente o 1% dos pontos mais extremos da distribuição.

Já na abordagem aplicada sobre os componentes principais (pós-PCA), mantendo o mesmo nível de significância porém agora com 2 graus de liberdade, detectaram-se 17 *outliers*, conforme ilustrado na Fig. 10. Essa redução no número

de *outliers* é consistente com o efeito da projeção em um subespaço de menor dimensionalidade, o qual tende a atenuar variações espúrias e ruídos presentes nas variáveis originais.

É válido ressaltar que a quantidade de *outliers* detectados foi relativamente baixa em relação ao total de amostras devido ao rigoroso critério de corte escolhido (1%). Caso seja de interesse identificar um conjunto maior de anomalias, faz-se necessário o uso de um nível de significância maior, como 5% por exemplo.

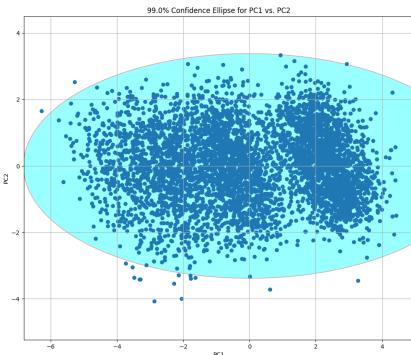


Fig. 10. Elipse de outliers após PCA

#### IV. CONCLUSÃO

Neste estudo, foi possível inferir como diferentes fatores interferem na qualidade do ar. Constatamos que as distribuições dos preditores eram, em geral, positivamente assimétricas. Além disso, a análise de correlação indicou que algumas variáveis apresentam uma correlação positiva, enquanto outras demonstram uma correlação negativa. Observou-se que altas concentrações de certos fatores são comuns em áreas com qualidade do ar danosa, ao passo que outros fatores parecem estar associados a uma melhoria nas condições.

O uso de técnicas de normalização se demonstrou essencial, possibilitando que pudéssemos aplicar métodos para diminuição de dimensionalidade e detecção de outliers. Ao aplicar a análise de componentes principais, conseguimos com sucesso reduzir o número de dimensões de 9 para 2 preservando 70% da variância dos dados.

Em síntese, os resultados obtidos neste estudo estabelecem uma base para trabalhos futuros voltadas ao desenvolvimento de modelos de aprendizado de máquina aplicados à análise da qualidade do ar. As informações geradas poderão ser utilizadas tanto em abordagens de regressão quanto de classificação, contribuindo para a construção de sistemas preditivos mais precisos e para o aprimoramento de estratégias de monitoramento ambiental.

#### REFERENCES

- [1] M. Kampa and E. Castanas, "Human health effects of air pollution," *Environmental Pollution*, vol. 151, no. 2, pp. 362–367, 2008, proceedings of the 4th International Workshop on Biomonitoring of Atmospheric Pollution (With Emphasis on Trace Elements). [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0269749107002849>
- [2] M. Owda, A. Y. Owda, and M. Fasli, "An exploratory data analysis and visualizations of underprivileged communities diabetes dataset for public good," in *2023 IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, 2023, pp. 581–585.
- [3] B. N. Vamshi, "Air quality analysis," *Journal of Data Analysis*, 07 2021.
- [4] W. Huang, T. Li, J. Liu, P. Xie, S. Du, and F. Teng, "An overview of air quality analysis by big data techniques: Monitoring, forecasting, and traceability," *Information Fusion*, vol. 75, pp. 28–40, 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1566253521000658>
- [5] C. Andrade, "Z scores, standard scores, and composite test scores explained," *Indian J. Psychol. Med.*, vol. 43, no. 6, pp. 555–557, Nov. 2021.
- [6] J. Osborne, "Improving your data transformations: Applying box-cox transformations as a best practice," *Pract Assess Res Eval*, vol. 15, pp. 1–9, 01 2010.
- [7] I.-K. Yeo and R. A. Johnson, "A new family of power transformations to improve normality or symmetry," *Biometrika*, vol. 87, no. 4, pp. 954–959, 2000. [Online]. Available: <http://www.jstor.org/stable/2673623>
- [8] A. Dürre, D. Vogel, and R. Fried, "Spatial sign correlation," *Journal of Multivariate Analysis*, vol. 135, pp. 89–105, 2015. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0047259X1400267X>
- [9] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philos. Trans. A Math. Phys. Eng. Sci.*, vol. 374, no. 2065, p. 20150202, Apr. 2016.
- [10] A. Maćkiewicz and W. Ratajczak, "Principal components analysis (pca)," *Computers Geosciences*, vol. 19, no. 3, pp. 303–342, 1993. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/009830049390090R>
- [11] P. J. Rousseeuw and B. C. van Zomeren, "Unmasking multivariate outliers and leverage points," *Journal of the American Statistical Association*, vol. 85, no. 411, pp. 633–639, 1990. [Online]. Available: <http://www.jstor.org/stable/2289995>