UNIVERSIDADE FEDERAL DE ALFENAS

BRUNO FELIPE ZANARDO

APLICAÇÕES DE ESTATÍSTICA MULTIVARIADA EM ANÁLISE DE DADOS EXPERIMENTAIS

POÇOS DE CALDAS/MG

2024

BRUNO FELIPE ZANARDO

APLICAÇÕES DE ESTATÍSTICA MULTIVARIADA EM ANÁLISE DE DADOS EXPERIMENTAIS

Dissertação apresentada como parte dos requisitos para obtenção do título de Mestre em Física pelo Programa de Pós-graduação em Física da Universidade Federal de Alfenas. Área de concentração: Física de Partículas e Campos.
Orientador: Prof. Dr. Cássius Anderson Miquele de Melo

POÇOS DE CALDAS/MG

2024

Sistema de Bibliotecas da Universidade Federal de Alfenas
Biblioteca Campus Poços de Caldas

Zanardo, Bruno Felipe.
Aplicações de estatística multivariada em análise de dados experimentais / Bruno Felipe Zanardo. - Poços de Caldas, MG, 2023.
88 f. : il. -

Orientador(a): Cássius Anderson Miquele de Melo.
Dissertação (Mestrado em Física) - Universidade Federal de Alfenas, Poços de Caldas, MG, 2023.
Bibliografia.

1. Estatística multivariada. 2. Análise de componentes principais. 3. Análise de correlação canônica. 4. Análise de componentes principais com incertezas experimentais. I. Melo, Cássius Anderson Miquele de, orient. II. Título.

Ficha gerada automaticamente com dados fornecidos pelo autor.

BRUNO FELIPE ZANARDO

APLICAÇÕES DE ESTATÍSTICA MULTIVARIADA EM ANÁLISE DE DADOS EXPERIMENTAIS

O Presidente da banca examinadora abaixo assina a aprovação da Dissertação apresentada como parte dos requisitos para a obtenção do título de Mestre em Física pela Universidade Federal de Alfenas. Área de concentração: Física de Partículas e Campos

Aprovada em: 25 de agosto de 2023.

Prof. Dr. Cássius Anderson Miquele de Melo Presidente da Banca Examinadora
Instituição: Universidade Federal de Alfenas

Profa. Dra. Iara Tosta e Melo
Instituição: Universitá Degli Studi di Catania

Prof. Dr. Gustavo do Amaral Valdiviesso Instituição: Universidade Federal de Alfenas

Dedico este trabalho, à minha família por sempre me apoiarem em todos os momentos de minha vida.

AGRADECIMENTOS

Agradeço aos meus pais, Ana e Marcos por sempre me darem todo o suporte necessário e que nunca me deixaram esquecer a importância que os estudos e a educação têm como poder de transformação na vida das pessoas.
A minha companheira de longa data, Tatyane, que sempre esteve ao meu lado e que com certeza ´e um grande estímulo para o meu desenvolvimento constante.
Ao professor Fernando Gardim, ao ex-aluno do mestrado de física Yuri Schneider e ao Leandro Henrique Pereira pela parceria e suporte com relação aos dados utilizados neste trabalho.
Ao meu orientador professor Cássius pelos ensinamentos, profissionalismo, companheirismo, disciplina e paciência que teve comigo durante os últimos anos. Sem dúvida, sua atuação foi um fator decisivo para o t´término desse projeto e para meu sucesso profissional.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001

RESUMO

A estatística multivariada é um ramo da estatística responsável por estudar situações em que se tem múltiplas variáveis e seus métodos podem ser aplicados nas mais diversas áreas do conhecimento auxiliando na tomada de decisão, isso porque seus métodos possuem como principais benefícios a redução de dimensionalidade do modelo estudado, tornando-o menos complexo, além de serem utilizados na construção de ´índices, classificação, associação entre as variáveis e inferência estatística. Neste trabalho, métodos de estatística multivariada foram aplicados em três situações distintas. Nas duas primeiras foi aplicado o método de correlação canônica e na terceira o método de análise de componentes principais (PCA) com incertezas experimentais. Sendo a primeira aplicação referente a uma análise socioambiental onde foi analisado a existência de uma correlação entre o índice de desenvolvimento humano (IDH) e seus subíndices com relação ao consumo de água e geração de esgoto dos municípios brasileiros. Enquanto a segunda análise está relacionada à Física de altas energias envolvendo a colisão de íons pesados de chumbo Pb-Pb. A terceira situação é referente a aplicação de PCA para a redução de dimensionalidade de um modelo de caracterização do meio interestelar. Como resultado foi possível gerar um modelo capaz de correlacionar o IDH com o consumo de água e a geração de esgoto com uma correlação canônica de 62,4% capaz de representar todo o país e um segundo modelo direcionado apenas para o estado de São Paulo com uma correlação canônica de 83%. Para o segundo cenário, envolvendo a colisão de íons pesados uma correlação canônica de 99,9% foi obtida, ratificando a correlação existente entre a entropia e o número de partículas carregadas. Uma segunda correlação proveniente do segundo par de variáveis canônicas retornou uma correlação também elevada com 96%, porém, neste modelo o momento transversal ficou com o maior peso canônico, podendo ser calculado a partir das demais variáveis estudadas como centralidade, energia e entropia. Com relação a análise de componentes principais foi possível reduzir o número de variáveis utilizadas na explicação do modelo de forma significativa, passando de 23 variáveis originais para 8 componentes principais, além de identificar que ao considerar a incerteza experimental durante a análise obtemos maior segurança com relação ao número de variáveis utilizadas para explicar o modelo.

Palavras-chaves: estatística multivariada; correlação canônica; componentes principais; análise socioambiental; física de altas energias; meio interestelar.

ABSTRACT

Multivariate statistics is a branch of statistics responsible for studying situations where statistics models presents more then one variable and their methods can be applied in the most diverse areas of knowledge assisting in decision making, because their methods have as main benefits the reduction of dimensionality of the model studied, making it less complex, in addition to being used in the construction of indexes, classification, association between variables and statistical inference. In this work Multivariate statistics methods were applied in tree different situations, the method of canonical correlation analysis was applied in two different situations and the principal component analysis with experimental errors method was applied in the last one. the first application referring to a socio-environmental analysis where the existence of a correlation between the human development index (HDI) and its sub-indices in relation to water consumption and sewage generation of Brazilian cities was analyzed. While the second analysis is related to high-energy physics involving the collision of heavy lead ions Pb-Pb. The third situation refers to the application of principal component analysis for the dimensionality reduction of a model of characterization of the interstellar medium. As a result, it was possible to generate a model capable of correlating the HDI with water consumption and sewage generation, with a canonical correlation of 62.4%, capable of representing the whole country, and a second model, directed only to the state of Sa˜o Paulo, with a canonical correlation of 83%. For the second scenario, involving the collision of heavy ions a canonical correlation of 99.9% was obtained, confirming the existing correlation between entropy and the number of charged particles. A second correlation, calculated from the second pair of canonical variables, returned a high correlation with 96%, however, in this model the cross-sectional moment had the highest canonical weight, and can be calculated from the other variables studied such as centrality, energy and entropy. Regarding the principal component analysis, it was possible to reduce the number of variables used in the explanation of the model significantly, from 23 original variables to 8 main components, in addition to identifying that when considering the experimental errors during the analysis we obtain greater security regarding the number of variables used to explain the model.

Key-words: multivariate statistics; canonical correlation; socio-environmental analysis; high-energy physics; principal component analysis; difuse interestelar bands.

LISTA DE TABELAS

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE QUADROS

SUMÁRIO

1 INTRODUÇÃO

A utilização dos dados vem se tornando cada vez mais frequente durante o processo de tomada de decisões nas mais diversas áreas do conhecimento e com o advento de novas tecnologias mais dados são gerados a cada dia e maior ´e o poder computacional para o tratamento deles.

A partir da aliança da estatística com a computação é possível realizarmos o processamento e análise de uma quantidade significativa de dados. Dentre as diversas áreas da estatística, os métodos relacionados a Estatística Multivariada tornam-se indispensáveis para tais análises.

A Estatística Multivariada é aplicável para os casos onde se tem múltiplas variáveis para uma mesma medição e é composta por um conjunto de métodos, que segundo [3] de forma geral possuem como objetivos, simplificar e facilitar a interpretação do modelo estudado a partir da criação de índices ou variáveis responsáveis por representar os dados originais do modelo; construir grupos de elementos que possuam similaridade entre si; investigar as relações de dependência entre as variáveis respostas associadas ao modelo e variáveis explicativas; comparar populações ou validar suposições a partir de testes de hipótese.

A análise de componentes principais e a análise de correlações canônicas são alguns dos métodos abordados pela Estatística Multivariada e que serão abordados neste trabalho.

Com relação a aplicabilidade da Estatística Multivariada, a mesma pode ser utilizada para a construção de índices, classificação, associação entre variáveis categóricas e inferência estatística.

Um exemplo prático pode ser observado na aplicação da análise de componentes principais sobre um conjunto de dados referente as variáveis que descrevem a migração de pássaros no intuito de reduzir a dimensão desse modelo através das componentes principais geradas por esse método. A PCA também pode ser aplicada para classificar qual o melhor solo para determinada cultura tendo como variáveis do modelo os elementos físico-químicos presentes no solo, [4] utilizou a PCA para a compressão de imagens digitais na medicina e [5] integrou essa técnica com ambientes de Data Warehouse no intuito de facilitar a caracterização e redução de dimensionalidade.

Outro exemplo se dá através da aplicação da análise de correlações canônicas para avaliarmos os resultados obtidos por um conjunto de estudantes no vestibular e seu desempenho durante o primeiro semestre de faculdade, descobrindo dessa forma se boas notas obtidas no vestibular podem ser traduzidas em um bom desempenho durante os primeiros meses da vida universitária do estudante.

Para termos uma ideia de quão amplo pode ser o uso da CCA nas diferentes áreas do conhecimento, temos a sua utilização pela [6] em seu trabalho para prever o tempo de vida do câncer de mama e [7] mostrou a efetividade em medir o nível de bem-estar de um adulto saudável através de medidas comportamentais e de cognição.

Porém, sabemos pela física que nenhuma medida ´e absoluta tendo uma incerteza experimental associada à sua medição e nenhum dos métodos da Estatística multivariada considera tais incertezas em seus cálculos.

Conforme descrito por [8] Taylor (2012, apud Flausino, 2018) a utilização da incerteza durante a análise e interpretação dos resultados se faz necessária, uma vez que dados com incertezas menores trazem maior confiabilidade e devem representar um peso maior quando comparados a dados com incertezas maiores. Dessa forma, conjuntos de dados, cujas incertezas experimentais não são conhecidas pode ser inutilizáveis devido à falta de confiança em seus valores.

Em seu trabalho, [2] desenvolve modificações de três métodos da Estatística Multivariada, de forma que as incertezas experimentais sejam levadas em consideração, esses métodos são a análise de componentes principais, análise de correlação canônica e a análise discriminante.

Devido à grande importância que a incerteza experimental possui na análise dos dados, esse trabalho tem como objetivo a aplicação dos métodos clássicos de Estatística Multivariada, PCA e CCA além dos métodos de PCA modificado de acordo com a metodologia proposta por [2] e realizar uma análise de seus resultados buscando identificar variações que justifiquem a mudança na tomada de decisão ao se aplicar o m´método modificado em relação ao m´método clássico.

Dessa forma o método de Análise de Componentes Principais modificado foi aplicado a um conjunto de 30 medições contendo 23 variáveis responsáveis pela explicação das bandas interestelares difusas conhecidas do inglês como (DIBs) no intuito de reduzir a dimensionalidade e identificar quais variáveis apresentam maior representatividade do modelo.

Realizamos a Análise de Correlação Canônica clássica em duas situações. Primeiro, em uma análise de interesse socioambiental, para identificar a correlação entre os conjuntos de dados referentes a geração de resíduo solido, geração de esgoto, consumo de água e os índices de IDH. E em um segundo momento a aplicação do método de CCA foi voltado para a física de altas energias envolvendo dados provenientes de simulações de colisão de partículas, resultantes da iteração forte entre quarks e gluóns, com variação da centralidade para cada simulação, tendo como objetivo identificar a quão significativa era a centralidade assim como qual o grau de significância das demais variáveis para o modelo.

2 ANÁLISE DE CORRELAÇÃO CANÔNICA

Dentre as diversas técnicas de estatística multivariada temos a análise de correlação canônica, do inglês *Canonical Correlatin Analysis* (CCA). Esta técnica, diferentemente das demais, realiza o estudo entre dois conjuntos de variáveis. Assim a CCA tem como finalidade encontrar a máxima correlação linear entre dois vetores que representam os conjuntos de variáveis estudados [9]. Para isso são realizadas as combinações lineares de cada conjunto analisado de forma que cada uma dessas combinações seja expressa a partir de uma única variável chamada de variável canônica. A correlação entre essas variáveis canônicas nos indica o grau de associação entre os conjuntos estudados e recebe o nome de correlação canônica [3].

A análise de correlação canônica pode ser utilizada tanto para redução de dimensionalidade de um modelo estatístico, determinando quais variáveis são mais relevantes na análise, quanto para identificar o grau de associação entre os conjuntos estudados que formam tal modelo [10]. O modelo teórico e a construção das variáveis canônicas sãos apresentados por [3], [10]) e [2] conforme abaixo. Para dois grupos de variáveis aleatórias, teremos os respectivos vetores para representá-las, $x = [x_{1,}x_{2,}..., x_p]^T$ de dimensão p × 1 e $y = [y_{1,}xy_{2,}..., y_p]^T$ de dimensão q × 1. As variâncias podem ser expressas conforme abaixo:

$$Cov(X) = \Sigma_{xx} \tag{2.1}$$

$$\mathrm{Cov(X,Y)} = \Sigma_{xy} \tag{2.2}$$

$$\mathrm{Cov(Y,X)} = \Sigma_{yx} \tag{2.3}$$

$$\mathrm{Cov(Y)} = \Sigma_{yy} \tag{2.4}$$

A partir de combinações lineares damos origem as variáveis canônicas U e V no intuito de resumir as associações existentes entre x e y em vez de usar as p × q covariâncias em Σ_{xy}

$$U = a^T x \tag{2.5}$$

$$U = b^T y \tag{2.6}$$

Onde a e b são vetores constantes de dimensões p×1 e q×1 provenientes da máxima correlação entre as variáveis canônicas U e V, a qual pode ser representada conforme abaixo:

$$Corr(U,V) = \frac{Cov(U,V)}{Var(U)\,Var(V)} \tag{2.7}$$

Sendo Cov (U,V) a covariância entre as variáveis canônicas e Var (U), Var (V) as variâncias.

Na análise de correlação canônica o número de pares de variáveis canônicas é limitado pelo menor números de variáveis de um dos conjuntos de variáveis estudado, de forma que o segundo par de variáveis canônicas é encontrado através da maximização da correlação entre U e V no conjunto das combinações lineares entre x e y que não são correlacionadas com o primeiro par. Assim, para uma análise contendo K variáveis canônicas, o k-ésimo par de variáveis canônicas pode ser definido como:

$$U_k = {a_K}^T x \tag{2.8}$$

$$V_k = {b_K}^T y \tag{2.9}$$

Os autovetores que maximizam a correlação entre U e V podem ser obtidos através da resolução do sistema de equações abaixo:

$$\left(\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx} - \lambda_k\Sigma_{xx}\right) a_K = 0 \tag{2.10}$$

$$\left(\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy} - \lambda_k\Sigma_{yy}\right) b_K = 0 \tag{2.11}$$

Dessa forma, λ_K ´e o k-ésimo maior autovalor das matrizes

$$(\Sigma_{xx}^{-1}\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}) \text{ e } (\Sigma_{yy}^{-1}\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}) \tag{2.12}$$

A correlação canônica é dada pela correlação em valor absoluto entre U_K e U_K e pode ser calculada como

$${p_k}^2 = \lambda_k \tag{2.13}$$

A análise correlação canônica segue os mesmos critérios da correlação de Pearson[1], ou seja, uma correlação de 1 significa que dois conjuntos de variáveis possuem uma correlação perfeita positiva, sendo está a máxima correlação atingível, caso essa correlação seja de -1 então os conjuntos de variáveis apresentam uma correlação perfeita negativa, isso significa que os dois conjuntos estão correlacionados, porém de forma oposta. Para o caso de o coeficiente de correlação ser igual a 0 então não há correlação linear alguma entre os grupos de variáveis podendo nesse caso haver outro tipo de correlação que não a linear.

Assim, quanto mais próximo de 1 e -1 for a correlação canônica mais correlacionados positivamente e negativamente os grupos de variáveis estudados estarão e quanto mais próximos de 0 menor a correlação entre eles. Sendo uma correlação aceitável algo relativo que dependerá muito da área de pesquisa, por exemplo, uma correlação de 0.7 ou 70% entre a aplicação de um procedimento cirúrgico em pacientes e a recuperação deles sem danos colaterais pode não fazer sentido uma vez que estamos lidando com vidas humanas e a margem de erro deve ser mínima. Porém se estivermos falando dessa mesma correlação em estudos de saneamento onde não há nenhuma outra opção disponível esses números podem ser utilizados para nortear as políticas públicas.

2.1 ANÁLISE DO CONSUMO DE ÁGUA E GERAÇÃO DE ESGOTO A PARTIR DO IDH

Nesta seção o método de CCA será plicado e discutido através da análise de dados relacionados ao consumo de água e geração de esgoto dos municípios e seus respectivos índices de desenvolvimento humano, IDH, e seus subíndices referentes a renda, longevidade e educação. Esse estudo é de extrema importância uma vez que, estabelecida uma correlação satisfatória entre os conjuntos de dados compostos pelas variáveis anteriormente descritas, tal correlação poderá ser utilizada para

[1] O coeficiente de correlação de Pearson é responsável por medir o grau da correlação linear entre duas variáveis quantitativas. É um índice adimensional com valores situados ente -1,0 e 1.0 inclusive, que reflete a intensidade de uma relação linear entre dois conjuntos de dados.

auxiliar nas políticas públicas voltadas para as questões de saneamento ambiental dos munícipios.

Primeiramente, foi realizado os levantamentos dos dados referentes aos indicadores de saneamento, consumo per capita de água e geração per capita de esgoto. Os dados foram obtidos através da plataforma do Sistema Nacional de Informação sobre Saneamento (SNIS) referente ao ano de 2010 em formato csv, o qual apresenta separadamente a população total atendida com abastecimento de água em número de habitantes e o volume em água consumido por essa população na unidade de 1000m3/ano. Sendo dessa forma necessário utilização da equação abaixo para o cálculo do consumo de água per capita em L/habitante.dia.

$$C_a = \frac{V_a .10^6}{P_a .365}, \tag{2.14}$$

onde Ca é o consumo de água per capita, P_a é a população total atendida pelo abastecimento de água e V_a é o volume de água consumido por ano por essa população.

De forma análoga foi realizado o cálculo da geração per capita de esgoto em L/habitante.dia conforme se segue.

$$G_e = \frac{V_e .10^6}{P_e .365}, \tag{2.15}$$

onde Ge é a geração de esgoto per capita, P_e é a população total atendida com esgotamento sanitário e V_e é o volume de esgoto coletado por ano.

Em um segundo momento foi realizado o levantamento dos dados referentes ao ´Indice de Desenvolvimento Humano, IDH, e seus subíndices através dos dados disponibilizados pelo instituto brasileiro de geografia e estatística, IBGE. Vale salientar, que devido a não realização da pesquisa censitária no ano de 2020, para este estudo foram utilizados os dados de IDH referentes a pesquisa censitária de 2010.

Com todos os dados necessários para o estudo disponíveis, foram retirados da base todos os municípios que não possuíam informações ou apresentavam valores zerados para o consumo de água e geração de esgoto, ficando assim com uma base

composta de 4930 cidades com informações de consumo de água e 1751 cidades com geração de esgoto.

Devido ao fato de serem encontrados dados na plataforma do SINIS com valores muito acima da média e implicando em possíveis medições ou lançamentos errôneos no sistema, um tratamento prévio foi realizado na base de dados de forma a remover os outliers[2]. Assim foi realizado um corte nas cidades que apresentaram valores de água ou esgoto distantes da ordem de três vezes o desvio padrão, restringindo a base às cidades que apresentassem até 300L/l.d−1 de consumo de água ou geração de esgoto. Com este tratamento é possível que alguns dados reais possam se perder, porém, devido ao tamanho da base espera-se que os resultados não sejam afetados de forma substancial.

O algoritmo para o cálculo da correlação canônica, assim como a construção dos gráficos de dispersão e calor, os quais serão apresentados neste trabalho, foram desenvolvidos com a utilização da linguagem de programação Python, o ambiente de desenvolvimento utilizado para a construção do código foi o Jupyter Notebook, pertencente ao Framework Anaconda. As bibliotecas pandas e numpy foram utilizadas para a construção do modelo estatístico como um todo, incluindo a construção de matrizes de covariância e correlação bem como a realização de cálculos matemáticos. Já para a análise exploratória e construção dos gráficos, foram utilizadas as bibliotecas Seaborn, Matplotlib e Plotly.

Os resultados aqui apresentados estão dispostos de tal forma que cada seção apresentará uma variável ou um conjunto de variáveis a serem explicadas. Tais variáveis serão referentes ao consumo de água, geração de esgoto ou as duas juntas. Em todas as situações o IDH se apresentará como variável explicadora e a depender da análise poderá ter em seu conjunto o consumo de água ou a geração de esgoto também.

Em todas as análises aqui apresentadas foram realizadas primeiramente uma análise exploratória dos dados, com seus resultados representados através de um gráfico de dispersão seguido por um gráfico de calor. Posteriormente a análise de correlação canônica foi aplicada obtendo-se os autovalores, autovetores e a correlação canônica entre os conjuntos.

---

[2] Observação anormal e extrema em uma amostra estatística ou série de dados que não é consistente com as demais medidas.

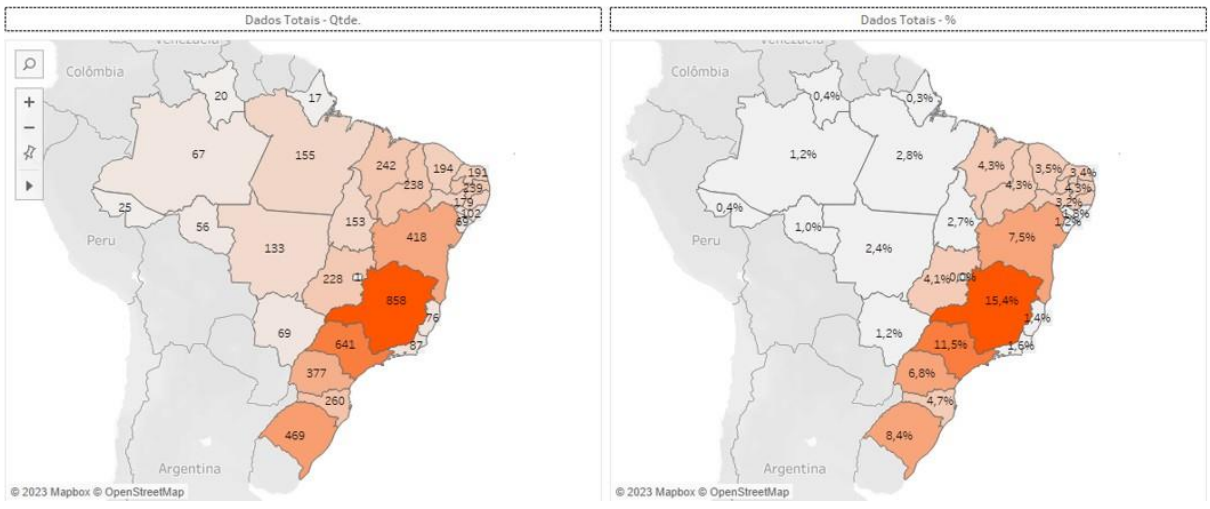
2.1.1 Análise Exploratória dos Dados

Ao longo das análises que irão se suceder nas próximas seções será possível notar uma variação com relação a volumetria dos dados utilizados em cada análise, isso ocorrera´ principalmente pelo fato de ter sido realizado um tratamento em algumas bases para que os outliers fossem retirados e pela restrição de informação de uma das variáveis utilizadas durante a análise com relação às outras variáveis, uma vez que a quantidade de dados para cada variável deve ser a mesma.

Esta seção tem como objetivo mostrar a evolução das mudanças das bases de dados com relação a quantidade de dados utilizados e evidenciar os outliers para que se tenha uma melhor compreensão dos tratamentos realizados nas bases.

Inicialmente foram levantados dados de IDH geral e seus subíndices de renda, educação e longevidade para um total de 5564 cidades disponibilizados pelo instituto brasileiro de geografia e estatística, IBGE, referente ao censo de 2010. A partir desses dados foi realizado uma análise exploratória composta por um mapa de calor mostrando a disposição geográfica dos municípios estudados, histogramas de cada variável assim como calculado a média e o desvio padrão.

Figura 1 – Mapas de calor - IDH Geral



Fonte: Autoria própria.
Legenda: Dados Totais – Qtde.: Distribuição geográfica dos dados em termos absolutos.
Dados Totais – %: Distribuição geográfica dos dados em termos percentuais.
Nota: As cores mais escuras representam as maiores quantidades de municípios e as mais claras as menores quantidades.

Através do mapa de calor acima é possível verificarmos que os estados com maior número de municípios em ordem decrescente são os de Minas Gerais, São Paulo, Rio Grande do Sul, Bahia e Paraná, representando juntos aproximadamente 50% de toda a base de dados. Abaixo segue a tabela contendo os dados estatísticos de todos os municípios apresentados no mapa de calor.

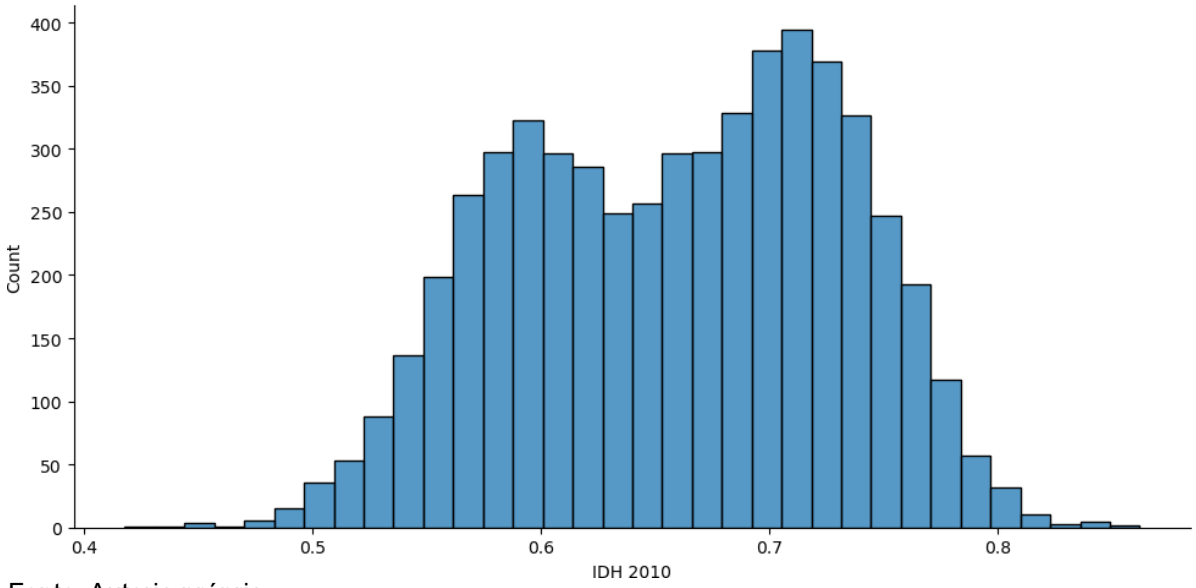
Tabela 1 – Análise exploratória dos dados de IDH

	IDH Geral	IDH Renda	IDH Longevidade	IDH Educação
Qtde.	5564	5564	5564	5564
Média	0,659	0,643	0,802	0,559
Desvio	0,072	0,081	0,045	0,093
Mín	0,418	0,400	0,672	0,207
25%	0,599	0,572	0,769	0,490
50%	0,655	0,654	0,808	0,560
75%	0,718	0,707	0,836	0,631
Máx	0,862	0,891	0,894	0,825

Fonte: Autoria própria.
Nota: Base completa contendo as informações de estatística descritiva referentes ao IDH geral, IDH Renda, IDH Longevidade e IDH educação de todos os 5564 municípios.

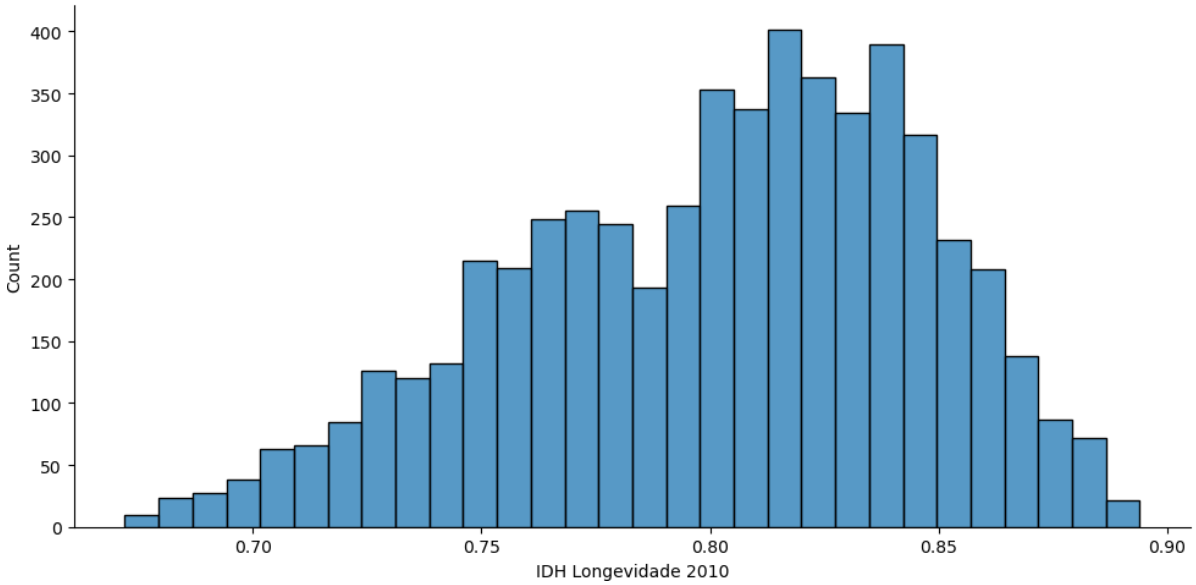
Através dos histogramas apresentados em seguida, é possível notar com clareza a presença de dois picos para os dados referentes à variável de IDH geral e IDH Renda, indicando não se tratar de uma distribuição gaussiana. Tal fato não interfere na aplicação do método de correlação canônica. Conforme será mostrado mais adiante nesta seção, a eliminação dos outliers contribuirá para deixar as distribuições mais simétricas.

Gráfico 1 – Histograma - IDH Geral



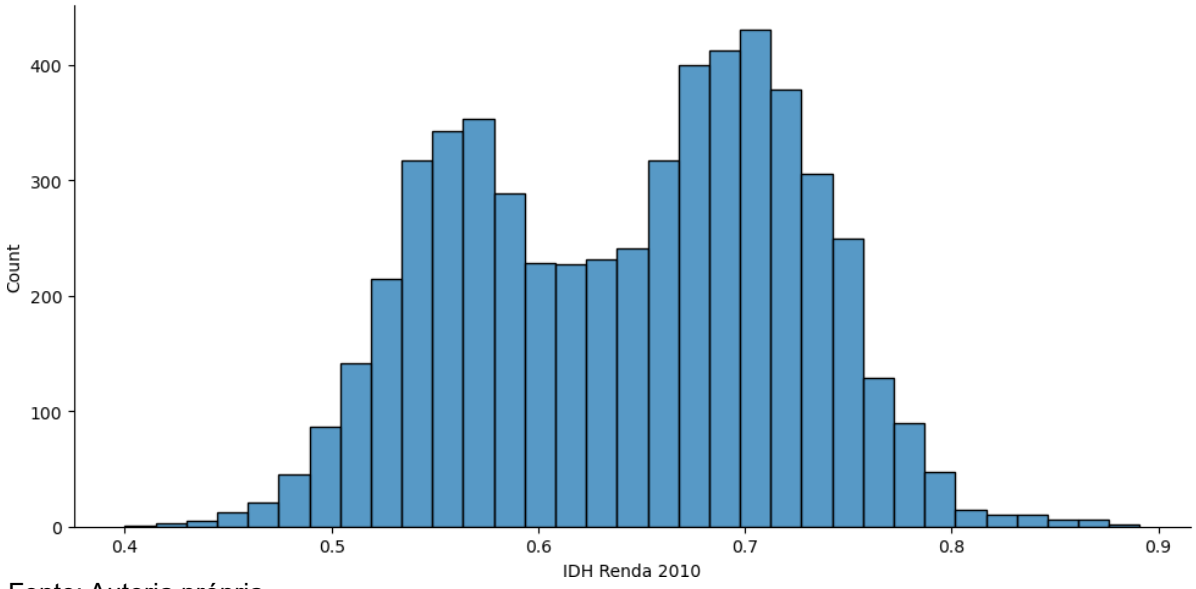
Fonte: Autoria própria.
Nota: Histograma correspondente aos dados de IDH geral de 2010 contendo as informações de todos os 5564 municípios.

Gráfico 2 – Histograma - IDH Longevidade



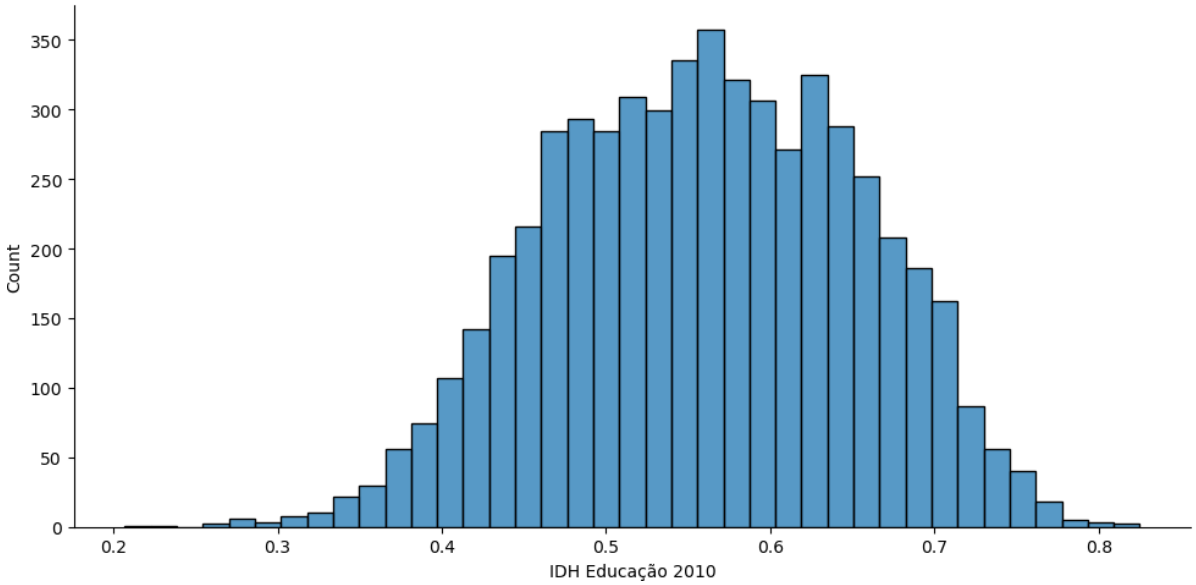
Fonte: Autoria própria.
Nota: Histograma correspondente aos dados de IDH Longevidade de 2010 contendo as informações de todos os 5564 municípios.

Gráfico 3 – Histograma - IDH Renda



Fonte: Autoria própria.
Nota: Histograma correspondente aos dados de IDH Renda de 2010 contendo as informações de todos os 5564 municípios.

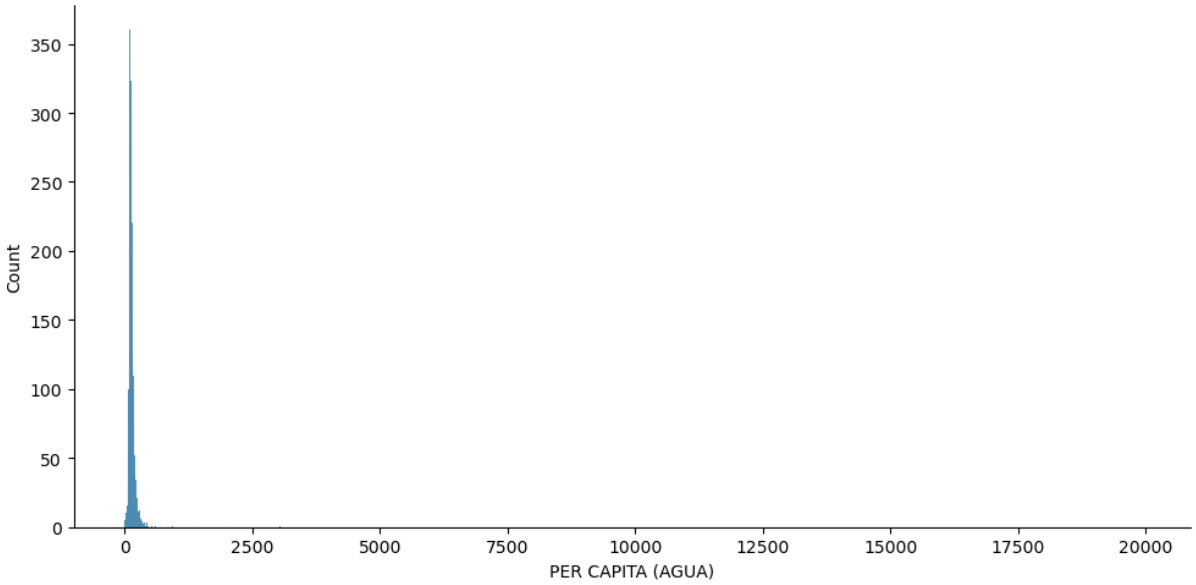
Gráfico 4 – Histograma - IDH Educação



Fonte: Autoria própria.
Nota: Histograma correspondente aos dados de IDH Educação de 2010 contendo as informações de todos os 5564 municípios.

Com relação a variável de consumo de água per capita foram levantados inicialmente dados para 4930 cidades. Segue abaixo o histograma e tabela contendo a média e desvio padrão para esses dados.

Gráfico 5 – Histograma - Água



Fonte: Autoria própria.
Nota: Histograma correspondente aos dados de consumo de água de todos os municípios que reportaram esse dado no Sistema Nacional de Informação sobre Saneamento para o ano de 2010 totalizando 4930 municípios.

Tabela 2 – Análise exploratória dos dados de consumo de água

	Água
Qtde	4930
Média	143
Desvio	324
Mín	2
25%	100
50%	122
75%	150
Máx	19882

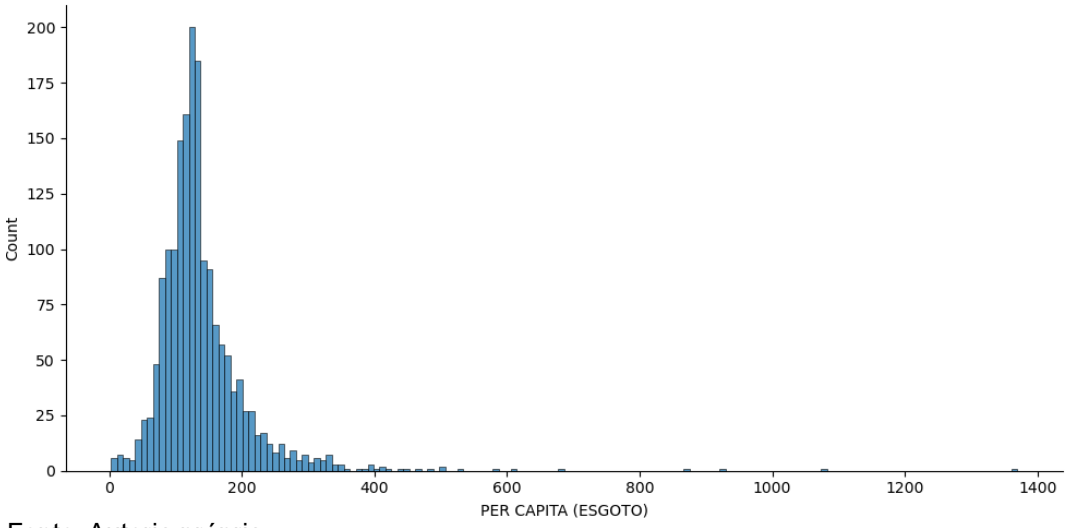
Fonte: Autoria própria.
Nota: Base completa contendo as informações de estatística descritiva referentes ao consumo de água per capita em litros/dia, de todos os 4930 municípios.

Tanto pelo histograma, quanto pela tabela contendo as informações estatísticas pode-se observar pontos muito acima da média, chegando a ter mais de 60 desvios

padrão de distância da média, indicando a presença de outliers, sendo o mais evidente o consumo de água per capita correspondente a 19881,9 litros que também é o nosso consumo máximo.

Para a variável de geração de esgoto per capita inicialmente foram levantados os dados para 1751 cidades. Segue abaixo Histograma e tabela contendo a média e o desvio padrão.

Gráfico 6 – Histograma - Esgoto



Fonte: Autoria própria.
Nota: Histograma correspondente aos dados de geração de esgoto de todos os municípios que reportaram esse dado no Sistema Nacional de Informação sobre Saneamento para o ano de 2010 totalizando 1751 municípios.

Tabela 3 – Análise exploratória dos dados de consumo de esgoto

	Esgoto
Qtde	1751
Média	140
Desvio	77
Mín	3
25%	103
50%	126
75%	157
Máx	1370

Fonte: Autoria própria.
Nota: Base completa contendo as informações de estatística descritiva referentes geração de esgoto de todos os 1751 municípios que apresentaram esse dado ao Sistema Nacional de Informação sobre Saneamento para o ano de 2010.

Assim como no caso do consumo de água, a geração de esgoto contendo todas as informações nos traz valores muito acima da média, sendo que a máxima geração de esgoto, com 1369,9 litros, esta´ a mais de 17 desvios padrão da média indicando outliers nessa base.

A presença de outliers nas bases de água e esgoto podem ser explicados em parte devido ao processo de lançamento e atualização desses dados no SNIS serem totalmente manuais, ocasionando assim possíveis erros de lançamento como vistos acima.

Devido aos *outliers* observados nas bases de água e esgoto, foi calculado a média e o desvio padrão utilizando os dados brutos e a partir desses valores foi realizado um tratamento na base eliminando todos os dados acima de três σ de distância da média. Possivelmente ainda teremos valores incorretos nas bases, porém considera-se que os valores estejam mais coerentes com a realidade das cidades após esses tratamentos.

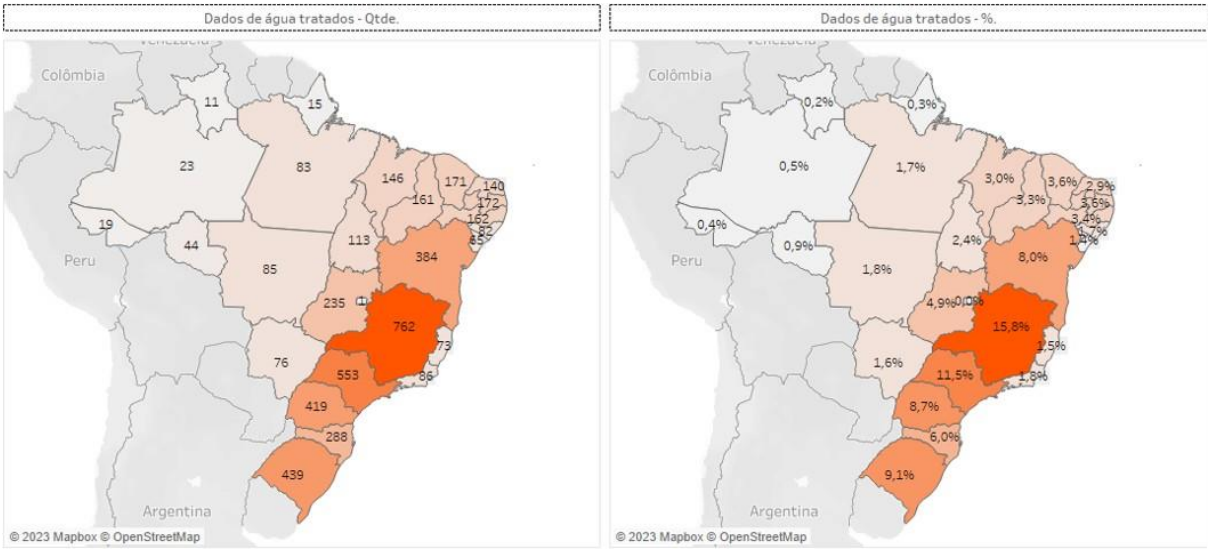
Após o tratamento dos dados de consumo de água a nova base apresentou informações referente a 4808 cidades, o que representa uma redução de 122 dados, correspondendo a aproximadamente 2,5% da base total de 4930 dados.

Já para os dados de IDH, a redução foi de 756 dados, correspondendo a aproximadamente 13,5% da base total de 5564 dados.

Dessa forma não é esperado que a base de IDH tenha um impacto relevante após o tratamento dos dados de água, uma vez que a redução da base não foi significativa.

Segue abaixo o novo mapa de calor referente a disposição geográfica dos 4808 municípios, o histograma para o consumo de água e a tabela com as informações de média e desvio padrão após o tratamento dos dados.

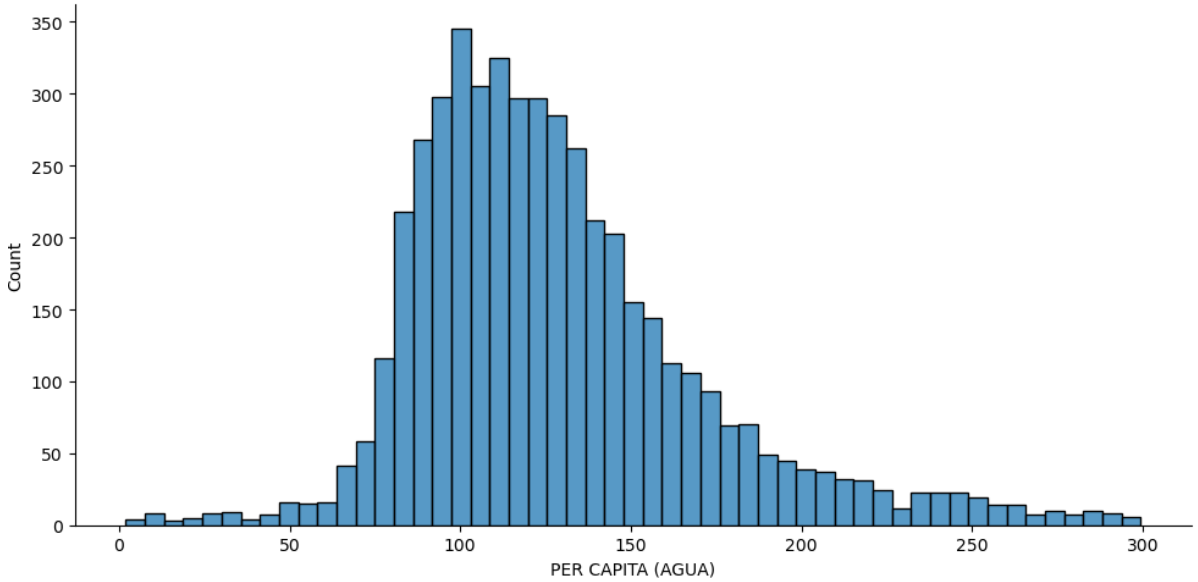
Figura 2 – Mapas de calor - Água



Fonte: Autoria própria.
Legenda: Dados de água tratados – Qtde: Distribuição geográfica dos dados em termos absolutos;
Dados de água tratados – %: Distribuição geográfica dos dados em termos percentuais.
Nota: Os mapas representam os dados de água após a retirada dos outliers de forma que as cores mais escuras representam as maiores quantidades de municípios e as mais claras as menores quantidades.

Após o tratamento dos dados de consumo de água pode-se observar que não houve uma mudança significativa com relação a representatividade dos estados na base com os estados de Minas Gerais e São Paulo ainda à frente com o maior número de municípios. Sendo que os cinco estados com os maiores números de municípios tiverem uma ligeira alta com relação ao percentual total, indo de aproximadamente 50% para cerca de 53% da base, enquanto os estados do norte e nordeste tiveram em sua grande maioria uma ligeira queda no percentual de representatividade.

Gráfico 7 – Histograma - Água



Fonte: Autoria própria.
Nota: Histograma correspondente aos dados de consumo de água de todos os municípios que reportaram esse dado no Sistema Nacional de Informação sobre Saneamento para o ano de 2010 após a retirada de outliers, totalizando 4808 municípios.

Tabela 4 – Análise exploratória dos dados de IDH e Água

	IDH Geral	IDH Renda	IDH Longevidade	IDH Educação	Água
Qtde.	4808	4808	4808	4808	4808
Média	0,660	0,643	0,801	0,561	128
Desvio	0,070	0,078	0,044	0,091	43
Mín	0,453	0,437	0,672	0,266	2
25%	0,601	0,573	0,769	0,492	99
50%	0,665	0,652	0,807	0,560	120
75%	0,717	0,706	0,836	0,629	147
Máx	0,862	0,891	0,894	0,811	299

Fonte: Autoria própria.
Nota: Base contendo as informações de estatística descritiva referentes ao IDH e consumo de água de 4808 municípios, após a retirada de outliers da base original.

Pela tabela apresentada é possível identificar que todos os pontos de mínimo referentes aos dados de IDH tiveram aumento, enquanto o todos os pontos de máximo, com exceção do IDH Educação, se mantiveram constantes. Sendo este um forte indicativo de que os outliers de água se encontravam nas cidades com menores IDH's.

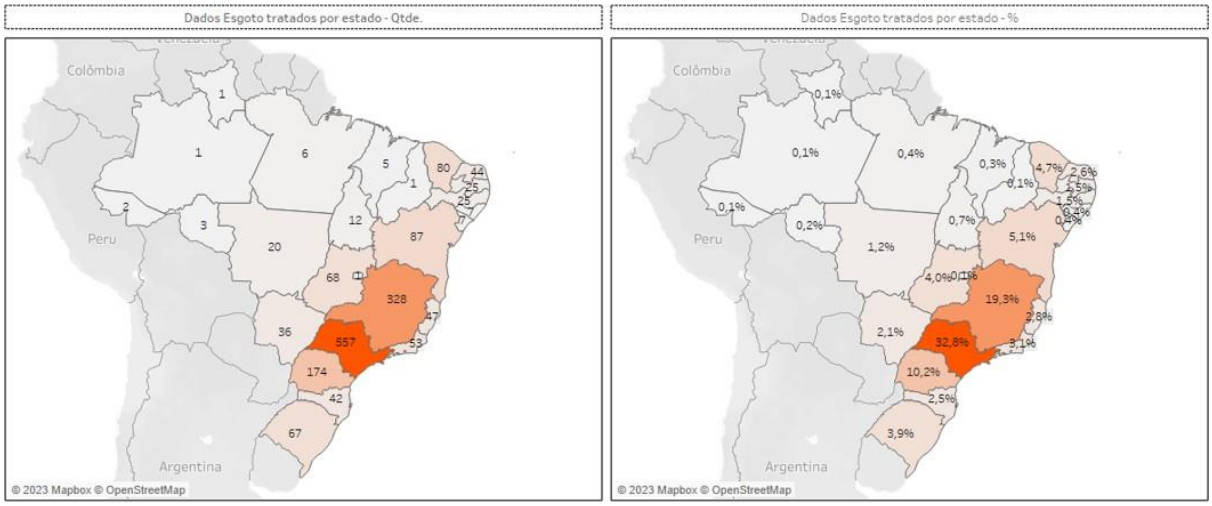
Com relação as médias apresentadas para o IDH geral e seus subíndices após o corte, não houve nenhuma mudança significativa, reforçando que a base de IDH não será impactada pela restrição dos dados ocasionado pelos dados de água.

Após o tratamento dos dados de geração de esgoto a nova base apresentou informações referente a 1699 cidades, o que representa uma redução de 52 dados, correspondendo a aproximadamente 3 % da base total de 1751 dados.

Devido à escassez de dados de geração de esgoto pela não atualização no SINIS por grande parte dos municípios é esperado uma mudança na distribuição dos dados de IDH com deslocamento das médias, uma vez que a base sofreu uma redução relevante, não devido ao tratamento dos outliers, que neste caso representam apenas 52 dados, mas sim pela falta de informação referente aos dados de esgoto no SINIS. A redução foi de 3813 dados, correspondendo a aproximadamente 69% da base total de 5564 dados.

Segue abaixo o novo mapa de calor referente a disposição geográfica dos 1699 municípios, histograma para a geração de esgoto e a tabela com as informações de média e desvio padrão após o tratamento dos dados.

Figura 3 – Mapas de calor - Esgoto

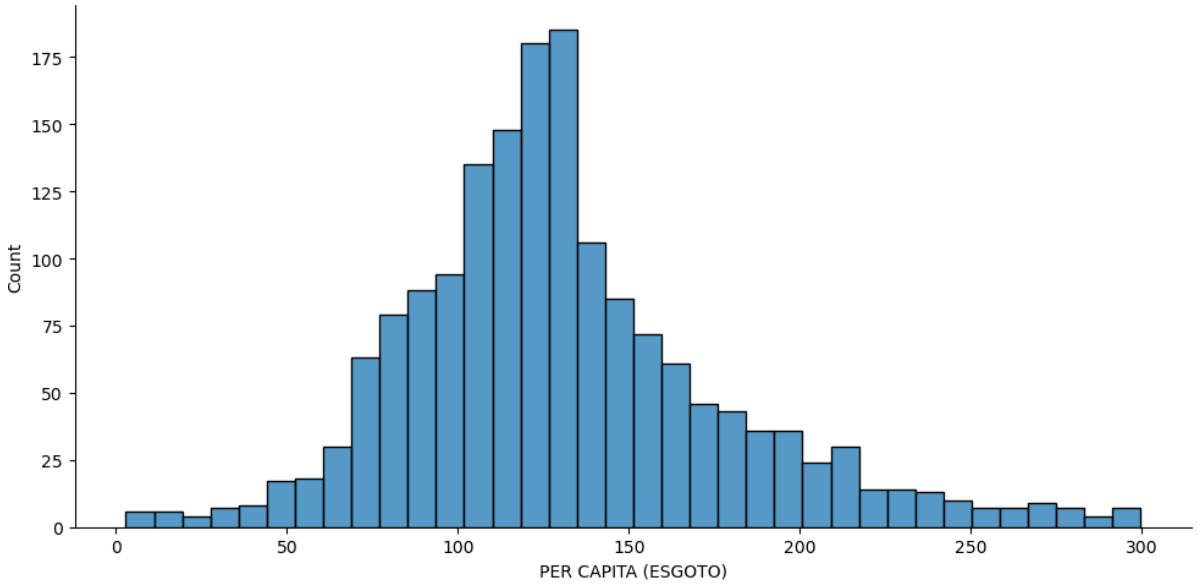


Fonte: Autoria própria.
Legenda: Dados Esgoto tratados por estado – Qtde: Distribuição geográfica dos dados em termos absolutos;
Dados Esgoto tratados por estado – %: Distribuição geográfica dos dados em termos percentuais.
Nota: Os mapas representam os dados de dados de Esgoto após a retirada dos outliers de forma que as cores mais escuras representam as maiores quantidades de municípios e as mais claras as menores quantidades, representando 1699 municípios.

Pelo mapa de calor acima é possível identificar uma redução com relação aos dados de geração de esgoto em todos os municípios de forma geral, com exceção do estado de São Paulo. É possível perceber também que essa redução é mais

perceptível nas regiões Norte e Nordeste o que explica a mudança de representatividade da base de dados com relação aos estados. Nessa nova configuração, os estados de São Paulo, Minas Gerais e Paraná representam mais de 60% de toda a base.

Gráfico 8 – Histograma – Esgoto



Fonte: Autoria própria.
Nota: Histograma correspondente aos dados de geração de esgoto após a retirada de *outliers* de todos os municípios que reportaram esse dado no Sistema Nacional de Informação sobre Saneamento para o ano de 2010 totalizando 1699 municípios.

Tabela 5 – Análise exploratória dos dados de IDH e Esgoto

	IDH Geral	IDH Renda	IDH Longevida	IDH Educação	Esgoto
Qtde.	1699	1699	1699	1699	1699
Média	0,700	0,683	0,821	0,613	131
Desvio	0,059	0,066	0,035	0,082	47
Mín	0,520	0,480	0,685	0,377	3
25%	0,663	0,646	0,803	0,588	102
50%	0,710	0,694	0,825	0,621	125
75%	0,741	0,727	0,846	0,675	153
Máx	0,862	0,891	0,894	0,811	300

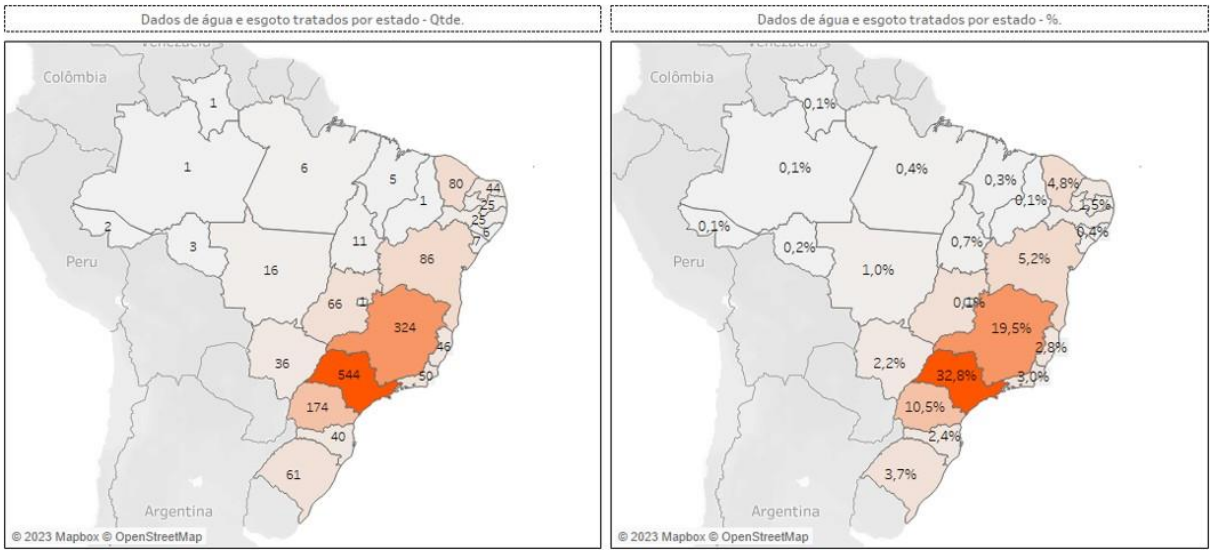
Fonte: Autoria própria.
Nota: Base contendo as informações de estatística descritiva referentes ao IDH e geração de esgoto de 1699 municípios, após a retirada de outliers da base original.

Assim como ocorreu após o tratamento dos dados de água, com os dados de esgoto também pode-se observar um aumento no mínimo para todos os IDH's e manutenção dos máximos com exceção do IDH Educação. Com relação às médias dos IDH's é possível verificar que todos tiveram um deslocamento para cima,

indicando que a maior parte dos municípios sem informação de esgoto apresentam menores IDH's.

Para a análise de correlação canônica que envolve as variáveis de consumo de água, geração de esgoto e os IDH's ao mesmo tempo temos uma intersecção entre as informações disponíveis de cada variável resultando em uma base com 1661 dados. Para esses novos dados foram gerados mapas de calor com a disposição geográfica dos dados, histogramas e cálculo da média e desvio padrão para todas as variáveis no intuito de verificar como os dados se comportam após o corte.

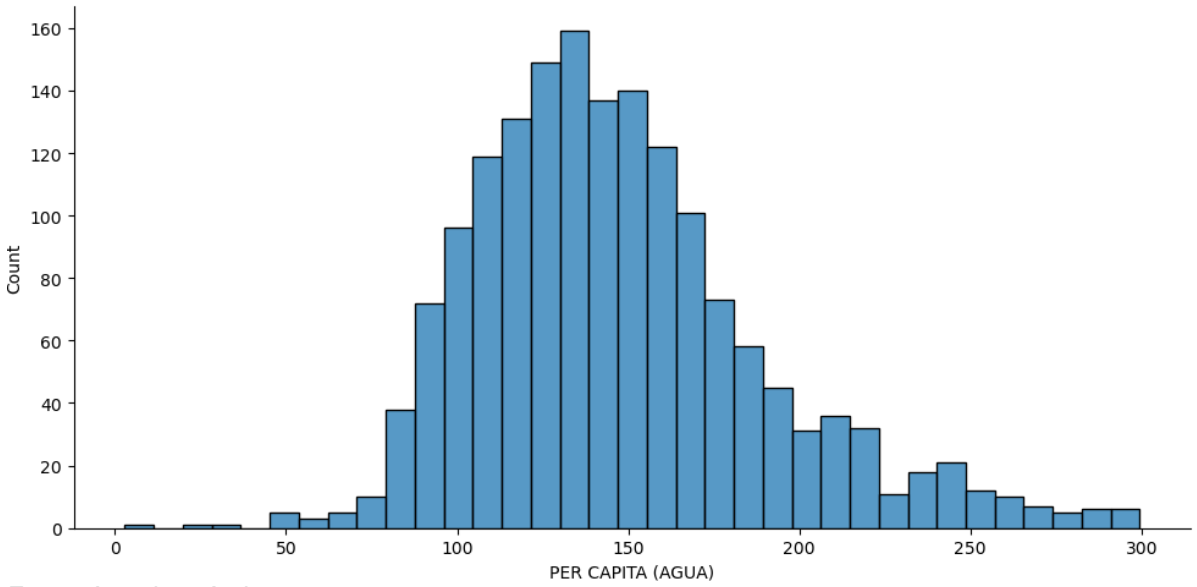
Figura 4 – Mapas de calor - Esgoto e Água



Fonte: Autoria própria.
Legenda: Dados de água e esgoto tratados por estado – Qtde: Distribuição geográfica dos dados em termos absolutos;
Dados de água e esgoto tratados por estado – %: Distribuição geográfica dos dados em termos percentuais.
Nota: Os mapas representam os dados de esgoto e água após a retirada dos outliers de forma que as cores mais escuras representam as maiores quantidades de municípios e as mais claras as menores quantidades.

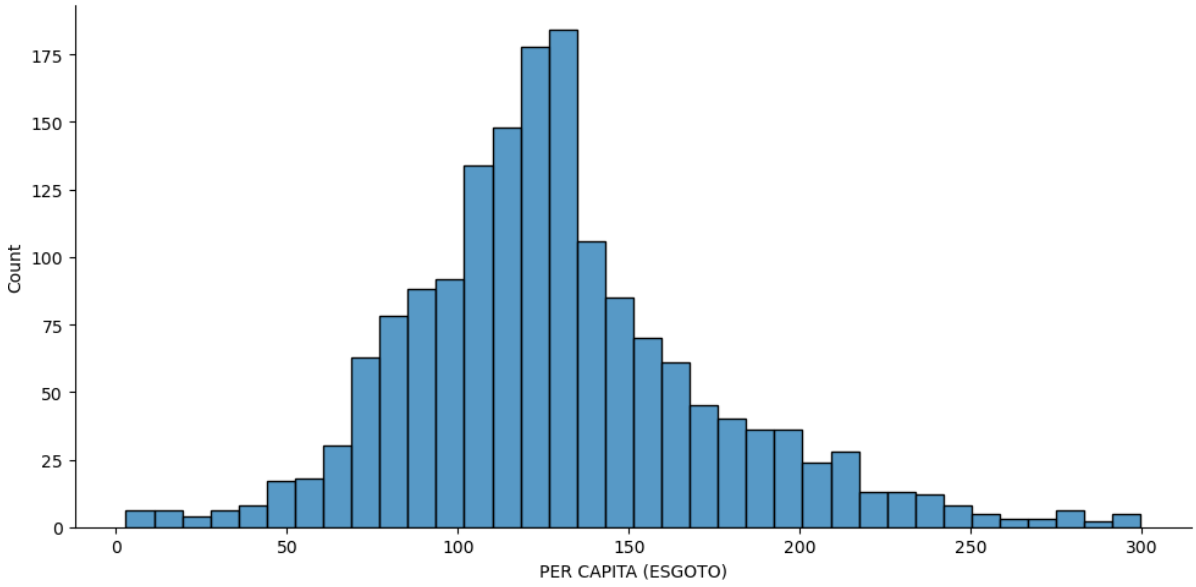
Devido à baixa variação dos dados, cerca de 2%, entre esta base envolvendo água e esgoto, 1661 dados, com relação a base discutida anteriormente com apenas esgoto, 1699 dados, nenhuma mudança significativa com relação a representatividade da base por estado foi observada, estando ainda responsáveis por representar mais de 60% da base os estados de São Paulo, Minas Gerais e Paraná.

Gráfico 9 – Histograma - Água



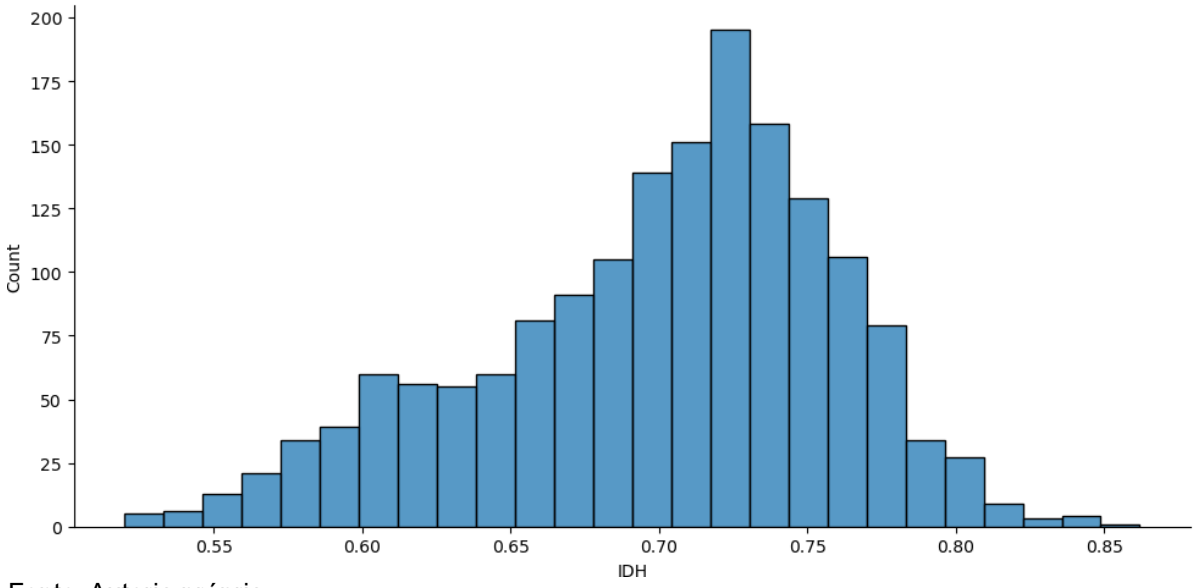
Fonte: Autoria própria.
Nota: Histograma correspondente ao consumo de água de todos os municípios que apresentaram dados tanto de consumo de água quanto para a geração de esgoto no Sistema Nacional de Informação sobre Saneamento para o ano de 2010 totalizando 1661 municípios.

Gráfico 10 – Histograma - Esgoto



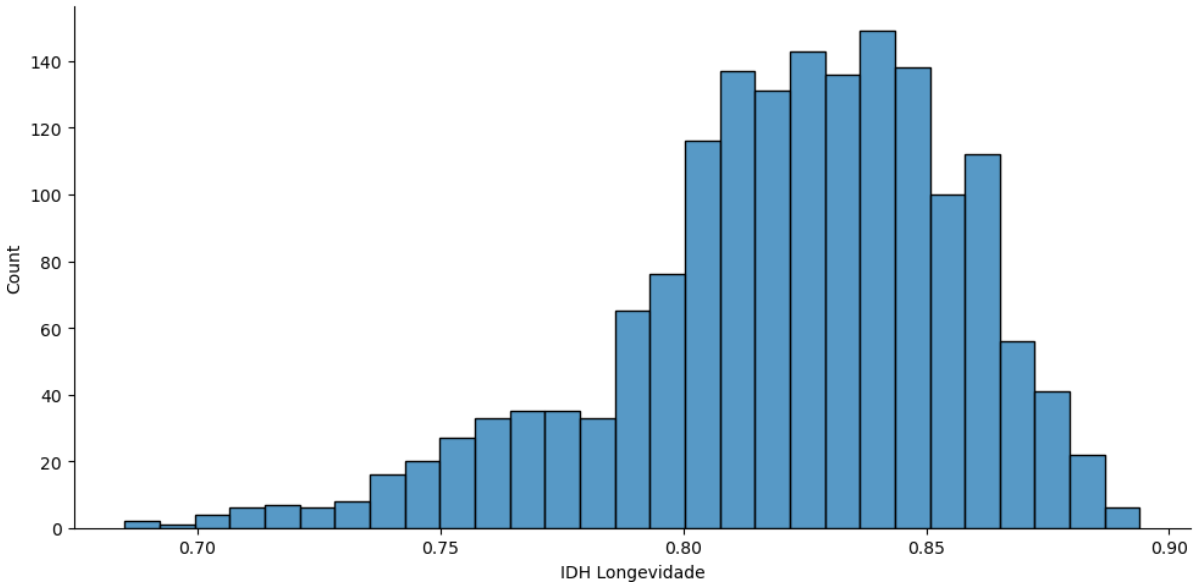
Fonte: Autoria própria.
Nota: Histograma correspondente a geração de esgoto de todos os municípios que apresentaram dados tanto de consumo de água quanto para a geração de esgoto no Sistema Nacional de Informação sobre Saneamento para o ano de 2010 totalizando 1661 municípios.

Gráfico 11 – Histograma - IDH Geral



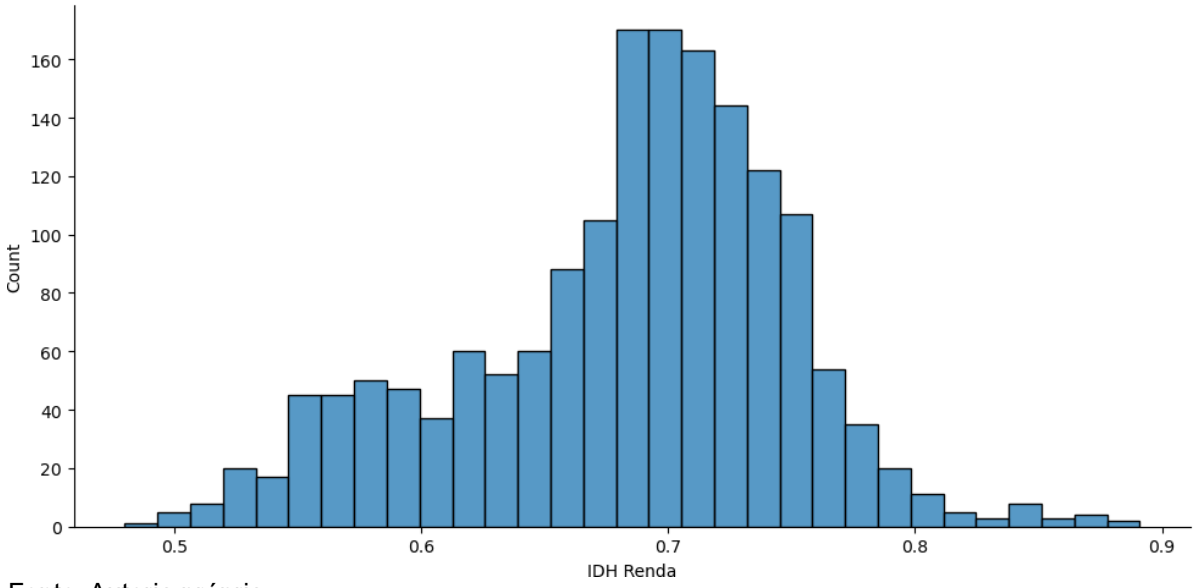
Fonte: Autoria própria.
Nota: Histograma correspondente aos dados IDH geral de todos os municípios que apresentaram dados tanto de consumo de água quanto para a geração de esgoto no Sistema Nacional de Informação sobre Saneamento para o ano de 2010 totalizando 1661 municípios.

Gráfico 12 – Histograma - IDH Longevidade



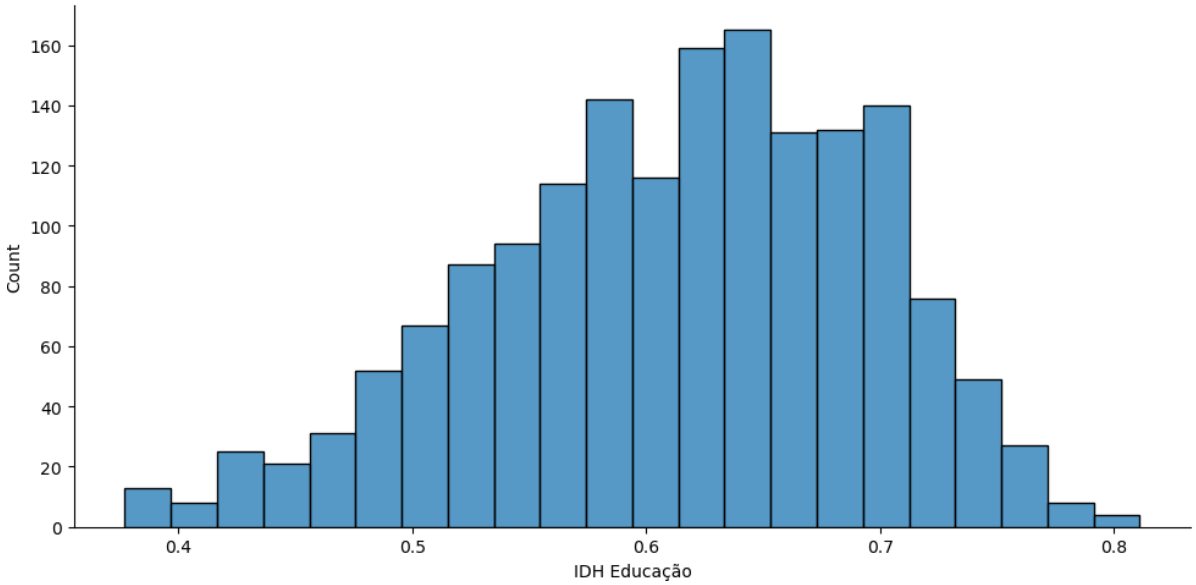
Fonte: Autoria própria.
Nota: Histograma correspondente aos dados IDH Longevidade de todos os municípios que apresentaram dados tanto de consumo de água quanto para a geração de esgoto no Sistema Nacional de Informação sobre Saneamento para o ano de 2010 totalizando 1661 municípios.

Gráfico 13 – Histograma - IDH Renda



Fonte: Autoria própria.
Nota: Histograma correspondente aos dados IDH Renda de todos os municípios que apresentaram dados tanto de consumo de água quanto para a geração de esgoto no Sistema Nacional de Informação sobre Saneamento para o ano de 2010 totalizando 1661 municípios.

Gráfico 14 – Histograma - IDH Educação



Fonte: Autoria própria.
Nota: Histograma correspondente aos dados IDH Educação de todos os municípios que apresentaram dados tanto de consumo de água quanto para a geração de esgoto no Sistema Nacional de Informação sobre Saneamento para o ano de 2010 totalizando 1661 municípios.

Tabela 6 – Análise exploratória dos dados de IDH, Água e Esgoto

	IDH Geral	Renda	Longevidade	Educação	Água	Esgoto
Qtde.	1661	1661	1661	1661	1661	1661
Média	0,699	0,683	0,821	0,613	147	129
Desvio	0,059	0,067	0,035	0,082	42	45
Mín	0,520	0,480	0,685	0,377	3	3
25%	0,663	0,646	0,803	0,557	117	102
50%	0,710	0,694	0,825	0,622	141	125
75%	0,741	0,727	0,846	0,675	169	151
Máx	0,862	0,891	0,894	0,811	299	300

Fonte: Autoria própria.
Nota: Base contendo as informações referentes ao IDH e geração de esgoto e consumo de água de 1661 municípios, após a retirada de outliers da base original.

Neste último conjunto de dados é possível verificar através dos histogramas e da tabela estatística que as distribuições de água e esgoto não tiveram mudanças significativas com relação aos conjuntos de dados tratados anteriormente para cada uma dessas variáveis, onde havia 4808 dados para o consumo de água e 1699 dados para a geração de esgoto.

Porém quando comparamos os histogramas dos IDH's após esse corte com os histogramas da base original que continha 5564 dados é nítido que ocorreu uma mudança na distribuição dos dados de IDH, não havendo agora mais dois picos para o IDH Geral e para o IDH Renda, mas apenas um pico. É possível observar também que o pico atual corresponde ao maior pico dos histogramas da base total, indicando que de fato os cortes realizados pelos tratamentos ou pela falta de informação das variáveis de água e esgoto estão diretamente relacionados as cidades com menores IDH's.

2.1.2 IDH x Subíndices

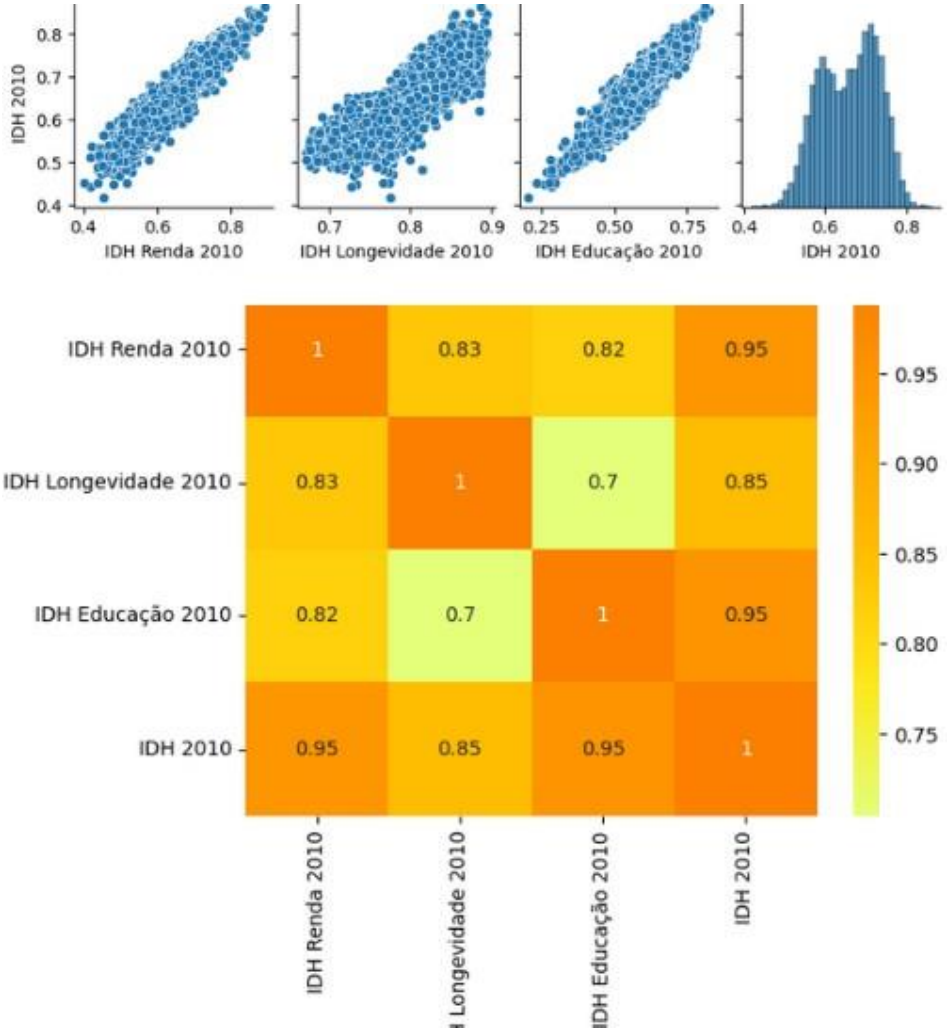
A primeira análise apresentada neste trabalho refere-se a análise de correlação canônica entre o índice de desenvolvimento humano, IDH e um conjunto de variáveis formado pelos seus subíndices, de renda, longevidade e educação.

Para esta análise foi utilizado a base de dados completa, sem a aplicação de nenhum corte ou restrição, contendo todos os 5564 municípios.

O intuito dessa análise é validar a eficácia do método. Uma vez que o IDH geral deriva da média geométrica dos subíndices apresentados, espera-se uma elevada correlação entre os grupos de variáveis.

Primeiramente uma análise preliminar foi realizada com os dados de IDH e seus subíndices, a partir da qual foram construídos os gráficos de dispersão e calor que se seguem na Fig. 19.

Figura 5 – IDH x Subíndices



Fonte: Autoria própria.
Nota: Gráfico de dispersão entre o IDH geral e seus subíndices de renda, longevidade e educação na parte superior seguido de mapa de calor na parte inferior, construídos a partir dos dados referentes ao ano de 2010 para 5564 municípios.

Tanto pelo gráfico de dispersão, quanto pelo mapa de calor é possível observar uma elevada correlação entre o IDH e todos os seus subíndices, assim como entre os próprios subíndices, sendo a maior correlação apresentada de 0,95 entre o IDH e o

subíndice de renda e entre o IDH e o subíndice de educação, ambas com mesmo valor.

Para a análise de correlação canônica foi definido que o grupo de variáveis explicadoras seria composto pelos subíndices de renda, educação e longevidade e que o grupo de variáveis a serem explicadas seria formado apenas pelo índice geral, IDH. Conforme discutido na seção de análise de correlação canônica, o vetor x representa as variáveis explicadoras e o vetor y as variáveis a serem explicadas, dando origem às variáveis canônicas U e V, de acordo com as equações (2.8) e (2.9) respectivamente.

A correlação canônica entre o IDH e seus subíndices foi calculada a partir da (2.13), onde λ ´e o maior autovalor responsável das matrizes apresentadas em (2.12) e a e b são os autovetores que garantem a máxima correlação entre as variáveis canônicas U E V e são obtidos através do sistema de equações composto por (2.10) e (2.11).

De acordo com a tabela mostrada na Fig. 20 abaixo, a correlação entre as variáveis canônicas U e V foi de 0,99 comprovando assim a eficácia do método. Além da elevada correlação também é possível verificarmos, através de seu autovetor a, que o componente de maior peso foi o subíndice da educação, seguido pelo subíndice de renda. O autovetor b apresentou apenas um componente de peso 1, uma vez que há apenas uma variável a ser explicada, o IDH geral.

Quadro 1 – IDH x Subíndices – Correlação canônica

Correlação Canônica	0,999		
	IDH Renda	IDH Longevidade	IDH Educação
Autovetor a	0,577	0,248	0,777
	IDH		
Autovetor b	1		

Fonte: Autoria própria.
Nota: Resultado da correlação canônica entre o IDH geral e seus subíndices de renda, longevidade e educação, assim como os valores dos autovetores da análise.

2.1.3 IDH X Água

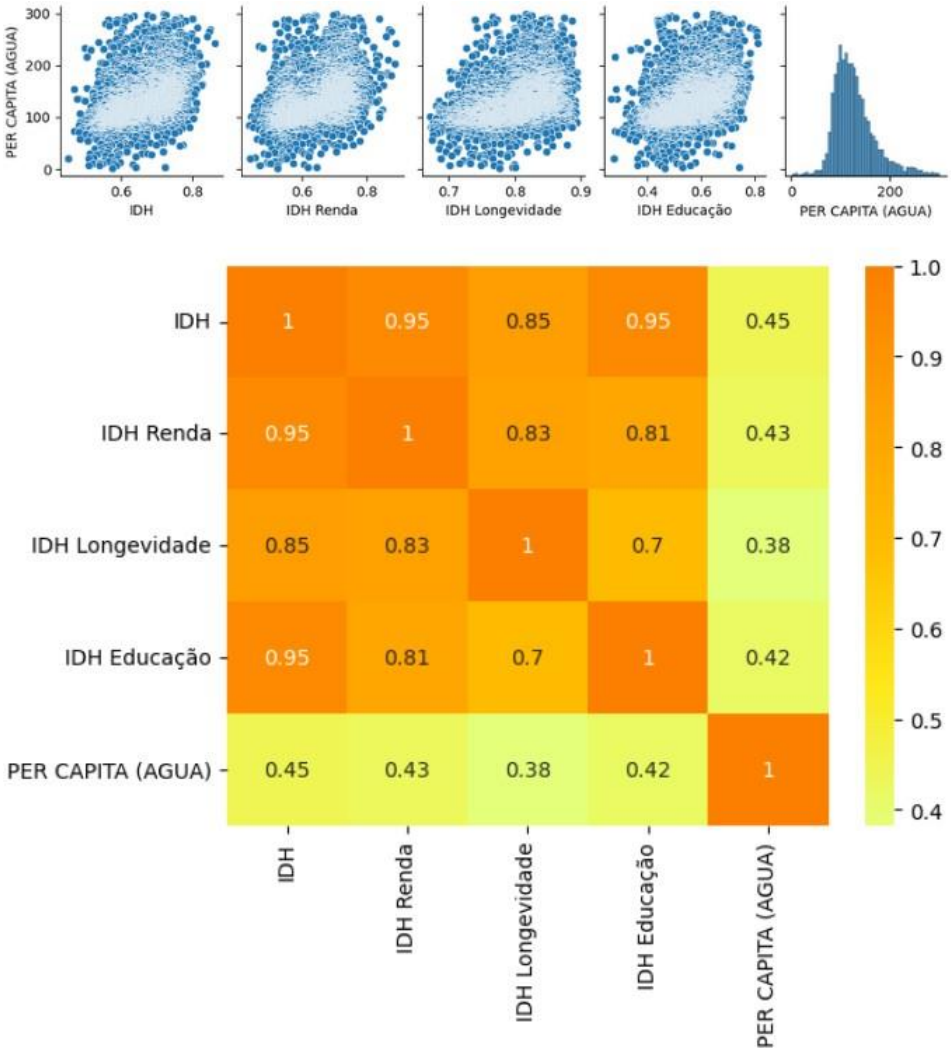
Nesta análise, o índice de IDH geral, juntamente de seus subíndices compõem o grupo de variáveis explicadoras representadas pelo vetor x, enquanto o grupo de

variáveis a ser explicadas é composto apenas pelo consumo de água per capita, representada pelo vetor y.

Conforme discutido na seção anterior de análise exploratória, para esta análise foram utilizados 4808 dados devido ao corte realizado na base de consumo de água no intuito de retirar os outliers e melhorar a qualidade dos dados.

Em uma primeira análise dos dados foi construído um gráfico de dispersão e um mapa de calor conforme a Fig. 21.

Figura 6 – IDH x Água



Fonte: Autoria própria.
Nota: Gráfico de dispersão entre os dados de consumo de água e o IDH geral e seus subíndices de renda, longevidade e educação na parte superior seguido de mapa de calor na parte inferior, construídos a partir dos dados referentes ao ano de 2010 para 4808 municípios que apresentaram informações sobre o consumo de água no SNIS.

Pode-se verificar que a máxima correlação existente entre as variáveis abordadas neste estudo que envolvem o consumo de água per capita se dá entre o

IDH geral e a água, no valor de 0,45 seguido pelos subíndices de renda, educação e apresentando a menor correlação com o subíndice de longevidade com 0,38.

Já a correlação canônica entre os grupos de variáveis, obtida a partir da (2.13), foi de 45,6%. O autovetor a responsável por dar origem a variável canônica U, conforme (2.8), apresenta como componente de maior peso o IDH geral com um valor de 0,800. Essa informação pode nos indicar que cidades com índices de desenvolvimento maiores tendem a ter mais acesso as infraestruturas de saneamento e distribuição de água e consequentemente terão um maior consumo de água per capita quando comparadas com cidades com valores de IDH menores. Ou até mesmo que essas cidades atingiram um IDH mais elevado pelo fato de se disporem mais facilmente ao acesso de recursos hídricos.

Porém, ao observarmos os componentes do autovetor a referente aos subíndices é possível notarmos que eles são negativos, mostrando que individualmente apresentam uma correlação inversa com o consumo de água sendo o componente de maior relevância o subíndice de educação, seguido do subíndice de renda. Indicando que cidades com índices de escolaridade maiores consomem menos água, esse fato pode estar ligado a questões de conscientização que evitam o desperdício de água. Já o autovetor b, responsável por originar a variável canônica V, de acordo com a (2.9), apresenta um único componente de peso 1, uma vez que há apenas uma variável a ser explicada, sendo está o consumo de água. Os dados das análises estão dispostos na Fig. 22.

Quadro 2 – IDH x Água - Correlação canônica

Correlação Canônica	0,456			
	IDH	IDH Renda	IDH Longevidade	IDH Educação
Autovetor a	0,800	-0,358	-0,148	-0,457
	Água			
Autovetor b	1			

Fonte: Autoria própria.
Nota: Resultado da correlação canônica entre os dados de consumo de água e o IDH geral e seus subíndices de renda, longevidade e educação, assim como os valores dos autovetores da análise.

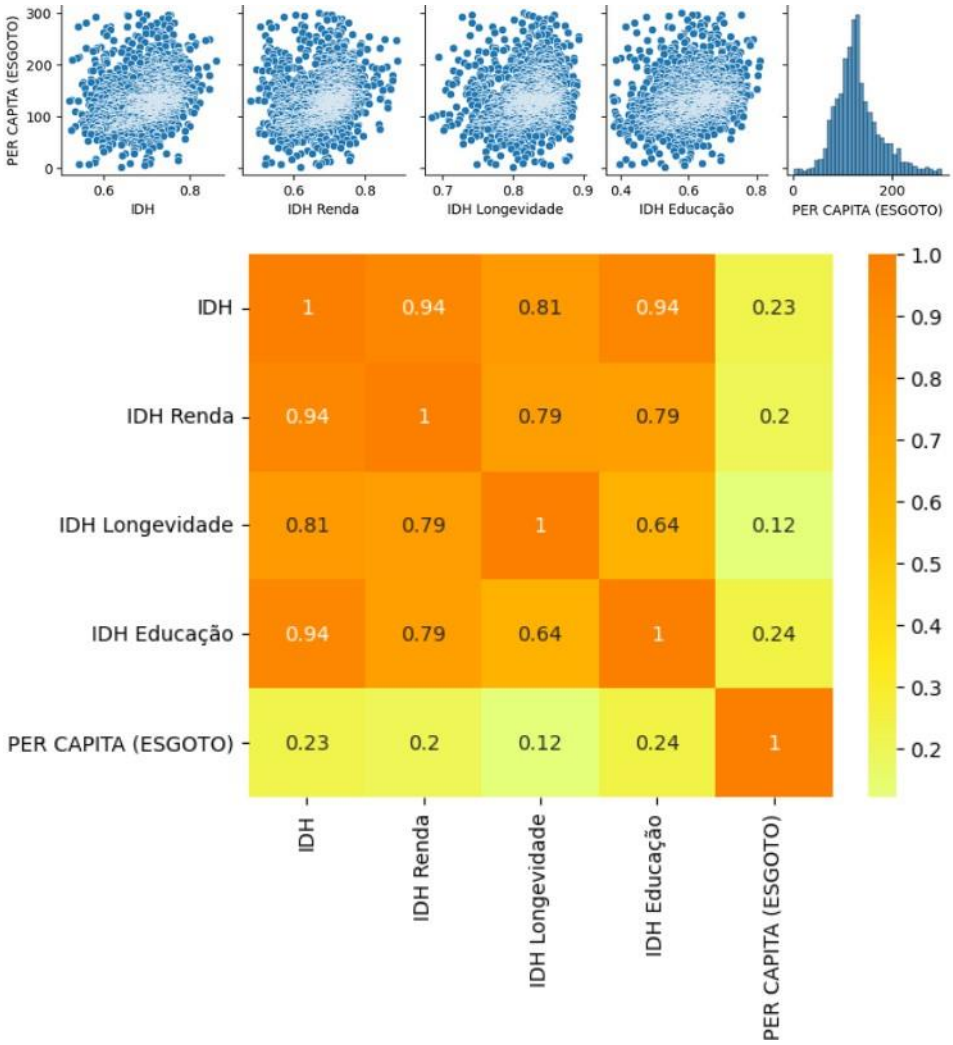
2.1.4 IDH X Esgoto

Nesta análise, assim como na primeira, teremos como variáveis explicadoras o IDH geral e seus subíndices, representados pelo vetor x e uma única variável a ser explicada, porém neste caso será a geração de esgoto per capita, representada através do vetor y.

Devido ao grande número de municípios que não atualizam o sistema do SINIS com os dados de geração de esgoto e pelo tratamento de dados realizado na base para a retirada dos outliers conforme explanado na seção de análise exploratória, nesta análise teremos uma quantidade de dados bastante menor quando comparada a análise anterior, com informações de 1699 municípios.

A análise de dados inicial composta pelo gráfico de dispersão e mapa de calor é apresentada na Fig. 23.

Figura 7 – IDH x Esgoto



Fonte: Autoria própria.
Nota: Gráfico de dispersão entre os dados de geração de esgoto e o IDH geral e seus subíndices de renda, longevidade e educação na parte superior seguido de mapa de calor na parte inferior,

construídos a partir dos dados referentes ao ano de 2010 para 1699 municípios que apresentaram informações de geração de esgoto ao SNIS.

Diferentemente da primeira análise entre IDH e consumo de água, neste estudo entre o IDH e a geração de esgoto não se pôde observar nenhuma correlação significativa entre as variáveis estudadas e a geração de esgoto, sendo a máxima correlação existente de 0,24 entre o subíndice de educação e a geração de esgoto per capita, o que vai em linha com o gráfico de dispersão apresentado, onde não se pode observar uma correlação linear bem definida entre as variáveis.

A correlação canônica calculada para esta análise através da (2.13) também foi bastante inferior com relação as análises prévias apresentadas neste trabalho, sendo de apenas 25,2%. O autovetor a responsável pela originação da variável canônica U através da (2.8), assim como nas outras análises realizadas até o momento também apresentou como componente de maior relevância o IDH geral com um peso de 0,763. As demais componentes referentes aos subíndices de IDH apresentaram valores negativos indicando uma correlação inversa como nas demais análises. O subíndice de educação foi a componente de maior peso com um valor de -0,534, indicando que pessoas com maior acesso à educação tendem a gerar menos esgoto. Como nesta análise há apenas uma única variável a ser explicada, o autovetor b possui apenas uma componente de peso 1 referente a geração de esgoto.

Quadro 3 – IDH x Esgoto - Correlação canônica

Correlação Canônica	0,252			
	IDH	IDH Renda	IDH Longevidade	IDH Educação
Autovetor a	0,763	-0,358	-0,064	-0,534
	Esgoto			
Autovetor b	1			

Fonte: Autoria própria.
Nota: Resultado da correlação canônica entre o IDH geral e seus subíndices de renda, longevidade e educação com relação aos dados de geração de esgoto, assim como os valores dos autovetores da análise.

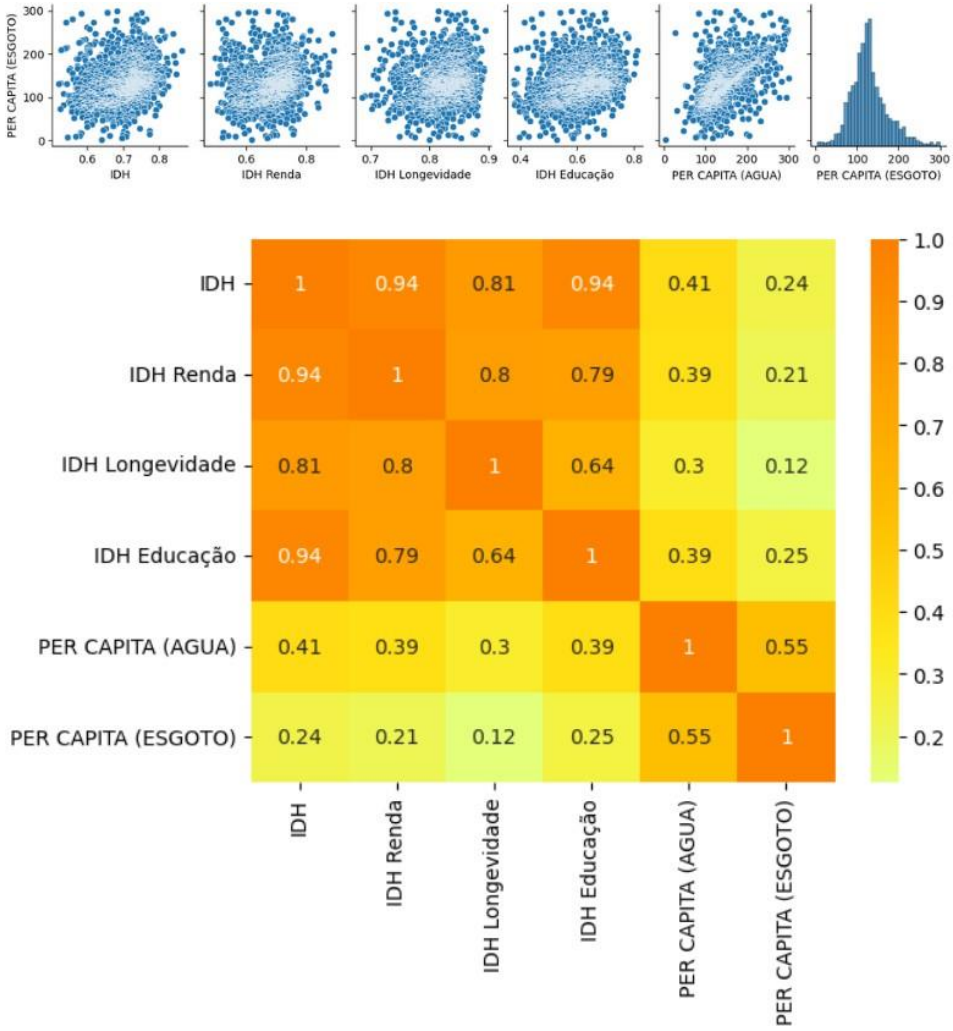
2.1.5 (IDH, Água) X Esgoto

Nesta análise o grupo de variáveis explicadoras será composto pelo IDH geral e seus três subíndices, de renda, educação e longevidade e a variável de consumo

de água per capita, representadas pelo vetor x, enquanto a variável a ser explicada será a geração de esgoto per capita, representada pelo vetor y.

Segue abaixo a análise inicial de dados composta pelo gráfico de dispersão e mapa de calor.

Figura 8 – IDH e Água x Esgoto



Fonte: Autoria própria.
Nota: Gráfico de dispersão entre os dados de geração de esgoto e o IDH geral e seus subíndices de renda, longevidade e educação e os dados de consumo de água na parte superior seguido de mapa de calor na parte inferior, construídos a partir dos dados referentes ao ano de 2010 para 1661 municípios.

Tanto pelo gráfico de dispersão quanto pelo mapa de calor pode-se observar que a melhor correlação apresentada se d´a entre o consumo de água per capita e a geração de esgoto per capita com valor de 0,55. Como já discutido anteriormente essa correlação pode ser explicada pelo impacto direto que a água tem na geração de esgoto durante as atividades básicas do cotidiano. Com relação às variáveis

relacionadas ao IDH e seus subíndices, a melhor correlação obtida entre variáveis de grupos distintos, ou seja, variáveis explicadoras e variáveis a serem explicadas, ocorreu entre o subíndice de educação e a geração de esgoto com valor de 0,25.

A correlação canônica calculada entre os grupos de variáveis estudadas através da (2.13), foi de 56%. O autovetor a responsável por dar origem a variável canônica U, conforme Eq 8, apresenta como componente de maior relevância o consumo de água per capita com um valor de 0,787 seguida pela componente de IDH geral com 0,517, ambas apresentando uma relação direta com a geração de esgoto. Já os subíndices apresentaram uma relação indireta com a geração de esgoto, ou seja, apresentaram componentes vetoriais com valores negativos, sendo o subíndice de longevidade a apresentar o maior peso com -0,200. O autovetor b, responsável por dar origem a variável canônica V, apresentou uma única componente de peso 1 por ter apenas a geração de esgoto a ser explicada.

Quadro 4 – IDH e Água x Esgoto - Correlação canônica

Correlação Canônica	0,560				
	IDH	IDH Renda	IDH Longevidade	IDH Educação	Água
Autovetor a	0,517	-0,200	-0,244	-0,116	0,787
	Esgoto				
Autovetor b	1				

Fonte: Autoria própria.
Nota: Resultado da correlação canônica entre os dados de geração de esgoto e o IDH geral e seus subíndices de renda, longevidade e educação e os dados de consumo de água, assim como os valores dos autovetores da análise.

Dessa forma entende-se que quanto maior o índice de desenvolvimento de uma cidade, assim como maior o consumo de água dessa cidade maior será geração de esgoto per capita. Os resultados aqui apresentados já eram esperados e seguem em linha com o coeficiente de retorno de 0,80 geração de água per capita sugerido pela NBR 9649, assim como diversos trabalhos compilados por [11] que apresentam coeficientes de retorno entre 0,5 e 0,9 da geração de água.

2.1.6 (IDH, Esgoto) X Água

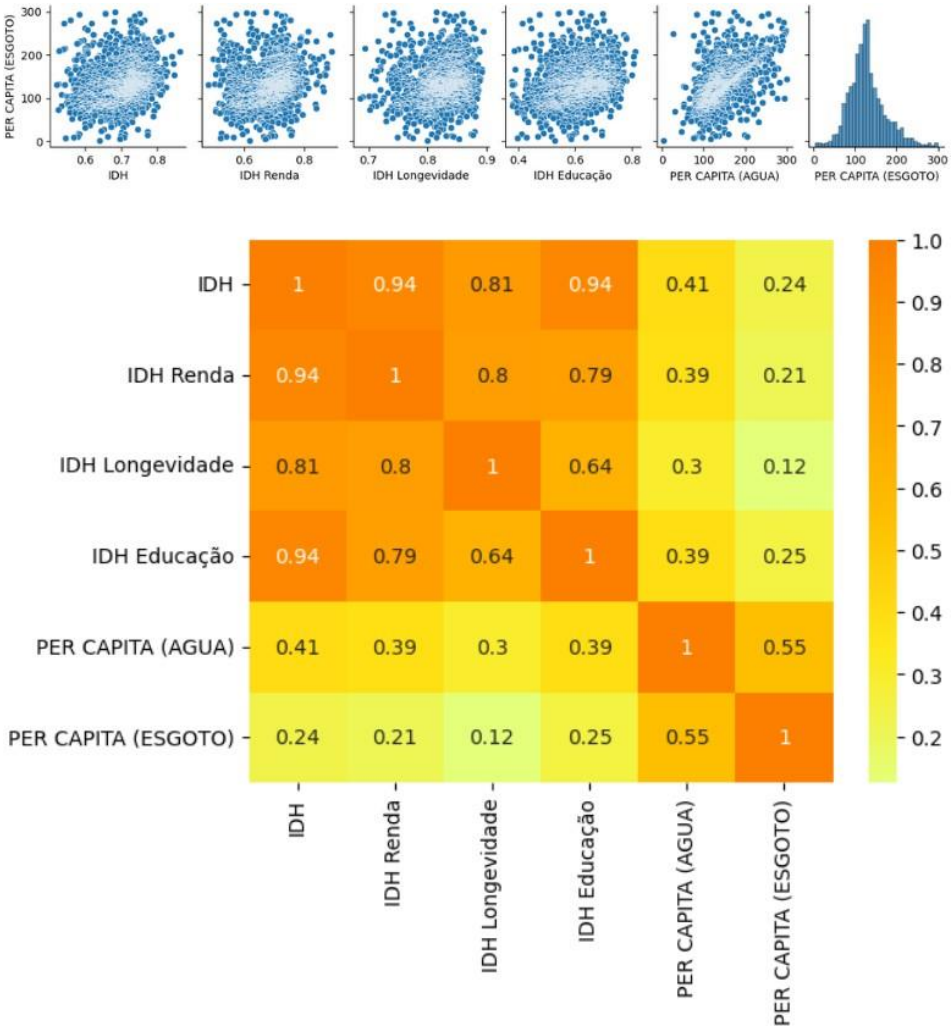
Nesta análise a água segue isolada como variável a ser explicada, representada pelo vetor y, porém foi acrescentado a geração de esgoto per capita

juntamente ao IDH geral e seus subíndices como variáveis explicadoras, representadas a partir do vetor x.

Conforme explicado na seção de análise exploratória, para esta análise foi realizado um tratamento dos dados com o objetivo de eliminar os outliers e melhorar a qualidade da base. Assim, a quantidade de dados utilizada foi de 1661.

Abaixo segue a análise de dados inicial composta pelo gráfico de dispersão e mapa de calor.

Figura 9 – IDH e Esgoto x Água



Fonte: Autoria própria.
Nota: Gráfico de dispersão entre os dados de consumo de água e o IDH geral e seus subíndices de renda, longevidade e educação e os dados de geração de esgoto na parte superior seguido de mapa de calor na parte inferior, construídos a partir dos dados referentes ao ano de 2010 para 1661 municípios.

De acordo com a análise preliminar dos dados apresentada acima, podemos observar que a máxima correlação entre elementos de grupos distintos, ou seja entre

uma variável do grupo explicador e outra do grupo a ser explicado, é dada entre o consumo de água per capita e a geração de esgoto per capita com o valor de 0,55 seguida pelo consumo de água e o IDH geral que está bem abaixo com 0,41.Essa maior correlação apresentada entre o consumo de água e a geração de esgoto é esperada, uma vez que sempre haverá um influência direta do consumo humano da ´agua na geração de seu esgoto durante a realização de atividades básicas do cotidiano.

Ao aplicarmos a análise de correlação canônica levando em consideração a geração de esgoto foi obtida uma correlação, a partir da (2.13), de 62,4%, um aumento significativo quando comparada a análise anterior, mostrando assim a importância e a relação direta que a geração de esgoto possuí com o consumo de água.

Ao observarmos os componentes gerados pela análise com relação ao autovetor a o qual da origem a variável canônica U através da (2.8), percebemos que assim como na análise anterior o maior peso ficou com o IDH geral apresentando um peso de 0,789, indicando mais uma vez que cidades mais desenvolvidas onde a população desfruta de melhor qualidade de vida tendem a consumir mais água. Ou ainda, que por motivos não explorados nesse estudo, como por exemplo geográficos e econômicos, podem ter favorecido o acesso a recursos hídricos o que favoreceu o desenvolvimento dessas cidades.

De forma análoga ao estudo anterior, os componentes referentes aos subíndices do IDH apresentaram seus valores negativos, sendo que novamente o componente de maior relevância foi o subíndice de educação com -0,458 seguido do subíndice de renda com -0,354 Indicando que uma população com mais acesso a educação tende a ser tornar mais conscientizada com relação ao consumo de água e assim evitar seus desperdícios. Aqui mais uma vez é apresentado apenas um componente de peso 1 para o autovetor b, devido ao fato de termos apenas uma única variável a ser explicada, neste caso o consumo de água.

Quadro 5 – IDH e Esgoto x Água - Correlação canônica

Correlação Canônica	0,624				
	IDH	IDH Renda	IDH Longevidade	IDH Educação	Esgoto
Autovetor a	0,789	-0,354	-0,142	-0,458	-0,148
	Água				
Autovetor b	1				

Fonte: Autoria própria.
Nota: Resultado da correlação canônica entre o IDH geral e seus subíndices de renda, longevidade e educação, consumo de água com relação a geração de esgoto, assim como os valores dos autovetores da análise.

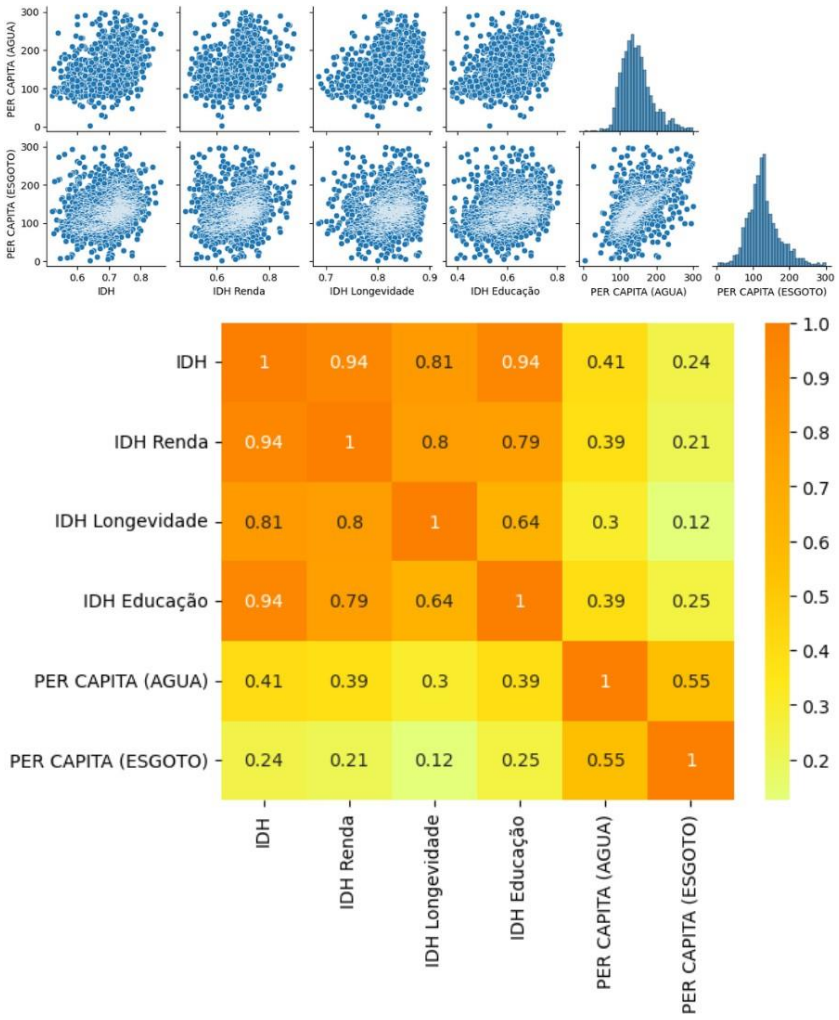
2.1.7 IDH X (Água, Esgoto)

Nesta análise teremos o IDH geral e seus subíndices de renda, longevidade e educação como variáveis explicadoras, representadas através do vetor x, enquanto o grupo de variáveis a ser explicadas será composto pelo consumo de água per capita e pela geração de esgoto per capita, representadas pela variável y.

Para esta análise foram utilizados 1661 dados, uma quantidade bastante inferior quando comparada com as análises envolvendo apenas IDH e água. Isso se deve devido a não atualização dos dados de geração de esgoto pelos municípios e pelo tratamento realizado na base de água e esgoto para a retirada dos outliers conforme discutido na seção de análise exploratória.

Segue abaixo os gráficos de dispersão e mapa de calor referentes a análise inicial dos dados.

Figura 10 – IDH x Água e Esgoto



Fonte: Autoria própria.
Nota: Gráfico de dispersão entre o IDH geral e seus subíndices de renda, longevidade e educação, consumo de água e geração de esgoto na parte superior seguido de mapa de calor na parte inferior, construídos a partir dos dados referentes ao ano de 2010 para 1661 municípios.

Através dessa primeira análise é possível observar que os dados de IDH não apresentam nenhuma correlação linear bem definida com os dados de geração de esgoto, sendo a maior correlação obtida entre o subíndice de educação e a geração de esgoto com valor de 0,25.

Já com relação aos dados de consumo de água, os valores de correlação são um pouco mais elevados chegando a atingir uma correlação máxima de 0,41 entre o IDH geral e o consumo de água.

A correlação canônica obtida nesta análise, a partir da (2.13), foi de 41,8%. O autovetor a apresentou todos os componentes referentes aos subíndices de IDH negativos, indicando uma relação indireta com o consumo de água e com a geração

de esgoto, sendo os componentes de maior peso o subíndice de educação com valor de -481 seguido pelo subíndice de renda com valor de -358. Já a componente referente ao IDH geral apresentou um valor positivo com um peso de 0,790. Isso nos indica que municípios com IDH's mais elevados tendem a gerar mais esgoto e a consumir mais água, entretanto, a partir dos subíndices, podemos esperar que pessoas com maior educação e com melhor poder aquisitivo sejam mais conscientes com relação ao consumo de água, evitando assim seu desperdício e consequentemente gerando menos esgoto.

Devido ao fato de termos duas variáveis a serem explicadas, temos também dois componentes para o autovetor b, sendo um componente referente ao consumo de água com valor de 0,997 e outro referente a geração de esgoto com valor de 0,072. O que nos evidencia que o componente de maior relevância para a correlação encontrada nesta análise é o consumo de água e que a contribuição da geração de esgoto foi mínima.

Quadro 6 – IDH x Água e Esgoto - Correlação canônica

Correlação Canônica	0,418			
	IDH	IDH Renda	IDH Longevidade	IDH Educação
Autovetor a	0,790	-0,358	-0,124	-0,481
	Água		Esgoto	
Autovetor b	0,997		0,072	

Fonte: Autoria própria.
Nota: Resultado da correlação canônica entre o IDH geral e seus subíndices de renda, longevidade e educação, com relação a geração de esgoto e consumo de água, assim como os valores dos autovetores da análise.

2.1.8 Conclusão Parcial: Análise de Consumo de água e geração de esgoto a partir do IDH

Após a análise exploratória dos dados pôde-se observar dois pontos importantes e relevantes para este estudo. O primeiro refere-se a escassez de informações referentes principalmente aos dados de geração de esgoto, de um total de 5564 municípios disponíveis na base de dados de IDH, apenas 1751 registraram a informação de geração de esgoto no SINIS, representando aproximadamente apenas 31% do total de municípios com IDH disponível. O segundo ponto refere-se à

qualidade dessas informações. Tanto para o consumo de água quanto para a geração de esgoto, foram identificados dados com valores que superam a média em mais de 60 desvios padrão para a água e mais de 17 desvios padrão para o esgoto, apresentando valores máximos de 19.881,9 L/hab.dia para consumo de água e de 19.881,9 L/hab.dia de geração de esgoto.

Geograficamente pôde-se verificar que os estados que possuem maior número de municípios estão, em sua maioria, localizados no sul e sudeste do país, tendo como destaque o estado de Minas Gerais com o maior número, representando 15,4% do total com 858 municípios. Já os estados que possuem menor número de municípios estão em sua maioria localizados no Norte, com destaque para o Amapá com o menor número, representando 0,3% do total com apenas 17 municípios.

Após o tratamento das bases de água e esgoto, foi possível notar um aumento da representatividade dos estados do sul e sudeste e uma redução da representatividade dos estados do norte e nordeste indicando assim que os estados do sul e sudeste apresentaram mais registros no SINIS e menos *outliers*, o que torna a base de dados desses estados mais confiáveis e coerentes com a realidade dos municípios. Sendo São Paulo, estado com maior registro no SINIS com 557 municípios, o que representa aproximadamente 33% de todos os registos nacionais.

Diante dessa análise exploratória, torna-se visível a falta de comprometimento da maioria dos municípios com relação a medição e lançamento dos dados no SINIS, indicando também possíveis erros e dificuldades de medição e gestão desses dados no sistema, que ocorrem de forma manual. Foi verificado também que, a maioria dos municípios que apresentaram dados incoerentes[3] ou falta de registros no SINIS, referem-se aos municípios com menores índices de IDH.

Dessa forma, torna-se indispensável que os estados deem assistência e fiscalizem seus municípios com relação a medição e gestão dos dados de consumo de água e geração de esgoto no SINIS, pois dessa forma uma base de dados mais completa e confiável será construída gerando análises mais assertivas que poderão ser utilizadas na tomada de decisão com relação às políticas públicas de cada município.

Com relação às análises de correlação canônica realizadas, pôde-se concluir a eficácia do método através da análise entre o IDH geral e seus subíndices de renda,

[3] Denominados aqui de *outliers*

longevidade e educação, o qual retornou uma correlação canônica de 99,99%. Nas análises envolvendo o IDH, consumo de água e geração de esgoto constatou-se que a correlação canônica entre os grupos de variáveis sempre superava a correlação individual entre as variáveis de grupos diferentes participantes da análise.

A análise que apresentou a melhor correlação canônica foi a que possuía como variáveis explicadoras o IDH geral, seus subíndices de renda, longevidade e educação e a geração de esgoto, pertencente ao grupo de variáveis a ser explicada estava apenas o consumo de água. Nessa análise obteve-se uma correlação canônica de 62,4%. A segunda análise com melhor correlação canônica apresentava como variáveis explicadoras o IDH geral, seus subíndices de renda, longevidade e educação e o consumo de água para explicar a geração de esgoto, com 56%.

Ambas as análises apresentaram correlações canônicas acima da correlação individual entre o consumo de água e geração de esgoto que foi de 55%, acima da correlação individual entre o IDH e o consumo de água que foi de 41% e acima da correlação individual entre o IDH e a geração de esgoto que foi de apenas 24%. Conclui-se dessa forma que o IDH geral e seus subíndices se tornam relevantes para a construção de um modelo que apresente uma melhor correlação envolvendo água e esgoto, do que simplesmente analisarmos água e esgoto isoladamente.

A menor correlação foi de 25,2% entre o IDH geral e seus subíndices e a geração de esgoto, sendo os IDH's variáveis explicadoras e o esgoto variável a ser explicada.

O IDH geral apresentou-se com um peso significativo em todas as análises realizadas e com um valor positivo, indicando uma relação direta com as variáveis a serem explicadas como o consumo de água e geração de esgoto, enquanto seus subíndices apresentaram valores negativos indicando uma relação inversa com o consumo de água e geração de esgoto. Conclui-se dessa forma que, quanto maior o IDH de um município maior tende a ser seu consumo de água e geração de esgoto, porém olhando para seus subíndices, todas as análises com exceção da análise que apresenta IDH e água como variáveis explicadoras para explicar o esgoto, apresentaram o IDH educação como o componente de maior peso e com valor negativo, indicando assim que, quanto maior o acesso à educação de uma população e consequente maior o grau de escolaridade, menor é o consumo de água sugerindo que uma população mais bem educada e informada é também mais conscientizada com relação ao desperdício e uso correto da água.

Para a análise que apresenta IDH e água como variáveis explicadoras para explicar o esgoto, as duas componentes de maior peso foram o IDH geral e o consumo de água, o que reforça o fato de municípios com IDH's mais elevados gerarem mais esgoto e a correlação já conhecida entre água e esgoto.

Após todas as análises terem sido realizadas com os dados totais disponíveis de acordo com cada variável, algumas análises foram refeitas considerando apenas os municípios com maior representatividade de suas respectivas bases de dados com o intuito de verificar a evolução dos resultados de correlação canônica uma vez que acredita-se que os estados com maior número de lançamentos de dados no SINIS são aqueles que possuem melhores recursos para a realização das medições por terem melhores IDH's conforme discutido anteriormente e também uma maior preocupação com essa questão, aumentando dessa forma a qualidade dos dados.

A metodologia adotada para a escolha das novas bases de dados foi a mesma para todas as análises que serão apresentadas a seguir e consiste em filtrar primeiramente os 5 estados de cada base com maior representatividade seguindo as informações apresentadas nos mapas de calor contendo a quantidade e o percentual de informações por estado na seção de análise exploratória. Posteriormente a mesma análise foi refeita considerando os 3 estados mais representativos, os 2 estados mais representativos e por fim com um único estado de maior representatividade.

Tabela 7 – Análise de correlação canônica envolvendo água e esgoto como variáveis a serem explicadas, variando a base de dados

Análise/base	Original	Top5	Top3	Top2	Top1
(IDH, ÁGUA) X ESGOTO	56%	63%	70%	75%	83%
(IDH, ESGOTO) X ÁGUA	62%	69%	71%	77%	83%

Fonte: Autoria própria.
Nota: A base Original refere-se a base completa contendo todos os 1661 dados de água e esgoto disponíveis para os municípios; O Top5 refere-se aos 5 estados de maior representatividade com relação a base Original, sendo estes os estados de MG, SP, PR, CE e BA, representando 72,8% da mesma com um total de 1208 dados; O Top3 refere-se aos 3 estados de maior representatividade com relação a base Original, sendo estes os estados de MG, SP e PR, representando 62,8% da mesma com um total de 1042 dados; O Top2 refere-se aos 2 estados de maior representatividade com relação a base Original, sendo estes os estados de MG e SP, representando 52,3% da mesma com um total de 868 dados; E o Top1 representa o estado de maior representatividade da base Original, sendo este o estado de SP com 544 dados representando 32,8%.

Tabela 8 – Análise de correlação canônica entre IDH e Consumo de água, variando a base de dados

Análise/base	Original	Top5	Top3	Top2	Top1
IDH X ÁGUA	46%	52%	44%	56%	45%

Fonte: Autoria própria.
Nota: A base Original refere-se a base completa contendo todos os 4808 dados de água disponíveis para os municípios; O Top5 refere-se aos 5 estados de maior representatividade com relação a base Original, sendo estes os estados de MG, SP, PR, RS e BA, representando 53,1% da mesma com um total de 2557 dados; O Top3 refere-se aos 3 estados de maior representatividade com relação a base Original, sendo estes os estados de MG, SP e RS, representando 36,4% da mesma com um total de 1754 dados; O Top2 refere-se aos 2 estados de maior representatividade com relação a base Original, sendo estes os estados de MG e SP, representando 27,3% da mesma com um total de 1315 dados; E o Top1 representa o estado de maior representatividade da base Original, sendo este o estado de MG com 762 dados representando 15,8%.

Tabela 9 – Análise de correlação canônica entre IDH e Geração de esgoto, variando a base de dados

Análise/base	Original	Top5	Top3	Top2	Top1
IDH X ESGOTO	25%	27%	32%	33%	29%

Fonte: Autoria própria.
Nota: Original: refere-se a base completa contendo todos os 1751 dados de esgoto disponíveis para os municípios; O Top5 refere-se aos 5 estados de maior representatividade com relação a base Original, sendo estes os estados de MG, SP, PR, BA e CE, representando 72,8% da mesma com um total de 1226 dados; O Top3 refere-se aos 3 estados de maior representatividade com relação a base Original, sendo estes os estados de MG, SP e PR, representando 62,3% da mesma com um total de 1059 dados; O Top2 refere-se aos 2 estados de maior representatividade com relação a base Original, sendo estes os estados de MG e SP, representando 52,1% da mesma com um total de 885 dados; E o Top1 representa o estado de maior representatividade da base Original, sendo este o estado de SP com 577 dados representando 32,8%.

Tabela 10 – Análise de correlação canônica entre IDH e (Água e Esgoto), variando a base de dados

Análise/base	Original	Top5	Top3	Top2	Top1
IDH X (ÁGUA, ESGOTO)	42%	46%	42%	44%	38%

Fonte: Autoria própria.
Nota: A base Original refere-se a base completa contendo todos os 1661 dados de água e esgoto disponíveis para os municípios; O Top5 refere-se aos 5 estados de maior representatividade com relação a base Original, sendo estes os estados de MG, SP, PR, CE e BA, representando 72,8% da mesma com um total de 1208 dados; O Top3 refere-se aos 3 estados de maior representatividade com relação a base Original, sendo estes os estados de MG, SP e PR, representando 62,8% da mesma com um total de 1042 dados; O Top2 refere-se aos 2 estados de maior representatividade com relação a base Original, sendo estes os estados de MG e SP, representando 52,3% da mesma com um total de 868 dados; E o Top1 representa o estado de maior representatividade da base Original, sendo este o estado de SP com 544 dados representando 32,8%.

Através das análises apresentadas acima podemos concluir que, de fato, há uma elevada correlação canônica entre os dados de IDH, água e esgoto quando o IDH se encontra no grupo de variáveis explicativas juntamente com a variável de água para explicar a geração de esgoto ou quando se encontra com a variável de esgoto para explicar o consumo de água, chegando a até 83% quando apenas os dados do estado de São Paulo foram utilizados. Vale destacar, mais uma vez, que o estado de São Paulo foi o que apresentou o maior número de registros da base utilizada na análise de água e esgoto com 32% de representatividade da base. Quanto maior a restrição de dados da base, respeitando a representatividade dos estados, maior foi a correlação canônica apresentada, corroborando com a hipótese de que os estados que apresentam maior número de dados registrados no SINIS apresentam também esses dados com uma qualidade superior com menos lançamentos incorretos que não condizem com a realidade.

Ao aplicarmos a análise para os cenários onde o grupo de variáveis explicadoras é formado apenas pelo IDH e seus subíndices para explicar o consumo de água ou a geração de esgoto obtemos valores de correlação canônica bem inferiores do que os observados na situação discutida no parágrafo anterior. O que indica que a correlação canônica é potencializada pela correlação existente entre o consumo de água e a geração de esgoto, devendo assim essas variáveis estarem em grupos diferentes para que o modelo apresente uma maior correlação canônica.

Outro ponto que também difere da primeira situação discutida é que não houve um aumento da correlação canônica conforme o fatiamento da base de acordo com a representatividade dos estados, mostrando que nem sempre uma base com melhor confiabilidade e qualidade dos dados apresentará uma maior correlação canônica devendo haver um equilíbrio entre o tamanho da base e a qualidade de seus dados.

2.2 ANÁLISE DE COLISÃO DE ÍONS PESADOS RELATIVÍSTICOS

Nesta seção será realizado o estudo e discussão das propriedades do estado inicial da matéria formada através das colisões de íons pesados relativísticos e sua correlação com seus respectivos observáveis finais a partir da aplicação do método de análise de correlações canônicas (CCA).

Este estudo é importante para o entendimento da matéria fortemente acoplada em condições extremas de temperatura e densidade. Matéria essa responsável por descrever o núcleo de estrelas de nêutrons e o universo em seus primeiros instantes.

O trabalho de [12] foi utilizado como base para este estudo. Nele, o plasma de quarks e glúons (QGP), considerado um dos maiores exemplos da atualidade sobre interação forte, é utilizado como ambiente de estudo. O QGP surge a partir da colisão de íons pesados, e neste trabalho foram simuladas colisões entre átomos de chumbo $P_b - P_b$ variando as condições iniciais possibilitando a análise de como a matéria se comporta em termos de densidade, excentricidade de expansão, número de partículas carregadas (N) e momento transversal médio (p_t).

Os dados utilizados para esta análise foram gerados por [12] a partir do simulador TRENTO (Reduced Tickness Event-by-event Nuclear Topology), o qual foi desenvolvido com o objetivo de simular colisões entre átomos pesados, sendo um parametrizador de condições iniciais via espessura reduzida e entropia, seguindo o modelo estatístico de Monte Carlo Glauber de sobreposição de depósitos de entropia, evento por evento de colisão.

Damos o nome de nucleons para os componentes da estrutura interna de um átomo, como seus prótons, nêutrons e demais hádrons [13]. A distribuição espacial dos nucleons dentro de um núcleo pode ser definida para um núcleo a partir de sua massa atômica. Assim, se considerarmos um núcleo simétrico e de massa atômica A define-se a densidade de probabilidade de um nucleons em determinada posição $\vec{r}$ ser encontrado [14].

$$p_{A(\vec{r})} = \frac{n_A(\vec{r})}{A}, \tag{2.16}$$

onde $n_A(\vec{r})$ é o número de nucleons por unidade de volume. Uma vez que ρA(⃗r) ´e uma função de densidade de probabilidade ela deve ser normalizada [14]

$$\int d^3\, p_A(\vec{r}) = 1 \tag{2.17}$$

$$\int_0^\infty r^2\, dr n_A(\vec{r}) = \frac{A}{4\pi} \tag{2.18}$$

Uma escolha comumente utilizada para $n_A(\vec{r})$ é a distribuição de Woods - Saxon:

$$n_A(\vec{r}) = \frac{n_0}{1+\exp\left(\frac{r-R_A}{a}\right)}, \quad (2.19)$$

onde n0 ´e a densidade de matéria normal da ordem de 0.16 fm^{-3}, a da ordem de 0.5 fm ambos com valores próximos entre prótons e neutrons, R é a distância do centro do núcleo até sua borda da ordem de 6.5 fm e r a posição em três coordenadas [13].

O termo “colisões binária” pode ser utilizado para definir a colisão entre dois nucleons pertencentes a átomos distintos que sofrem colisão somente um com o outro. O uso da descrição de colisões binárias auxilia na simplificação da complexidade do sistema de colisões [12] e pode ser calculado para um nucleon que atravessa um núcleo em seu eixo z pela equação abaixo [15].

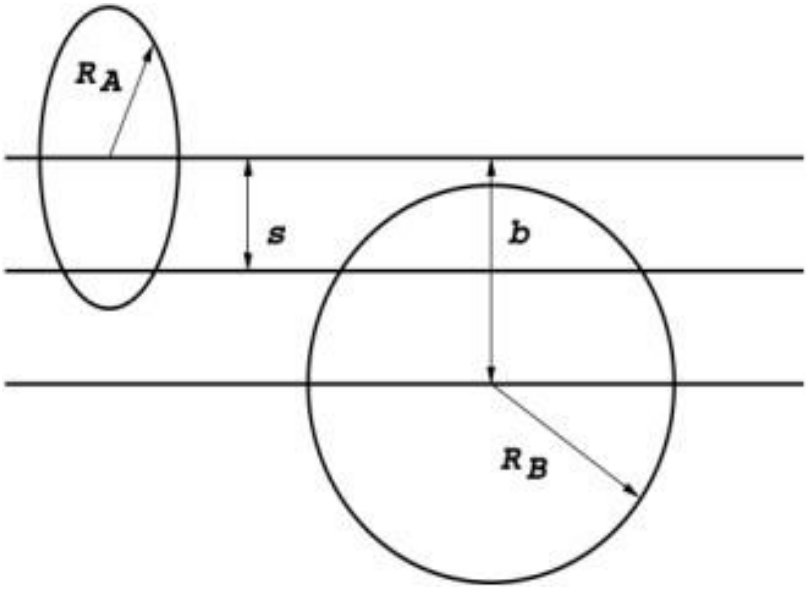
$$T_A(x,y) = \int_{-\infty}^{\infty} n_A(\vec{r})dz = \int_{-\infty}^{\infty} n_A(x,y,z)dz \quad (2.20)$$

Este conceito pode ser expandido para colisões núcleo-núcleo (AB), resultando na função de sobreposição nuclear [16].

$$T_{AB}(b) = \int d^2 s T_A(\vec{s}) T_B(\left|\vec{b}-\vec{s}\right|) \quad (2.21)$$

onde o parâmetro de impacto $\vec{b}$ conecta o centro dos núcleos e $\vec{s} = \vec{s}(x,y)$ aponta do centro de A para um ponto (x,y) em B conforme mostrado na figura31.

Figura 11 – Diagrama transversal de uma colisão Núcleo-Núcleo



Fonte:. [14]

Para a análise de correlação canônica aplicada neste estudo, foi utilizada uma base de dados proveniente da consolidação dos resultados obtidos através de nove simulações, onde em cada simulação foi utilizado um valor de centralidade diferente, com início em 1,25% e término em 55%.

A variável centralidade de colisão (C), pode ser definida como o alinhamento entre o centro de dois núcleos colisores no plano transversal, onde uma centralidade de 0% representa uma colisão mais central com o alinhamento total do centro dos núcleos e uma centralidade de 60% representa uma colisão periférica com colisão efetiva apenas das extremidades dos núcleos colisores.

Conforme discutido por [12], quanto menor a centralidade, ou seja, quanto mais central for a colisão, maior será o número de nucleons participantes e consequentemente maior será o número de partículas (N) no estado pós termalização devido a` maior energia de choque entre os nucleons.

Nos instantes iniciais após a colisão entre os íons de chumbo Pb-Pb o QGP é formado, sendo caracterizado por ser um fluido de alta temperatura, entropia (S) e energia (E), contido em um espaço de elevada densidade. A energia total (E) e a entropia total

(S) produzidas no momento da colisão podem ser calculadas via termodinâmica estatística através da utilização das equações de estado.

A energia total (E) e a entropia total (S) podem ser obtidas através das equações abaixo respectivamente:

$$E = \int \epsilon dV \quad (2.22)$$

$$S = \int s dV \quad (2.23)$$

onde ϵ é a densidade de energia e s é a densidade de entropia do QGP. Porém como o volume é desconhecido a solução da integral é substituída pela somatória da densidade de entropia/energia. A correlação existente entre essas duas propriedades pode ser encontrada através da utilização da relação que ambas possuem com a temperatura na fase QGP através das equações de estado.

2.2.1 (C, E, S, E/S, E/R^3, S/R^3) X (N, pt)

Nesta análise, a centralidade (C), energia (E), entropia (S), energia total por multiplicidade conservada (E/S), densidade de energia (E/R^3), densidade de entropia (E/S^3) apresentam-se como componentes do conjunto de variáveis explicadoras, representadas pelo vetor x, enquanto o número de partículas carregadas (N) e o momento transversal (p_t) compõem o grupo de variáveis a serem explicadas, representadas pelo vetor y.

Para esta análise foram utilizados um total de 900 dados, sendo 100 dados para cada simulação realizada com uma determinada centralidade. Segue abaixo a tabela referente aos resultados obtidos pela análise de correlação canônica entre o primeiro par de variáveis canônicas U1 e V1 conforme descrito respectivamente nas equações 2.8 e 2.9.

Quadro 7 – (C, E, S, E/S, E/R^3, S/R^3) X (N, p_t) - Correlação canônica

Correlaçao Canônica	0,999					
	C	E	S	E/S	E/R³	S/R³
Autovetor a	0,00499	-0,11888	-0,9926	0,00526	-0,01959	0,01304
	nº particulas (N)			Momento transversal (pt)		
Autovetor b	0,999998			0,001885		

Fonte: Autoria própria.
Nota: Resultado da correlação canônica entre o primeiro par de variáveis canônicas onde o grupo de variáveis explicadoras é composto pela centralidade (C), energia (E), entropia (S), energia total por multiplicidade conservada (E/S), densidade de energia (E/R3) e densidade de entropia (E/S3) e o grupo

de variáveis a serem explicadas é composto pelo número total de partículas carregadas e pelo momento transversal. Na parte inferior encontram-se os autovetores da análise.

A correlação canônica obtida nesta análise, a partir da (2.13), foi de 99,9%. Sendo a entropia total (S) o componente canônico de maior relevância do autovetor a, com um peso canônico de −0,9926, muito superior que os demais componentes analisados. Com relação ao autovetor b, o qual possuí apenas dois componentes, verificou-se que a variável de maior representatividade foi o número de partículas carregadas (N), com um peso canônico de 0,9999, muito superior que o segundo componente analisado, momento transversal (p_t), com um valor de 0,0018. Validando dessa forma a correlação direta entre entropia total (S) e número de partículas carregadas, constatado também por (ASSIS 2022) em seu trabalho.

Os dados referentes à correlação canônica do segundo par de variáveis canônicas U2 e V2 são apresentados na tabela abaixo:

Quadro 8 – (C, E, S, E/S, E/R^3, S/R^3) X (N, p_t) - correlação canônica

Correlaçao Canônica	0,959					
	C	E	S	E/S	E/R³	S/R³
Autovetor a	-0,04668	-0,64204	0,74558	0,15607	-0,06491	0,03386
	nº particulas (N)			Momento transversal (pt)		
Autovetor b	-0,18440			0,98285		

Fonte: Autoria própria

Nota: Resultado da correlação canônica entre o segundo par de variáveis canônicas onde o grupo de variáveis explicadoras ´e composto pela centralidade (C), energia (E), entropia (S), energia total por multiplicidade conservada (E/S), densidade de energia (E/R3) e densidade de entropia (E/S3) e o grupo de variáveis a serem explicadas é composto pelo número total de partículas carregadas e pelo momento transversal. Na parte inferior encontram-se os autovetores da análise.

Calculada a partir da (2.13), a correlação canônica apresentada para o segundo par de variáveis canônicas também é elevada, com um valor de 96%. Apesar da componente referente a entropia total (S) ainda ser a de maior relevância com um valor de 0,7455, é possível notar um aumento significativo em todos os demais componentes do autovetor a, com o segundo maior peso canônico com −0,64204 correspondente à variável de energia total (E). Já com relação ao autovetor b, houve uma inversão nos pesos canônicos quando comparado com o primeiro par de variáveis canônicas, sendo nesta análise a variável de momento transversal (p_t) o componente de maior peso com um valor de 0,9828 e não o número de partículas (N), apesar de seu peso ainda ser relevante para a análise. Portanto, para este estudo,

diferentemente do apresentado para o primeiro par de variáveis canônicas, nenhuma das variáveis apresentam pesos desprezíveis para a análise, possibilitando dessa forma, determinar o momento transversal a partir da combinação linear das variáveis explicadoras, algo inédito para na literatura.

2.2.2 (C, E, S, E/R^3, S/R^3) X (E/S, N, p_t)

Para esta análise umas das variáveis do grupo de variáveis explicadoras da análise anterior foi deslocada para o grupo de variáveis a serem explicadas. Assim, temos cinco variáveis explicadoras, sendo estas a centralidade (C), energia (E), entropia (S), densidade de energia (E/R^3) e densidade de entropia (E/S^3), responsáveis por compor o vetor x e três variáveis a serem explicadas compondo o vetor y, sendo estas o número de partículas carregadas (N), momento transversal (p_t) e energia total por multiplicidade conservada (E/S).

O conjunto de dados permanece inalterado com relação a análise anterior, contendo 900 dados, sendo 100 dados para cada simulação realizada com uma determinada centralidade. Segue abaixo a tabela referente aos resultados obtidos pela análise de correlação canônica entre o primeiro par de variáveis canônicas U1 e V1 conforme descrito respectivamente nas equações (2.8) e (2.9).

Quadro 9 – (C, E, S, E/R^3, S/R^3) X (E/S, N, p_t) - Correlação canônica

Correlação Canônica	0,999				
	C	E	S	E/R3	S/R3
Autovetor a	-0,00523	0,11282	0,99331	0,02005	-0,01335
	E/S		nº partículas (N)	Momento transversal (pt)	
Autovetor b	0,00361		0,99999	0,00127	

Fonte: Autoria própria
Nota: Resultado da correlação canônica entre o primeiro par de variáveis canônicas onde o grupo de variáveis explicadoras ´e composto pela centralidade (C), energia (E), entropia (S), densidade de energia (E/R3) e densidade de entropia (E/S3) e o grupo de variáveis a serem explicadas ´e composto pelo número total de partículas carregadas, momento transversal e energia total por multiplicidade conservada (E/S). Na parte inferior encontram-se os autovetores da análise.

Com uma correlação canônica de 99,9%, obtida através da (2.13), o primeiro par de variáveis canônicas apresentam como componentes de maior peso canônico a entropia total (S) com um valor de 0,9933 e o número de partículas (N) com um valor

de 0,9999 correspondendo respectivamente aos autovetores a e b. Sendo estes pesos canônicos muito superiores aos apresentados pelas demais componentes canônicas, demonstrando assim um resultado muito próximo ao obtido pelo primeiro par de variáveis canônicas da primeira análise onde possuíamos apenas duas variáveis a serem explicadas.

A nova variável a ser explicada, energia total por multiplicidade conservada (E/S), apresentou um peso canônico da mesma ordem da variável momento transversal (p_t), sendo estes respectivamente de 0,00361 e 0,00127.

Os dados referentes à correlação canônica do segundo par de variáveis canônicas U2 e V2 são apresentados na tabela abaixo:

Quadro 10 – (C, E, S, E/R^3, S/R^3) X (E/S, N, p_t)- Correlação canônica

Correlação Canônica	0,859				
	C	E	S	E/R3	S/R3
Autovetor a	-0,09395	-0,47526	0,57616	-0,56488	0,33801
	E/S	nº particulas (N)		Momento transversal (pt)	
Autovetor b	0,51798	0,25846		0,81541	

Fonte: Autoria própria.
Nota: Resultado da correlação canônica entre o segundo par de variáveis canônicas onde o grupo de variáveis explicadoras é composto pela centralidade (C), energia (E), entropia (S), densidade de energia (E/R3) e densidade de entropia (E/S3) e o grupo de variáveis a serem explicadas ´e composto pelo número total de partículas carregadas, momento transversal e energia total por multiplicidade conservada (E/S). Na parte inferior encontram-se os autovetores da análise.

A correlação canônica obtida entre o segundo par de variáveis canônicas, conforme apresentado em (2.13), foi de 85,9% tendo como a variável explicadora de maior peso canônico a entropia total (S) com um valor de 0,5716, porém é possível observar que as demais variáveis explicadoras apresentam pesos canônicos de mesma ordem e com valores aproximados ao da entropia (S), com exceção da centralidade (C), a qual apresentou um valor bastante inferior com um peso canônico de apenas 0,09395. Já com relação às variáveis a serem explicadas, todas apresentaram pesos canônicos não-desprezíveis, com valores de mesma ordem, com destaque para o momento transversal (p_t) com valor de 0,81541. Sendo possível, assim como comentado na análise anterior para o estudo com apenas duas variáveis explicadoras, realizar o cálculo do transversal (p_t) a partir da combinação linear das variáveis explicadoras.
Segue abaixo os dados referentes à correlação canônica do terceiro par de variáveis canônicas U3 e V3:

Quadro 11 – (C, E, S, E/R^3, S/R^3) X (E/S, N, p_t) - Correlação canônica

Correlação Canônica	0,782				
	C	E	S	E/R3	S/R3
Autovetor a	-0,02566	-0,73037	0,66035	0,14514	-0,09373
	E/S	nº particulas (N)		Momento transversal (pt)	
Autovetor b	0,76787	0,54447		0,33752	

Fonte: Autoria própria.
Nota: Resultado da correlação canônica entre o terceiro par de variáveis canônicas onde o grupo de variáveis explicadoras é composto pela centralidade (C), energia (E), entropia (S), densidade de energia (E/R3) e densidade de entropia (E/S3) e o grupo de variáveis a serem explicadas é composto pelo número total de partículas carregadas, momento transversal e energia total por multiplicidade conservada (E/S). Na parte inferior encontram-se os autovetores da análise.

A correlação obtida a partir da (2.13) para o terceiro par de variáveis canônicas foi de 78,2%, porém diferentemente do que foi apresentado nas análises anteriores com o primeiro e segundo par de variáveis canônicas, aqui pode-se observar que o componente canônico de maior peso agora ´e a energia total (E) com um valor de 0,73037 e não mais a entropia total (S). A variável canônica de menor peso foi, assim como nas análises anteriores, a centralidade (C) com valor de 0,02566. Com relação as variáveis a serem explicadas, a energia total por multiplicidade conservada (E/S) apresentou o maior peso canônico com valor de 0,76787.

2.2.3 (E, S, E/R^3, S/R^3, E/S) X (C, N, p_t)

Para esta análise foram utilizadas cinco variáveis explicadoras e três variáveis a serem explicadas. Dessa forma, energia (E), entropia (S), densidade de energia (E/R^3), densidade de entropia (E/S^3) e energia total por multiplicidade conservada (E/S) compõem o vetor x e centralidade (C), número de partículas carregadas (N) e momento transversal (p_t) compõem o vetor y.

O conjunto de dados segue o mesmo utilizado nas análises anteriores, contendo 900 dados, sendo 100 dados para cada simulação realizada com uma determinada centralidade. Segue abaixo a tabela referente aos resultados obtidos pela análise de correlação canônica entre o primeiro par de variáveis canônicas U1 e V1 conforme descrito respectivamente nas equações (2.8) e (2.9).

Quadro 12 – (E, S, E/R^3, S/R^3, E/S) X (C, N, p_t)- Correlação canônica

Correlação Canônica	0,999				
	E	S	E/R3	S/R3	E/S
Autovetor a	-0,11366	-0,99317	-0,02133	0,01434	0,00555
	C	nº particulas (N)		Momento transversal (pt)	
Autovetor b	-0,00328	-0,99999		-0,00211	

Fonte: Autoria própria.
Nota: Resultado da correlação canônica entre o primeiro par de variáveis canônicas onde o grupo de variáveis explicadoras ´e composto pela energia (E), entropia (S), densidade de energia (E/R3), densidade de entropia (E/S3) e energia total por multiplicidade conservada (E/S) e o grupo de variáveis a serem explicadas ´e composto pelo número total de partículas carregadas, momento transversal e centralidade (C). Na parte inferior encontram-se os autovetores da análise.

A correlação canônica obtida para o primeiro par de variáveis canônicas conforme (2.13) foi de 99,9%, apresentando como variáveis de maiores pesos canônicos a entropia total (S) e o número de partículas carregadas (N), pertencentes respectivamente ao conjunto de variáveis explicadoras e de variáveis a serem explicadas e valores de 0,99317 e 0,99999, tornando praticamente desprezíveis as demais componentes dos autovetores a e b.

Abaixo segue os resultados obtidos para a correlação canônica referente ao segundo par de variáveis canônicas U2 e V2:

Quadro 13 – (E, S, E/R^3, S/R^3, E/S) X (C, N, p_t)- Correlação canônica

Correlação Canônica	0,960				
	E	S	E/R3	S/R3	E/S
Autovetor a	-0,60863	0,76963	-0,08647	0,04952	0,16524
	C	nº particulas (N)		Momento transversal (pt)	
Autovetor b	-0,36694	-0,49812		0,78564	

Fonte: Autoria própria.
Nota: Resultado da correlação canônica entre o segundo par de variáveis canônicas onde o grupo de variáveis explicadoras é composto pela energia (E), entropia (S), densidade de energia (E/R3), densidade de entropia (E/S3) e energia total por multiplicidade conservada (E/S) e o grupo de variáveis a serem explicadas é composto pelo número total de partículas carregadas, momento transversal e centralidade (C). Na parte inferior encontram-se os autovetores da análise.

Para o segundo par de variáveis canônicas foi obtida uma correlação canônica de 96% através de (2.13). Com relação as variáveis explicadoras, os maiores pesos canônicas foram respectivamente da entropia total (S) com 0,76963 e da energia total (E) com 0,60836. No grupo das variáveis a serem explicadas, o momento transversal

(p_t) apresentou o maior peso canônico com valor de 0,78564 e as demais variáveis apresentaram pesos canônicos relevantes e de mesma ordem. É possível observar que ao deslocarmos a centralidade (C) para o grupo de variáveis explicadoras seu peso canônico cresce substancialmente quando comparado as demais análises apresentadas anteriormente onde está variável se encontrava no grupo de variáveis explicadoras.

Os dados obtidos a partir da correlação canônica entre o terceiro par de variáveis canônicas U3 e V3 é apresentado abaixo:

Quadro 14 – (E, S, E/R^3, S/R^3, E/S) X (C, N, p_t)- Correlação canônica

Correlação Canônica	0,698				
	E	S	E/R3	S/R3	E/S
Autovetor a	-0,70582	0,66995	0,18215	-0,14050	0,00815
	C	nº particulas (N)		Momento transversal (pt)	
Autovetor b	0,72635	0,68602		0,04236	

Fonte: Autoria própria.
Nota: Resultado da correlação canônica entre o terceiro par de variáveis canônicas onde o grupo de variáveis explicadoras é composto pela energia (E), entropia (S), densidade de energia (E/R3), densidade de entropia (E/S3) e energia total por multiplicidade conservada (E/S) e o grupo de variáveis a serem explicadas ´e composto pelo número total de partículas carregadas, momento transversal e centralidade (C). Na parte inferior encontram-se os autovetores da análise.

A partir da (2.13), a correlação canônica calculada para o terceiro par de variáveis canônicas foi de 69,8%, com a energia total (E) e entropia total (S) apresentando os maiores pesos canônicos com relação ao autovetor a com respectivamente 0,70582 e 0,66995. Com relação ao grupo das variáveis explicadoras, o maior peso canônico foi apresentado pela centralidade (C) com valor de 0,72635, porém, diferentemente do que foi observado nas análises anteriores para os estudos com três variáveis a serem explicadas, os todas as três componentes apresentaram pesos de mesma ordem, neste caso o momento transversal (p_t) apresentou um peso canônico muito inferior ao das outras duas componentes, com valor de apenas 0,04236.

2.2.4 Conclusão Parcial: Análise de Colisão de íons pesados relativísticos

Conforme apresentado e discutido por ([17]) e ([18]), experimentalmente a centralidade de colisão apresenta uma correlação direta com o número de partículas

carregadas (N), se considerarmos que toda entropia ´e produzida nos segundos iniciais da colisão espera-se também uma correlação direta entre a entropia (S) e multiplicidade observada. [12] em seu trabalho realiza o cálculo da correlação entre entropia total formada nos segundos iniciais pós colisão e o número de partículas carregadas ao final da evolução hidrodinâmica e encontra uma correlação bastante elevada para esse par de variáveis de 99,9%.

Ao utilizarmos o método de estatística multivariada de correlação canônica identificamos o mesmo resultado obtido por [12] a partir da análise do primeiro par de variáveis canônicas U1 e V1, o qual apresentou uma correlação canônica de 99,9% em todos os três cenários estudados, tanto com duas quanto com três variáveis a serem explicadas. Os resultados de [12] podem ser corroborados pelo método de análise de correlação canônica aplicado neste trabalho uma vez que as componentes que apresentaram maiores pesos canônicos para o primeiro par de variáveis canônicas foram a entropia total (S) e número de partículas carregadas (N), com valores substancialmente maiores que as demais variáveis envolvidas, mostrando assim a correlação direta entre elas e identificada também por [12].

As análises do segundo par de variáveis canônicas, U2 e V2, também apresentaram correlações canônicas elevadas, variando de 86% a 96% a depender do cenário estudado. Porém para este par de variáveis canônicas uma nova correlação entre as variáveis, desconhecida na literatura, foi obtida entre o momento transversal (p_t) e, principalmente, entre a energia total (E) e a entropia total (S). Pesos canônicos consideráveis também foram atribuídos às variáveis de densidade de energia (E/R^3) e densidade de entropia (S/R^3) para o cenário onde a multiplicidade conservada (E/S) foi deslocada para o grupo de variáveis a serem explicadas. Neste cenário a correlação canônica obtida foi de 86%, para os demais cenários estudados a correlação canônica foi de 96%.

Através das análises realizadas foi possível evidenciar que o terceiro par de variáveis canônicas U3 e V3 não apresentou uma correlação canônica elevada para nenhum dos cenários estudados quando comparado aos valores obtidos para os primeiros e segundos pares, não ultrapassando os 80%. Foram realizadas simulações com cenários contendo quatro variáveis a serem explicadas, porém a correlação obtida para o quarto par de variáveis canônicas foi ainda menor, não justificando a aplicabilidade desses modelos com quatro variáveis. Dessa forma podemos concluir que o conjunto de dados analisado comporta modelos com no máximo duas variáveis

a serem explicadas, sendo as mais significativas o número de partículas carregadas (N) e o momento transversal (p_t).

3 ANÁLISE DE COMPONENTES PRINCIPAIS

Este é um dos métodos de estatística multivariada mais conhecido e utilizado na atualidade. Dentre os benefícios de sua aplicação estão a classificação e a redução de dimensionalidade do modelo original em função da criação de novas variáveis provenientes de correlações lineares entre as variáveis originais, mantendo a maior parte da variância original. Essas correlações lineares recebem o nome de componentes principais, dessa forma, se temos p variáveis originais também poderemos obter p componentes principais. Porém um dos objetivos desse método ´e a redução da dimensionalidade do modelo, assim, as p variáveis originais devem ser substituídas por k componentes principais onde k < p [3].

Neste trabalho todas as análises foram realizadas admitindo uma distribuição de probabilidade normal, porém é importante ressaltar que o método de PCA não se restringe a uma distribuição de probabilidade específica, podendo ser aplicado para diversas distribuições.

A partir da aplicação deste método é possível observarmos correlações que não eram tão claras e evidentes entre as variáveis além de identificarmos, a partir da criação dos componentes principais, quais são as variáveis mais relevantes para o modelo em estudo.

A análise de componentes principais pode ser aplicada tanto a partir da matriz de covariância quanto pela matriz de correlação. O modelo teórico e a construção das componentes principais a partir da matriz de covariância de conjuntos de dados são apresentados por [3] e [2] conforme abaixo.

Admitindo-se um vetor aleatório $x = [x_{1,}x_{2,}..., x_p]^T$ contendo um vetor de médias $\mu = [\mu_{1,}\mu_{2,}..., \mu_p]^T$ e uma matriz de covariância Σ_{pxp} para um conjunto de dados de n medidas, tem-se como seus autovalores $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$ e seus respectivos autovetores $e_1, e_2 \ldots e_p$. Dessa forma a j-ésima componente principal da matriz de covariância Σ_{pxp}, j = 1,2...p é dada por:

$$y_j = e_j^T x \tag{3.1}$$

Sendo que $e_j{}^T$ seja o j-ésimo vetor a ser determinado de forma que satisfaça um critério de maximização, isso porque, a PCA rotaciona o sistema de eixos

coordenados em direção a maior variabilidade dos dados, determinando assim os p vetores que maximizam a variância das componentes principais conforme explicado por [19].

A esperança e a variância da j-ésimo componente principal é dada por

$$E[Y_j] = e_j^T \mu \tag{3.2}$$

$$Var[Y_j] = e_j^T \Sigma_{pxp} e_j = \lambda_j \tag{3.3}$$

Assim a covariância entre Y_j e Y_k , para j ≠ k e dada por

$$Cov[Y_j, Y_k] = Cov(e_j^T x, e_k^T x) \tag{3.4}$$

$$Cov[Y_j, Y_k] = e_j^T \Sigma_{ek} \tag{3.5}$$

É apresentado por [2] em seu trabalho, assim como as referências [3] e [19], que as restrições $e_j^T e_j = 1$ e $e_k^T e_k = 0$ são necessárias para que a maximização da variância ocorra conforme abaixo

$$\lambda_j = \max_{e_j} \frac{e_j^T \Sigma e_j}{e_j^T e_j} \tag{3.6}$$

Conforme equação acima e explicado por [20], a maior variância apresentada por uma componente principal corresponde ao maior autovalor, λ_j , pertencente a matriz Σ, e poderá ser calculado a partir do autovetor ej referente a este autovalor.

Assim, a componente principal mais relevante sempre estará associada ao autovetor correspondente ao maior autovalor da matriz Σ, o segundo componente estará associado ao autovetor correspondente a segunda maior autovalor e assim por diante, até o último componente.

Segundo [3] a variável mais importante para cada componente principal é sempre aquela que possuir o maior coeficiente, de forma que as componentes de cada autovetor da matriz Σ possam determinar a importância de cada variável.

Os valores médios dos dados amostrais não são conhecidos na prática, devendo assim ser estimado o vetor de média, μ, e sua matriz de covariância Σ através das equações descritas por [2], de forma que para um número n de vetores aleatórios coletados experimentalmente temos o vetor médio e matriz de covariância conforme segue

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{3.7}$$

$$S = \frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})\,(x_i - \bar{x})^T \tag{3.8}$$

De forma que $\hat{\lambda}_j$ e $\hat{e}_j$ sejam respectivamente o j-ésimo autovalor e seu autovetor correspondente [19].

Como um dos objetivos da aplicação do método de PCA é a redução de dimensionalidade do modelo através da utilização de componentes principais na substituição das variáveis originais, é importante que saibamos a variabilidade individual de cada componente principal com relação a variabilidade total dos dados originais.

Sendo a variância estimada de $\hat{Y}_j$ igual a $\hat{\lambda}_j$, $j = 1, 2, \ldots, p$ temos que o seu peso pode ser calculado de acordo com a equação abaixo

$$\hat{\lambda}_j = \frac{\hat{\lambda}_j}{\sum_{j=1}^{p}\lambda_j} \tag{3.9}$$

E que o peso total das k componentes escolhidas pode ser calculado por

$$V_T = \sum_{j}^{k}\lambda_j \tag{3.10}$$

Não há uma regra que estabeleça um número k de componentes principais que devem ser selecionadas para representar o modelo, porém, ao saber o peso individual de cada uma é possível restringir o modelo com aquelas componentes responsáveis por serem mais representativas, podendo dessa forma reduzir de forma significativa o número de variáveis utilizadas na explicação do modelo.

3.1 PCA COM INCERTEZAS EXPERIMENTAIS

Nos capítulos anteriores foram demonstrados e discutidos os métodos de estatística multivariada referentes a análise de componentes principais (PCA) e análise de correlação canônica (CCA). Porém, como visto, nenhum desses métodos

estatísticos leva em consideração as incertezas experimentais, decorrentes das incertezas estatísticas e instrumentais no momento da medição.

Em seu trabalho, Flausino [2] propõe uma nova abordagem estatística baseada nos métodos clássicos estudados até o momento, levando em consideração as incertezas experimentais. Dessa forma ao invés de termos um ponto no espaço multidimensional, onde cada ponto representa uma medição e cada dimensão representa uma variável do modelo, teremos uma nuvem na qual o valor verdadeiro estaria inserido, representando agora uma probabilidade deste ponto existir dentro da nuvem.

Para essa nova abordagem, tanto o método de componentes principais, quanto a análise de correlação canônica deverão ter um vetor contendo os valores das variáveis e outro vetor contendo suas respectivas incertezas conforme abaixo:

$$x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ x_{i3} \end{pmatrix} \tag{3.11}$$

$$\sigma_i = \begin{pmatrix} \sigma_{i1} \\ \sigma_{i2} \\ \sigma_{i3} \end{pmatrix} \tag{3.12}$$

Sendo i uma observação e p, o número de variáveis para cada observação

Outra característica em comum entre o método de PCA é o de CCA e que ambos utilizam a matriz de covariância para o cálculo de seus resultados e que esta depende diretamente do vetor médio das observações.

Conforme explanado por Flausino em [2], baseado nas discussões trazidas por [21] o método das máximas verossimilhanças é o mais adequado para o cálculo do vetor médio para um conjunto de n valores de uma mesma medida, cada qual com sua respectiva incerteza. E que a melhor estimativa para esta medida será a média ponderada dos dados onde os pesos são o inverso das variâncias.

Dessa forma, [2] utiliza a melhor estimativa do vetor médio, conforme descrito anteriormente, na construção das matrizes de covariância de ambos os métodos aqui tratados, de forma que cada componente da matriz é dada por

$$\bar{x}_a = \frac{\sum_{i=1}^{n} \frac{x_{ia}}{\sigma_{ia}^2}}{\sum_{i=1}^{n} \frac{1}{\sigma_{ia}^2}} \tag{3.13}$$

E sua variância

$$\sigma_{\bar{x}a}^2 = \frac{1}{\sum_{i=1}^{n} \frac{1}{\sigma_{ia}^2}} \tag{3.14}$$

Assim teremos, o vetor médio e sua variância respectivamente conforme abaixo

$$\bar{x} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \bar{x}_3 \end{pmatrix} \tag{3.15}$$

$$\sigma_{\bar{x}}^2 = \begin{pmatrix} \sigma_{\bar{x}_1}^2 \\ \sigma_{\bar{x}_2}^2 \\ \sigma_{\bar{x}_3}^2 \end{pmatrix} \tag{3.16}$$

sendo as incertezas das coordenadas do vetor médio, a raiz quadrada positiva dos elementos σ_{xj}^2.

3.1.1 Análise de componentes principais das bandas interestelares difusas (DIBs) com incertezas experimentais

Nesta seção aplicaremos o método de análise de componentes principais desenvolvido por [2], o qual considera em sua análise as incertezas experimentais, a um conjunto de dados contendo um total de 30 observações com 23 variáveis. Tais dados foram extraídos do trabalho de Ensor *et al*. [1] que aplicou o método de componentes principais tradicional com o intuito de reduzir a dimensionalidade e determinar a quantidade de parâmetros necessários para explicar a variação das bandas interestelares difusas.

As bandas interestelares difusas, conhecidas do inglês como DIBs (diffuse interestellar bands), foram detectadas pela primeira vez em 1897 nos espectros das estrelas Wolf-Rayet [22] e são bandas de absorção observadas no espectro de estrelas na qual a luz tenha atravessado uma quantidade significativa de material interestelar [23].

Não há um consenso sobre quais são os agentes causadores ou originadores das DIB's, também conhecidos como "carriers", porém os mais relevantes até o

momento são os grãos de poeira, pequenas moléculas de carbono, fulerenos e hidrocarbonetos aromáticos policíclicos (PAHs) [23]. Uma alternativa encontrada por pesquisadores da área para a identificação dos responsáveis pela produção das DIBs foi a separação das mesmas em famílias, sendo que cada família deve possuir comportamentos semelhantes [22]. Dessa forma podemos observar a separação de famílias por [24] a partir da correlação entre suas larguras equivalentes e o excesso de cor, obtendo dessa forma um grupo que apresenta uma correlação forte com o excesso de cor e outro que apresenta uma correlação fraca com o excesso de cor. Já [25] utilizou a intensidade apresentada pelas DIBs observadas para uma linha de visada fixa como parâmetro de distinção entre as famílias.

Com relação aos dados utilizados nesse trabalho, conforme descrito por [1], o mesmo realizou uma amostragem sobre um conjunto de dados maior contendo 243 estrelas no qual [26] realizou um estudo detalhado sobre depleção elementar. [1] relata que a amostragem foi realizada de forma que as condições físicas fossem razoavelmente bem definidas e que a amostra representasse as variações de espectro das DIBs de forma geral, sendo importante a seleção de estrelas de diferentes linhas de visada. Outro ponto levantado por [1] durante sua amostragem foi a necessidade de as informações referentes às variáveis auxiliares serem conhecidas e disponíveis.

Para um melhor entendimento do que afeta as variações das intensidades das DIBs, foi necessário a inclusão de outros parâmetros chamados de variáveis auxiliares no estudo. Esses parâmetros são o excesso de cor, E(B-V), o qual está relacionado a quantidade de poeira existente; parâmetro F, relacionado ao esgotamento total de diferentes elementos em uma linha de visada; N(HI) e N(H2) correspondendo respectivamente às informações de densidade de hidrogênio atômico neutro e densidade molecular de hidrogênio; f(H2), relacionado a quantidade de radiação UV interestelar; razão W (λ5797), que está relacionada a condições físicas em uma região do espaço; e, por fim, N(H), que representa o total de hidrogênio, este parâmetro não está listado na tabela de dados de amostra, porém pôde ser calculado a partir da equação abaixo [1].

$$N(H) = N(HI) + 2N(H_2) \tag{3.17}$$

Os dados completos e suas respectivas variáveis podem ser observados nas figuras abaixo:

Tabela 11 – Tabela com oito variáveis de comprimento de onda

Target	λ4428	λ4964	λ5494	λ5513	λ5545	λ5546	λ5769	λ5780
HD 15137	$1163 \pm \frac{106}{115}$	7.9 ± 2.5	11.1 ± 2.2	2.1 ± 3.0	6.9 ± 1.9	0.0 ± 1.9	3.9 ± 1.7	230.1 ± 9.1
HD 22951	$471 \pm \frac{68}{76}$	6.4 ± 1.1	2.0 ± 1.1	1.3 ± 1.5	6.2 ± 0.7	3.6 ± 1.0	0.7 ± 0.8	102.8 ± 3.6
HD 23180	$403 \pm \frac{45}{47}$	12.3 ± 1.4	6.4 ± 0.2	10.7 ± 1.7	10.3 ± 1.5	5.4 ± 1.5	7.2 ± 1.3	88.1 ± 5.0
HD 23630	$325 \pm \frac{48}{39}$	1.2 ± 1.0	2.4 ± 0.9	0.2 ± 1.5	0.8 ± 1.1	1.5 ± 1.0	2.3 ± 0.9	40.7 ± 4.8
HD 24398	$450 \pm \frac{61}{70}$	8.8 ± 0.9	5.4 ± 1.0	5.8 ± 1.1	6.1 ± 0.6	3.3 ± 1.0	2.5 ± 0.7	100.4 ± 2.7
HD 24534	$402 \pm \frac{49}{55}$	13.4 ± 1.6	7.6 ± 1.2	5.3 ± 1.9	9.4 ± 1.2	4.8 ± 1.6	7.1 ± 1.1	95.1 ± 5.0
HD 24760	$322 \pm \frac{41}{30}$	1.5 ± 0.8	3.3 ± 0.8	1.1 ± 1.0	1.5 ± 0.9	0.2 ± 0.8	1.6 ± 0.6	77.0 ± 3.4
HD 24912	$949 \pm \frac{89}{65}$	9.7 ± 1.3	7.0 ± 1.0	2.7 ± 1.2	8.9 ± 1.0	2.4 ± 1.2	2.4 ± 0.8	198.3 ± 3.1
HD 27778	$490 \pm \frac{74}{58}$	8.3 ± 1.4	4.6 ± 1.6	3.6 ± 1.4	8.0 ± 1.1	4.5 ± 1.0	2.2 ± 1.0	86.6 ± 4.6
HD 35149	$254 \pm \frac{43}{38}$	2.8 ± 1.3	2.8 ± 1.7	1.0 ± 1.9	2.6 ± 1.3	0.0 ± 1.4	1.7 ± 1.5	58.0 ± 5.5
HD 35715	$221 \pm \frac{47}{23}$	1.3 ± 0.8	1.1 ± 0.8	1.2 ± 0.9	1.1 ± 0.8	0.7 ± 0.9	0.7 ± 0.7	34.6 ± 3.6
HD 36822	$483 \pm \frac{78}{69}$	1.6 ± 2.4	1.4 ± 2.8	2.9 ± 3.0	2.0 ± 2.4	2.9 ± 2.4	1.0 ± 2.0	84.5 ± 9.6
HD 36861	$402 \pm \frac{49}{86}$	4.6 ± 1.0	3.2 ± 1.0	4.4 ± 1.1	3.2 ± 0.9	3.2 ± 0.9	1.5 ± 0.7	49.0 ± 3.5
HD 40111	$739 \pm \frac{109}{81}$	2.2 ± 4.7	2.7 ± 4.9	0.0 ± 7.2	3.6 ± 4.4	3.2 ± 4.9	3.3 ± 3.7	157.7 ± 19.5
HD 110432	$880 \pm \frac{64}{45}$	8.3 ± 1.0	4.1 ± 1.0	3.8 ± 1.4	5.2 ± 1.0	1.8 ± 0.8	0.3 ± 0.8	137.3 ± 3.7
HD 143275	$383 \pm \frac{21}{12}$	2.1 ± 1.0	5.1 ± 0.1	2.1 ± 1.5	5.2 ± 1.1	1.4 ± 1.2	1.9 ± 1.1	92.7 ± 4.2
HD 144217	$430 \pm \frac{54}{38}$	3.5 ± 0.8	2.6 ± 1.0	1.1 ± 1.6	4.1 ± 1.1	1.0 ± 1.0	0.7 ± 1.1	156.0 ± 4.9
HD 145502	$583 \pm \frac{50}{48}$	3.3 ± 1.2	6.3 ± 2.0	2.8 ± 2.5	4.4 ± 1.0	2.0 ± 1.2	3.0 ± 0.9	186.9 ± 5.2
HD 147165	$872 \pm \frac{50}{53}$	6.1 ± 1.0	8.2 ± 1.5	5.1 ± 1.6	4.5 ± 1.0	1.9 ± 1.2	0.8 ± 1.1	240.0 ± 4.2
HD 147933	$1254 \pm \frac{121}{77}$	20.0 ± 0.8	7.6 ± 0.5	13.8 ± 0.7	8.3 ± 0.5	6.9 ± 0.6	11.7 ± 2.8	209.8 ± 16.1
HD 149757	$576 \pm \frac{52}{47}$	6.6 ± 0.9	5.3 ± 1.1	3.0 ± 1.3	5.7 ± 0.9	2.5 ± 0.8	2.8 ± 1.1	65.1 ± 3.8
HD 164284	$686 \pm \frac{73}{53}$	2.5 ± 1.3	1.8 ± 1.4	2.3 ± 1.8	3.4 ± 1.1	1.5 ± 1.4	0.7 ± 1.0	94.4 ± 4.4
HD 170740	$834 \pm \frac{107}{91}$	10.5 ± 1.0	10.6 ± 1.0	8.6 ± 1.5	11.3 ± 1.0	4.6 ± 1.0	2.4 ± 0.8	240.3 ± 4.0
HD 198478	$1592 \pm \frac{191}{108}$	14.2 ± 2.0	10.5 ± 1.8	5.8 ± 2.1	11.8 ± 1.4	5.0 ± 1.5	1.5 ± 1.3	315.6 ± 5.8
HD 202904	$541 \pm \frac{92}{67}$	2.5 ± 1.5	2.6 ± 1.5	1.8 ± 1.6	1.0 ± 1.0	1.1 ± 1.2	1.0 ± 1.1	44.5 ± 4.6
HD 207198	$1282 \pm \frac{67}{89}$	24.6 ± 1.0	19.3 ± 0.9	16.6 ± 1.1	20.5 ± 0.9	9.9 ± 0.9	9.8 ± 0.7	249.0 ± 2.8
HD 209975	$1032 \pm \frac{182}{74}$	8.8 ± 1.4	13.0 ± 1.5	1.9 ± 1.9	11.2 ± 0.7	4.6 ± 1.6	0.4 ± 1.3	234.2 ± 4.7
HD 214680	$361 \pm \frac{55}{64}$	0.9 ± 1.0	4.4 ± 0.8	1.8 ± 1.4	1.7 ± 1.0	2.0 ± 0.9	0.6 ± 0.5	58.8 ± 2.8
HD 214993	$232 \pm \frac{63}{47}$	4.0 ± 0.7	0.2 ± 1.2	1.6 ± 1.3	1.4 ± 1.0	1.2 ± 0.9	0.6 ± 0.8	78.6 ± 4.8
HD 218376	$766 \pm \frac{64}{108}$	5.1 ± 1.0	5.9 ± 1.1	3.9 ± 1.2	6.2 ± 0.8	1.1 ± 1.1	1.1 ± 0.8	138.7 ± 4.4

Fonte: Adaptado de Ensor *et al*. (2017);

Tabela 12 – Tabela com oito variáveis de comprimento de onda

Target	λ5797	λ5850	λ6196	λ6270	λ6284	λ6376	λ6379	λ6614
HD 15137	68.1 ± 3.1	20.8 ± 2.8	19.9 ± 2.7	33.4 ± 4.6	298.6 ± 19.4	12.7 ± 3.4	36.2 ± 4.2	80.6 ± 4.1
HD 22951	35.9 ± 1.3	18.8 ± 1.2	10.5 ± 2.2	9.0 ± 2.9	130.8 ± 8.5	5.1 ± 1.3	23.8 ± 1.5	41.0 ± 2.1
HD 23180	57.7 ± 2.0	27.8 ± 1.3	12.8 ± 1.9	18.0 ± 3.3	95.4 ± 9.4	10.5 ± 2.1	41.3 ± 3.0	53.7 ± 3.4
HD 23630	6.7 ± 1.3	1.5 ± 1.0	1.9 ± 1.3	3.8 ± 3.3	21.0 ± 7.7	2.0 ± 2.0	3.0 ± 2.1	8.9 ± 2.8
HD 24398	55.5 ± 1.3	27.3 ± 1.1	15.2 ± 1.2	11.0 ± 2.5	94.1 ± 6.7	12.2 ± 1.8	46.3 ± 2.5	59.3 ± 1.9
HD 24534	58.9 ± 1.3	29.0 ± 1.4	15.2 ± 1.4	18.8 ± 3.5	78.2 ± 8.2	10.5 ± 3.7	40.3 ± 2.3	66.1 ± 2.4
HD 24760	13.5 ± 1.0	2.9 ± 0.8	6.0 ± 1.2	11.6 ± 2.0	105.9 ± 5.5	0.3 ± 1.3	8.2 ± 1.5	23.3 ± 2.1
HD 24912	51.4 ± 1.2	22.3 ± 1.7	21.7 ± 1.0	33.0 ± 1.7	272.4 ± 9.6	13.0 ± 1.9	30.1 ± 2.3	79.7 ± 1.8
HD 27778	37.4 ± 2.0	12.7 ± 1.3	10.8 ± 1.5	6.9 ± 3.2	117.8 ± 10.2	8.0 ± 1.8	17.4 ± 2.1	45.7 ± 2.7
HD 35149	11.8 ± 2.1	6.8 ± 1.3	7.1 ± 1.9	12.4 ± 3.7	78.0 ± 14.4	0.9 ± 2.4	6.0 ± 3.3	21.9 ± 4.6
HD 35715	3.3 ± 1.2	0.5 ± 0.7	2.4 ± 1.1	4.0 ± 2.0	55.4 ± 8.4	0.7 ± 2.0	2.8 ± 1.9	9.5 ± 1.9
HD 36822	16.4 ± 3.1	3.7 ± 2.2	8.1 ± 3.1	9.5 ± 8.8	106.6 ± 15.9	3.5 ± 3.3	10.1 ± 5.4	18.0 ± 6.2
HD 36861	23.3 ± 1.2	12.3 ± 0.8	4.9 ± 1.0	4.8 ± 2.1	51.6 ± 10.8	4.7 ± 1.8	6.2 ± 1.4	14.9 ± 1.8
HD 40111	32.3 ± 5.3	3.6 ± 3.1	13.0 ± 5.6	17.1 ± 1..0	211.1 ± 22.5	8.0 ± 7.6	12.9 ± 9.5	41.1 ± 9.6
HD 110432	35.0 ± 1.7	19.4 ± 1.0	18.0 ± 1.0	29.6 ± 2.0	185.1 ± 5.1	7.0 ± 1.8	32.4 ± 1.8	74.3 ± 2.1
HD 143275	17.4 ± 1.3	6.3 ± 1.1	7.6 ± 0.9	10.0 ± 3.8	118.9 ± 13.1	4.3 ± 1.8	10.1 ± 3.0	23.9 ± 1.6
HD 144217	17.3 ± 1.6	6.5 ± 1.1	13.5 ± 1.5	25.0 ± 2.3	159.3 ± 9.1	5.0 ± 2.4	14.0 ± 3.6	50.9 ± 1.7
HD 145502	33.7 ± 1.7	12.2 ± 1.2	14.1 ± 2.6	20.5 ± 2.5	199.6 ± 8.8	7.8 ± 2.0	30.0 ± 2.0	58.8 ± 2.5
HD 147165	31.3 ± 1.6	16.7 ± 1.1	17.5 ± 1.1	26.4 ± 2.7	214.2 ± 7.7	10.9 ± 2.0	21.1 ± 2.0	61.3 ± 2.3
HD 147933	57.2 ± 5.3	30.6 ± 2.6	17.0 ± 2.7	24.9 ± 5.0	173.8 ± 16.9	15.5 ± 2.8	28.0 ± 3.7	62.5 ± 3.6
HD 149757	32.6 ± 1.6	14.2 ± 1.1	10.3 ± 1.2	16.8 ± 2.9	72.0 ± 6.9	10.9 ± 2.0	16.7 ± 1.9	46.4 ± 2.0
HD 164284	13.8 ± 1.7	0.4 ± 1.3	6.8 ± 1.5	15.7 ± 3.0	111.3 ± 9.2	1.8 ± 2.0	11.3 ± 2.2	26.9 ± 2.7
HD 170740	63.3 ± 1.8	24.6 ± 1.1	26.3 ± 1.2	52.7 ± 2.6	249.6 ± 9.9	20.9 ± 1.6	60.7 ± 1.7	122.4 ± 2.2
HD 198478	75.0 ± 2.2	34.6 ± 1.6	33.1 ± 1.5	53.3 ± 4.2	379.5 ± 11.6	21.2 ± 3.5	46.7 ± 4.1	130.6 ± 3.4
HD 202904	5.7 ± 2.3	1.9 ± 1.7	3.6 ± 1.8	15.2 ± 3.1	82.2 ± 10.6	3.0 ± 2.6	11.7 ± 3.5	18.4 ± 2.7
HD 207198	132.6 ± 1.1	61.1 ± 0.7	32.3 ± 1.0	43.2 ± 1.7	227.2 ± 9.6	30.0 ± 1.8	71.8 ± 2.1	121.8 ± 1.9
HD 209975	71.5 ± 1.4	26.5 ± 1.6	26.9 ± 4.5	43.1 ± 3.1	240.2 ± 10.0	25.5 ± 2.7	45.5 ± 2.6	114.1 ± 3.1
HD 214680	20.1 ± 0.9	3.9 ± 0.9	5.4 ± 1.0	9.2 ± 1.6	68.7 ± 7.9	6.4 ± 1.5	4.5 ± 1.4	16.1 ± 2.0
HD 214993	13.6 ± 1.3	0.9 ± 0.7	7.6 ± 1.4	10.0 ± 2.2	107.1 ± 10.0	4.4 ± 1.7	13.9 ± 1.7	18.0 ± 2.3
HD 218376	38.7 ± 1.3	17.2 ± 1.0	14.2 ± 1.2	31.6 ± 2.3	175.7 ± 10.0	11.2 ± 2.0	37.0 ± 2.2	66.0 ± 2.2

Fonte: Adaptado de Ensor *et al*. (2017);

Tabela 13 – Tabela contendo a variáveis auxiliares

Target	E(B-V)	N(H I) $[10^{21}cm^{-2}]$	$N(H_2)$ $[10^{20}cm^{-2}]$	$f(H_2)$	F_*	$\frac{W(\lambda5797)}{W(\lambda5780)}$
HD 15137	0.24	$1.29^{+0.57}_{-0.40}$	$1.86^{+0.26}_{-0.12}$	$0.22^{+0.09}_{-0.06}$	0.37±0.09	0.30±0.02
HD 22951	0.19	$1.10^{+0.35}_{-0.32}$	$2.88^{+1.48}_{-0.98}$	$0.35^{+0.27}_{-0.18}$	0.73±0.05	0.35±0.02
HD 23180	0.22	$0.76^{+0.26}_{-0.23}$	$3.98^{+1.64}_{-1.16}$	$0.51^{+0.33}_{-0.24}$	0.84±0.06	0.65±0.04
HD 23630	0.05	$0.22^{+0.10}_{-0.07}$	$0.35^{+0.18}_{-0.12}$	$0.28^{+0.23}_{-0.15}$	0.89±0.10	0.16±0.04
HD 24398	0.27	$0.63^{+0.06}_{-0.07}$	$4.68^{+2.40}_{-1.59}$	$0.59^{+0.46}_{-0.31}$	0.88±0.05	0.55±0.02
HD 24534	0.31	$0.54^{+0.08}_{-0.07}$	$8.32^{+0.80}_{-0.73}$	$0.76^{+0.13}_{-0.11}$	0.90±0.06	0.62±0.04
HD 24760	0.07	$0.25^{+0.05}_{-0.05}$	$0.33^{+0.27}_{-0.15}$	$0.21^{+0.25}_{-0.14}$	0.68±0.04	0.18±0.02
HD 24912	0.26	$1.29^{+0.26}_{-0.24}$	$3.39^{+1.40}_{-0.99}$	$0.35^{+0.21}_{-0.15}$	0.83±0.02	0.26±0.01
HD 27778	0.34	$0.22^{+0.55}_{-0.22}$	$5.25^{+1.06}_{-0.88}$	$0.82^{+0.45}_{-0.27}$	1.19±0.07	0.43±0.03
HD 35149	0.08	$0.43^{+0.12}_{-0.13}$	$0.03^{+0.00}_{-0.03}$	$0.02^{+0.00}_{-0.02}$	0.54±0.11	0.20±0.04
HD 35715	0.03	$0.31^{+0.13}_{-0.13}$	$6\pm2\times10^{-6}$	$4\pm2\times10^{-6}$	0.66±0.11	0.10±0.04
HD 36822	0.07	$0.65^{+0.13}_{-0.12}$	$0.21^{+0.09}_{-0.06}$	$0.06^{+0.04}_{-0.03}$	0.74±0.08	0.19±0.04
HD 36861	0.10	$0.60^{+0.16}_{-0.16}$	$0.13^{+0.08}_{-0.05}$	$0.04^{+0.04}_{-0.02}$	0.57±0.04	0.48±0.04
HD 40111	0.10	$0.79^{+0.16}_{-0.15}$	$0.54^{+0.31}_{-0.20}$	$0.12^{+0.10}_{-0.07}$	0.49±0.04	0.20±0.04
HD 110432	0.39	$0.71^{+0.29}_{-0.21}$	$4.37^{+0.42}_{-0.38}$	$0.55^{+0.13}_{-0.11}$	1.17±0.11	0.25±0.01
HD 143275	0.00	$1.41^{+0.29}_{-0.29}$	$0.26^{+0.15}_{-0.09}$	$0.03^{+0.03}_{-0.02}$	0.90±0.03	0.19±0.02
HD 144217	0.18	$1.23^{+0.12}_{-0.11}$	$0.68^{+0.10}_{-0.09}$	$0.10^{+0.02}_{-0.02}$	0.81±0.02	0.11±0.01
HD 145502	0.20	$1.17^{+0.56}_{-0.59}$	$0.78^{+0.32}_{-0.23}$	$0.12^{+0.08}_{-0.07}$	0.80±0.11	0.18±0.01
HD 147165	0.31	$2.19^{+0.90}_{-0.87}$	$0.62^{+0.25}_{-0.18}$	$0.05^{+0.04}_{-0.03}$	0.76±0.06	0.13±0.01
HD 147933	0.37	$4.27^{+0.98}_{-0.80}$	$3.72^{+1.53}_{-1.09}$	$0.15^{+0.09}_{-0.07}$	1.09±0.08	0.27±0.03
HD 149757	0.29	$0.52^{+0.02}_{-0.04}$	$4.47^{+0.90}_{-0.75}$	$0.63^{+0.20}_{-0.17}$	1.05±0.02	0.50±0.04
HD 164284	0.11	$0.42^{+0.23}_{-0.39}$	$0.71^{+0.29}_{-0.21}$	$0.25^{+0.18}_{-0.20}$	0.89±0.18	0.15±0.02
HD 170740	0.38	$1.07^{+0.59}_{-0.47}$	$7.24^{+1.47}_{-1.22}$	$0.58^{+0.22}_{-0.18}$	1.02±0.11	0.26±0.01
HD 198478	0.43	$2.04^{+0.84}_{-0.63}$	$7.41^{+3.06}_{-2.17}$	$0.42^{+0.27}_{-0.20}$	0.81±0.05	0.24±0.01
HD 202904	0.09	$0.23^{+0.21}_{-0.23}$	$0.14^{+0.07}_{-0.05}$	$0.11^{+0.12}_{-0.10}$	0.39±0.11	0.13±0.05
HD 207198	0.47	$3.39^{+0.59}_{-0.50}$	$6.76^{+0.65}_{-0.59}$	$0.28^{+0.05}_{-0.05}$	0.90±0.03	0.53±0.01
HD 209975	0.27	$1.29^{+0.41}_{-0.38}$	$1.20^{+0.62}_{-0.41}$	$0.16^{+0.12}_{-0.09}$	0.57±0.26	0.31±0.01
HD 214680	0.08	$0.50^{+0.14}_{-0.15}$	$0.17^{+0.05}_{-0.04}$	$0.06^{+0.03}_{-0.03}$	0.50±0.06	0.34±0.02
HD 214993	0.06	$0.58^{+0.20}_{-0.18}$	$0.43^{+0.22}_{-0.14}$	$0.13^{+0.10}_{-0.07}$	0.68±0.10	0.17±0.02
HD 218376	0.16	$0.89^{+0.28}_{-0.26}$	$1.41^{+0.73}_{-0.48}$	$0.24^{+0.19}_{-0.13}$	0.60±0.06	0.28±0.01

Fonte: Adaptado de Ensor *et al*. (2017);

A partir dos dados apresentados nas figuras 39, 40 e 41, foi realizado o cálculo do autovalor e de seu respectivo desvio padrão assim como a construção de um

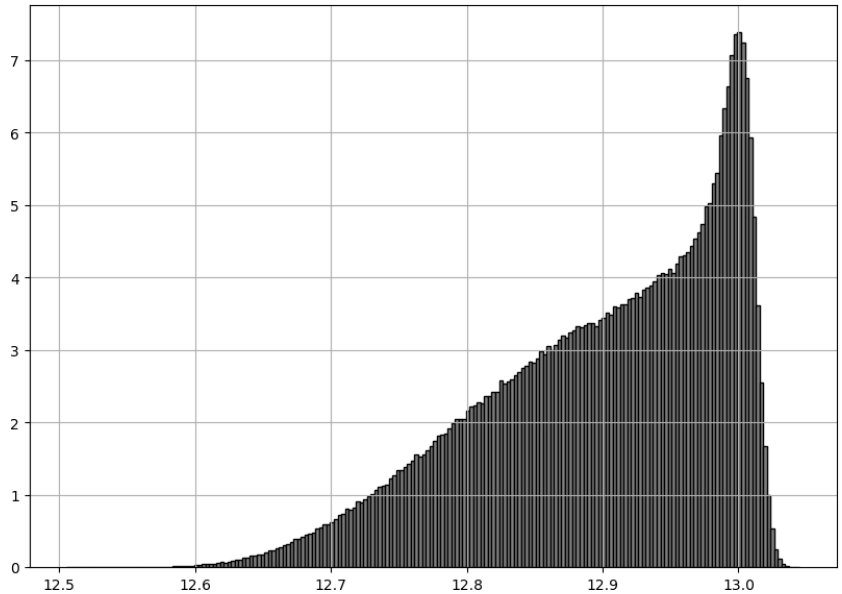
histograma para cada uma das 23 variáveis analisadas com o objetivo de verificarmos o impacto das incertezas. Para realizar a propagação de erros das componentes principais foi utilizado um algoritmo desenvolvido pelo autor em python, no qual foram utilizadas as bibliotecas xlrd para leitura e importação de arquivos; numpy para cálculos matriciais e estatísticos e matplotlib para a construção dos histogramas.

Como o método de PCA consiste na obtenção de autovalores e autovetores a partir da matriz de covariância 3.8, a metodologia aplicada, consiste em realizar um procedimento iterativo (três milhões de iterações) onde, em cada iteração, gera-se um vetor médio aleatório com distribuição gaussiana, em que a média e desvio padrão de cada componente são, respectivamente, as componentes da média ponderada e suas incertezas. Com isso, calcula-se uma matriz de covariância aleatória para cada iteração bem como seus autovalores e autovetores armazenando esses valores em histogramas.

A seguir será apresentado o autovalor, desvio padrão e histograma de cada componente principal, a qual chamaremos de PC.

PC_1

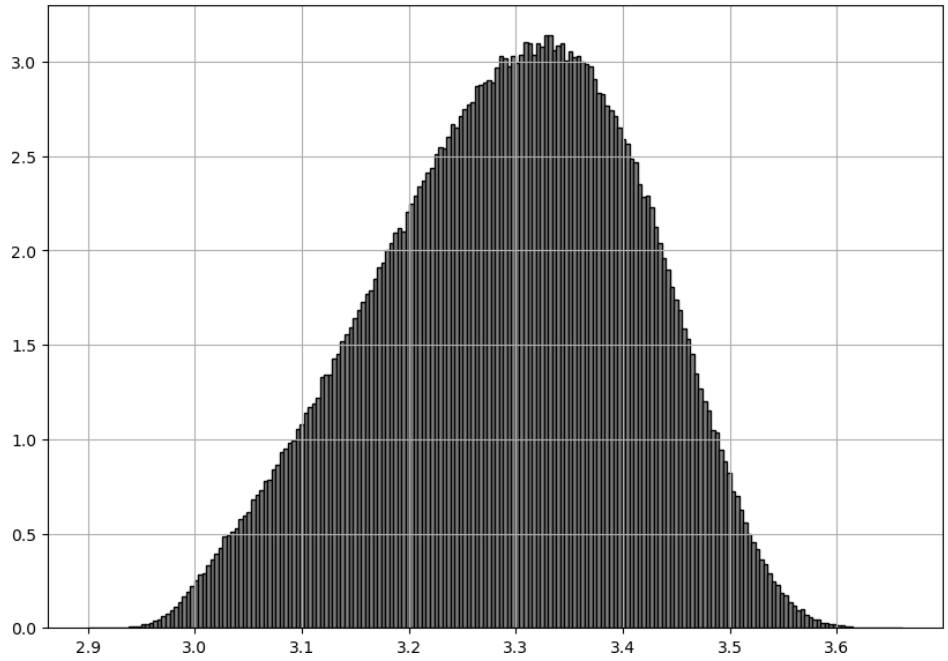
Gráfico 15 – Histograma referente a PC1



Fonte: Autoria própria.

PC_2

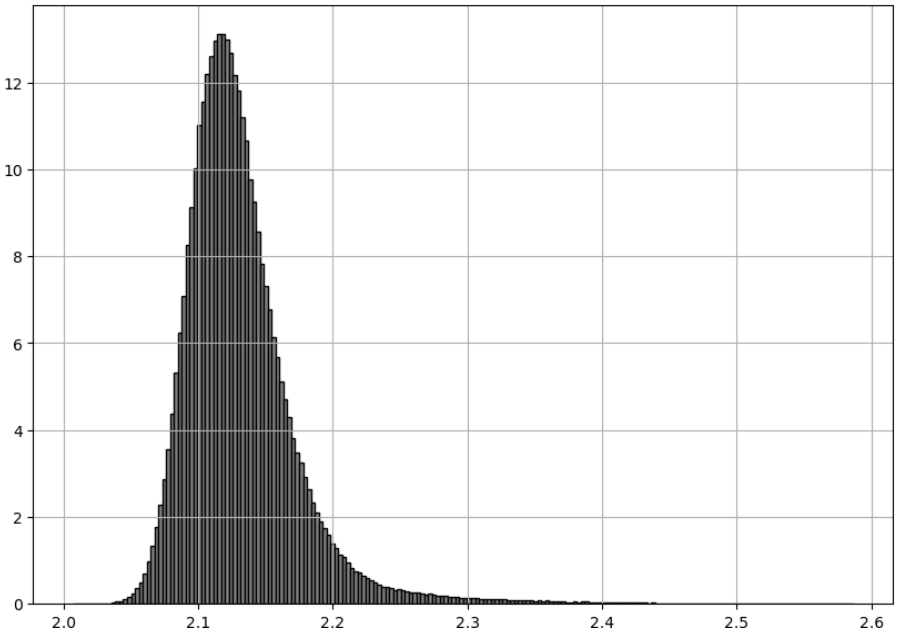
Gráfico 16 – Histograma referente a PC2



Fonte: Autoria própria.

PC_3

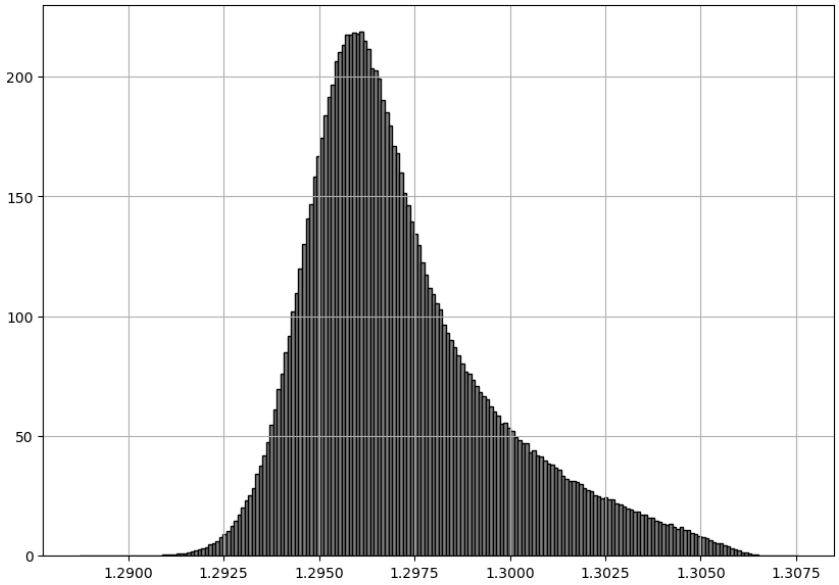
Gráfico 17 – Histograma referente a PC3



Fonte: Autoria própria.

PC_4

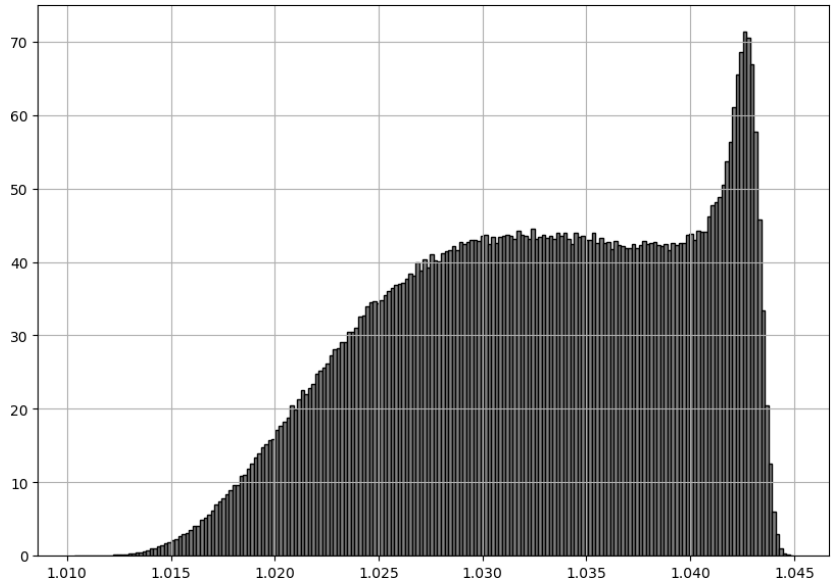
Gráfico 18 – Histograma referente a PC4



Fonte: Autoria própria.

PC_5

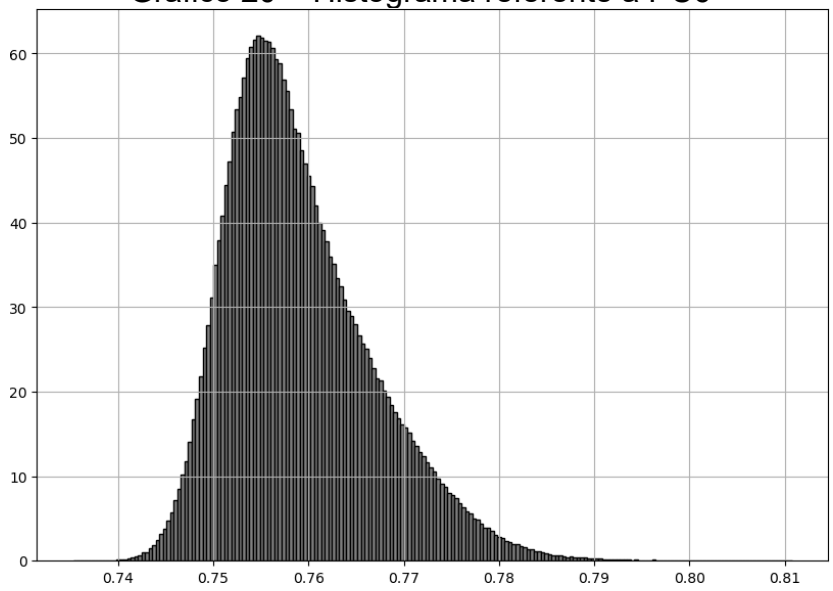
Gráfico 19 – Histograma referente a PC5



Fonte: Autoria própria.

PC_6

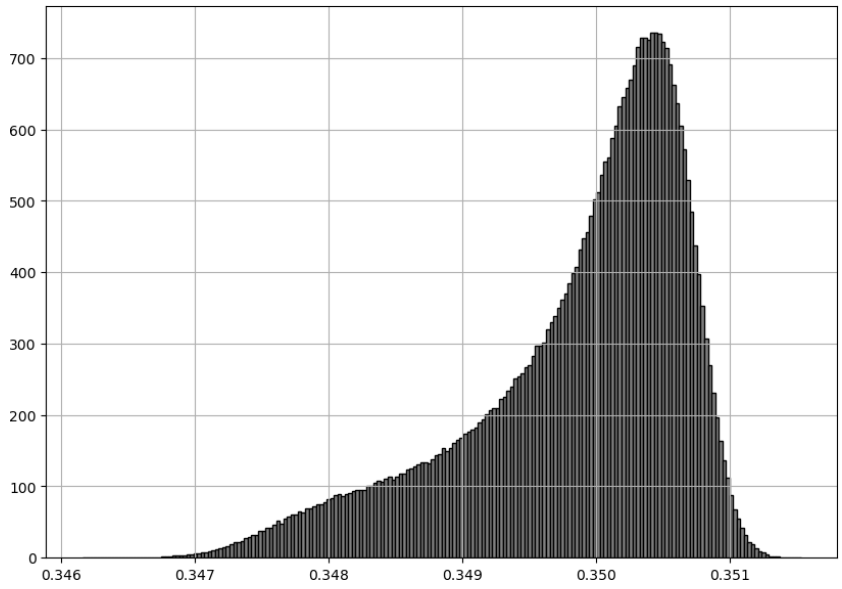
Gráfico 20 – Histograma referente a PC6



Fonte: Autoria própria.

PC_7

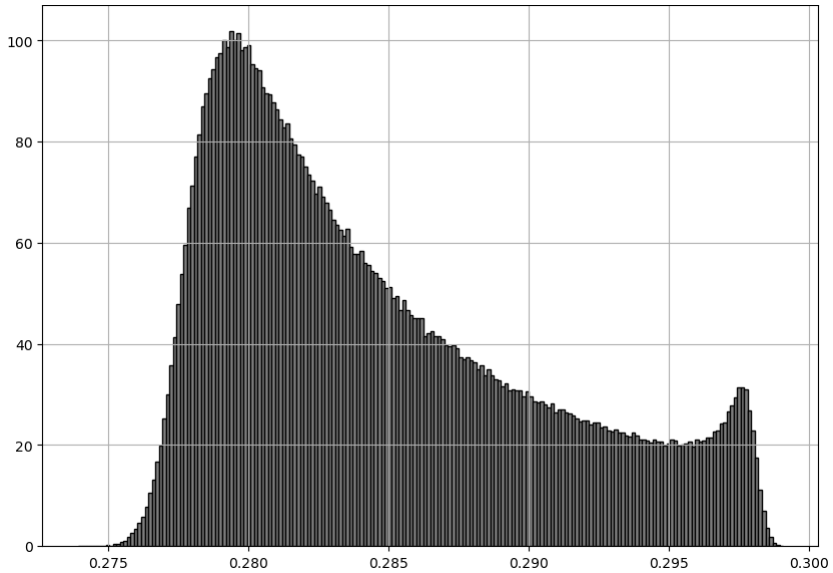
Gráfico 21 – Histograma referente a PC7



Fonte: Autoria própria.

PC_8

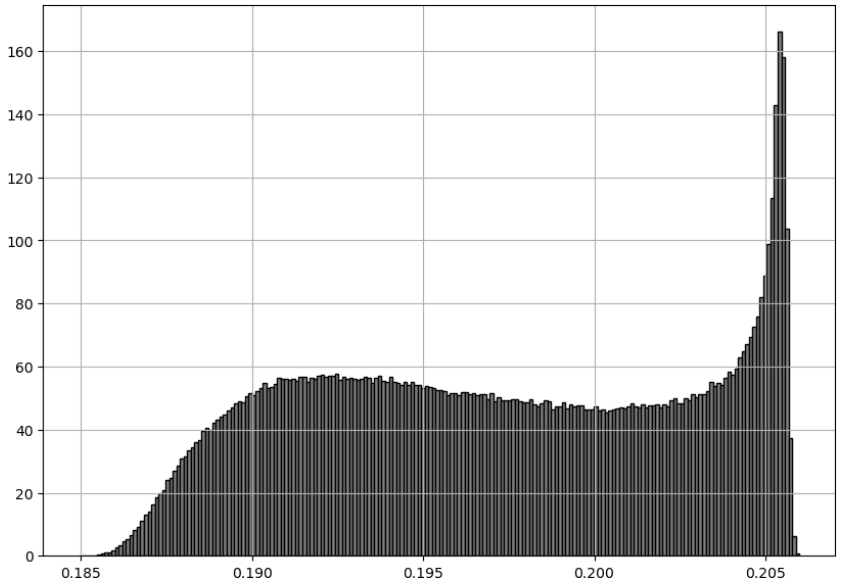
Gráfico 22 – Histograma referente a PC8



Fonte: Autoria própria.

PC_9

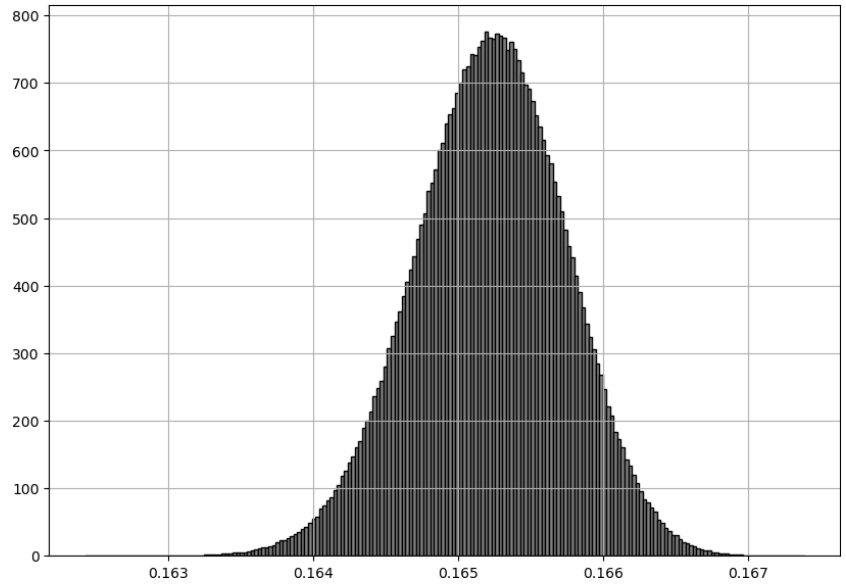
Gráfico 23 – Histograma referente a PC9



Fonte: Autoria própria.

PC_{10}

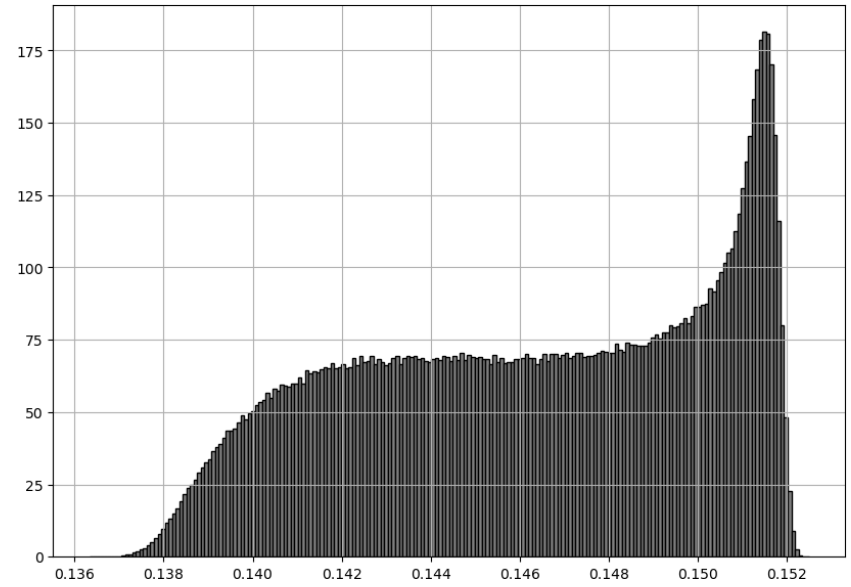
Gráfico 24 – Histograma referente a PC10



Fonte: Autoria própria.

PC_{11}

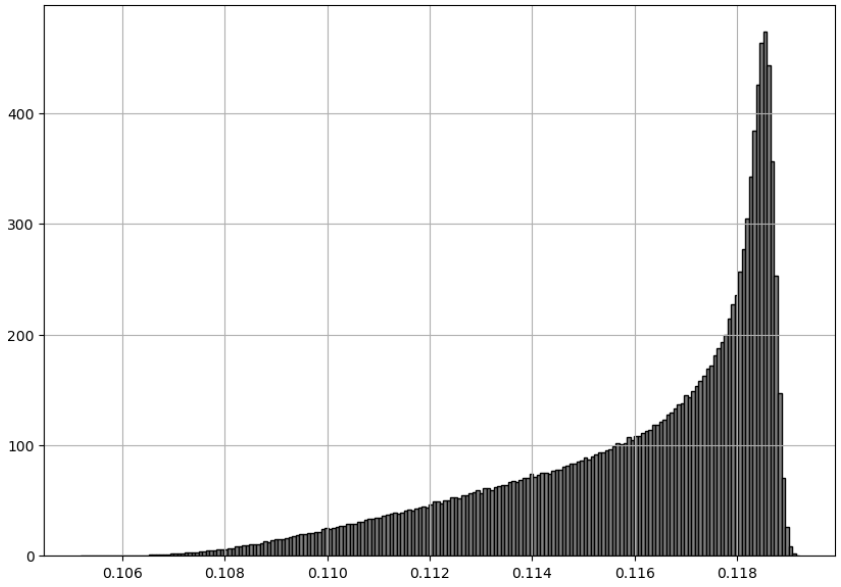
Gráfico 25 – Histograma referente a PC11



Fonte: Autoria própria.

PC_{12}

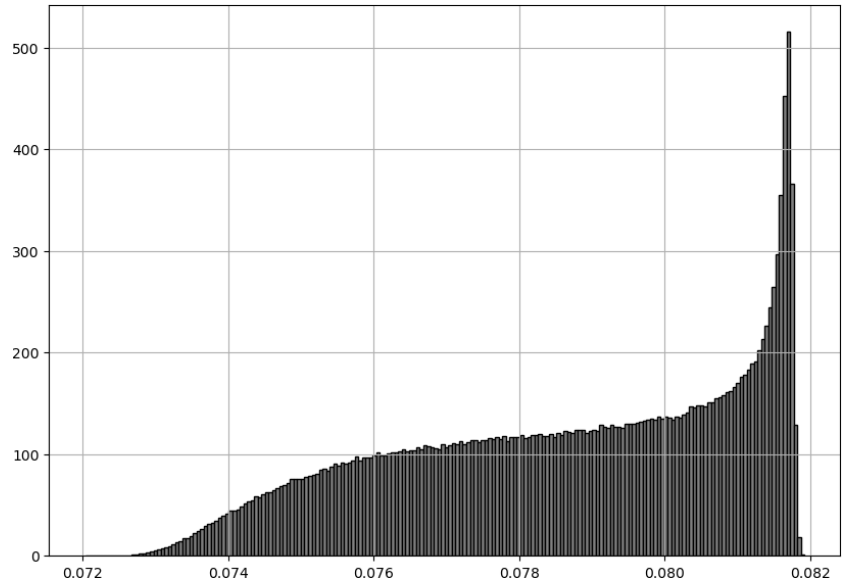
Gráfico 26 – Histograma referente a PC12



Fonte: Autoria própria.

PC_{13}

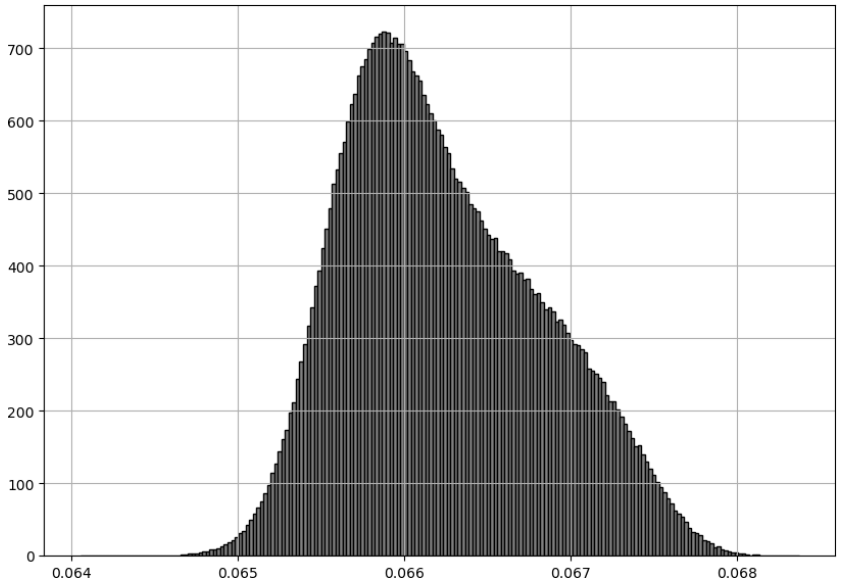
Gráfico 27 – Histograma referente a PC13



Fonte: Autoria própria.

PC_{14}

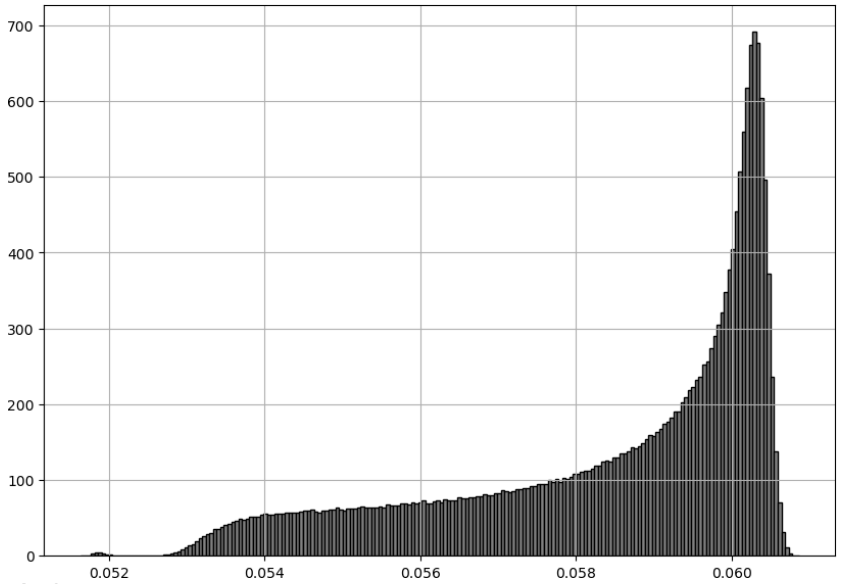
Gráfico 28 – Histograma referente a PC14



Fonte: Autoria própria.

PC_{15}

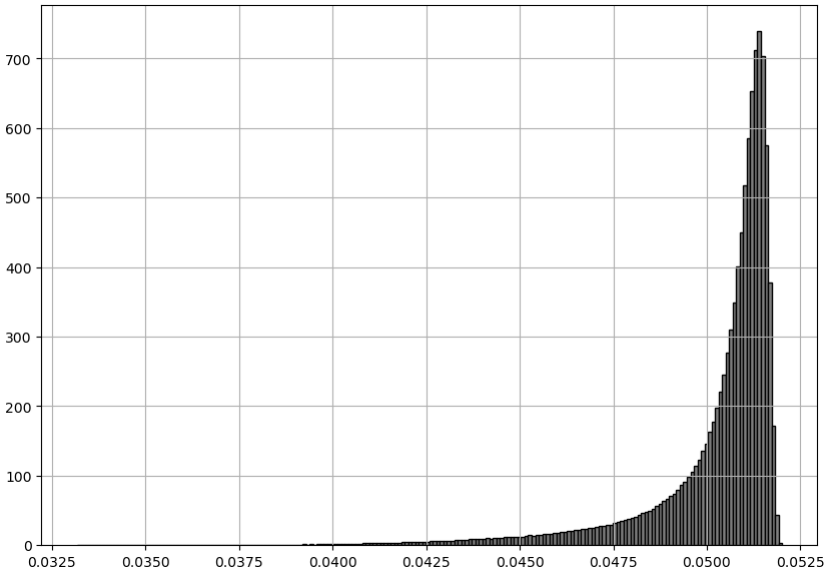
Gráfico 29 – Histograma referente a PC15



Fonte: Autoria própria.

PC_{16}

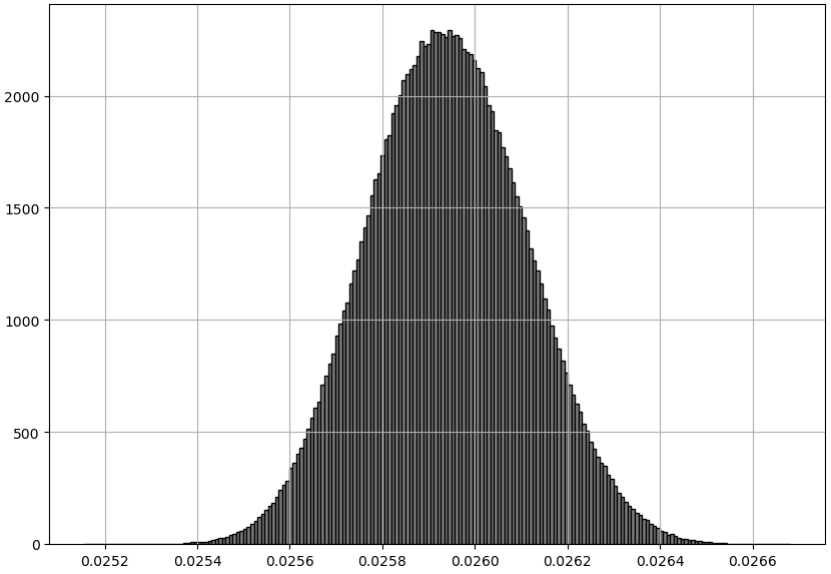
Gráfico 30 – Histograma referente a PC16



Fonte: Autoria própria.

PC_{17}

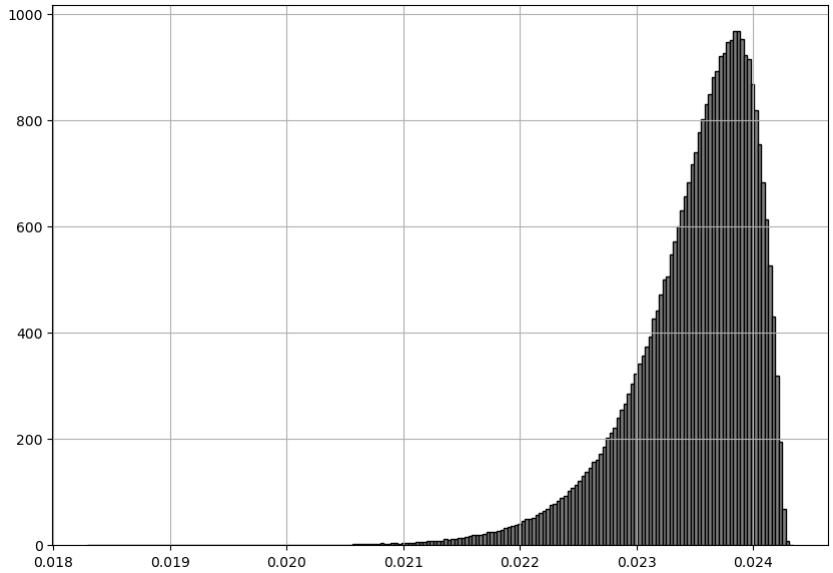
Gráfico 31 – Histograma referente a PC17



Fonte: Autoria própria.

PC_{18}

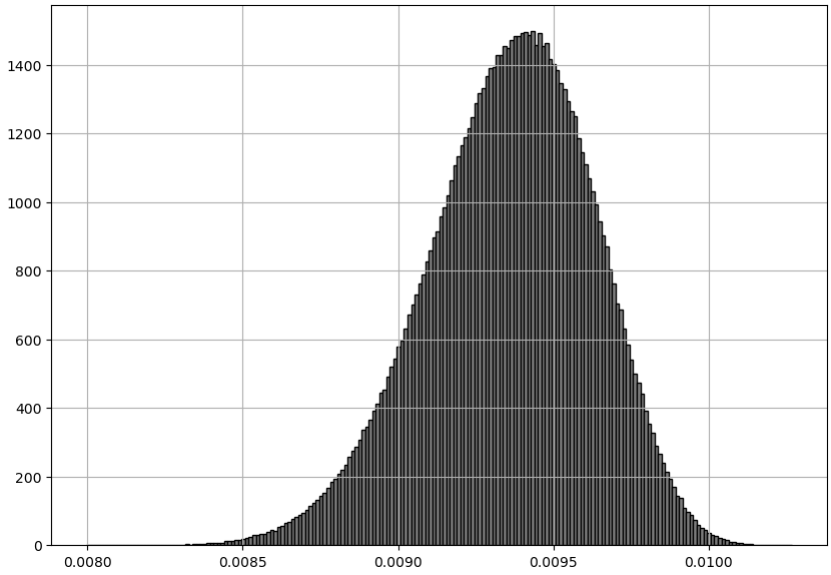
Gráfico 32 – Histograma referente a PC18



Fonte: Autoria própria.

PC_{19}

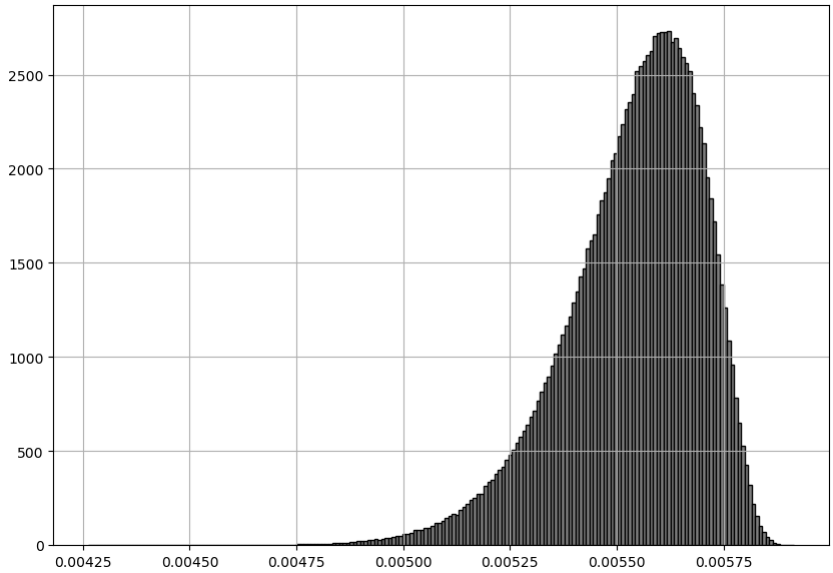
Gráfico 33 – Histograma referente a PC19



Fonte: Autoria própria.

PC_{20}

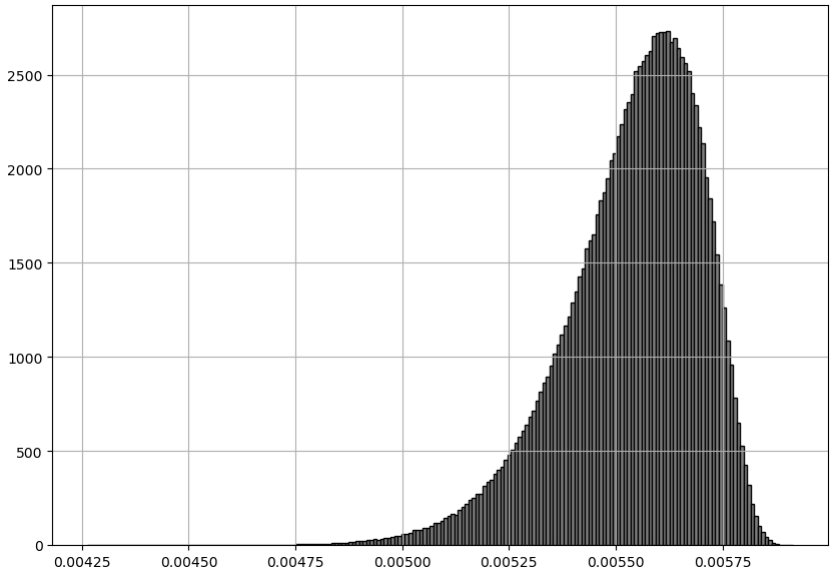
Gráfico 34 – Histograma referente a PC20



Fonte: Autoria própria.

PC_{21}

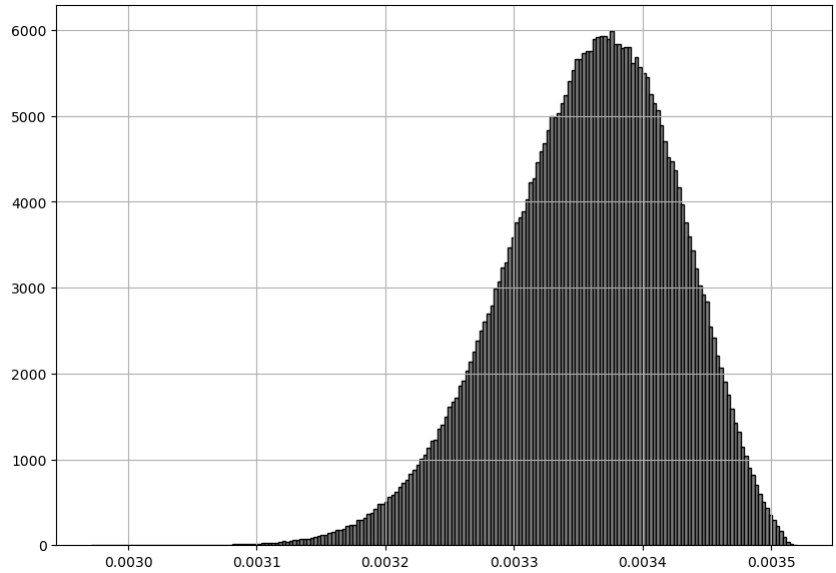
Gráfico 35 – Histograma referente a PC21



Fonte: Autoria própria.

PC_{22}

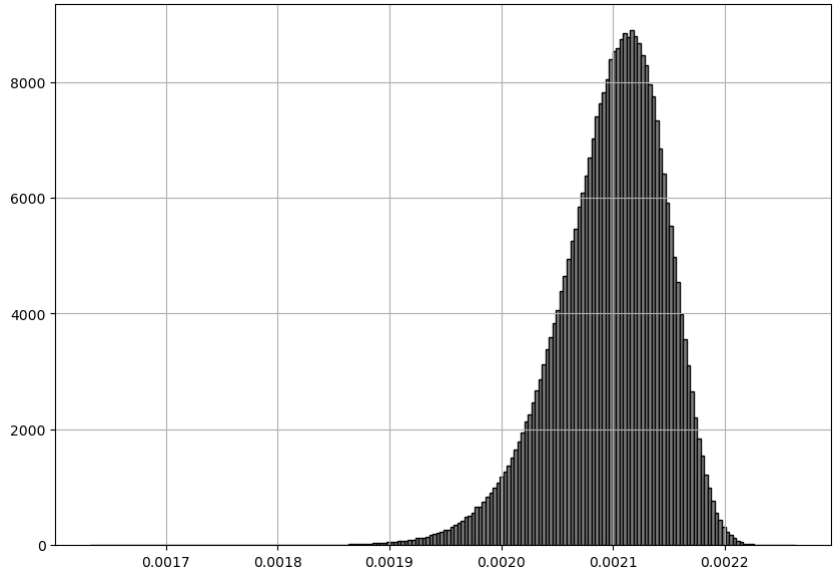
Gráfico 36 – Histograma referente a PC22



Fonte: Autoria própria.

PC_{23}

Gráfico 37 – Histograma referente a PC23



Fonte: Autoria própria.

Através das distribuições de frequência referentes aos autovalores das componentes principais apresentados acima podemos observar que várias componentes como é o caso da PC1, PC5, PC8, PC9, PC11, PC12, PC13, PC15 e

PC16 apresentam uma distribuição que não se assemelha a uma distribuição normal. Tal fato vai em linha com o que foi discutido por [1] com relação as variáveis estudadas onde ela menciona que todas as variáveis passaram por um teste de normalidade onde descobriu-se que algumas variáveis não apresentavam uma distribuição normal. Uma vez que os autovalores presentes nessas distribuições são calculados a partir das matrizes de correlação construídas por essas variáveis é de se esperar que sofram a influência de suas distribuições de acordo com a relevância que tais variáveis possuem no cálculo de cada componente principal.

A título de análise e comparação com os trabalhos desenvolvidos por [1] e [2] foi aplicado o método de PCA para as 23 variáveis apresentadas anteriormente. Os resultados para as 23 componentes principais geradas podem ser observados na tabela abaixo onde temos a apresentação do autovalor médio, desvio médio, seus respectivos valores em termos percentuais e cumulativos além de uma coluna contendo a incerteza relativa, o que nos auxilia na compreensão do quanto o valor de um autovalor pode ser sensibilizado pela sua incerteza.

Tabela 14– Resultados da PCA para as 23 variáveis

(continua)

PC	Autovalor	σ Autovalor	%Var	σ%Var	%VarCum	σ%VarCum	σ relativo
1	12.899	0.089	56.08	0.54	56.08	0.54	0.0096
2	3.29	0.12	14.31	0.53	70.39	0.76	0.0108
3	2.132	0.043	9.27	0.20	79.66	0.79	0.0099
4	1.2971	0.0025	5.640	0.040	85.30	0.79	0.0093
5	1.0326	0.0071	4.489	0.044	89.79	0.79	0.0088
6	0.7590	0.0077	3.300	0.040	93.09	0.79	0.0085
7	0.34986	0.00082	1.521	0.011	94.61	0.79	0.0084
8	0.2847	0.0057	1.238	0.026	95.85	0.79	0.0082
9	0.1972	0.0055	0.857	0.025	96.71	0.79	0.0082
10	0.16521	0.00052	0.7183	0.0054	97.43	0.79	0.0081
11	0.1462	0.0040	0.636	0.018	98.06	0.79	0.0081
12	0.1160	0.0026	0.504	0.012	98.57	0.79	0.0080
13	0.0786	0.0023	0.342	0.010	98.91	0.79	0.0080
14	0.06623	0.00059	0.2880	0.0032	99.20	0.79	0.0080
15	0.0584	0.0020	0.2540	0.0091	99.45	0.79	0.0079
16	0.0502	0.0019	0.2184	0.0084	99.67	0.79	0.0079
17	0.02594	0.00017	0.1128	0.0011	99.78	0.79	0.0079
18	0.02349	0.00054	0.1021	0.0024	99.88	0.79	0.0079
19	0.00936	0.00027	0.0407	0.0012	99.92	0.79	0.0079

(continuação)

20	0.006524	0.000075	0.02836	0.00038	99.95	0.79	0.0079
21	0.00554	0.00016	0.02407	0.00072	99.98	0.79	0.0079
22	0.003358	0.000067	0.01460	0.00031	99.99	0.79	0.0079
23	0.002097	0.000049	0.0091	0.00022	100.00	0.79	0.0079

Fonte: Autoria própria.
Nota: Resultados da análise de PCA para cada um dos 23 componentes principais gerados.

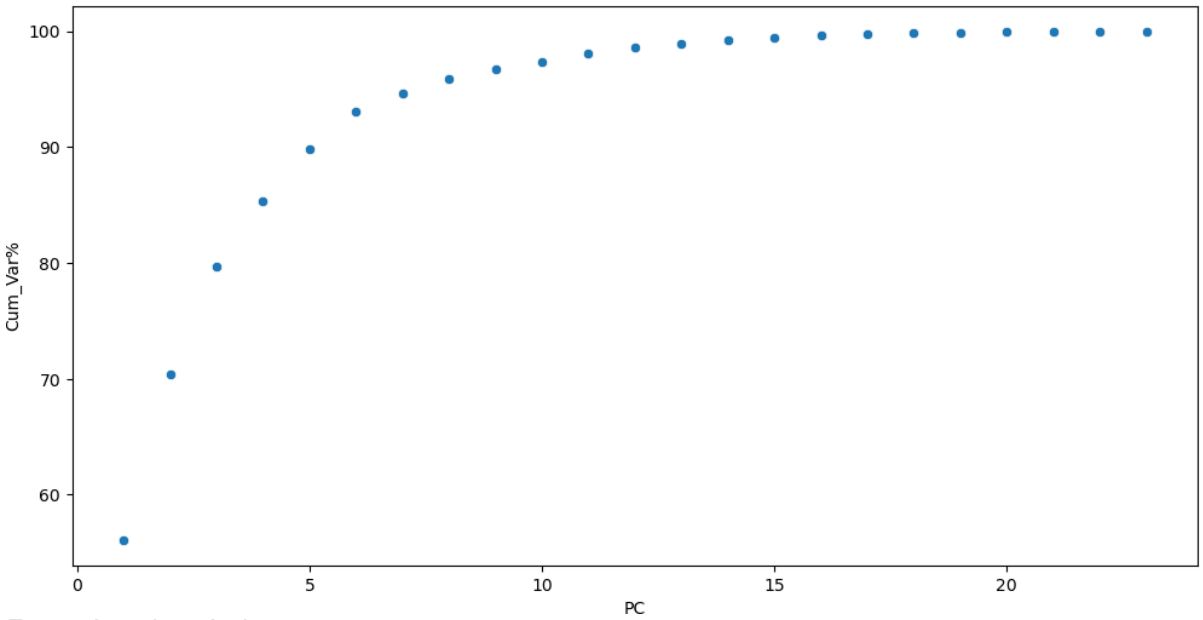
Através dos dados apresentados na tabela acima conseguimos visualizar um dos principais benefícios da PCA, que é a redução de dimensionalidade de um modelo. Ao utilizarmos as novas variáveis geradas conhecidas como componentes principais, aqui chamadas de PC, conseguimos reduzir para menos da metade o número de variáveis analisadas para explicar 95% do modelo referente a caracterização do meio interestelar quando comparamos com a utilização das variáveis originais.

Dessa forma, para que possamos explicar ao menos 95% do modelo devemos utilizar 8 componentes principais, isso porque estamos utilizando o método da PCA desenvolvido pelo [2], o qual considera as incertezas experimentais. É importante notarmos que se utilizarmos apenas 7 componentes principais devido a incerteza associada podemos não atingir os 95% de variância total desejada, isso porque com 7 componentes temos uma incerteza associada de ± 0,79%, podendo assim a variância acumulada estar entre 93,8% e 95,4%.

A importância e relevância da utilização das incertezas experimentais na utilização do método de PCA fica evidente a partir do exemplo acima, onde a não utilização de tais incertezas acarretaria a utilização de uma variável a menos do que o necessário, mostrando dessa forma que a utilização das incertezas experimentais funciona como uma margem de segurança e que nos informa um número de variáveis necessárias para a representação do modelo mais confiável que o método tradicional.

Outro fato observado na tabela 14 é que a partir da 11º componente principal, o acréscimo de novas componentes para a explicação do modelo torna-se muito pouco efetivo, apresentando variações de percentuais mínimos na variância total acumulada. Uma melhor visualização desse fenômeno é apresentada a partir do gráfico abaixo onde temos o avanço da variância total a partir da inclusão de novos componentes principais.

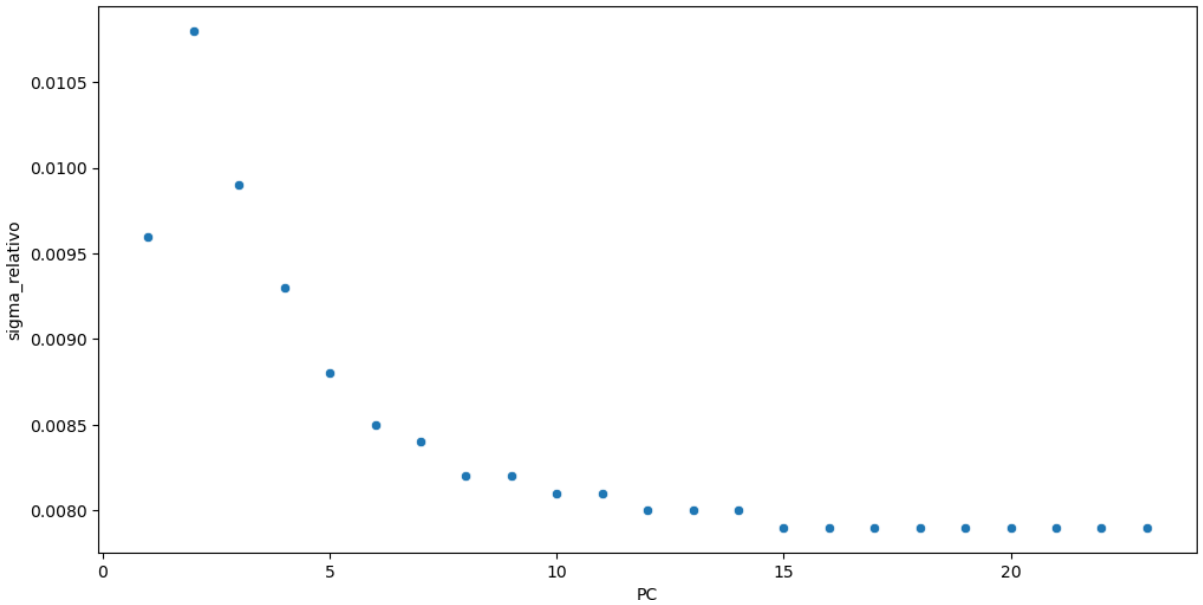
Gráfico 38 – Variância acumulada



Fonte: Autoria própria.

O comportamento obtido no gráfico da figura 66 também pode ser observado a partir da coluna σ%VarCum da tabela 11 que nos traz a informação da incerteza acumulada associada a variância acumulada, onde a partir da 11º componente principal temos muito pouca variação dos valores indicando que o acréscimo de novas componentes principais na análise não traz quase nenhum benefício.

Gráfico 39 – σ relativo



Fonte: Autoria própria.

Segue na Tabela 15 a análise realizada por [2].

Tabela 15– Análise de PCA desenvolvida por [2]

PC	Autovalor	σ_Autovalor	%Var	σ_%Var	%Cum	σ_%Cum	σ_Re
1	14,46	0,05	62,87	0,23	62,87	0,23	0,0036
2	2,863	0,013	12,45	0,06	75,32	0,24	0,0031
3	2,07	0,07	9,0	0,3	84,3	0,4	0,0047
4	1,289	0,010	5,60	0,04	89,9	0,4	0,0044
5	1,025	0,006	4,458	0,027	94,4	0,4	0,0042
6	0,260	0,004	1,130	0,019	95,5	0,4	0,0042
7	0,210	0,006	0,913	0,027	96,4	0,4	0,0042
8	0,178	0,003	0,772	0,014	97,2	0,4	0,0041
9	0,1554	0,0016	0,676	0,007	97,9	0,4	0,0041
10	0,1242	0,0019	0,540	0,008	98,4	0,4	0,0041
11	0,0976	0,0008	0,424	0,003	98,8	0,4	0,0041
12	0,0750	0,0007	0,326	0,003	99,2	0,4	0,0040
13	0,0599	0,0003	0,2602	0,0011	99,4	0,4	0,0040
14	0,0461	0,0020	0,200	0,009	99,6	0,4	0,0040
15	0,02709	0,00015	0,1178	0,0006	99,7	0,4	0,0040
16	0,02481	0,00029	0,1079	0,0013	99,9	0,4	0,0040
17	0,01090	0,00005	0,04739	0,00023	99,9	0,4	0,0040
18	0,00718	0,00013	0,0312	0,0006	99,9	0,4	0,0040
19	0,0066	0,0003	0,0287	0,0014	100,0	0,4	0,0040
20	0,00437	0,00007	0,0190	0,0003	100,0	0,4	0,0040
21	0,002235	0,000028	0,00972	0,00012	100,0	0,4	0,0040
22	0,00170	0,00016	0,0074	0,0007	100,0	0,4	0,0040
23	0,00015	0,00005	0,00065	0,00025	100,0	0,4	0,0040

Fonte: [2]
Nota: Resultados da análise de PCA para cada um dos 23 componentes principais gerados.

Com relação aos dados obtidos pela análise realizada por [2] é possível identificar os mesmos padrões comentados acima, porém com valores um pouco diferentes para os desvios e autovalores. Sendo que a principal diferença entre as duas aplicações se encontra nas ferramentas utilizadas e no número de iterações performadas em cada análise. Foi utilizado por [2] um algoritmo desenvolvido por ele juntamente com o framework de análise de dados chamado ROOT, desenvolvido em C++ pelo CERN, com um milhão de iterações, enquanto a análise desenvolvida nessa

dissertação foi realizada a partir de um algoritmo desenvolvido pelo autor em python com a utilização das bibliotecas pandas, numpy, matplotlib e seaborn através do jupyter como ambiente de desenvolvimento, com um total de três milhões de iterações.

Ao comparar as duas análises é possível perceber que se considerarmos a análise de [2] atingimos 95% de variância com a utilização de apenas 6 componentes principais, isso porque a incerteza acumulada encontrada por ele é bem menor, com apenas 0,4% incidindo sobre uma variância acumulada de 95,5%, assim a variância acumulada poderia estar entre 95,1% e 95,9%.

Segue abaixo na tabela 16 os dados obtidos por [1] em seu trabalho, no qual a incerteza experimental não é levada em consideração.

Tabela 16 – Resultados da PCA para as 23 variáveis realizados por [1]

PC	Autovalor	%Var	%Cum
1	15.248	63.30	66.30
2	3.158	13.73	80.03
3	1.801	7.83	87.86
4	1.139	4.95	92.81
5	0.355	1.54	94.35
6	0.262	1.14	95.49
7	0.192	0.84	96.33
8	0.186	0.81	97.14
9	0.157	0.68	97.82
10	0.117	0.51	98.33
11	0.096	0.42	98.75
12	0.074	0.32	99.07
13	0.066	0.29	99.36
14	0.055	0.24	99.60
15	0.032	0.14	99.74
16	0.025	0.11	99.85
17	0.012	0.05	99.90
18	0.008	0.03	99.93
19	0.006	0.03	99.96
20	0.005	0.02	99.98
21	0.003	0.01	99.99
22	0.002	0.01	100.00
23	0.000	0.00	100.00

Fonte:[1]
Nota: Resultados da análise de PCA para cada um dos 23 componentes principais gerados.

Ao verificarmos os dados obtidos por [1], a qual não considera as incertezas experimentais em sua análise, notamos que o autovalor da componente principal de maior valor, PC1, é maior que os valores obtidos nas outras análises que consideram as incertezas e que os autovalores obtidos para os demais componentes principais também diferem um pouco das outras análise. As divergências de valores entre as três análises é algo natural e esperado, uma vez que a metodologia desenvolvida por [2] para a obtenção da incerteza consiste em realizar o cálculo dos autovalores inúmeras vezes a partir de matrizes de covariância criadas a partir de vetores médios gerados aleatoriamente a cada iteração. Porém, é possível identificar as mesmas características de redução de dimensionalidade nos três casos e uma diminuição significativa do acréscimo de mais componentes principais para a explicação do modelo a partir da 11º componente.

3.1.2 Conclusão Parcial: Análise de componentes principais das bandas interestelares difusas (DIBs) com incertezas experimentais

A partir da análise desenvolvida neste trabalho e do comparativo dos resultados obtidos com os trabalhos de [1] e [2] pode-se verificar que, independentemente da análise considerar ou não as incertezas experimentais, em todas as análises ocorreu uma significativa redução de dimensionalidade do modelo, sendo necessário em todos os casos a utilização de menos da metade do número de variáveis originais para a obtenção de 95% da variância acumulada.

Porém, verificou-se que a utilização das incertezas a partir do método desenvolvido por [2] é recomendável e prudente para uma análise mais confiável, uma vez que, como demonstrado na seção anterior, traz mais segurança com relação ao número de componentes principais necessárias para a explicação do modelo quando optamos por trabalhar com os limites mínimos de incertezas.

Foi possível constatar também que, apesar da metodologia aplicada neste trabalho para a consideração das incertezas ser a mesma utilizada por [2], fatores como a quantidade de iterações e a ferramenta utilizada, juntamente com o fato de que para cada iteração um vetor médio aleatório é gerado e uma nova matriz de covariância é construída para a extração dos autovalores, resultarão em resultados ligeiramente diferentes para cada análise.

Entretanto, apesar de valores diferentes de autovalores terem sido encontrados para cada uma das três análises, os padrões identificados nas componentes principais foram os mesmos, indicando que a partir da 11º componente principal, a inclusão de mais componentes para a explicação do modelo era ineficiente. E para os casos em que a incerteza experimental foi considerada, a coluna de desvio acumulado nos mostrou que pode ser utilizada como critério na identificação dessa 'fadiga' das componentes principais, pois a partir de um determinado momento, a incerteza se tornará tão pequena que o valor acumulado se tornará constante, demonstrando que desse ponto em diante não é mais interessante considerarmos essas componentes principais na análise.

4 CONSIDERAÇÕES FINAIS

Neste trabalho, o método de estatística multivariada chamado de análise de correlação canônica foi aplicado em dois cenários completamente distintos enquanto o método de análise de componentes principais com incertezas experimentais foi aplicado para um terceiro cenário. Sendo o primeiro cenário estudado composto por variáveis de consumo de água, geração de esgoto, índice geral de IDH e seus subíndices de renda, educação e longevidade. Já o segundo cenário diz respeito `a colisão de íons pesados de chumbo ($P_b - P_b$), contando com variáveis como entropia, energia, densidade de energia, densidade de entropia, multiplicidade conservada, número de partículas carregadas e momento transversal. E o terceiro cenário trata-se do meio interestelar onde 23 variáveis como comprimentos de onda absorvidas, excesso de cor, densidade de hidrogênio atômico e molecular, hidrogênio total, quantidade de radiação UV entre outras são utilizadas para a redução de dimensionalidade do modelo.

Para o primeiro cenário estudado foram realizadas diversas análises com o intuito de identificar uma correlação entre o consumo de água, geração de esgoto e o IDH e seus subíndices. Porém, devido a qualidade suspeita dos dados e à falta de registros para algumas variáveis como geração de esgoto e consumo de água em alguns estados, alguns tratamentos prévios na base de dados geral tornaram-se necessários. Este foi um ponto importante e essencial para a continuidade do estudo, pois apesar de termos um método robusto e eficiente, se os dados não estiverem tratados corretamente e não forem fidedignos e condizentes com a realizada dos fatos o método poderá nos trazer resultados incoerentes. Dessa forma, o tratamento adequado dos dados e a confiabilidade deles tornam-se imprescindíveis para uma boa análise utilizando o método de correlação canônica.

Após o tratamento prévio da base de dados foi possível chegar a dois modelos nacionais de interesse público, os quais podem ser utilizados no auxílio de formulação e execução de projetos voltados para as políticas públicas. O primeiro é capaz de definir o consumo de água de um determinado município baseado nos seus índices de IDH e sua geração de esgoto, com uma correlação canônica de 62,4%. Já o segundo modelo é capaz de prever a geração de esgoto de um determinado município

a partir de seu consumo de água e seus índices de IDH, com uma correlação canônica de 54%. Uma outra análise foi realizada fatiando-se a base de dados através da representatividade dos estados, a partir dessa análise foi possível obter uma correlação para os dois modelos citados acima de 83% para o estado de São Paulo, estado que apresentou o maior número de registros da base utilizada na análise de água e esgoto com 32% de representatividade da base.

Já para o segundo cenário, a área de estudo foi a física de altas energias, um ramo da física dedicado a estudar e testar interações fundamentais, uma área completamente diferente do primeiro estudo, o que reforça a aplicabilidade do método em diferentes áreas do conhecimento. Neste estudo envolvendo as variáveis relacionadas ao processo de colisão de íons pesados de chumbo ($P_b - P_b$) obtivemos dois principais modelos com valores de correlação canônica elevados. Em ambos os modelos conseguimos obter o número de partículas carregadas (N) e o momento transversal (p_t) a partir das variáveis de entropia (S), energia (E), densidade de energia (E/R^3), densidade de entropia (S/R^3), multiplicidade conservada (E/S) e centralidade (C). A principal diferença entre eles é que no primeiro modelo obtivemos uma correlação canônica de 99,9%, onde as principais variáveis responsáveis por esses resultados eram a entropia (S) e o número de partículas carregadas (N), algo já esperado e conhecido de acordo com outros trabalhos como o de (Assis, 2022). Porém o segundo modelo, o qual apresentou uma correlação também elevada de 96% foi capaz de nos mostrar algo não tão óbvio e não discutido em outras literaturas, onde o momento transversal (p_t) foi a variável a ser explicada com maior peso canônico, podendo ser calculado através das demais variáveis envolvidas.

Para o terceiro cenário o ambiente de estudo foi o meio interestelar, para ser mais específico, foram analisadas variáveis relacionadas a formação das bandas interestelares difusas (DIBs). Durante a análise de componentes principais com incertezas experimentais conseguimos reduzir significativamente o número de variáveis para representar uma variância acumulada de 95%, passando de 23 variáveis originais para apenas 8 componentes principais. Além da redução de dimensionalidade também foi proposto um outro método para identificação das componentes principais mais significativas através da incerteza acumulada, uma vez a partir de determinado ponto a incerteza acumulada torna-se constante, demonstrando que as componentes principais associadas a essas incertezas não contribuem de forma efetiva para a explicação do modelo.

Através de uma comparação entre os métodos que considerem a incerteza experimental e o que não considera a incerteza em sua análise, foi identificado que ao utilizarmos os limites inferiores dessas incertezas associadas às componentes principais geradas, conseguimos trazer mais segurança na determinação do número de variáveis necessárias para a representação do modelo. Dessa forma, se não considerássemos a incerteza na análise, precisaríamos de 7 componentes principais para representar 95% do modelo. Porém, ao levar em consideração a incerteza, o número mínimo de variáveis passa a ser 8 para obter a mesma variância explicada de 95%. Assim, ao se utilizar a incerteza, teremos um número mínimo de componentes mais confiável para representar o modelo.

Dessa forma podemos constatar que o método de correlação canônica é bastante robusto, capaz de apresentar resultados não óbvios e não passiveis de serem calculados através do método convencional de correlação, isso porque ele é capaz de analisar todas as variáveis ao mesmo tempo gerando novas variáveis através de transformações lineares. Outra vantagem do método é que ele pode ser utilizado independentemente da área de conhecimento como pôde ser visto nas diferentes aplicações deste trabalho.

Com relação ao método de componentes principais (PCA) pôde-se constatar que, independentemente da consideração da incerteza experimental durante a análise, uma redução significativa de dimensionalidade do modelo será encontrada. Porém, é prudente e recomendável a utilização das incertezas experimentais durante as análises para a obtenção de um número de componentes principais mais confiável.

Como próximo passo a este trabalho fica a recomendação da aplicação do método de componentes principais (PCA) para a redução de dimensionalidade com um pequeno ajuste no algoritmo, de forma que o gerador de números aleatórios seja parametrizado para corresponder não a uma gaussiana, mas sim a um outro tipo de distribuição, explorando, assim, as distribuições apresentadas por algumas variáveis utilizadas no trabalho que não correspondiam a uma distribuição normal.

REFERÊNCIAS

1 ENSOR, T. *et al*. A principal component analysis of the diffuse interstellar bands. The Astrophysical Journal, **The American Astronomical Society**, v. 836, n. 2, p. 162, feb 2017. Disponível em: ⟨https://dx.doi.org/10.3847/1538-4357/aa5b84⟩. Acesso em: 18 dez. 2023.

2 FLAUSINO, F. S. **Nova formulação de ferramentas de estatística multivariada com incertezas experimentais.** 110 p. Dissertação (Mestrado em Física) — Universidade Federal de Alfenas, Poços de Caldas, 2018.

3 MINGOTI, S. **Análise de dados através de métodos de estatística multivariada**: uma abordagem aplicada. [S. l.]. Editora UFMG, 2005. ISBN 9788570414519. Disponível em: ⟨https://books.google.com.br/books?id=W7sZllHmmGIC⟩. Acesso em: 02 out. 2023.

4 SANTO, R. D. E. Principal component analysis applied to digital image compression. **Einstein (São Paulo)**, Instituto Israelita de Ensino e Pesquisa Albert Einstein, v. 10, n. 2, p. 135–139, Apr 2012. ISSN 1679-4508. Disponível em: https://www.scielo.br/j/eins/a/yzFzBTrdgGrv46hGsnzKssd/?lang=en. Acesso em: 15 nov. 2023.

5 ROSSI, R. G. **Análise de componentes principais em data warehouses**. 63 p. Dissertação (Mestrado em Ciências) — Universidade de São Paulo, São Paulo, 2017.

6 SUBRAMANIAN, V.; SYEDA-MAHMOOD, T.; DO, M. N. Multimodal fusion using sparse cca for breast cancer survival prediction. In: 2021 **IEEE 18th International Sympo- sium on Biomedical Imaging (ISBI)**. [S.l.: s.n.], 2021. p. 1429–1432.

7 AL., N. G. et. **The positive-negative mode link between brain connectivity, demographics, and behavior**: A pre-registered replication of smith et al. (2015). Royal Society, v. 9, n. 2, 2022.

8 TAYLOR, J.; ROQUE, W. **Introdução à Análise de Erros**: O Estudo de Incertezas em Medições Físicas. [S. l.]. BOOKMAN COMPANHIA ED, 2012. ISBN 9788540701366. Disponível em: ⟨https://books.google.com.br/books?id=DuCvuAAACAAJ⟩. Acesso em: 22 dez. 2023.

9 BENTON, A. *et al*. Deep Generalized Canonical Correlation Analysis. 2019. In **Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)**, pages 1–6, Florence, Italy. Association for Computational Linguistics.

10 LOESCH, C.; HOELTGEBAUM, M. **Métodos Estatísticos Multivariados**. [S. l.]. Saraiva Educação S.A., 2017. ISBN 9788502146112. Disponível em: ⟨https: //books.google.com.br/books?id=OylrDwAAQBAJ⟩. Acesso em: 24 nov. 2023.

11 TSUTIYA, M**. Abastecimento de água.** [S. l.]. Departamento de Engenharia Hidráulica e Sanitária da Escola Politécnica da Universidade de São Paulo, 2000. ISBN 9788590082361. Disponível em: ⟨https://books.google.com.br/books?id=LnETtwAACAAJ⟩. Acesso em: 10 dez. 2023.

12 ASSIS, Y. **Correlações entre propriedades do estado inicial e observáveis finais na colisão de íons pesados relativísticos.** 59 p. Dissertação (Mestrado em Física) — Universidade Federal de Alfenas, Poços de Caldas, 2022.

13 BUSZA, W.; RAJAGOPAL, K.; SCHEE, W. van der. Heavy ion collisions: The big picture and the big questions. **Annual Review of Nuclear and Particle Science**, Annual Reviews, v. 68, n. 1, p. 339–376, oct 2018. Disponível em: ⟨https://doi.org/10.1146%2Fannurev-nucl-101917-020852⟩. Acesso em: 26 out. 2023.

14 VOGT, R. **Ultrarelativistic Heavy-Ion Collisions.** Berkeley, CA, USA. Elsevier Science, 2007. ISBN 9780080525365. Disponível em: ⟨https://books.google.com.br/books?id= F1P8WMESgkMC⟩. Acesso em: 15 dez. 2023.

15 ESKOLA, K.; KAJANTIE, K.; LINDFORS, J. Quark and gluon production in high energy nucleus-nucleus collisions. **Nuclear Physics B**, v. 323, n. 1, p. 37–52, 1989. ISSN 0550-3213. Disponível em: ⟨https://www.sciencedirect.com/science/article/pii/ 0550321389905865⟩. Acesso em: 10 out. 2023.

16 CAMPOS, L. B. d. O. **Estudo da Modificação de Jatos em Colisões entre Íons-Pesados Relativísticos.** 76 p. Dissertação (Mestrado em Ciências) — Universidade de São Paulo, São Paulo, 2021.

17 AL., J. A. et. Measurement of an excess in the yield of j/ψ at very low pT in pb–pb collisions at $\sqrt{s}$ = 2.76 TeV. **Phys. Rev. Lett., American Physical Society**, v. 116, p. 222301, Jun 2016. Disponível em: ⟨https://link.aps.org/doi/10.1103/PhysRevLett.116.222301⟩. Acesso em: 20 dez. 2023.

18 ROMATSCHKE, P.; ROMATSCHKE, U. **Relativistic Fluid Dynamics in and out of Equilibrium:** And Applications to Relativistic Nuclear Collisions. United Kingdom. Cambridge University Press, 2019. (Cambridge Monographs on Mathematical Physics). ISBN 9781108483681. Disponível em: ⟨https://books.google.com.br/books?id=RoOWDwAAQBAJ⟩. Acesso em: 14 nov. 2023.

19 JOHNSON, R.; WICHERN, D. **Applied Multivariate Statistical Analysis**. Reino Unido: Prentice Hall, 2002. ISBN 9780130925534. Disponível em: ⟨https://books.google.com.br/books?id= VlcZAQAAIAAJ⟩. Acesso em: 18 nov. 2023.

20 HARDLE, W.; SIMAR, L. **Applied Multivariate Statistical Analysis.** Springer: Berlin Heidelberg, 2015. ISBN 9783662451717. Disponível em: ⟨https://books.google.com.br/ books?id=KI3dBgAAQBAJ⟩. Acesso em: 24 dez. 2023.

21 VUOLO, J. **Fundamentos da teoria dos erros.** [S. l.]. E. Blucher, 1996. ISBN 9788521200567. Disponível em: ⟨https://books.google.com.br/books?id=jOiTPgAACAAJ⟩. Acesso em: 14 dez. 2023.

22 COSTA, C. O. **Bandas Interestelares Difusas.** 152 p. Dissertação (Mestrado em Física e Matemática Aplicada) — Universidade Federal de Itajubá. Itajubá 2009.

23 HERBIG, G. H. The Diffuse Interstellar Bands**. Annual Review of Astronomy and Astrophysics,** v. 33, Vol. 33:19-73 (Volume publication date September 1995). Disponível em: https://doi.org/10.1146/annurev.aa.33.090195.000315. Acesso em: 10 dez. 2023.

24 CHLEWICKI, G. *et al*. Shapes and Correlations as Observational Discriminants for the Origin of Diffuse Bands. ,v. 305, p. 455, jun. **The Astrophysical Journal** 305:455-466. Disponível em: DOI:10.1086/164259 1986. Acesso em: 05 dez. 2023.

25 WESTERLUND, B. E.; KRE-lOWSKI, J. The division of diffuse interstellar bands into families. Astronomy and Astrophysics**,** v. 218, p. 216–220, **Astronomy and Astrophysics** 218:216-220. 1989. Disponível em: https://www.researchgate.net/publication/234170325_The_division_of_diffuse_interstellar_bands_into_families. Acesso em 17 dez. 2023.

26 JENKINS, E. B. A unified representation of gas-phase element depletions in the interstellar medium*. **The Astrophysical Journal**, The American Astronomical Society, v. 700, n. 2, p. 1299, jul 2009. Disponível em: ⟨https://dx.doi.org/10.1088/0004-637X/700/ 2/1299⟩. Acesso em: 19 dez. 2023.