Relatório: Análise de Componentes Principais (PCA) para importações do Porto de Santos da Rússia e Ucrânia

2º Ciclo - Ciência de Dados

Igor Silva de Carvalho

Novembro, 2024

1 INTRODUÇÃO

Em muitas circunstâncias se é necessário processar, analisar ou computar grandes conjuntos de dados. Em casos assim, pode ocorrer a chamada “maldição dos problemas de dimensionalidade”, onde para *datasets* cada vez maiores, – e com maiores dimensões (colunas) - existe um aumento exponencial do tempo e recursos para retirar dados significativos deles (GeeksforGeeks, 2024). Para lidar com os problemas de altas dimensionalidade, processos de redução de dimensão se demonstram uma opção viável.

De acordo com a IBM (2024) o chamado *Principal Component Analysis* (PCA) é um desses processos de redução de dimensionalidade, sendo utilizada principalmente para treinamento não supervisionada de sistemas de *Machine Learning* (ML).

Neste trabalho se aplicou a técnica de PCA sobre os *datasets* de “Carga” e “Carga Conteinerizada”, encontrados nos estatísticos aquaviários da ANTAQ - e extraídos do Portal de Dados Abertos (PDA). A aplicação do PCA sobre esses *datasets* foi feita com o intuito de reduzi-los aos seus componentes principais, afim de facilitar análises posteriores e também de melhor entender as relações entre as variáveis mais relevantes.

Os *datasets*, em formato *Comma Separeted Values* (CSV), foram extraídos com os seus respectivos *datasets* e dicionários de dados complementares. Mas, na hora da aplicação do PCA optou-se por utilizar apenas uma amostra correspondendo a cargas saídas da Rússia e Ucrânia e desembarcadas no Porto de Santos entre o período de 2018 e 2023. A escolha dessa amostra foi feita afim de comparações de perfil de cargas ao longo dos anos para um caso particular.

Os dados foram transformados e limpados utilizando-se a linguagem de programação Python e suas bibliotecas: Pandas, para a transformação dos dados; Scikit-learn para a normalização e aplicação do PCA; Matplotlib, para grafar os dados obtidos. Todos os gráficos e código relevantes encontram-se no repositório Github do projeto[1].

2 EMBASAMENTO TEÓRICO

A seguir são apresentados os conceitos necessários para a melhor compreensão das técnicas e ferramentas necessárias para se efetuar a análise de componente principal nos dados de Carga e Carga Conteinerizada.

2.1 PCA

De acordo com Azevedo (2021) A análise de componentes principais (ACP) – que de agora em diante será referido exclusivamente como PCA neste trabalho - é uma técnica de resolução de problemas multivariados (multidimensionais), mais especificamente, uma técnica de redução de dimensão. Ainda de acordo com o autor, seus benefícios são: (1) a possibilidade de utilização de metodologias univariadas, (2) a possibilidade de se trabalhar com um número menor de variáveis e também (3) a obtenção de detalhes de comportamento de dados que são difíceis de serem identificados a partir das variáveis originais.

O cálculo do PCA pode ser efetuado através das seguintes etapas (OLIVEIRA, 2022):

1 Acesso por aqui: https://github.com/igorzeck/PCA

1. Obtém-se apenas colunas numéricas para o dataset D de interesse. Isso pode ser feito ao vetorizar, categorizar, ou ignorar as variáveis não numéricas.
2. Normaliza-se o *dataset* caso este não esteja normalizado – com média zero e variância um.
3. Obtém-se a matriz de covariância, sendo esta uma matriz quadrada simétrica $n \times n$ com seus elementos sendo a permutação da covariância entre suas colunas:

$$C_{nn} = \begin{bmatrix} Cov(c_1 c_1) & \cdots & Cov(c_1 c_n) \\ \vdots & \ddots & \vdots \\ Cov(c_n c_1) & \cdots & Cov(c_n c_n) \end{bmatrix}$$

   De onde, a covariância de uma coluna k qualquer por uma coluna h qualquer é dada por:

$$Cov(c_k, c_h) = \frac{\sum_{i=1}^{n} (c_{ki} - \overline{c_k})(c_{hi} - \overline{c_h})}{n - 1}$$

   O valor da covariância indica uma relação de proporcionalidade entre as variáveis, com o sinal indicando o sentido da proporcionalidade, e a magnitude a intensidade da relação de variação de uma variável em função da outra (IBM, 2023).
4. Encontra-se os autovalores λ e autovetores v da transformação linear representada na matriz de covariância através da identidade $Cv = \lambda v$[2]:
   a) Os autovalores podem ser encontrados como raízes da determinante $det(C - \lambda \cdot I) = 0$, com até n raízes.
   b) Os autovetores – componentes dos *spans* geradores do autoespaço – podem ser encontrados a partir do *Kernel* para cada autovalor correspondente: $(C - \lambda \cdot I) \cdot v = 0$, de onde $v \in Ker((C - \lambda \cdot I))$ para cada λ.
5. Por fim, utiliza-se os k primeiros autovetores - em ordem decrescente de seus autovalores correspondentes – e os transpõem em uma matriz $n \times k$ que é então multiplicada à esquerda do *dataset* D original gerando o *dataset* D' projetado, contendo k componentes principais por colunas[3].

Os PCAs gerados indicam as direções de maior variância dos pontos de dados, com a maior variância capturada no seu componente, maior a informação retida (IBM, 2023).

2.2 Conjunto de Dados

Os conjuntos de dados escolhidos, de acesso aberto, contêm *datasets* diversos, abrangentes e pesquisados (ou elaborados) por um instituto de pesquisa respeitável, a Agência Nacional de Transportes Aquaviários (ANTAQ) vinculada ao Ministério de Portos e Aeroportos (MPA).

Neste trabalho foram utilizados, especificamente, os seguintes *datasets*[4]:

1. Carga – os arquivos de 2018 a 2023;

2 O exato valor dos autovalores pode variar com o cálculo utilizado, no entanto, a direção por eles representadas, e por extensão, o *span* deles é mesmo.

3 Neste trabalho foi utilizada a biblioteca Scikit-learn para a projeção, sendo que esta não utiliza este exato cálculo, o que leva a diferenças nos valores projetados.

2. Carga Conteinerizada – os arquivos de 2018 a 2023;
3. E também foram utilizados os arquivos de tabelas auxiliares:
   a) Instalação Origem;
   b) Instalação Destino;
   c) Mercadoria;
   d) Mercadoria Conteinerizada[5];

2.3 TERMINOLOGIA AQUAVIÁRIA

Afim de melhor entender certos termos apresentados e analisados, são apresentadas nesta seção terminologias aquaviárias.

2.3.1 NCM

A Nomenclatura Comum do Mercosul (NCM), de acordo com a Receita Federal brasileira, é um código de identificação de mercadorias (BRASIL, 2019). A despeito do nome seu uso não se limita apenas a mercadorias que tenham sido movimentadas pelo Mercosul.

Sua estrutura é expostas na Figura 1:

Figura 1: NCM

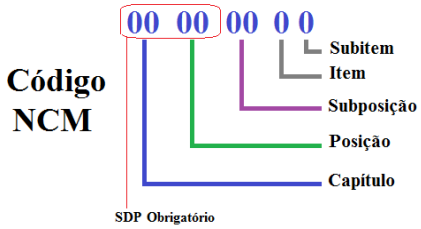


Figura: Documentação dos Estatísticos Aquaviários da ANTAQ (2014)

Onde os dígitos, da esquerda pra direita, demarcam maior grau de especificidade da mercadoria, sendo que nos *datasets* da ANTAQ, os códigos são identificados por seus quatro primeiros dígitos (código NCM SH 4)[6].

Nos conjuntos de dados da ANTAQ, o NCM é utilizado para a classificação das mercadorias contidas nas cargas não conteinerizadas. Já para as cargas conteinerizadas, se encontram apenas os códigos da *International Standard Organization* ISO para contêineres na tabela Carga, havendo na tabela Carga Conteinerizada os códigos NCM correspondentes das mercadorias (BRASIL, 2014).

Neste trabalho, o entendimento da estrutura dos códigos NCM se demonstra essencial afim de melhor compreender as variáveis de maior influência sobre os PCAs gerados após a redução de dimensões.

2.3.2 Tipos de Carga

De acordo com TOTVS (2023) os tipos de carga – relevantes para este trabalho – são:

1. Cargas Gerais: sendo o tipo de carga contável individualmente, como caixas, fardos, etc.

---

4 A modelagem de dados pode ser encontrada em: https://web3.antaq.gov.br/ea/sense/download.html#pt em "*Download* modelagem de dados".

5 Na prática, devido a limitação de recursos computacionais, optou-se por utilizar apenas a tabela de Mercadoria.

6 SH para Sistema Harmonizada, grupo de códigos do qual o NCM sucede (BRASIL, 2014).

2. Cargas Conteinerizadas: refere-se a carga com contêiner padronizado. Identificadas pelo **contêiner** enquanto, por exemplo, a Carga Geral seria identificada pela **mercadoria**[7].
3. Cargas a granel: Cargas que não precisam de um invólucro intermediário. São cargas que não podem ser contadas individualmente, pois são compostas por líquidos ou partículas homogêneas.

2.4 ETL

Ao analisar os dados extraídos dos *datasets* da ANTAQ foram utilizados processos e técnicas de Ciência de Dados e Álgebra Linear afim de organizar e interpretar corretamente os dados ali contidos - as técnicas de Álgebra Linear já foram elucidadas na seção 2.1. Para o processamento de dados brutos foram utilizadas técnicas de *Extract, Transform, Load* (ETL).

ETL é um processo de: (1) agregação de dados de múltiplas fontes; (2) transformações destes dados com algum intuito específico (como análise, por exemplo); (3) armazenamento em um conjunto de dados significativos (IBM, [2023?]). Nesta pesquisa, os processos de ETL descritos foram utilizados afim de preparar os dados para sua análise - após as extrações e transformações necessárias.

2.6 Ferramentas de ciência de dados

No que diz respeito a análise dos dados, foram utilizadas, nas etapas de coleta, filtragem e processamento de dados, uma série de ferramentas especializadas, sendo estas:

a) Python

Python é uma linguagem de programação interpretativa, interativa, orientada a objetos e funcional. Contém tipos de dados de alto grau de dinamicidade, tendo uma sintaxe clara e poderosa (Python, 2024). A escolha desta ferramenta foi feita levando-se em conta o grande acesso e acessibilidade a ferramentas de análise de dados, além da grande disponibilidade de bibliotecas externas existentes, e facilidade geral de uso.

b) Pandas

Pandas é uma flexível biblioteca código aberto de análise e manipulação de dados feita para a linguagem de programação Python (Pandas, 2024). Esta pesquisa utiliza a biblioteca Pandas com o objetivo de filtrar e processar os *datasets* coletados. Sendo estes *dataset* em formato CSV, o uso desta biblioteca se demonstra apropriado, uma vez que é possível a importação de dados brutos em CSV.

c) Jupyter Notebook

Criado a partir do projeto código aberto e sem fins lucrativos Project Jupyter em 2014 é um ambiente interativo de ciência de dados e computação científica. Neste trabalho foi utilizado em conjunto com a linguagem de programação Python para a criação dos *scripts* necessários para a extração, transformação e análise dos dados (JUPYTER, 2024).

---

7 Nos *datasets* da ANTAQ há a possibilidade de se haver múltiplos produtos por registro de Carga, ou seja, maior possibilidade de variedade de produtos.

3 METODOLOGIA

Nesta pesquisa utiliza-se de uma amostra não-probabilística, sendo esta extraída, transformada e processada por meio de processos ETL.

3.1 EXTRAÇÃO

Os dados dos estatísticos aquaviários da ANTAQ foram extraídos através do Portal de Dados Abertos (PDA) do Gov.br: acessando, no rodapé da página, a seção "conjunto de dados", e no campo de pesquisa inserindo "EA", selecionando o resultado "Estatísticos Aquaviários (EA)" da ANTAQ. Logo após "Recursos" e por fim "Acessar o recurso" tanto para o Estatístico quanto para os Metadados – este último para a utilização adequada dos *datasets*[8].

Neste trabalho os *datasets* serão disponibilizados em forma compactada - e com o *link* de acesso no documento “README.md” - no repositório do Github do projeto.

3.2 TRANSFORMAÇÃO

Para todos os *datasets* extraídos da ANTAQ, quando lidos no Pandas, tiveram na função “read_csv” o parâmetro “sep” (separador) com o valor do delimitador “;” - exceto para tabelas criadas durante o trabalho - e o parâmetro “decimal ”como o valor de “,”.

Para cada ano do período estudado (2018 – 2023), carregou-se através do Pandas as tabelas de Carga e aplicou-se as seguintes transformações:

1. Manteve-se apenas os registros contendo na coluna “Destino” o valor “BRSSZ”.
2. Modificou-se a coluna “Origem” para que ficasse com o valor de “CDBigramaOrigem” correspondente - da tabela Instalação Origem – nos casos em que o valor desta última fosse igual a “RU” para Rússia e “UA” para Ucrânia[9], descartando as demais entradas da tabela carga. Esta junção foi feito por meio da função merge do Pandas.
3. Adicionou-se a coluna “Ano” à tabela, contendo o mesmo valor referente ao ano do arquivo, para todos os registros da tabela.
4. Retirou-se os registros que tinha o valor de “ConteinerEstado” como “Vazio”.

O processo acima é então repetido juntado todos os DataFrames criados a um DataFrame final, sendo este utilizado no resto do arquivo para as demais transformações e análise.

Da tabela Carga manteve-se apenas os registros contendo na coluna “Sentido” o valor “Desembarque”, e “Tipos de Operação da Carga” contendo em sua *string* a palavra “Importação”. Converteu-se a coluna “Ano” para o tipo de dados *datetime*.

Juntou-se - com a função merge *e* com seu parâmetro “how*”* para “*left”* - a tabela Carga com as tabelas Carga Conteinerizadas, utilizando a coluna “IDCarga” como referência. Esta junção foi feita ano a ano, com a substituição dos valores da coluna “VLPesoCargaBruta” e “CDMercadoria” por seus correspondentes na tabela carga por meio da função “combine_first”.

---

8 Os dados podem também serem obtidos diretamente da ANTAQ por meio de https://web3.antaq.gov.br/ea/sense/download.html#pt

9 Código do Porto de Santos e código bigrama dos países conferidos na tabela “Instalação Origem” e confirmados em: https://web3.antaq.gov.br/portalv3/sdpv2servicosonline/ConsultarPorto.aspxCódigo

4 Análise

A análise foi feita ao selecionar um caso específico para efetuar a análise: cargas importadas da Rússia e da Ucrânia durante os anos de 2018 a 2023, ou seja, contidos neste período encontram-se os anos do ápice da pandemia de COVID-19 (2020 – 2021) e também o período da guerra na Ucrânia (2022 – Atual)[10].

4.1 Cargas Importadas em Santos da Rússia e Ucrânia

Ao analisar os dados de importação da Ucrânia e Rússia no Porto de Santos para o período de 2018 e 2023 podemos verificar alguns padrões na distribuição das entradas de cargas.

Nesta amostra estão contidas 10324 (95,72%) entradas da Rússia e 462 (4,28%) entradas ucranianas de um total de 10786 entradas. Vale denotar que para o ano 2023, não há entradas vindas da Ucrânia.

Retira-se a tabela criada na seção 3.2.2 as colunas: “IDAtrcacao” por ser apenas um identificador; “Destino” uma vez que possui apenas um valor “BRRSZ”; “Sentido”, uma vez que só possui um valor “Desembarcados”; ”Carga Geral Acondicionamento” e “ConteinerEstado” por suas redundâncias para com a coluna “Natureza de Carga”; “FlagConteinerTamanho” por sua redundância com a coluna “Natureza de Carga” e “TEU” (*Twenty Foot Equivalent*).

Após isto, categoriza-se o *dataset*:

1. Valores que já tenham o seu tipo (*dtype*) como “*category”* tem o seu código numérico de categoria extraído.
2. Para valores que sejam do *object* (tipo misto ou *string*) trata-os como se fossem do tipo “*category”* através da função do objeto *Series* “*astype*” e efetua o passo 1. O pandas categoriza os valores de tal forma que os códigos estejam em ordem crescente por valor[11].

Sobre o *dataset* categorizado utiliza-se a função *fillna* para o valor zero e normaliza-o utilizando a classe *StandardScaler* do módulo Scikit-learn. O normalizador é aplicado sobre o *dataset* através da função *fit_transform* que retorna o *array* que representa o *dataset* normalizado.

Através da classe PCA do Scikit-learn, cria-se o estimador que, por meio da função *fit_transform gera* um *dataframe* projetado com as colunas sendo os 21 componentes principais.

Segue abaixo uma tabela comparativa, tendo por linhas os PCAs e os três mais influentes *features* – variávies - originais, elucidando a influência (magnitude) que cada coluna possui sobre os dois maiores PCAs:

Tabela 1: Colunas de maior influência no PCA0 e PCA1[12]

PCA0		PCA1	
Colunas	Valores	Colunas	Valores
STSH2	-0,58	Natureza da Carga	0.52
STNaturezaCarga	-0,53	Ano	0.38
STSH4	-0,43	STSH4	-0.35

Fonte: Elaborada neste trabalho com os dados da ANTAQ (2018 - 2023).

10 Uma análise do dataset de Carga de 2018 como um todo pode ser encontrados no repositório do Github deste trabalho.

11 Em *strings* isso significa que valors que começam com “A” terão código menor que valores com “B”, criando uma sequência crescente.

12 neste trabalho as casas decimais serão todas separadas por “.” ao invés de “,”.

Através da Tabela 1 podemos concluir que para PCA0, o maior componente principal, os dois *features* que tiveram maior influência foram STSH2 – coluna que indica se houve movimentação de um único capítulo (exclusivo) ou não (compartilhado) na atracação de onde veio carga – seguido por STNaturezaCarga – coluna que indica se houve movimentação de um único tipo de natureza de carga (exclusivo) ou não (compartilhado) na atracação de onde veio carga. Já, para o segundo componente principal (PCA1), a coluna de maior influência foi a de Natureza da carga, seguida por Ano.

Em ambos componente a coluna STSH4 – que indica se houve movimentação de um único tipo de mercadoria (exclusivo) ou não (compartilhado) na atracação de onde veio carga – é a terceira mais influente, sendo negativa para os dois (indicando relação de inversa proporcionalidade para com os valores dos dois PCAs correspondentes).

A categorização foi feita de tal forma que o valor “exclusivo” corresponde ao número um e o valor “compartilhado” foi categorizado como sendo zero em seus códigos de categorias.

Sabendo das principais colunas, podemos melhor analisar os valores da matriz de covariância. Na Tabela 2 podemos ver a influência das três colunas mais influentes para o PCA0, entre si[13]:

Tabela 2: Valores de covariância para principais *features* de PCA0

	STSH2	STNaturezaCarga	STSH4
STSH2	1.000093	0.836816	0.675441
STNaturezaCarga	0.836816	1.000093	0.550732
STSH4	0.675441	0.550732	1.000093

Fonte: Elaborada neste trabalho com os dados da ANTAQ (2018 - 2023).

Pode-se perceber uma correlação de variância positiva entre as colunas STSH2 e STNaturezaCarga e STSH4. Isso pode implicar que, em geral, **atracações** com múltiplos capítulos de mercadoria são acompanhadas, também, por múltiplas naturezas de carga e também, em menor frequência, várias movimentações de mercadorias diferentes. Ou seja, através disso podemos classificar as cargas em dois grandes grupos: (1) aquelas em que a atracação registrada foi majoritariamente de uma categorias, (2) e aquelas nas quais houve maior variedade de categorias.

Já na Tabela 3 podemos observar as colunas de maior influência do componente principal PCA1:

Tabela 3: Valores de covariânica para principais *features* de PCA1

Colunas	Natureza da Carga	Ano	STSH4
Natureza da Carga	1.000093	0.473830	-0.099589
Ano	0.473830	1.000093	-0.011886
STSH4	-0.099589	-0.011886	1.000093

Fonte: Elaborada neste trabalho com os dados da ANTAQ (2018 - 2023).

Pode-se observar uma relação de proporcionalidade positiva (de carácter médio) entre Ano e Natureza da Carga:

13 A matriz de covariância completa encontra-se no arquivo Jupyter no repositório deste relatório.

1. Os valores possíveis de Natureza de Carga em ordem crescente de códigos de categoria, são: “Carga Conteinerizada”, “Carga Geral”, “Granel Líquido e Gasoso”, “Granel Sólido”.
2. E os valores possíveis de Ano são: 2018, 2019, 2020, 2021, 2022 e 2023.

Sabendo disso, essa correlação pode ser interpretada da seguinte forma, ao fixar a variável Ano e ao deixar Natureza de Carga variável: com o passar dos anos, houve uma tendência a ter mais cargas de Granel e menos Conteinerizadas. Podemos inferir que houve uma pressão contra o livre fluxo de mercadorias, já que as mais variadas são as de Carga Geral e Carga Conteinerizada.

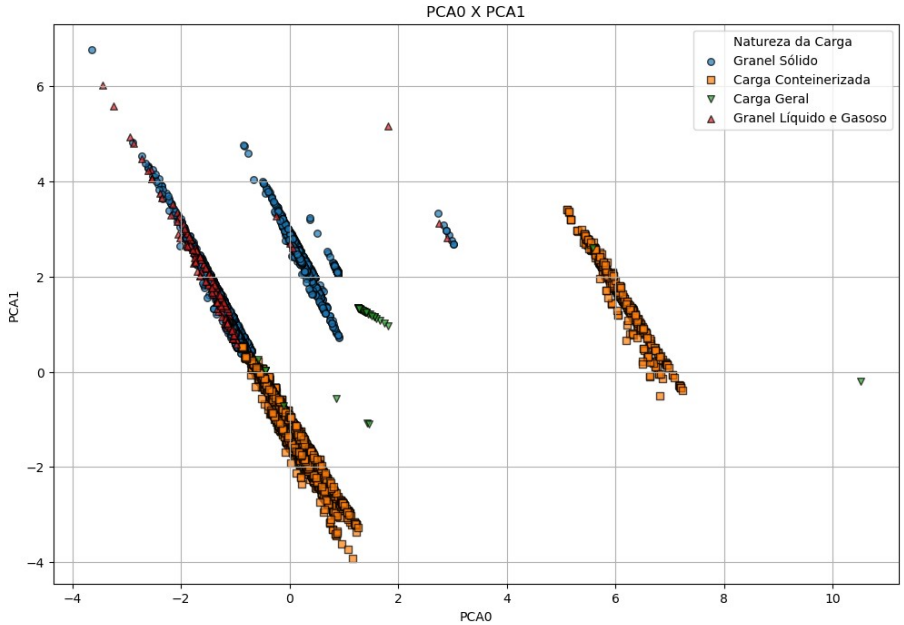
Na mesma tabela, pode se ver uma leve relação de inversa proporcionalidade entre ano e STSH4. As demais relações entre variáveis se demonstram insignificantes.

Por fim, os dois maiores autovalores e seus autovetores correspondentes:

i. Para o primeiro maior autovalor:
   1. λ_0 = 2.473235
   2. v_0 = (0.068, -0.029, -5.6e-17, 1.1e-16, 0.0, 0.0, 0.0, 0.0, 0.0, -3.4e-21, 2.1e-22, 0.0, -0.53, -0.58, -0.43, -0.26, 0.16, 0.071, -0.15, -0.19, 0.18)
ii. Para o segundo maior autovalor:
   1. λ_1 = 2.048641
   2. v_1 = (-0.24, 0.099, 0.0, 2.8e-17, -5.6e-17, 0.0, 0.0, -4.3e-19, 0.0, -6.8e-21, 8.5e-22, 0.0, -0.24, -0.19, -0.35, 0.52, -0.31, -0.057, 0.33, 0.38, -0.3)

Por fim, pode-se grafar os PCAs, tendo a linha o PCA0 e a coluna o PCA1. Na Figura 2 temos ambos coloridos por Natureza de Carga (Componente principal do PCA1):

Figura 2: Carga colorida por Natureza da Carga



Fonte: Elaborada neste trabalho com os dados da ANTAQ (2018 – 2023).

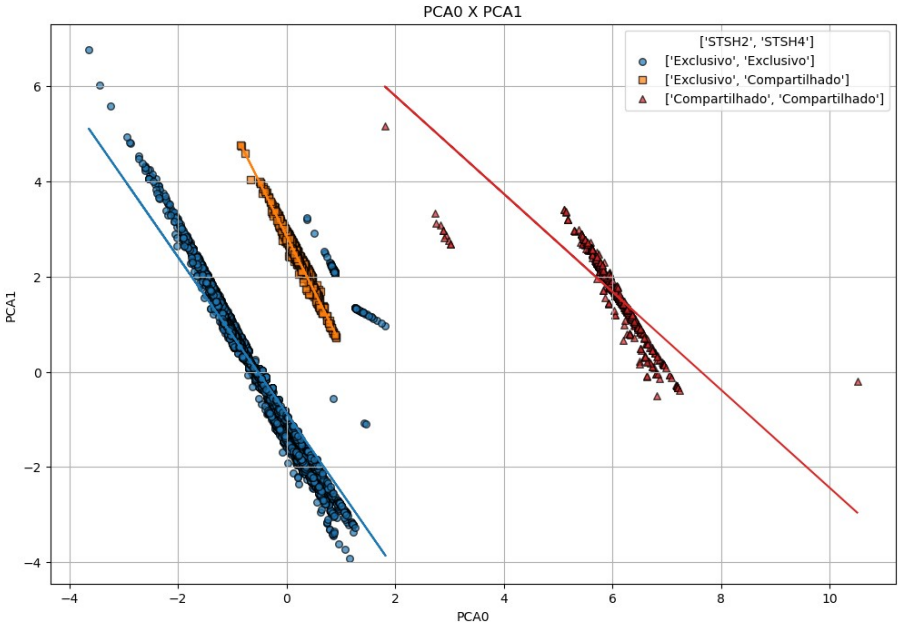
Pode-se observar na figura que há três grandes espaços (retas), e sendo uma delas populada inteiramente por cargas conteinerizadas, outra inteiramente por

Granel Sólido e por fim, a maior, mista com Granel Sólido, líquido e gasoso, assim como também carga conteinerizada.

Sabendo que o PCA0 é principalmente influenciado STSH2, STNaturezaCarga, STSH4, e sabendo que valores positivos desse PCA implicam em cargas com STSH2, STNaturezaCarga e STSH4 compartilhado – relação inversa - podemos concluir que, das cargas apresentadas, aquelas que são conteinerizadas encontram-se com PCA0 muito positivo, implicando serem de caráter compartilhado para todas as três variáveis, já para Granel sólido a distribuição é mais equilibrada entre exclusivo e compartilhado. O mesmo pode ser dito da seção conteinerizada da reta 1.

Afim de apurar a veracidade do exposto anteriormente, foi-se colorido as por STSH2 e STSH4 (a escolha deste acima do STNatureza se deve ao fato de que o STSH4 tem grande presença tanto no PCA0 quanto no PCA1). Segue abaixo a plotagem da Figura 3:

Figura 3: Carga por exclusividade em STSH2 e STSH4

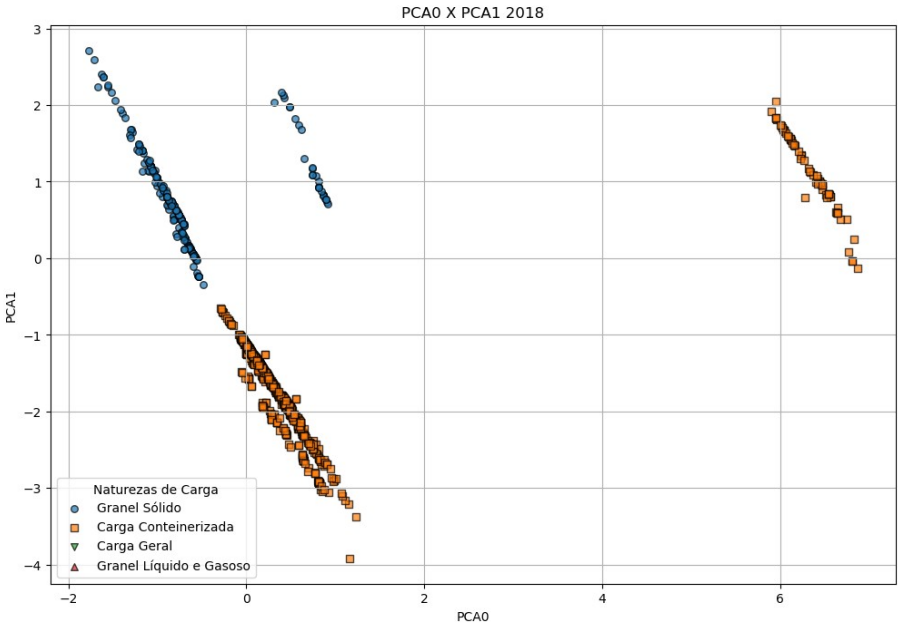


Fonte: Elaborada neste trabalho com os dados da ANTAQ (2018 – 2023).

Afim de melhor visualizar a tendência das retas, estas foram delineadas utilizando-se técnicas de regressão linear nos pontos de cada categoria – por meio da função “LinearRegression” do Scikit-learn. Lendo da esquerda para direita, a reta 1 é aquela que contém valores exclusivos para STSH2 e STSH4; a reta 2 é equilibrada e contém valor exclusivo apenas para STSH2; a reta 3 contém valores compartilhadas para ambas variáveis. Vale notar que a regressão linear da reta três é a que contém maior coeficiente angular, uma vez que esta tem valores mais dispersos devido a alguns *outliers*.

Por fim, verifica-se a análise feita a partir da matriz de covariância, no que diz respeito da relação da variável Ano e Natureza da Carga terem uma relação de proporcionalidade direta. Ao observar a plotagem do ano 2023 (maior valor para a coluna Ano) e comparar com 2018 (menor valor para coluna Ano) podemos observar as seguinte distribuição de Natureza das Cargas:

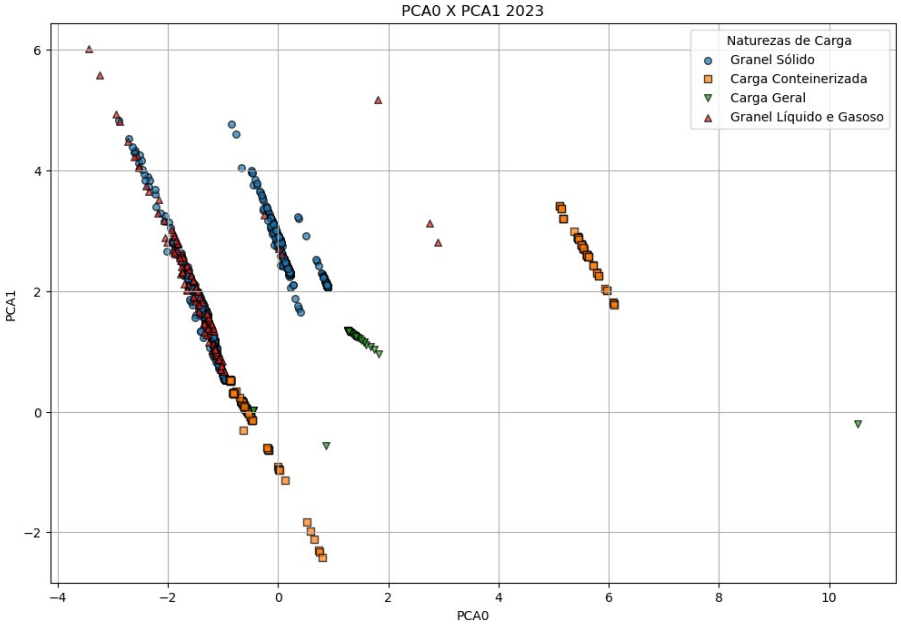
Figura 4: Carga colorida por natureza da carga (2018)



Fonte: Elaborada neste trabalho com os dados da ANTAQ (2018).

Na Figura 4 observa-se uma distribuição não muito dissimilar daquela na Figura 3. Já para o ano de 2023, observado na Figura 5:

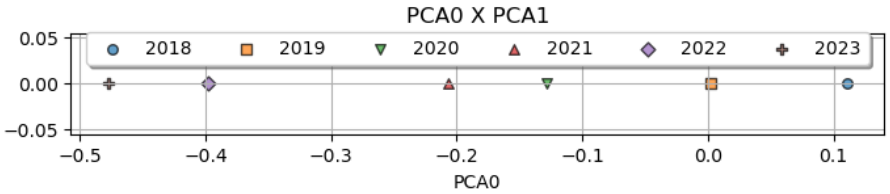
Figura 5: Carga colorida por natureza da carga (2023)



Fonte: Elaborada neste trabalho com os dados da ANTAQ (2023).

Verifica-se que há uma tendência a se ter menos cargas conteinerizadas e maior concentração de Granéis. Essa correlação é reforçada pelo fato de se haver mais cargas em 2023 (2164) do que no ano de 2018 (1228) e pela carga de menor variância - encontrado por menor distância da origem para todos os 21 PCAs - do ano de 2018 ter menor valor PCA0 do que a de 2023:

Figura 6: Carga de menor variância para cada ano



Fonte: Elaborada neste trabalho com os dados da ANTAQ (2023).

No que diz respeito ao PCA0, há uma clara tendência da carga de menor variância (mais representativa) para menor valores, o que implica em um aumento da exclusividade das colunas STs e menor variedade de mercadoria sendo movimentadas por atracação (atracações mais exclusivas)[14].

Conclui-se portanto que há uma tendência de cargas se tornarem mais exclusivas com o passar dos anos. No entanto, em termos absolutos essa tendência não é evidencia o suficiente para se tirar conclusões definitivas.

5 CONSIDERAÇÕES FINAIS

A análise PCA revelou possíveis tendências de exclusividade nas atracações ucranianas e russas, uma vez que ao reduzir a dimensionalidade do *dataset* de importações, os componentes principais mais expressivos demarcam um padrão de declínio de variedade de Natureza de Carga com o passar dos anos, além de um aumento da exclusividade das movimentações de pelo Porto de Santos.

Sabendo dos efeitos da guerra na Ucrânia, poderia induzir que houve uma diminuição do livre comércio entre os países em guerra e o Porto de Santos, dito isso, não houve o cessar absoluto do comércio marítimo por parte da Rússia. Isto pode ser verificado na matriz de covariância e depois verificado nos valores dos componentes principais.

No entanto, esta análise se demonstra passível de erros, em parte devido a amostra ser de poucas entradas relativa ao todo, dos período não ser muito abrangente, da escolha de colunas a serem removidas e por fim do fato de o *dataset* ser composto majoritariamente por cargas russas.

Em suma, a análise de componente principal se demonstra de grande auxílio para se melhor compreender as relações entre as variáveis em grandes datasets, agilizando a análise de padrões que poderiam se demonstrar custosos em serem identificados de formas mais tradicionais.

6 REFERÊNCIAS

IBM. **What is principal component analysis (PCA)?** 2023. Disponível em: https://www.ibm.com/topics/principal-component-analysis. Acesso em: 19 nov. 2024.

BRASIL. Gov.Br. Receita Federal. **NCM**. 2019. Disponível em: https://www.gov.br/receitafederal/pt-br/assuntos/aduana-e-comercio-exterior/classificacao-fiscal-de-mercadorias/ncm. Acesso em: 24 nov. 2024.

BRASIL. Antaq. Agência Nacional de Transportes Aquaviários. **Informações Metodológicas**. 2014. Disponível em: https://web3.antaq.gov.br/ea/sense/doc.html#pt. Acesso em: 24 nov. 2024.

14 Valores de menor variância foram encontrados por menor distância da origem em todas as dimensões (não só PCA0 e PCA1). Os detalhes do cálculo fogem do escopo deste trabalho.

GEEKSFORGEEKS. **Principal Component Analysis(PCA)**. 2024. Disponível em: https://www.geeksforgeeks.org/principal-component-analysis-pca/. Acesso em: 23 nov. 2024.

JAADI, Zaakaria. **Principal Component Analysis (PCA)**: a step-by-step explanation. Builtin. 2024. Disponível em: https://builtin.com/data-science/step-step-explanation-principal-component-analysis. Acesso em: 24 nov. 2024.

AZEVEDO, Caio. **Anáalise de componentes principais**: parte 1. Campinas: Unicamp, 2021. 50 slides, color. Disponível em: https://www.ime.unicamp.br/~%20cnaber/aula_ACP_Ana_Multi_P1_2S_2021.pdf. Acesso em: 24 nov. 2024.

IBM. **What is ETL (extract, transform, load)?** [2023?]. Disponível em: https://www.ibm.com/topics/etl. Acesso em: 24 nov. 2024.

OLIVEIRA, Alexandre Garcia de. Chapter 4: autovalores e autovetores. In: OLIVEIRA, Alexandre Garcia de. **Algebra Linear**: ciência de dados. Santos: [s.*n.*], 2022. Cap. 4. p. 73-77.

GEEKSFORGEEKS. **Linear Regression (Python Implementation)**. 2023. Disponível em: https://www.geeksforgeeks.org/linear-regression-python-implementation/. Acesso em: 24 nov. 2023.

TOTVS. **14 tipos de cargas para planejar a logística e gestão de frota do seu negócio**. 2023. Disponível em: https://www.totvs.com/blog/gestao-logistica/tipos-de-cargas/. Acesso em: 23 nov. 2024.

SCIKIT-LEARN (comp.). **PCA**. 2024. Disponível em: https://scikit-learn.org/dev/modules/generated/sklearn.decomposition.PCA.html. Acesso em: 19 nov. 2024.

PYTHON. *What is Python*. **PYHTON.ORG,** [*s.l.*], 19 set. 2024. Disponível em: https://docs.python.org/3/faq/general.html#what-is-python. Acesso em: 20 set. 2024.

PANDAS. *About Pandas: History of development*. **PANDAS,** [*s.l.*], 2024. Disponível em: https://pandas.pydata.org/about/. Acesso em: 20 set. 2024.

JUPYTER. **About Us**: project jupyter?s origins and governance. Project Jupyter’s origins and governance. 2024. Disponível em: https://jupyter.org/about. Acesso em: 24 nov. 2024.

Link do Github: https://github.com/igorzeck/PCA

Link da base de dados da ANTAQ (via PDA): https://dados.gov.br/dados/conjuntos-dados/estatistico-aquaviario-ea