**Modelo não supervisionada para classificação de blocos econômicos usando Pandemia do Coronavirus (COVID-19) entre 2020 e 2021.**

Rodrigo Da Costa¹\*;Renato Máximo Sátiro2

1 Georgia Institute of Technology. Mestrando em Data Analytics. Atlanta, Geórgia, Estados Unidos

2 Universidade Federal de Goiás (UFG). Doutorando em Administração. Goiânia, Goiás, Brasil

\*autor correspondente: costa@gatech.edu

**Modelo não supervisionada para classificação de blocos econômicos usando Pandemia do Coronavirus (COVID-19) entre 2020 e 2021.**

**Resumo**

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

**Palavras-chave:** k-means; processos de flutuação empírica; PCA (Análise do componente principal).

**Unsupervised model for classifying economic blocks using**

**Coronavirus Pandemic (COVID-19) between 2020 and 2021.**

**Abstract**

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

**Keywords:** k-means; empirical fluctuation processes; PCA (Principal component analysis)

**Introdução**

A classificação de blocos econômicos é um processo importante para entendermos a dinâmica macroeconômica dos países. A pandemia do coronavirus trouxe uma oportunidade única de compararmos os países com uma única lente (mesmo padrão) onde 63 métricas foram contabilizadas de Janeiro de 2020 à Novembro de 2021. Entre as métricas, temos informações que vão além da simples caracterização da doença e apresentam caraterísticas Políticas, Económicas, Sociais e Tecnológicas como por exemplo “life\_expectancy”, “handwashing\_facilities”, “human\_development\_index”, “gdp\_per\_capita”, “aged\_65\_older”, “stringency\_index” entre outros.

Uma das formas mais comuns de avaliação e classificação macroenômica é a utilização da técnica PEST ou PESTEL. Esta técnica porém faz parte do grupo de análises administrativas para mapear pontos fortes e fracos e como resultado pode ser bastante subjetiva.

Com base nos dados da pandemia COVID-19, podemos utilizar uma técnica de aprendizagem não supervisionada (k-means) para classificar os países em blocos econômicos de forma independente.

**Material e Métodos**

Os dados da pandemia COVID-19 serão obtidos de [1]. Com base nos dados e após um processo de limpeza, vamos utilizar a função efp do pacote R para identificar os pontos de separação para dividir os dados em blocos (antes e depois do ponto de quebra). Dentro destes blocos, podemos selecionar uma data aleatória e rodar a aprendizagem não supervisionada em k-means utilizando o pacote sklearn.cluster/KMeans em python.

No final, rodaremos uma análise PCA (Análise do componente principal) para apresentar os resultados da classificação em duas dimensões.

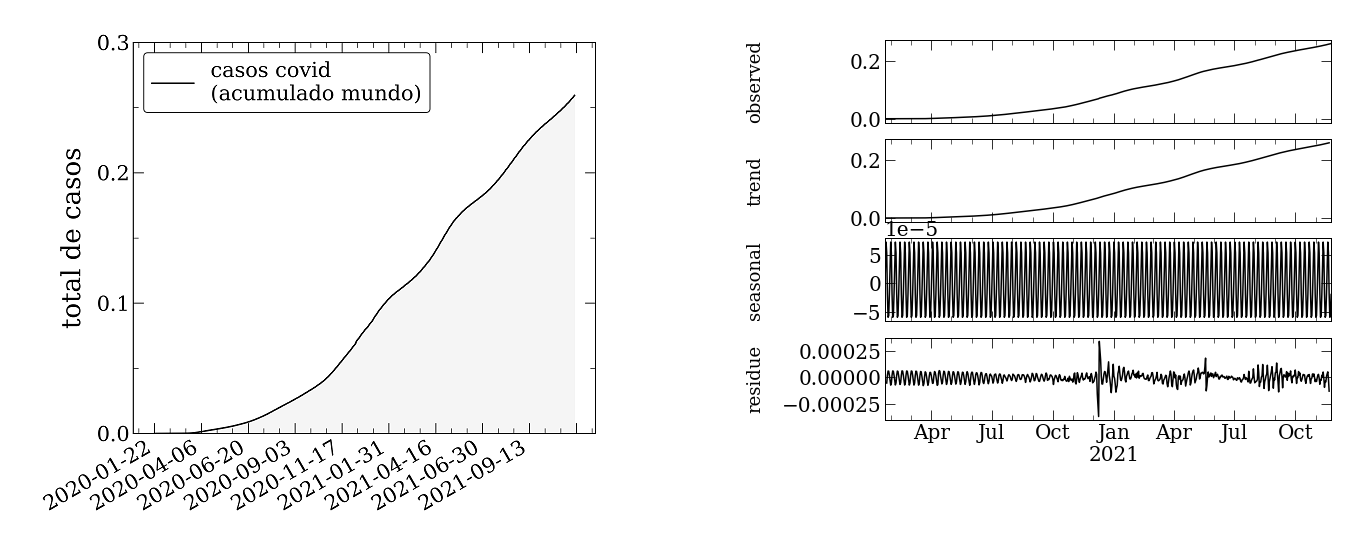
Coleta e Limpeza dos Dados

Os dados que utilizamos são curados pela equipe do ‘Global Change Data Lab’ em colaboração com a universidade de Oxford. Todos os dados produzidos pela equipe são curados e mantidos pela instituição e são disponibilizados sob a licença CC ‘creative commons’ onde qualquer pessoa pode utilizar os dados bastante atribuir o crédito para a instituição. O corpus atual dos dados contém 67 colunas onde 63 destas foram utilizadas para a nossa análise. Além disto, os dados apresentam 222 países com dados desde Janeiro de 2020 e com atualizações diárias e no nosso caso, utilizaremos os dados do 24 de Novembro de 2021 como base.

Para um bom resultado na execução da nossa análise, decidimos remover todas as colunas de informação categórica (texto) mantendo apenas um identificador do país. Para os dados que apresentam valores nulos (NaN), a decisão foi transformar estes valores em zero ao invés de remover a linha. Todos os valores também foram normalizados utilizando a biblioteca de ‘sklearn’ do python (MinMaxScaler).

Processos de Flutuação Empírica

O primeiro passo da análise foi selecionar datas chaves para a análise. Para isto, utilizamos uma análise de resíduo com decomposição de tendência, sazonalidade e resíduo.

**Figura 1.** Casos de COVID (número total) acumulado para todos os países   
e análise de decomposição

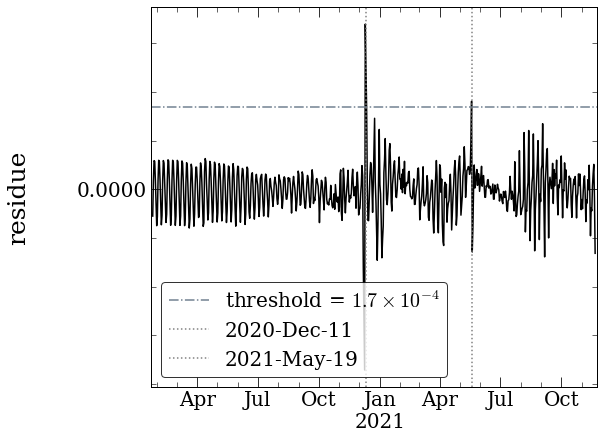
Com a avaliação do resíduo, traçamos uma linha de limite para identificar quais datas apresentaram maio variação. Utilizando o limite de , conseguimos identificar três datas:

|  |  |
| --- | --- |
| **Data** | **Resíduo** |
| 2020-12-11 |  |
| 2021-05-19 |  |
| 2021-11-24 | *data controle* |

**Tabela 1.** Identificação da data por resíduo (flutuação) e limite.

Estas datas foram também validadas visualmente através do gráfico de resíduos onde vemos claramente que para as datas marcados temos fortes evidências de mudanças de comportamento nos dados:

**Figura 2.** Análise de resíduo mostrando os limites de datas



Note que como data de controle a decisão foi utilizar o dia mais recente (24 de Novembro de 2021 para o este estudo).

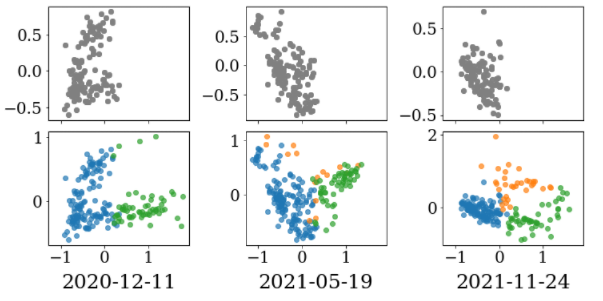
Análise de Cluster K-means

Após a coleta de dados, limpeza e definição das três datas que vão ser processadas, como modelo de classificação utilizaremos o modelo de agrupamento de k-means com valor do cluster sendo (três grupos). O modelo faz parte do grupo de modelos com aprendizado de máquina não supervisionado e foi escolhido uma vez que não queremos criar nenhum bias humano na classificação.

PCA (Análise do componente principal)

Por fim, também precisamos utilizar a análise de componente principal para reduzir os 63 componentes iniciais para apenas 2.

**Figura 3.** Análise de PCA para confirmar que os grupos de cores   
da classificação estão corretos e equivalentes.



Como mostra a figura 3 – a análise é importante para mostrar que todas as cores (agrupamentos do modelo de k-means) estão equivalentes.

**Resultados e Discussão**

O título da seção Resultados e Discussão deve ser alinhado à esquerda, grafado em negrito com as primeiras letras das palavras em letras maiúsculas. É permitido que a seção seja dividida em subtópicos com formatação de acordo com a descrição no item 1.1 Formato e margens, apresentados na mesma ordem da seção Material e Métodos. Nesta seção devem ser apresentados, discutidos e interpretados os resultados obtidos no trabalho, ou seja, autores devem fazer uma discussão comparativa dos resultados do seu trabalho com aqueles existentes na literatura e elaborar uma análise crítica dos dados, destacando as limitações e pontos positivos dos resultados.

**Conclusão**

O título da seção Conclusão(ões) ou Considerações Finais deve ser alinhado à esquerda e grafado em negrito. Fica a critério do aluno e do orientador a escolha de qual termo melhor se adequa ao trabalho. Esta seção deve conter frases curtas, apresentando as conclusões e inferências elaboradas a partir da discussão dos resultados. É importante que estas frases não sejam meras reproduções dos resultados, respondendo aos objetivos propostos no trabalho. Os autores não devem, em hipótese alguma, mencionar, citar ou reproduzir resultados de outros estudos na(s) conclusão(ões) ou considerações finais do TCC. Por fim, salienta-se que essa seção não deve conter tabelas ou figuras, sendo redigida de forma sucinta.

**Referências**

1. Hannah Ritchie, *et al*  - **"Coronavirus Pandemic (COVID-19)"**. Disponível em: 'https://ourworldindata.org/ (2020). Acesso em: 24 nov. 2021
2. Brown R.L., Durbin J., Evans J.M. (1975), **Techniques for testing constancy of regression relationships over time**, Journal of the Royal Statistal Society, B, 37, 149-163.
3. Kuan C.-M., Chen (1994**), Implementing the fluctuation and moving estimates tests in dynamic econometric models**, Economics Letters, 44, 235-239.
4. Arthur, D., Vassilvitskii, S.: **k-means++: The advantages of careful seeding**. Tech. rep., Stanford (2006)project
5. Zubair M., *et al* **An Efficient K-Means Clustering Algorithm for Analysing COVID-19.** Hybrid Intelligent Systems. vol 1375. Springer, Cham. (2021)