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

Rodrigo Da Costa[1]*; Renato Máximo Sátiro[2]

[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

Utilizando os dados da pandemia do Coronavirus entre 2020 e 2021 gerenciados pelo 'Global Change Data Lab' em colaboração com a universidade de Oxford. Utilizando análise de decomposição do número total de casos acumulado para todos os países e a técnica de flutuação empírica dos resíduos com limite de 1.7×10^{-4}, identificamos três datas chaves (Dezembro de 2020, Maio de 2021 e Novembro de 2021). Com as três datas definidas, aplicamos a análise de Cluster K-means sendo $k = 3$ (três grupos: verde, azul e laranja). Para cada um dos grupos, fizemos a análise de componentes principais (PCA) e apresentamos cada um dos dados em um mapa coroplético. Por fim, concluímos que verde são os 'países desenvolvidos', azul são os 'países não desenvolvidos' e laranja possui duas interpretações: (a) em Maio de 2021 são as 'super potências mundiais' e (b) laranja em Novembro de 2021 são as Mercado Emergentes.

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

Using data from the Coronavirus pandemic between 2020 and 2021 managed by the 'Global Change Data Lab' in collaboration with the University of Oxford. Using decomposition analysis of the accumulated total number of cases for all countries and the empirical fluctuation for residue with a 1.7×10^{-4} threshold, we identified three key dates (December 2020, May 2021 and November of 2021). With the three dates defined, we applied the Cluster K-means analysis being $k = 3$ (three groups: green, blue and orange). For each of the groups, we performed principal component analysis (PCA) and presented each of the data in a choropleth map. Finally, we conclude that green are developed countries, blue are 'undeveloped countries' and orange has two interpretations (a) in May 2021 are the 'world super powers' and (b) orange in November 2021 are the Emerging Markets.

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

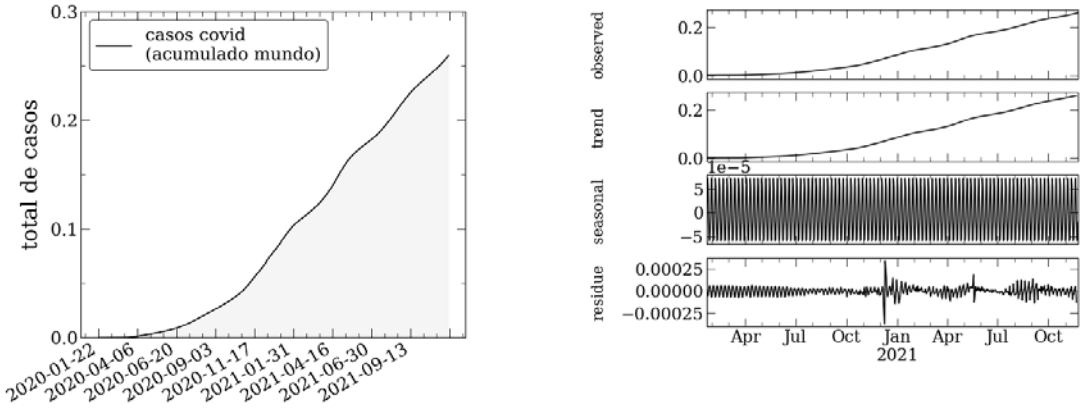
2020 e com atualizações diárias e no nosso caso, utilizaremos os dados do 24 de Novembro de 2021 como base.

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*



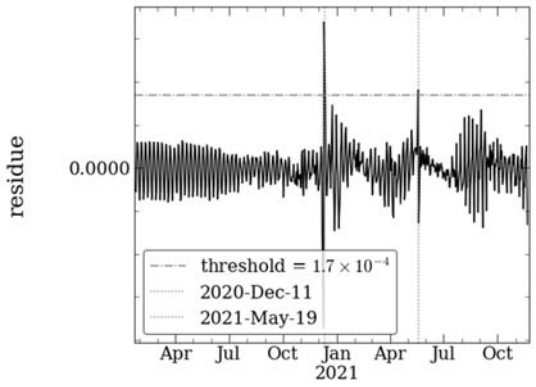
Com a avaliação do resíduo, traçamos uma linha de limite para identificar quais datas apresentaram maio variação. Utilizando o limite de 1.7×10^{-4}, conseguimos identificar três datas:

Data	Resíduo
2020-12-11	2.53×10^{-4}
2021-05-19	1.81×10^{-4}
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).

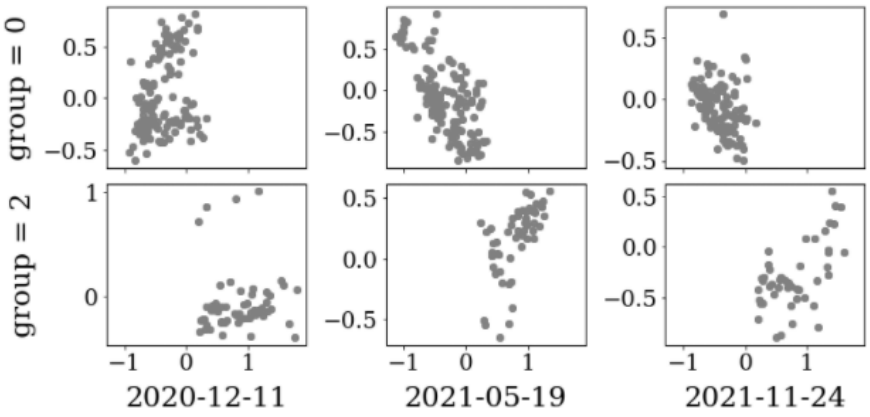
Análise de Cluster K-means

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 $k = 3$ (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.*



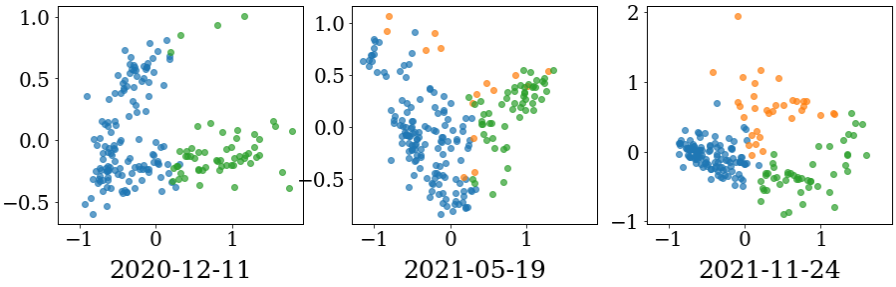
Como mostra a figura 3 – a análise é importante para mostrar que todas as cores (agrupamentos do modelo de k-means) estão equivalentes.

Resultados e Discussão

Para visualizarmos o modelo de predição, apresentamos os resultados da análise do componente principal (PCA) que utiliza passo de decomposição de valor singular (SVD). Com isto, conseguimos reduzir os dados de 63 componentes para apenas duas componentes (duas dimensões do gráfico, eixo x e y) onde apresentamos na figura 4. A figura 4 apresenta a classificação dos dados em três grupos (verde, azul e laranja) onde cada ponto no gráfico de resultados é um país e sua cor é definida com base no modelo de k-means.

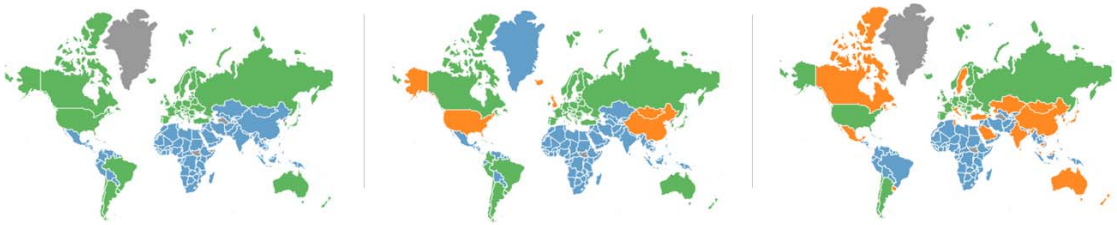
Note que para Dezembro de 2020 (primeira data), notamos que o modelo classificou os dados apenas em dois grupos, isto foi uma decisão do próprio modelo e os mesmos parâmetros foram usados para todas as datas ($k = 3$, três grupos).

Figura 4. *Resultado do modelo preditivo treinado para $k = 3$ (três agrupamentos).*



Com base nos resultados do modelo de classificação k-means, utilizamos a biblioteca gráfica D3 para gerar o mapa coroplético onde tomamos o cuidado de manter as mesmas cores apresentadas na análise de PCA do resultado do modole preditivo de classifição.

Figura 5. *Evolução dos grupos macroeconômicos da esquerda para a direita partindo de Dezembro de 2020, Maio de 2021 e finalmente Novembro de 2021.*

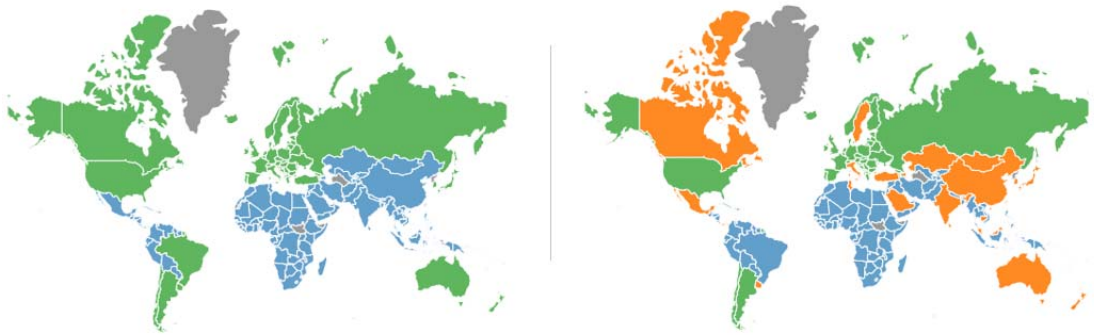


As três datas que escolhemos baseados na nossa análise de flutuação empírica dos resíduos são importantes para entender a evolução da pandemia. A primeira data (Dezembro de 2020) foi quando ainda não tínhamos nenhuma vacina e a pandemia atingia ponto crítico

– este data foi quando tivemos os Estados Unidos autorizando de forma emergencial o início da vacinação.

A segunda data (Maio de 2021) já apresenta um momento diference onde vários países já tinham administrado a segunda dose da vacina. Por fim, a terceira data (Novembro de 2021) é a data de controle e traz uma foto atual da classificação dos países.

Figura 6. *Comparativo lado a lado com o começo da análise (Dezembro de 2020) com o final da análise (Novembro de 2021)*



A análise feita aqui só considera três datas (três pontos na linha do tempo) e embora suficiente como comparação não representa o período crítico. Uma melhor análise aqui poderia considerar ao invés de uma única data um período de tempo e aplicar a técnica do voto majoritário das classificações do período. Infelizmente, uma análise deste tipo exige um grande esforço já que uma das características do modelo de k-means é a alocações de grupos aleatoriamente. Como a classificação é aleatória, processando a mesma data, teríamos os mesmos eixos do PCA mas as cores (valores de k) podem ser diferentes em cada regressão.

Para o caso aqui apresentado, tomamos o cuidado de manter os mesmos valores de k para as três datas de forma manual como apresentado na figura 3. Mas se expandirmos esta análise para mais datas ou para um período, este processo de garantir os valores de k (cores) precisa também ser automatizado.

Conclusão

Conseguimos classificar os países em clusters utilizando o modelo de k-means com $k = 3$ (três grupos) para três datas diference: Dezembro de 2020, Maio de 2021 e Novembro de 2021. Para Dezembro de 2020 notamos que o mundo era dividido em dois blocos econômicos onde o primeiro (em verde) inclui: América do Norte, Europa, alguns países da América do Sul e Oceania. Na literatura encontramos que a melhor aproximação para esta classificação da primeira data seria a de Países desenvolvidos (verde) e Países em desenvolvimento (azul).

Na segunda data temos a introdução agora de uma terceira cor (laranja). Aqui ressaltamos que três países aparecem como principais neste bloco: Estados Unidos, China e Inglaterra. A melhor interpretação para este bloco é o de 'super potências mundiais', estes foram os países que lideraram os avanços tecnológicos no mundo para o tratamento do COVID e como tal são bem representados neste grupo.

Para a terceira data (Novembro de 2021), o cor laranja não mais significa 'super potências mundiais' e uma melhor interpretação aqui seria a de mercados emergentes. Notamos que na terceira data o Brasil e alguns países da América da Sul recebem um rebaixamento (da cor verde para a azul) mostrando que ficam mais próximos do bloco de países considerados não desenvolvidos (sub desenvolvidos).

Por fim, concluímos que a utilização de inteligência artificial para a classificação de blocos econômicos é hoje uma realidade e muitas outras fontes de dados podem ser utilizadas aplicando as mesmas técnicas aqui descritas. Isto traz uma melhoria significativa ao processo atual com mais escala onde a classificação pode ser atualizada até mesmo em tempo real algo que seria inimaginável no passado.

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)

[6] Bostock, Mike. D3.js - Data-Driven Documents. Disponível em: 'http://d3js.org/'.

[7] O'Sullivan, Arthur. Economics : Principles in Action / Arthur O'Sullivan, Steven M. Sheffrin. Needham, Mass. :Prentice Hall, 2003.

[8] Costa, Rodrigo. MBA-USP-TCC. Disponível em: 'https://github.com/costargc/MBA-USP-TCC/'.