Uma nova regionalização do mundo, com base em critérios socio-econômicos

Gabriel Penha[1] Moisés Augusto[2]

[1]Universidade Federal da Bahia
{moises.augusto, penha.gabriel}@ufba.br

Resumo

Regionalização é um processo de divisão dos países em diferentes regiões (ou grupos). Historicamente, diversas regionalizações foram propostas e muitas delas são consideradas ainda presentemente. Neste trabalho, propomos uma nova regionalização, com base em critérios socio-econômicos. As seguintes características foram utilizadas na análise: população, taxa de crescimento populacional, taxa de urbanização, renda per capita, expectativa de vida, taxas de natalidade, de mortalidade e de mortalidade infantil além do índice de educação. O algoritmo K-means foi o procedimento utilizado para isto, através das variáveis criadas a partir de uma análise de componentes principais. Com esta análise de componentes, propusemos ainda um ranqueamento dos países com base em fatores socio-econômicos e avaliamos a clusterização obtida. No geral, os resultados encontrados não diferiram dos que eram esperados; as exceções foram ressaltadas durante o texto. Avaliamos que o trabalho poderia ser estendido com acesso a mais variáveis e com algumas modificações na metodologia proposta; ainda assim, os resultados foram considerados satisfatórios.

Palavras-chave: regionalização; análise de componentes principais; segmentação; ranqueamento; K-means.

1 Introdução

Regionalização é um processo de divisão dos países em diferentes regiões (ou grupos). Historicamente, diversas regionalizações foram propostas e muitas delas são consideradas ainda presentemente. De acordo com Sene and Moreira (1999), podemos citar os continentes, que dividem os países com base em critérios geográficos. Alternativamente, menciona-se a divisão histórica entre países do velho mundo, constituído por nações europeias, asiáticas e africanas, países do novo mundo, constituído pelas Américas e os do novíssimo mundo, com nações da Oceania. Podemos destacar ainda as divisões socio-econômicas que separavam os países entre primeiro (capitalistas desenvolvidos), segundo (socialistas ou de economia planificada) e terceiro (capitalistas subdesenvolvidos) mundos. Conforme o passar do tempo, alguns desses agrupamentos caíram em desuso, enquanto outros se mantiveram firmes. Em particular, o agrupamento socio-econômico mencionado não é mais utilizado. De fato, existem grupos - ou blocos - socio-econômicos atualmente; podemos citar o Mercosul, na América do Sul, a União Europeia, na Europa e o BRICS, composto por Brasil, Rússia, Índia, China e África do Sul, países economicamente emergentes. O leitor atento, no entanto, pode ter percebido pelos exemplos, que na formação deste tipo de blocos, com uma possível exceção ao BRICS,

não somente variáveis socio-econômicas são consideradas. No caso do Mercosul e da União Europeia, por exemplo, existe o fator geográfico.

Apesar de interessantes, esses tipos de divisão ocorrem apenas para um grupo de países específico. Podemos identificar as características similares aos BRICS, mas não temos outros grupos avaliados sob os mesmos critérios (neste caso, características econômicas, populacionais e, em certo nível, política) - isto é, o restante do mundo não está dividido em grupos caracterizados pelos mesmos fatores - o que poderia ser interessante para a formação de outros blocos cooperativos objetivando desenvolvimento mútuo.

Dito isso, neste trabalho, visamos propor uma nova regionalização, com base em critérios socio-econômicos. As seguintes características foram utilizadas na análise: população, taxa de crescimento populacional, taxa de urbanização, renda per capita, expectativa de vida, taxas de natalidade, de mortalidade e de mortalidade infantil além do índice de educação, medido através da taxa de alfabetização dos adultos e da taxa de escolarização combinada do primário, secundário e terciário bruto (com uma ponderação do terceiro). Uma definição mais detalhada de todas essas variáveis pode ser encontrada nos textos de Sagar and Najam (1998) e Sene and Moreira (1999).

Somado a isso, iremos propor um ranqueamento dos países com bases nas variáveis consideradas - isto é, do melhor, para o pior -. Obviamente, não listaremos todos os países aqui, mas o resultado será comentando e o endereço para verificá-lo na íntegra será disponibilizado mais a frente.

Este trabalho está dividido em 4 seções. A seção seguinte apresentará a metodologia. Nela descrevemos o processo de obtenção dos dados e os métodos utilizados para a realização da análise, para o ranqueamento e para a segmentação dos países. Na seção de resultados, apresentaremos o que foi obtido através das análises propostas. Finalmente, discutiremos os resultados obtidos e apresentaremos motivações para trabalhos futuros.

2 Metodologia

Antes de tudo, vale mencionar que as variáveis utilizadas para a realização da análise foram todas obtidas por fontes oficiais, em geral, por censos demográficos (em alguns casos, por estimativas oficiais disponibilizadas pela Organização das Nações Unidas - ONU -.). O critério de comparabilidade pode ter sido um pouco prejudicado neste ponto, pois os dados não necessariamente eram do mesmo período. Isto é, as informações não estavam centralizadas em um único conjunto de dados; elas foram agregadas de diferentes fontes (censos de diferentes países) e os dados mais atualizados para cada variável/país foram utilizados. Dessa forma, a informação de população, por exemplo, foi estimada entre 2020 e 2022 a depender do país, enquanto a de taxa de mortalidade infantil foi estimada entre 2015 e 2020.

Países com ao menos uma das informações muito desatualizadas ou não disponíveis não foram considerados na análise. Deste modo, 143 países passaram por todo o processo de ranqueamento e segmentação e 123 deles são visualizáveis em seus devidos segmentos no mapa da Figura 3.

Considerando a estrutura multivariada dos dados estudados, dividiu-se a análise em três partes: exploratória, análise de componentes principais e a segmentação (ou análise de *cluster*).

2.1 Análise exploratória

Na primeira parte, as variáveis foram avaliadas de modo descritivo (e em conjunto). Estudaram-se as estatísticas descritivas, as matrizes de covariância e correlação, além da normalidade

multivariada das informações. Para avaliação da normalidade multivariada, utilizou-se o teste de Henze-Zirkles, que se baseia numa distância funcional não negativa, medindo a distância entre a distribuição teórica e a empírica.

2.2 Análise de componentes principais

De acordo com Johnson and Wichern (2007), algebricamente a análise de componentes principais é uma combinação linear das p variáveis aleatórias $X_1, X_2, ...X_p$. É um procedimento que requer somente a matriz de covariâncias $\boldsymbol{\Sigma}$ (ou de correlações $\boldsymbol{\rho}$) e que visa reduzir a dimensionalidade dos dados (ou seja, explicar as p variáveis com $k < p$ combinações lineares delas) conservando o máximo possível de variabilidade. A i-ésima componente principal, Y_i é definida da seguinte forma: $Y_i = \boldsymbol{a_i'X}$, em que $\boldsymbol{a_i'}$ é o autovetor normalizado associado ao i-ésimo autovalor da matriz de covariâncias de $\boldsymbol{X}$. A normalização é importante para garantir que $\boldsymbol{a_i'a_i} = 1$, o que, como consequência, limita a variância de Y_i (e permite que $\sum_i^p Var(X_i) = \sum_i^p Var(Y_i)$), além de fazer com que $Cov(Y_i, Y_j) = 0$, com $i \neq j$.

Para a escolha do número de componentes principais, geralmente leva-se em conta a proporção da variabilidade explicada por cada componente. O procedimento é feito de modo que a primeira componente explique a maior variabilidade dentre as componentes; em seguida, a segunda componente será a que mais explica e assim em diante. Deste modo, como mencionado por Johnson and Wichern (2007), se as primeiras componentes explicam mais que 80% ou 90% da variabilidade, a depender do fenômeno estudado, elas podem ser utilizadas no lugar das variáveis originais.

Vale salientar que, neste trabalho, a análise de componentes principais foi realizada a partir da matriz de correlações das co-variáveis $\boldsymbol{X}$; isto é, a matriz de covariâncias do vetor aleatório $\boldsymbol{Z}$, com i-ésimo elemento definido da seguinte maneira:

$$Z_i = \frac{X_i - \mu}{\sigma}. \tag{1}$$

De acordo com Mardia *et al.* (1979), utilizar as variáveis $\boldsymbol{Z}$ ao invés de $\boldsymbol{X}$ é indicado quando as escalas das variáveis $\boldsymbol{X}$ são muito diferentes entre si. Neste trabalho, algumas variáveis estão no intervalo $[0-1]$, enquanto a população, por exemplo, tem dois países com mais de 1 bilhão de habitantes.

Finalmente, é importante mencionar o processo de ranqueamento dos países. O ranqueamento foi feito com base na primeira componente principal, que, como ficará claro na seção de Resultados, pode ser interpretada como uma componente socio-econômica. Os países com maior pontuação nessa componente eram melhor classificados no ranking proposto.

2.3 Análise de *cluster*

Para a segmentação dos grupos, utilizou-se o *K-means*. O *K-means* é um método de clusterização não hierárquico. Este tipo de método é desenhado para agrupar itens, ao invés de variáveis. O número de grupos, K, costuma ser otimizado por procedimentos como o método de Elbow, ou pela *Silhouette*. Johnson and Wichern (2007) e Mardia *et al.* (1979) descrevem mais detalhes sobre as abordagens em torno do método, mas a grosso modo, o objetivo deste tipo de análise é criar grupos de maneira que os indivíduos (neste caso, países) dentro de um mesmo cluster sejam os mais parecidos possíveis entre si e, simultaneamente, os mais diferentes possíveis comparado aos países dos outros clusters. Em particular, o *K-means* tenta minimizar a soma de quadrados das distâncias entre os objetos e seus centroides. Diversas distâncias podem ser utilizadas; no trabalho, a distância euclidiana foi a considerada.

Aqui vale ressaltar que a segmentação foi feita com as variáveis resultantes da análise de componentes principais descrita na subseção anterior. Poderíamos ter utilizado as variáveis originais, porém, além da análise de componentes principais reduzir a dimensionalidade (o que é bom para o desempenho deste tipo de método), de algum modo, as componentes são comparáveis (isto é, não possuem escalas muito diferentes), além de conterem a maioria da variabilidade presente nas variáveis. De outro lado, as componentes principais permitem uma interpretabilidade conjunta das variáveis; o que facilita a interpretação dos grupos segmentados na análise de cluster, isto é: é mais fácil interpretar, para cada um dos grupos criados, três componentes principais que dez variáveis.

3 Resultados

Primeiramente, é válido mencionar que todos os resultados e todo o processo de análise pode ser checado pelo leitor neste endereço do GitHub: https://github.com/mkyou/segmentacao-paises. Na pasta "*data*", além do conjunto de dados final utilizado (após o tratamento das variáveis), o leitor encontrará o arquivo com o nome "ranking_paises_pca1.csv", que apresenta os países ordenados pela primeira componente principal (os países no topo da lista foram considerados os melhores conforme as variáveis socio-econômicas utilizadas); por curiosidade, o Brasil é o quadragésimo nono na lista, o Peru é o sexagésimo terceiro, o melhor país da América Latina é o Chile, em trigésimo quinto e o melhor país, segundo estes critérios, é Singapura. Na pasta "*plots*", o leitor irá se deparar com a análise gráfica realizada que, em geral, serviu para verificar a dimensão e ter uma ideia da distribuição das co-variáveis. Além disso, ainda na mesma pasta, será possível visualizar os critérios gráficos utilizados para a escolha do número de componentes principais e do número de clusters (estes critérios serviram de guia, critérios numéricos como a proporção da variabilidade explicada pelas componentes, por exemplo, foram levados mais em conta que os primeiros).

Dito isso, os resultados mais importantes serão apresentados aqui; o acesso ao endereço especificado é opcional e não afetará o entendimento do leitor. Entre os resultados exploratórios, destacamos as diferentes escalas de mensuração das variáveis, já mencionadas na metodologia e a falta de normalidade multivariada nas informações estudadas. Tome, por exemplo, a Figura 1. Nela, podemos visualizar os histogramas do índice de educação, da expectativa de vida, do PIB (produto interno bruto) per capita e da população dos países. É possível observar uma assimetria relevante em 3 dessas variáveis - forte indício de falta de normalidade uni-variada e, consequentemente, falta de normalidade multivariada -. Os resultados são similares para as demais informações. A hipótese de normalidade multivariada também foi contra-indicada pelo teste de Henze-Zirkles; tanto para as co-variáveis originais, quanto para as componentes principais calculadas a partir destas. Deste modo, a realização de inferência e avaliação de testes de hipótese utilizando, por exemplo, testes como o T^2 de Hotelling, foram inviabilizadas.

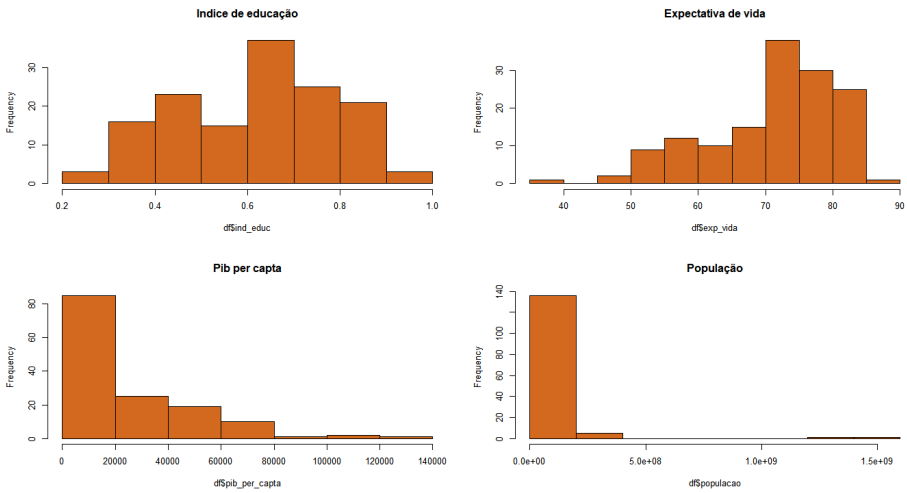

Figura 1: Histogramas empíricos dos índices de educação, da expectativa de vida, do PIB per capita e da população. Note a diferença nas escalas das co-variáveis e a assimetria nelas presentes

Fonte: Censos e estimativas de órgãos oficiais e da ONU.

Em relação à análise de componentes principais conduzida através da matriz de correlação das co-variáveis originais, salientamos que as três primeiras componentes explicavam mais de 80% das co-variáveis e foram, portanto, utilizadas para o restante do estudo. Verificaram-se os autovetores e as correlações das componentes principais para com as variáveis originais estudadas. A correlação das componentes com as variáveis são apresentadas na Tabela 1. Por ela, podemos observar que a primeira componente está muito correlacionada com variáveis como o índice de educação, a expectativa de vida, a taxa de urbanização (apesar de negativamente, neste caso - o que pode significar que países com maior expectativa de vida e índice de educação não se urbanizam mais por já estarem urbanizados -), entre outras - todas variáveis que, conforme a matriz de correlações estimada, são bastante correlacionadas entre si -. A segunda componente principal está bastante relacionada à população dos países, enquanto a terceira relaciona-se positivamente com a taxa de mortalidade e negativamente com o crescimento populacional. Dito de outra maneira, a primeira componente pode ser interpretada como uma componente socio-econômica; a segunda, como uma populacional, enquanto terceira é mais difícil de interpretar intuitivamente.

Variáveis/Componentes	CP1	CP2	CP3
Índice de educação	0.925547458	-0.12363960	0.09470436
Expectativa de vida	0.934241909	0.11424119	-0.21910375
PIB per capita	0.739285212	-0.16368518	-0.12968811
População	-0.008231919	0.87310911	0.34059982
Crescimento populacional	-0.769828418	0.06821896	-0.52501589
Percentual urbano	0.653517822	-0.15626906	-0.24555227
Taxa de urbanização	-0.879291297	0.05905690	-0.26018612
Taxa de mortalidade infantil	-0.926231152	-0.09807401	0.12775648
Taxa de mortalidade	-0.543597953	-0.42732967	0.61018615
Taxa de natalidade	-0.940992014	-0.02055044	-0.20539179

Tabela 1: Correlação das variáveis originais com as componentes principais.

Com isso posto, considerando que a primeira componente explica a variabilidade socio-econômica, o resultado do ranqueamento dos países mencionados no início da seção faz sentido intuitivo. Países que sabemos que são mais desenvolvidos figuraram nas posições mais altas do ranqueamento. Por outro lado, é importante ressaltar que as variáveis utilizadas não foram coletadas todas no mesmo período - lembre do exemplo utilizado, população obtida entre $2020-2022$ e taxa de mortalidade entre $2015-2020$; deste modo, caso considerássemos os índices/taxas atuais, é possível que obtivéssemos resultados um pouco diferentes.

A relação entre as componentes principais consideradas pode ser visualizada através dos *biplots* da Figura 2. O interessante desta figura é, justamente, comparar a variabilidade das componentes. Note que no gráfico superior esquerdo, a primeira componente possui uma variância muito maior (no eixo horizontal) que a segunda componente. Perceba ainda, que existem dois pontos muito discrepantes no eixo vertical. Considerando que a segunda componente é a populacional, não é difícil perceber que estes pontos representam a China e a Índia, os países mais populosos do mundo. Esta conclusão é reforçada pelo gráfico inferior esquerdo, em que agora, a segunda componente está no eixo horizontal, e estes dois pontos se destacam dos demais justamente neste eixo. Salientamos ainda que, desconsiderando os dois pontos mencionados, a variabilidade da terceira componente principal parece ser comparável a da segunda - uma hipótese a ser investigada -.

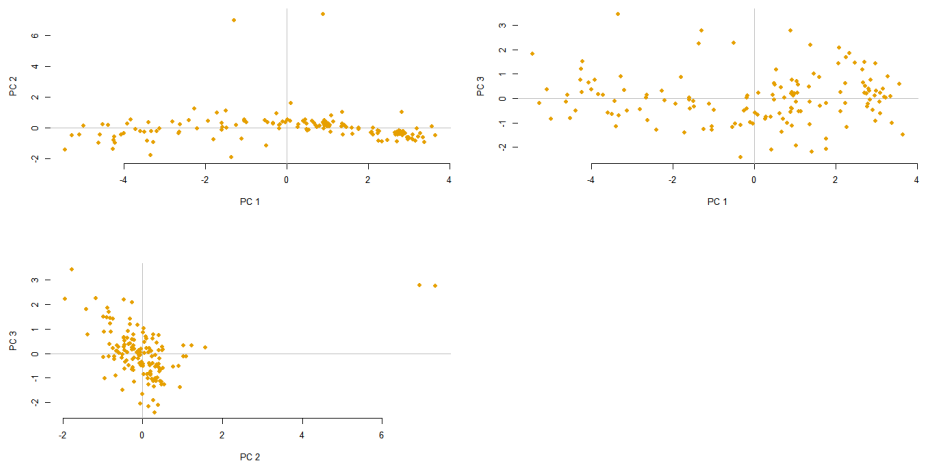

Figura 2: Relação das componentes principais. No gráfico superior esquerdo, estão as duas primeiras componentes; no superior direito, a primeira e a terceira componentes; no último a segunda e terceira componentes

Fonte: Censos e estimativas de órgãos oficiais e da ONU.

Finalmente, os resultados da segmentação considerado as três componentes principais pode ser visualizado na Figura 3. Vale salientar que o número de clusters escolhidos após a otimização foi $K = 4$.

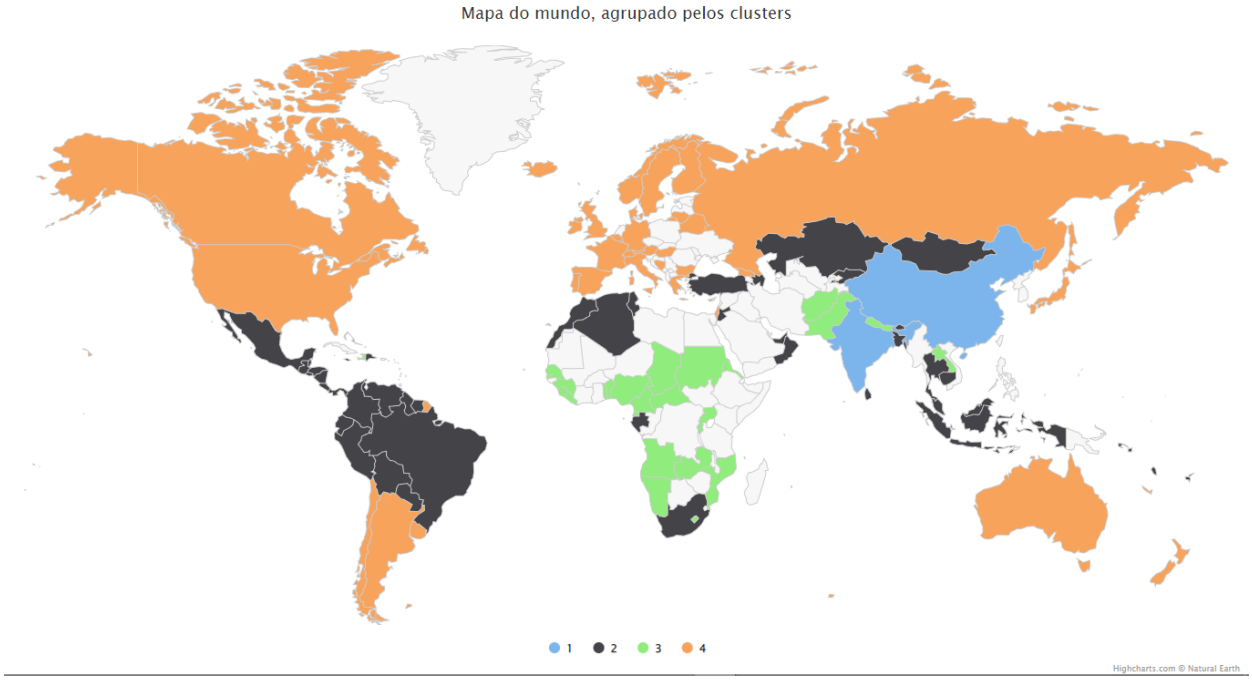

Figura 3: Segmentação mundial proposta
Fonte: Censos e estimativas de órgãos oficiais e da ONU.

Sobre a Figura 3, destacamos os seguintes fatores: os países em tom mais claro não foram segmentados; alguns por não possuírem informações atualizadas disponíveis publicamente (estes não fizeram parte de nenhum passo da análise após a exploratória), outros por limitações da metodologia utilizada para desenho do gráfico (estes foram, inclusive, ranqueados e são acessíveis no endereço disponibilizado no GitHub). Para a interpretação do gráfico, note que o cluster 1 só tem dois países: Índia e China. Possivelmente este cluster levou muito em conta a segunda componente principal; poderíamos chamá-lo de cluster dos países super-populosos.

O cluster dois é composto por vários países da América Latina, inclusive o Brasil. Alguns países africanos e outros mais orientais. Visualizando suas respectivas posições no ranking proposto, poderíamos relacioná-los, por exemplo, a antiga classificação de países em desenvolvimento. O cluster 3 é o que apresenta países que figuram nas piores posições do ranking. A maioria deles contidos no continente africano, com algumas poucas exceções. Novamente em uma eventual analogia a outras segmentações, este seria o grupo dos países mais pobres em desenvolvimento. Finalmente, o cluster 4 apresenta os países que figuram no topo do ranking e poderia ser chamado de cluster dos países desenvolvidos.

4 Discussões e conclusões

Diante do exposto, é válido que algumas coisas sejam comentadas.

Em primeiro lugar, os resultados obtidos em todos os passos da análise (componentes principais e segmentação) corresponderam as expectativas prévias; com algumas poucas exceções. Com base nisso, alguns poderiam inclusive afirmar que a análise realizada foi uma análise confirmatória de dados. Não consideramos que isso seja relevante e, ao invés de discutir o que já esperávamos, considerando inclusive o descrito por Sene and Moreira (1999), iremos focar a discussão no que era inesperado.

O agrupamento de Índia e China era inesperado. Apesar de as razões para isto ter acontecido serem claras - afinal são os dois países mais populosos do mundo e uma das três componentes principais era, basicamente, uma componente populacional - atualmente ambos se encontram em contextos socio-econômicos e geopolíticos diferentes. Este fator é uma das falhas no trabalho proposto: deveríamos ter utilizado a densidade demográfica do país, ao invés da população. Eventualmente seria possível que outros agrupamentos baseados apenas nessa variável fossem criados, mas possivelmente o peso desta componente seria menor que baseado na população.

Um segundo fator inesperado - este, não totalmente - é a divisão dos BRICS. Note que os 5 países foram divididos em 3 grupos diferentes. Em especial, a Rússia entrar no grupo de países desenvolvidos foi uma pequena surpresa, considerando alguns aspectos econômicos e sociais do país presentemente e as altas taxas de mortalidade e mortalidade infantil. Por outro lado, o alto índice de educação e de percentual urbano e a baixa taxa de natalidade podem ter contribuído para esta classificação. Outra hipótese pode ter relação com a segunda limitação deste trabalho, que envolvia não considerar dados igualmente recentes para todas as características estudadas. Isso também pode explicar a classificação da Argentina, que possivelmente não é, ao menos de forma consensual, um país que pode ser considerado desenvolvido atualmente.

Finalmente, os resultados da análise poderiam ter diferido se mais informações tivessem sido avaliadas. Isto é, uma eventual extensão do trabalho poderia considerar, por exemplo, taxa de mortalidade por armas de fogo, outros indicadores de educação, ou mesmo outros indicadores socio-econômicos no geral. Além disso, poderíamos ter considerado indicadores um pouco menos usuais, como o índice de felicidade e bem-estar ou indicadores de desigualdade.

Dito isso, esta é uma discussão não muito realizada no campo estatístico. O IBGE (2022) foi a fonte recente mais similar encontrada e apesar do trabalho fornecer informações de mais características (e, possivelmente, características atualizadas) dos países, não realiza um ranqueamento multivariado destas variáveis, tampouco, a análise de componentes principais. Não conseguimos o acesso às variáveis utilizadas pelo órgão; certamente, contribuiria muito para os objetivos aqui propostos.

Diante do exposto, é válido afirmar que os objetivos foram alcançados, ressaltar que os resultados obtidos foram similares aos esperados - exceto nos pontos já destacados - e que esta não precisa ser uma análise final isto é, alguns dos diversos pontos de melhoria foram apontados. Dito isso, apesar de agrupamentos com Índia e China, por exemplo, dificilmente resultar num bloco econômico, poderia resultar num bloco social. Certamente países com tão elevado número de habitantes possuem questões únicas e poderiam colaborar para aproveitá-las, ou resolvê-las.

Referências

IBGE (2022). Ibge — países. `https://paises.ibge.gov.br/#/`. (Accessed on 11/21/2022).

Johnson, R. A. and Wichern, D. W. (2007). Applied multivariate statistical analysis. 6th. *New Jersey, US: Pearson Prentice Hall*.

Mardia, K., Kent, J., and Bibby, J. (1979). *Multivariate analysis*. Probability and mathematical statistics. Acad. Press, London [u.a.].

Sagar, A. D. and Najam, A. (1998). The human development index: a critical review. *Ecological economics*, **25**(3), 249–264.

Sene, E. and Moreira, J. C. (1999). *Geografia Geral e do Brasil: espaço geográfico e globalização*. Scipione.