

Nicole Eduarda Baldin Colpani, Ricardo Semião e Castro , Yudi Siramidu Tujisoki

## **Agrupamento de países de acordo com dados sobre despesas no setor de saúde**

São Paulo

Novembro 2021

Nicole Eduarda Baldin Colpani, Ricardo Semião e Castro , Yudi Siramidu Tujisoki

**Agrupamento de países de acordo com dados sobre despesas  
no setor de saúde**

Fundação Getúlio Vargas

Orientador: Élia Yathie Matsumoto

São Paulo

Novembro 2021

# 1 Introdução

O presente artigo é motivado pela necessidade de entender se a classificação do Índice de Desenvolvimento Humano (IDH) é um ponto de referência adequado para instituições e governos quanto a quais países requerem maior atenção no setor de saúde. Com esse objetivo em mente, dados de 2017 e 2018 sobre despesas de saúde foram coletados para 189 países e dois métodos de clusterização foram utilizados para dividi-los em quatro grupos. Os resultados são comparados entre modelos e com o IDH do mesmo ano para verificar se são semelhantes ou não. Agrupamentos semelhantes ao índice reconhecido como a medida de desenvolvimento econômico pelo Programa das Nações Unidas para o Desenvolvimento (PNUD) implicam que o IDH contém informações satisfatórias sobre o dispêndio com saúde dos países. Nesse sentido, sua consideração seria suficiente quando organizações doadoras e governos precisam alocar fundos e/ou desenvolver melhores políticas públicas voltadas à saúde. Dito de outra maneira, resultados de clustering semelhantes com o IDH evidenciam que esse indicador é razoável para a avaliação objetiva e o planejamento sistemático por governos quanto a suas próprias carências em relação ao mundo no setor de saúde e por organizações internacionais quanto a quais países requerem mais atenção. A clusterização por  $K$ -médias apresentou resultados mais interpretáveis. Comparando o obtido por esse algoritmo com o IDH, verifica-se que as divisões são suficientemente similares.

## 2 Revisão Bibliográfica

De acordo com Gaertler (2005), sistemas de clusterização são utilizados para decompor um conjunto de entidades em “grupos naturais”. Esses procedimentos começaram a dar seus primeiros passos na década de 50 e, a partir da década de 70, a utilização desses algoritmos para análise inicial de dados tornou-se rotineira (GRODINS; YAMASHIRO, 1973). Do mesmo modo, a classificação de países em grupos é uma abordagem que está presente há algumas décadas na literatura. Abizadeh e Basilevsky (1986), por exemplo, baseiam-se em 21 variáveis socioeconômicas e utilizam análise de fator de máxima verossimilhança - método que os autores consideravam mais robusto, na época - para classificar 64 países.

Tratando do uso de métodos de clusterização para agrupamento de países, há grande diversificação das variáveis utilizadas e do objetivo de realizar essa classificação. Berlage e Terweduwe (1988) empregam análise de *clusters* - e de fatores - em diversas características nacionais para entender se a classificação de agências internacionais de quais países são menos desenvolvidos e quais são recentemente industrializados é baseada em critérios objetivos ou se ela se dá a partir de pressões políticas.

A decomposição do conjunto de nações pode ser fundamentada em questões culturais como fatores históricos, geográficos, atitude dos empregados no trabalho, preferências de estilo gerencial e/ou outras dimensões culturais (MENSAH; CHEN, 2013). Uma abordagem alternativa é agrupar com enfoque em variáveis ligadas ao crescimento econômico e desenvolvimento dos países como o feito por Mylevaganam (2017) e Basel, Gopakumar e Rao (2021). Ambos os papers citados buscam indicadores mais completos que o IDH para a classificação de países. Mylevaganam (2017) usa análise de componentes principais para obter pesos mais adequados aos índices que compõem o IDH e, então, aplica  $K$ -médias para classificar os países quanto ao seu nível de desenvolvimento. Basel, Gopakumar e Rao (2021), por sua vez, coletam quatro índices - índice econômico, social, de sustentabilidade e institucional - de 1996 a 2005 para 102 países e usam análise relacional cinza e  $K$ -médias com o mesmo objetivo de agrupar países com base em seus níveis de desenvolvimento.

Entre os artigos da literatura, Proksch et al. (2019) e Tang e Salvador (1986) mais se relacionam com o presente artigo. O primeiro pelo enfoque em variáveis relacionadas à saúde e o segundo por ter um intuito similar. Em Proksch et al. (2019), utiliza-se uma abordagem de multi-indicadores e clusterização hierárquica para agrupar 30 nações da OCDE em quatro *clusters* a partir de seus sistemas nacionais de inovação em saúde. Tang e Salvador (1986), por outro lado, empregam indicadores de desenvolvimento de interesse da Comissão Econômica e Social das Nações Unidas para a Ásia e o Pacífico (no inglês, ESCAP) e o algoritmo  $K$ -médias para agrupar os países com níveis de desenvolvimento similares na região. A iniciativa se deu para que a agência de desenvolvimento citada pudesse obter uma avaliação objetiva e realizar um planejamento sistemático da alocação de seus recursos limitados e dos recursos de outras instituições internacionais.

### 3 Dados

As variáveis coletadas para realização dos *clustering* incluem: despesas atuais com saúde como % do PIB, despesas domésticas com saúde como % das despesas atuais com saúde e despesas domésticas do governo geral como % das despesas atuais com saúde. Foram utilizados também dados de despesas domésticas privadas como % das despesas atuais com saúde, seguro de saúde voluntário como % das despesas atuais com saúde e despesas externas com saúde como % das despesas atuais com saúde.

Os dados referem-se aos anos de 2017 e 2018 para um conjunto de 189 países. Essas informações foram retiradas do Banco de Dados Global de Despesas com Saúde da Organização Mundial do Comércio. O Índice de Desenvolvimento Humano (IDH) dos países inseridos na análise, que foram usados após o processo de clusterização para fins de comparação, foram obtidos de um relatório de desenvolvimento humano publicado pelo Programa das Nações Unidas para o Desenvolvimento.

## 4 Metodologia

Nosso primeiro objetivo consiste em agrupar os países em diferentes grupos, de forma que a semelhança intragrupo de um país seja mais significativa do que a entre grupos, em relação aos gastos de saúde. De forma simples, isso representa um aprendizado não supervisionado. No aprendizado não supervisionado, trabalhamos com dados não rotulados (*unlabeled data*), isto é, temos inúmeras observações, cada qual com diversas características, entretanto, sem uma devida categorização. Para estudos como o nosso, em que trabalhamos com informações sobre o sistema de saúde de diversos países, é importante agrupar as observações em *clusters* usando métricas de similaridade - ou dissimilaridade - as quais verificam, de acordo com um critério pré-determinado, o grau de semelhança das informações que temos. Nesse sentido, obteríamos conjuntos categorizados conforme a similaridade existente dentro de cada *cluster*.

Os métodos utilizados nesse processo são chamados de métodos de agrupamento, ou clusterização. O primeiro método nesse artigo, chamado de  $K$ -médias, é amplamente utilizado em artigos de aprendizado não supervisionado relacionados à agrupamento de países. De partida, é muito importante escolher o número  $K$  de *clusters* que serão formados no final do aprendizado. Tendo isso decidido, são estabelecidos  $K$  pontos iniciais para cada grupo de países. A métrica de dissimilaridade utilizada nesse caso é a distância Euclidiana ao quadrado.

$$\begin{aligned} d(x_i, x_{i'}) &= \sum_{j=1}^p (x_{ij} - x_{i'j})^2 \\ &= ||x_i - x_{i'}||^2 \end{aligned}$$

Em que  $x_i$  representa uma observação da base de dados e  $j$  é o índice para as características disponíveis sobre as observações. O algoritmo de  $K$ -médias, partindo dos pontos iniciais pré-estabelecidos, amplia os *clusters* de tal modo que a distância Euclidiana entre a média empírica do *cluster* e os pontos existentes dentro dele seja minimizada. Assumindo que  $x_i$  represente um dos 189 países,  $j$  as informações sobre despesas de saúde de cada um,  $c_k$  um dos  $K$  *clusters* e  $\mu_k$  a média do *cluster*  $k$ , temos

$$J(c_k) = \sum_{x_i \in c_k} ||x_i - \mu_k||^2 \quad (4.1)$$

Assim, no final, quando todos os pontos estiverem alocados em pelo menos um *cluster*, esperamos que a soma de todas as distâncias Euclidianas de todos os  $K$  *clusters* seja minimizada.

O outro método - menos comum em outros artigos com análise semelhante - é o de agrupamento, ou clusterização, hierárquico. Diferentemente do anterior, não é necessário estabelecer previamente um número  $K$  de grupos que serão formados e, além disso, há uma maior flexibilidade na escolha do tipo de medida de dissimilaridade, já que existem, no total, quatro formas diferentes de verificar a semelhança dentro de cada grupo, sendo que a decisão depende da base de dados e do objetivo a ser alcançado.

Existem duas estratégias para esse sistema: aglomerativa ou divisiva. Nesse trabalho, focaremos na primeira delas. Nela, partimos da base da hierarquia em que cada país representa um grupo individual. A cada nível acima, os *clusters* são agrupados de acordo com a medida de dissimilaridade, sendo que, nesse artigo, a união é feita quando a distância entre os centroides de dois *clusters* sendo comparados é minimizada. No final, obtemos um dendrograma, no qual no topo dele está o grupo que inclui todos os países. Assim sendo, podemos observar diversas categorizações com um número variado de *clusters* para cada uma. A depender da análise sendo feita, corta-se a árvore no ponto em que a quantidade  $K$  de grupos é representada.

Após realizado o aprendizado não supervisionado, além de comparar as duas metodologias, realizamos uma comparação com a categorização de países feita de acordo com o Índice de Desenvolvimento Humano (IDH), o qual é amplamente difundido e discutido na bibliografia relacionada ao nosso trabalho. Esse índice é utilizado para classificar os mesmos países que estão sendo considerados na nossa base de dados em quatro categorias de acordo com o desenvolvimento humano: baixo (0-0.499), médio (0.500-0.799), alto (0.800-0.899) ou muito alto (maior que 0.900). Para fazer essa separação são consideradas, principalmente: (1) uma vida longa e saudável, medida pela expectativa de vida; (2) conhecimento, medido pela taxa de alfabetização de adultos e pela taxa de matrícula nos ensinos primário, secundário e terciário; e (3) um padrão de vida decente, representado pela Renda Nacional Bruta (RNB) per capita. Além de ser uma medida padronizada de bem-estar para classificação dos países, o IDH é uma ferramenta que pode facilitar a formulação de políticas econômicas e a alocação eficiente de recursos (KINNUNEN et al., 2019). Dessa forma, ele é utilizado por instituições



mundiais de apoio humanitário, como ONGs, para identificar países que necessitam de uma intervenção e por governos para criação ou melhoria de políticas públicas.

Assim sendo, como há quatro grupos principais estabelecidos pelo IDH, criamos quatro *clusters* com os métodos de agrupamento descritos acima. A partir da análise comparativa entre os grupos obtidos pelo aprendizado supervisionado e a categorização dada pelo IDH, nos anos de 2017 e 2018, esperamos que haja uma compatibilidade na divisão dos países. Caso isso seja verdade, o IDH pode ser usado como indicador de que países necessitam de ajuda humanitária das organizações ou de melhores políticas públicas internas pelo governo na área de saúde. Caso contrário, outros índices, como os utilizados nesse trabalho, são preferíveis para identificar a precariedade da saúde e/ou a urgência de uma intervenção.

## 5 Resultados

Em primeiro lugar, observamos o resultado do agrupamento pelo método de  $K$ -médias. Como determinado na seção anterior, nos anos de 2017 e 2018, os países foram divididos em quatro *clusters* diferentes partindo de pontos iniciais estabelecidos aleatoriamente. A categorização foi feita de acordo com a similaridade dos países em relação a seus dados sobre gastos no setor de saúde. Entretanto, para uma simples visualização dos *clusters*, os gráficos a seguir estão representados na segunda dimensão, quando, na verdade, a dimensão correta é aquela cujo valor é igual a quantidade de variáveis sendo consideradas no agrupamento, o que é impossível de observar na realidade. Nesse sentido, não dá para saber com certeza que dimensões estão sendo consideradas nos eixos.

Comparando os dois anos é possível notar que não há grandes diferenças na constituição dos *clusters*, e a maior parte dos países que mudam de grupo de um ano para outro, estavam na fronteira no ano de 2017, ou seja, já apresentavam semelhança com ambos os grupos. Focando no gráfico do ano de 2017, o grupo 1 tem como centro Serra Leoa, cujo gasto atual do governo com saúde (*Current Health Expenditure*, CHE) é de 17% do PIB, contudo 66% do CHE é doméstico - sendo 55% proveniente do setor privado e 11% do governo - e 34% do CHE são transferências externas. Isso transmite a ideia de que há uma necessidade de saúde significativa, mas o país não tem condições de garanti-la recorrendo à ajuda externa. Interessante notar que no mesmo grupo temos Brasil e Estados Unidos. Isso provavelmente ocorre, pois ambos apresentam CHE altos, 9% e 17% do seus PIBs, respectivamente, provavelmente, devido ao fato de terem populações grandes e territórios extensos, necessitando de um grande sistema de saúde para suprir o país inteiro. Entretanto, estão na fronteira e próximos aos outros grupos, já que não dependem de transferências externas, ambos com 100% do CHE doméstico, sendo que, no Brasil, 42% é sustentado pelo governo e, nos EUA, esse valor é de 50%.

No grupo 2, a Armênia está no centro. O país tem CHE de 10% do PIB, porém, 99% é doméstico, com 13% proveniente do governo e 85% do setor privado. Isso demonstra uma maior estabilidade do setor de saúde nacional, o qual não necessita de ajuda externa.

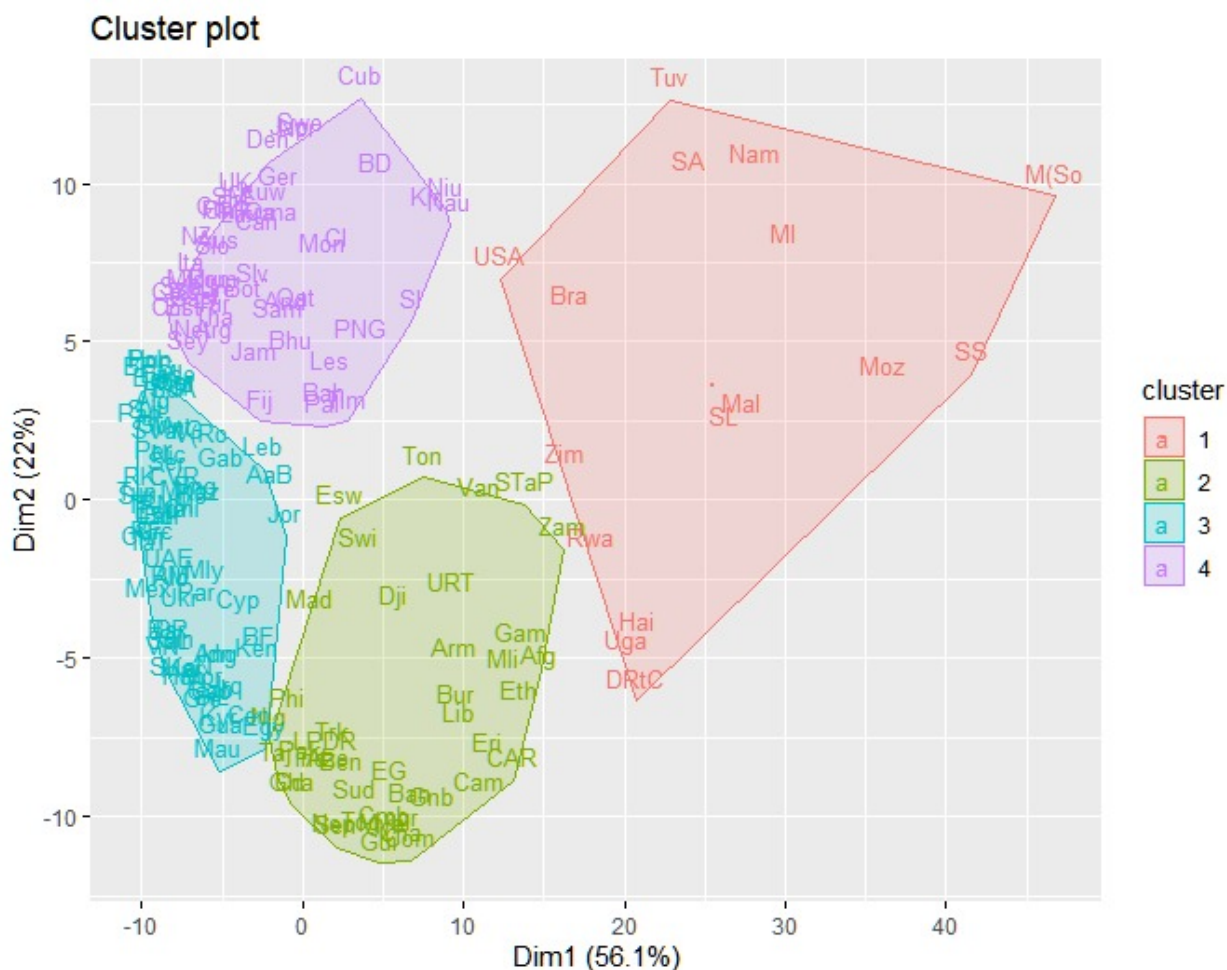


Figura 1 – 2017

Enquanto isso, no grupo 3, podemos escolher a Malásia como representante. O CHE é de 4% do PIB, sendo 100% de origem doméstica. O governo é o maior provedor, com 52% do CHE, em comparação com o setor privado, o qual fornece 48%. Isso pode demonstrar uma menor necessidade de investimento no setor de saúde ou a falta dele, contudo, é importante notar que o governo aparenta estar cumprindo o seu dever de prover o bem público.

Por último, no grupo 4, escolhemos o Qatar. O país apresenta CHE de 4% do PIB, sendo 100% de origem doméstica. O ponto diferencial é a grande participação do governo, com 72% do CHE doméstico. O governo, nesse caso, também aparenta estar fazendo sua parte, mas não conseguimos identificar se não há necessidade de maiores investimentos ou se a urgência é ignorada. Tendemos a acreditar que a primeira opção é mais provável, pois no mesmo *cluster* temos países como Japão e Alemanha.

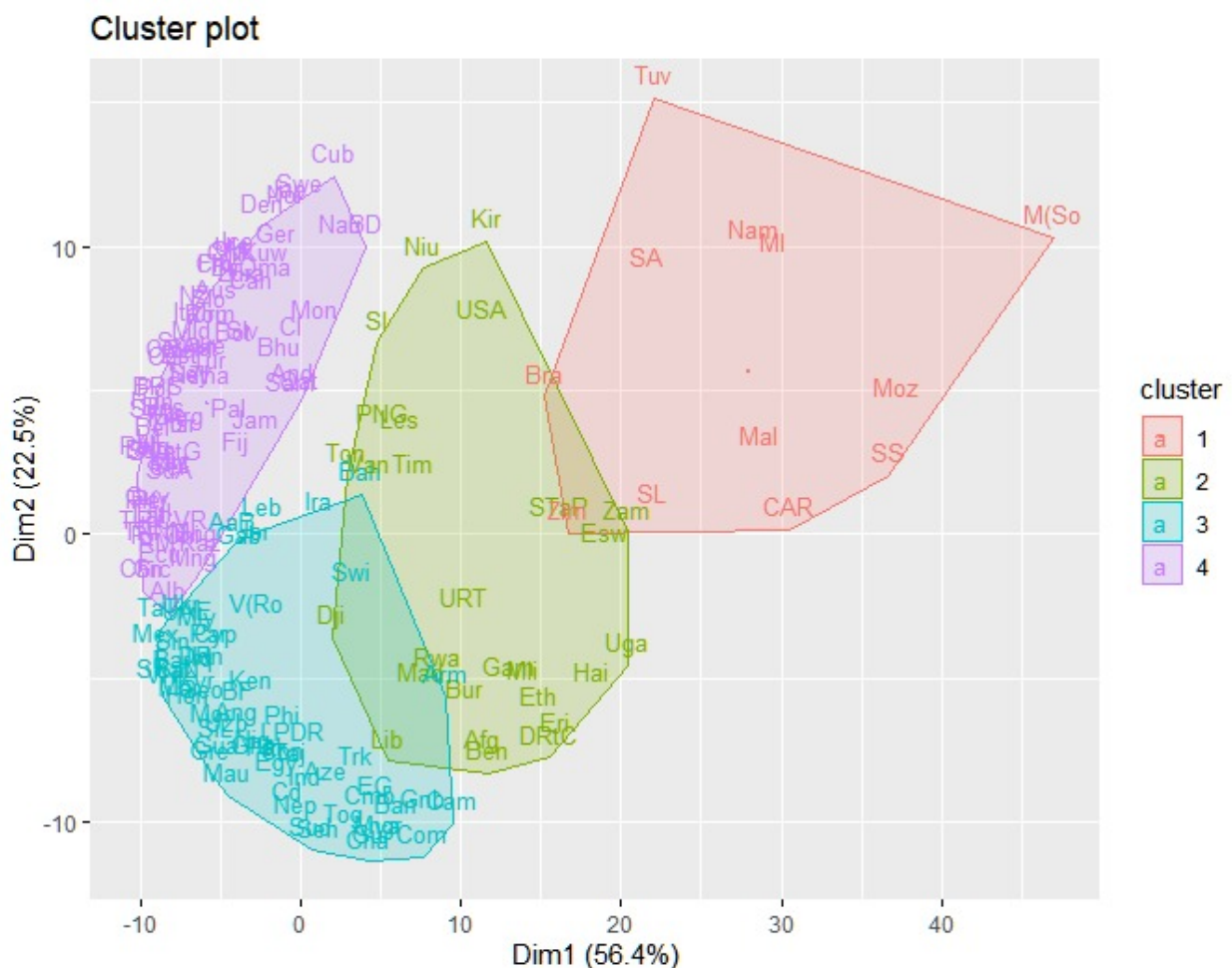


Figura 2 – 2018

Em segundo lugar, obtemos a divisão dos países por meio do método de clusterização hierárquica. Neste caso, construímos dois dendrogramas, um para cada ano, e cortamos a árvore de forma que fossem formados quatro *clusters*, isto de acordo com a medida de dissimilaridade baseada nos centroides. Analisando-os, é possível identificar, para ambos os períodos, três grupos bem definidos. Entretanto, isso acontece, pois incluem muitos poucos países ou, até mesmo, um único país, ou seja, os *clusters* são categorizados por características únicas. Um problema disso é que - apesar de poderem ser considerados como casos excepcionais - eles, provavelmente, não estão alocados de forma correta, isto é, ao invés de estarem isolados, deveriam estar próximos de países com os quais apresenta grande similaridade. A exemplo, temos a Micronésia (“M(So)”) no ano de 2017, a qual já no primeiro nível do dendrograma se separou em um *cluster* único diferente de todos os outros países. Algo que parece improvável,

dado o tamanho do país e a provável semelhança com outros países mais pobres. Ademais, outra característica a ser destacada nos gráficos obtidos são os *clusters* que incluem uma enorme quantidade de países. Isso dificulta muito a determinação de uma característica em comum entre todos integrantes do grupos, tornando a categorização mais fraca e de difícil generalização.

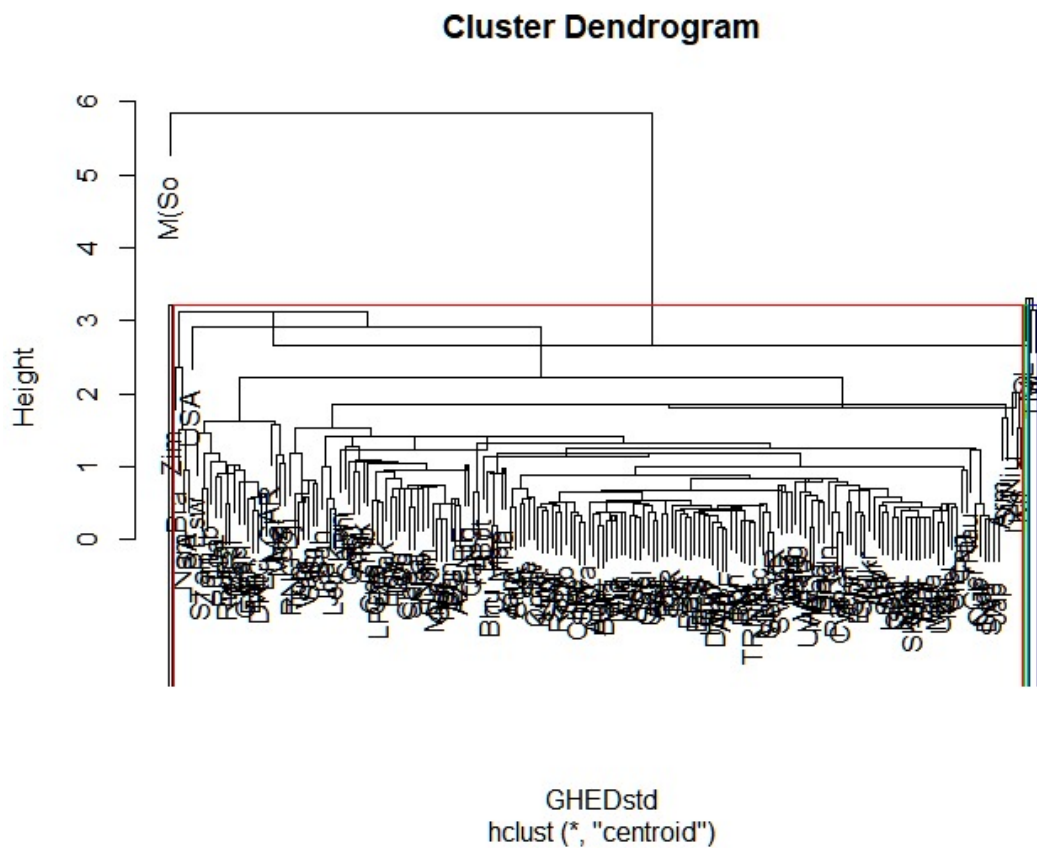


Figura 3 – 2017

Apesar disso tudo ser uma provável consequência da escolha de um  $K$  pequeno e os resultados não serem necessariamente errados, mas, sim, problemáticos, ainda assim esperávamos uma divisão mais clara, com grupos incluindo nem pouco e nem muitos países. Caso esse fosse o resultado obtido, a comparação da lista de países do IDH com os *clusters* da clusterização hierárquica seria mais simples e de melhor qualidade. Entretanto, como isso não aconteceu, é preferível focarmos nos resultados obtidos pelo método de  $K$ -médias.

Finalmente, comparando as categorias de países de acordo com IDH e os quatro

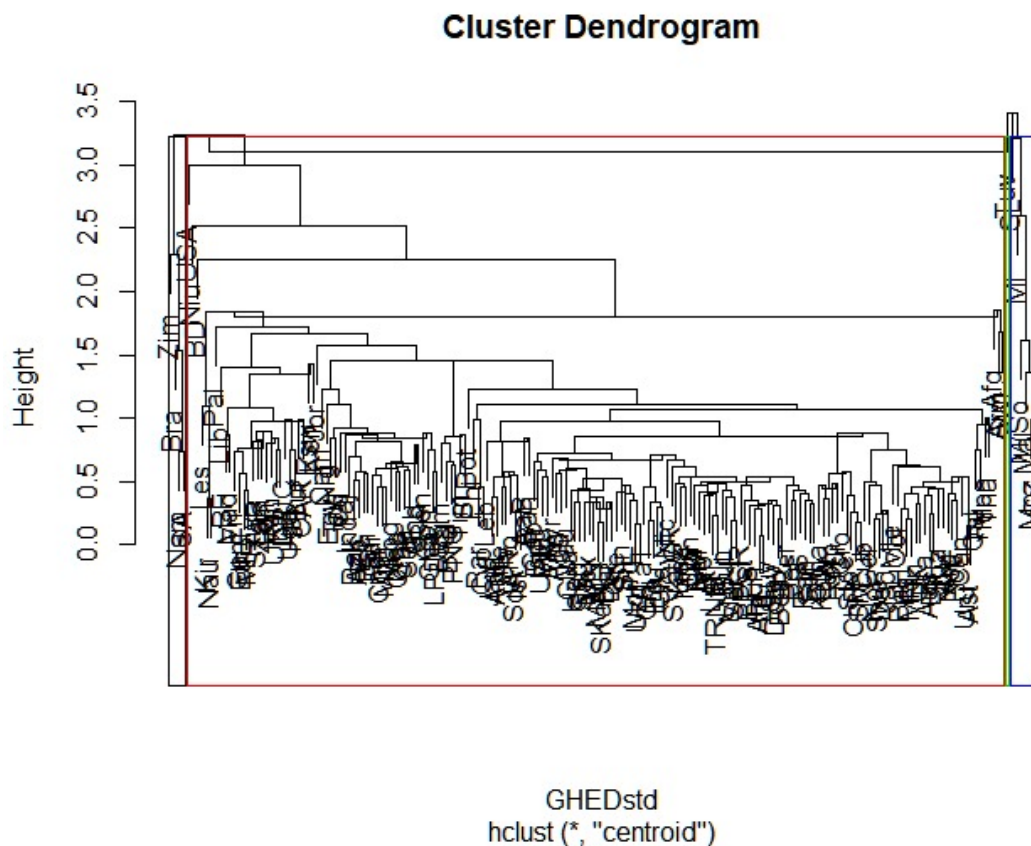


Figura 4 – 2018

*clusters* formados pelo método de  $K$ -médias, no geral, no grupo 4 temos países de IDH muito alto ou alto; no 3, estão alocados países de IDH alto, com alguns de IDH médio; no 2, a maior parte tem IDH médio, sendo alguns de IDH baixo; e no 1, a grande maioria apresenta IDH baixo.

Dadas algumas exceções, é possível afirmar que há uma compatibilidade entre as duas divisões abordadas, em que os países com maior IDH estão em grupos com menor necessidade de ajuda no setor de saúde e com governos como maiores provedores desse bem público. Nesse sentido, apesar de não ser de forma perfeita, o IDH pode ser utilizado para por governos e instituições como um ponto de referência para a análise dos países quanto ao setor de saúde. Todavia, não podemos excluir a viabilidade de outros índices de desenvolvimento ou de outros também focados na qualidade de saúde.

## 6 Conclusão

O objetivo principal do trabalho era utilizar os conhecimentos de aprendizado não supervisionado para agrupar 189 países a partir de informações obtidas sobre despesas no setor de saúde. Mais especificamente, foram usados dois métodos de *clustering*. O primeiro, amplamente utilizado em outras bibliografias que abordam temas semelhantes, é o  $K$ -médias. O resultado final são quatro *clusters*, em que o primeiro é constituído de países com maiores necessidades de ajuda no setor de saúde, dado que apresentam alto CHE, sustentado em grande parte por transferências externas. O segundo grupo, por sua vez, contém países com menores necessidades, recebendo menos transferências, mas ainda com o setor privado sendo um provedor significativo de saúde. Adicionalmente, o terceiro e o quarto grupos são caracterizados por países com maior segurança nesse setor da sociedade, sendo a intensidade de participação do governo no provimento desse bem, a principal diferença.

O segundo método abordado foi o de clusterização hierárquica. Esperávamos obter grupos mais bem definidos e categorizados. Contudo, obtivemos dendrogramas com dois tipos de *clusters*: o primeiro com apenas um ou poucos países e o segundo com inúmeros países, entre os quais seria difícil estabelecer uma conexão. Os resultados não estão necessariamente errados, mas não provêm respostas de interpretação intuitiva, de maneira que a comparação do o IDH não é realizada.

Concluimos que a divisão estabelecida pelo método de  $K$ -médias é compatível com a categorização do IDH. Isso fornece um argumento a favor da utilização desse índice como identificador de países que necessitam de ajuda externa para melhoria do setor de saúde e/ou como ponto de referência para melhores políticas públicas pelos Estados. Assim sendo, observamos o sucesso do método de  $K$ -médias enquanto a nossa tentativa de divergir da bibliografia existente, usando clusterização hierárquica, não nos forneceu resultados favoráveis. Isso pode ser visto como um possível ponto de melhora para pesquisas futuras. Além disso, nós nos limitamos à comparação ao IDH, pois ele é o índice mais comumente utilizado e é de simples interpretação. No entanto, poderíamos expandir nosso estudo ao comparar os resultados com outros índices que deem foco para o setor de saúde ou repetir os passos desse

artigo partindo de dados iniciais mais gerais que representem o desenvolvimento do país como um todo, por exemplo.



# Referências

- ABIZADEH, S.; BASILEVSKY, A. Socioeconomic classification of countries: a maximum likelihood factor analysis technique. *Social Science Research*, Elsevier, v. 15, n. 2, p. 97–112, 1986.
- BASEL, S.; GOPAKUMAR, K.; RAO, R. P. Classification of countries based on development indices by using k-means and grey relational analysis. *GeoJournal*, Springer, p. 1–19, 2021.
- BERLAGE, L.; TERWEDUWE, D. The classification of countries by cluster and by factor analysis. *World Development*, Elsevier, v. 16, n. 12, p. 1527–1545, 1988.
- GAERTLER, M. Clustering. In: *Network analysis*. [S.l.]: Springer, 2005. p. 178–215.
- GRODINS, F.; YAMASHIRO, S. Optimization of the mammalian respiratory gas transport system. *Annual review of biophysics and bioengineering*, Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA, v. 2, n. 1, p. 115–130, 1973.
- KINNUNEN, J. et al. The role of economic and political features in classification of countries in transition by human development index. *Informatica Economică*, v. 23, n. 4, p. 26–40, 2019.
- MENSAH, Y. M.; CHEN, H.-Y. Global clustering of countries by culture—an extension of the globe study. *Available at SSRN 2189904*, 2013.
- MYLEVAGANAM, S. The analysis of human development index (hdi) for categorizing the member states of the united nations (un). *Open Journal of Applied Sciences*, Scientific Research Publishing, v. 7, n. 12, p. 661–690, 2017.
- PROKSCH, D. et al. National health innovation systems: Clustering the oecd countries by innovative output in healthcare using a multi indicator approach. *Research Policy*, Elsevier, v. 48, n. 1, p. 169–179, 2019.
- TANG, J. C.; SALVADOR, P. Classification of countries for international development planning using cluster analysis. *Socio-Economic Planning Sciences*, Elsevier, v. 20, n. 4, p. 237–241, 1986.