Estatística Multivariada

Análise Global de Suicídios (1985-2016)

João Pedro de Noronha Funenga

MAEBD - Nº 61635

j.funenga@campus.fct.unl.pt

14 de janeiro de 2022

Conteúdo

Lista de Figuras

Lista de Tabelas

1 Introdução

Este projeto no âmbito da unidade curricular de *Estatística Multivariada* terá como objetivo analisar as taxas mundiais de suicídio e tirar algumas conclusões práticas a partir dos dados. Existem esterótipos e preconceitos pré-existentes acerca deste tema, como por exemplo, os homens serem o grupo predominante entre as vítimas ou pessoas de idade avançada tenderem a optar por esta saída (muitas das vezes devido a doenças crónicas). Neste trabalho irei testar se de facto os dados corroboram as teorias prévias ou não. Visto existirem países com condições e enquadramentos muito distintos, julgo que uma boa forma de abordar o problema será criando grupos (*clusters*) definidos por um conjunto de variáveis e veremos como estes de facto se encontram separados e existem diferenças significativas entre clusters diferentes.

2 Descrição do *dataset*

Como falado na introdução, para este trabalho decidi escolher um conjunto de dados relativos aos suicídios entre 1985 e 2016. Para além do número de suicídios, também existem as colunas relativas ao grupo etário existindo 5 (15-24, 25-34, 35-54, 55-74, 75+ anos), ao género, ao país e respetiva população e PIB per capita. Existem outras colunas adicionais mas que não utilizarei.

3 Metodologia

Para este projeto, utilizarei um dos mais conhecidos algoritmos de clustering, o ***k-means***. Este algoritmo funciona agrupando os dados em k grupos (sendo k um parâmetro que precisamos de passar). Cada um dos grupos será representado pelo seu centróide (ponto médio do cluster). Por sua vez, cada um dos pontos (que correspondem a uma observação do ficheiro *"suicides.csv"*), será atribuído ao cluster cuja **distância euclidiana ao centróide correspondente é menor**. Para cada iteração do algoritmo, a posição dos centróides vai alterando o que também afeta a que cluster é que os exemplos são alocados. Este algoritmo tem como condição terminal a não alteração da atribuição dos exemplos aos clusters entre iterações consecutivas. Por vezes, ao aplicarmos este algoritmo, temos como ideia o número de *clusters* finais que queremos. No entanto, este não é o caso. De modo a calcular o número de grupos ideal, irei recorrer a uma técnica falada em aula. Esta técnica é chamada de ***Silhouette Method*** e funciona da seguinte forma:

1. Realizamos a análise de clusters considerando um **intervalo de valores para k** (no meu caso k $\in$ [1,10])

2. Para cada valor de k **calcular a média dos coeficientes de *Silhouette***

3. Representar o gráfico que relaciona os k testados com a média dos coeficientes *Silhouette* entre todos os clusters.

4. O valor escolhido para k corresponderá ao **máximo da função** (em que a média dos coeficientes *Silhouette* têm o valor mais alto).

Depois de ter os clusters calculados e as respetivas observações agrupadas, irei analisar cada um deles e verificar se existe algum padrão que seja possível de revelar.

4 *Overview* Inicial

Primeiramente, iremos olhar para a evolução da taxa média de suicídios desde 1985 até 2016.

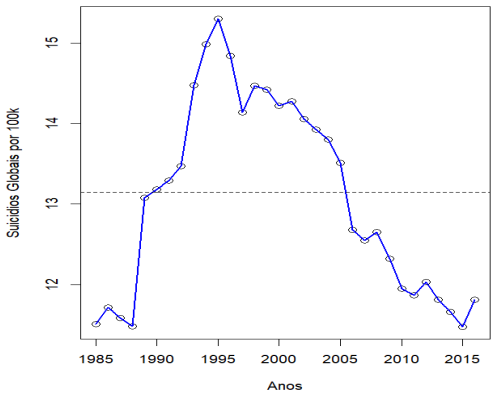

Figura 1: Taxa de suicídio ao longo dos anos

Como podemos observar, esta parece estar a decrescer desde aproximadamente 1995 e, anteriormente a esta data, os números eram bastante mais baixos. Possivelmente, isto pode ser explicado por nesta altura não haver um registo deste tipo de dados tão minucioso como agora e que uma boa parte de suicídios que aconteciam não eram sequer registados como tal. O decréscimo desde o final do milénio pode estar relacionado com a maior abertura da população em geral em falar abertamente da saúde mental. Para além disto, no gráfico também representei a **taxa média de suícidos global** com a linha a traçejado, que se situa nos **13.15 por 100.000 pessoas**.

5 Procedimento

Tendo a metodologia a usar sido explicada em cima, irei agora explicar como procedi. Primeiramente, decidi ajustar o *dataset* de modo a facilitar o uso depois. Algumas das alterações que realizei foi a alteração de nomes das colunas bem como descartar outras que não fariam sentido serem utilizadas. Um aspeto importante a referir foi a questão de necessitar de **padronizar os dados**. Ao verificar os atributos que irei analisar e que servirão para a constituição dos clusters, é possível perceber que estão em escalas e unidades completamente diferentes o que me leva a ter de os *standardizar* usando a função *scale*. Inicialmente por uma questão experimental, tentei perceber qual seria o resultado de não os escalar e, utilizando o número de grupos dado pelo método da *silhouette* (neste caso seriam 2 grupos), os dados apresentavam-se agrupados de uma forma que não fazia qualquer sentido, sendo assim outro fator que me levou a realizar este pré-processamento.

Como enunciado, o algoritmo de *k-means* tem como parâmetro o **número de *clusters*** e este foi obtido da forma explicada na metodologia. Dando uso a funções *built-in* do R, fiz a representação gráfica da relação entre o número de grupos e o valor médio do coeficiente de *Silhouette*.

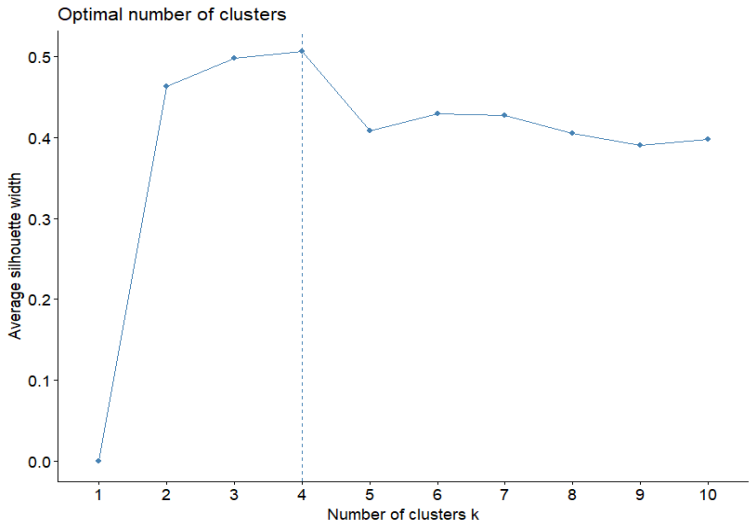

Figura 2: Coeficiente *Silhouette* vs nº de clusters

Como podemos ver pelo gráfico, o número ideal de *clusters* a utilizar é de 4, o que, comparado aos 2 dados utilizando os dados não *standardizados* faz muito mais sentido. Agora que tenho o valor para o parâmetro k, posso então utilizar a função do R para realizar o algoritmo em si. Esta função tem como **parâmetros** os **dados**

correspondentes aos atributos a serem utilizados, o **número de *clusters*** e o **número de inícios aleatórios** que serão feitos.

Para o primeiro argumento, os atributos que decidi utilizar para a representação dos clusters foram a **população do país**, o **número de suicídios por 100.000** pessoas bem como o ***PIB per capita* do país**. Julgo que a combinação destas três *features* será representativa para separar as observações nos vários grupos, isto porque são fatores que muitas das vezes estão relacionados entre si e têm uma fácil interpretação.

Para o segundo argumento, como vimos, o número de *clusters* a serem gerados é de 4, o correspondente ao máximo da imagem 2.

Relativamente ao último e terceiro argumento, este representa o número de inícios aleatórios que o algoritmo fará. Basicamente, o *k-means* começa por atribuir aleatoriamente a posição dos centróides que utilizará mas, como é natural, existem combinações de posições melhores que as outras. Por exemplo, caso os 4 centróides fossem todos colocados muito próximos uns dos outros, isto levaria a que os grupos gerados não fizessem grande sentido em termos de interpretação e por isto, com este argumento serão feitos n inícios e será escolhido o que melhor separar os dados (o que tiver uma variância intra-cluster baixa), para este parâmetro decidi utilizar um valor de 10 (cf. [1]).

Analisaremos então o gráfico que demonstra os 4 agrupamentos dos dados.

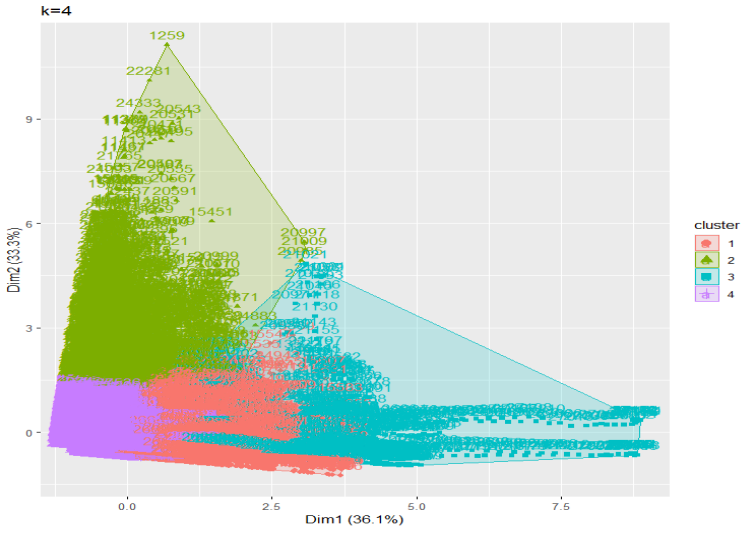

Figura 3: Clusters Gerados

Analisando o gráfico percebemos que os 4 grupos formados têm formas muito diferentes bem como o número de elementos que os constituem.

Tabela 1: Número de elementos dos clusters

Cluster 1	Cluster 2	Cluster 3	Cluster 4
4448	2321	1429	19622

A sobreposição que se verifica no gráfico está relacionada com o número de atributos selecionados (que correspondem à dimensão do espaço). Por ter escolhido 3 atributos e estar a fazer a representação num espaço bi-dimensional, existem pontos que se encontram sobrepostos.

6 Análise dos Clusters

Com os 4 clusters que obtive, irei analisar cada um e verificar o que, aproximadamente, cada um deles pode representar.

Tabela 2: Valores médios

	População	**Suicídios/100k**	**PIB *per capita* ($)**
Cluster 1	1384540.5	11.0	50607.8
Cluster 2	1130833.3	62.3	12041.4
Cluster 3	15487347.9	12.0	21008.5
Cluster 4	1040038.6	7.4	9486.9

6.1 Cluster 1

Analisando os valores médios para o primeiro *cluster*, percebemos que este tem um valor de aproximadamente 50k$, sendo assim o primeiro mais alto dos 4. Olhando agora para os suicidios por 100k pessoas, percebemos que este tem um valor de apenas 11 pessoas por cada cem mil, sendo assim o segundo com os números mais baixos. Relativamente à população média, os países que o constituem têm um valor aproximadamente de pouco menos de 1 milhão e meio de pessoas, o que é um número bastante baixo. Isto aparenta transmitir a mensagem de que quem vive em países não muito grandes com uma boa qualidade económica, não tende a optar por esta saída. Veremos agora, das observações, quais são os 5 países que mais aparecem neste cluster.

Tabela 3: Cluster 1 - Países Frequentes

País	Nº Ocorrências
Luxemburgo	300
Noruega	297
Dinamarca	252
Suiça	238
Islândia	226

Ao analisarmos a tabela supradita, percebemos que, para os países com mais ocorrências no cluster, todos eles são europeus. Isto vai ao encontro de o que podemos pensar. Países europeus têm um *PIB per capita* elevado e não apresentam um número muito elevado de suicídios.

Para além dos países que o constituem, veremos agora em que faixa etária é que há mais prevalência de suicídios.

Tabela 4: Cluster 1 - Frequências nas várias faixas etárias

Faixa Etária	Nº Ocorrências
25-34	793
15-24	785
5-14	755
55-74	734
35-54	691
75+	690

Pela tabela acima, percebemos que o número de suicídios é bastante homogéneo ao longo das várias faixas etárias. No entanto, tem uma maior prevalência nas gerações mais jovens que nas com mais idade. Por último, veremos a proporção de homens para mulheres.

Tabela 5: Cluster 1 - Frequências nos géneros

Género	Nº Ocorrências
Mulheres	2241
Homens	2207

Neste caso podemos observar que o número de suicídios em homens e mulheres foi muito semelhante e não aparenta representar o esterótipo que existe.

Relativamente ao valor de variabilidade intra-cluster, este apresenta um de 7158.1. Por este valor por si só não ter grande significado, será depois comparado quando chegar ao último *cluster*.

6.2 Cluster 2

Veremos agora o que acontece no segundo *cluster*. A primeira coisa que salta à vista é o elevado número de suicídios por cem mil habitantes. Para além disto, o *PIB per capita* é também bastante inferior ao primeiro, cerca de 5 vezes menos. Relativamente ao número médio da população, este é bastante semelhante ao primeiro. Este cluster aparenta representar o estereótipo de que países com menos capital sofrem de uma maior taxa de suicídio entre os seus habitantes. Com isto, analisaremos agora os 5 primeiros países.

Tabela 6: Cluster 2 - Países Frequentes

País	Nº Ocorrências
Cazaquistão	121
Ucrânia	111
Lituânia	107
Hungria	98
Bielorrússia	90

No primeiro cluster tivémos países maioritariamente da europa central e ocidental que, por norma, têm melhor reputação em diversos índices de felicidade (cf. [2]). Agora, ao vermos a tabela acima percebemos que todos os países que constituem este cluster ou são da europa de leste ou asiáticos, como é o caso do Cazaquistão. Estas observações vão de encontro com as ideias pré-formuladas que todos temos de nestas zonas, pela qualidade de vida ser menor, ressalvando ser uma generalização, aliada aos baixos vencimentos, o número de suicídios demonstra ser bastante mais alto que nos outros grupos. De seguida vamos ver quais são os grupos etários que mais aparecem neste *cluster*.

Tabela 7: Cluster 2 - Frequências nas várias faixas etárias

Faixa Etária	Nº Ocorrências
75+	1003
55-74	527
35-54	400
25-34	276
15-24	115
5-14	0

Ao contrário do que acontece no primeiro cluster relativamente à taxa de suicídios ser parecida nas várias faixas etárias, neste caso conseguimos claramente perceber que o número é aproximadamente o dobro para pessoas no final de vida. Este seria o pressuposto que muitos de nós faríamos, por haver uma panóplia de fatores que podem fazer alguém mais velho não querer continuar, sendo o caso mais comúm o da presença de doenças crónicas ou mesmo o da falta de apoio do estado no que toca a reformas (cf. [3]). Para as camadas mais jovens, felizmente o número não é tão elevado. Por último, veremos a proporção de homens para mulheres.

Tabela 8: Cluster 2 - Frequências nos géneros

Género	Nº Ocorrências
Homens	2222
Mulheres	99

Neste cluster já podemos ver claras diferenças entre os homens e as mulheres, sendo o dos homens muito mais elevado.

Relativamente ao valor de variabilidade intra-cluster, este apresenta um valor de 5667.3.

6.3 Cluster 3

Para o terceiro cluster, percebemos que a população média é bastante mais elevada que nos outros. Relativamente ao número de suicídios, este é o segundo mais alto mas longe do que acontece no segundo *cluster*. O PIB deste grupo é cerca do dobro do segundo *cluster*. Veremos quais são os países que lideram neste.

Tabela 9: Cluster 3 - Países Frequentes

País	Nº Ocorrências
EUA	338
Brasil	275
Rússia	192
México	186
Japão	182

Analisando a tabela, percebemos que os países que constituem este cluster são bastante diferentes em termos culturais e económicos, embora todos sejam bastante grandes e com uma elevada população, podendo ter sido este o fator prevalente em agrupá-los. O outro possivelmente foi o PIB, visto alguns destes serem as maiores potências mundiais. Esta divergência nos países que o constituem pode estar relacionada com, na representação gráfica 3 vermos que o cluster 3 se alonga bastante para o lado direito do gráfico. Observemos agora o que acontece com as faixas etárias.

Tabela 10: Cluster 3 - Frequências nas várias faixas etárias

Faixa Etária	Nº Ocorrências
35-54	453
15-24	256
5-14	249
55-74	226
25-34	214
75+	31

Assim como acontece no primeiro cluster, a taxa de suicídios para todos os grupos etários é bastante semelhante excepto para as pessoas no final de vida, acima dos 75 anos. Por último, veremos a proporção de homens para mulheres.

Tabela 11: Cluster 3 - Frequências nos géneros

Género	Nº Ocorrências
Mulheres	760
Homens	669

Assim como no primeiro, temos que o número de suicídios nas mulheres é ligeiramente mais alto, mas não com a proporção que existia no segundo.

Relativamente ao valor de variabilidade intra-cluster, este apresenta um valor de 6442.6.

6.4 Cluster 4

Finalmente, veremos o quarto cluster. Pela tabela dos valores médios 2, percebemos que os países que constituem este *cluster* têm também um número baixo de habitantes embora o seu PIB e taxa de suicídios sejam os mais baixos de todos. Olharemos agora para os países constituintes.

Tabela 12: Cluster 4 - Países Frequentes

País	Nº Ocorrências
Colômbia	372
Equador	372
Malta	370
Maurícia	366
Chile	361

Analisando a tabela supramencionada, vemos que os países ou são na América do Sul ou em África (tendo como outlier Malta), como é o caso das ilhas Maurícias. Isto reflete-se no PIB mais baixo deste cluster. Relativamente ao número de suicídios, mesmo sabendo que o PIB é baixo, os suicídios também o são (cf. [4]). Isto pode estar relacionado com uma forte crença religiosa (cf. [5]) que condena fortemente esta escolha ao contrário dos países europeus, onde esse fator não entra tanto em jogo na grande maioria deles. Veremos agora as idades com mais ocorrências.

Tabela 13: Cluster 4 - Frequências nas várias faixas etárias

Faixa Etária	Nº Ocorrências
5-14	3606
15-24	3486
25-34	3359
55-74	3155
35-54	3098
75+	2918

Para este último *cluster*, existem muitos mais suicídios em idades mais novas do que mais avançadas o que vai contra o esperado. Observemos a proporção de homens para mulheres.

Tabela 14: Cluster 4 - Frequências nos géneros

Género	Nº Ocorrências
Mulheres	10810
Homens	8812

Relativamente à proporção dos suicídios para os dois géneros, vemos que para este *cluster*, ao contrário do que é pensado, o número é mais elevado para as mulheres que para os homens embora não seja com a drástica diferença que existia no *cluster* 2.

Relativamente ao valor de variabilidade intra-cluster, este apresenta um valor de 11013.2, sendo assim o mais elevado de entre todos os clusters. Como sabemos, este valor também tenderá a ser mais elevado consoante o número de exemplos que foram atribuídos ao cluster (que é o caso).

7 Conclusão

Olhando agora de um modo geral sobre os agrupamentos gerados pelo algoritmo *k-means* sobre os nossos dados, percebemos que este conseguiu fazer a **separação das observações de uma forma lógica** e até de fácil interpretação. Existiram alguns **esterótipos que foram provados**, por exemplo, no *cluster* em que a taxa de suicídios é mais alta, coincide precisamente com os países da europa de leste que normalmente são associados a maiores números de suicídios e com qualidade financeira relativamente abaixo da média. Ou olhando para este ponto do outro lado da equação, que os países com o PIB mais alto têm taxas de suicídio mais baixas, sendo estes maioritariamente da europa central que é sabida ser financeiramente mais estável.

No entanto, também houve **surpresas** como é o caso das diferenças entre géneros. Em apenas um dos *clusters* é que os homens foram o grupo prevalente, ao contrário do que se acharia. Em todos os outros os números foram **superiores nas mulheres** embora por uma margem baixa.

Com esta análise, foi possível perceber a capacidade de um algoritmo como o *k-means* permitir facilmente visualizar a **topologia dos dados** de modo a tirarmos conlcusões e analisarmos um conjunto de observações que nem sabemos bem como se irá agrupar. É precisamente nestes casos que o uso de técnicas de *clustering* entram e revelam ser úteis.

Referências

[1] Kassambara A. (2020). *K-Means Clustering in R: Algorithm and Practical Examples*
Acedido a 13/01/2022, em
https://www.datanovia.com/en/lessons/k-means-clustering-in-r-algorith-and-practical-examples/

[2] Eurostat. (2019, 7 de Novembro). *Which EU country has the happiest people?*
Acedido a 13/01/2022, em
https://ec.europa.eu/eurostat/web/products-eurostat-news/-/ddn-20191107-1

[3] Conejero I., Olié E., Courtet P., Calati R. (2018, 20 de Abril). *Suicide in older adults: current perspectives*
Acedido a 13/01/2022, em
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5916258/

[4] Botega N., Stefanello S. (2009, Março). *Suicide attempts in South and Central America*
Acedido a 13/01/2022, em
https://oxfordmedicine.com/view/10.1093/med/9780198570059.001.0001/med-9780198570059-chapter-18

[5] Pew Research Center (2014, 13 de Novembro). *Religion in Latin America*
Acedido a 13/01/2022, em
https://www.pewforum.org/2014/11/13/religion-in-latin-america/