Fundação Getúlio Vargas - Escola de Matemática Aplicada
Disciplina: Análise Exploratória de Dados e Visualização
Prof. Asla Medeiros e Sá

Visualizando surtos epidêmicos

Cristiana Aparecida Nogueira Couto

30 junho de 2020

Resumo

O presente relatório é produto do trabalho final para a disciplina de Análise Exploratória de Dados e Visualização e contêm algumas visualizações utilizando um conjunto de dados de um surto epidêmico de sarampo em Hagelloch, Alemanha, em 1861. Algumas imagens foram produzidas com o objetivo de responder perguntas sobre o pico de casos, quantidade de famílias afetadas, localização espacial das famílias, dentre outros.

1 Introdução

Nesse momento em que o mundo passa por uma pandemia, muitas visualizações têm sido feitas com diferentes objetivos. A mais famosa delas é provavelmente a chamada *flatten the curve* (achatar a curva). O objetivo desse gráfico é trazer uma justificativa para o isolamento, mostrar que permanecer em casa e obedecer as regras de distanciamento social são importantes para fazer com que o sistema de saúde possa suportar os novos casos de COVID-19. Algumas perguntas são naturais nesse contexto. Todos os dias aparecem novas especulações sobre quando será o pico da curva, quando as coisas voltarão ao normal, quando as pessoas poderão sair de casa, entre outros. Outras visualizações comuns são aquelas contabilizando os casos acumulados, incidência e quantidade de óbitos. Todas elas são muito importantes para contar a linha do tempo da doença e ajudar a entender melhor algumas perguntas epidemiológicas.
A partir disso, foi decidido investigar algum surto epidêmico que já aconteceu em alguma época da história para analisá-lo tendo o começo, o meio e o fim. Além disso, foi escolhido investigar uma situação de menor proporção, um surto em uma região específica, ao invés de uma pandemia, que têm proporções globais, devido ao tamanho menor da base de dados, o que torna as manipulações necessárias mais simples. A escolha do *dataset* refletiu a possibilidade de criar visualizações interessantes, com as quais seja possível explorar diferentes variáveis.

2 Dataset

A base de dados observada para o trabalho é um complicado de dados de surtos de doenças em diferentes partes do mundo em diferentes épocas. A base completa é um pacote do próprio *r* chamado *outbreaks* e reúne mais de 20 surtos indo desde sarampo até norovírus e influenza.
O dataset escolhido informa sobre um surto de sarampo que aconteceu em 1861, num distrito administrativo de Tübingen, na Alemanha. A escolha desse dataset em específico se deu pelo fato de ter opções de variáveis que vão além de início da doença, data de morte etc.
Para cada indivíduo cuja as informações foram coletadas, há a data de início da erupção cutânea e a data de pródromo. Na medicina, pródromo é um sinal que indica o início de uma doença antes que sintomas mais específicos apareçam. Há também variáveis para gênero, idade, identificação familiar, se houve complicações, localização, data de morte (caso tenha ocorrido devido a doença) e por quem o indivíduo foi possivelmente infectado.
Esse surto epidêmico de sarampo ocorreu numa pequena vila isolada. Diariamente, os dados das casas e nomes dos indivíduos afetados eram registrados, além de outras informações como complicações, aparecimento e desenvolvimento de sintomas e morte. Segundo [4], a vila era composta de 577 habitantes, os quais 200 deles eram crianças com até quinze de idade que haviam escapado o surto anterior de sarampo ou ainda não eram nascidos na época. Algumas crianças foram isoladas durante o surto ou eram imigrantes e haviam sido infectadas previamente, do total de suscetíveis, 188 crianças foram infectadas.
Esse conjunto de dados foi coletado por Albert Pfeilsticker para sua tese de doutorado (1863) e posteriormente os dados foram reanalisados por Heike Oesterle (1962)[1].

	case_ID	infector	date_of_prodrome	date_of_rash	date_of_death	age	gender	family_ID	class	complications	x_loc	y_loc
1	1	45	1861-11-21	1861-11-25	NA	7	f	41	1	yes	142.5	100.0
2	2	45	1861-11-23	1861-11-27	NA	6	f	41	1	yes	142.5	100.0
3	3	172	1861-11-28	1861-12-02	NA	4	f	41	0	yes	142.5	100.0
4	4	180	1861-11-27	1861-11-28	NA	13	m	61	2	yes	165.0	102.5
5	5	45	1861-11-22	1861-11-27	NA	8	f	42	1	yes	145.0	120.0
6	6	180	1861-11-26	1861-11-29	NA	12	m	42	2	yes	145.0	120.0
7	7	42	1861-11-24	1861-11-28	NA	6	m	26	0	yes	272.5	147.5
8	8	45	1861-11-21	1861-11-26	NA	10	m	44	1	yes	97.5	155.0
9	9	182	1861-11-26	1861-11-30	NA	13	m	44	2	yes	97.5	155.0
10	10	45	1861-11-21	1861-11-25	NA	7	f	29	1	yes	240.0	75.0
11	11	182	1861-11-25	1861-11-30	NA	11	f	27	2	yes	270.0	135.0
12	12	45	1861-11-20	1861-11-25	NA	7	f	32	1	yes	195.0	27.5
13	13	12	1861-11-30	1861-12-05	NA	13	m	32	2	yes	195.0	27.5
14	14	181	1861-11-22	1861-11-29	NA	13	f	22	2	yes	227.5	185.0
15	15	45	1861-11-24	1861-11-29	NA	8	m	22	1	yes	227.5	185.0
16	16	181	1861-11-21	1861-11-25	NA	15	f	43	2	yes	172.5	172.5

Figure 1: Tabela com o conjunto de dados que será trabalhado.

A seguir estão abordadas mais detalhadamente as informações sobre cada uma das colunas:

- *case_ID*: Número de identificação do caso (não está em ordem temporal).

- *infector*: Número do paciente que é a provável origem da infecção desse caso.
- *date_of_prodrome*: Data de início dos sintomas prodrômicos.
- *date_of_rash*: Date de início das erupções cutâneas.
- *date_of_death*: Data de morte (NA implica recuperação).
- *age*: Idade em ano.
- *gender*: Genêro do indivíduo (alguns casos não possuem essa informação).
- *family_ID*: Número de identificação da família.
- *class*: Classe escolar (0 é pré-escola; 1, primeira classe; 2, segunda classe).
- *complications*: Complicações, variável binária (sim ou não);
- *x_loc*: Coordenada x da casa (em metros). A escala em metros é obtida multiplicando a coordenada original por 2.5.
- *y_loc*: Coordenada y da casa (em metro).

3 Visualizações

3.1 Quando foi o pico de casos?

Na visualização da figura 2 temos os casos diários reportados de crianças com sintomas iniciais e os casos de surgimento de erupção cutânea que é um sintoma característico do sarampo. Note que o pico de registros de sintomas iniciais precede o pico de casos reportados de início da erupção cutânea. Isso também ocorre nos dias inicias, onde os primeiros casos de erupção cutânea aparecem cerca de uma semana depois dos primeiros casos de sintoma iniciais. Podemos sugerir que os dados em azul representam os casos suspeitos e em verde estão os casos confirmados. Além disso, no segundo pico em torno dos dias 3 de dezembro à 10 de dezembro, o tempo que os indivíduos previamente suspeitos levam para desenvolverem a erupção cutânea é bem mais curto, o que explicaria o pico mais alto em verde do que o pico em azul.
Considerando que os dados estão completos [4], parece haver uma segunda onda de sarampo durante o surto. A doença infectou um grupo primeiro e pareceu desaparecer, em seguida as infecções aumentaram novamente resultando em uma segunda onda de infecções. Com isso, também podemos responder que o surto durou cerca de três meses, indo do final de outubro de 1861 até a última semana de janeiro de 1862, afetando um total de 188 crianças.

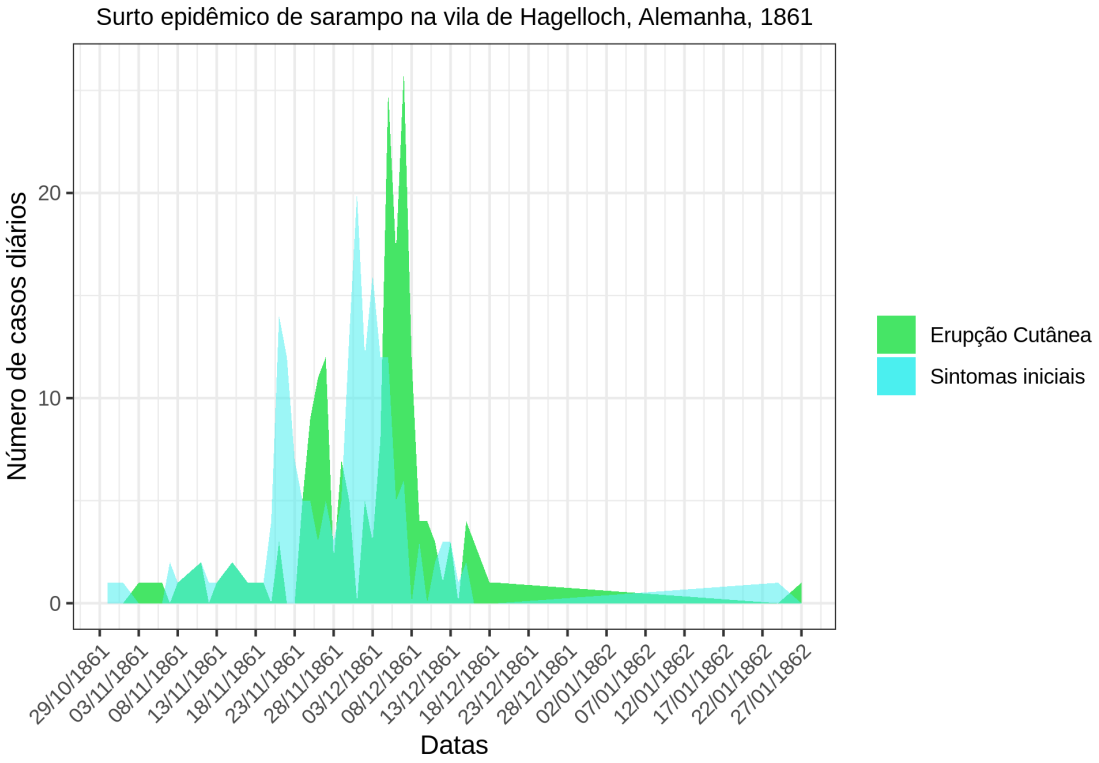

Figure 2: Gráfico comparando os registros diários de sintomas iniciais e início da erupção cutânea.

3.2 Para quantas pessoas cada indivíduo infectado transmitiu a doença?

O sarampo é uma doença infecto-contagiosa, ou seja, de fácil e alta transmissão, causada por um vírus do gênero *Morbillivirus*. Uma pessoa com sarampo pode infectar até 9 de 10 pessoas com as quais ela possui contato se elas não estiverem vacinadas [6]. No século XIX ainda não havia uma vacina disponível contra o sarampo, por isso identificar os indivíduos infectados e isolá-los o mais rapidamente possível era essencial para controlar o surto.
Na figura 3 temos a quantidade de pessoas infectadas por cada indivíduo. Caso um indivíduo infectado não tenha sido a possível fonte de nenhum outro caso ele não foi incluído na visualização. Foi optado pelo uso de um gráfico de pirulito, gráficos desse tipo são bastante utilizados quando num gráfico de barras há muitas colunas com a mesma altura, como é o caso dos dados de infectados que temos. A cor alaranjada e os círculos também fazem uma referência as "bolinhas" que surgem no corpo como sintoma da doença.

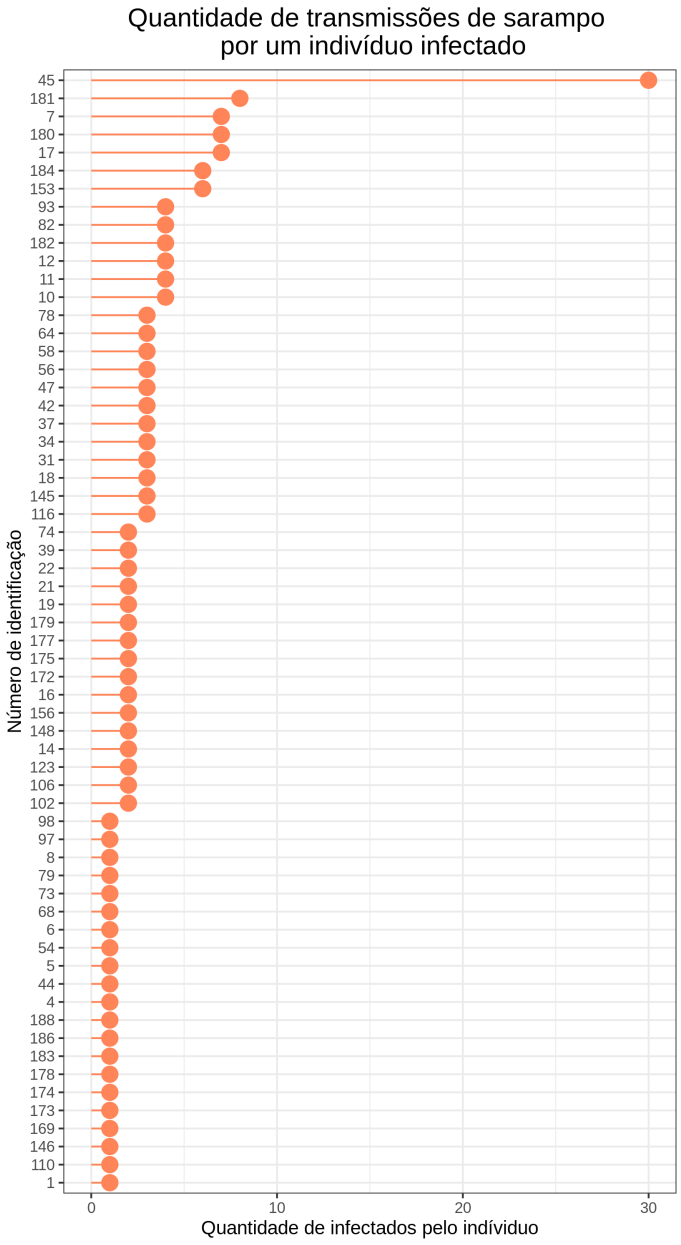

Figure 3: Visualização com o número de possíveis pessoas infectadas por cada indivíduo.

3.3 Quem foi infectado por quem?

A paleta de cores foi escolhida com o auxilio do site https://vis4.net/palettes que checa se ela é sensível para portadores de daltonismo. Apesar das classes serem sequenciais, foi escolhida uma paleta divergente de modo a ressaltar a diferença entre as classes. A mesma paleta de cores foi utilizada para as demais visualizações.
Com a visualização a seguir queremos responder a pergunta relacionada a origem da infecção de cada caso. A transmissão da doença se dá em um ambiente escolar. Há três classes, *0, 1* e *2*. Feita a diferenciação por cores podemos analisar o espalhamento da doença entre as classes.

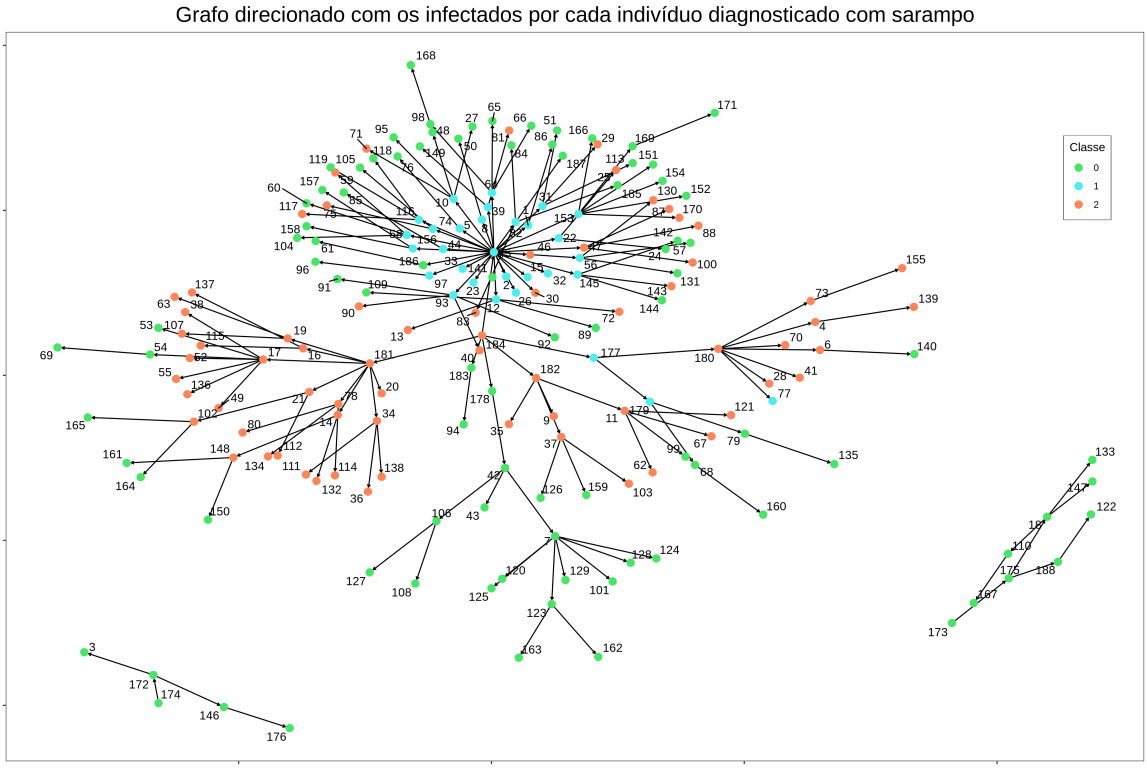

Figure 4: Grafo de origem das transmissões.

A partir desse grafo podemos supor que o individuo cujo número de identificação é 45 é um *super-spreader*, ou super-espalhador. Esse termo se refere a indivíduos que infectam desproporcionalmente outros indivíduos suscetíveis em comparação com outros que infectam poucos ou nenhum indivíduo. Saber identificar super-spreaders e investigar quais fatores o tornam mais propicio a espalhar uma infecção faz parte do conjunto de medidas para controlar uma doença infecciosa. Nesse surto epidêmico de sarampo, o caso 45 gerou 16%

dos casos. Observe ainda que se contarmos os casos gerados pelos indivíduos infectados pelo 45 essa porcentagem passa a ser de 46%.
Devido a quantidade de casos gerados pelo 45 a visualização dos casos gerados diretamente por ele fica comprometida, com isso a visualização com um grafo com arestas em formato de arco pode ser utilizadas para ressaltar esses casos. No eixo x estão os números de identificação dos casos. Todas as arestas saem de 45 e vão até os IDs dos individuos infectados.
Com o ID dos casos cronológico poderiamos explorar essa visualização com uma variável de tempo implicíta no eixo x. Assim, poderiamos supor em qual momento ocorreu a transmissão, dado que um indíduo infectado com sarampo leva de 8 a 13 dias para apresentar os primeiros sintomas [2].

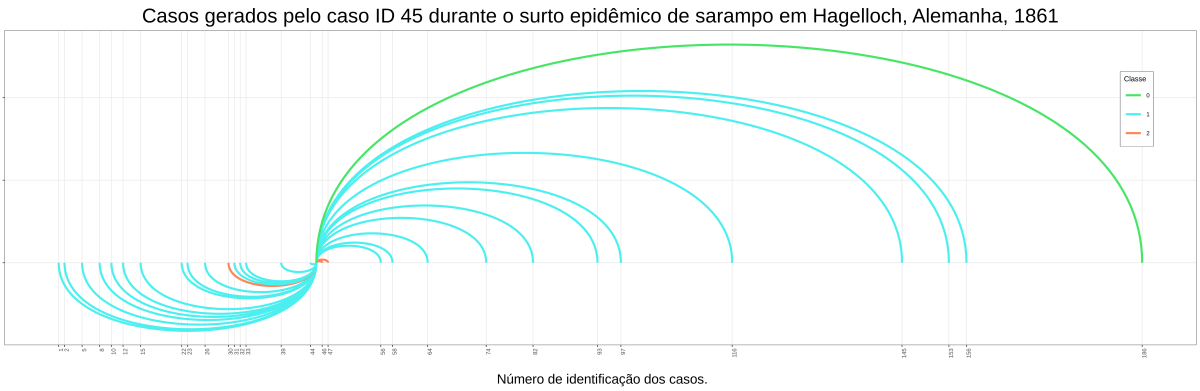

Figure 5: Possíveis transmissões do caso 45.

3.4 Qual a localização dos infectados?

Essa pergunta surge a partir da observação das colunas *x_loc* e *y_loc* contidas na tabela de dadas. Os dados de localização das casas das famílias afetadas já foram utilizados em modelos estatísticos em epidemiologia espacial [4]. Como indivíduos que são da mesma família possuem a mesma localização podemos responder com essa visualização quantas famílias foram afetadas pelo surto de sarampo. Os dados têm registradas 69 famílias, onde algumas delas possuem a mesma localização como é possível ver no gráfico. Podemos supor que essas famílias moravam juntas, ou pelo menos na mesma construção.

4 Conclusão

Utilizando esse *dataset* foi possível explorar algumas visualizações interessantes e levantar questionamentos sobre o surto epidêmico de sarampo numa pequena vila Alemã e sobre o papel da visualização em retratar a evolução de doenças infecciosas. Por fim, vale ressaltar que as visualizações foram feitas com um

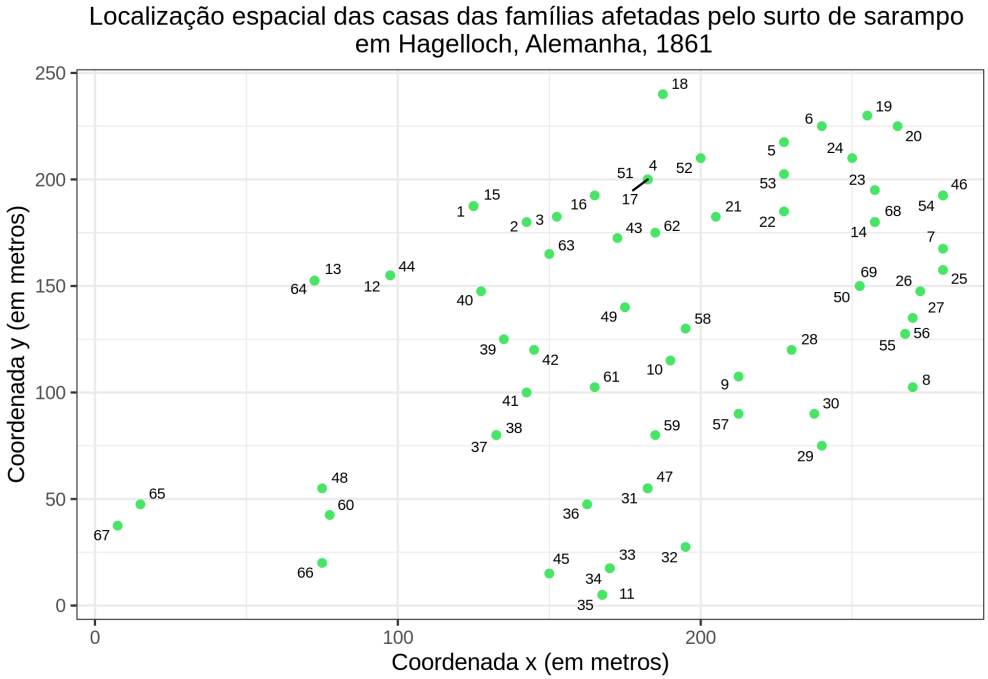

Figure 6: Localização das famílias.

caráter educativo e as suposições são baseadas apenas nas imagens produzidas. Qualquer afirmação concreta deve ser feita em conjunto com um(a) especialista da área de epidemiologia.

Referências

[1] Measles in hagelloch, germany, 1861. `http://www.repidemicsconsortium.org/outbreaks/reference/measles_hagelloch_1861.html`. [Online; Acessado em: 07/06/2020].

[2] O sarampo. `https://portal.fiocruz.br/noticia/o-sarampo`. [Online; Acessado em: 27/06/2020].

[3] Robert Kosara. The visual evolution of the "flattening the curve" information graphic. `https://eagereyes.org/blog/2020/the-visual-evolution-of-the-flattening-the-curve-information-graphic`. [Online; Acessado em: 08/06/2020].

[4] A.B. Lawson. *Statistical Methods in Spatial Epidemiology*. Wiley Series in Probability and Statistics. Wiley, 2013.

[5] Amanda Makulec. Ten considerations before you create another chart about covid-19. https://medium.com/nightingale/ten-considerations-before-you-create-another-chart-about-covid-19-27d3bd691be8. [Online; Acessado em: 08/06/2020].

[6] Mayra Malavé. O ressurgimento do sarampo: uma doença evitável. https://portal.fiocruz.br/noticia/o-ressurgimento-do-sarampo-uma-doenca-evitavel. [Online; Acessado em: 08/06/2020].

[7] Richard A. Stein. Super-spreaders in infectious diseases. *International Journal of Infectious Diseases*, 15(8):e510 – e513, 2011.