

UNIVERSIDADE FEDERAL DE JUIZ DE FORA
INSTITUTO DE CIÊNCIAS EXATAS
DEPARTAMENTO DE ESTATÍSTICA

Jéssica Vianelo Sell

**MODELO PARA DADOS LONGITUDINAIS DE CONTAGEM: UMA APLICAÇÃO
PARA DADOS DA COVID-19**

JUIZ DE FORA

2021

Jéssica Vianelo Sell

MODELO PARA DADOS LONGITUDINAIS DE CONTAGEM: UMA APLICAÇÃO
PARA DADOS DA COVID-19

Trabalho de Conclusão de Curso
apresentado ao Departamento de
Estatística da Universidade Federal de
Juiz de Fora, como requisito básico para
Conclusão do Curso de Estatística.

Orientador: Prof. Ph.D. Marcel de Toledo Vieira

JUIZ DE FORA

2021

Vianelo Sell, Jéssica.

Modelos para Dados Longitudinais de Contagem: uma Aplicação para Dados da Covid-19 / Jéssica Vianelo Sell. -- 2021.

96 p. : il.

Orientador: Marcel de Toledo Vieira

Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Juiz de Fora, Instituto de Ciências Exatas, 2021.

1. Óbitos. 2. Pandemia. 3. Análise de Dados Longitudinais. 4. Regressão de Poisson. I. de Toledo Vieira , Marcel, orient. II. Título.

Jéssica Vianelo Sell

MODELO PARA DADOS LONGITUDINAIS DE CONTAGEM: UMA APLICAÇÃO
PARA DADOS DA COVID-19

Trabalho de Conclusão de Curso
apresentado ao Departamento de
Estatística da Universidade Federal de
Juiz de Fora, como requisito básico para
Conclusão do Curso de Estatística.

BANCA EXAMINADORA

Prof. Ph.D. Marcel de Toledo Vieira - UFJF

Prof. Ph.D. Ronaldo Rocha Bastos - UFJF

Prof. D.Sc. Augusto Carvalho Souza - UFJF

AGRADECIMENTOS

Agradeço primeiro a Deus, que fez com que meus objetivos fossem alcançados durante esses anos.

Agradeço à minha mãe Gilmara, que sempre me apoiou e me incentivou a estudar, que é meu exemplo de mulher e sempre venceu todas as dificuldades. À minha vó Darci, que me apoiou desde o começo da faculdade. Ao meu padraсто Alexandre, pelo apoio com os obstáculos encontrados. Ao meu namorado Sávio, pelo apoio em todos os momentos. Agradeço as minhas tias Simone e Raquel, e aos meus tios José Eduardo e Marcelo, por me acolherem no período da faculdade, sem vocês esta jornada não seria possível. A toda a minha família por estar ao meu lado, e me incentivarem sempre.

Agradeço aos amigos que a faculdade me deu. Sem vocês esta jornada não seria a mesma, os jogos não seriam os mesmos, assim como as risadas.

Agradeço ao meu orientador Marcel, pela ajuda e auxílio no desenvolvimento desde trabalho, e pelos ensinamentos em sala. Aos meus professores do departamento pelo aprendizado, sempre à disposição para qualquer eventual dúvida, e pela paciência quando o desespero batia a porta.

À Universidade Federal de Juiz de Fora, por me acolher e proporcionar a oportunidade de estudo em uma instituição de ensino superior. Agradeço também à Plataforma JF Salvando Todos, pela disponibilização dos dados para esta monografia.

É com muita alegria que fecho mais um ciclo da minha vida. E que venha novas aventuras.

RESUMO

Desde dezembro de 2019, quando foi confirmado o primeiro caso de Covid-19 na cidade de Wuhan, China, muitos estudos sobre o novo coronavírus estão sendo produzidos. O presente trabalho tem a intenção de estudar o comportamento do número de casos de óbitos em cidades brasileiras com população entre 200 mil e 1 milhão de habitantes de acordo com as projeções do IBGE para 2020. A partir da Análise de Dados de Longitudinais e aplicação de um Modelo de Poisson para dados Longitudinais, foi analisada a relação entre o número de óbitos pela Covid-19 na população e um conjunto de variáveis explicativas, que são o número de casos confirmados, mobilidade social, R_t estimado, número de leitos e respiradores disponíveis. Neste trabalho, consideramos como variável tempo as semanas epidemiológicas, que começam da 1ª semana de 2020 até a 26ª semana epidemiológica 2021. Os resultados indicam que o R_t estimado, considerando os valores de 5 semanas atrás, é a variável preditora que tem mais influência no número médio de óbitos. E como já era esperado, os casos confirmados possui uma relação positiva com o número médio de vidas perdidas. Leitos Complementares, que são os leitos de UTI adulto e pediátrico, não tiveram influência significativa no número de óbitos. É verificado também, que a categoria de mobilidade social Residencial, possui uma relação positiva com os casos de óbitos, isso indica que quando o número de óbitos está elevado, as pessoas se sentem mais motivadas a permanecerem em casa.

Palavras-chave: Óbitos, Pandemia, Análise de Dados Longitudinais, Regressão de Poisson.

ABSTRACT

Since December 2019, when the first case of Covid-19 was confirmed in the city of Wuhan, China, many studies on the new coronavirus have been produced. The present work intends to study the behavior of the number of deaths in Brazilian cities with a population between 200 thousand and 1 million inhabitants according to IBGE projections for 2020. Based on the longitudinal data analysis and an application of a Poisson model for longitudinal data, we analyzed the relationship between the number of deaths by Covid-19 in the population and a set of explanatory variables, which are the number of confirmed cases, social mobility measures, estimated R_t , number of beds and available respirators. In this work, we considered as a time variable the epidemiological weeks, which start from the 1st week of 2020 to the 26th epidemiological week 2021. The results indicate that the estimated R_t , considering values from 5 weeks ago, is the predictor variable that has the most influence on the average number of deaths. And as expected, confirmed cases have a positive relationship with the average number of lives lost. Complementary beds, which are adult and pediatric ICU beds, had no significant influence on the number of deaths. It is also verified that the category of residential social mobility has a positive relationship with the cases of deaths, which indicates that when the number of deaths is high, people feel more motivated to stay at home.

Keywords: Deaths, Pandemic, Longitudinal Data Analysis, Poisson Regression.

LISTA DE TABELAS E QUADROS

Quadro 2.1. Indicadores de mudanças em dois tempos de J tempos medidos.....	23
Tabela 2.1. Função de Ligação Canônica para as Distribuições univariadas mais comuns.....	17
Tabela 3.1. Correlação no tempo j_i	28
Tabela 4.1. Apresentação e explicação das variáveis.....	29
Tabela 5.1 Coeficientes do Modelo.....	36
Tabela A1 Cidades estudadas e sua respectiva população.....	40

LISTA DE GRÁFICOS

Figura 4.1 – Gráfico de Perfil de casos acumulados de 10 municípios sorteados por AAS.....	30
Figura 4.2 – Gráfico de Perfil de casos acumulados com enfoque nos municípios de Itapevi e Cascavel.....	31
Figura 4.3 – Gráfico de Perfil de óbitos acumulados de 10 municípios sorteados por AAS.....	31
Figura 4.4 – Gráfico de Perfil de óbitos acumulados com enfoque nos municípios de Itapevi e Cascavel.....	32
Figura 4.5 – Box Plots de casos confirmados, óbitos, leitos de internação e complementar, respiradores em uso e estabelecimentos com respiradores.....	33
Figura 4.6 – Box Plots das categorias de mobilidade social.....	34
Figura 4.7 – Rt estimado.....	34
Figura 4.8 – Rt estimado modificado.....	34

SUMÁRIO

1. INTRODUÇÃO.....	11
2. REVISÃO NO CONTEXTO DOS DADOS TRANVERSAIS.....	13
2.1 Distribuição de Poisson.....	13
2.2 Modelo Lineares Generalizados para Dados Transversais.....	16
2.2.1 Família Exponencial.....	18
2.3 Modelo de Regressão Poisson para Dados Transversais.....	18
3. MODELOS PARA DADOS LONGITUDINAIS DE CONTAGENS.....	21
3.1 Estudos Longitudinais.....	22
3.2 Modelos para Dados Longitudinais.....	22
3.2.1 Distribuição de Poisson para Dados Longitudinais.....	24
3.2.2 Modelos Lineares Generalizados Mistos.....	25
3.2.3 Modelo de Regressão Poisson para Dados Longitudinais.....	26
4. BANCO DE DADOS.....	28
4.1 Apresentação e Explicação dos Dados.....	28
4.2 Análise Exploratória dos Dados.....	29
5. APLICAÇÕES E RESULTADOS.....	35
6. CONCLUSÃO.....	37
7. REFERÊNCIAS BIBLIOGRÁFICAS.....	38
ANEXOS.....	40
Anexo 1 – Cidades Estudadas.....	40
Anexo 2 – Modelo utilizando a função “ <i>wbm</i> ” implementado no software <i>R</i>	43
Anexo 3 – Script <i>R</i>	43

1. INTRODUÇÃO

Ao longo dos anos, surgiram na natureza diferentes vírus que se espalharam por todo o planeta, cada um com sua particularidade na transmissão e nos sintomas. Alguns são mais nocivos que outros, como por exemplo, o *Ebola*, que foi identificado pela primeira vez em 1976 na África Central. Sua transmissão é feita pelo contato de uma pessoa com sangue ou secreções infectadas pelos *Filovirus*, membros da família *Filoviridae* (JACOB; et. Al., 2020). Outro exemplo é o vírus da *Dengue*, que tem sua transmissão principalmente por meio da picada da fêmea do mosquito *Aedes Aegypti*. No Brasil, a primeira epidemia da Dengue ocorreu entre 1981 e 1982, no estado de Roraima (DIAS; et. Al., 2010).

Em dezembro de 2019, na cidade de Wuhan, China, algumas pessoas que tiveram contato com o mercado de Wuhan, conhecido por vender animais vivos, começaram a apresentar alguns sintomas de uma pneumonia de causa desconhecida. Após uma investigação do Centro Chinês para Controle e Prevenção de Doenças (*China CDC*), houve a identificação de um novo coronavírus nesses pacientes (HUANG; et. Al., 2020).

O novo coronavírus foi denominado de *sars-cov-2*, pois pertence à família coronavírus que, atualmente, conta com 6 membros, sendo que quatro destes levam a apenas resfriados, mas os outros dois podem causar sérios problemas respiratórios (BELASCO; FONSECA, 2020). A doença causada pelo vírus *sars-cov-2* é conhecida como Covid-19 (COVID significa Corona Vírus Disease – Doença do Coronavírus, e 19 pois teve seus primeiros casos confirmados no ano de 2019), e tem sua transmissão quando uma pessoa tem contato com gotículas expelidas por meio da tosse ou espirro, por outra pessoa que está contaminada.

De acordo com o artigo publicado por Huang et. Al (2020), pela revista científica *The Lancet*, o primeiro caso de COVID-19 no mundo foi confirmado no dia 1º de dezembro de 2019. Já no dia 11 de janeiro, foi registrada a primeira morte causada pelo novo coronavírus, na cidade de Wuhan, na China. O Comitê de Emergência da OMS declarou, no dia 31 de janeiro de 2020, emergência de saúde global com base na crescente taxa de contaminação do vírus, tanto na China quanto em outros países.

No Brasil, o primeiro caso de COVID-19 foi registrado no estado de São Paulo, no dia 26 de fevereiro de 2020, onde também foi registrado o primeiro óbito em 17 de março (JF SALVANDO TODOS, 2020).

Segundo a JF Salvando Todos (2020), a partir de dados do Ministério da Saúde, o maior registro de casos confirmados no Brasil, aconteceu no dia 23 de junho de 2021, chegando a 115.228 infectados pelo novo coronavírus. O primeiro pico da doença, tinha sido registrado em dia 29 de julho de 2020, com 69.074 casos confirmados.

O primeiro pico no número de vidas perdidas no país aconteceu no dia 29 de julho de 2020, com o registro de 1.595 óbitos em 24 horas. Já o segundo pico ocorreu no dia 8 de abril de 2021, que chegou a 4.249 óbitos registrados em apenas um dia.

Em Juiz de Fora, Minas Gerais, o primeiro caso confirmado foi de um homem, com histórico de viagem aos Estados Unidos, que foi confirmado por exame no dia 14 de março (LANDIM, 2020). A primeira morte também foi de um homem, e ocorreu no dia 7 de abril de acordo com a JF Salvando Todos (2020) a partir de dados da Prefeitura de Juiz de Fora. Desde novembro de 2020, os números de casos confirmados em Juiz de Fora vêm se mantendo em patamar elevado, chegando a registrar 650 novos casos apenas no dia 29 de abril de 2021. Antes disso, o maior registro tinha sido no dia 16 de julho de 2020, com 215 novos casos.

Os números de vidas perdidas para a COVID-19 em Juiz de Fora vinham apresentando pouca variação na maior parte do ano de 2020. Porém, a partir do mês de novembro de 2020, os registros de óbitos tiveram uma alta substancial chegando a 23 vidas perdidas apenas no dia 5 de janeiro de 2021. Já no dia 29 de março de 2021, houve o maior registro de vidas perdidas desde o início da pandemia em Juiz de Fora, com 42 óbitos. Logo depois desse pico, os casos de óbitos estão caindo no município.

Os dados que iremos utilizar neste trabalho são longitudinais e provenientes de cidades brasileiras com população entre 200 mil e 1 milhão de habitantes de acordo com as projeções do IBGE para 2020, disponibilizadas pela plataforma JF Salvando Todos (2020).

Nas últimas décadas, segundo Vieira (2012), tornou-se cada vez mais possível a geração e o armazenamento de grandes bases de dados, valorizando os estudos longitudinais. Ao se coletar dados dos mesmos indivíduos ou casos em diferentes momentos no tempo, contendo ou não as mesmas variáveis, teremos as pesquisas longitudinais. Estes tipos de pesquisa são importantes quando se deseja observar mudanças em certo tipo de população ou indivíduo ao longo do tempo. Também podemos, segundo Vieira (2012, p.9), distinguir “entre o grau de variação na variável resposta para um indivíduo ao longo do tempo e a variação entre diferentes indivíduos”.

Uma dificuldade que este tipo de pesquisa apresenta, é o alto custo. Já que é preciso acompanhar os mesmos indivíduos ou casos em vários momentos no tempo. Porém, neste

presente trabalho, isto não foi uma preocupação, já que os dados foram disponibilizados pela plataforma JF Salvando Todos (2020).

O objetivo desta monografia é estudar como o número de casos confirmados, as taxas de mobilidade social da população, o número de reprodução efetivo (R_t), número de leitos de internação, de Unidade de Tratamento Intensivo (UTI's) e de respiradores influenciam no número de óbitos, seja de forma positiva ou negativa. Além de verificar qual variável explicativa que mais influência a média de óbitos.

Nas próximas seções, são apresentadas uma metodologia para dados transversais na Seção 2; metodologia para dados longitudinais na Seção 3; resultados de uma análise exploratória dos dados, para se ter um entendimento sobre o comportamento das variáveis e da relação entre elas na Seção 4; assim como os resultados de um exercício de modelagem com o ajuste de um modelo para dados longitudinais de contagem na mesma seção; e algumas considerações finais a partir das análises feitas na Seção 5. Para análises e ajuste do modelo foi utilizado o software R (Ihaka e Gentleman, 1996).

2. REVISÃO NO CONTEXTO DOS DADOS TRANSVERSAIS

Nesta seção, apresentamos uma revisão da literatura sobre modelos para dados transversais. Os dados transversais são dados coletados em um determinado ponto no tempo (BUCHALLA; CARDOSO, 2005), enquanto no contexto longitudinal, os dados são coletados ao longo do tempo.

A seguir, na subseção 2.1 é apresentada a Distribuição de Poisson. Em seguida, na subseção 2.2 apresentamos sobre Modelos Lineares Generalizados para Dados Transversais. Na subseção 2.2.1 discorremos sobre Família Exponencial, e por fim, na subseção 2.3 apresentamos sobre o Modelo de Regressão Poisson para Dados Transversais.

2.1 Distribuição de Poisson

Os dados utilizamos neste trabalho são dados de contagem. Assim, temos que as variáveis podem assumir qualquer valor inteiro não negativo, como é o caso do número de óbitos por Covid-19 denotada por Y . Sendo assim, podemos dizer que a variável Y segue uma Distribuição de Poisson.

Segundo Casella e Berger (2010), a distribuição de Poisson possui um único parâmetro, denotado aqui por μ , e sua função de probabilidade é definida por

$$P(Y = y|\mu) = \frac{e^{-\mu}\mu^y}{y!}. \quad (2.1)$$

A variável aleatória Y possui apenas números inteiros não negativos, e chamaremos de Suporte da Distribuição de Y , o conjunto de valores que Y pode assumir, que é denotado por $A(y)$. Assim, temos

$$A(y) = \{y: 0,1,2,\dots\}.$$

Esta distribuição possui valor esperado e variância iguais a μ , em que μ indica o número esperado de ocorrências de certo evento por unidade de tempo, espaço ou outra medida de grandeza (CASELLA; BERGER, 2010). Sendo assim,

$$E(y) = \mu \text{ e } Var(y) = \mu.$$

O espaço paramétrico de μ , denotado por Θ , é o conjunto de todos os valores que o parâmetro μ pode assumir. Desta forma,

$$\Theta = \{\mu: 0 \leq \mu < \infty\}.$$

Resultado 2.1. A partir de Casella e Berger (2010), podemos chegar no valor esperado de Y da seguinte maneira

$$\begin{aligned} E(Y) &= \sum_{y=0}^{\infty} y \frac{e^{-\mu} \mu^y}{y!} \\ &= \mu e^{-\mu} \sum_{y=0}^{\infty} \frac{\mu^{y-1}}{(y-1)!}. \end{aligned}$$

Fazendo a substituição de $x = y-1$, temos

$$= \mu e^{-\mu} \sum_{x=1}^{\infty} \frac{\mu^x}{x!}, \text{ e}$$

sabendo que a expansão da série de Taylor é dada por

$$e^{\mu} = \sum_{x=1}^{\infty} \frac{\mu^x}{x!},$$

teremos por fim

$$E(Y) = \mu e^{(-\mu+\mu)}$$

$$E(Y) = \mu.$$

Resultado 2.2. A variância é dada pela seguinte equação

$$Var(Y) = E(Y^2) - [E(Y)]^2, \quad (2.2)$$

em que $E(Y^2) = \sum_y y^2 p(y)$.

Como estamos utilizando a Distribuição de Poisson, $E(Y^2)$ pode ser escrito da seguinte forma,

$$E(Y^2) = \sum_{y=0}^{\infty} y^2 \frac{e^{-\mu} \mu^y}{y!}$$

$$= \sum_{y=0}^{\infty} y \frac{e^{-\mu} \mu^y}{(y-1)!}. \quad (2.3)$$

Como considerado no Resultado 2.1, substituindo $x = y - 1$ na equação 2.3, temos

$$E(Y^2) = \sum_{x=1}^{\infty} (x+1) \frac{e^{-\mu} \mu^{(x+1)}}{x!}$$

$$= \sum_{x=1}^{\infty} x \frac{e^{-\mu} \mu^{(x+1)}}{x!} + \sum_{x=1}^{\infty} \frac{e^{-\mu} \mu^{(x+1)}}{x!}$$

$$= \left(\mu e^{-\mu} \sum_{x=1}^{\infty} x \frac{\mu^x}{x!} \right) + \left(\mu e^{-\mu} \sum_{x=1}^{\infty} \frac{\mu^x}{x!} \right). \quad (2.4)$$

Dividindo a equação 2.4 em duas equações, temos

- i. $\left(\mu e^{-\mu} \sum_{x=1}^{\infty} x \frac{\mu^x}{x!} \right)$, sabendo que

$$\left(\mu e^{-\mu} \sum_{x=1}^{\infty} x \frac{\mu^x}{x!}\right) = \left(\mu^2 e^{-\mu} \sum_{x=1}^{\infty} \frac{\mu^{(x-1)}}{(x-1)!}\right),$$

em que aplicando a expansão de Taylor, temos

$$\mu e^{-\mu} \sum_{x=1}^{\infty} x \frac{\mu^x}{x!} = \mu^2.$$

ii. $\left(\mu e^{-\mu} \sum_{x=1}^{\infty} \frac{\mu^x}{x!}\right),$

em que aplicando a expansão de Taylor, verificamos que

$$\mu e^{-\mu} \sum_{x=1}^{\infty} \frac{\mu^x}{x!} = \mu.$$

Assim, podemos substituir os valores encontrados acima na equação 2.4 obtendo o seguinte resultado,

$$E(Y^2) = \left(\mu e^{-\mu} \sum_{x=1}^{\infty} x \frac{\mu^x}{x!}\right) + \left(\mu e^{-\mu} \sum_{x=1}^{\infty} \frac{\mu^x}{x!}\right) = \mu^2 + \mu.$$

Portanto, substituindo os valores na equação 2.2, temos

$$\begin{aligned} Var(Y) &= (\mu^2 + \mu) - \mu^2 \\ Var(Y) &= \mu. \end{aligned}$$

2.2 Modelos Lineares Generalizados para Dados Transversais

Na modelagem estatística de dados desejamos encontrar o modelo que melhor se ajuste aos dados. E, apesar do Modelo Normal Linear ser um importante modelo, ele nem sempre consegue se ajustar a certos tipos de dados. Com isso, Nelder e Wedderburn (1972) propuseram os Modelos Lineares Generalizados (MLG's), unificando, não só Modelo Normal Linear, mas também o Logístico, Probit, entre outros. Os MLG's têm como ideia principal permitir que a variável resposta seja explicada levando-se em consideração alternativas que permitam a

consideração da sua distribuição, para o caso daquelas que pertencem à família exponencial. Como exemplo, temos as distribuições Normal, Poisson, Binomial, Gamma e Normal Inversa.

Os MLG's são constituídos por três componentes:

- *Componente aleatório*: é a variável resposta, que tem distribuição pertencente à família exponencial;
- *Componente sistemático*: é a combinação das variáveis explicativas combinadas linearmente com os parâmetros desconhecidos β , conhecido como preditor linear η ;
- *Função de Ligação*: é uma função que liga os componentes aleatórios e sistemáticos ($g(\cdot)$), podendo ser logarítmica para os modelos log-lineares, por exemplo.

A seguir, são apresentadas na Tabela 2.1 as ligações canônicas de algumas distribuições que pertencem à Família Exponencial.

Tabela 2.1. Função de Ligação Canônica para as Distribuições univariadas mais comuns

Distribuição	Ligação Canônica $g(\cdot)$	Nome da Ligação Canônica
Normal	$\eta = \mu$	Identidade
Poisson	$\eta = \log\{\mu\}$	Log
Binomial	$\eta = \log\left\{\frac{\mu}{1-\mu}\right\}$	Logit
Gamma	$\eta = \mu^{-1}$	Recíproca
Normal Inversa	$\eta = \mu^{-2}$	Quadrática

Fonte: Adaptação de Cordeiro e Demétrio (2008)

Idealmente, segundo Cordeiro e Demétrio (2008), o componente sistemático deve ser estabelecido durante a fase de planejamento do experimento. Já o componente aleatório é especificado assim que as medidas a serem realizadas são definidas.

A função de ligação mais simples dos modelos lineares clássicos, é aquela em que o preditor linear η é igual à média μ , chamada de função identidade que assume qualquer valor real para a variável dependente. Porém, como estamos trabalhando com dados de contagem com Distribuição Poisson, temos que a função identidade não pode ser utilizada, já que o parâmetro μ é maior ou igual a 0. Portanto, iremos utilizar a relação $\eta = \ln(\mu)$.

Segundo Nelder e Wedderburn (1972), o modelo base para o estudo da relação entre variáveis tem a forma

$$g(\mu) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p,$$

em que,

x_i , com $i = 1, 2, \dots, p$, são as variáveis explicativas com valores conhecidos;

β_i , com $i = 1, 2, \dots, p$, são parâmetros desconhecidos chamados de coeficientes de regressão.

2.2.1 Família Exponencial

No final do século XIX, os modelos para dados pertencentes à família exponencial foram desenvolvidos por Maxwell, Boltzmann e Gibbs (CORDEIRO; DEMÉTRIO, 2011). Porém o conceito de família exponencial foi introduzido na Estatística por Fisher. Segundo Cordeiro e Demétrio (2008), a família exponencial uniparamétrica, tem a seguinte função de probabilidade

$$f(x; \mu) = h(x) \exp[\eta(\mu) t(x) - b(\mu)],$$

em que, $\eta(\mu)$, $b(\mu)$, $t(x)$ e $h(x)$ apresentam valores em subconjuntos dos reais.

O parâmetro da Distribuição de Poisson tem suporte nos reais não negativos. Assim, a função de probabilidade desta distribuição pode ser escrita da seguinte maneira,

$$f(x; \mu) = \frac{e^{-\mu} \mu^x}{x!} = \frac{1}{x!} \exp[x \log(\mu) - \mu]. \quad (2.5)$$

Em 2.5, podemos ver que a função de probabilidade da Distribuição Poisson (Equação 2.1) pode ser escrita na forma da Família Exponencial. Portanto, temos que $\eta(\mu) = \log(\mu)$, $b(\mu) = \mu$, $t(x) = x$ e $h(x) = 1/x!$. Confirmando que a Distribuição Poisson pertence à Família Exponencial.

2.3 Modelo de Regressão Poisson para Dados Transversais

Poisson publicou, em 1837, a distribuição que levou seu nome. Assim, a variável aleatória $Y \sim \text{Poisson}, P(\mu)$, com $\mu > 0$, que é uma distribuição limite da Distribuição Binomial (CORDEIRO; DEMÉTRIO, 2011).

A Distribuição de Poisson tem um papel importante na análise de dados de contagem. Cordeiro e Demétrio (2008) mostram que as qualidades principais desse modelo são a boa descrição dos dados, sobretudo quando variância é proporcional a média, e, se seus eventos

ocorrem independentes e aleatoriamente no tempo, o modelo consegue determinar o número de eventos em um intervalo específico de tempo.

Como o parâmetro μ descreve tanto a média quanto a variância, podemos ter alguns casos em que a variância dos dados é maior do que a média, o que caracteriza um problema denominado superdispersão dos dados. Ao utilizar apenas um parâmetro, também podemos não estar captando toda a variabilidade dos dados estudados.

Uma solução possível quando se depara com a superdispersão dos dados, é o uso do modelo Binomial Negativa, já que também é usado para dados de contagem e possui suporte contido nos números inteiros positivos. Este modelo possui dois parâmetros, um para a média e outro para a variância. Apesar do modelo Binomial Negativa parecer melhor para explicar dados de contagem, pois supre as necessidades quando se tem superdispersão dos dados, uma desvantagem é dificuldade com a interpretação dos seus coeficientes, que pode ser muito trabalhosa.

Neste trabalho, optamos pela adoção do Modelo de Poisson. Assim, como foi dito na seção 2.2, para o caso transversal, temos os seguintes componentes de um MLG,

1. Componente Aleatório: Distribuição de Poisson $P(\mu)$;
2. Componente Sistemático: Combinação linear dos parâmetros desconhecido com as variáveis regressoras $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$, sendo $\boldsymbol{\eta}$ o vetor de dimensões $n \times 1$, $\boldsymbol{x}'_i = [x_{i1}, x_{i2}, \dots, x_{ip}]$, $i = 1, 2, \dots, n$, e $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)'$, $p < n$, é um vetor de $1 + p$ parâmetros desconhecidos a serem estimados, com

$$\mathbf{X} = \begin{bmatrix} \boldsymbol{x}'_1 \\ \boldsymbol{x}'_2 \\ \dots \\ \boldsymbol{x}'_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

sendo definido como uma matriz de covariáveis de dimensões $n \times p$. Além disso,

3. Função de Ligação: Log-linear $g(\mu) = \ln(\mu)$.

Utilizando a função de ligação log-linear, temos

$$\ln(\mu) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p. \quad (2.6)$$

Portanto, o modelo final ficaria da forma

$$\mu = e^{\beta_0} (e^{\beta_1})^{x_1} (e^{\beta_2})^{x_2} \dots (e^{\beta_p})^{x_p}, \quad (2.7)$$

Se utilizarmos apenas uma variável regressora na equação 2.7, teríamos

$$\mu = e^{\beta_0} (e^{\beta_1})^{x_1}.$$

Assim, dependendo do valor estimado para o coeficiente β , pode-se ter alteração na média μ , podemos ter o seguinte,

- Se $\beta_1 = 0$, isso não irá alterar o valor da média se houver alguma mudança na variável x_i ;
- Se $\beta_1 > 0$, qualquer alteração na variável regressora implicaria em um aumento na média μ ; e
- Se $\beta_1 < 0$, qualquer alteração na variável regressora implicaria na redução na média μ .

Atualmente vários *software* estatísticos como R, SAS e MATLAB, permitem a estimação dos coeficientes de regressão β_p do modelo (CORDEIRO; DEMÉTRICO, 2008). Mas, segundo Turkman e Silva (2000), podemos estimar os valores de β_1, \dots, β_p pelo método de Máxima Verossimilhança (MV) em função de β . Assim, temos

$$L(\beta) = \prod_{i=1}^n f(y_i). \quad (2.8)$$

Quando aplicamos a função logaritmo em 2.8, temos a log verossimilhança,

$$\ln L(\beta) = l(\beta). \quad (2.9)$$

Para encontrar uma estimativa de β_i , deremos derivar 2.9 e igualar a zero,

$$\frac{\partial l(\beta)}{\partial \beta_i} = 0. \quad (2.10)$$

Em seguida, fazer a segunda derivada de 2.10 para ver se o valor encontrado é um ponto de máximo,

$$\frac{\partial^2 l(\beta)}{\partial^2 \beta_i^2} < 0.$$

Segundo Turkman e Silva (2000), ao trabalharmos com uma Distribuição de Poisson, temos

$$Y_i \sim P(\mu_i), \quad i = 1, \dots, n, \text{ e}$$

$$f(y_i|\mu_i) = \exp\{y_i \ln(\mu_i) - \mu_i - \ln(y_i!)\}.$$

Como estamos utilizando a função de ligação $g(\mu) = \ln(\mu)$, podemos reescrever 2.6 como

$$\ln(\mu_i) = \mathbf{X}'_i \boldsymbol{\beta},$$

em que $\mathbf{X}_i = (x_{i1}, \dots, x_{ip})$ é um vetor de dimensão p associado às covariáveis x_i e $\boldsymbol{\beta}$ é um vetor de parâmetros desconhecidos β_i .

Portanto, temos a função de verossimilhança dada por

$$L(\boldsymbol{\beta}) = \exp\left\{\sum_{i=1}^n y_i \mathbf{X}'_i \boldsymbol{\beta} - \sum_{i=1}^n e^{(\mathbf{X}'_i \boldsymbol{\beta})} - \sum_{i=1}^n \ln(y_i!)\right\}.$$

Normalmente não encontramos equações lineares e soluções analíticas. Sendo assim, podemos utilizar métodos numéricos como o método de Newton-Raphson ou o método escore de Fisher (REITMAN, 2007) para a estimação de $\boldsymbol{\beta}$.

3. MODELOS PARA DADOS LONGITUDINAIS DE CONTAGENS

Dados Longitudinais, são dados que as mesmas, ou diferentes, variáveis são observadas em, pelo menos, dois pontos distintos do tempo para a mesma unidade. Existem muitas áreas que utilizam deste tipo de dado para pesquisa, como agricultura, epidemiologia e economia. Segunda Vieira (2012), os estudos envolvendo dados longitudinais estão ganhando espaço nos últimos anos. E estas pesquisas longitudinais são importantes para observar mudanças na característica de certa população ou indivíduo ao longo do tempo.

A seguir, na subseção 3.1 introduzimos o conceito de Estudos Longitudinais, na subseção 3.2 apresentaremos Modelos para Dados Longitudinais. Logo depois, na subseção 3.2.1 apresentamos a Distribuição de Poisson para Dados Longitudinais, seguida dos Modelos

Lineares Generalizados Mistos na subseção 3.2.2. Por fim, na subseção 3.2.3, apresentamos o Modelo de Regressão Poisson para Dados Longitudinais.

3.1 Estudos Longitudinais

Segundo Vieira (2012), podemos identificar dois tipos de estudos longitudinais:

- i. Estudos transversais repetidos: em que as mesmas variáveis são medidas repetidamente ao longo do tempo, para amostras diferentes;
- ii. Estudos do tipo painel: em que as mesmas variáveis e/ou diferentes são medidas em, no mínimo, dois tempos para os mesmos indivíduos.

No nosso trabalho, são considerados dados longitudinais do tipo painel, uma vez que são analisadas as mesmas variáveis para os mesmos indivíduos (municípios) ao longo do tempo.

A coleta de dados deste tipo pode demandar meses ou até anos, dependendo do contexto, resultando em custos elevados. Porém, neste trabalho não houve esta dificuldade, uma vez que os dados foram organizados e disponibilizados pela JF Salvando Todos (2020).

Outro problema que pode acontecer na coleta desses dados, é a não resposta ou a ausência de dados para um determinado caso, para uma determinada variável para um ou mais momentos no tempo. Esse fenômeno é chamado de não resposta longitudinal.

Neste estudo, não estamos trabalhando com dados individuais de pessoas, mas sim com dados agregados para o nível dos municípios. Logo a não resposta que poderemos ter, seria o município em questão não divulgar os dados de Covid-19. Para identificar se iremos realizar análises para dados balanceados, em que temos o mesmo número de observações ao longo do tempo para todas as cidades; ou desbalanceados, em que o número de observações varia (VIEIRA, 2012); será feita uma análise exploratória dos dados na Seção 4.

3.2 Modelos para Dados Longitudinais

Segundo Vieira (2012), os estudos longitudinais permitem diferir o grau de variação na variável resposta de um indivíduo ao longo do tempo e a variação entre indivíduos diferentes. Assim, podemos utilizar os modelos chamados de dois estágios, já que temos que levar em consideração a variação intra-indivíduo (estágio 1) e entre-indivíduos (estágio 2) (HAND; CROWDER, 1996).

A forma mais simples de um estudo longitudinal, é aquele em que a variável resposta é medida em apenas dois tempos (TWISK, 2003). Mas também podemos trabalhar com os dados, em que a variável resposta é medida em três ou mais tempos. Nessa situação podemos criar J –

1 tabelas de dupla entrada com dimensões $r \times c$, onde J é a quantidade de vezes que a variável resposta foi medida, e r (linha) e c (coluna) são as quantidades de categorias analisadas.

A seguir, são apresentados no Quadro 2.1 alguns exemplos de tabelas de duplas entradas. Suponha que desejamos estudar apenas a variação de uma variável resposta, sem levar em consideração as variáveis explicativas, em um grupo de J tempos diferentes. Logo, teríamos as seguintes tabelas de dupla-entrada:

Quadro 2.1. Indicadores de mudanças em dois tempos de J tempos medidos

		$j = 2$						$j = 3$						
		cat_1		...	cat_r				cat_1		...	cat_r		
$j = 1$	cat_1	n_{11}	...					n_{1r}	$j = 2$	cat_1	n_{11}	...	n_{1r}	
		
	cat_c	n_{c1}	...					n_{cr}		cat_c	n_{c1}	...	n_{cr}	
		$j = 4$						$j = J$						
		cat_1		...	cat_r				cat_1		...	cat_r		
$j = 3$	cat_1	n_{11}	...					n_{1r}	$j = J - 1$	cat_1	n_{11}	...	n_{1r}	
		
	cat_c	n_{c1}	...					n_{cr}		cat_c	n_{c1}	...	n_{cr}	

Fonte: Adaptação de Lima (2014).

Segundo Twisk (2003), quando se trabalha com dados longitudinais, é preciso verificar a relação existente entre a variável resposta com as variáveis predictoras e o tempo. Na escolha da técnica de modelagem a ser adotada, um critério poderia ser a natureza da variável resposta a ser analisada, se é discreta ou contínua. Assim, como a variável explicativa deste trabalho é do tipo contagem, podemos pressupor que ela tenha uma distribuição de Poisson, logo, podemos aplicar a regressão log-linear, mas levando em conta a variável tempo, aqui denotada por j , já que os dados são coletados repetidamente ao longo do tempo. Outra decisão importante é sobre como considerar a variável tempo, como discreta ou contínua. Neste estudo, assim como em

Vieira (2012), o tempo é considerado como discreto, com $j = 1, 2, \dots, J_i$. Quando os dados longitudinais são balanceados, $J_i = J \forall i$.

O modelo mais geral para os dados longitudinais, é da forma (VIEIRA, 2012)

$$Y_{ij} = x_{ij}\boldsymbol{\beta} + e_{ij},$$

em que Y_{ij} é a i -ésima observação no j -ésimo tempo, x_{ij} é um vetor de covariáveis fixas, e e_{ij} é o erro aleatório do modelo, que tem como característica valor esperado igual a zero.

Segundo Hand e Crowder (1996), quando se considera os efeitos fixos e aleatórios, o modelo para a i -ésima observação é dado por

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i, \quad (2.11)$$

em que \mathbf{b}_i é um vetor de efeitos aleatórios desconhecidos, permitindo uma contribuição combinada para \mathbf{Y}_i ($J \times 1$) através da matriz de design \mathbf{Z}_i ; $\boldsymbol{\beta}$ é um vetor de parâmetros de efeitos fixos desconhecidos de tamanho $1+p \times 1$, com matriz de delineamento conhecida \mathbf{X}_i , $J \times p$ para o i -ésimo indivíduo. A matriz \mathbf{Z}_i é constituída a partir de variáveis explicativas da mesma forma que uma matriz \mathbf{X}_i . Porém, \mathbf{Z}_i é diferente de \mathbf{X}_i , pois utilizamos \mathbf{X}_i para modelar a média $\boldsymbol{\mu}_i$, e \mathbf{Z}_i para modelar a covariância (HAND; CROWDER, 1996).

Note que, se $\mathbf{e}_i = \sigma^2\mathbf{I}$ e $\mathbf{Z} = \mathbf{0}$, o modelo misto em 2.11, se reduz ao modelo linear padrão. Segundo Crowder e Hand (1990, p. 89), temos para \mathbf{y}_i que

$$\boldsymbol{\mu}_i = E(\mathbf{y}_i) = \mathbf{X}_i\boldsymbol{\beta}, \text{ e}$$

$$\boldsymbol{\Sigma}_i = V(\mathbf{y}_i) = \mathbf{Z}_i\mathbf{B}\mathbf{Z}_i' + \mathbf{E}_i, \quad (2.12)$$

em que $E(\mathbf{e}_i) = \mathbf{0}$, $Var(\mathbf{e}_i) = \mathbf{E}_i$, $E(\mathbf{b}_i) = \mathbf{0}$, $Var(\mathbf{b}_i) = \mathbf{B}$, e $Cor(\mathbf{b}_i, \mathbf{e}_i) = \mathbf{0}$.

Podemos estimar $\boldsymbol{\beta}$ pelo método dos mínimos quadrados. Assim temos que

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^n \mathbf{X}_i' \boldsymbol{\Sigma}_i^{-1} \mathbf{X}_i \right)^{-1} \left(\sum_{i=1}^n \mathbf{X}_i' \boldsymbol{\Sigma}_i^{-1} \mathbf{y}_i \right)$$

em que $\boldsymbol{\Sigma}_i$ é dado pela expressão 2.12.

3.2.1 Distribuição de Poisson para dados Longitudinais

Considerando dados longitudinais, temos a variável aleatória \mathbf{Y}_i , sendo um vetor do i -ésimo indivíduo ($i = 1, 2, \dots, N$) na j -ésima medida longitudinal ($j = 1, 2, \dots, J_i$). Seja $\mathbf{Y}_i =$

$(Y_{i1}, \dots, Y_{ij_i})^T$ a medida de interesse para o caso i e \mathbf{x}_i um vetor de covariáveis associado a \mathbf{Y}_i . Para o modelo de Poisson com $Y_{ij} \sim \text{Poisson}(\mu_{ij})$, $\mu_{ij} > 0$, temos (RIZZATO, 2011)

$$P(\mathbf{Y}_i = \mathbf{y}_i) = \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!}, \text{ para } y_i = 0, 1, 2, \dots$$

Considerando

$$\ln(\mu_i) = \mathbf{x}'_i \boldsymbol{\beta},$$

temos

$$\begin{aligned} E(\mathbf{Y}_i | \mathbf{x}_i, \boldsymbol{\beta}) &= \boldsymbol{\mu}_i = \exp(\mathbf{x}'_i \boldsymbol{\beta}) \text{ e} \\ \text{Var}(\mathbf{Y}_i | \mathbf{x}_i, \boldsymbol{\beta}) &= \boldsymbol{\mu}_i = \exp(\mathbf{x}'_i \boldsymbol{\beta}). \end{aligned}$$

3.2.2 Modelos Lineares Generalizados Mistos

Uma metodologia que podemos utilizar para estudarmos o comportamento dos dados longitudinais, tem como base o Modelo Linear Generalizado Misto (MLGM), que pode ser entendido como uma extensão dos modelos Modelo Linear Misto (MLM) ou do MLG. O MLM leva em consideração os efeitos aleatórios e pressupõe uma distribuição normal para a variável resposta. Esse modelo pode ser usado em análises de medidas repetidas. Porém, a variável resposta pode assumir apenas uma única distribuição, a normal. Sendo assim, o que difere o MLM do MLGM, é o fato que no modelo MLGM a variável resposta pode assumir um conjunto de distribuições de probabilidades, discretas ou contínuas (OIKAWA, 2019).

Com a adoção do MLG, como foi visto na Seção 2.2, a variável resposta pode assumir outras distribuições, como Poisson, Gama, Binomial, entre outras. Entretanto, esse modelo considera apenas efeitos fixos. Portanto, podemos dizer que o MLGM é uma extensão tanto do MLM, no que tange a modelagem de dados longitudinais, quanto do MLG, quando a variável resposta segue distribuições não-normal.

Na teoria dos modelos lineares mistos, assumimos que

- i. os erros têm distribuição normal, e que
- ii. a variável resposta é uma combinação linear de efeitos fixos e aleatórios.

Assim, como foi explicado na Seção 2.2 para os MLG's, os componentes de um MLGM são (COSTA, 2003)

- i. *Componente aleatório*: variável resposta pertencente à família exponencial, e
- ii. *Componente Sistemático*: $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$, em que o preditor linear $\boldsymbol{\eta}$ é um vetor de dimensões $n \times 1$, \mathbf{X} é a matriz de covariáveis $n \times p$, e $\boldsymbol{\beta}$ é um vetor de parâmetros desconhecidos $1 \times p$.
- iii. *Função de Ligação*: $\eta(\boldsymbol{\mu}_i) = g(E(Y_i|\mathbf{b}_i)) = \mathbf{x}'_i\boldsymbol{\beta} + \mathbf{z}_i\mathbf{b}_i$.

A estimação dos parâmetros do MLGM, não é trivial uma vez que o processo de maximização da função de verossimilhança envolve uma grande quantidade de integrais, tornando a obtenção de uma forma fechada uma prática inviável.

Segundo Oikawa (2019), a função de verossimilhança para dados longitudinais, seria da seguinte forma

$$L(\boldsymbol{\beta}, \mathbf{B}, \phi) = \prod_{i=1}^N \int_{\mathbb{R}^k} \prod_{j=1}^{n_i} f_i(\mathbf{y}_i|\mathbf{b}_i; \boldsymbol{\beta}, \phi) f(\mathbf{b}_i|\mathbf{B}) d\mathbf{b}_i. \quad (2.13)$$

A expressão (2.13) envolve a densidade de uma probabilidade $f_i(\mathbf{y}_i|\mathbf{b}_i; \boldsymbol{\beta}, \phi)$ para a i -ésima observação, condicionada em seus efeitos aleatórios, para os efeitos fixos $\boldsymbol{\beta}$ e de dispersão ϕ .

3.2.3 Modelo de Regressão Poisson para Dados Longitudinais

Assim como para caso dos dados de contagem transversais, o modelo de regressão Poisson aplicado a dados longitudinais é constituído pelos mesmos componentes (ver Seção 2.3), mas tendo como variável dependente y_{ij} . A função de ligação em um contexto de regressão de Poisson, quando se considera dados em painel, é o logaritmo, de modo que

$$\boldsymbol{\eta}_i = \log(\boldsymbol{\mu}_i).$$

Podemos estimar as taxas médias pela inversa da própria função de ligação, a partir das estimativas dos coeficientes do modelo de regressão, por meio de $\hat{\boldsymbol{\mu}}_i = \exp(\hat{\boldsymbol{\eta}}_i)$ (OIKAWA, 2019).

Como foi mostrado acima, na equação 2.13, considerando um caso em que a variável resposta segue uma distribuição de Poisson ($\boldsymbol{\mu}_i$), podemos utilizar a função ligação “log” para interligar $\boldsymbol{\mu}_i$ com o preditor linear $\boldsymbol{\eta}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{b}_i$, de modo que $\log(\boldsymbol{\mu}_i) = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{b}_i$. Seja a

variância de b_i denotada por σ_b^2 , temos a função log-verossimilhança definida como (OIKAWA, 2019)

$$l(\boldsymbol{\beta}, \mathbf{B}, \phi) = \log \left(\prod_{i=1}^n \int_{-\infty}^{+\infty} \prod_{j=1}^{J_i} \frac{\mu_{ij}^{y_{ij}} e^{-\mu_{ij}}}{y_{ij}!} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2} b_i^2} db_i \right).$$

Como há uma dificuldade para obtenção de uma forma fechada para os estimadores a partir da equação acima, vários métodos numéricos foram desenvolvidos e podem ser utilizados, como por Aproximação de Laplace e Métodos de Monte Carlo (AGRESTI, 2015).

A análise de regressão de Poisson para Dados Longitudinais, é uma extensão da regressão de Poisson levando em conta a correlação intra-casos.

Como o valor esperado pode sofrer alterações ao longo do tempo, não é possível utilizar o preditor linear $\eta = \ln(\mu)$ como é no Modelo de Poisson, pois devemos levar em consideração o tempo (j). Assim, o preditor linear pode ser da seguinte forma, já que o valor de j é conhecido (FITZMAURICE et al, 2011)

$$\eta = \ln(\mu_i/j) = \ln(\mu_i) - \ln(j).$$

Como consequência da mudança no preditor linear, a equação do modelo passa a ser da seguinte forma

$$\ln(\mu_i/j) = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \dots + \beta_p x_{pj},$$

ou pode ser escrita como

$$\ln(\mu_i) = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \dots + \beta_p x_{pj} + \ln(j), \quad (2.14)$$

em que p é o número de covariáveis consideradas no modelo. Podemos ver que o termo $\ln(j)$ não inclui nenhum parâmetro. Sendo assim, mesmo adicionado a variável tempo, não se faz necessário estimar mais um parâmetro na expressão 2.14.

Segundo Twisk (2003), é necessário criar uma tabela para verificar as correlações intra-individual no decorrer do tempo. Assim, um exemplo para esta verificação é dado na Tabela 3.1 a seguir.

Tabela 3.1. Correlação no tempo j_i

	j_1	j_2	j_3	j_4	j_5	j_6
j_1		ρ_1	ρ_2	ρ_3	ρ_4	ρ_5
j_2	ρ_1		ρ_6	ρ_7	ρ_8	ρ_9
j_3	ρ_2	ρ_6		ρ_{10}	ρ_{11}	ρ_{12}
j_4	ρ_3	ρ_7	ρ_{10}		ρ_{13}	ρ_{14}
j_5	ρ_4	ρ_8	ρ_{11}	ρ_{13}		ρ_{15}
j_6	ρ_5	ρ_9	ρ_{12}	ρ_{14}	ρ_{15}	

Fonte: Twisk (2003)

4. BANCO DE DADOS

Como foi informado anteriormente, os dados que utilizados neste trabalho, foram disponibilizados pela plataforma JF Salvando Todos (2020). Nas subseções seguintes, são apresentados os dados (4.1) e uma análise exploratória deles (4.2).

4.1. Apresentação dos Dados

O banco de dados conta com 10.902 observações, disponibilizadas por semana epidemiológica (a semana epidemiológica começa no domingo e termina no sábado). Temos que os casos começam na 13ª semana de 2020 (22 a 28 de março) até a 26ª semana epidemiológica de 2021 (26 de junho a 3 de julho), sendo assim, temos no total 67 semanas com dados. Mas para aplicação do modelo, foi considerado da 1ª semana epidemiológica de 2020 (29 de dezembro de 2019 a 4 de janeiro de 2020) até a 26ª semana de 2021. Foram estudados 138 municípios que possuem entre 200 mil e 1 milhão de habitantes, segundo projeções do IBGE para 2020.

As variáveis utilizadas neste trabalho, são as que foram disponibilizadas pela Plataforma JF Salvando Todos (2020). A seguir, são apresentadas na Tabela 4.1 as variáveis consideradas neste trabalho.

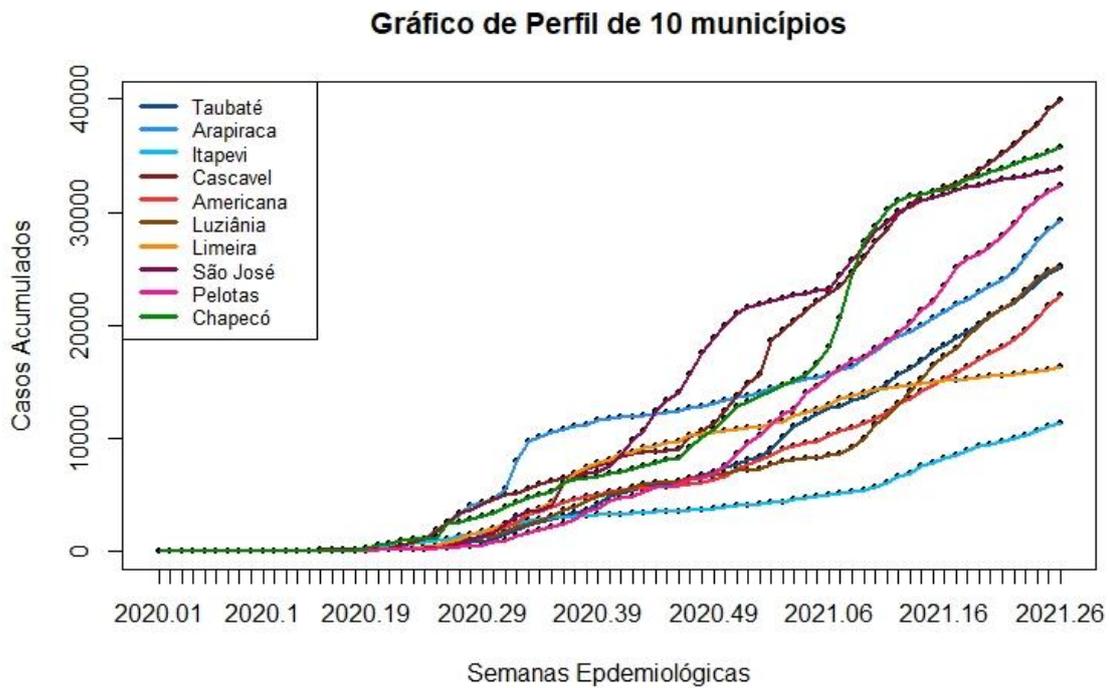
Tabela 4.1. Apresentação e explicação das variáveis.

Variável	Descrição
Casos confirmados	Número de pessoas que testaram positivos para a Covid-19
Casos de óbitos	Número de pessoas que faleceram decorrente da Covid-19
Mobilidade	Proporções de pessoas se deslocando para estabelecimentos de Varejo e Recreação, Mercearia e Farmácias, Parques, Estações de Trânsito, Locais de Trabalho e permanência na residência (Residencial), em relação a um período de referência pré-pandemia. Dados do Google Mobility.
Rt estimado	Número efetivo de reprodução para a Covid-19 (número médio de casos secundários produzidos por um único indivíduo infectado)
Leitos	Número de Leitos de Internação disponíveis para casos cirúrgicos, clínicos, obstétrico, pediátrico e outras especialidades
	Número de Leitos Complementares de UTI adulto II Covid-19 e UTI pediátrico II Covid-19.
Respiradores	Número de Respiradores em Uso pelos estabelecimentos
	Número de Estabelecimentos que possuem Respiradores

4.2. Análise Exploratória dos Dados

Nos gráficos a seguir, são apresentados os casos acumulados (figura 4.1) e óbitos acumulados (figura 4.2) por semana epidemiológica para 10 municípios selecionados aleatoriamente por Amostragem Aleatória Simples (AAS).

Figura 4.1 – Gráfico de Perfil de casos acumulados de 10 municípios sorteados por AAS



Os dados considerados desta monografia foram a partir da 13ª semana, como foi falado anteriormente, mas da 1ª a 12ª semana epidemiológica de 2020, consideramos os dados como 0. Assim, pela figura acima, as 12 primeiras semanas epidemiológicas estudadas não possuem dados, mas, a partir da 13ª semana, os casos confirmados começaram a ser registrados em algumas cidades. Podemos inferir que alguns municípios tiveram um crescimento de infectados maior que outras, como exemplo, temos a cidade de Cascavel (marrom escuro) que teve um crescimento exponencial mais acentuado que a cidade de Itapevi (azul mais claro).

Figura 4.2 – Gráfico de Perfil de casos acumulados com enfoque nos municípios de Itapevi e Cascavel

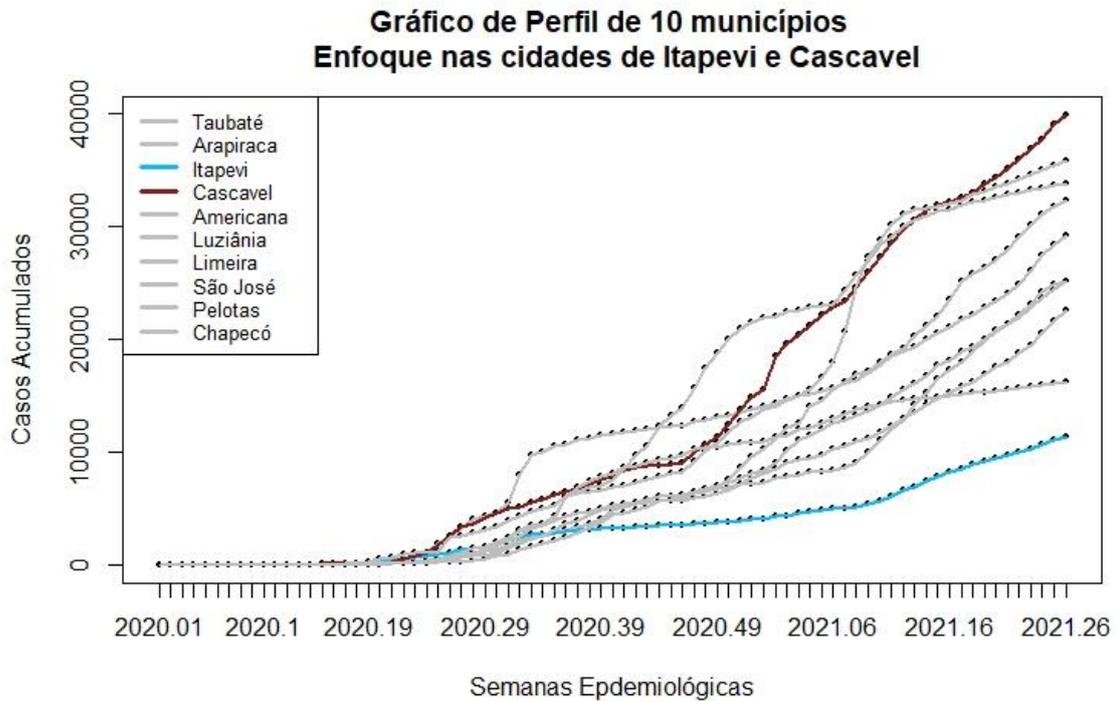
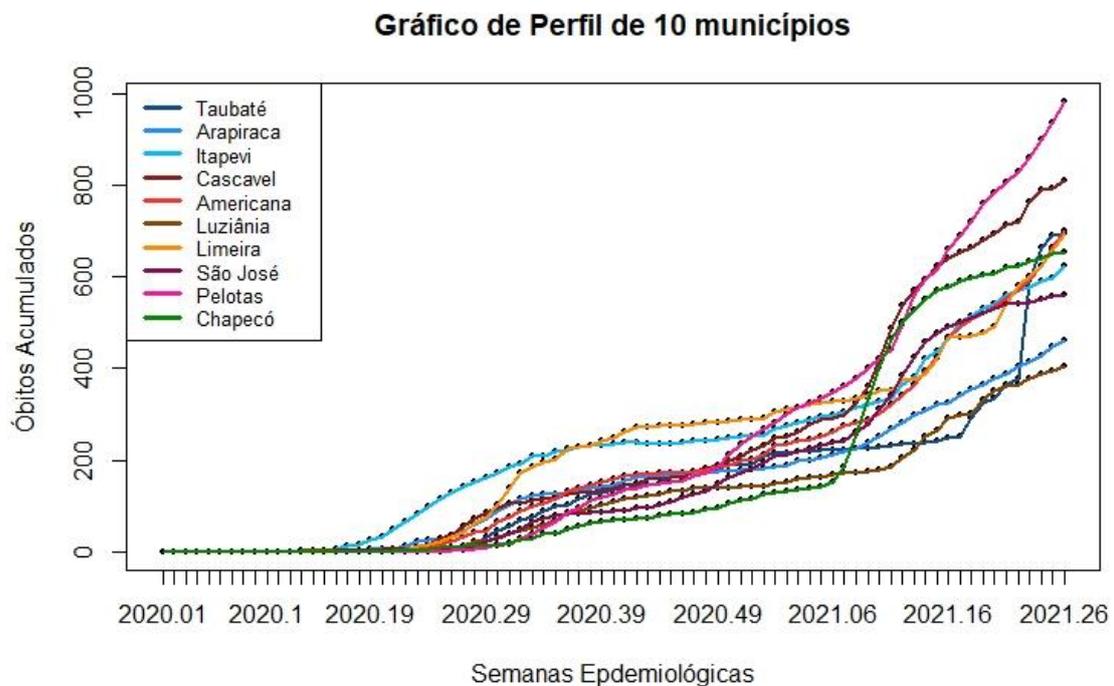


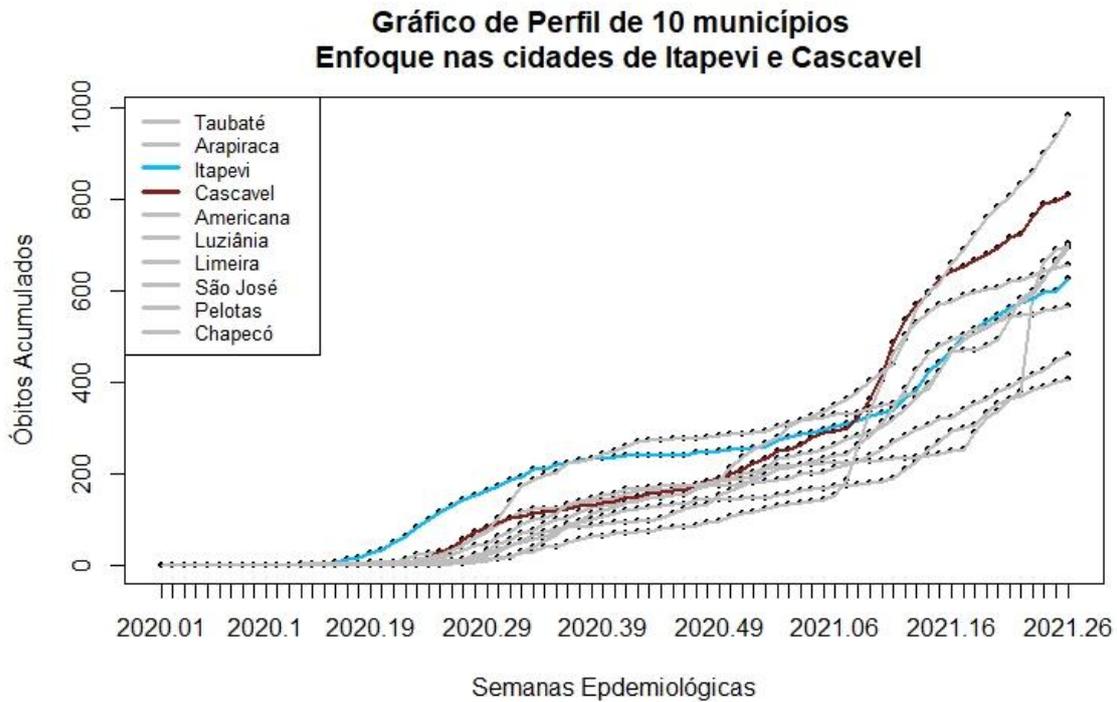
Figura 4.3 – Gráfico de Perfil de óbitos acumulados de 10 municípios sorteados por AAS



Em comparação com a figura 4.1, a cidade de Cascavel, veio com os registros de óbitos crescente como outras cidades, mas a partir da 6ª semana epidemiológica de 2021, há um salto

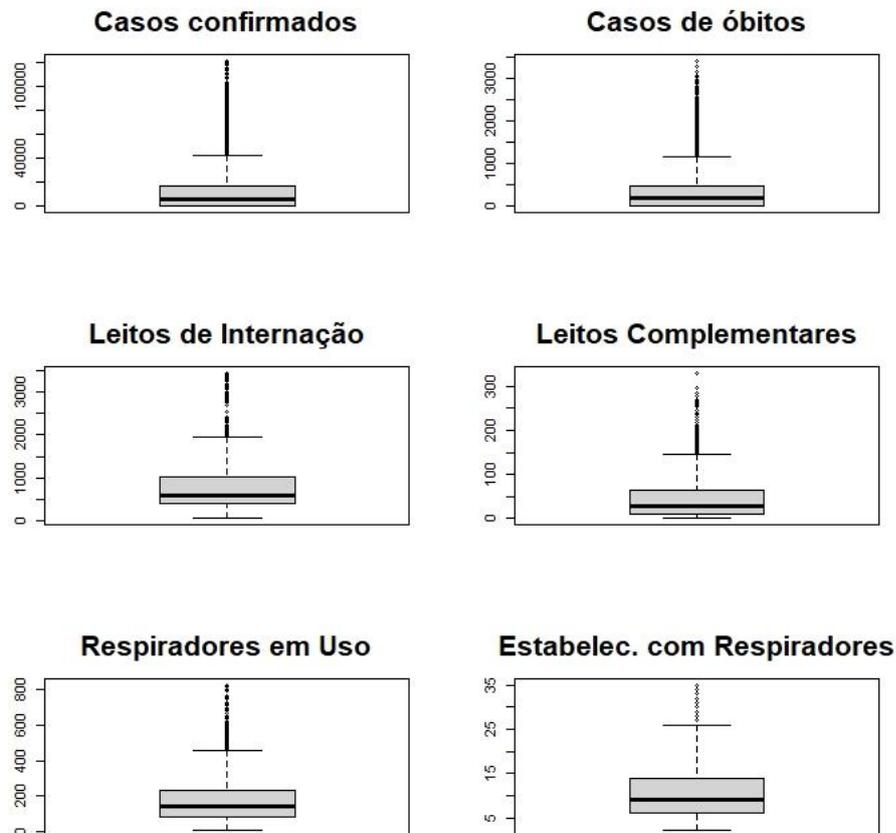
nos registros de vidas perdidas, e se torna a segunda cidade com mais casos de óbitos, dentre as 10 selecionadas. Já a cidade de Itapevi, que possuía menos registros de casos confirmados entre os municípios selecionados, ela se destaca por ser a cidade com mais casos de óbitos nas primeiras semanas estudadas. E mesmo depois de outras cidades aumentarem seus registros, ela ainda se destaca pelos casos de óbitos.

Figura 4.4 – Gráfico de Perfil de óbitos acumulados com enfoque nos municípios de Itapevi e Cascavel



Para uma melhor compreensão dos dados, a seguir são apresentados Box Plots para algumas variáveis apresentadas na tabela 4.1.

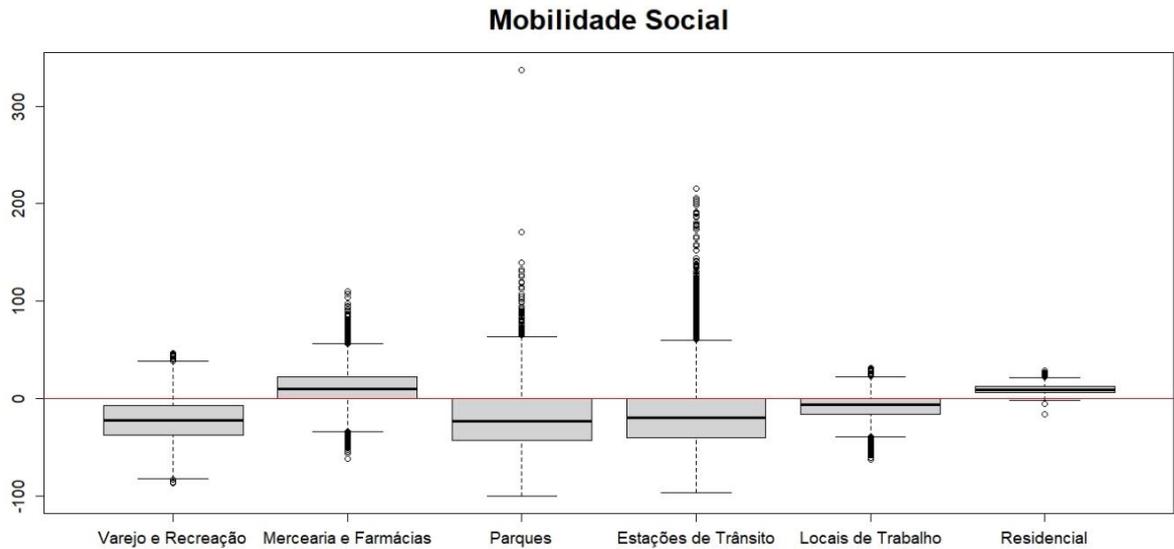
Figura 4.5 – Box Plots de casos confirmados, óbitos, leitos de internação e complementar, respiradores em uso e estabelecimentos com respiradores



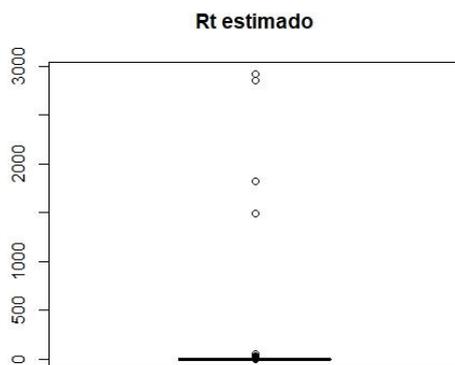
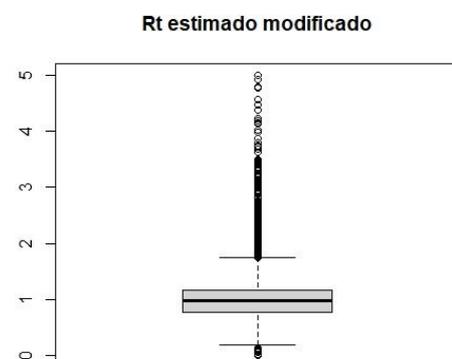
Podemos perceber que há muitos *outliers* acima do terceiro quartil em todas as variáveis apresentadas na figura 4.5. Na figura 4.6, são apresentados Box Plots para as categorias da mobilidade social.

Em relação à mobilidade social, a reta em vermelho na figura 4.6, é o valor de referência definido como mediana para o dia da semana calculada para o período de 5 semanas entre os dias 3 de janeiro e 6 de fevereiro de 2020. É possível perceber que 75% das semanas das categorias Varejo e Recreação, Parques, Estações de Trânsito e Locais de Trabalho, se encontram abaixo desse período de referência. Isso é algo esperado, pois com a pandemia, as pessoas foram informadas para permanecer em casa. E essa permanência é vista na categoria “Residencial”, em que quase todas as semanas se encontram acima do período pré-pandêmico. Mercarias e Farmácias, foi a categoria que teve uma grande procura, ficando 75% das semanas epidemiológicas acima do período de referência.

Figura 4.6 – Box Plots para as categorias de mobilidade social



Temos pela figura 4.7, que os dados de R_t estimados para as cidades selecionadas apresentam outliers muito grandes. E isso é algo que pode acontecer, pois os dados podem vir com algum erro do Ministério da Saúde e, em alguns casos, não há disponibilização dos dados corretos do dia, mas eles são disponibilizados de uma forma a suprir o erro da última divulgação. Outro fator que eleva o R_t , é a demora na divulgação dos casos. Assim, aparamos os outliers que eram maiores do que cinco, que passaram a ter esse valor. Na figura 4.8, os dados de R_t já estão na forma que será usado para estudo.

Figura 4.7 – R_t estimadoFigura 4.8 – R_t estimado modificado

5. APLICAÇÕES E RESULTADOS

Para ajustar um Modelo de Poisson Longitudinal com efeitos aleatórios, utilizamos a biblioteca “*panelr*”. Com a função *wbm*, ajustamos a regressão considerando óbitos como variável dependente e as demais variáveis como preditoras.

Após uma análise exploratória, vimos que a média e a variância dos dados foram bem diferentes, mas optamos pelo modelo de Poisson pela facilidade de interpretação dos coeficientes. Assim, foi informado na função “*wbm*”, que a família que seria utilizada era a Poisson, e consideramos como variável que não varia com o tempo a população de cada cidade.

Para utilização desse modelo, foi preciso identificar e informar a variável que corresponde aos municípios e a variável tempo, que no nosso caso, são as semanas epidemiológicas. Logo após a identificação, foram feitos sucessivos ajustes na função *wbm*, visando a seleção de preditores com coeficientes significativos ao nível de 10%. O modelo final é apresentado na Tabela 5.1 a seguir. Temos que a variável Leitos Complementares não foi significativa no modelo, assim ela não interfere no número médio de óbitos.

Foi preciso aplicar o comando *lag* em algumas variáveis explicativas para retardar seu efeito. Como por exemplo, se observarmos a Tabela 5.1 a seguir, temos *lag*(casos confirmados, 2), isso indica que estamos considerando os casos confirmados de 2 semanas epidemiológicas atrás. Considerando a experiência acumulada e resultados dos estudos desenvolvidos pela JF Salvando Todos (2020), de cuja equipe a autora do presente trabalho é membro, a definição do número de *lags* considerados para cada variável preditora foi feita de forma a identificar momentos temporais, efeitos e padrões que poderiam auxiliar os tomadores de decisão de políticas de saúde pública em medidas para o controle da pandemia. Assim, a seguir é apresentado uma tabela com algumas saídas do modelo. No Anexo 2 é apresentado o modelo que foi implementado no software R.

Tabela 5.1. Coeficientes do Modelo

Variáveis explicativas	Coefficiente estimado	Erro padrão	Estatística z	Valor p
(Intercepto)	2,21978	0,04187	53,02057	0,00001
lag(casos confirmados, 2)	0,00050	0,00000	120,26266	0,00001
lag(varejo e recreação, 11)	0,00112	0,00015	7,35957	0,00001
lag(mercearia e farmácias)	0,01997	0,00033	60,81085	0,00001
lag(parques, 2)	-0,01055	0,00019	-56,19351	0,00001
lag(estações de trânsito)	0,00026	0,00015	1,74492	0,08100
lag(locais de trabalho)	0,01680	0,00054	30,95729	0,00001
lag(residencial)	0,11045	0,00155	71,47839	0,00001
lag(rt estimado, 5)	0,23236	0,00484	47,98091	0,00001
lag(respiradores em Uso)	0,00347	0,00008	42,01749	0,00001
lag(estabelecimentos com respiradores)	0,00609	0,00330	1,84872	0,06450
leitos de internação	-0,00021	0,00002	-9,24809	0,00001

O modelo final inclui coeficientes significativos ao nível de 10%. Assim, foram selecionadas variáveis explicativas para o número de vidas perdidas pela Covid-19. Considerando o caráter longitudinal do modelo, a partir do valor estimado para o intercepto do modelo, podemos interpretar que se não houver variação em nenhuma variável explicativa, o número de óbitos por Covid-19 será de aproximadamente 2 em média em cada um dos municípios considerados.

Duas variáveis explicativas apresentaram coeficientes estimados com sinais negativos: a proporção de pessoas se deslocando para Parques e o número de leitos disponíveis para internação. Podemos interpretar que se espera que uma variação positiva de 10 pontos percentuais na proporção de pessoas se deslocando para parques duas semanas atrás esteja relacionado a uma redução média de aproximadamente 1 vida perdida. Este resultado é coerente uma vez a Covid-19 é transmitida principalmente pelas vias respiratórias e os parques são áreas abertas. Espera-se também que seria necessária uma variação positiva de 10.000 leitos de internação para se esperar uma redução média de 2 vidas perdidas por município e semana epidemiológica.

As demais variáveis explicativas apresentaram coeficientes estimados positivos. Aumentos no número de casos, como esperado, estão relacionados a aumentos no número de vidas perdidas. Variações positivas na proporção de pessoas se deslocando para pontos de venda de varejo, mercearias e farmácias, fazendo uso do transporte público e indo aos seus locais de trabalho estão relacionadas a aumentos no número de vidas perdida em média. Outro resultado que consideramos muito relevante e que também corrobora com a adoção de medidas de controle da circulação de pessoas é o coeficiente positivo para o Rt 5 semanas atrás, que indica que variações positivas na taxa de transmissão da Covid-19 estão relacionadas a aumentos do número médio de vidas perdidas.

Um resultado, à primeira vista contraintuitivo, é o coeficiente positivo para o percentual de pessoas que permanecem em casa. Uma possível interpretação é que em momentos em que o número de vidas perdidas está elevado em seu município de residência as pessoas se sentem mais motivadas a permanecerem em casa. Entretanto entendemos que seriam necessárias maiores investigações para podermos explorar melhor este resultado. Outros resultados que poderiam ser fonte de maiores reflexões são os coeficientes positivos para o número de respiradores disponíveis e respiradores em uso. Uma justificativa para este resultado é que a Covid-19 é uma enfermidade que apresenta alta mortalidade em pacientes internados em UTIs e fazendo uso de respiradores.

6. CONCLUSÃO

Nesta monografia conseguimos atingir os objetivos expostos na Seção 1. A seguir, destacamos alguns dos principais resultados alcançados. Identificamos uma relação positiva entre o número de casos confirmados e de vidas perdidas, algo que era esperado. O R_t foi a variável explicativa com o coeficiente com maior magnitude. Variações positivas no R_t estão relacionadas a aumentos na média do número de vidas perdidas. Essa relação positiva era esperada, já que como o cálculo do R_t utiliza o número de casos confirmados, e no nosso modelo a relação entre casos confirmados e vidas perdidas foi positiva. Este resultado pode também explicar a baixa magnitude do coeficiente para a variável número de casos uma vez que o seu efeito pode estar sendo captado pelo R_t . Como a Covid-19 pode impossibilitar as pessoas de conseguirem respirar sozinhas em estágios mais avançados, é preciso o uso de ventilação mecânica e equipar os estabelecimentos com respiradores. Entretanto, dada a alta mortalidade entre os pacientes com Covid-19 internados e com uso de respiradores, o modelo ajustado apresentou coeficientes positivos para as variáveis que captavam a quantidade e o uso de respiradores.

As maiores dificuldades enfrentadas para a execução deste trabalho, foram a organização da base de dados dada a sua complexidade e a utilização da função *wbm* no *software* R pela demora de sua aplicação. Os dados estudados nesta monografia, foram da 13ª semana epidemiológica de 2020 (22 a 28 de março) até a 26ª semana epidemiológica de 2021 (27 de junho a 3 de julho), sendo assim, para trabalhos futuros, poderia utilizar dados cada vez mais recentes para confirmar, entre outros aspectos, a relação dos óbitos com as variáveis preditoras estudadas. Além disso, alternativas de modelos poderiam ser consideradas em trabalhos futuros, modelos que lidam com a superdispersão dos dados, ou modelos com coeficientes aleatórios. Assim como inserir novas variáveis explicativas no modelo para remediar o problema da superdispersão e considerar as interações entre as variáveis preditoras.

7. REFERÊNCIAS BIBLIOGRÁFICAS

- AGRESTI, A. **Foundations of Linear and Generalized Linear Models**. Em: *Wiley Series in Probability and Statistics*. 2015.
- BELASCO, Angélica Gonçalves Silva; FONSECA, Cassiane Dezoti da. **Coronavírus 2020**. Revista Brasileira de Enfermagem. 2020.
- BUCHALLA, Cássia Maria; CARDOSO, Maria Regina Alves. **Principais desenhos de estudos epidemiológicos**. Em: *Epidemiologia dos agravos à saúde da mulher* [S.l: s.n.], 2005.
- CASELLA, George; Berger, Roger L. **Inferência Estatística**. 2ED. São Paulo: Cengage Learning, 2010.
- CORDEIRO, Gauss Moutinho; DEMÉTRIO, Clarice G. B. **Modelos Lineares Generalizados e Extensões**. Piracicaba, 2011.
- COSTA, Silvano Cesar da. **Modelos lineares generalizados mistos para dados longitudinais**. 2003. Universidade de São Paulo. Escola Superior de Agricultura “Luiz de Queiroz”.
- CROWDER, M. J.; HAND, D. J. **Analysis of Repeated Measures**. Chapman and Hall, 1990.
- DIAS, Larissa B. A.; ALMEIDA, Sérgio C. L. de; HAES, Tissiana M. de; MOTA, Letícia M.; RORIZ-FILHO, Jarbas S. **Dengue: transmissão, aspectos clínicos, diagnósticos e tratamento**. Universidade de São Paulo. Simpósio: “Condutas em enfermagem de clínica médica de hospital de média complexidade – 1ª Parte”. V. 43 n. 2, (2010), p. 143-152.
- HAND, David; CROWDER, Martin. **Practical Longitudinal Data Analysis**. Reino Unido: Chapman & Hall, 1996.
- HUANG, Chaolin; WANG, Yeming; LI, Xingwang; REN, Lili; ZHAO, Jianping; HU, Yi. **Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China**. Journal The Lancet, Vol. 395, No. 10223, (2020), p. 497-506.
- Ihaka, R. and Gentleman, R. (1996) **R: A Language for Data Analysis and Graphics**. Journal of Computational and Graphical Statistics, Vol. 5, n. 3, 299-314.
- JACOB, S.T., CROZIER, I., FISCHER, W.A. *et al.* **Ebola virus disease**. *Nat Rev Dis Primers* **6**, 13 (2020).
- JF SALVANDO TODOS, 2020. Plataforma JF Salvando Todos. Disponível em: <http://jfsalvandotodos.ufjf.br/>. Acesso em: 10 mai. 2021.
- LANDIM, Barbara. Boletim confirma primeiro caso de coronavírus em Juiz de Fora. **Tribuna de Minas**. Juiz de Fora, 14 de março de 2020. Disponível em: <https://tribunademinas.com.br/noticias/cidade/14-03-2020/boletim-confirma-primeiro-caso-de-coronavirus-em-juiz-de-fora.html>. Acesso em: 21 de agosto de 2021.
- LIMA, Marcos Alves de. **Estudo De Caso Em Um Veículo Publicitário Com Dado De Contagem Longitudinais**. 2014. Universidade Federal de Juiz de Fora. Instituto de Ciências Exatas. Departamento de Estatística.
- NELDER, J. A.; WEDDERBURN, R. W. M. **Generalized Linear Models**. Journal of the Royal Statistical Society. Serie A (General), Vol. 135, No. 3, (1972), p. 370-384.

OIKAWA, Koki Fernando. **Análise estatística de dados longitudinais e hierárquicos em Psicologia: uma análise comparativa entre GEE e GLMM**. 2019. Universidade de São Paulo. Instituto de Psicologia.

REITMAN, Diomedes Pael. **Uso de Métodos Clássicos e Bayesianos em Modelos de Regressão Beta**. 2007. Universidade Federal de São Carlos. Departamento de Estatística.

RIZZATO, Fernanda Buhner. **Modelos para análise de dados discretos longitudinais com superdispersão**. 2011. Universidade de São Paulo. Escola Superior de Agricultura “Luiz de Queiroz”.

SHAW, R. G.; MITCHELL-OLDS, T. **Anova for Unbalanced Data: na overview**. *Ecology*. Vol 74, No. 6, (1993), p. 1638-1645.

TWISK, Jos W.R. **Applied Longitudinal Data Analysis for Epidemiology: A Practical Guide**. Cambridge, EUA: Cambridge University Press, 2003.

ANEXOS

Anexo 1 – Cidades Estudadas

Tabela A1 – Cidades estudadas e sua respectiva população

Municípios	Estado	População
Duque de Caxias	Rio de Janeiro	924 624
Campo Grande	Mato Grosso do Sul	906 092
Natal	Rio Grande do Norte	890 480
Teresina	Piauí	868 075
São Bernardo do Campo	São Paulo	844 483
Nova Iguaçu	Rio de Janeiro	823 302
João Pessoa	Paraíba	817 511
São José dos Campos	São Paulo	729 737
Santo André	São Paulo	721 368
Ribeirão Preto	São Paulo	711 825
Jaboatão dos Guararapes	Pernambuco	706 867
Osasco	São Paulo	699 944
Uberlândia	Minas Gerais	699 097
Sorocaba	São Paulo	687 357
Contagem	Minas Gerais	668 949
Aracaju	Sergipe	664 908
Feira de Santana	Bahia	619 609
Cuiabá	Mato Grosso	618 124
Joinville	Santa Catarina	597 658
Aparecida de Goiânia	Goiás	590 146
Londrina	Paraná	575 377
Juiz de Fora	Minas Gerais	573 285
Porto Velho	Rondônia	539 354
Ananindeua	Pará	535 547
Serra	Espírito Santo	527 240
Caxias do Sul	Rio Grande do Sul	517 451
Niterói	Rio de Janeiro	515 317
Belford Roxo	Rio de Janeiro	513 118
Macapá	Amapá	512 902
Campos dos Goytacazes	Rio de Janeiro	511 168
Florianópolis	Santa Catarina	508 826
Vila Velha	Espírito Santo	501 325
Mauá	São Paulo	477 552
São João de Meriti	Rio de Janeiro	472 906

São José do Rio Preto	São Paulo	464 983
Mogi das Cruzes	São Paulo	450 785
Betim	Minas Gerais	444 784
Santos	São Paulo	433 656
Maringá	Paraná	430 157
Diadema	São Paulo	426 757
Jundiaí	São Paulo	423 006
Boa Vista	Roraima	419 652
Montes Claros	Minas Gerais	413 487
Rio Branco	Acre	413 418
Campina Grande	Paraíba	411 807
Piracicaba	São Paulo	407 252
Carapicuíba	São Paulo	403 183
Olinda	Pernambuco	393 115
Anápolis	Goiás	391 772
Cariacica	Espírito Santo	383 917
Bauru	São Paulo	379 297
Itaquaquecetuba	São Paulo	375 011
São Vicente	São Paulo	368 355
Vitória	Espírito Santo	365 855
Caruaru	Pernambuco	365 278
Caucaia	Ceará	365 212
Blumenau	Santa Catarina	361 855
Franca	São Paulo	355 901
Ponta Grossa	Paraná	355 336
Petrolina	Pernambuco	354 317
Canoas	Rio Grande do Sul	348 208
Pelotas	Rio Grande do Sul	343 132
Vitória da Conquista	Bahia	341 128
Ribeirão das Neves	Minas Gerais	338 197
Uberaba	Minas Gerais	337 092
Paulista	Pernambuco	334 376
Cascavel	Paraná	332 333
Praia Grande	São Paulo	330 845
São José dos Pinhais	Paraná	329 058
Guarujá	São Paulo	322 750
Taubaté	São Paulo	317 915
Limeira	São Paulo	308 482
Petrópolis	Rio de Janeiro	306 678
Santarém	Pará	306 480
Palmas	Tocantins	306 296
Camaçari	Bahia	304 302
Mossoró	Rio Grande do Norte	300 618

Suzano	São Paulo	300 559
Taboão da Serra	São Paulo	293 652
Várzea Grande	Mato Grosso	287 526
Sumaré	São Paulo	286 211
Santa Maria	Rio Grande do Sul	283 677
Gravatá	Rio Grande do Sul	283 620
Marabá	Pará	283 542
Governador Valadares	Minas Gerais	281 046
Barueri	São Paulo	276 982
Embu das Artes	São Paulo	276 535
Juazeiro do Norte	Ceará	276 264
Volta Redonda	Rio de Janeiro	273 988
Parnamirim	Rio Grande do Norte	267 036
Ipatinga	Minas Gerais	265 409
Macaé	Rio de Janeiro	261 501
Imperatriz	Maranhão	259 337
Foz do Iguaçu	Paraná	258 248
Viamão	Rio Grande do Sul	256 302
Indaiatuba	São Paulo	256 223
São Carlos	São Paulo	254 484
Cotia	São Paulo	253 608
São José	Santa Catarina	250 181
Novo Hamburgo	Rio Grande do Sul	247 032
Colombo	Paraná	246 540
Magé	Rio de Janeiro	246 433
Itaboraí	Rio de Janeiro	242 543
Americana	São Paulo	242 018
Sete Lagoas	Minas Gerais	241 835
Rio Verde	Goiás	241 518
Itapevi	São Paulo	240 961
Marília	São Paulo	240 590
Divinópolis	Minas Gerais	240 408
São Leopoldo	Rio Grande do Sul	238 648
Araraquara	São Paulo	238 339
Rondonópolis	Mato Grosso	236 042
Jacareí	São Paulo	235 416
Hortolândia	São Paulo	234 259
Arapiraca	Alagoas	233 047
Cabo Frio	Rio de Janeiro	230 378
Presidente Prudente	São Paulo	230 371
Maracanaú	Ceará	229 458
Dourados	Mato Grosso do Sul	225 495
Chapecó	Santa Catarina	224 013

Itajaí	Santa Catarina	223 112
Santa Luzia	Minas Gerais	220 444
Juazeiro	Bahia	218 162
Águas Lindas de Goiás	Goiás	217 698
Criciúma	Santa Catarina	217 311
Itabuna	Bahia	213 685
Parauapebas	Pará	213 576
Rio Grande	Rio Grande do Sul	211 965
Luziânia	Goiás	211 508
Alvorada	Rio Grande do Sul	211 352
Sobral	Ceará	210 711
Cachoeiro de Itapemirim	Espírito Santo	210 589
Cabo de Santo Agostinho	Pernambuco	208 944
Rio Claro	São Paulo	208 008
Angra dos Reis	Rio de Janeiro	207 044
Passo Fundo	Rio Grande do Sul	204 722
Castanhal	Pará	203 251
Lauro de Freitas	Bahia	201 635

Anexo 2 – Modelo utilizando a função “*wbm*’ implementado no software R

```
wbm(deaths ~ lag(confirmed, 2) + lag(varejo, 11) + lag(mercearia_e_farmacia) + lag(parques,
2) + lag(residencial) + lag(transito) + lag(trabalho) + lag(rt, 5) + lag(respUSO) + lag(respEstab)
+ leito_int | population, data = covid1, family = poisson, model = "within")
```

Anexo 3 – Script R

A seguir, é apresentado o script no R que foi utilizado para limpeza e criação do banco de dados utilizado nesta monografia.

```
##### Dados de casos confirmados, óbitos, mobilidade social e Rt#####
## Casos confirmados e óbitos
data1 <- read.csv("data_municipios_jessica.csv", sep = ";", dec = ",", header = TRUE)
str(data1)
data1 <- data1[order(data1$epi_week, decreasing = FALSE), ]
data1 <- data1[order(data1$name, decreasing = FALSE), ]
```

```

# quantos dias de notificação tem em cada semana epidemiológica?
qtdsemanas <- data1 %>%
  group_by(name, epi_week) %>%
  summarise(total = length(epi_week))

qtdsemanas <- as.data.frame(qtdsemanas)

# Casos confirmados por SE
a1 <- data.frame(matrix(0, ncol = 3, nrow = ((53-12+26)*138)))
colnames(a1) <- c("name", "epi_week", "confirmed")

k <- 0
for(i in 1:((53-12+26)*138)){
  a1[i,1] <- data1$name[qtdsemanas[i,3]+k]
  a1[i,2] <- data1$epi_week[qtdsemanas[i,3]+k]
  a1[i,3] <- data1$confirmed[qtdsemanas[i,3]+k]
  k <- k + qtdsemanas[i,3]
}

a1$epi_week <- as.character(a1$epi_week)

a1 <- a1[order(a1$epi_week, decreasing = FALSE), ]
a1 <- a1[order(a1$name, decreasing = FALSE), ]

# Óbitos por SE
b1 <- data.frame(matrix(0, ncol = 3, nrow = ((53-12+26)*138)))
colnames(b1) <- c("name", "epi_week", "confirmed")

k <- 0

```

```

for(i in 1:((53-12+26)*138)){
  b1[i,1] <- data1$name[qtdsemanas[i,3]+k]
  b1[i,2] <- data1$epi_week[qtdsemanas[i,3]+k]
  b1[i,3] <- data1$deaths[qtdsemanas[i,3]+k]
  k <- k + qtdsemanas[i,3]
}

b1$epi_week <- as.character(b1$epi_week)

b1 <- b1[order(b1$epi_week, decreasing = FALSE), ]
b1 <- b1[order(b1$name, decreasing = FALSE), ]

## Criando um data frame auxiliar para inserir as SE que faltam
unique(data1$epi_week) # faltam as 12 primeiro SE de 2020

aux1 <- as.data.frame(matrix(0, ncol = length(data1), nrow = 12*138))
colnames(aux1) <- names(data1)

aux1$code <- rep(unique(data1$code), rep(12, 138))
aux1$population <- rep(unique(data1$population), rep(12, 138))
aux1$id <- rep(unique(data1$id), rep(12, 138))
aux1$code_saude <- rep(unique(data1$code_saude), rep(12, 138))
aux1$name <- rep(unique(data1$name), rep(12, 138))

# SE 2020
epi20 <- read.csv("epi20.csv", sep = ";", header = TRUE)
epi20$Início <- as.Date(epi20$Início)
epi20$Término <- as.Date(epi20$Término)

aux1$epi_week <- rep(epi20[1:12,4], 138)

```

```
## Criando outro data frame auxiliar para inserir os dados da 13 SE de 2020 a 26 SE de 2021
```

```
aux11 <- as.data.frame(matrix(0, ncol = length(data1), nrow = 67*138))
```

```
colnames(aux11) <- names(data1)
```

```
aux11$code <- rep(unique(data1$code), rep(67, 138)) # 53-12+26
```

```
aux11$population <- rep(unique(data1$population), rep(67, 138))
```

```
aux11$id <- rep(unique(data1$id), rep(67, 138))
```

```
aux11$code_saude <- rep(unique(data1$code_saude), rep(67, 138))
```

```
aux11$name <- rep(unique(data1$name), rep(67, 138))
```

```
aux11$epi_week <- rep(unique(data1$epi_week), 138)
```

```
aux11$confirmed <- a1$confirmed
```

```
aux11$deaths <- b1$confirmed
```

```
## Juntado os data frames "aux" e "aux11"
```

```
casos <- merge(aux1, aux11, all = TRUE)
```

```
casos <- casos[order(casos$epi_week, decreasing = FALSE), ]
```

```
casos <- casos[order(casos$name, decreasing = FALSE), ]
```

```
##### Mobilidade Social#####
```

```
data2 <- read.csv("data_municipios_mobility.csv", sep = ";", dec = ",", header = TRUE)
```

```
data2 <- data2[order(data2$date, decreasing = FALSE), ]
```

```
data2 <- data2[order(data2$name, decreasing = FALSE), ]
```

```
# Contando as datas para cada cidades
```

```
repe <- data2 %>%
```

```

group_by(name, date) %>%
count()

# Aquelas cidades que tem duas datas repetidas
repe2 <- repe[(repe$n >= 2),]
repe2 <- as.data.frame(repe2)

# Aquelas cidades que tem três datas repetidas
repe3 <- repe[(repe$n >= 3),]

# Todas as cidades que repetem 3 vezes, repetem 2.
data2_so1 <- data2

for(i in 1:length(unique(repe2$name))){
  data2_so1 <- data2_so1[(data2_so1$name != unique(repe2$name)[i]) == TRUE, ]
}

# Criando uma matrix auxiliar
matriz_aux <- matrix(0, ncol = length(data2), nrow = length(data2$code))
colnames(matriz_aux) <- names(data2)

# Para as cidades com dias duplicados, eu peguei a 2ª aparição do dia ou a 3ª

## Cascavel

cascavel <- data2[(data2$name == "Cascavel") == TRUE, ]

j <- 1
for(i in 1:length(cascavel$code)){

```

```

if(cascavel$date[i] == cascavel$date[i+1]){
  matriz_aux[j,1] <- cascavel[i+1,1]
  matriz_aux[j,2] <- cascavel[i+1,2]
  matriz_aux[j,3] <- cascavel[i+1,3]
  matriz_aux[j,4] <- cascavel[i+1,4]

  matriz_aux[j,5] <- cascavel[i+1,5]
  matriz_aux[j,6] <- cascavel[i+1,6]
  matriz_aux[j,7] <- cascavel[i+1,7]
  matriz_aux[j,8] <- cascavel[i+1,8]
  matriz_aux[j,9] <- cascavel[i+1,9]
  matriz_aux[j,10] <- cascavel[i+1,10]
  matriz_aux[j,11] <- cascavel[i+1,11]
  matriz_aux[j,12] <- cascavel[i+1,12]

  j <- j + 1
}
}

repe1 <- repe[(repe$n == 1), ]
repe1C <- repe1[(repe1$name == "Cascavel"),]

w <- cascavel[(cascavel$name == repe1C$name[1]), 1:12]

for(k in 1:length(repe1C$name)){
  matriz_aux[j,1] <- w[(w$date == repe1C$date[k]), 1]
  matriz_aux[j,2] <- w[(w$date == repe1C$date[k]), 2]
  matriz_aux[j,3] <- w[(w$date == repe1C$date[k]), 3]
  matriz_aux[j,4] <- w[(w$date == repe1C$date[k]), 4]

  matriz_aux[j,5] <- w[(w$date == repe1C$date[k]), 5]

```

```

matriz_aux[j,6] <- w[(w$date == repe1C$date[k]), 6]
matriz_aux[j,7] <- w[(w$date == repe1C$date[k]), 7]
matriz_aux[j,8] <- w[(w$date == repe1C$date[k]), 8]
matriz_aux[j,9] <- w[(w$date == repe1C$date[k]), 9]
matriz_aux[j,10] <- w[(w$date == repe1C$date[k]), 10]
matriz_aux[j,11] <- w[(w$date == repe1C$date[k]), 11]
matriz_aux[j,12] <- w[(w$date == repe1C$date[k]), 12]
j <- j + 1
}

```

```
## Palmas
```

```
palmas <- data2[(data2$name == "Palmas") == TRUE, ]
```

```

for(i in 1:length(palmas$code)){
  if(palmas$date[i] == palmas$date[i+1]){
    matriz_aux[j,1] <- palmas[i+1,1]
    matriz_aux[j,2] <- palmas[i+1,2]
    matriz_aux[j,3] <- palmas[i+1,3]
    matriz_aux[j,4] <- palmas[i+1,4]

    matriz_aux[j,5] <- palmas[i+1,5]
    matriz_aux[j,6] <- palmas[i+1,6]
    matriz_aux[j,7] <- palmas[i+1,7]
    matriz_aux[j,8] <- palmas[i+1,8]
    matriz_aux[j,9] <- palmas[i+1,9]
    matriz_aux[j,10] <- palmas[i+1,10]
    matriz_aux[j,11] <- palmas[i+1,11]
    matriz_aux[j,12] <- palmas[i+1,12]
    j <- j + 1
  }
}

```

```

}
}

repe1P <- repe1[(repe1$name == "Palmas"),]

w <- palmas[(palmas$name == repe1P$name[1]), 1:12]

for(k in 1:length(repe1P$name)){
  matriz_aux[j,1] <- w[(w$date == repe1P$date[k]), 1]
  matriz_aux[j,2] <- w[(w$date == repe1P$date[k]), 2]
  matriz_aux[j,3] <- w[(w$date == repe1P$date[k]), 3]
  matriz_aux[j,4] <- w[(w$date == repe1P$date[k]), 4]

  matriz_aux[j,5] <- w[(w$date == repe1P$date[k]), 5]
  matriz_aux[j,6] <- w[(w$date == repe1P$date[k]), 6]
  matriz_aux[j,7] <- w[(w$date == repe1P$date[k]), 7]
  matriz_aux[j,8] <- w[(w$date == repe1P$date[k]), 8]
  matriz_aux[j,9] <- w[(w$date == repe1P$date[k]), 9]
  matriz_aux[j,10] <- w[(w$date == repe1P$date[k]), 10]
  matriz_aux[j,11] <- w[(w$date == repe1P$date[k]), 11]
  matriz_aux[j,12] <- w[(w$date == repe1P$date[k]), 12]
  j <- j + 1
}

## Praia Grande

praia <- data2[(data2$name == "Praia Grande") == TRUE, ]

for(i in 1:length(praia$code)){
  if(praia$date[i] == praia$date[i+1]){

```

```

matriz_aux[j,1] <- praia[i+1,1]
matriz_aux[j,2] <- praia[i+1,2]
matriz_aux[j,3] <- praia[i+1,3]
matriz_aux[j,4] <- praia[i+1,4]

matriz_aux[j,5] <- praia[i+1,5]
matriz_aux[j,6] <- praia[i+1,6]
matriz_aux[j,7] <- praia[i+1,7]
matriz_aux[j,8] <- praia[i+1,8]
matriz_aux[j,9] <- praia[i+1,9]
matriz_aux[j,10] <- praia[i+1,10]
matriz_aux[j,11] <- praia[i+1,11]
matriz_aux[j,12] <- praia[i+1,12]

j <- j + 1
}
}

repe1PG <- repe1[(repe1$name == "Praia Grande"),]

w <- praia[(praia$name == repe1PG$name[1]), 1:12]

for(k in 1:length(repe1PG$name)){
  matriz_aux[j,1] <- w[(w$date == repe1PG$date[k]), 1]
  matriz_aux[j,2] <- w[(w$date == repe1PG$date[k]), 2]
  matriz_aux[j,3] <- w[(w$date == repe1PG$date[k]), 3]
  matriz_aux[j,4] <- w[(w$date == repe1PG$date[k]), 4]

  matriz_aux[j,5] <- w[(w$date == repe1PG$date[k]), 5]
  matriz_aux[j,6] <- w[(w$date == repe1PG$date[k]), 6]
  matriz_aux[j,7] <- w[(w$date == repe1PG$date[k]), 7]

```

```

matriz_aux[j,8] <- w[(w$date == repe1PG$date[k]), 8]
matriz_aux[j,9] <- w[(w$date == repe1PG$date[k]), 9]
matriz_aux[j,10] <- w[(w$date == repe1PG$date[k]), 10]
matriz_aux[j,11] <- w[(w$date == repe1PG$date[k]), 11]
matriz_aux[j,12] <- w[(w$date == repe1PG$date[k]), 12]
j <- j + 1
}

```

```
## Rio claro
```

```
claro <- data2[(data2$name == "Rio Claro") == TRUE, ]
```

```

for(i in 1:length(claro$code)){
  if(claro$date[i] == claro$date[i+1]){
    matriz_aux[j,1] <- claro[i+1,1]
    matriz_aux[j,2] <- claro[i+1,2]
    matriz_aux[j,3] <- claro[i+1,3]
    matriz_aux[j,4] <- claro[i+1,4]

    matriz_aux[j,5] <- claro[i+1,5]
    matriz_aux[j,6] <- claro[i+1,6]
    matriz_aux[j,7] <- claro[i+1,7]
    matriz_aux[j,8] <- claro[i+1,8]
    matriz_aux[j,9] <- claro[i+1,9]
    matriz_aux[j,10] <- claro[i+1,10]
    matriz_aux[j,11] <- claro[i+1,11]
    matriz_aux[j,12] <- claro[i+1,12]
    j <- j + 1
  }
}

```

```

repe1RC <- repe1[(repe1$name == "Rio Claro"),]

w <- claro[(claro$name == repe1RC$name[1]), 1:12]

for(k in 1:length(repe1RC$name)){
  matriz_aux[j,1] <- w[(w$date == repe1RC$date[k]), 1]
  matriz_aux[j,2] <- w[(w$date == repe1RC$date[k]), 2]
  matriz_aux[j,3] <- w[(w$date == repe1RC$date[k]), 3]
  matriz_aux[j,4] <- w[(w$date == repe1RC$date[k]), 4]

  matriz_aux[j,5] <- w[(w$date == repe1RC$date[k]), 5]
  matriz_aux[j,6] <- w[(w$date == repe1RC$date[k]), 6]
  matriz_aux[j,7] <- w[(w$date == repe1RC$date[k]), 7]
  matriz_aux[j,8] <- w[(w$date == repe1RC$date[k]), 8]
  matriz_aux[j,9] <- w[(w$date == repe1RC$date[k]), 9]
  matriz_aux[j,10] <- w[(w$date == repe1RC$date[k]), 10]
  matriz_aux[j,11] <- w[(w$date == repe1RC$date[k]), 11]
  matriz_aux[j,12] <- w[(w$date == repe1RC$date[k]), 12]
  j <- j + 1
}

## São Carlos

carlos <- data2[(data2$name == "São Carlos") == TRUE, ]

for(i in 1:length(carlos$code)){
  if(carlos$date[i] == carlos$date[i+1]){
    matriz_aux[j,1] <- carlos[i+1,1]
    matriz_aux[j,2] <- carlos[i+1,2]
  }
}

```

```

matriz_aux[j,3] <- carlos[i+1,3]
matriz_aux[j,4] <- carlos[i+1,4]

matriz_aux[j,5] <- carlos[i+1,5]
matriz_aux[j,6] <- carlos[i+1,6]
matriz_aux[j,7] <- carlos[i+1,7]
matriz_aux[j,8] <- carlos[i+1,8]
matriz_aux[j,9] <- carlos[i+1,9]
matriz_aux[j,10] <- carlos[i+1,10]
matriz_aux[j,11] <- carlos[i+1,11]
matriz_aux[j,12] <- carlos[i+1,12]
j <- j + 1
}
}

repe1SC <- repe1[(repe1$name == "São Carlos"),]

w <- carlos[(carlos$name == repe1SC$name[1]), 1:12]

for(k in 1:length(repe1SC$name)){
  matriz_aux[j,1] <- w[(w$date == repe1SC$date[k]), 1]
  matriz_aux[j,2] <- w[(w$date == repe1SC$date[k]), 2]
  matriz_aux[j,3] <- w[(w$date == repe1SC$date[k]), 3]
  matriz_aux[j,4] <- w[(w$date == repe1SC$date[k]), 4]

  matriz_aux[j,5] <- w[(w$date == repe1SC$date[k]), 5]
  matriz_aux[j,6] <- w[(w$date == repe1SC$date[k]), 6]
  matriz_aux[j,7] <- w[(w$date == repe1SC$date[k]), 7]
  matriz_aux[j,8] <- w[(w$date == repe1SC$date[k]), 8]
}

```

```

matriz_aux[j,9] <- w[(w$date == repe1SC$date[k]), 9]
matriz_aux[j,10] <- w[(w$date == repe1SC$date[k]), 10]
matriz_aux[j,11] <- w[(w$date == repe1SC$date[k]), 11]
matriz_aux[j,12] <- w[(w$date == repe1SC$date[k]), 12]
j <- j + 1
}

```

```
## Santa Luzia
```

```
# tem dias que repete 1 vez, tem dias que repete duas
```

```
sl <- repe2[(repe2$name == "Santa Luzia") == TRUE, ]
```

```
sl2 <- sl[(sl$n == 2) == TRUE, ]
```

```
sl3 <- sl[(sl$n == 3) == TRUE, ]
```

```
luzia <- data2[(data2$name == "Santa Luzia") == TRUE, ]
```

```
luzia2 <- luzia[(luzia$date %in% sl2$date) == TRUE, ]
```

```
luzia3 <- luzia[(luzia$date %in% sl3$date) == TRUE, ]
```

```
for(i in 1:length(luzia2$code)){
  if(luzia2$date[i] == luzia2$date[i+1]){
```

```
    matriz_aux[j,1] <- luzia2[i+1,1]
```

```
    matriz_aux[j,2] <- luzia2[i+1,2]
```

```
    matriz_aux[j,3] <- luzia2[i+1,3]
```

```
    matriz_aux[j,4] <- luzia2[i+1,4]
```

```
    matriz_aux[j,5] <- luzia2[i+1,5]
```

```
    matriz_aux[j,6] <- luzia2[i+1,6]
```

```
    matriz_aux[j,7] <- luzia2[i+1,7]
```

```
    matriz_aux[j,8] <- luzia2[i+1,8]
```

```
    matriz_aux[j,9] <- luzia2[i+1,9]
```

```

    matriz_aux[j,10] <- luzia2[i+1,10]
    matriz_aux[j,11] <- luzia2[i+1,11]
    matriz_aux[j,12] <- luzia2[i+1,12]
    j <- j + 1
  }
}

for(i in 1:length(luzia3$code)){
  if(luzia3$date[i] == luzia3$date[i+2]){
    matriz_aux[j,1] <- luzia3[i+1,1]
    matriz_aux[j,2] <- luzia3[i+1,2]
    matriz_aux[j,3] <- luzia3[i+1,3]
    matriz_aux[j,4] <- luzia3[i+1,4]

    matriz_aux[j,5] <- luzia3[i+2,5]
    matriz_aux[j,6] <- luzia3[i+2,6]
    matriz_aux[j,7] <- luzia3[i+2,7]
    matriz_aux[j,8] <- luzia3[i+2,8]
    matriz_aux[j,9] <- luzia3[i+2,9]
    matriz_aux[j,10] <- luzia3[i+2,10]
    matriz_aux[j,11] <- luzia3[i+2,11]
    matriz_aux[j,12] <- luzia3[i+2,12]
    j <- j + 1
  }
}

repe1SL <- repe1[(repe1$name == "Santa Luzia"),]

w <- luzia[(luzia$name == repe1SL$name[1]), 1:12]

```

```

for(k in 1:length(repe1SL$name)){
  matriz_aux[j,1] <- w[(w$date == repe1SL$date[k]), 1]
  matriz_aux[j,2] <- w[(w$date == repe1SL$date[k]), 2]
  matriz_aux[j,3] <- w[(w$date == repe1SL$date[k]), 3]
  matriz_aux[j,4] <- w[(w$date == repe1SL$date[k]), 4]

  matriz_aux[j,5] <- w[(w$date == repe1SL$date[k]), 5]
  matriz_aux[j,6] <- w[(w$date == repe1SL$date[k]), 6]
  matriz_aux[j,7] <- w[(w$date == repe1SL$date[k]), 7]
  matriz_aux[j,8] <- w[(w$date == repe1SL$date[k]), 8]
  matriz_aux[j,9] <- w[(w$date == repe1SL$date[k]), 9]
  matriz_aux[j,10] <- w[(w$date == repe1SL$date[k]), 10]
  matriz_aux[j,11] <- w[(w$date == repe1SL$date[k]), 11]
  matriz_aux[j,12] <- w[(w$date == repe1SL$date[k]), 12]
  j <- j + 1
}

# Como fiz uma matriz grande demais, estou resumindo ela
matriz_aux <- matriz_aux[1:j-1,]

# Juntando "data2_so1" com "matriz_aux"
data2_modificado <- merge(data2_so1, matriz_aux, all = TRUE)
data2_modificado <- data2_modificado[order(data2_modificado$date, decreasing = FALSE),
]
data2_modificado <- data2_modificado[order(data2_modificado$name, decreasing = FALSE),
]

data2_modificado$epi_week <- 0

```

```
## Inserindo a SE
```

```
week2 <- c("01", "02", "03", "04", "05", "06", "07", "08", "09")
```

```
ano20 <- rep("2020", 53)
```

```
ep20 <- c(week2, epi20$Semana[10:53])
```

```
semana_ep20 <- paste(ano20, ep20, sep = ".")
```

```
# SE 2021
```

```
eipi21 <- read.csv("eipi21.csv", sep = ";", header = TRUE)
```

```
eipi21$Início <- as.Date(eipi21$Início)
```

```
eipi21$Término <- as.Date(eipi21$Término)
```

```
ano21 <- rep("2021", 26)
```

```
ep21 <- c(week2, eipi21$Semana[10:26])
```

```
semana_ep21 <- paste(ano21, ep21, sep = ".")
```

```
for(i in 1:length(data2_modificado$name)){
```

```
  if((year(data2_modificado$date[i]) == 2020) == TRUE){
```

```
    for(j in 1:length(epi20$Semana)){
```

```
      if(data2_modificado$date[i] >= epi20$Início[j] & data2_modificado$date[i] <= epi20$Término[j]){
```

```
        data2_modificado$epi_week[i] <- semana_ep20[j]
```

```
      }
```

```
    }
```

```
  }
```

```
  else{
```

```
    for(j in 1:length(eipi21$Semana)){
```

```
      if(data2_modificado$date[i] >= eipi21$Início[j] & data2_modificado$date[i] <= eipi21$Término[j]){
```

```

    data2_modificado$epi_week[i] <- semana_ep21[j]
  }
}
}
if(data2_modificado$date[i] == "2021-01-01" || data2_modificado$date[i] == "2021-01-02"){
  data2_modificado$epi_week[i] <- "2020.53"
}
}

```

Transformando em numérico

```

data2_modificado$varejo_e_recreacao <- as.numeric(data2_modificado$varejo_e_recreacao)
data2_modificado$mercearia_e_farmacia <-
as.numeric(data2_modificado$mercearia_e_farmacia)
data2_modificado$parques <- as.numeric(data2_modificado$parques)
data2_modificado$estacoes_de_transito <-
as.numeric(data2_modificado$estacoes_de_transito)
data2_modificado$locais_de_trabalho <- as.numeric(data2_modificado$locais_de_trabalho)
data2_modificado$residencial <- as.numeric(data2_modificado$residencial)

```

Média da Mobility por semana epidemiológica

Varejo e Recreação

```

c11 <- data2_modificado %>%
  group_by(name, epi_week) %>%
  summarise(Total = mean(varejo_e_recreacao, na.rm = TRUE),
            Count = n())

```

Mercearia e Farmácia

```

c22 <- data2_modificado %>%
  group_by(name, epi_week) %>%

```

```
summarise(Total = mean(mercearia_e_farmacia, na.rm = TRUE),  
          Count = n())
```

```
# Parques
```

```
c33 <- data2_modificado %>%  
  group_by(name, epi_week) %>%  
  summarise(Total = mean(parques, na.rm = TRUE),  
            Count = n())
```

```
# Estações de Trânsito
```

```
c44 <- data2_modificado %>%  
  group_by(name, epi_week) %>%  
  summarise(Total = mean(estacoes_de_transito, na.rm = TRUE),  
            Count = n())
```

```
# Locais de Trabalho
```

```
c55 <- data2_modificado %>%  
  group_by(name, epi_week) %>%  
  summarise(Total = mean(locais_de_trabalho, na.rm = TRUE),  
            Count = n())
```

```
# Residencial
```

```
c66 <- data2_modificado %>%  
  group_by(name, epi_week) %>%  
  summarise(Total = mean(residencial, na.rm = TRUE),  
            Count = n())
```

```
# quantidade de cidades em cada mes
```

```
data2_modificado %>%  
  group_by(year(data2_modificado$date), month(data2_modificado$date)) %>%
```

```
summarise(total = length(unique(name)))
```

```
# inserindo as SE que faltam das cidades
```

```
unique(data2_modificado$epi_week) # inserir as 6 primeiras SE de 2020
```

```
aux22 <- as.data.frame(matrix(0, ncol = length(data2_modificado), nrow = 6*136))
```

```
colnames(aux22) <- names(data2_modificado)
```

```
aux22$code <- rep(unique(data2_modificado$code), rep(6, 136))
```

```
aux22$population <- rep(unique(data2_modificado$population), rep(6, 136))
```

```
aux22$id <- rep(unique(data2_modificado$id), rep(6, 136))
```

```
aux22$code_saude <- rep(unique(data2_modificado$code_saude), rep(6, 136))
```

```
aux22$name <- rep(unique(data2_modificado$name), rep(6, 136))
```

```
aux22$epi_week <- rep(unique(epi20$SE[1:6]), 136)
```

```
# Criando um data frame com os resultados encontrado anteriormente
```

```
aux2 <- as.data.frame(matrix(0, ncol = length(data2_modificado), nrow = 73*136))
```

```
colnames(aux2) <- names(data2_modificado)
```

```
aux2$code <- rep(unique(data2_modificado$code), rep(73, 136))
```

```
aux2$population <- rep(unique(data2_modificado$population), rep(73, 136))
```

```
aux2$id <- rep(unique(data2_modificado$id), rep(73, 136))
```

```
aux2$code_saude <- rep(unique(data2_modificado$code_saude), rep(73, 136))
```

```

aux2$name <- rep(unique(data2_modificado$name), rep(73, 136))
aux2$epi_week <- rep(unique(data2_modificado$epi_week), 136)

aux2$varejo_e_recreacao <- c11$Total
aux2$mercearia_e_farmacia <- c22$Total
aux2$parques <- c33$Total
aux2$estacoes_de_transito <- c44$Total
aux2$locais_de_trabalho <- c55$Total
aux2$residencial <- c66$Total

# juntando "aux22" e "aux2"
aux2x2 <- merge(aux22, aux2, all = TRUE)
aux2x2 <- aux2x2[order(aux2x2$epi_week, decreasing = FALSE), ]
aux2x2 <- aux2x2[order(aux2x2$name, decreasing = FALSE), ]
mobilidade_dados <- aux2x2

# verificando quais cidades não tem dados em alguma categoria de mobilidade
somamob <- mobilidade_dados %>%
  group_by(name) %>%
  summarise(varejo = mean(varejo_e_recreacao, na.rm = TRUE),
            mercearia = mean(mercearia_e_farmacia, na.rm = TRUE),
            parques = mean(parques, na.rm = TRUE),
            estacoes = mean(estacoes_de_transito, na.rm = TRUE),
            trabalho = mean(locais_de_trabalho, na.rm = TRUE),
            residencia = mean(residencial, na.rm = TRUE),
            Count = n())

somamob[(somamob$varejo == 0) == TRUE, ] # todos tem dados
somamob[(somamob$mercearia == 0) == TRUE, ] # todos tem dados
somamob[(somamob$parques == 0) == TRUE, ] # 1 cidade não tem

```

```
somamob[(somamob$estacoes == 0) == TRUE, ] # 3 cidades não tem
somamob[(somamob$trabalho == 0) == TRUE, ] # todos tem dados
somamob[(somamob$residencia == 0) == TRUE, ] # todos tem dados
```

```
## inserindo as cidades embu das artes e são José dos campos - elas não tem nenhum dado de
mobilidade
```

```
# COLOCAQUEI TUDO ZERO PARA FICAR MAIS FACIL DE MUDAR DEPOIS
```

```
lnomes2 <- c("Embu das Artes", "São José dos Campos")
```

```
aux21 <- data.frame(matrix(0, ncol = length(data2_modificado), nrow = 2*79))
```

```
colnames(aux21) <- names(data2_modificado)
```

```
aux21$code <- c(data1[(data1$name == lnomes2[1]) == TRUE, 1][1:79],
               data1[(data1$name == lnomes2[2]) == TRUE, 1][1:79])
```

```
aux21$population <- c(data1[(data1$name == lnomes2[1]) == TRUE, 2][1:79],
                     data1[(data1$name == lnomes2[2]) == TRUE, 2][1:79])
```

```
aux21$id <- c(data1[(data1$name == lnomes2[1]) == TRUE, 3][1:79],
              data1[(data1$name == lnomes2[2]) == TRUE, 3][1:79])
```

```
aux21$code_saude <- c(data1[(data1$name == lnomes2[1]) == TRUE, 4][1:79],
                     data1[(data1$name == lnomes2[2]) == TRUE, 4][1:79])
```

```
aux21$name <- rep(lnomes2, rep(79,2))
```

```
aux21$epi_week <- c(epi20$SE, epi21$SE[1:26], epi20$SE, epi21$SE[1:26])
```

```
# juntando "mobilidade_dados" e "aux21"
```

```
mobilidadeDados <- merge(mobilidade_dados, aux21, all = TRUE)
```

```
## Inserindo as mobilidade dos estados, para os municípios que não tem nenhum dado em  
alguma categoria
```

```
# Os municípios são: Santa Luzia e Suzano (SP) - Parques
```

```
# Águas Lindas de Goiás (GO), Lauro de Freitas (BA) e Sobral (CE) - Estações de  
Trânsito
```

```
mobilidade <- read.csv("jessica_mobility_ufs.csv", sep = ";", header = TRUE)
```

```
str(mobilidade)
```

```
# pegando somente dados até 3 de julho de 2021
```

```
mobilidade <- mobilidade[(mobilidade$date < "2021-07-04") == TRUE, ]
```

```
# criando uma nova coluna para as SE
```

```
mobilidade$epi_week <- 0
```

```
## inserindo as semanas epidemiológicas
```

```
week2 <- c("01", "02", "03", "04", "05", "06", "07", "08", "09")
```

```
ano20 <- rep("2020", 53)
```

```
ep20 <- c(week2, epi20$Semana[10:53])
```

```
semana_ep20 <- paste(ano20, ep20, sep = ".")
```

```
ano21 <- rep("2021", 26)
```

```

ep21 <- c(week2, epi21$Semana[10:26])
semana_ep21 <- paste(ano21, ep21, sep = ".")

for(i in 1:length(mobilidade$name)){
  if((year(mobilidade$date[i]) == 2020) == TRUE){
    for(j in 1:length(epi20$Semana)){
      if(mobilidade$date[i] >= epi20$Inicio[j] & mobilidade$date[i] <= epi20$Termino[j]){
        mobilidade$epi_week[i] <- semana_ep20[j]
      }
    }
  }
  else{
    for(j in 1:length(epi21$Semana)){
      if(mobilidade$date[i] >= epi21$Inicio[j] & mobilidade$date[i] <= epi21$Termino[j]){
        mobilidade$epi_week[i] <- semana_ep21[j]
      }
    }
  }
  if(mobilidade$date[i] == "2021-01-01" || mobilidade$date[i] == "2021-01-02"){
    mobilidade$epi_week[i] <- "2020.53"
  }
}

# média para cada categoria da mobilidade
mob <- mobilidade %>%
  group_by(name, epi_week) %>%
  summarise(varejo = mean(varejo_e_recreacao),
            mercearia = mean(mercearia_e_farmacia),
            parques = mean(parques),
            estacoes = mean(estacoes_de_transito),

```

```

trabalho = mean(locais_de_trabalho),
residencia = mean(residencial),
Count = n())

# inserindo zero nas primeiras 6 semanas - pois não possui dados
banco_aux <- data.frame(matrix(0, ncol = 9, nrow = 4*6))
colnames(banco_aux) <- names(mob)
banco_aux$name <- rep(unique(mob$name), rep(6,4))
banco_aux$epi_week <- rep(semana_ep20[1:6], 4)

# juntando "banco_aux" e "mob"
mobilidadeSocial <- merge(banco_aux, mob, all = TRUE)

### Inserindo os dados criados nas respectivas cidades

# Embu das Artes e São José dos Campos não tem nenhum dado, eles são de SP
mobilidadeDados[(mobilidadeDados$name == "Embu das Artes") == TRUE, 7:12] <-
mobilidadeSocial[(mobilidadeSocial$name == "São Paulo") == TRUE, 3:8]
mobilidadeDados[(mobilidadeDados$name == "São José dos Campos") == TRUE, 7:12] <-
mobilidadeSocial[(mobilidadeSocial$name == "São Paulo") == TRUE, 3:8]

# Suzano não tem "parques", é de SP também
mobilidadeDados[(mobilidadeDados$name == "Suzano") == TRUE, 9] <-
mobilidadeSocial[(mobilidadeSocial$name == "São Paulo") == TRUE, 5]

# Águas lindas de Goiás (GO), Lauro de Freitas (BA), Sobral (CE) não tem "estações de
trânsito"
mobilidadeDados[(mobilidadeDados$name == "Águas Lindas de Goiás") == TRUE, 10] <-
mobilidadeSocial[(mobilidadeSocial$name == "Goiás") == TRUE, 6]
mobilidadeDados[(mobilidadeDados$name == "Lauro de Freitas") == TRUE, 10] <-
mobilidadeSocial[(mobilidadeSocial$name == "Bahia") == TRUE, 6]

```

```

mobilidadeDados[(mobilidadeDados$name == "Sobral") == TRUE, 10] <-
mobilidadeSocial[(mobilidadeSocial$name == "Ceará") == TRUE, 6]

```

```
##### Algumas cidades não possuem dados de mobilidade para alguma SE
```

```
#então iremos replicar os dados da semana anterior para as próximas que não possuem dados
```

```
# Adicionando uma coluna apenas para ficar mais fácil a replicação dos valores
```

```
mobilidadeDados$linha <- c(1:10902)
```

```
## Varejo e recreação
```

```
# Cidades que possuem dados de varejo e recreação faltantes
```

```
unique(mobilidadeDados[(mobilidadeDados$varejo_e_recreacao == "NaN") == TRUE, 5])
```

```
varejoDados <- mobilidadeDados[(mobilidadeDados$varejo_e_recreacao == "NaN") ==
TRUE, c(5,7:12)]
```

```
# quantos valores faltantes existem em cada cidade
```

```
varejoQtd <- varejoDados %>%
```

```
  group_by(name) %>%
```

```
  summarise(total = length(name))
```

```
nomes <- unique(mobilidadeDados[(mobilidadeDados$varejo_e_recreacao == "NaN") ==
TRUE, 5])
```

```
for(i in 1:length(nomes)){
```

```
  parquesub <- mobilidadeDados[(mobilidadeDados$name == nomes[i]) == TRUE, c(5,7,14)]
```

```
  d <- parquesub[(parquesub$varejo_e_recreacao == "NaN") == TRUE, 3]
```

```
  for(j in 1:length(d)){
```

```

    mobilidadeDados[d[j], 7] <- mobilidadeDados[d[j] - 1, 7]
  }
}

unique(mobilidadeDados[(mobilidadeDados$varejo_e_recreacao == "NaN") == TRUE, 5])

### mercearias e farmacias
unique(mobilidadeDados[(mobilidadeDados$mercearia_e_farmacia == "NaN") == TRUE, 5])
merceariaDados <- mobilidadeDados[(mobilidadeDados$mercearia_e_farmacia == "NaN") ==
TRUE, c(5,7:12)]

merceariaQtd <- merceariaDados %>%
  group_by(name) %>%
  summarise(total = length(name))
merceariaQtd

nomes <- unique(mobilidadeDados[(mobilidadeDados$mercearia_e_farmacia == "NaN") ==
TRUE, 5])
for(i in 1:length(nomes)){
  parquesub <- mobilidadeDados[(mobilidadeDados$name == nomes[i]) == TRUE, c(5,8,14)]
  d <- parquesub[(parquesub$mercearia_e_farmacia == "NaN") == TRUE, 3]

  for(j in 1:length(d)){
    mobilidadeDados[d[j], 8] <- mobilidadeDados[d[j] - 1, 8]
  }
}

unique(mobilidadeDados[(mobilidadeDados$mercearia_e_farmacia == "NaN") == TRUE, 5])

```

```

## parques
unique(mobilidadeDados[(mobilidadeDados$parques == "NaN") == TRUE, 5])
parquesDados <- mobilidadeDados[(mobilidadeDados$parques == "NaN") == TRUE,
c(5,7:12)]

parquesQtd <- parquesDados %>%
  group_by(name) %>%
  summarise(total = length(name))
parquesQtd

nomes <- unique(mobilidadeDados[(mobilidadeDados$parques == "NaN") == TRUE, 5])
for(i in 1:length(nomes)){
  parquessub <- mobilidadeDados[(mobilidadeDados$name == nomes[i]) == TRUE, c(5,9,14)]
  d <- parquessub[(parquessub$parques == "NaN") == TRUE, 3]

  for(j in 1:length(d)){
    mobilidadeDados[d[j], 9] <- mobilidadeDados[d[j] - 1, 9]
  }
}

unique(mobilidadeDados[(mobilidadeDados$parques == "NaN") == TRUE, 5])

# estações de transito
unique(mobilidadeDados[(mobilidadeDados$estacoes_de_transito == "NaN") == TRUE, 5])
estacoesDados <- mobilidadeDados[(mobilidadeDados$estacoes_de_transito == "NaN") ==
TRUE, c(5,7:12)]

```

```

estacoesQtd <- estacoesDados %>%
  group_by(name) %>%
  summarise(total = length(name))
estacoesQtd

```

```

nomes <- unique(mobilidadeDados[(mobilidadeDados$estacoes_de_transito == "NaN") ==
TRUE, 5])
for(i in 1:length(nomes)){
  parquesub <- mobilidadeDados[(mobilidadeDados$name == nomes[i]) == TRUE,
c(5,10,14)]
  d <- parquesub[(parquesub$estacoes_de_transito == "NaN") == TRUE, 3]

  for(j in 1:length(d)){
    mobilidadeDados[d[j], 10] <- mobilidadeDados[d[j] - 1, 10]
  }
}

```

```

unique(mobilidadeDados[(mobilidadeDados$estacoes_de_transito == "NaN") == TRUE, 5])

```

```

# locais de trabalho

```

```

unique(mobilidadeDados[(mobilidadeDados$locais_de_trabalho == "NaN") == TRUE, 5])
mobilidadeDados[(mobilidadeDados$locais_de_trabalho == "NaN") == TRUE, c(5,7:12)]

```

```

# residencial

```

```

unique(mobilidadeDados[(mobilidadeDados$residencial == "NaN") == TRUE, 5])
mobilidadeDados[(mobilidadeDados$residencial == "NaN") == TRUE, c(5,7:12)]

```

```

mobilidadeDados <- mobilidadeDados[order(mobilidadeDados$epi_week, decreasing =
FALSE), ]

mobilidadeDados <- mobilidadeDados[order(mobilidadeDados$name, decreasing = FALSE),
]

#####Rt estimado#####

data3 <- read.csv("data_municipios_rt.csv", sep = ";", dec = ",", header = TRUE)

# Verificando as cidades que repetem os dias
rt2 <- data3 %>%
  group_by(name, date) %>%
  count()

# Cidades com a data repetida 2 vezes
rt22 <- rt2[(rt2$n >= 2) == TRUE, ]
unique(rt22$name) # são as mesmas cidades que tem dias repetidos para mobilidade

rt1 <- rt2[(rt2$n == 1), ]

# Tirando as cidades que possuem dados repetidos
data3_so1 <- data3

for(i in 1:length(unique(rt22$name))){
  data3_so1 <- data3_so1[(data3_so1$name != unique(rt22$name)[i]) == TRUE, ]
}

## retirando os dias repetidos - vou pegar a primeira aparição de cada data repetida

```

```
rt_aux <- matrix(0, ncol = length(data3), nrow = length(data3$code))
colnames(rt_aux) <- names(data3)
```

```
## Cascavel
```

```
cascavel <- data3[(data3$name == "Cascavel") == TRUE, ]
```

```
j <- 1
```

```
for(i in 1:length(cascavel$code)){
  if(cascavel$date[i] == cascavel$date[i+1]){
    rt_aux[j,1] <- cascavel[i,1]
    rt_aux[j,2] <- cascavel[i,2]
    rt_aux[j,3] <- cascavel[i,3]
    rt_aux[j,4] <- cascavel[i,4]

    rt_aux[j,5] <- cascavel[i,5]
    rt_aux[j,6] <- cascavel[i,6]
    rt_aux[j,7] <- cascavel[i,7]
    j <- j + 1
  }
}
```

```
rt1C <- rt1[(rt1$name == "Cascavel"),]
```

```
w <- cascavel[(cascavel$name == rt1C$name[1]), 1:7]
```

```
for(k in 1:length(rt1C$name)){
  rt_aux[j,1] <- w[(w$date == rt1C$date[k]), 1]
  rt_aux[j,2] <- w[(w$date == rt1C$date[k]), 2]
```

```

rt_aux[j,3] <- w[(w$date == rt1C$date[k]), 3]
rt_aux[j,4] <- w[(w$date == rt1C$date[k]), 4]

rt_aux[j,5] <- w[(w$date == rt1C$date[k]), 5]
rt_aux[j,6] <- w[(w$date == rt1C$date[k]), 6]
rt_aux[j,7] <- w[(w$date == rt1C$date[k]), 7]
j <- j + 1
}

## Palmas

palmas <- data3[(data3$name == "Palmas") == TRUE, ]

for(i in 1:length(palmas$code)){
  if(palmas$date[i] == palmas$date[i+1]){
    rt_aux[j,1] <- palmas[i,1]
    rt_aux[j,2] <- palmas[i,2]
    rt_aux[j,3] <- palmas[i,3]
    rt_aux[j,4] <- palmas[i,4]

    rt_aux[j,5] <- palmas[i,5]
    rt_aux[j,6] <- palmas[i,6]
    rt_aux[j,7] <- palmas[i,7]
    j <- j + 1
  }
}

rt1P <- rt1[(rt1$name == "Palmas"),]

```

```

w <- palmas[(palmas$name == rt1P$name[1]), 1:7]

for(k in 1:length(rt1P$name)){
  rt_aux[j,1] <- w[(w$date == rt1P$date[k]), 1]
  rt_aux[j,2] <- w[(w$date == rt1P$date[k]), 2]
  rt_aux[j,3] <- w[(w$date == rt1P$date[k]), 3]
  rt_aux[j,4] <- w[(w$date == rt1P$date[k]), 4]

  rt_aux[j,5] <- w[(w$date == rt1P$date[k]), 5]
  rt_aux[j,6] <- w[(w$date == rt1P$date[k]), 6]
  rt_aux[j,7] <- w[(w$date == rt1P$date[k]), 7]
  j <- j + 1
}

## Praia Grande

praia <- data3[(data3$name == "Praia Grande") == TRUE, ]

for(i in 1:length(praia$code)){
  if(praia$date[i] == praia$date[i+1]){
    rt_aux[j,1] <- praia[i,1]
    rt_aux[j,2] <- praia[i,2]
    rt_aux[j,3] <- praia[i,3]
    rt_aux[j,4] <- praia[i,4]

    rt_aux[j,5] <- praia[i,5]
    rt_aux[j,6] <- praia[i,6]
    rt_aux[j,7] <- praia[i,7]
    j <- j + 1
  }
}

```

```
}
```

```
rt1PG <- rt1[(rt1$name == "Praia Grande"),]
```

```
w <- praia[(praia$name == rt1PG$name[1]), 1:7]
```

```
for(k in 1:length(rt1PG$name)){
```

```
  rt_aux[j,1] <- w[(w$date == rt1PG$date[k]), 1]
```

```
  rt_aux[j,2] <- w[(w$date == rt1PG$date[k]), 2]
```

```
  rt_aux[j,3] <- w[(w$date == rt1PG$date[k]), 3]
```

```
  rt_aux[j,4] <- w[(w$date == rt1PG$date[k]), 4]
```

```
  rt_aux[j,5] <- w[(w$date == rt1PG$date[k]), 5]
```

```
  rt_aux[j,6] <- w[(w$date == rt1PG$date[k]), 6]
```

```
  rt_aux[j,7] <- w[(w$date == rt1PG$date[k]), 7]
```

```
  j <- j + 1
```

```
}
```

```
## Rio claro
```

```
claro <- data3[(data3$name == "Rio Claro") == TRUE, ]
```

```
for(i in 1:length(claro$code)){
```

```
  if(claro$date[i] == claro$date[i+1]){
```

```
    rt_aux[j,1] <- claro[i,1]
```

```
    rt_aux[j,2] <- claro[i,2]
```

```
    rt_aux[j,3] <- claro[i,3]
```

```
    rt_aux[j,4] <- claro[i,4]
```

```

    rt_aux[j,5] <- claro[i,5]
    rt_aux[j,6] <- claro[i,6]
    rt_aux[j,7] <- claro[i,7]
    j <- j + 1
  }
}

rt1RC <- rt1[(rt1$name == "Rio Claro"),]

w <- claro[(claro$name == rt1RC$name[1]), 1:7]

for(k in 1:length(rt1RC$name)){
  rt_aux[j,1] <- w[(w$date == rt1RC$date[k]), 1]
  rt_aux[j,2] <- w[(w$date == rt1RC$date[k]), 2]
  rt_aux[j,3] <- w[(w$date == rt1RC$date[k]), 3]
  rt_aux[j,4] <- w[(w$date == rt1RC$date[k]), 4]

  rt_aux[j,5] <- w[(w$date == rt1RC$date[k]), 5]
  rt_aux[j,6] <- w[(w$date == rt1RC$date[k]), 6]
  rt_aux[j,7] <- w[(w$date == rt1RC$date[k]), 7]
  j <- j + 1
}

## São Carlos

carlos <- data3[(data3$name == "São Carlos") == TRUE, ]

for(i in 1:length(carlos$code)){
  if(carlos$date[i] == carlos$date[i+1]){

```

```

rt_aux[j,1] <- carlos[i,1]
rt_aux[j,2] <- carlos[i,2]
rt_aux[j,3] <- carlos[i,3]
rt_aux[j,4] <- carlos[i,4]

rt_aux[j,5] <- carlos[i,5]
rt_aux[j,6] <- carlos[i,6]
rt_aux[j,7] <- carlos[i,7]
j <- j + 1
}
}

rt1SC <- rt1[(rt1$name == "São Carlos"),]

w <- carlos[(carlos$name == rt1SC$name[1]), 1:7]

for(k in 1:length(rt1SC$name)){
  rt_aux[j,1] <- w[(w$date == rt1SC$date[k]), 1]
  rt_aux[j,2] <- w[(w$date == rt1SC$date[k]), 2]
  rt_aux[j,3] <- w[(w$date == rt1SC$date[k]), 3]
  rt_aux[j,4] <- w[(w$date == rt1SC$date[k]), 4]

  rt_aux[j,5] <- w[(w$date == rt1SC$date[k]), 5]
  rt_aux[j,6] <- w[(w$date == rt1SC$date[k]), 6]
  rt_aux[j,7] <- w[(w$date == rt1SC$date[k]), 7]
  j <- j + 1
}

## Santa Luzia

# tem dias que repete 1 vez, tem dias que repete duas

```

```

sl <- rt2[(rt2$name == "Santa Luzia") == TRUE, ]
sl2 <- sl[(sl$n == 2) == TRUE, ]
sl3 <- sl[(sl$n == 3) == TRUE, ]

luzia <- data3[(data3$name == "Santa Luzia") == TRUE, ]
luzia2 <- luzia[(luzia$date %in% sl2$date) == TRUE, ]
luzia3 <- luzia[(luzia$date %in% sl3$date) == TRUE, ]

for(i in 1:length(luzia2$code)){
  if(luzia2$date[i] == luzia2$date[i+1]){
    rt_aux[j,1] <- luzia2[i,1]
    rt_aux[j,2] <- luzia2[i,2]
    rt_aux[j,3] <- luzia2[i,3]
    rt_aux[j,4] <- luzia2[i,4]

    rt_aux[j,5] <- luzia2[i,5]
    rt_aux[j,6] <- luzia2[i,6]
    rt_aux[j,7] <- luzia2[i,7]
    j <- j + 1
  }
}

for(i in 1:length(luzia3$code)){
  if(luzia3$date[i] == luzia3$date[i+2]){
    rt_aux[j,1] <- luzia3[i,1]
    rt_aux[j,2] <- luzia3[i,2]
    rt_aux[j,3] <- luzia3[i,3]
    rt_aux[j,4] <- luzia3[i,4]

```

```

    rt_aux[j,5] <- luzia3[i,5]
    rt_aux[j,6] <- luzia3[i,6]
    rt_aux[j,7] <- luzia3[i,7]
    j <- j + 1
  }
}

rt1L <- rt1[(rt1$name == "Santa Luzia"),]

w <- luzia[(luzia$name == rt1L$name[1]), 1:7]

for(k in 1:length(rt1L$name)){
  rt_aux[j,1] <- w[(w$date == rt1L$date[k]), 1]
  rt_aux[j,2] <- w[(w$date == rt1L$date[k]), 2]
  rt_aux[j,3] <- w[(w$date == rt1L$date[k]), 3]
  rt_aux[j,4] <- w[(w$date == rt1L$date[k]), 4]

  rt_aux[j,5] <- w[(w$date == rt1L$date[k]), 5]
  rt_aux[j,6] <- w[(w$date == rt1L$date[k]), 6]
  rt_aux[j,7] <- w[(w$date == rt1L$date[k]), 7]
  j <- j + 1
}

# Ajustando minha matrix auxiliar
rt_aux <- rt_aux[1:j-1,]

# juntando "data3_so1" (que tem os dados originais) e "rt_aux" (que tirei os repetidos)
data3_modificado <- merge(data3_so1, rt_aux, all = TRUE)
data3_modificado <- data3_modificado[order(data3_modificado$name, decreasing = FALSE),
]

```

```

data3_modificado <- data3_modificado[(data3_modificado$date < "2021-07-04") == TRUE, ]
data3_modificado$epi_week <- 0
data3_modificado$date <- as.Date(data3_modificado$date)

## inserindo as SE

for(i in 1:length(data3_modificado$name)){
  if((year(data3_modificado$date[i]) == 2020) == TRUE){
    for(j in 1:length(epi20$Semana)){
      if(data3_modificado$date[i] >= epi20$Início[j] & data3_modificado$date[i] <=
epi20$Término[j]){
        data3_modificado$epi_week[i] <- semana_ep20[j]
      }
    }
  }
  else{
    for(j in 1:length(epi21$Semana)){
      if(data3_modificado$date[i] >= epi21$Início[j] & data3_modificado$date[i] <=
epi21$Término[j]){
        data3_modificado$epi_week[i] <- semana_ep21[j]
      }
    }
  }
  if(data3_modificado$date[i] == "2021-01-01" || data3_modificado$date[i] == "2021-01-02"){
    data3_modificado$epi_week[i] <- "2020.53"
  }
}

str(data3_modificado)

```

```

data3_modificado$rt <- as.numeric(data3_modificado$rt)

# media dos rt's por SE
rt1 <- data3_modificado %>%
  group_by(name, epi_week) %>%
  summarise(Total = mean(rt, na.rm = TRUE),
            Count = n())

# quantas SE tem cada cidade
r5 <- rt1 %>%
  group_by(name) %>%
  count()

## Fazendo um data frame com as SE que não tem dado - Isso varia de cidade para cidade
aux33 <- as.data.frame(matrix(0, ncol = length(data3_modificado), nrow = 1981))
colnames(aux33) <- names(data3_modificado)

# valores que tenho que replicar de cada cidade e colocar 0
dif <- 79 - r5$n

for(k in 1:5){
  vetaux <- 0
  j <- 1
  p <- 0
  vetaux2 <- unique(data3_modificado[,k])
  for(i in 1:length(dif)){
    vetaux[j: c(dif[i] + p)] <- rep(vetaux2[i], dif[i])
  }
}

```

```

    j <- dif[i] + p + 1
    p <- dif[i] + p
  }
  aux33[,k] <- vetaux

}

# epi_week
vetse <- 0
j <- 1
p <- 0
for(i in 1:length(dif)){
  vetse[j: c(dif[i] + p)] <- rep(epi20$SE[1: dif[i]], dif[i])
  j <- dif[i] + p + 1
  p <- dif[i] + p
}
aux33$epi_week <- vetse

# data frame para as SE que possui dados
rt_est <- as.data.frame(matrix(0, ncol = length(data3_modificado), nrow = sum(r5$n)))
colnames(rt_est) <- names(data3_modificado)

for(k in 1:5){
  vetaux <- 0
  j <- 1
  p <- 0
  vetaux2 <- unique(data3_modificado[,k])
  for(i in 1:length(r5$name)){

```

```

    vetaux[j: c(r5$n[i] + p)] <- rep(vetaux2[i], r5$n[i])
    j <- r5$n[i] + p + 1
    p <- r5$n[i] + p
  }
  rt_est[,k] <- vetaux
}

rt_est$rt <- rt1$Total

vetSE <- 0
j <- 1
p <- 0
for(i in 1:length(r5$name)){
  vetSE[j: c(r5$n[i] + p)] <- c( epi20$SE[(dif[i] + 1):53],      epi21$SE[1:26] )
  j <- r5$n[i] + p + 1
  p <- r5$n[i] + p
}

rt_est$epi_week <- vetSE

## Juntando "aux33" (com dados 0) e "rt_est" (com as medias de rt)
rtEst <- merge(aux33, rt_est, all = TRUE)
rtEst <- rtEst[order(rtEst$epi_week, decreasing = FALSE), ]
rtEst <- rtEst[order(rtEst$name, decreasing = FALSE), ]

#####Vacinas#####
# Antes da primeira pessoas ser vacinada no Brasil é 0;

```

Do dia 17 de janeiro de 2021 (3ª semana epidemiológica) até 17 de abril de 2021 (15ª SE) é 1;

Do dia 18 de abril de 2021 (16ª SE) até o dia 3 de julho de 2021 (26ª SE) é 2.

```
zero <- 53 + 2 # 53 SE de 2020 + 2 SE de 2021
```

```
um <- 26 - 2 - 11 # 26 SE de 2021 - 2 SE de 2021 que não tinha a vacina - 11 SE de 2021 que é "2"
```

```
dois <- 11
```

```
vacinas <- c(rep(0,zero), rep(1,um), rep(2,dois))
```

```
#####Montando data final#####
```

```
codigoSaude <- data1 %>%
```

```
  group_by(name) %>%
```

```
  summarise(codigo = code_saude)
```

```
cod <- unique(codigoSaude)
```

```
## "covid"
```

```
covid <- data.frame(matrix(rep(0,16), ncol = 16, nrow = 79*138)) # tenho 67 SE com dados
```

```
colnames(covid) <- c("code", "population", "id", "code_saude", "name", "epi_week", "vacinas",  
"confirmed", "deaths",
```

```
                "varejo_e_recreacao",          "mercearia_e_farmacia",          "parques",  
"estacoes_de_transito",
```

```
                "locais_de_trabalho", "residencial", "rt")
```

```
covid$code <- rep(unique(casos$code), rep(79,138))
```

```
covid$population <- rep(unique(casos$population), rep(79,138))
```

```
covid$id <- rep(unique(casos$id), rep(79,138))
```

```
covid$code_saude <- rep(cod$codigo, rep(79,138))
```

```
covid$name <- c(rep(unique(sort(casos$name)), rep(79,138)))
```

```
covid$epi_week <- casos$epi_week
```

```

covid$confirmed <- casos$confirmed
covid$deaths <- casos$deaths

covid$varejo_e_recreacao <- mobilidadeDados$varejo_e_recreacao
covid$mercearia_e_farmacia <- mobilidadeDados$mercearia_e_farmacia
covid$parques <- mobilidadeDados$parques
covid$estacoes_de_transito <- mobilidadeDados$estacoes_de_transito
covid$locais_de_trabalho <- mobilidadeDados$locais_de_trabalho
covid$residencial <- mobilidadeDados$residencial

covid$vacinas <- rep(vacinas, 138)

covid$rt <- rtEst$rt

covid <- covid[order(covid$name, decreasing = FALSE), ]

## OBS: para os dados de leitos e de respiradores, é preciso saber em qual mês começa a SE

# Pegando so os meses de 2020
meses20 <- month(epi20$Início)

# Verificando quantos SE tem em cada mês
tam20 <- c()
j <- 1
for(i in 1:12){
  tam20[i] <- length(meses20[meses20 == i])
}

#####3OBS#####

```

a 1ª SE de 2020 começa em dezembro de 2019. Então no vetor "tam20", é preciso tirar um mês de dezembro

```
tam20 <- c(tam20[1:11], 4)
```

Pegando os dados de dezembro de 2019

```
dois <- read.csv("leitos_internacao_dez_2019.csv", sep = ";", header = TRUE)
```

pegando somente as cidades que eu quero

```
cod_e_muni <- paste0(c("cod", "name"), 1:2)
```

Separando a coluna "Municípios" em duas: uma com o código e o outro com o nome dos municípios

```
dois <- separate(dois, "Município", into = cod_e_muni, sep = " ", remove = FALSE,
                 extra = "merge")
```

```
doism <- dois[(dois$cod1 %in% unique(data1$code_saude)) == TRUE, ]
```

```
for(i in 1:length(doism$name2)){
```

```
  doism$total_leitos[i] <- sum(doism$surgical[i], doism$clinical[i], doism$obstetric[i],
                              doism$pediatric[i], doism$other[i])
```

```
}
```

```
dados_2 <- doism %>%
```

```
  group_by(name2) %>%
```

```
  summarise(Total = sum(total_leitos, na.rm = TRUE),
```

```
            Count = length(total_leitos))
```

Pegando so os meses de 2021

```
meses21 <- month(epi21$Início)
```

```

# quantas semanas tem em cada mês
tam21 <- c()
j <- 1
for(i in 1:12){
  tam21[i] <- length(meses21[meses21 == i])
}

#####Dados de Leitos#####
## Leitos de Internação

leitos <- read.csv("jessica_leitos_internacao_novo.csv", sep = ";", dec = ".", header = TRUE)
str(leitos)
leitos$time <- as.Date(leitos$time, format = "%d/%m/%Y")

# Não tem tempo faltante
leitos[is.na(leitos$time) == TRUE,]

# Criando a coluna com a soma das colunas de 5 a 9: cirúrgico, clínico, obstétrico, pediátrico e
outros.
for(i in 1:length(leitos$name)){
  leitos$Total_leitos[i] <- sum(leitos$surgical[i], leitos$clinical[i], leitos$obstetric[i],
                              leitos$pediatric[i], leitos$other[i])
}

# colocando as datas em ordem
leitos <- leitos[order(leitos$time, decreasing = FALSE), ]

# pegando as colunas: code, population, id, code_saude, name, time, total
leitosN <- leitos[,c(1:5, 12, 21)]

# Agrupando por tempo e município, e somando sem distinção de SUS e privado

```

```

leitosAG <- leitosN %>%
  group_by(time, name) %>%
  summarise(Total = sum(Total_leitos, na.rm = TRUE),
            Count = n())

# quantos municípios tem em cada mês
tama <- leitosAG %>%
  group_by(time) %>%
  summarise(Total = length(month(time)))

## Replicando os dados de acordo com quantas SE tem em cada mês

# Para 2019
rep19 <- dados_2$Total
cidad19 <- dados_2$name2

# Para 2020
rep20 <- rep(leitosAG$Total[1:1656], rep(tam20, tama$Total[1:12]))
cidad20 <- rep(leitosAG$name[1:1656], rep(tam20, tama$Total[1:12]))

# Para 2021
rep21 <- rep(leitosAG$Total[1657:2484], rep(tam21[c(1:6)], tama$Total[13:18]))
cidad21 <- rep(leitosAG$name[1657:2484], rep(tam21[c(1:6)], tama$Total[13:18]))

# juntando os leitos de internação e as cidades
nomes <- c(cidad19, cidad20, cidad21)
vet_internacao <- c(rep19, rep20, rep21)

## Replicando as SE
week <- c("2020.01", "2020.02", "2020.03", "2020.04", "2020.05", "2020.06", "2020.07",
"2020.08", "2020.09", "2020.10", "2020.11", "2020.12")

```

```
# Para 2020
```

```
semanas <- c(week, (unique(data1$epi_week)[1:41]))
```

```
janeiro <- rep(semanas[2:5], tama$Total[1])
```

```
fevereiro <- rep(semanas[6:9], tama$Total[2])
```

```
marco <- rep(semanas[10:14], tama$Total[3])
```

```
abril <- rep(semanas[15:18], tama$Total[4])
```

```
maio <- rep(semanas[19:23], tama$Total[5])
```

```
junho <- rep(semanas[24:27], tama$Total[6])
```

```
julho <- rep(semanas[28:31], tama$Total[7])
```

```
agosto <- rep(semanas[32:36], tama$Total[8])
```

```
setembro <- rep(semanas[37:40], tama$Total[9])
```

```
outubro <- rep(semanas[41:44], tama$Total[10])
```

```
novembro <- rep(semanas[45:49], tama$Total[11])
```

```
dezembro <- rep(semanas[50:53], tama$Total[12])
```

```
meses20 <- c(janeiro, fevereiro, marco, abril, maio, junho, julho, agosto, setembro, outubro,
novembro, dezembro)
```

```
# Para 2021
```

```
semanas21 <- (unique(data1$epi_week)[42:67])
```

```
janeiro21 <- rep(semanas21[1:5], tama$Total[13])
```

```
fevereiro21 <- rep(semanas21[6:9], tama$Total[14])
```

```
marco21 <- rep(semanas21[10:13], tama$Total[15])
```

```
abril21 <- rep(semanas21[14:17], tama$Total[16])
```

```
maio21 <- rep(semanas21[18:22], tama$Total[17])
```

```
junho21 <- rep(semanas21[23:26], tama$Total[18])
```

```

meses21 <- c(janeiro21, fevereiro21, marco21, abril21, maio21, junho21)

# Juntando todas as SE de 2020 e 2021
mesestotal <- c(rep(semanas[1], 138), meses20, meses21)

#### Leitos Complementares
leitosComp <- read.csv("jessica_leitos_complementar_novo.csv", sep = ";", dec = ",", header =
TRUE)
leitosComp$time <- as.Date(leitosComp$time, format = "%d/%m/%Y")

# Somando os leitos de UTI adulta e pediátrico
for(i in 1:length(leitosComp$name)){
  leitosComp$Total_comp[i] <- sum(leitosComp$suti_adult_ii_covid19[i],
leitosComp$suti_pediatic_ii_covid19[i])
}

# criando um novo data frame somente com as colunas: code, population, id, code_saude, name,
time, Total_comp
leitosM <- leitosComp[,c(1:5, 12, 35)]
leitosM <- leitosM[order(leitosM$time, decreasing = FALSE), ]
leitosM$Total_comp[4800:4815]

# retirando a última linha que é a cidade de Colombo que não possui dado nem data
leitosM <- leitosM[-4815,]

compAG <- leitosM %>%
  group_by(time, name) %>%
  summarise(Total = sum(Total_comp, na.rm = TRUE),
            Count = n())

```

```

# quantas cidades tem em cada mês
qtcidC <- leitosM %>%
  group_by(time) %>%
  summarise(total = length(unique(name)))

## Quais cidades estão faltando em cada mês
r <- unique(leitos$time)

for(i in 1:18){ # tem 18 meses
  q <- compAG[(compAG$time == r[i]) == TRUE, ]
  compmeses <- leitosAG[(leitosAG$time == r[i]) == TRUE, ]
  print(r[i])
  print(compmeses[(compmeses$name %in% q$name) == FALSE, ])
  q <- 0
}

### OBS: A cidade de Colombo não tem dados de internação complementar

# Nem todos os meses têm todas as cidades, sendo assim, iremos completar os dados com 0.
compAG <- as.data.frame(compAG)

## pegando os dados de Águas Lindas de Goiás

leitosComp[(leitosComp$name == "Águas Lindas de Goiás") == TRUE, ]

# como no "compAG", existe 10 leitos em outubro, então vou replicar para novembro
novo1 <- data.frame(time = c("2020-01-01", "2020-02-01", "2020-03-01", "2020-04-01",
"2020-11-01"),
  name = rep("Águas Lindas de Goiás", 5),
  Total = c(0,0,0,0,10), Count = rep(1,5))

```

```

novo1$time <- as.Date(novo1$time)
compAG <- rbind(compAG, novo1)

```

```
## Embu das Artes
```

```

novo2 <- data.frame(time = c("2020-01-01", "2020-02-01", "2020-03-01"),
                    name = rep("Embu das Artes", 3),
                    Total = c(0,0,0), Count = rep(1,3))
novo2$time <- as.Date(novo2$time)
compAG <- rbind(compAG, novo2)

```

```
## Ribeirão as Neves
```

```

novo3 <- data.frame(time = c("2020-01-01", "2020-02-01", "2020-03-01"),
                    name = rep("Ribeirão das Neves", 3),
                    Total = c(0,0,0), Count = rep(1,3))
novo3$time <- as.Date(novo3$time)
compAG <- rbind(compAG, novo3)

```

```
## Santa Luzia
```

```

novo4 <- data.frame(time = c("2020-01-01", "2020-02-01", "2020-03-01", "2020-04-01"),
                    name = rep("Santa Luzia", 4),
                    Total = c(0,0,0,0), Count = rep(1,4))
novo4$time <- as.Date(novo4$time)
compAG <- rbind(compAG, novo4)

```

```
## adicionando a cidade de Colombo - pra ficar 138 cidades
```

```

novo5 <- data.frame(time = r, name = rep("Colombo",18), Total = rep(0,18), Count = rep(1,18))
compAG <- rbind(compAG, novo5)

```

```

## colocando em ordem de data
compAG <- compAG[order(compAG$name, decreasing = FALSE), ]
compAG <- compAG[order(compAG$time, decreasing = FALSE), ]

## Replicando os dados
# Para 2020
repC20 <- rep(compAG$Total[1:1656], rep(tam20, tama$Total[1:12]))
cidadC20 <- rep(compAG$name[1:1656], rep(tam20, tama$Total[1:12]))

# Para 2021
repC21 <- rep(compAG$Total[1657:2484], rep(tam21[c(1:6)], tama$Total[13:18]))
cidadC21 <- rep(compAG$name[1657:2484], rep(tam21[c(1:6)], tama$Total[13:18]))

#### OBS: não tinha dados de leitos complementares em dezembro de 2019
# Juntando os leitos complementares
vet_complementar <- c(rep(0, 138), repC20, repC21)

cidades19 <- sort(unique(compAG$name))
cidadesC <- c(cidades19, cidad20, cidad21)

# SE são as mesmas para leitos de internação
semanas_total <- c(rep("2020.1", 138), meses20, meses21)

### Juntando os dados de leitos de internação e complementar
internacao <- matrix(c(nomes, mesestotal, vet_internacao, vet_complementar), ncol = 4)
internacao <- as.data.frame(internacao)
colnames(internacao) <- c("names", "epi_week", "total_int", "total_comp")

```

```
#####Dados de respiradores#####
```

```
respiradores <- read.csv("jessica_respiradores_total.csv", sep = ";", header = TRUE)
```

```
respiradores <- respiradores[year(respiradores$time) != 2019, ]
```

```
# Agrupando por tempo e cidades os respiradores em uso e os estabelecimentos que possuem respiradores
```

```
respirador <- respiradores %>%
```

```
  group_by(time, name) %>%
```

```
  summarise(uso = sum(equip_uso),
```

```
            estab = sum(estab_equip),
```

```
            count = n())
```

```
# quantos municípios tem em cada mês
```

```
tamaResp <- respirador %>%
```

```
  group_by(time) %>%
```

```
  summarise(Total = length(month(time)),
```

```
            Count = n())
```

```
# replicando para 2020
```

```
resp20USO <- rep(respirador$uso[1:1656], rep(tam20, tamaResp$Total[1:12]))
```

```
resp20Estab <- rep(respirador$estab[1:1656], rep(tam20, tamaResp$Total[1:12]))
```

```
repcidades20 <- rep(respirador$name[1:1656], rep(tam20, tamaResp$Total[1:12]))
```

```
# replicando para 2021
```

```
resp21USO <- rep(respirador$uso[1657:2484], rep(tam21[1:6], tamaResp$Total[13:18]))
```

```
resp21Estab <- rep(respirador$estab[1657:2484], rep(tam21[1:6], tamaResp$Total[13:18]))
```

```
repcidades21 <- rep(respirador$name[1657:2484], rep(tam21[1:6], tamaResp$Total[13:18]))
```

```

## Pegando os respiradores de dezembro de 2019
resp2019 <- read.csv("respiradores_dez_2019.csv", sep = ";", header = TRUE)
resp2019 <- resp2019[1:2313,]
# separa coluna em duas
cod_e_muni <- paste0(c("cod", "name"), 1:2)
# Separando a coluna "Municípios" em duas: uma com o código e o outro com o nome dos
municípios
respirador19 <- separate(resp2019, "Município", into = cod_e_muni, sep = " ", remove =
FALSE,
                        extra = "merge")

# pegando somente as cidades que eu quero
respNovo <- respirador19[(respirador19$cod1 %in% unique(data1$code_saude)) == TRUE, ]
respNovo <- respNovo[order(respNovo$name2, decreasing = FALSE), ]

resp19USO <- respNovo$Equipamentos_em_Uso
resp19Estab <- respNovo$Estab_c._Equip_SUS
cidades2019 <- respNovo$name2

# juntando os dados de respiradores em uso e estabelecimentos com respiradores
respUSO <- c(resp19USO, resp20USO, resp21USO)
repEstab <- c(resp19Estab ,resp20Estab, resp21Estab)
cidadesResp <- c(cidades2019 ,repcidades20, repcidades21)

respiradores_ventiladores <- data.frame(matrix(c(cidadesResp, mesestotal, respUSO,
repEstab), ncol = 4))
colnames(respiradores_ventiladores) <- c("name", "epi_week", "respUSO", "respEstab")

```

```
#### Juntando os data frames: "internacao" e "respiradores_ventiladores" no data frame "covid"
```

```
# deixando os data frames em ordem de município
```

```
internacao <- internacao[order(internacao$names, decreasing = FALSE), ]
```

```
respiradores_ventiladores <- respiradores_ventiladores[order(respiradores_ventiladores$name,  
decreasing = FALSE), ]
```

```
covid$leito_int <- internacao$total_int
```

```
covid$leito_comp <- internacao$total_comp
```

```
covid$respUSO <- respiradores_ventiladores$respUSO
```

```
covid$respEstab <- respiradores_ventiladores$respEstab
```