Capítulo

8

A linguagem R na análise de dados: Um estudo de caso dos transportes públicos do RJ durante a pandemia da Covid-19

Julia Amaro Gonçalves Fagundes, Matheus Henrique de Sousa Oliveira e Vladimir Fagundes

Abstract

The use of the R Programming Language for statistical analysis using techniques of trend and seasonality reduction, Pearson correlation and cross correlation proved to be very relevant to solve real world problems. In this context, this chapter was developed in order to develop a tutorial applied to a relevant and current topic (the Covid-19 pandemic and public transport), aiming at a fast and targeted learning, so that the student quickly reaches his goals and ensure a solid foundation if you are interested in delving into both the subject and the language.

Resumo

A utilização da Linguagem de Programação R para elaboração de análise estatística utilizando técnicas de redução de tendência e sazonalidade, correlação de Pearson e correlação cruzada mostrou-se bastante relevante para resolver problemas do mundo real. Neste contexto, este capítulo foi desenvolvido com o intuito de apresentar um tutorial aplicado a um tema relevante e atual (a pandemia da COVID-19 e os transportes públicos), objetivando um aprendizado rápido e direcionado, de maneira que o aluno atinja rapidamente seus objetivos e garanta uma base sólida, caso possua interesse de se aprofundar tanto no assunto quanto na linguagem.

8.1. Introdução

Em março de 2020, à época com 118.326 casos confirmados e 4.292 óbitos, a Organização Mundial da Saúde (OMS) declarou a disseminação da Covid-19 como pandemia [OMS, 2020] . Desde então, o mundo enfrenta uma emergência de saúde pública, com poder devastador ainda não experimentado em um passado recente. O novo coronavírus é uma doença respiratória aguda e, por vezes, grave, podendo levar o infectado a óbito. De acordo com Perlan et al. (2020) a transmissão interpessoal ocorre pelo contato com secreções contaminadas, principalmente pelo contato com grandes gotículas respiratórias, mas também pode ocorrer por meio do contato com uma superfície contaminada.

Entretanto, apesar das características da transmissão de vírus, as medidas de limitação da circulação de pessoas para conter a propagação do Covid-19 não podem ser confundidas com a paralisação do transporte público. Apesar da redução do número de pessoas em circulação nas ruas, a interrupção dos serviços não é uma opção; pelo contrário. Nem todos os usuários de transporte público podem trabalhar remotamente ou possuem veículo próprio para se locomoverem quando necessário [Lima et al., 2020].

Diante desse cenário de incerteza, surge a necessidade de avaliar o comportamento da transmissão para que possam ser ofertadas alternativas para que profissionais dos serviços essenciais possam se locomover de maneira segura e adequada. Além disso, existe um grande número de dados espalhados em diversos órgãos que necessitam de análise para que se possa gerar indicadores que balizem tomadas de decisão com maior acurácia e possibilitem desenvolver políticas públicas que vão ao encontro dos anseios da sociedade.

Neste cenário, a associação de dados reais relativos à pandemia e aos transportes públicos permite uma análise mais assertiva em termos de estimativa da necessidade de ações preventivas, bem como para estratégias de retomada. Cabe ressaltar que uma maneira eficiente de se estimar a relação entre diferentes conjuntos de dados é a análise de correlação, onde são estudadas potenciais relações de causa e efeito dentre os dados disponíveis. Para tal estudo, podem ser utilizadas diferentes linguagens de programação como ferramentas para manipular e analisar dados tais como Python, Julia, Matlab, Java e R.

A linguagem de programação R se apresenta como uma alternativa robusta e eficiente para a ciência de dados em geral. Por ser uma linguagem desenvolvida por estatísticos, muitas funções e fórmulas que nas linguagens mais difundidas precisam ser desenvolvidas do zero, no R já vêm prontas para aplicação com breves comandos. Esta característica, além de permitir códigos mais simples, também contribui para minimização de erros. Cabe ainda ressaltar que o R é gratuito e de código aberto, o que gera potencial vantagem competitiva em relação a ferramentas como SAS e SPSS.

Dentre as principais funções da linguagem está extensa relação de modelos estatísticos, que vão desde a modelagem linear e não-linear, a análise de séries temporais, os testes estatísticos clássicos, análise de agrupamento e classificação, etc, além da apresentação gráfica dos resultados contando com variadas técnicas, passando

também pela criação e manipulação de mapas. Desta forma, a análise estatística da linguagem 'R' se apresenta como uma boa alternativa para análise de dados de forma simples e dinâmica, uma vez que dispõe de uma enorme quantidade de pacotes desenvolvidos com as principais manipulações mais comumente utilizadas na análise de dados, o que facilita tanto o aprendizado quanto a utilização da linguagem..

A escolha dos dados aqui aplicados foi feita a título de demonstração da utilização da linguagem, visando a aplicação em uma situação real que está sendo estudada ao redor de todo o mundo. Diante de diversas comparações entre o uso de transportes públicos e o aumento dos casos de Covid-19, foram selecionadas duas bases de dados reais (uma sobre transporte e outra sobre casos de Covid-19) para verificar a correlação entre as variáveis. Nesta aplicação, as duas bases serão tratadas como séries temporais regulares, abstraindo-se qualquer característica epidemiológica dos dados relacionados às contaminações.

Este capítulo apresenta um método de análise e sua aplicação prática para analisar dados temporais de transportes públicos e casos de COVID-19 no estado do Rio de Janeiro. Para tanto, será utilizada a linguagem R com todas as ferramentas necessárias para as análises que serão realizadas. Ao final do capítulo, o leitor terá uma experiência prática com o R que pode ser adaptada e reproduzida em outros cenários. Bem como obterá informações sobre a COVID-19 e sua relação com o transporte público o que, por sua vez, pode proporcionar insights ao leitor para se aprofundar no tema por meio de pesquisas e outras análises.

Durante o desenvolvimento, serão apresentadas ferramentas de visualização e manipulação de dados, juntamente com o procedimento empregado em cada função, bem como os resultados encontrados.

8.2. Metodologia

Nesta seção serão apresentados os principais passos e escolhas para análise. A cada etapa apresentada, também serão abordados os conceitos básicos necessários para sua compreensão. A Figura 8.1 apresenta as fases da pesquisa que será desenvolvida neste capítulo.

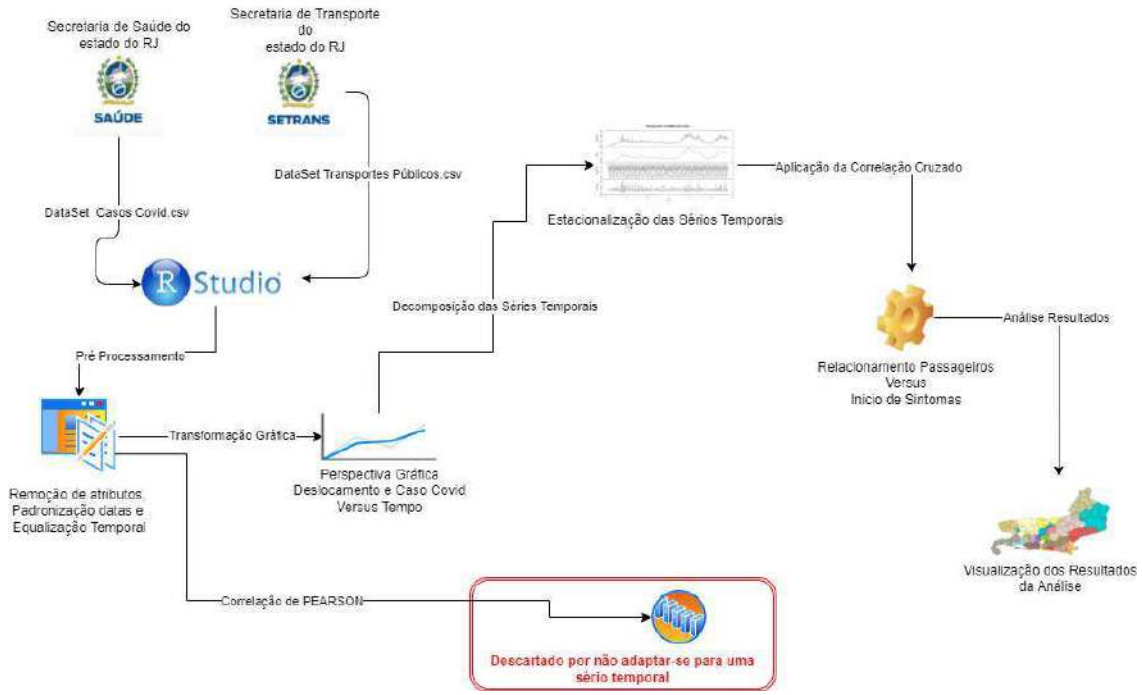

Figura 8.1. Fases da Pesquisa (Fonte: Autores)

8.2.1 Séries temporais

Segundo Ehlers (2007), uma série temporal consiste em um conjunto de observações feitas sequencialmente ao longo do tempo. A característica mais importante deste tipo de dados está no fato de que as observações vizinhas são dependentes, tornando necessário que essas dependências sejam levadas em consideração durante as análises e modelagens. Enquanto em modelos de regressão, por exemplo, a ordem das observações é irrelevante, para a análise em séries temporais a ordem dos dados é crucial. Vale ressaltar também que o tempo pode ser substituído por outras variáveis como espaço, profundidade, etc.

Como a maior parte dos procedimentos estatísticos foi desenvolvida para analisar observações independentes, o estudo de séries temporais requer o uso de técnicas específicas. Dados de séries temporais surgem em vários campos do conhecimento tais como: economia (preços diários de ações, taxa mensal de desemprego, produção industrial); medicina (eletrocardiograma, eletroencefalograma); epidemiologia (número mensal de novos casos da doença); meteorologia (precipitação pluviométrica, temperatura diária, velocidade do vento); etc.

Ehlers (2007) definiu algumas características particulares a este tipo de dados, por exemplo:

- Observações correlacionadas são mais difíceis de serem analisadas e, portanto, requerem técnicas específicas;
- É necessário levar em conta a ordem temporal das observações;
- Fatores complicadores como presença de tendências e variação sazonal ou cíclica podem ser difíceis de estimar ou remover;

- A seleção de modelos pode ser bastante complicada, e as ferramentas podem ser de difícil interpretação; e
- É mais difícil lidar com observações perdidas e dados discrepantes devido à natureza sequencial.

De um modo geral, os principais objetivos que podem ser alcançados ao se estudar séries temporais estão dispostos a seguir:

- Descrição: onde pretende-se descrever propriedades da série, tais como o padrão de tendência, a existência de variação sazonal ou cíclica, observações discrepantes (*outliers*), alterações estruturais como mudanças de padrão, etc.
- Explicação: onde se utiliza a variação em uma das séries para explicar o comportamento da outra.
- Predição: onde se objetiva predizer valores futuros com base em valores passados. Neste caso, assume-se que o futuro envolve incerteza, o que impede previsões exatas, porém se desenvolvem os estudos visando a máxima minimização dos erros de previsão.
- Controle: nos casos em que os valores da série temporal medem a "qualidade" de um processo e o objetivo está em controlar o mesmo. Como exemplo, pode-se citar o controle estatístico de qualidade onde as observações são representadas em cartas de controle.

Ainda de acordo com Ehlers (2007), uma das suposições mais frequentes que se faz a respeito de uma série temporal é a de que ela é estacionária, ou seja, ela se desenvolve no tempo aleatoriamente ao redor de uma média constante, refletindo alguma forma de equilíbrio estável. Todavia, a maior parte das séries na prática apresentam alguma forma de não estacionariedade. Como a maioria dos procedimentos de análise estatística de séries temporais supõe que estas sejam estacionárias, faz-se necessária a transformação dos dados originais antes da realização das análises, para tal, deve ser removida a tendência e a sazonalidade.

A tendência está autoexplicativamente relacionada com a tendência da série crescer ou decrescer ao longo do tempo. Já a sazonalidade está associada a um comportamento que tende a se repetir a cada período x de tempo em uma série temporal. A sazonalidade se apresenta em dois tipos, são eles:

- Aditiva: a série apresenta flutuações sazonais mais ou menos constantes não importando o nível global da série.
- Multiplicativa: o tamanho das flutuações sazonais varia dependendo do nível global da série.

8.2.2 Correlação

Neste estudo, pretende-se correlacionar os dados de utilização do transporte público com os casos de Covid-19 no Estado do Rio de Janeiro. Em geral, é comum para a análise de correlações se utilizar a correlação de Pearson. De acordo com Figueiredo Filho e Silva Junior (2009) essa forma de correlação se caracteriza como uma medida de associação linear entre variáveis. Em termos estatísticos, duas variáveis se associam

quando elas guardam semelhanças na distribuição dos seus escores, podendo se associar a partir da distribuição das frequências ou pelo compartilhamento de variância. No caso da correlação de Pearson, ela é uma medida da variância compartilhada entre duas variáveis. Por outro lado, o modelo linear supõe que o aumento ou decremento de uma unidade na variável X gera o mesmo impacto em Y . Em termos gráficos, por relação linear entende-se que a melhor forma de ilustrar o padrão de relacionamento entre duas variáveis é através de uma linha reta. Portanto, a correlação de Pearson exige um compartilhamento de variância e que essa variação seja distribuída linearmente.

O formato dos dados com observações diárias caracteriza uma série temporal, onde não cabe a aplicação da Correlação de Pearson, uma vez que este modelo considera a correlação linear entre duas variáveis, dificultando uma análise em dias diferentes, como é o caso das bases de dados a serem estudadas. Neste caso, o procedimento adequado se dá com a aplicação da correlação cruzada que permite identificar correlações que acontecem em qualquer período de tempo.

De acordo com Silva Filho (2014), o método de correlação cruzada é um método estatístico capaz de estimar o expoente que caracteriza a correlação de longo alcance entre duas séries temporais, em regime não estacionário. Dessa forma, a função de correlação cruzada estima a correlação entre duas séries temporais tendo o tempo incluído como uma variável. Dessa forma, é possível estimar a influência de uma série na outra mesmo que esta não tenha sido gerada no mesmo tempo, como por exemplo uma série alocada no tempo t influenciando no comportamento de uma série alocada no tempo t + k. O atraso ou defasagem é chamado de lag e no exemplo anterior está denotado como k, podendo adquirir valores positivos ou negativos. Cabe ressaltar que para a aplicação da função de correlação cruzada a série deve ser estacionária, sendo necessária portanto a remoção de quaisquer tendência e sazonalidade presentes em ambas as séries.

8.3 Fonte de Dados

A Secretaria de Transportes do estado do Rio de Janeiro (Setrans) é o órgão responsável por realizar estudos, pesquisas e planejamento do sistema de transportes do estado, bem como operar adequadamente os serviços de transportes e de terminais rodoviário de passageiros, metroviário, ferroviário e hidroviário. A Setrans disponibilizou para esta pesquisa dados de quantidade de passageiros diários para todos os modos de transporte por ela operados no período de 09/03/2020 até 11/04/2021.

Os modos de transporte público selecionados para esta análise foram: metrô, trem, barcas, ônibus municipais e ônibus intermunicipais. Além disso, para todos os modos foram considerados os valores totais de passageiros somando todas as linhas/ramais de cada modo.

Visando a comparação dos dados de utilização de transporte público com a evolução da pandemia da Covid-19 para o Estado do Rio de Janeiro, foram coletados dados referentes ao número de casos de Covid-19 no estado para o mesmo período dos dados fornecidos pela Setrans. A fonte dos dados foi a plataforma da Secretaria de Saúde do RJ que fornece uma base em formato csv com dados atualizados diariamente pelo Centro de Informações Estratégicas e Resposta de Vigilância em Saúde (CIEVS-RJ) da Secretaria de Saúde do Estado do Rio de Janeiro, a partir do sistema

esus-VE e SIVEP-Gripe, em articulação com as vigilâncias das secretarias municipais de saúde do Estado.

Buscando maior assertividade dos valores referentes aos números de casos, na tentativa de minimizar os erros referentes a atrasos da data de divulgação, foram selecionados os casos confirmados por data de início dos sintomas. A plataforma permite selecionar os parâmetros para gerar diferentes tabelas, as quais podem ser baixadas em formato csv. Para esta análise foi selecionado "Data de início dos sintomas" como parâmetro de linhas e "Casos confirmados" para as medidas.

Ao final da etapa de coleta de dados, foram geradas duas planilhas em formato csv, são elas: "12.04 - Dados.csv" para os dados de transporte público e "inicio_x_casos.csv" para dados de Covid-19.

O código completo e a base de dados de referência para os dados de casos de Covid-19 estão disponíveis no Github

<https://github.com/juliafagundescoc/Cap-tulo-8-SBC>

8.3.1 Importação dos dados

A IDE (*integrated development environment*) utilizada foi o RStudio que pode ser utilizada tanto para R quanto para Python. Em um primeiro momento, ambos os arquivos csv devem ser salvos na mesma pasta, a qual será definida como o diretório no RStudio como indicado na Figura 8.2.

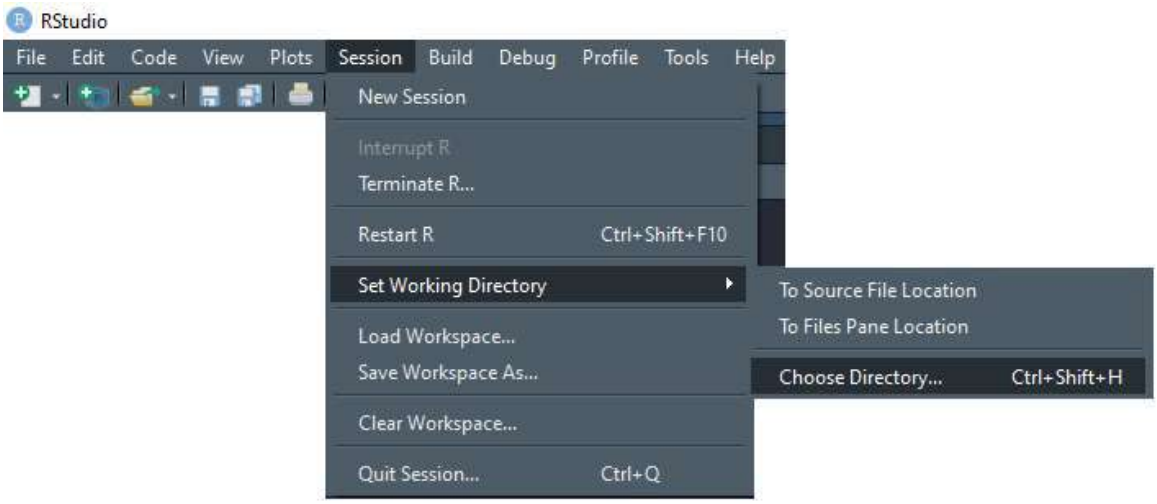

Figura 8.2. R Studio Seleção de Diretório (Fonte: Autores)

A Figura 8.3 ilustra a importação das bases de dados incluindo os parâmetros para adequação dos *Datasets*. Nesta etapa, foi selecionado o símbolo ";" como separador das colunas para a conversão do csv, além disso, o padrão de codificação (*encoding*) foi definido como "Latin1" para evitar erros com a presença de acentos nos nomes das variáveis. Nesta etapa também cabe incluir a definição do símbolo de separação de decimais, no caso da base enviada pela Setrans era a vírgula, por esse motivo o parâmetro "decimal_mark" foi definido como ",". Ao final, os tipos das colunas que estavam pré-definidas incorretamente (estavam como caracter

e eram numéricas) foram corrigidos utilizando o parâmetro “col_types”, igualando as respectivas colunas a “col_number()” para defiini-las como tipo numéricas.

```
inicio_x_casos <- read_delim("inicio_x_casos.csv",
                    ";", escape_double = FALSE,
                    locale = locale(encoding = "Latin1"),
                    trim_ws = TRUE)

X12_04_Dados <- read_delim("12.04 - Dados.csv",
                    ";", escape_double = FALSE,
                    locale = locale(decimal_mark = ",", encoding = "Latin1"),
                    trim_ws = TRUE,
                    col_types = cols(Metro_Tot = col_number(),
                                     Trem_Tot = col_number(),
                                     Barcas_Tot = col_number(),
                                     RMRJ = col_number(),
                                     Intermunicipais = col_number()))
```

Figura 8.3. Trecho do código importação das bases csv de Casos e Transportes Públicos (Fonte: Autores)

8.4. Desenvolvimento

Em um primeiro momento, cabe ressaltar que o R disponibiliza a documentação de cada função utilizando “?” antes da mesma, nesta é possível ter acesso às informações de modo de usar, parâmetros, aplicações, além de apresentar exemplos de aplicação.

No desenvolvimento deste estudo, de acordo com o contexto da proposta de ensinar a linguagem R, em uma aplicação prática, foram selecionadas as seguintes bibliotecas:

- **Dplyr** é o pacote mais útil para realizar transformação de dados, aliando simplicidade e eficiência de uma forma elegante. Os scripts em R que fazem uso inteligente dos verbos ‘dplyr’ e as facilidades do operador pipe tendem a ficar mais legíveis e organizados sem perder velocidade de execução. São suas principais funções: *select()* - seleciona colunas; *arrange()* - ordena a base; *filter()* - filtra linhas; *mutate()* - cria/modifica colunas; *group_by()* - agrupa a base; e *summarise()* - sumariza a base [Battisti e Smolski, 2019].
- **Lubridate** é o pacote que traz diversas funções para extrair os componentes de um objeto da classe date. Principais funções: second() - extrai os segundos; minute() - extrai os minutos; hour() - extrai a hora; wday() - extrai o dia da

semana.; mday() - extrai o dia do mês; month() - extrai o mês; e year() - extrai o ano [Battisti e Smolski, 2019].

- **Ggplot2** é o pacote que constrói um gráfico camada por camada. Este pacote permite criar o nosso canvas, um quadro em branco onde vamos colocar todas as outras camadas do gráfico. Dentro do R, isso corresponde a uma lista com as informações necessárias para a criação do gráfico. Cada camada adicionada ao ggplot adiciona mais informações nessa lista [Battisti e Smolski, 2019].
- **Ggcorrplot** é o pacote utilizado para visualizar facilmente uma matriz de correlação usando 'ggplot2'. Ele fornece uma solução para reordenar a matriz de correlação e exibe o nível de significância no gráfico. Também inclui uma função para calcular uma matriz de valores "p" de correlação [Battisti e Smolski, 2019].
- **Stats** é um pacote que facilita a visualização da correlação entre duas ou mais camadas [Battisti e Smolski, 2019].
- **Zoo** é um pacote para observações indexadas totalmente ordenadas. Destina-se particularmente a séries temporais irregulares de vetores / matrizes e fatores numéricos. Os principais objetivos do projeto do zoo são a independência de uma classe de índice / data / hora específica e consistência com ts e R de base, fornecendo métodos para estender os genéricos padrões [Battisti e Smolski, 2019].

8.4.1. Pré-processamento

Na etapa de pré-processamento, apresentada na Figura 8.4, foram removidas algumas observações e colunas obsoletas, bem como a adequação dos nomes das colunas para que os dois *Dataframes* ficassem harmonizados e com uma perspectiva comum. Além disso, foi incluída a variável "Municipais" que engloba a parcela dos ônibus que não fazem trajeto intermunicipal. Esta inclusão fez-se necessária para que não exista duplicidade de dados referentes aos ônibus, uma vez que a variável "RJMJ" englobava a soma dos ônibus municipais e intermunicipais. A variável "Total" também foi incluída, contemplando a soma de todos os modos de transportes.

Como a base de dados de casos de Covid-19 apresentava algumas datas ausentes e a característica dos dados não justificava imputação desses valores, optou-se por remover estas datas do *Dataframe* final denominado "df". Ainda de acordo com a característica dos dados não foram removidos potenciais outliers.

```
# Limpeza e ajuste do Dataframe de casos de Covid-19
casos=inicio_x_casos[-c(1,597),]
colnames(casos)[1] <- "Data"
colnames(casos)[2] <- "Casos"
casos$Data <- as.Date(casos$Data)
# Limpeza e ajuste do Dataframe de transportes
df=X12_04_Dados[-c(400:615),-c(1,4:6,8:17,19:26,28:31,34:57)]
df <- rename(df, Data = Dia)
df$Data <- as.Date(df$Data)
df["Municipais"] <- df$RMRJ - df$Intermunicipais
df["Total"] <- df$Metro_Tot + df$Trem_Tot + df$Barcas_Tot + df$RMRJ
df <- df[,-6]
# Remoção das datas com valores ausentes de número de casos
data = df$Data
data <- as_data_frame(data)
colnames (data)[1] <- "Data"
x = left_join(data,casos,by= "Data")
df["Casos"] <- x$Casos
```

Figura 8.4 Trecho do código limpeza e ajustes da base (Fonte: Autores)

A Figura 8.5 apresenta as 10 primeiras observações de "df" que apresenta 399 observações no total. Além das variáveis dos modos de transporte e de casos, a variável "Tipo_de_dia" assume valores de D.U. referente aos dias úteis, além de "Fim_de_semana", "Feriado" e "**Enforcado**".

	Data	Tipo_de_dia	Metro_Tot	Trem_Tot	Barcas_Tot	Intermunicipais	Municipais	Casos
1	2020-03-09	D.U.	887593	605061	77532	965632	2907782	42
2	2020-03-10	D.U.	901063	614043	77315	983721	2947688	154
3	2020-03-11	D.U.	918853	605926	73641	948307	2861275	66
4	2020-03-12	D.U.	879675	602627	75166	964846	2909712	74
5	2020-03-13	D.U.	903167	604208	75639	962760	2925060	112
6	2020-03-14	Fim_de_semana	407462	235571	24799	591606	1860578	127
7	2020-03-15	Fim_de_semana	228293	96007	11651	322015	1104619	306
8	2020-03-16	D.U.	581026	465951	51343	839176	2316611	220
9	2020-03-17	D.U.	448427	395721	34190	769230	2015355	201
10	2020-03-18	D.U.	360805	339936	25160	690204	1767611	214

Figura 8.5. Primeiras observações da base gerada (Fonte: Autores)

8.4.2. Visualizações

A função ggplot() da biblioteca "ggplot2" permite elaborar um gráfico de linhas com todas as variáveis utilizando o parâmetro "geom_line", utilizado para a elaboração de gráficos de linhas, a aplicação da função está apresentada na Figura 8.6.

```
#Evolução Diária por modo de transporte


ggplot(df, aes(x=Data, y=Metro_Tot))+
  geom_line(aes(col= "Metrô"))+
  geom_line(aes(y=Trem_Tot, col= "Trem"))+
  geom_line(aes(y=Barcas_Tot, col= "Barcas"))+
  geom_line(aes(y=Municipais, col= "Ônibus Municipais"))+
  geom_line(aes(y=Intermunicipais, col= "Ônibus Intermunicipais"))+
  geom_line(aes(y=Casos, col= "Casos"))+
  theme_classic()+
  labs(x="Data",
       y="Modo de Transporte",
       color=NULL)+
  theme(legend.position = "top")
```

Figura 8.6. Trecho do código Evolução diária por modo de transporte

A Figura 8.7 apresenta a plotagem do gráfico gerado com a aplicação da função anterior, neste é possível observar que a diferença de escala das variáveis dificulta a comparação entre o comportamento das mesmas. As variáveis "Casos" e "Barcas" por exemplo apresentam uma variação quase imperceptível em comparação com as demais variáveis, o que não condiz com a realidade.

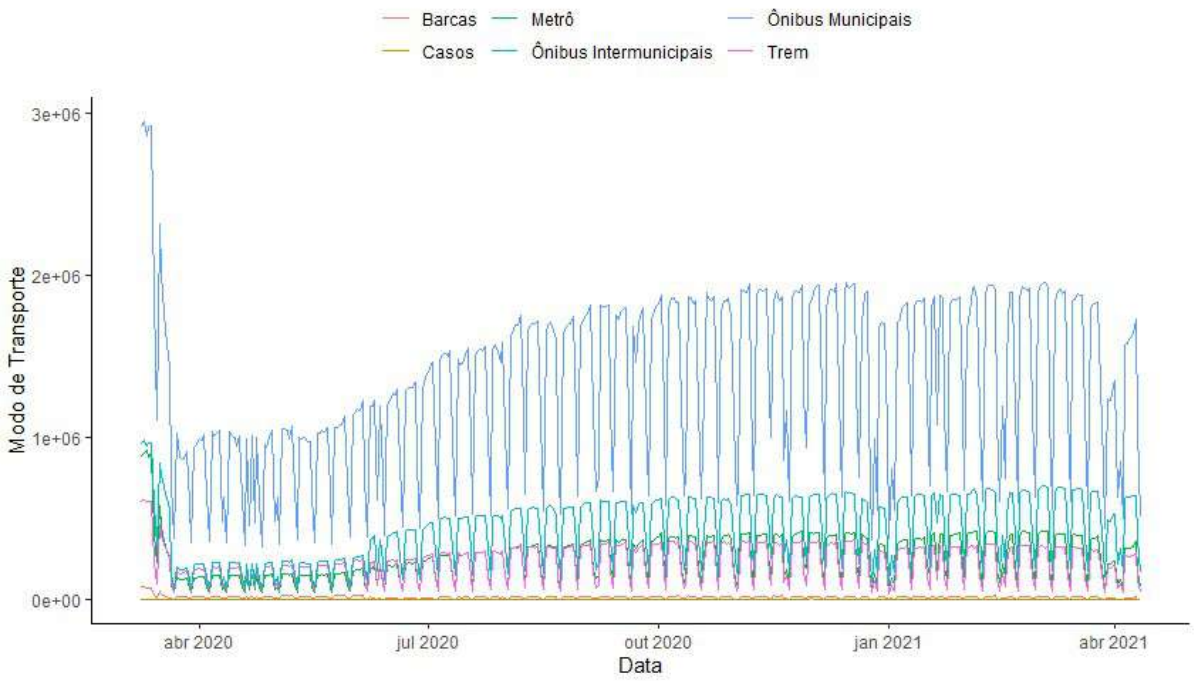

Figura 8.7. Evolução diária das variáveis

Para padronização dos gráficos de forma que seja possível analisar o comportamento dos mesmos foi aplicada a função scale() da base (Figura 8.8) para todas as variáveis exceto "Data", esta função executa o procedimento de centralizar em zero e alterar a escala para desvio padrão.

```
df_scale <- df
df_scale = data_frame("Data"          = df_scale$Data,
                      "Metro_Tot"     = scale(df_scale$Metro_Tot),
                      "Trem_Tot"      = scale(df_scale$Trem_Tot),
                      "Barcas_Tot"    = scale(df_scale$Barcas_Tot),
                      "Municipais"    = scale(df_scale$Municipais),
                      "Intermunicipais" = scale(df_scale$Intermunicipais),
                      "Casos"         = scale(df_scale$Casos)
                      )
```

Figura 8.8. Trecho do código Padronização das variáveis

A função ggplot() permite ainda a inclusão de marcos no gráfico. A Figura 8.9 apresenta a aplicação da função para os dados padronizados, incluindo ainda um marco representado por uma linha vettical, aplicado com a utilização do parâmetro "geom_vline" seguido do parâmetro "geom_text" que inclui a legenda no marco.

```
# Evolução diária padronizada para estudo do comportamento do gráfico
ggplot(df_scale, aes(x=Data, y=Metro_Tot))+
 geom_line(aes(col= "Metrô"))+
 geom_line(aes(y=Trem_Tot, col= "Trem"))+
 geom_line(aes(y=Barcas_Tot, col= "Barcas"))+
 geom_line(aes(y=Municipais, col= "Ônibus Municipais"))+
 geom_line(aes(y=Intermunicipais, col= "Ônibus Intermunicipais"))+
 geom_line(aes(y=Casos, col= "Casos"))+
 geom_vline(data = subset(df_scale, Data == "2020-03-16"),
         aes(xintercept = Data),       size = 0.3, colour = "black")+
 geom_text(data=subset(df_scale, Data == "2020-03-16"),
          mapping=aes(x=Data, y=0,    label= "Decretada situação de emergência"),
          size=4, angle=90, vjst=1, hjust=-0.7) +
 theme_classic()+  labs(x="Data",
     y="Modo de Transporte",
     color=NULL)+
 theme(legend.position = "top")
```

Figura 8.9 - Evolução diária padronizada para estudo do comportamento gráfico

O resultado da plotagem com os dados padronizados está apresentado na Figura 8.10. Neste, é possível observar uma queda bem definida após o início da pandemia, além disso, identifica-se um padrão semanal bem definido para todos os modos de transportes, com quebras pontuais que coincidem com feriados e recessos. Já os dados referentes aos casos de Covid-19 apresentam um comportamento mais caótico, onde visualmente não é possível identificar um padrão.

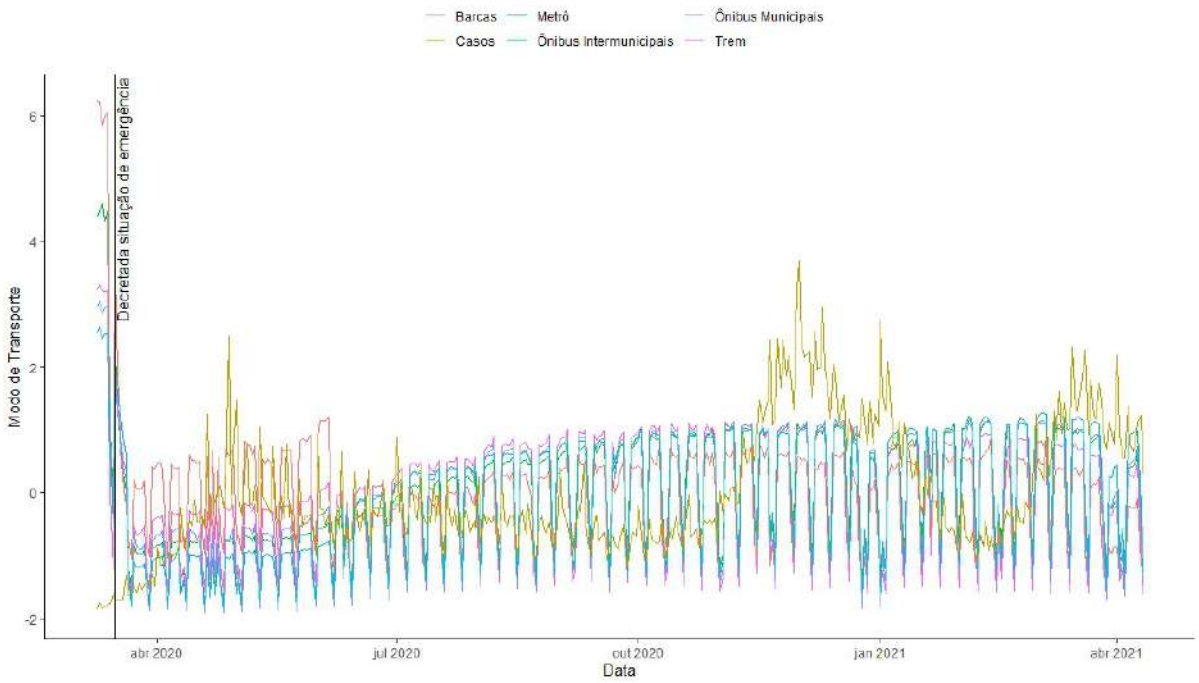

Figura 8.10. Gráfico de evolução diária com variáveis padronizadas

Em uma análise utilizando dados que não se apresentam no formato de uma série temporal, seria possível elaborar uma matriz de correlação entre todas as variáveis utilizando a Correlação de Pearson. Apesar de não ser adequada para este exemplo, esta será aplicada apenas a título de ilustração. A Figura 8.11 apresenta o procedimento para a criação da matriz, para tal, utiliza-se a função cor () da biblioteca "stats" para cálculo das correlações e, em seguida, estas são aplicadas na função ggcorrplot () da biblioteca "ggcorrplot" para a elaboração da matriz.

```
# Elaboração da Matriz de Correlação de Pearson

x = df_scale[,c("Metro_Tot","Trem_Tot","Barcas_Tot","Municipais",
        "Intermunicipais","Casos")]

matriz <- round(cor(x),2)
ggcorrplot(matriz,type = "lower",
        lab = TRUE,
        lab_size = 3,
        colors = c('firebrick',"white", 'dodgerblue4'),
        title = "Matriz de Correlação de Pearson",
        ggtheme = theme_classic)
```

Figura 8.11.Trecho do código Criação da Matriz de Correlação Pearson

A Figura 8.12 apresenta o resultado da matriz de correlação, nesta é possível observar uma correlação forte entre todos os modos de transportes, como era de se esperar pelo comportamento visual do gráfico disposto na Figura 8.10. É possível supor que a correlação de Pearson ilustra bem a correlação dos modos de transportes entre sí porque espera-se que esta ocorra no mesmo dia, sem defasagem de tempo. Por outro lado, a correlação encontrada entre o número de casos e todos os modos de transportes apresenta-se abaixo de 0,2, o que em geral se considera um resultado muito fraco para ser considerado. Neste caso, não é possível esperar um resultado fidedigno da correlação de Pearson uma vez que uma variável pode influenciar na outra em tempos diferentes, por exemplo, uma pessoa infectada em um modo de transporte no dia x só deve apresentar sintomas no dia x + t, sendo t o tempo desde a infecção até o aparecimento dos sintomas. Dessa forma, cabe a aplicação da correlação cruzada para identificar correlação em qualquer período de tempo, como será apresentado no item 1.3.3.

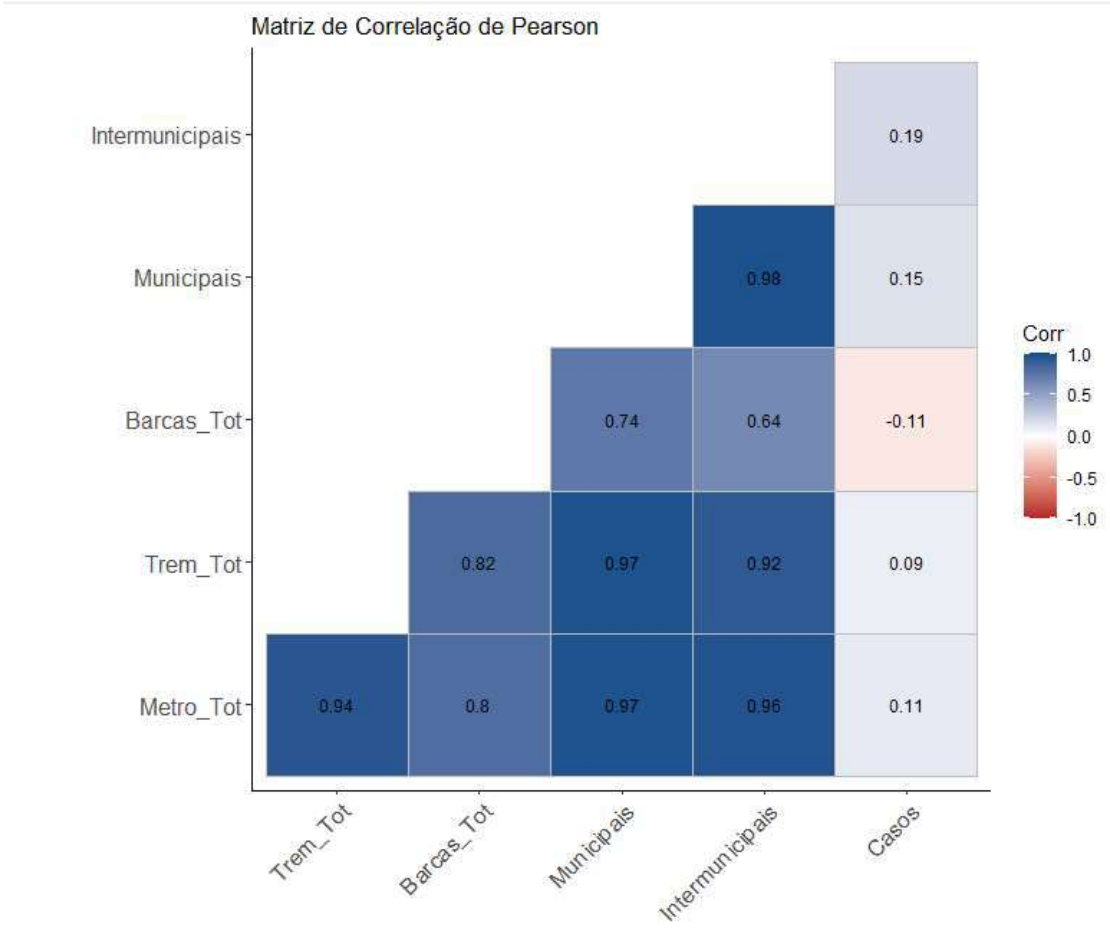

Figura 8.12 - Matriz de Correlação Pearson

8.4.3 Aplicação da Correlação Cruzada

Para a aplicação da correlação cruzada, a variável a ser comparada com o número de casos será "Total", que engloba a soma por dia de todos os modos de transportes. Por se tratar de uma série temporal, algumas etapas devem ser cumpridas para estacionarização da série de forma que se possa correlacionar as duas variáveis de maneira adequada. Em um primeiro momento, utiliza-se a função ts() para criar o objeto de série temporal como disposto na Figura 8.13, as variáveis utilizadas para comparação serão "Total" que engloba a soma de todos os modos de transportes e "Casos". Como a base de dados está organizada por dia, a frequência da série temporal, indicada pelo parâmetro "frequency", deve ser definida por 7, indicando um período natural de uma semana. O resultado da plotagem das duas séries temporais está apresentado na Figura 8.14.

```
# Séries Temporais

series <- df[,c("Data","Total","Casos")]

ts_total = ts(series$Total,frequency = 7)
ts_casos = ts(series$Casos,frequency = 7)

plot(cbind(ts_total,ts_casos), main="Séries Temporais")
```

Figura 8.13. Trecho do Código Séries Temporais

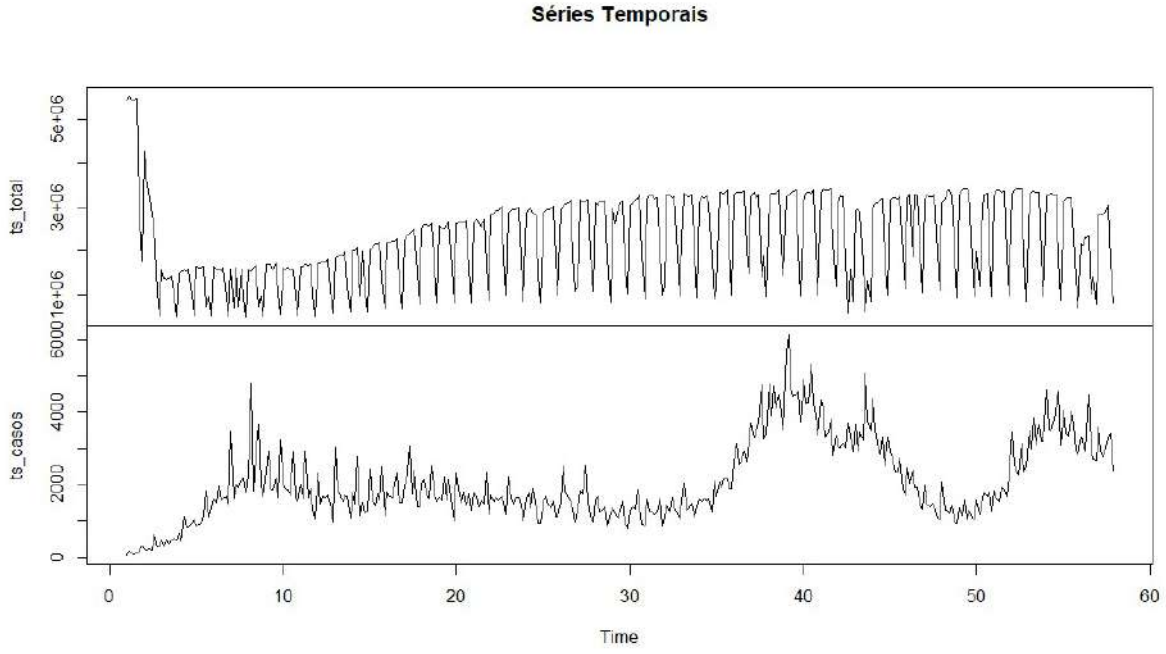

Figura 8.14. Gráfico Séries Temporais

A etapa seguinte consiste decompor as séries temporais com o objetivo de separar a tendência e a sazonalidade das mesmas, para tal, pode ser aplicada a função decompose() da biblioteca “stats” como apresentado na Figura 8.15. Como a variação sazonal apresenta constância ao longo do tempo, o parâmetro “type” recebe o argumento 'additive' relativo ao modelo aditivo, do contrário este deve receber o argumento ‘multiplicative’ referente ao modelo multiplicativo onde a variação sazonal aumenta ao longo do tempo.

```
# Remoção de tendência e de sazonalidade

decompose_total = decompose(ts_total, type = 'additive')
plot(decompose_total)
r_total = decompose_total$random

decompose_casos = decompose(ts_casos, type = 'additive')
plot(decompose_casos)
r_casos = decompose_casos$random

plot(cbind(r_total,r_casos), main=" ")
```

Figura 8.15. Trecho do Código remoção de tendência e sazonalidade

As Figuras 8.16 e 8.17 apresentam ,respectivamente, os resultados das plotagens da função decompose() aplicada à série temporal relativa ao total de transportes públicos e à série temporal relativa aos casos de Covid-19. A saída da função apresenta 4 gráficos, o primeiro denominado “observed” apresenta a série temporal original, o segundo denominado “trend” apresenta a tendência da série, o gráfico “seasonal” representa a sazonalidade da mesma e o último, “random” consiste no resultado final já randomizado com a remoção da tendência e da sazonalidade. Dessa forma, a série gerada como “random” para cada variável (Figura 8.18) passará a ser a série de entrada para a correlação, as séries randomizadas para “Total” e “Casos” foram armazenadas nas variáveis “r_total” e ‘r_casos”, respectivamente.

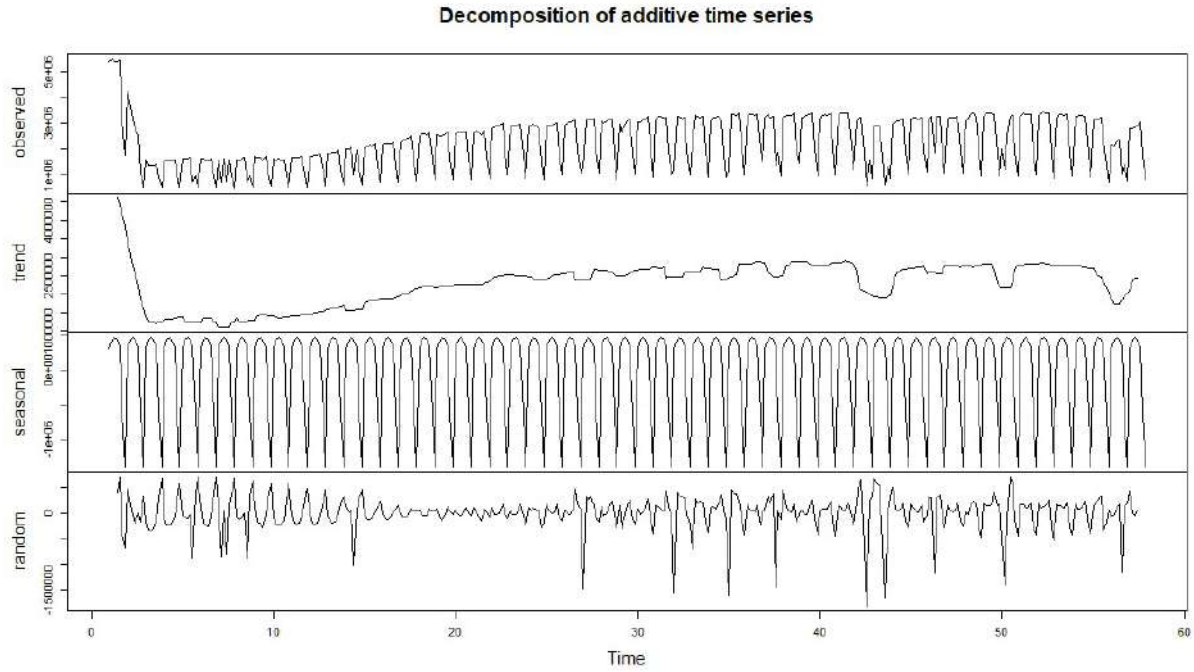

Figura 8.16. Gráfico Decomposição variável Total

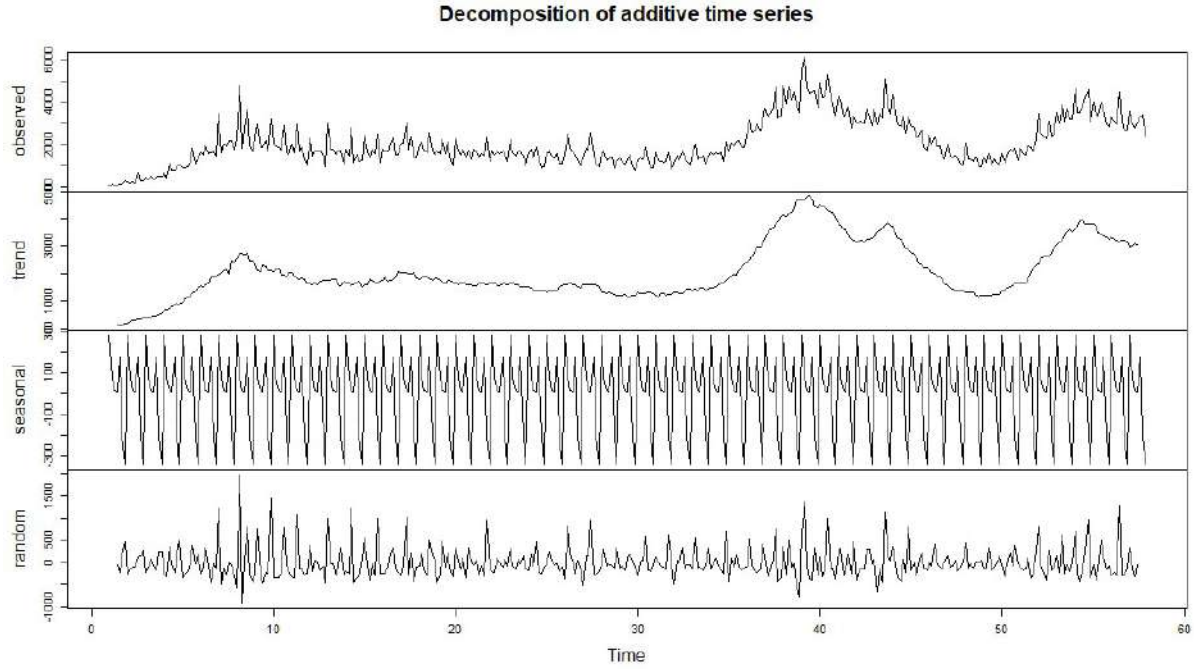

Figura 8.17. Gráfico Decomposição variável Casos

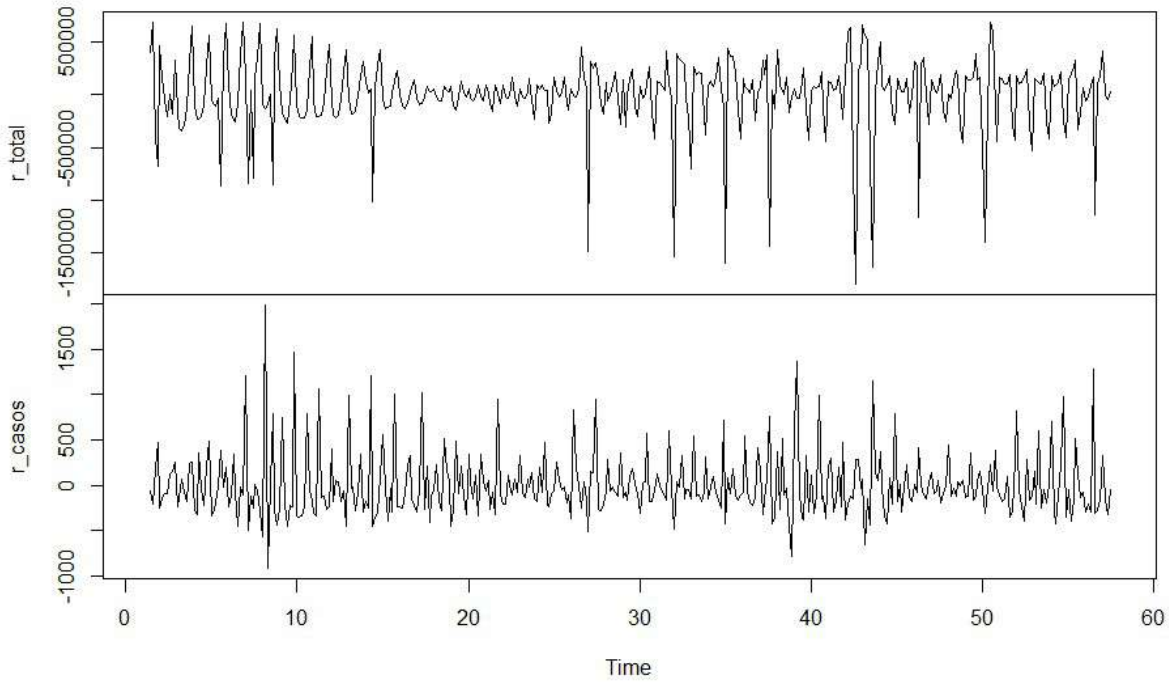

Figura 8.18. Gráfico Séries temporais Randomizadas

Como as variáveis a serem estudadas são séries temporais, a correlação mais adequada a ser aplicada é a correlação cruzada, nesta é possível identificar a correlação para diferentes *lags* (atrasos). No R, a função "ccf ()" da biblioteca "stats" permite a análise da correlação cruzada, nesta, as entradas assumem o formato ccf (x,y), onde x representa a variável líder e y a variável dependente. Nesta aplicação, em um primeiro momento, a variável r_total referente aos transportes públicos foi definida como líder e a variável r_casos, referente aos casos de Covid-19, como dependente (esta hipótese poderá ser confirmada após o resultado da correlação como será apresentado a seguir). A Figura 8.19 apresenta a aplicação da função ccf () seguida da definição do *lag* e correlação máximos, bem como da conversão do lag para o número de dias.

```
# Correlação Cruzada

ccf = ccf(r_total,r_casos, type = "correlation",
        na.action = na.pass,
        ylab = "Correlação",
        main = "Correlação Cruzada")

cor = ccf$acf[,,1]
lag = ccf$lag[,,1]
res = data.frame(cor,lag)
lag_max = res[which.max(res$cor),]$lag
lag_max
cor_max = res[which.max(res$cor),]$cor
cor_max
dias = lag_max *7
dias
```

Figura 8.19. Trecho do Código Correlação Cruzada

A Figura 8.20 apresenta o resultado da correlação cruzada, as linhas tracejadas azuis representam o intervalo de confiança calculado automaticamente pela função ccf () em função do tamanho da amostra e o *lag* máximo, os valores de correlação para além desse intervalo são considerados significativos. O *lag* negativo indica que a definição de variável líder foi correta, de forma que os valores referentes ao início de sintomas de covid estão relacionados com o uso do transporte público em um tempo anterior. Se o *lag* máximo fosse positivo, indicaria que a hipótese estava incorreta e as variáveis definidas inicialmente como líder e dependente deveriam ser invertidas. A correlação máxima foi de aproximadamente 0,12 identificada em um *lag* = - 0.8571429, como os intervalos são de 7 dias, o valor do *lag* em dias corresponde a - 6 dias.

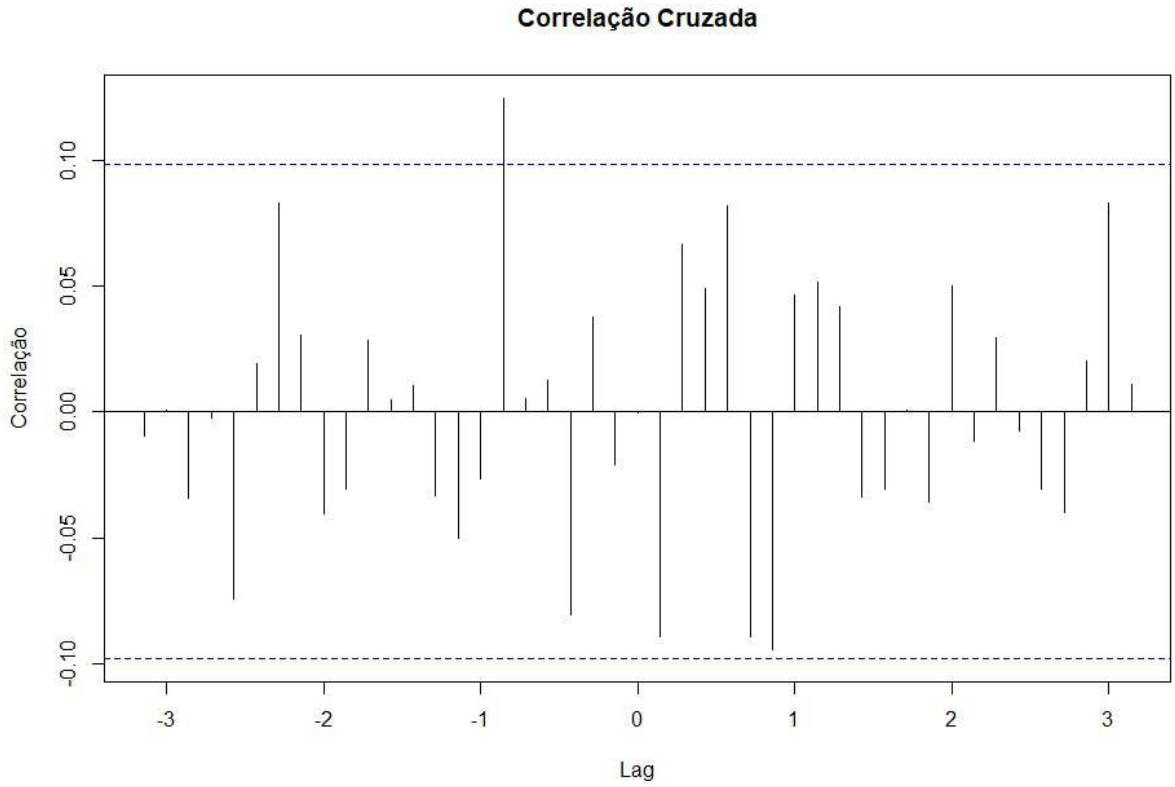

Figura 8.20. Gráfico Correlação Cruzada pós Estacionarização

A título de comparação, a Figura 8.21 apresenta a correlação aplicada entre as séries temporais antes da remoção da sazonalidade e da tendência, nesta pode-se observar diversos valores de correlação em diferentes *lags* tanto positivos quanto negativos, o que indica que a tendência e a sazonalidade das séries estão poluindo os dados, impedindo uma estimativa adequada das correlações. Cabe ressaltar que neste caso os *lags* máximos são positivos, o que acarretaria em uma interpretação incorreta da variável líder, reiterando a importância de uma série estacionária para aplicação da correlação cruzada.

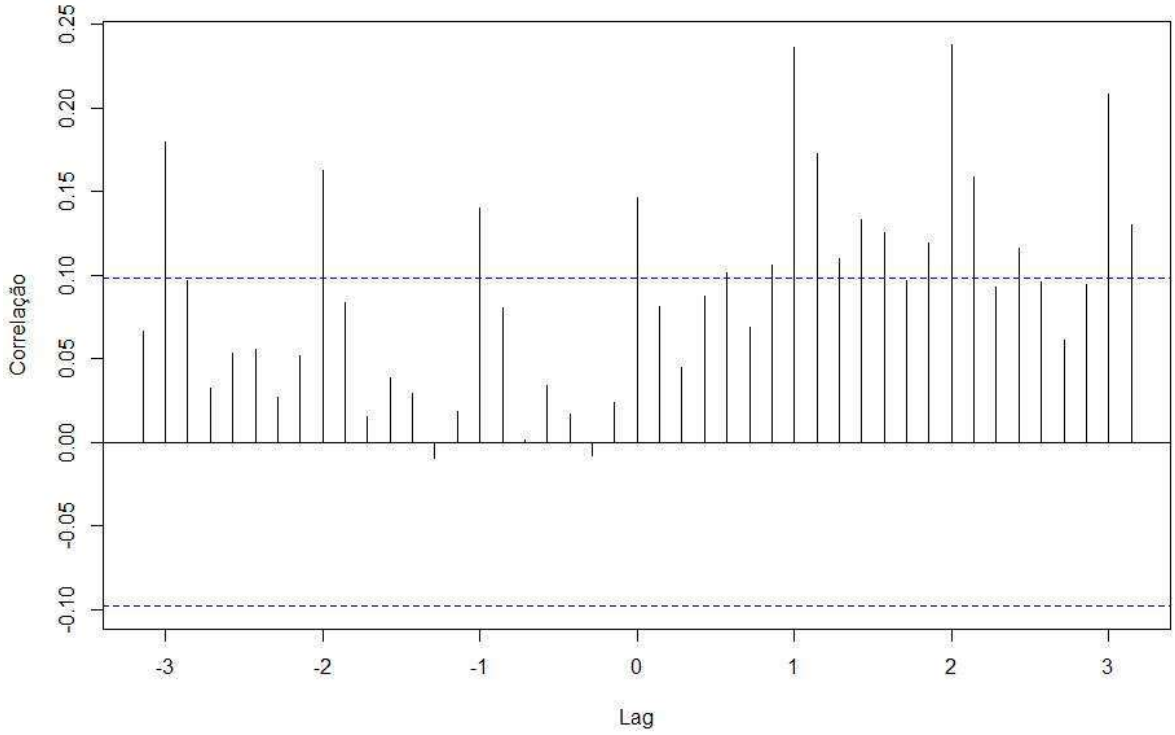

Figura 8.21. Gráfico de Correlação Cruzada com Tendência e Sazonalidade

Os resultados da Figura 8.21 indicam que existe uma correlação entre os dados de utilização de transportes públicos e os dados de contaminações por Covid-19, o *lag* indica que esta correlação de dá com 6 dias de antecedência, ou seja, um aumento de utilização de transportes públicos influencia no aumento do número de casos 6 dias depois. Em uma projeção para a realidade, algumas pesquisas de identificação do padrão de evolução da pandemia estimam que os sintomas da Covid-19 podem aparecer de 2 a 14 dias após o contágio, onde a maioria se encaixa no intervalo de 5 a 7 dias, dessa forma, o atraso de 6 dias encontrado está de acordo com o intervalo esperado.

8.5.Considerações Finais

Este capítulo teve como objeto principal apresentar o uso da linguagem R para análise de dados, com aplicação em séries temporais com dados de transportes públicos e casos de Covid-19. Uma vez que utilizamos base de dados reais, foi possível, além demonstrar técnicas estatísticas e como é a estrutura exigida pela ferramenta 'R', apresentamos resultados importantes para tomada de decisão pública.

Na etapa de pré-processamento, ao harmonizar-se os dados das duas bases recebidas em uma perspectiva integrada, informações sazonais como períodos de final de semana e feriados foram retiradas pois ofereciam distorções nos resultados.

Com a distribuição das informações em formato gráfico e a análise do comportamento da evolução temporal identifica-se um padrão semanal bem definido para todos os modos de transportes, com quebras pontuais que coincidem com feriados e recessos. Entretanto, os dados referentes aos casos de Covid-19 apresentam um

comportamento mais caótico, onde visualmente não é possível identificar um padrão. Desta forma, foi descartado a utilização da Correlação de Pearson pois a tendência e a sazonalidade das séries poluíram os dados, impedindo uma estimativa adequada das correlações.

Contudo, os resultados da correlação cruzada indicaram que existe uma correlação entre os dados de utilização de transportes públicos e os dados de contaminações por Covid-19, o *lag* indica que esta correlação de dá com 6 dias de antecedência, ou seja, um aumento de utilização de transportes públicos influencia no aumento do número de casos 6 dias depois. Esta descoberta vai ao encontro dos dados da Secretaria de Saúde do Estado do Espírito Santo que apontam que o período médio de incubação por coronavírus é de 05 dias [SESA, 2020].

Espera-se que este capítulo sirva de fonte para que mais pesquisadores e alunos que trabalham com análise de dados, ciência de dados ou pesquisa em geral construam uma base de conhecimento que facilite a evolução de futuras pesquisas.

8.6. Referências Bibliográficas

Battist, I. D. E. and Smolski, F. M. S. (2019) "Software R: Análise estatística de dados utilizando um programa livre", Editora Faith,Bag é, RS, 2019.Disponível em: <http://www.editorafaith.com.br/ebooks/grat/978-85-68221-44-0.pdf>.

Britto Dalson; Filho, Figueiredo; Alexander, José; et al. Desvendando os Mistérios do Coeficiente de Correlação de Pearson (r) . Revista Política Hoje, v. 18, n. 1, 2009. disponível em: <https://periodicos.ufpe.br/revistas/politicahoje/article/viewFile/3852/3156>.

EHLERS, Ricardo S. Análise de séries temporais. Laboratório de Estatística e Geoinformação. Universidade Federal do Paraná, v. 1, p. 1-118, 2007. Disponível em: <https://www.icmc.usp.br/pessoas/ehlers/stemp/stemp.pdf>

FIGUEIREDO FILHO, D. B.; SILVA JÚNIOR, J. A. Desvendando os Mistérios do Coeficiente de Correlação de Pearson (r). Revista Política Hoje. Disponível em: <https://periodicos.ufpe.br/revistas/politicahoje/article/viewFile/3852/3156>.

Lima, G. C. L. S., Schechtman, R., Brizon, L. C., Figueiredo, Z. M. (2020). Transporte público e COVID-19. O que pode ser feito?. Rio de Janeiro. Centro de Estudos em Regulação e Infraestrutura da Fundação Getúlio Vargas (FGV CERI).

Organização Mundial da Saúde- OMS (2020). Coronavirus disease 2019 (COVID19)-SITUATION REPORT 51. Acesso em: 08 de março de 2021. Disponível em: <https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200311-sitrep-51-covid-19.pdf>

Phelan, A. L., Katz, R., Gostin, L. O. The Novel Coronavirus Originating in Wuhan, China: Challenges for Global Health Governance, JAMA 2020, DOI:10.1001/jama.2020.1097

Secretaria de Saúde Estado do Espírito Santos - SESA (2020). Boletim Epidemiológico. Acesso em: 18 de julho de 2021. Disponível em: <https://coronavirus.es.gov.br/>

Secretaria de Saúde Estado do Rio de Janeiro - SESA-RJ (2021). Boletim Epidemiológico. Acesso em: 30 de julho de 2021. Disponível em: <http://sistemas.saude.rj.gov.br/tabnetbd/dhx.exe?covid19/esus_sivep.def>

SILVA FILHO, Aloísio Machado da. Autocorrelação e correlação cruzada: teorias e aplicações. 2014. Disponível em: <http://200.9.65.226/handle/fieb/766>