Curso de Vigilância em Saúde apoiado por plataforma BI

Aula 06- Analisando séries temporais com médias móveis

1. Apresentação

Olá!

Seja bem-vindo à quarta aula do curso de Vigilância em Saúde apoiado por plataforma *Business Intelligence* (BI)! Nesse curso estamos aprendendo alguns conceitos de Vigilância Epidemiológica apoiada por ferramentas de BI para análise de dados temporais.

O objetivo do nosso curso é ajudar a análise epidemiológica em um contexto de vigilância em serviços de saúde, utilizando ferramentas que facilitem a vida dos profissionais de saúde.

Nessa aula aprenderemos a calcular uma média móvel a partir de dados de dengue do Distrito Federal.

2. Introdução

Existem diversas dificuldades em lidar com dados temporais. Diante de alguns problemas de registro, digitação e sistemas de informação, podemos utilizar técnicas estatísticas para contornar problemas com os dados temporais. Essas técnicas podem auxiliar a realizar uma análise mais acurada e propor de forma mais assertiva ações de vigilância em saúde.

Em um cenário de pandemia, por exemplo, podemos enfrentar barreiras com a velocidade dos dados, resultando em uma distribuição temporal que não corresponde à realidade. Isso significa que poderemos ter um acúmulo sazonal de dados em determinados dias da semana e uma redução em outros dias, por causa do fluxo de informação dos sistemas de informação em saúde.

Em um caso de distribuições de dados temporais prejudicadas, podemos utilizar uma série de técnicas de suavização temporal. Essas técnicas se propõem a suavizar os dados, ou seja, redistribuir no tempo concentrações de dados que empiricamente não se comportam da mesma maneira. Nesse sentido, aplicaremos uma técnica chamada médias móveis.

3. Médias Móveis

Uma média móvel é uma técnica de suavização estatística que pode ser utilizada em dados temporais ou em dados espaciais. Essa técnica consiste em calcular uma média para um período redistribuindo as unidades no período em que foram registradas. Assim, todos os registros de um período x são somados e posteriormente divididos pelo número de unidades de tempo consideradas na análise (1).

Nem todos os dados necessitam de um tratamento por séries temporais, sendo que seu uso pode ser aplicado apenas a determinados dados. Quando se tem confiança nos dados analisados, o uso de médias móveis não faz sentido. Em um cenário em que você tem a certeza de que os dados de início dos sintomas estão corretos, por exemplo, é

preferível utilizar os dados reais de início dos sintomas para as análises de saúde.

Em casos em que há a suspeita de que os dados não foram distribuídos corretamente no tempo, a análise pode ser auxiliada pelo cálculo de médias móveis. Por exemplo, quando não temos acesso aos dados de início dos sintomas, podemos usar a data do registro e suavizar, considerando que o início dos sintomas deve ter ocorrido em um período adjacente. Quando temos lacunas nos dados temporais também podemos usar médias móveis, por exemplo, quando não há registro nos finais de semana.

4. Criando Médias Móveis com o R

Você deve ter aberto o R pelo arquivo Rproject que está disponível em nossa plataforma. Caso não se lembre de como realizar essa operação, retorne à aula 4 com as instruções detalhadas.

Para calcular a média móvel no R, vamos iniciar carregando pacotes. Primeiro, vamos instalar o pacote *Pacman*, que permite a instalação e carregamento de quaisquer pacotes que estejam no CRAN.

Após isso, vamos carregar os pacotes necessários para a análise da média móvel. Se os pacotes não estiverem instalados, a função *p_load* (do pacote *Pacman*) se encarregará de instalar esses pacotes.

```
#instalando e carregando pacotes
install.packages("pacman")

#verificando instalação de pacotes e carregando pacotes necessários
pacman::p_load(
  tidyverse,  # manipulação de dados
  slider,     # manipulando intervalos de tempo
  tidyquant,  # manipulando médias móveis
  read.dbc    # importando dados em dbc
)
```

Para esse exercício, nós vamos utilizar os dados de internação por dengue no Distrito Federal em 2020. Estamos usando dados de janeiro de 2020 até março de 2021 porque o SIHSUS tem um *delay* de três meses no registro dos dados. Para importar os dados, vamos lista-los e usar a função *for* para ler cada um dos arquivos e importá-los em um único *data.frame*.

```
#importando dados de internação

files          =          cbind(paste0("dados/sih_df_2020/",
list.files("dados/sih_df_2020/", pattern = "\\.dbc$")))

sihdf = NULL

for (i in 1:length(files)) {
  sihdf_temp = read.dbc(files[i])
  sihdf= data.frame(rbind(sihdf, sihdf_temp))
}

#visualizando variáveis
glimpse(sihdf)
```

Nós usamos a função *glimpse* (*Tidyverse*) para visualizar todas as variáveis que estão no *data.frame* que importamos. Depois disso, nós vamos usar as variáveis de diagnóstico principal (DIAG_PRINC) e data de internação (DT_INTER) para selecionar os casos de internação por dengue. Para usar a variável de data de internação, primeiramente vamos convertê-la em um formato de data reconhecida pelo R.

```
#separando dados de internação por dengue
sihden   =   subset(sihdf,   sihdf$DIAG_PRINC   ==   "A90"|
sihdf$DIAG_PRINC == "A91")
sihden$data = as.Date(sihden$DT_INTER, "%Y%m%d")
sihden  =  subset(sihden,  sihden$data  >=  "2020-01-01"  &
sihden$data <= "2020-12-31")
```

Após isso, nós vamos ciar um novo *data.frame*, que será utilizado para criamos nossa série temporal. Ele vai contar o número de internações por dia.

```
#criando data.frame de internações por dia
intden <- sihden %>%
  count(data, name = "internacoes")
```

Para compreender a distribuição dos dados, vamos usar a função *summary*, que mostrará as medidas descritivas de dispersão e tendência central das datas.

```
#cortando dias de internações para um ano
summary(intden$data)
```

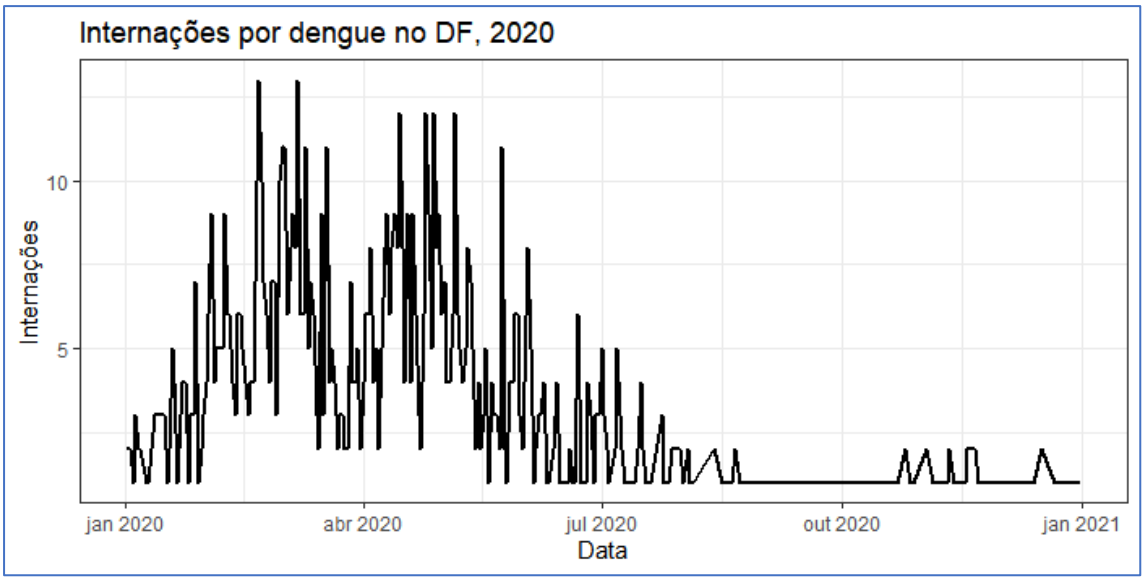
Nós receberemos um *output* com a data mínima e máxima a média e mediana das datas:

```
> summary(intden$data)
        Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
"2020-01-02" "2020-03-08" "2020-05-13" "2020-05-26" "2020-07-24" "2020-12-31"
>
```

Após criar esses dados e verificar que nossos dados realmente são do ano de 2020, vamos criar um gráfico com os casos de internação de dengue por dia, usando o *ggplot*.

```
#criando graficos com internacoes por dia
ggplot(data = intden)+
  geom_line(mapping = aes(x = data, y = internacoes), size
= 1)+
  ggtitle("Internações por dengue no DF, 2020")+
  theme_bw()+
  ylab("Internações")+
  xlab("Data")
```

Vamos receber um gráfico de linhas com os dados por dia. Perceba que no gráfico há muitas lacunas, como se não houvesse internações em determinados períodos. Suspeitamos que esses dados estejam assim por causa do registro, então vamos suavizar usando média móvel.



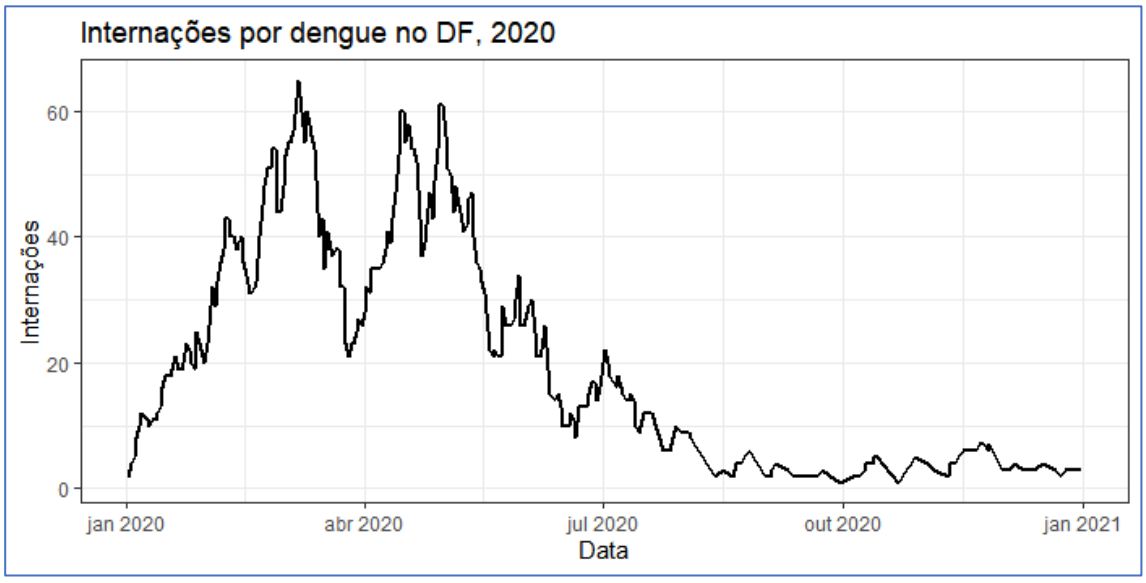
Nós vamos usar uma média móvel de 7 dias para preencher as datas que ficaram vazias. Usamos o parâmetro *before* como 6, pois estamos considerando os 6 dias posteriores para a criação da média móvel.

```
#preenchendo dias que não tiveram internações
pre <- intden %>%
  mutate(
    ma_7dias = slide_index_dbl(
      internacoes,                  # calculando internações
      .i = data,                      # indexando dados
      .f = ~sum(.x, na.rm = TRUE),     # somando valores e
considerando datas perdidas
      .before = days(6))               # criando média movel
de 7 dias para dias
  )
```

Após isso, vamos plotar novamente o gráfico para visualizar como ficou utilizando as médias móveis.

```
#gráfico de médias móveis
ggplot(data = pre)+
  geom_line(mapping = aes(x = data, y = ma_7dias), size =
1)+
  ggtitle("Internações por dengue no DF, 2020")+
  theme_bw()+
  ylab("Internações")+
  xlab("Data")
```

Após isso receberemos o seguinte gráfico. A distribuição ficou melhor distribuída e parece mais condizente com a realidade.



Nós vamos calcular as médias móveis por grupo de idade. Isso pode ser bem útil quando precisarmos lidar com subcategorias dos dados. Nesse caso, usaremos os grupos de idade, mas poderíamos usar estabelecimentos de saúde, territórios, sexo entre outros. A ideia é calcular uma média móvel para cada grupo.

Inicialmente, criaremos 4 grupos baseados em sua idade, usando a variável IDADE do SIH-SUS. Com o mesmo *data.frame*, vamos

```
#calculando média móvel por grupo de idade
sihden = sihden %>% mutate(age =
                             case_when(
                               IDADE < 10 ~ "Crianças",
                               IDADE  >9  &  IDADE  <  20  ~
"Adolescente",
                               IDADE  >19  &  IDADE  <  60  ~
"Adulto",
                               IDADE >59 ~ "Idoso"
                                        ))
```

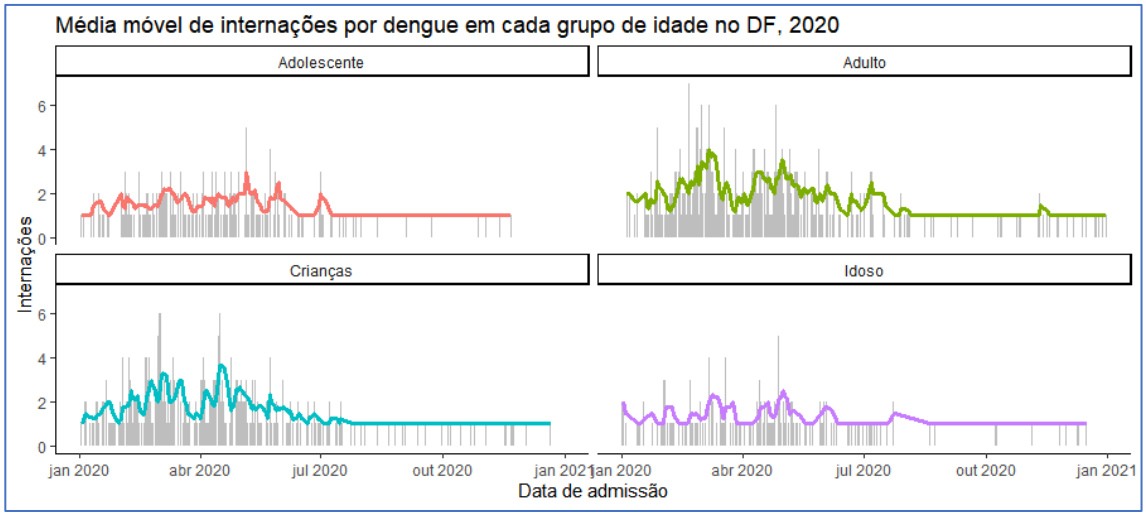
Após isso, vamos criar as médias móveis agrupadas por idade, também considerando uma média móvel de 7 dias.

```
#calculando média móvel por grupo
pretype <- sihden %>%

  count(age, data, name = "internacoes") %>%

  arrange(age, data) %>%           # organizando linhas por
grupo de idade

  group_by(age) %>%                # agrupando por grupo de
idade

  mutate(                          # calcualndo média móvel
por 7dias
    ma_7days = slide_index_dbl(
      .x = internacoes,              # contando o número de
casos por dia
      .i = data,                    # indexando por data de
admissão
      .f = mean,                       # usando a média de
valores
      .before = days(6)            # criando média movel de
7 dias para dias
    )
  )
```

Vamos criar um gráfico para isso, com quatro gráficos para cada um dos grupos criados.

```
#gerando gráficos de média móvel por grupo de idade
ggplot(data = pretype)+
  geom_col(                     # plotando os casos diários
em cinza
    mapping = aes(
      x = data,
      y = internacoes),
    fill = "grey",
    width = 1)+
  geom_line(                        # plotando os casos por
tgrupo de idade
    mapping = aes(
      x = data,
      y = ma_7days,
      color = age),
    size = 1.2)+
  facet_wrap(~age, ncol = 2)+       # criando mini gráficos
por grupo de idade
  theme_classic()+                  # mudando o tema
  theme(legend.position = "none")+ # removendo legenda
  labs(                               # incluindo título do
gráfico e dos eixos
    title = "Média móvel de internações por dengue em cada
grupo de idade no DF, 2020",
    x = "Data de admissão",
    y = "Internações")
```

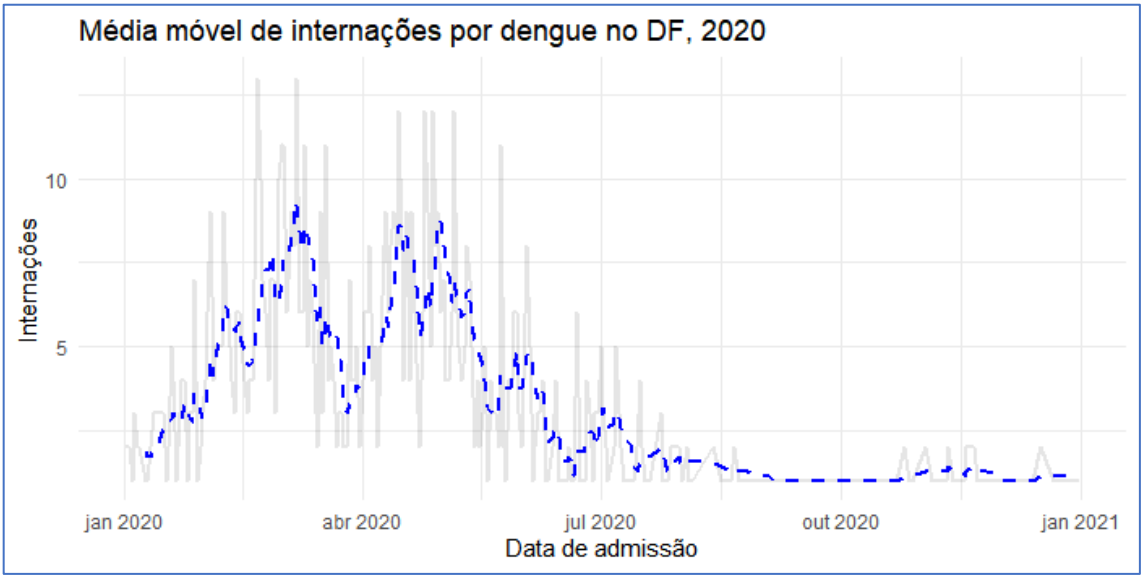
O gráfico obtido apresentará as médias móveis para cada um dos grupos que nós criamos com base na idade.



Por fim, vamos criar um gráfico que compara a média móvel diretamente com o gráfico original, permitindo comparar os dois gráficos. Nesse gráfico, nós criaremos a média móvel diretamente na visualização.

```
#calculando média móvel diretamente na criação do gráfico
sihden %>%
  count(data) %>%                          # contando número de
casos por dia
  drop_na(data) %>%
  ggplot(aes(x = data, y = n))+
  geom_line(
    size = 1,
    alpha = 0.1                          # colocando transparencia
na linha
  )+
  tidyquant::geom_ma(                      # plotando média móvel
    n = 7,
    size = 1,
    color = "blue")+
  theme_minimal()+
  labs(                                    # incluindo título do
gráfico e dos eixos
    title = "Média móvel de internações por dengue no DF,
2020",
    x = "Data de admissão",
    y = "Internações")
```

A visualização permitirá comparar as internações com a média móvel criada, conforme a imagem.



5. Referências Bibliográficas

1. NIELSEEN, Alieen. Análise Prática de Séries Temporais: Predição com estatística e aprendizado de máquina. Alta Books, Rio de Janeiro, 2021.

Atividades

1	Não há técnicas capazes de a realizar uma análise mais acurada e propor de forma mais assertiva ações de vigilância em saúde.	V	
		F	X
2	Existem diversas dificuldades em lidar com dados temporais.	V	X
		F	
3	Diante de alguns problemas de registro, digitação e sistemas de informação, podemos utilizar técnicas estatísticas para contornar problemas com os dados temporais.	V	X
		F	
4	Em um caso de distribuições de dados temporais prejudicadas não podemos utilizar uma série de técnicas de suavização temporal	V	
		F	X
5	Em um cenário de pandemia, por exemplo, podemos enfrentar barreiras com a velocidade dos dados, resultando em uma distribuição temporal que não corresponde à realidade	V	X
		F	
6	As técnicas de suavização não podem redistribuir no tempo concentrações de dados que empiricamente não se comportam da mesma maneira.	V	
		F	X
7	Uma média móvel é uma técnica de suavização estatística que pode ser utilizada em dados temporais ou em dados espaciais.	V	X
		F	
8	Média móvel consiste em calcular a dispersão para um período redistribuindo as unidades no período em que foram registradas	V	
		F	X
9	Em um caso de distribuições de dados temporais prejudicadas, podemos utilizar uma série de técnicas de suavização temporal	V	X
		F	
10	Nem todos os dados necessitam de um tratamento por séries temporais, sendo que seu uso pode ser aplicado apenas a determinados dados.	V	X
		F	
11	Todos os dados precisam passar por uma técnica de suavização de média móvel.	V	
		F	X
12		V	X

	As técnicas de suavização redistribuem no tempo concentrações de dados que empiricamente não se comportam da mesma maneira.	F	
13	Quando se tem confiança nos dados analisados, o uso de médias móveis não faz sentido	V	X
		F	
14	Em um cenário em que você tem a certeza de que os dados de início dos sintomas estão corretos, por exemplo, é preferível utilizar os dados reais de início dos sintomas para as análises de saúde	V	X
		F	
15	Em casos em que os dados foram distribuídos corretamente no tempo, a análise pode ser auxiliada pelo cálculo de médias móveis.	V	
		F	X
		F	X